PhytoOracle: Scalable, modular phenomic data processing pipelines

Emmanuel Gonzalez[a], Ariyan Zarei[b], Nathanial Hendler[a], Michele Cosi[a], Jeffrey Demieville[a], Sebastian Calleja[a], Travis Simmons[a], Holly Ellingson[c], Nirav Merchant[c], Eric Lyons[a,c], Duke Pauli[a,c]

[a]School of Plant Sciences, University of Arizona, Tucson, USA 85721; [b]Department of Computer Science, University of Arizona, Tucson, USA 85721; [c]Data Science Institute, University of Arizona, Tucson, USA 85721

ABSTRACT

Previous crop yield improvements have been largely due to the implementation of new management strategies, mechanization, and application of emerging technologies. While these approaches have led to stable, linear improvements, increases in crop yields are currently plateauing. The use and improvement of rapid, automated, and accurate phenomic selection methods leveraging high-resolution data collected throughout a growing season could help identify stress-adaptive traits to meet the growing global food demand. As the capacity of phenomics to generate larger and higher dimensional data sets improves, there is an urgent need to develop and implement robust and scalable data processing pipelines for rapid turnaround of processed results. Current phenomics processing pipelines lack modularity and the ability to exploit the distributed computational infrastructure required for machine learning (ML)-based workloads. To address these challenges, we developed PhytoOracle (PO), a suite of modular, scalable pipelines that aim to improve data processing efficiency for plant science research. PO integrates open-source frameworks for distributed task management on local, cloud, or high-performance computing (HPC) systems. Each pipeline component is available as a standalone container which can be independently deployed or linked into a pipeline. Additionally, researchers can swap between available containers or integrate new ones suited to their specific research. PO extracts phenotype trait values such as volume, height, canopy temperature, and maximum quantum efficiency (F_v/F_m) of photosystem II from data captured in field settings, enabling the study of phenotypic variation for elucidation of the genetic components of quantitative traits.

Keywords: image processing, machine learning, deep learning, automatic plant phenotyping, cyberinfrastructure, distributed computing, field phenomics, high performance computing

1. INTRODUCTION

1.1 Computational Technologies

The phenotyping datasets of the future pose new processing, storage, and analysis bottlenecks which can be addressed by leveraging both established and emerging technologies. Large scale phenomic data must be processed in a reproducible and timely manner to provide actionable insights. To address these bottlenecks, PO leverages a variety of computational technologies and resources. For example, data management systems such as CyVerse's Data Store, a cloud-based data management system built on the Integrated Rule-Oriented Data System (iRODS), provide data storage and cross-platform access during data processing[1]. Container technologies, such as Docker and Singularity, provide stand-alone environments with required dependencies pre-installed. High performance computers (HPCs) provide powerful processors, connected to fast memory, disk storage and networking to scale up processing tasks.

HPCs coupled with container technology provide a reproducible, scalable environment[2]. Larger datasets require distributed frameworks that leverage thousands of computers to process data within reasonable timeframes. CCTools[3], a suite of computing tools for deploying scalable applications, consists of Makeflow and Work Queue which provide the language and computational resource management, respectively, required to scale tasks across local, Cloud, or HPC computing environments. When coordinated, these computational resources can revolutionize the management and analysis of high-throughput phenotyping data, but significant software engineering is required.

1.2 Supported Data

PO provides an orchestration framework for processing RGB, thermal, photosystem II chlorophyll fluorescence (PSII), and 3D point cloud data irrespective of phenotyping platform. PO was developed for processing the large amount of phenomic data collected by the University of Arizona's Lemnatec Field Scanalyzer (FS), but it can also process phenomic data collected with other phenotyping platforms (e.g. drones, carts, and mobile phones). This is possible due to the modular nature of PO, which allows for the removal, rearrangement, or single deployment of pipeline components. PO can scale computation of large datasets by leveraging many processing cores on local, Cloud, or HPC clusters, allowing it to process among the largest phenomic datasets in the plant science research space. Comparisons between drone (DR) and FS data number and size highlight the need for distributed computing pipelines; for example, RGB data collections over the same region resulted in raw datasets of 458-532 images (3.3-3.9 GB) for DR and 9155-9270 images (104.0-141.0 GB) for FS. This amounts to an increase by factors of 17-20 in the number of images collected and 36-43 in the collection size.

2. MATERIALS AND METHODS

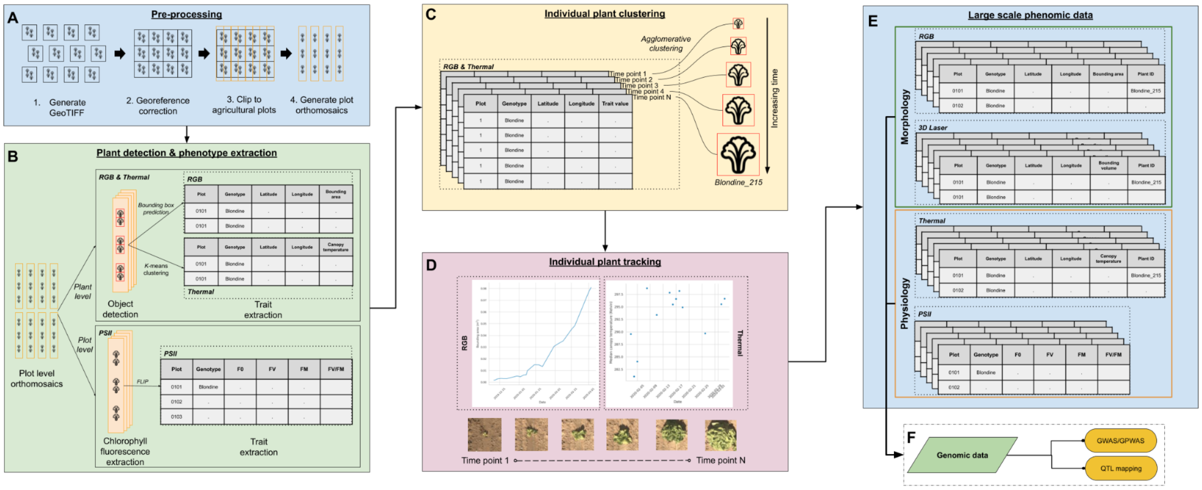

Figure 1. PO pipelines extract time-series phenotypic trait data from multiple data sources. (A) Preprocessing generates GeoTIFF images which are plot-clipped using a GeoJSON containing plot boundaries. Individual plot-clipped GeoTIFFs are mosaiced resulting in one orthomosaic per plot. (B) Plants are detected within RGB and thermal plot orthomosaics, which are subsequently used for the extraction of individual plants from 3D laser scanner data. (C) Agglomerative clustering is used to associate single plants detected throughout the growing season and (D) enable plant tracking. (E, F) PO provides time-series phenomic data for large scale QTL mapping and GWAS.

2.1 Data Processing Pipelines

RGB and thermal pipelines leverage sensor-specific Faster R-CNN detection models for plant detection, allowing for the localization of individual lettuce plants representing many genotypes (Figure 1). Upon detection, individual plants are isolated and analyzed for canopy temperature in thermal images and bounding area in RGB images. The geographical coordinates of each plant are collected, allowing for localization and extraction of individual plants in co-registered point clouds.

2.2 Model training and performance evaluation

For supervised ML training, RGB and thermal images were labeled using Labelbox, a web-based annotation tool. Faster R-CNN models were trained using the Detecto Python package and validated using labeled testing sets for each data type. The capture date of each image was included in the metadata to assess changes in model performance throughout the growing season. To quantify detection accuracy, the intersection over union (IoU), recall, precision, F1-score, and accuracy were calculated.

2.3 Benchmarking

CCTools Makeflow and Work Queue output processing logs, which provide information such as the total number of workers, tasks completed, and processing times. The relationship between processing time and number of processing cores was investigated using these data.

3. RESULTS

3.1 Model performance evaluation

RGB-FS achieved the greatest median intersection over union (IoU), with RGB-DR and Thermal-FS having comparable values (Table 1, Figure 2). A noticeable increase in intersection over union was observed over the growing season across all data types. Thermal-FS models had the greatest overall accuracy at 0.984 followed by RGB-DR at 0.970 and RGB-FS at 0.957 (Table 1).

Table 1. Performance metrics for Field Scanalyzer RGB (RGB-FS), drone RGB (RGB-DR), and thermal (Thermal-FS) Faster R-CNN detection models on Field Scanalyzer image data. TP=true positive, FP=false positive, and FN=false negative.

Data Type	Total detections	TP	FP	FN	Recall	Precision	F1-score	Accuracy
RGB - DR	4356	4097	182	77	0.982	0.957	0.969	0.970
RGB - FS	2752	2519	178	54	0.979	0.934	0.956	0.957
Thermal - FS	1450	1404	10	36	0.975	0.993	0.984	0.984

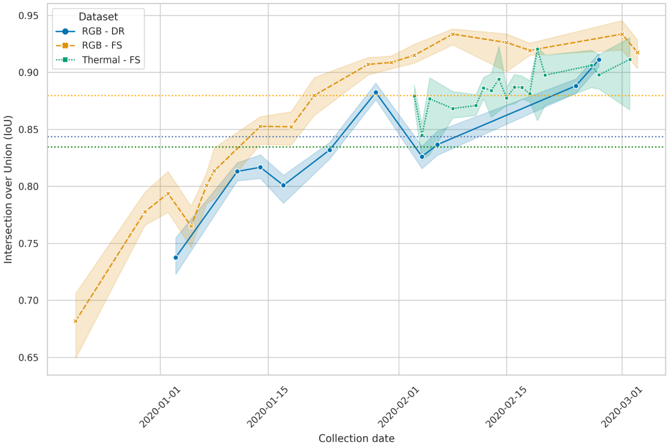

Figure 2. Change in IoU across data collections. RGB Field Scanalyzer scans (RGB-FS) and drone flights (RGB-DR) began earlier than thermal, allowing us to capture the temporal effect of collection date, a proxy to plant size, on median IoU. Blue, green, and orange dotted lines represent the median IoUs for RGB-FS, RGB-DR, and Thermal-FS test datasets, respectively. Error bands represent a 95% confidence interval.

3.2 Benchmarking

At the maximum number of workers tested in this study (1024 workers), processing times were: 235 minutes for 9,270 RGB images (140.7 GB), 235 minutes for 9,270 thermal images (5.4 GB), and 13 minutes for 39,678 PSII images (86.2 GB). These processing times include geo-correction of images and plant detection steps.

4. FUTURE DIRECTIONS

PO implements distributed computing, geospatial methods, and proximal sensing technologies to track plot or individual plant phenotypic traits throughout the growing season. The phenotypic trait data extracted from 2D image and 3D point cloud data provide large morphological and physiological phenomic datasets allowing for QTL mapping and GWAS studies. Additionally, point clouds generated by PO fill a gap in 3D ML training datasets, which can be used to develop and validate ML models for plant phenotyping applications[4]. Future directions include the development of a publicly-accessible, open-source VR experience that allows users to interact with 3D point clouds and visualize time-series phenotypic trait data within a single environment.

DATA AVAILABILITY STATEMENT

Raw, intermediate, and processed data are available on the CyVerse Data Commons. The PhytoOracle workflow repository can be found at https://github.com/LyonsLab/PhytoOracle and processing containers at https://github.com/phytooracle.

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Department of Energy Advanced Research Projects Agency - Energy OPEN program under Award Number DE-AR0001101, the U.S. Department of Energy Biological and Environmental Research program under Award Number DE-SC0020401, and National Science Foundation CyVerse project under Award Number DBI-1743442.

REFERENCES

[1] S. A. Goff *et al.*, “The iPlant collaborative: Cyberinfrastructure for plant biology,” *Front. Plant Sci.*, vol. 2, no. JUL, Jul. 2011, doi: 10.3389/fpls.2011.00034.

[2] G. M. Kurtzer, V. Sochat, and M. W. Bauer, “Singularity: Scientific containers for mobility of compute,” *PLOS ONE*, vol. 12, no. 5, p. e0177459, May 2017, doi: 10.1371/journal.pone.0177459.

[3] M. Albrecht, P. Donnelly, P. Bui, and D. Thain, *Makeflow: a portable abstraction for data intensive computing on clusters, clouds, and grids*. 2012, p. 13.

[4] N. Chebrolu, F. Magistri, T. Läbe, and C. Stachniss, “Registration of spatio-temporal point clouds of plants for phenotyping,” *PLOS ONE*, vol. 16, no. 2, p. e0247243, Feb. 2021, doi: 10.1371/journal.pone.0247243.

REVIEWS

Anonymous reviews will be published

RESPONSE TO REVIEWERS

Response to reviews will be published