Die approbierte Originalversion dieser Diplom-/Masterarbeit ist in der Hauptbibliothek der Technischen Universität Wien aufgestellt und zugänglich.
http://www.ub.tuwien.ac.at

The approved original version of this diploma or master thesis is available at the main library of the Vienna University of Technology.
http://www.ub.tuwien.ac.at/eng

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

# Extraction of Quantitative Traits from 2D Images of Mature Arabidopsis Plants

DIPLOMARBEIT

zur Erlangung des akademischen Grades

**Diplom-Ingenieur**

im Rahmen des Studiums

**Medizinische Informatik**

eingereicht von

**Marco Augustin, BSc.**
Matrikelnummer 0725924

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Univ.Ass. Dipl.-Ing. Dr.techn. Yll Haxhimusa

Wien, 27. Januar 2014

(Unterschrift Verfasserin)

(Unterschrift Betreuung)

FAKULTÄT
FÜR !NFORMATIK

Faculty of Informatics

# Extraction of Quantitative Traits from 2D Images of Mature Arabidopsis Plants

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

**Diplom-Ingenieur**

in

**Medical Informatics**

by

**Marco Augustin, BSc.**
Registration Number 0725924

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Univ.Ass. Dipl.-Ing. Dr.techn. Yll Haxhimusa

Vienna, 27. Januar 2014

(Signature of Author)

(Signature of Advisor)

---

# Erklärung zur Verfassung der Arbeit

Marco Augustin, BSc.
Rögergasse 4/8, 1090 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

| | |
|---|---|
| (Ort, Datum) | (Unterschrift Verfasserin) |

# Acknowledgements

When a journey comes to an end there is a moment when you reflect all the experiences you have made. In the end you look back on many positive things which made the journey exciting and wonderful, but also negative things which let you despair come back to your mind. To share these bright moments and to overcome the dark moments every traveller needs good companions. I am very thankful for all the great companions being by my side along the *"journey"* of writing this thesis and also during my years of study.

First and foremost I want to present my gratitude to my supervisor Yll Haxhimusa who was an excellent guide during the creation of this work.

Furthermore I want to thank all the people at PRIP for their valuable inputs and thoughts during common talks and presentations. Additionally, I want to thank the people from the Busch Group at the Gregor Mendel Institute of Molecular Plant Biology, especially Wolfgang Busch, for their collaborative support and for introducing me into the interesting field of plant biology.

I would particularly like to thank Angela for supporting me emotionally at any time of this work and during the duration of my studies. I would also like to thank all my friends with whom seemingly desperate situations have been solved and successes have been celebrated through all the years.

Finally I want to thank my parents. Without their support and guidance I would not have been able to enjoy all these experiences.

# Abstract

The functional analysis of genes is a popular and interesting challenge in natural sciences. The understanding about genes causing pathologies in humans and animals or genes causing an increasing crop yield are only two important and relevant applications.

High-throughput phenotyping studies are seeking to increase the understanding about the impact of the genotype of an organism on its appearance- the phenotype. To find this correlation, genetic sequenced data as well as the phenotypic characteristics, so called *traits*, have to be determined. The bottleneck in these large-scale studies is the manual manipulation of samples and the subsequent determination of traits.

*Arabidopsis thaliana* is a widespread, small, flowering plant and a popular model in functional genomics. In this work a framework is presented to extract geometrical and topological traits from 2D images of mature *Arabidopsis* (e.g. length of a stem, number of branches). Due to logistical reasons the plants were dried and pressed before the images were acquired. Therefore some parts of the plants are overlapping and the “realistic” architecture has to be reconstructed from the 2D images before the traits can be extracted. The reconstruction of the plants’ architecture is done in two steps. In the first step, a tracing approach is used for the extraction of the centerline of the plant. In the second step, continuity principles are used to group centerline segments and reconstruct the plants’ realistic architecture. The need for supervision along the pipeline is tried to be brought towards a minimum. Nevertheless, methods for minor corrective interventions are provided to rise the throughput rate.

The accuracy and the grade of automation during the plant reconstruction is depending on the morphological complexity of the plant structure. Unsupervised trait extraction using the framework is reserved to plants with a limited morphological complexity and images with a uniformly high contrast.

# Kurzfassung

Die Erforschung des Zusammenhangs zwischen Erbgut (Genotyp) eines Organismus und dessen Erscheinungsform (Phänotyp) ist von großem Interesse. Relevante Beispiele aus dem Bereich der Naturwissenschaften sind hierbei die Identifizierung der genetischen Mechanismen als Auslöser von Krankheiten oder genetische Veränderungen in Nutzpflanzen um diese robuster gegen ihre Umwelt zu machen. Es wird versucht mittels groß angelegten phänotypischen Studien die Korrelation zwischen dem Genotyp und dem Phänotyp eines Organismus zu erforschen. Während das genetische Sequenzieren von Modell-Organismen heutzutage effektiv durchgeführt werden kann, ist das manuelle Vermessen von Merkmalen, die den Phänotyp eines Organismus beschreiben, meist ineffektiv und mit hohen Kosten verbunden.

*Arabidopsis thaliana* (Acker-Schmalwand) ist eine weit verbreitete, kleine, blühende Pflanze und zählt zu den populärsten Modellen in der funktionellen Genforschung. In dieser Arbeit wird ein Framework zur phänotypischen Analyse von ausgewachsenen *Arabidopsis* Pflanzen auf Basis von 2D Bildern vorgestellt. Dabei werden quantitative, phänotypische Merkmale, die sowohl geometrische als auch topologische Eigenschaften der Architektur der Pflanze beschreiben, gemessen (z.B. Länge eines Pflanzenstamms, Anzahl von Stängel).

Aus logistischen Gründen wurden die Pflanzen getrocknet und gepresst bevor sie fotografiert wurden. Das führt dazu, dass Teile der Pflanzen überlappen und die Rekonstruktion auf Basis der 2D Bilder erschwert wird. Kernaufgabe dieser Pipeline ist es diese Überlappungen von einzelnen Ästen so zu rekonstruieren, dass sie dem realen Verzweigungsmuster der Pflanze entsprechen. Mittels der angewandten tracing-Methode wird in einem ersten Schritt die Mittelachse der Pflanze iterativ extrahiert. In einem weiteren Schritt werden Teile dieses "Skeletts" nach Kontinuitätskritieren gruppiert und zusammen gesetzt.

Bei der Entwicklung des Frameworks wurde besonders darauf geachtet, dass die Anzahl der nötigen Benutzer-Interaktionen minimiert wird. Um die Durchsatzrate des Prozesses zu erhöhen gibt es allerdings die Möglichkeit kleine Korrekturen an den Zwischenresultaten vorzunehmen. Die erzielte Genauigkeit bei der Bestimmung eines Merkmals ist von der Komplexität der Pflanze bzw. von den Bildeigenschaften abhängig. Eine voll-automatische Bestimmung von topologischen und geometrischen Merkmale bleibt Pflanzen mit einer begrenzten Komplexität vorbehalten.

# Contents

**1 Introduction 1**
1.1 Motivation . . . 2
1.2 Problem Statement . . . 2
1.3 Aim of the Work . . . 4
1.4 Outcome of this Work . . . 5
1.5 Structure of the Work . . . 5

**2 High-Throughput Phenotyping of Multicellular Organisms 7**
2.1 Functional Genomics: From Genotype to Phenotype . . . 8
2.2 High-throughput Phenotyping of Multicellular Organisms . . . 8
  2.2.1 *Caenorhabditis elegans* . . . 9
  2.2.2 Crop Plants . . . 9
2.3 Digital Phenotyping of *Arabidopsis* . . . 11
2.4 Summary . . . 13

**3 Networks of Curvilinear Structures 15**
3.1 Overview . . . 15
3.2 Analysis of Curvilinear Structures . . . 18
3.3 Direct Exploratory Approaches . . . 19
3.4 Direct Exploratory vs. Non-Direct Exploratory Approaches . . . 22
3.5 Labeling and Grouping in Networks of Curvilinear Structures . . . 22
3.6 Summary . . . 24

**4 Pre-processing Strategies for Efficient Digital Plant Phenotyping 25**
4.1 Imaging Modality . . . 25
  4.1.1 Image Acquisition Setup . . . 26
  4.1.2 Limitations . . . 26
4.2 Methods for Pre-Processing . . . 28
  4.2.1 Color Spaces . . . 28
  4.2.2 Mathematical Morphology: Operators . . . 29
  4.2.3 Gray-level Thresholding . . . 31
4.3 Color Transformation and Alignment . . . 32
4.4 Automated Identification of Regions-of-Interest . . . 32

4.5 Spatial Calibration . . . 38
4.6 Summary . . . 39

**5 Quantitative Trait Extraction 41**
5.1 Plant Model . . . 42
5.1.1 Plant Architecture . . . 42
5.1.2 Model of Mature *Arabidopsis* . . . 42
5.2 Seed Point Identification . . . 43
5.2.1 Manual Seed Point Selection . . . 44
5.2.2 Automated Seed Point Selection . . . 44
5.3 Tracing of Stems and Branches . . . 50
5.3.1 Initialization . . . 53
5.3.2 Estimation . . . 54
5.3.3 Identification . . . 55
5.3.4 Validation . . . 57
5.3.5 Repetitious Tracing Prevention . . . 58
5.3.6 Stopping Criteria . . . 59
5.3.7 Supplementary Routines . . . 59
5.4 Plant Reconstruction . . . 61
5.4.1 Filtering . . . 61
5.4.2 Clustering . . . 62
5.4.3 Grouping . . . 64
5.5 Geometrical and Topological Trait Extraction . . . 69
5.5.1 Topological Traits . . . 69
5.5.2 Geometrical Traits . . . 70
5.5.3 Silique Detection . . . 70
5.6 Summary . . . 72

**6 Experiments and Evaluation 73**
6.1 Evaluation Data . . . 74
6.1.1 Image Dataset . . . 74
6.1.2 Ground Truth from Manual Phenotyping (GT UK) . . . 74
6.1.3 Ground Truth from Digital Phenotyping (GT Fiji) . . . 76
6.1.4 Evaluation and Development Environment . . . 77
6.2 Monochromatic Image Representation - Experiments . . . 78
6.2.1 Color . . . 78
6.2.2 Gray-scale images . . . 78
6.3 Evaluation of Pre-Processing . . . 79
6.3.1 Correctness of Automated ROI Identification . . . 80
6.3.2 Disk Space Savings . . . 81
6.3.3 Spatial Calibration . . . 83
6.4 Automated Seed Point Identification . . . 86
6.5 Centerline extraction . . . 87
6.5.1 Completeness . . . 87

6.5.2 Medial Axis Transformation . . . . . . . . . . . . . . . . . . . . . . . . . 88
6.6 Plant Reconstruction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 91
6.7 Quantitative Trait Extraction . . . . . . . . . . . . . . . . . . . . . . . . . . 96
6.7.1 Silique Number . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 96
6.7.2 Plant and Branches' Size . . . . . . . . . . . . . . . . . . . . . . . . . 99
6.7.3 Branching Pattern Analysis . . . . . . . . . . . . . . . . . . . . . . . . 100
6.7.4 Siliques Distribution . . . . . . . . . . . . . . . . . . . . . . . . . . . 102
6.8 Executability . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 103
6.8.1 Grade of Automation . . . . . . . . . . . . . . . . . . . . . . . . . . . 103
6.8.2 Time Performance in General . . . . . . . . . . . . . . . . . . . . . . 104
6.9 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 105

**7 Conclusion** **109**

**A Plant Image Samples** **113**

**Bibliography** **121**

# CHAPTER 1

# Introduction

The functional understanding of genes and their correlation to the appearance (the phenotype) of an organism is needed in different research fields [20]. Whole genome sequences for the most popular model organisms have already been determined. The bottleneck of present large-scale-genomic studies is the manual manipulation of samples and the subsequent determination of quantitative properties describing the phenotype of an organism. The use of appropriate computer vision technologies combined with the interdisciplinary work between biologists and computer scientists can maybe help to find the link between genotype and phenotype of multicellular organisms [47].

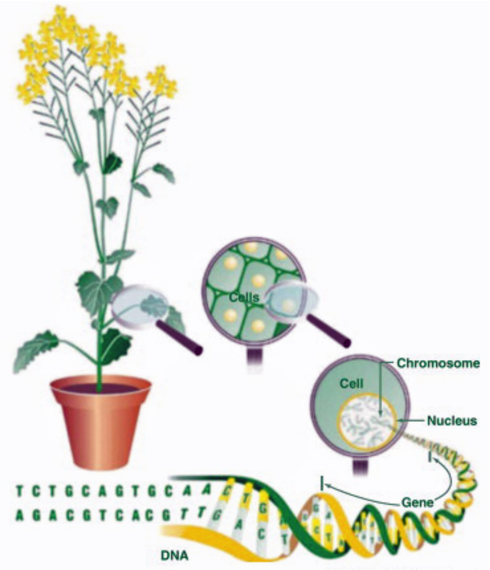


Figure 1.1: Finding the correlation between the genotype and the phenotype of multicellular organisms is needed in different fields, e.g. medicine or plant biology [20].

## 1.1 Motivation

The functional analysis of genes became a popular and interesting challenge in natural sciences over the past 30 years [57]. The correlation between genetic data and the likelihood for certain pathologies in humans or genes causing an increasing crop yield are only two important and relevant examples inside this field [9].
Whole genome sequences of model organisms like the fruit fly *Drosophilia melanogaster* or the plants *Arabidopsis thaliana* and rice is available nowadays. Inducing a specific gene mutation and studying its impact on the phenotype of these multicellular organisms is one of the most effective ways to understand the function of a gene [47].
While sequencing the genome of model organisms can be solved effectively nowadays, the automatic extraction of quantitative characteristics describing the phenotype became the bottleneck for many large-scale functional genomic studies [12]. The phenotypic characteristics, e.g. the size of a plant, are denoted as *traits*. Modern image acquisition tools and the increasing knowledge in the field of computer vision offer new possibilities for high-throughput phenotyping studies. The development of computer vision based phenotyping pipelines can help to overcome the current drawbacks of large-scale genetic studies, which are the manual manipulation of samples and the subsequent manual determination of relevant traits [47].
The demand of digital phenotyping technologies in biology builds up an interesting interdisciplinary field for researches in the field of computer science and especially in the field of computer vision [12]. *Arabidopsis thaliana* is a small flowering plant and popular model organism. The extraction of quantitative traits during its different stages of growing became an important challenge during large-scale genetic studies [20]. As the manual extraction of traits in high-throughput phenotyping studies, where up to 20,000 sample are analysed, is a time consuming and expensive task, a computer-aided extraction and analysis of traits would lower the costs.

## 1.2 Problem Statement

This work is located in the field of high-throughput phenotyping of plants. More specific, 2D images of *Arabidopsis thaliana* are analysed and phenotypic properties (*traits*) concerning the architecture of the plant are extracted. The 2D images are taken after the last stage of growing. During the last stages of growing, the plant mostly remains out of its stems and siliques (see sample image in Figure 1.2). By phenotyping mature *Arabidopsis*, traits concerning the final architecture of the plant can be extracted. Relevant traits describing the architecture are for instance [22]:

- Number of main stems and side branches (branching pattern)
- Length of main stems and side branches
- Average width of main stems and side branches

Further important traits are related to the plants' siliques. Siliques are the *fruits* of *Arabidopsis* and contain the seeds of the plant in the mature stage. The size of the siliques is an indicator

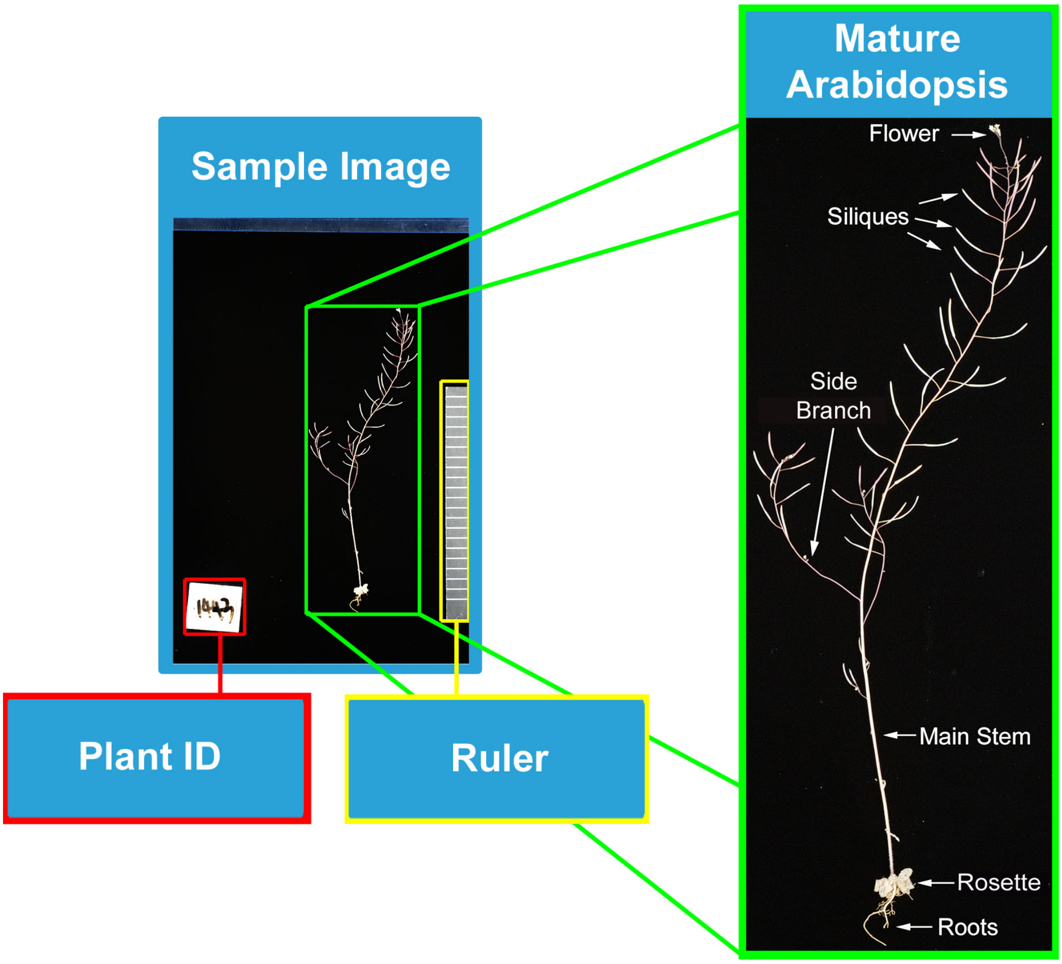


Figure 1.2: Sample image of the dataset used in this work. *Arabidopsis* in the senescent stage of growing, where the plant mainly consists out of stems and siliques. Some remaining flowers can hide relevant structures and are therefore identified as a source of error. The analysis of the rosette and the roots of the plant are not of interest for this work. A ruler and a plant ID sign are added during image acquisition.

for the amount of seeds produced. This is an important factor regarding the crop yield of a plant [30]. Relevant traits concerning the siliques are:

- Total number of siliques
- Siliques distribution: Number of siliques on the individual branches of the plant
- Length and average width of the siliques

The plants used in this work were growing in a natural environment where the surrounding conditions were well known. Due to storage and shipping reasons the plants dried and were pressed before the 2D images were taken. Because of that some stems are crossing and a correct reconstruction of the plants' "realistic" architecture has to be done before the relevant traits can be extracted. The grade of complexity of this problem increases with the number of critical points which can be identified as:

- **Branching point:** One stem branches into multiple sub-branches. Most frequent are branching points where one branch is branching off from another branch.
- **Crossing point:** Multiple stems are overlapping without a correlating physical connection in nature. Due to the projection into 2D space they appear as branching points in the images. The distinction between branching points and crossing points is crucial for the correct reconstruction of the branching topology of a plant.
- **Termination point:** Start and end points of the plant. Start points are located nearby the rosette while end points are mainly marked by siliques.

Based on the problems and the necessary tasks mentioned above, the problem statement can be divided into three parts:

1. ***Pre-Processing of the image data:*** High-throughput digital phenotyping studies imply high storage- and computational costs. To counteract these two challenges the regions-of-interest (ROI) in the images are cropped to its bounding box. The ROIs in this work are identified as the plant itself, a plant ID and a ruler.
2. ***Architectural reconstruction of the plant:*** Due to the overlappings which originate from logistics and the image acquisition, the reconstruction of the plants' realistic architecture must be done before traits can be extracted. This is mainly done in two step. In a first step the centerline of the plant is extracted from the images. In the second step, continuity principles are used to group centerline segments and reconstruct the plants' realistic architecture.
3. ***Extraction of quantitative traits:*** While geometrical traits are quantifying the geometry of the plant, topological traits are used to describe the branching architecture of a plant. Both types of traits are extracted after the "realistic" architecture of the plant was reconstructed.

## 1.3 Aim of the Work

The aim of this work is to develop a framework which is able to analyse 2D images of mature *Arabidopsis*. With use of the developed framework the time for the extraction of traits for biologists should be reduced and brought towards a minimum. The framework must provide methods for minor corrective interventions by the user during the different steps of processing. The most relevant traits which are extracted are:

- Branching pattern of the plant: Occurrence of different branch types

- Plant size: Length of the individual stems and branches
- Total number of siliques
- Length of siliques
- Distribution of siliques: Number of siliques on the individual branches of the plant

Further, this work should highlight approaches from computer vision which are useful in the field of digital phenotyping of multicellular organisms. A special focus is hereby set on the analysis of curvilinear objects which form tree-like structures. The concepts of the proposed pipeline for phenotyping the branching topology of the shoot system (parts of a plant growing above the ground) should be adaptable for the analysis of similar structures, e.g. like the root system of plants.

Another goal of this work is to obtain new insights concerning the architecture of mature *Arabidopsis thaliana* from a technical point of view. This can be valuable for the definition of a model for the design of future high-throughput phenotyping pipelines.

## 1.4 Outcome of this Work

In this work a framework is presented to extract geometrical and topological traits from 2D images of mature *Arabidopsis*.

The resulting pipeline performs every step for phenotyping the plant. It starts with pre-processing the image data and ends with the final extraction of quantitative traits. The architectural reconstruction of the plant is achieved by extracting the centerline of the plant using a tracing approach and hierarchically reconstructing the plant's architecture based on continuity principles.

The evaluation shows that geometrical traits like the main stem length or the silique length can be measured with an accuracy of over 90 %. The extraction of complex topological traits (e.g. the number of side branches at different depths) are more error-prone as they suffer from error-propagation along the pipeline. Unsupervised trait extraction using the framework is reserved to plants with a limited morphological complexity and images showing an uniformly high contrast.

## 1.5 Structure of the Work

This work is structured as follows. In Chapter 2 a brief introduction into phenotyping of multicellular organisms is given for a better understanding of the overall goal of this work. Important aspects of recent approaches are highlighted and relevant methods are described roughly. Available tools for extracting different traits of *Arabidopsis* are reviewed and compared.

Chapter 3 provides an overview of approaches for the extraction and analysis of networks of curvilinear objects from 2D images. Different approaches to extract the medial axis of an object are discussed and the general strategies within these approaches are reviewed.

After reviewing the relevant literature for this work in Chapter 2 and 3, the image data for the evaluation of the proposed framework as well as the used methods for pre-processing are described in Chapter 4.

Chapter 5 describes the approaches used for the reconstruction of the plants' architecture and the subsequent extraction of quantitative traits. In Chapter 6 the evaluation of the framework and experiments are discussed. The thesis' last section includes a short summary of the thesis as well as a discussion for future works.

Each chapter, except Chapter 1 and 7, starts with a short introduction which highlights the topics to be discussed in the chapters' sub-sections. Each chapter is shortly summarized in the chapters' last section.

## CHAPTER 2

# High-Throughput Phenotyping of Multicellular Organisms

The functional understanding of genes and their correlation to the appearance (the phenotype) of an organism is needed in different research fields. In an agricultural context one major focus is to know which genetic configuration can lead to an increasing crop yield or which genes can make a plant more robust to different natural environments. Researchers in the field of medicine study the likelihood for a specific pathology when being carrier of a specific gene [9, 47].

One of the most relevant approaches to reach this goal is mutational analysis. In these studies a specific gene mutation is induced and its impact on the phenotype is observed. Alternatively, natural strains with many mutations are investigated and correlations of DNA sequences and phenotypes are investigated. To find a significant correlation between the genotype and the phenotype, the genetic sequenced data as well as the phenotypic description of an organism have to be determined. The phenotypic description is done by quantifying phenotypic characteristics (e.g. length, width of a plant), the so called *traits*. While whole genome sequences for the most popular multi-cellular organisms have been determined, the bottleneck of large-scale-genomic studies in the present is the manual manipulation of samples and the subsequent screening by eye [47].

Increasing computational power as well as new technologies in image acquisition made it possible to automatize some tasks during a phenotyping pipeline. Approaches from the field of computer vision are used to extract and quantify the traits of interest [47]. This chapter highlights selected existing approaches in high-throughput phenotyping studies and is structured as follows. In Section 2.1 the basic principles behind functional genomic studies are roughly described for a better understanding of the overall goal of this work. This general section is followed by examples of phenotyping pipelines which were developed for different multicellular organisms like *Caenorhabditis elegans* (*C. elegans* worm) and crop plants like *rice* or *maize* (see Section 2.2). Section 2.3 focusses on the latest developments of digitally phenotyping *Arabidopsis*. This includes a short review of selected tools and methods as well as highlighting the quantitative traits which are currently in focus of research.

## 2.1 Functional Genomics: From Genotype to Phenotype

Traditional methods in molecular biology focused on the experimental analysis of single genes or proteins in depth. In modern biology this approach changed to experiments on the whole collection of DNA (the genome), RNA (the transcriptome) or protein (the proteome). *"Functional genomics is the genome-wide study of the function of DNA (including genes and nongenic elements), as well as the nucleic acid and protein products encoded by DNA" [40].*

While the genotype of an organism is comprised by its DNA, the phenotype is the *"outward manifestation in terms of properties such as size, shape, movement, and physiology" [40].* These quantitative phenotypic properties are called *traits*.

A fundamental and challenging task in the field of biology is to find the correlation between the genotype and the phenotype of organisms. Gaining a clearer understanding concerning this relation often implies modern high-throughput studies which further can be complemented by traditional methods.

Two different principles exist for high-throughput genetic screening: *reverse* and *forward genetics*. In reverse genetics ("gene-driven") a large number of genes is systematically inhibited and phenotypic traits are measured. In the "phenotype-driven" approach (forward genetics) the trait of interest is defined beforehand, e.g. the growth rate of a plant in presence of specific environmental conditions (e.g. chemicals). Then different mutants are created and the phenotype of interest is observed. If there is one class of mutants which shows a significant different behaviour compared to others, more specific investigations can be applied in further experiments [40].

Both approaches need a high laboratory effort during phenotyping the organisms. While sequencing of organisms became an inexpensive task, the manual trait extraction of thousands of phenotypes seems to be the bottleneck in large-scale functional genomic studies. These studies are demanding computer aided phenotyping pipelines which can make high-throughput phenotyping easier and cheaper [24]. Improvements in digital phenotyping are playing a key-role in finding the link between the genotype and the phenotype of organisms. In the following chapters some last developments of digital phenotyping approaches are listed.

## 2.2 High-throughput Phenotyping of Multicellular Organisms

A number of animals and plants became popular models for large-scale genomic studies, e.g. *Caenorhabditis elegans* (a worm), *Drosophila melanogaster* (fruit fly), maize or rice. The first example of high-throughput phenotyping approaches in this work shows the phenotyping of *Caenorhabditis elegans*. This work is interesting for two different aspects. First, it shows that high-throughput phenotyping studies are not limited to non-moving (or rather slow-moving) objects like plants. Second, there are similarities concerning the digital phenotyping setup between plants and worms. As well as plant seedlings, in some setups also worms are living in a chamber with a fluid material. Within this chamber the worms and other objects can overlap during image acquisition. This makes the phenotyping of single objects more difficult.

### 2.2.1 *Caenorhabditis elegans*

*C. elegans* is one of the most popular animal model organism. The genome of this worm was the first whole genome of an animal which was ever sequenced. The size and the environmental robustness made this animal popular for large scale genomic studies [58]. Traits of interest are for example properties describing the locomotion or the grouping behaviour of worms. Another major phenotype to be extracted during phenotyping studies of *C. elegans* is survival. This means different environmental conditions (e.g. heat, small chemical molecules) are tested during the development of the worms and the ratio of number of larvae to number of larvae and eggs is determined at the end of the experiment. For this approach *White et. al.* [58] developed a method to automatically quantify the number of adult worms, larvae and eggs/embryos within seconds. Images from a 96 microwell plate (flat plate with multiple wells which can hold fluids used in testing laboratories) in which the worms are observed are acquired. These images are processed in a hierarchical image processing pipeline, where large and small segments are segmented, objects are decomposed into parts and extracted features of these parts are classified by a SVM (Support-Vector-Machine). In a last step the single parts are merged together according to their similarity. Figure 2.1 shows the phenotyping pipeline as well as some example result images [58].

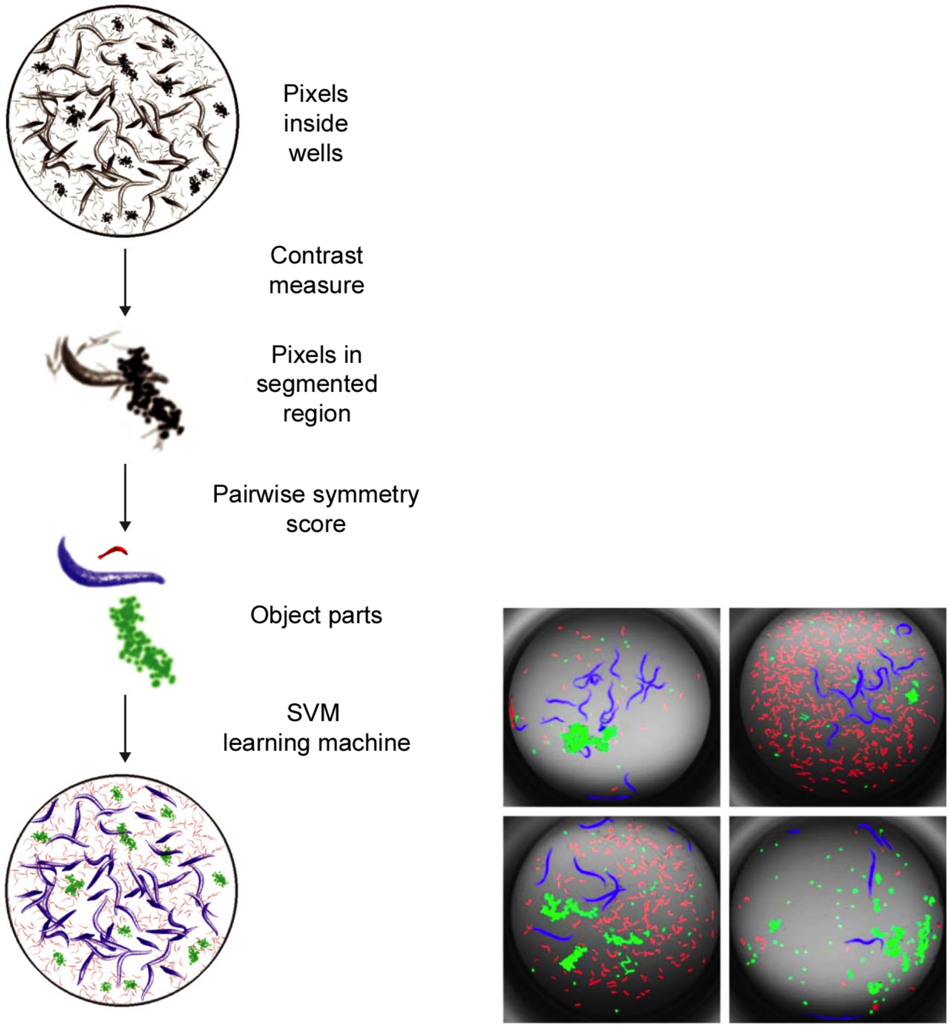
Another example where algorithms of computer vision are used during phenotyping of *C. elegans* is the extraction of quantitative traits concerning the moving behaviour of worms. *Restif et. al* [43] extracted different traits concerning the moving behaviour (e.g. speed) with use of tracking.

### 2.2.2 Crop Plants

The ongoing changes in the global climate as well as food security and the discovery of plant-based biofuel feedstocks are continuously increasing the demand on crop yield. As regular breeding programs are not longer sufficient to face these major challenges, modern plant breeding technologies are required. One promising field is the functional analysis of genes [20]. Gaining genetic data from popular plant models became an inexpensive task. The current bottleneck of large-scale genetic studies is the extraction of high quality phenotypic data from a high number of samples [24]. For this purpose different high-throughput phenotyping pipelines were already developed for the three major crop plants: maize, rice and wheat. *Grift et al.* [24] developed a method to determine two root trait characteristics which are the fractal dimension (FD) for describing the complexity of the roots and the root top angle (RTA). With use of the developed method up to 700 roots can be analysed per day. The roots are placed into a soft box, containing two fixed monochrome cameras and 5 images (one top view, four lateral views) are taken. For the lateral images, the root is automatically rotated by 90 degrees. Additionally, two background images (one top view, one lateral view) are taken. Background subtraction is used for segmenting the root and extracting the traits [24].

While this approach analyses the root traits of mature maize, most of the digital plant phenotyping pipelines are extracting traits from plants during an earlier stage of growing. A typical sketch of a digital phenotyping pipeline in a laboratory environment is shown in Figure 2.2.

*Galkovskyi et al.* [21] developed a framework called GiARoots, which can be part of such a



(a) Hierarchical image processing pipeline developed for phenotyping C. *elegans* survival rate [47].

(b) Resulting images, with labeled results: adult worms (blue), larva (red) and embryo (green) [58].

Figure 2.1: Image processing pipeline and results for automatically phenotyping the characteristic survival for C.*elegans*.

phenotyping pipeline. The tool is able to extract up to 19 root system architecture (RSA) traits from root images, e.g. maize roots. This tools' special focus is the applicability to varying root images. Several parameters can be set for an image stack or individual images. Further, the user can choose between three segmentation algorithms and up to 19 traits can be set. GiA Roots offers the possibility to expand or replace segmentation algorithms or expand the list of root traits. The standard segmentation algorithms available in the framework are global thresholding, adaptive thresholding and double adaptive thresholding. To extract some of the traits a skeletonization of the segmented image is needed. Morphological thinning is used to extract the medial axis of the roots by default. RSA traits which can be chosen via a graphical user interface (GUI) are for example the average width of a root, bushiness, convex area, major ellipse axis length, network depth, network surface area or root length. The evaluation of this tool was done by comparing the results to a set of other algorithms. The dataset comprised 2393 images from 12 different genotypes of maize [21].

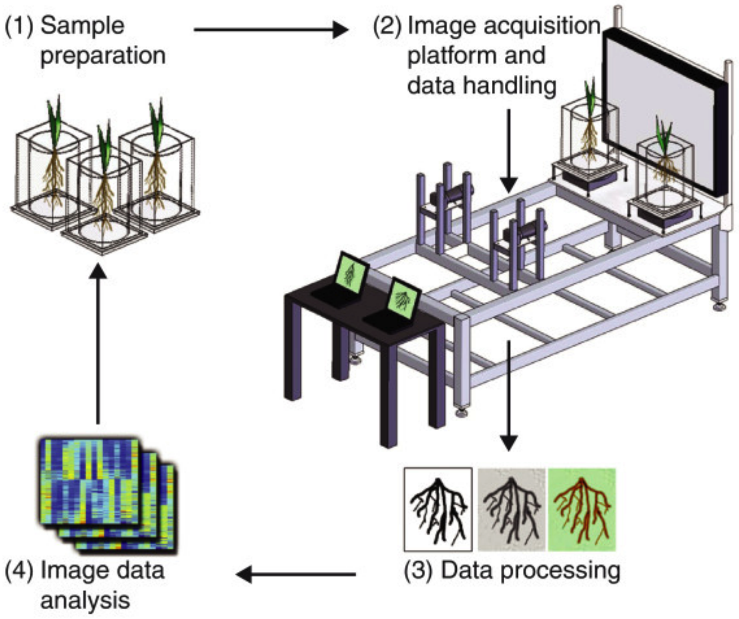


Figure 2.2: A digital plant phenotyping pipeline in a laboratory environment: (1) Plants are growing in cylinder or in plates filled with a gel-based media. (2) Image acquisition: Varies from a regular camera or scanner up to a fully computer control image acquisition setup. (3) Image data is processed and traits are extracted. (4) Extracted traits are analysed [47].

## 2.3 Digital Phenotyping of *Arabidopsis*

Due to its small size and relatively small genome consisting out of approximately 135 megabase-pairs [1] (comparison to human genome: approximately 3000 megabase-pairs [13]) *Arabidopsis* became the most popular plant model for functional genomic researches. Further, the existence of insertional mutations for most of the genes made it a popular model for phenotypic studies. Computer based phenotyping pipelines especially focus on root growth rates and root architecture characteristics like root length or root angle [47]. Few digital phenotyping pipelines are known gathering non-root characteristics as for example leaf size and shape [3]. In the following, present digital phenotyping methods for *Arabidopsis* are roughly described. The focus during this discussion is on the methods' underlying image processing pipeline as well as on the traits which were extracted. An overview of which technique extracts which phenotypic properties is shown in Table 2.2.

*Subramanin et al.* [49] developed a high throughput robot system for analysing the ability of plant roots to react on gravity changes. These kind of experiments are called root gravitropic experiments. A robot based image acquisition setup is presented which is able to fully automatic process 36 Petri dishes (cell cultural dishes). Each Petri dish, filled with a fluid gel, can hold between two and four plant seedlings. The plants are tracked during this curvature adaption in a 3 minute interval over a 8 hour period. A specific challenge is the correct focusing on the plant seedlings because of the camera robot system. This is done using focus measuring methods. The segmentation of the plant root is achieved using pixel probabilities based on two assumptions: (a) the root pixels are darker than the background pixels and (b) the root does not occupy more than 3 % of the image. These assumptions are used to define the likelihood and the prior in terms of the Bayes theorem. With use of the obtained probability matrix, patches from similar high probabilities are clustered and analysed. Each patch gets an heuristic driven weight based on basic moment analysis (area, major and minor axis). Patches with a root similar appearance (large, long, thin) get a higher weight than others. The patch with the highest score is identified as the most probable root patch [49].

*French et al.* [19] developed RootTrace which is an open source framework for high-throughput phenotyping the root of *Arabidopsis* during the first days of growing (day 1 to day 5). In contrast to the approach described by *Subramanin et al.* [49] the image acquisition setup consists of a regular camera (Canon G9) which is placed at a fixed distance to a dish in which multiple plants are growing in a fluid gel. Parameters which are extracted throughout large time series are the growth rate, curvature measurements, root bend detection and the number of lateral roots. RootTrace uses a tracing approach, where the centerline of a structure is extracted similar to human behaviour when following a line with a pen. A start point is defined by the user and the root is followed until the root tip (termination point) is reached. RootTrace uses an automatic tracking procedure, normally used for tracking moving objects, called condensation. This particle-filter based method represents probability density functions by using a set of discrete weighted hypothesis. This hypothesis can be for example assumptions about a possible next position or assumptions concerning the velocity. A weight for a possible hypothesized location is assigned during evaluation and a graph representing these weights is build. The centerline of the root can be extracted by finding a way through the graph which fulfils the combination of shortest distance and highest probability. For each plant and each time series the start point has only to be chosen once [19, 36].

A tool for analysing single images of *Arabdisopsis* roots, growing in vertical Petri dishes is called EZ-Rhizo [2]. The method is semi-automatic and the user is guided through the image processing pipeline. The pipeline consists out of gray-value thresholding segmentation, noise-reduction, cropping and skeletonization. In each step the user can choose applied methods and parameter values. A re-touch step allows the user to manually correct small discontinuities, which may come up during the segmentation process. Compared to the previously described tools, this tool is developed for the analysis of root system architecture characteristics in a later stage of growing *Arabidopsis* (day 3 to day 9) [2]. A commercial software for *Arabidopsis* leaf and root phenotyping is WinRHIZO [41]. The latest pro-version (2012b) is able to extract RSA characteristics from main- and lateral- roots of *Arabidopsis* as well as some characteristics concerning the leafs and seeds (see Table 2.2 for details).

KineRoot [5] is a framework developed for the specific purpose of high-temporal and high-spatial root growth analysis. Grayscale images of *Arabidopsis* are gathered by using a compound microscope with infrared light. Time series pictures (every 5 min) are processed by using tracking of marker points. These marker points have to be set manually for each time-series set. With use of KineRoot the growth velocity as well as the relative elongation rate can be determined [5]. *Arvidsson et al.* [3] developed a phenotyping pipeline for high-throughput studies which focusses on the analysis of the leaf growth behaviour. The image acquisition is done in an imaging chamber, where pictures are taken by a camera, mounted on a robotic arm. The plants for this analysis are growing in regular 6 cm plant pots. The images are binarized using gray-value thresholding. Mathematical morphological operators are used for post-processing before the traits are extracted. The area of the leaf was of special interest. With use of this setup approximately 7000 plants can be analysed per day [3].

A more extended list of existing frameworks is available on the *Plant Image Analysis* website of *G. Lobet*[1]. The interested reader is referred to the work of *Joshua N. Cobb et al.* [12], which are reviewing an extended list of frameworks for *next-generation phenotyping*.

## 2.4 Summary

Multiple disciplines such as medicine or plant biology have an increasing need of knowledge, explaining the correlation between the genotype and the phenotype of multicellular organisms. Modern genotyping techniques make the genome analysis of different organism an effective and accurate task. The bottleneck in large-scale functional genomic studies is the effective and accurate extraction of phenotypic properties describing the phenotype of an organism.

Digital phenotyping pipelines have already been developed for different multicellular organisms like *C. elegans* or crop plants. With use of these methods the effort for biologists during phenotyping can significantly be reduced. Concerning the digital phenotyping of *Arabidopsis* the latest developments are focussing on the accurate and high-throughput extraction of quantitative traits describing the root architecture and root development. There is a lack of sophisticated methods for analysing the phenotypic expression of *Arabidopsis* in later growing stages. Further, most methods focus on *simple* traits like curvature characteristics, the length or the area of roots. More complex traits like branching patterns or branching orders are barely noted. Improvements concerning the image acquisition as well as the development of accurate tools are needed for a better understanding of the correlation between the genotype and the phenotype of multicellular organisms.

---

[1] `http://www.plant-image-analysis.org/` *(last accessed 2014/01/13)*

| Reference | Extracted Traits | Purpose |
|---|---|---|
| Subramanian et al. [49] | Root growth<br>Curvature characteristics<br>Tip angle | Root gravitropic experiments |
| *French et al.* [19, 36] | Root growth<br>Curvature characteristics<br>Bend detection<br>Number of lateral roots | Root gravitropic experiments<br>Root development studies |
| *Armengaud et al.* [2] | Root Length (geodesic, euclidean)<br>Plant angle<br>Number of lateral roots<br>Tortuosity<br>Network depth<br>Lateral roots density<br>Basal-, branched- and apical zone length<br>Lateral roots: position, length (geodesic, euclidean), angle | Root System Architecture (RSA) analysis |
| WinRHIZO [41] | Number of end-, branching-, crossing-points<br>Main & lateral root lengths (geodesic, euclidean)<br>Area<br>Seedlings count<br>Leaf length & area | RSA analysis<br>Leaf analysis |
| KineRoot [5] | Growth velocity<br>Relative elongation rate | Root development studies |
| *Arvidsson et al.* [3] | Rosette area<br>Convex hull<br>Compactness<br>Relative leaf growth rate<br>Number of leafs | Plant development |

Table 2.2: Selected digital phenotyping approaches for *Arabidopsis*.

CHAPTER 3

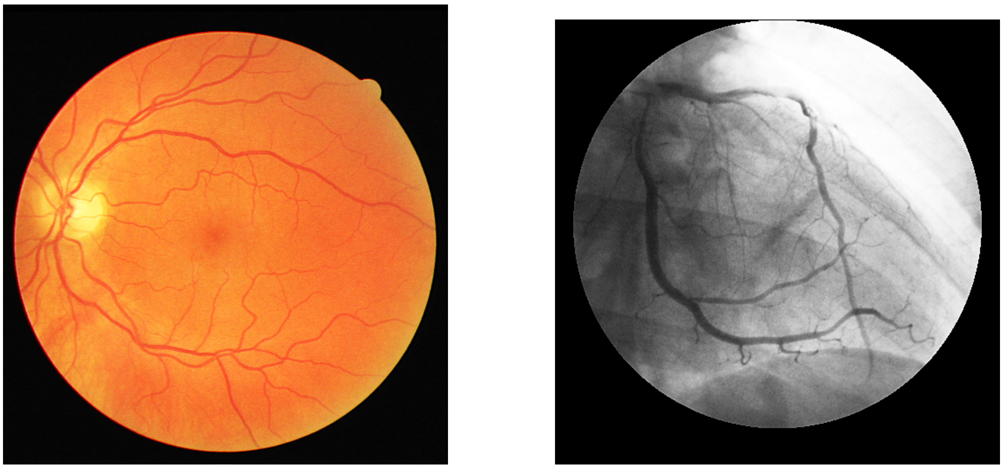
# Networks of Curvilinear Structures

The extraction and analysis of networks of curvilinear structures is one of the earliest studied problems in computer vision [54]. Their occurrence in a wide range of scale and their appearance in man-made as well as in biological/natural systems made their study attractive. While the detailed analysis of these networks in plant biology are so far limited to the analysis of “simple” traits describing the root system (see Section 2.2.2 and 2.3), approaches in other fields have already successfully studied more sophisticated characteristics, e.g. branching patterns. This chapter highlights some state-of-the-art approaches for a detailed analysis of networks of curvilinear structures from different research areas. An overview of the different fields and their relevance for this work are described in the first section of this chapter.

In Section 3.2 the different strategies to analyse networks of curvilinear structures are discussed. Approaches for the segmentation of the whole network are in focus of this section. Section 3.3 focuses on tracing approaches which are directly extracting the centerline of an network without the need of a preceding segmentation step. The benefits and drawbacks of a direct extraction of the centerline of a network of curvilinear structures is discussed in Section 3.4. In Section 3.5 approaches for labeling and grouping of trees and sub-trees in networks of curvilinear structures are discussed.

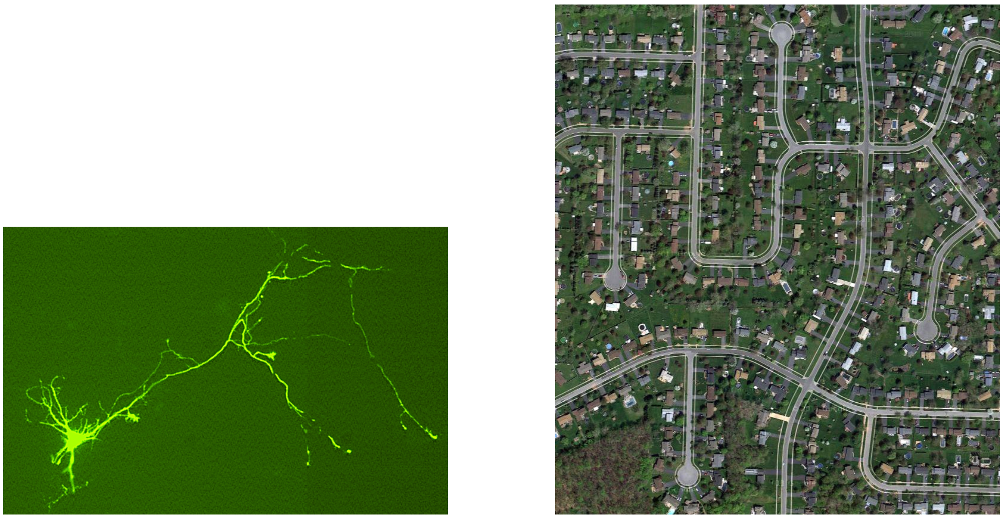
## 3.1 Overview

Curvilinear structures are formed frequently in artificial (man-made) constructions or in nature. Their appearance in networks show a huge variation in size as well as in the morphological complexity of the structure. While man-made curvilinear structures mostly appear in a large scale (e.g. roads, drawings, etc.), curvilinear structures in nature often appear in a small or even microscopic scale (e.g. biological neural networks, blood vessels, plant root system, e.g.). Due to the frequent occurrence of these particular objects, the extraction and analysis of curvilinear structures became one of the first studied problems in computer vision. Since then, the automatic extraction of these kind of structures is limited by noise and the morphological complexity of the structures [54]. Sample images containing curvilinear structures of different research areas are given in Figure 3.1.



(a) Image of the human retina from the *DRIVE (Digital Retinal Images for Vessel Extraction)* database [48].

(b) A coronary angiogram sample image from `http://www.osirix-viewer.com/datasets/` *(last accessed: 2014/01/23).*



(c) Picture of a biological neural network. The picture is part of a test set for a neuron tracer tool called NeuronJ: `http://www.imagescience.org/meijering/software/neuronj/` *(last accessed: 2014/01/23).*

(d) Aerial image of a road network [6].

Figure 3.1: Networks of curvilinear structures. Their scale in real world varies from kilometres in aerial images of roads to nanometers in microscopic images of biological neural networks.

As mentioned before, curvilinear structures can be categorized into the man-made domain and the natural/biological domain. While artificially created structures mostly follow a design principle (e.g. standardized symbols and line types in electrical schematics or drawings) which is known a-priori, there is a lack of knowledge about a design principle in biological objects. Over the years different natural objects have been studied and different models are nowadays defined for a limited amount of biological objects, e.g. blood vessels [14].

A further distinction between different types of curvilinear networks can be made between time-dependent systems and static systems. The approaches for the analysis of artificial curvilinear structures often serve a one-time use. An example of this domain is the automatic measurement of streets in aerial images for cartography or land-use planning.

In the field of plant biology root gravitropic experiments (see also Section 2.3) or root development studies [19, 36, 49] are analysing the plants' appearance over a certain period of time (e.g. minutes, days). The change of morphological attributes are often in focus of these works. The images which are analysed in this work serve a one-time use.

The study of a single piece of a curvilinear structure is often not of special interest. More often the analysis of a whole network of curvilinear structures is of interest. These networks (or often called trees) can be formed by a single object which branches into smaller objects (e.g. blood vessels [18], plants [2]) or by multiple objects (e.g. biological neural networks [35]) where each object itself can branch into smaller objects.

Quantitative morphological attributes like the tortuosity, the width or the area of these networks are extracted and used as a description of a tree. As an example, where these morphological attributes are of relevance, the field of medical imaging should be mentioned. A special attention was hereby given to the extraction of retinal blood vessels in 2D images of the eye. The change of morphological attributes of these blood vessel trees can be an indicator for different diseases (e.g. diabetes) [18].

Another interesting field which is focusing on the extraction of similar trees is the area of neuron imaging. The reconstruction of neuronal morphology can have high impact on the understanding of nervous communication processes, for example inside the brain [15].

While in medicine the use of 3D image acquisition setups (e.g. magnetic resonance tomography, computer tomography) is very common, there are fields where the image acquisition is limited to the 2D image space. For example, the plants in this work were dried and pressed for storage and transportation reasons. A 3D image acquisition setup would not overcome the problems which arise because of overlapping plant regions.

When projecting the 3D object into the 2D image space, overlapping of parts of the structure are frequent and the branching analysis of the network gets complicated. Solving these overlappings and reconstructing the original ("realistic") topological structure of the object is a challenging task in the field of computer vision [18, 35].

Due to the biological background of this work, the following sections focus on the segmentation and morphological analysis of biological curvilinear structures. Further, a special focus was set on the processing of 2D images, as the analysed images in this work are acquired with a regular DSLR (digital single-lens reflex) camera.

## 3.2 Analysis of Curvilinear Structures

For the analysis of networks of curvilinear structures, the medial axis, also denoted as skeleton or centerline, of these objects is an efficient and informative representation. With use of the skeleton and its critical points, like branching points and termination points, important attributes like length, tortuosity and branching patterns can be determined.
Different approaches to determine the centerline of an object exist in literature. Typically the centerline of an object is extracted after separating the foreground object from its background and subsequently applying a medial axis transformation to the foreground object. In contrast to these approaches, direct exploratory methods are extracting the centerline without the need of a preceding segmentation approach. While the objective and methods of the first approach are discussed in the following, the direct exploratory approach is discussed in Section 3.3. In this work, a direct exploratory approach was used to extract the centerline of the plant. Although the following approaches were not used for extracting the centerline in this work, some general assumptions and considerations in these approaches where valuable for the design of the phenotyping pipeline.
A wide range of different approaches exist for the segmentation of curvilinear structures, e.g. pattern recognition techniques, multi-scale approaches, mathematical morphology, matched filtering or model based approaches [18]. The objective of these methods is to separate fore- and background pixels in an image by applying a specific function to each pixel of an image. For example, single-value thresholding can be stated as a very simple segmentation method. All pixels beyond a specific intensity threshold are labeled as foreground, all pixels beneath that threshold are labeled as background. Regarding to the extraction of curvilinear structures more sophisticated methods are for example based on mathematical morphology [60] or focus on line detection [37].
In order to find the skeleton of a foreground object, the distance transform or thinning approaches can be used. Representing objects by its skeleton was an early studied problem in the field of computer vision, there exist several efficient ways of extracting and storing the skeleton of an object [35].
In the following, different approaches to segment networks of curvilinear structures and their methodology are discussed.

*Zana et al.* [60] define a blood vessel as a bright, local linear and piece-wise connected structure. This definition makes processing of the image using mathematical morphology suitable. A linear, 15 pixel long structuring element is used in different directions (every 15°) for a stepwise noise reduction and contrast enhancement of the tree-like structure. In a first step morphological opening in 12 directions is applied to the image. The use of geodesic reconstruction of the supremum of the opened images into the original image leads to noise reduction while the small capillaries are preserved. A sum of top-hat operation using the reconstructed image and the opened image reduces small bright noise and improves the contrast of all linear parts. A further enhancement is done to reach the final segmentation by using a Laplacian filter and an alternating filter based on morphological opening and closing by reconstruction [60]. The basic assumptions concerning the shape of a blood vessel in the given approach, influenced the design of a basic plant model of mature *Arabidopsis* in this work (see Section 5.1.2). Further, the use of mathematical morphology operators using a line structuring element in specific directions comes into use for the extraction of the ruler units in this work (see Section 4.5).

Another model-based approach is presented by *Nguyen et al.* [37]. The presented method is an extension to basic line detection techniques. At each pixel a fixed sized window is identified. Within this window the intensity profile along a rotating linear line (every $15^\circ$) is determined for each of the 12 directions. The line with the maximum average intensity value is the so called winning line and a line response value is calculated for the pixel. Drawbacks like merging too close vessels or producing false vessels nearby strong vessels is overcome by using different scales of the line and a more sophisticated calculation of the response value, which is presented in [37].

*Martínez-Pérez et al.* [33] present an approach based on region growing. Region growing is an iterative process, where based on an initial set of pixels, neighbouring pixels are merged to a specific class if a certain condition is fulfilled. The presented approach is interesting considering the extraction of networks of curvilinear structures in a noisy environment. In the given paper, the region growing process is based on two local features of the pixels which are the edge strength and ridge strength. The edge strength is measured by the magnitude of the gradient of the image. Ridge points are identified using the Hessian of the image. The maximum eigenvalue of the Hessian corresponds to the maximum principle curvature and is a parameter for the ridge strength. As the blood vessel width and strength varies over an image, the image is convolved with a set of different Gaussian kernels (different scales of $\sigma$). The features are determined for each scale and the maximum of the values is taken as the feature of a pixel. The scale space should vary from the minimum width of a blood vessel inside an image to its maximum, e.g. 2 to 20 pixels. With use of this features and pre-defined conditions all pixels in an image are iteratively classified into the region class *background* or *vessel* [33]. On basis of this segmentation approach, *Martínez-Pérez et al.* [34] developed an approach to quantify the tree morphology semi-automatically. This approach is discussed in Section 3.5.

## 3.3 Direct Exploratory Approaches

In contrast to the previous methods, direct exploration methods are only treating pixels in a neighbourhood of relevant structures instead of processing all pixels in an image. These algorithms are also often denoted as *tracing* algorithms [35]. The principle behind the tracing algorithms is quiet similar to the human behaviour when following a line with a pen. A start point is selected and the line structure is followed according to its continuity properties until a certain stopping criteria (e.g. end point) is reached. These sort of algorithms generally need an initialization step, where at least a single start point is specified. Specifying additional parameters like tracing direction or probable end points can be useful during the tracing procedure. Different methods following this principle can be found in the field of plant biology [19], neuroscience [35, 61] or medical imaging [7, 14, 26, 50, 59]. Tracing algorithms focus on the iterative extraction of the objects' skeleton without the need of a preceding segmentation step. Important features like length and width for each centerline-segment can be determined during this extraction. The centerline and its local geometrical features can already be a sufficient representation for the analysis of morphological attributes of an object. For that reason tracing is often referred as a single-pass operation [50]. A general centerline extraction procedure can be described in 5 steps. Figure 3.2 gives an overview of tracing a centerline from one specific starting point. Methods found in literature differ from each other in some of the steps and are suited to certain applications.

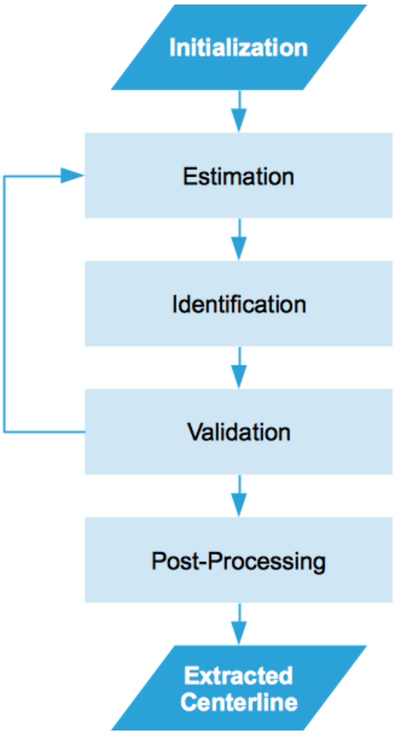


Figure 3.2: General tracing schema for centerline extraction.

***Initialization*** A few input parameters are necessary for tracing algorithms. Start points (also called seed points), end points, tracing direction and maximum width of the structure are possible parameters. While in early years of tracing development [50,53] the start point was set manually by the operator, modern approaches [7, 14] use automatic seed point identification algorithms and specify automatic stopping criteria. These automatic seed point identification algorithms are based on finding ridge candidates [14, 61] or edge pair candidates [7] on a fixed grid over the whole image. The candidates are validated against predefined heuristics and the remaining points are used as initial points for the tracing algorithm. Not all seed points are finally traced as some get eliminated during the tracing procedure. Methods with an automated seed point detection generally initialize two tracing directions for each seed point. This work falls into the group of automated seed point detection approaches (see Section 5.2.2).

***Estimation*** The next step is to find an appropriate neighbourhood in which the next centerline point can be expected. *Sun et al.* [50] propose to estimate the next position by a linear extrapolation of the current point with a self-adopting step size towards the tracing direction. The linear gray value intensity profile perpendicular to the tracing direction is taken as the neighbourhood in which the next point is expected. The length of the determined intensity profile as well as the step size is adopted to the current vessel width [50]. *Zhang et al.* [61] are using the same approach for tracing neuron cells. *Boroujeni et al.* [7] propose to use a semi-circular neighbourhood for finding the next probable centerline point. The size of the semi-circle is also adopted to the actual vessel width. Further, *Borounjeni et al.* [7] are using the output of a *vesselness filter* developed by *Frangi et al.* [16] instead of the intensity values of the image to overcome problems with noise in x-ray images. *Yin et al.* [59] expand the semi-circular neighbourhood by using a semi-elliptical scanning profile, where the length of minor axis is adapted to the width of the vessel and the major axis is adopted to the curvature change of the vessel. The intensity profile along a full circle is extracted by the approach of *Haris et al.* [26]. Using semi-circular or even circular scanning profiles instead of linear or rectangular scanning profiles guarantees the same look-ahead distance in all directions. This is especially useful when branching points or crossing points have to be identified [7].

***Identification*** The main goal in this step is to identify the next centerline point of the current segment. This is achieved by finding the boundary (pair of edge points) within the neighbourhood extracted in the previous step. The boundary detection can be done by different edge detection techniques. Finding the maximum of the first derivative of the Gaussian filtered intensity profile is used by the approaches described in [26], [7] and [61]. *Sun et al.* [50] are defining the edges by using a roll-off point based on average signal and background intensity levels of the profile. *Yin et al.* [59] are using a Bayesian method with a maximum posteriori probability criterion based on gray value statistics of the profile for defining the edge points. The midpoint of the determined boundary pair is taken as the next centerline point. Further the tracing direction is calculated as the directional vector from the current point to the just identified point. Features like the intensity value or the diameter/radius can be extracted and used for further analysis or the self-adopting step size schema.

***Validation*** During this step the tracing procedure is validated by checking predefined stopping criteria. The stopping criteria can be based on the actual tracing progress or based on considerations concerning the intensity values of the current profile. For instance, the identified centerline point is checked if it was already traced during previous iterations to avoid back tracing. Another stopping criteria, based on the intensity values of the profile, is checking the recent percent dynamic range against some predefined threshold. This is often used for identification of termination points [7, 50]. Other stopping criteria are more general, for example if one point on the scanning profile is found to be outside of the image range or if there was no valid centerline point found in the previous step [7]. If non of the stopping criteria is fulfilled the tracing procedure starts again with the estimation of the next centerline points. Otherwise the tracing for the current segment is stopped.

***Post-processing*** Depending on the application different post-processing steps can be applied to the extracted centerline. Curve smoothing, filtering of small segments or handling discontinuities are commonly used [7].

## 3.4 Direct Exploratory vs. Non-Direct Exploratory Approaches

Non-direct exploratory methods identify each pixel in an image either as foreground/object pixel or background pixel. To extract the centerline of the foreground object a skeletonization process is typically used. The result of this step is often reported as a source of error because of unwanted gaps, loops, branches or crossings. Especially while processing 2D data unwanted loops and crossings are a source of error. Due to this fact a rectification step is required, in which the skeleton is post-processed by removing loops or small branches. The critical points (termination-, branching- and crossing- points) are identified in an additional step [18, 35].

In contrast, direct exploratory techniques are extracting the centerline as a single-pass operation without the need of a preceding segmentation. During the centerline extraction local features like length and width of the currently traced segment can be extracted. Further, if we assume that the extracted tree is a connected object, information about the branching pattern can be collected while tracing the object [26]. The main disadvantage of direct exploration methods is that only structures get segmented where an initial seed point was identified. For example, if a branch is missed during tracing the segmentation result will be incomplete.

Concerning the computational efficiency, tracing approaches are reported as more effective especially if the ratio between the number of foreground pixels to the number of background pixels is rather low [18, 35].

## 3.5 Labeling and Grouping in Networks of Curvilinear Structures

After the segmentation of an image each pixel inside the image is identified either as part of a foreground object or a background object, i.e. connected components. If multiple connected components were identified during the segmentation of an image and the morphological attributes of each single component is of interest, each region has to be labeled. Labeling algorithms transform a binary image to a multi-level image, where the background is represented by zero values and the different regions are represented by distinct non-zero labels [46].

Having all objects identified, quantitative characteristics (e.g. area), also denoted as region descriptors, can be determine for each connected component. In an other step, similar regions can now be classified into groups. Members of one group are characterized by a high similarity between each other while members of distinct groups show a low similarity with each other. Similarity factors for instance can be based on topological characteristics of each region [46]. In the following this process is referred to *grouping*.

As a simple example the labeling of blood vessels into veins or arteries in an image, containing multiple blood vessel, can be considered. After the segmentation each pixel in the image is classified either as vessel or non-vessel (binary image). The labeling process identifies each blood vessel (region) inside the image and the region descriptors for each vessel can be calculated. With use of a proper region descriptor, e.g. gradient magnitude, the classification into vein or artery can be done. Blood vessels with a higher gradient magnitude are labeled as veins and blood vessels with a lower gradient magnitude are labeled as arteries [28]. Figure 3.3 shows the result of applying such a labeling on a sample image.

In the analysis of networks of curvilinear structures the labeling of trees and sub-trees is

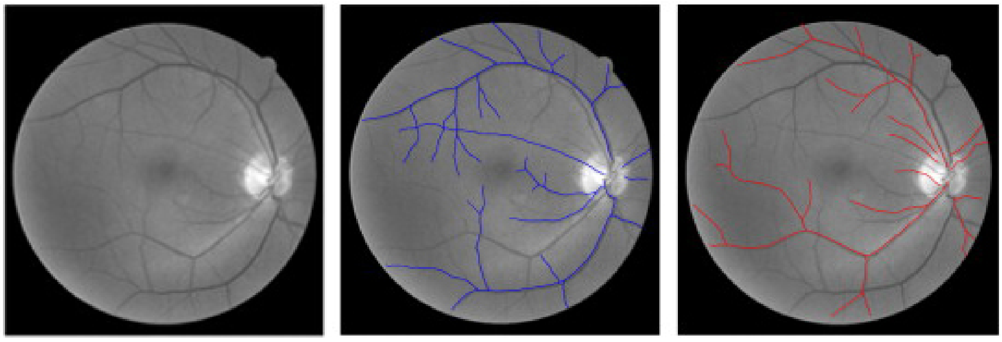


Figure 3.3: Labeling of blood vessels into veins and arteries. Veins are overlaid in blue and arteries are overlaid in red [28].

often in focus. Different methods related to the labeling and grouping of curvilinear structures were developed. A typical application of labeling and grouping concerning the networks for curvilinear structures is the determination of the branching pattern of a tree. Similar curvilinear structures are grouped at critical points (e.g. branching point, crossing points) and the realistic reconstruction of the branching pattern is possible by defining the root of the tree.

While direct exploratory approaches can label the different regions (centerline segments) already during the extraction of the centerline [28], global processing methods need to detect critical points and the different regions in an additional step [34].

*Martínez-Pérez et al.* [34] developed a semi-automatic method for grouping different trees in retinal images. After thinning and extracting critical points of the result from the segmentation process, critical points are classified as termination points, branching points and crossing points under specified conditions. For example, they assume that crossings always appear as neighbouring branching points and that vessels always split into two daughter vessel at branching points. After classifying each critical point the user selects possible start points and the structure is finally traced until a termination point is reached. With use of this method different topological properties like the Strahler number or the number of external edges and internal edges are extracted [34]. *Huang et al.* [28] use similar constraints for classifying critical points as crossing points or branching points. As tracing was used for extracting the vascular retinal tree in this method the termination points are already known. A classification of the termination points into end points and start points is done based on the location of the optical disk in the retinal image. After this classification all the retinal trees in the image are traced from the starting points to its end points. With use of an automatic seed point detection this approach is reported as fully automatic [28].

*Kai-Shun et al.* [31] also present an approach which is based on the segmentation using tracing. After tracing, an extended Kalman filter is used for grouping centerline segments at the position of critical points. The features which are used for the Kalman filter are based on the continuity of the curvature as well as on width and intensity changes. The vessel grouping process iteratively merges single segments until each individual segment is part of a group [31]. A similar procedure for a different purpose is also described by *White et al.* [58] and was already mentioned in

Section 2.2.1. In their application worms have to be classified into their developmental stage - adult, larva and egg. After applying a global image segmentation algorithm the resulting objects are split into parts using a symmetry axis algorithm. Afterwards each part is classified using a Support Vector Machine before they are finally grouped together and get their final object labels. A typical result of this application can be seen in Figure 2.1b [58].

## 3.6 Summary

The segmentation and analysis of networks of curvilinear structure has always been in focus since the early years of computer vision. Demanded from different fields like medicine or geography, researches in the field of computer vision developed different methods to extract these structures, e.g. global processing approaches, direct exploratory approaches. An accurate automated extraction is still limited by the morphological complexity of the network as well as by noise.

Most of the segmentation methods used in the digital phenotyping of plants are based on global processing methods (compare Section 2.2.2 and 2.3). Tracing methods were successfully applied for extracting blood vessel networks or biological neural networks. Especially if sophisticated features like branching patterns are in focus, the feature extraction which is already done during tracing makes direct exploration techniques a convenient choice of method. Further, critical points which are necessary to reconstruct the “realistic” network topology can already be determined during the centerline extraction.

While there already exists a great amount of approaches to extract the whole network of curvilinear structures or its skeleton, the amount of current methods for labeling and grouping of trees or sub-trees inside these networks is limited.

CHAPTER 4

# Pre-processing Strategies for Efficient Digital Plant Phenotyping

Digital high-throughput phenotyping pipelines require a finely tuned image processing pipeline to be efficient enough to analyse a large number of samples. This starts with an adequate image acquisition setup and ends with an efficient way to store and reuse the analyzed data. The following chapter covers the parts which are usually located at the head of an image processing pipeline - the image acquisition and the pre-processing.
The chapter starts with a description of the image data which was analysed during the evaluation of the framework. In this context, the image acquisition setup as well as the resulting limitations are discussed in Section 4.1. This section is followed by a theoretic description of the most relevant methods which were used during pre-processing. This includes methods such as mathematical morphology or fundamental segmentation strategies.
The monochromatic representation of the images and their alignment are discussed in Section 4.3. Section 4.4 and Section 4.5 present the two most important pre-processing modules. In Section 4.4 the automated region-of-interest identification method and its benefits for the remaining pipeline are described. In Section 4.5 an approach for an automatic extraction of the conversion factor is proposed. This conversion factor is necessary to transform the *pixel* values of the image into *real-world* metric units. The chapter is summarized in the last section.

## 4.1 Imaging Modality

The image data used in this work was provided by Wolfgang Busch from the Gregor Mendel Institute (GMI) of Molecular Plant Biology in Vienna[1]. Originally the images were taken as part of a project entitled *"The molecular basis of local adaption in A. thaliana"* led by Benjamin Brachi (Bergelson Lab, University of Chicago, US). This project involves the groups of

---

[1] `http://www.gmi.oeaw.ac.at/research-groups/wolfgang-busch` *(last accessed: 2014/01/16)*

Magnus Nordberg (GMI, AT), Joy Bergelson (University of Chicago, US), Caroline Dean (John Innes Centre in Norwich, UK) and Svante Holm (Midsweden University in Sundsvall, SE). In the following, the image acquisition setup as well as the plants' preparation and their resulting limitations are discussed.

### 4.1.1 Image Acquisition Setup

The plants which were analysed in this work, were growing outdoors in a natural environment where the environmental conditions were well known. The images of the plants were taken indoors using a Nikon D700 DSLR (digit single-lens reflex) camera. The flattened plants were put on a black velvet board to guarantee a high contrast between object and background. A wipeable plant identification (ID) sign as well as a ruler were added to the velvet for later analysis. Two flashlights were used in a darkened room to ensure equal light conditions during the whole image acquisition process. The imaging setup is illustrated in Figure 4.1. Images are taken with a resolution of 12,1 megapixels and are stored in RGB color mode with 8 bit color depth. This results in a final image size of $4284 \times 2844$ pixels which leads to a file size of 36,6 MB per image. The images are stored as TIFF (the tagged image file format) files. Sample images are shown in Section 6.1.

### 4.1.2 Limitations

Due to storage and shipping reasons the plants were dried and pressed. For this reason, some traits concerning the architecture of the plant can not be described, as their expression was already influenced before the images were taken. For instance, branching angles are not anymore well defined. Similar problems occur when projecting a 3D scene onto a 2D image plane. Nevertheless, a 3D image acquisition would not help to overcome the limitations in this work, because the problem is that the plants are already manipulated before the images are acquired. A complete and clear reconstruction of these plants would only be possible if acquiring the images by using a 3D setup on the field.
To summarize the limitations:

- The set of traits which can be extracted is limited, e.g. the branching angles are distorted.
- Parts of the flattened plant, like stems, lateral branches or siliques are overlapping. During segmentation of the 2D images these crossings appear as branching points and the "realistic" branching architecture is distorted. The more complex this branching morphology gets, the more difficult the reconstruction of the "realistic" branching architecture becomes.
- Automated determination of the plant ID, which was written by hand onto the wipeable sign, is limited to sophisticated OCR (optical character recognition) algorithms. As this was not the focus of this work, the automated determination of the plant ID is not done in this work. The plant ID in this work was set manually.

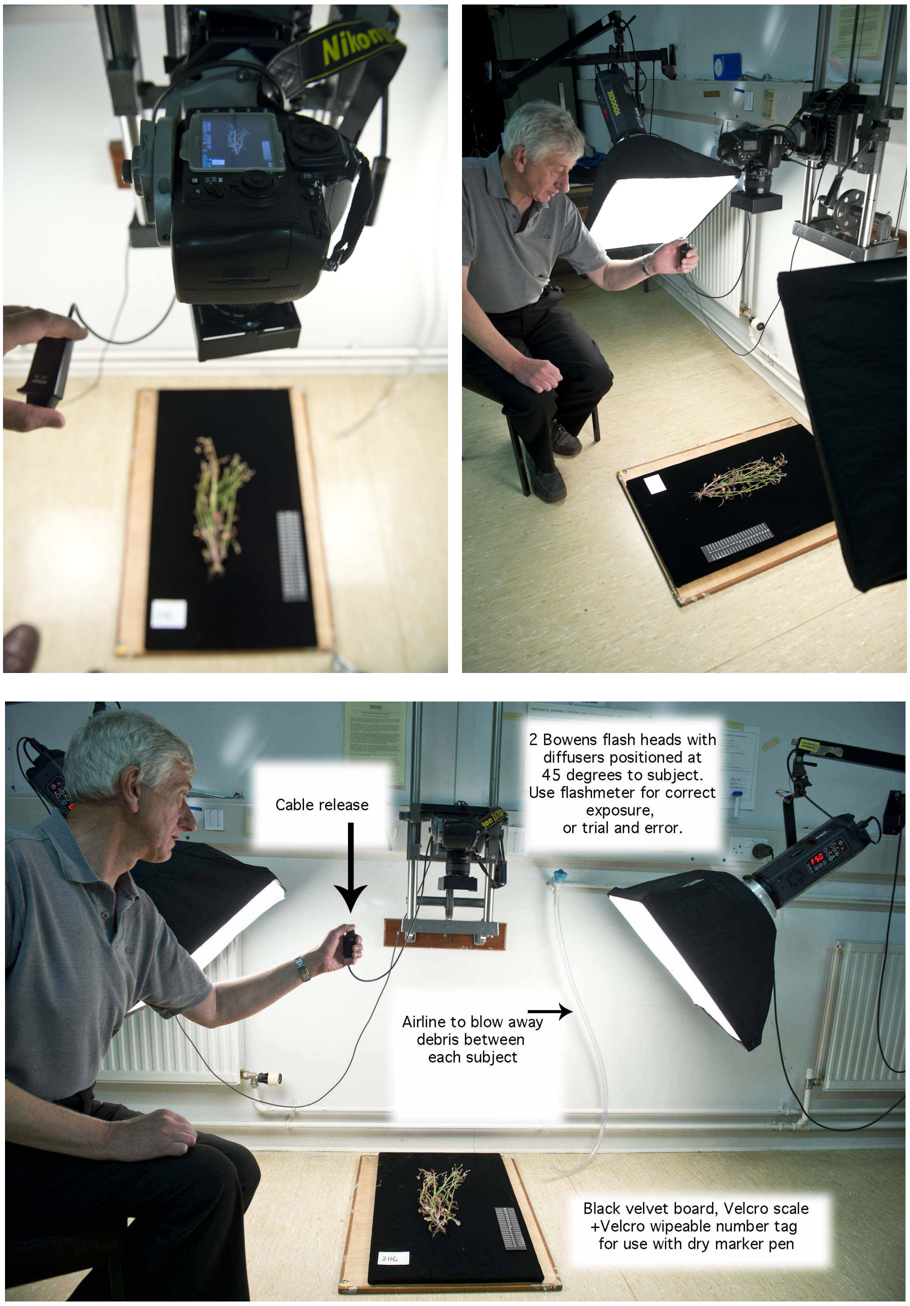

Nikon
Cable release
2 Bowens flash heads with diffusers positioned at 45 degrees to subject. Use flashmeter for correct exposure, or trial and error.
Airline to blow away debris between each subject
Black velvet board, Velcro scale +Velcro wipeable number tag for use with dry marker pen

Figure 4.1: Setup for image acquisition of mature *Arabidopsis* under controlled conditions to ensure image quality.

## 4.2 Methods for Pre-Processing

This section points out and summarizes the theory behind shared methods used during different pre-processing steps. Different color spaces where considered for an alternative representation of the original RGB plant images. These alternative color spaces are described in Section 4.2.1. Mathematical morphology was used during different parts of pre-processing. Their basic operations as well as their use are discussed in Section 4.2.2. The section is completed by the definition of gray-level thresholding image segmentation.

### 4.2.1 Color Spaces

Color can be a useful feature for differentiating between specific regions in an image. A better discrimination between colors can often be achieved by transforming the images from the original color space into another color space. Different color spaces exist and are used in practice. They differ from each other in their primary colors and their defined gamut (range of colors) [46, pp. 37-39]. Further, when searching for a monochromatic representation of a color image, the transformation into another color space and extraction of one specific color channel is common in practice [17]. A selection of the most common color spaces, which were considered as alternatives to the original RGB color space of the plant images, are given in the following. Experiments concerning the best choice of a color space are discussed in Section 6.2.

#### RGB

The RGB colour space is defined as an additive mixing of the three primary colors, which are red (*R*), green (*G*) and blue (*B*). One specific color is described by a vector containing three elements. Each element describes the intensity of one primary color. Using 8-bit color depth per channel the possible intensity values are defined between 0 and 255. The color red is then represented by the vector $[255\ 0\ 0]$. RGB color space is frequently used in storing-, processing- and coding- applications as wells as in monitors [46, pp. 37-39].

#### HSV

HSV- *Hue*, *Saturation* and *Value (Brightness)* is the triplet representing a color in the HSV color space. The intensity information is isolated from the color. A specific color can be defined by choosing pure hue ($S = 1, V = 1$) and varying (decreasing) the saturation and the brightness (value) until the desired color is reached. Hue and saturation are fitted to the human perception and for this the HSV model is a more intuitive color representation than the RGB color model. The HSV color space, represented as a cylinder or hexcone, is often used in image processing softwares' user interfaces for color picking [46, pp. 37-39].

#### L*a*b*

The L*a*b* (sometimes also called CIELAB) color space is defined by a complex non-linear transformation from the RGB color space. The *L**-channel is representing the *lightness*, while the $a*$ and $b*$ are representing the color opponent channels. The L*a*b* color space is a perceptually uniform color space. This means that little change in color or brightness leads to the same amount of change in the visual importance of a viewer. To reach this perceptual uniformity the price of increased computational effort for the transformation of an image from RGB to L*a*b* space has to be paid. The usage of L*a*b* has its roots in colorimetry where absolute values and differences of colors are studied [29,51, pp. 83-84].

### YIQ

While the Y component is describing the intensity, the I and Q components are representing the color information. Similar to RGB, YIQ is based on additive color mixing. The chrominance values I and Q are corresponding to the amounts of blue and red in the color. The Y component is providing all the information which is necessary for a representation on a monochromatic display. Similar to L*a*b*, this color model it is fitted to the human perception of color, particularly to the human sensitivity to luminance. The YIQ color model corresponds closely to the YUV color model which is used in PAL television [46, pp. 37-38].

## 4.2.2 Mathematical Morphology: Operators

Operations based on mathematical morphology are often used to enhance object structures during image pre-processing as well as for image segmentation [26,60]. It is commonly used when a specific shape of objects is an issue. Mathematical morphology can be used for binary images as well as for gray-scale images [46, pp. 657-672]. In this section the most important transformations which are used in this work are defined for binary images.

Generally, a morphological operation $\Psi$ is defined by the relation of an image $X$ and a smaller point set $B$, called a structuring element. The image $X$ as well as the structuring element $B$ are modelled as points in a N-dimensional euclidean space. The domain for 2D images is the two dimensional euclidean space $\varepsilon^2$.

The structuring element $B$ has a defined local origin $O$. During the transformation $\Psi(X)$ the structuring element $B$ is systematically moved over each pixel of the image $X$. The pixel in the image which corresponds to the local origin $O$ of the structuring element $B$ is defined as the *current pixel*. The result of the relation is stored in an output image at the position of the current pixel [46, pp. 657-672].

The basic operations of mathematical morphology are defined and discussed for their use in the following.

**Dilation** $(\oplus)$ combines two sets by using vector addition of set elements. The dilation $X \oplus B$ is the set of all possible vector additions of pairs of elements, one coming from $X$ and one coming from $B$ [25,46, pp. 661-662]:

$$X \oplus B = \left\{p \in \varepsilon^2 : p = x + b, \;\; x \in X \text{ and } b \in B\right\} \tag{4.1}$$

Dilation is increasing the object size and therefore is used to fill small holes and to narrow gaps in objects. To preserve the objects original size it is often combined with morphological eroding, which is described in the following [46, pp. 661-662].

**Erosion** $(\ominus)$ combines two sets by using vector subtraction of set elements. The erosion $X \ominus B$ is the set of all elements $p$ for which $x + b \in X$ for every $b \in B$ is valid [25, 46, pp. 662-664]:

$$A \ominus B = \left\{ p \in \varepsilon^2 : p = x + b \in X \;\; \text{for every} \;\; b \in B \right\} \tag{4.2}$$

Erosion is decreasing the object size and therefore is often used to break up structures or decompose complicated objects into several simpler ones. Erosion is the morphological dual transformation to dilation. Neither of these transformations are invertible [25, 46, pp. 662-664].

**Opening** $(\circ)$ **and closing** $(\bullet)$ are morphological transformations based on dilation and erosion. As erosion and dilation are non inverse transformations an erosion followed by dilation is not reconstructing the original image. Instead, a simplified and less detailed image of the original image is obtained. This morphological transformation of an image $X$ with a structuring element $B$ is called **opening** and is defined as follows [46, pp. 665-667]:

$$X \circ B = (X \ominus B) \oplus B \tag{4.3}$$

Applying dilation before erosion is called **closing** and is defined as:

$$X \bullet B = (X \oplus B) \ominus B \tag{4.4}$$

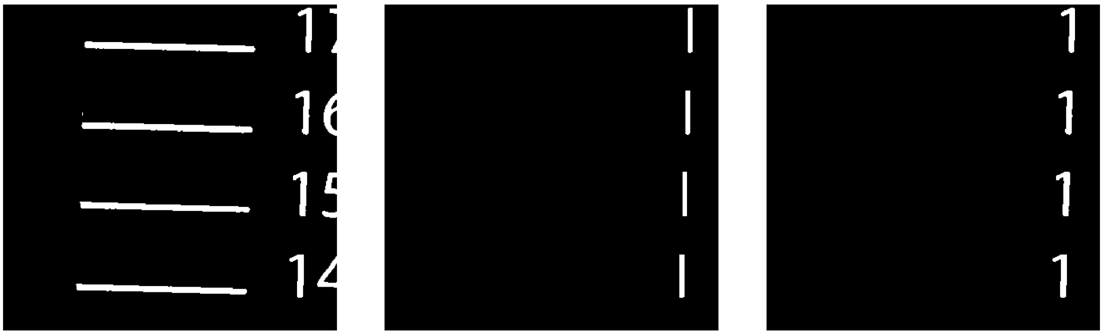
With use of an isotropic structuring element $B$ opening is used for eliminating structures in the image which are smaller than the specified structuring element. Further, small gaps get widened and contours can be smoothed. By use of morphological closing, with an isotropic structuring element, minor gaps can be closed, small holes are eliminated and little holes in a contour are closed. Both operations leave the global shape of the objects unchanged [25, 46, pp. 665-667].

**Morphological reconstruction** $R_{mask}\left(marker\right)$ is used when the exact reconstruction of a specific set of connected components in an image is wanted. In contrast to the transformations above, two images $X$ and $Y$ as well as a structuring element $B$ are used. One of the images is used as a marker (starting position) and the other image is used as a mask which constrains the transformation process. The structuring element $B$ is specifying the connectivity. By defining $X$ as mask and $Y$ as marker, the reconstruction $R_X\left(Y\right)$ of $X$ from $Y$ is defined by the following procedure ($\cap$ denotes the set intersection of two images):

```
Initialize $h_1$ to be the marker image $Y$;
Create the structuring element $B$;
repeat
    $h_{k+1} = (h_k \oplus B) \cap X$;
until $h_{k+1} = h_k$;
$R_X\left(Y\right) = h_{k+1}$
```

Conceptually, morphological reconstruction can be considered as a repeated dilation of the marker image $Y$ until the contour of the mask image $X$ is reached. The definition, given above,

uses dilation for transforming the marker image $Y$. A similar operation can also be defined by using erosion [23, pp. 656-664].
An example of morphological reconstruction is illustrated in Figure 4.2. Morphological opening and morphological reconstruction are used to extract the number *"1"* in the images. A vertical line structuring element is used, as the number *"1"* is the only object whose major direction is vertical. Morphological opening would result in an inexact segmentation of the number *1* (top is removed). Morphological reconstruction of the original image (mask) by using the opened image as a marker results in an exact segmentation of the number *"1"*.



(a) Binary image of ruler snippet.

(b) Image after morphological opening with a vertical line structuring element.

(c) Image after morphological reconstructing the original image (mask) with the opened image (marker).

Figure 4.2: Morphological opening and morphological reconstruction by opening used for segmenting the number *"1"* of the binarized ruler snippet.

### 4.2.3 Gray-level Thresholding

*"Image segmentation is one of the most important steps leading to the analysis of processed image data - its main goal is to divide an image into parts that have a strong correlation with objects or areas of the real world contained in the image"* [46, p. 175]. Gray-level thresholding, or simply *thresholding*, is the easiest process to fulfill this task. Objects or image regions can be characterized by similar intensity values. Specifying a constant intensity value or **threshhold t** which can be used to segment the image into object regions and background regions is the principle behind gray-level thresholding. The transformation of a gray-scale image $I(x,y)$ into a (segmented) binary image $BW(x,y)$ is defined as [46]:

$$BW(x,y) = \begin{cases} 1 & if\ I(x,y) > t \\ 0 & else \end{cases} \tag{4.5}$$

This means, all pixels with an intensity bigger than the given threshold *t* are set to 1, the remaining pixels are set to 0. Gray-level thresholding is limited to simple applications, but still widely used because of its simplicity and efficiency. Different methods were developed for detecting the threshold of a gray-scale image. While the manual selection of a threshold (e.g. on basis of the image histogram) is often not successful, *Otsu et al.* [38] developed a method for detecting the optimal threshold of an image based on gray-value statistics [46].

## 4.3 Color Transformation and Alignment

### Monochromatic Representation

The images obtained by the image acquisition setup (see Section 4.1.1) are stored in the RGB color space. As the color information was found to not be helpful during the further processing pipeline (see Experiments in Section 6.2), the images are reduced to a monchromatic color representation. The best contrast was achieved when transforming the RGB image into the L*a*b* color space and extracting the luminance (L) channel. The transformation into the L*a*b* color space is computationally intensive. For this reason a sufficient solution can also be achieved by only using the red (R) channel of the RGB image (see Section 6.2), which is done in this work.

### Alignment

The alignment of the images is done according to the plants' growth direction. As all the images are taken with the same acquisition setup, all the plants are rotated with 90° counter-clockwise.

## 4.4 Automated Identification of Regions-of-Interest

The images to be processed contain three different regions(objects)-of-interest which are the *plant ID sign*, the *ruler* and the *plant* itself. The automated identification of these objects is valuable for different reasons. As the plants vary a lot in size, but the image acquisition setup is always the same, there is sometimes a lot of unused background space in the images (compare Section 6.3.2). This space can be removed without loosing any information about the plant and therefore is valuable in terms of disk space consumption of the whole phenotyping pipeline. Further, if the plant images are processed using a non-direct exploratory approach (see Section 3.2) the computational cost is reduced. But, also concerning direct exploratory approaches the removing of interfering foreground objects is valuable as automated seed-point detection algorithms can be disturbed by these objects.

Concerning the ruler, its extraction is important for the determination of the conversion factor. After determining the conversion factor correctly, there is no reason to keep the image of the ruler.

The extraction of the plant ID sign is valuable for assigning an unique ID to the plant for the phenotyping pipeline. In this work, the automated analysis of the plant ID is not done as sophisticated OCR techniques would be necessary and this was not in focus of this work. Future phenotyping pipelines should use machine-written plant ID signs to make this automated processing more comfortable. Nevertheless, the plant ID sign was extracted as well to show the potential for future applications.

Due to the constant image acquisition setup a few assumptions can be stated which makes the automated identification of the objects easier:

- ***Plant ID sign:*** The sign is always placed in the bottom-left region of the image. The object's intensity values are near to maximum brightness (except the number itself) and the shape is almost quadratic. A slight variation concerning the orientation of the sign can be noticed, when looking through a set of images.
- ***Plant:*** The plant itself is the most varying object in terms of size in the image series. The plant is the most central object of interest in the images. The contrast of the plant decreases with the width of the stem. Small structures, especially in the plants end-stems, can suffer concerning this property.
- ***Ruler:*** The ruler is always placed at the very right side of the image. The structure has a rectangular shape and its orientation is corresponding with the growth direction of the plant. The ruler mainly consists of two *colors* which are white (ruler units and numbers) and mid-gray. The appearance of the ruler varies as sometimes the numbers of the ruler are visible and sometimes they disappear.

The ROI extraction is based on image processing principles like image pyramids [23, pp. 463-466], bit-plane slicing [23, pp. 117-119] and mathematical morphology (see Section 4.2.2, [23, 25, 46]. In the following, the developed method is presented and divided into the most important steps. An overview of the automated ROI identification procedure is illustrated in Figure 4.3. The explanations are illustrated in corresponding figures. Input of the ROI procedure is a gray-scale image, which originates from the extraction of the R-channel from the RGB color space (see Section 4.3).

1. ***Downsampling***: To reduce the computational cost during the pre-processing steps, the gray-scale images are downsampled to 1/8 of its original size. This corresponds to the 3rd level of an image pyramid [23, pp. 463-466]. The evaluation (see Section 6.3.1) shows that this resolution is sufficient for correctly extracting the ROIs. The downsampling is achieved by a bicubic interpolation (output pixel value is weighted average of nearest $4 \times 4$ neighbourhood pixel intensities).
2. ***Bit-plane Slicing:*** Enables to highlight the contribution of specific bits to the total image appearance [23, pp. 117-119]. An 8-bit image can be considered as a composition of eight 1-bit planes. Bit-plane 1 is containing the LSB (least significant bit) and bit-plane 8 the MSB (most significant bit). This process is illustrated in Figure 4.5. Information regarding to the foreground objects is mostly represented by the bit-planes 5 to 8. Using the sum of the bit-plane images ($I_{bit}$) 5 to 8 results in a binary image $BW$ showing all relevant foreground object:

$$BW\left(x,y\right) = min\left(1, \sum_{bit=5}^{8} I_{bit}\left(x,y\right)\right) \tag{4.6}$$

   In this case, this operation is equivalent to a gray-level thresholding at level 15. This analysis is similar to the approach by *Fraz et al.* to get a shape and orientation map for blood vessels [17].

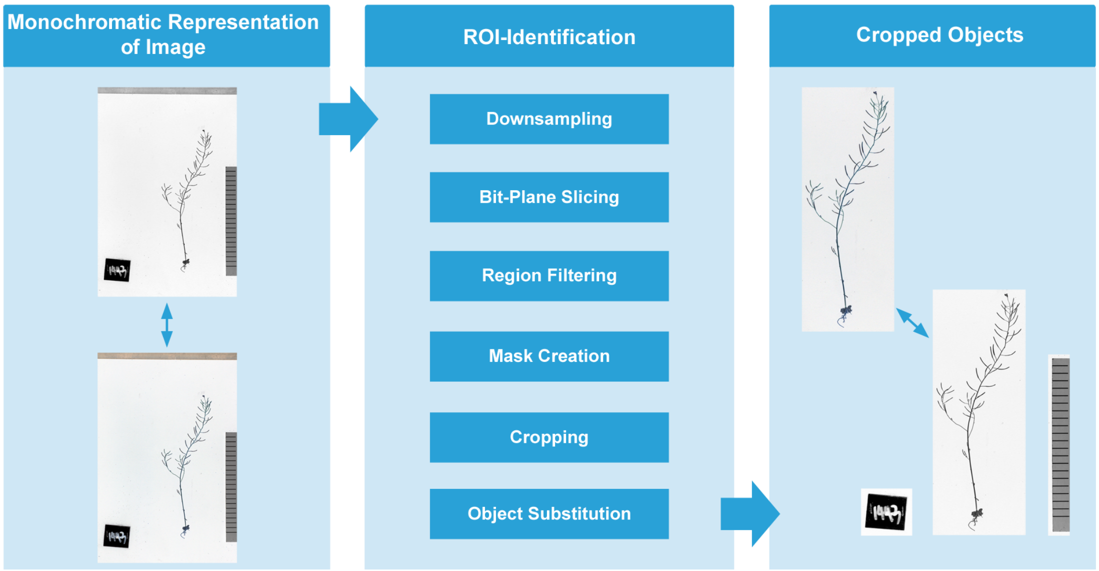


Figure 4.3: Overview of the proposed automated ROI identification process. The input of the procedure is a gray-scale image. In the end, the plant is extracted in the gray-scale mode as well as in the original RGB mode. The other objects are only extracted in the gray-scale mode. Images were inverted for illustration reasons.

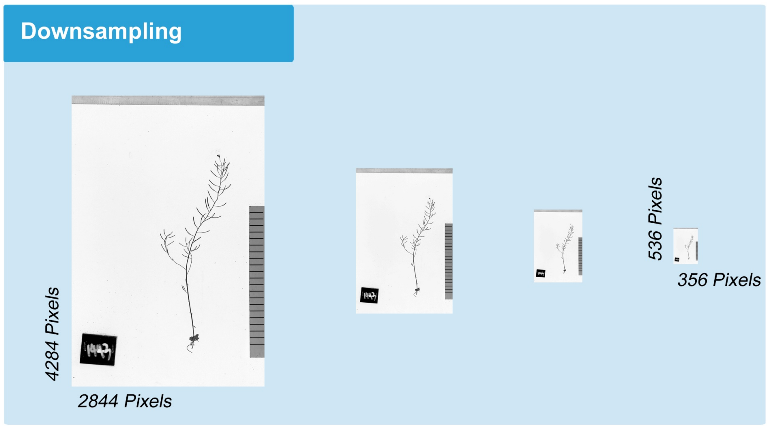


Figure 4.4: Downsampling the image to 1/8 of its original size reduces the computational cost in the following pre-processing steps. Images were inverted for illustration reasons.

3. ***Region Descriptors***: In this step, region descriptors for each connected component (object) in the binary image are determined:
    - *Perimeter*: The length of the boundary of a region.

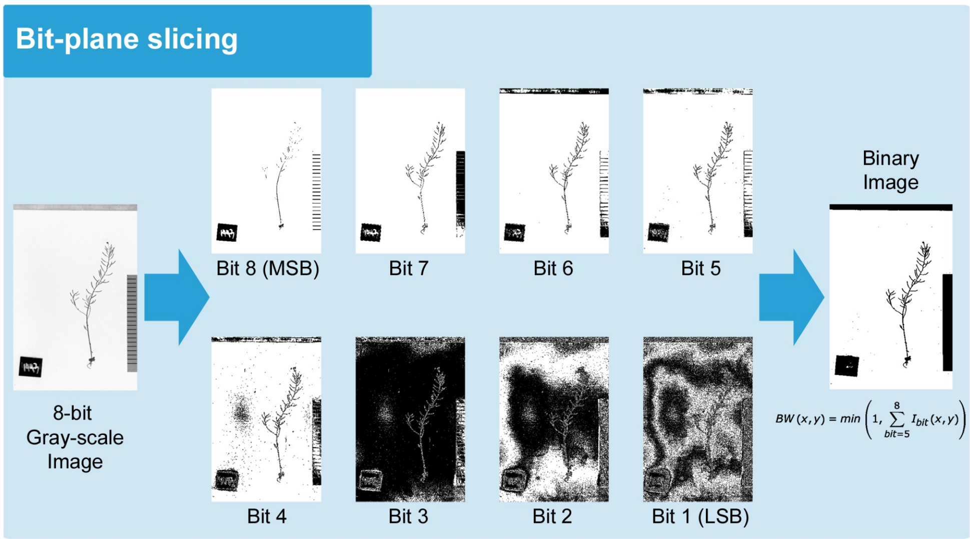


Figure 4.5: Binarization of gray-scale image on basis of bit-plane slicing. Most relevant bit-planes (bit 5 to bit 8) are summed up for a binarization of the gray-scale image. Images were inverted for illustration reasons.

- *Centroid*: The center of mass of a region.
- *Bounding Box*: The smallest rectangle containing all elements of a region.

4. ***Region Filtering:*** After the binarization during step 2 multiple foreground regions are segmented, but only three of them are defined as ROI. With use of the region descriptors small regions can be filtered by thresholding the perimeter. All objects with a perimeter measuring $5\ \%$ or less than the perimeter of the biggest region are filtered. Further, an interfering object at the top of the image (a dark ruler) which is present in the images is removed. This is done by removing the object with the lowest centroid's $y$- value (most top object). After filtering all unwanted objects, the remaining regions are sorted by their centroids' $x$-value. This step is illustrated in Figure 4.6.

5. ***Mask Creation:*** In this step the final masks to extract the objects-of-interest are created. For this reason, the bounding box of each region is filled and an individual mask is created for each object. With use of morphological dilation $\oplus$ (squared structuring element with $width = 10\ Px$), the masks are extended to be sure that no relevant structures are removed in the final extraction. This procedure is illustrated in Figure 4.7.

6. ***Cropping:*** The masks created during the previous steps are resized to the originals image size and the images are cropped. While the ruler as well as the plant ID are only extracted in gray-scale, the plant region is cropped in gray-scale as well as in the RGB mode. These final steps are shown in Figure 4.8.

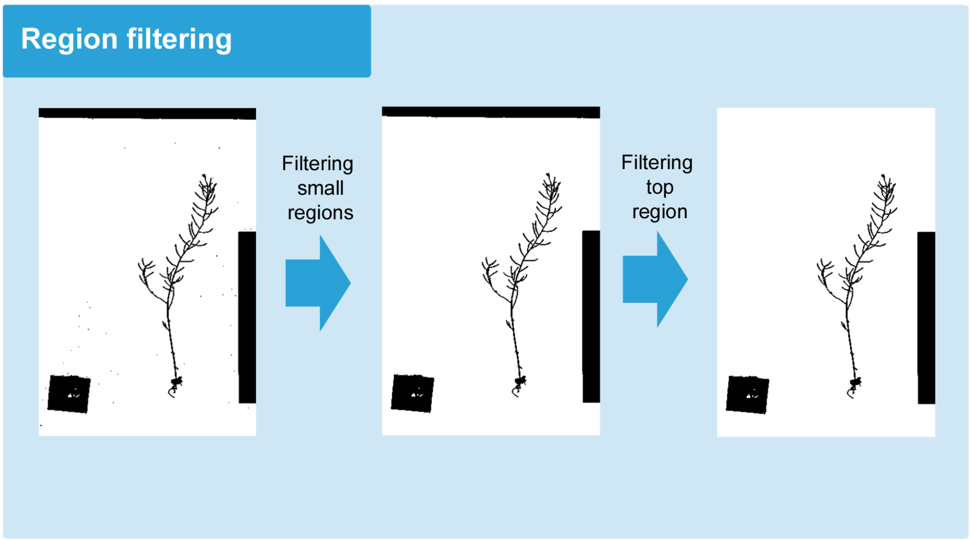


Figure 4.6: Filtering small regions and interfering top region of the binary image. All objects with a perimeter measuring $5\ \%$ or less than the perimeter of the biggest region are filtered. Further the interfering image on the top of the image is filtered on basis of the $y$-value of the regions' centroids. Images were inverted for illustration reasons.

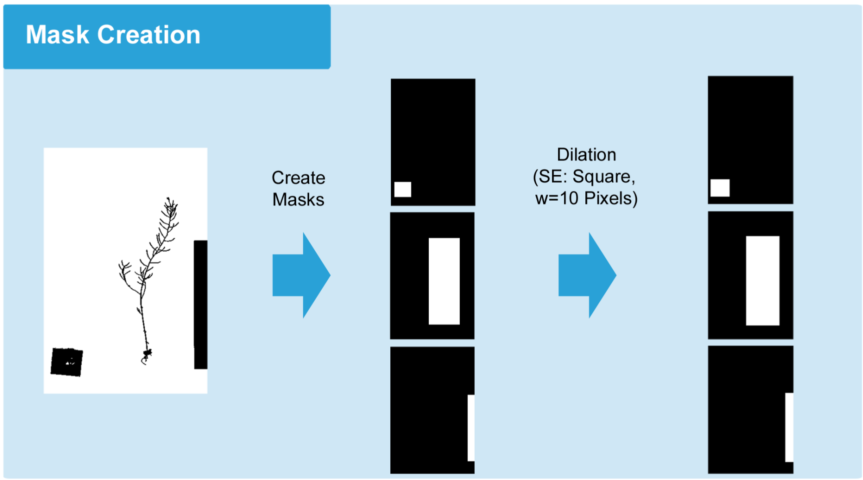


Figure 4.7: An individual mask is created for each object-of-interest. The initial creation of the mask is based on the bounding box of each object. The mask is further dilated to ensure that all relevant structures are preserved while cropping. Image showing the plant was inverted for illustration reasons.

7. ***Object Substitution:*** In some cases, the top branches of a plant are using almost the

whole $x$-range of the images. In this case, the bounding boxes of the plant and the bounding boxes of the plant ID or ruler are intersecting. Cropping the plant would result in an image, with interfering objects still present. To overcome this problem, the bounding boxes are checked for intersection. In case of intersecting bounding boxes, the intersecting region inside the plant's bounding box is filled with background values created by a simple background model. The background model is based on the mean and the standard deviation of all background elements in the gray-scale image (masked using the binarized gray-scale image).

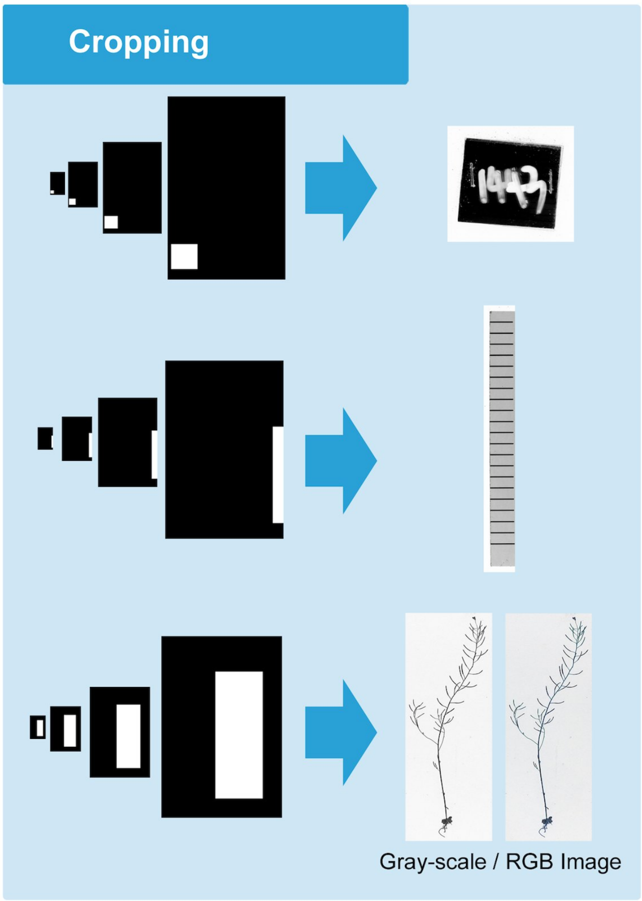


Figure 4.8: Cropping of the individual ROIs is done after resizing the masks to the image's initial size. Ruler and plant ID are only extracted in gray-scale, while the plant is extracted in gray-scale and in the original RGB space. Gray-scale images were inverted for illustration reasons.

By using the developed method, the disk space consumption for the original RGB images was reduced by 82 %. Further the cropping method did not fail in any case and all regions were identified correctly. See Section 6.3.1 and 6.3.2 for a detailed evaluation of this method.

## 4.5 Spatial Calibration

To make quantitative traits comparable to results of other studies which may have used a different image acquisition setup, the pixel units have to be transformed to *real-world metric units*, e.g. *mm*. The conversion factor which is used for this spatial calibration is determined by use of the ruler. The length of this ruler in real-world units is known and accounts 200 mm. The spatial calibration is done for every single image as there is a small variation in between the images. A method which is able to extract the conversion factor automatically is proposed and explained in the following. Figure 4.10 gives an overview of the described steps and their intermediate results.

1. ***Ruler Extraction:*** This is done by the ROI method described in Section 4.4. The cropping procedure supplies the ruler as 2D gray-scale image.

2. ***Binarization:*** To determine the conversion factor, the distance between the ruler's graduations has to be extracted. The ruler graduations appear as the brightest regions in the image. The other regions are identified as background regions with mainly two different intensity values (mid-gray and black). As the image acquisition setup provides a very constant illumination, the regions can be separated effectively by gray-level thresholding. Looking at the histogram (see Figure 4.9) of a sample ruler image, three peaks can be identified. The threshold value $t$ was set between the second and the third peak (*t=175*). The resulting binary image can be seen in Figure 4.10.

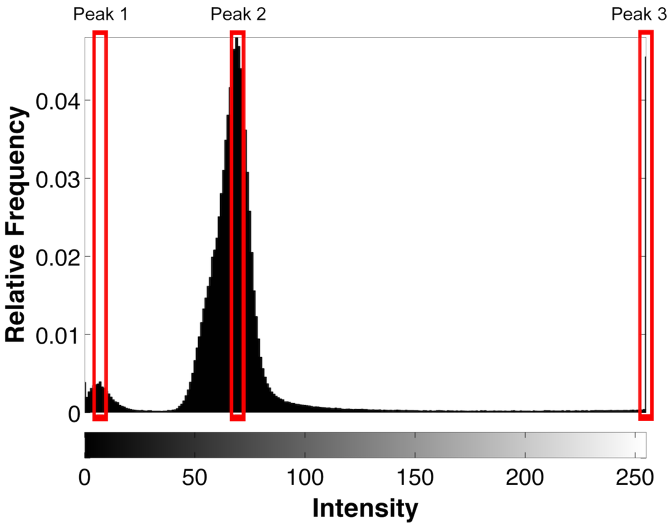


Figure 4.9: Gray-scale histogram of ruler from sample image. Three peaks can be identified around the colors *black*, *mid-gray* and *white*.

3. ***Ruler Graduations Extraction:*** As numbers or other smaller regions (noise) are present after binarization of the ruler, further processing steps are required. Extracting the straight-lined *ruler graduations* is done by *morphological reconstruction*. The marker image is

produced using morphological eroding ($\ominus$) with a horizontal, straight line as a structuring element. The width of the structuring element is set to 1/3 of the total image width. The mask is defined as the binary image from the previous step. The resulting image, which contains only the ruler graduation lines can be seen in Figure 4.10 (3).

4. ***Determination of the Average Ruler Graduations Distance:*** To determine the average euclidean distance between the ruler graduations, the *centroids* of the graduation regions are determined in a first step. Further, the *N* centroids $c_i$ are sorted by its *y*-value and the the pairwise euclidean differences

   $$d(c_{(i)}, c_{(i)+1}) = \sqrt{\left(c_{(i)} - c_{(i+1)}\right)^2} \qquad c_{(i)},\ c_{(i+1)}\ \in \mathbb{N}^2 \tag{4.7}$$

   between each adjacent centroids $c_{(i)}$ are calculated. The centroids and their corresponding distances can be seen in Figure 4.10 (4). The centroids are highlighted in purple, the euclidean distances are printed in blue. The final ruler-graduation distance $d_{RG_{Px}}$ of the image is defined as the average distance between the *(N-1)*-centroids' pairwise distances:

   $$d_{RG_{Px}} = \sum_{i=1}^{N-1} \frac{d\left(c_{(i)}, c_{(i)+1}\right)}{N-1} \tag{4.8}$$

5. ***Calculation of the Conversion Factor:*** The conversion factor $c_f$ between pixel values and real-world-units is calculated with use of the a-priori known ruler-graduation distance in mm $d_{RG_{mm}}$:

   $$c_f = \frac{d_{RG_{mm}}}{d_{RG_{Px}}} \tag{4.9}$$

   This conversion factors says, that $1\ pixel$ in the image equals $c_f\ milimeters$ in *real world.* Taking the example image of figure 4.10, $d_{RG_{Px}}$ was determined to be $97.64\ Pixels$. This equals $d_{RG_{mm}} = 10\ mm$ which means that $1\ Pixel$ in the image equals $0.1\ mm$ in real world.

The evaluation of the spatial calibration method is described in Section 6.3.3.

## 4.6 Summary

Basic image processing methods like mathematical morphology or gray-scale thresholding showed their proper use during pre-processing of the images. This can be seen as a result from the well prepared image acquiring setup and the limited complexity of the imaging modality. The limitations, resulting from the preparation of the plants, will be discussed after the quantitative traits are extracted and the plant is analyzed in detail.

The output of the pre-processing part is the cropped plant image, a unique plant ID and a conversion factor used for the spatial calibration. The pre-processing part can be seen as consistent system which is practicable on its own. For instance, it does not matter which processing technique will be used for a subsequent centerline extraction along the phenotyping pipeline. The

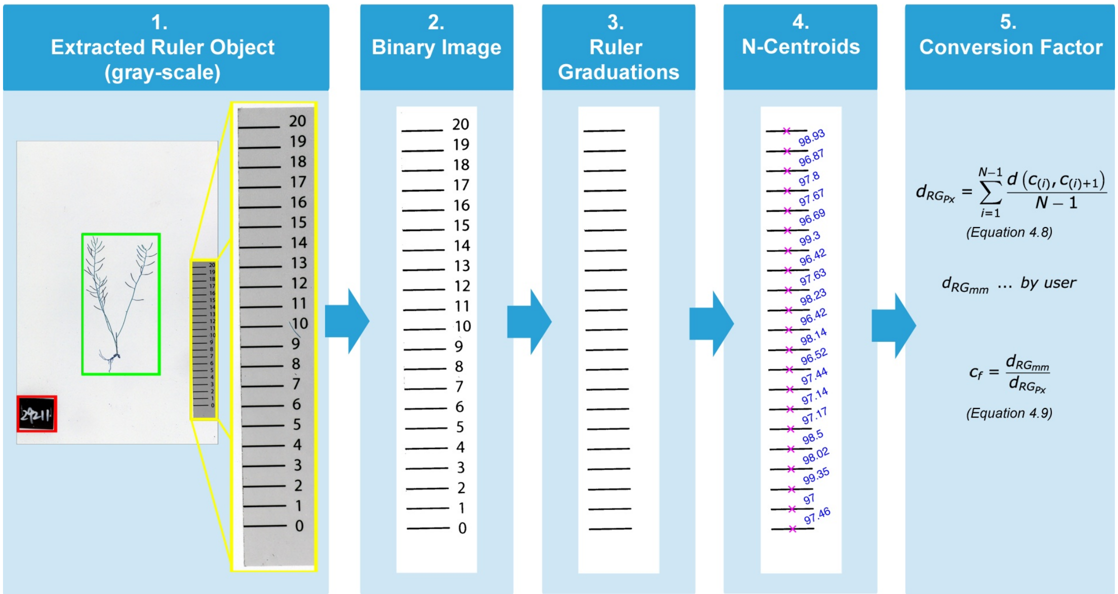


Figure 4.10: Procedure to determine the conversion factor for transforming pixel values to mm. The intermediate results are ordered accordingly to the steps 1 to 5 from the left to the right. Images were inverted for illustration reasons.

reduced disk space consumption and the appropriately cropped plant are functional for direct exploratory approaches as well as for other segmentation approaches.

In the proposed framework the plant ID is set manually. This drawback should be overcome in future phenotyping pipelines, by using machine-written number tags or QR-codes (Quick Response). The other parts of the pre-processing pipeline can be run unsupervised with the exception of quality control step. Thereby, the suggested results need to be approved by an expert. The pre-processing pipeline is developed by using well defined modules. This means that methods can easily be substituted by others or new modules can be added if the image acquisition setup changes. This makes the pre-processing strategy non-rigid and other strategies can be developed on basis of the actual pre-processing pipeline.

CHAPTER 5

# Quantitative Trait Extraction

After pre-processing the images during the first part of the digital phenotyping pipeline, the plant images are now used to extract the quantitative traits. These traits are a description of the plants' phenotype. In focus of this work are traits which describe the geometry of the plant structure as well as properties which describe the *"network topology"* of the plant. These properties are denoted as geometrical and topological traits in the following. Topological traits are for example the occurrence of different branch types (different depths) or the number of siliques on the different branch types. Geometrical traits are for example the length of the individual siliques. To extract topological and geometrical traits of sub-regions of the plant, the plant's realistic branching architecture has to be reconstructed from the 2D images.

This chapter deals with the methodology used for extracting the centerline of the plants in a first step, reconstructing the architecture in a second step and determining the quantitative traits in a final step.

In Section 5.1 the general architecture of mature *Arabidopsis* is described. Further, a model which can be derived from the general description is formulated. This model is used during the centerline extraction as well as the reconstruction procedure.

In the subsequent Section 5.2 two different methods for the identification of the main stems of a plant are presented. A manual approach as well as an automatic procedure are described.

Section 5.3 covers the description of the tracing procedure which was used for the extraction of the centerline of the plant. During the tracing procedure the skeleton as well as *local* geometrical and topological features are determined. These features combined with continuity principles of the plant's centerline segments are used for reconstructing the plant hierarchically. The hierarchical reconstruction procedure with its different layers is described in Section 5.4.

The final determination of quantitative traits is discussed in Section 5.5. The chapter is summarized in the last section.

## 5.1 Plant Model

A simplified representation of the complex structure of a plant is useful while developing methods to analyse it. Further, some constraints are useful to achieve a consistent representation of the final results. In the subsequent sections a basic knowledge concerning the plants' architecture is summarized and a model, based on this descriptions, is defined.

### 5.1.1 Plant Architecture

The term *"architecture" of a plant* is used for describing the different components of a plant regarding to space and time. Concerning the topological characteristics, this architecture can range from simple plants (low number of critical points) to complex plants (high number of critical points) [22]. A branching point, where one branch splits up into several branches is an example of a critical point.

Describing the plant's architecture is done by using topological as well as geometrical information. Topological information is describing the physical connection between different parts (e.g. siliques, leafs, flowers, branches) of the plant. Geometric information is used for a detailed description of the plant's components. Size, shape, orientation as well as the spatial location of the components are used [22]. Figure 5.1 illustrates the most relevant components of mature *Arabidopsis*. The quantitative traits extracted in this work comprise topological and geometrical information. As an example for a quantitative trait describing topological information the *siliques distribution* can be given. The siliques distribution provides information about the number of siliques located on different branches. An example for a quantitative trait concerning the geometrical information is the individual *silique's length*.

### 5.1.2 Model of Mature *Arabidopsis*

To simplify the task of analysing the architecture of a plant a few assumptions are made on basis of preliminary experiments. Such characteristics are already defined for similar curvilinear structures like blood vessels. For instance, *Zana et al.* [60] define a blood vessel as a piecewise linear and connected object, which forms a tree-like structure. *Heneghan et al.* [27] state that the cross-sectional gray level profile of blood vessels in images approximate a Gaussian shape. Further, they assume that the gray level as well as the direction of a blood vessel do not change abruptly. In addition *Heneghan et al.* [27] define that blood vessels (especially in retinal images) originate from a specific area. Similar constraints are used to define the model of mature *Arabidopsis* in this work :

i) Stems and branches are piecewise linear and their medial axis can be represented by connected line segments.

ii) The stems' and branches' crosswise intensity profile approximates a Gaussian profile.

iii) The gray level changes along stems and branches are smooth.

iv) The direction of the stems and branches does not change abruptly.

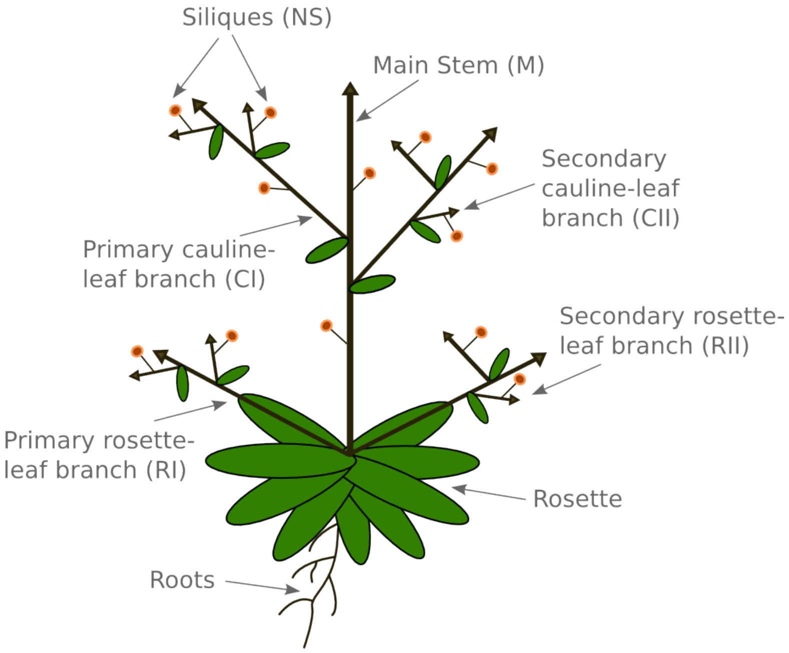

Siliques (NS)
Main Stem (M)
Secondary cauline-leaf branch (CII)
Primary cauline-leaf branch (CI)
Secondary rosette-leaf branch (RII)
Primary rosette-leaf branch (RI)
Rosette
Roots

Figure 5.1: Model of mature *Arabidopsis* plant showing the most important components of the plant.

v) The plant is connected and forms a tree-like structure.

vi) The growing direction of the plants is identical on all images of the plants.

vii) The stems' and branches' diameter decreases coming closer to the plants end points (shoots' tips) and does not change abruptly.

viii) The siliques' diameter increases before it decreases until a termination point is reached.

ix) The area of the rosette is defined as the origin of all main stems.

Figure 5.2 shows a 3D intensity profile plot of specific parts (branching point, crossing point and silique) of a plant to illustrate some of the characteristics defined above. For a better illustration the resolution of the image parts was raised and the images were filtered with a Gaussian kernel (size: $13 \times 13$, $\sigma = 1$).

## 5.2 Seed Point Identification

To reconstruct the realistic architecture of the plant, initial points specifying the origin (root) of the plant have to be specified. In this work, we are only interested in the parts of the plant which are growing above ground (shoot system of a plant). The initial points, also denoted as *seed points*, are located on the main stems nearby the rosette. Further, a tracing procedure is

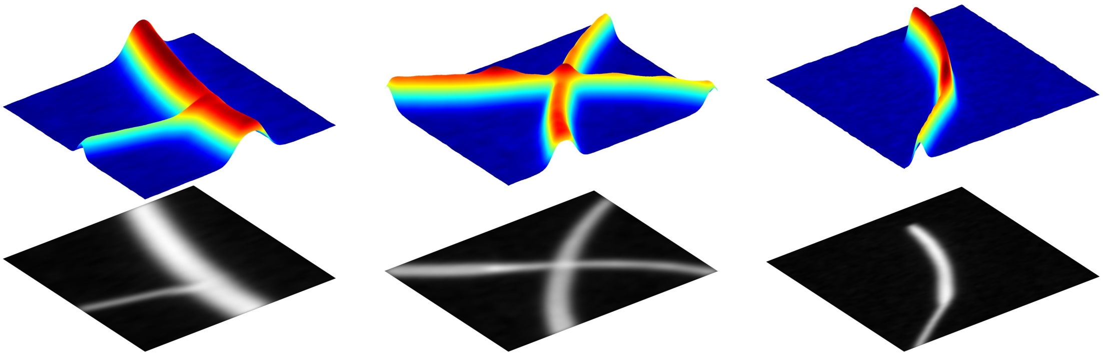


Figure 5.2: Intensity profile of specific parts of mature Arabdiopsis (from left to right): branching point, crossing point and silique. The approximated Gaussian profile as well as the smooth direction and gray level changes can be noticed. Further, the characterizing local change of the diameter along a silique is illustrated.

used to extract the centerline of the plant. As tracing is a direct exploration approach, some initial points are required for the initialization. As we are expecting the plant as a connected tree-like structure, one seed point for each main stem should be sufficient to initialize the tracing procedure.

Based on this considerations, the identification of seed points is a crucial task in this work as different parts rely on them. Methods to identify an initial set of points inside an image can be found in different tracing approaches. While early developments used a manual seed point identification done by the operator [34, 50], recent developments use automated seed point identification routines [7, 28]. In this work both principles are used.

### 5.2.1 Manual Seed Point Selection

Approaches using a manual seed point selection usually provide a graphical user interface (GUI), where the operator can manually set initial seed points. The advantages of this method are the low number of false positive seed points and the involved validation by the user. The main disadvantage is the time consumption for manually setting the points. *Sun et al.* [50] are using a manual seed point selection for the tracing of coronary arteries. *French et al.* [19, 36] use a manual seed point selection for the tracing of *Arabidopsis* roots in root growth studies.

In this work the manual seed point selection is used when the automated seed point selection procedure fails.

### 5.2.2 Automated Seed Point Selection

Seed point identification routines which are used in automated tracing approaches are focusing on the identification of ridge points or edge pairs along scan lines of a pre-defined grid. The method of *Boroujeni et al.* [8] is collecting edge points along grid lines before the edge points are validated and edge pairs are identified. For each seed point the initial direction (based on

the gradient) as well as the initial location is determined. *Delibasis et al.* [14] are choosing one ridge point in each block of a pre-defined size as possible seed point. *Huang et al.* [28] are using the location of the optical disk to identify the start points in their topological analysis of a retinal vessel network.
The method developed during this work, is based on the assumption that one seed point for each sub-tree of the network must be sufficient for a complete centerline extraction. This assumptions should hold under regular conditions: low gray-value discontinuities and little noise in the images. This means that one seed point must be identified for each main stem of the plant. As all main stems originate from the rosette, the identification of the rosette and its main stems is in focus of this method. The method is divided into two parts:

i) *Seed Point Identification:* Approximative segmentation of the rosette and main stem (seed point) identification.

ii) *Seed Point Validation:* Validation of seed points and initial feature set calculation.

### Seed Point Identification

This part of the method fulfils the task of locating the rosette in the image and identifying possible seed points located on the main stems of the plant. The rosette can be expected to be a structure with a circular or elliptical shape. The diameter of the rosette is expected to be higher than the width of the thickest stem structure of the plant. These differences in shape between the stems and the rosette are exploited to locate the rosette. An overview of the procedure and the most relevant intermediate results are shown in Figure 5.3.

The automated seed point identification method is structured as follows:

1. ***Search Area Restriction:*** The location of the rosette can be expected to be in the lower part of the image as the images are aligned according to the plant's growing direction (see Section 4.3). Therefore, the search area (possible location of rosette) is reduced to the lower third of the image (chosen empirically). During this step the plant image is resized to half of its original size to reduce the computational cost. This corresponds to the 1st level of an image pyramid [23, pp. 463-466].

2. ***Binarization:*** Similar to the binarization during the automated identification of ROIs (see Section 4.4), the binarization was done by using the sum of bit-planes 5 to 8 from the plant image, which equals a gray-value threshold of 15.

3. ***Rosette Identification:*** While stems and branches are approximated by piecewise linear line segments, the rosette with its leafs is roughly approximated by elliptical or circular shapes. Further, the rosettes' diameter can be expected bigger than the main stems' diameters. These characteristics are exploited for an approximate segmentation of the rosette. The segmentation is solved as an iterative process, where morphological opening using a flat, disk-shaped structuring element is applied to the binary image in each step. During

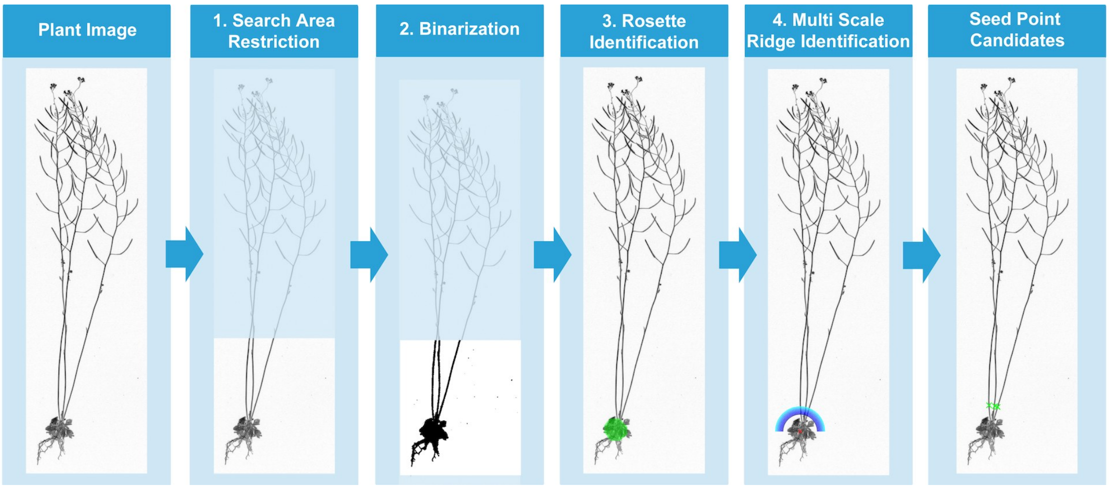

Plant Image
1. Search Area Restriction
2. Binarization
3. Rosette Identification
4. Multi Scale Ridge Identification
Seed Point Candidates

Figure 5.3: An overview of the seed point identification method which identifies possible seed points in a plant image. One seed point is identified for each main stem. Therefore, the seed point identification routine can be used for the tracing procedure as well as for reconstructing the realistic topology of the plant during the trait determination. The multi scale ridge identification process is illustrated in detail in Figure 5.4. For illustration reasons the images are shown inverted.

each step the radius of the disk is reduced by 1 $pixel$ until exactly one connected component was found in the resulting binary image. The initial radius of the disk was set to 20 $pixels$. The procedure is illustrated in Algorithm 5.1.

4. ***Multi Scale Ridge Identification:*** The segmentation of the previous step provides a rough and simplified representation of the rosette. An additional step is required to identify the seed points, they equal the main stems which originate from the rosette. On basis of the centroid $p = (p_x, p_y)$ and the equivalent diameter $d_{eqv}$ (diameter of a circle with an area equalling the rosette's region) a semi-circular neighbourhood is considered to identify possible ridge points. Specifically, a semi-circular gray-value profile $S_r(p)$ with radius $r$ from point $p$ is defined as a sequence of equally spaced samples $n$ taken along the circumference of a semi-circle [7]

$$S_r(p) = \{c_i,\ i = 0, 1, \cdots, n-1\} \tag{5.1}$$

where $c_i$ is defined as:

$$c_i = I\left(p_x + r \cdot cos\left(i \cdot \delta\theta\right), p_y + r \cdot sin\left(i \cdot \delta\theta\right)\right) \qquad \delta\theta = \frac{\pi}{n-1} \tag{5.2}$$

Since $I$ is defined in the discrete space the values of $c_i$ are determined using the nearest pixel values. The gray-value profile is convoluted with an 1-D Gaussian filter of size 13

```
Input: Binary Image of Plants' Lower Part (size $w \times l$) $BW$
Output: Binary Image of Rosette (size $w \times l$) $rosetteBW$
$diskRadius \leftarrow 20$;
$rosetteBW \leftarrow$ initialize blank rosette image as binary image with size of BW;
for $i \leftarrow 1$ to $diskRadius$ do
    $B \leftarrow$ Create flat, disk-shaped structuring element with radius diskRadius;
    $openedBW \leftarrow$ Morphological opening of $BW$ using $B$;
    $N_{CC} \leftarrow$ Determine number of connected components in $openedBW$;
    if $N_{CC} \neq 1$ then
        $diskRadius \leftarrow diskRadius - 1$;
    else
        $rosetteBW \leftarrow openedBW$;
        break;
    end
end
if $rosetteBW$ is a blank image then
    Manual seed point selection has to be done;
end
```

**Algorithm 5.1:** Procedure for the identification of the rosette in a binarized plant image.

with $\sigma = 1$. The gray-value profile is determined iteratively for different values of r, starting with $d_{eqv}$ and raised by 5 pixels during each iteration. During each iteration the local maxima along the gray-value profile are identified. The sampling rate $n$ is set to $d_{eqv} \times \pi$ pixels.

The local maxima can be assumed as possible ridge points (seed points). After 5 iterations the number of local maxima for each iteration are analysed. If there was no change in the number of local maxima during the past 5 iterations it can be assumed that the number of local maxima matches the number of main stems. This assumption is based on the topological structure of the plant. It can be observed, that in a certain region after the rosette the main stems do not tend to branch. If there was a change in number of local maxima the iterative process is continued until a maximum number of iterations is reached. The procedure is illustrated in Algorithm 5.2. Figure 5.4 shows the gray level profiles during each iteration for different radii. The pixel locations of the local maxima during the last iteration are stored as possible seed points. The seed points are validated in the second part of the automated seed point identification routine, which is described in the following section.

### Seed Point Validation

The seed points identified during the previous step are validated in an additional step. This validation includes the accurate detection of the stems' border pairs as well as the initial feature calculation for each seed point. The procedure is motivated by the procedure described by *Haris*

```
Input: Equivalent Diameter $d_{eqv}$, Centroid $p$
Output: Seed Point Candidates $seeds$
$radius \leftarrow 0,75 \cdot d_{eqv}$;
$maxIts \leftarrow 25$;
$iterator \leftarrow 1$;
$bSeeds \leftarrow false$;
while $bSeeds == false$ do
    $S_r(p) \leftarrow$ get semi-circle gray-value profile;
    Smooth $S_r(p)$ using a 1-D Gaussian filter;
    $nPeaks(iterator) \leftarrow$ get number of peaks in smoothed signal;
    $seeds(iterator) \leftarrow$ get coordinates of local maxima;
    if $iterator \geq 5$ then
        check if the number of peaks changed during the last 5 iterations;
        if no change in number of peaks then
            $seeds \leftarrow seeds(iterator)$;
            $bSeeds \leftarrow true$ ;
        else
            if $maxIts == iterator$ then
                break;
            else
                $radius \leftarrow radius + 5$;
                $iterator \leftarrow iterator + 1$;
            end
        end
    else
        $radius \leftarrow radius + 5$;
        $iterator \leftarrow iterator + 1$;
    end
end
if $bSeeds == false$ then
    Manual seed point selection has to be done;
end
```

**Algorithm 5.2:** *Multi Scale Ridge Identification:* Procedure for identifying possible seed points.

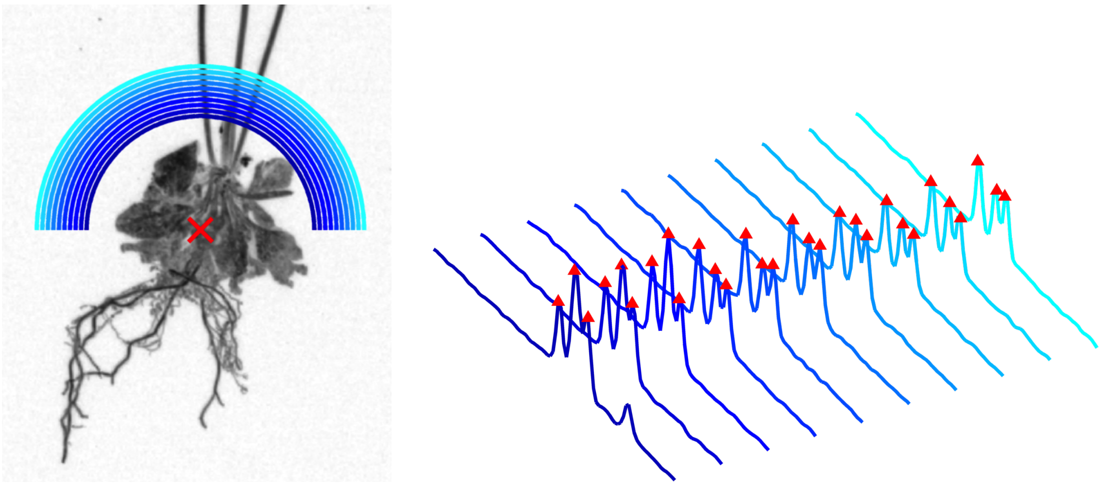


Figure 5.4: Semi-circular neighbourhood for different radii with corresponding gray-value profiles. The red cross in the left image indicates the centroid of the rosette. The red triangles in the right image are marking the local maxima which are considered as ridge points. The rosette image is inverted for illustration reasons.

*et al.* [26] used for their iterative tracing procedure. The intensity profile along the circumference of a full circle is analysed for the identification of the edge pairs. Parts of the seed point validation procedure appear also in the tracing procedure of this work (see Section 5.3), especially the way of calculating the features. The procedure for the seed point validation is structured as follows:

1. ***Ridge Identification:*** Similar to the procedure for the seed point identification a circular neighbourhood around a point $p$ is defined. This can easily be done by expanding the parameter $\delta\theta$ to $\frac{2\pi}{n-1}$ in Equation 5.2. The intensity profile is filtered by an 1-D Gaussian kernel (size 13, $\sigma = 1$).

2. ***Finding Ridge Pairs:*** It can be assumed that each seed point $p$ is surrounded by one pair of ridge points $\langle q_1,\ q_2\rangle$. One ridge point is located towards the growing direction and one ridge point is located towards the rosette. If two main stems are very close to each other, wrong ridge points can be identified. To avoid a wrong validation the best pair for each seed point is identified by determining the gradient's magnitude along the connection between each pair of ridge points. The ridge points having the lowest change in the gradients magnitude on its connection are identified as ridge pair.

3. ***Feature Determination:*** On basis of the ridge point pair $\langle q_1,\ q_2\rangle$, where $q_1$ denotes the ridge point towards the growing direction of the plant and $q_2$ denotes the ridge point closer to the rosette, the following features are calculated for each seed point (illustration see Figure 5.5) [7]:

- *Initial Direction* $\overrightarrow{u}$*:*

$$\overrightarrow{u} = \frac{q_1 - q_2}{\|q_1 - q_2\|} \tag{5.3}$$

  The initial direction $\overrightarrow{u}$ is an unit vector, where $\|\cdot\|$ denotes the magnitude of the vector.

- *Edge points* $e_L$ *and* $e_R$*:* The edge points of the stem are determined by finding the maximum gradient magnitude along a linear gray-value profile perpendicular to the initial direction at the point $q$. A linear gray-value profile $P$ is determined for the left edge $P_L$ as well as for the right edge $P_R$. The length $w$ of the profiles is set to be the initial radius which was also used for the circular gray-value profile.

- *Initial Seed Point Location* $p$*:* After localizing the edge points, the initial seed point location is refined. The initial seed point is calculated as the point between the edge points $e_L$ and $e_R$.

- *Radius* $R$*:* The radius of the stem is calculated as half of the euclidean distance between the left and the right edge point.

- *Signal Level* $s$*:* The signal level (intensity value) of the image $I$ at location $p$.

- *Percent Dynamic Range* $\gamma$*:* To calculate the percent dynamic range of the cross-sectional profile, the signal level $S_g$ between the edge points is determined as:

$$S_g = \frac{1}{e_L + e_R + 1} \left\{ \sum_{i=1}^{e_L} I\left(P_L[i]\right) + \sum_{i=2}^{e_R} I\left(P_R[i]\right) \right\} \tag{5.4}$$

  The background level $B_k$ is defined by:

$$B_k = \frac{1}{2w - e_L - e_R - 2} \left\{ \sum_{i=e_L+1}^{w} I\left(P_L[i]\right) + \sum_{i=e_R+1}^{w} I\left(P_R[i]\right) \right\} \tag{5.5}$$

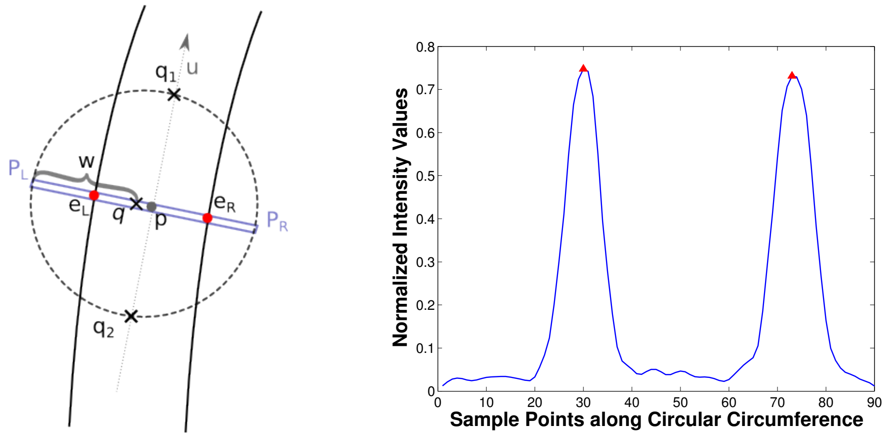
  Based on the above definition the percent dynamic range $\gamma$ is defined as:

$$\gamma = \frac{S_g - B_k}{B_k} \cdot 100\,\% \tag{5.6}$$

After determining the features for each seed point, the seed points for one plant are stored in a list which is used for initializing the tracing algorithm (described in the next section). If any of the features could not be determined the seed point is rejected. The user can validate the result of the automated seed point identification routine to spot undetected or wrongly detected main stems. The detailed evaluation of the automated seed point selection is discussed in Section 6.4.

## 5.3 Tracing of Stems and Branches

Tracing is a direct exploratory centerline extraction (segmentation) approach where a structure (often curvilinear/tubular) is followed from a starting point to an end point. The tracing procedure provides a sequence of extracted centerline points, which represent the medial axis of



(a) Illustration of the seed point validation procedure. On basis of the estimated seed point $q$ from the previous step and the determined neighbouring ridge points $q_1$ and $q_2$ the initial feature set for the validated, initial seed point $p$ are calculated.

(b) Typical gray-value profile along the circumference of the circle. Two peaks marking the neighbouring ridge points of $p$ can be identified.

Figure 5.5: Seed point validation based on the a detected ridge pair along a circular gray-value profile.

the object of interest. Local features like tracing direction, radius or intensity at the centerline points are extracted during the centerline determination and are used for travelling along the object. Further, this local features can be used for reconstructing the architecture of the plant or calculating region descriptors like lengths or widths. The process of extracting these sequences using tracing can be subdivided in four steps and is executed iteratively:

i) *Initialization:* A start point $p^k$ for the current iteration $k$ is taken from a queue of possible starting points. At the beginning of the tracing procedure the queue contains only the seed points which where identified during the seed point identification (see Section 5.2) . During the tracing procedure, additional points, e.g. branching points, are added to the queue.

ii) *Estimation:* An estimation of the location for the next centerline point along the structure is performed by using a dynamical semi-circular search window. This semi-circular search window guarantees a constant look-ahead distance in all directions.

iii) *Identification:* The next centerline point is expected to be located along the circumference of the semi-circle. To identify possible centerline points, the intensity profile along the circumference is searched for local maxima. Each local maximum is representing a ridge point at a plant's stem. These ridge points are also denoted as candidate points.

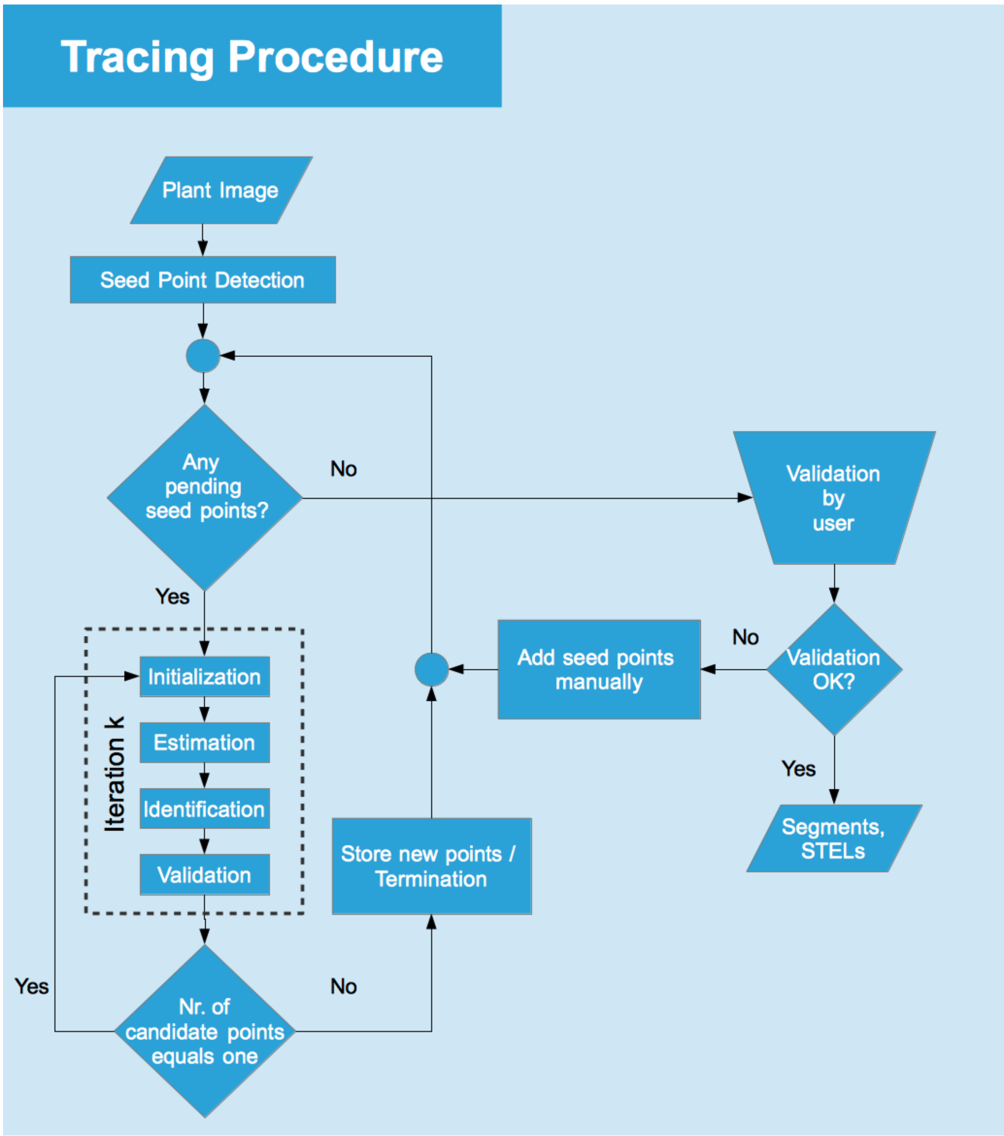


Figure 5.6: Flowchart of the tracing procedure including a manual validation step.

iv) *Validation:* The candidate points are validated and local features for each point are calculated. If only one candidate point $p^{k+1}$ is validated, this point is used to initialize the next iteration of the algorithm. Otherwise, the current iteration is terminated and the initialization starts with the next element in the queue.

The tracing method used in this work is mainly based on the procedure described by *Boroujeni et al.* [7]. In contrast to this work, *Boroujeni et al.* are using the output of an enhancement filter instead of the intensity values of an image during the tracing procedure. The enhancement filter is based on the eigenvalues of the Hessian Matrix and was proposed by *Frangi et al.* [16].

*Boroujeni et al.* are tracing blood vessel in (noisy) X-ray images. The main steps of the algorithm are described in the following sections. A flowchart of the tracing procedure is shown in Figure 5.6.

### 5.3.1 Initialization

The tracing procedure is initialized locally at validated positions. These points are so called starting points or seed points. At the beginning of the centerline extraction process the initial points are derived from the seed point identification procedure (see Section 5.2.1 and 5.2.2). Further points are identified during tracing at significant/critical positions which are e.g. branching points or crossing points.

Each of these points is defined as a start point for one centerline segment, which is a sequence of stem centerline points. Based on the initial features of the seed point the tracing procedure is invoked. The features for the following centerline points along the centerline segment are extracted iteratively. The centerline points are also denoted as *STELs (STem-ELements)*. Figure 5.7 is illustrating the terms centerline segments (or simply segments) and STELs at the location of a branching point. A STEL for the current iteration *k* is defined as a set of 7 local geometrical features and 3 topological parameters:

$$STEL_k = \left\{p^k, \overrightarrow{u}^k, R^k, e_L^k, e_R^k, s^k, \gamma^k\right\} \tag{5.7}$$

The features which are determined during each iteration are (for a detailed description see Section 5.2.2 - Seed Point Validation):

- $p^k$: Location of current $STEL_k$ in pixel values.
- $\overrightarrow{u}^k$: Directional vector (unit vector) at point $p^k$ pointing towards the current tracing direction.
- $R^k$: Radius of the stem at position $p^k$.
- $e_L^k$ and $e_R^k$: Position of edge points, which are located perpendicular to the growing direction $\overrightarrow{u}^k$ with the distance $R^k$.
- $s^k$: Signal level (normalized intensity level) of the image $I$ at point $p^k$.
- $\gamma^k$: Percent dynamic range from linear intensity profile, perpendicular to the growing direction.

Additional parameters to describe the topological connection between STELs are used for reconstructing the plant. These parameters are:

- $ID$: A unique ID for each STEL.
- $Parent$: The ID of the previous STEL is stored.
- $Type$: Specific STELs, like root or termination STELs, are tagged.

A similar approach was used by *Haris et al.* [26], where coronary angiograms where traced using a circular template and the extracted and labelled blood vessels were represented as trees.

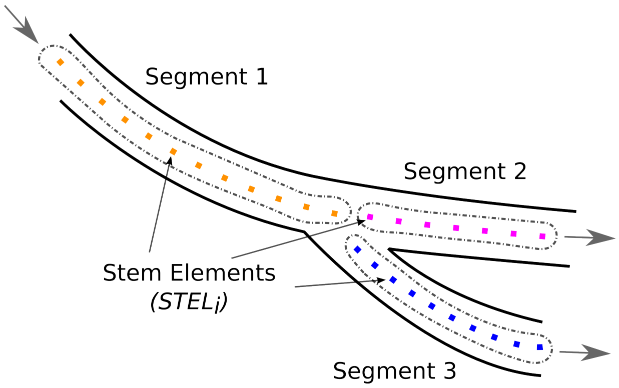


Figure 5.7: Each centerline segment consists out of multiple STELs, which are the centerline points, detected during tracing. A STEL is defined through a set of 7 features describing the geometrical characteristics of a centerline point and 3 additional parameters which are used to describe the relation between STELs.

### 5.3.2 Estimation

The tracing procedure extracts points along the centerline of a structure. The location of the next point along the structure is estimated using a semi-circular search window. The next STEL is expected to be located along the circumference of this semi-circle. Different types of templates can be found in literature for defining the neighbourhood estimator. Approaches vary from simple linear extrapolation of one point [50] to circular [26] and semi-elliptical [59] templates. This work follows the approach of *Bouroujeni et al.* [7] who propose to use a semi-circular template.

The semi-circular gray-value profile $S_r^k\left(p^k\right)$ with radius $r^k$ from point $p^k$ is defined as a sequence of equally spaced samples $n$ taken along the circumference of a semi-circle

$$S_r^k\left(p^k\right) = \{c_i,\ i = 0, 1, \cdots, n-1\} \tag{5.8}$$

where $c_i$ is defined as:

$$c_i = V\left(p_x^k + r^k \cdot cos\left(i \cdot \delta\theta\right), p_y^k + r^k \cdot sin\left(i \cdot \delta\theta\right)\right), \qquad \delta\theta = \frac{\pi}{n-1} \tag{5.9}$$

Since $V$ (normalized intensity values of the image *I*) is defined in the discrete space the values of $c_i$ are determined using the nearest pixel values. To smooth the intensity profile, the gray-values along the circumference are convolved with a 1-D Gaussian filter of size 13 with $\sigma = 1$. The sampling rate $n$ was set to be 45. The radius $r^k$ of the semi-circle must be defined “big” enough to cover all different stem widths and must be “small” enough to not detect points along neighbouring structures. For this reason, the radius $r^k$ is adapted dynamically to the current structure at each iteration of the algorithm. It is defined by

$$r^k = \rho \cdot \left[max\{R^k, R^{k+1}\right] \tag{5.10}$$

where $R^k$ and $R^{k+1}$ are the radius of the stem at the current and the next STEL position. The constant factor $\rho$ should be defined bigger than 1, to be sure to cover the whole width of a stem. In this work, $\rho$ was defined to be 2. Further, the values $r^k$ should not fall below a threshold of 5 pixels. Figure 5.8 shows two sample images of the semi-circular search window when tracing along stems.

### 5.3.3 Identification

During this step, possible ridge points along the semi-circular scanning profile are identified. In this work the normalized intensity values *V* from the image *I* are extracted along the semi-circular circumference. When considering other images with a worse image quality (e.g. more noise, less contrast) like x-ray images, other energy functions can be considered. For instance *Boroujeni et al.* [7] are using a vesselness filter by Frangi [16] to identify possible ride points at different scales of the vesselness filter. This approach is also used by *Martínez-Pérez et al.* [33]. Other approaches use morphological pre-processing methods as suggested in the work of *Haris et al.* [26].

In this work, the stems appear as bright structures on a dark background and the intensity profile across the stems tend to appear as a Gaussian-like shape (see Section 5.1.2). Due to that reason, we expect that peaks along the intensity profile represent possible centerline candidates (see sample images in Figure 5.8). The identification of relevant candidate points is done by using local maximum determination. In this work, each sample point along the intensity profile is identified as one of the following (see Figure 5.9 for illustration):

**Stem Point**

Is a point which is located along the medial axis of the stems. If more than one possible stem point $q_i$ (candidate point) is identified, a branching or crossing of stems can be expected. All candidate points are refined before they are considered as valid.

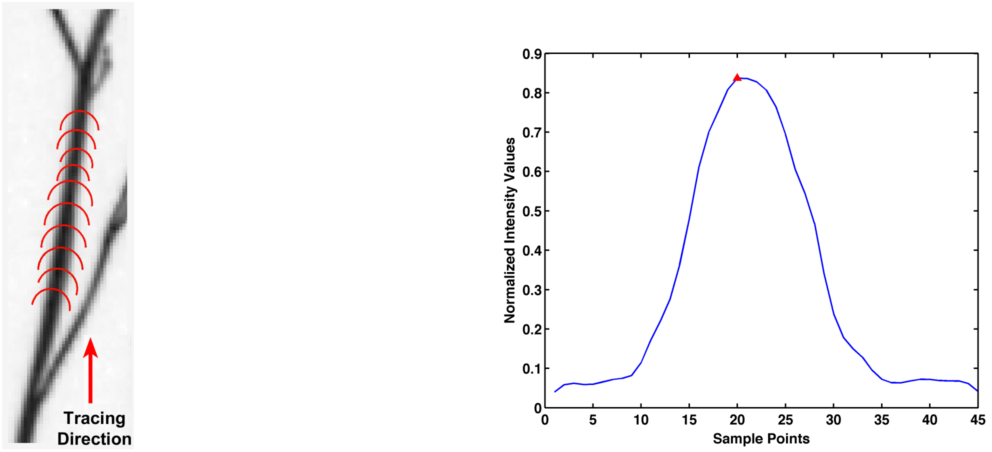
**Non-stem Point**

Is a point which belongs to the background of the image.

**Outlier-stem Point**

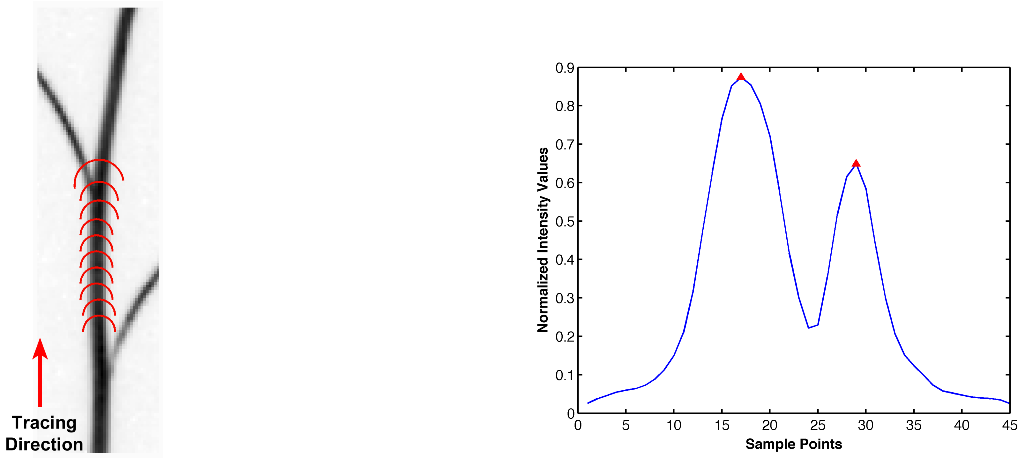
If two or more stems run very close to each other, it can occur that neighbouring stem points are identified as stem points. This mistake can be avoided by choosing a small value for $r^k$ which specifies the size of the semi-circular search window in Equation 5.2. To identify an outlier-stem point, the intensity values along the connection between $p^k$ and the candidate points $q_i$ are scanned for values below a certain threshold value *t*. This connectivity-test is also used in the approach of *Huang et al.* [28]. The threshold value *t* is set to 20 % of the normalized intensity values.

The differentiation between outlier stem points and branching/crossing stem points is useful for prioritization reasons when initializing new seed points. Branching and crossing points have a higher priority than outlier-stem points. The outlier-stem points are only stored to reach a higher completeness of the tracing algorithm and are initialized after all other critical points are



(a) Semi-circular template propagating along a regular stem, where no critical points are detected.

(b) The intensity profile along the semi-circular circumference of the dynamical search window (last position in image a). The red triangle indicates the identified candidate point during this iteration.

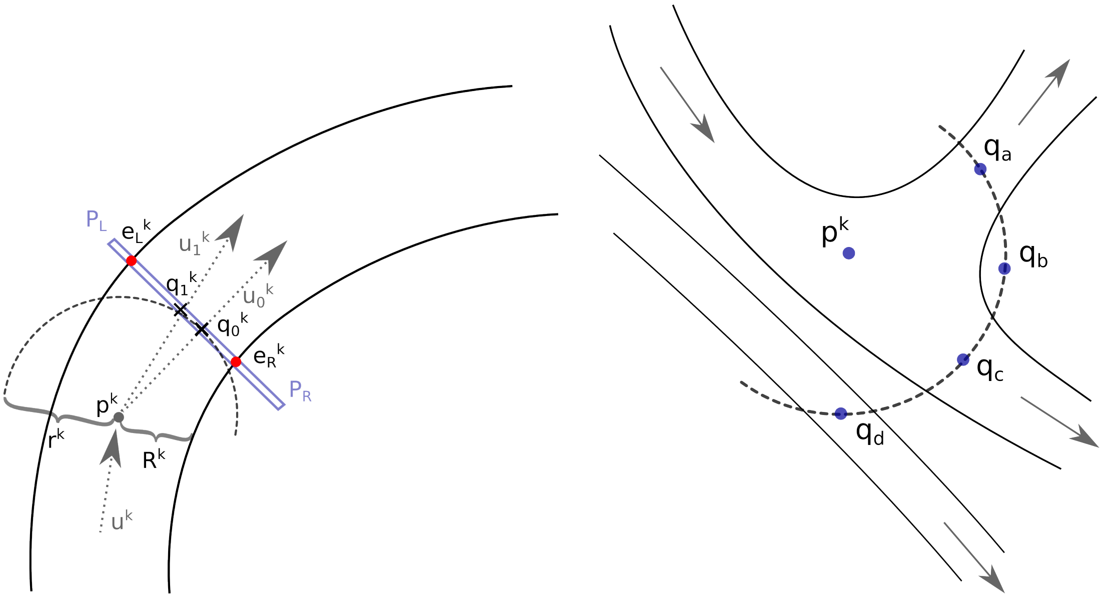


(c) Semi-circular template propagating along a regular stem and the detection of a branching point.

(d) The intensity profile along the semi-circular circumference of the dynamical search window (last position in image c). The red triangles are indicating the identified candidate points during this iteration. Two candidate points were identified.

Figure 5.8: Sample images showing the iterative process of tracing along a plants' stem.

processed. In contrast to stem points, outlier-stem points are not validated in the next step, as they are validated during their own initialization phase (see Section 5.3.7).



(a) Identification of a valid candidate point. The geometric determination and extraction of the final local feature of the candidate point are illustrated.

(b) Different types of candidate points can be identified along the circumference of the search window. Stem points ($q_a$, $q_c$) as well as outlier points ($q_d$) and non-stem points ($q_b$) can occur while tracing along a plants' stem.

Figure 5.9: Illustration of centerline point identification and determination of the STELs final features.

### 5.3.4 Validation

The accuracy of the identified ridge points is depending on the tracing direction as well as on small intensity variations across the stems. To reach a higher accuracy concerning the position of the points, the first guess (position of the local maxima) of centerline points is refined by taking into account the stem's edge points. This refinement step is proposed by different tracing approaches [7, 28, 50]. The calculations during the refinement are done for each stem point in the same way (see Figure 5.9 for illustration):

1. The geometric direction of the connecting vector between the points $p^k$ and $q_0^k$ (candidate from the estimation step) is calculated:

$$\overrightarrow{u_0^k} = \frac{p^k - q_0^k}{\|p^k - q_0^k\|} \tag{5.11}$$

2. Two linear intensity profiles $P_L$ and $P_R$ are drawn at location $q_0^k$ perpendicular to the direction $\overrightarrow{u_0^k}$. On each of these profiles an edge-point at the transition between stem and non-stem points is expected. To determine the edge points, different methods can be found in literature. *Sun et al.* [50] try to find the edge point as a roll-of point between the signal level and the background level. The approach which is used in this work is based on the gradients magnitude along the linear intensity profiles $P_L$ and $P_R$. The location of the edge point is expected to be at the position of the gradients maximum magnitude [7]. If $P_L$ and $P_R$ are the left and right intensity profile at the point $q_0^k$ with length $w$, the left edge $e_L$ point can respectively be defined as:

   $$\begin{aligned} e_L^k &= argmax\left(|\nabla_x m| + |\nabla_y m|\right), \\ m \in P_L &= \{1, 2, \ldots, w\} \end{aligned} \tag{5.12}$$

   The right edge point $e_R^k$ is calculated in the same way. $|\nabla_x m| + |\nabla_y m|$ is an estimate of the local gradient. The value for *w* is chosen to be the same value as $r^k$, as this value already satisfies the criteria to span the whole stem cross-section [7].

3. After identifying the edge points $e_L^k$ and $e_R^k$, the location of the final stem point $q_1^k$ can be calculated as the medial point between the edge points. The radius $R^{k+1}$ is also calculated during this step.

4. The final tracing direction is updated once again using the final stem points location:

   $$\overrightarrow{u_1^k} = \frac{p^k - q_1^k}{\|p^k - q_1^k\|} \tag{5.13}$$

The final position of the validated centerline and edge points can be influenced by a parameter $\alpha$ which is regulating the step size according to:

$$p^{k+1} = p^k + \alpha\overrightarrow{u_1^k} \tag{5.14}$$

If $\alpha$ equals 1, the step size equals $r^k$. A smaller step size can be considered to compensate high variations of the stem direction. In this work the step size was set to 0.9.

If all the calculations during this step were done properly the stem point is considered as valid. If only one point is validated, the tracing procedure is initialized again using the currently validated point. If none or more than one valid point was determined the tracing procedure is terminated. Further termination criteria are defined in Section 5.3.6.

### 5.3.5 Repetitious Tracing Prevention

To prevent tracing of already traced stem parts a centerline image is created and updated during the tracing procedure. The centerline image is a binary image with the size of the plants' image *I*. It is initialized empty (each pixel value is zero). During the tracing procedure the pixel values of the current stem-segments are updated and set to one. To connect two STELs, the Bresenham line drawing algorithm [10] is used. The centerline image is used in two situations for preventing repetitious tracing:

i) *During initialization* of a seed point a small neighbourhood (e.g. $5 \times 5$) in the centerline image is searched for non-zero elements. If this is the case the seed point is removed from the queue and the tracing procedure is not invoked.

ii) *During validation* of possible stem points the connection between the current centerline point $p^k$ and the candidate point $q_1^k$ is searched for non-zero elements in the centerline image. If a non-zero element is detected, the tracing procedure is terminated at the current location.

This approach is described in the approach by *Boroujen et al* [7]. An even more extended approach, where even an *edge-line* image is used for the intersection test is described in [14].

### 5.3.6 Stopping Criteria

The tracing process is an iterative process, where the skeleton of the plant structure is extracted piece by piece until certain criteria are fulfilled. These stopping criteria are used for a correct termination of the tracing process as well as to ensure an organized tracing procedure. The tracing procedure, described in the previous sub-sections is stopped if one of the following criteria is fulfilled:

i) Any of the pixels along the semi-circular neighbourhood is outside of the image range.

ii) No candidate points are determined during the current tracing iteration (last STEL is tagged as termination-STEL).

iii) No *valid* candidate points are determined during the current tracing iteration.

iv) More than one valid candidate point is found (branching- or crossing- point).

v) The connection between the current point $p^k$ and one of the current candidate points $q_i^k$ intersects the already extracted skeleton.

vi) The percent-dynamic-range falls below a certain threshold (10 $\%$ is used in this work).

### 5.3.7 Supplementary Routines

The general tracing approach is completed by supplementary routines with are fitted to the images and the application of this work. These routines are based on experiments and experiences which were made while designing and developing the tracing procedure.

#### Tracing Towards the Rossette

The seed point identification routine provides a validated start point along each main stem of a plant. These starting points are not located at the exact transition between rosette and stems as they are estimated in a certain look ahead distance. To reach a higher accuracy concerning the stem lengths, each initial seed point is initialized in two directions. One directional vector is pointing towards the rosette and one directional vector is pointing towards the growing direction

of the plant.
Before the regular tracing procedure is invoked, each initial point, which faces the rosette is used and the section between the initial point and the rosette is traced. If any critical point is found during tracing down the stem, the tracing procedure of the current segment is stopped. The critical points are *not* added to queue of initial seed points.

### Exact-Ending Routine

If there are no local maxima detected along the semi-circular intensity profile, the current point can be expected to be a termination point. As the search window (estimation phase) is analysing pixel values in a certain look-ahead distance of the current point, the current point is not implicitly the actual termination point of the segment. To provide a more exact ending of the termination segments a linear intensity profile with length $r^k$ is determined in direction of tracing. The exact termination point is defined as the pixel having the maximum gradients magnitude (compare to Section 5.3.4). The width of this last STEL is set to be 1 pixel.

### Outlier-stem Seed Point Initialization

There exist two separated queues containing seed points which are used for the initialization process of the tracing algorithm. The first queue contains start points which were detected as branching- or crossing- points. These points have a valid connection to their parent STEL. The second list of initial points contain start points which were detected as outlier-stem points. These points have no topological information to other STELs and therefore have a lower priority. Regarding to this, the points of the stem-point list are taken first to iteratively initialize the tracing algorithm. If this list is empty, the algorithm is initialized with points from the outlier-stem list. In contrast to the regular stem points, the initial parameters are not well defined for these points, as their originating elements are not known. For this reason the outlier-stem start points have to be initialized differently to the regular stem points.:

i) Seed point is validated by scanning a small neighbourhood (e.g. $5 \times 5$) at the point's position in the centerline image for non-zero elements. If a non-zero element is found the seed point is rejected.

ii) The next point along the outlier-segment can be expected to be in the neighbourhood of the seed point. As the origin of the seed point is not known a circular neighbourhood is used instead of a semi-circular neighbourhood to estimate the next STEL's position. The radius $r$ of the circular neighbourhood is fixed and was chosen to be 10 pixels. The radius should be chosen bigger than the maximum radius of the stems. The candidate points along the intensity profile are identified the same way as during the regular tracing algorithm. Outlier-stem candidates as well as already traced stem candidates are rejected.

iii) If there still exists a valid candidate point, the features are calculated the same way as during the tracing algorithm (see Section 5.3.4). The outlier-stem seed point is validated.

After initializing the algorithm by using the determined valid seed point, the regular tracing algorithm is continued.

**Manual-intervention Routine**

When both queues of initial points are empty the user can set additional seed points in case the tracing result is incomplete. This happens e.g. when critical points are not detected correctly or the contrast of the stem structures is too low. In this case, the user can choose additional seed points and the initial features for these points are calculated the same way as they are calculated during the automated seed-point detection method (see Section 5.2.2).

## 5.4 Plant Reconstruction

The tracing procedure, as described in the previous section, results in a set of *unconnected* centerline stem segments. The property of each segment is defined by geometrical and topological features which are determined during tracing. Due to the fact that branching patterns can be very complex and do not follow an a-priori known principle, the reconstruction of the connected network can not be done during the tracing procedure. Further, the restricted *local* knowledge (tracing directions, radii, intensity values, etc.) at branching and crossing points is not sufficient to reconstruct the plants' real, *overall* architecture during tracing. It should be considered that the tracing direction does not always match to the growing direction of the plant, as there is no distinction made between branching points and crossing points during tracing. To reconstruct the plants' realistic architecture, the distinction between these significant points must be made and the segments have to be grouped and connected.

The result of the tracing approach is represented as an undirected graph $G = (V, E)$, where each centerline segment is represented as a node $V_i$ and a possible connection between two centerline segments is represented as an edge $E_i$. A plant can be represented as a set of trees $T = (T_1, \ldots, T_n)$, where *n* equals the number of main stems. Each tree *T* consists of branches *B* and branching-points *BP*, i.e. $T = (B, BP)$. The plant reconstruction can be seen as a transformation from the undirected graph *G* into the set of trees *T*. For this purpose a hierarchical reconstruction procedure is proposed in this work. Each processing step during reconstruction of the plant is treated as one layer. Each layer is built upon the other without loosing the detailed information from the lower layer. This principle of a *graph pyramid* allows corrective actions until the layer which represents the highest grade of detail. Corrective actions may be taken after the reconstruction process, e.g. during a manual validation phase.

The lowest layer covers the result of the tracing algorithm represented as graph *G*. The layer at the very top is the final representation using trees *T*. With use of the final representation the topological as well as the geometrical traits can be extracted. The interior layer represents the intermediate step of grouping and connecting stem segments. An overview of this transformation process is shown in Figure 5.10. All layers and processing steps are described in the following.

### 5.4.1 Filtering

Each centerline segment which comprises less than 5 pixel is removed from the list of segments.

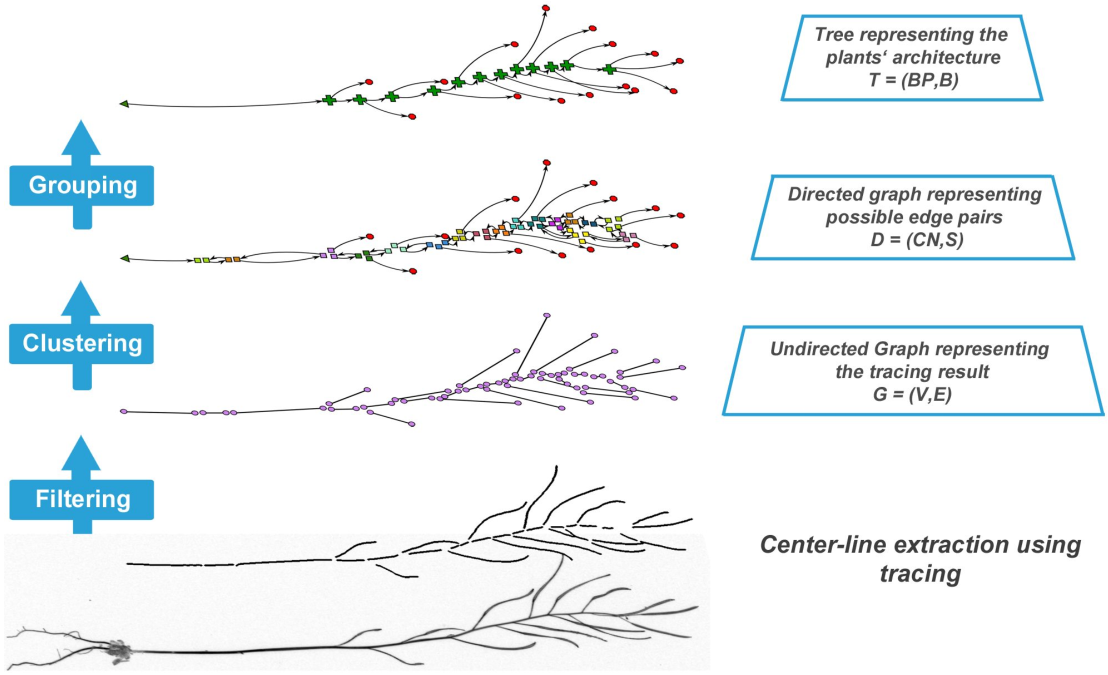

Tree representing the plants‘ architecture
T = (BP,B)
Grouping
Directed graph representing possible edge pairs
D = (CN,S)
Clustering
Undirected Graph representing the tracing result
G = (V,E)
Filtering
Center-line extraction using tracing

Figure 5.10: Hierarchical reconstruction of the plants’ realistic architecture. The base layer is representing the extracted centerline segments during the tracing procedure. Each centerline segment has two end-STELs (head and tail- represented by purple circles in the second layer). These STELs are used for describing possible connections between the centerline segments. The clustering of end-STELs into cluster nodes is based on the topological information (gained during tracing) and the positions of the STELs (distance measurement). STELs of one cluster node are represented by coloured squares in the third layer. The resulting cluster nodes contain a set of possible edge pairs (connections) between the segments. The top layer is representing the plant with use of final branches and branching points. Start nodes are represented by a green triangle, end nodes are represented by a red circle and branching points are represented by a green “+”.

### 5.4.2 Clustering

If we assume, that each centerline segment $V_i$ in the undirected graph $G$ can be connected to any other centerline segment $V_j$ we obtain an over-complete graph. As we e.g. know that a segment at the very top of the image is unlikely to be connected to a segment at the very bottom of the image, we are looking for a reduced set of possible connections. To reach this goal we build a layer on top of the graph $G$.

In this representation, the first and last STEL of each segment (representing the head and the tail of a segment) is clustered to a start- or end-STEL of another segment. The clustering has to be done in a way that STELs with a high probability of an existing connection between the them are clustered in one node. At this level each centerline segment is represented as an edge, denoted as segment *S* and a set of possible connections between adjacent centerline segments as nodes, denoted as cluster-nodes *CN*. A directed graph $D = (CN, S)$ including cycles is used for representing the data in this layer.

A possibility to create such a subset is the use of a Region-Adjacency-Graph (RAG), where each segment is handled as region and neighbouring connections between neighbouring regions are also possible connection between the segments.

Due to the gained topological relations between segments during the centerline extraction, we have the possibility to define an even more reduced set of connections. For each STEL a related parent STEL is stored during the tracing procedure. This means that also at the position of critical points like branchings or crossings we know which segments are related with each other. Further, at termination points we already now, that there are no outgoing connections from these termination STELs/segments. With use of this information we obtain a minimized connectivity representation, i.e. a smaller graph.

Due to the fact that there is no distinction between crossings and branchings made during tracing, intersections and *unrelated* stems can occur. For these segments/STELs different criteria have to be found to define possible neighbouring segments. This is done using the euclidean distance for each start- and end-STEL of a segment. If no neighbouring cluster node can be found (distance greater 50 pixels), a new cluster node is created. Using the euclidean distance to merge nearby critical points was already described in other approaches [11, 28, 34].

The clustering process is divided into the following steps:

1. The start- and end-coordinates for each segment are determined. These are the locations of each first and last STEL of a segment. The group of STELs which are either the first or the last STEL of a segment are also denoted as end-STELs.

2. Root and termination STELs are tagged during the tracing procedure. With use of these tags a cluster node is created for each STEL which is tagged either as root or termination STEL.

3. End-STELs which are related to each other are clustered into cluster nodes. A valid relation between end-STELs is either child and parent STEL or grand child and grand parent STEL.

4. All remaining end-STELs are added to the nearest cluster node based on the euclidean distance. If no nearby cluster exists (distance greater 50 pixels), a new cluster is created.

A few heuristics are defined under which end-STELs are not clustered into one cluster node:

- It is not allowed that a start and an end-STEL of one segment are in the same cluster node. This is likely to occur with small segments.

- If the amount of segments in a cluster node equals two, and both segments have the same parent or grand parent, a nearby cluster is searched and the cluster nodes are merged. Otherwise there would be no valid way to enter or leave the cluster node.

For each cluster node CN the following properties are stored:

- *CNID:* Each cluster node gets an unique ID, which is denoted as CNID.
- *EdgeIDs:* The IDs of the centerline segments which are cluster are stored. This is the set of possible edge-pairs.
- *STELIDs:* The IDs of the end-STELs which were clustered.
- *Edgecount:* Number of edges involved in the cluster node.
- *Location:* The location of the cluster node is stored as the mean location of the involved end-STELs.

### 5.4.3 Grouping

After clustering the critical points during the previous step, each cluster node must be visited and a decision which edges are connected with each other must be made. In psychology, the *Gestalt laws* are known as rules which help the human vision to perceive groups or sub-groups in a set of smaller segments/objects. A number of principles are formulated which describe the correlation between the shape of the individual segments and the human perception of building groups or sub-groups [52]. The continuity principle is used in this work for grouping sets of centerline segments. For this reason a cost term $c\left(S_i, S_j\right)$ which can be seen as travelling costs from one segment $S_i$ to another segment $S_j$ has to be determined. Edges with minimal costs are connected. Connecting two edges is also denoted as grouping. Further, a direction of each segment must be determined.

#### Continuity Principle

One of the most obvious characteristics of the plants' architecture is the small grade of change relating to the curvature, the intensity level as well as the radius of the stems. This continuity properties can be used for grouping segments inside a cluster node. In detail, the following geometric terms are calculated for an edge pair.

**Edge Direction Similarity**

The edge direction similarity is determined as the angular difference between two branches in degrees [55]. This can be considered as the term with the highest relevance when connecting an edge pair. The angle $\theta\left(S_1, S_2\right)$ between two segments $S_1$, $S_2$ and their directions $\overrightarrow{u}_1$, $\overrightarrow{u}_2$ is defined as:

$$\theta\left(S_1, S_2\right) = acos\left(\frac{\overrightarrow{u}_1 \bullet \overrightarrow{u}_2}{|\overrightarrow{u}_1| \cdot |\overrightarrow{u}_1|}\right) \tag{5.15}$$

The "•" operator is denoting the dot product between two vectors. As unit vectors are used as directional vectors for each segment, the denominator will always be 1 and therefore can be excluded from the calculation. Figure 5.11a is illustrating the edge direction similarity.

**Tortuosity**

The direction similarity does not include characteristics concerning the curvature. To treat these characteristic the tortuosity of the final connected branch $B_{12}$ is calculated. The tortuosity is calculated as the ratio between the path length $l_{12}$ and euclidean distance between the endpoints $d_{12}$ of the resulting branch $B_{12}$ [55]:

$$tor\left(S_1, S_2\right) = \frac{l_{12}}{d_{12}} \tag{5.16}$$

The tortuosity of a straight line equals 1. An illustration of the tortuosity is shown in Figure 5.11b.

**Edge Linkage Distance**

Is the euclidean distance $d_l$ between the location of the end-STEL of branch $S_1$ and the location of the start-STEL of branch $S_2$ (or vice versa). An illustration is shown in Figure 5.11c

**More Characteristics**

Other characteristics which can be considered for modelling the cost function to connect two branches are the *width consistency* [55] or the *intensity consistency*. Both characteristics are not used in the final cost function of this work.

**Cost Function**

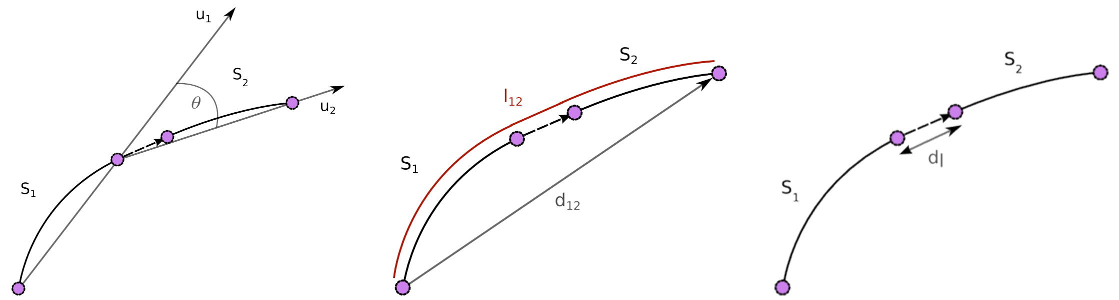
The final cost function used in this work was modelled empirically. During the first trials only the edge direction similarity was used to find the best edge pairs in a cluster node. A few cases, especially when side branches branch off from the main stem, showed that the cost function will fail. For this reason an additional term using the tortuosity and the edge linkage distance was added. As the value range of the tortuosity term is very low, the exponential function is used to model the tortuosity term. The final cost $c\left(S_1, S_2\right)$ function for connecting two segments $S_1$ and $S_2$ is defined as:

$$c\left(S_1, S_2\right) = \theta\left(S_1, S_2\right) + d_l \cdot \exp^{tor\left(S_1, S_2\right)} \tag{5.17}$$

**Grouping Process**

The grouping procedure is initialized using a directed graph including cycles $D = (CN, S)$. For all segments, except the terminating one, the directions are unknown. During the grouping procedure the directed graph *D* is transformed into a set of trees *T* as branching points and crossing points are resolved. The realistic branching topology of the plant is resolved at the end of this step. To solve this task, a local approach is used in which each cluster node *CN* is visited and resolved. Another approach would be to solve this problem globally with use of graph based approaches, e.g. minimum spanning tree.

The advantage of a local grouping procedure is that less combinations of edge-pair scores have to be determined, as some combinations can be rejected due to updated directional information. The disadvantage is that wrong local decisions can influence other local decisions and the overall reconstruction fails. The idea behind the developed grouping procedure is similar to the



(a) Edge direction similarity between two segments $S_1$ and $S_2$, calculated as the angle between the directional vectors $\overrightarrow{u}_1$ and $\overrightarrow{u}_2$.

(b) The tortuosity of the resulting branch $B_{12}$ is calculated as the ratio between the path length $l_{12}$ and the euclidean distance between the segments end points $d_{12}$.

(c) The linkage distance is calculated as the euclidean distance between two branches end points.

Figure 5.11: Continuity properties which are used to model the costs of grouping two segments $S_1$ and $S_2$.

approach presented by *Quek et al.* [42]. Their idea to segment vessel trees is to propagate a wave towards the flow direction of the blood vessel and calculating a cost function along the vessel structure. After all end points were reached the wave (cost function) is traced back and evaluated until the initial position of the vessel tree. In this way the branching structure of the vessel can be reconstructed [42].

The approach developed in this work is based on a similar idea. The branching structure of the exterior regions of the plant can be expected to be less complex (terminating branches) than those of the interior regions of the plant. For this reason the cluster nodes are processed from the outside to the inside, from the top to the bottom of the plant. Further, the cluster nodes are visited from less critical to more critical to encounter the disadvantage of propagating errors. The definition of cluster node complexity is discussed in the following.

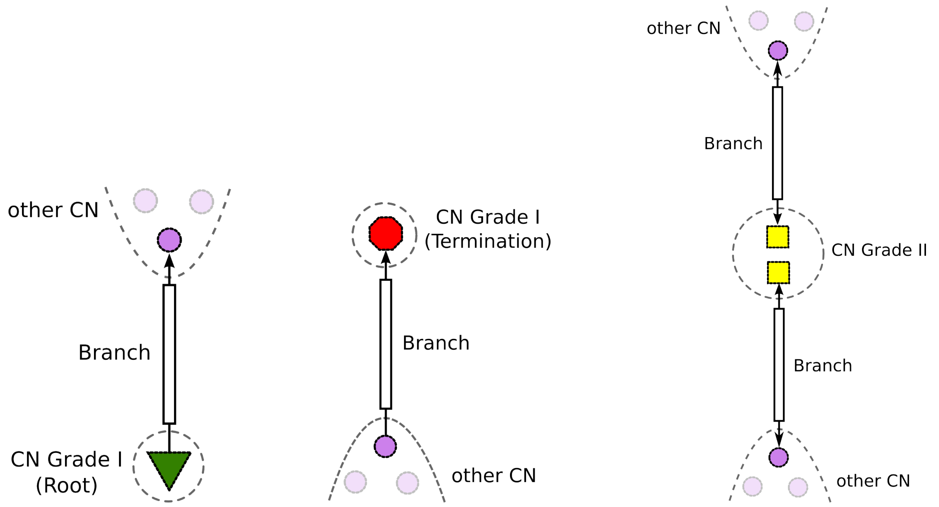
### Cluster Node Complexity

The amount of possible edge pairs is depending on the amount of edges which are clustered in a cluster node. This relation can be expressed by the binomial coefficient $\binom{n}{k}$, where *n* is denoting the number of edges in a cluster node and *k* is 2 (pairs of edges). This means, if we consider a cluster node with 3 edges, we obtain 3 different possible edge pairs. Considering a cluster node with 4 edges, we obtain 6 different configurations. The more configurations are possible in a cluster node, the lower is the probability for a correct decision (if we assume that each configuration has the same probability). Therefore the complexity of correctly solving a cluster node depends on the amount of edges involved in the cluster. Thus the grade of complexity is assigned for each cluster node. In detail, the cluster nodes are divided into cluster nodes of grade I to IV+. Grade I means that only one edge is involved (termination branches), grade IV+ means that four or more edges are involved in the cluster node.

Further, the grade of complexity in a cluster node is lowered if the direction of any of the involved edges is already known. For cluster nodes of grade III and higher, it is required that the direction of at least one of the edges is already known before the cluster is solved. For each grade of cluster node some specific topological configurations appear repeatedly. Constraints concerning this specific configurations as well as the different grades itself are discussed in the following.

### Cluster Nodes Grade I

Segments which are included in cluster nodes grade I are either root segments or termination segments. Generally, root and termination segments are already identified during the tracing procedure. Some additional termination segments can come up during the clustering process, e.g. because of filtering of small segments. The directions of the termination segments is validated while visiting all cluster nodes grade I. Figure 5.12a and 5.12b are illustrating the typical configuration of cluster nodes with grade I.



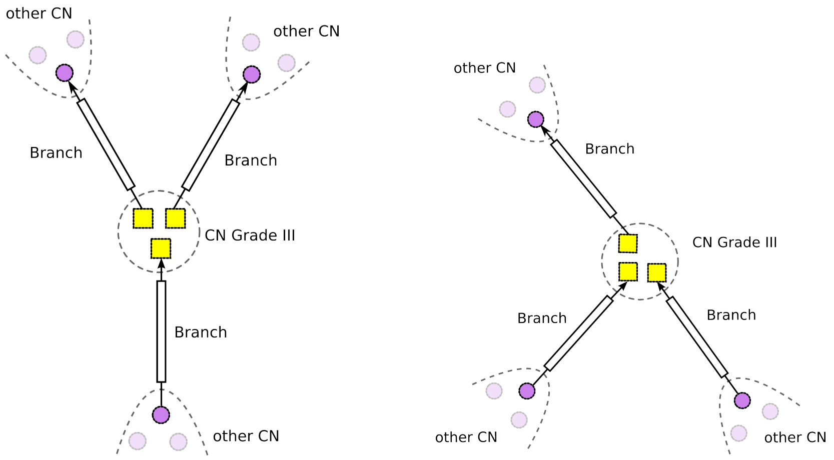
(a) Root segments are defined during the automatic seed point intitialization.

(b) Termination segments are identified during the tracing procedure as well as during the cluster building process.

(c) The connection between the STELs in the cluster node of grade two is checked for non-stem points. Further, the score of the resulting branch is determined to validate the connection process.

Figure 5.12: Typical segments which are part of cluster nodes grade I are termination and root branches. The direction is already defined during the tracing procedure. Cluster nodes of grade II typically occur at discontinuities, where e.g. small segments were filtered.

**Cluster Nodes Grade II** These cluster nodes occur due to filtering of small segments or tracing irregularities. Segments which are involved are regularly discontinuities of one branch. During visiting these type of cluster node, the connection (intensity profile) between the corresponding STELs is validated if it undercuts a threshold value of 20 %. Further, the costs for merging the segments is determined. If their is no connection between the points or the score



(a) A cluster node grade III resolved as a branching point, where one stem branches off of another branch. The direction of at least one segment (branch) has to be known for resolving these type of cluster nodes.

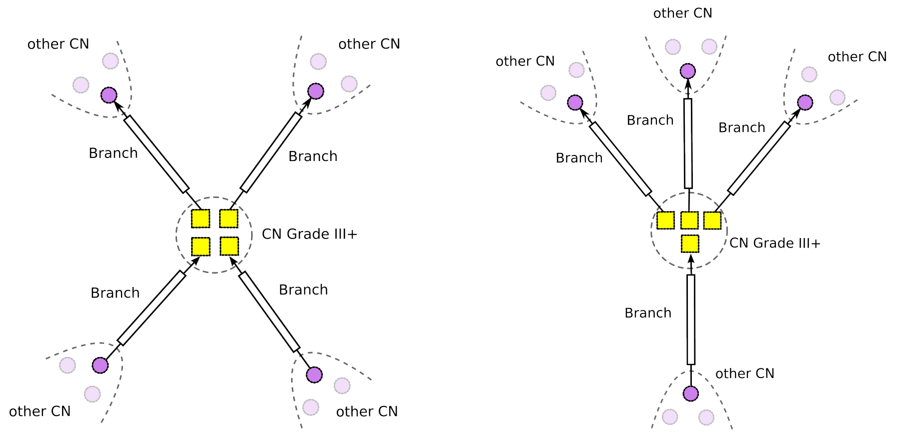
(b) A cluster node grade III resolved as terminating point. This occurs when one stem is terminating nearby another stem.

Figure 5.13: Most frequent type of cluster node is the cluster node with grade III complexity. Generally this configuration can be resolved as a branching or a termination.

exceeds a given score-threshold $s_c$, the neighbourhood is checked for the occurrence of other cluster nodes. If neighbouring cluster nodes can be found the cluster nodes are merged. The score-threshold $s_c$ is set to *165* in this work. If one of the involved branches was already validated as a termination branch, the final direction of the combined branch is prescribed. In Figure 5.12c a symbolic illustration of a cluster node grade II is shown.

**Cluster Nodes Grade III** Three segments occur regularly if one stem branches off from another stem or if one stem is terminating (overlapping) on another branch (illustrated in Figure 5.13). The final configuration can be determined by calculating the score for all possible edge pairs. If two branches are already known, the direction of the remaining edge is determined by finding the best edge pair with the unknown branch. If the directions of all branches are already known, the decision between *branching* and *termination* can be achieved by determining the degree of the cluster node. A negative degree (two outgoing edges) is indicating a branching, a positive degree is indicating a termination. To minimize the amount of operations, necessary for calculating the edge pair scores, cluster nodes which contain already a maximum of predefined branches are handled first. Figure 5.13 is illustrating a termination as well as a branching configuration of a cluster node with grade III.

**Cluster Nodes Grade IV+** These type of clusters contain four or even more than four segments. A sub-group which can be identified are crossings, where four segments are grouped pairwise. This happens when two stems overlap due to the image acquisition setup and the



(a) Crossings occur when two stems are overlapping. Due to the projection into the 2D image space as well as during the shipping and storage of the plants these configurations occur.

(b) Two very close branchings of stems sometimes appear as 3 tier branch when four edges are clustered.

Figure 5.14: Cluster nodes which contain more than 3 edges are often mixtures of very close termination-, crossing- and branching- points. These configurations are the biggest source of errors during the grouping process.

preparation of the plants. Another sub-group are 3-tier branching points, where 2 branches branch off from 1 branch in a very close neighbourhood.
All other configurations (especially when containing more than 4 four edges) are a combination of the already described configurations. These cluster nodes are solved by searching minimal cost paths from already defined branches (direction is known) to other branches.

## 5.5 Geometrical and Topological Trait Extraction

After reconstructing the realistic topological architecture of the plant, the quantitative traits for each plant can be determined. The traits are divided into topological traits and geometrical traits.

### 5.5.1 Topological Traits

Topological traits are traits which are relevant to quantify the "network topology" of the plant. Especially the number of different branch-types as well as the number of siliques are of interest. The following branch types are considered for a branch:

- *Main Stem:* A main stem (MS) is originating from the rosette and is usually the thickest, longest and most central branch. If more then one main stem is originating from the rosette area, all these branches are denoted as main stems and are identified by a continuous number, e.g. *MS_1, MS_2,...*.

- *Siliques:* Each terminating branch can either be classified as silique or leaf. Siliques can generally be identified by a specific length and a higher variation in width. A detailed discussion about the identification of siliques is done in Section 5.5.3.
- *Leafs:* All terminating branches which can not be classified as siliques are denoted as leafs. Leafs in this sense denote mainly flowers and small exterior branches.
- *Side Branches:* The remaining interior branches are denoted as side branches. To differentiate between different types of side branches, the depth of the side branches is used. This means that side branches branching off a main stem (MS), are denoted as primary side branches or side branches depth one (SB I). Side branches branching off from a SB I are denoted as secondary side branches or side branches depth two (SB II). Side branches from a higher order are quite rare, all remaining side branches are summarized as side branches with depth three or higher, denoted as SB III+.

For each type of branch the number of branches is determined. Another topological trait which is of interest is the *siliques distribution*. The siliques distribution is defined as the number of siliques originating from each branch type. Same types of side branches are grouped and the number of siliques is summarized.

### 5.5.2 Geometrical Traits

Beside the topological traits, geometrical traits describing the actual size of the plants are determined. The geometrical traits which are calculated for each branch *B* are:

- *Path Length* $l_B$*:* The length of the branches' centerline is determined by the number of odd and even direction codes along the centerline. $N_o$ is denoting the number of odd direction codes along the centerline, $N_e$ is denoting the number of even direction codes along the centerline. The final path length can be calculated by [34]:

$$l_B = N_o + N_e \cdot \sqrt{2} \tag{5.18}$$

- *Euclidean Length* $d_B$*:* The euclidean length of a branch is defined as the euclidean distance between the branches' end points [34].
- *Tortuosity* $tor_B$*:* The tortuosity of a branch is an indicator of how curved a branch is. The tortuosity of a branch is defined as the ratio between the path length $l_B$ and the euclidean length $d_B$ [34].

### 5.5.3 Silique Detection

The classification between siliques and leafs is based on the assumption that the exterior parts of the plant are mainly siliques. Due to the mature state of the plant, where *biological* leafs and flowers were already dropped, this assumption should hold for most of the plants. Further, siliques generally tend to a have a higher variation in width along its skeleton and the average siliques length on one plant has less variation compared to other leafs.

To classify between siliques and other leafs the following geometrical features can be considered:

- Euclidean distance between end points of a leaf
- Path length
- Tortuosity
- Mean and standard deviation of the signal level along the leaf
- Average width and standard deviation along the leaf
- Directional variation in degrees
- Area of the silique

The most relevant features being considered in this work are the path length and the width variation. The classification methodology is based on the detection of outliers in a set of observations. Two different methods are used for detecting outliers.

**Minimum Covariance Determinant (MCD) Estimator**

The MCD estimator calculates a robust estimation for multivariate location and variation. The algorithm finds a subset *h* out of *n* samples (observations) having the smallest covariance determinant. The mean of this subset *h* and their covariance are taken as the robust estimation of location and variation. The minimum size of *h* is

$$\frac{n+p+1}{2} \tag{5.19}$$

where *p* is the number of variables used. The maximum value for *h* is *n*. These estimators are then used to calculate robust distances to all samples. Samples with a large distance can be expected as outliers [44]. In this work *h* was specified as $0.75 \cdot n$.
The calculation of the MCD estimator is part of LIBRA, which is a MATLAB library for robust analysis [1] [56].

**MADe Method**

Is a simple, robust method to identify outliers for univariate variables. This method is used if the number of leafs is smaller than 20. The MADe method is based on the median absolute deviation (MAD), which is calculated as follows [45]:

$$MAD = median(|X_i - median(X)|) \tag{5.20}$$

An outlier is detected if the score $M_i$ is outside the value range $[-3, 3]$, where $M_i$ is calculated as:

$$M_i = \frac{0.6745 \cdot (X_i - median(X))}{MAD} \tag{5.21}$$

The path length is used for detecting outliers if the number of the plants is less than 20.

---

[1] `http://wis.kuleuven.be/stat/robust/LIBRA/`

## 5.6 Summary

Deriving topological and geometrical traits from 2D images, where parts of the objects tend to overlap, usually requires a segmentation step as well as a reconstruction step before final characteristics can be extracted.

This work is focusing on the extraction of the plants' centerline. With use of the centerline, the topological traits as well as the geometrical traits can be extracted. In contrast to common methods, where the whole object is segmented in a first step and the centerline is extracted in an additional step, a direct exploratory approach was used in this work. The advantages of this tracing procedure are a less erroneous medial axis representation (no pruning required) and the fact that topological relations between certain parts of the centerline are already determined during the centerline extraction. Further, local geometrical features, e.g. directional changes, radii or segment lengths are also already determined during tracing the plants' stems.

These *local* attributes can subsequently be used for reconstructing the realistic architecture of the plant.

The reconstruction is based on the local information of the tracing algorithm and continuity principles during the grouping of centerline segments. A hierarchical reconstruction (graph pyramid) approach is proposed in this work.

The tracing procedure requires an initial set of points (seed points) for initialization. The automated seed point initialization procedure, which is proposed in this work, is providing one seed point for each main stem. Furthermore, this seed points are necessary to determine the quantitative traits during the final step. The accuracy of the quantitative trait extraction, where topological as well as geometrical traits are extracted, is depending on the accuracy of the tracing procedure as well as on the correctness of the reconstruction procedure.

CHAPTER 6

# Experiments and Evaluation

In this work a framework is proposed to extract quantitative traits of 2D images from mature *Arabidopsis*. To determine the benefits and drawbacks of the applied methods, the framework was evaluated using a dataset containing 106 images and two different types of ground truths.
In Section 6.1 the characteristics of the images are summarized and the ground truths including their different traits are presented. The chapter continuous with Section 6.2 which contains experiments concerning the use of different gray-scale transformations and analysing the contrast of the images.
Section 6.3 contains the evaluation regarding the pre-processing methods. This starts with evaluating the region-of-interest identification and showing the benefits which are achieved by an efficient pre-processing of the images. Further, the unsupervised method to apply a spatial calibration is evaluated by comparing it to a manual spatial calibration.
The automated seed point identification is evaluated in Section 6.4 . The focus during this evaluation was set on the determination of the precision and the grade of automation which can be achieved. The seed points of this procedure are used as initial points for the tracing procedure which is evaluated in Section 6.5. The completeness of the centerline extraction is determined when only one seed point per main stem is used. Further, the benefits and drawbacks concerning the resulting medial axis using tracing are discussed.
In Section 6.6 the unsupervised plant reconstruction is analysed. The throughput rate of this stage is determined and the biggest sources of error are discussed. The extraction of the quantitative traits using the unmodified results after the reconstruction step is done in the subsequent section. The accuracy of the extracted topological and geometrical traits is determined and discussed. In addition, the grade of automation and the overall performance of the framework is discussed in Section 6.8. Especially the grade of automation is analysed as it is one of the key-characteristics to determine if the framework could be used in laboratory environment or not.

## 6.1 Evaluation Data

The evaluation of the framework is done by using 106 images of 106 different plants. The image dataset, the different types of ground truths as well as the evaluation and development environment are discussed in the following.

### 6.1.1 Image Dataset

The images which are used in this work are originating from a project entitled *The molecular basis of local adaptation in A. thaliana*. A first set of images (including those used in this work) were analysed manually and certain traits (see Section 6.1.2) were extracted by one student. The image acquisition setup for all images remained the same and is discussed in detail in Section 4.1.1.

The evaluation of each step during the digital phenotyping pipeline in this work was done with 106 images. All the images are having the same characteristics:

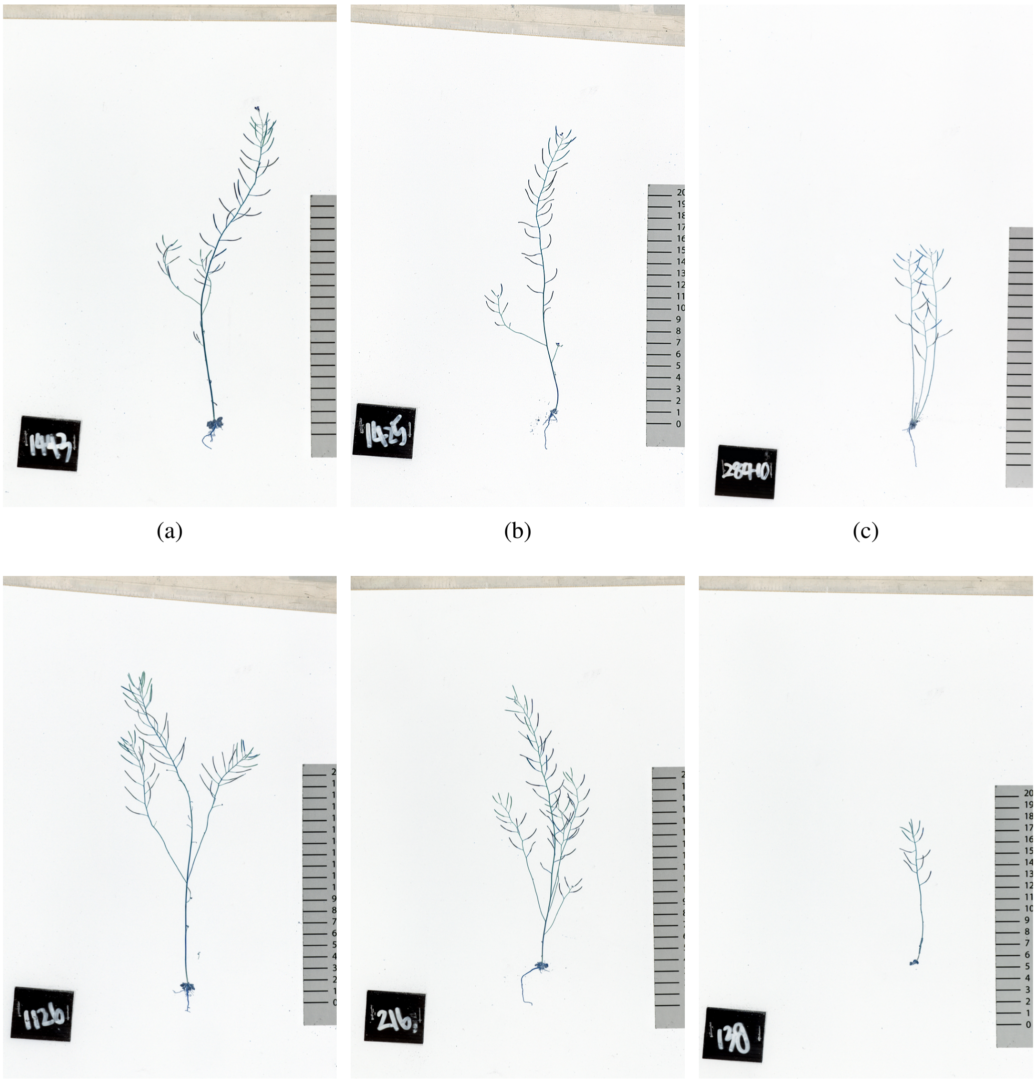
- Image resolution: 12.1 Megapixels
- Color mode: RGB 8 Bit
- Image size: $4284 \times 2844 Px$
- File size: 36.6 MB
- File format: TIFF

The plant images were taken for the purpose of manually determining the traits as part of a pre-study for a subsequent, bigger study. Due to this fact, there was no special attention given to the overlapping of certain branches. The success of correctly reconstructing a plants' realistic architecture is highly depending on the number of crossings/overlappings. Some sample images are shown in Figure 6.1. The images were inverted and rotated for illustration reasons.

### 6.1.2 Ground Truth from Manual Phenotyping (GT UK)

The given plant images were already analysed manually by a student at the John Innes Centre (JIC) in Norwich, UK. The following traits were extracted:

- Number of main stems
- Number of basal branches
- Number of cauline branches
- Number of cauline nodes
- Number of all side branches
- Number of siliques



(a) (b) (c)

(d) (e) (f)

Figure 6.1: Sample images of the dataset which was used for the evaluation of the phenotyping framework. The plants differ a lot in size as well as in their morphological complexity. Images are rotated and inverted for illustration reasons.

- Average length of siliques (in cm)
- Infertile numbers of siliques

The data is provided as a table, where each plant can be identified by the plant ID which is visible on the plants' images (see Figure 6.1). Originally, the traits were extracted from a set containing 1763 plant images. Concerning the dataset which was used for evaluating the framework, a manual ground truth was available for 105 out of 106 images.

### 6.1.3 Ground Truth from Digital Phenotyping (GT Fiji)

The ground truth from the manual phenotyping (GT UK) is not replicable and limited concerning the geometrical properties of the individual branch types as well as concerning the topological traits (e.g. distribution of siliques) of the plant. For this reason, a second ground truth was created to evaluate the framework. This ground truth was determined by using Fiji [1] in combination with a plug-in called *Simple Neurite Tracer* [32]. With use of the Simple Neurite Tracer the centerline of each branch is extracted by manually defining points along the plants structure. These points are connected by minimizing a cost function which is based on the euclidean distance between the points. The euclidean distance is scaled by the intensity values of the image or a measure of *tubeness* [32]. All 106 images were analysed in this way.

The ground truth for each plant is stored in a separate folder including the extracted centerline of the image, a csv-sheet including the branches geometrical properties as well as a swc file for reproducing the annotation process.

In contrast to the manual ground truth, where the branching structure was not analysed in detail, the branches definitions for the semi-automatic ground truth are slightly different. The definitions follow the principle which was formulated during the quantitative trait extraction of the framework (details see Section 5.5). Each branch is denoted as one of the following:

- Main Stem (MS)
- Side Branch Depth I (SB I)
- Side Branch Depth II (SB II)
- Side Branch Depth III+ (SB III+)
- Silique
- Leaf

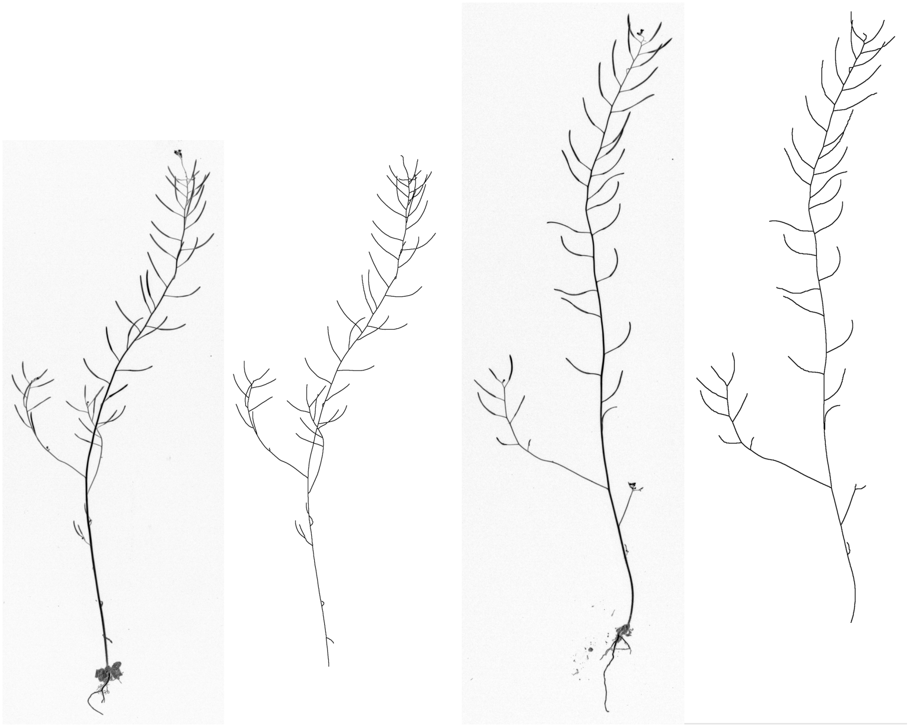
The topological and geometrical traits as defined in Section 5.5 were determined for each plant. The path length $l_B$ for each branch was already provided by the Simple Neurite Tracer. The euclidean length $d_B$, the tortuosity $tor_B$ as well as the topological summaries were determined in an additional step using Matlab[2].

Sample images and their ground truth are shown in Figure 6.2.

---

[1] http://fiji.sc/Fiji

[2] http://www.mathworks.de/products/matlab/



(a) (b)

Figure 6.2: Sample images and their corresponding, semi-automatically extracted centerlines.

### 6.1.4 Evaluation and Development Environment

The development of the framework as well as its evaluation was done on a single machine. A MacBook Pro (13-inch, Mid 2009) with the following specifications was used:

- *Processor:* 2.53 GHz Intel Core 2 Duo processor
- *Memory:* 4GB of 1066 MHz DDR3 SDRAM memory
- *Hard Drive:* 250GB Serial ATA; 5400 rpm
- *Graphics:* NVIDIA GeForce 9400M graphics processor with 256MB of DDR3 SDRAM (shared with main memory)
- *OS:* OS X Mountain Lion (10.8.5)

The framework was developed using Matlab R2012b (8.0.0.783).

## 6.2 Monochromatic Image Representation - Experiments

Color can be a useful feature for differentiating between specific regions in an image. A better discrimination between colors can often be achieved by transforming the images from the original color space into another (see Section 4.3). This section deals with experiments concerning the color as well as a suitable monochromatic (gray-scale) representation of the images.

### 6.2.1 Color

As the plants got dried, due to transportation and storage reasons, they lost their natural color (variations of green) and turned into light-brown. On a first view, the color of the siliques seems to differ from the color of the stems. By a detailed analysis of different images someone can notice, that this constraint is not valid for every silique and the color variation is too "big" to be used reliably. Even on a single image the color information of the siliques varies often.

Preliminary experiments with the image data showed that the color information of the plants is not useful for discriminating between different parts of the plant. For this reason there is no loss of useful information when transforming the images into a monochromatic representation. The processing steps for analysing the plants get more simple and efficient after this reduction.

### 6.2.2 Gray-scale images

There are several ways of transforming the RGB color image into a gray-scale image. A common way is to reduce the color information by selecting one specific channel from the color space. As already indicated in the previous paragraph, the color information of the image is mainly represented by luminance. For this, transforming the plant into a color space, where luminance and chrominance are separated was considered. The following monochromatic representations were tested and are denoted as follows:

- ***R:*** Red (R) color channel from the RGB color space.
- ***G:*** Green (G) color channel from the RGB color space.
- ***B:*** Blue (B) color channel from the RGB color space.
- ***V:*** Brightness (value V) channel from the HSV color space.
- ***L:*** Luminance (L) channel from the L*a*b* color space.
- ***Y:*** Luminance (Y) channel from the YIQ color system.

An accurate image segmentation result can be achieved if there is a high discrimination between the object and its background. This is the case, if the contrast between these regions is high. For this reason the different gray-scale transformations were analysed concerning the contrast of the final gray-scale image. As the analysis of the contrast from a few samples by eye brought no clearance, a contrast measurement was considered and realised. To measure the contrast of an image, the fore- and background object have to be identified. The foreground object was defined as the plant and estimated using the bit-plane slicing method described in Section

4.4. All objects except the plant were filtered after the binarization step. The background object was defined as all elements which are not part of the four biggest objects which are the plant ID sign, the ruler, the plant and the interfering object at the top of each image. The reference mask was created by using the luminance channel in the L*a*b* color space, as it can be expected to be the perceptually most uniform representation.

Different types of contrast measurements can be considered for images. A contrast measurement, based on the Weber ratio and referred as Weber contrast *is used to measure the local contrast of a single target of uniform luminance seen against an uniform background* [39]. The Weber contrast is defined as:

$$C = \frac{\Delta L}{L} \qquad \Delta L = L_f - L_b \qquad L = L_b \tag{6.1}$$

$\Delta L$ in this experiment is calculated as the difference between the foreground Luminance $L_f$ and the background luminance $L_b$. The luminance is measured as the average gray-level intensity. A high perceptual contrast in an image results in a high Weber contrast value.

Further, the computational effort for each transformation was measured by the computation time the color space transformation required.

***Results:*** The contrast was measured for all images of the dataset and the Weber contrast was calculated for each image and each gray-scale representation. A boxplot showing the Weber contrast for each gray-scale representation is shown in Figure 6.3. The L-channel extracted from the L*a*b* color space shows the highest contrast in average. Further the computation time ($12,7\ s$ per image) is significantly higher compared to other transformations (e.g. transformation into YIQ: $0,6\ s$ per image).

***Discussion:*** As expected, the L*a*b* color space seems to be the best choice for representing the image in gray-scale. In a time-critical system the choice of this representation is unsuitable since the computational cost for the transformation is much higher compared to the other considered transformations. In this work, a trade-off between the achieved contrast and the computational time was made by choosing the R-channel of the RGB space as a gray-scale representation.

## 6.3 Evaluation of Pre-Processing

The pre-processing pipeline, described in Section 4.3, 4.4 and 4.5 can be considered as independent from the remaining processing pipeline. The most important methods during pre-processing are the automated ROI identification (see Section 4.4) and the automated spatial calibration (see Section 4.5).

The accurate identification and cropping of the regions-of-interest are crucial tasks in respect of computational cost and disk space savings. The effect on the computational cost during analysis of the plant with use of tracing methods is rather small. Nevertheless, if methods are used for an automated seed point detection, objects which are non-plant objects can interfere the rational

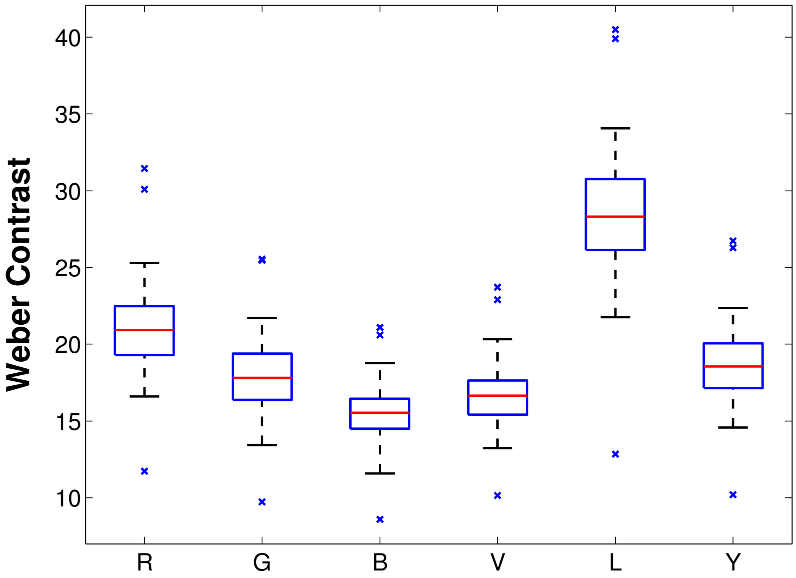


Figure 6.3: Boxplot showing the Weber contrast value for all images of the dataset in the individual monochromatic representation. The L channel from the L*a*b* color space shows the highest contrast of all representations. Certainly the computational effort for the color transformation is also the highest.

choice of seed points. Further, if data is someday reused and analyzed using e.g. a global segmentation method the computational effort will be reduced significantly.
The images are stored, although they were already analyzed, for a later use and for the reproducibility of the analysis. As digital high-throughput phenotyping systems are producing thousands of images in a high resolution, the disk space capacity can quickly be depleted.
Therefore the automated ROI identification is evaluated concerning the disk space savings which can be achieved by cropping the plants and the correctness of identifying the different regions.
The evaluation concerning the spatial calibration was focusing on a comparison of a manual spatial calibration and the automated spatial calibration (see Section 4.5).

### 6.3.1 Correctness of Automated ROI Identification

To determine the correctness of the ROI identfication method all images of the dataset were manually controlled. Images, showing a mask-overlay of the region-of-interest were produced for this evaluation. Figure 6.4 shows four example images for determining the correctness of the algorithm.
A region is identified as *correct* if both of the following two conditions are given:

1. Region of type $T$ is correctly identified as region $T$, where the set of $T$ comprises

$$\{\mathit{Plant}, \mathit{Plant\text{-}ID}, \mathit{Ruler}\} \tag{6.2}$$

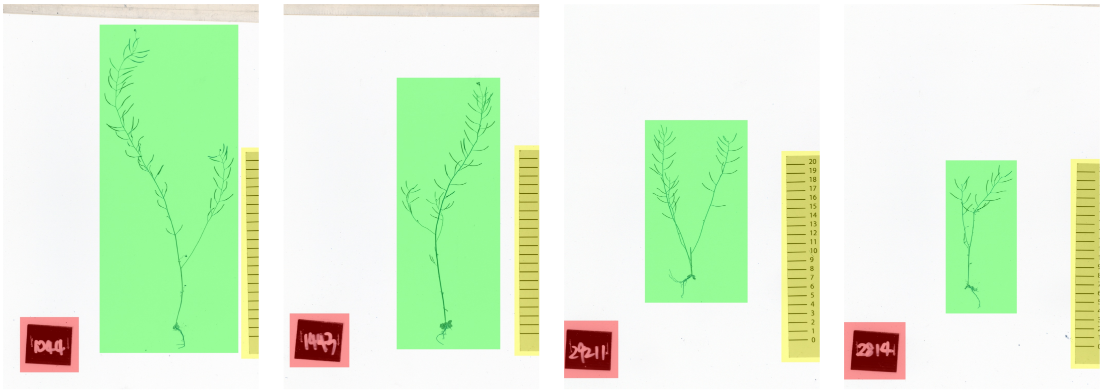


Figure 6.4: Evaluation images for determining the correctness of the algorithm. The individual masks are shown as an overlay (*red* = plant ID, *green* = plant, *yellow* = ruler). Each image of the dataset was controlled visually for its correctness.

.

2. The extracted window is *not* removing relevant details of the region, e.g. removing parts of the plant.

***Results:*** With use of the proposed methods, all regions of the whole dataset (106 images) were correctly classified and cropped. This means:

- All plant-ID regions were identified as plant-ID regions.
- All plant-regions were identified as plant-regions.
- All of the ruler-regions were identified as ruler-regions.
- All parts of a region were extracted.

***Discussion:*** The use of basic image processing techniques showed to be very effective during an automated ROI identification. Although there is a slight variation during the image acquisition, all regions were identified correctly. The ROI extraction is fitted to the current image acquisition setup. Changes in the acquisition setup may lead to errors in the ROI identification. If the setup changes, the a-priori assumptions for designing the method (see Section 4.4) have to be reconsidered. As the method was split into different modules, some modules could then be modified or substituted by other methods.

### 6.3.2 Disk Space Savings

The disk space savings are evaluated by using characteristics which are related to data compression evaluation. The disk space savings are a result of removing *redundant data*- data which is not useful for representing any information. The amount of redundant data is defined as follows. *"Let b and b' denote the number of bits (or information-carrying units) in two representations of the same information, the relative data redundancy R of the representation with b bits is*

$$R = 1 - \frac{1}{C} \tag{6.3}$$

*where C, commonly called the compression ratio, is defined as"*

$$C = \frac{b}{b'} \tag{6.4}$$

[23, pp. 526-531]. These characteristics can be determined for individual or multiple images. The relative data redundancy is equivalent to the relative disc space savings.

***Results:*** Using the proposed cropping method, the total disk space consumption of the dataset could be reduced from 3874.6 $MB$ to 770.7 $MB$ (corresponds to $R = 81.92$ %). The average space savings for an individual image is 29.9 $MB$, which means that the average image is carrying 81.92 % of redundant information. Table 6.2 is providing a detailed overview of the disk space savings which were achieved.

| | Disk Space Consumption | | Compression Rate *(C)* | Data Redundancy | |
|---|---|---|---|---|---|
| | *(b)* [MB] | *(b')* [MB] | | *b - b'* [MB] | *(R)* [%] |
| Total | 3874.6 | 700.7 | 5.53:1 | 3174.0 | 81.92 |
| Mean $\overline{x}$ | 36.6 | 6.6 | 9.27:1 | 29.9 | 81.92 |
| Std. Deviation $s$ | 0 | 4.5 | 8.82:1 | 4.5 | 12.43 |
| $x_{max}$ | 36.6 | 26.2 | 58.77:1 | 35.9 | 98.30 |
| $x_{min}$ | 36.6 | 0.6 | 1.40:1 | 10.4 | 28.45 |

Table 6.2: Statistics concerning the initial disk space consumption *b* and obtained disk space consumption *b'* for the dataset. The compression rate (C) as well as the absolute and the relative data redundancy *R* are illustrated for all images (total) and the average image (mean $\overline{x}$). Additionally the extrema ($x_{max}$, $x_{min}$) of each parameter were determined.

The histogram in Figure 6.5 shows the distribution of the relative data redundancy *R*. The histogram is illustrating the wide range of the individual space saving rates, which can be explained by the varying size of the plants in the test set. As the cropping procedure is cropping the plants to its bounding box the space savings for big plants are smaller compared to the space savings for small plants.

***Discussion:*** The cropping procedure is extracting only the relevant information of the images. Redundant information, which also a human observer would not need for analysing the plants is removed. As the plants vary a lot in size, but the image acquisition setup stays (almost) the same during the acquisition of the images, a lot of unused background is "produced". Removing this

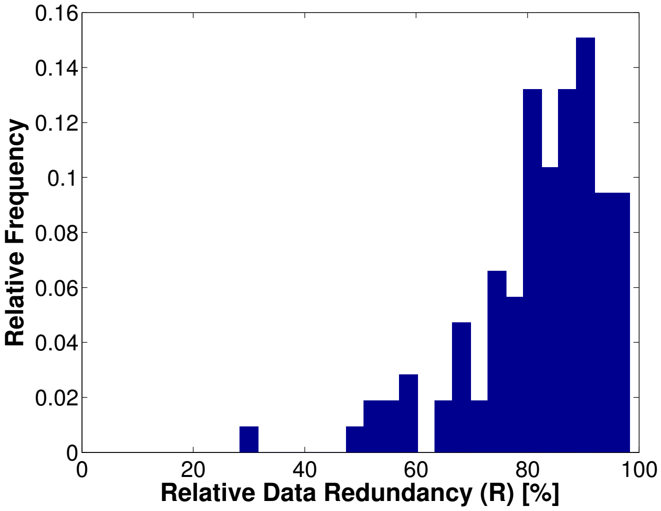


Figure 6.5: Histogram showing the distribution of all relative space savings gained from the whole dataset.

background can be done without loosing any relevant information. The evaluation shows that this step is indispensable for an efficient data storage handling. As already mentioned before, the space saving achievement is dependent from the actual plant size. Nevertheless, considering a test set of rather big plants, the redundant information can still be expected to be around 25 $\%$ (compare minimum *R* value in Table 6.2).

A text file including the plant ID as well as the conversion factor for each image is stored additionally to the plant image itself. The size for this text file was not included in the previous calculation as it would not show an effect on the results.

### 6.3.3 Spatial Calibration

The focus of this evaluation is to get to know the difference between a manual and an automated extraction of the scale (pixel/mm ratio - conversion factor). For this comparison the scale (representative: the ruler-graduation distance $d_{RG}$) was once extracted manually by using *imtool* from *Matlab's Image Processing Toolbox* and once using the automated method described in Section 4.5. The manual determination was achieved by measuring the euclidean distance $d\left(p_{(1)}, p_{(N)}\right)$ between the first and the last ruler graduation of the ruler image. This distance is further divided by the number of ruler graduations $N$, which was always 20.

$$d_{RG}^{(r)} = \frac{d\left(p_{(1)}, p_{(N)}\right)}{N} \tag{6.5}$$

The following parameters are determined for evaluating the automated procedure:

- *Relative/Absolute Change:* Parameter which describes the deviation of two quantitative measurements. In quantitative mathematics also referred as error [4]. The manually extracted values are taken as reference values $d_{RG}^{(r)}$ and are compared to the measurements

$d_{RG}^{(a)}$ achieved by the automated procedure. The relative change $\delta_d$ in $\%$ is calculated as follows:

$$\delta_d = \frac{\left| d_{RG}^{(a)} - d_{RG}^{(r)} \right|}{d_{RG}^{(r)}} \cdot 100\ \% \tag{6.6}$$

The expression in the numerator is denoted as *absolute change* $|\Delta d|$. When comparing the final conversion factors $c_f$, which result from Equation 4.9, the relative change between $c_f^{(r)}$ and $c_f^{(a)}$ is denoted as $\delta_c$.

- *Measures of central tendency and statistical dispersion:* The *arithmetic mean* $\overline{x}$, the *standard deviation* $s$, the *maximum* $x_{max}$ and the *minimum* $x_{min}$ are used for describing and analysing the relative and absolute changes on the whole dataset.

***Results:*** With use of the developed spatial calibration method all the images of the dataset were analyzed automatically. The statistics from the manual extractions as well as from the automated extraction are shown in Table 6.3. The determined mean for the manual ruler-graduations distance was 97.107 Pixels ($\pm 0.388$) while the automatic approach yield to an average of 97.112 Pixels ($\pm 0.398$). When calculating the conversion factor $c_f$ on basis of the ruler distance, the results were even more similar. There was no difference noticeable until the fourth decimal place. The average conversion factor was determined with $0.1030$ $Px/mm$ ($\pm 0.0004$).

| | Manual | | Automated | |
|---|---|---|---|---|
| | Distance $d_{RG}^{(r)}[Px]$ | Scale $c_f^{(r)}[mm/Px]$ | Distance $d_{RG}^{(a)}[Px]$ | Scale $c_f^{(a)}[mm/Px]$ |
| Mean $\overline{x}$ | 97.107 | 0.1030 | 97.112 | 0.1030 |
| Std. Deviation $s$ | 0.388 | 0.0004 | 0.398 | 0.0004 |
| $x_{max}$ | 97.740 | 0.1042 | 97.768 | 0.1042 |
| $x_{min}$ | 95.996 | 0.1023 | 96.013 | 0.1023 |

Table 6.3: Measures of central tendency and statistical dispersion for manual and automated extraction of ruler unit distances $d_{RG}$ and conversion factors $c_f$.

Statistics concerning the relative and absolute changes for the manual and the automated routine are shown in Table 6.4. While there is a small absolute and relative change noticeable when comparing the ruler unit distances $d_{RG}$, there is only an extremely slight difference when comparing the conversion factors $c_f$. The maximum absolute difference of ruler units differences observed, was less then 1 Pixel (0.163). This difference is hardly noticeable when calculating the final scale constants. Further, the maximal impact (maximum error/change) regarding to the actual image sizes was determined. Assuming a plant with maximum size (maximum height of the image) the difference between manual and automated extraction would effect in 0.116 Pixels (in average).

Figure 6.6 is showing 2 plots comparing the individual manual and the automated spatial calibration. Both plots are showing the extracted ruler unit distances. In the scatter plot (left) the

|  | Ruler Unit Distance |  | Scale |  | Worst Case |
|---|---|---|---|---|---|
|  | $\left\|\Delta d^{(r)}\right\|$ $[Px]$ | $\delta_d$ $[\%]$ | $\left\|\Delta d^{(a)}\right\|$ $[10^{-6}Px]$ | $\delta_c$ $[10^{-6}\%]$ | (h = 4284 $Px$) $[Px]$ |
| Mean $\overline{x}$ | 0.026 | 0.0263 | 27.076 | 263.03 | 0.116 |
| Std. Deviation $s$ | 0.025 | 0.0259 | 26.527 | 258.7 | 0.114 |
| $x_{max}$ | 0.163 | 0.1663 | 104.0701 | 1665.343 | 0.730 |
| $x_{min}$ | 0.0001 | 0.0001 | 0.1056 | 1.0275 | 0.0005 |

Table 6.4: Relative and absolute change statistics for the dataset. The last column is showing the maximum possible impact of an automate extraction of the scale compared to the manual extraction. This is the theoretic case, when the plant is as large as the image height, which is 4284 pixels.

blue line denotes the *ideal* match between the manual and the automated extraction. The scatter plot shows, that the two measuring strategies are almost identical. The Bland-Altman plot (right) is comparing the mean value of each measured pair $i$ with its difference $\Delta d_i$. Additionally two lines at position $\overline{\Delta d_i} + 1.96SD$ and $\overline{\Delta d_i} - 1.96SD$ are shown.

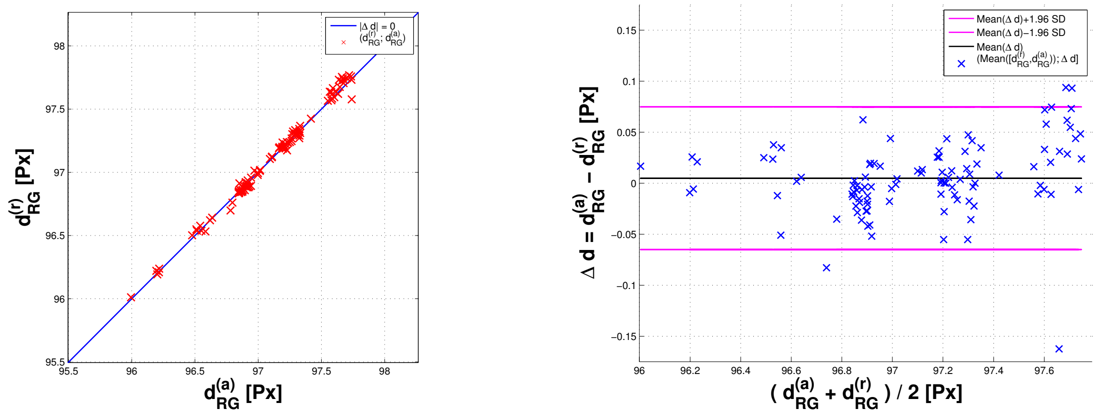


Figure 6.6: Scatter plot (left) showing the manual against the automated ruler graduations distances. The blue line indicates the case that the two measuring methods are the same. It can be noticed that the deviations from this blue line are in a small range. The Bland-Altman plot (right) is comparing the mean values $\overline{\Delta d_i}$ of each measuring pair $i$ against its difference. Only four samples can be noticed to be outside the range $\overline{\Delta d_i} + / - 1.96SD$.

***Discussion:*** The evaluation showed, that there is no noticeable difference between the manual and the automated extraction of the conversion factor. It also showed that the image acquiring conditions were very constant. The resulting ruler unit distance varies between 95.996 and 97.740 Pixels (manually extracted). This results in a scale variation between 0.1023 and 0.1042 $mm/Px$.

Gray-level thresholding at a level of *t = 175* was used for the binarization of the ruler. If the acquisition setup is changing (other ruler) this parameter may need to be adapted. Otherwise, the use of gray-level thresholding should be sufficient enough for a whole range of rulers as they regularly show a good local contrast.

## 6.4 Automated Seed Point Identification

An automated seed point identification procedure was developed and is described in Section 5.2.2. Whenever the automated seed point identification procedure fails to identify the seed points correctly, the seed points can be chosen manually by the user. This evaluation is used to determine how often an user intervention was needed and how precise the unsupervised method works.

The performance of the method can be determined by quantifying the ability of identifying a seed point correctly. A seed point is correctly identified when it is located on the main stem nearby the rosette. The following characteristics are measured:

- *True Positives (TP):* The number of detected seed points which are located on the main stem.
- *False Positives (FP):* The number of detected seed points which are not located on the main stem.
- *False Negatives (FN):* The number of seed points which were not detected using the automated seed point identification routine.
- *Precision (P):* Percentage of detected seed points which are located on the main stems in reality:
$$P = \frac{TP}{TP + FP} \tag{6.7}$$
- *Recall (R):* Percentage of seed points which are detected by the procedure.
$$R = \frac{TP}{TP + FN} \tag{6.8}$$
- *F-value (F):* A measure of balance between precision and recall:
$$F = \frac{2 \times R \times P}{R + P} \tag{6.9}$$

***Results:*** The characteristics were obtained by controlling the output of the automated seed point procedure manually. The amount of TP, FP and FN were determined for each plant. Table 6.5 is summarizing the precision and the recall rate as well as the F-value. The evaluation was done using two different sizes for the initial structuring element (disk). The best result was achieved using an initial disk size of 20 pixels. The resulting F-value was obtained in 97.01 $\%$. In 7 out of 106 plant images($\simeq$ 6 $\%$) a manual re-touch of the results was needed as some seed points were wrongly detected or missed.

| Initial Disk Size [Px] | Precision | Recall | F-value |
|---|---|---|---|
| 15 | 0.935 | 0.992 | 0.963 |
| 20 | 0.949 | 0.992 | 0.970 |

Table 6.5: Performance measurements for the automated seed point identification routine.

***Discussion:*** The automated seed point detection is working properly in most of the cases ($\simeq$ 94 $\%$ of the dataset). The more main stems are originating from the rosettes location, the more challenging the accurate identification of the main stems gets, as the branches intend to cross each other already nearby the rosette. Problems also can arise if the rosette of the plant is not dominant enough, which means that there are no leaves left and the root network of the plant directly leads into the main stems. In this case a rosette can not be identified and the whole procedure fails. In regular cases, where the main stems originate from the rosette in an organized way and the rosette is dominant enough the automated procedure can be used.

## 6.5 Centerline extraction

The extraction of the centerline of the plants was done using a tracing approach as presented in Section 5.3. The main advantage of this approach is that local features, describing the object, are already determined during the segmentation. Further, tracing procedures are reported as less sensitive against discontinuities along an object's contour and the resulting medial axis is reported as less error-prone [18, 35]. The main disadvantage is that branching points can be missed during the iterative tracing procedure and the obtained centerline will be incomplete. The evaluation of this procedure as well as a short comparison between the medial axis transform achieved using tracing and a regular skeletonization approach is done in this section.
The evaluation concerning the accuracy of the segmentation approach is done after the reconstruction process in Section 6.7.2.

### 6.5.1 Completeness

Intensity variations along the stems or bad positions of the search window can lead to undetected branches. In those cases the obtained centerline will be incomplete. To evaluate the completeness of the algorithm, the centerline results from the tracing procedure were validated manually. In those cases, where the tracing result was incomplete, additional seed points were set manually until the segmentation result was complete. The number of additional seed points was determined for each image and is used as an indicator for the completeness of the algorithm.

***Results:*** The histogram in Figure 6.7 illustrates the number of additional seed points which were necessary to complete the segmentation result. Images requiring the same amount of seed points where grouped into bins. A manual intervention was needed in $48.11\ \%\ (51/106)$ of the plants. A detailed look at the histogram shows that in $70.75\ \%$ of the images zero or one clicks were needed. Ordinarily, in these cases only a single silique was missed during the tracing. Further, in $92.45\ \%\ (98/106)$ of the images less then 4 clicks were necessary.

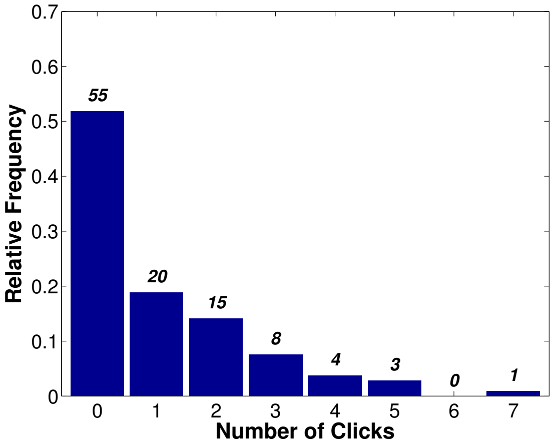


Figure 6.7: The histogram shows how many clicks were needed to complete the segmentation result in each image.

***Discussion:*** The evaluation showed that using one seed point per main stem results in an incomplete segmentation in approximately 50 % of the images. In most of the cases, a single silique or leaf was missing.

10 images were identified where bigger parts of the plant remained unsegmented. To counteract the problem of incomplete segmentation results a few corrective actions can overcome this drawback. Additional seed points can be added during the automatic seed point identification approach. These points could be initialized along a fixed grid over the whole image. The use of such techniques is already reported in literature by different approaches [8, 28, 61]. When specifying a grid for images of mature *Arabidopsis*, an appropriate grid size can for example be half the length of the average expected silique size.

### 6.5.2 Medial Axis Transformation

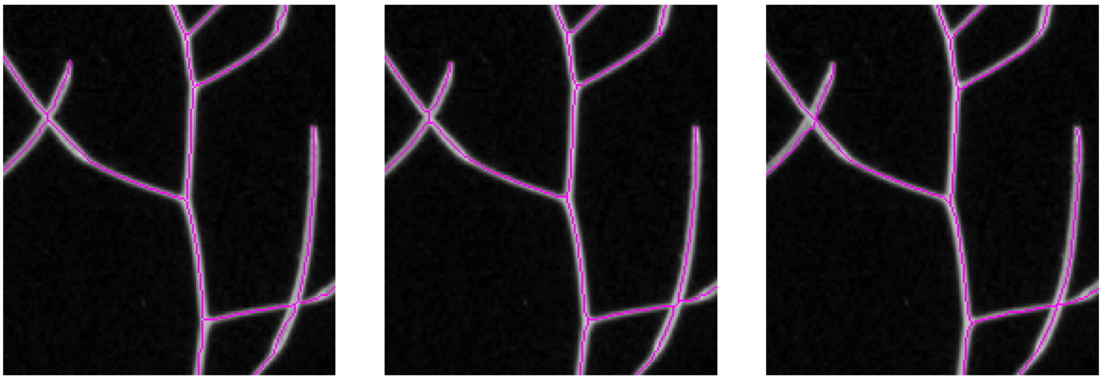
As already discussed in Chapter 3, different approaches exist to extract the centerline of curvilinear structures. A common technique is to segment the whole object in a first step and use a medial axis transformation in a second step. Common problems which are reported with the use of this approaches are resulting loops, occurrence of small branches or unwanted gaps [18, 35]. Tracing overcomes this problem by identifying critical points at first and later connecting the points if they are validated. Therefore this technique is reported to be less sensitive to noise and more robust against small variations in the contour of the object. To evaluate these phenomena 2 sample images are visually compared.

***Results:*** The comparison is done by using the extracted centerline of the tracing approach used in this work, the semi-automatically generated ground truth skeleton and a skeleton which was created by using a regular skeletonization procedure after globally segmenting the image using Otsu's method [38]. The differences between the technique are illustrated in Figure 6.8.

***Discussion:*** While analysing the centerlines, different aspects concerning the final skeleton can be noticed. Representative samples for the most obvious differences are shown in Figure 6.8. The images of the first row show a branching pattern, where the distances between critical points are large enough. Both techniques provide valid results when comparing the centerlines to the ground truth. During the centerline extraction using tracing, the critical points are already determined and the segments are split. Using the global approach, the bifurcation points in the skeleton have to be determined in an additional step. It can be noticed that each crossing point is turned into two branching points using the global segmentation/skeletonization approach.

The second row shows a sample, where the branching structure is more complex due to very close and thin branches and the presence of interfering objects like a flower. In these images, the differences between the methods are more obvious. While the regular skeletonization procedure produces loops and small branches the results of the tracing procedure is less "noisy". This reduction of noise can be seen as an advantage, because there is no need for post processing the skeleton to remove small branches or loops. On the other hand if big regions with complex structures occur in an image the result will remain incomplete without a manual intervention.

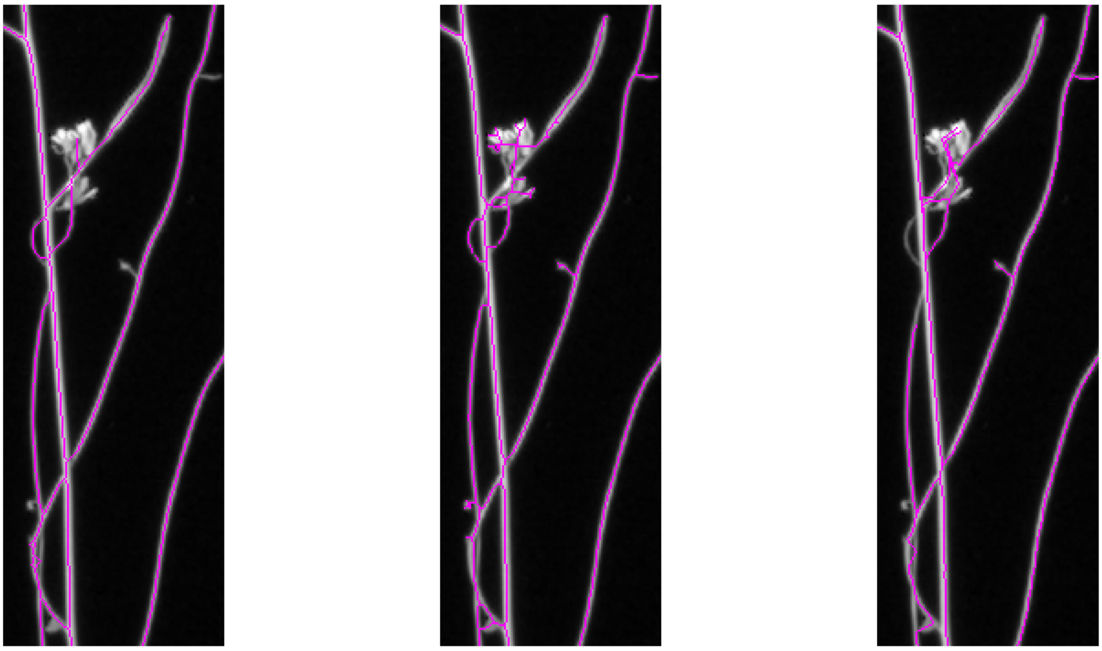
Summarized, the main advantage of the centerline extraction using tracing is the determination of local geometrical and topological features while segmenting the plant. With use of these features the reconstruction of the plant is possible directly after segmenting the image. If branching structures get very complex, which means many critical points in a small area, the tracing procedure will fail as the correct relation between centerline points can not be determined. This limitation could be overcome during the image acquisition setup by untangling very bushy plants.



(a) Ground truth skeleton (GT Fiji)

(b) Medial axis transformation after using Otsu [38] thresholding

(c) Centerline of the tracing-approach



(d) Ground truth skeleton (GT Fiji)

(e) Medial axis transformation after using Otsu [38] thresholding

(f) Centerline of the tracing-approach

Figure 6.8: Comparison of the ground truth skeleton with a skeletonization approach after using Otsu's thresholding [38] and the centerline extraction using tracing. The images of the first row show a branching pattern, where the distances between critical points are large enough. The second row shows a sample, where the branching structure is more complex due to narrow, tiny branches and interfering object like flowers.

## 6.6 Plant Reconstruction

Grouping of centerline segments is a crucial task while reconstructing the plants realistic architecture and is the basis for the final trait extraction. The success of the grouping process depends on the accurate segmentation of the centerline, the clustering process which defines the set of possible connections between segments and the strength of the cost function which was defined in Section 5.4.3.

The grouping procedure is applied systematically starting with segments from the exterior of the plant. The grouping procedure is continued until all segments were visited. Thus, intermediate results of one cluster node can directly be used for simplifying the grouping decision in an adjacent cluster node. The drawback of this method is the propagation of wrong decisions. This can lead to unconnected regions of the plant in the worst case, as not all groups are forced to be connected to the main stem. This fact was in focus of the evaluation and is evaluated by the throughput rate which was achieved during the unsupervised reconstruction process.

Further, the most likely configurations (defined in Section 5.4.3) are analysed and the success of reconstructing them was analysed visually. The most frequent problems were identified and are discussed in this section. The grouping process was evaluated using all images of the dataset. Images which were not completely segmented during tracing where re-traced by manually setting additional seed points.

***Results:*** In 89 out of 106 images the reconstruction procedure turned out in a complete result, where each main stem is represented by a single tree. This equals a throughput rate of $83,96\ \%$ for the unsupervised reconstruction step. The success of correctly grouping centerline segments during cluster nodes until grade III could be observed as less error-prone. Examples of almost perfect reconstructed parts of a plant are illustrated in Figure 6.9 and 6.10. The main sources of error during the reconstruction of the plant, which could be identified were:

- *Extensive Overlappings:* The occurrence of extensive overlappings often result in a configuration as shown in Figure 6.11. The segmentation process provides 5 instead of 4 segments. The grouping process is resolving this configuration as two bifurcation points, where one point is turned into a branching point and the other point is turned into a termination point. This has a major impact on the topological characteristics of the plant, as it gets resolved deeper than it actually is.
- *Segmentation Inaccuracies:* Very close stems or siliques as well as interfering flowers and a sporadically low contrast (which could be observed especially in the upper parts of the plant) are causing segmentation inaccuracies. This leads to the absence of short segments, merging of close stems or discontinuities. Figure 6.13 is showing an example for an error-prone grouping process at the very top of a plant. Small segments are missing due to bad local segmentation. As a consequence, the grouping process remains error-prone as wrong segments get merged, branching points are identified at crossing points and segments are terminating where they should not.
- *High Significant Points Density:* If multiple segments branch or overlap in a restricted area the correct reconstruction of the branching architecture is getting complicated. This

occurs for example if a stem is crossing nearby a branching point of an other stem. Figure 6.12 is showing an example for this problem.

- *Siliques Shape Variation:* While the main stems and side branches have a very low change of the curvature over long distances, the shapes of the siliques is much more unpredictable and varying. This leads to problems if two siliques are crossing at their lower parts as their shape after the crossing is hard to predict. In such cases, the limitations of the cost functions are reached.
- *Small Segments:* If relatively small segments (below 20 pixels) were part of a cluster node, the success of a correctly grouping the segment to its parent segment is limited as the definition of the segment is often not sufficient.

***Discussions:*** The evaluation concerning the plant reconstruction showed that a unsupervised reconstruction of the plants realistic branching architecture is restricted to cases where the morphological complexity is limited. Overlappings which originate from the image acquisition were identified as the biggest source of error. To unravel some of the branches before taking the images would make the reconstruction process more successful.
Further, the evaluation highlighted a drawback of the grouping procedure which is the propagation of errors due to wrong local decisions. A global procedure for reconstructing the plant may rise the throughput rate of the unsupervised reconstruction.

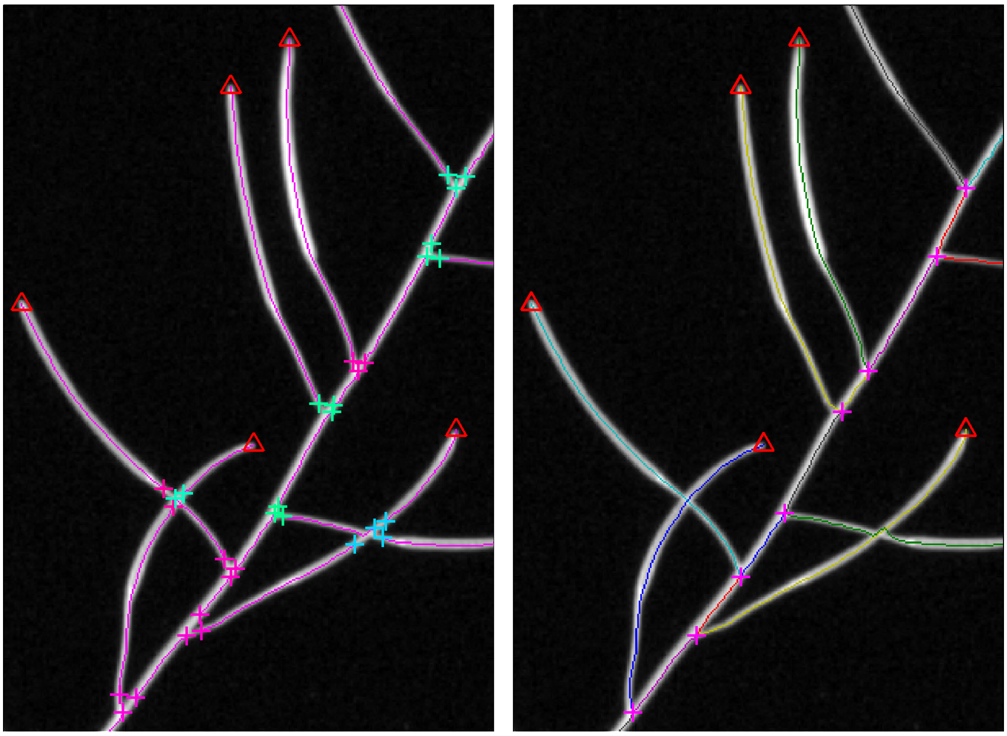


Figure 6.9: Parts of a plant where the success rate of correctly grouping a segment to its parent is 100 %. Branching points as well as crossing points in this image can be resolved correctly.

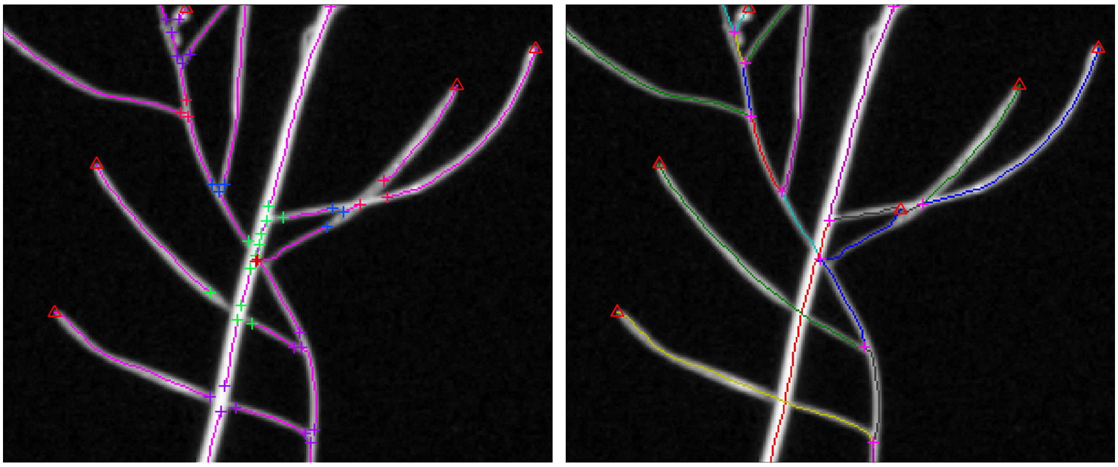


Figure 6.10: Example of a successful reconstruction of the plants realistic topological characteristics. Only one extensive crossing from the siliques to the very right is resolved incorrectly. The remaining crossings in this example are solved correctly as the crossing angle between the segments is big enough to group the involved segments.

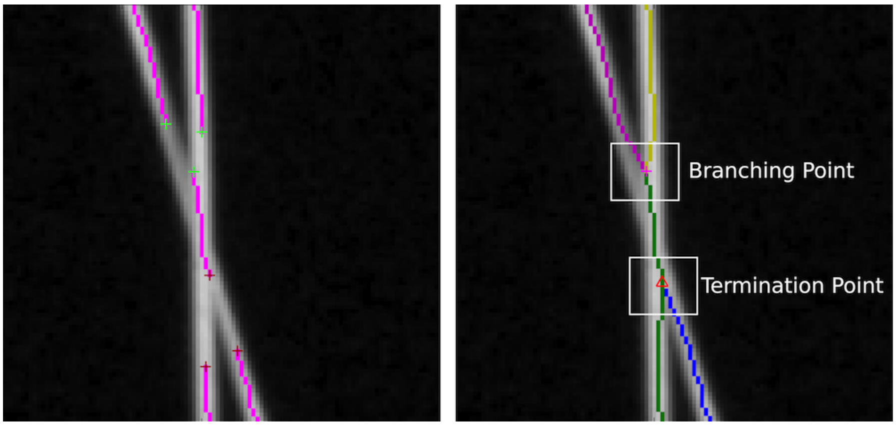


Figure 6.11: Extensive overlappings result into a wrong topological reconstruction of the plant. A branching point and a termination point are identified at a point where normally two pairs of segments should be grouped. The left image shows a region of a plant after the clustering process. The right image shows the same region after the grouping procedure. Branching points are marked with a pink coloured "+", termination points are marked with a red triangle.

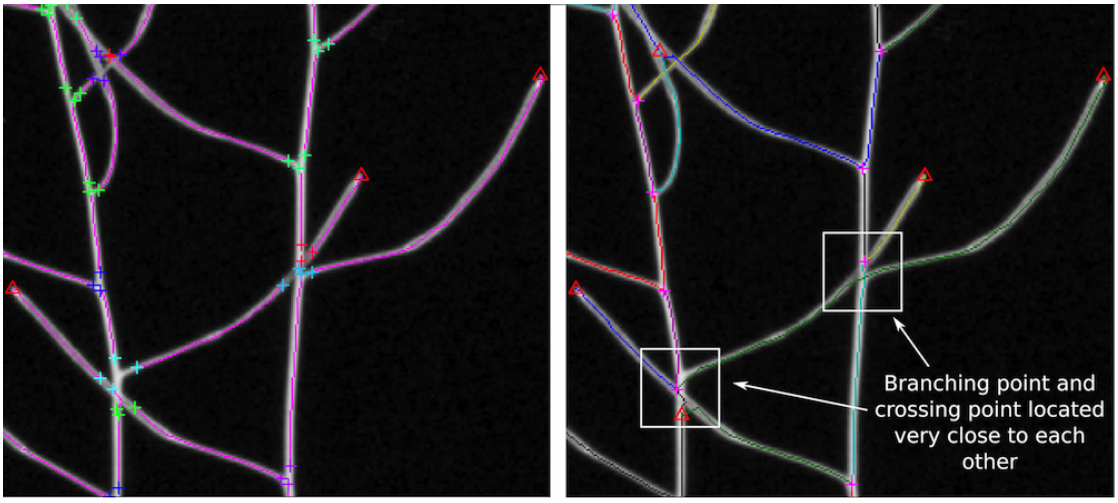

Branching point and crossing point located very close to each other

Figure 6.12: Crossing points are identified as the biggest source of error during the reconstruction of the plant. The process of grouping a segment to its correct parent segment is getting complicated at regions where branching points as well as crossing points occur.

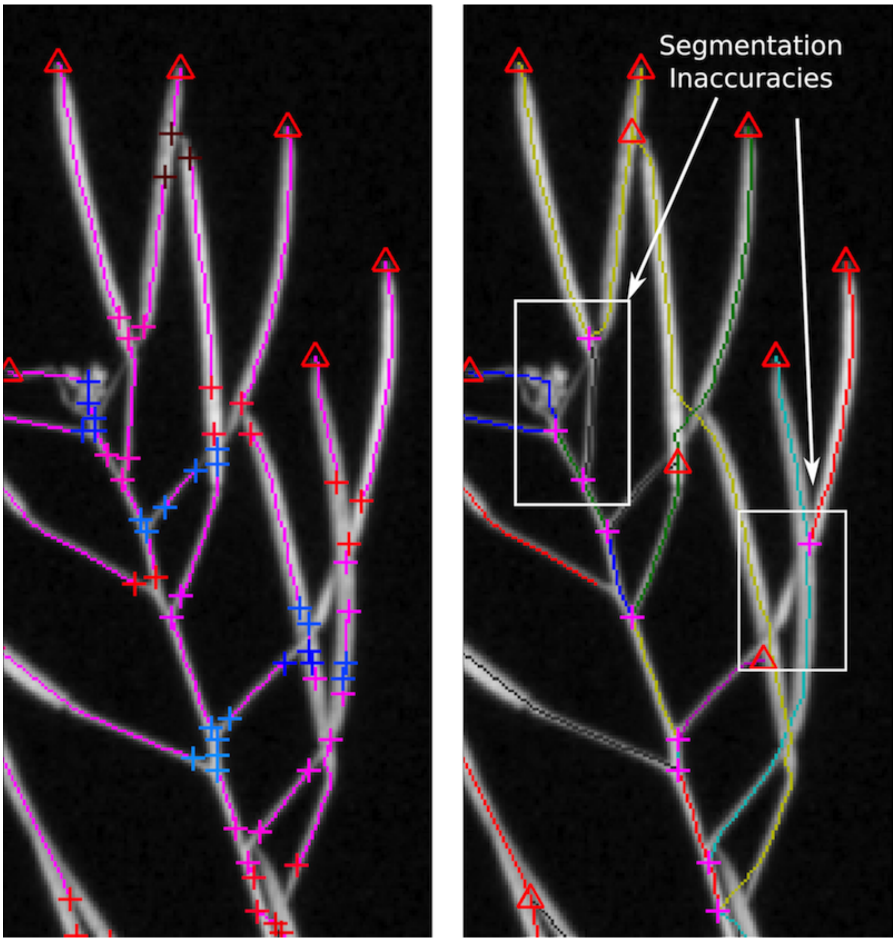


Figure 6.13: Segmentation inaccuracies in areas of very close stems lead to a error-prone reconstruction of the branching architecture. The left image shows a region of a plant after the clustering process.

## 6.7 Quantitative Trait Extraction

The topological and geometrical traits of the plants are extracted after connecting the centerline segments during the reconstruction procedure. There was no manual correction done after the grouping process. For this reason, the results of the quantitative trait extraction are highly dependent on the success of the reconstruction process. The evaluation of the final trait extraction was done using only images, where the grouping process resulted in one grouped object where the majority of the plant was reconstructed (see Section 6.6). In 89 out of 106 images the grouping process produced such a result. This reduced set is the basis for the evaluation of the quantitative trait extraction.

The evaluation was done by comparing the results of the framework and two independent ground truths (see Section 6.1.2 and 6.1.3). The two ground truths (GT) are shortened as GT Fiji (semi-automatically approach using Fiji) and GT UK (result of manual phenotyping the plants in Norwich, UK). Results which were produced using the developed framework are shortened with FW.

While the evaluation in this section is determining parameters concerning the set of 89 images, a number of resulting individual plant images and their corresponding traits are shown in the Appendix A of this work.

### 6.7.1 Silique Number

The identification of the siliques is one of the major interest for biologists. As not only the number of siliques is in focus of this work, but also their location on plant, the extraction and determination of the siliques has to be done on the plant. This makes the process more complex compared to studies where the number of harvested siliques are determined on an image.

For each image *i* the total number of siliques $SN_i$ is extracted. The relative error for each plant *i* is calculated by the difference between the siliques number of the ground truth $SN_i^{(GT)}$ and the siliques number determined by the framework $SN_i^{(FW)}$:

$$\eta_i = \frac{SN_i^{(GT)} - SN_i^{(FW)}}{SN_i^{(GT)}} \cdot 100\ \% \tag{6.10}$$

***Results:*** The mean relative error for all plants in the dataset was calculated by the mean of the relative errors' absolute values. The mean relative errors compared with the ground truths are shown in Table 6.6. The most reliable comparison is expected between GT Fiji and the FW results. The mean relative error was $11.66\ \%$ ($\pm$9.09). The relative error between the two independent GTs is surprisingly high. An explanation for this value is given in Figure 6.14a. This scatter plot shows the number of siliques provided by GT Fiji and GT UK. While there is a small variation between most of the silique numbers, two outliers (ID 9032 and ID 1120) are identified. The outliers were identified as errors which were made during the manual phenotyping of the plants (GT UK).

In the following the traits which were extracted by the framework are only compared with GT Fiji. Figure 6.14b is showing the scatter plot for comparing the values of the GT Fiji with the

| **Mean Relative Errors [%]** | GT (UK) | GT (Fiji) | FW |
|---|---|---|---|
| GT (UK) | 0 (±0) | X | X |
| GT (Fiji) | 8.09 (±25.15) | 0 (±0) | X |
| FW | 10.48 (±10.56) | 11.09 (±9.09) | 0 (±0) |

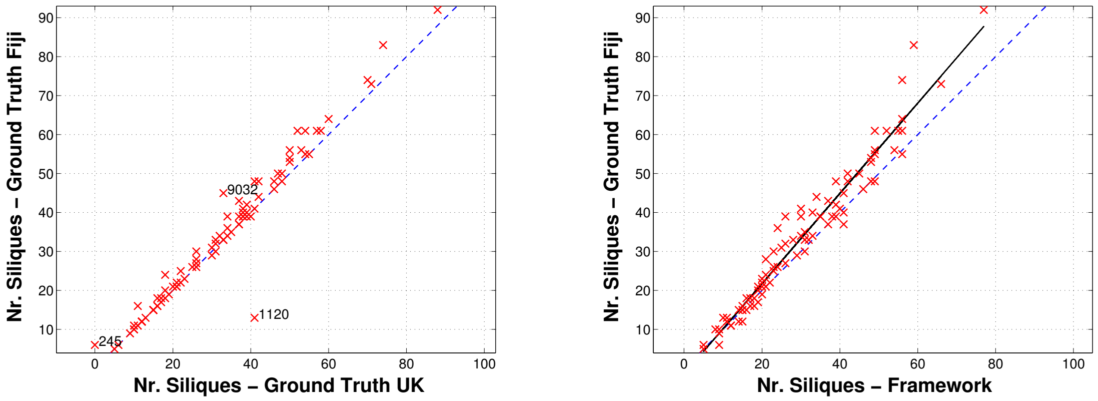
Table 6.6: The average relative errors which were achieved by comparing the different measurements.

values of the FW. Additionally to the blue line, which is indicating the regression line for an exact match between the methodologies, the regression line for the measurements is illustrated by the black line. The pearson correlation coefficient was determined with $97.6\ \%$ $(p < 0.01)$. It can be notified that the FW is detecting too less silique if the number of the siliques on a plant is rising. This systematic error comes more clear in a Bland-Altmann plot, which is shown in Figure 6.14c. The absolute differences are shown on the y-axis. It can be noticed that much more samples are spread below the zero line, especially when the number of siliques rises.

***Discussion:*** The evaluation concerning the total number of siliques which were identified using the framework shows that a systematical error is noticeable by an increasing number of siliques. An explanation for this behaviour is the increase of overlappings in big plants containing many siliques, which could be observed.

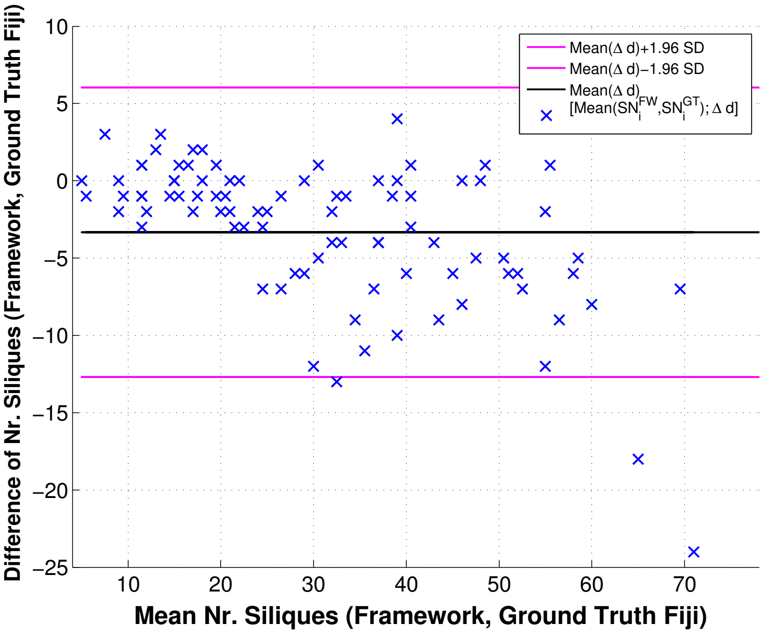
The comparison of the two sets of GTs showed also a small variation concerning the number of siliques which where identified. This shows that even for a human observer the correct identification is not always clear.

The relative error of the number of siliques does not provide information concerning the precision of the silique classification. Certainly, this type of evaluation would require an absolutely correct result after the grouping procedure and subsequent determination of TP, FP and FN based on the classification. In this work, the correctness of the reconstruction can not be guaranteed and therefore siliques or leafs may get split or merged with other parts. An evaluation after the unsupervised reconstruction would not be valid. This extended evaluation would go beyond the scope of this thesis and therefore remains open for future works.



(a) Scatter plot comparing the existing GTs. A small variation between the distinct set of ground truths can be observed. The sample with ID 245 was missing in the GT UK. Further, two errors in the GT UK were identified, which may cause the high value of the mean relative error when comparing the GTs.

(b) Scatter plot comparing the number of siliques which where extracted using the framework and the ground truth which was determined using Fiji. The black line is illustrating the regression line between the measurements. The pearson correlation coefficient was determined with 97.6 % ($p < 0.01$).



(c) The difference between $SN_i^{(FW)}$ and $SN_i^{(GT)}$ for each plant *i* is shown on the y-axis of the Bland-Altmann plot. The mean value of $SN_i^{(FW)}$ and $SN_i^{(GT)}$ is shown on the x-axis. It can be observed that the deviation between the methods rises by an increasing number of siliques on a plant.

Figure 6.14: Evaluation of the identification of the number of siliques.

### 6.7.2 Plant and Branches' Size

The length of the individual branches is evaluated in this section. For this reason the length of the certain branch types (MS, SB I, SB II, SB III+ and Siliques) were averaged per plant and compared to the Fiji ground truth. Determining the length of the branches of the reconstructed plant is again very dependent on the success of the reconstruction procedure. On the one hand the evaluation concerning the size is an indicator for the accuracy of the segmentation process on the other hand deficits concerning the reconstruction process will be visible.

***Results:*** The evaluation concerning the stems' lengths is summarized in Table 6.7. The average relative error concerning the MS length between the two measurements was $3.64\ \%\ (\pm 3.19)$. The average relative error concerning the siliques average length was $6.73\ \%\ (\pm 4.77)$. Scatter plots for both measurements are shown in Figure 6.15.
Concerning the interior branches of the plant (SB) the variation between the framework and the ground truth is worse. Looking at Table 6.7, it can be observed, that the traits, extracted using the framework, are much lower than the GT values.
While analysing this error, two main sources of error were identified: extensive overlappings and high significant points density (compare with Section 6.6). These sources of error were confirmed implicitly by applying a threshold value $\epsilon$ on the branches lengths and the corresponding change of results. The threshold is calculated for each plant as the average size of the siliques lengths' and adding one time the standard deviation. The change in results is visible in Table 6.7. It can be noticed that the number of side branches below the threshold value $\epsilon$ must have been very high as the average length of the side branches is noticeably increasing.
The partially wrong reconstructions also lead to a bad determination concerning the branching architecture of the plant. While in the GT no side branches with a depth deeper than 2 were identified, the framework identifies some branches with a depth of three and higher. To make the results comparable the side branches of the different depths were summarized as SB. In Figure 6.15c the change of the results concerning the lengths of all interior branches SB is visualized in a scatter plot. The results are shown before and after applying the threshold value.
Figure 6.15d illustrates the difference between the independent ground truths, while Table 6.8 summarizes the average relative error between the different methods of measurement. A systematical error can be identified by the transition of the linear regression line. This systematical error can be explained by the two different methods of measuring the siliques. While the length of a silique was measured from the beginning of the carpel in GT UK, the length of the silique in the FW was measured as the path between the branching point where the silique originated and the end of the silique.

***Discussion:*** Traits concerning the size of the plant are highly dependent on the previous steps of the phenotyping pipeline. This starts with the image acquisition where overlapping arise by reducing the plants to the 2D image space and with the erroneous grouping in critical cluster nodes. While the exterior branches (siliques, main stems) of the plant can be reconstructed with an accuracy of more than 90 %, the exact extraction of the interior stem lengths is errorprone. Applying a threshold which is specifying a minimum length of a side branch results in a noticeable change of the results. This is an indicator that too many short side branches emerged

| Branch Type | Ground Truth Fiji | | Framework | |
|---|---|---|---|---|
| | without $\epsilon$ [Px] | with $\epsilon$ [Px] | without $\epsilon$ [Px] | with $\epsilon$ [Px] |
| MS | 1983 (±731) | 1987 (±728) | 2030 (±756) | 2041 (±743) |
| SB I | 650 (±413) | 779 (±386) | 258 (±187) | 791 (±467) |
| SB II | 150 (±126) | 331 (±0) | 230 (±210) | 816 (±370) |
| SB III+ | X | X | 168 (±105) | 621 (±308) |
| SB | 667 (±441) | 778 (±400) | 253 (±367) | 730 (±450) |
| Siliques | 179 (±24) | X | 168 (±21) | X |

Table 6.7: Results of the evaluation concerning the branches' sizes for the dataset. Applying a threshold value $\epsilon$ improves the final results of the trait extraction concerning the size for the interior branches.

| **Mean Relative Errors [%]** | GT (Chicago) | GT (Fiji) | FW |
|---|---|---|---|
| GT (Chicago) | 0 (±0) | X | X |
| GT (Fiji) | 27.37 (±6.03) | 0 (±0) | X |
| FW | 29.86 (±12.31) | 6.73 (±4.77) | 0 (±0) |

Table 6.8: Comparison of the siliques' length determined using the framework and the ground truths.

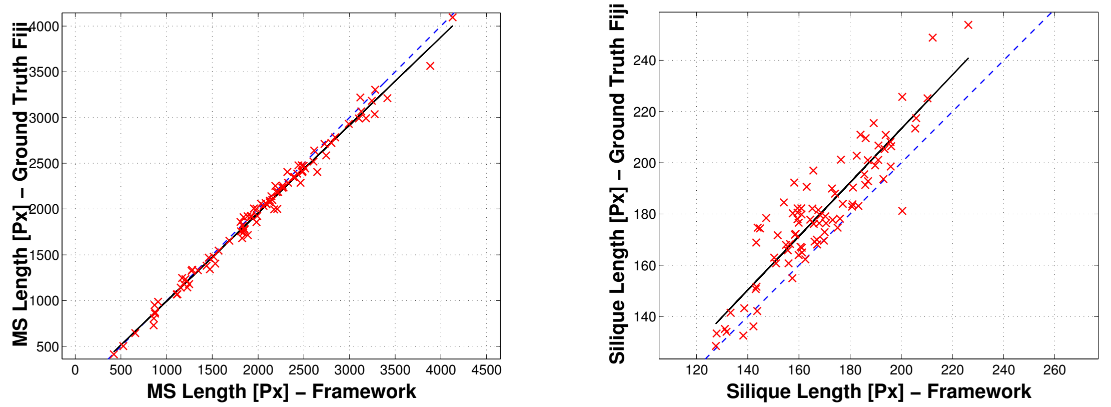
during the reconstruction procedure. A source of error are for example extensive overlappings of two siliques, where one short side branch is wrongly created.

### 6.7.3 Branching Pattern Analysis

The topological structure of the plant is characterized by the number of main stems and side branches of different depths in this work. In biology an even more detailed differentiation between types of main stems and side branches is done. To compare the results of the framework with the results generated using Fiji, the absolute numbers of the detected branch types were determined.

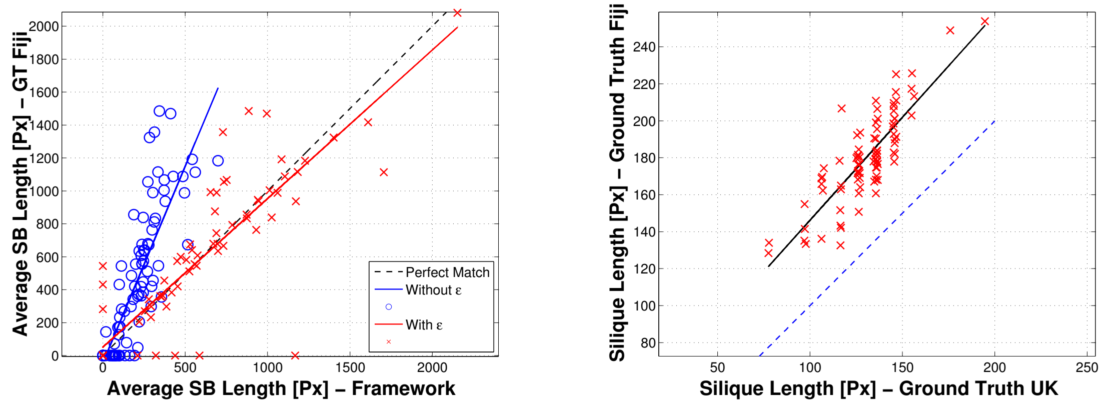
***Results:*** The number of certain stem types which were determined while analysing the plant are depending on the correctness of the grouping procedure. As already shown in Section 6.7.2, wrong local grouping decision in the interior of the plant can cause a propagation of the error until the exterior of the plant. Table 6.9 summarizes the absolute numbers for the dataset. Again, the threshold value $\epsilon$, specifying a minimum branch length, is necessary to obtain valuable results.

***Discussion:*** The detection of the topological structure is strongly related to the complexity of the images and the successful reconstruction of the plants realistic architecture. Similar to the evaluation concerning the plant's size, the exact identification of the topological internal archi-



(a) The main stem lengths for each plant determined by the framework are compared with the ground truth from Fiji. The average relative error was determined with $3.64\ \% \ (\pm 3.19)$. The correlation coefficient was determined with 0.99 (p<0.01).

(b) The average silique lengths for each plant determined by the framework are compared with the ground truth from Fiji. The average relative error was determined with $6.73\ \% \ (\pm 4.77)$. The correlation coefficient was determined with 0.90 (p<0.01). A bias can be observed which indicates a systematical error. The silique lengths determined by the framework intend to be smaller when comparing with the ground truth.



(c) Applying a threshold on the side branches lengths improves the extraction noticeable. Nevertheless the results show still a large deviation. The correlation coefficient changes from 0.83 (p<0.01) to 0.89 (p<0.01).

(d) A comparison of the ground truths showed a systemtical difference in determining the lengths of the siliques. Apart from the systematical difference the variation of the manual trait extraction is still higher compared to the extraction using the framework. The correlation coefficient was determined with 0.83 (p<0.01).

Figure 6.15: Evaluation of the different branch types lengths.

| Branch Type | Ground Truth Fiji | | Framework | | Absolute Differnce | |
|---|---|---|---|---|---|---|
| | without $\epsilon$ | with $\epsilon$ | without $\epsilon$ | with $\epsilon$ | without $\epsilon$ | with $\epsilon$ |
| MS | 99 | 98 | 97 | 95 | 2 | 3 |
| SB I | 139 | 118 | 372 | 98 | -233 | 20 |
| SB II | 4 | 1 | 89 | 23 | -85 | -22 |
| SB III+ | 0 | 0 | 79 | 13 | -79 | -13 |
| SB | 143 | 119 | 540 | 134 | -397 | -15 |

Table 6.9: The architecture of the plant is quantified by the number of primary stems (MS), internal stems (SB) and their leafs (siliques, leafs). Errors during the trait extraction using the framework are leading to a high deviation between the stem numbers extracted by the framework and the ground truth.

tecture of the plant is not possible using the unsupervised method of the framework. Applying a threshold value to the internal branch sizes shows a noticeable improvement of the results.

### 6.7.4 Siliques Distribution

Not only the geometrical traits concerning the siliques are of interest but also the topological traits. Especially the distribution of the siliques is in focus. The siliques distribution is the number of siliques located on the different types of branches. The correct topological position of a silique, which is located on the exterior of the plant requires a correct reconstruction of all interior segments of a plant. Wrong decisions in the interior of a plant are influencing the results of the final quantitative traits massively as the error is propagated to the next level. To evaluate the siliques distribution, the number of siliques originating from the different types of branches were determined and compared.

***Results:*** As already expected, the siliques distribution is containing errors and is deviating from the siliques distribution which was determined during the ground truth extraction. Figure 6.16 is showing the number of siliques which were detected using the framework compared to the number of siliques from the ground truth. What can be noticed is the already mentioned propagation of error along the depth of the plant. While there remain siliques undetected at the main stem level, many siliques are detected on *deeper* side branches.

***Discussion:*** The siliques distribution which was determined using the framework is error-prone as it is dependent on the correct reconstruction of the plant in all lower levels of the branching tree. The error propagation is clearly visible when analysing the siliques distribution achieved using the framework.

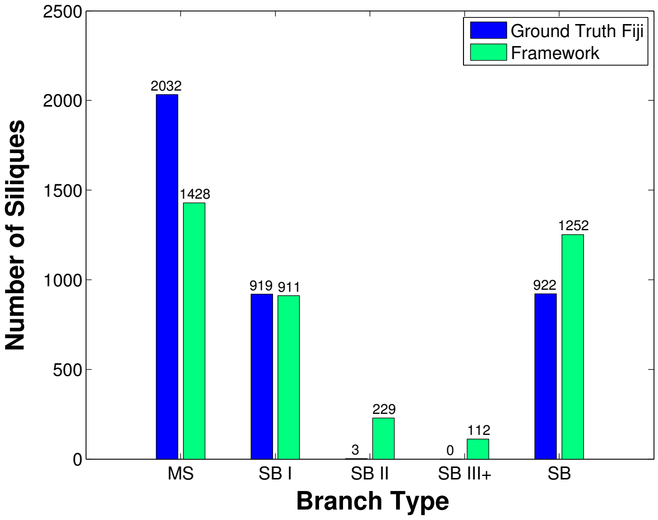


Figure 6.16: The siliques distribution represented by the number of siliques which were detected on the different types of branches.

## 6.8 Executability

Beside the accuracy of the achieved results, the actual use of a digital phenotyping system in a laboratory environment is depended on the exectuability of the developed system. In high-throughput phenotyping the most important aspects concerning the executability are the grade of automation and the performance regarding to time. Both aspects were evaluated and are described in the following. An overview of the results is shown in Figure 6.17 at the end of this section.

### 6.8.1 Grade of Automation

Each manual intervention during the digital phenotyping pipeline is causing costs, as it implies the need of an expert to manually control and correct the results. For this reason the manual intervention during the different steps of this framework was tried to be brought towards a minimum. To *measure* the grade of automation the number of *clicks* which have to be made by an operator were counted. The grade of automation was evaluated by using all 106 images of the dataset.

***Results:*** The number of clicks was determined for the first three modules of the pipeline:

I) Pre-Processing: No Clicks were necessary during the Pre-Processing of the images.

II) Seed Point Detection: During the automated seed point detection, some false positives or false negative seed points were detected. To manually correct these points **7** clicks were necessary.

III) Centerline Extraction: Due to the miss-detected stems the centerline extraction process remained partially incomplete in $48.11\ \%$ $(51/106)$ of the images. As already described in Section 6.5.1, to complete the results after this step **112** clicks were necessary.

There was no manual intervention done during the modules *IV)* and *V)*, where the plant is reconstructed and the traits are extracted. The grade of automation during these steps is depending on the complexity of the final traits which should be extracted using the framework and the specified fault tolerance.

***Discussion:*** The grade of automation which was achieved during the pre-processing as well as during the automated seed point detection would make the developed framework applicable in a laboratory environment. More than $90\ \%$ of the images could be analysed without a manual intervention during these steps. Nevertheless, a manual validation step would be required. Concerning the centerline extraction approach, the grade of automation is unsatisfying. A manual intervention in approximately $50\ \%$ of the cases is not applicable in a real, laboratory environment. To counteract the problem of undetected stems during tracing, future developments could use additional seed points which are located on pre-defined grids. This approach is already described as helpful by different approaches [8, 28, 61].

### 6.8.2 Time Performance in General

The computational time for the different modules of the phenotyping pipeline was evaluated for those image which were processed fully automatically and where the reconstruction process was identified as satisfying. This reduced set comprises 47 images. For each module the total computation time as well as the average computational time per image were determined. Further, the ratio of the computational time for each module to the computational costs in total was determined. There were no optimization techniques used for reducing the computational time. This could be considered for future works.

***Results:*** An overview of the computational times for each module and the overall computational times is presented in Table 6.10. Further, the results are shown in Figure 6.17, where the overall executability evaluation is presented.

| Module | Avg. Time / Image [s] | Total Time [s] | Overall Ratio [%] |
|---|---|---|---|
| I) Pre-Processing | 3.45 | 162.11 | 7.56 |
| II) Seed Point Initialization | 0.59 | 27.69 | 1.29 |
| III) Centerline Extraction | 28.49 | 1339 | 62.47 |
| IV) Plant Reconstruction | 10.99 | 516.33 | 24.09 |
| V) Trait Extraction | 2.09 | 98.21 | 4.58 |

Table 6.10: The computational costs were evaluated for each module. The total time as well as the average time per image for those image which were processed fully automatically are presented.

***Discussion:*** The evaluation concerning the computational costs showed that the tracing procedure as well as the plant reconstruction module are computationally too expensive for a validation by an operator during processing the images. For this reason, a final digital phenotyping system should consider to provide intermediate results for validation after each module for a set of images.
With use of the framework 47 images could be processed in approximately 35 minutes (without optimizing the code concerning computational time). Creating the ground truth images using Fiji took approximately 10 minutes per image (time saving by using the framework of approximately 90 %).

## 6.9 Summary

The evaluation of the framework which is proposed in this work highlighted the benefits and drawbacks along the certain modules of the phenotyping pipeline. The benefits of using 2D images of the plants are a high throughput rate during the image acquisition. The biggest drawback and the main problem during analysing the plants are the overlapping of plant regions, which originate from the projection from the 3D object to the 2D image space as well as from pressing the plants due to logistical reasons.
While the throughput rate after the pre-processing stage is still 100 % the throughput rate starts to decrease with identifying the main stems. During identifying the initial points along the main stems for the tracing algorithm as well as for the final trait extraction the throughput rate drops to 94 %. The effort to re-raise the throughput rate to 100 % was determined with 7 clicks.
Using a direct exploratory centerline extraction approach has the benefit of extracting local geometrical and topological features concerning the medial axis of the object already during the segmentation step. Another benefit of the tracing procedure is that only the local neighbourhood of an object is explored. Interfering, unconnected objects in the background will not influence the segmentation process. On the other hand if mistakes happen in identifying the neighbourhood, the results may remain incomplete. This was the case in 48 % of the images where parts of the plant were missing. In most cases, single siliques were missing. These missing regions can be added by manually adding additional seed points by clicking on the plants structure. In 92 % of the images, less then 4 clicks were necessary to complete the segmentation. Adding additional points, e.g. along a specified grid, already at the beginning of the tracing procedure would bring the throughput rate of the unsupervised centerline extraction procedure towards 100 % and should be considered in future works.
The reconstruction of the plants' real branching architecture is identified as the most critical task during the pipeline. During this step the inaccuracies from the segmentation process which originate from the limitations of the image acquisition come to light. Wrong local decisions in the interior of the plant are propagated to the exterior of the plant and can cause a bad reconstruction of the plant. This effect is noticeable while analysing the final traits of the plant. While the length of the main stems and the length of the siliques can be extracted with a relative error from approximately 5 %, a detailed topological reconstruction of the interior branches and stems of the plant is error-prone. Re-analysing the traits with use of knowledge gained during the trait extraction, e.g. specifying that side branches have to be longer than the average siliques

length are improving the traits concerning the interior of the plant. A manual intervention during the plant reconstruction would raise the accuracy of the final traits but should be considered as an expensive task for experts. A trade-off between accuracy respectively complexity of the extracted traits and arising costs for manual interventions has to be planned for future works.

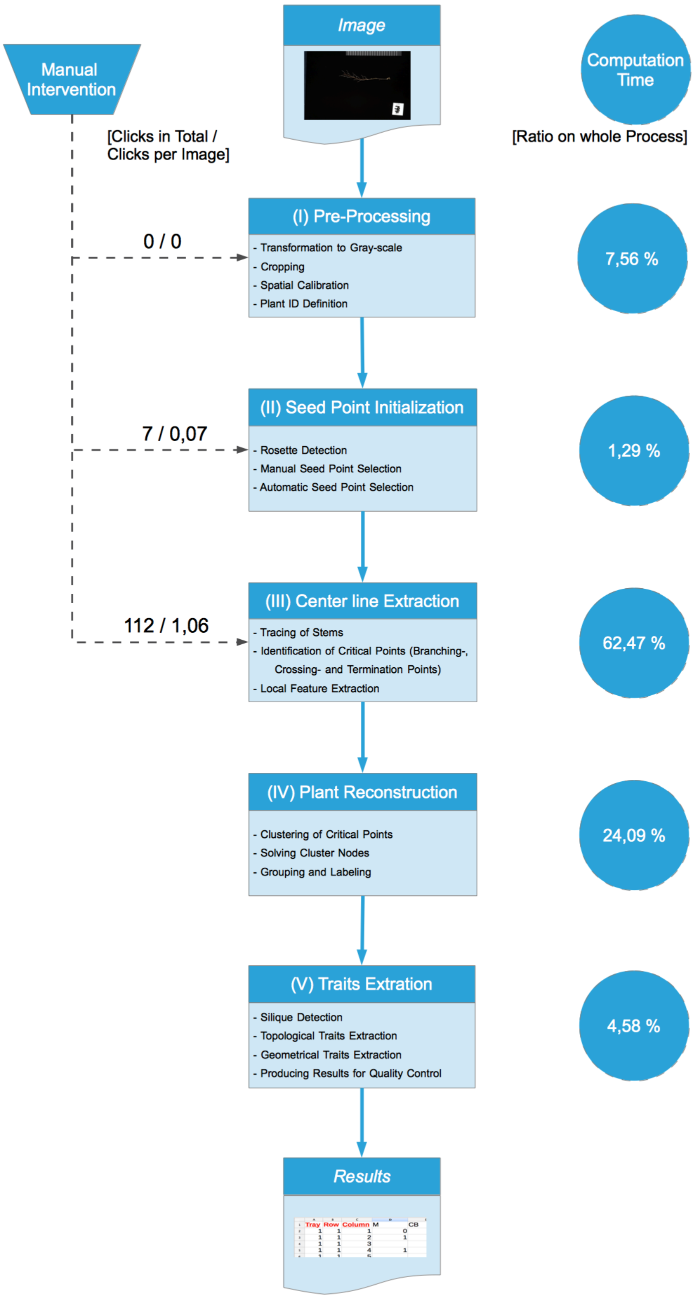

Manual Intervention
[Clicks in Total / Clicks per Image]
Image
Computation Time
[Ratio on whole Process]
0 / 0
(I) Pre-Processing
- Transformation to Gray-scale
- Cropping
- Spatial Calibration
- Plant ID Definition
7,56 %
7 / 0,07
(II) Seed Point Initialization
- Rosette Detection
- Manual Seed Point Selection
- Automatic Seed Point Selection
1,29 %
112 / 1,06
(III) Center line Extraction
- Tracing of Stems
- Identification of Critical Points (Branching-, Crossing- and Termination Points)
- Local Feature Extraction
62,47 %
(IV) Plant Reconstruction
- Clustering of Critical Points
- Solving Cluster Nodes
- Grouping and Labeling
24,09 %
(V) Traits Extration
- Silique Detection
- Topological Traits Extraction
- Geometrical Traits Extraction
- Producing Results for Quality Control
4,58 %
Results
Tray Row Column M CB

Figure 6.17: Framework overview.

## CHAPTER 7

# Conclusion

This chapter summarizes the work which was done during this thesis. The benefits and drawbacks of the framework are discussed and prospects for future works are given.

In this work, a semi-automatic framework for the extraction of geometrical and topological traits from 2D images of mature *Arabidopsis* is presented. Due to logistical reasons and the image acquisition setup parts of the plant are overlapping. This makes the extraction of topological and geometrical traits challenging.
The developed pipeline performs every single step for phenotyping the plant. This starts with a region-of-interest (ROI) identification used for cropping the images. The ROI methods is based on globally segmenting the down-sampled images using the most relevant bit planes of the gray-scaled images. Mathematical morphology operators are subsequently used to identify the objects of interest which are the plant itself, a plant ID as well as a ruler. The ruler is used for an automatic spatial calibration which is necessary to make the results of the framework comparable to a provided ground truth and future studies.
A tracing approach using a dynamical, semi-circular search window is used to extract the centerline of the plant. During this extraction, local features like the centerline-segments' direction, the width or the intensity of the centerline are identified. The motivation behind the use of a direct exploratory method for the extraction of the centerline is to gain geometrical and topological local features already during the segmentation process. This information can subsequently be used for reconstructing the plants' "realistic" branching architecture. The hierarchical reconstruction approach systematically identifies critical points as branching points or crossing points and groups the centerline segments of the segmentation process. The grouping of centerline segments is based on continuity principles, e.g. the edge direction similarity. The grouping process is invoked at the exterior leafs of the plant and is continued until all segments are grouped.
Further, an approach for the automatic identification of the main stems of a plant is presented. This procedure is based on the identification of the rosette using mathematical morphology operators and a multi-scale ridge identification in the rosettes' extended neighbourhood. The location of the main stems is used for initializing the tracing procedure as well as for the final quantitative

trait extraction.
While the pre-processing methods can be ran unsupervised, the remaining parts of the pipeline are requiring the validation of an user after each processing step along the pipeline. The validation guarantees the correct identification of main stems as well as the completeness of the centerline extraction process. If the validation is negative, the user can interact with the framework and manually correct the preliminary results. There is no manual correction implemented after the reconstruction procedure and the subsequent trait extraction.
The evaluation of the framework was done using $106$ images of $106$ different plants. The throughput rate of the final steps in the pipeline is $83,96$ % $(89/106)$, as wrong decisions during the plant reconstruction led to partially incomplete final plants. These plants were excluded from the dataset before the accuracy of the final traits was evaluated. The evaluation showed that the geometrical traits of the exterior parts of the plants are extracted with an relative error of $3,64$ % $(\pm 3,19)$ for the main stems' length and $6,73$ % $(\pm 4,77)$ for the siliques' length. Concerning the topological and geometrical traits in the interior of the plant (mainly side branches) the results are error-prone as wrong local decisions are propagated to the next level. For this reason, the correct topological reconstruction of the interior of the plant is not possible without supervision during the reconstruction procedure. This step has to be considered in future works.

We also want to reflect on the choice of methods being used along the pipeline. Using a tracing approach to gain local geometrical and topological information already during the segmentation of curvilinear objects showed to be useful when topological characteristics are in focus of the analysis. The main drawback of this technique is the incompleteness of the results if branches remain undetected during tracing along the object. To overcome this drawback, additional seed points (e.g. along a pre-defined grid) can be added before the tracing procedure is invoked. The use of such techniques is already reported in literature by different approaches [8, 28, 61].
The "local" systematic classification between branching points and crossing points achieves satisfying results if the number of involved edges in a cluster node are smaller or equal 4. The procedure gets error-prone if the number of involved segments rises or the involved segments in a cluster node are relatively small. These are typical configurations which appear in morphologically complex plants. Another drawback of this procedure is the side-effect of the error propagation. A wrong, local decision in a cluster node can lead to a propagation of this error to other cluster nodes. Substituting the local reconstruction method with a global approach, e.g. a graph based approach like finding a minimum spanning tree, could raise the quality of the results and reduce the need for an user interaction during the plant reconstruction process.

Computer vision approaches used in this work showed that these technologies are promising for the future of high-throughput phenotyping studies. While the unsupervised trait extraction using this framework is limited by the image properties and the morphological complexity of the plant, future developments may open the way to fully unsupervised trait extraction.

More general considerations for future works can be found regarding the image acquisition setup and the trait extraction. The major challenges of this work originate from the image acquisition setup. As shown during the evaluation, the biggest source of error are overlapping

branches and stems. If the plants could be unravelled before taking the images, the quality of the trait extraction would be improved. Raising the costs of unravelling the plant before acquiring the images would lower the costs for a manual correction of the final results. The definition of a trade-off between these options is needed.
An accurate topological reconstruction would rise the number of extractable traits in future works. Such traits can be the bifurcation ratio of the branching pattern or the internodal distance between types of branches. Further, with use of reliable topological and geometrical traits a context between these properties could be established. Quantitative traits describing this context can be a valuable for the understanding of the correlation between the genotype and the phenotype.

## APPENDIX A

# Plant Image Samples

In the following, results of the reconstruction process as well as the corresponding quantitative traits for 6 plant images of the dataset are shown. The plant images are overlaid with the (labeled) centerline of the plant. For illustration reasons the plant image was inverted and the centerline was thickened. Different branches are marked in different colors. The centerline of the siliques are coloured in turquoise, leafs are coloured in blue. The markers are identifying significant points along the plant:

- *Green Triangle:* Start of the main stem
- *Red Triangle:* Termination point
- *Purple "+":* Branching point

Each sample image is provided with a table summarizing the following traits:

| | |
|---|---|
| $N_{MS}$ | Number of main stems (MS) |
| $N_{SBI}$ | Number of side branches with depth 1 (SB I) |
| $N_{SBII}$ | Number of side branches with depth 2 (SB II) |
| $N_{SBIII+}$ | Number of side branches with depth 3 or deeper (SB III+) |
| $N_{Silique}$ | Total number of siliques |
| $\bar{l}_{MS}$ [Px] | Average length of MS |
| $\bar{l}_{Silique}$ [Px] | Average length of siliques |

Table A.1: Abbreviations of traits

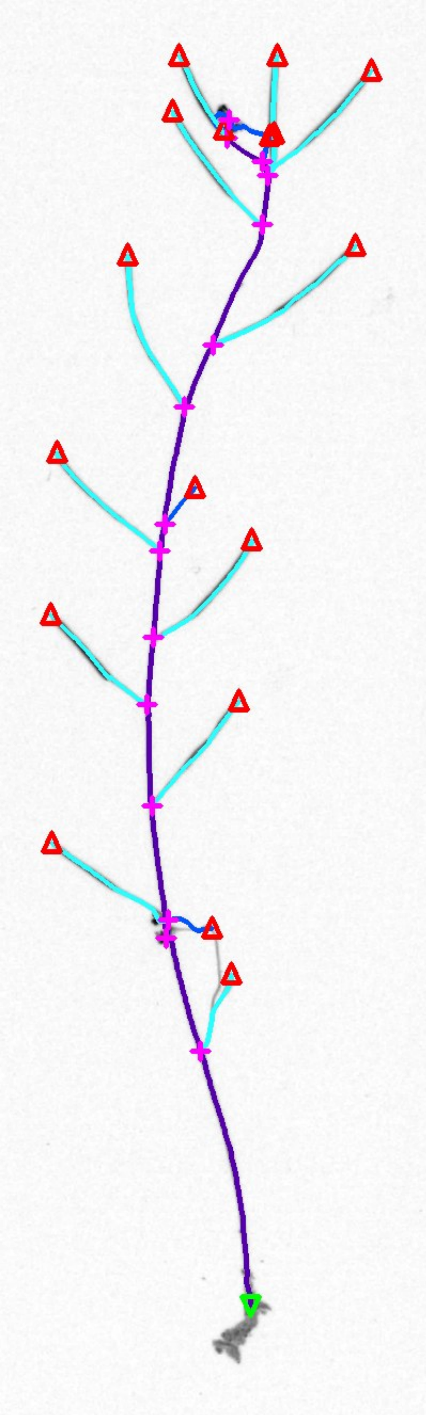


Figure A.1: *_DSC1331.tiff* - Plant Image

| Trait | GT Fiji | FW |
|---|---|---|
| $N_{MS}$ | 1 | 1 |
| $N_{SBI}$ | 1 | 1 |
| $N_{SBII}$ | 0 | 0 |
| $N_{SBIII+}$ | 0 | 0 |
| $N_{Silique}$ | 11 | 12 |
| $\bar{l}_{MS}$ [Px] | 1144 | 1225 |
| $\bar{l}_{Silique}$ [Px] | 141 | 133 |

Table A.2: *_DSC1331.tiff* - Traits

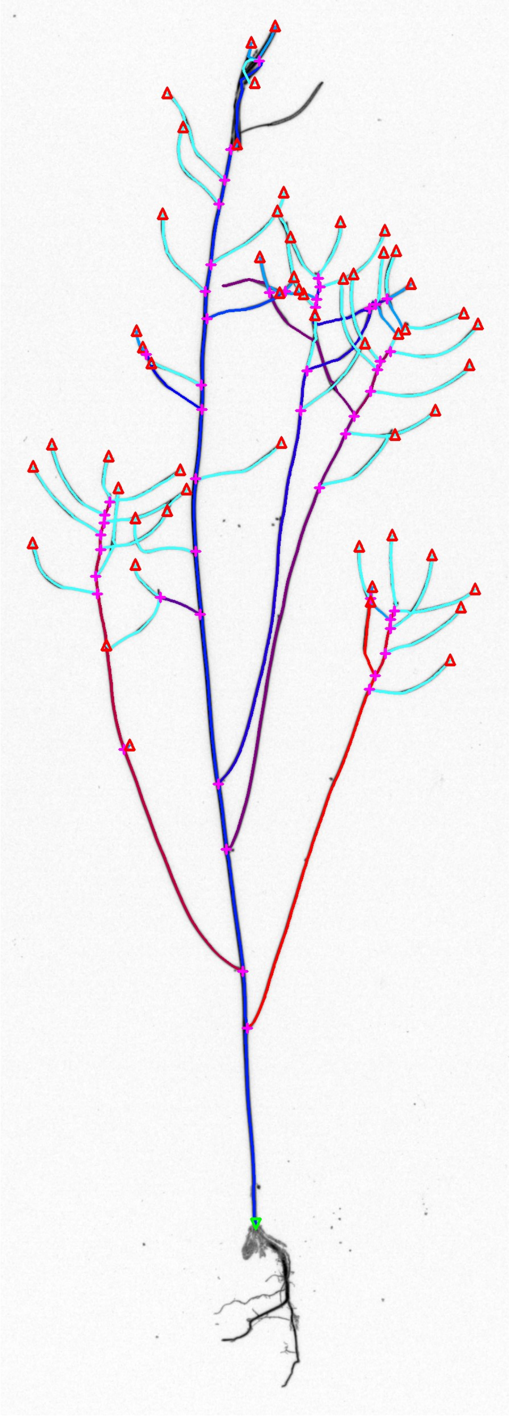


Figure A.2: *_DSC1337.tiff* - Plant Image

| Trait | GT Fiji | FW |
|---|---|---|
| $N_{MS}$ | 1 | 1 |
| $N_{SBI}$ | 3 | 5 |
| $N_{SBII}$ | 0 | 2 |
| $N_{SBIII+}$ | 0 | 3 |
| $N_{Silique}$ | 48 | 39 |
| $\bar{l}_{MS}$ [Px] | 1997 | 2176 |
| $\bar{l}_{Silique}$ [Px] | 185 | 154 |

Table A.3: *_DSC1337.tiff* - Traits

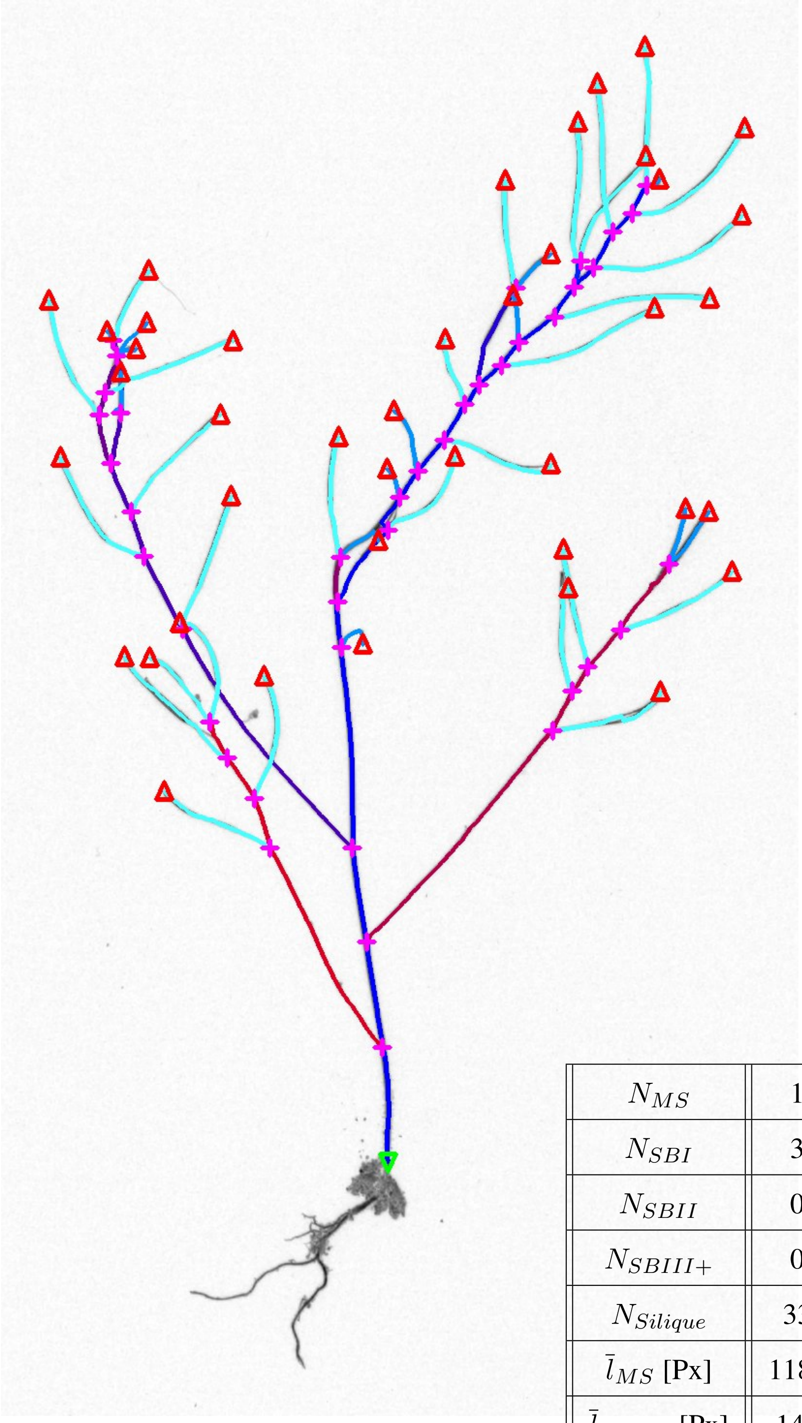


Figure A.3: *_DSC1753.tiff* - Plant Image

| | | |
|---|---|---|
| $N_{MS}$ | 1 | 1 |
| $N_{SBI}$ | 3 | 6 |
| $N_{SBII}$ | 0 | 1 |
| $N_{SBIII+}$ | 0 | 0 |
| $N_{Silique}$ | 33 | 28 |
| $\bar{l}_{MS}$ [Px] | 1185 | 1185 |
| $\bar{l}_{Silique}$ [Px] | 143 | 139 |

Table A.4: *_DSC1753.tiff* - Traits

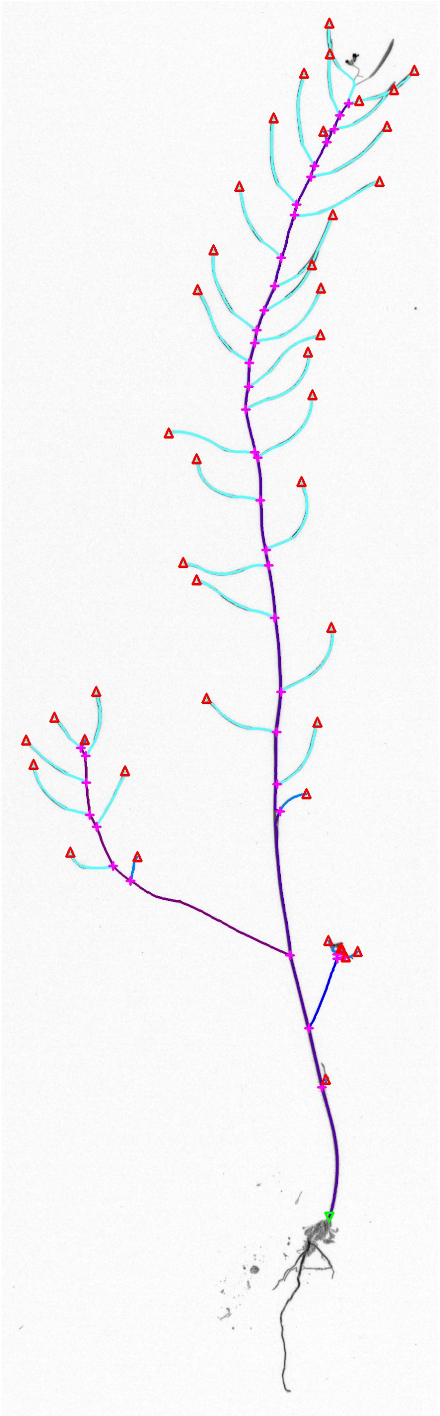


Figure A.4: *_DSC1826.tiff* - Plant Image

| Trait | GT Fiji | FW |
|---|---|---|
| $N_{MS}$ | 1 | 1 |
| $N_{SBI}$ | 2 | 2 |
| $N_{SBII}$ | 0 | 1 |
| $N_{SBIII+}$ | 0 | 0 |
| $N_{Silique}$ | 33 | 31 |
| $\bar{l}_{MS}$ [Px] | 2479 | 2441 |
| $\bar{l}_{Silique}$ [Px] | 175 | 175 |

Table A.5: *_DSC1826.tiff* - Traits

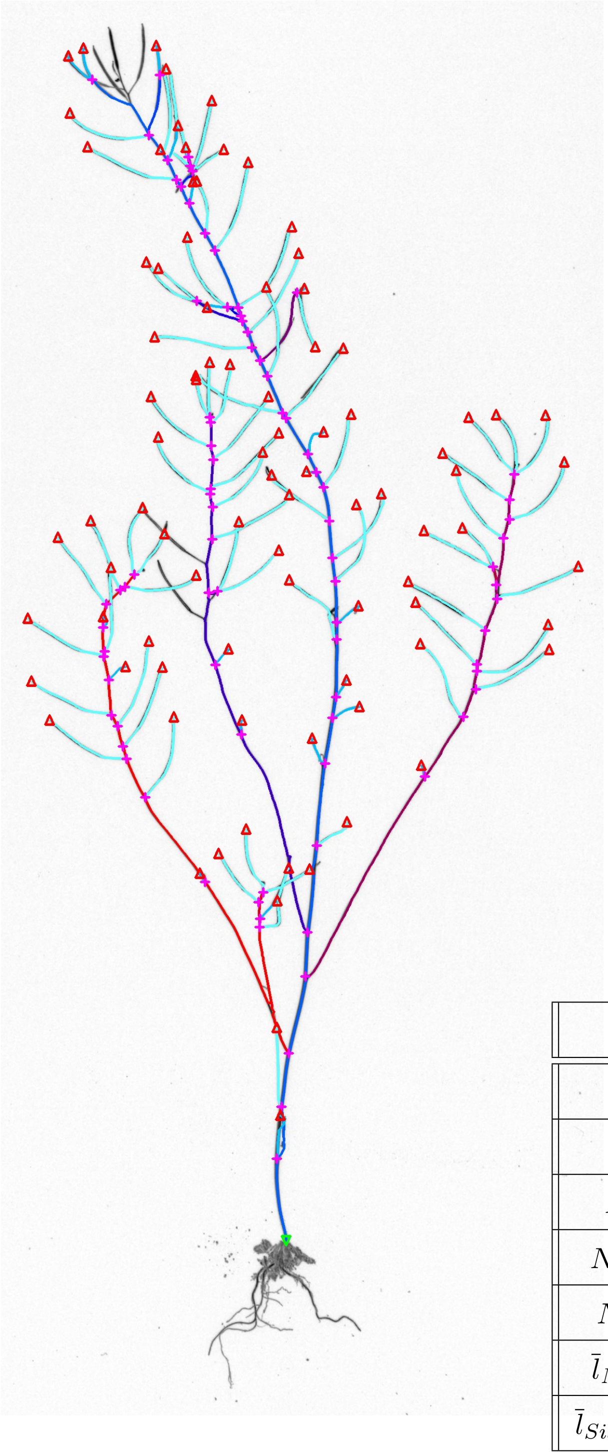


Figure A.5: *_DSC7838.tiff* - Plant Image

| Trait | GT Fiji | FW |
|---|---|---|
| $N_{MS}$ | 1 | 1 |
| $N_{SBI}$ | 4 | 6 |
| $N_{SBII}$ | 0 | 2 |
| $N_{SBIII+}$ | 0 | 3 |
| $N_{Silique}$ | 73 | 66 |
| $\bar{l}_{MS}$ [Px] | 2445 | 2516 |
| $\bar{l}_{Silique}$ [Px] | 167 | 161 |

Table A.6: *_DSC7838.tiff* - Traits

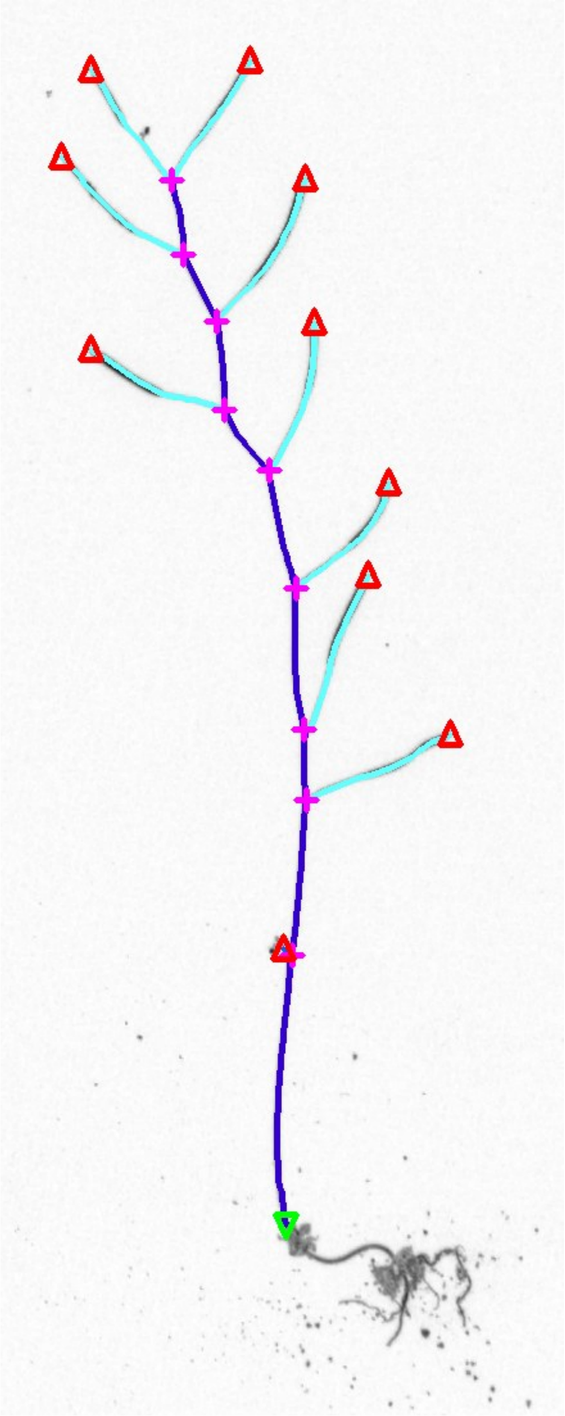


Figure A.6: *_DSC8129.tiff* - Plant Image

| | | |
|---|---|---|
| $N_{MS}$ | 1 | 1 |
| $N_{SBI}$ | 0 | 0 |
| $N_{SBII}$ | 0 | 0 |
| $N_{SBIII+}$ | 0 | 0 |
| $N_{Silique}$ | 9 | 9 |
| $\bar{l}_{MS}$ [Px] | 874 | 873 |
| $\bar{l}_{Silique}$ [Px] | 128 | 128 |

Table A.7: *_DSC8129.tiff* - Traits

# Bibliography

[1] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000.

[2] P. Armengaud, K. Zambaux, A. Hills, R. Sulpice, R. J. Pattison, M. R. Blatt, and A. Amtmann. EZ-Rhizo: integrated software for the fast and accurate measurement of root system architecture. *The Plant Journal*, 57(5):945–956, 2009.

[3] S. Arvidsson, P. Pérez-Rodríguez, and B. Mueller-Roeber. A growth phenotyping pipeline for Arabidopsis thaliana integrating image analysis and rosette area modeling for robust quantification of genotype effects. *New Phytologist*, 191(3):895–907, 2011.

[4] H.-J. Bartsch. *Taschenbuch mathematischer Formeln. 21.Auflage*. Fachbuchverlag Leipzig im Carl Hanser Verlag, 2007.

[5] P. Basu, A. Pal, J. P. Lynch, and K. M. Brown. A novel image-analysis technique for kinematic study of growth and curvature. *Plant Physiology*, 145(2):305–316, 2007.

[6] F. Benmansour, P. Fua, and E. Turetken. Automated reconstruction of tree structures using path classifiers and mixed integer programming. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 0:566–573, 2012.

[7] F. Z. Boroujeni, O. Rahmat, R. Wirza, N. Mustapha, L. S. Affendey, and O. Maskon. Coronary artery center-line extraction using second order local features. *Comp. Math. Methods in Medicine*, 2012.

[8] F. Z. Boroujeni, R. Wirza, O. Rahmat, N. Mustapha, L. S. Affendey, and O. Maskon. Automatic selection of initial points for exploratory vessel tracing in fluoroscopic images. *Defence Science Journal*, 61:443–451, 2011.

[9] B. Brachi, G. Morris, and J. Borevitz. Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, 12(10):232–240, 2011.

[10] J. E. Bresenham. Algorithm for computer control of a digital plotter. *IBM Systems Journal*, 4(1):25–30, 1965.

[11] D. Calvo, M. Ortega, M. G. Penedo, and J. Rouco. Automatic detection and characterisation of retinal vessel tree bifurcations and crossovers in eye fundus images. *Computer Methods and Programs in Biomedicine*, 103(1):28–38, 2011.

[12] J. N. Cobb, G. DeClerck, A. Greenberg, R. Clark, and S. McCouch. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. *Theoretical and Applied Genetics*, 126(4):867–887, 2013.

[13] Human Genome Sequencing ConsortiumInternational. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[14] K. K. Delibasis, A. I. Kechriniotis, C. Tsonos, and N. Assimakis. Automatic model-based tracing algorithm for vessel segmentation and diameter estimation. *Computer Methods and Programs in Biomedicine*, 100(2):108 – 122, 2010.

[15] D.E. Donohue and G.A. Ascoli. Automated reconstruction of neuronal morphology: An overview. *Brain Res Rev*, 67(1-2):94–102, 2011.

[16] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever. Multiscale vessel enhancement filtering. In *MICCAI*, pages 130–137. Springer-Verlag, 1998.

[17] M. M. Fraz, S. A. Barman, P. Remagnino, A. Hoppe, A. Basit, B. Uyyanonvara, A. R. Rudnicka, and C. G. Owen. An approach to localize the retinal blood vessels using bit planes and centerline detection. *Comput. Methods Prog. Biomed.*, 108(2):600–616, 2012.

[18] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman. Blood vessel segmentation methodologies in retinal images - a survey. *Comput. Methods Prog. Biomed.*, 108(1):407–433, 2012.

[19] A. P. French, S. Ubeda-Tomas, T. Holman, M. Bennett, and T. Pridmore. High-throughput quantification of root growth using a novel image-analysis tool. *Plant Physiology*, 150(4):1784–1795, 2009.

[20] R. T. Furbank and M. Tester. Phenomics – technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, 16(12):635–644, 2011.

[21] T. Galkovskyi, Y. Mileyko, A. Bucksch, B. Moore, O. Symonova, C. Price, C. Topp, A. I. Pascuzzi, P. Zurek, S. Fang, J. Harer, P. Benfey, and J. Weitz. GiA Roots: software for the high throughput analysis of plant root system architecture. *BMC Plant Biology*, 12(1):116–127, 2012.

[22] C. Godin, E. Costes, and H. Sinoquet. A method for describing plant architecture which integrates topology and geometry. *Annals of Botany*, 84(3):343–357, 1999.

[23] R. C. Gonzalez and R. E. Woods. *Digital image processing (3rd Edition)*. Prentice Hall, 3 edition, August 2007.

[24] T. E. Grift, J. Novais, and M. Bohn. High-throughput phenotyping technology for maize roots. *Biosystems Engineering*, 110(1):40 – 48, 2011.

[25] R.M. Haralick, S. R. Sternberg, and X. Zhuang. Image analysis using mathematical morphology. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-9(4):532–550, 1987.

[26] K. Haris, S. N. Efstratiadis, N. Maglaveros, J. Gourassas, and G. Louridas. Model-based morphological segmentation and labeling of coronary angiograms. *IEEE Trans. Med. Imaging*, 18(10):1003–1015, 1999.

[27] C. Heneghan. Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis. *Medical Image Analysis*, 6(4):407–429, 2002.

[28] Y. Huang, J. Zhang, and Y. Huang. An automated computational framework for retinal vascular network labeling and branching order analysis. *Microvascular Research*, 84(2):169 – 177, 2012.

[29] P. Kakumanu, S. Makrogiannis, and N. Bourbakis. A survey of skin-color modeling and detection methods. *Pattern Recogn.*, 40(3):1106–1122, 2007.

[30] K.M. Leon-Kloosterziel, C.J. Keijzer, and M. Koornneef. A seed shape mutant of arabidopsis that is affected in integument development. *Plant Cell*, 6(3):385–392, 1994.

[31] K.-S. Lin, C.-L. Tsai, C.-H. Tsai, M. Sofka, S.-J. Chen, and W.-Y. Lin. Retinal vascular tree reconstruction with anatomical realism. *Biomedical Engineering, IEEE Transactions on*, 59(12):3337–3347, 2012.

[32] M. H. Longair, D. A. Baker, and J. D. Armstrong. Simple neurite tracer: Open source software for reconstruction, visualization and analysis of neuronal processes. *Bioinformatics*, 2011.

[33] M. E. Martínez-Pérez, A. D. Hughes, A. V. Stanton, S. A. Thom, A. A. Bharath, and K. H. Parker. Retinal blood vessel segmentation by means of scale-space analysis and region growing. In *MICCAI*, pages 90–97, 1999.

[34] M. E. Martínez-Pérez, A. D. Hughes, A. V. Stanton, S. A. Thom, N. Chapman, A. A. Bharath, and K. H. Parker. Retinal vascular tree morphology: a semi-automatic quantification. *IEEE Trans. Biomed. Engineering*, 49(8):912–917, 2002.

[35] E. Meijering. Neuron tracing in perspective. *Cytometry Part A*, 77A(7):693–704, 2010.

[36] A. Naeem, A. P. French, D. M. Wells, and T. Pridmore. High-throughput feature counting and measurement of roots. *Bioinformatics*, 27(9):1337–1338, 2011.

[37] U. T. V. Nguyen, A. Bhuiyan, L. A. F. Park, and K. Ramamohanarao. An effective retinal blood vessel segmentation method using multi-scale line detection. *Pattern Recognition*, 46(3):703 – 715, 2013.

[38] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.

[39] E. Peli. Contrast in complex images. *J. Opt. Soc. Am. A*, 7(10):2032–2040, 1990.

[40] J. Pevsner. *Bioinformatics and functional genomics*. Wiley-Blackwell, 2009.

[41] WinRHIZO Pro. Winrhizo pro 2004a software: Root analysis. *Regent Instruments Inc., Quebec, Canada*, 2004.

[42] F. K. H. Quek and C. Kirbas. Vessel extraction in medical images by wave-propagation and traceback. *IEEE Transactions on Medical Imaging*, 20:117–131, 2001.

[43] C. Restif, C. Ibanez-Ventoso, M. Driscoll, and D. N. Metaxas. Tracking c. elegans swimming for high-throughput phenotyping. In *ISBI - International Symposium on Biomedical Imaging*, pages 1542–1548, 2011.

[44] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[45] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. John Wiley & Sons, Inc., New York, NY, USA, 1987.

[46] M. Sonka, V. Hlavac, and R. Boyle. *Image processing, analysis, and machine vision*. Thomson-Engineering, 2007.

[47] R. Sozzani and P. Benfey. High-throughput phenotyping of multicellular organisms: finding the link between genotype and phenotype. *Genome Biology*, 12(3):219–225, 2011.

[48] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging*, 23(4):501–509, 2004.

[49] R. Subramanian, E. P. Spalding, and N. J. Ferrier. A high throughput robot system for machine vision based plant phenotype studies. *Machine Vision and Applications*, 24(3):619–636, 2013.

[50] Y Sun. Automated identification of vessel contours in coronary arteriograms by an adaptive tracking algorithm. *IEEE Trans Med Imaging*, 8(1):78–88, 1989.

[51] R. Szeliski. *Computer vision: Algorithms and applications*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[52] D. Todorovic. Gestalt principles. *Scholarpedia*, 3(12):5345, 2008.

[53] Y.A. Tolias and S.M. Panas. A fuzzy vessel tracking algorithm for retinal images based on fuzzy clustering. *Medical Imaging, IEEE Transactions on*, 17(2):263–273, 1998.

[54] E. Türetken, F. Benmansour, B. Andres, H. Pfister, and Fua. Reconstructing loopy curvilinear structures using integer programming. In *CVPR*, pages 1822–1829, 2013.

[55] E. Türetken, G. González, C. Blum, and P. Fua. Automated reconstruction of dendritic and axonal trees by global optimization with geometric priors. *Neuroinformatics*, 9(2-3):279–302, 2011.

[56] S. Verboven and M. Hubert. LIBRA: a MATLAB library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75:127–136, 2004.

[57] D. Weigel. Natural variation in arabidopsis: From molecular genetics to ecological genomics. *Plant Physiology*, 158(1):2–22, 2012.

[58] A. G. White, P. G. Cipriani, H.-L. Kao, B. Lees, D. Geiger, E. Sontag, K. C. Gunsalus, and F. Piano. Rapid and accurate developmental stage recognition of c. elegans from high-throughput image data. In *CVPR*, pages 3089–3096. IEEE, 2010.

[59] Y. Yin, M. Adel, and S. Bourennane. Retinal vessel segmentation using a probabilistic tracking method. *Pattern Recognition*, 45(4):1235–1244, 2012.

[60] F. Zana and J. C. Klein. Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation. *Trans. Img. Proc.*, 10(7):1010–1019, 2001.

[61] Y. Zhang, X. Zhou, A. Degterev, M. Lipinski, D. Adjeroh, J. Yuan, and S.T.C. Wong. A novel tracing algorithm for high throughput imaging screening of neuron-based assays. *J Neurosci Methods*, 160(1):149–62, 2007.