Guest Editorial
Special Issue on Visual SLAM

I. Introduction

When a world observing sensor moves through a scene, analysis of the data it captures permits estimation of its egomotion, even if nothing is known in advance about the contents of the scene around it. The key is to make multiple observations of repeatably measurable aspects of the environment as the sensor moves and use assumptions, normally that the scene is mostly static, to infer both the sensor's motion and knowledge of the geometrical layout of those measured scene entities. In recent years, the field of simultaneous localization mapping (SLAM) has provided understanding and practical algorithms for tackling such problems, with a view, in particular, to applications in autonomous mobile robotics.

SLAM has been one of the most active research fields in robotics for over ten years and a major topic at leading international conferences. Excellent results have been reported by many researchers, but until recently, the most successful systems required the use of laser range-finder sensors and predominantly aimed to build 2-D maps of planar environments.

It was considered more challenging to attempt SLAM with standard cameras as the main sensory input, since the essential geometry of the world does not "pop out" of images the same way as it does from laser data. The challenges of solving problems such as robust feature detection, data association, and computationally efficient large-scale state estimation are much harder than using a sensor such as a Sick laser scanner. However, cameras offer the advantages of full 3-D, low cost, and compactness as well as the intuitive "human-like" appeal of purely visual sensing. Perhaps, more importantly, they provide seemingly limitless potential for the extraction of detail, both geometric and photometric, and for semantic and other higher level scene interpretation, as continually demonstrated by the flourishing computer vision research community.

Therefore, inevitably, there has been a surge in research in a vision-based SLAM over the past five years. Much of the current state-of-the-art research in robot localization and mapping relies primarily, or exclusively, on the use of cameras for exteroceptive sensing. Recent robotic systems that use single cameras, stereo rigs, or omnidirectional cameras, often in combination with odometry or inertial sensing, have demonstrated reliable and accurate localization and mapping and shown that cameras are highly viable sensors for SLAM.

Approaches based on well-known SLAM filtering algorithms from the robotics literature such as extended Kalman filtering, particle filtering, and submapping have proven to be effective in sequentially building consistent maps of certain domains, often in real time and on increasingly large scales. Coming from another direction, and inspired by computer vision research on multiple view geometry, have been "visual odometry" approaches that obtain highly accurate local motion estimates but do not necessarily aim to build globally consistent maps.

As the mathematical character of the generic "SLAM problem" as large-scale Bayesian inference over a graph of constraints becomes widely understood, these two different approaches have now increasingly begun to meet, and state-of-the-art algorithms use ideas and estimation techniques from both. In particular, as in SLAM research based on other sensors, "pose-based" or "view-based" methods in which a map of historical sensor locations, rather than explicit scene feature positions that are built, have proven to be very successful.

Another important realization has been that visual sensing can offer both accurate local motion estimation via visual odometry and robust detection of "loop-closure" events when previously observed parts of an environment are revisited. This latter capability is provided by visual recognition methods relying on groups of salient features or whole image statistics.

Vision-based SLAM systems, including the ones presented in this special issue, are now showing the potential to build detailed maps on scales as large as a building or a city block. Vision-based SLAM has been implemented on a variety of robot platforms operating in a diverse range of environments, including on the ground, in the air, or underwater. This study has the potential to guide autonomous robots on their explorations and operations in large and complex environments or more limited environments such as homes where low-cost service robots might be deployed. Vision-based SLAM, however, has many other potential applications beyond traditional mobile robotics scenarios, due to the capability it can give a camera to serve as a general-purpose 3-D position sensor.

It has been said that anytime one moves a sensor along an uncertain trajectory, in an uncertain environment, one encounters a SLAM problem; the application scenarios involving cameras moving along uncertain trajectories with incomplete prior information about the environment are limitless. Applications can be envisaged in domains such as wearable computing, user interfaces, augmented reality, or robotic surgery.

The focus of this Special Issue of the IEEE Transactions on Robotics (TRO) is to publish outstanding recent results in the rapidly progressing subject of visual SLAM. This issue contains papers that describe vision-only systems as well as systems that accommodate information provided by other sensors such as inertial units and odometry. The contributions presented include improvements to "classical SLAM algorithms" such as feature-based and view-based SLAM based on extended Kalman and extended information filters, as well as recent developments

Digital Object Identifier 10.1109/TRO.2008.2004620

1552-3098/$25.00 © 2008 IEEE

such as visual odometry, appearance-based methods for the detection of loop closure, and biologically inspired SLAM algorithms. Finally, there are papers that put many of these methods together into convincing systems.

II. Guide to the Special Issue

The first three papers of the Special Issue present ideas to extend the capabilities of feature-based SLAM systems based on the extended Kalman filter (EKF). The paper by Civera *et al.* presents the inverse depth parametrization, i.e., a novel representation for point features within monocular SLAM that allows undelayed initialization of features within the standard EKF framework. They demonstrate that there is a high degree of linearity in this parametrization and that, therefore, a Gaussian distribution represents uncertainty very accurately. This parametrization can cope with features which are so far from the camera that they present little parallax during motion, leaving open the possibility that these points "come in" from infinity if the camera makes larger motions.

The paper by Paz *et al.* describes a system that can carry out SLAM in large indoor and outdoor environments using a stereo pair moving in 3-D as the only sensor. Features sufficiently close to the stereo pair are incorporated as normal 3-D points, and the inverse parametrization proposed in the paper by Civera *et al.* is used for distant features that show little disparity or features only seen by one camera. Real-time execution is attained by building sequences of conditionally independent local maps that can share valuable information about the camera state and features tracked during the transition. These local maps are joined using the divide-and-conquer algorithm adapted for conditionally independent local maps to provide a complete map whenever required.

The paper by Solà *et al.* explores the idea of considering a multicamera system as several independent monocular bearing-only sensors instead of a monolithic supersensor. This idea is close to that of the paper by Paz *et al.* (including the use of the inverse parametrization). It allows the system to map nearby objects while still making use of the bearing information provided by the observation of remote ones. The greater degree of flexibility in the consideration of the cameras in this system results in additional advantages: possible desynchronization of the sensors firing, allowing the use of several unequal cameras, and self-calibration.

The next two papers concentrate on going beyond the usual point-based feature maps of SLAM and aim to determine more detailed representations of a scene's surface geometry. The paper by Silveira *et al.* presents a structure and motion algorithm for visual SLAM that considers illumination parameters (and additional constraints, such as positive depth) to simultaneously estimate image alignment, camera poses (through optical flow), and the structure of planar surfaces. Iterative nonlinear optimization is carried out to minimize the norm of the vector of discrepancies in intensity. Details are given on how the system can be initialized and how problems of local minima can be avoided. Both simulated and real urban sequences are used to illustrate the result of this method.

The paper by Gee *et al.* presents a framework for incorporating higher level structure, such as planes and lines, into a real-time EKF-based SLAM system. Such structures are "discovered" by first mapping low-level features (points and edgelets) and then searching for sets that agree with the hypothesis of a plane of a line. The multiple low-level features can then be replaced in the SLAM map by a single high level feature, transferring the uncertainty in their locations rigorously into uncertainty in the high-level feature representation. The advantages here are in efficiency and completeness of scene representation and increased semantic understanding. The methods are illustrated with simulations and real experiments, demonstrating, in particular, how the large planar surfaces of a room viewed with a hand-held camera can be automatically determined during real-time operation.

The next two papers are concerned with view-based SLAM approaches, where vision is used as a correcting influence on trajectories obtained from other sensors. The paper by Andreasson *et al.* describes a visual view-based SLAM system that relies on odometry measurements and panoramic images to obtain a topologically correct and geometrically accurate representation of the sensor pose graph. The system computes a similarity measure between pairs of images to detect those that are likely to correspond to nearby poses. Relative pose estimations are obtained in such cases by computing the relative orientation using histograms of the matched features. These, along with the odometry measurements, are fed to a multilevel relaxation algorithm to optimize the pose graph. Indoor and outdoor experiments of 618 m and 1.4 km are discussed, comparing the results of the proposed system with differential GPS data as well as by providing accurate grid maps computed with laser measurements.

The paper by Mahon *et al.* contains the main theoretical contribution of the use of Cholesky modifications to enable efficient incremental state and covariance recovery during pose-based SLAM using information filtering. This development is demonstrated effectively in the context of underwater SLAM, combining local motion estimation based on a package of inertial and other sensors and vision-based loop closure detection and pose correction based on feature matching in images obtained from a stereo rig observing the seabed. The paper's experimental results demonstrate accurate mapping of a large area of seabed crisscrossed by the underwater robot and, in particular, the large improvement obtained by incorporating the information gained from visual loop closures.

Visual odometry and visual loop closure detection are two of the key recent developments in visual SLAM, and the following two papers present contributions in these two areas. The paper by Scaramuzza and Siegwart describes a real-time visual odometry algorithm for computing the egomotion of a vehicle relative to the road. The only input used are images provided by a single omnidirectional camera mounted on the roof of the vehicle. Two different trackers are combined: a homography-based tracker that detects and matches robust scale invariant features that most likely belong to the ground plane and an appearance based approach that gives high-resolution estimates of the rotation of the vehicle. Image mosaicing is used to obtain a textured 2-D reconstruction of the estimated path. Experiments with a 400-m camera trajectory estimated purely from omnidirectional images show the potential of omnidirectional cameras in visual odometry applications.

The paper by Angeli *et al.* concerns an approach to appearance-based visual SLAM, where a topologically correct scene map is built fully automatically with very few constraints on the type of camera or motion. A "bag-of-words" image classification approach, using local shape and color information, is used in combination with a temporal Bayesian filter to give robust probabilistic estimates of the occurrence of loop-closure events when scene locations are revisited. In comparison with other recent work on visual loop-closure detection, the method presented in this paper requires no prior training step and learns the visual dictionary of important image characteristics online. Results are presented on both indoor and outdoor sequences taken with a hand-held camera.

The next two papers present rather different approaches to visual SLAM with some degree of biological inspiration. In the paper by Milford and Wyeth, the authors' biologically inspired *RatSLAM* algorithm is used in combination with a surprisingly simple computer vision front end that supplies both local motion estimates and evidence of loop closures to produce a capable pose-based SLAM system. The paper presents striking results showing real-time mapping of the road network of a whole city suburb based purely on the images acquired from a single forward-facing camera, mounted on a car driven at natural speeds on a loopy trajectory. While not aiming at the metric local accuracy of the methods in some of the other papers, the paper's original viewpoint and the method's large-scale robustness offer an intriguing alternative.

The paper by Frintrop and Jensfelt revisits the neglected but compelling idea of actively controlling the pointing direction of a robot-mounted camera with the aim of improving the visual SLAM performance. An important part of the method is the use of a biologically motivated visual attention algorithm to detect salient and unique landmarks for SLAM that are sparse but of particularly high quality. Active camera control is then used to explore the environment to find landmarks, track them continuously as the robot moves, and force their redetection after periods of neglect for loop closure. The paper proves the ideas with a fully implemented real-time system demonstrating robust indoor localization and mapping.

Finally, there are three papers that put modern techniques together by using vision for local motion estimation, feature mapping, and view-based trajectory optimization. The paper by Konolige and Agrawal describes *FrameSLAM*, i.e., an algorithm that combines the best of visual odometry and view-based SLAM into a uniform approach based on nonlinear optimization. Visual odometry is used to estimate precise frame-to-frame motion of a binocular stereo rig with optional inertial measurement unit (IMU) assistance if available. The dense graph of poses produced is reduced to a skeleton of camera poses joined by nonlinear constraints, which is sparse enough to be updated in real time in response to the detection of loop closure events between nearby nonconsecutive frames, enforcing global consistency. Indoor and outdoor experiments demonstrate the real-time viability of *FrameSLAM* for highly accurate mapping of 3-D trajectories of up to 10 km.

The paper by Zhou *et al.* presents a system that uses a conventional camera in combination with a range camera to produce 3-D dense maps of unstructured environments traversed by the sensor set moving freely in 3-D. Both features and camera poses are estimated using an extended information filter. Instead of updating the map every time new sensor information is available, updates are carried out only when enough information will be gained by the system from maximally informative observations. This way, this system can maintain a compact representation of the environment, including feature locations, in contrast with view-based SLAM systems.

Finally, the short paper by Steder *et al.* presents a system that allows aerial vehicles, such as small blimps or helicopters, to acquire visual maps of large environments using an attitude sensor and low-quality cameras (monocular or stereo) pointing downward. Correspondences between features are obtained using a variant of the progressive sample consensus (PROSAC) algorithm that are used to compute spatial constraints between camera poses, which can correspond to visual odometry or place revisiting. The final map is computed using a graph-based maximum-likelihood mapping algorithm. The paper discusses several large indoor and outdoor experiments. If a stereo setup is available, the system is able to learn visual elevation maps of the ground.

Acknowledgment

The Guest Editors wish to express their enormous gratitude to L. Parker, the IEEE Transactions on Robotics (TRO) Editor supervising the review process of this Special Issue, for her professional and patient work throughout the whole process of editing this Special Issue. They also thank Editor-in-Chief A. De Luca as well as Editorial Assistant D. Lo Tenero for their constant support. Finally, they are grateful to the many colleagues who agreed to provide reviews for the 50 papers submitted to the Special Issue and sent in thorough and helpful reviews.

José Neira, *Guest Editor*
Robotics Group
Universidad de Zaragoza
50009 Zaragoza, Spain

Andrew J. Davison, *Guest Editor*
Visual Information Processing (VIP) Research Group
Department of Computing
Imperial College London
London SW7 2DD, U.K.

John J. Leonard, *Guest Editor*
Computer Science and Artificial Intelligence Laboratory
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 4307, USA