Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 59-23 (2025) 257–262

Evaluation of next-best-view planning to deal with occlusions in tomato plants

Gert Kootstra * **Akshay K. Burusa** *

* *Agricultural Biosystems Engineering, Wageningen University and Research, Wageningen, The Netherlands (e-mail: gert.kootstra@wur.nl, akshaykburusa@gmail.com).*

Abstract: Robots can address challenges in agriculture. However, agricultural environments are complex, resulting in significant perception challenges due to occlusions. This paper evaluates a semantic-aware next-best-view (NBV) planner designed to improve the detection of tomato plant nodes for harvesting and deleafing tasks. The method actively selects viewpoints based on expected information gain, utilizing a probabilistic Semantic OctoMap, integrating spatial attention and semantic awareness. The method was evaluated in a real-world greenhouse, analyzing its robustness against higher occlusion levels, uninformative viewpoints, and reduced field-of-view. Experimental results demonstrated that the Semantic NBV planner outperformed baseline planners, achieving a better reconstruction with fewer viewpoints.

Keywords: Agricultural robotics, Perception and sensing, Active vision, Next-best-view planning, Object detection, Semantic scene understanding

1. INTRODUCTION

Agriculture is facing big challenges, with the need to increase production, while reducing the environmental impact and dealing with labour shortages. Robots can contribute to the mitigation of these challenges. However, most agricultural robotic solutions are not commercially available yet. One of the biggest reasons for this is that current robots cannot deal well enough with the complex agricultural environment, which is characterized by high variability and clutter (Kootstra et al., 2020). Specifically, for the robot's perception system, this results in the challenges of variation and incomplete information (Kootstra, 2023). Many of the current agricultural robots perceive the environment passively or using a pre-defined set of viewpoints. In a complex agricultural environment, this results in important information often being occluded, hampering the task execution of the robot. This paper, instead, focuses on the active capabilities of a robot to change viewpoint, in order to get the required information to execute the tasks at hand.

An occluded object is hidden from the camera view because another object is in front of it, or because of self occlusion. In cluttered environments, such as orchards and greenhouses, occlusions pose a significant challenge, as relevant information is often not observable for the robot. For a plant-phenotyping use-case, for instance, Boogaard et al. (2020) showed that 36 viewpoints per plant were needed to reliably observe the leaf and fruit nodes of cucumber plants. Similarly, for a bell-pepper harvesting use-case, Hemming et al. (2014) showed that at best 69% of the fruits were visible in the images, even with a requirement of only 50% visibility.

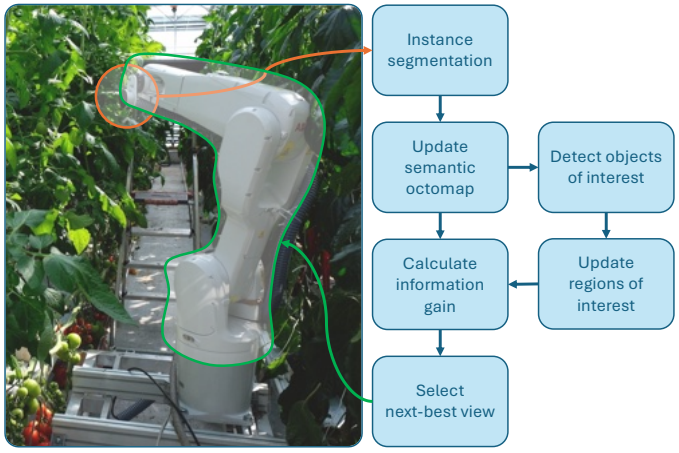

Fig. 1. Overview of the semantic NBV planner.

Observing the scene from multiple viewpoints can form a solution. By combining five viewpoints, the bell-pepper detection rate, for instance, could be increased to 90% (Hemming et al., 2014). In (Sa et al., 2017; Barth et al., 2016) fixed scan paths were used to get a more complete 3D reconstruction of the to-be-harvested fruits. For fruit-counting tasks, multi-object tracking methods have been developed to combine information from multiple camera images (Smitt et al., 2021; Halstead et al., 2021; Rapado-Rincón et al., 2023, 2024). Although these multi-view approaches alleviate the problem of occlusion, it is not guaranteed that the required information is captured as the sets of viewpoints were pre-defined.

Instead of relying on a pre-defined set of viewpoints, active perception methods have been shown to better deal with occlusions (e.g. Lehnert et al., 2019; Zapotezny-Anderson and Lehnert, 2019). In particular, next-best-view (NBV) methods show great opportunities, as they plan the next

⋆ This work was financed by the Dutch Research Council through the TTW Perspectief program "FlexCRAFT" (P17-01).

10.1016/j.ifacol.2025.11.796

viewpoint based on the current reconstruction, in order to optimize information gain (Zaenker et al., 2021; Burusa et al., 2024b). Although also indirect, deep-learning-based methods exist (e.g. Zeng et al., 2022; Ci et al., 2024), typically, a three-dimensional (3D) voxel space is used to represent the uncertainty in the 3D reconstruction of the scene, with unobserved parts having the maximum uncertainty. The next view that provides the most information, calculated using ray tracing, is then selected.

In (Burusa et al., 2024a), we proposed a novel NBV method, which includes spatial attention and semantic awareness. Where earlier work dealt with efficient reconstruction of the complete workspace, our Semantic-Aware NBV planner focusses on the efficient detection and reconstruction of specific objects. The method is based on a probabilistic representation of 3D space, including the uncertainty on the classification of detected objects. Initially, a large region of interest (RoI) is set to indicate the area where the objects are expected, which becomes more refined over time when more information about the scene has been gathered, in order to guide attention to the most likely locations to find objects of interest. The uncertainty and RoIs are used in the NBV planning to calculate the expected information gain (IG) for a set of candidate viewpoints, to then select the candidate with the highest IG as the next-best view. We showed that this resulted in improved performance in the detection of fruit and leaf nodes in tomato plants (Burusa et al., 2024a). In that study, the proposed method was evaluated thoroughly in simulation, with an additional real-world experiment as proof-of-principle, leaving some questions about the practical considerations to use NBV planning in the real world.

In this paper, we provide an comprehensive evaluation of the Semantic NBV method in the real world in a tomato-harvesting and deleafing use-case. The task of the NBV method is to detect, classify and localize the fruit and leaf nodes of tomato plants, which can be used in downstream tasks to harvest the fruits and remove the leaves. To provide insight in the functioning of the method, we investigated a number of important aspects: (i) the difference between fruit-node and leaf-node detection, (ii) how the level of occlusion relates to the detection performance, (iii) how important the selection of candidate viewpoints is, and (iv) what the relationship is between the camera field of view and the NBV performance.

2. MATERIALS AND METHODS

2.1 Semantic next-best-view planner

This subsection provides a high-level description of the Semantic NBV Planner. For a detailed description, we refer to (Burusa et al., 2024a).

Figure 1 gives an overview of the Semantic NBV Planner. Every iteration, the planner gets an observation from the RGB-D camera (Intel Realsense L515). The colour image is processed using Mask R-CNN (He et al., 2017), which was trained to detect the fruit and leaf nodes. Combined with the depth information, this gives a semantic point cloud, which is then used to update the probabilistic 3D scene reconstruction using the *Semantic OctoMap*

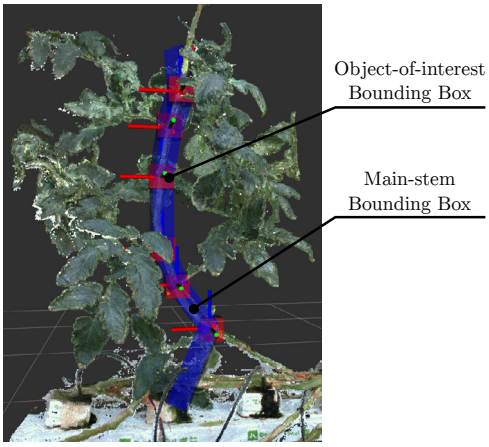

Fig. 2. Attention mechanism to guide view planning. Blue shows the RoIs around the main stem, red indicates RoIs on the plant nodes.

(Section 2.1.1). In the OctoMap, the objects of interest (OOIs) – the fruit and leaf nodes in our use-case – are detected by clustering the voxels using OPTICS (Ankerst et al., 1999), with the minimum number of points set to 20 and the maximum distance to 40mm, based on the size of a single node. These OOIs are then used to update the set of RoIs (Section 2.1.2). Using the updated Semantic OctoMap and RoIs, the information gain is then calculated for a set of candidate viewpoints, resulting in the candidate with the highest IG to be selected as the next viewpoint (Section 2.1.3).

Semantic OctoMap: The 3D workspace is represented using OctoMap (Hornung et al., 2013), a probabilistic occupancy map. In our experiments, we used a resolution of 3 mm for a voxel. Each voxel stores the occupancy probability, with a value of $p_o(x) = 0$ for an empty voxel x, $p_o(x) = 1$ for certainly occupied, and values in between for uncertain reconstructions, with a maximum uncertainty for $p_o(x) = 0.5$. Unobserved voxels are not represented in the OctoMap, and are considered unknown with $p_o = 0.5$. The occupancy probabilities are updated when new sensor measurements are received. When a sensor measurement indicates that a voxel is occupied, the occupancy probability is increased by 0.7 in a Bayesian manner, and reduced by 0.4 when a sensor measurement indicates that a voxel is free (Hornung et al., 2013).

In addition to the occupancy information, the OctoMap stores semantic information. Apart from the occupancy probability, each voxel holds a semantic class label, $c_s(x)$ and an associated confidence score, $p_s(x)$. For voxels that are observed for the first time, the class label and confidence are set based on the corresponding Mask R-CNN output. If a voxel was already in the Semantic OctoMap, the values were updated with the Mask R-CNN output using the max-fusion method (for details, see Burusa et al. (2024a)). Note that $p_s(x)$ is updated for all points related to the objects detected by Mask R-CNN. All other observed points are considered background, $c_s(x) = -1$, and get $p_s(x) = 1$, to make sure that the information gain for those is zero (see Eq. 2).

Regions of Interest: Figure 2 illustrates the RoIs that guide the attention of the Semantic NBV Planner. The red

RoIs are placed around the nodes (the OOIs) detected in the Semantic OctoMap, guiding the planner to reconstruct the nodes in more detail. Every time a new node is detected, a new RoI is placed. The blue RoIs are centred around the main stem and run between the detected nodes. At the start, the stem RoI is initialized as an elongated vertical box centred in the workspace. When nodes are detected, the stem RoI can be estimated more accurately.

Information Gain: Using the Semantic OctoMap and the RoIs, the *expected information gain* can be calculated for a given viewpoint, ξ, using ray tracing:

$$G_{\text{sem}}(\xi) = \sum_{x \in (\mathcal{X}_\xi \cap \mathcal{B})} I_{\text{sem}}(x), \tag{1}$$

where $\mathcal{X}_\xi$ is the set of voxel within the camera field-of-view of viewpoint ξ and $\mathcal{B}$ is the set of all voxels in the attention RoIs. The expected information gain for a voxel, $I_{\text{sem}}(x)$, is determined using entropy:

$$I_{\text{sem}}(x) = -p_s(x)\log_2(p_s(x)) - (1-p_s(x))\log_2(1-p_s(x)), \tag{2}$$

Note that $I_{\text{sem}}(x) = 0$ for all irrelevant parts of the plant because $p_s(x) = 1$ for voxels that are not associated with outputs of the image-based object detector, in order to focus view planning on the relevant semantic classes.

In order to minimize motion of the robot, informative viewpoints close to the current viewpoint are promoted by calculating the view utility U:

$$U_{\text{sem}} = G_{\text{sem}}(\xi) \times e^{-0.5d}, \tag{3}$$

where d is the Euclidean distance cost.

The next-best viewpoint is selected as the viewpoint with the highest utility from a sampled set of viewpoint candidates $\mathcal{V}$:

$$\xi_{\text{best}} = \arg\max_{\xi \in \mathcal{V}} U_{\text{sem}}(\xi). \tag{4}$$

2.2 Experimental setup

Data acquisition: To compare the different planners in identical circumstances and to allow a repeated sensitivity analysis, we set up an offline experimental setup. For this, a large number of camera images from a diverse set of camera views were collected of eight plants in a production greenhouse. A total of 600 camera images per plant were collected, at 600 different positions (20 columns × 30 rows on a view plane with a distance of 40-60 cm from the plant's centre). At each position, one image was collected with the orientation perpendicular to the view plane. A total of $600 \times 8 = 4{,}800$ camera images with a resolution of 960×540 pixels were collected for the offline experiments.

Per plant, for every viewpoint, the set of 30 viewpoint candidates, $\mathcal{V}$, was selected pseudo randomly from the 600 camera images. The view plane was equally divided into a 3-by-3 grid and candidates were sampled from each grid to ensure that candidates were sampled from all parts of the view plane, similar to Burusa et al. (2024b). Note that this results in a set of viewpoint candidates at different distances d from the current viewpoint (see Eq. 3). For the random planner, the next viewpoint was selected at random from the set of camera images.

Ground-truth reconstruction: A ground-truth 3D reconstruction was made of every plant using a structure-from-motion and multi-view stereo reconstruction (Schonberger and Frahm, 2016) based on all 600 camera images. Using a voxel filter, the resolution of the ground-truth point clouds was reduced to 3mm, the same resolution as used for the OctoMap. Because the 600 viewpoints were all from one side of the plant, the ground-truth reconstruction is not fully complete and misses parts of the back side. However, as the NBV planners selected the viewpoints from the same set of 600 images, the evaluation is still valid.

Evaluation: The OctoMap reconstruction was converted into a point cloud and then compared to the ground-truth plant reconstruction. As this paper focuses on the detection and reconstruction of fruit and leaf nodes, the evaluation also focused on the reconstruction of these parts of the plant. To that end, the position of all nodes were manually annotated and all points within a cube with sides of 40 mm around the nodes were used for comparison.

For each node, the F1-score was calculated by comparing the reconstructed point cloud, R, to the ground-truth point cloud, G. For the F1-score calculation, a point in R was considered a true positive, if there was a corresponding point in R within 6mm distance (twice the point-cloud resolution). All points in R without correspondence were false positives, and all points in G without correspondence were false negatives. If a node was reconstructed with a F1-score greater of equal to 62.5%, the node was considered *correctly detected*. As a final metric indicating the quality of the reconstruction of all nodes on a plant, we then used the *percentage of correctly detected objects* (PCO):

$$\text{PCO} = \frac{\text{Correctly detected objects}}{\text{Total number of objects}} \times 100. \tag{5}$$

The threshold of 62.5% corresponded roughly to 50% reconstruction of a complete node. The threshold is higher, because the ground-truth reconstruction missed parts of the back side, as discussed earlier.

Experiments: The performance of the planners was evaluated on the detection of fruit nodes and the detection of leaf nodes separately. The tomato plants had fewer tomato nodes than leaf nodes, on average 2.5 and 5.5 per plant, respectively. The experiments were performed on the eight plants with twelve repetitions per plant to even out the random selection effects, leading to total of 96 trials. The PCO was calculated for every viewpoint and each trial was terminated after ten viewpoints.

The Semantic NBV planner presented in Section 2.1 was compared to the Volumetric NBV planner (Burusa et al., 2024b), which calculates the information gain based only on the occupancy probability, p_o, and a random planner, selecting random viewpoints from the set of candidates. In the sensitivity analysis, three different conditions were applied: (1) the addition of extra occluding objects in the scene, (2) the addition of uninformative viewpoints in the set of viewpoint candidates, and (3) a reduction of the field of view of the camera. Details on the experiments and the results are presented in the next Section.

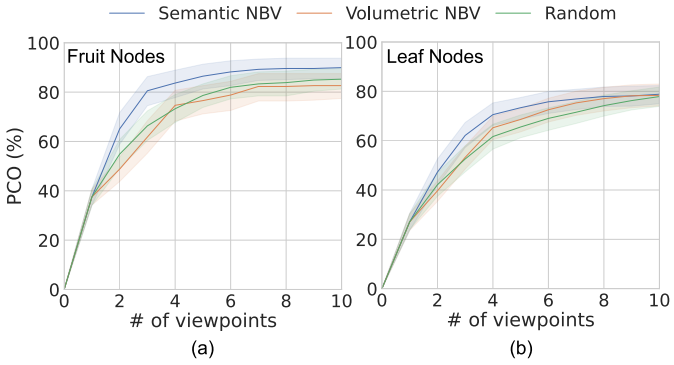

Fig. 3. Performance of the planners on the perception of (a) fruit nodes and (b) leaf nodes when a tomato plant was placed in front of the robot. The error bands show the 95% confidence interval of the mean over all 96 experiments.

3. RESULTS

3.1 Semantic NBV

Figure 3 (a) and (b) show the performance of the planners on the detection of fruit and leaf nodes, respectively. The Semantic NBV planner outperforms the other planners, achieving a higher PCO with viewer viewpoints. After four viewpoints, the planner achieves 10 and 9 percent point higher performance compared to the random planner for fruit and leaf nodes respectively. The performance on the fruit nodes exceeds that of the leaf nodes.

3.2 Impact of adding more occlusion

The impact of occlusion on the performance of the planners was tested by adding black boxes of 0.05m × 0.01m × 0.05m at a distance of 0.05m in front of the plant nodes, with a deviation of 0.15m to the left or right. These boxes were rendered in the camera images and depth map, as illustrated in Figure 4.

The results in Figure 5 show that the reconstruction with the added occlusion is slightly worse, but without a significant difference in the PCO. For the leaf nodes, the impact is a bit more severe, resulting in a small yet non-significant lower PCO for all planners. The Semantic NBV planner still outperforms the other planners.

Further inspection showed that the added occlusions only blocked the view on the nodes for a small set of all collected viewpoints, which explains the small impact on the PCO for all planners.

Fig. 4. To test the impact of occlusion on planner performance, more occlusion was simulated by adding black boxes in front of the plant nodes. The figure shows the rendered boxes in the camera image.

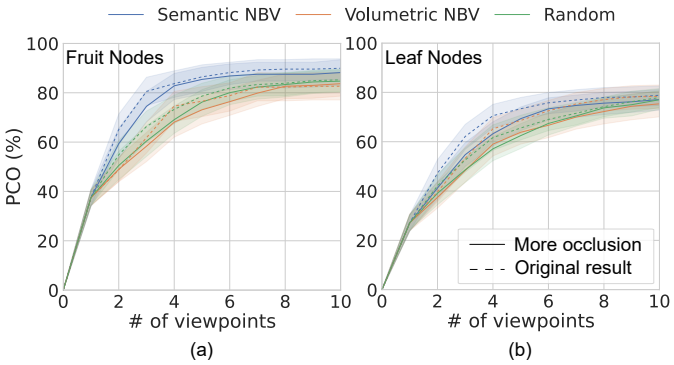

Fig. 5. Impact of adding more occlusion on the planners' performance on the perception of (a) fruit nodes and (b) leaf nodes. The error bands show the 95% confidence interval of the mean over all 96 experiments.

3.3 Impact of adding uninformative viewpoint candidates

The NBV planners are designed to choose the most informative viewpoints. To test if this is also happening in practice, 30 uninformative viewpoints were added to the set of candidate viewpoints, $\mathcal{V}$, at each step. The uninformative candidates were facing away from the plant and did not have any plant node in view.

The results in Figure 6 clearly show that the added uninformative views do not influence the two NBV planners, but have a significant negative effect on the random planner. This indicates that the NBV planners effectively assess the information gain of candidate viewpoints.

3.4 Impact of reducing the camera's field-of-view

The original field-of-view (FOV) of the camera was quite big, capturing a large part of the plant. To study the effect of a smaller FOV on the performance of the planners, we reduced it by a factor of four, as shown in Figure 7. The image size was reduced from 960 × 540 pixels to 480 × 270 pixels.

Figure 8 shows that the smaller FOV significantly impacts all the planners. The NBV planners, however, still outperform the random planner. For the fruit nodes, both NBV planners were less impacted by the reduced FOV than the

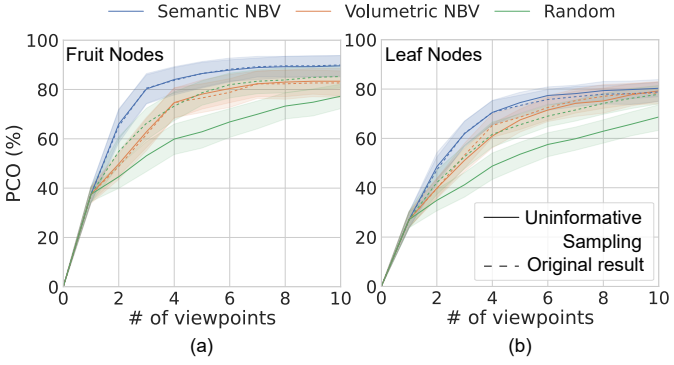

Fig. 6. Impact of adding uninformative viewpoint candidates on the planners' performance on the perception of (a) fruit nodes and (b) leaf nodes. The error bands show the 95% confidence interval of the mean over all 96 experiments.

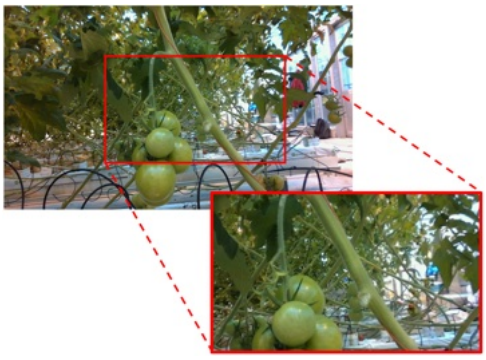

Fig. 7. To test the impact of the camera's field-of-view on planner performance, the field-of-view was reduced by a factor of four.

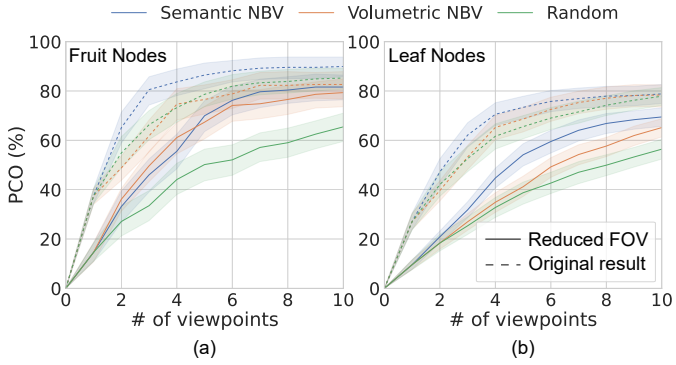

Fig. 8. Impact of reducing the camera's field-of-view on the planners' performance on the perception of (a) fruit nodes and (b) leaf nodes. The error bands show the 95% confidence interval of the mean over all 96 experiments.

random planner, while for the leaf nodes the impact was less for the Semantic NBV planner.

4. DISCUSSION

This work demonstrated the advantage of using semantic information in addition to 3D volumetric information to estimate the information gain of a next viewpoint for targeted perception of plant nodes. In this study, the Semantic NBV planner was evaluated on the 3D reconstruction quality of fruit and leaf nodes of real tomato plants in a production greenhouse and compared to a volumetric NBV planner and a random planner. The results show that the Semantic NBV planner is most effective in locating and reconstructing the nodes, which is in agreement with the results of Burusa et al. (2024a).

The sensitivity of the planners was evaluated under varying conditions, specifically, higher levels of occlusion, uninformative viewpoint candidates, and a reduced field-of-view. These conditions occur in real-world applications, depending on the crop and the robotic setup, and can strongly influence the performance of the viewpoint planners. In general, the results demonstrated that the Semantic NBV planner outperformed the other planners in all conditions and that it generally dealt better with the perturbations, thereby further validating the effectiveness of the approach for real-world greenhouse environments.

Both the addition of more occluding objects and the reduction of the field of view resulted in a lower PCO. This was to be expected, as in both cases there are fewer viewpoints with a good view on the plant nodes. In our experiments, we used 30 randomly selected candidate viewpoints for the NBV planners. The drop in performance due to the perturbations could be prevented by using a higher number of candidate viewpoints. However, that would come with addition computational costs, as the complexity of the NBV planners is linear in the number of candidate viewpoints.

The experiment with the added level of occlusion showed only a minor effect, caused by the occluding objects being too small. With large objects, resulting in more severe occlusions, it is expected that the NBV planners, compared to the random planner, will be less impacted, due to the use of information-gain estimation.

A key limitation of the experiments was that all available viewpoints for the planners were collected on a planar surface, with the camera always pointing perpendicular to that plane. This setup severely constrained the performance of the NBV planners, as oblique views can provide more favourable perspectives on the nodes. If the robot can sample viewpoints with more variation in the pose, the NBV planners are expected to outperform the random planner by a larger margin.

The presented method selects the next viewpoint without considering future viewpoints. This can result in suboptimal behavior. Future improvements, therefore, should include future viewpoints, for instance by utilizing a receding-horizon planner Lodel et al. (2022). To deal with the additional computation time, learning-based NBV methods should be considered **?**.

The plants used in this study were growing in normal greenhouse conditions, so in rows with overlapping plants and with a complex background consisting of consecutive rows. However, the plant reconstruction was done per plant independently. The set of 600 viewpoints were collected from a plane centred in front of the plant and the working space of the OctoMap was centered around the plant. Future work needs to extend the work to deal with the reconstruction of multiple tomato plants. This requires also a method to link the fruit and leaf nodes a specific plant ID, for instance by detecting the main stem.

The accurate and efficient perception of plant nodes, as demonstrated in this work, directly improves the performance and productivity of robotic harvesting and deleafing in greenhouses. The accurate perception of nodes can enable robots to effectively localize and cut the nodes, increasing their success rate. Also, by efficiently overcoming occlusion using fewer viewpoints can improve the speed of harvesting and deleafing operation by the robots.

5. CONCLUSIONS

This work presented an evaluation of a semantics-aware next-best-view planner for robotic perception of tomato plant parts. The planner used semantic information in addition to 3D volumetric information to focus on task-relevant plant-part reconstruction. The results indicate that the Semantic NBV planner could gather information efficiently, even in cluttered and occluded environments.

In the real-world experiments, the Semantic NBV planner perceived 10 percent points more fruit nodes and 9 percent points more leaf nodes after four viewpoints compared to the baseline Volumetric NBV and Random planners. Additional experiments highlighted the method's robustness to changes in conditions related to increased occlusion, the presence of uninformative viewpoints, and reduced camera field-of-view.

This study provides strong evidence that Semantic NBV planning is a viable solution for improving robotic perception in complex agricultural environments. By integrating semantic and volumetric information into viewpoint planning, agricultural robots can deal better with occlusions, paving the way for enhanced automation in crop monitoring and harvesting.

REFERENCES

Ankerst, M., Breunig, M.M., Kriegel, H.P., and Sander, J. (1999). Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2), 49–60.

Barth, R., Hemming, J., and van Henten, E.J. (2016). Design of an eye-in-hand sensing and servo control framework for harvesting robotics in dense vegetation. *Biosystems Engineering*, 146, 71–84. doi: 10.1016/j.biosystemseng.2015.12.001.

Boogaard, F.P., Rongen, K.S., and Kootstra, G. (2020). Robust node detection and tracking in fruit-vegetable crops using deep learning and multi-view imaging. *Biosystems Engineering*, 192, 117 – 132. doi: 10.1016/j.biosystemseng.2020.01.023.

Burusa, A., Scholten, J., Wang, X., Rincon, D.R., van Henten, E.J., and Kootstra, G. (2024a). Semantics-aware next-best-view planning for efficient search and detection of task-relevant plant parts. *Biosystems Engineering*, 248, 1–14. doi:10.1016/j.biosystemseng.2024.09.018.

Burusa, A., van Henten, E.J., and Kootstra, G. (2024b). Attention-driven next-best-view planning for efficient reconstruction of plants and targeted plant parts. *Biosystems Engineering*, 246, 248–262. doi: 10.1016/j.biosystemseng.2024.08.002.

Ci, J., van Henten, E.J., Wang, X., Burusa, A.K., and Kootstra, G. (2024). SSL-NBV: A self-supervised learning-based nbv algorithm for efficient 3D plant reconstruction by a robot. *ArXiv*. doi: 10.48550/arXiv.2410.14790.

Halstead, M., Ahmadi, A., Smitt, C., Schmittmann, O., and McCool, C. (2021). Crop agnostic monitoring driven by deep learning. *Frontiers in Plant Science*, 12. doi: 10.3389/fpls.2021.786702.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988. doi: 10.1109/ICCV.2017.322.

Hemming, J., Ruizendaal, J., Hofstee, J.W., and van Henten, E.J. (2014). Fruit detectability analysis for different camera positions in sweet-pepper. *Sensors*, 14, 6032–6044. doi:10.3390/s140406032.

Hornung, A., Wurm, K.M., Bennewitz, M., Stachniss, C., and Burgard, W. (2013). Octomap: An efficient probabilistic 3d mapping framework based on octrees. *Autonomous robots*, 34(3), 189–206. doi:10.1007/s10514-012-9321-0.

Kootstra, G. (2023). Advances in visual perception for agricultural robotics. In E.J. van Henten and Y. Edan (eds.), *Advances in Agri-Food Robotics*. Burleigh Dodds.

Kootstra, G., Bender, A., Perez, T., and van Henten, E.J. (2020). *Robotics in Agriculture*, 1–19. Springer Berlin Heidelberg. doi:10.1007/978-3-642-41610-1_43-1.

Lehnert, C., Tsai, D., Eriksson, A., and McCool, C. (2019). 3d move to see: Multi-perspective visual servoing towards the next best view within unstructured and occluded environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3890–3897. doi:10.1109/IROS40897.2019.8967918.

Lodel, M., Brito, B., Serra-Gómez, A., Ferranti, L., Babuška, R., and Alonso-Mora, J. (2022). Where to look next: Learning viewpoint recommendations for informative trajectory planning. In *2022 International Conference on Robotics and Automation (ICRA)*, 4466–4472. IEEE.

Rapado-Rincón, D., Nap, H., Smolenova, K., van Henten, E.J., and Kootstra, G. (2024). MOT-DETR: 3D single shot detection and tracking with transformers to build 3D representations for agro-food robots. *Computers and Electronics in Agriculture*, 225, 109275. doi: 10.1016/j.compag.2024.109275.

Rapado-Rincón, D., van Henten, E.J., and Kootstra, G. (2023). MinkSORT: A 3D deep feature extractor using sparse convolutions to improve 3d multi-object tracking in greenhouse tomato plants. *Biosystems Engineering*, 236, 193–200. doi:10.1016/j.biosystemseng.2023.11.003.

Sa, I., Lehnert, C., English, A., McCool, C., Dayoub, F., Upcroft, B., and Perez, T. (2017). Peduncle detection of sweet pepper for autonomous crop harvesting—combined color and 3-d information. *IEEE Robotics and Automation Letters*, 2(2), 765–772.

Schonberger, J.L. and Frahm, J.M. (2016). Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113. doi:10.1109/CVPR.2016.445.

Smitt, C., Halstead, M., Zaenker, T., Bennewitz, M., and McCool, C. (2021). Pathobot: A robot for glasshouse crop phenotyping and intervention. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2324–2330. doi: 10.1109/ICRA48506.2021.9562047.

Zaenker, T., Smitt, C., McCool, C., and Bennewitz, M. (2021). Viewpoint planning for fruit size and position estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3271–3277. IEEE.

Zapotezny-Anderson, P. and Lehnert, C. (2019). Towards active robotic vision in agriculture: A deep learning approach to visual servoing in occluded and unstructured protected cropping environments. *Proceedings of AgriControl 2019*, 52(30), 120–125. doi: 10.1016/j.ifacol.2019.12.508.

Zeng, X., Zaenker, T., and Bennewitz, M. (2022). Deep reinforcement learning for next-best-view planning in agricultural applications. In *2022 International Conference on Robotics and Automation (ICRA)*, 2323–2329. IEEE.