Robotics and Autonomous Systems 61 (2013) 1144–1158

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Real-time 6-DOF multi-session visual SLAM over large-scale environments

J. McDonald [a,*], M. Kaess [c], C. Cadena [b], J. Neira [b], J.J. Leonard [c]

[a] Department of Computer Science, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

[b] Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Zaragoza 50018, Spain

[c] Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Available online 30 August 2012

Keywords:
Bundle adjustment
Place recognition
Stereo
Anchor nodes
iSAM

ABSTRACT

This paper describes a system for performing real-time multi-session visual mapping in large-scale environments. Multi-session mapping considers the problem of combining the results of multiple simultaneous localisation and mapping (SLAM) missions performed repeatedly over time in the same environment. The goal is to robustly combine multiple maps in a common metrical coordinate system, with consistent estimates of uncertainty. Our work employs incremental smoothing and mapping (iSAM) as the underlying SLAM state estimator and uses an improved appearance-based method for detecting loop closures within single mapping sessions and across multiple sessions. To stitch together pose graph maps from multiple visual mapping sessions, we employ spatial separator variables, called anchor nodes, to link together multiple relative pose graphs.

The system architecture consists of a separate front-end for computing visual odometry and windowed bundle adjustment on individual sessions, in conjunction with a back-end for performing the place recognition and multi-session mapping. We provide experimental results for real-time multi-session visual mapping on wheeled and handheld datasets in the MIT Stata Center. These results demonstrate key capabilities that will serve as a foundation for future work in large-scale persistent visual mapping.

1. Introduction

Despite substantial recent progress in visual simultaneous localisation and mapping (SLAM) [1], many issues remain to be solved before a robust, general visual mapping and navigation solution can be widely deployed. A key issue in our view is that of *persistence*—the capability for a robot to operate robustly for long periods of time. As a robot makes repeated transits through previously visited areas, it cannot simply treat each mission as a completely new experiment, not making use of previously built maps. However, nor can the robot treat its complete lifetime experience as "one big mission", with all data considered as a single pose graph and processed in a single batch optimisation. We seek to develop a framework that achieves a balance between these two extremes, enabling the robot to leverage off the results of previous missions, while still adding in new areas as they are uncovered and improving its map over time.

The overall problem of persistent visual SLAM involves several difficult challenges not encountered in the basic SLAM problem. One issue is dealing with dynamic environments, requiring the robot to correct for long-term changes, such as furniture and other objects being moved, in its internal representation; this issue is not addressed in this paper. Another critical issue, which is addressed in this paper, is how to pose the state estimation problem for combining the results of multiple mapping missions efficiently and robustly.

Cummins defines the multi-session mapping problem as "the task of aligning two partial maps of the environment collected by the robot during different periods of operation [2]". We consider multi-session mapping in the broader context of life-long, persistent autonomous navigation, in which we would anticipate tens or hundreds of repeated missions in the same environment over time. As noted by Cummins, the "kidnapped robot problem" is closely related to multi-session mapping. In the kidnapped robot problem, the goal is to estimate the robot's position with respect to a prior map given no *a priori* information about the robot's position.

Also closely related to the multi-session mapping problem is the multi-robot mapping problem. In fact, multi-session mapping can be considered as a more restricted case of multi-robot mapping in which there are no direct encounters between robots (only indirect encounters, via observations made of the same environmental structure). Kim et al. presented an extension to iSAM to facilitate online multi-robot mapping based on multiple

* Corresponding author. Tel.: +353 1 708 4589; fax: +353 1 708 3848.
E-mail addresses: johnmcd@cs.nuim.ie (J. McDonald), kaess@mit.edu (M. Kaess), cesarcadena.lerma@gmail.com (C. Cadena), jneira@unizar.es (J. Neira), jleonard@mit.edu (J.J. Leonard).

0921-8890/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2012.08.008

pose graphs [3]. This work utilised "anchor nodes", equivalent to the "base nodes" introduced by Ni and Dellaert for decomposition of large pose graph SLAM problems into submaps of efficient batch optimisation [4], in an approach called Tectonic Smoothing and Mapping (T-SAM). Our work builds on the approach of Kim et al. [3] to perform multi-session visual mapping by incorporating a stereo odometry frontend in conjunction with a place-recognition system for identifying inter- and intra-session loop closures.

This paper makes a number of extensions to the work presented in [5]. In particular, in [5] we provided preliminary results of a multi-session visual SLAM system based on the architecture shown in Fig. 1. Here we expand the discussion and give details of changes that we have made to increase the system's overall robustness and to permit real-time processing over large scale environments. Results are provided demonstrating robust multi-session visual SLAM processing on datasets including (i) up to four separate sessions (totalling $>$ 45 min of video at 20 fps), (ii) wheeled and handheld sensors, (iii) indoor and outdoor sequences, and, (iv) sequences involving full 6-DOF-motion (i.e. ascending and descending stairs). We also present a comprehensive quantitative evaluation of the system using the above datasets.

The remainder of the paper is organised as follows. In the next section we review related work in the area, focussing on multi-session and multi-robot approaches to localisation and mapping. Section 3 provides an overview of the system architecture, with details of front-end processing including the stereo odometry and single-session visual SLAM given in Sections 3.1 and 3.2, respectively. In Section 3.3 we describe the approach used to integrating the quaternion-based representation for rotations and homongeneous point representation into the iSAM optimisation process. The back-end modules including visual place recognition and multi-session visual SLAM are explained in Sections 3.4 and 3.5, respectively. Experimental results and a comprehensive quantitative analysis of the system's performance is provided in Section 4. Finally, Section 5 provides concluding remarks and potential future directions for the research.

2. Related work

Several vision researchers have demonstrated the operation of visual mapping systems that achieve persistent operation in a limited environment. Examples of recent real-time visual SLAM systems that can operate persistently in a small-scale environment include Klein and Murray [6], Eade and Drummond [7], and Davison et al. [8,9]. Klein and Murray's system is highly representative of this work, and is targeted at the task of facilitating augmented reality applications in small-scale workspaces (such as a desktop). In this approach, the processes of tracking and mapping are performed in two parallel threads. Mapping is performed using bundle adjustment. Robust performance was achieved in an environment as large as a single office. While impressive, these systems are not designed for multi-session missions or for mapping of large-scale spaces (e.g., the interior of a building).

One exception to this has been the extension to PTAM developed by Castle et al. [10] to permit several cameras to work in multiple maps, both separately or simultaneously. Here the approach to dealing with large-scale environments is to permit the user to decide what regions to map. Each map is bounded in size and operates independently of other maps in the system. To switch between maps the current frame is matched against a set of subsampled keyframes from all existing maps. The authors provide impressive results of the technique's operation in a building scale environment. A key difference between this approach and our work is that the system does not estimate the transformation between the submaps and therefore does not provide a global estimate of the environment.

There have also been a number of approaches reported for large-scale visual mapping. Although a comprehensive survey is beyond the scope of this paper, we do draw attention to the more relevant stereo-based approaches. Perhaps the earliest of these was the work of Nistér et al. [11] on stereo odometry. In the robotics literature, large-scale multi-session mapping has been the focus of recent work of Konolige et al. in developing view-based mapping systems [12,13]. Our research is closely related to this work, but has several differences. A crucial new aspect of our work in relation to [13] is the method we use for joining the pose graphs from different mapping sessions. In the view-based mapping approach, Konolige et al. employ the Toro incremental optimisation algorithm to allow for real-time performance. Due to the fact that Toro requires that all poses are connected in a single graph, at the beginning of each new session the new pose-graph is immediately connected to the last pose from the previous session through what they refer to as a "weak link". The weak links are added with a very high covariance and subsequently deleted after place recognition is used to join the pose graphs [13]. In our approach, which extends [3] to full 6-DOF, we use anchor nodes as an alternative to weak links. Here each session is represented initially as a disjoint pose-graph with each pose stored relative to that pose-graph's anchor node. When the place recognitions system identifies an encounter between two separate sessions, the encounter induces a constraint between the two associated poses and the anchor nodes of the associated pose-graphs. Since the pose-graphs are each represented relative to the anchors nodes, their use provides a more efficient and consistent way to stitch together the multiple pose graphs resulting from multiple mapping sessions. Further details on this aspect of our system are provided in Section 3.5. In addition, our system has been applied to hybrid indoor/outdoor scenes, with hand-carried (full 6-DOF) camera motion.

3. System overview

In this section we describe the architecture and components of a complete multi-session stereo visual SLAM system. This includes a stereo visual SLAM frontend, a place recognition system for detecting single and multi-session loop closures, and a multi-session state-estimation system. A schematic of the system architecture is shown in Fig. 1. The system uses a sub-mapping approach in conjunction with a global multi-session pose graph representation. Optimisation is performed by applying incremental and batch SAM to the pose graph and the constituent submaps, respectively. Each submap is constructed over consecutive sets of frames, where both the motion of the sensor and a feature-based map of the scene is estimated. Once the current submap reaches a user-defined maximum number of poses, 15 in our system, the global pose graph is augmented with the resultant poses.

In parallel to the above, as each frame is processed, the visual SLAM frontend communicates with a global place recognition system for intra- and inter-session loop closure detection. When a loop closure is detected, pose estimation is performed on the matched frames, with the resultant pose and frame-id's passed to the multi-session pose graph optimisation module.

3.1. Stereo odometry

Within each submap the inter-frame motion and associated scene structure is estimated via a stereo odometry frontend. The most immediate benefit of the use of stereo vision is that it avoids issues associated with monocular systems, including the inability to estimate scale and indirect depth estimation. The