DOCTOR OF PHILOSOPHY

Automated Analysis of Digital Gonioscopy Images Using Deep Learning

Peroni, Andrea

Award date:
2022

Automated Analysis of Digital Gonioscopy Images Using Deep Learning

University of Dundee

Andrea Peroni

School of Science and Engineering

University of Dundee

This dissertation is submitted for the degree of

Doctor of Philosophy

November 2022

Declaration

Candidate's Declaration

I, Andrea Peroni, hereby declare that I am the author of this thesis; that I have consulted all references cited; that I have done all the work recorded by this thesis; and that it has not been previously accepted for a degree.

	30/11/2022
Signed	Date

Supervisor's Declaration

I, Emanuele Trucco, hereby declare that I am the supervisor of the candidate, and that the conditions of the relevant Ordinance and Regulations have been fulfilled.

	30/11/2022
Signed	Date

Acknowledgements

First and foremost I am extremely grateful to my supervisor Prof. Emanuele Trucco for his support and encouragement during my PhD study. His knowledge, passion and enthusiasm have always inspired and motivated me.

I want to thank the members of the CVIP and the Vampire research groups, both at the University of Dundee and at the University of Edinburgh, for their conviviality and for being always willing to share their experience and provide invaluable feedbacks.

I would like to thank Dr. Michael Crabb and Dr. Iain Martin for their precious assistance as members of my Thesis Monitoring Committee.

I want to express my gratitude to all the clinical collaborators, especially Dr. Stewart Gillan (Ninewells Hospital, Dundee, UK), Dr. Andrew Tatham (Princess Alexandra Eye Pavilion, Edinburgh, UK), Prof. Carlo Enrico Traverso (Clinica Oculistica, DINOGMI, University of Genoa, Genoa, Italy), Prof. Luís Abegão Pinto (Hospital de Santa Maria, Lisbon, Portugal) and their colleagues for their remarkably valuable involvement over the entire duration of this research project, for sharing and annotating the data and for the courtesy of providing the images shown in this thesis.

This research was made possible by an industrial grant from NIDEK Technologies Srl. (Albignasego, Italy). The company has not just acted as funder, its highly qualified members have always been available to share technical know-how and resources and to discuss research requirements and outcomes.

Finally, I would like to thank my family, my closer friends and those special people I was so lucky to spend most of the lockdown duration with for their continued support during this sometimes difficult, but overall wonderful and unforgettable experience.

Abstract

Glaucoma is a high prevalence optic neuropathy that may lead to irreversible blindness. The assessment of the anterior chamber angle of the eye is recommended by international guidelines to evaluate risk factors, categorise the disease and decide treatment strategies. The clinical-standard technique, called gonioscopy, is a difficult manual examination, seldom practised in primary care settings.

The NIDEK GS-1 digital imaging device has automated several phases of gonioscopy making it available to non-expert operators and allowing to store high quality pictures of the eye region. However, images of the angle still need extensive knowledge and time to be interpreted correctly to produce a diagnosis.

We aimed to research whether deep learning systems, currently studied in many other applications in medicine and ophthalmology, could effectively support the analysis of digital gonioscopic images, enabling patient screening in primary care settings. Experienced ophthalmologists have been involved in the collection and evaluation of clinical requirements to focus our work, leading to the selection of two main image analysis tasks to investigate.

The first is the semantic segmentation of anatomical layers in gonioscopic images. We designed and developed an algorithm capable of providing, for the first time, rich morphological information about the eye region. The algorithm can deal with specific image (e.g., peripheral blur and vignette) and ground truth (e.g., partial annotations) characteristics that would make other state-of-the-art systems ineffective, returning an overall segmentation accuracy above 90% on our test set. Moreover, the system may estimate the reliability of its results by generating uncertainty maps through *Monte Carlo dropout* that proved to be effective at detecting small segmentation flaws.

The second is the automatic grading of angle aperture, a measure of clinical relevance. Differently from existing literature, in which the angle aperture is estimated over large angle quadrants (90°-wide), we explored solutions to grade smaller angle regions (about 5°-wide) for an increased sensitivity at detecting local angle closures. Our results highlight potentials and limitations of this approach and provide useful hints for future development in this field.

Both development phases followed inter-domain sessions with clinical and technical collaborators for the formulation of annotation protocols and the generation of ground truth.

Moreover, we have studied inter-annotator variability of ground truth delineations of angle layers to provide a comprehensive context for evaluating the performance of automated systems. In particular, we compare the performance of our segmentation model with the average per-layer variability among experts finding good correlation. Results suggest that the identification and delineation of some of the anatomical layers of the angle is difficult even to experts and that, as a consequence, the limited agreement among annotators reflects on the algorithm's performance.

Despite the limited size of our annotated datasets, we demonstrated that deep learning systems can aid the analysis of digital gonioscopic images, possibly encouraging further research in this field. Semi-automated imaging devices and automatic analysis software may offer an effective and efficient alternative to conventional gonioscopy (the current clinical-standard) and help both prevent the development of glaucoma through better screening plans, and manage its treatment.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 About this Chapter

This chapter provides the motivation and the clinical background of this research, the research questions, the main contributions to the field and the structure of this thesis.

1.2 Motivation

Glaucomas are a family of optic neuropathies (i.e., diseases damaging the optic nerve) and are the second leading cause of irreversible blindness worldwide after cataract [84] affecting nowadays between 70 and 80 million people [94], a number that is predicted to considerably increase to more than 110 million within twenty years [125]. It is commonly referred to as *the silent thief of sight* since it often remains asymptomatic for years and is diagnosed only at a late stage of progression when some degree of irreversible visual impairment has already been caused. It has been estimated that, depending on the geographical area, 50% to 90% of the people affected by glaucoma may be unaware they have it [69, 99, 43, 105, 9]. Glaucoma reduces patients' quality of life [93] and impacts healthcare costs [128, 95]. It is thus fundamental to discover, evaluate and monitor risk factors in advance and diagnose the disease at the earliest stage possible. Several factors have been associated with higher chances of developing glaucoma. Among them, age, ethnicity, the morphology of the *anterior*

chamber angle of the eye (described in the next section) and an elevated *intraocular fluid pressure* (IoP) [135]. The assessment of the anterior chamber angle is particularly important and recommended by current international guidelines [31, 1].

The inspection of the anterior chamber angle is performed through an examination called *gonioscopy* [4], that will be better described in a later section of this chapter. It is however worth highlighting here that the conventional way to perform this exam has several known limitations. The current clinical standard technique for gonioscopy is a fully manual procedure which requires significant time, patient cooperation, and operator expertise [44, 23]. It does not enable to acquire images of the eye region (also known as *gonio-photographs*) easily, thus preventing comparisons and follow-up. The result is that gonioscopy is often performed less frequently than recommended [17] and is seldom practised by optometrists in primary-care settings. To overcome some of these disadvantages, few imaging devices for digital gonioscopy have been developed in recent years [87, 114].

Our work aimed to investigate solutions for the automated analysis of digital gonio-photographs to support the detection and evaluation of glaucoma-related conditions, considering requirements collected from clinical experts. It was motivated by the importance of gonioscopy, a recommended clinical practice related to a high-prevalence disease (glaucoma), and by the clinical need (common to many fields in medicine) for a more efficient analysis of data, which has not been tackled yet for gonioscopy by previous contributions in the literature.

The medical background of this research, with a description of the anterior chamber angle, of gonioscopy, and of complementary examinations to image this eye region, is briefly described in the next few sections to provide the reader with the information useful to contextualise the technical work.

1.3 The Anterior Chamber Angle

The anterior chamber angle (also called *irido-corneal angle* or *drainage angle*) is the region of the anterior chamber of the eye located at the interface between the iris and the cornea. It

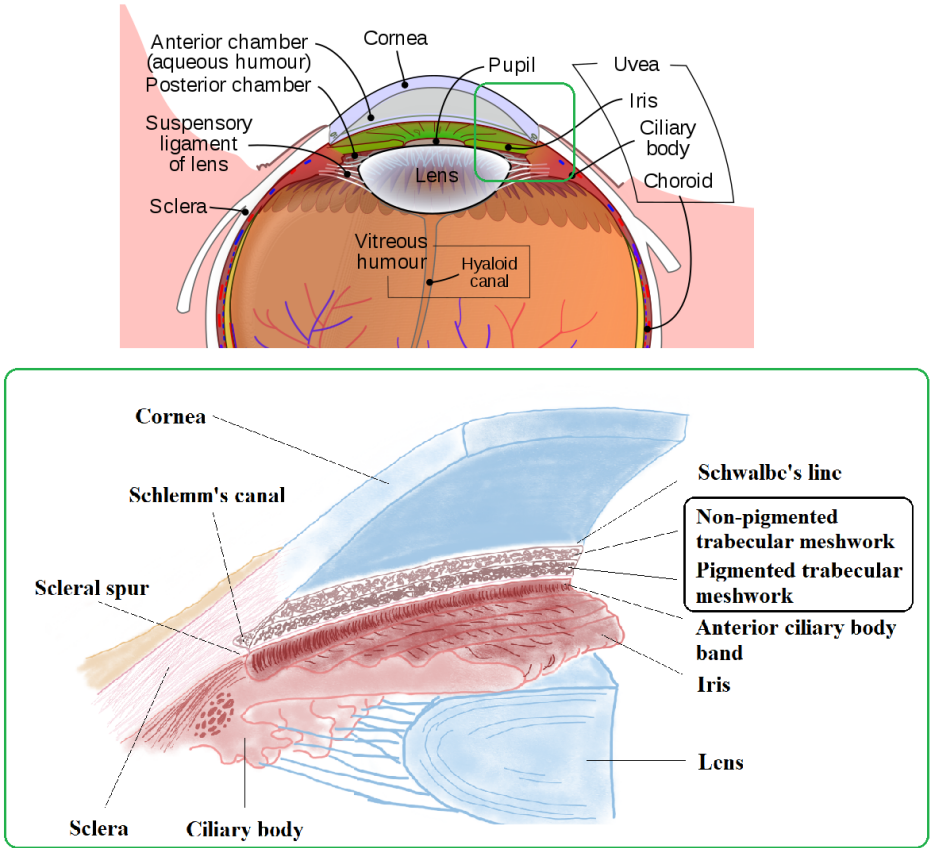

Fig. 1.1 *Top: a schematic representation of a human eye anterior section (adapted and reproduced under CC BY-SA 3.0, source: https://commons.Wikimedia.org/wiki/File:Schematic_diagram_of_the_human_eye_en.svg). Bottom: the anterior chamber angle as magnification of the area highlighted by the green rectangle.*

is composed of several anatomical *layers* that extend along the entire circular periphery of the iris. With reference to Figure 1.1, moving from the cornea towards the iris, we traverse the main structures of the angle:

- *cornea*: the transparent front of the eye that covers the iris, the pupil and the whole anterior chamber;
- *Schwalbe's line*: a thin layer made up of collagen that separates the cornea from the trabecular meshwork. It usually appears white, but it may be pigmented in some cases;
- *trabecular meshwork*: a mesh of collagen fibres and epithelium. It is divided into two parts, a non-functional and usually paler one (*anterior* or *non-pigmented* trabecular meshwork) and a functional pigmented one (*posterior* or *pigmented* trabecular

meshwork). The functional part of the trabecular meshwork is located in front of the Schlemm's canal;

- *Schlemm's canal*: a circular vessel that surrounds the entire drainage angle circumference along the border located behind the posterior trabecular meshwork and thus normally invisible in a gonioscopic exam;
- *scleral spur*: a ridge of white collagen fibres of the sclera, between the anterior ciliary body and the trabecular meshwork. It is usually relatively bright with respect to the neighbouring regions;
- *anterior ciliary body band*: longitudinal fibres of the ciliary body. Normally, it appears as a light brown or grey band;
- *iris root*: the peripheral region of the iris originated from the anterior ciliary body.

The layers described above are always present in the anterior chamber angle but, depending on the morphology of the region, some of them may not be visible during a gonioscopic examination. This happens for example when there are local adhesions of the iris root with the trabecular meshwork or the cornea, called *synechiae*, or when the iris root is very close (i.e., in *apposition*) to the cornea. The impossibility of seeing some of the layers through gonioscopy may relate with an increased risks of developing certain sub-types of glaucoma, as will be discussed shortly.

The trabecular meshwork is a layer with an important physiological role. It regulates the drainage of the *aqueous humour*, the fluid that fills the anterior chamber of the eye and that is continuously produced by the ciliary processes in the posterior eye chamber (a small space between the iris and the lens). The aqueous humour flows through the meshwork towards the Schlemm's canal and is then collected by the circulatory system. By allowing the correct amount of fluid to filter through, the trabecular meshwork is the structure mainly responsible for the control of the overall IoP.

Some pathological conditions may lead the trabecular meshwork to lose efficiency at draining the aqueous humour or prevent it from functioning properly. This might be due to

structural degradation or to some extent of physical occlusion, e.g. due to iris apposition. As previously mentioned, it is well known that an increase in IoP is one of the main risk factors of developing glaucoma [135], hence being able to see the anterior chamber angle and evaluate its conditions is important.

According to the characteristics of the drainage angle, two most common broad categories of glaucoma may be identified. They are:

- *open-angle glaucoma*: this is the most common type of glaucoma. It is caused by a slight imbalance between the average amounts of aqueous humour produced by the ciliary processes and that filtered by the trabecular meshwork. This imbalance makes the IoP increase slowly and painlessly and can take years to cause symptoms detectable by the patient;

- *angle-closure glaucoma*: it is caused by an extended coverage of the trabecular meshwork by the iris. The progression is often slow and painless (primary angle-closure), but sometimes may evolve suddenly and painfully, especially when caused by traumas or infections (secondary angle-closure). Though, worldwide, open angle glaucoma is more common, angle-closure glaucoma is responsible for a disproportionate number of patients with severe vision loss [94, 24].

In a study published in 2014 [125] the overall prevalence of glaucoma worldwide among people aged 40-80 years (2.33 billions in total in 2013, based on the *World Population Prospects*) was estimated to be the 3.54% (3.04% of open-angle and 0.5% of angle-closure glaucoma). However, when considering different geographical regions, these values may vary. For instance, the combined prevalence was estimated to be higher than average in Africa and Latin America (4.79% and 4.51% respectively) and lower in Europe and Oceania (2.93% and 2.97% respectively). Angle-closure glaucoma is more prevalent in Asia (1.09%) and less in North America (0.39%). The number of glaucoma cases is predicted to grow between 2020 and 2040 mainly due to increases in Asia and Africa (+49.5% and +87.3% respectively). Statistics from a study published in 2001 [24] report that the angle-closure type is responsible for 91% of total bilateral blindness cases due to glaucoma in China,

highlighting that, despite being less common, it is a much more severe condition that needs to be diagnosed and treated promptly.

Glaucoma may also develop in patients with no increased ocular pressure, in which case it is called normal-tension glaucoma. In this case the anterior chamber angle is not directly involved in the pathogenesis and, therefore, this thesis will not further discuss this condition.

A study conducted in the United Kingdom [95] estimated that the average per-patient annual cost of glaucoma treatment (comprising both non-drug and drug costs) is 475£, however, according to the authors *"the wider societal costs from caring for the visually impaired patient are not covered in the database and these are excluded from the cost analyses, but could be expected to represent a significant aspect of the overall cost burden of glaucoma"* (Rahman *et al.*, 2013, p. 1). Another European study [128] conducted in 2005 reports that the annual cost of a glaucoma patient depends on the severity of the disease and may span from 455€ to 969€.

Blindness caused by glaucoma is not curable. However, if risk factors are found, preventive actions can be taken, or, if the disease is diagnosed in its early stages, it is possible to effectively slow down its development through medications, laser treatments or surgery.

1.4 Visualise the Anterior Chamber Angle

1.4.1 Gonioscopy

IoP can be measured through an easy and uninvasive exam called *tonometry*. However, when an increased IoP is found, understanding the cause may not be straightforward and a more detailed assessment must be performed. Gonioscopy is the clinical standard examination for drainage angle assessment according to international guidelines [31, 1].

The aim of gonioscopy is the visual assessment of the anterior chamber angle to evaluate the appearance and morphology of its anatomical structures. Characteristics such as the visibility of layers, the shape and extent of their occlusion, the pigmentation of both the trabecular meshwork and the Schwalbe's line, and the presence and shape of vessels guide

the evaluation of risks of glaucoma development, the categorization of the disease and, eventually, the choice for the best treatment strategy.

For example, the visibility of anatomical structure and the extent and type of occlusions (i.e., the concept of *angle aperture* that will be better explained in the following chapters) are fundamental for angle-closure glaucoma diagnosis which, in turn, may require specific surgery if compared to open-angle glaucoma that is, instead, mainly treated using drugs.

The grading of trabecular meshwork pigmentation, instead, may be used to tune the power of laser treatments, since the darker the area to target, the higher the energy that will be absorbed by tissues.

Conventional Gonioscopy

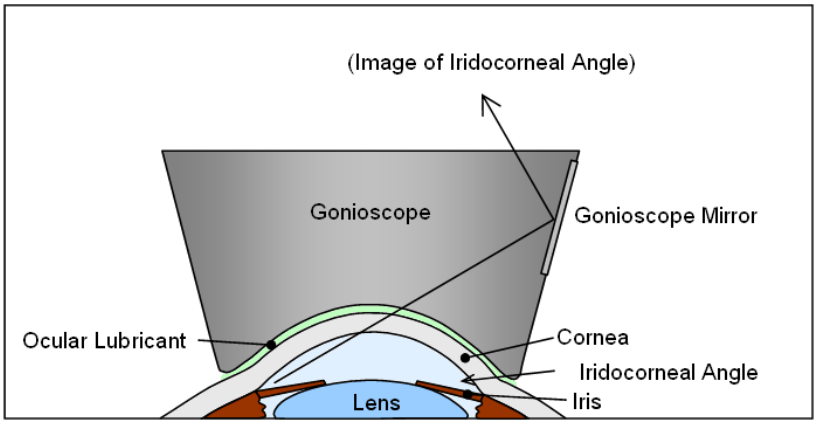

Fig. 1.2 *Schematic representation of the working principle of indirect gonioscopy (reproduced under CC BY-SA 3.0, source: https://commons.Wikimedia.org/wiki/File:Gonio.png).*

Conventional gonioscopy is performed manually, using a slit-lamp or an operating microscope together with a special lens, called *goniolens*. The lens allows to inspect the surface of the drainage angle directly or indirectly (using mirrors) (Figure 1.2). The procedure requires to hold the lens in contact with the patient's cornea, using a thin layer of gel as interface. The gel lubricates the surface of contact and allows the ophthalmologist to adjust the lens position in order to inspect the angle without damaging the cornea.

This examination technique has not significantly evolved over the past fifty years. It is not easy and requires substantial experience [44, 23]. Moreover, it requires time, has a low

repeatability and it does not enable an easy acquisition of digital pictures of the eye region making the follow up of patients difficult. Because of all this, it may procure discomfort to the patient, that, in turn, makes the exam itself more difficult and less effective.

Conventional gonioscopy is not performed as frequently as recommended [17] especially in primary-care settings, potentially missing patients at risk. In order to overcome some of its limitations, digital imaging devices have been developed [87, 114].

Digital Gonioscopy: the NIDEK GS-1

For the purpose of this research, only the NIDEK GS-1 digital device has been used and a brief description of the way it works is provided here.

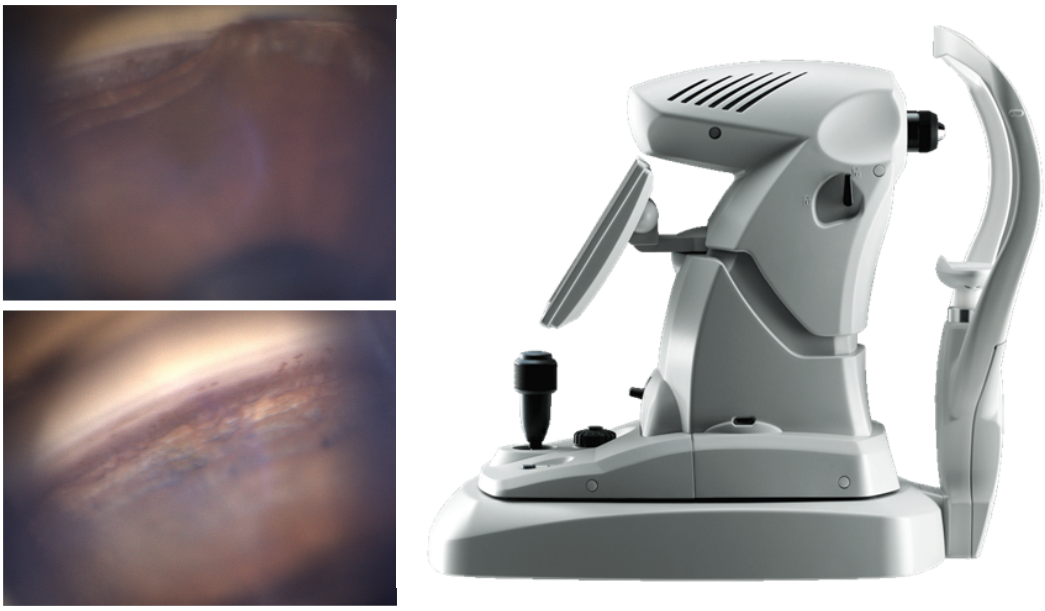

Fig. 1.3 *Left: two images of drainage angle sectors; at the top a partially closed angle due to a synechia, at the bottom a healthy, open angle. Right: an image of the NIDEK GS-1 device.*

The GS-1 device shares the same basic working principles of manual indirect gonioscopy, such as the use of the multi-mirrored prism and the lubricating gel. It aims to improve the conventional examination by reducing both the operator's experience required and the time needed. It can actively support the operator during the examination, detecting the angle automatically and, based on that, adjusting the alignment to speed-up the acquisition and

increase reproducibility. Several studies have reported the clinical utility of this device [114, 76, 8].

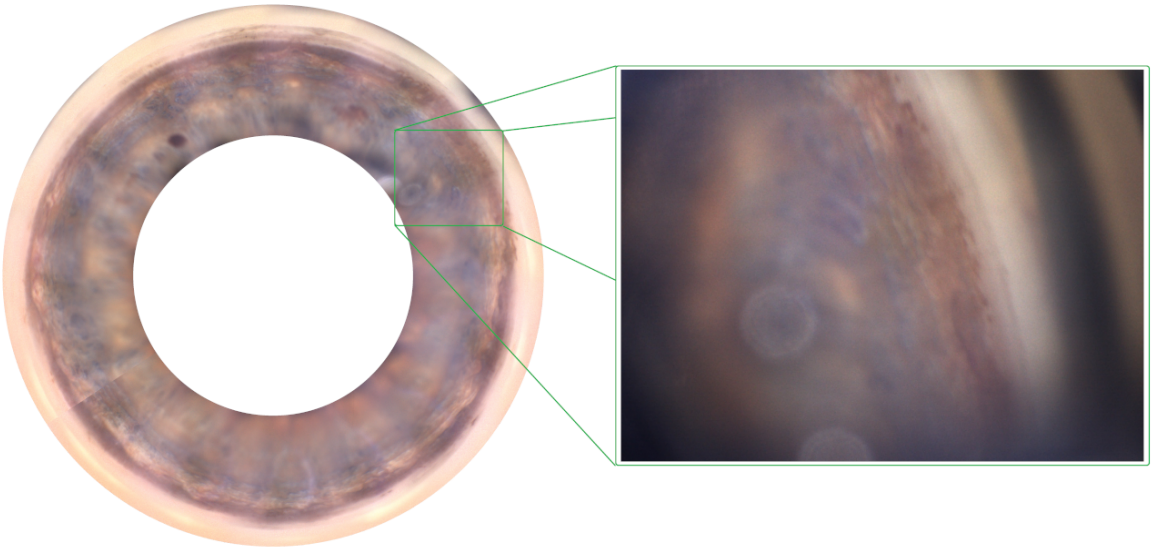

Fig. 1.4 *Left: the whole exam acquired by the NIDEK GS-1, represented as the software-generated circular stitching of multiple sector images. Right: a single sector acquisition.*

The whole 360° drainage angle region is assessed by acquiring digital images of sixteen 22.5°-wide sectors of the interface. Since the depth-of-field of each shot is limited, each sector is imaged through several pictures at different focal distance. Digital images are immediately saved and stored on a hard drive and it is easily possible to use them for comparisons and analysis. This represents an important advantage with respect to the conventional examination technique. An example of software-generated stitching image capturing the whole drainage angle circumference is shown in Figure 1.4 (left) together with one of the angle sector acquisition (right) used to create it.

The NIDEK GS-1 can effectively simplify and speed up the examination by automating some of the most difficult steps of conventional gonioscopy. However, the data analysis to identify conditions that may need medical intervention is still a time-consuming practice that must be performed by experts.

1.4.2 Optical Coherence Tomography

The *anterior segment optical coherence tomography* (AS-OCT) [49] is the main alternative to gonioscopy for assessing the drainage angle. It is a non-invasive and non-contact diagnosis method based on a principle similar to ultrasonography, but it uses light instead of ultrasounds. It extracts longitudinal sections of the anterior chamber (called *B-scans*) measuring the interference patterns produced by light reflected by different tissues and transforming these in distances between anatomical surfaces.

From each of these sections it is possible to estimate the angle width, as an indicator for diagnosing angle-closure glaucoma. This technique has a primary disadvantage with respect to gonioscopy: it doesn't allow direct access to the surface of the angle, that is a precious source of information (e.g., to evaluate trabecular meshwork pigmentation and to monitor post-surgical healing of tissues). Moreover, it is not able to acquire a complete overview of the angle (along the circumference where iris and cornea meet, Figure 1.1), but only one section at a time; in practice, the sampling of radial B-scans along the drainage angle generates gaps, possibly missing localized morphological features of interest (e.g., synechiae). Finally, important landmarks, like the scleral spur, may not be visible clearly.

AS-OCT and gonioscopy have different advantages for the detection and evaluation of glaucoma-related conditions in the anterior chamber angle, remaining complementary examination approaches in clinical practice [18].

1.5 Research Questions and Contributions

Current clinical set-ups, e.g., virtual clinics and telemedicine, often require the separation of data-acquisition (performed by a technician) and data-analysis (performed by a clinician) and conventional gonioscopy is not suitable for this scenario. Digital imaging devices may be used by medical photographers to efficiently examine patients, but the large amount of images they collect must be viewed by an expert. As it happens for other medical specialities, software may help data evaluation, speeding up the process and saving substantial time. However, we

must keep in mind that applications in medicine are critical in terms of reliability of results and safety of the patients.

For these reasons, our work focused on the following research questions:

- *Can deep learning algorithms perform a clinically motivated analysis of digital gonioscopic images, supporting diagnosis?*
- *What performance assessment of such algorithms can support clinical acceptability, and what clinical ground truth is needed?*

and produced the following contributions:

- **the first study on inter-annotator variability at delineating anatomical layers of the drainage angle in gonio-photographs (Chapter 4)**: this is the first study aiming at evaluating inter-annotator agreement using delineations of anatomical layers provided by multiple clinical experts. Importantly, a meaningful evaluation of medical image analysis algorithms requires first to assess inter-annotator variability at performing the same task. Ideally, intra-annotator variability would be needed too, but it was impossible to secure it given the limited annotator time available compared to the time required by each annotation. This analysis enables to put algorithm performance in the correct perspective and helps one to understand the potential and limitations of such systems for real applications. The output of this study has been published in *Translational Vision Science and Technology* [89];
- **the development and evaluation of a new approach for the semantic segmentation of anatomical layers of the drainage angle in gonio-photographs (Chapter 5)**: despite similar systems have been deployed for other clinical applications, the deep learning algorithm reported here is the first specifically addressing digital gonioscopic images. It has been comprehensively evaluated and compared to the results of the inter-annotator variability study to interpret correctly its performance. This research has been published in *Communications in Computer and Information Science* [88] and *BMJ Open Ophthalmology* [90] and in *Investigative Ophthalmology & Visual Science* [91] as a conference abstract;

- **the investigation on advantages and limitations of a system for automatic *local* angle aperture grading (Chapter 6)**: a dataset of digital gonioscopic exams has been annotated by multiple experts. Each exam sector has been divided into multiple sub-sectors and classified according to its aperture, obtaining local ground truth for angle aperture (about every 5-7°). A custom deep learning classification algorithm has been evaluated for the local classification of angle aperture in digital gonio-photographs with the aim to investigate the potential and challenges of performing this task and to provide a baseline for future work.

It is worth mentioning that the semantic segmentation of angle layers and the grading of angle aperture are somehow correlated tasks. The grading of angle aperture, in fact, may depend, among other criteria, on the visibility of anatomical layers, which is an obvious information from segmentation outputs.

The semantic segmentation system is, potentially, extremely powerful and versatile and may be a baseline for a large number of future studies comprising the grading of angle aperture, the measurement of layers' width and synechiae's extent, and even the development of advanced automatic alignment and online target detection (i.e., augmented visualization) systems during surgery. However, mainly due to the difficulty in obtaining ground truth for semantic segmentation (due to the complexity of the annotation task itself and to the COVID-19 pandemic, which largely limited the availability of experts to support our work), we decided to investigate a separate system for grading angle aperture at this stage of research. Moreover, a stand-alone angle aperture classifier has been considered of greater interest based on considerations about its translational potential since it may be designed to have lower memory-related requirements. These facts motivated our decision to keep the two systems independent at this stage of research, but our intention in the future is to investigate what advantages may be obtained by making them inter-operate.

1.6 Disclosure

This research has been fully funded by NIDEK Technologies Srl. (Albignasego, Italy) and has been conducted with the collaboration of several international clinical centres:

- Ninewells Hospital, NHS Tayside, Dundee, United Kingdom;
- Princess Alexandra Eye Pavilion, NHS Lothian, Edinburgh, United Kingdom;
- Hospital de Santa Maria, Lisbon, Portugal;
- Clinica Oculistica, Di.N.O.G.M.I., University of Genoa, Genoa, Italy;

and research groups working on image analysis from two universities:

- CVIP/Vampire group, University of Dundee, Dundee, United Kingdom;
- Vampire group, University of Edinburgh, Edinburgh, United Kingdom.

1.7 Structure of the Thesis

The thesis is structured as follows:

- **Related Work**: Chapter 2 briefly summarizes literature on the evolution of image analysis with deep learning, focusing on the approaches and techniques that are more relevant for our research and providing examples of applications similar to our study case whenever possible;
- **Clinical Requirements and Annotations**: Chapter 3 describes the collection and selection of clinical requirements, and the consequent annotation of gonio-photographs for the image analysis tasks selected;
- **Inter-annotator Variability Study**: Chapter 4 presents the inter-annotator variability study on delineations of anatomical layers in digital gonio-photographs as a baseline for the evaluation of automatic analysis systems performance;

- **Semantic Segmentation of Drainage Angle Layers**: Chapter 5 describes the design, the development and the evaluation of the first semantic segmentation algorithm for digital gonioscopic pictures with particular focus on the approaches devised to deal with specific data characteristics and clinical usability issues;
- **Angle Aperture Classification**: Chapter 6 reports the pilot study on the local angle aperture classification algorithm, with experiments and considerations about current challenges and future opportunities;
- **Discussion and Future Work**: Chapter 7 concludes the thesis and discusses this research highlighting achievements and limitations, with suggestions for future work.

Chapter 2

Related Work

2.1 About this Chapter

This chapter discusses the background and the literature most relevant to our research. The topics are: (i) general basic deep learning concepts and processing units, (ii) convolutional neural networks for image classification, (iii) convolutional neural networks for semantic segmentation, (iv) architecture improvements common to both tasks, (v) introduction to epistemic uncertainty estimation, (vi) applications of classification and semantic segmentation systems in ophthalmology (with focus on digital gonio-photographs whenever possible), and (vii) published work on inter-annotator variability highlighting the issue of ground truth reliability and its effects on the assessment of automatic systems performance.

2.2 General Deep Learning Concepts

Deep learning models (also known as neural networks) are non-linear computational systems developed to learn mapping functions between inputs and outputs to address specific analysis tasks (e.g., classification or translation of data). Neural networks consist of structured ensemble of processing units (described later), some of which have *learnable weights*, parameters to be optimized. The optimization is performed during a process called *training*. Depending on the task to solve, it may be required to include *ground truth* data (also

called *annotations*) in the training process. They are used to supervise the optimization by comparing network's outputs with references through a loss function. Learnable weights are iteratively updated to decrease the loss function value. Once the network has been trained, it can be used to process new, unknown data, a step called *inference*.

Early studies on neural networks started in the late fifties with the *perceptron* [98] and saw key advances in the eighties and nineties, developing some of the fundamental components of all modern architectures for image analysis, e.g., the introduction of *convolutions* [29] and *back-propagation* learning [101, 66, 67]. Despite the encouraging results obtained, limitations in computational power and unresolved issues regarding the training process (e.g., over-fitting and vanishing gradients) slowed down the development of this technologies until about 2005 when powerful GPU systems started to become available and new studies improved the training process addressing some of the known issues [45, 46, 32, 33].

This chapter focuses on convolutional neural networks (CNNs) as they are the family of deep learning models our work is based on. Other recent important deep learning systems comprise *vision transformers* [57] and, more specifically, *convolutional vision transformers* [136].

The main advantage of convolutional neural networks over conventional machine learning approaches is their capability of processing raw data without requiring hand-crafted features extraction and, ideally, any kind of ad-hoc pre-processing. These systems process input images using a minimum set of basic constituent modules, that are:

- ***convolutional layers***: weight matrices used to filter inputs through the convolution operation. Filters have an assigned size (e.g., 3x3 pixels) along input's dimensions, which defines their *receptive field*, and may extend across several channels. The main advantages of convolutional layers are: (i) *sparse interactions*, meaning that, considering a specific convolutional filter, each value of its input feature maps affects only a limited region of its output feature maps, proportional to the kernel size, and (ii) *weight sharing*, which is a way to reduce the number of weights in a network since small convolutional kernels are used to filter the whole input. These two concepts improve models' invariance against the location of discriminative features in the data;

- ***fully connected layers***: mainly used in classification networks, they are filters that map each input point into every output point. Several of these layers are usually stacked to reduce the dimensionality of data representations and match the number of possible output classes;

- ***non-linear operators***: also called *activation functions*, they allow the network to learn and model non-linear decision class boundaries (e.g., ReLU and LeakyReLU [81, 74]);

- ***pooling layers***: layers that reduce the width and size of intermediate features to increase the receptive field of convolutional kernels and enable to capture more context, introducing also additional non-linearities. They are not essential, but used by almost all deep learning systems.

Remarkable results have been reached so far in a wide variety of tasks using deep learning models, e.g., natural language processing [85], image classification [132], object detection [71] and image segmentation [78].

2.3 CNNs for Image Classification

CNNs for image classification have been designed and improved over the years to address a basic and important task for automatic image analysis which is the categorization of a whole input sample (an image) into one of a set of pre-defined classes. Large datasets have been generated and used to evaluate new approaches [61, 19] since the annotation of images showing natural objects, e.g., vehicles and animals, for classification purposes is relatively fast and does not require specialized knowledge.

For the purposes of this thesis, the focus will be on fundamental classifier architectures, still widely used as baseline for the development of application-specific systems.

2.3.1 Examples

An overview of a few examples of deep CNN classifier architectures, important for the evolution of these systems over the years, is proposed in this section to introduce the solutions and improvements that characterise these models and to provide useful references.

Additional information can be found in recent reviews, e.g., [11, 113].

Approaches that are particularly relevant for our research and common to both classification and segmentation systems are discussed more in detail in Section 2.5.

AlexNet

AlexNet [62] is the neural network that won the *ILSVRC 2010* [102] image classification contest on a subset of the ImageNet dataset [19] and contributed to generate new interest for deep learning. The architecture consisted of five convolutional layers, with variable decreasing size (from 11x11 to 3x3 pixels) and three fully connected layers. It used *rectifier linear units* (ReLU) as non-linear activations [81, 74], *Dropout* [119] (see Section 2.5) and label-preserving data augmentation.

VGG Nets

VGG Nets [116] are a group of network configurations designed to explore the effects of increased model depth (number of successive convolutional filters) on performance. The number of convolutions considered ranges from 8 to 16 and all models terminate with three fully connected layers. Groups of two or three convolutional layers are followed by max-pooling layers and, moving towards the output, they return an increasing number of feature maps. In order to limit computational requirements, all kernels have size equal to 3x3 pixels. The study demonstrated that depth is a crucial characteristic for achieving better performance and allows to use small convolutional kernels.

GoogLeNet and Inception Nets

GoogLeNet [120] has been devised to reduce the problem of vanishing gradient. It consists of 22 layers and introduces a new processing block called *inception module* which performs multiple parallel convolutions each using a different receptive field size, also making the model less sensible to target objects' dimension in the image. The inception module proposed in [120] has been further refined and led to the design of a family of *Inception Nets* [121, 122] characterized by different implementations of the inception module.

ResNets ad DenseNets

ResNets [38, 39] are a family of very deep (up to more that 100 layers) classifiers designed to be effectively trained end-to-end by incorporating short skip connections that provide a preferential path for the back-propagation signal to flow during optimization. A hybrid *Inception-ResNet* model has also been proposed [122]. The success of ResNets highlighted the importance of skip connections and led to the design of *DenseNet* [47], where feature maps are forwarded to multiple convolutional layers through skip connections to encourage a better reuse of encoded information. More details on the processing solutions characterising residual and dense models are provided in Section 2.5.

Some of the classifiers introduced above (e.g., VGG Nets, Inception Nets, ResNets and DenseNets) can still be considered important architectures as they are among the most frequently used either off-the-shelf (e.g., [124]) or through some kind of adaptation (e.g., [104, 103, 60]) for most practical applications, e.g., in medical image analysis, and are often the baseline for the evaluation of new approaches.

2.4 CNNs for Image Segmentation

One particularly sophisticated application of deep learning algorithms consists in the segmentation of multi-dimensional data arrays, that is the automatic delineation of targets in images according to one of the following approaches:

- *semantic segmentation*: each spatial data point (pixel) in an image is assigned a label indicating the specific class of targets it belongs to;
- *instance segmentation*: data points (pixels) belonging to *one* class of targets are grouped according to the particular class instance they belong to;
- *panoptic segmentation*: each spatial data point (pixel) in an image is assigned both its class and instance labels.

Although much of the literature discussed in this chapter is suitable for all the above mentioned tasks, the main focus will be on semantic segmentation [123]. One of the main contributions of this work is, in fact, the development of a new semantic segmentation algorithm for digital gonio-photographs (Chapter 5). In this context, only one instance of each of the segmentation targets may be visible in an image (possibly the only instance may be partially occluded thus being split into several regions), thus excluding the need to discriminate between multiple objects of the same class.

2.4.1 Examples

To address the semantic segmentation task, several CNNs have been proposed over about the past ten years. Similarly to what has been done in Section 2.3, a few baseline state-of-the-art models are discussed here, leaving a more detailed description of relevant architectural improvements to Section 2.5.

Patch-based Segmentation

Proceeding in chronological order, one of the first ways to tackle this task was presented by Ciresan *et al.* in 2012 [16]. In this article, the authors developed a system to segment neuron membranes in serial-section transmitted electron microscopy. An image classifier was used to label each image pixel as membrane or non-membrane based on the context within a squared window centred on it. They also studied the performance of network ensembles, finding that they led to better results than a single network. The main drawback of this approach is that it

does not exploit the local correlation of image features. In fact, image patches analysed to label adjacent pixels share most of the information but need to be processed separately by the system.

Fully Convolutional Networks

Long *et al.* [73] adapted several classification models for the semantic segmentation task and compared them using the *PASCAL* dataset [22]. Their intuition was to replace the fully connected layers of classification networks with additional convolutional layers and up-sampling layers. They would output coarse feature maps exploiting a wide context to assess what is visible in the input data. They incorporated (learnable) deconvolutional filters to upsample the feature maps from layers located at different depths of the architecture and merged global and local features to return the final semantic segmentation map. The model was trained end-to-end using dense per-pixel annotations.

DeconvNet and SegNet

While in [73] the upsampling path of the network is quite simple (i.e., only one deconvolutional layer regardless of the dimension of the map to up-sample), other authors proposed to expand the decoder to reconstruct a better high resolution segmentation map of the input. Both Noh *et al.* [82] and Badrinarayanan *et al.* [7] independently designed deep encoder-decoder models built upon the VGG-16 classifier [116], and using *un-pooling* as upsample technique. Un-pooling layers receive the indexes of the pixels selected by the corresponding max-pooling layers and use them to resample feature maps considering the original location of detected features. This makes the network reconstruct target edges more accurately.

Noh *et al.* [82] used the whole VGG-16 architecture (including the fully connected layers) as encoder, which alone consists of more than 100 million parameters, and a specular deconvolutional decoder (made up of deconvolutions), obtaining a complex model. The convolutional encoder was initialized with weights pre-trained on the ImageNet dataset [19]. To facilitate the training process and to correctly segment targets of variable size, they first generated multiple *candidate proposals* (i.e. image crops possibly containing an object),

segmented them and then aggregated the results to obtain the overall input segmentation. They also verified that an ensemble of the architecture in [73] and theirs usually led to better segmentations than either method alone. Badrinarayanan *et al.* [7], instead, removed the fully connected layers from the VGG-16 model, obtaining a much lighter network (the encoder consists of about 15 million parameters) that was possible to train end-to-end. They used convolutions instead of deconvolutions in the deep decoder to densify the upsampled, sparse feature maps.

U-net

Introduced by Ronneberger *et al.* in 2015 [97], *U-net* set a new gold-standard in semantic segmentation tasks. The model consists in a contracting path to capture context (global image features) and a symmetric expanding path to refine the final segmentation map with local details. *Long skip connections* between corresponding layers of the contracting and expanding paths are included to propagate and reuse detail information from shallow layers of the encoder. The very effective architecture, the use of data augmentation through elastic deformations as well as the use of a weighted loss led U-net to achieve the best performance in both the *ISBI 2012* segmentation challenge (previous best result was obtained by Ciresan *et.al.* [16]) and the *ISBI 2015* cell tracking challenges. U-net has become the fundamental backbone for several modern semantic segmentation architectures [115].

Attention U-net

The *attention U-net* is an architecture capable of refining feature maps generated by the encoder layers before they are forwarded and processed by the corresponding decoder layers through addition of *attention gates* [83, 107]. These modules can trim features that are not relevant for a specific task and, thus, refine the final segmentation map. Attention gates can be included in standard U-net models as well as in more complex and sophisticated variants as those introduced below.

Inception, Residual and Dense U-nets

Inception [120–122], *residual* [38, 39] and *dense* [47] processing blocks, initially conceived in the context of classification, have been introduced into U-net architectures (replacing simple convolutional layers) to make them inherit their advantages. In particular *inception* blocks may help improve the segmentation performance whenever target objects are represented at different scales in the inputs, *residual* blocks enable to design very deep segmentation networks, and *dense* blocks optimise the use of information and help reduce the issue concerning vanishing gradients during optimization. More details on the processing solutions characterising residual and dense variants are provided in Section 2.5.

U-net++

U-net++ [140] is an architecture inspired by *dense* blocks [47], in which long skip connections between each level of the encoder and the corresponding level of the decoder are replaced by a sequence of densely connected convolutional layers linked by short skip connections. Short skip connections may also include the upsampling of feature maps when placed between convolutional layers at different network's levels. The long skip connections of the conventional U-net do not ensure that feature maps originated from the encoder are effectively used by the decoder. U-net++ aims at making the information encoded by the downsampling path of the network more useful for the decoder, by adaptively refining it through intermediate filters.

2.5 Architecture Improvements

Most of the recently designed CNNs systems for both image classification and semantic segmentation tasks are based on the architectures presented in the previous sections and may comprise different combinations of the advanced modules and optimization techniques that are detailed hereafter. The approaches and network modules most relevant for this research thesis are described in the next paragraphs.

Features Normalization

In convolutional neural networks, input data are sequentially processed and transformed by model filters, meaning that each layer adapts its parameters according to the distribution of the outputs returned by the previous layer(s) over the training process. Learnable weights are updated multiple times proportionally to their effect on the loss value, leading to continuous changes in the distribution of intermediate feature maps. This makes the epoch-wise optimization suboptimal and increases time to convergence. In order to prevent this phenomenon and speed up training, several techniques have been proposed to normalize features distribution and make it more independent from the parameters update.

Data features generated by a specific layer may be seen as a multidimensional array $F \in \mathbb{R}^{D_1 \cdot D_2 \cdot \ldots \cdot D_n \cdot C \cdot B}$ where F is the feature array, $D_1, D_2, ..., D_n$ are the dimensions of each feature ($n = 2$ for images), C is the number of features and B is the mini-batch size (i.e., the number of input data processed before each optimization step).

Depending on the dimensions over which the normalization (i.e., subtraction of mean value and division by standard deviation) is applied, some of the most used approaches are:

- *batch normalization* [48]: over $D_1, D_2, ..., D_n$ and B;
- *layer normalization* [6]: over $D_1, D_2, ..., D_n$ and C;
- *instance normalization* [130]: over $D_1, D_2, ..., D_n$.

Each of the previous techniques may include additional learnable parameters (usually scale and shift). Data normalization techniques are effective at reducing the variability of features distribution over training, thus leading to better and faster convergence.

Dropout

Increasing depth and width of convolutional neural networks (i.e., their capacity) is usually a sensible way to improve their performance. The large number of trainable parameters may, however, cause over-fitting on the training set samples and reduce the generalization of the model to unknown data. Ensembles of networks have proven to considerably increase

both performance and generalization, but they come with an increased computational cost. *Dropout* [119] is a convenient way to regularize network training by randomly selecting feature maps values to be zeroed. In case of 2D (or spatial) dropout [127], entire feature maps are zeroed, while others are not altered. Dropout prevents processing units (e.g., convolutional layers) from strongly relying on the output of specific previous filters (i.e., memorize the training set) since they may be disabled by dropout, so the overall architecture must adapt to learn better and more general data representations reducing the issue of over-fitting.

Residual and Dense Blocks

The depth of a convolutional neural network is strongly associated with its capability to exploit context information. However, as extensively verified, for instance, by He *et al.* [38], deeper networks are usually affected by a performance degradation, not associated either with over-fitting or with vanishing/exploding gradients. Long sequences of convolutional filters and non-linearities hamper the back-propagation of the error signal and prevent an effective optimization of model's weights. Several studies have presented a common solution to this problem: the introduction of *signal shortcuts* (also known as *short skip connections*). He *et al.* proposed [38] and subsequently improved [39] a new network module, called *residual block*, to learn residual transformations. A residual transformation reinterprets the relation between the input x and the output y, usually written as $y = f(x)$, as $y = x + g(x)$. This relation is made up of two parts, the identity one and the residual one. The former ensures a preferential route for back-propagation without increasing computational costs or complexity.

The concept was further expanded by Huang *et al.* [47] designing *densely connected blocks* within which the feature maps from the first $n-1$ layers are concatenated and forwarded to the next one. The result is a high reuse rate of features. Thanks to this improvement, dense networks achieved state-of-the-art performance while requiring a reduced number of trainable parameters since the convolutional layers could be made up of fewer kernels. This approach can also be interpreted as *deep supervision* [68], since the back-propagation signal flows easily even to the layers furthest from the output.

Short signal shortcuts have been used to improve the conventional U-net, especially for clinical applications, e.g., by using residual blocks ([70, 58]) and dense blocks ([50, 131]). Short skip connections have proven to be fundamental for effectively propagating the error signal in deep networks [20].

2.6 Epistemic Uncertainty Estimation

During inference, deep learning models for classification and semantic segmentation are used to process unknown data and obtain the corresponding output. This standard, fully deterministic approach makes it impossible to know whether the model has a sufficient comprehension of the input for returning a reliable result. For this reason, techniques for estimating decision uncertainty have gained considerable attention in recent years [2], especially in those applications where the consequences of false positives/negatives is high, e.g., for clinical purposes.

Sources of uncertainty may be divided into two main classes: *epistemic* (or model's uncertainty) if it can be reduced by increasing the knowledge about the problem (e.g. by improving the model or increasing the amount of data), and *aleatoric* (or observations' intrinsic noise) if it cannot be reduced either by adding knowledge or improving the model. The sources of aleatoric and epistemic uncertainties are not always well defined and may be application-dependent [59]. Only epistemic uncertainty will be considered in this thesis, as more relevant in safety-critical applications (e.g. to detect out-of-distribution examples) and when limited data are available [56].

According to literature [64, 72], the most widely used approaches for epistemic uncertainty estimation in image analysis-related tasks are *Monte Carlo dropout* (MC dropout) [30] and *deep ensemble* [63] methods. Both techniques aims at modelling the variability among several predictions for the same input sample as a measure of epistemic uncertainty. The former method randomly switches-off intermediate neurons (or entire feature maps) of a single trained neural network using dropout [119, 127], while the latter uses the outputs from

multiple models (e.g., a set of identical architectures initialized randomly and thus having reached different local minima over optimization).

Our research will focus on MC dropout by Gal *et al.* [30] as it is more convenient in terms of memory requirements (only one set of weights must be stored). The authors have conceived a practical way to approximate the Bayesian framework (i.e., the gold-standard approach). It consists in activating dropout layers during inference and generating several output candidates by sampling weights from their approximated posterior distribution.

The final classification is obtained by averaging the softmax activations of the candidates and the epistemic uncertainty estimated by their pixel-wise variance or entropy. If the model has sufficient knowledge about the input to provide a reliable output, we expect final activations not to change much due to dropout action, thus returning a low uncertainty. Conversely, if the dropout action heavily affects output values, results are likely unreliable.

Uncertainty estimation is closely related to another property of deep learning models, called *calibration*, defined as the capability of a model to output *confidence* values (i.e., the outputs of the final softmax activation layer) that follow the actual classification accuracy. For instance, when a well-calibrated model classifies 100 samples with 0.9 confidence, only 90 of their output classes should ideally be correct. Bad calibration may lead networks to be very confident even when a result is wrong (or underconfident when it is right), making estimating uncertainty as the variance (or entropy) of multiple output candidates pointless. The effect of calibration on segmentations obtained with deep learning systems and on the estimation of uncertainty in several clinical applications has been investigated by Mehrtash *et al.* [77]. Model calibration can be verified both qualitatively (e.g., using *calibration plots*) and quantitatively. The quantitative evaluation of miscalibration can be performed using *expected calibration error* (Naeini *et al.* [80], Guo *et al.* [35]) by calculating the average value of the absolute difference between model's accuracy and confidence over a number of samples.

2.7 Clinical Applications

2.7.1 Image Classification

Convolutional neural networks have been deployed successfully for the detection or grading of many pathologies of the eye, the organ on which this thesis concentrates. Some examples of disease-specific applications comprise diabetic retinopathy [34], macular degeneration [86, 137], glaucoma [3, 5]. Multi-disease detection models have also been developed [126].

According to existing literature, the main application of classification systems to the analysis of anterior chamber angle characteristics concerns the classification of its aperture, as it is related to the risk of developing glaucoma and to its categorization (see Chapter 1).

Angle Aperture Classification in AS-OCT Data

Generally speaking, the angle aperture grading task concerns the classification of a visual representation of the irido-corneal region (e.g., an AS-OCT image or a gonio-photograph) into one of a set of clinical grades that associate its morphology (e.g., the curvature of the iris in AS-OCT scans or the visibility of angle layers in gonio-photographs) with the risk of trabecular meshwork obstruction (and consequently of an increased intra-ocular pressure). Several deep learning systems for angle aperture grading have been developed in recent years using AS-OCT B-scans as input data.

Fu *et al.* proposed several solutions [26, 25, 27]. The simplest one [26] was an adaptation of a VGG-16 [116] network which was pre-trained on the ImageNet dataset [19] and then fine-tuned on AS-OCT scans. In this work the authors supported results with *attention maps* to show what region of the image was affecting classifications the most. In [25, 27] multi-level parallel classification paths based on VGG-16 are proposed to exploit important image features of regions of interest increasingly focusing on the irido-corneal junction. In [25] two regions of interest are considered, the smallest one being extracted using an automated segmentation algorithm which also provides clinical parameters of the anterior chamber that are processed through a linear support vector machine (SVM). The final classification is obtained by considering both the SVM classifier and the two-level deep classifier outputs. In

[27] a three-level deep classification network was investigated by comparing the performance using different backbone architectures (e.g. VGG-16 and ResNet [38])

Porporato *et al.* [92] implemented a VGG-16 network to classify multiple AS-OCT B-scans as *"Open"* or *"Closed"* and provided a global classification of the eye that was then compared with one obtained using conventional gonioscopy finding good agreement ($AUC = 0.85$) with high sensitivity and specificity (83% and 87% respectively).

Angle Aperture Classification in Gonio-photographs

To our best knowledge, only one publication describes so far the application of a deep learning model for the classification of angle aperture in gonio-photographs [14]. Images were acquired using an EyeCam device (Clarity Medical Systems, Pleasanton, California, USA) and depict quadrants (i.e. 90°-wide sectors) of the drainage angle. Authors use a ResNet-50 network [38] pre-trained on ImageNet [19] and fine-tuned on EyeCam gonio-photographs to detect angle-closure, defined as the absence of the pigmented trabecular meshwork (see Chapter 1) in more than half of the image. The training dataset (not publicly available) was composed of 33,635 quadrant images from more than 4000 patients. Training was performed ensuring that the training-test split of data was performed at patient level. Automatic classifications were compared with expert's gradings and the authors found very good agreement ($AUC = 0.96$). As reported in the paper, the training dataset is representative only for Chinese patients with limited variability of some characteristics of the eye region (e.g., iris colour). The authors also suggest to investigate similar solutions for other types of gonio-photographs, e.g., those obtained with a NIDEK GS- 1 device, as the EyeCam system takes a long time to acquire data. This work will be considered when describing our angle aperture classification algorithm in Chapter 6.

2.7.2 Image Semantic Segmentation

CNNs for semantic segmentation have been deployed in many clinical applications and on different image modalities [36, 51, 139, 79].

One case close to the purpose of our research, i.e., to apply semantic segmentation techniques to gonioscopic images, is the segmentation of retinal layers in morphological OCT data.

Semantic Segmentation of OCT Retinal Data

Both OCT B-scans of the retina and digital gonio-photographs show layered anatomical structures and for this reason it is meaningful to report relevant work that investigated the semantic segmentation of retinal layers in OCT scans.

The first network specifically designed and trained for this purpose is *ReLayNet* by Roy *et al.* [100], where a U-net backbone model [97] is enriched with un-poolings [82, 7], and a combination of Dice and multi-class logistic losses is used to improve the segmentation performance. The loss accounts for class pixel-count imbalance and gives more importance to layers' boundaries.

Subsequently Wei and Peng [134] presented a modified version of *ReLayNet* replacing concatenations with depth max-pooling layers and introducing a new mutex Dice loss. The loss accounts for the morphological properties of retinal layers (e.g., their order) to guide the segmentation by penalizing the predictions proportionately to their distance from the annotation. It is worth mentioning that a similar approach cannot be used when segmenting gonio-photographs since the possible visual occlusion of one or multiple layers in one image would make it unreliable. Fu *et al.* [28] designed a multi-prediction guided network based on U-net [97] using *feature refinement modules* to adaptively weigh encoder features before they are concatenated with the corresponding decoder ones, and *multi-prediction guided attention modules* to refine and deeply supervise [68] the feature maps along the decoding path.

Several work on OCT data has also employed epistemic uncertainty estimation through Monte Carlo dropout [30], e.g. to detect anomalies [110], to improve results visualization and explainability [108] and to improve semi-supervised training frameworks [109].

There are important differences between OCT data and digital gonio-photographs. While OCT B-scans depict a 2D longitudinal scan and the sequence of layers does not change, digital gonioscopic images show a 3D region and, even if the order of layers is constant,

some of them may be completely or partially hidden. Image modalities are also different, with digital gonio-photographs being affected by shallow depth of field and vignetting. Even if the acquisitions for the two anatomical regions may look alike to some extent (both show layered anatomical structures), the profound differences in both acquisition modality and morphology (3D for the anterior chamber angle and 2D for the retinal B-scans) make the methods reported in the existing literature on semantic segmentation of OCT data unsuitable for off-the-shelf deployment in digital gonioscopy.

2.8 Inter-annotator Variability

In order for automatic algorithms to be deployed in real-life scenarios, they need first to be *validated*, meaning that their outputs have to be compared with a reference standard to evaluate how well they perform. In the case of artificially generated data or phantoms, a perfect reference standard can be obtained easily. However, when dealing with real applications, e.g., medical image analysis, ground truth information is usually impossible to obtain, unless perhaps in some cases through invasive surgery. For this reason, ground truth measurements are usually approximated by experts' annotations and the two terms will be used interchangeably in this thesis. Human annotations are, however, affected both by *systematic bias* (due, e.g., to training) and *noise* (random variations due, e.g., to tiredness).

These aspects lead to some degree of variability among experts, called *inter-annotator variability*, and also among multiple annotations from the same person at different times, called *intra-annotator variability*. When ground truth from multiple experts is available, the evaluation of variability is essential to interpret the results of the validation process including the comparison between different algorithms [75].

In the remainder of this thesis, only inter-annotator variability will be discussed, since multiple annotations from the same rater were not obtained for this study as too time consuming. Despite its importance, the effects of inter-annotator variability on automatic systems validation has been investigated rather sparsely in the literature. Moreover, frameworks and metrics to quantify variability are highly heterogeneous [129].

Lampert *et al.* [65] investigated inter-observer variability in four different scenarios and how ground truth fusion methods, e.g., consensus voting [54] and STAPLE [133], might affects algorithms' performance. *Consensus annotations*, i.e., the image areas annotated by a minimum number k of experts, were proven to considerably decrease, as k increased, in all the considered case scenarios, with consensus worsening the more linear the structures were. Moreover, ground truth obtained through consensus voting usually comprise only more obvious features, possibly leading to overoptimism when evaluating an algorithm. STAPLE does not perform well when limited data is available or there is high variability. Annotations from multiple experts or methods to merge them affect the evaluation of models and their comparison. The authors of this study suggest to validate algorithms with ground truth from multiple experts whenever possible.

Joskowicz *et al.* [52] studied inter-annotator variability when segmenting various organs in computerized tomography (CT) scans. They found that the range of variability between pairs of raters may vary a lot (5% - 57%) and that two or three annotators may be not sufficient to estimate the ground truth variability for a specific segmentation problem and to establish a reliable standard for the evaluation of automatic algorithms.

Ribeiro *et al.* [96] tackled the inter-annotator variability issue in skin lesion segmentation. They argued that the accuracy of annotations imposes an upper bound to the performance of an algorithm and that is important to decide whether it is worth to keep working on a model. Inter-annotator variability may be a proxy for annotation accuracy in this sense. They also studied simple ways to pre-process annotations to reduce variability, e.g., though morphological operators, assuming that the detail remains suitable for the specific task.

Recent studies have also investigated the relation between inter-annotator variability and the epistemic uncertainty estimated by probabilistic models (e.g., [55]).

Jungo *et al.* [53] postulated that, whenever disagreement is observed among annotators, supervised models need to be able to reflect it through uncertainty estimation in order to provide useful results. The authors explored the effects of ground truth fusion strategies on the way models infer epistemic uncertainty. They found that common fusion approaches reduce models' capabilities of reproducing experts' variability that is fundamental to detect

unreliable segmented areas and trigger additional human intervention. For this reason, they suggest to train automatic systems using multiple references from different annotators.

Chotzoglou and Kainz [15] studied the correlation among annotators' variability, probabilistic models' predictive epistemic uncertainty (computed either as the variance or the entropy of the segmentation candidates) and segmentation quality (quantified using Dice score). They found a negative correlation between segmentation quality and epistemic uncertainty meaning that, as expected, worse segmentations are associated with larger predicted uncertainty. Epistemic uncertainty was also found related to experts' variability meaning that probabilistic models may return reliable estimates of human uncertainty.

2.9 Discussion and Conclusions

Deep learning models have been demonstrated to be powerful tools for automatically analyse data and return relevant information for detection, localization and delineation of features. State-of-the-art architectures (e.g., ResNets, DenseNets and their segmentation counterparts Residual and Dense U-nets) enabled to reach remarkable performance in the analysis of different types of natural images.

Many opportunities exist for applying these systems to healthcare to support disease diagnosis and grading and the interest within the clinical community has considerably increased with time. However, the deployment of these technologies to the medical field requires particular attention about results interpretability and data curation. Since the output of an automatic system may have important (potentially dangerous) consequences on patients lives, there must be a way to interpret and understand model's confidence through reliable uncertainty estimation techniques. Moreover, ground truth for clinical applications is often impossible to obtain and its approximation relies on experts' opinions that may be affected by bias (inter-annotator variability) and noise (intra-annotator variability). Estimating variability in annotators' choices is important to assess the performance of automatic systems for data analysis.

While several papers on classification and segmentation of OCT data exist in literature, the available work on the analysis of gonio-photographs is very limited (classification) or even absent (semantic segmentation). Considering the importance of gonioscopy as clinical-standard, this gap motivated our research.

Chapter 3

Clinical Requirements and Annotations

3.1 About this Chapter

This chapter describes the preliminary phases of this research on the design and the development of deep learning algorithms for digital gonioscopy, collecting and evaluating clinicians' preferences on the type of automatic systems that would best support data analysis. The clinical requirements obtained were used to focus the research questions and lead to the contributions listed in Chapter 1. The chapter then presents the activities related to the annotation processes designed to meet the requirements and provide ground truths for the training and the evaluation of the systems to be developed.

3.2 Requirements Collection and Evaluation

The automatic analysis of digital gonioscopic images both through conventional and machine learning-based systems is a complex and almost unexplored research field, as highlighted in Chapter 2. As the pool of tasks possibly helpful in clinical practice and never addressed on gonio-photographs is large, clinical experts were inquired on what the most useful analysis tools would be.

Eight clinical collaborators at six international clinical centres (in the United Kingdom, Portugal, Italy, France, United States and Japan) with extensive experience in gonioscopy

rated a list of image analysis tasks in order to identify those with higher consensus. A copy of the questionnaire shared with the clinicians is reported in Appendix A. Many of the tasks in the list were collected by NIDEK Technologies Srl. during clinical conferences and meetings with clinical partners over a period of time antecedent to the beginning of this research. Some of them were included after a preliminary review of deep learning applications in image analysis for medicine and presented to experts to evaluate the potential interest. The tasks considered ranged from the detection/localization of anatomical landmarks of the anterior chamber angle to the assessment of more general features such as its aperture. Divided, for simplicity, into three broad categories, they were:

1. ***detection/localization of features***:

    - *synechiae*: iris tissue locally adhering to the trabecular meshwork or cornea, thus preventing the aqueous humour outflow (see Chapter 1);
    - *neo-vascularizations*: blood vessels growing in the drainage angle, possible symptom of some glaucoma sub-types;
    - *Schwalbe's line*: one of the main landmarks of the anterior chamber angle located between the trabecular meshwork and the cornea, relevant for angle aperture grading (see Chapter 1);
    - *scleral spur*: another important landmark of the anterior chamber angle located between the ciliary body and the trabecular meshwork, fundamental for angle aperture grading (see Chapter 1);

2. ***classification/grading of features***:

    - *drainage angle aperture*: according to one of the available clinical grading systems (e.g., Spaeth's [117] or Scheie's [106]) to evaluate risks of angle closure;
    - *trabecular meshwork pigmentation*: according to one of the available clinical grading systems (e.g., Scheie's [106]) as a preliminary step for laser surgical procedures (e.g., laser trabeculoplasty);

3. ***others***:

    - *semantic segmentation of layers*: to distinguish between different anatomical tissues visible in the images;
    - *extraction of the angle profile*: to approximate the 3D structure of the region;
    - *estimation of synechiae size in degrees*: important to estimate the total extension of angle closure;
    - *angle aperture in degrees*: to evaluate the aperture of the angle geometrically and not based on other features, e.g., on the visibility of the layers;
    - *classification of image focus*: to flag poor quality images.

Ophthalmologists assigned a priority score (*High*, *Medium* or *Low*) to each task. *N/A* could be assigned if the clinician was not sure about the answer. The results are reported in Table 3.1.

The three tasks that received the largest percentages of *High* priority rates are the **detection/localization of synechiae**, the **detection/localization of the scleral spur** and the **semantic segmentation of the anterior chamber angle layers**. In particular, the detection of the Scleral spur relates to the need for a way to estimate angle aperture, as its visibility is the discriminant characteristic of a fully open drainage angle.

Among the three tasks with the highest priority according to experts, the semantic segmentation of angle layers was selected as first task to tackle for the automatic analysis of digital gonio-photographs in this research project. In fact, the automatic delineation of anatomical layers of the anterior chamber angle can be a suitable pre-processing step to address many other tasks in the list that were considered useful by the ophthalmologists (comprising the two with the highest priority).

In particular, once the layer interfaces are obtained, it is possible to infer the localization of both the Schwalbe's line and the scleral spur. Synechiae may be detected by evaluating the intersections of the outer boundary of the iris with other layers (e.g., if at any point of the segmentation map the ciliary body and the scleral spur are hidden by the iris, that is a synechia) and their size (in degree) could be estimated from the segmentation map by

Table 3.1 *Task ratings as provided by the ophthalmologists involved.*

Group	Task	Priority			
		High	**Medium**	**Low**	**N/A**
Features Detection/Localization	**Synechiae**	**87.5%**	**12.5%**	**0%**	**0%**
	Neo-vascularizations	50%	12.5%	37.5%	0%
	Schwalbe's Line	37.5%	25%	37.5%	0%
	Scleral Spur	**75%**	**25%**	**0%**	**0%**
Classification	Angle Aperture	50%	37.5%	0%	12.5%
	Trabecular Meshwork Pigmentation	25%	75%	0%	0%
Other	**Semantic Segmentation**	**62.5%**	**25%**	**0%**	**12.5%**
	Angle Profile	0%	87.5%	0%	12.5%
	Synechiae Size in Degrees	37.5%	62.5%	0%	0%
	Angle Aperture in Degrees	50%	50%	0%	0%
	Image Focus Classification	25%	37.5%	25%	12.5%

accounting for additional information on the geometry of the eye (e.g., average iris radius) and of the optical system (e.g., field of view). Angle aperture may be graded, for example, according to the visibility of the layers (directly, using Scheie's grading system [106] or indirectly, according to the apparent level of iris insertion, using Spaeth's system [117]) across the whole extension of the drainage interface. The trabecular meshwork could be segmented as pre-processing step for grading its pigmentation (e.g., according to Scheie's system [106]), which is an important practice prior to laser treatments.

Moreover, given the time required to obtain data annotations, the choice of the semantic segmentation task was deemed more effective since it requires ground truth that may be also suitable for other development purposes (see discussion above).

We remind the reader that this research thesis has been funded by an industrial grant and that its objectives reflect both academic and industrial (i.e., translational potential) interests. For this reason the second task selected consists in the direct classification of angle aperture, without necessarily relying on image segmentation.

The training and evaluation of machine learning algorithm requires a curated set of annotated images. The tools and protocols for collecting annotations to develop solutions for the two analysis tasks selected above are the topic of the next sections.

3.3 Annotations for Semantic Segmentation

In the case of semantic segmentation algorithms, the annotations consist in the boundaries of targets delineated on the best-focus frame selected from the stack of acquisitions for each angle sector. We remind that the best-focus frame of an angle sector is the image of the focus stack (the device acquires multiple shots at different focus planes) showing the best sharpness in the region centred on the trabecular meshwork (or on the outer iris boundary if the meshwork is not visible). Best-focus shots were automatically chosen by the imaging device software through proprietary algorithms that first locate the interface point along the oriented (according to the sector location) median segment [10] and then evaluate the focus of a region of interest centred on that point [138]. Best focus frames were always inspected

by the author of this thesis and corrected in the unlikely case they were suboptimal. The choice of the best-focus image is fundamental in order to provide the segmentation system with the most informative (better detailed) data as input.

To facilitate the annotation process and increase the consistency of semantic segmentation ground truth from different annotators, a suitable annotation tool has been selected from the public domain and an annotation protocol has been devised. This work was conducted in close collaboration with the clinicians involved.

3.3.1 Annotation Tool

The use of third-party software to annotate medical data introduces constraints:

- *file sharing*: medical data may not be shared with any non-authorized individual/institution according to the *General Data Protection Regulation* (*GDPR*). Thus, the data should be stored locally and the annotation tool should not forward any information remotely;
- *operative system-independent*: the annotation tool should be usable with most operative systems in use in the clinical world transparently.

Our choice was the *VGG Image Annotator* (version 2.0.8) [21], an open source application suitable for both research and industrial purposes, developed by the Visual Geometry Group (VGG) at Oxford University. It consists in a portable *.html* file, it neither requires nor uses internet at any time and is usable with all the most common operative systems, since it only requires a browser interface to be run on. The annotation tool is suitable for segmenting target regions in images through the delineation of polygons and allows the user to assign a label to each of them. Annotations can be exported in a convenient tabular format (in a *.json* file) storing the coordinates of the vertices of each polygon and the corresponding label.

Figure 3.1 shows the annotation tool user interface, a digital gonio-photograph and the annotation of one of the layers.

The annotation tool provides a standard framework for general-purpose annotation of images. Specific region labels were defined to adapt the tool to the purpose of this annotation task.

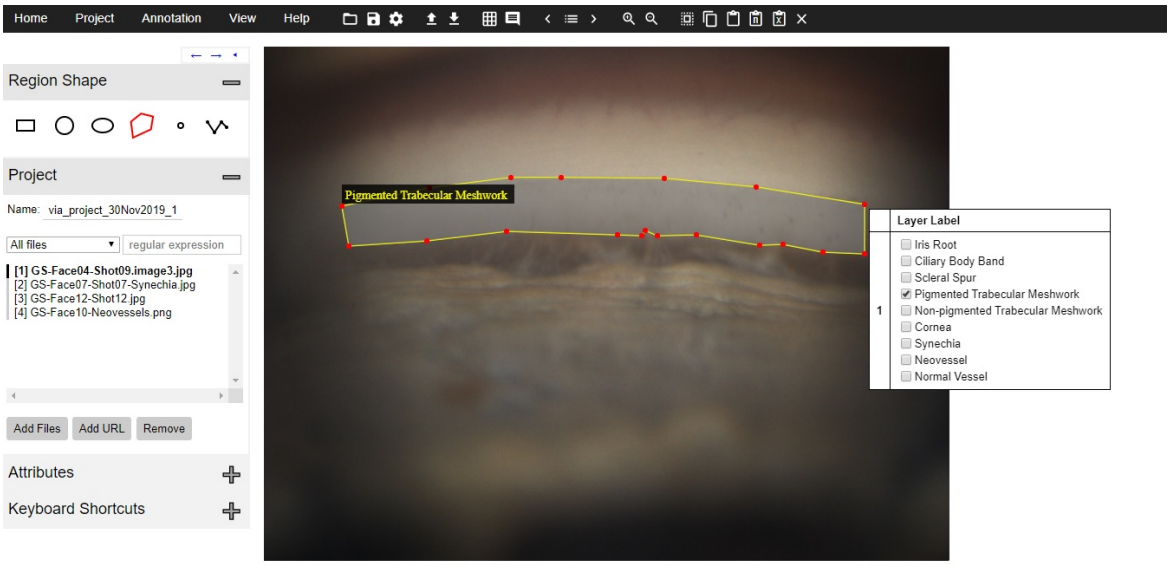

Fig. 3.1 *VGG Image Annotator user interface with an example of digital gonio-photograph showing a sector of the anterior chamber angle and the annotation of one layer.*

3.3.2 Annotation Protocol

The annotation protocol was conceived to deal with specific target and image features in order to promote consistency across annotators.

The full protocol document, provided to the clinical annotators, can be found in Appendix B. The protocol concerns the annotation of the six anatomical layers, namely: the iris root, the ciliary body band, the scleral spur, the pigmented (or posterior) trabecular meshwork, the non-pigmented (or anterior) trabecular meshwork and the cornea. Please refer to Chapter 1 for a description of the layers. Note that not all of them may be visible simultaneously in each gonio-photograph.

Importantly, the protocol states that layers must be annotated only if the clinician is confident in detecting their boundaries with adjacent tissues by looking solely at the images and not by making hypothesis on their location based on medical knowledge. In fact, layers of the anterior chamber angle have a consistent structural sequence and the presence of one of them may be deducted by the visibility of the adjacent ones even if its boundaries are not clearly detectable in an image. Sometimes a layer may not be visible at all meaning that it is

occluded by the iris (appositional or synechial closure) or by another foreign object, e.g., a stent (a small implant that facilitates the aqueous humour outflow).

The protocol also includes guidelines aimed at dealing with specific digital gonio-photographs characteristics, mainly due to the limited depth-of-field that usually leads to a blurred (out of focus) and dark (vignette) image periphery.

The main guidelines are reported here, with examples:

- annotate only the part of the image that is bright enough and in-focus so that layers boundaries can be identified (Figure 3.2);

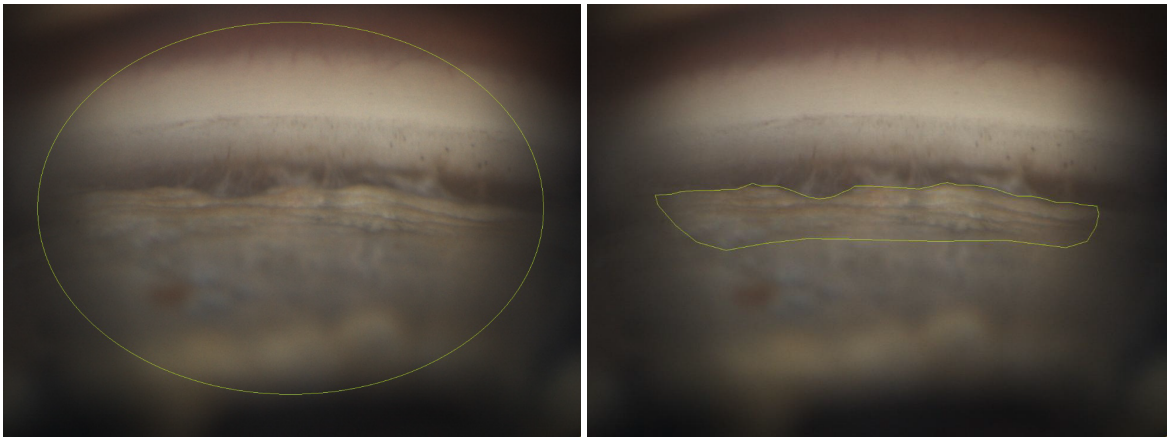

Fig. 3.2 *Examples of: (a) bright image region and (b) in-focus part of the iris-root.*

- do not overlap delineations of different layers;
- annotate layer-layer interfaces as accurately as possible, without gaps (Figure 3.3);
- choose the number and placement of polygon vertices so that the delineation does not include pixels from adjacent layers.

As brightness and focus vary smoothly across the frame, the identification of the *informative* image region is subjective and does not rely on the morphology of the angle layers. Despite the protocol guidelines, annotations could present overlapping regions. We considered the size of each of these overlapping regions to decide how to deal with the issue time-by-time. In case of small (a few pixel wide) regions, the overlapping area was arbitrarily

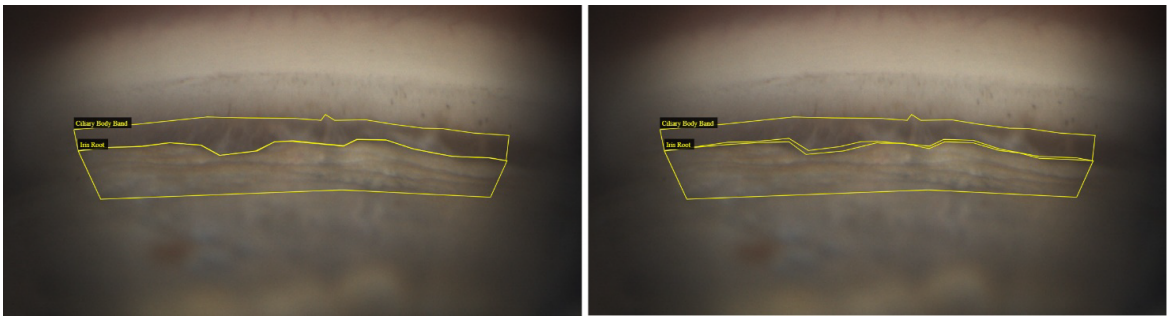

Fig. 3.3 *Examples of annotations without (a) and with (b) gaps between delineations.*

assigned to the smaller layer. In case of larger overlapping areas, the clinician was asked to refine the ground truth. Gaps between delineations were to avoid whenever the local detail of the image was sufficient to identify interfaces. However, clinicians might ignore image areas they could not identify clearly. This was not an issue for our studies, as the un-annotated image areas were managed properly as described in Chapter 4 and Chapter 5 in order not to compromise the reliability of our results.

The annotation protocol also reports instructions on how to annotate other features of interest in digital gonio-photographs. They aim to locate pathological traits like neo-vessels and synechiae and to provide a classification for the Schwalbe's line (which can not be segmented easily using a polygon due to its very limited thickness). The instructions for annotating these additional features are not discussed here, since they have not been used in the research conducted so far.

Annotations of pigmented and non-pigmented trabecular meshwork have been merged before use as a way to circumvent the low agreement among ophthalmologists on the location of their mutual interface (see Chapter 4 for more details).

Different sets of images have been annotated by possibly non-identical groups of annotators depending on the purpose of the studies described in the following chapters (either the development of the semantic segmentation algorithm or the inter-annotator variability study). A description of the datasets and the number and affiliations of the annotators who provided the ground truths for a specific study will be always provided.

3.4 Annotations for Aperture Classification

Annotations for the image classification task consist in labels associated to specific sub-areas to be processed independently from others, as discussed below for our angle aperture system.

In this case, annotations were performed by clinicians on a software-generated all-in-focus version of each exam sector image, obtained using Navis-Ex (version 1.11.0.6) and the GS-1 Viewer (version 1.0.0.3, not commercially available yet), two proprietary software developed by NIDEK Co., LTD. The algorithm that produces all-in-focus sector representations uses the full stack of images acquired by the NIDEK GS-1 device at different focus planes, and merges the in-focus area of each of them in a single frame. This approach has been considered advantageous in this case since all-in-focus representations allow the clinician to evaluate angle aperture more clearly. Software-generated all-in-focus images would have not been suitable for segmentation purposes, since the blending algorithm may produce artefacts not perceivable by humans but possibly affecting algorithms designed to return pixel-level delineations of targets.

3.4.1 Annotation Tool

The annotation tool used to grade angle aperture in digital gonio-photographs was developed by NIDEK Technologies Srl. and made available for this research. An annotation protocol has been devised in collaboration with clinical experts. The tool complies with GDPR and can be run on Microsoft Windows devices.

The user interface of the annotation tool with examples of images and annotations thereof is depicted in Figure 3.4.

Sector images were first rotated so that they showed the iris at the bottom of the frame, and a region of interest (960 x 960 pixels) was selected to highlight the central area of each picture. A grid overlaid on the sector image identified three *sub-sectors* to be annotated independently. The reason of this choice was to generate local aperture ground truth as opposed to coarser aperture characterizations in previous work, as will be clarified in Chapter 6 when discussing our study on algorithms for angle aperture classification in gonio-photographs. The choice

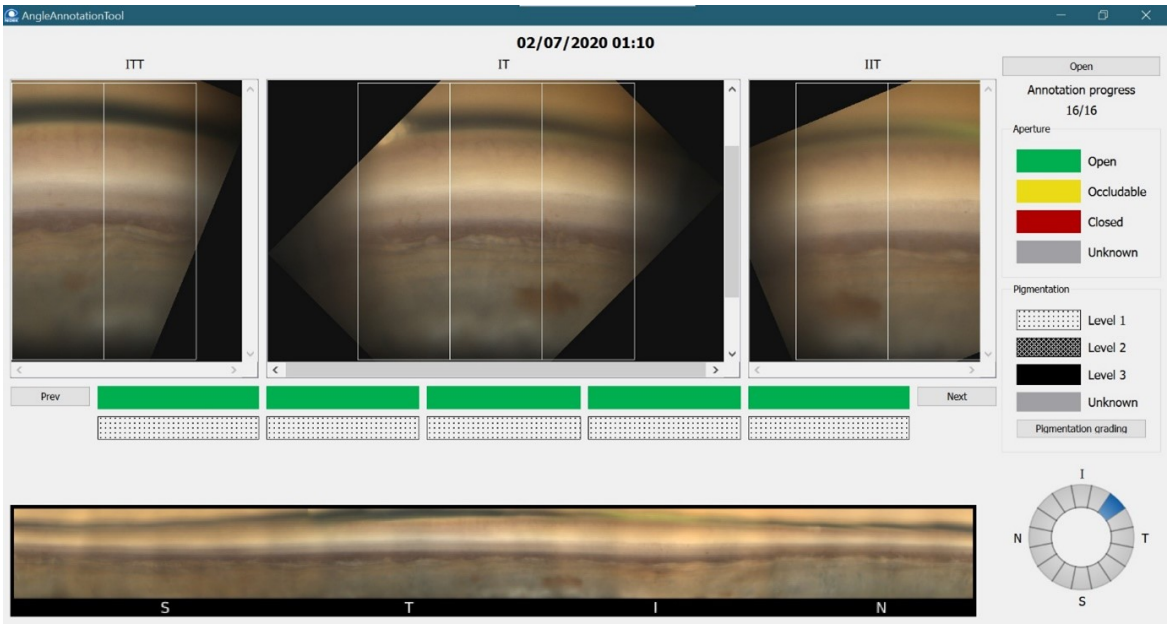

Fig. 3.4 *Annotation tool for grading angle aperture in digital gonio-photographs acquired with the NIDEK GS-1 device. It shows examples of gonio-photographs and the angle aperture annotations for their sub-sectors (the first row of coloured rectangles).*

of dividing each sector image into three sub-sectors was a compromise between several technical and clinical considerations. An aperture classification based on three sub-sectors each covering 5°-7° degrees of the angle interface would be a huge improvement with respect to previous proposed algorithms (classification of 90°-wide sectors) as it would ideally allow to spot small synechiae. It is also a good compromise in terms of time spent annotating the data, and likely provides enough context for the deep learning model to associate anatomical traits to the corresponding aperture grades.

The annotation tool allows to navigate among images of the same exam, to inspect them, select and review labels. Annotations were conveniently stored in *.xml* files.

3.4.2 Annotation Protocol

The protocol for this task concerns the annotation of angle sub-sectors into three aperture classes based on the visible angle layers. The full protocol document, provided to the clinical annotators, can be found in Appendix C.

The classification criteria rely on a clinical grading system called *Spaeth's system* [117] and an additional class aggregation criteria approved by experts. The Spaeth's system grades the angle aperture into five classes according to the apparent iris insertion (which implicitly relies on the visibility of angle layers through gonioscopy, refer to Appendix C for more details). Our aggregation simplifies the problem to three classes and an additional class is used for un-gradable sub-sectors (e.g., if the area is hidden by artefacts like bubbles or shadows). The classes are:

- ***open*** *(colour code: GREEN)*: scleral spur and / or ciliary body band visible (Spaeth's grades D and E);
- ***occludable*** *(colour code: YELLOW)*: Schwalbe's line and trabecular meshwork visible to some extent (either posterior or anterior), scleral spur and ciliary body band not visible (Spaeth's grade C);
- ***closed*** *(colour code: RED)*: trabecular meshwork not visible. Schwalbe's line can be visible or not (Spaeth's grades A and B);
- ***unknown*** *(colour code: GREY)*: the angle is not visible due to misalignment, because its view is prevented from obstacles (e.g. bubbles, eye lashes), or the quality of the sub-sector (e.g. sharpness and / or illumination) is not good enough to evaluate the structures.

Spaeth's grade C considers the trabecular meshwork as a single layer (pigmented and non-pigmented parts together). It is advantageous since it is often very difficult to discern between the two areas (as found in Chapter 4).

The protocol also reports how to deal with ambiguous cases: *"In case the angle aperture varies in the considered sub-sector, the classification that better describes it (that is, applicable to the most part of it) shall be chosen. In case it is not possible to assess which classification is predominant in the sub-sector, the one corresponding to the more anterior iris insertion shall be chosen (e.g., if part sub-sector is Open and part Occludable, the classification shall be Occludable; if some is Occludable and some Closed, the classification*

shall be Closed). If a sub-sector is in part gradable (Open, Occludable or Closed) and in part un-gradable (Unknown) the classification shall be that of the gradable part". The choice of assigning the more severe class when it is not possible to decide which aperture grade is predominant in a sub-sector was made to increase the sensitivity of the system to pathological cases (i.e., to make it favour more pessimistic output classes in ambiguous cases), as it is preferable according to ophthalmologists.

The annotation protocol also reports instructions on how to label angle sub-sectors according to the variability of another feature of clinical interest in digital gonio-photographs, which is the trabecular meshwork pigmentation. These annotations have not been used in our research yet and will not be discussed in this thesis.

3.5 Discussion and Conclusions

Clinical requirements have been collected from a network of experts located at different institutions worldwide. The tasks to investigate solutions for and thus focus our research questions have been selected based on: (i) the results of the questionnaire (the template of which is reported in Appendix A), (ii) the opportunities they could give for future developments (e.g., potential utility for multiple applications) and (iii) their industrial interest,. Ophthalmologists have been involved in the design of the annotation protocols for the different research phases and agreed on their final versions. They were consulted when technical compromises (e.g., class aggregation criteria) were deemed necessary and approved them. This approach ensured that annotations were clinically meaningful and suitable for the development of automatic systems for image analysis.

Chapter 4

Inter-annotator Variability Study

4.1 About this Chapter

Obtaining ground-truth (i.e., annotations) for training and validating machine learning algorithms for real-life applications is usually costly and difficult. This is particularly challenging in medical data analysis where the exact classification of tissues, their detection or their segmentation may be impossible without invasive procedures (e.g., biopsy) and usually rely on non-invasive inspections, e.g., MRI or CT scans. This is, for example, the case of mole classification (as benign or malign) often performed, at least initially, by visual evaluation rather than through biopsy.

The reliability of the ground truth obtained in this way may be affected by many factors involving both the data and the observer(s). For example, the quality of data (e.g., the signal-to-noise ratio of an EEG or the sharpness of a picture) is crucial for inferring parameters of clinical relevance with confidence. In turn, quality itself may depend on several factors, among which the acquisition protocol, the experience of the person responsible for data collection and the differences between instruments (e.g., different makers).

Observers may also be conditioned by their specific experience/preparation, that may introduce systematic bias between experts (i.e., *inter-annotator variability*), and other factors that may affect the repeatability of their own analysis (i.e., *intra-annotator variability*), such as tiredness.

The variability in the ground truth provided by a group of experts has direct consequences on the design, training and validation of systems for automated data analysis and must be considered carefully [75]. Collecting ground truth from multiple observers is encouraged in the current literature and modelling annotation variability is a fundamental requisite for meaningful software validation [52].

This chapter reports, to our best knowledge, the first inter-annotator variability study on manual delineations of anterior chamber angle layers in digital gonio-photographs. It aims to provide a comprehensive context for the correct evaluation and validation of algorithms performing an automated analysis of digital gonio-photographs, in particular for detection and segmentation of anatomical layers.

Note that current literature about automatic analysis systems for gonio-photographs is limited to the classification of images using labels describing large anatomical regions [14]. However, supported by the opinion of experts, we argue that the assessment of other clinically relevant features could benefit from a local, rather than global, characterization of the anterior chamber angle anatomy (i.e., a pixel-wise classification or segmentation). A precise delineation of anterior chamber angle layers could be advantageous, for example, for measuring synechial closure extension and its changes over time, or segmenting the trabecular meshwork to allow automatic pigmentation grading (e.g., prior to laser trabeculoplasty). Moreover, auto-alignment and auto-tracking systems based on layers segmentation could improve examinations in remote and virtual clinics, which are currently gaining importance due to the COVID-19 pandemic. This inter-annotator study on dense annotations of anterior chamber angle layers has been motivated by all these premises.

Intra-annotator variability was not investigated due to the prohibitive effort required to collect multiple annotations for a representative set of images from the experts and to the issues caused to healthcare systems by the pandemic.

This study has been published in *Translational Vision Science and Technology* [89].

4.2 Materials

Data

As part of the data annotation process described in Section 3.3.2 for the development of the semantic segmentation algorithm, a sub-set of the digital gonio-photographs was chosen to be annotated by multiple ophthalmologists. 20 angle sector images, each depicting a 22.5 degree wide portion of the irido-corneal interface, were selected from 18 eyes of 17 patients and used to study inter-annotator variability.

Digital gonio-photographs of the anterior chamber angle consist of 1280 x 960 (width, height), RGB images stored in *.jpeg* image format, acquired using a NIDEK GS-1 device at two clinical sites located in Genova (Italy) and Dundee (United Kingdom).

Figure 4.1 shows two anterior chamber angle sectors, acquired with a NIDEK GS-1.

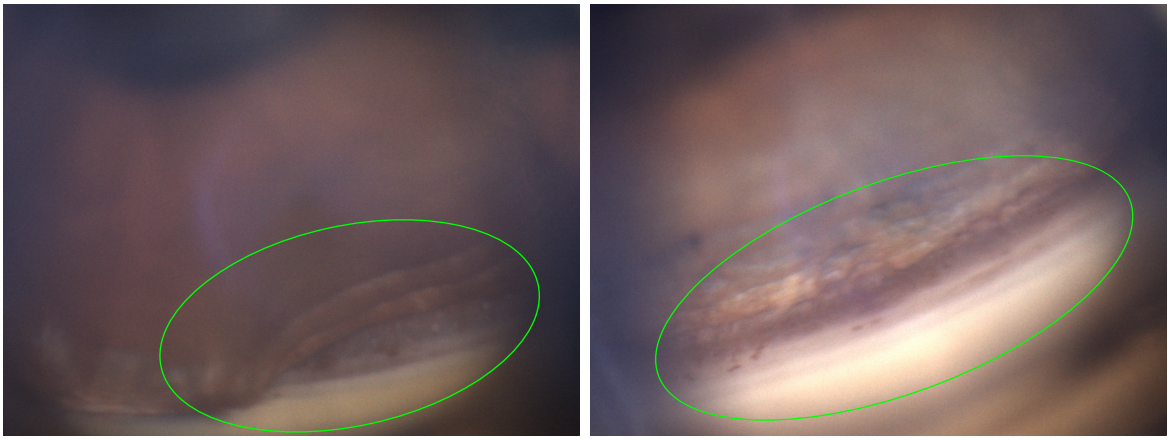

Fig. 4.1 *Example of two images of anterior chamber angle sectors taken with a NIDEK GS-1 device. The in-focus and bright areas are highlighted by green ellipses.*

As described in Chapter 1, the NIDEK GS-1 takes several shots per angle sector by progressively changing the focal plane in order to acquire the full depth of the three-dimensional drainage angle. For each sector considered, the image with the focus on the edge of the drainage angle (i.e., on the ciliary body band or the scleral spur if the angle sector was open or on the iris–cornea interface if it was closed) was selected to provide the sharpest (highest contrast) visualization of the layer interfaces. As previously mentioned, the acquisition device automatically selects from the acquisition stack the shot with the estimated best focus

on the edge of the angle, through a proprietary algorithm [10, 138]; automatically selected best-shot images have been inspected anyway by the author of this thesis and modified in the unlikely case they were not optimal. The iridocorneal interface region may be not always centred in the frame. Note that the limited depth of field implies that the inner portion of the iris and the outer portion of the cornea may appear blurred and that a vignette is visible in most images, whereby the image periphery appears darker than its centre. Both phenomenon are visible outside the green ellipses in Figure 4.1.

Image selection was not based on acquisition conditions or the patient's diagnosis (e.g., ocular hypertension, glaucoma) but only on local layers morphology, as this study aimed to assess inter-annotator variability on descriptive image features and not to relate these features to diagnosis. A clinical stratification of patients was, thus, not relevant and is not provided.

Rather, the images selected are representative of the variability of visual features of the eye region observed in clinical practice, such as iris colour and trabecular meshwork pigmentation, and include relevant local variations of layers interfaces, such as appositional angle closure and peripheral anterior synechiae (see Chapter 1 for a description of these conditions). More in detail, the images included 6 light and 14 dark irises (where blue or green eyes were considered light and brown eyes were considered dark); 5 highly and 11 slightly pigmented trabecular meshworks in non-closed angle sectors (where slightly pigmented corresponds to Scheie's pigmentation grades[1] none, 1, and 2, and highly pigmented corresponds to Scheie's grades 3 and 4); four angle-closure images defined as appositional irido-corneal contact in at least 50% of the sector (Scheie's aperture grade[2] 4); and four images showing anterior synechiae.

Annotations

Images were annotated according to the annotation protocol described in Section 3.3.2. We remind that "to annotate" means to trace the contours of the layers visible in the image and

[1]The Scheie's pigmentation grading scale is a clinical standard based on pigmentation density. It ranges from 0 (or None) to 4, the higher the number the more pigmented the trabecular meshwork [106].

[2]The Scheie's aperture grading scale is a clinical standard based on the visibility of angle layers. It ranges from 0 (or Wide) to 4, the higher the number the narrower the anterior chamber angle [106].

assign them the correct label using the annotation tool (VGG Image Annotator 2.0.8 [21]). Image regions were highlighted using polygonal shapes, and labels selected from a list of pre-defined entries.

The annotation of a single image took, on average, 10 minutes. It was unfeasible to obtain multiple annotations for more images at this stage, since the delineation of layers is a time-consuming process and the availability of clinical experts to annotate data was limited.

The annotation protocol stipulates to annotate only the sharp (in-focus) and bright image areas; to avoid tracing the contours of target layers that were not clearly identifiable with respect to neighbouring ones; and to trace layer–layer interfaces as precisely as possible. We highlight that this protocol is designed to generate annotations suitable for validating automatic segmentation systems, and not to necessarily reflect the normal practice of clinicians.

The specific subset of images used to evaluate inter-annotator variability was annotated by five ophthalmologists from four clinical institutions located in Genoa, Italy; Lisbon, Portugal; Dundee, United Kingdom; and Los Angeles, California, USA. Clinicians had different levels of experience; this was not included as a parameter to model inter-annotator variability. When they performed the annotations, two annotators were, respectively, year 4 and 7 specialty trainees with experience in gonioscopy; one was a clinical study investigator with 5 years of experience in an image reading centre; one was a glaucoma specialist with 5 years of clinical experience in glaucoma management; and one had more than 10 years of clinical experience in tertiary referral centres.

Annotations consist of a set of non-overlapping polygonal contours enclosing the bright and sharp area of each anterior chamber angle layer considered in this study: iris root (IR), ciliary body band (CBB), scleral spur (SS), trabecular meshwork (TM), and cornea (C). Since the annotation tool did not provide a way to automatically avoid different polygons to overlap, the raw annotations were pre-processed in order to deal with small overlapping areas (width of few pixels) by being arbitrarily assigned to the layer with the smaller surface. Larger overlapping regions were re-annotated by the clinical expert. Gaps between layers are not an issue according to our methods (please refer to the next section for detailed information). As anticipated in Section 3.3.2 pigmented and non-pigmented trabecular meshwork regions,

although annotated independently, were merged together before use. This was done to increase annotations reliability when training the semantic segmentation system (Chapter 5) and to conduct the inter-annotator variability study accordingly. Even if the discussion on inter-annotator variability will mainly consider the trabecular meshwork as a whole, a few useful comparisons with the results obtained without merging the two sub-regions will be provided to justify our choice.

Image characteristics and protocol guidelines had important consequences on the annotations. For example, even if the layers span the whole image from side to side, vignetting and blur often make the annotators ignore the image periphery. As a result, most images have annotations concentrated in a limited image area (Figure 4.2). A degree of subjectivity was always present, for instance when locating the transition between in-focus and blurred regions of the iris or the transition between well-lit and dark regions within the trabecular meshwork. Moreover, annotators could choose not to annotate part of an image if they did not feel sufficiently confident (e.g., if they judged the quality of an image region too poor). This did not necessarily indicate disagreement with the other annotators, but resulted in part of each image being left unannotated (labelled *NA*). Overall, each clinician annotated at least one anatomical layer in every image.

4.3 Methods

In our analysis, inter-annotator variability only accounted for annotated image regions and was not affected by un-annotated areas. This ensured that variability measures reflected only differences in clinical judgement made with confidence.

An example of digital gonio-photograph and an annotation thereof is shown in Figure 4.2. All of the image pixels within a delineation were labelled as pertaining to a single layer. The two pixels highlighted in the figure belong to the same anterior chamber angle layer (iris root), but the point labelled “2” in the image was not included in the annotated region, given the (subjectively) estimated border between well-lit and dark regions. This is a departure from many inter-annotator variability studies for segmentation systems in medical image

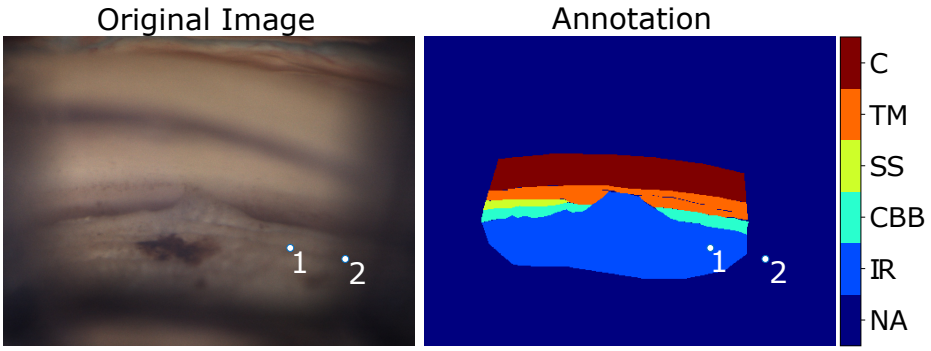

Fig. 4.2 *Original gonio-photograph (left) and an annotation (right). Points 1 and 2 highlight two pixels in the iris root. Point 2 has been excluded from the annotation, given the subjective estimation of the transition between the bright and the dark image regions.*

analysis [52, 96], which typically expect the full extent of targets to be annotated. These choices imply that standard analysis methods, such as consensus and comparison metrics, such as the Dice score, required adapting in order to be used consistently. Inter-annotator variability was analysed in three experiments, reported below.

4.3.1 Layer-wise Annotation Frequency

Layer-wise annotation frequency refers to the number of times each clinician delineated the contours of each structure, as a measure of their confidence at recognizing and locating drainage angle layers in digital gonio-photographs. Annotators were instructed to trace contours only when they judged them to be clearly visible. Occasionally, some layers were not visible at all; for example, the scleral spur was not visible in the case of angle closure. For this experiment, we were only interested in the existence of a layer annotation, not in its geometry; hence, two annotators could be equally confident in delineating a specific layer even if the two actual contours differed. This methodology was designed to provide insight into the variability in experts' confidence in identifying anatomical layers from local image features of digital gonio-photographs.

4.3.2 Layer-wise Consensus

This analysis examined consensus by the number of pixels agreed to be part of a given layer by a minimum number of annotators. Its size was plotted as a function of the minimum number of agreeing annotators, i.e., the *consensus threshold*. The purpose was to obtain an indication of which layers were annotated with high and low consistency among the annotators in terms of location and size. In the literature [52], the consensus of multiple annotations of the same image is usually computed as the subset of pixels labelled in the same way by at least n (consensus threshold) annotators, with all of the other pixels considered as background. The consensus is always maximum when n equals 1. In that case, in fact, the consensus area for a given target is trivially the union of all image pixels annotated as belonging to that target by at least one of the annotators. Consensus is expected to decrease as the threshold n increases unless the delineations drawn by several annotators coincide perfectly.

We adapted this concept for our study to deal correctly with the un-annotated areas (i.e., not background), defining a three-category label for each pixel, as follows:

1. *consensus region (label 1)*: the set of pixels annotated as appertaining to a given layer by at least n observers. For example, given 5 annotators and $n = 3$, the iris root consensus region would be the set of pixels classified consistently as iris by at least 3 experts, regardless of who these experts are;

2. *disagreement region (label –1)*: the set of pixels annotated as the given layer by k annotators, with $1 \leq k < n$, and differently (i.e., belonging to another layer) by at least one. According to the example introduced earlier, given 5 annotators and $n = 3$, this region would collect all the pixels classified as iris by one or two annotators and differently, e.g., ciliary body band, by at least one expert;

3. *ignored region (label 0)*: the set of pixels annotated as the given layer by k annotators, with $k < n$, and left un-annotated by the others; this region is ignored when computing consensus size variations, as its variability does not necessarily reflect changes of the actual agreement level. In the example, given 5 annotators and $n = 3$, this area would

comprise all those pixels annotated as iris by less than 3 experts and left un-annotated by the others.

It follows that NA image regions do not affect consensus computation according to our experimental design.

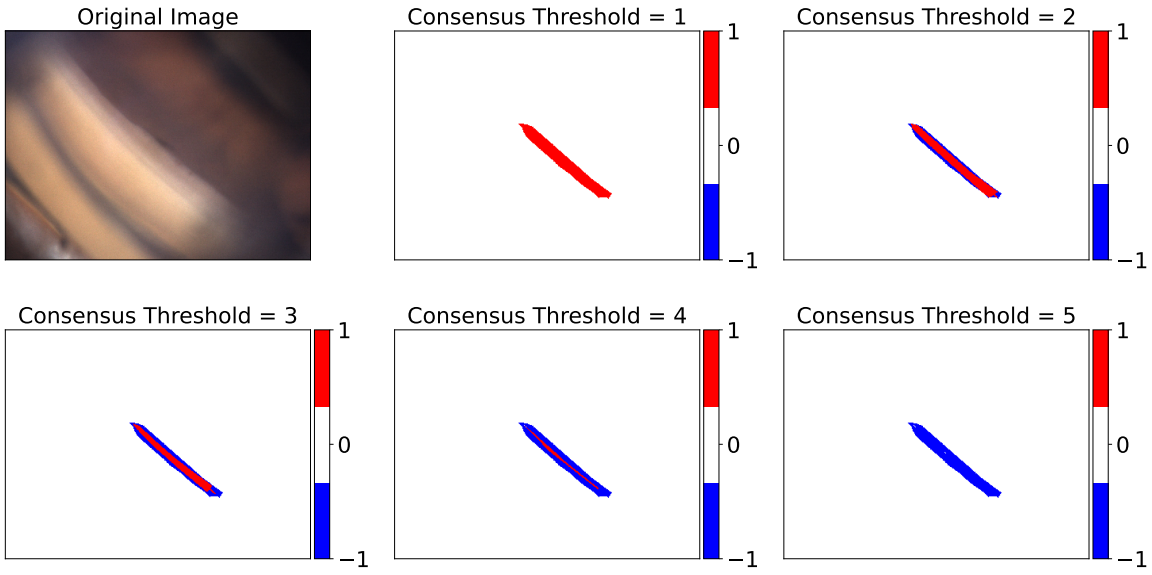

Fig. 4.3 *Original RGB image (top left) and the five scleral spur consensus maps as the consensus threshold varies. Label 1 is the consensus region, –1 is the disagreement region, and 0 is the ignored region.*

An example of how the consensus region varied with the consensus threshold is shown in Figure 4.3. The extent of disagreement increased with the consensus threshold value, as expected. In the ideal case of perfect agreement among all the annotators, the agreement region size would not change varying the threshold.

After generating five consensus maps for each target layer, one per threshold value (when the threshold equalled 1 it led to the union of annotated pixels, and when it equalled 5 it led to their intersection), we studied how the consensus size decreased as the threshold increased.

4.3.3 Agreement Analysis

This analysis compared agreement between pairs of annotators, with one annotator chosen as reference for each pair. Each comparison yielded a 5×5 confusion matrix, given that there were five target anatomical classes. Un-annotated regions were excluded from the

computation so that intersections between areas that were annotated by one annotator but not by the other one did not affect the results. Layer-wise precision, sensitivity, and Dice scores of each annotator were calculated as follows:

- *Precision*: $TP/(TP+FP)$, where TP is the true positives and FP is the false positives.
- *Sensitivity*: $TP/(TP+FN)$, where FN is the false negatives.
- *Dice score*: $2 \cdot (precision * sensitivity)/(precision + sensitivity)$

Average values and standard deviations of each annotator were computed as an overall measure of inter-annotator agreement.

4.4 Results

4.4.1 Layer-Wise Annotation Frequency

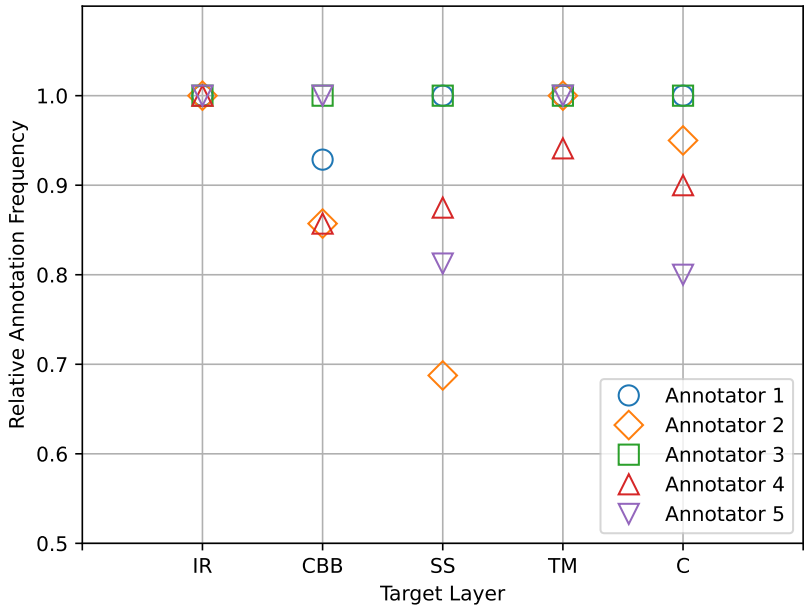

Fig. 4.4 *Plot of per-layer annotation frequencies for each annotator.*

Figure 4.4 shows the per-layer annotation frequency of each annotator, a measure of how confidently they identified and traced contours. Note that the fixed sequence of the layers provides expectations about what layers are present in a given location, but segmentation (contours) depends on local image features. The iris root was the only region segmented by all participants the same number of times (i.e., most consistently). This can be explained by considering that the boundary between the iris and the next visible layer is usually sharp and thus well identifiable, but this may not be true for other layers. The relative segmentation frequency measured for the remaining layers varied, up to a maximum percent difference of 31% for the scleral spur (annotator 2 vs. annotators 1 and 3). Only one participant (annotator 3) provided the observed maximum number of annotations for all of the layers.

4.4.2 Layer-wise Consensus

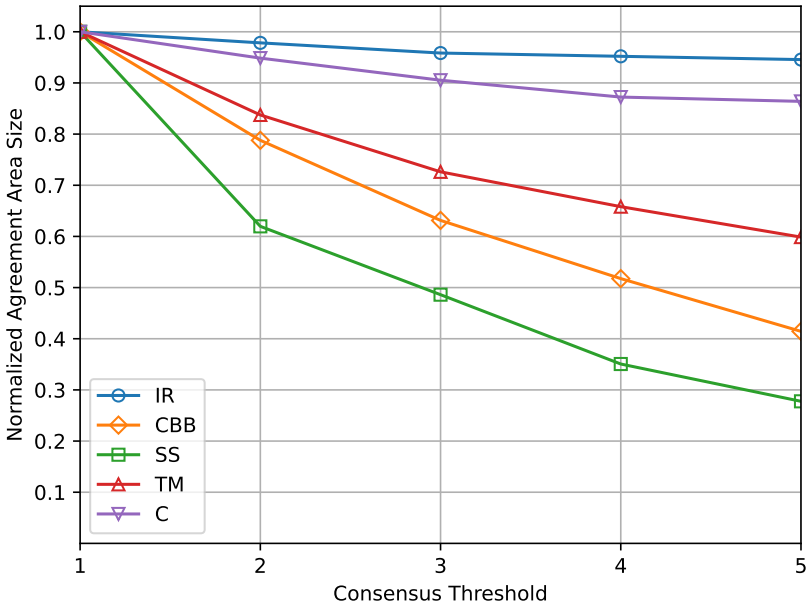

Fig. 4.5 *Plot of the ratio between consensus pixels and annotated pixels against the consensus threshold (minimum number of annotators agreeing).*

Figure 4.5 shows how the area of the consensus region for each layer decreased as the consensus threshold (minimum number of annotators agreeing) increased. Per-layer

consensus areas have been normalized to [0, 1] where 1 means that, for a given layer and consensus threshold, all annotated pixels are also agreement pixels. Thus, the normalized consensus value does not depend on the size of the annotations.

As previously mentioned, a consensus variation only occurred if the consensus threshold exceeded the actual number of agreeing annotators for a given pixel and at least one of them disagreed. This ensured that we only considered actual pixel-wise classification differences and not the subjective choice to not annotate an image region. The plot suggests that the consensus levels on some layers were low; for example, the minimum average agreement (i.e., the agreement calculated when the consensus threshold equals the number of annotators, averaged over all the images in the dataset) was only about 28% of the annotated pixels for the scleral spur. It is also worth noticing that, although the consensus for the cornea and iris root converged to an almost stable percentage for high consensus threshold values, the consensus for the trabecular meshwork, scleral spur, and ciliary body band kept decreasing approximately linearly (for thresholds ≥ 2).

4.4.3 Agreement Analysis

The ground truth provided by every pair of participants was compared, generating a set of 5×5 confusion matrices, given that five was the number of anterior chamber angle layers considered. The cell $C_{i,j}$ of a confusion matrix gives the number of pixels that belong to the intersection between the annotation of target i by the first annotator and the annotation of target j by the second annotator, taken as reference. Perfect agreement would result in a diagonal confusion matrix. Three layer-wise metrics of inter-annotator agreement were computed from each confusion matrix: precision, sensitivity, and Dice score. Specificity was not considered since layer annotations are often small compared to the image area and the variability would only slightly affect the specificity values thus leading to overoptimistic results. Finally, per-annotator mean values and standard deviations were obtained.

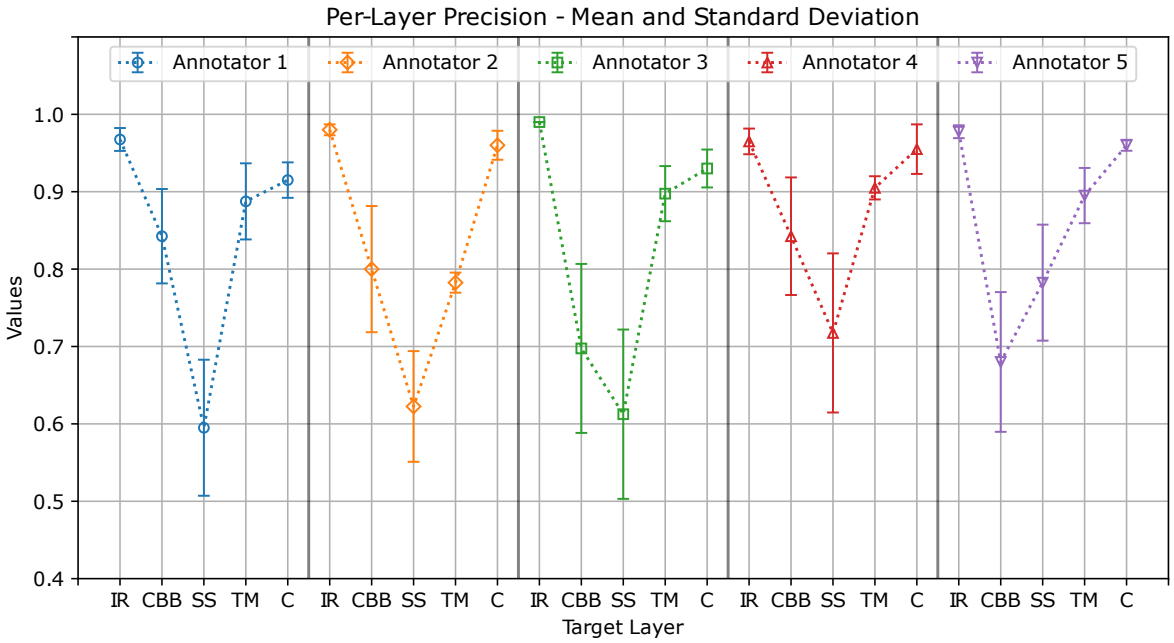

Fig. 4.6 *Annotators' average precision (plot points) and standard deviation (whiskers) when annotating each layer. Layers acronyms are reported in the x axis according to their anatomical topology. The order of annotators is not relevant.*

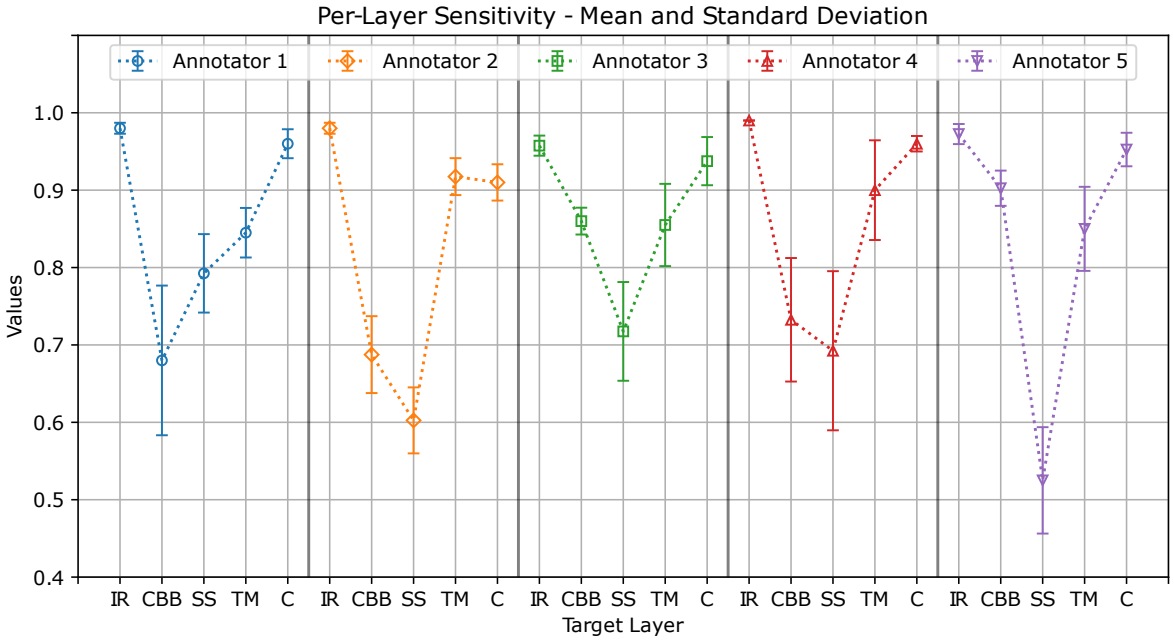

Fig. 4.7 *Annotators' average sensitivity (plot points) and standard deviation (whiskers) when annotating each layer. Layers acronyms are reported in the x axis according to their anatomical topology. The order of annotators is not relevant.*

The annotators' mean precision and standard deviation values for each layer are reported in Figure 4.6. The ciliary body band and scleral spur overall show the lowest mean precision

values and/or the highest standard deviations. Figure 4.7 shows the mean sensitivity values and standard deviations. As in the case of precision, the maximum overall variability occurred for the ciliary body band and scleral spur.

Comparing Figures 4.6 and 4.7 gives us an insight into the differences between annotations from different clinicians. For a specific annotator and target, good precision but low sensitivity suggests that, compared with contours traced by others, the area delineated was thinner but centred on average, thus generating a prevalence of false-negative classifications (e.g., for annotator 1, ciliary body band). Good sensitivity but lower precision suggests that the delineated area was larger but centred on average, generating a prevalence of false-positive classifications (e.g., for annotator 5, ciliary body band). Low precision and sensitivity suggest that the delineated area was displaced from the average (e.g., for annotator 2, scleral spur).

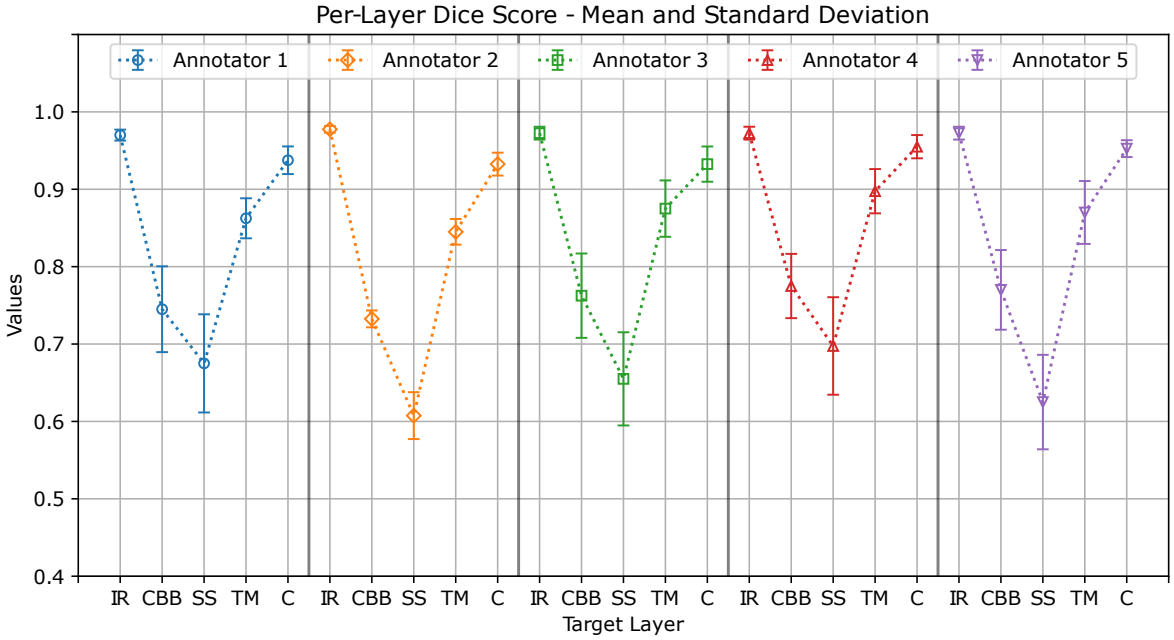

Fig. 4.8 *Annotators' average Dice score (plot points) and standard deviation (whiskers) when annotating each layer. Layers acronyms are reported in the x axis according to their anatomical topology. The order of annotators is not relevant.*

Figure 4.8 shows the mean Dice scores and corresponding standard deviation values, providing a quantitative metric to compare annotations from different clinicians. The graphs show a common pattern: good agreement on the iris root, trabecular meshwork, and cornea

(with average Dice score ranges of 0.97–0.98, 0.84–0.9, and 0.93–0.96, respectively) and lower agreement on the scleral spur and ciliary body band (with average Dice score ranges of 0.61–0.7 and 0.73–0.78, respectively).

In all previous plots, the segment of the horizontal axis referring to any specific annotator reports the *ordered* sequence of angle layers according to their natural topology. The order of annotators is not relevant instead. It is worth highlighting that the iris root and the cornea (the first and the last in the sequence) have only the top or, respectively, the bottom annotation border bounded by other layers, making it reasonable to expect higher average precision, sensitivity and Dice score values as confirmed by our results. In case of the iris root, its interface with the following layer (or layers if the morphology is pathological, e.g., in case of a synechia) is usually much sharper than those between any two other layers and easier to delineate in our images. Its variability was low in general in our dataset, except when the iris and the ciliary body band are characterized by a very similar shade of brown which makes the identification of their mutual boundary more difficult.

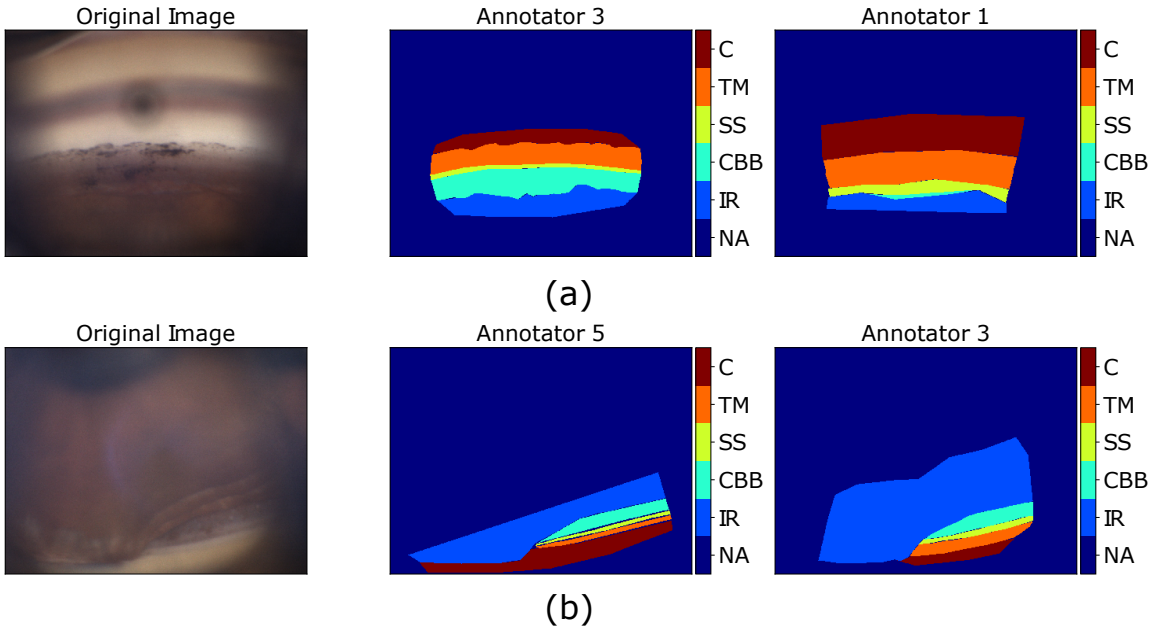

Fig. 4.9 *Visual representation of cases that led to low agreement metric values. (a) Low-precision CBB annotation and low Dice score SS annotation. (b) Low-sensitivity TM annotation.*

Figure 4.9 shows two examples of annotations that led to low values of per-layer agreement metrics. Figure 4.9a compares annotator 3 and annotator 1. The ciliary body band

annotation provided by annotator 3 included that of annotator 1, but it was larger and generated false positives in regions annotated differently by annotator 1 and, in turn, a low precision score. In the same comparison, the two scleral spur annotations cover different regions of the image, thus returning a low Dice score (low sensitivity and low precision). Figure 4.9b compares annotator 5 with annotator 3. The trabecular meshwork annotation of annotator 5 was included in that of annotator 3 but it was thinner, which caused false negatives in part of the region annotated as cornea by annotator 5, thus returning a low sensitivity score.

4.5 On the Effect of Splitting the Trabecular Meshwork

In Section 3.3.2 we have anticipated that, although the annotation process for producing semantic segmentation ground truths initially considered two sub-regions of the trabecular meshwork separately (i.e., pigmented and non-pigmented), the segmentation algorithm (see Chapter 5) has been trained and evaluated after merging these two sub-regions in a single layer (simply called trabecular meshwork). The inter-annotator variability study presented so far has been conducted according to that choice, since our aim was to study the issue of ground-truth variability to provide the correct context for the evaluation of our segmentation algorithm.

However, in this section we briefly discuss how inter-annotator variability would vary if the annotations of the pigmented and non-pigmented trabecular meshworks were kept apart.

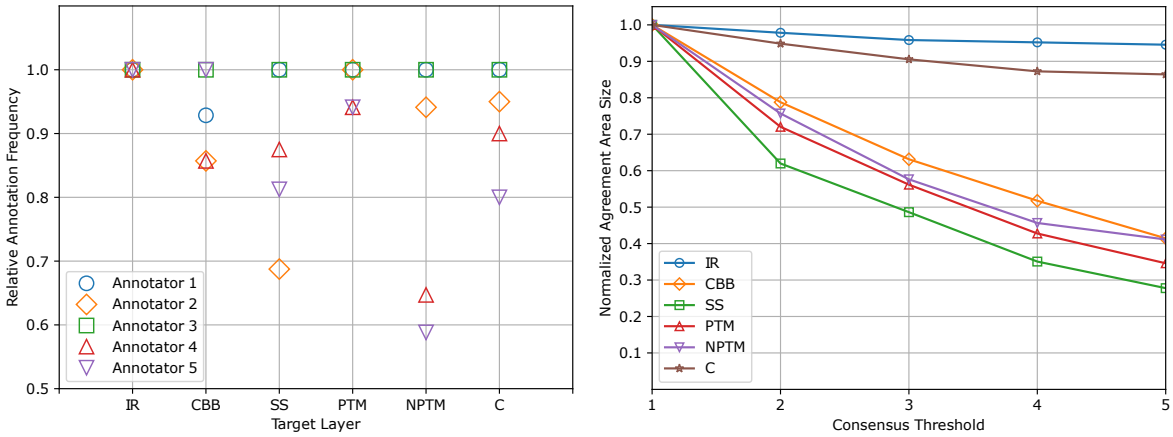

Fig. 4.10 *Layer annotation frequency plot (left) and consensus plot (right) when the trabecular meshwork annotation is split into its pigmented (PTM) and non-pigmented (NPTM) parts.*

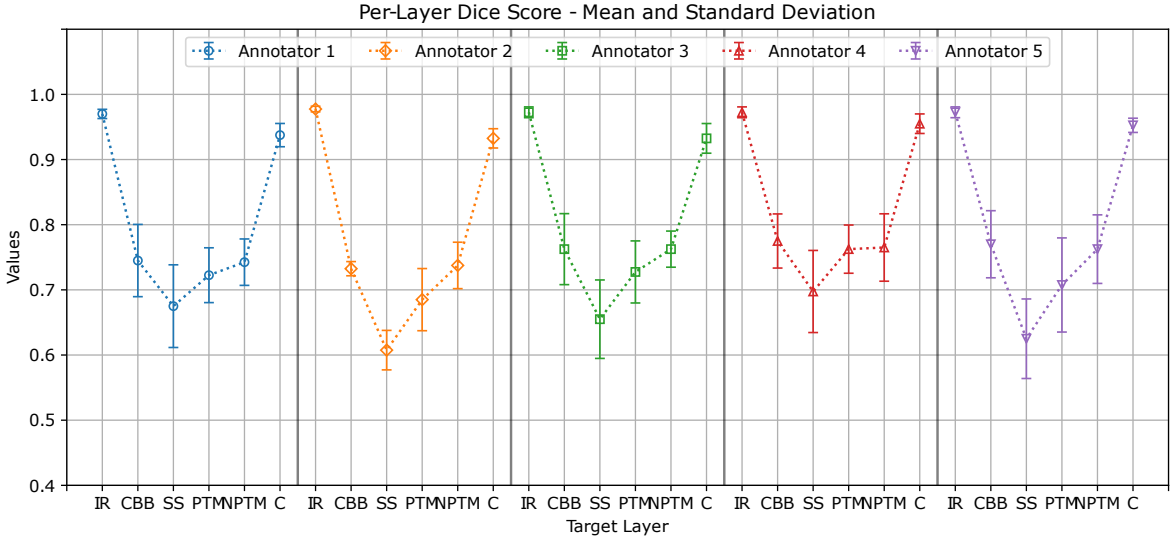

Fig. 4.11 *Annotators' average Dice score (plot points) and standard deviation (whiskers) when annotating each layer. Pigmented (PTM) and non-pigmented (NPTM) trabecular meshwork sub-regions are now two independent annotations.*

From the annotation frequency plot (Figure 4.10 (left)) it is possible to appreciate how the non-pigmented trabecular meshwork is now the layer annotated less frequently by some of the experts, with a maximum relative difference in the number of times it was delineated in the dataset equal to 41%. By looking at Figure 4.10 (right), the pigmented and non-pigmented regions show much worse consensus compared to when they are merged together (Figure 4.5).

Figure 4.11 shows that the average per-annotator Dice scores for pigmented and non-pigmented trabecular meshwork regions is lower that the values obtained when considering their surfaces as a whole.

Given the clinical importance of the trabecular meshwork and, in particular, of its pigmented (functional) component, these findings lead to an interesting hypothesis to be confirmed by future studies: since the decrease in metrics may be only due to the variability of the interface delineation between the pigmented and the non-pigmented trabecular meshworks, our results suggest that, despite the whole trabecular meshwork being delineated with fairly high consensus, the mutual boundary between its two sub-structures is difficult to localize precisely even by experienced ophthalmologists. This is to be accounted for when assessing algorithms for angle aperture grading since a few studies in the literature (e.g., [14]) consider the visibility of the pigmented part of the meshwork as the criterium to grade an angle section as *Open*.

4.6 Discussion and Conclusions

Well-designed data annotations are a crucial component of the development of reliable machine learning algorithms. When annotations from different experts are available, modelling their variability is important to interpret algorithm performance correctly, especially in the field of medical data analysis, where it is often impossible to obtain objective ground truth. In this study, to our best knowledge, we have presented the first inter-observer variability study on segmentations of anatomical layers in digital images of the anterior chamber angle. This study has been designed to support the evaluation of the deep learning algorithm for semantic segmentation of drainage angle layers presented in Chapter 5.

From the analysis of the annotations provided by five experienced ophthalmologists we obtained a detailed, quantitative description of the inter-annotator variability that can be summarized in the following points:

- providing contours of anterior chamber angle structures in digital gonio-photographs is challenging due to target feature variability (e.g., pigmentation, colour shades) and

image quality (e.g., illumination, sharpness, focus). This led to differences in the number of times the participants felt sufficiently confident to delineate target structures even where their presence was expected from anatomical knowledge;

- the consensus area of per-layer segmentation regions, defined as the number of pixels labelled the same by a minimum number of annotators, was much smaller for the scleral spur and ciliary body band compared with other layers (only about 28% and 41% of the pixels annotated as such by at least one expert). This result is particularly relevant because the scleral spur is an important marker to classify a drainage angle sector as fully open;

- the average values of agreement metrics showed a common pattern among annotators. High agreement values were found for structures with boundaries better characterized in terms of visual features of the images (e.g., contrast, colour, texture), namely, the iris root, trabecular meshwork, and cornea. Low agreement values were found for the ciliary body band and scleral spur regions.

This study has some limitations, in particular the limited numbers of images and ophthalmologists involved, although many papers in the literature of ophthalmic image analysis report experiments involving up to only three or four annotators. The reason for this is that generating annotations is time consuming, and clinical time is at a premium.

Nevertheless, our results provide important information on inter-annotator variability at delineating anatomical layers of the drainage angle in digital gonio-photographs, at least in two ways. First, they provide a quantitative context for interpreting values of assessment measures obtained when evaluating automatic systems (e.g., our semantic segmentation system in Chapter 5). Second, they give a first insight into the consensus of clinicians analysing digital gonio-photographs, which seems clearly dependent on specific layers. Given the variability of annotations by different experts, training and validating systems for automated gonioscopy with data acquired from several annotators seems strongly advisable to improve generalization. Estimating output uncertainty is necessary to highlight image features that are more difficult to classify (and possibly linked to increased inter-annotator

variability), thus improving interpretability and ultimately clinicians' trust in these algorithms. We included uncertainty estimation in our segmentation algorithm that will be discussed in Chapter 5.

Larger studies are advisable to firm up our conclusions to obtain truly reliable validation of artificial intelligence and machine learning applications for computer-aided analysis of gonioscopic images in the framework of evaluation of risk factors associated with glaucoma development, categorization of the disease, and support for the choice of treatments.

Chapter 5

Semantic Segmentation of Drainage Angle Layers

5.1 About this Chapter

This chapter aims to describe the deep learning model developed to perform the semantic segmentation of anatomical layers of the anterior chamber angle in digital gonio-photographs acquired with the NIDEK GS-1 device. The choice of this research target was based on the results of a consultation with participating clinicians and the evaluation of the advantages that this system could provide in practice, as discussed in Chapter 3.

As explained in Chapter 2, only very limited literature exists about the automatic analysis of gonio-photographs and only concerns the direct classification of angle aperture. We remind the reader that the grading of angle aperture implicitly relies on the visibility of the anatomical layers, the characteristics of those are thus expected to be somehow modelled by the classifier during training. Other information of clinical interest relies, at least partially, not just on the visibility of layers but also on their shape, size and location in the acquired images. This is, for example, the case of systems for grading the pigmentation of the trabecular meshwork (necessary to tune laser treatments), auto-alignment and auto-tracking algorithms for automating the examination in non-contact clinical settings, and augmented visualization of the eye region during the implantation of stents, to help the ophthalmologist insert the

device into the trabecular meshwork. Thus, the semantic segmentation of layers may provide a more exhaustive description of the relevant features of the iridocorneal angle and may be a powerful backbone for further analysis tools to support digital gonioscopy.

The algorithm we propose has been designed to deal with specific data characteristics (e.g., vignette and shallow depth-of-field) that would prevent other state-of-the-art segmentation systems from being effective. It aims to fill a gap in the literature about clinical applications of deep learning to gonioscopy and to provide a cornerstone for future developments.

This study has been published at different stages of development in the *Communications in Computer and Information Science* [88] and *BMJ Open Ophthalmology* [90] journals and in the *Investigative Ophthalmology & Visual Science* [91] as a conference abstract.

5.2 Materials

Data

A pilot dataset of digital gonio-photographs (1280 x 960 pixels, RGB), acquired with a NIDEK GS-1 device was selected from the databases shared by the clinical sites in Genova (Italy), Lisbon (Portugal) and Dundee (United Kingdom). A total of 274 sector images from 214 exams of 162 patients was annotated by four clinical experts (from Ninewells Hospitals, Dundee, UK; Hospital de Santa Maria, Lisbon, Portugal; and Ospedale San Martino, Genoa, Italy) according to the annotation protocol described in Section 3.3.2. Each sector image was annotated by only one expert, except for a subset of 20 images which was annotated by all the clinicians involved (plus another one who annotated only this subset) and used to study inter-annotator variability (Chapter 4). For the purposes of this study, whenever multiple annotations were available for a sector image, the one provided by the clinician with more experience (measured in years) was selected.

All images were acquired with patients' agreement and following General Data Protection Regulation rules (including anonymisation at source) during routine clinical examinations. Since this work focuses on the development of a software tool and is not an association study requiring cross-linked patient data, none was sought.

Table 5.1 *Segmentation dataset features distribution (% rounded at first decimal). Each image has been categorized by two main visual traits: the iris colour (rows) and an additional predominant feature of the sector (columns).*

	Anterior Synechiae	**Appositional Angle Closure**	**Highly Pigmented TM**	**Slightly Pigmented TM**
Light Iris	7 (2.6 %)	1 (0.4 %)	43 (15.7 %)	45 (16.4 %)
Dark Iris	43 (15.7 %)	24 (8.8 %)	60 (21.9 %)	51 (18.6 %)

Table 5.1 shows the distribution of the features of interest in our dataset by two main traits. The first is the iris colour, light (blue or green) or dark (brown); the second is the predominant feature of the angle sector, that can be only one of the following four: (1) the presence of anterior synechiae, (2) appositional closure of the angle (in at least half of the frame), or (3, 4) the trabecular meshwork pigmentation grade in all the other images (two cases: highly pigmented, corresponding to Scheie's grades II, III and IV; slightly pigmented, Scheie's grades None and I). Dark irises are predominant (65%), especially in the subsets representing synechiae and angle closures (86% and 96%). Moreover, structural changes in the angle layers represented by synechiae and angle closures are considerably under-represented in the dataset (27.5%).

Annotations

We remind the reader that these digital gonio-photographs show a narrow depth-of-field and vignetting, so that only part of them can be evaluated confidently. Deciding which part to evaluate is left to the annotators, introducing a degree of subjectivity in the ground truth and features correlation between the annotated and the un-annotated (label NA) image regions (refer to Section 3.3.2 and Chapter 4 for a detailed discussion on this). These limitations must be carefully accounted for when designing and training a segmentation algorithm. In fact, semantic segmentation depends on anatomical features visible in the images, e.g. layers' interfaces, while the (subjective) boundaries between the annotated and un-annotated regions of the image depend on the gradual reduction of local information content due to de-focusing and vignetting.

The results of our analysis on inter-annotator agreement, discussed in Chapter 4, will be fundamental when assessing system performance.

5.3 Methods

5.3.1 Pre-processing and Data Augmentation

Sector images were first rotated to a common orientation of layers (horizontal, iris at the bottom) to make the segmentation task insensitive to the sector location along the angle circumference. The deterministic rotation of sector images was possible by associating each sector number (reported in the file name) with the corresponding acquisition angle along the 360° iridocorneal interface. As previously explained, each examined eye provides 16 best-focus sector images, meaning that the rotation angle of each sector is a multiple of 22.5° (360° / 16). They were then resized from 1280 x 960 to 320 x 240 pixels (width, height) through nearest neighbour interpolation and preserving the aspect ratio. Each colour channel was divided by 255 (the maximum possible value stored in an unsigned byte) to normalize inputs to the [0, 1] range.

The augmentation pipeline comprised both geometric and photometric transformations. All transformation parameters were randomly extracted from uniform distributions at each new training epoch. Geometric transformations consisted of: translations along x and y axis (ranges $[0, image\_width/3]$ and $[0, image\_height/3]$ respectively); rotations (range $[-30°, 30°]$); shears along x and y axis (range $[-10°, 10°]$); and magnification (range $[0.8, 1.2]$). Zero-padding was used whenever needed to make the resolution of the transformed images correspond to 320 x 240 pixels. Photometric transformations consisted of: contrast variations (range $[0.8, 1.2]$); Gaussian noise injection (zero mean, standard deviation range $[0, 0.02]$); uniform brightness variation (range $[0.8, 1.2]$); and non-uniform (sinusoidal) brightness variations (mixture of two sinusoids with amplitude, frequency and phase ranges tuned based on a qualitative analysis of augmented images). In particular, the sinusoidal brightness perturbation introduces random vertical shadows in the image that resemble those caused, for example, by eye lashes or blurred gel bubbles and aims to increase the insensitivity

of the network to local brightness variations and improve the continuity of interfaces between layers in the final segmentation map. An example of sinusoidal brightness perturbation is provided in Figure 5.1.

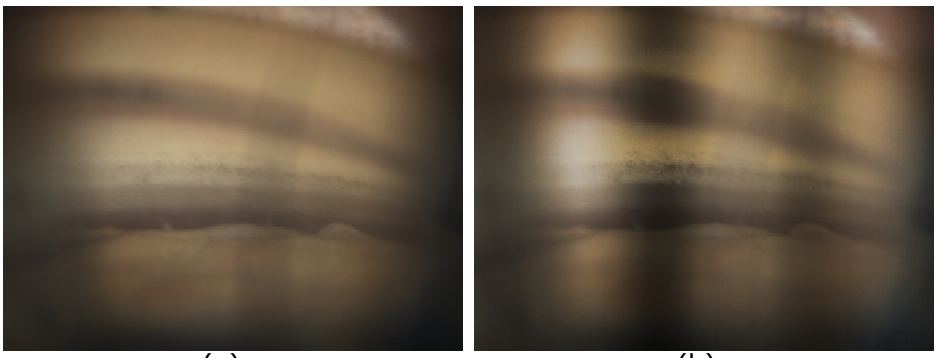

Fig. 5.1 *Example of original sector image (a) and its augmented version using the sinusoidal brightness perturbation (b).*

Except for the sinusoidal brightness perturbation and the Gaussian noise injection that were developed from scratch, the other augmentation techniques and the ranges of their parameters refer to the implementations available in the Pytorch (version 1.7.0) library called *torchvision* (version 0.11.2). Ranges of transformations have been determined empirically based on a trial and error approach and the qualitative evaluations of augmented data.

Both geometric and photometric transformations were meant to simulate the reasonable effects of slight device misalignments during examination and the variability of acquisition conditions (e.g., focus, illumination, possible movements of the patient).

5.3.2 Network Architecture

Figure 5.2 summarizes our approach to provide an accurate segmentation map of anterior chamber angle layers and deal with the characteristics of our ground truth (e.g., vignetting and blurring) effectively.

We remind the reader the acronyms used to identify the different areas in both the annotations and the segmentation maps returned by the algorithm: NA, un-annotated region; IR, iris root; CBB, ciliary body band; SS, scleral spur; TM, trabecular meshwork; C, cornea.

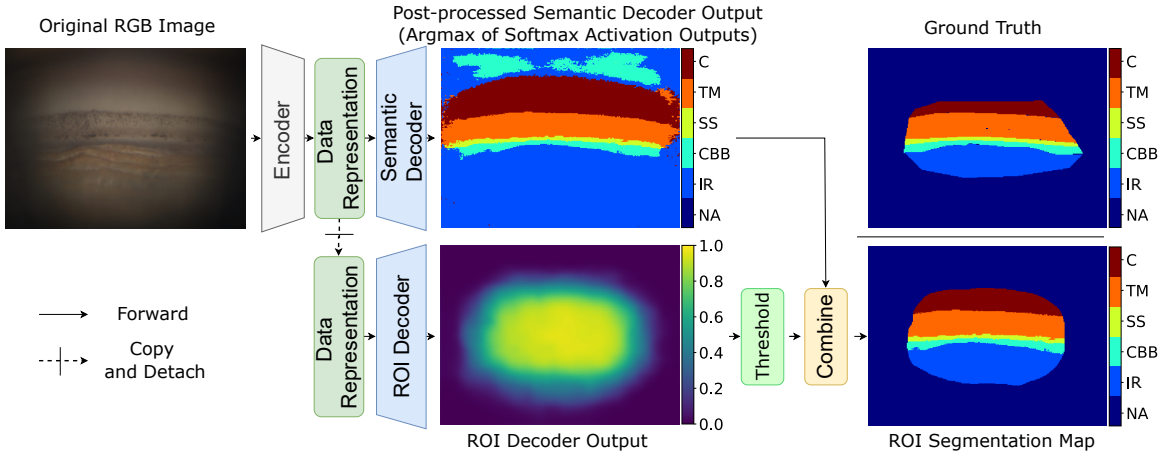

Fig. 5.2 *Overview of network architecture with examples of input, intermediate and final results, also compared with the ground truth.*

The data representation generated by the network encoder is processed in parallel by two independent network units: the semantic decoder, which returns the estimated class for each image pixel (the segmentation map), and the region of interest (ROI) decoder, which highlights an image ROI (the sharp and well-lit area). The two outputs are combined to provide a final segmentation map within the estimated image ROI.

The basic processing blocks are shown in Figure 5.3 (a). The convolutional block (Figure 5.3 (a1)) (ConvB) is a sequence of dropout [119] (0.2 drop probability), 2D convolutional layer (3x3 kernel size, with Kaiming-He uniform initialization [37], zero-padding), instance normalization [130] and Leaky-ReLU activation [74] (0.01 negative slope). Dense blocks (Figure 5.3 (a2)) (DenseB), inspired by [47], are sequences of four ConvBs with two intermediate concatenations of output pairs. The encoder block (Figure 5.3 (a3)) (EncB) is a DenseB followed by an input-output concatenation. The decoder block (Figure 5.3 (a4)) (DecB) is a DenseB followed by a ConvB that reduces the number of feature maps. Figure 5.3 (b) captures the three network components in detail. The encoder (Figure 5.3 (b1)) is composed of two initial ConvBs with 8 filters (dropout in the first ConvB is disabled), followed by three combinations of max pooling and EncB, each doubling the number of feature maps. A final DenseB generates the latent data representation. The semantic decoder (Figure 5.3 (b2)) upsamples feature maps via max un-pooling [82] and concatenates them with those forwarded by the encoder. The resulting feature maps are processed through DecBs, each reducing

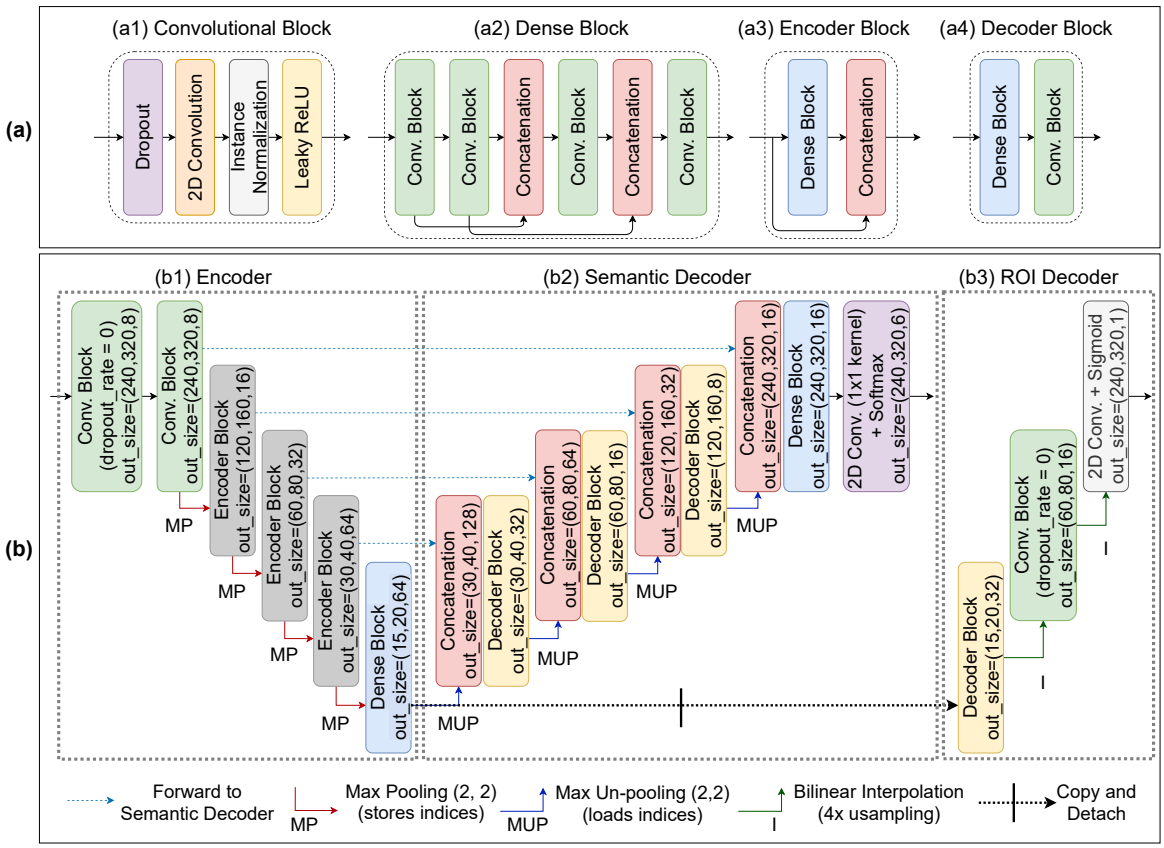

Fig. 5.3 *Basic processing blocks (a): convolutional (a1), dense (a2), encoder (a3) and decoder (a4) blocks; detail of the proposed architecture (b): encoder (b1), semantic decoder (b2) and ROI decoder (b3).*

by a factor 4 the number of feature maps. The processing ends with a DenseB and a 1x1 2D convolution, followed by a six-class softmax activation. The un-annotated (NA) label is considered for consistency but does not contribute to semantic decoder optimization. In fact, the un-annotated image area, does not have any anatomical semantics and only refers to the region that was not possible to annotate with confidence by clinicians.

The ROI decoder (Figure 5.3 (b3)) is our solution to filter out the artefacts *expected* in the segmentation map periphery (since dark and blurred image areas are not informative enough to be correctly classified). A detached copy of the data representation is processed through a DecB, a ConvB (dropout is deactivated) and a 2D convolution with a sigmoid activation (more details on this in the following section). The intermediate feature maps are up-sampled by a factor 4 using bilinear interpolation.

After ROI map binarization (threshold 0.5 in our results), the ROI map and the semantic map are multiplied. Semantic segmentation of anatomical layers of the drainage angle and ROI localization are two un-correlated tasks that rely on different interpretations of the same image. The former looks for patterns and textures that characterize the different anatomical layers; the latter evaluates sharpness and illumination variations across the frame. We verified experimentally that other systems, like attention mechanisms [83], are not effective at addressing this problem since they do not deal with the two tasks independently.

5.3.3 Network Training

The sequence of encoder and semantic decoder can be interpreted as a segmentation U-Net [97] trained end-to-end via weighted categorical cross-entropy with equal weights for all the anatomical layers considered and weight 0 for the region of the image not annotated by the experts. In our first implementation of this model [88], class weight proportional to the average target sizes were used to compensate for the overall imbalance in the average size of the different anatomical layers considered. However, this approach led to worse performance in terms of poorer precision (many false positives) for the thinner layers of the eye region, i.e. the scleral spur and the ciliary body band. This can be explained considering that the scleral spur and the ciliary body band are the layers of the angle that returned the largest inter-annotator variability (Chapter 4). Assigning larger loss weights to these classes is likely to lead to overfitting and poor generalization.

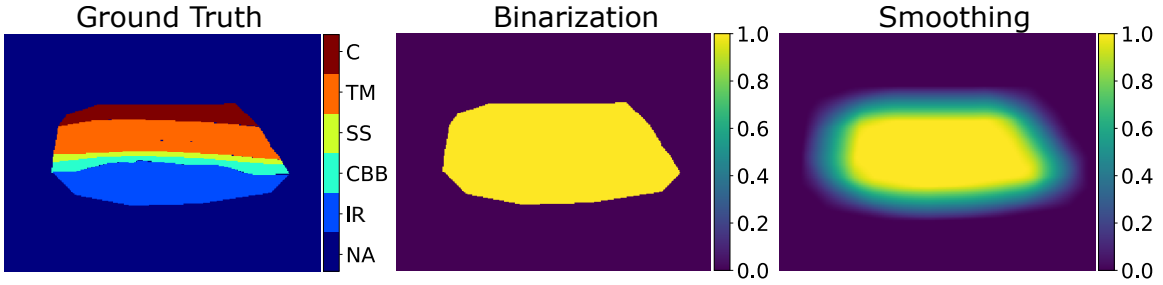

Fig. 5.4 *Example of ROI likelihood map generation. The semantic ground truth (left) is binarized first (annotated region = 1; un-annotated region = 0) and gaps between adjacent layers are filled-in (centre). The binarized image is then smoothed to simulate a distribution of clinician's confidence when annotating the image and obtain the final ROI likelihood map (right) that will be used to train the ROI Decoder.*

To train the ROI decoder we first generated *ROI likelihood maps* from the original semantic annotations: all annotated pixels were first assigned value 1 (binarization), gaps between layers (if any) were filled using the closure morphological operator to obtain a dense region that was then smoothed to simulate a probabilistic distribution of annotator's evaluation confidence (Figure 5.4). Smoothing was performed using an averaging filter with kernel size (*input_width / 6*, *input_height / 6*). ROI likelihood maps were used as reference when computing the mean squared error (MSE) loss for the ROI decoder, independently from the semantic segmentation optimization. The estimation of the most informative image ROI has been treated as a regression, rather than a classification, problem. The ROIs delineated by experts are, in fact, just a subjective representation of slowly varying features of the images, i.e., brightness and focus, that do not have sharp boundaries. Despite the ROI decoder being optimized through MSE error loss minimization we opted to include a final sigmoid activation to constrain outputs. We also verified experimentally that, in our case study, a final sigmoid activation makes the ROI decoder less sensitive to local image features and provides more homogeneous regions of interest than a linear activation. The loss affects only the update of ROI decoder weights and not of the network encoder. This is done to ensure that the encoder can focus on the more complex features that discriminate layers.

Optimization of model's weights was performed through stochastic gradient descent with Nesterov momentum equal to 0.9 (as per PyTorch implementation), learning rate equal to 0.01 and 8 images per batch. Model development and training was carried out in Python (version 3.7.9) and PyTorch (version 1.7.0).

5.3.4 Epistemic Uncertainty Estimation

We included dropout layers in our model and used Monte Carlo dropout [30] to generate multiple softmax activations for an input image at inference time. This approach allows to assess how small variations in the network structure affect the segmentations, thus suggesting whether local output values are the result of the generalization of useful features over training or just the specialization of specific nodes (likely due to overfitting).

The predicted class for each pixel is the argmax of the average activations, and the activations' variance for the assigned class estimates the model (epistemic) uncertainty. If the pixel activations for a given final class are consistent across several output candidates, the variance is low; otherwise, the variance is high (high uncertainty).

5.4 Results

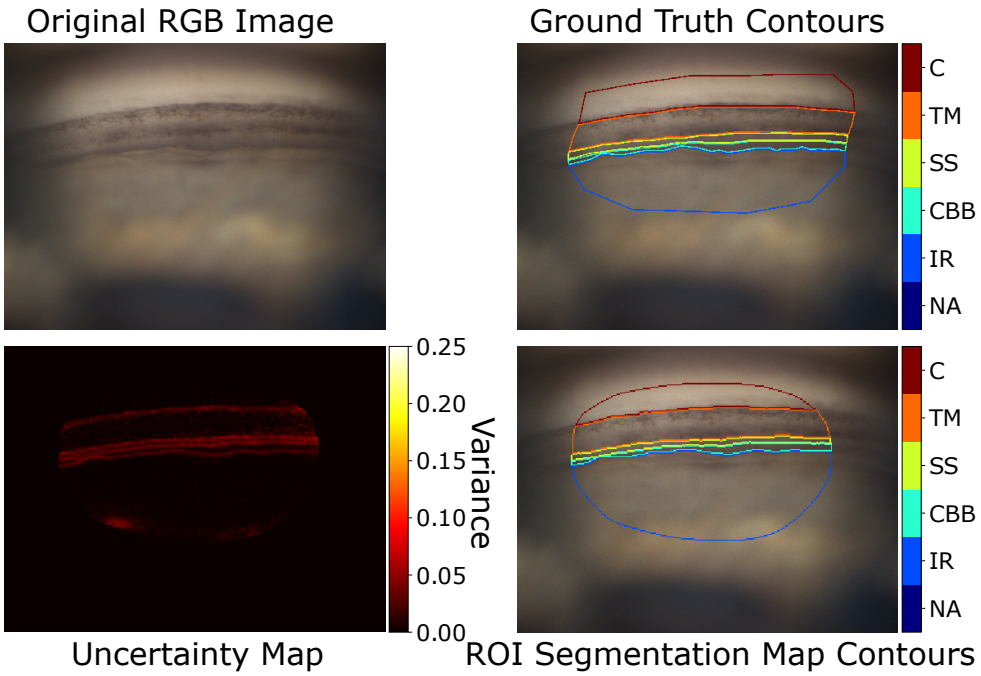

Fig. 5.5 *Example of gonio-photograph (top left) and ground truth delineations of the visible layers (top right); boundaries of the segmentation map output by the semantic decoder and refined by the ROI (bottom right); uncertainty (variance) map (bottom left). Results obtained using 7 predictions through Monte Carlo dropout.*

Figure 5.5 compares an example of combined network output (edges of the segmentation map refined by the ROI, bottom right) with the ground truth delineation (top right). The segmentation is noticeably accurate and the ROI very similar to that highlighted by the annotator. The uncertainty (variance) map provides useful information about the model confidence in the results. In this case, the variance map only highlights layers interfaces, as expected even when the segmentation is accurate, since layers boundaries are often not very sharp features.

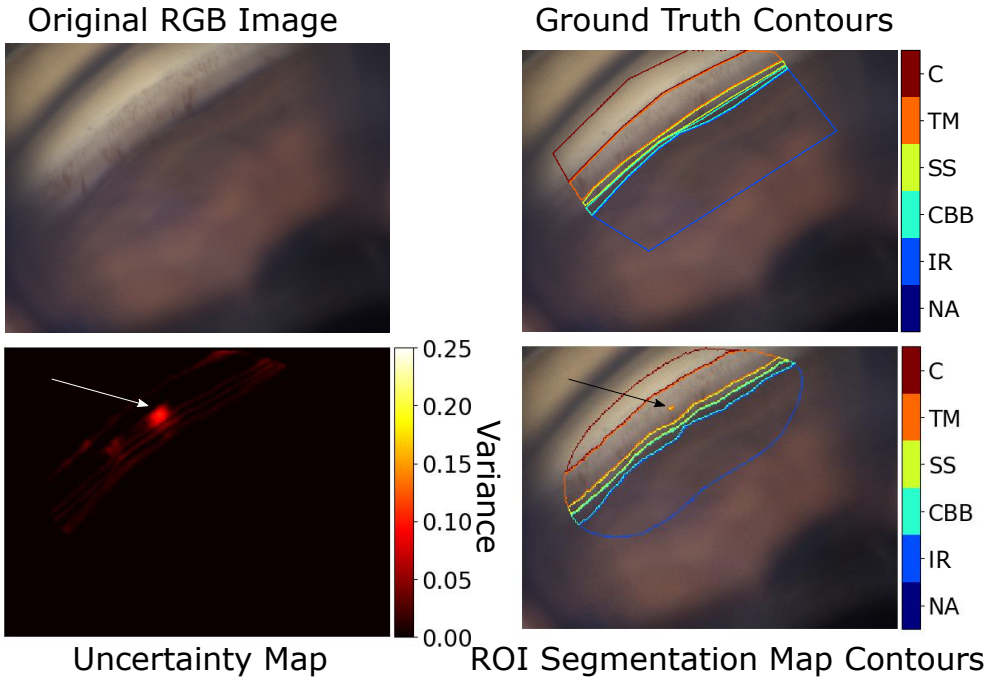

Fig. 5.6 *Example of gonio-photograph (top left) and ground truth delineations of the visible layers (top right); boundaries of the segmentation map output by the semantic decoder and refined by the ROI (bottom right); uncertainty (variance) map (bottom left). The arrows indicate a small misclassified area (bottom right) and the corresponding local uncertainty (bottom left). Results obtained using 30 predictions through Monte Carlo dropout.*

Figure 5.6 shows another example of overall good segmentation performed by our deep learning model. In this example it is possible to notice a small area within the trabecular meshwork that has been mislabelled as scleral spur (indicated by the black arrow in the bottom right image). The corresponding area of the variance map (bottom left) informs the user of the uncertainty associated with that segmented region, suggesting that an expert re-evaluation is advisable.

From both previous examples it is possible to appreciate the visual characterisation of target layers. Their boundaries are not sharp edges but usually just low-contrast transitions or changes in the density of pigmentation.

The performance of the segmentation model was evaluated. First, we split the dataset into a test set (31 images from 25 exams of 17 patients) and a training-validation set (243 images from 189 exams of 145 patients), with similar distributions of the features described in Section 5.2. We then configured the training pipeline to randomly split the training-validation

set into 5 folds of 29 patients each, to cross-validate the model, each fold consisting of a variable number of images (mean: 48.6, std: 5.7). Importantly, according to our methods the training, the validation and the test sets are always granted to contain data from non-overlapping groups of patients.

We considered the segmentation accuracy *within the annotated images area* over training and implemented a policy for storing model weights after each epoch returning an increased validation accuracy. Accuracy only accounted for pixels within the annotated image regions, since segmentation performance on the un-annotated area is, by definition, not measurable. We noticed that using maximum accuracy rather than minimum loss as early stopping criterium caused the training to stop earlier on average and, thus, made the model less sensitive to noise in the data (i.e., the saved parameters are less affected by overfitting on the training set).

Table 5.2 reports average (fold-wise) per-layer performance of the segmentation model computed on the hold-out test set. Precision, sensitivity and Dice scores were the metrics considered. It is worth noticing that the average metrics resemble the values of inter-annotator agreement presented in Chapter 4 highlighting that model performance is likely affected by the degree of variability in specific layer annotations in the dataset, as expected.

The average (fold-wise) per-pixel segmentation accuracy within the annotated image region was about 91%. ROIs estimated by the model and those identified by the annotators cannot be compared quantitatively since their delineation depends on slow-varying features (brightness and focus) and, thus, subjective. To qualitatively validate the ROI decoder of our model we asked a clinician, blinded to our ground truth data, to verify that ROIs estimated by the model did not leave out any anatomical feature of clinical interest present in the original image. The result was that our model was judged capable to highlight an appropriate ROI in every test image and in every cross-validation fold even when the angle interface was not centred in the frame, suggesting a stable, reliable approach.

Model calibration was verified on the hold-out test set after every cross-validation fold by computing the expected calibration errors (ECE) [80]. An average (across folds) ECE of

Table 5.2 *Layer-wise segmentation model performance. Mean precision, sensitivity and Dice score values and standard deviations obtained from the comparison between network's semantic decoder outputs (argmax of softmax actvation outputs) and experts' annotations (un-annotated regions do not affect the results) over a 5-fold cross validation experiment.*

		Mean (%)	**Std (%)**
I	**Prec.**	92.8	1.3
	Sens.	97.4	0.8
	Dice	94.8	0.4
CBB	**Prec.**	84.8	1.0
	Sens.	64.0	4.3
	Dice	72.8	2.9
SS	**Prec.**	68.4	1.7
	Sens.	69.6	3.3
	Dice	68.8	1.2
TM	**Prec.**	84.6	2.0
	Sens.	91.2	1.2
	Dice	87.6	0.5
C	**Prec.**	96.4	0.5
	Sens.	92.2	1.2
	Dice	94.2	0.4

0.01 was obtained, suggesting that our model is well calibrated and that activation variance may be used to estimate epistemic uncertainty.

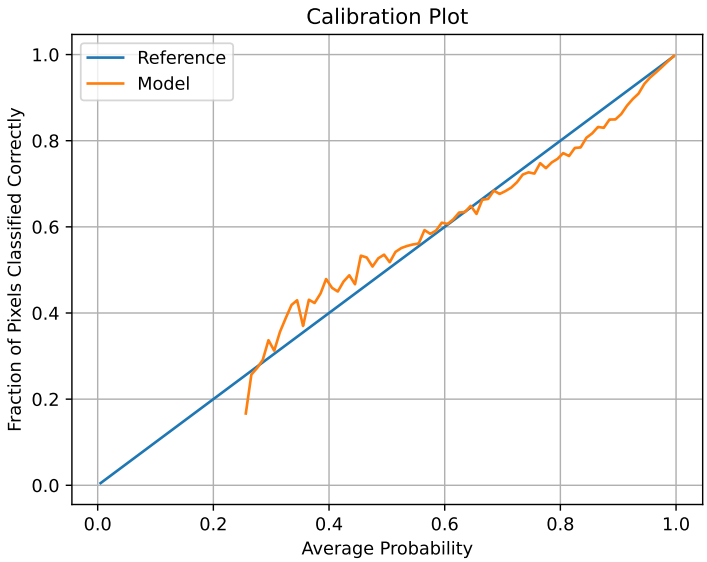

Fig. 5.7 *Example of calibration plot; the fraction of pixels classified correctly is plotted against the average value of final network activations within equally ranged intervals (100 bins in this example).*

Figure 5.7 shows an example of calibration plot obtained during the cross-validation of our model. Te calibration plot starts at an average probability value of about 0.25, this is expected since the minimum activation value necessary for classifying a pixel in a 5-class (the un-annotated class is never predicted by the model) segmentation task must be > 0.2.

Uncertainty estimation highlighted local segmentation artefacts correctly in our tests set.

5.5 Discussion and Conclusions

The morphology of the anterior chamber angle layers is clinically relevant as related to a high-prevalence disease (glaucoma), therefore automatic systems for its analysis will be increasingly important. Existing papers on deep learning applications for the automatic analysis of gonio-photographs consider only the direct angle aperture classification task (e.g., [14]) assigning a single state to each image (a label that globally grades a, possibly, large angle region) that may obscure relevant local conditions. Moreover, there are clinical

needs (according to ophthalmologists) that require the localization and delineation of angle structures and not just their visibility, e.g., precisely highlighting the trabecular meshwork as augmented visualization mode during the insertion of draining implants. In order to assess the angle morphology accurately and account for local variations of layers' interfaces, a segmentation approach is certainly more suitable.

The theoretical approach implemented by most state-of-the art semantic segmentation networks (comprising attention models [83]) is not suitable for dealing with limitations posed currently by the annotation of our gonio-photographic images, i.e., the partial annotation of target layers due to vignette and blur, without modifications. In fact, the simultaneous but independent analysis of both anatomical (e.g., layer boundaries) and qualitative (e.g., brightness and blur) features variation across a single frame is not required in many real-life applications. In particular, our un-annotated class cannot be treated as many segmentation network do with the background class, as its characterization is profoundly different.

The solution presented in this thesis overcomes the limitations we found by adaptively identifying a ROI to refine segmentation maps and improve results readability. Moreover, the calibrated model can support data analysis and interpretation further by highlighting uncertain segmentation areas. This is done by estimating pixel-wise epistemic uncertainty as the activation variance of the final predicted class over multiple segmentation candidates. The overall segmentation accuracy (91%) is promising, albeit within a limited dataset. The layer-wise Dice scores on the test set correlate well with those resulted from the inter-annotator variability study in Chapter 4.

The current limitations of this research are as follows. Firstly, the limited amount of data available must be acknowledged. Although the dataset is representative for a variety of anterior chamber angle features, larger datasets of annotated images are needed to train and evaluate comprehensively deep learning systems for gonio-photographs analysis in clinical practice. The relative novelty and still limited use of digital devices for gonioscopy currently limits the availability of data for rarer conditions. In particular, more images representing complex layers morphologies, e.g. synechiae, shall be collected and annotated in the future to further improve the generalization capabilities of our models.

Secondly, our algorithm does not currently consider any prior knowledge on angle topology (e.g., the fixed order of layers) that could potentially improve the quality of segmentation outputs. Some work on this topic has been conducted using OCT retinal scans (e.g., [40, 41]) and shall be investigated in the future.

The segmentation model presented here fills a gap in current applications of deep learning for ophthalmology, ideally enabling a much more informative analysis of digital gonio-photographs than classification algorithms for angle closure and possibly providing a processing backbone for the measurements of clinical parameters, for evaluating changes of layers morphology over time and for developing auto-tracking systems for device alignment in non-contact clinical settings.

Despite being theoretically very powerful and capable of supporting clinicians in a wide variety of analysis and measurements, our semantic segmentation system currently suffers from the lack of reliable annotated data to be trained on. For this reason, we decided to investigate the problem of direct angle aperture classification similarly to the limited existing literature, as a stand-alone task independent from the segmentation of layers. This is the topic of Chapter 6. It is an important purpose of our future work to develop a reliable angle aperture classifier based on or capable of interacting with the segmentation network. Encouraging steps have already been performed to improve the image annotation process. A proprietary annotation tool has been developed by NIDEK Technologies Srl to specifically delineate layer interfaces in gonio-photographs, although it could be adopted for many more practical applications [118]. This new annotation tool will likely make the delineation of anatomical layers in digital gonio-photographs much easier and effective, and allow to collect a much larger dataset to train our segmentation models.

Chapter 6

Angle Aperture Classification

6.1 About this Chapter

Angle aperture and the importance of its estimation in the screening and categorization of glaucoma have been presented in Chapter 1. We remind that a narrow angle is a condition which increases the chances of developing angle-closure glaucoma. The assessment of angle aperture is, thus, fundamental in clinical practice to evaluate risks and to choose what procedures are most suitable to prevent or treat the disease. The characterization of angle closure, for instance, its morphology and extent, must be taken into account when assessing gonioscopic exams as the preferred treatment of extended appositional closures may differ from that of local synechial closures. As the evaluation of a large volume of digital gonio-photographs requires a long time, machine learning systems could support data analysis, enabling the screening of a larger number of patients, more effectively.

We found only few previous studies on machine learning systems based on support vector machines (SVM) for angle aperture classification in gonio-photographs (Cheng *et al.* [13, 12]) and, as reported in Chapter 2, just one very recent work aiming to address this task using a deep learning model (Chiang *et al.* [14]). This chapter will focus on the deep learning system as it exceeds the performance of the SVM-based ones in this task and is more relevant for discussing our research.

The system described in [14] is trained to detect angle closure (defined as the absence of the pigmented trabecular meshwork in at least half of the angle interface) on gonio-photographs taken with an EyeCam device (Clarity Medical Systems, Pleasanton, California, USA) and representing about 90°-wide quadrants of the drainage angle.

This approach may not be sufficiently sensitive to local features of interest that may be only a few degrees wide (e.g., synechiae), and provides only an approximate estimate of the patient's condition. Thus, the exploration of possible solutions for a more detailed evaluation of angle aperture, capable of detecting, among others, small synechial closures, has been deemed of both research and commercial interest. Moreover, in the conclusions of their work, Chiang *et al.* recommend to investigate systems for the estimation of angle aperture in digital gonio-photographs acquired with faster and less operator-dependant devices such as the NIDEK GS-1.

According to clinical assessment, the measure of angle aperture can be approximated by estimating the apparent iris insertion, which, in turn, often relies on the visibility of angle layers. Several aperture grading scales include the visibility of layers or the apparent iris insertion in their classification criteria (e.g., Scheie's [106] and Spaeth's [117] scales).

From a computational point of view the problem of assessing the visibility of anterior chamber angle layers (and, thus, angle aperture) may be addressed through deep learning approaches at least in two main ways:

- *semantic segmentation*: this approach enables an arbitrarily dense (even column-wise) estimation of angle aperture based on the segmentation map. It is however extremely costly both in terms of ground truth generation (i.e., delineating the contour of each layer in each image) and processing resources (especially memory, since the segmentation network requires intermediate feature maps to be stored and shared between the encoder and the decoder);

- *classification*: this approach allows to estimate the angle aperture by processing patches of gonio-photographs. The size of these patches may be selected to optimize the trade-off between clinical needs, effort required to obtain ground truths and computational

power for training and inference phases. Here, the ground truth consists of labels (e.g., *Open*, *Occludable*, *Closed* and *Unknown* in our study).

The semantic segmentation algorithm presented in Chapter 5 showed promising overall performance. It could be used as a pre-processing step to generate segmentation maps of angle sections, which could then be classified according to the layers visible. Its application to the angle aperture task is discussed in Chapter 7 as part of our future work. However, the segmentation network is computationally expensive in terms of memory required (which may be an issue in case of deployment on the acquisition device) and the generation of new ground truth annotations to consolidate its performance requires resources and time. Therefore, solutions requiring less computational resources and/or reducing the need for large sets of dense ground truth have been considered preferential, from a translational perspective, to ensure scalability.

This chapter aims to describe the pilot experiments that have been carried out and the results obtained so far for the classification of angle aperture in digital gonio-photographs acquired with the NIDEK GS-1 device, without relying on the semantic segmentation algorithm. We compare our approach with the previously published work [14] highlighting clinical advantages and discussing limitations and suggestions for future work.

6.2 Materials

Data

110 digital gonioscopic exams (16 sectors each) acquired with a NIDEK GS-1 device were selected from the available pool of data shared by the clinical sites in Vienna (Austria), Los Angeles (US), Lisbon (Portugal) and Genova (Italy).

The selection was carried out by the author of this thesis by visually inspecting the available exams to identify those showing pathological morphologies (in terms of angle closure) of the iridocorneal angle. The aim was to optimize the distribution of instances showing different degrees of angle closure. The variability of the iris colour was taken into

account; 68% of data were exams of dark irises (i.e. brown eyes). The variability of the trabecular meshwork pigmentation was not quantified.

The cardinality of the dataset to annotate was mainly limited by the actual availability of exams showing, to some extent, *Occludable* and *Closed* angle sections. Including additional *Open* (healthy) exams was possible but would just likely worsened the issues related to using an already highly un-balanced dataset in terms of angle grades distribution (see next paragraph for more details).

Exams were acquired with patients' agreement and following General Data Protection Regulation rules (including anonymisation at source) during routine clinical examinations.

Annotations

Annotations were performed using an annotation tool developed by NIDEK Technologies Srl. according to the annotation protocol described in Section 3.4.2. Two experienced ophthalmologists performed the annotations. Two disjoint sub-sets, consisting of 50 exams each, were randomly extracted from the whole dataset and assigned to the experts. The remaining 10 exams were labelled by both the experts with the purpose of assessing inter-annotator variability, similarly to what was done for the semantic segmentation task. However we did not have the time to carry out this study, but it remains among the purposes of our future research. We decided to choose as ground truth of the 10 common exams the annotations provided by the ophthalmologists with more years of experience.

Despite the effort to make the dataset as balanced as possible with respect to the pre-defined angle aperture classes, that are *Open*, *Occludable* and *Closed* (see Section 3.4.2 for definitions), only about 12.6% and 15.3% of all the sub-sectors were representative for the *Occludable* and the *Closed* classes respectively. Also, about 7.4% of all the sub-sectors were labelled as *Unknown* by the annotators, meaning that the image quality was not sufficient for grading or the expert was not confident enough.

6.3 Methods

6.3.1 Pre-processing and Data Augmentation

Differently from the strategy adopted when addressing the semantic segmentation task, which consisted in choosing the best-focus image (i.e., the image of the acquired focus stack with the focus plane centred on the outer boundary of the iris) of each angle sector considered, for this classification task we chose to use the artificially reconstructed all-in-focus images for both training and evaluation. We remind that all-in-focus images are obtained from the whole acquisition stack by merging the most informative areas of each focus plane using an algorithm included in the NIDEK Navis-Ex software. The advantage of all-in-focus images is that layer-layer interfaces are more likely sharper than in best-focus images and may, ideally, better guide the aperture grading system. Conversely, they may be affected by pixel-level artefacts due to the reconstruction algorithm and could decrease the quality of boundary delineations returned by the segmentation task.

1280 x 960 pixels (width, height) RGB images depicting sectors of the anterior chamber angle were rotated so that the iris was at the bottom of the frame, and then re-sized to 640 x 480 pixels. Both rotation and re-sizing used bicubic interpolation to reduce negative effects on image quality. The rotation filled the empty corners of the resulting image with zeros. Images were then cropped centrally to obtain 480 x 480 pixels pictures, equivalent (a part from the scale) to the regions highlighted to be labelled by annotators (see Section 3.4.2 for detailed information on the annotation process). The coordinates of the trabecular meshwork (or those of the direct iridocorneal interface if the meshwork was not visible) in each original sector image were estimated by the device during acquisition [10]. They were extracted from the exam metadata and geometrically transformed to account for the pre-processing already performed on the original images in order to locate the interface correctly.

Transformed trabecular meshwork coordinates were used to select a 480 x 160 pixels region-of-interest (ROI) from each image, vertically centred on the trabecular meshwork as shown in Figure 6.1 (if the trabecular meshwork coordinates are too close to the upper or lower bound of the image, the ROI selected is the upper or lower 480 x 160 pixels area of the

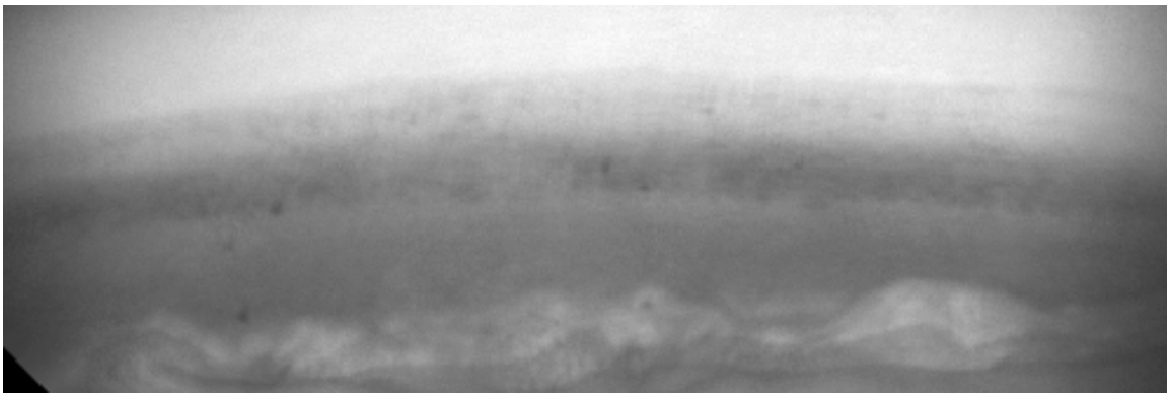

Fig. 6.1 *Example of greyscale 480 x 160 (width, height) ROI extracted using the trabecular meshwork coordinates.*

image respectively). Each ROI was divided into three 160 x 160 sub-sectors, corresponding to the three angle interface regions of each sector image independently annotated by the experts (see Section 3.4.2) and were then paired with the corresponding aperture class.

At the beginning of each training epoch, training images (i.e., 160 x 160 pixel sub-sector images) were independently augmented by applying geometric and photometric transformations, as follows:

1. *affine transformation*: comprising random rotation (uniform distribution, range $[-30°, 30°]$); random translations along x and y axes (uniform distributions, ranges $[-width/6, width/6]$ and $[-height/6, height/6]$ respectively); random shears (uniform distribution, range $[-10°, 10°]$, random scale (uniform distribution, range $[0.8, 1.2]$); and random horizontal flip (Bernoulli distribution, $p = 0.5$). Affine transformation is currently performed using bilinear interpolation which has been considered an acceptable trade-off between quality of results and speed of computation;

2. *sharpening/blurring*: sharpness enhancement / blur by a factor randomly selected from a uniform distribution, range $[0.8, 1.2]$;

3. *brightness variation*: global brightness increase / reduction by a factor randomly selected from a uniform distribution, range $[0.8, 1.2]$;

4. *Gaussian noise injection*: additive Gaussian noise with 0 mean and standard deviation randomly selected from a uniform distribution with, $[0, 3]$.

Except for the Gaussian noise injection which was inherited from the semantic segmentation algorithm pre-processing pipeline, augmentation techniques and the ranges of their parameters refer to the implementations available in the Pytorch (version 1.7.0) library called *torchvision*. Ranges of transformations have been determined empirically.

At the beginning of each new epoch, every training image was augmented with a probability equal to 0.9 so that the original version of each sub-sector image was available to the network on average once every ten epochs. Given the high correlation between adjacent sub-sectors of the angle, the actual variability of features among training samples (e.g., iris colour, trabecular meshwork pigmentation) from the same exam was often very low. Data augmentation was used to alleviate over-fitting and regularize the training process. Test images were never augmented, but both training and test images were normalized to the [0, 1] range before being fed to the convolutional neural network, by dividing each 8-bit colour channel by 255, its maximum representable value.

6.3.2 Network Architecture

The model we used is a custom convolutional neural network with ten convolutional layers and three fully connected layers. The basic processing blocks as well as the overall network architecture are depicted in Figure 6.2.

All the convolutional layers consist of multiple 3 x 3 2D convolutional kernels with 1 x 1 zero-padding. They are initialized using the Kaiming-He uniform distribution [37] and are followed, in this order, by a layer normalization [6], a leaky-ReLU activation [74] with negative slope equal to 0.01, and a 2D dropout [127] with drop probability equal to 0.2. After two convolutions a max-pooling layer halves the width and height of feature maps. The unit comprising two sequences of convolutional layer, layer normalization, activation and 2D dropout, and followed by a max-pooling layer will be called, from now on, *convolutional block* (Figure 6.2 (a)). The first convolutional layer in a convolutional block doubles the input number of feature maps, except for the one receiving the network's input that always returns 8 feature maps, despite the number of channels of the input image (RGB or greyscale).

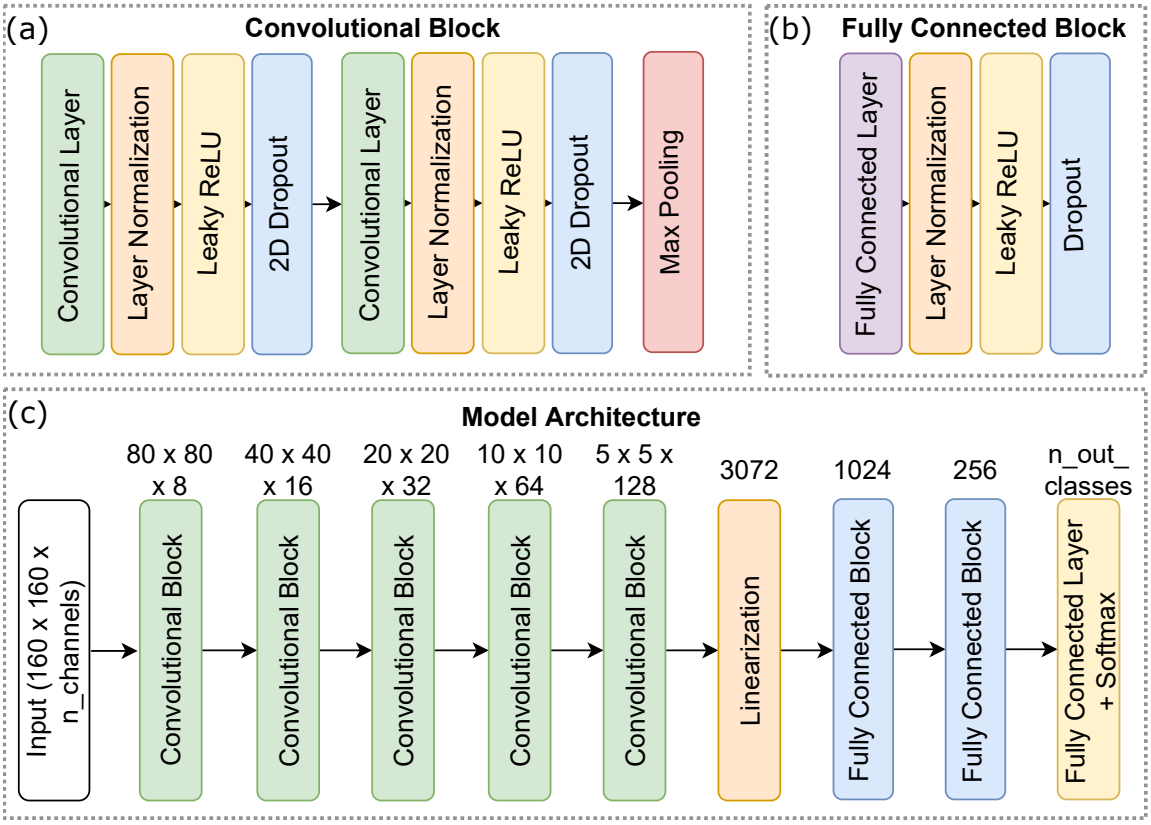

Fig. 6.2 *Basic constituent blocks and overall convoutional neural network architecture.*

The first two fully connected layers are followed by layer normalization, leaky-ReLU activation and dropout [119] (0.2 drop probability). Each sequence of fully connected layer, layer normalization and leaky-ReLU will be called, from now on, *fully connected block* (Figure 6.2 (b)). The first fully connected block receives as input the vector obtained after linearizing the 128, 5 x 5 feature maps and outputs 1024 nodes. The second fully connected block reduces the number of nodes from 1024 to 256. The last fully connected layer is followed by a softmax activation that returns normalized scores that associate each input angle patch with the considered set of output classes. The number of input channels and output classes is not specified in the scheme of the overall model architecture (Figure 6.2 (c)), as it may vary with the experiment (as described in Section 6.3.4).

The depth of the network (number of blocks in sequence) as well as its width (the number of convolutional kernels in each processing block) were tuned over an optimization experiment aiming at assessing how variations in the overall capacity of the model would

affect its performance in our classification task. We started with a very simple network made up of two convolutional blocks and one fully connected block (both defined above) and gradually included additional modules until the performance saturated, thus obtaining the proposed architecture.

We acknowledge that more sophisticated architectures, e.g., attention mechanisms [83], could lead to further improvements. Moreover, network pre-training could allow to increase the complexity of the classifier without incurring in over-fitting. Additional investigation on alternative network implementations is recommended in the future.

6.3.3 Network Training

Training consisted of 500 epochs using batches of 128 sub-sector images. We verified experimentally that increasing the number of training epochs did not lead to any performance improvements. The stochastic gradient descent optimizer had a learning rate of 0.01 and Nesterov momentum equal to 0.9 (according to PyTorch implementation). The loss function was a weighted categorical cross-entropy with all weights equal to 1 (except when otherwise stated in Section 6.4) but that of the *Unknown* class, which is always 0. This means that sub-sectors that were deemed unclassifiable by experts do not affect the learning process. They are however kept in the dataset to ensure consistency in the number of input sub-sectors for each exam.

Training was always performed from scratch, with weights initialized randomly according to the selected distributions (e.g., Kaiming-He uniform for convolutional kernels).

6.3.4 Evaluation Set-up and Metrics

Several possible frameworks for angle aperture classification have been considered and compared to evaluate the performance of the classification model. In particular, the following variations have been tested:

- comparison between RGB and greyscale inputs for the three-class classification problem (*Open*, *Occludable* and *Closed*);

- comparison between different aggregations of output classes. In particular, by merging classes *Open and Occludable*, and *Occludable and Closed*;
- comparison between class weights in the loss function that either ignores or account for the unbalance of the dataset in the case of *Open* vs *Occludable + Closed* class aggregation.

5-fold cross-validation experiments have been carried out in each case. Training and test sets were split by patient to avoid data correlation. Average (computed over each cross-validation experiment) per-class sensitivity and precision (with standard deviations) have been used in the three-grade classifier case, and average test and training losses (with standard deviations) have been plotted to analyse the training process. Average receiver operating characteristics (ROC) curve and area under the ROC curve (AUC) have been used in the two-grade classification set-up.

ROC curves were obtained by calculating sensitivity and specificity values respectively for the *Closed* or the aggregation between the *Occludable and Closed* classes at 100 classification thresholds, i.e., the minimum value required to classify an input as appertaining to the class, equally spaced between 0 and 1.

6.4 Results

Greyscale vs RGB Input Data

The first comparison involved training using either greyscale (one channel) or RGB (three channels) input data. All processing and network hyper-parameters were kept un-altered except for the number of input channels.

The reason for comparing the greyscale and RGB inputs is to verify whether the information on colours may help the grading of angle aperture.

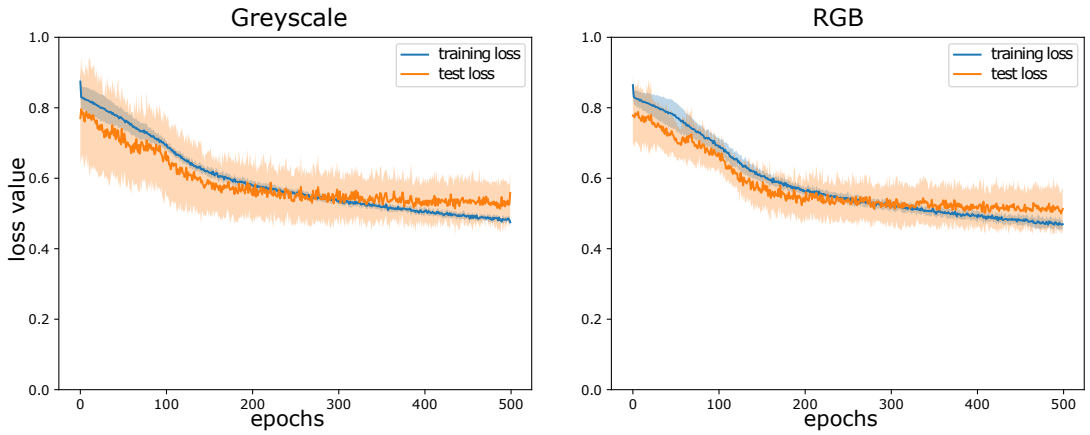

Fig. 6.3 *Average training and loss functions (solid lines) and their standard deviations (shaded areas) computed over the cross-validation experiments on greyscale and RGB input data.*

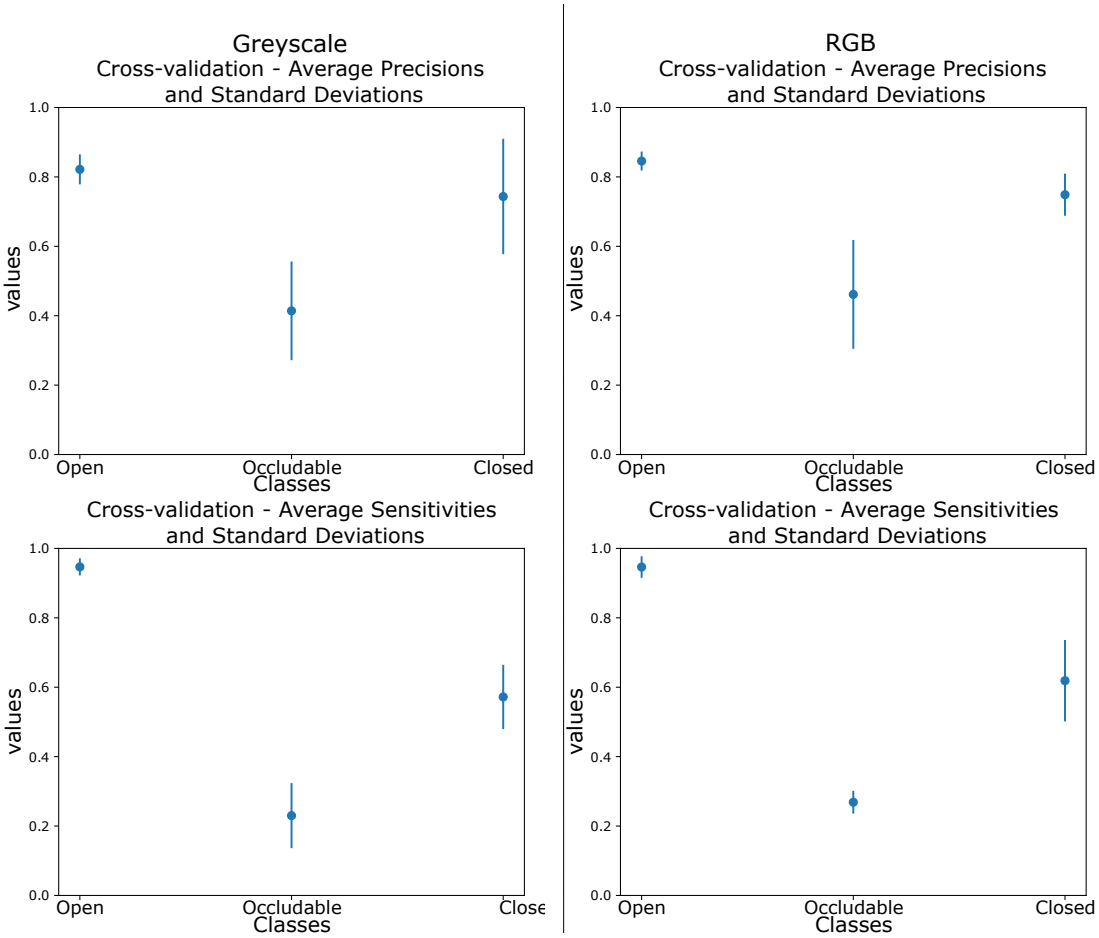

Fig. 6.4 *Per-class precisions and sensitivities (means and standard deviations over the cross-validation experiments) for training using greyscale and RGB input data.*

Figure 6.3 shows the average training and test losses (solid lines) and the corresponding standard deviations (shaded areas) computed over the cross-validation experiments in both cases. They show very similar trends and, even if a reduced training stability may be observed at the initial stages in the case of greyscale inputs (larger standard deviation). It is possible to notice that the test loss reached a plateau very soon during the training process.

Figure 6.4 reports per-class precisions and sensitivities (means and standard deviations over the cross-validation experiments) in the two cases. The cross-validation experiment on RGB inputs resulted in only marginally better values at an increased computational cost. Therefore it has been deemed reasonable to run all the following experiments on greyscale images.

Class Aggregation

The second experiment aimed to evaluate classification performance after aggregating the output classes in two ways: *Open* vs *Occludable* + *Closed* and *Open* + *Occludable* vs *Closed* (where the sign + is used to indicate aggregation). In both cases class instances were aggregated before training the model, to obtain a binary classification problem; the *Unknown* class is always an output class but is never predicted by the network since it does not contribute to the loss function. The two-class problem allowed us to use ROC curves and the AUC metric for evaluation of results. Everything else was left unaltered.

Figure 6.5 shows the ROC curves in the two cases. The average AUC value for the cross-validation using the *Open* + *Occludable* aggregation was 0.91, while for the cross-validation using the *Occludable* + *Closed* aggregation it was 0.88. The difference may be due to the higher imbalance in class examples in the first case. Average ROC curves were computed by calculating the mean sensitivity and specificity values at each threshold (a sample point of the ROC). Standard deviations were computed at each threshold value from the corresponding distribution of sensitivities across validation folds.

Clinical considerations must be taken into account when deciding if these aggregations are meaningful. A discussion with clinicians on this matter has not been possible yet because of lack of time, but will be part of our future work.

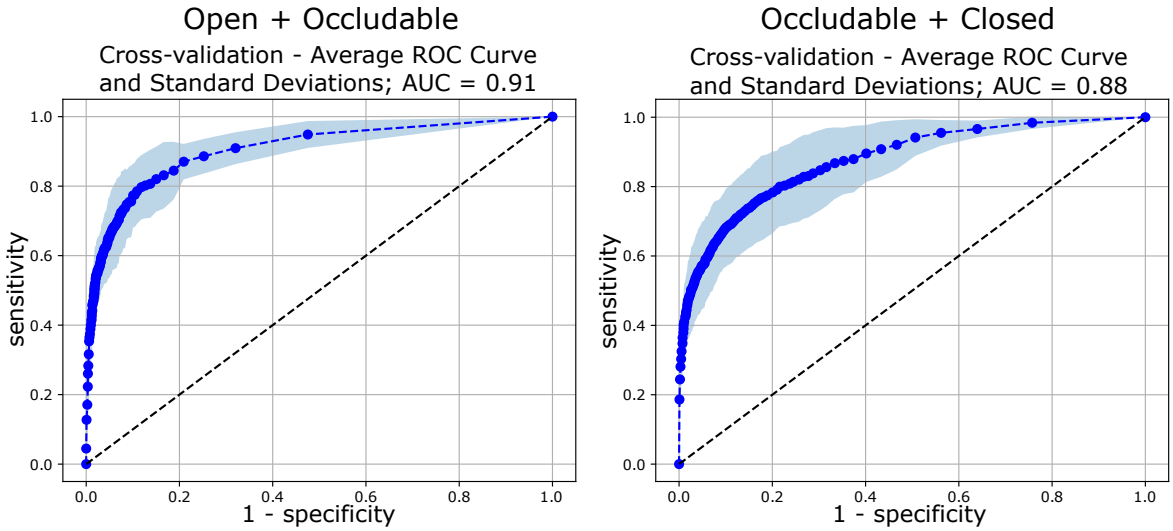

Fig. 6.5 *Average ROC curves (solid lines) and standard deviations (shaded areas) for the Open + Occludable and the Occludable + Closed aggregations. AUC values are reported in the titles.*

Balanced vs Unbalanced Class Weights

In this last experiment we tested how assigning loss weights proportional to the cardinality imbalance between classes affects the ROC curves and AUC values in the case of *Open* vs *Occludable + Closed* class aggregation. We compared the metric values obtained when the class weights in the loss function either ignore or account for class imbalance (the weight for the *Unknown* class is always 0). In the former case the loss weights were (1, 1, 0); in the latter, the weight for the *Occludable + Closed* class equalled the ratio between the number of elements in class *Open* and the number of elements in class *Occludable + Closed*. This ratio was 2.3 and the weights were thus (1, 2.3, 0).

Results are shown in Figure 6.6. The two cross-validation experiments returned comparable AUC (0.88 and 0.89 for the non weighted and weighted cases respectively), but in the weighted case the variability among folds (the standard deviation of sensitivities, depicted as the shaded area) was reduced considerably. This may be motivated by the fact that, when increasing its weight in the loss computation, the sensitivity of the model to the *Occludable + Closed* grade is less affected by the variable imbalance among classes in each cross-validation fold, i.e., the variability of false negatives count decreases.

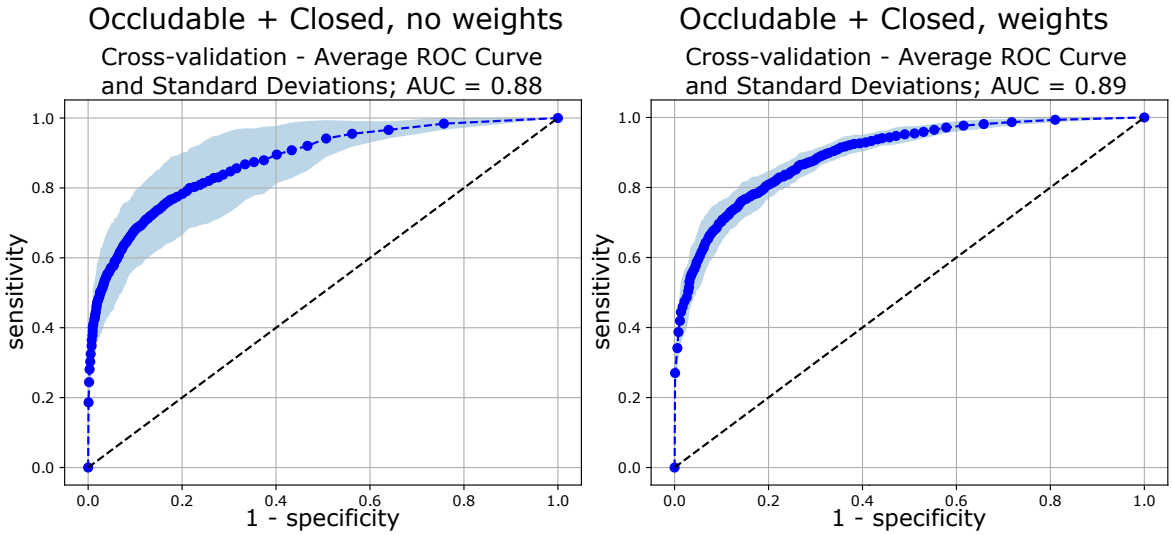

Fig. 6.6 *Average ROC curves (solid lines) and standard deviations (shaded areas) for the non weighted and weighted loss trainings for the Occludable + Closed aggregation. AUC values are reported in the titles.*

6.5 Discussion and Conclusions

We discuss our study comparing data and framework with those reported in the previously mentioned research article by Chiang *et al.* [14] on angle aperture classification in digital gonio-photographs. Briefly, in [14] the authors fine-tuned a ResNet-50 convolutional neural network [38], pre-trained (on ImageNet [19]), to solve the two-class problem concerning the detection of angle closures in digital gonio-photographs acquired with an EyeCam device and representing quadrants (90°-wide sectors) of the anterior chamber angle. Their dataset (training and test) was made up of 33635 quadrant images ($\sim$30300 open and $\sim$3300 closed) from 4152 patients, and angle closure was defined as the absence of the pigmented trabecular meshwork in at least half of the quadrant image. The training-test split was performed at patient level thus ensuring the un-correlation of data between the two sets. The performance reported in the article is very good with AUC higher than 0.96. Our classifier obtained, depending on the experiment, AUC values ranging between 0.88 and 0.91.

The classification of angle aperture in very-wide angle sectors may easily fail to detect local pathological tissue (e.g., synechiae) with negative consequences concerning the diagnosis and the choice of treatment. This limitation motivated our investigation to focus on grading

much narrower angle regions, thus returning an, ideally, more detailed characterization of the iridocorneal interface capable of highlighting the presence of local tissue abnormalities.

Several differences between our study and the one by Chiang *et al.* must be considered to contextualize the findings. Firstly, although our dataset consisted of about 1/6 of their overall number of images (5280 sub-sector images vs 33600 quadrant images), it was generated by only 1/38 of the patients. This means that the intra-patient correlation of features (e.g., iris colour and texture, trabecular meshwork colour and texture, characterization of pathological conditions) is much higher in our dataset than in theirs; we have, in fact, 48 sub-sector images per exam, one exam per patient, while they have 4 quadrant images per exam and 1 or 2 exams per patients (from different eyes).

On the other hand, their dataset, although much larger than ours, is less representative for the variation of visual features linked to ethnicity (as they acknowledge in the Discussion section) since it only comprises patients that have identified themselves as Chinese Americans. Our dataset, instead, accounts for a larger ethnical variability given both the geographical distribution of the clinical sites that acquired the exams and the absence of any ethnicity constraint for patient selection.

Secondly, the field of view of their quadrant images is very different from that of the sub-sector images we extracted from the original GS-1 acquisitions. The higher magnification of our images may cause small textures to affect the classification and make the distribution of input data more complex to learn and generalize. For example, the trabecular meshwork appears in their images as an almost homogeneous brown-ish stripe, while in ours it is a highly heterogeneous region characterized by complex textures (e.g., the pigmentation). However, the approach discussed in our study is suitable for a local classification of angle sections that, in turn, can improve the clinical value of this system.

Thirdly, the availability of a larger volume of less correlated input data, presumably enabled the authors of [14] to use a much larger network ($\sim$23 vs $\sim$4 million parameters) without incurring in over-fitting. It is likely that a deeper network trained on a larger amount of data could lead to better performance, as generally suggested in the deep learning literature. We must highlight that in [14] the network was pre-trained on the ImageNet dataset [19],

while in all our experiments it was trained from scratch. Pre-training our model or using a larger, pre-trained state-of-the-art classification network could reduce the over-fitting issue in our study case as well, and this approach will be tested in our future work.

It is also important to highlight that in [14] the closure of an angle sector was defined as the absence of the pigmented trabecular meshwork in at least half of a quadrant image, but the grades in our study did not distinguish between pigmented and non-pigmented meshwork, making the two approaches not directly comparable. The reason for our choice was that, according to our inter-annotator variability study (see Section 4.5), the difference between pigmented and non-pigmented trabecular meshwork may be unclear, even to the expert ophthalmologist.

It must be acknowledged that, despite the relatively small size of the drainage angle regions we considered and annotated, changes of state (e.g., *Open* to *Occludable*) may still be present within the same sub-sector, so that a single class label is not always an accurate characterization and it may be difficult to evaluate quantitatively whether a misclassification is actually a mistake or just the result of the network focusing on a different image region than that considered by the annotator. For example, a sub-sector image equally representing an *Open* and an *Occludable* areas and annotated as *Open* could be predicted as *Occludable* by the model (and quantitatively counted as a mistake) without actually being entirely wrong. This is a situation quite common in our annotated dataset. The implementation of systems capable of explaining classification outputs (e.g., Grad-Cam [111, 112]) may help us better understand model's behaviour. According to their annotation protocol, the issue of multiple aperture grades coexisting within the same image reasonably exists also in [14], however the authors do not make any mention about it.

The very low precision and sensitivity of the *Occludable* class observed in the three-class experiments (Section 6.4) may be caused by the weak features, computationally speaking, that differentiate it from the *Open* class, making the model often mislabel them, both with false positives and negatives. Moreover, according to our inter-annotator variability study, the scleral spur and the ciliary body band (the visibility of which discriminates between the *Open* and the *Occludable* classes) are those with highest variability among experts. The *Open*

class metrics were just slightly affected by this trend since the cardinality of this set is much larger and most of its instances were correctly classified. In order to verify this theory, it will be advisable to conduct an inter-annotator study using a common dataset to be annotated by multiple experts.

The approach we adopted when building the neural network, consisting in gradually adding processing blocks and consequently evaluating the model, showed that performance saturates quickly (when the classifier structure is still very simple), making any addition to the architecture only worsen over-fitting. This suggests that the size of the dataset (110 exams from 110 different patients), considering also the high correlation of its samples (48 sub-sector images for each exam), may not be sufficiently informative for addressing this task more effectively. Repeating this experiment including both pre-training and more sophisticated networks (like in [14]) is recommended in the future to verify this hypothesis. Moreover, when multiple aperture grades are (about) equally represented in an sub-sector image, a single label may not describe the input properly, possibly leading to noise in the data. However, especially when aggregating classes, performance indicates that further research and effort in collecting and annotating new data may lead to good results.

Even if the training process could be driven by changing the class weights to account for class cardinality differences, it must be acknowledged that further work in collaboration with clinicians is needed to properly evaluate the minimum requirements in terms of discrimination power. A different approach based on the semantic segmentation algorithm presented in Chapter 5 will be considered in future research. Once obtained, segmentation maps could be scanned column by column (given that all images have been pre-processed to show the iris at the bottom of the frame) and the classification of angle sub-sectors might be obtained by directly applying the deterministic rules that are based on layers visibility and by considering the additional segmentation uncertainty to refine and explain results.

Chapter 7

Discussion and Future Work

7.1 About this Chapter

This chapter first reviews and summarises the work conducted for this thesis, then discusses its limitations and possible solutions to be considered in the future.

7.2 Summary of the Thesis

The assessment of the anterior chamber angle is fundamental for the diagnosis, categorization and management of glaucoma, a high prevalence disease that leads to severe visual impairment and even irreversible blindness. One of the current clinical standard examinations, gonioscopy, requires extensive experience and time, thus not being performed as often as necessary, causing misdiagnoses and preventing people in need of treatment from receiving it.

The NIDEK GS-1 device enables easier and faster examinations with the additional advantage of automatically storing digital images of the drainage angle. It can be used by inexperienced ophthalmologists, by optometrists, and by medical photographers in the context of virtual clinics. Although the acquisition phase is fast, images of the angle must be still viewed and evaluated by an expert to produce a diagnosis.

The aim of our research was to investigate possible applications of deep learning systems to the analysis of digital gonio-photographs to support their evaluation. Despite deep learning algorithms have been developed for and deployed to many image processing tasks in medicine (and, more specifically, in ophthalmology), very limited contributions exist about the automated analysis of gonio-photographs. In fact, only very few studies have been published so far and most of them only consider the problem of angle aperture grading in digital gonio-photographs acquired with a different device (EyeCam). All of them except one are based on conventional image processing techniques or support vector machine classifiers.

Our research started by involving clinical experts from several sites located worldwide in a requirements collection process. They were asked to fill in a questionnaire assigning a priority score to several analysis tasks on digital gonio-photographs. The feasibility of the tasks with highest average priority score was assessed to select development aims. The commercial interest in, potentially, translating the technologies developed was also considered during selection.

Two analysis tasks were selected to be studied: the semantic segmentation of drainage angle layers, and the grading of local angle aperture. In both cases, the earliest design phase concerned the formulation of an annotation protocol to generate ground truth data for supervised model training.

Annotations for the semantic segmentation task consisted of manual delineations of anatomical layers visible in a pilot set of digital gonio-photographs that were selected to account, as much as possible, for the distribution of features of the drainage angle. These features comprised both physiological (e.g., iris colour, trabecular meshwork pigmentation) and pathological (e.g., the shape of synechiae) characteristics of the angle. Ground truth was provided by four ophthalmologists who also annotated a set of 20 common images for an inter-annotator variability study. An additional annotator provided ground truth just for the 20 images thus being included only in the variability study.

Annotations for the angle aperture grading task consisted of local aperture labels (*Open*, *Occludable* or *Closed*) assigned to equally wide sub-sectors of digital gonio-photographs (three sub-sectors per image). Ground truth was provided by two experts.

7.2.1 The Inter-annotator Variability Study

The effectiveness of deep learning algorithms largely depends on the quality of the ground truth. In particular, the variability in annotations provided by multiple experts imposes an upper bound to system's performance. This is a phenomenon that too often is not considered or studied when evaluating automatic systems, thus making their performance hardly interpretable correctly. In order to evaluate the semantic segmentation network consistently, we studied the inter-annotator variability of manual delineation of anterior chamber angle layers in our digital gonio-photographs. This study was described in Chapter 4.

We measured variability in three ways: (i) by counting the number of times each expert was confident at detecting layers; (ii) by measuring the variability of consensus (i.e., the area annotated by many clinicians accordingly) when increasing the consensus threshold; and (iii) by computing average per-layer annotators' precision, sensitivity and Dice score when comparing their ground truth with others'.

Our results suggest that even expert ophthalmologists don't feel always confident at detecting drainage angle layers in digital gonio-photographs, especially the scleral spur and the non-pigmented part of the trabecular meshwork, when considering its pigmented and non-pigmented components separately. Consensus decreases considerably, and linearly, for scleral spur, ciliary body band and trabecular meshwork (or its two components when considered separately). Per-layer metric trends confirm that scleral spur and ciliary body band delineations show highest overall variability among annotators.

This is the first time an inter-annotator variability study on manual delineation of layers in digital gonio-photographs has been conducted and published, being potentially a reference for future research not only concerning automatic systems for the analysis of gonio-photographs but also clinical studies.

7.2.2 The Semantic Segmentation Algorithm

The design and development of a semantic segmentation algorithm for anatomical layers of the drainage angle was motivated by the interest of clinicians and the advantages it could have in future applications. This work was presented in Chapter 5.

The rich morphological information provided by semantic segmentations may be, in fact, useful as preprocessing step for a wide variety of measurements (e.g., the identification of pathological tissues or the estimation of trabecular meshwork pigmentation, necessary in case of laser treatments). Segmentation systems are being developed and tested for layer segmentation on OCT scans of the retina for similar purposes, however they can not be used off-the-shelf on digital gonio-photographs because of the specific characteristics of these images (e.g., de-focusing and vignetting) and the effects they have on the ground truth (i.e., partial annotations). The vignette and blur visible in our digital gonio-photographs prevented annotators from delineating the whole extent of layers, thus creating inter-class feature correlation between layer annotations and the un-annotated image region. To solve this issue, we designed a network architecture with one decoder and two decoders that perform two independent tasks. The first one assigns each image pixel a class label, without considering the un-annotated image region. This may cause segmentation artefacts over the image periphery, where there is not enough information for providing a meaningful classification of pixels. The second decoder evaluates image sharpness and brightness and returns a region of interest to refine segmentation results. The overall network is trained so that the encoder is optimized only by the back-propagation signal coming from the semantic decoder. Combined results proved to provide accurate segmentation of layers and reliable ROIs, even when the area of interest was not centred in the frame. We verified model's calibration so that the adoption of the Monte Carlo dropout approach allows to estimate pixel-wise epistemic uncertainty maps that were consistent with local segmentation flaws in our tests. Overall segmentation accuracy was above 90% and layer-wise metrics well correlated with those in the inter-annotator variability study suggesting the possibility to obtain better results with refined ground truth.

This is the first published work on semantic segmentation algorithms specifically devised for digital gonio-photographs.

7.2.3 The Angle Aperture Classification Algorithm

The estimation of the angle aperture is important for the categorization and management of glaucoma and for the correct assessment of risk factors. Previous work on automatic angle aperture grading has been based on gonio-photographs acquired with the EyeCam device and reported good overall performance. However the images considered in those studies represent very wide angle sectors (about 90°-wide) and their grading is not suitable for the identification of local angle closures (e.g., synechiae). Motivated by the clinical need for a more precise evaluation of angle closures we generated a pilot dataset of local (about 5°-wide) angle aperture annotations and trained and tested a custom baseline deep learning classifier to have a preliminary understanding of potentials and limitations of this approach. This work was presented in Chapter 6.

We started considering a three-class grading problem, with the *Open*, *Occludable* and *Closed* classes based on a simplified version of the Spaeth's clinical grading system. We noticed that the *Occludable* class is the one returning worse classification performance likely because of the weak characterization of the features that differentiate it from the *Open* class, that rely on the visibility of ciliary body band and/or the scleral spur (the layers showing the largest variability in annotations from multiple experts according to our previous study).

We then found that different class aggregation strategies and loss weighting may improve classification performance obtaining useful information to guide future developments.

7.3 Contributions

This research thesis contributes to the very limited existing literature on automatic systems for the analysis of gonio-photographs both with new technical developments and with the generation and analysis of new annotated datasets. The key contributions may be summarised as follows.

- The first study on inter-annotator variability at delineating anatomical layers of the drainage angle in gonio-photographs (Chapter 4) relevant both for the development and the evaluation of automatic systems and for future clinical studies.
- The development and evaluation of a new approach for the semantic segmentation of anatomical layers of the drainage angle in gonio-photographs (Chapter 5) capable of effectively dealing with limitations of ground truth posed by the characteristics of the images under study.
- The completion of a pilot study aiming at investigating advantages and limitations of a prototype deep learning classifier for automatic *local* angle aperture grading (Chapter 6) in digital gonio-photographs.

7.4 Limitations and Future Work

This section discusses the main limitations of our work and considers possible solutions to be explored in the future.

7.4.1 Datasets

The most important limitation of our research is the size of the datasets that have been collected and annotated.

The NIDEK GS-1 is still quite a novel imaging device and the availability of exam databases is very limited. Moreover, the vast majority of available exams show healthy anterior chamber angles that are not representative for the extremely large variability of pathological features, e.g., the shape of synechiae. The under-representativeness of pathological exams makes the training and evaluation of deep learning algorithms problematic, with substantial risk of overfitting the training set and obtain poor generalization.

The annotation of data is costly in terms of time and must be performed by experts with limited availability. This has been particularly true in our work because of the COVID-19

pandemic which has focused most of the efforts and time of our clinical collaborators towards more urgent activities over an extended period of time.

The activity of annotating images is particularly difficult and tedious in the case of generating semantic segmentation ground truth. Our choice of using an already available annotation tool [21], despite this software being very effective in a wide variety of scenarios and adequately customizable to meet specific requirements, was mainly dictated by the lack of time necessary to develop our own tool and was a sub-optimal solution in our case. In particular the annotation of adjacent angle layers led the experts to delineate layer-layer interfaces twice. However a new semi-automatic proprietary annotation tool has been recently developed by NIDEK Technologies Srl. (Spagnuolo and De Giusti [118]) specifically to ease and speed up the delineation of targets in images and will be used to obtain ground truth for our future work, thus allowing to generate more annotations in the same amount of time spent in this activity.

The selection and annotation of more pathological cases is advisable in future research to better train and evaluate the performance of automatic analysis systems and also to confirm the baseline obtained with the inter-annotator variability study.

7.4.2 Variability of Annotations

Our study on inter-annotator variability of anatomical layers delineations provides a means to specifically assess the performance of the segmentation model meaningfully, but also insights useful for the evaluation of other analysis systems for our digital gonio-photographs. We acknowledge, however, that a really comprehensive study on the variability of annotations should, possibly, involve more experts and, more importantly, comprise an analysis on intra-annotator variability. This would provide an estimate of the ground truth variability due to the random contribution of, among others, human attention and tiredness, and of human-machine interactions (e.g., eye-hand coordination) rather than due to experts' bias. This information might be then compared with the inter-annotator variability to understand which of the two is the most relevant.

7.4.3 Semantic Segmentation

Our semantic segmentation network performs well in most of the examined cases, and, from a preliminary qualitative analysis on a limited set of images, appears to provide a reliable interpretation of its outputs that can highlight flaws and suggest the user to perform further actions in case of uncertain results. A more accurate assessment of this has to be carried out on a larger test dataset and under clinicians' supervision.

It must be said that most of the current imperfections returned by the segmentation model involve local (and usually very limited) patches of pixels that do not comply with the topology of the angle layers. A suitable strategy for incorporating constraints about the topology of the anterior chamber angle should be devised, to further improve the performance of the system. Few research papers on topologically correct segmentations of retinal OCT layers have been published and should be considered as baseline [41, 42].

7.4.4 Angle Aperture Grading

The deep learning classifier trained and tested in our work receives sub-sectors of the anterior chamber angle as inputs and returns the predicted aperture class. The variability of the visual features in our gonioscopic images is extremely large and most of them do not directly concern the aperture of the angle. This means that the classifier must identify few relevant characteristics of the images, while learning to generalise the useless (for this task) information provided by many others. This makes the training process very difficult given the limited size of the dataset, an issue already discussed previously in this chapter.

A possible solution could consist in preprocessing the input data to reduce their complexity, e.g., by filtering out features that are not discriminative for angle aperture grading, and so allow the model to focus on better detecting relevant ones. Additional experiments using more sophisticated, pre-trained networks are recommended in the future to better evaluate whether the current limitations are mainly due to the size of the annotated dataset.

The assignment of a single aperture class label to images that could show features associable to more than one morphological condition makes the dataset noisy and the evaluation

of results difficult. The implementation of systems to better interpret model's outputs, e.g., Grad-Cam [111], could help better interpret model's behaviour and support predictions. Moreover, inter and intra-annotator variability studies are advisable to quantitatively assess the consistency of ground truth.

The evaluation of relevant features should be conducted involving clinical experts to ensure a clinically sound approach to this issue.

Additional investigation should be carried out on using the semantic segmentation algorithm as backbone for the estimation of angle aperture. This approach could be advantageous for several reasons; among them the fact that the morphological information obtained trough segmentation could return a deterministic and arbitrarily dense aperture grading based on a selected clinical system relying on the visibility of angle layers. This allows criteria for angle grading to be adjusted without needing either to obtain new ground truth or retrain the network.

References

[1] (2017). European glaucoma society terminology and guidelines for glaucoma, 4th edition - part 1, supported by the EGS foundation. *British Journal of Ophthalmology*, 101(4):1–72.

[2] Abdar, M. et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.

[3] Ahn, J. M. et al. (2018). A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PLOS ONE*, 13(11):e0207982.

[4] Alward, W. L. M. and Longmuir, R. A. (2008). *Color Atlas of Gonioscopy*. American Academy of Ophthalmology.

[5] Asaoka, R. et al. (2019). Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *American Journal of Ophthalmology*, 198:136–145.

[6] Ba, J. L. et al. (2016). Layer normalization. *Arxiv Preprint Arxiv:1607.06450*.

[7] Badrinarayanan, V. et al. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495.

[8] Barão, R. C. et al. (2020). Automated gonioscopy assessment of XEN45 gel stent angle location after isolated XEN or combined phaco-XEN procedures: Clinical implications. *Journal of Glaucoma*, 29(10):932–940.

[9] Budenz, D. L. et al. (2013). Prevalence of glaucoma in an urban west african population: The tema eye survey. *Jama Ophthalmology*, 131(5):651–658.

[10] Cappellari, L. et al. (2020). Deep learning based iridocorneal angle detection for automated gonioscopy. *Investigative Ophthalmology & Visual Science*, 61(7):1620–1620.

[11] Chen, L. et al. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22):4712.

[12] Cheng, J. et al. (2011a). Focal biologically inspired feature for glaucoma type classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 91–98. Springer.

[13] Cheng, J. et al. (2011b). Focal edge association to glaucoma diagnosis. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4481–4484. IEEE.

[14] Chiang, M. et al. (2021). Glaucoma expert-level detection of angle closure in goniophotographs with convolutional neural networks: The chinese american eye study. *American Journal of Ophthalmology*, 226:100–107.

[15] Chotzoglou, E. and Kainz, B. (2019). Exploring the relationship between segmentation uncertainty, segmentation performance and inter-observer variability with probabilistic networks. In *Lecture Notes in Computer Science*, pages 51–60. Springer International Publishing.

[16] Ciresan, D. et al. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in Neural Information Processing Systems*, 25:2843–2851.

[17] Coleman, A. L. et al. (2006). Use of gonioscopy in medicare beneficiaries before glaucoma surgery. *Journal of Glaucoma*, 15(6):486–493.

[18] Cutolo, C. A. et al. (2021). Moving beyond the slit-lamp gonioscopy: Challenges and future opportunities. *Diagnostics*, 11(12):2279.

[19] Deng, J. et al. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

[20] Drozdzal, M. et al. (2016). The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer International Publishing.

[21] Dutta, A. and Zisserman, A. (2019). The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM.

[22] Everingham, M. et al. (2010). The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338.

[23] Feng, R. et al. (2019). Perceptions of training in gonioscopy. *Eye*, 33(11):1798–1802.

[24] Foster, P. J. (2001). Glaucoma in china: how big is the problem? *British Journal of Ophthalmology*, 85(11):1277–1282.

[25] Fu, H. et al. (2018). Multi-context deep network for angle-closure glaucoma screening in anterior segment OCT. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 356–363. Springer International Publishing.

[26] Fu, H. et al. (2019). A deep learning system for automated angle-closure detection in anterior segment optical coherence tomography images. *American Journal of Ophthalmology*, 203:37–45.

[27] Fu, H. et al. (2020). Angle-closure detection in anterior segment OCT based on multilevel deep network. *IEEE Transactions on Cybernetics*, 50(7):3358–3366.

[28] Fu, Z. et al. (2021). MPG-net: Multi-prediction guided network for segmentation of retinal layers in OCT images. In *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE.

[29] Fukushima, K. and Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, 15(6):455–469.

[30] Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

[31] Gedde, S. J. et al. (2021). Primary angle-closure disease preferred practice pattern®. *Ophthalmology*, 128(1):P30–P70.

[32] Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

[33] Glorot, X. et al. (2011). Deep sparse rectifier neural networks. In Gordon, G., Dunson, D., and Dudík, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA. PMLR.

[34] Gulshan, V. et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402.

[35] Guo, C. et al. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

[36] Havaei, M. et al. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31.

[37] He, K. et al. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[38] He, K. et al. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[39] He, K. et al. (2016b). Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, pages 630–645. Springer International Publishing.

[40] He, Y. et al. (2017). Towards topological correct segmentation of macular OCT from cascaded FCNs. In *Fetal, Infant and Ophthalmic Medical Image Analysis*, pages 202–209. Springer International Publishing.

[41] He, Y. et al. (2019). Deep learning based topology guaranteed surface and MME segmentation of multiple sclerosis subjects from retinal OCT. *Biomedical Optics Express*, 10(10):5042.

[42] He, Y. et al. (2021). Structured layer surface segmentation for retina OCT using fully convolutional regression networks. *Medical Image Analysis*, 68:101856.

[43] Hennis, A. et al. (2007). Awareness of incident open-angle glaucoma in a population study: The barbados eye studies. *Ophthalmology*, 114(10):1816–1821.

[44] Hertzog, L. H. et al. (1996). Glaucoma care and conformance with preferred practice patterns. *Ophthalmology*, 103(7):1009–1013.

[45] Hinton, G. E. et al. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

[46] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

[47] Huang, G. et al. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[48] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France. PMLR.

[49] Izatt, J. A. (1994). Micrometer-scale resolution imaging of the anterior eye in vivo with optical coherence tomography. *Archives of Ophthalmology*, 112(12):1584.

[50] Jegou, S. et al. (2017). The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE.

[51] Jin, Q. et al. (2019). DUNet: A deformable network for retinal vessel segmentation. *Knowledge-based Systems*, 178:149–162.

[52] Joskowicz, L. et al. (2018). Inter-observer variability of manual contour delineation of structures in CT. *European Radiology*, 29(3):1391–1399.

[53] Jungo, A. et al. (2018). On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 682–690. Springer International Publishing.

[54] Kauppi, T. et al. (2009). Fusion of multiple expert annotations and overall score selection for medical image diagnosis. In *Scandinavian Conference on Image Analysis*, pages 760–769. Springer.

[55] Kendall, A. et al. (2017). Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *Procedings of the British Machine Vision Conference 2017*. British Machine Vision Association.

[56] Kendall, A. and Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Arxiv Preprint Arxiv:1703.04977*.

[57] Khan, S. et al. (2022). Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s):1–41.

[58] Khanna, A. et al. (2020). A deep residual u-net convolutional neural network for automated lung segmentation in computed tomography images. *Biocybernetics and Biomedical Engineering*, 40(3):1314–1327.

[59] Kiureghian, A. D. and Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112.

[60] Kokkalla, S. et al. (2021). Three-class brain tumor classification using deep dense inception residual network. *Soft Computing*, 25(13):8721–8729.

[61] Krizhevsky, A. et al. (2009). Learning multiple layers of features from tiny images.

[62] Krizhevsky, A. et al. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

[63] Lakshminarayanan, B. et al. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[64] Lambert, B. et al. (2022). Trustworthy clinical ai solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis.

[65] Lampert, T. A. et al. (2016). An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. *IEEE Transactions on Image Processing*, 25(6):2557–2572.

[66] LeCun, Y. et al. (1989). Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems*, 2.

[67] LeCun, Y. et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[68] Lee, C.-Y. et al. (2015). Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. PMLR.

[69] Leite, M. T. et al. (2011). Managing glaucoma in developing countries.

[70] Li, D. et al. (2019). Residual u-net for retinal vessel segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE.

[71] Liu, L. et al. (2019). Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128(2):261–318.

[72] Loftus, T. J. et al. (2022). Uncertainty-aware deep learning in healthcare: A scoping review. *PLOS Digital Health*, 1(8):e0000085.

[73] Long, J. et al. (2015). Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[74] Maas, A. L. et al. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer.

[75] Maier-Hein, L. et al. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1).

[76] Matsuo, M. et al. (2019). Automated anterior chamber angle pigmentation analyses using 360° gonioscopy. *British Journal of Ophthalmology*, 104(5):636–641.

[77] Mehrtash, A. et al. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878.

[78] Minaee, S. et al. (2021). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[79] Mookiah, M. R. K. et al. (2021). A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. *Medical Image Analysis*, 68:101905.

[80] Naeini, M. P. et al. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[81] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Icml*.

[82] Noh, H. et al. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528.

[83] Oktay, O. et al. (2018). Attention u-net: Learning where to look for the pancreas. *Arxiv Preprint Arxiv:1804.03999*.

[84] Organization, W. H. et al. (2012). Global data on visual impairments 2010. *Geneva: Who*, pages 1–5.

[85] Otter, D. W. et al. (2021). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624.

[86] Peng, Y. et al. (2019). DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*, 126(4):565–575.

[87] Perera, S. A. et al. (2010). Use of EyeCam for imaging the anterior chamber angle. *Investigative Opthalmology & Visual Science*, 51(6):2993.

[88] Peroni, A. et al. (2020). A deep learning approach for semantic segmentation of gonioscopic images to support glaucoma categorization. In *Communications in Computer and Information Science*, pages 373–386. Springer International Publishing.

[89] Peroni, A. et al. (2021a). On clinical agreement on the visibility and extent of anatomical layers in digital gonio photographs. *Translational Vision Science & Technology*, 10(11):1.

[90] Peroni, A. et al. (2021b). Semantic segmentation of gonio-photographs via adaptive ROI localisation and uncertainty estimation. *BMJ Open Ophthalmology*, 6(1):e000898.

[91] Peroni, A. et al. (2021c). Semantic segmentation of gonioscopic images exploiting adaptive roi localization and uncertainty estimation. *Investigative Ophthalmology & Visual Science*, 62(8):382–382.

[92] Porporato, N. et al. (2021). Towards 'automated gonioscopy': A deep learning algorithm for 360° angle assessment by swept-source optical coherence tomography. *British Journal of Ophthalmology*, pages bjophthalmol–2020–318275.

[93] Quaranta, L. et al. (2016). Quality of life in glaucoma: A review of the literature. *Advances in Therapy*, 33(6):959–981.

[94] Quigley, H. A. (2006). The number of people with glaucoma worldwide in 2010 and 2020. *British Journal of Ophthalmology*, 90(3):262–267.

[95] Rahman, M. Q. et al. (2013). Direct healthcare costs of glaucoma treatment. *British Journal of Ophthalmology*, 97(6):720–724.

[96] Ribeiro, V. et al. (2019). Handling inter-annotator agreement for automated skin lesion segmentation. *Arxiv Preprint Arxiv:1906.02415*.

[97] Ronneberger, O. et al. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

[98] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

[99] Rotchford, A. P. et al. (2003). Temba glaucoma study: A population-based cross-sectional survey in urban south africa. *Ophthalmology*, 110(2):376–382.

[100] Roy, A. G. et al. (2017). Relaynet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomedical Optics Express*, 8(8):3627–3642.

[101] Rumelhart, D. E. et al. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

[102] Russakovsky, O. et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

[103] Saeed, Z. et al. (2021). Classification of pulmonary viruses x-ray and detection of COVID-19 based on invariant of inception-v 3 deep learning model. In *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*. IEEE.

[104] Sarwinda, D. et al. (2021). Deep learning in image classification using residual network (ResNet) variants for detection of colorectal cancer. *Procedia Computer Science*, 179:423–431.

[105] Sathyamangalam, R. V. et al. (2009). Determinants of glaucoma awareness and knowledge in urban chennai. *Indian Journal of Ophthalmology*, 57(5):355.

[106] Scheie, H. G. (1957). Width and pigmentation of the angle of the anterior chamber: A system of grading by gonioscopy. *Jama Archives of Ophthalmology*, 58(4):510–512.

[107] Schlemper, J. et al. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53:197–207.

[108] Sedai, S. et al. (2018). Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using Bayesian deep learning. In *Computational Pathology and Ophthalmic Medical Image Analysis*, pages 219–227. Springer International Publishing.

[109] Sedai, S. et al. (2019). Uncertainty guided semi-supervised segmentation of retinal layers in OCT images. In *Lecture Notes in Computer Science*, pages 282–290. Springer International Publishing.

[110] Seebock, P. et al. (2020). Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal OCT. *IEEE Transactions on Medical Imaging*, 39(1):87–98.

[111] Selvaraju, R. R. et al. (2016). Grad-cam: Why did you say that?

[112] Selvaraju, R. R. et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[113] Sharma, S. and Guleria, K. (2022). Deep learning models for image classification: Comparison and applications. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 1733–1738. IEEE.

[114] Shi, Y. et al. (2019). Novel and semiautomated 360-degree gonioscopic anterior chamber angle imaging in under 60 seconds. *Ophthalmology Glaucoma*, 2(4):215–223.

[115] Siddique, N. et al. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057.

[116] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Arxiv Preprint Arxiv:1409.1556*.

[117] Spaeth, G. L. (1971). The normal development of the human anterior chamber angle: A new system of descriptive grading. *Transactions of the Ophthalmological Societies of the United Kingdom*, 91:709–739.

[118] Spagnuolo, G. and De Giusti, A. (2022). Semi-automatic segmentation for gonioscopic images. *Investigative Ophthalmology & Visual Science*, 63(7):223–F0070.

[119] Srivastava, N. et al. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

[120] Szegedy, C. et al. (2015a). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[121] Szegedy, C. et al. (2015b). Rethinking the inception architecture for computer vision.

[122] Szegedy, C. et al. (2016). Inception-v4, inception-resnet and the impact of residual connections on learning.

[123] Taghanaki, S. A. et al. (2020). Deep semantic segmentation of natural and medical images: A review. *Artificial Intelligence Review*, 54(1):137–178.

[124] Tan, W. et al. (2021). Classification of covid-19 pneumonia from chest ct images based on reconstructed super-resolution images and vgg neural network. *Health Information Science and Systems*, 9(1):1–12.

[125] Tham, Y.-C. et al. (2014). Global Prevalence of Glaucoma and Projections of Glaucoma Burden through 2040. *Ophthalmology*, 121(11):2081–2090.

[126] Ting, D. S. W. et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22):2211.

[127] Tompson, J. et al. (2014). Efficient object localization using convolutional networks.

[128] Traverso, C. E. (2005). Direct costs of glaucoma and severity of the disease: A multinational long term study of resource utilisation in europe. *British Journal of Ophthalmology*, 89(10):1245–1249.

[129] Trucco, E. et al. (2013). Validating retinal fundus image analysis algorithms: Issues and a proposal. *Investigative Opthalmology & Visual Science*, 54(5):3546.

[130] Ulyanov, D. et al. (2016). Instance normalization: The missing ingredient for fast stylization. *Arxiv Preprint Arxiv:1607.08022*.

[131] Wang, C. et al. (2019). Dense u-net based on patch-based learning for retinal vessel segmentation. *Entropy*, 21(2):168.

[132] Wang, W. and Yang, Y. (2019). Development of convolutional neural network and its application in image classification: A survey. *Optical Engineering*, 58(04):1.

[133] Warfield, S. K. et al. (2004). Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921.

[134] Wei, H. and Peng, P. (2020). The segmentation of retinal layer and fluid in sd-oct images using mutex dice loss based fully convolutional networks. *Ieee Access*, 8:60929–60939.

[135] Weinreb, R. N. et al. (2014). The Pathophysiology and Treatment of Glaucoma. *Jama*, 311(18):1901.

[136] Wu, H. et al. (2021). CvT: Introducing convolutions to vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

[137] Yim, J. et al. (2020). Predicting conversion to wet age-related macular degeneration using deep learning. *Nature Medicine*, 26(6):892–899.

[138] Zandonà, A. and Cappellari, L. (2021). Deep learning based best focus image detection in automated gonioscopy. *Investigative Ophthalmology & Visual Science*, 62(8):2136–2136.

[139] Zhou, T. et al. (2021). Automatic covid-19 ct segmentation using u-net integrated spatial and channel attention mechanism. *International Journal of Imaging Systems and Technology*, 31(1):16–27.

[140] Zhou, Z. et al. (2018). UNet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer International Publishing.

Appendix A

This appendix reports the questionnaire that was shared with ophthalmologists to collect their opinion on what automatic analysis tools would be the most useful for supporting the clinical evaluation of digital gonio-pohotographs acquired with the NIDEK GS-1 device.

Results were used to focus our research as it is discussed in Chapter 3

NIDEK GS-1 Automatic Image Analysis – Questionnaire

The present questionnaire has the purpose of collecting the clinicians' valued opinion on which would be the most clinically useful automatic analysis tools for NIDEK GS-1. The results will lay the base for successive feasibility studies.

1. *In your opinion, how important/useful would be to automatically detect and localize the following anatomical structures?*

	Priority			Estimated number of cases seen in a month (if pathological)
	High	Medium	Low	
1.1. Synechiae	☐	☐	☐	———
1.2. Neovascularization	☐	☐	☐	———
1.3. Schwalbe's line	☐	☐	☐	
1.4. Scleral spur	☐	☐	☐	

Notes:

_____

_____

2. *In your opinion, how important/useful would be to automatically classify the following elements?*

	Priority		
	High	Medium	Low
2.1. Angle aperture classification, according to one of the standards (i.e., Spaeth)	☐	☐	☐
2.2. Trabecular meshwork pigmentation	☐	☐	☐

Notes:

_____

_____

3. *In your opinion, how important/useful would be to perform the following analyses?*

	Priority		
	High	Medium	Low
3.1. Automatic identification of the angle layers (sclera, trabecular meshwork, ciliary body, etc)	☐	☐	☐
3.2. Angle profile extraction	☐	☐	☐
3.3. Estimation of the synechiae size in degrees	☐	☐	☐
3.4. Angle aperture measurement in degrees	☐	☐	☐
3.5. Image focus level classification (on focus, blurred, etc)	☐	☐	☐

Notes:

_____

_____

Appendix B

This appendix reports the final version of the annotation protocol devised for the delineation of anatomical layers in digital gonio-photographs. More details on this can be found in Chapter 3. Ground truth obtained according to this annotation protocol have been used in the study on inter-annotator variability (Chapter 4) and in the development of the semantic segmentation algorithm (Chapter 5).

G.A.I.A. Project – Gonioscope Automatic Image Analysis

Phase 1: Semantic Segmentation of the Irido-corneal Angle Structures

ANNOTATION TOOL AND PROTOCOL

Andrea Peroni

October 2019

University of Dundee

Contents

1. Project Overview

1.1. G.A.I.A. Project General Aims

The G.A.I.A project is a collaboration between the CVIP/Vampire research group at the University of Dundee, NIDEK Technologies S.r.l. and clinical structures in Dundee, Edinburgh, Genova and Lisbon.

Its target is the development of machine learning algorithms for a new ophthalmic device conceived to perform gonioscopy, called GS-1. These algorithms will support the diagnosis procedure providing information of interest to clinicians.

The project is divided into several phases, each with different purposes. In this document, the first phase is described together with the annotation tool user guide and the annotation protocol.

1.2. Phase 1: Semantic Segmentation of Structures

The Phase 1 target is the design of a **semantic segmentation algorithm** capable of identifying all the different structures of interest located in the irido-corneal angle region in order to provide information on the layers interfaces location and their visibility.

It is also a general-purpose algorithm that can be used to achieve more specific outcomes in subsequent project phases.

1.3. Phase 1: Annotation Task

Neural networks need many annotated images in order to learn how to identify features associated with the specific regions that we want to automatically locate in GS-1 acquisitions.

The aim of the annotation task for this particular purpose is to provide accurate information about the boundaries between the irido-corneal angle structures.

These annotations will be used to **train and validate** a semantic segmentation algorithm which is already under development.

2. Scope of Annotations

Targets for annotations are three: *angle layers*, *synechiae* and *neovessels*.

2.1. Angle Layers

For the purposes of this study, seven physiological angle layers have been selected to be annotated.

Two different kind of annotations shall be performed, depending on the considered angle layer.

2.1.1. Segmentation of Layers

This kind of annotation must be used to segment six out of the seven angle layers.

Starting from the iris and moving towards the cornea, they are:

- *Iris root*
- *Ciliary body band*
- *Scleral spur*
- *Pigmented trabecular meshwork*
- *Non-pigmented trabecular meshwork*
- *Cornea*

Annotations shall consist of polygons outlining the contours of the above-mentioned regions; the polygon points shall be manually selected by the annotator and shall follow the contours of each visible layer at their best.

2.1.2. Information on Visibility – Schwalbe's Line

A different kind of annotations shall be provided for the *Schwalbe's line*.

In this case a mutually exclusive choice among three possible options is requested, after an accurate inspection of the image.

The available options are:

- Visible & Pigmented: if the Schwalbe's line is visible and pigmented.
- Visible & Not Pigmented: if it is visible and not pigmented.
- Not Visible: if it is not clearly identifiable in the image.

2.2. Synechiae and Vessels

An additional annotation is required if synechiae or vessels are present.

Synechiae shall be annotated using a bounding-box that includes the synechia and part of the surrounding image region.

Vessels can be annotated using a bounding-box or a polygon. As a rule of thumb, a bounding-box is appropriate if the vessel length is approximately <1/10 of the image width, otherwise a polygonal shape is to be preferred.

As for synechiae, the bounding-box or the polygon that highlights the vessel must include some context. Vessels shall be labelled as *"Normal Vessel"* or *"Neovessel"*.

3. Configuring Google Chrome

NOTE: this procedure needs to be performed only once, before starting the annotation process for the first time; on subsequent annotation sessions, it shall not be necessary to repeat it.

Before starting the annotation process, ***Google Chrome*** must be properly configured to comply with the way annotations must be saved, which is described in depth in Section 4.8.

Start up *Google Chrome*, click on the *Customize and control Google Chrome* icon on the top right of the browser interface window and select *Settings* (Figure 1).

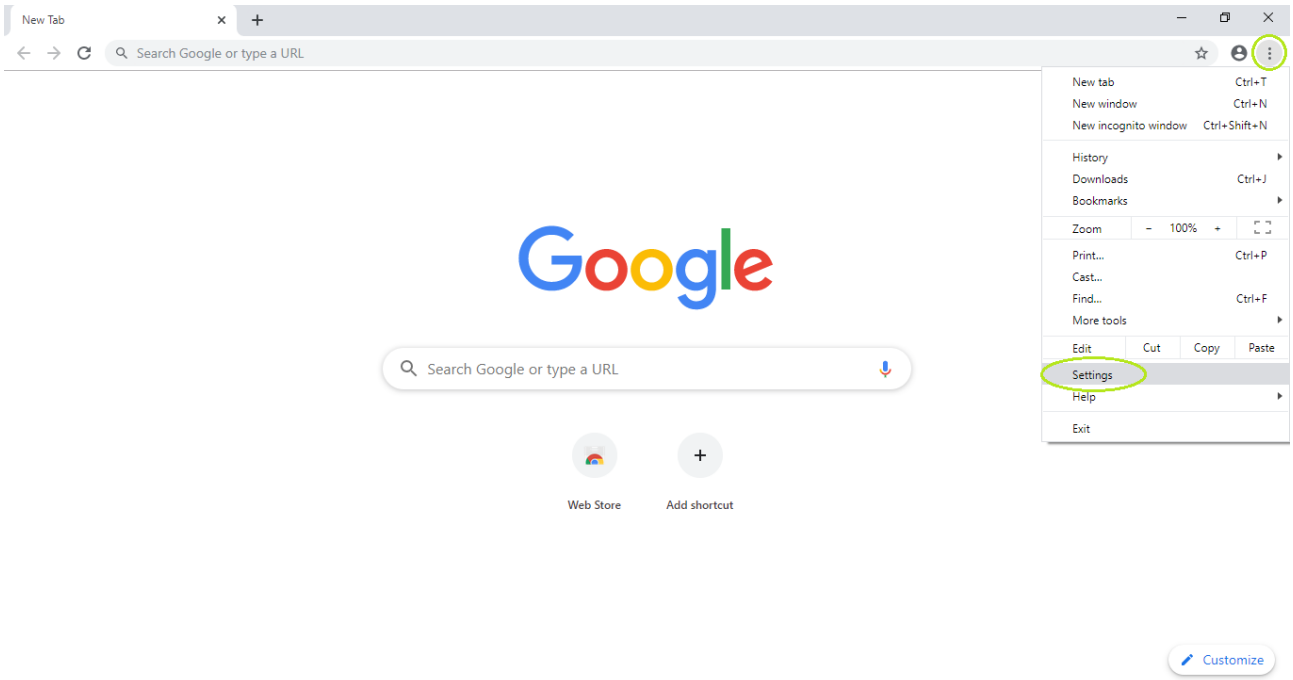

Figure 1: Opening Google Chrome Settings

The *Settings* menu appears. Select *Advanced*, then *Downloads* from the side menu (Figure 2).

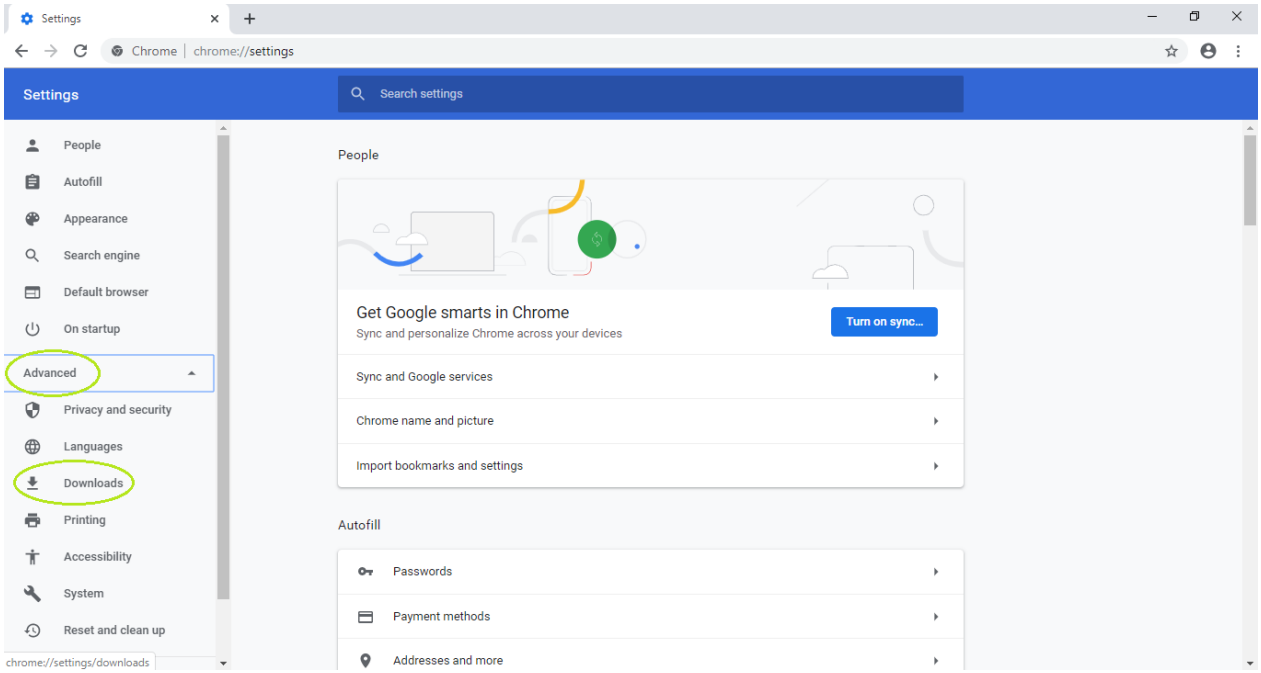

Figure 2: Opening Google Chrome Download menu

From the *Download* menu, enable the *"Ask where to save each file before downloading"* feature by clicking on the icon highlighted in Figure 3.

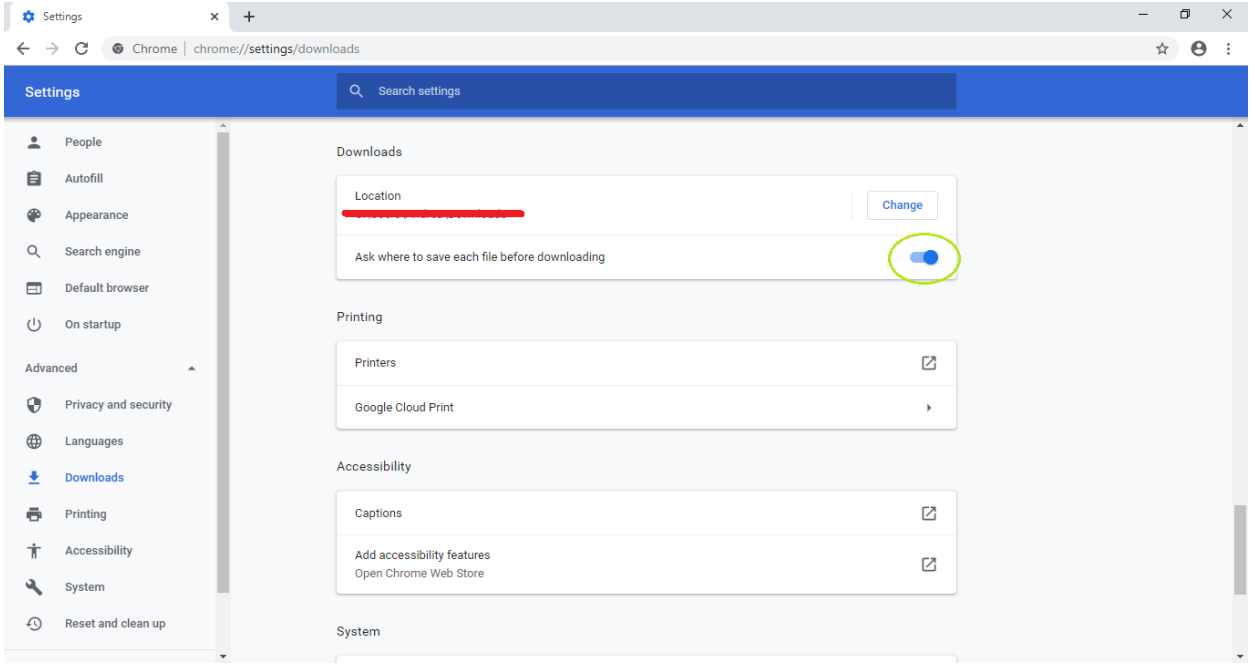

Figure 3: Enabling "Ask where to save file" feature

You can now quit the *Settings* tab; Google Chrome browser is now properly configured.

4. Annotation Tool and Protocol

4.1. Provided Files

In order to annotate a set of GS-1 images, a folder called *GS-1_Dataset* will be provided.

It contains the following files:

- The annotation tool, called *AnnotationTool.html*.
- A file called *Annotations.json* which stores all the pre-defined settings for the *AnnotationTool.html* and will store all the annotations as soon as they are available.
- A folder called *images* that stores all the images to be annotated.
- A folder called *exams* that contains a list of subfolders. Each of these subdirectories is called like one of the images in the dataset and stores the image itself as well as the complete exam it has been selected from. The exam can be inspected in order to exploit a wider context while annotating the selected irido-corneal angle sector.

The annotation tool shall be started by simply opening the corresponding .html file with *Google Chrome*, which must have previously configured as described in Section 3. Please note that all the information provided in the following paragraphs specifically refers to using this browser.

4.2. Annotation Process Flowchart

In this paragraph the annotation process flowchart is reported (Figure 4).

The flowchart is structured so that actions (blue rectangles) and choices (green rhombuses) are well identifiable.

In particular, the first choice ("*Will you annotate a new image?*") may lead to the end of an annotation session (a condition in which the images may not have all been annotated yet), while the second one ("*Have all images been annotated?*") checks whether the entire annotation task is terminated or not (all the images of the dataset have been annotated).

Close to the actions' rectangles, the references to the document sections describing the action in detail are reported.

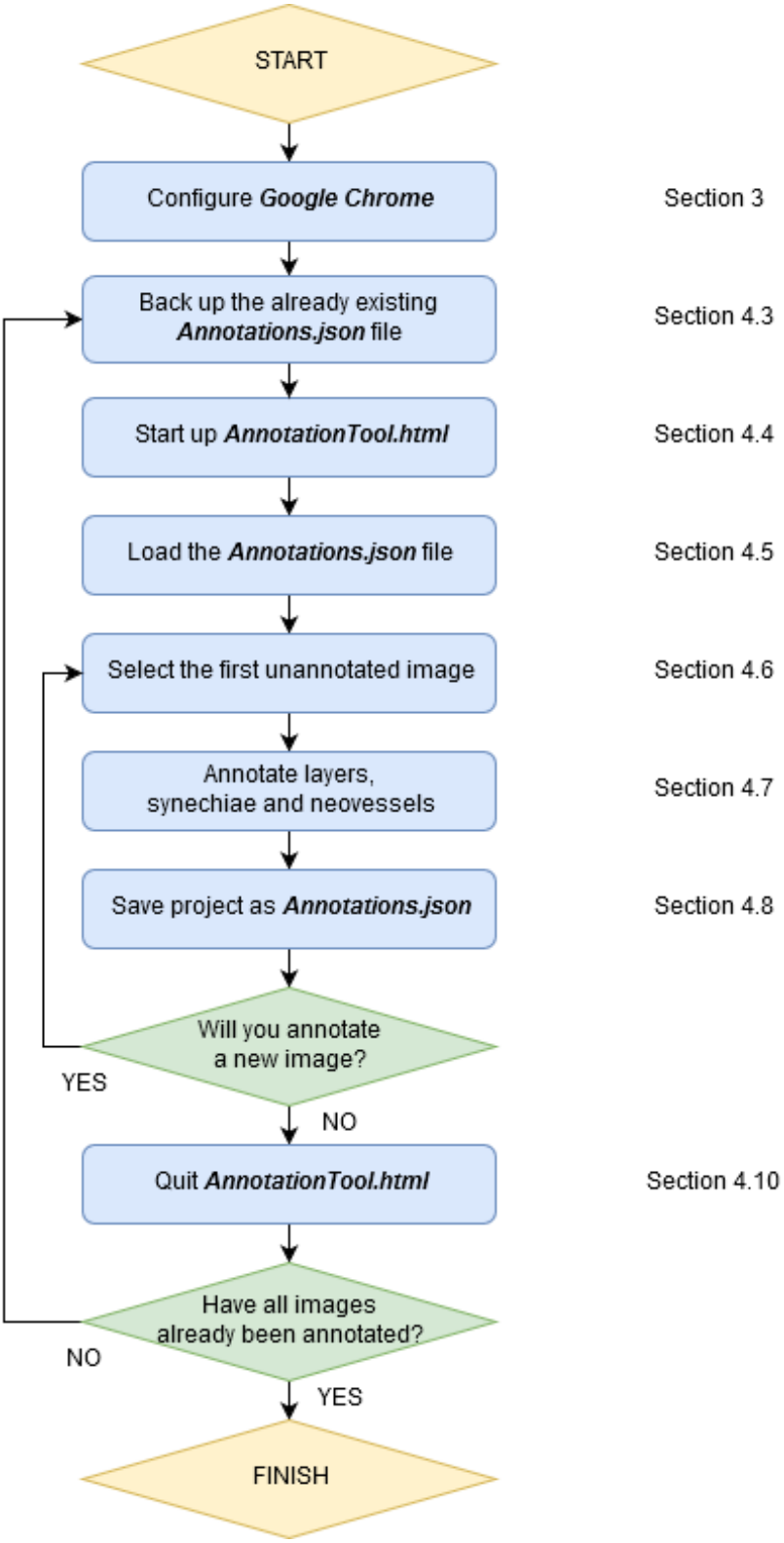

Figure 4: Annotation process flowchart

4.3. Backing up the Annotations.json file

Every time a new annotation session starts, the already existing ***Annotations.json*** file must be backed up.

In order to store a sequence of *Annotations.json* backup files (each corresponding to a different annotation session) a copy of this file must be first created and must be then renamed accordingly to the following template:

Annotations_backup_date_##.json

Where *date* must be replaced by the current date in the format *yyyymmdd* and *##* specifies the session number in that date.

For example, if two annotation sessions were performed on December 1st, 2019, the corresponding backups must be saved as:

- Annotations_backup_20191201_01.json
- Annotations_backup_20191201_02.json

To do so, open the *GS-1_Dataset* folder, select the *Annotations.json* file and then *copy-and-paste* (*Ctrl + C, Ctrl + V*) it in the same folder, a file called *Annotations – Copy.json* is created (Figure 5).

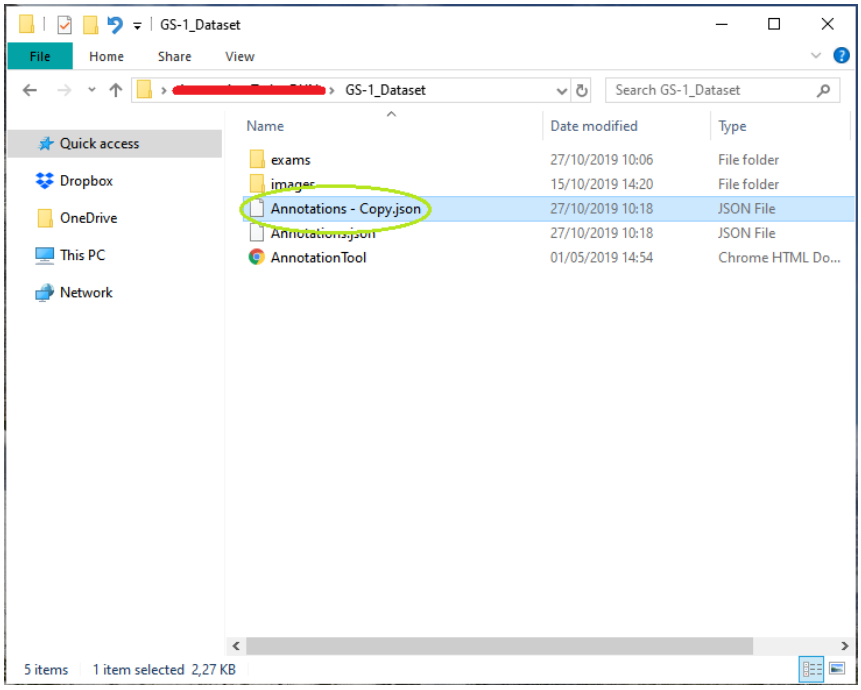

Figure 5: Annotations - Copy.json file

Right-click on the *Annotations - Copy.json* file and select *Rename*. Write the new file name as described above and press *Enter* to confirm (Figure 6).

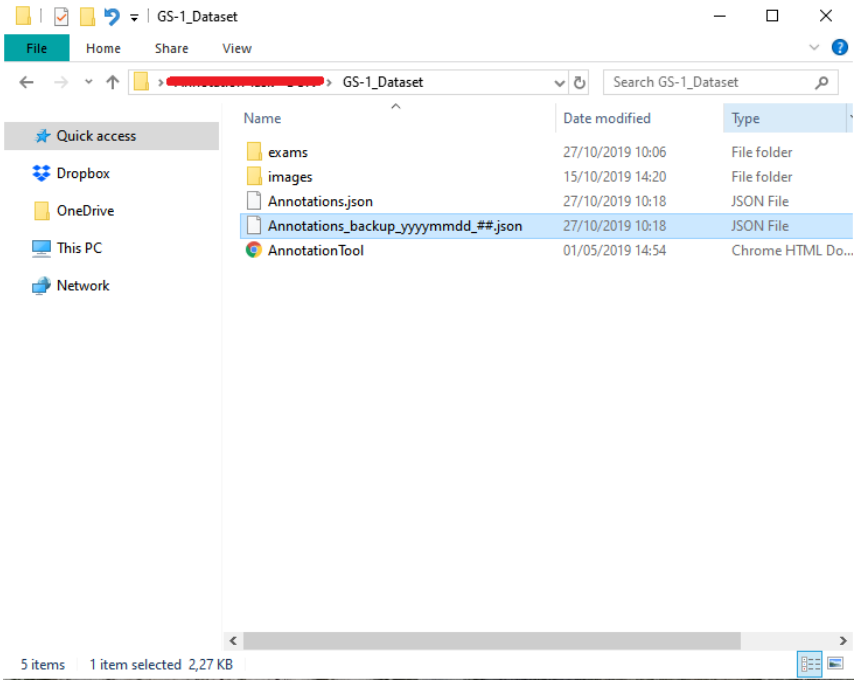

Figure 6: Renaming the Annotations - Copy.json file

4.4. Annotation Tool User Interface

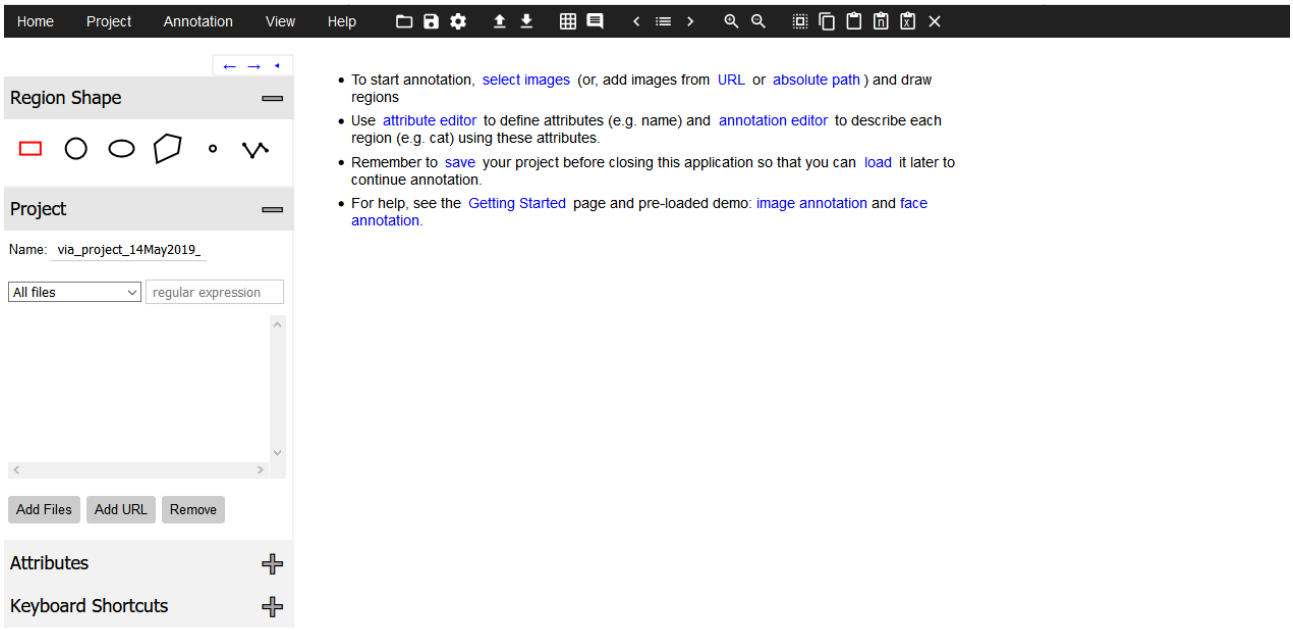

Figure 7: Annotation Tool – home page

The annotation tool selected for this task is the VGG Image Annotator (version 2.0.8) developed by the Visual Geometry Group (VGG) at Oxford University and available for academic and commercial projects[1].

Once the annotation tool is started up by clicking on the *AnnotationTool.html* file, a *Google Chrome* window opens and the interface represented in Figure 7 is shown. It is composed by three main sections:

- The top bar, which comprises project related menus together with view and annotation options. It can be used for:
    - Loading and saving the project, through the "Project" menu
    - Showing/hiding region labels, through the "View" menu
    - Deleting a polygon

- The side menu, where the images are listed, and both the annotation shape and the annotated regions attributes can be handled. It can be used for:
    - Visualizing the list of images
    - Checking what images are still to be annotated
    - Checking the currently visualized image
    - Selecting the region shape
    - Listing the keyboard shortcuts, by expanding the "Keyboard Shortcuts" menu

- The central panel, where the image to annotate will be shown once selected. It can be used for:
    - Performing/visualizing the annotations on the current image.

[1] *Abhishek Dutta and Andrew Zisserman. 2019. The VIA Annotation Software for Images, Audio and Video. In Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3343031.3350535*.

4.5. Project Initialization

NOTE: the following procedure shall be repeated every time an annotation session is started.

In order to load the pre-generated settings, select the "Project" menu in the top bar, then select "Load". A local folder browser appears.

Navigate to the *GS-1_Dataset* folder, select the *Annotations.json* file and press "Open" (Figure 8).

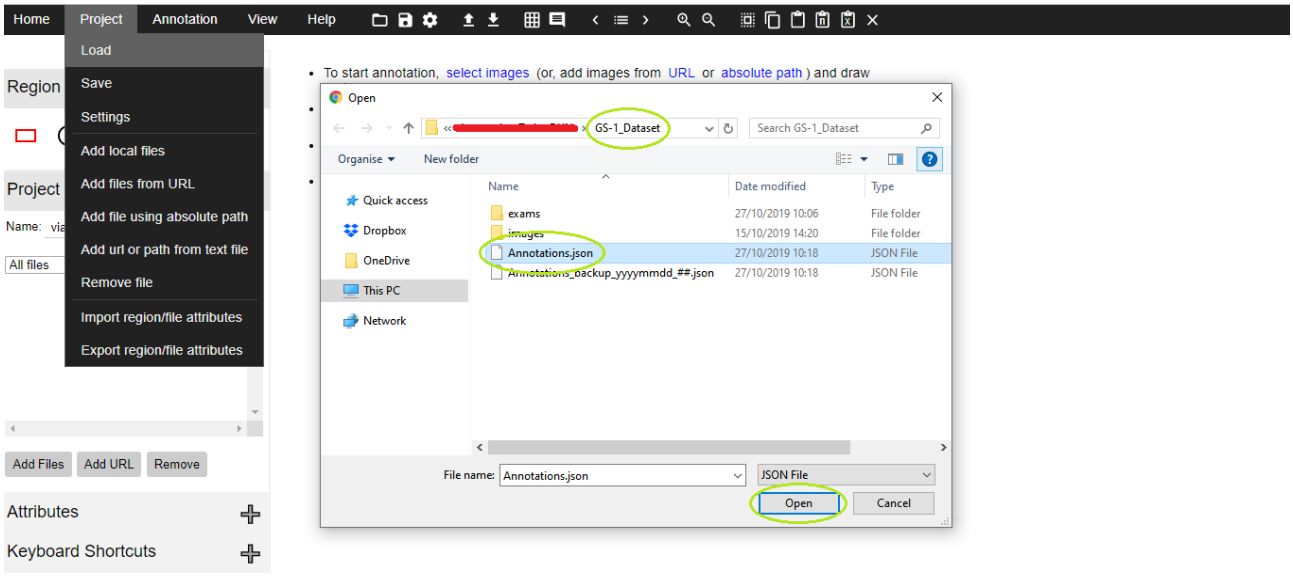

Figure 8: Loading the settings file

Please note that every time you load the Annotations.json file at the beginning of a new annotation session*, a backup copy of that file must have already been created.*

All the images in the *images* folder (both annotated and not) will be listed in the "Project" side menu.

Please note that in this example there are only two images in the folder, but they will be more in real applications.

The first image of the list (alphabetic order) is highlighted in bold type and is shown in the central panel of the tool as reported in Figure 9.

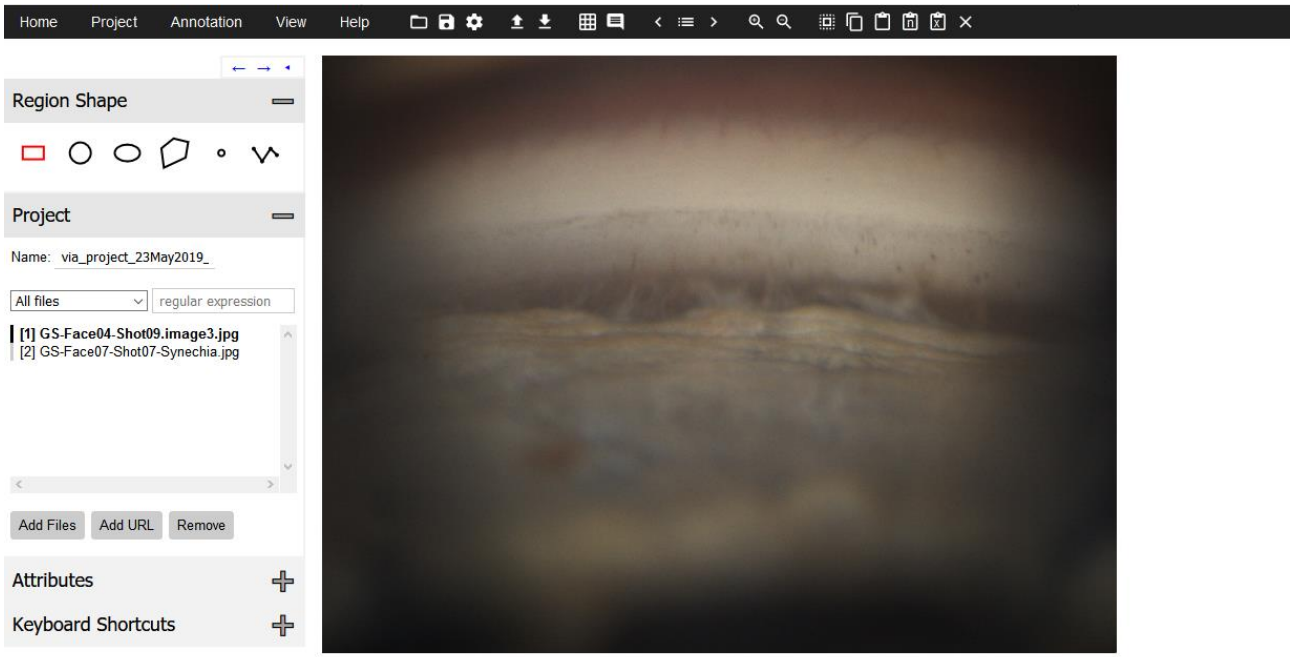

Figure 9: Dataset loaded

If you want the image list to show only the unannotated pictures, select "*Show files without regions*" from the drop-down list in the "Project" side menu. To show again the entire list of images, select "*All files*" from the same drop-down list. The other options in the drop-down list will always lead to an empty list and should not be selected.

4.6. Layers Annotation

4.6.1. Segmentation of Layers

It is now possible to select the "Polygon region shape" from the "Region shape" in the side menu and start creating the contour of **all the layers except the Schwalbe's line** by marking a sequence of points that belong to its boundary.

Layers must be annotated only if they are clearly identifiable in the image with respect to their adjacent structures. If, considering only the visual information the image provides, a layer is not well visible and there is no confidence about the exact location of its boundaries, don't annotate it.

Note: before starting to draw a new annotation, be sure that no other regions are currently highlighted. If so, they could be dragged over the image.

Once a layer annotation is terminated press "Enter" on your keyboard in order to confirm the polygon. To make the labels checklist appear, click on an unannotated region of the picture first,

then click again inside the polygon. Choose the correct name from the checklist, as shown in Figure 10.

Note: if, by clicking on the image, an unwanted new polygon is created, just press the "Esc" button of your keyboard to delete it.

Sometimes, the layer labels may hide an image region you are interested in interacting with. To temporarily remove the labels, select the "View" menu from the top bar and click on "Show/hide region labels (I)". To make the labels appear again, re-click on the same option.

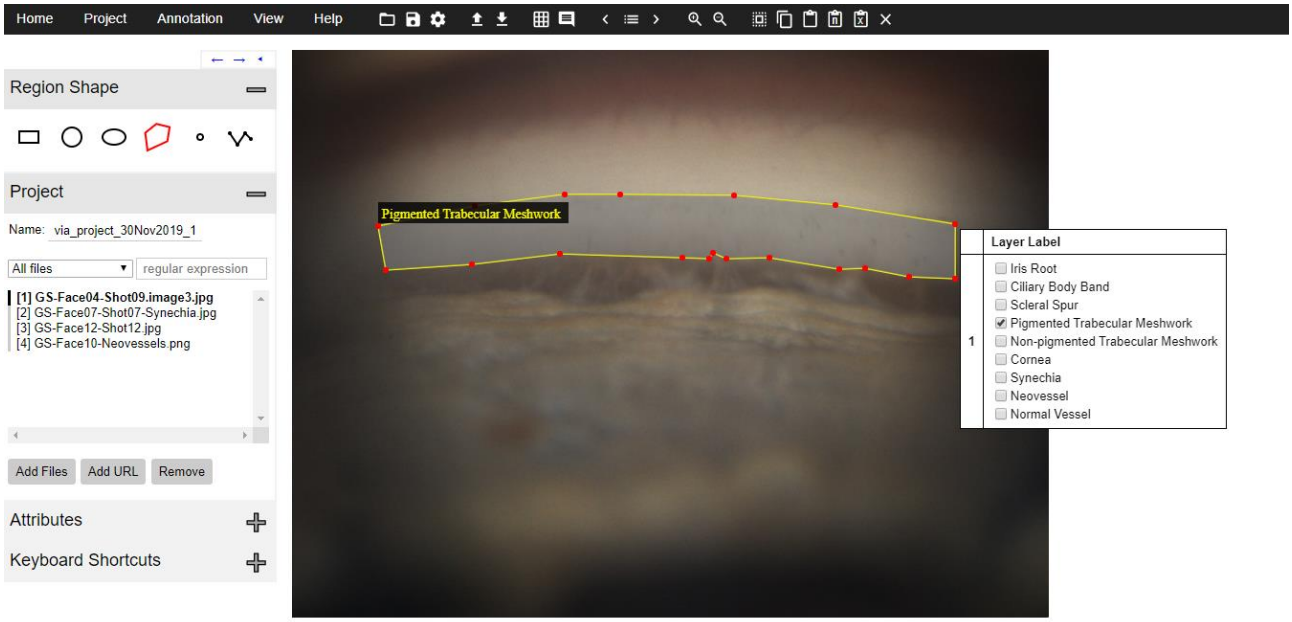

Figure 10: Example of layer annotation

Polygon points are easily and freely adjustable after the polygon has been drawn. To adjust a polygon vertex, keep it clicked and move it over the image.

On Windows, it is possible to add or remove specific polygon vertices once it has been drawn on the image and confirmed.

To add a vertex, keep the *Ctrl* button of the keyboard pressed and left-click with your mouse over the edge of the polygon you want to add the point to.

To remove a vertex, keep the *Ctrl* button of the keyboard pressed and left-click with your mouse over the vertex you want to delete.

Unfortunately, adding/removing a vertex is not possible on Macs.

In "Appendix A: how to choose the correct number of vertices" some examples about the suggested way to draw annotation polygons can be found.

If you want to delete an entire annotated region after its creation, click a point inside the corresponding polygon (the region is now highlighted and all its vertices are visible) and press the "X" button in the right end of the top bar (shown in Figure 11).

<u>Do not use the "Remove" button from the side menu, it will delete the entire image from the dataset.</u>

Figure 11: "Delete region" button

In the case a layer label needs to be edited, select the corresponding polygon to make its checklist appear and change the label.

The following general rules must be considered in order to provide consistent annotations:

- Annotate and label only the visible layers in every image.
- Annotate the image portion that is reasonably bright (e.g. the region inside the ellipse in Figure 12).

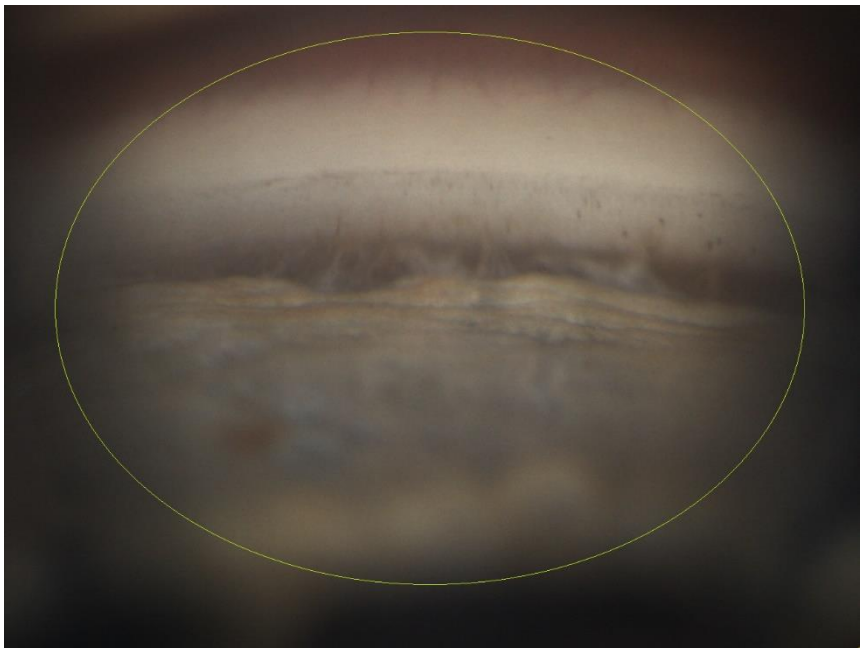

Figure 12: Bright region in a picture

- Annotate only the reasonably in-focus part of the iris (as shown in Figure 13).

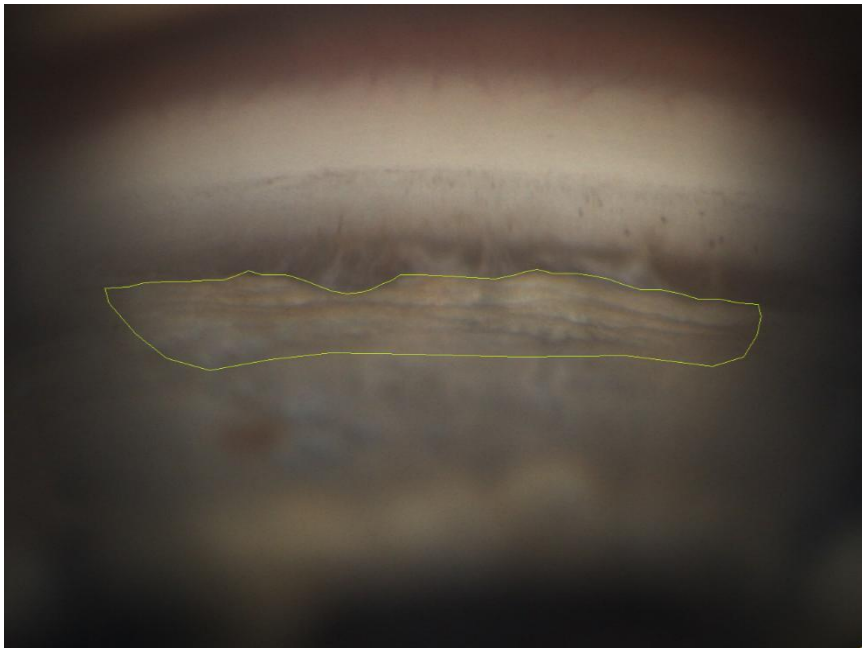

Figure 13: In-focus part of the iris

- Most of the images show a darker layer in the cornea region. If it is present, annotate only the region of the cornea that is externally bounded by this darker layer, as reported in Figure 14.

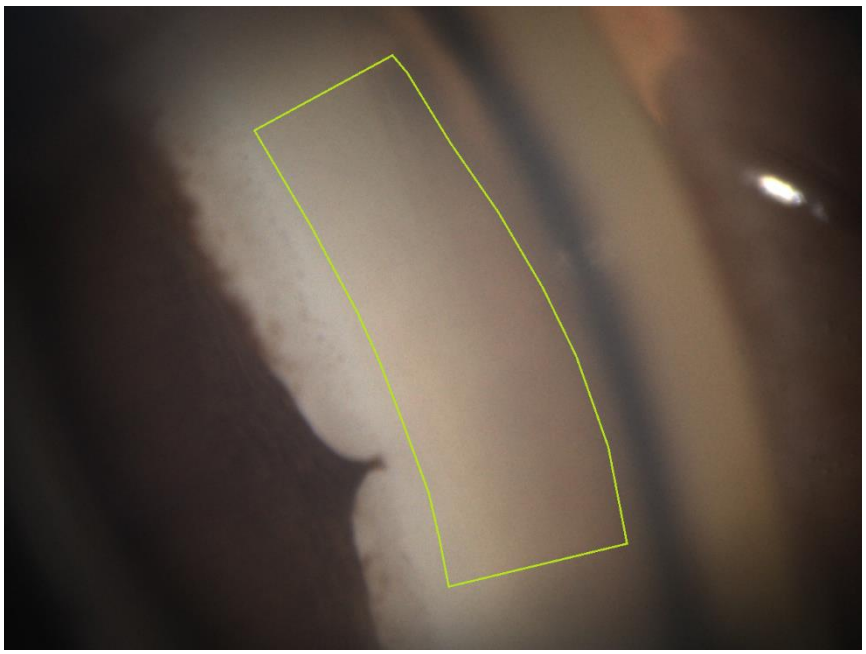

Figure 14: Correct annotation of the cornea

- If something (e.g. MIGs or synechiae) occludes a structure, thus dividing it into several separate regions, draw multiple polygons and assign the same label to all of them. Two

examples can be found in Figure 18 (synechia) and in Appendix C: How to deal with implants in images.

- Polygons that refer to different regions should not overlap.

- Every time there exists an interface between two angle layers, any gap between the annotations of the two subsequent layers must be avoided and the common boundary must correspond to the interface itself (example in Figure 15).

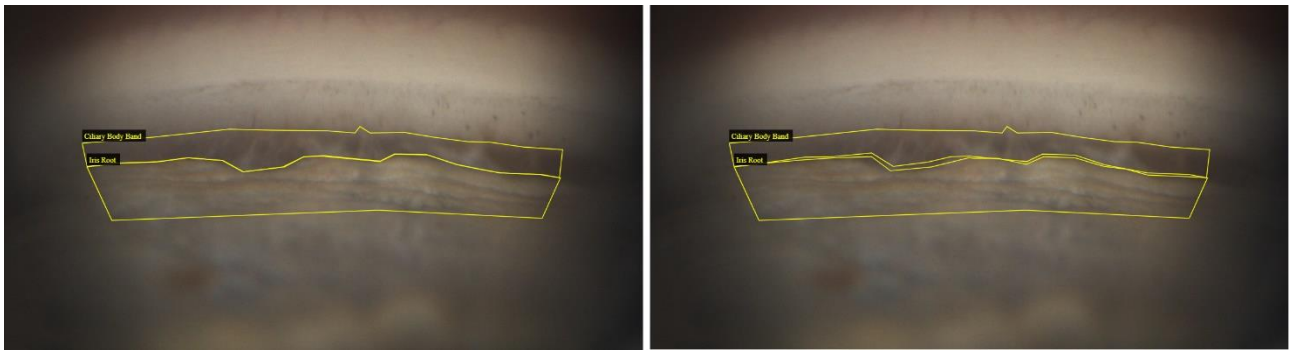

Figure 15: A good interface annotation (left) vs a wrong one (right)

A list of additional useful keyboard shortcuts is reported in "Appendix B: useful keyboard shortcuts".

4.6.2. Information on Visibility – Schwalbe's Line

The Schwalbe's line shall be annotated differently from the other layers since its characteristics make it difficult to be segmented using a polygon.

In this case, the information about its visibility in the image and its pigmentation grade must be provided.

It can be done by selecting the correct choice among the three available options:

- Visible & Pigmented: if the Schwalbe's line is visible and pigmented
- Visible & Not Pigmented: if it is visible and not pigmented
- Not Visible: if it is not clearly identifiable in the image

Please note that the choices are mutually exclusive, therefore only one of them must be selected.

In order to perform this annotation, press the *Space* button of the keyboard first.

A new menu appears in the bottom side of your screen, as reported in Figure 16.

Figure 16: Bottom menu layout

It is divided into two subsections:

- The first one is called *Region Annotations* and shows all the region annotations of the current image (if already performed)
- The second one is called *File Annotations* and shows the checkbox for annotating the Schwalbe's line

Select the *Image Annotations* tab and tick the correct choice (Figure 17).

Figure 17: Schwalbe's line checkbox

Once the correct option is selected, the annotation is correctly performed.

Before quitting the bottom menu, select the *Region Annotations* tab again. If you don't do so, you won't be able to assign a label to further polygons.

You can now make the bottom menu disappear by pressing again the *Space* button on your keyboard.

4.7. Synechiae and Vessels Annotation

In images in which one or more synechiae or vessels are present, an additional annotation must be performed.

Note: before starting to draw a new annotation, be sure that no other regions are currently highlighted. If so, they could be dragged over the image.

Every synechia annotation is drawn using the *"Rectangular region shape"* from the *"Region shape"* side menu, must include the entire synechia and part of the context around it (see the example in Figure 18) and must be labelled as *"Synechia".*

Please note that once a rectangle is placed, the checkbox appears immediately. Every vessel annotation must highlight a single vessel and comprise a limited region that surrounds it. For this reason, two different approaches may be adopted:

1. The vessel length is approximately < 1/10 of image width: in this case the annotation is performed using the *"Rectangular region shape"* from the *"Region shape"* side menu.
2. The vessel is longer: the annotation shall be performed using the *"Polygon region shape"* from the *"Region shape"* side menu.

If there are several vessels in a small image area, draw multiple annotations, each one centred on a specific neovessel (see some examples in Figure 19).

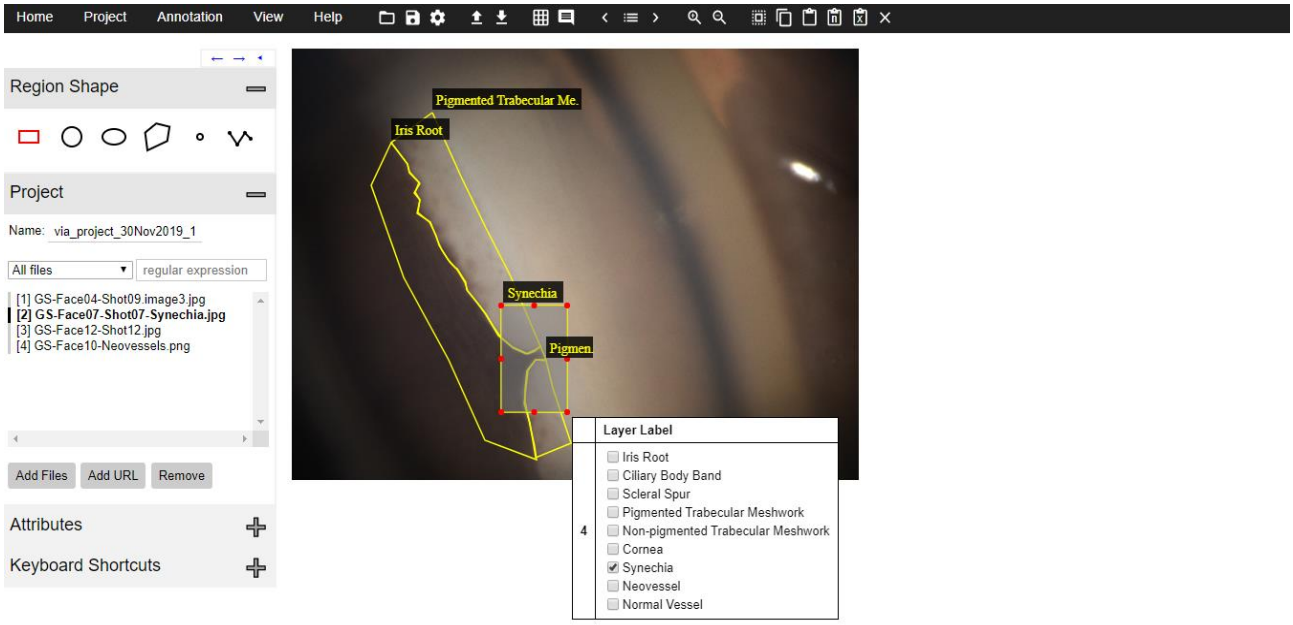

Figure 18: Example of synechia annotation

Figure 19: Example of vessels annotation

4.8. Saving Annotations

Every time an image has been completely annotated (see an example in Figure 20), save the annotations through the "Project" menu in the top bar.

Figure 20: Annotated image

Select "Save". The "**Save Project"** dialog window will appear, as reported in Figure 21.

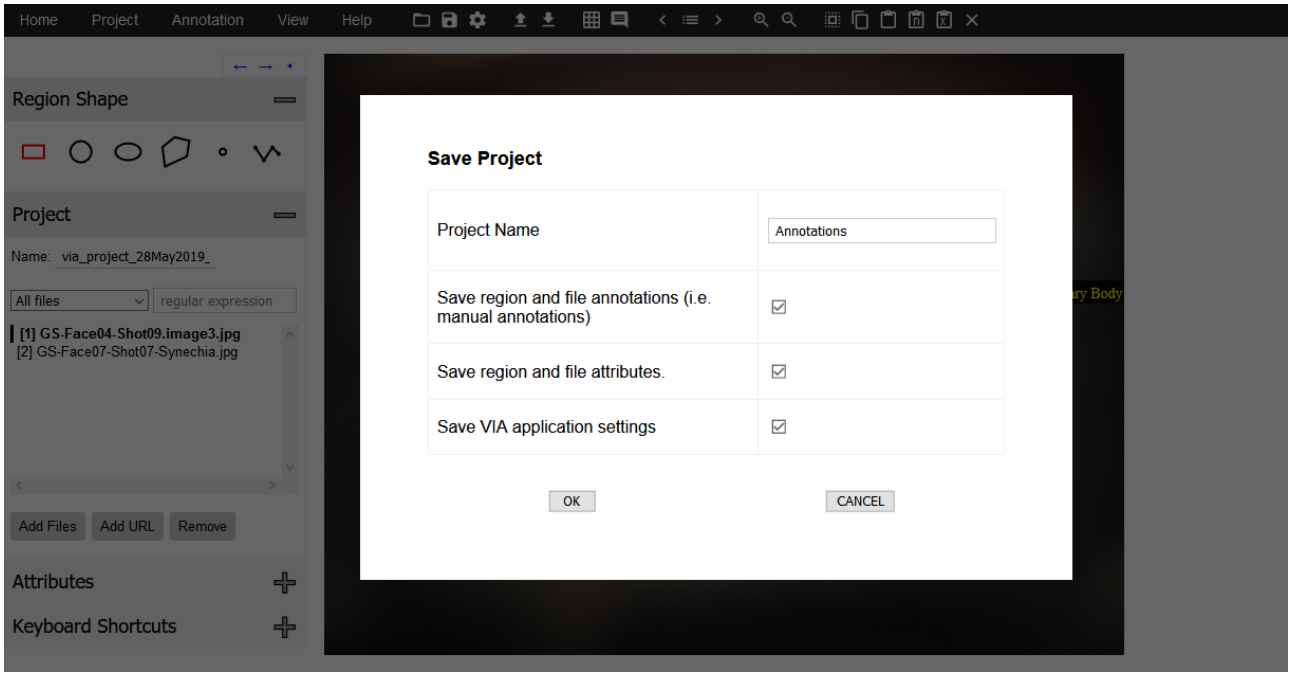

Figure 21: "Save Project" dialog window

Select "OK" **without changing any of the other options**.

The window reported in Figure 22 appears.

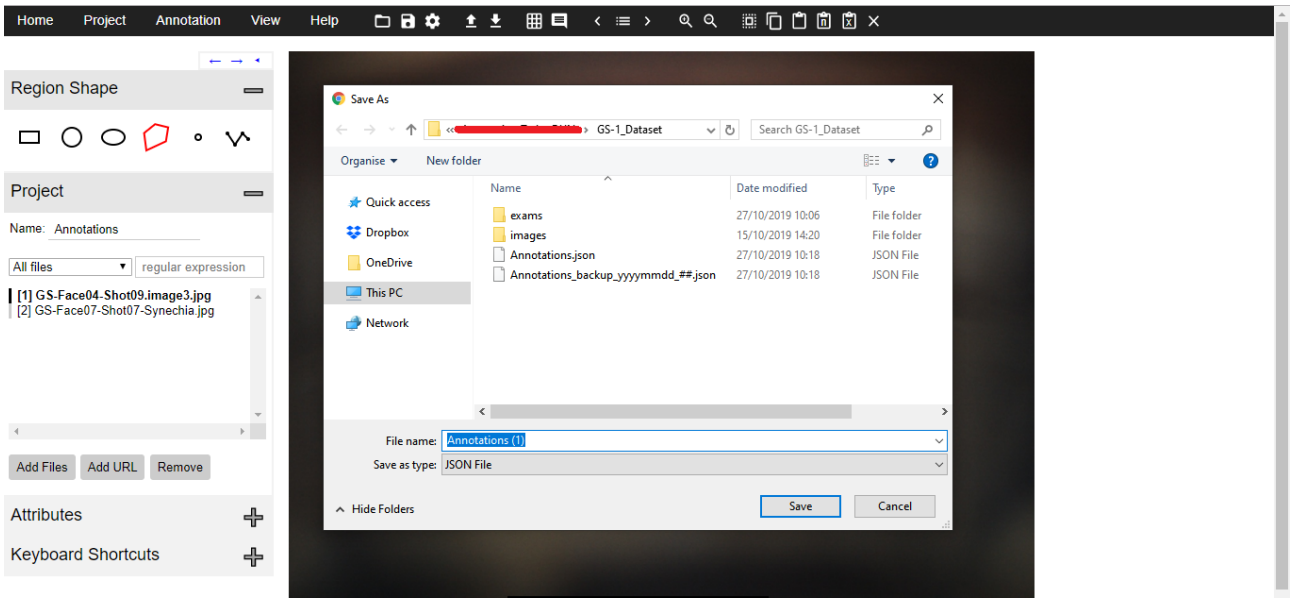

Figure 22: Path selection

Here, the correct path to the *GS-1_Dataset* folder must be provided.

If *Google Chrome* suggests you rename the file you are going to save by adding a number in round brackets, remove it. The filename must be *Annotations*, the type extension (*.json*) is automatically added and must not be written explicitly

Press *Save* and confirm that you want to overwrite the existing *Annotations.json* file (Figure 23).

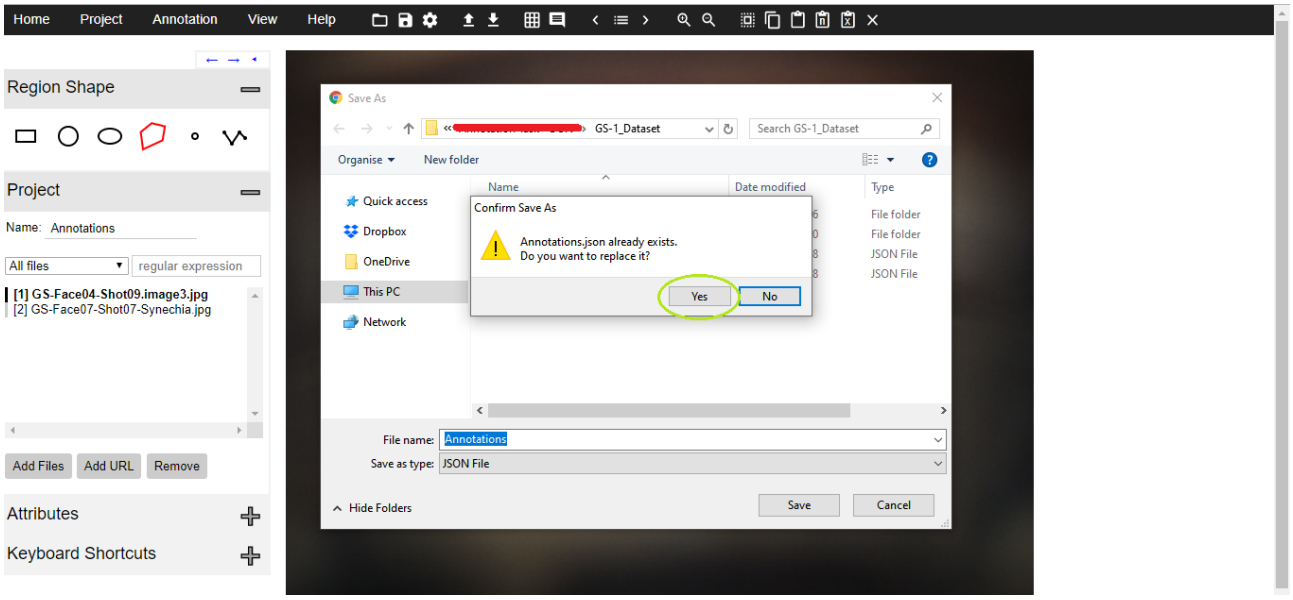

Figure 23: Confirmation window

The updated *Annotations.json* file has now been correctly saved in your *GS-1_Dataset* folder.

You can now annotate a new image or quit the annotation tool.

4.9. Pausing the Annotation Process

If you want to pause the annotation process, save the project as reported in Section 4.8 before quitting the tool.

At the next start up, load the *Annotations.json* file, as in Section 4.5, and restart from where you had stopped.

4.10. Quitting the Annotation Tool

After the project has been saved, it is possible to quit the annotation tool by closing the browser.

If you are asked whether to leave the current page or not, confirm.

4.11. Feedback on the Annotation Process

The annotation of a large batch of images is a process that may require a lot of time. For this reason, it is requested to provide periodic feedbacks in order to evaluate annotations, discuss about possible issues and align the involved clinical partners.

Each periodic feedback shall refer to a specific number of annotated images, and shall consist in a backup of the latest version of the annotation file, created as explained in Section 4.3. The *Annotations_backup_yyyymmdd_##.json* file shall be sent by email to [redacted].

Appendix A: how to choose the correct number of vertices

In this appendix, good practices for an efficient and correct annotation are explained.

It is important to notice that, to make the annotation of a specific structure useful for this research purposes, it should not contain pixels that belong to other regions.

Let's show some examples in order to evaluate the possible consequences of different kind of annotations

In Figure 24, an inaccurate annotation is shown. It contains pixels that belong to different layers. Such an annotation is **not useful and even confusing** for a neural network that tries to learn the characteristics associated with a specific layer. This kind of annotation must be avoided.

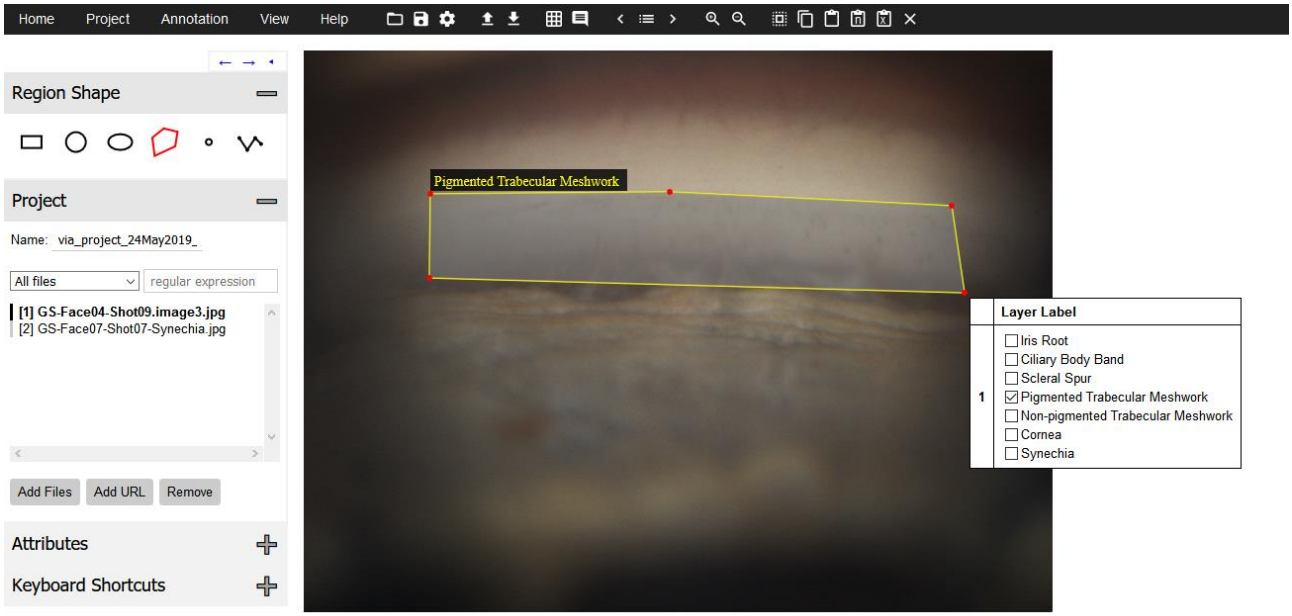

Figure 24: Inaccurate annotation

In Figure 25, an example of an even too accurate annotation is shown.

In this case, there are no negative side effects concerning the learning process of the neural network, but this annotation style is inefficient and time consuming.

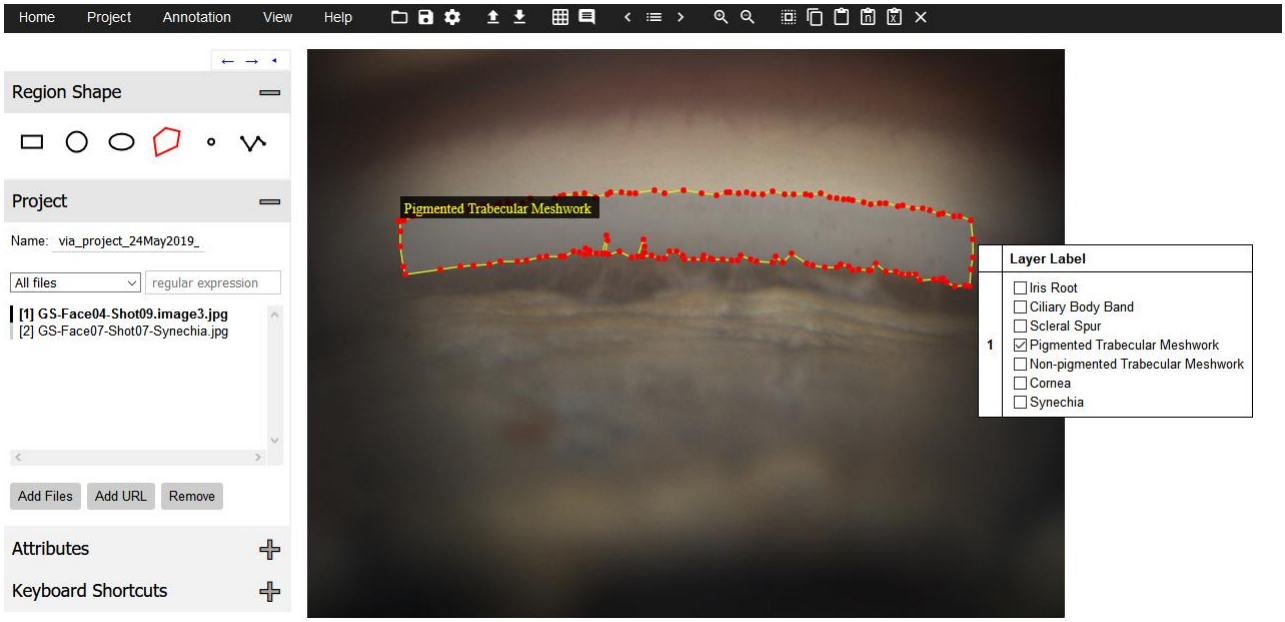

Figure 25: Too accurate annotation

The best solution is an adaptive accuracy approach, characterized by a sequence of vertices that follow the real boundary between two regions, without being too accurate but correctly avoiding the inclusion of different layers.

An example of correct annotation is shown in Figure 26.

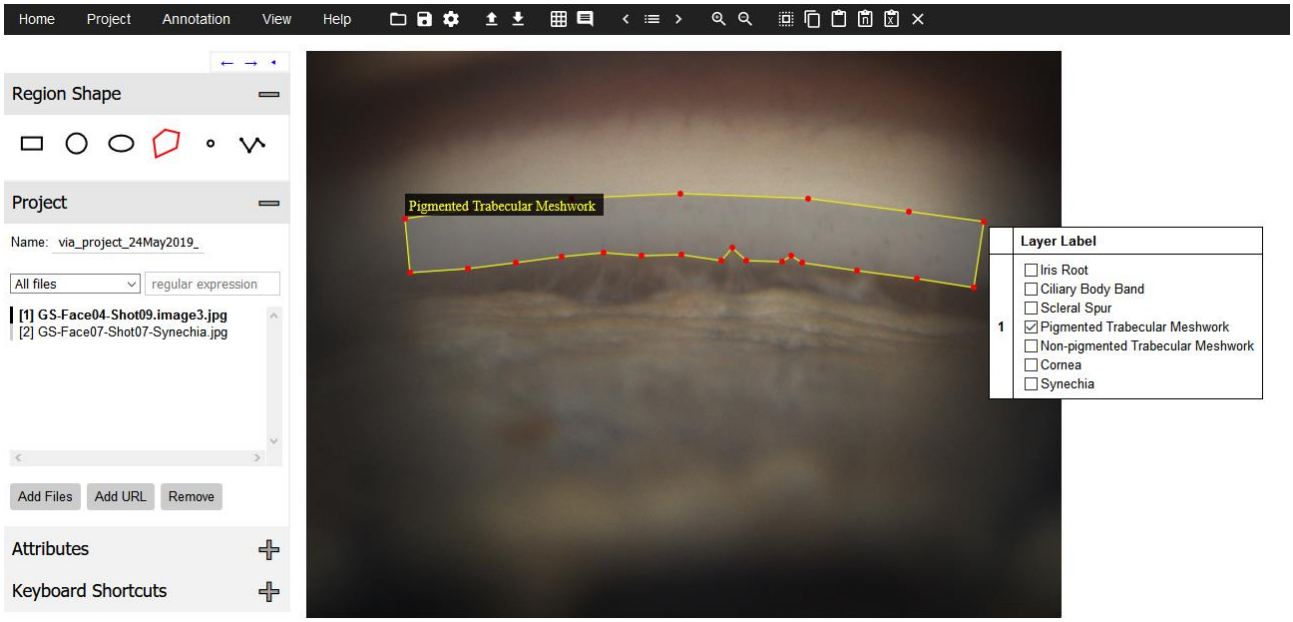

Figure 26: Correct annotation

Appendix B: useful keyboard shortcuts

Here, a list of useful keyboard shortcuts is reported (Figure 27).

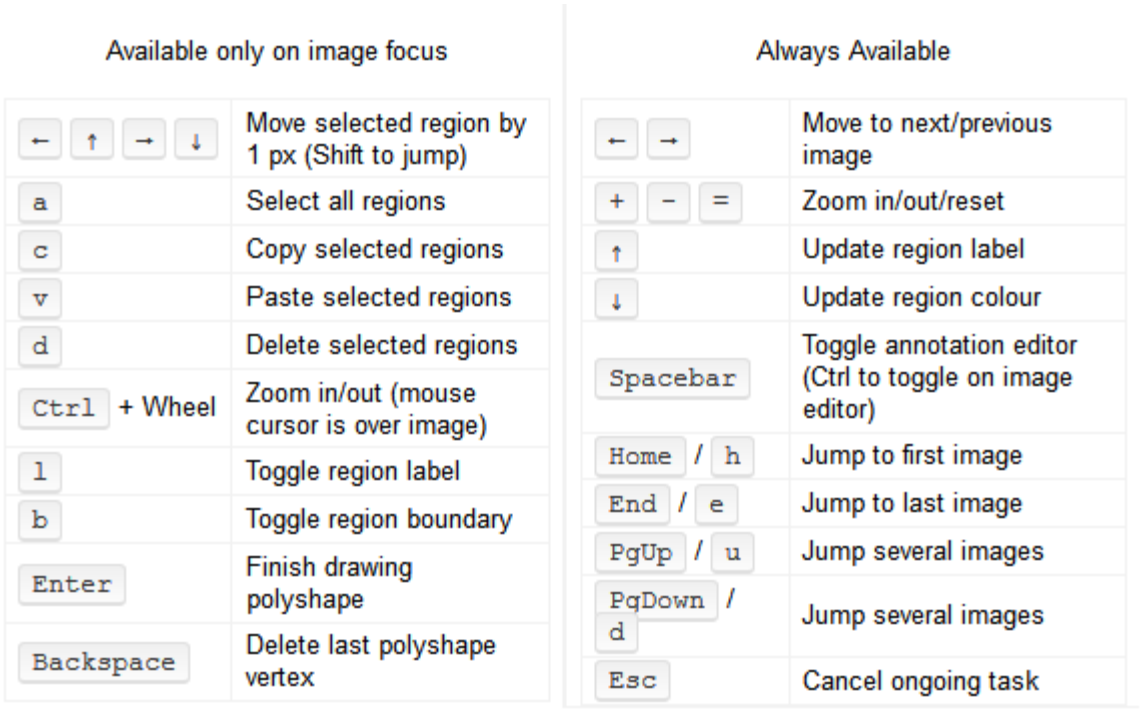

Figure 27: Keyboard shortcuts

Appendix C: How to deal with implants in images

If one or more implants are visible in an image, they must be cut out from the annotations of every layer by following their contours while drawing the region polygons.

If an implant divides a layer into two separate regions, two polygons must be drawn and the same label must be assigned to both of them, as described in Section 4.6.1.

An example on how to deal with their presence is shown below (Figure 28).

Please note that all the layers except for the iris and the cornea have been annotated using two polygons.

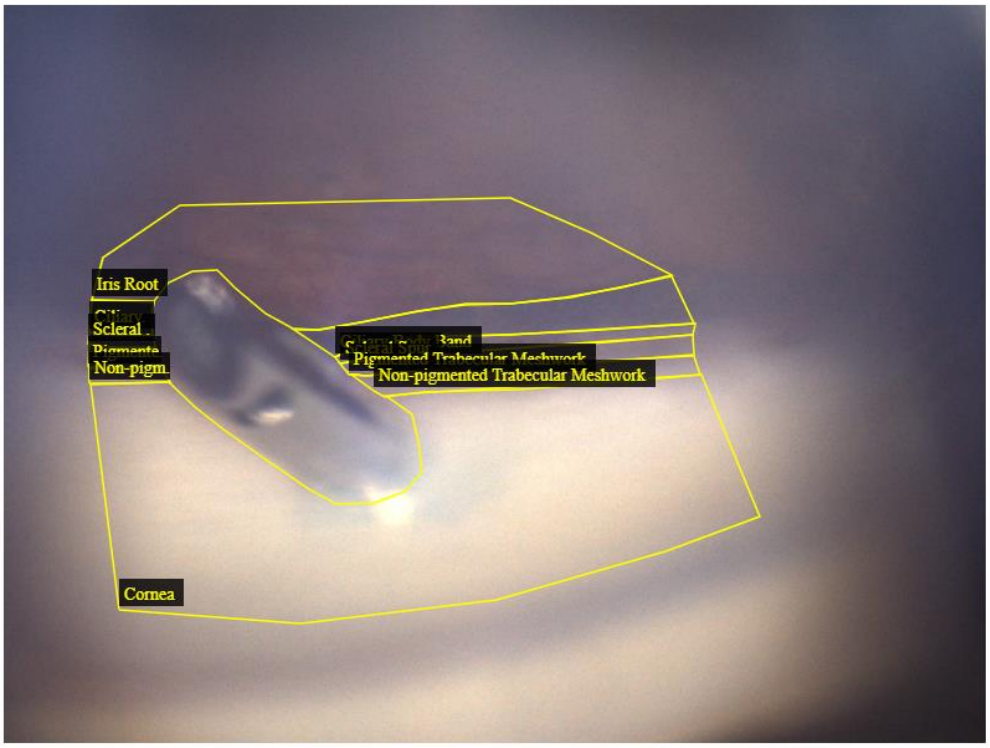

Figure 28: How to cut out an implant

Appendix C

This appendix reports the final version of the annotation protocol devised for the local grading of angle aperture in digital gonio-photographs. More details on this can be found in Chapter 3. Ground truth obtained according to this annotation protocol have been used in the pilot study on local angle aperture classification (Chapter 6).

G.A.I.A. Project – Gonioscope Automatic Image Analysis

Phase 2: Angle Aperture and Trabecular Meshwork Pigmentation Grading

ANNOTATION TOOL AND PROTOCOL

Andrea Peroni

July 2021

NIDEK University of Dundee DUNDEE Vampire

Contents

1 Project Overview

1.1 G.A.I.A. Project General Aims

The G.A.I.A project is a collaboration between the CVIP/Vampire research group at the University of Dundee, NIDEK Technologies S.r.l. and clinical sites in Dundee, Edinburgh, Genoa and Lisbon.

Its target is the development of machine learning algorithms for a new ophthalmic device conceived to perform gonioscopy, called GS-1. These algorithms will support the diagnosis procedure providing information of interest to clinicians.

The project is divided into several phases, each with different purposes. In this document, the second phase is described together with the annotation tool user guide and the annotation protocol.

1.2 Phase 2: Angle Aperture and Trabecular Meshwork Pigmentation Grading

The target of G.A.I.A. Project Phase 2 is the design and development of machine learning algorithms for the automatic classification of *angle aperture* and *trabecular meshwork pigmentation* in radial, equally wide sub-sectors of GS-1 acquisitions depicting the irido-corneal angle.

1.3 Phase 2: Annotation Task

Developing machine learning algorithms requires annotated data, i.e. data previously evaluated by an expert and graded into categories (classes) of interest for a given task. Algorithms can be trained to identify the features (e.g. signal patterns) that better characterize each of the output classes, and can be subsequently used to automatically evaluate unknown data, aiding the analysis of large amounts of raw information.

The aim of this annotation task is to grade radial, equally wide sub-sectors of GS-1 acquisitions in terms of irido-corneal angle aperture and trabecular meshwork pigmentation (when visible), as explained in the following paragraphs.

The resulting annotations shall be used to **train and validate** classification algorithms to automatically evaluate and grade angle aperture and trabecular meshwork pigmentation of irido-corneal angle sub-sectors of GS-1 acquisitions.

2 Scope of Annotations

GS-1 full-mode exams consist of 16 partially overlapping sectors of the irido-corneal interface of an eye, each acquired as a stack of several shots with different focal planes. A representative image of each sector is considered for the purposes of this work. All sector images are pre-processed to uniform the visualization of data by rotating them so that the iris root is always shown at the bottom of the image and the cornea at its top. Three radial, parallel and equally wide sub-sectors are highlighted in the best-lit region of each sector image; these are the sub-sectors to annotate (Figure 1).

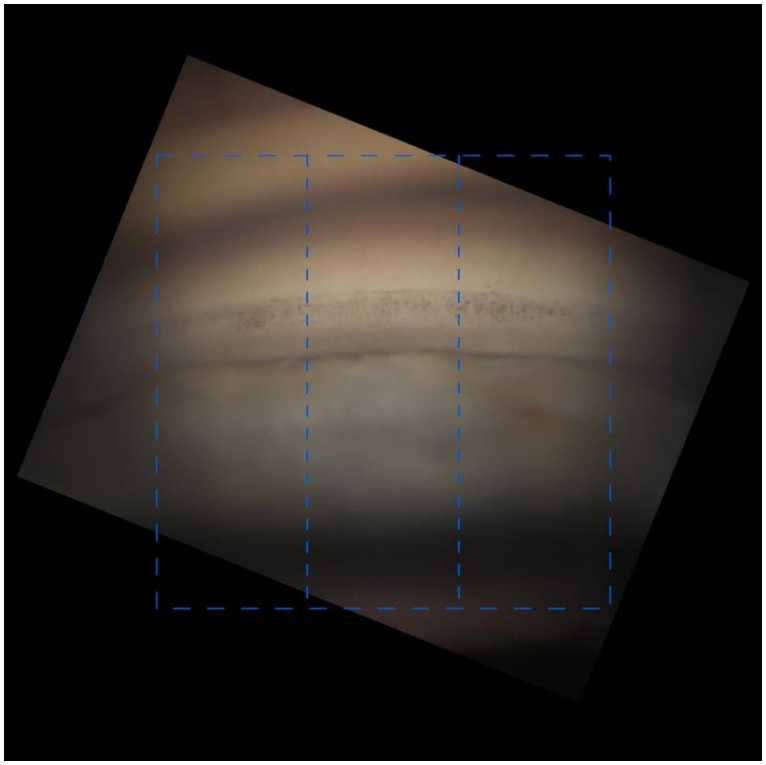

Figure 1: pre-processed sector image; the three sub-sectors of the region to annotate are highlighted by the blue dashed lines.

For each of these sub-sectors, the annotator is asked to grade two anatomical features: the local *aperture of the irido-corneal angle* and the *pigmentation of the trabecular meshwork* (if visible).

2.1 Angle Aperture

The local aperture of each irido-corneal angle sub-sector of a given GS-1 image shall be graded according to the apparent iris insertion. Three classes are defined based on the visible angle structures. An additional class is used for un-gradable sub-sectors.

- *Open*: scleral spur and / or ciliary body band visible (D, E)

- *Occludable*: Schwalbe's line visible, trabecular meshwork visible to some extent (either posterior or anterior), scleral spur and ciliary body band not visible (C)
- *Closed*: Trabecular meshwork not visible. Schwalbe's line can be visible or not (A, B)
- *Unknown*: the angle is not visible due to misalignment, because its view is prevented from obstacles (e.g. bubbles, eye lashes), or the quality of the sub-sector (e.g. sharpness and / or illumination) is not good enough to evaluate the structures.

Annotations shall be performed by selecting the most appropriate class of each sub-sector of a given GS-1 image from a combo box menu in the annotation tool. All the sub-sectors of a given GS-1 image must be assigned one of the classes described above.

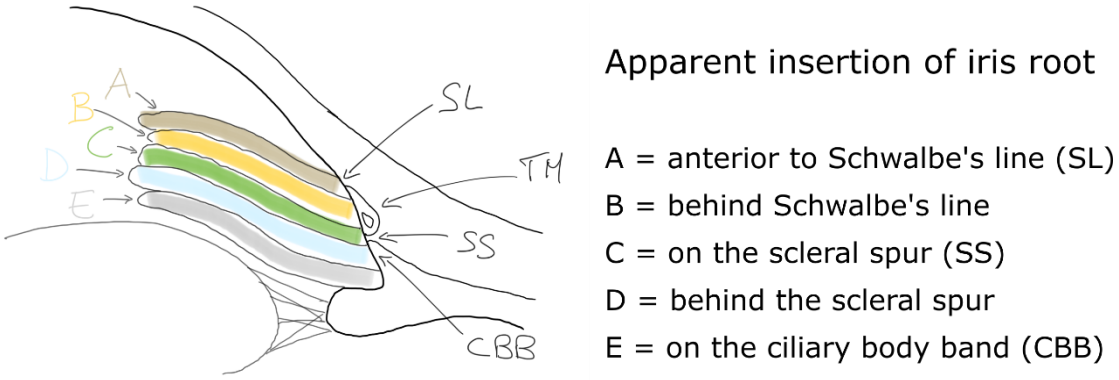

Figure 2:The Spaeth Grading System of gonioscopic finding.

In case the angle aperture varies in the considered sub-sector, the classification that better describes it (that is, applicable to the most part of it) shall be chosen. In case it is not possible to assess which classification is predominant in the sub-sector, the one corresponding to the more anterior iris insertion shall be chosen (e.g., if part sub-sector is *Open* and part *Occludable*, the classification shall be *Occludable*; if some is *Occludable* and some *Closed*, the classification shall be Closed). If a sub-sector is in part gradable (*Open*, *Occludable* or *Closed*) and in part un-gradable (*Unknown*) the classification shall be that of the gradable part.

2.2 Trabecular Meshwork Pigmentation

The local trabecular meshwork pigmentation of each irido-corneal angle sub-sector of a given GS-1 image shall be graded according to four categories including a simplified three-class Scheie's pigmentation scale and an additional class for un-gradable cases, being the classes defined as:

- *Level 1*: absent-to-low pigmentation (Scheie's classes None and 1)
- *Level 2*: mid pigmentation (Scheie's class 2)
- *Level 3*: high-to-very-high pigmentation (Scheie's classes 3 and 4)
- *Unknown*: the trabecular meshwork is not visible (e.g. in angle closure sub-sectors)

The original Scheie's scale for the grading of trabecular meshwork pigmentation is represented in Figure 3.

Annotations shall be performed by selecting the most appropriate class from a combo box menu in the annotation tool. All the sub-sectors of a given GS-1 image must be assigned one of the classes described above.

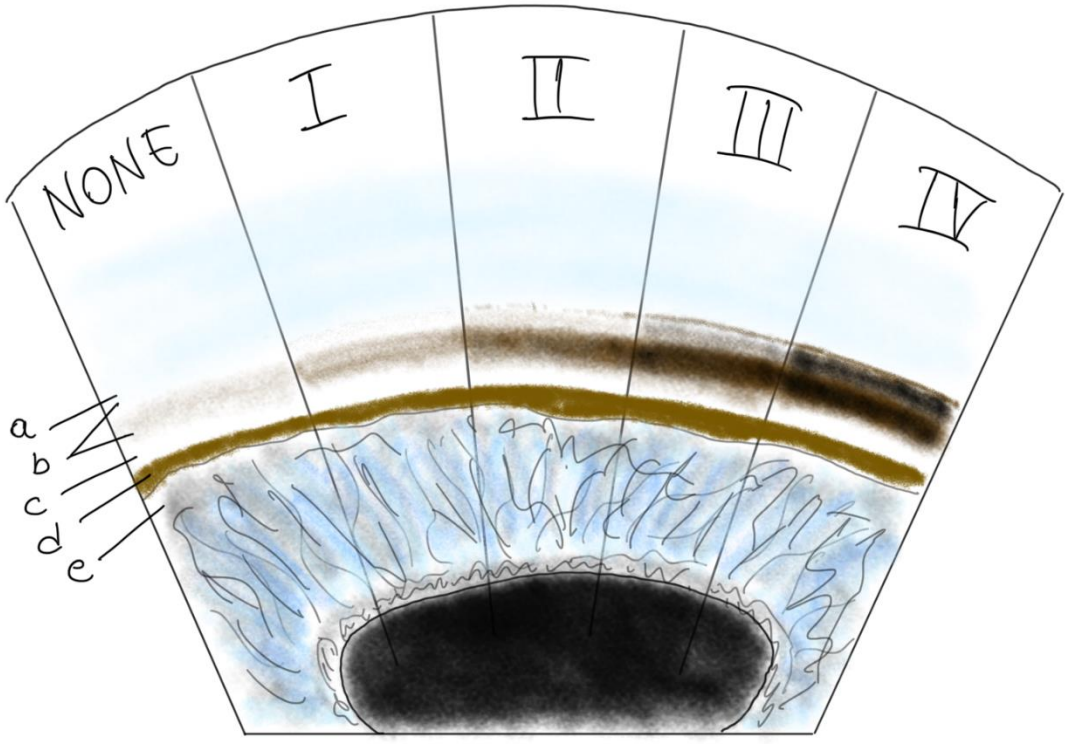

Figure 3: Scheie's system of grading trabecular meshwork pigmentation. a, Schwalbe' s line; b, anterior and posterior trabecular meshwork; c, scleral spur; d, ciliary body band; e, iris root.

In case the pigmentation varies in the considered sub-sector, the classification that better describes it (that is, applicable to the most part of it) shall be chosen. In case it is not possible to assess which classification is predominant in the sub-sector, the one corresponding to the highest Scheie's grade shall be chosen (e.g., if part sub-sector is *Level 1* and part *Level 2*, the classification shall be *Level 2*; if some is *Level 2* and some *Level 3*, the classification shall be *Level 3*). If a sub-sector is in part gradable (*Level 1*, *Level 2* or *Level 3*) and in part un-gradable (*Unknown*) the classification shall be that of the gradable part.

3 Annotation Tool and Protocol

3.1 Annotation Process Flowchart

The annotation process flowchart is reported in Figure 4.

The flowchart is structured so that *actions* (blue rectangles) and *conditions* (green rhombuses) are well identifiable.

In particular, the first condition ("Exam Annotation Completed?") may lead to the opening of a new exam to annotate / check it, or to the end of the annotation session even if the current exam annotation has not been completed yet.

Next to each action rectangle, the reference to the document section describing said action in detail is reported.

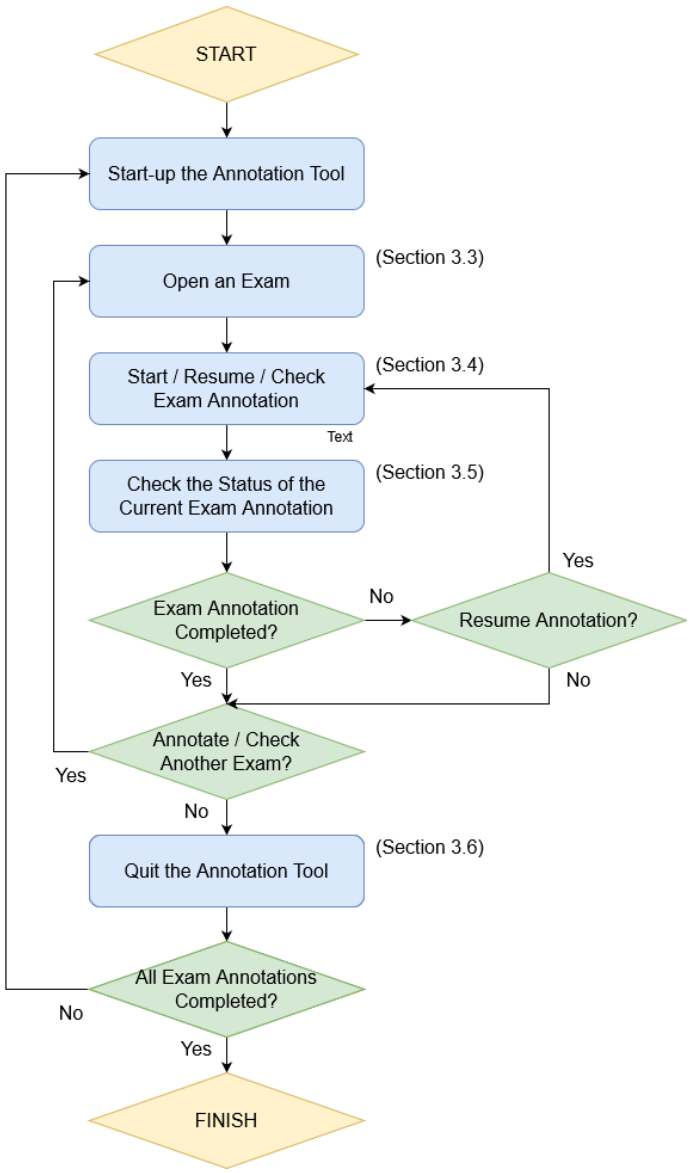

Figure 4: Annotation process flowchart

3.2 Annotation Tool User Interface

The user interface of the main page of the Annotation Tool is shown in Figure 5.

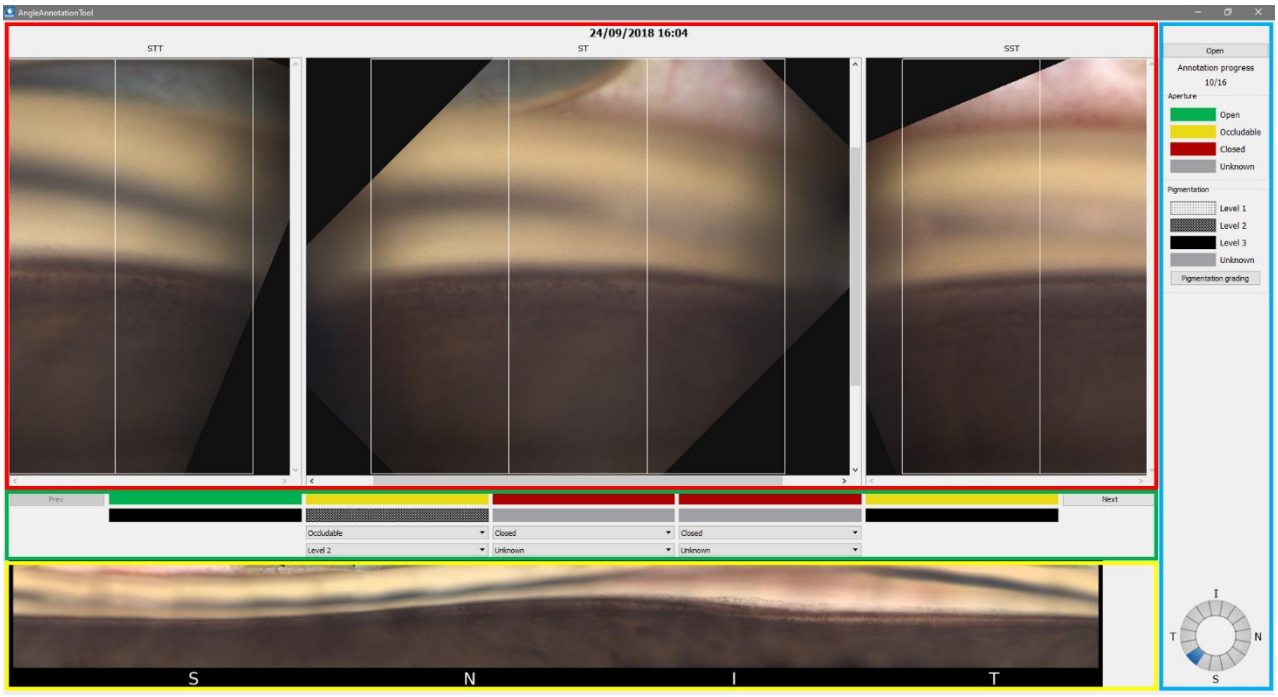

Figure 5: Annotation Tool user interface - main page. (Aperture and pigmentation classes in the image have been assigned randomly and may not be clinically correct)

The interface may be conveniently divided into 4 panels characterized by their location and functions, highlighted in Figure 5 using rectangles of different colours.

GS-1 sector view panel (red rectangle in Figure 5). It shows the following information:

- Date-time of the currently open exam
- Three views representing adjacent angle sectors of the currently open exam. All the sectors are pre-processed to uniform the order and orientation of angle layers in the frames (iris at the bottom and cornea at the top) and to highlight a region of interest for each sector.
  The middle view shows the entire 960x960 pixels ROI of the angle sector to annotate and highlights the three sub-sector boundaries. The user can zoom in/out to better visualize features of interest and pan horizontally and vertically.
  The view on the left shows the right-most half of the ROI for the preceding image in the sequence, with its right-most sub-sector highlighted; the view on the right shows the left-most half of the ROI for the following image in the sequence, with its left-most sub-sector highlighted; the user can not interact with these images.
- Initial(s) of the sector location on the irido-corneal interface circumference.

Annotation panel (green rectangle in Figure 5; reported in detail in Figure 6).

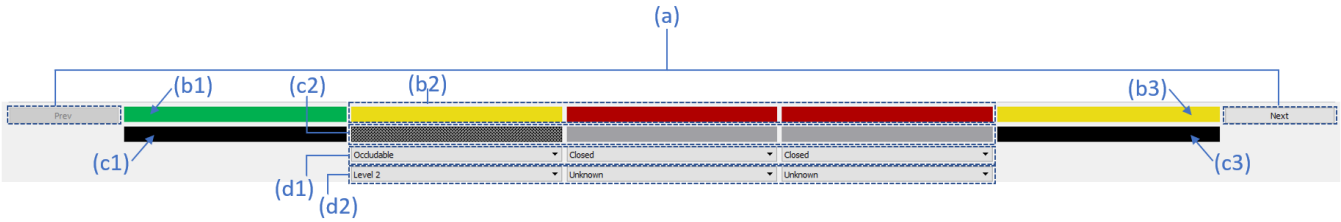

Figure 6: Annotation panel. (Aperture and pigmentation classes in the image have been assigned randomly and may not be clinically correct)

It shows the following information:

- (a): "Prev" and "Next" buttons to move back and forth along the sequence of sectors to annotate for the currently open GS-1 exam. "Prev" ("Next") button is disabled when the left-most (right-most) sector of the sequence is reached (according to linear stitching representation).
- (b1): the aperture classification of the right-most sub-sector of the preceding sector in the sequence (if already annotated, empty otherwise) [1] .
- (b2): the aperture classifications of the three sub-sectors of the current sector in the sequence (if already annotated, empty otherwise) [1], with the left-most classification corresponding to the left-most sub-sector, the central classification to the central sub-sector, the right-most classification to the right-most sub-sector.
- (b3): the aperture classification of the left-most sub-sector of the following image in the sequence (if already annotated, empty otherwise) [1] .
- (c1): the trabecular meshwork pigmentation classification of the right-most sub-sector of the preceding sector in the sequence (if already annotated, empty otherwise) [2] .
- (c2): the trabecular meshwork pigmentation classifications of the three sub-sectors of the current sector in the sequence (if already annotated, empty otherwise) [2] , with the left-most classification corresponding to the left-most sub-sector, the central classification to the central sub-sector, the right-most classification to the right-most sub-sector.
- (c3): the trabecular meshwork pigmentation classification of the left-most sub-sector of the following sector in the sequence (if already annotated, empty otherwise) [2] .
- (d1): the three combo box menus for selecting the aperture class for the three sub-sectors of the current sector image according to the guidelines provided in Section 2.1.
- (d2): the three combo box menus for selecting the trabecular meshwork pigmentation class for the three sub-sectors of the current sector image according to the guidelines provided in Section 2.2.

Side menu: this is the one highlighted by the blue rectangle in Figure 5. From top to bottom, it shows the following information:

- *Open* exam button: when pressed, an additional floating window appears (Figure 7) showing the complete list of exams date-times, together with their annotation status reporting the

[1] Colour coded according to the aperture legend reported in the side menu.
[2] Pattern coded according to the trabecular meshwork pigmentation legend reported in the side menu.

annotation status (*Absent*, *Partial*, *Complete*, see Section 3.3). The user can select an exam and open it by pressing "OK", or go back to the main Annotation Tool page by pressing "Cancel". If the listed exams are too many to fit in the window, the user can scroll the list up/downwards. While this floating window is open, interaction with the main Annotation Tool page on the background is disabled.

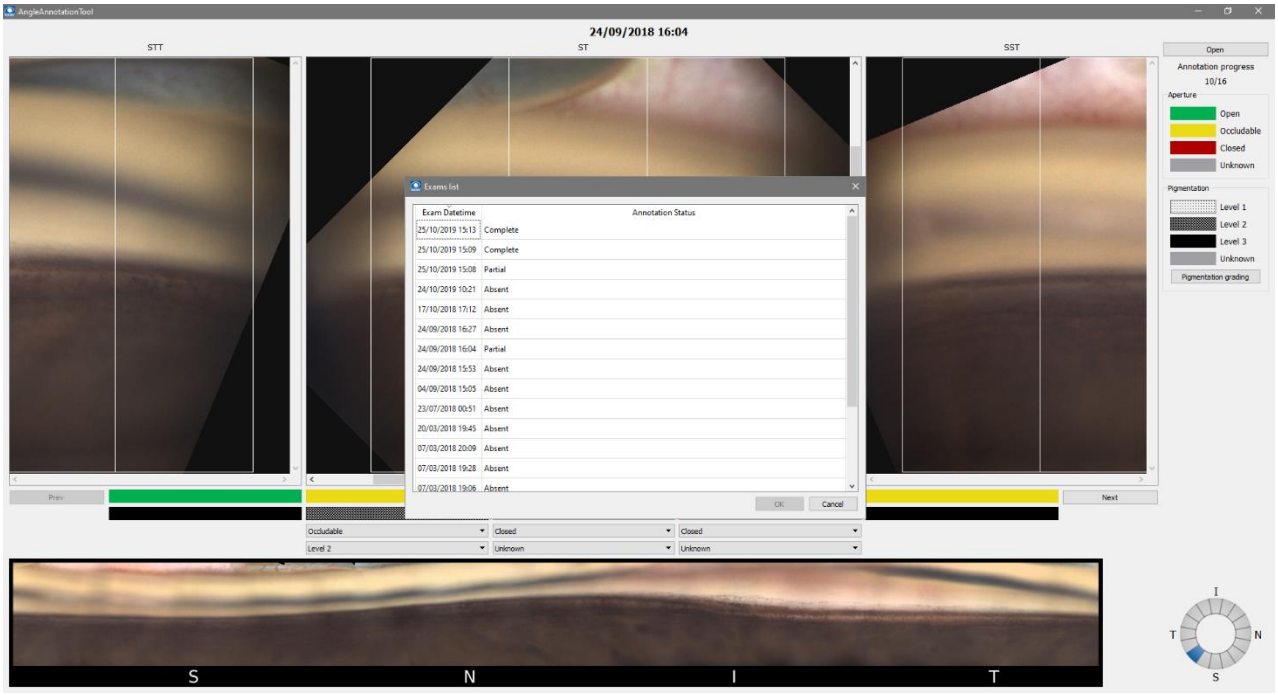

Figure 7: Open exam floating window. (Aperture and pigmentation classes in the image have been assigned randomly and may not be clinically correct)

- *Annotation progress* counter: the fraction of sectors of the currently open exam that have been fully annotated. A sector is considered fully annotated only when both the aperture and pigmentation class have been selected for all its sub-sectors.
- *Aperture* legend: it links the colour codes to the corresponding aperture classes, defined in Section 2.1.
- *Pigmentation* legend: it links the pattern codes to the corresponding trabecular meshwork pigmentation classes, defined in Section 2.2.
- *Pigmentation grading* button: when pressed, an additional floating window appears (Figure 8) showing several image patches as reference for different trabecular meshwork pigmentation grades, according to their definition in Section 2.2. When this floating window is open, the user can still interact with the main Annotation Tool page.

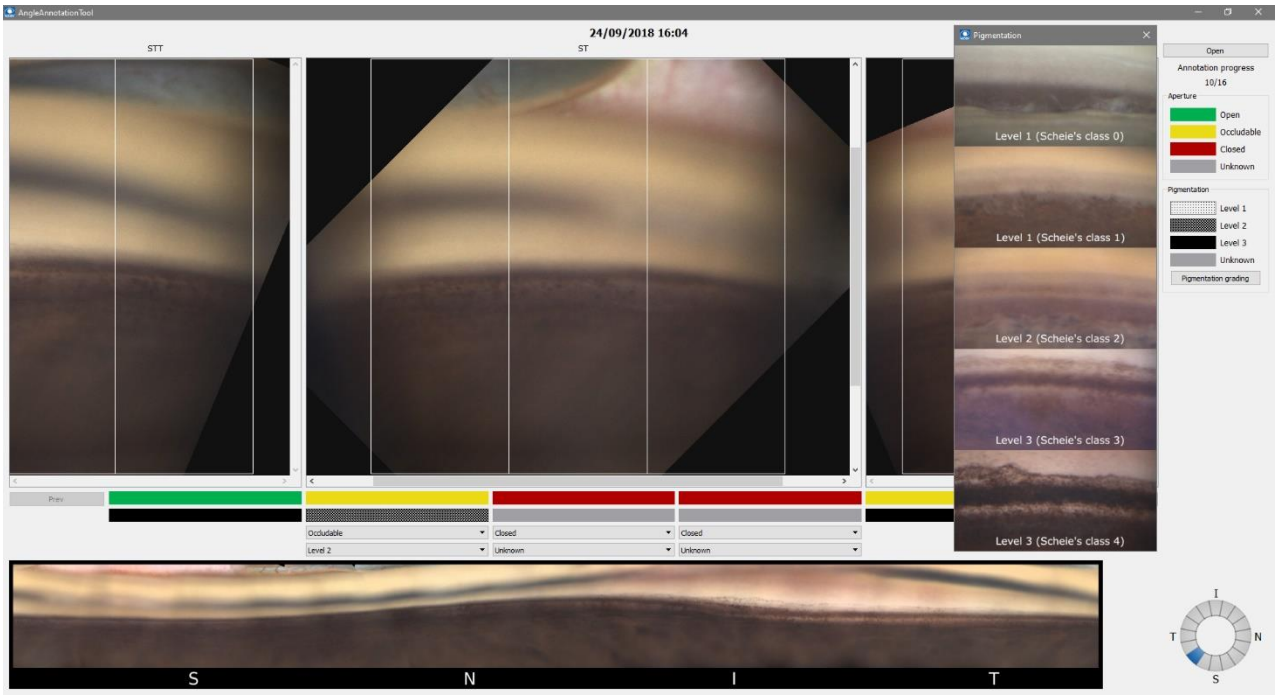

Figure 8: Pigmentation grading reference window. (Aperture and pigmentation classes in the image have been assigned randomly and may not be clinically correct)

- *Selected facet indicator*: it shows the location of the current sector image on the irido-corneal angle circumference.

Stitching panel: the one highlighted by the yellow rectangle in Figure 5. It shows the linear stitching of the currently open exam. The stitching can be used to quickly inspect the overall patient's condition. It is not possible to interact (e.g. to zoom or pan) with this panel.

Note: on start-up, the Annotation Tool does not show any exam, all fields are empty and all commands are disabled except for the "Open" button to select an exam to annotate.

3.3 Opening a GS-1 Exam

Opening an exam is possible at any moment during the annotation process.

By pressing the "Open" button from the side menu, a new floating window appears (Figure 7) showing the list of all exams to be annotated together with an indication of their annotation status:

- *absent* means that no aperture or pigmentation class has been assigned to any sub-sector of the exam;
- *partial* means that aperture or pigmentation class has been assigned to at least one sub-sector of the exam;
- *complete* means that all exam sub-sectors have been assigned both aperture and pigmentation classes.

After selecting an exam from the list, select “Ok” to open it. The Annotation Tool now shows the first sector image in the middle view and all the commands are enabled. To go back to the main Annotation Tool page without opening a new exam, press “Cancel”.

Despite the status, any exam can be selected and opened to start, resume, amend or just check its annotations.

Note: if a new exam is opened while another exam was being annotated, all the performed annotations for the previous exam are automatically saved and the new exam is opened. It is not possible to open more than one exam at the same time.

3.4 Performing Annotations

Once an exam has been opened, the first sector image (corresponding to the left-most part of the linear stitching representation) is shown in the middle view and all the commands are enabled.

The user can assign aperture and pigmentation classes to a given image sub-sector using the corresponding combo box menu located below the middle sector image. To do so, click on the combo box to show the available classes and then select the most appropriate one.

In Figure 9 the correspondence between combo box menus and sub-sectors of the image in the central view is shown.

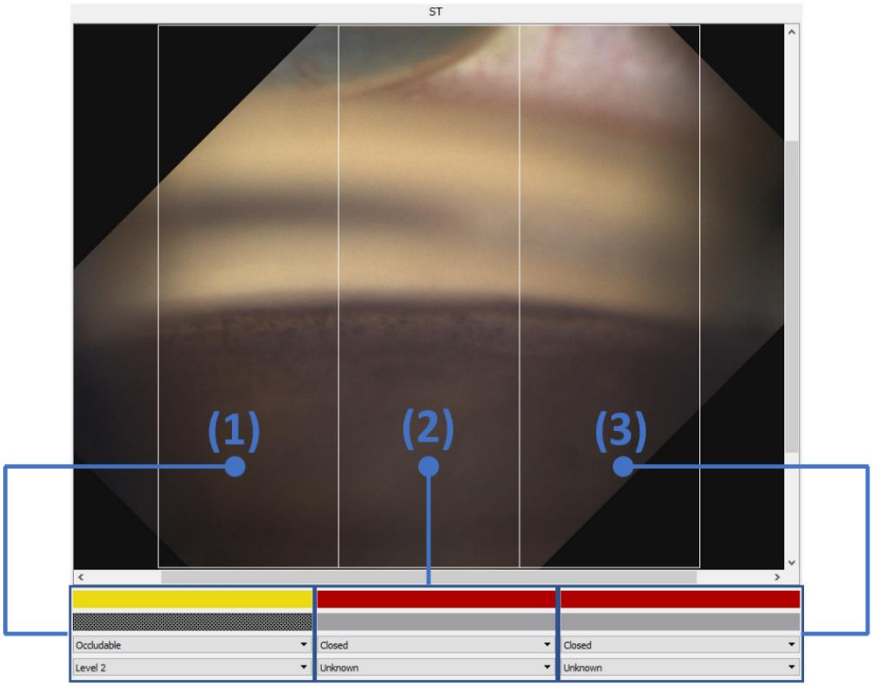

Figure 9: correspondences between combo box menus (and the corresponding annotation values) and the sub-sectors in the central view. (Aperture and pigmentation classes in the image have been assigned randomly and may not be clinically correct)

The user can also interact with the middle view by zooming in/out and panning horizontally and vertically. To go back to the original field of view, double-click anywhere on the middle view; the 960x960 pixels sector ROI will automatically fit into the view.

Note: be careful to consider that the correspondence between combo box menus and image subsectors can be affected by the interaction with the image (by zooming or panning).

The user can select or amend aperture and pigmentation classes for any of the exam sub-sectors at any moment and in any order. However, annotating both aperture and pigmentation sequentially from left to right enables handy shortcuts.

By annotating an exam sequentially from left to right, the user can exploit the expected correlation between adjacent sectors and propagate the already performed annotations (either aperture or pigmentation or both) to the right.

This works for both aperture and pigmentation annotations independently and is possible in two case scenarios only, here explained:

- the right-most sub-sector of the previous sector image in the sequence (left view) has been annotated and all three the sub-sectors of the current sector image in the sequence (central view) have not yet. By right-clicking on the annotation for the right-most sub-sector of the previous sector image in the sequence (labelled as *b1* or *c1* in the "Annotation panel" paragraphs of Section 3.2), the user can select the *"Copy aperture"* or *"Copy pigmentation"* command to copy the annotation to all three the sub-sectors of the current sector image in the sequence;
- the left-most sub-sector of the current sector image in the sequence (central view) has been annotated and the other two sub-sectors of the current sector image in the sequence (central view) have not yet. By right-clicking on the annotation for the left-most sub-sector of the current sector image in the sequence (the left-most rectangles among those labelled as *b2* or *c2* in the "Annotation panel" paragraphs of Section 3.2), the user can select the *"Copy aperture"* or *"Copy pigmentation"* command to copy the annotation to the other two sub-sectors of the current sector image in the sequence.

An example of *"Copy aperture* / *Copy pigmentation"* context menu is shown in Figure 10.

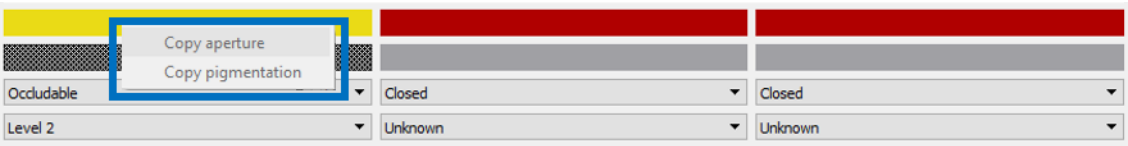

Figure 10: sub-set of Annotation Panel commands for the current sector image in the sequence with the "Copy aperture / Copy pigmentation" context menu highlighted by the blue rectangle.

In fact, aperture and trabecular meshwork pigmentation are features that usually vary slowly or do not change at all in an exam. These two commands can be used to propagate the right-most sub-sector annotation of the preceding sector or the first sub-sector annotation of the current sector to

all the sub-sectors of the current sector. Once propagated, annotations can be amended, e.g. to account for local variations of angle aperture.

The user can inspect the linear stitching to have a general idea of the patient's condition, but all the annotations shall be based on the images highlighting the sub-sectors of interest, located above the combo box menus.

Note: *by design of the GS-1 acquisition process, the outermost sub-sectors of adjacent sector images are partially overlapped. The annotation could be, however, different based on local image characteristics.*

The user shall always keep in mind the definitions of aperture and pigmentation classes and shall use the legends in the side menu, as well as the reference image for the pigmentation grading (shown after pressing the *pigmentation grading* button), so as to make sure their annotations conform with the grading adopted for this task.

3.5 Checking the Annotation Status

Before quitting the Annotation Tool or opening a new exam, the user may want to check the annotation status for the current exam.

This is possible by simply checking the *Annotation progress* value from the side menu. If all the exam sectors have been fully annotated, meaning that all the sub-sectors have been assigned both an aperture and a pigmentation class, the reported value is 16/16. The ratio decreases by one unit for each sector of the currently open exam that has not been fully annotated yet, meaning that at least one of its sub-sectors has not been assigned either the aperture or the pigmentation class (or both).

If, for example, the *Annotation progress* ratio is 13/16, it means that three sectors of the exam have not been fully annotated yet.

3.6 Saving Annotations and Quitting the Tool

The user can quit the Annotation Tool at any moment by simply pressing the [X] button at the top-right end of the Annotation Tool window. All the performed annotations are saved before the application is terminated.

Annotations are also automatically saved every time a new exam is opened.