University of Stuttgart
Institute for Photogrammetry

Master Thesis: **Performance Test of Open-Source SLAM for Indoor Navigation**

Author: Ramtin Keshtkarrezaei

Examined by: Prof. Dr.-Ing. Norbert Haala

Supervised by: Dipl.-Ing. Tom Rudolph (Aibotix)

Submission Date: October 25, 2016

1. Introduction

One of the major requirements in development of unmanned aerial systems (UASs) is being able to sense the environments with a high degree of autonomy. In this master thesis, we are focusing on a Hexacopter "Aibot X6 V2", which intends to localize itself in an indoor environment only with the means of vision. The "Aibot X6 V2" uses a monocular camera, to determine its 3D Pose (Position and Orientation) and to generate semi-dense/sparse point clouds of its environment. Because of the fact that both its path and map of the environment are unknown initially, we use a probability distribution to estimate its pose and how the environment looks like. The problem of estimation the location of the Aibot and providing the map of the environment at the same time is so called, simultaneous localization and mapping (SLAM), which is considered as a fundamental problem for truly autonomous robots. In an outdoor environment, we normally use Global Positioning System/Global Navigation Satellite System (GPS/GNSS) for navigation but because of the fact that in an indoor environment, no GPS/GNSS signal is available, therefore it inspires us to use vision, in order to sense the environment. Cameras can also be fused with different sensors like laserscanners or IMUs, but here the goal is to localize the "Aibot X6 V2" based on Vision-Only Sensing (Monocular Camera). The objective of the master thesis is to select a suitable solution to calculate the camera pose for visual indoor positioning of the Hexacopter "Aibot X6 V2" with a monocular camera, therefore three open-source SLAM approaches have been introduced, including a direct (feature-less) approach of "LSD-SLAM: Large-Scale Direct Monocular SLAM" [1], an indirect (feature-based) approach of "ORB-SLAM: A Versatile and Accurate Monocular SLAM System" [2], and a Kinect-based approach of "RGBD-SLAM" [3]. Furthermore "Omnidirectional LSD-SLAM" [4] is proposed, which has better accuracy than LSD-SLAM [1]. In this master thesis, two of the main open-source visual SLAM algorithms (LSD-SLAM [1], ORB-SLAM [2]) have been compared and evaluated. Afterwards, one of the SLAM approaches, was selected by the Thesis Author for indoor positioning of the "Aibot X6 V2", based on 40 evaluations and accuracy analysis on 6 public datasets, each with three different camera calibrations and 2 own datasets with two individual camera calibrations on both SLAM- and Static-Scenes.

2. Approach

The Localization in an unknown indoor environment with the available objects only through monocular vision, defines the problem of SLAM in this master thesis. It is considered as a chicken-and-egg problem, because of the fact that localization and mapping are both unknown and dependent on each other, meaning "3D Mapping" is required for the localization and "3D Pose" is required for "3D Mapping". The pose of the camera consists of the "3D Position" and the "3D Orientation", which forms the six degrees of freedom (6DOF) without considering the scale. About the scale, I would like to mention that it should be considered to form an overall 7DOF for SLAM, because of the fact that in the evaluation section part of this thesis, we can see, how important it is for a SLAM-approach to be "Scale-Aware", in order to be to endure severe camera movements.

2.1 ORB-SLAM [2]**:** It is a feature-based visual SLAM approach, which uses the ORB feature detector [5] to extract features in an indoor environment. In this master thesis, ORB-SLAM [2] is used with a monocular camera and the purpose is to track this camera in a GPS-denied environment only through the means of vision. In the Figure 1, the procedure of "Tracking" can be seen. As it is illustrated after set of features are extracted, in the next step, features are tracked between image pairs, which defines the process of "Map Initialization". It is an important step, because ORB-SLAM [2] has failed to initialize in the SLAM-Scenes in this master thesis, therefore it lost track in lots of conditions. After the map is initialized, camera poses can be estimated as the third step. Besides the Bundle Adjustment (BA) is used to track and optimize the camera locations.

[1] "LSD-SLAM: Large-Scale Direct Monocular SLAM" J. Engel, T. Schöps, D. Cremers, 2014. **[2] "ORB-SLAM: A Versatile and Accurate Monocular SLAM System"** Raúl Mur-Artal, J. M. M. Montiel and Juan D. Tardós, 2015. **[3] "An Evaluation of the RGB-D SLAM System"** F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, W. Burgard, 2012. **[4] "Large-Scale Direct SLAM for Omnidirectional Cameras"** D. Caruso, J. Engel, D. Cremers, IROS, 2015. **[5] "ORB: an efficient alternative to SIFT or SURF"** Ethan Rublee Vincent Rabaud, Kurt Konolige, Gary Bradski, Willow Garage, Menlo Park, California, 2011.

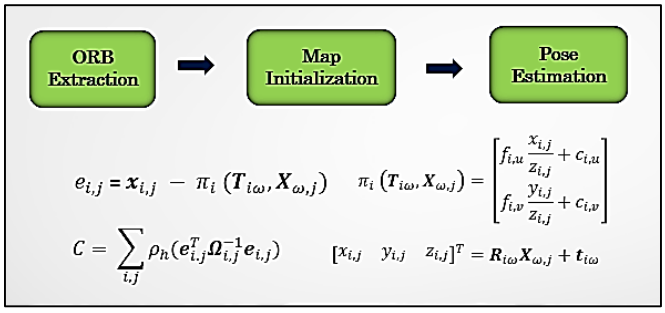

Figure 1: Tracking Pipeline of ORB-SLAM, [2] "ORB-SLAM: A Versatile and Accurate Monocular SLAM System" Raúl Mur-Artal, J. M. M. Montiel and Juan D. Tardós, 2015.

2.2 LSD-SLAM [1]: The "Large-Scale Direct Monocular SLAM" [1] is direct (feature-less) visual SLAM approach, which optimizes the geometry directly on the image intensities, which enables using all information in the image, as a result, it generates a Semi-Dense map of the environment instead of just sparse point clouds, that is produced by ORB-SLAM [2]. LSD-SLAM [1] minimizes the photometric error by the Gauss-Newton method to track the camera, known as "SE(3) Tracking". In the Figure 2, it can be seen that variance-normalized photometric error, illustrated as $E(\xi_{ji})$, which for the case of SE(3) Tracking includes just the black parts of the equation. Because of the fact that it is highly important for a monocular SLAM system to have Scale-Awareness, in the next step, we consider scale for LSD-SLAM [1]. The difference is that in "SE(3) Tracking", the camera pose is computed just for the current frame, that's why scale cannot be estimated. In "Sim(3) Tracking", it can be seen in $E(\xi_{ji})$, that apart from the existing photometric residual (shown in black), the red part (depth residual) is added to form the equation of "Sim(3) Tracking". About the point $p' := \omega(p, D_i(p), \xi_{ji})$, I would like to mention that it's the transformed point of p, after camera's pose ("Rotation" and "Translation"), defined by ξ_{ji}. Furthermore $D_i(p)$ defines the inverse depth of point p and ω is a projective warp function.

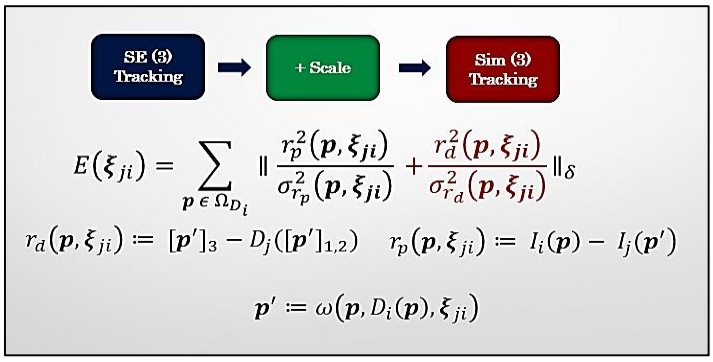

Figure 2: Tracking Pipeline of LSD-SLAM, [1] "LSD-SLAM: Large-Scale Direct Monocular SLAM" J. Engel, T. Schöps, D. Cremers, 2014.

2.3 RGBD-SLAM [3]: It is a visual SLAM approach, which computes the trajectory of the Kinect-Sensor and 3D Map of the environment, using its RGB camera and a depth sensor including an "Infrared-Projector" (IR Pattern) and an "Infrared Camera" to read the projected laser pattern. RGBD-SLAM [3] is considered as a feature-based method, because of the fact that it uses various feature detectors like SIFT [6], SURF [7] and ORB [2] to extract features and at the same time it uses its depth sensor to generate dense map of the environment. Because of its depth sensor, RGBD-SLAM [3] was not considered in the evaluation process of this master thesis, since the objective is using a monocular camera. In the Figure 3, RGBD-SLAM's pipeline is shown. The first step is "Pairwise Feature Matching", that searches for common features in image pairs and tries to find good correspondences. In the next step "Pairwise 6D Transformation Estimation", in order to cope with noisy matches, RANSAC (Random Sample Consensus) [8] is used. After matching of features in two frames, three matched feature pairs are randomly selected, which have minimal number of rigid body transformation in SE(3). "6D Transformation Estimation" include 3D Position and 3D Orientation, which are computed by SE(3) as the camera pose. This test avoid outliers, which their pairwise Euclidean distances do not match. This applies to all matched features. In the next step, "Global Pose Graph Optimization", in order to have globally consistent trajectory, we optimize the pose graph using an optimizer. At the final step "3D Point Clouds" are generated, which uses octree-based mapping to build the map of the environment.

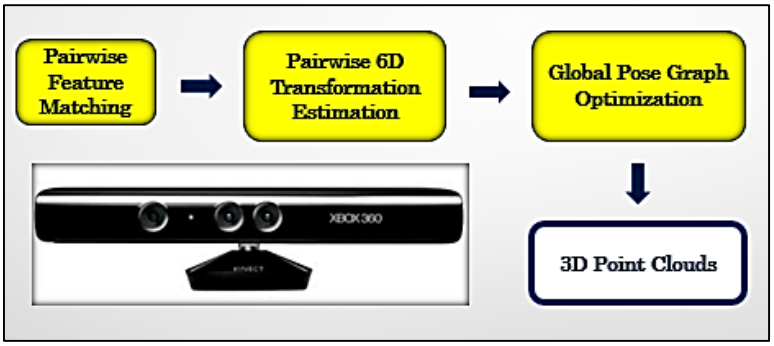

Figure 3: RGBD-SLAM Pipeline, [3] "An Evaluation of the RGB-D SLAM System" F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, W. Burgard, 2012

3. Evaluation

In this Master Thesis, both SLAM approaches (LSD-SLAM [1], ORB-SLAM [2]) have been accurately evaluated in SLAM and Static Scenes. After 40 Evaluations on 8 different Datasets, LSD-SLAM [1] is selected as a Monocular Visual-SLAM Approach for Indoor Positioning of the Aibot X6 V2 because of its better performance in the evaluated SLAM-Scenes of TUM Public Datasets and the first sequence of my own dataset. On TUM Public Datasets, SLAM-Scenes were captured in an indoor environment with several objects around to be tracked, so that a camera can localizes itself, by Direct or Indirect Tracking. As a matter of fact, ORB-SLAM [2] failed to initialize and lost track in 6 out of 9 evaluations, therefore it could not provide any camera pose information in those 6 sequences. In the other 3 evaluations of SLAM-Scenes, ORB-SLAM [2] could provide just few camera poses out of hundreds of images with less accuracy "**0.596597 m**" than LSD-SLAM [1], which had better accuracy "**0.369048 m**" with almost all of the possible camera poses and a performance without even one "Tracking Lost" in all of the SLAM-Scenes. In the first sequence of my own dataset, images were taken in an office at Aibotix GmbH, including very distinctive objects with regular camera motions, however ORB-SLAM [2] failed to initialize and lost track, while LSD-SLAM [1] presented a fast performance with a reasonable accuracy. In the Static-Scenes, ORB-SLAM [2] also did not function well in all the cases and it fluctuated with unreasonable accuracy values. In the second sequence of my own dataset, I have used another strategy for evaluation of camera poses. I have let the camera be stationary at a known coordinate system, and I moved the object with small displacements in different scales. Therefore all the evaluated camera poses in both LSD-SLAM [1] and ORB-SLAM [2] at ideal case should show the same camera poses, which was not the case, therefore RMSE (Absolute Trajectory Error) was calculated for both approaches to evaluate their accuracy. Furthermore I would like to mention that, although ORB-SLAM [2] showed better results in some Static-Scenes under special circumstances, but LSD-SLAM [1] is selected, because of its reliability and good performance on the evaluated SLAM- and Static-Scenes, on both Public and Own Datasets, which proves it as an appropriate choice for Indoor-Positioning of the Aibot X6 V2, that has real SLAM-Scenes as a Hexacopter, which moves through an indoor environment and should be able to localize itself with the means of a robust SLAM-Algorithm.

4. Conclusion

In order to increase the accuracy of LSD-SLAM [1], a fisheye lens can be added to its monocular camera, which defines it as "Omnidirectional LSD-SLAM" [4]. Because of its wider field of view, it can navigate better in an indoor environment and handle severe camera rotations. Strategy of tracking is almost the same like LSD-SLAM [1], but here we used another projection, which is called "Unified Omnidirectional Camera Model" and instead of using inverse depth, inverse distance is used to model points behind the camera. "Sim(3) Tracking" is used for estimation of the camera poses including scale. In the Figure 4, the "Unified Omnidirectional Camera Model" and its projection can be seen.

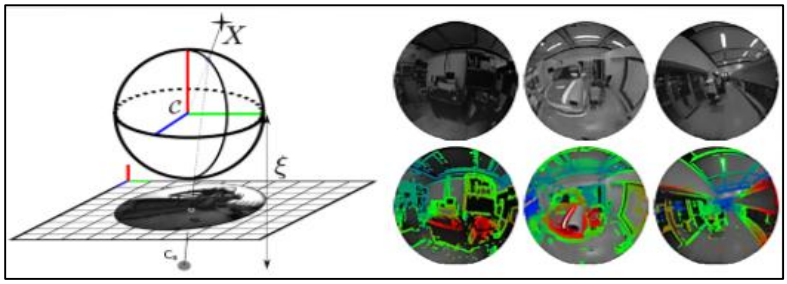

Figure 4: Omnidirectional LSD-SLAM, [4] "Large-Scale Direct SLAM for Omnidirectional Cameras" D. Caruso, J. Engel, D. Cremers, IROS 2015.

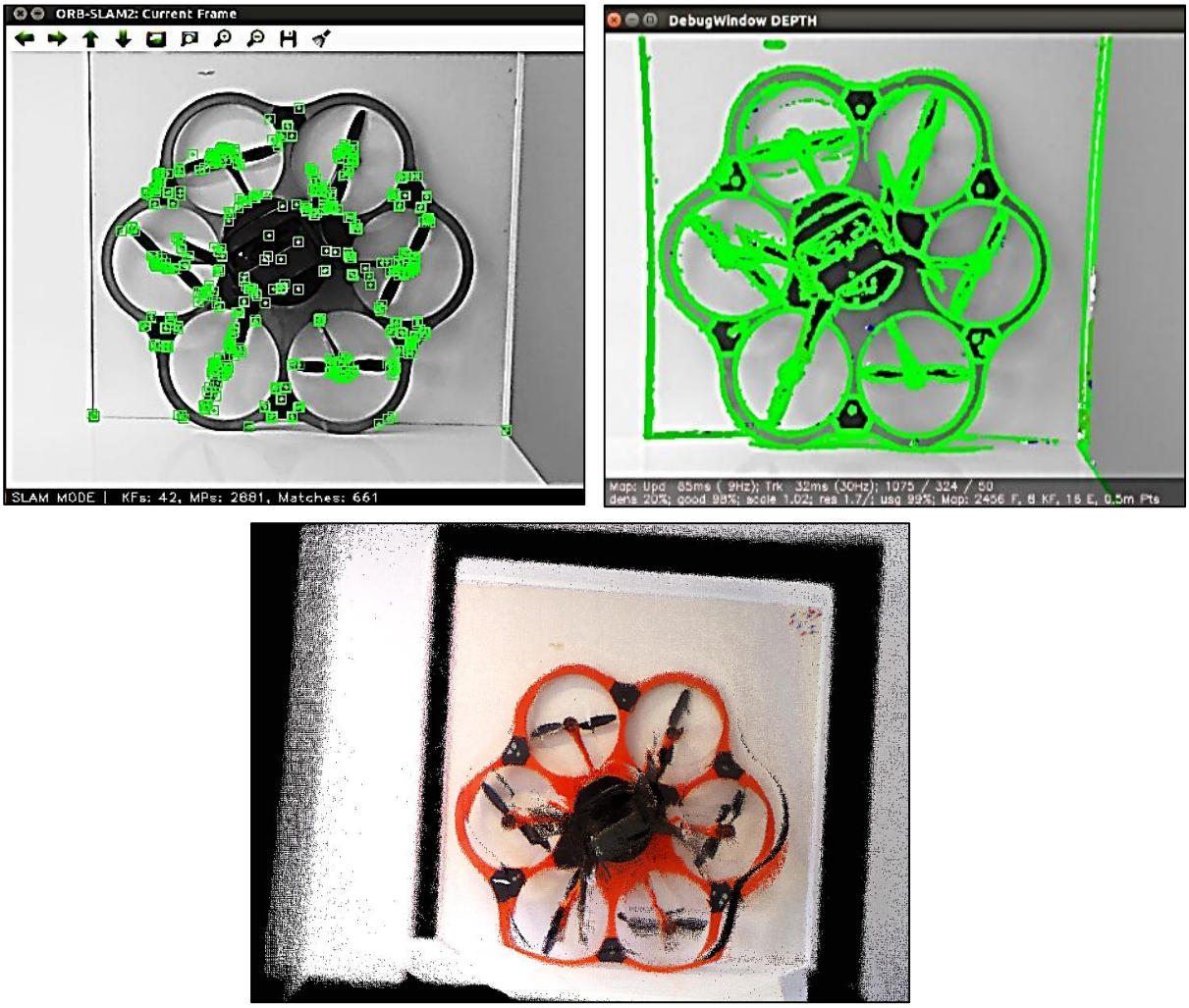

Figure 5: ORB-SLAM [2], LSD-SLAM [1], RGBD-SLAM [3] (from top to down), Object is the "Aibot X6 V2"

Reference

[1] "LSD-SLAM: Large-Scale Direct Monocular SLAM" J. Engel, T. Schöps, D. Cremers, 2014.
[2] "ORB-SLAM: A Versatile and Accurate Monocular SLAM System" Raúl Mur-Artal, J. M. M. Montiel and J. D. Tardós, 2015.
[3] "An Evaluation of the RGB-D SLAM System" F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, W. Burgard, 2012.
[4] "Large-Scale Direct SLAM for Omnidirectional Cameras" D. Caruso, J. Engel, D. Cremers, IROS, 2015.
[5] "ORB: an efficient alternative to SIFT or SURF" Ethan Rublee Vincent Rabaud, Kurt Konolige, Gary Bradski, Willow Garage, Menlo Park, California, 2011.
[6] "Object Recognition from Local Scale-Invariant Features" David G. Lowe, Computer Science Department, University of British Columbia, Vancouver, B.C., V6T 1Z4, Canada, 1999. **"Distinctive Image Features from Scale-Invariant Keypoints"** David G. Lowe, Computer Science Department, University of British Columbia, Vancouver, B.C., Canada, 2004.
[7] "Speeded-Up Robust Features (SURF)" Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, ETH Zurich, BIWI,Switzerland, K. U. Leuven, ESAT-PSI, Belgium, 2008.
[8] "Random Sample Consensus (RANSAC)"
Martin A. Fischler & Robert C. Bolles "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", 1981.