Masterarbeit

Technische Hochschule Deggendorf

Fakultt Angewandte Naturwissenschaften und Wirtschaftsingenieurwesen

Studiengang Mechatronische und cyber-physische Systeme

Implementierung und Evaluierung kamerabasierter Odometriemethoden auf einer mobilen Plattform

Implementation and Evaluation of Camera-based Odometry Methods on a Mobile Platform

Masterarbeit zur Erlangung des akademischen Grades:

Master of Engineering (M.Eng.)

vogelegt von: **Gautam Dobariya**

Matrikelnummer: **777723**

prfer: **Prof. Dr. Ing. Wolfgang Aumer**

Alexander

Cham 1st November 2020

Declaration of Integrity

I hereby confirm, that I have written the Masters thesis at hand independently,that I have not used any sources or materials other than those stated, nor availed myself of any unauthorized resources, and that I have not submitted this Masters thesis in any form as an examination paper before, neither in this country, nor abroad, and hat the electronic copy of this Masters thesis and the printed versions are identical.

Cham, _____ _____

Date Signature

Abstract

In industrial automation, the market needs reliable and scalable solutions for autonomous transportation in production and logistic processes. To address this need, SICK AG e.g. offers reliable LiDaR localization solutions and a big portfolio of LiDaR sensors. However, the increased usage of small mobile platforms in swarm applications introduces additional requirements compared to historically bigger autonomous vehicles: The cost factor regarding number and type of used sensor systems increases, the performance of the hardware is limited and the environment changes.The main target of this work is the investigation, evaluation and adaption of suitable visual Odometry algorithm for warehouse application of AGC.

Keywords:

Standard Standard Standard Standard Standard Standard Standard Standard Standard Standard Standard Standard Standard Standard Standard Standard

Contents

Acronyms

BA Bundle Adjustment. 3, 7

ORB Oriented Fast and Rotated Brief. V, 7

PTAM Parallel Tracking and Mapping. 7

RANSAC Random Sample Consesus. 4

SFM Structure From Motion. 2, 3

SLAM Simultaneous Localization and Mapping. V, 2, 7

V-SLAM Visual Simultaneous Localization and Mapping. 7

VO Visual Odometry. 2–4, 6, 7, 9–12

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In industrial automation, the market needs reliable and scalable solutions for autonomous transportation in production and logistic processes. To address this need, SICK AG e.g. offers reliable LiDaR localization solutions and a big portfolio of LiDaR sensors. However, the increased usage of small autonomous guided carts (AGC) introduces additional requirements compared to historically bigger autonomous vehicles: The cost factor regarding number and type of used sensor systems increases, the performance of the hardware is limited and the environment changes. The main target of this work is the investigation, adaption and evaluation of camera based visual Odometry algorithms for warehouse application in autonomous guided carts (AGC).

1.2 Thesis Structure

The thesis is structured into seven chapters. The Chapter 1 which serves as an introduction.

Chapter 2 covers the basics of Visual Odometry (VO). It explains the different methods of VO. Furthermore it describes state-of-the-Art and explains in depth the selected algorithms namely ORB-SLAM, Direct Sparse Odometry(DSO), Semi-direct Visual Odometry (SVO).

Chapter 3 provides some information about types of camera used in Visual Odometry, Important characteristics and the selection of cameras used in the implementation. It also includes an important task of calibration which affects the result of Visual Odometry.

Chapter 4 describes the experimental setup, data collection and proposed implementation. All algorithms are then implemented. Some improvements are explained and The results are then discussed.

Chapter 5 evaluates the algorithms based on evaluation criteria. They are compared with LiDaR odometry and the best performed algorithm is selected for future implementation.

Chapter 6 concludes the thesis by summarizing the results and suggestions for future works.

Chapter 2

Visual Odometry (VO)

This Chapter gives an overview of Visual Odometry, its working, different types of approaches and state-of-the-art of Visual Odometry.

2.1 Basics

Localization of a robot is a fundamental challenge and one of the most important tasks. For autonomous navigation, motion tracking, and obstacle detection and avoidance, a robot must know of its position in real time. Vision-based Odometry is a novel and robust solution utilized for this purpose.[11] It allows a robot to localize itself accurately by using only a stream of images captured by a camera attached to the vehicle.

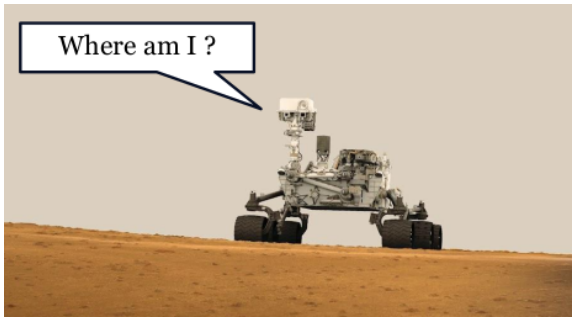

Figure 2.1: NASA Path finder robot[1]]

2.1.1 What is Visaul Odometry ?

VO is defined as the process of estimating the egomotion (translation and rotation with respect to a reference frame) of an Agent(e.g. vehicle, human and robot) by observing a sequence of images using single or multiple cameras attached to it.[8]] VO is a particular case of a technique known as Structure From Motion (SFM) in Computer Vision that tackles the problem of 3D reconstruction of environment and camera poses from set of images[14]. VO mainly focuses on 3-D motion of the camera sequentially in real time (sequential SFM).VO mainly differs with SLAM in terms of global mapping. VO focuses on local consistency and incrementally estimate the path of camera/robot pose, and some local optimization whereas SLAM performs both localization and global mapping.

2.1.2 VO Pipeline

The VO pipeline is summarized in Figure 2.2. For every new image I(or image pair for stereo case), the first two steps consist of detecting and matching 2-D features with those from the previous frames. 2-D features that are the reprojection of the same 3-D feature across different frames are called image correspondences. The feature detection consists of detecting features independently in all the images and then then feature matching will find the same features in sequence of images and then tracks them using a local search technique, such as correlation. The next step consists of computing the relative motion(translation and rotation) between the two consecutive time instants. There are three different approaches for motion estimation depending on the correspondences specified in 3-D or 2-D. Current camera pose is then computed by concatenation of the previous pose. Finally, an iterative local optimization known as BA can be done over the last m frames to obtain a more accurate estimate of the local trajectory. Each steps are discussed further in next section.

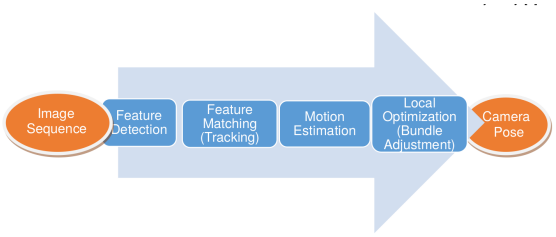

Figure 2.2: Visual Odometry Pipeline

2.1.3 Types of VO

VO mainly classified based on types of camera used for Such as stereo, monocular, omnidirectional, and RGB-D cameras (Fig.). Monocular VO suffers from scale ambiguity because of unknown depth information of images. Stereo VO solves this scaling problem by retrieving depth information using two cameras at little on distance known as baseline. Stereo case can be degraded to monocular if the baseline is much smaller than distances to the scene from the camera.
These methods are then further divided according to their approach as Indirect method(Feature based), Direct method(Intensity based) and Hybrid Approach (mixture of both approaches).

Indirect Approach

This is a classical approach for VO and SFM. The Indirect or Feature-based method involves extraction of some features such as corners, edges etc. from the images frames. See Figure2.3 These features are then matched and tracked among two consecutive image frames. Based on the feature tracking motion of camera is estimated. This approach can typically divided into two steps: 1) Feature detection and matching, 2) geometric optimization on the computed point correspondences. In first step an image is matched with a previous one by comparing each feature in both images and calculating the Euclidean distance of feature vectors to find the candidate matching features.[11] In second step using these match correspondences

the camera motion and surrounding 3D geometry can be estimated. In case of stereo VO the features are first compared with each image pair and thus depth information of feature can be estimated. In this approach the reprojection error is minimized using Bundle Adjustment because keypoints positions(geometric quantities) are used to compute camera pose. The Bundle adjustment problem is described as below.

$$T_{k,k-1} = argmin_T \sum_i \|u_i^{'} - u_i\|_{\Sigma}^2$$

where $u_i^{'} = \pi(P_i, T_{k,k-1})$, u_i is i^{th} pixel 2D positions and $u_i^{'}$ is reprojected 2D pixel position using 3D projection (π).

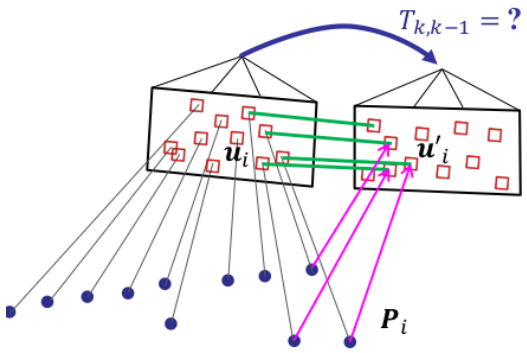

Figure 2.3: Indirect method Optimizes Reprojection Error

The disadvantage of feature-based approaches is their low speed due to feature extraction and matching at every frame, the necessity for robust estimation techniques that deal with erroneous correspondences (e.g., RANSAC), and the fact that most feature detectors are optimized for speed rather than precision.[4]

Direct Approach

Direct method uses directly the pixel intensity as an information instead of extracting features and tracking them for motion estimation. Direct methods are based on assumption that Brightness remains constant in all image frames.[6] Direct methods are also known as Appearance based approach as it monitors the appearance of image in consecutive frames. The camera motion then can be estimated by Optical-flow algorithms which determines the displacement of brightness patterns of a group of pixels using intensity values from one image to another.[11] There are two types of such algorithms based on selection of number of image pixels for calculation called as Dense and Sparse Optical-flow methods. Dense algorithms are less robust to noise as compared to Sparse based. Sparse algorithms select only those features which have more variance than others in particular image region. One of the most used sparse based algorithms for tracking is Lucas-Kanade method.[10] As There is no feature extraction step is involved direct approach minimizes directly the photometric error formulated as below.

$$T_{k,k-1} = argmin_T \sum_i \|I_k(u_i^{'}) - I_{k-1}(u_i)\|_{\sigma}^2$$

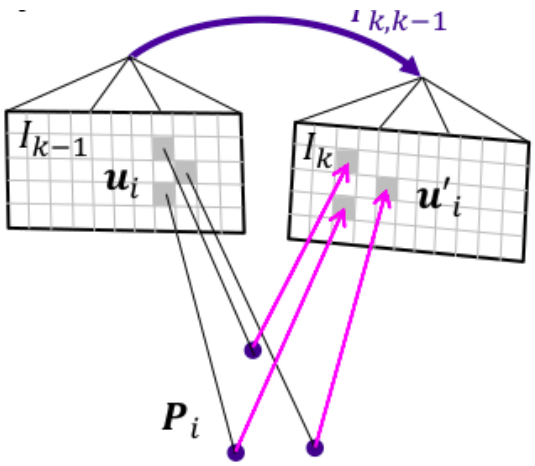

Figure 2.4: Direct Method Optimizes Photometric Error

where $u_i^{'} = \pi(P_i, T_{k,k-1})$ and I_k is k_{th} image. see Figure 2.4

Depending upon the number of feature selection for calculating 3D geometry Direct methods can be divided into three types such as Dense, Semi-dense and Sparse methods.A graphical Overview of these methods can be seen in Figure 2.5. Dense approaches use every pixel in the image, where as semi-dense use just the pixels with high intensity gradient, and the proposed and sparse approach uses selected pixels at corners or along intensity gradient edges.[3]

Figure 2.5: Image-to-model alignment (marked in green for corners and magenta for edgelets) for sparse, semi-dense, and dense methods. [4]

As Direct methods minimize the photometric error (intensity difference) for tracking between two images they required a well calibrated camera as compared to Indirect methods because they minimized the image pixel positions on images. A simple process comparison is described in the Figure.2.6. Indirect methods have been very popular for a long time but recent advances in direct methods have shown better accuracy and robustness, especially when the images do not contain enough explicit corner features.[2] The robustness in the direct approach comes from the joint estimation of motion and correspondences as well as the ability to also use non-corner pixels, corresponding to edges, or even smooth image regions.[5]

Hybrid Approach

The indirect approach fails to deal with texture-less or low-textured environments of a single pattern such as sandy soil, asphalt, and concrete etc.. The less feature

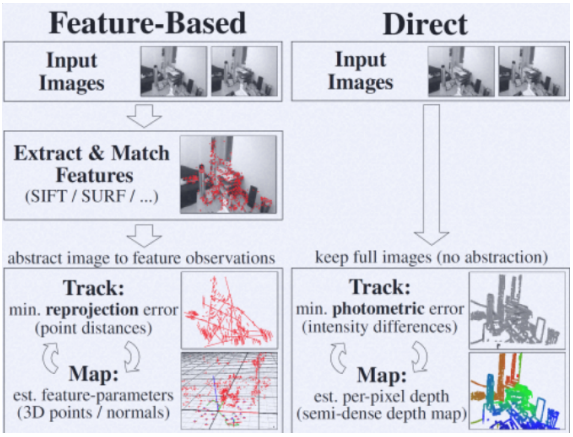

Figure 2.6: Process comparison between Direct and Indirect methods source:[3]

detection in these type of environments make this approach inefficient. While the direct approach is more robust and better than indirect approach in low-textured or single pattern environment but they are not much robust to image occlusions or inconsistencies in system (e.g rolling shutter). To have a advantages of both approaches the best solution is to use the combination of both approaches which combines tracking of salient features over over frames and use the pixel intensity as an information[11]. Forster et al.[4] proposed a hybrid approach in which they use a 4x4 patch around features and estimate the camera pose by minimizing the photometric error of these patches. For pose and structure refinement, the reprojection error of every feature is calculated with respect to the nearest keyframe that has observed the feature at nearly the same angle.

Monocular vs Stereo

In Monocular VO a single camera is used for the whole pipeline. In this case features need to be observed in subsequent frames in order to track motion properly. Features observed in the first frame are triangulated into 3D points with help of second frame, and then transformation can be calculated using third frame [8]. While in stereo case, 3D points can be reconstructed (by triangulation) only by observing the features in the left and right images of a single pair simultaneously. Motion is estimated by observing features in two successive frames (both in left and right). Stereo approach has advantage of depth information of environment because it can obtain the disparity in scene. Where as monocular cameras can measure motion using pixel information only with no knowledge of scene depth. When the distance between scene to stereo camera becomes very long compared to the baseline(distance between left and right camera), the stereo case can be degraded to monocular because its very erroneous to measure the depth for far scene.

2.1.4 State of the Art

VO has been very active research topic in recent years. There has been many algorithms published based on the approaches discussed in section 2.1.3 and research is still ongoing. There are several papers which describes the current state

of VO. They are [11], [8] ,[14]. Currently research is focused based on the deep learning methods [15] [16] which is not in the scope of this thesis. Though VO and V-SLAM are widely researched topics it is still difficult to get an overview of all the algorithms. A list of various VO and V-SLAM algorithms, references and code if available can be found at A.1. This list has been referred from [7]. The list presents mostly all algorithms invented so far. This thesis is based on the algorithms which work only based on camera. Also ,these algorithms can be classified according to their approaches. Considering these facts and other criteria such as

1. runs on CPU
2. open source availability
3. real-time
4. state-of-the-art

three algorithms (one from each approach)

1. ORBSLAM [12]
2. Direct sparse odometry with loop closure (LDSO)[5]
3. Semi-direct visual odometry(SVO) [4]

are selected for implementation and later evaluation. In the next section these three algorithms their approaches, pros and cons compared to each other are described.

ORBSLAM

ORBSLAM is a very popular feature(Indirect) based visual SLAM approach. It uses an open-source ORB feature descriptors as feature extraction and matching, which was developed by Rublee et al.[13].These ORB features are robust to rotation and scale and also provides good invariance to auto-exposure and illumination changes. Further more they are fast to extract and match which makes them suitable for real-time applications[12]. Mur-Artal et al. [12] used an approach of parallel threads for ORBSLAM similar to that of used in PTAM [9]. It uses three main parallel threads:1) tracking 2) local mapping 3) loop closing. Loop closing is the thread of performing full BA. A detailed approach is given in the figure 2.7.

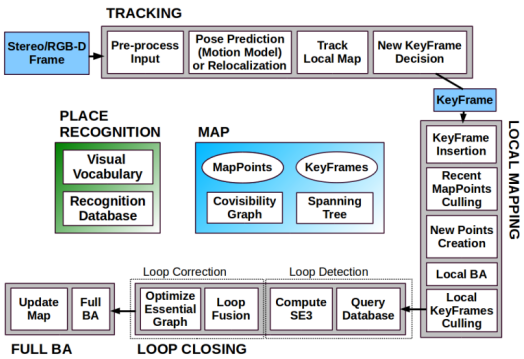

Figure 2.7: A process overview of ORBSLAM [12]

Direct Sparse Odometry

Semi-direct Visual Odometry

Chapter 3

Camera

This chapter covers details of camera, the main hardware used in VO. Different types of camera and important properties for VO implementation, selection of camera and calibration are explained.

3.1 Classification of Camera

Camera can be classified mainly into two types

a. Passive Camera

i. Monocular
ii. Stereo
iii. Omnidirectional

b. Active Camera

i. Time-of-flight (TOF)
ii. RGB-Depth

Passive camera are mostly used in VO implementation. Some types are shown in below figure.

3.2 Important Properties for Selection

There are some criteria to select the proper camera model in order to get satisfactory results. First important property is shutter technology. The rolling shutter cameras have geometric distortions in the image for high frame rate. As direct approaches of VO are not meant to optimize geometric noise. Another important factor is Field of view (FOV) of Camera. By having a large FOV there would be enough information (features) to estimate trajectory and the algorithm will not be crashed or struggle to relocalize. Higher camera resolution will increase the accuracy of 3D pose of the features but at the same time it will increase the computation cost as the image size will be bigger. Direct approaches are also not robust to automatic exposure changes. Therefore, for the direct approach implementation manual focus with fixed exposure time will provide better result. Lastly the suitable lens object with manual focus can reduce the effect of Vignetting which occurs due to blockage of light due to some camera elements or hoods attached to lens.

3.3 Selection of Camera

Based on discussion of camera properties above and availability, two different type of cameras with different properties were selected so that results can be compared and the better performed camera will be selected for the adaptation and further research. In table 3.3 the properties of these cameras are compared with each other and with Ids UI-3241LE which is used in TUM- benchmark dataset.

Model	SICK Picocam I2D304C-RCA11	Genius Widecam F100	Ids UI-3241LE benchmark)
Shutter technology	global	rolling shutter	both (rolling and
Lens type	C-mount	attached	S-mount
Max. fps	19	30	60
Max. Resolution	2048*2048 (4.19 MP)	1280 * 720	1280 *1024 (1.31
Exposure time (ms)	0.0009 - 2000	auto	
Sensor size (mm)	11.26 x 11.26	6.784 x 5.427	
FOV (degree)	90 Diagonal	120	98 x 79
Lens	Kowa, LM8HC		Lensagon BM40

3.4 Camera Calibration

For any VO method the camera calibration is an important part of preparation. Though some cameras are manufactured very well, they still have some distortions. Using cameras directly without doing calibration can lead to wrong trajectory and VO will not perform well. camera calibration can be classified into two types 1. Geometric and 2. Photometric. Photometric calibration covers the effect of shutter speed, motion blur and vignette. It is mostly recommended for cameras which have rolling shutter and auto-exposure technology and for direct approach which uses directly image pixel intensity values for tracking [17]. This section will discuss only geometric calibration as photometric calibration is not done in the experiment due to its complexity and non-necessity.

3.4.1 Pinhole Camera Model

3.4.2 Intrinsic Parameters

3.4.3 Extrinsic

3.4.4 recommendations for good calibration

3.4.5 calibration experiments and result

result is shown in appendix ...

Chapter 4

Experimental Setup

This chapter covers hardware part of the thesis which is camera. Different types of camera and important properties for VO implementation are explained.

4.1 Experimental Setup

Robot setup figures and homogeneous transformation with details in appendix

4.1.1 Proposed Implementation

explain figures of implementation in software framework.

4.1.2 Data acquisition

How data is collected and figures of different sequences (with or w/o loop closing etc.). warehouse setup figures. manual or autonomous navigation.

Some conditions or requirements of good data collection e.g. lighting etc.

4.2 Modifications and changes in implementation

e.g with less features, less RANSAC iteration, window search size, no. of keyframes etc.... with figures.

4.3 results

Chapter 5

Evaluation

5.1 Evaluation

5.1.1 Evaluation criteria

some criteria like accuracy(absolute trajectory error), time, efficiency(computation cost), robustness(re-localization), ability to find loop closure, environment condition such as lighting, feature less area.

5.1.2 Comparison

comparison with LiDaR , ground truth(?) ... graphs of 2D trajectory and results of evaluation parameters. and selection of best performer.

5.1.3 Observations

why it works well and why not others problems with others.

5.1.4 suggestions

if possible any suggestions for further improvements

Chapter 6

Conclusion

conclusion

References

[1] author. *Title of Citation*. source:www.jpl.nasa.gov.

[2] J. Engel, V. Koltun, and D. Cremers. "Direct Sparse Odometry". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Mar. 2018).

[3] J. Engel, T. Schöps, and D. Cremers. "LSD-SLAM: Large-Scale Direct Monocular SLAM". In: *European Conference on Computer Vision (ECCV)*. Sept. 2014.

[4] C. Forster et al. "SVO: Semidirect Visual Odometry for Monocular and Multicamera Systems". In: *IEEE Transactions on Robotics* 33.2 (2017), pp. 249–265.

[5] X. Gao et al. "LDSO: Direct Sparse Odometry with Loop Closure". In: *iros*. Oct. 2018.

[6] M. Irani and P. Anandan. "All About Direct Methods". In: *ICCV '99: Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*. 1999, pp. 267–277.

[7] Chris Kahlefendt. "Implementation and Evaluation of Monocular SLAM for an Underwater Robot". type. Hamburg University of Technology.

[8] Reza Hoseinnezhad Khalid Yousif Alireza Bab-Hadiashar. *An Overview to Visual Odometry and Visual SLAM: Applications to Mobile Robotics*. 2015. DOI: 10.1007/s40903-015-0032-7.

[9] G. Klein and D. Murray. "Parallel Tracking and Mapping for Small AR Workspaces". In: *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. 2007, pp. 225–234.

[10] Bruce D. Lucas and Takeo Kanade. "An iterative image registration technique with an application to stereo vision". In: *In IJCAI81*. 1981, pp. 674–679.

[11] M.Iqbal Saripan Mohammad O.A.Aqel Mohammad H.Marhaban and Napsiah Bt.Ismail. *Review of visual odometry: types, approaches, challenges, and applications*. 2016. DOI: 10.1186/s40064-016-3573-7.

[12] Raul Mur-Artal and Juan D. Tardós. "ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras". In: *CoRR* abs/1610.06475 (2016). arXiv: 1610.06475. URL: http://arxiv.org/abs/1610.06475.

[13] E. Rublee et al. "ORB: An efficient alternative to SIFT or SURF". In: *2011 International Conference on Computer Vision*. 2011, pp. 2564–2571.

[14] D. Scaramuzza and F. Fraundorfer. "Visual Odometry [Tutorial]". In: *IEEE Robotics Automation Magazine* 18.4 (2011), pp. 80–92.

[15] S. Wang et al. "DeepVO: Towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks". In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 2017, pp. 2043–2050.

[16] N. Yang et al. "D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. arXiv: 2003.01060 [cs.CV].

[17] Nan Yang et al. *Challenges in Monocular Visual Odometry: Photometric Calibration, Motion Bias and Rolling Shutter Effect*. 2018. arXiv: 1705.04300 [cs.CV].

Appendix A

Title of appendix A

A.1 Section title