Sharing the right data right: a symbiosis with machine learning

Sotirios A. Tsaftaris [1,*], and Hanno Scharr [2]

[1]Institute for Digital Communications, School of Engineering, University of Edinburgh, Edinburgh, EH9 3FG, UK

[2]Institute of Bio- and Geosciences: Plant Sciences (IBG-2) Forschungszentrum Jülich GmbH, D-52425, Jülich, Germany

[*] Correspondence: S.Tsaftaris@ed.ac.uk (S.A. Tsaftaris), URL: http://tsaftaris.com; h.scharr@fz-juelich.de (H. Scharr), URL: http://www.fz-juelich.de/ibg/ibg-2/EN/Staff/Enabling%20Technologies/Scharr_Hanno/Scharr.html

Keywords: open data, machine learning, plant phenotyping

Abstract

In 2014 plant phenotyping research was not benefiting from the machine learning revolution as appropriate data were lacking. We report the success of the first open dataset in image-based plant phenotyping suitable for machine learning, fuelling a true interdisciplinary symbiosis, increased awareness and steep performance improvements in key phenotyping tasks.

Advancing plant phenotyping by sharing 'problems'

Appropriate training and testing data are at the heart of computer vision (CV) and machine learning (ML) research as means for developing and evaluating novel approaches. However, in 2014, appropriate data for image-based phenotyping problems were lacking. Thus, plant phenotyping was not benefiting from this data-driven revolution [1] and CV/ML researchers were largely unaware of phenotyping applications. To address these limitations, we opened a collection of plant data for several phenotyping tasks including two 'hot' CV/ML problems: leaf segmentation a 'multi-instance segmentation' problem and leaf counting an 'object counting' problem [2,3]. Our objectives were to measure how well broad ML algorithms could solve major phenotyping tasks but also enlarge the community of scientists considering phenotyping applications. Within four years our datasets became very popular, even reaching 'standard dataset' status for multi-instance segmentation and object counting. Most importantly we saw tremendous improvement in performance in solving these two tasks. Here, we report on our efforts to promote the dataset, which we consider were vital for its success and aim to offer advice on making problems interesting and corresponding data useful for the CV/ML community.

Setting the problem: the competitions

The current race for publications in ML and CV is mostly won by the 'best performing' solution and we wanted to leverage this to maximise attention to plant phenotyping problems. We, thus, organised a 'competition' (typically in the ML community refer to as 'challenge') as a rapid and visible publication route to attract CV/ML researchers. Prerequisite for a successful challenge is a challenging but doable problem, not too hard and not yet solved. Leaf segmentation and counting are ideal problems as they were amenable to several approaches and can be understood by everyone. As clarity is mission critical, clear and easy to interpret performance measures are a must. This is trivial for counting, where we decided to use average absolute count difference and average count difference. Evaluating multi-instance segmentation results is less simple. Suitable measures for single instance segmentation are well established e.g. the 'intersection of the union' score or the very similar and classical Dice score. Both range between zero and one and are thus easily interpretable as success rates being perfect at 100%. Since no well-established criteria existed when we established the competition, we introduced a multi-instance version of the Dice score. We ensured that neither reordering of instances nor exchanging roles of ground truth and test solutions have an influence on the result. The resulting 'Symmetric best Dice' measure is thus as easily interpretable as the single instance version. We made participants' lives as easy as we could: accompanying our datasets were scripts to load and pre-process data and code evaluating the proposed metrics for measuring performance. In combination, our strategy not only lowered the barrier to entry but also standardized the results and their presentation ultimately allowing direct comparison of methods (see Box 1).

Setting the stage: the workshops

To incentivise participation, we organized the first challenge in 2014 as part of a workshop accepting also full-length papers describing challenge submissions. To make this publication avenue as attractive as possible we organized the workshop together with an internationally renowned and high-ranking computer vision conference to maximise visibility. In addition, full-length papers were published together with main conference proceedings [4] –a great value for CV researchers– and extended versions were bundled in a special issue [5] of a computer vision journal.

As a community building measure and to promote application-related problems, the workshop also accepted 'problem statement' papers. These papers do not describe solutions, as is common for CV/ML conferences, but properly describe unsolved, but relevant, problems in an application area. Thus, application scientists, experts in these problems, were engaged without being experts in CV or ML.

This workshop format was so well received across the disciplines that we proceeded to organize it almost yearly in conjunction with top CV conferences (e.g. ICCV, ECCV, and BMVC (twice)).

Setting the baseline: the collation study

We compiled a collation study summarizing the results, where all challenge contributors served as co-authors [6]. This showed the quality bandwidth of the applied solutions, but also set the baseline for comparing future approaches.

Releasing the data: the paper and the website

Given the success of the first workshop and challenge we decided to publish a paper describing, for the first time, a large collection of image analysis problems that arise in plant phenotyping; and to offer accompanying data, performance metrics and several baseline methods demonstrating the challenging aspect of the data. This coincided with the continuous requests from colleagues for the challenge data and to perform evaluations on the test set. Together with the release of the paper [3], we therefore created a website[ii] allowing the download of data after registration. This registration information allows us to track the success and impact of the dataset (see Figure 1B-D).

The impact

- As of September 2018, 1600 requests were recorded, with an overall exponential growth (Figure 1B), doubling approximately every seven months.
- Approximately 70% of requests originate from users not actually working in plant phenotyping (Figure 1C), so we do attract new people and raise awareness of phenotyping.
- Undergraduate students are the largest group requesting our data (27%). This is exciting: we are introducing the problem very early in the academic development of the community. Equally encouraging is that 10% of requests are associated with industry, showing the potential exploitation impact. The remaining requests (63%) are almost equally split among researchers in higher education (MSc, PhD, postdocs/faculty).
- As of September 2018, the paper [3] has received 47 citations, according to Google Scholar. Thanks to pioneering work from Romera-Paredes & Torr [7] and Ren & Zemmel [8] our dataset has become almost a benchmark in multi-instance segmentation setting a trend for using our dataset to test and evaluate pipelines for the benefit of ML research.
- This community effort led to steep performance increases (Figure 1D). From early results at 74.4% in leaf segmentation accuracy, measuring overlap of the algorithm's result with ground truth delineation, we now reached 90% (by leveraging also synthetic data [12]). This is a remarkable 20% relative improvement within 4 years. On counting the performance gain has been even more astounding with a 3.7-fold reduction in error.

What's next

These remarkable gains are mostly seen for Arabidopsis wild-type where we had released adequate number of training data. We should now collectively focus to translate such performance on other plants and cultivars as well.

However, when do we stop and what do we have to do to get there? Typically, in CV/ML a limit is met when automated algorithms achieve or surpass human level performance. For leaf counting, a recent study showed that 0.29 is the current human expert performance [9], clearly, we are not there yet but coming closer. For leaf segmentation such study is currently lacking as is tedious to perform. Deep learning approaches need more annotated data to help improve performance yet obtaining annotated data of significant variability and size is difficult. Options to synthesise data [10-12] or use citizen scientist platforms to collect annotations can help increase scalability and is an area of active research and activity [9].

To enable continuous evaluation, we have now set up an online system[iv] that evaluates performance of approaches based on a held-out testing set. The ideal would be to use a sandbox virtual system that executes code on a dedicated server, but this requires dedicated funding. Finally, to improve performance on other cultivars, plants and tasks, appropriate training data are necessary which the community together should provide (see Box 1 for advice).

Conclusion

We want to emphasize that opening data is good practice as it allows reproducibility of science. In the plant sciences this is starting to take place. Yet we still hesitate to publish data when we don't have a good solution. This article demonstrates that potential for impact exists even if we do not yet have the perfect solution –in fact it is better to show the limitations of current approaches and demonstrate difficulty. If we open our data the 'right' way, the CV/ML community can and will solve our problems 'for free'–a true symbiosis.

Resources

i. http://www.plant-phenotyping.org/CVPPP2014
ii. https://www.plant-phenotyping.org/datasets
iii. https://www.plant-phenotyping.org/datasets-impact
iv. https://competitions.codalab.org/competitions/18405
v. https://zenodo.org/

References

[1] Minervini, M. et al. (2015) Image analysis: The new bottleneck in plant phenotyping [applications corner]. IEEE Signal Processing Magazine **32**, pp. 126–131.

[2] Scharr, H. et al. (2014) Annotated Image Datasets of Rosette Plants, Forschungszentrum Jülich, Report No.: FZJ-2014-03837, pp. 1—16

[3] Minervini, M. et al. (2016) Finely-grained annotated datasets for image-based plant phenotyping. Pattern Recognition Letters 81, pp. 80–89

[4] Agapito, L. et al. (eds.) (2015) *Computer Vision - ECCV 2014 Workshops, Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part IV*, Springer International Publishing, Vol. 8928, 396 pages

[5] Scharr, H. et al. (2016) Special issue on computer vision and image analysis in plant phenotyping. Machine Vision and Applications 27, pp. 607–609

[6] Scharr, H. et al. (2016) Leaf segmentation in plant phenotyping: A collation study. Machine Vision and Applications 27, pp. 585–606. Special Issue on Computer Vision and Image Analysis in Plant Phenotyping.

[7] Romera-Paredes, B. and Torr, P. H. S. (2016) Recurrent Instance Segmentation. Proceedings of European Conference on Computer Vision, pp 312-329. First version: Arxiv (2015) https://arxiv.org/abs/1511.08250

[8] Ren, M. and Zemel, R. S. (2017) End-To-End Instance Segmentation With Recurrent Attention. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6656-6664

[9] Giuffrida, M. V. et al. (2018) Citizen crowds and experts: observer variability in image-based plant phenotyping. Plant methods **14**, pp. 1–14.

[10] Giuffrida, M. V. et al. (2017) ARIGAN: Synthetic arabidopsis plants using generative adversarial network. Proceedings of the 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), Venice, Italy, pp. 22-29.

[11] Zhu, Y. et al. (2018) Data Augmentation using Conditional Generative Adversarial Networks for Leaf Counting in Arabidopsis Plants. Computer Vision Problems in Plant Phenotyping (CVPPP2018) pp. 1–11.

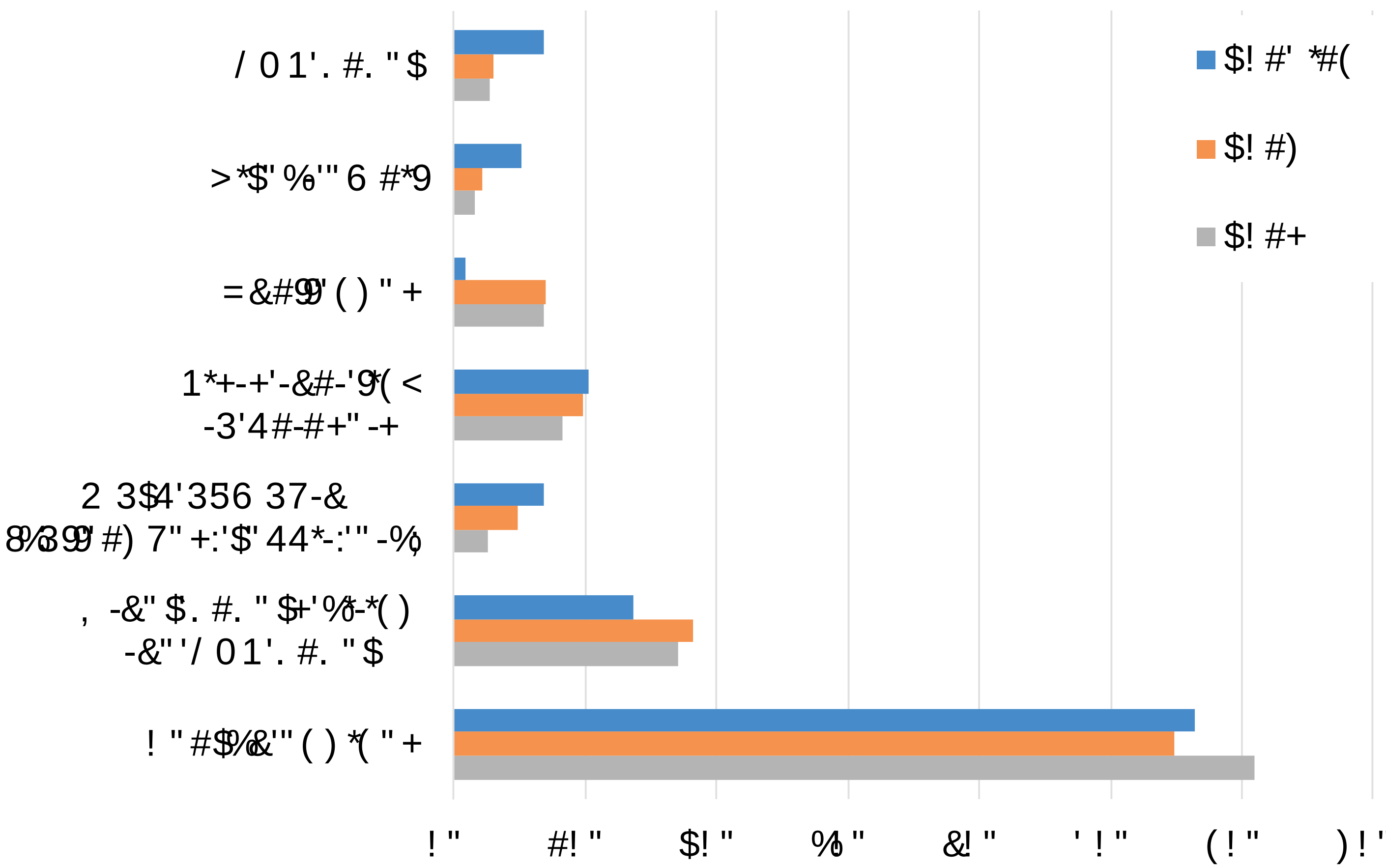
[12] Ward, D. et al. (2018) Deep Leaf Segmentation Using Synthetic Data. Computer Vision Problems in Plant Phenotyping (CVPPP2018) pp. 1–13.

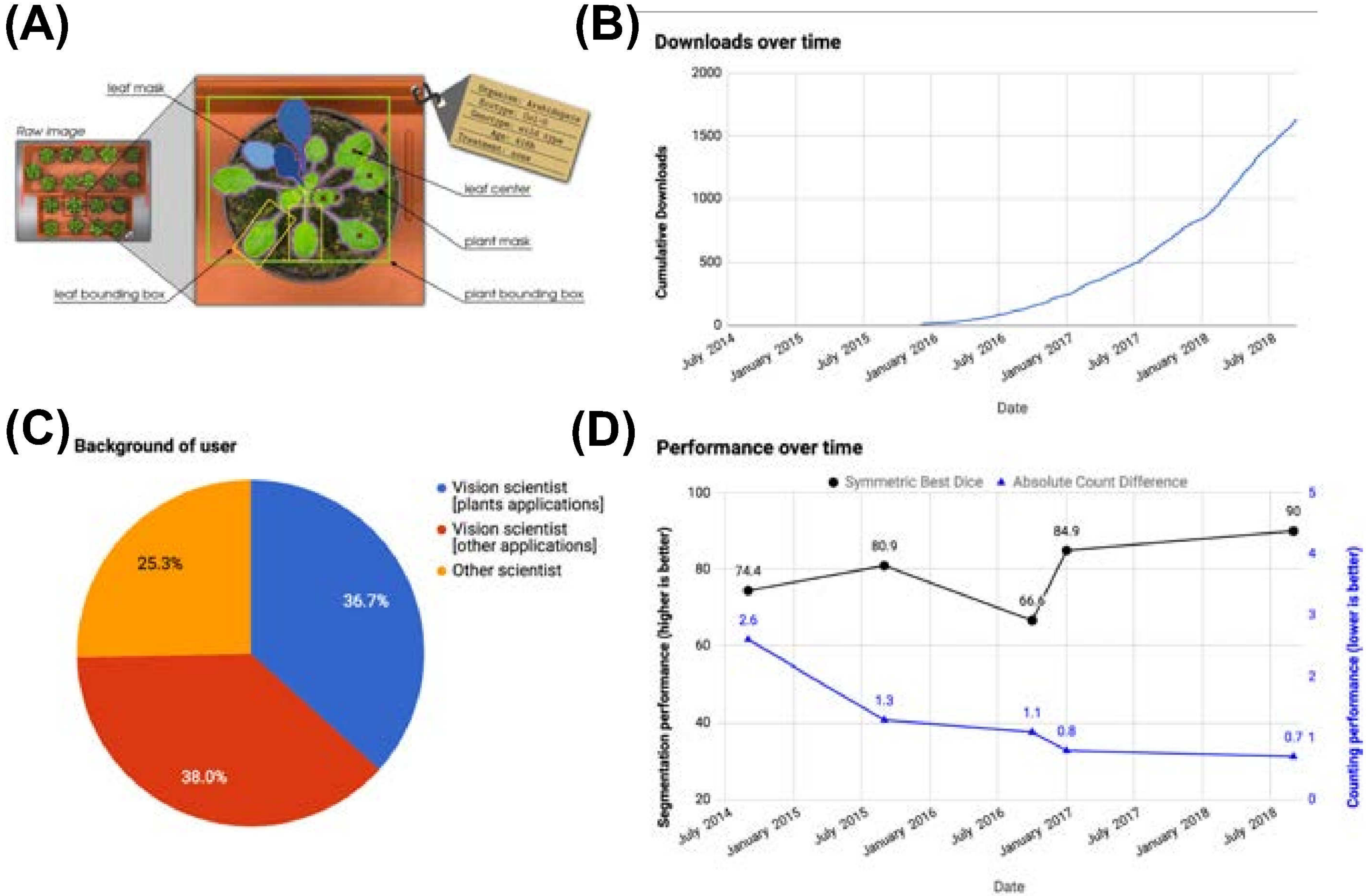
Figure 1. (A) An example of the content and annotations available in the dataset (figure adapted and modified from [3]); (**B)** Cumulative download requests since release (December 2015, axis starts earlier for visualisation purpose); **(C)** Background (expertise) of users downloading the data over 1600 requests; (**D)** Performance in leaf segmentation (Symmetric Best Dice) and leaf counting (absolute count difference) as evolution of time, showing exemplars of performance taken from papers that use the datasets (see also our website on impact[iii]). Early points before the dataset release refer to challenge contributions.

Box 1. How to share data successfully by making it useful for others

- *Open the RIGHT data for the RIGHT problem (neither undoable nor solved)*: Publish data not just as means of verification but as "problem statements" [2,3] and give a baseline [6]. Observe the balance: if the baseline performs already too well, it is not a challenging problem.
- *Appropriate data for the problem*: Data without accompanying annotations are not very useful for the machine learning era of today.
- *Use terminology attractive for the intended audience:* For example, leaf segmentation is a multi-instance object segmentation.
- *Adhere to Findable, Accessible, Interoperable, and Re-usable (FAIR) principles:* We opened the data, described their origin, metadata, how to read/write, and how to share findings.
- *Decide on suitable metrics:* Ideally a single, well established, and easy to interpret measure should be provided to evaluate performance.
- *Offer implementations of error metrics and keep it standardised:* Open training data sets together with ground truth, and test sets without. Offer code for metrics, especially for those non-trivial to compute such as the multi-instance version of the Dice score. Perform test set evaluations for participants. It promotes standardization and credibility of results.
- *Pick the right sharing platform:* We build our own website and used a survey form to collect requested information (which essentially powered this paper's analysis, see website on impact[iii]).
- *Work on disseminating the dataset and the problem*: see Figure I.

Figure I. How users heard about the data, shown as percentage over three periods. The first four options correspond to actions that we had direct influence over, whereas the rest were out of our hands. **PRL paper:** We observe that publishing a tech report [2] satisfies the FAIR data but is just an entry point. A peer-reviewed paper gave us better visibility [3]. **Challenges:** We organised workshops and challenges, were inclusive and rallied the community with the collation study paper [6]. **Invitation:** We directly emailed people after building our own contact list and ensured that the data are listed in relevant databases (**Lists**). This is what fed the initial growth, which of course changed completely as more people discovered the data as soon as the first "high-impact" computer vision paper cited the dataset. Then growth over the years was largely fuelled by **papers citing the PRL paper** [3] and of course by **search engines**, so having a website helps immensely. (The search engine numbers could be inflated as search may have been the means and not the origin.)




(A)
Raw image
leaf mask
leaf center
plant mask
plant bounding box
leaf bounding box
(B)
Downloads over time
Cumulative Downloads
2000
1500
1000
500
0
July 2014
January 2015
July 2015
January 2016
July 2016
January 2017
July 2017
January 2018
July 2018
Date
(C)
Background of user
Vision scientist [plants applications]
Vision scientist [other applications]
Other scientist
36.7%
38.0%
25.3%
(D)
Performance over time
Symmetric Best Dice
Absolute Count Difference
Segmentation performance (higher is better)
Counting performance (lower is better)
74.4
80.9
66.6
84.9
90
2.6
1.3
1.1
0.8
0.7
Date