Visual Servoing in the Large

Daniele Fontanelli
Antonio Danesi
Felipe A. W. Belo
Paolo Salaris
Antonio Bicchi

Interdepartment Research Center 'Enrico Piaggio'
University of Pisa
via Diotisalvi 2
56100 Pisa, Italy
daniele.fontanelli@esa.int

Abstract

In this paper we consider the problem of maneuvering an autonomous robot in complex unknown environments using vision. The goal is to accurately servo a wheeled vehicle to a desired posture using only feedback from an on-board camera, taking into account the non-holonomic nature of the vehicle kinematics and the limited field-of-view of the camera. With respect to existing visual servoing schemes, which achieve similar goals locally (i.e. when the desired and actual camera views are sufficiently similar), we propose a method to visually navigate the robot through an extended visual map before eventually reaching the desired goal. The map comprises a set of images, previously stored in an exploratory phase, that convey both topological and metric information regarding the connectivity through feasible robot paths and the geometry of the environment, respectively. Experimental results on a laboratory setup are reported showing the practicality of the proposed approach.

KEY WORDS—cooperation, mobile robots, simultaneous localization and mapping, visual servoing

1. Introduction

One of the main obstacles that still hinder penetration of mobile robots into wide consumer markets is the unavailability of powerful, versatile and cheap sensing. Vision technology is potentially a clear winner as far as the ratio of information provided versus cost is considered. Cameras of acceptable accuracy are currently sold at a price which is one to two orders of magnitude less than, for example, laser scanners. As a consequence, much attention is being devoted to solving the non-trivial problem of using visual information for controlling and localizing robots in a visually mapped environment.

This paper deals with the problem of using off-the-shelf cameras fixed on inexpensive mobile platforms to enable navigation and accurate control to goal configurations in space based on visual maps of the environment, which can be contextually built in the process. To this purpose, powerful tools have been recently provided in the research literature in three main fields: autonomous vehicle localization and map building, visual feature processing and visual servoing for mobile robots. Our effort is mainly focused on the integration of visual servoing techniques for wheeled vehicles with advanced techniques for exploring and representing the environment.

Visual servoing of vehicles is an attractive solution for the estimation/control problem when implementing feedback directly on output measurements, i.e. grabbed images. Different settings have been considered in the visual servoing literature, using omnidirectional cameras (Thompson et al. 1999; Hadj-Abdefkader et al. 2006; Mariottini et af. 2006), pan-tilt heads (Tsakiris et af. 1997; Hespanha 2000), zooming cameras (Benhimane and Mafis 2003) or cameras carried by an articulated arm mounted on the robot (Tsakiris et al. 1997b).

In this paper, however, we assume the use of conventional cameras fixed on-board. This solution, which is the simplest and most economically viable, is also the most challenging from a technical point of view. The combination of the projective geometry underpinning camera information generation and the non-holonomic kinematics of wheeled vehicles produces an intrinsically nonlinear dynamical system, whose stabilization has attracted the attention of researchers since the last decade (Hashimoto and Noritsugu 1997; Conti-celli et al. 2000). An even harder set of problems is posed by conven-

The International Journal of Robotics Research
Vol. 28, No. 6, June 2009, pp. 802–814
DOI: 10.1177/0278364908097660

Figures 5–14 appear in color online: http://ijr.sagepub.com

tional cameras fixed on-board, because of the limited field-of-view (FOV) constraint they impose on the motion of image features while the vehicle maneuvers. Visual servoing with FOV constraints has been considered more recently in the robotics literature, with the earliest contributions (to the best of our knowledge) provided by Kantor and Rizzi (2003) and Murrieri et al. (2002, 2004). Recent advances on optimal feedback control (Bhattacharya et al. 2007; López-Nicolás et al. 2007) and in servoing in the presence of obstacles (Lopes and Koditschek 2007) were made.

To our knowledge, all methods for visual servoing have so far focused on local stabilization i.e. the initial and desired conditions of the system are assumed to be close enough so that a significant number of features remain in view all along the maneuver. In this work, we start from a modified version of the switching visual controller for non-holonomic vehicles with limited FOV (Murrieri et al. 2004), with the purpose of using it to servo the vehicle *in the large*, i.e. across paths connecting totally different initial and final views.

The visual servoing in the large is thus feasible if the robot has access to information that allows it to self-localize with respect to a sufficient number of waypoints which can be used to topologically connect the initial and desired images. Moreover, it also needs to collect sufficient metric data to reach one waypoint from another under visual servoing. A representation of the environment that conveys these metric and topological information will be referred to as a hybrid visual map. The construction of such a map, and its update with data obtained during robot servoed operations in an uncertain environment, are the subjects of investigation in this paper.

The literature on the problem of simultaneous visual-based localization and map building (v-SLAM) is rather extensive (e.g. Rekleitis et al. 2001; Chiuso et al. 2002; Davison 2003; Royer et al. 2005) and COTS software is already available (Karlsson et al. 2005). These results are clearly fundamental to our goals. However, accurate servoing of a vehicle with non-holonomic kinematics and FOV constraints is not considered in typical v-SLAM references, and represents the main original contribution of this paper.

Our approach in this paper consists of two main phases: a mapping phase and a navigation phase. During the mapping phase, the robot collects and tracks image features extracted from the camera's images. The initially unknown three-dimensional (3D) positions of the image features are estimated, possibly using iterated robot motions until the estimate accuracy reaches a desired level. The estimates of the features' positions are represented in the metric map, while robot initial pose can be connected with the final pose of the mapping process with a link to a new node in the topological map. During the navigation phase, which relies on visual servoing, the robot traverses the waypoints saved in the map while it localizes itself and updates the map information. In the following, we describe the components of the described strategy in detail.

2. Notation

Let $\langle W \rangle$ denote a global reference frame with respect to which all features are motionless (see Figure 1). Consider a calibrated monocular camera fixed onboard the vehicle, and let $\langle C \rangle$ denote the camera frame. We assume that the principal axis of the camera, denoted by Z_c, is aligned with the forward motion direction of the vehicle. Without loss of generality, we also assume that in the initial position, the origin of $\langle W \rangle$ and $\langle C \rangle$ coincide, and that $X_w = Z_c$ and $Y_w = Y_c$.

Let the robot's posture be denoted ${}^W\xi = {}^W[\xi_1, \xi_2, \xi_3]^T \in \Re^2 \times S$. More precisely, ($\xi_1$,$\xi_2$) are the cartesian coordinates of the middle point of the vehicle and ξ_3 is the angle between the Z_c axis and the X_w axis (Figure 2). Let the absolute position of the ith feature be ${}^W P_i = {}^W [p_1^i, p_2^i, p_3^i]^T \in \mathbb{R}^3$.

The position (x_i, y_i) of the features in the image plane is described by the well-known perspective projection mapping $\Upsilon : \Re^3 \to \Re^2$

$$\Upsilon : {}^C P_i \to \begin{bmatrix} x_i \\ \\ y_i \end{bmatrix} = \begin{bmatrix} \alpha_x \dfrac{{}^C p_1^i}{{}^C p_3^i} \\ \\ \alpha_y \dfrac{{}^C p_2^i}{{}^C p_3^i} \end{bmatrix}, \tag{1}$$

where α_x and α_y are camera calibration parameters representing the focal length multiplied by the pixel dimension scale factor for each axis of the image. The coordinates of the ith feature point in the camera frame $\langle C \rangle$ is ${}^C P_i = {}^C [p_1^i, p_2^i, p_3^i]^T$.

Let I_A denote the image observed from a given robot position ${}^W\xi_A$ (or A for short), and let F_A denote a set of features extracted from I_A. We assume that a number n_A of feature descriptors are included in F_A, which is obtained by using a robust feature extraction algorithm based on scale invariants (Se et al. 2002) (we used the implementation described in Karlsson et al. 2005). To each feature set F_A, a set of coordinates in the image plane ${}^{Im}F_A = \{[x_1, y_1]^T, \ldots, [x_{n_A}, y_{n_A}]^T\}$ and a set of coordinates in the world frame ${}^W F_A = \{{}^W[p_1^1, p_2^1, p_3^1]^T, \ldots, {}^W[p_1^{n_A}, p_2^{n_A}, p_3^{n_A}]^T\}$ are also associated.

Given two images I_A and I_B and their associated feature descriptor sets F_A and F_B, we let F_{AB} denote the set of corresponding features. More precisely, two features in F_A and F_B are considered to be corresponding if their distance (as the reciprocal of a weighted similarity, considering also e.g. luminosity or bitmap correlation) is below a chosen threshold.

3. Visual Servoing with Limited FOV

The baseline of the visual scheme adopted in this work is the switching visual servoing scheme presented in Murrieri et al. (2004). This controller is termed 'hybrid' in the previous work, referring to the mixed continuous-discrete nature of the dynamics involving the physics of the robot and the supervising logic. To avoid confusion with the mixed metric-topological

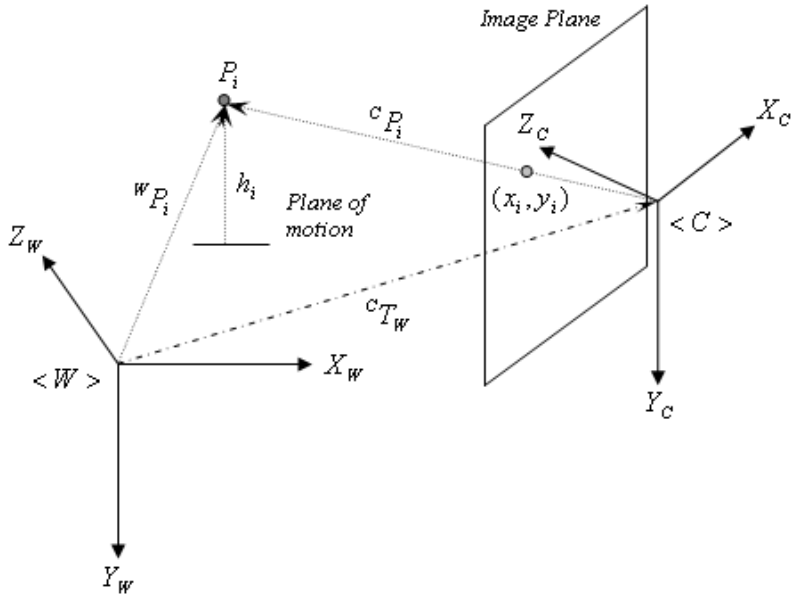


Fig. 1. Fixed frame $\langle W \rangle$, camera frame $\langle C \rangle$ and relative feature coordinates.

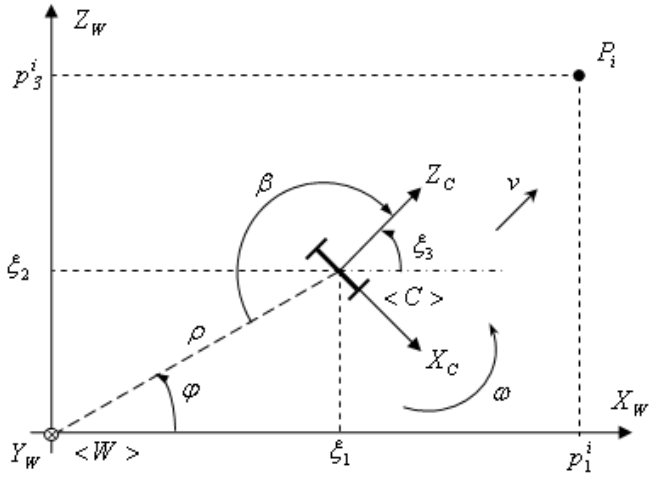


Fig. 2. Fixed frame $\langle W \rangle$, camera frame $\langle C \rangle$, and relative coordinates (ξ_1, ξ_2, ξ_3) and (ρ, ϕ, β).

nature of hybrid maps in this context, we will refer to the visual control scheme as 'switching'. We briefly recap the structure and the effectiveness of the controller, while skipping proofs and details to be found in the full article.

In the following, the image I_K with corresponding image feature positions ${}^{Im}F_K$ and 3D feature positions ${}^{W}F_K$ is assumed to be grabbed from the robot position ${}^{W}\xi_K$.

Let the unicycle dynamics be described in polar coordinates by

$$\begin{bmatrix} \dot{\rho} \\ \dot{\phi} \\ \dot{\beta} \end{bmatrix} = \begin{bmatrix} -\rho\ \cos\beta \\ \sin\beta \\ \sin\beta \end{bmatrix} u + \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \omega, \qquad (2)$$

where $\rho = \sqrt{\xi_1^2 + \xi_2^2}$, $\phi = \arctan \xi_2/\xi_1$ and $\beta = \pi + \phi - \xi_3$.

Let I_A denote the image observed from the robot current position and I_B denote the image from the desired robot position, which we consider to be in $\rho_B = \phi_B = 0$, $\beta_B = \pi$. We assume that F_{AB} contains at least $n \geq 4$ corresponding features, for which the coordinates ${}^{Im}F_A$ are measured on the current image, while ${}^{Im}F_B$ are known from a visual map of the environment. We temporarily assume also that the coordinates in the world frame ${}^{W}F_A = \{{}^{W}[p_1^1, p_2^1, p_3^1]^T, \ldots, {}^{W}[p_1^{n_A}, p_2^{n_A}, p_3^{n_A}]^T\}$ are available to the controller (this assumption will be removed later).

The constraint on the angle under which the camera views the kth tracked feature ${}^{Im}P_k = {}^{Im}[x_k, y_k]^T$ is expressed in these coordinates as

$$\begin{aligned} \gamma(\rho, \phi, \beta) &= \phi - \pi - \beta - \arctan \frac{{}^{W}p_3^k + \rho \sin\alpha}{{}^{W}p_1^k + \rho \cos\alpha} \\ &= \arctan \frac{{}^{Im}x_k}{\alpha_x} \in [-\Delta, \Delta] \end{aligned} \qquad (3)$$

where the limited FOV is described by a symmetric cone centered in the optical axis Z_c with semi-aperture Δ.

The switching controller is expressed in a set of different polar coordinates, which are conveniently denoted by intro-

ducing the two vectors $\tilde{\beta} = [\beta, \beta - \pi, \beta - \pi, \beta + \pi, \beta + \pi]$ and $\tilde{\phi} = [\phi - \pi, \phi - 2\pi, \phi, \phi - 2\pi, \phi]$. Correspondingly, a set of five distinct candidate Lyapunov functions can be written as

$$V_i(\rho, \alpha, \beta) = \frac{\rho^2}{2} + \frac{\tilde{\phi}_i^2}{2} + \frac{\tilde{\beta}_i^2}{2}, \tag{4}$$

with $i = 1, \ldots, 5$. The control law choice, i.e.

$$\begin{cases} u = \cos\beta \\ \omega = \lambda\tilde{\beta}_i + \dfrac{\sin\beta\cos\beta}{\tilde{\beta}_i}(\tilde{\phi}_i + \tilde{\beta}_i) \end{cases}, \tag{5}$$

is such that all the Lyapunov candidates have negative semi-definite time derivatives and (by La Salle's invariant set principle) are asymptotically stable. These five different control laws (parameterized by λ) define in turn five different controlled dynamics (analogous to Equation (2)) that are globally asymptotically stable in the state manifold $\Re^+ \times S^2$. Although none of these control laws alone can guarantee that the FOV constraint is satisfied throughout the parking maneuver, it is shown that a suitable switching logic among the control laws achieves this goal. The switching law is triggered when, during the stabilization with one of the five control laws, a feature approaches the border of the field of view by a threshold $\Delta_j < \Delta$, i.e. when $|\gamma| \geq \Delta_j$.

It should be noticed that the switching logic between different controllers could be triggered ever more frequently when ρ approaches zero. To avoid this so-called Zeno phenomenon, a dead zone is introduced in the controller for $\rho \leq \rho_D$, within which the forward velocity control u is set to zero. This implies that maneuvers are stopped when the desired accuracy ρ_D is reached.

4. Map Building

To apply the visual servoing scheme in the large, a map for the environment has to be built. This map must contain both metric information on a set of targets and waypoint images and topological information on the physical possibility of executing a motion (with the given kinematic and FOV constraints) from one waypoint to another. In our hybrid map representation, the metric information is represented by a set of robot postures, along with the corresponding 3D position estimates for the features observed from such postures. The topological information is represented by an undirected reachability graph (indeed, we assume that possible environment changes do not affect the traversability of the space by the robot).

More specifically, the graph is described as $G = (\bar{F}, \bar{\xi}, \bar{S})$ where $\bar{F}$ is the set of features subsets F_K and each F_K is associated to a node K. $\bar{\xi}$ is the set of poses ξ_K, and each ξ_K denotes the robot posture where the image I_K was taken. Finally, to each arc $S_{i,j}$ we associate a weight corresponding to the complexity of the maneuver to reach node i from node j. These weights could in principle be associated to minimum time, minimum distance or minimum control effort. However, because the actual execution of the robot motion will not follow an optimal strategy, we simply used the Euclidean distance between the nodes.

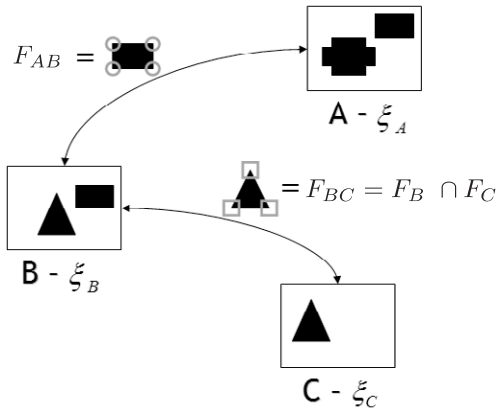


Fig. 3. Hybrid image map: grabbed images are indicated with a capital letter, say I_A, I_B, I_C. Each grabbed image corresponds to both a certain robot configuration in the metric map (see Figure 4) and a node in the topological map. The nodes A and B are connected if and only if the set of features $F_{AB} = F_A \cap F_B$ is not empty.

4.1. Overview of the Hybrid Map Construction Method

1. From the initial unknown position of the vehicle (i.e. ${}^W\xi_A = {}^W[0, 0, 0]^T$) an image I_A of a portion of the scene in view is grabbed and stored in the first node A of the hybrid map (see Figure 3).
2. From the image in view, a subset F_A of n_A features is selected.
3. The vehicle moves, avoiding obstacles with proximity sensors, in an arbitrary direction using a simple control law that keeps the image point features in view.
4. An extended Kalman filter is implemented using odometry and camera measurements to estimate the relative spatial position of the feature in camera frame $\langle C \rangle$. The estimated EKF state is

$$\begin{aligned} S &= [S_1^r, S_2^r, S_3^r, S_1^f, S_2^f, S_3^f, \ldots, S_{3n-1}^f, S_{3n}^f]^T \\ &= [{}^W\xi_1, {}^W\xi_2, {}^W\xi_3, {}^C p_1^1, {}^C p_2^1, {}^C p_3^1, \ldots, {}^C p_2^n, {}^C p_3^n]^T, \end{aligned}$$

i.e. the n features coordinates to estimate in the $\langle C \rangle$ camera frame.

5. Once 3D feature position estimates have converged to a value under a given level of uncertainty determined by the covariance matrix, the robot stops moving and a new node corresponding to the current pose is added to the hybrid map.

6. To add new nodes from the already created ones, the procedure starts again from step 2.

It is important to mention that the exploration strategy is not considered in this paper, while the focus is on both the gathered information representation and on the visual servoing control aspects. Therefore, we adopt a quite simple exploration strategy to add new nodes: between step 6 and step 2, the robot turns on the spot, choosing randomly between counterclockwise or clockwise rotations, until at most four features are within the FOV (note that only robot localization is needed at this step). Then, a new set of features is selected and estimated while the robot starts moving back and forth again. An additional image, after the robot rotation on the spot, is therefore also added to the map.

The described procedure is instantiated until the whole map is constructed. Notice that the use of simple EKF estimators is sufficient when sufficiently robust feature extraction and tracking techniques are available, such as was the case in our experimental setting. However, should feature outliers occur in the process, more robust filtering should be used in place of simple EKF (e.g. Vedaldi et al. 2005; Lu et al. 2006).

In the following, we provide more details on the methods used to build the two parts of the hybrid map.

4.1.1. Metric Data

In our experimental setup, the vehicle motion is assumed to be constrained on the ${}^{C}X \times {}^{C}Z$ plane (or equivalently ${}^{W}X \times {}^{W}Z$ plane). This hypothesis is correct when the robot moves at a fixed level, e.g. on the floor of an office or factory space, and implies that the coordinate ${}^{c}p_2^i = h_i$ of each feature is constant and represents the height of the feature on the plane of motion (see Figure 1). The initial guess for the extended Kalman filter is computed making the hypothesis that each feature is at the same generic height on the plane of motion and inverting the perspective projection Equation (1). The initial model covariance matrix is block diagonal and given by:

$$P_0 = \begin{bmatrix} P_0^r & 0 & \dots & 0 \\ 0 & P_0^1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P_0^n \end{bmatrix}.$$

In the first mapping phase, the environment is completely unknown. Therefore, all the entries of $P_0^r \in \Re^{3\times3}$ are set to zero since the frame $\langle W \rangle$ is positioned in the initial robot position. $P_0^i \in \Re^{3\times3}$, $\forall i = 1, \dots, n$, are initialized depending on feature mean estimation errors (e.g. $\sim$1 m for the feature distance from the camera and $\sim$20 cm for the other two coordinates in our experimental setting) and weighted with the relative feature distance from the camera, since the estimation accuracy reduces whenever the distance increases. (Note that the ith feature distance is estimated using the hypothesis that feature heights are fixed.)

The discrete nonlinear model of the feature dynamics and the unicycle kinematic model are assumed for state prediction:

$$\begin{bmatrix} \hat{S}_1^r(k+1) \\ \hat{S}_2^r(k+1) \\ \hat{S}_3^r(k+1) \\ \hat{S}_1^f(k+1) \\ \hat{S}_2^f(k+1) \\ \hat{S}_3^f(k+1) \\ \vdots \\ \hat{S}_{3n-1}^f(k+1) \\ \hat{S}_{3n}^f(k+1) \end{bmatrix} = \begin{bmatrix} \hat{S}_1^r(k) + \cos\left(\hat{S}_3^r(k) + \frac{u_2(k)}{2}\right) u_1(k) \\ \hat{S}_2^r(k) + \sin\left(\hat{S}_3^r(k) + \frac{u_2(k)}{2}\right) u_1(k) \\ \hat{S}_3^r(k) + u_2(k) \\ \hat{S}_1^f(k) + \hat{S}_3^f(k) u_2(k) \\ \hat{S}_2^f(k) \\ \hat{S}_3^f(k) - u_1(k) - \hat{S}_1^f(k) u_2(k) \\ \vdots \\ \hat{S}_{3n-1}^f(k) \\ \hat{S}_{3n}^f(k) - u_1(k) - \hat{S}_{3n-1}^f(k) u_2(k) \end{bmatrix}$$

where $U(k) = [u_1(k), u_2(k)]^T$ are the encoder measurements for forward and angular velocity, obtained from

$$u_1(k) = R\frac{\omega_r(k) + \omega_l(k)}{2}$$

and

$$u_2(k) = R\frac{\omega_r(k) - \omega_l(k)}{L},$$

respectively. ω_r and ω_l are the rotational encoder for the right and left wheel, R is the wheel radius and L is the length of the wheel axle.

Two different noise sources are taken into account. The model dynamical errors $\eta = (\eta_1^r, \eta_2^r, \eta_3^r, \eta_1, \eta_2, \eta_3, \dots, \eta_{3n-1}, \eta_{3n})^T$ are modeled as additive and zero mean gaussian noises with covariance matrix $Q = P_0$. Systematic errors are assumed to be removed by suitable calibration, hence non-zero mean errors are not modeled. The odometry errors γ_r, γ_l, for the right and left wheel respectively, are assumed to be zero mean and gaussian distributed. They are also assumed to be equal for both wheels. This is a simple but easily verified assumption in respect of generic unicycle-like vehicles, and is computed taking into account lack of accuracy in odometry (typically due to wheels slipping and skidding). The covariance matrix of the prior estimate (model prediction) is then calculated by the formula

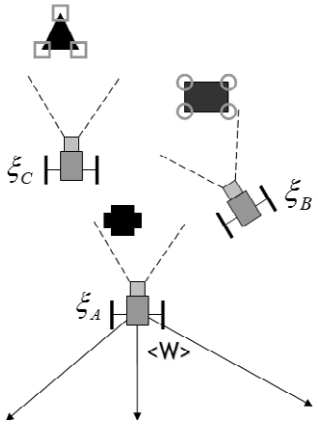


Fig. 4. Metric map: the positions A, B, C, representing each image node in the topological map (see Figure 3), ξ_A, ξ_B and ξ_C are a set of 3D robot postures.

$$P_k^- = A_k P_{k-1} A_k^T + \sigma_\gamma^2 B_k B_k^T + Q$$

where σ_γ^2 is the input variance ($\gamma = \gamma_r = \gamma_l$ by assumption), A_k and B_k are the model Jacobians and P_{k-1} is the model covariance matrix at the previous step.

By inverting Equation (1), filter-predicted measurements are obtained while the measurement corrections are the actual image feature positions, considering a zero mean gaussian noise to represent the inaccuracy of image coordinates extraction.

4.1.2. Topological Data

Once three-dimensional feature position estimates have converged to a value with a low level of uncertainty, the image I_B, grabbed from the previously unknown position ξ_B reached at the end of the estimation process, is added as a new node B in the hybrid map. The set F_{AB} represents the feature set used by the visual servoing controller to steer the vehicle from position ξ_A to position ξ_B in the metric mapped space (see Figure 4) or, equivalently, from node A to node B in the topological space (see Figure 3).

It is worthwhile to note that the relation between image nodes J and metric robot positions ξ_J has the so-called *downward/upward solution property* that guarantees consistency between the mapped spaces (Thrun and Biicken 1996). In particular, the property satisfaction ensures that the planned maneuvers, i.e. a path planning in the topological map or a visually servoed path in the metric map, are consistent. The visual servoing technique underpinning our method verifies this property by construction.

The mapping process will be able to continue starting from position ξ_B (or again from ξ_A), adding a new node in the hybrid map, say C, and an image I_C, together with a new vehicle position in the metric map ξ_C and a new set of estimated features F_{BC} (or F_{AC}), and so on. Note that the global set of features in view from node B is $F_B \supset F_{AB} \cup F_{BC}$ and that $F_{AB} \cap F_{BC} \neq \emptyset$ implies $F_{AC} \neq \emptyset$, directly enforcing a connection between the positions ξ_A and ξ_C. Furthermore, the set F_{AB} can be used to travel from ξ_A to ξ_B and vice versa by inverting desired and initial image feature positions in the position based visual servoing controller. A graphical example of a hybrid map produced by a vehicle that has traversed positions ξ_A, ξ_B and ξ_C of an unknown environment is depicted in Figures 3 and 4.

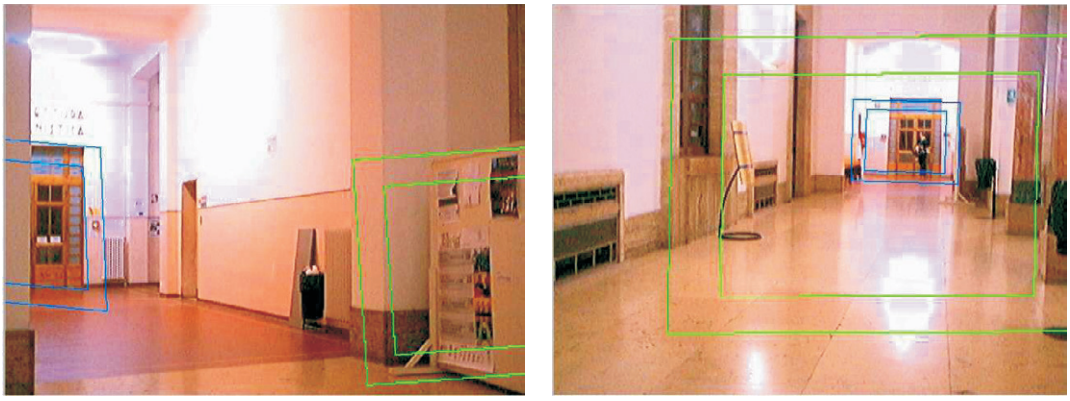
5. Navigation through Topological Waypoints

Let the robot be in a generic mapped position, say ${}^W\xi_A = {}^W[\xi_1, \xi_2, \xi_3]^T$ (or a node A with image I_A). Suppose that the robot has to reach a new position, say ξ_K, expressed in the metric map (whose fixed reference frame is $\langle W \rangle$). If ${}^W\xi_K$ corresponds to a topologically mapped location K, which has an associated image I_K, a standard graph visiting algorithm is used for the path selection from A to K in the image map. This therefore permits the vehicle to steer through the map nodes using the servoing presented in Section 3. One possibility in order to implement a minimum traveling path algorithm through actual and desired nodes is using the A* algorithm (Hart et al. 1968). This algorithm demands an admissible heuristic estimate of the distance, to be saved in the node link weight $S_{A,B}$. In our case, the distance between node A and node B is the physical distance given by the relative node positions ξ_A and ξ_B. The vehicle thus travels from A to K through a set of mapped images.

It is worthwhile to note that the goal of the visual servoing is expressed with an image mapped in the topological map. Indeed, the hybrid map representation ensures that each image is labeled with the corresponding graph node and each graph node is connected to the robot metric space. Moreover, it is also important to note that paths taken during exploration are likely to be a straight line in Cartesian space. However, the visual servoed path is not necessarily so.

Remembering that a single connection between two nodes is used to travel between the node images (desired image and final image are interchangeable), graph navigation is deadlock free in static environments.

It is worthwhile to note that, in the case of multiple agents in the same unknown environment, each agent builds its hybrid map that can be successively fused with the other agents' maps. Indeed, fusing two topological maps comes relatively easy when it is possible to identify a visual servoing path between nodes relative to two different maps, avoiding more complex metric map merging techniques (Leonard et al. 2002;



(a) (b)

Fig. 5. Identifying nodes in the environment: (a) a kidnapped robot identifies two possible nodes in the known map that can be reached; (b) from position one (outer green contours), it is possible to reach position two (inner blue contours), thus enforcing a path closure or map merging.

Thrun et al. 2002). Since topological map merging relies on SIFT features to detect similar images by determining the set of common features between the image in view from the robot and the mapped image, the same idea can be used to solve the kidnapped robot problem (Figure 5).

Similarly, if the robot recognizes a previously mapped image node during the map building process and, if the node is reachable, a new edge is created and the closed path is generated. Again, consistency of the closed loop connections between the path planners and the hybrid map is ensured by the visual servoing controller and the upward/downward solution property.

6. Robot Localization and Map Update

During the navigation on the map, the robot perceives new information and can continuously update its estimate of its own posture. An update of the metric map on which localization is based is also possible. To do so, we again adopt an Extended Kalman Filter. Selecting a set of $n \geq 4$ estimated feature points (Murrieri et al. 2004), the EKF state will be

$$\begin{aligned} S &= [S_1^r, S_2^r, S_3^r, S_1, S_2, S_3, \ldots, S_{2n-1}, S_{2n}]^T \\ &= [{}^W\xi_1, {}^W\xi_2, {}^W\xi_3, {}^C p_1^1, {}^C p_3^1, \ldots, {}^C p_1^n, {}^C p_3^n]^T, \end{aligned}$$

where the first three elements represent the vehicle state space. The feature heights come from previously estimated values.

Estimated state initial guess is computed using the least mean squares static localization proposed in Murrieri et al. (2004) for the vehicle and inverting Equation (1) for the feature. The initial model covariance matrix is again block diagonal and is initialized depending on vehicle localization errors (∼1 m for robot cartesian position and ∼1 radian for the orientation in our experimental setting).

The state prediction model, incorporating the vehicle kinematic, is

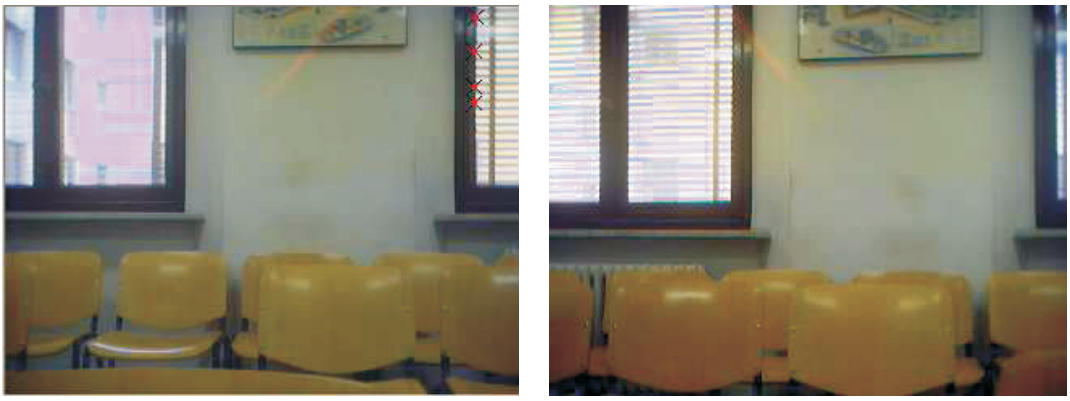
$$\begin{bmatrix} \hat{S}_1^r(k+1) \\ \hat{S}_2^r(k+1) \\ \hat{S}_3^r(k+1) \\ \hat{S}_1(k+1) \\ \hat{S}_2(k+1) \\ \vdots \\ \hat{S}_{2n-1}(k+1) \\ \hat{S}_{2n}(k+1) \end{bmatrix} = \begin{bmatrix} \hat{S}_1^r(k) + \cos\left(\hat{S}_3^r(k) + \frac{u_2(k)}{2}\right) u_1(k) \\ \hat{S}_2^r(k) + \sin\left(\hat{S}_3^r(k) + \frac{u_2(k)}{2}\right) u_1(k) \\ \hat{S}_3^r(k) + u_2(k) \\ \hat{S}_1 + \hat{S}_2 u_2(k) \\ \hat{S}_2 - u_1(k) - \hat{S}_1 u_2(k) \\ \vdots \\ \hat{S}_{2n-1} + \hat{S}_{2n} u_2(k) \\ \hat{S}_{2n} - u_1(k) - \hat{S}_{2n-1} u_2(k) \end{bmatrix}$$

where $U(k) = [u_1(k), u_2(k)]^T$ are, again, the encoder measurements for forward and angular velocity. Noisy measurements, modeling errors

$$\eta = (\eta_1^r, \eta_2^r, \eta_3^r, \eta_1, \ldots, \eta_{2n})^T$$

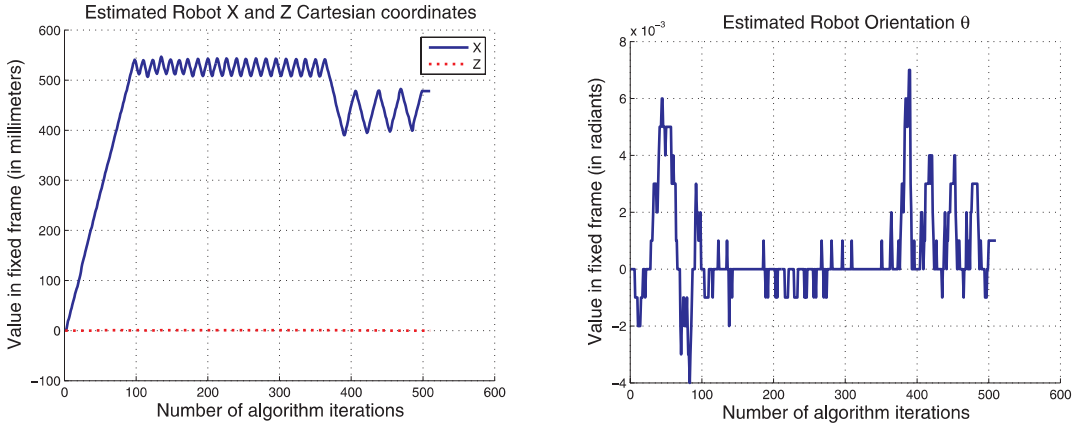
and noisy odometry data are considered in a similar way as in the map building filter.

It is worthwhile noting that during the servoed path it is possible to enforce the feature estimation obtained in the mapping process and to add a more reachable set of 3D features from unmapped entities.



(a) (b)

Fig. 6. Images grabbed by the robot camera in (a) the starting position and (b) final position of the estimation process. These images will be added as image nodes to the hybrid map.



(a) (b)

Fig. 7. Estimated (a) robot positions and (b) orientation during the mapping process are reported (back and forth, image based, mapping phase). The estimation algorithm adopted is the extended Kalman filter.

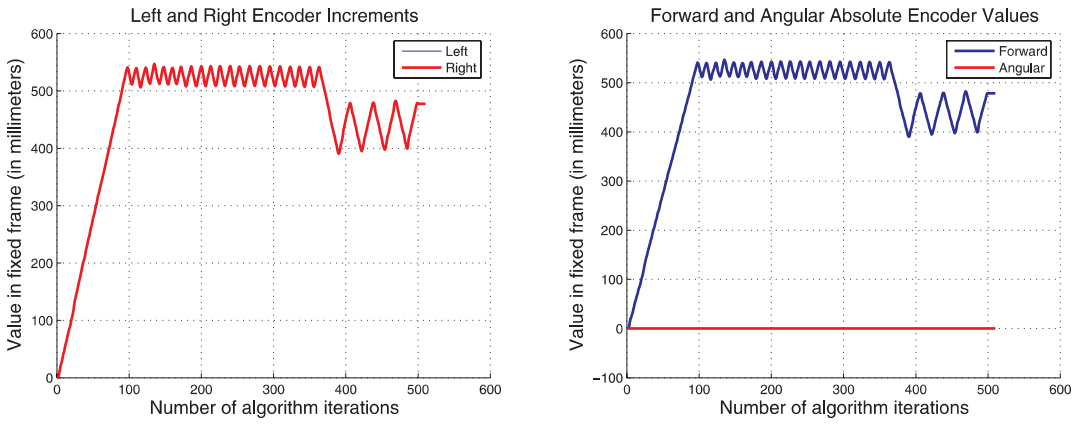
7. Experimental Results

7.1. Hybrid Map Experiments

A low-cost apparatus was employed to highlight the potential of the proposed technique. The experimental setup comprises of a K-Team Koala vehicle (www.k-team.com/robots/koala/index.html), equipped with a commercial webcam placed on the front part of the robot platform. The vehicle has two symmetric rows of three wheels on its sides, each actuated by a single low-resolution stepper-motor actuator. The construction implies that slipping and skidding of some of the wheels occurs whenever the vehicle moves along a curved trajectory. Such conditions make it hard to use odometry for localization and control, and strongly motivate the use of visual servoing. The controller is implemented under Windows XP on a 1130 MHz Pentium III laptop mounted onboard. SIFT elaboration is performed using ERSP vision library (Goncalves et al. 2005; Karlsson et al. 2005). The Intel OpenCV (www.intel.com/research/mrl/research/opencv/) library was used to compute optical flow and to track features. The hardware communication between the robot and the laptop is performed by a RS-232 serial cable.

7.1.1. Map Building

The image-based controller used for exploration avoids feature occlusions and obstacles and it is able to take into account



(a) (b)

Fig. 8. Encoder values for (a) the left and right wheel and (b) the resulting forward and angular paths, i.e. the time integrals of the unicycle forward and angular velocities.

the limited field of view constraint. More precisely, the image-based controller simply steers the vehicle back and forth to estimate the feature 3D coordinates (using an EKF estimator with erroneous feature height initial guesses) and to satisfy the previously mentioned visual constraints. Then, at the end of each feature visual estimation, it rotates and starts the mapping phase again with two other images.

In the experiments both the mapping phase (topological and metric) and the navigation phase (visual servoing) are reported. In the mapping experiment, the robot collects a set of images and, for each pair of images, say I_i and I_j, it estimates the 3D coordinates of the image feature points F_{ij} and the 3D robot positions ξ_i and ξ_j. The images I_4 (Figure 6a) and I_5 (Figure 6b) are grabbed by the robot camera at the beginning and at the end of the estimation process, respectively.

Figure 7 depicts the robot estimated position during the mapping: (ξ_1, ξ_2) cartesian positions are reported on the left, while the orientation ξ_3 is on the right.

The encoder values for left and right wheel and for forward and steering path are also reported (see Figure 8). Note that the robot has traveled an almost linear path, moving back and forth, therefore left and right wheel encoders are almost identical and angular encoder value is always zero. Moreover, the forward and angular encoder values are actual measures made available to the EKF, computed from the left and right wheel encoders. An image-based controller was employed, together with an obstacle avoidance controller based on proximity sensors (notice the correction in the trajectory at $t \sim 35$ s).

It should be noted that the particularly smooth path traveled by the robot during the mapping process dramatically reduces odometry lack of accuracy, allowing a more accurate and faster feature estimation. Finally, notice how the EKF estimator strongly relies on the dead-reckoning data from the wheel encoders.

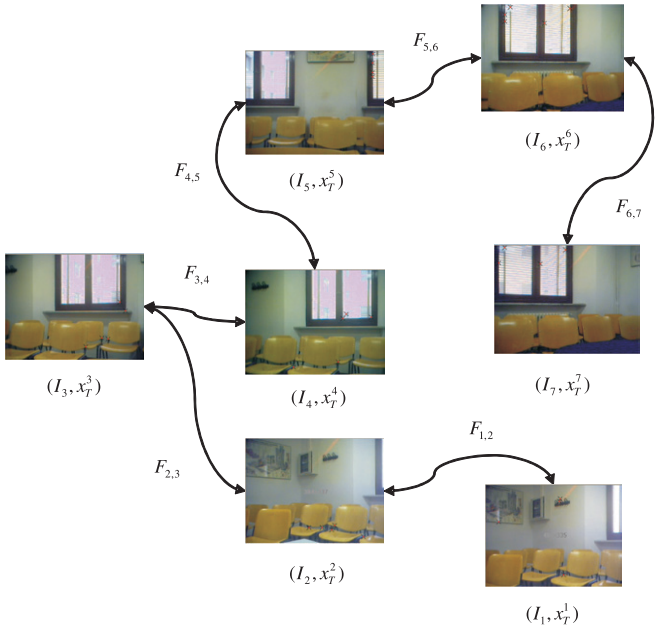


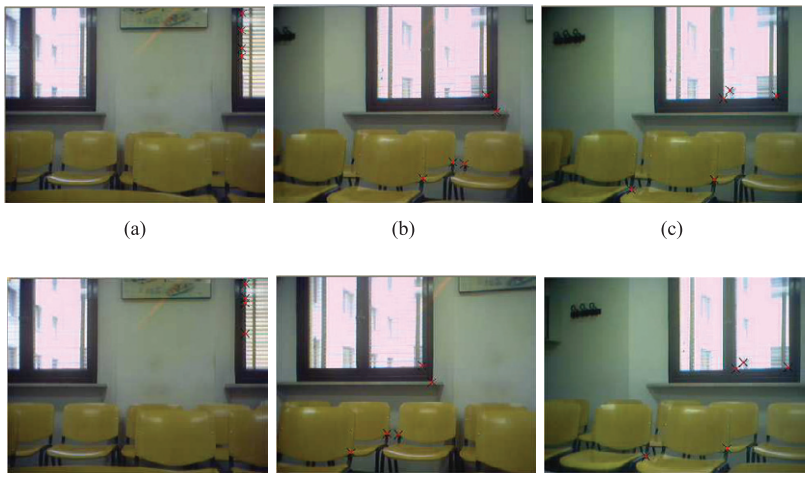
Fig. 9. The topological image-based map.

The experimental mapping process runs for 50 s and the sampling period (i.e. the inverse of the frequency of the EKF steps) is $T = 0.1$ s. The sampling period T is determined by the worst-case frame rate available for commercial webcams. Although even low cost cameras ensure about 20–30 frames per second, the rate changes depending on ambient illumination variations.

Features are represented using a patch from the captured image. It is worthwhile to note that all the processing is carried out online. Figure 9 reports an image graph created during a mapping traveling.

(a) (b)

Fig. 10. Map navigation: (a) desired image and (b) initial image.



(a) (b) (c)

(d) (e) (f)

Fig. 11. (a–c) Desired images from each topological map node and (d–e) images grabbed from the camera after each visually servoed path.

7.1.2. Localization and Navigation

The visual servoing controller has been used to travel the distance between the mapped images, parking the vehicle in the position x_T^3 (position I_3 in the topological map, see Figure 10a). The initial robot position is unknown, but the architecture solves the kidnapped robot problem identifying the topological position x_T^5 (see Figure 10b). Hence, the visual servoing path corresponds to a travel between image node I_5 to I_2.

In Figure 11, the nodes crossed by the robot during the parking are represented. In Figures 11a–c, the images stored in the topological map (i.e. desired images for the visual servoing) are represented. In Figures 11d–e, the images grabbed from the camera after each path are depicted. Once it is possible to localize and track features of the next node to be reached, an intermediate node is no longer approached.

A wide movement in the mapped environment comprises several limited movements between each pair of images (Figure 11). Nevertheless, the visual servoed motion between successive images is still quite small. Indeed, it is well known in the literature (Chaumette and Hutchinson 2006, 2007) that large image errors (hence, large robot movements) decrease accuracy and robustness of the visual servo controller. In the proposed architecture, the granularity of the topological map is related to the visual servoing accuracy.

Figure 12 reports image feature coordinates during the parking maneuver between two images of the topological image map, while Figure 13 reports the angle of attention of each feature (Murrieri et al. 2004).

Notice that the rather large oscillations that can be observed in Figures 12 and 13 describe motions in the image plane of the observed features corresponding to the small maneuvers of the non-holonomic vehicle that are necessary to cope with

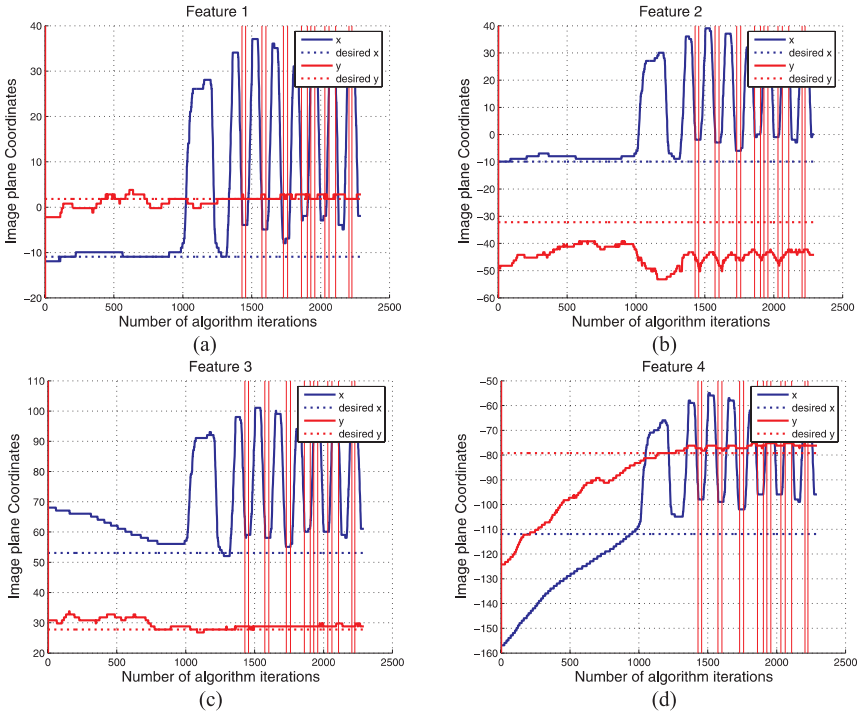


Fig. 12. Feature image plane coordinates, reported in pixels.

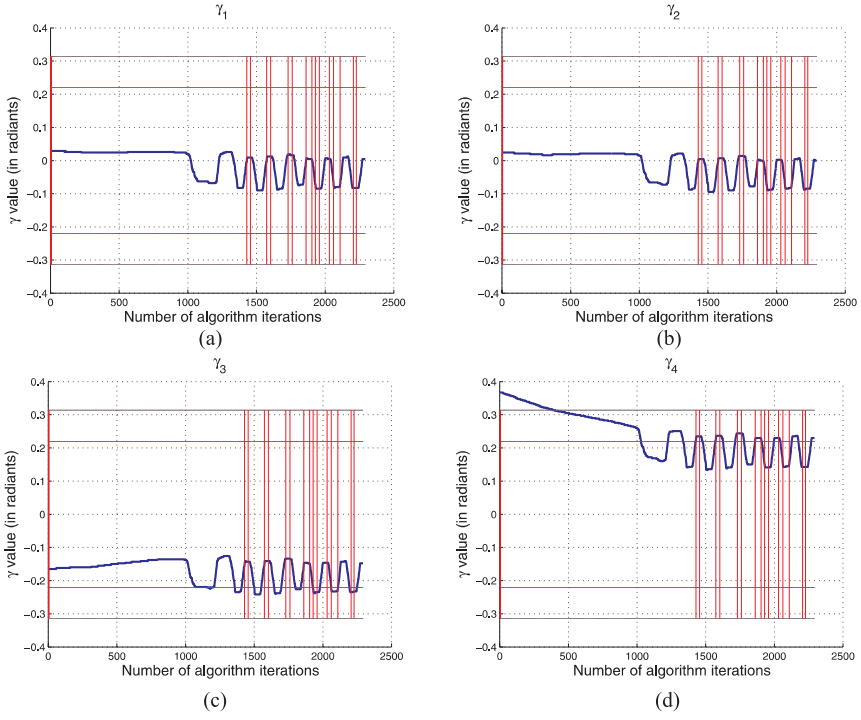
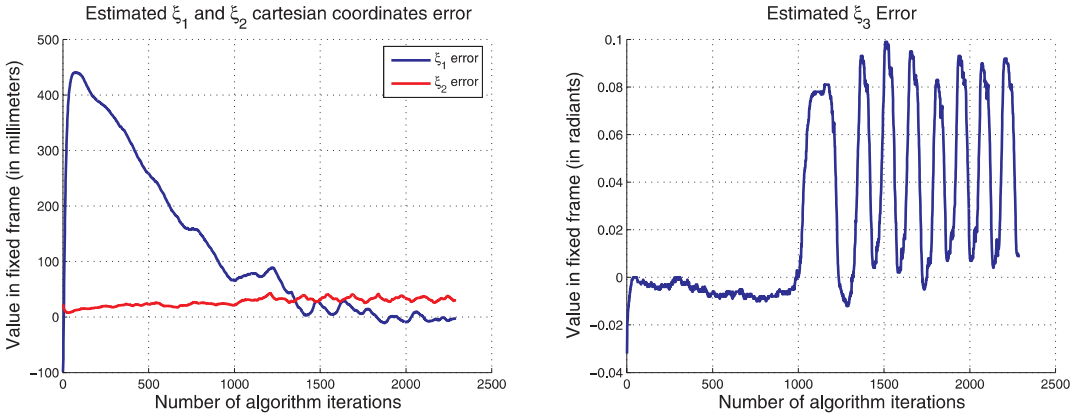


Fig. 13. γ angles, reported in radians.



(a) (b)

Fig. 14. Visual servoing error compensation.

the camera-limited FOV when near the final position (Murrieri et al. 2004). As described above, the control switching policy stops the robot when the desired distance threshold ρ_D is reached, as shown in Figures 12 and 13.

Finally, Figure 14 reports the parking errors for Cartesian coordinates (ξ_1, ξ_2) and orientation ξ_3 during a parking task accomplishment. The total length of the parking maneuver is about 22 s with sampling time in the parking task $T = 0.1$ s. It should be noted that, in Figure 14a, the rather ample feature motions observed before in the image plane (Figure 12) correspond to quite moderate actual motions of the vehicle center, and are caused by the orientation changes in the parking maneuver (Figure 14b).

8. Conclusions

In this paper, we have proposed a visual servoing scheme for a non-holonomic vehicle in unknown indoor environments. The proposed approach gives a solution to the problem of autonomously building a map for servoing purposes. The solution is based on previously developed control schemes and a map capable of overcoming the limits of these schemes in a large environment. The work could be regarded as an attempt to connect control techniques (action) and sensorial data interpretation (perception). The method has the advantage that the maps produced and stored are rather small in terms of memory occupancy, hence in communication bandwidth requirements. This permits map sharing among mobile agents, which is the goal of future work. Many other interesting developments of the present work are possible, including the adoption of a purely appearance-based navigation and mapping scheme for accurate servoing.

Acknowledgements

Financial support was provided by EC Contract IST-2004-004536 (IP RUNES) and IST-2004-511368 (N.o.E. HYCON).

References

Benhimane, S. and Mails, E. (2003). Vision-based control with respect to planar and non-planar objects using a zooming camera. In *Proceedings of IEEE International Conference on Robotics and Automation,* Taipei, Taiwan, pp. 991–996.

Bhattacharya, S., Murrieta-Cid, R., and Hutchinson, S. (2007). Optimal paths for landmark-based navigation by differential-drive vehicles with field-of-view constraints. *IEEE Transactions on Robotics,* **23**(1): 47–59.

Chaumette, F. and Hutchinson, S. (2006). Visual servo control, Part I: Basic approaches. *IEEE Robotics and Automation Magazine,* **13**(4): 82–90.

Chaumette, F. and Hutchinson, S. (2007). Visual servo control, Part II: Advanced approaches. *IEEE Robotics and Automation Magazine,* **14**(1): 109–118.

Chiuso, A., Favaro, P., Jin, H., and Soatto, S. (2002). Structure from motion casually integrated over time. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **24**(4): 523–535.

Conticelli, F., Prattichizzo, D., Guidi, F., and Bicchi, A. (2000). Vision-based dynamic estimation and set-point stabilization of nonholonomic vehicles. In *Proceedings of IEEE International Conference on Robotics and Automation,* San Francisco, CA, USA, pp. 2771–2776.

Davison, A. J. (2003). Real-time simultaneous localization and mapping with a single camera. In *Proceedings of IEEE International Conference on Computer Vision,* Volume 2, pp. 1403–1410.

Goncalves, L., DiBernardo, E., Benson, D., Svedman, M., Ostrowski, J., Karlsson, N., and Pirjanian, P. (2005). A visual front-end for simultaneous localization and mapping. In *Proceedings of IEEE International Conference on Robotics and Automation,* Barcelona, Spain, pp. 44–49.

Hadj-Abdelkader, H., Mezouar, Y., Aldreff, N., and Martinet, P. (2006). Omnidirectional visual servoing from polar lines. In *Proceedings of IEEE International Conference on Robotics and Automation,* pp. 2385–2390.

Hart, P., Nilsson, N., and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics,* **4**(2): 100–107.

Hashimoto, K. and Noritsugu, T. (1997). Visual servoing of nonholonomic cart. In *Proceedings of International Conference on Robotics and Automation,* pp. 1719–1724.

Hespanha, J. P. (2000). Single camera visual servoing. In *Proceedings of IEEE International Conference on Decision and Control,* pp. 2533–2538.

Kantor, G. and Rizzi, A. (2003). Feedback control of underactuated systems via sequential composition: Visually guided control of a unicycle. In *Proceedings of International Symposium on Robotics Research,* pp. 281–290.

Karlsson, N., DiBernardo, E., Ostrowski, J., Goncalves, L., Pirjanian, P., and Munich, M. E. (2005). The vSLAM algorithm for robust localization and mapping. In *Proceedings of IEEE International Conference on Robotics and Automation,* Barcelona, Spain, pp. 24–29.

Leonard, J., Rikoski, R., Newman, P., and Bosse, M. (2002). Mapping partially observable features from multiple uncertain vantage points. *International Journal of Robotics Research,* **21**(11): 943–975.

Lopes, G. and Koditschek, D. (2007). Visual servoing for nonholonomically constrained three degree of freedom kinematic systems. *International Journal of Robotics Research,* **26**(7): 715–736.

López-Nicolás, G., Bhattacharya, S., Guerrero, J., Sagües, C., and Hutchinson, S. (2007). Switched homography-based visual control of differential drive vehicles with field-of-view constraints. In *Proceedings of IEEE International Conference on Robotics and Automation,* Rome, pp. 4238–4244.

Lu, L., Dai, X., and Hager, G. D. (2006). Efficient particle filtering using RANSAC with application to 3D face tracking. *Image and Vision Computing,* **24**(6): 581–592.

Mariottini, G., Prattichizzo, D., and Cerbella, A. (2006). Image-based visual servoing for mobile robots with catadioptric camera. In *European Robotics Symposium,* pp. 159–170.

Murrieri, P., Fontanelli, D., and Bicchi, A. (2002). Visual-servoed parking with limited view angle. In B. Siciliano and P. Dario (Eds.), *Experimental Robotics VIII,* Volume 5 of *Springer Tracts in Advanced Robotics (STAR),* pp. 254–263. Springer Verlag.

Murrieri, P., Fontanelli, D., and Bicchi, A. (2004). A hybrid-control approach to the parking problem of a wheeled vehicle using limited view-angle visual feedback. *International Journal of Robotics Research,* **23**(4–5): 437–448.

Rekleitis, I., Sim, R., Dudek, G., and Milios, E. (2001). Collaborative exploration for the construction of visual maps. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems,* pp. 1269–1274.

Royer, E., Bom, J., Dhome, M., Thuilot, B., Lhuillier, M., and Marmoiton, F. (2005). Outdoor autonomous navigation using monocular vision. In *Proceeding of the IEEE/RSJ Conference on Intelligent Robots and Systems,* pp. 1253–1258.

Se, S., Lowe, D., and Little, J. (2002). Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research,* **21**(8): 735–758.

Thompson, S., Zelinsky, A., and Srinivasan, M. (1999). Automatic landmark selection for navigation with panoramic vision. In *Proceedings of Australian Conference on Robotics and Automation AGRA'99.*

Thrun, S. and Biicken, A. (1996). Integrating grid-based and topological maps for mobile robot navigation. In *Proceedings of the AAAI Thirteenth National Conference on Artificial Intelligence,* Portland, Oregon.

Thrun, S., Roller, D., Ghahmarani, Z., and DurrantWhyte, H. (2002). SLAM Updates require constant time. Technical report, School of Computer Science, Carnegie Mellon University.

Tsakiris, D., Rives, P., and Samson, C. (1997). Applying visual servoing techniques to control nonholonomic mobile robots. In *International Conference on Intelligent Robots and Systems.* Workshop on 'New Trends in Image-based Robot Servoing'.

Vedaldi, A., Jin, H., Favaro, P., and Soatto, S. (2005). KALMANSAC: Robust filtering by consensus. In *Proceedings of the International Conference on Computer Vision (ICCV),* Volume 1, pp. 633–640.