Response letter on MVAP-D-15-00159, "Multi-modality Imagery Database for Plant Phenotyping"

July 27, 2015

We thank the Guest Editors for the efforts of handling our submission and two reviewers for providing the constructive and valuable feedback. The detailed comments help us further improve the quality of our manuscript. In the following, we list our detailed responses to all the comments. For your convenience, we add a separate reference list in the end that is used in this response letter.

REVIEWER 1

Comment 1: *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.*

Response: We have proofread it very carefully and corrected some typos and grammar errors.

Comment 2: *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.*

Response: We have added two paragraphs with related references to better connect plant phenotyping and computer vision tasks. Please refer to paragraph 2 and 3 in Sec. 1.

Comment 3: *The authors should clarify what is intended by "leaf alignment" and why it is important, since this feature is probably less common in plant phenotyping applications than the others that are mentioned.*

Response: Leaf alignment aims to estimate the structure of a leaf. In the case where the leaf structure is simple (e.g., ellipsoid-shaped leaf), leaf alignment amounts to estimating the inner and outer tips of the leaf. For more complex leaf structure (e.g., maple leaf), leaf alignment will estimate additional landmarks on the leaf contour that will consistently appear in all leaves. Leaf alignment is important for two reasons: First, it goes beyond leaf segmentation and provides structural information of a leaf - a discriminative signature for a leaf. For example, [1] uses a polygonal leaf model for leaf segmentation and plant identification. Second, plant biologists often hypothesize that different local parts of a leaf may have different photosynthetic efficiency. This leaf structure information can be used to quantitatively study this hypothesis and its implication to the overall photosynthetic efficiency of a plant.

Comment 4: *Since multi-modality is a strength of the proposed dataset, the authors should motivate better*

their choice of imaging modalities (i.e. fluorescence, infrared, RGB color, and depth), highlight 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*.*

Response: We have inserted a brief explanation with references in the last paragraph of Sec. 2.

Comment 5: *Figure* 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 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.*

Response: We have augmented Figure 1 by adding one zoom-in view for one Arabidopsis plant. Since bean dataset only contains one plant with good view, we do not provide a zoom-in view.

Comment 6: *Each of the databases mentioned in Table* 1 *should be accompanied by a reference.*

Response: We have the corresponding references in Sec. 2, which follows the order as they appear in Table 2. To make it complete, we also added the references to Table 2.

Comment 7: *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 databases exist (e.g., ImageCLEF, ICL leaf database) which are omitted from the review. The authors should adjust language and/or expand the literature review.*

Response: We have adjusted the language by the statement of "we summarize existing publicly available databases that are most related to our work".

Comment 8: *In the second paragraph in Sec.* 2*, 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.*

Response: It is correct that the dataset (Haug and Ostermann, 2014) is not about single leaves imaged in a constrained environment. We still category it in the first type because it does not include leaf segmentation. All segmentation tasks are based on plant level. We have reformulated this in the second paragraph in Sec. 2.

Comment 9: *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?*

Response: It is true that we do not compare group differences, based on morphology (for example). However, we are using this database to address a fundamental issue with visual phenotyping: accurate and automated identification and tracking of individual leaves over developmental time scales (weeks). The reason this is important is highlighted in other sections, in particular in text we are adding to address Comment 2 and 4. The inherent challenge is that leaves change in size, position, and shape as they emerge and grow. Also they may overlap or be overlapped by other leaves. We emphasize more on leaf identification and tracking rather than developing methods/algorithms to define group differences. We have added one paragraph in

Conclusion to better reflect this.

Comment 10: *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.*

Response: We have revised paragraph 2 in Sec. 3.2.1 and added 2 references to support this statement.

Comment 11: *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?*

Response: The depth sensor uses a a flash near IR illuminator to obtain depth estimates. The near IR illuminator will not influence the plant growth. The color image does not use any special lighting so it is not harmful as well. The IR and fluorescence images use a saturation light, which will damage the plants only if at high frequency. However, as we are taking images once per hour, the influence is minimal and can be ignored.

Comment 12: *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.*

Response: In this use case scenario, the training part of leaf segmentation algorithms may have both the RGB and depth modalities, while during the testing, we may only have the RGB modality available. This scenario is useful because it leverages additional information during the training, while does not increase the hardware/sensor cost during the testing. Specifically, we may use two types of approaches to implement this scenario. One is called learning with side information, which is the depth modality in this example (see [2]). The other is to use transfer learning that handles missing modality in the target domain (see [3]).

Comment 13: *In Sec.* 5 *no baseline method or result is reported for bean images. Could the authors comment on this?*

Response: The previous developed leaf alignment, segmentation, and tracking algorithm is designed particularly for rosette plants like Arabidopsis and tobacco. The rosette plant structure is used as a prior in the development of the algorithm. However, bean plant does not belong to rosette plants. Therefore, we did not apply the baseline method to bean plants. We have added this statement in the beginning of Sec. 5.

Comment 14: *The conclusions (Sec.* 6*) should be expanded.*

Response: We have expanded the conclusions.

Comment 15: *The acronym "MSU-PID" should be clearly defined when it first appears in text (p.* 2*, line* 51*)*

Response: Except in the Abstract, the "MSU-PID" first appears in paragraph 4 of Sec. 1. We have clearly defined it as short for Michigan State University Plant Image Database.

Comment 16: *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.

Response: We have changed the inaccurate annotations to adhere to the SI style conventions.

Comment 17: *In Sec.* 3.2.1*, the "a" of "chlorophyll a" could be italicized to improve clarity.*

Response: We have changed "chlorophyll a" to "chlorophyll *a*" all over the paper.

Comment 18: *It is not clear on p.* 8*, line* 15*, if the size of the database is* 380 *megabytes (MB) or* 380 *megabits (Mb).*

Response: It is 380 MB. We have changed it in the text.

Comment 19: *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.*

Response: t_1 represents the outer leaf tip and t_2 represents the inner leaf tip. This representation is the same as the released leaf tip labels as illustrated in Sec. 4.4. The equation does not rely on the correspondence of inner or outer tips as long as the estimated tips correspond to the labeled tips.

Comment 20: *Sec.* 5.1*, p.* 9*, for completeness the authors should mention the approach they used to find a threshold for plant segmentation.*

Response: We use some training images with groundtruth plant segmentation masks to learn a threshold for binary segmentation. The learnt threshold is then applied to both training images and testing images to generate the image mask. We have added this statement to the paper.

REVIEWER 2

Comment 21: *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 labeling is available several secondary annotations can be derived so in some sense Table* 1 *should be annotated accordingly. Nevertheless, I recommend the authors for arranging information as in Table* 1*.*

Response: We refer to Scharr et al 2014 that the Leaf Segmentation Challenge database only includes independent images without tracking context. Please refer to Comment 9 for the reason we do not include mutants.

Comment 22: *The authors do a nice work of calibrating the cameras and measuring noise in the depth camera. However, in page* 2 *introduction says that this calibration allows for the explicit correspondence*

between pixels of any 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 ~ 70 *cm even when the camera sensors are really close to each other (less than* 5 *cm) 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.*

Response: We add one more section (Sec. 4.3) in the paper to describe the labeling and the propagation to other modalities. The data labeling is performed in the fluorescence image modality. The infrared image is from the same camera and so pixel labeling will be identical. The labels are propagated to the RGB color and depth modalities using planar homographies. The homography between the fluorescence image and the RGB color image is calculated by fitting matching features in these modalities, and the homography to the depth image is found by fitting a plane through the 3D plant points. As the reviewer noted, using these homographies will result in pixel segmentation errors due to parallax for plant regions outside of the modeled planes. We evaluate these errors qualitatively and quantitatively. Figure 6 illustrates the label propagations from the fluorescence image to the other modalities and from the color image to other modalities with two different line colors. Errors occur where the lines fail to overlap and are due to propagation effects as well as boundary selection variations by the labeler between modalities. We compare the magnitude of these two effects in Table 4. The performance of label propagation is only slightly worse than the performance of another annotator. Therefore, we propagate the fluorescence labels to other modalities for Arabidopsis plants. As for bean plant, the homograph-based mapping performed poorly due to the large within-plant depth variations. So we only provide the labels for bean plant in fluorescence and IR as they use the same camera. Please refer to Sec. 4.3 for more details.

Comment 23: *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.*

Response: We have this clarification in the paper. In Sec. $3.2.3$, we describes our image collection process, where the capture of all four modalities takes 4 minutes. During this short period, no substantial movement or growth were observed.

Comment 24: *Figure* 3 *(a) ... From Figure* 2 *it appears the distance from plant to sensor to be greater than* $60\ mm$ *(*$6\ cm$*) !!! ie., to me it looks close to* $60\ cm$*, but either Figure* 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?*

Response: Yes, the distance to the plant is roughly $620\ mm$. The image planes are plotted not at the plant location, but at a distance proportional to the focal length of each camera. We added dashed lines to the

figure to make it clearer that these planes do not correspond to the plant depth, and updated the caption to better explain it.

Comment 25: *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.*

Response: We added one paragraph in the end of Sec. 4.1 to address the trade-off between image resolution and experiment throughput.

Comment 26: *On manual annotation process, did you use an extra annotator or a supervisor?*

Response: The released label is labeled by one annotator. However, we do use an extra annotator to label several images and the annotation error between different annotators is evaluated. As shown in Table 4, the average annotation similarity in case of *SBD* score is around 0.847 for fluorescence and 0.858 for RGB images. Different annotators have more agreement on the labeling as the plant grows. This is due to the evaluation metric *SBD*, which favors bigger leaves.

Comment 27: *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.*

Response: Our annotation is not interactive. It is totally based on user input without any computer feedback except for visualization. Leaf label is implemented by clicking several points along the boundary of the leaf. Tip label is implemented by clicking the outer and inner tip points respectively. We did explore other options for labeling and found this simple naive way worked the best.

Comment 28: *It is unclear if you release this annotation tool? If not, please do. It will be extremely useful for the community.*

Response: We will release the labeling tool together with the database.

Comment 29: *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.*

Response: Please refer to Comment 22.

Comment 30: *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?*

Response: The annotated image is an image the same size as the original image. And each pixel is annotated with a number. The same number represents the same leaf and 0 represents the background. For overlapping leaves, we can visualize the relative location and annotate accordingly. Please refer to Figure $9(b)$ for examples of the labeled images.

Comment 31: *Based on your annotations can you find which leaf occludes which?*

Response: During annotation, we can observe which leaf occludes which and annotate the overlapping area

to the leaf on top. However, it is hard to estimate which leaf occludes which from the annotations.

Comment 32: *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 labeling 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.*

Response: We have applied our algorithm on the RGB images while the groudtruth label is generated via label propagation. The performance on RGB images is shown in Figure 9 and 10.

Comment 33: *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.*

Response: We agree and have removed the term "dense" from the depth map image.

Comment 34: *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.*

Response: It is right that Erblichkeit is part of the title and the author's name is Wilhelm Ludwig Johannsen. We have corrected it.

Comment 35: *In page* 4*, line* 4*, maybe you should put Chlorophyll a either in quotes or italics so a is not confused as an article.*

Response: We have put it in italics to avoid confusion.

Comment 36: *Figure* 2 *if possible can you please also include an image showing the cameras from the view of the plant (LED off)?*

Response: Figure 2 is updated with extended details. Please refer to the paper.

Comment 37: *Collimated ... is a nice special term, but please put in parenthesis the definition to help broad readership.*

Response: "Collimated" means to make the light rays parallel. We have added this definition in the paper.

Comment 38: *Please also define object albedo again to broaden readership.*

Response: Daniel

Comment 39: *Couple of typos:*

1. *page* 6*, line* 27 2*nd col, please add "are" before present*
2. *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*

3. *page* 8*, line* 30*, left col, consider using performance instead of performances*

Response: We have corrected those typos.

References

[1] Guillaume Cerutti, Laure Tougne, Julien Mille, Antoine Vacavant, and Didier Coquin, "Understanding leaves in natural images–a model-based approach for tree species identification," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1482–1501, 2013.

[2] Jixu Chen, Xiaoming Liu, and Siwei Lyu, "Boosting with side information," in *Proc. Asian Conf. Computer Vision (ACCV)*, pp. 563–577. Springer, 2013.

[3] Zhengming Ding, Shao Ming, and Yun Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, 2014.