Robotics and Autonomous Systems 93 (2017) 13–26

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

Sparse optimization for robust and efficient loop closing

Yasir Latif [a,*], Guoquan Huang [b], John Leonard [c], José Neira [d]

[a] ARC Center for Robotic Vision, University of Adelaide, Adelaide, SA 5005, Australia
[b] Department of Mechanical Engineering, University of Delaware, Newark, DE 19716, USA
[c] Computer Science and Artificial Intelligence Laboratory Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[d] Instituto de Investigación en Ingeniería de Aragón (I3A) Universidad de Zaragoza, Zaragoza, Spain

HIGHLIGHTS

- Loop closure detection methods based on sparse optimization.
- Useful in cases where there is high degree of perceptual aliasing.
- General framework that allows flexible representation.

ARTICLE INFO

Article history:
Available online 11 April 2017

Keywords:
SLAM
Place recognition
Relocalization
Sparse optimization

ABSTRACT

It is essential for a robot to be able to detect revisits or *loop closures* for long-term visual navigation. A key insight explored in this work is that the loop-closing event inherently occurs sparsely, i.e., the image currently being taken matches with only a small subset (if any) of previous images. Based on this observation, we formulate the problem of loop-closure detection as a *sparse, convex* ℓ_1-minimization problem. By leveraging fast convex optimization techniques, we are able to efficiently find loop closures, thus enabling real-time robot navigation. This novel formulation requires no offline dictionary learning, as required by most existing approaches, and thus allows *online incremental* operation. Our approach ensures a *unique* hypothesis by choosing only a single globally optimal match when making a loop-closure decision. Furthermore, the proposed formulation enjoys a *flexible* representation with *no* restriction imposed on how images should be represented, while requiring only that the representations are "close" to each other when the corresponding images are visually similar. The proposed algorithm is validated extensively using real-world datasets.

1. Introduction

With a growing demand for autonomous robots in a range of applications, such as search and rescue [1,2], and space and underwater exploration [3], it is essential for the robots to be able to navigate accurately for an extended period of time in order to accomplish the assigned tasks. To this end, the ability to detect revisits (i.e., *loop closure* or place recognition) becomes necessary, since it allows the robots to bound the errors and uncertainty in the estimates of their positions and orientations (poses). In this work, we particularly focus on loop closure during visual navigation, i.e., given a camera stream we aim to efficiently determine whether the robot has previously seen the current place or not.

Even though the problem of loop closure has been extensively studied in the visual-SLAM literature (e.g., see [4–6]), a vast majority of existing algorithms typically require the *offline* training of visual words (dictionary) from *a priori* images that are acquired previously in visually similar environments. Clearly, this is not always the case when a robot operates in an unknown, drastically different environment. In general, it is difficult to reliably find loops in (visual) appearance space. One particular challenge is the perceptual aliasing – that is, while images may be similar in appearance, they might be coming from different places. To mitigate this issue, both temporal (i.e., loops will only be considered closed if there are other loops closed nearby) and geometric constraints (i.e., if a loop has to be considered closed, a valid transformation must exist between the matched images) can be employed [6]. It is important to point out that the approach of [6] decides on the quality of a match *locally* – If the match with the highest score (in some distance measure) is away from the second highest, it is considered a valid candidate. However, the local information may lead to incorrect loop-closure decisions because both temporal and geometric conditions can easily fail in highly self-similar environments such as corridors in a hotel.

* Corresponding author.
E-mail address: yasir.latif@adelaide.edu.au (Y. Latif).

http://dx.doi.org/10.1016/j.robot.2017.03.016
0921-8890/

To address the aforementioned issues, in this paper we introduce a *general*, *online* loop-closure approach for vision-based robot navigation. In particular, by realizing that loops typically occur intermittently in a navigation scenario, we, for the first time ever, formulate loop-closure detection as a sparse ℓ_1-minimization problem that is convex. This is opposed to the current methods that cast loop closure detection as an image retrieval problem [5,6]. By leveraging the fast convex optimization techniques, we subsequently solve the problem efficiently and achieve real-time frame-rate generation of loop-closure hypotheses. Furthermore, the proposed formulation enjoys *flexible* representations and can produce loop-closure hypotheses regardless of what the extracted features represent –that is, any discriminative information, such as descriptors, Bag of Words (BoW), or even whole images, can be used for detecting loops. Lastly, we shall stress that our proposed approach declares a loop that is valid only when it is *globally unique*, which ensures that if perceptual aliasing is being caused by more than one previous image, *no* loop closure will be declared. Although this is conservative in some cases, since a false loop closing can be catastrophic while missing a loop closure generally is not, ensuring such global uniqueness is necessary and important, in particular, in highly self-similar environments.

The rest of the paper is organized as follows: After reviewing the related work, we formulate loop-closure detection as a sparse ℓ_1-minimization problem in Section 3. In Section 4 we present in detail the application of this formulation to visual navigation, which is validated via the real-world experiments in Section 5. Finally, Section 6 concludes this work as well as outlines the possible directions for future research.

2. Related work

The problem of loop-closure detection has been extensively studied in the SLAM literature and many different solutions have been proposed over the years (e.g., see [4,7] and references therein). In what follows, we briefly overview the work that closely relates to the proposed approach.

In particular, the FAB-MAP [5] is a probabilistic appearance-based approach using visual BoW for place recognition, and was shown to work robustly over trajectories up to 1000 km. Similarly, the Binary-BoW (BBoW)-based method [6] detects the FAST keypoints [8] and employs a variation of the BRIEF descriptors [9] to construct the BoW. A verification step is further enforced to geometrically check the features extracted from the matched images. It should be pointed out that both methods [5,6] are based on the similar ideas of text-retrieval [10,11]: These methods learn the BoW dictionaries beforehand, which are used later for detecting loop closures when the robots actually operates in the field. This restricts the expressive power of the dictionary in cases where it has to operate in environments drastically different from where the dictionary was constructed. In contrast, the proposed approach builds the dictionary *online* as the robot explores an unknown environment, while at the same time efficiently detecting loops (if any). Moreover, rather than solely relying on the descriptors-based BoW, our method is flexible and can utilize *all* pixel information to discriminate places even in presence of dynamic objects (encoded as sparse errors), any descriptor that can represent similar places, or any combination of such descriptors.

Some recent work has focused on loop closure under extreme changes in the environment such as different weather and/or lighting conditions at different times of the day. For example, Milford and Wyeth [12] proposed the SeqSLAM that is able to localize with drastic lighting and weather changes by matching sequences of images with each other as opposed to single images. Churchill and Newman [13] introduced the experience-based maps that learn the different appearances of the same place as it gradually changes in order to perform long-term localization. Building upon [13], Paul and Newman [14] also discovered new images to attain better localization. In addition, Lee et al. [15, 16] have explored geometric features such as lines for the task of loop closure detection in both indoor and outdoor scenarios. Note that if the information invariant to such changes can be extracted as in [5,6,12–16], the proposed formulation can also be used to obtain loop-closure hypotheses. Essentially, in this work we focus on finding loop closures given some discriminative descriptions such as descriptors and whole images, assuming *no* specific type of image representations.

More recently, with the rediscovery of efficient machine learning techniques, Convolutional Neural Networks (CNNs) [17,18] have been exploited to address loop closure detection [19,20]. These networks are multi-layered architectures that are typically trained on millions of images for tasks such as object detection and scene classification. The internal representations at each layer are learned from the data itself and therefore can be used as features to replace hand-crafted features. Based on this approach, Sünderhauf et al. [19] extract features from different layers in the network and identify the layers that are useful for view-point and illumination invariant place recognition. Moreover, in [20] landmarks are treated as objects by finding object proposals in the images and features are extracted for them using deep networks. These features then allow for view-point invariant place categorization by matching different objects from varied viewpoints. In these CNN-based place categorization techniques, the networks are used as feature extractors followed by some form of matching. In this paper, we show that these deep features can also be utilized in the proposed framework of loop-closure detection.

It should be noted that in our previous conference publication [21], we have preliminarily shown that the proposed loop-closing framework is general and can employ most hand-crafted features. Recently, Shakeri and Zhang [22] extended this sparse-optimization based framework to an incremental formulation allowing for the use of the previous solution of the sparse optimization to jump start the next one, while Zhang et al. [23] further extended it to a multi-step delayed detection of loops (instead of single-step detection as in our prior work [21]) in order to exploit the structured sparsity of the problem. In this paper, we present more detailed analysis and thorough performance evaluations, including new experiments using deep features and validations in challenging multiple-revisit scenarios, as well as new comparisons against the well-known nearest neighbor (NN) search.

3. Sparse optimization for loop closure

In this section, we formulate loop-closure detection as a sparse optimization problem based on a sparse and redundant representation. Such representations have been widely used in computer vision for problems such as denoising [24], deblurring [25], and face recognition [26]. Similarly, Casafranca et al. [27, 28] formulated the back-end of graph SLAM as an ℓ_1-minimization problem. However, *no* prior work has yet investigated this powerful technique for loop closure detection in robot navigation. The key idea of this approach is to represent the problem *redundantly*, from which a *sparse* solution is sought for a given observation.

Suppose that we have the current image represented by a vector $\boldsymbol{b} \in \mathcal{R}^n$, which can be either the vectorized full raw image or descriptors extracted from the image. Assume that we also have a dictionary denoted by $\mathbf{B} = [\boldsymbol{b}_1 \cdots \boldsymbol{b}_m] \in \mathcal{R}^{n\times m}$, which consists of m basis vectors of the same type as $\boldsymbol{b}$. Thus, solving the linear system $\mathbf{B}\boldsymbol{x} = \boldsymbol{b}$ yields the representation of $\boldsymbol{b}$ in the base $\mathbf{B}$ in the form of the vector $\boldsymbol{x} \in \mathcal{R}^m$. Elements of $\boldsymbol{x}$ indicate which basis vectors, $\boldsymbol{b}_i$, best explain $\boldsymbol{b}$ and how much the corresponding contributions are. A zero contribution ($x_i = 0$) simply implies that the corresponding basis vector $\boldsymbol{b}_i$ is irrelevant to $\boldsymbol{b}$. One trivial