Large-Scale 6-DOF SLAM With Stereo-in-Hand

Lina M. Paz, *Member, IEEE*, Pedro Piniés, *Member, IEEE*, Juan D. Tardós, *Member, IEEE*, and José Neira, *Member, IEEE*

***Abstract*—In this paper, we describe a system that can carry out simultaneous localization and mapping (SLAM) in large indoor and outdoor environments using a stereo pair moving with 6 DOF as the only sensor. Unlike current visual SLAM systems that use either bearing-only monocular information or 3-D stereo information, our system accommodates both monocular and stereo. Textured point features are extracted from the images and stored as 3-D points if seen in both images with sufficient disparity, or stored as inverse depth points otherwise. This allows the system to map both near and far features: the first provide distance and orientation, and the second provide orientation information. Unlike other vision-only SLAM systems, stereo does not suffer from "scale drift" because of unobservability problems, and thus, no other information such as gyroscopes or accelerometers is required in our system. Our SLAM algorithm generates sequences of conditionally independent local maps that can share information related to the camera motion and common features being tracked. The system computes the full map using the novel conditionally independent divide and conquer algorithm, which allows constant time operation most of the time, with linear time updates to compute the full map. To demonstrate the robustness and scalability of our system, we show experimental results in indoor and outdoor urban environments of 210 m and 140 m loop trajectories, with the stereo camera being carried in hand by a person walking at normal walking speeds of 4–5 km/h.**

***Index Terms*—Linear time, scalability, stereo vision, visual SLAM.**

I. Introduction: State-of-the Art in Visual SLAM

THE INTEREST in using cameras in simultaneous localization and mapping (SLAM) has grown tremendously in recent times. Cameras have become much more inexpensive than lasers, and also provide texture rich information about scene elements at practically any distance from the camera. 6-DOF SLAM systems based on 3-D laser scanners plus odometry have been demonstrated feasible both indoors and outdoors [2], [3], as well as vision aided by laser without odometry [4] and vision aided by an inertial navigation system [5], [6]. But in applications where it is not practical to carry heavy and bulky sensors, such as egomotion for people tracking and environment modeling in rescue operations, cameras seem the only light weight sensors that can be easily adapted to helmets used by rescuers, or simply worn.

Current visual SLAM research has been focused on the use of either monocular or stereo vision to obtain 3-D information from the environment. Quite a few monocular visual SLAM systems have been demonstrated to be viable for small environments [7]–[16]. Most are essentially standard extended Kalman filter (EKF) SLAM systems, and vary in the technique used to initialize a feature, given the partiality of the bearing only information provided by one camera, or in the type of interest points extracted from the images (be it Harris corners, Shi–Tomasi corners, scale-invariant feature transform (SIFT) features, or some combination). Some works have also considered segment features [17], [18]. Larger environments have been tackled in hierarchical visual SLAM [19].

A single camera is used in all of these systems, and although very distant features are potentially detectable, scale unobservability is a fundamental limitation. Either the scale is fixed in some way (for example, by observing a known object [16]), or drift in scale can occur as is reported in the hierarchical visual SLAM system [19]. Panoramic cameras are also being used in visual SLAM [20], [21]. Here, the limitation of scale unobservability is overcome using an additional stereo vision bench for motion estimation between consecutive frames. In the work of Royer *et at.* [22], only monocular images are used. Mapping is achieved using a batch hierarchical bundle adjustment algorithm to compute all camera as well as interest points locations. The scale is introduced in the system by manually entering the length of the path.

Manuscript received December 17, 2007; revised July 3, 2008. This paper was recommended for publication by Associate Editor J. Leonard and Editor L. Parker upon evaluation of the reviewers' comments. This work was supported in part by the European Union under Project RAWSEEDS FP6-IST-045144 and in part by the Dirección General de Investigación of Spain under Project SLAM6DOF DPI2006-13578. A preliminary version of this work was presented at the *Visual SLAM: An Emerging Technology* workshop held at the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, 2008.

The authors are with the Department of Computer Science and Systems Engineering, Universidad de Zaragoza, Zaragoza E-50018, Spain (e-mail: linapaz@unizar.es; ppinies@unizar.es; tardos@unizar.es; jneira@unizar.es).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the author. This material includes the following video files. *VSLAM_stereo_distribution.avi:* Suppose that a point is located in front of a stereo camera at different distances ranging from 1 to 20 m. The video shows the uncertainty distributions estimated by different methods when the point position is reconstructed from a noisy stereo pair observation. Cloud points are samples obtained from the real distribution, red ellipses represent uncertainties estimated using depth point parameterization, and the interior of the green regions represent the uncertainties estimated using inverse depth parameterization. It is clear from the video that the depth point parameterization is overconfident when the original point is far from the stereo. *VSLAM_local_map.avi:* In this video, the process of building a local map is shown. The video reproduces the sequence of images captured by the stereo as well as a top and lateral perspectives of the local map built. *VSLAM_outdoor.avi*, *VSLAM_indoor.avi:* To demonstrate the robustness and scalability of our system, we show experimental results in indoor and outdoor urban environments of 210- and 140-m loop trajectories, with the stereo camera being carried in hand by a person walking at normal walking speeds of 4–5 km/h. The colors used in the camera images are green for features predicted and matched, blue for predicted but not matched, and red for matchings rejected by JCBB. *Note:* The videos VSLAM_local_map.avi, VSLAM_outdoor.avi, VSLAM_indoor.avi, and VSLAM_stereo_distribution.avi have been tested on VLC media plater: http://www.videolan.org; Windows Media Player with xvid codec from: http//www.xvid.org/

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

Digital Object Identifier 10.1109/TRO.2008.2004637

1552-3098/$25.00 © 2008 IEEE

Stereo visual systems provide scale through the baseline between the cameras, known from calibration. Davison and Murray demonstrated the first active stereo visual SLAM system [23]–[25]. It is based on standard EKF, and thus, has low scalability also. Under restrictive planar environment assumptions, Iocchi *et al.* built an environment map using stereo [26]. Se *et al.* demonstrated a visual stereo SLAM system using SIFT features in a small laboratory environment [27]. This system is also unlikely to scale adequately to large environments or work in more challenging outdoor scenarios as cross-correlations were neglected for computational reasons. In [28] and [29], the authors demonstrate an autonomous blimp system for terrain mapping using stereo as the only sensor, also using a standard EKF SLAM algorithm. Saez *et al.* [30] presented a 6-DOF stereo visual SLAM system, where egomotion estimation is done by a 3-D point matching algorithm, and mapping through a global entropy minimization algorithm in indoor orthogonal scenarios, with difficult extension to more complex nonorthogonal environments.

In [31] and [32], Sim *et al.* describe a dense visual SLAM system using Rao–Blackwellized particle filters and SIFT features (a similar effort in using Rao–Blackwellized particle filters and SIFT features for visual SLAM was reported in [15]). Visual odometry [structure from motion (SFM)] is used to generate proposals for the sensor motion and global pose estimation algorithms for loop closing. This system works in either monocular or stereo mode, with cameras mounted on a robot moving in 2-D; sensor trajectories with 6 DOF will require large amounts of particles for their representation. In [33], the authors also compare the advantages of separate monocular and stereo approaches in traditional SLAM frameworks.

In this paper, we show the advantages of being able to accommodate both monocular and stereo information in carrying out a 6-DOF SLAM with a handheld camera. In the works of Sola *et al.* [34] and Lemaire *et al.* [20], it is also pointed out that combining visual information at close range as well as at infinity should improve the performance of visual SLAM.

Since the initial results of [35], great progress has been made in the related problem of visual odometry [36]–[39]. Visual odometry systems have the important advantage of constant time execution. Furthermore, during exploratory trajectories, in which an environment feature is seen for a certain window of time and never more, visual odometry can obtain the same precision in the estimation of the sensor location as a SLAM system, with a great reduction in cost. Unfortunately, visual odometry does not cope with loop closings, and thus, eventual drift in these cases is inevitable. Stereo visual odometry combined with GPS can result in a mapping system that avoids long-term drift [40], [41], but unfortunately GPS is not always available. Improving the precision in sensor location through loop closing is one of the main advantages of SLAM.

An important limitation of current SLAM systems that use the standard EKF algorithm is that when mapping large environments very soon, they face computational as well as consistency problems [42]. Many efforts have been invested in reducing the $O(n^2)$ cost of the EKF updates. In [43], an information filter, the dual of the Kalman filter, was used, allowing constant time updates irrespective of the size of the map. An approximation is carried out to sparsify the information matrix, which may lead to map divergency [44]. The treemap algorithm [45] performs updates in $O(\log n)$ also by forcing information matrix sparseness by weak link breakage. In more complicated trajectories, such as lawn mowing, the cost can be more than log linear [46]. In the smoothing and mapping method [47], the authors observed that the information matrix is exactly sparse when all vehicle locations are considered in the stochastic map, and thus, very efficient techniques can be used to compute the batch solution (a recent incremental version is described in [48]).

All of these algorithms use the information form, and thus, the state and covariance are not readily available. There are alternatives that work on the covariance form, such as the map joining algorithm [49]. It works on a sequence of local maps of limited size, and thus, it can cut down the cost of EKF SLAM considerably, although remaining $O(n^2)$. It has the additional advantage of improving the consistency of the resulting estimation [42]. The divide and conquer algorithm [50] is able to compute the covariance form of the stochastic map in an amortized time linear with the size of the map, improving further the consistency of the solution. However, in these systems, local maps are required to be statistically independent. This requires creating a new local map from scratch every time the current local map size limit has been reached. Consequently, no sharing of valuable information is possible in a 6-DOF visual SLAM, such as the camera velocity, or information about features currently being tracked. This issue has been tackled in a recent work [51] by using the *conditional independence* property.

In this paper, we describe a robust and scalable 6-DOF visual SLAM system that can be carried in hand at normal walking speeds of 4–5 km/h, and used to map large indoor and outdoor environments. In Section II, we summarize the main characteristics of our system. In Section III, we describe the details of the visual SLAM system that provides the sequence of conditionally independent (CI) local maps, the basic building blocks of our mapping algorithm. This algorithm, CI divide and conquer (D&C) SLAM, is explained in Section IV. In Section V, we describe the two experiments carried out to test the system, an indoor 200 m loop and an outdoor 140 m loop. In Section VI, we discuss the results obtained, and finally, in Section VII, we draw the main conclusions of our work.

II. Our Proposal

The fundamental characteristics of the system that we describe in this paper are as follows.

1) Unlike any other visual SLAM system, we consider information from features, both close and far from the cameras. A stereo provides 3-D information from nearby scene points, and each camera can also provide bearing only information from distant scene points. Both types of information are incorporated into the map and used to improve the estimation of both the camera pose and velocity, as well as the map.

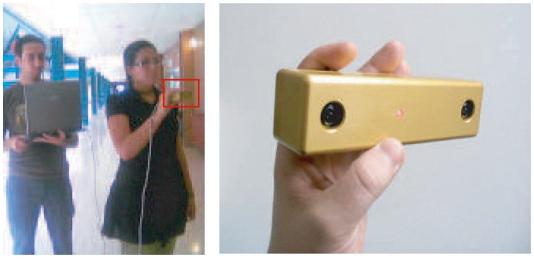


Fig. 1. Stereo vision system used to acquire the image sequences. (Left) Experimental setup during the data acquisition for the indoor experiment.

2) Nearby scene points provide scale information through the stereo baseline, eliminating the intrinsic scale unobservability problem of monocular systems.
3) We use Conditionally Independent Divide and Conquer SLAM algorithm that allows the system to maintain both camera velocity information and current feature information during local map initialization. This adds robustness to the system without sacrificing precision or consistency in any way. Being a D&C algorithm, it also allows linear time execution, enabling the system to be used for large-scale indoor/outdoor SLAM.

Our 6-DOF hardware system consists of a stereo camera carried in hand and a laptop to record and process a sequence of images (see Fig. 1). Since the camera moves in 6 DOF, we define the camera state using 12 variables: camera position in 3-D Cartesian coordinates, camera orientation in Euler angles, and linear and angular velocities. It is known that a stereo camera can provide depth estimation of points up to a certain distance determined by the baseline between left and right cameras. Therefore, two regions can be differentiated: a region close to the cameras and visible by both, in which the stereo behaves as a range and bearing sensor. The second is the region of features far from the cameras or seen by only one, in which the stereo becomes a monocular camera, providing bearing only measurements of such points. To take advantage of both types of information, we combine 3-D points and inverse depth (ID) points (introduced in [52]) in the state vector in order to build a map and estimate the camera trajectory. The system produces sequences of local maps of limited size containing both types of features using an EKF SLAM algorithm. As we detail in Section IV, these local maps are joined into a full map using the CI D&C SLAM algorithm, obtaining as final result a full stochastic map containing all tracked features, and the final and intermediate camera states from each local map. This system is highly scalable: local maps are built in constant time, regardless of the size of the environment, and the CI D&C algorithm requires amortized linear time.

During the feature tracking process, the right image is chosen as reference to initialize new features. Interest points are detected and classified according to their disparity with the left image. Those points whose disparity reveals a close distance are initialized as 3-D features, otherwise they are modeled as ID points and initialized using the bearing information obtained from the right image. When the camera moves, these features are tracked in order to update the filter and produce the corresponding corrections. To track a feature, its position is predicted in both images inside a bounded region given by the uncertainty in the camera motion and the corresponding uncertainty of the feature.

The process to select, initialize, and manage these features is detailed in the next section.

III. Visual SLAM System

A. State Representation

The state vector that represents a local submap $\mathbf{x}_B$ contains the final camera location $\mathbf{x}_c$ and the location of all features $\mathbf{x}_{f_{1:n}}$ with respect to the map base reference B, the initial camera location. Some features are codified using the *ID parametrization* that model points that are at infinity in $\mathbf{x}_{\mathrm{ID}}$. Additionally, Cartesian *3-D parametrization* is used to represent depth points in $\mathbf{x}_{\mathrm{3D}}$:

$$\mathbf{x}_B = \begin{bmatrix}\mathbf{x}_c\\ \mathbf{x}_{f_{1:n}}\end{bmatrix} = \begin{bmatrix}\mathbf{x}_c\\ \mathbf{x}_{ID}\\ \mathbf{x}_{3D}\end{bmatrix}. \tag{1}$$

The camera is described by the position of its optical center in Cartesian coordinates $\mathbf{r}$, its orientation in Euler angles Ψ, its linear velocity $\mathbf{v}$, and its angular velocity $\mathbf{w}$. In order to carry out the prediction process, the camera motion follows a constant velocity model with zero mean Gaussian noise in the linear and angular accelerations

$$\mathbf{x}_c = \begin{bmatrix}\mathbf{r}\\ \Psi\\ \mathbf{v}\\ \mathbf{w}\end{bmatrix}. \tag{2}$$

Image corners classified as depth points are transformed to 3-D points, given the disparity information provided by the stereo pair. Section III-D describes the criterion adopted to select points as depth points. Since the stereo camera provides rectified images, the backprojection equations to obtain a 3-D point are based on a pinhole camera model that relates image points and 3-D points using the following transformation function:

$$\begin{aligned}\mathbf{x}_{\mathrm{3D}} &= f(u_r, v_r, u_l, v_l)\\ &= [x, y, z]^T\\ &= \left[\frac{b(u_r - u_0)}{d}, \frac{b(v_r - v_0)}{d}, \frac{fb}{d}\right]^T\end{aligned} \tag{3}$$

where (u_r, v_r) and (u_l, v_l) are the pixels on the right and left images, and $d = (u_l - u_r)$ is the horizontal disparity. The remainder terms in the equations are the calibrated parameters of the camera, i.e., the central pixel of the image (u_0, v_0), the baseline b, and the focal length f.

Given the camera location $\mathbf{x}_{c_i}$, an ID point is defined in [52] as

$$\mathbf{x}_{\mathrm{ID}} = \begin{bmatrix}\mathbf{r}_i\\ \theta_i\\ \phi_i\\ \rho_i\end{bmatrix}. \tag{4}$$

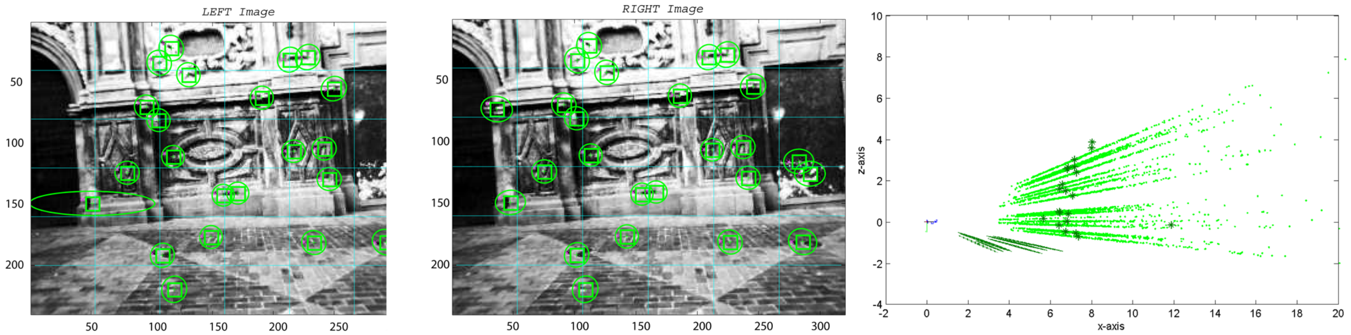


Fig. 2. Points detected using a stereo camera. Projection of map features on both left and middle images. (Right) We show feature uncertainties from a lateral perspective. 3-D feature uncertainties are drawn using darker ellipses whereas we use samples to show the ID feature uncertainties. The accompanying video `VSLAM_local_map.avi` illustrates the process of building a single local submap.

This vector depends on the optical center $\mathbf{r}_i$ of the camera from which the feature was first observed, the direction of the ray passing through the image point (i.e., azimuth θ_i, elevation ϕ_i), and the inverse of its depth, $\rho_i = 1/d_i$.

B. Selection and Management of Trackable Points

To ensure tracking stability of map features, distinctive points have to be selected. Following a similar idea as the one presented in [53], we use the Shi–Tomasi variation of the Harris corner detector to select good trackable image points and their corresponding 11×11 surrounding patch.

From the first step, the right image is split using a regular grid; the point with the best detector response per grid cell is selected (see Fig. 2). At each step, we use only those features that fall in the field of view (FOV) of the camera when they are projected along with their uncertainties on right and left images. Using the patch associated with each feature, a matching search based on normalized cross-correlation is performed inside the projected uncertainty region, as introduced in [24]. During the following steps, those cells that become and remain empty for a given time are monitorized to initialize a new feature when a good point is detected. In this way, features can be uniformly distributed in the image, improving the amount of information gathered from the scene, and therefore the map estimate. The approach is accompanied by a feature management strategy so that nonpersistent features are deleted from the state vector to avoid an unnecessary growth in population.

C. Measurement Equation

At each step, we apply the active search process described before such that, for each projected feature in the stereo image, a match is found after performing normalized cross-correlation. Thus, a new observation $\mathbf{z}$ given by the matched pixel is used to update the state of the camera and the map.

In the right camera, the equation that defines the relation between the ith ID feature $\mathbf{x}^i_{\mathrm{ID}}$ and its observation $\mathbf{z}^{r_i}_{\mathrm{ID}}$ is given by the following measurement equation:

$$\begin{aligned}\mathbf{z}^{r_i}_{\mathrm{ID}} &= h^r_{\mathrm{ID}}(\mathbf{x}_c, \mathbf{x}^i_{\mathrm{ID}}) + \upsilon \\ &= \mathrm{projection}(\ominus \mathbf{x}_c \oplus \mathbf{x}^i_{\mathrm{ID}}) + \upsilon \qquad (5)\end{aligned}$$

where h^r_{ID} is the function that projects the ID feature to the right camera and υ is a zero mean Gaussian noise with σ_p standard deviation that represents the projection error in pixels. Alternatively, we can define the measurement equation that relates the inverse point observation on the left image by

$$\begin{aligned}\mathbf{z}^{l_i}_{\mathrm{ID}} &= h^l_{\mathrm{ID}}(\mathbf{x}_c, \mathbf{x}^i_{\mathrm{ID}}) + \upsilon \\ &= \mathrm{projection}(\ominus \mathbf{x}_c \oplus \mathbf{x}_{c_r c_l} \oplus \mathbf{x}^i_{\mathrm{ID}}) + \upsilon \qquad (6)\end{aligned}$$

where the displacement of the left camera optical center with respect to the right camera is given by the rigid transformation $\mathbf{x}_{c_r c_l} = [0\ b\ 0]^T$.

In a similar way, we describe observations corresponding to 3-D map features in the right and left cameras as

$$\begin{aligned}\mathbf{z}^{r_i}_{3\mathrm{D}} &= h^r_{3\mathrm{D}}(\mathbf{x}_c, \mathbf{x}^i_{3\mathrm{D}}) + \upsilon \\ &= \mathrm{projection}(\ominus \mathbf{x}_c \oplus \mathbf{x}^i_{3\mathrm{D}}) + \upsilon \\ \mathbf{z}^{l_i}_{3\mathrm{D}} &= h^l_{3\mathrm{D}}(\mathbf{x}_c, \mathbf{x}^i_{3\mathrm{D}}) \\ &= \mathrm{projection}(\ominus \mathbf{x}_c \oplus \mathbf{x}_{c_r c_l} \oplus \mathbf{x}^i_{\mathrm{ID}}) + \upsilon.\end{aligned}$$

Note that we use $\oplus$ and $\ominus$ operators in order to denote the corresponding compositions and inversions of transformations. They represent different transformations depending on the kind of parametrization used to express a feature. In [49], the definitions for 2-D transformations were introduced, dealing mainly with point features and line features. In [54], the operations have been extended for 3-D ID and depth points. Details of the calculation of the corresponding Jacobians to propagate the uncertainties correctly can also be found in [54].

Fig. 2 shows the prediction of these 3-D ID features that fall inside the FOV of each of the cameras. A good advantage of using a stereo camera is that although a feature can disappear from the FOV of one camera, information to update the state is available if the feature can be still found in the other. As it will be shown in the experiments, this fact is of extreme importance when the camera rotates or turns around a corner, since features escape very fast from the FOV of a single camera, making the estimation of the camera location in these moments very weak.

D. Depth Points Versus ID Points

Current research on monocular SLAM has shown that the ID parametrization is suitable to represent the distribution of

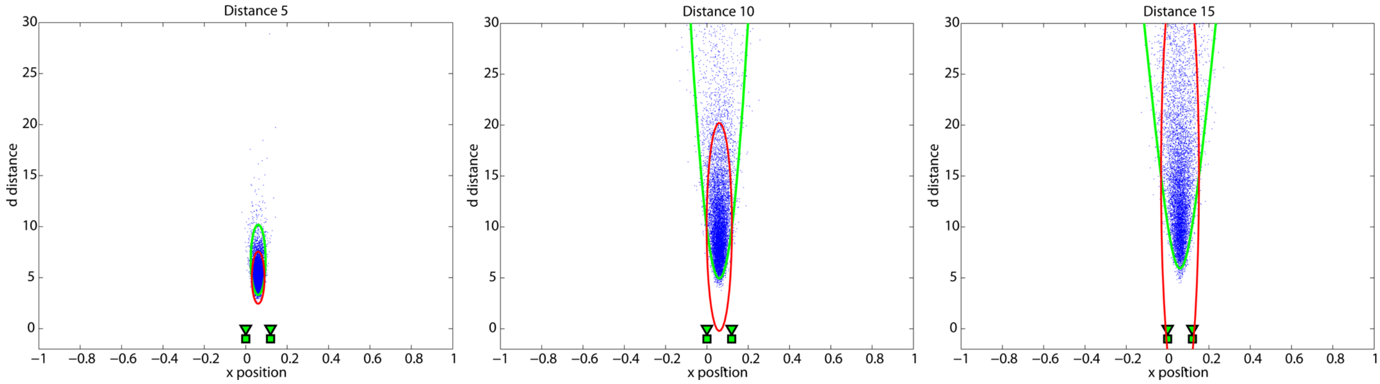


Fig. 3. Simulated experiment of a point reconstruction from a stereo pair observation for a point at (left) 5-, (middle) 10-, and (right) 15-m distance. The point clouds are samples from the real distribution of the point location, given that the pixel noise in the images is Gaussian. Dark red ellipses represent the uncertainty region for the point location when the back projection equations of a depth point are linearized. Light green regions represent the uncertainty in the point using the ID parametrization. The accompanying video `VSLAM_stereo_distribution.avi` shows the real and approximate uncertainties.

features at infinity as well as close points, allowing to perform an undelayed initialization of features. Despite its properties, each ID point needs an overparametrization of six values instead of a simpler three coordinates spatial representation [55]. This produces a computational overhead in the EKF. Working with a stereo camera, which can estimate the depth of points close to the camera, raises the subtle question of when a feature should be initialized using a 3-D or an ID representation.

In order to clarify this issue, we have designed a simulated experiment to study the effect of the linearization in both representations when a point is initialized using the stereo information. In this simulated experiment, the variance of the pixel noise ($\sigma_p = 1$ pixel) and the actual intrinsic parameters of the stereo camera used, such as the baseline, are taken into account to implement the simulation. The experimental setup consists of a stereo pair where the left camera is located at the origin of the reference frame, with its principal axis pointing along Z and the X axes pointing to the right. The right camera is at $b = 12$ cm in X. We consider a point that is in the middle between both cameras at different distances in Z. Given a noisy pixel observation, the uncertainty region of a reconstructed point is sampled and plotted in Fig. 3 for three different point distances: 5,10, and 15 m. The uncertainty region of the 3-D representation, which is calculated using a linearization of (3) and evaluated in the ground truth, is represented by the dark red ellipse. The corresponding uncertainty region of the linearized ID representation is bounded by the light gray lines in the plot. Notice that the ID parametrization models very accurately the real uncertainty for the studied distances. However, although the dark ellipse covers the real distribution at 5 m quite accurately, for longer distances, the ellipse overestimates the uncertainty in the region close to the cameras and is overconfident for far distances.

This empirical analysis suggests choosing a threshold of 5 m. A point closer than 5 m is initialized using a 3-D representation, a more distant point is parameterized as an ID point.

ID features can be transitioned to 3-D points, reducing significantly the number of DOF. Conversion requires an analysis of the linearity of the functions that model both depth point and ID point distributions. In [55], this issue is considered by using a linearity index. Such analysis makes it possible to decide when an inverse point distribution is well approximated with the overparameterized coding. Switching from ID to depth depends on a linearity threshold derived from the analysis.

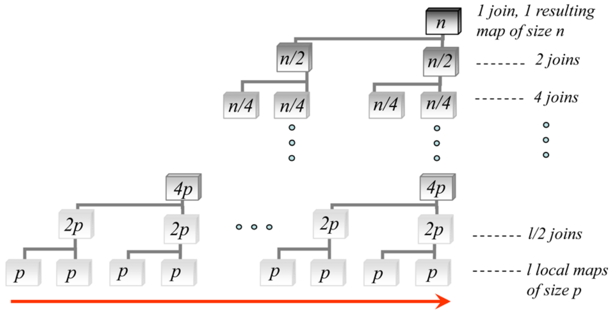


Fig. 4. Binary tree representing the hierarchy of maps that are created and joined in D&C SLAM. The red line shows the sequence in which maps are created and joined.

IV. CI D&C SLAM

D&C SLAM has proved to be a good algorithm in minimizing the computational complexity of EKF-based SLAM and improving consistency of the resulting estimate [50]. The algorithm allows us to efficiently join several local maps into a single state vector using map joining in a hierarchical tree structure (see Fig. 4). Local maps can be obtained in constant time, regardless of the size of the environment, and the map joining operations can be performed in an amortized linear time. The D&C SLAM algorithm was, however, conceived for statistically independent sequences of local maps. This requires creating a new local map from scratch every time the current local map size limit has been reached. Consequently, it is not possible to share valuable information in a 6-DOF visual SLAM, such as the camera velocity, or information about features currently being tracked.

In this section, we describe the *CI* D&C SLAM algorithm, which is able to work with maps that are not statistically independent, but rather *conditionally independent*, and thus, allow sharing of the valuable information with no increment in computational cost or loss of precision whatsoever.

A. CI Local Maps

In visual SLAM, it can be very useful to share some state vector components between consecutive submaps: some camera states, such as linear and angular velocities, as well as features that are in the transition region between adjacent submaps and are currently being tracked. This allows us to improve the estimate of relative location between the submaps and continue tracking the observed features with no interruptions. Nevertheless, special care is needed to join the submaps in a single map since their estimates are not independent anymore.

The novel technique to achieve these requirements is based on the concept of CI local maps presented in [51]. Here, we present a brief summary of the technique.

Suppose that a local map 1 has been built and we want to start a new submap 2 not from scratch, but sharing some elements in common with 1. Submap 1 is described by the following probability density function:

$$p(\mathbf{x}_A, \mathbf{x}_C | \mathbf{z}_a) = \mathcal{N}\left(\begin{bmatrix} \hat{\mathbf{x}}_{A_a} \\ \hat{\mathbf{x}}_{C_a} \end{bmatrix}, \begin{bmatrix} P_{A_a} & P_{AC_a} \\ P_{CA_a} & P_{C_a} \end{bmatrix} \right) \tag{7}$$

where $\mathbf{x}_A$ are the components of the current submap that only belong to map 1, $\mathbf{x}_C$ are the elements that will be shared with map 2, and $\mathbf{z}_a$ the observations gathered during the map construction. Notice that upper case subindices are for state vector components whereas lower case subindices describe which observations $\mathbf{z}$ have been used to obtain the estimate.

Submap 2 is then initialized with the result of marginalizing out the noncommon elements from submap 1:

$$p(\mathbf{x}_C | \mathbf{z}_a) = \int p(\mathbf{x}_A, \mathbf{x}_C | \mathbf{z}_a)\, d\mathbf{x}_A = \mathcal{N}(\hat{\mathbf{x}}_{C_a}, P_{C_a}). \tag{8}$$

During the trajectory along map 2, new observations $\mathbf{z}_b$ are gathered about the common components $\mathbf{x}_C$ as well as observations of new elements $\mathbf{x}_B$ that are incorporated into the map. When map 2 is finished, its estimate is finally described by

$$p(\mathbf{x}_C, \mathbf{x}_B | \mathbf{z}_a, \mathbf{z}_b) = \mathcal{N}\left(\begin{bmatrix} \hat{\mathbf{x}}_{C_{ab}} \\ \hat{\mathbf{x}}_{B_{ab}} \end{bmatrix}, \begin{bmatrix} P_{C_{ab}} & P_{CB_{ab}} \\ P_{BC_{ab}} & P_{B_{ab}} \end{bmatrix} \right) \tag{9}$$

where the subindices in the estimates $\hat{\mathbf{x}}_{C_{ab}}$ and $\hat{\mathbf{x}}_{B_{ab}}$ reveal that both sets of observations $\mathbf{z}_a$ and $\mathbf{z}_b$ have been used in the estimation process. This means that submap 2 is updated with all the information gathered by the sensor. But observe that map 1 in (7) has been updated with the observation $\mathbf{z}_a$ but not with the more recent observations $\mathbf{z}_b$.

Fig. 5 shows a Bayesian network that describes the probabilistic dependencies between elements of submaps 1 and 2. As it can be seen, the only connection between the set of nodes ($\mathbf{x}_A$, $\mathbf{z}_a$) and ($\mathbf{x}_B$, $\mathbf{z}_b$) is through node $\mathbf{x}_C$, i.e., both subgraphs are *d-separated* given $\mathbf{x}_C$ [56]. This implies that nodes $\mathbf{x}_A$ and $\mathbf{z}_a$ are *CI* of nodes $\mathbf{x}_B$ and $\mathbf{z}_b$ given node $\mathbf{x}_C$. Intuitively, this means that if $\mathbf{x}_C$ is known, submaps 1 and 2 do not carry any additional information about each other.

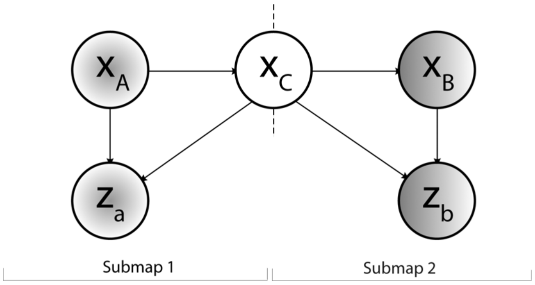


Fig. 5. Bayesian network that describes the relations between two consecutive submaps.

B. CI Map Joining

Consider two consecutive CI local maps. We are interested in joining the maps into a single stochastic map described by

$$p(\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C | \mathbf{z}_a, \mathbf{z}_b) = \mathcal{N}\left(\begin{bmatrix} \hat{\mathbf{x}}_{A_{ab}} \\ \hat{\mathbf{x}}_{C_{ab}} \\ \hat{\mathbf{x}}_{B_{ab}} \end{bmatrix}, \begin{bmatrix} P_{A_{ab}} & P_{AC_{ab}} & P_{AB_{ab}} \\ P_{CA_{ab}} & P_{C_{ab}} & P_{CB_{ab}} \\ P_{BA_{ab}} & P_{BC_{ab}} & P_{B_{ab}} \end{bmatrix} \right). \tag{10}$$

Taking into account the submap conditional independence property, it can be demonstrated [51] that the optimal map result of the joining can be computed using

$$\begin{aligned} K &= P_{AC_a} P_{C_a}^{-1} \\ &= P_{AC_{ab}} P_{C_{ab}}^{-1} && (11) \\ \hat{\mathbf{x}}_{A_{ab}} &= \hat{\mathbf{x}}_{A_a} + K(\hat{\mathbf{x}}_{C_{ab}} - \hat{\mathbf{x}}_{C_a}) && (12) \\ P_{A_{ab}} &= P_{A_a} + K(P_{CA_{ab}} - P_{CA_a}) && (13) \\ P_{AC_{ab}} &= K P_{C_{ab}} && (14) \\ P_{AB_{ab}} &= K P_{CB_{ab}}. && (15) \end{aligned}$$

Using this technique, we can build local maps that have elements in common, and then retrieve the global information in a consistent manner. After the joining, the elements belonging to the second map are transformed to the base reference of the first map.

C. Actual Implementation for Stereo

The D&C SLAM algorithm of [50] can be adapted to work with conditional independent local maps simply by using the CI map joining operation described before. As we mentioned before, since the camera moves in 6 DOF, the camera state is composed of its position using 3-D Cartesian coordinates, the orientation in Euler angles, and its linear and angular velocities. 3-D points and ID points are included as features in the state vector. When a local map $\mathbf{m}_i$ is finished, the final map estimate is given by

$$\mathbf{m}_i.\hat{\mathbf{x}} = \begin{bmatrix} \hat{\mathbf{x}}_{R_i R_j} \\ \hat{\mathbf{v}}_{R_i R_j} \\ \hat{\mathbf{x}}_{R_i F_{1:m}} \\ \hat{\mathbf{x}}_{R_i F_{m+1:n}} \end{bmatrix} \tag{16}$$

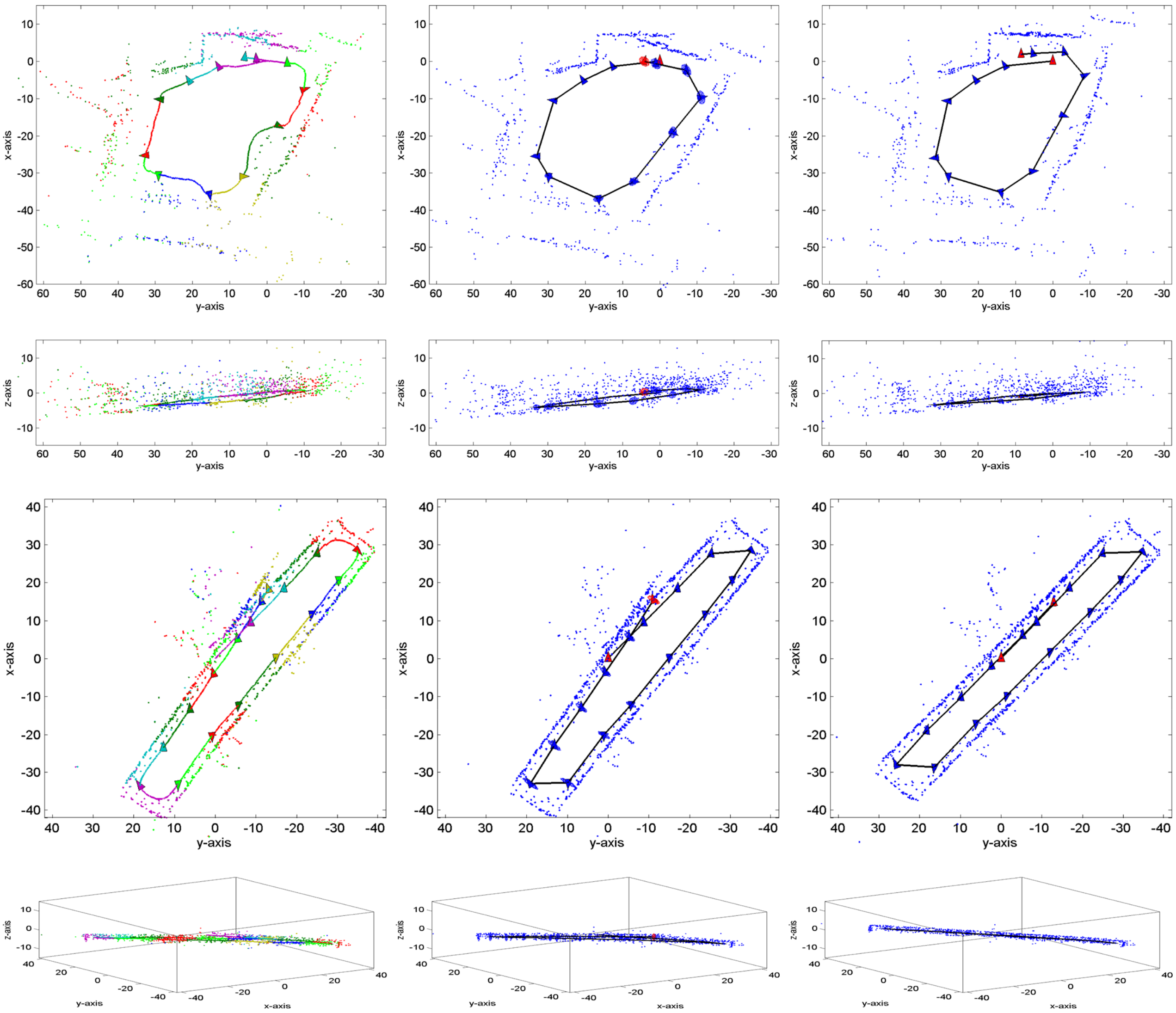


Fig. 6. (a) Outdoors experiment: 6-DOF stereo SLAM on a public square. (Top row) Both XY projection and (top-middle row) YZ projection are shown in order to illustrate the precision obtained. (b) Indoor experiment along a building environment. (Bottom-middle row) XY projection and (bottom row) YZ projection. (Left column) The sequence of CI local maps is represented with respect to the initial reference; (middle column) results obtained after running the D&C algorithm that joins and corrects the estimates; (right column) final map obtained when the loop closing constraint is imposed. The scale factor and camera positions are well recovered due to the combined observations of 3-D points and ID points. The accompanying videos `VSLAM_video_outdoor.avi` and `VSLAM_video_indoor.avi` show the full execution of the outdoor and indoor experiments.

where $\hat{\mathbf{x}}_{R_i R_j}$ is the final camera location R_j with respect to the initial one, R_i and $\hat{\mathbf{v}}_{R_i R_j}$ are the linear and angular velocities, $\hat{\mathbf{x}}_{R_i F_{1:m}}$ are 3-D and ID features that will only remain in the current map, and $\hat{\mathbf{x}}_{R_i F_{m+1:n}}$ are 3-D and ID features that will be shared with the next submap $\mathbf{m}_j$.

Since the current camera velocity $\hat{\mathbf{v}}_{R_i R_j}$ and some features $\hat{\mathbf{x}}_{R_i F_{m+1:n}}$ are used to initialize the next local map, these elements have to be computed with respect to the base reference of the second map R_j:

$$\mathbf{m}_i.\hat{\mathbf{x}} = \begin{bmatrix} \hat{\mathbf{x}}_{R_i R_j} \\ \hat{\mathbf{v}}_{R_i R_j} \\ \hat{\mathbf{x}}_{R_i F_{1:m}} \\ \hat{\mathbf{x}}_{R_i F_{m+1:n}} \\ \cdots \\ \ominus\hat{\mathbf{x}}_{R_i R_j} \oplus \hat{\mathbf{v}}_{R_i R_j} \\ \ominus\hat{\mathbf{x}}_{R_i R_j} \oplus \hat{\mathbf{x}}_{R_i F_{m+1:n}} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{x}}_{A_a} \\ \cdots \\ \hat{\mathbf{x}}_{C_a} \end{bmatrix} \qquad (17)$$

where the new elements define the common part $\hat{\mathbf{x}}_{C_a}$ and the original map defines $\hat{\mathbf{x}}_{A_a}$. Notice that the appropriate composition operation has to be applied for each transformed component and that the corresponding covariance elements have to be added to the map.

In local mapping, a base reference has to be identified to start a new map. This common reference is represented by the final vehicle position, which is the case of R_j between $\mathbf{m}_i$ and $\mathbf{m}_j$.

The initial state vector of the next submap is then given by

$$\mathbf{m}_j.\hat{\mathbf{x}} = \begin{bmatrix} \hat{\mathbf{x}}_{R_j R_j} \\ \ominus\hat{\mathbf{x}}_{R_i R_j} \oplus \hat{\mathbf{v}}_{R_i R_j} \\ \ominus\hat{\mathbf{x}}_{R_i R_j} \oplus \hat{\mathbf{v}}_{R_i R_j} \\ \ominus\hat{\mathbf{x}}_{R_i R_j} \oplus \hat{\mathbf{x}}_{R_i F_{m+1:n}} \end{bmatrix} \qquad (18)$$

where $\hat{\mathbf{x}}_{R_j R_j}$ represents the location of the camera in the new reference frame with initial zero uncertainty and zero correlation with the rest of the elements of the initial map. Notice that the initial velocity brought from the previous map has been replicated twice. One of the copies will change as the camera moves through the new map carrying the current camera velocity. The other copy will remain fixed and, together with the transformed features, will be the common elements with the previous map. The same process is successively repeated with all local maps.

D. Continuous Data Association in Each Local Map

Recent work on large environments [19] has shown that the joint compatibility test [57] helps avoiding map corruption in the visual SLAM by rejecting measurements that come from moving objects. This framework is suitable in environments with a limited number of observations. However, a branch and bound algorithm implementation of (*JCBB*) has limited use when the number of observations per step is large. In this paper, we have obtained more efficient results using the *randomized joint compatibility* version *RJC* proposed in [50], in which, in the spirit of RANSAC, a *joint compatibility (JC)* test is run with a fixed set of p randomly selected measurements. In this case, correlation between patches and individual χ^2 tests is used to obtain candidate matches. If all p measurements and their matches are jointly compatible, we apply the nearest neighbor rule to match the remaining measurements. Once a full hypothesis H is obtained, we check *JC* to avoid false positives. The process is repeated t times with adaptive RANSAC, limiting the probability of missing a correct association.

E. Map Matching

The property of sharing common elements solves the data association problem between consecutive local maps [50]. This requires us to solve data association only in loop closing situations. We use the map matching algorithm of [19] in order to detect a previously visited area. The algorithm finds correspondences between features in different local maps, taking into account the texture and the relative geometry between the features. If sufficient corresponding features are found, an ideal measurement equation that imposes the loop closing constraint is applied in the final map.

V. Experiments in Urban Outdoor and Indoor Environments

In order to demonstrate the robustness and scalability of the visual SLAM system that we propose, we have gathered two 320 × 240 image sequences with a stereo system (see Fig. 1). The system provides a 65 × 50 degree FOV per camera, and has a baseline of 12 cm, limiting the 3-D point features initialization up to a distance close to 5 m.

An indoor loop (at 48 fps) and an urban outdoor (at 25 fps) loop sequences were captured carrying the camera in hand, at normal walking speeds of 4–5 km/h. Both sequences were processed in MATLAB with the proposed algorithms on a desktop computer with an Intel 4 processor at 2.4 GHz. The higher frame rate for the indoor experiment helps in reducing the probability of mismatches given that the environment includes brick walls providing ambiguous texture information.

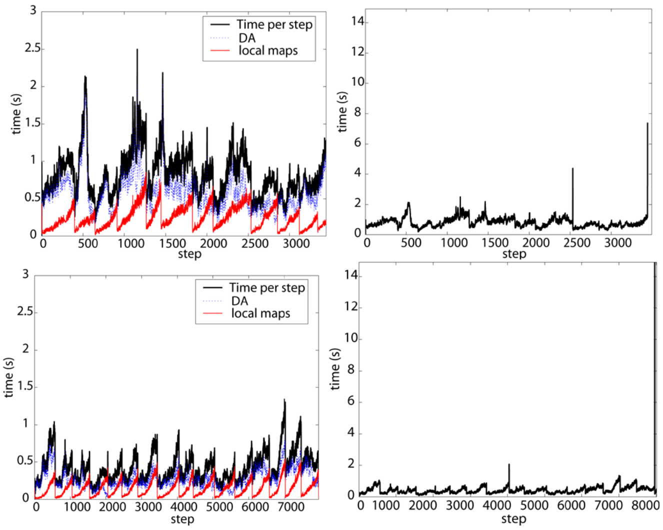


Fig. 7. Running time per step of all associated processes; a detailed analysis of (left) the features extraction, local mapping (labeled as local maps), and data association (DA) times; (right) total time per step where the peaks represent the joins performed by the CI D&C algorithm. (Top) Outdoor environment: the public square. (Bottom) Indoor environment.

The outdoor sequence is composed of 3441 stereo pairs gathered in a public square of our home town (see Fig. 6 top row). The full trajectory is approximately 140 m long from the initial camera position. Fig. 6, left column, shows the sequence of conditional independent local maps obtained with the technique described in Section IV-A. Each map contains 100 features combining ID and 3-D points. The total number of maps built during the stereo sequence is 11. The result of D&C without applying the loop closing constraint is shown in Fig. 6, middle column. As it can be observed, the precision of the map obtained is good enough to almost align the first and last submaps after all the trajectory has been traversed, even without applying loop closing constraints. Fig. 6, right column, presents the final result after closing the loop.

The second experiment was carried out inside one of our campus buildings in a walk of approximately 210 m (see Fig. 6, bottom row). The same process was run in order to obtain a full map from 8135 stereo pairs. This environment has a particular degree of difficulty due to ambiguous texture and the presence of extensive zones of glass windows such as offices, corridors, and cafeterias. This can be noticed in the long distance points estimated in some of the maps, which are actually inside offices and the cafeteria (see Fig. 6, left column). The result of CI D&C is shown in Fig. 6, middle column, and the final result after loop closing is shown in Fig. 6, right column.

Our 6-DOF SLAM system, even implemented in MATLAB, does not exceed 2 s per step, which is the worst case when building CI local maps. Fig. 7 shows how the system running time remains constant in most of the steps. Moreover, time peaks that appear when CI D&C takes place are below 8 s for the square experiment and 14 s for the indoor experiment, which are the maximum times required in the last step.

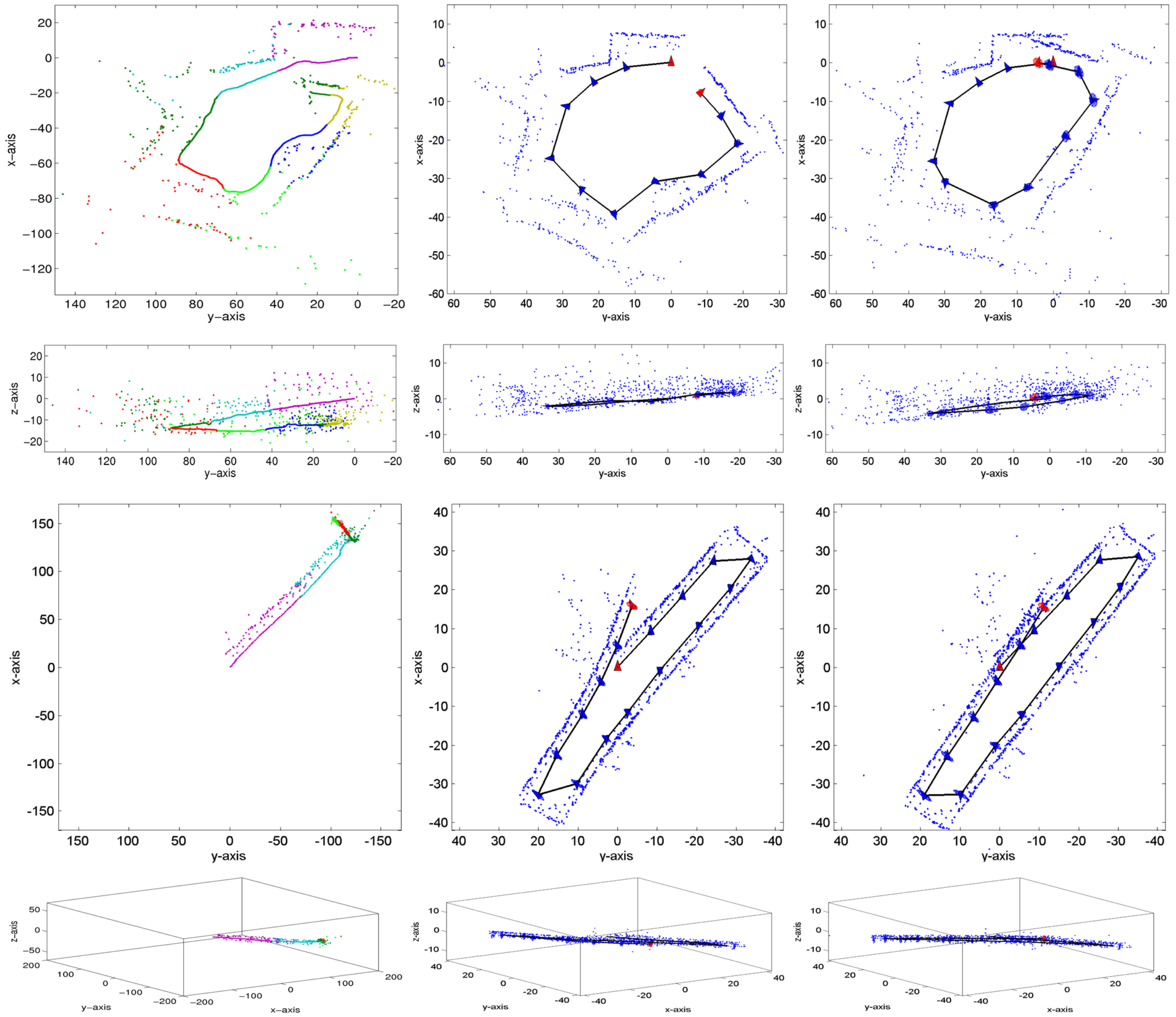


Fig. 8. Comparison of the outdoor and indoor maps obtained before the loop closure using three different techniques. (Left) Monocular SLAM with ID points, (middle) stereo SLAM with 3-D points, and (right) the proposed stereo SLAM with 3-D points and ID points.

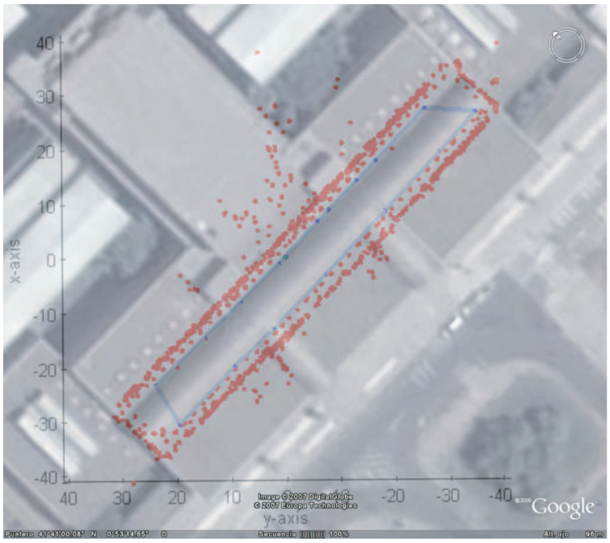
Using the Google Earth tool, we can see that the map scale obtained and the trajectory followed by the camera is very close to the real scale. Fig. 9 illustrates comparative results. We loaded the MATLAB figure in Google Earth and set the scale parameter to the real scale. Given that we had neither GPS nor compass measurements for the initial locations of the camera that are the base reference of each map, the position and orientation of the figure over the map were adjusted by hand. It can be noticed that angles between the square sides and the shape of the walls of the surrounding environment have been captured with precision.

VI. Discussion

As presented in Section I, several works have demonstrated successful visual SLAM systems in small environments using monocular or stereo cameras. There are several important factors that limit the extension of these results to large-scale environments.

First, the computational complexity and consistency of the underlying SLAM technique. In this paper, we have presented a novel algorithm that builds CI local maps in constant time and combines them in an optimal way in amortized linear time. Although the experiments presented here were processed in MATLAB, we expect that the extension to stereo of our current real-time implementation [19] will be able to build local maps up to 100 features in real time, with updates at 25 Hz. The D&C map joining, loop detection, and loop closing can be implemented on a separate thread, taking advantage of current multiple core processors.

In the case of monocular SLAM, another important limiting factor is the intrinsic unobservability of the scale. This problem can be addressed using additional sensors such as the vehicle



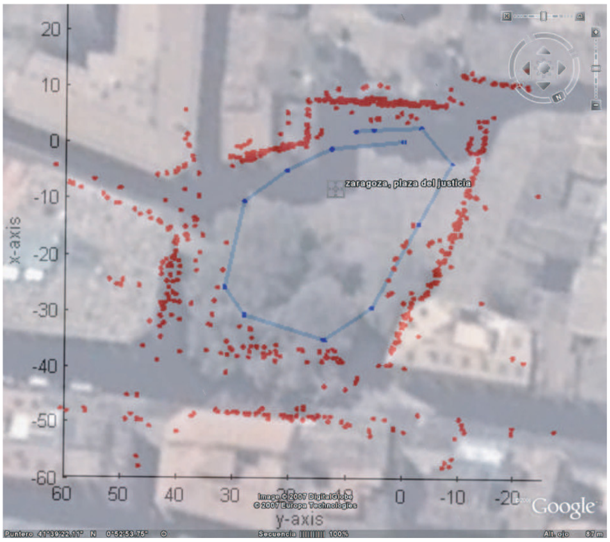

20
10
0
-10
-20
-30
-40
-50
-60
x-axis
60 50 40 30 0 -10 -20
y-axis
zaragoza, plaza del justicia
Image © 2007 DigitalGlobe
© 2007 Europa Technologies
Google

Fig. 9. Stereo visual SLAM recovers the true scale. (Top) Building environment and (bottom) the public square overlapping Google Earth.

odometry, GPS, or inertial units. When they are not available, the scale can be initialized using some *a priori* knowledge about the environment such as the size of a known object visible at the start [16] or the initial speed of the camera. However, in large environments, unless scale information is injected on the system periodically, the scale of the map can slowly drift (see, for example, the experiments in [19]). Another critical issue appears when the scene is mostly planar and perpendicular to the optical axis. In this situation, with a monocular camera, it is very difficult to distinguish between camera translation and rotation, unless a wide FOV is used.

To illustrate these difficulties, we have processed our indoor and outdoor experiments using only the information from the right camera. As we are now using a bearing only system, all the features are initialized using the ID representation. To bootstrap the system, we have introduced a initial estimated speed for the camera of 1 m/s. Apart from that, our visual SLAM algorithm remains unchanged. The resulting maps are represented in the left column of Fig. 8. As it can be seen, the scale obtained by the system drifts (compare the beginning of the loop with the end). Also, in the outdoor experiment, at a certain point, the system misinterprets the camera translation as a rotation, and the map gets corrupted. Here, we are using a camera with FOV of 65°. The results obtained in the same environment with an FOV of 90° are significantly more robust [51]. In the indoor experiment with a monocular camera, as the objects are much closer to the camera, most of the features disappear fast from the FOV when the camera turns, leading to a bad estimation of its position and consequently divergence in the map estimate.

We have also processed the sequences with our SLAM algorithm using conventional stereo, i.e., changed to initialize all the features whose disparity is larger than one pixel as 3-D points. Features without disparity are discarded because its depth cannot be computed by stereo. The immediate benefit is that the true environment scale is observable and the map corruption disappears (Fig. 8, middle column). However, for points that are more than 10 m away from the camera, a Gaussian in xyz is a bad approximation for its true uncertainty. This is the reason for the map deformation that is clearly visible in the lower part of the outdoor experiment, where many features are at about 20 m from the camera.

The proposed system (Fig. 8, right column) combines the advantages of stereo and bearing only vision. On the one hand, the true scale is precisely obtained due to the 3-D information obtained by the stereo camera from close point features. On the other hand, the region with useful point features extends up to infinity due to the ID representation developed for bearing-only SLAM. The depth of the features that are far from the camera can be precisely recovered by the system if they are seen from viewpoints that are separated enough. In that case, they can be upgraded to 3-D points for better efficiency [55]. Otherwise, they remain as ID points and still provide very valuable orientation information that improves map precision and keeps the SLAM system stable when few close features are observed.

VII. Conclusion

In this paper, we have shown that 6-DOF visual mapping of large environments can be efficiently and accurately carried out using a stereo camera as the only sensor. One of the contributions of the paper is that information from features nearby and far from the cameras can be simultaneously incorporated to represent the 3-D structure more precisely. Using close points provides scale information through the stereo baseline avoiding “scale-drift,” while ID points are useful to obtain angular information from distant scene points.

Another contribution of the paper is the combination of two recent local mapping techniques to improve consistency and reduce complexity in the SLAM process. Using CI local maps [51], our system is able to properly share information related to the camera motion model and common features between consecutive maps. Smoother transitions from map to map are achieved as well as better relative locations between local maps. By means of the simplicity and efficiency of the CI D&C SLAM algorithm, we can recover the full map very efficiently. The combination of both techniques adds robustness to the process without sacrificing precision.

In [50], we describe the performance of D&C SLAM when the vehicle carries out different types of trajectories. For some trajectories, the cost of map joining can increase at some steps, depending of the size of the overlap between the maps to be joined: doing exploration, the overlap is constant and the cost of map joining is small, when completing a loop traversal for a second time the overlap between the maps is total and the cost of joining will be much higher. Although we are able to close large indoor and outdoor loops, the algorithm used for loop closing strongly depends on detecting sets of features already stored in the map when the same area is revisited. It would be interesting to analyze other types of algorithms for loop closing, for instance, the image to map algorithm proposed in [58].

Moreover, as we assume smooth motions, the relocation algorithm presented in [58] would enable the system to avoid failures in case of jitter.

There is also a restriction of the system to estimate pitch orientation due to the use of Euler angles. A combined solution using quaternions can mitigate the problem. This will be part of our future research.

Apart from upward looking cameras and jitter, there are no limitations to manoeuver the camera freely: it can be used in environments that include stairs and other terrain accidents. This kind of experiment will be part of the evaluation process for future work.

We will also focus on comparing our system with other stereo vision techniques such as visual odometry. We are very interested in studying the fusion of the stereo camera with other sensors like GPS or inertial systems in order to compare the precision obtained. We will consider other types of feature detectors as well, and their effect in the final result.

Acknowledgment

The authors would like to thank J. M. M. Montiel and J. Civera for the fruitful discussions.

References

[1] L. M. Paz, P. Piniés, J. D. Tardós, and J. Neira, “6DOF SLAM with stereo camera in hand,” presented at the Conf. Vis. SLAM—Emerging Technol., IROS 2008, San Diego, CA, 2008.

[2] J. Folkesson and H. Christensen, “Graphical SLAM for outdoor applications,” *J. Field Robot.*, vol. 23, no. 1, pp. 51–70, 2006.

[3] A. Nüchter, K. Lingemann, J. Hertzberg, and H. Surmann, “6-D SLAM-3D mapping outdoor environments: Research Articles,” *J. Field Robot.*, vol. 24, no. 8/9, pp. 699–722, 2007.

[4] L. Ellekilde, “Dense 3-D map construction for indoor search and rescue,” *J. Field Robot.*, vol. 24, no. 1/2, pp. 71–89, Feb. 2007.

[5] J. Kim and S. Sukkarieh, “Airborne simultaneous localisation and map building,” in *Proc. IEEE Int. Conf. Robot. Autom., (ICRA 2003)*, Sep. 14–19, vol. 1, pp. 406–411.

[6] M. Bryson and S. Sukkarieh, “Building a robust implementation of bearing-only inertial SLAM for a UAV,” *J. Field Robot.*, vol. 24, no. 1/2, pp. 113–143, 2007.

[7] M. Deans and M. Hebert, “Experimental comparison of techniques for localization and mapping using a bearing-only sensor,” in *Proc. Int. Symp. Exp. Robot., (ISER 2000).*, Lecture Notes in Control and Information Science, vol. 271, S. S. D. Rus, Ed. Honolulu, HI: Springer-Verlag, 2000, pp. 395–404.

[8] T. Fitzgibbons and E. Nebot, “Bearing only SLAM using colour-based feature tracking,” presented at the 2002 Aust. Conf. Robot. Autom., Auckland, New Zealand, 2002.

[9] T. Bailey, “Constrained initialisation for bearing-only SLAM,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA 2003)*, Sep. 14–19, vol. 2, pp. 1966–1971.

[10] N. Kwok and G. Dissanayake, “An efficient multiple hypothesis filter for bearing-only SLAM,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., (IROS 2004)*, Sep. 28–Oct. 2, vol. 1, pp. 736–741.

[11] N. Kwok, G. Dissanayake, and Q. Ha, “Bearing-only SLAM using a SPRT based Gaussian sum filter,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA 2005)*, Apr. 12–22, 2005, pp. 1109–1114.

[12] J. Sola, A. Monin, M. Devy, and T. Lemaire, “Undelayed initialization in bearing only SLAM,” in *Proc. IEEE/RSJ Int. Con. Intell. Robots Syst. (IROS 2005)*, Aug. 21–26, 2005, pp. 2499–2504.

[13] T. Lemaire, S. Lacroix, and J. Sola, “A practical 3D bearing-only SLAM algorithm,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS 2005)*, Aug. 2–6, 2005, pp. 2449–2454.

[14] P. Jensfelt, D. Kragic, J. Folkesson, and M. Bjorkman, “A framework for vision based bearing only 3-D SLAM,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA 2006)*, May 15–19, 2006, pp. 1944–1950.

[15] A. Gil, O. Reinoso, O. Martínez-Mozos, C. Stachniss, and W. Burgard, “Improving data association in vision-based SLAM,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS 2006)*, Beijing, China, Oct. 2006, pp. 2076–2081.

[16] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, “MonoSLAM: Real-time single camera SLAM,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1052–1067, Jun. 2007.

[17] J. Folkesson, P. Jensfelt, and H. Christensen, “Vision SLAM in the measurement subspace,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA 2005)*, Apr. 18–22, 2005, pp. 30–35.

[18] P. Smith, I. Reid, and A. Davison, “Real-time monocular SLAM with straight lines,” in *Proc. Brit. Mach. Vis. Conf.*, 2006, vol. 1, pp. 17–26.

[19] L. Clemente, A. J. Davison, I. D. Reid, J. Neira, and J. D. Tardós, “Mapping large loops with a single hand-held camera,” in *Proc. Robotics: Sci. Syst.*, Atlanta, GA, Jun. 2007.

[20] T. Lemaire and S. Lacroix, “SLAM with panoramic vision,” *J. Field Robot.*, vol. 24, no. 1/2, pp. 91–111, 2007.

[21] T. Goedemé, M. Nuttin, T. Tuytelaars, and L. Van Gool, “Omnidirectional vision based topological navigation,” *Int. J. Comput. Vis.*, vol. 74, no. 3, pp. 219–236, 2007.

[22] E. Royer, M. Lhuillier, M. Dhome, and J. Lavest, “Monocular vision for mobile robot localization and autonomous navigation,” *Int. J. Comput. Vis.*, vol. 74, no. 3, pp. 237–260, 2007.

[23] A. Davison, “Mobile robot navigation using active vision,” Ph.D. dissertation, Univ. Oxford, Oxford, U.K., 1998.

[24] A. J. Davison and D. W. Murray, “Simultaneous localization and map-building using active vision,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 865–880, Jul. 2002.

[25] A. Davison and N. Kita, “3-D simultaneous localisation and map-building using active vision for a robot moving on undulating terrain,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn. (CVPR 2001)*, vol. 1, pp. I-384–I-391.

[26] L. Iocchi, K. Konolige, and M. Bajracharya, “Visually realistic mapping of a planar environment with stereo,” in *Proc. Int. Symp. Exp. Robot. (ISER’00)*, pp. 521–532.

[27] S. Se, D. Lowe, and J. Little, “Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks,” *Int. J. Robot. Res.*, vol. 21, no. 8, pp. 735–758, 2002.

[28] I. Jung and S. Lacroix, “High resolution terrain mapping using low altitude aerial stereo imagery,” in *Proc. 9th Int. Conf. Comput. Vis.*, Nice, France, Oct. 13–16, 2003, vol. 2, pp. 946–951.

[29] E. Hygounenc, I. Jung, P. Soueres, and S. Lacroix, “The autonomous blimp project of LAAS-CNRS: Achievements in flight control and terrain mapping,” *Int. J. Robot. Res.*, vol. 23, no. 4, pp. 473–511, 2004.

[30] J. Saez, F. Escolano, and A. Penalver, “First Steps towards Stereo-based 6DOF SLAM for the visually impaired,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. (CVPR’05)-Workshops*, vol. 3, Washington, DC: IEEE Computer Society, Jun. 20–26, 2005, pp. 23–23.

[31] R. Sim, P. Elinas, M. Griffin, and J. Little, “Vision-based SLAM using the Rao–Blackwellised particle filter,” in *Proc. IJCAI Workshop Reason. Uncertainty Robot. (RUR)*, Edinburgh, U.K., 2005, pp. 9–16.

[32] R. Sim, P. Elinas, and J. Little, “A study of the Rao–Blackwellised particle filter for efficient and accurate vision-based SLAM,” *Int. J. Comput. Vis.*, vol. 74, no. 3, pp. 303–318, 2007.

[33] T. Lemaire, C. Berger, I. Jung, and S. Lacroix, "Vision-based SLAM: Stereo and monocular approaches," *Int. J. Comput. Vis.*, vol. 74, no. 3, pp. 343–364, 2007.

[34] J. Sola, A. Monin, and M. Devy, "BiCamSLAM: Two times mono is more than stereo," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA 2007)*, Rome, Italy, Apr. 10–14, 2007, pp. 4795–4800.

[35] Z. Zhang and O. Faugueras, "Three-dimensional motion computation and object segmentation in a long sequence of stereo frames," *Int. J. Comput. Vis.*, vol. 7, no. 3, pp. 211–241, 1992.

[36] N. Simond and P. Rives, "Trajectography of an uncalibrated stereo rig in urban environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS 2004).*, Sep. 28–Oct. 2, 2004, vol. 4, pp. 3381–3386.

[37] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry for ground vehicle applications," *J. Field Robot.*, vol. 23, no. 1, pp. 3–20, 2006.

[38] A. Comport, E. Malis, and P. Rives, "Accurate quadri-focal tracking for robust 3-D visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom., (ICRA 2007)*, Roma, Italy, Apr. 10–14, 2007, pp. 40–45.

[39] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the mars exploration rovers," *J. Field Robot.*, vol. 24, no. 3, pp. 169–186, 2007.

[40] M. Agrawal and K. Konolige, "Real-time localization in outdoor environments using stereo vision and inexpensive GPS," in *Proc. Int. Conf. Pattern Recogn. (ICPR 2006)*, Hong Kong, vol. 3, pp. 1063–1068.

[41] K. Konolige, M. Agrawal, R. Bolles, C. Cowan, M. Fischler, and B. Gerkey, "Outdoor mapping and navigation using stereo vision," presented at the Int. Symp. Exp. Robot., Rio de Janeiro, Brazil, Jul. 2006.

[42] J. Castellanos, R. Martinez-Cantin, J. Tardós, and J. Neira, "Robocentric map joining: Improving the consistency of EKF-SLAM," *Robot. Auton. Syst.*, vol. 55, no. 1, pp. 21–29, Jan. 2007.

[43] S. Thrun, Y. Liu, D. Koller, A. Y. Ng, Z. Ghahramani, and H. Durrant-Whyte, "Simultaneous localization and mapping with sparse extended information filters," *Int. J. Robot. Res.*, vol. 23, no. 7/8, pp. 693–716, 2004.

[44] R. Eustice, M. Walter, and J. Leonard, "Sparse extended information filters: Insights into sparsification," in *Proc. IEEE/RJS Int. Conf. Intell. Robots Syst.*, Edmonton, AB, Canada, 2–6, Aug. 2005, pp. 3281–3288.

[45] U. Frese, *Treemap: An o(logn) Algorithm for Simultaneous Localization and Mapping*. New York: Springer-Verlag, 2005, pp. 455–476, ch. Spatial Cognition IV.

[46] U. Frese, "Efficient 6DOF SLAM with treemap as a generic backend," in *Proc. IEEE Int. Conf. Robot. Autom.* Apr. 10–14, 2007, pp. 4814–4819.

[47] F. Dellaert and M. Kaess, "Square root SAM: Simultaneous localization and mapping via square root information smoothing," *Int. J. Robot. Res.*, vol. 25, no. 12, pp. 1181–1203, Dec. 2006.

[48] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Fast incremental smoothing and mapping with efficient data association," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA 2007)*, Roma, Italy, Apr. 10–14, 2007, pp. 1670–1677.

[49] J. D. Tardós, J. Neira, P. M. Newman, and J. J. Leonard, "Robust mapping and localization in indoor environments using sonar data," *Int. J. Robot. Res.*, vol. 21, no. 4, pp. 311–330, 2002.

[50] L. M. Paz, J. D. Tardós, and J. Neira, "Divide and conquer: EKF SLAM in $O(n)$," *IEEE Trans. Robot.*, vol. 24, no. 5, Oct. 2008.

[51] P. Piniés and J. D. Tardós, "Large scale SLAM building conditionally independent local maps: Application to monocular vision," *IEEE Trans. Robot.*, vol. 24, no. 5, Oct. 2008.

[52] J. Civera, A. J. Davison, and J. M. M. Montiel, "Inverse depth parametrization for monocular SLAM," *IEEE Trans. Robot.*, vol. 24, no. 5, Oct. 2008.

[53] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proc. Int. Conf. Comput. Vis.*, Nice, France, Oct. 13–16, 2003, vol. 2, pp. 1403–1410.

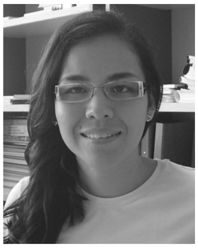
[54] L. M. Paz, P. Piniés, J. D. Tardós, and J. Neira, "Measurement equation for inverse depth points and depth points," Dept. Inf. e Ingeniería de Sistemas, Univ. de Zaragoza, Zaragoza, Spain, Internal Rep. RR-08-06, 2008.

[55] J. Civera, A. Davison, and J. Montiel, "Inverse depth to depth conversion for monocular SLAM," in *Proc. IEEE Int. Conf. Robot. Autom.*, Roma, Italy, Apr. 10–14, 2007, pp. 2778–2783.

[56] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[57] J. Neira and J. D. Tardós, "Data association in stochastic mapping using the joint compatibility test," *IEEE Trans. Robot. Autom.*, vol. 17, no. 6, pp. 890–897, Dec. 2001.

[58] B. Williams, P. Smith, and I. Reid, "Automatic relocalisation for a single-camera simultaneous localisation and mapping system," in *Proc. IEEE Int. Conf. Robot. Autom.*, Roma, Italy, Apr. 10–14, 2007, pp. 2784–2790.



Lina M. Paz (M'08) was born in Cali, Colombia, in 1980. She received the M.S. degree in electronic engineering from the Universidad del Valle, Cali, in 2003. Since 2004, she has been working toward the Ph.D. degree in computer science with the Department of Computer Science and Systems Engineering, University of Zaragoza, Zaragoza, Spain.

Her current research interests include mobile robotics, computer vision for environment modeling, and simultaneous localization and mapping (SLAM).

Pedro Piniés (M'08) was born in Bilbao, Spain, in 1979. He received the M.S. degree in telecommunication engineering in 2004 from the University of Zaragoza, Zaragoza, Spain, where he is currently working toward the Ph.D. degree with the Robotics, Perception, and Real Time Group.

His current research interests include simultaneous localization and mapping (SLAM), mobile robotics, computer vision, and probabilistic inference.

Juan D. Tardós (M'05) was born in Huesca, Spain, in 1961. He received the M.S. and Ph.D. degrees in electrical engineering from the University of Zaragoza, Zaragoza, Spain, in 1985 and 1991, respectively.

He is currently a Full Professor with the Department of Computer Science and Systems Engineering, University of Zaragoza, where he is in charge of courses in robotics, computer vision, and artificial intelligence. His current research interests include simultaneous localization and mapping (SLAM), perception, and mobile robotics.

José Neira (M'07) was born in Bogotà, Colombia, in 1963. He received the M.S. degree from the Universidad de los Andes, Bogotá, in 1986 and the Ph.D. degree from the University of Zaragoza, Zaragoza, Spain, in 1993, both in computer science.

He is currently an Associate Professor with the Department of Computer Science and Systems Engineering, University of Zaragoza, where he teaches courses in compiler theory, computer vision, and mobile robotics. His current research interests include autonomous robots, data association, and environment modeling.