2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

On-Line Estimation of Feature Depth for Image-Based Visual Servoing Schemes

Alessandro De Luca Giuseppe Oriolo Paolo Robuffo Giordano
Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza"
Via Eudossiana 18, 00184 Roma, Italy
{deluca,oriolo,robuffo}@dis.uniroma1.it

***Abstract*— In the image-based visual servoing framework, error signals are directly computed from image feature parameters, thus obtaining control schemes which do not need neither a 3-D model of the scene, nor a perfect knowledge of the camera calibration matrix. However, the current value of the depth Z for each considered feature must be known. We propose a method to estimate on-line the value of Z for point features while the camera is moving through the scene, by using tools from nonlinear observer theory. By interpreting Z as a continuous unknown state with known dynamics, we build an estimator which asymptotically recovers the actual depth value for the selected feature.**

I. INTRODUCTION

The introduction of visual information in the control loop of robot systems has increased the flexibility and the accuracy of the tasks commonly performed by these systems [1], [2], by providing higher position accuracy, robustness to sensor noise and calibration uncertainties, and reactiveness to environmental changes. This is especially true for the class of mobile robots, where the elaboration of visual cues is often crucial for self-localization and navigation. Another interesting use of visual feedback is the possibility to specify a robotic task in terms of some image features extracted from a target object while the camera/robot is moving through the scene.

Two basic approaches have been proposed in the past years to deal with this kind of task, namely *position-based visual servoing* (PBVS) and *image-based visual servoing* (IBVS) [1]. In PBVS, the image features are processed in order to estimate the relative 3D pose between the camera and the target, which is then used as an error signal for controlling the motion of the robot/camera system toward its desired goal [3]. In IBVS the error is directly computed in terms of the features, whose motion on the image plane is related to the velocity twist of the camera via the *interaction matrix*. The advantages of IBVS over PBVS are the following: *(i)* a 3D model of the target is not needed; *(ii)* performance is robust with respect to perturbations of the robot/camera models, in particular to calibration errors [4]; *(iii)* it is easier to devise feature-based motion strategies aimed at keeping the target always in the field of view of the camera [5]. However, there are also some drawbacks to be considered. Apart from situations where the interaction matrix loses rank during the motion, local minima of the task error function [6] may be encountered when trying to impose an (unfeasible) independent motion to a large number of image features [7]. Moreover, the feature depths are unknown in a pure IBVS setting, and must be estimated during servoing in order to correctly compute the interaction matrix (a common choice is to simply use their constant value at the desired pose). Thus, only local stability can be guaranteed for most IBVS schemes [8].

In this paper, we address the estimation problem of the unknown depth Z of a static point in an IBVS scheme. Our starting idea is that, since the motion of the feature on the image plane depends upon the current value of Z, it is possible to estimate this value by comparing the measured motion with the one predicted by using the current estimate of Z, under the assumption of a perfect knowledge of the camera 3D motion and of its intrinsic parameters. This is a typical issue of the more general paradigm of *motion and structure reconstruction*, whose purpose is to design an identification scheme to estimate both the camera motion and the structure (i.e., the 3D geometry) of the scene. Our work assumes a known relative motion among the camera and the target, which can be achieved, for example, if the point feature is fixed in the world and the camera is mounted on the end-effector of a robot manipulator.

In the last years, several works have addressed the structure identification with known motion. Chaumette et al. [9] propose a general methodology to recover the 3D information of several geometric primitives (points, lines, cylinders, spheres, etc.) by measuring the current values of the features, of the image motion (the feature time derivatives) and of the camera velocity twist. However, due to the presence of noise and discrete sampling, the extraction of the image motion is not trivial, and some constraints on the allowed camera motions must be considered. In [10], two Kalman filter-based algorithms are derived and compared, the first estimating a continuous depth map of the scene, and the second extracting the depth of a discrete set of features. Both methods need the computation of the current image motion, and impose several constraints on the camera motion in order to simplify the problem. In particular, the second method assumes a camera which translates orthogonally to the optical axis (without rotations), so that the depth of the features is kept constant and the problem is considerably simplified. A similar approach is found in [11], where, again, only lateral camera motions are allowed. Adaptive IBVS

1-4244-0602-1/07/$20.00 ©2007 IEEE.

schemes are devised in [12], [13] for a camera mounted on a nonholonomic mobile robot via an on-line estimation of a constant unknown parameter (the height of the object points and the depth of the target plane at the desired pose, respectively). In fact, whenever the depth is kept constant during the camera motion, or the value of Z is related to any other fixed quantity as in, e.g., [10]–[13], the problem of depth identification can be formulated in the adaptive control context, where several tools allow, under suitable hypothesis, to estimate an unknown constant parameter.

With respect to these works, in this paper we tackle the problem of depth identification for static features point without any preliminary constraint on the camera motion, and without the explicit need for image motion estimation; thus, the only information used is the current value of the features measured on the image plane. This is obtained by recasting the problem into the nonlinear observer framework, which provides techniques to estimate unmeasurable time-varying states of known dynamical systems.

The novel contribution of the paper are therefore:

- the derivation of a nonlinear observer built upon the exact kinematic equations of the camera-target system,
- the elimination of unnecessary hypotheses on the camera motion often present in the past literature, e.g., motion along a plane or to keep a constant depth,
- the removal of image motion from the quantities strictly needed by the depth estimation process.

The paper is organized as follows: in Sect. II we recall the basic kinematic relationships of the camera/target system, while in Sect. III we design a nonlinear observer to estimate the unknown value of Z. Finally, in Sect. IV some simulations are presented in order to show the performance of the proposed method.

II. Perspective camera model

An *image feature* is any real-valued quantity associated to a selected primitive (e.g., the coordinates of a point, the area of an ellipse, the angular coefficient of a line, etc.) in the image plane. Given a vector of features $f = [f_1 \dots f_k]^T \in \mathbb{R}^k$, the velocity twist (V, ω) of the camera is mapped to $\dot{f}$ by a $k \times 6$ matrix $J(f, Z)$ called *interaction matrix*

$$\dot{f} = J(f, Z) \begin{bmatrix} V \\ \omega \end{bmatrix},$$

where $Z \in \mathbb{R}^k$ is the vector of the depths associated to each feature in f [14]. It is possible to determine the interaction matrix for many features of interest, see [2] for the case of points, lines, planes, circles, etc., and [15] for the set of image moments. Since in this work we address the depth identification for point features, we will focus on the interaction matrix linking the camera velocity twist to the 2D point velocity on the image plane.

With reference to Fig. 1, consider a world reference frame $\mathcal{F}_O : \{O; \overrightarrow{X}_O, \overrightarrow{Y}_O, \overrightarrow{Z}_O\}$ and a pin-hole camera associated to the moving frame $\mathcal{F}_C : \{O_C; \overrightarrow{X}_C, \overrightarrow{Y}_C, \overrightarrow{Z}_C\}$ with Z_C coincident with the camera optical axis. The image plane,

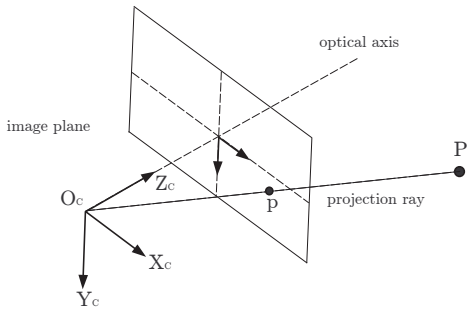

Fig. 1. World and camera frame definitions.

perpendicular to the optical axis, lies at a distance λ (the focal length) from O_C, and is endowed with a 2D reference frame $\mathcal{F}_I : \{O_I; \overrightarrow{u}, \overrightarrow{v}\}$ with axes parallel to $\overrightarrow{X}_C$ and $\overrightarrow{Y}_C$, respectively.

Consider a fixed 3D point P whose coordinates in the camera frame $\mathcal{F}_C$ are $({}^CX, {}^CY, {}^CZ)$. The velocity of P in $\mathcal{F}_C$ is expressed as [14]

$$\begin{bmatrix} {}^C\dot{X} \\ {}^C\dot{Y} \\ {}^C\dot{Z} \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 & 0 & -{}^CZ & {}^CY \\ 0 & -1 & 0 & {}^CZ & 0 & -{}^CX \\ 0 & 0 & -1 & -{}^CY & {}^CX & 0 \end{bmatrix} \begin{bmatrix} V \\ \omega \end{bmatrix}. \tag{1}$$

In the following, we will drop the dependency on $\mathcal{F}_C$ since we will always refer to quantities expressed in the camera frame, unless otherwise stated.

The pin-hole camera model projects a 3D point P into the image point $p = [f_1\, f_2]^T$ determined by the intersection of the projection ray with the image plane (see Fig. 1). A simple but widely used projection model is [14]:

$$\begin{aligned} f_1 &= \lambda \frac{X}{Z} \\ f_2 &= \lambda \frac{Y}{Z}. \end{aligned} \tag{2}$$

By differentiating (2) w.r.t. time and using (1) we get the well-known relationship

$$\begin{aligned} \begin{bmatrix} \dot{f}_1 \\ \dot{f}_2 \end{bmatrix} &= \begin{bmatrix} -\frac{\lambda}{Z} & 0 & \frac{f_1}{Z} & \frac{f_1 f_2}{\lambda} & -\left(\lambda + \frac{f_1^2}{\lambda}\right) & f_2 \\ 0 & -\frac{\lambda}{Z} & \frac{f_2}{Z} & \lambda + \frac{f_2^2}{\lambda} & -\frac{f_1 f_2}{\lambda} & -f_1 \end{bmatrix} \begin{bmatrix} V \\ \omega \end{bmatrix} \\ &= J_p(f_1, f_2, Z) \begin{bmatrix} V \\ \omega \end{bmatrix} \end{aligned} \tag{3}$$

where $J_p(f_1, f_2, Z)$ is the 2×6 point feature *interaction matrix*. Note that, since only the first three columns of J_p are affected by the value of Z, a pure camera rotation does not bring any information useful for depth estimation: a camera translation must be necessarily present. This intuitive conclusion, already well known in the context of the observability of dynamical systems with perspective outputs (see, e.g., [16]), will be reobtained in Sect. III in terms of the *persistency of excitation* condition that will explicitly characterize which camera motions are useless for the depth estimation.

III. Design of the nonlinear observer

The purpose of this section is to design a nonlinear observer which will estimate the value of Z during the motion of the camera. We will assume a calibrated camera, i.e., the value of λ is known.

It is convenient to rewrite eqs. (1) and (3) in a more canonical form. Let $x = [f_1\ f_2\ Z]^T \in \mathbb{R}^3$ be the state vector and $u = [V\ \omega]^T \in \mathbb{R}^6$ be the input vector. Hence, using (3) and the last row of (1), the state dynamics are expressed by the driftless system

$$\dot{x} = \begin{bmatrix} -\frac{\lambda}{x_3} & 0 & \frac{x_1}{x_3} & \frac{x_1 x_2}{\lambda} & -\left(\lambda + \frac{x_1^2}{\lambda}\right) & x_2 \\ 0 & -\frac{\lambda}{x_3} & \frac{x_2}{x_3} & \lambda + \frac{x_2^2}{\lambda} & -\frac{x_1 x_2}{\lambda} & -x_1 \\ 0 & 0 & -1 & -\frac{x_2 x_3}{\lambda} & \frac{x_1 x_3}{\lambda} & 0 \end{bmatrix} u$$
$$y = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \tag{4}$$

where the output vector $y \in \mathbb{R}^2$ represents the measurable variables, i.e., the coordinates of the point p on the image plane. Consider the change of coordinates

$$\widetilde{x} = \begin{bmatrix} x_1 \\ x_2 \\ \frac{1}{x_3} \end{bmatrix},$$

which is globally defined since $x_3(t) > \lambda > 0$ (i.e., the point P is supposed to lie always in front of the image plane, otherwise the visual servoing would fail). In the new coordinates, system (4) becomes

$$\dot{\widetilde{x}} = \begin{bmatrix} -\lambda\widetilde{x}_3 & 0 & \widetilde{x}_1\widetilde{x}_3 & \frac{\widetilde{x}_1\widetilde{x}_2}{\lambda} & -\left(\lambda + \frac{\widetilde{x}_1^2}{\lambda}\right) & \widetilde{x}_2 \\ 0 & -\lambda\widetilde{x}_3 & \widetilde{x}_2\widetilde{x}_3 & \lambda + \frac{\widetilde{x}_2^2}{\lambda} & -\frac{\widetilde{x}_1\widetilde{x}_2}{\lambda} & -\widetilde{x}_1 \\ 0 & 0 & \widetilde{x}_3^2 & \frac{\widetilde{x}_2\widetilde{x}_3}{\lambda} & -\frac{\widetilde{x}_1\widetilde{x}_3}{\lambda} & 0 \end{bmatrix} u$$
$$y = \begin{bmatrix} \widetilde{x}_1 \\ \widetilde{x}_2 \end{bmatrix}. \tag{5}$$

Since (5) is driftless and the output has dimension smaller than the state, its linear approximation at any point is unobservable. This is a consequence of the intrinsic nonlinear nature of (5) in the sense that any linear time-invariant approximation will lose the observability property. A suitable nonlinear observer is then needed in order to correctly address the problem.

Let $\widehat{x} \in \mathbb{R}^3$ be the estimate of the (partially) unknown state $\widetilde{x}$. We seek an update law in the form

$$\dot{\widehat{x}} = \alpha(\widehat{x}, y)u + \beta(\widehat{x}, y, u) \tag{6}$$

which guarantees $\lim_{t\to\infty} \|\widehat{x}(t) - \widetilde{x}(t)\| = 0, \ \forall \widehat{x}(t_0)$. The design of the functions $\alpha(\widehat{x}, y)$ and $\beta(\widehat{x}, y, u)$ will be based upon the following result, known as the *persistency of excitation* lemma. For the reader's convenience, we will report the full statement whose proof can be found in [17, Lemma B.2.3].

Lemma 1: Consider the linear time-varying system

$$\begin{cases} \dot{\xi} = H\xi + \Omega^T(t)z, & \xi \in \mathbb{R}^n \\ \dot{z} = -\Lambda\Omega(t)P\xi, & z \in \mathbb{R}^p \end{cases} \tag{7}$$

where H is an $n\times n$ Hurwitz matrix, P is an $n\times n$ symmetric positive definite matrix such that $H^TP + PH = -Q$, with Q symmetric positive definite, and Λ is a $p\times p$ symmetric positive definite matrix. If $\|\Omega(t)\|$, $\|\dot{\Omega}(t)\|$ are uniformly bounded and the *persistency of excitation* condition is satisfied, i.e., there exist two positive real numbers T and γ such that

$$\int_t^{t+T} \Omega(\tau)\Omega^T(\tau)\mathrm{d}\tau \geq \gamma I > 0, \qquad \forall\, t \geq t_0, \tag{8}$$

then $(\xi,\, z) = 0$ is a globally exponentially stable equilibrium point. ■

We now perform some manipulation in order to be able to apply Lemma 1 to our case. Let $e = \widetilde{x} - \widehat{x}$ be the error vector, and note that the subvector $[e_1\ e_2]^T$ is directly accessible for measurements. Thus, if we define the observer as in (6) with

$$\alpha(\widehat{x}, y) = \begin{bmatrix} -\lambda\widehat{x}_3 & 0 & y_1\widehat{x}_3 & \frac{y_1y_2}{\lambda} & -\left(\lambda + \frac{y_1^2}{\lambda}\right) & y_2 \\ 0 & -\lambda\widehat{x}_3 & y_2\widehat{x}_3 & \lambda + \frac{y_2^2}{\lambda} & -\frac{y_1y_2}{\lambda} & -y_1 \\ 0 & 0 & \widehat{x}_3^2 & \frac{y_2\widehat{x}_3}{\lambda} & -\frac{y_1\widehat{x}_3}{\lambda} & 0 \end{bmatrix}$$
$$\beta(\widehat{x}, y, u) = \begin{bmatrix} k_1e_1 \\ k_2e_2 \\ k_3((-\lambda u_1 + y_1u_3)e_1 + (-\lambda u_2 + y_2u_3)e_2) \end{bmatrix} \tag{9}$$

with $k_1, k_2, k_3 > 0$, we get the error dynamics

$$\begin{aligned} \dot{e}_1 &= -k_1e_1 + (-\lambda u_1 + y_1u_3)e_3 \\ \dot{e}_2 &= -k_2e_2 + (-\lambda u_2 + y_2u_3)e_3 \\ \dot{e}_3 &= -k_3\begin{bmatrix} -\lambda u_1 + y_1u_3 & -\lambda u_2 + y_2u_3 \end{bmatrix}\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + \\ &\quad (\widetilde{x}_3^2 - \widehat{x}_3^2)u_3 + \left(\frac{y_2u_4 - y_1u_5}{\lambda}\right)e_3. \end{aligned} \tag{10}$$

If we set

$$\begin{aligned} \xi &= [e_1\ e_2]^T \\ z &= e_3 \\ H &= \begin{bmatrix} -k_1 & 0 \\ 0 & -k_2 \end{bmatrix} \\ \Omega(t) &= \begin{bmatrix} -\lambda u_1 + y_1u_3 & -\lambda u_2 + y_2u_3 \end{bmatrix} \\ \Lambda &= k_3 \\ P &= I, \end{aligned}$$

system (10) is very close to the formulation in (7), the only differences being the last two terms in the e_3 dynamics.

It is worth noting that, when

$$u_3(t) \equiv u_4(t) \equiv u_5(t) \equiv 0, \tag{11}$$

the two formulations match exactly and the global exponential stability of e is guaranteed, as long as the conditions of Lemma 1 are met. While we will thoroughly discuss such conditions in the forthcoming analysis, we would like to emphasize that (11) corresponds to a camera motion which

keeps the depth Z constant. As explained in Sect. I, in this case the problem is considerably simplified and can be attacked with various techniques. The purpose of our analysis is to show that (10) can converge also when (11) does not hold.

Proposition 1: Using the observer (6)–(9), the origin of the error system (10) is exponentially stable as long as the conditions of Lemma 1 are verified, in particular condition (8).

Proof: It is useful to rewrite (10) as $\dot{e} = A(t)e + g(e,t)$ with

$$A(t) = \begin{bmatrix} -k_1 & 0 & \Omega_1(t) \\ 0 & -k_2 & \Omega_2(t) \\ -k_3\Omega_1(t) & -k_3\Omega_2(t) & 0 \end{bmatrix}$$

$$g(e,t) = \begin{bmatrix} 0 \\ 0 \\ (\widetilde{x}_3^2 - \widehat{x}_3^2)u_3 + \left(\frac{y_2u_4 - y_1u_5}{\lambda}\right) e_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \left(2\widetilde{x}_3u_3 + \frac{y_2u_4 - y_1u_5}{\lambda}\right) e_3 - u_3e_3^2 \end{bmatrix}. \quad (12)$$

We can consider $g(e,t)$ as a perturbation term of the nominal system $\dot{e} = A(t)e$ which, if (8) holds, is guaranteed by Lemma 1 to be globally exponentially stable. Note that $g(e,t)$ is a vanishing perturbation, i.e., $g(0,t) = 0,\ \forall t$. Several tools are available for the stability analysis of globally exponentially stable systems with vanishing perturbations (see [18] for an overview). Generally, if $\|g(e,t)\|$ is sufficiently small, the exponential stability is preserved, at least locally. Due to the boundedness of $\|\Omega(t)\|$ and $\|\dot{\Omega}(t)\|$, the system $\dot{e} = A(t)e$ is an exponentially stable slowly varying linear system, and therefore there exists a suitable Lyapunov function $V(e,t)$ such that

$$c_1e^Te \le V \le c_2e^Te$$
$$\dot{V}(e,\,t) = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial e}A(t)e \le -c_3\|e\|^2$$
$$\left\|\frac{\partial V}{\partial e}\right\| \le c_4\|e\|,$$

with $c_1 \dots c_4$ positive constants. To derive bounds on $g(e,t)$ note that $0 < \widetilde{x}_3(t) < 1/\lambda$, $|y_1(t)| \le M_1$ and $|y_2(t)| \le M_2$ where M_1, M_2 are the (finite) dimensions of the image plane[1], and, since $\|\Omega(t)\|$ is bounded, there exists a positive constant M_3 such that $|u_3(t)| \le M_3$, $|u_4(t)| \le M_3$ and $|u_5(t)| \le M_3$. Finally, by defining E as the maximum value of $|e_3(t)|$, we have $|e_3^2| \le E|e_3| \le E\|e\|$. Hence,

$$\|g(e,t)\| \le \left(\frac{(2 + M_1 + M_2)}{\lambda} + E\right) M_3\|e\| = \gamma\|e\|.$$

Using the Lyapunov candidate $V(e,t)$ for the perturbated system (12), we get

$$\dot{V}(e,t) \le -c_3\|e\|^2 + \left\|\frac{\partial V}{\partial e}\right\| \|g(e,t)\| \le -c_3\|e\|^2 + c_4\gamma\|e\|^2.$$

[1]We are implicitly assuming a camera motion such that the object of interest is always kept in the field of view.

If γ is small enough to satisfy the bound $\gamma < c_3/c_4$, $\dot{V}$ is negative definite and system (10) is exponentially stable. Note that, for given camera parameters (focal length and image plane size) and motion (u_3, u_4, u_5) the constant γ only depends on E, i.e., the maximum value of $|e_3(t)|$. This can be made arbitrarily small by choosing the initial condition $e_3(t_0)$ inside a suitable level set $S_c = \{e \in \mathbb{R}^3 |\, V(e,t_0) \le c\}$, since we have

$$E \le \|e(t)\| \le \|e(t_0)\| \le \frac{V(e,\,t_0)}{c_1} \le \frac{c}{c_1}. \quad \blacksquare$$

Note that the above stability result is of a local nature, since convergence of the error to zero is only guaranteed in a suitable neighborhood of the origin. A less conservative estimate of this neighborhood may be obtained by considering that our observer will be obviously initialized with the measured values of the feature, so that $e_1(t_0) = e_2(t_0) = 0$. This implies that $|e_3(t)| \le |e(t_0)| = E$, so that

$$|e_3(t_0)| \le \frac{c_3}{c_4M_3} - \frac{2 + M_1 + M_2}{\lambda}$$

guarantees exponential error convergence.

The conditions of Lemma 1 deserve some additional considerations. First of all the boundedness of $\|\Omega(t)\|$ and $\|\dot{\Omega}(t)\|$ requires that the input signal $u(t)$ is bounded with bounded derivatives. As for condition (8), it has a deeper meaning: it essentially states that there must not exist a $\bar{t}$ such that $\forall t > \bar{t},\ \|\Omega(t)\| \equiv 0$. By direct inspection of the expression of $\Omega(t)$, we can see that this can only happen if

1) $\exists\, \bar{t} \,|\, \forall\, t > \bar{t}:\quad u_1(t) \equiv 0,\ u_2(t) \equiv 0,\ u_3(t) \equiv 0$, i.e., if no translations are involved in the camera motion;
2) $\exists\, \bar{t} \,|\, \forall\, t > \bar{t}:\quad \lambda u_1 = y_1u_3,\ \lambda u_2 = y_2u_3$, which is equivalent to

$$\frac{u_1}{u_3} = \frac{X}{Z},\ \frac{u_2}{u_3} = \frac{Y}{Z}.$$

This means that the camera is translating along the projection ray of the selected point p, and, thus, the projection of P on the image plane is kept constant during the motion.

It is interesting to note that such persistency of excitation condition, essential for the estimation convergence, is basically due to the scale ambiguity present in every perspective system. Indeed, it is well known (see, e.g., [14]) that within a perspective system it is impossible to distinguish an object from the same object twice as big, twice as far and moving twice as fast. The condition of nonzero (and known) camera translational velocity introduces a scale information which is essential to disambiguate among all the equivalent states and to successfully recover the actual feature depth. A somehow related excitation condition is also derived in [19] as a byproduct of the proposed minimum-energy estimator for perspective systems. It is also worth citing [20], where a conceptually similar scheme is proposed, although not directly based on the persistency of excitation principle.

IV. Simulations

In this section we will present some simulations in order to show the performance of the observer (6)–(9) derived in Sect. III. We begin with the simple case of a camera performing a sinusoidal motion along the $\overrightarrow{Z}$-axis, a case that, e.g., could not be addressed with the methods in [10], [11] (see Sect. I). The simulation data are:

$$\begin{aligned} \widetilde{x}(t_0) &= [24 \quad -5 \quad 2]^T \\ \widehat{x}(t_0) &= [24 \quad -5 \quad 1]^T \\ u_3(t) &= 0.5\cos\pi t \\ k_1 &= 20 \\ k_2 &= 20 \\ k_3 &= 0.5 \\ \lambda &= 128 \end{aligned} \quad .$$

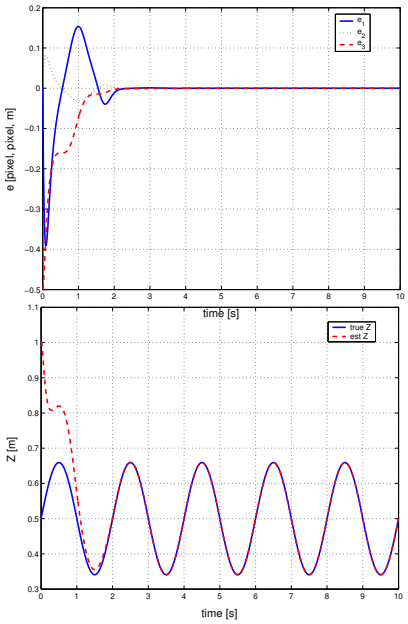

Fig. 2. First simulation. Above: error behaviour. Below: true (solid blue line) and estimated (dashed red line) Z (we recall that $Z = x_3 = 1/\widetilde{x}_3$).

Note that the first two components of the estimation error are initially zero because the feature position is measured. Figure 2 depicts the behaviour of $e(t)$ during the simulation and shows how the estimate of Z approaches the true value. Convergence is reached after few seconds of motion.

Our second experiment involves a more complex camera motion consisting of a translation and a rotation about the $\overrightarrow{X}$ and $\overrightarrow{Z}$ axes. We set

$$\begin{aligned} \widetilde{x}(t_0) &= [10 \quad -10 \quad 2]^T \\ \widehat{x}(t_0) &= [10 \quad -10 \quad 1]^T \\ u_1(t) &= 0.1\cos 2\pi t \\ u_3(t) &= 0.5\cos\pi t \\ u_4(t) &= 0.6\cos\pi/2\, t \\ u_6(t) &= 1 \\ k_1 &= 20 \\ k_2 &= 20 \\ k_3 &= 0.5 \\ \lambda &= 128 \end{aligned} \quad .$$

Figure 3 shows again a good convergence behaviour even if in this case the camera motion is quite complex.

As an additional case study, we implemented the proposed algorithm in the Webots environment [21] by considering a camera with $\lambda = 128$ pixels mounted on the end-effector of a mobile manipulator made of a unicycle-like platform carrying a 3R spatial arm (see Fig. 4). Video clips of this simulation can be found at `www.dis.uniroma1.it/~labrob/research/depth_est.html` and are also included in the video attachment to this paper. The idea was to test the performance of the proposed observer against the command/measurement noise automatically introduced by the webots engine (roughly equivalent to a white noise with std. deviation $\sigma = 0.1$ pixels added to the extracted feature coordinates). The objective is to estimate the depth of the target point (the red dot on the cube in Fig. 4), while the first and second link of the manipulator move according to the velocity profiles:

$$\begin{aligned} \dot{q}_1 &= 0.2\sin 0.4\pi t \\ \dot{q}_2 &= 0.1\sin 0.8\pi t. \end{aligned}$$

The initial value of the estimated depth is set at $1/\widehat{x}_3(t_0) = 0.05$ and the gains were chosen as $k_1 = k_2 = 10$ and $k_3 = 0.8$. Despite the noise, the observer is able to estimate accurately the actual value of the depth Z, as shown in Fig. 5.

V. Conclusions

By borrowing techniques from the nonlinear observer theory, we were able to design a nonlinear observer which estimates the unknown depth of a point feature during the motion of the camera. In contrast with previous works, we do not need to estimate the image motion of the features, nor to constrain the camera motion along any particular

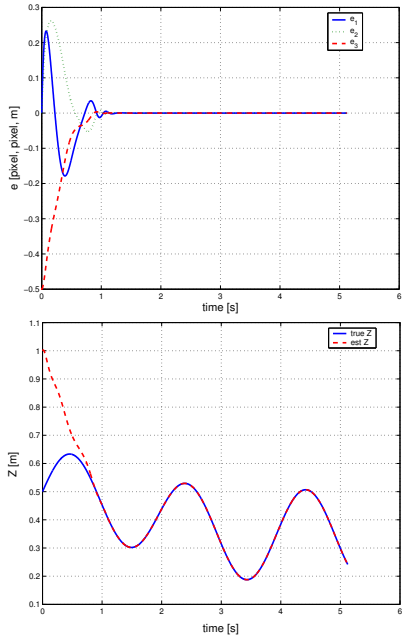

Fig. 3. Second simulation. Above: error behaviour. Below: true (solid blue line) and estimated (dashed red line) Z.

Fig. 4. Webots simulation environment with a mobile manipulator carrying a camera mounted on the end-effector.

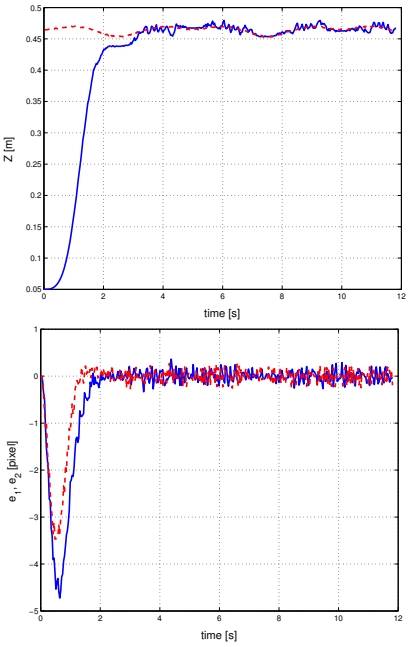

Fig. 5. Webots simulation. Above: true (dashed red line) and estimated (solid blue line) Z. Below: behavior of e_1 (solid blue line) and e_2 (dashed red line) vs time.

direction. Due to the good convergence performance of the algorithm, it is possible to integrate it into classical image-based visual servoing schemes. For example, we also included our depth estimation algorithm in a visual servoing scheme for a mobile manipulator with a camera on the end-effector [22]. Video clips of this set-up can be found at the website `www.dis.uniroma1.it/~labrob/research/NMM_IBVS.html`.

In our future developments we will try to remove the assumption of known camera motion which was needed in this work. This could be achieved, for instance, by first estimating the relative camera/target motion as in [23], and then feeding this information to our observer algorithm for scene structure identification. Finally, we are currently planning to implement the proposed approach (visual servoing + depth estimation) on a real mobile robot with a pan-tilt camera mounted on its top.

REFERENCES

[1] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, 1996.

[2] B. Espiau, F. Chaumette, and P. Rives, "A new approach to visual servoing in robotics," *IEEE Transactions on Robotics and Automation*, vol. 8, no. 3, pp. 313–326, 1992.

[3] W. J. Wilson, C. C. W. Hulls, and G. S. Bell, "Relative end-effector control using cartesian position based visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 684–696, 1996.

[4] B. Espiau, "Effect of camera calibration errors on visual servoing in robotics," in *3rd International Symposium on Experimental Robotics*, 1993, pp. 182–192.

[5] P. I. Corke and S. A. Hutchinson, "A new partitioned approach to image-based visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 4, pp. 507–515, 2001.

[6] C. Samson, B. Espiau, and M. L. Borgne, *Robot Control: The Task Function Approach*. Oxford University Press, 1991.

[7] F. Chaumette, "Potential problems of stability and convergence in image-based and position-based visual servoing," in *The Confluence of Vision and Control*, ser. LNCIS, D. Kriegman, G. Hager, and A. Morse, Eds., vol. 237. Springer Verlag, 1998, pp. 66–78.

[8] E. Malis and P. Rives, "Robustness of image-based visual servoing with respect to depth distribution errors," *Proc. of the 2003 IEEE International Conference on Robotics and Automation*, vol. 1, pp. 1056–1061, 2003.

[9] F. Chaumette, S. Boukir, P. Bouthemy, and D. Juvin, "Structure from controlled motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 5, pp. 492–504, 1996.

[10] L. Matthies, R. Szelinski, and T. Kanade, "Kalman filter-based algorithms for estimating depth from image sequences," *International Journal of Computer Vision*, vol. 3, pp. 209–236, 1989.

[11] C. E. Smith and N. P. Papanikolopoulos, "Computation of shape through controlled active exploration," *Proc. of the 1994 IEEE International Conference on Robotics and Automation*, vol. 3, pp. 2516–2521, 1994.

[12] F. Conticelli, B. Allotta, and P. K. Khosla, "Image-based visual servoing of nonholonomic mobile robots," in *Proc. of the 38th Conference on Decision and Control*, 1999, pp. 3496–3501.

[13] Y. Fang, W. E. Dixon, D. M. Dawson, and P. Chawda, "Homography-based visual servo regulation of mobile robots," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 35, no. 5, pp. 1041–1050, 2005.

[14] Y. Ma, S. Soatto, J. Košecká, and S. S. Sastry, *An Invitation to 3-D Vision*, S. Antman, J. Marsden, L. Sirovich, and S. Wiggins, Eds. Springer–Verlag New York, 2004, vol. 26.

[15] F. Chaumette, "Image moments: A general and useful set of features for visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 20, no. 4, pp. 713–723, 2004.

[16] H. Inaba, A. Yoshida, R. Abdursul, and B. K. Ghosh, "Observability of perspective dynamical systems," *Proc. of the 39th IEEE Conference on Decision and Control*, pp. 5157–5162, 2000.

[17] R. Marino and P. Tomei, *Nonlinear Control Design: Geometric, Adaptive and Robust*. Prentice Hall, London, 1995.

[18] H. K. Khalil, *Nonlinear Systems*, 2nd ed. Prentice-Hall, 1996.

[19] A. P. Aguiar and J. P. Hespanha, "Minimum-energy state estimation for systems with perspective outputs," *IEEE Transactions on Automatic Control*, vol. 51, no. 2, pp. 226–241, 2006.

[20] W. E. Dixon, Y. Fang, D. M. Dawson, and T. J. Flynn, "Range identification for perspective vision systems," *IEEE Transactions on Automatic Control*, vol. 48, no. 12, pp. 2232–2238, 2003.

[21] http://www.cyberbotics.com.

[22] A. De Luca, G. Oriolo, and P. Robuffo Giordano, "Image-based visual servoing schemes for nonholonomic mobile manipulators," *to appear in Robotica*.

[23] S. Soatto, R. Frezza, and P. Perona, "Motion estimation via dynamic vision," *IEEE Transactions on Automatic Control*, vol. 41, no. 3, pp. 393–413, 1996.