Stereo vision-specific models for Particle Filter-based SLAM

F.A. Moreno *, J.L. Blanco, J. Gonzalez
University of Malaga

Abstract

This work addresses the SLAM problem for stereo vision systems under the unified formulation of particle filter methods. In contrast to most existing approaches to visual SLAM, the present method does not rely on restrictive smooth camera motion models, but on computing incremental 6D pose differences from the image flow through a probabilistic visual odometry method. Moreover, our observation model, which considers both the 3D positions and the SIFT descriptors of the landmarks, avoids explicit data association between the observations and the map by marginalizing the observation likelihood over all the possible associations. We have experimentally validated our research with two experiments in indoor scenarios.

Key words: computer vision, stereo vision, SLAM, robot localization, particle filters

1. Introduction

Vision systems have acquired growing importance in mobile robotics during the last years due to their low cost and the rich information that cameras provide in comparison with traditional robotic sensors, like laser scanners or sonars. Vision-based systems are employed in a wide range of robotic applications such as object recognition [4,5,28], obstacle avoidance [20,33,44], navigation [38,39], topological global localization [25,52] and, more recently, in simultaneous localization and mapping (SLAM) [10,41,48], which has become a prominent research area in mobile robotics since the early nineties.

SLAM is one of the most challenging open problems for developing truly autonomous robots. It can be stated as the problem of a robot building a map of an unknown environment while simultaneously tracking its position using the partially built map. Most of the approaches to SLAM have been proposed for two kind of sensory data: raw range scans [30,15], and features (or *landmarks*) either extracted from scans [12,36] or from images, having in this case the so-called visual SLAM [8,10,27,41,46,48].

In visual SLAM, the inherent errors in the formation of images and in the detection of features introduce uncertainty in the observed landmarks, which must be properly managed by means of probabilistic Bayesian filtering, extensively discussed elsewhere [35,51]. The underlying principle of those filters is the Bayes' rule, which states how to update a prior belief about a variable x given a new observation z and an observation model:

$$\underbrace{p\left(x|\, z\right)}_{\text{posterior}} \propto \underbrace{p\left(x\right)}_{\text{prior}}\ \underbrace{p\left(z|x\right)}_{\text{obs.model}} \tag{1}$$

This filter can be implemented for on-line operation by iteratively executing prediction and update steps. In the prediction, the system state (comprising the robot pose) is propagated in time according to some given transition model (the motion model), leading to the *prior distribution* of the system state. In the update step, this prior is refined according to a given observation model, obtaining the *posterior probability density*, which already includes the information available up to some given time step. The choice of suitable motion and observation models means a cornerstone in the development of robust probabilistic SLAM approaches.

Two widely employed implementations of Bayesian filters are the Extended Kalman Filter (EKF) [22], and the family of sequential Monte Carlo (SMC) methods, or *particle filters* (PFs) [1]. EKF is the mainstream approach for SLAM [9,12], but it is limited by the assumption of Gaussianity in the state and observations. This makes EKF specially inappropriate for global localization. On the other hand, PFs can cope with complex, non-parametric and even multi-modal distributions. With the introduction of

* Corresponding author.
Email addresses: famoreno@isa.uma.es (F.A. Moreno), jlblanco@ctima.uma.es (J.L. Blanco), jgonzalez@ctima.uma.es (J. Gonzalez).

the Rao-Blackwellised Particle Filters (RBPFs) [37], SMC methods have become a unified framework for SLAM and global localization.

This article addresses the problems of SLAM and global localization for stereo vision systems. Our contribution consists in providing appropriate probabilistic models for the motion and observations of a stereo camera, both suitable for PF methods. Our proposal takes a sequence of stereo images as the only input, and does neither rely on any other sensory data (odometers, IMU, etc) nor assume a priori knowledge about the camera movement. More concretely, the novelties with regard to each of these models are:

– The robot ego-motion estimation is addressed by performing 6D visual odometry through a reliable 3D landmark registration method which models the uncertainty in the pose increment estimation as a Gaussian. This motion model provides three main advantages in relation to other proposals: i) it is applicable to either, wheeled and not-wheeled robots that navigate on any type of surface (even flying robots), ii) it does not suffer from systematic errors, and iii) it overcomes the divergence and local-minima problems which suffer the iterative approaches to visual odometry, as well as the need for an initial estimation. Moreover, the method is efficient and it may be applied between small incremental poses, hence the Gaussianity assumption justifies.
– The observation model avoids explicit data association by applying marginalization over all the possible associations, thus discarding the possibility of incorrect correspondences between the observed landmarks and the map.

Our approach has been validated by experiments with a real robot that has been driven within an office-like scenario where it builds a map of the environment and tracks its position simultaneously. To test the performance of our method, the estimated path has been compared with that computed from a ICP approach which employs as input the data provided by a scanner laser. In addition, a visual SLAM experiment with a camera describing fully unconstrained 6 DoF movements is also presented.

Next we present a survey on works related to the topic of visual SLAM. In section 3 we state the SLAM problem, while the notation and the landmark extraction process employed in this work are introduced in section 4. Then, we present a visual odometry approach for the probabilistic motion model. Section 6 describes our proposal for the observation model. Some experimental results are presented in section 7 and, finally, we provide some conclusions and future work.

2. Related Research

A number of works in the technical literature have addressed robot localization and SLAM using vision sensors, including omnidirectional, monocular, stereo, and trinocular cameras.

In [32] an omnidirectional camera is used to estimate the distance of the closest color transition in the environment, mimicking laser rangefinders performance. These measurements are introduced into a particle filter to determine the position of the robot within a previously constructed map. Tamini *et al.* [50] also present an omnidirectional camera-based global localization approach for mobile robots using a modified version of the SIFT features that decreases the number of detected points and, therefore, the computation time of the localization process. The work in [11] presents a vision-based robot localization approach with just one camera which obtains a visual map of the ceiling and localizes the robot using a simple scalar brightness measurements as input. The robot localization within the map is carried out by a particle filter-based algorithm. In [54], an image retrieval system based on invariant features is combined with particle filter-based localization. These approaches only address global localization and do not deal with SLAM.

The SLAM problem is tackled in the paper series [8–10] using a single camera (called MonoSLAM). The proposed method, which performs in real time, extracts a reduced but enough number of salient image features through the operator of Shi and Tomasi [47], which are identified by their associated image patches. The scale factor, which represents one of the main limitations of monocular SLAM, is resolved by initializing the system looking at a pattern of known size. Other monocular SLAM approaches [6] introduce the inverse depth parametrization for the undelayed initialization of features. Similarly, the work in [42] adds an Inertial Measurement Unit (IMU) to an implementation of the inverse depth-based monocular SLAM, reporting an improved accuracy in the estimation of the scale factor of the map. A RBPF-based method for performing monocular SLAM is reported in [27], which extracts SIFT features from the images and applies an Unscented Kalman Filter (UKF) within the robot localization algorithm to sample new particles poses as well as to update the observed landmarks. Typically, monocular approaches to SLAM employ motion models which assume smooth paths for the camera by restricting its velocities and accelerations. In addition, they suffer from ambiguity when estimating small displacements and rotations of the camera.

Stereo and trinocular systems elude the above-mentioned problems by exploiting the special characteristics of the epipolar geometry to directly extract 3D information from the detected features in the images. Hence, these camera configurations are widely extended in vision-based systems for robot localization and SLAM [7,14,46,48,49].

A trinocular camera is employed in [46] to address SLAM by tracking SIFT visual features [29] in unmodified environments. The ego-motion estimation is computed from the robot odometry (as an initial estimation) and a least-squares procedure that finds the camera movement with the best alignment between the observed and the predicted image coordinates of the 3D landmarks in the map. The iterative nature of this method contrasts with our closed-form solution to perform visual odometry, while the usage

of the encoder-based robot odometry as initial estimation restricts their method to wheeled robots. The spatial uncertainty of the landmarks in the map is modeled by a Kalman filter.

In [7] it is also proposed a trinocular SLAM system which uses 3D line segments as the elements of the map (instead of point features). They approximate the distribution of the robot pose with a particle filter and model the uncertainty in the 3D segments of the map with a Gaussian distribution which is updated over time with an EKF. An experiment in a simulated environment is presented to validate the results of this approach. Although it is an interesting variation of the traditional approaches, its application is limited to environments where straight lines can be easily found. Since these approaches employ the information provided by three images at each time step, there exists an improvement in the robustness of the matching process, but, on the other hand, the computational burden of the method increases, which becomes a significant problem in visual SLAM.

The works in [48] and [49] extract SIFT features from stereo images and compute their 3D correspondent points in space, which are taken as landmarks for a map built through a RBPF. In their approach, the weights of the particles are computed from the distance between the positions of the observed landmarks and the predicted positions (based on the particles pose) of their corresponding landmarks in the map. The matches are determined by computing the Euclidean distance between a reduced 36-D version of SIFT descriptors of the 3D landmarks. In that work, the motion model is based on an iterative Levenberg-Marquardt non-linear optimization algorithm which minimizes the re-projection error of the 3D coordinates of the landmarks on the images. Similarly, another RBPF approach which also employs SIFT features is presented in [14] as a vision-based solution to SLAM. In this case, the motion model relies on the robot odometry, while the observation model is derived from the Mahalanobis distance between the positions of the observed and mapped landmarks. Data association is determined from the Mahalanobis distance between the SIFT descriptors of the 3D landmarks assuming independence between the elements of the 128D SIFT vector.

A key issue of trinocular and stereo vision-based proposals is solving correspondences between features in the map and those being observed by the robot. Compared to these approaches, the present work avoids explicit data association by marginalizing the observation model over all the possible associations, hence avoiding potential incorrect associations between mapped and observed landmarks. Furthermore, unlike encoder-based odometry and unlike iterative algorithms for visual odometry [46,48,49], we propose a motion model based on a closed-form visual odometry algorithm which performs in 6D and relies only on the stereo images gathered by the camera.

3. The Particle Filter Approach to SLAM

Let x_t, u_t and z_t be the robot pose, the action, and the observation at time step t, respectively, and let m be the map of the environment. The aim of the full SLAM problem [51] is to estimate the joint distribution of both the robot path and the map, i.e. to compute $p\left(x_{1:t}, m|z_{1:t}, u_{1:t}\right)$ where $z_{1:t} = \{z_1, \ldots, z_t\}$, $u_{1:t} = \{u_1, \ldots, u_t\}$ and $x_{1:t} = \{x_1, \ldots, x_t\}$. In this work, u_t represents the robot pose change between time steps $t-1$ and t, which, in our case, is unknown and will be estimated by means of visual odometry.

A way of efficiently dealing with the high dimensionality of the system state in SLAM, is to employ a Rao-Blackwellized particle filter, which reduces the complexity of the estimation problem by sampling over a subset of the state variables. Thus, by factoring $p\left(x_{1:t}, m|z_{1:t}, u_{1:t}\right)$ we can sample the distribution of the possible robot paths and compute the map distribution from those samples [13,51]:

$$p\left(x_{1:t}, m|z_{1:t}, u_{1:t}\right) = \underbrace{p\left(x_{1:t}|z_{1:t}, u_{1:t}\right)}_{\text{robot path}} \underbrace{p\left(m|x_{1:t}, z_{1:t}, u_{1:t}\right)}_{\text{map}} \tag{2}$$

According to this approach, there is a map distribution associated to each sample of the robot path.

Notice that $u_{1:t}$ can be eliminated from the second term in (2) since the robot path $x_{1:t}$ d-separates the map m and the actions, hence they become conditionally independent (please, refer to [43] for an extensive explanation). Moreover, the term regarding the map in (2) can be further factored due to the conditional independence between the landmarks in the map, given a robot path hypothesis:

$$p\left(m|x_{1:t}, z_{1:t}\right) = \prod_{j=1}^{M} p\left(m_j|x_{1:t}, z_{1:t}\right) \tag{3}$$

The above expressions state that the joint probability density of the robot path and the map, given the set of measurements, can be computed using one estimator for the robot path and M for the landmarks in the map for each of the P particles. In this work, we use a particle filter to estimate $p\left(x_{1:t}|z_{1:t}\right)$, and a Kalman filter (KF) to update the positions of the landmarks at each time step.

Regarding the robot path estimation, it is updated at each time step by appending the latest robot pose, which is computed from:

$$\underbrace{p\left(x_t|z_{1:t}, u_{1:t}\right)}_{\substack{\text{pose estimation}\\ \text{at time } t}} \propto \underbrace{p\left(z_t|x_t\right)}_{\text{observation model}} \cdot \overbrace{\int \underbrace{p\left(x_t|x_{t-1}, u_t\right)}_{\text{transition model}} \underbrace{p\left(x_{t-1}|z_{1:t-1}, u_{1:t-1}\right)}_{\substack{\text{pose estimation}\\ \text{at time } t-1}} dx_{t-1}}^{\text{pose } \textit{prior} \text{ estimation}} \tag{4}$$

where $p(x_t|z_{1:t})$ is approximated by a set of particles, each of them representing a possible robot pose. In short, the Rao-Blackwellized Particle Filter that estimates $p(x_t, m|z_{1:t}, u_{1:t})$ evolves as follows:

(i) The robot path particles are propagated according to the transition model which, for our case, is the mobile robot *motion model*. In this work we estimate the motion between consecutive time steps through a visual odometry algorithm, explained in detail in section 5.

(ii) These particles are subsequently weighted according to the observation model which estimates the likelihood of obtaining the current observation from the pose hypothesis hold by each particle. The observation model, based on 3D landmarks with SIFT descriptors, will be exposed in section 6.

(iii) Next, a resampling stage is performed (if necessary) over the particles. The probability of surviving for each particle is proportional to its importance weight.

(iv) Finally, update the map associated to each particle.

4. Map and Observations

This section presents the process for obtaining 3D visual landmarks from the environment and the associated notation. These landmarks will be the elements of both the observations and the map.

4.1. *Notation and Definitions*

In this work, an observation z_t and the map m are defined as sets of 3D landmarks:

$$\begin{aligned} z_t &= \left\{z_t^i\right\}_{i=\{1,\ldots,N\}} \quad \text{where} \quad z_t^i = \left\langle \mathbf{X}_t^i, \mathbf{F}_t^i \right\rangle \\ m &= \left\{m^j\right\}_{j=\{1,\ldots,M\}} \text{where } m^j = \left\langle \mathbf{X}_m^j, \mathbf{F}_m^j \right\rangle \end{aligned} \tag{5}$$

Each landmark, either of an observation or in the map, comprises a 3D location $\mathbf{X}$, and an associated SIFT descriptor $\mathbf{F}$ [29]. The uncertainty in the 3D positions of the landmarks is modeled by normal distributions with mean μ and a 3×3 covariance matrix Σ:

$$\mathbf{X}_t^i \sim N\left(\mu_t^i, \Sigma_t^i\right) \; \mathbf{X}_m^j \sim N\left(\mu_m^j, \Sigma_m^j\right) \tag{6}$$

The SIFT descriptor $\mathbf{F}$ of each landmark is also assumed to be normally distributed with a diagonal covariance matrix containing a constant value for each dimension, say $\left(\sigma_{\mathrm{S1}}^2, \ldots, \sigma_{\mathrm{S128}}^2\right)$.

$$\mathbf{F}_t^i \sim N\left(\mu_{\mathbf{F}_t^i}, \Sigma_{\mathbf{F}_t^i}\right) \; \mathbf{F}_m^j \sim N\left(\mu_{\mathbf{F}_m^j}, \Sigma_{\mathbf{F}_m^j}\right) \tag{7}$$

Summarizing, we define a generic 3D landmark l^i as a normally distributed random variable with the following statistics:

$$\begin{aligned} l^i &\sim N\left(\left\langle \mu_{\mathbf{X}}^i, \mu_{\mathbf{F}}^i \right\rangle, \left(\begin{array}{c|c} \Sigma_{\mathbf{X}}^i & \mathbf{0} \\ \hline \mathbf{0}^{\mathrm{T}} & \Sigma_{\mathbf{F}}^i \end{array}\right)\right) \\ &= N\left(\left\langle \begin{pmatrix} X^i \\ Y^i \\ Z^i \end{pmatrix}, \begin{pmatrix} d_1^i \\ \vdots \\ d_{128}^i \end{pmatrix} \right\rangle, \left(\begin{array}{ccc|ccc} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} & & & \\ \sigma_{YX} & \sigma_Y^2 & \sigma_{YZ} & & \mathbf{0} & \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_Z^2 & & & \\ \hline & & & \sigma_{\mathrm{S1}}^2 & \cdots & 0 \\ & \mathbf{0}^{\mathrm{T}} & & \vdots & \ddots & \vdots \\ & & & 0 & \cdots & \sigma_{\mathrm{S128}}^2 \end{array}\right)\right) \end{aligned} \tag{8}$$

where $\mathbf{0}$ is a 3×128 null matrix, $\mu_{\mathbf{X}}^i$ stands for the mean of the 3D landmark position, $\mu_{\mathbf{F}}^i$ denotes the mean of the 128D SIFT descriptor, and $\Sigma_{\mathbf{X}}^i$ and $\Sigma_{\mathbf{F}}^i$ are their associated covariance matrices, respectively. The parameters $\left(\sigma_{\mathrm{S1}}^2, \ldots, \sigma_{\mathrm{S128}}^2\right)$ stand for the variance in each of the dimensions of the SIFT descriptor, and model the uncertainty when computing the descriptor of the same feature from different points of view. Their values have been determined empirically from an independent experiment which tracks a set of 576 features in a sequence of 50 images while computing their SIFT descriptors at each time step. In this experiment, a feature is visible and tracked in an average of 26 images. The standard deviation of their descriptors as they vary with time in each of the descriptor dimensions are taken as the values of $(\sigma_{\mathrm{S1}}, \ldots, \sigma_{\mathrm{S128}})$ (see Figure 1).

Once the landmark notation has been introduced, we address the process of obtaining those landmarks from the stereo images.

4.2. *Extraction of Reliable Observation Landmarks*

To obtain a set of 3D landmarks from a pair of stereo images we need to find feature points in both images, to match them, and to estimate their corresponding 3D locations. Next we describe the whole process in more detail.

Several methods have been proposed in the literature for extracting interest points from images, as the well-known detectors of Kitchen & Rosenfeld [24] and Harris [16], based on the first and the second-order derivatives of images, re-

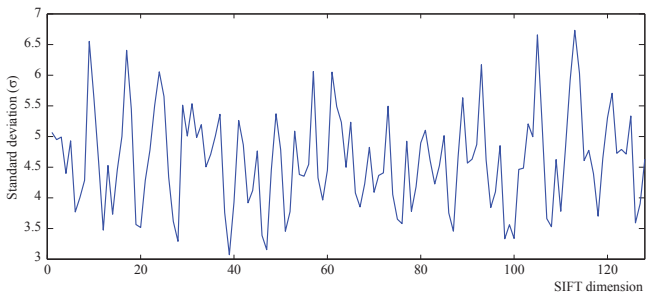


Fig. 1. Estimation of the standard deviation for each dimension of the SIFT descriptor.

spectively. More recently, the SIFT detector proposed by Lowe [29] deals with the detection process by identifying local extrema in a pyramid of Difference of Gaussians (DoG). It also provides the detected features with a descriptor that exhibits invariance to rotation and scale, and partial invariance to lighting changes and affine distortions. The evaluations performed in [2] and [34] reveal the SIFT descriptor as one of the best methods for identifying features. However, the SIFT detector is outperformed by the Harris-based ones when extracting and tracking features through a sequence of images, since its repeatability is not high enough [3]. These papers present the SIFT approach as one of the best descriptors but not as the best detector for visual SLAM. In our work, the detection of interest points in the images is carried out by the method proposed by Shi and Tomasi [47], which is strongly based on the Harris approach. It searches for points in the image with special characteristics that make them easy to be tracked in subsequent images, which is one of the cornerstones of our visual odometry approach stated in section 5. In short, this method computes the eigenvalues of the local autocorrelation matrix around the image points, and compares them with a predefined threshold to detect interest points. Once they are detected, their corresponding 128D SIFT descriptor is also computed to make them sufficiently distinguishable and to improve the robustness of the stereo matching process. Since the Shi and Tomasi detector does not provide any scale information for the detected points, the SIFT descriptor computation is accomplished in one scale only (i.e. the original image) losing, therefore, the scale invariance. However, the resulting SIFT descriptor has been proved to be distinguishable enough for performing stereo matching and for measuring the similarity between the projected landmarks when detected from different points of view in indoor navigation. Although a shorter key vector may be employed for stereo matching with good results, in [29] it is suggested the usage of a 128D to achieve the best matching performance, which is specially interesting to perform the *loop closure* in SLAM, that is, when the robot *realizes* that it has reached an already visited area.

After detecting the set of keypoints in each image, they are matched according to both the similarity of their descriptors and the restriction imposed by the epipolar geometry. In concrete, for each keypoint in the left image, the Euclidean distance between its descriptor and those of the keypoints in the right image is computed. For a pair of conjugate keypoints to be considered a candidate match, both the minimum distance must be below a fixed threshold and the second lowest distance must be sufficiently apart from the minimum (see Figure 2(b)). This method differs from the one proposed by Lowe in [29], which imposes that the ratio of the two minimum distances between descriptors is below a certain threshold. Although this criterion is effective for stereo matching, it may establish correspondences between descriptors that have different levels of distinctiveness, and, therefore, not easily identifiable. In our method, the lower threshold ensures that matched points have very

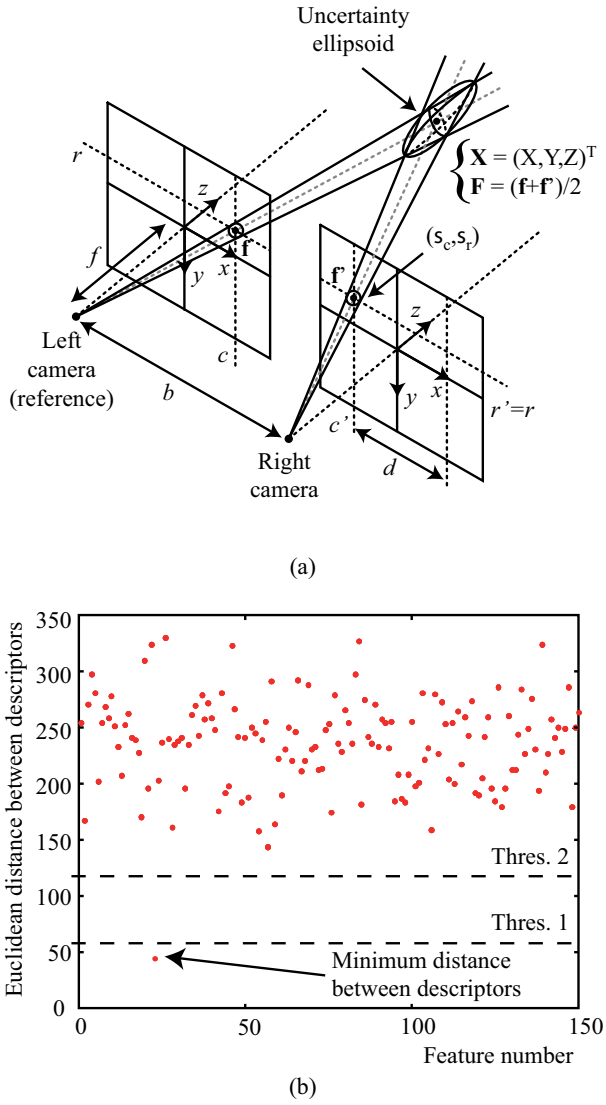


Fig. 2. (a) Configuration of a stereo vision system and schematic representation of the uncertainty in the localization of a 3D landmark. (b) Euclidean distance between the descriptors of a feature in the left image and all the features in the right one.

distinctive descriptors (which leads to very low absolute Euclidean distance) while the second threshold is set to avoid ambiguous correspondences.

In addition, the points must fulfill the epipolar constraint, i.e. they have to lay on their conjugate epipolar lines. In a stereo vision system with parallel optical axis as the one shown in Figure 2(a), epipolar lines are parallel and horizontal, thus, the epipolar constraint reduces to check that both features are in the same row of the image.

Once the matching has been established, the most likely 3D coordinates of the landmark is estimated by projecting them back to space [17,45]. We also consider here the uncertainty in the 3D landmark position due to errors in the image quantization and in the feature detection process. Assuming a stereo system with parallel optical axes and a pinhole camera model (see Figure 2(a)), the 3D coordinates (X, Y, Z) of a landmark can be computed from two

matched points in the left and right images by [45]:

$$X = (c - c_0)\frac{b}{d};\; Y = (r - r_0)\frac{b}{d};\; Z = f\frac{b}{d} \tag{9}$$

where (r, c) are the coordinates of the interest point in the reference image (say, the left one), (r_0, c_0) represents the coordinates of the principal point in the reference image, and b, d and f stands for the baseline, the disparity, and the focal length of the stereo rig, respectively (please, refer to Figure 2(a)).

Errors in the variables r, c, and d, are usually modelled as uncorrelated zero-mean Gaussian random variables [31]. Using a first-order error propagation to approximate the distribution of the variables in (9) as multivariate Gaussians, we obtain the following covariance matrix for the X, Y and Z coordinates:

$$\Sigma_{\mathbf{X}} \approx \mathbf{J}\, diag\left(\sigma_c^2, \sigma_r^2, \sigma_d^2\right) \mathbf{J}^{\mathrm{T}} \tag{10}$$

where $\mathbf{J}$ stands for the Jacobian matrix of the functions in (9), and σ_c^2, σ_r^2 and σ_d^2 are the variances of the corresponding variables. Expanding (10) we come to the following expression for $\Sigma_{\mathbf{X}}$:

$$\Sigma_{\mathbf{X}} = \left(\frac{b}{d}\right)^2 \tag{11}$$
$$\begin{pmatrix} \sigma_c^2 + \frac{\sigma_d^2 (c-c_0)^2}{d^2} & \frac{\sigma_d^2 (c-c_0)(r-r_0)}{d^2} & \frac{\sigma_d^2 (c-c_0) f}{d^2} \\ \frac{\sigma_d^2 (c-c_0)(r-r_0)}{d^2} & \sigma_r^2 + \frac{\sigma_d^2 (r-r_0)^2}{d^2} & \frac{\sigma_d^2 (r-r_0) f}{d^2} \\ \frac{\sigma_d^2 (c-c_0) f}{d^2} & \frac{\sigma_d^2 (r-r_0) f}{d^2} & \frac{\sigma_d^2 f^2}{d^2} \end{pmatrix}$$

which models the uncertainty in the 3D coordinates of landmarks computed from the noisy measurements of the stereo system.

Please, note that the uncertainties of the camera intrinsic parameters, i.e. the baseline, the focal length and the principal point coordinates, are not taken into account, since the camera employed in our experiments is supposed to be accurately calibrated by the manufacturer and, therefore, the errors in these parameters may be considered negligible, as is common in the literature. However, they could be easily introduced by linear uncertainty propagation using the equations in (9).

In order to validate this error model, we have performed an experiment where the real density of the landmark location (derived from a Monte-Carlo simulation) has been compared with the approximated density from the linearized model. For that purpose, we have chosen by hand a set of matches in a pair of stereo images and computed their disparity, having in this way the values of r, c and d in equations (9) and (11). For each stereo match, the Monte-Carlo simulation has been performed by drawing a set of 10.000 samples from the Gaussians distributions of r, c and d (assuming a variance of $\sigma_r^2 = \sigma_c^2 = 1$ and $\sigma_d^2 = 2$ pixels, respectively), and by projecting them through equations (9), to yield a set of 10.000 samples of the landmark 3D position. These samples accurately model the real density and allow the estimation of the real means and covariances. Figure 3 shows the planar projections of the 3D uncertainty ellipses associated to both estimations for one of the landmarks, where the black-thick one corresponds to the Monte-Carlo method and the blue-dotted one corresponds to the linearized model.

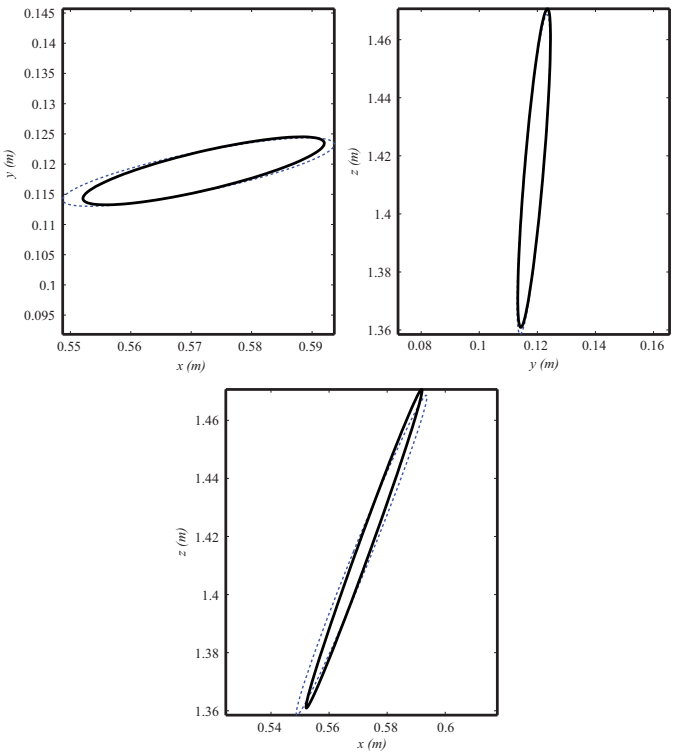


Fig. 3. Planar projections of the uncertainty ellipses associated to both Monte-Carlo simulation (black-thick line) and linearized error model estimations (blue-dotted line).

To measure the similarity between these two density distributions, we have employed the Kullback-Leibler divergence (D_{KL}) [26], widely employed in statistics for that purpose. We must remark that, since the involved distributions are Gaussians, D_{KL} can be computed through a closed-form solution [18]; this measure yields an average value of 0.36, which is similar to that obtained from two normalized one-dimensional Gaussians with an offset of about 0.8σ in their means.

We have checked other ways of obtaining the covariance matrix $\Sigma_{\mathbf{X}}$ like the Unscented Transformation (UT) [53] and the scaled UT [21]. We have contrasted them to the real density and, since the comparison of the results achieved by the three aforementioned methods (linear, UT and scaled UT) shows similar performance, we have opted for the linear approach (summarized by equations (9) and (11)) because of its efficiency.

Finally, each 3D landmark is assigned a 128D SIFT descriptor which is simply computed as the average value of

the descriptors (f and f') from each image: $\mathbf{F} = (f + f')/2$.

The so-computed 3D landmarks are the elements of both the observations and the map, and constitute the basis of the probabilistic SLAM method proposed in this work.

4.3. *Map Initialization and Update*

This section addresses the management of the map built during the SLAM process, which entails the insertion, update and deletion of landmarks.

In short, the map management can be summarized by this sequence of steps, which will be further explained next:

(i) In each iteration, it is performed a data association procedure to obtain a set of correspondences between the observed landmarks and those in the map.
   (a) The positions of the landmarks in the map with a correspondence are updated.
   (b) The landmarks in the map without a correspondence are not modified.
   (c) The observed landmarks with no correspondences are introduced into the map.

(ii) Landmarks which have not been observed a significant number of times are deleted from the map, since they are considered as *non-stable*.

Every landmark in the map has two associated attributes indicating the number of times and the last time step it has been observed, respectively.

At the beginning of the SLAM process, all the landmarks which are detected in the first observation are introduced into an initially empty map, and their associated counters are initialized accordingly. Then, as new observations are gathered, the detected landmarks at each time step are compared with those in the map in order to obtain a set of matches. To that purpose, their probability of being in correspondence is evaluated from the distance between both their 3D positions and their SIFT descriptors, whereas the matching decision is taken according to a certain threshold. This correspondence measure is computed from the same expression which is proposed for our observation model (to be described in section 6).

Once the correspondences have been established, the 3D positions of the landmarks with positive matches are updated through the Kalman Filter equations [23], which, according to our observation model, come to:

$$\begin{aligned} \mu_m &= \Sigma_m \left(\Sigma_{\tilde{m}}^{-1} \mu_{\tilde{m}} + \Sigma_t^{-1} \mu_t \right) \\ \Sigma_m &= \left(\Sigma_{\tilde{m}}^{-1} + \Sigma_t^{-1} \right)^{-1} \end{aligned} \tag{12}$$

where $(\mu_{\tilde{m}}, \Sigma_{\tilde{m}})$ and (μ_m, Σ_m) represent the distributions of the landmark position before and after the update process, respectively, while (μ_t, Σ_t) stands for the observed position of the landmark at time step t. Please, refer to appendix B for a complete derivation of these expressions from the Kalman Filter equations.

Finally, we update the counters of all the landmarks in the map, and delete those which are considered to be non-stable, i.e. those landmarks which have not been detected neither in recent iterations nor a suficient number of times.

5. Motion Model: Visual Odometry

Typically, in robot localization and SLAM, the motion model is given by a probabilistic characterization of the robot displacement obtained from encoder-based odometry. In this section we describe a motion model which does not rely on the robot odometry but is based on matching 3D landmarks between two consecutive robot poses. This motion model is not restricted to planar robot motion since it estimates the incremental change in 6D: x, y, z, *yaw*, *pitch*, and *roll*. The algorithm takes as inputs two sets of 3D points computed at different time steps, with known correspondences and coordinates relative to each robot pose, and estimates their relative change through a closed-form solution derived in [19]. Figure 4 shows a schematic representation of the visual odometry approach whose main stages are briefly depicted next.

In the first pair of stereo images, two sets of features are extracted and matched according to both their SIFT descriptors and the epipolar constraint (as explained in the previous section). The matched pairs are subsequently projected into space and their 3D spatial uncertainty is also computed.

Then, the features are tracked in the next pair of stereo images (using the KLT method), which produces to a new set of matched image features at this time step. Notice that, since the correspondences between the tracked features in the left and right images are already known, it is not necessary to match each other again. This speeds up the process significantly and reduces the computational burden of the whole visual odometry procedure. Finally, the new set of matched points is also projected into 3D space.

We must remark that, due to the tracking process, the associations between the two sets of 3D landmarks are also known at each time step. This set of 3D landmark pairs is taken as input for the closed-form solution which computes the 6D robot pose increment and its associated uncertainty, as explained next.

5.1. *Statement of the Visual Odometry Problem*

Let $q_{t,t+1}$ be a random variable which models the pose increment between time steps t and $t+1$ as a function of the sets of 3D landmarks X_t and X_{t+1} (as they were defined in section 4.1):

$$q_{t,t+1} = f\left(\mathbf{X}_t, \mathbf{X}_{t+1}\right) \qquad q_{t,t+1} \sim N\left(\mu_q, \Sigma_q\right) \tag{13}$$

Assuming a linear propagation of errors, $q_{t,t+1}$ follows a Gaussian distribution with covariance matrix Σ_q and mean:

$$\mu_q = \begin{pmatrix} \Delta x & \Delta y & \Delta z & \Delta\alpha & \Delta\beta & \Delta\gamma \end{pmatrix}^{\mathrm{T}} \tag{14}$$

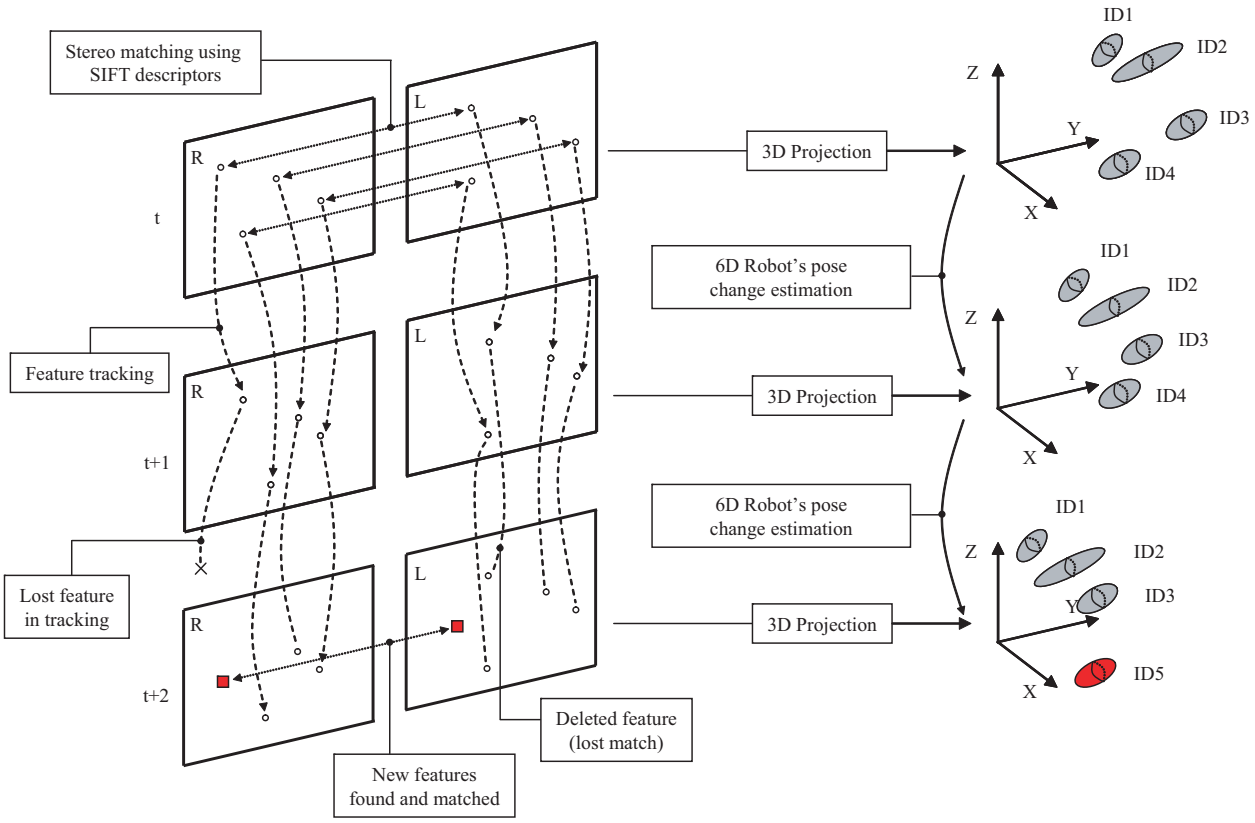


Fig. 4. Scheme of the visual odometry algorithm.

where $\Delta x, \Delta y$, and Δz, are the increments in the X, Y, and Z coordinates respectively, and $\Delta\alpha, \Delta\beta$, and $\Delta\gamma$ stand for the increments in the *yaw*, *pitch*, and *roll* angles, respectively.

The visual odometry problem consists of computing μ_q and Σ_q from equation (13), as described next.

5.2. *Computation of the Mean*

The mean of the pose increment (μ_q) is computed through the method reported by Horn in [19], where it is derived a closed-form solution to the least-squares problem of finding the relationship between two coordinate systems given a number of points in both systems. In our work these points will be assumed to be the mean values μ_t^i of the 3D landmarks (please, refer to equation (6)).

This closed-form solution is in contrast to other proposals for visual odometry based on iterative methods [40,48] which require an initial estimation, and may have convergence problems. Horn's algorithm, as applied to this problem, is depicted in appendix A.

5.3. *Computation of the Covariance*

Covariance matrices are usually obtained through a linear approximation of the functions involved in a given transformation between variables (see, for example, section 4.2). However, the closed-form solution described in appendix A cannot be linearized since it involves the computation of eigenvectors. Instead, we propose the following solution for estimating the covariance matrix of the pose increment Σ_q.

The rigid transformation that relates two sets of N corresponding points in two instants of time t and $t+1$ can be expressed as:

$$\mathbf{X}_{t+1}^i = f\left(\mu_q, \mathbf{X}_t^i\right) \; i = 1, \ldots, N \tag{15}$$

which becomes linear when using homogeneous coordinates:

$$\begin{pmatrix} X_{t+1}^i \\ Y_{t+1}^i \\ Z_{t+1}^i \\ 1 \end{pmatrix} = \left(\begin{array}{c|c} \mathbf{R} & \mathbf{t} \\ \hline 0\;\;0\;\;0 & 1 \end{array}\right) \begin{pmatrix} X_t^i \\ Y_t^i \\ Z_t^i \\ 1 \end{pmatrix} \tag{16}$$

Our interest is to compute the 6×6 covariance matrix Σ_q, given μ_q and the 3×3 covariances of all the 3D landmarks in correspondence between t and $t+1$ (Σ_t^i and Σ_{t+1}^i, respectively). This can be accomplished as follows:

First, let Σ be a $3N \times 3N$ block diagonal matrix comprising the information about Σ_t^i and Σ_{t+1}^i:

$$\Sigma = \begin{pmatrix} \Sigma^1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Sigma^2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \Sigma^N \end{pmatrix} \tag{17}$$

where N stands for the number of landmarks, $\mathbf{0}$ is a 3×3 null matrix, and:

$$\Sigma^i = \Sigma^i_t + \Sigma^i_{t+1} \tag{18}$$

On the other hand, let $\mathbf{H}$ be a $3 \times 6N$ matrix composed of a set of 3×6 submatrixes H^i:

$$\mathbf{H} = \begin{pmatrix} H^1 & \dots & H^i & \dots & H^N \end{pmatrix} \tag{19}$$

where each H^i is the Jacobian of f in (15), with respect to the components of μ_q, for each pair i of corresponding 3D points:

$$H^i = \begin{pmatrix} \frac{\partial f_X}{\partial \Delta x} & \frac{\partial f_X}{\partial \Delta y} & \frac{\partial f_X}{\partial \Delta z} & \frac{\partial f_X}{\partial \Delta \alpha} & \frac{\partial f_X}{\partial \Delta \beta} & \frac{\partial f_X}{\partial \Delta \gamma} \\ \frac{\partial f_Y}{\partial \Delta x} & \frac{\partial f_Y}{\partial \Delta y} & \frac{\partial f_Y}{\partial \Delta z} & \frac{\partial f_Y}{\partial \Delta \alpha} & \frac{\partial f_Y}{\partial \Delta \beta} & \frac{\partial f_Y}{\partial \Delta \gamma} \\ \frac{\partial f_Z}{\partial \Delta x} & \frac{\partial f_Z}{\partial \Delta y} & \frac{\partial f_Z}{\partial \Delta z} & \frac{\partial f_Z}{\partial \Delta \alpha} & \frac{\partial f_Z}{\partial \Delta \beta} & \frac{\partial f_Z}{\partial \Delta \gamma} \end{pmatrix} \tag{20}$$

In this expression, f_X, f_Y, and f_Z are derived from (15) and denote the functions which determine the X, Y and Z coordinates, respectively, of the 3D point at time step $t+1$.

The resulting covariance matrix Σ_q is computed using the above-mentioned matrices as:

$$\Sigma_q = \left(\mathbf{H}^{\mathrm{T}} \Sigma^{-1} \mathbf{H}\right)^{-1} \tag{21}$$

Notice that, due to the block diagonal structure of Σ, its inverse matrix is also block diagonal and contains the inverse of the submatrixes Σ^i:

$$\Sigma^{-1} = \begin{pmatrix} (\Sigma^1)^{-1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & (\Sigma^2)^{-1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & (\Sigma^N)^{-1} \end{pmatrix} \tag{22}$$

Thus, the expression (21) can be split into the sum of its block diagonal elements, yielding the final expression of the covariance matrix Σ_q:

$$\Sigma_q = \left(\sum_{i=1}^{N} H^{iT} \left(\Sigma^i\right)^{-1} H^i \right)^{-1} \tag{23}$$

It must be remarked that the sum of the two covariance matrices in (18) has no meaning but its inverse $\left(\Sigma^i\right)^{-1}$ in (22) may be interpreted as a logical way of weighting the contribution of each pair of points to the pose increment uncertainty. Thus, the covariance of the pose increment, shown in (23), can be interpreted as the weighted sum of the uncertainty in q given by the different pairs of landmarks.

The computed mean and covariance constitute an estimation of the robot displacement between two time steps, which will be used as the motion model for the Bayesian filter discussed in the present work.

6. An Observation Model for Stereo Vision

In the following we introduce our proposal for a probabilistic observation model $p\left(z_t \middle| x_t\right)$, which stands for the likelihood of an observation at time t, given the robot pose x_t. Notice that, as each particle $x_t^{[k]}$ represents a hypothesis of the robot pose (k represents the index of the particle), this likelihood will be evaluated at each particle in the filter. In the following formulation, however, for clarity, we will omit the particle indexes.

Firstly, assuming conditional independency between the errors in the detection of the individual landmarks z_t^i, the likelihood function can be factorized as follows:

$$p\left(z_t|x_t\right) \overset{\text{cond.ind}}{=} \prod_i p\left(z_t^i|x_t\right) \tag{24}$$

To avoid explicit data association between landmarks in the observation and in the map, we apply next the law of total probability to marginalize out the observation likelihood of individual landmarks by considering all the possible associations:

$$p\left(z_t^i|x_t\right) = \sum_{\psi=\{1,\dots,M,\phi\}} p\left(z_t^i|x_t, c_i = \psi\right) \underbrace{P\left(c_i = \psi|x_t\right)}_{\eta} \tag{25}$$

where c_i is an unknown discrete variable that represents the correspondence of the ith observed landmark. Its possible values are $1, \dots, M$ for map landmarks, or ϕ for no correspondence with the map. Notice that the *a priori* probability of any given correspondence $P\left(c_i = \psi|x_t\right)$ is a constant since it does not depend on the actual observation z_t^i. If we do not have any other information, we can assume the same probability for all the possible correspondences, including the null one:

$$p\left(z_t^i|x_t\right) = \eta \sum_{\psi=\{1,\dots,M,\phi\}} p\left(z_t^i|x_t, c_i = \psi\right) \tag{26}$$

The term $p\left(z_t^i|x_t, c_i = \psi\right)$ can be seen as the probability of the observed landmark z_t^i and its corresponding landmark m_ψ to coincide in both the 3D space of the position and the 128-dimensional space of the SIFT descriptors. This can be computed by simply evaluating at the origin a Gaussian distribution whose mean μ is the difference between the means of z_t^i and m_ψ and the covariance Σ is the sum of their covariance matrices:

$$p\left(z_t^i|x_t, c_i=\psi\right) = N\left(0; \underbrace{\bar{z}_t^i - \bar{m}_\psi}_{\mu}, \underbrace{\Sigma_{z_t^i} + \Sigma_{m_\psi}}_{\Sigma}\right) \tag{27}$$

$$= \eta' \exp\left\{-\frac{1}{2}\left(\bar{z}_t^i - \bar{m}_\psi\right)^{\mathrm{T}} \left(\Sigma_{z_t^i} + \Sigma_{m_\psi}\right)^{-1} \left(\bar{z}_t^i - \bar{m}_\psi\right)\right\}$$

where

$$\eta' = \left(2\pi \left|\Sigma_{z_t^i} + \Sigma_{m_\psi}\right|\right)^{-\frac{1}{2}} \tag{28}$$

Due to the particular structure of the mean μ and the covariance matrix Σ (similar to those shown in expression (8)), the exponential term in (27) can be split in two factors related to the position and descriptor dimensions of the random variable, respectively:

$$\begin{aligned} & N\left(0; \mu, \Sigma\right) \\ & = \eta' \exp\left\{-\frac{1}{2}\left(\mu_{\mathbf{X}}^{\mathrm{T}} \middle| \mu_{\mathbf{F}}^{\mathrm{T}}\right) \left(\begin{array}{c|c} \Sigma_{\mathbf{X}} & \mathbf{0} \\ \hline \mathbf{0}^{\mathrm{T}} & \Sigma_{\mathbf{F}} \end{array}\right)^{-1} \left(\begin{array}{c} \mu_{\mathbf{X}} \\ \hline \mu_{\mathbf{F}} \end{array}\right)\right\} \\ & = \eta' \exp\left\{-\frac{1}{2}\left(\mu_{\mathbf{X}}^{\mathrm{T}} \Sigma_{\mathbf{X}}^{-1} \mu_{\mathbf{X}} + \mu_{\mathbf{F}}^{\mathrm{T}} \Sigma_{\mathbf{F}}^{-1} \mu_{\mathbf{F}}\right)\right\} \\ & = \eta' \underbrace{\exp\left\{-\frac{1}{2}\mu_{\mathbf{X}}^{\mathrm{T}} \Sigma_{\mathbf{X}}^{-1} \mu_{\mathbf{X}}\right\}}_{\text{position}} \underbrace{\exp\left\{-\frac{1}{2}\mu_{\mathbf{F}}^{\mathrm{T}} \Sigma_{\mathbf{F}}^{-1} \mu_{\mathbf{F}}\right\}}_{\text{descriptor}} \end{aligned} \tag{29}$$

Furthermore, since $\Sigma_{\mathbf{F}}$ is a diagonal matrix containing constant values $\left(\sigma_{S1}^2, \ldots, \sigma_{S128}^2\right)$, the exponent of the descriptor term in (29) is proportional to the squared Mahalanobis distance between the descriptors and its computation is greatly simplified.

7. Experimental Results

The proposed method for performing visual SLAM within a particle filter framework has been tested in two experiments involving different kinds of camera movements:

(i) In the first experiment, Sancho, one of our mobile robots equipped with a BumbleBee [1] stereo vision system, was manually driven following an almost circular trajectory in an office environment while gathering stereo images.

(ii) The second experiment shows the performance of our approach when coping with data from a hand-held camera describing a totally unconstrained trajectory.

In both experiments, the camera ego-motion was computed using the visual odometry process described in section 5 and the estimations were stored in a log file in order to be subsequently used as the motion model of the SLAM process, which does not operate in real time.

[1] http://www.ptgrey.com

7.1. *Office-like Environment Experiment*

Figures 5(a-b) show a plan and a snapshot of the environment where this first experiment has been carried out. This environment offers some interesting features for testing the robustness of the proposed visual SLAM approach. Thus, for example, the white board shown at the top of the image 5(b) means a place where the detection of visual landmarks is very unlikely. On the other hand, there are some bookshelves at the left side of the room which produces many reliable landmarks. Finally, the bottom-right zone of the room is not well illuminated, which hampers the extraction of landmarks. The results achieved by our SLAM method in all these different situations are shown in the next section, while a video showing the complete evolution of this experiment can be watched online [2].

In the feature extraction process, it has been assumed that errors in the variables r and c (i.e. the row and the column in the image, respectively, of the detected interest points) have a variance of 1 pixel, while the errors in d (i.e. the disparity) is considered to be 2 pixels. This value arises from the assumption of independence between the errors in the estimation of the column variables for both images; thereby the variance of the disparity $(d = c_{left} - c_{right})$ becomes the sum of the variances of both c variables.

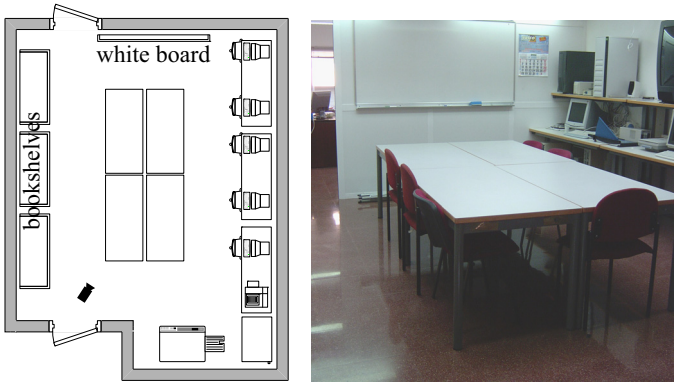
Others parameters of the camera configuration and for the experiment setup are summarized in Table 1.

7.1.1. *Map Building*

In this experiment, the recorded sequence of stereo images is used to build a map of the environment through the Rao-Blackwellized Particle Filter (RBPF) described in section 3.

The evolution of the constructed map as the robot moves is shown in Figure 6, where a top view of the 3D landmarks of the map being built (shown as ellipses representing 95% confidence intervals) is displayed at six different time steps.

[2] http://www.youtube.com/watch?v=m3L8OfbTXH0



(a) (b)

Fig. 5. Experiment scenario. (a) Plan and (b) a picture of the environment.

Table 1
Camera Configuration and Experimental Setup

Camera Configuration *PointGrey BumbleBee*	
Baseline (b)	11.9 cm
Focal length (f)	507.808 px
Principal Point coordinates (c_0, r_0)	(252.922, 356.237) px
Image size	640×480 px
Experimental Setup	
Path length (approx.)	40 m
Total number of stereo images	1000
Image capture rate	3 Hz
$\left(\sigma_c^2, \sigma_r^2, \sigma_d^2\right)$	(1,1,2) px
Number of particles	80
Number of landmarks in the final map	927

In addition, the hypothesis of the robot path estimated by each particle is also shown in the figure.

The uncertainty in both the robot pose and the map depends on the number of observed landmarks at each time step and the number of times they are detected. Thereby, the uncertainty in the landmark positions decreases as they are detected in successive observations.

The zones A, B and C indicated in Figure 6(f) correspond to the interest zones which where mentioned in the previous section: the white board, the bookshelves and the poorly illuminated zone, respectively. As can be seen, the group of observed landmarks in each situation are sensibly different. Thus, in zone A, the number of landmarks is quite low, as it was expected from the visual characteristics of the environment. In zone B there is a high number of observed landmarks with low uncertainty, which manifests the adequate visual characteristics of that zone. Finally, in zone C, because of the poor illumination, the landmarks are observed only a few number of times and, therefore, the initial uncertainty of their position do not reduce much. Regarding the uncertainty of the robot pose estimation, directly related to the size of the red covariance ellipse in Figure 6, it grows during the first part of the experiment since the vast majority of the observed landmarks are new. This situation can be appreciated in the first part of the plot in Figure 7(a), which shows the evolution of the determinant of the particles' covariance matrix. The substantial increase around iteration 225 is due to the combination of two adverse factors: i) the robot is performing a turn (which entails significant errors in the visual odometry estimations) and ii) the scarce visual information available in that zone since the camera sees an almost totally texture-less surface (the above-mentioned white board).

However, notice that this increasing trend changes drastically around iteration 240 (at some point between Figures 6(c) and (d)) since the robot *closes the loop*, reaching an already navigated position. Now, most of the observed landmarks correspond to those previously stored in the map,

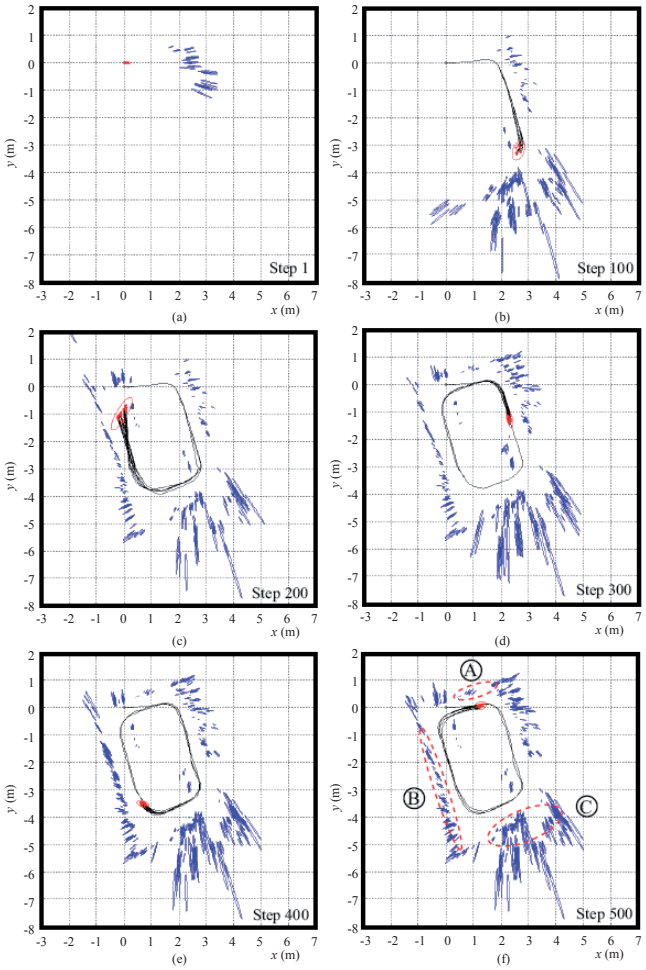


Fig. 6. Map building with a Rao-Blackwellized Particle Filter. Map representation at different time steps where a 95% Gaussian confidence interval of the landmark positions is represented by blue ellipses. The red ellipse surrounding the particle set represents the uncertainty in the robot pose corresponding with a 95% Gaussian confidence interval

hence the estimation of the robot position improves (the particles converge towards the real robot location) and, therefore, the determinant of the covariance matrix reduces significantly. This *loop closure* can be also appreciated in Figure 7(b) where it is represented the IDs of the observed landmarks with correspondences in the map for each time step. In this work, the management of the landmark ID is accomplished as follows:

– As the robot explores the environment, a unique ID is assigned (in an increasing order) to each new observed landmark.
– A landmark which was observed previously keeps its original ID while its position and covariance are updated with the new observation.

Thus, the increasing zone at the beginning of the Figure 7(b) (up to time step 225 approx.) illustrates the initial situation where the robot navigates an area for the first time while observing both new landmarks (the majority) and landmarks which were recently observed and stored in

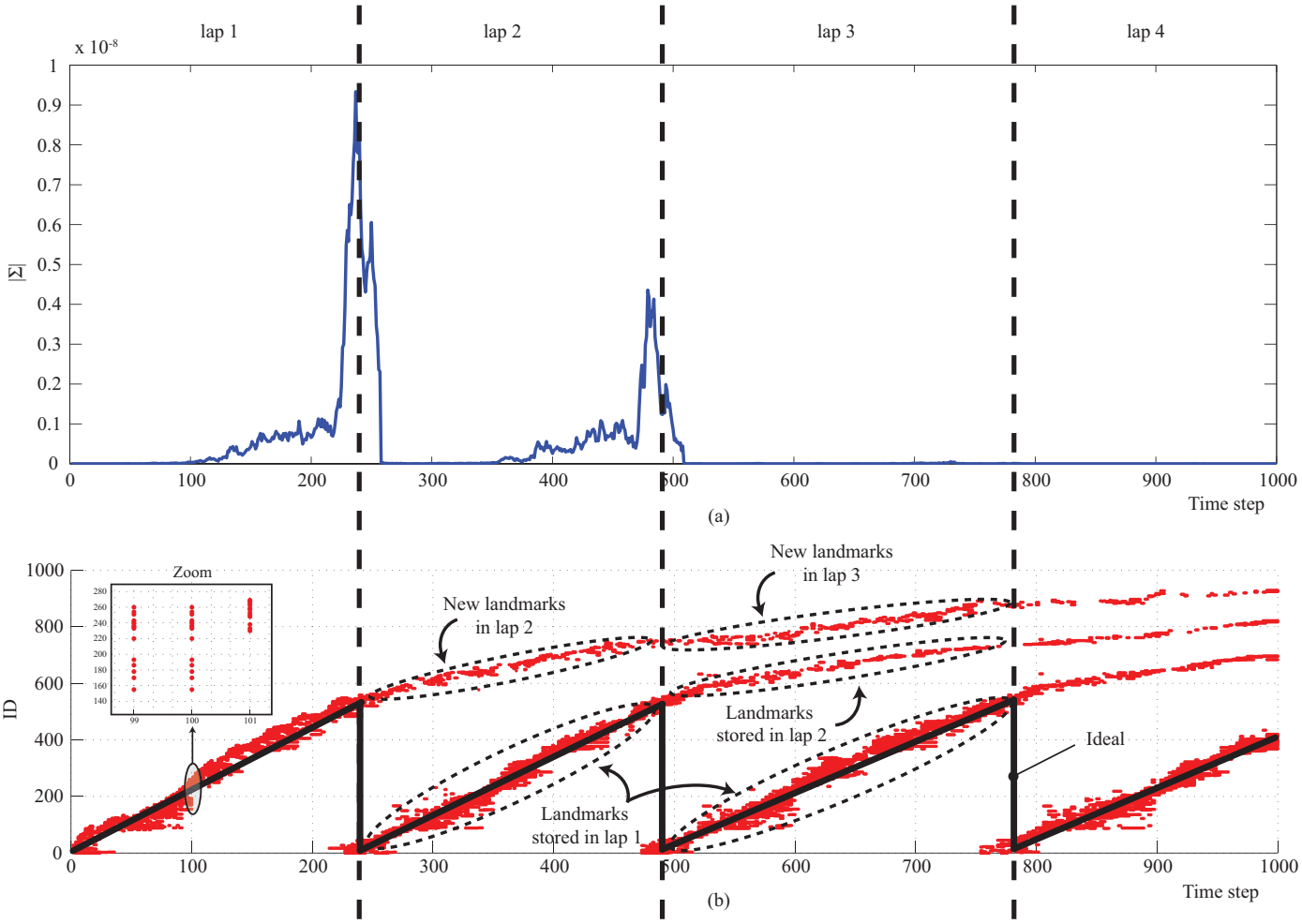


Fig. 7. (a) Determinant of the particles' covariance matrix through time and (b) the landmarks IDs at each time step

the map.

At loop closure (approx. at time step 240), the robot observes a large number of already stored landmarks while the number of new landmarks reduces significantly. The already observed landmarks correspond to those observed at the beginning of the experiment (with the lowest IDs). This manifests in Figure 7(b) as two different groups of IDs: one for the new landmarks and other for those already stored in the map. This situation repeats every time that the robot reaches the initial position of the path.

Please note that, in an ideal situation, with the robot observing all the possible landmarks in the environment and establishing correct correspondences with the map at each time step, the representation of the IDs in this experiment should be sawtooth-shaped (the solid black line in the figure). In real situations, although the robot moves through an already explored zone, it observes new landmarks which are assigned new IDs, giving the particular shape of Figure 7(b).

Finally, it must be remarked that there exists another interval of iterations (between time steps 375 and 500) where the determinant of the particles' covariance matrix grows (see Figure 7(a)). The initial increasing zone is caused by the high uncertainty of the landmarks in that zone of the map (zone B in Figure 6(f)) while the peak at time step 475 comes again from the combination of a turn and the presence of the texture-less white board (zone A in Figure 6(f)). The magnitude of this increase is sensibly lower than that in the first lap of the experiment since some correspondences between the observed landmarks and those of the map have been established, hence reducing the landmark uncertainty and, therefore, improving the robot pose estimation. Following this trend, notice that the influence of this situation is practically negligible in the third and the fourth laps.

7.1.2. *Visual SLAM Performance*

In order to evaluate the performance of our visual SLAM approach, we have compared the robot path estimates from both a RBPF algorithm based on laser data, gathered with a SICK LMS-200 laser scanner, and our visual SLAM approach (see Figure 8(a)). In this work, we have considered the former as the *ground truth* of the robot path since laser data is highly accurate. We must remark that both laser data and stereo images have been gathered simultaneously during the navigation in order to ensure that the path fol-

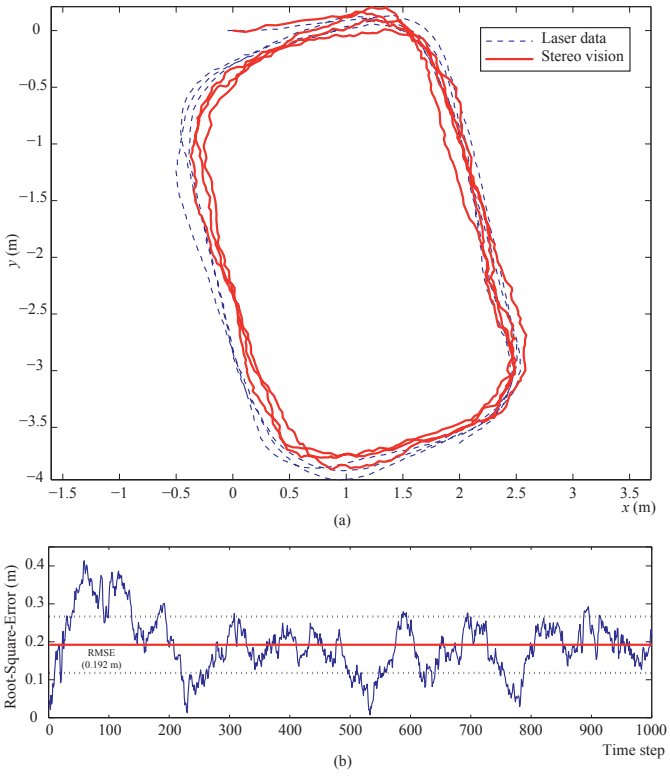


Fig. 8. (a) Comparison between the robot path estimated through a RBPF based on laser data (blue, dashed line) and our visual SLAM approach (red, thick line). (b) Root-Square-Error committed by the visual SLAM method (the horizontal thick line indicates the RMSE $\approx$ 19.2 cm.)

lowed by the robot is exactly the same for both types of data.

Figure 8(b) shows the root-square-error committed by the proposed method when estimating the robot path, yielding a Root-Mean-Square-Error (RMSE) of 19.2 cm with a standard deviation of 7.4 cm.

Regarding the computational time, this experiment has been carried out on a Desktop PC Intel Pentium 4 at 2.60 GHz running under Windows XP SP2 with 2 GB of RAM memory. The processing time of each iteration of the particle filter is represented in Figure 9, where it can be seen the increasing tendency due to the growing amount of stored landmarks in the map, which involves an increment in the processing time of both computing the likelihood and inserting the observations into the map.

Our implementation of the particle filter does not accomplish the insertion stage at every iteration, but only when the robot has moved a certain distance. This leads to the noisy appearance of the Figure 9, since iterations which perform and do not perform the insertion stage alternate through time spending a mean time of 22.4 and 11.65 seconds, respectively (shown in the figure as dotted red lines), while the overall mean time for each iteration is 15.3 seconds (solid red line in the figure).

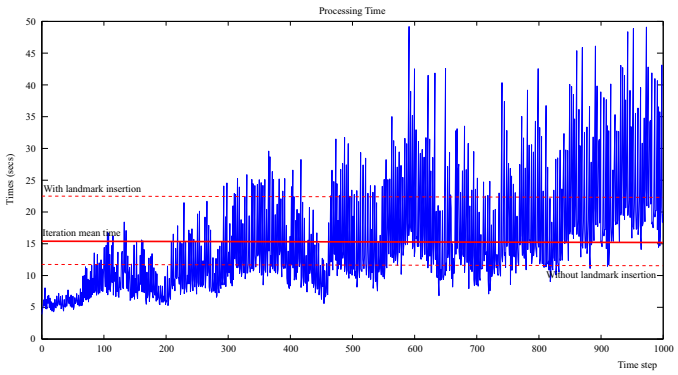


Fig. 9. Processing time evolution for each iteration of the PF. The dotted red lines represent the mean time of an iteration in the cases of inserting and not inserting the observed landmarks into the map while the solid red line indicates the overall mean time.

7.2. *6 DoF Camera Experiment*

This second experiment has been carried in order to test the suitability of our SLAM approach when handling totally unconstrained movements. Thus, a stereo camera has been moved by hand following an arbitrary 6 DoF trajectory in one of our labs, while gathering images and computing visual odometry from them. The estimated ego-motion information is subsequently employed as the motion model of our SLAM proposal.

Figures 10 (a)-(b) show a representative image employed in this experiment and a 3D representation of the estimated path, respectively. Please note that it is difficult to obtain any kind of *ground truth* in this type of experiments, since the camera real trajectory cannot be estimated from typical exteroceptive sensors such as laser scans or sonars. Therefore, the performance of our method may only be estimated by visual inspection from a video [3] showing the sequence of the stereo images captured by the camera, as well as the evolution of both the constructed map and the camera estimated path.

As can be seen in the video, the camera is initially located on a table, and later on it is lifted up while describing a pair of turns in the air. The estimated 6 DoF path of the

[3] http://www.youtube.com/watch?v=b73W53Kwgjw

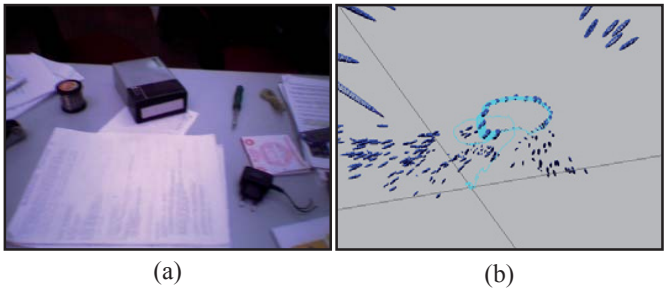


Fig. 10. Our 6 DoF camera experiment. (a) A representative image employed in this experiment, and (b) the 3D estimated path of the camera.

camera closely resembles this trajectory, thus validating our method for arbitrary 6D movements.

8. Conclusions and Future Work

In this paper we have addressed the SLAM problem for stereo vision systems within the probabilistic framework of PF methods.

The robot ego-motion is estimated through a closed-form formulation to perform 6D visual odometry which models the uncertainty of the pose increment estimation as a Gaussian distribution. This ego-motion estimation relies only on the visual information provided by the stereo camera and does not assume any restriction in the robot movement (e.g. smooth paths, navigation across planar surfaces, etc.). Moreover, it avoids the divergence and local-minima problems of the iterative approaches to visual odometry.

On the other hand, our observation model for the RBPF algorithm considers observations as sets of landmarks determined by their 3D positions and their SIFT descriptors, as well as their associated uncertainty. As an important contribution, we avoid explicit data association by marginalizing out the observation likelihood over all the possible associations, thus overcoming the problems derived from establishing incorrect correspondences between the observed landmarks and those in the map.

An experiment with a real robot has been performed in order to validate our proposal in the context of map building. The experimental results illustrate its adequate performance when coping with the SLAM problem and reveal the proposed models as promising approaches for stereo vision in robotics. The MSE committed by our method in comparison to a RBPF approach employing laser data is approximately 19.2 cm.

We are currently working on reducing the computational time of the algorithm with the aim of applying our SLAM approach in real time. Moreover, we are also studying the suitability of other different approaches for extracting image features in order to achieve robust, efficient and more accurate implementations.

Appendix A. Computation of the Pose Increment through a Closed-form Solution

The closed-form solution that computes the pose increment:

$$\mu_q = \langle \Delta x, \Delta y, \Delta z, \Delta\alpha, \Delta\beta, \Delta\gamma \rangle \tag{A.1}$$

between two reference systems given the coordinates of a set of N 3D points in both systems, $\{\mathbf{X}_t^i\}$ and $\{\mathbf{X}_{t+1}^i\}$, can be summarized as follows:

1. Compute the centroids ($\mathbf{c}_t$ and $\mathbf{c}_{t+1}$) of the two sets of points and subtract them from their coordinates in order to deal only with coordinates relative to their centroids:

$$\begin{aligned}
\mathbf{c}_t &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_t^i \\
\mathbf{c}_{t+1} &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_{t+1}^i \\
\overline{\mathbf{X}}_t^i &= \left(\overline{X}_t^i, \overline{Y}_t^i, \overline{Z}_t^i\right)^{\mathrm{T}} = \mathbf{X}_t^i - \mathbf{c}_t \\
\overline{\mathbf{X}}_{t+1}^i &= \left(\overline{X}_{t+1}^i, \overline{Y}_{t+1}^i, \overline{Z}_{t+1}^i\right)^{\mathrm{T}} = \mathbf{X}_{t+1}^i - \mathbf{c}_{t+1}
\end{aligned} \tag{A.2}$$

2. For the ith 3D point, compute the following nine products of its coordinates at time t and $t+1$:

$$\begin{aligned}
P_{XX}^i &= \overline{X}_t^i\overline{X}_{t+1}^i \quad P_{YX}^i = \overline{Y}_t^i\overline{X}_{t+1}^i \\
P_{XY}^i &= \overline{X}_t^i\overline{Y}_{t+1}^i \quad P_{YY}^i = \overline{Y}_t^i\overline{Y}_{t+1}^i \\
P_{XZ}^i &= \overline{X}_t^i\overline{Z}_{t+1}^i \quad P_{YZ}^i = \overline{Y}_t^i\overline{Z}_{t+1}^i \\
P_{ZX}^i &= \overline{Z}_t^i\overline{X}_{t+1}^i \quad P_{ZY}^i = \overline{Z}_t^i\overline{Y}_{t+1}^i \\
P_{ZZ}^i &= \overline{Z}_t^i\overline{Z}_{t+1}^i
\end{aligned} \tag{A.3}$$

3. Accumulate the products in (A.3) for all the 3D points to end up with the following nine values:

$$\begin{aligned}
S_{XX} &= \sum_i P_{XX}^i \quad S_{YX} = \sum_i P_{YX}^i \\
S_{XY} &= \sum_i P_{XY}^i \quad S_{YY} = \sum_i P_{YY}^i \\
S_{XZ} &= \sum_i P_{XZ}^i \quad S_{YZ} = \sum_i P_{YZ}^i \\
S_{ZX} &= \sum_i P_{ZX}^i \quad S_{ZY} = \sum_i P_{ZY}^i \\
S_{ZZ} &= \sum_i P_{ZZ}^i
\end{aligned} \tag{A.4}$$

4. Form a 4x4 symmetric matrix with the elements in (A.4):

$$\mathbf{N} = \begin{pmatrix} N_{11} & N_{12} & N_{13} & N_{14} \\ N_{21} & N_{22} & N_{23} & N_{24} \\ N_{31} & N_{32} & N_{33} & N_{34} \\ N_{41} & N_{42} & N_{43} & N_{44} \end{pmatrix} \tag{A.5}$$

where

$$
\begin{aligned}
N_{11} &= S_{XX} + S_{YY} + S_{ZZ} \\
N_{12} &= N_{21} = S_{YZ} - S_{ZY} \\
N_{13} &= N_{31} = S_{ZX} - S_{XZ} \\
N_{14} &= N_{41} = S_{XY} - S_{YX} \\
N_{22} &= S_{XX} - S_{YY} - S_{ZZ} \\
N_{23} &= N_{32} = S_{XY} + S_{YX} \\
N_{24} &= N_{42} = S_{ZX} + S_{XZ} \\
N_{33} &= -S_{XX} + S_{YY} - S_{ZZ} \\
N_{34} &= N_{43} = S_{YZ} + S_{ZY} \\
N_{44} &= -S_{XX} - S_{YY} + S_{ZZ}
\end{aligned}
\tag{A.6}
$$

5. Find the eigenvector corresponding to the largest eigenvalue of $\mathbf{N}$, which will be the quaternion that determines the optimal rotation between the two sets of points.
6. Compute the rotation matrix ($\mathbf{R}$) associated to the so obtained quaternion, and compute the translation $\mathbf{t} = (\Delta x, \Delta y, \Delta z)^T$ as the difference between the centroid at time t and the rotated centroid at time $t+1$:

$$
\mathbf{t} = \mathbf{c}_t - \mathbf{R}\mathbf{c}_{t+1} \tag{A.7}
$$

7. Finally, we extract the values of the increments in *yaw*, *pitch*, and *roll* angles $\langle \Delta\alpha, \Delta\beta, \Delta\gamma \rangle$ between poses from the rotation matrix $\mathbf{R}$, having in this way all the components of μ_q.

Appendix B. Derivation of the landmark update equations

In this paper, we have employed a linear Kalman filter to estimate the 3D position of the landmarks in the map constructed during the SLAM process. This appendix addresses the derivation of the position update equations, from those of the Kalman Filter.

As stated in [23], the Kalman filter method comprises two different phases: *prediction* and *update*. In the prediction phase it is estimated the state of the system at the current time step $\left(\mathbf{x_{k|k-1}}\right)$ from the previous estimate $\left(\mathbf{x_{k-1|k-1}}\right)$ and a control action $(\mathbf{u_k})$, whereas the update phase refines the state estimation by introducing the information provided by the observation $(\mathbf{y_k})$, yielding a more accurate estimation $\left(\mathbf{x_{k|k}}\right)$. In addition, the KF provides a covariance matrix associated to the estimation uncertainty $\left(\mathbf{P_{k|k}}\right)$.

The equations stated for the discrete KF algorithm are as follows:

Prediction

$$
\begin{aligned}
\mathbf{x_{k|k-1}} &= \mathbf{F_k x_{k-1|k-1}} + \mathbf{B_k u_k} \\
\mathbf{P_{k|k-1}} &= \mathbf{F_k P_{k-1|k-1}} \mathbf{F_k}^{\mathrm{T}} + \mathbf{Q_k}
\end{aligned}
\tag{B.1}
$$

where $\mathbf{F_k}$ relates the state at the previous time step to that at the current step (also known as the *transition model*), $\mathbf{B_k}$ relates the control action with the state, and $\mathbf{Q_k}$ is the covariance of a zero-mean multi-variate gaussian distributed noise which affects the *a priori* estimation.

Update

$$
\begin{aligned}
\mathbf{x_{k|k}} &= \mathbf{x_{k|k-1}} + \mathbf{K_k y_k} \\
\mathbf{P_{k|k}} &= \left(\mathbf{I} - \mathbf{K_k H_k}\right) \mathbf{P_{k|k-1}}
\end{aligned}
\tag{B.2}
$$

where $\mathbf{H_k}$ is the observation model, which relates the state of the system to the observation at the current time step, $\mathbf{y_k}$ stands for the innovation, and $\mathbf{K_k}$ represents the Kalman filter gain, with expressions:

$$
\begin{aligned}
\mathbf{y_k} &= \mathbf{z_k} - \mathbf{H_k x_{k|k-1}} \\
\mathbf{K_k} &= \mathbf{P_{k|k-1}} \mathbf{H_k}^{\mathrm{T}} \left(\mathbf{H_k P_{k|k-1}} \mathbf{H_k}^{\mathrm{T}} + \mathbf{R_k}\right)^{-1}
\end{aligned}
\tag{B.3}
$$

being $\mathbf{R_k}$ the covariance of the zero-mean multi-variate gaussian distribution affecting the measurement process.

Note that, in this work, $\mathbf{x_{k|k-1}}$ and $\mathbf{x_{k|k}}$ represents the mean of the estimated position of a landmark in the map before and after the update process, respectively, while $\mathbf{z_k}$ stand for the observed position of the landmark at time step k. Besides, $\mathbf{P_{k|k-1}}$, $\mathbf{P_{k|k}}$ and $\mathbf{R_k}$ are the covariance matrixes of their associated uncertainties.

For clarity, let us take up again the notation introduced in section 4.3, renaming these variables as follows:

$$
\begin{aligned}
\mathbf{x_{k|k}} &\equiv \mu_m & \mathbf{P_{k|k}} &\equiv \Sigma_m \\
\mathbf{x_{k|k-1}} &\equiv \mu_{\tilde{m}} & \mathbf{P_{k|k-1}} &\equiv \Sigma_{\tilde{m}} \\
\mathbf{z_k} &\equiv \mu_k & \mathbf{R_k} &\equiv \Sigma_k
\end{aligned}
$$

Thus, with this change in the nomenclature and merging expressions (B.2) and (B.3), the KF prediction and update equations become:

$$
\begin{aligned}
\mu_{\tilde{m}} &= \mathbf{F}\mu_{\tilde{m}'} + \mathbf{B u_k} \\
\Sigma_{\tilde{m}} &= \mathbf{F}\Sigma_{\tilde{m}'}\mathbf{F}^{\mathrm{T}} + \mathbf{Q_k}
\end{aligned}
\tag{B.4}
$$

being $(\mu_{\tilde{m}'}, \Sigma_{\tilde{m}'})$ the estimation of the landmark position at the previous time step, and

$$
\begin{aligned}
\mu_m &= \mu_{\tilde{m}} + \Sigma_{\tilde{m}}\mathbf{H}^{\mathrm{T}} \left(\mathbf{H}\Sigma_{\tilde{m}}\mathbf{H}^{\mathrm{T}} + \Sigma_k\right)^{-1} \left(\mu_k - \mathbf{H}\mu_{\tilde{m}}\right) \\
\Sigma_m &= \left(\mathbf{I} - \Sigma_{\tilde{m}}\mathbf{H}^{\mathrm{T}} \left(\mathbf{H}\Sigma_{\tilde{m}}\mathbf{H}^{\mathrm{T}} + \Sigma_k\right)^{-1} \mathbf{H}\right) \Sigma_{\tilde{m}}
\end{aligned}
\tag{B.5}
$$

Note that the subscript $\mathbf{k}$ in matrixes $\mathbf{F}$, $\mathbf{B}$ and $\mathbf{H}$ has been dropped, since they are considered to be constant through time. Furthermore, in our work, the transition model and the control action for the landmarks in the map are very simple, since they remain static with time. Therefore, we do not consider $\mathbf{u_k}$ and take $\mathbf{F} = \mathbf{I}$, which entails that the predicted landmark positions are identical to their estimations at the previous time step:

$$\mu_{\tilde{m}} = \mu_{\tilde{m}'}$$
$$\Sigma_{\tilde{m}} = \Sigma_{\tilde{m}'}$$

On the other hand, as the observations in our work directly determine the 3D spatial coordinates of the landmarks model, we have $\mathbf{H} = \mathbf{I}$, thereby simplifying equations (B.5) considerably:

$$\mu_m = \mu_{\tilde{m}} + \Sigma_{\tilde{m}} \left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1} \left(\mu_k - \mu_{\tilde{m}}\right) \tag{B.6}$$

$$\Sigma_m = \left(\mathbf{I} - \Sigma_{\tilde{m}} \left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1}\right) \Sigma_{\tilde{m}} \tag{B.7}$$

These simplifications can be further transformed in order to get the more compact expressions shown in equations (12) in section 4.3, as derived next.

Regarding the covariance equation (B.7), we can substitute the identity matrix by $\left(\Sigma_{\tilde{m}} + \Sigma_k\right)\left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1}$ and factor out to obtain:

$$\begin{aligned}\Sigma_m &= \left(\left(\Sigma_{\tilde{m}} + \Sigma_k\right)\left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1} - \Sigma_{\tilde{m}} \left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1}\right) \Sigma_{\tilde{m}} \\ &= \Sigma_k \left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1} \Sigma_{\tilde{m}}\end{aligned} \tag{B.8}$$

which is equivalent to:

$$\Sigma_m = \left(\Sigma_k^{-1} + \Sigma_{\tilde{m}}^{-1}\right)^{-1} \tag{B.9}$$

On the other hand, the equation of the mean (B.6) can be expanded as follows:

$$\begin{aligned}\mu_m = \mu_{\tilde{m}} &+ \Sigma_{\tilde{m}} \left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1} \mu_k - \\ &- \Sigma_{\tilde{m}} \left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1} \mu_{\tilde{m}}\end{aligned} \tag{B.10}$$

and by factoring out the $\mu_{\tilde{m}}$ variable, it becomes:

$$\begin{aligned}\mu_m = &\left(\mathbf{I} - \Sigma_{\tilde{m}} \left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1}\right) \mu_{\tilde{m}} + \\ &+ \Sigma_{\tilde{m}} \left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1} \mu_k\end{aligned} \tag{B.11}$$

Now, the identity matrix can be substituted in the same way as with the covariance expression, and by factorizing out again, we obtain:

$$\mu_m = \Sigma_k \left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1} \mu_{\tilde{m}} + \Sigma_{\tilde{m}} \left(\Sigma_{\tilde{m}} + \Sigma_k\right)^{-1} \mu_k \tag{B.12}$$

Finally, by using equation (B.8), we can conclude that:

$$\begin{aligned}\mu_m &= \Sigma_m \Sigma_{\tilde{m}}^{-1} \mu_{\tilde{m}} + \Sigma_m \Sigma_k^{-1} \mu_k \\ &= \Sigma_m \left(\Sigma_{\tilde{m}}^{-1} \mu_{\tilde{m}} + \Sigma_k^{-1} \mu_k\right)\end{aligned} \tag{B.13}$$

which, together with equation (B.9), constitute the simplified expressions presented in section 4.3.

References

[1] M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, D. Sci, T. Organ, and S.A. Adelaide. A tutorial on particle filters for online nonlinear/non-GaussianBayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[2] M. Ballesta, A. Gil, O. Martinez-Mozos, and O. Reinoso. Local descriptors for visual SLAM. In *Workshop on Robotics and Mathematics*, Coimbra, Portugal, September 2007.

[3] M. Ballesta, A. Gil, O. Reinoso, and O. Martinez Mozos. Evaluation of interest point detectors for visual SLAM. *International Journal of Factory Automation, Robotics and Soft Computing*, 4:86–95, 2007.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[5] AC Berg, TL Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1, 2005.

[6] J. Civera, A.J. Davison, and J. Montiel. Inverse depth parametrization for monocular slam. *IEEE Transactions on Robotics*, 24(5):932–945, Oct. 2008.

[7] M.N. Dailey and M. Parnichkun. Landmark-based simultaneous localization and mapping with stereo vision. In *Proceedings of the Asian Conference on Industrial Automation and Robotics*, 2005.

[8] A.J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pages 1403–1410, 2003.

[9] A.J. Davison, Y.G. Cid, and N. Kita. Real-time 3D SLAM with wide-angle vision. *Proc. IFAC Symposium on Intelligent Autonomous Vehicles, Lisbon*, 2004.

[10] A.J. Davison, I. Reid, N. Molton, and O. Stasse. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, June 2007.

[11] F. Dellaert, W. Burgard, D. Fox, and S. Thrun. Using the CONDENSATION algorithm for robust, vision-based mobile robot localization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:588–594, 1999.

[12] M. Dissanayake, P. Newman, S. Clark, H.F. Durrant-Whyte, and M. Csorba. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotics and Automation*, 17(3):229–241, 2001.

[13] A. Doucet, N. de Freitas, K. Murphy, and S. Russell. Rao-Blackwellised particle filtering for dynamic Bayesian networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 176–183, 2000.

[14] A. Gil, Ó. Reinoso, O. Martinez Mozos, C. Stachniss, and W. Burgard. Improving Data Association in Vision-based SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2076–2081, 2006.

[15] G. Grisetti, C. Stachniss, and W. Burgard. Improved Techniques for Grid Mapping With Rao-Blackwellized Particle Filters. *IEEE Transactions on Robotics*, 23:34–46, February 2007.

[16] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of Alvey Vision Conference*, volume 15, 1988.

[17] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[18] J.R. Hershey and P.A. Olsen. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. volume 4, pages 317–320, 2007.

[19] B.K.P. Horn et al. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642, 1987.

[20] K. Hosoda, K. Sakamoto, and M. Asada. Trajectory generation for obstacle avoidance of uncalibrated stereovisual servoing without 3D reconstruction. *Proceedings of IEEE/RSJ International Conference on the Intelligent Robots and Systems (IROS).'Human Robot Interaction and Cooperative Robots'*, 1, 1995.

[21] S.J. Julier. The scaled unscented transformation. In *Proceedings of the American Control Conference*, volume 6, pages 4555–4559, 2002.

[22] S.J. Julier and J.K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls*, 3, 1997.

[23] R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[24] L. Kitchen and A. Rosenfeld. Gray-level corner detection. *Pattern Recognition Letters*, 1(2):95–102, 1982.

[25] J. Kosecka and F. Li. Vision based topological Markov localization. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2, 2004.

[26] S. Kullback. *Information Theory and Statistics*. Dover Publications, 1997.

[27] M. Li, B. Hong, Z. Cai, and R. Luo. Novel Rao-Blackwellized Particle Filter for Mobile Robot SLAM Using Monocular Vision. *International Journal of Intelligent Technology*, 1(1):63–69, 2006.

[28] DG Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, 1999.

[29] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[30] F. Lu and E. Milios. Robot Pose Estimation in Unknown Environments by Matching 2D Range Scans. *Journal of Intelligent and Robotic Systems*, 18(3):249–275, 1997.

[31] L. Matthies and S. Shafer. Error modeling in stereo navigation. *IEEE Journal of Robotics and Automation*, 3(3):239–248, 1987.

[32] E. Menegatti, A. Pretto, A. Scarpa, and E. Pagello. Omnidirectional vision scan matching for robot localization in dynamic environments. *IEEE Transactions on Robotics*, 22(3):523–535, 2006.

[33] J. Michels, A. Saxena, and A.Y. Ng. High speed obstacle avoidance using monocular vision and reinforcement learning. In *Proceedings of the 22nd international conference on Machine learning*, volume 2, pages 593–600. ACM Press New York, NY, USA, 2005.

[34] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1615–1630, 2005.

[35] M. Montemerlo. *FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem With Unknown Data Association*. PhD thesis, University of Washington, 2003.

[36] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, pages 593–598, 2002.

[37] K. Murphy. Bayesian map learning in dynamic environments. *Advances in Neural Information Processing Systems (NIPS)*, 12:1015–1021, 1999.

[38] D. Murray and J.J. Little. Using Real-Time Stereo Vision for Mobile Robot Navigation. *Autonomous Robots*, 8(2):161–171, 2000.

[39] I. Ohya, A. Kosaka, and A. Kak. Vision-based navigation by a mobile robot with obstacle avoidance using single-camera vision and ultrasonic sensing. *IEEE Transactions on Robotics and Automation*, 14(6):969–978, 1998.

[40] C.F. Olson, L.H. Matthies, M. Schoppers, and M.W. Maimone. Rover navigation using stereo ego-motion. *Robotics and Autonomous Systems*, 43(4):215–229, 2003.

[41] L.M. Paz, P. Pinies, J.D. Tardos, and J. Neira. Large-scale 6-dof slam with stereo-in-hand. *IEEE Transactions on Robotics*, 24(5):946–957, Oct. 2008.

[42] P. Pinies, T. Lupton, S. Sukkarieh, and J.D. Tardós. Inertial Aiding of Inverse Depth SLAM using a Monocular Camera. *IEEE International Conference on Robotics and Automation*, pages 2797–2802, 2007.

[43] S.J. Russell and P. Norvig. *Artificial intelligence: a modern approach, chapter 14*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1995.

[44] K. Sabe, M. Fukuchi, J.S. Gutmann, T. Ohashi, K. Kawamoto, and T. Yoshigahara. Obstacle avoidance and path planning for humanoid robots using stereo vision. In *Proccedings of the IEEE International Conference on Robotics and Automation (ICRA)*, volume 1, 2004.

[45] S. Se, D. Lowe, and J. Little. Local and global localization for mobile robots using visuallandmarks. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, 2001.

[46] S. Se, D. Lowe, and J. Little. Mobile Robot Localization and Mapping with Uncertainty using Scale-Invariant Visual Landmarks. *The International Journal of Robotics Research*, 21(8):735, 2002.

[47] J. Shi and C. Tomasi. Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[48] R. Sim, P. Elinas, M. Griffin, and J.J. Little. Vision-based SLAM using the Rao-Blackwellised particle filter. *IJCAI Workshop on Reasoning with Uncertainty in Robotics (RUR)*, 2005.

[49] R. Sim, P. Elinas, M. Griffin, A. Shyr, and J.J. Little. Design and analysis of a framework for real-time vision-based SLAM using Rao-Blackwellised particle filters. In *Proceedings of the 3rd Canadian Conference on Computer and Robot Vision*, pages 21–29, 2006.

[50] H. Tamimi, H. Andreasson, A. Treptow, T. Duckett, and A. Zell. Localization of mobile robots with omnidirectional vision using Particle Filter and iterative SIFT. *Robotics and Autonomous Systems*, 54:758–765, 2006.

[51] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. The MIT Press, September 2005.

[52] I. Ulrich and I.R. Nourbakhsh. Appearance-Based Place Recognition for Topological Localization. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1023–1029, 2000.

[53] E.A. Wan and R. Van Der Merwe. The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pages 153–158, 2000.

[54] J. Wolf, W. Burgard, and H. Burkhardt. Robust vision-based localization for mobile robots using an image retrieval system based on invariant features. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 1, 2002.