Managing Data Association in visual SLAM using SIFT features

Arturo Gil, Óscar Reinoso, Luis Payá, Mónica Ballesta, José M. Pedrero
Systems Engineering Department
Miguel Hernández University
03202 Elche (Alicante), SPAIN
Email: {arturo.gil,o.reinoso}@umh.es

Abstract: **This paper describes an approach to solve the Simultaneous Localization and Mapping (SLAM) problem for autonomous mobile robots using visual landmarks and a Rao-Blackwellized particle filter. Our map is represented by a set of three dimensional landmarks referred to a global reference frame. We use significant points extracted from stereo images as natural landmarks. In particular we employ SIFT features found in the environment. Each landmark is associated with a visual descriptor that partially differentiates it from others. We concentrate on a reduced set of highly stable landmarks. In order to do that, we track a visual feature for a significant number of frames prior to integrating it in the filter. As a result, we obtain different examples that represent the same natural landmark. Using this procedure, a better model for each landmark is obtained, which lets us improve data association among the landmarks in the map.**

Key-Words: Mobile Robots, Visual SLAM, Visual Landmarks, Particle Filter

1. INTRODUCTION

Building an accurate map of a given environment is one of the hardest tasks for a mobile robot. It is inherently difficult, since noise in the estimation of the robot pose leads to errors in the estimation of the map and vice versa. Here, we consider the problem of Simultaneous Localization and Mapping (SLAM) using a Rao-Blackwellized Particle Filter (RBPF).

Most work on SLAM so far has focussed on building 2D maps of environments using range sensors such as SONAR and laser [1], [2]. Recently, Rao-Blackwellized particle filters have been used as an effective mean of solving the SLAM problem using occupancy grid maps [3]. In this approach, each particle constructs its own map based on the observations and the trajectory for that particle.

Recently, some authors have been concentrating on building three dimensional maps using visual information extracted from cameras. Cameras are typically less expensive than laser sensors and are able to provide 3D information from the scene using stereo vision. In this scenario, the map is represented by a set of three dimensional landmarks related to a global reference frame. In [4] and [5] stereo vision is used to track 3D visual landmarks extracted from the environment. During the exploration phase, the robot extracts SIFT features from stereo images and calculates relative measurements to them. Landmarks are then integrated in the map with an EKF associated to each one. The work in [6] deploys a Rao-Blackwellized particle filter to estimate both the path and the map. Three dimensional SIFT features are extracted from the environment and integrated in the filter.

The major contribution of this paper is twofold. First, we present a mechanism to deal with the data association problem in the context of visual SLAM. Second, our approach tracks landmarks prior to integrating them in the map. As a result, only those landmarks that are more stable are incorporated in the map. By using this approach, our map typically consists of a reduced number of landmarks compared to those of [5] and [6], for comparable map sizes. The work presented here differs mainly from the work in [6] in two ways: First, we track each landmark for consecutive frames prior to integrating it in the filter, thus concentrate on a reduced set of highly stable landmarks. Second, we deploy an improved method to manage Data Association among the landmarks in the map, which improves the quality of the estimated path and the map.

The remainder of the paper is structured as follows. Section 2 deals with visual landmarks and their utility in SLAM. Section 3 explains the basics of the Rao-Blackwellized particle filter. Next, Section 4 presents our solution to the data association problem in the context of visual landmarks. In Section 5 we present our experimental results. Finally, Section 6 sums up the most important conclusions and proposes future extensions.

2. VISUAL LANDMARKS

In the approach presented here, we use SIFT (Scale Invariant Feature Transform) features as natural landmarks in the environment. SIFT features were developed for image feature generation, and used initially in object recognition applications (see [7] and [8] for further details). SIFT features are located at maxima and minima of a difference of Gaussian function applied in scale space. They are computed by building an image pyramid with resampling between each level. SIFT locations extracted with this procedure may be understood as significant points in space that are highly distinctive. In addition, each SIFT location is given a descriptor that provides invariance to image translation, scaling, rotation and partial invariance to illumination changes and view point changes. Thus, this fact enables the same points in the space to be recognized from different viewpoints, if the viewing angle does not differ too much. Lately, SIFT features have been used in robotic applications as visual landmarks for localization and SLAM tasks ([4], [5], [6]).

Given two images, captured with a stereo system, we extract natural landmarks which correspond to points in the 3-dimensional space. In order to do that, we extract SIFT features from the left and the right image of the stereo cameras. Each location is accompanied by its SIFT descriptor. Following, we find the correspondence for the points across images. The correspondence is constrained by the epipolar geometry of the stereo system. In addition, a comparison between SIFT descriptors associated to the keypoints is used to avoid false correspondences. As a result, at a time t we obtain a set of B observations denoted by $z_t = \{z_{t,1}, z_{t,2}, \ldots, z_{t,B}\}$. Each observation is constituted by $z_{t,k} = (v_{t,k}, d_{t,k})$, where $v_{t,k} = (X^l, Y^l, Z^l)$ is a three dimensional vector referred to the left camera reference frame and $d_{t,k}$ is the SIFT descriptor associated to that point.

3. RAO-BLACKWELLIZED SLAM

Following the usual nomenclature in Rao-Blackwellized SLAM, we call x_t the robot pose at time t. On the other hand, the robot path until time t will be denoted as $x^t = \{x_1, x_2, \ldots, x_t\}$, the set of observations made by the robot until time t will be denoted $z^t = \{z_1, z_2, \ldots, z_t\}$ and the set of actions $u^t = \{u_1, u_2, \ldots, u_t\}$. We formulate the SLAM problem as that of determining the location of all landmarks in the map m and robot poses x^t from a set of measurements z^t and robot actions u^t. Thus, it can be stated as the estimation of the posterior over robot paths and maps $p(x^t, m|z^t, u^t, c^t)$.

While exploring the environment, the robot has to determine whether a particular observation $z_{t,k} = (v_{t,k}, d_{t,k})$ corresponds to a previously seen landmark or to a new one. This problem is known as the Data Association problem and will be further explained in Section 4. Provided that, at a time t the map is formed by N landmarks, the correspondence is represented by $c_t = \{c_{t,1}, c_{t,2}, \ldots, c_{t,B}\}$, where $c_{t,i} \in [1 \ldots N]$. In consequence, at a time t the observation $z_{t,k}$ corresponds to the landmark $c_{t,k}$ in the map. When no correspondence is found we denote it as $c_{t,i} = N + 1$, indicating that a new landmark should be initialized.

The map m is represented by a collection of N landmarks $m = \{\theta_1, \theta_2, ..., \theta_N\}$. Each landmark is described as: $\theta_k = \{\mu_k, \Sigma_k, d_k\}$, where $\mu_k = (X^g_k, Y^g_k, Z^g_k)$ is a vector describing the position of the landmark k referred to a global reference frame with associated covariance matrix Σ_k. In addition, each landmark θ_k is associated with a SIFT descriptor d_k. This map representation is compact and has previously been used to localize a robot in indoor environments [9].

3.1. Particle filter estimation

The conditional independence property of the SLAM problem implies that the posterior over robot paths and maps can be factored as [10]:

$$p(x^t, m|z^t, u^t, c^t) = p(x^t|z^t, u^t, c^t) \prod_{k=1}^{N} p(\theta_k|x^t, z^t, u^t, c^t) \tag{3.1}$$

This equation states that the full SLAM posterior is decomposed into two parts: one estimator over robot paths, and N independent estimators over landmark positions, each conditioned on the path estimate. We approximate the posterior $p(x^t|z^t, u^t, c^t)$ using a set of M particles, each particle having N independent landmark estimators (implemented as EKFs), one for each landmark in the map. Each particle is thus defined as $S^{[m]}_t = \{x^{t,[m]}, \mu^{[m]}_{t,1}, \Sigma^{[m]}_{t,1}, \ldots, \mu^{[m]}_{t,N}, \Sigma^{[m]}_{t,N}\}$, where $\mu^{[m]}_{t,i}$

is the best estimation at time t for the position of landmark θ_i based on the path of the particle m and $\Sigma_{t,i}^{[m]}$ is the associated covariance matrix. The particle set $S_t = \{S_t^{[1]}, S_t^{[2]}, \ldots, S_t^{[M]},\}$ is calculated incrementally from the set S_{t-1} at time $t-1$ and the robot control u_t. Thus, each particle is sampled from a proposal distribution $x_t^{[m]} \sim p(x_t | x_{t-1}, u_t)$. Next, and following the approach of [10] each particle is then assigned a weight according to:

$$\omega_{t,i}^{[m]} = \frac{1}{\sqrt{|2\pi Z_{c_{t,i}}|}} e^{\{-\frac{1}{2}(v_{t,i} - \hat{v}_{t,c_{t,i}})^T [Z_{c_{t,i}}]^{-1} (v_{t,i} - \hat{v}_{t,c_{t,i}})\}} \tag{3.2}$$

Where $v_{t,i}$ is the current measurement and $\hat{v}_{t,c_{t,i}}$ is the predicted measurement for the landmark $c_{t,i}$ based on the pose $x_t^{[i]}$. The matrix $Z_{c_{t,i}}$ is the covariance matrix associated with the innovation $(v_{t,i} - \hat{v}_{t,c_{t,i}})$. Note that we implicitly assume that each measurement $v_{t,i}$ has been assigned to the landmark $c_{t,i}$ of the map. This problem is, in general, hard to solve, since similar-looking landmarks may exist. In section 4. we describe our approach to this problem. In the case that B observations from different landmarks exist at a time t, we calculate the total weight assigned to the particle multiplying the weights computed using Equation (3.2).

4. DATA ASSOCIATION

While the robot moves through the environment, it must decide whether the observation $z_{t,k} = (v_{t,k}, d_{t,k})$ corresponds to a previously mapped landmark or to a different landmark. In most existing approaches ([4], [5], [7]) the data association is performed using the squared Euclidean distance between SIFT descriptors

$$E = (d_i - d_j)(d_i - d_j)^T, \tag{4.1}$$

where d_j is the SIFT descriptor associated with the current measurement, while d_i is the descriptor associated with a landmark in the map. Then, the landmark in the map that minimizes the distance E is regarded as the correct data association. Whenever the distance E is below a certain threshold, the two landmarks are considered to be the same. Otherwise, a new landmark is created. Figure 1 shows the same point in space as seen from different viewpoints. We experimentally compared the SIFT descriptor of the same point in the different frames. When the same point is viewed from slightly different viewpoints and distances (e.g. Figure 1(a)-(c)), the distance E remains low. However, when the same point is viewed from significantly different viewpoints (e.g. Figure 1(a)-(d)) the difference in the descriptor is remarkable and the check using the Euclidean distance is likely to produce a wrong data association. Figure 2(a) compares the two SIFT vectors computed for the views (a) and (b) of the point in Figure 1. We can observe that, in views (a)-(b) the vectors remain similar. Figure 2(b) compares the two SIFT vectors computed for the views (a) and (d). We can clearly observe that the vectors are significantly different in the latter case.

We propose a different method to deal with the data association in the context of SIFT features. We address the problem from a pattern classification point of view. We consider the problem of assigning a pattern d_j to a class C_i, where each class C_i models a landmark. Whenever a landmark is found, it is tracked along p frames, and its descriptors $d_1, d_2, \ldots, d_p$ are stored. We consider different views of the same visual landmark as different elements of class C_i and compute a mean value $\bar{d}_i$ that represents the prototype of the class C_i. A covariance matrix S_i is estimated, assuming the elements in the SIFT vector independent of each other. Thus, S_i is a diagonal matrix whose elements are the variance of each element in the SIFT vector, computed using the p example vectors of the landmark. Whenever a new landmark d_j is found, we compute the squared Mahalanobis distance to each stored landmark, represented by $\bar{d}_i$ and S_i as

$$L = (\bar{d}_i - d_j) S_i^{-1} (\bar{d}_i - d_j)^T. \tag{4.2}$$

We compute the distance L for the landmarks in the map of each particle and assign the correspondence to the landmark that minimizes L. If none of the values exceeds a predefined threshold, we create a new landmark. As we will show in the experiments, this technique allows us to make better data associations and, as a result, produce better maps of the environment.

In order to test this distance function we have recorded a set of images with slight variations of viewpoint and distance (see Figure 1). SIFT landmarks are easily tracked across consecutive frames, since the variance in the descriptor is low. In addition, we visually judged the correspondence across images. Based on these data we computed the matrix S_i for each SIFT point tracked for more than 5 frames. Following, we computed the distance to the same class using Equation (4.1) and (4.2). For each experiment, we select the class that minimizes the distance function. Since we already know the truth correspondences, we can compute the number of mistakes and correct matches. A total of 3000 examples were used for this purpose. Using the Euclidean distance we obtained a 83.85% of correct matches. When using the squared Mahalanobis distance, a 94.04% of correct matches were obtained. The number of correct matches is significantly increased by using the squared Mahalanobis distance (4.2). Since most of the false correspondences are avoided, we can obtain better estimates of the map and the path traversed by the robot.

5. EXPERIMENTAL RESULTS

During the experiments we have used a B21r robot equipped with a stereo head and a LMS laser range finder. We manually steered the robot along the rooms of the building 79 of the University of Freiburg. A total of 507 stereo images at a resolution of 320x240 were collected. The total traversed distance of the robot is approximately $80m$. For each pair of stereo images a number of correspondences were established and observations were obtained. After stereo correspondence, each point is tracked for a number of frames. In a practical way, when a landmark has been tracked for more than 5 frames it is considered a new observation and is integrated in the filter.

Figure 4 shows a map built using 100 particles. A total number of 1500 landmarks were estimated. It can be seen that, with only 100 particles, the map is topologically correct. Some areas of the map do not possess any landmark, which correspond to feature-less areas (i.e. texture-less walls), where no SIFT features were found.

In order to test the quality of our results, we compared the estimated pose of our method with the estimated pose using laser data recorded during exploration. In order to calculate the pose based on laser measurements, the method exposed in [3] was used. Figure 3(a) shows the error in localization for each movement of the robot during exploration using 100 particles.

In addition, we have compared both approaches to data association as described in Section 4.. To do this, we have made a number of simulations varying the number of particles used in each simulation. The process was repeated using both data association methods. As can be seen in Figure 3(b) for the same number of particles, better localization results are obtained when the squared Mahalanobis distance is used, thus improving the quality of the estimated map.

Finally, our maps typically consist of about 1500 landmarks, a much more compact representation than the presented in [6], where the map contains typically around 10.000 landmarks.

6. CONCLUSION

We have presented a solution to SLAM based on a Rao-Blackwellized particle filter that uses visual information extracted from cameras. In particular SIFT features have been used as natural landmarks. The method is able to build 3D maps of an indoor environment using relative measurements extracted from a stereo pair of cameras.

We also have proposed an alternative method to deal with the data association problem in the context of visual landmarks. When different examples of a particular SIFT descriptor exist (belonging to the same landmark) we obtain a probabilistic model for it. By this procedure, the data association is improved, and consequently, better results are obtained since most of the false correspondences are avoided.

Maps created by this procedure do not directly represent the occupied or free areas of the environment. In consequence, the map can be used to effectively localize the robot, but cannot be directly used for navigation. This fact is derived from the nature of the sensors and it is not a failure of the proposed approach. For navigation tasks other low-cost sensors such as SONAR should be deployed.

Acknowledgments

This research has been supported by the spanish government through project DPI2004-07433-C02-01 (Ministerio de Educación y Ciencia. Título: HERRAMIENTAS DE TELEOPERACIÓN COLABORATIVA. APLICACIÓN AL CONTROL COOPERATIVO DE ROBOTS), and project PCT-G54016977-2005 (FUNDACIÓN QUORUM: PARQUE CIENTÍFICO Y EMPRESARIAL DE LA UNIVERSIDAD MIGUEL HERNÁNDEZ. Título: ROBOTS COOPERATIVOS PARA LA VIGILANCIA E INSPECCIÓN DE EDIFICIOS E INSTALACIONES INDUSTRIALES).

References

[1] O. Wijk and H. I. Christensen. Localization and navigation of a mobile robot using natural point landmarkd extracted from sonar data. *Robotics and Autonomous Systems*, 1(31):31–42, 2000.

[2] S. Thrun. A probabilistic online mapping algorithm for teams of mobile robots. *International Journal of Robotics Research*, 20(5):335–363, 2001.

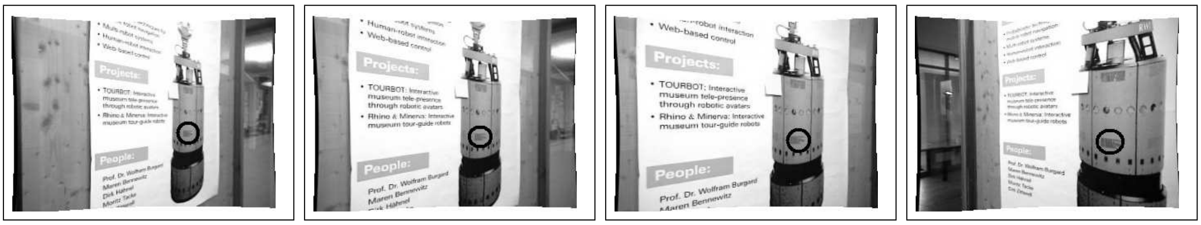
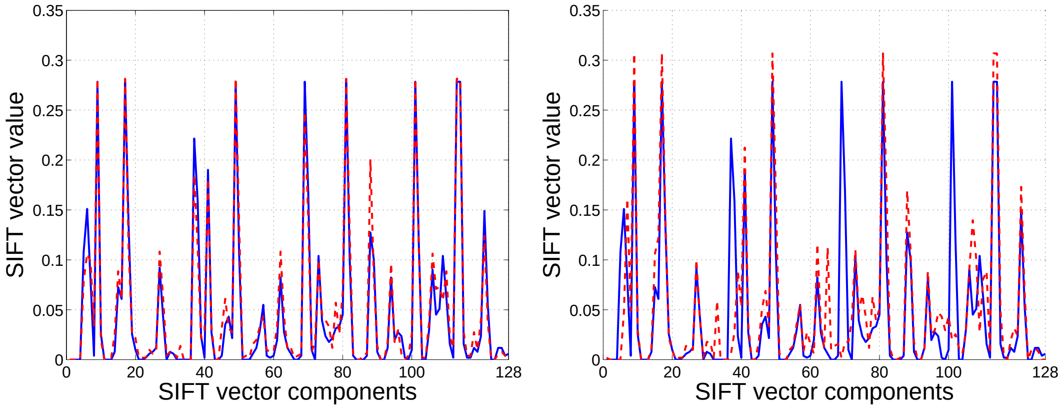


Figure 1: The images depict the same landmark (marked with a circle) viewed from different viewpoints. The squared Euclidean distance between consecutive images is between 0.03 and 0.06. In contrast to that, the squared Euclidean distance between non-consecutive images is between 0.3 and 0.4, which is around one order of magnitude larger.



(a)

(b)

Figure 2: Figure (a) compares two SIFT vectors corresponding to close views of the same point. Figure (b) compares two SIFT vectors corresponding to separate views of the same point. As can be seen, the SIFT descriptor does not provide total invariance to viewpoint changes.

[3] C. Stachniss, D. Haehnel, and W. Burgard. Exploration with active loop-closing for FastSLAM. In *IEEE/RSJ Int. Conference on Intelligent Robots and Systems*, 2004.

[4] J. Little, S. Se, and D. Lowe. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2051–2058, 2001.

[5] J. Little, S. Se, and D. Lowe. Global localization using distinctive visual features. In *Proceedings of the 2002 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems*, 2002.

[6] R. Sim, P. Elinas, M. Griffin, and J. J. Little. Vision-based slam using the rao-blackwellised particle filter. In *IJCAI Workshop on Reasoning with Uncertainty in Robotics*, 2005.

[7] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60):91–110, 2004.

[8] D. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, 1999.

[9] A. Gil, O. Reinoso, A. Vicente, C. Fernández, and L. Payá. Monte carlo localization using sift features. *Lecture Notes in Computer Science (LNCS)*, 1(3523):623–630, 2005.

[10] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. Fastslam: A factored solution to the simultaneous localization and mapping problem. In *AAAI*, 2002.

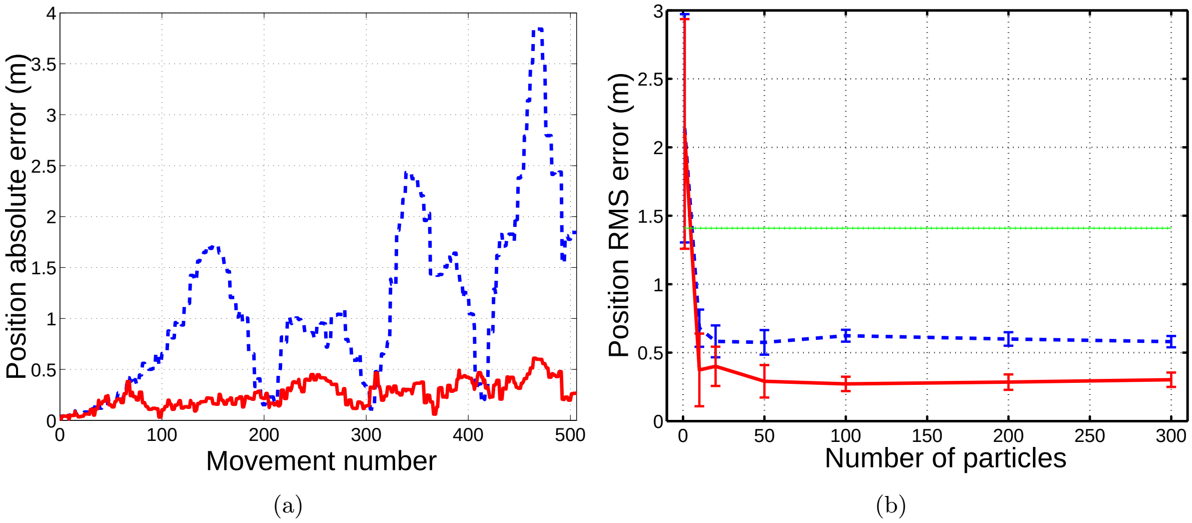


Figure 3: Figure (a) shows the absolute position error during the SLAM process. Figure (b) shows the error in localization when varying the number M of particles. The RMS error in odometry is shown as a dotted line. The results using Equation (4.1) are shown as a dashed line and results using Equation (4.2) are shown as a continuous line.

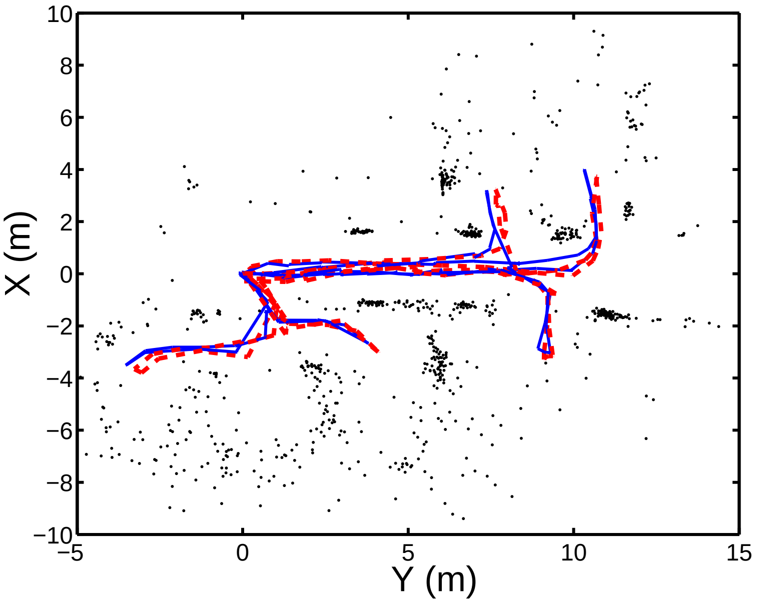


Figure 4: The figure shows a map created using 100 particles. Each black point represents a landmark. We also show superimposed the ground truth path estimated using laser range data (continuous) and the estimated path using our approach (dashed).