Feature extraction for Plant Phenotyping

Project Proposal Bachelor Thesis

Iris Verweij
10633421
Supervisor: Roberto Valenti

April 22, 2016

Contents

1 Phenotyping: Key in ensuring Food Security

On the 21th of March, 2016, the total number of Dutch inhabitants has exceeded 17 million [2]. Surpassing this seventeen million benchmark illustrates the global trend in population growth [22] and life expectancy [23]. In order to provide food security for this ever growing population, a production increase of at least 50 % is required before the year 2050 [24]. Moreover, crop yields may decline under the stress of climate change and food scarcity might lead to depletion of natural capital [15, 24]. Increasing the yield and crop quality requires more insight into plants expressions, phenotype, under different environmental conditions [12, 4, 7].

A crops phenotype is determined by both its genotype, genetic code, and the environmental conditions to which it is exposed [13]. The phenotype of a plant influences amongst others the growth rate, seed yield and resource acquisition [4]. Especially the parameters involving production resources are of interest in the scope of food security [4, 7, 13, 12, 18]. Extensive plant models which include both phenotype as genotype information will increase understanding of biology and are key in ensuring food security[13, 21]. Although systematic qualification of phenotypes appear to be a fundamental step in the evolution of food production, the rate at which crops phenotypes are extracted lags behind on genotyping [13, 12, 4, 7, 18].

The lack of software [13] has constrained the development of automated algorithms for image processing and feature extraction [9, 13]. To help generating robust algorithms and improve the body of literature on phenologics, a number of open source platforms arise [11, 8, 9, 21]. Since the commerce of these research encouraging platforms, many studies have addressed feature extraction.

The importance for feature extraction has been illustrated by Bruno et al. [1]. Images of individual leaves have been analyzed as a mean to generate the structure of a leaf. Based on the idea that objects possess fractional dimensions, which can be seen in both the structure of a leaf, but also in the structure of a complete crop [16], the morphological features of a plant have been used for a binary classification algorithm. The features of patterns which have been extracted are highly characteristic [1].

Nonetheless, most studies regarding plant feature extraction are limited in scope. For example, the leaf growth tracker of Dellen, Scharr and Torras [5] assumes circular shaped leaves. Therefore this technique is not suitable for other Rosette breeds. Similar assumptions where made in the research of Pape and Klukas [14], since very small leaves and leaves with strong overlap remain difficult to detect with high probability. Experiments with several methods for leaf segmentation by Scharr et al. [18] has helped to gain insight in the difficulties of this task. Scharr et al. [18] point out that the correct localization of a leafs center is key for a highly accurate segmentation. Unfortunately, once leaves heavily overlap, which is more likely to happen in more advanced growth stages, segmentation fails. Thus, the difficulty in extracting features from a complete plant puts constrains on the robustness of the methods.

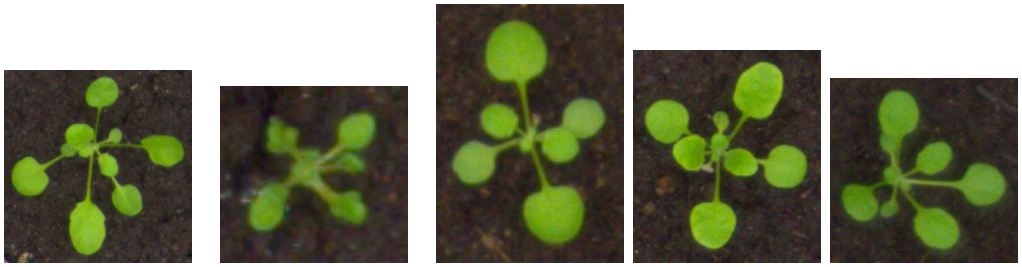
Due to the complexity of a plant and the wide variety across breeds, the idea emerged to create a feature extraction method which surpasses the traditional characteristic elements of a plant. A new representation of a plant that can be used for various types of biology research would help in gaining insight into the connection between a plants phenotype and genotype. In order to propose a new framework for plant representation the following research question emerged. *What features are required for a plant representation which is breed, rotation and growth rate invariant?*

By comparing the new proposed representation with researches which use other feature extraction methods, the representation can be evaluated.

2 Method

A method for feature extraction is proposed in order to create a representation of the plants exterior. The programming language Python with its extensions in OpenCV will be used to do so. The features will be based on salient point pairs. First the data will described, thereafter the process of feature extraction will be set out.

2.1 Data



(a) Col-0 (b) crt (c) ein2 (d) adh1 (e) pgm

Figure 1: Images from the plant Arabidopsis Thalia 2013 Canon database, captions specify the mutant types [11, 12].

For this research a database with raw and annotated visual data from the Arabidopsis Thalia will be used. This database has been retrieved from the open-source platform `plant-phenotyping.org`[11, 12]. 165 RGB images of individual Arabidopsis Thalia plants (see figure 1), from the year 2013, with 4 different mutant types, in several growth stages, are featured in the first part of the research. The plant representations will be matched across the tray images, shown in figure 2, from the same dataset for evaluation. In addition to the visual data, `plant-phenotyping.org` has included metadata which shows the plant type, a possible treatment and its growth stage in hour.

2.2 Plantlet

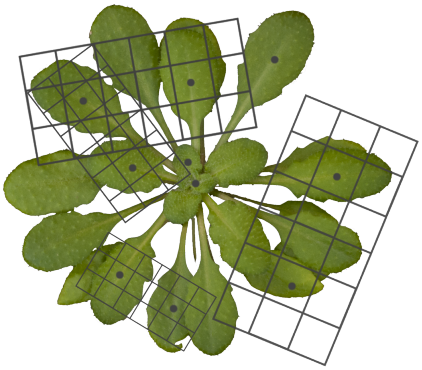
A representation of a plant, a Plantlet, will be created by using features retrieved from salient point pairs. Similar to the keypoints of Lowe's [10] SIFT method, salient points locate characteristic local structure in an image [25, 6]. These points are independent from plant specific elements, such as leaves, the core or size. Surpassing the traditional approaches, a representation of an entire plant's structure should not focus on localizing elements but identify characteristics.

By assigning salient points, using a method from Laurent Itti [19] which has been proven and reviewed, features can be extracted. Figure 3 shows an simplified illustration of the feature extraction. The assigned points will be paired with every other salient point in the same image. Every point pair will generate a feature by using a Histogram of Oriented Gradients (HoG) [10]. A schematic support of the feature extraction to construct a Plantlet is set out in figure 4.

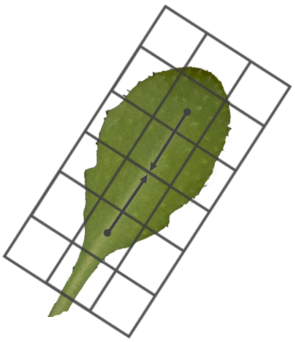
Figure 2: Image from the Arabidopsis Thalia 2013 tray database [12, 11].

The HoG grid consists of two 3x3 connected grids, every square produces an histogram of eight parameters, in this research plan the histogram will be referred to as being one big parameter to prevent complexity. After the HoG method 18 parameters have been identified. Thereafter two additional parameters are included, the length of the grid and the normalized growth stage in time retrieved from the metadata. The growth stage is added as an parameter of the feature to be able to distinguish between small plants and young plants.

After the extraction, all features of a Plantlet are compared with a visual 'Bag-of-Words' method [26]. The top ten features with the highest similarity will be selected, these features are expected to give the best representation of the plant. Another more experimental method to find fractal structure might be applied [16]. This method tries to find the translation matrix which is needed to transform features. However, this method has never been applied to a similar task and might result in unsatisfiable outcomes. Therefore 'Bag-of-Words' will primarily be used for selecting the best features of a Plantlet.



(a) Arabidopsis Thalia Col-0 plant with illustrative imaginary salient points. Point pairs have already been created for a small section.



(b) Leaf of the Arabidopsis Thalia with imaginary salient point pair. A grid is alligned to create the HoG which will be used to extract 18 parameters. In addition the length of the grid and the growth stage in time are included to create a feature with 20 parameters.

Figure 3: Simplified illustration of the salient point pairs and the feature extraction.

2.3 Evaluation of a Plantlet

The next section will briefly discuss more evaluation methods, this paragraph will give more insight in finalizing the Plantlets.

In order to check whether the Plantlets are indeed good representations of a plants image, the Plantlets will be matched within trays and mutants. This means that both identical as rotated images of the exact same mutants will be represented by Plantlets, and matched. This revise will give insight into the representative value. Precision and recall of the matching results will be used as an evaluation method. If the results do not exceed the baseline, results expected from a dummy classifier, more time and effort is required in choosing the Plantlets features and parameters.

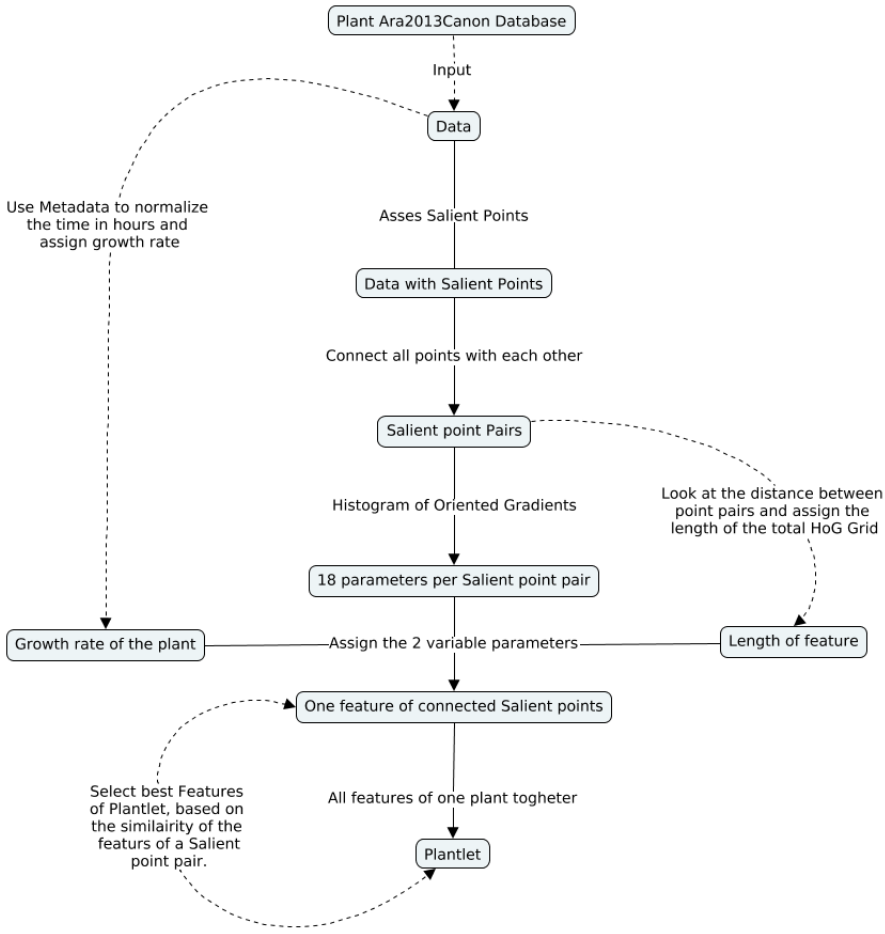


Figure 4: Approach for the creation of Plantlet plant representation by salient point pairs.

3 Evaluation

The representation of the plants are required to be breed, rotation and growth stage invariant. The following evaluation and research methods are applied to check whether the Plantlets indeed achieve in being a better representation than features which have been extracted in other projects.

First a classification of the Arabidopsis Thalia 2013 mutants is conducted. As the evaluation section in the previous paragraph already addressed, Plantlets are checked on rotation invariance by matching within mutants. However, if the representation is breed invariant a classification across trays should also yield successful results. The results of the classification across mutants and time will give insight in the classification across breeds, since a low classification rate across mutants implies a low classification rate across breeds.

The results of cross tray matching will be compared to the outcomes of Silva et al., [20] and Chen et al. [3]. Precision and Recall will also point out whether the classifier based on the Plantlets outperforms a dummy classifier.

A second examination is based on the time invariance of the representation. By using the WEKA module of Java, machine learning techniques will be applied on the data. In contrast to the normal Plantlets, the normalized growth stage in time will now be set as the unknown parameter. Several algorithms, such as support vector machines, naive bayes, neural network and linear regression will be applied to predict the growth stage. Silva et al. [20] have also used similar machine learning techniques, these insights and outcomes will be used during the testing phase.

The mean average error with standard deviation of the predictions are used as a measure to compare machine learning algorithms and to select the best method. Furthermore, the mean average error determines whether the outcome is sufficient. In addition, the research of Rahaman et al. [17] enhances multiple options for comparison.

4 Schedule

The following section will address the order of activities in which they are presented in the GANTT chart, see figure 5.

4.1 Project Design

The first four weeks of the project are used for the specification of the project plan. Besides room for discussing the subjects with a supervisor, these weeks also allows literature study. Two two critical casual relations may be designated: the problem definition could only be created after the allocation of the subjects and the data retrieval has been crucial for the start of the project plan. This project proposal is the final step of the first phase.

4.2 Research

The majority of the time has been reserved for the actual research. The first two weeks are the most important part of the entire research, since this time will be used for the creation of the Plantlets. If the evaluation of the Plantlets do not succeed in surpassing the baseline the entire schedule and plan will have to be altered. However, if this does occur, a partially new literature study will be required to select relevant additions to the parameters of the features.

Nonetheless, I am assuming that two weeks is more than enough time to create the Plantlets. The representations will be used to classify across trays. Finally an prediction of the growth stage is included. Even though the amount of time should be sufficient for the execution of both task, bearing in mind the amount of data, a possible alteration to the schedule for the intermediate presentation on May 27, 2016 might exclude one of the two final tasks. This would, of course, be undesirable. But if this may occurs I would also include an additional literature study to re-investigate relevant projects which might help to overcome issues.

I have planned to start writing during a week after the start of the research. Because insights into the data will help to improve the thesis and over the course of time it will be easier to adjust the text since the results and insights are still relatively new and easily evoked.

4.3 Finalizing

The final stage of the research is used to finalize the thesis and optionally adjust the style in order for it to be published. I do tend to write the thesis in such a style from the beginning, however I might need more time for it to match state of the art requirements. I also want to prepare the final presentation and spend enough time on the content and lay out of the slides.

Furthermore I want to use this time for an external spelling check, since I have been diagnosed with Dyslexia and may need to alter some of the sections due to grammatical mistakes.

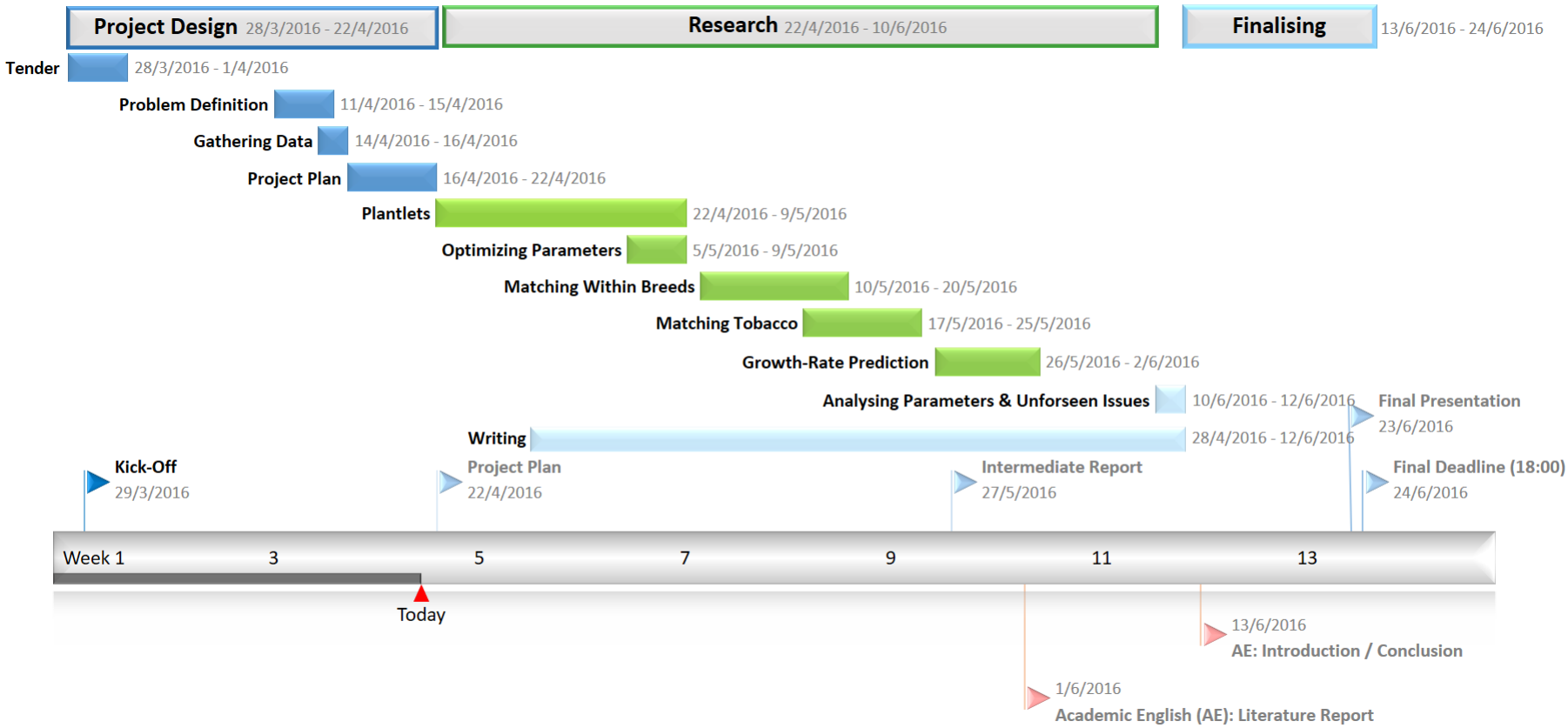


Figure 5: GANTT-Chart which gives an organized overview of the research, including the most important events whom are featured on the bottom bar.

References

[1] Odemir Martinez Bruno, Rodrigo de Oliveira Plotze, Mauricio Falvo, and Mário de Castro. Fractal dimension applied to plant identification. *Information Sciences*, 178(12):2722–2733, 2008.

[2] Centraal Bureau voor Statistiek. Nederland telt 17 miljoen inwoners, March 2016. Accessed: April 19th, 2016.

[3] Dijun Chen, Kerstin Neumann, Swetlana Friedel, Benjamin Kilian, Ming Chen, Thomas Altmann, and Christian Klukas. Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. *The Plant Cell*, 26(12):4636–4655, 2014.

[4] Jeffrey A Cruz, Xi Yin, Xiaoming Liu, Saif M Imran, Daniel D Morris, David M Kramer, and Jin Chen. Multi-modality imagery database for plant phenotyping. *Machine Vision and Applications*, pages 1–15, 2015.

[5] Babette Dellen, Hanno Scharr, and Carme Torras. Growth signatures of rosette plants from time-lapse video. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 12(6):1470–1478, 2015.

[6] Lior Elazary and Laurent Itti. Interesting objects are visually salient. *Journal of vision*, 8(3):3–3, 2008.

[7] Robert T Furbank and Mark Tester. Phenomics–technologies to relieve the phenotyping bottleneck. *Trends in plant science*, 16(12):635–644, 2011.

[8] Caleb Harper and Mario Siller. Openag: A globally distributed network of food computing [spotlight], 2015.

[9] Christian Klukas, Dijun Chen, and Jean-Michel Pape. Integrated analysis platform: an open-source information system for high-throughput plant phenotyping. *Plant physiology*, 165(2):506–518, 2014.

[10] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[11] M. Minervini, A. Fischbach, H. Scharr, and S.A. Tsaftaris. Plant phenotyping datasets, 2015. Accessed: April 15th, 2016.

[12] Massimo Minervini, Andreas Fischbach, Hanno Scharr, and Sotirios A. Tsaftaris. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern Recognition Letters*, pages 1–10, 2015.

[13] Massimo Minervini, Hanno Scharr, and Sotirios A Tsaftaris. Image analysis: The new bottleneck in plant phenotyping [applications corner]. *Signal Processing Magazine, IEEE*, 32(4):126–131, 2015.

[14] Jean-Michel Pape and Christian Klukas. 3-d histogram-based segmentation and leaf detection for rosette plants. In *Computer Vision-ECCV 2014 Workshops*, pages 61–74. Springer, 2014.

[15] John Roy Porter, Liyong Xie, Andrew J Challinor, Kevern Cochrane, S Mark Howden, Muhammed Mohsin Iqbal, David B Lobell, and Maria Isabel Travasso. Chapter 7: Food security and food production systems. Technical report, Cambridge University Press, 2014.

[16] Przemyslaw Prusinkiewicz and Aristid Lindenmayer. Fractal properties of plants. In *The Algorithmic Beauty of Plants*, pages 175–189. Springer, 1990.

[17] Md Matiur Rahaman, Dijun Chen, Zeeshan Gillani, Christian Klukas, and Ming Chen. Advanced phenotyping and phenotype data analysis for the study of plant growth and development. *Frontiers in plant science*, 6, 2015.

[18] Hanno Scharr, Massimo Minervini, Andrew P French, Christian Klukas, David M Kramer, Xiaoming Liu, Imanol Luengo, Jean-Michel Pape, Gerrit Polder, Danijela Vukadinovic, et al. Leaf segmentation in plant phenotyping: a collation study. *Machine vision and applications*, pages 1–22, 2015.

[19] Christian Siagian and Laurent Itti. Biologically-inspired robotics vision monte-carlo localization in the outdoor environment. In *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pages 1723–1730. IEEE, 2007.

[20] LOLA Silva, ML Koga, CE Cugnasca, and AHR Costa. Comparative assessment of feature selection and classification techniques for visual inspection of pot plant seedlings. *Computers and electronics in agriculture*, 97:47–55, 2013.

[21] Dan Stanzione. The iplant collaborative: Cyberinfrastructure to feed the world. *Computer*, 11:44–52, 2011.

[22] The World Bank. Data: Population, total, 2014. Accessed: April 19th, 2016.

[23] The World Bank. Data: World development indicators: Mortality, 2014. Accessed: April 1th, 2016.

[24] The World Bank. Food security: Overview, March 2016. Accessed: April 1th, 2016.

[25] Roberto Valenti, Nicu Sebe, and Theo Gevers. Image saliency by isocentric curvedness and color. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2185–2192. IEEE, 2009.

[26] Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206. ACM, 2007.