push in all changes in July 14th

Reviewer #1: The manuscript by Cruz et al. presents a new image dataset of Arabidopsis and bean plants, acquired using different imaging modalities (namely, fluorescence, infrared, RGB colour, and depth). Plant material, experimental setup, sensor calibration, and annotation procedure are described. To set baseline performance on this new dataset, a leaf segmentation and tracking algorithm previously published by some of the authors is adopted.

The dataset presented in this manuscript is timely and very relevant to both the computer vision community and the plant community as well. Description of experimental protocol is rather complete, however, the manuscript presents several issues that must be addressed for it to be considered for publication in this journal.
Please find below a list of my main concerns/comments.

Major comments:

- The text (including the abstract) contains many typographical and grammatical errors, which are distracting and render the manuscript difficult to read. The authors should proofread the document carefully.

ALL

- In the introduction, four example applications are mentioned for which the dataset can be used. According to the title of the paper the dataset is proposed for phenotyping applications. Thus, the authors should elaborate on why and how the computer vision tasks they mention are relevant to plant phenotyping. The authors should also include few key references to works that use (or discuss) such features for plant phenotyping.

Jin/Jeff

- In relation to the previous comment, the authors should clarify what is intended by "leaf alignment" and why is important, since this feature is probably less common in plant phenotyping applications than the others that are mentioned.

Liu

- Since multi-modality is a strength of the proposed dataset, the authors should motivate better their choice of imaging modalities (i.e. fluore scence, infrared, RGB color, and depth), highlighting their importance and what information they convey on plant structure and functions. Apparently, this is only partly done for depth at the end of Sec. 2.

Jin/Jeff

- Fig. 1 should be augmented (or a separate figure should be created) with zoom in details showing how the same plant parts (or scene portions) appear in the different modalities. This would help the reader get a better idea of the contrast and the type of information conveyed by each of the modalities adopted by the authors.

Liu

- Each of the databases mentioned in Table 1 should be accompanied by a reference.

Liu

- The second paragraph of Sec. 2 begins with "We summarize all existing publicly available databases that are related to plant imagery". This is a strong statement, particularly because other plant related image datasets exist (e.g., ImageCLEF, ICL leaf database) which are omitted from the review. The authors should adjust language and/or expand the literature review.

Liu

- Sec. 2, p. 3, end of second paragraph. I believe the observation on single leaves imaged in a constrained environment does not apply to the dataset by (Haug and Ostermann, 2014). Please reformulate.

Liu

- It appears from Sec. 3.1 that all plant subjects (respectively for Arabidopsis and bean) belong to the same genotype and that no treatments were performed. This would entail that the proposed dataset cannot be used to investigate computer vision algorithms or imaging modalities in relation to group differences. Could the authors comment on this?

Jin

- In Sec. 3.2.1 the authors remark that light used for night image acquisition does not influence plant development. A reference should be included to support this statement.

Jin

- As for the other lighting and imaging protocols adopted by the authors (Sec. 3.2.1 and 3.2.2), do they interfere with plant growth?

Jin

- At the end of the first paragraph of Sec. 4.4, p. 8, the authors envision a use of the multiple modalities in their dataset in which algorithms are developed to handle missing modalities. Briefly presenting a use case scenario or citing relevant works may help clarify the importance of this point.

Liu

- In Sec. 5 no baseline method or result is reported for bean images. Could the authors comment on this?

Liu

- The conclusions (Sec. 6) should be expanded.

Liu

Minor comments:

- The acronym "MSU-PID" should be clearly defined when it first appears in text (p. 2, line 51).

- According to the SI (International System of Units), units of measure should be written in roman type, while italic type is reserved for variables. Besides, when they follow a number, a space should be included between numerical value and unit symbol. To improve clarity of the manuscript, the authors are therefore advised to adhere to the SI style conventions.

- In Sec. 3.2.1, the "a" of "chlorophyll a" could be italicized to improve clarity.

- It is not clear on p. 8, line 15, if the size of the database is 380 megabytes (MB) or 380 megabits (Mb). In the former case "MB" should be used instead of "Mb", while in the latter case the authors should consider reporting the value in megabytes (MB) which is more common.

- In Eq. (3), p. 8, the authors should better define the symbols used. In particular, it is not clear if subscripts 1 and 2 refer to inner and outer leaf tips respectively, and in which order.

- Sec. 5.1, p. 9. For completeness the authors should mention the approach they used to find a threshold for plant segmentation.

Liu

Reviewer #2: The manuscript presents a collection of imagery within a plant phenotyping context to support the development of computer vision algorithms. The paper's main strength and novelty is the introduction of a multimodal database with annotations. However, it does have some issues that need to be clarified prior to publication.
Those issues are discussed below.

==Originality and overlap with other works==
Overall the paper appears, although some aspects of the work have appeared in previous publications of the authors. However, what the authors present here can stand in it self and appears complete.

==Major Strengths==
* A collection of multimodal image database in two different plants.
* Appropriate annotation and evaluation metrics.
* If also the annotation tool is (will be publicly) available, this will serve as an additional plus.
* Thorough description of methodology
* An "example use case" included
* Although it is not clear how (and where) the data will be shared, given that they will be publicly available it will have a significant impact in the field of plant phenotyping

==Major weakness and issues==
A) The authors argue that the most closely related work is that of Scharr et al 2014. They identify as weaknesses that it i) uses only RGB images and ii) can be used for a few vision problems. However, upon reading the Scharr paper it is clear that the authors of that work, do mention that they collect additional data to be used in tracking context and for other applications. It is not clear if the authors here refer to what is available data from Scharr et al or what is described in the paper? Furthermore, the data in Scharr et al appear to originate from different mutants and under different treatments, and also imaged with different cameras. Since this work does use only wildtypes and single cultivars this distinction should be made. Furthermore, once leaf labelings are available several secondary annotations can be derived so in some sense Table 1 should be annotated accordingly.
Nevertheless, I commend the authors for arranging information as in Table 1

Liu

B1) The authors do a nice work of calibrating the cameras and measuring noise in the depth camera. However, in page 2 introduction say that this calibration allows for the explicit correspondence between pixels of an y modality. The authors rely on this to annotate data in one modality and propagate labels in the other (at least this is what I understand from later on description of methodology). However, this is a VERY strong assumption and depends completely on the distance between the cameras, the angles, the distance between object and sensors and the actual object arrangement. From our experience even when imaging co-planar plants (such as young arabidopsis) at a distance of ~70cm even when the camera sensors are really close to each other (less than 5cm) some differences in view are there and occlusions are present. The authors should comment on this and should show as supportive evidence examples of plants at different growth stages in all 4 modalities in raw and annotated form to show how close this correspondence is matching and how the propagated labels. Furthermore, additional supporting evidence could be obtained either by arranging for two external and blind annotators to label some data (different age, different placement in the tray to show the effect of angle) in another modality (e.g., optical) and then measure inter-observer variability. Then they can propagate annotations from fluorescence to images of that modality and measure agreement. If this agreement is better than the in between rater variability then you could argue that propagating annotations is ok to do and actually beneficial. Nevertheless, you should definitely mention this limitation of your work.

Daniel

B2) Furthermore, it assumes that cameras are perfectly synchronized which i) is not mentioned and ii) it cannot be done for some modalities since the same camera (with different filter) is used so some delay is expected. Granted the plants may not move in between but this should be clarified and mentioned.

Jin

C) Fig 3 (a) ... From Fig 2 it appears the distance from plant to sensor to be greater than 60mm (6cm) !!! ie., to me it looks close to 60cm, but either Fig 3 has wrong axis range or something else is going on. Can you please explain/update?
Also same figure (3), shouldn't the images overlap? why are shown translated?

Daniel

D) On table 3 you list resolutions for the acquired image data. This image resolution is much lower than the LSC database of Scharr et al. This limitation should be mentioned.

Jin

E1) On manual annotation process: Did you use an extra annotator or a supervisor?

Daniel

E2) Your annotation process is completely interactive. Any way to "save interaction" by interactive segmentation approaches? I applaud your comment on the difficulty of merging super-pixel results, but I would add at least somewhere that easier ways to annotate will be beneficial.

Liu

E3) It is unclear if you release this annotation tool? If not, please do ... it will be extremely useful for the community.

Liu

E4) Again going back to the problem of lack of exact 1-1 correspondence between modalities, how can you guarantee that labels on one are good on the other given such high differences in resolutions among modalities? Please include at least some visual examples.

Daniel

F) Does the annotation _hour_YY_label.png contain a label for each pixel or only for the boundaries? If yes, then can you obtain single boundary definition in between overlapping leaves?

Liu

G) Based on your annotations can you find which leaf occludes which?

Liu

H) The authors test an algorithm (developed by them) on the fluorescence part of the data. Granted this algorithm is presented elsewhere on the same data. I was wondering if the authors can either a) use the same algorithm on one of the other modalities, and how it would perform (assuming ground truth labelings obtained via propagation)? OR b) if they can apply another benchmark method even from the broad CV literature on one of the CV problems considered in another modality. I think doing one of the two, time permitting, will greatly assist the paper.

Liu

==Minor issues==

A) I do not subscribe to the term dense for such low resolution depth cameras. In my book dense refers to high res depth maps. Maybe the authors can reconsider the use of the term.

Daniel

B) I am not sure the authors used the correct reference for Erblichkeit 1903. I think Erblichkeit is part of the title. A google search reveals this:
http://caliban.mpipz.mpg.de/johannsen/erblichkeit/
With the author being: Wilhelm Ludwig Johannsen
Please check

C) page 4, line 4, maybe you should put Chlorophyll a either in quotes or italics so a is not confused as an article?

D) Figure 2 if possible can you please also include an image showing the cameras from the view of the plant (LED off)?

Jin

E) Collimated ... is a nice special term, but please put in parenthesis the definition to help broad readership.

F) Please also define object albedo again to broaden readership

G) couple of typos:
page 6, line 27 2nd col, please add "are" before present
page 7, lines 56-69, left col, the last two sentences should be together and not be separated by a period, but by a comma
page 8, line 30 (left col) consider using performance instead of performances