

# BVE + EKF: A Viewpoint Estimator for the Estimation of the Object's Position in the 3D Task Space Using Extended Kalman Filters

Sandro Costa Magalhães<sup>1,2</sup><sup>a</sup>, António Paulo Moreira<sup>1,2</sup><sup>b</sup>, Filipe Neves dos Santos<sup>1</sup><sup>c</sup>  
and Jorge Dias<sup>3,4</sup><sup>d</sup>

<sup>1</sup>INESC TEC, Porto, Portugal

<sup>2</sup>FEUP, Porto, Portugal

<sup>3</sup>ISR, University of Coimbra, Coimbra, Portugal

<sup>4</sup>KUCARS, Khalifa University, Abu Dhabi, U.A.E.

**Keywords:** Viewpoint Selection, 3D Position Estimation, Pose Estimation, Statistics, Kalman Filter, Active Perception, Active Sensing.

**Abstract:** RGB-D sensors face multiple challenges operating under open-field environments because of their sensitivity to external perturbations such as radiation or rain. Multiple works are approaching the challenge of perceiving the three-dimensional (3D) position of objects using monocular cameras. However, most of these works focus mainly on deep learning-based solutions, which are complex, data-driven, and difficult to predict. So, we aim to approach the problem of predicting the three-dimensional (3D) objects' position using a Gaussian viewpoint estimator named best viewpoint estimator (BVE), powered by an extended Kalman filter (EKF). The algorithm proved efficient on the tasks and reached a maximum average Euclidean error of about 32 mm. The experiments were deployed and evaluated in MATLAB using artificial Gaussian noise. Future work aims to implement the system in a robotic system.

## 1 INTRODUCTION

Agriculture is a critical sector that has been facing several difficulties over time. That constraints are well-designed by several organizations, such as the United Nations (UN) in the objectives for sustainable development (ODS) (General Assembly, 2015). However, their solution is still a challenge.


The increased food demand promoted by the population growth (FAO, 2023) and labor shortage require improved and efficient agricultural technologies that may speed up the execution of farming tasks. Monitoring and harvesting are some tasks that may benefit from robotization in the area.


Autonomous robots require robust perception systems to detect and identify fruits and other agricultural objects. Several works use RGB-D<sup>1</sup> cameras to see the objects and infer their three-dimensional (3D)


position (Kumar and Mohan, 2022; Magalhães et al., 2022). However, RGB-D sensor can malfunction under open-field environments due to external interferences (Kumar and Mohan, 2022; Gené-Mola et al., 2020; Ringdahl et al., 2019), such as radiation or rain. To overcome this challenge, several works designed solutions to infer the position of the objects from monocular sensors. The most common systems are based on deep learning to infer the six-dimensional (6D) or 3D pose of objects (Li et al., 2023; Parisotto et al., 2023; Chang et al., 2021; Wang et al., 2021) or estimate their depth (Ma et al., 2019; Birkl et al., 2023). Deep learning deploys, although complex, algorithms that are very data-dependent, usually supervised, and hard to predict and modify.

Despite the tendency for deep-learning solutions, we still can use Bayesian algorithms to infer the 3D position of objects. In previous work, (Magalhães et al., 2024) designed the MonoVisual3DFilter to estimate the position of objects using histogram filters. However, the algorithm still requires the manual definition of viewpoints to estimate the position of the fruits.

<sup>a</sup> <https://orcid.org/0000-0002-3095-197X>

<sup>b</sup> <https://orcid.org/0000-0001-8573-3147>

<sup>c</sup> <https://orcid.org/0000-0002-8486-6113>

<sup>d</sup> <https://orcid.org/0000-0002-2725-8867>

<sup>1</sup>Red, green, blue and depth sensor

Therefore, in this work, we challenge to identify a mechanism that can autonomously infer the 3D position of fruits without the demand for manually defining viewpoints.

We approach our question with the challenge of autonomously identifying the position of fruits, such as tomatoes, in the tomato plant for precision monitoring or harvesting tasks. We assume the system has a manipulator with a monocular camera configured in an eye-hand manner.

In the following sections, this article explores the proposed solution. The section 2 details the implementation of the best viewpoint estimator (BVE) powered by the extended Kalman filter (EKF) to estimate the 3D position of the objects in the tasks space. This section also defines some experiments to evaluate the algorithm. The section 3 illustrates the results for the various experiments and some algorithm limitations. Finally, section 4 concludes and summarizes this study and introduces future work.

## 2 MATERIALS AND METHODS

### 2.1 BVE

A statistical optimization approach guides towards a solution for this problem. The observation of a fruit from a viewpoint has an associated observation error. The goal is to identify a subsequent viewpoint that minimizes this error. Thus, the problem is the minimization of a loss function related to the intersection of Gaussians distributions (1), where  $N_i(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$  denotes a Gaussian distribution. The index  $i \in \mathbb{N}$  corresponds to each observation viewpoint.

$$N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) = N_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1) \cdot \dots \cdot N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}_n) \quad (1)$$

The Gaussian distribution's product (1) presents significant computational complexity. Nevertheless, (Petersen and Pedersen, 2012) posits that we can decompose the product of Gaussians into two distinct equations—addressing the mean values and the covariance. Because we expect the fruit to remain stopped while hung on the tree, this solution proposes that the position of the tomato,  $\mathbf{k}$ , remains invariant, thus  $\boldsymbol{\mu}_i = \mathbf{k}$ . Hence, we simplify the optimization problem to (2).

$$\boldsymbol{\Sigma}_p = (\boldsymbol{\Sigma}_1^{-1} + \dots + \boldsymbol{\Sigma}_n^{-1})^{-1} \quad (2)$$

The observation noise covariance is predominantly a characteristic intrinsic to the camera. Consequently, the camera's covariance  $\boldsymbol{\Sigma}_c$  remains constant within the camera's frame,  $C$ . The movement of

the camera to different poses,  $\mathbf{c}$ , changes the observation noise in the fixed world frame  $W$ . So, the model requires an observation covariance matrix expressed within the main frame  $W$  (3) to correlate the multiple observations. The matrix  $\mathbf{R}_C^W$  represents a rotation matrix that delineates the relationship between the camera's frame  $C$  and the main frame  $W$ .

$$\boldsymbol{\Sigma}_n = \mathbf{R}_C^W \boldsymbol{\Sigma}_c \mathbf{R}_C^{W\top} \quad (3)$$

In concluding the initial problem definition, we should recognize that the covariance matrix undergoes modification with each iteration of the algorithm as a consequence of the computations performed in equations (2) and (3). To generalize the system's initial conditions, a generic covariance matrix,  $\boldsymbol{\Sigma}_o$ , is considered. This matrix represents the culmination of all previous intersections of covariance matrices up to the point  $k - 1$ .

#### 2.1.1 Definition of the Rotation Matrix

The observation covariance matrix  $\boldsymbol{\Sigma}_c$  is initially defined into the camera's frame. We can convert data between frames using homogeneous transformations, namely rotation matrices, because the translation is irrelevant. Figure 1 illustrates a possible generic relationship between frames. The camera's frame, denoted as  $O_{x_C y_C z_C}$ , is centered at the sensor, and the  $\mathbf{x}_C$  axis indicates the camera's viewing direction. For simplicity, we assume that  $\mathbf{y}_C$  is always parallel to the plane defined by  ${}_{x_W} O_{y_W}$ . This simplification is possible because the covariance matrix is ideally symmetrical in the  $\mathbf{x}_C$  axis, and the other axis's orientation is irrelevant.

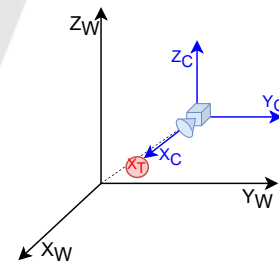


Figure 1: Definition of the camera's frame and the main frame.

Given the fruit position's estimation in the main frame,  $\hat{\mathbf{k}}$ ,  $\mathbf{e}_{x_C^W}$  denotes the unit vector of  $\mathbf{x}_C$  axis (4), where  $\mathbf{c}$  is the camera's position. We can define the camera frame's axes in the main frame through (5), (6), and (7). The rotation matrix  $\mathbf{R}_C^W$  that relates the camera frame to the main frame is obtained from (8). In (8),  $\mathbf{e}_{x_C^W}$ ,  $\mathbf{e}_{y_C^W}$ , and  $\mathbf{e}_{z_C^W}$  are the unit vector of  $\mathbf{x}_C^W$ ,  $\mathbf{y}_C^W$ , and  $\mathbf{z}_C^W$ , respectively.

$$\mathbf{e}_{x_c^W} = \frac{\hat{\mathbf{k}} - \mathbf{c}}{\|\hat{\mathbf{k}} - \mathbf{c}\|} \quad (4)$$

$$\mathbf{x}_C^W = \mathbf{e}_{x_c^W} = [x_1 \quad x_2 \quad x_3]^\top \quad (5)$$

$$\mathbf{y}_C^W = [-x_2 \quad x_1 \quad 0]^\top \quad (6)$$

$$\mathbf{z}_C^W = \mathbf{x}_C^W \times \mathbf{y}_C^W \quad (7)$$

$$\mathbf{R}_C^W = \begin{bmatrix} \mathbf{e}_{x_c^W} & \mathbf{e}_{y_c^W} & \mathbf{e}_{z_c^W} \end{bmatrix} \quad (8)$$

### 2.1.2 Definition of the Objective Function

We aim to minimize a function related to the product of Gaussian distributions (2). This endeavor requires a loss function directly contingent upon the Gaussian intersection. The optimizer predicates a scalar output from the loss function we designed as a dispersion dependency.

For each observation, (9) characterizes the intersection of two covariance matrices in a fixed main frame. The computation of this intersection necessitates the calculation of three inverse matrices, which is a computationally demanding operation.

$$\Sigma_u = (\Sigma_o^{-1} + \Sigma_n^{-1})^{-1} \quad (9)$$

We reduced the number of these operations through the precision matrix ( $\mathbf{P} = \Sigma_u^{-1}$ ). Then, the precision's concentration ( $c = \det(\mathbf{P})$ ) translates the matrix into a scalar. So, we can define the objective function as the dispersion ( $1/c$ ), because, according to the properties of the determinants,  $\det(\mathbf{P}^{-1}) = \det(\mathbf{P})^{-1}$ . Due to the low magnitude of the loss function, we scaled the dispersion into the logarithmic scale (10).  $\mathbf{P}$  and  $\Sigma_n$  are dependent on  $\hat{\mathbf{c}}$ , the next estimated position of the camera, which we aim to optimize.

$$f(\hat{\mathbf{c}}) = -\ln(\det(\Sigma_n^{-1} + \Sigma_o^{-1})) \quad (10)$$

Alternatively to the loss function 10, we can minimize the absolute maximum eigen value of the covariance matrix if we have enough computing power to compute (9). While using this loss function, we should remember that it is highly non-linear and whose derivative function varies at each step because of the maximum function.

$$f(\hat{\mathbf{c}}) = \max(|\text{eig}(\Sigma_u)|) \quad (11)$$

We can use optimization algorithms operating with non-linear restrictions and loss functions to solve the problem using both functions. For the current analysis, we opted to use an interior-point algorithm (Nocedal et al., 2014), already implemented in MATLAB's optimization toolbox(The MathWorks, Inc., 2024).

We also intend to effectively drive the camera to the objects to perform tasks, while estimating the position of the fruit. Towards that, we added an additional component to the loss function (13). The  $\text{act}(i, a, b)$  is a sigmoid activation function (12). The sigmoid activates the additional component, forcing the camera to approximate the object. In the activation function (12),  $a$  and  $b$  are control parameters that set its aggressiveness and its set point (i.e., the value of the function for  $\text{act}(\cdot) = 0.5$ ), respectively;  $i$  is the iteration number of the procedure. Through this strategy, we can activate gradually the Euclidean error to the fruit according to the evolution of the estimation procedure.

$$\text{act}(i, a, b) = \frac{1}{1 + e^{-a \cdot (i - b)}} \quad (12)$$

$$F(\hat{\mathbf{c}}) = f(\hat{\mathbf{c}}) + \text{act}(i, a, b) \cdot \|\hat{\mathbf{k}} - \hat{\mathbf{c}}\| \quad (13)$$

### 2.1.3 Definition of the Restrictions

The proposed algorithm can effectively estimate the best camera poses that maximize the observability of the fruits. However, some restrictions should be implemented to match the environment and hardware constraints. So, we defined that the selected poses must be inside the working space of a 6 DoF<sup>2</sup> manipulator. A spheric model simplifies this workability restriction. Considering a manipulator with a working space centered in  $\mathbf{m}$  and with a radius  $r_m$ , in meters, the camera's pose  $\hat{\mathbf{c}}$  must be inside (14). We only estimate the center position of the fruit but mislead its volume. An additional condition forces the camera to be outside the fruit space. Thus, considering an average fruit radius  $r_k$ , centered in  $\mathbf{k}$ , the camera's pose has to comply with (15).

$$(\hat{\mathbf{c}} - \mathbf{m}) \cdot (\hat{\mathbf{c}} - \mathbf{m})^\top - r_m^2 \leq 0 \quad (14)$$

$$-(\hat{\mathbf{c}} - \hat{\mathbf{k}}) \cdot (\hat{\mathbf{c}} - \hat{\mathbf{k}})^\top + r_k^2 \leq 0 \quad (15)$$

The camera's orientation is also relevant to ensure it looks towards the fruit. The algorithm only focuses in estimating the best position for the camera, but also the orientation of it should be constrained, ensuring the camera is looking towards the fruits. The camera has a conical vision profile. So, we constrained the fruits to be inside the camera's field of view, with a conical shape, (19) and figure 2, where  $HFOV$  is the angle of the horizontal field of view of the camera in radians.

$$\mathbf{e}_c = \frac{\hat{\mathbf{k}} - \hat{\mathbf{c}}}{\|\hat{\mathbf{k}} - \hat{\mathbf{c}}\|} \quad (16)$$

<sup>2</sup>Degrees of freedom

$$\mathbf{e}_{c_{\perp}} = [-e_{c,2} \ e_{c,1} \ e_{c,3}]^T \quad (17)$$

$$\mathbf{x}_{\text{lim}} = \hat{\mathbf{k}} + r_k \cdot \mathbf{e}_{c_{\perp}} \quad (18)$$

$$0 \geq \frac{\mathbf{x}_{\text{lim}} - \hat{\mathbf{c}}}{\|\mathbf{x}_{\text{lim}} - \hat{\mathbf{c}}\|} \cdot \mathbf{e}_c - \cos\left(\frac{HFOV}{2}\right) \quad (19)$$

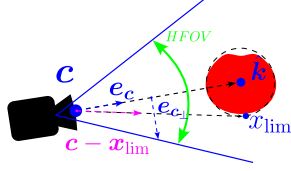


Figure 2: Restriction—fruit inside the camera's field of view.

In a tomato greenhouse, where plants are aligned in rows, the robot must avoid crossing these rows to prevent damage. These rows make a wall of uncrossable tomato plants. Avoiding crossing the designed walls is managed by defining a restriction (22), modelling the plant rows as a planar boundary to keep the robot on one side, set at a distance  $d$  meters from the fruit, as illustrated in the figure 3. The plane's orientation is determined by the normal unit vector  $\mathbf{e}_{n_{\text{plane}}}$ , which represents the normal vector  $\mathbf{n}_{\text{plane}}$ .

$$\mathbf{n}_{\text{plane}} = [\hat{k}_0 \ \hat{k}_1 \ 0]^T \quad (20)$$

$$\mathbf{w} = \hat{\mathbf{k}} - d \cdot \mathbf{e}_{n_{\text{plane}}} \quad (21)$$

$$0 \geq \mathbf{e}_{n_{\text{plane}}} \cdot (\hat{\mathbf{c}} - \mathbf{w}) \quad (22)$$

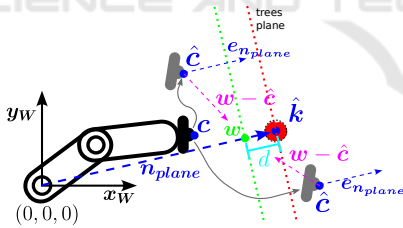


Figure 3: Restriction—camera cannot cross the fruits' trees.

In addition to the previous restrictions, we designed extra ones based on the manipulator's specific features. These ensure that only valid poses are selected, making the kinematics computable, which varies with the manipulator's kinematics.

The previously designed constraints are very specific and complex for the designed target. Conducting essays with simplified algorithms should be advantageous in understanding the benefits and the limitations of a more complete scene design. So, we also conducted experiments with simplified restrictions, considering just one: the distance between the camera and the fruit, denoted as  $l_{\text{dist}}$  in (23).

$$l_{\text{dist}} - \varepsilon < \|\hat{\mathbf{c}} - \hat{\mathbf{k}}\| < l_{\text{dist}} + \varepsilon \quad (23)$$

## 2.2 Fruit Pose Estimation Using the EKF

The BVE computes the best observability viewpoints but does not estimate the 3D position of the objects. Based on an initial rough estimation of the position of the fruit, the EKF can provide iterative refinement of the objects' position.

To ensure a good correct operation of the EKF, the camera should move smoothly while looking at the fruit to correct the fruit position estimation iteratively. So, an additional restriction is implemented to the BVE to ensure that the camera moves to the next best viewpoint in a radius of  $r_d$  meters, (24).

$$\|\hat{\mathbf{c}}_{k+1} - \mathbf{c}_k\| - r_d < 0 \quad (24)$$

The EKF is divided into two main steps: the prediction phase and the correction phase. The fruit position is continuously estimated during prediction, attending dynamics and predictive movement. At the correction phase, the fruit is observed by a dedicated sensor, and so its position is corrected according to the measurements performed.

**Prediction.** During the prediction phase, we estimate the fruit's position, attending to its zero dynamics. The EKF should expect the fruit to not move. So, the predicted position of the fruit is the same as the previous one (25). Besides, the EKF also has to propagate the covariance estimation error (26).

$$\hat{\mathbf{x}}_{k|k-1} = f(\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{u}_{k-1}) = \mathbf{I} \cdot \hat{\mathbf{x}}_{k-1} \quad (25)$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_k \cdot \mathbf{P}_{k-1} \cdot \mathbf{F}_k^T + \mathbf{Q}_k = \mathbf{P}_{k-1} + \mathbf{Q}_k \quad (26)$$

$$\mathbf{F}_k = \left. \frac{\partial f}{\partial \mathbf{x}} \right|_{\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{u}_k} = 1 \quad (27)$$

**Correction.** Assuming that the camera always observes the fruit, the EKF always has a correction phase after a prediction phase. During this phase, the EKF corrects the estimation of the fruit's position (35), acknowledging the measures obtained from the camera to the sensor (28). The EKF uses the same rough initial estimation method based on the fruit's average size and the camera's distortion. The correction phase also corrects the covariance propagation error (33). In these equations,  $\hat{\mathbf{x}}$  is the estimated position of the fruit for each instance, and  $\varepsilon$  is a random noise variable added to create noise for the simulated environment (under real conditions, this value is realistically measured and should be ignored).

$$h(\hat{\mathbf{x}}_{k|k-1}) = \|\hat{\mathbf{x}}_{k-1} - \mathbf{c}\| \quad (28)$$

$$\mathbf{z}_k = \|\mathbf{k} - \mathbf{c}\| + \varepsilon \cdot \sqrt{\sigma_{xx}} \quad (29)$$

$$\mathbf{H}_k = \nabla h(\hat{\mathbf{x}}_{k|k-1}) = \frac{\hat{\mathbf{x}}_{k-1} - \mathbf{c}}{\|\hat{\mathbf{x}}_{k-1} - \mathbf{c}\|} \quad (30)$$

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \cdot \mathbf{H}_k^\top \cdot (\mathbf{H}_k \cdot \mathbf{P}_{k|k-1} \cdot \mathbf{H}_k^\top + R_k)^{-1} \quad (31)$$

$$R_k = \sigma_{xx} \quad (32)$$

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k \cdot \mathbf{H}_k) \cdot \mathbf{P}_{k|k-1} \quad (33)$$

$$\tilde{\mathbf{y}}_k = \mathbf{z}_k - h(\hat{\mathbf{x}}_{k|k-1}) \quad (34)$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \cdot \tilde{\mathbf{y}}_k \quad (35)$$

### 2.3 Experiments

Multiple essays were made under a simulation context in MATLAB to validate the designed algorithms. We deployed an iterative protocol that adds restrictions to the optimizer. That helps to understand the limitations with the increase of the optimization difficulties. Below, we systematize the different experiments and the restrictions considered for each one. Mind that in all cases, the BVE always considers the restriction (24) in  $r_d$  of 0.2 m. The EKF uses a near-realistic covariance matrix for the camera's observations of the fruit, corresponding to a diagonal matrix and a bigger observation variance in the  $xx$  axis.

**E1.** For this experiment, we used the loss function (10) and restricted the BVE's behavior with limited the position of the camera  $l_{dist}$  of  $(1.0 \pm 0.1)$  m to the fruit, (23).

**E2.** In this experiment, we repeated the previous essay, but we also considered the restriction (14) that assures that the camera is inside the manipulator's working space, considering the Robotis Manipulator-H with its center  $\mathbf{m}$  in  $[0 \ 0 \ 0.159]^\top$  m and a maximum range  $r_m$  of 0.645 m.

**E3.** In this experiment, we consider the loss function (10) and the restrictions (14), (15) with the average tomato size  $r_k$ , and (19).

**E4** This experiment considers the restrictions and the loss function of E3 and adds the restriction (22), considering  $d = 0.1$  m;

**E5.** This experiment repeats the previous experiment, adding the kinematics constraints, ensuring that the camera's pose is always a valid pose for the manipulator;

**E6.** Repeats the experiment E1, considering the loss function (11), based on the minimization of the maximum covariance, instead of the dispersion-based loss function (10);

**E7.** Repeats the experiment E2, considering the loss function (11);

**E8.** Repeats the experiment E3, considering the loss function (11);

**E9.** Repeats the experiment E4, considering the loss function (11); and

**E10.** Repeats the experiment E5, considering the loss function (11).

We executed the simulation code for 100 runs for each of these experiments. In each run, we consider a random position for the tomato  $\mathbf{k}$ ,  $k_i \in [-1, 1]$  m, and a random initial position for the camera  $\mathbf{c}$ ,  $c_i \in [-2, 2]$  m. The initial estimation of the fruit was initialized in its real position  $\mathbf{k}$  added by a random bias between  $[-0.15; 0.15]$  m for each axis.

We assessed the algorithm's performance using the mean absolute percentage error (MAPE) (36), mean absolute error (MAE) (37), root mean square error (RMSE) (39), and mean square error (MSE) (38).

$$\text{MAPE}(\mu_j, \hat{\mu}_j) = \frac{1}{N \cdot M} \sum_i \sum_j \left| \frac{\mu_{ij} - \hat{\mu}_{ij}}{\mu_{ij}} \right| \times 100 \quad \forall j \in \mathbb{N} : \{1..M\} \quad (36)$$

$$\text{MAE}(\mu_j, \hat{\mu}_j) = \frac{1}{N \cdot M} \sum_i \sum_j |\mu_{ij} - \hat{\mu}_{ij}| \quad \forall j \in \mathbb{N} : \{1..M\} \quad (37)$$

$$\text{MSE}(\mu_j, \hat{\mu}_j) = \frac{1}{N \cdot M} \sum_i \sum_j (\mu_{ij} - \hat{\mu}_{ij})^2 \quad \forall j \in \mathbb{N} : \{1..M\} \quad (38)$$

$$\text{RMSE}(\mu_j, \hat{\mu}_j) = \sqrt{\frac{1}{N \cdot M} \sum_i \sum_j (\mu_{ij} - \hat{\mu}_{ij})^2} \quad \forall j \in \mathbb{N} : \{1..M\} \quad (39)$$

## 3 RESULTS AND DISCUSSION

The BVE powered with EKF can effectively estimate the fruits' position while using a monocular camera. We organized ten experiments, as described in the section 2.3. Table 1 reports the average errors for the different experiments. Figure 4 illustrates sample paths produced by the optimizer for experiments E2, E5, E7, and E10.

Analyzing the table 1, we verify that simpler and more flexible restrictions result in smaller estimation errors. However, discarding E1, all the experiments conducted in similar estimation errors with an Euclidean error of about 30 mm, if considering the loss

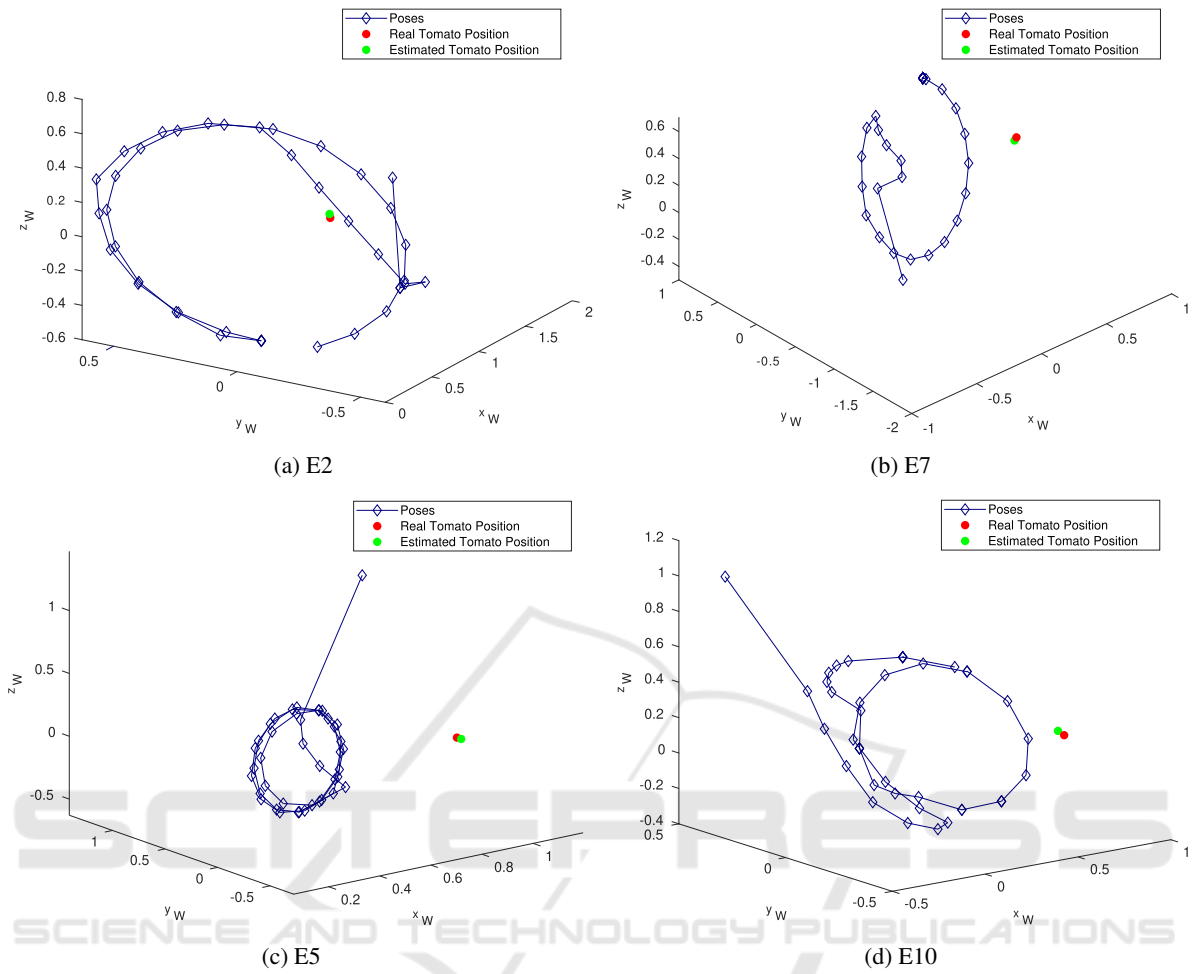


Figure 4: Sample paths generated by the different experiments to assess the fruit’s position. Light blue – poses; red – real fruit position; green – estimated fruit position.

Table 1: Error computations to the centre estimation using the BVE and the EKF.

	MAPE (%)	MAE (mm)	RMSE (mm)	MSE (mm) <sup>2</sup>
<b>E1</b>	5.12	37.1	15.5	241.64
<b>E2</b>	12.86	52.5	25.5	650.47
<b>E3</b>	8.64	53.8	28.1	790.79
<b>E4</b>	11.65	57.2	29.1	848.40
<b>E5</b>	10.62	60.9	31.2	971.44
<b>E6</b>	32.38	53.5	26.3	689.63
<b>E7</b>	14.79	53.1	24.7	612.17
<b>E8</b>	12.62	48.0	21.2	447.57
<b>E9</b>	10.44	53.0	23.4	548.17
<b>E10</b>	20.06	62.3	31.3	982.65

function (10). Despite constraining, the BVE + EKF can perform similarly while increasing the constraining difficulty. Differently, the loss function 11 has a

more progressive behavior, having better results than (10) for less constraining restrictions.

In a general evaluation, we can conclude that using a differentiable loss function (experiments E1 to E5) such as the dispersion (10) brings advantages over a none differentiable loss function (experiments E6 to E10) such as (11), which depends on a maximum operation. Besides, empirical analysis of the performance of both loss functions under the same conditions concludes that the dispersion loss function was also faster to compute because it has one less inversion matrix to calculate. Simpler and less restrictive conditions deliver lower errors while estimating the position of the fruits once the camera has more freedom to navigate around the region of interest. In both strategies, the BVE tends to plan an approximated circular path in the case where the algorithm is free to design its path to the most restricted cases (Figures 4). These circular paths do not always happen in the same plan but in various plans, even transversal ones.

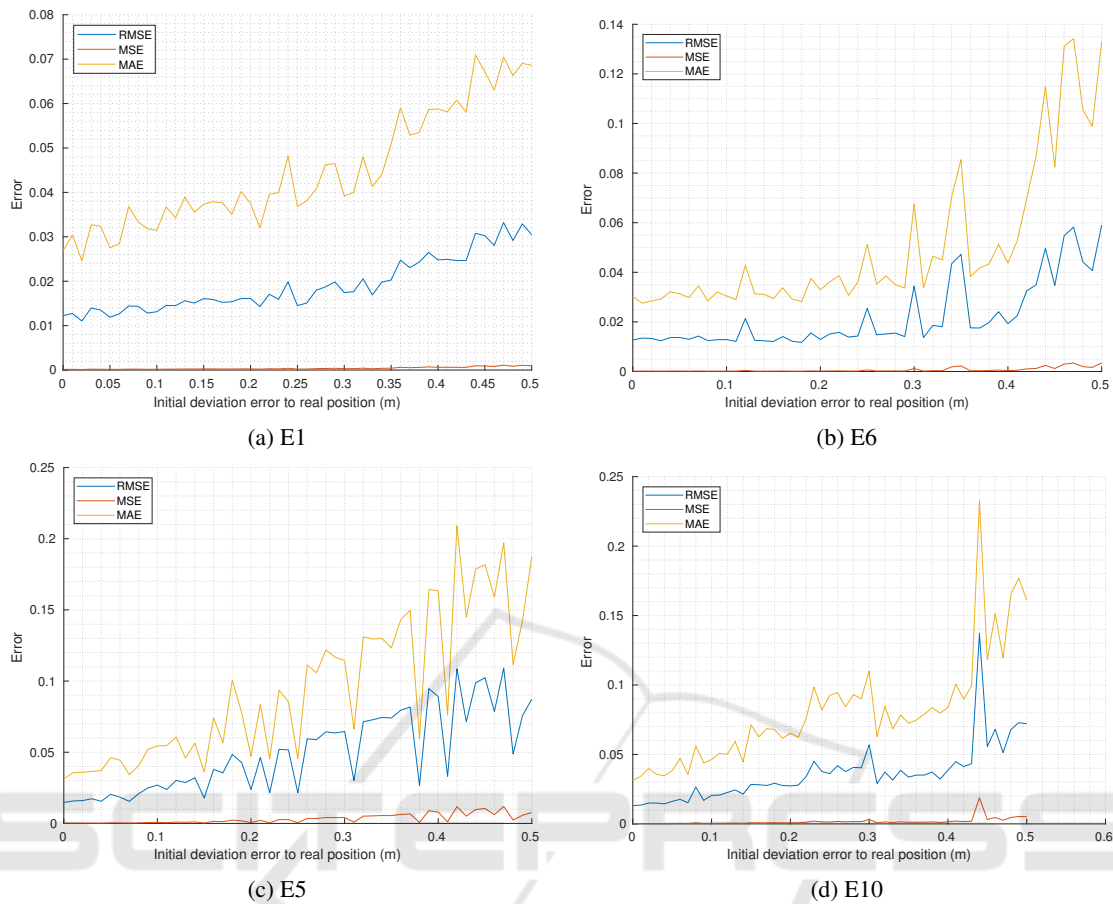


Figure 5: Average error for the recoverability of the loss functions for the BVE + EKF considering different initial estimation errors. Blue – RMSE; red – MSE; yellow – MAE.

For some applications, such as precise and careful fruit picking with cutting tools, the reported estimation errors may not be small enough. To improve the error, the system should apply strategies and algorithms that reduce the covariance. Freer systems also prove to have smaller errors because of the liberty of the algorithm to choose the best observation positions. Besides, implementing historical knowledge should optimize innovation and promote the acquisition of new scene perspectives.

The previous conclusions are enough to understand the performance of both models but do not allow us to understand their limitations and recovery capacity. So, we also performed a recoverability analysis for the loss function to approximate the fruit's position correctly. To achieve this aim, we made multiple essays for estimating the real position of the fruits, considering an initial estimation error between 0 m to 0.5 m in steps of 1 cm. We considered ten simulations for each initial estimation error and computed the average result. Figure 5 illustrates the average er-

rors, given the initial conditions for experiments E1 and E5.

From these experiments, we can conclude that both loss functions perform similarly under the most complex and demanding restrictions. Still, the algorithm can tolerate bigger initial estimation errors by using the dispersion minimization loss function (10). Besides, this loss function is also easier to compute, and the next viewpoint is estimated quickly and easily.

As has been observed, the algorithm is effective in searching for the best viewpoint to estimate the position of the fruits. However, task execution is also relevant to ensure the sensory apparatus moves toward the object. The algorithms can approximate the object in a two-step procedure: positioning the fruit and moving to it. However, using a properly designed loss function such as (13), the BVE + EKF algorithms can iteratively refine the fruit's position while moving towards it. Figure 7 illustrates a possible path to move the sensors from the starting pose to the object, con-

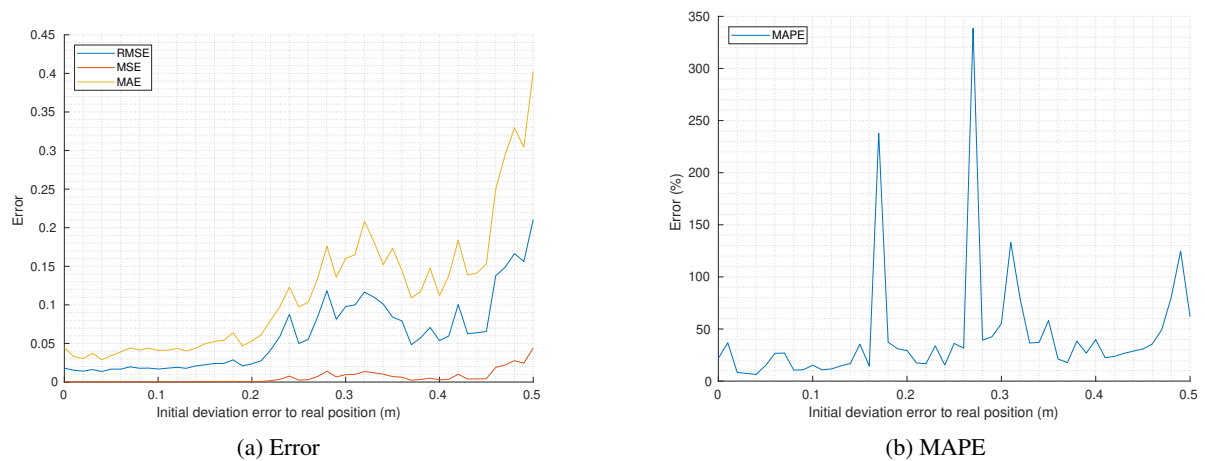


Figure 6: Average Error and MAPE for the recoverability of the loss functions for the BVE + EKF considering different initial estimation errors for the loss function (13) and the restrictions such as in E5. On left (blue – RMSE; red – MSE; yellow – MAE); On Right (blue – MAPE).

sidering the restrictions E5. This scheme shows that the algorithm tends to have a circular path while correcting the fruit’s position.

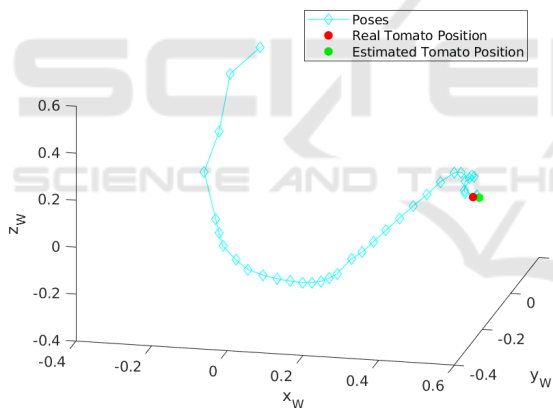


Figure 7: Sensor approximation’s path using the loss function (13) and the restrictions such as in E5. Light blue – poses; red – real fruit position; green – estimated fruit position.

Similar to the previous examples, we also performed a recoverability analysis of the algorithm running the loss function (13). The behavioral results are illustrated in figure 6 plots. The algorithm tends to perform worse and accumulate more errors for this function. Here, we can rely on an initial estimation error until about 15 cm. Initial estimation errors bigger than that will result in a final significant estimation error.

## 4 CONCLUSIONS

The robotization of agricultural fields is an approach that can help to overcome some current societal challenges, such as labor shortages in the field or improved crops. However, that requires the implementation of robust 3D or 6D perception systems independent of depth sensors because of their sensibility to external perturbances.

To approach the problem, we studied a Gaussian-based solution to minimize the observation covariance over the fruits, which we called BVE. We powered the BVE with an EKF that iteratively approximates the position of the fruits. The essay was deployed and tested in mathematical simulation over MATLAB. We designed two loss functions to optimize the resulting observability error: a covariance dispersion-based function and the maximum variance of the covariance matrix. The system reached reasonable results with average Euclidean errors lower than 31.2 mm. A more distinctive analysis concludes that the maximum covariance function is more sensitive to restrictions, so a lower error with fewer constraining restrictions. On the other hand, the dispersion-based function is empirically faster to compute and more robust.

Additional evaluations were conducted to assess the algorithm’s robustness to different initial conditions, which show that both loss functions perform similarly. A variant loss function that drives the sensor to the object proves the robot can perform both tasks simultaneously.

Future work should focus on developing the system in a robotic system under controlled environ-



ments. Besides using EKF, other equivalent algorithms may be tested, such as the Unscented Kalman Filter (UKF). Additional improvements may also be studied, such as implementing historical knowledge that promotes the selection of newer innovative poses.

## ACKNOWLEDGEMENTS

This work is financed by National Funds through the FCT – Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project OmicBots, with reference PTDC/ASP-HOR/1338/2021 (DOI:10.54499/PTDC/ASP-HOR/1338/2021).

Sandro Costa Magalhães is granted by the Portuguese Foundation for Science and Technology (FCT) through the ESF integrated into NORTE2020, under scholarship agreement SFRH/BD/147117/2019 (DOI:10.54499/SFRH/BD/147117/2019).

## REFERENCES

- Birkel, R., Wofk, D., and Müller, M. (2023). MiDaS v3.1 – a model zoo for robust monocular relative depth estimation.
- Chang, J., Kim, M., Kang, S., Han, H., Hong, S., Jang, K., and Kang, S. (2021). GhostPose: Multi-view Pose Estimation of Transparent Objects for Robot Hand Grasping. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.
- FAO (2023). FAOSTAT Statistical Database. 2024-05-21.
- General Assembly (2015). Transforming our world: the 2030 Agenda for sustainable development. Resolution A/RES/70/1.
- Gené-Mola, J., Llorens, J., Rosell-Polo, J. R., Gregorio, E., Arnó, J., Solanelles, F., Martínez-Casasnovas, J. A., and Escolà, A. (2020). Assessing the performance of RGB-D sensors for 3D fruit crop canopy characterization under different operating and lighting conditions. *Sensors-basel.*, 20(24):7072.
- Kumar, M. S. and Mohan, S. (2022). Selective fruit harvesting: Research, trends and developments towards fruit detection and localization – a review. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 237(6):1405–1444.
- Li, F., Vutukur, S. R., Yu, H., Shugurov, I., Busam, B., Yang, S., and Ilic, S. (2023). NeRF-Pose: A First-Reconstruct-Then-Regress Approach for Weakly-supervised 6D Object Pose Estimation. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2115–2125.
- Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., and Fan, X. (2019). Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6850–6859, Seoul, Korea (South). IEEE.
- Magalhães, S. A., Moreira, A. P., dos Santos, F. N., and Dias, J. (2022). Active perception fruit harvesting robots — a systematic review. *Journal of Intelligent & Robotic Systems*, 105(14).
- Magalhães, S. A. C., dos Santos, F. N., Moreira, A. P., & Dias, J. M. M. (2024). MonoVisual3DFilter: 3D tomatoes' localisation with monocular cameras using histogram filters. *Robotica*, 1–20. doi:10.1017/S0263574724000936
- Nocedal, J., Öztoprak, F., and Waltz, R. A. (2014). An interior point method for nonlinear programming with infeasibility detection capabilities. *Optim. Methods Softw.*, 29(4):837–854.
- Parisotto, T., Mukherjee, S., and Kasaei, H. (2023). MORE: simultaneous multi-view 3D object recognition and pose estimation. *Intelligent Service Robotics*, 16(4):497–508.
- Petersen, K. B. and Pedersen, M. S. (2012). *The Matrix Cookbook*. Version 20121115.
- Ringdahl, O., Kurtser, P., and Edan, Y. (2019). Performance of RGB-D camera for different object types in greenhouse conditions. In *2019 European Conference on Mobile Robots (ECMR)*, pages 1–6, Prague, Czech Republic. IEEE.
- The MathWorks, Inc. (2024). MATLAB 9.14.0.2206163 (R2023a).
- Wang, H., Dong, L., Zhou, H., Luo, L., Lin, G., Wu, J., and Tang, Y. (2021). YOLOv3-litchi detection method of densely distributed litchi in large vision scenes. *Math. Probl. Eng.*, 2021:1–11.