Höhere Technische Bundeslehranstalt
und Bundesfachschule
im Hermann Fuchs Bundesschulzentrum

Autonomous Car Mapping and Tracking

Diploma Documentation

School autonomous focus on Mobile Computing and Software Engineering

Performed in school year 2019/2020 by:

Alexander Voglsperger (AV), 5AHELS

Simon Moharitsch (SM), 5AHELS

Advisors:

Dipl. Ing. Müller Gerhard

February 12, 2020

Thema:
Autonomous Car Mapping and Tracking

Subtopics and Editor:

Implementing SLAMS and DeepTAM, Image Pre-Processing
Alexander Voglsperger, 5AHELS
Advisors: Dipl. Ing. Müller Gerhard

Implementing DeepTAM, Gathering Trainingdata
Simon Moharitsch, 5AHELS
Advisors: Dipl. Ing. Müller Gerhard

Projectpartner:

Designation:	Johannes Kepler University - Artificial Intelligence Lab
Address:	Altenberger Straße 69
ZIP, location:	4040 Linz, Austria
Contact person:	Dr. Nessler Bernhard
Phone:	+43 (0)732 2468 4539
E-Mail:	nessler@ml.jku.at

Höhere technische Bundeslehranstalt
und Bundesfachschule Braunau
Elektronik und Technische Informatik
School autonomous focus on Mobile Computing
and Software Engineering

DIPLOMA DOCUMENTATION

Author Alexander Voglsperger, Simon Moharitsch

Vintage Schoolyear 5AHELS 2019/2020

Topic of the diploma documentation Autonomous Car Mapping and Tracking

Cooperation-partner Johannes Kepler University - Artificial Intelligence Lab

Taskdefinition A camera delivers a sequence of 2D pictures of the environment in front of a car. Only sing these pictures the programm should generate a 3D map. Since the pictures don't contain any depth information a SLAM (Simultaneous Localization and Mapping) should be applied.

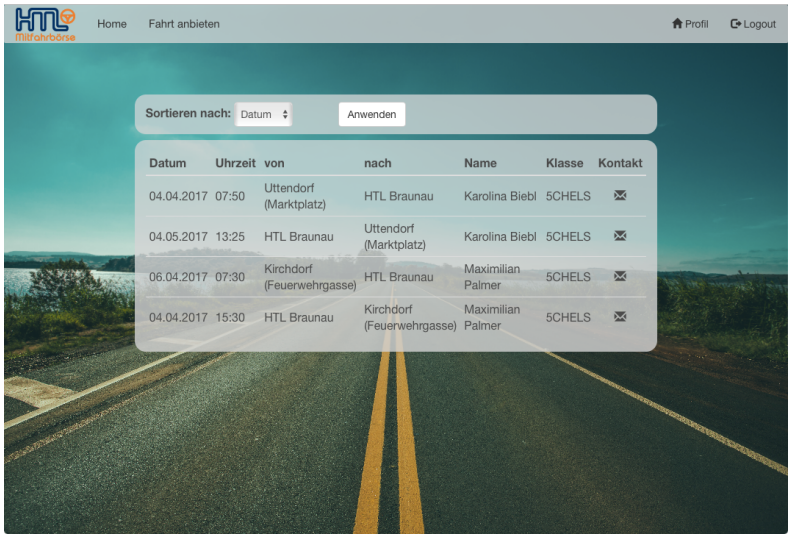
Realization As a foundation ROS was used because it is freely available and has an active community supporting the project. The ORB SLAM and LSD SLAM have already been implemented in ROS as nodes and can be used with changing a few things to get it working. DeepTAM is fairly new and hasn't been implemented into ROS yet.

Outcome Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Höhere technische Bundeslehranstalt
und Bundesfachschule Braunau
Elektronik und Technische Informatik
School autonomous focus on Mobile Computing
and Software Engineering

Illustrative graph, photo (incl. explanation)

Landing page of HTL-carpooling:



Accessibility of diploma thesis

HTL Braunau archive, or
`https://diplomarbeiten.berufsbildendeschulen.at/`

Approval (date / signature)

Examiner	Head of College / Department

Statement

I declare in lieu of oath that I have written this diploma thesis independently and without outside help, have not used sources and aids other than those stated directly and have made the sources used verbatim and in terms of content taken as such recognizable.

Braunau/Inn, 12.02.2020	Alexander Voglsperger	_____
Location, Date	*Author*	*Signature*

Braunau/Inn, 12.02.2020	Simon Moharitsch	_____
Location, Date	*Author*	*Signature*

Contents

Abstract

Im Vorwort teilt der Bearbeiter dem Leser wichtige Tatsachen mit, die Erklärungen zu seiner Arbeit beinhalten – z.B. die Motivation für die Bearbeitung des Themas oder besondere Schwierigkeiten bei der Bearbeitung und/oder Materialbeschaffung.

Hier können auch Mitteilungen persönlicher Natur enthalten sein – z.B. Dank an Institutionen/Personen für die geleistete Unterstützung.

Summary

Die *Zusammenfassung* oder auch *Kurzfassung* soll den Inhalt der Diplomarbeit auf maximal einer halben Seite zusammenfassen.

Dieses Dokument dient als Vorlage und Beschreibung für die Dokumentation der Diplomarbeit. Es werden Hinweise zur Erstellung einer guten Dokumentation gegeben. Dies betrifft welchen Inhalt die Arbeit haben soll genauso wie welche Regeln eingehalten werden müssen und mit welchen technischen Mitteln das Dokument erstellt werden kann.

Beim Inhalt dieser Arbeit wurden alle grundlegenden Qualitätsregeln eingehalten und kann daher als Musterlösung gesehen werden. Zum Erstellen wurde das Textsatzsystem LaTeXverwendet. Es ist vorgesehen, dass der LaTeXQuelltext dieses Dokuments als Ausgangspunkt für die eigene Dokumentation verwendet wird.

Dieses Dokument sollte unbedingt aufmerksam gelesen werden ehe mit der eigenen Arbeit begonnen wird.

1 SLAM^{AV}

1.1 What is SLAM?

SLAM is an acronym and stands for **S**iultaneous **L**ocalisation **A**nd **M**apping. *"SLAM is concerned with the problem of building a map of an unknown environment by a mobile robot while at the same time navigating the environment using the map."* [1]
This problem is thus a chicken-and-egg problem because neither the map or location are known, and have to be estimated at the same time. Cameras, ultrasonic sensors and laser radar (Lidar) sensors are most commonly used for fetching the 2D and 3D data of the robot's surroundings [2].
There are several algorithms out there, which try to solve this problem using algorithms and some even deep learning. Mostly they achieve an approximate map, which is done in a reasonable time span. Many popular SLAM-algorithms use methods that include *particle filters*, *extended Kalman filter* and *co-variance intersection*[1] [3].

1.2 Application

The biggest selling point for using SLAM implementations is pretty simple. Many places where autonomous robots may be required don't have good enough maps that are up-to-date , if the exist at all or it might be in an environment where positioning for instance GPS can't be used properly [4]. If slams weren't be used then someone would have to go to the place and make a map. This would delay the mission and add to the costs.
With a robot, that is capable of using a SLAM method to detect and locate itself in the unknown surroundings this wouldn't be an issue. The robot could go in, generate a map that updates itself and use it to navigate around.
Existing approaches that are used are in self-driving cars, unmanned aerial vehicles, autonomous underwater vehicles, planetary rovers and newer domestic robots.

1.3 History

The decisive work in SLAMs was tone in the research by R.C. Smith and P. Cheeseman who worked on the representation and estimation of spatial uncertainty in 1986. Another major work in this area was done by a research group with the head being *Hough F. Durrant-Whyte*. Durrant-Wyhte and his group showed that answer to SLAMs lies in the nearly infinite amount of data that can be used. This lead to the motivation of finding algorithms which are trackable and approximate in a time realistic manner.
Sebastian Thrun was playing.

1.4 Existing Methods

There exits a big variety of SLAM methods, that try to achieve the same goal using different approaches [5]. Most known or popular are the following:

- EKF SLAM
  Utilizes the extended Kalman filter. The algorithm uses the likely-hood for data association. It was the go-to SLAM from 1990 to the early 2000s until Fast SLAM was introduced [6].
- Fast SLAM
  Works recursively so it scales logarithmically to the scale of the landmark. It can handle much bigger landmarks than the EKF-SLAM ever could without requiring as much computing power [6].
- ORB-SLAM2
  It's a real-time SLAM library for Monocular, Stereo and RGB-D cameras. It can detect loops and relocate the camera in real-time. It uses camera trajectory and sparse 3D reconstruction to get information out of the image sequence [7].
- DVO-SLAM
  Implements a *dense visual SLAM* system for RGB-D cameras. It's based on *Dense Visual Odometry* and was extended to include frame-to-key matching with loop closure to older key-frames [8].
- RGB-D SLAM
  Utilizes the depth information of a RGB-D camera, e.g., Microsoft Kinect or Intel Real-Sense Cameras [9].
- LSD-SLAM
  It's a direct monocular SLAM. It tracks the *direct image alignment* and estimates geometry in form of *semi-dense depth map* instead of relying on keypoints [10].

2 Robot Operating System[AV]

2.1 What is the Robot Operating System?

The Robot Operating System, which is also known as ROS is a flexible framework for writing software that gets utilized on robots. It was founded by Willow Garage in 2012 and gets primarily maintained by the Open Source Robot Foundation (OSRF) [11]. In Europe the project gets coordinated by the Fraunhofer IPA in form of the *ROS Industrial Consortium Europe*. ROS is a middle-ware which is not a operating system but provides services that manage hardware abstraction, low-level device control, message-passing between processes and package management. *"It is a collection of tools, libraries, and conventions that aim to simplify the task of creating complex and robust robot behavior across a wide variety of robotic platforms."* [12]

2.2 Design

The processes of ROS are represented in nodes which are in a graph structure. Everything gets managed by a single process called *ROS Master*, to whom all other nodes register on startup. But instead of sending all of the messages over the master, the master sets up a peer-to-peer connection between the nodes. This decentralized architecture is helpful as many robots consists of many computer hardware which is connected via a network and are likely to transfer big messages [13].

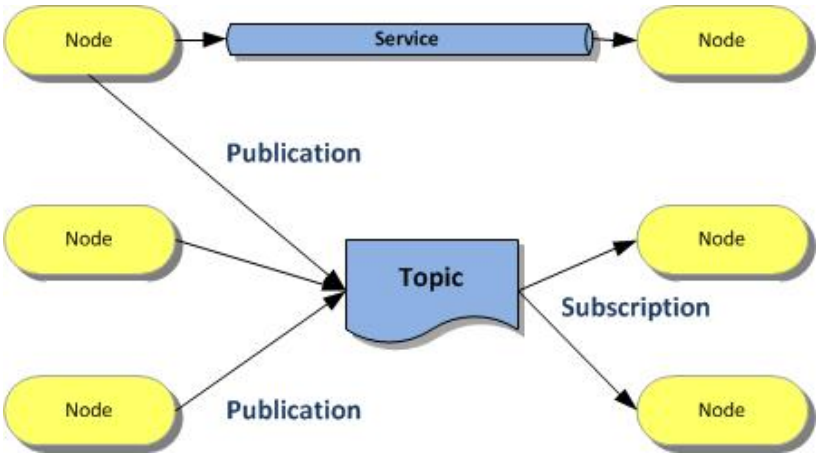


Figure 2.1: ROS structure
Source: `https://tinyurl.com/tkf2smq`

2.2.1 Topics

It is based on a topic system, where a topic acts like a bus over which nodes send and receive messages. Each topic must be unique in it's name, which is usually set by the developer. The process of publishing and subscribing is handles anonymously so that no node knows which nodes are sending and receiving messages on a certain topic.

2.2.2 Nodes

A node, which represent a single running process, can provide data using a matching topic and publish it to the system, where theoretically every other node can subscribe to it, to get the data.

2.3 Licenses and OS

The language-independent tools and the main client libraries have men released under the BSD license and as such they are open source software for commercial and research use. The majority of 3rd party packages are released under several other open-source licenses.
The ROS libraries are geared toward a UNIX-System which is mainly due to their dependence on a large collection of open source software and libraries. For example *Ubuntu* is in the list of supported operating systems, while others like *Fedora, Mac OS and Windows* are "experimental" and are mainly supported by the community [14].

2.4 Tools

One of the core functionalities that ROS provides are the tools which allow the developers to visualize 2D and 3D data, record data, easily navigating ROS packages, creating complex scripts that configure and setup processes. Thanks to this tools it simplifies and provides solution for common robotic development.

2.4.1 Rosbag

Rosbag is a tool that can be used over the command line to record, playback and store ROS message data. The data gets stored in a file called bag, where it records the messages as they come in. It's possible to play these bag files. By doing this the recorded messages get published into the system, as they where live. It's very handy if you need data for later development or to use a bunch of different scenarios for testing

2.4.2 RQt

RQt provides a graphical overview of the ROS computation graph. It shows the nodes and how they are connected to each other. It also shows if a node is even subscribing to a topic or publishes something. Other than that it can be used to subscribe to different topics and show them directly in RQt.

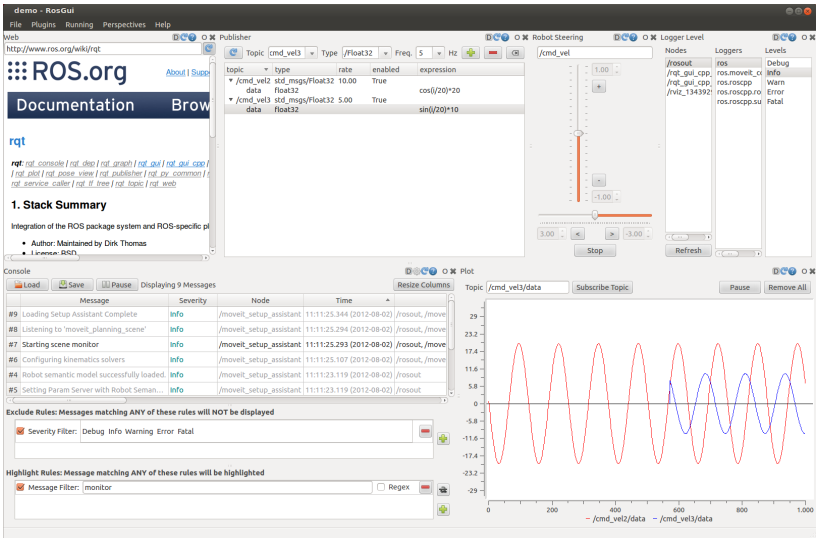


Figure 2.2: RQt interface
Source: `https://wiki.ros.org/RQt`

2.4.3 CatKin

Catkin is the newer ROS build system, which compiles the files in the source folder. It is based on CMake and is cross-platform and language-independent as most other ROS tools.

2.4.4 Rviz

A visualizer for three-dimensional data where robots, environments and sensor information can be visualized. It is highly customise able with display many types of visualisation and plugin support.

2.4.5 Roslaunch

Roslaunch is a tool for launching multiple ROS nodes and setting parameters on startup. It can be used to launch nodes locally or remotely on a server. The configuration for a start script is written in a launch file using XML. In these files it's easy to make a automated startup and configuration process to be executed with one command. It's possible to execute launch files in other launch files to chain them together.

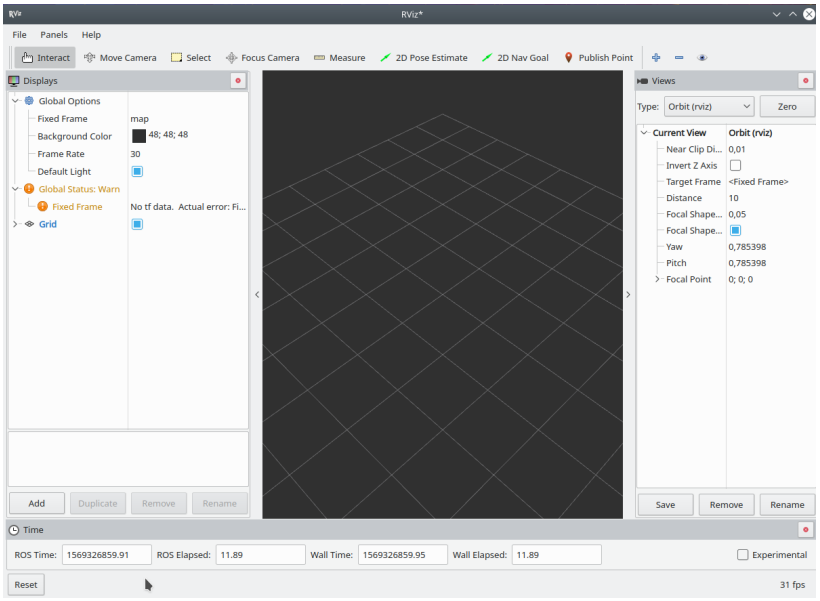


Figure 2.3: Rviz interface

3 Artificial Neural Networks[SM]

3.1 What is a Artificial Neural Networks?

Artificial **N**eural **N**etworks(ANN) are inspired by biological neural networks that constitute animal brains. Important to notice is that they are not faithful models of biologic neural or cognitive phenomena. In fact most of these models are more closely related to mathematical and/or statistical models(For Example: clustering algorithms). Such systems "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules.

3.2 Areas of Application

ANN are viable computational models for a wide variety of problems, including pattern classification, speech synthesis and recognition, adaptive interfaces between human and complex physical system, function approximation, associative memory, clustering, forecasting and prediction, combinatorial optimization, nonlinear system modeling, and control [15]

3.3 Components of an ANN

Simplified a ANN consists of three main components(neurons, connection and the weight associated with them) the propagation function and a bias. In the following topics I will give you a short summary what these components are and afterward I will explain how they work together.

3.3.1 Neurons

Neurons are elementary units in an ANN. A neuron gets one ore more inputs and depending on the value of the inputs the output is set. A neuron can get its inputs from other neurons or, if its at the beginning, from the source of the data that needs to be processed. Depending on the Type of ANN they are placed in different structures. In most cases the output of a Neuron is a number between 0 and 1.

3.3.2 Connection and weights

These Neurons are Connected. Which neurons are connected with others depends on the structure. The weights characterize how important a connection between neurons is.

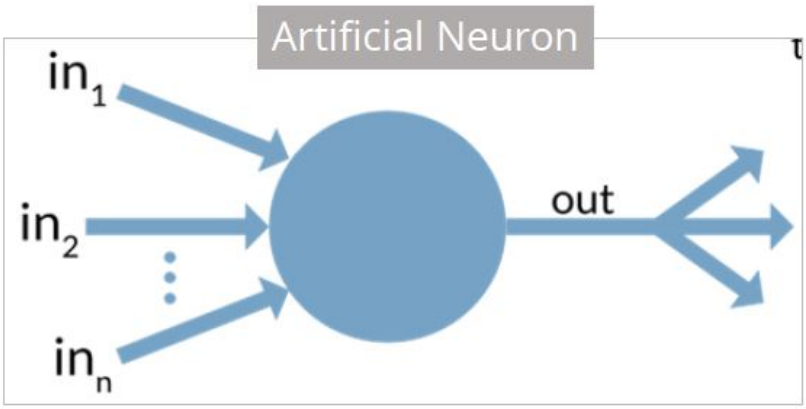


Figure 3.1: General view of Neurons
Source: https://tinyurl.com/yyfthk7c

For example:
A neuron, we call it base for this example, has two neurons connected to it as inputs. The weight of the connection of the first neuron has a bigger weight than the second connection. That means the output of the base depends more on the input of the first neuron

3.3.3 Propagation function and activation function

This is a function witch takes the Inputs of a neuron, the weight of these connections and the bias and adds them up. The resulting value is the processed by the activation function which sets the output. One of the most common activation function is the sigmoid function because it is not a step function wich means the output doesn't change instantaneously. Thats important for the training algorithmus.

3.3.4 Bias

The bias is a Neuron which has no Inputs. A bias is used to shift the decision boundary to the left or right.

3.4 Organization

A artificial neural network can be organized in many different ways.

3.4.1 Feed Forward ANN

The following picture demonstrates a feed forward ANN .There are a variable number of hidden layers depending on the purpose of the neural network. Nothing in the hidden layer is visible.

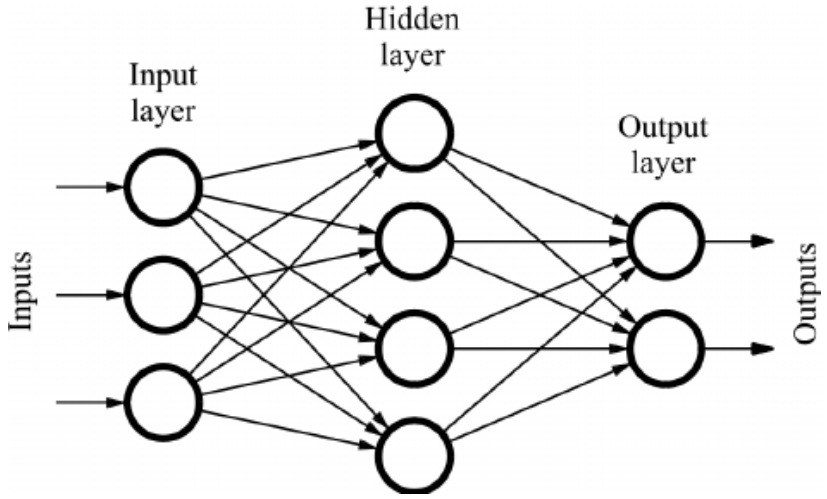


Figure 3.2: Feed forward network

3.4.2 CNN

Convolution Neural Network (CNN) is the most important Network organization for our task. It is based on the human visual cortex and perfect for image and video recognition. CNN is based on the human visual cortex. The components of a CNN are a series of convolution and sub-sampling layers followed by a fully connected layer and a normalizing layer.

How a CNN works will be explained in the following example. The same example like in "Review of Deep Learning Algorithms and Architectures" will be used[16, 17]

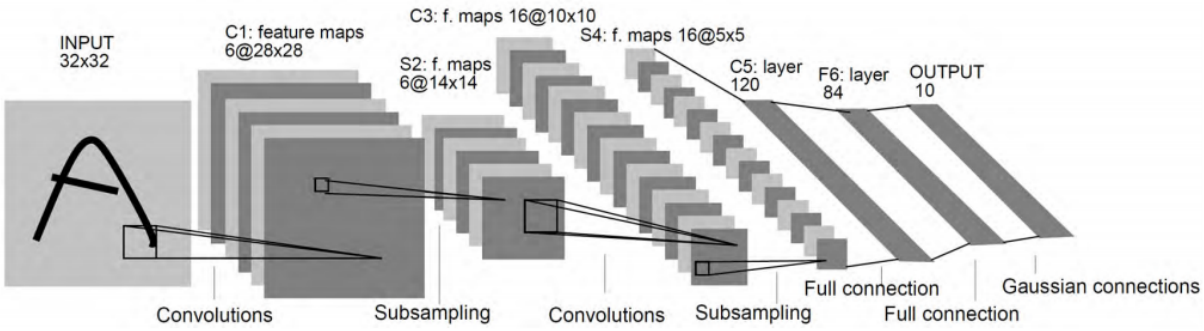


Figure 3.3: 7-layer architecture of CNN for character recognition
[16, 17, Fig. 4.]

Progressively more refined feature extraction at every layer is performed by the series of multiple convolution layers. This process is moving from input to output layer. After the convolution layers there are fully connected layer that perform classification. There is also the possibility of putting Sub-sampling or pooling layers between each convolution layer. The input of a CNN is 2D n*n pixelated image. Each layer consists of filters or kernels (groups of 2D neurons). In most neural networks neurons in each feature extraction layer are connected to all neurons in the adjacent layers. But in a CNN they are only connected to the spatially

mapped fixed sized and partially overlapping neurons in the previous layer's input image or feature map.

3.5 Encoder-Decoder-Based Architecture

4 ORB-SLAM2^{AV}

4.1 What is ORB-SLAM?

ORB-SLAM2 is a versatile, real-time SLAM implementation which uses Mono-, Stereo- and RGB-D cameras. It's designed to generate a 3D map from prominent points in the picture and keypoints. It features loop closing, re-localization and a reusable map [18]. It works in a wide variety of use cases. The SLAM can be used on a small hand-held camera or drones up to self driving cars. ORB-SLAM2 is based on ORB-SLAM and was inter alia developed by Raúl Mur-Artal who already worked on ORB-SLAM.

Figure 4.1: ORB-SLAM Example image
Source: `https://tinyurl.com/ruvnj39`

4.2 How does the ORB-SLAM work?

4.2.1 Extracting Keypoints

The SLAM uses a feature-based method. This means that it extracts features on prominent keypoints throughout the image input. These feature information is then distributed to all operations which handle them independent from the camera type. [18]

This is how finding these keypoints works on different camera types:

- Stereo Image
  For a stereo camera setup the keypoints get extracted for both images separately and then the left keypoints are searched on the right image. Then the found points get compared to the original ones that where found on the right side

- RGB-D Image
  On a RGB-D camera keypoints get extracted using prominent keypoints and then calculating the approximate position using the depth information from the information from the depth sensor.
- Monocular Image
  On a Monocular image the approximate position gets triangulated by using multiple images. The Disadvantage is that they don't provide a scale information and only do rotational and transnational movement estimations.

4.2.2 Loop-closing and Bundle Adjustments

Loop-closing and bundle adjustments are performed in two steps. First the loop-closing will happen when the system detects overlapping environments where the system changes scaling to reconnect certain parts as scale drifting will occur on monocular cameras.
Second step is the bundle adjustment, which gets executed after a successful loop-closing, where the system tries to optimize all keypoints and keyframes using the Levenberg-Marquardt method (alternative to Gauss-Newton method).[19] Also the camera orientation and position will be optimized to compensate errors in tracking. All the bundle adjustment is done in a separate thread since this is a heavier task.
When finished, the updated and optimized keyframes and keypoints get merged into the original keyframes and keypoints [18].

4.2.3 Localization

When a area has been mapped well in the past the *Localization Mode* can be turned on which deactivates the Local Mapping and the Loop Closing thread and this saving computing power. Locating is done by continuously comparing the previous points with the current points of the image. This works when an area is unmapped but drifting might add up. Matching the current points with the one on the map will ensure that it is drift-free [18].

4.2.4 Input/Output

Input Data: rectified Monochrome/Color images
Output Data: Rough 3D map with pixel-points and image with current prominent key-points

5 LSD-SLAM^{AV}

5.1 What is LSD-SLAM?

LSD-SLAM stands for **L**arge-**S**cale **D**irect Monocular SLAM and is a fairly new real-time monocular SLAM that is fully direct-driven instead of relying on keypoint/keyfeature [10]. The algorithm works on the image intensity for tracking and mapping at the same time. This method allows building large-scale maps on normal processors that are consistent when comparing it to current state-of-the-art algorithms.

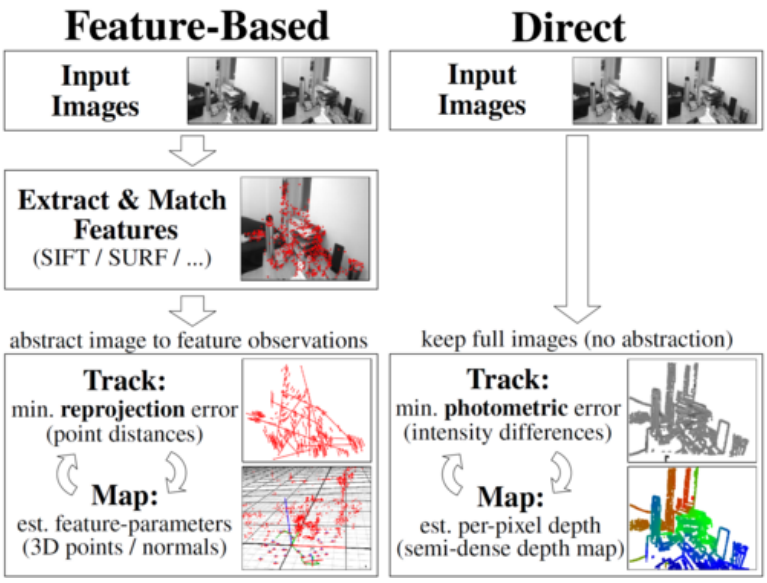


Figure 5.1: Feature-Based and Direct Difference
Source: `https://tinyurl.com/ycmjhb9d`

5.2 Difference Feature-Based and Direct

- Feature-based
  A feature-based SLAM (e.g. ORB-SLAM 4) looks for distinctive points in the image and then uses only these keypoints to process the information. When there are only a few prominent keypoints (e.g indoor, tunnels) the result will not be that accurate.
- Direct
  Direct based SLAMs use all information that's provided in the image. This does not

only include distinctive points but also edges and sometimes surfaces and thus makes it possible to create a more accurate and denser 3D map [10].

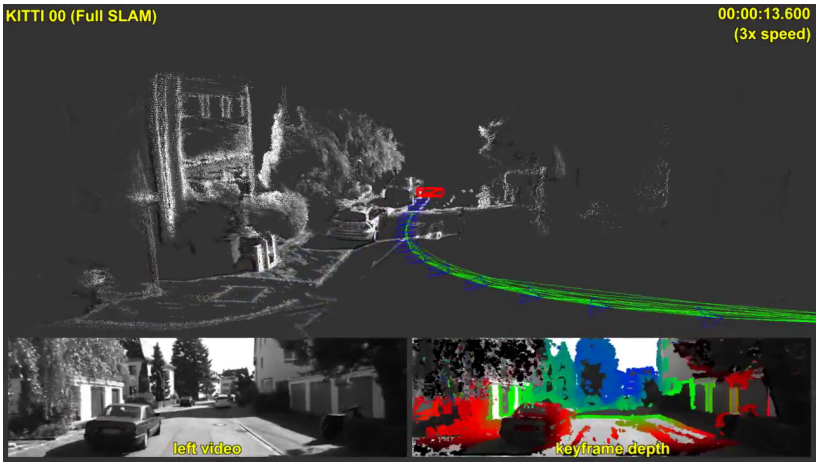


Figure 5.2: LSD-SLAM Example image
Source: `https://tinyurl.com/qkuamyy`

5.3 How does the LSD-SLAM work?

5.3.1 Components that make up the LSD-SLAM

- Tracker
  The *Tracking* component uses the current frame in relation to the last frame to continuously track the camera movement.
- Depth Map Estimation
  Depth map estimation is done using tracked frames to refine or replace the current frame. Depth information is calculated by filtering over a small per-pixel baseline. If the camera has moved too far or the image has changed too much a new keyframe is initialized [10].
- Map Optimization
  When a keyframe gets replaced as a tracking reference, the refinement process stops and it gets included in the 3D depth map. *Map optimization* then starts working to detect loop-closures or scale-drifts. This is done by a similarity transformation to frames that where taken nearby.

5.3.2 Depth Map Estimation

New keyframes are created when there where no frames before or the camera has moved/rotated so far that the set threshold has been exceeded. When this happens the new latest frame is chosen to become the new keyframe and the keypoints from the previous keyframe get projected onto the new one. The process is followed by scaling to fit the needs of the Direct Image Alignment. When that is done the keyframe replaces the previous ones and gets used to track the subsequent frames.

Not every frame results in a new keyframe. Frames that don't make it to a new keyframe are used to improve and refine the current keyframe. The refinement is done by using a small stereo comparison for regions in an image where the expected use for advancement is higher. The result of this comparison then gets merged into the existing 3D point cloud to add potentially new information or refining existing pixels [10].

5.3.3 Map optimization

Scaling, rotation and movement isn't always perfect, which results in drift. Even if the drift effect is little it adds up, which might result in some very off map [10]. The Pose-Graph-Optimization is a optimization algorithm which aims to fix these drifts with pretty good results. The advantages of the pose-graph-optimization are that it's fast and vulnerable for poor initialization estimates [20].

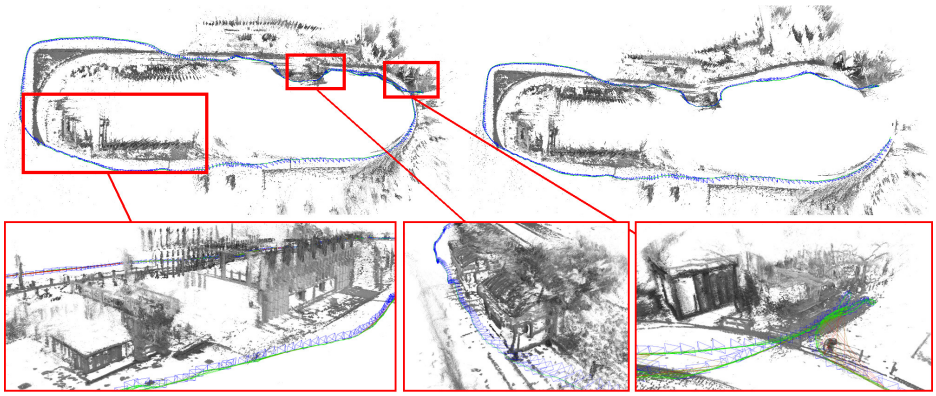


Figure 5.3: LSD-SLAM Pose-Graph-Optimization (left after loop-closure, right before loop-closure)
Source: `https://tinyurl.com/qkuamyy`

5.3.4 Input/Output

Input Data: rectified monocular image, camera info
Output Data: image with probability colored points, 3D point cloud

6 DeepTAM^{SM}

This Chapter is based on the work of:"DeepTAM: Deep Tracking and Mapping with Convolutional Neural Networks".[21]

6.1 What is DeepTAM?

DeepTAM provides a keyframe-based dense camera tracking and depth map estimation system that is entirely learned. The idea of DeepTAM is based on DTAM [22].The generic idea is: drift-free camera tracking via a dense depth map towards a keyframe and aggregation of depth over time. But the way to implement this concept is different. In the DeepTAM deep networks are used for tracking and mapping.These networks learn only from data. It also processes more than two images for the 6 DOF egomotion and depth estimation. With that it can avoid the drift of the use of keyframes and as more keyframes come in it can refine the depth map.

6.2 Tracking

The main objective is to estimate a 4 x 4 transformation matrix T. This matrix maps a point in the keyframe coordinate system to the coordinate system of the current camera frame. DeepTAM uses an more efficient way. It generates a virtual keyframe and tries to predict the increment instead of trying to estimate T. For more details see [21].

6.2.1 Network Architecture

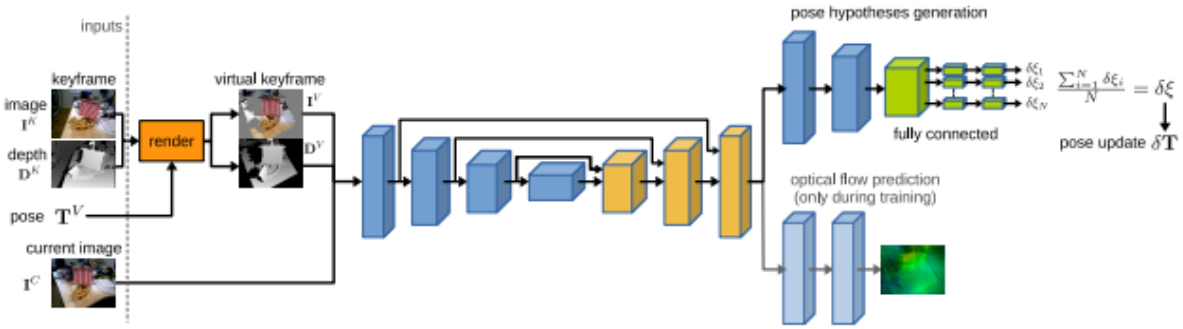


Figure 6.1: Schematic of DeepTAM
Source: http://lmb.informatik.uni-freiburg.de/Publications/2019/ZUB19a

To estimate the 6 DOF pose between a keyframe and an image the encoder-decoder-based architecture is used. To estimate camera motion you have to relate the keyframe to the current image. Because of that DeepTAM uses optical flow as an supportive task. With this optical flow the network is ensured to take advantage of the relationship between both frames. It uses two network branches for predicting the pose. One is the optical flow prediction and the other is the pose hypotheses generation. This improves the accuracy for the pose prediction.

6.3 Mapping

DeepTAM computes a set of depth maps every keyframe. For good quality depth maps, information will be accumulated in a cost volume. From this cost volume the depth map will be extracted by means of a convolutional neural network.

Normally the cost volume is taken as data term and because of that a depth map can be obtained by searching for the minimum cost. Using this method, because of the noise in the cost volume, there must be various optimization techniques and sophisticated regularization terms included to extract the depth in a robust manner.DeepTAM instead has a network which is trained to use the matching cost information in the cost volume and simultaneously combine it with the image-based scene priors to obtain more accurate and more robust depth estimates.

The accuracy is limited by the number of depth labels for cost-volume-based methods. That is why there is an adaptive narrow band strategy used to keep number of labels constant while increase the sampling density. The cost volume for the narrow band recomputes for a small selection of frames and searches again for a better depth estimate. The narrow band requires a good initialization and regularization to keep the band in the right place but allows to recover more details in the depth map.

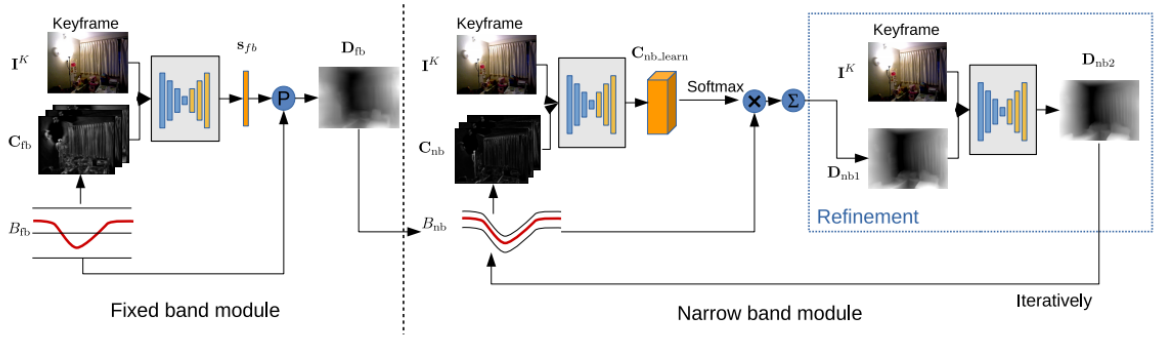


Figure 6.2: Mapping Networks Overview
Source: http://lmb.informatik.uni-freiburg.de/Publications/2019/ZUB19a

6.3.1 Network Architecture

The network is trained to predict the keyframe inverse depth from the keyframe image and the cost volume which is computed from a set of images and camera poses. The keyframe inverse depth is represented as inverse depth, which enables a more precise representation with closer distance. There is also a coarse-to-fine strategy along the depth axis applied. Mapping is divided into a fixed band module and a narrow band module. The narrow band cost volume centers at the current depth estimation and accumulates information in a small band close to the estimate, while the fixed band module builds a cost volume with depth labels evenly spaced in the whole depth range.

Between the minimum and maximum depth label the fixed band module regresses an interpolation factor as output. Because of this the network cannot reason about the absolute scale of the scenes, which makes the network more flexible and generalize better. The fixed band contains a set of fronto-parallel planes as depth labels. Conversely the narrow band contains discrete labels which are individual for each pixel.The prediction of interpolation factors is not useful since the network in the narrow band module has no knowledge of the band's shape. The narrow band is not provided with the band shape, because the network tends to ignore the cost information in the cost. This makes the depth regularization difficult. Therefore there is another network appended, which focuses on this problem.

6.3.2 Training

In about 8 days in total the training of the mapping network can be acomplished on a NVIDIA GTX 1080Ti.

7 Workflow

7.1 Used Hardware[AV]

For video capturing a *Raspberry Pi 3B+* with a *Pi Camera V2* or a *Pi Wide Angle Lens Camera* are used. The Raspberry Pi sends the video feed to a separate more power full PC over WiFi using a Python3 script.
For processing a *Lenovo Think-Station S20* or a *Lenovo W550s* are used depending on the amount of processing power is required. For more intense work a server access at the Johannes Kepler University was supplied to work on their system.
As the work is based around implementing it on the Audi Autonomous Driving Cup (AADC) car a remote controlled model car was borrowed for a few weeks.

7.2 Used Software[AV]

7.2.1 Raspberry Pi

The Raspberry Pi is running Raspbian Buster since it is well optimized for the mini computer and only required to be able to execute a Python script to send the raw video feed over http to the the processing device.

7.2.2 PC

The Think-Station and the laptop are running Kubuntu 18.04, which is basically Ubuntu but has a GUI that's a more like Windows and is supported until May 2023.
The Think-Station has a eight core Intel Xeon CPU, a GTX 1660TI and 12GB of RAM inside.
The Laptop has a four core Intel i7 and 8GB of RAM built in.

ADTF
At first Ubuntu 16.04 with Automotive Data and Time-Triggered Framework (ADTF) was used since it's the recommended environment by the AADC car manufacturer DigitalWerk. There were many compatibility ans stability issues and it is very difficult to get into the whole system as it's not very beginner friendly. After trying to get the basics of ADTF working it was clear that switching to ROS might be better. The main problems with ADTF are, that ADTF isn't running very stable, requires certain packages to be in a non-standard folders and not having them in the regular location and it is very difficult at the when using it for the first time.

ROS
Running ROS Melodic on Kubuntu 18.04 was pretty straight forward. The instructions on the ROS website are very clear and can be directly copied without issues. The principle of the workspace is also easy to understand. In the source folder the modules get put in and when compiling the modules automatically generates a setup file to use them. Usage is very easy as the framework already does a lot in the background and using nodes is nearly always setting input and output with a few parameters.

7.3 Setup[AV]

As the PC and laptop are not the best idea to run around with, a Raspberry Pi is used instead to stream the video over WiFi to the PC/laptop which are connected to the router over LAN. This makes the camera setup very portable as the pi, camera and powerbank are packed together and don't have much weight. The PC can sit somewhere where and just processing the received video signal.

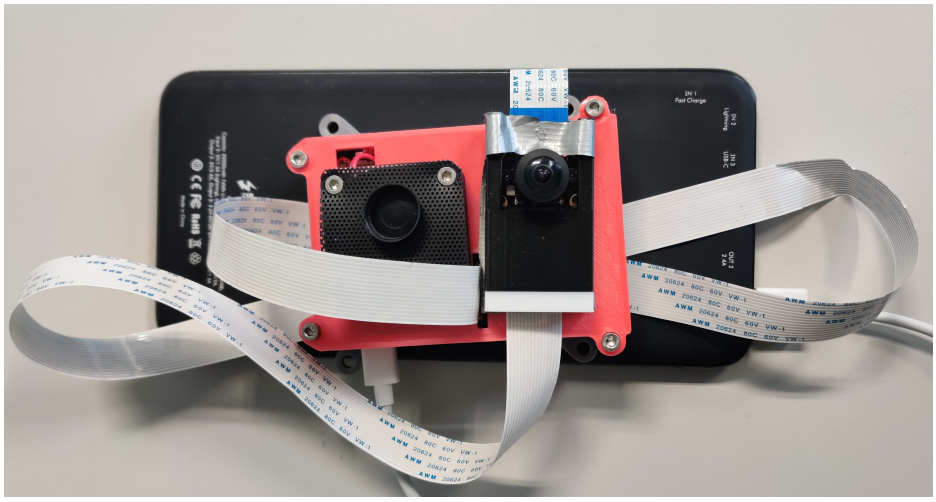


Figure 7.1: Stream setup with Raspberry Pi, camera and powerbank

7.4 Streaming video from Pi to PC[AV]

Enabling Camera
To use a camera on a Raspberry Pi the interface needs to be enabled first. This can be done in the built-in tool called *raspi-config*. In this tool under the subsection called *Interfacing Options* there is a option with the name *Camera*. When this is done the camera can be used after a restart.

Python Script
In the code snipped 7.1 at first a *piCamera* instance with the name *cam* is created. As parameters the resolution gets set to *1280x720* pixels and the frame rate is set to *30* frames per second (FPS). If needed the image can be rotated, e.g the camera is mounted upside down. When starting the camera a output and format are expected. For the output a

separate class is used which sets how and when a new frame can be published and for the format the *mjpg* video codec is chosen, as a pack for getting MPEG-streams already exists in ROS and it's not power hungry when running it on the Raspberry Pi.
After the camera "recording" has started successfully the server is started to make the stream accessible to other devices. The server runs until the user closes the script using *CTRL + C*. After closing the server the *finally* block gets called, where the camera "recording" is stopped so that other programs can use the camera again.

```
#Only resolution and framerate in Constructorparameters
with picamera.PiCamera(resolution=RESOLUTION, framerate=FPS) as cam:
    output = streamingOutput()
    # Rotation if needed
    cam.rotation = ROTATION
    #start stream
    cam.start_recording(output, format='mjpeg')

    try:
        #IP, Port
        hostAndPort = ('',PORT) # '' as IP automatically get's ip
        server = streamingServer(hostAndPort, streamingHandler)
        server.serve_forever() # Serves as long as script is running
    finally:
        cam.stop_recording() # Stops stream to make camera accessable from other
             apps again
```

Listing 7.1: Main Function of Camera Feed

The streamingHandler that is shown in snipped 7.2 handles the actions that are taken when client connects to the Raspberry Pi. At the beginning it checks if the client is requesting the */stream.mjpg* file. If the client is not requesting that specific file a *404 Not Found* Error is returned. But if the correct file is requested at first a *200 OK* code. In addition to the status code headers are send, which tell the client to not use cache. After sending the HTTP OK to the client a permanent loop is started which always waits until a new image from the camera is ready and then sends it to the client as an JPEG image. The loop ensures that the always client receives the latest image and so creates a video. Should the client drop the connection an exception is raised which causes the loop to stop and end the handler for that specific client until the client connects again.

```
class streamingHandler(server.BaseHTTPRequestHandler):
    def do_GET(self):
        # Check if user is requesting stream
        if self.path == '/stream.mjpg':
            # yes --> Send Stream
            # send header for beginning streaming
            self.send_response(200) # HTTP OK
            self.send_header('Age', 0)
            self.send_header('Cache-Control', 'no-cache, private')
            self.send_header('Pragma', 'no-cache')
            self.send_header('Content-Type', 'multipart/x-mixed-replace;
                boundary=FRAME')
            self.end_headers()

            # start Stream
            try:
                while True:
                    # Send new image when available
                    with output.condition:
```

```
                        output.condition.wait()
                        frame = output.frame
                    # Header for sending image
                    self.wfile.write(b'--FRAME\r\n')
                    self.send_header('Content-Type', 'image/jpeg')
                    self.send_header('Content-Length', len(frame))
                    self.end_headers()
                    self.wfile.write(frame)
                    self.wfile.write(b'\r\n')
            except Exception as e:
                print('Removed streaming client %s: %s', self.client_address,
                    str(e))
        else:
            # User requested something else --> Send 404 Page
            self.send_error(404)
            self.end_headers()
```

Listing 7.2: Streaming Handler of CamStream

The behavior when a new image from the pi camera is ready to send is defined in the snipped 7.3. At the beginning it initializes itself with basically no image. *piCamera* class constantly writes into this, as it's defined as the output. When a new image is ready the *picamera* class sends "\xff\xd8" as binary to the output class to notify it. The output class then cuts the buffer so it only contains the current image and sets it in the *frame* variable. To let everybody else know that a new image is ready it sends out a notification.

```
class streamingOutput(object):
    def __init__(self):
        self.frame = None
        self.buffer = io.BytesIO()
        self.condition = Condition()

    def write(self, buf):
        if buf.startswith(b'\xff\xd8'):
            # New frame available to get to buffer --> Notify all clients
            self.buffer.truncate()
            with self.condition:
                self.frame = self.buffer.getvalue()
                self.condition.notify_all()
            self.buffer.seek(0)
        return self.buffer.write(buf)
```

Listing 7.3: Streaming Output of CamStream

7.5 Receiving images on PC and Laptop[AV]

For receiving the images on the PC or Laptop an existing ROS-node is used. Video-Stream-OpenCV is designed to publish videos in the ROS network which are received from different sources, e.g. USB-cameras, video-files, network cameras and video-streams [23].

7.5.1 MJPG-Stream receiver

To automate the startup procedure of the node a roslaunch file is used to start the node is written for and automatically sets the required parameters.

The launchfile shown in 7.4 published the received image stream on a topic called *camera*. The video stream provider are the Raspberry Pi's IP-address, port and */stream.mjpg* directory. 30 FPS are used since they provide a good balance between amount of traffic and amount of detail in the movement.

```
<!-- launch video stream -->
<include file="$(find video_stream_opencv)/launch/camera.launch" >
    <!-- node name and ros graph name -->
    <arg name="camera_name" value="camera" />
    <!-- url of the video stream -->
    <arg name="video_stream_provider" value="http://192.168.2.109:5000/stream.
        mjpg" />
    <!-- set camera fps to (probably does nothing on a mjpeg stream)-->
    <arg name="set_camera_fps" value="30"/>
      <!-- set buffer queue size of frame capturing to -->
      <arg name="buffer_queue_size" value="100" />
    <!-- throttling the querying of frames to -->
    <arg name="fps" value="30" />
```

Listing 7.4: MJPG-Stream receiver Launch file

7.6 Cameras[AV]

Nearly every camera has some kind of distortion where the proportions of the image are different to the real world. This is especially noticeable on wide angle lenses which can capture a bigger part of the environment while sitting in the same spot. This can be seen in figure 7.2 that the normal camera only captures a small portion compared to the wide angle lens camera but the wide angle lens creates distortions when getting to the edges.

Figure 7.2: **Left Image:** normal Raspberry Pi Camera. **Right Image:** wide angle lens camera

7.6.1 Calibration

To get rid of the distortions on a wide angle lens camera calibration is needed, which is a mask that gets applied on the image to remove these distortions and rectify it. For calibration the *camera calibration*-node is used. Calibration is done by moving and rotating a checkerboard is used since it has good contrast between the tiles and the size of a tile is known and always

the same. The node recognizes the checkerboard and calculates the distortion-factors from a series of pictures that have been taken.
The command for starting the node is the following, where amount of tiles, size of tiles in millimeter and camera are set:

```
rosrun camera_calibration cameracalibrator.py --size 8x6 --square 0.026 --no-
    service-check image:=/camera/image_raw camera:=/camera
```

Listing 7.5: Start Calibration Node

This command opens a window which shows the live image feed from the camera and highlights edges on the checkerboard which is shown in figure 7.3. These highlighted edges are points which are used to calculate the distortion parameters using an algorithm that was developed by OpenCV [24].

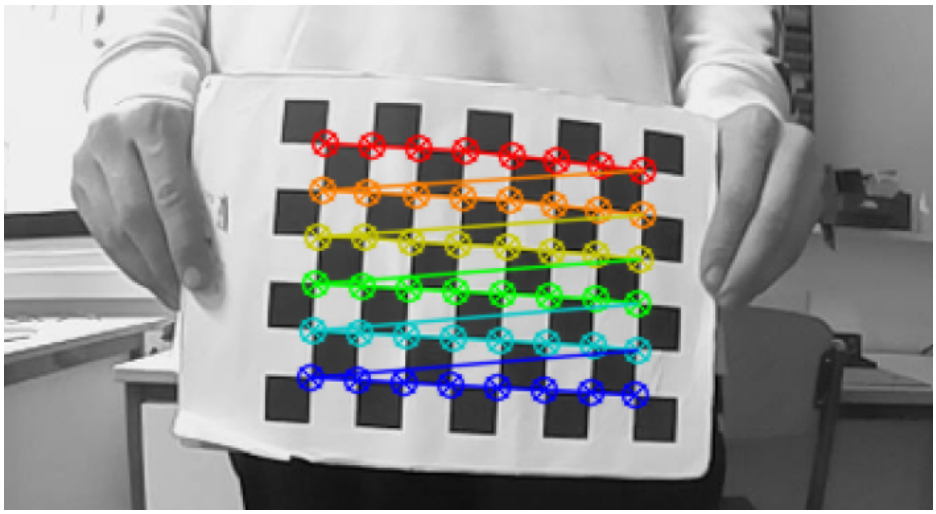


Figure 7.3: Camera calibration

When the node has enough reference images the computing of the parameters can start. The duration of the calculations depends on the CPU and how many images have been taken. But most likely it will take around 5 Minutes until the computing is finished. The output data contains two formats of the same data which will look something like shown in figure 7.6. The output files are normally compatible with all applications without any issues.

```
image_width: 1280
image_height: 720
camera_name: narrow_stereo
camera_matrix:
  rows: 3
  cols: 3
  data: [ 600.43227,    0.     ,  650.29343,
            0.     ,  605.883  ,  253.24984,
            0.     ,    0.     ,    1.     ]
camera_model: plumb_bob
distortion_coefficients:
  rows: 1
  cols: 5
  data: [-0.253310, 0.043398, 0.005101, -0.003291, 0.000000]
```

Listing 7.6: Calibration file

7.7 Integrating ORB-SLAM2

Integrating ORB-SLAM2 4 is pretty straight forward since there is already ROS version of it which is very well maintained by **appliedAI-Initiative** on GitHub [25].

8 Fazit und Persönliche Erfahrungen

8.1 Fazit

Zusammenfassung der Projektergebnisse. Besondere Erkenntnisse. Beurteilung des Lösungswegs. Eventuelle Alternativen und möglicher Erweiterungen.

8.2 Persönliche Erfahrungen

Hier (und nur hier) darf aus der Ich-Perspektive geschrieben werden.

Glossary

AADC Audi Autonomous Driving Cup. 19

ADTF Automotive Data and Time-Triggered Framework. 19

FPS Frames per Second. 20, 23

Lidar Laser radar. 1

OSRF Open Source Robotics Foundation. 3

RGB-D camera Camera which captures an image that contains depth information besides the normal image. 2

ROS Robot Operating System. 3–5, 21, 22, 25, 30

SLAM Simultaneous Localization and Mapping. 1, 2, 11, 13

Bibliography

[1] S. Riisgaard and M. R. Blas, "Slam for Dummies." https://tinyurl.com/y32jtecm.

[2] S. Prabhu, "Introduction to slam (simultaneous localisation and mapping)." ARreverie, 2019. https://tinyurl.com/y5an9jq9.

[3] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (slam): part ii," *IEEE Robotics Automation Magazine*, vol. 13, pp. 108–117, Sep. 2006.

[4] S. Martin, "What is simultaneous localization and mapping?." Techapeek, 2019. https://tinyurl.com/y5yajv5u.

[5] C. Stachniss, U. Frese, and G. Grisetti, "OpenSLAM Website." Openslam, 2019. https://openslam-org.github.io/.

[6] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "Fastslam: A factored solution to the simultaneous localization and mapping problem," in *In Proceedings of the AAAI National Conference on Artificial Intelligence*, pp. 593–598, AAAI, 2002.

[7] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, pp. 1147–1163, Oct 2015.

[8] F. Steinbruecker, J. Sturm, and D. Cremers, "Real-time visual odometry from dense rgb-d images," in *Workshop on Live Dense Reconstruction with Moving Cameras at the Intl. Conf. on Computer Vision (ICCV)*, 2011.

[9] F. Endrees, J. Hess, and N. Engelhard, "Rgb-d slam." ROS Wiki, 2019. https://tinyurl.com/yynyqwg2.

[10] J. Engel, T. Schops, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European Conference on Computer Vision (ECCV)*, September 2014. https://vision.in.tum.de/research/vslam/lsdslam?redirect=1.

[11] O. Foundation, "Osr foundation homepage." Website, 2019. https://www.osrfoundation.org/.

[12] ROS, "About ros." ROS, 2019. https://www.ros.org/about-ros/.

[13] C. Robotics, "Ros 101: Intro to the robot operating system." Robohub, 2014. https://tinyurl.com/y4q6paak.

[14] ROS, "Is ros for me?," 2020. https://www.ros.org/is-ros-for-me/.

[15] M. H. Hassoun, *Fundamentals of Artifical Neural Networks*. The MIT Press, 1995.

[16] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019.

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, Nov 1998.

[18] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, pp. 1255–1262, Oct 2017.

[19] E. Weisstein, "Levenberg-marquardt method," 2020. `http://mathworuld.wolfram.com/Levenberg-MarquardtMethod.html`.

[20] E. Olson, J. Leonard, and S. Teller, "Fast iterative optimization of pose graphs with poor initial estimates," pp. 2262–2269, 2006.

[21] H. Zhou, B. Ummenhofer, and T.Brox, "Deeptam: Deep tracking and mapping with convolutional neural networks," *International Journal of Computer Vision*, 2019.

[22] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *2011 International Conference on Computer Vision*, pp. 2320–2327, Nov 2011.

[23] R. D. Group, "video stream opencv," 2020. `https://github.com/ros-drivers/video_stream_opencv`.

[24] O. D. Team, "Camera calibration with opencv," 2020. `https://docs.opencv.org/2.4/doc/tutorials/calib3d/camera_calibration/camera_calibration.html`.

[25] appliedAI Initiative, "Orb-slam 2 ros," 2020. `https://github.com/appliedAI-Initiative/orb_slam_2_ros`.

List of Figures

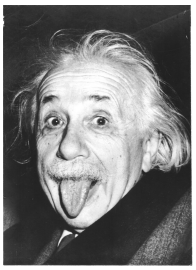
Listings

Authors

Alexander Voglsperger

Birthday, Place of birth:	25.03.2001, Ried im Innkreis
School education:	Volksschule Aurolzmünster Informatik Hauptschule Aurolzmünster HTL Braunau
Internship:	Team7 Natürlich Wohnen GmbH, 4 Weeks, IT Krankenhaus Ried im Innkreis, 4 Weeks, IT Johannes Kepler University - AI Lab, 4 Weeks, Mapping and Tracking on self-driving car
Address:	Forchtenau 196 4971, Aurolzmünster Österreich
E-Mail:	alexander.voglsperger@gmail.com

Simon Moharitsch



Birthday, Place of birth:	01.01.1970, Braunau am Inn
School education:	Volksschule Hauptschule HTL
Internship:	Firmenname, Zeit, Tätigkeit
Address:	Strasse Nummer PLZ, Ort Österreich
E-Mail:	max@mustermann.com