Visual servoing over unknown, unstructured, large-scale scenes

G. Silveira, Ezio Malis, P. Rives

Visual Servoing over Unknown, Unstructured, Large-scale Scenes

Geraldo Silveira [*,†], Ezio Malis [*], Patrick Rives [*]

[*] INRIA Sophia-Antipolis – Project ICARE
2004 Route des Lucioles, BP 93
06902 Sophia-Antipolis Cedex, France
FirstName.LastName@sophia.inria.fr

[†] CenPRA Research Center – DRVC Division
Rod. Dom Pedro I, km 143,6, Amarais
CEP 13069-901, Campinas/SP, Brazil
Geraldo.Silveira@cenpra.gov.br

***Abstract*— This work proposes a new vision-based framework to control a robot within model-free large-scale scenes, where the desired pose has never been attained beforehand. Thus, the desired image is not available. It is important to remark that existing visual servoing techniques cannot be applied in this context. The rigid, unknown scene (i.e. the metric model is also not available) is represented as a collection of planar regions, which may leave the field-of-view continuously as the robot moves toward its distant goal. Hence, a novel approach to detect new planes that enter the field-of-view, which is robust to large camera calibration errors, is then deployed here. In fact, it is well-known that representing the scene as composed by planes, the estimation processes are improved in terms of accuracy, stability, and rate of convergence. This Extended 3D vision-based control technique is also based on an efficient second-order method for plane-based tracking and pose reconstruction. The framework is validated by using simulated data with artificially created scenes as well as with real images, and accurate navigation tasks are shown.**

I. Introduction

The use of visual information to control dynamic systems in closed loop has been widely deployed during the last decade. Indeed, several vision-based controllers have been proposed by the robotics community. In any case however, the control objective of visual servoing systems is to drive the robot from an initial pose to a reference (desired) pose, by using appropriate information extracted from image data. Generally, those systems are designed such that the initial pose is considered to be in a neighborhood of the desired one. The present work is different from the previous ones in many aspects. First of all, it is focused on the control of a single camera over large-scale scenes where the desired pose has never been attained by the robot before (see Fig. 1). Thus, the desired image to be acquired is not available. In addition, it is dealt here with unknown scenes, i.e. the metric model of the scene is also not available a priori. Hence, it is not possible to render the desired image. Nevertheless, a model-free pose-based visual servoing can be envisaged in this case. There exist various visual servoing strategies where the control error is defined in the Cartesian space. As for the case of model-based approaches, the reader is referred to e.g. [1]. Concerning the model-free schemes, for example the methods proposed in [2], [3] and [4], the authors use the current and the desired images in order to recover the epipolar geometry that relates those images. Indeed, the translation and rotation motions can be derived from such information. However, besides the need of the desired image, the strategy proposed in [2] may not be the most adequate one when the scene is planar since the required essential matrix is degenerate. In contrast, the approach devised here copes with planar scenes indistinguishably from other scenes. With respect to [3], also besides the need of the desired image, the authors assume that sufficient information is available in the images so that the homography at the infinity can be recovered, which is not a trivial issue. The visual servoing approach proposed here is more related to the work accomplished in [4] and [5], where an unknown, unstructured scene is considered as well. However, the former work requires the desired image and, albeit in the latter one there is no need of the desired image, it also relies on a non-planar scene. In fact, it is well-known that representing the scene as composed by planes, the estimation processes are improved in terms of accuracy, stability, and rate of convergence [6]. In this case, the number of planes to be considered in the entire scene can be viewed as a trade-off between accuracy and computational load. Hence, the unknown scene is represented in this work as a collection of planar regions, which may leave the field-of-view continuously as the robot moves toward its distant goal. Thus, complex strategies to deal with the visibility constraints are not required at all. In fact, the unknown desired image may not have anything in common with the initial one, but the desired Cartesian path may still be followed accordingly. The proposed Extended 3D (E-3D) vision-based control framework relies mainly on two key techniques: on a novel approach to detect new planes in the image as the robot evolves, so that the known planes may leave the field-of-view; and on an efficient second-order method for plane-based tracking and pose reconstruction. In addition, the proposed approach is based on a hybrid strategy that combines image features and image templates, so that the sensitivity of pose-based techniques with respect to image measurement errors is drastically minimized. The proposed approach is also different from other vision-based SLAM techniques, whose majority of works do not control the robot. For example, the scheme conceived in [7], besides not controlling the camera, it assumes that small image patches are observations of planar regions, and whose normal vector is initially assigned to a "best guess" orientation. With respect to the plane detection algorithm used here, besides its robustness against large camera calibration errors, a closed-form solution to determine the normal vector is presented. In addition, the necessary and sufficient conditions to allow for identifying new planes that enter the image are also provided. Results for navigation tasks are shown and very small Cartesian errors were obtained. Also, experimental results in different scenarios demonstrate the robustness characteristics of the method.

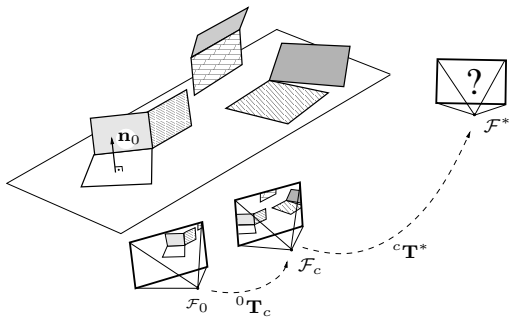

Fig. 1. The objective of the approach: to perform a vision-based navigation task over an extensive scene, considered as piecewise planar, where neither the desired image (corresponding to the desired pose) nor the scene model are available.

The remainder of this work is arranged as follows. Section II reviews some basic theoretical aspects, as well as it introduces the proposed long-term navigation framework. The vision aspects involved in the strategy is presented in the Section III, while the control aspects are developed in Section IV. The results are then shown and discussed in the Section V. Finally, the conclusions are presented in the Section VI, and some references are given for further details.

II. Modeling

Let $\mathcal{F}$ be the camera frame whose origin $\mathcal{O}$ coincides with its center of projection $\mathcal{C}$. Suppose that $\mathcal{F}$ is displaced with respect to another frame $\mathcal{F}'$ (which is not necessarily the initial frame $\mathcal{F}_0$, nor the desired frame to be aligned $\mathcal{F}^*$) in the Euclidean space by $\mathbf{R} \in SO(3)$ and $\mathbf{t} = [t_x, t_y, t_z]^T \in \mathbb{R}^3$, respectively the rotation matrix and the translation vector. Consider the angle-axis representation of the rotation matrix. By using the matrix exponential, $\mathbf{R} = \exp([\mathbf{r}]_\times)$, where $\mathbf{r} = \mathbf{u}\theta$ is the vector containing the angle of rotation $\theta \in [0, 2\pi)$, and the axis of rotation $\mathbf{u} \in \mathbb{R}^3 : \|\mathbf{u}\| = 1$. The notation $[\mathbf{r}]_\times$ represents the skew symmetric matrix associated to vector $\mathbf{r}$. Hence, the camera pose can be defined with respect to frame $\mathcal{F}'$ by a (6×1)-vector $\boldsymbol{\xi} = \left[\mathbf{t}^T, \mathbf{r}^T\right]^T$, containing the global coordinates of an open subset of $\mathbb{R}^3 \times SO(3)$.

A. Camera Model

Consider the pinhole camera model. In this case, a 3D point with homogeneous coordinates $\mathbf{P}_i = \left[X_i, Y_i, Z_i, 1\right]^T$ defined with respect to frame $\mathcal{F}$, $i = 1, 2, \ldots, n$, is projected onto the image space $\boldsymbol{\mathcal{I}} \subset \mathbb{R}^2$ as a point with pixels homogeneous coordinates $\mathbf{p}_i \in \mathbb{P}^2$ through

$$\mathbf{p}_i = [u_i, v_i, 1]^T \propto \mathbf{K} \left[\begin{array}{cc} \mathbf{I}_3 & \mathbf{0} \end{array}\right] \mathbf{P}_i, \tag{1}$$

where $\mathbf{K} \in \mathbb{R}^{3\times3}$ is an upper triangular matrix that gathers the camera intrinsic parameters

$$\mathbf{K} = \begin{bmatrix} \alpha_u & s & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

with focal lengths $\alpha_u, \alpha_v > 0$ in pixel dimensions, principal point $\mathbf{p}_0 = [u_0, v_0, 1]^T$ in pixels, and skew s. Correspondingly, the same point $\mathbf{P}_i \in \mathbb{P}^3$ is projected onto the image space $\boldsymbol{\mathcal{I}}' \subset \mathbb{R}^2$ associated to $\mathcal{F}'$ as

$$\mathbf{p}'_i = [u'_i, v'_i, 1]^T \propto \mathbf{K} \left[\begin{array}{cc} \mathbf{R} & \mathbf{t} \end{array}\right] \mathbf{P}_i. \tag{3}$$

Then, from the general rigid-body equation of motion along with (1) and (3), it is possible to obtain the fundamental relation that links the projection of $\mathbf{P}_i$ onto both images:

$$\mathbf{p}'_i \propto \mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1}\mathbf{p}_i + \frac{1}{Z_i}\,\mathbf{K}\,\mathbf{t}. \tag{4}$$

B. Plane-based Two-view Geometry

Consider the normal vector description of a plane $\boldsymbol{\pi} = \left[\mathbf{n}^T, -d\right]^T \in \mathbb{R}^4 : \|\mathbf{n}\| = 1, d > 0$. Let $\boldsymbol{\pi}$ (resp. $\boldsymbol{\pi}'$) be defined with respect to frame $\mathcal{F}$ (resp. $\mathcal{F}'$). If a 3D point $\mathbf{P}_i$ lies on such planar surface then

$$\mathbf{n}^T\,\mathbf{P}_i = \mathbf{n}^T Z_i\,\mathbf{K}^{-1}\mathbf{p}_i = d, \tag{5}$$

and hence

$$\frac{1}{Z_i} = \frac{\mathbf{n}^T\mathbf{K}^{-1}\mathbf{p}_i}{d}. \tag{6}$$

By plugging (6) into (4), a projective mapping $\mathbf{G} \in PL(2) : \mathbb{P}^2 \mapsto \mathbb{P}^2$ (also referred to as the projective homography) defined up to a non-zero scale factor is achieved:

$$\mathbf{p}'_i \propto \mathbf{G}\,\mathbf{p}_i. \tag{7}$$

In addition, it can be noticed that $\mathbf{G}$ encompasses an Euclidean homography $\mathbf{H} \in \mathbb{R}^{3\times3}$ for the case of internally calibrated cameras. That is, for normalized homogeneous coordinates $\mathbf{m}_i = \mathbf{K}^{-1}\,\mathbf{p}_i$, Eq. (7) becomes

$$\mathbf{m}'_i \propto \underbrace{\mathbf{R} + d^{-1}\,\mathbf{t}\,\mathbf{n}^T}_{\mathbf{H}}\,\mathbf{m}_i. \tag{8}$$

As a remark, it is well-known that the same expressions are obtained, independently if the object is planar or not, if the camera undergoes a pure rotation motion (i.e. $\forall\,\mathbf{R} \in SO(3)$ but $\mathbf{t} = \mathbf{0}$) since depth information is completely lost.

C. Navigation Formulation

Visual servoing systems are usually designed such that the desired frame to be attained $\mathcal{F}^*$ is aligned with the absolute frame $\mathcal{F}_w$. Indeed, the aim is to promote adequate motions such that $\mathcal{F} \to \mathcal{F}^*$. On effect, this leads then to be $\boldsymbol{\xi}^* = \mathbf{0}$ and the control objective to drive $\boldsymbol{\xi} \to \mathbf{0}$ as $t \to \infty$. However, since the purpose in this work is to navigate the robotic platform (see Fig. 1), the absolute frame is then set to coincide with the initial frame, i.e. $\mathcal{F}_0 = \mathcal{F}_w$ and thus $\boldsymbol{\xi}_0 = \mathbf{0}$. Hence, the current and desired poses are here defined w.r.t. $\mathcal{F}_0$, what leads to a desired $\boldsymbol{\xi}^* = \left[\mathbf{t}^{*T}, \mathbf{r}^{*T}\right]^T$ and the control objective to be

$$\boldsymbol{\xi} \to \boldsymbol{\xi}^* \text{ as } t \to \infty. \tag{9}$$

In fact, after the proper specification of the navigation task, a change of coordinate system back to the usual one can obviously be made. Also, as already stated, the proposed framework is based on the representation of the scene as a

collection of planar regions. It is well-known that such constraint allows for implementing much more stable and accurate pose reconstruction algorithms [6]. Indeed, the core of the proposed navigation framework is basically given as follows. Provided $\mathbf{K}$ and a set of planes $\{\boldsymbol{\pi}\}$, the control objective (9) can be perfectly achieved by regulating a Cartesian-based error function (track a Cartesian-based path) constructed from images:

$$\mathbf{e} = \mathbf{e}\big(\boldsymbol{\mathcal{I}}, \{\boldsymbol{\pi}\}, \mathbf{K}, \boldsymbol{\xi}^*, t\big), \quad \forall t \in [0, T]. \tag{10}$$

The control aspects are further discussed in Section IV. From such definition of the error function, let us present an overview of the proposed method to perform vision-based control tasks over large-scale unknown scenes, for some sufficiently small $\epsilon > 0$:

Algorithm 1. The E-3D visual servoing framework.

1: define plane $\boldsymbol{\pi}_0$ in the first image $\boldsymbol{\mathcal{I}}_0$
2: **repeat**
3: apply control law
4: track known planes and recover pose
5: **if** conditions in the Proposition 3.1 are verified **then**
6: by using $\{\mathbf{K}, \widehat{\mathbf{R}}, \widehat{\mathbf{t}}\}$, identify new planes that enter $\boldsymbol{\mathcal{I}}$
7: **end if**
8: **until** $\|\mathbf{e}\| < \epsilon$

The procedures stated from line 4 to 6 of the Algorithm 1 are further detailed in the next section.

III. Planes Detection and Tracking

A. Pose Reconstruction from Multiple Planes

This subsection intends to present how multiple planes are tracked in the image space, as well as how the camera pose is recovered. Both tasks are treated as belonging to a single block since the rigidity of the scene is taken into consideration to achieve superior tracking performance, and to provide more accurate pose estimates. However, due to paper length restrictions, only an overview of the scheme will be described here. The reader is referred to [8] for more details.

Consider that at least one planar object is observed in the image, and that a reference template corresponding to a given frame $\mathcal{F}'$ has been selected. How to cluster those planar regions in the image will be described in the next subsections. Also, in order to perform the mapping between the projective and the Euclidean spaces, the camera is supposed to be calibrated. By using such efficient second-order minimization technique, every template is then optimally tracked in the image space. It is an efficient algorithm since only first image derivatives are used, and the Hessians are not explicitly computed. Indeed, its two main advantages are the high convergence rate and the avoidance of local minima. Then, after finding the optimal homography $\mathbf{H}_j$ (i.e., the solution of the optimization problem), its decomposition into $\mathbf{R}_j$ and $\mathbf{t}_j$ for every template is performed. The rigidity constraint of the scene is thus imposed a posteriori. That is, the relative pose between two frames $\mathcal{F}'$ and $\mathcal{F}$ must be the same for all planes to yield the pose estimate $\{\widehat{\mathbf{R}}, \widehat{\mathbf{t}}\}$.

B. Detection of New Planes

Since the known planes will eventually get out of the image during a long-term navigation, one must identify new planes that enter the field-of-view (and track them optimally over the sequence). In this subsection, the method used to detect planar regions in a *pair* of images is presented. The interest in finding planar regions in images is not new, and a number of different approaches have been proposed by the computer vision community. However, the majority of the approaches in the literature relies on a preliminary step of 3D scene reconstruction (i.e. the depth map is required, as in e.g. [9]). Those methods are in general too time-consuming, or demand several images to converge, or they rely on scene assumptions (e.g. structured scenes [10], perpendicularity assumptions), or even on heuristic searches. In order to circumvent those constraints, the used algorithm is based on an efficient voting procedure directly from the solution of a linear system, which is derived from the following. Equation (4) along with (6) allow for rewriting the fundamental equation that links the projection of the same 3D point onto $\boldsymbol{\mathcal{I}}$ and $\boldsymbol{\mathcal{I}}'$ as

$$\mathbf{p}'_i \propto \mathbf{G}_\infty \mathbf{p}_i + \mathbf{e}_p \, \bar{\mathbf{n}}^T \mathbf{K}^{-1} \mathbf{p}_i, \tag{11}$$

where $\mathbf{G}_\infty = \mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1}$ is the homography at the infinity, $\mathbf{e}_p = \mathbf{K}\,\mathbf{t}$ is the epipole in the second view, and $\bar{\mathbf{n}} = \mathbf{n}/d$ is the normal vector scaled by the distance to $\mathcal{F}$. Then, triplet of corresponding interest points (e.g. Harris) are managed in order to form linear systems whose solutions are used in a progressive Hough-like transform, and in order to respect the real-time constraints. A template is formed by means of the convex hull of the clustered points. In addition, it is well-known that the Hough Transform (and its variants) is one of the most important robust techniques in computer vision [11]. As it will be shown, even if the set of camera parameters $\{\mathbf{K}, \mathbf{R}, \mathbf{t}\}$ are miscalibrated, i.e. only an estimated set $\{\widehat{\mathbf{K}}, \widehat{\mathbf{R}}, \widehat{\mathbf{t}}\}$ is provided, and even if there also exist mismatched corresponding points (outliers), it is still possible to cluster planar regions in the image (see next subsection for the necessary and sufficient conditions). This robustness property is an attractive characteristic of the approach since it is able to tolerate large errors in its inputs. Furthermore, besides the explicit clustering of planar regions, there is no "best guess" initialization regarding the normal vector of the plane (e.g. [7], where the authors assume that small image patches are observations of planar regions and whose vector, after such initialization, is refined based on a gradient descent technique). In the next subsection, a closed-form solution to determine the equations of the new clustered planes will be presented.

C. Determination of the Equations of the New Planes

To this point, a set of new planes $\{\boldsymbol{\pi}_j\}$ (resp. $\{\boldsymbol{\pi}'_j\}$) are segmented in the image $\boldsymbol{\mathcal{I}}$ (resp. $\boldsymbol{\mathcal{I}}'$), and their corresponding homographies $\{\mathbf{G}_j\}$ are found robustly and optimally. In addition, the relative pose between $\mathcal{F}'$ and $\mathcal{F}$ is also provided, which must be the same $\forall \boldsymbol{\pi}$ projected onto $\boldsymbol{\mathcal{I}}$ if the scene is rigid. However, in order to include them in the pose reconstruction algorithm, it is needed to determine each $\mathbf{n}_j$ in the 3D space. On effect, manipulating Eqs. (7) and (8) $\mathbf{H} = \alpha\,\mathbf{K}^{-1}\,\mathbf{G}\,\mathbf{K}$, the following expression is obtained:

$$\mathbf{t}_{d_j}\mathbf{n}_j^T = \alpha_j\,\mathbf{K}^{-1}\,\mathbf{G}_j\,\mathbf{K} - \mathbf{R}. \tag{12}$$

Multiplying both members of (12) by the transpose of the reconstructed scaled translation vector $\mathbf{t}_{d_j}^T = \mathbf{t}^T/d_j$, a closed-form solution for determining the normal vector w.r.t. $\mathcal{F}$ of each segmented $\boldsymbol{\pi}_j$ is achieved:

$$\mathbf{n}_j^T = \left(\mathbf{t}_{d_j}^T\,\mathbf{t}_{d_j}\right)^{-1}\mathbf{t}_{d_j}^T\left(\alpha_j\,\mathbf{K}^{-1}\,\mathbf{G}_j\,\mathbf{K} - \mathbf{R}\right). \tag{13}$$

Given that $\text{svd}(\mathbf{H}) = [\sigma_1, \sigma_2, \sigma_3]^T$ are the singular values of $\mathbf{H}$ in decreasing order, $\sigma_1 \geq \sigma_2 \geq \sigma_3 > 0$, and that such homography can be normalized by the median singular value [12], it is possible to use the facts that $x = \text{sgn}(x)\,|x|$, $\forall x \in \mathbb{R}$, $\det(\mathbf{H}) = \prod_{i=1}^{3}\lambda_i(\mathbf{H})$, and that σ_i are the square-roots of $\lambda(\mathbf{H}^T\mathbf{H})$, so that the scale factor $\alpha_j \in \mathbb{R}$ is given as

$$\alpha_j = \frac{\text{sgn}(\det(\mathbf{H}_j))}{\sigma_2(\mathbf{H}_j)}, \tag{14}$$

where $\text{sgn}(\cdot)$ denotes the signum function.

Proposition 3.1 (Normal Vector Determination): The necessary and sufficient conditions for the normal vector determination (13) are such that:

- $\mathbf{t} \neq \mathbf{0}$ so that $(\mathbf{t}_{d_j}^T\mathbf{t}_{d_j})^{-1} = d_j^2\,(\mathbf{t}^T\mathbf{t})^{-1}$ exists. Obviously, $d_j > 0$, $\forall j$, so that all the planes are in front of the camera;
- $|\det(\mathbf{G})| > 0$, so that the plane is not in a degenerate configuration (i.e. projected as a line), and $\alpha \neq 0$.

The last condition of the Proposition 3.1 is due to $\det(\mathbf{H}/\alpha) = \frac{1}{\det(\mathbf{K})}\det(\mathbf{G})\det(\mathbf{K}) = \det(\mathbf{G})$, if the second condition holds. It is also important to remark that the last condition can then be used as a measure of degeneracy, and that explains why the projective homography $\mathbf{G}$ was not parameterized here as a member of the $SL(3)$ (the Special Linear group). The $SL(3)$ is the group of (3×3) matrices that has the determinant equal to 1.

IV. Control Aspects

Let the robot be controlled in velocity $\mathbf{v} = \left[\boldsymbol{\upsilon}^T, \boldsymbol{\omega}^T\right]^T \in \mathbb{R}^q$, respectively the linear and angular velocities, with $q \leq 6$ dofs. As already stated, the rigidity assumption of the scene is imposed so that the relative displacement between $\mathcal{F}'$ and $\mathcal{F}$ are the same for all tracked planes, which is performed directly in the Euclidean space. In addition, since a known plane can leave the field-of-view *without* destabilizing the system (since it is possible to detect and reconstruct new planes), a pose-based visual servoing technique is the appropriate choice for the task. Hence, the error vector is constructed from the knowledge of the current and desired poses (extracted from ${}^0\mathbf{T}_c$ and ${}^0\mathbf{T}^*$, respectively), and then expressed both with respect to $\mathcal{F}^*$ (to conform to the usual absolute frame). Thus, the control error (10) is here defined as

$$\mathbf{e} = \left[\mathbf{e}_\upsilon^T, \mathbf{e}_\omega^T\right]^T = \left[{}^*\mathbf{t}_c^T, {}^*\mathbf{r}_c^T\right]^T = \left[\mathbf{t}^T, \mathbf{u}^T\theta\right]^T \quad \in \mathbb{R}^q, \tag{15}$$

denoting the error in translation and in the rotation respectively. Considering a positioning task, the derivative of (15) yields

$$\dot{\mathbf{e}} = \mathbf{L}(\boldsymbol{\xi})\,\mathbf{W}(\boldsymbol{\xi})\,\mathbf{v}, \tag{16}$$

with the interaction matrix

$$\mathbf{L}(\boldsymbol{\xi}) = \begin{bmatrix} \mathbf{I}_3 & -[\mathbf{e}_\upsilon]_\times \\ \mathbf{0} & \mathbf{L}_\omega \end{bmatrix}. \tag{17}$$

The $\mathbf{L}_\omega$ is the interaction matrix related to the parametrization of the rotation: $\frac{d(\mathbf{u}\theta)}{dt} = \mathbf{L}_\omega\,\boldsymbol{\omega}$. By using the Rodrigues' formula for expressing the rotation matrix, it can be shown that

$$\mathbf{L}_\omega = \mathbf{I}_3 - \frac{\theta}{2}[\mathbf{u}]_\times + \left(1 - \frac{\text{sinc}(\theta)}{\text{sinc}^2(\frac{\theta}{2})}\right)[\mathbf{u}]_\times^2, \tag{18}$$

where the function $\text{sinc}(\cdot)$ is the so-called sine cardinal or sampling function. Also, it can be noticed that

$$\det(\mathbf{L}_\omega) = \text{sinc}^{-2}(\theta/2), \tag{19}$$

whose singularities are for $\theta = 2k\pi$, $\forall k \in \mathbb{Z}_+$, and hence the largest possible domain: $\theta \in [0, 2\pi)$. In addition, the upper-block triangular matrix $\mathbf{W}(\boldsymbol{\xi}) \in \mathbb{R}^{6\times 6}$ in (16) represents the transformation

$$\mathbf{W}(\boldsymbol{\xi}) = \begin{bmatrix} \mathbf{I}_3 & [{}^*\mathbf{t}_c]_\times \\ \mathbf{0} & \mathbf{I}_3 \end{bmatrix}\begin{bmatrix} {}^*\mathbf{R}_c & \mathbf{0} \\ \mathbf{0} & {}^*\mathbf{R}_c \end{bmatrix} = \begin{bmatrix} {}^*\mathbf{R}_c & [{}^*\mathbf{t}_c]_\times{}^*\mathbf{R}_c \\ \mathbf{0} & {}^*\mathbf{R}_c \end{bmatrix}, \tag{20}$$

since the control input $\mathbf{v}$ is defined in camera frame $\mathcal{F}_c$ and the error is expressed in $\mathcal{F}^*$. With respect to the control law, if it is imposed an exponential decrease for the error

$$\dot{\mathbf{e}} = -\lambda_\mathbf{v}\,\mathbf{e}, \quad \lambda_\mathbf{v} > 0, \tag{21}$$

then its substitution into (16) by using (15) permits to achieve

$$\mathbf{v} = -\lambda_\mathbf{v}\,\mathbf{W}^{-1}(\boldsymbol{\xi})\,\mathbf{L}^{-1}(\boldsymbol{\xi})\,\mathbf{e} \tag{22}$$

$$= -\lambda_\mathbf{v}\begin{bmatrix} {}^c\mathbf{R}^* & -{}^c\mathbf{R}^*[{}^*\mathbf{t}_c]_\times \\ \mathbf{0} & {}^c\mathbf{R}^* \end{bmatrix}\begin{bmatrix} \mathbf{I}_3 & [{}^*\mathbf{t}_c]_\times\mathbf{L}_\omega^{-1} \\ \mathbf{0} & \mathbf{L}_\omega^{-1} \end{bmatrix}\mathbf{e}. \tag{23}$$

Such an expression can be further simplified. Given that $[\mathbf{u}]_\times^k\,\mathbf{u} = \mathbf{0}$, $\forall k > 0$, it yields $\mathbf{L}_\omega^{-1}\mathbf{e}_\omega = \mathbf{e}_\omega$, $\forall \mathbf{e}_\omega$, with

$$\mathbf{L}_\omega^{-1} = \mathbf{I}_3 + \frac{\theta}{2}\text{sinc}^2\left(\frac{\theta}{2}\right)[\mathbf{u}]_\times + (1 - \text{sinc}(\theta))\,[\mathbf{u}]_\times^2, \tag{24}$$

and the final control law is achieved as

$$\mathbf{v} = -\lambda_\mathbf{v}\begin{bmatrix} {}^c\mathbf{R}^* & \mathbf{0} \\ \mathbf{0} & {}^c\mathbf{R}^* \end{bmatrix}\mathbf{e}. \tag{25}$$

As a remark, the control law (25), besides the full decoupling of translational and rotational motions (it has a block diagonal matrix), it promotes a straight-line path linking $\overrightarrow{\mathcal{O}\mathcal{O}^*}$ in Cartesian space since $\dot{\mathbf{t}} = {}^*\mathbf{R}_c\,\boldsymbol{\upsilon} = -\lambda_\mathbf{v}{}^*\mathbf{R}_c{}^c\mathbf{R}^*\,\mathbf{t} = -\lambda_\mathbf{v}\mathbf{t}$.

V. Results

In this section, the results obtained with the E-3D visual servoing technique are shown and discussed. Concerning the image features (used by the plane detection algorithm), the Harris detector was applied in this work. Then, all the detected templates (corresponding to the convex hull of the clustered points) are used by the pose recovery technique, which also tracks them simultaneously during navigation. With respect to the method for detecting new planes, various pairs of images were used for testing purposes and some results can be seen in Fig. 2, which agree with the expectations: detected planes are actual planes. Due to real-time requirements, only a portion of the entire plane is clustered and tracked. Nevertheless, a region growing process based on the plane equations could be used to partition the entire plane. Furthermore, since the true camera calibration parameters (both intrinsic and extrinsic ones) were not available, it was used for *all* tested pairs of images: $\alpha_u = \alpha_v = 500$ pixels with principal point as the middle of the image, as well as $\mathbf{R} = \mathbf{I}_3$ and $\mathbf{t} = [-0.1, 0, -1]^T$m for the rotation and translation motions, respectively. Albeit these parameters are not the true ones, the actual planes were detected. Therefore, the robustness properties of the approach was thus also verified.

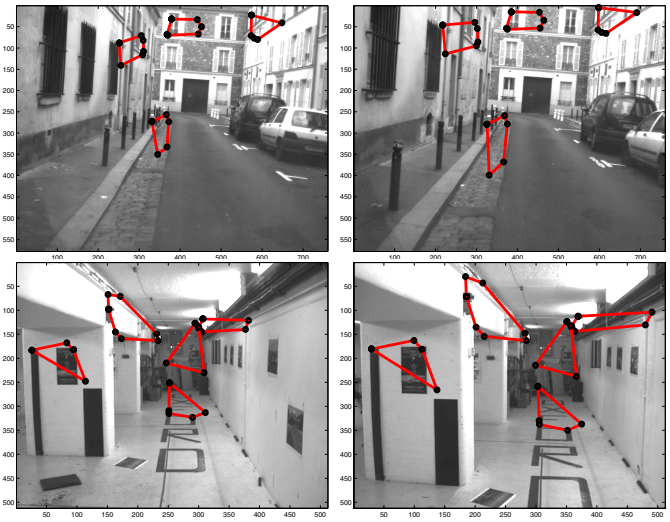

Fig. 2. Some results obtained by using the plane detection algorithm, where the detected planar regions are surrounded by red lines. Due to real-time requirements, only a portion of the planes is clustered and tracked.

In order to have a ground truth for the proposed vision-based control technique, a textured scene was constructed: its base is composed of four planes disposed in pyramidal form, but cut by another plane on its top. Onto each one of the five plans, a different texture was applied (see Fig. 3). With respect to the navigation task, the control gain was set to $\lambda_{\mathbf{v}} = 0.5$ and a closed, arbitrary Cartesian trajectory was specified and afterwards subdivided into 10 elementary positioning tasks. It is shown in Fig. 4 the obtained images at the convergence for some of tasks, where the detected planar regions for recovering the pose are superposed. A remark is valuable here: one may notice that the known plane (shown in the first image) leaves the field-of-view but the entire navigation task could be

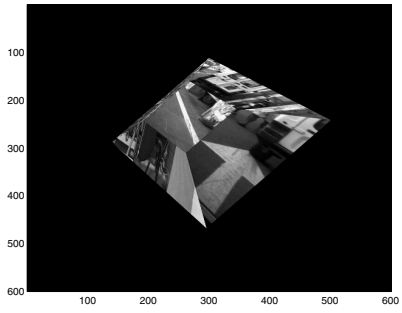

Fig. 3. Image of the artificially created, textured, piecewise planar scene.

completed accordingly, since new planes have been identified. In addition, when such plane reenters the image it is newly determined. An elementary task is said to be completed here when the translational error drops below a certain precision (it was set when $\|\mathbf{e}_v\| < 0.1$mm). Notice that in this case where the desired image is not available, existing model-free visual servoing techniques cannot be applied. As for the evolution of the task, both the exponential decrease of the norm of the control error for some of the specified tasks, as well as the computed control signals can also be seen in Fig. 4. The true errors obtained in the pose recovery process along the entire task are depicted in Fig. 5, since the real ground truth is known. One can observe that when the image loses resolution (i.e. the camera moves away from the object), the precision of the reconstruction also decreases and vice-versa. Nevertheless, one important result comes from performing the closed-loop trajectory (which has a displacement of ≈3.3m): errors smaller than 0.1mm in translation and than 0.01° in rotation were obtained after the camera comes back to the same pose at the beginning (compare first and last images of the Fig. 4). Such result demonstrates the precision achieved by the framework.

Another important result from the approach is that the scene can be reconstructed in 3D space (up to a scale factor). Such result is shown in Fig. 6 for different views of the scene. It pictures that the E-3D visual servoing approach can be also used as a Plane-based Structure from Controlled Motion technique, improving the stability, the accuracy and the rate of convergence of Structure From Motion methods.

VI. Conclusions

This work proposes a new visual servoing approach for large-scale scenes, where the desired image to be acquired (corresponding to the desired pose) is not available beforehand. In addition, it was dealt here with unknown scenes, which are represented as a collection of planar regions. By taking that into consideration, an accurate real-time pose reconstruction is deployed. As the robot evolves, since the known planes will eventually get out of the field-of-view, new planes in the scene are detected and then used by the pose recovery algorithm. Hence, distant goals may be specified. Navigation tasks were performed and only negligible Cartesian errors were obtained. In addition, it is shown that the proposed vision-based control scheme can be used as a Plane-based Structure from Controlled Motion technique as well.

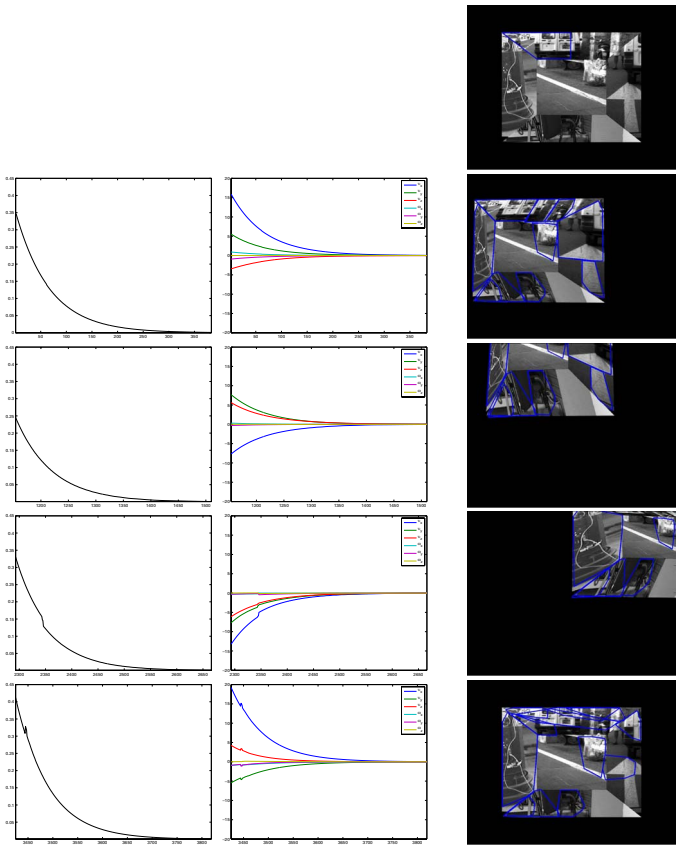

Fig. 4. A plane is initialized in the first image. For each elementary task shown, the norm of the error and the control signals (in [cm/s] and [deg/s]) vs. number of iterations are drawn. At the right, the corresponding obtained images at the convergence, which are superposed by the detected planar regions (in blue), are shown. Observe that a plane leaves the field-of-view (3th and 4th images) but when it reenters it is newly identified (5th image).

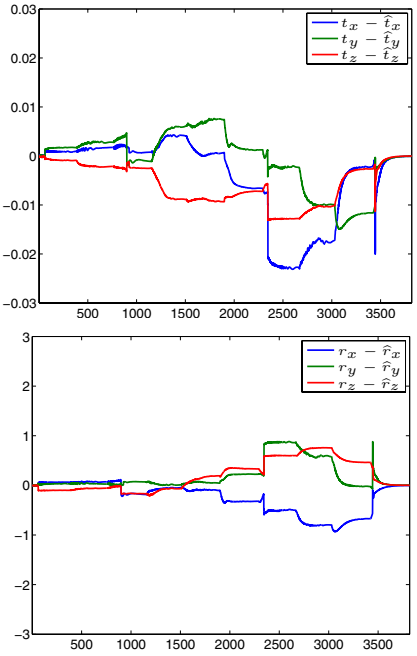

Fig. 5. Errors in the pose recovery (position [m] and attitude [deg], respectively) vs. number of iterations along the entire navigation task.

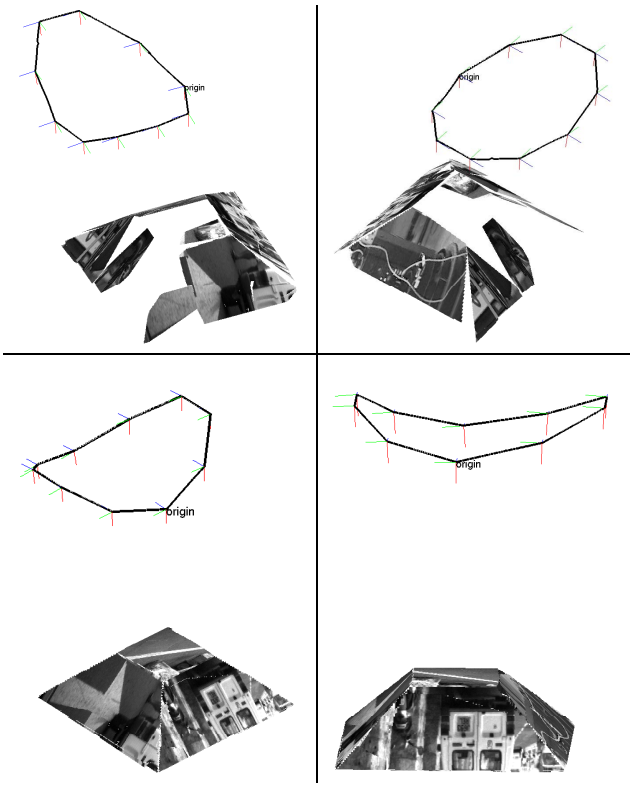

Fig. 6. The desired poses, the performed trajectory, and the 3D reconstructed scene as seen from different viewpoints (first row: the scene with the used planes only and, at the bottom, the scene after performing a region growing).

ACKNOWLEDGMENTS

This work is also partially supported by the CAPES Foundation under grant no. 1886/03-7, and by the international agreement FAPESP-INRIA under grant no. 04/13467-5.

REFERENCES

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

[2] R. Basri, E. Rivlin, and I. Shimshoni, "Visual homing: surfing on the epipoles," *Int. Journal of Comp. Vision*, vol. 33, no. 2, pp. 22–39, 1999.

[3] C. J. Taylor and J. P. Ostrowski, "Robust vision-based pose control," in *Proc. IEEE Int. Conf. on Robot. and Automat.*, 2000, pp. 2734–2740.

[4] E. Malis and F. Chaumette, "Theoretical improvements in the stability analysis of a new class of model-free visual servoing methods," *IEEE Trans. on Robotics and Automation*, vol. 18, no. 2, pp. 176–186, 2002.

[5] P. Rives, "Visual servoing based on epipolar geometry," in *Proc. of the IEEE/RSJ Int. Conf. on Intell. Robots and Systems*, 2000, pp. 602–607.

[6] R. Szeliski and P. H. S. Torr, "Geometrically constrained structure from motion: points on planes," in *Proc. of the Eur. Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, 1998, pp. 171 – 186.

[7] N. D. Molton, A. J. Davison, and I. D. Reid, "Locally planar patch features for real-time structure from motion," in *Proc. BMVC*, 2004.

[8] S. Benhimane and E. Malis, "Real-time image-based tracking of planes using Efficient Second-order Minimization," in *IEEE/RSJ International Conference on Intelligent Robots Systems*, Japan, October 2004.

[9] K. Okada *et al.*, "Plane segment finder: Algorithm, implementation and applications," in *Proc. of the IEEE ICRA*, 2001, pp. 2120–2125.

[10] C. Baillard and A. Zisserman, "Automatic reconstruction of piecewise planar models from multiple views," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1999, pp. 559–565.

[11] C. V. Stewart, "Robust parameter estimation in computer vision," *SIAM Rev.*, vol. 41, pp. 513–537, 1999.

[12] Z. Zhang and A. R. Hanson, "Scaled Euclidean 3D reconstruction based on externally uncalibrated cameras," in *IEEE Symposium on Computer Vision*, 1995, pp. 37–42.