

# The Sorghum-100 Dataset

Justin Dulay, Chao Ren, Greg Rolwes and Abby Stylianou

Saint Louis University

justin.dulay|chao.ren|greg.rolwes|astylianou@slu.edu

## Abstract

*Automated high throughput plant phenotyping involves leveraging sensors, such as RGB, thermal and hyperspectral cameras, to make large scale and rapid measurements of the physical properties of plants for the purpose of better understanding the difference between crops and facilitating rapid plant breeding programs. One of the most basic phenotyping tasks is to determine the cultivar, or species, in a particular sensor product. This simple phenotype can be used to detect errors in planting and to learn the most differentiating features between cultivars. It is also a challenging visual recognition task, as a large number of highly related crops are grown simultaneously, leading to a classification problem with low inter-class variance. Here, we describe the Sorghum-100 dataset, a large dataset of RGB imagery of sorghum captured by a state-of-the-art gantry system.*

## 1. Introduction

Sorghum is widely used as an agricultural feed substitute, a gluten-free ancient grain, a source of bio-fuel, and even as popcorn in some food communities. Demand for sorghum for a variety of purposes has risen with the need for better food and energy sources, motivating the need to rigorous plant breeding strategies to select for traits that are valued for each purpose (e.g., more grain for food uses or more biomass for bio-fuel production).

Automated high throughput plant phenotyping involves leveraging sensors, such as RGB, thermal and hyperspectral cameras (among others), to make large scale and rapid measurements of the physical properties of plants for the purpose of better understanding the difference between crops and facilitating rapid plant breeding programs. One of the most basic phenotyping tasks is to determine the cultivar (or species) in a particular sensor product. In experiments with a large number of related cultivars being grown simultaneously, this is a challenging fine-grained visual categorization task due to the low inter-class variability.



Figure 1: The TERRA-REF Field and Gantry-based Field Scanner in Maricopa, Arizona (top), with sorghum being grown in the field. Sorghum is a hugely important cereal crop, widely used as a source of grain, agricultural feed, and even bio-fuel. Over several seasons, hundreds of varieties of both bio-energy and grain sorghum were grown in the TERRA-REF field (middle and bottom), and were imaged daily for the purpose of high throughput phenotyping.

## 2. Background

### 2.1. TERRA-REF Field and Gantry-based Field Scanner

In 2016, the Advanced Research Project Agency–Energy (ARPA-E) funded the Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform, or TERRA-REF[2]. The TERRA-REF project stood up a state-of-the-art gantry based system for monitoring the full growth cycle of over an acre of crops with a cutting-edge suite of imaging sensors, including stereo-RGB, thermal, short- and long-wave hyperspectral, and laser 3D-scanner sensors. The goal of the TERRA-REF gantry was to perform in-field automated high throughput plant phenotyping, the process of making phenotypic measurements of the physical properties of plants at large scale and with high temporal resolution, for the purpose of better understanding the difference between crops and facilitating rapid plant breeding programs. Due to the technical demands of high-throughput phenotyping, it is most often performed in controlled environments (e.g., greenhouses with imaging platforms). Controlled environments play a very important role in understanding plant performance by providing management of the abiotic environment, and the ability to reproduce experimental conditions year-round, but plant performance in field settings, both in terms of growth and yield parameters, is strongly influenced by variability in weather, soil conditions and other environmental parameters that cannot be observed in the greenhouse. The TERRA-REF gantry system was designed to meet the sort of technical requirements for high throughput phenotyping in a field setting. The TERRA-REF field and gantry system are shown in Figure 1, and example data captured from its RGB, 3D-scanner and thermal cameras are shown in Figure 2.

Over the course of its first several years in operation, the TERRA-REF platform collected multiple petabytes of sensor data capturing the full growing cycle of sorghum plants from the sorghum Bioenergy Association Panel [1], a set of 390 sorghum cultivars whose genomes have been fully sequenced and which show promise for bio-energy usage.

### 2.2. Phenotyping from Aerial Data

While in-field sorghum phenotypes have been determined using aerial RGB data from drones[8, 6, 3, 10, 4, 7], UAV datasets are limited in their temporal resolution due to the labor required in capturing the data, and their spatial resolution is limited both by the sensors that are able to be mounted on board a drone and the time constraints of the drone operator (flights from lower to the ground are higher resolution but take longer to complete). By comparison, the data from the TERRA-REF Gantry-based Field Scanner has both high spatial resolution (the gantry has extremely high

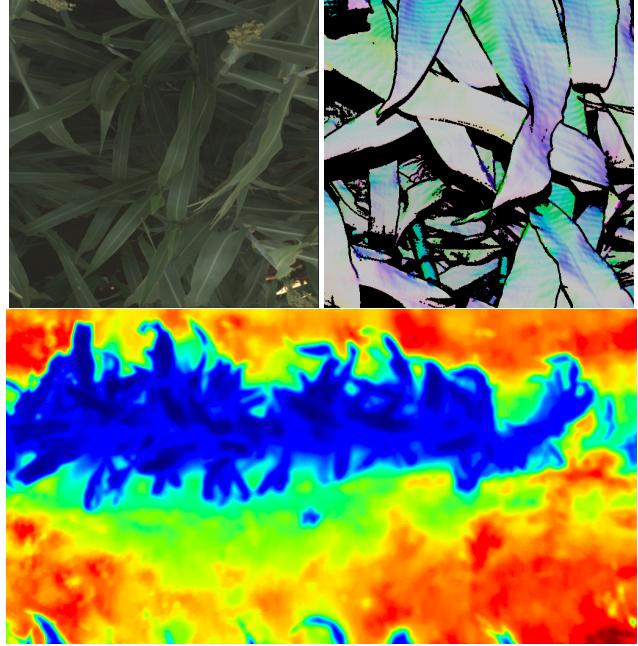


Figure 2: **Example data from the TERRA-REF gantry system.** (top-left) RGB data. (top-right) 3D-scanner data (false color, where color indicates the surface normal and value indicates depth from the scanner). (bottom) Thermal data. In this paper we focus on data from the RGB camera.

quality sensors and the height of the gantry is placed to optimally image the plants throughout the growing cycle), and temporal resolution (data is captured every day).

## 3. Dataset & Classification Task

In this paper, we describe the Sorghum-100 dataset, a curated subset of the RGB imagery captured during the TERRA-REF experiments, labeled by cultivar and day after planting. The dataset will be released publicly and there will be a corresponding Kaggle competition. This data could be used to develop and assess a variety of plant phenotyping models which seek to answer questions relating to the presence or absence of desirable traits (e.g., “does this plant exhibit signs of water stress?”). In this paper we focus on the question: “What cultivar is shown in this image?” Predicting the cultivar in an image is an especially good challenge problem for familiarizing the machine learning community with the TERRA-REF data. At first blush, the task of predicting the cultivar from an image of a plant may not seem to be the most biologically compelling question to answer – in the context of plant breeding, the cultivar, or parental lines are typically known. A high accuracy machine learning predictor of the species captured by the sensor data, however, can be used to determine where errors in the planting process may have occurred. For example,

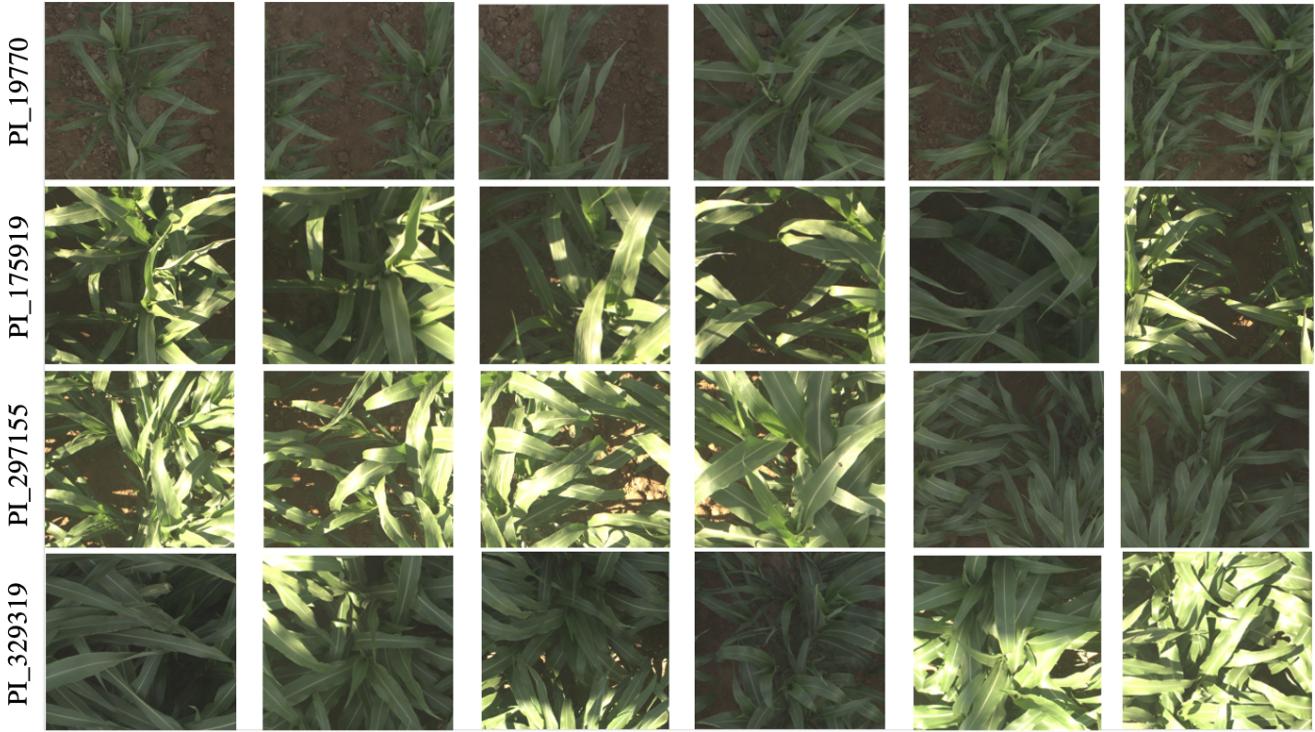


Figure 3: Multiple images from the dataset representing different cultivars. The rows represent different cultivars. The columns represent different captured dates respectively: June 1st, 3rd, 7th, 17th, 19th and 27th, 2017.

seed may be mislabeled prior to planting, or planters may get jammed, depositing seeds non-uniformly in a field [9]. Both types of errors are surprisingly common and can cause major problems when processing data from large-scale field experiments with hundreds of cultivars and complex field planting layouts.

The Sorghum-100 dataset consists of 48,106 images and 100 different sorghum cultivars grown in June of 2017 (the images come from the middle of the growing season when the plants were quite large but not yet lodging – or falling over). In Figure 3, we show a sample of images from four different cultivars. Each row includes six images from different dates in June. This figure highlights the high inter-class visual similarity between the different classes, as well as the high variety in the imaging conditions from one day to the next, or even over the course of a day.

The dataset is divided into a training dataset and a testing dataset. Each cultivar was grown in two separate plots in the TERRA-REF field as shown in Figure 4 (top) to account for extremely local field or soil conditions that might impact the growth of plants in one particular plot. We leverage this natural split in the data when dividing our dataset between train and test – images for a given cultivar in the training dataset come from one plot, while the test images from that same cultivar come from the other plot. This means that a model cannot achieve high performance by memorizing

features that aren’t meaningful phenotypes (e.g., by memorizing patterns observed in the dirt). The training dataset consists of 22,635 images, and the testing dataset consists of 25,471 images (which plot was included in training vs. test was randomly selected).

## 4. Baseline Results

To provide a reasonable baseline on the Sorghum-100 dataset, we trained a ResNet-50 model [5] (pre-trained on ImageNet). During training we resize the original images to be 512 pixels on its shortest side, and then take a random  $512 \times 512$  crop (at test time, we take a center crop). We normalize by channel means and standard deviations and perform random horizontal and vertical flips. We use global average pooling and train with cross entropy loss. This baseline approach achieves 72.12% top-1 classification accuracy on the test set.

## 5. Conclusion

In this paper, we introduced the Sorghum-100 cultivar classification dataset which includes tens of thousands of images from 100 bio-energy lines of sorghum grown in the TERRA-REF field. This is the first gantry-based sorghum dataset, which has higher temporal and spatial resolution than similar UAV based datasets, making the data suitable



**Figure 4: TERRA-REF Field Organization.** (top) Each experimental cultivar was planted in two different plots at distant locations in the field (borders were planted with well-known cultivars). In the top figure, each plot is labeled by its cultivar name, and plots from the same cultivar have matching colors. For each non-border cultivar, we include images from one of the plots in our training data, and images from the other plot in our test data, requiring models to generalize across field locations (as opposed to, for example, overfitting on unique ground features that are not relevant to the cultivar). (bottom) Original data from the sensor is pre-processed to crop regions that confidently only consist of plants from a single plot. The blue rectangle in the image above shows the ground boundaries of a plot projected onto the image.

for generating models for true high-throughput phenotyping in the field. While this paper presents the dataset in the context of cultivar classification, we hope to in the future release more data products and metadata, including data from additional sensors and both hand- and algorithm-generated phenotypes with the goal of supporting the development of machine learning models for more sophisticated high-throughput phenotyping tasks.

The dataset and corresponding competition can be found at <https://www.kaggle.com/c/sorghum-100>.

## References

- [1] Zachary W Brenton, Elizabeth A Cooper, Mathew T Myers, Richard E Boyles, Nadia Shakoor, Kelsey J Zielinski, Bradley L Rauh, William C Bridges, Geoffrey P Morris, and Stephen Kresovich. A genomic resource for the development, improvement, and exploitation of sorghum for bioenergy. *Genetics*, 204(1):21–33, 2016. [2](#)
  - [2] Maxwell Burnette, Rob Kooper, J. D. Maloney, Gareth S. Rohde, Jeffrey A. Terstriepl, 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. [2](#)
  - [3] Yuhao Chen, Javier Ribera, Christopher Boomsma, and Edward J Delp. Plant leaf segmentation for estimating phenotypic traits. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3884–3888. IEEE, 2017. [2](#)
  - [4] Wei Guo, Bangyou Zheng, Andries B. Potgieter, Julien Diot, Kakeru Watanabe, Koji Noshita, David R. Jordan, Xuemin Wang, James Watson, Seishi Ni-nomiya, and Scott C. Chapman. Aerial imagery analysis – quantifying appearance and number of sorghum heads for applications in breeding and agronomy. *Frontiers in Plant Science*, 9:1544, 2018. [2](#)
  - [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. [3](#)
  - [6] Ali Masjedi, Jieqiong Zhao, Addie M Thompson, Kai-Wei Yang, John E Flatt, Melba M Crawford, David S Ebert, Mitchell R Tuinstra, Graeme Hammer, and Scott Chapman. Sorghum biomass prediction using uav-based remote sensing data and crop model simulation. In *IGARSS 2018-2018 IEEE International Geo-*

*science and Remote Sensing Symposium*, pages 7719–7722, 2018. 2

- [7] Andries B. Potgieter, Barbara George-Jaeggli, Scott C. Chapman, Kenneth Laws, Luz A. Suárez Cadavid, Jemima Wixted, James Watson, Mark Eldridge, David R. Jordan, and Graeme L. Hammer. Multispectral imaging from an unmanned aerial vehicle enables the assessment of seasonal leaf area dynamics of sorghum breeding lines. *Frontiers in Plant Science*, 8:1532, 2017. 2
- [8] Javier Ribera, Fangning He, Yuhao Chen, Ayman F Habib, and Edward J Delp. Estimating phenotypic traits from uav based rgb imagery. *arXiv preprint arXiv:1807.00498*, 2018. 2
- [9] Davinder Sharma, Jagadish Rane, Rajender Singh, Vijay Kumar Gupta, and Ratan Tiwari. Comparison of different planting methods to determine the precision of phenotyping wheat in field experiments. In *Advances in Plant & Microbial Biotechnology*, pages 77–83. Springer, 2019. 3
- [10] Z. Zhang, A. Masjedi, J. Zhao, and M. M. Crawford. Prediction of sorghum biomass based on image based features derived from time series of uav images. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6154–6157, 2017. 2