

DISS. ETH NO. 26039

EFFICIENT VISUAL LOCALIZATION FOR GROUND VEHICLES IN OUTDOOR ENVIRONMENTS

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

MATHIAS BÜRKI

MSc. ETH ME

born on October 28, 1987
citizen of Ennetbaden, Switzerland

accepted on the recommendation of
Prof. Dr. Roland Siegwart, Examiner
Prof. Dr. Juan Tardós, Co-examiner

2019

Autonomous Systems Lab
Department of Mechanical and Process Engineering
ETH Zurich
Switzerland

© 2019 Mathias Bürki. All rights reserved.

Abstract

Visual (self)localization enables Autonomous Ground Vehicles (AGVs) to assess their position and orientation within an environment with up to centimeter level accuracy, using only cost-effective camera sensors. Especially for high precision maneuvering in GNSS-denied environments, using cameras for localization may be the best suited option for budget- or weight constrained platforms. However, particularly in outdoor environments, camera images are subject to various forms of appearance change. This renders it challenging to reliably localize a vehicle against a map previously built from sensor data recorded under different appearance conditions. A powerful approach to deal with these appearance changes is to enhance the map with visual data from several recordings, each collected under different appearance conditions. The amount of data generated following this approach, however, scales with the number of recordings collected over time, and thus unveils a need for smart algorithms managing this data and ensuring efficient use of computation, storage and network bandwidth resources. The contributions of this thesis are centered around the research questions addressing this need for a resource-efficient and reliable visual localization system for AGVs in outdoor environments.

In Part A, we propose an algorithm to dynamically select small amounts of map data matching the current appearance condition, thereby lowering network bandwidth consumption, and reducing computational demands on the vehicle platforms. We show that exploiting co-observability statistics allows for performing this appearance-based map data selection in a highly effective manner, without the need to explicitly model or enumerate the different appearance conditions.

Part B is devoted to the development of a practical map management process for a visual localization system targeted at long-term use. Our experiments have revealed that multi-session maps converge to a relatively stable state after several months of collecting recordings under varying appearance conditions. Furthermore, through a tight integration of appearance-based map data selection with offline map summarization, a completely scalable visual localization and mapping framework is reached that can be used for indefinite periods of time.

In Part C, we present the visual localization system developed within the UP-Drive project¹ for autonomous cars in urban outdoor environments. Thereby, a special focus has been placed on robustness against outdoor and long-term appearance change, and on a careful evaluation of the localization accuracy. We demonstrate that reliable and accurate visual localization is feasible in structured outdoor environments, even over long time spans, across vastly different seasonal,

¹<https://www.up-drive.eu>

Abstract

weather, and lighting conditions including at night-time, and with local point features with binary descriptors on a CPU-only computer architecture.

Zusammenfassung

Visuelle (Selbst-)Lokalisierung ermöglicht autonomen Landfahrzeugen ihre Position und Orientierung in einer Umgebung mithilfe von kostengünstigen Kamerasensoren zentimetergenau zu bestimmen. Speziell für hoch präzises Navigieren in Umgebungen, in denen kein GNSS Signal verfügbar ist, können Kameras die bestmögliche Sensorwahl sein für die Lokalisierung von Fahrzeugen, bei deren Ausstattung Kosten oder Gewicht ein wichtiger Faktor sind. Gerade in Freilandumgebungen unterliegen Kamerabilder jedoch verschiedenen Formen von Erscheinungsveränderungen. Dies erschwert die Lokalisierung eines Fahrzeuges in einer Karte, die zuvor mithilfe von Sensordaten, die unter anderen Erscheinungsbedingungen aufgezeichnet worden sind, erstellt wurde. Eine bewährte Vorgehensweise, um diesen Erscheinungsveränderungen Herr zu werden, erweitert die Karte mit visuellen Daten von mehreren Aufzeichnungen, die jeweils unter anderen Erscheinungsbedingungen aufgezeichnet wurden. Die Menge an Daten, die dabei generiert wird, skaliert jedoch mit der Anzahl Aufzeichnungen, die über die Zeit gesammelt und in die Karte integriert werden. Aus diesem Grund sind intelligente Datenverarbeitungsalgorithmen gefragt, die ökonomischen Ressourcenverzehr im Bezug auf Rechenleistung, Speicher, und Netzwerklast sicherstellen. Die Beiträge dieser Arbeit behandeln diese Forschungsfrage nach einem ressourceneffizienten und zuverlässigen visuellen Lokalisierungssystem für autonome Landfahrzeuge in Freilandumgebungen.

Im Teil A präsentieren wir einen Algorithmus zur dynamischen Selektion von kleinen Mengen an Kartendaten, die zu den aktuell vorherrschenden Erscheinungsbedingungen passen. Dies reduziert zum einen die Menge an Daten, die über ein Netzwerk ausgetauscht werden muss, und senkt zum anderen die Anforderungen an die Rechenleistung der Fahrzeuge. Wir zeigen, dass diese Aufgabe sehr effizient mithilfe von Statistiken erreicht werden kann, die das gemeinsame Beobachten von Landmarken abbilden. Im Speziellen ist keine explizite Codierung von Erscheinungsbedingungen erforderlich.

Teil B befasst sich mit der Entwicklung eines praktikablen Karten-Managementprozesses für visuelle Lokalisierungssysteme, die für Langzeitgebrauch ausgelegt sind. Unsere Experimente haben gezeigt, dass die multi-session Karten mit der stetigen Integration von Aufzeichnungen bei unterschiedlichen Erscheinungsbedingungen zu einem stabilen Zustand hin konvergieren. Der Zeitraum für diesen Konvergenzprozess liegt in der Größenordnung von wenigen Monaten. Des Weiteren kann durch eine enge Integration der ercheinungsabhängigen Kartendaten Selektion mit Techniken zur offline Kartenzusammenfassung ein komplett skalierbares visuelles Lokalisierungssystem aufgebaut werden, das für beliebig lange Zeiträume einsetzbar ist.

In Teil C präsentieren wir das visuelle Lokalisierungssystem, das während des UP-Drive Projekts für autonome Autos in urbanen Umgebungen entwickelt wurde. Dabei ist ein spezieller Fokus auf die Robustheit gegen Erscheinungsveränderungen in Freilandumgebungen über längere Zeiträume, und auf eine gründliche Evaluierung der Lokalisierungsgenauigkeit gelegt worden. Wir zeigen, dass zuverlässige und präzise visuelle Lokalisierung möglich ist, in strukturierten Freilandumgebungen, über lange Zeiträume, und trotz starker Erscheinungsveränderungen bedingt durch den Jahreszeitenwechsel, in unterschiedlichen Wetter- und Lichtverhältnissen inklusive Dunkelheit nachts, mit Punkt-Features mit binären Deskriptoren, und auf einer Rechnerplattform ohne Grafikkarte.

Acknowledgements

Thank you Prof. Dr. Roland Siegwart for the opportunity to conduct my doctoral study at the Autonomous Systems Lab. The positive environment you create in your lab is unique, and your always encouraging and supportive mindset truly inspirational. It is an honor and a privilege to be part of your group.

Thank you, Prof. Dr. Juan Tardós for reviewing and examining my thesis, and for your valuable feedback and the fruitful discussion at the defense.

Thank you Dr. Juan Nieto and Dr. Cesar Cadena, for your tremendous support with writing this thesis. There have been many ups and downs for me during this doctoral study, but at any point in time, I could always count on your help and your valuable advice, your guidance in writing this thesis, and your time for reviewing and proof-reading our publications, all for which I am deeply grateful.

Thank you Dr. Igor Gilitschenksi, Dr. Marcin Dymczyk, Dr. Renaud Dubé, Dr. Elena Stumm, and Lukas Schaupp for your valuable scientific contributions, and for your time to review and proof-read our publications. It is a true pleasure to work with you.

Thank you Lukas Schaupp and Mathias Gehrig for your valuable contributions to the UP-Drive project. Without your help, it would have not been possible to both write this thesis, and at the same time deliver our contributions to the research project.

Thank you Dr. Thomas Schneider, Dr. Marcin Dymczyk, Marius Fehr and Dr. Simon Lynen for your collaboration on the ASL mapping team. Your work on the ASL visual SLAM framework has been the basis upon which our contributions on lifelong localization and mapping in UP-Drive were developed. Especially the development of VIZARD[10] has significantly benefited from the ground work on sliding-window estimation carried out by Thomas.

Thank you Dr. Paul Furgale, Dr. Peter Mühlfellner, Dr. Ulrich Schwesinger, Woitek Derendarz, Fabian Pucks, and Ulrich Krebs for the close collaboration and teamwork on both the V-Charge and UP-Drive projects. It is a true pleasure to work with you.

Thank you Hannes Sommer. Always I could rely on your help, be it regarding advice for mathematical questions, or be it for your support of the ASL IT infrastructure. Your readiness to help others and your competence are second to none.

Thank you Woitek Derendarz and Dr. Martin Rufli. Without your strong commitment and dedication, the UP-Drive project would have never come about.

Thank you Cornelia Della Casa, and Lucy Borsatti for your administrative support at ASL in general, and in the UP-Drive project in particular. It is a true

Acknowledgements

pleasure to work with you.

Thank you Jan Viriden, Markus Bühler, Dr. Jérôme Maye, and Dr. Ralf Kästner for taking excellent care of the Kermit and JanETH research vehicles.

Thank you Eva and Peter Bürki. You have raised me to believe in myself. Without this, I would have never begun this doctoral study in the first place.

Thank you Anita for your infinite patience. I love you.

27th August, 2019

Mathias Bürki

Financial Support

The research conducted during this doctoral study has received funding from the European Unions's Seventh Framework Programme under grant-agreement No. 269916 (V-Charge), and under the European Union's Horizon 2020 Programme under grant-agreement No. 688652 (UP-Drive), and from the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0284 (UP-Drive).

Contents

Abstract	i
Zusammenfassung	iii
Acknowledgements	v
Preface	1
1 Introduction	3
1.1 Scope and Objectives	4
1.2 Approach	5
1.2.1 Part A: Online Landmark Selection	6
1.2.2 Part B: Efficient Map Management	7
1.2.3 Part C: Reliable Visual Localization in <i>UP-Drive</i>	8
2 Contributions	9
2.1 Part A: Online Landmark Selection	9
2.2 Part B: Efficient Map Management	12
2.3 Part C: Reliable Visual Localization in <i>UP-Drive</i>	13
2.4 List of Publications	15
2.5 List of Supervised Students	16
3 Conclusions and Outlook	19
3.1 Part A: Online Landmark Selection	20
3.2 Part B: Efficient Map Management	21
3.3 Part C: Reliable Visual Localization in <i>UP-Drive</i>	22
A. ONLINE LANDMARK SELECTION	25
Paper I: Appearance-Based Landmark Selection for Efficient Long-Term Visual Localization	27
1 Introduction	28
2 Related Work	30
3 Problem Statement	31
4 Probabilistic Landmark Ranking	32

Contents

5	Evaluation	34
6	Conclusion	42
Paper II: Appearance-Based Landmark Selection for Long-Term Visual Localization		45
1	Introduction	46
2	Related Work	48
3	Background	51
4	Appearance-Based Landmark Selection	53
5	Evaluation	57
6	Conclusions	79
B.	EFFICIENT MAP MANAGEMENT	97
Paper III: Map Management for Efficient Long-Term Visual Localization in Outdoor Environments		99
1	Introduction	100
2	Related Work	101
3	Methodology	103
4	Evaluation	107
5	Conclusions	112
C.	RELIABLE VISUAL LOCALIZATION IN <i>UP-DRIVE</i>	115
Paper IV: VIZARD: Reliable Visual Localization for Autonomous Vehicles in Urban Outdoor Environments		117
1	Introduction	118
2	Related Work	119
3	Methodology	121
4	Evaluation	124
5	Conclusions	133
	Bibliography	135
	Curriculum Vitae	143

Preface

This is a cumulative doctoral thesis and as such consists of the most relevant publications. The publications are grouped into three parts and attached at the end.

In addition to the individual publications an overarching introduction is provided in Chapter 1. We start with explaining the relevance of this thesis, followed by the objectives and the approach taken to fulfill these. For each contributing publication we explain how it embeds into the overall goals of this thesis and highlight the relevance of the research work in Chapter 2. Furthermore, we show how each paper is related to our other publications. We close this thesis by a summary of the achievements and provide an outlook for future directions and research in Chapter 3.

Chapter 1

Introduction

Knowledge of its own location within an environment constitutes one of the core competences of any mobile autonomous robot. Only by knowing its current location it becomes possible to infer where to go, and how to get there. Localization is thus a prerequisite for any goal-oriented planning and navigation capabilities.

It serves, however, also a second purpose. In complex surroundings, information gathered by sensors on-the-fly may not, in every situation, be sufficient for proper and safe interaction with the environment. In these cases, accurate localization allows exploiting prior information about the environment, thereby lowering the dependence on the robot's on-board sensory capabilities.

The establishment of satellite navigation systems (GNSS) approximately forty years ago have in principle enabled localization in outdoor environments everywhere around the globe. However, a lack in accuracy and unpredictable failure modes especially in urban environments render GNSS-based localization solutions unsuited for tasks requiring reliable and highly accurate localization, such as, for example, precise parking maneuvers of autonomous cars, or precise driving in (semi-)structured environments such as roads without lane markings, on sidewalks, or in open pedestrian areas in city centers. In addition to that, GNSS localization may be extremely inaccurate near high-rise buildings or even entirely unavailable in underground parking garages and tunnels.

As an alternative to GNSS-based localization, mobile robots may be equipped with exteroceptive sensors that directly perceive the near distance environment and allow to infer the robot's pose by relating the current observations with previously recorded data (i.e., the map). In the context of ground vehicles in outdoor environments, cameras and LiDAR sensors have received most attention for this role in the respective research communities in recent years. While LiDAR sensors are able to precisely measure the local geometry and are largely unaffected by appearance change, they are comparatively expensive. Cameras, on the other hand, are cost-efficient, yet still offer very rich information about the environment,

both in regard to appearance, but through Structure-from-Motion also in regard to geometry. They are thus an attractive choice of sensor for localizing autonomous ground vehicles in outdoor environments. It is the aim of this thesis to investigate the use of cameras to accurately localize ground vehicles in outdoor environments and address specific challenges arising in this context.

1.1 Scope and Objectives

For a visual localization system to be useful in real-world applications, a series of requirements need to be met. *a) Accuracy:* Firstly, the estimates of the vehicle's pose in its environment must be sufficiently accurate. For the applications targeted in this thesis, we aim at estimating all six degrees of freedom of the vehicle's pose with centimeter level accuracy, such that safely steering an autonomous vehicle on urban roads can be guaranteed. *b) Reliability:* Secondly, a visual localization system needs to be able to reliably provide pose estimates. Short periods of localization failure may be bridged by forward-propagating self-motion from wheel-odometry. This, however, quickly accumulates drift, leading to inaccurate and uncertain pose estimates already after a few meters. Therefore, a high localization recall is pivotal. In outdoor environments, localization recall is mainly challenged by changing appearance conditions, rendering it difficult to match visual cues in the current camera images with map data recorded previously under different conditions. *c) Efficiency:* Thirdly, pose estimates must be provided promptly and frequently, using the limited computational, memory and network bandwidth resources available on the mobile platform, the server-based map backend, and the communication infrastructure in-between. This requires fast algorithms, and compact data representations.

In addition to that, visual localization is closely related to mapping, with the aforementioned requirements on *accuracy*, *reliability*, and *efficiency* imposing direct implications on the map representation, and the process of building, extending, and curating visual maps. Accurate localization is only possible if the respective map data is mapped accurately. High localization recall, on the other hand, may require the map to contain visual cues from multiple recordings of an environment under differing appearance conditions. This might increase the size of the map, rendering it difficult to optimize, transmit, store, and load it into memory. Real-time localization further requires fast access to map data on the vehicles.

The development of a visual localization system that meets all of our criteria mentioned above, that is thus both highly accurate, reliable, and efficient, exceeds the scope of this thesis. However, a number of specific problems related to these three goals have been investigated in depth, and are presented in the following three parts. Part A is devoted to online selection of visual map data matching the current appearance condition. This part thus addresses efficiency, reducing both map transmission costs and computational resource demands on the vehicle. In Part B, the challenge of building and managing maps over long time spans within a completely scalable visual localization and mapping framework is addressed. A

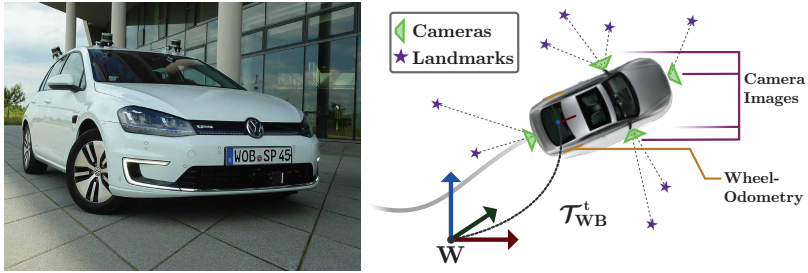


Figure 1.1: One of the UP-Drive autonomous cars (left), and an illustration of the sensor set-up used for visual localization in this thesis (right). A multi-camera rig, consisting of four wide angle cameras with fisheye lenses, is used together with wheel-odometry to provide accurate estimates of the 6DoF transformation between the vehicle body frame (B), and the map frame of reference (W) for every time where a set of camera images is captured (t).

main focus of this part is thus efficiency and reliability. In Part C, we present VIZARD, the visual localization system developed in the UP-Drive project. With this, we primarily address the need for accuracy and reliability.

1.2 Approach

The primary application targeted in this thesis are autonomous cars in urban outdoor environments. The research and technology is, however, directly applicable to arbitrary Autonomous Ground Vehicles (AGVs), both indoors and outdoors. Furthermore, the full potential of some of the algorithms presented, such as the online selection of map data based on the current appearance condition, or the underlying multi-session mapping framework, can best be exploited in outdoor environments, and in scenarios involving a fleet of vehicles sharing a map for localization.

We additionally assume that the operating environment is known and may thus be mapped a priori, and that the AGVs are equipped with one or multiple cameras, and a sensor providing self-motion estimates, such as wheel-odometry. One of the UP-Drive autonomous cars, together with an illustration of its sensor set-up as it is used for the visual localization algorithms presented in this thesis, is depicted in Figure 1.1. Furthermore, a visualization of the car localizing inside our map consisting of 3D point-feature landmarks is shown in Figure 1.2.

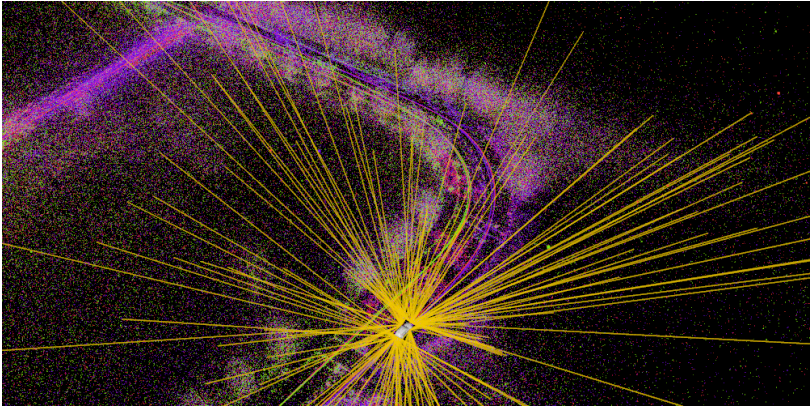


Figure 1.2: Illustration of the localization and mapping framework used in this thesis. The multi-session map consists of sparse 3D points, dubbed landmarks, triangulated from 2D point features (e.g., FREAK[3]) tracked in successive camera images. Different colors are used to represent the map session a landmark has been generated from. Feature points extracted on the live camera images on the vehicle are matched against map landmarks to form 2D-3D geometric constraints, from which the vehicle 6DoF pose with respect to the map reference frame is inferred using Non-Linear Least-Squares optimization. Inlier landmark observations are depicted as dark yellow lines between the camera, and the respective 3D map landmark.

1.2.1 Part A: Online Landmark Selection

Appearance conditions in outdoor environments can be drastically different, due to changes in weather, season, or illumination between day-time and night-time. These conditions may be so diverse that it is not possible to use a single set of visual landmarks for localization under all the possible different appearance conditions. To illustrate this, it suffices to note that not only feature descriptors of the same physical structure may be different under changing appearance conditions, but also the location of interest points may be vastly different at day time as opposed to at night. A map allowing for reliable visual localization under any appearance condition in these environments thus requires incorporating landmarks from multiple, different appearance conditions. We refer to such a map as a multi-session map. However, localization at any given point in time does not require all map landmarks, as only those representing the current appearance condition can be matched with features observed in the current camera images. This offers a potential to optimize for resource efficiency by selecting landmarks representing the current appearance condition in an online fashion, prior to matching them with features extracted from the current camera images. This on the one hand reduces the computational demands on the vehicles, as only a small fraction of landmarks

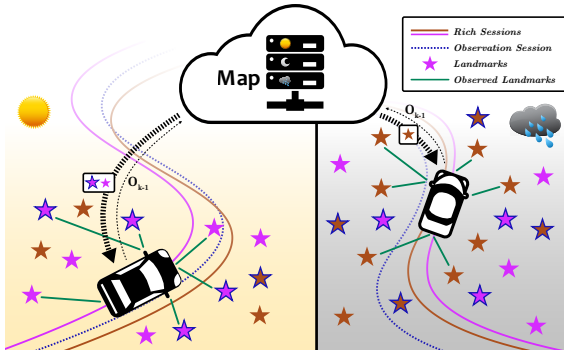


Figure 1.3: Shared-map scenario motivating the proposed appearance-based landmark selection. One large map containing landmarks recorded under different appearance conditions is stored and maintained on a cloud-based map backend. Vehicles en route under different appearance conditions retrieve selected landmarks matching their currently encountered appearance condition (thick dashed arrow), use those landmarks for visual localization (turquoise lines), and report back a set of recently observed landmark identifiers (thin dashed arrow).

needs to be processed in every iteration of the localization algorithm. On the other hand, it also allows for significantly reducing network bandwidth usage required for transmitting map data from a cloud-based server to the vehicle. The latter is of special interest in a scenario, where a fleet of vehicles uses a common map for localization, as it is anticipated for autonomous cars in the near future. A schematic illustration motivating the online landmark selection is depicted in Figure 1.3.

1.2.2 Part B: Efficient Map Management

In Part A, we assume a multi-session map with recordings covering all appearance conditions to be available in advance. In practice, this is not the case. Instead, the multi-session map needs to be built incrementally and curated over time, as gradually, data recorded under different appearance conditions becomes available. As some of the dynamics of appearance change, such as seasonal variations, occur on a very slow time scale, it may require a substantial amount of time until the multi-session map has reached sufficient appearance coverage. In addition to that, even on a cloud-based server, the computational and storage budget is finite. Therefore, conditions have to be defined for deciding which recordings of the environment should be added to the map, and algorithms have to be employed that tackle an indefinite growth of the map. In this part, we address these needs and investigate research questions such as how long it may take for a multi-session map to reach sufficient appearance coverage in outdoor environments, and what implications

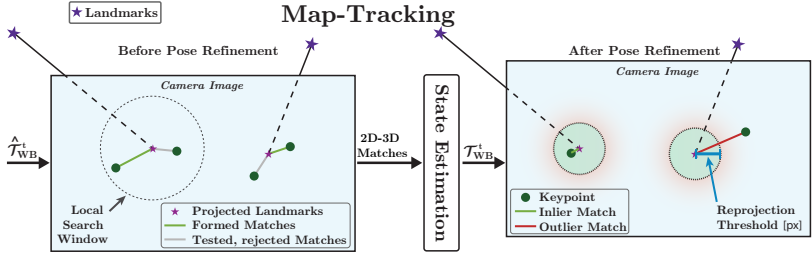


Figure 1.4: Local iterative localization paradigm, dubbed “Map-Tracking”, used in all three parts of this thesis. A rough estimate (\hat{T}_{WB_t}) of the vehicle’s position and orientation is used to project 3D map landmarks into the camera images and form 2D-3D matches using a local search window in the image space, and the descriptor distance (left side, “Before Pose Optimization”). These 2D-3D matches are used in a state estimation module to refine the vehicle pose estimate (\hat{T}_{WB_t}). In Part A and B, a simple estimate of the vehicle pose using only the visual 2D-3D constraints is used. In contrast to that, VIZARD in Part C fuses the visual 2D-3D constraints with wheel-odometry measurements in a probabilistic manner, in order to achieve temporally smoother pose estimates. The refined pose estimate is further used to distinguish between inlier and outlier landmark observations, employing a *Reprojection Threshold* ρ [px] (right side, “After Pose Refinement”).

the use of map summarization techniques may have on visual localization using appearance-based landmark selection.

1.2.3 Part C: Reliable Visual Localization in *UP-Drive*

A substantial effort during this doctoral study has been devoted to developing a highly reliable and accurate visual localization system, dubbed VIZARD, for the autonomous cars in the UP-Drive project. To achieve this, we on the one hand fuse wheel-odometry measurements with visual localization constraints using an Information Filter, which corresponds to the dual formulation of the (Extended) Kalman Filter. This allows for smooth and accurate pose estimates at all times. We further employ a local localization algorithm, referred to as map-tracking, in combination with multi-session maps. This enables highly reliable localization across vastly different appearance conditions. A detailed illustration of the this local localization algorithm can be found in Figure 1.4.

Apart from presenting the algorithmic components of VIZARD, we further focus on an extensive evaluation of its long-term performance in outdoor environments.

Chapter 2

Contributions

In this chapter, the scientific contributions achieved during this doctoral study are presented in relation to four first-author publications, which constitute the core of this thesis. All of the presented work has been conducted in strong collaboration with the co-authors and supervisors.

2.1 Part A: Online Landmark Selection

The following two papers address a specific aspect of the scalability challenge subject to visual localization systems, namely that of adaptive selection of map data on the vehicle side. This allows to save network bandwidth, and reduces the on-board computational load on the vehicles. It is in contrast in particular to offline map summarization[20, 24, 35, 65], which addresses an orthogonal aspect of the scalability challenge by performing an offline selection of map data on the map backend (server) side.

Paper I

Mathias Bürki, Igor Gilitschenski, Elena Stumm, Roland Siegwart and Juan Nieto,
“Appearance-Based Landmark Selection for Efficient Long-Term Visual Localization”,
presented at IROS 2016 in Daejeon, South Korea

Context

Using a sparse map with binary visual feature descriptors, it is not feasible to cover all appearance conditions encountered in outdoor environments using a single set of landmarks. Instead, a significant fraction of the landmarks are specific to the appearance condition encountered during the respective sortie through the mapped

environment. That is, these landmarks may only be re-detected and matched under similar appearance conditions. It follows from this, that at any given point in time, not all of the landmarks available in the map are useful for localization, as only a subset of landmarks matches the appearance condition currently present. Evaluating this subset of useful landmarks in an online manner given the currently encountered appearance conditions allows for efficient use of network bandwidth and computational resources.

Contribution

We propose a landmark selection scheme to select landmarks from a multi-session map matching the current appearance condition. At its core, a ranking function assigns a score to each landmark based on the co-observability relation with recently observed landmarks. It is thus an unsupervised distinction between useful and not useful landmarks under the current appearance condition, solely based on the implicit appearance coherence encoded by the co-observability relation. In particular, no explicit modeling of different appearance conditions is necessary. In contrast to related work by Linegar et al.[45], and MacTavish et al.[52], our method is able to evaluate the appearance coherence on the level of individual landmarks, and is thus able to exploit multi-session maps containing landmarks observed from more than one map session. In a thorough evaluation using two outdoor dataset collections, covering day-time conditions over the course of a full year, and the transition from day-time to night-time respectively, we demonstrate the potential of appearance-based landmark selection to significantly reduce network bandwidth usage while maintaining as high a localization performance as when using all map landmarks instead. Furthermore, an analysis of jointly selected landmarks across different datasets shows that our landmark selection algorithm uses different sets of landmarks for different times of the year, or different times of day respectively. This supports our hypothesis of needing more than a single set of landmarks to cover the various appearance conditions in outdoor environments.

Interrelations

The work presented in Paper I introduces an efficient algorithm for selecting landmarks based on the current appearance condition, but the conference format prevents an in-depth analysis and evaluation of the proposed selection strategy. Furthermore, the sensor setup available on the vehicle used for collecting the datasets in Paper I does not allow for an assessment of the localization accuracy. These aspects are addressed in Paper II, and the reader is kindly referred to Section 2.1.

In addition to that, the works presented in Paper I and Paper II assume an a-priori available multi-session map covering all appearance conditions. Building such a map in practice may be a time consuming and long lasting task, as for some modes of outdoor appearance change, such as seasonal variations, it may take up to one year until all necessary data can be collected. It is thus crucial to investigate a chronological, incremental multi-session map building process. Furthermore, as

described in Section 2.1, appearance-based landmark selection only addresses one aspect of the visual localization scalability challenge. In that regard, it is of special interest to evaluate the interaction of online appearance-based landmark selection with offline map summarization. These research questions are addressed in Paper III in Part B, and the reader is kindly referred to Section 2.2.

Paper II

Mathias Bürki, Cesar Cadena, Igor Gilitschenski, Roland Siegwart and Juan Nieto,
“**Appearance-Based Landmark Selection for Visual Localization**”,
published in the Journal of Field Robotics, 2019

Context

This paper builds upon and extends the work presented in Paper I. We aim at selecting small amounts of landmarks from a multi-session map matching the current appearance condition, in order to minimize map data exchange, and computational resource demand on the vehicle platforms. For more details on the context, the reader is kindly referred to Section 2.1.

Contributions

This papers investigates the characteristics of the appearance-based landmark selection introduced in Paper I in more depth. In particular, the influence of *observation sessions*, a technique introduced in Paper III to collect more statistical evidence for the co-observability relation without increasing the size of the map, is analyzed in detail. Furthermore, an additional landmark ranking function based on appearance equivalence classes is derived and evaluated. This ranking function is agnostic to the number of *rich-* and *observation sessions* present in the map, or the number of landmarks associated with any of the map sessions, and thus more ubiquitously usable in practice. We further relate and compare the proposed appearance-based landmark ranking functions with ranking schemes commonly used in the field of Information Retrieval. In addition to that, we make use of the *NCLT* dataset collection, which offers ground-truth poses. This allows evaluating and comparing the localization accuracy using different landmark ranking functions. We further evaluate the computation times needed for the individual modules in the localization pipeline. This has revealed a substantial potential to save computational resources on the vehicle platform by employing appearance-based landmark selection, since the latter allows to discard a large fraction of landmarks mismatching the current appearance condition at an early stage in the localization algorithm. As a result, the runtime of the localization algorithm on the vehicle can be significantly reduced compared to when all the landmarks from all appearance conditions have to be processed. Apart from saving computational resources on the vehicles, this further leads to a decoupling of the online localization runtime from the number of landmarks or sessions present in the multi-session map, and thus

improves the scalability of the localization and mapping system.

Interrelations

The work presented in this paper constitutes an extension, both theoretically and experimentally, of Paper I. Several practical aspects, such as the combination of online appearance-based landmark selection with offline map summarization, or the chronological, incremental building of the respective multi-session maps, are investigated in Paper III in Part B, presented in Section 2.2.

2.2 Part B: Efficient Map Management

Appearance-based landmark selection on the one hand only addressed one aspect of the visual localization scalability challenge, and on the other hand assumes an a-priori availability of a multi-session map covering all appearance conditions of a given outdoor environment. In practical applications, however, algorithms like appearance-based landmark selection, which optimize a specific part of the localization pipeline, must be combined with other modules optimizing orthogonal constraints to form a complete scalable localization framework. Furthermore, managing multi-session maps over long time spans poses a challenge on itself, as the map may need to be extended with data from new sessions as they become available over time.

Paper III

Mathias Bürki, Marcin Dymczyk, Igor Gilitschenski, Cesar Cadena, Roland Siegwart, and Juan Nieto,

“Map Management for Efficient Long-Term Visual Localization in Outdoor Environments”,

presented at IV 2018, in Changshu, China

Context

The work in this paper is driven by research questions addressing the practicability of a scalable visual localization system for real-world applications with a fleet of Autonomous Ground Vehicles. A map ought to be shared among multiple vehicles to capitalize on data collection synergies, and prevent data duplication. It may contain multiple sessions to cover a wide range of appearance conditions, but it must be built incrementally, and chronologically. The visual localization system as a whole must be scalable and resource efficient. That is, the map may not grow indefinitely over time, and network bandwidth needed for exchanging map data, as well as computational resources, both on a map backend, but also on the vehicle side, are limited and must be used economically. Furthermore, a simple and effective procedure for updating and curating the map is needed in long-term operations.

Contribution

We propose a decision criteria for adding a new dataset to a (multi-session) map based on the translation error resulting from localizing the new dataset against the existing map in an offline process. Indefinitely adding new session to the map may, however, at some point exceed the boundaries of the computational resources on the cloud-based map backend. We tackle this challenge by employing map summarization techniques, thereby enforcing an upper bound on the total number of landmarks in the map. Furthermore, we propose the concept of *observation sessions*, which allow to significantly increase the co-observability statistics between map landmarks without increasing the size of the map. In our long-term evaluation, we show that online appearance-based landmark selection, and offline map summarization, can be successfully deployed in combination, leading to highly efficient online visual localization in combination with a highly efficient multi-session mapping backend.

Interrelations

This paper addresses several practical aspects of developing and deploying a completely scalable visual localization and mapping framework. It is thus related to Paper I and Paper II presented in Section 2.1 and Section 2.1, which present and evaluate one of the key components of our scalable localization and mapping framework, namely online appearance-based landmark selection. In addition to that, the findings related to chronological, incremental map building and management are applicable to the work presented in Paper IV in Part C 2.3, which also employs multi-session maps to gain robustness against appearance change in long-term operations.

2.3 Part C: Reliable Visual Localization in *UP-Drive*

This part presents our efforts within the UP-Drive project to develop a highly reliable, robust, and accurate visual localization system for autonomous cars in urban outdoor environments. In contrast to Part A, and B, where we have addressed specific (sub-)modules of a localization system, we are in Part C interested in criteria concerning a high performing localization system as a whole. Reliability in this context means our localization system can be trusted to be functional regardless of the current weather, lighting or seasonal conditions. It is further crucial for the localization system to run for indefinite periods of time on the vehicles without interruptions or interventions. To achieve this robustness, careful software design, error handling, and memory management is necessary. In addition to that, the control stack in the UP-Drive cars solely relies on our visual localization system for actuating the steering wheel. A centimeter level localization accuracy at all times is thus required, in order to keep the cars safely within the lane boundaries.

Paper IV

Mathias Bürki, Lukas Schaupp, Marcin Dymczyk, Renaud Dubé, Cesar Cadena, Roland Siegwart, and Juan Nieto,

“VIZARD: Reliable Visual Localization for Autonomous Vehicles in Urban Outdoor Environments”,
presented at IV 2019 in Paris, France

Context

With VIZARD, we present a visual localization system for urban outdoor environments. A special focus is set on robustness against weather, lighting, and seasonal appearance change, as they are common to urban outdoor environments. We are further constrained to use a CPU-only computational platform on the UP-Drive cars, limiting us to the use of binary feature descriptors, such as FREAK[3]. The development of VIZARD has further been driven by high requirements on the metric localization accuracy and the need for a real-time capable localization system.

Contribution

We outline the components of our proposed visual localization system in detail, and describe our methodology of fusing wheel-odometry measurements with visual constraints stemming from map-tracking, our local localization module. The main contribution in this part, however, is a thorough parameter study and extensive evaluation in several challenging urban outdoor environments over multiple years and several hundreds of driving kilometers. We derive optimal parameters for our map-tracking module, allowing to maximize localization recall while maintaining high localization accuracy. Additionally, we compare the use of different binary descriptors, and demonstrate the benefit in localization recall attainable by performing local localization, as opposed to global localization. This work shows that visual localization using point features with binary descriptors is able to provide accurate metric pose estimates with nearly 100% localization recall across different weather and seasonal conditions, and even a night-time under artificial street lighting.

Interrelations

VIZARD addresses the probabilistic fusion of wheel-odometry and visual localization constraints neglected in the proof-of-concept localization systems used in Part A and B. This allows computing smoother vehicle poses, an important requirement of control algorithms relying on the pose estimates from a localization system. Furthermore, it is complementary to the work presented in Part A and B, as it addresses more development and integration related aspects, whereas the the work presented in Part A and B is more research oriented.

2.4 List of Publications

In the context of the author’s doctoral studies the following publications were achieved. They are presented in chronological order.

- *Lionel Heng, MATHIAS BÜRKI, Gim Hee Lee, Paul Furgale, Roland Siegwart, Marc Pollefeys*, **“Infrastructure-Based Calibration of a Multi-Camera Rig”**, ICRA, 2014
- *Hugo Grimmert, MATHIAS BÜRKI, Lina Paz, Pedro Pinies, Paul Furgale, Ingmar Posner, Paul Newman*, **“Integrating Metric and Semantic Maps for Vision-Only Automated Parking”**, ICRA, 2015
- *Ulrich Schwesinger, MATHIAS BÜRKI, Julian Timpner, Stephan Rottmann, Lars Wolf, Lina Maria Paz, Hugo Grimmert, Ingmar Posner, Paul Newman, Christian Häne, Lionel Heng, Gim Hee Lee, Torsten Sattler, Marc Pollefeys, Marco Allodi, Francesco Valenti, Keiji Mimura, Bernd Goebelsmann, Wojciech Derendarz, Peter Mühlfellner, Stefan Wonneberger, Rene Waldmann, Sebastian Grysczyk, Carsten Last, Stefan Brüning, Sven Horstmann, Marc Bartholomäus, Clemens Brummer, Martin Stellmacher, Fabian Pucks, Marcel Nicklas, Roland Siegwart*, **“Automated Valet Parking and Aharging for e-Mobility”**, IV, 2016
- *Peter Mühlfellner, MATHIAS BÜRKI, Michael Bosse, Wojciech Derendarz, Roland Philippsen, and Paul Furgale*, **“Summary Maps for Lifelong Visual Localization”**, JFR, 2016
- *MATHIAS BÜRKI, Igor Gilitschenski, Elena Stumm, Roland Siegwart and Juan Nieto*, **“Appearance-Based Landmark Selection for Efficient Long-Term Visual Localization”**, IROS, 2016
- *Miguel Valls, Hubertus Hendriks, Victor Reijgwart, Fabio Meier, Inkyu Sa, Renaud Dubé, Abel Gawel, MATHIAS BÜRKI, Roland Siegwart*, **“Design of an Autonomous Racecar: Perception, State Estimation and System Integration”**, ICRA, 2018
- *MATHIAS BÜRKI, Marcin Dymczyk, Igor Gilitschenski, Cesar Cadena, Roland Siegwart, and Juan Nieto*, **“Map Management for Efficient Long-Term Visual Localization in Outdoor Environments”**, IV, 2018
- *Nikhil Bharadwaj Gosala, Andreas Bühler, Manish Prajapat, Claas Ehmke, Mehak Gupta, Ramya Sivanesan, Abel Gawel, Mark Pfeiffer, MATHIAS BÜRKI, Inkyu Sa, Renaud Dubé, Roland Siegwart*, **“Redundant Perception and State Estimation for Reliable Autonomous Racing”**, ICRA, 2019

- *MATHIAS BÜRKI, Cesar Cadena, Igor Gilitschenski, Roland Siegwart and Juan Nieto*, “**Appearance-Based Landmark Selection for Visual Localization**”, JFR, 2019
- *MATHIAS BÜRKI, Lukas Schaupp, Marcin Dymczyk, Renaud Dubé, Cesar Cadena, Roland Siegwart, and Juan Nieto*, “**VIZARD: Reliable Visual Localization for Autonomous Vehicles in Urban Outdoor Environments**”, IV, 2019
- *Lukas Schaupp, MATHIAS BÜRKI, Renaud Dubé, Roland Siegwart, Cesar Cadena*, “**OREOS: Oriented Recognition of 3D Point Clouds in Outdoor Scenarios**”, under review for IROS, 2019

2.5 List of Supervised Students

Throughout the author’s doctoral studies, a substantial effort has been spent on supervising a series of student projects.

Master’s Thesis

Master student, six months full time

- Dino Hüllmann, “Continuous-Time SLAM using Hermithian Splines”
- Lukas Fröhlich, “Improving Multi-Sensor Data Fusion for Localization of Automated Vehicles”, awarded with the *ETH Medal* for an outstanding Master’s Thesis, nominated for the *Johann Puch Automotive Award*
- David Vogt, “Outdoor Global Localization for an Autonomous Car”,
- Leonie Traffelet, “Hybrid Vision-LiDAR Localization for Autonomous Ground Vehicles”,
- Lukas Schaupp, “Place Recognition with Data-Driven Descriptors using 3D Point Clouds”,
- Gregory Bättig, “One Shot Learning for Traffic Sign Recognition”,
- Jannic Veith, “Visual Localization for an Autonomous Car in a 3D LiDAR Map”

Semester Projects

Master student, three to four months part time

- Franz Thurnhofer, “Comparison of State-of-the-Art Methods for Localization of Self-Driving Cars”,

- Baldur Yngvason, “Let’s Drive! Mission Planning for a Self-Driving Car”,
- Victor Reijgwart, “System Integration for an Autonomous Racecar”,
- Miguel de la Iglesia, “State Estimation and Sensor Fusion for an Autonomous Racing Car”,
- Manish Prajapat, “Sensor Fusion and Velocity Estimation for an Autonomous Race Car”,
- Fynn von Kistowski, “Distillation of Keypoint Detection and Description Networks for Mobile Applications”,
- Niclas Vödisch, “Cone Detection and Classification using Cameras and a LiDAR”,
- Shashank Shing, “Perception and Fusion of Cone Measurements for an Autonomous Racing Car”,
- Patrick Pfreundschuh, “Map-Tracking using LiDAR Sensors for Localization of an Autonomous Car”

Computational Science and Engineering (CSE) Seminar

Master student, literature review, three to four months part time

- Shoshana Jakobovits, “RatSLAM: a Review”,
- Stefano D’Apolito, “Feature Detection and Extraction in 3D LiDAR Point-Clouds”,

Probabilistic Learning for Robotics (PLR) Project

Master student, three to four months part time

- Niclas Vödisch, and Andreas Bühler, “Vision-LiDAR Inter-Modality Representation Learning using Generative Adversarial Networks (GANs)”,
- Hao-Chih Lin and Juan Lin, “Vision-LiDAR Inter-Modality Representation Learning using Superpoint”,

Bachelor’s Projects

Bachelor student, three to four months part time

- Tobias Grundmann, “Long-Term Evaluation of Keypoint Descriptor Stability in Outdoor Environments”,
- Mathieu Rohner, “Integration and Evaluation of a Binary Bag-of-Words for Outdoor Loop Closure”,

Study on Mechatronics

Bachelor student, literature review, three to four months part time

- Kornel Eggenschwiler, “State-of-the-Art in Visual Outdoor Loop Closure Detection”
- Maurice Grunder, “State-of-the-Art in Large-Scale Bundle-Adjustment for Visual SLAM”

Conclusions and Outlook

During these doctoral studies, we have been able to push forward the state-of-the-art in visual localization in outdoor environments. Thereby, experience and insights have been gained, which are shared in this chapter.

As a primary conclusion, we can state that accurate and reliable visual localization is *feasible* in challenging outdoor environments, despite the considerable change in appearance conditions encountered in long-term operation. We have been able to demonstrate this in particular in Part C, using classic Computer Vision tools such as point features and local descriptors, and on a CPU-only computational platform. As our experiments have shown, a high localization recall and thus a high reliability can be attributed to a large extent to the use of a local localization algorithm, such as map-tracking, as compared to global localization.

Nevertheless, and in spite of the efforts made in Part A and Part B towards efficiency and economic resource use, scalability remains a major challenge. Even when employing online appearance-based landmark selection, combined with offline map summarization, the visual localization system as presented in this thesis and used in UP-Drive may reach its limitations in areas considerably larger than several square kilometers. The recent trend towards more abstract, and thus more semantically meaningful visual features may help to significantly improve the scalability of visual localization in the future. It remains, however, an open challenge and future research focus to demonstrate that visual localization with more abstract feature representations is able to achieve similarly accurate pose estimation results as with classic point features.

It the remainder of this chapter, concluding remarks and possible future research directions are discussed for a selection of subtopics of special interest for each of the three parts of this thesis.

3.1 Part A: Online Landmark Selection

We have presented an algorithm for dynamically selecting landmarks from multi-session maps matching the current appearance conditions. With this, map data transmission costs can be lowered, and the requirements on the computational platform on the vehicles is reduced, without significantly sacrificing localization precision or recall.

How to encode map data from multiple appearance conditions in the map?

A limitation of many algorithms addressing online adaptive selection of map data is a often strong and inherent dependence on the underlying map representation. That is, the appearance-based landmark selection we have proposed in Paper I and Paper II is targeted at a multi-session map representation, where sparse 3D point landmarks are all expressed in the same frame of reference. It is only transferable to a limited degree to other map representations. In recent years, various other paradigms have been followed in order to represent visual maps used under different appearance conditions[16, 52, 71], without a clearly visible trend towards a more generic or standardized map data representation. Future efforts to harmonize map representation may foster the development of online and offline map selection algorithms with a broader, more generic applicability.

Evaluation of Appearance Condition As described in Paper II, the consistency with the current appearance condition may need to be re-evaluated repeatedly along a traversal through the mapped area (see “reset” in Section 5.7). A naive strategy of re-evaluating after a fixed number of localization iterations is employed for simplicity and proof-of-concept in Paper II. This may not be an optimal strategy for real-world applications. Instead, more adaptive algorithms should be investigated that may, for example, trigger “resets” only when the trajectories of *rich sessions* deviate (e.g., in one *rich session*, the car drives straight, in another, it took a right turn), or when there are temporal discontinuities in the observed appearance conditions (e.g., when driving in or out of a tunnel).

Explicit Encoding of Appearance Conditions The appearance-based landmark selection proposed in Paper I and Paper II exploits an implicit encoding of appearance condition solely based on the co-observability relation of landmarks in the map. This may allow for optimal selection performance, as the ranking of landmarks is determined in a data-driven manner. It further simplifies the design and integration of the appearance-based landmark selection module into a visual localization and mapping framework, as no explicit modeling of appearance conditions is necessary. However, an explicit encoding, employing a pre-defined enumeration and classification of appearance conditions (e.g., night-time, sunny, spring, etc.), may also exhibit certain advantages and may thus be worth exploring in the future. For example, an explicit encoding may

alleviate the need for “resetting”. It may further allow for a more in-advance selection of map data matching the current appearance condition, e.g., already in the garage prior to a sortie of the vehicle, based on date, time of day, or weather forecast.

3.2 Part B: Efficient Map Management

A thorough investigation of employing appearance-based landmark selection in combination with offline map summarization has shown that it is possible to build a completely scalable visual localization and mapping framework, and that the respective multi-session maps can be incrementally built over the time span of several months.

Map Post-Processing During these doctoral studies, building the multi-session maps has often proven to be the by far most challenging task, as compared to developing the localization algorithms. While for localization, a standard algorithmic approach is applicable, the various mapping post-processing and optimization steps have often revealed to require substantial manual inspection and verification, and have thus remained difficult to automate. Future research ought to be dedicated towards automated detection of (geometrical) inconsistencies among multiple map sessions in the map, followed by automatic mitigation, through, e.g., additional outlier rejection steps.

Map Tiling Providing highly accurate visual localization for continuous areas significantly larger than a few square kilometers may become challenging with a single, monolithic multi-session map, even when employing efficient map management strategies, such as appearance-based landmark selection and offline map summarizing as they are described in Paper I, II, and III. While either cloud-based map data streaming, as motivated in Part A and B, or alternatively, standard disk caching mechanism if the map is stored on the vehicle, may allow for efficiently performing online localization in virtually arbitrarily spacious environments, managing and optimizing the maps on the map backend may reach its limitations. Splitting up the map into separate, or only loosely coupled, sub-maps on the backend may thus be inevitable if the map coverage should exceed a certain spatial extent. Therefore, future research ought to discuss the question of when to trigger a split, how large a sub-(multi-session)-map should be, how to prevent open-loop trajectory segments, and how to handle sup-map switching in online operation.

Data Integrity If a fleet of vehicles collaboratively contribute to extending and improving a shared map, as described in Part B, data integrity and security may become an important aspect. How can it be guaranteed that the mapping data streamed to the cloud-based map backend can be trusted? How are faulty sensors, or faulty sensor calibrations, detected and mitigated? How can a malicious attempt to temper with the shared map be prevented? While

aspects like these may be ignored when focusing on developing the proof-of-concept (S)LAM algorithms, they certainly are of importance when targeting an industry grade product deployment, and many questions regarding reliable detection of faulty data are still unsolved. A direct collaboration between SLAM roboticists and IT network security experts may effectively address these open questions in the future, and allow for safe and secure exchange of map data among a fleet of vehicles.

Data Compression In the transmission direction from a cloud-based map backend to the vehicle, compressing the map data further than what is proposed in Paper I and Paper II may not be feasible, as the 3D position of all landmarks used for localization is required in order to be able to infer the metric pose of the vehicle. In the opposite direction, that is from the vehicle to the map backend, there may be potential to further decrease the data transmission, as not all recently observed landmarks may be necessary to successfully infer the appearance condition using the co-observability relation. More research into automatically extracting a smaller subset of observed landmarks representative of the current appearance condition may thus allow to further reduce network upload traffic without a decrease in localization performance with appearance-based landmark selection.

3.3 Part C: Reliable Visual Localization in *UP-Drive*

We have shown that with a CPU-only computational platform, and by employing point features and binary descriptors, highly reliable localization is feasible, even across vastly different appearance conditions as they occur over long periods of time in outdoor environments.

The Landmark Representation Point features with local descriptors have proven to be a powerful couple for highly accurate metric visual localization. However, as we have seen in this thesis, as well as in many related work [12, 16, 24, 49, 52, 65], a large number of features have to be extracted, tracked, matched and managed in order to cover an outdoor environment both spatially and in appearance. This renders map management challenging and requires extensive and for certain applications prohibitive computational and storage resources, even if algorithms optimizing resource use are employed. Furthermore, point features lack semantic meaning, and are thus not ideal for coarse grained localization and navigation, or global localization in spatially large environments. For these reasons, substantial efforts have been made in recent years to use more abstract, semantically meaningful features for visual localization [28, 68, 76, 89, 90]. While this has shown promising results for coarse grained global localization, centimeter-level accurate and reliable metric pose estimation with other features than points remains an open challenge. Recent work by DeTone [23] exploiting advancements in machine learning have, however, further pushed forward the quality of point features.

This may remedy some of the weaknesses of classic handcrafted point features as used in this thesis. In particular, it may require considerably fewer map sessions to cover the same diversity of appearance conditions. It does, however, also require a high-end GPU to be available on the vehicles for online localization. Further research into deriving more compact, but similarly well performing Artificial Neural Network models could mitigate this problem and make algorithms like Superpoint useable on CPU-only platforms in the future.

Local vs. Global Localization The high localization recall attained with VIZARD in Paper IV has to be attributed to a large degree to the fact that the visual localization algorithm only solves a local localization problem. This drastically reduces the search space for feature correspondences, to an extent that allows to be relatively permissive in regard to the allowed descriptor distance of 2D-3D matches. Once the vehicle is localized for the first time, local localization is well justified. Subsequent queries along a trajectory are highly correlated in space, and even the incremental motion between the capturing of two subsequent set of camera images is well observable with wheel-odometry or an IMU sensor. Depending on the application and environment, bootstrapping local localization may, however, impose a challenge. Fortunately, in outdoor environments, which are most challenging for visual global localization algorithms, there is often a GPS signal available. As our experience in UP-Drive shows, for its application domain of autonomous cars in urban outdoor environments, using a consumer grade GPS sensor to bootstrap VIZARD has proven to be a reliable and robust solution. In contrast to that, state-of-the-art global localization algorithms, employing e.g. a visual Bag-of-Words, may provide satisfactory performance for bootstrapping local localization indoors, as there is considerably less variance in appearance conditions.

Sensors During the work on the V-Charge¹, UP-Drive², and AMZ Driverless Car Racing³ projects, experience has been gathered with visual localization of AGVs using different kinds of sensor set-ups. It has repeatedly proven valuable to use a multi-camera rig, covering an extensive field of view of, ideally, 360 degrees. This increases the number of visible features, and the robustness towards occlusion, dynamic objects, adverse sunlight, and sensor failure. In addition to camera(s), the visual localization algorithms presented in this thesis further rely on the availability of odometry measurements, in order to forward propagate the state estimates, for temporal smoothing, and for bridging short-time localization failures with dead-reckoning. For slow dynamics, as encountered with autonomous cars in urban environments, the presence of wheel-odometry has proven to be very valuable for this task. It is built in per default in cars, of low cost, allows for an unambiguous

¹<https://cvg.ethz.ch/research/v-charge/>

²<https://www.up-drive.eu>

³<http://driverless.amzracing.ch/en/home>

detection of dynamic objects in the scene, and provides - in contrast to an IMU - fully observable self-motion even in the case of planar constant velocity motion. Nevertheless, the use of an IMU in addition to wheel-odometry may in practice still be worthwhile, as this may improve the estimation of roll and pitch motion, allow for an assessment of the gravity direction, and provide redundancy. In addition to that, in applications with more dynamic driving motion, such as in the AMZ Driverless Racing Competition, or in more high speed driving conditions, such as on high-ways, wheel-odometry may be much more subject to slipping as compared to driving in urban environments. In these situations, the presence of an IMU, or even only a gyroscope, may significantly improve the odometric yaw rate estimation.

Global Shutter vs. Rolling Shutter Visual localization in outdoor environments can be especially challenging under poor lighting conditions as they occur for example at night-time. In these situations, a high dynamic range of the camera chip is paramount, as this prevents motion blur. In that regard, CMOS camera sensor chips usually offer a considerably higher dynamic range than their CCD counterparts. However, the former are commonly subject to rolling shutter artifacts, while the latter have a global shutter circuitry. In our experiments presented in Paper IV, we have tested our visual localization system both with a global shutter camera system on the NCLT datasets, and with a rolling shutter camera system on the UP-Drive datasets. In the driving dynamics present in the UP-Drive datasets, that is with driving speeds of up to 35km/h , there has been no noticeable compromise in localization performance that could be clearly attributed to the rolling shutter mechanism. The difference in image quality under night-time conditions is, however, very clearly pronounced. The NCLT night-time dataset from December 1st 2012 exhibits considerable motion blur, even under the comparatively slow motion of the Segway platform driving at approximately 5km/h . In contrast to that, the rolling shutter cameras in the UP-Drive datasets provide crisp and feature-rich images even at night-time, thereby considerably facilitating localization under artificial street lighting. For slow moving, and slow turning ground vehicles such as autonomous cars in urban environments, the benefit of increased dynamic range with CMOS rolling shutter cameras may thus outweigh the potential loss in localization performance caused by rolling shutter motion artifacts. It is, however, important to note that this observation may not be applicable for higher driving speeds, as they are common on rural roads or high-ways, since in these scenarios, motion distortion due to the rolling-shutter may be more pronounced.

Part A

ONLINE LANDMARK SELECTION

Appearance-Based Landmark Selection for Efficient Long-Term Visual Localization

Mathias Bürki, Igor Gilitschenski, Elena Stumm, Roland Siegwart and Juan Nieto

Abstract

In this paper, we present an online landmark selection method for distributed long-term visual localization systems in bandwidth-constrained environments. Sharing a common map for online localization provides a fleet of autonomous vehicles with the possibility to maintain and access a consistent map source, and therefore reduce redundancy while increasing efficiency. However, connectivity over a mobile network imposes strict bandwidth constraints and thus the need to minimize the amount of exchanged data. The wide range of varying appearance conditions encountered during long-term visual localization offers the potential to reduce data usage by extracting only those visual cues which are relevant at the given time. Motivated by this, we propose an unsupervised method of adaptively selecting landmarks according to how likely these landmarks are to be observable under the prevailing appearance condition. The ranking function this selection is based upon exploits landmark co-observability statistics collected in past traversals through the mapped area. Evaluation is performed over different outdoor environments, large time-scales and varying appearance conditions, including the extreme transition from day-time to night-time, demonstrating that with our appearance-dependent selection method, we can significantly reduce the amount of landmarks used for localization while maintaining or even improving the localization performance.

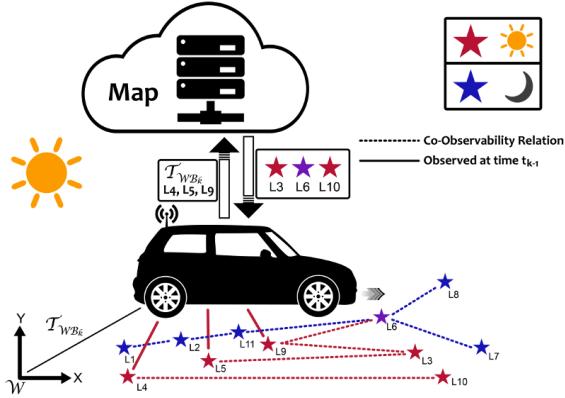


Figure 4.1: A schematic illustration of a distributed visual localization system using online landmark selection. A vehicle continually receives selective visual landmarks for localization from a cloud-based map server during operation, while transmitting back its pose and information about recently observed landmarks. In the depicted situation, landmarks L_4 L_5 and L_9 have recently been observed, therefore their IDs, plus a rough initial estimate of the vehicle’s current pose are transmitted to the map server. In response, a subset of relevant landmarks, consisting of L_3 , L_6 and L_{10} are transmitted back to the vehicle and used for subsequent localization.

1 Introduction

A fundamental problem to be tackled to enable fully autonomous driving is the cooperation and coordination among multiple vehicles, including sharing and exchanging information. This will be a key aspect for the success in coping with the complexity, variability, and volatility of typical urban environments. Especially for the task of localization and mapping, sharing and maintaining a common map offers a high potential for reducing data redundancy and for providing timely up-to-date maps. Vehicles will be required to exchange data among themselves, and/or with a common cloud-based map-service. Since bandwidth on mobile data networks is a scarce resource, it is pivotal to minimize the amount of information exchanged. This is particularly important for visual localization and mapping, where appearance variations generate the need to store many different representations for each location [16], [65].

To approach this problem, we propose an online landmark selection method which - without losing localization performance - is able to significantly reduce the amount of data exchanged between the vehicle and its map source.

The general principle of the method can be summarized as follows:

- The prevailing appearance condition of the environment is inferred from landmarks observed in recent localization attempts during a traversal through the mapped area.
- Using this information, all landmarks in a spatially local neighborhood (the candidate landmarks) are ranked according to how likely they are to be observed in subsequent localization attempts along the same traversal.
- A selected subset of top-ranked candidate landmarks is transferred back to the vehicle and used for localization.

Localization can then be performed on the vehicle, based on the selection of suitable landmarks, which is computed on a remote map server and sent to the vehicle. The data exchanged during each localization attempt consists of a rough initial estimate of the vehicle's current pose, references (e.g. IDs) to recently observed landmarks, and a reduced set of selected landmarks. A schematic illustration of this distributed localization paradigm can be found in figure 4.1.

The key for an effective landmark selection is the ranking process. In our approach, this ranking is performed in an unsupervised manner, based on co-observability statistics between a candidate landmark and a set of recently observed landmarks collected in past traversals through the same area.

The main contribution of the proposed approach is the derivation of an online landmark selection method based on co-observability statistics. The motivation for this work is the need for a localization and mapping strategy, that can deal with the bandwidth-constrained settings found in distributed systems operating in changing environments. In particular, our approach provides the following features:

- Efficient and accurate localization using only an appearance-dependent subset of landmarks inferred at runtime in an unsupervised manner.
- The size of the selected subset is adaptable to prevailing bandwidth restrictions.
- Computational demands on the vehicle are reduced by significantly cutting down the amount of input data used for localization.

We evaluate our approach in two complementary scenarios. In the first scenario, our landmark selection method is evaluated in a long-term experiment on an outdoor parking-lot, covering day-time conditions observed over the time frame of one year. In the second scenario, our method is evaluated in a small city environment, covering extreme appearance changes from day-time to night-time over the time frame of one day. The results validate our approach by showing that with our landmark selection method, we significantly reduce the amount of data exchanged between the vehicle and the map, while maintaining comparable or even better localization performance than if all data is used.

Note that despite us partly drawing the motivation for this work from a cooperative multi-vehicle scenario, the algorithm is evaluated in a distributed single-vehicle set-up. That is, a single vehicle localizes against a potentially remote cloud-based map-server, attempting to minimize the bandwidth usage while maintaining the localization performance. The method shown readily generalizes to and taps its full potential in a multi-vehicle set-up.

The remainder of this paper is structured as follows: In section 2, our proposed selection method is put into context with other related work. Section 3 and 4 derive the underlying appearance-based landmark ranking function the selection method is based upon, before an evaluation thereof is presented in section 5. To conclude, we summarize our findings and discuss future work in section 6.

2 Related Work

Extensive efforts have been made in the past years to adapt visual localization systems for long-term operation and resource-constrained environments. The methods presented in [20], [35], [57] all involve an adaptive selection of either landmarks or visual views in order to bound the growth of maps, while accounting for an environment subject to appearance change. This selection may be based on a short-term/long-term memory model [20], on clustering techniques [35], or on random pruning in neighbourhoods of high data density [57]. Similarly, the summary-mapping techniques proposed in [65] and [25] aim at maintaining as compact and small a map representation as possible, while at the same time covering a high degree of variance in appearance. All of these methods have in common, that the selection is an offline process performed prior to and/or independent of the robot's next operation. In contrast to that, our proposed selection method is an online process, selecting landmarks at runtime according to the prevailing appearance conditions, without modifying the underlying map.

In [45], an online selection algorithm is presented that is, as in our case, adaptive to appearance conditions. Rather than reasoning over relevant landmarks, different visual "experiences" are prioritized for localization on resource constrained platforms. In contrast to this setting based on "experiences", all landmarks that we select are expressed in a common coordinate frame, which allows the poses of the vehicle to also be estimated in a common frame, independent of what landmarks are selected and hence what appearance condition the vehicle is exposed to. This enables a seamless integration of our method with other modules of an autonomous vehicle, such as planning, navigation, and control. Furthermore, by performing selection at the level of individual landmarks, our approach is more closely linked to the underlying environmental features. In this way, accounting for the fact that many landmarks may be shared among similar appearance conditions while others may be very distinct to certain conditions is implicitly handled by our framework.

Landmark selection has also been studied in connection with specific tasks like path-planning and obstacle avoidance. The method presented by [62] selects those landmark measurements from a map, which maximize the utility wrt. a predefined

task, such as collision-free navigation. In contrast, the method we present selects landmarks based on the appearance condition the robot is exposed to during operation.

Recently, landmark co-occurrence statistics have been increasingly exploited in the context of place-recognition. In [19] co-occurrence information is used to infer which types of features are often seen together, as this helps distinguishing places. Furthermore, in [32], [33], and [85] places are described and identified by constellations of visible landmarks or features grouped based on co-observability, therefore incorporating pseudo-geometric information in their representation. Similarly, the works of [44] and [61] rely on landmark co-occurrence statistics for prioritizing relevant landmarks or environments for improved place-recognition efficiency. The clear correlation between the appearance of the environment and the co-observability of landmarks demonstrated in these works has inspired the selection algorithm presented in this paper. However, we propose to use the co-observability statistics in order to achieve a different goal, namely to infer which landmarks are likely to be observable in the near future during online operation, allowing to minimize data exchange.

Along that line, the work presented in [14] is similar to ours, as they learn co-observability relationships across different appearance conditions in order to predict the current operating condition of the robot. The main difference is that their co-observability prediction is performed on the level of camera images, whereas we propose to exploit co-observability on the level of individual 3D landmarks, contained in a sparse geometric visual map.

3 Problem Statement

We consider a scenario in which iterative visual localization systems, such as the ones described in [65], [39] or [16], are used for periodic correction of pose estimates obtained from odometry. The underlying map is assumed to be stored as a pose-graph in a multi-session SLAM framework, as described in [17], which contains information about landmarks (position estimates and feature descriptors of respective observations). Additionally, bundle adjustment and loop-closure [50] have been performed to merge identical landmarks observed in multiple mapping sessions, register the maps to each other and refine the resulting joint map.

Each of the mapping sessions that is used for generating the multi-session map may have been recorded at different times, with possibly very different appearance conditions. Therefore, the observability of individual landmarks is highly variable and not all are equally useful for localization under a specific appearance condition. In a scenario as described in section 1, where the map is located on a cloud-based server, a decision has to be made about which landmarks to use, and hence transmit to the vehicle, for each localization attempt during online operation. In order to support this decision, the vehicle provides the server with a rough initial pose estimate and information on which landmarks have recently been observed along the trajectory. Based upon this information, we propose a landmark selection

method aimed at only selecting those landmarks for transmission to the vehicle, which are deemed likely observable, and thus useful for localization.

In particular, we are interested in the following landmark ranking function:

$$f_{\hat{T}_{WB_k}, \mathcal{O}_{k-1}}(l) := P(l | \hat{T}_{WB_k}, \mathcal{O}_{k-1}) \quad (4.1)$$

It denotes the probability of observing landmark l at time t_k , given an initial estimate of the vehicle's pose denoted by \hat{T}_{WB_k} , and a list of recently observed landmarks before time t_k denoted by \mathcal{O}_{k-1} . In order to improve readability, we use abbreviated symbols $\hat{T}k$ and \mathcal{O} for the remainder of this section.

4 Probabilistic Landmark Ranking

Using Bayes' rule, expression 4.1 can be reformulated as

$$P(l | \hat{T}k, \mathcal{O}) = \frac{P(\mathcal{O} | l, \hat{T}k) \cdot P(l | \hat{T}k)}{P(\mathcal{O} | \hat{T}k)} \quad (4.2)$$

Probability $P(\mathcal{O} | \hat{T}k)$ is a fixed constant and does not influence the ranking of landmarks, whereas $P(l | \hat{T}k)$ denotes the pose dependent probability of observing landmark l at time t_k . We model the latter with a uniform distribution over all landmarks observed from within a given radius r around the estimate of the vehicle's pose $\hat{T}k$, and zero for other landmarks. In practice, this allows retrieving an appearance-independent tight spatial subset of possibly observable candidate landmarks, denoted by \mathcal{C}_k , as described in [64], which are then ranked according to $P(\mathcal{O} | l, \hat{T}k)$.

The term $P(\mathcal{O} | l, \hat{T}k)$ can be interpreted as the probability of having recently observed the set of landmarks \mathcal{O} , given an estimate of the vehicle's current pose $\hat{T}k$ and that landmark l is observed at time t_k . Since past landmark observations are independent of the current vehicle's pose, we can reformulate as follows:

$$P(\mathcal{O} | l, \hat{T}k) = P(\mathcal{O} | l) \quad (4.3)$$

From a frequentist's perspective, $P(\mathcal{O} | l)$ could be approximated by the number of times, *all* landmarks in \mathcal{O} and l have been observed together in the past, divided by how often l was observed. For such a quantification to hold as a good approximation, the amount of co-observation data must be very high and no "appearance-outliers" (i.e. landmarks recently observed although they do not conform with the overall prevailing appearance condition) may be present in \mathcal{O} - two requirements unlikely met in practical applications. We therefore propose to approximate $P(\mathcal{O} | l)$ by explicitly accounting for limited statistical data and possible "appearance-outliers".

We assume the multi-session map has been generated from datasets representing traversals through the mapped area under different appearance conditions, possibly

augmented with additional co-observation statistics from further traversals through the map. Thus, we interpret the set of all past traversals, denoted by \mathbf{Z} , as an enumeration over appearance conditions represented in the multi-session map. With this, we can use the law of total probability in order to obtain the following decomposition:

$$P(\mathcal{O} | l) = \sum_{z \in \mathbf{Z}} P(\mathcal{O} | z, l) \cdot P(z | l) \quad (4.4)$$

We model $P(z | l)$ with a uniform distribution over all traversals z in which l was observed, and zero for all other traversals. For a traversal observing l , the likelihood $P(\mathcal{O} | z, l)$ becomes independent of l , and we can thus reformulate (4.4) as:

$$P(\mathcal{O} | l) = \frac{1}{|\mathbf{Z}'|} \sum_{z \in \mathbf{Z}'} P(\mathcal{O} | z) \quad (4.5)$$

where \mathbf{Z}' denotes all traversals in \mathbf{Z} where l was observed in. Due to the fact that this appearance term $P(\mathcal{O} | l)$ is only evaluated for a spatially local subset of landmarks \mathcal{C} (retrieved evaluating $P(l | \hat{\mathcal{T}}k)$) and the appearance condition is assumed to be locally stable, both in a spatial and temporal manner, it suffices to consider a landmark as observed in traversal z if it has been observed at least once along the traversal, regardless of the place or time. Analogously, two landmarks are considered co-observed in a traversal z , if both of them have been observed at least once in z , at potentially different times and places.

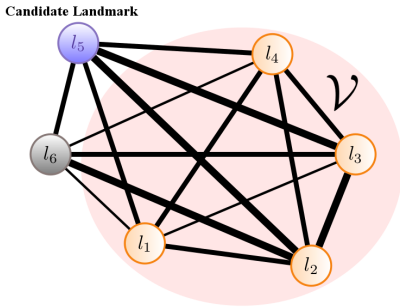


Figure 4.2: The co-observability graph represents which landmarks (vertices) have been co-observed how often in the past (edges). From knowing which landmarks have recently been observed along the current traversal (\mathcal{V} , orange), our goal is to decide how likely a candidate landmark (blue) is to be observed at the current time-step.

For each of these past traversals in \mathbf{Z}' either none, some or even all landmarks in \mathcal{O} were observed. To account for potential "appearance-outliers", we model the probability $P(\mathcal{O} | z)$, namely the probability of observing \mathcal{O} in traversal z , to be equal to the fraction of landmarks in \mathcal{O} actually observed in traversal z .

In conclusion, we can express our ranking function as follows:

$$f_{\hat{\mathcal{T}}k, \mathcal{O}}(l) = \frac{1}{|\mathbf{Z}'|} \sum_{z \in \mathbf{Z}'} |\mathcal{O}_z| \quad (4.6)$$

where $|\mathcal{O}_z|$ denotes the number of landmarks of \mathcal{O} that were observed in traversal z . For simplicity, the constant denominator $|\mathcal{O}|$ is omitted from the sum as it does not influence the ranking.

An intuitive graphical interpretation of this ranking function is shown in Figure 4.2, where landmarks are represented as vertices and the co-observation relation as weighted edges connecting them. The landmarks colored in orange denote the recent observations \mathcal{O} , while the candidate landmark is colored in blue. The score of candidate l according to the presented ranking function corresponds to the sum of co-observation connections into \mathcal{O} , normalized by the total number of traversals observing l . It represents how tightly a candidate is connected to the set \mathcal{O} in the pair-wise co-observation graph. Hence, candidates with a strong connection into \mathcal{O} are favored over those with only a weak connection, relating to how likely the given candidates are co-observed with \mathcal{O} .

5 Evaluation

The proposed landmark selection method exploits varying appearance conditions expressed in a single multi-session map of sparse landmarks. In order to be able to build such multi-session maps, sufficient data must be collected during the mapping phase, that is diverse enough to cover several different conditions, while exhibiting also some overlap in appearance. To the best of our knowledge, no publicly available datasets fulfill these criteria. We therefore evaluate our selection method in two complementary experimental scenarios using our own datasets recorded for the purpose of evaluating long-term visual localization and mapping.

In scenario A, a multi-session map of an open-space parking lot area is created, with datasets spanning over one year, covering the entire range of weather conditions and seasonal change. In scenario B, a city environment is mapped over the course of six hours from day-time to night, covering the most extreme change in appearance from daylight to night-time under artificial street lighting.

A total of 31 traversals of the parking lot environment (roughly 155m each) and 26 traversals of the city environment (roughly 455m each) were recorded, resulting in an accumulated driving distance of about 16.5km. For each environment, half of the recordings distributed over the respective time spans were used to build the map and augment the co-observability data, while the other half ($\approx 8km$) were used for the evaluation. Example images from each of the two environments can be seen in figures 4.3 and 4.4.

The vehicle’s sensor setup consists of four wide-angle fish-eye cameras - one in each cardinal direction - and wheel odometry sensors. The cameras run at a frame-rate of 12.5Hz. All images were recorded in gray-scale and down-scaled to 640px x 480px.

During each traversal, localization is performed iteratively. For each image, a rough initial pose estimate is calculated (based on the previous pose estimate and integrated wheel odometry), a candidate set \mathcal{C}_k is retrieved, from where a top-ranked subset of landmarks is selected yielding \mathcal{S}_k , landmark-keypoint matches are formed, the initial pose estimate is refined using a non-linear least-squares estimator, and a final match classification step distinguishes between inliers and outliers. The landmarks associated with these inlier matches are considered the



Figure 4.3: Example images from the parking-lot environment, showing the varying appearance conditions induced by changes in lighting, weather, as well as foliage.



Figure 4.4: Example images from the city environment, showing the changes in appearance from day to night.

observed landmarks at a given time t_k , as described in section 3, and are denoted by \mathcal{O}_k .

In \mathcal{O} , we only keep observed landmarks from the previous localization (i.e. from time t_{k-1}), since in our experimental scenarios, no significant improvement was observable when extending \mathcal{O} over a longer time window.

5.1 Ranking Function and Selection Policies

We aim at demonstrating that with our selection method, we can significantly reduce the number of landmarks used for localization while simultaneously maintaining a similar localization performance. For this, we evaluate several performance metrics for three different selection policies: i) using the ranking function derived in section 3 and 4, ii) random selection, and iii) simply selecting all landmarks. The latter marks the baseline for our experiments, while random selection constitutes a lower bound for the quality of our ranking-based selection.

We formally define the selection policy as follows:

$$\Omega(\mathcal{C}, f(), r, m) := \text{Select } n \text{ top-ranked landmarks}$$

where $n = \min(r * |\mathcal{C}|, m)$, based on a selection ratio r and a maximum number of landmarks m . Consequently, $\Omega(\mathcal{C}, f(), 1.0, \infty)$ corresponds to selecting all landmarks. While parameter m directly relates to some fixed constraint on the available network

bandwidth, the selection ratio r prevents the algorithm to select poorly ranked landmarks in spatial locations of generally few visual cues (small $|\mathcal{C}|$). For the sake of notational brevity, we abbreviate $\Omega(\mathcal{C}, f(), r, m)$ by $\Omega(f(), r, m)$ in the plots shown. With $f_{rank}()$ we refer to the ranking function derived in section 3 and 4, while $f_{rand}()$ denotes a random uniform ranking across \mathcal{C} .

5.2 Metrics

The following metrics are evaluated and respective experimental results are presented in the remaining subsections.

a) *Ratio between the number of selected landmarks and the number of candidate landmarks:*

$$r_k^{sel} := \frac{|\mathcal{S}_k|}{|\mathcal{C}_k|}$$

This metric directly relates to the amount of data transmission saved by performing landmark selection.

b) *Ratio between the number of observed landmarks with and without a selection at a given time t_k :*

$$r_k^{obs} := \frac{|\mathcal{O}_k^{\Omega(f_{rank}(), r, m)}|}{|\mathcal{O}_k^{\Omega(f(), 1.0, \infty)}|}$$

The number of observed landmarks constitutes a good indicator of the resulting pose estimate's accuracy (see [65]) and the ratio r_k^{obs} is directly related to how well the selection predicts the current appearance condition. An ideal landmark selection method would achieve a ratio close to 1.0, with a significantly reduced number of selected landmarks.

c) *RMS errors for translation and orientation wrt. wheel-odometry:*

For each localization attempt, the transformation between the initial rough pose estimate, based on the visual pose estimate from t_{k-1} and forward integrated wheel-odometry, and the refined visual pose estimate from t_k , can be computed, and is denoted by $T_{B_k^{est} B_k^{odo}}$. Conceptually, this transformation corresponds to the odometry drift correction. While wheel-odometry accumulates drift over time, it is locally very smooth. Since this refined visual pose estimate is only based on the positions of the matched landmarks, and in particular no odometry fusion is performed, the magnitude of $T_{B_k^{est} B_k^{odo}}$ is dominated by the uncertainty of the visual estimate.

We compute separate RMS errors for both the translational and rotational component of $T_{B_k^{est} B_k^{odo}}$. Note that this metric does not describe the absolute localization accuracy. It only constitutes an indicator for the relative uncertainty of the visual pose estimates allowing a comparison between the three cases of selecting all landmarks, random selection, and ranking-based selection.

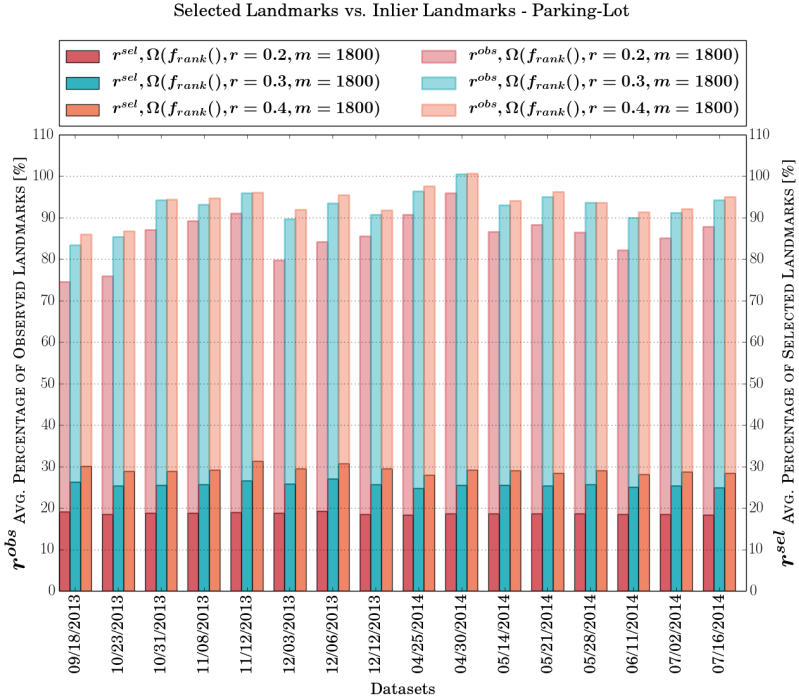


Figure 4.5: Illustration of the relation between the average number of selected landmarks and the average number of observed landmarks for datasets from the parking-lot scenario. The lower bars (between 20-30%) correspond to the average percentage of selected landmarks r^{sel} , while the upper bars (between 75-100%) show the average percentage of observed landmarks r^{obs} .

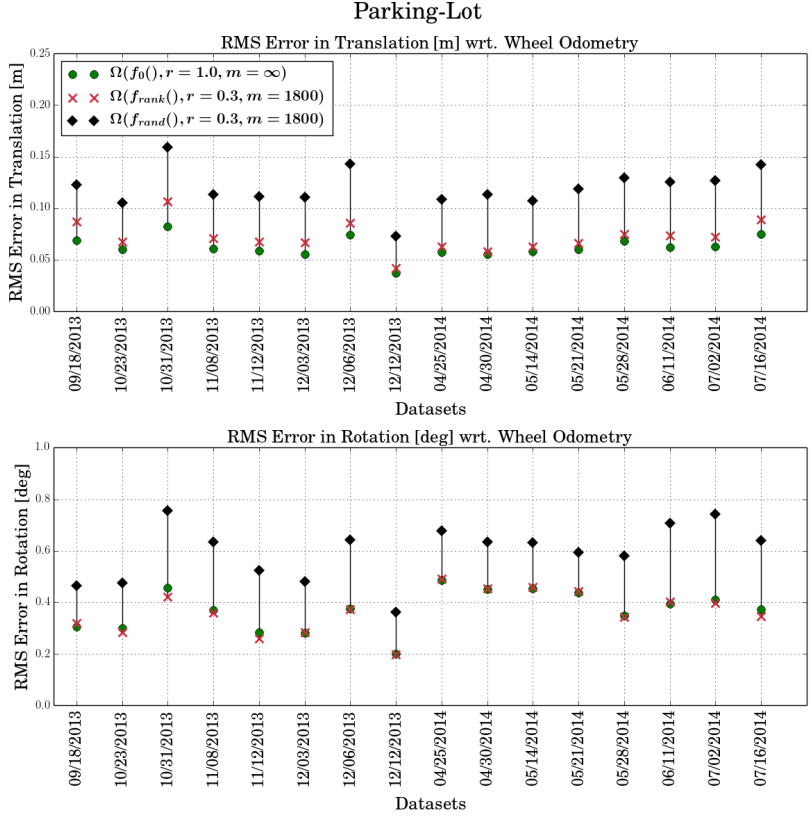


Figure 4.6: RMS error in translation and rotation wrt. wheel odometry for the parking-lot scenario. In green, the RMS error is shown for the case where all landmarks are used, while black diamonds indicate results from random selection, and red crosses for the proposed ranking-based selection method.

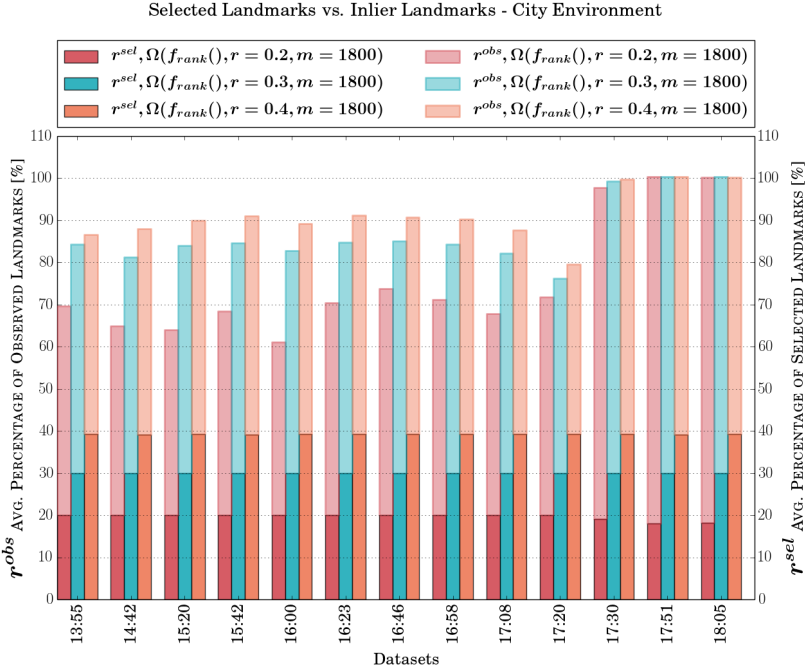


Figure 4.7: Illustration of the relation between the average ratio of selected r^{sel} and the average ratio of observed landmarks r^{obs} for the city environment datasets.

5.3 Parking-Lot Experiments

Figure 4.5 shows the relation between selected landmarks and observed landmarks for the parking-lot experiment. For each dataset, three different sets of selection policy parameters are evaluated, corresponding to more and less strict landmark selection. While on average only 20-30% of the total landmarks are used for localization, the ratio of observed landmarks with and without selection still remains between about 75-100%. For the dataset recorded on April 30th 2014, the average r^{obs} value even lies slightly above 100%. This is due to the fact that by eliminating landmarks inconsistent with the current appearance prior to the 2D-3D matching, the chance of wrong keypoint-landmark associations is reduced, potentially yielding even more observed landmarks in the case of ranking-based selection as compared to if all landmarks are selected.

In addition, figure 4.6 shows the RMS error for translation and orientation for the

three cases of using all landmarks, ranking-based selection, and random selection - the latter two with $r = 0.3$ and $m = 1800$. From this plot, we see that the rotational component is mainly unaffected by the landmark selection, while there is a slight increase in the RMS error for the translational part. In effect, the decrease in the number of observed landmarks results in a slightly less well constrained position estimate, whereas the orientation remains well constrained even with fewer observed landmarks. This is due to the fact, that, for a pure visual pose estimate, the translational component strongly depends on the spatial distribution of observed landmarks, especially on their distance from the vehicle, while the orientation does not. However, the translational RMS error remains significantly lower than for the case of random selection, indicating meaningful landmark selection with the proposed ranking function.

5.4 City Environment Experiments

Figure 4.7 again shows the relationship between the ratio of selected and observed landmarks, this time for the city environment. During daytime, an average observation ratio between 60% and 90% is achieved, depending on the strictness of selection, while at night-time, 100% is reached almost independent of how many landmarks were selected. In contrast to the year-long parking-lot scenario with a high number of varying appearance conditions, in this scenario, we essentially have two very distinct conditions, namely day-time, and night-time, with a far greater total number of landmarks at day-time than at night-time. Therefore, selecting even as much as 40% of the candidate landmarks at day-time may still exclude valid day-time landmarks, simply because of the limited number of selected landmarks. At night-time, the opposite is true, where even a very strict selection of below 20% allows selecting all relevant landmarks under this condition. This effect is also well visible in the RMS error plots in figure 4.8. At day-time, even a random selection performs relatively well, since the day-time landmarks are in vast majority. At night-time, however, our ranking function not only outperforms a random selection, but even achieves slightly better results than when all landmarks are selected. As already mentioned above, this is due to the reduced chance of forming wrong keypoint-landmark associations, allowing to achieve a more robust pose estimate.

5.5 Shared vs. Appearance-Specific Landmarks

In order to demonstrate that our selection method favors different landmarks under different appearance conditions, we evaluate the pair-wise fraction of jointly selected landmarks between two datasets.

The results are depicted in figure 4.9 for the two scenarios and a selection ratio $r = 0.25$ and maximum number of landmarks $m = 1800$.

For the parking-lot scenario, a clear seasonal pattern can be observed, whereas for the city environment, a shift from day-time to night-time landmarks is visible. About 10% of the landmarks selected in any dataset are jointly selected in all

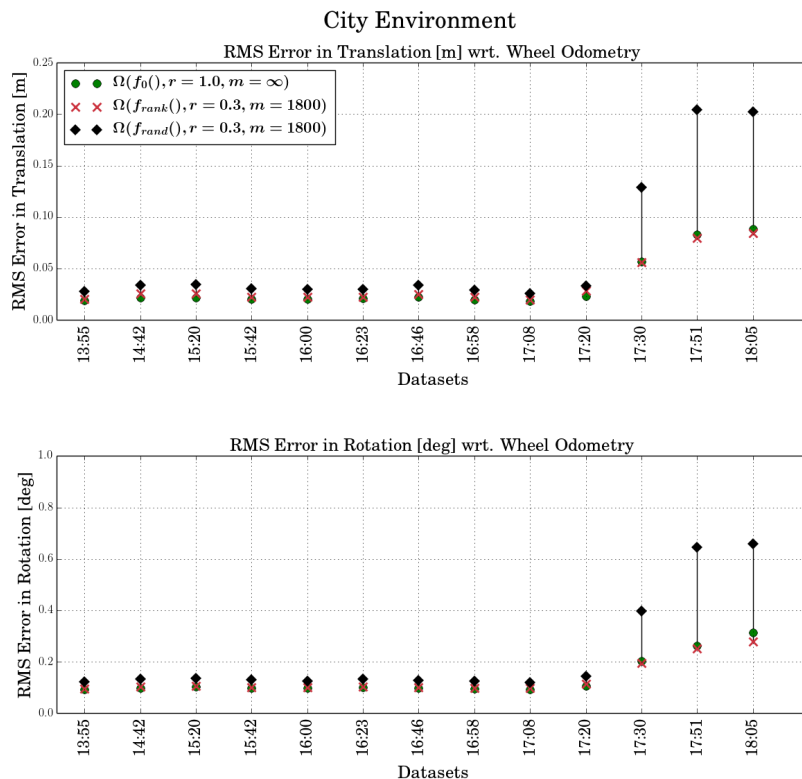


Figure 4.8: RMS error in translation and rotation wrt. wheel odometry for the city environment scenario. Green corresponds to using all landmarks, while the black diamonds indicate results from random selection, and red crosses from the proposed ranking-based selection method.

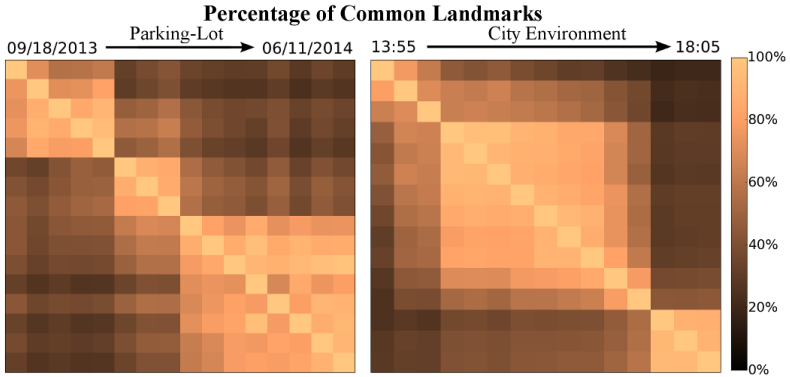


Figure 4.9: Fraction of jointly selected landmarks between individual datasets.

datasets for the parking-lot scenario, whereas this fraction is as low as 2.5% for the city-environment.

6 Conclusion

We have presented an appearance-based landmark selection method for visual localization systems allowing to significantly reduce the data exchange during online operation between a vehicle and a cloud-based map server. Using a simple ranking function, we can distinguish between landmarks that are useful and not useful for localization under the current appearance conditions, using co-observation statistics from previous traversals through the mapped area. The selection method is evaluated in two environments undergoing long-term seasonal and weather change on the one side, and a full transition from day- to night-time on the other side, in combination covering a large extent of possible appearance variations for a visual localization system. The number of landmarks used for localization under a specific appearance condition can be reduced to as little as 30% while still achieving localization performance comparable to when all landmarks are used instead. Importantly, in environments undergoing extreme changes in appearance with a clear association of landmarks to the appearance (e.g. day-time and night-time) a very precise selection is possible, even outperforming the case where all available landmarks are used. However, the results of the day/night experiment further show that defining an appearance-independent number of landmarks to select at each time-step may not adequately account for the potentially very unbalanced number of landmarks useful under a certain appearance condition. Therefore, in future work, more complex selection policies adapting the number of landmarks to select to the prevailing appearance condition ought to be investigated. In addition

to that, more sophisticated appearance outlier detection could further improve the results. Last but not least, extending our appearance-based ranking function with further aspects, such as the spatial distribution and uncertainty of landmark positions, and/or combining it with summary-map techniques such as the ones presented in [65] or [25], could significantly boost the performance and yield better localization accuracy with even fewer selected landmarks.



Appearance-Based Landmark Selection for Long-Term Visual Localization

Mathias Bürki, Cesar Cadena, Igor Gilitschenski, Roland Siegwart and Juan Nieto

Abstract

Visual localization in outdoor environments is subject to varying appearance conditions rendering it difficult to match current camera images against a previously recorded map. Although it is possible to extend the respective maps to allow precise localization across a wide range of differing appearance conditions, these maps quickly grow in size and become impractical to handle on a mobile robotic platform. To address this problem, we present a landmark selection algorithm that exploits appearance co-observability for efficient visual localization in outdoor environments. Based on the appearance condition inferred from recently observed landmarks, a small fraction of landmarks useful under the current appearance condition is selected and used for localization. This allows to greatly reduce the bandwidth consumption between the mobile platform and a map backend in a shared-map scenario, and significantly lowers the demands on the computational resources on said mobile platform. We derive a landmark ranking function that exhibits high performance under vastly changing appearance conditions and is agnostic to the distribution of landmarks across the different map sessions. Furthermore, we relate and compare our proposed appearance-based landmark ranking function to popular ranking schemes from Information Retrieval, and validate our results on the challenging NCLT datasets, including an evaluation of the localization accuracy using ground-truth poses. In addition to that, we investigate the computational and bandwidth resource demands. Our results show that by selecting 20% – 30% of landmarks using our proposed approach, a similar localization performance as the baseline strategy using all landmarks is achieved.

1 Introduction

Visual localization systems are able to provide centimeter-accurate pose estimations of mobile robots with a low-cost sensor setup. This renders visual localization an attractive alternative to LiDAR-based localization which today still require mechanically complex and thus expensive hardware. However, and in contrast to aforementioned LiDAR localization, visual localization systems targeting long-term usage suffer from variations in appearance conditions which render matching between currently observed visual cues and landmarks stored in the map difficult. A promising approach to address this problem has been proposed in the form of multi-session maps [16, 65, 71] that incorporate visual cues from more than one appearance condition. The resulting maps, however, quickly grow in size and become impractical to handle on the mobile robotic platform. In order to mitigate this problem, the map can be stored on a cloud-based backend and made available to the robots in operation over a mobile data network. Apart from relieving the mobile platforms from storing large maps, such a shared-map scenario offers further advantages such as the reduction of redundant data, more efficient map maintenance, and an increased potential for collaboration between the robots. However, it also requires map data to be exchanged between the map backend and the robots in operation over bandwidth constrained mobile data networks. This renders it important to only exchange map data that can be used for localization at the particular time and place of operation. For this purpose, it may be sufficient to only transmit a fraction of map data available in the multi-session map, since the latter must cover all possible appearance conditions, while the robots in operation are exposed to only one condition at a certain time and place. It is the aim of this work to exploit this potential and select landmarks for localization based on the current appearance condition. This serves the following two purposes: a) Keep data exchange between the map backend and the mobile platform, and therewith the bandwidth consumption on a mobile network, as low as possible, and b) lower the computational resource demands on the mobile platform, increasing the real-time capability of visual localization. At the same time, a localization performance as good as if all landmarks are used ought to be maintained. Additionally, the appearance-based landmark selection enables decoupling of the localization performance from map management. While the multi-session map at the backend may be large, and resource intensive to maintain, localization on the vehicles remains as efficient as if only one map session of the current appearance condition was available.

In summary, we present a complete visual localization system yielding $6DoF$ pose estimates at each time-step with the capability to perform efficient online data-association through appearance-based landmark selection.

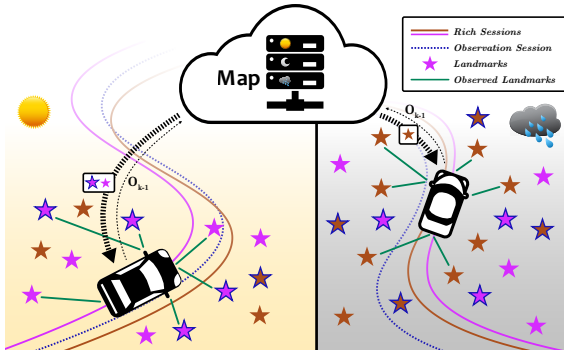


Figure 5.1: Shared-map scenario motivating our work. One large map containing landmarks from multiple *rich*- and *observation sessions* is stored and maintained on a cloud-based map backend. Vehicles en route under different appearance conditions retrieve selected landmarks matching their operation conditions (thick dashed arrow), use those landmarks for visual localization (turquoise lines), and report back a set of recently observed landmark identifiers (thin dashed arrow).

The main contributions of this paper are as follows:

- We derive, analyze and compare a ranking function for appearance-based landmark selection based on appearance equivalence classes, which can be shown to maximize the number of observed landmarks with respect to the current appearance condition.
- We investigate in detail the impact of the incorporation of *observation sessions*, a lightweight extension to the visual maps boosting the landmark selection performance.
- In an extensive evaluation involving three collections of outdoor datasets, one of them publicly available, we thoroughly investigate the performance of the appearance-based landmark selection in real-world conditions, and compare against related popular ranking schemes from Information Retrieval.
- An analysis of the computational performance demonstrates the real-time capability of the appearance-based landmark selection and reveals its potential to reduce the computational load on the vehicle platforms.

This paper builds upon our previous work on appearance-based landmark selection presented in [8, 9] and extends it in several aspects: We derive several appearance-based ranking functions, relate them to popular ranking schemes from Information Retrieval, and evaluate the expected performance of our proposed solution on a related state-of-the-art SLAM framework which keeps separate maps for different appearance conditions. In addition to that, we present an extensive evaluation

on the publicly available *NCLT* dataset collection, including an assessment of the localization *accuracy* with respect to ground-truth. The evaluation on the *NCLT* dataset collection further demonstrates the applicability of our proposed appearance-based landmark selection on a second robotic platform in highly challenging long-term outdoor conditions, and with a considerably different camera system than the one on the vehicle used in the *Parking-Lot* and *City Environment*. A detailed investigation of the computational performance further not only shows the real-time capability of the localization pipeline, but also reveals lower computational resource demands as a second benefit of our proposed appearance-based landmark selection apart from reduced bandwidth consumption.

2 Related Work

Outdoor environments are subject to appearance change, such as change in illumination, as well as change in weather and seasonal conditions. This has a severe impact on long-term operations of outdoor visual localization systems, as in many environments, change in appearance is much more pronounced than structural change, and already with relatively small time offsets of only several hours between mapping and localization it may become difficult to match currently observed visual cues against a visual map. The approaches to overcome this can in general be distinguished into two categories: a) Initiatives to overcome the appearance dependency, and b) attempts to collect and organize appearance-dependent visual features from differing conditions. We first present an overview over relevant work associated with category a), before investigating approach b) in detail in the remainder of this section.

In [41] Lategahn et al. propose a local feature descriptor named DIRD which exhibits illumination invariance superior to other popular local features such as SURF[5] or BRIEF[13]. Nevertheless, the ability to cover appearance change is ultimately still limited in situations with such strong differences in illumination that let already the location of keypoints be different. In another approach to reduce the appearance change in images, Maddern et al. make use of the spectral properties of color cameras in order to apply an illumination invariant gray-scale transformation to images, effectively removing shadows and reducing the appearance variation due to sunlight [18, 54, 55, 69]. This on the one hand requires a photometrically calibrated color camera, and on the other hand is only able to reduce the appearance change due to sunlight. Any other source of appearance change, such as seasonal change, or day-time vs. night-time, are not tackled. In [56], McManus et al. propose to learn location-dependent detectors that retrieve large patches in images deemed descriptive for the respective place. While this shows promising re-detection performance across vastly different appearance conditions, it is not able to allow as precise a metric localization compared to using local corner-based features.

As mentioned above, an alternative approach to tackle the challenge of appearance change lies in the attempt to enrich a visual map with features from varying conditions in order to extent its appearance coverage and allow localization across

a wide range of differing conditions. In [35], Konolige and Bowman present a visual mapping algorithm that is able to aggregate visual cues from different states of the environment into so-called “Views”, which are managed over long time spans. Their system, however, mainly targets structural changes in dynamic indoor environments.

In a similar vein, Milford and Prasser have extended RatSLAM[58] in [73] and [60] to include “Local View Cells” and abstract “Experience Maps” which allow associating previously visited places under varying appearance with the same physical location on the one hand, and the creation and maintenance of a spatially consistent map representations across different environmental states on the other hand. However, the ability to yield a precise metric pose estimate of the robot in a Euclidean coordinate system is limited. In contrast to that, Churchill et al. propose a visual mapping framework called “Experience-Based Mapping” which explicitly creates and maintains separate and detached visual maps for varying outdoor environmental conditions [16]. While this allows precise metric localization under essentially any appearance condition, the visual pose estimate can only be expressed with respect to a Euclidean coordinate system that is unique to each experience. Any interpretation in a common coordinate frame requires links between experiences based on additional sensor modalities, such as (differential) GPS, which may considerably deteriorate the accuracy of the resulting pose estimate. For this reason, attempts have been made to represent visual features - or landmarks respectively - from different appearance conditions in a single Euclidean coordinate frame. Paton et al. present a visual mapping framework able to incorporate and co-relate landmarks from different appearance conditions in outdoor environments with respect to a manually taught reference path [71]. This enables a mobile robot to autonomously repeat the reference route in vastly different appearance conditions. The principle behind the multi-session mapping framework proposed by Mühlfellner et al. in [64, 65] is similar. However, there is no notion of a privileged path, or session respectively, in the map. Instead, the resulting map offers accurate metric localization under any appearance condition represented by the map sessions with respect to a single coordinate frame.

While incorporating landmarks from varying environmental states into a single map can successfully enable visual localization in vastly different appearance conditions, the resulting maps quickly grow in size and become impractical to maintain. Therefore, considerable efforts have been made to optimize map representations such that keeping redundant landmarks is avoided and only a minimal set of landmarks that allow localization across different appearance conditions is maintained. In [21], a long-term short-term memory model is proposed to dynamically distinguish useful from outdated landmarks. Such a model of change is especially suited to environments that exhibit some fraction of features stable in appearance (e.g., corners on the ceiling), but does not have the ability to represent multiple environment states at the same time. In contrast to that, [31, 35] and [57] employ clustering of images, or landmark respectively, in order to keep the number of visual cues bounded. While Konolige and Bowman use a similarity measure between local feature clusters to discard redundant “Views” [35], Hochdorfer et al. remove visual data on the landmark level by assessing the usefulness of individual landmarks inside a local

feature cluster based on position uncertainty [31]. Milford and Wyeth, on the other hand, simply discard landmarks randomly to keep the data density within a cluster bounded [57]. More recent and advanced approaches to bounding the map size for metric visual localization systems are presented in [65], [24]. Mühlfellner et al. compare a number of different algorithms to prune landmarks in a multi-session map, demonstrating selection criteria involving the number of observed sessions, and the total number of observations of a landmark to yield good metric localization performance over long time spans while keeping the map size limited [65]. Along a similar vein, Dymczyk et al. propose to solve an Integer Linear Problem with cost terms favoring landmarks with a large number of observations on the one hand, and guaranteeing a minimal number of landmarks observed from every keyframe on the other hand [24].

In contrast to metric localization, efficient map representations and landmark selection has also been studied in the context of place recognition. In [27] and [81], only the SIFT[48] features contributing the most to the distinctiveness of places are retained in the map. Similarly, in [44], [19], [85], [32] and [33], co-visibility of features is used to efficiently and effectively solve the place recognition problem.

While all of these works describe successful approaches to mitigate the problem of ever-growing visual maps, they only address offline map maintenance with the goal of computing as small a map representation as possible while at the same time maintaining the appearance coverage over different conditions. However, as mentioned in Section 1, in long-term operations in outdoor environments, the map must cover a far wider range of appearance conditions than what the robots in operation require at a given point in time. This offers a potential to further optimize data usage and minimize computational demands on the robot platforms by distinguishing currently useful data based on the observed appearance conditions in an online fashion. In this regard, Linegar et al. [45] have presented an algorithm for the Experience-Based Mapping framework which adaptively selects the best matching “Experience” in an online fashion. While their work addresses a similar motivation as ours, there are substantial differences as a consequence of the different underlying map representation and mapping framework. For instance, the different appearance conditions are represented as individual maps, and therefore their selection of useful map data occurs on the level of “Experiences”. In contrast to that, and due to the fact that our landmarks in the map from the different appearance conditions are all expressed with respect to a single coordinate frame, we are able to select map data matching the current appearance conditions on the level of individual landmarks. In addition to that, we may also select landmarks from more than one session in the map at a time, allowing to benefit from potentially overlapping appearance conditions. In a similar vein, MacTavish et al. propose an online selection of useful map data for their Visual Teach & Repeat framework [52]. Analogous to [45] and in contrast to our work, they perform the selection on the level of “Experiences”, are, however, able to simultaneously use more than one “Experience” for localization. Their work differs further to ours in the methodology at the basis of the selection algorithm. While they compute and compare current images to their map images employing a visual Bag-of-Words representation, we

evaluate the current appearance conformity on the basis of co-observability of recently observed landmarks. This relieves us from having to train and rely on a vocabulary.

3 Background

In this section we briefly introduce the components of our localization and mapping system. This overview supports and facilitates the understanding of subsequent sections in this paper. We first describe the mapping process and the resulting map structure, before presenting our visual localization module in detail.

3.1 Mapping

Mapping is performed in an offline process. We track FREAK[3] features¹ from one camera frame to the next, and triangulate the position of these landmarks using wheel-odometry. With this, a map is generated with a graph of the vehicle’s poses (position and orientation) at image acquisition times, as well as the landmark positions in 3D space. If necessary, loops are closed using the matching-based loop-closure algorithm [79]. Finally, both the poses of the vehicle and the positions of the landmarks are jointly optimized in a Bundle-Adjustment routine.

Further mapping sessions are added by first localizing the new dataset in an offline process against the pre-existing map. This generates both initial pose estimates for the vehicle in the new dataset and associations between features from the camera images of the new dataset and landmarks of the pre-existing map. In addition, new landmarks are spawned from features of the new dataset that failed to find a matching map landmark. Finally, the resulting multi-session map is optimized again with Bundle Adjustment. Note that all information, i.e., both the landmark positions and vehicle poses, of all map sessions, are expressed in the same metric three-dimensional coordinate frame of reference, denoted by $\mathcal{F}\mathcal{F}_W$.

3.2 Localization

The aim of the localization module is to estimate the vehicle’s 6DoF pose with respect to the map coordinate frame of reference $\mathcal{F}\mathcal{F}_W$, given one or more camera images acquired at a specific point in time, and some rough prior knowledge about the current vehicle’s location. We refer to this localization paradigm as local iterative localization, in contrast to global localization or loop-closure where no a priori knowledge of the vehicle’s pose is available.

Let $map := \{V, L, E\}$ denote the map containing a set of vertices V (robot’s poses), a set of landmarks’ positions L , and a set of edges E capturing the observation relation between vertices and landmarks. Let further $\mathcal{F}\mathcal{F}_B$ denote the vehicle body

¹As we demonstrate in the Appendix in Section 6.2, our appearance-based landmark selection algorithm is agnostic to the type of local feature descriptor used. However, in practice, not every descriptor may be equivalently well suited for building multi-session maps, and the choice of descriptor can further be restricted by computational constraints.

coordinate frame. Image acquisitions occur repeatedly along a traversal through the mapped area at a given frequency. Instead of referring to the time of image acquisition, we enumerate them with index k , and refer to the set of images recorded at the k^{th} acquisition with I_k . With this, we can formulate our local iterative localization problem as follows:

$$\tilde{\mathcal{T}}_{WB_k} = \text{localize}(I_k, \hat{\mathcal{T}}_{WB_k}, \text{map}) \quad (5.1)$$

with $\tilde{\mathcal{T}}_{WB_k}$ denoting the estimate of the vehicle’s pose expressed in the map coordinate frame of reference. Analogously, $\hat{\mathcal{T}}_{WB_k}$ denotes the rough prior guess of the same quantity. Using $\hat{\mathcal{T}}_{WB_k}$, landmarks are retrieved from the map that have been observed from near-by, and their respective 3D points are back-projected into the camera image plane, where they are matched against the feature descriptors extracted on the query images based on pixel and descriptor distance. The refined pose estimate $\tilde{\mathcal{T}}_{WB_k}$ is calculated from solving a non-linear least squares optimization problem involving an image-plane projection error constraint with a robust cost function for every keypoint-landmark match. Observations under a pre-defined back-projection error are considered inliers of the localization iteration k , and the respective landmarks form the set of observed landmarks \mathcal{O}_k . The prior guess of the pose for the subsequent localization at iteration $k + 1$ is readily obtained from forward propagating the previous pose estimate with the use of wheel-odometry:

$$\hat{\mathcal{T}}_{WB_{k+1}} := \tilde{\mathcal{T}}_{WB_k} \mathcal{T}_{B_k B_{k+1}}^{odo} \quad (5.2)$$

The main steps of the localization module are summarized in Algorithm!1 in Section!4.

Note that the matching in image space between 2D features and 3D landmarks requires an association of one feature descriptor for every landmark in the map. For our experiments, we group all observations associated with the same 3D point based on their association with the respective *rich session* (see Section 3.3). For every group, we then evaluate the one observation with the smallest accumulated descriptor distance to all other descriptors of the same group, and have the descriptor of this observation, together with respective 3D point, form a landmark used for selection and matching.

3.3 Rich- and Observation Sessions

In section 3.1, we have described how a map can be enriched with landmarks from multiple sessions by localizing a dataset against the map in an offline process. We refer to a dataset added to a map in this fashion as a *rich session*. Adding a *rich session* to a map extends the appearance coverage of the map with the conditions present in the respective dataset. At the same time, however, the size of the map, and the complexity and runtime of the optimization with Bundle-Adjustment is considerably increased.

In contrast to that, a dataset can also be added to the map without the addition of new landmarks. For this, the dataset is localized against the map, and the

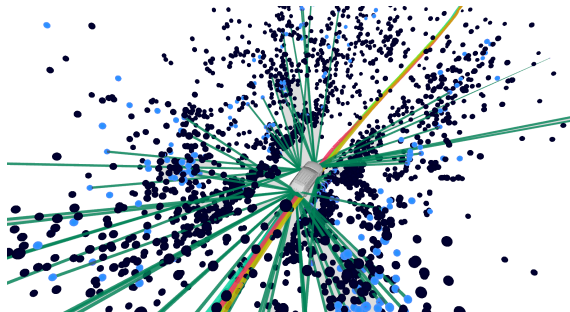


Figure 5.2: Snapshot visualization of our landmark selection. The thick colored lines depict the pose-graph of the map, while the candidate landmarks C are shown as black-, and selected landmarks S as blue spheres. The turquoise lines indicate inlier observations between the four cameras and some of the selected landmarks after the pose refinement step.

vertex poses along the trajectory are added to the pose-graph of the map, analogous to adding a *rich session*. Instead of tracking and triangulating new landmarks, however, only the relation between keypoints from the new dataset and observed pre-existing map landmarks is registered. This barely increases the size of the map and does not have an impact on the complexity of Bundle-Adjustment. Although this does not extend the appearance coverage either, it increases the landmark co-observation statistics, which can be beneficial for the performance of appearance-based landmark selection. A dataset added to the map in this fashion is referred to as an *observation session*.

4 Appearance-Based Landmark Selection

In this section, the selection of landmarks for localization based on appearance is described in detail. After formally presenting the problem at hand, we introduce a landmark ranking function used to prioritize relevant landmarks for the selection process. We conclude this section by relating our problem of appearance-based landmark selection to popular ranking schemes from Information Retrieval.

4.1 Problem Formulation

The goal of appearance-based landmark selection is to decide which of the landmarks in the map are likely to be seen under the present appearance condition. In a generalized manner, this problem can be formulated as follows:

$$S_k = \text{selectLandmarks}(f, C_k, n, \mathcal{A}), \text{ with } S_k \subseteq C_k, \quad (5.3)$$

where C_k denotes the set of geometrically visible candidate landmarks, S_k denotes the set of selected landmarks, f refers to the landmark ranking function, n to the number landmarks to select, and the current appearance condition is expressed as \mathcal{A} . The ranking function f maps a landmark l to a score, i.e.

$$f : l \rightarrow [0, 1], \forall l \in C_k . \quad (5.4)$$

Whereas a landmark is defined as a three-tuple:

$$l := (p_l, d_l, Z_l), \text{ with } Z_l \subseteq Z, \forall l \in L$$

with p_{l_i} denoting the 3D point expressed in the frame of reference \mathcal{FW} , d_l denoting the descriptor associated with landmark l , and Z_l denoting the set of map sessions in which the landmark was observed. The set of all map sessions is denoted by Z .

The set of selected landmarks S_k is formed by applying the ranking function f to every landmark $l \in C_k$, before selecting n top-ranked landmarks. In this work, we choose n to be relative to the number of candidate landmarks available at iteration k , formally expressed as follows:

$$\begin{aligned} U_k &:= \{l \in C_k \mid f(l) > 0\} \\ n &:= \min(\alpha|C_k|, |U_k|), \text{ with } \alpha \in [0, 1] . \end{aligned} \quad (5.5)$$

Pre-selecting the candidate landmarks based on the condition $f(l) > 0$ allows the ranking function to exclude certain landmarks from being selected. This property is used by the ranking function f_{MRS} as described in Section 5.3.

In the following section, we elaborate in detail on how to find a tangible expression for \mathcal{A} and propose a formulation for a ranking function.

4.2 The Ranking Function

The ranking function ought to reflect the probability of successfully forming a match between a map landmark and a feature extracted from the current set of images under the current appearance condition \mathcal{A} .

In order to motivate the formulation for our proposed appearance-based landmark ranking function, we introduce it from a probabilistic perspective. We are thus interested in evaluating the following quantity: $P(l \in \mathcal{O} \mid \mathcal{A})$. This denotes the probability of observing landmark l under the current appearance condition \mathcal{A} . By ranking all candidate landmarks according to this probability, and selecting some number of top-ranked landmarks, we achieve our goal of maximizing the number of observed landmarks.

Ranking Landmarks Based on Appearance Equivalence Classes

Unfortunately, \mathcal{A} is an abstract, intangible entity and not directly observable. However, as every traversal through the environment is related to the particular appearance condition present during that time, all available information regarding

the probability of observing landmark l under some appearance condition \mathcal{A} is encoded in the map session observation relation of landmarks. That is, if l_i and l_j were observed in the same sessions, i.e. $Z_{l_i} = Z_{l_j}$, it can be assumed that

$$P(l_i | \mathcal{A}) = P(l_j | \mathcal{A}) . \quad (5.6)$$

This allows approximation by substituting the current appearance condition \mathcal{A} by the respective set of map sessions a landmark has been observed in, i.e.

$$P(l \in \mathcal{O} | \mathcal{A}) \approx P(l \in \mathcal{O} | Z_l) . \quad (5.7)$$

This renders the conditioning on the appearance condition tangible, as the observing map session relations of landmarks are well-defined and countable. Note that we employ a common abuse of notation by interpreting the expression $P(l \in \mathcal{O} | Z_l)$ as the probability of observing landmark l , given it has been observed in the past in the map sessions Z_l . We can thus group all landmarks into distinct equivalence classes, and model the observation likelihood with a simple Bernoulli distribution, i.e.

$$P(l \in \mathcal{O} | Z_l) \sim \text{Ber}(\theta^{[l]}), \text{ with } [l] := \{l_j \in L | Z_{l_j} = Z_l\} \quad (5.8)$$

It remains to estimate the appearance dependent parameters $\theta^{[l]}$. For this, we employ the principle of local temporal stability of appearance conditions: Whenever the mapped area is traversed, the appearance conditions are expected to change along the route in the same manner as they have in previous traversals. Following this principle, we thus expect to again observe the same landmarks together with those that have already in the past been co-observed. This allows to compute a Maximum Likelihood Estimate for $\theta^{[l]}$ using recently selected and observed landmarks from previous localization iterations. For this, we add subscript k to refer to localization iteration k , as described in Section 3.2:

$$\theta_k^{[l]} = P(l_o \in \mathcal{O}_k | l_o \in [l]) = \frac{P(l_o \in \mathcal{O}_k, l_o \in [l])}{P(l_o \in [l])} \approx \frac{|\mathcal{O}_{k-1}^{[l]}|}{|S_{k-1}^{[l]}|}, \text{ with} \quad (5.9)$$

$$\mathcal{O}_{k-1}^{[l]} := \{l_o \in \mathcal{O}_{k-1} | l_o \in [l]\}, \quad (5.10)$$

$$S_{k-1}^{[l]} := \{l_s \in S_{k-1} | l_s \in [l]\} \quad (5.11)$$

We can interpret this quantity as the estimated relevance of appearance equivalence class $[l]$, based on recently collected statistical samples. With a limited budget of landmarks to select, prioritizing the selection according to this ranking function maximizes the number of expected observed landmarks under the current appearance condition. Note, however, that this statement of optimality only refers to the selection of landmarks based on appearance. There are further non-appearance related effects (e.g., geometry, occlusion, etc.) having an impact on whether a landmark is observed or not.

For our experiments, we use a temporal smoothing of $\theta^{[l]}$ over the $N = 50$ most recent iterations and define our ranking function accordingly:

$$f_{AEC}(l) := \frac{1}{N} \sum_{w=0}^{N-1} \theta_{k-w}^{[l]} \quad (5.12)$$

4.3 Relation to Information Retrieval

In this section, we relate our proposed appearance-based landmark ranking approach to common concepts in the field of Information Retrieval. With this, we aim at providing further theoretical context and facilitating the understanding and interpretation of the ranking function described in equation 5.12.

The principles of Information Retrieval are usually stated in a linguistic context, where the overall goal is to retrieve a set of text documents most relevant to a given search query consisting of a set of query words [77]. Analogous to appearance-based landmark selection for visual localization, a ranking function is required, which assigns a relevance score to each document in the collection, according to how well the document matches the query words. It has thereby proven to be most successful to take two distinct aspects of relevance into consideration when assessing the relevance of a query word to a document. The *term frequency* aspect reflects how well a given query term represents the given document, while the *inverse document frequency* aspect attempts to reflect the overall discriminatory power of a word with respect to the entire document collection. These two aspects form two separate terms, whose product is assigned as the relevance weight of a query word with respect to a document. The overall ranking score can readily be computed either by summing over all relevance weights, or by representing the relevance weights in vector form and employing cosine similarity [77]. The result is the well-known *tf-idf* ranking scheme. Drawing the analogy with appearance-based landmark selection, we can interpret recently observed landmarks as the query. This allows expressing the appearance-based ranking function f_{AEC} described in equation 5.12 as follows:

$$tf(l_o, l) := \begin{cases} 1 & \text{if } [l_o] = [l], \\ 0 & \text{otherwise} \end{cases}, \quad idf(l, S_{k-1}) := \frac{1}{|S_{k-1}^{[l]}|} \quad (5.13)$$

$$f_{AEC}(l) = \sum_{l_o} tf(l_o, l) idf(l_o, S_{k-1}) \quad (5.14)$$

A unary *term frequency* only considers query landmark relevant if they belong to the same appearance equivalence class. The *inverse document frequency* term downweights contributions of landmarks if a large quantity of landmarks from the same appearance equivalence class have recently been selected. We note, however, that this interpretation of the *idf* term deviates from the text-book definition. This is because in the context of appearance-based landmark selection, we are rather interested in weighting the query words in relation to the set of recently

selected landmarks, as opposed to the set of candidate landmarks. We further note that there are countless variations in how to formulate tf and idf terms in order to achieve optimal retrieval performance in a given application [2, 77]. In Section 5.3, we introduce further sensible formulations that we compare against in our experiments.

Algorithm 1 Iterative Local Localization. The retrieval of nearby vertices from the pose-graph employs a distance δ and yaw angle discrepancy ϕ around the pose guess $\hat{\mathcal{T}}_{WB_k}$.

```

1: function LOCALIZE( $I_k, \hat{\mathcal{T}}_{WB_k}, map, \mathcal{O}_{k-1}$ )
2:    $K \leftarrow \text{extractFeatures}(I_k)$ 
3:    $V_k \leftarrow \text{retrieveNearbyVertices}(\hat{\mathcal{T}}_{WB_k}, \delta, \phi, map)$ 
4:    $C_k \leftarrow \text{getLandmarksObservedFromVertices}(V_k, map)$ 
5:    $S_k \leftarrow \text{selectLandmarks}(C_k, \mathcal{O}_{k-1}, f)$ 
6:    $M \leftarrow \text{match2D3D}(K, S_k)$ 
7:    $\hat{\mathcal{T}}_{WB_k}, \mathcal{O}_k \leftarrow \text{estimatePose}(M, \hat{\mathcal{T}}_{WB_k})$ 
8: end function

```

An overview of the localization with appearance-based landmark selection in pseudo-code can be seen in Algorithm 1.

5 Evaluation

In this section, we present the results of our evaluation, focusing on a) demonstrating the effectiveness of selecting landmarks using the appearance-based ranking function presented in Section 4.2 in multiple challenging long-term outdoor environments, b) comparing our proposed ranking function with related popular ranking schemes, c) reporting on the resulting localization precision and accuracy, and c) analyzing the computational performance of the respective localization algorithm.

In order to facilitate the navigation within and reading of this section, we first present a concise summary of the conducted experiments. Subsequently, the dataset collections, respective sensor configurations, and evaluation metrics are introduced, before the various experiments are presented in detail. A paragraph containing our key findings concludes the evaluation section.

Please note that a direct comparison of our appearance-based landmark selection performance with the most related works [45, 52] is inherently difficult, as the underlying mapping framework and visual feature representations are fundamentally different, and the selection of relevant data on the level of individual landmarks constitutes a unique feature of our method. With the ranking function f_{MRS} , as introduced in Section 5.3, we aim at comparing our method with the performance that is to be expected with an ‘‘Experience-Based’’ mapping framework, which creates and maintains separate maps for each map-session. In addition to that, comparisons of the localization performance with selecting landmarks randomly,

and with the localization performance using all landmarks, serve as lower- and upper-bounds for properly assessing the effectiveness of our proposed landmark selection on the one hand, and the extent of saving mobile network bandwidth on the other hand. We further assess and compare the selection performance with various ranking schemes inspired by the *tf-idf* concept in Information Retrieval.

In order to keep the evaluation section as concise as possible, we prefer to present metrics aggregated over all datasets of the respective dataset collection. However, the interested reader is kindly invited to study the graphs showing the performance on each dataset separately in the Appendix in Section 6.1.

5.1 Experiments Overview

Our experiments can be divided into four groups as follows.

Rich Sessions Only

We first investigate the effectiveness of the proposed appearance-based landmark selection and the resulting localization precision with maps containing only *rich sessions*. This allows us to restrict the landmark selection to select from at most one *rich session* at any localization iteration along the trajectory with the ranking function f_{MRS} , as described in Section 5.3. It corresponds to the localization performance attainable with mapping frameworks that keep separate maps for every session, such as the Experience-Based mapping framework by Churchill and Newman [16]. In reverse, it shows the benefit in localization precision achievable in a multi-session mapping framework as the one used for this work, which expresses all landmarks from all sessions in a common reference coordinate frame and thus allows selecting landmarks from more than one *rich session* at the same time. The respective experiments can be found in Section 5.5, Figures 5.4 and 5.5.

Observation Sessions

With the presence of *observation sessions*, the selection performance of different appearance-based landmark ranking functions exhibit more pronounced variance. Therefore, the experiments in this section aim at analyzing these differences in performance and relate them to the varying environmental conditions. Note that since *observation sessions* span across multiple *rich sessions*, the ranking function f_{MRS} is no longer properly defined and is thus not included in these experiments. The experiments can be found in Section 5.6, Figures 5.6, 5.7 and 5.8.

Localization Accuracy

The *NCLT* dataset collection provides ground-truth pose estimates. This allows us to evaluate the localization accuracy along the trajectories of all *NCLT* datasets. Apart from yielding an absolute estimate of the localization accuracy achieved by the different selection policies and landmark ranking functions, we can further investigate and validate the relation between the localization accuracy and other

performance influencing metrics such as the distance from the map trajectories, or the number of observed landmarks. The respective experiments can be found in Section 5.7, and Table 5.5. Furthermore, two special phenomena are analyzed in detail in two case studies in Figures 5.10 and 5.11.

Computational Performance Analysis

The potential to significantly reduce the computational requirements on the vehicle side constitutes - apart from a reduction in mobile network bandwidth consumption - a second strong incentive to employ the proposed appearance-based landmark selection. In order to support this claim, we have measured and analyzed the computational costs involved for the different components of our visual localization pipeline, both with, and without appearance-based landmark selection. The respective results are presented in Section 5.8 and Figure 5.12.

5.2 Dataset Collections

The selection of datasets for evaluating the performance of the proposed appearance-based landmark selection has been driven by the following main criteria: a) The dataset collection ought to cover a wide range of varying appearance conditions, with still sufficient appearance overlap allowing to build a multi-session map. b) The sensor set-up must include an odometry sensor, which we require for the forward propagation of the pose states in our iterative localization pipeline. c) Ideally, the dataset collection offers ground-truth poses, which enable an evaluation of the localization accuracy. Many popular publicly available dataset collections fail to meet these criteria. With the *NCLT* datasets, however, there exists a dataset collections offering all features relevant for us. Furthermore, the appearance conditions covered by the *NCLT* datasets are diverse and very challenging, with changing weather conditions, often a setting sun, or strong shadows in the field of view. They thus provide an ideal settings for putting the different appearance-based landmark ranking functions through their paces.

We extend the evaluation with two self-collected datasets, named *Parking-Lot* and *City Environment*. Similar to the *NCLT* datasets, the *Parking-Lot* datasets cover long-term appearance change during day-time. The respective sensor set-up and platform dynamics differ, however, which adds further variation to the evaluation scenarios. In contrast to the *NCLT* and *Parking-Lot* datasets, the *City Environment* datasets cover a very specific scenario of appearance-change, namely that of the change from day-time to night time.

NCLT

In the *The University of Michigan North Campus Long-Term Vision and LIDAR datasets* [15], a *Ladybug 3* camera is used, together with wheel-odometry from the Segway platform. All images are undistorted and down-scaled to dimensions of $808px \times 616px$ in order to be comparable in resolution to the images recorded in

the *Parking-Lot* and *City Environment* collection respectively. The 27 datasets from the *NCLT* collection were recorded between January 2012 and April 2013 on the north campus of the Michigan University in Ann-Arbor. The route and direction of traversal followed during the individual recordings, however, varies considerably between the different datasets. For the purpose of this evaluation, we have extracted an approximately 750m long segment of the routes that has been traversed in all datasets, except the one recorded on January 10th 2013. Furthermore, the dataset from December 1st 2012 has been excluded from the evaluation since it comprises the only night-time recording. Due to a lack of any recordings from transitioning conditions at dusk or dawn, it is not possible to extend the appearance coverage of the map to an extent that would allow proper localization at night-time. The traversing direction of all the remaining datasets is the same, except for the recordings from February 4th 2012, November 4th 2012 and February 23rd 2013 which traverse the mapped area in opposite direction. These datasets can be successfully localized, even though the respective precision and accuracy are worse compared to the other datasets.

Parking-Lot

The *Parking-Lot* datasets cover a circular traversals of a car on a open space parking lot. A total of 28 datasets recorded between August 2013 and July 2014 cover a vast variety of different weather and seasonal conditions during day-time. Among others, they include low-standing sun, rain and wet snow, as well as scattered clouds and clear skies.

City Environment

In order to cover the extreme change in appearance from day-time to night-time, 23 drivings in a *City Environment* have been recorded during the course of a day, starting around noon, and ending around 6pm in the evening. While the weather condition across these datasets remains static, illumination undergoes drastic change from diffuse daylight to night-time with artificial street lighting.

The sensor set-up used in the *Parking-Lot* and *City Environment* datasets consists of four fish-eye cameras mounted on a car (facing front, left, rear and right), and wheel-odometry. Images are recorded at 12.5hz in gray-scale at a resolution of 640px x 400px.

An overview over the weather conditions, the usage of each dataset in the corresponding multi-session maps, as well as example images for all three dataset collections can be found in the Appendix in Table 5.6, 5.7, and 5.8. More sample images of the *Parking-Lot* datasets can be found in [65], and in [15] for the *NCLT* datasets.

5.3 Ranking Functions

Before presenting the metrics and experimental results, we introduce additional ranking functions used for comparison and as baselines in the evaluation.

We employ localization with the following pseudo ranking function and selection fraction $\alpha = 1.0$ as a baseline to evaluate the performance of our proposed appearance-based ranking functions:

$$f_0(l) := 1 \quad \forall l \in C, \alpha = 1.0 \quad (5.15)$$

This corresponds to using all landmarks in the candidate set C for localization, and in general serves as an upper-bound for the performance of any other ranking function with $\alpha < 1.0$.

As a lower-bound for the performance of landmark selection, we further compare against selecting landmarks randomly:

$$f_{random}(l) := v, v \sim \mathcal{U}[0, 1] \quad (5.16)$$

In addition to that, we also compare the performance of the appearance-based ranking functions introduced in Section 4.2 to the performance of the following ranking function:

$$f_{MRS}(l_i) := \begin{cases} 1, & \text{if } p([l_i] | \mathcal{A}) = \max_{[l]}(p([l] | \mathcal{A})) \\ 0, & \text{otherwise} \end{cases} \quad (5.17)$$

This ranking function selects at most $n = \alpha|C|$ landmarks observed from the *rich session* with currently the best conformity with the encountered appearance condition. While switching the selection of landmarks from one *rich session* to another is allowed along the traversal, selecting landmarks from more than one *rich session* for a specific localization iteration is prohibited. It thus demonstrates the localization performance attainable with separate maps from each *rich session*, in contrast to having all landmarks and observer vertices expressed in one common coordinate frame of reference.

We further include the following appearance-based ranking functions introduced in [8] in our comparison:

$$f_{NCV}(l) := \frac{1}{|Z_l|} \sum_{z \in Z_l} |\mathcal{O}_{k-1}^z| \quad (5.18)$$

$$\text{with } \mathcal{O}_{k-1}^z := \{l \in \mathcal{O}_{k-1} \mid z \in Z_l\} \quad (5.19)$$

It corresponds to a normalized voting-based ranking. Every landmark observed in the previous localization iteration casts a vote for each of its observing sessions. In order to prevent landmarks observed from multiple map sessions to always dominate over landmarks observed from fewer or only one map session, the accumulated votes are normalized by the number of map observer sessions.

In addition to that, we compare our proposed appearance-based ranking functions with different variations of the *tf-idf* ranking scheme used in Information Retrieval.

$v_l := [x_i] \in \mathbb{R}^{ Z_l } \text{ with } x_i = \begin{cases} 1, & \text{if } z_i \in Z_l \\ 0, & \text{otherwise} \end{cases}$ $v_q := \sum_{l \in \mathcal{O}_{k-1}} v_l$ $f_{AV}(l) := \text{cosine}(v_l, v_q) \forall l \in C$	$tf(l, z)$	$\begin{cases} 1, & \text{if } z \in Z_l \\ 0, & \text{otherwise} \end{cases}$
	$idf()$	1.0

The ranking function f_{AV} uses a vector space representation for landmarks with a binary tf term representing the observing map session relation. A *cosine* similarity metric is employed as the ranking score.

$w(l, l_o) := \sum_{z \in Z_{l_o}} tf(z, l)idf(z, S_{k-1})$ $f_{TFIDFA}(l) := \sum_{l_o \in \mathcal{O}_{k-1}} w(l, l_o)$	$tf(l, z)$	$\begin{cases} 1, & \text{if } z \in Z_l \\ 0, & \text{otherwise} \end{cases}$
	$idf(z, C)$	$\log(\frac{ C }{ C^z })$, with $C^z := \{l \in C \mid z \in Z_l\}$

Ranking function f_{TFIDFA} follows an analogy with text document retrieval where landmarks are interpreted as documents containing words in the form of observing map sessions. The *multi-set* of query words is built from all observing map sessions from the set of recently observed landmarks. Further, a standard *inverse document frequency* term is employed, down-weighting the contribution of map sessions frequently present in the observing map sessions of the candidate landmarks.

$r(z) := \sum_{l \in \mathcal{O}_{k-1}} tf(l, z)idf(l, Z)$ $f_{TFIDFB}(l) := \sum_{z \in Z_l} r(z)$	$tf(l, z)$	$\begin{cases} 1 & \text{if } z \in Z_l \\ 0, & \text{otherwise} \end{cases}$
	$idf(l, Z)$	$\log(\frac{ Z }{ Z_l })$

In contrast to f_{TFIDFA} , the ranking function f_{TFIDFB} first attempts to rank map sessions, instead of directly ranking landmarks. For this, roles are switched, and map sessions are interpreted as documents, containing words in the form of landmarks observed in the respective session. The set of recently observed landmarks \mathcal{O}_{k-1} forms the set of query words, upon which the map sessions are ranked, following a standard *tf-idf* scheme. The ranking score for a candidate landmark is eventually

formed as the sum of the respective observing session relevances.

$w(l, z) := tf(l, z)idf(z, S_{k-1})$ $r(z) := \sum_{l \in \mathcal{O}_{k-1}} w(l, z)$ $f_{WRS}(l) := \sum_{z \in Z_l} r(z)$	$tf(l, z)$	$\begin{cases} 1 & \text{if } z \in Z_l \\ 0, & \text{otherwise} \end{cases}$
	$idf(z, S_{k-1})$	$\frac{1}{ S_{k-1}^z }$ with $S_{k-1}^z := \{l \in S_{k-1} \mid z \in Z_l\}$

Ranking function f_{WRS} is defined similar to f_{AEC} , evaluates, however, the relevances of individual map sessions $r(z)$, instead of appearance equivalence classes. Analogous to f_{TfIdfB} , a sum over all observing session relevances is used as the final ranking score of a candidate landmark.

5.4 Metrics

An informative measure for the quality of the ranking function is the comparison between the number of observed landmarks using only some percentage of selected landmarks, and the number of observed landmarks using all landmarks for localization at a given iteration k . This ratio is denoted r_k^{obs} and formally defined as follows:

$$r_k^{obs} := \frac{|\mathcal{O}_k^{f, \alpha}|}{|\mathcal{O}_k^{f_0, \alpha=1.0}|} \quad (5.20)$$

An ideal ranking function f achieves an observation ratio r_k^{obs} close to 1.0 with a selection fraction α as low as possible. This would indicate that only landmarks currently observable receive a high score and are selected, whereas unobservable landmarks receive a low score and are discarded.

The *NCLT* dataset collection provides ground-truth poses based on fused and globally optimized pose estimates computed from the 3D LiDAR scans and the differential GPS sensor measurements. We make use of this in order to evaluate the accuracy of the localized poses within a local neighborhood of the map [7]. For every localization iteration along a traversal of a dataset, we compare the transformation between the pose resulting from solving our visual localization optimization problem, $\tilde{\mathcal{T}}_{WB_k}$, and the pose of the nearest vertex in the map, \mathcal{T}_{WnnV_k} , with the transformation between the ground-truth pose for the current image, \mathcal{T}_{GB_k} , and the ground-truth pose of the same nearest vertex in the map, \mathcal{T}_{GnnV_k} . This results in the following formulation for the local error transformation:

$$\mathcal{T}_{nnVB_k}^W = \mathcal{T}_{WnnV_k}^{-1} \tilde{\mathcal{T}}_{WB_k} \quad (5.21)$$

$$\mathcal{T}_{nnVB_k}^G = \mathcal{T}_{GnnV_k}^{-1} \mathcal{T}_{GB_k} \quad (5.22)$$

$$\mathcal{T}_{LEGT_k} := \mathcal{T}_{nnVB_k}^{W-1} \mathcal{T}_{nnVB_k}^G \quad (5.23)$$

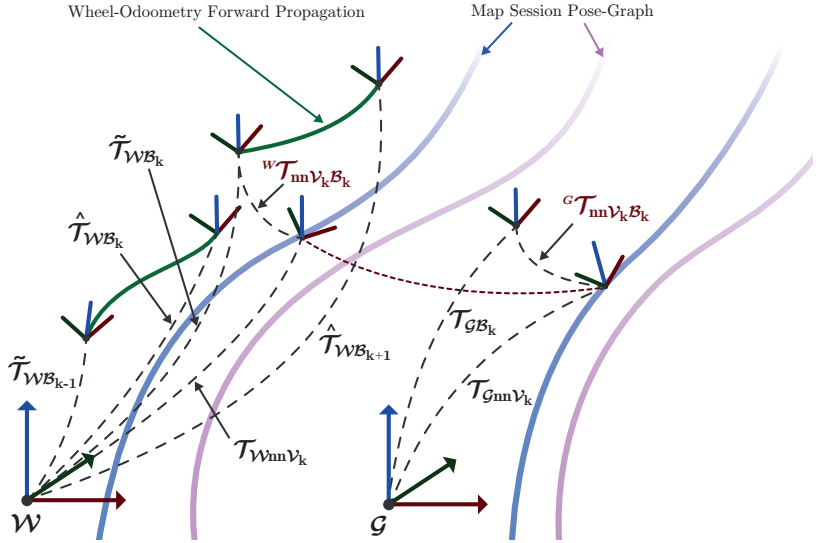


Figure 5.3: The two coordinate systems \mathcal{FW} and \mathcal{FG} and all relevant transformations used for the calculation of the local localization precision with respect to the wheel-odometry, and the local localization accuracy with respect to the ground-truth solution.

All involved transformations are schematically depicted in Figure 5.3.

Apart from the inaccuracy of the visual localization, there are further sources of errors affecting \mathcal{T}_{LEGT} , such as a) inherent inaccuracies of the ground-truth transformations, b) time synchronization, c) sensor intrinsics and extrinsics calibration, d) scale and space distortions between the two involved coordinate systems \mathcal{FW} and \mathcal{FG} , and e) inconsistencies in the pose-graph of the visual map. The effect of the distortion between the involved coordinate systems is almost entirely mitigated by employing local errors as described above. In order to eliminate any errors due to inconsistencies in the pose-graph of the visual map, we optimize the poses of the *NCLT* maps with an additional prior constraint linked to the ground-truth transformation closest in time. The inherent inaccuracies of the ground-truth solution are expected to be considerably lower than the localization accuracies from the visual localization system, as the former is computed from a globally optimized SLAM solution using the 3D LiDAR scans and differential GPS, with all datasets cross-registered, and a manual removal of wrong loop-closure constraints [15]. As a consequence, we expect \mathcal{T}_{LEGT} to reflect primarily the (in-)accuracy of the visual localization.

The local-error transformation \mathcal{T}_{LEGT}_k is further decomposed into the corre-

sponding three dimensional translation and rotation vector, denoted by p_{LEGT_k} , and a_{LEGT_k} respectively.

In the cases of the *Parking-Lot* and *City Environment* datasets, no ground-truth solution is available. Since in each localization iteration, a visual-only pose optimization problem is solved (see Section 3.2) we can still assess how well the resulting pose estimate is constrained along a dataset by computing statistics on the transformations between the pose estimates, and the respective pose guess of the same iteration:

$$\mathcal{T}_{LEO_k} := \hat{\mathcal{T}}_{WB_k}^{-1} \bar{\mathcal{T}}_{WB_k} \quad (5.24)$$

We refer to this as localization precision, as opposed to localization accuracy as described in equation 5.23. In addition to the error induced by the visual localization itself on the current and previous pose estimate, the local error transformation \mathcal{T}_{LEO_k} also contains the local drift of the wheel-odometry in-between. However, the latter is expected to be at least one magnitude smaller, leaving the magnitude of \mathcal{T}_{LEO_k} to be dominated by the visual localization errors.

5.5 Rich Sessions Only

We first present the ratios of observed and selected landmarks, as well as the precision results, for all three dataset collections, whereas the presented values are aggregated over all datasets of the respective collection.

In Figure 5.4, the relation of observed vs. selected landmarks is shown for selection fractions between 10% and 40%. Since there is a significant discrepancy in the observation percentage during day-time as opposed to at night, we further show the observation percentage in the *City Environment* aggregated over day-time datasets, that is, up and including 17:30, and over the remaining night-time datasets, separately. In addition to that, Figure 5.5 shows a comparison of the localization precision.

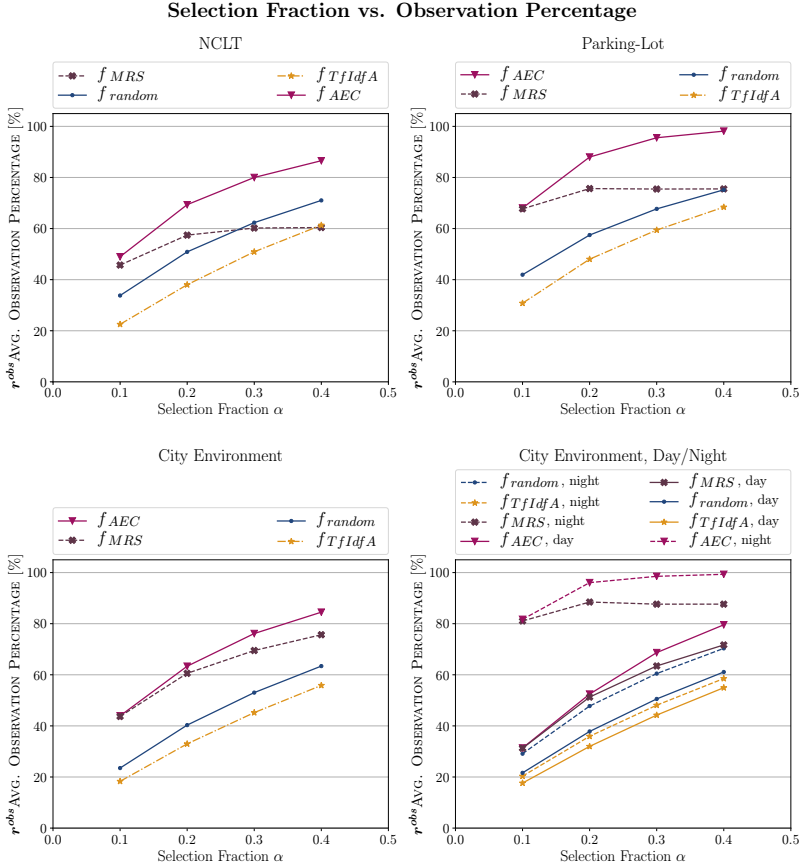


Figure 5.4: The average observation percentage r_{obs}^{avg} in relation to the selection fraction α for different choices of ranking functions, and for all three dataset collections against maps containing only *rich sessions*. In the *City Environment*, datasets are further split up into day-time datasets (up until 17:30), and night-time datasets.

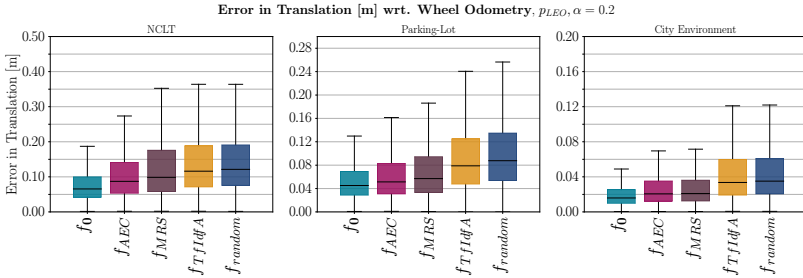


Figure 5.5: The aggregated localization translation precision for all three dataset collection against the map containing only *rich sessions*. The following ranking functions are shown: Localization using all landmarks, f_0 , $\alpha = 1.0$, appearance-based landmark selection with f_{AEC} , f_{TfIdfA} , f_{MRS} , and random selection with f_{random} , all with a selection fraction of $\alpha = 0.2$.

Note that in this scenario with a map containing only *rich sessions*, it is straightforward to see that the ranking score with the appearance-based ranking functions f_{AEC} , f_{WRS} , f_{AV} , and f_{TfIdfB} is identical. We therefore only show the results for the ranking function f_{AEC} .

In all three environments, ranking landmarks with f_{AEC} yields a consistently high observation percentage, and localization precision close to the one achieved using all landmarks. In contrast to that, ranking landmarks using f_{TfIdfA} fails, as it yields a consistently lower observation percentage than random selection, and precision values considerably worse than the other ranking functions. The *idf* term of f_{TfIdfA} follows the text-book definition of *inverse document frequency*, thus down-weighting the influence of map sessions if there are many candidate landmarks observed in the respective session. As described in Section 4.3, this criteria does not well reflect the appearance conformity of a landmark, and instead tends to favor map sessions with only few landmarks.

We further note the performance limitations of f_{MRS} . With low selection fractions, the attained observation percentage is on the same level as other well-performing ranking functions, such as f_{AEC} , f_{AV} , and f_{TfIdfB} . However, the restriction to only select from one *rich session* results in performance saturation for larger selection fractions. The respective loss in precision is clearly visible in case on the *NCLT* and *Parking-Lot* datasets at a selection fraction of 20%, and demonstrates one of the benefits of having all landmarks, even from multiple *rich sessions*, registered in one common coordinate frame of reference. Note that this loss in precision is less pronounced on the *City Environment* datasets, as in this case, there are only few different appearance conditions represented in the map, with a clear best-matching *rich session* at any time.

It can further be observed that the overall observation percentage and localization precision in the *NCLT* environment is lower as compared to the *Parking-Lot*, despite

both environments reflecting long-term day-time conditions. This discrepancy suggests that there is a larger difference in encountered appearances over the year in relation to the number of *rich sessions* in the map in the *NCLT* scenario as compared to the *Parking-Lot* scenario. Precision in the *NCLT* environment further pays tribute to the fact that the trajectories in the *NCLT* datasets often do not follow the exact same route and exhibit lateral offsets of up to 12m. This renders the visual localization considerably more challenging as opposed to the *Parking-Lot* scenario, where there is a quite precisely repeated driving pattern on the parking lot.

In the *City Environment*, the comparatively low observation percentage is attributed to the fact that there are fewer diverse appearance conditions covered, resulting in a lower number of *rich sessions* present in the map. During day-time, where there are considerably more landmarks than at dusk and night-time, even a selection fraction of up to 40% may not be sufficient to select all landmarks matching the current appearance condition. This effect is supported by the graph in Figure 5.4 showing the observation percentage aggregated separately over day-time and night-time datasets. All ranking functions, including random selection, exhibit a significantly higher observation percentage at night as opposed to during day-time. The increase at night is, however, most pronounced for the appearance-based ranking functions f_{AEC} , and f_{MRS} .

5.6 Observation Sessions

Adding *observation sessions* to the map can further improve the performance of the appearance-based landmark selection at a negligible additional map storage or computational burden.

It can be seen in Figure 5.6 and Figure 5.7 that for low selection fractions, the best observation percentage and localization precision is achieved using the proposed f_{AEC} ranking function. In addition to that, ranking functions f_{AV} and f_{TfIdfB} also achieve favorable performance for low selection fractions, and even achieve the best observation percentage on the *NCLT* and *Parking-Lot* datasets for higher selection fractions. We attribute this phenomena to saturation effects with higher selection fractions on the one hand, and to dataset specific artifacts, such as the *lock-in* effect discussed in Section 5.7 on the other hand. Both may undermine the theoretical optimality of f_{AEC} under ideal conditions.

In addition to that, the ranking function f_{NCV} exhibits the highest variance in performance. As can be seen in Figure 5.6, and Figure 5.8, this ranking function fails during day-time, yielding an observation percentage worse than that of random selection. It further performs poorly in general for low selection fractions on the other two dataset collections too. However, for high selection fractions, the opposite is the case, and appearance-based landmark selection with f_{NCV} achieves the best performance on the *NCLT* and *Parking-Lot* datasets.

We further observe ranking functions f_{AV} and f_{TfIdfB} perform very similarly in all three environments. This is remarkable, as the respective expressions of the ranking functions are considerably different.

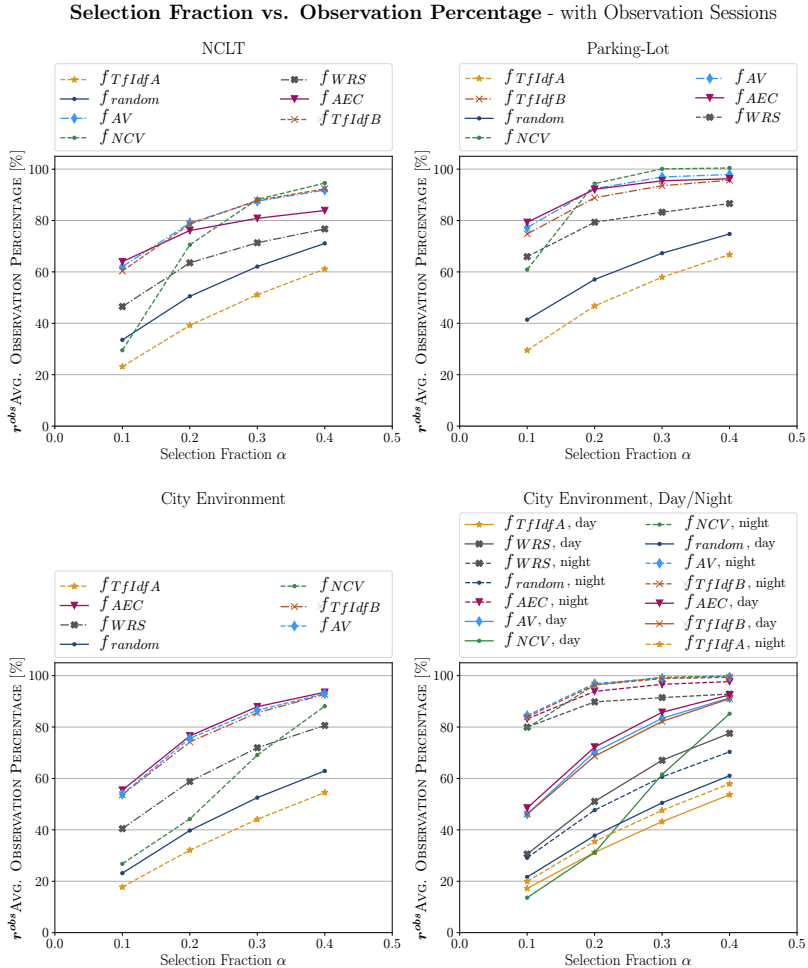


Figure 5.6: The average observation percentage r^{obs} in relation to the selection fraction α for different choices of ranking functions, and for all three dataset collections against maps containing *observation sessions*. In the *City Environment*, datasets are further split up into day-time datasets (up until 17:30), and night-time datasets.

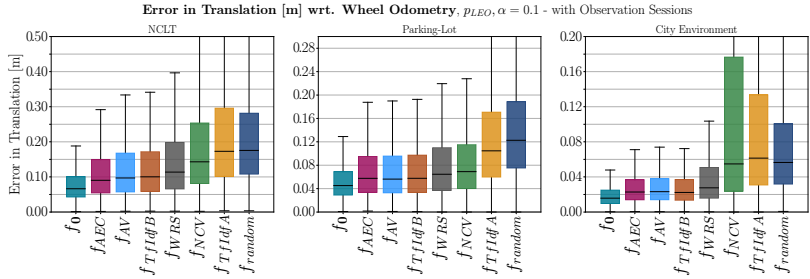


Figure 5.7: The aggregated localization translation precision for all three dataset collection against the map containing *observation sessions*. The following ranking functions are shown: Localization using all landmarks, f_0 , $\alpha = 1.0$, appearance-based landmark selection with f_{AEC} , f_{AV} , f_{TfIdfB} , f_{NCV} , f_{TfIdfA} , and random selection with f_{random} , all with a selection fraction of $\alpha = 0.1$.

Similar as with maps containing only *rich sessions*, the ranking function f_{WRS} outperforms random selection, but falls short of any of the before mentioned appearance-based ranking functions. The presence of *observation sessions* further is not able to improve the poor performance of f_{TfIdfA} .

The benefit of ranking based on appearance equivalence classes is most pronounced in the *City Environment* at dusk around 17:25, as can be seen in Figure 5.8. Despite this being only one dataset, it is exemplary for a general phenomena: The heavy bias in the number of landmarks towards day-time *rich sessions* lets most other ranking functions preferably select landmarks from day-time. However, night-time landmarks would, although fewer in absolute numbers, already yield more inlier observations relative to the number of selected night-time landmarks. Only the two ranking functions f_{AEC} and f_{WRS} are able to exploit this and achieve almost 20% more landmark observations in this case. The ranking function f_{WRS} , however, suffers from sub-optimal performance during day-time, leaving the ranking function based on appearance equivalence classes as the only one with high observation percentage all the time.

Before elaborating on the localization accuracy evaluation in the subsequent section, we summarize the key findings of the different ranking function’s performance. On all three dataset collections, a significant boost in observation percentage by the use of *observation sessions* is well visible. In addition to that, the ranking function f_{AEC} exhibits the best performance at low selection fractions, while the performance of f_{NCV} is most susceptible to the selection fraction, performing poorly for low selection fractions, but even outperforming all other ranking functions by a small margin at a selection fraction of 40%.

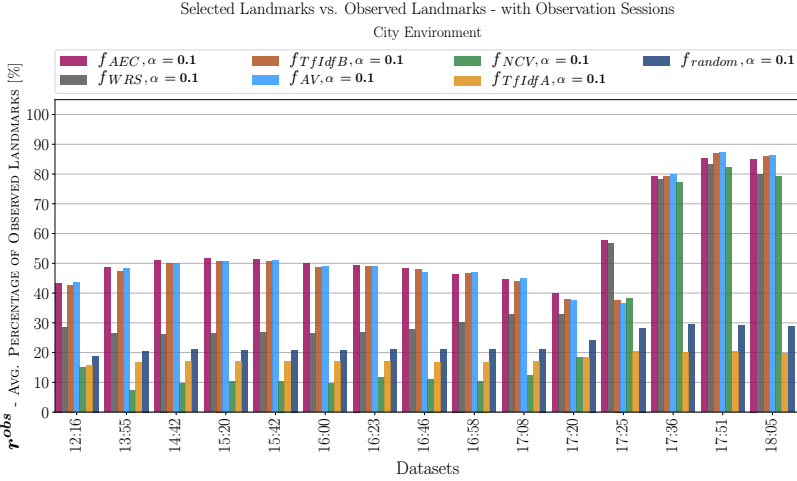


Figure 5.8: The average observation percentage r^{obs} for different choices of ranking functions and a selection fraction $\alpha = 0.1$, for all datasets of the *City Environment* against the map containing *observation sessions*.

5.7 Localization Accuracy

In this section, we present the visual localization accuracy results of the *NCLT* datasets using the ground-truth solution based on 3D LiDAR and differential GPS as a reference.

The median errors in translation of \mathcal{T}_{LECT} , denoted by p_{LECT} are listed in Table 5.5 for localization using all landmarks, f_0 , as well as appearance-based selection with the ranking functions f_{AEC} , f_{NCV} , f_{MRS} , and random selection, f_{random} ; all with a selection fraction of 10%. Furthermore, the translation accuracy using the ranking function f_{AEC} is listed both when localizing against the map containing only *rich sessions*, as well as when localizing against the map containing both *rich sessions* and *observation sessions*. For the latter, the ranking function is denoted by “ f_{AEC} os.”.

We first note that the median translation accuracy of the reference localization using all landmarks exhibits a rather large span, ranging from 11cm up to 44cm. This is on the one hand due to the varying trajectories of the respective datasets. On the other hand, not every appearance condition encountered in the datasets used for evaluation is equally well covered by the *rich sessions* in the map, resulting in differing visual localization performance. The most important factor for deteriorated localization performance, however, is the direction of traversal, resulting in the datasets from November 4th 2012 and February 23rd 2013 to perform considerably

Table 5.5: The translation localization accuracy for the *NCLT* datasets, using the ground-truth poses as a reference. The columns show the translation median error in meters for the following six ranking functions: Localization using all landmark ($f_0, \alpha = 1.0$), appearance-based landmark selection with f_{AEC} , f_{NCV} , and f_{MRS} , and random selection with f_{random} . For appearance-based and random selection, a selection fraction $\alpha = 0.1$ is used. All ranking function localize against the map containing only *rich sessions*, except for $f_{AEC, os}$ which localizes against the map containing both *rich sessions* and *observation sessions*.

Date	Median $pLEGT[m], \alpha = 0.1$					
	$f_0(\alpha = 1.0)$	f_{AEC}	$f_{AEC, os}$	f_{NCV}	f_{MRS}	f_{random}
8th January, 2012	0.155	0.2	0.181	0.2	0.206	0.247
15th January, 2012	0.215	0.317		0.35	0.363	0.329
2nd February, 2012	0.14	0.196	0.145	0.222	0.203	0.187
12th February, 2012	0.179	0.263	0.211	0.268	0.266	0.3
18th February, 2012	0.18	0.247	0.182	0.28	0.287	0.268
19th February, 2012	0.129	0.191		0.215	0.211	0.191
17th March, 2012	0.129	0.196	0.155	0.235	0.204	0.212
25th March, 2012	0.265	0.304	0.271	0.301	0.294	0.329
31st March, 2012	0.112	0.189		0.197	0.196	0.163
29th April, 2012	0.141	0.196	0.161	0.202	0.203	0.216
26th May, 2012	0.121	0.151	0.135	0.154	0.161	0.181
4th August, 2012	0.139	0.159	0.157	0.16	0.158	0.233
28th September, 2012	0.124	0.18	0.149	0.259	0.194	0.202
28th October, 2012	0.138	0.201		0.251	0.236	0.219
4th November, 2012	0.308	0.396	0.363	0.458	2.282	0.41
17th November, 2012	0.174	0.215	0.185	0.219	0.24	0.253
23rd February, 2013	0.445	0.485		0.544	0.524	0.477
5th April, 2013	0.168	0.225	0.188	0.218	0.237	0.253

worse than any other dataset.

We further observe that the accuracy using the appearance-based ranking function f_{AEC} on the map containing only *rich sessions* slightly outperforms selecting landmarks based on f_{NCV} , but both perform significantly better than using f_{MRS} for landmark selection. This again demonstrates the gain in performance due to the ability to select landmarks from more than one *rich session* at a time.

In addition to that, there is a clearly pronounced boost in localization accuracy when using the map with additional *observation sessions* for localization, with landmark selection based on f_{AEC} even achieving accuracy values close to those of the reference localization using all landmarks for certain datasets.

It is noticeable, however, that random selection often achieves accuracy values close to those of appearance-based selection, at least in the case of using the map with no *observation sessions*. In this regard, we notice that the random selection of landmarks occurs for every localization iteration along the trajectory. Despite only selecting 10% at each iteration, even after short traversals of a few meters, almost

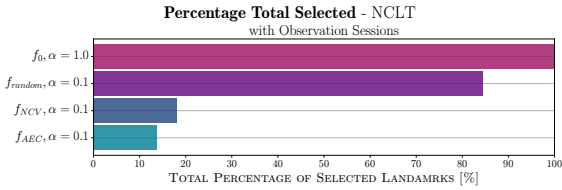


Figure 5.9: The total percentage of unique landmarks selected over the course of the entire trajectory of a dataset. This percentage directly conforms to the overall savings in data transmissions between a map backend and a mobile robot in a shared-map scenario. While appearance-based selection only uses a percentage of landmarks approximately equivalent to the respective selection fraction in each iteration, random selection makes use of almost all landmarks at least once along the trajectory.

all landmarks available in the vicinity of the respective map segment have been selected at least once by f_{random} . This effect is well visible in Figure 5.9 which displays the total number of selected landmarks for each of the different ranking functions and selection policies. While all appearance-based ranking functions only select a fraction of all landmarks across the entire dataset trajectory approximately equal to the selection fraction at each iteration of 10%, random selection selects up 85% of all landmarks in the map. For this reason, localization using random selection may be considerably less precise, but its accuracy is not in the same extent worse compared to both appearance-based landmark selection and localization using all landmarks.

In addition to that, the challenging route selection of the *NCLT* datasets lead to varying localization performance along a trajectory of a specific dataset. Even though the influence of outliers in the localization performance on the overall accuracy is limited by our choice of the median error, the magnitude of the latter is often still heavily influenced by short segments of the trajectory with very poor localization performance. In order to render these effects more tangible, we investigate the localization accuracy in relation to the number of observed landmarks and the distance to the nearest vertex in the map in detail for the two datasets of January 8th and February 2nd 2012.

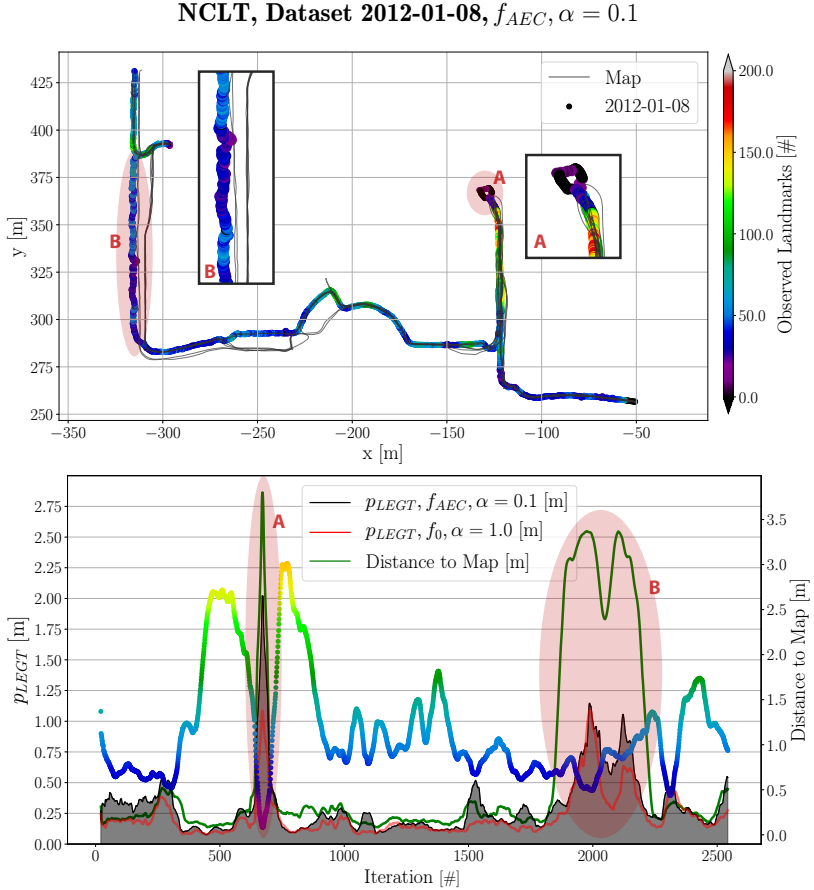


Figure 5.10: Bird’s eye perspective onto the mapped segment of the *NCLT* scenario, together with a temporal analysis of the number of observed landmarks, the translation localization accuracy, and the distance to the nearest vertex in the map, referred to as “Distance to Map”. On the left-hand side, the trajectories of the *richsessions* in the map are drawn in light gray, while the localized poses of the dataset from January 8th 2012 is drawn in color indicating the number of observed landmarks along the route. Two particular situations along the trajectory of this dataset are marked by capital letters “A” and “B” and analyzed in detail in Section 5.7.

The left-hand side of Figure 5.10 shows a birds-eye perspective of the mapped segment used in the *NCLT* scenario. The trajectories of the *rich sessions* present in the map are drawn in gray, whereas the trajectory of the dataset being localized from January 8th 2012 is drawn in color indicating the number of landmark observations along the trajectory. In contrast to that, the bottom half depicts the relation between time - or iteration index respectively - on the one hand, and the number of observed landmarks, the localization accuracy, and the distance to the nearest vertex in the map on the other hand.

The trajectories from all *rich sessions* in the map follow up and down a long aisle between iteration 300 and 900. While the trajectory from January 8th in general follows the same pattern, the turning point at the back of the aisle occurs a few meters farther into the aisle compared to the trajectories present in the map. This situation is marked with the letter “A” in Figure 5.10. While the distance to the nearest vertex in the map suddenly increases from approximately 30cm up to almost 3m, the number of observed landmarks drops to almost zero. At the same time, the localization accuracy is greatly reduced, both in case of the reference localization with f_0 , as well as and considerably more severely in the case of appearance-based localization. In this regard, situation “A” also serves as a good example for the strong correlation between the number of observed landmarks and the localization precision and accuracy respectively.

In situation “B”, the number of observed landmarks is only slightly lower than in the preceding trajectory segment. The distance to the map, however, is considerably increased, since the trajectory of the dataset travels along the street instead of on the parallel sidewalk as in all the *rich sessions* in the map. This again results in a peak degradation of the localization accuracy and demonstrates the correlation between the distance to the map, and the localization performance.

The dataset from February 2nd exhibits even slightly lower localization accuracy with the map containing only *rich sessions* for appearance-based landmark selection with f_{AEC} compared to performing random selection with f_{random} . As can be seen in Figure 5.11, the errors in localization are mainly attributed to two peaks, again marked with a capital letter “A” and “B”.

In situation “A”, the trajectory from February 2nd directly crosses the street, while all the datasets used for building the map take a right turn up and down the aisle. This leads to a sudden increase in the distance to the nearest vertex in the map, and as a result of that a simultaneous drop in the number of observed landmarks and the respective localization accuracy.

In contrast to that, the peak drop in localization accuracy in situation “B” originates from a *lock-in* effect inherent to the presented appearance-based landmark selection. In order to understand the cause of this peak drop, we point out that the relevance of the different appearance equivalence classes is evaluated based on recently observed landmarks. The latter themselves are a subset of recently selected landmarks. This does not constitute a problem as long as the availability of landmarks from all *rich sessions* in the map along the trajectory is maintained, and the appearance conditions encountered do not show any abrupt change that is not also reflected in the respective *rich-* or *observation sessions*. In the *NCLT* scenario,

NCLT, Dataset 2012-02-02, f_{AEC} , $\alpha = 0.1$

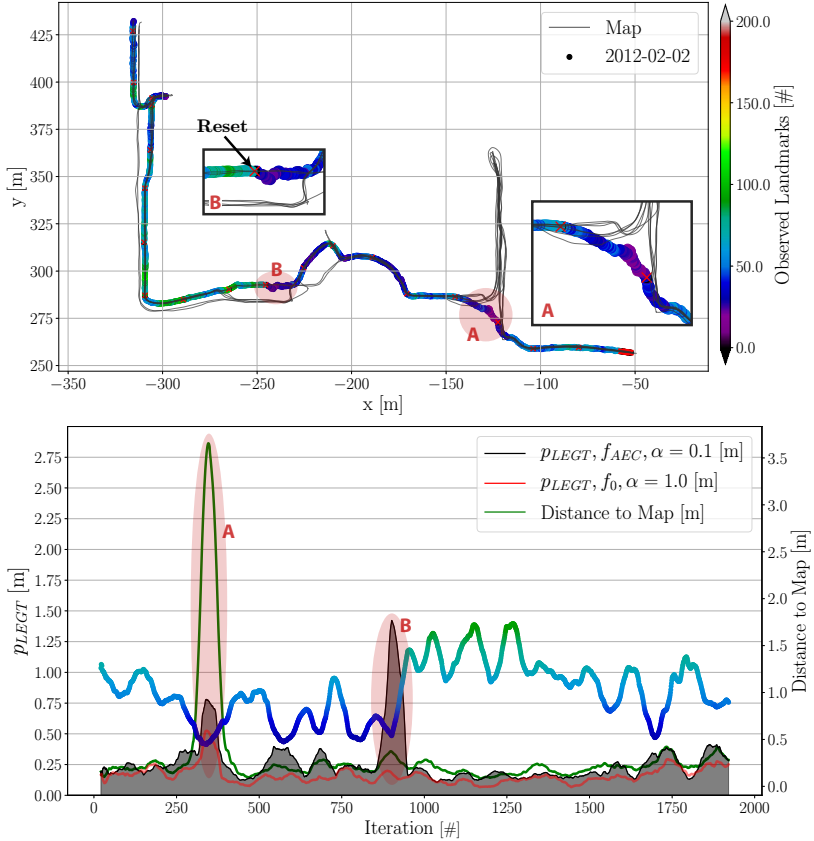


Figure 5.11: Bird’s eye perspective onto the mapped segment of the *NCLT* scenario, together with a temporal analysis of the number of observed landmarks, the translation localization accuracy, and the distance to the nearest vertex in the map, referred to as “Distance to Map”. On the left-hand side, the trajectories of the *richsessions* in the map are drawn in light gray, while the localized poses of the dataset from February 2nd 2012 is drawn in color indicating the number of observed landmarks along the route. Two particular situations along the trajectory of this dataset are marked by capital letters “A” and “B” and analyzed in detail in Section 5.7.

however, the trajectory segment between iteration 600 and 1100 is characterized by the different datasets taking varying routes. Thus, the *rich session* with the best appearance conformity may at once exhibit a large lateral offset, or not be available at all temporarily, resulting in the number of observed landmarks to decrease and the localization accuracy to drop. In order to recover from this *lock-in* situation, the appearance-based landmark selection can be “reset”. For this, all candidate landmark are used for localization of a single iteration, allowing to properly re-evaluate the relevance of all available appearance equivalence classes. For the presented experiments, such a “reset” is set to occur at every 100th iteration, and its effect is clearly visible in situation “B”: After the “reset”, the number of inliers swiftly increases from approximately 20 up to 100, and the respective localization accuracy recovers. In practice, it is advisable to link the triggering of “resets” to a metric reflecting the condition of poor localization performance in situations where localization is expected to perform reasonably well (e.g., when the assumed location is close to the mapped trajectories). Such a metric is, however, application and use-case specific.

Apart from the exemplary situations mentioned and discussed in detail above causing the localization accuracy to degrade, Figure 5.10 and 5.11 also indicate that under normal circumstances, that is, with the localized trajectory and all map *sessions* following the same route, a localization accuracy of around 10cm is achieved, which is in accordance with the results found in [65].

5.8 Computational Performance Analysis

We conclude the evaluation section by analyzing the computational time spent on the major blocks of our localization pipeline. All computations have been carried out on a Lenovo W530 with an Intel i7 CPU, and without the use a GPU. In addition to the incentive of lower data bandwidth usage, the computational performance analysis reveals a second benefit of appearance-based landmark selection in the form of reduced computational demands on the mobile platform side. Figure 5.12 shows the execution times of the following building blocks:

- **Feature Tracking:** The time to extract keypoints and compute FREAK descriptors on all involved cameras.
- **Landmark Retrieval:** The time to retrieve all near-by pose-graph vertices, and their observed landmarks.
- **Landmark Ranking:** The time to apply f on all candidate landmarks.
- **Landmark Selection:** The time to select n top-ranked candidate landmarks, yielding S_k .
- **Landmark Back-Projection:** This step involves the look-up of the landmark descriptor for each selected landmark, and the back-projection of the landmark’s 3D point into the camera image planes, using the pose guess \tilde{T}_{WB_k} .

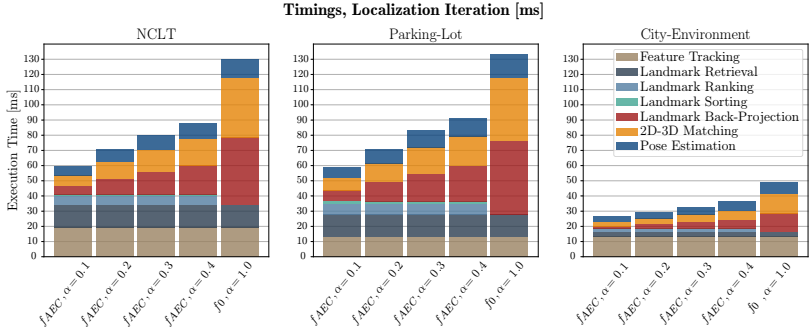


Figure 5.12: The execution times of the individual building blocks of our localization pipeline, including the modules specific to appearance-based landmark selection. While the overall execution times vary in the different scenarios, in all cases localization with appearance-based landmark selection is able to perform significantly faster compared to localization without landmark selection. This is due to the fact that the time invested into landmark ranking and sorting is more than compensated by the time saved in back-projecting and matching of the landmarks in the subsequent modules of the localization pipeline.

- **2D-3D Matching:** The formation of associations between the FREAK features in the current camera images, and the back-projected map landmarks.
- **Pose Estimation:** Refinement of the pose estimation employing a non-linear least-squares optimization problem, yielding \tilde{T}_{WB_k} .

Apart from the feature tracking, the overall execution times in the *City Environment* are by more than a factor 2 lower. This is due to the fact that there are only 4 *rich sessions* in the map, with only one from bright day-light. Thus, the resulting number of candidate landmarks being retrieved in each iteration is considerably lower as compared to the *NCLT* datasets.

We further note that the feature tracking, the landmark retrieval, and the pose estimation step all have to be carried out both in case of localization with, as well as without appearance-based landmark selection, and that their running time is mostly independent of the number of selected landmarks. Nevertheless, these blocks are included in Figure 5.12 in order to give a comprehensive overview over the running time and real-time capability of the localization pipeline.

The computational differences between localization with and without appearance-based landmark selection can be summarized as follows: In contrast to localization without landmark selection, localization with landmark selection has to invest time in ranking and sorting the candidate landmarks. In exchange, however, considerably fewer landmarks have to be back-projected and matched against the features in the image, reducing the runtime of these modules in relation to the respective selection

fraction. As can be seen in Figure 5.12, in both scenarios, even with a selection fraction of 40%, the total runtime of appearance-based selection is considerably lower than without landmark selection. The computational load on the mobile platform side is even further reduced in a shared-map scenario as motivated in Section 1, where ranking and sorting of candidate landmarks is carried out on the backend side.

From the accumulated running times in the *NCLT* scenario, it can be deduced that localization with appearance-based landmark selection is able to run at $10 - 15Hz$, while localization without landmark selection may not be able to exceed $8Hz$. Only accounting for the modules running on the mobile platform in case of a shared-map scenario, namely feature-tracking, landmark back-projection, matching and pose estimation, the resulting difference in runtime performance is increased to $15 - 30Hz$ with landmark selection, as opposed to only $10Hz$ without landmark selection.

6 Conclusions

In this section, we summarize our key findings and draw conclusions for the use of appearance-based landmark selection in practice.

At first, we note that substantial differences in the camera set-up in the *NCLT* datasets, such as the lack of fish-eye distortion, does not have a significant effect on the performance of the appearance-based landmark selection. Similar as with the *City Environment* and the *Parking-Lot* datasets, an appearance-based selection of 20% to 30% of the available landmarks allows achieving a localization performance similar to using all landmarks.

Furthermore, we have analyzed in detail the performance of several appearance-based landmark ranking function in combination with maps with, and without *observation sessions*. Selecting landmarks using the proposed f_{AEC} ranking function yields the best performance, especially for low selection fractions. However, other formulations for the ranking functions, most notably f_{AV} and f_{TfidB} , achieve favorable performance too. This observation, together with the independence with respect to the distribution of landmarks in map sessions, let f_{AEC} be the ranking function of choice in general.

With the *lock-in* effect observed on the dataset example depicted in Figure 5.11, we have analyzed and described a potential pitfall inherent to the use of appearance-based landmark selection. In practice, an application and use-case specific monitoring of the observed localization performance in relation to what performance is to be expected is pivotal in order to swiftly detect a *lock-in* situation and initiate a “reset” of the appearance-based landmark selection. Easily trackable metrics, such as the number of observed landmarks, and the distance from the nearest vertex in the map, may serve as potent indicators to distinguish *lock-in* situation from poor localization due to too large divergence from the mapped territory. This suggestion is supported by the strong correlations between the aforementioned metrics and the localization performance, as shown in Figure 5.10 and 5.11.

The localization accuracy achieved in the *NCLT* scenario is in general in accordance with the respective precision, although the magnitude of the former is slightly higher. This is attributed to the fact that there are more sources of error involved, such as the error of the ground-truth solution itself, and inaccuracies of the intrinsic and extrinsic sensor calibrations. It is in this regard important to note again that the pose estimated in each iteration is computed from solving a non-linear least squares optimization problem only containing constraints between the image keypoints and the matched 3D landmarks from the map. In particular, there is no temporal smoothing or sensor fusion, which would prevent immediate degradation of accuracy in many situations where temporarily only few landmark are observed.

In a detailed computational performance analysis, we have shown that our localization pipeline with appearance-based landmark selection is able to run in real-time. Furthermore, the use of appearance-based landmark selection significantly lowers the computational demand on the mobile platform, as only a fraction of landmarks have to be processed in each localization iteration.

Appendix

We present the observation percentage and localization precision separately for each dataset of the three dataset collections in Section 6.1. This relates to Section 5.5, and Section 5.6, where the same metrics are shown in aggregated form.

Furthermore, we compare the localization performance with different choices of feature descriptors on the *Parking-Lot* datasets. The respective results can be found in Section 6.2.

We conclude the appendix with a list of all datasets used in this evaluation, the respective weather conditions, and some sample images in Section 6.3.

6.1 Individual Dataset Performance Analysis

In Figure 5.13, the observation percentage is shown with ranking function f_{AEC} for selection fractions of 10% – 40% with maps containing only *rich sessions*. It can be observed that for certain datasets of the *Parking-Lot* collection, the average number of observed landmarks with a 30% or 40% selection fraction can even exceed the average number of observed landmarks when using all candidate landmarks. This exhibits a saturation effect, resulting in occasionally achieving a higher number of observed landmarks with only a subset of selected landmarks, as opposed to using all candidate landmarks. While counter-intuitive at first, this is due to the fact that including more candidate landmarks increases the chance of forming wrong 2D-3D matches. After the subsequent pose estimation step, these wrong matches are then classified as outliers, resulting in a potentially lower number of observed landmarks.

Furthermore, the different observation percentage characteristics during day-time as opposed to at night are clearly visible in the *City Environment*. During the day, a selection of 40% of the landmarks is not sufficient for an observation percentage of more than 90%, while at night-time, even 20% of selected landmarks achieve almost an observation percentage of 100%.

The localization precision using different ranking functions and with a selection fraction of 20% are shown for each dataset of all three collections in Figure 5.14, 5.15, and 5.16 respectively. The results reflect the patterns visible in Figure 5.13, and in Section 5.5. The best performance is achieved using f_{AEC} , f_{AV} and f_{TfIdfB} for ranking landmarks, with precision values often close to that of using all landmarks for localization instead. While the precision using f_{MRS} can vary considerably between different datasets, ranking functions f_{TfIdfA} fails, resulting in occasionally even worse precision than selecting landmarks randomly.

Enriching the maps with *observation sessions* results in a higher variance of performance between different ranking functions, as can be seen in Figure 5.17 for the *NCLT* and *Parking-Lot* collection, and in Section 5.6 in Figure 5.8 for the *City Environment*. The respective localization precision results are shown in Figure 5.18, Figure 5.19, and Figure 5.20. Most notable is the failure of the ranking function f_{NCV} during day-time in the *City Environment*. As discussed in Section 5.6, ranking function f_{AEC} is the only one achieving consistently high localization precision in the *City Environment* both during day-time, at dusk, as well as at night-time.

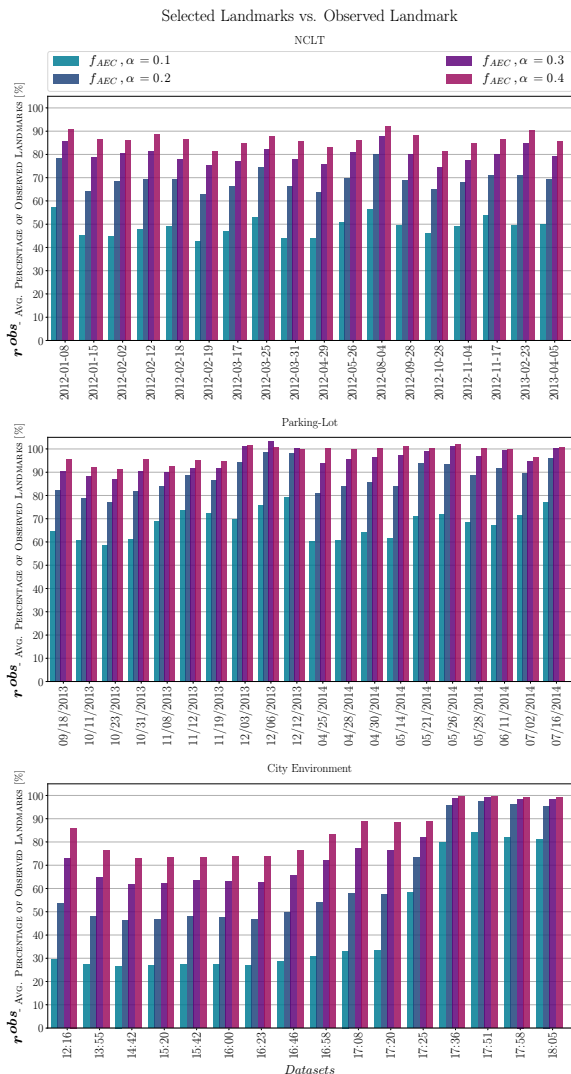


Figure 5.13: The average observation percentage r^{obs} for selection fractions between 10% and 40%, for every dataset of the *NCLT* (top), *Parking-Lot* (middle), and *City Environment* dataset collection against maps containing only *rich sessions*.

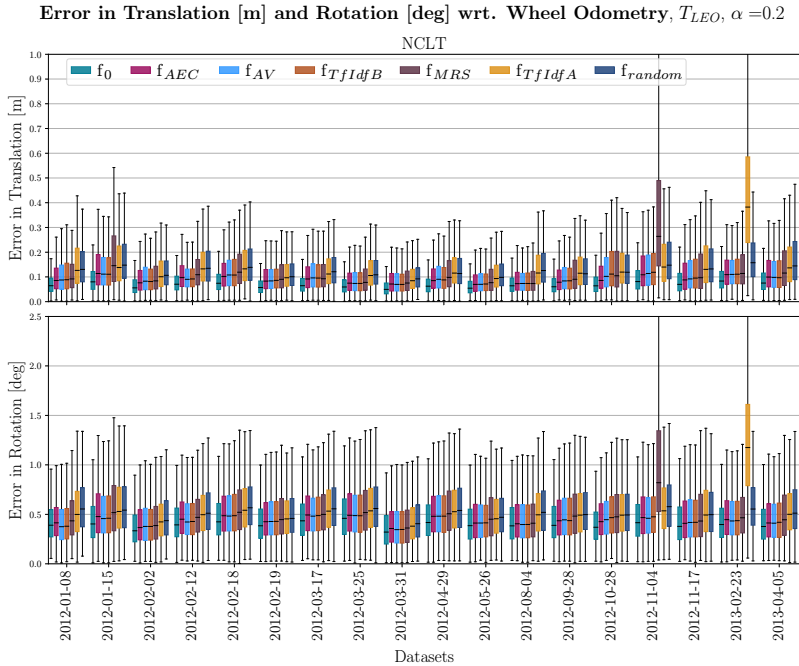


Figure 5.14: The localization precision for a selection fraction of 20%, for every dataset of the *NCLT* collection against the map containing only *rich sessions*.

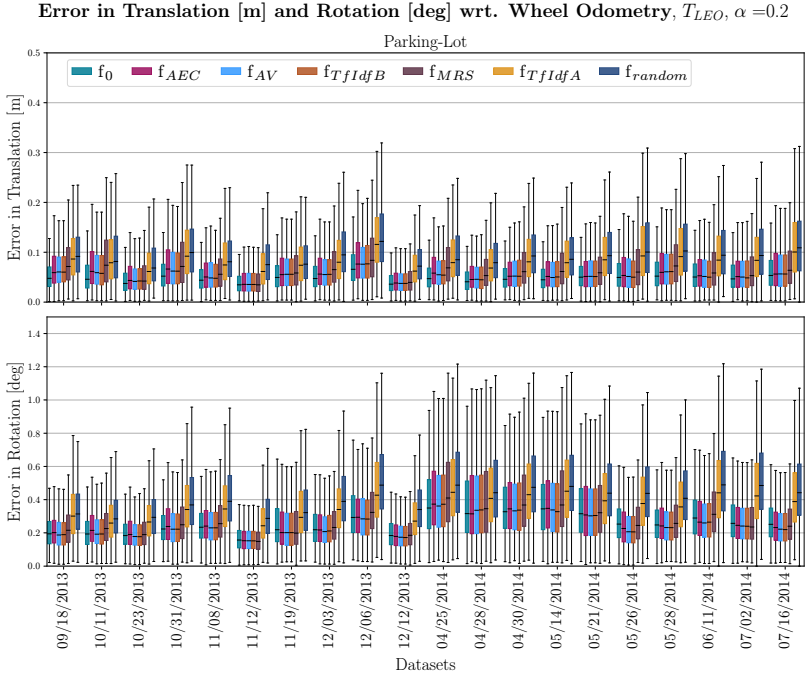


Figure 5.15: The localization precision for a selection fraction of 20%, for every dataset of the *Parking-Lot* collection against the map containing only *rich sessions*.

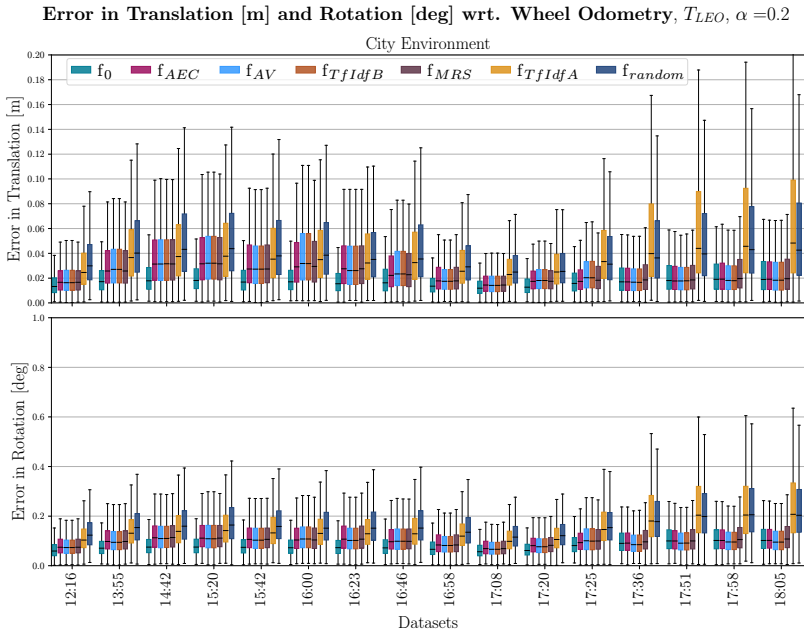


Figure 5.16: The localization precision for a selection fraction of 20%, for every dataset of the *City Environment* collection against the map containing only *rich sessions*.

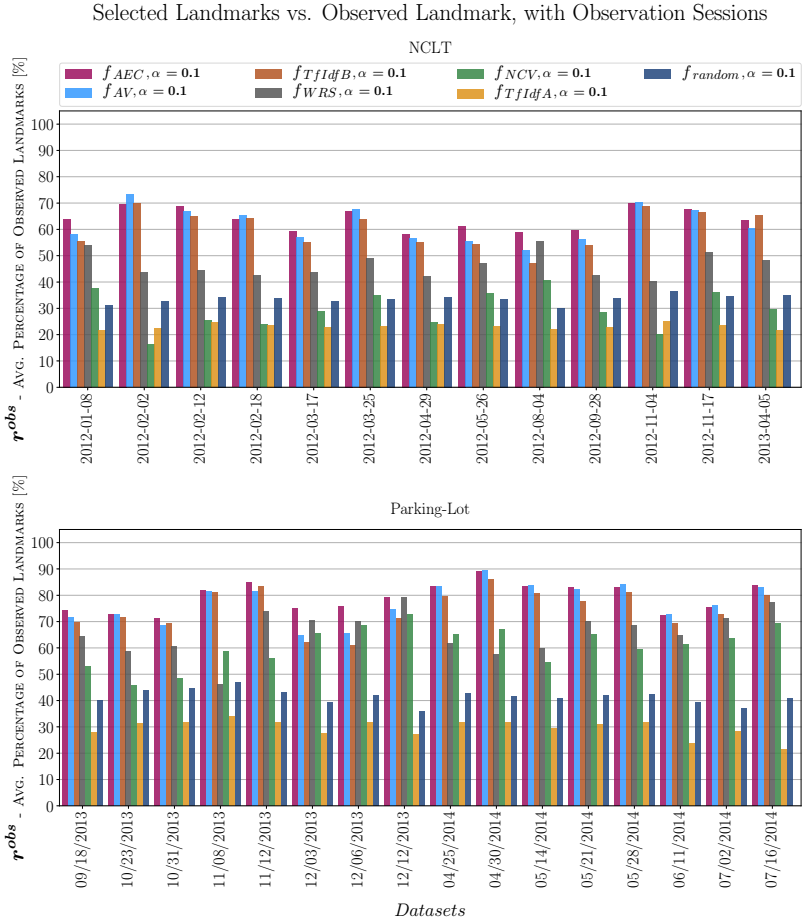


Figure 5.17: The average observation percentage r^{obs} for a selection fraction of 10%, for every dataset of the *NCLT* and *Parking-Lot* collection against the map with *observation sessions*.

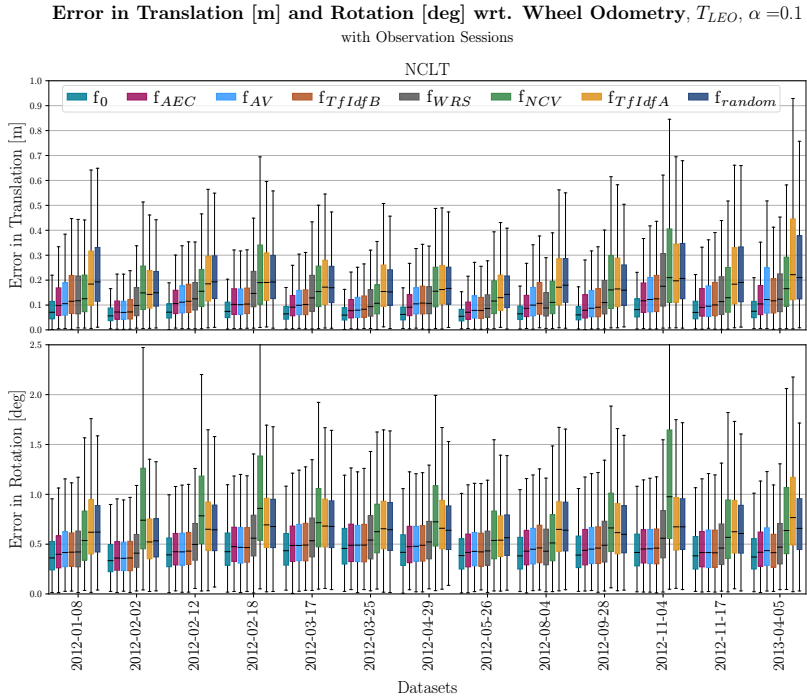


Figure 5.18: The localization precision for a selection fraction of 10%, for every dataset of the *NCLT* collection against the map with *observation sessions*.

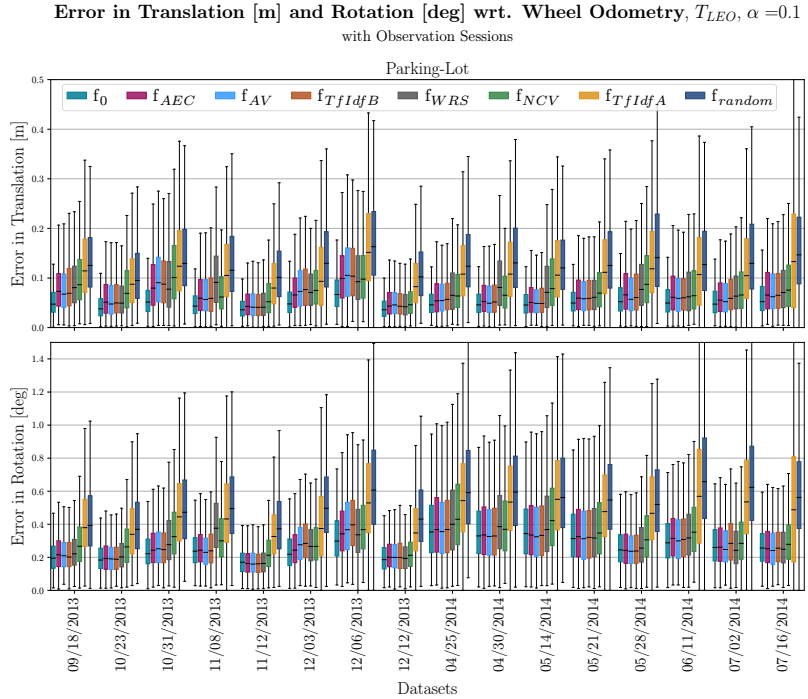


Figure 5.19: The localization precision for a selection fraction of 10%, for every dataset of the *Parking-Lot* collection against the map with *observation sessions*.

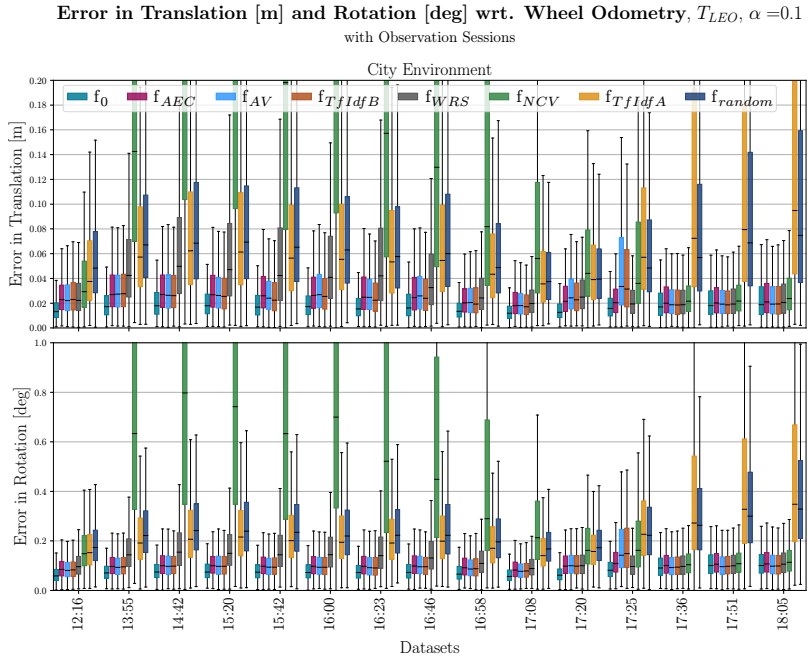


Figure 5.20: The localization precision for a selection fraction of 10%, for every dataset of the *City Environment* collection against the map with *observation sessions*.

6.2 Feature Descriptor Comparison

Our proposed appearance-based landmark ranking functions are per construction independent of the local feature descriptor used for mapping and localization, as they only take the co-observability patterns of landmarks into account. Nevertheless, the feature descriptor is an integral part of the localization pipeline, and thus the resulting performance of the localization with appearance-based landmark selection may not be identical with every choice of local feature descriptor. We have therefore evaluated the localization performance with popular choices of different local feature descriptors on the *Parking-Lot* dataset collection. The results are shown in Figure 5.21, 5.22 and 5.23. As expected, the results are similar regardless of the choice of descriptor.

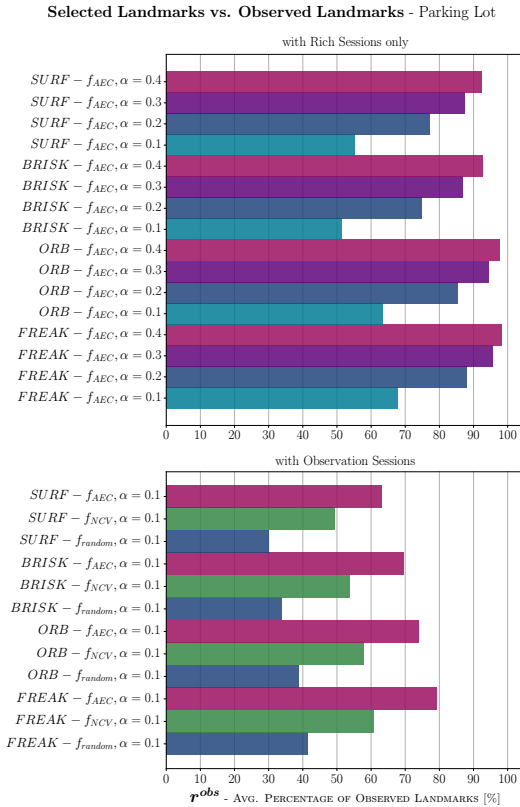


Figure 5.21: Observation percentage for different choices of feature descriptors, with a selection fraction of 20% against the map with only *rich sessions*, aggregated over all datasets of the *Parking-Lot* collection.

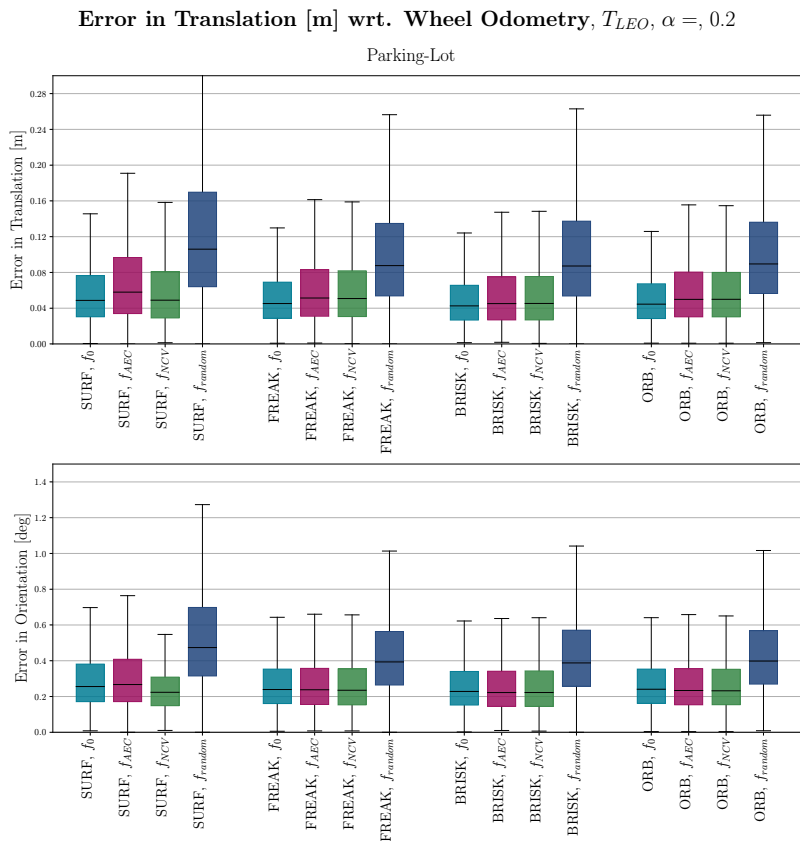


Figure 5.22: Localization precision with different choices of feature descriptors, a selection fraction of 20% against the map with only *rich sessions*, aggregated over all datasets of the *Parking-Lot* collection.

**Error in Translation [m] wrt. Wheel Odometry, T_{LEO} , $\alpha = 0.1$
with Observation Sessions**

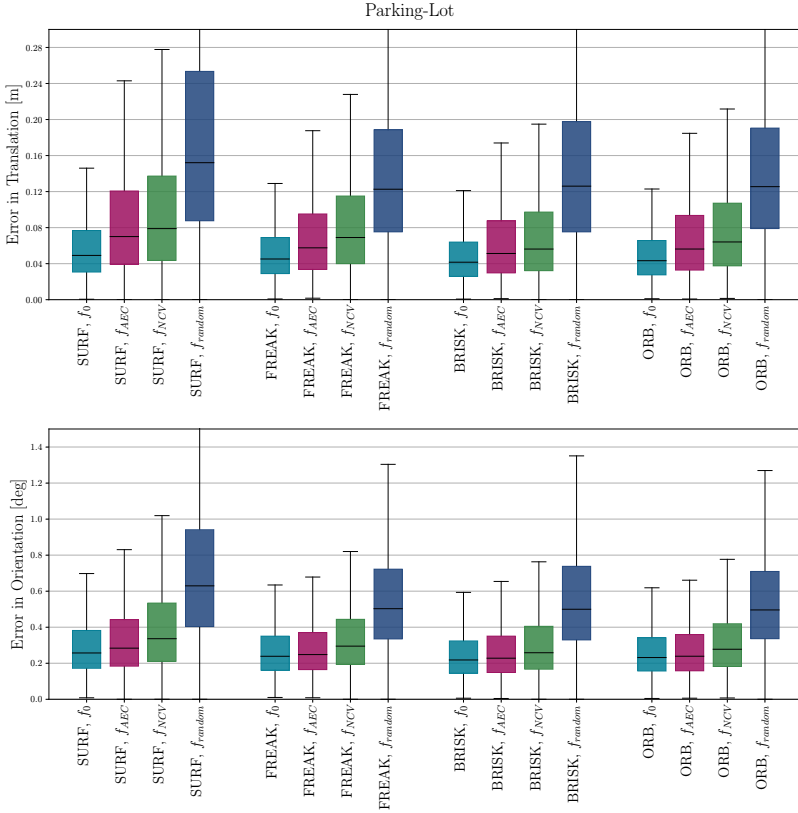


Figure 5.23: Localization precision with different choices of feature descriptors, a selection fraction of 10% against the map with *observation sessions*, aggregated over all datasets of the *Parking-Lot* collection.

6.3 Sample Images

Table 5.6: List of the *Parking-Lot* datasets with their respective weather condition and usage in the maps. The lower-case “*r*” and “*o*” indicate that the dataset has been added to the map as a *rich-* and *observation sessions* respectively.






























Date	W.	U.	Example Images
20th August, 2013		<i>r</i>	
17th September, 2013			
18th September, 2013			
11th October, 2013		<i>o</i>	
16th October, 2013		<i>r</i>	
23rd October, 2013			
31st October, 2013			
7th November, 2013		<i>r</i>	
8th November, 2013			
12th November, 2013			
19th November, 2013		<i>o</i>	
3rd December, 2013			
6th December, 2013			
10th December, 2013		<i>r</i>	
12th December, 2013			
14th January, 2014		<i>r</i>	
25th April, 2014			
28th April, 2014		<i>o</i>	
30th April, 2014			
5th May, 2014		<i>r</i>	
14th May, 2014			
21st May, 2014			
26th May, 2014		<i>o</i>	
28th May, 2014			
11th June, 2014			
30th June, 2014		<i>r</i>	
2nd July, 2014			
16th July, 2014			

Table 5.7: List of the *City Environment* datasets with their respective weather condition and usage in the maps. The lower-case “*r*” and “*o*” indicate that the dataset has been added to the map as a *rich-* and *observation sessions* respectively.

Date	W.	U.	Example Images
11:49	☀	<i>r</i>	
12:16	☀		
13:55	☀		
14:31	☀	<i>o</i>	
14:42	☀		
15:20	☀		
15:42	☀		
15:56	☀	<i>o</i>	
16:00	☀		
16:23	☀		
16:46	☀		
16:58	☁		
17:03	☁	<i>o</i>	
17:08	☁		
17:15	☁	<i>r</i>	
17:20	☁		
17:25	☁		
17:30	☁	<i>r</i>	
17:36	●		
17:43	●	<i>r</i>	
17:51	●		
17:58	●	<i>o</i>	
18:05	●		

Table 5.8: List of the *NCLT* datasets with their respective weather condition and usage in the maps. The lower-case “*r*” and “*o*” indicate that the dataset has been added to the map as a *rich-* and *observation sessions* respectively.

Date	W.	U.	Example Images
8th January, 2012	☀		
15th January, 2012	☀	<i>o</i>	
22nd January, 2012	☁	<i>r</i>	
2nd February, 2012	☁		
4th February, 2012	☀	<i>r</i>	
5th February, 2012	☀	<i>r</i>	
12th February, 2012	☀		
18th February, 2012	☀		
19th February, 2012	☁	<i>o</i>	
17th March, 2012	☀		
25th March, 2012	☀		
31st March, 2012	☁	<i>o</i>	
29th April, 2012	☁		
11th May, 2012	☀	<i>r</i>	
26th May, 2012	☀		
16th June, 2012	☀	<i>r</i>	
4th August, 2012	☀		
20th August, 2012	☀	<i>r</i>	
28th September, 2014	☁		
28th October, 2014	☁	<i>o</i>	
4th November, 2014	☁		
16th November, 2014	☀	<i>r</i>	
17th November, 2014	☀		
23rd February, 2013	☁	<i>o</i>	
5th April, 2013	☀		

Part B

EFFICIENT MAP MANAGEMENT

Paper



Map Management for Efficient Long-Term Visual Localization in Outdoor Environments

Mathias Bürki, Marcin Dymczyk, Igor Gilitschenski, Cesar Cadena, Roland Siegwart and Juan Nieto

Abstract

We present a complete map management process for a visual localization system designed for multi-vehicle long-term operations in resource constrained outdoor environments. Outdoor visual localization generates large amounts of data that need to be incorporated into a lifelong visual map in order to allow localization at all times and under all appearance conditions. Processing these large quantities of data is non-trivial, as it is subject to limited computational and storage capabilities both on the vehicle and on the mapping backend. We address this problem with a two-fold map update paradigm capable of, either, adding new visual cues to the map, or updating co-observation statistics. The former, in combination with offline map summarization techniques, allows enhancing the appearance coverage of the lifelong map while keeping the map size limited. On the other hand, the latter is able to significantly boost the appearance-based landmark selection for efficient online localization without incurring any additional computational or storage burden. Our evaluation in challenging outdoor conditions shows that our proposed map management process allows building and maintaining maps for precise visual localization over long time spans in a tractable and scalable fashion.

Published in:
Intelligent Vehicles Symposium (IV), 2018
DOI: 10.1109/IVS.2018.8500432

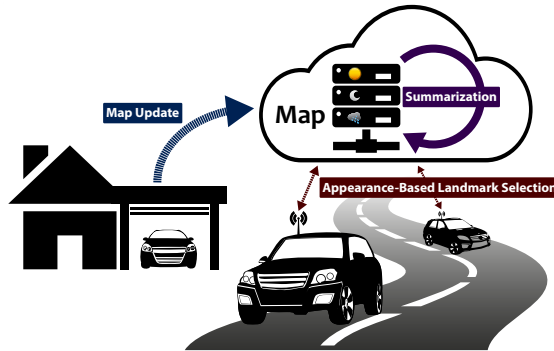


Figure 6.1: The model scenario motivating this work. Multiple vehicles simultaneously localize using a shared remote map, which is accessed over a mobile communication link. Appearance-based landmark selection allows only querying landmarks from the map which are in accordance with the current appearance condition, ensuring efficient usage of the available bandwidth. Once a vehicle returns from a sortie, collected map data is uploaded to the backend and incorporated into the map. Subsequent map summarization on the backend ensures the map size never exceeds a fixed number of landmarks, and thus guarantees limited storage requirements and a tractable map maintenance process in long-term perspective.

1 Introduction

Visual localization systems for mobile robots constitute an attractive alternative to laser-based systems, as the former can offer accurate localization performance with a low-cost sensor setup. This is especially true for future autonomous cars, where mass-production renders the sensor suite a sensitive matter of expense. However, visual localization systems generate large amounts of data that need to be processed both online on the vehicles as well as offline through map building and map maintenance. The inherent sensitivity of visual systems with respect to changing appearance conditions further exacerbates this problem in the context of long-term autonomy, as multiple appearances of the mapped places need to be stored and managed in order to be able to localize with satisfying accuracy across all conditions.

In recent years, methods have been presented to address the scalability and efficiency issues of individual components of a visual localization system ([8, 16, 24, 26, 65, 71]). However, little attention has been paid to how to combine the different components to form a complete and tractable localization framework, how the differently optimized methods interact, how visual maps are to be built and managed over indefinite time spans, and most importantly: how the large amount of data accumulated over time can be processed and incorporated in an optimal way; everything with the purpose of allowing precise visual localization

in outdoor environments at all times. It is the aim of this paper to address these questions. For this, we have built a scalable and efficient visual localization system for multi-vehicle outdoor shared-map scenarios as depicted in Figure 6.1 and anticipated for autonomous cars in the near future. It employs both appearance-based landmark selection on the vehicle side, as well as offline map summarization on the cloud-based mapping backend. We demonstrate how the visual maps can be managed and improved over time as the vehicles are exposed to vastly different appearance conditions. With the novel concept of an *observation session*, together with a modified formulation of the ranking function for appearance-based landmark selection, we propose a lightweight procedure to handle large quantities of frequently collected sensor data with the aim of improving the landmark selection performance without increasing the size of the map.

Our main contributions can be summarized as follows:

1. Demonstration of a complete map management procedure for an efficient visual localization and mapping system designed for long-term outdoor use.
2. Introduction of *observation sessions* and proposal of a new formulation of the ranking function for appearance-based landmark selection, allowing to exploit frequently collected sensor data to significantly increase the landmark selection performance without increasing the map size.

In an extensive evaluation in two real-world scenarios, covering weather and seasonal changes at daylight over the course of one year, and the extreme illumination change from day-time to night-time over the course of one day, we validate, first, the practicability of the proposed map management procedure in challenging outdoor conditions, and second, we show how additional co-observability statistics can improve the appearance-based online localization, and where the limitations thereof lie.

The rest of this paper is structured as follows: After an overview over related literature, we present our map management procedure in detail in Section 3, before presenting an extensive evaluation of the system’s performance in Section 4. Summarizing remarks about our key findings conclude the paper in Section 5.

2 Related Work

Ever since the advent of SLAM systems, maintaining map representations for enabling long-term operations has been a key focus, with a variety of different approaches evolving over time. The methods described in [88], [21] and [20] aim at maintaining a most up-to-date representation of the environment over longer times. These approaches, however, reach their limits whenever a map is required to represent multiple different representations at the same time, as is the case for outdoor visual localization applications.

For this reason, substantial efforts have been made to permanently augment visual maps with data from differing environment conditions. Churchill et al. [16]

introduced the Experience-Based mapping framework, which on-demand adds new “sub-maps” (referred to as “Experiences”) of the environment under newly observed appearance conditions. In a similar vein, the Multi-Session mapping proposed by Mühlfellner et al. [65] and the Multi-Experience Localization proposed by Paton et al. [70] both allow adding multiple datasets of an environment, collected during differing appearance conditions, to a common map representation. All of these approaches in their basic form, however, suffer from increased and ultimately unbounded storage, memory and computational resource requirements. To address this deficit, map summarization techniques have been developed, aiming at maintaining as small a map representation as possible, while at the same time still providing as good a localization performance across as far ranging appearance conditions as possible. Early works in this field include identifying reliable, geometrically-consistent features in the image retrieval context [87], and suppressing confusing features from certain regions of the database images [34]. Follow-up approaches vary from clustering [35] or random pruning [57] of visual “Views”, to selection at the landmarks level based on various landmark ranking functions [24, 25, 47, 65, 72].

The selection must not necessarily be carried out on the backend side, but instead may as well be performed already in an online fashion on the robot, prior to uploading new map data [26, 30, 74]. In general, these contributions focus on constructing a reliable set of landmarks for all possible environment states, reducing the runtime of tracking and localization, and/or the uplink bandwidth requirements.

In contrast to that, online landmark selection algorithms further allow decreasing the resource demands on the vehicle and on the communication downlink by having the vehicle query the map only for a selective fraction of landmarks which are deemed useful under the current operating conditions. Previous work by the authors [8] and by Linegar et al. [45] have successfully demonstrated such algorithms in the context of Autonomous Driving – a use-case especially prone to visual appearance change and applicable to the shared-map scenario. In relation to that, the work by Krajník [37], [36] aims at predicting the current state of the environment based on previously observed and learned temporal patterns. While this approach is promising for dynamic indoor applications, it is only partially applicable to outdoor environments with often non-periodic changes.

We believe that an ultimately efficient visual localization system must do justice to constrained resources along the whole pipeline, that is, on the mapping backend side, as well as on the mobile platform and the communication link in-between. None of the aforementioned works, however, address all of these constraints simultaneously, whereas in this paper, we present a map management procedure that allows reaching an entirely scalable and efficient visual localization system for long-term use. Furthermore, and in contrast to [45] and [71], our metric multi-session map representation (see [65], [25]) keeps all map data (vertices, landmarks), even from multiple appearance conditions, expressed in a single map reference frame. This not only facilitates higher level tasks of autonomous operation, such as path planning and control, but also allows implementing the online landmark-selection and the offline map summarization on the level of individual landmarks.

3 Methodology

In this section, we present the theoretical concepts of the three main components of our localization and mapping system: (i) the map update, (ii) appearance-based landmark selection with *observation sessions*, and offline map summarization.

3.1 Map Update

The methodology of our map management procedure is based on a map update process as depicted in Figure 6.2. Sensor data, consisting of camera images and wheel-odometry measurements, is collected during a sortie of a vehicle and processed after the vehicle has returned to its home-base. The newly collected dataset is first localized in an offline process against the map available at that time. In case the performance of this localization is worse than a pre-defined threshold, the map is considered to not cover the appearance condition encountered during this sortie sufficiently well and new landmarks are tracked and triangulated from the dataset. A dataset added to the map in this fashion is referred to as a *rich session*. A subsequent map summarization step ensures the total number of landmarks to remain below a fixed number, guaranteeing a bounded map size at all times. If, on the other hand, localization has performed sufficiently well, the map is considered to cover the encountered conditions and no new landmarks are added to the map. In this case, however, the localization still reveals useful information about what landmarks in the map have been observed during the sortie. This information is added to the map in the form of an *observation session*. In contrast to the *rich session* described above, adding an *observation session* conforms to merely marking existing landmarks as observed in the respective sortie. However, both for future online localization as well as future map summarization steps, this additional statistical data is valuable, as it allows a better distinction between useful and not useful landmarks. The resulting updated map is then used for localization of subsequent sorties.

In order to benefit from the *observation sessions* during online localization with appearance-based landmark selection, a modified formulation of the landmark ranking function is required and described in the following subsection.

3.2 Appearance-Based Landmark Selection with Observation Sessions

In our previous work presented in [8], a method to tackle the problem of only querying useful map data during an outdoor operation has been introduced on the basis of appearance-based landmark selection. Following an iterative localization paradigm, such as the ones described in [65] and [39], a ranking function $f(l)$ assigns a score to each landmark of a candidate set C_k (pre-selected based on spatial proximity), according to how likely l is observable under the current appearance condition. Then, a small subset S_k of top-ranked landmarks are selected using a selection policy $\Omega(f(), \dots)$, transmitted to the vehicle, and used for localization at

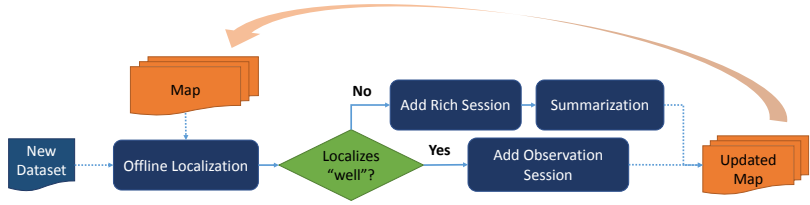


Figure 6.2: Schematic illustration of the map update process on the backend. A newly received dataset is first localized against the available map. In case the localization performance is below a defined threshold, new landmarks are created from the dataset and added to the map, followed by a summarization step, which reduces the total number of landmarks in the map again to a fixed number. In the other case, the co-observation statistics of all landmarks observed during the localization are updated, but no new landmarks are added.

iteration k . The ranking function described in [8] adaptively weights the different sessions present in the pre-built map based on the session-affiliation of recently observed landmarks along the traversal. Although successfully reducing the amount of landmarks used for localization, it relies on the map to be created a priori with all sessions approximately uniformly distributed across the appearance space.

In a practical scenario, however, the map sessions may not be uniformly distributed, but they are rather added once a “new” appearance condition is encountered for the first time. In addition to that, whenever the vehicle traverses through the mapped area under an appearance condition already well-covered in the map, additional co-observability information can be gathered in the form of *observation sessions*.

As our experiments presented later show (see Figure 6.6), the original ranking function from [8], denoted by $f_{orig}()$, is not well suited to incorporate these additional *observation sessions*. We thus propose a new formulation of the ranking function that is agnostic to how the mapping sessions are distributed across the appearance space, and the number and distribution of additional *observation sessions* present in the map.

Let \mathcal{A} denote the current appearance condition. We are interested in evaluating $p(l | \mathcal{A})$, corresponding to the probability of observing landmark l under the current appearance condition. Let further Z denote the set of all sessions present in the map, both *rich sessions* and *observation sessions*, and L denote the set of all landmarks in the map. With every landmark $l \in L$, we associate the set Z_l , corresponding to all sessions that have observed landmark l .

We note that $p(l \mid \mathcal{A})$ directly depends on Z_l , that is, $Z_{l_i} = Z_{l_j} \Rightarrow p(l_i \mid \mathcal{A}) = p(l_j \mid \mathcal{A})$. We can thus group landmarks into appearance equivalence classes, according to:

$$[l_i] := \{l_j \in L \mid Z_{l_j} = Z_{l_i}\} \quad (6.1)$$

with

$$p(l_j \mid \mathcal{A}) = p(l_i \mid \mathcal{A}) \quad \forall l_j \in [l_i] \quad (6.2)$$

Hence, evaluating $p(l \mid \mathcal{A})$ amounts to evaluating $p([l] \mid \mathcal{A})$, which can be interpreted as the relevance of appearance equivalence class $[l]$ under appearance condition \mathcal{A} .

The abstract appearance condition \mathcal{A} is not directly observable. However, it can be approximated by the means of recently selected and observed landmarks as follows:

$$p([l_i] \mid \mathcal{A}) \approx \begin{cases} \frac{|\mathcal{O}_{[i]}|}{|S_{[i]}|}, & \text{if } |S_{[i]}| > 0 \\ 0, & \text{otherwise} \end{cases} \quad (6.3)$$

where $\mathcal{O}_{[i]}$ and $S_{[i]}$ denote the sets of recently observed and selected landmarks of appearance equivalence class $[l_i]$ respectively. Accordingly, we define our new landmark ranking function as:

$$f_{rank}(l) := p([l] \mid \mathcal{A}), \quad \forall l \in L \quad (6.4)$$

In Figure 6.3, a comparison of our modified ranking function with the originally proposed formulation on the experimental set-up used in [8] is shown, ensuring that our modified formulation does not reveal regressive performance under these conditions. Further evaluation results demonstrating the merit of our new formulation are presented in Section 4.2.

3.3 Offline Map-Summarization

Whenever a *rich session* is added to the map, the total number of landmarks increases, and therewith also the size of the map. We therefore apply the map summarization techniques proposed by Dymczyk et al. in [24] to keep the map size bounded at all times.

They suggest to reduce the number of landmarks by solving the following integer-based optimization problem:

$$\text{minimize } \mathbf{q}^T \mathbf{x} + \lambda \mathbf{1}^T \boldsymbol{\zeta}, \text{ subject to} \quad (6.5)$$

$$\sum_{i=1}^N \mathbf{x}_i = n_{desired} \quad (6.6)$$

$$\mathbf{A} \mathbf{x} + \boldsymbol{\zeta} \geq b \mathbf{1} \quad (6.7)$$

$$\boldsymbol{\zeta} \in \{\{0\} \cup \mathbb{Z}^+\}^M. \quad (6.8)$$

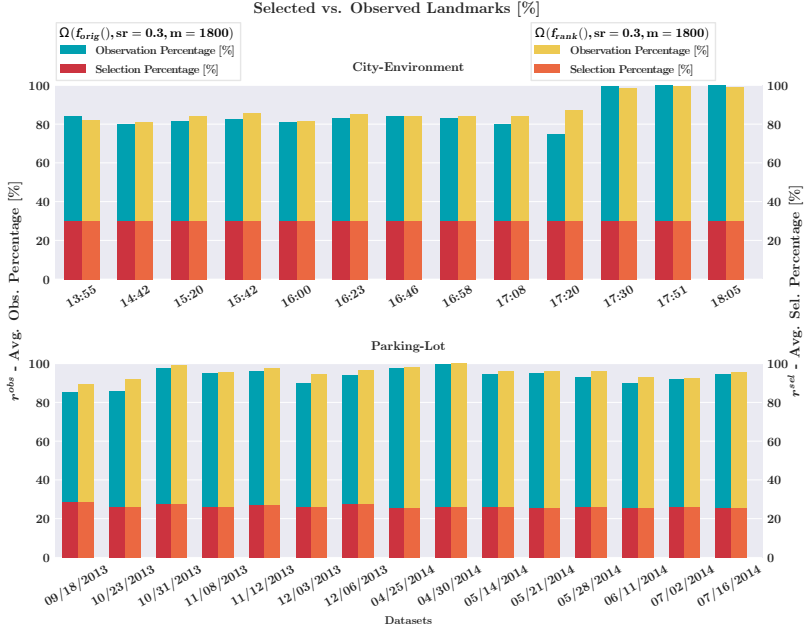


Figure 6.3: Direct comparison of our new appearance-based ranking function $f_{rank}()$ with the original ranking function proposed in [8] $f_{orig}()$, ensuring the former to perform at least as well on the same map and selection policy used for the experiments in [8]. For each dataset of the City-Environment and the Parking-Lot scenario, the observation percentage r^{obs} (blue, yellow bars) is shown for a selection ratio of 30% and a maximum number of selected landmarks of $m = 1800$. In this setting (a priori built map without *observation sessions*) both ranking functions perform equally well. The benefit of our modified ranking function $f_{orig}()$ is shown in further experiments with *observation sessions* presented in section 4.2.

Each landmark is assigned a corresponding binary switch variable $x_i \in \{0, 1\}$, indicating whether the landmark should be kept or removed. The landmarks are selected based on the cost vector \mathbf{q} , estimated using the number of sessions a landmark was observed in and the total number of observations. Additional constraints ensure some desired total number of landmarks ($n_{desired}$) to remain in the map, and a sufficient number of landmarks (b) visible from every vertex. Matrix \mathbf{A} encodes the vertex-landmark co-observability, while the slack variable ζ allows to relax this constraint (at cost λ), ensuring a solution to the optimization problem can be found in all cases.

4 Evaluation

Our evaluation is structured into two parts as follows: We first present our findings related to the map management process and offline summarization, thereby looking into how many *rich sessions* are added over time, and how the degree of map summarization affects the localization performance. In the second part, we show the performance of the online appearance-based landmark selection on the incrementally improved maps over time, focusing on a comparison between our modified, more generic ranking function proposed in Section 3.2, and the original ranking function proposed in [8] under the influence of additional *observation sessions*.

The data for the evaluation has been collected in two complementary real-world scenarios. The first one covers weather and seasonal change over the course of a full year at day-time on an outdoor parking-lot, while the second one covers extreme lighting change from full day-light to complete night-time in a city environment. The sensor suite consists of four fish-eye cameras, one facing in each cardinal direction, running at 12.5Hz, and wheel-odometry. The images are scaled down to 640×400 pixels prior to processing. Example images can be found in [8] and in the video contributions available online^{1,2}. All computations have been performed on a consumer-grade laptop with an Intel i7 CPU. Localization runs in real-time at > 5 Hz.

4.1 Map Update and Summarization

As a metric for assessing the quality of the generated maps over time, we employ translation RMS errors between the rough pose estimate from forward-propagated wheel-odometry and the refined pose after optimization. In accordance with the results found in [8], we omit the presentation of RMS errors in orientation, as they highly correlate with the translation errors and are of negligible magnitude in any case ($\ll 2^\circ$). Note that the refined pose at each iteration is obtained from solving a vision-only non-linear least-squares optimization problem. The resulting RMS error thus approximates the standard deviation of the localization along the trajectory.

¹<https://youtu.be/TJMqGSHtIjU>

²https://youtu.be/JL_5zMEQKYc

In outdoor environments, the updated map must still be able to cover the range of appearance conditions represented by the incorporated *rich sessions*, even after summarization. The number of landmarks required to achieve this not only depends on the sensor setup and the spatial extent of the map, but also on the variance in appearance conditions encountered. The City-Environment scenario covers extreme lighting changes from day-time to night-time, but the overall variance is still considerably smaller than in the year-long day-time Parking-Lot scenario. To guide our map update process described in Section 3.1 and Figure 6.2, we have therefore chosen to perform map summarization with a maximum number of 75k (75'000) and 150k landmarks in the City-Environment and the Parking-Lot scenario respectively, and use a 10cm threshold on the translation RMS error on these maps as a decision criterion to add the dataset at hand either as a *rich session*, or as an *observation session*. The choice of the 10cm threshold is motivated by recent work ([8, 65]) suggesting this to be a reasonable and realistic upper bound for localization precision with the given sensor suite.

The evolution of the localization performance resulting from this map update regime is shown in Figure 6.5, where localization has been evaluated with the following three combinations of selection policies and ranking functions: a) $\Omega(f_0(), \alpha = 1.0)$, using all candidate landmarks for localization, b) $\Omega(f_{rank}(), \alpha = 0.2)$, appearance-based landmark selection with a selection ratio of 20%, and c) $\Omega(f_{rand}(), \alpha = 0.2)$, the corresponding random selection. As described in 3.2 and [8], the selection policy $\Omega(f(), \alpha = \alpha)$ selects some fraction α of top-ranked landmarks from C_k which are then used for localization at the given iteration k .

In order to thoroughly assess the influence of the offline map summarization on the localization performance, we have further evaluated the latter against more strictly summarized maps (with 50k for the City-Environment, and 100k landmarks for the Parking-Lot environment respectively), as well as against indefinitely growing unsummarized maps.

In the City-Environment scenario, appearance conditions appear stable throughout the afternoon until the beginning of dusk shortly after 5pm. At 5:15pm, 5:30pm and 5:43pm, additional *rich sessions* are added, gradually expanding the appearance coverage of the map until, finally, night-time localization is feasible at 6pm.

In contrast to that, in the Parking-Lot scenario, the appearance patterns are much less clear. Already the initial map, built from the first dataset from August 20th, does not allow sufficient localization performance for the second dataset from September 17th. In general, it seems to be necessary to have a *rich session* present for every month of the year. Nevertheless, for the second half of the year, the map clearly shows converging tendencies, with only occasional datasets just barely above the 10cm precision threshold, and the spread between reference localization using all landmarks, and the random selection decreasing.

Figure 6.4 further shows the number of landmarks associated with each *rich session* at each stage of the incremental map building process, both for the summarized map and the unsummarized map. The *rich sessions* added in the City-Environment scenario in dusk naturally contain considerably fewer landmarks, which is also

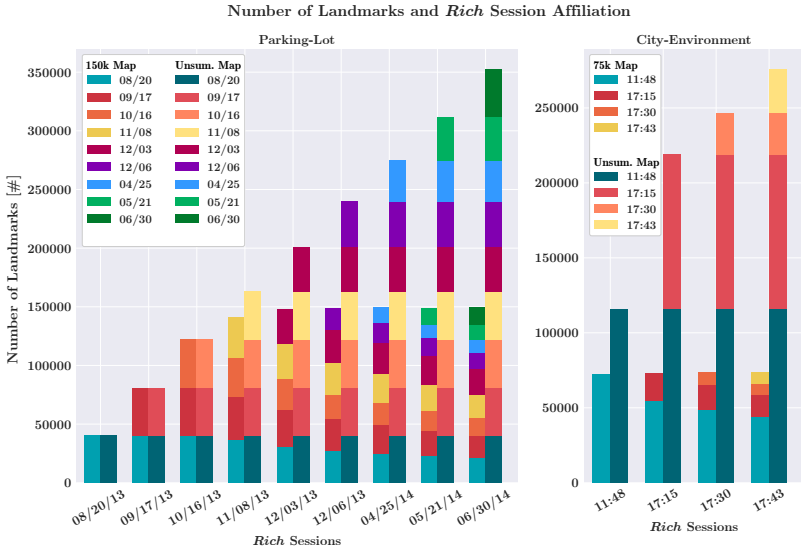


Figure 6.4: The number of landmarks and their affiliation with the corresponding *rich session*, both for the summarized and the unsummarized maps, is presented for the two evaluation scenarios. On the x-axis, every dataset which is added as a *rich session* is listed in chronological order, whereas the y-axis shows the absolute number of landmarks the summarized and unsummarized maps contain at this stage. The colors indicate how many landmarks belong to what *rich session*.

reflected in both the summarized and the unsummarized map. In contrast to that, the *rich sessions* added in the Parking-Lot scenario all contain a similar number of landmarks, and summarization reduces already present sessions more or less equally as new sessions are added.

Ideally, the localization precision is preserved after map summarization. If this is the case, the summarization only removes noisy landmarks from the map which are not re-observable under any of the encountered appearance conditions. In the City-Environment scenario, the performance difference related to map summarization is best visible at night-time, where the unsummarized map shows significantly better performance in case of the random selection. However, with the appearance-based selection, almost the same precision is attainable as if all landmarks were used. This shows that the summarization algorithm successfully removes redundant and noisy landmarks while maintaining a good coverage over the different appearance conditions. Similar results can be observed also for the Parking-Lot scenario. The more *rich sessions* are added, and hence the fewer landmarks of an individual

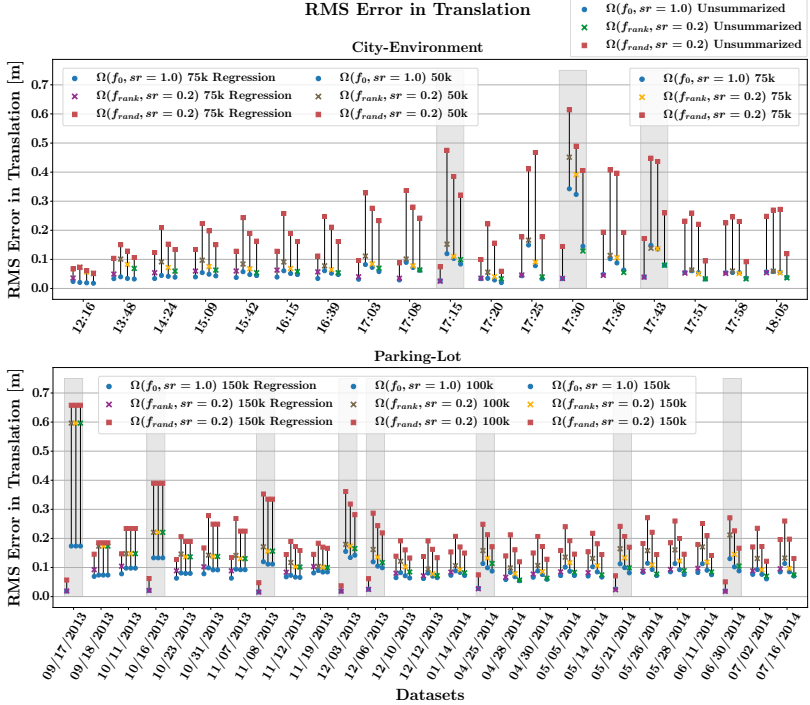


Figure 6.5: Translation RMS error for all the datasets of both scenarios, differently summarized maps, and different ranking functions and selection policies. The datasets marked in dark gray are added as *rich sessions* while all other datasets are added as *observation sessions* instead. The blue circles correspond to the precision achieved when using all landmarks for localization ($\Omega(f_0), \alpha = 1.0$), whereas the red cubes correspond to 20% random selection ($\Omega(f_{rand}), \alpha = 0.2$), while the colored crosses represent the appearance-based selection ($\Omega(f_{rank}()), \alpha = 0.2$). These results are shown for four different map configurations in each of the scenarios: The unsummarized map, two degrees of summarization, and with respect to the final summarized map (“Regression”).

rich session can be present in the summarized map, the larger the performance gap between the differently summarized and the unsimplified map becomes. It can further be observed that for the 100k map, the appearance-based localization cannot keep up with the performance compared to the 150k map, indicating that for this scenario and this time span, more than 100k landmarks are required in order to maintain sufficient appearance space coverage.

Since in these experiments we deliberately choose to build the maps incrementally and in chronological fashion, the performance evaluation of a certain dataset only uses the map available at that point in time. To demonstrate that the summarization algorithm in fact creates maps that maintain usability across all previously encountered appearance conditions, we evaluate the performance of all datasets in retrospect using the final map created after having processed the last dataset in chronological order. The results of this “regression” test are shown in Figure 6.5, with the corresponding map labelled with “Regression”. As can be seen, all datasets achieve at least as high a precision as if the map available by that time is used instead. Note, however, that for all datasets added as *rich sessions* this “regression” test in principle corresponds to self-localization. Hence the artificially high precision in these cases.

4.2 Appearance-Based Landmark Selection

The goal of the appearance-based landmark selection is to achieve a high online localization performance with as few landmarks selected as possible. This can best be evaluated by comparing respective selection ratios α with the corresponding observation ratios r_k^{obs} :

$$r_k^{obs} := \frac{|\mathcal{O}_k^\Omega(f_{rank}, \alpha=\alpha)|}{|\mathcal{O}_k^\Omega(f_0, \alpha=1.0)|} \quad (6.9)$$

The selection ratio α denotes the fraction of candidate landmarks used for localization at iteration k . In contrast to that, the observation ratio r_k^{obs} compares the number of observed landmarks, using a specified ranking function $f_{rank}()$ and selection ratio α , to the number of observed landmarks when all candidate landmarks are used.

In [8], it has been shown that with a pre-built map and selection ratios between 20-40%, a localization performance comparable to using all landmarks can be achieved. In contrast to that, in this paper, we aim at investigating how the relation between selection ratio and localization performance evolves in a scenario where datasets are chronologically processed and the map is built-up incrementally, with both *rich sessions* and *observation sessions*.

Figure 6.6 shows the observation percentage for a selection ratio of 20% for the three cases of using the ranking function originally proposed in [8], and using our new ranking function introduced in Section 3.2 with and without *observation sessions*.

In early stages of the map building with only few *rich sessions* present in the map, the benefit of using the *observation sessions* is most pronounced. As

soon as more *rich sessions* become available though, the selection not using the *observation sessions* performs more and more similarly. In case of the City-Environment scenario, this is due to the fact that towards night-time even a pure selection on the 5:30pm and 5:43pm *rich sessions* allows achieving virtually 100% observation percentage already. In contrast to that, the appearance conditions in the Parking-Lot scenario are much more diverse and unrelated from one dataset to the next one. After having some number of *rich sessions* present in the map, additional co-observability information in potentially only weakly related appearance conditions is only of minor or no help anymore.

5 Conclusions

We have presented a complete map management process for a visual localization system tailored to long-term operations in resource constraint outdoor environments. Offline map summarization guarantees maps of bounded size at all times, while online localization with appearance-based landmark selection allows only transmitting and processing the map data required and useful under the current appearance condition. With the incorporation of landmark co-observation statistics in the form of *observation sessions* in combination with a new formulation for the appearance-based landmark ranking function $f_{rank}()$, we have proposed a lightweight mechanism to improve the appearance-based landmark selection during online localization at negligible storage or computational costs. An extensive evaluation in real-world conditions has shown that these additional *observation sessions* have the potential to significantly improve the landmark selection performance. However, their usefulness degrades as more and more *rich sessions* are available in the map. We have further evaluated the localization performance on the maps with different degrees of summarization resulting from the proposed map management paradigm, and shown that precise localization is possible over long time frames and across vastly different appearance conditions while keeping the map size bounded.

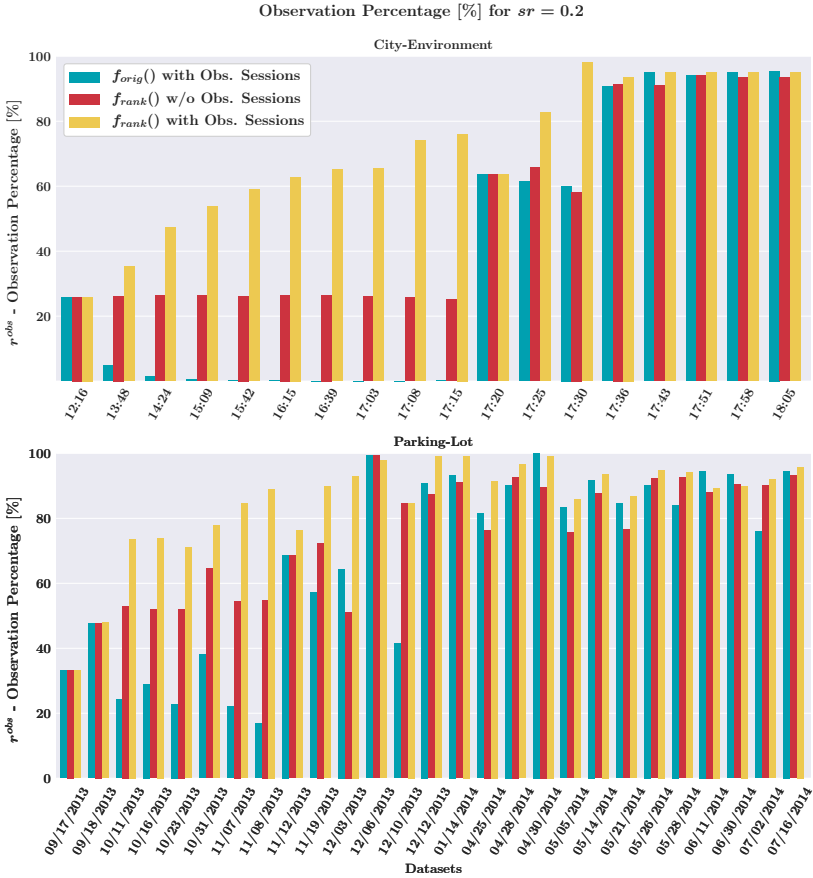


Figure 6.6: The observation percentage r^{obs} is shown for both scenarios, a selection ratio of 20% and for the following three ranking functions and selection policies: a) $f_{orig}()$ with observation sessions, b) $f_{rank}()$ with observation sessions, and c) $f_{rank}()$ without observation sessions. Especially in the early stages where the map still contains only very few rich sessions, the observation sessions allow a significant boost of the observation percentage.

Part C

RELIABLE VISUAL LOCALIZATION IN
UP-DRIVE

VIZARD: Reliable Visual Localization for Autonomous Vehicles in Urban Outdoor Environments

Mathias Bürki, Lukas Schaupp, Marcin Dymczyk, Renaud Dubé,
Cesar Cadena, Roland Siegwart, and Juan Nieto

Abstract

Changes in appearance is one of the main sources of failure in visual localization systems in outdoor environments. To address this challenge, we present VIZARD, a visual localization system for urban outdoor environments. By combining a local localization algorithm with the use of multi-session maps, a high localization recall can be achieved across vastly different appearance conditions. The fusion of the visual localization constraints with wheel-odometry in a state estimation framework further guarantees smooth and accurate pose estimates. In an extensive experimental evaluation on several hundreds of driving kilometers in challenging urban outdoor environments, we analyze the recall and accuracy of our localization system, investigate its key parameters and boundary conditions, and compare different types of feature descriptors. Our results show that VIZARD is able to achieve nearly 100% recall with a localization accuracy below $0.5m$ under varying outdoor appearance conditions, including at night-time.



Figure 7.1: We aim at accurately localizing the *UP-Drive* vehicle depicted in the upper-left corner in a map of visual features depicted on the right side. Features are extracted from images of the surround-view camera system (lower-left corner) and matched against 3D landmarks in the map. Inlier matches, centered on the estimated $6DoF$ pose of the vehicle in the map, are illustrated as dark yellow lines on the right side.

1 Introduction

Localization is a pivotal capability of any autonomous vehicle. By knowing their precise location, vehicles are able to plan a path to a next goal location, navigate safely in the environment, and eventually successfully complete their mission. Especially for autonomous vehicles in urban environments, localization is challenging, as GNSS based localization systems fail to provide reliable and precise enough localization near buildings due to multi-path effects, or in tunnels or parking garages due to a lack of visible satellites. Alternative exteroceptive sensor modalities are therefore necessary to accomplish this task, of which LiDARs and cameras have received most attention in recent years [12, 49]. While LiDARs have become more suited for mass market adoption, we believe there are still significant advantages with camera-based localization systems, despite the challenges related to long-term appearance change in outdoor environments. Cameras remain considerably more cost-effective than LiDAR sensors, allowing them to be deployed in multitudes and in a flexible way on a large quantity of vehicles. Furthermore, they can be used for sensing both appearance and geometric information of the environment, and are often better suited for global localization and loop-closure detection, which are necessary capabilities for bootstrapping any local localization algorithm, and to maintain geometrically consistent maps in lifelong operation [12].

For these reasons, we have developed a visual localization system, dubbed

VIZARD, for the self-driving cars in the *UP-Drive* project¹, with the following main features:

1. We employ map-tracking, a local localization algorithm able to generate both accurate *6DoF* pose estimates and achieve high localization recall.
2. Multi-session mapping techniques enable us to successfully tackle the challenge of long-term appearance change in outdoor environments, and even allow for localizing in night-time conditions.
3. The use of binary descriptors and an efficient sensor fusion backend renders real-time localization with CPU-only hardware set-ups feasible.

In a thorough evaluation of all crucial aspects of our localization system using two long-term outdoor dataset collections, one of them publicly available, we carefully analyze the most important parameters in our pipeline, compare the use of different binary descriptors, investigate key performance metrics such as localization accuracy and recall and relate to a state-of-the-art metric global localization algorithm. We see the main added value of this paper in sharing with the community the insights gained in this long-term study.

The contributions of this paper are thus as follows:

- We thoroughly study the critical parameters of our localization system, analyze their boundary conditions, and share our gained insights.
- From a comparison of the localization performance using different binary descriptors, we show which descriptors are best suited for map-tracking across long-term appearance change.
- In an extensive evaluation on multiple long-term dataset collections, we demonstrate state-of-the-art localization performance across vastly different appearance conditions in outdoor environments, including at night time.

2 Related Work

Visual localization systems can be divided into two main categories. Global localization systems are able to retrieve the location of a robot with no prior knowledge of the robot's pose. In contrast to that, local localization systems exploit a motion model to compute a prior on the robot location, thereby reducing the search space in the map.

¹The **UP-Drive** project is a research endeavor funded by the European Commission, aiming at advancing research and development towards fully autonomous cars in urban environment. See www.up-drive.eu.

Global Localization

Early visual global localization systems have been presented in the context of offline geometric scene reconstruction from a large number of images collected from varying viewpoints [1, 81]. These works led the foundation for many subsequent *6DoF* global localization algorithms, and have been improved in numerous follow-up works [19, 44, 46, 50, 79]. More recently, deep learning techniques have given rise to novel global localization algorithms with remarkable robustness against drastic appearance change [23, 78]. They require, however, high-end GPUs in order to achieve real-time operation.

In general, the aforementioned global localization algorithms are capable of achieving reliable localization across significant appearance change in outdoor environments. However, as shown in [80], they often fall short of providing high recall with localization accuracies below $0.5m$, and are thus not well suited for our application, where we aim at permanently localizing our vehicle with sufficient accuracy to prevent deviation from the road lane boundaries. Note that there has also been a substantial amount of work on global localization in the realm of place recognition, or image retrieval [4, 53, 59, 67, 86]. These approaches, however, only provide a best matching image candidate in a map, instead of a *6DoF* metric pose, and are thus addressing a different problem than ours.

Local Localization

Local localization algorithms take prior information on the robot pose into account, in order to reduce the localization search space and increase recall. This is well motivated in practice, as subsequent localization attempts of a mobile robot are far from independent, but in fact highly correlated in space, with the incremental motion between images often observable, although with drift, from odometry sensors such as wheel-odometry or IMUs. As a consequence, instead of regarding localization as an independent module, it can be directly integrated into the state estimation framework that optimizes the robot’s pose in its environment by fusing odometry measurements and localization constraints. The ORB-SLAM [66] framework with its *localization mode* offers a local localization system similar to ours. They lack, however, the capability to integrate multiple sessions into a map, which greatly limits the robustness towards appearance change in outdoor environments. Lategahn and Schiller present a hierarchical visual localization system for outdoor environments that combines a global with a local localization module [40]. They achieve robustness against appearance change by employing DIRD descriptors [41]. Their experimental evaluation, however, only spans across six weeks, and it thus remains unclear, how well their system performs over long-term appearance change. Instead of employing an illumination invariant descriptor, Paton et al. use color-constant images to gain robustness against appearance change [69]. However, color-constancy primarily removes shadows under sunlight, but does not tackle other variations in appearance, such as seasonal change, or transitions from day to night-time.

Multi-Session Mapping

A common technique to achieve robustness against arbitrary long-term appearance change incorporates visual cues from multiple sorties through the environment in the map. We refer to this as multi-session mapping. Schneider et al. have presented a state estimation framework that fuses visual-inertial odometry with metric global localization [6, 51, 82]. While their mapping framework allows merging several sessions, they use a feature based global localization algorithm which prohibits sufficient localization recall in outdoor environments. Paton et al. present a visual localization system using multi-session mapping in [71]. Their application is, however, restricted to a teach-and-repeat scenario. In contrast to that, we employ loop-closure detection and bundle adjustment in order to get geometrically consistent multi-session maps, which adds additional flexibility in route planning and navigation. The “Experienced-Based Mapping” framework developed by Churchill et al. maintains separate map instances for diverse appearance conditions [16]. This allows visual localization in arbitrarily appearance conditions in an elegant and efficient manner. However, the maintenance of separate maps for differing conditions renders it impossible to share visual cues between sessions, which can increase recall. Furthermore, an integration of the localization module into a complete navigation stack is more challenging, as the visual pose estimates are expressed with respect to separate, disconnected coordinate frames.

Similarly to our localization system, the works presented in [42, 63, 65, 83] use a local localization algorithm with multi-session maps for localization. As opposed to our work, Mühlfellner et al., refrain from fusing their visual pose estimates with wheel-odometry, which limits the accuracy and smoothness of their pose estimation framework, while Sons et al. do not report on localization accuracy and recall in long-term experiments.

3 Methodology

The VIZARD system consists of the following main components, presented at the beginning of this section. a) We employ a state estimation framework for fusing wheel-odometry and visual localization constraints. b) Our map-tracking module matches keypoints extracted from current camera images to landmarks from the map. Furthermore, key information regarding our mapping pipeline is provided at the end of this section, and a schematic overview of VIZARD can be found in Figure 7.2.

3.1 State Estimator

At the core of our localization system we employ a state estimation framework in information form, the dual representation of the (Extended) Kalman-Filter [11, 84]. Our state representation entails an estimate of the current transformation between the vehicle body coordinate frame \mathcal{F}_B and the map reference frame \mathcal{F}_W for each time-step t : $\mathbf{x}_t := [T_{WB}^t]$. Note that T_{WB} is an element of $SE(3)$, and thus

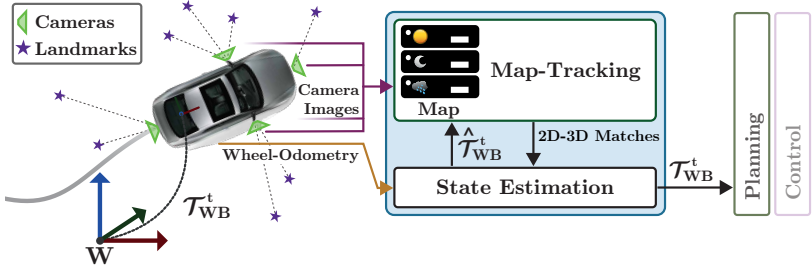


Figure 7.2: The map-tracking module extracts 2D features from current camera images, and matches them with 3D map landmarks locally in image space using a pose prior $\hat{T}t$. The state estimation module fuses the visual 2D-3D matches with the current wheel-odometry measurement to obtain a current vehicle pose estimate T_{WB}^t .

represents all six degrees of freedom. The corresponding rotations are represented by unit quaternions. At every time-step t , a set of n simultaneously recorded camera images, and a relative odometry transformation measurement $T_{B_{t-1}B_t}^{odo}$ are received. A new state is created by forward-propagating the previous state estimate using the odometry measurement: $\hat{T}t := T_{WB}^t T_{B_{t-1}B_t}^{odo}$. It is used both in the map-tracking module as a pose prior, and as an initial linearization point in the filter update. After finding 2D-3D matches in the current set of images using map-tracking, the states are updated by retrieving the *MAP* estimate of the following cost function:

$$\begin{aligned}
 c(T_{WB}^t, T_{WB}^{t-1}) := & \|f_{prior}(T_{WB}^{t-1})\|_{P_{t-1}}^2 \\
 & + \|f_{odo}(T_{WB}^t, T_{WB}^{t-1})\|_Q^2 \\
 & + \sum_{i=1}^m \varphi(f_{loc}(T_{WB}^t))
 \end{aligned}$$

The prior pose and odometry factors, f_{odo} and f_{prior} respectively, follow a standard quadratic loss expression, while the localization re-projection factors f_{loc} employ a *Huber* robust cost function φ to account for possible wrong keypoint-landmark associations. All factors follow a standard graph SLAM formulation, as described in [12]. We retrieve the *MAP* estimate by iteratively minimizing the cost function c using the Levenberg-Marquardt in the GTSAM framework [22].

3.2 Map-Tracking

At every timestep t , the forward-propagated pose $\hat{T}t$ represents a rough estimate of the vehicle's location at time t in the map. With this, we can retrieve all landmarks from the map that have been observed from within a given distance around $\hat{T}t$.

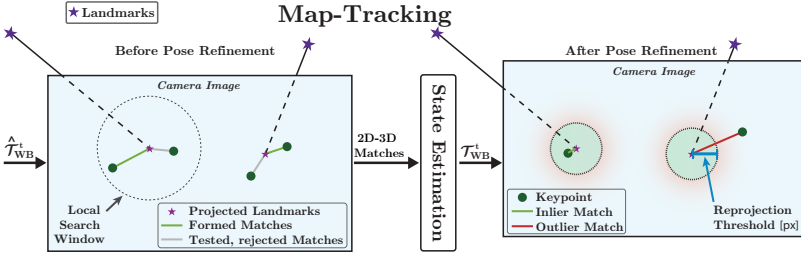


Figure 7.3: The map-tracking module extracts 2D features from current camera images, and matches them with 3D map landmarks locally in image space using a pose prior \hat{T}_{WB}^t . The state estimation module fuses the visual 2D-3D matches with the current wheel-odometry measurement to obtain a current vehicle pose estimate T_{WB}^t .

Using \hat{T}^t and the extrinsics calibration between the vehicle body and the individual camera frames, the landmark 3D points are projected into the current set of images, and matched with extracted keypoints in the following way: A landmark and a keypoint are only considered as a match candidate if their image space distance is smaller than $40px$. This avoids forming geometrically inconsistent matches. Further, a keypoint is preferably matched with the candidate landmark whose FREAK [3] descriptor is closest to the FREAK descriptor of the keypoint, using the Hamming distance metric. A *descriptor distance threshold* $\delta[px]$ is employed to limit the distance between the two descriptors, thus ensuring appearance consistency. The resulting 2D-3D matches are fed back into the state estimator where they form the visual localization constraints for the state update at time t . After optimizing the vehicle pose T_{WB}^t , the geometric consistency of every localization constraint is re-evaluated. For this, a *reprojection threshold* $\rho[px]$ is used to distinguish between inlier and outlier landmark observations. While the localization factors of outlier observations are removed, the localization factors of inlier observations are marginalized out together with the previous pose T_{WB}^{t-1} . An illustration of the map-tracking algorithm can be found in Figure 7.3.

3.3 Mapping

A base-map is built by tracking and triangulating local features along the trajectory of the first-session dataset. The resulting 3D landmark points are added to the map together with their median feature descriptors. Subsequently, more map sessions are added by localizing further datasets against the available (multi-)session map using map-tracking. Note that all landmarks in the resulting multi-session map are expressed in one common frame of reference \mathcal{F}_W . Similar local localization and mapping algorithms have been used in our previous work [8, 65], to which we kindly refer the interested reader for more details.

4 Evaluation

This section presents evaluation results on the following three key aspects. a) In a parameter study, the optimal values for the most important parameters of our localization system are identified. b) We further investigate the influence of different binary descriptors on the localization performance. c) In long-term experiments across vastly different appearance conditions in outdoor environments, the localization accuracy and recall using map-tracking are evaluated, and compared with the accuracy and recall resulting from using a global localization algorithm.

The subsequent section first describes the *UP-Drive* vehicle platform, including the sensor set-up, computing infrastructure, and provides details on the online operation. Additional sections are devoted to a brief description of the three dataset collections, and the evaluation metrics used in our experiments.

4.1 The *UP-Drive* Platform

The *UP-Drive* vehicle is equipped with a surround-view camera system consisting of four cameras with fish-eye distorted lenses. Images are recorded at $30Hz$ with a resolution of 640×400 pixels in gray-scale. Furthermore, wheel tick encoders and a low-end IMU provide odometry measurements, which are fused with the visual localization constraints as described in Section 3.2. The vehicle and sample images from the camera system are depicted in Figure 7.1. Localization is run in real-time at $10Hz$ on a consumer-grade computer with an Intel i7 CPU and 16GB of RAM. In particular, no GPU is required, neither for mapping, nor for localization. Furthermore, for bootstrapping map-tracking, a position prior is generated with a consumer-grade GPS sensor, while the orientation hypothesis is generated from orientations of near-by map poses.

4.2 Dataset Collections

UP-Drive

The *UP-Drive* dataset collection consists of 32 drives on the Volkswagen factory premises in Wolfsburg, Germany, recorded between December 2017 and December 2018. The total driving distance is approximately $300km$. The scenery resembles an urban environment, with busy streets, buses, zebra crossings, and pedestrians². This dataset collection not only covers seasonal appearance changes and a wide range of different weather conditions, it also contains datasets recorded at dusk and night-time. Five datasets, three from day-time, one at dusk, and one at night, are used to build a multi-session map. The remaining 27 datasets are used for evaluating the localization.

² Sample images can found online at https://github.com/ethz-asl/up-drive_visual_dataset/wiki/Sample-Images

NCLT

The *NCLT* [15] dataset collection consists of 27 recordings collected with a Segway platform on the Michigan University campus between January 2012 and April 2013. Analogous to the *UP-Drive* datasets, odometry poses based on wheel-tick encoders and an IMU sensor are fused in the state estimation framework. A *Ladybug 3* camera system is used, collecting images at 5Hz which are undistorted and down-scaled to a resolution of 808×616 pixels prior to being fed into our framework. The visited routes vary considerably from dataset to dataset. However, there is an approximately 750m long outdoor segment that is traversed, with some minor deviations, in almost all datasets in either one or the opposite direction. We therefore use this sub-segment of the campus for building a multi-session map using seven of the datasets. The remainder of the datasets are used for evaluating the localization. Similar to the *UP-Drive* datasets, the *NCLT* datasets cover seasonal and weather changes over an annual cycle.

KITTI

We further use *Sequence 00* of the *KITTI* [29] visual odometry benchmark dataset in our evaluation. It is the only *KITTI* dataset with significant segments of the trajectory revisited. We split the dataset in two parts, and use the first 170 seconds for mapping, and the remainder for localization. As opposed to the *UP-Drive* and the *NCLT* datasets, the appearance conditions in the *KITTI* drive thus remain similar between mapping and localization.

4.3 Metrics

Localization Recall

We measure localization recall $\mathbf{r}[\%]$ as the distance traveled while localized in relation to the total distance traveled in the respective dataset. Localization at time t is deemed successful if there are at least 10 inlier landmark observations after the pose optimization.

Localization Accuracy

The 6DoF localization accuracy is evaluated for each successfully localized set of images along the trajectory of a dataset by comparing the relative transformation between the estimated pose T_{WB}^t and the nearest vertex in the map, with the same quantity estimated by a reference solution [7]. For the *NCLT* datasets, ground-truth poses are available, which we employ to evaluate both the translation accuracy \mathbf{p}_{xyz}^e [m], and orientation accuracy θ_{xyz}^e [deg]. Note that the availability of ground-truth poses is a unique feature of *NCLT*, and the primary reason why we have decided to evaluate on the *NCLT* datasets, in addition to our self-collected *UP-Drive* datasets.

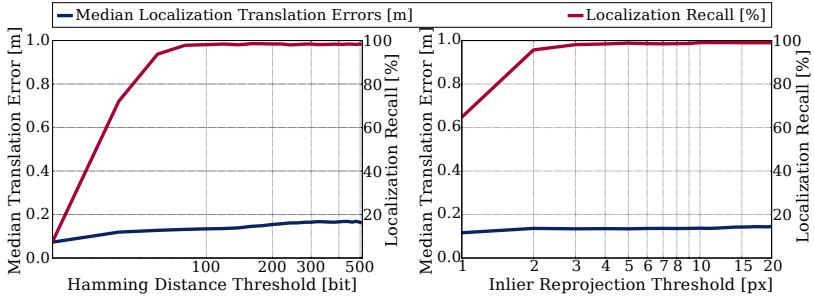


Figure 7.4: Localization recall r_{mt} (red) and median translation accuracy \bar{p}_{xyz}^e (blue) on the *NCLT* dataset from January 8th 2012, in relation to increasing values of the *descriptor distance threshold* δ (left), and *reprojection threshold* ρ (right) respectively, on a logarithmic scale. Even for very high values of δ and ρ , the vehicle remains accurately localized.

For the *UP-Drive* and *KITTI* datasets, no ground-truth poses are available. Both dataset collections provide, however, poses estimated by an *RTK GPS* sensor, which we use for producing a rough estimate of the localization accuracy on these datasets. Since the *RTK GPS* altitude estimates are unreliable, we only report on planar \mathbf{p}_{xy}^e [m] and lateral translation errors p_y^e [deg] on the *UP-Drive* and *KITTI* datasets.

4.4 Map-Tracking Parameter Study

As described in Section 3.2, there are two main parameters guiding the formation of 2D-3D localization constraints in the map-tracking module, namely the *descriptor distance threshold* δ [bits], and the *reprojection threshold* ρ [px]. While the *descriptor distance threshold* ensures appearance consistency by setting an upper bound on the descriptor distance for matching 2D keypoints with a 3D landmarks, the *reprojection threshold* enforces geometric consistency by discarding localization constraints if their respective reprojection error after the pose update is more than ρ pixels.

In Figure 7.4, the localization recall and median localization error are shown for increasing values of δ , and ρ respectively, for the *NCLT 2012-01-08* dataset. A fixed value of $\delta = 100$ bits, and $\rho = 3$ px is used unless the respective parameter is varied as indicated on the x-axis. As expected, localization recall quickly rises with increasing δ and ρ . Interestingly, the localization accuracy remains approximately constant, even for high values of δ and ρ . This may appear counter-intuitive at first. A descriptor distance threshold greater than 25% of the total descriptor length clearly allows for many wrong matches to be formed, and eventually ought to lead to false positive localizations. In order to understand why this scenario does not occur, it is important to note that, as described in Section 3.2, the *descriptor distance threshold* only serves to discard matches whose descriptor distance is above

δ bits. If there are multiple match candidates for a given keypoint in the image, the matching algorithm still attempts to pick the landmark with the lowest descriptor distance. Therefore, as long as there *are* sufficiently many correct matches that can be formed, our algorithm *will* find them, even with a very lean *descriptor distance threshold* δ .

A similar effect exists for the *reprojection threshold* too. As long as the pose prior is close to correct and there are sufficiently many valid 2D-3D matches, localization will not deviate from the correct trajectory, even with a very high ρ threshold.

Hence, as long as the vehicle is correctly localized, and there are enough valid localization matches *possible*, our localization system will *remain* correctly localized,

However, too high a value for δ and ρ may indeed derail the localization system if the pose prior is sufficiently wrong. In the remainder of this section, we therefore aim at finding the range of values for δ and ρ that guarantee no false-positive localization, even if the pose prior is wrong. Knowing this range is important in two ways. Firstly, it defines a safe operating space for choosing δ and ρ where a positive localization feedback, such as a certain number of inlier landmark observations, can be trusted. Secondly, it reveals a maximum degree of pose prior disturbance that can be tolerated when bootstrapping the map-tracking algorithm with any kind of auxiliary global localization input such as consumer grade GPS, or a place-recognition module. In order to evaluate these properties, we have conducted a parameter sweep experiment, varying both values for δ and ρ , as well as increasing the disturbance of the prior pose in yaw-angle, longitudinal, and lateral dimension separately. The resulting range of safe operating conditions is shown in Figure 7.5. The colors indicate the maximum disturbance, before either bootstrapping map-tracking is no longer possible, or, marked with an ‘X’, bootstrapping resulted in false-positive localization instead. It can be seen that for all three modes for disturbance, there is a safe range for both δ , and ρ , that guarantee convergence to correct localizations, even for considerably inaccurate prior poses with up to 10 degrees in yaw angle, and 3m meters in longitudinal and lateral direction. Furthermore, taking the results from both Figure 7.4, and 7.5, we find with $\delta = 100bits$, and $\rho = 3.0px$, a safe choice of parameters yielding maximum recall and sufficient robustness for bootstrapping map-tracking with a consumer-grade GPS sensor.

4.5 Binary Descriptor Comparison

In addition to the *descriptor distance threshold* and the *reprojection threshold*, the type of descriptor is another pivotal design choice, as it influences not only the localization recall, but also the size of the map. We restrict ourselves to the use of binary descriptors, as they can be matched very efficiently on a CPU-only platform, and compare the localization performance for three popular choices of binary local feature descriptors, namely FREAK [3], BRISK [43], and ORB [75]. Krajník et al. have evaluated the influence of various local feature descriptors for visual teach-and-repeat in [38]. They have, however, employed a global matching algorithm to find correspondences between two images recorded at the same location.

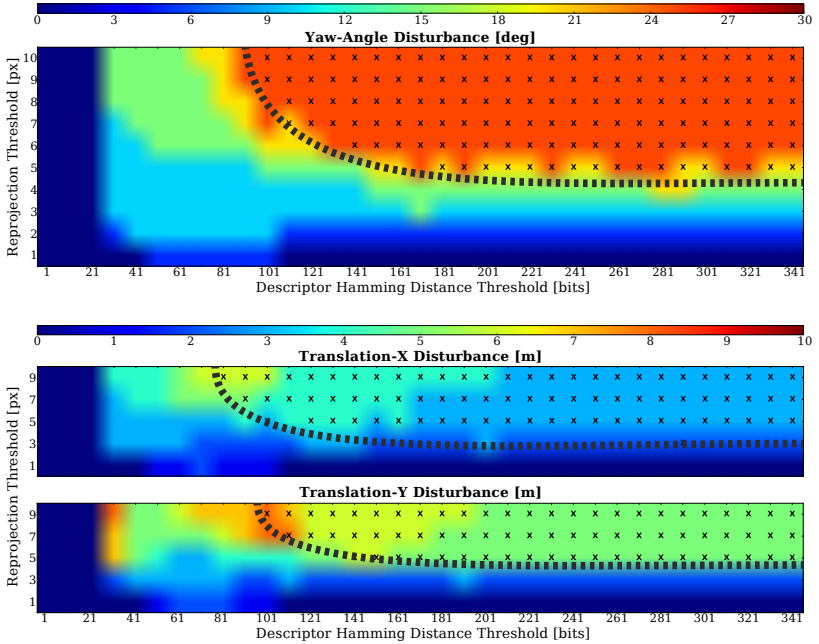


Figure 7.5: Sensitivity of the map-tracking pose prior in relation to increasing values for the *descriptor distance threshold* δ , and *reprojection threshold* ρ . The colors represent the maximum admissible degree of disturbance in yaw angle (top), longitudinal (middle), and lateral direction (bottom) leading to convergence of the localization to the true pose. The parameter combinations marked with 'X' denote unsafe operating regions, where high prior pose disturbances lead to false-positive localizations. In the remaining operating regions, localization fails if the prior pose disturbance is larger than the degree represented by the respective color. All three modes of disturbances reveal a safe region for the choice of δ and ρ allowing for guaranteed convergence towards the correct pose, while tolerant to significant disturbance in the prior pose.

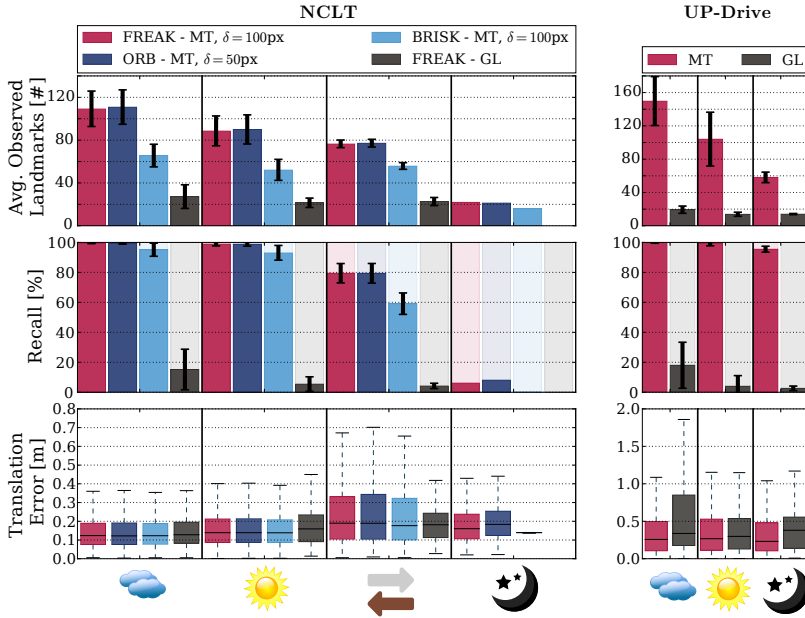


Figure 7.6: Average number of observed landmarks (top), localization recall (middle), and translation accuracy on the *NCLT* datasets (left), and *UP-Drive* datasets (right). For the *NCLT* datasets, the translation localization accuracy is evaluated using the *ground-truth* poses, while for the *UP-Drive* datasets, the planar translation errors with respect to the *RTK GPS* poses are shown. The datasets are grouped into categories according to appearance conditions (cloudy or rainy, sunny, and night-time) and traversal direction (indicated by the two opposing arrows). The localization performance using map-tracking (MT) is compared with global localization (GL). On the *NCLT* datasets, the map-tracking performance is further compared with different choices of binary descriptors.

In contrast to that, our map-tracking algorithm employs a pose prior and performs a local search in the image space. This variation in methodology leads to differing results as compared to [38]. While the evaluation by Krajník et al. suggests superior performance of BRISK as compared to FREAK and ORB, our experiments reveal worse performance of map-tracking with BRISK than with FREAK or ORB. This emphasizes the strong influence of the specific feature matching algorithm with respect to the localization performance using different types of features. In Table 7.1, the localization recall, the average number of observed landmarks, and the localization accuracy, are presented for different choices of descriptors, aggregated over all day-time *NCLT* datasets. Note that the *descriptor distance threshold* δ is set at 100bits for the two 64 byte long descriptors FREAK and BRISK, and at 50bit for the 32 byte long ORB descriptors, thereby allowing the same relative fraction of bits to be different when forming localization matches in all three cases. The performance using FREAK and ORB is nearly identical. This is remarkable, as the descriptor size of the latter is only half of that of FREAK. With BRISK, on the other hand, the average number of observed landmarks and the localization recall is significantly worse. However, the impact on the localization accuracy is marginal, as only the pose estimates of successful localizations are considered.

A more detailed evaluation of the descriptor comparison can be found in Figure 7.6, which shows the aforementioned metrics evaluated separately for groups of datasets formed according to the four categories exhibiting differing localization performance. The category *Cloudy* includes four, and the category *Sunny* 12, datasets labeled as (partially) cloudy, and sunny respectively, according to [15]. The category *Opposite* contains the two day-time datasets *2012-11-04*, and *2013-02-23* traversing the map in the opposite direction, while the *Night* category represents the only night-time dataset from December 1st 2012. The loss in recall with BRISK is primarily attributed to the two datasets traversing the map in opposite direction, where the recall with BRISK is approximately 20% lower than with FREAK or ORB. Contrary to that, the average number of observed landmarks remains roughly the same with BRISK across all three day-time categories, while FREAK and ORB observe significantly more landmarks when traversing the map in the predominant direction, both under cloudy skies, and in sunny conditions.

Based on these experiences, we suggest to use ORB as a binary descriptor for map-tracking, or FREAK in case there are no restrictions with respect to the map size.

4.6 Localization Accuracy and Recall

In order to fully rely on our visual localization system to control the car in the *UP-Drive* project, a high localization recall with an accuracy below 0.5m is paramount, as only short driving segments with no localization may be bridged with wheel-odometry before the car may deviate from its designated lane. We compare the localization recall and accuracy of our localization system using map-tracking with the metric global localization algorithm based on the work presented in [51] and available in the *maplab* framework [82]. We refer to the results using

	FREAK	ORB	BRISK
$\mathbf{r}_{mt}[\%]$	96.89 +/- 6.62	96.76 +/- 6.61	89.75 +/- 12.0
$\mathbf{obs}_{\emptyset}[\#]$	92.97 +/- 49.94	94.28 +/- 51.38	56.28 +/- 33.12
$\bar{\mathbf{p}}_{xyz}^e[\text{m}]$	0.14 [0.32]	0.14 [0.32]	0.14 [0.3]

Table 7.1: Descriptor comparison on the *NCLT* datasets, showing the average recall with map-tracking $\mathbf{r}_{mt}[\%]$, the average number of observed landmarks $\mathbf{obs}_{\emptyset}[\#]$, with standard deviations denoted by “+/-”, and the median translation localization accuracy $\bar{\mathbf{p}}_{xyz}^e[\text{m}]$, with the 90 percentile denoted in square brackets.

	$\mathbf{r}_{mt}[\%]$	$\mathbf{r}_{gl}[\%]$	$\bar{\mathbf{p}}_{xyz}^e / \bar{\mathbf{p}}_{xy}^e, \bar{\mathbf{p}}_y^e [\text{m}]$		$\bar{\theta}_{xyz}^e [^\circ]$
<i>NCLT</i>	96.89 +/-6.62	7.49 +/-8.59	0.14 [0.32]		1.23 [1.8]
<i>UP-Drive</i>	99.23 +/-1.75	8.94 +/-12.62	0.26 [0.88]	0.12 [0.58]	0.21 [0.33]
<i>KITTI</i>	96.05	94.24	0.43 [0.8]	0.31 [0.62]	0.26 [0.59]

Table 7.2: The aggregated localization performance on the *NCLT*, *UP-Drive*, and *KITTI* dataset(s), showing average localization recall with map-tracking \mathbf{r}_{mt} , and with global localization \mathbf{r}_{gl} , and the median translation ($\bar{\mathbf{p}}_{xyz}^e$) and orientation ($\bar{\theta}_{xyz}^e$) accuracy. For *UP-Drive* and *KITTI*, planar $\bar{\mathbf{p}}_{xy}^e$ and lateral $\bar{\mathbf{p}}_y^e$ errors are shown instead of full 3DoF translation errors. Standard deviations are denoted by “+/-”, and the 90 percentile is shown in square brackets.

this algorithm with *GL* in the respective figures and tables. Both algorithms, that is map-tracking and global localization, operate on the same multi-session maps, using the same landmarks. Note, however, that the global localization algorithm is fundamentally different to the map-tracking module presented in this paper, as in contrast to the former, the latter is able to exploit a pose prior. By including this comparison, we aim at highlighting the gain in localization recall attainable by using a local localization algorithm, as opposed to relying only a global localization algorithm. Map-tracking does, however, require *some* global localization module for bootstrapping, or re-localizations. As described in Section 4.1, a consumer grade GPS sensor serves this role on the *UP-Drive* vehicles.

The localization recall with map-tracking $\mathbf{r}_{mt}[\%]$, and with global localization $\mathbf{r}_{gl}[\%]$, and the localization accuracy are shown in Table 7.2, aggregated over all datasets of the three collections. Note that the *NCLT* night-time dataset from December 1st is excluded in the table. While map-tracking attains close to 100% recall on all three dataset collections, global localization fails for extended periods on the *NCLT* and *UP-Drive* datasets which both exhibit pronounced appearance change. On the *KITTI* drive, however, the appearance condition only undergo minor change, and global localization achieves with 94% a similarly high recall as map-tracking. This illustrates the challenge in finding enough correct feature matches with a global localization algorithm in multi-session maps that cover outdoor environments with various different appearance conditions. Solely relying on a global localization algorithm in these environments may thus not be sufficient to guarantee reliable localization in real-world applications. As our results show, exploiting a pose prior can help to significantly increase the reliability of the localization.

We further note that the planar median localization accuracy in *UP-Drive* and *KITTI* are below $0.5m$. Note that due to the different kind of reference sensors, these numbers are not directly comparable with the localization accuracy attained on the *NCLT* datasets, with the latter exhibiting a median translational accuracy of $11cm$. Furthermore, the *KITTI* vehicle is equipped with only a forward facing camera, while the *UP-Drive* vehicle has a surround view camera rig. This results in less strictly constraint position estimates on the *KITTI* dataset, which translate into significantly lower planar and lateral localization accuracy. For the driving performance, the lateral errors are most important. On the *UP-Drive* datasets, the median lateral error is below $15cm$, which is sufficient for a smooth steering of the car.

The median orientation errors are less effected by the difference in the camera rigs, and are well below 0.5 degrees for both the *UP-Drive* and *KITTI* datasets. In contrast to that, the orientation errors on the *NCLT* datasets are higher due to more vivid roll and pitch motions of the Segway platform, as compared to the car platforms in case of *UP-Drive* and *KITTI*.

A more detailed analysis of the localization recall and planar accuracy on the *NCLT* and *UP-Drive* datasets is shown in Figure 7.6, with datasets grouped into categories as described in Section 4.5. There are 10 drives of the *UP-Drive* dataset collection in the *Cloudy*, 15 in the *Sunny*, and two in the *Night* category respectively. Recordings in rainy conditions are categorized as *Cloudy*, since there is little difference in performance on rainy datasets as opposed to in dry cloudy conditions. Map-tracking reaches virtually 100% recall with a median localization accuracy of around $10cm$ for all the *NCLT* day-time datasets that traverse the map in the primary direction. The same high recall is also achieved for all day-time *UP-Drive* datasets, with a planar median localization accuracy with respect to RTK GPS of approximately $20cm$. The additional challenge for visual localization in sunny conditions is, however, reflected in a lower average number of observed landmarks in case of map-tracking, and in a significantly worse recall using global localization. Recall using map-tracking remains, however, unaffected.

In contrast to that, map-tracking performs significantly worse on the two *NCLT* datasets that traverse the map in the opposite direction, with considerably lower recall, and slightly lower localization accuracy. This is understandable, given that there is only one map session in opposite direction, whereas there are six traversing the map in the primary direction. However, this also reveals the limitations of matching landmarks projected into the cameras field-of-views under considerable viewpoint change. Here it is important to note the asymmetry of the *Ladybug* camera rig when traversing in the opposite direction, as the surround view is covered by an odd number of five cameras.

Furthermore, the only *NCLT* night-time dataset from December 1st 2012 fails to localize along most parts of the trajectory. Not only is this the only available recording under night-time conditions, but the Segway also traverses the map in the opposite direction, further exacerbating localization. A lack of more recordings from dusk or night-time renders it impossible to augment the multi-session map with the appearance conditions at night-time, and thus the conditions in this dataset lie



Figure 7.7: On the left, a sample image of a trajectory segment that fails to localize at night. A lack of structure and street lighting renders it unfeasible to match a sufficient number of map landmarks. A few meters later, street lighting is present (right side), and localization is picked up again.

outside the appearance coverage of the map. In contrast to the *NCLT* datasets, the *UP-Drive* datasets contain multiple recordings under both dusk and night-time conditions, allowing to extend the appearance coverage of the multi-session map with these conditions. Therefore, localization at night is successful in this case, even though the respective average recall is slightly less than 100% for the *UP-Drive* night-time datasets. This minor drop in recall is mainly attributed to the night-time recording from December 11th, which only attains a recall of 92%. Sample images of the route segment where localization fails on this dataset are depicted in Figure 7.7. In this part of the route, the car is driving up North on a ramp crossing numerous rail tracks. With a lack of both street lamps and near-by building structures, there are hardly any stable visual cues in this section, and our localization system fails to match sufficiently many landmarks from the map. Only later along the ramp, once artificial lighting on the railing to the left and right of the road boundary is present, localization is picked up again. This example demonstrates the current limitations of visual localization in night-time conditions. Even with high-performance CMOS cameras providing remarkably bright images at night, a certain amount of artificial street lighting and human made structure in the vicinity is required.

5 Conclusions

This paper presented a reliable visual localization system for urban outdoor environments. An extensive evaluation on several hundreds of kilometers of real-world driving conditions over the course of more than a year has demonstrated that our localization system is able to meet the requirement of high localization recall at high accuracy. Thereby, the appearance conditions encountered in our experiments not only cover various challenging weather conditions, wet road surfaces, sun reflections, and seasonal changes, but also night-time conditions. A comparison with a state-of-the-art global metric localization algorithm has revealed a large increase

in recall attainable by instead employing a local localization algorithm, such as the map-tracking algorithm described in this paper. Additionally, a comparison of binary feature descriptors suggests superior performance of map-tracking when using FREAK or ORB, as compared to using BRISK. In a thorough parameter study, we have further investigated the boundary conditions of our map-tracking module and validated a safe range for selecting the most critical parameters in order to guarantee reliable localization.

Bibliography

- [1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building Rome in a day. In *IEEE International Conference on Computer Vision*, 2009.
- [2] A. Aizawa. An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 2003.
- [3] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [4] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [5] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, 2006.
- [6] M. Bloesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart. Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback. *International Journal of Robotics Research*, 2017.
- [7] W. Burgard, C. Stachniss, G. Grisetti, B. Steder, R. Kümmerle, C. Dornhege, M. Ruhnke, A. Kleiner, and J. D. Tardós. A comparison of SLAM algorithms based on a graph of relations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.
- [8] M. Bürki, I. Gilitschenski, E. Stumm, R. Siegwart, and J. Nieto. Appearance-based landmark selection for efficient long-term visual localization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.
- [9] M. Bürki, M. Dymczyk, I. Gilitschenski, C. Cadena, R. Siegwart, and J. Nieto. Map management for efficient long-term visual localization in outdoor environments. In *IEEE Intelligent Vehicles Symposium*, 2018.
- [10] M. Bürki, L. Schaupp, M. Dymczyk, R. Dubé, C. Cadena, R. Siegwart, and J. Nieto. Vizard: Reliable visual localization for autonomous vehicles in urban outdoor environments. *arXiv preprint*, 2019.

- [11] M. Burri, M. Bloesch, D. Schindler, I. Gilitschenski, Z. Taylor, and R. Siegwart. Generalized information filtering for mav parameter estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.
- [12] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 2016.
- [13] M. Calonder, V. Lepetit, M. Özuysal, T. Trzcinski, C. Strecha, and P. Fua. BRIEF: Computing a local binary descriptor very fast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.
- [14] N. Carlevaris-Bianco and R. M. Eustice. Learning temporal co-observability relationships for lifelong robotic mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, Workshops*, page 15, 2012.
- [15] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice. University of Michigan North Campus long-term vision and lidar dataset. *International Journal of Robotics Research*, 2016.
- [16] W. Churchill and P. Newman. Experience-based Navigation for Long-Term Localisation. *International Journal of Robotics Research*, 2013.
- [17] T. Cieslewski, S. Lynen, M. Dymczyk, S. Magnenat, and R. Siegwart. Map api-scalable decentralized map building for robots. In *IEEE International Conference on Robotics and Automation*, 2015.
- [18] L. Clement, J. Kelly, and T. D. Barfoot. Robust Monocular Visual Teach and Repeat Aided by Local Ground Planarity and Color-constant Imagery. *Journal of Field Robotics*, 2017.
- [19] M. Cummins and P. Newman. Appearance-only SLAM at large scale with FAB-MAP 2.0. *International Journal of Robotics Research*, 2011.
- [20] F. Dayoub and T. Duckett. An adaptive appearance-based map for long-term topological localization of mobile robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008.
- [21] F. Dayoub, G. Cielniak, and T. Duckett. Long-term experiments with an adaptive spherical view representation for navigation in changing environments. *Robotics and Autonomous Systems*, 2011.
- [22] F. Dellaert. Factor graphs and gtsam: A hands-on introduction. Technical report, Georgia Institute of Technology, 2012.

-
- [23] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. *arXiv*, 2017.
- [24] M. Dymczyk, S. Lynen, M. Bosse, and R. Siegwart. Keep it brief: Scalable creation of compressed localization maps. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.
- [25] M. Dymczyk, S. Lynen, T. Cieslewski, M. Bosse, R. Siegwart, and P. Furgale. The gist of maps - Summarizing experience for lifelong localization. In *IEEE International Conference on Robotics and Automation*, 2015.
- [26] M. Dymczyk, T. Schneider, I. Gilitschenski, R. Siegwart, and E. Stumm. Erasing bad memories: Agent-side summarization for long-term mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.
- [27] L. Fayin and J. Košecká. Probabilistic location recognition using reduced feature set. In *IEEE International Conference on Robotics and Automation*, 2006.
- [28] D. Gálvez-López, M. Salas, J. D. Tardós, and J. Montiel. Real-time monocular object slam. *Robotics and Autonomous Systems*, 2016.
- [29] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, 2013.
- [30] W. Hartmann, M. Havlena, and K. Schindler. Predicting Matchability, 2014.
- [31] S. Hochdorfer and C. Schlegel. Towards a robust visual SLAM approach: Addressing the challenge of life-long operation. In *International Conference on Advanced Robotics*, 2009.
- [32] E. Johns and G. Z. Yang. Feature Co-occurrence Maps: Appearance-based localisation throughout the day. In *IEEE International Conference on Robotics and Automation*, 2013.
- [33] E. Johns and G. Z. Yang. Generative methods for long-term place recognition in dynamic scenes. *International Journal of Computer Vision*, 2014.
- [34] J. Knopp, J. Sivic, and T. Pajdla. Avoiding Confusing Features in Place Recognition. In *European Conference on Computer Vision*, 2010.
- [35] K. Konolige and J. Bowman. Towards lifelong visual maps. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.
- [36] T. Krajník, J. P. Fentanes, O. M. Mozos, T. Duckett, J. Ekekrantz, and M. Hanheide. Long-term topological localisation for service robots in dynamic environments using spectral maps. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014.

- [37] T. Krajník, J. Pulido Fentanes, M. Hanheide, and T. Duckett. Persistent localization and life-long mapping in changing environments using the Frequency Map Enhancement. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.
- [38] T. Krajník, P. Cristoforis, K. Kusumam, P. Neubert, and T. Duckett. Image features for visual teach-and-repeat navigation in changing environments. *Robotics and Autonomous Systems*, 2017.
- [39] H. Lategahn and C. Stiller. City GPS using stereo vision. In *International Conference on Vehicular Electronics and Safety*, 2012.
- [40] H. Lategahn and C. Stiller. Vision-only localization. *Transactions on Intelligent Transportation Systems*, 2014.
- [41] H. Lategahn, J. Beck, and C. Stiller. DIRD is an illumination robust descriptor. In *IEEE Intelligent Vehicles Symposium*, 2014.
- [42] M. Lauer, C. G. Keller, C. Stiller, et al. Mapping and localization using surround view. In *IEEE Intelligent Vehicles Symposium*, 2017.
- [43] S. Leutenegger, M. Chli, and R. Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *IEEE International Conference on Computer Vision*, 2011.
- [44] Y. Li, N. Snavely, and D. P. Huttenlocher. Location Recognition using Prioritized Feature Matching. *European Conference on Computer Vision*, 2010.
- [45] C. Linegar, W. Churchill, and P. Newman. Work Smart , Not Hard : Recalling Relevant Experiences for Vast-Scale but Time-Constrained Localisation. *IEEE International Conference on Robotics and Automation*, 2015.
- [46] L. Liu, H. Li, and Y. Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *IEEE International Conference on Computer Vision*, 2017.
- [47] A. Loquercio, M. Dymczyk, B. Zeisl, S. Lynen, I. Gilitschenski, and R. Siegwart. Efficient descriptor learning for large scale localization. In *IEEE International Conference on Robotics and Automation*, 2017.
- [48] D. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision*, 1999.
- [49] S. Lowry, N. Sunderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual Place Recognition: A Survey. *IEEE Transactions on Robotics*, 2016.

-
- [50] S. Lynen, M. Bosse, P. Furgale, and R. Siegwart. Placeless Place-Recognition. In *International Conference on 3D Vision*, 2014.
- [51] S. Lynen, T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart. Get out of my lab: Large-scale, real-time visual-inertial localization. In *Robotics: Science and Systems*, 2015.
- [52] K. Mactavish, M. Paton, and T. D. Barfoot. Visual triage: A bag-of-words experience selector for long-term visual route following. In *IEEE International Conference on Robotics and Automation*, 2017.
- [53] W. Maddern, M. Milford, and G. Wyeth. Cat-slam: probabilistic localisation and mapping using a continuous appearance-based trajectory. *International Journal of Robotics Research*, 2012.
- [54] W. Maddern, A. D. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman. Illumination Invariant Imaging: Applications in Robust Vision-based Localisation, Mapping and Classification for Autonomous Vehicles. In *IEEE International Conference on Robotics and Automation, Workshops*, 2014.
- [55] C. McManus, W. Churchill, W. Maddern, A. D. Stewart, and P. Newman. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *IEEE International Conference on Robotics and Automation*, 2014.
- [56] C. McManus, B. Upcroft, and P. Newman. Scene Signatures : Localised and Point-less Features for Localisation. In *Robotics Science and Systems*, 2014.
- [57] M. Milford and G. Wyeth. Persistent navigation and mapping using a biologically inspired slam system. *International Journal of Robotics Research*, 2010.
- [58] M. Milford, G. Wyeth, and D. Prasser. RatSLAM: a hippocampal model for simultaneous localization and mapping. In *IEEE International Conference on Robotics and Automation*, 2004.
- [59] M. J. Milford and G. F. Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *IEEE International Conference on Robotics and Automation*, 2012.
- [60] M. J. Milford, D. Prasser, and G. F. Wyeth. Experience Mapping : Producing Spatially Continuous Environment Representations using RatSLAM. In *Australasian Conference on Robotics and Automation*, 2005.
- [61] M. Mohan, D. Gálvez-López, C. Monteleoni, and G. Sibley. Environment selection and hierarchical place recognition. In *IEEE International Conference on Robotics and Automation*, 2015.

- [62] B. Mu, A.-a. Agha-mohammadi, L. Paull, M. Graham, J. How, and J. Leonard. Two-Stage Focused Inference for Resource-Constrained Collision-Free Navigation. 2015.
- [63] P. Muehlfellner, P. Furgale, W. Derendarz, and R. Philippsen. Evaluation of fisheye-camera based visual multi-session localization in a real-world scenario. In *IEEE Intelligent Vehicles Symposium*, 2013.
- [64] P. Mühlfellner, P. Furgale, W. Derendarz, and R. Philippsen. Designing a Relational Database for Long-Term Visual Mapping. 2015.
- [65] P. Mühlfellner, M. Bürki, M. Bosse, W. Derendarz, R. Philippsen, and P. Furgale. Summary Maps for Lifelong Visual Localization. *Journal of Field Robotics*, 2016.
- [66] R. Mur-Artal and J. D. Tardós. Visual-inertial monocular slam with map reuse. *IEEE Robotics and Automation Letters*, 2017.
- [67] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss. Robust visual robot localization across seasons using network flows. In *AAAI Conference on Artificial Intelligence*, 2014.
- [68] L. Nicholson, M. Milford, and N. Sünderhauf. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters*, 2019.
- [69] M. Paton, K. MacTavish, C. J. Ostafew, and T. D. Barfoot. It’s not easy seeing green: Lighting-resistant stereo Visual Teach & Repeat using color-constant images. In *IEEE International Conference on Robotics and Automation*, 2015.
- [70] M. Paton, K. MacTavish, M. Warren, and T. D. Barfoot. Bridging the appearance gap: Multi-experience localization for long-term visual teach and repeat. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.
- [71] M. Paton, K. Mactavish, M. Warren, and T. D. Barfoot. Bridging the appearance gap: Multi-experience localization for long-term visual teach and repeat. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.
- [72] K. Pirker, M. Ruther, and H. Bischof. CD SLAM - continuous localization and mapping in a dynamic world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011.
- [73] D. Prasser, M. Milford, and G. Wyeth. Outdoor simultaneous localisation and mapping using RatSLAM. *Field and Service Robotics*, 2006.

-
- [74] D. M. Rosen, J. Mason, and J. J. Leonard. Towards lifelong feature-based mapping in semi-static environments. In *IEEE International Conference on Robotics and Automation*, 2016.
- [75] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *IEEE International Conference on Computer Vision*, 2011.
- [76] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [77] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 1988.
- [78] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. *arXiv*, 2018.
- [79] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. In *IEEE International Conference on Computer Vision*, 2011.
- [80] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [81] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [82] T. Schneider, M. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart. maplab: An open framework for research in visual-inertial mapping and localization. *IEEE Robotics and Automation Letters*, 2018.
- [83] M. Sons and C. Stiller. Efficient multi-drive map optimization towards life-long localization using surround view. In *International Conference on Intelligent Transportation Systems*, 2018.
- [84] H. Strasdat, J. M. Montiel, and A. J. Davison. Visual slam: why filter? *Image and Vision Computing*, 2012.
- [85] E. Stumm, C. Mei, S. Lacroix, and M. Chli. Location graphs for visual place recognition. In *IEEE International Conference on Robotics and Automation*, 2015.

- [86] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [87] P. Turcot and D. G. Lowe. Better matching with fewer features: The selection of useful features in large database recognition problems. In *IEEE International Conference on Computer Vision, Workshops*, 2009.
- [88] A. Walcott-Bryant, M. Kaess, H. Johannsson, and J. J. Leonard. Dynamic pose graph SLAM: Long-term mapping in low dynamic environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.
- [89] S. Yang and S. Scherer. Direct monocular odometry using points and lines. In *IEEE International Conference on Robotics and Automation*, 2017.
- [90] F. Zhang, T. Rui, C. Yang, and J. Shi. Lap-slam: A line-assisted point-based monocular vslam. *Electronics*, 2019.

Curriculum Vitae

Mathias Bürki

born October 28, 1987

citizen of Ennetbaden AG, Switzerland

- 2013–2019 *ETH Zürich, Switzerland*
Doctoral studies at the Autonomous Systems Lab, Supervised
by Prof. Roland Siegwart
- 2011–2013 *ETH Zürich, Switzerland*
Master of Science in Robotics, Systems and Control
- 2008–2011 *ETH Zürich, Switzerland*
Bachelor of Science in Computer Science
- 2004–2007 *Secondary School, Alte Kantonsschule Aarau, Switzerland*
Matura, Focus on Natural Sciences
- 1996–2004 *Primary School, Suhr, Switzerland*