Norwegian University
of Life Sciences

Master's Thesis 2020 30 ECTS
Faculty for Science and Technology (REALTEK)

Automation of the labeling of images of sugar beet cultivation with hyperspectral imaging

Annika Jäkel
Data Science

Supevisors

Prof. Dr. Ingunn Burud
Faculty of Science and Technology, Norwegian University Of Life Sciences, Ås, Norway

Dr. Julia Osten
Fraunhofer Institute for Transportation and Infrastructure Systems (IVI), Dresden, Germany

Contents

List of Tables v
List of Figures vii
List of Abbreviations xi
1 Introduction 1
1.1 Challenges for agriculture 1
1.2 Importance and techniques of weed control 1
1.3 Autonomous weeding robots 2
1.4 Sugar beet 3
1.5 Contribution and outline of this thesis 3
2 Theory and related work 5
2.1 Theory 5
2.1.1 Computer vision 5
2.1.2 Spectral imaging 6
2.1.3 Classification methods for spectral data 11
2.1.4 Image registration and alignment 15
2.2 Related work 18
2.2.1 Crop and weed detection without spectral imaging 18
2.2.2 Crop and weed detection with spectral imaging 19
2.2.3 Automated labeling for crop and weed detection 20
3 Material and Methods 22
3.1 Plants 22
3.2 Cameras 23
3.3 Data acquisition 23
3.4 Analysis and classification of the spectral data 26
3.4.1 Pre-processing of the spectral data 26
3.4.2 Scatter correction and normalization 28
3.4.3 Ground truth segmentation masks 28
3.4.4 Filter tests 28
3.4.5 Classifier: Neural networks 28
3.4.6 Train-test-evaluation split 30
3.5 Image registration of the RGB and spectral images 32
4 Results 37
4.1 Analysis and classification of the spectral data 37
4.1.1 Determining the best bandpass filter for distinguishing between sugar beet and weed 37
4.1.2 Classification based on the spectral data 37
4.1.3 Classification of the spectral data 38
4.2 Image registration of the RGB and the spectral images 52
4.2.1 Filtering of homographies 53
4.3 Projection of multispectral classification data onto the RGB plane . 55

5 Discussion . . . 57
5.1 Materials and Methods . . . 57
5.2 Analysis and classification of the spectral data . . . 58
5.2.1 Spectral imaging for plant classification . . . 58
5.2.2 Deep learning for plant classification . . . 59
5.2.3 Features for the classification of plant species . . . 61
5.3 Image registration . . . 62
6 Conclusions . . . 63
References . . . 65
Appendix . . . 74

List of Tables

1 Sensed wavebands indicated by ”x” for different bandpass filter options for the hyperspectral camera MV1-D2048x1088-HS02-96-G2 (Photonfocus AG) [88], lens: Edmund Optics, 35 mm/F1.65 67716 VIS-NIR. 25

2 Apertures tested and corresponding exposure times for the hyperspectral camera MV1-D2048x1088-HS02-96-G2 (Photonfocus AG), lens: Edmund Optics, 35 mm/F1.65 67716 VIS-NIR, bandpass filters 600 - 875 nm. . . . 27

3 Dataset for testing different bandpass filter options for the hyperspectral camera MV1-D2048x1088-HS02-96-G2 (Photonfocus AG) and bandpass filters from Edmund Optics (35 mm/F1.65 67716 VIS-NIR), half of the number of images are sugar beet, the other half weed. 31

4 Number and percentage of pixels for a pixel classifier divided into background, sugar beet and weed of a timeseries of hyperspectral images (15 bands, red - near infrared region, size of one image: 409 x 216 pixels) for different apertures, spectral data acquired with MV1-D2048x1088-HS02-96-G2 (Photonfocus AG), lens:Edmund Optics, fcocal length (f): 35 mm/F1.65 67716 VIS-NIR, bandpass filters 600 - 875 nm. 31

5 Grades and corresponding criteria for visual assessment of superimposed image of a RGB image and the corresponding, aligned hyperspectral image, ghosting refers to blur of not perfectly overlapping parts. 34

6 Confusion matrix of Partial Least Squares - Discriminant Analysis pixel classifications of calibrated multispectral images of sugar beet and weed plants, cultivated in the lab ($n_{\text{images}} = 863$), all numbers in % based on the true labels. 39

7 Confusion matrix for the evaluation dataset pixel classifications of a neural network with one convolution over the spectral domain for multispectral images (15 channels) ($n_{\text{images}} = 863$, each image 409 x 216 pixels, 15 % of pixels belong to the evaluation dataset), all numbers in % based on the true labels. 40

8 Confusion matrix for the evaluation dataset for pixel classifications of a neural network with ten fully connected layers for distinghuishing between soil and plant pixels of multispectral images (15 channels) ($n_{\text{images}} = 863$, each image 409 x 216 pixels, 15 % of pixels belong to the evaluation and validation dataset respectively), separate refers to sugar beet and weed in separate pots and with different exposure times and mixed to sugar beet and weed in one pot with the same exposure time, all numbers in % based on the true labels, dice coefficient for the evalutation set. 40

9 Confusion matrix for the evaluation dataset for pixel classifications of a neural network with ten fully connected layers for distinghuishing between sugar beet and weed pixels of multispectral images (15 channels) ($n_{\text{images f/2.8}} = 318, n_{\text{images f/6}} = 109, n_{\text{images f/11}} = 179, n_{\text{images f/8}} = 168, n_{\text{images f/14}} = 89$, each image 409 x 216 pixels, 15 % of plant pixels belong to the calibration and evaluation dataset respectively), separate refers to sugar beet and weed in separate pots and with different exposure times and mixed to sugar beet and weed in one pot with the same exposure time, all numbers in % based on the true labels, dice coefficient for the evalutation set. 42

10 Counts of folders for which certain metrics were useful to discriminate between classes of image alignment performance based on visual assessment of boxplots, maximum: 17, n_{folder} between 20 and 60, spatial refers to a the position and order of the projected four corner points, sharpness refers to the variance of the laplacian filtered image. 54

11 Average of visual assessment grade for different homography filtering options, based on 15 randomly picked images of each of the 17 folders ($n_{\text{per folder}} = 15, n_{total} = 255$, assessment grade 1: best possible alignment, 2: good - okay alignment with ghosting but no doppelgaenger, 3: alignment with doppelgaenger, MSE means Mean square error, sum corr refers to the sum of the correlation matrix between the overlapping part of the aligned images and spatial filtering was a check of the order and angles of the warped four corner points. 55

12 Performance of best filtering option for homography and the ground truth for the visual assessment grades, assessment grade 1: best possible alignment, 2: good - okay alignment with ghosting but no doppelgaenger, 3: alignment with doppelgaenger, MSE means Mean square error, sum corr refers to the sum of the correlation matrix between the overlapping part of the aligned images and spatial filtering was a check of the order and angles of the warped four corner points. 55

List of Figures

1 Types of computer vision tasks, credits: Waleed Abdulla, source: [3] 5

2 Difference between structure of image arrays or cubes depending on the sensor type, source: see graphic. 6

3 Difference between spectra of soil and plant reflectance values based on sensor, source: see graphic. 7

4 Decreasing apertures and increasing f-number which is the denominator in the term f/f-number illustrated with camera lense, credits: Wikipedia user KoeppiK, source: [59]. 9

5 Aberrations of convex lenses: (a) Spherical aberration, Andrei Stroe, source: [6] , (b) Chromatic aberration, credits: Wikipedia user Andres 06, source: [5]. 9

6 Sketch of the reflectance in % of two made-up spectra relative to a reference spectrum, graph based on Stefansson, 2019, p. 43 [102]. 11

7 Sketch of a simple, fully connected neural network with an example activation function for one node, based on a sketch of Adrian Rosebrock [92], edited. 13

8 Sketch of projection, one type of 2D planar transformations, content based on Szeliski 2011, p. 311 [105]. 15

9 Plants for experiments about classifying sugar beet and weed pixels with hyperspectral imaging in the near infrared region, 12.2.2020. 22

10 Snapshot mosaic CMV2K-SM5x5-NIR sensor (25 channels, 600 - 900 nm) of Photonfocus AG's camera MV1-D2048x1088-HS02-96-G2, source: [88]. . 24

11 Measurement setup in the lab for the acquisition of optical and hyperspectral images of sugar beet and weed with. 24

12 Dataset of image pairs of optical and hyperspectral images of sugar beet and weed in different pots, images acquired with different exposure times for the hyperspectral camera (left) and sugar beet and weed photographed with the same exposure time for the hyperspectral camera (right) and two pots with sugar beet and weed plants together. 27

13 Examples of projections of a region of interest of a hyperspectral image onto a RGB image plane, images acquired with a stereo camera system (camera casings 0.5 cm apart), orange indicates a wrong projection, based on angles between contour lines and order of corner points, green indicates a right projection based on the same criteria. 35

14 Examples for superimposed, aligned RGB and spectral images of plants and corresponding visual assessment grades (1: Best possible quality, 2: Good-okay, 3: Not usable) for the quality of the projection as a result of homography based image registration of an RGB and a hyperspectral image (near infrared region), red box for (a) indicates parallax, red box for (c) doppelgaenger. 35

15 Boxplots of mean squared error (mse) grouped by visual assessment grade of image alignment (1: best, 2: good, 3:very bad), homography based image alignment of RGB and hyperspectral images ($n_{imagepairs} = 33$, one folder) with opencv 36

16 Extended multiplicative scatter corrected and L2 normalized, calibrated average spectra of weed and sugar beet plants cultivated in the lab for (a) No bandpass filter ($n_{images} = 46$), (b) Bandpass filter 600 - 875 nm ($n_{images} = 68$), (c) Bandpass filter 675 - 975 nm ($n_{images} = 64$), each image consisted in 409 x 216 pixels with varying fraction of plant pixels. 38

17 Average calibrated absorbance for different apertures of background (potting soil, plastic pots, table), sugar beet and weed plants [%] grown in pots, $n_{\text{images separate}} = 606$, $n_{\text{images mixed}} = 257$, $n_{\text{images background}} = 863$, resolution each image: 409 x 216 pixels with varying fraction of plant pixels. 39

18 Extended multiplicative scatter corrected (EMSC) and L2 normalized, calibrated average spectra of weed and sugar beet plants cultivated in the lab for (a) degree(emsc) = 0, (b) degree(emsc) = 1, (c) degree(emsc) = 2, each image consisted in 409 x 216 pixels with varying fraction of plant pixels, $n_{\text{images}} = 863$. 43

19 a: Original image, b: Prediction of a neural network with 10 fully connected layers for distinghuishing between soil and plant for multispectral images (15 channels), purple: background, yellow: plant 44

20 Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/2.8 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 318 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out. 45

21 Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/6 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 109 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out. 46

22 Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/11 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 179 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized multispectral images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out. 47

23 Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/2.8, f/6, f/11 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 606 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized multispectral images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out. 48

24 Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/8 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 168 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized multispectral images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out. 49

25 Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/14 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 89 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized multispectral images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out. 50

26 Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/14 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 89 calibrated, extended multiplicative scatter corrected (degree = 2) and L2 normalized multispectral images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out. 51

27 Superimposed grayscale versions of RGB image and aligned multispectral image (15 bands, 675 - 800 nm) of several weed species with the imregdemon algorithm from matlab. 52

28 Correct and incorrect matches of image registration process of grayscake versions of an RGB image and a multispectral image (15 bands, 675 - 800 nm). 52

29 Counts of how many times a channel of multispectral images (15 bands, 675 - 800 nm) was used for grayscale conversion that led to most matches between and optical image and multispectral image during image registration process with Scale invariant feature transform feature detection (opencv 3.4.2). 53

30 All predictions, pixel or bounding boxes were projected from a multispectral image onto the image plane of a multispectral image, classes (sugar beet or weed) were not taken into account for determining the quality of the projection, Figure a: Warped pixel predictions, b: Bounding boxes for overlapping plants, c: Empty bounding box after warping, d: Succesuful bounding box projection. 56

31 Average calibrated reflectance in % for different apertures of background (potting soil, plastic pots, table), sugar beet and weed plants, $n_{\text{images separate}=606}$, $n_{\text{images mixed}=257}$, $n_{\text{images background}=863}$, resolution one image: 409 x 216 pixels with varying number of plant pixels. 74

List of abbreviations

fig. Figure

tab. Table

nm Nanometer

ms Milliseconds

f Focal Length

RGB Normal color image with a red, a green and a blue channel

PCA Principal Component Analysis

KNN k nearest neighbours classifier

SVM Support Vector Machine

ANN Artificial Neural Network

CNN Convolutional Neural Network

PLS-DA Partial Least Square - Discriminant Analysis

NDVI Normalized Difference Vegetation index

NIR Near-infrared

MSE Mean Squared Error

SSIM Structural Similarity Index

RANSAC Random Sample Consensus algorithm

SIFT Scale Invariant Feature Transform

TP True Positives

FP False Positives

TN True Negatives

FN False Negatives

Acknowledgements

A big thank you to everyone who accompanied my learning journey of this Master's thesis: Thank you for the constructive feedback and guidance to my main supervisors Ingunn Burud and Julia Osten - and Andreas Backhaus, Hans-Christian Klück and Sebastian Warnemünde who became important mentors. It was a pleasure to work with you. Three other special thank yous go to Denise Ackermann, Jasper Zevering and my Mom who supported me along the way. Last but not least, thanks to everyone at Fraunhofer IVI in Dresden for the very good and supporting working atmosphere.

Abstract

The objective of this thesis was to develop an automated labeling system for RGB images (red green blue) of sugar beet and weed plants with the help of multispectral imaging. 863 image pairs of sugar beet and 18 weed species, consisting each in one RGB and one multispectral image of the same plants, were acquired in the lab. Different apertures and bandpass filters were tested and the multispectral camera captured 15 wavebands between 654 and 866 nm. The pixels of the multispectral images were classified with a pipeline of two fully connected artificial neural networks (ANN) of the same architecture of ten hidden layers. The first ANN distinguished between plants and background, and the second one between sugar beet and weed. The transfer of the classifications based on the multispectral images onto the RGB images was attempted with a local motion model (imregdemon, Matlab) and a global motion model (projection). One projection matrix (homography) was computed for each acquisition session during which the camera position did not change and plants had a similar height. The best homographies were chosen based on spatial parameters, the mean squared error and the sum of the correlation matrix between the RGB and the warped spectral image.

The classification accuracy for the background versus plant classifier were ¿ 98 % for both classes. The sugar beet versus weed classifier reached a per-class accuracy between 73 and 95 % and dice coefficients between 0.71 and 0.92 for the evaluation data set. On a plant level, classification results were very satisfying if plants of the same age (sugar beet) and species (weed) had been included in the training data set. The application of the local motion model failed most likely due to huge differences in resolution, reflectance values and image sections. After the filtering of the projection matrices, 95 % of the image pairs reached a satisfying projection of bounding boxes. The accuracy of projection was not high enough for conveying pixel segmentation masks. This could be achieved by further applying a local-motion model.

The overall goal, to automatically label sugar beet and weed plants, was achieved for bounding boxes. Nevertheless, the system, especially the image registration can be further improved regarding reliability and performance. The developed labeling system will be tested with field data.

1 Introduction

1.1 Challenges for agriculture

Current projections for 2050 estimate that food production has to increase by 70 % in order to meet the increasing demand for food caused by a growing world population [31] [110]. Another factor for the rising food demand is the global trend to eat more animal products due to a higher income and resulting diet changes [31, 110]. Meeting this demand with limited and decreasing natural resources like arable land, oil, water, fertilizers, and climate change will be challenging and can be achieved to 80 % by increasing the intensity of farming and not the expansion of agricultural land [31, p. 1] [58, p. I]. Furthermore, agricultural production has to become more resource-efficient and sustainable in order to protect hardly renewable resources like fertile soil, atmosphere, biodiversity and groundwater levels [58, p. I]. Also, the workforce for the agricultural sector will become scarcer since higher income opportunities await employees in the cities [32, p. 6]. The current agricultural production systems can not fulfill these requirements and need to be transformed [110, p. 1]. Organic agriculture claims to use fewer resources and results in 30 % more biodiversity than conventional farming systems, but also produces 20 - 25 % less yield [58, p. 24]. Therefore, using only the current organic farming techniques would not solve the problem. One possible solution could be precision farming which uses technology to save resources by, for instance, identifying variability regarding nitrogen or weeds in a field and precisely treating every region with the minimal amount of fertilizer or pesticides necessary [121, p. 172]. Optimal plant cultivation, reduced costs, higher yields, and lower environmental impact are the promises of precision farming [34, p. 667] [116, p. 1] [17]. Precision farming combines sensors, robotics, computer science, computer vision, agricultural sciences and remote sensing [47, p. 218].

1.2 Importance and techniques of weed control

Plants are the basis of our nutrition as the largest contributor to human daily calorie-intake [32, p. 24] or as fodder for the consumed animal products. In order to maintain and increase yields, plants must be protected against abiotic stressors, such as drought, and biotic stressors like pests, diseases and weeds. Among the biotic stressors, weeds cause the highest potential yield losses [83]. Weeds are unwanted plants on agricultural fields that compete for resources like light, water and nutrients with the crop [116, p. 2]. Without weed control, yields will be lower, and the extent of yield loss depends on the crop, the climate and the growing system [58, p. 11]. For row crops such as sugar beet and potato, 40 to 50 % of the yield would be lost without weeding, for other crops that stand closer together 30% of loss were estimated [58, p. 11], [99], [83]. Weeds can be controlled via herbicides, mechanical removal (machines or manually) and crop rotation, among others [58, p. 11]. Herbicides are widely and generously used since 1950/60 because spraying is faster and cheaper than weed removal by manual labor or tillage, especially in developed countries [43, pp. 1099, 1103] [44, p. 1] [13, pp. 1048-1049]. Reduced tillage is considered beneficial for soil protection and causes less water loss, which is relevant in arid regions [13, pp. 1048-1049]. Also, mechanical weed removal in-row without manual labor is often impossible. Consequently, herbicides have the highest share regarding active substances in crop protection globally [58, p. 4] with 37 % of active substances in chemical plant protection worldwide [37, p. 1] and in Germany with 35 % [114] [60, p. 712]. Even though

the infestation with weeds is spatially heterogeneous between and within fields, herbicides are usually applied homogeneously over the whole field [49, p. 637] [121, p. 172].
The heavy use of herbicides entails several negative consequences. Herbicide resistance has occurred in over 200 weed species worldwide, and there are more examples of site-specific herbicide resistance [48, p. 1306]. The reason is presumably that some farmers exclusively used one type of chemical plant protection instead of varying between different chemical formulas and including mechanical weed removal [13, p. 1037] [41, p. 390]. Studies have indicated that fewer herbicides could be used without any yield losses and even economic gains [37, p. 1] [101, p. 4](see review in [58, p. 13]). The reduced usage of herbicides without yield loss would also benefit the farmers who regard pesticides as pricey, but lack effective and cheaper alternatives [58, p. 17]. Lately, potential environmental and health hazards of pesticides have been discussed controversially by the public and the scientific community [95] [58, pp. 14-16]. Regulations for the application and storage of pesticides are stringent in Europe and might become even stricter due to public pressure. The amount of pesticides and their derivatives is checked regularly. In Germany, an alarming drop in the number and diversity of insects was reported in 2017, and the main driver is assumed to be landscape use with less plant diversity and pesticide applications [97, p.1] [93]. Furthermore, the German Federal Environmental Agency and the European Parliamentary Research Service stated that the current amount of pesticides might have led to damage of insects, birds, soil microbiota and pollinators and other animals through the food chain [114] [58, p. 8]. In countries with fewer restrictions for chemical plant protection, more pesticides and their residues might occur in the groundwater, soil and food.

The current industrialized agricultural system is designed to be workforce- and cost-efficient and was developed during a time with cheap petroleum [81], enabling the majority of society to work in other areas. This resulted in vast monocultures that can be cultivated easily with large and heavy machinery that traverses the fields several times a year. For one, this results in soil degradation through compaction which reduces fertility [16, pp. 515 - 517], meaning the ability of the soil to nurture and host plants [16, p. 379]. A workforce-efficient solution to avoid soil degradation and petroleum usage by heavy machines could be light, electrified and autonomous field robots.

1.3 Autonomous weeding robots

The environmental and economical costs caused by homogeneous spraying of chemical plant protection can be alleviated by precision weed management [49, p. 637]. Precision herbicide applications would reduce most expenditures of the cultivation of cereals, sugar beet and maize [87, p. 194]. One possible solution for the described problems consists of autonomous, electrical weeding robots that remove weed either mechanically or with the lowest amount of herbicides possible. In the case of electricity from renewable sources, this can additionally decrease petroleum usage in agriculture. Many research groups [121, p. 176] and companies have already attempted to develop such robots, but the variability of agricultural fields makes it a challenging task [121]. Few first robots are already on the market or ready for sales but will be most likely used by a small number of early adopters until the technology matures and becomes cheaper (examples: Contadino by Continental (prototype) [23], Farmdroid [33], Naio Technologies Dino [107]). Even though such robots

are expensive, they could pay off for European farmers due to the trends towards growing field sizes and decreasing workforce in agriculture. Furthermore, more farmers are highly educated and see themselves as entrepreneurs who may be willing to adopt high-tech solutions [64, p. 2], especially for crops with a high market value such as sugar beet.
One of the main obstructions for field robots has been the reliable and precise weed and crop detection with computer vision due to the high variability in plants and the environment in agriculture [99] [19, p. 1] [10, p. 153]. Especially for the highly variable conditions in fields and plant phenotypes, deep learning computer vision methods are well suited and convolutional neural networks (CNN) are state of the art for image classification and segmentation [112, 11]. The biggest obstacle to training a CNN that can robustly classify weeds and crops under various conditions is the large amount of training data needed, consisting in images with annotations of crops and weeds that cover a large range of natural variability [9, p. 1] [123, pp. 5128, 5134]. Manually annotating thousands of images is very time-consuming and expensive.

1.4 Sugar beet

In Germany, one important cash-crop is sugar beet [124], the alternative to sugar cane for colder climates [15, pp. 174 - 176]. The breeding and cultivation of sugar beet have been incentivized by German authorities from the 18th century on to become independent of sugar cane deliveries [124], because the supply with sugar from sugar cane was often stalled by wars and trade blockades [35]. Nowadays, sugar cane is still the main source for sugar with 75 %, sugar beet contributes the other 25 % to the world's sugar production [15]. Based on FAO statistics in 2019, 8.4 % of the daily human energy intake is covered by sugar and sweeteners, which makes it the 4th most important energy source for humans [32, p. 24]. This contribution is higher than milk products (4.8 %) and meat (5 %) [32, p. 24]. For a high sugar beet yield, weed control is crucial since the youth development is very slow and the distance between plants is 45 - 50 cm [30, p. 127], which provides enough space and time for weeds to grow [60, p. 712]. In Germany, one herbicide application before sowing and three to four applications in a two-week interval after germination are common practice in order to secure yields [60, pp. 711-712] [90] [115]. Most prevalent and resistant weeds are part of the *Chenopodium* genus, the *Polygoneae* family, canola, weed turnips and volunteer potatoes [60, p. 709]. Due to the big potential in herbicide reduction, and the high market value of sugar beet, the use of electrified weeding robots is attractive.

1.5 Contribution and outline of this thesis

This thesis is written in collaboration with the Fraunhofer Institute for traffic and infrastructure systems (Fraunhofer IVI, Dresden) and with support of the Fraunhofer Institute for Factory Operation and Automation (Fraunhofer IFF, Magdeburg), which are two of the participating institutes in the Fraunhofer framework project Cognitive agriculture (COGNAC, www.cognitive-agriculture.de). The project aims at developing the digital and electrical infrastructure and corresponding tools for agriculture like a data space, sensors and automation concepts for farming [36]. One part of the project is to develop an autonomous, electrical weeding robot for sugar beet. The robot is supposed to recognize weed and sugar beet with a red-green-blue (RGB) camera and deep learning. To train the deep learning application, a large amount of images is needed.

This thesis's contribution is to develop an automated labeling system for RGB-images of sugar beet and weed with hyperspectral images (600 - 900 nm). Because two separate cameras were used, one RGB-camera and one hyperspectral camera, spatial matching of the images of the two cameras was necessary. Therefore, the thesis is split into two parts: The classification of sugar beet, weed and background based on the spectral data and the transfer of the classifications onto the RGB's. Following questions will be evaluated:

Classification based on spectral data

- Can hyperspectral imaging be used for safely labeling sugarbeet and weed without considering spatial features?
- What are the best camera configurations for the used camera regarding wavebands and aperture for the classification of sugar beet and weed?
- What classification method is most successful?

Image registration

- Is it possible to automatically transfer classification masks or bounding boxes from the hyperspectral images with the RGB images?
- How well does the image alignment work?
- What methods are most suitable for image alignment for this case?

The image data for this thesis was acquired in the laboratory. Data acquisition under field conditions began ends of April 2020 since sugar beet was sown in the beginning/midth of April 2020. Therefore, field data could not be included in this thesis. First, an overview of the used techniques such as spectral imaging and deep learning will be given, as well as a review of related scientific work. This is ensued by the description of the used equipment, plants and methods and results of the analysis. Then, the results of this work will be discussed and compared to similar scientific works. The outlined questions will be answered in the conclusions part, along with a summary of the thesis.

2 Theory and related work

2.1 Theory

2.1.1 Computer vision

Computer vision aims at enabling computers to analyze and understand image data similar to human beings [52, p. 1] [100, p. 1]. The input of a computer vision system consists of images or video frames and optionally additional information like camera position or geo-coordinates [18, p. 2]. The output can be a transformed representation of the input image, such as the removal of blurriness in medical images caused by movements of the patient. Another output form is a ”decision” like classifying the image content, detecting faces or determine the number of apples in a picture [18, p. 2]. The main goal of computer vision is to replace human vision and thereby humans in many tasks like analysis of medical images, driving or quality control in industrial production [126, ch. 13]. These kinds of tasks are very challenging for computers [18, p. 4] that are better with static concepts and forms than abstract concepts with a variety of possible forms. For instance, chairs exist in many distinct designs, and additionally, an image of a chair looks very different depending on the illumination and perspective. Computer vision can be divided into four types of tasks [3], (see fig. 1):

- **Image classification:** What is that an image of? (Example: Balloons.)
- **Semantic segmentation:** To which class does this pixel belong?
- **Object detection:** Where are different objects located, what size are they and what type of object is it?
- **Instance segmentation**: Where are the objects in the image, which pixels belong to each individual object and what type of object is it?

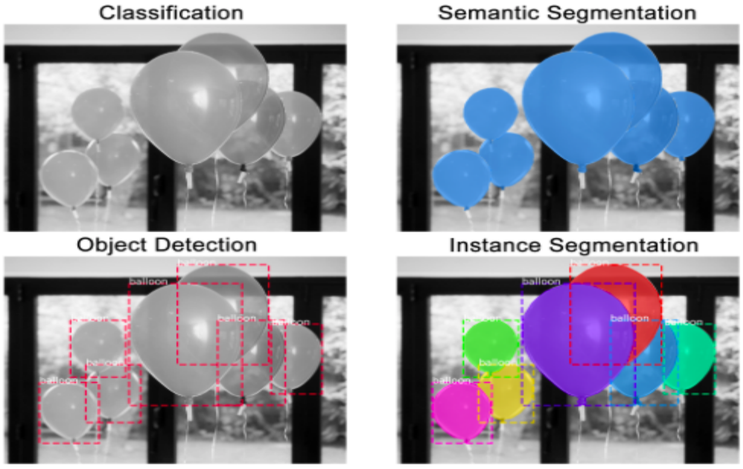


Figure 1: Types of computer vision tasks, credits: Waleed Abdulla, source: [3]

2.1.2 Spectral imaging

Each chemical element and molecule absorbs or reflects certain wavelengths especially strongly, resulting in a characteristic spectral profile [72, pp. 117 - 118]. Since imaging sensors measure the radiance intensity for specific wavebands, and transform it first to an electrical and then a digital signal, the strength of reflectance (or emission) of the photographed object is captured for the specified wavebands [72, p. 2]. Consequently, images contain spectral and spatial information about the photographed objects [72, p. 117]. For instance, healthy plants with a lot of chlorophyll reflect wavelengths around 520 - 540 nm strongly and appear green because of that [72, p. 144]. Multispectral cameras have a broadband sensor that can capture three to ten spectral bands that cover more than 20 nanometer (nm) each [64, p. 2]. By this definition, RGB images are also multispectral images of the region of the electromagnetic spectrum that is visible for humans. Common RGB cameras have sensors that measure the three wavebands that humans perceive as red (sensitivity peak around 600 - 625 nm), green (sensitivity peak around 520 - 540 nm) and blue (peak around 450 - 470 nm) (see fig. 3). This is possible by adding the Bayer filter array to the sensor [12] (”Bayer-Pattern”), which is an array of spectral filters for the colors red, green and blue. Hence, each pixel can only measure one of the three colors and the majority of the pixels senses ”green” [12, sheet 4] because human vision also relies heavily on green reflectance [12]. In order to obtain information about a broader range of the electromagnetic spectrum and a higher spectral resolution, hyperspectral cameras are used (see fig. 3). Hyperspectral cameras measure more than 20 and up to several hundred spectral bands that are comparably narrow with 10 nm [64, p. 2]. The main advantage of spectral imaging is that more information about the photographed material is available [104, p. 2] (see fig. 3). The data structure of spectral images can be imagined as a 3D-matrix, with the first two dimensions representing the spatial dimensions and the third the spectral dimension, as depicted in fig. 2 [72, p. 9]. In the following, multispectral imaging with more than three channels and hyperspectral imaging will be referred to as ”spectral imaging” as opposed to RGB images.

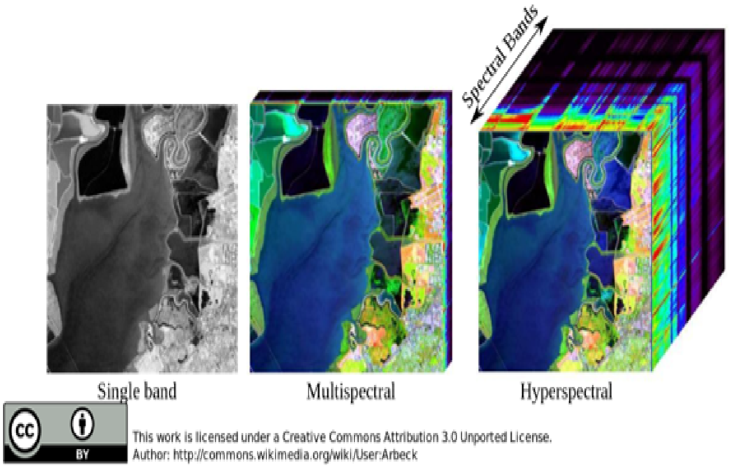


Figure 2: Difference between structure of image arrays or cubes depending on the sensor type, source: see graphic.

Beer’s law (see eq. 1) describes the relationship between the absorbance A at a certain

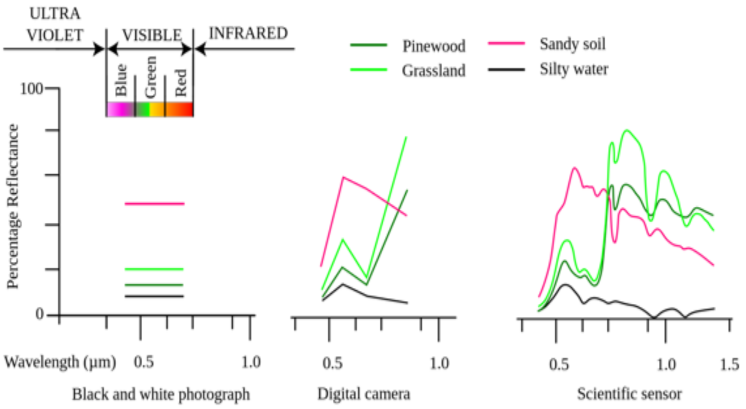


Figure 3: Difference between spectra of soil and plant reflectance values based on sensor, source: see graphic.

wavelength λ and the molar extinction coefficient e, the path length of the light l and the concentration c of the absorbing chemical compound [102][p. 15].

$$A(\lambda) = \eta(\lambda) \, * \, l \, * \, c \tag{1}$$

Even if not all requirements of Beer's law are met, absorbance values still give a reasonable estimate of the concentration [102][p. 16], supporting the approximation of the chemical composition of materials with hyperspectral imaging [41, p. 21].

2.1.2.1 Measuring plant properties with spectral imaging

Especially spectral images of the near-infrared region, ranging from 700 to 1100 nm, and the shortwave-infrared region, ranging from 1100 - 2500 nm, can be used to distinguish between materials, and therefore, also plants since different plant species have characteristic chemical compositions and surface structure [14, p. 100] [56] [34, p. 669] [113, p. 95] [54, p. 3]. Plants in general have a very distinct spectral footprint (see fig. 3) with a reflectance peak at green wavebands, low reflectance for blue and red light, a sharp rise of reflectance at the red edge region (around 700 nm) and then continuous high reflectance for the near-infrared region (750 - 1300 nm) [72, p. 145]. These characteristics enable an easy differentiation between plants and non-plants, for example, with the normalized difference vegetation index (NDVI), which exploits the huge difference of reflectance between red and near-infrared wavebands (NIR) of plant's spectra (see eq. 2) [72, p. 147].

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{2}$$

The chemical composition and ergo the spectral response of plants depends on many factors such as nutrition status, water content [29], thickness of the material and mass [80, p. 213] [45] [103], surface parameters (leaf hairs, wax layer) [80, p. 213], age [21] [63][50, p. 59] [103] [40] [4], leaf inclination and shadowing [14, p. 99], humidity [54, p. 3] and infestation with diseases. Many studies concerning plant health have used hyperspectral imaging for detecting a change in plant health [54, p. 3], e.g. the infection with a certain funghi before the symptoms could be perceived by humans [118]. Another challenge is that related plant species, such as sugar beet and lambsquarters, have similar spectra [9, p. 2] [104, p. 6] [113, p. 95]. Furthermore, each pixel contains information from the neighbouring pixel, especially since some materials are transparent or semi-transparent for certain wavelengths [49, p. 650]. Due to all these influence factors, even the spectra of one plant can vary a lot [117, p. 67].

2.1.2.2 Hardware for spectral imaging

The properties of data depend on the measurement system that was used for the data acquisition. Spectral data is collected with imaging spectrometers that obtain information about the space (spatial), spectrum (spectral) and strength of reflectance (radiometric)[72, p. 8]. Spectral imaging spectrometers are often distinguished based on how they obtain spatial and spectral data [70, p. 010901–3]. There are generally three options [70, p. 010901–3]:

- Spatial scanning
- Spectral scanning
- non-scanning

With spatial scanning methods, all wavebands are captured at once for one spatial unit, which refers to one pixel for point-scanning cameras (whisk-broom) and one line for line-scanning cameras (push-broom) [70, p. 010901–3] [72, p. 10]. For spatial scanning instruments, the target and camera have to move relative to each other in a stable way and with the same velocity. Spectral scanning devices scan the whole field of view for one waveband within one exposure time by using the corresponding bandpass filter on a wheel that contains all bandpass filters [70, p. 010901–3] [72, p. 11]. Spectral scanning methods are also called staring-imaging and they require static scenes and relatively long exposure times [70, p. 010901–3]. There are several approaches for non-scanning methods, but they all have in common that the spectral and spatial information is obtained during one exposure time [46]. The snapshot methods enables faster and easier image acquisition. Further, all non-scanning devices share the drawback of decreased spatial or spectral resolution or quality [46, p. 090901-19]. One approach for the technical implementation for a non-scanning spectral camera is to extend and improve the idea of the Bayer filter to use a spectral filter mosaic [46, p. 090901-11].

2.1.2.3 Aperture and aberration

Since different apertures were tested for this thesis, the meaning and effect of different apertures is explained in the following and illustrated in fig. 4. The term effectual aperture refers to the diameter of the circle through which light can enter the camera lens [85]. The term aperture is defined as $\frac{focal\ length}{f-number}$ and the f-number is defined as $\frac{focal\ length}{effectual\ aperture}$ [85]. In this thesis, "f/f-number" refers to the aperture. A small f-number results in a large aperture (see fig. 4). A high f-number means that most of the outer part of the lens is covered by the diaphragm, causing a small aperture, meaning less light enters the camera (see fig. 4). The spatial and spectral quality of images is often not ideal due to spherical and chromatic aberration of camera lenses. Spherical aberration describes spatial confusions in the image, caused by the lens refracting rays that fall on the outer part of the lens stronger than rays that enter the lens at the center [25, pp. 78 - 79] (see

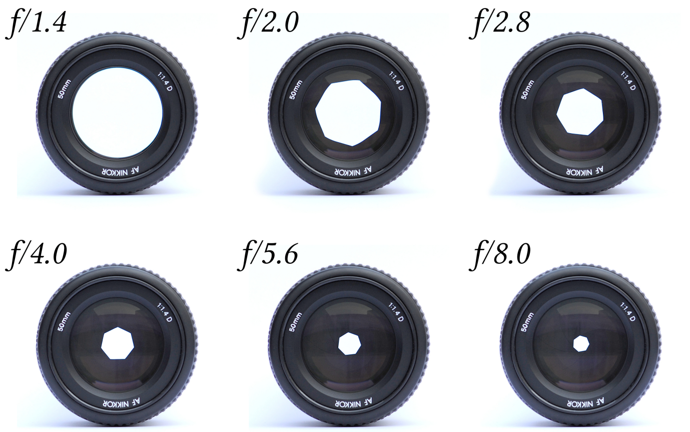


Figure 4: Decreasing apertures and increasing f-number which is the denominator in the term f/f-number illustrated with camera lense, credits: Wikipedia user KoeppiK, source: [59].

fig. 5(a)). A counteraction could be to increase the f-number, which means shadowing the outer part of the lens, but this leads to a reduction in resolution and brightness, if the exposure time is not increased accordingly [62, p. 8]. Chromatic aberration describes the phenomenon that the strength of the refraction by the same lens is dependent on the wavelength, resulting in slightly different focal lengths [25, p. 75] (see fig. 5(b)). The manual of the hyperspectral camera indicated that higher f-numbers were preferable for better spectral quality and pointed out that with small f-numbers, the spectral response was shifted towards lower wavelengths [89, p. 3].

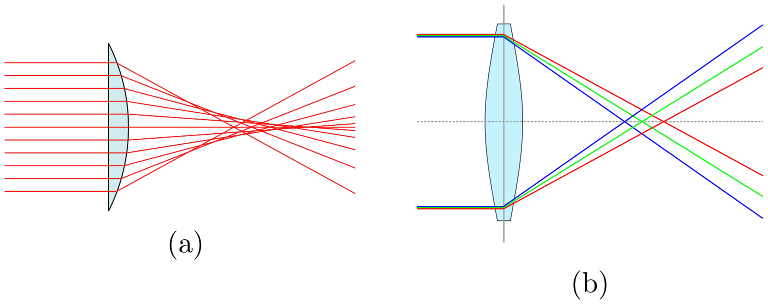


Figure 5: Aberrations of convex lenses: (a) Spherical aberration, Andrei Stroe, source: [6] , (b) Chromatic aberration, credits: Wikipedia user Andres 06, source: [5].

2.1.2.4 Calibration and pre-processing of spectral data

The raw image data straight out of the camera is influenced by many factors such as [77, p. 53]:

- Lighting conditions that vary within and between images
- Temperature of the camera (dark current or thermal signal)
- Lens properties
- Aperture

Furthermore, the raw images straight out of the spectral camera have no physical unit [72, p. 17]. The transformation of the raw data *raw* into reflectance R and the correction for different illumination conditions and the dark current is achieved by the calibration formula shown in eq. 3 [102, p. 24]. The white reference w mentioned in eq. 3 has known reflective properties [72, p. 19] and should reflect close to 100 % of the incoming rays [102, p. 24]. The dark reference d is obtained by blocking the lens from light and thereby only measuring the thermal signal [72, p. 17].

$$R = \frac{raw - d}{w - d} \tag{3}$$

The absorbance A can be approximated from the reflectance R with eq. 4, which is not entirely correct but has shown high functional correlations with the concentration of chemical compounds based on Beer's law [102, p. 26].

$$A = \log_{10}(R) \tag{4}$$

Particularly for close-range spectral imaging of uneven objects like plants, the difference in distance to and angles towards the imaging system, differences in the roughness of surfaces and the distinct geometries cause light scatter and influence the spectral data greatly [77, p. 55] [78, p. 121]. Therefore, pre-processing techniques like Standard Normal Variate (SNV) and (Extended) Multiplicative Scatter Correction (E)MSC that mitigate scatter and other disturbing effects that do not account for differences in (bio) chemical composition are necessary [77, p. 55] [78, p. 121]. SNV correction is widely used and delivers good results [78, p. 124], [77, p. 55]. The equation 5 describes how SNV transforms a pixel p with the spatial coordinates x and y and spectral channel c. The advantage of SNV is that no reference spectrum is required for the correction [77, p. 55].

$$p(x, y, c)_{SNV} = \frac{p(x, y, c)\ -\ mean(c)}{standard\ \ deviation(c)} \tag{5}$$

The formula, and often also the result of Multiplicative Scatter Correction (MSC), is similar to SNV with the difference that the spectra is corrected based on a reference spectrum without scattering [77, p. 55] [102, p. 40]. The main idea is to estimate the additive (intercept, β_0) and multiplicative deviations (slope, β_1), which are assumed to be caused by, e.g., light scatter, of the spectrum relative to the reference spectrum [102, pp. 40 - 43] (see fig. 6). Additive effects are represented by the intercept of the spectrum 1 and 2 in fig. 6 and refer to the reflectance values of a certain spectrum being higher or lower by a constant due to a disturbing effect [102, p. 42]. Multiplicative effect means that the reflectance of all wavebands is influenced by a factor, represented by the slope of spectra 1 and 2 in fig. 6 [102, p. 42]. Instead of an acquired reference spectrum, the mean spectrum can be used [102, p. 40].

$$p(x, y, c)_{MSC} = \frac{p(x, y, c) - \beta_0}{\beta_1} \tag{6}$$

EMSC includes additional, higher polynomial degrees in the nominator [102, p. 43]. $Degree(EMSC) = 0$ will refer to the correction with only β_0 and β_1, $degree(EMSC) = 1$ to further subtracting the term $\beta_2\mathrm{v}$ of polynomial degree 1 in eq. 7, and so on. v in eq. 7 is a vector that either contains artificial features or spectra of disturbing chemical components, such as water [102, p. 44].

$$p(x, y, c)_{EMSC} = \frac{p(x, y, c) - \beta_0 - \beta_2\mathrm{v} - \beta_3\mathrm{v}^2}{\beta_1} \tag{7}$$

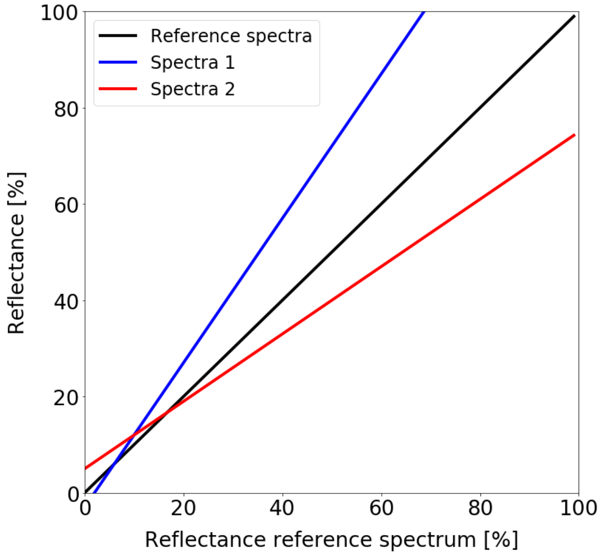


Figure 6: Sketch of the reflectance in % of two made-up spectra relative to a reference spectrum, graph based on Stefansson, 2019, p. 43 [102].

2.1.2.5 Assessing spectral quality

So far, frameworks for the objective assessment of spectral quality have not received much attention, even though it is decisive for the tasks hyperspectral imaging is used for [98, p. 23]. One method is to compare a hyperspectral frame to a reference spectrum [98, p. 24]. Task-based quality is another approach, meaning the evaluation of the spectral quality based on the performance at the task the spectral data was acquired for [98, p. 30], for instance, the classification of background, sugar beet and weed pixels.

2.1.3 Classification methods for spectral data

The following section briefly covers the analysis methods for classification used in this thesis or by research teams in the related work section.

2.1.3.1 Statistical methods

A k-nearest neighbours (KNN) classifier remembers all training samples and their corresponding class [91, p. 102]. An integer for k and a distance metric has to be chosen [91, p. 102]. A new sample is classified by examining the k training samples that are closest to it in the n-dimensional feature-space, based on the chosen distance metric [91, p. 102]. The new sample is assigned to the class that most of its k closest neighbours belong to [91, p. 102].

A Support Vector Machine (SVM) is a classification method that maximizes the distance between hyperplanes that enclose the different classes [91, pp. 76 -77]. The hyperplanes act as separating boundaries between classes and depend on the support vectors, samples that are close or contained in the hyperplane [91, p. 76].

Principal component analysis (PCA) is a useful technique for feature extraction and the exploration of informative subspaces within the data [91, p. 142]. PCA relies on the idea that the directions in which the explanatory variables show the most variation are the ones that contain the most essential information [91, p. 142]. Mathematically, PCA is based on singular value decomposition and the eigenvectors and eigenvalues of the covariance matrix of the dataset [91, p. 144]. PCA transformed data is often used as input for regression and classification models because it has fewer dimensions than the original data while maintaining the most important information.

Partial Least Squares (PLS) is similar to regression analysis of PCA transformed data in many ways, but the most significant difference is that not the subspaces of highest variation are taken into account but the hyperplanes that separate the given classes best [65, p. 25]. It is an iterative and supervised approach, since the y-data is also part of the input, for finding interesting subspaces for the explanatory and the response variables [65, p. 23].

"The goal of linear discriminant analysis is to find the feature subspace that optimizes class separability" [91, p. 155]. It is similar to PCA in the way that a new "coordinate system" is created out of a linear combination of the original features [91, p. 155]. However, not by computing eigenvectors and eigenvalues of the covariance matrix, which results in sorting the results based on the direction with the largest variance, as for PCA. Instead, the within and between-class-scatter matrices of the classes are the basis for discriminant analysis [91, pp. 158 - 159]. Thereby, the biggest eigenvector of the eigendecomposition points in the direction where the classes are easiest to separate [91, p. 156].

Bayesian classifiers rely on the assumption that data and its distribution from the past can help to determine the class of a new sample with probability calculations [24, pp.774]. For example, if previous studies had shown that men were generally taller than women and the height distribution of both genders were known, a Bayesian classifier would classify an unknown, relatively tall person of 1.96 m as male.

2.1.3.2 Deep learning

For many classification or regression tasks, artificial neural networks (ANN's) are outperforming other algorithms [91, p.380] in robustness and versatility. This is especially true for complex tasks like speech or image recognition [91, p. 380]. Neural networks are often referred to with the expression "deep learning" where "deep" refers to a high number of layers and nodes, and, therefore, trainable coefficients ("weights") (see fig. 7). The huge number of trainable weights enables deep nets to solve complex problems [91, p. 73]. The main obstruction for ANN's is that with too few training samples, the network tends to learn the patterns of the training data too well and performs badly on other data [91, p. 73]. This is why it is crucial to have high quality and quantity training data for neural networks. For most types of ANN's, training data consists in the explanatory variables and the response variables, called "ground truth" in general and "labels" for classification problems. ANN's can also be used for classifying pixels of spectral images.

An ANN mimics some processes and structures of the brain, like ANN's neurons, called nodes, that are connected and receive, process and send signals [91, p. 384]. The individual nodes are linear functions of the input signals wrapped in a non-linear function ("activation function"), like a sigmoid or a rectified linear unit function (see fig. 7 and eq. 8) [91, p. 444]. By using non-linear functions as a wrapper, the ANN can capture more complicated patterns. The rectified linear unit (Relu) is defined in eq. 8 and is one of the most used and best-performing activation functions [91, pp. 449 - 450]. The coefficients and biases of the linear equation of each node in this complex net of functions are changed, based on how correct the prediction of the network compared to the ground truth data was [91, p. 387]. This is determined by the so-called cost function. The cost function and its gradient determine how much and in which direction the weights of the network have to change since the goal is to find the global minimum of the cost function [91, pp. 35 - 36]. The process of following the negative gradient in order to get to the global minimum of the cost function is called gradient descent [91, pp. 35 - 36]. The optimizer and the learning rate are responsible for how and how fast the weights are updated [91, p. 429]. Nadam optimizer (Nesterov-accelerated Adaptive Moment Estimation) is based on gradient descent, but incorporates two improvements resulting in higher speed and quality: Adaptive moment estimation means that the learning is accelerated when the negative gradient is very steep, and the other way round, and the learning rate is adapted for each parameter [27, pp. 1 - 2]. Nesterov acceleration can be imagined as looking one gradient step ahead and thereby determining the best direction [27, p. 3].

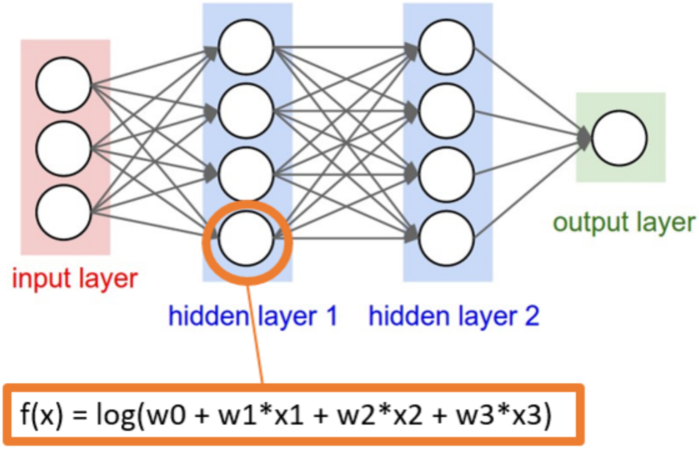


Figure 7: Sketch of a simple, fully connected neural network with an example activation function for one node, based on a sketch of Adrian Rosebrock [92], edited.

$$relu(x) = max(0, x) \tag{8}$$

Overfitting is a considerable problem with deep ANN's for hyperspectral images because labeled training data is scarce [22, p. 6233], [39, p. 2]. There exist several methods to reduce overfitting for deep networks, such as L2 regularization [91, p. 408] of the weights, dropout [91, p. 512] or batch normalization [53, p. 5]. A batch refers to a subset of the training data based on which the weights are updated. Normalization, in this case, refers to the linear conversion of each feature by subtracting the mean and dividing by variance of the feature (zero mean, variance of 1) [53, p. 3]. Batch normalization also has the advantage that the network converges faster because the distribution and scale of the data remain relatively stable over different batches[53, p. 1]. The regularization effect of batch normalization was not the primary intention but occurs since not the absolute values of each sample are used but scaled versions that depend on the other samples of the batch whose composition changes during training [53, p. 5].

One special type of ANN's that is mostly used for image analysis are convolutional neural networks (CNN's) [91, p. 494]. Their main advantage is that they combine the analysis of the reflectance values with the spatial information and can maintain the input image's spatial information. They use filtering as a "traditional" image analysis method to gain information in a spatial context [91, p. 495]. An image filter is a sliding window with a specific pixel size that summarizes the values in its field of perception in a specific way and outputs a filtered image with new pixel values [91, pp. 496 - 498]. One example is a 3 x 3 mean filter, where the new pixel value at position (x,y) in the output image is the mean of the nine reflectance values inside the sliding window with the center pixel at (x,y) of the input image. In several layers, the image data or its descendants is convolved with one or several filters whose weights can be trained based on the cost function [91, p. 494]. The result of a convolution is called feature map because a filter extracts specific patterns of the image that are sometimes incomprehensible to humans and other times apparent properties, like horizontal edges [61, 112]. CNN's are often built hierarchically, with the first layers extracting low-level features of small perceptive fields which serves as input for the layers that extract higher-level features or classify based on feature maps [91, pp. 494 - 495].

CNN's are designed for and most used for image analysis in the spatial domain, which could also be applied to hyperspectral images. But in the case of many spectral channels, convolution in the spectral domain can be beneficial as shown by Hu et al. [51] who worked with spectral data that captured between 103 and 224 spectral bands. The spectral CNN with only one convolutional layer was tested against a Radial Basis Function SVM (RBF-SVM) and three different types of "normal" ANN's, meaning only fully connected layers of different depths and structures [51, p. 7]. Three remote sensing datasets were evaluated, and for each class, 200 random pixels were chosen for the training dataset, which represented between 4 - 21 % of the total amount of pixels [51, p. 6]. It is interesting to note that the improved SVM version outperformed or performed similarly well as the two shallowest ANN's [51, p. 7]. The proposed CNN, even though it was quite shallow and needed less time for calculating predictions than two of three other ANN's, achieved a 1 - 2.5 % higher accuracy than the RBF-SVM and a 2 - 3.6 % higher accuracy compared to the other ANN's [51, p. 7]. Luo et al. [71] developed this approach further by taking the eight neighboring pixels of the one center pixel into account and performing convolutions over the spectral and spatial domain [71, p. 3]. This method is based on the knowledge that adjacent pixels have very similar spectral properties and are therefore highly correlated [71, p. 2]. Hu et al. used the same data sets [71, p. 4]), and their model reached accuracies around 99 %. A similar approach has been applied and tested by Gao et al. .[39], Chen et al. [22] and Santara et al. [94], among others.

2.1.3.3 Performance measures for binary classifiers

The performance of a binary classifier in the presence of class imbalances is often measured with recall, which is also called true positive rate (eq. 9), precision (eq. 10), false-positive rate (eq. 11) and dice coefficient (eq. 12). Class imbalance refers to a (big) difference between the number of samples for the different classes. This can decrease the meaningfulness of metrics like accuracy because if, for instance, 99 % of all samples belong to class A, an accuracy of 97 % is a poor performance if only class A samples were classified correctly. All those parameters use the number of true positives (TP, "positive" sample correctly predicted as "positive"), false positives (FP, a "negative" sample was predicted as "positive"), true negatives (TN, "negative" sample correctly predicted as "negative") and false negatives (FN, "positive" sample was predicted as "negative") [91, p. 206].

$$recall = \frac{TP}{\text{FN + TP}} \tag{9}$$

$$precision = \frac{TP}{\text{TP + FP}} \tag{10}$$

$$\text{false positive rate} = \frac{FP}{FP + TN} \tag{11}$$

$$dice = \frac{\text{precision x recall}}{\text{precision + recall}} \tag{12}$$

(Source of equations 10, 9, 12: [91, p. 208])

2.1.4 Image registration and alignment

Image registration or image alignment is "the computation of 2D and 3D transformations that map features in one image to another" [105, p. 311]. This is used, among others, to stabilize video frames or for creating a panorama image out of many single images [105, p. 528]. There are two types of transformations: Global motion models compute one transformation matrix for all pixels in one image, and local motion models determine different transformations for each pixel, which can be represented as a vector displacement field [105, p. 170]. The imregdemon algorithm from Matlab iteratively estimates a displacement field to align two images based on Thirion's approach to consider image alignment a diffusion process [111]. The main idea is that certain pixels of image A are control-points, called demons, that can determine (e.g., based on a gradient) whether or not they are "inside" or "outside" their target area in image B [111, pp. 246 - 247].
One example of a global motion model is the 2D transformation called projection that preserves straight lines but not angles or size (see fig. 8), and the transformation matrix is known as homography [106, p. 4]. This type of transformation can be used for planar objects [106, pp. 7 - 8] and when the camera only rotated around its axis without any other movement because then all points can be assumed to be on the same plane in infinity [106, pp. 8 - 9]. In general, there are two main approaches to compute a global motion model: A reflectance-based and a feature-based approach [106, pp. 1 - 2]. A reflectance, or pixel-based, approach tests how similar the pixel values are and is therefore susceptible when different color channels are chosen [106, p. 15]. A feature-based global motion model is computed based on a set of matching keypoints, in most cases, very distinctive

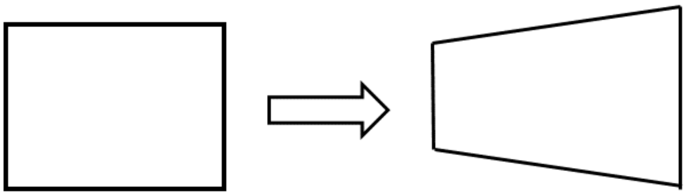


Figure 8: Sketch of projection, one type of 2D planar transformations, content based on Szeliski 2011, p. 311 [105].

points like corners, in both images [105, p. 207]. For the feature-based approach, the first step consists in finding and describing characteristic points like corners in each image that are in the best case, invariant to rotation, scale and not too sensitive towards changing lighting conditions [106, p. 33].

2.1.4.1 Feature detection with SIFT

The feature detection algorithm SIFT was developed in 1999 by David Lowe [68]. SIFT stands for **S**cale **I**nvariant **F**eature **T**ransform and its features are invariant to scale of interest points, rotation and translation and also stable towards some changes in illumination and changes in object's appearance due to different perspectives, called parallax [68, p. 1]. Parallax is the perceived displacement of an object relative to the background or another object when viewed from two different viewpoints. Parallax effects are especially strong for close-range images but can be removed with different approaches, most of them require a 3D camera calibration [105, pp. 445 - 446].

SIFT and its descendants have proven to outperform other descriptors [75, p. 1615] [106, p. 37]. SIFT achieves this by looking for key points in several versions of the image [106, p. 37]: First, the image gets scaled up by the factor of 2 and then gets halved in size several times ("different octaves") [122]. For each scale, the Gaussian blur filter is applied with different strengths [122]. Then the Difference of Gaussians is computed [122]. Minima and Maxima of adjacent pixels in the image at hand are taken into account but also the same pixels of adjacent images in the same octave and the corresponding pixels in the scale level above and below [122]. The key points are stored relative to the gradient which makes SIFT features rotation-invariant [106, p. 37].

Nevertheless, when SIFT is confronted with greater changes in reflectance, as it is the case when comparing images acquired at different wavebands, the number of wrong matches increases [125, p. 1]. Some sort of filtering has to be applied either on the matches or the suggested homographies [125, p. 1]. One way to filter out mismatched keypoints is to use a global distance (similarity) threshold between the keypoints in the two images [69, p. 104]. Another, more effective filtering option is thresholding Lowe's distance ratio [69, p. 104]. For each keypoint, the matching algorithm normally finds several possible matches in the other image [69, p. 104]. The similarity, or distance, of each matching pair is computed based on the description of the keypoints. Due to the robustness of SIFT towards, e.g., slight parallax and differences in lighting, the second-best match of a correct match is most likely very close to the best match in the spatial dimension but also regarding the distance metric [69, p. 104]. The opposite is the case for wrong matches, then the second-best match is most likely wrong, too, and the dissimilarity between keypoints is probably even bigger than for the best match [69, p. 104]. Lowe's ratio is described in eq. 13 with s_1 as the shortest (best) distance for a matching keypoint pair and s_2 as the second-shortest distance (see eq. 13). This can be used for filtering the matches with a threshold t for Lowe's distance ratio as shown in eq. 14. Lowe's distance ratio has values between 0 and 1 and the higher it is, the more likely it is to be a correct match. Therefore a good ratio threshold t is 0.8 [69, p. 104]. Lowe stated that 90 % of the false matches could be removed by a collateral removal of 5 % of correct matches.

$$Lowe's\ distance\ ratio = \frac{s_1}{s_2} \tag{13}$$

$$t\ *\ s_2 >= s_1 \tag{14}$$

2.1.4.2 Computation of the homography

The homography projects the pixels of image plane A onto image plane B, as shown in eq. 15 [106, p. 4]. It is obtained from the filtered, matching keypoints by solving a linear system of equations. [106, pp. 41-42]. At least four matching key points are necessary in order to compute a homography.

$$\begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = H * \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} * \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} \tag{15}$$

Since some of the matching keypoints are probably wrong, the **RA**ndom **SA**mple **C**onsensus algorithm (RANSAC) is often used to compute the homography matrix [106, p. 39]. This iterative algorithm expects outliers and needs more data than minimal necessary for the solution of the equation [106, p. 39]. The algorithm has four main steps [106, p. 39], [74, pp. 3 - 4]:
1. Choose a random subset of data points of the minimum size necessary to solve the equation.
2. Compute the solution
3. Based on a distance/tolerance threshold: Include all other data points that support the model
4. Repeat 1 - 3 until nr of iterations/other criteria is reached
5. Choose the model that had most inliers.

The more iterations RANSAC runs for, the more likely it becomes that the best solution is found [106, p. 39], provided that the distance metric is reasonable.

2.1.4.3 Quality assessment of image registration

An automated and objective method for evaluating the quality of image alignment is a challenging task [79, p. 240]. There are two main approaches: The comparison of the differences of the pixel values in the overlapping area (pixel level) and the comparison of the overlapping area on a structural level by examining edges and geometrical features [79, p. 236]. Many quality assessment systems combine several parameters, on a pixel level and a structural level, to achieve good performance for a variety of images [79, p. 237] [57, p. 1]. An example of a pixel-level approach is to compute the mean squared error (MSE) between the two aligned images i_1 and i_2 with pixel coordinates x and y (see eq. 16). The lower the MSE, the more similar are the pixel values, but this method is not reliable in all cases [79, p. 234].

$$MSE = \frac{1}{n} * \sum_{n=0}^{n} (i_1(x,y) - i_2(x,y)^2) \tag{16}$$

Correlation or rather cross-correlation is often used for measuring the error of image alignment, too [106, p. 18]. The Structural Similarity Index (SSIM) is computed for windows of the image and composited of three parts: Comparison of luminance, contrast and structure [119, p. 604]. Luminance is compared using mean reflectance (μ in eq. 17), contrasts based on variance (σ^2) and structure using correlation (σ_{xy}) (see eq. 17) [119, pp. 604 - 605]. SSIM is a popular metric for determining the similarity between two images [57, p. 3][79, p. 236].

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{17}$$

Apart from the registered images, the superimposed images can be examined, which means that the transformed image is blended over the corresponding region of the other image [57, p. 1]. When the images' alignment has worked properly, the superimposed images do not have many shadows, often called ghosting. Ghosted superimposed images appear blurry to the viewer. Blurriness of an image means that edges are not sharply pronounced. Applying an edge operator on a blurry image would, therefore, deliver fewer edges and less variance of the edge image. One edge operator is the Laplacian operator which uses the second partial derivatives for finding edges [20, pp. 139 - 140]. An edge is indicated by a sharp change in the intensity values, which is a local minimum or maximum of the first derivative [20, pp. 139 - 140]. This leads to values equal to 0 for the second derivative at the planes surrounding the edge and at the peak/valley of the first derivative [20, pp. 139 - 140]. The Laplacian kernel is defined as[20, p. 140]:

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

The variance of the Laplacian of an image is therefore high when there are many edges and low for few pronounced edges, which indicates blur [86, pp. 315 - 316].

2.2 Related work

There have been many attempts to develop robust computer vision systems for crop and weed detection under field conditions [87, pp. 193 - 194] [116, p. 2] [120, p. 1] [109, p. 521] [117, p. 63] [10, 96]. Since this thesis is very practice-oriented and there are only a few papers about automated labeling of crops and weed with spectral imaging, several other weed versus crop classification approaches have been studied in order to gain a broader understanding of techniques and their benefits and disadvantages. In order to cover other features that can be used in addition to spectral features, the next subsection also covers crop and weed detection approaches that do not (solely) rely on spectral data.

2.2.1 Crop and weed detection without spectral imaging

Thinking of how humans identify plant species mainly based on color, shape and surface characteristics, it is a valid hypothesis that computers could use the same spectral region and the same features for the classification of plant species [55, p. 602].

Astrand proposed a system based on color and shape features of RGB and grayscale images for the segmentation of sugar beet and weed, feeding those features into five KNN-classifiers, which reached up to 96 % of correct classifications [8]. The color features were the main contributors to the classification success [8].

Tellaeche et al. worked with corn, and used area and structure parameters of 340 RGB images and a Bayesian classifier in order to determine whether or not a certain grid cell of an image contained weed [109]. Further, images were binarized into plant and background and rows were detected [108, p. 523]. Spectral features were only utilized for plant detection and afterward ignored since the spectra of weed and wheat were too similar [108, p. 523]. The highest accuracy was 92 % [108, p.529]. The system was not completely robust towards changes in illumination [108, p.529].

Mao et al. exploited the information of hue and saturation values of 500 RGB images and classified 90% of the wheat plants and 88 % of the weeds correctly with discriminant and regression analysis [73]. Hue denotes the dominant wavelength and is therefore associated with the color [73, p. 961]. Saturation refers to the strength of the color in question [73, p. 961]. The use of H-S-variables mitigated the variation of illumination [73, p. 959].

Dos Santos Ferreira et al. reached 99% accuracy for distinguishing between broadleaf weeds, grass weeds, soybean and soil in 15.000 RGB images of soybean fields captured by a drone using shape, color and texture features and a CNN [26].

Arakeri et al. deployed an artificial neural net and trained it on color features of grid fields of RGB images with 5305 median filtered images of an onion field and obtained an accuracy of 99 % [7, pp. 1203-1204]. The authors and team built an artificial onion field where different weed and onion settings, illumination and wind conditions were taken into account [7, p. 1203].

Nejati et al. exploited the density of leave edges and leaf vein structure in RGB images, enhanced by the Fast Fourier Transform, to differentiate between corn and weed in real-time. Fourier transformed data is in the frequency space, applied on an image. This means that high frequencies represent areas with lots of variation in intensity (for instance, at edges) and low frequencies mean little variation of intensity values (all smooth parts) [1]. This method reached an accuracy of 92 % on the 80 images that served as test set [82, p. 1].

2.2.2 Crop and weed detection with spectral imaging

Lottes et al. [66] distinguished between sugar beet and weed, based on four-channel-images (RGB + NIR) [66, p. 2] by separating plants from the background with NDVI, computing several statistical features of reflectance, gradients, texture and spatial based on multiple plant pixels and channels for the plants and then using a random forest for the classification [66, p. 1]. Pixel classifications were smoothed based on the classifications of the adjacent pixels [66, p. 1]. For sugar beet (positives), on a plant level, the model reached a true positive rate (TP/recall) between 90 % - 96 % and precision of 95 % [66, pp. 6 - 7]. For weed (negatives), True Negative Rates of 90 %, precision between 82 % [66, pp. 6 - 7]. The model was trained and evaluated on images with sugar beet in the 4-leaf stage [66, p. 6]. When this model was applied to a dataset where sugar beet was in a late 2-leaf/early 4-leaf stage, the performance of the classifier declined as expected, resulting in 85 % of true positive rate for sugar beet and True negative rate of 79 % [66, p. 7].

Milioto et al. utilized four-channel RGB-NIR color images of sugar beet fields, extracted the vegetation areas with the NDVI thresholding and regions of single or overlapping plants with blob analysis to mitigate the effect of size on the CNN [76, pp.3 - 4]. The blobs were classified with a CNN, resulting in a 97 % precision and a recall of 98 % [76,

p. 4].

Gerhards et al. detected and identified weeds and crops in fields of rape, maize or sugar beet, based on spectral data of the NIR and visible spectrum, and area, diameter and shape features that were compared to a database with plant species characteristic shape features [41, 42]. This resulted in a detection rate of 82 % [42, pp. 32-40].

Zhang et al. used hyperspectral data (400 - 795 nm) and three different Bayesian classifiers developed for three seasons for tomato plants and the associated weeds in real field conditions and also reached accuracies of 96% [127]. Only the spectra were used as features since shapes of plants change during the development and can be distorted by leaf damage (holes, ragged edges) or overlapping leaves [127, p. 66].

Vrindts et al. evaluated the differentiation between maize, sugar beet and seven weed species in the lab and in the field with stepwise discriminant analysis and spectral data between 400 and 2000 nm in the lab and 400 to 820 nm in the field [117]. The lab experiments yielded accuracies of 97 % and the field trials a success rate of 90 % of correctly classified plants [117, p. 63].

Feyaerts et al. [34] used spectral imaging to distinguish between sugar beet (4 - 6 leaf-stage) and weed under real field conditions. Six wavelengths of the visible and NIR spectra were used: 441, 446, 459, 883, 924 and 988 nm [34, pp. 669 - 670]. A fully connected neural network reached the best classification accuracy of the five tested methods with 80 % correct classifications for sugar beet and 91 % for weed [34, p. 679].

Okamoto et al. [84] acquired hyperspectral images with wavelengths between 400 - 1000 nm (resolution 10 nm) in the field to classify sugar beet and four common weed species in sugar beet fields [84, p. 32]. For training, 75 pixels from 3 plants per species were used for each species. Wavelet transformed pixel data was subjected to variable selection and classified with linear discriminant analysis [84, pp. 34 - 35]. 81.3 % of the sugar beet pixels were classified correctly and, depending on the species, between 74.7 % to 97.3% of the weeds [84, p. 36].

2.2.3 Automated labeling for crop and weed detection

Louargent et al. used the spatial information about sugar beet and maize fields to choose their training data for weeds only from plants between rows and the sugar beet/maize training data from the plants precisely on the sowing line [67, pp. 7 - 8]. Row orientation was detected with a Fast Fourier Transform, the vegetation was identified with NDVI thresholding and the row position was determined [67, p. 7]. Shape features such as area and orientation were applied to further distinguish weed from crop [67, p. 7]. Four channels, between 550 and 790 nm were utilized as features for a SVM classifier [67, p. 5]. This system was able to classify 81 % of the sugar beets and 92 % of the weeds correctly [67, p. 13].

Wendel et al. used a similar approach for vegetable fields, using wavebands between 391 and 887 nm (spectral resolution of 2 nm) [123, p. 5130] and a crop line detection algorithm, NDVI to separate vegetation from the soil, PCA for feature extraction and a Linear Discriminant Analysis for the classification [123, p. 5130]. They also used a manually labeled training set as reference. The automatically labeled training set yielded a dice score of 0.85 compared to the manually annotated one with a dice score of 0.95 [123, p. 5135].

Gao et al. applied a partially unsupervised maize weed classification on RGB images, utilizing intra-row plants as the weed training set, with a random forest for feature selection, and achieved an overall accuracy of 95 % [38]. Pixel- and geometry-based features were utilized [38]. The data was only captured from one field and might be unreliable for other plants or fields. [9].

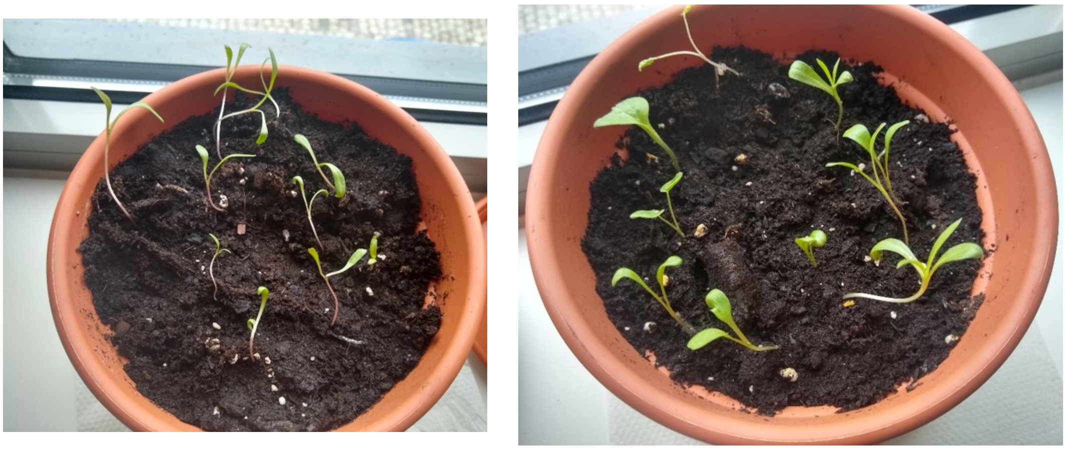
3 Material and Methods

The experiments described in the following were the basis for the camera setup and image analysis pipeline for the data acquisition campaigns in the field from April until June 2020. If not indicated otherwise, all data were analyzed with Python 3.7.4 and mentioned Python libraries.

3.1 Plants

Plant seeds of sugar beet and 18 common weed species in sugar beet fields were cultivated from seed. Because of the lacking offer of weed seeds and weeds on the fields, not all critical weed species of sugar beet fields could be obtained. Following weed species were included: Nettle, dandelion, sorrel (Polygonaceae), ribwort, buttercup, chickweed, daisies, slender meadow foxtail, wild carrot, tansy, wild rocket, mugwort, borage, milk thistle, safflower, globe thistle, sea holly plants (eryngium), lambsquarters (Chenopodium album, Chenopodium genus), field bindweed and knotweed (Polygoneae family).

Sugar beet was sown the 16th of January 2020 and the 30th of January in order to gather data from different age groups. All weed species except for lambsquarters, field bindweed and knotweed, were sown the 23rd of January 2020. Lambsquarters, field bindweed and knotweed (Polygonaceae) were sown the 31st of January due to later delivery of the seeds. Lambsquarters has an average germination time of 30 days, resulting in a 2-leaf stage being the highest development stage photographed. The germination of field bindweed and knotweed did not succeed because they need specific cold stimuli for germination, which could not be provided appropriately. The plants were cultivated in room temperature (18 - 20 °C) under natural lighting conditions and irrigated twice a day. There were 54 sugar beet plants, 88 weed plants, 28 of them lambsquarters. The growing medium was potting soil for gardening purposes, already fertilized. The soil also contained perlites, a white volcanic mineral that improves the airing of the soil. From February on, black gnats (Sciaridae) started to infest the potting soil of all pots. The insects were most likely imported within the potting soil. The larvas live in the soil and partially use plant roots as nutrition.



(a) Sugar beet.

(b) Weed.

Figure 9: Plants for experiments about classifying sugar beet and weed pixels with hyperspectral imaging in the near infrared region, 12.2.2020.

3.2 Cameras

The following two cameras were used for the image acquisition: The red-green-blue (RGB) camera, the UI-5240CP-C-HQ from IDS Imaging Development Systems GmbH with 1.31 megapixel. The lense for the RGB camera was the Schneider-Kreuznach Cinegon 1.4/12 CCTV-Lens with a focal length of 12 mm. The near-infrared region was filtered out for the RGB camera. Moreover, one snapshot hyperspectral camera, the MV1-D2048x1088-HS02-96-G2 from Photonfocus AG with 2.2 megapixels, resulting in 2048 x 1088 pixels in total and 409 x 216 pixels for each of the 25 spectral bands between 600 and 975 nm. The hyperspectral camera was equipped with the Edmund Optics (focal length 35 mm/F1.65 67716 VIS-NIR) lens. Photonfocus' MV1-D2048x1088-HS02-96-G2 uses the IMEC snapshot mosaic, meaning that several 5 x 5 mosaics of sensors for all 25 wavelengths are next to each other (see fig. 10), enabling fast image acquisition without the necessity to move the camera relative to the object. Additionally, two hardware bandpass filters from for the hyperspectral camera were tested as described in section '"Spectral data", subsection "Filter tests". One bandpass filter was transmissive for specific wavelengths between 600 and 875 nm, the other one for wavelengths between 675 and 975 nm, see tab. 1 for details. Photonfocus highly recommended the use of a bandpass filter because, without a filter, the signal for one waveband would contain a lot of unwanted noise from other wavebands due to the occurrence of cross-talk and second-order signals from other waveband [89, p. 5]. Tab. 1 describes which wavebands were measured with and without the hardware filters.

3.3 Data acquisition

All measurements were performed in the darkened lab at Fraunhofer IVI with two halogen lamps (Kaiser, series nr: 001691, 300 Watt, OSRAM 64516 bulbs) as the only source of light (see fig. 11). Homogeneous lighting with a slight angle towards the target, as recommended by Photonfocus' manual [89], was ensured (see fig. 11). The RGB and the hyperspectral camera had the same angle towards the target and the two camera cases were 0.5 cm apart in order to ensure that the RGB camera captured the whole field of view of the hyperspectral camera. Due to different lenses and focal lengths of the cameras, the covered region of the RGB camera was bigger compared to the hyperspectral camera. RGB and hyperspectral camera took images simultaneously. Therefore, the UTC timestamp was used in the name for RGB and spectral images in order to guarantee the correct matching of the image pairs afterward. The bottom of the camera lenses had a distance of around 30 - 35 cm to the target, depending on the height of the plants. For the RGB camera, 6.8 pixels covered 1 mm in real life. In the hyperspectral images, 1 mm of real life was represented by 3.5 pixels. The aperture of the RGB camera was f/16 and the automatic white correction was activated. The exposure time for the hyperspectral camera was active with a threshold of a maximum of 2 % of saturated pixels. Different apertures between f/2.8 and f/14 were tested for the hyperspectral camera. Because, on the one hand, a small aperture results in better spectral quality and sharpness, which is desirable for a proper classification (see p. 9). On the other hand, with low apertures, a longer exposure time is required if image brightness is supposed to be maintained which is not always feasible for high-throughput data acquisitions in the field (see p. 9).

Between the 18th and 20th of February, the different bandpass filter options (no filter, 600 - 875 nm, 675 - 975 nm) were tested. The hyperspectral images were acquired as

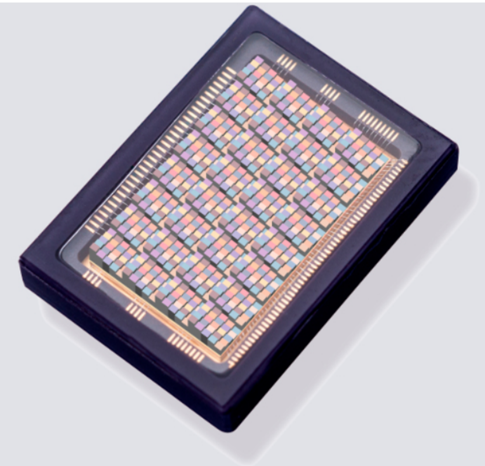


Figure 10: Snapshot mosaic CMV2K-SM5x5-NIR sensor (25 channels, 600 - 900 nm) of Photonfocus AG's camera MV1-D2048x1088-HS02-96-G2, source: [88].

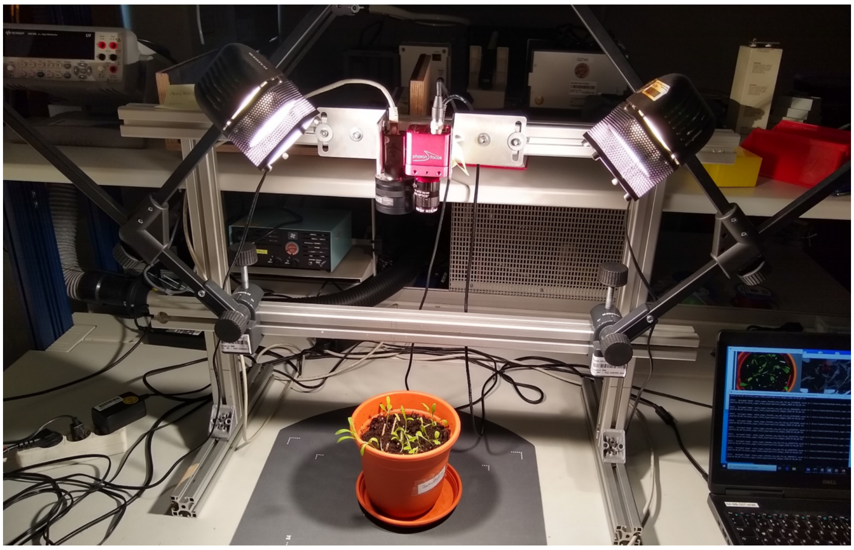


Figure 11: Measurement setup in the lab for the acquisition of optical and hyperspectral images of sugar beet and weed with.

Table 1: Sensed wavebands indicated by ”x” for different bandpass filter options for the hyperspectral camera MV1-D2048x1088-HS02-96-G2 (Photonfocus AG) [88], lens: Edmund Optics, 35 mm/F1.65 67716 VIS-NIR.

Wavebands [nm]	No bandpass filter	Filter 600 - 875 nm	Filter 675 - 975 nm
600	x		
607	x		
616	x		
629	x		
638	x		
646	x		
655	x	x	
663	x		
671	x		
680	x		
687	x	x	x
701	x	x	
727	x	x	x
740	x	x	x
753	x	x	x
767	x	x	x
779	x	x	x
792	x	x	x
804	x	x	x
816	x	x	x
835	x	x	x
846	x	x	x
857	x	x	
867	x	x	
877	x		
Number of channels	25	15	11

described in the manual provided by Photonfocus [89, pp. 7 - 8]:
1. The white reference and a dark reference were acquired with optimized exposure time for the white reference.
2. The target is shown to the camera and the exposure time is adjusted, respectively. Another dark reference at target exposure time is taken.

Between the 20.2.2020 and the 3.3.2020, sugar beet and weed were cultivated in separate pots since this made it easy to differentiate between plant species and hence simplified the creation of a training data set. As weeds had a faster youth development and, therefore, more highly reflecting leave areas early on, the exposure time for sugar beet and weed plants was adjusted individually, resulting in different exposure times for sugar beet and weed (see tab. 2 and fig. 12). The images captured with weed and sugar beet in separate pots and with separate exposure time, as just mentioned, will be referred to as "separate data set". Then, the majority of the plants was planted together in two big pots and photographed with the same exposure time in order to make the setting more realistic (see fig. 12). This image data set will be called "mixed data set". The images of the mixed pots were acquired on the 5.3.20, 6.3.20 and 9.3.20. The labeled data of the mixed data set consisted mainly of the plants of the separate pots, which hosted 2-leaf or early 4-leaf stages for sugar beet and a subset of all weed species (see fig. 12). To make the data set more representative, some sugar beet and weed pixels of the mixed pots were manually labeled. The mixed pots contained older sugar beet plants in the 4-leaf to 6-leaf stage. Nevertheless, not all weed species in the mixed pots could be labeled due to the high similarity in appearance with sugar beet. Moreover, small plant parts like shoots were labeled less frequently in the mixed pot since the annotation tool was rectangular which made labeling of small parts very tedious.
The separate data set made up the majority of the entire data set. The background was defined as everything that was not a plant and consisted in potting soil, including white perlite pellets, red plastic pots, parts of a hand and a grey table. Including test images for determining the best bandpass filter, over 900 image pairs, each consisting in one RGB image and one corresponding spectral image, were taken. For the classification and automated labeling, 863 image pairs were captured.

3.4 Analysis and classification of the spectral data

3.4.1 Pre-processing of the spectral data

The spatial demosaicing, the calibration and the de-noising based on the used bandpass filter of the hyperspectral raw data was performed with the software *HyperSpectral SDK* from Photonfocus AG, 2016. Spatial demosaicing is necessary due to the already explained mosaic structure of the sensor. demosaicing means, that the mosaic of 5 x 5 fields that each contained all wavebands was split up into one spatial image for each waveband by using bilinear interpolation [89, p. 6]. The calibration formula eq. 18 that takes into account the different exposure times for white target ($t0$) and actual target ($t1$) is defined by [89, p. 6]:

$$\text{calibrated reflectance} = \frac{frame_{t1} - dark_{t1}}{white_{t0} - dark_{t0}} * \frac{t1}{t0} \tag{18}$$

Saturated pixels were defined as reflectance values higher than 98 % of the possible maximum 2^{15} and replaced with the median of the eight surrounding pixels to maintain the

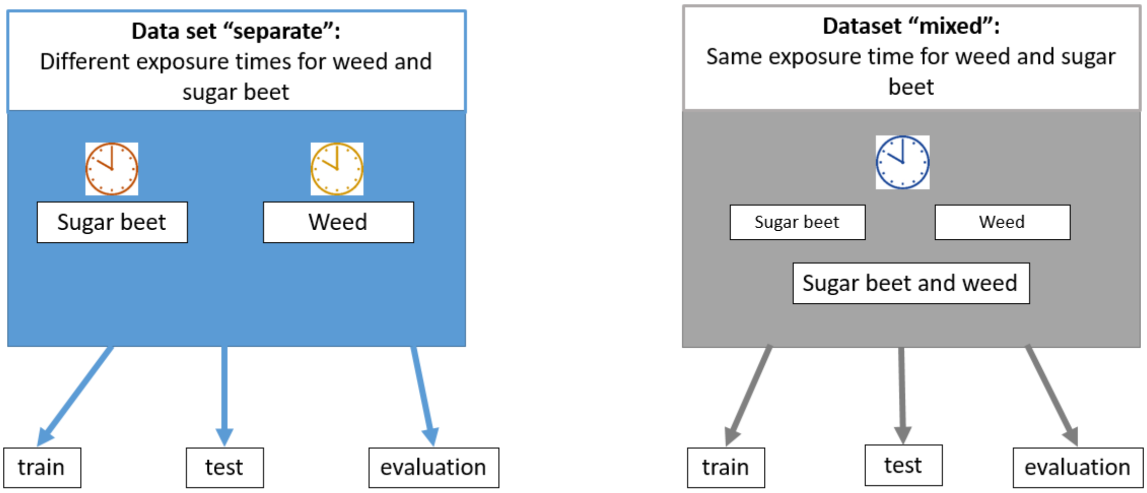


Figure 12: Dataset of image pairs of optical and hyperspectral images of sugar beet and weed in different pots, images acquired with different exposure times for the hyperspectral camera (left) and sugar beet and weed photographed with the same exposure time for the hyperspectral camera (right) and two pots with sugar beet and weed plants together.

Table 2: Apertures tested and corresponding exposure times for the hyperspectral camera MV1-D2048x1088-HS02-96-G2 (Photonfocus AG), lens: Edmund Optics, 35 mm/F1.65 67716 VIS-NIR, bandpass filters 600 - 875 nm.

Date	Aperture	sugar	weed
		exposure time [ms]	exposure time [ms]
20.2.20	f/2.8	21	14
24.2.20	f/2.8	33	20
25.2.20	f/2.8	27	17
26.2.20	f/2.8	21	13
28.2.20	f/6	88	52
2.3.20	f/11	262	158
3.3.20	f/11	326	179
5.3.20	f/14	349	349
6.3.20	f/8	127	127
9.3.2020	f/8	131	131

spatial context for the segmentation masks.

Absorbance A was calculated from reflectance values + 1 with eq. 4 and scaled afterwards to values between 0 and 1 with min-max-scaling (see eq. 19) since values between 0 and 1 are beneficial for the training speed of ANN's. Reflectance values were increased by 1 to avoid values approaching negative infinity through the log-transform (see eq. 4) since this caused problems for the EMSC correction.

$$A_{scaled} = \frac{A - min(A)}{max(A) - min(A)} \tag{19}$$

3.4.2 Scatter correction and normalization

There were two final classifiers, one that distinguished between plants and background and one that differentiated between weed and sugar beet. No scatter correction was performed for the plant versus background classifier since the difference between plant and soil is large enough. For the sugar beet versus weed classifier, scatter correction was performed with either SNV (see eq. 5) or EMSC (see eq. 7) with degrees between 0 and 2 for all plant pixels for each aperture separately. SNV and EMSC were evaluated based on the classification results, and the best option was chosen for the final classifier. In order to reduce magnitude effects cause by different overall brightness of a pixel, L2 normalization based on formula eq. 20 was applied. Given a vector v of length three with elements v_1,v_2,v_3, the L2-norm of v is defined by:

$$v_{l2} = v * \frac{1}{\sqrt{v_1^2 + v_2^2 + v_3^2}} \tag{20}$$

3.4.3 Ground truth segmentation masks

Ground truth masks were generated automatically by applying NDVI using equation 2 and afterward manually determined thresholds for binarizing the NDVI-images into plants and background. For the mixed data set, this was done for the separate pots and additionally, sugar beet and weed pixels were manually assigned in the mixed pots with the rectangular annotation tool from spectralpython's spectral.imshow.

3.4.4 Filter tests

Spectral quality was defined regarding the task to distinguish between sugar beet and weed based on the spectral information only. Images for the spectral quality tests were captured with aperture f/2.8. The spectral quality of the different filter options was tested based on visual assessment of the calibrated and EMSC- and L2-corrected mean spectra, including the standard deviation. For the filter tests, the same number of randomly selected sugar beet and weed images was used, so half of the indicated total number of images in tab. 3 were crop and weed images respectively.

3.4.5 Classifier: Neural networks

All neural networks were built with the TensorFlow API, version 2.1.0. All classifiers were applied pixel-wise such that one sample was a vector with the 15 corrected wavebands.

Only spectral features were used to keep the model simpler and to determine the contribution of spectral information towards classification. Hyperparameters such as batch size and learning rate were optimized based on the learning curve of dice coefficient and loss function. A batch size of 32, a learning rate of 1e-5, Nadam optimizer and class weights w were used for the final classifiers. Class weights are a valid option to counteract class imbalances by providing the fraction for all classes (see eq. 21). In eq. 21, n refers to the number of samples

$$w_{class\ 1} = \frac{n_{total}}{n_{class\ 1}} \tag{21}$$

The data was shuffled before each epoch to change the batch composition and, hence, increase the classifier's robustness (see p. 13). The parameters precision (see eq. 10), recall (see eq. 9) and dice coefficient (see eq. 12) were monitored since they are all suited for classifications with imbalanced classes which was the case for this data set (see tab. 4). The priority was a high recall. Because, explained with the example of the weed (”negative class”) versus sugar beet (”positive class”) classifier, a high recall implies that few sugar beet pixels are mislabeled as weed (see pp. 14 - 15) since this would lead to the sugar beet being removed in the worst case which would be a higher economical damage than one unrecognized weed plant. Secondly, a high precision means that only a few weed pixels are misclassified as sugar beets, leading to most weeds being recognized by the robot and extinguished (see pp. 14 - 15). Since the dice coefficient is a combination of recall and precision, a high dice coefficient was the overall goal.

3.4.5.1 Three-classes-CNN

A simple convolutional neural network described by Hu et al. [51] (see p. 14) for all three classes (background, sugar beet, weed) was built and trained in a first attempt. The loss function was categorical-cross entropy and the input data was not scatter-corrected or L2-normalized since spectra of background and plants are very distinct. Also, the average uncorrected sugar beet and weed spectra that was only calibrated had shown differences between plant species. For comparison, the same data was classified with PLS-DA. The network architecture was:

1. Input layer (number of wavebands)
2. Convolutional layer among spectral domain (filter size 3, 6 filters)
3. Batch normalization
4. Relu activation
5. Max Pooling layer
6. Flatten layer
7. Dense layer (6 nodes)
8. Batch normalization
9. Output layer (Dense Layer, 3 nodes: sugar beet, weed, background)

3.4.5.2 Two binary ANNs

The second classifier consisted of two ANNs, both with the same architecture and only fully connected layers. Each fully connected layer was followed by Batch-normalization and an activation with relu. One of these building blocks of fully connected layer, batch normalization layer and activation layer will be called DenseNorm.

1. DenseNorm (128 nodes)
2. DenseNorm (64 nodes)
3. DenseNorm (64 nodes)
4. DenseNorm (32 nodes)
5. DenseNorm (32 nodes)
6. DenseNorm (16 nodes)
7. DenseNorm (16 nodes)
8. DenseNorm (8 nodes)
9. DenseNorm (8 nodes)
10. DenseNorm (4 nodes)
11. Output: Dense, activation sigmoid (1 node, values between 0 (weed) and 1 (sugar beet))

The loss function was binary-cross-entropy. The plant versus soil classifier was trained on all apertures together and the data that was only calibrated since differences between soil and plants are big enough even with scatter. For the separate data set, one sugar beet versus weed classifier was trained once for all apertures together and then a net with the same architecture was trained for all apertures separately. For the mixed data set, sugar beet versus weed ANN's with the same architecture were trained separated by aperture.

3.4.6 Train-test-evaluation split

The input data for the plant versus soil classifier and the three-classes-CNN consisted in all pixels with the distributions described in tab. 4. For the sugar beet versus weed classifier, only plant pixels were used as input.

For all classifiers, 70 % of the data was used as training data, 15 % for testing and 15% for evaluation. The same fraction of each class of the original data set (see tab. 4) was maintained in the train, test and evaluation set. Additionally, the unlabeled portion of the mixed data set was used for evaluation.
All shown classification masks were median filtered in order to emphasize the overall classification and suppress noise.

Table 3: Dataset for testing different bandpass filter options for the hyperspectral camera MV1-D2048x1088-HS02-96-G2 (Photonfocus AG) and bandpass filters from Edmund Optics (35 mm/F1.65 67716 VIS-NIR), half of the number of images are sugar beet, the other half weed.

Date	**Filter**	**Number of images**
20.02.2020	No filter	46
20.02.2020	600 - 875 nm	68
20.02.2020	675 - 975 nm	64

Table 4: Number and percentage of pixels for a pixel classifier divided into background, sugar beet and weed of a timeseries of hyperspectral images (15 bands, red - near infrared region, size of one image: 409 x 216 pixels) for different apertures, spectral data acquired with MV1-D2048x1088-HS02-96-G2 (Photonfocus AG), lens:Edmund Optics, fcocal length (f): 35 mm/F1.65 67716 VIS-NIR, bandpass filters 600 - 875 nm.

Aperture	**images**	**Pixels**	**Sugar beet**	**Weed**	**Background**
	[number]	[number]	[%]	[%]	[%]
f/2.8	318	28,093,392	2,10	8,20	89,60
f/6	109	9,629,496	2,80	7,70	89,40
f/11	179	15,813,576	2,70	12,60	84,70
f/14	89	7,862,616	1,40	2,90	95,80
f/8	168	14,841,792	1,70	4,10	94,20
Total	863	76,240,872			

3.5 Image registration of the RGB and spectral images

The reflectance values were used for all image registration approaches because the RGB image values were also given as reflectance, so no transformation to absorbance was required. Additionally, the estimation of absorbance from reflectance was not necessary as no conclusions about the chemical composition were drawn. Two image registration methods were applied and assessed. First, the function imregdemon from Matlab was utilized as a local motion image registration approach, which mitigates parallax effects (see p. 15). Since imregdemon requires grayscale images, different channels and combinations of three channels were tried for one RGB and corresponding spectral image. Also, the region of interest was cropped out from the RGB images.

The second image registration method applied the concept of projection and involved therefore, the computation of a homography matrix based on matching keypoints in the two images to be aligned (see pp. 15 - 17). Furthermore, the homography-based image registration method took advantage of the fact that the camera positions to each other and relative to the target did not change during one image acquisition session and that plant heights were mostly similar for one session. Therefore, homographies were calculated for each image, and then the homography for one session was computed using the element-wise mean over the best homographies. The best homographies were found by using several evaluation metrics. The registration process was performed with OpenCV, version 3.4.2 and an OpenCV function in this thesis starts with "cv2". One image pair refers to the RGB and corresponding spectral image. The overview over the developed workflow is given now and certain steps are explained more detailed afterward:

1. Convert RGB and spectral images to grayscale with best color options for each image pair with cv2.cvtColor and the mode cv2.BGR2gray.
2. Detect key points with SIFT (cv2.xfeatures2d.SIFT_create) for each image pair.
3. Match keypoints with Bruteforce matcher (cv2.BFMatcher(cv2.NORM_L1)) for each image pair.
4. Disregard keypoints with a distance bigger than 60 and Lowe's distance ratio using eq. 14 with a threshold for each image pair.
5. Compute the homography for each image pair based on the remaining keypoints using RANSAC algorithm and a tolerance threshold of 5 pixels using cv2.findHomography for each image pair and save the computed homography matrix.
6. Determine the projection of the edges of spectral image onto the RGB with cv2.perspectiveTransform for each image pair.
7. Determine the warped spectral image with cv2.warpPerspective for each image pair, based on the homography computed in step 6.
8. Assess the quality of the match for each image pair with MSE, SSIM, the sum of the correlation matrix, sharpness (variance of the Laplacian), spatial filtering and save the results along with the corresponding homography.

9. For each folder, filter out the homographies that passed the spatial filtering test (sum of errors = 0) and then the homographies with the best values for all the other metrics (lowest MSE, highest sharpness, highest SSIM, the sum of correlation matrix).

10. Compute the session's homography by taking the element-wise mean over the filtered homographies.

The distance of 60 for matching keypoint was set relatively high because some correct matches had a huge distance due to different reflectance values of RGB and spectral image. For the whole workflow of image registration, grayscale versions of the RGB and the spectral image were required. The RGB images were transformed to grayscale by taking into account all three channels in a weighted fashion with the green channel having the strongest influence (see OpenCV documentation: [2]). The assumption for the spectral images was that combining information of three channels to one grayscale image instead of using one channel would provide a better result for the registration since more wavebands would contribute to contrasts and edges. Consequently, for each spectral image, every possible three-channel combination out of all spectral channels was tried out and assessed. For the assessment, two methods were applied in a brute force way over a representative subset of all images (n = 250). First, the SSIM for the grayscale RGB and the grayscale spectral image was computed for every possible three-channel-combination for the spectral image. SSIM was computed with the function from the library skimage, version 0.15.0 (skimage.measure.compare_ssim). The occurrence of channels that resulted in the highest SSIM for each image was determined and stored. The second approach was to apply the image registration process until step 4, meaning finding and filtering the matching keypoints between the images, and then safe the channel combinations for the grayscale conversion of the spectral images that produced the highest number of matches for each image pair. Another tried approach was to use automated cropping of the overlapping region of the RGB images in order to improve the quality of the matches by excluding possible mismatches in advance.

In order to find the best homography for each session, several metrics were used for filtering. The metric called spatial filtering was explicitly developed for this project and is based on the spectral image's four corner points being projected onto the RGB image plane. There were five checkpoints for the spatial filtering that are illustrated in fig. 13:

1. The corner points have to be in the same order as before.
2. Vertical alignment of the upper two points: vertical difference in pixels ¡ threshold
3. Vertical alignment of the lower two points: vertical difference in pixels ¡ threshold
4. Horizontal alignment of the left two points: horizontal difference in pixels ¡ threshold
5. Horizontal alignment of the right two points: horizontal difference in pixels ¡ threshold

The violation of the first checkpoint resulted immediately in the highest error score of 4 and no further checks were performed since this homography was obviously completely wrong (see top scetch in fig. 13). The violation of one of the other four criteria resulted in adding 1 to the error score (see middle row of fig. 13). So, the best spatial filtering result was an error score of 0 and the worst an error score of 4. Thereby, spatial filtering prevented the worst homographies.

The quality assessment metrics SSIM, MSE and sum of correlation matrix were determined using the warped spectral image onto the RGB plane and the corresponding part of the RGB image. SSIM was calculated with the already mentioned skimage's SSIM function. The correlation between RGB and the aligned spectral image was computed as the cross-correlation between the two images and the sum of correlation between the two images numpy.corrcoef (NumPy version 1.16.5). The mean squared error (MSE) was calculated with eq. 16. The "sharpness", or rather variance of the Laplacian, was computed based on the superimposed image, meaning the original RGB and the aligned spectral image on top of each other (see fig. 14). The variance of the Laplacian was obtained with the OpenCV function cv2.Laplacian() and NumPy's variance numpy.var(). The kernel sizes 3 x 3, 5 x 5, 7 x 7 and 9 x 9 for the Laplacian of the superimposed image were calculated and treated as different parameters.

One person visually assessed each superimposed image by giving a grade for the quality of the alignment, bearing in mind the overall task to convey bounding boxes from the spectral image plane onto the RGB image plane. Visual assessment grades for the superimposed images were defined with grades between 1 and 3 with the criteria for each grade described in tab. 5 and illustrated in fig. 14. The first step of determining the best filtering parameters was to visually assess all superimposed images and then to compare the visual assessment grade with the boxplots of different parameters and counting how often low or high values of the parameter were associated with grade 1 or 2 (boxplot example: fig. 15). For example, fig. 15 shows an association between low values for MSE and a better image alignment. The parameters that showed the ability to separate good and bad matches by consistently being low or high for good matches were further evaluated in the second step. The second step was to find the best combination of filtering metrics (step 9 in the workflow). This was achieved by applying several different combinations of quality assessment parameters on a subset of 15 randomly sampled images per session (255 out of 863 images in total) and evaluating the resulting superimposed images with visual assessment grades. The mean visual assessment grade was used for comparison between different filtering combinations. The ground truth data for each folder, the filtering combinations were compared to, consisted in handpicked superimposed images with visual assessment grade 1 and the mean of those best corresponding homographies. Therefore, the ground truth grade represented the best possible result for the projection method.

Table 5: Grades and corresponding criteria for visual assessment of superimposed image of a RGB image and the corresponding, aligned hyperspectral image, ghosting refers to blur of not perfectly overlapping parts.

Grade	**Criteria**
1	No or little ghosting, no doppelgänger only parallax
2	Ghosting, but bounding boxes would still work, no doppelgaenger
3	Lot of ghosting, doppelgaenger

413 RGB images with the projected bounding boxes were assessed visually regarding the bounding boxes' spatial quality. A good bounding box encloses the plant tightly, a bad bounding box does not contain the plant or only a small fraction of the plant. The problem that overlapping plants of the same species are within one bounding box was not tackled within this thesis.

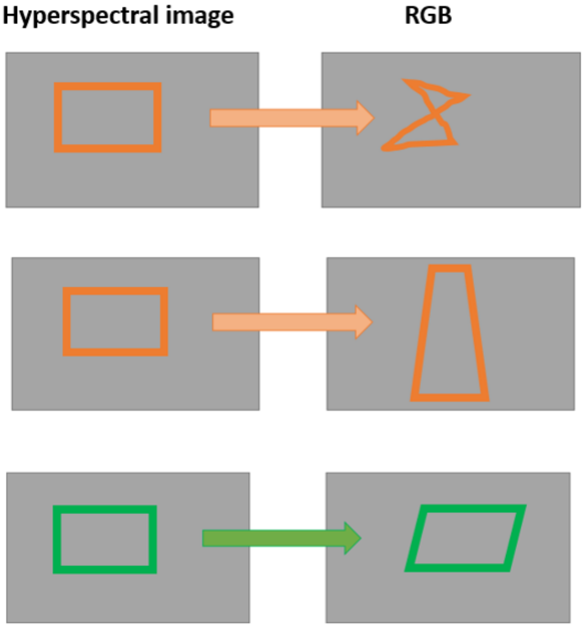
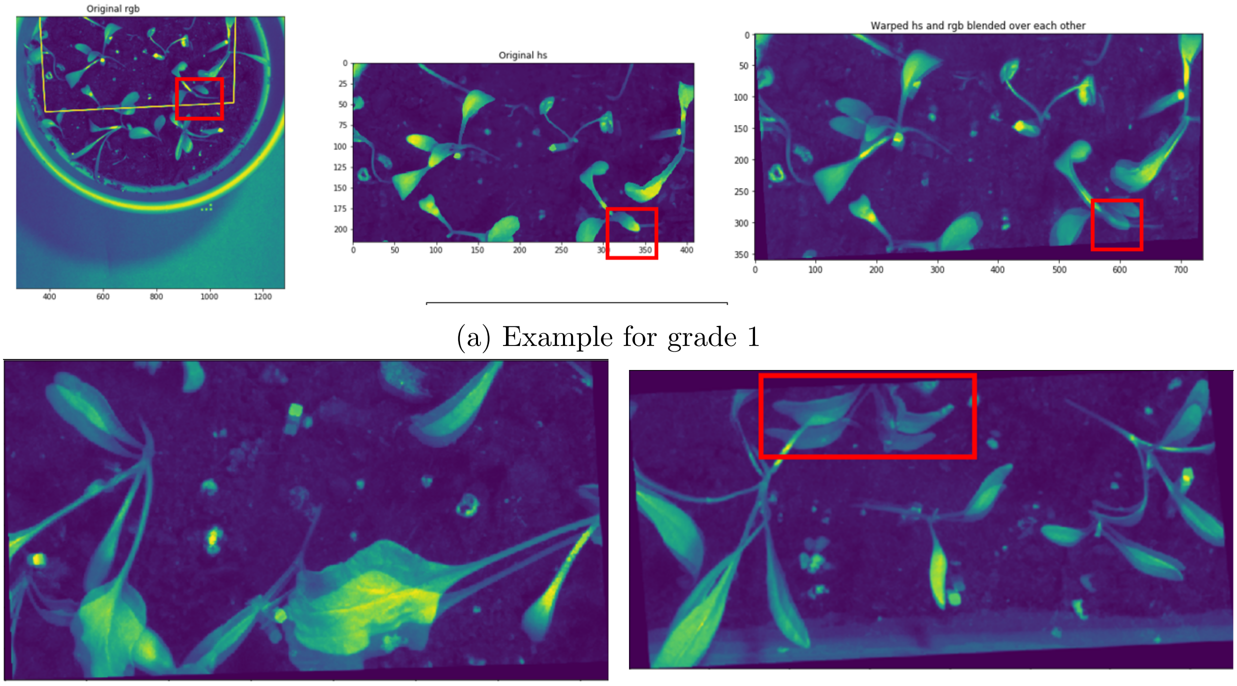


Figure 13: Examples of projections of a region of interest of a hyperspectral image onto a RGB image plane, images acquired with a stereo camera system (camera casings 0.5 cm apart), orange indicates a wrong projection, based on angles between contour lines and order of corner points, green indicates a right projection based on the same criteria.



(a) Example for grade 1

(b) Example for grade 2

(c) Example for grade 3

Figure 14: Examples for superimposed, aligned RGB and spectral images of plants and corresponding visual assessment grades (1: Best possible quality, 2: Good-okay, 3: Not usable) for the quality of the projection as a result of homography based image registration of an RGB and a hyperspectral image (near infrared region), red box for (a) indicates parallax, red box for (c) doppelgaenger.

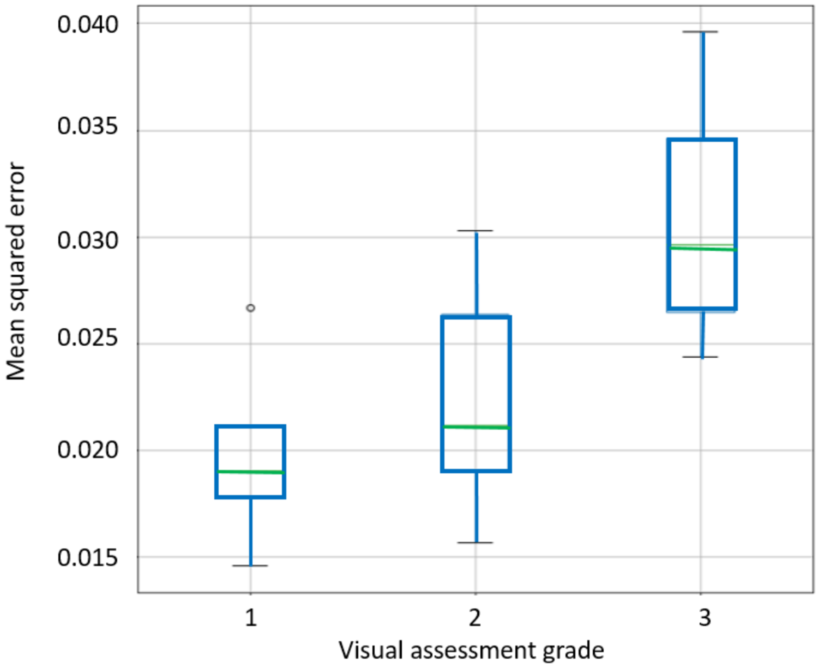


Figure 15: Boxplots of mean squared error (mse) grouped by visual assessment grade of image alignment (1: best, 2: good, 3:very bad), homography based image alignment of RGB and hyperspectral images ($n_{imagepairs} = 33$, one folder) with opencv

4 Results

Due to the lack of light during January in Dresden, Germany, all plants developed remarkably long stems. The following section also covers the results of failed or not completely satisfying attempts as documentation for future research, but the corresponding graphics or tables are not shown in every case.

4.1 Analysis and classification of the spectral data

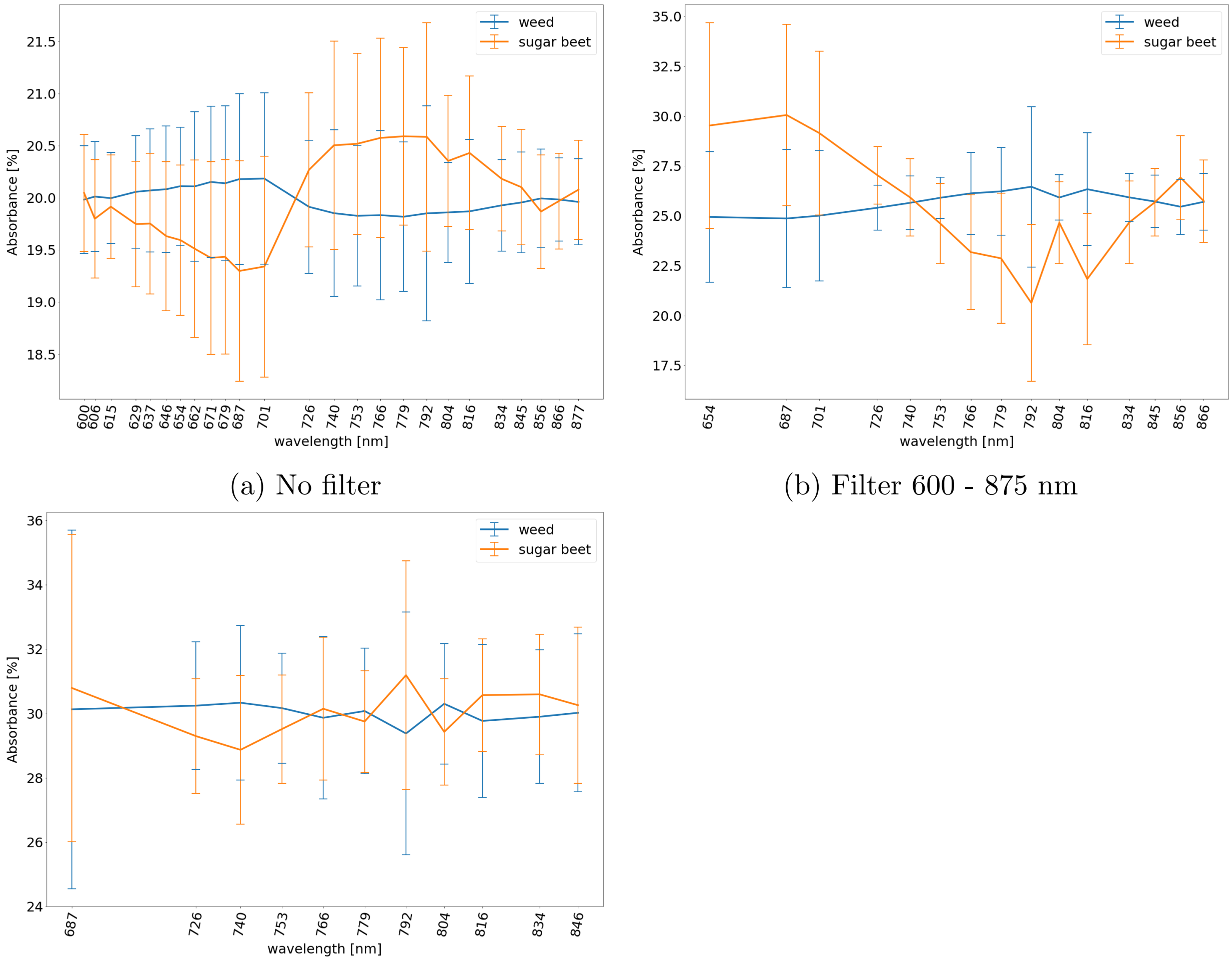
4.1.1 Determining the best bandpass filter for distinguishing between sugar beet and weed

Fig. 16 shows the average absorbance in % of sugar beet and weed without a bandpass filter and two different bandpass filters. In fig. 16(a) it can be seen that without a bandpass filter, the absorbance values of sugar beet and weed were very similar with 0.7 % being the highest difference between averages. Furthermore, the standard deviations overlapped greatly (see fig. 16(a)). Sugar beet and weed absorbance values with the filter 675 - 975 nm (see fig. 16(c)) differed maximum by under 2 %, and the standard deviations often went beyond the average absorbance of the other plant type. Average sugar beet and weed spectra for the filter option 600 - 875 nm differed by around 5 % for wavebands 654, 687, 792 nm (see fig. 16 (b)). Standard deviation error bars were overlapping but not crossing the other plant types mean spectra as this was the case for no filter or the filter 675 - 975 nm (see fig. 16). Overall, the average weed and sugar beet spectra had the most distinct absorbance values when using the bandpass filter 600 - 875 nm.

4.1.2 Classification based on the spectral data

Average reflectance values for plant pixels were relatively low (see fig. 31 in the appendix).The data visualized in fig. 17 was calibrated and converted into absorbance. The mean absorbance of the background was higher for all plant spectra for wavebands longer than 700 nm except for the plant pixels that belong to the aperture f/14 dataset. For wavebands greater than 726 nm, weed absorbance of the separate dataset (left side in fig. 17) was generally lower than sugar beet absorbance. For the mixed dataset, the opposite was the case with higher weed than sugar beet absorbance (see fig. 17, middle). It is noteworthy that all absorbance values except for aperture f/14 were between 10 and 25 percent, except for aperture f/14 tha had a lot higher absorbance values between 20 and 46 % (see fig. 17). Nevertheless, the basic pattern of all plant absorbance values was the same, showing higher values for wavebands 654 and 687 nm and then a decrease to a lower absorbance level for the rest of the wavelengths (see fig. 17). The decrease was particularly sharp for aperture f/14.

SNV and EMSC with degree(EMSC) = 0 were very similar and since higher degrees of EMSC were also tested, EMSC was chosen as the scatter correction technique. Fig. 18 illustrates the calibrated, EMSC corrected and L2-normalized spectra of weed and sugar beet aperture-wise. From fig. 18, it gets evident that the increasing degree of EMSC did not change the overall patterns of sugar beet and weed spectra. But the range of the corrected values became narrower for the separate dataset and stayed around the same for the mixed dataset for increasing degrees for EMSC (see fig. 18). Generally, the values for weed and sugar beet were more distinct for the separate dataset than for the mixed dataset (see fig. 18). Furthermore, the curve shapes for weed and sugar beet were re-



(a) No filter

(b) Filter 600 - 875 nm

(c) Filter 675 - 975 nm

Figure 16: Extended multiplicative scatter corrected and L2 normalized, calibrated average spectra of weed and sugar beet plants cultivated in the lab for (a) No bandpass filter ($n_{images} = 46$), (b) Bandpass filter 600 - 875 nm ($n_{images} = 68$), (c) Bandpass filter 675 - 975 nm ($n_{images} = 64$), each image consisted in 409 x 216 pixels with varying fraction of plant pixels.

verse for the separate and the mixed data set in some parts (see fig. 18). For the mixed dataset, the difference between weed and sugar beet was slightly bigger for aperture f/14 than for aperture f/8, particularly for wavebands 654 and 687 nm for degree(EMSC) = 1 and degree(EMSC) = 2, and for wavelength 792 nm for all degrees (see fig. 18). For the separate dataset, the patterns of weed absorbance were similar for all apertures, the same applied to sugar beet (see fig. 18). The average sugar beet and weed absorbance were the closest together for aperture f/2.8 and the most distinct for aperture f/6. With the increasing degree of EMSC, the difference between the quite similar spectra for aperture f/6 and f/11 and aperture f/2.8 became more evident (see fig. 18).

4.1.3 Classification of the spectral data

The PLS-DA misclassified few background pixels as plants, but 16 % of sugar beet and 4 % of weed pixels as background (see tab. 6). The confusion between sugar beet and weed was 12 % of mislabeled sugar beet and 14 % of mislabeled weed pixels (see tab. 6).

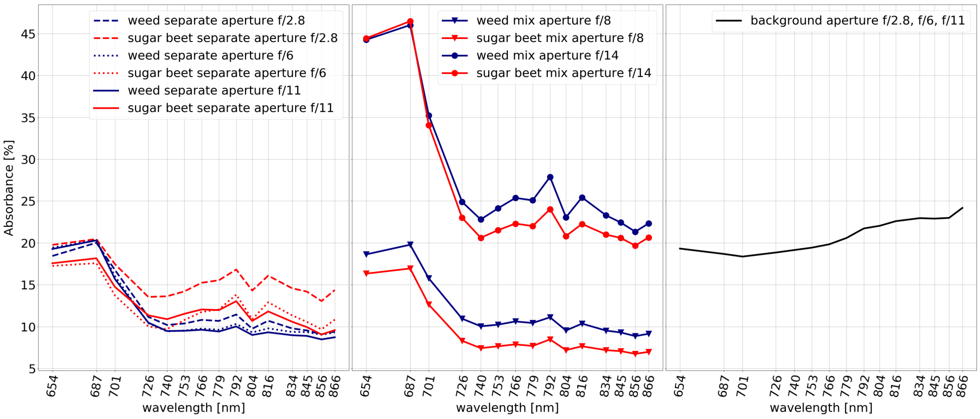


Figure 17: Average calibrated absorbance for different apertures of background (potting soil, plastic pots, table), sugar beet and weed plants [%] grown in pots, $n_{\text{images separate}} = 606$, $n_{\text{images mixed}} = 257$, $n_{\text{images background}} = 863$, resolution each image: 409 x 216 pixels with varying fraction of plant pixels.

Table 6: Confusion matrix of Partial Least Squares - Discriminant Analysis pixel classifications of calibrated multispectral images of sugar beet and weed plants, cultivated in the lab ($n_{\text{images}} = 863$), all numbers in % based on the true labels.

True label	**Predicted label [% of true label]**		
	Background	Sugar beet	Weed
Background	97.7	1.9	0.4
Sugar beet	16.5	71.3	12.3
Weed	4.4	14.6	81.1

The classification performance of the three-classes CNN for soil against plants worked out best with 95 % of soil pixels classified correctly (see tab. 7). Of the evaluation dataset, 89 % of sugar beet and weed pixels, respectively, were classified correctly, and most misclassifications occurred between sugar beet and weed (see tab. 7).

The final classification pipeline consisted of one classifier separating the plant from soil pixels, and classifying plant pixels into weed and sugar beet. The results are shown separated by aperture.

The plant versus background net reached correct classifications for over 98 % of plants and soil pixels (see tab. 8). The dice coefficients of validation and evaluation dataset were quite similar and both high with 0.87 for the validation set and 0.9 for the evaluation set (see tab. 8). From fig. 19, it gets apparent that most misclassifications happened at the margins of plants.

The sugar beet versus weed classifier was trained aperture-wise and on all degrees of EMSC separately. Only degrees of EMSC that produced the best results regarding the dice coefficient and visual assessment of the segmentation masks are presented in tab. 9. Solely, the confusion matrix for the evaluation set is shown since the class-wise accuracy

Table 7: Confusion matrix for the evaluation dataset pixel classifications of a neural network with one convolution over the spectral domain for multispectral images (15 channels) (n_{images} = 863, each image 409 x 216 pixels, 15 % of pixels belong to the evaluation dataset), all numbers in % based on the true labels.

True label	**Predicted label [% of true label]**		
	Background	Sugar beet	Weed
Background	95.8	3.8	0.4
Sugar beet	3.5	88.6	7.9
Weed	0.4	10.7	88.9

Table 8: Confusion matrix for the evaluation dataset for pixel classifications of a neural network with ten fully connected layers for distinghuishing between soil and plant pixels of multispectral images (15 channels) (n_{images} = 863, each image 409 x 216 pixels, 15 % of pixels belong to the evaluation and validation dataset respectively), separate refers to sugar beet and weed in separate pots and with different exposure times and mixed to sugar beet and weed in one pot with the same exposure time, all numbers in % based on the true labels, dice coefficient for the evalutation set.

True label	**Predicted label [% of true label]**	
	Background	Plant
Background	98.4	1.6
Plant	98.3	1.7
Dice coefficient validation set	0.87	
Dice coefficient evaluation set	0.90	

was representative of the whole dataset. The dice coefficient and percentage of correctly predicted labels were generally higher for the separate dataset compared to the mixed dataset (see tab. 9). Within the separate dataset, the aperture f/6 set had the highest dice coefficient and 95 % of sugar beet, such as 94 % of weed classified correctly (see tab. 9). The aperture f/2.8 set produced the second-best result of the separate dataset with around 90 % of plants classified correctly (see tab. 9). For aperture f/11, 94 % of sugar beet plants were classified correctly, but only 87 % of weed plants, resulting in the lowest dice coefficients for the separate dataset (see tab. 9). The results for all separate apertures together were better than for aperture f/2.8 and f/11 alone but worse than for aperture f/6 with around 90 % of sugar beet and weed pixels being classified correctly and an evaluation set dice coefficient of 0.79 % (see tab. 9).

For the mixed dataset, aperture f/14 resulted in slightly better classifications than aperture f/8 (see tab. 9). Even though aperture f/14 had better classification results for degree(EMSC) = 2 on a pixel level, the visual assessment of the images showed that the classification for sugar beet was clearer and better, but more plants were misclassified in general (see fig. 26). As the example classification masks for aperture f/2.8 (see fig. 20) show, sugar beet of different development stages (2-leaf stage and 4-leaf stage) and different weed species were mainly classified correctly on plant level. Problems for aperture f/2.8 were fuzzy margins of leaves especially at the edges of the image (see fig. 20 (a) and (b)) and shadowed shoot parts for weed (see fig. 20 (c)).

Fig. 21 confirms the outstanding high classification accuracy and dice coefficients for aperture f/6 in tab. 9 for sugar beet plants of different age and different weed species. Still, mainly shoots of weed plants were misclassified.

Even though the classification results for aperture f/11 were the worst for the separate dataset, the overall results on a plant level were satisfying for weed and sugar beet, as depicted in fig. 22.

Further, the segmentation masks resulting from training and testing for all separate apertures together were as satisfying as the classification accuracy and dice coefficients. Even though some plant pixels were assigned to the wrong class (see fig. 23), on a plant level, the classifications were generally correct.

The sugar beet plant pixels that were part of the training dataset of aperture f/8 were mostly classified correctly (see fig. 24(a)). However, a noticeable fraction of weed pixels of the training dataset (see fig. 24(d)) was not labeled rightly. Even with annotated pixels of sugar beet and weed pixels in the mixed pot, the classifier did not succeed in distinguishing between sugar beet and weed in the mixed pot (see fig. 24(b) and (c)).

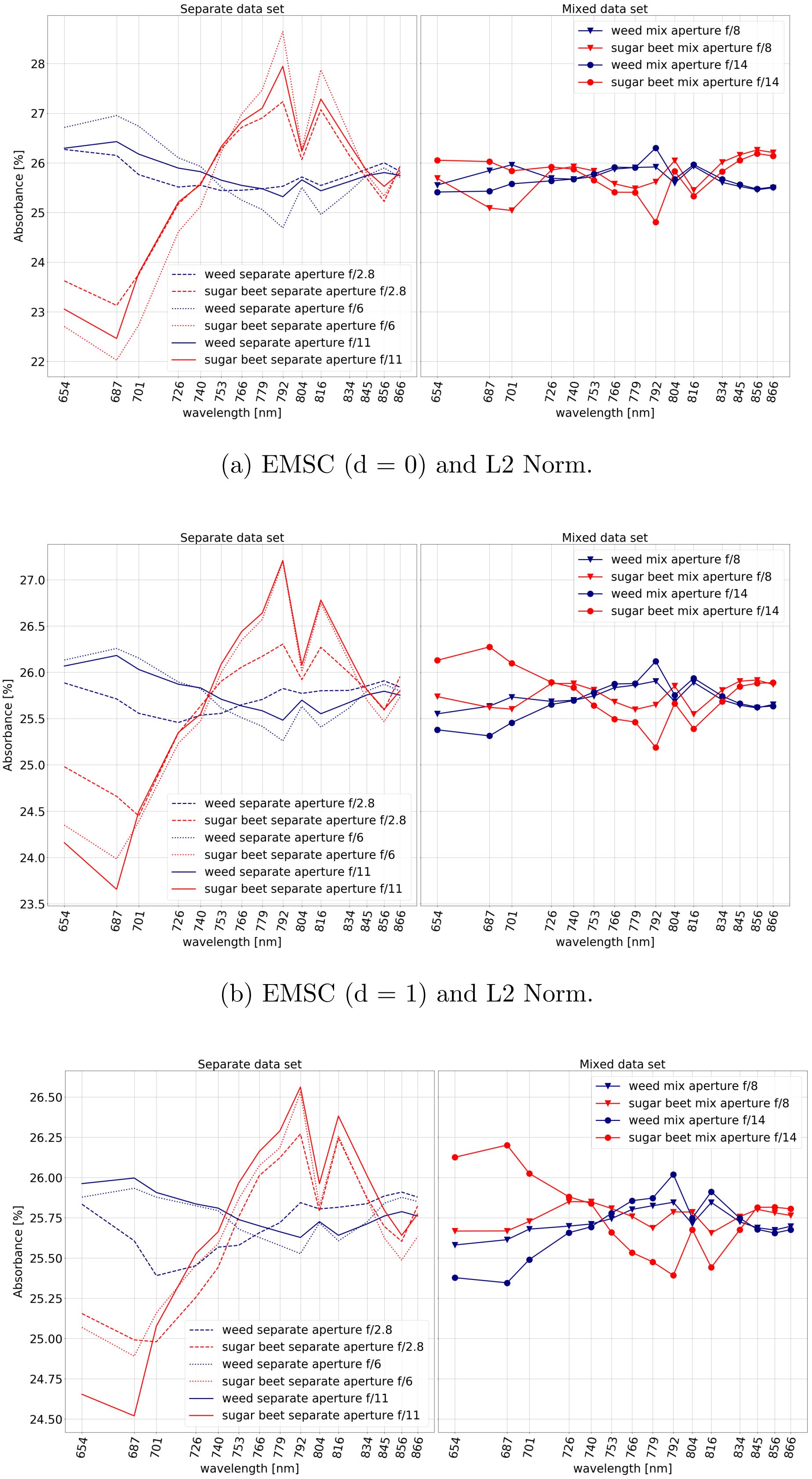
This was better for aperture f/14 and degree(EMSC) = 0, as it can be seen in fig. 25. Sugar beet pixels of the separate (see fig. 25(a)) and the mixed pot (see fig. 25(b)) and different development stages were mostly labeled correctly. Still, shadow areas and young leaves were mostly predicted incorrectly (see fig. 25(a) and (b)). That the classification results on a pixel level were worse than the separate dataset (see tab. 9) was reflected on plant level for weeds, especially for small plants and shadow areas (see fig. 25(c) and (d)). Nonetheless, when annotating a label on a plant level, the majority of weeds were labeled correctly (see fig. 25(c) and (d)).

When correcting the dataset of aperture f/14 with EMSC of degree 2, the classification results on pixel-level improved (see tab. 9). In the segmentation masks, it can be observed that the labeling for sugar beet did indeed improve with fewer pixels misclassified as weed (see fig. 26(a), (b) and (d)). On the other hand, the performance for weed declined, again, particularly for short plants and angled plant parts (see fig. 26(c) and (d)).

For all classifiers, the weed species lambsquarters was often misclassified.

Table 9: Confusion matrix for the evaluation dataset for pixel classifications of a neural network with ten fully connected layers for distinghuishing between sugar beet and weed pixels of multispectral images (15 channels) ($n_{\text{images f/2.8}} = 318$, $n_{\text{images f/6}} = 109$,$n_{\text{images f/11}} = 179$, $n_{\text{images f/8}} = 168$,$n_{\text{images f/14}} = 89$, each image 409 x 216 pixels, 15 % of plant pixels belong to the calibration and evaluation dataset respectively), separate refers to sugar beet and weed in separate pots and with different exposure times and mixed to sugar beet and weed in one pot with the same exposure time, all numbers in % based on the true labels, dice coefficient for the evalutation set.

dataset /d(emsc)	**Aperture**	**True label**	**Predicted label [%]**		**Dice coefficient**
			sugar beet	weed	evaluation set
separate d = 0	f/2.8	Sugar beet	89	11	0.81
		Weed	92	8	
separate d = 0	f/6	Sugar beet	95	5	0.92
		Weed	94	6	
separate d = 0	f/11	Sugar beet	94	6	0.75
		Weed	87	13	
separate d = 0	f/2.8, f/6, f/11	Sugar beet	91	9	0.79
		Weed	90	10	
mixed d = 0	f/8	Sugar beet	81	19	0.66
		Weed	72	28	
mixed d = 0	f/14	Sugar beet	73	27	0.68
		Weed	75	25	
mixed d = 2	f/14	Sugar beet	79	21	0.71
		Weed	79	21	

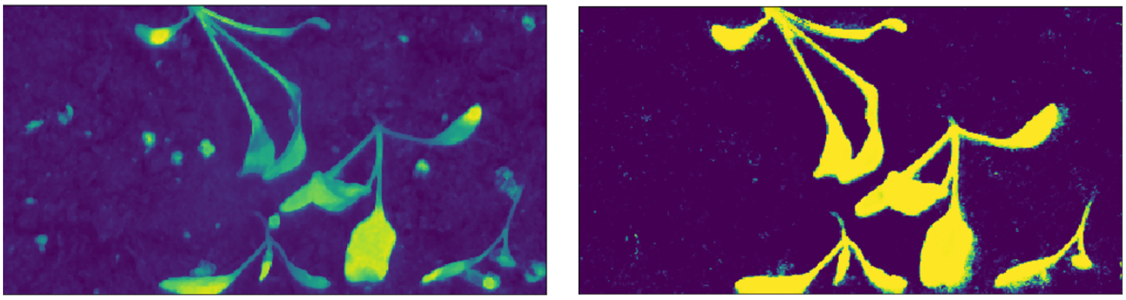


(a) EMSC (d = 0) and L2 Norm.

(b) EMSC (d = 1) and L2 Norm.

(c) EMSC (d = 2) and L2 Norm.

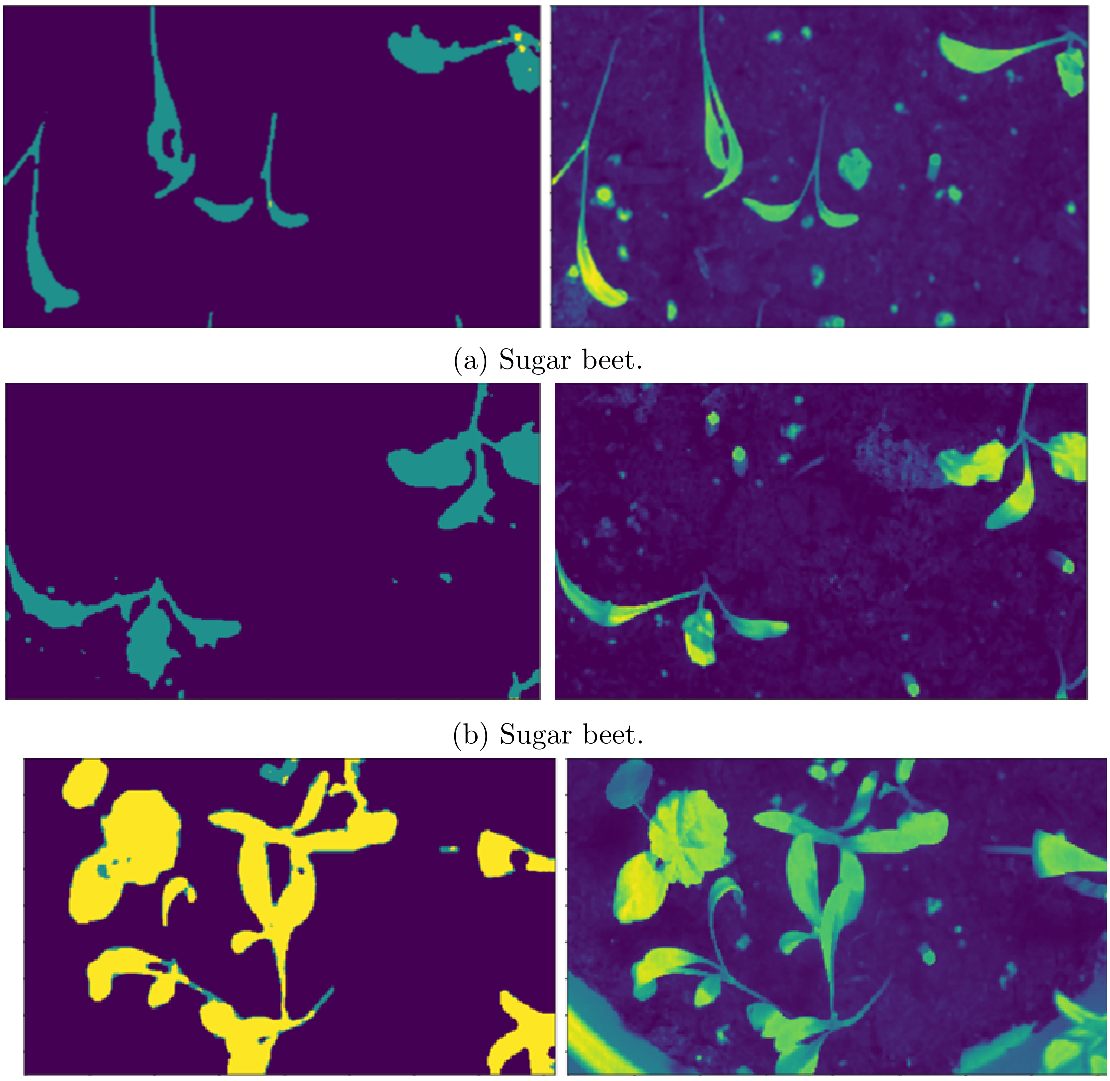
Figure 18: Extended multiplicative scatter corrected (EMSC) and L2 normalized, calibrated average spectra of weed and sugar beet plants cultivated in the lab for (a) degree(emsc) = 0, (b) degree(emsc) = 1, (c) degree(emsc) = 2, each image consisted in 409 x 216 pixels with varying fraction of plant pixels, $n_{\text{images}} = 863$.



(a)

(b)

Figure 19: a: Original image, b: Prediction of a neural network with 10 fully connected layers for distinghuishing between soil and plant for multispectral images (15 channels), purple: background, yellow: plant

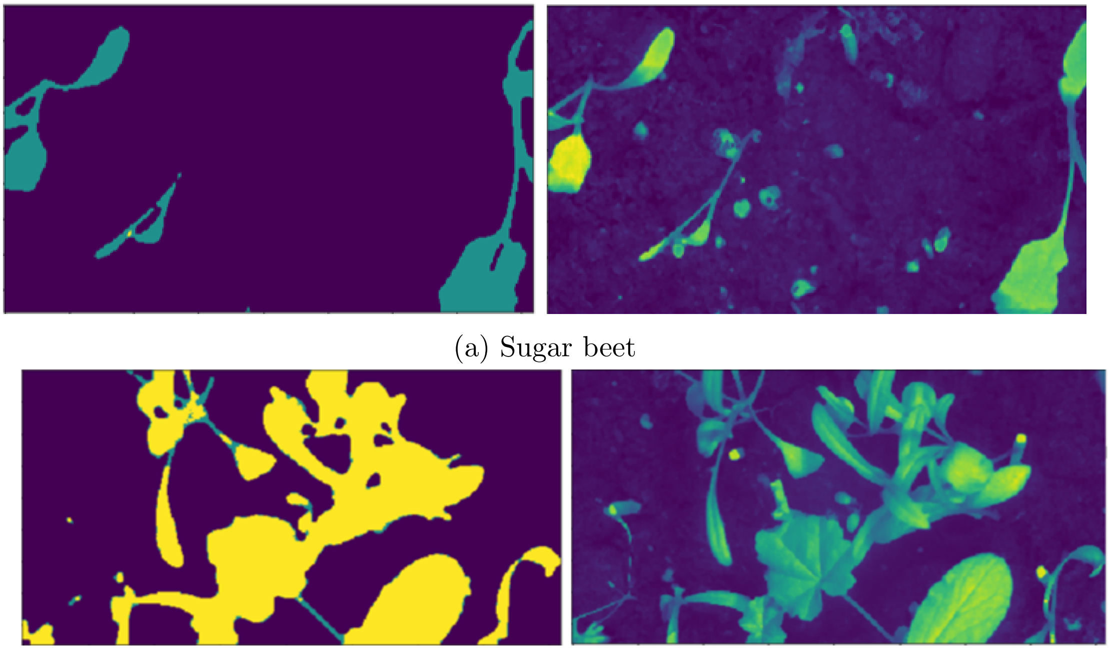


(a) Sugar beet.

(b) Sugar beet.

(c) Weed.

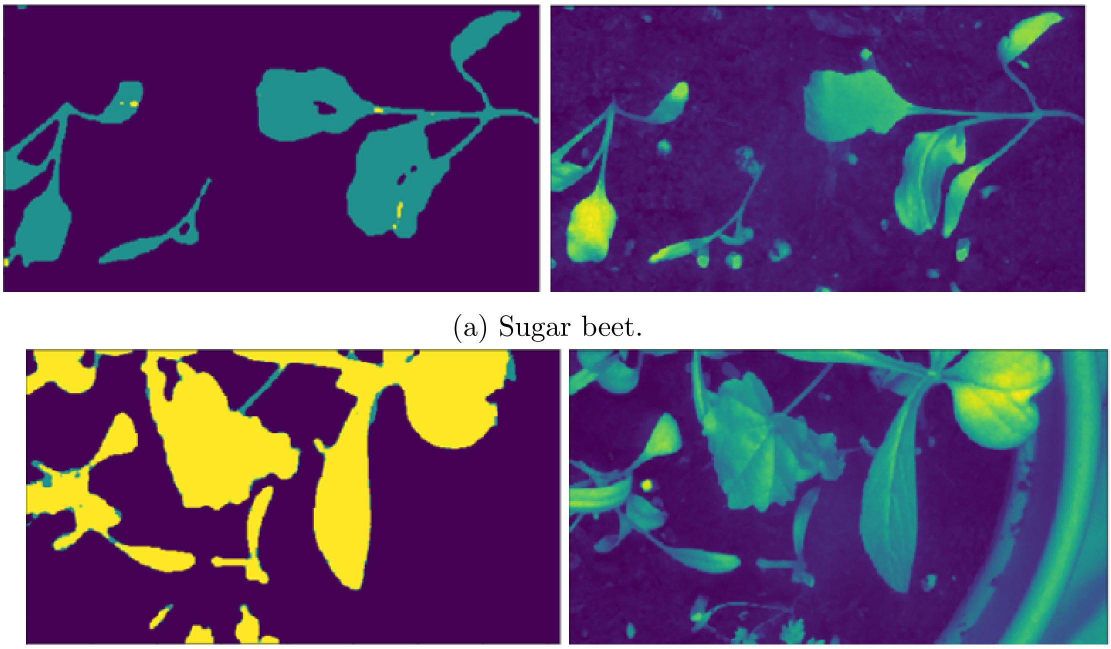
Figure 20: Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/2.8 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 318 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out.



(a) Sugar beet

(b) Weed

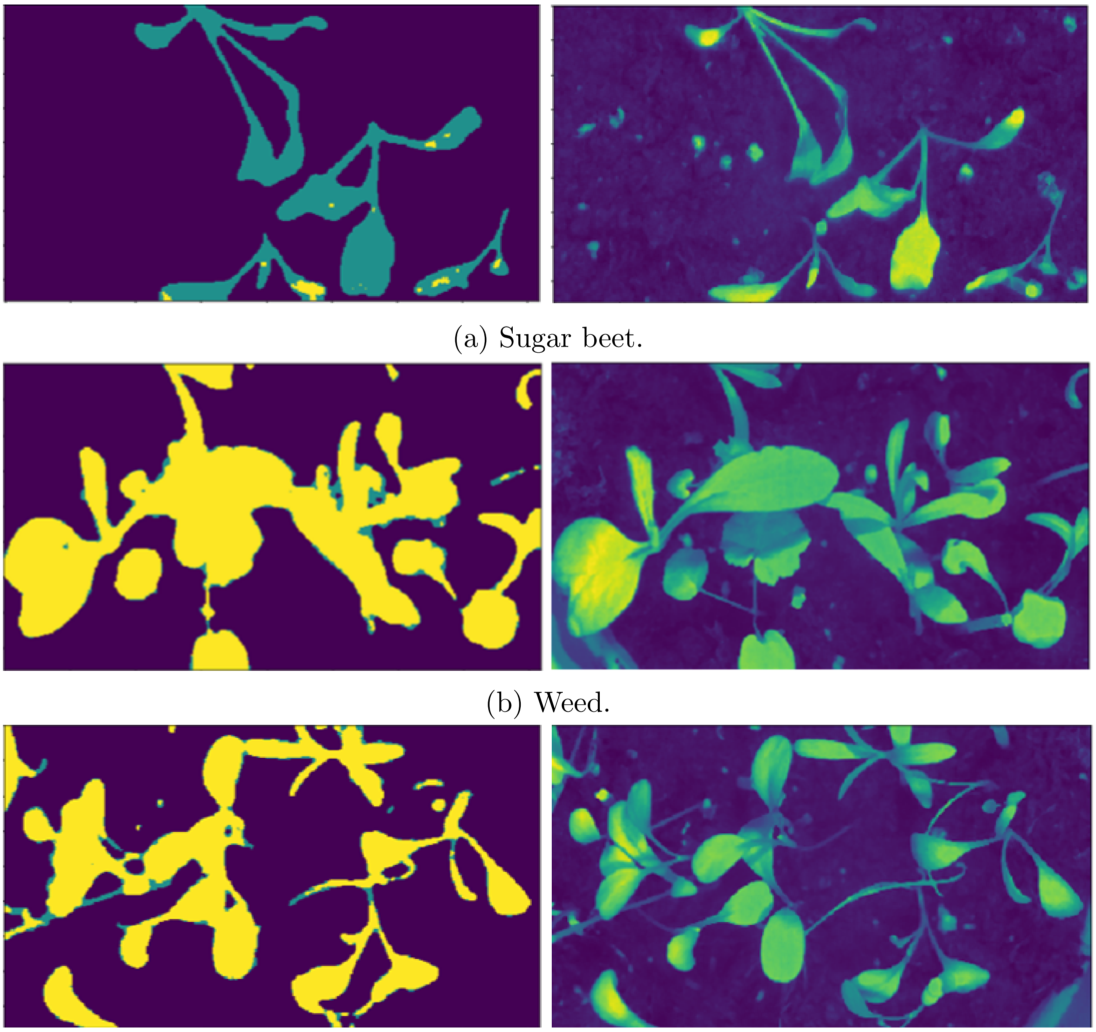
Figure 21: Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/6 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 109 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out.



(a) Sugar beet.

(b) Weed.

Figure 22: Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/11 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 179 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized multispectral images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out.

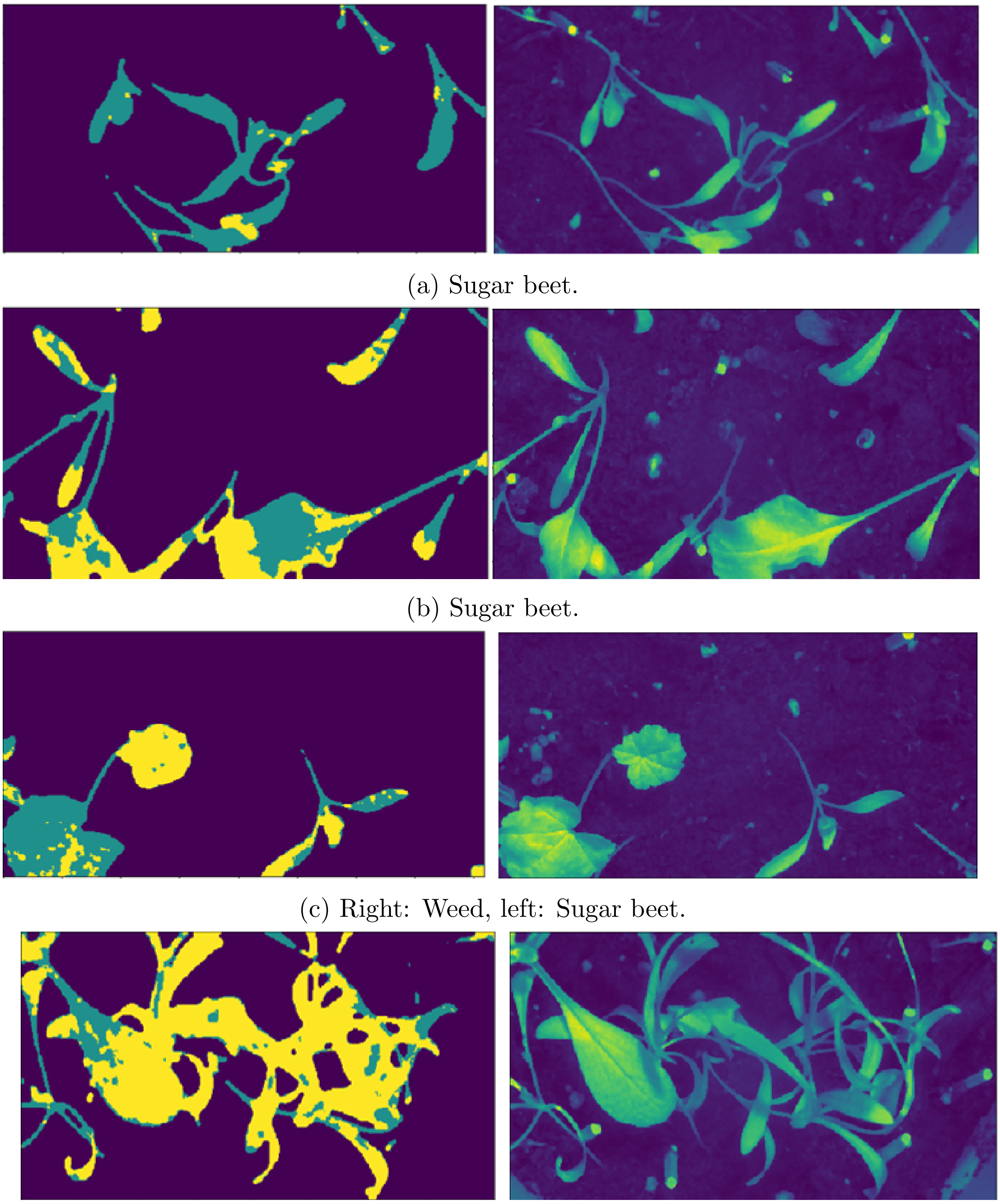


(a) Sugar beet.

(b) Weed.

(c) Weed.

Figure 23: Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/2.8, f/6, f/11 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 606 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized multispectral images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out.



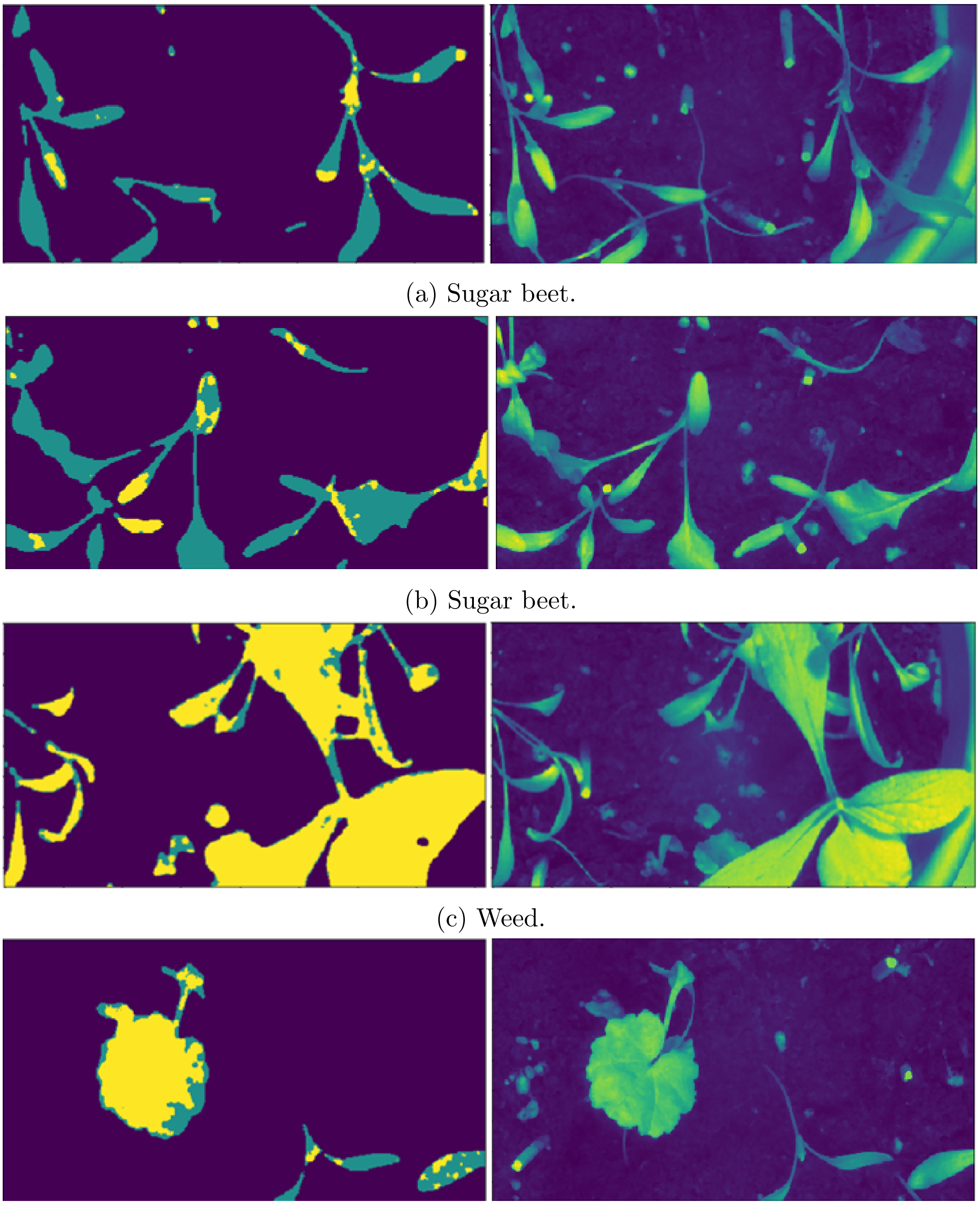
(a) Sugar beet.

(b) Sugar beet.

(c) Right: Weed, left: Sugar beet.

(d) Weed.

Figure 24: Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/8 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 168 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized multispectral images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out.



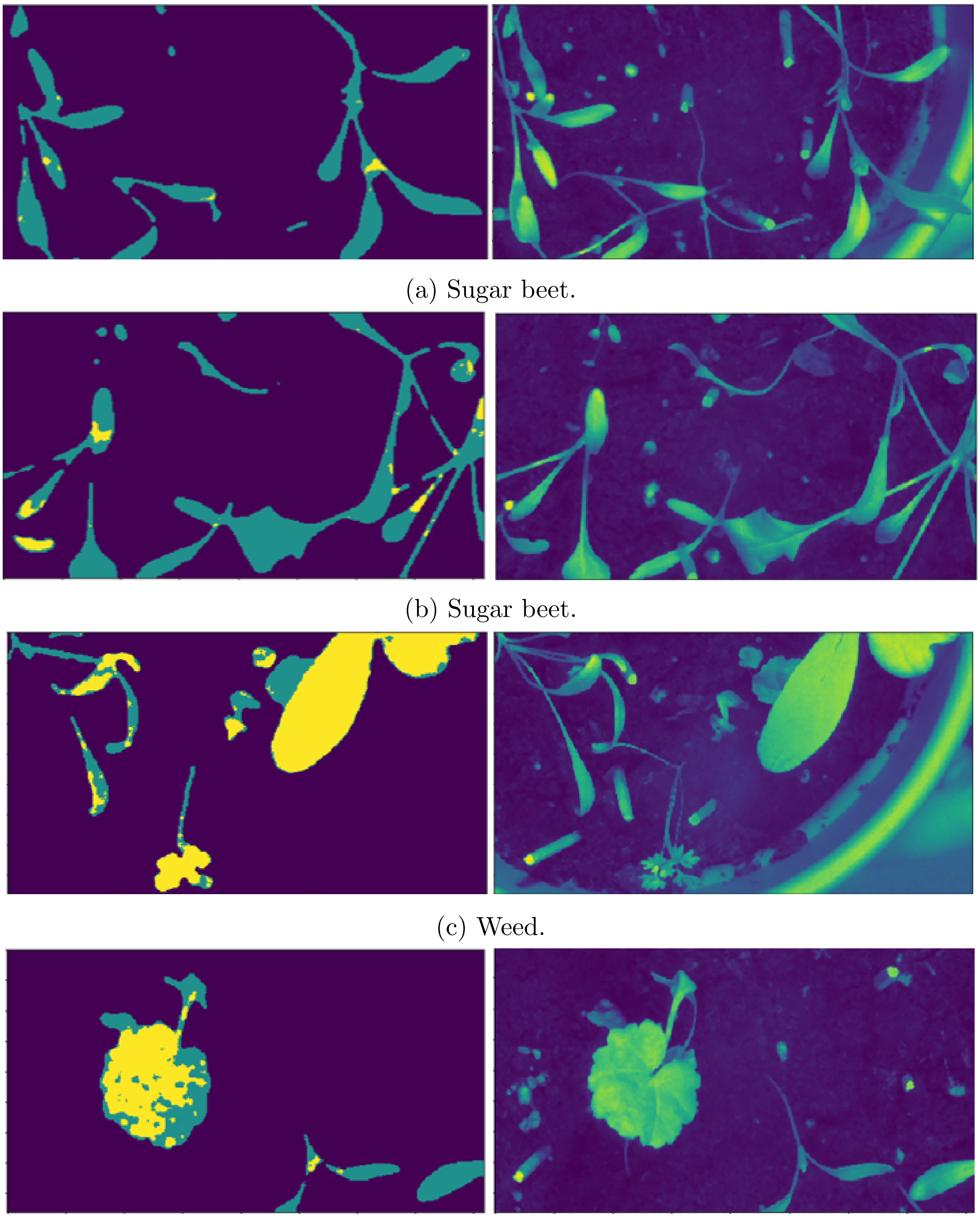
(a) Sugar beet.

(b) Sugar beet.

(c) Weed.

(d) Left: Weed, right: Sugar beet.

Figure 25: Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/14 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 89 calibrated, extended multiplicative scatter corrected (degree = 0) and L2 normalized multispectral images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out.



(a) Sugar beet.

(b) Sugar beet.

(c) Weed.

(d)

Figure 26: Examples for original grayscale images on the right and on the left predicted masks (yellow: weed, turuqoise: sugar beet, purple: background) for aperture f/14 of a ten layer neural network that distinguishes between sugar beet and weed, dataset consisted in 89 calibrated, extended multiplicative scatter corrected (degree = 2) and L2 normalized multispectral images with 409 x 216 pixels each with varying amount of plant pixels which were filtered out.

4.2 Image registration of the RGB and the spectral images

The imregdemon algorithm did not succeed in aligning the spectral image with the cropped RGB for none of the chosen channels or combinations of three channels (see fig. 27).

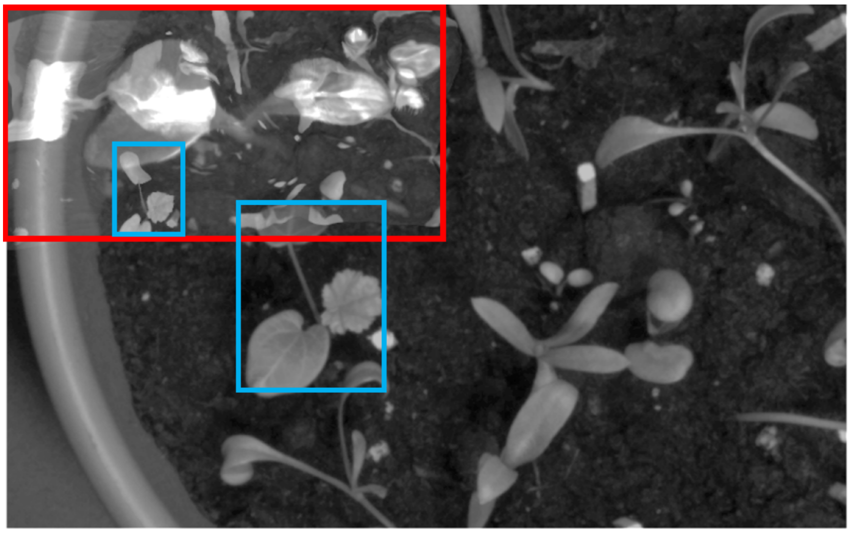


Figure 27: Superimposed grayscale versions of RGB image and aligned multispectral image (15 bands, 675 - 800 nm) of several weed species with the imregdemon algorithm from matlab.

The feature point matcher produced mismatches for most of the images (see fig. 28). This made the use of RANSAC and later filtering of the homographies necessary.

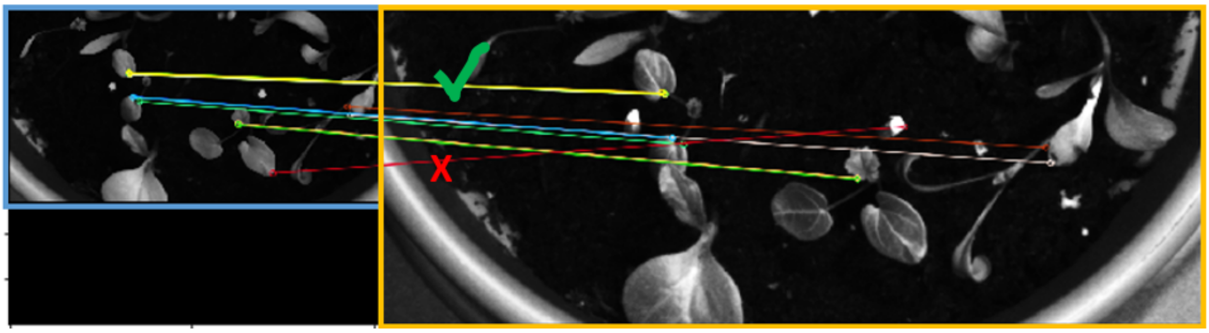


Figure 28: Correct and incorrect matches of image registration process of grayscake versions of an RGB image and a multispectral image (15 bands, 675 - 800 nm).

The best SSIM as an indicator for finding the most favorable channels of the spectral images for image registration with the RGB images, turned out to be inaccurate (results not shown). The parameter of highest number of matches between RGB and spectral image was utile and revealed that the waveband of 740 nm was best suited (see fig. 29) and that use of all three channels of the RGB images was the best option. Further, pre-cropping the RGB images before registration did not improve the image alignment (results not shown).

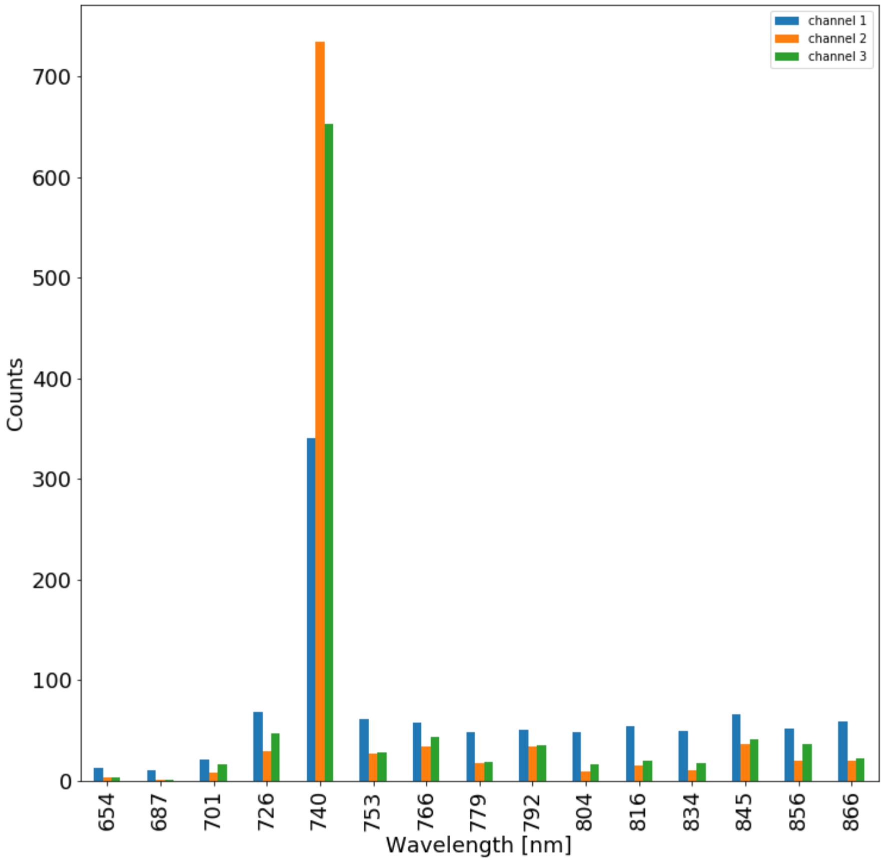


Figure 29: Counts of how many times a channel of multispectral images (15 bands, 675 - 800 nm) was used for grayscale conversion that led to most matches between and optical image and multispectral image during image registration process with Scale invariant feature transform feature detection (opencv 3.4.2).

4.2.1 Filtering of homographies

Lowe's distance ratio removed many correct matches and not all of the wrong matches, leaving only few or too few matches (beteen 0 and 30) for computing the homography (results not shown). The resulting homographies were often not sufficiently good, therefore the distance ratio threshold was set to 1 and thereby deactivated.
To determine the pixel off-set was tried to be calculated as the difference between edge images or segmentation masks. This was not applied because converting RGB and spectral data to masks or edge images led to too much different noise in both pictures. The cross correlation required long computing times and delivered similar results to the sum of the correlation coefficients between the RGB image and the superimposed image.

The results of the performance of different parameters for homography filtering, summarized in tab. 10, indicate that high quality in the superimposed images, as a result of adequate homographies, is assosciated with low values for the spatial filtering and MSE and high values for the sharpness measures for kernel sizes 5 x 5,7 x 7 and 9 x 9, SSIM and the sum of the correlation matrix (see tab. 10). Sharpness measures with kernel size 3 x 3 did not clearly indicate a good quality of the superimposed images (see tab. 10). The spatial filtering only yielded low results for the best homographies or indifferent results and the other parameters often had a tendency for the best homographies but also good values for wrong homographies (see tab. 10. Therefore, only homographies with the best spatial filtering results were kept and on those images, the other metrics were re-applied (see tab. 10, column "Counts after position filtering"). With this pre-step, high quality of superimposed images was associated with low values for MSE, high values for the sum of the correlation matrix and high values for the sharpness measure with kernel size 7 x 7 (sharpness 7) and 9 x 9 (sharpness 9) (see tab. 10, column "Counts after position filtering"). SSIM and sharpness measures with kernel size 3 x 3 and 5 x 5 did not deliver as clear results. Sharpness 7 x 7 and 9 x 9 accomplished very similar results for most images ((see tab. 10, column "Counts after position filtering")).

Table 10: Counts of folders for which certain metrics were useful to discriminate between classes of image alignment performance based on visual assessment of boxplots, maximum: 17, n_{folder} between 20 and 60, spatial refers to a the position and order of the projected four corner points, sharpness refers to the variance of the laplacian filtered image.

Parameter	**Counts**			**Counts after spatial filtering**		
Class 1 values	lower	higher	same	lower	higher	same
Spatial	**12**	0	5			
Sharpness kernel size 3 x 3	4	8	5	5	4	8
Sharpness kernel size 5 x 5	3	**12**	2	3	6	8
Sharpness kernel size 7 x 7	1	**13**	3	3	10	4
Sharpness kernel size 9 x 9	1	**13**	3	3	10	4
Mean Squared Error	**13**	3	1	**12**	2	3
Structual Similarity Index	4	**11**	2	5	9	3
Sum of correlation matrix	3	**11**	3	3	**11**	3

Tab. 11 displays the average visual assessment grades for several filtering options and for the ground truth. Sharpness in this table refers to sharpness measure with kernel size 7x7. This dataset yielded the best average grade with 1.5 (see tab. 11) and only 4 % of the superimposed images showing doppelgaenger (see tab. 12). The results of the spatial filtering alone were not satisfying but they showed that taking the mean for each element of the homography from all filtered homographies of this folder yielded better results than the median (see tab. 11). The combination of spatial filtering first and then choosing the homographies that had high values for sharpness and sum of correlation matrix and a low MSE at the same time, achieved similar results as only applying spatial filtering even though fewer homographies were involved for each folder (see tab. 11). Applying spatial pre-filtering and then averaging over the homographies with the lowest MSE's and the highest sharpness values showed with an average assessment grade of 1.87 better results then previously described filtering approaches (see tab. 11). The best results with an average of 1.64 delivered the pre-filtering with the spatial criteria and then picking the homographies with highest sharpness and sum of the correlation matrix between the RGB and the warped spectral image (see tab. 11). This was close to the ground truth average. In tab. 12, it can be seen that even though the means of the ground truth and the spatial + sharpness + sum_corr filter were close, the distribution of visual assessment grades among classes was slightly different. For the ground truth data, there were 10 % more images that got grade 1 than for the automatically filtered homographies (see tab. 12). The reverse was the case for grade 2 with 9 % more images for the filtering system compared to the ground truth (see tab. 12). With 6 % of the assessed superimposed images showing doppelgaenger, the filtering system performed only slightly worse than the ground truth.

Table 11: Average of visual assessment grade for different homography filtering options, based on 15 randomly picked images of each of the 17 folders ($n_{\text{per folder}} = 15, n_{total} = 255$, assessment grade 1: best possible alignment, 2: good - okay alignment with ghosting but no doppelgaenger, 3: alignment with doppelgaenger, MSE means Mean square error, sum corr refers to the sum of the correlation matrix between the overlapping part of the aligned images and spatial filtering was a check of the order and angles of the warped four corner points.

Type of homography filtering	**Combination**	**Average visual grade**
Ground truth	mean	1.5
Spatial	median	2.2
Spatial	mean	2.1
Spatial + sharpness + MSE + sum corr	mean	2.1
Spatial + sharpness + MSE	mean	1.87
Spatial + sharpness + sum corr	mean	1.64

Table 12: Performance of best filtering option for homography and the ground truth for the visual assessment grades, assessment grade 1: best possible alignment, 2: good - okay alignment with ghosting but no doppelgaenger, 3: alignment with doppelgaenger, MSE means Mean square error, sum corr refers to the sum of the correlation matrix between the overlapping part of the aligned images and spatial filtering was a check of the order and angles of the warped four corner points.

Grade	**Ground truth**		**Filter spatial + MSE + sum corr**	
	Number of images	Percent [%]	Number of images	Percent [%]
1	136	53	108	42
2	108	42	131	51
3	11	4	16	6

4.3 Projection of multispectral classification data onto the RGB plane

A representative example of the projection of pixel predictions of the spectral classifier can be seen in fig. 30(a). Some pixels match correctly but most show an off-set (seen fig. 30(a)). On a bounding box level, the main problems were the recognition of individual plants when plant parts overlapped (see fig. 30(b) and fig. 30(c)). Another problem was the off-set between box and plant caused by the image registration that was too big (see fig. 30(c)). Nevertheless, an off-set between bounding box and plant leading to an empty bounding box occured only for 4.4 % of 413 assessed images. And in most cases, for each image with empty bounding box, only one to two plants were affected. Fig. 30(d) illustrates a successful bounding box projection, the majority of plant parts are inside the bounding boxes. Nevertheless, the bounding boxes did not fit perfectly tight (see fig. 30(d)).

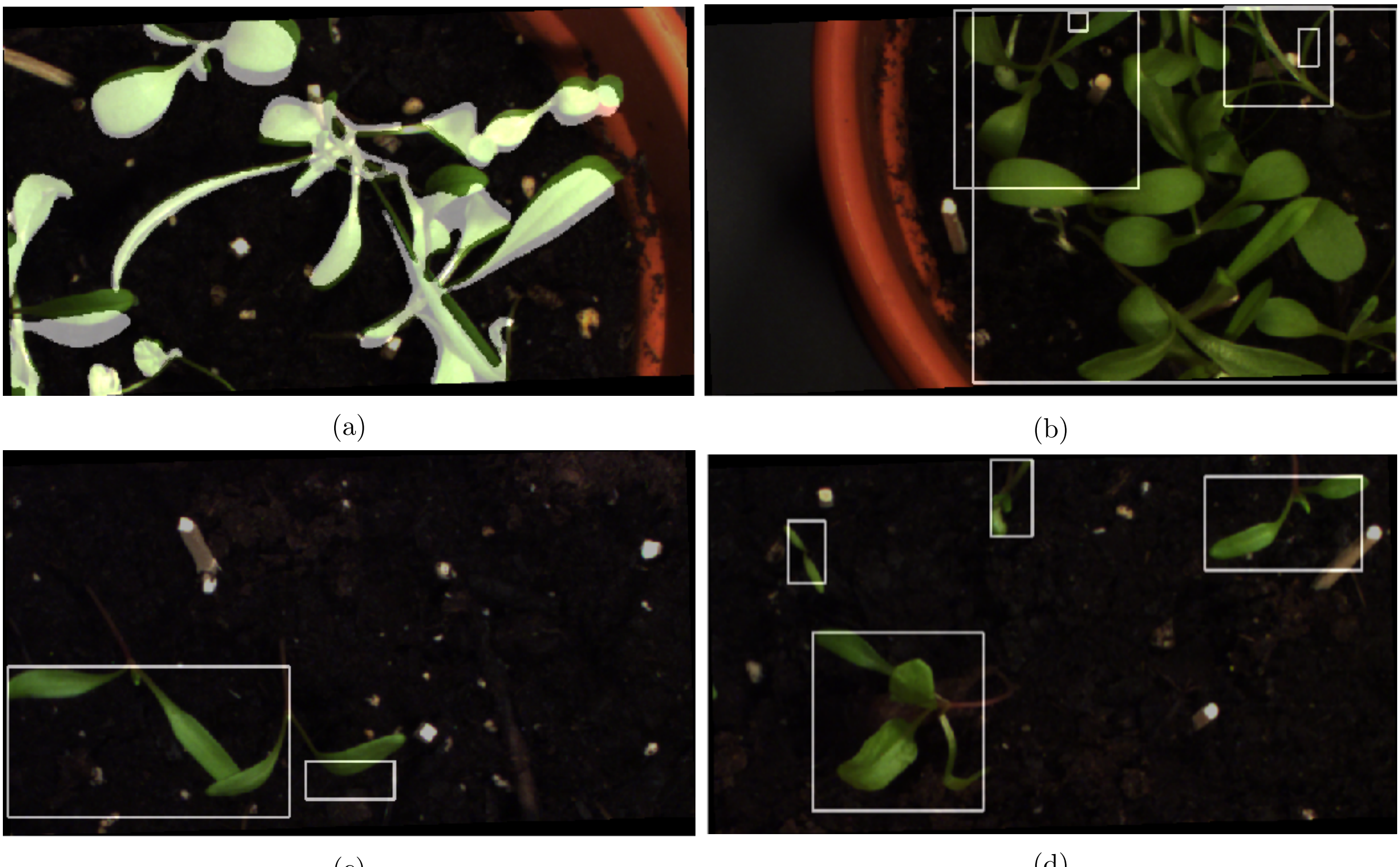


Figure 30: All predictions, pixel or bounding boxes were projected from a multispectral image onto the image plane of a multispectral image, classes (sugar beet or weed) were not taken into account for determining the quality of the projection, Figure a: Warped pixel predictions, b: Bounding boxes for overlapping plants, c: Empty bounding box after warping, d: Succesuful bounding box projection.

5 Discussion

Since the number of plants was small and results of lab experiments are normally not directly transferable to the field, due to different conditions for the plants and the cameras, the results at hand have to be confirmed with field data. Nevertheless, the lab experiments provided important information regarding the choice of methodology, and the developed Python software can be applied on field data.

5.1 Materials and Methods

More stable growing conditions for the plants would have led to more "realistic" plants without the long shoots that partially caused an unnatural angle towards the camera. Furthermore, soil without black gnat larva that attack the roots of the plants is preferable. Even though no symptoms of illness were observed in the plants, the black gnat larva could have had an influence on plant health and therefore the plant spectra. Also, potting soil without white perlites is most likely desirable because the bright perlites might have influenced the automated exposure time adjustment towards a shorter exposure time. This could explain the relatively low reflectance values for the average plant pixels.

The snapshot technology of the used spectral camera had benefits such as no need of movement of the target relative to the camera and fast image acquisition [70]. The biggest drawback was that due to the necessary bandpass filter, the number of wavebands was reduced from 25 possible channels to 15. Still, the for plant classification crucial NIR region [34, p. 669] [14, p. 100] [56] [34, p. 669] [113, p. 95] [54, p. 3] was covered partially and sugar beet and weed spectra were distinguishable in most cases. Regarding the different apertures, a more standardized procedure as it was performed, would have been better. This includes trying different apertures on the same day from the same plants in order to exclude other sources of variance that could influence the results. Such a standardized approach was used for the bandpass filter tests.
The same applies to the CNN classifier that could have been tested further for the two classifier approach, the first distinguishing between soil and plant pixels and the second one between sugar beet and weed with the EMSC and L2 corrected data. This was not carried out due to shortage of time and strong recommendations to use a fully connected network instead by researcher from Fraunhofer IFF.

Some spectral cameras cover the visible and the NIR region which would make the image registration process, that increases the risk of errors, redundant [67, 123]. Further, a 3D camera system could have been used for avoiding parallax effects. Nevertheless, having a pipeline for spatial matching between images of different spectra and sensors increases the independency from specific camera systems and possible associated costs and limitations of wavebands. Regarding the visual assessment of the spatial matching, more than one assessment person would have increased the reliability of the results.

5.2 Analysis and classification of the spectral data

5.2.1 Spectral imaging for plant classification

Acquiring high quality multi- or hyperspectral images of uneven objects is difficult due to varying angles and distances to the sensor and shadows. Therefore, scatter correction is crucial in order to uncover the non-scatter related differences between plant species. Vrindts mitigated spectral variations that occurred even within one plant, due to different angles towards light and sensor, by utilizing waveband ratios instead of scatter corrected wavebands themselves as features [117, p. 67]. This example emphasizes the need for scatter correction for close-range spectral imaging, along with Mishra et al. [77, p. 55], Mohd et al. [78, p. 121] and the results of this thesis. Applying EMSC and L2 normalization uncovered for the separate data set that the curve shapes of average weed and sugar beet absorbance were very different, which was not visible from the spectral data that was only calibrated and transformed to absorbance. For the mixed data set, the absorbance values of sugar beet and weed for aperture f/8 and f/14 became more alike regarding range of values and more unlike regarding curve trajectories after the EMSC and L2 transform, compared to before those two transformations. The more distinct curve shapes after the corrections of the mixed data set are in alignment with the observations of the separate data set and the literature. The reasons for the different range and reverse curve shapes of weed and sugar beet averages of separate and mixed data set might be that only a small, not-representative subset of the mixed data set was used by mainly utilizing the plants that were in the species-segregated pots (see fig. 12). The plants of the mixed dataset that were in the species-segregated pots were not representative regarding the development stage of sugar beet and the weed species composition. Additionally, some weed species could not be distinguished from sugar beet in the pictures and were therefore not included in the training set for the mixed dataset either. Moreover, because the annotation tool used for the mixed pot was rectangular, margins, small leaves and shoots were under-represented. It is not likely that the exposure time was the decisive factor for the differences between separate and mixed data set because for the separate data set, different exposure times did not lead to very distinct patterns for the same plant group, meaning that calibration and pre-processing worked adequately. Since there were only two days between the image acquisition for the separate and the mixed data set, age or plant illness are also unlikely to play a role.

Although hyperspectral imaging is prone to erroneous signals, hyperspectral data contains a lot more (spectral) information than RGB images, while maintaining the spatial information for certain camera types, and has therefore huge potential for plant classification. This was shown, amongst others, by Vrindts [117], who exclusively used spectral data to classify sugar beet, maize, rape and seven weed species with good results (accuracy of 90 %). Despite the season dependency of the Bayesian classifiers, Zhang's et al. work demonstrated that merely hyperspectral data contained enough information to discriminate with high accuracy between crop and weed under field conditions [127]. Okamoto et al. [84] worked with few pixel samples and tried to discriminate between five plant species which increased the chance for wrong predictions. Those might have been the reasons for the comparably low classification accuracy, reinforcing the necessity of many training samples. Notwithstanding, Okamoto's classification accuracy referred to pixels for which misclassifications are common and often reduced by smoothing methods [66, p. 5]. On a plant-level, Okomato's lowest accuracy of 75 % of correctly classified pixels [84, p. 36] might be enough to infer the correct plant species.

5.2.2 Deep learning for plant classification

The results of this thesis showed in accordance with several scienific studies that deep learning methods perform better at classifying background, sugar beet and weed pixels compared to statistical classification methods. This was shown by comparing the performance of the spectral three-classes CNN and PLS-DA on the same dataset. Even though statistical classifiers for plant species can reach a classification accuracy higher than 95 % and are able to deal with some variance, they might not be flexible enough for the amount of variability in plants that occurs under different growing and field conditions. Within some weeks, the phenotype (form) and spectral response of plants changes naturally [21] [63][50, p. 59] [103] [40] [4], even more when the health condition of the plants changes, too [54, p. 3][118]. Additional variables are lighting conditions [14, p. 99], dust, humidity [54, p. 3], changing soil types and different weed species between and within fields. This is supported by the work of Zhang et al. [127] who had to use three distinct Bayesian classifer for spectral data of three different seasons in order to achieve satisfying classification results [127]. The use of several different classifiers for disinct seasons is not a feasible solution for a commercial weeding robot that might work globally. Additionally, many characteristic seasonal changes in the plant phenotype are rather wheather-dependent, and the wheather of one season is not consistent over the years. A deep learning approach could have been able to perform well for all three seasons due to its greater flexibility. Even though Lottes et al. reached okay performance on sugar beet of another growth stage, the authors did not consider it good enough for mechanical removal [66, p. 7]. This highlights the limitations of their training data set and their modeling approach, which relied on circumferences and texture, which might change during plant development. Arakeri et al. strengthened the hypothesis that ANN's have an enormous potential for plant classification, even if it is only trained on RGB color features [7]. The excellent classification performance of 99 % accuracy of Arakeri's net might decrease when challenged with more variation regarding plants' development stage and field, but a loss of few percents would still be usable for plant classification. Furthermore, Feyaerts et al. [34] reached the best classification results with a deep learning approach, even though their ANN only consisted in three hidden layers. However, scatter correction or normalization of the spectral data was not mentioned by Feyearts et al., which might have caused the relatively low classification accuracy of 80 % for sugar beet and 91 % [34] for weed on a plant-level. To distinguish between overlapping plants is especially challenging for computer vision systems. A CNN trained on 17,000 weed annotations could detect weeds with an abundance of overlapping plant parts [28], emphasizing the potential of deep learning for in-field plant detection and the need for high quality and quantity training data.

5.2.2.1 Spectral CNN for pixel classification

The usage of a convolution over the spectral domain worked, but might have not been completely suited for this data set with only 15 channels, as it was developed for more than 100 spectral bands [51]. Also a combination of spectral and spatial features, in this case convolutions, could have been more successful as shown by Luo et al. [71], Gao et al.[39], Chen et al. [22] and Santara et al. [94]. Another reason for the worse results of the CNN compared to the system of two fully connected ANN's is probably that the whole capacity of the a neural network could be used to distinguish between two classes instead of three. It was probably because of that, that all of the discussed scientific papers in the theory part first distinguished between plants and background before classifying between different plant species. Moreover, the ANN had more trainable weights and the training plant pixels for the CNN were not EMSC or L2 transformed. Nevertheless, classification results of the CNN on the pixel-level were mostly good enough for correct classifications on the plant-level. Therefore, a neural network with convolutions over the spectral domain is an approach worth exploring further, particularly with a higher number of channels.

5.2.2.2 Soil versus plant ANN

Since the background and plant spectra are very distinct, as suggested by the data and the literature [72, p. 147], it was not surprising that the soil versus plant classifier performed very well and better than sugar beet versus beet classifier. The misclassification of the margins of plants as the biggest problem makes sense, based on the fact that spectra of neighboring pixels are highly correlated, and each pixel contains information of adjacent pixels [49, p. 650]. This phenomenon might have been further enforced by the bilinear interpolation used for the demosaicing of the spectral images, which infers pixel values based on neighboring pixels. Furthermore, fringed plant edges in the classification results were especially problematic for aperture f/2.8 which can be explained with spherical aberration.

5.2.2.3 Sugar beet versus weed ANN

That the results for all datasets were best for degree(EMCS) = 0 is logical, based on the fact that the average spectra showed the biggest difference in percent between weed and sugar beet for degree(EMSC) = 0, especially for the separate data set. On the plant-level, the results for the different apertures of the separate dataset were quite similar, with classification accuracy for each class between 89 and 95 % for sugar beet and between 87 and 94 % for weed, aligning with the similarity of the average absorbance spectra between apertures. From these spectra and the knowledge about spatial and chromatic aberration that were visible in some masks of the f/2.8 dataset, it was expected that the classification results would not be the best compared to the other apertures. One of the reasons for the aperture f/6 dataset performing best could be that this was data from one acquisition day and not several as for the other aperture sets. Therefore, there might have been less variance. Based on the average absorbance spectra and due to a higher f-number and, therefore, less confusions caused by spherical and chromatic aberrations, aperture f/11 was expected to reach better classification results than aperture f/2.8. This was only true for sugar beet pixels but not for weed pixels. There are no obvious reasons regarding the plants or acquisition method that explain why weed pixels of the aperture f/11 dataset were misclassified more often than for the other apertures of the separate dataset.

Because of the chromatic and spherical aberration effects of different apertures, it is not a clean approach to train and test a classifier on a dataset with images with different apertures, especially with higher differences as in this case (f-numbers between 2.8 - 11). Though, from a pragmatic point of view, it was interesting to test this approach, too. Since especially field data is often subjected to different lighting conditions or speed requirements that might make it desirable to change the aperture, even though it is not recommendable. Further, the average spectra for sugar beet and weed of the different apertures were alike. The classification results for the whole separate data set were in between the results for the different aperture subsets of the separate dataset, which is reasonable.

Based on the overall smaller distance between mean spectra of sugar beet and weed of the mixed data set after EMSC and L2 correction, it was to be expected that classification results would be worse than for the separate data set. The plant-based classifications for aperture 8 were not reliable; the ones for aperture 14 were, if the degree of EMSC correction was 0. The better results for aperture f/14 could be due to better spectral quality because of chromatic aberration effects. Regarding the mixed data set, considering the greater difference between the mean spectra of sugar beet and weed for aperture f/14, it was no surprise that the sugar beet versus weed classifier performed better for the aperture f/14 dataset than for the aperture f/8 data set. It was interesting to see that for aperture f/14, a higher degree for EMSC produced a higher dice coefficient and cleaner results with regards to shadows and angles for sugar beet - as expected, since those lead to scattering - but worse overall classification results on a plant-level. The worse results can not be explained directly by the average spectra since the gap between sugar beet and weed spectra remained similar to d(EMSC) = 0.
One of the reasons for the overall worse classification results for the mixed data set, particularly for the mixed pot, was most likely that the training and test data set were not fully representative with regards to age or rather development stage (sugar beet) and species composition (weed). This hypothesis is supported by the fact that the classification for the test and evaluation data of the mixed dataset with the same age and species distribution was acceptable on a plant-level, but a lot worse for sugar beet of different age and unknown weed species, which were planted mainly in the mixed pot.

With optimization, better classification results might be possible, but 90 % of correctly classified pixels per plant should be enough to determine the right species with a majority voting principle. Further, misclassification of some pixels are common, as already pointed out [66, p. 5]. All in all, it can be concluded that the combination of information-rich hyperspectral images and deep learning is very potent.

5.2.3 Features for the classification of plant species

There is evidence that successful classifications for weed and crop can also be performed only using RGB channels. For instance, Mao et al. [73] (wheat) and Arakeri et al. (onions) [7] used solely color features of RGB images and reached around 90 % and 99 % correct classification respectively. On the other hand, Telleache et al. stated that the spectral information of the RGB camera did not provide enough distinctive information to differentiate between corn and weed [108, p. 523]. Astrand (sugar beet) [8], Dos Santos Ferreira (soybean) [26], Gerhards et al. [42] and Lottes et al. [66](sugar beet) used both, spectral and spatial features. Even though shape features might be useful for discriminating between grass-like species and leafy-species, the differentiation between a grass-like crop, like wheat, and weed species, like foxtail, is more difficult with only shape-features. Moreover, shape features are not very reliable over time and over many species since the shape of the same leave changes over time [127, p. 66] or different leaves at different parts of the plant have different forms. Also, shapes are very susceptible to mechanical lesions [127, p. 66] and angles of view. This is why Zhang et al. refrained completely from using spatial features. An argument supporting shape features is that, normally, one would try to remove weed at a specific growth stage, but this constraint

should not limit a mechanical weeding robot.
On the other hand, spectral data has also limitations with regards to discriminative power, because the plant's spectrum changes with age, health status and moisture, as already discussed. The limitations of spectral data could be observed in the data set on hand, too, since lambsquarters, a weed species that has a very similar spectral response to sugar beet [9, p. 2] [104, p. 6] [113, p. 95], was misclassified very often as sugar beet. Another reason could have been that there were very few pixels available from lambsquarters because the plants had just germinated when they were photographed. Okamoto et al. have also observed that some weed species are harder to distinguish from sugar beet than others [84, p. 36].

As pointed out, both, spectral and spatial features have their benefits and limitations when it comes to their distinctive power for weed versus crop classification. Based on that, the combination of both feature types seems to be the best solution. Likewise, an expansion by other features such as the Fourier transform of the spectral or spatial data could be valuable, as demonstrated by Nejati et al. [82]. Dos Santos Ferreira et al. showed [26] that combining a variety of features, like shape, texture and color, with a CNN, can lead to an accuracy of 99 %.
There were little other approaches for automated labeling, but those that existed all exploited spatial features and crop lines for determining the weed and sugar beet training datasets and ground truth [67, 123, 38].

5.3 Image registration

For image registration of close-range images of uneven objects with different heights like plants, a 3D camera system would be ideal for optimal alignment results. To avoid 3D calibration of the cameras and achieve image alignment results that are exact on the plant level, a local motion approach like imregdemons from Matlab can be a good solution. However, the imregdemon algorithm failed for the presented data set, most likely due to the different resolutions and the bigger differences in reflectance values between RGB and spectral images. Probably, these differences were also the reason that Lowe's distance ratio was not useful for this data set. Another reason for imregdemon's failure could be that matching areas did not overlap in the beginning, which was mentioned in Thirion [111, p. 247] as a pre-requisite. This hypothesis could not be confirmed since a more detailed description of Matlab's imregdemon algorithm is not publicly available.

One reason for the 740 nm waveband providing the best results for image registration with RGBs could be that the leaves showed high reflectance values for this waveband, which was also the case for the grayscale RGB images with a big contrbution of the green waveband. Since this was also true for all wavebands bigger than 740 nm, there might have been other spectral and spatial quality influences leading to higher contrast or more similar reflectance values to the RGB images.

The developed filtering system for the homographies based on several quality assessment parameters has proven to choose the suitable homographies for one session automatically. Only one parameter, the pixel threshold for the spatial filtering, might have to be re-adjusted for other relative camera settings for optimal spatial filtering. Still, as long as the pixel threshold is not too small, the other filtering steps should compensate for a too

high threshold.
The visual assessment grades for the filtering system, resulting in 16 images being not suited for bounding box projections, aligned with the results of the final bounding box evaluation, that found 18 images with empty bounding boxes. By combining several homographies, the results became more stable and it was possible to perform image registration for image pairs with too few matches, which occurred especially for the images of the tiny lambsquarters plants. This was only possible because the position of the two cameras relative to each other did not change for the same acquisition session.
Due to the failure of the reflectance-based image matching approach and the lack of a 3D calibration of the camera systems, ghosting and parallax caused inaccuracies could not be prevented or mitigated entirely. This happened because a 2D motion model for planar surfaces was applied to very close targets of different heights. Nevertheless, the height difference of the plants was only a few centimeters, which is why for most images and plants, the projection of bounding boxes worked. More thorough testing of the system regarding the number of satisfying matches and limits would be worthwhile. An interesting approach would be to combine the homography-based image registration for finding the coarse overlapping areas as the first step, with a displacement field approach for finer adjustments and removal of parallax and ghosting as a second step. Also, more preprocessing techniques regarding reflectance values could be explored in order to facilitate a displacement field calculation. The projected bounding boxes can be easily transformed into the coco-format or another image labeling format. By comparing automated labeled data with manually labeled data, Wendel et al. [123] also showed that automated labeling of crops was possible and can reach good classification results, but was not as good as manual labeling.

The problem with completely overlapping plants of the same class being perceived as one object can hardly be solved with traditional image analysis techniques. One possibility to automatically create a labeled training data set with overlapping plants would be to crop out segmented and already classified plant individuals of images, and place them in an overlapping manner into an artificial image.
There was no optimization performed regarding computing time or memory usage of the code because the system does not have to operate in real-time since it only serves the purpose to create a training data set automatically. Nevertheless, these are points of improvement.

6 Conclusions

The overall goal of this thesis was to develop a system for automatically labeling RGB images based on hyperspectral imaging of sugar beet and weed plants. This task had to main parts: First, the detection and classification of sugar beet and weed plants in the hyperspectral images. Second, the transfer of this information onto the RGB image.
This thesis showed that spectral data from only 15 wavebands can be sufficient for labeling weed and sugar beet plants safely under lab conditions. Since a high-quality training data set for any type of classifier is crucial for the future classification success, spatial features could be added to increase the reliability and robustness of the classifiers. This is especially important for spectral imaging data that was acquired under field conditions with a higher amount of erroneous signals.
Based on the literature and the comparison between a CNN and PLS-DA analysis, deep

learning methods seem to be best suited for distinguishing between plant species because they can capture the large variability in plants and fields.
For the utilized spectral snapshot camera, the bandpass filter 600 - 875 nm delivered the best results. Because of spherical and chromatic aberration effects, a high f-number is recommendable. However, it is not easy to give concrete recommendations for the used camera since aperture f/6 led to the best results for the separate data set, but for the mixed data set, the highest f-number, 14, performed best. Further investigations, especially under field conditions are necessary in order to determine the best aperture settings.

It was also demonstrated that it is possible to transfer bounding boxes based on the hyperspectral image onto the RGB image, even with a 2D motion model applied on close-range images with 3D objects. Nevertheless, the method is not directly suited for this task, and therefore it does not work in all cases and a general loss in accuracy was observed. Further filtering of the homographies combined with the use of a local motion model, or a 3D calibrated camera system could solve the existing problems.

This thesis has shown that reliable, automated labeling of RGB images with spectral imaging is viable, and within the scope of the thesis, a prototype of such a system was developed with Python. The labeling system can be used for future images, even though this prototype should be developed further in order to increase the performance, robustness and user-friendliness. Especially, a quality check for empty bounding boxes should be added. Furthermore, the problem with overlapping plants of the same species that are perceived as one plant should be tackled by, e.g., artificially producing images of overlapping, but as individuals labeled plants. Also, the neural network has to be retrained on the field data because of the differences to lab conditions.
Computer vision with cheap RGB cameras and deep learning has a huge potential for agriculture, and other areas. For instance, autonomous weeding robots with good computer vision systems can contribute to a more sustainable agriculture by reducing costs, time and the amount of herbicides used while securing yields. Currently, manual image labeling is the bottleneck for the development and application of computer vision systems for such weeding robots. Therefore, further research into automated labeling is desirable. For such automated labeling systems, hyperspectral imaging can play a big role because spectral data alone contains enough information to distinguish between plant species, or other types of objects.

Bibliography

[1] URL: https://www.princeton.edu/~cuff/ele201/kulkarni_text/frequency.pdf (visited on 02/02/2020).

[2] URL: https://docs.opencv.org/2.4/modules/imgproc/doc/miscellaneous_transformations.html (visited on 02/02/2020).

[3] Waleed Abdulla. *Splash of Color: Instance Segmentation with Mask R-CNN and TensorFlow.* URL: https://engineering.matterport.com/splash-of-color-instance-segmentation-with-mask-r-cnn-and-tensorflow-7c761e238b46 (visited on 12/03/2019).

[4] M. L. Adams et al. "Fluorescence and reflectance characteristics of manganese deficient soybean leaves: effects of leaf age and choice of leaflet". In: *Plant Nutrition — from Genetic Engineering to Field Practice.* Ed. by N. J. Barrow. Vol. 44. Dordrecht: Springer Netherlands, 1993, pp. 261–264. ISBN: 978-94-010-4832-3. DOI: 10.1007/978-94-011-1880-451.

[5] Wikipedia user Andreas 06. *Chromatic aberration.* 2006. URL: https://de.m.wikipedia.org/wiki/Datei:Chromatic_aberration_convex.svg (visited on 05/24/2020).

[6] Wikipedia user Andrei Stroe. *Spherical aberration.* 2008. URL: https://commons.wikimedia.org/wiki/File:Spherical_aberration_2.svg (visited on 05/24/2020).

[7] Megha. P. Arakeri et al. "Computer vision based robotic weed control system for precision agriculture". In: *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI).* IEEE, 2017, pp. 1201–1205. ISBN: 978-1-5090-6367-3. DOI: 10.1109/ICACCI.2017.8126005.

[8] Björn Åstrand and Albert-Jan Baerveldt. "A mobile robot for mechanical weed control". In: *International Sugar Journal* 105.1250 (2003), pp. 89–95.

[9] M. Bah, Adel Hafiane, and Raphael Canals. "Deep Learning with Unsupervised Data Labeling for Weed Detection in Line Crops in UAV Images". In: *Remote Sensing* 10.11 (2018), p. 1690. DOI: 10.3390/rs10111690.

[10] Adel Bakhshipour and Abdolabbas Jafari. "Evaluation of support vector machine and artificial neural networks in weed detection using shape features". In: *Computers and Electronics in Agriculture* 145 (2018), pp. 153–160. ISSN: 01681699. DOI: 10.1016/j.compag.2017.12.032.

[11] Pierre Barré et al. "LeafNet: A computer vision system for automatic plant species identification". In: *Ecological Informatics* 40 (2017), pp. 50–56. ISSN: 1574-9541.

[12] B. E. Bayer. "Color imaging array". US3971065A. 1976.

[13] Hugh J. Beckie. "Herbicide-resistant weed management: focus on glyphosate". In: *Pest management science* 67.9 (2011), pp. 1037–1048. DOI: 10.1002/ps.2195.

[14] Jan Behmann, Jörg Steinrücken, and Lutz Plümer. "Detection of early plant stress responses in hyperspectral images". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 93 (2014), pp. 98–111. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2014.03.016.

[15] Enrico Biancardi et al. "Sugar Beet". In: *Root and Tuber Crops*. Ed. by J.E Bradshaw. Vol. 94. New York, NY: Springer New York, 2010, pp. 173–219. ISBN: 978-0-387-92764-0. DOI: 10.1007/978-0-387-92765-76.

[16] Hans-Peter Blume et al. *Scheffer/Schachtschabel: Lehrbuch der Bodenkunde*. 16th ed. 2010. Berlin, Heidelberg: Springer Berlin Heidelberg, Imprint, and Springer Spektrum, 2010. ISBN: 9783662499603.

[17] R. Bongiovanni and J. Lowenberg-Deboer. "Precision Agriculture and Sustainability". In: *Precision Agriculture* 5.4 (2004), pp. 359–387. ISSN: 1385-2256. DOI: 10.1023/B:PRAG.0000040806.39604.aa.

[18] Gary R. Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. 1. ed., [Nachdr.] Software that sees. Beijing: O'Reilly, 2011. ISBN: 978-0-596-51613-0.

[19] Ralph B. Brown and Scott D. Noble. "Site-specific weed management: sensing requirements—what do we need to see?" In: *Weed Science* 53.2 (2005), pp. 252–258. ISSN: 0043-1745.

[20] Wilhelm Burger and Mark James Burge. *Digital image processing: An algorithmic introduction using Java*. Second Edition. Texts in computer science. London: Springer, 2016. ISBN: 978-1-4471-6683-2.

[21] G. A. CARTER et al. "Effect of competition and leaf age on visible and infrared reflectance in pine foliage". In: *Plant, Cell and Environment* 12.3 (1989), pp. 309–315. ISSN: 0140-7791. DOI: 10.1111/j.1365-3040.1989.tb01945.x.

[22] Yushi Chen et al. "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks". In: *IEEE Transactions on Geoscience and Remote Sensing* 54.10 (2016), pp. 6232–6251. ISSN: 0196-2892. DOI: 10.1109/TGRS.2016.2584107.

[23] Continental. *Back to Business: Dank dem neuen Agrarroboter von Continental können Landwirte sich wieder auf das Wesentliche konzentrieren*. 2019. URL: https://www.continental.com/de/presse/messen-events/agritechnica-2019/agrarroboter-contadino-197202.

[24] Michael J. Crawley. *The R book*. 2. ed. Chichester: Wiley, 2013. ISBN: 9781118448922. DOI: 10.1002/9781118448908. URL: %5Curl%7Bhttp://lib.myilibrary.com?id=450278%7D.

[25] Rainer Dohlus. *Technische Optik*. De Gruyter Studium. Berlin: De Gruyter, 2015. ISBN: 9783110351439. URL: http://gbv.eblib.com/patron/FullRecord.aspx?p=4006804.

[26] Alessandro dos Santos Ferreira et al. "Weed detection in soybean crops using ConvNets". In: *Computers and Electronics in Agriculture* 143 (2017), pp. 314–324. ISSN: 01681699. DOI: 10.1016/j.compag.2017.10.027.

[27] Timothy Dozat. “Incorporating nesterov momentum into adam”. In: (2016).

[28] M. Dyrmann, R. N. Jørgensen, and H. S. Midtiby. “RoboWeedSupport - Detection of weed locations in leaf occluded cereal crops using a fully convolutional neural network”. In: *Advances in Animal Biosciences* 8.2 (2017), pp. 842–847. ISSN: 2040-4700. DOI: 10.1017/S2040470017000206.

[29] CHRISTOPHER D. ELVIDGE. “Visible and near infrared reflectance characteristics of dry plant materials”. In: *International Journal of Remote Sensing* 11.10 (1990), pp. 1775–1795. ISSN: 0143-1161. DOI: 10.1080/01431169008955129.

[30] *Fachstufe Landwirt.* 9., überarb. Aufl., Neuausg. 2012. ISBN: 978-3-8354-0526-4.

[31] FAO. *How to Feed the World in 2050.* 2009.

[32] FAO. *WORLD FOOD AND AGRICULTURE - STATISTICAL POCKETBOOK 2019.* [S.l.]: FOOD & AGRICULTURE ORG, 2019. ISBN: 978-92-5-131849-2.

[33] farmdroid. *Willkommen bei Farmdroid.* unkown. URL: http://farmdroid.dk/de/willkommen/ (visited on 02/05/2020).

[34] Filip Feyaerts and Luc van Gool. “Multi-spectral vision system for weed detection”. In: *Pattern Recognition Letters* 22.6-7 (2001), pp. 667–674. ISSN: 0167-8655.

[35] C. Fischer. *Infoblatt Zuckerrübe.* Leipzig, 2003. URL: https://www.klett.de/alias/1010375 (visited on 02/05/2020).

[36] Fraunhofer Institut. *Fraunhofer-Leitprojekt ≫Cognitive Agriculture≪: Schaffung eines umfassenden informationsbasierten Ökosystems für den Agrarbereich.* 2018. URL: https://www.iese.fraunhofer.de/de/innovation_trends/SmartFarming/cognitive-agriculture.html (visited on 02/05/2020).

[37] Sabrina Gaba et al. “Herbicides do not ensure for higher wheat yield, but eliminate rare plant species”. In: *Scientific reports* 6 (2016), p. 30112. DOI: 10.1038/srep30112.

[38] Junfeng Gao et al. “Fusion of pixel and object-based features for weed mapping using unmanned aerial vehicle imagery”. In: *International Journal of Applied Earth Observation and Geoinformation* 67 (2018), pp. 43–53. ISSN: 03032434. DOI: 10.1016/j.jag.2017.12.012.

[39] Qishuo Gao, Samsung Lim, and Xiuping Jia. “Hyperspectral Image Classification Using Convolutional Neural Networks and Multiple Feature Learning”. In: *Remote Sensing* 10.2 (2018), p. 299. DOI: 10.3390/rs10020299.

[40] H. W. Gausman et al. “Age Effects of Cotton Leaves on Light Reflectance, Transmittance, and Absorptance and on Water Content and Thickness 1”. In: *Agronomy Journal* 63.3 (1971), pp. 465–469. ISSN: 00021962. DOI: 10.2134/agronj1971.00021962006300030035x.

[41] R. Gerhards and Svend Christensen. “Real–time weed detection, decision making and patch spraying in maize, sugarbeet, winter wheat and winter barley”. In: *Weed research* 43.6 (2003), pp. 385–392. ISSN: 0043-1737.

[42] R. Gerhards, H. Oebel, and M. Sökefeld. *Teilschlagspezifische Unkrautbekämpfung durch raumbezogene Bildverarbeitung im Offline- (und Online-) Verfahren (TURBO): Anschlussbericht für das Forschungs- und Entiwcklungdvorhaben.* Ed. by Universität Hohenheim, Institut für Phytomedizin, Fachgebiet Herbologie. Hohenheim, 2007.

[43] Leonard P. Gianessi. "The increasing importance of herbicides in worldwide crop production". In: *Pest management science* 69.10 (2013), pp. 1099–1105. DOI: \url{10.1002/ps.3598}.

[44] Leonard P. Gianessi and Nathan P. Reigner. "The Value of Herbicides in U.S. Crop Production". In: *Weed Technology* 21.2 (2007), pp. 559–566. ISSN: 0890-037X. DOI: 10.1614/WT-06-130.1.

[45] Driss Haboudane et al. "Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture". In: *Remote Sensing of Environment* 81.2-3 (2002), pp. 416–426. ISSN: 00344257. DOI: 10.1016/S0034-4257(02)00018-4.

[46] Nathan Hagen and Michael W. Kudenov. "Review of snapshot spectral imaging technologies". In: *Optical Engineering* 52.9 (2013), p. 090901. ISSN: 0091-3286. DOI: 10.1117/1.OE.52.9.090901.

[47] Karl D. Hansen et al. "An autonomous robotic system for mapping weeds in fields". In: *IFAC Intelligent Autonomous Vehicles Symposium.* Ed. by The International Federation of Automatic Control. 2013, pp. 217–224. ISBN: 978-3-902823-36-6.

[48] Ian Heap. "Global perspective of herbicide-resistant weeds". In: *Pest management science* 70.9 (2014), pp. 1306–1315. DOI: 10.1002/ps.3696.

[49] I. Herrmann et al. "Ground-level hyperspectral imagery for detecting weeds in wheat fields". In: *Precision Agriculture* 14.6 (2013), pp. 637–659. ISSN: 1385-2256. DOI: 10.1007/s11119-013-9321-x.

[50] X. U.E. Li-hong et al. "Canopy spectral reflectance characteristics of rice with different cultural practices and their fuzzy cluster analysis". In: *Rice Science* 12.1 (2005), pp. 57–62. ISSN: 1672-6308.

[51] Wei Hu et al. "Deep Convolutional Neural Networks for Hyperspectral Image Classification". In: *Journal of Sensors* 2015.2 (2015), pp. 1–12. ISSN: 1687-725X. DOI: 10.1155/2015/258619.

[52] T. Huang. *Computer Vision: Evolution And Promise.* 1996. DOI: 10.5170/CERN-1996-008.21.

[53] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.* URL: http://arxiv.org/pdf/1502.03167v3.

[54] Jacquemoud, Stephane, and Susan L. Ustin. "Leaf optical properties: A state of the art". In: *8th International Symposium of Physical Measurements & Signatures in Remote Sensing.* 2001.

[55] Abdolabbas Jafari et al. "Weed detection in sugar beet fields using machine vision". In: *Int. J. Agric. Biol* 8.5 (2006), pp. 602–605.

[56] John R. Jensen. *Remote sensing of the environment: An earth resource perspective.* Pearson Education India, 2009. ISBN: 8131716805.

[57] Jichao Jiao et al. "A structural similarity-inspired performance assessment model for multisensor image registration algorithms". In: *International Journal of Advanced Robotic Systems* 14.4 (2017), p. 172988141771705. ISSN: 1729-8814. DOI: 10.1177/1729881417717059.

[58] Wannes Keulemans, Dany Bylemans, and Barbara de Coninck. *Farming without plant protection products: Can we grow without using herbicides, fungicides and insecticides?* Luxembourg: Publications Office of the European Union, 2019. ISBN: 978-92-846-3993-9.

[59] Wikipedia user KoeppiK. *Lenses with different apertures.* 2019. URL: `https://de.m.wikipedia.org/wiki/Datei:Lenses_with_different_apertures.jpg` (visited on 05/24/2020).

[60] Erwin Ladewig et al. "Pflanzenschutz im Zuckerrübenanbau in Deutschland–Situationsanalyse 2018". In: *Sugar Industry* 143.12 (2018), pp. 708–722.

[61] LeCun, Yann and Boser, Bernhard E and Denker, John S and Henderson, Donnie and Howard, Richard E and Hubbard, Wayne E and Jackel, Lawrence D. "Handwritten digit recognition with a back-propagation network". In: ().

[62] M. Lehmann, S. Socher, and R. Schwierz. *Optische Abbildungen: Physikalisches Grundpraktikum Technische Universität Dresden.* 2016.

[63] Fei Li et al. "Evaluating hyperspectral vegetation indices for estimating nitrogen concentration of winter wheat at different growth stages". In: *Precision Agriculture* 11.4 (2010), pp. 335–357. ISSN: 1385-2256. DOI: `10.1007/s11119-010-9165-6`.

[64] F. LÓPEZ-GRANADOS. "Weed detection for site-specific weed management: mapping and real-time approaches". In: *Weed research* 51.1 (2011), pp. 1–11. ISSN: 0043-1737. DOI: `10.1111/j.1365-3180.2010.00829.x`.

[65] Avraham Lorber, Lawrence E. Wangen, and Bruce R. Kowalski. "A theoretical foundation for the PLS algorithm". In: *Journal of Chemometrics* 1.1 (1987), pp. 19–31. ISSN: 0886-9383. DOI: `10.1002/cem.1180010105`.

[66] Philipp Lottes et al. ""An effective classification system for separating sugar beets and weeds for precision farming applications."" In: *IEEE International Conference on Robotics and Automation (ICRA).* 2016, pp. 5157–5163. URL: `http://www.ipb.uni-bonn.de/wp-content/papercite-data/pdf/lottes16icra.pdf` (visited on 01/13/2020).

[67] Marine Louargant et al. "Unsupervised classification algorithm for early weed detection in row-crops by combining spatial and spectral information". In: *Remote Sensing* 10.5 (2018), p. 761.

[68] D. G. Lowe, ed. *Object recognition from local scale-invariant features'.* 2. 1999. URL: `https://www.cs.ubc.ca/~lowe/papers/iccv99.pdf`.

[69] David G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2 (2004), pp. 91–110. ISSN: 0920-5691. DOI: `10.1023/B:VISI.0000029664.99615.94`.

[70] Guolan Lu and Baowei Fei. "Medical hyperspectral imaging: a review". In: *Journal of biomedical optics* 19.1 (2014), p. 10901. DOI: `10.1117/1.JBO.19.1.010901`.

[71] Yanan Luo et al. *HSI-CNN: A Novel Convolution Neural Network for Hyperspectral Image.* URL: `http://arxiv.org/pdf/1802.10478v1`.

[72] Dimitris G. Manolakis, Ronald B. Lockwood, and Thomas W. Cooley. *Hyperspectral imaging remote sensing: Physics, sensors, and algorithms.* Cambridge: Cambridge University Press, 2016. ISBN: 9781316017876. DOI: `10.1017/CBO9781316017876`.

[73] Wenhua Mao, Xiaoan Hu, and Xiaochao Zhang. "Weed detection based on the optimized segmentation line of crop and weed". In: *International Conference on Computer and Computing Technologies in Agriculture*. Springer. 2007, pp. 959–967.

[74] Martin A. Fischler und Robert C. Bolles. *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography.* 1980.

[75] Krystian Mikolajczyk and Cordelia Schmid. "Performance evaluation of local descriptors". In: *Electronics Letters* 27.10 (2005), pp. 1615–1630. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2005.188.

[76] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. "Real-time blob-wise sugar beets vs weeds classification for monitoring fields using convolutional neural networks". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 4 (2017), p. 41. ISSN: 2194-9042.

[77] Puneet Mishra et al. "Close range hyperspectral imaging of plants: A review". In: *Biosystems Engineering* 164 (2017), pp. 49–67. ISSN: 15375110. DOI: 10.1016/j.biosystemseng.2017.09.009.

[78] Mohd Shahrimie Mohd Asaari et al. "Close-range hyperspectral image analysis for the early detection of stress responses in individual plants in a high-throughput phenotyping platform". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 138 (2018), pp. 121–138. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2018.02.003.

[79] B. Möller, R. Garcia, and S. Posch. "Towards objective quality assessment of image registration results." In: *VISAPP* (2007), pp. 233–242.

[80] D. Moshou, H. Ramon, and J. de Baerdemaeker. "A weed species spectral detector based on neural networks". In: *Precision Agriculture* 3.3 (2002), pp. 209–223. ISSN: 1385-2256.

[81] Roni A. Neff et al. "Peak oil, food systems, and public health". In: *American journal of public health* 101.9 (2011), pp. 1587–1597. DOI: 10.2105/AJPH.2011.300123.

[82] Hossein Nejati, Zohreh Azimifar, and Mohsen Zamani. "Using fast fourier transform for weed detection in corn fields". In: *2008 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2008, pp. 1215–1219. ISBN: 978-1-4244-2383-5. DOI: 10.1109/ICSMC.2008.4811448.

[83] E.-C. OERKE. "Crop losses to pests". In: *The Journal of Agricultural Science* 144.1 (2006), pp. 31–43. ISSN: 0021-8596. DOI: 10.1017/S0021859605005708.

[84] H. Okamoto et al. "Plant classification for weed detection using hyperspectral imaging with wavelet analysis". In: *Weed Biology and Management* 7.1 (2007), pp. 31–37. ISSN: 1444-6162. DOI: 10.1111/j.1445-6664.2006.00234.x.

[85] Edmund Optics. unknown. URL: https://www.edmundoptics.de/knowledge-center/application-notes/imaging/lens-iris-aperture-setting/ (visited on 05/27/2020).

[86] J. L. Pech-Pacheco et al. "Diatom autofocusing in brightfield microscopy: a comparative study". In: *15th International Conference on Pattern Recognition.* Ed. by Alberto Sanfeliu. Los Alamitos, Calif: IEEE Computer Society Press, 2000, pp. 314–317. ISBN: 0-7695-0750-6. DOI: 10.1109/ICPR.2000.903548.

[87] Gerassimos G. Peteinatos et al. "Potential use of ground-based sensor technologies for weed detection". In: *Pest management science* 70.2 (2014), pp. 190–199. DOI: 10.1002/ps.3677.

[88] Photonfocus. *MV1-D2048x1088-HS02-96-G2.* https://www.photonfocus.com/de/produkte/kame d2048x1088-hs02-96-g2. unknown. (Visited on 05/24/2020).

[89] Photonfocus AG. *Hypesprectal Imaging.* Ed. by Photonfocus AG. unknown.

[90] proplanta. *Zuckerrüben: Pflanzenschutz.* unknown. URL: https://www.proplanta.de/Zuckerruebe/Pflanzenschutz-Pflanzenbauliche-Basisinformationen-Zuckerruebe_Pflanze1155217045.html (visited on 01/28/2020).

[91] Sebastian Raschka and Vahid Mirjalili. *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow.* Second edition, fourth release,[fully revised and updated]. Expert insight. Birmingham and Mumbai: Packt Publishing, 2018. ISBN: 9781787125933.

[92] Adrian Rosebrock. *Simple neural network.* unknown. URL: https://pyimagesearch.com/wp-content/uploads/2016/08/simple_neural_network_header.jpg (visited on 05/24/2020).

[93] Francisco Sánchez-Bayo and Kris A.G. Wyckhuys. "Worldwide decline of the entomofauna: A review of its drivers". In: *Biological Conservation* 232 (2019), pp. 8–27. ISSN: 00063207. DOI: 10.1016/j.biocon.2019.01.020.

[94] Anirban Santara et al. "BASS Net: Band-Adaptive Spectral-Spatial Feature Learning Neural Network for Hyperspectral Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 55.9 (2017), pp. 5293–5301. ISSN: 0196-2892. DOI: 10.1109/TGRS.2017.2705073.

[95] F. Schmidt. *Münchener Studie bestätigt starkes Insektensterben in Deutschland: Ehrenamtliche Insektenkundler hatten vor zwei Jahren Alarm geschlagen: Die Zahl der Fluginsekten sei drastisch eingebrochen. Nun bestätigt eine neue Studie zu drei deutschen Naturregionen die Befürchtungen.* Ed. by Deutsche Welle. 2019. URL: https://www.dw.com/de/m%5C%C3%5C%BCnchener-studie-best%5C%C3%5C%A4tigt-starkes-insektensterben-in-deutschland/a-51051311 (visited on 02/05/2020).

[96] D. Scott and Ralph B. Brown. "The use of spectral properties for weed detection and identification-a review". In: *AIC 2002 - Science: process or product?* Ed. by CSAE/SCGR. Citeseer, 2002.

[97] Sebastian Seibold et al. "Arthropod decline in grasslands and forests is associated with landscape-level drivers". In: *Nature* 574.7780 (2019), pp. 671–674. DOI: 10.1038/s41586-019-1684-3.

[98] R. Shrestha et al. "Quality evaluation in spectral imaging – Quality factors and metrics." In: *Journal of the International Colour Association* 10 (2014), pp. 22–35. URL: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/197995.

[99] D. C. Slaughter, D. K. Giles, and D. Downey. “Autonomous robotic weed control systems: A review”. In: *Computers and Electronics in Agriculture* 61.1 (2008), pp. 63–78. ISSN: 01681699. DOI: 10.1016/j.compag.2007.05.008.

[100] Milan Sonka, Václav Hlaváč, and Roger Boyle. *Image processing, analysis, and machine vision.* 3. ed. Toronto: Thomson Learning, 2008. ISBN: 978-0-495-08252-1.

[101] S. Spycher and A. Bosshard. “Evaluation von Massnahmen in der Land-wirtschaft zur Reduktion der Gewässerbe-lastung mit Pflanzenschutzmitteln”. In: (2015).

[102] Petter Stefansson. “Hyperspectral imaging: Algorithmic advances in variable selection and application to wood science”. Dissertation. As: Norwegian University of Life Sciences, 2019.

[103] Daniela Stroppiana et al. “Plant nitrogen concentration in paddy rice from field canopy hyperspectral radiometry”. In: *Field Crops Research* 111.1-2 (2009), pp. 119–129. ISSN: 03784290. DOI: 10.1016/j.fcr.2008.11.004.

[104] Yumiko Suzuki, HIROSHI OKAMOTO, and TAKASHI KATAOKA. “Image segmentation between crop and weed using hyperspectral imaging for weed detection in soybean field”. In: *Environmental Control in Biology* 46.3 (2008), pp. 163–173. ISSN: 1880-554X.

[105] Richard Szeliski. *Computer vision: Algorithms and applications.* Texts in computer science. London: Springer, 2011. ISBN: 9781848829343. DOI: 10.1007/978-1-84882-935-0. URL: http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10421311.

[106] Richard Szeliski. “Image Alignment and Stitching: A Tutorial”. In: *Foundations and Trends® in Computer Graphics and Vision* 2.1 (2006), pp. 1–104. ISSN: 1572-2740. DOI: 10.1561/0600000009.

[107] Naio Technologies. *AUTONOMOUS VEGETABLE WEEDING ROBOT - DINO: The ideal robot for vegetable farms.* URL: https://www.naio-technologies.com/en/agricultural-equipment/large-scale-vegetable-weeding-robot (visited on 02/06/2020).

[108] Alberto Tellaeche et al. “A computer vision approach for weeds identification through Support Vector Machines”. In: *Applied Soft Computing* 11.1 (2011), pp. 908–915. ISSN: 15684946. DOI: 10.1016/j.asoc.2010.01.011.

[109] Alberto Tellaeche et al. “A vision-based method for weeds identification through the Bayesian decision theory”. In: *Pattern Recognition* 41.2 (2008), pp. 521–530. ISSN: 00313203. DOI: 10.1016/j.patcog.2007.07.007.

[110] *The future of food and agriculture: Trends and challenges.* Rome: Food and Agriculture Organization of the United Nations, 2017. ISBN: 978-92-5-109551-5.

[111] J.-P. Thirion. “Image matching as a diffusion process: an analogy with Maxwell’s demons”. In: *Medical Image Analysis* 2.3 (1998), pp. 243–260. ISSN: 13618415. DOI: 10.1016/S1361-8415(98)80022-4.

[112] Jordan R. Ubbens and Ian Stavness. “Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks”. In: *Frontiers in plant science* 8 (2017), p. 1190. ISSN: 1664-462X.

[113] Saleem Ullah et al. “Identifying plant species using mid-wave infrared (2.5–6mm) and thermal infrared (8–14mm) emissivity spectra”. In: *Remote Sensing of Environment* 118 (2012), pp. 95–102. ISSN: 00344257. DOI: 10.1016/j.rse.2011.11.008.

[114] Umweltbundsamt. *Pflanzenschutzmittelverwendung in der Landwirtschaft.* 2019. URL: https://www.umweltbundesamt.de/daten/land-forstwirtschaft/pflanzenschutzmittelverwendung-in-der#textpart-1 (visited on 01/28/2020).

[115] Unknown. *Zuckerherstellung: Herstellung-Anbau, Pflanzenschutz und Ernte.* 2019. URL: https://www.landschafftleben.at/lebensmittel/zucker/herstellung/anbau-pflanzenschutz-und-ernte (visited on 01/28/2020).

[116] Anup Vibhute and Shrikant K. Bodhe. “Applications of image processing in agriculture: a survey”. In: *International Journal of Computer Applications* 52.2 (2012). ISSN: 0975-8887.

[117] E. Vrindts, J. de Baerdemaeker, and H. Ramon. “Weed detection using canopy reflection”. In: *Precision Agriculture* 3.1 (2002), pp. 63–80. ISSN: 1385-2256.

[118] Mirwaes Wahabzada et al. “Plant Phenotyping using Probabilistic Topic Models: Uncovering the Hyperspectral Language of Plants”. In: *Scientific reports* 6 (2016), p. 22482. DOI: 10.1038/srep22482.

[119] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 13.4 (2004), pp. 600–612. ISSN: 1057-7149. DOI: 10.1109/tip.2003.819861.

[120] Ukrit Watchareeruetai et al. “Computer Vision Based Methods for Detecting Weeds in Lawns”. In: *2006 IEEE Conference on Cybernetics and Intelligent Systems.* IEEE, 2006, pp. 1–6. ISBN: 1-4244-0022-8. DOI: 10.1109/ICCIS.2006.252275.

[121] Martin Weis et al. “Precision farming for weed management: techniques”. In: *Gesunde Pflanzen* 60.4 (2008), pp. 171–181. ISSN: 0367-4223. DOI: 10.1007/s10343-008-0195-1.

[122] Edmund Weitz. 2016. URL: http://weitz.de/sift/ (visited on 02/02/2020).

[123] Alexander Wendel and James Underwood. “Self-supervised weed detection in vegetable crops using ground based hyperspectral imaging”. In: *Conference band International Conference on Robotics and Automation.* Ed. by ICRA. IEEE, 2016, pp. 5128–5135. ISBN: 1467380261.

[124] Wikipedia. *Zuckerrübe - Wikipedia, Die freie Enzyklopädie.* 2020. URL: https://de.wikipedia.org/w/index.php?title=Zuckerr%5C%C3%5C%BCbe&oldid=199778424 (visited on 05/24/2020).

[125] Z. Yi, C. Zhiguo, and X. Yang. “Multi-spectral remote image registration based on SIFT”. In: *Electronics Letters* 44.2 (2008), p. 107. ISSN: 0162-8828. DOI: 10.1049/el:20082477.

[126] Aston Zhang et al. *Dive into Deep Learning.* http://www.d2l.ai. 2019.

[127] Yun Zhang, David C. Slaughter, and Erik S. Staab. “Robust hyperspectral vision-based classification for multi-season weed mapping”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 69 (2012), pp. 65–73. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2012.02.006.

Appendix

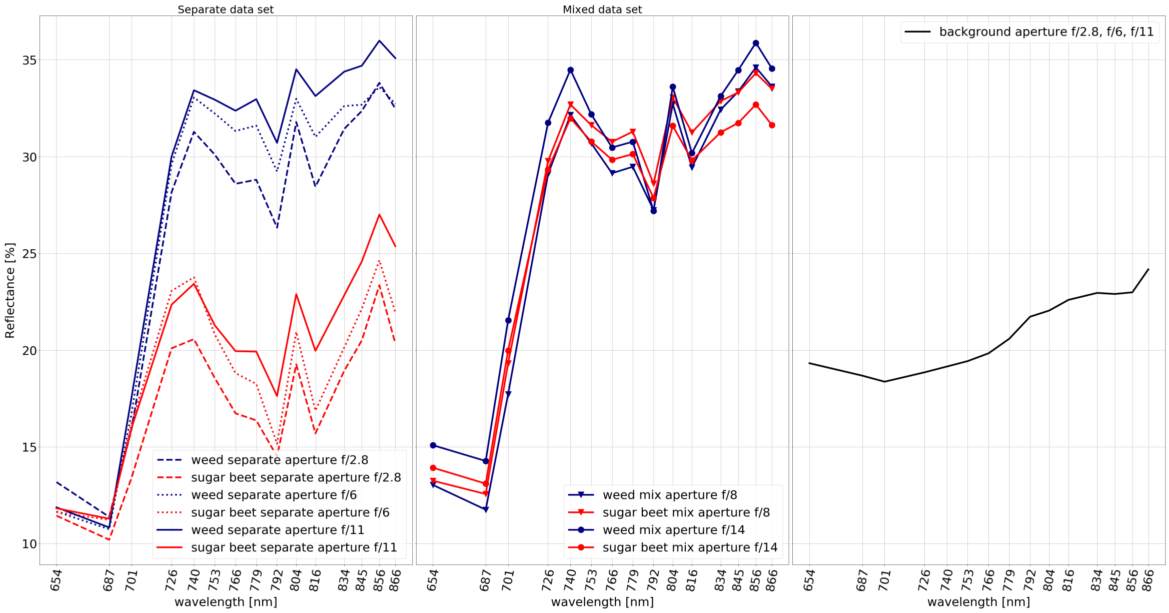


Figure 31: Average calibrated reflectance in % for different apertures of background (potting soil, plastic pots, table), sugar beet and weed plants, $n_{\text{images separate=606}}$, $n_{\text{images mixed=257}}$, $n_{\text{images background=863}}$, resolution one image: 409 x 216 pixels with varying number of plant pixels.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway