I-RIM Conference 2019
October 18 - 20, Rome, Italy
ISBN: 9788894580501
DOI: 10.5281/zenodo.4782481

Weakly Supervised Fruit Counting for Yield Estimation using Spatial Consistency

1st Enrico Bellocchio
Department of Engineering
University of Perugia
Perugia, Italy
enrico.bellocchio@gmail.com

2nd Gabriele Costante
Department of Engineering
University of Perugia
Perugia, Italy
gabriele.costante@unipg.it

3rd Thomas A. Ciarfuglia
Department of Engineering
University of Perugia
Perugia, Italy
thomas.ciarfuglia@unipg.it

4th Paolo Valigi
Department of Engineering
University of Perugia
Perugia, Italy
paolo.valigi@unipg.it

***Abstract*—Fruit counting is a fundamental component for yield estimation.Most of the SotA approaches address the problem using fruit models or by explicitly learning to count. In this paper, we tackle the problem by proposing a framework that learns to count fruits without the need for task-specific supervision labels. The experiments on different varieties of fruits show that our approach reaches performances that are comparable with SotA approaches based on the supervised paradigm.**

***Index Terms*—Agricultural Automation, Computer Vision for Other Robotic Applications, Deep Learning in Robotics and Automation, Robotics in Agriculture and Forestry, Visual Learning.**

I. Introduction

Among the multitude of agricultural processes that draw the attention of computer science and robotics researchers, an important role is certainly played by yield estimation. An accurate estimation of the yield of a culture facilitates the farmer efficiently in planning for harvesting operations and crop sales. Despite its importance, the standard practice to yield estimation often relies only on coarse measurements and direct inspection, a practice that has high costs and low accuracy.

In this work we aim to remove the need for explicit instance or density labelling. We achieve this by proposing a weakly supervised deep architecture that relies only on an image level binary classifier, *i.e.,* the sole supervision label that we need is whether the image contains instances of the fruit or not. Since this information by itself is not sufficient to allow the network to learn to count, we propose a novel objective function that imposes consistency between the image level classifier predictions computed at different spatial locations and scales. We demonstrate that our approach reaches performance that are comparable to fully supervised baselines with respect to three different fruit species, namely apples, almonds and olives.

II. Contribution

In this work, we give a new twist to the fruit counting problem by removing the strong assumption of having labelled data samples to train the object count models. We only require a simple image level binary classifier that predicts whether the image contains instances of the fruit or not, *i.e.,* we do not need the number of objects in the image or bounding box labels or density maps during training. To achieve this, we build an objective function that combines the classifier output at different locations and scales of the image with a spatial consistency term. This forces the model to learn to count without the need for task specific supervision signals.

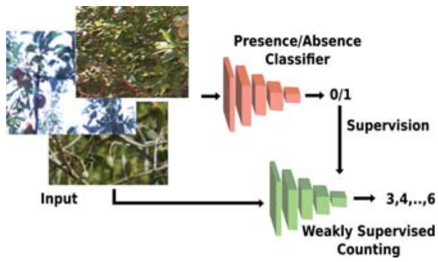

Fig. 1: Our weakly supervised counting network. The weak supervision comes in the form of a simple binary presence classifier, that requires less data than any fully supervised method.

With an extensive set of experiments on different fruit varieties (olives, almonds and apples) we show how our weakly supervised approach is able to achieve a performance similar to its fully supervised counterpart and to SotA approaches. Furthermore, we release a new dataset containing images of olive groves with ground truth information for comparison and further research.

III. Proposed Approach

Ideally we want to remove this label and have a network that is able to learn from the images what and how to count. Stated like this the problem is ill-posed, because we should give the network at least a slight hint about what to learn. For this reason, we introduce the multi-branch counting CNN (MBC-CNN) that operates on different image sub-windows at different levels. More precisely, it works on three scales, the whole image, the image divided into quadrants, and the image divided into 16 parts. While we do not use the correct number of fruits in the image as a supervisory signal, we impose the constraint that at each scale the total count regressed on the corresponding tiles must be consistent with the total count of other levels.

Since these labels cannot be used naively to train a counting network, we introduce an image level binary classifier, which will be referred to as PAC (Presence-Absence Classifier), and use it to train the actual counting network. The key intuition, in addition to the counting consistency, is to force consistency between the output of each counting branch and the prediction of the PAC. If the classifier predicts the presence of object instances, the counter should output a number greater than zero. Conversely, when the absence of fruits is estimated, the count must be zero.

IV. Experiments

A. Datasets

The proposed WS-COUNT framework is evaluated with respect to three different fruit species: apples, almonds and olives. For apples and almonds we use the datasets provided by [2]. Hence, we collected and manually labelled a set of images of olive trees.

To build the olive dataset we collected 28 high resolution images with 5456×3632 pixels by using a high quality camera. The images capture the full olive tree shape. We processed the images by randomly picking smaller tiles of 606×403 pixels to obtain a dataset of 1402 images, of which 1298 were used for training and 104 for testing and bounding box labelling was performed.

B. Baselines

To prove the effectiveness of our approach, we compare it against four different baselines. As a fully SotA supervised baseline we use the approach proposed by [2], which is trained on bounding-box instance labels. This method uses the Faster-RCNN object detector to count the fruit instances in an image. In addition to this, we consider also two supervised neworks trained in an end-to-end fashion by using the instance counts as supervision signal. The first one is the S-COUNT architecture, that is a CNN trained to solve directly the counting task in a regression fashion and using directly the RMSE loss. Since WS-COUNT exploits a multi-branch structure, we decided to compare it also against a multi-branch version of S-COUNT (which we named MBS-COUNT) to provide a fair comparison.

Finally, since the PAC network is used as a source of supervision, we asked ourselves whether simply counting the binary prediction of the presence-absence classifier (PAC) could provide good estimates.

C. Results and Discussion

To evaluate the performance of both the baseline methods and our approach, we compare their count estimates with the ground truth value by using the RMSE metric. We start our discussion by commenting the average RMSE obtained by each model over the whole test sets for each fruit dataset. The results are presented in Table I. It can be observed that the best performance are achieved by [2]. This is to be expected, since their model is specifically trained with the most informative labels (bounding boxes on instances).The S-COUNT and MBS-COUNT networks give higher, but still comparable, errors with respect to [2], showing that end to end counting on total number of instances is effective. The most important result is that WS-COUNT, despite being trained in a weakly supervised manner, achieves performances that are close to the supervised baselines. It is also important to observe that the errors obtained by the MB-PAC-Only baseline are considerably higher than WS-COUNT, which proves that the combination of the classifier consistency and the spatial consistency losses gives the network a better capability to count the fruit instances.

Method	**Olives**	**Almonds**	**Apples**
Bargoti et al. [2]	1.57	1.90	1.22
S-COUNT	2.03	3.40	2.00
MBS-COUNT	1.91	3.24	1.88
MB-PAC-Only	3.30	3.65	2.22
WS-COUNT	2.44	3.61	2.03

TABLE I: The Table shows the RMSE on the test sets.

V. Conclusion

In this work we proposed a novel weakly-supervised framework for fruit counting in agricultural applications. The WS-COUNT strategy is able to learn to count without requiring task-specific supervision labels, such as manually labelled object bounding boxes or total instance count. The experiments run on three different fruit species clearly show that our approach guarantees performances that are comparable to those of fully supervised baselines.

References

[1] M. Stein, S. Bargoti, and J. Underwood, "Image based mango fruit detection, localisation and yield estimation using multiple view geometry," *Sensors*, vol. 16, no. 11, p. 1915, 2016.

[2] S. Bargoti and J. Underwood, "Deep fruit detection in orchards," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 3626–3633.

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

[4] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in neural information processing systems*, 2010, pp. 1324–1332.

[5] M. Rahnemoonfar and C. Sheppard, "Deep count: fruit counting based on deep simulated learning," *Sensors*, vol. 17, no. 4, p. 905, 2017.

[6] S. W. Chen, S. S. Shivakumar, S. Dcunha, J. Das, E. Okon, C. Qu, C. J. Taylor, and V. Kumar, "Counting apples and oranges with deep learning: A data-driven approach," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 781–788, 2017.

[7] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in *2017 IEEE International Conference on Computer Vision (ICCV)*, vol. 00, Oct. 2018, pp. 5899–5907.

[8] T. Durand, T. Mordan, N. Thome, and M. Cord, "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017.

[9] S. Bargoti, "Pychet labeller - an object annotation toolbox." 2016. [Online]. Available: https://github.com/sbargoti/pychetlabeller

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.