New Depth From Focus Filters in Active Monocular Vision Systems for Indoor 3-D Tracking

Tiago Gaspar, *Student Member, IEEE*, and Paulo Oliveira, *Senior Member, IEEE*

***Abstract*— In this paper, new methodologies for the estimation of the depth of a target with unknown dimensions, based on depth from focus strategies, are proposed. The measurements are extracted from images acquired with a single camera, resorting to the minimization of a new functional, deeply rooted on the optical characteristics of the lens system. The analysis and synthesis of two complementary filters and a linear parametrically varying observer are discussed in detail. These estimators use information present on the boundary of the target, which is assumed to be on a plane parallel to the camera sensor, and whose dimensions are considered to remain constant over time. This paper complements a single pan and tilt camera-based indoor positioning and tracking system. To assess the performance of the proposed solutions, a series of indoor experimental tests for a range of operation of up to ten meters, which included tracking and localizing a small unmanned aerial vehicle with unknown dimensions, was carried out. Depth estimates with accuracies on the order of a few centimeters were obtained.**

***Index Terms*— Complementary filtering, depth from focus, linear-parameter-varying (LPV) observers, monocular vision systems, positioning and tracking, sensor fusion.**

I. Introduction

With the development of autonomous robotic vehicles, localization and tracking have become fundamental issues that must be addressed to provide autonomous capabilities to a robot. The availability of reliable estimates for the position of a robot is essential to its navigation and control systems, which justifies the significant effort that has been put into this domain [1]–[6].

Successfully exploited techniques have been reported, such as infrared radiation, ultrasound, radio frequency, and vision [1]. The indoor tracking system addressed in this paper is vision based, since this approach has a growing domain of applicability and leads to interesting results with a very low investment; see the comprehensive survey on monocular 3-D tracking in [7]. This system estimates in real time the position, velocity, and acceleration of a target that evolves along an unknown trajectory in the 3-D world, as well as its angular speed. These estimates are obtained using suboptimal stochastic multiple-model adaptive estimation (MMAE) techniques that exploit information provided by a single camera [8].

In monocular configurations, the problem of estimating the position of a target strongly depends on the accuracy of the algorithms used to estimate the target depth, i.e., its distance with respect to the camera. These algorithms play a key role in many areas, such as 3-D reconstruction [9], obstacle detection [10], and video surveillance [11]. In multicamera approaches, triangulation strategies are the most effective solution to estimate the position of a target and therefore its depth. However, for systems with small baselines, i.e., with small depth sensitivities, or for systems with only one camera, triangulation methods are not the best option [12]. In these situations, monocular depth estimation strategies should be considered. Moreover, multicamera systems have two significant disadvantages: the image-to-image matching problem, perhaps the major source of errors in this type of strategies, and the missing part problem (it is not possible to estimate the depth of points that are visible only in images acquired by one static camera).

When a single camera is used, the depth of a point in the 3-D world can be estimated by exploiting the relation between this quantity and the amount of blur that corrupts the projection of the point into the images. This is done by modeling the influence that some of the camera intrinsic parameters have on images acquired with a small depth of field. Based upon this principle, there are three main strategies that have been exploited: depth from blur by focusing [13], [14], depth from blur by zooming [15], and depth from blur by irising [16]. In this paper, we are mainly concerned with the depth estimation from blur by focusing. Two different techniques based upon this approach can be found in the literature: depth from defocus [14], [16], [17], and depth from focus [13], [18]–[20]. The depth estimation strategy that is proposed here is based on this latter method, since this approach does not require a mathematical model for the blurring process of the camera, i.e., the point spread function (PSF) responsible for the blurring does not need to be modeled.

This paper is an evolution of a framework recently proposed for target tracking and positioning [8], where a low-cost single pan and tilt camera-based indoor positioning and tracking system was presented. The depth estimation strategy used in that system was a simple visual looming technique that required the knowledge of the dimensions of the target.

Manuscript received July 11, 2014; revised November 26, 2014; accepted December 27, 2014. Date of publication February 10, 2015; date of current version August 7, 2015. Manuscript received in final form January 2, 2015. This work was supported in part by the Fundação para a Ciência e a Tecnologia (FCT) through Mechanical Engineering Institute/Associated Laboratory for Energy, Transports and Aeronautics under Grant PEst-OE/EME/LA0022 and Grant PTDC/EEA-CRO/111197/2009 and in part by the Institute for Systems and Robotics. The work of T. Gaspar was supported by the Ph.D. Student Scholarship under Grant SFRH/BD/46860/2008 through FCT. Recommended by Associate Editor G. Antonelli.

The authors are with the Instituto Superior Técnico, Universidade de Lisboa, Lisbon 1049-001, Portugal (e-mail: tgaspar@isr.ist.utl.pt; p.oliveira@dem.ist.utl.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCST.2015.2388956

In this paper, indoor positioning and tracking systems play a minor role. The emphasis is put on the depth estimation strategy. Two novel approaches to estimate the depth of a target [21], [22] are described: one based on a complementary filter and the other based on a linear-parameter-varying (LPV) observer. These strategies combine the measurements of the depth of the target, obtained with algorithms that exploit the concept of depth from focus, with the measurements of the dimensions of the images of the target. In the complementary filter approach, the measurements of the dimensions of the projection of the target into acquired images are used to infer the value of the target depth derivative over time, i.e., the velocity of the target along the camera principal axis. This quantity is corrupted by a bias, which is the result of assuming an incorrect value for the target unknown dimensions. However, since the proposed complementary filter estimates this bias, it provides estimates for the instantaneous depth of a target describing arbitrary trajectories in the 3-D world, without requiring the availability of further information about its dimensions and shape. In the observer-based approach, the dynamics of the depth of the target is written as a function of a parameter that depends on the dimensions of the projection of the target into acquired images, which leads to an LPV observer for the depth of targets with unknown dimensions. The use of the proposed depth estimation strategies leads to a new monocular indoor positioning and tracking system, which estimates in real time the 3-D position, linear velocity, linear acceleration, and angular speed of targets with unknown dimensions.

The main contributions in this paper are as follows:

1) the complete process of synthesis, analysis, implementation, and validation in real time of two new depth estimators for target tracking and positioning;
2) a detailed study of the influence that some of the camera parameters, such as the focal length and aperture, have on the proposed algorithms;
3) experimental results comparing the performance of the proposed methods with other monocular and stereo-based depth-estimation strategies;
4) experimental results illustrating the behavior of the whole positioning and tracking system when the target is a small indoor unmanned aerial vehicle (UAV).

This paper is organized as follows. A brief overview of the positioning and tracking system architecture is presented in Section II, and in Section III, the process of obtaining the image-based measurements used by the depth estimation algorithms is addressed. The proposed complementary filter and the LPV observer are studied in Sections IV and V, respectively, where their design and analysis are detailed. In Section VI, experimental results illustrating the performance of the proposed depth estimation algorithms and the behavior of the whole positioning and tracking system are presented, and in Section VII, the concluding remarks are provided.

II. System Architecture

The indoor positioning and tracking system proposed in [8] has three main modules: 1) that addresses the interface with the camera; 2) that implements image processing algorithms; and 3) that is responsible for dynamic system state estimations. The architecture of this system is presented in Fig. 1, where some quantities are introduced informally to augment the legibility of the document.

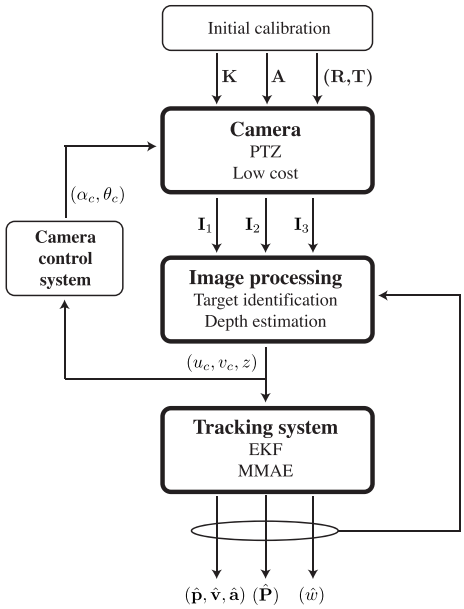


Fig. 1. Tracking system architecture.

The extraction of physical information from an image acquired by a camera requires the knowledge of its intrinsic and extrinsic parameters, which are computed during the initial calibration process. In this paper, the classical direct linear transformation method was used [23]. The camera intrinsic parameters are denoted by the matrix $\mathbf{A}$ and its extrinsic parameters are denoted by $\mathbf{R}$ and $\mathbf{T}$, where $\mathbf{R}$ is a rotation matrix and $\mathbf{T}$ a translation vector. These parameters depend, respectively, on the orientation and position of the camera reference frame $\{\mathcal{C}\}$ with respect to an inertial reference frame $\{\mathcal{I}\}$. This calibration was preceded by an independent determination of a set of parameters $\mathbf{K}$ that compensates for the distortion introduced by the lens of the camera. With this purpose, the lens calibration method proposed in [8] was used. The pan, tilt, and zoom (PTZ) low-cost camera is responsible for providing three images ($\mathbf{I}_1$, $\mathbf{I}_2$, and $\mathbf{I}_3$) to the image processing module, in each iteration of the algorithm. The reason why three images are used per iteration will become clear in the next section.

The image processing module has two main purposes: to identify the target in each image and to estimate its distance in relation to the camera. The target is identified using active contours, which are also known as snakes [24], [25]. This approach consists in finding the target contour and use it to compute the target center coordinates (u_c, v_c). The measurements of the target depth z are provided by depth from focus strategies, as explained in Section III, and processed either by the complementary filter proposed in Section IV or by the observer proposed in Section V. These three quantities (u_c, v_c, and z) correspond to the measurements that are used to obtain the estimates $\hat{\mathbf{p}}$, $\hat{\mathbf{v}}$, and $\hat{\mathbf{a}}$, of the target position $\mathbf{p}$, velocity $\mathbf{v}$, and acceleration $\mathbf{a}$, in the inertial reference frame $\{\mathcal{I}\}$.

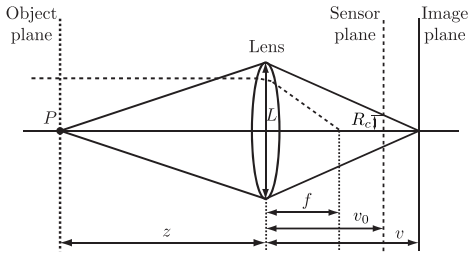


Fig. 2. Model for the imaging process of a thin lens.

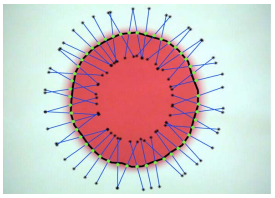


Fig. 3. Example of the boundary, in black, and lines approximately orthogonal to it, in blue, used to estimate the depth of a simple target (in this case a red circle in a white background).

To obtain estimates for the state and parameters of the dynamical model of the target, an estimation problem is formulated and solved. The target dynamical model is linear on the system state (**p**, **v**, and **a**), but nonlinear on the target angular speed w. On the contrary, the sensor model is nonlinear on the system state. Therefore, a set of Extended Kalman Filters (EKFs), each one associated with a different angular speed value, were included in a MMAE architecture. These filters provide estimates for the system state for the state error covariance $\hat{\mathbf{P}}$ and for the target unknown angular speed $\hat{w}$ (Fig. 1).

The command of the camera is the result of solving a decision problem with the purpose of maintaining the target close to the image center. The implemented decision system consists in computing the pan and tilt angles (α_c and θ_c, respectively) that should be sent to the camera at each moment. Large distances between the target and the center of acquired images are avoided. Thus, the capability of the overall system to follow targets is increased.

III. Depth Measurements

In this section, the process of obtaining the measurements used by the depth estimation algorithms proposed in the remaining of this paper is described. The measurements of the depth of the target are obtained resorting to depth from focus strategies, and the target depth derivative is inferred from the variations in the boundary of the images of the target.

A. Target Depth

The idea of inferring depth from focus is based on the concept of depth of field, which is a consequence of the inability of one lens to simultaneously focus planes on the scene at different depths. Depth of field corresponds to the distance between the farthest and the nearest planes on the scene whose points appear in acquired images with a satisfactory definition, according to a certain criterion.

Considering a thin model for the lens of the camera [26], it is possible to establish a nonlinear relation between the distance z from the lens to the plane that the camera can exactly focus at each instant of time (the object plane) and the distance v between the lens and the image plane at which the projection of points in the object plane appears sharply focused (Fig. 2). To complete the relation, the focal length f of the lens is considered. This relation is known as the Gaussian lens formula [26] and can be written as

$$z = \frac{fv}{v - f}. \tag{1}$$

The use of (1) to estimate the depth of a target moving in the scene requires the knowledge of both the focal length of the camera and the value of v, i.e., the value of the camera focus that minimizes the amount of blur that corrupts the projection of the target into acquired images. The estimation of this quantity requires the definition of a metric that quantifies the sharpness of a transition in an image. Metrics related with high-frequency energy contents in the image, the Fourier transform, image gradient, or the Laplacian, are detailed in [18]. Our goal is to estimate the depth of a target, and therefore, the proposed metric aims to maximize the image gradient magnitude across lines orthogonal to the target boundary, which, as described in [8], is obtained using snakes [24], [25]. This approach considers that the real target boundary is on a plane perpendicular to the camera principal axis, which is the plane that appears sharply focused when the camera focus value v_0 [i.e., the distance between the lens and the plane of the camera charge-coupled device (CCD) sensor] is the one that optimizes the proposed metric. This assumption is not too restrictive, since for typical applications, the difference between the depths of the points in the target contour is usually small compared with the accuracy of the depth estimation algorithm. The plane in which the target boundary is considered to be (the object plane) is the plane that specifies the depth of the target. The problem at hand can be written as

$$v = \arg\min_{v_0} \; g(v_0)$$

where the cost function

$$g(v_0) = \left[\frac{1}{N_l} \sum_{i=1}^{N_l} \max_{(x,y)\in l_i} ||\nabla I_{v_0}(x,y)||^2 \right]^{-1} \tag{2}$$

is the inverse of the mean of the square of the image gradient magnitude maximum values across lines approximately orthogonal to the target boundary (Fig. 3). Moreover, N_l denotes the number of used lines, l_i the ith line, ∇ the gradient operator, $||\cdot||$ the Euclidean norm, and $I_{v_0}(x,y)$ the intensity of the image acquired with the focus value v_0 at point (x,y). The formulation of this problem as the minimization of $g(v_0)$, instead of the maximization of its inverse, is based on the model that will be proposed for this function.

To gain some insight into how to model the cost function, consider a scene consisting of a plane at a given depth. In this case, images acquired with a focus value v_0 can be obtained by convolving the ideal sharply focused image $I^f(x,y)$ of the

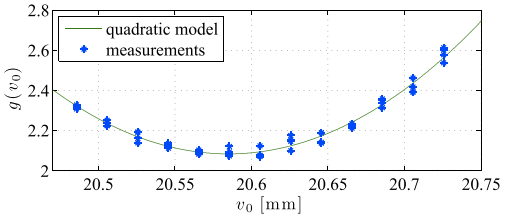


Fig. 4. Cost function for an AXIS 215 PTZ, when the camera focal length is 29 mm and the target is 3 m away from the lens.

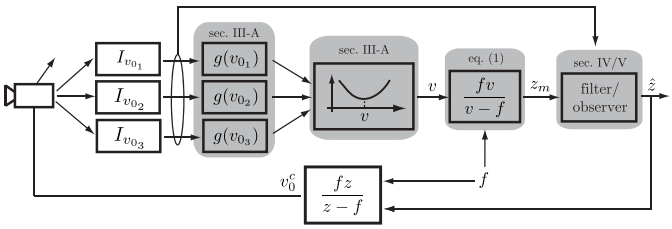


Fig. 5. Architecture of the proposed depth estimation algorithm.

plane with the PSF $h(x, y)$ of the lens system for the depth of the plane, i.e., with the function that models the camera blurring process for the plane depth, $I_{v_0}(x, y) = I^f(x, y) * h(x, y)$.

A common model for the PSF is a circle of constant intensity. Let, in this situation, the PSF be

$$h(x, y) = \begin{cases} \dfrac{1}{\pi R_c^2}, & x^2 + y^2 \leq R_c^2 \\ 0, & x^2 + y^2 > R_c^2 \end{cases}$$

where R_c is the radius of the circle and consider the existence of a vertical step in the sharply focused image of the form $I^f_{v_0}(x, y) = a_1 + a_2\, u(x - x_0)$, where $u(x - x_0)$ is the standard *unit step function* centered at point x_0, a_1 is the intensity of the image when $x < x_0$, and a_2 is the magnitude of the step.

In this situation, it is easy to show that the partial derivative of $I_{v_0}(x, y)$ with respect to y is 0, since differentiation and convolution are linear operations, and thus they commute. After some mathematical manipulation, it is also possible to show that the partial derivative of $I_{v_0}(x, y)$ with respect to x is

$$\begin{cases} 0, & |x - x_0| > R_c \\ \dfrac{2a_2}{\pi R_c^2}\sqrt{R_c^2 - (x - x_0)^2}, & |x - x_0| \leq R_c. \end{cases}$$

By considering a line l orthogonal to the boundary of the target, we can conclude that

$$\max_{(x,y)\in l} ||\nabla I_{v_0}(x, y)||^2 = ||\nabla I_{v_0}(x, y)||^2\Big|_{x=x_0} = \left(\frac{2a_2}{\pi R_c}\right)^2.$$

If some trigonometric manipulations are used, the value of R_c can be written as a function of f, z, and v_0, and the diameter of the lens L; see [16] for details. The replacement of the value of R_c in $(2a_2/\pi R_c)^2$ by its expression allows us to write the cost function proposed in (2) in the form

$$g(v_0) = \frac{(f - z)^2 v_0^2 + 2fz(f - z)v_0 + (fz)^2}{[4fza_2/(L\pi)]^2}. \tag{3}$$

According to the discussion above, which is supported by Fig. 4, the cost function in (2) is expected to depend quadratically on v_0. Thus, a quadratic model is considered for this function. Since three coefficients are enough to define a quadratic function, the acquisition of at least three images with different focus values provides at least three measurements of $g(v_0)$, one per focus value, which are enough to estimate the three coefficients of the cost function model. If three or more images are acquired, a system of linear equations results, which can be solved using the standard linear least squares method [27]. The linear dependence of this model on the parameters to be estimated is the reason why the minimization of $g(v_0)$ was considered, instead of the maximization of its inverse, which seemed to be more intuitive. The estimated coefficients can be easily converted into the estimates of $v = \arg\min_{v_0} g(v_0)$, i.e., the estimates of the focus value that minimizes the cost function for a given depth of the target, since this value corresponds to the one that minimizes the quadratic function. By repeating this procedure over time, the successive estimates for the value of v result, and consequently, the estimates for the instantaneous depth z of the target can be computed using (1).

To study the sensitivity of the estimates of the depth z of the target with respect to the estimates of the value of v, let us compute the partial derivative $(\partial z/\partial v)$ of z with respect to v, which is given by

$$\frac{\partial z}{\partial v} = -\left(\frac{z}{f} - 1\right)^2$$

according to (1). If δz and δv are used to denote small perturbations in the value of z and v, respectively, around some fixed points, we have that

$$\delta z \approx \frac{\partial z}{\partial v}\delta v \approx -\left(\frac{z}{f} - 1\right)^2 \delta v$$

where the symbol $\approx$ is used to indicate that the two members of the equation are approximately equal. As can be seen, the uncertainty in the estimation of z grows quadratically with the depth of the target, if the accuracy in the estimation of v is assumed to be the same for all possible depths.

A simplified version of the architecture of the proposed depth estimation strategy is shown in Fig. 5. In this figure, z_m denotes the measurements of the target depth provided by the algorithm described in this section, which can be written in the form $z_m = z + z_d$, where z_d is the noise that corrupts the measurements of z. Moreover, $\hat{z}$ denotes the target depth estimates obtained from one of the algorithms that will be proposed in Sections IV and V, and $I_{v_{0_i}}$ and $g(v_{0_i})$, $i = 1, 2, 3$, denote, respectively, the three images used by the described depth estimation algorithm and the cost function measurements extracted from these images. It is assumed that the depth of the target does not change during the acquisition of these three images, which is not true in many situations. However, the impact of this assumption in the performance of the proposed algorithms should not be significant when compared with the accuracy of the estimates of the target depth, as long as the target does not perform aggressive maneuvers in the direction of the camera principal axis. The value of v_0^c

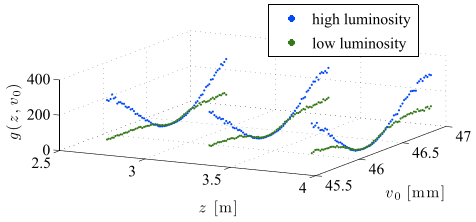


Fig. 6. Influence of the scene illumination on the cost function for several target depths when the image intensity is scaled (the results obtained with an AXIS 215 PTZ and $f = 45.6$ mm).

corresponds to the focus value used to command the focus of the camera.

The accuracy of the target depth estimates obtained with the strategy proposed in this section depends on the depth of field D_{of} of the optical system, which can be written as

$$D_{of} = \frac{4z(z-f)fLR_c}{(fL)^2 - 4(zR_c)^2} \tag{4}$$

where

$$(fL)^2 - 4(zR_c)^2 > 0 \qquad z - f > 0.$$

These expressions can be derived from (1), if some simple trigonometric relations that can be inferred from Fig. 2 are used. As can be seen, larger target depths lead to larger depths of field, as they decrease and increase, respectively, the values of the denominator and numerator in (4). The influence that the camera focal length and aperture have on the depth of field of the an optical system is on the opposite direction, i.e., large focal lengths and large apertures lead to small depths of field [28]. The smaller the depth of field, the more accurate are the depth estimates. Thus, estimates of depths of targets that are close to the lens, obtained using large focal lengths and large apertures, tend to be more accurate.

The accuracy of the depth estimates also depends on the shape of the cost function. It is difficult to compute the minimum of a flat cost function, for instance. Thus, it is important to understand the influence that some quantities have on this function, namely, the scene illumination, and the camera focal length and aperture.

The illumination of the scene influences the shape of the cost function in (3) through the value of a_2. This influence can be minimized by scaling the intensity of acquired images along the lines l in such a way that the magnitude a_2 of the step functions is unitary. As shown in Fig. 6, this strategy minimizes the impact of the scene illumination in the shape of the cost function, especially in the vicinity of its minimum, which is the region of interest. This is not true outside this region, since the noise corrupting the derivative of the image intensity is not negligible when the target boundary is too blurred and the image intensity values are small, which occurs when the scene is poorly illuminated.

To study the influence of the camera focal length and aperture on the shape of the cost function, let the quadratic function in (3) be written in the form $g(v_0) = av_0^2 + bv_0 + c$. Flat cost functions are associated with small values of a, and large values of a lead to cost functions with narrow concavities, whose minimum is easier to compute. Therefore, the value

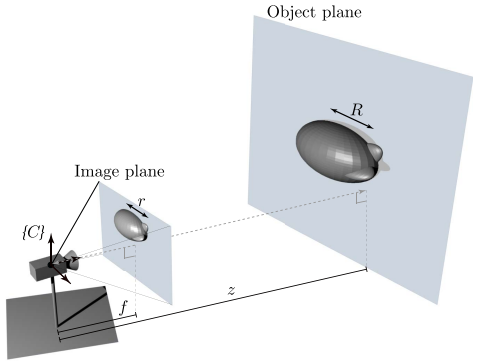


Fig. 7. Relation between the dimensions of the target measured on the object plane and on the image plane, R and r, respectively, for a pinhole camera.

of a defines how difficult is to find the minimum of the cost function. The partial derivatives of a with respect to the camera focal length and aperture are given by

$$\frac{\partial a}{\partial f} = 2z\left(\frac{L\pi}{4zfa_2}\right)^2\left(1 - \frac{z}{f}\right)$$
$$\frac{\partial a}{\partial L} = 2L\left(\frac{\pi}{4za_2}\right)^2\left(1 - \frac{z}{f}\right)^2.$$

Since it is difficult to imagine a situation in which the depth of the target is less than or equal to the focal length, which is on the order of some millimeters, we can assume that $z > f$ and, consequently, $(\partial a/\partial f) < 0$ (even if such situation occurred, it would not have any relevance for the positioning problem that is addressed in this paper). This indicates that small focal lengths lead to cost functions with narrow concavities. However, as previously stated, small focal lengths are associated with large depths of field. Thus, there is a tradeoff between the depth of field and the precision in the computation of the cost function minimum that must be considered when choosing the focal length. The balance between these two aspects depends on the particular optical system in use.

Regarding the camera aperture, large values improve the accuracy of the depth estimates as they lead to cost functions with narrow concavities, $(\partial a/\partial L) > 0$, and to small depths of field.

Note that the measurements obtained according to the proposed strategies are robust to variations in the camera focal length and aperture, which may change the shape of the cost function, as long as they do not occur during the acquisition of a set of images used to obtain a depth estimate. This flexibility results from the computation of new parabola coefficients in each iteration of the algorithm, which leads to the adaptation of the cost function to the new parameters.

B. Target Depth Derivative

By considering a pinhole model for the camera [23] (Fig. 7), the relation between the cartesian coordinates (x, y, z) of a point in the camera reference frame $\{C\}$ and the coordinates (x_p, y_p) of its projection into the image plane is given by $x_p = fx/z$ and $y_p = fy/z$, where the origin of the camera reference frame is considered to be coincident with the camera

optical center, and the origin of the image frame is in the principal point.

From the expressions of x_p and y_p, it is straightforward to show that the distance R between two points in a plane at a distance z from the camera and the distance r between the projection of these points into the image plane are related by

$$r = fR/z. \tag{5}$$

Consider that the coordinates of the points in the boundary of the image of the target are discrete random variables, and let the square root of the trace of the covariance matrix associated with such variables be used as a measure of the target dimensions. In this case, the dimensions of the target verify (5) and are invariant to rotations of the target image boundary. More details about this topic can be found in [21].

According to relation (5) and assuming that f and R remain constant, it is possible to write the derivative of the depth of the target with respect to time in the form

$$\dot{z} = -\frac{\dot{r}}{r^2}Rf \tag{6}$$

where r and $\dot{r}$ denote the square roots of the trace of the covariance matrix associated with the boundary of the image of the target and its derivative with respect to time, respectively. Both quantities follow from the boundary of the target in the image, and their measurements are here denoted as r_m and $\dot{r}_m$.

Relation (6) is a function of the value of R, which corresponds to the dimensions of the real target measured on the object plane. When this quantity is not known, i.e., when the dimensions of the target are not available, an extra term γ that considers this uncertainty must be added to the value of R, resulting in

$$\dot{z}' = \underbrace{-\frac{\dot{r}}{r^2}Rf}_{\dot{z}} \underbrace{-\frac{\dot{r}}{r^2}\gamma f}_{\beta}$$

for the target depth derivative. The value of $\dot{z}$ corresponds to the real target velocity in the direction of the camera principal axis, and β corresponds to a bias term that results from considering γ. The new quantity $\dot{z}'$ denotes a biased version of the target depth derivative $\dot{z}$.

The measurements ψ_m of the target depth derivative provided by the previously described method have the form

$$\psi_m = \psi + \beta + \psi_d + \beta_d \tag{7}$$

where ψ denotes the real target depth derivative, ψ_d is the noise that corrupts the measurements of this quantity, and β_d is a disturbance related to the bias value. The noise may come, for instance, from errors in the segmentation of the target.

The implementation of the discrete-time depth estimation algorithms proposed in the Sections IV and V requires the availability of discrete-time versions of the measurements derived in this section. If T is the sampling interval, the values of the target depth z_{m_k}, at time instants kT (where $k = k_0, k_0+1, \ldots$, and k_0 is associated with the initial instant k_0T), are obtained from the depth from focus algorithm, and the values of the target depth derivative ψ_{m_k}, at the same instants, are computed according to $\psi_{m_k} = -fR'\dot{r}_{m_k}/r_{m_k}^2$, with $r_{m_k} = (\mathrm{tr}(\Sigma_{\mathbf{x}_k}))^{1/2}$ and $\dot{r}_{m_k} = (r_{m_k} - r_{m_{k-1}})/T$, where $\Sigma_{\mathbf{x}_k}$ is the covariance matrix associated with the boundary of the projection of the target into the image acquired at instant kT. Ideally, ψ_{m_k} would be obtained using the target real dimensions R measured in the object plane. However, since this quantity is not known, R is replaced by a constant R' chosen by the user. This quantity can be written as $R' = R+\gamma$, where both R and γ are unknown. As explained before, γ corresponds to the difference between the guessed and real target dimensions.

From (5), it is possible to conclude that at a given time instant kT, an estimate $\hat{R}_k$ for the dimensions of the target can be easily obtained according to $\hat{R}_k = r_{m_k}\hat{z}_k/f$, where $\hat{z}_k$ denotes the value of an estimate $\hat{z}$ for the depth of the target at the same time instant. In particular, in experiments where the dimensions of the target do not vary over time, a global estimate $\hat{R}$ for R can be obtained as the mean of the estimates $\hat{R}_k$ computed along the whole experiment.

At this point, several similar quantities have been introduced. To augment the legibility of the document, a list of these quantities, and their meaning, is presented here:

- z target depth, i.e., distance from the lens to the object plane associated with the target;
- $\hat{z}$ general estimate of the target depth z;
- $\hat{z}_k$ estimate of the target depth z at time instant kT;
- r dimensions of the image of the target, i.e., dimensions of the target measured in the image plane;
- R dimensions of the target measured in the object plane;
- $\hat{R}_k$ estimate of R computed using the measurements obtained at the time instant kT;
- $\hat{R}$ estimate of R computed using the measurements obtained along the whole experiment;
- R' parameter used to replace R, which is unknown;
- z' biased version of the target depth that results from replacing the target dimensions R, in (5), with R';
- R_c PSF radius.

IV. Depth Complementary Filter

In this section, a complementary filter that provides estimates for the depth of a moving target is proposed. Initially, for motivation, a simple continuous-time complementary structure for situations where the dimensions of the target are known is presented. Afterward, this structure is modified to address the same problem when the dimensions of the target are not known, and the process of obtaining a discrete-time version of the filter that results is described. A rigorous formulation of the problem addressed in this section is presented next.

Problem Statement 1: Consider a moving target with unknown dimensions and unknown position (x, y, z) in the camera reference frame $\{\mathcal{C}\}$. Suppose that the measurements

$$\begin{cases} z_m = z + z_d \\ \psi_m = \psi + \beta + \psi_d + \beta_d \end{cases}$$

of the target depth and its derivative are obtained from the images acquired with a single camera and that both quantities are corrupted by noise (z_d and ψ_d, respectively) in complementary frequency regions. These quantities are measured

in relation to the camera reference frame. The value of the target depth derivative is affected by a bias term β, which results from the unknown nature of the target dimensions, and which is corrupted by a disturbance β_d. Given these assumptions, design a filter that provides an optimal solution in the minimum mean square error sense for the problem of estimating the instantaneous depth of the moving target.

A. First-Order: Known Target Dimensions

When the target dimensions R on the object plane are known, the measurements of the target depth derivative presented in (7) are not biased, since $\gamma = \beta = 0$. In this case, a filter with gain $k > 0$ and state-space realization

$$\dot{\hat{z}} = \psi_m + k(z_m - \hat{z}) \tag{8}$$

can be used to obtain the estimates $\hat{z}$ of the target depth from the measurements z_m and ψ_m. As shown in [21], such estimates consist of an undistorted copy of the original signal z, corrupted by the measurement noises z_d and ψ_d

$$\hat{z} = z + \mathcal{F}_z z_d + \mathcal{F}_\psi \psi_d \tag{9}$$

where $\mathcal{F}_z$ and $\mathcal{F}_\psi$ are the linear time-invariant operators.

The proposed filter blends the information provided by the depth from focus algorithm at low frequencies with that from the target depth derivative at high frequencies. This combination is appropriate since the depth from focus measurements are more reliable at low frequencies, whereas the target depth derivative measurements may be corrupted by a bias in the same frequency region (as exemplified in the next section), which makes it useful at higher frequencies.

For the system in study, the stochastic underlying process model, here called $\mathcal{M}$, can be written relying on the realization

$$\sum_{\mathcal{M}} := \begin{cases} \dot{z} = \psi_m - \psi_d \\ z_m = z + z_d \end{cases}$$

where ψ_d and z_d play the roles of process and measurement noises, respectively. To design the gain of the filter, consider an $\mathcal{H}_2$ estimation framework, in which the objective is to minimize the asymptotic variance of the system estimation error, when the input of the system is white Gaussian noise; see [29] for details about $\mathcal{H}_2$ filtering. In this case, the goal is to minimize the asymptotic variance of the depth estimation error $z - \hat{z}$ for given values of the covariances of ψ_d and z_d. The optimal solution to this problem has the complementary structure described in (8). In a deterministic setup, where the aim is to shape the filter closed-form transfer function, the filter can be designed using any efficient method, and the analysis of the filter can be performed in the frequency domain using the Bode plots.

B. Second-Order: Unknown Target Dimensions

In most situations, there is no information about the dimensions of the target. Therefore, the value of γ and, as a consequence, the value of β are not known, and the measurements of the target depth derivative presented in (7) are biased. The simple complementary structure described previously does not allow steady-state bias estimation. However, a modified version of its structure, augmented with an extra integrator, will meet this additional constraint. This strategy results in a new complementary filter, described in the remainder of this section, with the realization

$$\sum_{\mathcal{M}} := \begin{cases} \begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \end{bmatrix} = \begin{bmatrix} -k_1 & 1 \\ -k_2 & 0 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} + \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} z_m + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \psi_m \\ \hat{z} = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} \end{cases} \tag{10}$$

where x_1 and x_2 denote the states associated with the depth of the target and with the bias term, respectively, and k_1 and k_2 are the filter gains. Note that as mentioned in the end of Section III-B, the bias term models the uncertainty resulting from having a target with unknown dimensions. It is possible to show that this bias can also be seen as a term that comprises slowly time-varying perturbations in the dimensions of the target. Therefore, the estimation of the extra state variable x_2 mitigates the effect of assuming constant target dimensions.

From (10), it is possible to show that the estimated target depth can be rewritten as in (9), where the transfer functions of $\mathcal{T}_1$ and $\mathcal{T}_2$ take the form

$$T_1(s) = \frac{k_1 s + k_2}{s^2 + k_1 s + k_2} \qquad T_2(s) = \frac{s^2}{s^2 + k_1 s + k_2}$$

and the intensity of the noise term is given by $\mathcal{F}_z z_d + \mathcal{F}_\psi (\psi_d + \beta_d)$. This term depends on the transfer functions $F_z(s) = T_1(s)$ and $F_\psi(s) = s/(s^2 + k_1 s + k_2)$. As before, $T_1(s) + T_2(s) = I$, where $T_1(s)$ and $T_2(s)$ correspond to low- and high-pass filters, respectively. The second-order complementary filter proposed blends the information provided by the depth from focus algorithm at low frequency regions with that of the target depth derivative in the complementary frequency range, leaving the original signal z undistorted. Therefore, low-frequency bias in the disturbances that corrupt ψ_m will be naturally rejected at the output. Note also that the filter rejects high-frequency noise present in z_m.

In this situation, the underlying process model can be written relying on the realization

$$\sum_{\mathcal{M}} := \begin{cases} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \psi_m - \begin{bmatrix} \psi_d \\ \beta_d \end{bmatrix} \\ z_m = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + z_d \end{cases} \tag{11}$$

where ψ_d and β_d correspond to the process noise and z_d to the measurement noise. One of the two measurements available (ψ_m) is used as an input to the differential equation that models the process and the other (z_m) as the observation. So far, no assumption has been made about the noise terms. However, if the process and measurement noises are stationary, white, and Gaussian processes with zero mean, then as stated in Lemma 1, the complementary filter described in this section corresponds to a stationary Kalman filter for the realization presented in (11). Therefore, according to the properties of Kalman filters [27], the proposed complementary filter provides a stable and optimal solution, in the minimum mean square error sense, for the problem of estimating the depth of a target evolving according to the presented underlying

process model. Note that under the aforementioned Gaussian assumptions, the bias term is modeled as a Wiener process.

Lemma 1: Let the process and observation noises in realization (11) correspond to stationary white Gaussian noises with zero mean and spectral densities σ_ψ^2, σ_β^2, and σ_z^2, respectively (i.e., $\psi_d \sim \mathcal{N}(0, \sigma_\psi^2)$, $\beta_d \sim \mathcal{N}(0, \sigma_\beta^2)$, and $z_d \sim \mathcal{N}(0, \sigma_z^2)$), and β denote a low-frequency bias that corrupts the measurements of the target depth derivative. Then the complementary filter in (10) is the stationary Kalman filter for the system (11) if $k_1 = (2\sigma_\beta/\sigma_z + (\sigma_\psi/\sigma_z)^2)^{1/2}$ and $k_2 = \sigma_\beta/\sigma_z$.

Proof: Let the system realization (11) assume the form

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u + \mathbf{L}\eta \\ y = \mathbf{C}\mathbf{x} + \vartheta \end{cases}$$

where $\mathbf{x} = [x_1\, x_2]^T$, $u = \psi_m$, $y = z_m$, $\eta = [\psi_d\ \beta_d]^T$, and $\vartheta = z_d$ and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

Consider that the process noise η and the observation noise ϑ correspond to stationary white Gaussian noises with zero mean and spectral densities $\mathbf{Q} = \text{diag}[\sigma_\psi^2, \sigma_\beta^2]$ and $\mathbf{R} = \sigma_z^2$, respectively. The notation $\text{diag}[\sigma_\psi^2, \sigma_\beta^2]$ represents a diagonal matrix with the elements σ_ψ^2 and σ_β^2 in its diagonal. In this situation, the estimation error covariance matrix $\mathbf{P}$ of the Kalman filter for the system (11) is the solution of the Riccati equation $\dot{\mathbf{P}} = \mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^T + \mathbf{L}\mathbf{Q}\mathbf{L}^T - \mathbf{P}\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C}\mathbf{P}$; see [27] for more details. The stationary Kalman filter is obtained by setting $\dot{\mathbf{P}} = 0$ in this equation. Considering the general expression $\mathbf{P} = \left(\begin{smallmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{smallmatrix}\right)$ for the estimation error covariance matrix, where $p_{12} = p_{21}$, by the properties of symmetry of covariance matrices, the solution of the Riccati equation in steady-state leads to $p_{11} = \sigma_z(2\sigma_\beta\sigma_z + \sigma_\psi{}^2)^{1/2}$, $p_{12} = p_{21} = \sigma_\beta\sigma_z$, and $p_{22} = \sigma_\beta\sigma_z(2\sigma_\beta\sigma_z + \sigma_\psi{}^2)^{1/2}$. The Kalman filter gain that follows from this solution is $\mathbf{K} = \mathbf{P}\mathbf{C}^T\mathbf{R}^{-1} = [(2\sigma_\beta/\sigma_z + (\sigma_\psi/\sigma_z)^2)^{1/2}\ \sigma_\beta/\sigma_z]^T$, which results in:

$$\begin{aligned} \dot{\hat{\mathbf{x}}} &= \mathbf{A}\hat{\mathbf{x}} + \mathbf{B}u + \mathbf{K}(y - \mathbf{C}\hat{\mathbf{x}}) \\ &= \begin{bmatrix} -\sqrt{2\frac{\sigma_\beta}{\sigma_z} + \left(\frac{\sigma_\psi}{\sigma_z}\right)^2} & 1 \\ -\frac{\sigma_\beta}{\sigma_z} & 0 \end{bmatrix} \hat{\mathbf{x}} + \begin{bmatrix} \sqrt{2\frac{\sigma_\beta}{\sigma_z} + \left(\frac{\sigma_\psi}{\sigma_z}\right)^2} \\ \frac{\sigma_\beta}{\sigma_z} \end{bmatrix} z_m \\ &\quad + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \psi_m \end{aligned}$$

for the state estimate provided by the Kalman filter. This equation is equivalent to the one that provides the state estimate according to the complementary filter in (10), when $k_1 = (2\sigma_\beta/\sigma_z + (\sigma_\psi/\sigma_z)^2)^{1/2}$ and $k_2 = \sigma_\beta/\sigma_z$. Therefore, under the stated assumptions, the complementary filter proposed in this section corresponds to a stationary Kalman filter. ∎

In an $\mathcal{H}_2$ setting, the goal is to minimize the state estimation error for given values of the covariances of ψ_d, β_d, and z_d. As mentioned, the optimal solution to this problem has the complementary structure described in (10). Therefore, the covariances of ψ_d, β_d, and z_d can be regarded as design parameters to vary the cutoff frequency of the filter.

In summary, two strategies can be used to obtain the gains of the filter: a deterministic approach in which classical strategies based on pole placement, for instance, are used [30] and a stochastic approach, in which the noise variances are determined and used to compute the filter gains.

The discrete-time equivalent of the proposed second-order complementary filter is obtained by sampling the solution of the state equation in (10) at time instants kT, where T is the sampling interval and $k = k_0, k_0 + 1, \ldots$; the value of k_0 is associated with the initial time instant k_0T. By assuming that the inputs z_m and ψ_m of the filter remain constant between sampling instants, a discrete time-invariant linear state equation for the filter results; see [30] for more details.

The discrete-time complementary filter proposed in this section provides the estimates for the depth of a target, with unknown dimensions, moving in a 3-D scene. This filter is suitable for the tracking system described in Section II, since it is appropriate for implementation in a digital computer and its computational complexity is very small.

V. Depth LPV Observer

In this section, a state-space formulation for the evolution of the target depth is derived and an observer for the state of the LPV system that results is proposed.

To derive a state-space formulation for the evolution of the depth of the target, consider that both R and f in (5) do not vary over time and that the measurements r_m and $\dot{r}_m$ of r and $\dot{r}$, respectively, are exact. Under these assumptions, it is straightforward to show that the derivative of the depth of the target with respect to time can be written in the form

$$\dot{z} = -\frac{\dot{r}_m}{r_m} z. \tag{12}$$

By denoting the quotient $-\dot{r}_m/r_m$ by a parameter α, and by considering that exact measurements z_m of the target depth are available, a deterministic LPV system with the realization

$$\begin{cases} \dot{z} = \alpha z \\ z_m = z \end{cases}$$

results. An observer for the state z of this system can be written in the form [30]

$$\dot{\hat{z}} = \alpha\hat{z} + h(z_m - \hat{z}), \quad \hat{z}(t_0) = \hat{z}_0 \tag{13}$$

where $\hat{z}$ and $\dot{\hat{z}}$ are the target depth estimate and its derivative with respect to time, respectively; h is the observer gain, t_0 is the initial time instant, and $\hat{z}_0$ is the initial estimate for the target depth.

From the considerations above, it is easy to show that the state estimation error $\tilde{z} = z - \hat{z}$ satisfies the linear equation

$$\dot{\tilde{z}} = (\alpha - h)\tilde{z}, \quad \tilde{z}(t_0) = z_0 - \hat{z}_0 \tag{14}$$

where $\dot{\tilde{z}}$ denotes the derivative of the estimation error with respect to time. The unknown parameter α depends on the target velocity and on the target depth, as it can be written in the form $\alpha = \dot{z}/z$. Since all physical systems have limitations in terms of maximum velocity, $\dot{z}$ is bounded. Moreover, it is

physically impossible that the target depth goes below a certain value (the target and the camera cannot coincide), and thus α is upper bounded and the bound depends on the type of motion of the target. According to this reasoning, we can expect that the gain h must be chosen according to the experiment at hand to guarantee the stability of the observer (Lemma 2).

Lemma 2: The linear state equation (14) is uniformly exponentially stable if the gain h of the observer verifies $h \geq \alpha_{\max} + \nu/q$, where $\alpha_{\max}$ is the upper bound of α, and ν and q are finite positive constants.

Proof: The proof of this lemma can be found in [22]. ■

From (8) and (13), it is possible to infer that the structure of the first-order complementary filter and the structure of the proposed LPV observer are very similar. The main difference is in the computation of the target depth derivative. In the complementary filter, it is based on (6), which is a function of the target dimensions R measured on the object plane, and in the LPV observer, it is based on (12), which is a function of the state variable z. The two expressions are equivalent; however, the latter does not require any knowledge about the dimensions of the target. This is an important advantage since the performance of both strategies is similar, as will be seen in the next section.

The discrete-time version of the proposed observer is obtained by sampling the solution of (13) at discrete-time instants and by assuming that z_m, r_m, and $\dot{r}_m$ remain constant between consecutive sampling times; see [22] for details about this discretization. The proposed observer provides estimates for the depth of a target, with unknown dimensions, moving in a 3-D scene. Thus, it is suitable for the proposed tracking system, since it is appropriate for implementation in a digital computer and its computational complexity is very small.

VI. Experimental Results

In this section, the proposed depth estimation algorithms are assessed using experimental data. This assessment consists of the following:

1) a comparison with several monocular depth estimation strategies, based on the information provided by the authors in the corresponding articles;
2) a comparison with a monocular depth estimation method that is state of the art;
3) a comparison with a classical stereo-based approach;
4) an experimental evaluation of the proposed methods when the target describes two common trajectories (in this case a straight line and a circumference);
5) an example of application of the presented depth estimation strategies in a real positioning and tracking system.

The results presented in this section were obtained with the 215 PTZ camera from AXIS. The images with the spatial resolution 704×576 pixels were used. Since image segmentation is a complex domain, which is not the main focus of this paper, targets with easily identifiable colors were considered.

As in most cameras, the value of the distance v_0 between the plane of the CCD sensor of the used camera and the lens of the camera is not accessible to the operator. Instead, a different parameter ranging from 1 to 9999 is available. This parameter is specified by the manufacturer and is usually known as the camera focus setting. The use of the proposed depth estimation algorithms requires the calibration of the relation between these two quantities; see [31] for details about this procedure.

TABLE I
Comparison Between Several Monocular Depth Estimation Methods for Static Scenes and Real Data

Method	RMSE (%)	Depth range [m]	Real time positioning
Pentland et al. [14]	2.5	up to ~ 1	yes
Ens et al. [16]	1.3	0.8 to 0.95	no
Subbarao et al. [35][1]	2.3 to 20	0.6 to 5	yes
Favaro et al. [34][2]	~ 1	3 to 4	no
Krotkov [18]	~ 1	1 to 3	no
Our method	~ 1	3 to 4	yes

The accuracy of the proposed algorithms depends on the extraction of information from the blur present in the images. To guarantee that this blur results only from the distance of the target to the camera and not from the motion of the CCD sensor, we have to wait some time for this sensor to stop completely after each movement. This interval is approximately 0.4 s for the camera in use, and therefore, the acquisition of an image takes approximately 0.4 s. This time is used to perform image processing tasks, such as identifying the target. Since three images are required per iteration of the algorithm, each iteration takes approximately 1.2/1.3 s, and thus the nominal sampling interval T for the application was set to 1.3 s. Obtaining accurate depth estimates with this sampling interval imposes a constraint in the velocity at which the target can change its depth. However, as explained, this limitation is mostly imposed by the slow speed at which the CCD sensor can be moved and not by the proposed algorithms. Note that the camera in use is a regular camera, not modified for this type of tasks. Tracking other types of targets, namely, targets that perform more aggressive maneuvers, could be achieved using a camera that allowed faster movements of the sensor. Moreover, the aforementioned sampling interval is in accordance with typical update rates for indoor 3-D positioning systems. The *Cricket* system, for instance, which is a reference in terms of indoor positioning, works at 1 Hz [32], and the vision-based system presented in [33] works at 0.1 Hz.

The results provided in this section were obtained by discretizing the LPV observer according to the algorithm described in detail in [22]. The same strategy was used to discretize the complementary filters and the aforementioned sampling interval was used.

A. Comparison With Alternative Depth Estimation Strategies

A comparison between several depth from focus and depth from defocus methods is presented in Table I. The root-mean-square error (RMSE), expressed as a percentage of the distance to the camera, is provided, as well as the range of depths

[1]The error increases linearly from 2.3%, at 0.6 m, to approximately 20%, at 5 m.

[2]The results presented in the table for the algorithm proposed in [34] were obtained using the code that Favaro and Soatto made available on the internet, since the accuracy evaluation provided in [34] is for synthetic data. These results were obtained using the experimental setup described in this section.

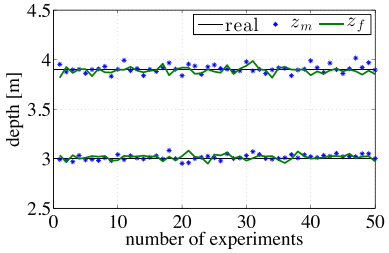


Fig. 8. Comparison between the depth estimation strategies proposed in Section III (RMSE = 1.03%) and [34] (RMSE = 0.94%).

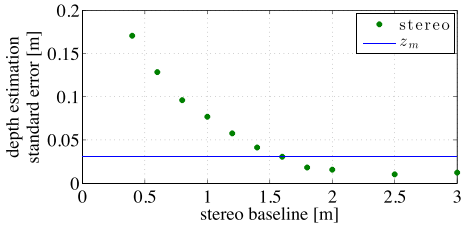


Fig. 9. Comparison between the depth estimation strategy proposed in Section III and a stereo-based strategy.

associated with the reported accuracies. The last column indicates if the implementation of the algorithms in a real-time positioning system is realistic. The first four methods in the table are based on depth from defocus, whereas the last two use a depth from focus approach. Apart from the information that concerns our method and the method proposed in [34], which we obtained by conducting a set of experiments with a static target, all the other information in the table is based on the values indicated in [14], [16], [18], and [35].

The depth estimation strategy detailed in Section III is compared here with the one proposed by Favaro and Soatto [34]. This method learns a set of projection operators from blurred images, which are then applied to novel acquired (blurred) images. Their approach does not use any information about the PSF and consists in minimizing the Euclidean norm of the difference between the estimated and the observed images. Depth is inferred from the operator that leads to the output with the lowest energy. For the results presented in this section, a Gaussian kernel for patches of 3×3 pixels was considered, and 50 equifocal planes placed equidistantly in the range between 2.7 and 4.2 m, in front of the camera, were used. Larger patches could have been considered to obtain better accuracies; however, this would make the algorithm slower. This algorithm provides a depth map of the scene, thus we used a mean of the depth map values in the vicinity of the target contour to estimate its depth.

In Fig. 8, the comparison between the novel strategy proposed in this paper and the aforementioned one is illustrated. A target placed at two different depths, 3 and 3.9 m, was used, and 50 independent experiments were performed with $f = 45.6$ mm. In the figure, z_m denotes the target depth measurements obtained with the strategy presented in Section III-A and z_f the measurements obtained with the strategy proposed in [34]. The values in brackets in the caption correspond to the RMSEs associated with each approach, expressed as a percentage of the target depth. These values were computed by dividing the RMSEs by the target real depth (both experiments were considered). The performance of the algorithm proposed in [34] is slightly superior to the performance of our algorithm. However, despite requiring only two images to estimate the depth of the target (one image less than the strategy that we propose), this algorithm is slower, which is a critical issue in real time applications. In particular, our method is approximately three times faster than the one proposed in [34], for the experiments reported in Fig. 8.

Overall, the depth estimation approach proposed in this document seems to have the best properties for real-time target tracking, when compared with all the monocular strategies presented in Table I. The algorithm proposed in [18] attains similar accuracies for depths on the order of a few meters; however, according to the details provided in [18], this strategy is not appropriate for real-time applications since it is computationally too expensive.

In Fig. 9, a comparison between the depth estimation strategy proposed in Section III and a classical stereo-based system is presented. The cameras were calibrated using the strategy described in [36], and the experimental setup used to obtain the results was similar to the one used to obtain the results in Fig. 8. A target was placed in front of two PTZ cameras with a depth of 3 m, and the cameras were oriented in such a way that the projection of the target was always close to the center of the images. In the monocular configuration the focal length of the camera was 45.6 mm, and in the stereo approach, it was 6.9 mm for both cameras. Most classical stereo methods assume that the stereo system has a nonverged geometry, i.e., that the epipolar lines are parallel to each other; see [37] for details about nonverged and verged geometries. Since this assumption does not hold in many stereo systems and, in particular, does not hold in this case, the image pairs were rectified to a nonverged geometry; see [38] for details on the topic of image rectification. In this paper, the center of the target image boundary is used as the point that determines the target position. The depth z of the target is obtained from the disparity d_P of this point (i.e., from the displacement of this point in the images acquired with one camera with respect to the images acquired with the other camera) according to the expression $z = fB_c/d_P$, where f is the focal length of the cameras and B_c the baseline of the system. Several baselines were tested, and 20 experiments were performed for each baseline. As can be observed in the figure, the standard error of our monocular depth estimation strategy is smaller than the one of the stereo-based approach for baselines smaller than approximately 1.6 m. This result is in accordance with the idea that monocular depth estimation is a good option when the baselines in multicamera systems are small.

Schechner and Kiryati [12] show that depth from focus strategies can be seen as a realization of the geometric triangulation principle, by considering that the diameter of the lens aperture corresponds to the baseline between two cameras in a stereo system. The main difference is in the physical dimensions of the systems, since lens apertures are typically one or two orders of magnitude smaller than stereo baselines,

(a) (b)

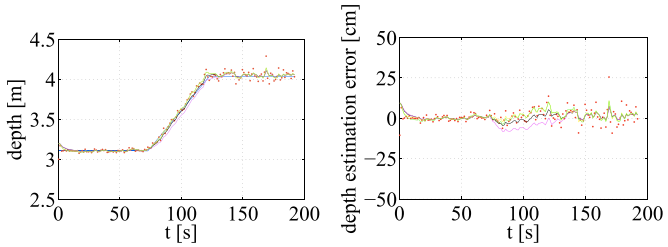
Fig. 10. Real-time target tracking. (a) Experimental setup. (b) Target identification, where the initial snake, its temporal evolution, and the final contour estimate are presented in black, red, and blue, respectively.

which leads to smaller depth sensitivities in the monocular case. However, due to the 2-D nature of lens apertures, the (implied) triangulation in depth from focus does not use only two marginal points, as in stereo, but a continuum of points, which makes the triangulation more robust. This is why the monocular strategy outperforms stereo for small baselines.

B. Performance of the Proposed Algorithms

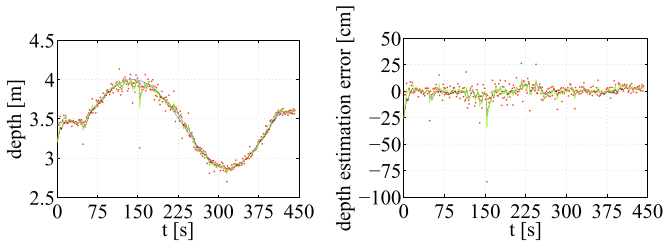
In the sequel, three experiments are reported: two in which the target, a balloon attached to a robot *Pioneer P3-DX* as in Fig. 10, moves along a straight line and a circumference, both with known coordinates, and a third in which the target, the UAV shown in Fig. 14(a), describes a trajectory in the 3-D space with unknown coordinates. The first two experiments are used to study the performance of the proposed depth estimation algorithms, and the third illustrates the behavior of the whole positioning and tracking system that results when depth is inferred from these strategies.

The depth from focus estimates were obtained using as many lines, crossing the target contour estimate, as the number of points of this contour. In the three experiments, 40 pixel wide lines were used. In the first two experiments, the focal length f of the lens and the value considered for the target unknown dimensions R' were 45.6 and 10 mm, respectively, and in the third experiment 20.5 mm and 1 mm, respectively. The dynamics of the discrete-time versions of the second-order complementary filter and LPV observer were derived from their continuous-time equivalents by setting the gains k_1, k_2, and h to 0.4, 0.04, and 0.4, respectively. In the case of the complementary filter, the gains were chosen so that the transfer functions $T_1(s)$ and $T_2(s)$ had two real poles in -0.2. The gain of the LPV observer was chosen according to the relation presented in Lemma 2, where $\alpha_{\max}$ was set to 0.4 (this value can be adjusted for different experiments). Since α can be written in the form $\alpha = \dot{z}/z$, the constraint $\dot{z} \leq 0.4z$, on the target depth derivative, is always verified. For performance comparison purposes, the results obtained in the first two experiments with a discrete-time version of the first-order complementary filter derived in Section IV-A are also presented. This filter was discretized according to the strategy described in Section IV-B for the second-order complementary filter. The gain of the filter was set to 0.4, i.e., equal to the LPV observer gain, and the real dimensions R of the target used in the first two experiments were 35.76 mm. To compute these dimensions, a set of experiments was performed,



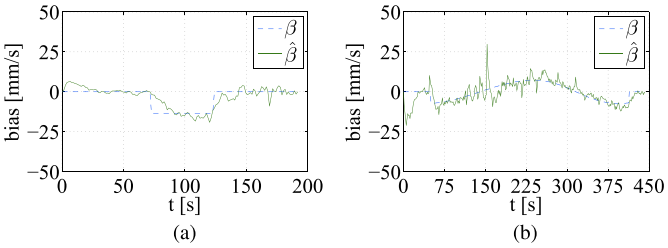
(a) (b)

Fig. 11. Experimental evaluation of the performances of the depth from focus algorithm (in red, $\sigma_{ss} = 45.5$ mm), LPV observer (in black, $\sigma_{ss} = 20.7$ mm), first-order complementary filter (in yellow, $\sigma_{ss} = 24.5$ mm), second-order complementary filter (in green, $\sigma_{ss} = 27.4$ mm), and standard filter (in magenta, $\sigma_{ss} = 33.8$ mm) in the straight-line trajectory experiment. The target real depth is shown in blue. (a) Depth estimation. (b) Depth estimation error.



(a) (b)

Fig. 12. Experimental evaluation of the performances of the depth from focus algorithm (in red, $\sigma_{ss} = 79.8$ mm), LPV observer (in black, $\sigma_{ss} = 37.7$ mm), first-order complementary filter (in yellow, $\sigma_{ss} = 43.8$ mm), and second-order complementary filter (in green, $\sigma_{ss} = 48.5$ mm) in the circular trajectory experiment. The target real depth is shown in blue. (a) Depth estimation. (b) Depth estimation error.



(a) (b)

Fig. 13. Comparison between the real value of the bias that corrupts the measurements of the target depth derivative and the bias estimates obtained with the second-order complementary filter. (a) Straight-line trajectory. (b) Circular trajectory.

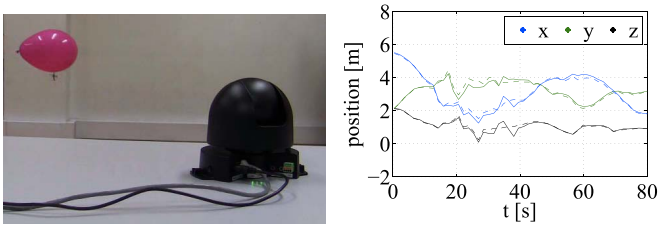
where the target was moved along a calibrated trajectory. The measurements of the dimensions of the projection of the target into acquired images were combined with the calibrated target depths to obtain the real dimensions of the target.

In Figs. 11–13, the performance of the proposed complementary filters and observer is addressed. For comparison purposes, the results obtained using a standard filtering approach that models the dynamics of the target with a single integrator are presented for the straight line trajectory experiment. The gain of this filter was set to 0.4, i.e., equal to the one of the LPV observer. From Figs. 11(a) and 12(a), it is possible to conclude that the estimates provided by all the strategies converge to the vicinity of the target real depth. In other words, the depth estimation errors, shown in Figs. 11(b) and 12(b), converge to the vicinity of zero.

As can be observed from the standard deviations σ_{ss} of the steady-state depth estimation errors presented in Figs. 11(b) and 12(b), the three proposed depth estimation algorithms perform better than the direct measurements provided by the depth from focus strategy. In particular, the LPV observer is the one that leads to the smallest standard deviations of the steady-state depth estimation errors (20.7 mm in the straight line trajectory and 37.7 mm in the circular trajectory), when compared with the first-order complementary filter (24.5 mm in the straight line trajectory and 43.8 mm in the circular trajectory) and with the second-order complementary filter (27.4 mm in the straight line trajectory and 48.5 mm in the circular trajectory). This was unexpected since the first-order complementary filter resorts to additional information (the real target dimensions) that is not used by the observer. A possible explanation for this fact has to do with the similarities between the structure of this filter and the structure of the observer, which were detailed in Section V. The main difference between the two is in the computation of the target depth derivative. In the complementary filter, it depends on the square of r_m, and in the observer, it depends directly on r_m. Thus, the influence of the noise that corrupts this quantity is smaller in the observer. This is a significant advantage for the observer-based approach, which does not require any knowledge about the dimensions of the target. Since the computation of the target depth derivative in the second-order complementary filter also depends on the square of r_m, a similar reasoning can be used to explain why the observer outperforms this strategy.

From Fig. 11, it is possible to conclude that the performance of the proposed estimators is better than the performance of the aforementioned standard filtering approach, especially when the target is moving (approximately between the 70 s and the 120 s). This difference results from the fact that this simple strategy does not use any information about the dynamics of the target. This is not the case with the proposed filters that use measurements of variations in the dimensions of the image of the target to obtain such information. The difference in performance between this approach and the proposed ones is more significant if the RMSE is used for evaluation purposes, as the depth estimates provided by the standard filter have an offset when the target is moving [Fig. 11(b)]. If only this period is considered, the RMSE of the standard filter is 55.7 mm, which is even worse than the performance of the depth from focus measurements (RMSE = 43.2 mm). The RMSEs associated with the LPV observer, firstorder complementary filter, and second-order complementary filter are 24.1, 35.9, and 37.7 mm, respectively.

There are several reasons that can explain the errors observed in Figs. 11(b) and 12(b): 1) uncertainty associated with the characterization of the real trajectory of the target; 2) errors resulting from the fitting of the cost function; 3) errors in the calibration of the camera intrinsic parameters, whose values vary with changes in the camera focus setting; and 4) uncertainty associated with the calibration of the relation between the focus value and the focus setting of the camera.



(a) (b)

Fig. 14. Real-time tracking of a UAV. (a) UAV and AXIS 215 PTZ. (b) Position estimation. In (b), solid and dashed lines represent the estimates of the UAV position obtained using depth estimates provided by the second-order complementary filter and LPV observer, respectively.

The values of the bias estimates $\hat{\beta}$, provided by the second-order complementary filter, result from the unknown nature of the target dimensions and are shown in Fig. 13.

The third experiment, in which a small UAV with unknown dimensions moves in the 3-D space, is shown in Fig. 14. In Fig. 14(a), the UAV and the used camera are shown, and in Fig. 14(b), the real-time estimates of the UAV position $\mathbf{p}$ in the inertial reference frame $\{\mathcal{I}\}$, provided by the tracking system described in Section II, are presented. An analysis of the performance of the system, i.e., a comparison between the real and estimated target positions, is not provided for this experiment since the coordinates of the trajectory of the UAV are not known. A demo movie illustrating the behavior of the system in this case can be found online in [39].

As explained in the end of Section III, it is possible to estimate the dimensions of the target when they do not vary over time. To confirm this statement, a target with dimensions $R = 73.19$ mm was placed 3 m away from the camera. The use of the approach described in Section III to estimate the dimensions of this target leads to $\hat{R} = 73.23$ mm, when the depth estimates provided by the second-order complementary filter are used, and to $\hat{R} = 73.22$ mm, when the observer depth estimates are considered. As can be seen, these values are very close to the target real dimensions. The standard deviations of the errors associated with the target dimension estimates obtained with the filter and with the observer during the 65 s of the experiment were 1.18 and 1.07 mm, respectively.

VII. Conclusion

In this paper, new methodologies for the estimation of the depth of a target with unknown dimensions were proposed. The measurements of the target depth, extracted from images acquired with a single camera and based upon depth from focus techniques were considered. These measurements were processed using a complementary filter and a LPV observer, whose analysis and synthesis were provided. This paper complements an inexpensive single pan and tilt camera-based indoor positioning and tracking system, as it can be used to estimate the instantaneous depth of a moving target with unknown dimensions. The performance of the overall system was assessed using a series of indoor experimental tests. A centimetric accuracy was obtained for a range of operation of up to ten meters. The proposed system was also used to track and localize in real time a small indoor UAV with

unknown dimensions. In the future, this system will be used to track and estimate in real time the 3-D position of marine animals under captivity for behavioral studies.

ACKNOWLEDGMENT

The authors would like to thank P. Favaro and S. Soatto who provided the MATLAB code used to obtain the comparison results presented in Section VI-A.

REFERENCES

[1] K. W. Kolodziej and J. Hjelm, *Local Positioning Systems: LBS Applications and Services*. Boca Raton, FL, USA: CRC Press, 2006.
[2] Y. Bar-Shalom, X. Rong Li, and T. Kirubarajan, *Estimation With Applications to Tracking and Navigation: Theory Algorithms and Software*. New York, NY, USA: Wiley, 2001.
[3] P. Saeedi, P. D. Lawrence, and D. G. Lowe, "Vision-based 3-D trajectory tracking for unknown environments," *IEEE Trans. Robot.*, vol. 22, no. 1, pp. 119–136, Feb. 2006.
[4] S. Rezaei and R. Sengupta, "Kalman filter-based integration of DGPS and vehicle sensors for localization," *IEEE Trans. Control Syst. Technol.*, vol. 15, no. 6, pp. 1080–1088, Nov. 2007.
[5] G. Antonelli and S. Chiaverini, "A deterministic filter for simultaneous localization and odometry calibration of differential-drive mobile robots," in *Proc. 3rd IEEE Eur. Conf. Mobile Robots*, Sep. 2007.
[6] M. Linderoth, A. Robertsson, K. Åström, and R. Johansson, "Object tracking with measurements from single or multiple cameras," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 4525–4530.
[7] V. Lepetit and P. Fua, "Monocular model-based 3D tracking of rigid objects," *Found. Trends Comput. Graph. Vis.*, vol. 1, no. 1, pp. 1–89, 2005.
[8] T. Gaspar and P. Oliveira, "Single pan and tilt camera indoor positioning and tracking system," *Eur. J. Control*, vol. 17, no. 4, pp. 414–428, 2011.
[9] L. Bertelli, P. Ghosh, B. S. Manjunath, and F. Gibou, "Robust depth estimation for efficient 3D face reconstruction," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 1516–1519.
[10] A. Discant, A. Rogozan, C. Rusu, and A. Bensrhair, "Sensors for obstacle detection—A survey," in *Proc. 30th Int. Spring Seminar Electron. Technol.*, May 2007, pp. 100–105.
[11] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: Real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809–830, Aug. 2000.
[12] Y. Y. Schechner and N. Kiryati, "Depth from defocus vs. stereo: How different really are they?" in *Proc. 14th IEEE Int. Conf. Pattern Recognit.*, Nov. 1998, pp. 1784–1786.
[13] H. Q. H. Viet, M. Miwa, H. Maruta, and M. Sato, "Recognition of motion in depth by a fixed camera," in *Proc. 7th Digit. Image Comput., Techn. Appl.*, Dec. 2003, pp. 205–214.
[14] A. Pentland, T. Darrell, M. Turk, and W. Huang, "A simple, real-time range camera," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1989, pp. 256–261.
[15] N. Asada, M. Baba, and A. Oda, "Depth from blur by zooming," in *Proc. Vis. Inter. Annu. Conf.*, May 2001, pp. 165–172.
[16] J. Ens and P. Lawrence, "An investigation of methods for determining depth from focus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 2, pp. 97–108, Feb. 1993.
[17] R. Ben-Ari, "A unified approach for registration and depth in depth from defocus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1041–1055, Jun. 2014.
[18] E. Krotkov, "Focusing," *Int. J. Comput. Vis.*, vol. 1, no. 3, pp. 223–237, Oct. 1987.
[19] S. K. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 8, pp. 824–831, Aug. 1994.
[20] R. R. Sahay and A. N. Rajagopalan, "Dealing with parallax in shape-from-focus," *IEEE Trans. Image Process.*, vol. 20, no. 2, pp. 558–569, Feb. 2011.
[21] T. Gaspar and P. Oliveira, "Monocular depth from focus estimation with complementary filters," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 4986–4991.
[22] T. Gaspar and P. Oliveira, "New dynamic estimation of depth from focus in active vision systems," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, 2011, pp. 484–491.
[23] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[24] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.
[25] A. Blake and M. Isard, *Active Contours*, 1st ed. New York, NY, USA: Springer-Verlag, 2000.
[26] E. Hecht, *Optics*, 4th ed. Reading, MA, USA: Addison-Wesley, 2001.
[27] R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*. New York, NY, USA: Wiley, 1997.
[28] S. F. Ray, *Applied Photographic Optics*, 3rd ed. Waltham, MA, USA: Focal Press, 2002.
[29] K. Sun and A. Packard, "Robust H_2 and H_∞ filters for uncertain LFT systems," *IEEE Trans. Autom. Control*, vol. 50, no. 5, pp. 715–720, May 2005.
[30] W. J. Rugh, *Linear System Theory*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.
[31] K. Tarabanis, R. Y. Tsai, and D. S. Goodman, "Modeling of a computer-controlled zoom lens," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 2. May 1992, pp. 1545–1551.
[32] N. B. Priyantha, "The cricket indoor location system," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2005.
[33] H. Hile and G. Borriello, "Positioning and orientation in indoor environments using camera phones," *IEEE Comput. Graph. Appl.*, vol. 28, no. 4, pp. 32–39, Jul./Aug. 2008.
[34] P. Favaro and S. Soatto, "A geometric approach to shape from defocus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 406–417, Mar. 2005.
[35] M. Subbarao and G. Surya, "Depth from defocus: A spatial domain approach," *Int. J. Comput. Vis.*, vol. 13, no. 3, pp. 271–294, 1994.
[36] J. Bouguet. *Camera Calibration Toolbox for MATLAB*. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/, accessed Jan. 2015.
[37] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 993–1008, Aug. 2003.
[38] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*, vol. 201. Upper Saddle River, NJ, USA: Prentice-Hall, 1998.
[39] T. Gaspar and P. Oliveira. *Depth Estimation in Active Monocular Vision Systems for Indoor 3D Tracking*. [Online]. Available: http://users.isr.ist.utl.pt/~tgaspar/DFF_tracker_movie, accessed Jan. 2015.

Tiago Gaspar (S'11) received the M.Sc. degree in electrical and computer engineering from the Instituto Superior Técnico, Lisbon, Portugal, in 2008, where he is currently pursuing the Ph.D. degree in electrical and computer engineering.

His current research interests include positioning and tracking systems, sensor and signal fusion, nonlinear estimation, and video synchronization.

Paulo Oliveira (M'92–SM'11) received the Ph.D. degree in electrical and computer engineering from the Instituto Superior Técnico (IST), Lisbon, Portugal, in 2002.

He participated in more than 30 European and Portuguese research projects, over the last 25 years. He is currently an Associate Professor with the Department of Mechanical Engineering, IST, where he is a Researcher with the Institute of Mechanical Engineering. He also collaborates with the Institute for Systems and Robotics, IST. He co-authored over 50 journal and 150 conference papers. His current research interests include mechatronics with a special focus on the fields of autonomous vehicles, robotics, sensor fusion, navigation, positioning, and nonlinear estimation.