Automated visual fruit detection for harvest estimation and robotic harvesting

Steven Puttemans[1], Yasmin Vanbrabant[2], Laurent Tits[3] and Toon Goedemé[1]
[1] EAVISE, KU Leuven - Campus De Nayer, Jan Pieter De Nayerlaan 5, BE-2800 Sint-Katelijne-Waver, Belgium
e-mail: [steven.puttemans, toon.goedeme]@kuleuven.be
[2] KU Leuven, Department of Earth and Environmental Sciences, Division Forest, Nature and Landscape,
Celestijnenlaan 200E, BE-3001 Leuven, Belgium, e-mail: yasmin.vanbrabant@kuleuven.be
[3] Flemisch Institute for Technological Research (VITO), Remote Sensing Unit, Boeretang 200, BE-2400 Mol, Belgium
e-mail: laurent.tits@vito.be

***Abstract*— Fully automated detection and localisation of fruit in orchards are key components in creating automated robotic harvesting systems. During recent years a lot of research on this topic has been performed, either using basic computer vision techniques, like colour based segmentation, or by resorting to other sensors, like LWIR, hyperspectral or 3D. Recent advances in computer vision present a broad range of advanced object detection techniques that could improve the quality of fruit detection from RGB images drastically. We suggest to use an object categorisation framework based on boosted cascades of weak classifiers to implement a fully automated semi-supervised fruit detector and demonstrate it on both strawberries and apples. Compared to existing techniques we improved fruit detection, mainly in the case of fruit clusters, using a supervised machine learning instead of hand crafting image filters specific to the application. Moreover we integrate application specific colour information to ensure a more stable output of our fully automated detection algorithm. Finally we make suggestions for efficient fruit cluster separation. The developed technique is validated on both strawberries and apples and is proven to have large benefits in the field of automated harvest and crop estimation.**

***Keywords*— Object Categorisation, Application Specific Constraints, Integrated Pre-filtering, Autonomous Harvesting**

I. Introduction

Autonomous robotic harvesting is a rising trend in agricultural applications, like the automated harvesting of fruit and vegetables. Farmers continuously look for solutions to upgrade their production, at reduced running costs and with less personnel. This is where harvesting robots come into play. While the mechanics of grabbing objects is a well documented problem with many proposed solutions, robustly localizing objects stills seems to be very challenging, due to natural variations in shape and size, occlusion and uncontrolled lighting conditions. Multiple solutions, discussed in more detail in section II, tried tackling the detection and localisation of single object instances, yielding only moderate success.

Object categorisation as a technique is well studied in the field of computer vision. However outside of the conventional applications like face or pedestrian detection, other fields of research could also benefit from the power of these techniques, able to detect an object class including its intra-class variation (shape, size, colour, texture, etc.), instead of sticking to basic segmentation based approaches. Therefore we propose a fully automated semi-supervised (during the object model learning part) system, able to identify unique object instances in unseen images. Furthermore we improve the detection of single object instances within clusters, by suggesting two approaches for separating clusters into individual object instances. Finally we speed up the complete process using scene specific knowledge.

II. Related Work

The state-of-the-art in automated fruit detection and localisation focusses either on 2D segmentation based approaches like [12], adds extra information to the problem using either a thermal LWIR camera [13], a hyperspectral sensor [7] or even the power of 3D scanning techniques [15] to add an extra layer of knowledge. The downside of adding extra layers of information is the increase in computational complexity and thus these techniques become quite time consuming, especially when calculating of a full high density 3D point cloud. Furthermore using hyperspectral imaging demands expensive hardware, with only limited resolution at reasonable prices. Therefore one of the main goals of our research is to prove that 2D information can be sufficient for efficient object detection and localisation. Furthermore, many of these additive techniques are quite depending on very controlled (lab) environment settings, that are quite impossible to achieve in open air or greenhouse agricultural set-ups.

The segmentation based approach of [13] uses thermal imaging (LWIR) for the counting and analysis of apple cultivars in orchards. Their research is based on the different IR radiation between fruit and leafs. To ensure a decent detection accuracy, images need to be acquired in the afternoon, to achieve a large temperature gradient between the apples and the background. This is a downside when creating a universal applicable approach since fruit exposed to less direct sunlight, results in partial and incomplete apple detections. Subsequently, a simple colour channel based segmentation approach is used, requiring the definition of hard thresholds, which could vary over time. This is further improved in [12] where the possibility of using pure RGB based colour and shape segmentation for apple detection and analysis is investigated. Wherever the basic segmentation does not work, specific image transformations on separate colour channels

are applied to achieve a better contrast between leafs and fruit. Finally they apply a parameter based blob analysis to identify fruit instances. This works fine for fully visible object instances, but fails automatically when objects get partially occluded. In comparison with our approach, object categorization techniques are able to detect partially occluded instances, since these variations are also included in the training data, and thus ensure that more objects will be found.

The work of [15] uses a 3D stereo set-up, applying colour based segmentation on the intensity image, successfully separating clusters on tomato plants. We would like to retrieve individual objects, which is impossible for this cluster based segmentation approach, mainly due to the fact that separate instances in a cluster share a similar depth profile.

Finally [5] proposes a fruit detection and classification method for a strawberry harvesting robot, closely relating to one of our test cases. They introduce the use of OTHA colour spaces, on which they apply segmentation based algorithms to extract strawberries from the background, which is proved to work in single object instances. However. expanding this to a harvesting set-up on a real robot, is nearly impossible, because the segmentation will most likely not work due to object clusters, occlusions of leafs and different lighting conditions. Similar research [3] in lab conditions is performed on individual object instances, against a clean background, successfully segmenting and analysing the apple region.

The downsides of segmentation based approaches can be countered by the proposed object categorisation techniques. These techniques ([14], [2], [4]) train an object model from training data including intra-class variation (e.g. shape, colour, occlusion, defects). This allows for a more variate and robust detection in the wild. The work of [11] on recognising and counting peppers in cluttered greenhouses, based on a bag of visual words combined with a sliding window approach, is a first attempt of using more advanced computer vision techniques for solving the task. They start by locating fruit in individual images, then aggregate the estimates from multiple images using a novel statistical approach to cluster repeated and incomplete observations. Compared to our technique, where only a single view from the set-up is needed, this approach can only work successfully if multiple views are provided to support the different observation hypothesis.

The work of Viola and Jones on cascades of weak classifiers has previously proved to be very efficient in industrial object detection set-ups ([10], [9]), and is thus an ideal technique to explore for our proposed solution. Furthermore the technique suits our need of incorporating scene specific knowledge to improve the detection accuracy.

III. Datasets

Since no publicly available annotated datasets of fruit in unconstrained conditions exist, the first step in implementing our object detection algorithm, is gathering the necessary data for training and testing the specific object models. This research focusses on two separate cases: strawberry picking and apple harvest estimation. Positive training data is gathered containing different representations of the strawberries and apples (e.g. different illumination conditions, different orientations, partially occluded by branches and leafs, different viewpoints, ...) whereas for the negative training data the objects are removed and the remaining image pixels are used as background information, maintaining application specific background knowledge. This does limit the ability of using the trained models outside the intended set-up but it decreases the training time immensely compared to learning a set-up independent object model.

For the strawberry picking case, the goal is to provide the location of all ripe strawberries given an RGB colour input image of the scene. A trinocular stereo set-up is used in order to grab different viewpoints (bottom-up and side-view) of the strawberries. For the apple harvesting estimation application, two apples cultivars are tested: Gala and Red Delicious. For both cultivars a training dataset is gathered using side-views of the apple trees inside the orchards. For the evaluation of the apple detection models, images with a thirty degrees angle inclination where used, giving us a clear separate test set, maintaining the objectivity in evaluating the models. For both apple cultivars, models were trained using the original RGB data and separate models using the transformed data based on our scene specific constraints, as discussed in section IV-B.

Table 1 gives a detailed overview of the train and test images collected, the number of annotated objects, the model dimensions and the amount of negative samples used at each stage of weak classifiers. All annotations were manually made by a domain expert. For the strawberry case, images were captured with two AVT Manta cameras both having a 1292×964 pixel resolution. For the apple cultivar case, images were captured using a Samsung NX3000 with a 3648×5472 pixel resolution.

IV. Suggested Approach

In this section we describe the complete pipeline for building both the strawberry and the apple detector models, and prove that using scene constraints increases the accuracy of the detection output. The approach is mainly developed on top of the strawberry case, whereas the apple case is used to verify the approach for other agricultural cases.

Table 1. Data overview for both applications: number of images, number of annotations, model dimensions and the amount of negative window samples.

		strawberry		appleGala		appleRedDelicious	
		train	test	train	test	train	test
pos	#images	205	750	30	30	32	32
	#labeled	1500	/	1595	625	1075	1160
	dimensions	35x38		65x65		62x66	
neg	#images	200	/	30	/	30	/
	#windows	5000	/	4000	/	3000	/

Fig. 1. Strawberry model trained with both ripe and unripe strawberries.

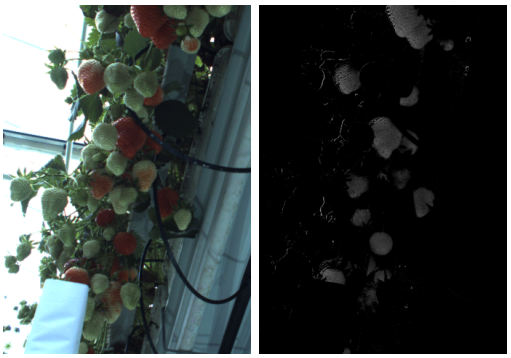

Fig. 2. I_{RG} colour transformation used to segment ripe strawberry areas.

A. A cascade of weak classifiers based on greyscale data

To start off we build a cascade classifier model using AdaBoost [14] with local binary patterns [6] as local feature descriptor. This type of classifier ignores colour information and focusses on gradient information instead, by comparing regions of pixel intensities in the greyscale image. Furthermore each input sample, evaluated by the learned model during both training and test phase, is preprocessed using a histogram equalisation step to account for varying lighting conditions. For each image a set of manually positioned annotations is supplied, from which the local feature descriptors are learned used to separate the positive and negative training data.

In this context we trained two models using the strawberry training data. We first build a model using all strawberries (both ripe and unripe) annotations. This results in a model, able to separate strawberries from the background and the plant with mediocre results, as shown in Figure 1. However, our focus is to separate ripe from unripe strawberries. Training a second model where the unripe strawberries are used as negatives failed, since colour information is ignored by the boosting process based on greyscale feature descriptor.

We tested a similar approach on the apple case, since ripeness was not an issue here. We noticed that training a model on greyscale data yielded quite a lot of false positive detections, while missing actual apples, as seen in Figure 3(a). This is mainly due to the simple shape and structure of an apple, yielding feature descriptors that are not unique enough compared to the background information. This is a problem that can only be solved by adding more scene or application specific information, like the colour of the fruit.

B. Adding scene specific information to improve the detector

To avoid valuable colour information being ignored and to reduce false positive detections, we investigate possibilities to include colour information as an extra filter. This idea is inspired by the work of [2], where adding colour descriptors to the learning process increases the performance of pedestrian detection drastically. The addition of colour should make it possible to separate between ripe dark red and unripe green strawberries and to post-filter valid red apple detections, in order to drop the amount of false positive detections. To incorporate colour information for our set-up, we derive an application specific colour transformation seen in equation 1.

$$I_{RG} \begin{cases} 0 & if\ I_R - I_G < 0 \\ I_R - I_G & if\ I_R - I_G > 0 \end{cases} \quad (1)$$

The derived equation supports red coloured regions, while it ignores the greener regions, by subtracting the green channel I_G from the red channel I_R and clipping negative values. This boils down to projecting the RGB colour cube on an axis connecting these two colours G(0,1,0), which represents leafs, branches and unripe strawberries and R(1,0,0), representing ripe strawberries. This idea can be applied to any coloured object with a distinct colour difference to the background. Applying the equation results in an image with a visible difference between ripe and unripe strawberries on the one hand and background information on the other hand. This is visualised in Figure 2, and only applies if the background does not consists of bright red colours (greenhouse conditions).

The gain of information obtained by performing the colour transformation can be used in two ways. It can be used as a post-processing filter after applying a general colour ignorant object detector, as seen in Figure 3(b). This works in case of the strawberry case, where building a cascade classifier for ripe strawberries only is not possible. After each colourless detection, a post-processing filter can define if in the I_{RG} image the response is high enough to decide that it is an actual ripe strawberry. We apply an Otsu thresholding [8] on the I_{RG} image to receive a binary image and calculate the amount of white pixels. If more than 50% of the detected pixels yield a response in the binary map, we allow the detection as a ripe strawberry. This post-detection candidate removal, leads to an increase in processing time due to an extra filter step, but reduces the amount of false positive detections. To avoid this increase in processing time, we integrated the colour based knowledge as a pre-filter, by retraining the object model on both I_{RG} transformed positive and negative data sets. Compared to using a post-processing filter this results in faster processing, a higher amount of true positive detections and a lower amount of false positive detections (Figure 3(c)).

Since our object detection model is applied on a multi-scale basis, we restrict the object size search range of the

(a) Model trained on RGB input data.

(b) Model trained on RGB input data with I_{RG} postprocessing.

(c) Model trained on I_{RG} preprocessed input data.

Fig. 3. Adding scene knowledge to the detection process as post- or preprocessing step.

object detector, due to the known distance of the camera set-up compared to the object itself. This again results in a speed-up and a decrease in false positive detections, since multiple layers of the processed scale pyramid during multi-scale detection can be ignored in the search for object candidates. As seen in Figure 3, it only makes sense to detect apples in a certain range, depending on the distance to the camera, removing any false positives of undesired scales.

C. Splitting object clusters into separate object instances

A reoccurring issue during fruit localisation is the existence of fruit clusters. Segmentation based approaches are in most cases unable to find the small borders between connecting objects. This is where our object detection framework comes in handy. Since overlapping and partially occluded objects are also part of the positive training set, the model is able to identify single objects within clusters. This means that the output of the detector can be used to successfully separate clusters. In this paper we suggest two possible approaches following our object detection pipeline, that use the location of the found detections to further segment object clusters.

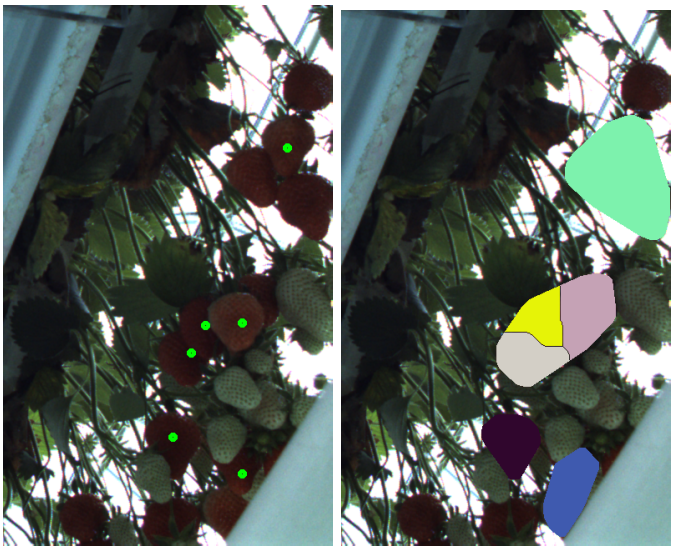

Fig. 4. Watershed based segmentation for separating object clusters.

*1) **Watershed based segmentation**:* We use the Otsu thresholded I_{RG} image, containing blobs with possible ripe strawberry pixels. The object detection returns regions of interest containing individual strawberry detections. White pixels falling together in a single object detection, are merged together, to cope with the negative effect of the colour transformation where clutter covering parts of the object, splits single object instances into separate blobs. Once a merged binary image is produced, the centres of all detections are used as initialisation positions for a watershed based segmentation [1], combined with a separate randomly positioned seed point for the background. The watershed will split larger blobs into separate objects and identify them with a unique ID, as illustrated in Figure 4. The borders of each region are defined by the merged binary image, pinpointing areas with possible strawberry pixels. A downside of the watershed based approach can be the harsh boundaries between two consecutive objects, but this heavily depends on the watershed implementation used.

*2) **Tinocular stereo triangulation based segmentation**:* In order to avoid the harsh boundaries of the watershed segmentation we propose a second solution, based on a calibrated trinocular stereo set-up, like in our strawberry case. By performing the strawberry detector in all three camera views

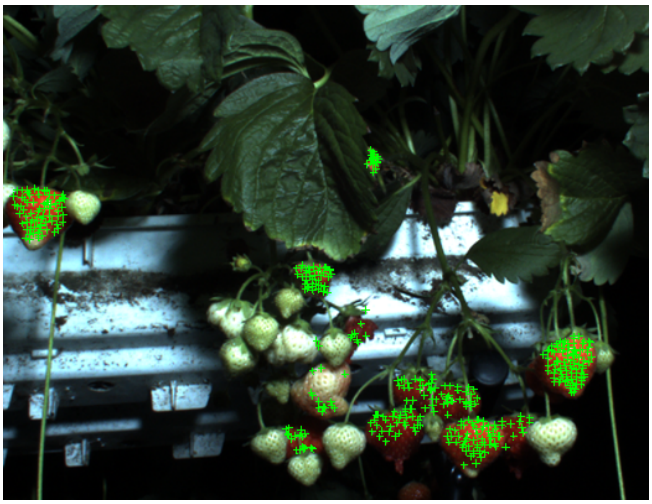

Fig. 5. Applying DoG filtering to identify seed positions inside detections.

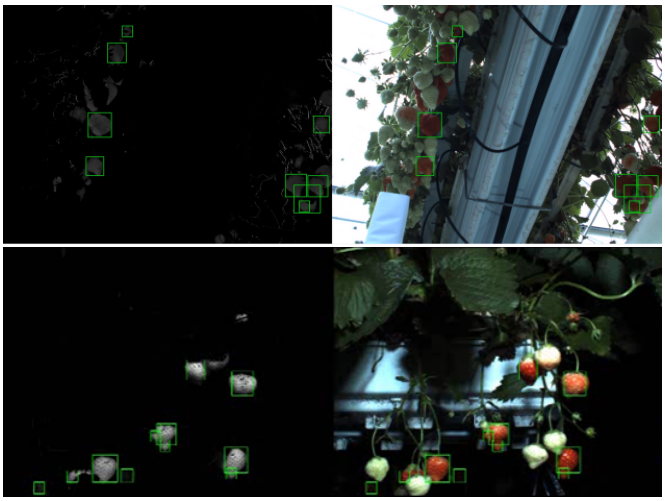

Fig. 6. Visual detections for the strawberry picking case.

we know where strawberry are located in the given 2D images. We then perform a Difference of Gaussians (DoG) filter on the found regions (see Figure 5), which successfully locates the strawberry seeds. By performing 3D triangulation on those seeds inside detections, a local 3D map of the strawberries can be generated. Combined with the segmentation data from the I_{RG} image, the depth edges can then be efficiently used to separate strawberries in clusters. Evidently this approach only works for objects that have identifiable unique textures, which yield enough unique points for the 3D triangulation. In case of flat texture-less objects the watershed based segmentation approach will still achieve better results.

V. Results

A. Results on the strawberry test case

Due to the lack of decent ground truth data provided on the strawberry test sets, producing objective accuracy results on the produced object detectors is not that straightforward. One of the main challenges lies in defining what we have to classify as a ripe strawberry and what not. This differs from strawberry to strawberry cultivar, and thus introduces bias when people start annotating ripe strawberries. Even application experts have troubles uniquely defining a ripe strawberry using objective criteria. However visual results, as seen in Figure 6, clearly show that we are able to accurately detect visible and partially covered strawberries. Furthermore, all 'pick-able' strawberries are clearly located. Much depends on the extend of the used training set, and the quality of the annotation inside that given set of images. We clearly notice that using application-specific colour information improves the detection output and that we are able to uniquely identify objects inside object clusters, which can be used as seed points for our proposed segmentation approaches. A video of the validation pipeline on a strawberry test sequence can be seen at `https://youtu.be/XFa11nF63gk`.

B. Results on the apple cultivar test case

In contrast to the strawberry case, the apple cultivars test sets are accompanied by a complete set of manually provided object annotations. This allows us to perform a qualitative accuracy analysis in the form of precision-recall curves, as seen in Figure 7. For both apple cultivars, Gala and Red Delicious, we evaluated both greyscale (striped curves) and I_{RG} (filled curves) models on the available test data. The green curves show the results of the models trained for the Gala cultivar, while the red curves correspond to the Red Delicious cultivar. We clearly prove that adding the I_{RG} colour transformation to the model training pipeline improves the detection performance of the model compared to using the raw greyscale input image data. This is compared to a pure segmentation based approach, seen as the black curve inside Figure 7, which proves to be much worse then our

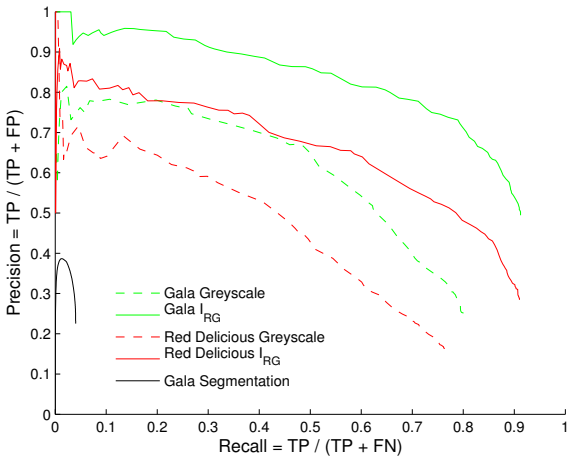

Fig. 7. Precision - Recall curves for both apple cultivars.

suggested approaches. For this we applied hard thresholds on the I_{RG} image, applied some erosion and dilation operators to remove noise, followed by a blob detection algorithm. Each detected blob is evaluated on its dimensions, ensuring that only blobs inside a specific scale range are allowed. Visual results in Figure 8 clearly show a well performing algorithm in challenging conditions like occlusion, background clutter, etc.

VI. Discussion and conclusion

While existing segmentation based object detection approaches seem to work for agricultural applications in very constrained 'lab' conditions, they tend to fail when the environment gets more challenging. Adding extra sensors like IR, hyperspectral or 3D sensors could cope with this, but have proven to work less efficiently in outdoor conditions than 2D image processing. This was our motivation for developing a robust and promising technique for agricultural object detection. By smartly combining an object categorization framework based on a boosted cascade of weak classifiers, with scene and application specific pre-filtering and efficient cluster segmentation, we created a promising pipeline for applications like strawberry picking, apple harvest estimation, etc. Furthermore, the lack of publicly available and annotated datasets, limits the possibility of comparing to existing techniques in the field of fruit picking and harvest estimation.

We acknowledge that several parts of the pipeline could be further improved. For the detection part, we could investigate the influence of adding more training data and tweaking the different training parameters. Considering today's tend, deep learning techniques could be considered as the next step. They seem to promise very high detection rates on cases like face and pedestrian detection. However it is known that these techniques require huge amounts of training time and processing power to be able to learn an object model. Improving the separation of clustered objects should be a primary focus in future work. We can already provide a very robust centre position of found objects, but getting the exact outline would help building the robot picker, in order to avoid damaging as less fruit as possible. We also discovered the problem of annotation bias, whereas the annotator needs to be well informed on what objects to annotate corresponding to the training data used for the model. If not, the annotations can lead to very misleading results on the precision recall curves, showing worse results than the actual accuracy of the trained object detector. It became clear that defining the actual objects to be found can be very challenging, even for domain experts.

Fig. 8. Visual results for Gala (first row) and Red Delicious (second row) test images.

Acknowledgement

This work is partially supported by the KU Leuven, Campus De Nayer and the Flanders Innovation & Entrepeneurship (AIO). We would like to thank Octinion, for providing the strawberry datasets. Furthermore we would like to thank VITO and the Research Center for Fruit, for sharing apple data.

References

[1] A. Bleau and L. J. Leon. Watershed-based segmentation and region merging. *Computer Vision and Image Understanding*, 77(3):317–370, 2000.

[2] P. Dollár, Z. Tu, et al. Integral channel features. In *BMVC*, 2009.

[3] S. R. Dubey and A. S. Jalal. Detection and classification of apple fruit diseases using complete local binary patterns. In *ICCCT*, pages 346–351. IEEE, 2012.

[4] P. F. Felzenszwalb, R. B. Girshick, et al. Cascade object detection with deformable part models. In *CVPR*, pages 2241–2248. IEEE, 2010.

[5] G. Feng, C. Qixin, et al. Fruit detachment and classification method for strawberry harvesting robot. *International Journal of Advanced Robotic Systems*, 5(1):41–48, 2008.

[6] S. Liao, X. Zhu, et al. Learning multi-scale block local binary patterns for face recognition. In *Advances in Biometrics*, pages 828–837. Springer, 2007.

[7] H. Okamoto and W. S. Lee. Green citrus detection using hyperspectral imaging. *Computers and Electronics in Agriculture*, 66(2):201–208, 2009.

[8] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[9] S. Puttemans and T. Goedemé. Visual detection and species classification of orchid flowers. In *MVA*, pages 505–509. IEEE, 2015.

[10] S. Puttemans, T. Goedemé, et al. Automated walking aid detector based on indoor video recordings. In *EMBC*, pages 5040–5045. IEEE, 2015.

[11] Y. Song, C. Glasbey, et al. Automatic fruit recognition and counting from multiple images. *Biosystems Engineering*, 118:203–215, 2014.

[12] D. Stajnko and Z. Čmelik. Modelling of apple fruit growth by application of image analysis. *Agriculturae Conspectus Scientificus*, 70(2):59–64, 2005.

[13] D. Stajnko, M. Lakota, et al. Estimation of number and diameter of apple fruits in an orchard during the growing season by thermal imaging. *Computers and Electronics in Agriculture*, 42(1):31–42, 2004.

[14] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages I–511. IEEE, 2001.

[15] L. Yang, J. Dickinson, et al. A fruit recognition method for automatic harvesting. In *M2VIP*, pages 152–157. IEEE, 2007.