Chapter 2

Background

Vision-based SLAM is an active research area, but thus far most of the existing systems construct either occupancy grids or topological maps (see [10] for a survey and [11] for a more recent example), and these approaches are inappropriate for large scale metric mapping. For large scale metric maps, the simplest approach is to represent the world with a sparse collection of *landmarks*. These landmarks could be distinctive-looking 3D points or more complex objects such as lines, curves, corners, and so on.

Early vision-based SLAM systems did use sparse features, but they typically compressed the map to 2D. For example, Kriegman, Triendl, and Binford's system [12] uses a stereo sensor to extract vertical lines from the environment. Observed lines are used to reduce odometric uncertainty using an extended Kalman filter (EKF), then the observations are in turn used to update an environment map containing 2D point features representing the observed vertical lines. Yagi, Nishizawa, and Yachida's system [13] took a similar approach but used a single omnidirectional vision sensor and accumulation of measurements over time, rather than stereo, to determine the positions of vertical line landmarks. These systems and others have amply demonstrated the efficacy of vision-based SLAM based on line landmarks in constrained indoor environments with smooth floors.

Faugeras and colleagues [14, 15] were the first to develop a vision-based SLAM system storing a sparse 3D map. Their system first constructs a "local" 3D line segment map of the current scene using trinocular stereo. It explicitly represents the uncertainty about each feature's robot-relative pose in the form of a covariance matrix. The new local map is registered against the current global map and used to update an estimate of the robot's position using an EKF. Finally, assuming the robot's position, the global map is be updated with the freshly observed features, again using EKFs.

Se, Lowe, and Little [6] demonstrate the use of SIFT (the scale invariant feature transform) [16] point features as landmarks for the vision-baaed SLAM problem. Their system also uses a trinocular stereo camera rig and models the positional uncertainty of the landmarks with Kalman filters.

Sim and Dudek [17] take a different approach; rather than prespecifying the features (lines, points, corners, and so on) that should be used for map building and localization, their system learns generative models for the appearance of salient features during exploration.

Until quite recently, most vision-based SLAM systems limited themselves by separating the motion estimation and map estimation problems. Typically, at each step,

the robot's location would be estimated via Bayesian inference or some other estimation technique, then that position would be assumed for the map update. While this approach leads to fast algorithms, not considering alternative robot poses when estimating landmark positions is suboptimal. Other researchers in the robotics community took a formal probabilistic approach and explored the possibility of representing, at each point in time, the full joint posterior distribution over robot trajectories and landmark positions. Smith, Self, and Cheeseman [18] introduced the "stochastic map," which represents not only the positions of landmarks in the world with their associated uncertainties, but also the uncertainty of the robot's position, the covariance between each pair of landmarks, and the covariance between the robot's position and each landmark. This seminal theoretical work inspired many successful SLAM systems, e.g. [5, 19, 20]. Lemaire et al. [8] demonstrate successful loop closure for a ground rover using an EKF with both stereo and monocular sensors. In a particularly impressive demonstration of the power of the stochastic map approach, Davison and colleagues [2–4] have solved the vision-based SLAM problem with point landmarks extracted from a single camera without odometry. Their system runs in real time at 30 Hz.

However, moving to larger-scale environments (especially outdoor environments) is difficult for mobile robots with vision sensors because it generally requires the capability of storing a huge number of landmarks, and at the same time, maintaining them in a manner consistent with erroneous robot trajectories and the global geometric structure of environments. While the EKF-based stochastic map very accurately represents all of the available information about landmark and robot positions (within the limits of the Gaussian approximation), the method unfortunately cannot scale to the thousands of landmarks needed for large-scale environments, due to the quadratic size of the full covariance matrix. Lemaire et al. [8] point out another practical difficulty, that EKF-based SLAM is overly sensitive to modeling and linearization errors.

One approach to making the large-scale mapping and localization problem tractable is to separate the two steps entirely. Royer et al. [21] learn a landmark map from a prerecorded video acquired in a large-scale outdoor environment using bundle adjustment (a batch structure from motion technique) and demonstrate real-time localization and path following based on this a priori map.

Murphy [22], however, provides an alternative solution within the framework of SLAM. He recognized that map elements are conditionally independent given the robot's trajectory through time. He used this insight in the design of the Rao-Blackwellised particle filter (RBPF), in which the joint posterior over robot trajectories and maps is represented by a set of samples or particles, each particle containing one possible robot trajectory and the corresponding stochastic map. The fact that the robot's trajectory is fixed for a given particle has an important consequence: all of the covariances between different map elements in the stochastic map become 0. For a landmark map, this means the covariance matrix for each individual landmark is sufficient to represent all of the available knowledge of the environment.

Murphy only demonstrated the RBPF on a toy problem, but more recent work has applied the technique to the real world with immense success. Montemerlo, Thrun, and colleagues [1] use the RBPF and 2D point landmarks measured by a laser scanner to construct large-scale 2D maps. In their system, each particle represents a possible

robot trajectory, set of data associations, and landmark map. The maps are stored in a tree structure that allows sharing subtrees between particles, allowing a real-time implementation that scales to thousands of landmarks. Eliazar and Parr [23] also use the RBPF and a laser scanner for SLAM, but build a 2D occupancy grid rather than a landmark database. Their algorithm also requires a sophisticated data structure that allows sharing maps between particles.

One of the most promising approaches to large-scale metric vision-based SLAM, then, is to use the RBPF as the underlying estimation algorithm. There has been some recent work applying stereo vision head and the RBPF to the vision-based SLAM problem [7, 24]. These systems use Thrun et al.'s FastSLAM algorithm [1] for the RBPF to create sparse landmark maps organized by k-D trees for efficient search and modification.

There have been successful demonstrations of SLAM in indoor environments using point landmarks constructed using SIFT [16]. In Se et al. [6], a mobile robot with trinocular stereo vision maps a 10 m $\times$ 10 m laboratory employing the Kalman Filter to refine landmark positions and robot poses. In Sim et al. [25], the RBPF is used for a robot with stereo vision to map a large-scale indoor environment, demonstrating the RBPF's scalability to large numbers of visual landmarks and robust loop-closing capabilities. However, SIFT is computationally expensive. Combined with the computational complexity of maintaining many robot path estimates in the RBPF, systems based on SIFT and the RBPF are going to be expensive or slow for several years to come.

Towards the goal of solving large-scale online SLAM problems, we are going to propose a novel approach: the combination of the performance-proven RBPF and a lightweight sensor model that does not rely on computationally expensive SIFT features. Our proposed vision-based sensor model using trinocular stereo and Shi-Tomasi image point features is one of the most promising solutions that could meet practical online SLAM requirements.