Visual Bearing-Only Simultaneous Localization and Mapping with Improved Feature Matching

Hauke Strasdat, Cyrill Stachniss, Maren Bennewitz, and Wolfram Burgard

Computer Science Institute, University of Freiburg, Germany

Abstract. In this this paper, we present a solution to the simultaneous localization and mapping (SLAM) problem for a robot equipped with a single perspective camera. We track extracted features over multiple frames to estimate the depth information. To represent the joint posterior about the trajectory of the robot and a map of the environment, we apply a Rao-Blackwellized particle filter. We present a novel method to match features using a cost function that takes into account differences between the feature descriptor vectors as well as spatial information. To find an optimal matching between observed features, we apply a global optimization algorithm. Experimental results obtained with a real robot show that our approach is robust and tolerant to noise in the odometry information of the robot. Furthermore, we present experiments that demonstrate the superior performance of our feature matching technique compared to other approaches.

1 Introduction

Mapping is one of the fundamental problems in mobile robotics since representations of the environment are needed for a series of high level applications. Without an appropriate model of the environment, for example, delivery tasks cannot be carried out efficiently. A large group of researchers investigated the so-called simultaneous localization and mapping (SLAM) problem. The majority of approaches focuses on proximity sensors to perceive the environment such as laser range finders, sonars, radars, or stereo vision cameras.

In this paper, we address the problem of learning maps using a mobile robot equipped with a single perspective camera only. Compared to a laser range finder, cameras have the advantage that they are cheap and lightweight. One of the problems, however, is the missing distance information to observed landmarks. This information is not provided by a perspective camera. We present a mapping system that can use this sensor setup to learn maps of the environment. Our approach applies a Rao-Blackwellized particle filter to maintain the joint posterior about the trajectory of the robot and the map of the environment. We furthermore present a novel method to establish the data association between features. It takes into account the individual feature descriptor vectors as well as spatial constraints. Our approach is able to compute the optimal matching between observed and already tracked features. To achieve this, we apply the Hungarian method which is an efficient global optimization algorithm. Experiments carried out with a real robot illustrate the advantages of our technique for learning maps with robots using a single perspective camera.

2 Related Work

Davison et al. [1,2] proposed a visual SLAM approach using a single camera that does not require odometry information. The system works reliable in room-size environments but is restricted in the number of landmarks it can handle. Landmarks are matched by looking back into the image at the expected region and by perfoming a local match. Sim et al. [3] use a stereo camera in combination with FastSLAM [4]. SIFT features [5] in both cameras are matched using their description vectors as well as the epipolar geometry of the stereo system. The matching between observations and landmarks is done using the SIFT descriptor only. In the bearing-only algorithm of Lemaire et al. [6], the feature depth is estimated using a mixture of Gaussians. The Gaussians are initialized along the first observation and they are pruned in the following frames.

3 Visual SLAM and Feature Matching

The joint posterior about the robot's trajectory and the map is represented by a Rao-Blackwellized Particle Filter (RBPF) similar to FastSLAM [4]. It allows the robot to efficiently model the joint posterior in a sampled fashion.

To obtain landmarks, we extract speeded-up visual features (SURF) [7] out of the camera images. These features are invariant to translation and scale. They can be extracted using a Fast-Hessian keypoint detector. The 64-dimensional feature descriptor vector d is computed using horizontal and vertical Haar wavelet responses. A rotational dependent version of SURF is used since the roll angle of the camera is fixed when it is attached to a wheeled robot.

In order to obtain spherical coordinates of a feature given its position in the image, we apply a standard camera model. In this way, pixel coordinates of detected keypoints are transformed into the azimuthal angle θ and the spherical angle ϕ. The distance ρ to the observed feature cannot be measured since we use only a monocular camera. The tuple (θ, ϕ) is referred to as bearing-only observation $\mathbf{z}$.

3.1 Observation Model

In this section, we assume that a map of 3D-landmark is given. Each landmark l is modeled by a 3D Gaussian (μ, Σ). Moreover, we assume data association problem between observed features and landmarks is solved. These assumption are relaxed in the subsequent sections.

For each particle k, each observation $\mathbf{z} = (\theta, \phi)^T$ perceived in the current frame is matched with a landmark $l \in M^{[k]}$, where $M^{[k]}$ is the map carried by particle k. For each complete assignment of the currently observed features to map features, an update of the Rao-Blackwellized particle filter is carried out.

In order to determine the likelihood of an observation $\mathbf{z}$ in the update step of the particle filter, we need to compute the predicted observation $\hat{\mathbf{z}}$ of landmark $l = (\mu, \Sigma)$ for particle k. To achieve this, we have to apply two transformations. First, we transform world coordinates $\mu = (\mu_1, \mu_2, \mu_3)^T$ into camera-centric coordinates $c = (c_1, c_2, c_3)^T$

using the function g. Afterwards, the camera-centric Cartesian point c is transformed into camera-centric spherical coordinates $\hat{\mathbf{z}} = (\hat{\rho}, \hat{\theta}, \hat{\phi})^T$ using the function h:

$$\hat{\mathbf{z}} = h(g(\mu_l, \mathbf{x}^{[k]})) \tag{1}$$

Here, $\mathbf{x}^{[k]}$ is the current pose of particle k. The corresponding measurement uncertainty Q is predicted using the Jacobean $G = h'(g'(\mu_l, x^{[k]}))$ as

$$Q = G \cdot \Sigma \cdot G^T + \mathrm{diag}(\sigma_\rho, \sigma_\theta, \sigma_\phi). \tag{2}$$

Here, Σ is the covariance matrix corresponding to landmark l, and σ_θ and σ_ϕ represent the uncertainty over the two spherical angles. The uncertainty over the depth σ_ρ is set to a high value in order to represent the bearing-only aspect of the update. The observation likelihood λ is based on a Gaussian model as

$$\lambda = |Q|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \begin{pmatrix} 0 \\ \theta - \hat{\theta} \\ \phi - \hat{\phi} \end{pmatrix}^T Q^{-1} \begin{pmatrix} 0 \\ \theta - \hat{\theta} \\ \phi - \hat{\phi} \end{pmatrix} \right). \tag{3}$$

Since the depth ρ to the observed feature is unknown, the pretended innovation $(\rho - \hat{\rho})$ is set to zero. We weight each particle k with respect to its observation likelihood λ.

Finally, the Kalman gain is calculated by $K = \Sigma \cdot G^T \cdot Q^{-1}$ so that the landmark (μ, Σ) can be updated using an *extended Kalman filter* (EKF) approach.

3.2 Depth estimation and Landmark Initialization

Although it is possible to integrate bearing-only observations into the RBPF, the full 3D information is necessary in order to initialize landmarks in a 3D map. We track features over consecutive frames and estimate the depth ρ of the features using discrete probability distributions similar to [1] but in a bottom-up manner. When a feature f is initially observed, a 3D ray is cast from the camera origin o towards the observed feature. Equally weighted bins $b^{[j]}$ – representing different distances $\rho^{[j]}$ – are distributed uniformly along this ray within a certain interval. This reflects the fact that initially the distance to the feature is unknown. To get an estimate about the depth of the features, they are tracked over consecutive frames (the next subsection explains the feature matching process). In case the initial feature f is matched with a feature $\bar{f}$ in the consecutive frame, the bins are projected back into that frame. They lie on the so-called *epipolar line* [8], the projection of the 3D ray into the image. The depth hypotheses $\rho^{[j]}$ are weighted according to the distance to the pixel location of feature $\bar{f}$ using a Gaussian model. Figure 1 illustrates the estimation process for two features in consecutive frames. As soon as the variance of the histogram $Var(\rho^{[j]})$ falls below a certain threshold, the depth is estimated by the weighted average over the histogram $\rho = \sum_j h^{[j]} \cdot \rho^{[j]}$. If it is not possible to initialize a landmark within n frames, the corresponding feature is discarded (here n=5).

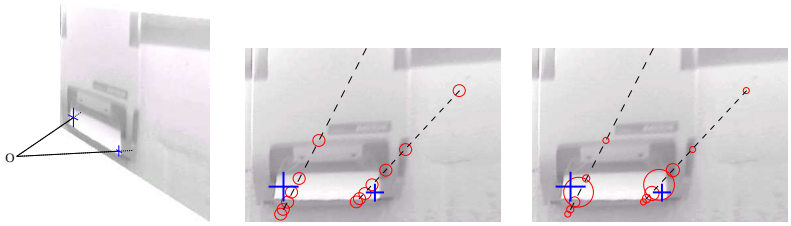

Fig. 1. This figure shows the depth estimation process for two features (crosses). Left: A ray is cast from the camera origin through the initial feature. Center: The ray is re-projected in the consecutive frame. This line (dashed) is called *epipolar line*. Depth hypotheses (circles) were distributed uniformly on the ray in the Cartesian space, which results in an irregular distribution in the image space. Right: Hypotheses are weighted according to their distance to the corresponding feature.

Depth Estimation as Preprocessing Step The robot's pose at the point in time, when the corresponding feature is observed initially, determines where the 3D ray is located in the world. Naïvely, for each particle a histogram of depth hypotheses has to be maintained so that the bins can be updated accordingly to the individual particle poses. However, this would lead to an overhead in computation time and memory. Fortunately, it is possible to maintain a depth histogram independently of the particles. The 3D ray is described by the angles θ, ϕ and an arbitrary origin o. Over the following n frames, the relative motion is added to o, so that the projection of the hypotheses' positions into the current frame can be calculated. Since the motion noise for wheeled robots is negligible within n frames, it can be omitted for the depth estimation process only.

Landmark Initialization Once a feature is reliably tracked and the depth of a feature is estimated, a landmark l is initialized. This has to be done for each particle k individually. Using the particle pose at the time t_f when feature f was observed for the first time, the global Cartesian landmark position μ can be calculated by

$$\mu_l := g^{-1}\left(h^{-1}\left(\rho_f, \theta_f, \phi_f\right), \mathbf{x}_{t_f}^{[k]}\right), \tag{4}$$

whereas the landmark uncertainty results from

$$\Sigma := G^{-T} \cdot H^{-1} \cdot R \cdot H^{-T} \cdot G^{-1}. \tag{5}$$

The uncertainty over the depth estimation process is reflected by the diagonal covariance matrix $R = \mathrm{diag}(Var(\rho_f^{[j]}), \sigma_\theta, \sigma_\phi)$.

3.3 Data Association

Finally, we describe how to match the current feature observations with landmarks in the map as well as with tracked features which are not yet contained in the map. This is done using the Hungarian method [9]. The Hungarian method is a general method to determine the optimal assignment under a given cost function. In our case, we use a cost function that takes into account differences of the feature descriptor vectors as well as well as spatial information to determine matches between observations and tracked features as well as between observations and landmarks.

Feature Matching Intuitively, features that are tracked to obtain the corresponding depth ρ can be matched based on their descriptor vectors using the Euclidean distance. However, this approach has a serious short-coming. Its performance is low on similar looking features since it completely ignores the feature positions. Thus, we instead use the distance of the descriptor vectors as a hard constraint. Only if the Euclidean distance falls below a certain threshold, a matching is considered. We define the cost function by means of the epipolar line introduced in Section 3.2. By setting the matching cost to the distance of the feature to this line in the image space (see Figure 2), not only the pixel locations of the comparing features are considered but also the relative movement of the robot between the corresponding frames is incorporated.

Landmark Matching Using Observation Likelihoods Similarly, during the matching process between landmarks and observations we use the distance of the descriptor vectors as a constraint. Landmarks are matched with observations using their positions. Since the observations are bearing-only, the distance to the landmark position cannot be computed directly. For this reason, the observation likelihood in Eq. (3) is used. It is high if and only if the distance between the observation and prediction is small. Thus, the cost is defined by the reciprocal of the observation likelihood $1/\lambda$. If the observation likelihood lies below a certain threshold, the cost are set to a maximum value. This refers to the fact that the features are regarded as different features with probability one.

4 Experimental Results

The first mapping experiment is performed on a wheeled robot equipped with a perspective camera and a laser range scanner (see Figure 3). The robot was steered through a 10m by 15m office environment for around 10 minutes. Two camera frames per second and odometry data was recorded. In addition, laser range data is stored in order to calculate a ground truth estimate of the robot's trajectory using scan matching on the laser data [10].

The results are illustrated in Figure 4. Following the presented approach, the average error of the robot path in terms of the Euclidean distance in the x/y-plane is 0.28m. The error in the orientation averages 3.9°. Using the odometry of the robot only, one obtains an average error of 1.69m in the x/y-plane and 22.8° in orientation.

We compared our feature matching approach using the Hungarian method on the distance to the projected line to other three techniques. Figure 5 shows a qualitative

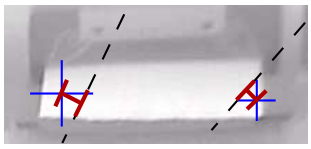

Fig. 2. Hungarian Matching: The cost function is set to the distance between the epipolar line and the feature location in the image space.

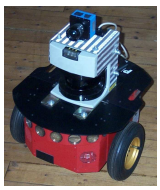

Fig. 3. Wheeled robot equipped with a perspective camera.

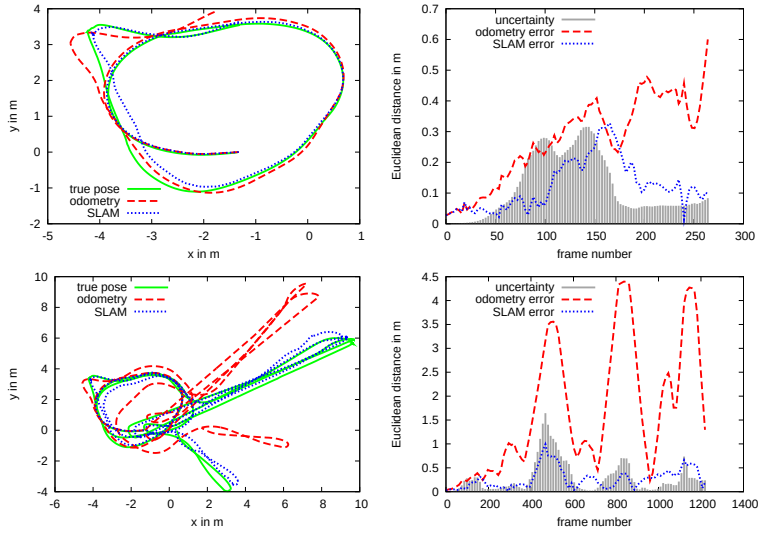

Fig. 4. The robot's trajectory is shown on the left, the corresponding error functions and uncertainty are shown on the right. Top: If the robot explores an unknown environment, the error values go up as well as the uncertainty. As soon as the loop is closed, the estimation error and uncertainty decreases, whereas the odometry error still goes up. Bottom: Complete trajectory.

evaluation of our approach on a difficult example – a heater which has a number of very similar looking features close to each other. Furthermore, we compare the Hungarian method quantitatively to the nearest neighbor approach, both using the distance to the epipolar line as cost function. If the Hungarian method is used, approximately 2% more landmarks are initialized. This number – obtained by four different sequences of 500 images each – can be explained by the fact that mismatches are likely to yield too high variances in the depth estimation. Landmarks, however, are initialized only if the depth can be estimated with low variance. By manual inspection, one can see that the data association has less errors than the nearest neighbor approach (see Figure 5).

5 Conclusions

In this paper, we presented a novel technique for learning maps with a mobile robot equipped with a single perspective camera only. Our approach applies a RBPF to maintain the joint posterior about the trajectory of the robot and the map of the environment. Using our approach, the robot is able to compute the optimal data association between observed and already mapped features by applying the Hungarian method. Experiments carried out with real a robot showed the effectiveness of our approach.

Acknowledgment

This work has partialy been supported by the German Research Foundation (DFG) under the contract numbers SFB/TR-8 and BE 2556/2-1. Special thanks to Dieter Fox,

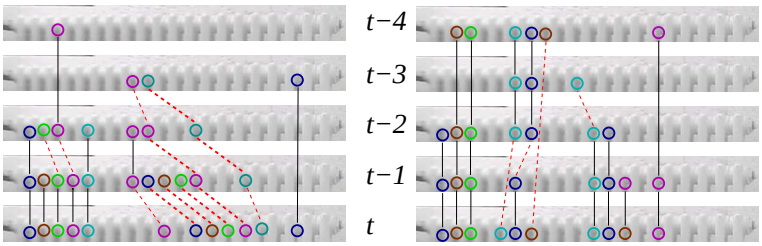

Fig. 5. Starting from the current frame at time *t*, we look back to evaluate how many features in the current frame were reliable tracked over the last four frames. The nearest neighbor assignment on SURF descriptors (left) results in 10 matches and 12 mismatches, whereas our approach results in 20 matches and 4 mismatches (right). Our approach also outperforms the two other combinations: nearest neighbor assignment using the projected line (14 matches, 7 mismatches) and the Hungarian method on the SURF descriptors (14 matches, 4 mismatches).

who supervised the first author in the early stages of developing the presented framework. We would like to thank Herbert Bay and Luc Van Gool for making the SURF binaries publicly available.

References

1. Davison A: Real-time simultaneous localization and mapping with a single camera. In Proc. of European Conf. on Computer Vision (ECCV), 2003.
2. Davision A, Reid I, Molton N, and Stasse O: MonoSLAM: Real-time single camera SLAM. IEEE Transaction on Pattern Analysis and Machine Intelligence 29(6), 2007.
3. Sim R, Elinas P, Griffin M, and Little J: Vision-based SLAM using a Rao-Blackwellized particle filter. In Proc. of IJCAI Workshop on Reasoning with Uncertainty in Robotics, 2005.
4. Montemerlo M, Thrun S, Koller D, and Wegbreit B: FastSLAM: A factored solution the to simultanieous localization and mapping problem. In Proc. of National Conference on Artificial Intelligence (AAAI), 2002.
5. Lowe D: Distinctive image feature from scale-invariant keypoints. In Proc. of International Journal of Computer Vision (IJCV), 2003.
6. Lemaire T, Lacroix S, and Sol J: A practical bearing-only SLAM algorithm. In Proc. of IEEE International Conf. on Intelligent Robots and Systems (IROS), 2005.
7. Bay H, Tuytelaars T, and Van Gool L: SURF: Speeded up robust features. In Proc. of European Conf. on Computer Vision (ECCV), 2006.
8. Hartley R and Zisserman A: Multiple View Geometry in Computer Vision. Cambirdige university press, second edition, 2003.
9. Kuhn H: The Hungarian method for the assigment problem. Naval Research Logistic Quaterly, 2:83-97, 1955.
10. Lu F and Milios E: Robot pose estimation in unkown environments by matching 2d range scans. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 935-938, 1994.