Active camera positioning utilizing guarded motion control to obtain a frontal view of a tomato truss enabling ripeness detection

den Hartog, C.M.

Award date:
2021

ACTIVE CAMERA POSITIONING UTILIZING GUARDED MOTION CONTROL TO OBTAIN A FRONTAL VIEW OF A TOMATO TRUSS ENABLING RIPENESS DETECTION

MASTER THESIS

C.M. (Stan) den Hartog
0953184

Mechanical Engineering
Control Systems Technology
Systems and Control
Supervisor: Dr. ir. M.J.G. van de Molengraft
Co-supervisor: Prof. dr. ir. H.P.J. Bruyninckx
Coach: PDEng. ir. J.P.F. Senden
CST number: CST2021.014

Eindhoven, April 16th 2021

Declaration concerning the TU/e Code of Scientific Conduct for the Master's thesis

I have read the TU/e Code of Scientific Conduct[i].

I hereby declare that my Master's thesis has been carried out in accordance with the rules of the TU/e Code of Scientific Conduct

Date 3/23/21

Name Constantijn Marc den Hartog

ID-number 0953184

Signature CM den Hartog

Insert this document in your Master Thesis report (2nd page) and submit it on Sharepoint

[i] See: http://www.tue.nl/en/university/about-the-university/integrity/scientific-integrity/
The Netherlands Code of Conduct for Academic Practice of the VSNU can be found here also.
More information about scientific integrity is published on the websites of TU/e and VSNU

Active camera positioning utilizing guarded motion control to obtain a frontal view of a tomato truss enabling ripeness detection

C.M. den Hartog
Control Systems Technology Group, *Department of Mechanical Engineering*, *Technical University of Eindhoven*

***Abstract*—Automation of deleafing and harvesting fruits still can not successfully manage complexity, caused by occlusion by leaves or fruits, after two decades. For tomato plants, reliable detection is required for: 1) petioles of (clusters of) leaves for deleafing, 2) ripeness of individual tomatoes for a truss, 3) cutting the peduncle of a tomato truss for harvesting. This paper introduces a novel concept of active positioning of a RGB-camera via guarded motion control around a tomato truss using a-priori geometrical relations. The guards are implemented on truss position and size to keep it within field-of-view while manually rotating around the truss in an artificial test-setup. The bounding box of the truss shows a maximum width over length when the camera obtains a frontal view of the truss which enables ripeness detection (part of use case 2). Successively, the truss length is estimated solely from the camera position motion data (without using a direct distance measurement to the truss) using an Extended Kalman filter and minimizing the squared mean length estimation errors. Analysis reveals that camera distortion and calibration, and vertical camera off-sets, trusses hanging non-vertically and tomato diameter play a role in creating systematic errors. It will be shown that this method measures the truss length to be structurally higher than its actual truss length by 9.8 ± 2.1 [%]. At a typical camera-to-truss distance of 25 [cm], this translates into a distance inaccuracy of 2.5 ± 0.5 [cm] which needs to be taken into account in the next step in the process of truss harvesting.**

***Index Terms*—Robotic harvesting, active sensor positioning, active perception, ripeness detection**

I. Introduction

In 1996, the first two tomato-picking robot prototypes [1][2] were created to harvest the fruits, and since then several start-ups aimed to robotize the maintenance of e.g., tomato plants [3]. However the variance in phenotype between the plants (e.g., growth speed, orientation of trusses or leaf-clusters, and resulting plant placement) creates complex sensor conditions for a robot to reliably deleaf and harvest fruits e.g., tomato trusses. Reliable deleafing and harvesting means that all necessary leafs are cut off and ripe trusses are harvested, while minimizing losses of tomatoes and causing no damage to the plants. Current developed systems follow a pre-determined path along the rails between the different rows of plants in order to capture images using a RGB(-D)-camera for deleafing and harvesting. This approach functions for specific cases where the detection conditions are identical in phenotype and handled identically, and thus predetermined sensor positions can capture the required information. Vision detection for tomato plants encounters many occlusions of petioles, peduncles or even tomatoes due to their leaves or trusses (Figure 1). Hence, reliable detection, i.e., appropriate positioning of the sensor and accompanying object identification software, is required for the following use cases:

Fig. 1: Occlusion for the three uses-cases as highlighted by red circles: deleafing with petioles hidden by leaves (left), ripeness detection of all tomatoes in a truss including the ones hidden behind the visible tomatoes (center) and cutting off the hidden peduncle for harvesting of tomato truss (right). Photographs taken in Vereijken's greenhouse.

1) petioles of (a cluster of) leafs for deleafing,
2) ripeness of individual tomatoes for each individual truss,
3) picking at the peduncle of tomato truss.

This paper mainly focuses on the active positioning of the visual sensor (i.e., camera) to enable ripeness detection (use case 2). Between 1966 and 1972 the first robot had been created that could reason about its actions, planning how to use and position its sensing resources [4], and current implementations of greenhouse robotics do not adaptively plan their camera placement for ripeness detection (section III). An approach will be taken where the a-priori environment knowledge is combined with the information gained from the images in order to determine the following action, i.e., active positioning of the sensor. Active positioning entails that the current state of the sensor determines what the next action will be, concerning the deployment of its sensor resources, based on expectations generated by its state [5]. For harvesting of a tomato truss, the ripeness has to be evaluated and the peduncle has to be identified, which is possible when seeing the tomato truss from a frontal view (Figure 2 (center)). The aim in this paper is thus to gain a frontal view of the tomato truss with a sufficiently large visible area on all individual

tomatoes enabling ripeness detection, starting from an initial image of a tomato cluster. A strategy for camera placement is implemented, validated using a RGB-camera moving around the tomato truss to obtain a frontal view as well as the camera-to-truss distance estimation and its accuracy.

II. Use-cases

The economic need for automation of greenhouse crop-growing is mainly caused by a lack of skilled workers, increased labor costs and the increasing world population [6][7][8]. The most labor-intensive tasks at a greenhouse are deleafing the plant (15 [%] of total labor) and harvesting the fruits (31 [%] of total labor), with combined raised-platform work, consisting of wiring and rehanging the plant, pruning branches and the flowers (35 [%] of the total labor) [9][10]. Deleafing and ripeness estimation and harvesting of fruits are thus the main areas of interest for automation in the greenhouse. These tasks have a similar nature of object identification and cutting its attachment to the stem, and large contributions to the total amount of work.

Deleafing. The tomato-plants in greenhouses grow rapidly like weeds and are hung in high-wire system, holding the plant up vertically via a wire. The lowest leafs are less photosynthetically active due to the obstruction of incident light of the higher leaves. The lowest leaves are removed which also increases the airflow throughout the crop, ensures that the energy in the plant goes to growing the tomatoes, and facilitates the harvesting of the trusses in an unobstructed view [11]. The leaf is cut at its stem-attachment, the petiole, within 1 [cm] of the stem, as to prevent rotting of the stem (see Figure 2). The to-be-cut leafs can be damaged, whereas the tomato trusses, wires and stems should not be harmed in any manner. Typically two or three leafs are grown around each tomato truss, and as reference manual labor removal costs about 0.02 [€/leaf] [9]. Note that tomato plants do not have identical growth speed, hence after weekly re-hanging at the high-wire still some stems can be positioned closer together. Here, clusters of leafs may occur, creating obstructed views of the petioles and thus (machine) vision is not able to provide a reliable detection of their petioles ($\sim$10 [%] of the cases) [9]. In contrast, human workers mostly use tactile information from the hands by moving along the stem and feel where to cut off leafs [9].

Ripeness detection. Customers desire complete tomato trusses for purchase, thus the trusses have to be harvested as a whole which means that the ripeness of an entire truss has to be determined [9]. Using machine vision, different methods have been demonstrated to be effective to estimate truss ripeness, using machine learning [12], color space segmentation [13] or a combination of the two [14], and one ripeness estimator in the field makes use of harvesting cards, comparing the ripeness of the image to certified examples [15]. The minimum requirement on the spot size area A to detect ripeness has been indicated to range from 200 to 2500 [pixels] [16][18] with a minimum roundness indicator P ($P = \frac{4\pi A}{C^2}$, where C is the circumference [pixels]) in the range from 0.5 to 0.7 [-] with a success rate of 94 [%]. However, this was determined by the tomato object detection itself, being impacted by inhomogeneous light distribution [16][18], and occlusions or overlapping of the randomly oriented trusses [17][18]. Hence, ripeness detection for larger than the minimum spot size ranges and minimum roundness P is expected to be reliable when measured at a clearly defined unobstructed frontal view.

As the tomato-plants hang vertically, the tomato trusses, due to their weight, hang down the side of the stem, away from the direction that the stem hangs in (e.g., Figure 2), further explained in subsection IV-B. The tomato trusses that are ripe enough to be harvested hang below and at eye-height, below the leafs. Thus the ripe trusses are only possibly occluded by other trusses, stems, or greenhouse components (e.g., pillars, tubes) or have their peduncle occluded simply due to their orientation towards the sensor. In order for a truss to be harvested, its ripeness needs to be assessed, based on the ripeness of the individual tomatoes belonging to the truss as this is the common practice in the ripeness estimator machines as described above. Using visual sensors the ripeness of individual tomatoes can be determined through assessing its color and spectral chlorophyll fluorescence, which are both dependent on the size of the tomato on the image [19]. The gradient of the ripeness of the individual tomatoes of a truss is utilized to assess the ripeness of the complete truss. The tomatoes of a single tomato truss have a zigzag pattern around the peduncle, for the higher tomatoes of a tomato truss (Figure 2), whereas the lower tomatoes are clustered together. Assessing the ripeness of the truss is thus dependent on identifying which tomatoes belong to a single truss, and identifying the ripeness of each individual tomato. Visual obstructions e.g., Figure 1 or multiple trusses close together, thus create the difficulties for ripeness assessment.

Harvesting. Ripe trusses are required to be harvested after which they will be distributed for consumption. Different mechanical solutions are designed, directed to single-tomato

Fig. 2: Typical desired views without occlusions, as highlighted by green circles, for the three uses-cases: deleafing with petioles attached to the left and right on the stem (left), ripeness detection of all tomatoes in a truss including a visualization of the zigzag pattern (center) and cutting of the peduncle for harvesting of tomato truss (right). Photographs taken in Vereijken's greenhouse.

harvesting and truss harvesting. Single-tomato harvesting is done through suction [20][21], gripping [20] or a combination of the two [20][22]. Tomato truss harvesting is currently done by holding the peduncle and cutting it from the stem [20][23][24]. For the initiation of harvesting or grabbing the peduncle no quantitative requirements are provided for neither the distance to the peduncle from the sensor or actuator, nor its placement accuracy [20][23][24][25][26][27]. Manual labor cost for harvesting a truss is about 0.05 [€/truss] [9]. The trusses are pruned from the stem at less than 1 [cm] distance in order to prevent rotting. Due to the possibility for tomatoes to detach from the truss, the truss needs to be carefully handled in order to prevent any damage.

In summary, three use-cases have been identified being deleafing, tomato ripeness detection and truss harvesting. The main difficulties of these use-cases are caused by inhomogeneous light distribution (detection), by (partial) occlusions of the peduncle, petiole or tomatoes (environment), and by species having short peduncles or petioles (mechanical) [24]. This study will focus on solving occlusions (environment) by active positioning of the visual sensor. In particular, the ripeness detection use-case will be investigated, since active sensor positioning might bring the largest benefit here in eliminating occluded views by moving to a position where the ripeness of each individual tomato of the truss can be measured.

III. Related work

Machines and research that do the identification of ripe fruits or fruit-clusters, leafs or leaf clusters and their attachments to the stem (e.g., peduncle and pedicel for tomato-plants) based on visual sensors i.e., cameras are taken into account. This set of related work is chosen due to the required multi-functionality of ripeness and peduncle identification, which is most easily possible to do simultaneously through visual imaging.

Tens of machines are designed for harvesting of soft fruits, e.g., tomatoes [23][25][28], peppers [29] apples [30][31], raspberries [32] and strawberries [33][34][35][36], and for deleafing of different plants, e.g., tomatoes [37], cucumber [38] and grapes [39]. These machines all assess the ripeness of the fruit using a similar structure, with cameras rigidly fixed to the robot that have pre-determined pathing of the complete robot. The imagery along this pathing is used to assess the case-by-case-varying environment, identify fruits and giving sensory feedback before and during the harvesting and deleafing. Inherent to the chosen control strategy with the fixed positioning on the robot of the visual sensor e.g., camera, i.e., passive positioning of the sensor through feedforward control of the complete machine, the robot acts only upon the collected information available along its pre-determined path. The machines vary when it comes to the harvesting mechanism, as some machines have moving cameras (e.g., [32]) in order to center the fruit and to assess the distance to the fruit after it has been decided that the fruit is ripe and is required to be harvested. However, this is for fruits that can be harvested from any side and where only a single fruit is harvested at a time (e.g., raspberries). It is not indicated in the robot how the distance to the fruit is assessed, e.g., RGB-D cameras or utilizing the normal size of the fruit. The moving cameras are not implemented for fruit localization and ripeness detection.

Due to occlusions of petioles or peduncles, not enough data (e.g., ripeness, amount of fruits, peduncle identification) is available in more complex, occluded cases in order to harvest all trusses in the case of body-frame-fixed cameras [40]. Actively positioning the sensor to gain information to improve the reliability of the task of cutting leafs, assessing ripeness or harvesting a truss is a possible solution.

Active sensor placement includes decision-making to decide the next sensor placement, with early efforts being based around finite state machines [5]. Active positioning of sensors is attempted through cost-function optimization using grid-based positions [41]. The workspace of the robot is discretized, and the cost-function considering a specific uncertainty or information gain and distance criterion in its environment is iteratively optimized over the grid of the workspace [42]. The nature of this approach is prone to failure of proper placement due to the discretization, having finite options for placement of sensors, and not taking into account information that could be gained during the movement. Recursive optimization, moving the sensor only to neighboring points in the grid on the way to the next viewpoint, is an alternative [43]. The discretization remains the largest issue, as it is computationally expensive to obtain a small grid, and additionally it is assumed that from at least a single point on the grid the fruit can be seen that satisfies the requirements set by the task, which follows from the decision on how the grid is placed and with which discretization size.

In summary, in order to reliably perform tasks enough information has to be gained through the sensors, which are currently attached rigidly to the entire robot. Active sensor placement has been attempted through cost-function optimization over a grid. In contrast, in this thesis active positioning through open-loop movements using object detection from the image sensor information is proposed, which is not limited to discretized grid-based positions and hence not blocking potential valid options for placement of the sensor.

IV. Proposed approach

In this section, the concept and benefits will be explained high-level in subsection IV-A. In subsection IV-B the geometrical knowledge from the greenhouse and the resulting initial camera placement are explained in detail, placing the approach in the context of ripeness detection. The resulting strategy for the active placement of the camera based on the knowledge from the previous subsection is given in subsection IV-C. The values of the exact implementation are dependent on the used set-up within a specific environment and therefore will be given in section V.

A. General description

Active sensor placement provides a possible advantage in complex environments to retrieve additional information and have enough to perform the task, e.g., in order to cut petioles, assess ripeness of a truss and harvesting the truss, and move from the examples in Figure 1 to the solved occlusions in Figure 2. More generally, the goal of active sensor placement is obtaining more information to perform the eventual task of harvesting reliably, which requires relevant knowledge from a complex environment to be simplified when possible, utilizing a-priori, geometrical knowledge. Since these relations within and between plants do not depend on distances between objects in Euclidean space, these measurements are not necessary and will solely complicate the method of the task. This geometrical knowledge can be used to resolve occlusions and re-position the sensor accordingly. As it cannot be guaranteed that a grid can be provided (properly placed and with adequate discretization size) that will always have a non-occluded view of the target (in this case the peduncle or front of the truss), open-loop movements with guards to stop and redirect motion. Guarded motion utilizes sensory conditions (guards) that determine when and how to re-position the camera [44], in this case based on open-loop movements. Due to the nature of open-loop movements, there is no grid and thus the complete range of motion of the robot can be accessed which offers more views on the target.

B. A-priori geometrical relations and resulting camera initial position

1) A-priori geometrical relations in tomato plants: The systematic manner in which the tomato plants are placed and maintained, ensures that the tomato trusses are also hanging in a systematic manner. There is a difference in position and orientation due to the limited variation in phenotype. Due to the re-hanging to the right (Figure 3 (left)) to compensate for growth of the plants in combination with the deleafing (section II), the tomato trusses hang between 1.3 and 1.8 [m] height, unobstructed by leafs. In combination with the weight of the tomatoes of the truss, the trusses hang to the right ('underneath' the stem due to gravity), and have a higher probabilities of hanging in the middle of the right-side for these particular tomato varieties and this greenhouse [9].

2) Resulting initial camera placement: The initial placement of the camera is in the right-bottom (Figure 3 (left-bottom)), above the rails in the walking path between the different rows of plants, such that the camera will not touch other plants while moving from one plant to the next. The likelihood is higher that the trusses are initially seen from a frontal view when observing from the right, compared to having an initialization in the left-bottom. The initialization position is thus based on the relation between the plants, and the relation between the stem and the trusses.

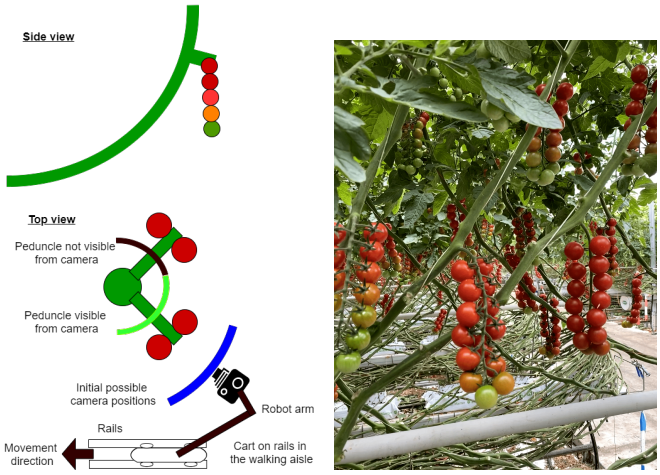

Fig. 3: Left: schematic overview of the manner that trusses are attached to the stem (with a side-view (top) and top-view (bottom)), and with the blue range indicating from which initial angle the camera normally views the trusses from the walking path. Right: a real-life example of the manner in which trusses (another type of tomatoes than the ones shown in Figures 1 and 2) hang in a greenhouse, displaying the various orientations in which the trusses can hang. Photograph taken in Vereijken's greenhouse.

C. Camera motion method towards truss ripeness detection position

Firstly, it will be explained in subsubsection IV-C1 that the front of the truss can be localized through maximization of the width of the truss normalized to its length. In subsubsection IV-C2 the movement to perform the normalized width maximization is elaborated upon. Lastly, in subsubsection IV-C3 a length estimation of the truss utilizing depth through motion methods is explained without direct Euclidean camera-to-truss distance measurements, which is to be used for the subsequent harvesting mechanism placement.

1) Motion to maximize normalized truss width: For the case in the greenhouse, the initial placement of the camera is directed to the next harvestable truss, thus having a random orientation and position compared to that truss within the range indicated in subsection IV-B. To be able to determine the ripeness of a truss, the size of the individual tomatoes in the image has to be sufficiently large ($A > 2500$ [pixels], $P > 0.7$ [-]), and the tomato truss is to be observed frontally with a visible peduncle (section II). For the purpose of camera placement, the tomato truss is to be identified as a whole, since the information concerning individual tomatoes is only required during the ripeness estimation which occurs once the camera is placed in front of the truss.

Based on the initial image, two optimizations take place consecutively in order to gain a frontal view of the tomato truss from any position of the camera while viewing the truss: the camera needs to move to the appropriate height and to the correct orientation concerning the truss. For the first optimiza-

Fig. 4: Three typical picture examples with the placement of the blue bounding boxes by an object detection algorithm, where the width and length of the bounding box are respectively the horizontal and vertical sizes measured in pixels: rotating from a view on the truss on the left-side (left), to a left-front view (center) and finally a frontal view (right). Photographs taken in Vereijken's greenhouse.

tion, the camera is positioned at the height corresponding to the middle of the truss, where the length against the width of the bounding box of the truss is maximized through a vertical rotation around the truss in the plane of the side view as given in Figure 3 (left-top) (maximizing the normalized length at each iteration i, as denoted in Equation 1a). The second optimization uses a rotation and movement in the horizontal plane (the top view in Figure 3 (left-bottom)) around the truss in order to obtain a frontal view of the truss, where the normalized width at iteration i is defined as the width divided by the height of the bounding box (Equation 1b). When rotating around the truss, the normalized width exhibits a maximum when the camera is in front of the truss. The validity of the active camera positioning method will be proven by demonstration of this maximum in the normalized width.

$$l_{n,i} = \frac{l_i}{w_i} \tag{1a}$$

$$w_{n,i} = \frac{w_i}{l_i} \tag{1b}$$

To identify the tomato truss as a whole enabling these optimizations, object detection in the form of bounding boxes (Figure 4) is a simple form of image classification and object localization. This form of object detection has a low computational load in comparison to image segmentation, where each pixel belong to the object is given as an output [45][46]. The bounding box gives a sufficient amount of information for localization of the front of the tomato truss (the visible width and length of the truss), and is thus the preferred measurement of the tomato truss because of its simplicity and low computational load.

As a starting point, the initial maximization of the length against the width is assumed to be completed, as these two optimizations are consecutive. This reduces the 6 Degree Of Freedom (DOF) movement to a 3 DOF movement, utilizing movements in the horizontal plane and yaw rotations to rotate around the truss, with locked pitch, roll and vertical motion. Furthermore, it is assumed that on the images the size and placement of the bounding box is equal to the ground truth (with measurement noise), since faulty measurements will have to be compensated for utilizing different methods and perhaps different sensors. Next, the tracking software attached to the camera can track the bounding box of the tomato until a guard (e.g., a maximum or minimum size of the truss in the image) is reached, creating discrete decision positions. In order to position the camera sensor, no depth information gained from specific depth sensors will be utilized, as this would unnecessarily complicate the implementation due to sensor-fusion. Furthermore, due to the nature of the depth-sensor, models based on the geometrical properties of individual tomatoes and the complete truss composition would be necessary. Depth estimation utilizing the bounding box and the positions of the camera would not require additional modelling and can give sufficient depth estimates.

2) Algorithm flow: The steps taken at every iteration of images is indicated in Figure 5 (left). Considering the image at a certain iteration, the knowledge gained from the positioning and the image of the camera is updated: the width and length of the truss are evaluated and the camera position is updated. If the bounding box is outside of the guards, the camera is repositioned (Figure 6). If no guards are crossed, the normalized width is evaluated and the camera state machine as

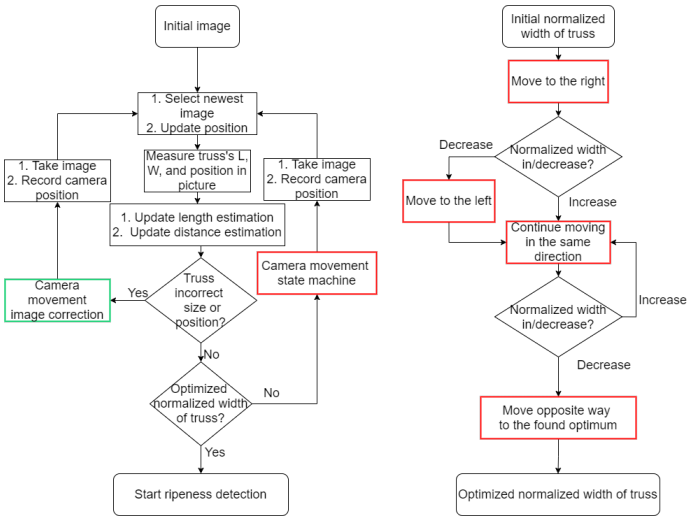

Fig. 5: Left: flow diagram of the iterative steps that are required to be taken at every image, deciding which movement is required to fulfill the task of active sensor placement. The movement directions that are given in green and red, are further explained, respectively, on the right and in Figure 6. Right: flow diagram of the movement direction in the case where no guards are crossed and specific movements towards the optimum of the normalized width of the truss can be performed, assessed at every iteration capturing new images (deciding the movement direction of the red movement box on the left).

described in Figure 5 (right) controls the next motion direction for the camera sensor.

The movement direction based on crossing the guards set on the images, consists of two parts: size (both width and length) and position (Figure 6). The guards for the size of the truss require the bounding box to be of a certain minimum and maximum size in pixels based on the specific implementation (section V), leading to forward or backwards motion respectively. The guards for the position of the truss require the bounding box to not be within a certain amount of pixels near the edges of the image, in order to ensure the full truss to be on the image. This leads to either rotation or backwards motion, for respectively single or opposing edges that are touched.

The movement direction of the camera (Figure 5 (right)), if the truss is appropriately visible (meaning no guards are crossed) in the images, is a rotation around the truss. The initial translation is to the right, due to the higher probability of the truss hanging in that direction (Figure 3). This rotation is created through adjustments as follow from Figure 6. Since the normalized width is to be optimized (compared to the length of the truss), the movement decisions are based on this criteria. The camera is required to move in the direction of the increase of the normalized width, until a decrease is measured, indicating that the maximum is identified. To prevent a false positive of the maximum due to measurement noise, a threshold is implemented that ensures that the decline in normalized width is statistically relevant.

3) Truss length estimation: During the motion, additional information regarding the truss can be determined through the measured position of the camera and the (change in) size of the truss and thus its bounding box (Figure 4), such as the distance to the truss [cm] and the length of the truss [cm] (Figure 7, Equation 2). This technique is called "depth from motion". This data is useful for subsequent sensor or actuator placement for the harvesting task. The length of the truss in the view remains constant and is only dependent on the distance

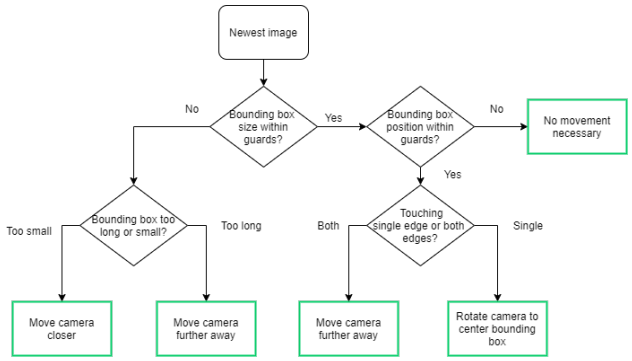

Fig. 6: Flow diagram of the movement direction in the case where a guard is crossed: the truss in the image is either not within the guards of the position or of the size, which is assessed at every iteration of the images (Green decision box in Figure 5 (left) flow diagram for the movement direction).

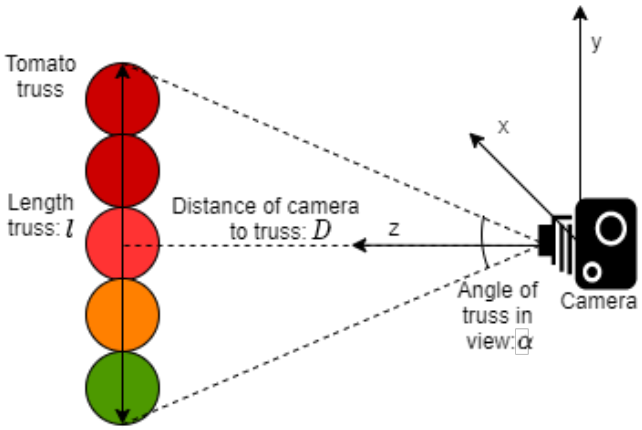

Fig. 7: Schematic drawing of a camera viewing a tomato truss of length l, at a distance D from the camera, with a viewing angle of α (left).

to the truss, due to the initial positioning at the height of the middle of the truss and the horizontal motions performed during the rotation. The length of the bounding box is directly correlated through trigonometry with the angle α [rad] based on the amount of pixels on the image, the distance D [m] to the truss and the length of the truss l [m] (Equation 2).

$$D_i = \frac{l_i}{2\tan\left(\frac{\alpha_i}{2}\right)} \tag{2}$$

Utilizing the schematic analysis based on the 2-D model (Figure 7) and the step sizes between images in the camera-frame m_z and m_x in the z and x directions respectively (Figure 7), the dynamics of the angle of the truss is determined (Equation 3), including the measured range of pixels y_i that is converted to α_i [rad] through the pixels per degree ppd constant of the camera.

$$x_{i+1} = \alpha_{i+1} = f(\alpha_i, l_i) \tag{3a}$$

$$\alpha_{i+1} = 2\tan^{-1}\left(\frac{l_i}{2\sqrt{(D_i - m_{z,i})^2 + m_{x,i}^2}}\right) \tag{3b}$$

$$y_i = ppd \cdot \alpha_i \tag{3c}$$

Squared Mean Error (SME) minimization: A method for estimation of the length of the truss l is through minimization utilizing the squared error between the measured angle α_i and the estimated angle $\hat{\alpha}_i$. The sum of the squared errors is minimized over the parameter l over the total amount of measurements n (Equation 4) [47]. This method is an a-posteriori method functioning best after all measurements have been gathered, creating more certainty due to the larger amounts of measurements. However, due to limited available

data and model inaccuracies, this method is prone to over-fitting for the parameter l.

$$\min_{l} \sum_{i=1}^{n} E_i^2 \tag{4a}$$

$$E_i = \alpha_i - f(\alpha_{i-1}, l) = \frac{y_i}{ppd} - f(\frac{y_{i-1}}{ppd}, l) \tag{4b}$$

$$E_i = \alpha_i - 2\tan^{-1}\left(\frac{\frac{l}{2}}{\sqrt{\left(D_{i-1} - m_{z,i-1}\right)^2 + m_{x,i-1}^2}}\right) \tag{4c}$$

Extended Kalman Filter: In the case that the distance is required at any iteration i, an observer can be built to estimate the states of Equation 3 in order to compensate for measurement errors, wrong initial estimation or camera positioning errors, which minimizes propagation errors due to modeling errors [48]. This prevents the over-fitting that occurs during the SME minimization. The length of the truss l is added to the states for parameter estimation due to the uncertainty on its value. The feedback gain K_i of the observer is obtained from the standard Extended Kalman Filter [49]. The Extended Kalman Filter is further explained in Appendix A.

Both methods of truss length estimation will be implemented and compared for performance. The algorithm for the movement in combination with the length and distance estimations (subsection IV-C), will be experimentally validated starting from the initial camera position as denoted by the environment (subsection IV-B).

V. Experiment Design

A setup is used to perform the experiments where the position displacements based on the control algorithm (subsection IV-C) are made manually, (as described in section IV). In subsection V-A, subsection V-B and subsection V-C, respectively, first the experimental setup is described, after which the determination of the guards is explained and subsequently the results of the motion around the truss and its length estimation is shown.

A. Experimental set-up

The artificial set-up consists of an artificial stem, with a tomato truss (Figure 8 (center)) hanging underneath the stem (Figure 8 (left)) and a camera which can be manually moved in the horizontal plane and rotated around the yaw-axis, while fixed in the other DOF (Figure 8 (right)).

The angle of the artificial stem resembles the tomato plant utilizing the black wooden frame, however does not resemble its weight (Figure 8 (left)). The plastic tomato truss has a similar orientation compared to the stem and is of similar composition as normal trusses albeit the tomatoes are lighter (Figure 8 (center)).

The camera is the Intel® RealSense™ D435 (Figure 8 (right)). The Field of View (FoV) of the camera is

Fig. 8: The artificial setup displayed in its totality (left), with a close-up of a tomato truss attached to the stem (center) and the visual sensor (right), set up according to Figure 3 (left).

$69.4° \times 42.5° \times 77°$ ($H \times V \times D$), with a resolution of 640×480 [pixels] (see discussion in subsection V-C with respect to calibration and FoV cropping due to VGA format). The supplier specifications of the FoV are a maximum accuracy of 3 [°], the image has a maximum distortion of 1.5 [%] and the camera is calibrated in the factory [50][51]. The built-in stereo-function, i.e., the depth-signal, is not used in order to minimize the saved data during the motion. The camera is attached to a mobile pole, around which it can be rotated. The camera position is measured with respect to the black wooden frame using a carpenter's square and a thread and needle attached to the center of the camera (Figure 8 (right)). Its yaw rotation (in the horizontal plane) is measured using this needle pointing to a protractor at the bottom of the camera pole.

B. Methodology to set guards

The motion is performed manually with two types of guards determining its open-loop movement direction. Firstly, the distance between the camera and the truss is guarded to be within a range of 15 and 22 [cm], respectively, when the truss fully fills the vertical FoV and about 70 [%] (based on the camera specifications). Secondly, guards at a distance of 2 [%] of the total width and length of the image from the edges of the images ensure that the truss stays inside the FoV in the horizontal and vertical directions by, respectively, requesting a yaw rotation and a horizontal backwards motion when the bounding box around the truss hits these guards. The smallest rectangular bounding box around the truss is manually selected for every image as shown in Figure 4. Note that the aspect ratio of the camera FoV is larger than that of the tomato truss, so that the maximum width of the truss will always fit into the FoV when the truss length fits.

C. Results of the maximization of the normalized truss width and truss-length estimation

1) Motion to maximize normalized truss width: The camera was initialized at a random position (within its starting region) pointing towards the truss (Figure 10 (left)). The motion profile in the x,y-plane can be seen in Figure 9 (left), where the guards guide the camera position forward from the initial position, after which the rotational motion starts. In the second

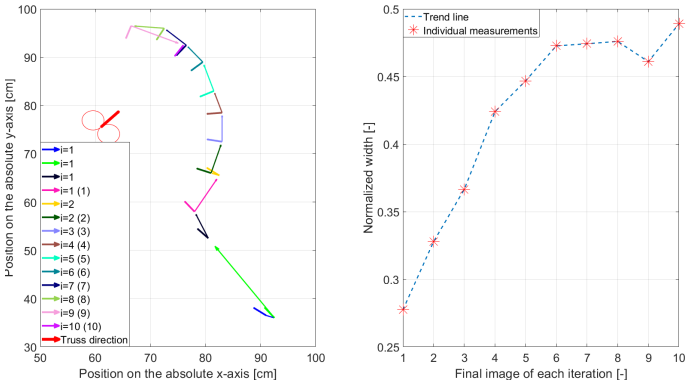

Fig. 9: The top view of the positions of the camera and its direction (left), and the measurement of the normalized width during this rotation at every image that satisfied the guards during the motion (right). A maximum normalized width in iterations 8 and 10 is visible as during the 10^{th} iteration the camera is rotated back to the maximum normalized width (right).

Fig. 10: Acquired images during the experiment that fulfilled the requirements set by the guards during the rotational movement around the truss, corresponding to the initial (left) with an occluded peduncle, fifth (center) and final tenth iteration (right).

iteration the camera is guided forward, closer to the truss. The measured normalized width increases throughout iteration 1 to 8, as visible in Figure 9 (right). The continued rotation around the truss in measurement of iteration 9 shows a decrease in normalized width, hence the maximum in normalized width can be found by rotating back. In iteration 10, the camera is rotated back to this position of the maximum normalized width and successfully shows a frontal view of the truss with full access to the tomato truss to start ripeness detection, see Figure 10 (right). The utilized threshold to prevent a false positive of the maximum due to measurement noise is set to 0.01 [-] based on a random measurement error of ± 4 [pixels] on the selection of the bounding box. The final camera placements ensure that the tomatoes are visible with an average area of 6600 ± 1100 [pixels], where the minimum requirement is 2500 [pixels]. The average roundness of the localized tomatoes is 0.92 ± 0.05 [-], where the minimum requirement is 0.7 [-]. Therefor, a camera position is obtained that enables ripeness detection.

2) Truss-length estimation: To use the Extended Kalman Filter (EKF) (Equations 7 and 9) to estimate the length of the truss it requires to be initialized with specific starting parameters (explained in section A), whereas the Squared Mean Error (SME) minimalization does not need to be initialized. For the EKF, the $\hat{x}_{0|0}$ starting parameters are chosen for the length l to be equal to the average truss length of 19.1 [cm], which is known with a standard deviation of ± 0.1 [cm]. For the angle α a starting value of 0.4 [rad] is used which corresponds to the normal distance of the camera to the next truss, being the distance between two stems which is typically 45 [cm] in greenhouses [9]. Note that taking other initial values changes only the starting point and initial convergence, but does not affect the asymptotic value of the length and angle estimation in later iterations in steady-state [52]. The process noise on the angle has possible unmodelled behavior which is included in the process noise (typical error of $< 5 \cdot 10^{-2}$ [rad] at each iteration) and there is 0.2 [cm] process noise on the exact length of the truss (2σ), so $Q = \text{diag}[2 \cdot 10^{-3}, 10^{-2}]$, as those errors are taken to be 2σ. These values have been measured minimally 20 times, and therefor the 95 [%] confidence interval is at 2σ. The random measurement error of the manual indication of the size of the truss bounding box is low, namely ± 4 [pixels] (or about 0.6 [%] of the FoV, which is taken as 2σ), which results in $R = 1.4 \cdot 10^{-5}$. The estimated angle α is known with high certainty ($\pm 7.6 \cdot 10^{-3}$ [rad]) as it is measured directly and the error in estimating it follows from an initial faulty estimation of the length of the truss. Therefor, the covariance of the angle is of < 0.3 [rad] average initial error whereas the covariance of the length is chosen to be high (10^3) as usual in parameter estimation, $P_{0|0} = \begin{bmatrix} 0.09 & 9.5 \\ 9.5 & 10^3 \end{bmatrix}$ [49].

The EKF approximates the angle α_i with minimal error due to the direct measurement feedback of y_i, and the length of the truss l_i is estimated to be at a steady state value of around 21 [cm] (Figure 11). The SME minimization of the length of the truss (Equation 4) over the different iterations, displays a similar behavior by converging to a steady state of l_i after two images of around 21.3 [cm].

Even though both length estimators converge to a similar value of 21.3 ± 0.3 [cm], this displays a systematic error as the actual length of the truss is 19.1 ± 0.1 [cm]. Two additional sets of experiments have been conducted with that truss and also show a systematic error where the length of the truss is estimated to be, respectively, 20.5 ± 0.4 [cm] and 21.3 ± 0.2 [cm]. Therefor, on average the estimated length of 21.0 ± 0.3 [cm] shows a systematic error compared to the the actual value of 19.1 ± 0.1 [cm].

Moreover, similar results are found for three datasets where the shape of the tomato truss has been altered: one truss with the bottom two tomatoes missing, one truss with the top tomato missing and one truss with the top two tomatoes

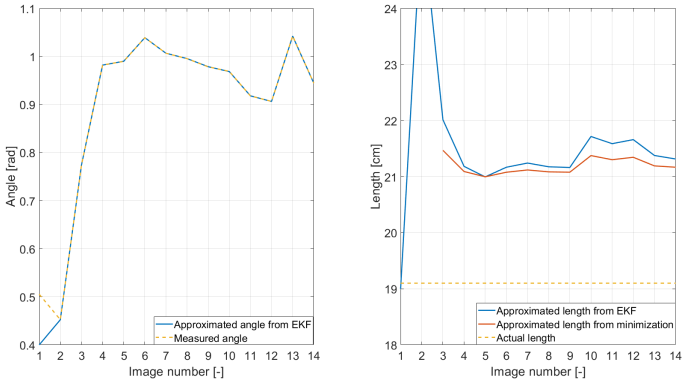

Fig. 11: The results of the measured angle of the bounding box of the truss on the image (left) and the length estimation of the tomato truss (right), acquired during each of the images where the bounding box of the truss is fully in view, thus not crossing the guards set on the edges of the image. Note that the Extended Kalman Filter provides a consistent length value in steady state (i.e., after 3rd image), and similarly the minimalisation is also shown for steady state.

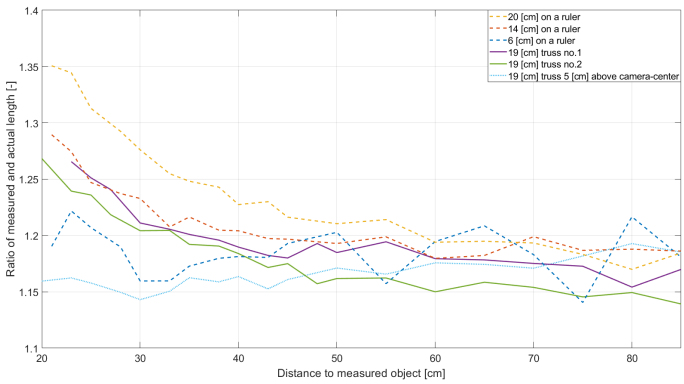

Fig. 12: The ratio's between the measured and actual length of four experiments, three with a ruler at different lengths, two of the same truss hanging at the same height as the camera (rehung perfectly vertical between measurements) and one of the truss hanging 5 [cm] off-center from the camera.

missing when compared to Figure 10 (right). The results of these three datasets are given in Table I. The estimated length of the truss for these datasets also shows a similar systematic overestimation of the actual respective truss lengths.

	Measured length [cm]	Real length [cm]
Bottom two tomatoes missing	17.0±0.3	15.3±0.2
Top one tomato missing	19.0±0.2	17.0±0.2
Top two tomatoes missing	16.9±0.4	15.7±0.2

TABLE I: Measurements of 3 trusses that are missing tomatoes in the truss as displayed in Figure 10 (right), showing a similar systematic error factor in the measured length using the SME minimization and EKF versus the real truss length.

The systematic error is found to have (at least) two types of contributions: firstly the (non-)linear camera performance and secondly the geometrical position deviations like off-center placement of the camera with respect to the truss and/or non-vertical hanging trusses, and the effect of the diameter of the tomatoes.

1) The Intel® RealSense™ D435 is to be calibrated with a linear conversion factor of $\frac{69.4}{640}$ $[\frac{^\circ}{\text{pixel}}]$ and is used in VGA mode. This has been checked by placing the camera at a certain distance from a ruler, see results in Figure 12. At large distances (right side in Figure 12) the sizes on the ruler are measured in the center of the camera and show a ratio of about 1.15 [-] with respect to the real length. Hence, this camera calibration needs to be corrected, where OEM calibration is described to be done in all modes at 60 [cm] at 15 [°] off-center in FoV [51]. Note that the VGA image has a lower aspect ratio compared with (full) HD and hence is cropped in its horizontal angle (vertical in our application). When the camera is moving closer to the ruler, the measured ratio increases which indicates that lens aberrations of the camera start to contribute. This aberration is known as pincushion distortion where the deviation increases quadratically with the distance to the center of the image [53]. The guards for the bounding box of the truss ensure the truss to fill > 70 [%] of FoV, driving the measurements in the range where distortion contributes. The typical camera distance during rotation around the truss is about 20 to 30 [cm] corresponding to a correction factor of 1.3 to 1.2 (see Figure 12). This would result in a correction of the estimated length of 21.0 ± 0.3 [cm] into 16.8 ± 0.7 [cm], which is then significantly shorter than the real truss length of 19.1 ± 0.1 [cm].

As a next step the geometrical contributions are investigated for potential contributions to the remaining systematic error.

1) In the ideal case the camera is perfectly aligned to the center of the truss height, whereas in practice an off-center placement error in the vertical direction can occur. For the extreme case as shown in Figure 13 (top-left) where the camera is aligned to the bottom of the truss. The effective length shrinks when this offset increases due to effectively increasing distance to the center of the truss to the camera. Theoretically, the delta in truss length for a vertical off-set of 5 [cm] is 0.9 and 0.5 [cm] respectively at distances of 20 and 30 [cm] for a 19.1 [cm] truss. This has been tested in an calibration test where the center of the truss was 5 [cm] above the camera (which was slightly tilted with a pitch to keep the truss within the FoV at smaller distance), see light blue curve in Figure 12. As can be seen at large distance the same calibration correction factor of 1.15 [-] is observed, but at smaller distance this factor is not increasing due to distortion but even slightly decreasing by about 10 [%] (which is about double of the calculated contribution).

2) The truss does not hang completely vertically, or for example the tomatoes of the bounding box do not hang at the same distance (e.g., from a side-view, the top tomato and bottom tomato hang at different distances to the camera) (Figure 13 (bottom-left)). The length of the truss effectively shrinks, as the vertical length decreases due to the angle. Normal ranges of the tilt of the truss are 0 to 10 [°]. For a 10 [°] tilt of a 19.1 [cm] truss, the delta in estimated truss length is 0.1 and 0.2 [cm], respectively, for distances of 20 and 30 [cm] when viewed from a frontal view. From a side view, the difference in estimated length is 0.3 [cm].
3) The tomatoes can be represented by spheres with a certain radius. Due to this radius, the tangent moves forward with the increase of the angle from the camera, i.e., the closer the camera is and thus the wider the angle is, the closer the truss appears to be (Figure 13 (right)). The ratio between the measured and actual length of the truss is affected, estimating the truss to be smaller than it is in reality (Figure 12). This effect thus lowers the ratio. For a vertical-hanging truss of 19.1 [cm], and a tomato radius of 2 [cm] and with the vertically-centered camera at a distance between 20 and 30 [cm], the effective difference in estimated size is 0.2 [cm].

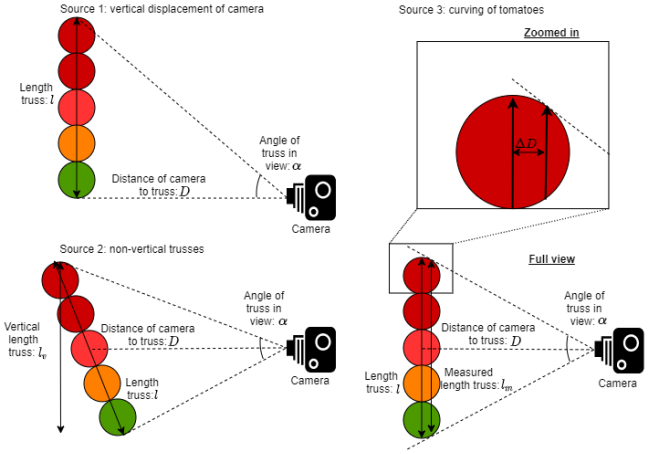

Fig. 13: Simplified drawings of the three geometrical sources of the systematic error in the length estimation of the tomato truss: the difference in vertical camera placement between the normal placement (Figure 7) and an extreme variant (left-top), the difference between vertical-hanging normally-placed trusses and trusses hanging at an angle (left-bottom) and the difference due to the curving of the tomatoes including a zoomed-in view (right).

The combined effect of the four sources creates a complex mix where the ratio depends on the magnitudes of the different disturbances and their interdependent influences. Taking all sources into account, the ratio between the measured and estimated length might be between 1.3 and 1.15 [-], thus estimating the original truss to be respectively 16.2 and 18.3 [cm]. Note that after this correction the highest value of 18.3 [cm] is still lower than the actual length of 19.2 [cm], which indicates the presence of other systematic errors. The same holds for the variety of other trusses that have been experimented with (Table I). Additionally, the effect of ambient temperature has been tested on the camera performance, however this caused no change in calibration or drift. Therefor, although large parts of the initial systematic error have been investigated, the complete systematic error has not been fully identified.

The distance estimation is based on the approximation of the length of the truss and the measured angle (Equation 2). This current model of the tomato truss and these dynamics of the interaction between the camera movement and the truss overestimate the length of the truss with an average error of 9.8 ± 2.1 [%] between the estimated and actual length based on six experiments as described above. The distance to the truss is identically overestimated by 2.5 ± 0.5 [cm] for a typical camera-to-truss rotation distance of 25 [cm], since errors form the truss length estimate propagate linearly into this distance measurement (Equation 2). In case where this accuracy is insufficient the four aforementioned sources are to be compensated for: the camera is to be well-calibrated and the dynamics model is required to be extended in order to include the three geometric sources as described above, and if necessary, including adjustments in the modelling of the tomato truss e.g,. gaining extra knowledge from the vertical angle in which the truss hangs (source 3) instead of solely taking into account the bounding box, or utilizing a model where individual tomatoes are used to estimate the distance through tracking, as the chosen tomatoes can be centered on the image and thus have minimal distortion (minimizing source 1). Nevertheless, the current measurement setup has demonstrated an systematic overestimation of the distance by 9.8 ± 2.1 [%] by solely utilizing RGB-camera images and the positions of the camera: a good starting point for the next process step of harvesting.

VI. Conclusion and Future Work

This novel concept of active camera positioning with guarded motion as motion controller based on the gathered data combined with the a-priori geometrical knowledge from the plants in the greenhouse has been tested using an artificial test setup with limited variation. The camera position in front of the truss has been reached by optimizing the motion such that truss is seen at its widest compared to its length. Robustness against truss variations (different amount of tomatoes, truss lengths and compositions) has been displayed in the experiments, as long as the frontal view on the truss has the maximum normalized width. The size of the individual tomatoes in this frontal position has been checked to be sufficient for ripeness detection. For the next process step to harvest the tomato truss the Euclidean distance to the truss is relevant. Although this harvesting step is for future research and development, this research has investigated the accuracy of the distance estimation which can be reached by only using the positions of the camera and the images at these positions,

where the position of and distance to the truss are thus a-priori unknown. First, the truss length is estimated and compared to its real length. It has been found that the truss length is about 9.8 ± 2.1 [%] *higher* than their actual lengths. This structural error has been investigated, and next to distortion and incorrect calibration of the camera image, also the offset position and angle of the camera with respect to the truss and even the size of the tomato have a relevant contribution. With calibration correction including distortion and geometric error sources (see Figure 12), this estimated length is estimated to be between 15.2 and 4.5 [%] *lower* than the actual truss length. Note that the interdependencies between these geometrical deviations could play a role in this remaining systematical error, which would require further investigation. Secondly, the distance from the camera position to the truss is solely deduced from the camera image using the actual truss height and its angle to convert this into the Euclidean distance to the truss. Since the angle measurement is directly measured and accurate (< 3 [°] error on the sensor), the same structural offset error is transferred to the Euclidean distance of the truss to the frontal camera position, which for a typical distance of 25 [cm] yields a position inaccuracy of 2.5 ± 0.5 [cm]. Requirement analysis of the next process step to harvest the truss needs to show whether this provided position accuracy is sufficient.

Practically, the future work should be to continue testing in a greenhouse. The active camera placement method has presently been tested in an artificial setting, and it should be further validated in the greenhouse which has natural variation (e.g., in species, size and amount of tomatoes, truss orientations, length of truss, partial occlusions). Preferably, this implementation should utilize a robot arm (5 DOF) which would significantly improve the accuracy regarding the positioning of the camera. The accuracy of the distance estimation needs to be assessed whether it is at the current level adequate for subsequent harvesting (usecase 3, section II). Additionally, a correctly calibrated camera without significant lens distortion is recommended to minimize systematic corrections. Furthermore, this manner of active camera placement can be further investigated to compare it to currently implemented methods in terms of reliability, time usage and accuracy.

Fundamental investigations in the future work can cover other use-cases within the greenhouse or industrial setting with limited variation can be researched utilizing a similar method, for example: deleafing (use-case 1, section II), or tomato truss harvesting with peduncle cutting (use-case 3, section II). Note that other potential industry applications for this active perception method could be e.g., chicken filet cutting (differing in orientation of the chicken), or chips bag sorting (differing in orientation and shape of the bags), etc. The deleafing task is currently done through a combination of visual and tactile information by the human workers as mentioned in section II, and hence a purely image-based implementation is not likely. However, an active sensor placement method can be created by combining tactile and visual sensors, and by determining and simplifying the geometrical relations and determining a simple maximization utilizing guarded motion motion control. E.g., a tactile mechanical gripper concept with flex-hinges could be used such that it centers around the peduncle and can cut the peduncle at about 1 [cm] from the stem. Note that the peduncle and truss can sustain a small motion by the gripper as long as it is not damaged. Of course, other sensors (e.g., tactile or depth) providing more accurate and direct depth information could be explored in case this would be necessary. For the peduncle recognition and also for the initial tomato truss detection investigations are also required to generate a robust automated object detection such that the process of robotized harvesting can be completed successfully.

ACKNOWLEDGMENTS

Throughout my thesis I have received a great deal of support and assistance from my supervisors René van de Molengraft, Herman Bruyninckx and Jordy Senden, for which I am very thankful since it enabled me to grow and develop my research and personal skills. Foremost, I would like to express my sincere gratitude to my supervisor René for his continuous focus on defining the requirements and placing the research in a broader framework. Furthermore, I am thankful for the new insights I obtained from Herman, who always challenged me to take a deeper look at the implementations, sharpen my thinking and bringing my research to a higher level. I would like to thank Jordy for his patient support, his guidance and inputs throughout each stage of the process and for all the opportunities I was given to develop myself. My sincere thanks also goes to Ronald Zeelen from Priva, whose help was pivotal in gaining insight concerning the use-cases and helped me further define my project. Next to this, I am grateful for the opportunity to have had a tour in the Vereijken greenhouses by Gert-jan van Rixtel, which helped me to gain a more practical view of the greenhouses and helped me further refine my use-cases as explained by Ronald Zeelen. I would like to thank Xin Wang, Akshay Burusa, David Rapado Rincón and Robert van der Kruk for sharing the knowledge gained from their respective projects and for brainstorming and discussing about the application of active perception. Of course this acknowledgement would not be complete without expressing my gratitude to my girlfriend Maggie, who has been patient and loving throughout my thesis, and my family and friends, for their positive energy.

REFERENCES

[1] O. Sakaue and S. Hayashi, "Tomato harvesting by robotic system", ASAE, pp. 96-3067, 1996.

[2] N. Kondo, Y. Nishitsuji, P. Ling and K. Ting, "Visual Feedback Guided Robotic Cherry Tomato Harvesting", Transactions of the ASAE, vol. 39, no. 6, pp. 2331-2338, 1996. Available: 10.13031/2013.27744.

[3] E. van Henten, "Greenhouse mechanization: state of the art and future perspective.", International Symposium on Greenhouses, Environmental Controls and In-house Mechanization for Crop Production in the Tropics, no. 710, pp. 55-70, 2004. Available: 10.17660/actahortic.2006.710.3.

[4] N. Nilsson, "A mobile automato: an application of artificial intelligence techniques", Sri International Menlo Park Ca Artificial Intelligence Center, 1969.

[5] R. Bajcsy, Y. Aloimonos and J. Tsotsos, "Revisiting active perception", Autonomous Robots, vol. 42, no. 2, pp. 177-196, 2017. Available: 10.1007/s10514-017-9615-3.

[6] M. Roser, "Employment in Agriculture", Our World in Data, 2013. [Online]. Available: https://ourworldindata.org/employment-in-agriculture.

[7] M. Castillo and S. Simnitt, "Farm Labor", Ers.usda.gov, 2020. [Online]. Available: https://www.ers.usda.gov/topics/farm-economy/farm-labor/.

[8] M. Roser, "World Population Growth", Our World in Data, 2013. [Online]. Available: https://ourworldindata.org/world-population-growth.

[9] C. den Hartog and G. van Rixtel, "Report visit Vereijken greenhouse," Technical University Eindhoven, unpublished, 2020.

[10] H. Kurosaki, H. Ohmori, H. Hamamoto and Y. Iwasaki, "Work hours and yield for large-scale tomato production in Japan", Acta Horticulturae, no. 1037, pp. 753-758, 2014. Available: 10.17660/actahortic.2014.1037.98.

[11] C. Gagne, N. Mattson and D. Kovach, "Your guide to High-wire Tomato Growing", Greenhousegrower.com, 2020. [Online]. Available: https://www.greenhousegrower.com/production/your-guide-to-high-wire-tomato-growing/.

[12] N. El-Bendary, E. El Hariri, A. Hassanien and A. Badr, "Using machine learning techniques for evaluating tomato ripeness", Expert Systems with Applications, vol. 42, no. 4, pp. 1892-1905, 2015. Available: 10.1016/j.eswa.2014.09.057.

[13] F. Zhang, "Ripe Tomato Recognition with Computer Vision", Proceedings of the 2015 International Industrial Informatics and Computer Engineering Conference, 2015. Available: 10.2991/iiicec-15.2015.107.

[14] A. Arefi, A. Motlagh, K. Mollazade and R. Teimourlou, "Recognition and localization of ripen tomato based on machine vision", Australian Journal of Crop Science, vol. 5, no. 10, pp. 1144-1149, 2011.

[15] "Plantalyzer® - HortiKey", HortiKey, 2020. [Online]. Available: https://hortikey.nl/plantalyzer/.

[16] C. Ji, J. Zhang, T. Yuan and W. Li, "Research on Key Technology of Truss Tomato Harvesting Robot in Greenhouse", Applied Mechanics and Materials, vol. 442, pp. 480-486, 2013. Available: 10.4028/www.scientific.net/amm.442.480.

[17] H. Yin, Y. Chai, S. X. Yang and G. S. Mittal, "Technical Note: Ripe Tomato Detection for Robotic Vision Harvesting Systems in Greenhouses", Transactions of the ASABE, vol. 54, no. 4, pp. 1539-1546, 2011. Available: 10.13031/2013.39005.

[18] Y. Zhao, L. Gong, Y. Huang and C. Liu, "Robust Tomato Recognition for Robotic Harvesting Using Feature Images Fusion", Sensors, vol. 16, no. 2, p. 173, 2016. Available: 10.3390/s16020173.

[19] E. Pekkeriet, "Glas 4.0: 4 robotprojecten + (PicknPack, Phenobot, Trimbot & Sweeper) [Presentatieslides]", Glas40.nl, 2018. [Online]. Available: https://glas40.nl/Portals/42/presentaties/Act/Robot-projects-WUR-Glas4.0%20Public%20-%20Erik%20Pekkeriet.pdf?ver=2018-05-31-140659-230.

[20] C. Morar, I. Doroftei, I. Doroftei and M. Hagan, "Robotic applications on agricultural industry. A review", IOP Conference Series: Materials Science and Engineering, vol. 997, no. 1, p. 012081, 2020. Available: 10.1088/1757-899x/997/1/012081.

[21] Y. Zhao, L. Gong, C. Liu and Y. Huang, "Dual-arm Robot Design and Testing for Harvesting Tomato in Greenhouse", IFAC-PapersOnLine, vol. 49, no. 16, pp. 161-165, 2016. Available: 10.1016/j.ifacol.2016.10.030.

[22] N. Kondo, M. Monta, T. KC, G. GA and L. PP, "Harvesting robot system for single truss upside down tomato production", JOURNAL of the JAPANESE SOCIETY of AGRICULTURAL MACHINERY, vol. 58, pp. 467-470, 1996. Available: https://doi.org/10.11357/jsam1937.58.Supplement_467.

[23] A. Leichman, "World's first tomato-picking robot set to be rolled out - ISRAEL21c", ISRAEL21c, 2019. [Online]. Available: https://www.israel21c.org/israeli-startup-develops-first-tomato-picking-robot/.

[24] C. Ji, J. Zhang, T. Yuan and W. Li, "Research on Key Technology of Truss Tomato Harvesting Robot in Greenhouse", Applied Mechanics and Materials, vol. 442, pp. 480-486, 2013. Available: 10.4028/www.scientific.net/amm.442.480.

[25] "Introducing AI-equipped Tomato Harvesting Robots to Farms May Help to Create Jobs — Solutions Wherever You Go — Panasonic Newsroom Global", Panasonic Newsroom Global, 2018. [Online]. Available: https://news.panasonic.com/global/stories/2018/57801.html.

[26] L. Luo, Y. Tang, Q. Lu, X. Chen, P. Zhang and X. Zou, "A vision methodology for harvesting robot to detect cutting points on peduncles of double overlapping grape clusters in a vineyard", Computers in Industry, vol. 99, pp. 130-139, 2018. Available: 10.1016/j.compind.2018.03.017.

[27] T. de Haan, P. Kulkarni and R. Babuska, "Geometry-based grasping of vine tomatoes", 2021. Available: arXiv:2103.01272.

[28] E. Black and L. Kolodny, "This robot can pick tomatoes without bruising them and detect ripeness better than humans", CNBC, 2019. [Online]. Available: https://www.cnbc.com/2019/05/11/root-ai-unveils-its-tomato-picking-robot-virgo.html.

[29] "Sweet Pepper Harvesting Robot", Sweeper-robot.eu, 2018. [Online]. Available: http://www.sweeper-robot.eu/.

[30] "The Future of Fresh Fruit Harvest", Ffrobotics.com, 2019. [Online]. Available: https://www.ffrobotics.com/.

[31] "Abundant Robotics", Abundantrobotics.com, 2017. [Online]. Available: https://www.abundantrobotics.com/.

[32] J. Kollewe and R. Davies, "Robocrop: world's first raspberry-picking robot set to work", the Guardian, 2019. [Online]. Available: https://www.theguardian.com/technology/2019/may/26/world-first-fruit-picking-robot-set-to-work-artificial-intelligence-farming.

[33] "Fruit picking robots", Into Robotics, 2017. [Online]. Available: https://www.intorobotics.com/fruit-harvesting-robots/.

[34] S. Staff, "Japan robot can pick strawberry fields forever for farmer", Phys.org, 2013. [Online]. Available: https://phys.org/news/2013-09-japan-robot-strawberry-fields-farmer.html.

[35] "Agricultural Robotics", Harvestcroo.com, 2013. [Online]. Available: https://harvestcroo.com/.

[36] "Octinion presents the world's first strawberry picking robot", Octinion, 2019. [Online]. Available: http://octinion.com/news/press-release-octinion-presents-world%E2%80%99s-first-strawberry-picking-robot.

[37] "Priva Kompano Deleaf Line", Priva, 2016. [Online]. Available: https://www.priva.com/nl/ontdek-priva/blijf-op-de-hoogte/nieuws/priva-kompano-deleaf-line.

[38] E. Van Henten et al., "An Autonomous Robot for De-leafing Cucumber Plants grown in a High-wire Cultivation System", Biosystems Engineering, vol. 94, no. 3, pp. 317-323, 2006.

[39] T. Botterill et al., "A Robot System for Pruning Grape Vines", Journal of Field Robotics, vol. 34, no. 6, pp. 1100-1122, 2016. Available: 10.1002/rob.21680

[40] C. den Hartog and R. Zeelen, "Report conversation Ronald Zeelen," Technical University Eindhoven, unpublished, 2020.

[41] P. Roy and V. Isler, "Active view planning for counting apples in orchards", 2017 IROS, pp. 6027-6032, 2017. Available: 10.1109/iros.2017.8206500.

[42] J. Vasquez-Gomez, L. Sucar, R. Murrieta-Cid and J. Herrera-Lozada, "Tree-based search of the next best view/state for three-dimensional object reconstruction", International Journal of Advanced Robotic Systems, vol. 15, no. 1, 2018. Available: 10.1177/1729881418754575.

[43] E. Dunn, J. van den Berg and J. Frahm, "Developing visual sensing strategies through next best view planning", 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4001-4008, 2009. Available: 10.1109/iros.2009.5354179.

[44] K. Pratt and R. Murphy, "Protection from Human Error: Guarded Motion Methodologies for Mobile Robots", IEEE Robotics & Automation Magazine, vol. 19, no. 4, pp. 36-47, 2012. Available: 10.1109/mra.2012.2220500.

[45] P. Sharma, "Image Classification vs Object Detection vs Image Segmentation", Medium, 2019. Available: https://medium.com/analytics-vidhya/image-classification-vs-object-detection-vs-image-segmentation-f36db85fe81.

[46] M. Rajchl et al., "DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks", IEEE Transactions on Medical Imaging, vol. 36, no. 2, pp. 674-683, 2017. Available: 10.1109/tmi.2016.2621185.

[47] S. Kourouklis, "A New Estimator of the Variance Based on Minimizing Mean Squared Error", The American Statistician, vol. 66, no. 4, pp. 234-236, 2012. Available: 10.1080/00031305.2012.735209.

[48] G. Ellis, "Observers in Control Systems". Burlington: Elsevier, 2002.

[49] B. Schnekenburger, "A modified extended kalman filter as a parameter estimator for linear discrete-time systems", Master, New Jersey Institute of Technology, 1988.

[50] Intel® RealSense™ D400 Series Product Family. 2019. Available: intel.com/content/dam/support/us/en/documents/emerging-technologies/intel-realsense-technology/Intel-RealSense-D400-Series-Datasheet.pdf.

[51] "Best Known Methods for Optimal Camera Performance over Lifetime – Intel® RealSense™ Depth and Tracking Cameras", Intel® RealSense™ Depth and Tracking Cameras, 2021. [Online]. Available: https://www.intelrealsense.com/best-known-methods-for-optimal-camera-performance-over-lifetime/#3.6.

Note: calibration is dependent on display format. The calibration is noticed to be different in VGA-mode, leading to cropping of the horizontal FoV of 69.4 [°] to approximately 52.1 [°].

[52] C. Johnson, "Optimal initial conditions for full-order observers", International Journal of Control, vol. 48, no. 3, pp. 857-864, 1988. Available: 10.1080/00207178808906222.

[53] "Distortion (legacy) Module", imatest, 2020. [Online]. Available: https://www.imatest.com/docs/distortion_instructions/.

[54] I. Reid and H. Term, "Estimation ii", University of Oxford, 2001. [Online]. Available: https://www.robots.ox.ac.uk/~ian/Teaching/Estimation/LectureNotes2.pdf.

APPENDIX A
THE EXTENDED KALMAN FILTER

The Extended Kalman Filter consists of a predict step (Equation 7) and an update step (Equation 9). The predict step utilizes the dynamics of the sytem (Equation 3) in order to predict the state $\hat{x}_{i|i-1}$, and predict the covariance estimate $P_{i|i-1}$, where Q is the process noise covariance matrix [49]. The covariance estimates P indicate the accuracy belonging to certain state estimates, and can be initialized through the average error between the initial value and the initial estimate (Equation 5) [54]. The process noise covariance matrix Q is the average size of the process noise w squared, having a mean of zero [54].

$$P_{0|0} = E\left[\left(\mathbf{x}\left(t_0\right) - \hat{\mathbf{x}}\left(t_0\right)\right)\left(\mathbf{x}\left(t_0\right) - \hat{\mathbf{x}}\left(t_0\right)\right)^T\right] \tag{5}$$

$$E\left[w_k w_l^T\right] = \begin{cases} Q_k & k = l \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$\hat{x}_{i|i-1} = f(\hat{x}_{i-1|i-1}) \tag{7a}$$

$$F_i = \left.\frac{\delta f}{\delta x}\right|_{\hat{x}_{i-1|i-1}} \tag{7b}$$

$$P_{i|i-1} = F_i P_{i-1|i-1} F_i^T + Q \tag{7c}$$

The update step (Equation 9) calculates the gain K_i based on the updated covariance estimate $P_{i|i-1}$ and utilizes the new measurement error $\tilde{y}_i$ in order to update the states $\hat{x}_{i|i}$ and covariance estimates $P_{i|i}$, where R is the measurement noise covariance and σ_m is the standard deviation of the measurement error (Equation 8) [49]. The matrix H is defined by $y_i = Hx_{i_i}$, meaning $H = \frac{1}{ppd}$.

$$R = \sigma_m^T \sigma_m \tag{8}$$

$$\tilde{y}_i = y_i - H\hat{x}_{i|i-1} \tag{9a}$$

$$K_i = P_{i|i-1} H^T \left(H P_{i|i-1} H^T + R^T\right)^{-1} \tag{9b}$$

$$P_{i|i} = \left(I - K_i H\right) P_{i|i-1} \tag{9c}$$

The eventual state estimation $\hat{x}_{i|i}$ is then calculated as given in Equation 10, utilizing the feedback gain K_i and the system as described in Equation 3, including the length of the truss l as a state.

$$\hat{x}_{i|i} = \hat{x}_{i|i-1} + K_i \tilde{y}_i \tag{10}$$