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. The images taken from plants in fields are all from a top-down perspective and contain usually multiple plants per image. For these images metadata is also available. In this paper we describe both datasets’ characteristics with respect to plant variety, plant age, and number of images. We further introduce an open-access sample of the indoor-dataset that contains 1,000 images of each species covered in our dataset. These, in total 14,000 images, had been selected, such that they form a representative sample with respect to plant age and individual plants per species. This sample serves as a quick entry point for new users to the dataset, allowing them to explore the data on a small scale and find the parameters of data most useful for their application without having to deal with hundreds of thousands of individual images.

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, and its health. Plants also react in appearance to outside factors such as drought, time of day, temperature, humidity, and sunlight available. Furthermore, the correct classification of plants requires expert knowledge, which cannot easily be crowdsourced. 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. This research area running under the names of precision agriculture, digital agriculture, smart farming, or Agriculture 4.0 has the potential to increase yields while reducing the usage of resources, such as water, fertilizer, and herbicides [16–28]. This next revolution in agriculture is fuelled by data, in particular labelled image-data with rich metadata information.

In this paper 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. This dataset consists of images of individual plants, images showing several plants with and without bounding boxes, and metadata for each image and...
Table 1. Overview on available datasets.

| Dataset   | Type                              | Image Count       |
|-----------|-----------------------------------|-------------------|
| Lab-data  | Indoor grown crops and weeds      | 1.2 million images|
| Subsample | Sample of Lab data                | 14,000 images     |
| Field-data| Crops and weeds in farmland       | 540,000 images    |

Plant. At the time of writing more than 1.2 million images have been added to this dataset, since April 2020. All of these images had been captured and automatically labelled by using a robotic system as in [29]. The second dataset, the field-data, consists of images taken in the field in the growing seasons of 2019 and 2020. These images show a top down perspective on the crops (and weeds) as they grow in cultivated farmland.

We present here these two datasets with respect to some of their key metrics, such as the number of images per species imaged. Further we provide a sample from the lab-data consisting of 14,000 images. This sample (around 1% of the total data) is intended to give researchers an overview on how the data is structured with respect to available plant ages and growing stages, as well as the number of individual plants imaged by the system. For this, we carefully selected images from the entire dataset, such that the age distribution per species is preserved and a wide range of individual plants is represented. Following best-practices for data-accessibility, as outlined in [11], we give immediate open access to this data-sample, provide the metadata, and give insights on how the data was collected. In terms of long-term data storage and accessibility both full datasets will eventually be fully open-source following a data management plan of stagewise release. Our goal is to provide researchers in digital agriculture with labeled data to facilitate data-driven innovation in the field.

The rest of this paper is structured as follows: In Chapter 2 we describe the general structure of both datasets and their metadata. In Chapter 3 we visualize and describe key characteristics of the datasets, such as the number of individual plants per species. In Chapter 4 we describe the structure of the sample. We conclude the paper in Chapter 5 with information on the planned growth of the datasets and how to obtain the sample and the original dataset.

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 below.

See Figure 1 for examples on a plain image, the respective image with bounding boxes, and cropped out single images.

Each JSON-file contains information about the plain image and the respective single plant images as follows:

- version: An internally used version number
- file_name: The filename of the original image (the plain image)
- bb_file_name: The filename of the bounding box image
- singles, bb_images, source_file_path: These fields are internally used filepath information for retrieving data from an object storage.
- date: A string containing the day of imaging in the format YYYY-(M)M-(D)D. The string does not contain leading zeroes for months or days.
- time: A string containing the time of imaging in the format (h)h:(m)m:(s)s. The string does not contain leading zeroes for hours, minutes, or seconds. The timezone is CST (Central Standard Time; UTC-06:00) and CDT (Central Daylight Time; UTC-05:00), respectively.
- room, institute, camera, lens: These fields encode the location and camera-equipment used.
- vertical_res, horizontal_res: The resolution given in pixels along the vertical and horizontal image-axis, respectively.
- camera_pose: This field contains the following subfields: x, y, z, polar_angle, azimuthal_angle. The first three subfields describe the camera position in cm with respect to an origin-point inside the imagable volume of the system. The latter two subfields describe the camera’s pan (polar_angle) and tilt (azimuthal_angle). See Figure 3 for details.
Figure 1. Two images taken by the system with drawn calculated bounding boxes. Below four single plant images cropped from the first image.

Figure 2. Example of an image taken in the field, showing several Soybean plants.

- bounding_boxes: This field contains a list of elements each corresponding to one plant around which a bounding box was drawn by the system (see above for the description of bounding box images and single plant images). Each element in this list contains the same list of subfields as follows:

  - plant_id: Each plant is identified through an individual identifier. This allows to differentiate individual plants from the same breed and species.
  - label: A common name label attached to the plant, such as “Canola”
  - scientific name: The scientific name of the plant, such as Brassica napus
  - subimage_file_name: The filename of the respective single plant image. This field is used to associate the single plant images with the plain image from which they had been cropped from or to the respective JSON-file (by replacing the file-ending from .jpg to .json)
  - date_planted: The day at which the plant was planted. This information is used to determine a plant’s age, which in our case is defined as the number of days that have elapsed between planting and imaging
  - position_id: An internally used identifier for the spatial position on which the plant was located when imaged
  - x_min, y_min, x_max, y_max: The coordinates of the upper-left and lower-right corner of the calculated bounding box, respectively.

The field-data collection is in structure similar to the above. However, as there are no labels attached to individual plants we have only the following fields in use: version,
file_name, date, time, room, institute, camera, lens, label, vertical_res, horizontal_res, source_file_path. Note that the label field is with respect to the entire image, whereas for lab-data it was associated with a cropped out single plant image. The entry under label thus describes the type of crop that is cultivated on the farmland from which the image had been taken. Indeed, as imaging also took place before the application of herbicides, we can see weeds between the dominant crops on some images. Other fields in the JSON-files have the same interpretation and usage as above. An example for a field-data image is given in Figure 2.

3. Dataset Characteristics

We now describe the lab-data with respect to different metrics, followed by metrics on the field-data.

Figure 4 shows the number of images taken by the system starting from April 2020 (imaging before April 2020 took place and is available on request. It is not included in this dataset due to experimentation with the imaging system itself). As of this writing we have taken more than 446,000 plain images (that is images that contain multiple plants) and cropped out over 1.2 million single plant images from these. The system is run several times per week producing thousands of new plain images. The data acquisition rate drops significantly after October 2020 as access was restricted due to the COVID pandemic. We anticipate the acquisition rate to raise as accessibility to our facilities improve.

The third panel in 4 shows the imaged plants’ age-distribution. We define age as the number of days elapsed between seeding the individual plant and the time the image had been taken. The histogram uses a binning size of 7 days. The majority of plants had been imaged, when they were between 7 and 35 days old. This corresponds to the growth stage in farmlands at which it is critical to distinguish between weeds and crops. Thus, our emphasis on plants of this age matches the data needed for important applications in digital agriculture such as estimating germination rate of crops, quantifying weeds, and automated weeding. By our definition of age the germination time itself influences the age-value in our metadata. Since germination times vary between species, so do their age-distributions. For example, the age distribution of weeds is generally shifted towards “older” plants by one or more weeks compared to crops. Indeed, in our efforts to grow both of them, we have encountered that weeds require a longer time and more care to germinate. Most plants are only imaged towards a point at which they “outgrow” the system, i.e., where the plants’ size and shape lead to overlaps and inaccuracies when calculating bounding boxes.

Table 2 lists how many single plant images per species are in the dataset. The number varies strongly by species, which is due to availability of seeds, germination success, and access to our facilities.

In Table 3 we list how many individual plants we have
Table 2. Number of single plant images per species.

| Common Name       | Scientific Name      | Image count |
|-------------------|----------------------|-------------|
| Barley            | Hordeum vulgare      | 30597       |
| Barnyard Grass    | Echinochloa crus-galli | 76258    |
| Common Bean       | Phaseolus vulgaris   | 159217      |
| Canada Thistle    | Cirsium arvense      | 89731       |
| Canola            | Brassica napus       | 255004      |
| Dandelion         | Taraxacum officinale | 87426      |
| Field Pea         | Pisum sativum        | 68658       |
| Oat               | Avena sativa         | 59153       |
| Smartweed         | Persicaria spp.      | 99650       |
| Soybean           | Glycine max          | 203980      |
| Wheat             | Triticum aestivum    | 120417      |
| Wild Buckwheat    | Fallopia convolvulus | 24973      |
| Wild Oat          | Avena fatua          | 7065        |
| Yellow Foxtail    | Setaria pumila       | 14815       |

Table 3. Number of individual plants per species.

| Common Name       | Scientific Name      | Individual plants |
|-------------------|----------------------|-------------------|
| Barley            | Hordeum vulgare      | 14                |
| Barnyard Grass    | Echinochloa crus-galli | 21             |
| Common Bean       | Phaseolus vulgaris   | 53                |
| Canada Thistle    | Cirsium arvense      | 14                |
| Canola            | Brassica napus       | 128               |
| Dandelion         | Taraxacum officinale | 16                |
| Field Pea         | Pisum sativum        | 24                |
| Oat               | Avena sativa         | 28                |
| Smartweed         | Persicaria spp.      | 21                |
| Soybean           | Glycine max          | 84                |
| Wheat             | Triticum aestivum    | 47                |
| Wild Buckwheat    | Fallopia convolvulus | 15                |
| Wild Oat          | Avena fatua          | 3                 |
| Yellow Foxtail    | Setaria pumila       | 5                 |

Table 4. Number of field images per month.

| Year | Month | Image count |
|------|-------|-------------|
| 2019 | June  | 45954       |
|      | July  | 84033       |
| 2020 | May   | 14084       |
|      | June  | 197980      |
|      | July  | 167896      |
|      | August| 32230       |

Table 5. Number of field images per crop.

| Common Name       | Scientific Name      | Individual plants |
|-------------------|----------------------|-------------------|
| Canola            | Brassica napus       | 137551            |
| Faba bean         | Vicia faba           | 24147             |
| Oat               | Avena sativa         | 18578             |
| Soybean           | Glycine max          | 342780            |
| Wheat             | Triticum aestivum    | 19121             |

imaged per species. Again the numbers vary due to availability and germination success of seeds and access to our facilities.

We now give a short description of the field-data. The collection of field-data was performed by imaging the field via a stereoscopic camera mounted to a tractor. The camera, pointed straight down, records a video as the tractor drives through the field. We chose one of the two video channels to extract frames as images. These images form the field-data collection. Here the amount of images extracted is chosen such that consecutive images show some overlap with respect to the area imaged. We also provide the video data itself for the user to extract images under their own timing conditions or to work on the video directly. Table 4 and Table 5 give a breakdown by the number of images extracted from the videos per month and species, respectively. Field-data from the 2020 growing season is further accompanied by metadata information about temperature, wind-speed, cloud coverage, and camera height from ground.

4. Description of subsample

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

For each species listed in Table 2 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 does only contain single plant images, 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).
5. 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. After describing the data-structure we presented 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-point clouds 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.
[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.
[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.
[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.
[5] R. Oberti and A. Shapiro, “Advances in robotic agriculture for crops,” Biosystems Engineering, vol. 100, no. 146, pp. 1–2, 2016.
[6] A. Bechar and C. Vigneault, “Agricultural robots for field operations. part 2: Operations and systems,” Biosystems Engineering, vol. 153, pp. 110–128, 2017.
[7] A. Bechar and C. Vigneault, “Agricultural robots for field operations. part 2: Operations and systems,” Biosystems Engineering, vol. 153, pp. 110–128, 2017.
[8] T. Duckett, S. Pearson, S. Blackmore, and B. Grieve, “Agricultural robotics: The future of robotic agriculture,” CoRR, vol. abs/1806.06762, 2018.
[9] R. R. Shamshiri, C. Weltzien, I. A. Hameed, I. J. Yule, T. E. Grift, S. K. Balasundram, L. Pitonakova, D. Ahmad, and G. Choudhary, “Research and development in agricultural robotics: a perspective of digital farming,” International Journal of Agricultural and Biological Engineering, vol. 11, pp. 1–14, 2018.
[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,” NIAS - Wageningen Journal of Life Sciences, vol. 90-91, p. 100307, 2019.
[11] G. Lobet, “Image analysis in plant sciences: Publish then perish,” Trends in Plant Science, vol. 22, no. 7, pp. 559–566, 2017.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[19] P. Bosilj, T. Duckett, and G. Cierniak, “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.
[20] J. G. A. Barbedo, “Digital image processing techniques for detecting, quantifying and classifying plant diseases,” SpringerPlus, vol. 2, p. 660, 2013.
[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.

[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.

[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.

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

[25] M. V. Giuffrida, F. Chen, H. Schar, 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.

[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.

[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.

[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.

[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, 2020.

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