ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

RSL Robotic Systems Lab
Prof. Dr. Marco Hutter

Master Thesis

Autonomous Vision-based Safe Proximity Operation of a Future Mars Rotorcraft

Autumn Term 2024

Supervised by:
Roland Brockers
Robert Hewitt
Marco Hutter

Author:
Lasse Fierz

Contents

List of Acronyms

- **UAV:** Unmanned Aerial Vehicle
- **SFM:** Structure From Motion
- **LSD:** Landing Site Detection
- **LS:** Landing Site
- **BA:** Bundle Adjustment
- **DEM:** Dense Elevation Map
- **OMG:** Optimal Mixture of Gaussian
- **LOD:** Level Of Detail
- **HiRISE:** High Resolution Imaging Science Experiment
  (High Resolution Satellite Imagery on the Mars Reconnaissance Orbiter (MRO))
- **LRF:** Laser Range Finder
- **GT:** Ground Truth
- **LSM:** Landing Site Manager
- **FSM:** Finite State Machine
- **BT:** Behavior Tree

Abstract

An autonomous rotorcraft literally stands or falls on its reliable landing capabilities. When that same rotorcraft is on Mars, this procedure cannot fail even once. The LORNA (Long Range Navigation) project tackles this problem by introducing a Landing Site Detection (LSD) mechanism which aggregates Structure From Motion (SFM) point clouds into a multi resolution elevation map and performs landing site segmentation on the collected depth information. In this master's thesis we incorporated this landing site detection pipeline into an autonomy framework and implemented a behavior tree based landing mechanism to safely and efficiently select, verify and discard detected landing sites. Furthermore, the pipeline was enhanced using a stereo camera depth input alternative to SFM for lower altitudes to remove the necessity of lateral motion in order to perceive depth. The software was tested extensively in a gazebo simulation on different synthetic as well as recorded environments and different behaviors were considered and analyzed throughout various Monte Carlo iterations. The contributions in this work aim at enabling future Mars rotorcrafts to autonomously and reliably land at safe locations thus enabling a more daring aerial exploration of the red planet.

Chapter 1

Introduction

With the Ingenuity rotorcraft's lifecycle coming to an end, the question about future Mars rotorcrafts and their capabilities draws ever closer.
For the future NASA develops two different rotorcraft Mars concepts.

- Mars sample return helicopter concept

  The first is associated with the Mars Sample Return mission in which NASA's Perseverance rover collects Mars rock and sand samples in test tubes. These tubes are subsequently brought back to a sample retrieval landing platform from which a small rocket (Mars Ascent Vehicle) launches them into Mars orbit. There, ESA's Earth Return Orbiter will enclose them in a highly secure containment capsule and deliver them to Earth.

  In case that the Perseverance rover won't be able to collect the test samples and deliver them to the lander, a Mars Sample Return helicopter concept is envisioned. This is a rotorcraft for low altitudes equipped with a mechanical gripper in order to offer an alternative way of transporting Mars samples to the retrieval station should the Perseverance rover fail to do so.

- Mars Science Helicopter (MSH)

  Secondly, for future exploratory large distance missions, NASA is conceptualizing a Mars Science Helicopter (MSH) project. The aspirations for such a rotorcraft are on one hand to cover farther distances at high altitudes with accurate state estimation and on the other to land safely, autonomously and reliably in previously unknown terrain. These two feats allow a helicopter to perform much advanced science missions compared to Ingenuity.

The Long Range Navigation (LORNA) project that I have been involved with is working on a concept to tackle the second project's challenges while dealing with the constraints that rotorcraft missions on Mars provide us with. These are namely a limitation on the size and weight of the drone, a constraint on computational power due to the deployment on limited embedded processors and lastly a large delay in communication which makes adaptive remote control from Earth impossible.

1.1 Objective

The conclusive high level objective of the LORNA science concept is the achievement of long range safe navigation including global localization, safe landing site detection and full system autonomy. The navigation endeavor is tackled using a laser-range-finder augmented visual-inertial odometry state estimator which uses

map based localization to achieve global localization. Landing site acquisition is achieved using a structure from motion based 3D reconstruction of the terrain which is fed into the creation of a multi resolution dense elevation map used for landing zone segmentation. Finally, a state machine-based autonomy framework orchestrates the entire procedural workflow.
The endeavor in this thesis was to create a front to back landing mechanism that combines the existing vision based landing site detection algorithm with the autonomy framework. In order to accomplish this, both the landing site detection algorithm and the autonomy had to be altered. Last but not least, given that the structure from motion depth generation depends on lateral movement, which is less desirable for a drone navigating at low altitudes in unfamiliar surroundings, the utilization of a stereo camera for low altitude 3D reconstruction presents a viable solution to attain real-time depth perception without necessitating lateral displacement. A stereo camera is thus a light-weight solution which allows a drone to perceive depth statically as well as in vertical and lateral motion.

1.2 My Contribution

In this work, I established the interface between the vision based landing site detection algorithm and the autonomy framework in order to make informed landing decisions based on detected landing sites. A safe and efficient landing mechanism was implemented in the existing autonomy. This mechanism utilizes a novel stereo camera 3D reconstruction procedure to avoid lateral motion at low altitudes.

- **Stereo Camera Depth**

  A stereo camera was implemented in the simulated drone model in order to get stereo camera images. Additionally, a stereo camera depth node was put in place as an augmentation of SFM to supply the landing site detection algorithm with a point cloud at low altitudes without the need for lateral motion.

  An automatic switch was inserted between the SFM node and the stereo camera depth node by utilizing the already present laser range finder sensor on board. This allows for minimal computational overhead as only one depth creation node runs at a time.

- **Ground Truth**

  A ground truth depth node was created for two reasons. First it allowed the validation of the stereo camera depth output. Second, having a perfect point cloud of the terrain, specific testing of the autonomous landing behavior itself was possible.

- **Autonomy LSD Interface and Landing Site Handling**

  The landing site detection output initially only consisted of a one single site's location. This output was enhanced to utilize many more characteristics already present in the landing site detection algorithm in order for the autonomy to make a more informed decision in regard to what spot to select. These landing site properties are

  - Terrain roughness
  - Size
  - Terrain uncertainty
  - Detection altitude

  - Obstacle height

  The autonomy framework was expanded to correctly receive and sort incoming landing sites based on their individual heuristic score. Additional landing site handling mechanisms like re-detection, verification and banishment were introduced.

- **Behavior Tree for Adaptive Decisions**

  Using the existing behavior tree structure from the autonomy, an adaptive landing procedure consisting of both existing and novel actions was implemented. The implementation of the landing behavior optimized for both safety and efficiency.

- **Simulation Setup**

  As just recently the switch was made to Gazebo Garden the entire visual pipeline (SFM + LSD) had never run with this simulation environment before. Therefore, I implemented the changes necessary to run the landing site detection procedure on the Gazebo sensor input.

- **Deployment of LSD Pipeline onto an Embedded Processor**

  Currently, the used processor on the drone is modalAI's voxl2. Both the structure from motion and the landing site detection software did not run out of the box having an incompatible dependency handling with the voxl's AARCH architecture. Resolving these dependency issues, I was able to run the landing site detection pipeline with the structure from motion depth supply on the voxl2 using a collected rosbag of images and respective poses from the xVIO state estimator.

1.3 Organization of this Thesis

- **Related Work**

  As is custom, I will introduce the reader to what has been done in this area. Main focus will be placed on vision-based landing site detection procedures and previous work on autonomous landing.

- **System Overview**

  The entire project architecture will be outlined. Emphasis lies on the methods that I have heavily interacted with in this thesis. These are mainly the structure from motion depth generation, the landing site detection mechanism and the autonomy framework.

- **Methodology**

  Here I conceptually lay out the high level structure of the implemented work in this thesis. The two key contributions stereo depth and autonomous landing are introduced as well as the ground truth used.

  I will go into the reasoning of why a stereo camera is necessary as a low altitude depth alternative. Additionally, I analyze the stereo option theoretically and conclude its usage domain.

  I explain the ground truth depth source used in this work both in order to compare stereo with and to test the autonomous pipeline without the possibility of insufficient depth information. Additionally, I analyze the ground truth's comparability to SFM to ensure adequate testing of the autonomous landing pipeline.

Lastly, I outline the prerequisites for the implementation of the autonomous landing behavior and introduce the methodology of the final implementation in the behavior tree structure of the autonomy.

- **Stereo Camera Depth Alternative**

  This chapter elaborates on the stereo camera depth implementation. A coordinate frame overview is given and the entire process from sensor data handling to point cloud generation is discussed. Lastly it is qualitatively compared to a depth camera based ground truth.

- **Autonomous Landing Procedure**

  Here I will lay out the core contribution of this project which combines the existing system with the novel contributions of this work in order to put together a front to back autonomous landing procedure. First, I describe the interface between the autonomy and the landing site detection pipeline. Then I introduce the conceptual implementation of the landing procedure before I show its implementation in the form of a set of actions structured in a behavior tree. Lastly the working pipeline is shown in a case example of a science mission flown in simulation.

- **Evaluation**

  Here I introduce the test setup according to which I performed repeated randomized simulation flights. I introduce the outcome defining metrics and the results of the test flights. Lastly these results are analyzed numerically as well as visually considering the final choices of landing sites.

- **Conclusion**

  I summarize the novel contributions of this work and conclusively assess the characteristics and quality of the final landing pipeline. Shortcomings of the approach are pointed out and remedies are discussed.

- **Outlook**

  Further enhancements of the current systems are laid out and alternatives for future iterations are discussed. Also, emphasis is placed on current insufficiencies and the necessity of resolving them.

Chapter 2

Stereo Camera Depth Node Implementation

2.1 Theoretical Analysis

When it comes to depth perception the obvious drawback of a stereo camera is its limited baseline. It only perceives depth accurately for objects within a certain proximity to the lens.
Assuming a perfectly calibrated and rectified camera there is still always an inaccuracy in the depth estimation arising from the disparity error.
The depth error is estimated using the following derivation. The formula to derive a depth value from a calculated disparity is:

$$z = \frac{f \cdot b}{d} \tag{2.1}$$

Where b is the z is the depth estimate, b is the baseline, f is the focal length and d is the disparity value.
Taking the derivative of z w.r.t. d we get

$$\frac{\partial z}{\partial d} = -\frac{f \cdot b}{z^2} \tag{2.2}$$

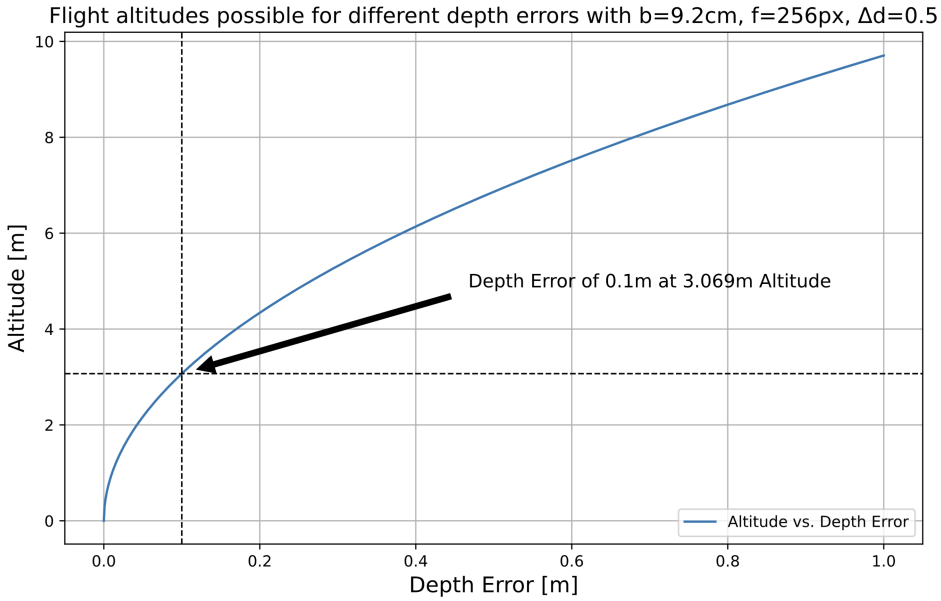
And substituting (eq. (1.1)) we get:

$$\partial z = \frac{z^2}{f \cdot b} \partial d \tag{2.3}$$

Where the sign was left away as for our application there lies equal danger in a point being perceived too close and too far away.
For the maximum altitude given a maximum allowable depth error this yields:

$$z_{\max} = \sqrt{\frac{\Delta z_{\max} \cdot b \cdot f}{\Delta d}} \tag{2.4}$$

Where Δz is the depth error and Δd the disparity error.
The stereo camera mounted on the drone in JPL's aerial vehicle lab had a baseline of about 10 cm and a focal length of 256.
With these properties and estimating a subpixel-precision disparity error of 0.5 pixels the depth error at varying altitudes looks as follows:
Let's assume we allow a maximum depth error of 10 cm. Considering this constraint we can fly at a maximum altitude of about 3 m as indicated in section 1.1.



This limitation has to be kept in mind. However, it is neither too surprising nor is it too restrictive as the stereo camera is simply a depth alternative for low altitude flight maneuvers. In the context of an entire science mission it is almost exclusively used for landing site verification purposes.

2.2 Implementation

Like Structure from Motion, the stereo depth instance is a ROS node which is given images and image poses from the xVIO state estimator. As the state estimator was in its final development stages during my thesis, camera images and a ground truth camera pose from the simulation were used instead as input for the stereo algorithm. Note that only one camera pose is given as the second one is derived in a straight forward manner, given the fixed hardware baseline.

2.2.1 Stereo Setup Overview

fig. 1.1 shows the drone setup in the simulation with the stereo camera. The stereo camera pair is indicated with the opaque boxes. The significant distance to the drone's core is necessary to avoid capturing the landing feet in the image due to the simulation model's discrepancy to the physical drone. As presented in **??** the drone hardware has landing skids that are spread significantly farther apart than for the simulated model. That is why, when using the stereo camera mounted at the core of the physical drone, the mainstays are not visible in the images detected. In the simulation they would be detected unless the stereo cameras are positioned further from the rotorcraft's center.

Frames

A critical part of navigation is always the consistency of the coordinate systems in which quantities are represented. Hereafter in fig. 1.2 the present coordinate systems of the stereo camera setup are displayed.
Notably there are three important frames:

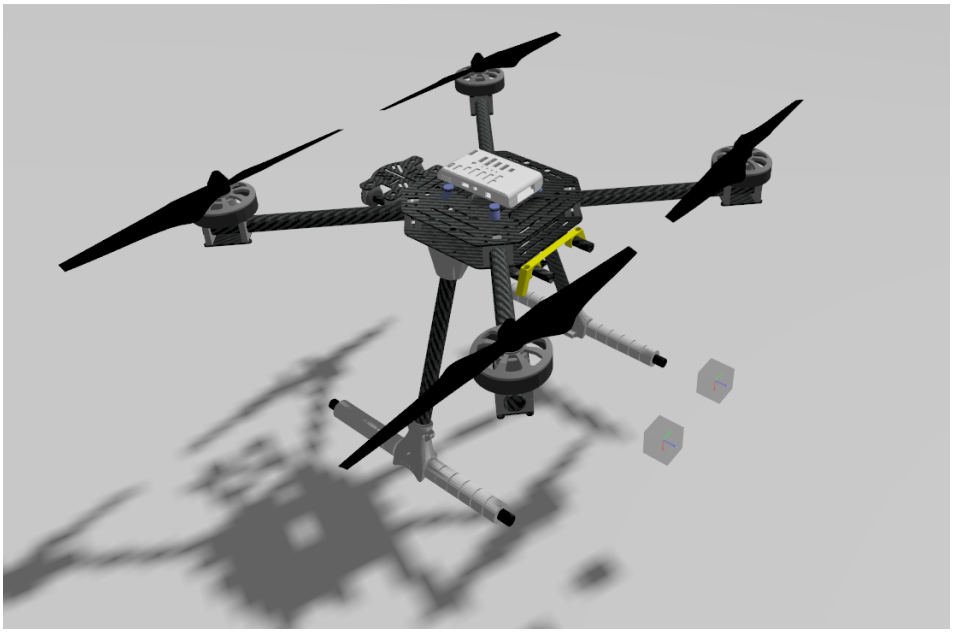


Figure 2.1: Stereo camera on drone indicated by opaque boxes

- The reference world frame W

  This is the global frame relative to which the drone flies and the point clouds are created.

- The drone (base link frame) D

  This is the pose of the moving drone throughout a mission. It is constantly published by the simulation.

- The camera pose C

  The camera pose is the frame in which the point cloud is created and aggregated. The relative pose of the cameras are set in the simulated drone's setup file. This transformation is static and is applied to each incoming pose message directly.

Hereafter, the following notation is used:

- t_{dc}^{W}: Transformation from drone to camera, represented in the World frame
- R_{DC}: Active Rotation from the Drone to the Camera frame
- r_{wp}^{W}: Position vector from the world frame to the detected point, represented in the World frame.

One more thing to note in fig. 1.1 is that in Gazebo's camera convention, the optical axis of a camera points along the x-axis. Therefore, the base link pose was subscribed to and the pose was converted to the camera pose using the adequate transformation. Neither Gazebo's base link nor camera coordinate conventions are relevant as long as we correctly track the pose of a downwards facing camera.

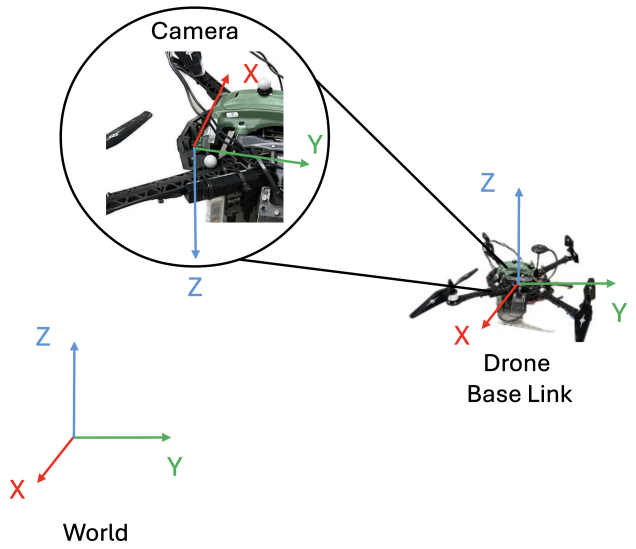


Figure 2.2: Coordinate frames in stereo camera depth setup

2.2.2 Input Handling

The input of the images as well as the base link pose are received using ROS subscribers. Despite publishing these simulated Gazebo topics at equal rates however, they did not arrive simultaneously. Note that this would be resolved when working with the actual xVIO state estimator as it uses the tracking camera's image and supplies further nodes with that same image as well as the pose with synchronized timestamps.
The pose is only required for the transfer of the created points from the camera frame to the world frame. Therefore, the two input images were processed into a depth image in a single step and only after were all three inputs used in order to create the point cloud.
fig. 1.3 schematically shows how the stereo camera depth node handled this shortcoming of asynchronous sensor messages by manually picking the pose which temporally closest corresponds to the image's timestamp. This was possible as the image processing step dominates the computation time of the input handling and the pose can thus be continuously updated in the meantime to best fit the images.

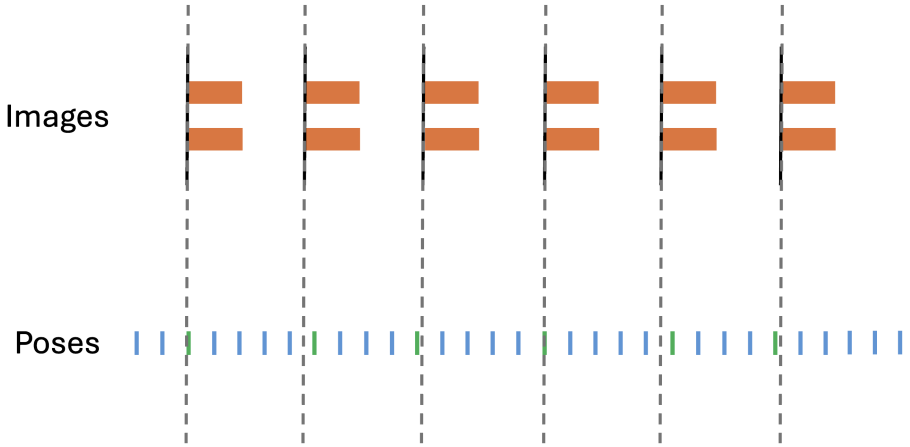


Figure 2.3: Schematic visualization of input synchronization used in the stereo depth node

2.2.3 Disparity Creation

The chosen StereoSGBM (Semi-Global Block Matching) algorithm creates a disparity map from two horizontally aligned images using an approach based on Heiko Hirschmüller's stereo approach introduced in [**?**].
The resulting disparity image is shown in **??**.

Figure 2.4: Disparity image from the stereo camera using StereoSGBM.

The StereoSGBM algorithm cuts off the left boundary as can be seen in **??**. This is due to the maximum disparity parameter which prevents the pixels on the left border of the left image to be matched. As the left image is the reference image, the border is removed in the output. Additionally, artifacts could be seen on the right-hand side of the image. Both of these border artifacts were neglected using a mask.

2.2.4 Point Cloud Creation

Having created a disparity image using the approach laid out in section 1.2.3, the disparity pixels are first converted to the depth values using the classic disparity depth relation 1.1 shown in section 1.1.
From the created depth image, the pose and the camera parameters, the 3D locations of each detected point can be derived. This is done by simply tracing the projection line of the detected point as indicated in fig. 1.4 and using the similar triangles as shown in fig. 1.5
The formulas for the point coordinates in the camera frame are shown in eq. (1.5).

$$x_{cp}^{C} = \text{depth} * \frac{(u - o_x)}{f_x} \tag{2.5}$$

$$y_{cp}^{C} = \text{depth} * \frac{(v - o_y)}{f_y} \tag{2.6}$$

$$z_{cp}^{C} = \text{depth} \quad \text{Depth value is the same} \tag{2.7}$$

Figure 2.5: Stereo depth image

Lastly, to get the point cloud in the world frame, the coordinate transform is applied:

$$r_{wp}^{W} = r_{wc}^{W} + R_{WC} \cdot r_{cp}^{C} \tag{2.8}$$
$$= r_{wd}^{W} + R_{WD} \cdot r_{dc}^{D} + R_{WC} \cdot r_{cp}^{C} \tag{2.9}$$

Where

$$r_{cp}^{C} = \begin{pmatrix} x_{cp}^{C} \\ y_{cp}^{C} \\ z_{cp}^{C} \end{pmatrix} \tag{2.10}$$

The final point cloud created using this process is shown in **??**.
The final output of the node is a generated point cloud in the world frame together with two poses representing the camera locations of the generated point cloud.

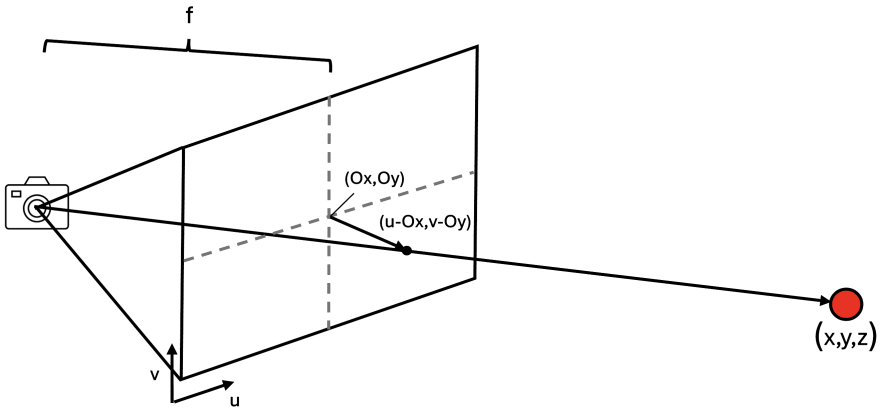


Figure 2.6: Schematic of the line projection procedure to derive the 3D location of detected points

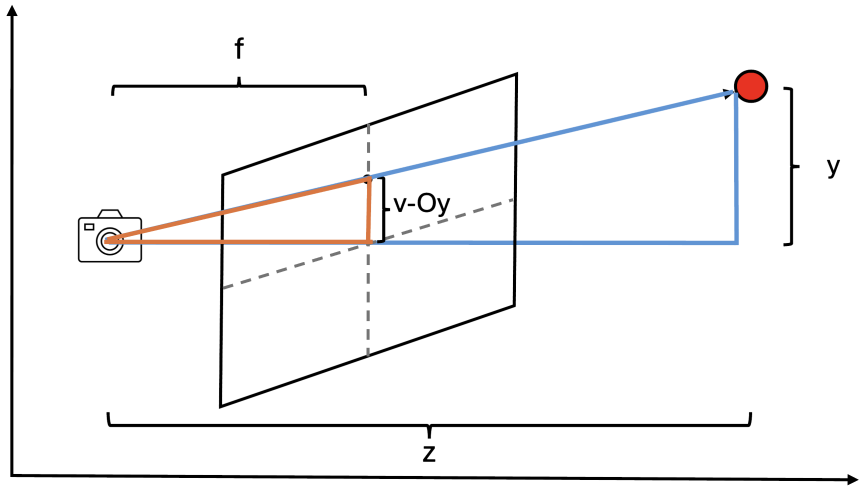


Figure 2.7: Schematic of the line projection procedure to derive the 3D location of detected points

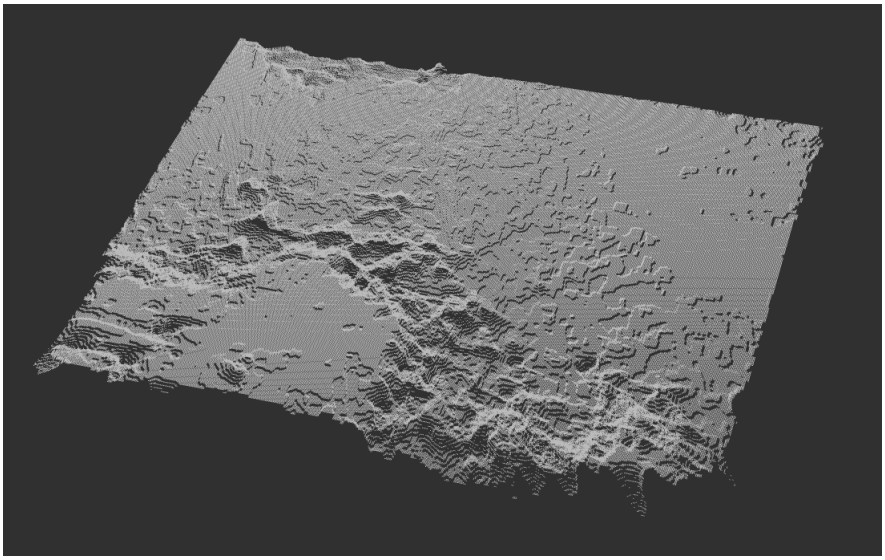


Figure 2.8: RViz visualization of stereo camera point cloud

StereoSGBM vs StereoBM

Apart from the StereoSGBM implementation, OpenCV provides the StereoBM algorithm which in general is faster but less precise.
Implementing and comparing the two algorithms the following results were seen:

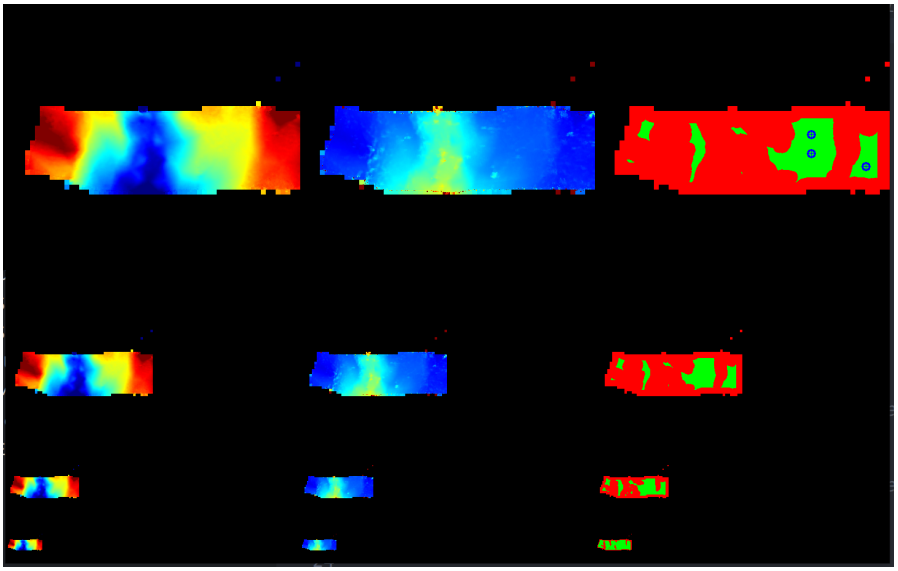


Figure 2.9: LSD debug image when using the StereoBM algorithm

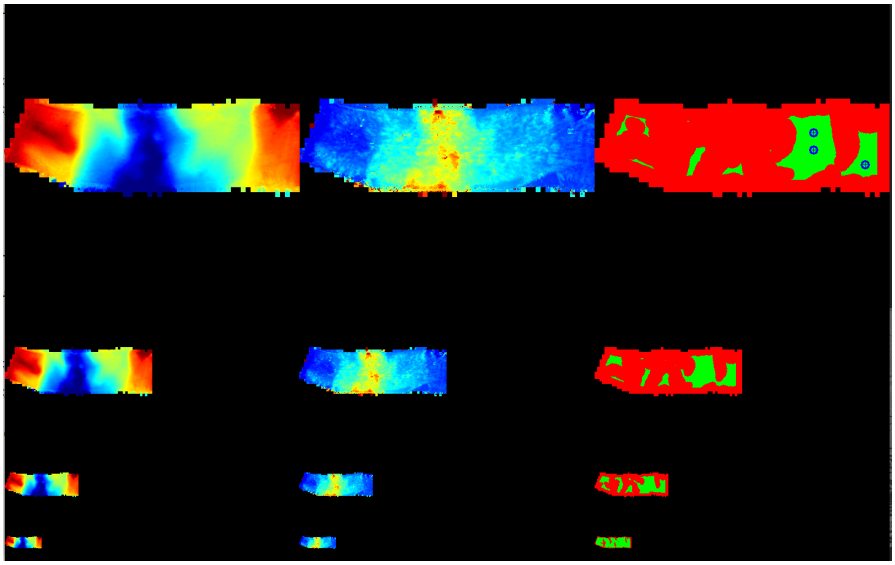


Figure 2.10: LSD debug image when using the StereoSGBM algorithm

As can be seen in the figures **??** and **??**, StereoBM did not in general perform bad. However, quite regularly, StereoBM produced erroneous points. In general, a swift algorithm like StereoBM is preferable for our purposes. Nevertheless, precise terrain reconstruction is a vital component of the pipeline, and thus, the quality must not be compromised. Therefore, StereoSGBM was retained as the algorithm of choice for the stereo depth creation.

2.2.5 Switching

In order to achieve the final desired perception mechanism of flying laterally with SFM and using a stereo camera depth node at low altitudes, one needs to switch between the two alternatives.
The obvious decision to use in the switching mechanism is the drone's current altitude above ground. This could be achieved by analyzing the generated point cloud at a given iteration to determine the median altitude which indicates the altitude above ground. This however is avoidable computational overhead.
As mentioned in **??** the drone has a laser range finder on board. This allows us to get an estimate of the altitude above ground at any given moment without the need for image processing.
Therefore, the switching is performed by using a separate ROS subscriber which continuously checks the LRF's measurement and activates or deactivates the SFM node and stereo node respectively.

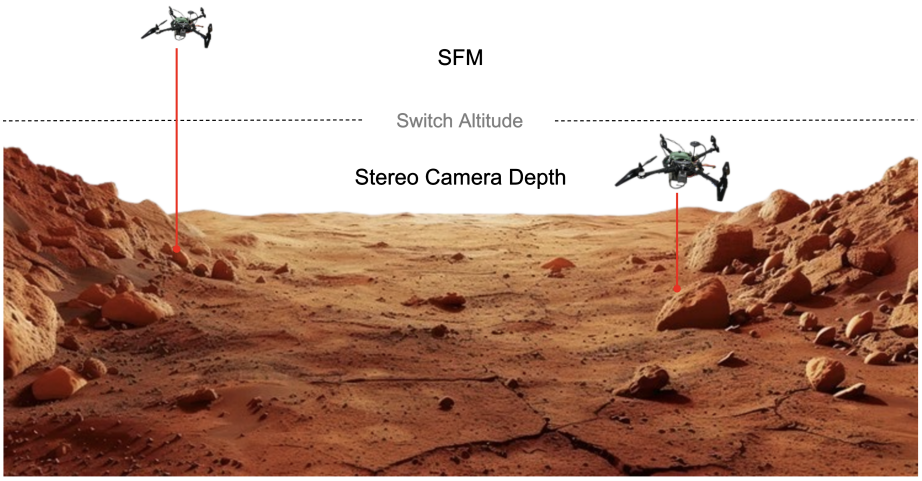


Figure 2.11: Laser Ranger Finder Based Switch between Depth Sources

2.2.6 Landing Site Detection without Lateral Motion

Taking off vertically with the drone in the simulation, the first landing site without lateral motion was found.

Figure 2.12: Drone during vertical ascent in simulation

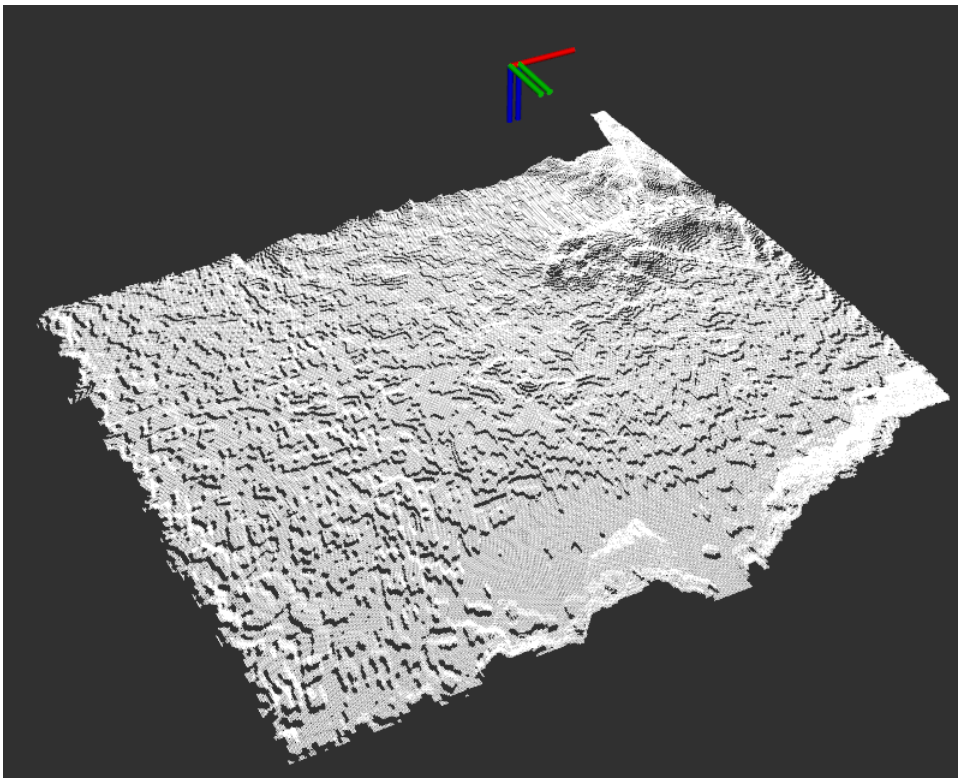


Figure 2.13: RViz visualization of created point cloud from stereo camera

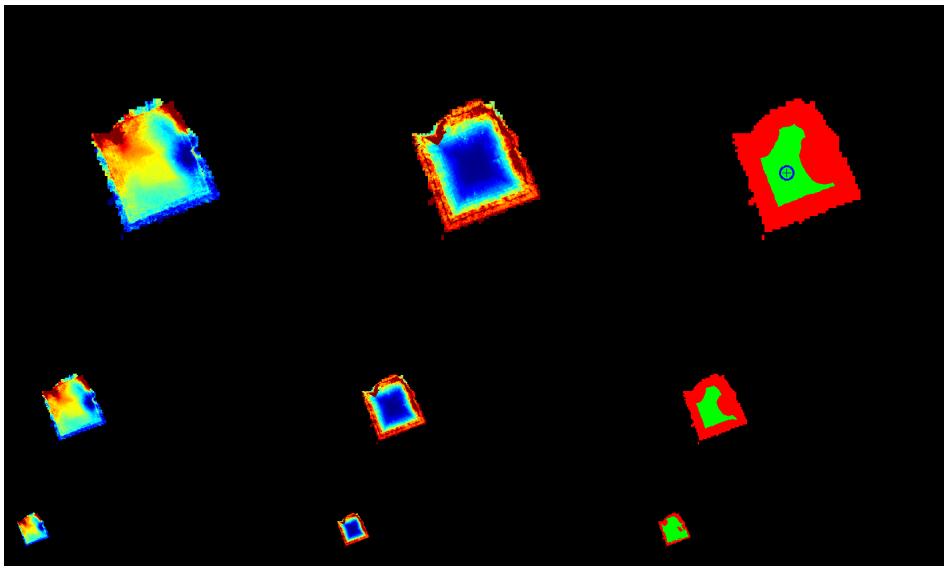


Figure 2.14: LSD Debug output displaying LORNA's first detected landing site during vertical motion

2.3 Qualitative Practical Analysis

Once implemented the landing site detection instance could be supplied by the stereo depth node. The result thereof can be seen below:

Figure 2.15: Considered terrain patch in Gazebo simulation

When comparing the result to the ground truth LSD output it can be seen that LSD creates a very accurate DEM from the stereo camera depth input. The landing sites detected are reasonable when compared to the terrain reference.

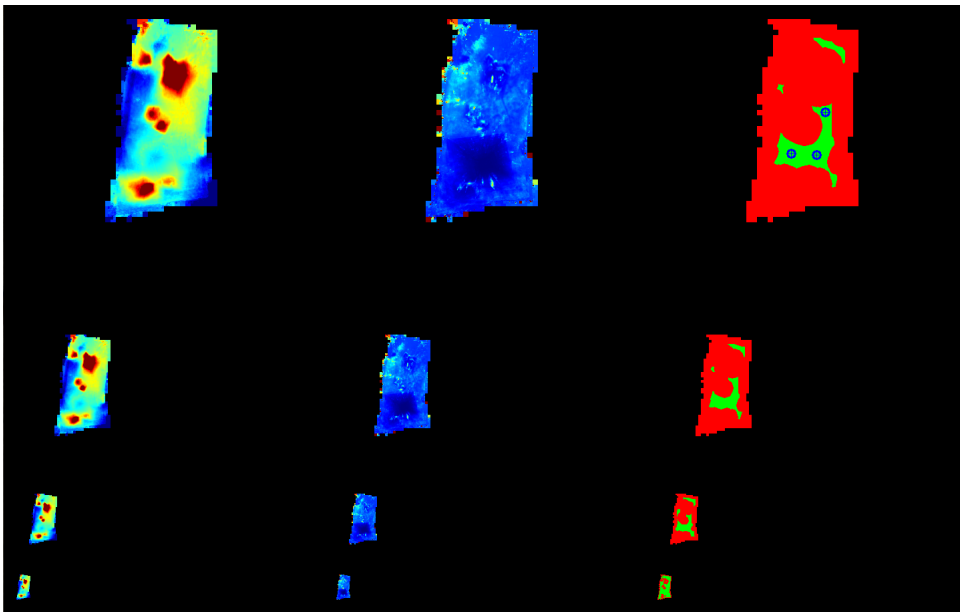


Figure 2.16: Stereo camera depth supplied LSD debug image at 2.5 m altitude

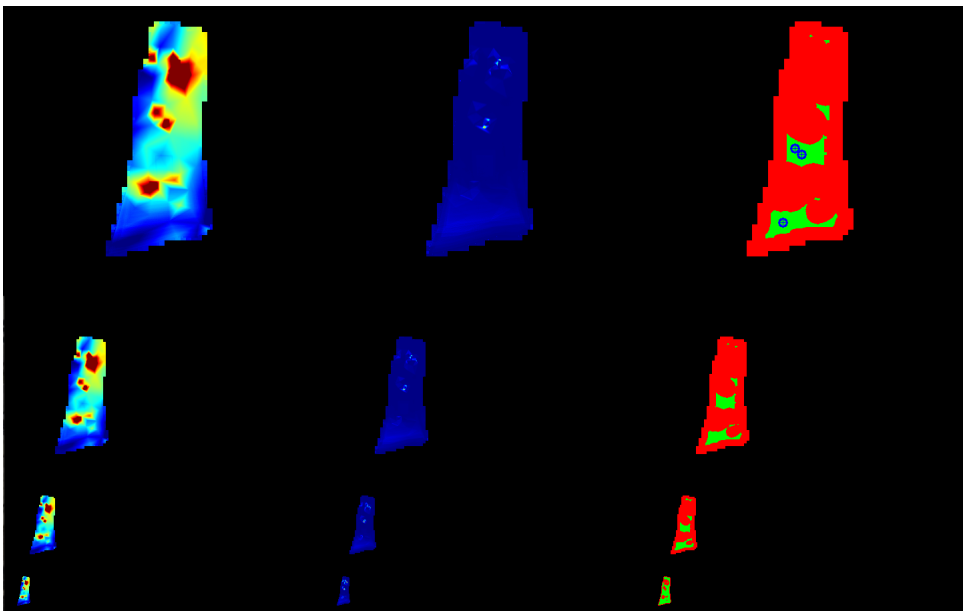


Figure 2.17: GT depth supplied LSD debug image at 2.5 m altitude

As the stereo camera has a relatively small, fixed baseline. The usage domain is restricted to low altitudes. The residual part of a mission is still flown using SFM.

Bibliography

[1] M. Vukobratović and B. Borovac, "Zero-moment point — thirty five years of its life," *International Journal of Humanoid Robotics*, vol. 1, no. 01, pp. 157–173, 2004.

[2] M. Raibert, *Legged Robots That Balance*. Cambridge, MA: MIT Press, 1986.

[3] G. A. Pratt and M. M. Williamson, "Series elastic actuators," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1995, pp. 3137–3181.