An extensive lab- and field-image dataset of crops and weeds for computer vision tasks in agriculture

Michael A. Beck
University of Winnipeg
Winnipeg, MB Canada
m.beck@uwinnipeg.ca

Chen-Yi Liu
University of Winnipeg
Winnipeg, MB Canada

Christopher P. Bidinosti
University of Winnipeg
Winnipeg, MB Canada

Christopher J. Henry
University of Winnipeg
Winnipeg, MB Canada

Cara M. Godee
University of Winnipeg
Winnipeg, MB Canada

Manisha Ajmani
University of Winnipeg
Winnipeg, MB Canada

Abstract

We present two large datasets of labelled plant-images that are suited towards the training of machine learning and computer vision models. The first dataset encompasses as the day of writing over 1.2 million images of indoor-grown crops and weeds common to the Canadian Prairies and many US states. The second dataset consists of over 540,000 images of plants imaged in farmland. All indoor plant images are labelled by species and we provide rich metadata on the level of individual images. This comprehensive database allows to filter the datasets under user-defined specifications such as for example the crop-type or the age of the plant. Furthermore, the indoor dataset contains images of plants taken from a wide variety of angles, including profile shots, top-down shots, and angled perspectives. We further introduce a 14,000 images sample, intended as a quick entry point for the indoor-dataset.

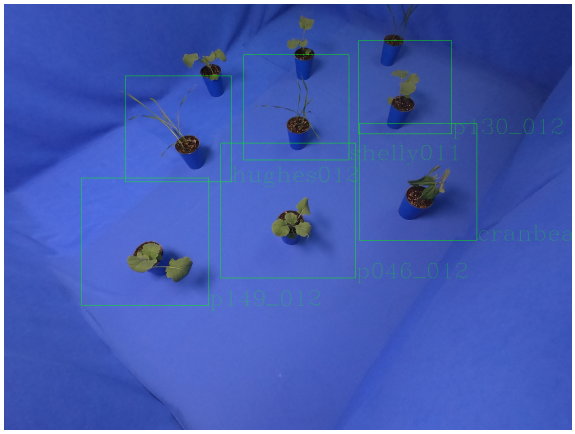

Figure 1. An image taken by the system with drawn calculated bounding boxes.

1. Introduction

A sufficient amount of labelled data is critical for machine-learning based models and a lack of training data often forms the bottleneck in the development of new algorithms. This problem is magnified in the area of digital agriculture as the objects of interest – plants – have a wide variety in appearance that stems from the plants' growing stage, its specific cultivar, its health, and the current environment. Furthermore, the correct classification of plants requires expert knowledge, which cannot easily be crowd-sourced. All of this frames the labelling of plant-data as a challenge that is significantly harder compared to similar image-labelling tasks. Yet, as we witness the introduction of sensors [1–4], robotics [5–10], and machine learning [11–16] to agricultural applications, there is a strong demand for such training data. Precision agriculture (also known as digital agriculture, smart farming, or Agriculture 4.0) has the potential to increase yields while reducing hte usage of resources [16–28].

Here we describe two datasets, each consisting of hundreds of thousands of images, suitable for machine-learning and computer vision applications. The first dataset, the *lab-data*, consists of indoor-grown crops and weeds that had been imaged from a wide variety of angles. The plants selected are species common on farmlands in the Canadian prairies and many US states. All indoor images had been captured and automatically labelled by our robotic system described in [29]. The second dataset consists of images taken in the field in the growing seasons of 2019 and 2020.

Table 1. Image counts per species

Common Name	Image count	Individual plants
Barley	30597	14
Barnyard Grass	76258	21
Common Bean	159217	53
Canada Thistle	89731	14
Canola	255004	128
Dandelion	87426	16
Field Pea	68658	24
Oat	59153	28
Smartweed	99650	21
Soybean	203980	84
Wheat	120417	47
Wild Buckwheat	24973	15
Wild Oat	7065	3
Yellow Foxtail	14815	5

Further we provide a 14,000 images sample from the lab-data.

2. Data- and Metadata-Structure

The lab-data can be divided into 4 different kind of files that relate to each other as follows.

- Plain images: These are the images as captured by the camera. They typically show several plants in the same image.
- Bounding box images: These images are the same as the original images with the difference that they are overlaid with visible bounding boxes around the plants as calculated by the system. Plants too close to the border of the image or overlapping too far into each other are not being bounded by the system.
- Single plant images: These are images cropped out from plain images according to the calculated bounding boxes. Only plants for which a bounding box has been drawn are cropped out as individual image.
- JSON-files: These files contain the metadata associated with each plain image and are described in more detail in the readme.txt of the data-sample.

See Figure 1 for examples of an image with bounding boxes. The field-data collection is in structure similar to the above and also described in the readme-file.

3. Description of subsample

To create a visual overview on the lab-data we created a subsample that is structured as follows:

For each species listed in Table 1 we have selected 1,000 single plant images, thus the subsample contains 14,000 images. Furthermore, within each of these categories we have selected images, such that the age-distribution of the 1,000 images closely matches the age-distribution of all available images for that species. In addition we selected images such that all individual plants grown are represented in the subsample with the following exceptions: There are 51 individual Common Beans present in the subsample (instead of 53 in the entire dataset), as well as 113 Canola plants (of 128), 51 Soybean plants (of 84) and 37 Wheat plants (of 47). The distribution of the image dimensions (width, height) for the subsamples resembles the size distribution of the entire dataset, we did however not select images to directly optimize the sample under that criteria.

The total size on disk of the subsample is approximately 2.2 GB. The subsample contains single plant images only, which are organized in one subfolder per species. We consider this subsample as a good entry point into the entirety of the dataset, which can be used to train some initial models. For example, simple models that differentiate between species or classes of species (e.g., monocots versus dicots, crops versus weeds) or younger versus older plants.

4. Conclusion and data availability

In this paper we presented an extensive dataset of labelled plant images. These images show crops and weeds as common in the Canadian prairies and northern US states. We described and published a subsample that mirrors the full dataset in key characteristics, but is smaller in overall size and thus more tractable. We are actively growing the dataset into several dimensions: New field- and lab-data is being acquired and processed as of writing. Furthermore, additional data-sources such as the generation of 3d-pointclouds and hyperspectral scans are being tested and developed. Additional field-data sources are also being explored, including imagery from UAVs and a semi-autonomous rover. Data from these sources will accompany the datasets presented in this paper in the near future.

The 14,000 images sample is available on https://doi.org/10.25739/rwcw-ex45 at the CyVerse Data Store, a portal for full data lifecycle management. The full dataset which contains 1.2 million single plant images (and counting) is made available to researchers and industry through the data-portal hosted by EMILI under http://emilicanada.com/ (Digital Agriculture Asset Map). The authors take Lobet's general critique [11] on data-driven research in digital agriculture (or any research field) seriously. We further created a datasheet following the guidelines of Gebru et al. [30]

References

[1] M. Vázquez-Arellano, H. W. Griepentrog, D. Reiser, and D. S. Paraforos, "3-d imaging systems for agricultural applications - a review," *Sensors*, vol. 16, no. 5, 2016. 1

[2] F. Yandun Narvaez, G. Reina, M. Torres-Torriti, G. Kantor, and F. A. Cheein, "A survey of ranging and imaging techniques for precision agriculture phenotyping," *IEEE/ASME Transactions on Mechatronics*, vol. 22, no. 6, pp. 2428–2439, 2017. 1

[3] A. Antonacci, F. Arduini, D. Moscone, G. Palleschi, and V. Scognamiglio, "Nanostructured (bio)sensors for smart agriculture," *TrAC Trends in Analytical Chemistry*, vol. 98, pp. 95–103, 2018. 1

[4] A. Khanna and S. Kaur, "Evolution of internet of things (IoT) and its significant impact in the field of precision agriculture," *Computers and Electronics in Agriculture*, vol. 157, pp. 218–231, 2019. 1

[5] R. Oberti and A. Shapiro, "Advances in robotic agriculture for crops," *Biosystems Engineering*, vol. 100, no. 146, pp. 1–2, 2016. 1

[6] A. Bechar and C. Vigneault, "Agricultural robots for field operations. part 2: Operations and systems," *Biosystems Engineering*, vol. 153, pp. 110–128, 2017. 1

[7] A. Bechar and C. Vigneault, "Agricultural robots for field operations. part 2: Operations and systems," *Biosystems Engineering*, vol. 153, pp. 110–128, 2017. 1

[8] T. Duckett, S. Pearson, S. Blackmore, and B. Grieve, "Agricultural robotics: The future of robotic agriculture," *CoRR*, vol. abs/1806.06762, 2018. 1

[9] R. R. Shamshiri, C. Weltzien, I. A. Hameed, I. J. Yule, T. E. Grift, S. K. Balasundram, L. Pitonakova, D. Ahmad, and G. Chowdhary, "Research and development in agricultural robotics: a perspective of digital farming.," *International Journal of Agricultural and Biological Engineering*, vol. 11, pp. 1–14, 2018. 1

[10] J. Relf-Eckstein, A. T. Ballantyne, and P. W. Phillips, "Farming reimagined: a case study of autonomous farm equipment and creating an innovation opportunity space for broadacre smart farming," *NJAS - Wageningen Journal of Life Sciences*, vol. 90-91, p. 100307, 2019. 1

[11] G. Lobet, "Image analysis in plant sciences: Publish then perish," *Trends in Plant Science*, vol. 22, no. 7, pp. 559–566, 2017. 1, 2

[12] J. Wäldchen, M. Rzanny, M. Seeland, and P. Mäder, "Automated plant species identification-trends and future directions," *PLOS Computational Biology*, vol. 14, pp. 1–19, 04 2018. 1

[13] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, 2018. 1

[14] D. I. Patrício and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review," *Computers and Electronics in Agriculture*, vol. 153, pp. 69–81, 2018. 1

[15] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2018. 1

[16] K. Jha, A. Doshi, P. Patel, and M. Shah, "A comprehensive review on automation in agriculture using artificial intelligence," *Artificial Intelligence in Agriculture*, vol. 2, pp. 1–12, 2019. 1

[17] A. Binch and C. Fox, "Controlled comparison of machine vision algorithms for rumex and urtica detection in grassland," *Computers and Electronics in Agriculture*, vol. 140, pp. 123–138, 2017. 1

[18] M. D. Bah, A. Hafiane, and R. Canals, "Deep learning with unsupervised data labeling for weed detection in line crops in uav images," *Remote Sensing*, vol. 10, no. 11, 2018. 1

[19] P. Bosilj, T. Duckett, and G. Cielniak, "Analysis of morphology-based features for classification of crop and weeds in precision agriculture," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 2950–2956, 2018. 1

[20] J. G. A. Barbedo, "Digital image processing techniques for detecting, quantifying and classifying plant diseases," *SpringerPlus*, vol. 2, p. 660, 2013. 1

[21] N. Fahlgren, M. A. Gehan, and I. Baxter, "Lights, camera, action: high-throughput plant phenotyping is ready for a close-up," *Current Opinion in Plant Biology*, vol. 24, pp. 93–99, 2015. 1

[22] A. Singh, B. Ganapathysubramanian, A. K. Singh, and S. Sarkar, "Machine learning for high-throughput stress phenotyping in plants," *Trends in Plant Science*, vol. 21, no. 2, pp. 110–124, 2016. 1

[23] N. Shakoor, S. Lee, and T. C. Mockler, "High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field," *Current Opinion in Plant Biology*, vol. 38, pp. 184–192, 2017. 38 Biotic interactions 2017. 1

[24] M. A. Gehan and E. A. Kellogg, "High-throughput phenotyping," *American Journal of Botany*, vol. 104, no. 4, pp. 505–508, 2017. 1

[25] M. V. Giuffrida, F. Chen, H. Scharr, and S. A. Tsaftaris, "Citizen crowds and experts: observer variability in image-based plant phenotyping," *Plant methods*, vol. 14, no. 1, pp. 1–14, 2018. 1

[26] F. Tardieu, L. Cabrera-Bosquet, T. Pridmore, and M. Bennett, "Plant phenomics, from sensors to knowledge," *Current Biology*, vol. 27, no. 15, pp. R770–R783, 2017. 1

[27] M. Bacco, P. Barsocchi, E. Ferro, A. Gotta, and M. Ruggeri, "The digitisation of agriculture: a survey of research activities on smart farming," *Array*, vol. 3-4, p. 100009, 2019. 1

[28] I. Charania and X. Li, "Smart farming: Agriculture's shift from a labor intensive to technology native industry," *Internet of Things*, vol. 9, p. 100142, 2020. 1

[29] M. A. Beck, C.-Y. Liu, C. P. Bidinosti, C. J. Henry, C. M. Godee, and M. Ajmani, "An embedded system for the automated generation of labeled plant images to enable machine learning applications in agriculture," *PLOS ONE*, vol. 15, pp. 1–23, 12 2020. 1

[30] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. I. au2, and K. Crawford, “Datasheets for datasets,” 2020. 2