Metric Learning on Field Scale Sorghum Experiments

Zeyu Zhang
George Washington University
zeyu@gwu.edzu

Abby Stylianou
Saint Louis University
abby.stylianou@slu.edu

Robert Pless
George Washington University
pless@gwu.edu

Abstract

We explore a metric-learning approach to automatically create representations of images of sorghum grown in a field setting. We train a metric learning algorithm so that images from the same cultivar are mapped to similar locations in feature space. While this task is not itself intrinsically useful, we demonstrate that the features learned in this task support estimation of standard phenotypes (height and leaf length and width), and genetic marker classification.

1. Introduction

High throughput phenotyping in agriculture seeks to automatic the measurement of plant phenotypes by automating aspects of the measurement process. For example, in field grown sorghum, measurements like plant-height and canopy cover, and geometric measurements of leaf length, width and angles, and visual phenotypes that may be measurable automatically. Characterizing these phenotypes over time in large scale field trials opens up opportunities for improving plant-breeding and better understanding of the genotype-phenotype relationships. Attempts to automate these analysis with computer vision pipelines based on classical approaches (edge detection, segmentation and explicit coding for more complicated analysis) such as PlantCV [3] have made many analytics possible, but such approaches often struggle in field conditions.

Here we approach this problem from another viewpoint. We hypothesize that if a convolutional neural network is trained to differentiate between a large number of different cultivars based on imagery, then it must be learning visual features that capture relevant visual phenotypes. This extended abstract captures our initial experiments in this direction. We propose one approach calculate these features, and show preliminary results that those features can be used to predict the phenotypes and the presence or absence of several common genetic markers. Figure 1 gives a high-level overview of the process.

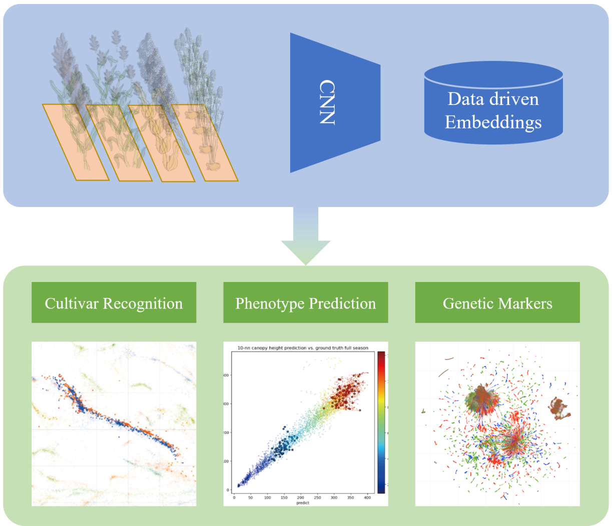

Figure 1. We explore a deep metric learning based approach to producing data-driven phenotypes. Based on a very large dataset of images captured nearly daily from 350 varieties of sorghum grown in 700 plots, we follow a two step process. First, we use a proxy-loss based metric learning approach to to map images to features so that images from the same cultivar are mapped to similar locations in feature space. We then show that this embedding approach generalizes to unseen cultivars, and demonstrate that the features it produces can support image-based cultivar recognition, estimation of standard phenotypes, and prediction of genomic information.

2. Approach Overview

- We create a curated RGB-image dataset based on TERRA [1, 5]. Our dataset contains images from 700 plots, 350 sorghum cultivars, along with their genetic and phenotype labels.
- We build an image embedding using a convolutional neural network a standard Resnet-50 architecture.
- We characterize the ability of these embedding features to support phenomic estimation and genomic-marker classification. There are many tools for these processes; in this proof of concept we use k-NN based approaches.

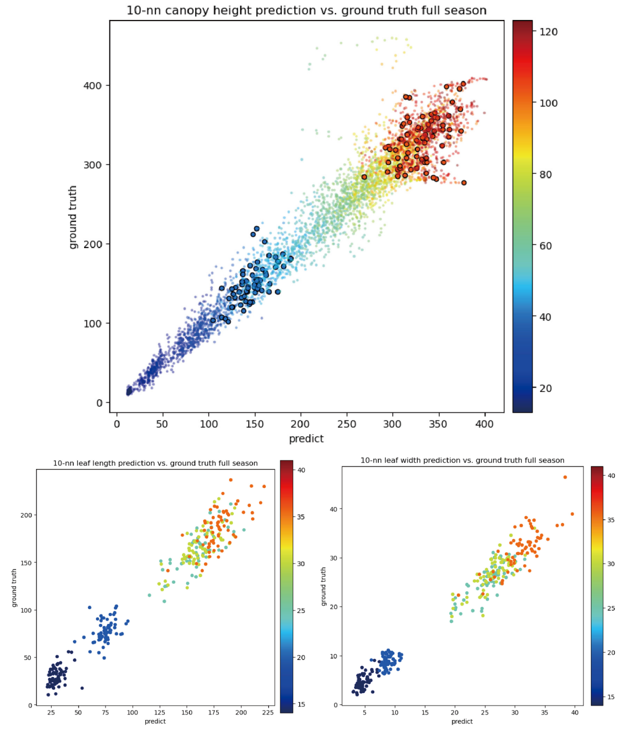

Figure 2. Prediction of phenotypes from test data, based on the weighted average of the k-nearest neighbors in the embedding feature space. Top shows the canopy height prediction, bottom left shows leaf length and bottom right shows the leaf width. The leaf-length and leaf-width predictions are only available in the early season because the leaf-length and width phenotype data in the TERRA-REF dataset are unreliable after canopy-closure.

In this case, for images from test data (images of a cultivar not used during the training of the network), we compute the image embedding, and then predict based on the labels of images with the most similar embeddings.

3. Results

To compute phenotype predictions throughout the season, we embed images from a given plot on a given data into the feature embedding space, find the k-nearest neighbors in all of the training set. We use a weighted sum of each of the training neighbor phenotypes to get the final prediction. We use this approach to predict canopy height, leaf length and width from RGB data (comparing with ground truth labels extracted from 3D laser scanner data).

We also explore the ability of the embedding features to be useful to predict genetic markers for images in the test set. The sorghum lines in the TERRA-REF projects are fully sequenced in the TERRA-REF project, and there are a set of well-known genetic markers families: a leaf wax group that controls leaf wax composition [7]; the dw group that is known to control plant height, stem length and internode length [10, 4]; the dry stalk (d) locus group that is known to control height and sugar composition [9]; the ma group that controls flowering time and maturity [2, 6]; and the tan group which controls pigmentation and tannin production [8].

	accuracy			
genetic marker family	k=1	k=5	k=10	k=15
ma	0.765	0.768	0.777	0.763
tan	0.654	0.664	0.660	0.672
d locus	0.686	0.698	0.703	0.690
dw	0.737	0.747	0.765	0.771
leaf wax	0.664	0.663	0.685	0.684

Table 1. Genetic marker classification accuracy using k-NN with different k. The experiment is designed so that chance performance scores 0.5.

We consider the task of predicting whether an image shows a variety of sorghum that has the 'reference' or 'alternate' version for a particular genetic marker family. For genetic marker classification, we use per-plot, per-day feature max pooling and find the nearest neighbors. After the nearest neighbors acquired, the genetic marker of neighbors vote to get the prediction (reference vs. alternate) of query plot. Table 1 shows that when testing on cultivars not used during training, we perform substantially above chance at the task of predicting reference/alternate for all five of these marker families.

4. Conclusion

We demonstrate training a network on the task of metric-learning — mapping images from the same cultivar into nearby locations in an embedding space — creates features that are useful for phenotype prediction and genetic marker classification. This approach takes advantage the ability to capture large amounts of data in real field conditions, using labels that may be easy to get (what cultivar is grown at each location in the field) instead of labels that may be harder to find (such as hand measurements of phenotypes). Despite using these weaker labels, the features capture visual characteristics related to plant phenotypes, in ways that are good enough to perform above chance at genetic marker classification.

References

[1] Maxwell Burnette, Rob Kooper, J. D. Maloney, Gareth S. Rohde, Jeffrey A. Terstriep, Craig Willis, Noah Fahlgren, Todd Mockler, Maria Newcomb, Vasit Sagan, Pedro Andrade-Sanchez, Nadia Shakoor, Paheding Sidike, Rick Ward, and David LeBauer. TERRA-REF data processing infrastructure. In Sergiu Sanielevici, editor, *Proceedings of the Practice and Experience on Advanced Research Computing, PEARC 2018, Pittsburgh, PA, USA, July 22-26, 2018*, pages 27:1–27:7. ACM, 2018. 1

[2] Hugo E. Cuevas, Chengbo Zhou, Haibao Tang, Prashant P. Khadke, Sayan Das, Yann-Rong Lin, Zhengxiang Ge, Thomas Clemente, Hari D. Upadhyaya, C. Thomas

Hash, and Andrew H. Paterson. The Evolution of Photoperiod-Insensitive Flowering in Sorghum, A Genomic Model for Panicoid Grasses. *Molecular Biology and Evolution*, 33(9):2417–2428, 06 2016. 2

[3] Malia A Gehan, Noah Fahlgren, Arash Abbasi, Jeffrey C Berry, Steven T Callen, Leonardo Chavez, Andrew N Doust, Max J Feldman, Kerrigan B Gilbert, John G Hodge, et al. Plantcv v2: Image analysis software for high-throughput plant phenotyping. *PeerJ*, 5:e4088, 2017. 1

[4] Josie L. Hilley, Brock D. Weers, Sandra K. Truong, Ryan F. McCormick, Ashley J. Mattison, Brian A. McKinley, Daryl T. Morishige, and John E. Mullet. Sorghum dw2 encodes a protein kinase regulator of stem internode length. *Scientific Reports*, 7(1), 7 2017. 2

[5] David LeBauer, Maxwell A. Burnette, Jeffrey Demieville, Noah Fahlgren, Andrew N. French, Roman Garnett, Zhenbin Hu, Kimberly Huynh, Rob Kooper, Zongyang Li, Maitiniyazi Maimaitijiang, Jerome Mao, Todd C. Mockler, Geoffrey Morris, Maria Newcomb, Michael J Ottman, Philip Ozersky, Sidike Paheding, Duke Pauli, Robert Pless, Wei Qin, Kristina Riemer, Gareth Scott Rohde, William L. Rooney, Vasit Sagan, Nadia Shakoor, Abby Stylianou, Kelly Thorp, Richard Ward, Jeffrey W White, Craig Willis, and Charles S Zender. TERRA-REF, An Open Reference Data Set From High Resolution Genomics, Phenomics, and Imaging Sensors. https://datadryad.org/stash/dataset/doi:10.5061/dryad.4b8gtht99, 2020. 1

[6] Rebecca L. Murphy, Daryl T. Morishige, Jeff A. Brady, William L. Rooney, Shanshan Yang, Patricia E. Klein, and John E. Mullet. Ghd7 (ma6) represses sorghum flowering in long days: Ghd7 alleles enhance biomass accumulation and grain production. *The Plant Genome*, 7(2):plantgenome2013.11.0040, 2014. 2

[7] Anurag Uttam, Praveen Madgula, Yechuri Rao, Vilas Tonapi, and Ragimasalawada Madhusudhana. Molecular mapping and candidate gene analysis of a new epicuticular wax locus in sorghum (sorghum bicolor l. moench). *Theoretical and Applied Genetics*, 130, 10 2017. 2

[8] Yuye Wu, Xianran Li, Wenwen Xiang, Chengsong Zhu, Zhongwei Lin, Yun Wu, Jiarui Li, Satchidanand Pandravada, Dustan D. Ridder, Guihua Bai, Ming L. Wang, Harold N. Trick, Scott R. Bean, Mitchell R. Tuinstra, Tesfaye T. Tesso, and Jianming Yu. Presence of tannins in sorghum grains is conditioned by different natural alleles of tannin1. *Proceedings of the National Academy of Sciences*, 109(26):10281–10286, 2012. 2

[9] Jingnu Xia, Yunjun Zhao, Payne Burks, Markus Pauly, and Patrick J. Brown. A sorghum nac gene is associated with variation in biomass properties and yield potential. *Plant Direct*, 2(7):e00070, 2018. 2

[10] Miki Yamaguchi, Haruka Fujimoto, Ko Hirano, Satoko Araki-Nakamura, Kozue Ohmae-Shinohara, Akihiro Fujii, Masako Tsunashima, Xian Song, Yusuke Ito, Rie Nagae, Jianzhong wu, Hiroshi Mizuno, Jun-Ichi Yonemaru, Takashi Matsumoto, Hidemi Kitano, Makoto Matsuoka, Shigemitsu Kasuga, and Takashi Sazuka. Sorghum dw1, an agronomically important gene for lodging resistance, encodes a novel protein involved in cell proliferation. *Scientific Reports*, 6:28366, 06 2016. 2