Robust Visual Servo Control for Robotic Fruit Picking in the Presence of Parametric Uncertainties and Complex Scenarios

Juan D. Gamba and Antonio C. Leite

Abstract— **In this paper, we address a robust hybrid visual servoing problem for a fixed target using an eye-in-hand camera configuration, when the camera calibration parameters are uncertain. The proposed solution combines the strengths of image-based and position-based visual servoing approaches, by defining a control error in both the image and task spaces, in order to perform successful robotic fruit harvesting tasks. A pre-trained Deep Convolutional Neural Network (DCNN) encoder-decoder based on a minimized Segnet version is used to perform the strawberries segmentation and extraction from complex backgrounds. An image-based localization method, based on ORB and BFMatcher algorithms, are used to extract the image feature of the fruit for the vision-based control algorithm. A robust control design is employed to cope with the uncertainties in the calibration parameters of the camera-robot system. To deal with possible singular configurations that may arise during the task execution, we employ an inverse-kinematics algorithm based on the transpose Jacobian and interaction matrices. Simulation results are included to demonstrate the effectiveness of the proposed methodology.**

I. INTRODUCTION

Over the last years, there is a strong trend towards the designing of autonomous robotic systems able to perform a wide range of agricultural tasks in orchards, vineyards, polytunnels and farms [1]. For example, weeds species recognition and killing, soft-fruit recognition and picking, as well as plant phenotyping are just a few examples of how robots are ruling fields around the world. The agricultural environment introduces several challenges and difficulties, particularly, for robotic harvesting and 3D navigation of mobile robots [2]. Indeed, changes in seasons and weather conditions, crop growth and rotation, dense vegetation, different maturity levels of fruits, the existence of diseases and fungi in plants, all these factors create a dynamic and poorly structured environment. Thus, the automatic fruit harvesting system has to incorporate perception and cognition capabilities in the gripper design [3] as well as intelligent sensors and systems for fruit detection, recognition and localisation [4].

Applications for image segmentation have demonstrated the effectiveness of classical computer vision algorithms for detecting and localizing objects in well structured environments. Applications for object segmentation have demonstrate the effectiveness of classic computer vision algorithms for detecting and localizing objects in very controllable situations. Mehta and Burks [5] have designed a vision-based estimation and control system for robotic citrus harvesting based on the combination of large field-of-view of a fixed camera and the accuracy of a mobile camera. Barth *et al.* [6] have proposed a visual servoing approach which uses the eye-to-hand camera configuration for sweet pepper harvesting in dense vegetation. Image recognition and segmentation for fruit detection in complex scenarios is still considered as an open research problem, due to the occlusions, variable lighting conditions and presence of modeling uncertainties [4]. Fruit recognition and segmentation applications into non-controllable scenarios still considered as an open research topic, due to the high exposure to noise, light changes and complex uncertainties. Deep encoder-decoder architectures have been used recently to perform semantic segmentations using complex backgrounds, due to its ability for learning textures and image features of a given interest object. These algorithms may be promising for carrying out detection and localization tasks on unstructured scenarios [7]. Deep encoder-decoder architectures have been used recently to perform semantic segmentations into very complex backgrounds, due to its ability for learning textures and image features of a given interest object, these algorithms demonstrate the possibility of carrying out detection and localization tasks into non-controllable scenarios [7]. Hung *et al.* [8] introduce a segmentation scheme using multispectral images, sparse autoencoders, and Support Vector Machine (SVM) schemes to segment leaves. Dias *et al.* [9] have proposed a robust flower identification algorithm based on Fully Convolutional Neural Networks (FCN), to demonstrate how FCNs are able to deal with challenging image segmentation tasks. Dias *et al.* [9] have proposed a robust flower identification algorithm based on fully convolutional neuronal networks (FCNs), they have demonstrated how FCNs are able to deal with very challenging segmentation assignments.

In this work, we address the soft-fruit harvesting problem by using a visual servoing approach based on the combination of computer vision, machine learning, and control theory methodologies. A robust vision-based control scheme which combines the Image-Based Visual Servoing (IBVS) and the Position-Based Visual Servoing (PBVS) approaches is designed to cope with parametric uncertainties in the camera-robot system. In order to detect and localize different strawberries into a scene, we use a pre-trained DCNN encoder-decoder algorithm [10]. To achieve the segmentation stage, the fruit localization method uses a combination of the Oriented FAST and Rotated BRIEF (ORB) [11] and Brute-Force Matching (BFMatcher) algorithms, due to their well-known properties of robustness, fastness, and accuracy. 3D Computer simulations results obtained with a Mitsubishi RV-

[1]Juan D. Gamba and Antonio C. Leite are with the Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Postal Code 22451-900, Rio de Janeiro RJ, Brazil. email: juan212@ele.puc-rio.br, antonio@ele.puc-rio.br

2AJ robot, performing a picking task of multiple and isolated strawberries (to avoid clustering and other more complex situations); are also included to illustrate the performance and effectiveness of the proposed visual servoing scheme.

II. PROBLEM FORMULATION

In this work, we address the robotic fruit harvesting problem using a visual servoing scheme with an RGB-D stereo camera mounted on the robot end-effector (Fig. 1). Here, the following notation is considered: $p_{ij} \in \mathbb{R}^3$ and $R_{ij} \in \mathbb{SO}(3)$ denote respectively the position vector and orientation matrix of the frame $\mathcal{F}_j$ with respect to frame $\mathcal{F}_i$; $T_{ij} \in \mathbb{R}^{4\times 4}$ is the homogeneous transformation matrix, which denotes the pose of the frame $\mathcal{F}_j$ with respect to frame $\mathcal{F}_i$. In this context, the pose of the camera frame $\mathcal{F}_c$ with respect to the base frame $\mathcal{F}_b$ is given by $T_{bc} = T_{be}\, T_{ec}$, say:

$$T_{bc} = \begin{bmatrix} R_{bc} & p_{bc} \\ 0^{\mathsf{T}} & 1 \end{bmatrix} = \begin{bmatrix} R_{be}\, R_{ec} & R_{be}\, p_{ec} + p_{be} \\ 0^{\mathsf{T}} & 1 \end{bmatrix} \quad (1)$$

Here, we assume that (A1) the homogeneous transformation matrix T_{be} can be obtained from the *forward kinematics* map by using, for example, the Denavit-Hartenberg convention. In this case, implies that $p_{be} = p_{be}(q)$ and $R_{be} = R_{be}(q)$. For simplicity, we also assume that (A2) the camera frame $\mathcal{F}_c$ and the end-effector frame $\mathcal{F}_e$ are aligned only with respect to z-axis, but the relative translation between their origins and the relative orientation of their z-axes, denoted by ϕ, may be uncertain. In this context, implies that $R_{ec} = R_{ec}(\phi)$.

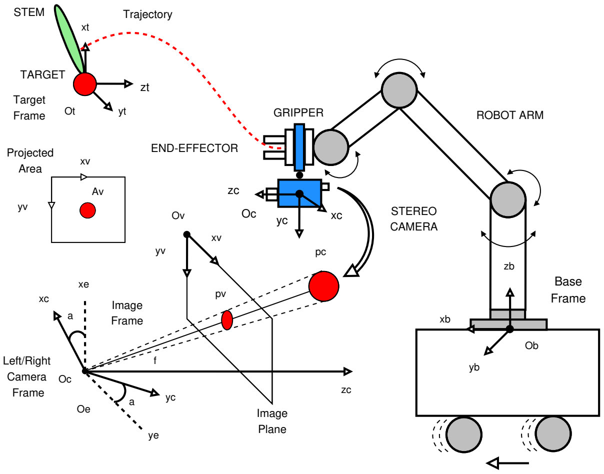

Fig. 1: Visual servoing system for robotic harvesting tasks.

The fruit harvesting task we consider consists of moving the robot arm to the vicinity of the fruit, cut the stem using a suitable device attached to the robot end-effector and store the fruit in a storage device. The first goal is to solve the image segmentation and interpretation problem, that is, to detect and recognize the target object located in the robot workspace by using a fully convolutional neural network (FCNN) algorithm. Once the target object is tracked by using a stereo vision system, the next step is to solve the correspondence and 3D reconstruction problem, that is, to compute the 3D coordinates of the fruit with respect to the camera by using, for example, a simple triangulation technique [12]. Thus, the pose of the target frame $\mathcal{F}_t$ with respect to the base frame $\mathcal{F}_b$ is given by $T_{bt} = T_{bc}\, T_{ct}$, say:

$$T_{bt} = \begin{bmatrix} R_{bt} & p_{bt} \\ 0^{\mathsf{T}} & 1 \end{bmatrix} = \begin{bmatrix} R_{bc}\, R_{ct} & R_{bc}\, p_{ct} + p_{bc} \\ 0^{\mathsf{T}} & 1 \end{bmatrix}. \quad (2)$$

where T_{ct} is the homogeneous transformation matrix whose entries can be computed from the application of the FCNN algorithm and the triangulation technique. Finally, since the homogeneous transformation matrix T_{bt} is computed, we can employ an *inverse kinematics-based algorithm* to transform the motion specifications, assigned to the robot end-effector in the task space, into the corresponding joint space motions, allowing for the successful execution of the desired motion.

III. SEGMENTATION APPROACH

The following section illustrates how the DCNN encoder-decoder is used to learn a certain manifold or small sets of such manifold, this applied to a segmentation assignment, which consists in a pixel-wise classification task. A simplified DCNN encoder-decoder version based on SegNet architecture is used to perform the segmentation process. At the encoder side, there are four convolutional layers with a fixed kernel dimension of (7×7), activation function $ReLU$ [13] and filter with a dimension of 64, there are also three max-pooling layers after every convolutional layer to add small translation invariance to the model. Decoder side has four deconvolutional layers with same kernel dimensions to preserve the symmetry into the model, then three upsampling layers are added before every deconvolutional layer.

A. Trainning Data-set

For application needs, it is quite difficult to obtain an already-made data-set for strawberries semantic segmentation. A manual selection process with super-pixels [14] was used to create the different images annotations. An annotation relates to the desired output for each input image, it contains every pixel membership to a certain class.

Fig. 2: Data-set samples.

The custom data-set (Fig. 2) contains fifty images of 480×360 dimensions, where ninety and ten percent of it are used for training and validation purposes.

B. Training

During segmentation, pixels are classified into background, strawberry and strawberry leaf. By adopting a specific class for background, it is possible to facilitate the fruit extraction from background. The DCNN encoder-decoder is not only able to learn textures and features information from fruits but also from the background, that can become very complex in some situations. The algorithm obtains information from both classes which helps to make a more accurate segmentation in comparison with the OHTA cascade segmentation method introduced by Wei *et al.* [15]. Due to the low-sampling obtained from the custom dataset a $Dropout$ layer was added after every convolutional and deconvolutional layer, to avoid interdependent learning among the neurons and take more advantage of the encoder-decoder model [16].

Network training was carried out on a Desktop PC with an Intel Core i7-7700 Processor, 8 GB RAM and a Nvidia Geforce GTX 1080 GPU. Training was executed by running 220,000 steps, with four images by step, a dropout of 0.3 and a learning rate of 0.001, $Adam$ optimizer was chosen due to its computationally efficient architecture and ability to deal with very noisy and/or sparse gradients [17], training took around of twenty hours approximately into an Ubuntu 18.04 OS, Python, and TensorFlow-GPU framework.

After training the algorithm obtained an accuracy of 97% and a Mean Intersection Over Union ($MIoU$) of 60% wih training examples, with validation examples the algorithm obtained an accuracy of 96% and a $MIoU$ of 59.5%.

C. Results

The following section presents the results obtained with the simplified SegNet model at the strawberry segmentation task. Fig. 3 demonstrates the results obtained from the DCNN encoder-decoder algorithm with test images (unknown images for the network), images at (i) the column of the left side demonstrates the algorithm input X, (ii) the middle column denotes the expected output Y, (iii) the right column shows the algorithm output $\hat{Y}$. The background class is denoted by black color, the strawberry class by yellow color and the strawberry leaves class by blue color. In this context, it is also possible to see the difficulties to extract the strawberry leaves due to its big similarity to the background class. Furthermore, the algorithm is able to recognize and segment strawberries successfully, which demonstrates its capacity for generalizing over unknown data to make accurate predictions.

IV. VISUAL SERVOING APPROACH

In this work, we consider an RGB-D stereo camera attached to the robot end-effector. Let $p_{ct} = [\,x_c \;\; y_c \;\; z_c\,]^{\mathsf{T}}$ be the coordinates of a 3D point expressed in the camera frame $\mathcal{F}_c$. From the perspective projection model, the 3D point is projected in the image space as a 2D point with the coordinates $p_v = [\,x_v \;\; y_v\,]^{\mathsf{T}}$ expressed in pixels, say:

$$\begin{bmatrix} x_v \\ y_v \end{bmatrix} = \frac{f}{z_c} \begin{bmatrix} \alpha_x & 0 \\ 0 & \alpha_y \end{bmatrix} \begin{bmatrix} x_c \\ y_c \end{bmatrix} + \begin{bmatrix} x_{v0} \\ y_{v0} \end{bmatrix}, \qquad (3)$$

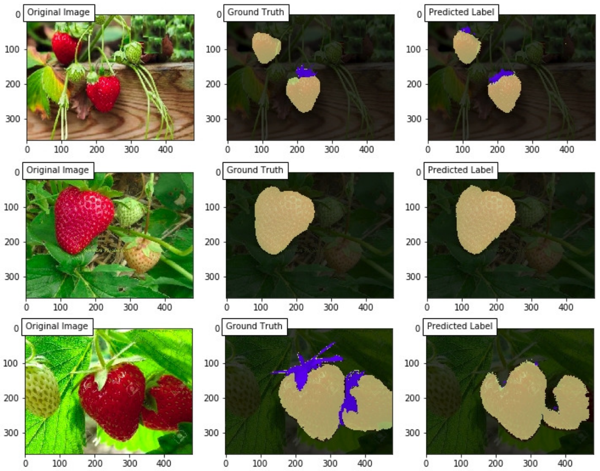

Fig. 3: Results test samples.

where $\{x_{v0}, y_{v0}, f, \alpha_x, \alpha_y\}$ is the set of camera intrinsic parameters: x_{v0} and y_{v0} are the coordinates of the principal point; f is the focal length; α_x and α_y are the scaling factors in pixel per millimeter. The 3D point is projected in the image plane as a 2D point with normalized coordinates $p_p = [\,x_p \;\; y_p\,]^{\mathsf{T}}$ given by:

$$x_p = \frac{x_v - x_{v0}}{f\alpha_x}\,, \qquad y_p = \frac{y_v - y_{v0}}{f\alpha_y}\,. \qquad (4)$$

Now, we suppose that the robot end-effector is moving with linear velocity $v_c \in \mathbb{R}^3$ and angular velocity $\omega_c \in \mathbb{R}^3$ both expressed with respect to the (instantaneous) camera frame $\mathcal{F}_c$. Then, using the well-known relationship of *velocity transformation* between the target frame $\mathcal{F}_t$ and the camera frame $\mathcal{F}_c$, we obtain the following motion equation [18]:

$$\dot{p}_{ct} = -v_c - Q(\omega_c)\, p_{ct}\,, \qquad (5)$$

with

$$v_c = R_{bc}^{\mathsf{T}}\, \dot{p}_{bc}\,, \qquad \omega_c = R_{bc}^{\mathsf{T}} Q({}^{b}\omega_{bc}) R_{bc}\,.$$

Now, the key idea consists of computing the position in the scene of the 3D points projected on the image plane of the two cameras using a triangulation technique [12]. Let $\bar{z}_c := \ln\left(z_c/z_d\right) \in \mathbb{R}$ be a supplementary normalized depth coordinate, where z_d is a depth scaling factor and $\ln(\cdot)$ denotes the natural logarithm function. Combining (4) into (5) and adding $\bar{z}_c$, yields:

$$\dot{w} = L_w\, \mathbf{v}_c\,, \qquad \begin{bmatrix} \dot{x}_p \\ \dot{y}_p \\ \dot{\bar{z}}_c \end{bmatrix} = L_w \begin{bmatrix} v_c \\ \omega_c \end{bmatrix}, \qquad (6)$$

with

$$L_w = \begin{bmatrix} -\dfrac{1}{z_c} & 0 & \dfrac{x_p}{z_c} & x_p\, y_p & -(1+x_p^2) & y_p \\ 0 & -\dfrac{1}{z_c} & \dfrac{y_p}{z_c} & (1+y_p^2) & -x_p\, y_p & -x_p \\ 0 & 0 & -\dfrac{1}{z_c} & -\dfrac{y_p}{z_c} & \dfrac{x_p}{z_c} & 0 \end{bmatrix},$$

where $L_w \in \mathbb{R}^{3\times 6}$ is the *interaction matrix* related to $w \in \mathbb{R}^3$, which denotes the 3D point coordinates expressed in the image and operational spaces.

Notice that, since the target object is assumed to be fixed with respect to the base frame $\mathcal{F}_b$ the desired values for image features are assumed to be constant, and changes in s and w depend only on camera motion.

A. HVS control design

The control goal is to drive a set of features w to the desired values of the hybrid features w_d say:

$$w \to w_d\,, \qquad e_w = w_d - w \to 0\,, \tag{7}$$

where $e_w \in \mathbb{R}^3$ is the hybrid image feature error. Since the camera is attached to the robot end-effector (i.e., eye-in-hand configuration) and $\mathbf{v}_c = R_{bc}^{\top}\,\mathbf{v}$, we can expand (6), obtaining the following control system:

$$\dot{w} = L_w\, R_{bc}^{\top}\, J(q)\, u\,, \tag{8}$$

where $J(q) \in \mathbb{R}^{6\times n}$ is the *geometric Jacobian* of the robot manipulator, and $u \in \mathbb{R}^n$ represents the joint control velocity signal of the robotic arm $u := \dot{q}$ [12]. From the time-derivative of (7), we can design the velocity control signal u as:

$$u := J^*(q)\, R_{bc}\, L_w^*\, \Lambda_w\, e_w\,, \qquad \Lambda_w = \Lambda_w^{\top} > 0\,, \tag{9}$$

where Λ_w is a proportional gain matrix, J^* and L_w^* are generic matrices to be properly defined in order to guarantee the asymptotic convergence of the image feature error e_w to zero, that is, $\lim_{t\to\infty} e_w(t) = 0$. Notice that, to avoid linearization of the error system, the algorithm can be designed using the *transpose* of the Jacobian and interaction matrices, $J^{\top}$ and $L_w^{\top}$. As a consequence, the algorithm is computationally more efficient and the error dynamics will be governed by a first-order nonlinear differential equation. In this case, implies that: (i) if $J^*(q) \equiv J^{\top}(q)$, it is possible to deal with kinematic singularities since the control algorithm does not require matrix inversion; (ii) if $L_w^* \equiv L_w^{\top}$, it is possible to cope with the ill-conditioning problem of the interaction matrix L_w, even if the HVS system uses only two image features to perform a 6-DoF task.

B. Stability and Robustness Analysis

In this section, we consider the fundamental issues related to the stability and robustness of the proposed HVS control scheme. As previously discussed, the designed control algorithm (9) is able to ensure the global or local stability of HVS system, provided that the calibration parameters of the camera-robot system are fully *known*. However, it is well-known that the complete knowledge of the camera calibration parameters and robot kinematics may not be satisfied from the practical point of view, particularly when the eye-in-hand camera configuration is used. Therefore, an approximation or an estimation of the interaction matrix L_w, the rotation matrix R_{bc} as well as the Jacobian matrix $J(q)$ must be realized [18]. In this case, the velocity control signal (9) takes the form:

$$u := \hat{J}^*(q)\, \hat{R}_{bc} \hat{L}_w^*\, \Lambda_w\, e_w\,, \tag{10}$$

where $\hat{R}_{bc} = R_{be}\hat{R}_{ec}(\phi)$, where $\phi \in \mathbb{R}$ is the misalignment angle between the camera frame $\mathcal{F}_c$ and the end-effector frame $\mathcal{F}_e$ around the z-axes, assumed to be uncertain. To analyse the stability and convergence properties of the HVS control scheme, we use the Lyapunov stability formalism. Consider the following positive-definite candidate Lyapunov function defined by $2V(e_w) = e_w^{\top}\Lambda_w e_w$. The time-derivative of V along the system trajectories (7), (8) and (10), is given by

$$\dot{V}(e_w) = -e_w^{\top}\,\Lambda_w\, L_w\, R_{bc}^{\top}\, J(q)\, \hat{J}^*(q)\, \hat{R}_{bc}\, \hat{L}_w^*\, \Lambda_w\, e_w\,. \tag{11}$$

The time-derivative of V is definite negative if the matrices $\hat{J}^*(q)$ and $\hat{L}_w^*$ are assumed to be full-rank, which can be guaranteed respectively if the robot arm has redundant degrees of freedom and the HVS control scheme uses one or more than two image features. The condition $\dot{V} < 0$ with $V > 0$ implies that the system trajectories uniformly converge to $e_w = 0$, that is, the error system is asymptotically stable.

Moreover, since the robot kinematic and camera calibration intrinsic parameters are positive values, the presence of uncertainties in any or both matrices, $J^*(q)$ and L_w^*, is not capable to violate the condition of negative definiteness (11). Accordingly, the misalignment angle ϕ between the camera and the end-effector frames needs to be less than $\frac{\pi}{2}$ rad, in order to ensure the positive-definiteness property of the matrix $\hat{R}_{ec}$. Under these assumptions it is possible to guarantee, the asymptotic stability of the HSV system for regulation tasks. Conversely, for tracking tasks, robust and adaptive control strategies could be used to overcome the restriction of the camera misalignment angle [19] and deal with the existence of parametric uncertainties in the camera-robot system [20], [21].

V. Design and Implementation

In this section, we describe the practical aspects for designing and implementation of the proposed Hybrid Visual Servoing (HVS) approach for robotic fruit harvesting. It is well-known that from the calculation of the same image features in both images obtained from the stereo vision system, we can compute the 3D point coordinates of the target in the operational space by using a triangulation technique [12]. Moreover, recognizing and matching points that belong to the same image feature in different scenes may be a difficult task, along with object extraction or image segmentation in complex backgrounds [22]. Segmentation and extraction processes are executed with a pre-trained DCNN encoder-decoder.

After extracting the object of interest from the scene, we use the ORB algorithm [11] in order to obtain features from both images and, then, the BF matcher algorithm is selected due to its simplicity and acceptable performance for identifying matching points in both images. BF matcher is a searching algorithm, which finds the closest descriptor in the second image set by point-by-point testing. Then the best ten results are chosen to triangulate the object Cartesian position. As noted, it is not possible to guarantee the absence of outliers during this process, thus, performing a triangulation with outliers features may result in a wrong target. Therefore, it is not always possible to ensure the system stability by performing an end-point open-loop PBVS control scheme, to reach the corresponding 3D point coordinates obtained from the triangulation.

Under this constraint, a HVS control scheme seems to be the more appropriate to ensure the reliability and safety of tasks performed by vision-based controllers. By having a stable control loop for collecting one strawberry, the real situation yields on collecting more than one strawberry in a complex real-world scenario, where many strawberries may be on the screen and a harvest-planning phase may not be trivial for collecting the strawberries as fast as possible. As expected, while performing the HVS control scheme to approach an object of interest, the others objects are considered as external disturbances to the control loop, which can lead to system instability.

In order to deal with the disturbances presented during the approaching phase towards a single object of interest in the presence of other fruits, a self-updating tracking window [23] is used to ensures the object visibility and be able to compute its corresponding image features. After reaching the desired distance, a final image-based height adjustment could be carried out for positioning the gripper close to the fruit stem, cut it and store the fruit in a storage box, completing the harvesting task successfully.

VI. CONTROL VERIFICATION AND VALIDATION

In this section, we present simulation results for a robotic fruit harvesting task. The simulations tests were carried out considering the presence of parametric uncertainties in the Jacobian matrix J, interaction matrix L_w and rotation matrix R_{ec}. We also assumed that robot end-effector is moving in the neighborhood of singular configurations.

The robustness of the control algorithm will be evaluated by choosing the misalignment angle between the camera and the end-effector frames around the z-axis as $\phi = \pi/6\ rad$ and 10% of uncertainty in the length of the last link of the robot arm as well as in the camera intrinsic parameters. The control goal is to drive the object image feature w to the desired image feature w_d located at the camera center point (x_{v0}, y_{v0}) with desired depth $z_d = 0.1\ m$. The numerical simulations were executed in MATLAB and V-REP robot simulator (see Fig.4). To illustrate the performance and effectiveness of the HVS control scheme, simulations results are shown in Fig. 5-7.

Fig. 5(a) and (b) shows the behavior over time of the position of the image feature and the position error during the regulation tasks, where we can observe a slight disturbance at the beginning of the simulation due to the existence of the misalignment between the camera and end-effector frames in the z-axis. We can also note the asymptotic convergence of both signals to zero. The time history of the HVS control signal is illustrated in Fig. 6(a) and (b), where it is possible to verify the stable behavior of the linear and angular velocity signals, obtained using the HVS control scheme. Fig. 7 shows the motion of multiple image features (strawberries) from a given initial position "$*$" to a desired final position "$\circ$", it is possible to verify the satisfactory performance of the proposed control scheme, in spite of the existence of parametric uncertainties in the camera-robot system.

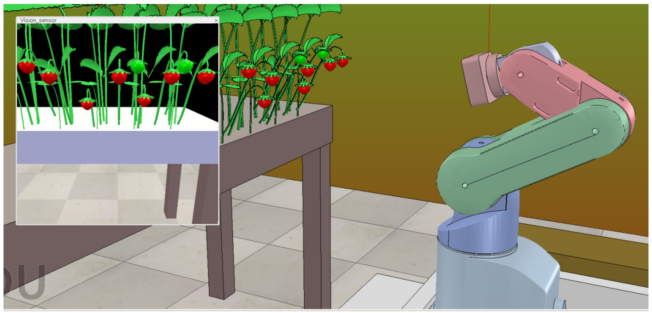

Fig. 4: Robotic fruit picking tasks on V-REP robot simulator.

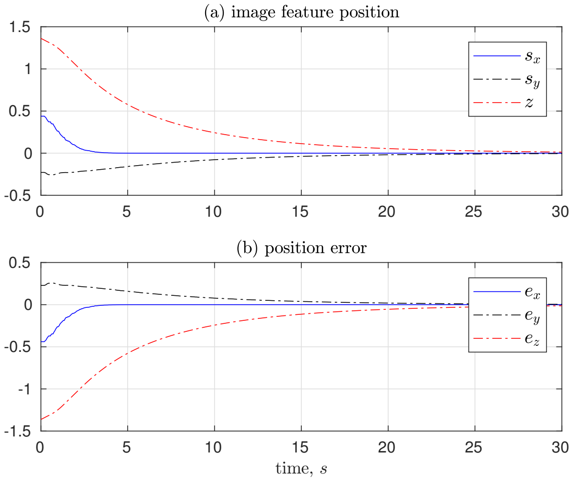

Fig. 5: Regulation task: (a) image feature position, (b) position error.

VII. CONCLUDING REMARKS

In this work, we have developed a hybrid visual servoing control combining the benefits of PBVS and IBVS visual servoing approaches. We have shown that the HVS approach has robustness properties to cope with the ill-conditioning problem of the Jacobian and interaction matrices, and also to deal with parametric uncertainties for regulation tasks. By using a self-updating tracking window approach it is possible to guarantee a satisfactory performance of the proposed solution during the task execution, even in the presence of many targets. Numerical simulations and preliminary practical results have shown the efficiency and feasibility of using robot arms to perform semi-autonomous harvesting tasks for soft fruits in orchards and poly-tunnels. Some proposed topics for further investigation involve: (i) improve the DCNN encoder-decoder algorithm to improve stem detection

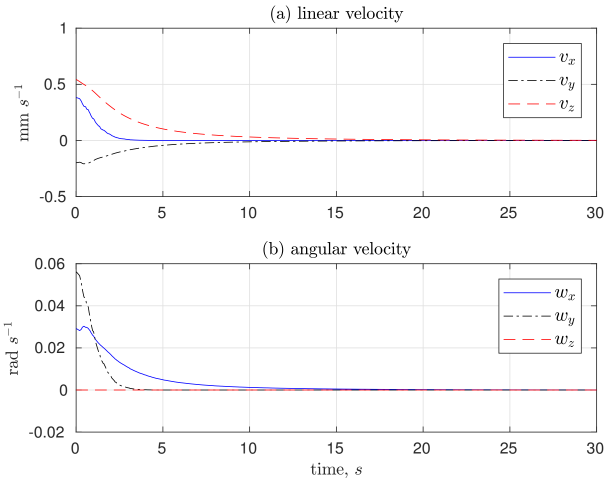

Fig. 6: HVS control signal: (a) linear velocity; (b) angular velocity.

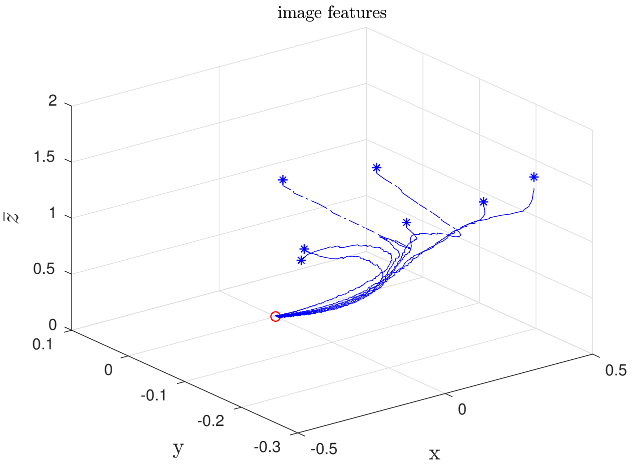

Fig. 7: Trajectory of the image features: initial position "$*$" and final position "$\circ$".

and segmentation to perform fruit harvesting under different stems configurations and rotations. (ii) examine other DCNN encoder-decoder algorithms for fruits recognition and image segmentation based on atrous convolution schemes, in order to avoid the loss of information at downsampling phases, also test instance segmentation to deal with recognition problems presented in occlusion and clustering situations;

References

[1] Y. Edan, S. Han, and N. Kondo, "Automation in Agriculture," in *Springer Handbook of Automation*, S. Y. Nof, Ed. Springer-Verlag Berlin Heidelberg, 2009, pp. 1095–1128.

[2] C. W. Bac, E. J. van Henten, J. Hemming, and Y. Edan, "Harvesting Robots for High-value Crops: State-of-the-art Review and Challenges Ahead," *Journal of Field Robotics*, vol. 31, no. 6, pp. 888–911, 2014.

[3] D. Eizicovits, B. van Tuijl, S. Berman, and Y. Edan, "Integration of Perception Capabilities in Gripper Design using Graspability Maps," *Biosystems Engineering*, vol. 146, pp. 98–113, 2016, special Issue: Advances in Robotic Agriculture for Crops.

[4] A. Gongal, S. Amatya, M. Karkee, Q. Zhang, and K. Lewis, "Sensors and Systems for Fruit Detection and Localization: A Review," *Computers and Electronics in Agriculture*, vol. 116, pp. 8–19, 2015.

[5] S. S. Mehta and T. F. Burks, "Vision-based Control of Robotic Manipulator for Citrus Harvesting," *Computers and Electronics in Agriculture*, vol. 102, pp. 146–158, 2014.

[6] R. Barth, J. Hemming, and E. J. van Henten, "Design of an Eye-in-hand Sensing and Servo Control Framework for Harvesting Robotics in Dense Vegetation," *Biosystems Engineering*, vol. 146, pp. 71–84, 2016, special Issue: Advances in Robotic Agriculture for Crops.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[8] C. Hung, J. Nieto, Z. Taylor, J. Underwood, and S. Sukkarieh, "Orchard fruit segmentation using multi-spectral feature learning," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 5314–5320.

[9] P. A. Dias, A. Tabb, and H. Medeiros, "Multispecies fruit flower detection using a refined semantic segmentation network," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3003–3010, Oct 2018.

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[11] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," pp. 2564–2571, Nov 2011.

[12] B. Siciliano, L. Sciavicco, L. Villani, and G. Oriolo, *Robotics: Modelling, Planning and Control*. Springer Publishing Company, Inc., 2009.

[13] A. F. Agarap, "Deep learning using rectified linear units (relu)," *CoRR*, vol. abs/1803.08375, 2018. [Online]. Available: http://arxiv.org/abs/1803.08375

[14] S.-F. Chang, "Segmentation using superpixels: A bipartite graph partitioning approach," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ser. CVPR '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 789–796. [Online]. Available: http://dl.acm.org/citation.cfm?id=2354409.2355103

[15] X. Wei, K. Jia, J. Lan, Y. Li, Y. Zeng, and C. Wang, "Automatic Method of Fruit Object Extraction under Complex Agricultural Background for Vision System of Fruit Picking Robot," *Optik - International Journal for Light and Electron Optics*, vol. 125, no. 19, pp. 5684–5689, 2014.

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[18] F. Chaumette and S. Hutchinson, "Visual Servo Control - Part I: Basic approaches," *IEEE Robotics Automation Magazine*, vol. 13, no. 4, pp. 82–90, Dec 2006.

[19] T. R. Oliveira, A. C. Leite, A. J. Peixoto, and L. Hsu, "overcoming limitations of uncalibrated robotics visual servoing by means of sliding mode control and switching monitoring scheme."

[20] A. C. Leite, F. Lizarralde, and L. Hsu, "Hybrid Adaptive Vision-Force Control for Robot Manipulators Interacting with Unknown Surfaces," *The International Journal of Robotics Research*, vol. 28, no. 7, pp. 911–926, 2009.

[21] A. C. Leite and F. Lizarralde, "Passivity-based Adaptive 3D Visual Servoing without Depth and Image Velocity Measurements for Uncertain Robot Manipulators," *International Journal of Adaptive Control and Signal Processing*, vol. 30, no. 8-10, pp. 1269–1297, 2016.

[22] D. A. Forsyth and J. Ponce, *Computer Vision - A Modern Approach*, 2nd ed. Pearson Inc., 2012.

[23] H. Qian, Y. Mao, J. Geng, and Z. Wang, "Object tracking with self-updating tracking window," pp. 82–93, 2007.