Utrecht University

Master's Thesis

submitted in partial fulfilment of the
requirements for the course "Sustainable Development"

Application of a Self-learning Algorithm to Analyse Microscopic Images of Stomata

Theresa Pflüger

Utrecht University
Faculty of Geosciences
Copernicus Institute of Sustainable Development

First Supervisor: Dr. Hugo Jan de Boer
Second Supervisor: Prof. dr. Friederike Wagner-Cremer

7 August 2020

Abstract

The stomata on plant leaves are crucial for gas exchange and, when observed on (sub)fossil leaves, also provide insight into historic plant growth conditions such as atmospheric CO_2 and humidity. Current methods to quantify stomatal traits rely on manual analysis of microscopic images, which is labour-intensive and requires expert knowledge. The use of computer-aided methods has the advantage to increase efficiency in data acquisition due to time-savings as the sample throughput is higher in a shorter time and thereby the potential of more reproducible results. This study applied a machine-learning approach developed by Jayakody, Liu, Whitty, & Petrie (2017). The aim was to test the effectiveness of the automated detection method for identifying stomata in microscopic images and its sensitivity towards stomata size and image quality. Two major experiments have been conducted using the originally introduced image dataset of grapevines from Jayakody et al. (2017), and an additional set of images from different plant types consisting of ferns and grasses, since specifically grasses feature a more complex stomata type. The outcome was compared to results by Jayakody et al. (2017) and manual counts of the fern and grass images. The images of the original dataset were manipulated by undergoing treatments of enlargement and reduction to mimic stomata size differences. For testing the influence of image quality, the original images were downsampled to achieve quality loss, while the fern and grass image dataset was classified into quality categories based on the perceived image quality by a human viewer. Additionally, comparisons between the detection accuracy of the different stomata types were carried out. It has been found that the algorithm reaches a limit in respect to stomata size leading to greater numbers of missed stomata indicating a lower tolerance towards variations in stomata size. In contrast, the method shows a high level of robustness in terms of image quality for the downsampled images generating a low number of incorrect detections. Also, the image quality of the fern and grass images does not seem to have a significant influence on the effectiveness of detecting stomata. In terms of stomata type, the algorithm handled both types well and no significant difference in accuracy has been found.

Keywords: Stomata, Microscopic imagery, Automatic stomata detection, Machine learning, Stomata counting

Acknowledgements

I would like to thank Hugo de Boer for his encouraging supervision, support and guidance throughout my thesis work during this difficult time and exceptional circumstances. I am thankful to him for taking out time to discuss the problems regarding the approach of the thesis and the set-up of my experiments while providing his valuable suggestions.

I am grateful to Hiranya Jayakody who did not only develop the foundation of this project but supported me through generous and clarifying feedback regarding certain issues about the methodology.

Also, I would like to thank everyone I met during my lab work who created a productive yet relaxed atmosphere. They were excited about my work and showed their interest and support. I experienced a friendly and warm working atmosphere in the lab.

Last but not the least, I am thankful to Martin Smit who took the time to help me collect plant material from Hortus Botanicus Amsterdam. He too showed great interest and excitement about my project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

When plants left their aquatic environment and colonised terrestrial areas around 450 million years ago (Willis, 2014), they had to adapt and develop structures that helped them to survive on land. They had to make sure that there is adequate water supply – without excessive water loss – while making it possible to take up carbon dioxide (CO_2) to be able to perform vital photosynthesis. This is assumed to be the birth of the cuticle, stomata and vascular tissue of terrestrial plants. Stomata have been found on plant fossils of the age of more than 400 million years (Vatén & Bergmann, 2012). Stomata are pores on the surface of the plant's aerial structures and their main function is to facilitate gas exchange between the plant interior and the atmosphere (Willmer & Fricker, 1996a).

1.1 The importance and functions of stomata

To highlight the importance of stomata research, key functions of stomata are elucidated here. Understanding stomata, their characteristics and behaviour plays an important role in predicting the health of plants (e.g. Bhugra et al., 2018; Jayakody, Liu, Whitty, & Petrie, 2017) and therefore their capacity to affect the carbon cycle of the planet.

Stomata fulfil a variety of key functions for plants to be able to survive in the terrestrial environment. The capability of stomatal opening and closure goes beyond the mere facilitation of gas exchange between plants and the atmosphere involving CO_2, O_2 and water vapour required for photosynthesis. Restriction of water loss from the plant's tissue delaying desiccation on terrestrial grounds while optimising the carbon gain per unit water loss are one of the other crucial functions. Especially the capacity to limit water loss is crucial for plants to thrive in environments with fluctuating water supply and is assumed to be one of the earliest functions of stomata. Furthermore, possible transpiration through stomata and related transport of solutes across the plant leading to increased delivery of important nutrients to various sites in the shoot (Raven, 2002) is an essential role involving stomata.

Other selective pressures on stomata are the variations in atmospheric CO_2 concentration (C_a) and changes in moisture. Hence, the need to adapt and increase the conductance to CO_2 diffusion to meet the desired CO_2 flux, and regulating the balance between water transpired and

CO_2 fixed by enhancing transpiration efficiency (Franks & Farquhar, 2007). In respect to environmental change and a changing climate, the first stomata played a key role in enabling terrestrial vegetation and changing the climate thereof (Berry, Beerling, & Franks, 2010). Through their capacity to regulate the pore's aperture for water and CO_2 fluxes to occur, stomata are having a significant influence on the global water and carbon cycles (e.g. Hetherington & Woodward, 2003). Early climate change took place due to the feedback between the increased transpiration and rise in precipitation over land and thus leading to the expansion of climate zones facilitating the spread of plants (Berry et al., 2010).

1.2 Stomatal characteristics

Stomata being key features to survive terrestrial conditions have undergone changes throughout evolutionary time creating a variety in stomata morphology likely due to different divergence times and evolutionary pressure. Also, between distinct clades such as ferns, gymnosperms and angiosperms there are stomatal differences; and even amongst different angiosperms with grasses being outstanding. Stomata morphology is characterised by varying shapes, sizes and distributions (i.e. densities) (e.g. Franks & Beerling, 2009; Vaten & Bergmann, 2012; Willmer & Fricker, 1996b). While the specific stomata anatomy of the individual stomatal complex remained mostly unaltered over geologic time, changes occurred in respect to the number of stomata on the leaf epidermis as well as the pore size and stomatal density (e.g. Beerling & Franks, 2009; de Boer, Eppinga, Wassen, & Dekker, 2012; Franks & Beerling, 2009; Vaten & Bergmann, 2012). For example, the same valve mechanism of turgor pressure is operating in stomata for the last hundreds of millions of years (Berry et al., 2010). With the emergence of different species and with that accompanied variety in stomatal morphology, environmental change played a role in this stomatal development. It is known that stomatal changes are correlated to the atmosphere's C_a with increasing stomatal density while C_a is decreasing (Franks et al., 2013). The reason for developing smaller stomata over time is shown to be of positive influence on the carbon uptake under low C_a while leading to a rise in productivity. Additionally, smaller stomata allow more of the leaf's epidermal space for other cell types and functions such as subsidiary cells or oil cells (de Boer et al., 2012; Franks & Beerling, 2009) which are important adaptations. Being able to conduct stomata analysis in an efficient manner while including a variety of plants, i.e. stomata types and sizes, is crucial when it comes to gaining new insights. New and advanced methods of analysis using artificial intelligence such as machine learning, for example, can support the research (see also Section 1.3).

Stomata types

In respect to stomata morphology, there are two basic types: the elliptical type (kidney-shaped) and the graminaceous type (dumbbell-shaped, typical for grasses) (Jones, 2014; Willmer & Fricker, 1996b). Particularly grass stomata are important to be included in studies on stomatal morphology since they feature different characteristics such as distinctive dumbbell-shaped guard cells arranged in rows with parallel running subsidiary cells next to each guard cell unlike, for instance, fern stomata, which feature the other type of kidney-shaped stomata (Franks & Farquhar, 2007; Jones, 2014). It was of interest to test an automated detection method on both stomata types.

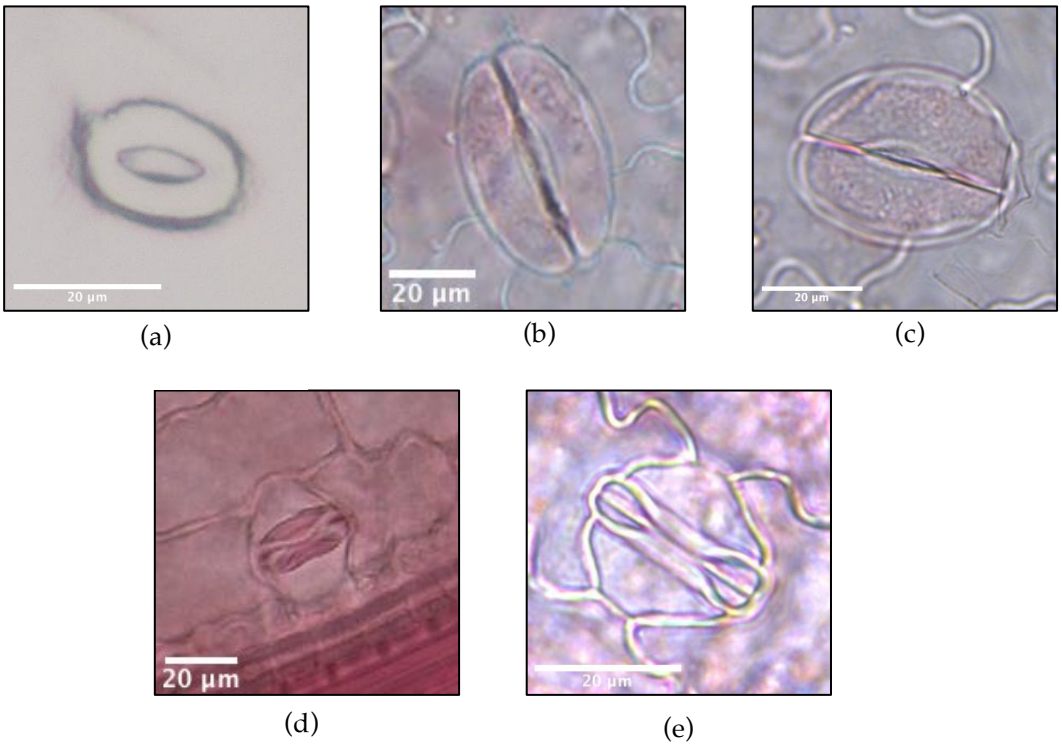


Figure 1.1: Examples of stomata from different plant types showing their distinct shape. **a** Sample stoma from the image dataset by Jayakody et al. (2017). **b** and **c** Fern stomata at 20× and 63× magnification, respectively. **d** and **e** Grass stomata at 20× and 63× magnification, respectively.

1.3 Stomata images-based research

Since stomata are microscopic structures on plant leaves, stomata research depends microscopic images for studying their behaviour and carrying out measurements. Ongoing research has contributed to the investigation of stomatal traits of various plant types by using microscopic images while manually counting and measuring stomata (e.g. Casado-García, Heras, &

Sanz-Saéz, 2020; de Boer et al., 2016; Fetter, Eberhardt, Barclay, Wing, & Keller, 2019; Jayakody et al., 2017; Laga, Shahinnia, & Fleury, 2014). However, this process is expensive as a single image of the plant epidermis may contain dozens of stomata. Thus, automated computer-aided methods to detect and count stomata may offer benefits such as saving time compared to manual counting, leading to an increase in efficiency in data acquisition and the subsequent analysis (see Fetter et al., 2019; Jayakody et al., 2017).

Developments involving algorithms to accelerate the resource-intensive process have been made. Previous research applied the problem of stomata detection in microscopic images such as Laga et al. (2014) and Liu, Tang, Petrie, and Whitty (2016) who both used samples of only one species (i.e. wheat and grapevine, respectively). Jayakody et al. (2017) also only tested their automated stomata detection method on only one species (grapevines). Thus, this research aimed to add a variety of species for testing the algorithm developed by Jayakody et al. (2017).

1.3.1 Theoretical groundwork

At the basis of these aforementioned developments is the theoretical groundwork. The algorithms which form the foundation of each method for the automation of object detection in images were established by previous researchers from different fields.

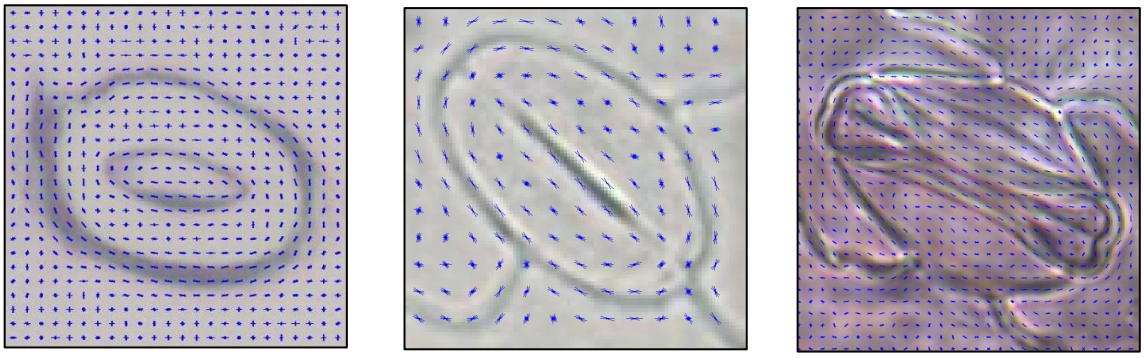
The algorithm on which the customised classifier by Jayakody et al. (2017) is based on is the so-called Viola-Jones algorithm (Viola & Jones, 2001). This algorithm was initially developed to detect faces but was re-trained for the purpose of detecting stomata using a different feature type as learning descriptor: histogram of oriented gradients (HOG) instead of Haar-like wavelet features (Jayakody et al., 2017; Viola & Jones, 2001). The implementation of HOG descriptors for machine-learning algorithms was pioneered by Dalal & Triggs (2005). HOG features perform well when it comes to the detection involving a shape-based object classification where the shape of the object of interest occurs in many variations such as direction, for example, since HOG capture the overall shape of the object (Dalal & Triggs, 2005; The MathWorks, 2020). A visual explanation of HOG features based on the image dataset used in this research can be found in Section 1.3.2.

1.3.2 Automated stomata detection

Below, the basic workflow of appearance-based object detection involved in identifying stomata in microscopic images is briefly outlined.

Firstly, a dataset of images needs to be acquired which is representative of the data that is subject to testing at a later stage. Secondly, the algorithm needs to be trained on a set of images which is similar to the test images. One set features the object of interest (i.e. single or multiple stomata) and the other set must not contain any stomata but be of similar make-up as the first dataset (i.e. microscopic images of the leaf's cuticle with veins, air bubbles, dust particles, etc.). Next, manual annotation of the regions containing the object of interest indicates the positions of objects of interest in these training images. During manual labelling it is important to capture variations in object appearance to create a representative dataset. The training leads to the algorithm learning to distinguish between positive and negative samples on which the object detection is based on. The final stage is the evaluation of the object detection using metrics such as precision and recall and the *F*-measure. These are commonly used metrics to evaluate results of image analyses (e.g. Casado-García et al., 2020; Tharwat, 2018). Details on the exact method of stomata imaging, object detection and the training process tailored to this case, as well as the evaluation of the algorithm's performance can be found in Section 2.4 and 2.6.

The algorithm used here implementing a Cascade Object Detector was developed by Jayakody et al., 2017 using the program Matlab®. For this research, Matlab® R2019b was used. Training the classifier involves a feature type on which the detection is based on. Jayakody et al. (2017) makes use of the feature type of histogram of oriented gradients (HOG) and thus building a HOG descriptor that, after training, is able to detect stomata as regions of interest (ROI) based on HOG features (see Figure 1.2) in a microscopic image.



(a) (b) (c)

Figure 1.2: Examples of HOG feature visualisation for positive samples from each dataset. **a** Grapevine stoma (from original dataset by Jayakody et al. (2017). **b** Fern stoma. **c** Grass stoma.

1.4 Scientific relevance

As mentioned previously, stomata research is an important part in determining the vegetation's influence on the carbon cycle. This is especially crucial in a changing climate affected by rising CO_2 levels due to the anthropogenic impact. Thus, stomata research has a predictive power in respect to future developments. Yet, paleoecologists and paleoclimatologists are also interested stomatal characteristics. This is due to the stomata's function in atmospheric gas exchange. Stomatal density measurements from fossil plants can act as an indicator of paleoatmospheric CO_2 concentration and thus serving as a prediction tool for reconstructing paleoclimates (Fetter et al., 2019). Such measurements can be accelerated by making use of computer-based methods as it has the potential to decrease resource-intensive and human labour and the need for expert knowledge in identifying stomata in microscopic images (Casado-García et al., 2020; Jayakody et al., 2017). Additionally, when measuring cell properties based on micrographs it is crucial to (spatially) calibrate the image to obtain correct measurements in real units (e.g. µm). This intermediate step could also be avoided by setting up the automated measurement method beforehand.

The application of computer-based analysis is getting increasingly widespread attention due to time-savings and the ability to process greater amounts of data. Here, the automated detection method is expected to analyse microscopic images and detect objects of interest in a faster manner than a human would be able to. Thus, automated methods improve the efficiency in data acquisition as the sample throughput is higher in a shorter amount of time (e.g. Bhugra et al., 2018; Fetter et al., 2019) and can therefore advance research and allow the focus to be on the interpretation of results provided by the automated analysis and in turn leading to possible quality improvements in that regard. Additionally, there is a greater potential of more reproducible results as the method has already been established (Bhugra et al., 2018). Testing the performance of such a method is crucial to identify limitations and the necessity of improvement to achieve a better applicability, also to being able to implement the method in various fields of research like, for instance, detecting and identifying pollen grains from microscopic images. Similarly, testing on a broad range of input data (test data) with varying conditions is essential to understand the requirements the data needs to fulfil for the method to be successful, i.e. image quality requirements, but also regarding the most suitable method of preparing samples as different methods generate varying results not only in terms of quality but also regarding the time investment (e.g. epidermal leaf impressions or bleached leaf cuticles) (see also Yuan et al., 2020).

In terms of having large sample sizes of images, the problem of storing them arises as some microscopic image acquisition techniques, such as stacked image generation for instance, produce large files that take up a great amount of memory. By knowing the requirements for successfully applying automated image analysis in this case, the usage of memory can be kept to a minimum. Savings in memory also allow for obtaining a greater amount of data leading to a higher degree of completeness, as well as greater robustness in respect to results and statistical models through obtaining more data (Jayakody et al., 2017; Laga et al., 2014).

1.5 Objectives and research questions

This thesis aimed to subject the computer-aided method developed by Jayakody et al. (2017) using a self-learning algorithm to detect objects of interest (i.e. stomata) in microscopic images to various tests. Under these different conditions, the performance and limitations of the algorithm were assessed. The research project consists of two main parts:

Part I – Expansive testing of the original image dataset provided by Jayakody et al. (2017) by undertaking the images various treatments and manipulations, and
Part II – Testing the algorithm on a new sample of different plant types such as ferns and grasses.

For each of the two parts, the algorithm was tested in respect to its sensitivity towards stomata size and image quality. Firstly, size differences were mimicked by manipulating the image (i.e. enlarging or reducing the original images), and by using microscopic images taken at different magnification levels (fern and grass images). Secondly, image quality was being defined as, on the one hand, a quantifiable criterion of image resolution (downsampling the manipulated dataset), and on the other hand, as a perceived quality by a human viewer (fern and grass images). A detailed elucidation of the different experiments can be found in Section 2.1.

Regarding part II, the stomata type plays a major role as well in testing the algorithm's performance as specifically grass stomata are differing from grapevine of the original set of images and fern images (see also Section 1.2).

Research questions
To achieve the aforementioned research objectives, the following research question is the main focus of the thesis:

Can the automated detection method be successfully applied to a variety of microscopic images featuring stomata including the more complex type of grass stomata, and differing stomata sizes as well as changes in image quality to correctly identify stomata?

And to be able to answer the main research question, these sub-questions are formulated:

1. To what extent is the algorithm sensitive towards differences in size of the object of interest (i.e. stomata) and how does this sensitivity influence detection accuracy?
2. To what extent is the algorithm sensitive when it comes to the quality of the image (i.e. differing resolution or disturbance) and how does this sensitivity affect detection accuracy?
3. To what extent does the complexity of the object of interest (i.e. distinct stomata type of grasses) influence detection accuracy?

To answer each of these research questions, various experiments and tests have been conducted with assessing the results using a common set of metrics and comparing the result to the outcome of the non-manipulated original images (part I), and an established reference for the fern and grass images (part II) See also Section 2.3.

Chapter 2

Methods

In this chapter, the exact methodology including collection of plant material, preparation of microscopic images and image acquisition, as well as statistical testing and evaluation will be described.

Sample preparation for taking microscopic images took place in the GeoLab at Utrecht University. In this research, the standard protocol for the preparation of microscope slides of leaf samples was followed. The exact methodology for this will be described in Section 2.2.2. The implementation and testing of the computer-aided method is based on the algorithm's script provided by Jayakody et al. (2017).

Firstly, to achieve part I of the research objective, a number of different test images was established by performing various manipulations on the original image dataset from Jayakody et al. (2017). The kinds of treatments are described in Section 2.1.1. Secondly, regarding part II and testing whether the algorithm can successfully detect stomata in microscopic images of different plant types, a new sample collection of ferns and grasses was obtained by preparing microscopic images. A list of all collected species can be found in Appendix A.

For both parts of the research, the application of the computer-aided method is equal. Both, training and testing on the image datasets has been carried out, followed by manual counting of the algorithm's image output of annotated detected regions of interest in the image. Statistical analysis has been performed to evaluate the algorithm's accuracy using common evaluation metrics (see Section 2.5).

2.1 Data collection

The data collection mainly involved two steps. Firstly, the extension of the original dataset by Jayakody et al. (2017) by means of manipulation of these images. And secondly, the addition of newly collected microscopic images covering other plant types that have not yet been tested as input for the self-learning algorithm (i.e. ferns and grasses).

2.1.1 Manipulation of original image dataset

As mentioned in Section 1.5, testing the algorithm's performance under various conditions involved the manipulation of the original image dataset. To cover a wide range of possibilities and to prompt the algorithm's performance and accuracy, a number of different experiments were conducted which are given below in Table 2.1.

Table 2.1: Experiments for manipulating the original image dataset by Jayakody et al. (2017) by means of enlarging, reducing or downsampling (changing the resolution) the image. The percentage refers to the size in relation to the original image with corresponding pixel dimensions where 4800×3600 being the original pixel dimensions. Downsampling was done leading to the respective percentage of the original image with resizing to the original size afterwards.

Enlargement	**Reduction**	**Downsampling**
150 (72005400)	75 (36002700)	24001800 (50)
200 (96007200)	60 (28802160)	1200900 (25)
250 (120009000)	50 (24001800)	600450 (12.5)
300 (1440010800)	40 (19201440)	480360 (10)
400 (1920014400)	30 (14401080)	240180 (5)
		144108 (3)
		192144 (4)
		9672 (2)

2.1.2 Addition of fern and grass images

For the purpose of testing the algorithm on a wider range of microscopic images featuring a variety of stomata types and magnification levels, in addition to the original image dataset by Jayakody et al. (2017), new microscopic images of a broad sample including ferns and grasses were obtained. This addition posed a more appropriate database for the algorithm and lead to a suitable sample size to perform statistical analysis. Collection of plant material from the botanical gardens of Utrecht University and Amsterdam was carried out, generating a sample of 14 species with five fern and 9 grass species (see Appendix A, Table 1). From all samples images were obtained which were used to train and test the automated stomata detection algorithm.

First, sample slides of the collected fern and grass leaf samples were prepared which was followed by acquiring microscopic images of these samples. The sample slides show the leaf's cuticle and stomata of a bleached leaf fragment. For more details on preparation and make-up of the samples see also Section 2.2.2. These microscopic images served as input for training and testing the automated algorithm (see Section 2.3).

2.2 Materials and preparation

The preparation of appropriate images involved two aspects: (i) the creation of manipulated images based on the original dataset, (ii) the preparation and creation of new microscopic images by means of bleaching the whole leaf or producing an epidermal impression depending on the quality of the leaf sample. The following section elucidates the process of the image preparation, specifically of the fern and grass images, and the different treatments to obtain a larger sample of test images of the original image dataset.

2.2.1 Manipulated images

As the algorithm analyses pixels in an image, manipulations were a feasible approach to examine the algorithm's performance. To mimic differences in image quality as well as size differences of stomata in microscopic images taken using various methods and at different magnifications, the manipulation of the original published dataset by Jayakody et al. (2017) was conducted. This available set of images was used as it was proven that the object detection of the self-learning algorithm is successful in relatively accurately detecting stomata. Manipulation of the original image dataset was done by using Adobe® Photoshop® CC 2018. Three different experiments were conducted on a number of 12 images from the original image dataset. These 12 images were selected based on subjective assessment regarding quality to build a relatively representative sample size for the experiments that followed. The treatments of enlargement and reduction were used to test the sensitivity to image size (i.e. stomata size) and the downsampling treatments were used testing the sensitivity to image quality.

Enlargement

With respect to the experiment of enlargement, five treatments were chosen that represent a percentage change in image size as shown above in Section 2.1.1. The automatic setting for the resampling method was selected to increase the image size in Photoshop®. These enlarged images were then used as input to the algorithm.

Reduction

For the experiment of reducing the image, five treatments were chosen that represent a percentage change in image size as specified earlier. The automatic setting for the resampling method was selected while decreasing the image size in Photoshop®. Similarly, these decreased images were then used as input to the algorithm.

Downsampling

The third experiment involved image scaling in the form of downsampling generating a smaller image from the higher-resolution image and thus leading to a decrease in image quality and lost detail in the final image (i.e. lower number of pixels). Firstly, to reduce image quality, the resampling method "bicubic sharper (reduction)" was selected to downsample the image in Photoshop® according to the aforementioned experiments. Each experiment involves to a two-step approach. The first step corresponds to a change in image size in % and, as a second step, rescaling of the images to the original pixel dimensions equal to the original images' pixel dimensions of 4800×3600. The final images served as input for the algorithm.

2.2.2 Fern and grass images

Establishing samples to acquire suitable images for the image detection, the samples of the plant material needed to be prepared. The following describes the process of microscopic sample preparation by using a bleaching technique as well as a superficial impression method. Moreover, to test the algorithm's sensitivity to quality, two quality categories were established to which the selected test images have been allocated to.

Bleached leaf samples

During lab work and the preparation of the microscope slides, as well as the acquisition of the images the following materials end equipment was used:

- Plant material: collected fern and grass leaves
- Clear nail polish
- Clear scotch tape
- Chlorine bleach 5 %
- Forceps, spatulas, slides and cover slips.
- Colouring agent safranin
- Glycerin jelly
- Microscope system Leica® DM6000 B
- Software Leica® Application Suite (version 4.12.0) for Windows®

Generally, the standard procedure to obtain suitable microscopic images of stomata from leaf samples was followed by conducting these steps:

1. Per leaf a number of 3 samples was taken.
2. Soaking the leaf sample in chlorine bleach or a solution of bleach and water.
3. Thoroughly rinsing the now white or transparent sample to wash off any bleach to stop the bleaching process.
4. Adding safranin as counterstain where necessary to aid in visibility (i.e. contrast) of the stomata of the sample.

5. When sample was appropriate, it was made permanent by mounting it in glycerin onto the slide.

To observe stomata under the microscope, it is often necessary to bleach the leaves beforehand as the thickness and the pigments in the leaf cause difficulties in viewing stomata and epidermal cells clearly (Sharma, 2017). The rate of the bleaching is dependent on several factors such as species, the freshness of the leaf or whether it was dried, the thickness of the leaf, or if the cuticle features a protective, waxy layer. Generally, dry leaves and leaves from fern species were bleached in less amount of time than grass leaves. Bleaching time varied between a few hours to two nights depending on the concentration of bleach and water as well as the above-mentioned factors. Dilution in various levels was undertaken to make the solution less aggressive keeping the bleaching process under better control. To accelerate the bleaching process, the samples in tubes with bleach were put in a block heater at a temperature between 30 °C and 50°C. The bleaching process was completed as soon as the leaf turned transparent and lost its green pigment. The leaf samples were rinsed thoroughly and were then ready to be mounted on a microscope slide. Glycerin jelly was melted, the sample mounted within and a cover slip was added. This method preserves the samples to be stored permanently.

In some cases (e.g. *Oryza sativa*, *Milium effusum L.*), the preparation of a suitable slide of a bleached leaf fragment was not possible, and an epidermal impression using nail polish was produced. For details, see the following section.

Epidermal impressions

Producing a nail polish impression of the leaf's epidermis aided to determine whether the stomata are on the abaxial (lower) or the adaxial (upper) side of the leaf's surface in a fast manner. Hereby, a thin layer of clear nail polish was painted on both sides of the epidermis. After the nail polish layer had completed dried, the tape was applied to carefully peel off the nail polish film which was sticking to the tape. This was then be pressed tightly onto a microscope slide and thus the permanent leaf epidermis imprint has been generated from which an image can be taken.

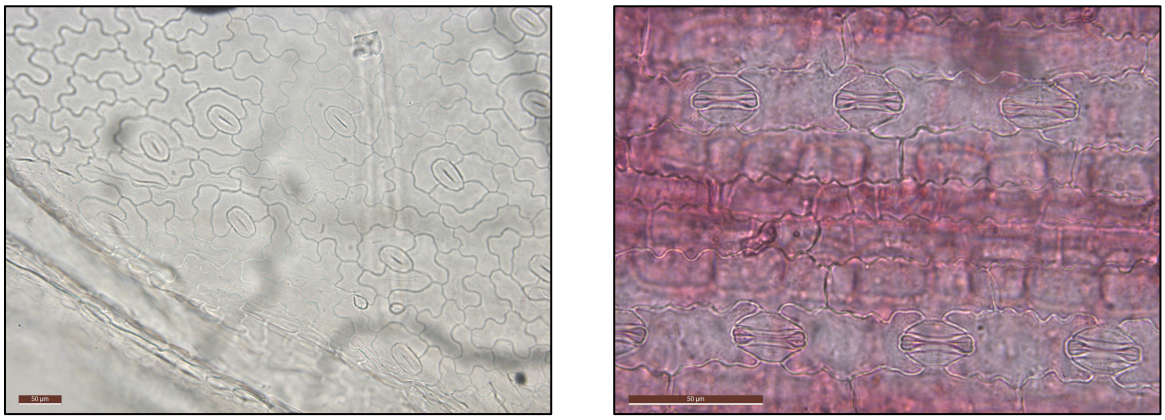
The collected ferns are hypostomatous and thus are only equipped with stomata on the abaxial side of the leaf. In respect to the grasses, it is dependent on the species. Some are hypostomatous, while some are amphistomatous, with the former meaning to feature stomata on only one side, mostly the abaxial one, and the latter having stomata on both sides of the leaf surface. Thus, samples were made from each side of the leaf of amphistomatous species.

Quality categories

To test the algorithm's sensitivity to image quality, two quality categories were established. Only the test images were assigned to either one of the quality categories. The distinction of the categories was based on the following definitions (see Table 2.2). Note, that these quality categories depend on the perceptual assessment of a human viewer. Figure 2.1 shows sample images for each quality category.

Table 2.2: Criteria for defining the two quality categories of low to medium and medium to high for the test images of the fern and grass image dataset.

Low to medium quality	**Medium to high quality**
• Large parts (50%) of the image is out of focus	• Minor parts (25%) of the image out of focus
• Great disturbance and noise in the image (i.e. air bubbles, cell tissue, etc.)	• Clear cell boundaries
• Blurry cell boundaries	• Stomata clearly visible
• Stomata often times hardly visible	• No disturbance like air bubbles, cell tissue, etc.



(a) (b)

Figure 2.1: Sample images for each quality category. **a** Sample fern image of the lower image quality. **b** Sample grass image of the higher image quality.

Image acquisition

Once the microscope sample slides were completed, images were acquired using the fully automated upright microscope system Leica® DM6000 B and the accompanied software Leica Application Suite® (version 4.12.0) for Windows®. Images at two different objective lens

magnifications (20× and 63×) were obtained. Two different magnifications were chosen for the purpose of testing the algorithm's sensitivity to size of the object of interest.

2.3 Manual counting

To establish a benchmark that represents the expected result which serves as the ground truth, manual countings of stomata in each test image of both the manipulated images and the additional fern and grass species were necessary. All visible and partly visible but clearly recognisable stomata of the entire image were counted by using the software ImageJ® and recorded as *ground truth*. In the case of the manipulated images, the manual countings of detected regions of interest (ROI) annotated by bounding boxes (i.e. *true positives*, *false positives* and *false negatives*, see Section 2.6 for a detailed explanation) of the results from the original (non-manipulated) images were used as reference point since the algorithm was tested on its sensitivity to changes in size and resolution. In respect to the fern and grass image dataset, the initially established counting represents the reference point. Note, that there is a large variance in the number of stomata between the test images as the images were of varying quality and partly blurred possibly hiding stomata.

2.4 Automated image processing and object detection

The algorithm developed by Jayakody et al. (2017), detects stomata as regions of interest (ROI) based on HOG features in a microscopic image after being trained utilising a specified training dataset. Details of the training process will be explained in the following section.

After automatically detecting stomata in the given microscopic image, the algorithm creates a bounding box around the ROI which will be found on the final output image as a result (see Figure 3.1 for a visual example). These bounding boxes were determined as correct or incorrect classification and recorded. Based on this data, the metrics of *precision* (or *positive prediction value*), *recall* (or *hit rate*) and the *F-measure* (or F_1-score) could be calculated. More details regarding the evaluation of the outcome can be found in Section 2.6.

The original image dataset and the corresponding results act as reference points when it comes to comparing and evaluating the results of the manipulated image results of all three experiments. In regard to the expanded image dataset of fern and grass species, the manually generated ground truth functions as reference for evaluation.

Another factor that influences the result of the object detection is a parameter that can be set before running the script and testing the algorithm. Since the algorithm makes use of a sliding window that searches the images for the ROI based on the HOG descriptor, it can be fine-tuned to meet the requirements of the test image with the respective object of interest. This parameter is given as an area range in pixels of the ROI in the test images. It should roughly resemble the size of the stomata found in the microscopic images, as otherwise the detector is searching for an object that does not relate to the object of interest and the outcome is not valid.

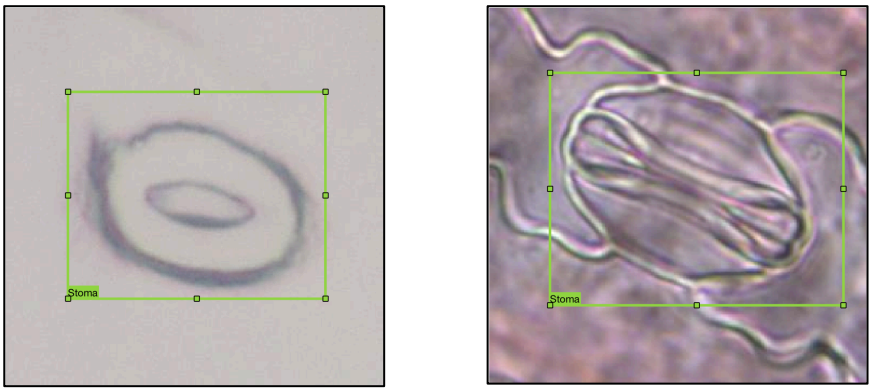
2.4.1 Training

Training of the algorithm is necessary to build a customised classifier that detects objects of interest – in our case stomata – with adequate accuracy. The training dataset needs to be representative of the images subject to the application of the object detector model (Casado-García et al., 2020).

The training involves building a training dataset beforehand made up of positive and negative images that the algorithm uses as reference to learn and distinguish ROI as positive meaning an object of interest has been found, or to classify the region as negative meaning no object has been reported. An object of interest has been found when it represents the specified feature of the histogram of gradients (HOG).

For both datasets, positive and negative samples were generated by utilising a script developed by Jayakody et al. (2017) that, through manual indication, crops the original image and thus generates separate smaller samples that together are used as training data. Each positive sample contains a single stoma. To generate representative positive samples, stomata in varying quality and size were selected in a random manner.

To build the classifier, the algorithm draws on the training dataset as well as a ground truth specified by using the Image Labeler App® provided by Matlab®. The ground truth contains rectangular ROI that have been manually labelled in each of the positive samples with a bounding box (see Figure 2.2).



(a) (b)

Figure 2.2: Examples of the labelled ground truth using Matlab's® Image Labeler App®. **a** Sample stoma from the original image dataset. **b** Sample stoma from the fern image dataset.

Manipulated images

The training dataset was used to train the classifier for testing the manipulated images. This classifier was built based on 8 cascade stages with a *False Alarm Rate* of 0.1, and a *Negative Sample Factor* of 1. These parameters yielded acceptable results without a large number of false detections (*false positives*). The training dataset is made up of 555 positive samples and 309 negative samples. The classifier was only trained on stomata from the original size image and not on stomata from the manipulated images.

Fern and grass images

The complete image dataset of 278 fern and grass images was split in two with each 139 images. Each set of 139 images was used as either training or testing dataset to achieve differing image datasets for the purpose of establishing an 'unbiased' classifier. To build a classifier that yields acceptable results, all of the respective 139 images were used for the training process, while for the testing only a selection of images (n=20) was used after being separated into the two quality categories (see Section 2.2.2, Table 2.2).

The training dataset contains 2072 positive samples, and 1154 negative samples of both ferns and grasses including both magnifications. Overall, it has been found that a classifier with 10 cascade stages, a *False Alarm Rate* of 0.1, and a *Negative Sample Factor* of 2 generates acceptable results regarding the ratio of correct and false detections with a relative low number of *false positives*.

2.4.2 Testing

The actual test of the algorithm's performance was done by running the code written by Jayakody et al. (2017). The execution of the script utilises the previously built classifier to analyse the given test images by using a sliding window approach. This means, according to The MathWorks (2020), that a window to detect the object of interest slides over the image while deciding whether the current region of the location of the sliding window contains the object of interest or if it does not, i.e. labelling the region as *positive* or *negative*, respectively. Based on these labels, the classifier passes the region to the next stage of the Cascade Classifier and only if the last stage labels the region as *positive*, an object of interest has been found and the region is labelled as ROI and a bounding box on the final output image is created. If the label is *negative*, the region's classification is complete and the window slides to the next location and continues to label each region (The MathWorks, 2020).

Manipulated images

In respect to creating a reasonable sample size for testing the algorithm and to start to see a pattern in the outcome, 12 images for each experiment have been chosen. The trained classifier was also tested on the same 12 images of the original dataset for reference purposes. The ROI area range that specifies the area range of the stomata found in the images was between 14000 pixels and 150000 pixels. This area range has been kept the same throughout all experiments to test the performance of the algorithm without adjusting the parameters to the respective stomata size in the enlarged or reduced images. As expected, the larger images (250%, 300% and 400%) took a longer time to be analysed and classified since the images have larger pixel dimensions and a great file size and thus big amount of data to be analysed.

Fern and grass images

A number of 20 images of each quality category, i.e. low to medium, and medium to high quality, was selected to test the performance of the detection algorithm. All 40 of those images encompassed both magnifications, however in varying proportions. There were 11:9 (20×:63×) images for the lower quality image dataset. For the higher quality images, the ratio was 17:3 (20×:63×). An additional 2 images have been used resulting in a total of five 63× images for the purpose of having a slightly bigger sample size when testing the algorithm's performance for each magnification.

Three test rounds were conducted. All tests have been done including both fern and grass images. The only distinction between test rounds was based on the difference in magnification

and lower or higher quality. One of the test rounds was done by using a generic ROI area. The generic ROI area parameter covers the full range of possible ROIs in all images containing both magnifications. The two other test rounds made use of an adjusted ROI area parameter that was tailored to the respective magnification level (i.e. the stomata size range in the image). This adjustment was performed for the purpose of testing whether a fine-tuning to the test image will be yielding an improvement in accuracy. The results will be shown in the following chapter.

2.5 Statistical analysis

As part of both research objectives, the effectiveness of the algorithm for all tests was compared between the ground truth, i.e. results of the original, non-manipulated image dataset (part I) and the respective manipulated images (i.e. enlarged, reduced or downsampled), and the generic outcome of the fern and grass image dataset (part II) with the adjusted fern and grass image results, respectively. To achieve comparability between tests of the different manipulation experiments, and to determine significant results among all tests, statistical testing methods were applied. Statistical analysis was performed using R (version 4.0.2).

Manipulated images

The original test image dataset was used as the best achievable result and thus used as reference point to which the experiments' results were compared to. To determine to what extent the algorithm's effectiveness in terms of mean difference m of precision and recall changes for each manipulation experiment (i.e. the significance of each treatment), multiple paired t-tests with a sample size of 12 were conducted. The null hypothesis H_0 is that there is no difference in precision or recall after each treatment when compared to the original image (H_0: $m = 0$). The alternative hypothesis H_a is that a difference can be found (H_a: $m \neq 0$).

Fern and grass images

To compare the algorithm's performance between quality categories, a Welch t-test was conducted. This compared the means of the two independent groups of the two quality categories (i.e. lower and higher quality images). The sample size of each testing group was 20. It was tested whether the mean m of the lower quality (LQ) image results (m_{LQ}) is different to the mean m of the higher quality (HQ) images (m_{HQ}). The expectation was that the higher quality images show higher precision values as there is less disturbance in the image that could lead to false positive detections. The null hypothesis H_0 is that there is a difference in precision or

recall when compared to each other (H_0: $m_{LQ} \neq m_{HQ}$). Whereas, the alternative hypothesis H_a is that no difference can be found (H_a: $m_{LQ} = m_{HQ}$).

Similar to the original and manipulation experiments, when comparing the generic and the adjusted ROI area parameter, a paired t-test was applied. Here, the null hypothesis H_0 is that there is a difference in precision or recall after adjusting the parameter when comparing to the generic outcome (H_0: $m \neq 0$). The alternative hypothesis H_a is that no difference can be found (H_a: $m = 0$). The expectation is that there is an improvement in accuracy as the detector was fine-tuned to its test input.

Moreover, a statistical test determining whether there is a difference in the algorithm's performance between the two stomata types of ferns and the more complex type of the grasses was conducted. A Welch t-test in the same manner as the significance test for image quality compared the mean difference between both groups.

2.6 Evaluation

To assess the algorithm's performance, the metrics of precision and recall as well as the *F*-measure, or F_1-score, of the test's accuracy were determined. These are commonly used metrics to evaluate results of image analyses (e.g. Casado-García et al., 2020; Tharwat, 2018). Precision "represents the proportion of positive samples that were correctly classified to the total number of positive predicted samples" (Tharwat, 2018, p. 4), while recall "represents the positive correctly classified samples to the total number of positive samples" (Tharwat, 2018, p. 3, 4) where positive samples are all stomata in the image irrespective of being detected or not. The *F*-measure can be determined using the previous metrics and it "represents the harmonic mean of precision and recall" (Tharwat, 2018, p. 5).

Each metric makes use of the previously determined correct or incorrect classifications involving *true positives* (*TP*), *false positives* (*FP*) and *false negatives* (*FN*). True positives are positive samples correctly classified as positive, i.e. true positives equal the number of detected stomata (ROIs). False positives are negative samples that have been classified as positive, i.e. corresponding to the number of bounding boxes that were incorrectly classified as stomata. False negatives are positive samples (ROIs) that have been mistakenly classified as negative, i.e. stomata that the algorithm missed. Equations 1-3 show how precision (Eq. 1), recall (Eq. 2) and the F_1-score (Eq. 3) were calculated:

$$P = \frac{TP}{TP+FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F - measure = \frac{2P \times R}{P \times R} = \frac{2TP}{2TP + FP + FN} \tag{3}$$

The results from each of these statistical tests are presented in the following chapter.

Chapter 3

Results

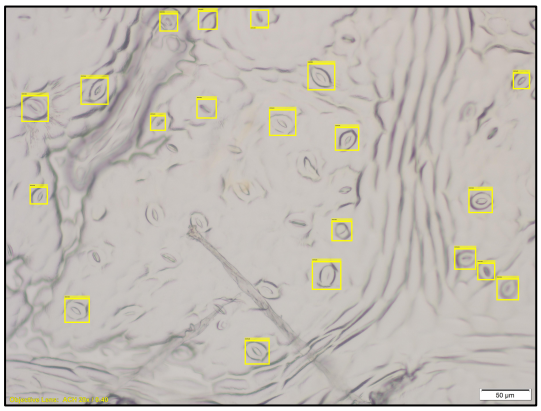
In this chapter, all the results obtained from testing the algorithm on the manipulation experiments as well as the fern and grass images are shown.

The numerical results refer to the manual counting involving correct stomata detections (true positives), incorrect predictions (false positives), and missed stomata (false negatives). The statistical tests make use of the previously described metrics showing the significance of the manipulation treatments while comparing to the original results. The fern and grass image results were compared in relation to lower or higher image quality and using a test involving a generic and an adjusted ROI parameter setting of the algorithm.

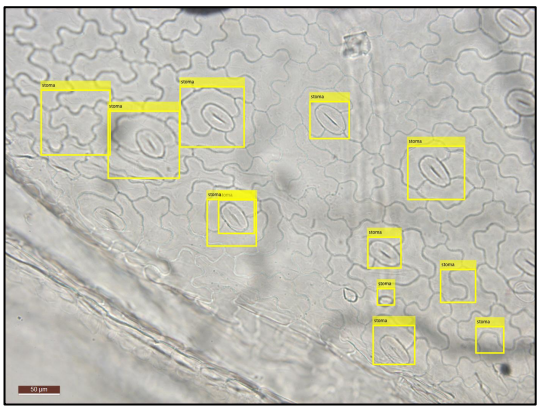
Additionally, results comparing the two stomata types are presented as it was expected that the more complex stomata type of the grasses might pose difficulties to the algorithm due to more complex HOG feature descriptor, consequently leading to a poorer performance. These results can be found at the end of this chapter.

3.1 Numerical results

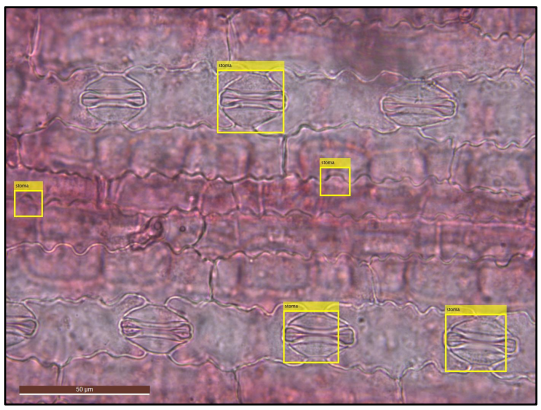
Here, the outcome of the manual counting of all tests is presented and briefly outlined. Figure 3.1 shows the results generated by the stomata detection method for sample microscopic images for the original, the fern and the grass image datasets of lower and higher quality, respectively.



(a)



(b)



(c)

Figure 3.1: Stomata identification results for three sample images of each of the datasets of original, fern and grass images. The yellow bounding boxes show automatically detected regions of interest (ROI). **a** Result of an original image of the dataset by Jayakody et al. (2017). **b** Result of a fern image (lower quality). **c** Result of a grass image (higher quality).

3.1.1 Manipulated images

A set of 12 images containing 395 stomata was tested. Table 3.1 shows the manually counted variables for each experiment with Original being the reference.

Overall did the various experiments involving the enlargement, reduction and downsampling of the original image dataset from Jayakody et al. (2017) yield varying results. The algorithm generated the highest number of true positives (TP) and the least number of false positives (FP) and false negatives (FN, i.e. missed stomata) for the non-manipulated original images. This was expected and thus these images were also used as reference for comparison of the experiments' outcome. The highest number of TP and lowest number of FP and FN was

achieved by the downsampling experiment. The reduction experiments yielded the lowest results for TP and FP while having the highest number of missed stomata (FN) (see Table 3.1).

Table 3.1: Numerical results obtained for the original image dataset and its manipulated experiments (i.e. enlargement, reduction and downsampling treatment) using 12 microscopic images containing 395 stomata. [1]An average value (combining all 12 images and experiments within each treatment) for each variable was calculated.

	Actual number of stomata	ROIs detected	True positive	False positive	False negative
Original	395	324	266	58	129
Enlargement[1]	395	260	148	112	247
Reduction[1]	395	102	75	27	320
Downsampling[1]	395	406	245	161	150

3.1.2 Fern and grass images

A set of 20 images for each quality category was tested, where the lower quality images contain 360 stomata, and the higher quality images contain 128 stomata. This difference in the number of stomata is due to fact that images of higher quality were images of larger magnification (63×) in most cases. Table 3.2 shows the manually counted variables for each quality category as well as results of the adjusted tests.

The results for both quality categories vary when comparing all recorded variables. The lower quality images did achieve a higher number of TP but also a greater number of FP and FN than the higher quality images.

Table 3.2: Numerical results obtained for the fern and grass dataset for each quality category using 20 microscopic images containing 360 stomata for the lower quality images, and 128 for the higher quality images. [1]Adjusted to the magnification levels of 20×, and 63×, respectively.

	Actual number of stomata	ROIs detected	True positive	False positive	False negative
Lower quality	360	260	117	143	243
LQ (20)[1]	312	191	102	89	210
LQ (63)[1]	48	3	1	2	47
Higher quality	128	157	52	105	76
HQ (20)[1]	77	80	29	51	48
HQ (63)[1]	58	9	7	2	51

3.2 Statistical results

As the interest of this research is to test the algorithm's accuracy in respect to stomata size and image quality, each treatment (i.e. manipulations of enlargement, reduction or downsampling) was compared to the non-manipulated, original test images while fern and grass images of different quality were compared to each other. To determine the significance of each test variable (size or quality) for each of the datasets (i.e. original and the fern and grass images), statistical tests were conducted. Paired t-tests reveal the significance in changes of precision and recall for the manipulations and for the adjustment of the ROI area for the fern and grass images. A Welch t-test determines the significance in differences of precision and recall when taking both image quality categories into account.

For all statistical tests, a significance level α of 0.05 was applied, meaning the null hypothesis was rejected for a p-value being inferior or equal to the significance level α, and thus the alternative hypothesis was accepted, respectively. Further details on this can be found in Section 4.2.3 when discussing which hypotheses were accepted.

3.2.1 Manipulated images

All treatments of manipulated experiments were compared to the original, non-manipulated results using paired t-tests. As mentioned in the previous chapter (see Section 2.5), the following statistical hypotheses were stated:

$$\text{Null hypothesis } H_0\text{: } m = 0$$
$$\text{Alternative hypothesis } H_a\text{: } m \neq 0$$

The results are shown in Table 3.3, as well as in Figures 3.2, 3.3 and 3.4. The respective results for individual treatments of each experiment (i.e. enlargement, reduction and downsampling) can be found in Appendix B.

For all manipulation experiments, the F_1-score indicates a varying performance. The highest score was obtained by testing the algorithm on the original images, as expected. The second highest score was achieved by the downsampling experiments, followed by the enlargement and finally the reduction experiments.

Table 3.3: Statistical results obtained for the original image dataset and the manipulation experiments (enlargement, reduction and downsampling). Each value represents the mean of all 12 tested images for each experiment combined. Note, the sample size of the mean of each manipulation experiment is greater than 12 as each experiment is made of these 12 images with varying treatments (see 2.1.1).

	Precision	Recall	F_1-score
Original	0.80	0.67	0.72
Enlargement	0.48	0.36	0.40
Reduction	0.54	0.19	0.25
Downsampling	0.64	0.62	0.61

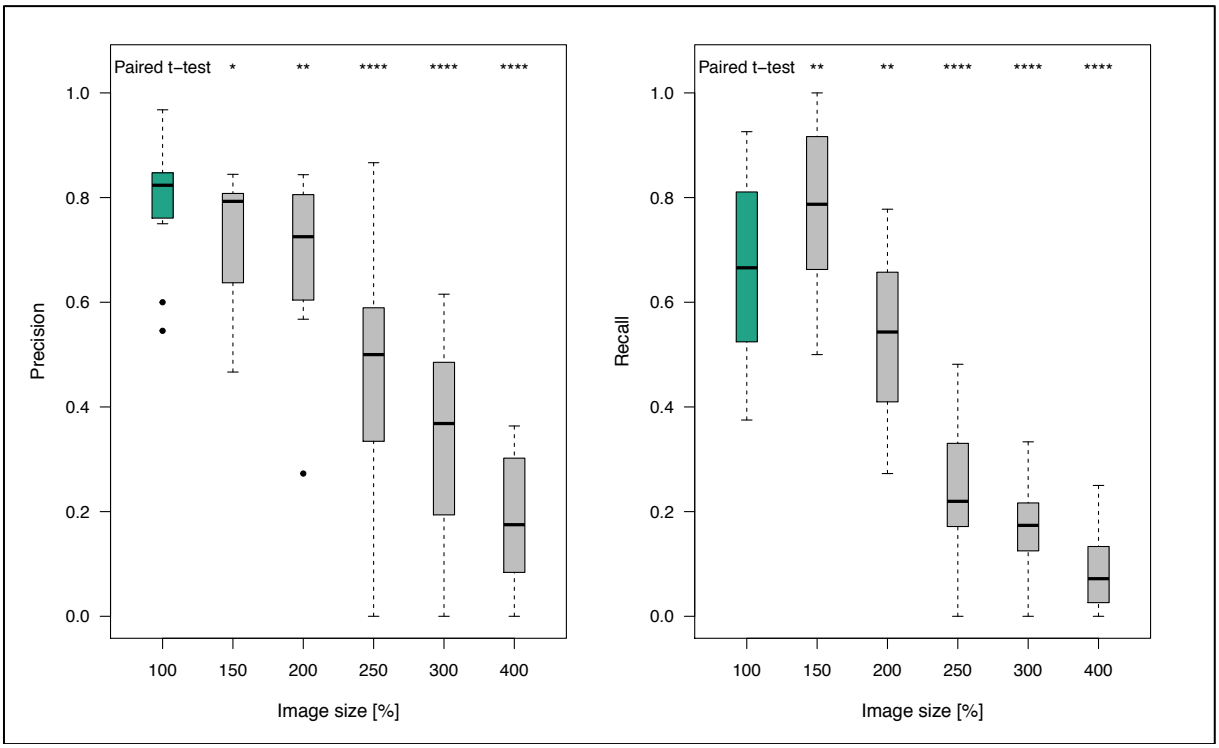


Figure 3.2: Precision and recall for manipulation experiment of **enlargement** with significance levels from the paired t-test when compared to the original image results.

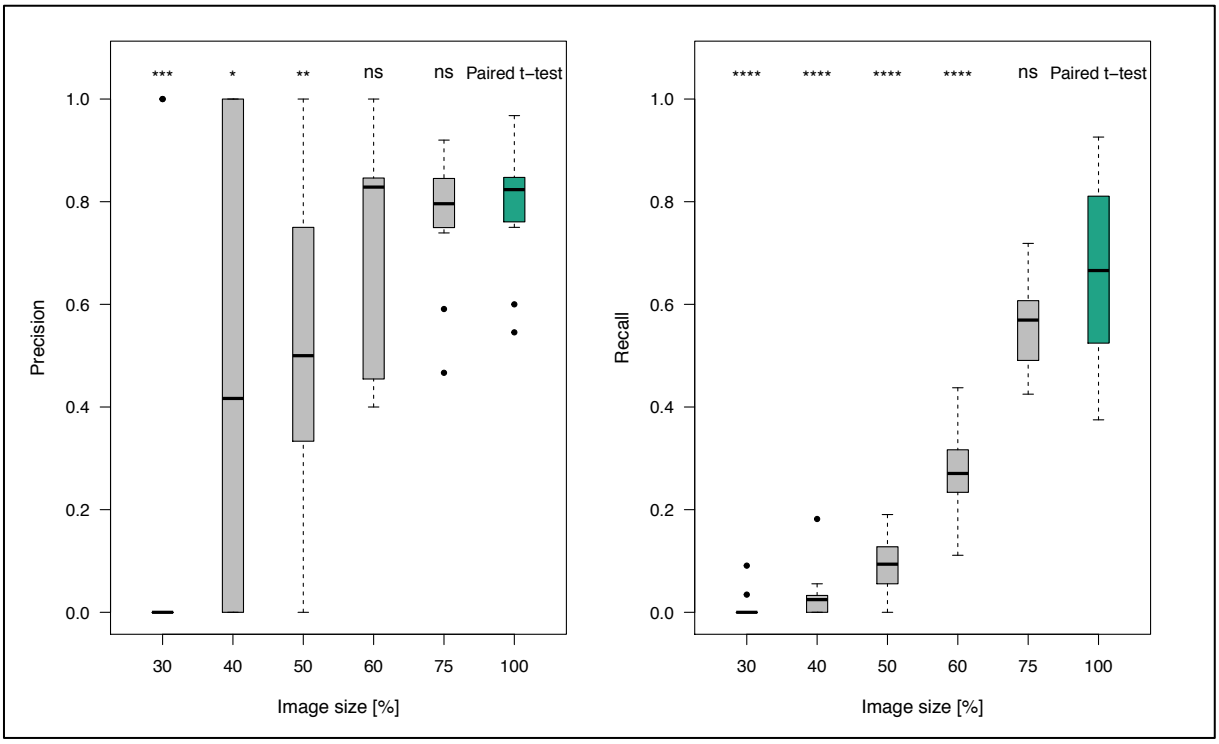


Figure 3.3: Precision and recall for manipulation experiment of **reduction** with significance levels from the paired t-test when compared to the original image results.

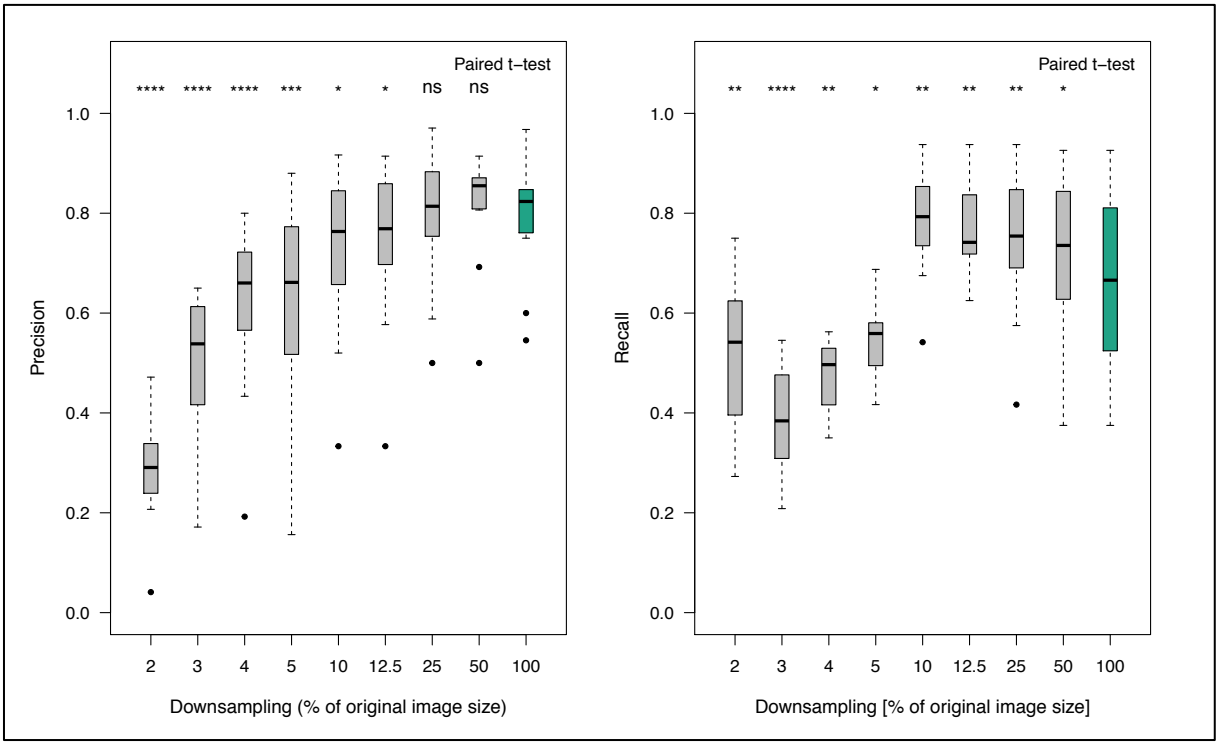


Figure 3.4: Precision and recall for manipulation experiment of **downsampling** with significance levels from the paired t-test when compared to the original image results.

3.2.2 Fern and grass images

Using a Welch t-test, the lower quality images were compared with the higher quality images in terms of precision and recall. As described before, paired t-tests were conducted to determine a difference between the generic and an adjusted ROI area parameter for both image qualities. The basis of the statistical tests were the following hypotheses:

$$\text{Null hypothesis } H_0\text{: } m_{LQ} \neq m_{HQ}$$
$$\text{Alternative hypothesis } H_a\text{: } m_{LQ} = m_{HQ}$$

The results can be found in Table 3.4 as well as in Figures 3.5 and 3.6.

For all tests of both image quality categories, the F_1-score indicates a varying performance. The highest score was obtained by testing the algorithm on the 20× images of both image qualities and the overall score of the generic ROI test of the higher quality images. The other tests involving the generic ROI test and the 63× images achieve lower F_1-scores, where the 63× yield the lowest.

Table 3.4: Statistical results obtained for the fern and grass image dataset for each quality category. Each value represents the mean of all 20 tested images. [1]Adjusted to the magnification levels of 20×, and 63×, respectively.

	Precision	Recall	F_1-score
Lower quality	0.34	0.31	0.29
LQ (20)[1]	0.49	0.32	0.35
LQ (63)[1]	0.11	0.04	0.06
Higher quality	0.45	0.54	0.37
HQ (20)[1]	0.42	0.64	0.41
HQ (63)[1]	0.29	0.17	0.20

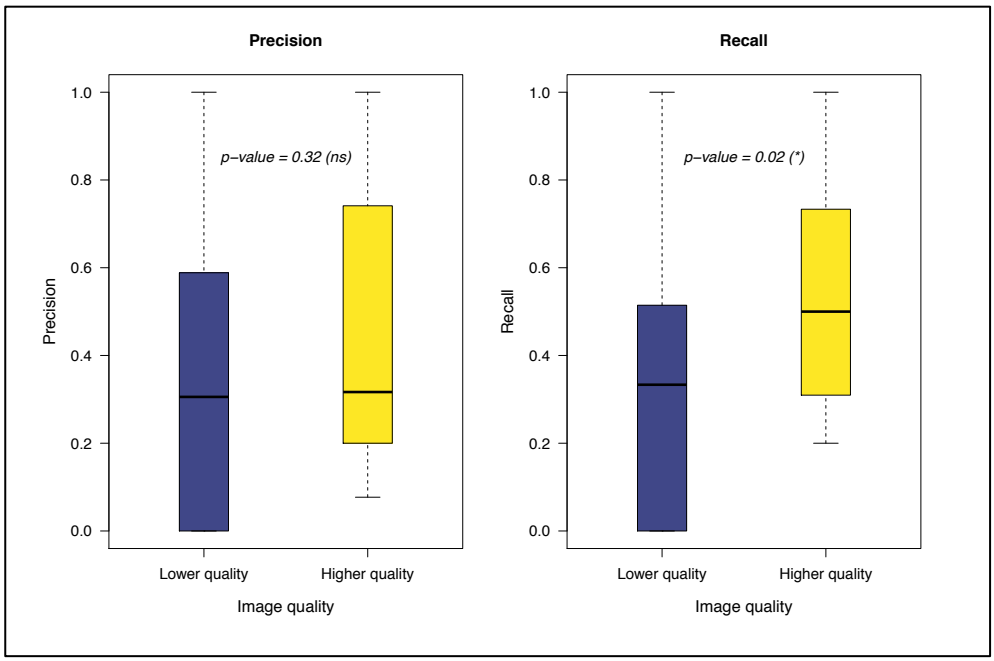


Figure 3.5: Precision and recall for each quality category with significance levels of the Welch t-test.

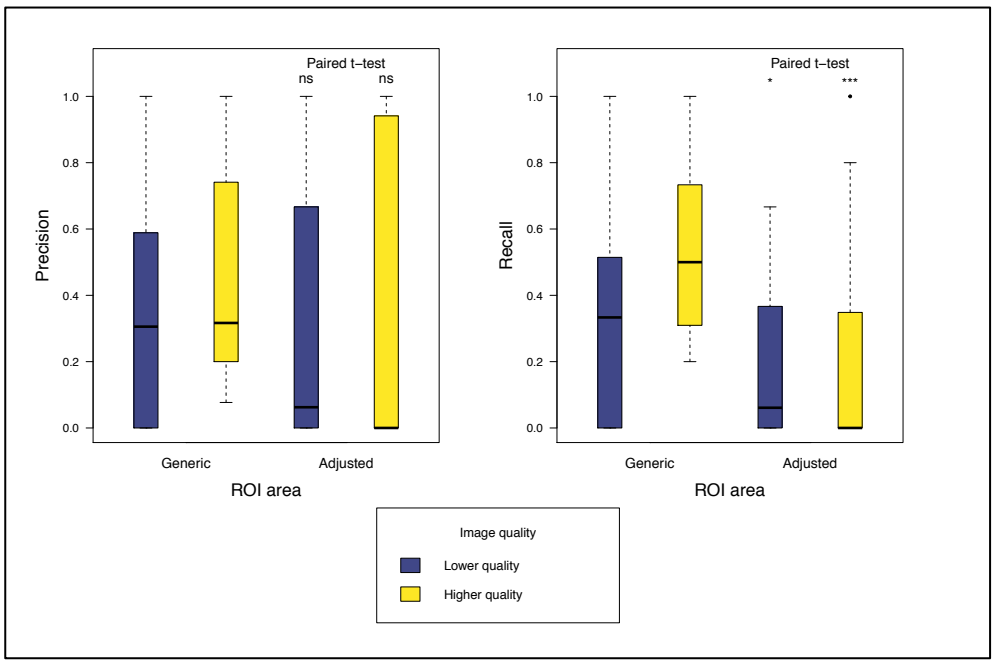


Figure 3.6: Precision and recall for the generic and adjusted ROI area for each quality category with significance levels from the paired t-test.

3.3 Results comparing stomata types

Here, the results comparing the two stomata types of the fern and grass families are presented.

3.3.1 Numerical results

A set of a total of 40 images with 16 and 24 for each family of ferns and grasses, respectively, was tested. The fern images contain 119 stomata, while the grass images contain 369 stomata. This difference in the number of stomata is due to fact that ferns have a lower stomata density than grasses. Table 3.5 shows the manually counted variables for each family combining both magnification levels (i.e. 20× and 63×).

Relative to the overall number of total detected ROI, the grass images yield a greater number of TP compared to the fern images. Also, less stomata are missed (FN) in the grass images.

Table 3.5: Numerical results obtained for the combined fern and grass images (i.e. 20× and 63× for each family together) using 40 microscopic images containing 488 stomata in total, with 16 images containing 119 (ferns) and and 24 images containing 369 (grasses).

	Actual number of stomata	ROIs detected	True positive	False positive	False negative
Ferns	119	144	48	96	71
Grasses	369	273	121	152	248

3.3.2 Statistical results

Using a Welch t-test, the fern images were compared with the grass images in terms of precision and recall. Due to the greater complexity in stoma type (i.e. the object of interest), it was expected that the performance of the algorithm was poorer and thus a difference between the mean precision and recall can be found. The basis of the statistical test were the following hypotheses:

$$\text{Null hypothesis } H_0\text{: } m_{Ferns} \neq m_{Grasses}$$
$$\text{Alternative hypothesis } H_a\text{: } m_{Ferns} = m_{Grasses}$$

The Welch t-test revealed a p-value of p=0.59 for precision and p=0.60 for recall. Both p-values are not significant meaning neither precision nor recall differ significantly between the fern and grass images (see Table 3.6 below).

Table 3.6: Statistical results obtained for the fern and grass image dataset combined (i.e. 20× and 63× together). Each value represents the mean of the 16 fern and 24 grass test images.

	Precision	Recall	F1-score
Ferns	0.36	0.40	0.33
Grasses	0.42	0.45	0.33

Chapter 4

Discussion

Studying stomata is an important part of research laying groundwork in predicting past and future climates, assessing their influence on the water and carbon cycles as well as the health and productivity of agricultural settings, among others. In this research, an automated machine-learning method was tested on a variety of plant types and stomata sizes but also different image qualities of the microscopic images that make up the foundation of stomata research.

In this chapter, the methodology of the data collection involving sample preparation and the training process of the object detector are reviewed. Additionally, the evaluation of the detector's performance is discussed, and limitations are highlighted. Finally, possible improvements and current advancements are briefly presented.

4.1 Methodology – Sample preparation

There are a variety of methods for microscopic sample preparation and image acquisition. Obvious advantages and disadvantages of the selected sampling methods are discussed in the following.

Such methods range from less complex methods such as applying a layer of nail polish on the leaf's epidermis and lifting the print using the clear tape and thus obtaining a leaf epidermis impression; or a more complicated approach of bleaching the leaf fragment and mounting each sample on to the microscopic slide permanently fixating the sample with glycerin. The former method seems to be the quickest and easiest method to carry out as the time investment is merely several minutes for the nail polish to dry. Yet, this method did not work well for all samples collected here. Especially the ferns are flimsy and fragile and thus, while attempting to remove the dried nail polish layer, the impression was not successful. Also, it is important to be careful in applying the tape as air bubbles are easily captured in between leading to an unclear sample when mounted on the microscopic slide. When it come to the leaf bleaching method, the time investment has to be taken into account when collecting data for future analysis. The amount of time a leaf needs to be bleached depends on the freshness and the kind of the leaf sample, as mentioned before in Section 2.2.2. Additionally, even after bleaching the leaf sample for a significant amount of time, tissue can still remain in the sample and lead to

noise in the final sample. However, the leaf bleaching method leads to more detailed results as the cell boundaries are captured better compared to the epidermal impressions.

4.2 Automated stomata detection

Here, various aspects that affect the training and performance of the automated stomata detection are discussed. Particular factors that have been tested and found to influence the outcome are described as well as the statistical results of the manipulated dataset and the fern and grass dataset are interpreted.

4.2.1 Influencing factors

There is a range of factors that seem to influence the effect of successful stomata detection in the microscopic test images for both testes datasets of manipulated and fern and grass images. In the following these factors are discussed.

ROI area range parameter
The ROI area range parameter affected the performance of the algorithm by means of presetting and thus aligning the detector's sliding window search and thus looking for the correct object of interest. If the parameter was deliberately kept at a range that does not represent the rough size of the stomata in the test image, there were either almost no reports of detected objects (parameter too large) or a too big number of bounding boxes in the final outcome image (parameter too small). Thus, it is necessary for the detector to successfully detect stomata that the ROI area parameter is set correctly. This, however, poses a limitation of the applied detection method which is further explained in Section 4.4.

To test the performance when fine-tuning the ROI parameter, it has been specifically adjusted to the different magnification levels of the fern and grass dataset. Adjustment took place for the 20× and 63× magnification, respectively. Both test results were compared to a test that used a generic ROI parameter which covers both magnifications by means of statistical tests. These results are discussed in Section 4.2.3.

Image size
The factor of image size is more or less directly related to the ROI parameter setting. Specifically, when testing the enlarged or reduced images from the original dataset, there seems to be a tipping point when the algorithm is not anymore able to detect stomata as well as in the original image. It is expected that this is related to the ROI area range which was not adjusted

when conducting the manipulation experiments and kept at the original setting. This is further discussed in Section 4.2.3.

Image quality

As the image quality was one of the objectives for testing the algorithm's performance, it has been found that – similarly to image size – when there seems to be a threshold where the difference in precision and recall compared to the original results changes in significance or drops to a lower value, respectively. Detailed explanations can be found in Section 4.2.3.

Note, that in terms of image quality, it has to be distinguished between the tests of the algorithm's sensitivity towards the quantifiable quality (i.e. resolution of the manipulated dataset) and the perceived quality (fern and grass dataset). Both appear to have a different influence on the detector's performance. The manipulated image experiments seem to show a tipping point at which the algorithm is not anymore able to detect stomata correctly, similar to the image size experiments. At this point, the difference in precision, for instance, changes and the values drop significantly. When testing the fern and grass image dataset while looking at the lower or higher image quality, the difference in precision, for example, appears not to be of great significance. Relevant statistical results are discussed below.

4.2.2 Training process

The training process for building a customised classifier for detecting stomata in different microscopic images of different plant types and species worked well and the short amount of time the training takes is a great advantage of the algorithm compared to other automated methods (see Jayakody et al., 2017). Training takes up to a few minutes at most. This enables researchers to process a great amount of data in an efficient manner accelerating the data analysis.

When training the object detector, it is assumed that the vast majority of the image – and thus the sliding window searching for the object of interest – does not contain stomata (The MathWorks, 2020). This, however, poses difficulties analysing microscopic images of some plant species with high stomata densities. Therefore, creating a training dataset of positive images is complicated, especially for the grass species, as the density of stomata is very high often times leading to the cropped images of the individual stomata for positive samples containing more than one stoma or at least fractions of other stomata leading to a misconception of "negative area" around the object of interest in the image. This may be one of the reasons, the

training is less accurate, and the outcome shows poorer results compared to the original dataset by Jayakody et al. (2017).

The purpose of selecting differing images for the training and test dataset was to test the method's effectiveness without being biased by having classified the same images in the preceding training process leading to a seemingly distorted accuracy. Similarly, this approach is also applicable in a practical approach in utilising this method for future object detection in images as only a limited set of images will be needed to be prepared for the training stage while the remaining images will be analysed. Only this way, the method would truly aid in a more efficient data acquisition and analysis and thus more reproducible results.

4.2.3 Performance and evaluation

The performance of the automated stomata detection was evaluated using the metrics of precision and recall as well as the *F*-measure. In the following, the numerical and statistical test results are interpreted and discussed.

For all statistical tests, a significance level α of 0.05 was applied, meaning the null hypothesis was rejected for a *p*-value being inferior or equal to the significance level α, and thus the alternative hypothesis was accepted, respectively.

Manipulated images

The algorithm generated the best results in terms of correct detections (i.e. highest number of true positives, least number of false positives and false negatives) for the non-manipulated original images. This was expected and thus these images were also used as reference for comparison of the experiments' outcome. Thus, paired t-tests were carried out to determine the significance of the treatment on precision and recall. Both of these metrics would be ideal when the value is one.

Enlargement

The statistical tests reveal a significant difference between the result of all treatments when each was compared to the original outcome. There appears to be a tipping point where the difference represents a higher significance (250% and larger) after treatment. This pattern is true for both, precision and recall. Thus, the null hypothesis H_0 ($m = 0$) was rejected and the alternative hypothesis H_a ($m \neq 0$) was accepted.

Reduction

Note, that the increments for reducing the image size are smaller than for the enlargement. The reason for this was that it was expected that the algorithm seemed to have a lower tolerance towards reducing the image size.

The results of the statistical paired t-tests reveal a more complex result for the treatments of reduction. A significant difference between the original result for precision and the image size of 50% and smaller was found but not for 60% and 75%. Note, that the level of significance is lower for 40% but again greater for 30%. This may be due to the great variance of precision values of each of the reduced 12 images. In this case, some outcome images of the reduction experiment show values of zero or one for precision. Yet, when precision equals one, it seems to imply that the algorithm was 100 percent successful in detecting all objects of interest in the image. However, it has to be kept in mind that precision equals one when there are zero false positives. Thus, even when the algorithm detected one single stoma only, but has not yielded any false detections, precision equals one. Seemingly perfect results like this example, have to be interpreted carefully when interpreting the algorithm's performance and also the significance of the treatment compared to the original. Results like these could be improved by having a larger sample. The metric recall, however, shows a clear pattern where all reduction treatments were significant except for the 75%.

As shown above, for the case of reduction, the null hypothesis H_0 ($m = 0$) was rejected and the alternative hypothesis H_a ($m \neq 0$) was accepted as well.

Downsampling

The downsampling experiment includes the most extensive number of treatments thus leading to a greater sample size which may have an influence on the result. It has been found that a tipping point is existent, similar to the enlargement experiments. The change in significance occurs at a very low image resolution where the pixel dimensions were 5% of the original image. Also, the values for recall drop at the 5%-mark although the difference is not significant which may be due to the great variance of the original values itself. Yet, the value is lower than the original one indicating a visible difference. Overall, this change in significance and the decrease can be interpreted as the algorithm being relatively insensitive towards image quality.

However, for the downsampling treatments, the null hypothesis H_0 ($m = 0$) was rejected as well and thus the alternative hypothesis H_a ($m \neq 0$) was accepted. Yet, this statement is solely based on the fact that six out of eight treatments showed a significant difference in terms of

precision when compared to the original result, despite only four showing the greatest significance and two being not significant. Conclusively it can be noted, there is a change in precision is existent, but the overall sensitivity is the least in terms of image quality (resolution).

The F-measure for all three manipulation experiments

The *F*-measure, or the F_1-score, ranges from zero to one with high values indicating high classification performance (Tharwat, 2018). For all manipulation experiments, the F_1-score indicates varying performance. As expected, the highest score was obtained for the original images. The second highest score was achieved by the downsampling experiments, followed by the enlargement and finally the reduction experiments. Accordingly, the F_1-score indicates the best classification performance for the downsampling treatment out of all three manipulation experiments. This insight can be interpreted that the algorithm may be least sensitive to image quality, similar to what the precision values show as mentioned before.

In conclusion, the manipulation experiments indicate that the object detector seems to be more sensitive to image size than image quality. Yet, this will be tested with the fern and grass dataset. Note, that the resolution (pixel dimensions) of the fern and grass images is lower than the dataset by Jayakody et al. (2017) to begin with. The images can be compared to around 30% of the downsampling treatment. Accordingly, as the downsampling experiments show the significance of difference in performance show up only from 5% and lower, it is expected that the numerical (quantifiable) image quality was not having an impact. Therefore, it has to be kept in mind that the perceived image quality has been isolated as main influencing factor when performing the statistical tests involving the fern and grass image dataset.

Fern and grass images

The object detector generated the best results in terms of correct detections (i.e. highest number of true positives, least number of false positives and false negatives) for the lower image quality. When it comes to the ROI adjustment, the 63×-higher-quality images score the highest number of TP compared to all detected ROIs.

Two different comparisons using statistical tests have been carried out determining whether there is a significant difference between the two image qualities itself (Welch t-test), and if a significant difference is present when adjusting the ROI area parameter and comparing it to a generic one (paired t-tests).

Comparison of lower and higher quality images

The Welch t-test reveals no significant difference between both image qualities in terms of precision, but a significant difference for recall. Thus, for precision the null hypothesis H_0 ($m_{LQ\ Precision} \neq m_{HQ\ Precision}$) was rejected and thus the alternative hypothesis H_a ($m_{LQ\ Precision} = m_{HQ\ Precision}$) was accepted. For recall, the alternative hypothesis H_a ($m_{LQ\ Recall} = m_{HQ\ Recall}$) was accepted. Note, that the overall number of FN was greater than the overall FP (see Table 3.2) as recall relies on the number of FN compared to precision depending on the number of FP. This finding indicates one of the average recall values – here the 63× lower quality image results (see Table 3.4) – was lower than the 63× higher quality image results meaning that more stomata have been missed in the lower quality images leading to a more significant difference despite not being of great significance. The F_1-score supports this result. This is explained in more detail below.

Comparison of a generic and an adjusted ROI parameter

As the ROI parameter being a main influencing factor on the detector's performance, it was tested whether an adjustment affects precision and recall in a significant manner when comparing to the non-adjusted (generic) results. Paired t-tests were conducted and it has been found that there is no significant difference for precision and thus the null hypothesis H_0 ($m_{Precision} \neq 0$) was rejected and the alternative hypothesis H_a ($m_{Precision} = 0$) was accepted. In terms of recall, a significant difference has been found, where the higher image quality results show a greater significance than the lower image quality images. In this case, the null hypothesis H_0 ($m_{Recall} \neq 0$) was accepted.

The above described result of recall showing a significant difference after adjusting the ROI parameter, but precision showing no significance, is similar to the Welch t-test's result and that the metric of recall depends on the number of FN. This implies that the number of missed stomata (FN) is greater for the adjusted ROI parameter than for the generic one. This result seems contradictory as adjustment of the ROI parameter was expected to improve the outcome, or at least yield the same results as a test involving the generic ROI parameter. A possible explanation for this is that the adjustment constrains the algorithm to an extent that leads to the sliding window method being less effective as the window scales up and down within these constraints during the search. Subsequently, the search window inspects the image within the constraints and thus having less flexibility leading to missed stomata.

The *F*-measure

For all tests involving both image qualities as well as the generic and adjusted ROI parameter, the F_1-score indicates a varying performance. The overall values were lowest for the lower quality images. Although the 20× images of both image qualities yield the highest scores, the

overall F_1-score does not surpass 0.5 indicating a mediocre performance of the classification and object detection.

The F_1-score may indicate an overall poorer performance for the fern and grass dataset when compared to the original dataset by Jayakody et al. (2017). This may be due to different method of stomata imaging using a bleached-leaf approach instead of an epidermal leaf impression leading to a greater noise and a feature-rich background in the final image. This, in turn, results in more incorrect detections as these features function as distraction to the algorithm due to their similarity of the shape to the object of interest.

In conclusion, testing the algorithm's performance using fern and grass microscopic images, indicate that there is no significant sensitivity when it comes to image quality. Consequently, it can be said that for practical reasons it is not required to take microscopic images with a very high resolution and thus requiring a greater memory space. This finding can be taken into account when applying this particular object detector in future research where preparation of new images is necessary.

Comparison of fern and grass stomata type

The Welch t-test reveals no significant difference between both families in terms of precision and recall. Thus, in both cases, the null hypothesis H_0 ($m_{Ferns} \neq m_{Grasses}$) was rejected and thus was the alternative hypothesis H_a ($m_{Ferns} = m_{Grasses}$) accepted. This finding indicates no significant influence of the stomata type when it comes to automatically detecting stomata in microscopic images despite the more complex HOG feature descriptor of the grass stomata as object of interest. This is crucial as this outcome shows the wide applicability of the automated stomata detection method to a variety of stomata which is valuable for the research involving a broad sample.

4.3 Conclusion – Comparing findings of manipulated images with fern and grass images

When comparing the findings of the tests involving the manipulated image dataset and the fern and grass images, it can be stated that image quality of both, the quantifiable and the perceived quality, had the least significant influence on the correct detection of stomata in microscopic images. Precision and recall of the downsampling experiment show the highest numbers overall compared to the enlargement and reduction experiments. Also, no significant difference has been found when comparing the lower and higher image qualities of the fern and grass images.

In contrast, image size, i.e. mimicked or actual stomata size, appears to play a more important role when it comes to automatically detecting stomata in microscopic images. There are many more missed stomata (FN) in the enlarged or reduced images.

Additionally, the stomata type appears to not be a crucial factor for the automated stomata detection. As long as the training involves both types, the algorithm shows an adequate performance. Yet there are some limitations of the method which could be addressed to improve the overall detection accuracy, especially for the fern and grass image dataset based on the microscopic images taken for this research. Further remarks on this can be found in Section 4.5.

4.4 Limitations

In respect to limitations of the object detection method, there are several aspects that make it somewhat cumbersome and time-consuming although the method attempts to accelerate the process of data analysis. In the following, these limitations are elucidated.

ROI area range parameter
As mentioned in Section 1.4, automated detection methods are assumed to reduce the amount of time needed to analyse microscopic images since spatial calibration is not needed as the algorithm was trained on the objects of interest and is thus analysing the image for the respective size. However, this algorithm needed to be given a parameter that pre-sets the area range of the ROIs (i.e. stomata found in the test images). This adds an additional step requiring measurements of several stomata beforehand to determine the rough size which can then be set as ROI area range parameter for the detector.

Setting a parameter that does not correspond with the images subject to testing, yields inaccurate results. Either, the detector determines wrong objects that relatively closely resembles the previously defined feature type, or the detector does not find correct objects that fit to what it has learned during the training process. For example, if the parameter was set smaller than the rough size of the stomata in the test images, any object of elliptical shape (as defined by the HOG feature type) such as air bubbles will be found to be a ROI and annotated with a bounding box. Furthermore, if the parameter is larger than the stomata in question, the detector reports, for instance, epidermal cell boundaries that form a shape similar to the HOG descriptor, as positive detection. In both cases, the results are not valid and the ROI needs adjustment in accordance with the images subject for testing. Thus, for the method to work

successfully, it is required to know the rough size of the object of interest (i.e. stomata) in the test images which does not lead to time savings in terms of avoiding manual measurements.

Collection of training data

To build a customised object detector, training on the image dataset in question is essential. For this, a set of positive and negative samples needs to be established. As suggested by Jayakody et al. (2017), training samples can be obtained by manually cropping individual stomata from the training image dataset. Negative samples were obtained from the same training dataset in a similar manner. This process, however, is time-consuming and might need expert knowledge (Casado-García et al., 2020), but it is crucial for the training of the algorithm and thus inevitable. Note, that a sufficient number (approx. hundreds) of training samples is required to build an effective object detector. After all, the size of the training dataset has improved using this method compared other classifiers that need thousands of samples to build an accurate classifier (Jayakody et al., 2017). Note, that the samples should be representative of the stomata found in the test dataset, i.e. including as many variations of appearance as possible.

Moreover, annotating the positions of ROI in the positive training samples is crucial for the training to work properly as this step makes up for the ground truth. For this, rectangular annotations were manually added to the positive samples using the Image Labeler App® provided by Matlab® (see also Section 2.4.1). This step also consumes a reasonable amount of time and must be considered when preparing the data.

Evaluation

Similar to the collection and annotation of training data, the evaluation involves a substantial amount of manual labour. The process of obtaining the numerical results involved counting the bounding boxes in the outcome images. True positives, false positives and false negatives have to be determined and recorded. Yet, this step can only be done involving human assessment but it is essential to get the results for determining the evaluation metrics of precision, recall and the *F*-measure.

In conclusion, several aspects such as knowing the size of stomata in the variety of test images and adjusting the algorithm's parameters, as well as manually annotating the ROI for the ground truth input for the algorithm leads to an overall great amount of manual labour that is still required to establish the training datasets and ground truth but also evaluating the final results by manually counting the bounding boxes are limitations of the training and testing of this particular machine-learning-based detector.

4.5 Improvements and advancements

Although the here tested method works appropriately, and automated detection methods based on HOG features still achieve high accuracies (see also Aono et al., 2019 and Jayakody et al., 2017), there have been advancements in the work on detecting stomata in microscopic images progressing form machine-learning approaches, like the one used in this research, to deep- leaning methods (e.g. Li et al., 2019). Similarly, Jayakody et al. (2017) is working on a so-called region based convolutional neural network (RCNN) method to automatically detect stomata (Jayakody et al., in prep.). Detection methods that implement deep leaning features outperform the classifiers build based on HOG features (Aono et al., 2019).

Especially, working with microscopic images having a feature-rich background, such as the fern and grass images dataset used here, can be a challenge for the HOG-feature-based detection method as shapes similar to stomata can be found in these images. In this case it is proposed that the more advanced methods using convolutional neural networks perform better and show high accuracy such as shown by (Bhugra et al., 2018).

Chapter 5

Conclusion

The aim of this research was to test the effectiveness of an automated self-learning algorithm in detecting stomata in microscopic images and how this is influenced by variations in stomata size and image quality using images of different plant types.

The research questions whether the automated detection method can be successfully applied to a variety of microscopic images featuring stomata including the more complex type of grass stomata, and differing stomata sizes as well as changes in image quality to correctly identify stomata were answered. For both parts of the research objective and for all involved experiments of enlarging and reducing, as well as testing on different image qualities by downsampling and classifying images in terms of perceived quality, the algorithm proved to be successfully applicable in detecting stomata in all images, however, with varying accuracy.

First, a sensitivity towards size differences has been tested by mimicking different sizes by manipulating images from the image dataset that was originally established alongside the implemented method developed by Jayakody et al. (2017). A significant effect involving a tipping point has been found indicating a sensitivity towards stomata size. Secondly, the algorithm seems to be least sensitive towards image quality. Noe, that there is a significant effect in the detection accuracy but this only occurs at a low level and is thus disregarded as crucial influencing factor. Yet, it has to be kept in mind that the image quality of the fern and grass dataset and, therefore the algorithm's performance, could have been improved by using a different method of sample preparation leading to a less feature-rich background with less distraction for the algorithm. More advanced methods are being developed that seem to overcome this issue. Finally, a variety of stomata types was introduced and tested by obtaining a new set of microscopic images of fern and grass species where the grass stomata feature the more complex stomata type. Here, the algorithm did not show a significant difference between both stomata types in the tested images indicating a wide applicability of the automated detection method to a variety of stomata which is a key finding of this research.

References

Aono, A. H., Nagai, J. S., da SM Dickel, G., Marinho, R. C., de Oliveira, P. E., & Faria, F. A. (2019). A stomata classification and detection system in microscope images of maize cultivars. bioRxiv, 538165.

Beerling, D. J., & Franks, P. J. (2009). Evolution of stomatal function in 'lower' land plants. *The New Phytologist*, *183*(4), 921–925.

Berry, J. A., Beerling, D. J., & Franks, P. J. (2010). Stomata: key players in the earth system, past and present. *Current Opinion in Plant Biology*, *13*(3), 233–240.

Bhugra, S., Mishra, D., Anupama, A., Chaudhury, S., Lall, B., Chugh, A., & Chinnusamy, V. (2018). Deep convolutional neural networks based framework for estimation of stomata density and structure from microscopic images. In Proceedings of the european conference on computer vision (eccv) (pp. 0–0).

Casado-García, A., Heras, J., & Sanz-Sáez, A. (2020). Google colaboratory for quantifying stomata in images. In R. Moreno-Díaz, F. Pichler, & A. Quesada-Arencibia (Eds.), *Computer aided systems theory – eurocast 2019* (pp. 231–238). Cham: Springer International Publishing.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 ieee computer society conference on computer vision and pattern recognition (cvpr'05)* (Vol. 1, p. 886-893 vol. 1).

de Boer, H. J., Eppinga, M. B., Wassen, M. J., & Dekker, S. C. (2012). A critical transition in leaf evolution facilitated the cretaceous angiosperm revolution. *Nature Communications*, *3*(1221).

de Boer, H. J., Price, C. A., Wagner-Cremer, F., Dekker, S. C., Franks, P. J., & Veneklaas, E. J. (2016). Optimal allocation of leaf epidermal area for gas exchange. *New Phytologist*, *210*(4),1219–1228.

Fetter, K. C., Eberhardt, S., Barclay, R. S., Wing, S., & Keller, S. R. (2019). Stomatacounter: a neural network for automatic stomata identification and counting. *New Phytologist*, *223*(3), 1671–1681.

Franks, P. J., Adams, M. A., Amthor, J. S., Barbour, M. M., Berry, J. A., Ellsworth, D. S., . . . others (2013). Sensitivity of plants to changing atmospheric co2 concentration: from the geological past to the next century. *New Phytologist*, *197*(4), 1077–1094.

Franks, P. J., & Beerling, D. J. (2009). Maximum leaf conductance driven by co2 effects on stomatal size and density over geologic time. *Proceedings of the National Academy of Sciences*, *106*(25), 10343–10347.

Franks, P. J., & Farquhar, G. D. (2007). The mechanical diversity of stomata and its significance in gas-exchange control. *Plant Physiology*, *143*(1), 78–87.

Hetherington, A. M., & Woodward, F. I. (2003). The role of stomata in sensing and driving environmental change. *Nature*, *424*(6951), 901–908.

Jayakody, H., Liu, S., Whitty, M., & Petrie, P. (2017). Microscope image based fully automated stomata detection and pore measurement method for grapevines. *Plant methods*, *13*(1), 94.

Jones, H. G. (2014). Stomata. In Plants and microclimate: a quantitative approach to environmental plant physiology (Second edition ed., pp. 122–152). Cambridge University Press.

Laga, H., Shahinnia, F., & Fleury, D. (2014). Image-based plant stomata phenotyping. In *2014 13th international conference on control automation robotics & vision (icarcv)* (pp. 217–222).

Li, K., Huang, J., Song, W., Wang, J., Lv, S., & Wang, X. (2019). Automatic segmentation and measurement methods of living stomata of plants based on the cv model. *Plant methods*, *15*(1), 67.

Liu, S., Tang, J., Petrie, P., & Whitty, M. (2016). A fast method to measure stomatal aperture by mser on smart mobile phone. In *Applied industrial optics: Spectroscopy, imaging and metrology* (pp. AIW2B–2).

Raven, J. A. (2002). Selection pressures on stomatal evolution. *New Phytologist*, *153*(3), 371–386.

Sharma, N. (2017). *Leaf clearing protocol to observe stomata and other cells on leaf surface.* Retrieved from https://bio-protocol.org/bio101/e2538 doi: 10.21769/BioProtoc.2538

The MathWorks, I. (2020). *Train a cascade object detector.* Retrieved from https://nl.mathworks.com/help/vision/ug/train-a-cascade-object-detector.html

Vaten, A., & Bergmann, D. C. (2012). Mechanisms of stomatal development: an evolutionary view. *EvoDevo*, *3*(1), 11.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 ieee computer society conference on computer vision and pattern recognition. cvpr 2001* (Vol. 1, p. I-I).

Willis, J., KJ andMcElwain. (2014). The evolutionary record and methods of reconstruction – introduction. In The evolution of plants (Second edition ed., pp. 1–26). Oxford University Press.

Willmer, C., & Fricker, M. (1996a). Introduction. In Stomata (Second edition ed., pp. 1–11). Chapman & Hall.

Willmer, C., & Fricker, M. (1996b). The structure and development of stomata. In Stomata (Second edition ed., pp. 36–91). Chapman & Hall.

Yuan, J., Wang, X., Zhou, H., Li, Y., Zhang, J., Yu, S., . . . others (2020). Comparison of sample preparation techniques for inspection of leaf epidermises using light microscopy and scanning electronic microscopy. Frontiers in Plant Science, 11, 133.

Appendix

Appendix A

The following tables present the exact species name of the collected plant material and the sample collection location.

Table 1: List of sampled fern species with location of collection.

Species	Location
Adiantum caudatum	Hortus Botanicum Amsterdam
Lygodium japonicum (Thunb.) Sw.	Hortus Botanicum Amsterdam
Matteuccia struthiopteris (L.) Tod.	Hortus Botanicum Amsterdam
Nephrolepis multiflora (Roxb.) F.M.Jarrett ex C.V.Morton	Hortus Botanicum Amsterdam
Pteridium aquilinum	Utrecht Botanic Gardens

Table 2: List of sampled grass species with location of collection.

Species	Location
Arundo donax	Utrecht Botanic Gardens
Cortaderia selloana	Utrecht Botanic Gardens
Festuca flavescens	Utrecht Botanic Gardens
Melica uniflora Retz.	Hortus Botanicum Amsterdam
Milium effusum L.	Hortus Botanicum Amsterdam
Oryza sativa	Utrecht Botanic Gardens
Pharus latifolius L.	Hortus Botanicum Amsterdam
Panicum virgatum	Utrecht Botanic Gardens
Sorghum bicolor	Utrecht Botanic Gardens

Appendix B

Here, the results are reported separated by experiment and each treatment. Each value represents an average calculated by combining all 12 tested images.

Table 3: Numerical results obtained for the original image dataset and each **enlargement** experiments using 12 microscopic images containing 395 stomata.

	Actual number of stomata	ROIs detected	True positive	False positive	False negative
Enlargement (150%)	395	411	310	101	85
Enlargement (200%)	395	308	219	89	176
Enlargement (250%)	395	198	101	97	294
Enlargement (300%)	395	196	70	126	325
Enlargement (400%)	395	183	38	145	357

Table 4: Numerical results obtained for the original image dataset and each **reduction** experiments using 12 microscopic images containing 395 stomata.

	Actual number of stomata	ROIs detected	True positive	False positive	False negative
Reduction (75%)	395	279	218	61	177
Reduction (60%)	395	149	109	40	286
Reduction (50%)	395	60	37	23	358
Reduction (40%)	395	19	10	9	385
Reduction (30%)	395	4	2	2	393

Table 5: Numerical results obtained for the original image dataset and each **downsampling** experiments using 12 microscopic images containing 395 stomata.

	Actual number of stomata	ROIs detected	True positive	False positive	False negative
Downsampling (50%)	395	347	290	57	105
Downsampling (25%)	395	363	296	67	99
Downsampling (12.5%)	395	403	306	97	89
Downsampling (10%)	395	417	310	107	85
Downsampling (5%)	395	352	219	133	176
Downsampling (4%)	395	303	184	119	211
Downsampling (3%)	395	306	147	159	248
Downsampling (2%)	395	758	211	547	184

Table 6: Statistical results obtained for the original image dataset and the **enlargement** experiments. Each value represents the mean of all 12 tested images.

	Precision	Recall	F_1-score
Enlargement (150%)	0.73	0.77	0.74
Enlargement (200%)	0.68	0.54	0.59
Enlargement (250%)	0.47	0.24	0.32
Enlargement (300%)	0.34	0.17	0.22
Enlargement (400%)	0.18	0.09	0.12

Table 7: Statistical results obtained for the original image dataset and the **reduction** experiments. Each value represents the mean of all 12 tested images.

	Precision	Recall	F_1-score
Reduction (75%)	0.77	0.56	0.64
Reduction (60%)	0.70	0.27	0.39
Reduction (50%)	0.55	0.09	0.16
Reduction (40%)	0.51	0.03	0.06
Reduction (30%)	0.17	0.01	0.02

Table 8: Statistical results obtained for the original image dataset and the **downsampling** experiments. Each value represents the mean of all 12 tested images.

	Precision	Recall	F_1-score
Downsampling (50%)	0.81	0.72	0.76
Downsampling (25%)	0.79	0.75	0.76
Downsampling (12.5%)	0.74	0.77	0.75
Downsampling (10%)	0.72	0.79	0.75
Downsampling (5%)	0.62	0.54	0.57
Downsampling (4%)	0.62	0.47	0.52
Downsampling (3%)	0.49	0.38	0.41
Downsampling (2%)	0.28	0.53	0.36