Visual Simultaneous Localisation And Mapping:Challenges and Improvements

Nhlalala Maluleke

Supervisor(s):
Dr Ritesh Ajoodha

A research report submitted in partial fulfillment of the requirements for the degree of Computer Science (BScHons)

in the

Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg

28 November 2021

Declaration

I, Nhlalala Maluleke, declare that this research report is my own, unaided work. It is being submitted for the degree of Computer Science (BScHons) at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

Nhlalala Maluleke

28 November 2021

Abstract

The recent discovery of SLAM algorithms has propelled the world of computer vision and robotics to newer territories which has helped the technological world we live in today. SLAM solves the problem of mapping and localising a robot(agent) in an unknown area based on sensor information. There are many different kinds of implementations of SLAM which require a scanner mounted on an agent which is able to make sense of its environment. There are many ways in which an agent can acquire scanned information, like using a depth camera(e.g RealSense, Xbox'x Kinect , Xtion, etc.) and using laser-based sensors (e.g LDS,LiDAR,LRF). In this paper visual SLAM is explored using a monocular camera to scan the environment. Because using a monocamera is much cheaper than using a laser-based sensor the aim of this paper is to investigate the challenges and improvements of visual-based SLAM rather than the traditional laser-based SLAM. In this paper we need to confirm that different SLAM algorithms give different results depending on whether the SLAM algorithm uses a feature-based approach or a direct-based approach. In this paper we want to see that the feature-based methods do not develop a denser representation of the environment whereas the direct-based methods develop a much denser representation of the environment. We would also want to discover that direct-based methods are more computationally expensive compared to feature-based methods. This paper will act as a gateway introduction to visual SLAM which would help an individual with little to no knowledge of the proposed research.

Acknowledgements

I would like to thank Dr. Ritesh Ajoodah for allowing me to select my own research topic which allowed me to explore more of what interests me.

Contents

Chapter 1

Introduction

The discovery of cameras in the 19th century has spawned multiple fields in human society. From recording a video in black and white to having a digitized quality video that represents an environment in its entirety, cameras have evolved better than the human eye can see. Cameras are used in various fields. We are able to capture images to keep in digital storages so we can feel nostalgic in being able to travel in time just by looking at a simple photo. Modern day artists no longer need to paint a person or an object in real time, they can simply take a photo and paint at their own leisure. The list goes on for the many applications of cameras. With the discovery of Artificial Intelligence cameras became one of the major contributors of AI's applications. Because they provide vision, we are able to perform vision-based machine learning. These applications are implemented on modern day cameras which are able to detect objects in the environment, filter out unwanted things in the environment and many other operations. In recent times however the combination of AI and cameras has been used in fields where an agent has to navigate through an environment, localising itself and building a map simultaneously this is called SLAM. The rise of self driving car companies like Tesla and comma ai are greatly impacted by the rise of better quality cameras with quality Machine Learning algorithms.

This rise of self driving vehicles which uses multiple cameras to navigate the environment has also led to more research conducted on using a single camera on an agent. Agents which use monocular cameras are mostly smaller agents like mobile robots and UAVs(Unmanned Aerial Vehicles) which are drones. These mini robots are used together with SLAM to perform multiple operations. The type of SLAM which is used together with a camera/monocular camera is called vision SLAM or vSLAM for short. Places where vSLAM hasn't been a commonplace is indoor

inspections. Robots are used to inspect storage facilities, caves and other indoor environments, however these robots still require trained pilots to use them. Because indoor environments have more obstacles it is much harder for autonomous and manually trained pilots to operate in them. In outdoor environments the use of GNNS is used.

In comparison to laser SLAM which uses a laser to generate a 2D occupancy map is not ideal as MAV/UAV robots are able to move around in 3D spaces. Another drawback of using laser SLAM is that lasers are very expensive. Visual SLAM on the other hand is much cheaper and it produces a 3D occupancy map.

Using inexpensive cameras for SLAM not only reduces cost it also reduces the weight of the robot and thus reduces battery life which is essential to keep the robot functioning to accomplish its duties.

As mentioned in the abstract, there are two different approaches of performing vSLAM we can use feature-based or direct-based SLAM , the difference between the two being that feature-based SLAM produces less dense maps whilst direct-based SLAM produces much denser maps, the drawback of using a much denser representation of a map is that it is computationally expensive to produce whereas if we use a less dense representation we don't get a full representation of the map, thus navigation and obstacle avoidance becomes hard.

Chapter 2

Background

2.1 Introduction

To understand the concepts of visual-slam and thus the research report, a context is necessary which would detail the ideas of SLAM and its applications.

2.1.1 SLAM

Background

Simultaneous Localisation And Mapping(SLAM) is a problem of an agent which is placed in an unknown environment to be able to build a map while being able to simultaneously know its location on the map.

2.1.2 Preliminaries

Lets consider an agent which moves through an environment with a sensor that helps in observation the environment.At time k we define the following:

- x_k : State vector which describes the orientation and location of agent.
- u_k : Control vector which is applied at k-1 which allows to move the agent from k-1 to state x_k.
- m_i : Location vector which describes the location of the ith landmark.
- z_{ik} : Observations taken from the agent's location of the ith landmark.

Lets define the following sets:

- $X_{0:k} = x_0, x_1, ..., x_k = X_{0,k-1}, x_k$: describes the history of the agent's location.
- $U_{0:k}$=$u_1, u_2, .., u_k = U$.
- m=$m_1, m_2, ..., m_n$ set of landmarks.
- $Z_{0:k}$=$z_1, z_2, ...z_k$ =$Z_{0:k-1}, z_k$ set of observations.

2.1.3 Probabilistic SLAM

The probability representation of SLAM is described as the following distribution: $P(x_k, m|Z_{0:k}, U_{0:k}, x_0)$ it will be computed k times. This distribution describes landmarks to the agent's state at time k given observations with control.

we now describe the **observation distribution** which is the probability of making an observation z_k given the agent's location and landmarks: $P(z_k|x_k, m)$

Lets now describe a motion model as: $P(x_k|x_{k-1}, u_k)$.

2.1.4 Monocular-SLAM

Monocular-SLAM is the ability to use SLAM algorithms with single camera.

2.1.5 Maneuvering

To move around one location to the next using SLAM is a challenge but In an article published by Shen et al. [2012] which explored the exploration of starting at an unknown location then the algorithm gradually builds a map of its surroundings.These surroundings are later labelled as seen locations thus extending the map representation.The approach used is known as a frontier-based exploration. In this paper it is found that the sensors providing incomplete information of the surrounding envi- ronment often fail to accurately capture environments which are unoccupied and unknown. Therefore frontier-based exploration fails because it relies on the boundary between unoccupied and unknown areas to determine the next exploratory step of the UAV this led to making a much denser map which was computationally expensive. Therefore to solve these issues an approach was devised which did not require a dense representation in which there's an observation and assumption that unstructured or uncluttered regions of the map that correlate

to the unexplored regions of indoor environments. This paper extends flight in unknown indoor environments Bachrach et al. [2009] which describes the difficulties in achieving fully autonomous flight for UAVs. The difference between ground robots and flying robots is also compared.

2.1.6 Mapping the environment

In order to map an environment a reconstructed mesh from sparse-clouds created by feature-based SLAM a certain solution was proposed In a paper published by Piazza et al. [2018] where a real-time algorithm is performed which is able to reconstruct a manifold mesh using a single core of a CPU while the other cores are used by the camera and for sparse data estimation via feature-based SLAM.For this experiment the recon- struction algorithm was tested on the KITTI dataset. Four stereo sequences of the visual odometry datasets which were labelled as : 00,01,02 and 05 were used, the hardware used was an Intel(R) Core(TM) i7-4770S at 3.10GHz with an 8GB of DDR3 RAM with the reconstruction algorithm running in a single core. Thus the proposed algorithm performed better than its competitors. This article extends from Chen and Medioni [1992] which focuses on creating a model of a physical model, it is relevant to the paper because to map complex indoor environments a model that represents the physical objects is required.

2.1.7 Scaling the environment

To scale the environment and thus creating a much denser representation from a feature-based vSLAM we would require a Convolutional Neural Network(CNN) In a paper published by Tateno et al. [2017] visual slam algorithms are fused with Convolutional Neural Networks(CNN) to improve the depth maps of the scene.This fusion improves the depth predictions in image locations where visual slam tends to fail, for example low textured areas are hard to visualise for SLAM algorithms thus CNN improves the quality of detecting these areas. The use of absolute scale of the reconstruction is demonstrated and thus a benchmark is done to show the accuracy of CNN.In this paper a comparison of LSD-SLAM and ORB-SLAM is conducted and the performance between the two was different because of the fact that LSD-SLAM is a direct method which is denser than the feature based LSD-SLAM.

The two articles that deal with navigation are Engel et al. [2014] and Mur-Artal et al. [2015].

2.1.8 Generalising using different Lighting

To explore more on mapping a complex indoor environments with different lighting conditions and using a low-cost depth camera and hardware a paper published by Newcombe et al. [2011] which comprised of components that make up the entire system which are the surface measurement,surface reconstruction update,surface prediction and the sensor pose estimation of which the initial depth measurements are obtained using a kinect device, thus a vertex map is created at the surface measurement stage, the surface measurement is then integrated into the scene model maintained with a volumetric,truncated signed distance funcion(TSDF) representation. In surface prediction, the loop between mapping and localisation is closed by tracking the live depth frame. Finally using multi-scale ICP alignment between the predicted surface and current sensor measurements in the sensor pose estimation makes sure that Live sensor tracking is achieved.

2.1.9 Add protective gear to the UAV

We can add protective gear to our robot(UAV) in a paper published by Caroti et al. [2018] the use of protective equipment. In this case a cage which surrounded a UAV that served as protection was used,real world data was used. Due to the presence of the rotating cage every image displayed was always changing the portion of the scene. The exploration of the use of protective structures like cages around the UAV when used indoors. Since more obstacles are present in- doors protection is investigated.

2.1.10 Adding more cameras

we can also add more cameras to the robot(UAV) to improve the SLAM algorithm. In an article published by Houben et al. [2016] the adding more cameras to the UAV improved SLAM and thus improving the movement around the environment by adding more visualisation to the UAV. In this paper the improved algorithm was more robust than the original monocular ORB-SLAM algo- rithm, by dynamically

changing the keyframes per frame performance escalated.The algo- rithm proved to be sufficient for path planning. Though with incred- ible improvements, the algorithm would give random results when it was operating near obstacles.This paper extends from Mur-Artal et al. [2015] which explains ORB-SLAM, the authors added Multi-cameras,Integration of an Inertial Measurement Unit(IMU) filter that supports visual tracking they also added enhancements to the original algorithm which were: local map estimation and keyframe creation.

2.1.11 LSD-SLAM

For LSD-SLAM we have a paper published by Engel et al. [2014] LSD-SLAM is described as a direct-based algorithm which produces dense and accurate maps which are consistent and also has accurate pose estimations.

2.1.12 ORB-SLAM

For ORB-SLAM we have a paper published by Mur-Artal et al. [2015] ORB-SLAM is described as a feature-based algorithm which produces robust and trackable maps, it is also good at loop closing and relocalization which means that if a UAV moved to a different location it is able to come back to its original location

Chapter 3

Research Methodology

When generating a map that will be used for navigation there are certain advantages and sacrifices to be made, either we use the denser point cloud representation provided by direct SLAM methods or we use the less dense representation provided by feature-based SLAM methods. In the case of using direct-based methods we would have an advantage of having a denser representation which means that every obstacle on the map is more visible and thus we would be able to navigate the world freely, but we also sacrifice the performance of compute power because using direct SLAM methods is expensive. When we use feature-based SLAM methods we have an advantage of being able to use a method which doesn't require much computing power but we sacrifice map representation, because feature-based SLAM produces less dense maps some places on the map are less represented thus feature-based SLAM methods have more chances of not showing some obstacles.

To solve the problem that feature-based SLAM methods have is to densify the map that it produces. This would still have its drawbacks because it would require sensors to do scale estimation, sensors which are expensive. Thus to solve this problem we would have to fuse feature-based SLAM with a neural network called a Convolutional Neural Network(CNN) which would take care of scale estimation. This fusion of SLAM and CNN would improve the depth predictions because feature-based SLAM produces sparse point clouds. Thus the fusion of SLAM and CNN will increase or will generate more representable dense obstacles on the map , making the map more navigable for path planning.

3.1 Research design

For this research the use of feature-based ORB-SLAM will be used as it is simple to implement given that it doesn't require more computing power than its direct-based competitor methods like LSD-SLAM. The research is designed to use a UAV as a robot which will implement SLAM. The UAV that's going to be used is a DJI Tello, a small UAV dedicated for educational purposes, taking pictures and videos and other uses. The DJI Tello is better suited to integrate it with SLAM, it has a single camera mounted at the front which will be able to capture a video that will be used by SLAM. In this review the UAV is used for indoor exploration, the UAV communicates via WI-FI, although it doesn't have a wide WIFI range it is suitable for the research. The DJI tello is mainly chosen because of its low price in UAV terms.

The system will be implemented on ROS which is a robot framework which uses nodes to communicate using a subscribe/publish model. The DJI tello will be using ORB-SLAM while its inner workings like the controller are managed by ROS. The DJI Tello will use ORB-SLAM because of its robustness compared to other SLAM methods.

3.2 Methods

The DJI tello will provide a video stream to the video node as RGB image input which is received by the ORB-SLAM node, the map will consist of feature points which are less dense which is unsuitable for obstacle avoidance. Therefore the map generated by ORB-SLAM wont have a correct scale.

To solve this issue of not having a scaled map and not having a dense representation of the map a CNN is used. The CNN will be in the middle of the input RGB image video stream and the SLAM node. In this case the CNN will receive input as RGB images then uses these images to produce a much denser depth map.

Here is a breakdown of all the components to be used:

ROS(Robotic Operating System): ROS provides a communication framework between programs which are called nodes, each node communicates by publishing to a topics with a messages, another node will subscribe to the topics to get the messages sent by the publishing node.

UAV: For test purposes a simulated UAV is used to capture a video stream it uses WIFI(2.4GHz) to communicate, it comes with an SDK for developers.Its camera specifications are:

Photo = 5MP(2592 x 1936) FOV = 82.6 degrees Video = HD720P at 30fps Format = JPG(Photo);MP4(Video) Image Stabilization = Electronic Image Stabilization(EIS)

SLAM: As mentioned before ORB-SLAM will be used for the research, LSD-SLAM will be used for comparison.

CNN: A ResNet-50 Architecture will be used.

To train the CNN a dataset is required, an NYU Depth dataset will be used as it contains indoor images which are required in our research. This dataset contains up to 5000 images of indoor environments. These images have been recorded by a kinect camera. This dataset includes different types of rooms like basements, dining rooms, offices , bathrooms and offices. These images will be enough to train the network these images are taken from different areas .

While training the dataset we provide it with depth images corresponding to a sequence of RGB images. So The network uses RGB images as input and evaluates using images with depth as the ground truth.we determine if the depth estimation network is optimal using ORB-SLAM depth samples. This information is used in networks to improve depth predictions. The downside of ORB-SLAM is that untextured spaces aren't very helpful, the precision will be lower ,this can be minimized by filtering out areas with low textures. In this way, the network is trained more strongly to resist changes in the samples.

3.3 Limitations

The maps produced are static, making ORB-SLAM to have difficulties in performing the loop closure process, loop closure is the process of moving from one position to another then coming back to the exact initial position. The maps are easily affected by noise making it difficult to have proper depth maps.

Chapter 4

Results

When ORB-SLAM was performed twice, using localization mode and non-localisation mode, the camera properties were:

- Camera: Pinhole
- fx: 435.20468139648438
- fy: 435.20468139648438
- cx: 367.45172119140625
- cy: 252.20085144042969
- k1: 0
- k2: 0
- p1: 0
- p2: 0
- fps: 20
- color order: RGB (ignored if grayscale)

And the ORB extractor parameters were:

- Number of Features: 1200
- Scale Levels: 8
- Scale Factor: 1.2000000476837158

- Initial Fast Threshold: 20
- Minimum Fast Threshold: 7
- First KF:0; Map init KF:0

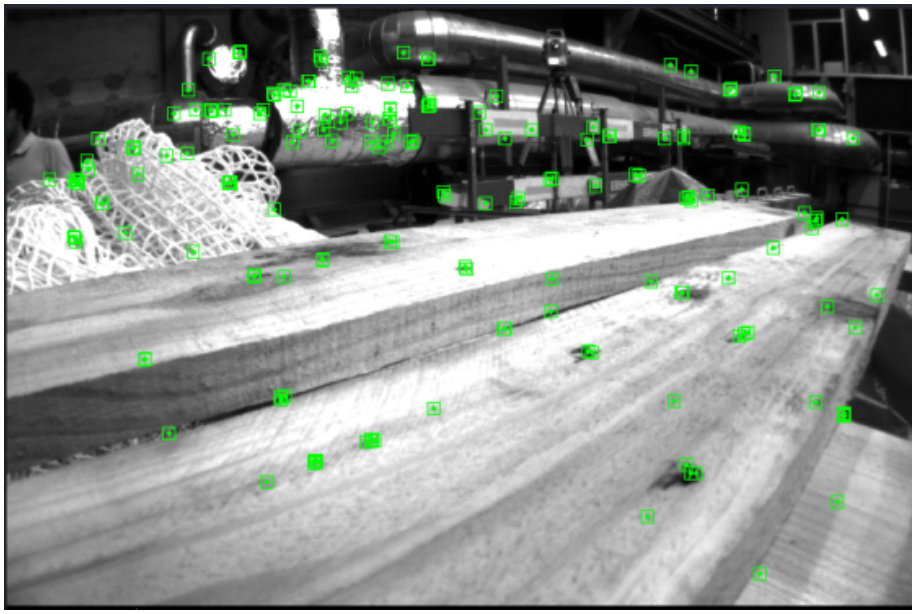

FIGURE 4.1: ORB-SLAM without Localization mode

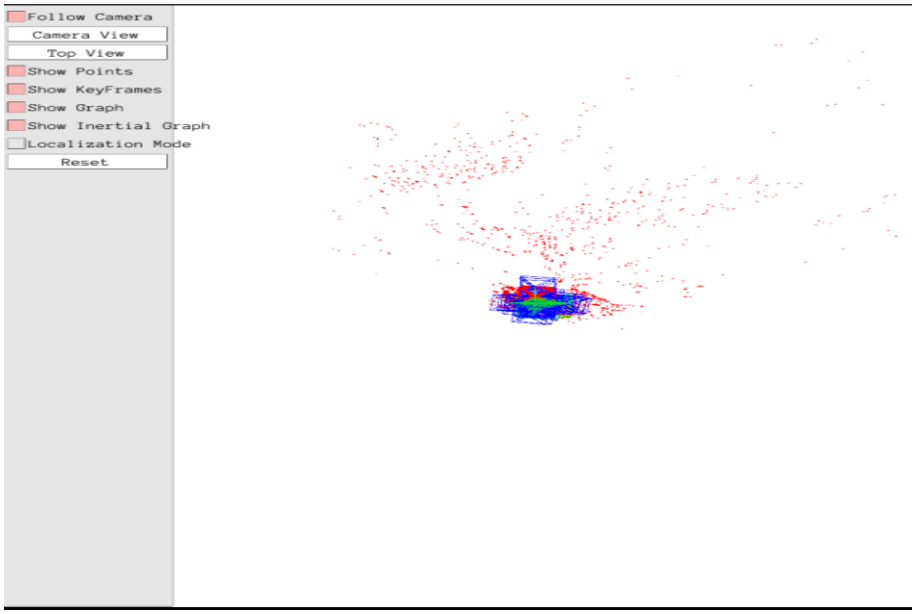

FIGURE 4.2: ORB-SLAM without Localization mode

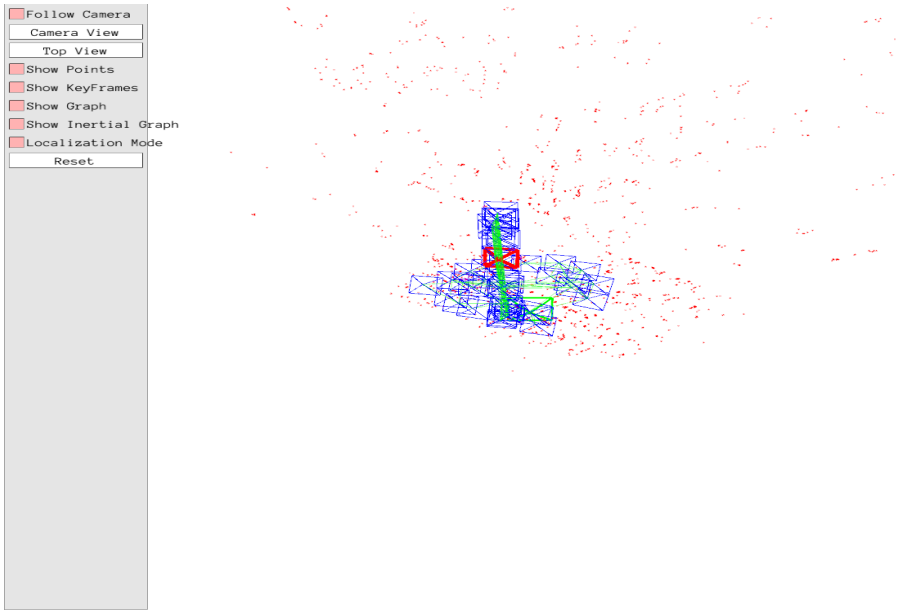

FIGURE 4.3: ORB-SLAM with Localization mode

Chapter 5

Conclusion

Monocular Camera or visual SLAM was used in most of the studies. Different SLAM algorithms are available, direct-based vs feature-based. There is also an increase in Machine learning which is used for Position and orientation estimation.The two popular SLAM methods are ORBSLAM and LSD-SLAM. The key difference is that ORB SLAM is feature-based with distributed point clouds, while LSD-SLAM is a direct-based method, which produces semi-dense point clouds. ORB SLAM is more powerful than LSD-SLAM.Creating a live navigable map or creating a live map using feature-based SLAM rarely creates a point cloud. Another option is to simply use the SLAM system for location and use the structure of the motion algorithm to generate the map of the environment. However, due to their high density approach, they are more often used in post processing after encountering performance difficulties in real time. Construction of maps in real time is a demanding procedure which takes time. The most common method used to scale a map is to use an external sensor such as a depth camera. Recently, CNN depth estimation is more often used for improvement. The SLAM system provides a scale, facilitates initialization, and improves the pose(position+orientation) estimation. For example it can also be generated by leveraging the high-density depth prediction of these networks. Mapping of environmental barriers or obstacles is much easier with the fusion of ORB-SLAM and CNN. The depth predictions can also be combined with semantics. To create information and make the map more deeper.

Visual Simultaneous Localisation And Mapping will always be a continuous topic that computer scientists will continue to improve. With SLAM algorithms still being algorithms that still do not cater for some environments, vision-based UAVs that uses SLAM will always struggle in mapping those environments which are not

"seen" by the UAV. But as we saw in the literature review that we can improve these SLAM algorithms by adding more cameras to the UAV which increases the visual which then improves the SLAM algorithms to be able to map environments much better, we can also improve the SLAM algorithms(ORB-SLAM to be specific) by combining it with a Convolutional Neural Network which performs depth estimations and scale estimations to mitigate the effects of using a less denser SLAM algorithm(ORB-SLAM) because using a denser SLAM algorithm(LSD-SLAM) would be computationally expensive. An improvement to the UAV drone in general would be to add protective gear (like cages around it) so that even if it comes into contact with obstacles it is well protected.However as mentioned vision-based UAV localisation and navigation still need further research as some factors affecting the UAV or rather factors which affect the SLAM algorithms still poses a huge problem. Factors such as loop-closure which states that if an agent(UAV) moves from an original position to another position , it should be able to come back to its original position at some time.

Bibliography

[1] Abraham Bachrach, Ruijie He, and Nicholas Roy. "Autonomous flight in unknown indoor environments". In: *International Journal of Micro Air Vehicles* 1.4 (2009), pp. 217–228.

[2] G Caroti et al. "Indoor photogrammetry using UAVs with protective structures: issues and precision tests". In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42.3/W4 (2018).

[3] Yang Chen and Gérard Medioni. "Object modelling by registration of multiple range images". In: *Image and vision computing* 10.3 (1992), pp. 145–155.

[4] Jakob Engel, Thomas Schöps, and Daniel Cremers. "LSD-SLAM: Large-scale direct monocular SLAM". In: *European conference on computer vision*. Springer. 2014, pp. 834–849.

[5] Christian Eschmann et al. "Unmanned aircraft systems for remote building inspection and monitoring". In: (2012).

[6] Sebastian Houben et al. "Efficient multi-camera visual-inertial SLAM for micro aerial vehicles". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 1616–1622.

[7] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. "ORB-SLAM: a versatile and accurate monocular SLAM system". In: *IEEE transactions on robotics* 31.5 (2015), pp. 1147–1163.

[8] Enrico Piazza, Andrea Romanoni, and Matteo Matteucci. "Real-time cpu-based large-scale 3D mesh reconstruction". In: *arXiv preprint arXiv:1801.05230* (2018).

[9] Shaojie Shen, Nathan Michael, and Vijay Kumar. "Autonomous indoor 3D exploration with a micro-aerial vehicle". In: *2012 IEEE international conference on robotics and automation*. IEEE. 2012, pp. 9–15.

[10] Keisuke Tateno et al. “Cnn-slam: Real-time dense monocular slam with learned depth prediction”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6243–6252.

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10]