Title goes here

Anonymous CVPR submission

Paper ID Occular Tension

Abstract

Talk about overall problem of vision-based SLAM. Then give 1 sentence summary of each paper. Then tell that we've implemented one and show our results.

1. Introduction

Introduction goes here. Define problem of Visual Simultaneous Localization And Mapping (VSLAM) which estimates a 3D model or map of the environment in which a camera moves along a trajectory in [?].

2. Analysis of 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 VSLAM.

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.