Title goes here

Anonymous CVPR submission

Paper ID Occular Tension

Abstract

The overall problem of Visual Simultaneous Localization and Mapping (VSLAM) and its different aspects have been researched in the computer vision community extensively. This paper critiques three papers on VSLAM and provides an analysis of the innovative contributions of each. Robust Large Scale Monocular Simultaneous Localization and Mapping proposes a framework for implementing VSLAM using monocular cameras. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects proposes a method for VLSAM using prior knowledge of the environment and object recognition. Good Features to Track for Visual SLAM (GF-SLAM) focuses on the aspect of finding good visual features to use in VSLAM based on their temporal observability. This paper also provides an analysis of the implementation of the GF-SLAM paper and the recreation of the paper's key experiment.

1. Introduction

The problem of estimating a 3D model the environment as sensed by a camera as well as estimating the camera's trajectory is Visual Simultaneous Localization and Mapping (VSLAM). The computer vision community has researched and provided innovative solutions that focus on different aspects of the overall problem. In this paper, we analyze the following three CVPR conference papers on VSLAM, discussing their innovative solution to a particular aspect of VLSAM or the problem as a whole, implementation, and subsequent conclusions: 1) Robust Large Scale Monocular Visual SLAM [1], 2) SLAM++: Simultaneous Localisation and Mapping at the Level of Objects [2], and 3) Good Features to Track for Visual SLAM [3]. We also provide an in-depth analysis of our implementation of Zhang *et al*. [3] and the conclusions drawn from the reproduction of the paper's key experiment.

2. Robust Large Scale Monocular Visual SLAM

2.1. Problem Statement

The paper focuses on the problem of using calibrated monocular cameras to perform VSLAM while making the algorithm robust, accurate, and scalable. Monocular VSLAM comes with the challenge of not being able to observe the scale of the scene of the environment. In order to overcome this, loop closures (which occur when the camera returns to a previously observed location) need to be detected. This is an issue in large environments where many scenes look alike, and results in an erroneous 3D model if loop closures are not detected properly. Thus, the paper focuses not only on the general problem of monocular VSLAM but also tackles a key subproblem of dealing with loop closure.

2.2. Innovative Contribution

To solve the problem of monocular VSLAM, the authors propose a framework consisting of three parts: 1) a Structure from Motion (SfM) algorithm based on the *Known Rotation Problem* [?] is used to estimate submaps which are parts of the camera trajectory and the unknown environment [?], 2) a loopy belief propagation algorithm is used to efficiently aligns many submaps based on a graph of relative 3D similarities to produce a global map that is consistent up to a scale factor, and 3) an outlier removal algorithm that detects and removes outliers in the relative 3D similarity graph is used to reject wrong loop closures.

2.3. Proposed Method

The paper proposes a four-part framework to implement the innovations that solve monocular VSLAM: keyframe selection, submap reconstruction, pairwise similarity estimation, and large scale relative similarity averaging.

Keyframe selection: For each frame in the captured video, Harris Points of Interest (PoI) are detected and tracked using a Lucas-Kanade tracker. When the Euclidean distance between the PoI of the current frame and previously selected keyframe is greater than a specified threshold, the frame is selected as a keypoint used as input to VS-

LAM.

Submap reconstruction: Consecutive keyframes are clustered, and using the *Known Rotation Problem*, a SfM algorithm is applied to each one by first extracting the SURF PoI [?] from all member keyframes. Loops are closed inside of each submap by matching these PoI between pairs of keyframes. The epipolar geometry is then calculated using the 5-point algorithm with RANSAC and bundle adjustment [?] between consecutive pairs of images using the SURF matches and tracked Harris PoI. The local 3D orientations are then extracted are used to estimate the global 3D orientation. With this, known tracks of PoI are built and a linear program is used to solve the *Known Rotation Problem* to estimate the camera pose at each keyframe and the associated 3D point to reconstruct the submap [?].

Pairwise Similarity Estimation: Loop closures in the reconstructed submaps are detected by first applying a bag of words approach on the SURF descriptors of the 3D points of all submaps to give each submap a unique descriptor. After that, the relative 3D similarities between each keyframe and its 10 nearest nieghboors is estimated by matching SURF descriptors with the 3D points of each submap using a k-d tree, and then using the 3-points algorithm with RANSAC and nonlinear refinement on those matches.

Large Scale Similarity Averaging: To align the submaps by estimating thier global 3D similarity to the global reference frame, a cost function on relative similarities is minimized by transforming the problem to a graph inference problem. Outliers in the graphs (representing wrong loop closures) are rejected by the *outlier removal algorithm* in which loop closures are incrementally checked by finding the shorted loop of inliers and adding them to the overall graph of inliers if their cycle error and covariance are within specified bounds. Once the outliers are removed, the *loopy belief propagation algorithm* performs the graph inference by accumulating the measurements and variances on temporal subgraphs of the original graph as it builds up final average global similarity. This algorithm is parallelized, so it can be applied on a large scale of submaps.

2.4. Experimental Evaluation

To evaluate the proposed VSLAM framework, the authors compared its performance with that of state of the art algorithms [?] and [?] on the TUM and KITTI datasets and with four different cameras with different resolutions on indoor videos they captured. Each experiment used the same optimized parameters for the various parts of the algorithm. When evaluating the results of the algorithms with respect to ground truth, a 3D similarity obtained from the minimum distance between the estimated and actual camera trajectories was used. When compared to the [?] on the TUM RGB-D dataset, the author's approach resulted in a lower RMSE for camera trajectories than [?]. When compared to [?] on the KITTI dataset, the author's algorithm estimated camera trajectories that were closer to ground truth and thus performed better. When compared to [?] on their own videos, the proposed method outperformed [?] with respect to the ground truth motion of the camera. In addition, the paper discusses the limitations of the framework in not being able to estimate a pure rotation of the camera, the necessity for the sensed environment to be static, and the necessary for consecutive relative similarities to be outlier free. However, the framework still has reasonable performance when applied to datasets that involve some moving objects.

2.5. Subsequent Conclusions

The performance evaluation of the method shows that the authors' proposed monocular VSLAM framework does substantiate their claim. Robust, independent submap generation is achieved by the visual odometry approach based on the *Known Rotation Problem*, and these submaps can be processed and aligned to form the global map and camera trajectory estimates with loop closure through the outlier removal and loopy belief propagation algorithms. Even with the described limitations, the evaluations show the innovative framework does provide a robust, accurate, and scalable solution to loop closure and the overall problem monocular VSLAM.

3. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects

3.1. Problem Statement

The authors of this paper propose an approach to the VSLAM problem using a combination of the KinectFusion algorithm along with an efficient graph based 3D object recognition system. According to them, this approach offers several advantages of existing VSLAM systems in operation that operate at the level of low level primitives (i.e. points, lines, etc).

3.2. Innovative Contribution

The VSLAM problem has been approached from the perspective of 3D object recognition before. However these methods generally reveal huge amounts of wasted computational effort via repeated low level geometry processing of the 3D objects. To counter this, the authors propose the building of pose graph maps based on an "object-oriented" approach that directly encodes the positions of recognized 3D structures. With each new measurement, the graph is continually optimized with new measurements from the sensors and allows for efficient tracking of the camera system based on recognized landmarks. In addition to this, the algorithms make the assumption that the world has "intrinsic symmetry in the form of repetitive objects" thereby allowing for the the objects in a scene to be identified and seg-

mented as salient repeated elements. The algorithm leverages this repetetiveness along with the efficient use of GPU architectures to provide a real time processing system.

3.3. Proposed Method

Creating an Object Database: The authors first create a database of repeatedly occurring objects via known KinectFusion algorithms. These are objects that are subsequently recognized and used in their VSLAM process.

SLAM Map Representation: The authors represent the world via a graph where each node stores the 6DOF pose of discovered objects relative to a fixed world frame as well as an annotation of the type of the object from the earlier created database.

Real-Time Object Recognition: This portion of the method recognizes objects in the world based on standard mesh recognition algorithms. The implementation is parallelized on GPUs to allow the real-time detection of multiple instances of multiple objects. These correspondences are obtained via the use of Point-Pair Features (PPFs) which are four dimensional descriptors.

Camera Tracking and Object Pose Estimation: The iterative closest point (ICP) algorithm is used to to track the pose of the camera model based on the earlier computed object based locations. A Huber penalty function is used to in this optimization process. Criteria is developed to ensure successful convergence of the tracking error.

Graph Optimization: The poses of the static object is now viewed as a graph optimization problem which minimizes the sum over all the measurement constraints based on the known features of each object.

Relocalization: The system accounts for a loss in camera tracking by re-initilazing localization based on matching at least 3 of the objects seen in the previously tracked long-term graph.

3.4. Experimental Evaluation

The authors reference a video submitted along with this paper to CVPR as a better description of the advantages of their method.

Loop Closure: Small loop closures are detected and compensated for by the ICP algorithm. Larger loop closures are compensated for via the use of the relocalization method.

Large Scale Mapping: Scaled mapping of a large room (15mX10mX3m) was obtained along with the mapping of 34 different objects around the room. The algorithm uses no priors regarding the original placement of these objects.

Moved Object Detection: The algorithm also displays the ability to track these objects while they, themselves are in motion.

System Statistics: The algorithm displays the amount of storage used as compared to the more traditional KinectFusion algorithm. The given mapped rooms is stored in about 20MB of space as compared to the 1.4GB used by KinectFusion. The resultant compression ratio is 1/70 which is a dramatic improvement.

3.5. Subsequent Conclusions

The paper makes several bold claims with regard to its own contributions to the literature. The graph based optimization method does indeed seem novel and the system has significantly large data compression ratio when compared to the KinectFusion algorithm. UnfortunatelyThe experimentally evaluated conclusions are quite sparse when compared to the dense and well written introduction and methodology. The biggest issue pertains from the fact that no standard metric is used to compare the system's advantages and efficiently to other 3D object recognition based VSLAM approaches. Thus, though several claims are made about the paper's VSLAM advantages over other methods, there are no easy ways to determine the validity of these claims.