DISS. ETH NO. 26162

VIEWPOINT-TOLERANT PLACE RECOGNITION FOR UNMANNED AERIAL VEHICLES USING VISION

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

FABÍOLA ALVARES RODRIGUES DE SOUZA MAFFRA

Mestra em Informática,
Pontifícia Universidade Católica do Rio de Janeiro

born on May 21, 1982
citizen of Brazil

accepted on the recommendation of

Prof. Dr. Margarita Chli, Examiner
Prof. Dr. Marc Pollefeys, Co-examiner
Prof. Dr. José María Martínez Montiel, Co-examiner

2019

Vision for Robotics Lab
Department of Mechanical and Process Engineering
ETH Zurich
Switzerland

Abstract

With the emergence of powerful techniques for robotic egomotion estimation and map building that follow the SLAM (Simultaneous Localization And Mapping) paradigm, Place Recognition has become of fundamental importance for robotic autonomy. Addressing Place Recognition by determining whether a robot returns to a previously visited location, which is widely known as the Loop-closure Detection problem, is a key competence to enable the creation of accurate maps and even recovery from complete localization failures, essentially opening up the way towards long-term autonomous robot navigation. However, the deployment of robotic platforms for long periods of time or for multiple missions taking place months or years apart from each other, can pose major challenges in Place Recognition, due mainly by the large appearance variability that a place may experience over time, such as seasonal and lighting changes, weather conditions as well as human activity. Trajectory and viewpoint variations are common even in shorter-term missions. The views of a street from a car, for example, when it is navigating from the opposite directions can be rather different. Considering the navigation in a scene using an Unmanned Aerial Vehicle (UAV), however, the viewpoint changes experienced are far more challenging, and this is especially the case in flights with rotorcraft UAVs, which are able to move with great agility in 3D.

In this thesis, we address the problem of viewpoint-tolerant Place Recognition for autonomous robot navigation. More specifically, we have focused our efforts on the development of approaches that are suitable for small UAVs with restricted payload onboard and as a result, limited computational capabilities. Deep learning approaches addressing Place Recognition have been demonstrated to perform very well under isolated variations in appearance. The power of these methods, however, stems from specific training on the expected scene variations and complex computational effort. This, in turn, imposes the need for extensive training datasets and powerful Graphics Processing Units (GPUs), which are often unavailable onboard small aircraft, rendering the use of such methods at least impractical in aerial navigation. On the other hand, more affordable and scalable feature-based techniques building on the efficient Bag-of-Words (BoW) representation exist in the literature, however, these methods are known to fail dramatically in the presence of large appearance and viewpoint changes. This is largely due to the fact that BoW approaches discard all geometric information of the scene structure by design.

Inspired by the need for lightweight and effective techniques for Place Recognition onboard small aircraft, this thesis investigates ways to render feature-based approaches capable of coping with the variability of places when experienced from such small aircraft, while bounding the onboard computation effort for real-time

operation. As a result, this thesis describes a set of novel approaches for viewpoint-tolerant Place Recognition progressively building on top of each other, achieving unprecedented robustness with relation to the state-of-the-art. Assuming that a nominal, keyframe- and vision-based SLAM framework is running onboard the robot, this thesis advocates the power of exploiting both 2D visual information inherent in images, as well as the often noisy estimates of the local 3D geometry captured by SLAM in deciding on whether the robot is in the presence of a loop. Across all approaches proposed here, a BoW image representation is used in combination with efficient binary image features to enable fast image retrieval. Any loop-closure candidates from the database of all robot's experiences matching a query image is then subjected to geometric verification. This entails a test for matching constellations of the visible image features in an attempt to reject false appearance matches returned by the image retrieval step. Along with investigating efficient and robust geometric tests to avoid false positive loop-closures, different image and scene representations have been investigated. Namely, the first approach proposed, employs orthophotos to create a well-conditioned problem to address orientation tolerance, demonstrating better recall than counterpart methods relying on perspective images in urban environments, where the presence of large planar structures can be assumed. Pushing for more general scenarios, and relaxing this assumption, the second method for lightweight Place Recognition proposed in this thesis is a new, carefully designed pipeline to support low-burden computation and to take advantage of any scale and rotation invariance offered by binary descriptors by using combined geometric checks that make use of both 2D and 3D information. Tests in both hand-held and aerial datasets exhibiting large viewpoint and appearance changes have revealed unprecedented recall for perfect precision for this pipeline in comparison to the state of the art. However, it was only with the extension of this pipeline with a scene-depth completion module to densify the map of the local scene (i.e. "place"), described in the final method proposed here, that indeed tolerance to extreme viewpoint changes of up to 45° was achieved. This comprises a drastic improvement in viewpoint tolerance when compared with the state of the art today, demonstrating that feature-based approaches still have a lot to offer in Place Recognition at extreme viewpoint changes.

Throughout the research conducted for this thesis, several synthetic and real datasets, with both hand-held and aerial footage, were captured and made publicly available. Inspired by the lack of such datasets in the literature and the need to benchmark methods, these datasets were designed to present large appearance changes and extreme viewpoint variations (0-45°). In particular, our synthetic datasets are, to the best of our knowledge, the first to isolate the problem of viewpoint changes for Place Recognition, addressing a crucial gap in the literature. Tackling real-time, viewpoint-tolerant Place Recognition for lightweight single- or multi-robot applications, as well as releasing novel benchmarking datasets, the research findings of this thesis push the boundaries of vision-based aerial navigation, but also shed light to new research directions towards long-term robot autonomy in real missions. The prospect of leveraging the benefits of both feature- and learning-based approaches to go beyond viewpoint-tolerance and addressing the

open problem of combined tolerance to common challenges, such as seasonal and illumination changes as well as higher-level reasoning for perceptual aliasing, opens up exciting opportunities for added robotic intelligence and autonomy.

Riassunto

Con l'emergere di tecniche sempre più potenti per la stima del movimento di un robot e la costruzione di mappe seguendo i paradigmi dello SLAM (Simultaneous Localization And Mapping), il problema del Place Recognition ha assunto un'importanza fondamentale nell'attività di ricerca rivolta all'autonomia dei robot. Cercare di determinare tramite tecniche di Place Recognition se un robot abbia fatto ritorno in un luogo precedentemente visitato, problema conosciuto comunemente come Loop-closure Detection, è una capacità essenziale per permettere la costruzione di mappe accurate e di recuperare la stima della posizione in caso di errori nel processo di localizzazione, aprendo di fatto la via verso la navigazione autonoma a lungo termine dei robot. Ciononostante, l'utilizzo di robot in missioni che richiedono lunghi periodi di tempo o che sono da svolgersi a distanza di mesi o anni tra di loro, può porre delle grandi difficoltà nel Place Recognition, principalmente a causa della notevole variabilità a cui un luogo può essere soggetto nel tempo, dovuta ad esempio a cambiamenti stagionali o delle condizioni di illuminazione, alle condizioni meteo e all'attività umana. Le variazioni di traiettoria e di punto di osservazione sono tuttavia comuni anche nelle missioni di breve periodo. Ad esempio, le esperienze visive di una strada vista da un'auto possono essere notevolmente differenti a seconda della direzione di guida. Considerando la navigazione di un Unmanned Aerial Vehicle (UAV), tuttavia, i cambiamenti del punto di osservazione sono nettamente più difficoltosi, specialmente in caso di volo di UAV multirotori, a causa della loro agilità nei movimenti in 3D.

In questa tesi, viene affrontato il problema del Place Recognition robusto al cambiamento del punto di osservazione della scena per la navigazione autonoma dei robot. Più precisamente, abbiamo concentrato i nostri sforzi sullo sviluppo di approcci che siano adatti a UAV di piccole dimensioni, con ridotta capacità di carico utile e, di conseguenza, con potere computazionale limitato. È stato dimostrato che gli approcci di Deep Learning che affrontano il problema del Place Recognition ottengono ottimi risultati in caso di variazioni isolate nell'aspetto dei luoghi precedentemente visitati. La forza di questi metodi tuttavia deriva da un training, o allenamento, specifico sui cambiamenti attesi e da un notevole sforzo computazionale. Queste caratteristiche impongono di conseguenza la necessità di vasti dataset di allenamento e di potenti GPU (Graphics Processing Units), che spesso non sono disponibili a bordo di veicoli di piccole dimensioni, rendendo l'uso di tali metodologie non pratiche in caso di navigazione aerea. D'altra parte, in letteratura esistono tecniche computazionalmente meno costose e meglio scalabili basate sull'estrazione di features, come l'efficiente Bag-of-Words (BoW). Tuttavia, è noto che questi metodi falliscono in modo drastico in caso di grandi cambiamenti

di aspetto e del punto di osservazione. Questo limite è dovuto al fatto che, per scelte di progettazione, l'approccio del BoW scarta tutte le informazioni relative alla struttura geometrica della scena.

Partendo dalla necessità di tecniche computazionalmente trattabili ed efficaci da utilizzarsi a bordo di veicoli di dimensioni ridotte per il problema del Place Recognition, questa tesi si concentra sugli approcci basati sull'estrazione di features al fine di renderli capaci di affrontare i cambiamenti nell'aspetto dei luoghi quando sperimentati da droni, cercando di limitare lo sforzo computazionale per permettere operazioni in tempo reale. Come risultato, questa tesi introduce una serie di nuove metodologie per Place Recognition robuste ai cambiamenti del punto di osservazione che si vanno a basare progressivamente l'una sull'altra, raggiungendo un livello di robustezza senza precedenti nello stato dell'arte. Supponendo che un sistema di SLAM, basato esclusivamente su fotocamere e sull'estrazione di keyframes, sia operativo a bordo del robot, questa tesi sostiene l'importanza di sfruttare sia le informazioni visive 2D inerenti alle immagini, sia le misure della geometria 3D catturata dallo SLAM, seppur sottoposte a rumore, al fine di decidere se un robot sia in presenza di un luogo precedentemente visitato, o Loop-Closure. In tutti gli approcci proposti, una rappresentazione basata su BoW è usata in combinazione con features binarie estratte dall'immagine per consentire un rapido recupero delle informazioni. Ogni candidato per un Loop-Closure, estratto dal database di tutte le esperienze passate di un robot e corrispondente a una immagine di query, è soggetto a una verifica di tipo geometrico. Ciò impone un test per verificare la corrispondenza delle costellazioni delle features visibili nelle immagini, nel tentativo di rifiutare le corrispondenze errate ottenute nel passaggio di recupero dell'immagine dal database. Oltre allo sviluppo di test geometrici efficienti e robusti per evitare falsi positivi nei Loop-Closures, sono state investigate diverse rappresentazioni delle immagini e delle scene. Il primo approccio proposto utilizza ortofoto al fine di costruire un problema ben posto per migliorare la robustezza alla variazione di orientazione del punto di vista, mostrando come in ambienti urbani, dove è possibile ipotizzare la presenza di grandi strutture piane, sia possibile ottenere un miglior recall rispetto ad altri metodi basati su immagini proiettive. Muovendosi verso scenari più generici e rilassando l'ipotesi di planarità, il secondo metodo per Place Recognition proposto in questa tesi è un nuovo approccio attentamente progettato, in modo tale da essere caratterizzato da un basso carico computazionale e dalla capacità di sfruttare l'invarianza alla scala e alla rotazione dei descrittori binari, utilizzando una combinazione di controlli geometrici che impiegano sia le informazioni 2D sia 3D. I test nei dataset, sia nei casi hand-held sia aerei, caratterizzati da grandi cambiamenti nella direzione di osservazione ed nell'aspetto dei luoghi, mostrano, data una precisione perfetta, un recall senza precedenti nello stato dell'arte. Tuttavia, solo con l'estensione dell'approccio con un modulo per il calcolo della profondità della scena osservata, al fine di densificare la mappa locale (i.e. il "luogo"), descritto nell'ultimo metodo qui introdotto, la tolleranza a cambiamenti estremi del punto di vista riesce a raggiungere i 45°. Ciò ha portato ad un miglioramento drastico rispetto allo stato dell'arte odierno della robustezza al cambiamento della direzione di osservazione, mostrando come gli approcci basati

sull'estrazione di features abbiano ancora molto da offrire al problema del Place Recognition soggetto a cambiamenti estremi del punto di vista.

Durante la ricerca condotta per questa tesi, diversi dataset, sia hand-held sia arei, sono stati raccolti e resi pubblici. Ispirati dalla mancanza di tali dataset in letteratura e dalla necessità di metodi per effettuare comparazioni dei risultati, questi dati sono caratterizzati da grandi cambiamenti di aspetto e da variazioni estreme del punto di osservazione (0-45°). In particolare, i nostri dataset sintetici sono, al meglio delle nostre conoscenze, i primi a isolare il problema delle variazioni del punto di vista per il Place Recognition, colmando una lacuna cruciale presente in letteratura. Affrontando il problema del Place Recognition robusto ai cambiamenti dei punti di osservazione, utilizzabile in tempo reale in uno o più robot di dimensioni ridotte, e allo stesso tempo pubblicando nuovi dataset per attività di comparazione, i risultati di questa tesi non solo spingono i limiti della navigazione aerea basata su sensori visivi, ma gettano luce anche su nuove direzioni di ricerca verso l'uso di robot autonomi in missioni a lungo termine nel mondo reale. La prospettiva di sfruttare i benefici dei metodi basati sia sull'estrazione di features sia sul Deep Learning per affrontare la tolleranza ai cambiamenti dei punti di osservazione e altri problemi comuni non ancora risolti, come cambiamenti stagionali e di illuminazione, nonché il ragionamento ad alto livello per l'aliasing percettivo, apre nuove eccitanti opportunità per lo sviluppo dell'intelligenza e dell'autonomia dei robot.

Acknowledgements

First, I would like to express my deepest gratitude to Prof. Dr. Margarita Chli, who trusted me and gave me the opportunity to do my PhD studies at the Vision for Robotics Lab. I dearly appreciate her patient guidance, encouragement and useful criticisms throughout the course of my research. Her willingness to give her time so generously has been very much appreciated.

I thank my examiners Prof. José María Martínez Montiel and Prof. Marc Pollefeys, for their time and for their insightful comments and questions during my doctoral exam.

I thank my fellow labmates for making my experience at Vision for Robotics Lab even more exciting and fun. The inspiring discussions and fun moments will never be forgotten.

Some special words of gratitude go to my friends who have always been a major source of support and happiness. In particular, I would like to thank my friends Luana, Vanessa and Roberto. Thanks guys for always being there for me.

I would like to thank my beloved Prof. Dr. Marcelo Gattass whose encouragement was key to make me start a PhD. His guidance and support since my college days will always be remembered.

I wish to express my deepest gratitude to my family, especially my parents for their support and unconditional love over all these years. I wish to thank my brother Sergio who persuaded me to follow in his footsteps, my sister Rebeca and my brother-in-law Eduardo for being always present in my life. I also would like to thank my extended family, Lya, Jorge, Marcia, Vladimir and Ananda, for all the support and encouragement over the last 15 years. Finally, my deepest debt of gratitude goes to my husband Lucas, who stayed on my side through the most difficult times of this journey and beyond. Without you this accomplishment would not be half as sweet.

November 29, 2019 *Fabiola Maffra*

Financial Support

The research leading to these results has received funding from the Swiss National Science Foundation (SNSF, Agreement no. PP00P2 157585), EC's Horizon 2020 Programme under grant agreement n. 644128 (AEROWORKS) and NCCR Robotics.

Contents

Preface

This is a cumulative doctoral thesis and as such, it comprises the most relevant works published by the author during her doctoral studies. Chapter 1 introduces the overall problem addressed in this thesis, Chapter 2 details the main contributions achieved, and Chapter 3 concludes this thesis by presenting a summary of the main achievements and discussing future directions to address remaining open problems in long-term Place Recognition for UAVs and other robots. All peer-reviewed papers composing the main contributions of this thesis are attached at the end of this document.

Chapter 1

Introduction

Recent years have been marked by a sharp growth of the UAV industry, with several off-the-shelf consumer UAVs appearing in the market and commercial solutions that go beyond aerial photography and media coverage. With applications ranging from digitization of archaeological sites to search-and-rescue, there are many tasks where autonomous UAVs can make a real difference. Motivated by their potentially great impact, booming research attention has been dedicated in automating the navigation of UAVs driven by both Academia and Industry.

The ability of a mobile robot to navigate in the environment and reliably find its way between known locations, while avoiding obstacles and unsafe conditions is a key capability for realistic autonomy. While many competencies are required by a robot to successfully navigate in its environment, accurate localization is a fundamental requirement for navigational autonomy [100]. By determining its pose in the environment, the robot can make decisions based on what it perceives, and given this information, it can plan its future actions. While the use of external guidance, such as beacons and GPS signals, are common ways to address localization for robot navigation, such cues are not always available or reliable. For example, while GPS signals can be partially occluded in the close vicinity of large structures or even be entirely unavailable, such as indoors or underground, the use of beacons requires modifications in the environment, which is not always possible. Consequently, for a robot to be truly autonomous it needs to rely on its own sensor-suite in order to gain sufficient understanding of its surroundings to determine its pose within the environment. Localization can be addressed with Place Recognition, which formally, is the problem of determining whether a robot is re-visiting an area, where it has been to before. This is typically addressed by comparing the sensing cues (e.g. the current image) that the robot experiences at its current location against all past sensing experiences that live in a database and are associated to specific 'places' in the world. A 'place', in this context, can be a location in a topographic map of the scene, or even refer to the pose of a robot in

a metric map. In some cases, such as in robot localization, the outcome of a Place Recognition instance is not only a boolean decision on whether the robot has been to the current place in the past, but also a pose transformation relating the two places, the current and the matched one.

As Place Recognition addresses the comparison of the representation of different places in the world, different sensing cues can be employed to this end. Laser- and sonar-based approaches [30, 31], for example, compare local signatures of the scene structure, however, it is now well understood that vision provides a very favourable balance between richness of information encoded in images and the affordability, portability and ubiquity of cameras. As a result, vision-based Place Recognition as addressed in this thesis, is now the most common practice in the field.

Vision-based Place Recognition, however, is a challenging task due to the large variability in a scene's appearance that can be observed in the real world, caused by changes in the time of the day, weather or seasonal effects, human activities and dynamic objects. Figure 1.1 illustrates some challenging recognition tasks due to environmental changes. Conversely, different locations can appear identical, as shown in Figure 1.2. This so-called "perceptual aliasing" is a key problem in robotic perception and Place Recognition, attracting great research interest. On top of all these challenges that the research community has been trying to address over the past years, tackling Place Recognition for a small UAV adds on to the challenge, as the same scene can be experienced from drastically different viewpoints, resulting to very different depictions of the same place. Figure 1.3 depicts some viewpoint changes experienced by a UAV that pose major challenges when assessing image similarity for Place Recognition. Moreover, if onboard computation is a requirement, the limited resources onboard a UAV have to be taken into account, driving research towards lightweight solutions.

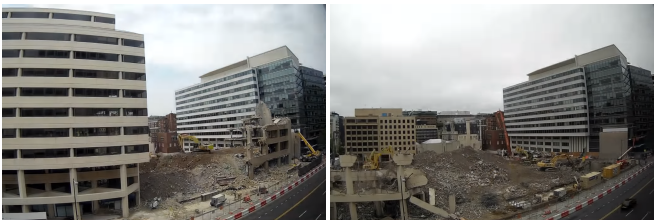
Inspired by the challenges of Place Recognition from aerial imagery and as aerial vehicles are some of the most dynamic and challenging platforms for robotic perception today, the focus of this thesis is on this problem as it promises great impact to the Robotics community. By pushing the state-of-the-art in vision-based Place Recognition, it is possible to enable longer-term autonomous operation of robotic systems in real world scenarios, where major appearance and viewpoint changes are usually present. As such, this thesis addresses vision-based Place Recognition for a small UAV dealing not only with common appearance changes, but also under usual to extreme viewpoint changes among captures. By focusing our research on aerial robots, we believe that the portability of the theoretical and practical outcomes of this thesis to other platforms (e.g. ground robots) with simpler motion and computational constraints should be straightforward.

1.1 Motivation and Objectives

Over the last couple of decades, robots have been evolving quickly, from stationary machines to powerful and sophisticated mobile platforms capable of performing challenging tasks in a wide range of environments. While, in the past, the use of

(a) Seasonal changes

(b) Illumination changes



(c) Physical changes in the environment

(d) Changes due to dynamic objects

Figure 1.1: Example challenges in vision-based Place Recognition illustrating the difficulty in assessing image similarity in the presence of common appearance changes that the same 'place' can experience.

Figure 1.2: Examples of a difficult Place Recognition task, given the existence of self-similar structures in the environment, a problem known as perceptual aliasing. Despite the major similarities between each pair of images in each column, the depicted buildings are in the reality located in different physical locations. Such problems could be addressed by extending the description of a place to incorporate more views of each scene to provide extended context in the comparison.

mobile robots was restricted to controlled environments offering a certain level of structure, rather expected variations, and often a known map, such as in factories and warehouses, now their use is becoming a reality in more general environments. From navigating hospital halls to deliver meals and medication [56] to performing search-and-rescue missions [79], there are many applications, in which long-term deployment of autonomous mobile robots in unknown and often unstructured environments is becoming increasingly important. In fact, to achieve autonomy in such scenarios, mobile robots often need to be able to self-localize, even when faced with dynamic and possibly self-similar environments.

In applications where the map of the environment is unknown and external guidance (e.g. GPS) is not available, the robot needs to employ SLAM to incrementally build a map of its workspace while keeping track of its pose within this map. To this end, several approaches for SLAM have been proposed in the literature employing different sensors, such as single cameras [34, 77, 90] or visual and inertial cues [14, 61, 87], as this is the core ability of spatial understanding necessary for

Figure 1.3: Example viewpoint changes experienced by a UAV. As illustrated here, images from the same place can look very different when captured from different viewpoints, posing major challenges when assessing images similarity.

autonomous robot navigation.

Clearly, the quality of the map and the accuracy of the localization estimates are closely related to the sensors used to perceive the environment. However, sensor readings are usually subject to noise and bias and it is inevitable that in all SLAM approaches, small estimation errors accumulate over time. In effect, this means that the robot's understanding of the world and its trajectory in it diverge from reality, especially over long exploratory missions and as a result, both the map and the robot's location can become inconsistent. Detecting when the robot returns to a previously visited location to close a loop in its trajectory, can join corresponding points of the map, reducing the drift of estimates, thus aiding the creation of more accurate maps and improving localization accuracy. Similarly, Place Recognition can be performed to enable relocalization in case of SLAM failure. In this case, a new SLAM map is often triggered, while attempts to match the robot's current surroundings to all past robot experiences are running in the background. As soon as a match is identified, the two maps can be merged and SLAM can continue on the joint map.

With SLAM approaches reaching maturity, the deployment of multiple robotic platforms to perform a task collaboratively has been attracting increasing attention from the Robotics community. Multi-robot systems offer great promise in a variety of applications, not only to speed up the execution of a task, but also to perform tasks that would not be possible using a single robot (e.g. lift heavy loads collaboratively). With each robot performing SLAM and building its own map independently, a global map of the environment can be obtained by performing inter-map Place Recognition. By detecting when one robot returns to a place already visited by another robot can form the basis of any collaboration among them. In a search-and-rescue scenario, for example, where time is critical and the environment can pose great hazards, the search area to be inspected can be split among a team of heterogeneous robots [39], consisting of UAVs and ground robots, to assist first responders by building a map of the environment. This can be used to aid decision-making, minimizing the exposure of rescuers to unnecessary risks to inspect the area manually. To this end, UAVs, for example, can provide the overview of the scene, while ground robots can be used to explore occluded and

confined spaces. However, in order to bring estimates together to build a joint map, efficient collaboration is necessary, which is only possible if the robots know where they are with respect to each other. This highlights the challenge of viewpoint-tolerant Place Recognition to identify loops across the trajectories of the different robots, experiencing the scene from such different viewpoints.

Apart from multi-robot scenarios, another interesting application of Place Recognition is the ability of performing localization in multi-mission scenarios from single robots. A vacuum cleaner robot, for example, performs the same task in different occasions at different times of the day over the same area. Assuming basic SLAM functionality onboard, mapping of the environment can be performed during a first exploratory mission. However, once a map of the environment is obtained, the robot can use it to plan its path more effectively in subsequent missions. While the scene can be largely expected to remain the same, some parts are due to change (e.g. chairs can be moved), highlighting the need for Place Recognition to identify place matches against the previous acquired map. Place Recognition in such scenarios can become particularly challenging as not only furniture that can be frequently moved to another place, but also illumination conditions are constantly changing throughout the day and night (e.g. the same scene can look very different in direct sunlight and when lit by a lamp in the evening). A projection of similar challenges can be envisioned in street-level localization with Google Maps AR, where the map is augmented with accurate navigation information to assist humans while navigating in an environment. Building on the ideas of [114], GPS information is firstly used to recover Street View images from the person's nearby area. Then, given an image of the person's current location, Place Recognition is performed by matching the query image against all Street View images, enabling more precise pose estimation. Robustness to common scene changes in the Place Recognition pipeline employed, is imperative for smooth operation.

Overall, Place Recognition plays a key role in both (i) the longer-term autonomy of a robot triggering drift correction and relocalization, but also (ii) the co-localization of robots in multi-robot setup, effectively enabling collaboration. Following the realization that vision-based approaches suffer from lack of robustness and generality, the main goal in this thesis is to address Place Recognition under extreme viewpoint changes, while maintaining robustness to common appearance changes and affordable computation.

1.2 Related Work

Feature-based solutions using local feature detectors and descriptors, such as SURF [12] and SIFT [66], have been widely applied in Place Recognition. However, identifying whether a robot is revisiting a place by directly matching a query image against all images in a database is a very inefficient process. To this end, a lot of early research effort has been focusing on image retrieval techniques to efficiently search a database containing all the previous experiences of the robot for loop-closure candidates that are similar in appearance to a given query image. In-

spired by text retrieval techniques, the pioneering work in [102] gave rise to what is widely known as the Bag-of-Words (BoW) approach. This technique relies on quantizing local feature descriptors generated using a set of training images to build a dictionary of visual words and then representing new images as a set of visual words it contains. The Bag-of-Words approach is usually combined with an inverted-file-index linking visual words to frames they appeared in for fast image retrieval [27]. Several well-performing algorithms for Place Recognition using a BoW representation were proposed in the literature. One of the most influential works employing the BoW approach is the open-source FAB-MAP framework for Place Recognition [27], which proposes a probabilistic model of place appearance using a generative model of visual words observations and a sensor model that explains missed observations of visual words. While FAB-MAP discretizes a feature space of SURF [12] descriptors, the work in [36] exploits the use of efficient binary features for Place Recognition and proposes a binary BoW representation. Motivated by the need for lightweight solutions able to run onboard of robotic mobile platforms, the works in [77] and [87] make use of this binary BoW representation to perform loop-closure detection during SLAM. Although several well-performing feature-based approaches have been proposed for Place Recognition, the extraction of unique and repeatable features has been proven to be far from trivial [63], and large appearance changes usually pose major challenges for feature-based methods.

Instead of using traditional feature-based representations, another popular approach to image representation is to consider whole-image descriptors, such as GIST [83] and HOG [29], successfully employed for Place Recognition [65, 75, 78]. These methods are usually considered more robust to appearance changes than feature-based approaches, which usually suffer from the lack of repeatability of local descriptors when changes in appearance occur. However, whole-image approaches often lack invariance to viewpoint changes as whole-image descriptor comparison methods tend to assume that the images compared are captured from a similar viewpoint. SeqSLAM [75], for example, makes use of whole-image descriptors and proposes to address Place Recognition by matching image sequences instead of single observations, achieving impressive recall rates on scenes with dramatic changes in lighting (e.g. day/night). In robotics applications it is common to have access to a sequence of images captured along the robot's trajectory rather than just single-image observations of places. Therefore, matching image sequences instead of individual images can offer useful cues to increase the number of correct matches and filter out incorrect associations. Place Recognition approaches in general benefit from exploiting image sequences, especially during SLAM, where false positive detections can cause large mapping errors.

In the last years, Convolutional Neural Networks (CNNs) have been successfully employed to learn compact image representations suitable for Place Recognition [7] or even to regress a 6-DoF (position and orientation) pose directly from an image [51]. Motivated by their ability to learn generic features that can be deployed for a variety of related, but different visual tasks [84, 98], the work in [21] was the first to exploit the utility of CNNs for Place Recognition. Giving an entire image as input to a pre-trained network, whole-image descriptors were directly extracted from its

activation layers and subsequently employed for image comparison. While several approaches using pre-trained CNNs as whole-image feature extractors for Place Recognition were proposed in the literature [8, 21, 104], these methods are quite sensitive to changes in viewpoint and partial occlusions. More viewpoint-tolerant representations were proposed combining traditional local feature extractors with CNN descriptors to match image patches over large appearance and viewpoint changes [48, 105]. However, these methods require to apply the pre-trained network for every extracted region, resulting in high computational cost. Instead of relying on external feature detectors, the work in [23] directly identifies salient regions from the CNN layer activations and only needs to run the network once for each image, significantly reducing the computational cost. On the other hand, several approaches reported a gain in performance by training a CNN specifically for the Place Recognition problem [7, 22]. While these methods usually perform well under large appearance changes, the impact of extreme viewpoint changes remains largely unexplored, as large datasets addressing both appearance and viewpoint variations are very difficult to obtain. More recently, absolute pose regression approaches have become popular [16, 50, 80, 95]. Given a set of training images and their corresponding poses, these approaches train a CNN to regress the camera pose directly from an image. Requiring only one forward pass through the network, these methods are usually very efficient. However, their performance is still significantly less accurate than feature-based approaches augmented with 3D information, which was shown to achieve higher pose accuracy by explicitly establishing feature correspondences between 2D pixel positions and a 3D map of the scene [15, 95, 97, 110]. While CNN-based features have demonstrated high invariance to changes in the scene, deep learning techniques usually rely on powerful GPUs, rendering them computationally too expensive to run onboard small aircraft. Besides this, they most often rely on very extensive, annotated datasets to cover all possible variations, which are very hard, if not impossible, to obtain.

1.3 Approach

Place Recognition onboard a small UAV is inherently rather different from the traditional Place Recognition problem for a car navigating in the streets of a city as commonly addressed in the literature. While a multitude of powerful and heavy sensors can be carried onboard a car, the limited payload of small UAVs needs to be taken into account when choosing the best sensors for the task at hand. As a result, this thesis tackles the problem of viewpoint-tolerant Place Recognition using vision as the robot's main sensing modality as cameras offer rich information about the environment and are compact and lightweight to be carried onboard small aircraft. As inertial sensors are commonly available onboard UAVs, in this thesis, we make use of both visual cues from a single camera, as the single exteroceptive sensor employed onboard, and feeds from an inertial sensor (i.e. providing acceleration and gyroscopic measurements, and in effect making scale observable), which, together with cameras, comprise the most commonly used sensor setup for

UAV navigation.

The dynamicity and agility of a small UAV (i.e. especially rotorcraft) means that it is very likely to approach the same scene from a wide range of viewpoints, which is by definition fatal for techniques employing whole-image representations, while feature based approaches also struggle greatly. As a result, current feature-based methods attempt to circumvent major changes in appearance by using high quality feature detectors and descriptors, such as SIFT and SURF. These features, however, are typically far too expensive to employ onboard a small UAV, which renders most of these techniques unusable. With deep learning approaches usually demanding powerful GPUs that cannot be carried onboard small aircraft, and considering the advantages and drawbacks of other solutions, the approaches proposed throughout this thesis make use of a binary BoW approach (i.e. operating on features with binary descriptors, such as BRISK [59] and ORB [92]) combined with an inverted-file-index linking visual words to frames they appeared in, for fast image retrieval.

As SLAM is a prerequisite for autonomous robot navigation, the approaches for Place Recognition proposed throughout this thesis are interfaced with a nominal visual-inertial SLAM system employing the keyframes approximation, such as [61, 87]. On the quest, to build more robust Place Recognition solutions, we exploit not only the 2D information inherent in images, but also a local 3D map of the environment that is anyway computed by SLAM that is assumed to be running in the background.

In summary, assuming a nominal, visual-inertial, keyframe-based SLAM system running onboard the robot, in this thesis we make use of low cost binary features that are computationally more suitable for UAV navigation, while exploiting both 2D and 3D information to improve feature-based image matching to enable accurate Place Recognition at large viewpoint changes (i.e. up to 45°). All the works proposed in this thesis make use of a binary BoW approach to retrieve loop-closure candidates from a large database of images containing all the previous locations visited by the robot. The set of candidates is then reduced by first employing appearance checks, followed by geometric ones in order to decide whether the robot is currently at a place experienced already (i.e. present in the database of places) or a new place and also to compute an accurate pose of the robot.

To accomplish these objectives, we break the overall goal of this thesis down to three main parts. First, we develop an approach for viewpoint-tolerant Place Recognition specific to urban scenes. Assuming the existence of major planar structures in the scene, the proposed approach transforms the current image into an orthophoto [69], in an attempt to compensate for the camera rotation and achieve a repeatable image representation. Using this method as a basis, the approach in [107] performs drift correction and relocalization in order to enable robust relative pose estimation between two UAVs. In an attempt to relax the planarity assumption of [69], more generic scenarios are considered in [70] proposing a new Place Recognition framework that puts the focus on robust geometric checks to avoid false positive detections that often occur with appearance-only checks. This framework is then extended in [71] to tackle more extreme viewpoints by

improving feature-based image matching using a depth completion approach.

In order to evaluate the proposed approaches, several datasets, containing visual and inertial information, as well as ground-truth, were created throughout this thesis. While datasets containing visual and inertial information, such as KITTI [41], exist in the literature, most of the available sequences exhibit mainly forward camera motion with a front-looking camera, rendering data labelling for ground-truth very difficult. For this reason, real outdoor sequences were recorded especially for Place Recognition applications using flying and hand-held setups with a side-looking camera, permitting clear decisions on ground-truth labelling. These sequences exhibit moderate appearance changes and large viewpoint variations. To isolate the problem of viewpoint changes in Place Recognition, while keeping full control of the test conditions, photo-realistic synthetic datasets, exhibiting extreme changes in viewpoint, were also produced. The synthetic datasets were created using 3D models obtained by photogrammetric reconstruction and a UAV physical simulator. These benchmarking datasets [70, 71] capture the same area repeatedly, at different times of the day and the year, and with different platforms, experiencing the scene from different viewpoints. Some of these datasets have already been used in [113], presenting an evaluation of visual Place Recognition techniques and advocating the particular difficulties when employed for aerial navigation, resonating the arguments presented in this thesis.

Chapter 2

Contribution

This chapter presents the scientific contributions of this thesis. The overall context and the main contributions of each of the main publications are summarized, and the interrelation among them is discussed. Other related works, made in collaboration with others throughout the course of this thesis, are also listed. The list of students supervised by the author is provided at the end of this chapter.

2.1 Core Publications

Paper I

Fabiola Maffra, Lucas Teixeira, Zetao Chen, Margarita Chli, "Loop-Closure Detection in Urban Scenes for Autonomous Robot Navigation". In *International Conference on 3D Vision (3DV)*, 2017.

Context

Place Recognition is commonly addressed as an image retrieval problem, and the success of the BoW approach in searching for similar images in a database had led to its wide use. This technique relies on building a dictionary of visual words by clustering locally invariant feature descriptors appearing in a collection of model images and then representing each new image as the set of visual words it contains. A BoW approach combined with an inverted-file-index is usually applied to efficiently search for loop-closure candidates in a database of images containing all the previous experiences of the robot. While standard strategies store visual information captured by a perspective camera, some approaches rely on the creation of orthophotos to enable much better condition when evaluating images similarity. By detecting vanishing points and generating gravity-aligned orthophotos, the approach in [10] corrects for the rotation of the camera when capturing a place,

converting Place Recognition to a homothetic problem involving only scale and an offset in a 2D plane. In a similar spirit, the work in [20] demonstrates a gain in performance by combining both unmodified perspective images and their corresponding orthophotos for Place Recognition. However, both works assume the existence of a prior 3D model of the environment, which is unrealistic in some cases. Assuming a high-density of largely planar structures, common in man-made environments, in this paper we propose a novel Place Recognition approach for autonomous robot navigation in urban scenes. Generating a mesh of the robot's local surroundings in real-time, the proposed approach estimates the most salient plane in the current view to generate its corresponding orthophoto. Loosely integrated with a keyframe-based SLAM, no previous knowledge of the environment is required by the proposed system.

Contribution

This paper [69] has two main contributions. First, we propose a new approach to perform loop-closure detection for robot navigation in urban scenarios that does not require any previous knowledge of the environment nor does it impose unrealistic assumptions (e.g. a Manhattan world). This approach was demonstrated to be fast enough to run onboard a small UAV with limited computational capabilities. Second, we create a new approach to generate orthophotos in real-time from sparse features provided by a visual-inertial SLAM algorithm. Orthophotos are commonly computed by extracting line segments in an image in order to detect its vanishing points. Estimating vanishing points is usually an ill-conditioned problem given the very small intersection angles among the lines, while deciding which line segment is associated to which vanishing point present additional source of errors [47]. One advantage of the proposed approach is that it does not rely on the estimation of vanishing points, instead it estimates the most salient plane in the scene, using the local 3D map provided by SLAM, to create its corresponding orthophoto.

Interrelations

The Place Recognition approach presented in this work generates a 3D mesh of a location out of the 3D landmarks provided by a SLAM system. While in this paper a mesh of the local scene is estimated to provide the basis to identify the most prominent plane in the scene (which is later used to compute the corresponding orthophoto), in Paper III, the estimated mesh is used to densify the 3D map provided by the SLAM algorithm to enable feature-based matching between images captured from very different viewpoints.

The Place Recognition approach proposed in this paper was put to the test in [107], in which we propose a collaborative pose estimation between two UAVs to enable aerial manipulation and close-up inspection of structures of interest with low or no texture. In this scenario, a master UAV carrying a known constellation of LEDs (Light-Emitting Diodes) is equipped with the capability to run SLAM onboard and is assumed to fly at a distance from a structure to be inspected

(e.g. a wind turbine). Tracking the LEDs to estimate its relative distance to the master, a slave UAV is employed to perform the close inspection of the structure, as its field of view is then too limited to perform SLAM reliably, onboard. The orthophoto-based Place Recognition is used in this setup to keep the drift of the SLAM estimation bounded, as well as for relocalization in case of a SLAM failure.

Paper II

Fabiola Maffra, Zetao Chen, Margarita Chli, "Viewpoint-tolerant Place Recognition combining 2D and 3D information for UAV navigation". In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

Context

Place Recognition onboard a small UAV is particularly challenging; the dynamicity and agility of a small UAV means that it is very likely to approach the same scene from a wide range of viewpoints, which is very problematic, at best, for feature-based approaches and fatal for global image-representation techniques. Current feature-based BoW approaches try to circumvent major changes in appearance by using high-quality feature detectors and descriptors, such as SIFT [66] and SURF [12]. These features, however, are typically far too expensive to employ onboard a small UAV, which renders most of the existing Place Recognition techniques unusable. Interestingly, features with binary descriptors, such as ORB [14], BRISK [15] and FREAK [16] promise similar matching performance to SIFT or SURF at a dramatically low computation, however, it becomes far more difficult to cluster them into visual words in a BoW approach. The work in [17] was the first in the literature to use binary features for Place Recognition, however, the precision-recall characteristics of this method show that it is still very sensitive to noise. The BoW approach discards all spatial information between visual words by definition, accepting as a match two images having the same words regardless of their geometric constellation in the image space. While in ground robot navigation scenarios, where the scene is expected to always be experienced up-right, this might be enough [4], in UAV navigation, where very different viewpoints are expected, geometric verification of an appearance match is imperative. In this paper, we present a scalable framework to identify loop-closures in a robot's trajectory using low cost, binary features suitable for UAV navigation. While the first priority in Place Recognition is to avoid false positive loop detections, false negatives become of particular interest in viewpoint-challenging cases as they occur far more commonly than in any other scenario, effectively limiting our ability to correct for accumulated drift. In this spirit, here we propose to first use the 3D-3D Horn's geometric verification [43] and if this proves unsuccessful, a follow-up check for 2D-3D geometric consistency using the method of [53] is performed. As the 3D map data is usually sparser than the 2D image data, by expanding the set of correspondences to be considered the proposed approach aims to increase the recall rates while still maintaining perfect precision.

Contribution

In this paper, we propose a novel pipeline for viewpoint-tolerant Place Recognition that makes use of promising leads from existing works, combining them in a way that enables unprecedented robustness to a wide range of common challenges (i.e. tolerance to viewpoint, illumination changes, occlusions, perceptual aliasing, etc). The proposed pipeline was carefully designed to support low-burden computation and to take advantage of any scale and rotation invariance offered by the BRISK features, using combined geometric checks that exploit not only the 2D information inherent in images, but also the 3D information provided by the SLAM system that is assumed to be running in the background. Besides this, new datasets with visual-inertial information and manually-annotated ground-truth were made publicly available with this paper [70]. The new datasets exhibit viewpoint, illumination and situational changes, suitable to test Place Recognition approaches.

Interrelations

The Place Recognition pipeline proposed in this paper served as a basis for the Paper III. In Paper III, the pipeline of Paper II was augmented with a depth completion approach for map densification, in order to improve the geometric checks proposed here, enabling Place Recognition at more dramatic viewpoint changes.

Paper III

Fabiola Maffra, Lucas Teixeira, Zetao Chen, Margarita Chli, "Real-time Wide-baseline Place Recognition using Depth Completion". In *IEEE Robotics and Automation Letters (RA-L)*, 2019.

Context

Extreme changes in appearance and viewpoint can pose a significant challenge for feature-based approaches. As a result, several approaches attempt to combine both 2D information from images and 3D geometry of the scene, either by using cameras and LIDAR sensors [99] or directly extracting 3D information from the images using Structure-From-Motion [94] or SLAM [77]. Usually, these methods rely on image retrieval techniques that make use of 2D information to recover loop-closure candidates to a given query image, before testing for geometric consistency using 3D information. In these cases, the scene is usually represented by a 3D map, in which each 3D point is associated with one or more local descriptors in the image space, and 3D correspondences between two places are usually obtained via descriptors matching in the image space. However, under extreme viewpoint changes, feature-based image matching is strongly affected by affine distortions and occlusions, resulting in a reduced number of correspondences between the query's and the candidate's keypoints. When only keypoints with 3D information are considered in the geometric check, this problem gets even worse, especially when relying on sparse 3D maps obtained by SLAM. As such, in this paper, we propose

a new Place Recognition pipeline, building on the findings and the experience of our previous works, that employs depth-completion on sparse feature maps obtained during SLAM to perform map densification in the hope of augmenting the 3D information available in the scene. While state-of-the-art algorithms in depth completion make use of CNNs to accomplish impressive accuracy, such as [109] and [88], these approaches require powerful GPUs that are not suitable for small aircraft. Although some CPU-only approaches (e.g. [57]) can also achieve reasonable accuracy, these methods usually rely on good quality and not very sparse depth information as input in order to create a dense representation of the scene. In this paper, we rely on an efficient CPU-based approach tailored to SLAM input. Despite lower accuracy, this method can handle arbitrary sparse maps with a certain amount of noise and compute a dense representation of the input image in about 7ms.

Contribution

In this paper, we propose a novel, real-time pipeline for loop-closure detection that employs depth-completion to enable feature-based matching between images captured from very different viewpoints. This approach is shown to be capable of addressing dramatic changes in viewpoint (of up to 45°), demonstrating that wide-baseline image matching is possible using feature-based approaches. In addition, we released new photo-realistic datasets exhibiting dramatic viewpoint changes in simulation that isolate for the first time the problem of viewpoint changes in Place Recognition from other challenges, such as scale variance, dynamicity of the scene, and illumination. Together with our synthetic datasets, we also published real datasets capturing similarly large viewpoints using both aerial and ground footage. In particular, the air-ground sequence is especially interesting to test Place Recognition in scenarios exhibiting common appearance challenges, such as illumination changes and perceptual aliasing.

Interrelations

Driven by the shortcomings of the approach in Paper II in addressing wide-baseline viewpoint changes, this work builds on top of the pipeline of Paper II to enable more robust geometric checks. In particular, here we propose an additional step to perform depth completion for map densification to improve the establishment of 3D correspondences enabling feature-based matching across images captured under very wide baselines. The depth completion step relies in creating a mesh of the robots' surrounding from the 3D landmarks estimated by a SLAM system, using the same approach as Paper I.

2.2 List of Publications

In the context of the author's doctoral studies the following publications were achieved.

- **F. Maffra**, L.Teixeira, Z. Chen, and M. Chli. Loop-closure detection in urban scenes for autonomous robot navigation. In 2017 International Conference on 3D Vision (3DV), pages 356-364. IEEE, 2017

- **F. Maffra**, Z. Chen, and M. Chli. Viewpoint-tolerant Place Recognition combining 2D and 3D information for UAV navigation. In 2018 International Conference on Robotics and Automation (ICRA), pages 2542-2549. IEEE, 2018

- **F. Maffra**, L. Teixeira, Z. Chen, and M. Chli. Real-time Wide-baseline Place Recognition using Depth Completion. IEEE Robotics and Automation Letters (RAL), 4(2):1525-1532, 2019

- Z. Chen, **F. Maffra**, I. Sa, and M. Chli. Only look once, mining distinctive landmarks from convnet for visual Place Recognition. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 9-16. IEEE, 2017

- M. Keller, Z. Chen, **F. Maffra**, P. Schmuck, and M. Chli. Learning deep descriptors with scale-aware triplet networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2762-2770, 2018

- L. Teixeira, **F. Maffra**, M. Moos, and M. Chli. VI-RPE: Visual-Inertial Relative Pose Estimation for Aerial Vehicles. IEEE Robotics and Automation Letters (RAL), 3(4):2770-2777, 2018

2.3 List of Supervised Students

Throughout the author's doctoral studies, she has supervised students at the bachelor and master levels listed below, for their summer, semester, and master projects. For projects that resulted in a publication, the relevant citation is given.

Master Theses

Master students, 6 months, full-time

- Marlin Strub (Fall 2017): "Exploring Continuous Representation of the World for Place Recognition"

- Luca Bartolomei (Fall 2018): "3D Radiation Mapping using a small UAV"

- Michel Keller (Spring 2017): "On the Importance of Scale in Feature Descriptor Learning" [49]

Semester Theses

Master students, 3-4 months, part-time

- Marlin Strub (Spring 2016): "Towards Efficient Place Recognition for UAV Navigation"
- Luca Bartolomei (Spring 2017): "Calibration and set-up of two independent cameras mounted on servos for UAVs relative pose estimation"
- Josua Bögli (Spring 2017): "Content-aware Geometric Check Towards Robust Place Recognition"
- Michael Reto (Spring 2017): "Investigating Feature Invariance for Long-Term Place Recognition"
- Anna Dai (Spring 2018): "Deep Learning-based Semantic Segmentation for 3D mapping"

Summer Project

Bachelor student, 3 months, full-time

- Sining Qin (Summer 2018): "State-of-the-art Object Detection and Segmentation on NVIDIA Jetson TX2"

Chapter 3

Conclusion and Outlook

This thesis addresses the problem of appearance-based Place Recognition by primarily investigating the task of recognizing a place from very different viewpoints. With the aim of developing general and practical systems for autonomous UAV navigation, the proposed methods use efficient binary features that are suitable for small aircraft restricted to small payload and limited computational capabilities. A key contribution is that all methods proposed in this thesis make the most of the SLAM estimation processes that are typically running onboard a robot attempting to navigate with autonomy of some degree. This work enables loop-closure detection for drift correction in the SLAM estimation, relocalization in cases of SLAM failures and/or map merging in multi-robot scenarios. Thus, these methods exploit not only the 2D appearance information inherent in images, but also the noisy estimates of the local 3D geometry provided by SLAM.

At first, a loop-closure detection framework for robot navigation in urban scenarios is proposed in Paper I [69], and orthophotos are generated out of sparse features provided by SLAM to eliminate the effect of the rotation of the camera from the Place Recognition problem, creating much better conditions to assess image similarity. Orthophotos have shown to achieve consistently better performance than their corresponding counterparts based on perspective images, in scenarios where a major plane could be detected. This orthophoto-based Place Recognition approach is used to perform drift correction and relocalization in a pipeline developed for collaborative pose estimation between two UAVs [107] for inspection of a structure of interested. In this pipeline, global pose estimation is performed from one of the UAVs, with errors in the order of 0.2m in the estimated distance to the structure of interest (building facade), 0.5m in the tangent plane to the facade and nearly zero in yaw. As these errors reveal, this orthophoto-based method for Place Recognition can provide a reliable means to estimate a relative pose between the UAVs, while keeping the drift in SLAM bounded and performing relocalization when needed.

Aiming for robot navigation in more general scenarios, in Paper II [70], we relax the planarity assumption of the scene and propose a new pipeline, which makes use of a database of perspective images. This was carefully designed to support low-burden computation and to take advantage of any scale and rotation invariance offered by the BRISK features. The proposed pipeline makes use of a binary BoW approach followed by candidate filtering and combined geometric checks to enable robust loop-closure detection, achieving real-time performance and higher recall, at perfect precision, than the state of the art in challenging, newly obtained datasets. Finally, in Paper III [71] this work is further extended with a new scene depth-completion approach to improve the establishment of 3D correspondences during geometric checks. This approach was key to enable feature-based matching across images captured from very wide baselines. Evaluation on synthetic and real datasets with both hand-held and aerial footage, showed significant improvement in precision-recall rates in comparison to the state of the art, while keeping on-board computation affordable for autonomous UAV navigation. In particular, the method proposed in Paper III outperforms the approach proposed in Paper I even for datasets that are commonly described as mostly planar when visualized from a frontal view (e.g. L'Agout 0-45°). The main reason is that 3D structures that are not noticeable from a specific viewpoint (e.g. at 0°), such as roofs, can become largely visible at extreme viewpoint changes (i.e. at 45°), causing major issues when assessing image similarity. The results obtained throughout this thesis, especially in Paper III, demonstrate that feature-based techniques still have a lot to offer in Place Recognition at extreme viewpoint changes. The research in this thesis, gave rise to several real [70, 71] and photo-realistic simulated [71] datasets for Place Recognition with visual and inertial information, as well as ground-truth whenever possible, which we made publicly available. In particular, our synthetic datasets are, to the best of our knowledge, the first to isolate the challenge of viewpoint changes for Place Recognition, addressing a crucial gap in the literature.

The rest of this chapter discusses some relevant aspects of the algorithms implemented throughout this thesis and sketch possible ideas for extensions and improvements to the proposed methods, for future research in Place Recognition towards robust autonomous robot navigation.

Orthophotos

In Paper I [69], the image database used for Place Recognition consists of orthophotos generated from perspective images captured using a traditional perspective camera. After image retrieval, the set of loop-closure candidates proceeds to geometric verification. To this end, in Paper I we followed a procedure similar to the one proposed in [10], in which the authors exploit the fact that they are solving a homothetic problem and propose to replace a traditional RANSAC-based geometric check by an efficient 1D voting scheme. In this case, scale as well as a horizontal and vertical displacement, are estimated separately. In our approach, metric scale is obtained during the creation of each orthophoto (thanks to the visual-inertial SLAM estimation), allowing us to rescale a query-candidate image

pair to a common scale and then apply the 1D voting scheme to estimate only the horizontal and vertical offsets. However, an adaptation to vote for the scale as in [10], could be very beneficial to the system, as new features need to be computed when resizing an image during the geometric verification step, which is usually a very expensive procedure for real-time applications. As such, the scale computed during the creation of the orthophoto could be used to confirm the correctness of the scale estimated by the 1D voting scheme, increasing the robustness of the method.

Image Retrieval

When the robot captures a new image from its current location, a BoW approach is employed in order to search for loop-closure candidates in an image database containing all places previously visited by the robot. The number of images retrieved from the database is pre-defined in all the approaches proposed throughout this thesis, and it is usually set to retrieve the 50 most similar images to a given query. However, the performance of the image retrieval employed in our pipelines decreases under extreme viewpoint changes, and as a result the number of correct candidates proceeding to geometric verification also shrinks, as demonstrated in Paper III [71]. This problem can be diminished by retrieving more images from the database, but in this case, more images will need to be tested for geometric consistency, increasing the time required by the system to determine whether the current image closes a loop with a previously visited location. Although the BoW is considered state-of-the-art in feature-based Place Recognition, quite a few techniques to boost its performance have already been proposed in the literature [24, 44, 45, 86]. Among them is burstiness weighting [44] to suppress the saturation of the BoW vectors in representing images by repetitive patterns and hamming embedding [5, 45, 93] to a more precise representation of the descriptors. Any of these techniques could be attempted in order to improve the proposed approaches [69–71]. Moreover, besides to the BoW approach, other techniques for image retrieval, such as Locality Sensitive Hashing [58], exist in the literature. Analysing the behaviour of such existing image retrieval techniques under extreme viewpoint changes would be a promising topic of research.

Perceptual Aliasing

Although providing efficient image retrieval, by design, the BoW approach discards all spatial information when comparing sets of words, in effect, the discriminative power of images is reduced and thus, renders loop-closure decisions more prone to perceptual aliasing. For this reason, in all Place Recognition approaches proposed in this thesis, geometric verification of an appearance match between a query-candidate image pair is performed by testing any appearance match for matches in the relative configuration of features, alleviating the problem of perceptual aliasing. Another approach to better cope with this, as used in Paper II [70] and III [71], is to impose a requirement on matching a sequence of images in appearance space

before allowing a loop-closure candidate to proceed to the geometric verification step. While this approach is usually enough for ground robots, which usually revisit the same location following a very similar trajectory, either in the same direction or not, for UAVs that typically follow very different trajectories when revisiting a place, more robust techniques might be required. A popular way to address this problem is to consider the few last possible loop-closure detections and the robot displacement between them. Then, the sequential check can be done by verifying whether the current detection position less the displacement is equal to the previous detection position. Although this can be done using all previous detections, the higher the influence of the drift in SLAM, the lower should be the number of previous frames taken into account. This idea is usually implemented with more sophisticated methods, such as Markov localization [81] and Monte Carlo localization [67].

Ultra-wide baseline feature-based matching

During the geometric checks in Paper II [70], the camera pose of the query image is obtained by establishing 3D-3D or 3D-2D correspondences between a query-candidate pair. By assuming that the scene is represented by a sparse 3D map obtained from SLAM, and each 3D point is associated with one or more local descriptors in the image space, 3D-3D and 3D-2D correspondences are obtained via descriptor matching in the image space. Under extreme viewpoint changes, feature-based image matching is strongly affected by affine distortions and occlusions, resulting in a reduced number of correspondences between the query's and the candidate's keypoints. This problem becomes even worse due to the fact that only a reduced number of features can be tracked during SLAM in order to keep its real-time performance. Moreover, only keypoints successfully tracked by SLAM have a 3D landmark associated with them. While it is possible to extract all the keypoints needed for Place Recognition during SLAM, only a small number of them arrives at the geometric check carrying a 3D information, making it very difficult to establish enough correspondences for loop-closure detection between the query and the candidate images. In Paper III [71], this was circumvented by using a depth-completion approach to perform local map densification, in which interpolated 3D landmarks were estimated for the 2D keypoints that had no depth-estimates associated, improving the establishment of 3D-3D and 3D-2D correspondences for images captured from very different viewpoints. It must be noted that this approach is restricted to the viewpoint-invariance offered by the local feature descriptor used during geometric check, in this case BRISK [59]. To allow more drastic viewpoint changes, higher quality feature detectors and descriptors, such as SURF [12] and SIFT [66] can be used in the pipeline. However, the bigger accuracy comes at the cost of longer run-times, when compared to BRISK, but real-time can still be obtained if fewer keypoints are used during the geometric verification step or fewer images are selected from the image database.

Place Recognition invariant to viewpoint and appearance changes

It is widely accepted in the research community that feature-based approaches usually perform poorly in the presence of large changes in appearance, such as those caused by weather conditions, seasonal and illumination changes (e.g. day/night). The appearance variations of a particular place addressed in this thesis, as well as the viewpoint changes in capturing the scene in an image (or a sequence of images), are restricted to the invariance provided by the local feature detectors and descriptors used in the pipeline. More dramatic changes in appearance usually rely on the use of whole-image descriptors or in deep learning based techniques. While deep learning approaches have been demonstrating impressive results under extreme appearance changes, most of the methods proposed in the literature rely on holistic image descriptors and usually, struggle greatly at extreme viewpoint changes. In this thesis, we demonstrated that feature-based techniques still have a lot to offer in Place Recognition at extreme viewpoint changes. Combining both deep learning approaches to tackle appearance changes and feature-based approaches to handle viewpoint changes is probably a nice path to follow to drive research forward towards Place Recognition with invariance to both common scene changes and variations in the capture of the scene in images.

Semantic Place Recognition

While humans are capable of recognizing objects and places following a holistic approach, feature-based Place Recognition focusses on a set of small regions of interest in the image when assessing image similarity, disregarding the general context of the scene, often resulting to greater ambiguity. For example, the corner of a window on a building facade can be locally identical to the corner of any other window in the building or even a window of a car. In this case, even state-of-the-art algorithms in feature-based image matching are not able to distinguish between two identical regions belonging to different types of objects, rendering them more prone to wrong associations.

Motivated by the remarkable progress that deep neural networks have been experiencing towards semantic scene understanding [91], several approaches have attempted to incorporate semantic knowledge to improve Place Recognition [37, 38] and visual localization [97, 108]. A common strategy is to use semantic information to improve image matching. While some approaches rely on augmenting traditional feature descriptors with semantic information to improve the matching step [6, 55, 101], others attempt to directly learn a descriptor by encoding semantic knowledge within it [97]. By augmenting the appearance information of a place with semantic labels (e.g. window, car, person), it would be straightforward to adapt the proposed algorithms to use this additional information to improve feature-based matching in a way that wrong associations between descriptors belonging to different classes could be avoided by simply checking their associated semantic label. Image retrieval could also benefit from semantic information. Based on the principle that semantics do not change in the presence of appearance and

viewpoint changes, loop-closure candidates could be selected based on their semantic characteristics. In addition, when combined with creating a map, semantic information can be used to improve SLAM. For example, by masking out both moving structures (e.g. pedestrians and cars) and commonly occurring regions (e.g. sky), while preserving more persistent and discriminative structures, such as buildings, poles and road marks, tracking could be made more robust and more useful features would be retained for long-term Place Recognition [13, 90].

In general, Place Recognition can also benefit from research in object recognition and scene classification. Objects can offer relevant cues about a location, especially in indoor environments, where the function of a place, such as 'kitchen' and 'bedroom', can be directly estimated by the objects contained within it [9, 26, 112]. In addition, scene classification can be used to narrow down the search space for Place Recognition, ensuring scalability and opening the way to long-term deployment of robotic platforms.

The main idea of augmenting Place Recognition with semantic information is based on the fact that semantics are resilient to transient variations of the appearance of a place and the conditions of the scene when capturing it in an image (e.g. viewpoint, illumination). However, deep neural networks specifically trained for the task of Place Recognition are commonly trained on datasets depicting appearance changes, while viewpoint changes are usually not available. The lack of large datasets in which both appearance and viewpoint changes are present would pose the greatest obstacle to the use of semantic approaches in scenes where large viewpoint changes occur, as these networks would probably not generalize well to these conditions. As such, the use of semantic information using deep learning approaches for Place Recognition under extreme viewpoint changes still an open problem.

Paper

Loop-Closure Detection in Urban Scenes for Autonomous Robot Navigation

Fabiola Maffra, Lucas Teixeira, Zetao Chen, Margarita Chli

Abstract

Relocalization is a vital process for autonomous robot navigation, typically running in the background of sequential localization and mapping to detect loops in the robot's trajectory. Such loop-closure detections enable corrections for drift accumulated during the estimation processes and even recovery from complete localization failures. In this work, we present a novel approach loosely integrated with a keyframe-based SLAM system to perform loop-closure detection in urban scenarios for autonomous robot navigation. Generating a mesh of the current robot's surroundings in real-time using monocular and inertial cues, the proposed method estimates the most salient plane in the current view, enabling the creation of the corresponding orthophoto for this plane. Evaluating image similarity on orthophotos forms a much better conditioned problem for relocalization, minimizing effects from viewpoint changes. Employing binary image descriptors and tests on their relative constellation in the image, the proposed approach exhibits robustness also to illumination and situational variations common in real scenes, overall resulting to significant improvement in loop-closure detection performance in urban scenes with respect to the state of the art.

Published in:
International Conference on 3D Vision (3DV), 2017
DOI: 10.1109/3DV.2017.00048

1 Introduction

The emergence of powerful techniques for robotic egomotion estimation and map building that follow the SLAM (Simultaneous Localization And Mapping) paradigm has been drawing research and industrial interest in recent years, as this is the core ability of spatial understanding for autonomous robot navigation. With the aim of developing general and practical systems, the use of external tracking or unreliable positioning systems (e.g. GPS) is typically avoided, albeit restricting the scalability of approaches for robot navigation as drift inevitably accumulates over time during sequential processing (especially during exploratory trajectories). Detecting when a robot returns to a previously visited place has long been known to offer useful cues for diminishing the effects of drift and similarly, detecting when one robot returns to a place already visited by another robot can also form the basis of any collaboration amongst them [96]. In both cases, it is the problem of Place Recognition that needs to be addressed, *aka* Loop-Closure detection. Following such a loop detection, new pose-to-pose and pose-to-features constraints are established in the SLAM graph, subject to non-linear optimization, such that the loop closure is enforced and the effects of the drift correction are propagated back to the rest of the SLAM graph.

Primarily addressed using visual cues, place recognition is a challenging task, due to the large appearance variations that the same physical place in the world can exhibit. Illumination and situational variations in the scene's appearance become an issue even at different times of the same day and are certainly caused by weather or seasonal changes, while viewpoint changes or dynamic objects add on to the challenge of identifying place similarity. While impressive works exist in the literature addressing some of these variations in isolation, it is still very challenging to simultaneously address them all together, which is key in enabling robust robot navigation in real tasks. In this spirit, this paper proposes a new orthophoto-based approach for loop-closure detection in the presence of viewpoint, illumination and situational variations. With the rationale that comparing orthophotos instead of perspective images poses a far better conditioned query for place recognition, the proposed approach achieves high recall, while filtering out ill posed queries effectively, and thus, minimizing the probability of false positives.

In this paper, we specifically study the problem of urban robot navigation with the outlook of employing such a system for automating the navigation of small Unmanned Aerial Vehicles (UAVs), which are restricted to small payload and limited computational capacity. Moreover, exhibiting great agility, they highlight the need for viewpoint tolerant place recognition techniques. In urban scenarios, we assume the presence of a high density of structures that are largely planar and are common in man-made environments. This assumption allows us to utilize a planarity prior on the scene and harvest the robustness it can bring to place recognition in this scenery. As a result, the main contribution of this work are:

- a new approach to generate orthophotos in nearly real-time from sparse features provided by a SLAM algorithm, and

- a novel loop-closure detection framework for robot navigation in urban scenes, which does not require any previous knowledge of the environment nor does it impose unrealistic assumptions (e.g. Manhattan world).

Evaluated on challenging datasets, the proposed approach achieves higher precision and recall with respect to the state of the art, exhibiting unprecedented robustness to viewpoint, illumination and situational changes.

2 Related Work

Place recognition is most often addressed using appearance-based cues and as a result, draws inspiration from Image Retrieval from the Computer Vision literature. Identifying whether a query image is present in the database (i.e. containing all past experiences of the robot in the robot navigation paradigm) can be a very inefficient process, so for this purpose, visual dictionaries have been devised to retrieve matching images with high probability. Inspired by text retrieval techniques, the pioneering work in [102] gave rise to what is widely known as the Bag Of Words (BOW) approach. This technique relies on building a dictionary of visual words by clustering locally invariant feature descriptors, such as SIFT [66], appearing in a set of model images and then representing each image as the set of visual words it contains. The use of this representation, permits the analogous application of many theoretical developments such as TF-IDF (Term Frequency - Inverse Document Frequency) and probabilistic naive Bayes [74] from the fields of text retrieval and classification on images [27, 102]. Such techniques, naturally, apply well to place recognition for mobile robots, and are generally well-established in the field, including extended generative models for location observations [3, 28].

The success of BOW approaches in searching for similar images in a database has led to their wide use, however, it was soon realised that their performance decreases with the size of the vocabulary, not only affecting complexity, but also encouraging misclassification. The FABMAP framework [28], partially alleviating the latter by learning the dependencies between visual words, is a framework that is currently considered one of the highest performing pipelines for loop-closure detection in robot navigation scenarios. Its reliance on computationally expensive image features (i.e. SURF [12]) and intolerance to even small viewpoint changes restricts the applicability of FABMAP to scenarios targeting ground robots with large computational capabilities. As with FABMAP, a common source of error in the vast majority of place recognition systems is that they discard most of the geometric information in the image/scene when comparing feature sets. As a consequence, the discriminative nature of the model is reduced, typically resulting in either perceptual aliasing or reduced recall. Full feature-based comparisons can be computationally expensive, and therefore most of the underlying structure and geometry between features is generally ignored, such as in [27]. Following this realization, a handful of works [36, 77] have investigated ways of incorporating some geometric information into the location models. A common approach is to perform RANSAC [33] to compute a transformation between a query and match

candidate images [77].

Probably the most relevant work to this paper is the work in [36], who employ the binary features ORB [92] in a BOW approach and demonstrate its successful applicability to ground robot navigation. As binary features are computationally drastically more efficient than their floating point counterparts (e.g. SURF), they are most commonly used during SLAM [61, 77]. As a result, re-using them for place recognition promises to eliminate unnecessary computational effort, however, the robustness of place recognition systems based on binary features to common scene variations is limited. Inspired by these limitations, in this work we propose to make use of SLAM's 3D estimation to recover a mesh of the local workspace of the robot, which in turn enables the estimation of an orthophoto of the current view. By forming place recognition queries employing binary features in orthophotos, the problem of assessing image similarity using binary descriptors is shown to become more stable and achieve improved performance.

The underlying assumption of largely planar scenery made in this work has also been used to generate orthomosaics from aerial imaging. Orthorectification, essentially facilitates the alignment of images taken from different viewpoints to form a larger mosaic, and as shown in Baatz et al [10] the overlapping part of two orthophotos of the same place is typically very similar resulting to their straightforward alignment. Testing on imagery of buildings facades, [10] factorize the rotation out of the recognition problem by generating gravity-aligned orthophotos outperforming purely 2D-based methods. In a similar spirit, Chen et al [20] demonstrate a gain in place recognition by combining both unmodified perspective images and their corresponding orthophotos. However, both works assume the existence of a prior 3D environment model, which can be unrealistic in some applications.

Inspired by [10], in this paper, instead of searching for a perfect alignment between images, we aim to verify whether the configuration of features shared by two orthophotos presents a consistent layout. This step is known as a geometrical check in loop-closure algorithms. Moreover, in this work, the orthophoto plane is directly extracted from the 3D landmarks used for the robot visual navigation system without the need of computing lines and extracting vanishing points, as in [10].

3 Methodology

As visual-inertial (VI) SLAM is typical in robot navigation, and UAV navigation in particular, the proposed system is interfaced with a nominal VI SLAM system processing cues from a single camera and an Inertial Measurement Unit (IMU). The pipeline, however, is largely agnostic to the type of vision-based SLAM used, with the only requirements of knowledge of the gravity direction and the metric scale. Generating a mesh in 3D out of the local SLAM landmarks, the predominant plane in the scene is identified and the orthophoto corresponding to the current view (i.e. Q in Fig. 4.1) is generated. Extracting binary features on this orthophoto, the pipeline queries the orthophotos database for an appearance based

map identifying possible loop-closure candidates. These are then subjected to a geometric check seeking candidates with matching relative constellation of features in the orthophoto space. Considering the robot navigation paradigm, in the following, we assume that the robotic platform at hand has a monocular-inertial sensor suite onboard.

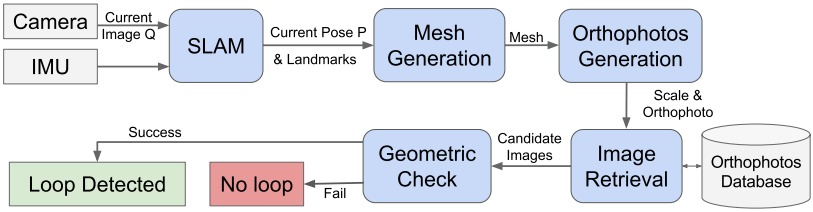


Figure 4.1: The proposed pipeline for place recognition employing mesh-based orthophoto generation with appearance and geometric checks to determine whether the current image Q forms a loop closure with an image in the database containing past robot experiences.

3.1 Real-Time Visual-Inertial Scene Estimation

In this work, we use the open-source keyframe-based VI SLAM algorithm OKVIS [61], which estimates the trajectory of the robot considering a limited window of past poses, and as a result has no loop-closure detection or correction scheme. OKVIS provides in real-time, the current robot pose P and a 3D map comprising of the estimated locations of 3D visual landmarks extracted from the image feed. These are fed into the open-source mesh generation pipeline of [106], which was demonstrated to robustly compute the 3D mesh of the landmarks visible from P in a computationally very lightweight manner, providing a denser scene representation. Filtering out inconsistent landmark measurements from SLAM, the mesh generation algorithm applies a local Laplace filter, implicitly enforcing local smoothness. This is crucial for robust orthophoto generation, as it has a direct effect on the detection of the most salient plane in the scene.

3.2 Orthophotos Generation

An orthophoto of a largely planar scene is the orthogonal projection of this scene onto the most dominant plane of the scene; so in essence, the orthophoto of a perspective image corrects for the camera tilt and the terrain relief. Fig. 4.2 illustrates an example of an orthophoto generated by the algorithm from the Old City dataset introduced in Section 4.1. Although the environment is usually not planar in general, in urban scenes structures are largely planar and aligned to the

gravity direction. In this work, we select the biggest gravity-aligned plane in the image as the orthophoto plane. The rationale behind this is that when viewing the same place at different times, from different viewpoints, most of times, the same orthophoto-plane can be extracted, and as a result, place recognition can be effectively performed.

Figure 4.2: An example of an orthophoto (on the right) generated automatically by the proposed framework for the corresponding original image shown on the left, by estimating a local mesh illustrated in Figure 4.3.

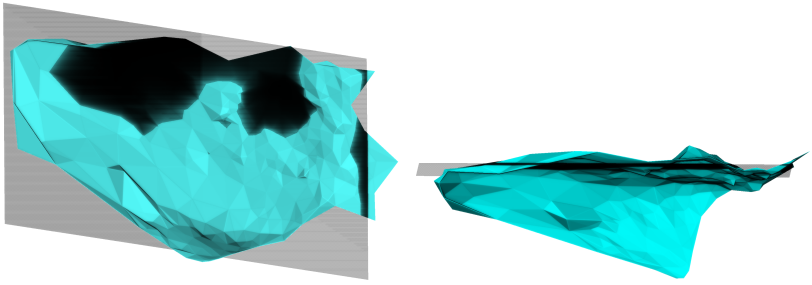


Figure 4.3: The mesh of the scene of Figure 4.3 in cyan and the main plane extracted from it in grey. The side view can be seen on the left and the top view on the right.

The generation of orthophotos first requires the estimation of the most predominant plane in the image, that will serve as the orthophoto plane. This estimation is facilitated by the 3D mesh provided by the VI-SLAM and the Mesh generation module. Aligning the mesh's coordinate frame with gravity (OKVIS already provides a gravity aligned map), we project the 3D mesh to the 2D top view of the scene. The longest line in this view corresponds to the largest vertical plane in the 3D scene. In order to recover this line, we use an iterative Huber M-Estimator to

fit a line to the 2D SLAM points (i.e. the mesh's vertices) considering any point within a pre-specified distance to the estimated line (here $40cm$) as inliers. Upon discovering the longest line in the top view of the scene, we set the middle of it to correspond to the center of the orthophoto-plane. The normal of this plane is selected as the normal of the line that points to the direction of the camera in the gravity aligned SLAM coordinate frame.

In order to project the current perspective image to the estimated orthophoto plane, we first find where the four corners of the frustum of the camera intersects with this plane (i.e. points P1, P2, P3 and P4 in Fig. 4.4). With this information, we form a homography to transform this plane from image coordinates to metric coordinates and use this to project the perspective image onto the orthophoto plane, forming the orthophoto. In order to restrict the size of this orthophoto, we rescale it to the maximum of twice of the original resolution. We impose this restriction because robot cameras have very low resolution, in our case 752x480. So a higher rescaling factor in addition with the orthogonalization of the image creates a very distorted image.

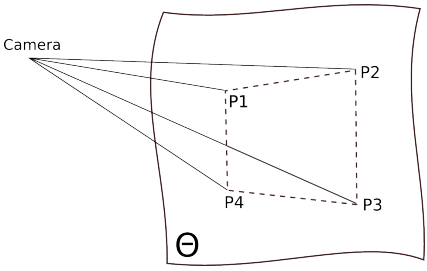


Figure 4.4: The intersection of the camera frustum with the estimated orthophoto-plane Θ to be used in order to form the homography to be applied on the perspective image for the generation of the corresponding orthophoto.

3.3 Image Retrieval

In order to detect revisited places we make use of a hierarchical Bag of Binary Words (BoBW) visual vocabulary, describing an image as a collection of visual words combined with an inverted file index. In this work, the visual database consists of orthophotos generated from perspective images captured using a traditional perspective camera. Each entry in the database comprises an appearance signature of the corresponding image, namely its BoBW descriptor. Following the approach suggested by Galvez and Tardos [36], we adapt for the binary features used in OKVIS, namely the BRISK features [59]. Namely, we build a visual vocabulary by discretizing the 48-byte BRISK descriptors' space. In order to train

this vocabulary, we used about 6000 images comprising both indoor and outdoor environments, different from the ones used for testing. The vocabulary tree built has 10 branches and 6 depth levels, resulting to a vocabulary of one million visual words.

In order to query the orthophoto of the current view Q for appearance matches in the orthophotos database, BRISK features are detected and the BoBW descriptor for Q is formed. The vocabulary tree is used to score the L1-distance of this descriptor against the entries in the orthophotos database using a TF-IDF weighting scheme [28] to suppress commonly occurring words and form the set of matching image candidates.

3.4 Geometric Check

The BOW approach discards all spatial information of visual words by definition, effectively accepting as any match two images having the similar visual features regardless of their relative constellation in the image space. A geometric check based on a RANSAC scheme is usually applied after appearance matching to improve loop closure detection by verifying whether the configuration of features belonging to these two images presents a consistent layout. When matching gravity-compatible orthophotos, Baatz et al [10] reduce the 6 DOF perspective recognition problem to a homothetic problem involving only scale and a translation in a 2D plane. By exploiting the fact that they are solving a homothetic problem, [10] suggests to replace the computationally expensive RANSAC-based geometric check by an efficient 1D voting scheme, where scale as well as a horizontal and a vertical offset are estimated separately.

The proposed approach conducts geometric verification to every query-candidate orthophotos pair that is shortlisted by the image retrieval module. By making use of the metric scale provided for each orthophoto during their creation, we first convert both images to a common scale, which allows us to use the 1D voting scheme for both horizontal and vertical displacement. With both the query and the candidate matching orthophotos in the same scale, we establish BRISK correspondences across features detected in both images.

Following the approach of [10], we estimate the horizontal x and vertical y components of the relative translation between the query Q and the candidate C, independently. Every pair of corresponding points (x_C, y_C) and (x_Q, y_Q) contributes with one vote for the x-displacement $(x(i) = x_C(i) - x_Q(i))$ and one vote for the y-displacement $(y(i) = y_C(i) - y_Q(i))$. The global displacement of the x-coordinate is determined by fitting a probabilistic density function to all the votes computed along the axis x. To this end we use a Kernel density estimation (KDE) supported by a Gaussian kernel, where each offset in x contributes with a Gaussian probability density function with mean centered at $x(i)$ and a standard deviation defined by a translation tolerance in meters. The probability density function is then computed by summing up all these contributions and the global maximum of this distribution is used as the global displacement in this direction. The corresponding points whose coordinate differences are within a certain distance from

the global displacement in x are considered inliers. Since all the coordinates are expressed in meters it is easy to define a distance tolerance to compute the inliers set. The same procedure is then applied to compute the displacement in y. The intersection of the two resulting inlier sets constitutes the final inliers of the geometric check. The number of inliers is then used as a metric to decide whether a candidate should be accepted as a loop closure to match the query image. The different thresholds applied are analysed in section 4 by means of precision-recall curves.

4 Experiments and Results

While there do not exist directly comparable methods for place recognition using orthophotos, as a baseline algorithm, we form a variant of the proposed pipeline adapted to use perspective images as done traditionally in robot navigation scenarios, as the monocular-based ORB-SLAM [77]. This enables fairness of comparisons as we ensure that all tests use the same features and are subject to the same quality of SLAM estimation. As the geometric verification voting scheme is not suitable when using perspective images for the variant pipeline that we refer to as $Persp_{FM}$, we implemented the strategy used in the BoBW approach in [36]. This consists in computing a spatial transformation between the matched images by estimating the fundamental matrix using RANSAC for the variant algorithm. The proposed method from here onwards is referred to as $Ortho_{TR}$ to denote the use of Orthophotos with the voting scheme used to estimate translation.

4.1 Datasets

Existing place recognition datasets normally only contain visual information, however, in order to put our proposed approach to the test, we need visual and inertial sensing information, as well as ground truth. Outdoor visual-inertial datasets, such as KITTI [41] are designed for motion estimation and are not suited for testing place recognition as they exhibit mainly forward camera motion with a front-looking camera, rendering it very difficult to label the images for ground truth in loop closures.

All the datasets used in this paper were recorded using a visual-inertial sensor [82] providing grayscale global-shutter images at 20 Hz synchronized with inertial measurements. For our experiments, we use information from only one of the two cameras of the sensor to conduct monocular-inertial estimation. The datasets were recorded using a hand-held setup with the camera facing perpendicular to the direction of motion (i.e. side-looking). All imagery was labelled for ground-truth loop closures, by first using any priors on GPS information whenever available to suggest potential loops and then manually correcting these suggestions. Below, we describe in detail all the datasets used in this paper.

(a)

(b)

Figure 4.5: Example loop-closuring pairs from the Old City dataset identified with the proposed approach. Each group of four images shows the original perspective views in the top row and the respective orthophotos in the bottom row.

Shopping Street sequences 1 & 2

Two datasets were recorded when walking down a busy shopping street with many pedestrians. Examples are shown in Figures 4.9 and 4.10. Shopping Street 1 was recorded with the sensor held at eye-level height and exhibits loops with small viewpoint changes, perceptual aliasing and changes in the scene appearance. Shopping Street 2 was recorded along the same street a few months later with the sensor mounted at the top of a $4m$-long rod held vertically in order to capture the scenery captured in Shopping Street 1, at least partially, but from different viewpoints. By combining these two sequences, a very challenging place recognition dataset is created, where the scene is not only revisited from very different viewpoints, but due to the large time interval between recordings, strong appearance variations can also be observed with most of the restaurants and shop windows in different configurations; e.g. shutters closed, window displays and even store logos changed. Moreover, parts of Shopping Street 2 exhibit large variance in illumination conditions, making it hard even for humans to detect whether it is the same place visited in the first sequence. These sequences have a total of approximately 1200 meters and 26 mins.

Old City sequences

Two sequences were recorded at the end of the day in an old city area, exhibiting similar characteristics as the Shopping Street datasets, albeit with more challenging viewpoint variations. This dataset comprises two traverses along the same route, each one covering a distance of approximately 230 meters. In total, 10 minutes of data were recorded for this dataset. Example images are shown in Figure 4.5.

4.2 Orthophotos versus Perspective Images

Aiming to verify whether using the orthophotos generated by $Ortho_{TR}$ can perform better than their perspective counterparts in a place recognition, we test for loop closures within Shopping Street 1, which comprises two different traverses one the same day, along the same route. Images from the first traverse are used to populate the database of images, and using images from the second traverse this database is queried for loop closures. Parts of these trajectories does not overlap and in that case loop closures should not be detected.

Each of $Ortho_{TR}$ and $Persp_{FM}$ builds their own, separate database of images for retrieval; $Ortho_{TR}$ builds a database of orthophotos, while $Persp_{FM}$'s database comprises of the perspective images. It is important to note that only the perspective images with more than 30% of inliers have their corresponding orthophotos computed. If an image does not meet this requirement, it is is not considered neither $Ortho_{TR}$ nor $Persp_{FM}$ during this test. For both pipelines the image retrieval step considers the top 10 best images recovered from the database, while the corresponding geometric check is run for every pair of query-candidate images. Their performances are illustrated by the dashed lines plotted in Fig. 4.6,

demonstrating that $Ortho_{TR}$ performs consistently better than $Persp_{FM}$ in this scenario. This attests to our earlier claim that using orthophotos, place recognition can be more robust and accurate, compared to employing perspective images for the same tests. The performance of the two systems in a more general scenario is recorded, in which all the images in the Shopping Street 1 sequence images are considered, shown in the solid lines in Fig. 4.6. As expected, in this case the recall for $Ortho_{TR}$ decreases, but still performs systematically better than $Persp_{FM}$.

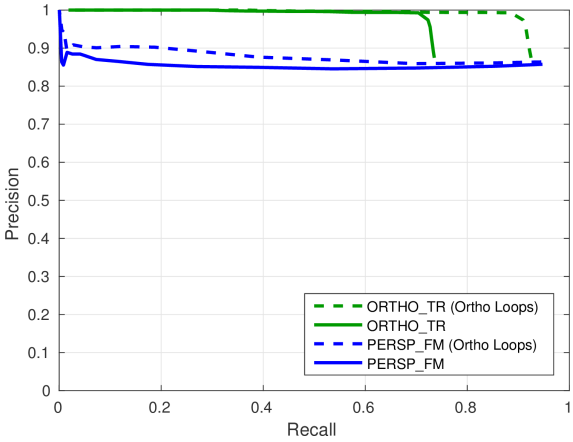


Figure 4.6: Precision-recall curves on Shopping Street 1 comparing the performance of using orthophoto (green) and perspective (blue) images for place recognition. The dashed lines indicate the respective performances when considering only the images with enough inliers to generate orthophotos, while the solid lines illustrate performances when all the images of the sequence are considered. The reference and test traverses are collected at the same day along the same route, with small viewpoint changes.

We also compute precision-recall curves for Shopping Street 1 + 2 and Old City using both $Ortho_{TR}$ and $Persp_{FM}$ algorithms as can be seen in Fig. 4.7 and Fig. 4.8, respectively. As previously done, the first sequence of each dataset is inserted into the database of images, while the second one is used as image queries. For Shopping Street 1 + 2, we insert the first loop of Shopping street 1 into the database and use as query images the Shopping Street 2 sequence. $Ortho_{TR}$ is evidently able to maintain higher precision than $Persp_{FM}$, essentially attesting to better consistency of performance, rendering it more trustworthy in closing loops during autonomous robot navigation. Example loop closures for Shopping Street 1 + 2 and Old City are shown in Fig. 4.9 and Fig. 4.5, respectively.

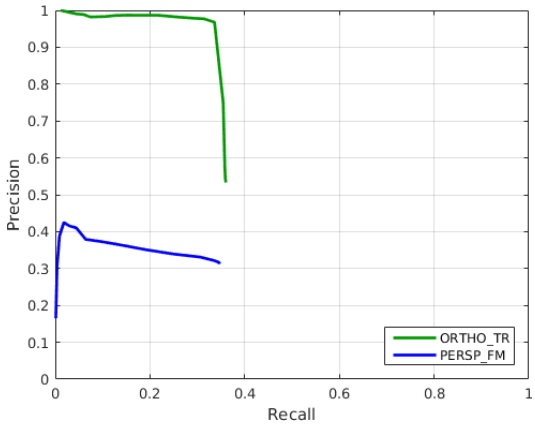


Figure 4.7: Precision-recall curves on the combined Shopping Street 1 + 2 dataset, comparing performances when using orthophoto (green) and the perspective (blue) images for place recognition. While the reference traverse in the same as the one used in Fig. 4.6, the test traverse is collected at the same route after four months, thus exhibiting much stronger condition variations.

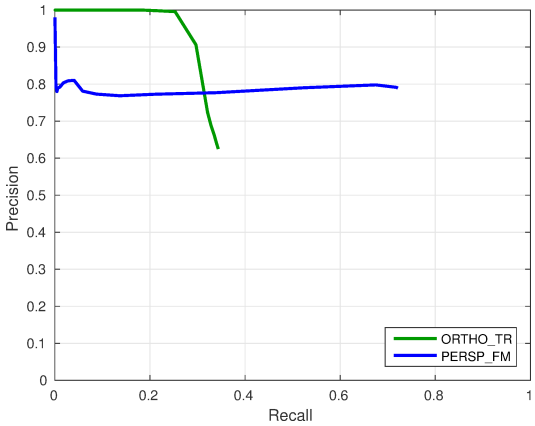


Figure 4.8: Precision-recall curves on the Old City dataset comparing performances of using orthophoto (green) or perspective (blue) images for place recognition. It is clear that orthophoto-based place recognition achieves much higher precision at the same recall rate.

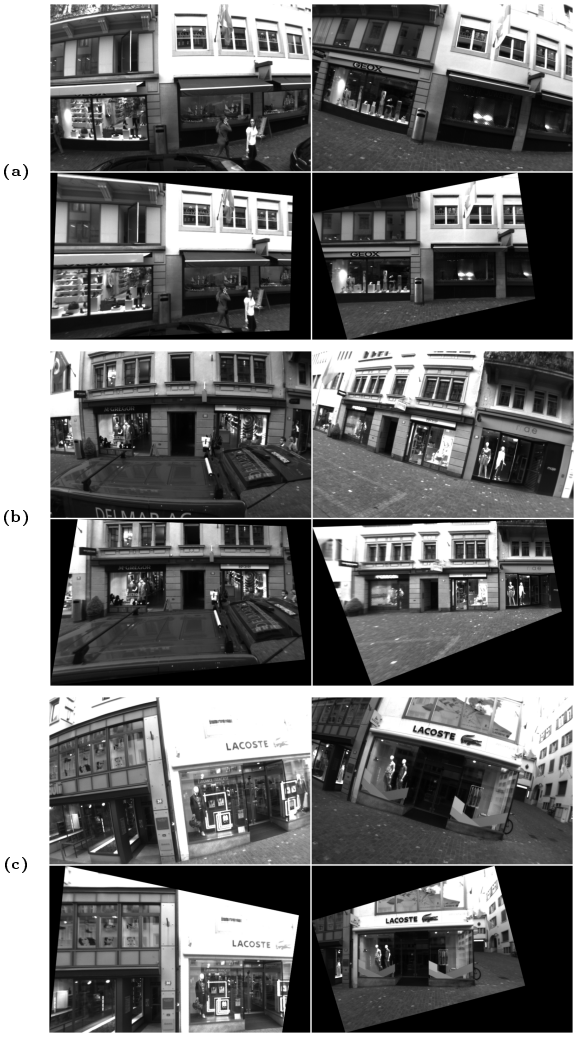
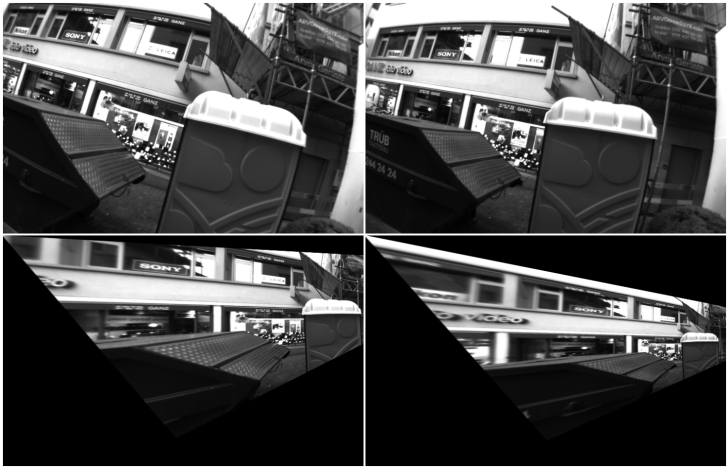


Figure 4.9: Example loop-closures from the combined Shopping Street 1 + 2 dataset shown in each row, as identified by the proposed approach. In each group of images, the top row illustrates to the original perspective images, while the respective orthophotos are in the bottom row. In (a) and (b) we can observe large viewpoint and situational changes, with pedestrians and a major occlusion by a car, while (b) and (c) show difficult lighting conditions.



(a)

(b)

Figure 4.10: Example loop-closures from the Shopping Street 1 dataset tested with the proposed approach and their respective orthophotos. The images show that it is possible to compute the orthophotos if more than one plane is present in the scene, even when the environment consists of different depths.

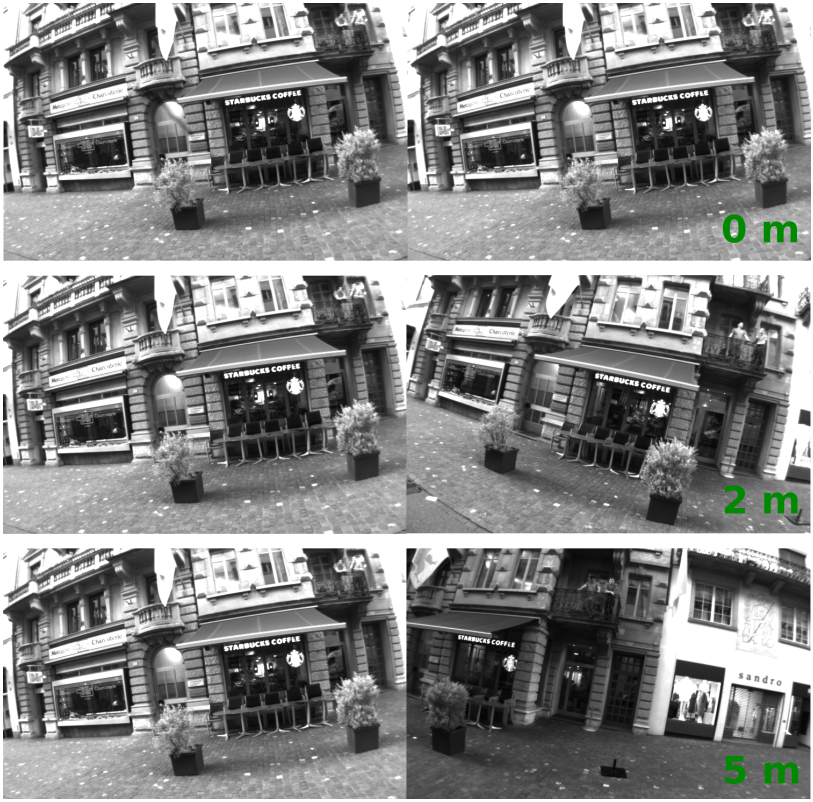


Figure 4.11: Example illustration two consecutive images for different spacing strategies. From top to bottom the figure depicts no gap, 2 meters and 5 meters between the images.

4.3 Viewpoint Changes and System Scalability

In order to test different extents of viewpoint variations using both perspective images and orthophotos, we implemented three different spacing policies between consecutive images when populating our database. In the first setting, all the keyframes used by OKVIS are inserted into our database of images, resulting in a big overlap between consecutive images in our sequence. In the second setting, an

image is only inserted if it is at least 2 meters away from the previously inserted image. In the last setting, a distance of 5 meters between consecutive images is considered, leading to a much more challenging place recognition scenario as illustrated in Fig. 4.11.

Both pipelines, $Ortho_{TR}$ and $Persp_{FM}$, were tested using these three different policies. Using the same strategy as before, in a first step all relevant images from the first traverse are used to populate the corresponding image database and then all the images in the second traverse are used to query that database. Fig. 4.12 shows the respective precision-recall curves for each case. $Persp_{FM}$ presents a sharp drop in precision-recall rates when the gap between images increases, while $Ortho_{TR}$ is still able to maintain much better recall for perfect precision. This illustrates that the orthophotos generated automatically are more robust against viewpoint variations than the perspective images, as expected. Based on these findings, it would be possible to augment the pipeline to select non-overlapping images in order to build a less confusing database of images (i.e. places), while making the place recognition problem more scalable.

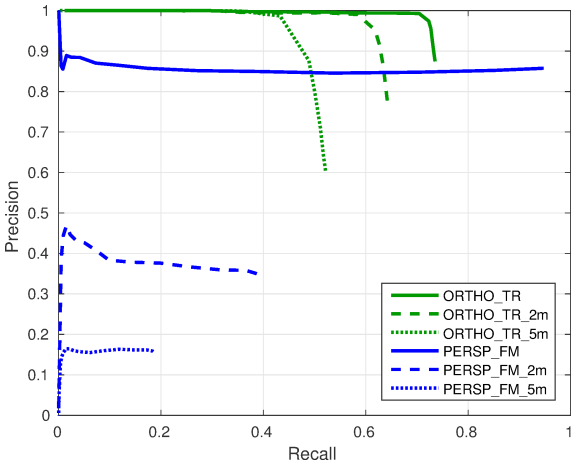


Figure 4.12: Precision-recall curves on the Shopping Street 1 dataset, comparing performances of using orthophoto (green) and perspective (blue) images for place recognition with different sampling spacings between consecutive images in the reference traverse (solid lines: original spacing, dashed: $2m$ spacing, dotted: $5m$ spacing across the camera's trajectory).

5 Timings

Table 4.1 shows timings of each individual component in the proposed pipeline averaged over all the runs in the experiments. As evident, the proposed approach is about twice real-time, with the bottleneck on the feature detection and matching. As loop-closure detection and correction usually runs on a background thread in most SLAM systems, real-time is not a requirement. It is worth noting that an adaptation to vote for the scale as in [10], the processing time can be reduced, as it eliminates the need to rescale the image and recompute features in it.

Step	Average time per image
Image Rescaling	5 ms
Features Detection	40 ms
Features Matching	21 ms
KDE	0.2 ms
Total	66.2 ms

Table 4.1: Average timings for the online component of $Ortho_{TR}$.

6 Conclusion

This paper presents an efficient and precise algorithm to tackle the loop-closure detection problem based on orthophotos automatically generated online. Evaluation against a baseline approach employing perspective images and combined appearance and geometric checks, the proposed approach achieves consistently better precision-recall characteristics in challenging datasets exhibiting viewpoint, illumination and situation changes simultaneously. Tailored for robot navigation in urban scenarios and aiming for low computational complexity, this approach makes the most of a SLAM system that is typically already running in the background in such scenarios.

Further directions include interfacing this pipeline with a global mapping algorithm to enable loop-closure correction within SLAM and harvest the benefits of a robust loop-closure detection pipeline in robot navigation.

Paper II

Viewpoint-tolerant Place Recognition combining 2D and 3D information for UAV navigation

Fabiola Maffra, Zetao Chen, Margarita Chli

Abstract

The booming interest in Unmanned Aerial Vehicles (UAVs) is fed by their potentially great impact, however progress is hindered by their limited perception capabilities. While vision-based odometry was shown to run successfully onboard UAVs, loop-closure detection to correct for drift or to recover from tracking failures, has so far, proven particularly challenging for UAVs. At the heart of this is the problem of viewpoint-tolerant place recognition; in stark difference to ground robots, UAVs can revisit a scene from very different viewpoints. As a result, existing approaches struggle greatly as the task at hand violates underlying assumptions in assessing scene similarity. In this paper, we propose a place recognition framework, which exploits both efficient binary features and noisy estimates of the local 3D geometry, which are anyway computed for visual-inertial odometry onboard the UAV. Attaching both an appearance and a geometry signature to each 'location', the proposed approach demonstrates unprecedented recall for perfect precision as well as high quality loop-closing transformations on both flying and hand-held datasets exhibiting large viewpoint and appearance changes as well as perceptual aliasing. Upon acceptance, these datasets will be made publicly available.

Published in:
IEEE International Conference on Robotics and Automation (ICRA), 2018
DOI: 10.1109/ICRA.2018.8460786

1 Introduction

With small Unmanned Aerial Vehicles (UAVs) sparking great interest for a plethora of potential applications ranging from digitization of archaeological sites to search-and-rescue, there has been an increasing body of research dedicated in automating their navigation. As Spatial understanding forms the basis of autonomous robot navigation, a variety of techniques for robotic egomotion estimation and map building that perform SLAM (Simultaneous Localization and Mapping) have been proposed in the literature. In addition, addressing place recognition by determining whether a robot returns to a previously visited place is a key competence to enable the creation of accurate maps, relocalization and even collaboration between different robots performing SLAM, essentially opening up the way towards long-term operation of robotic platforms in real world scenarios. However, the agility and portability of small aircraft comes at the cost of small payload and as a result, limited computational capabilities. Current solutions involve restricting the onboard memory of past experiences by limiting the size of the SLAM map (e.g. as in [111]). As small estimation errors are usually accumulated over time, restricting the estimation process to a limited window accentuates the problem of drift even further, highlighting the need for suitable place recognition techniques. Moreover, the agility and dynamicity of UAV manoeuvres pose particular challenges in place recognition, as the same place needs to be estimated from very different viewpoints.

Inspired by the challenges of place recognition from aerial imagery, in this paper, we present a scalable framework to identify loop-closures in a robot's trajectory using low cost, binary features suitable for UAV navigation. As UAV navigation is one of the hardest scenarios for place recognition, the portability of the proposed method to other platforms (e.g. a ground robot) with simpler motion and computational constraints should be straightforward. Moreover, while the vast majority of works in this domain restrict their operation to a decision whether there has been a loop closure or not, here, we go a step further to accurately estimate the transformation between the matching robot poses, which can be directly used in a subsequent optimization step. Designed to be interfaced with a keyframe- and vision-based odometry system, the proposed pipeline is shown to outperform the state of the art on both indoor, aerial sequences evaluated on ground-truth data from a highly accurate tracking system (i.e. Vicon), as well as outdoor hand-held and aerial urban sequences against GPS position information. To encourage further research and benchmarking in viewpoint-tolerant place recognition our challenging datasets are being made publicly available. Fig. 5.1 illustrates an example of a successful loop detected by the proposed approach designed to cope with large viewpoint changes and perceptual aliasing. The main contributions of this work are:

- a new, carefully designed place recognition pipeline especially developed for robot navigation, which avoids false positive loop closures at all costs, exhibiting robustness to viewpoint changes, and
- new datasets with visual-inertial information and manually-annotated ground-

truth capturing viewpoint, illumination and situational changes, suitable to test place recognition approaches.

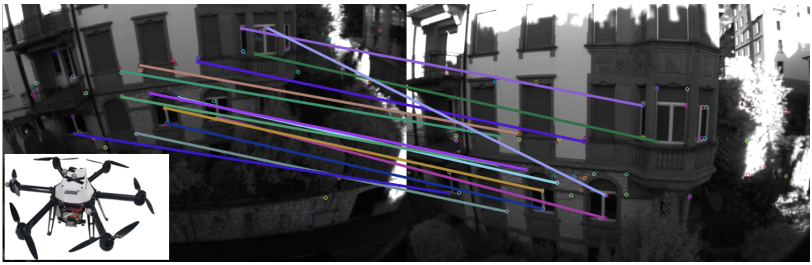


Figure 5.1: A loop in the UAV dataset correctly detected by the proposed approach, despite the large viewpoint change and the mismatches caused by repetitive scene structure. This challenging dataset was captured with the UAV in the inset and it is being made publicly available.

2 Related Work

Place recognition, also referred to as loop closure detection, is usually addressed using appearance-based cues. Typically, two main tasks must be accomplished to address place recognition: (a) query a database of images to find possible similar locations and then (b) determine which, if any, of these images represents the same place as the query. Identifying whether a robot is revisiting a place by directly matching a query image to all images into a database containing its previously visited locations is very inefficient. For this reason, either a Bag of Words approach (BoW) approach [102] with an inverted file index or a descriptor voting scheme [40] are usually applied in the first task followed by a geometric consistency check in the latter one. The widely known BoW approach relies on discretizing the space of feature descriptors generated using a set of training images to build a dictionary of visual words and then representing new images as a set of visual words it contains. Several well-performing algorithms using a BoW representation were proposed in the literature, with FABMAP [28] considered to be one of the most successful pipeline for place recognition.

A less popular approach is to consider global image representations, instead of traditional feature-based representations. In PTAM [52] for example, a smaller and blurred version of the original keyframe image was used as a descriptor of a place, which implies that for relocalization (i.e. loop-closure detection) an exhaustive search across the entire database of images is necessary to identify a potential correlation match. SeqSLAM [75] has demonstrated very impressive recall rates

on scenes with dramatic changes in lighting (day/night), however, the method still lacks invariance (e.g. in viewpoint) and relies on using long sequences of images to tackle perceptual aliasing of the query location. Moreover, the scalability of such methods is more limiting than with feature-based BOW approaches, where indexing and searching for matches can be done more efficiently.

More recently, Convolutional Neural Networks (CNNs) have been successfully applied to solve the place recognition problem under extreme changes in appearance (e.g. time of the day, weather, seasons as well as human activity and occlusions). While [7] and [89] train a CNN to learn a compact image representation suitable to place recognition, another common strategy cast the place recognition problem as a classification task [19, 42]. While impressive results have been obtained by using deep learning techniques, this approach still very computational expensive. While efforts to reduce the computational complexity exist [23], place recognition using deep learning remains unsuitable for real-time estimation onboard a small UAV with small payload and limited computational capabilities.

Place recognition onboard a small UAV is a particularly challenging problem; the dynamicity and agility of a small UAV means that it is very likely to approach the same scene from a wide range of viewpoints, which is by definition fatal for global image-representation techniques, while feature based BoW approaches also struggle greatly. This is inherently a very different problem from the traditional place recognition on a car in the streets of a city as addressed in [28] and [75]. The need for unique and repeatably recognizable features is all the more important in order to allow viewpoint-invariant recognition. As a result, current methods choose to work with the highest quality of feature detectors and descriptors, such as SIFT [66] and SURF [12]. These features, however, are typically far too expensive to employ onboard a small UAV, which renders most of the existing place recognition techniques unusable.

Interestingly, features with binary descriptors, such as ORB [92], BRISK [59] and FREAK [1] promise similar matching performance to SIFT or SURF at a dramatically low computation, however, it becomes far more difficult to cluster them into visual words in a BoW approach. The work in [36] was the first in the literature to use binary features for place recognition, however, the precision-recall characteristics of this method still very sensitive to noise.

Another interesting line of research that has recently appeared makes use of learning techniques to overcome the large viewpoint differences from ground to aerial images. While these wide baselines are not usually addressed in place recognition systems, novel algorithms to air-ground matching have been proposed in complementary areas [2, 64, 72]. Despite the impressive aforementioned algorithms, we still lack a robust solution that overcomes the large viewpoint differences between images captured from a UAV, while keeping onboard computation affordable for a long-term place recognition system.

Finally, while most of the place recognition systems ignores the underlying structure and geometry between features when comparing features sets, a handful of works have investigated how to incorporate some geometric information in their location models, such as in [85], where locations are represented by both visual

landmarks and a distribution of the distances between them in 3D coming from range-finders or stereo cameras. Instead of relying on additional sensors to obtain 3D landmark positions, in [103] landmarks are tracked between successive images using a single camera, recording the binary covisibility between landmarks in a graph-based map of the world. In the general case, the graph matching problem in undirected graphs is an NP-hard problem. As a result, there are still open questions on how such techniques can be efficiently and sufficiently approximated to provide the robustness necessary for place recognition for a UAV.

3 Methodology

The proposed framework is designed to be employed within the loop of robot navigation, so we assume that a vision-based SLAM/odometry system using a keyframes paradigm runs on a separate thread. A hierarchical Bag of Binary Words (BoBW) visual vocabulary is formed in binary descriptors' space with an inverted file index to efficiently query at runtime, the database of keyframes captured during the robot's trajectory for loop-closures. The workflow comprises of two consecutive checks as illustrated in Fig. 5.2; an Appearance Check making use of the keyframe-covisibility information captured by SLAM refines and removes erroneous loop-closure candidates suggested by the BoBW descriptors, before a Geometric Check tests for matches in the configuration of features (in 3D and in 2D) in the candidate keyframe matches that survive the Appearance Check. A successful Geometric Check denotes loop closure detection; in this case the system does not only provide the matched keyframes, but also the best rigid transformation found between them.

3.1 Visual-Inertial Keyframe-based SLAM

With the ultimate goal of place recognition for a UAV, we assume that a nominal monocular-inertial SLAM system is running in the background, as this is a widely accepted sensor setup for small aircraft with limited payload [111], permitting absolute scale estimation. The proposed system, however, is agnostic to the keyframe- and vision-based SLAM system to be used (i.e. no inertial sensing is necessary). In this work and throughout our experiments, we employ the open-sourced OKVIS visual-inertial SLAM/odometry framework of [60, 61], while we have developed a Covisibility Graph data-structure similarly to [77], where any two keyframes (nodes) share an edge if they share enough 3D landmarks. This approach is more adaptive than choosing a fixed number of consecutive images to represent a location. As SLAM keyframes can provide both the detected features (in this case BRISK [59]) in image space and the local 3D map, a new entry is created in the Image Database for every new SLAM keyframe. Each such entry comprises of an appearance signature of the corresponding keyframe, namely its BoBW descriptor and a geometry signature that is the local, sparse 3D map of keypoints that this keyframe has been associated with.

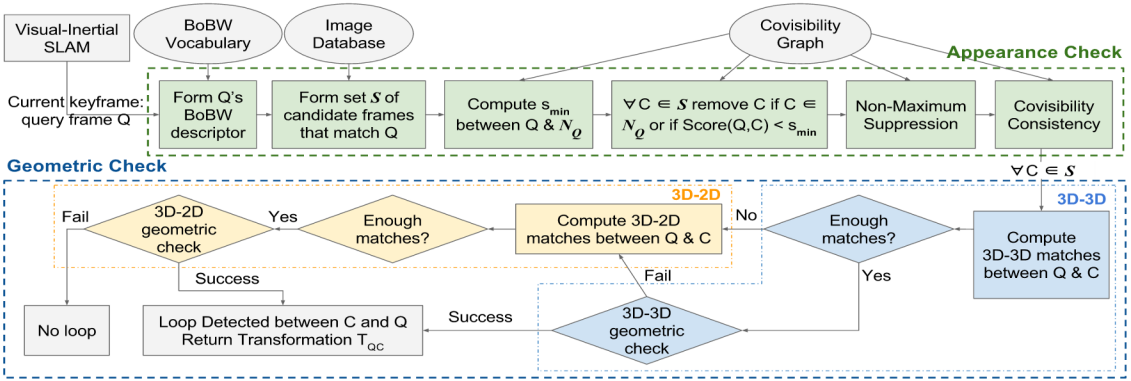


Figure 5.2: The proposed Place Recognition Pipeline, first queries the current keyframe Q for an appearance match consulting the BoBW vocabulary, the database of visited keyframes (i.e. known 'locations') and the keyframes' covisibility graph maintained by the SLAM system. If Q appears similar enough to a database image, the candidate matching keyframes are checked for similarity in their geometry of features both in 3D (local map space) and in 2D (image space).

3.2 Building the BoBW Visual Vocabulary

Opting for a hierarchical visual vocabulary [36], the proposed method describes an image as a collection of words combined with an inverted file index allowing efficient retrieval in a large database of images. While features, such as SURF [12] and SIFT [66] are well-established and known to provide stable detection and high quality description of corresponding image areas, their prohibitively high computational cost has been driving research in robotic visual perception towards the computationally cheaper binary-descriptor alternatives, such as BRISK, ORB [92] and BRIEF [18]. Clustering binary instead of floating point descriptors, however, to form visual words is still subject to research, as a bit flip could potentially change a descriptor's mapping to the words space, resulting to low word repeatability and thus, violating one of the basic assumptions of the bag-of-words approach.

Following the approach suggested by Galvez and Tardos [36], we create a visual vocabulary adapted to the binary features used by OKVIS, namely BRISK [59], effectively reusing any features extracted in the loop of SLAM. The aim here is to exploit any scale and rotation invariance offered by BRISK. It should be noted that the descriptor size used within OKVIS consists of 48 bytes (instead of 64 as in the original implementation) and the feature orientation is aligned with the gravity since the inertial sensor provides this. To compute the BoBW vocabulary we discretize the 48-byte BRISK descriptors' space using about 3500 training images in total. These depict indoor and outdoor environments and are different to the ones used at runtime. The resulting vocabulary tree has 10 branches and 6 depth levels resulting to a vocabulary of a million words.

3.3 Appearance Check

The first step to place recognition is to check the current query keyframe Q against the Image Database for any entry with similar appearance. To this end, the BoBW descriptor of Q is scored based on its L1-distance to Database entries, using a 'term frequency–inverse document frequency' (tf-idf) weighting scheme [27] to suppress commonly occurring words to form the set $\mathcal{S}$ of matching keyframe candidates. Following the approach of [77], the set $\mathcal{N}_\mathcal{Q}$ of immediate neighbours of Q in the Covisibility Graph (i.e. depicting common scene structure) is formed, recording the minimum similarity score s_{min} between Q and any member of $\mathcal{N}_\mathcal{Q}$ computed as the L1-distance between their BoBW descriptors. Any candidate matching keyframes in $\mathcal{S}$ that score lower similarity to Q than s_{min} or already belong to $\mathcal{N}_\mathcal{Q}$ are removed from $\mathcal{S}$. All remaining members of $\mathcal{S}$ undergo non-maximum suppression within their immediate neighbourhood in the Covisibility Graph; all members of such a covisibility group are scored for their similarity to Q and the corresponding entry in $\mathcal{S}$ is replaced with the highest scoring keyframe in each group. If the sum of the highest N scores in one such group does not reach at least 75% of the best score across all groups, the corresponding entry in $\mathcal{S}$ is removed entirely. Finally, every surviving candidate in $\mathcal{S}$ is checked for covisibility consistency with at least 3 candidate matches surviving the last Appearance Checks (i.e. corresponding to

the two previous query keyframes). Two keyframes are defined to be covisibility-consistent if their covisibility groups share at least one keyframe. This last step aims to eliminate candidates in $\mathcal{S}$ that do not share similar appearance with the previous query keyframes.

3.4 Geometric Check

The BoW approach discards all spatial information between visual words by definition, accepting as a match two different images having the same words regardless of their constellation. While in ground robot navigation scenarios this might be enough [28], in UAV navigation, where very different viewpoints are expected, geometric verification of an appearance match is imperative. Moreover, while traditionally, place recognition techniques stop short of estimating a relative transformation between the matching frames (e.g. this would be enough in image retrieval), in robot navigation, this information constitutes very useful input to a subsequent optimization step to enforce the loop closure that is detected and avoid local minima. Realising this, [77] implement a geometrical validation step employing the Horn method [43], which given two sets of 3D map points with known correspondences, estimates a 3D rigid transformation between them if enough inliers are found. However, for dynamic camera motion with large viewpoint changes, SLAM systems struggle to find enough correct 3D map points needed for a successful Horn test resulting to much fewer loops detected than actually experienced.

The first priority in place recognition is to avoid false positive loop detections, however, false negatives become of particular interest in viewpoint-challenging cases as they occur far more commonly than in any other scenario, effectively limiting our ability to correct for accumulated drift. In this spirit, here we propose to first use the 3D-3D Horn's geometric verification and if this proves unsuccessful, check for a 2D-3D geometric consistency using the method of [53]. This provides a closed-form solution to the Perspective-Three-Point (P3P) problem for the full transformation between two camera poses in the world reference frame using at least three 2D-3D point correspondences.

For every keyframe candidate C (member of $\mathcal{S}$) to match Q that reaches the Geometric Check we compute the BRISK correspondences between them, limiting the correspondence search only to the keypoints that have a 3D landmark associated with them. Erroneous correspondences are removed using a second Nearest Neighbour (2nd NN) test [66], while we also apply bidirectional matching to discard ambiguous matches. If enough 3D-3D correspondences are found, we attempt to verify the 3D-3D geometry between Q and C by estimating their rigid transformation T_{QC} using Horn within a RANSAC scheme. However, if this approach fails to estimate a transformation with at least N inliers the 2D-3D geometry verification is attempted. In order to expand the set of correspondences to consider, the 2D keypoints in Q are tested for matches with the image projections of all 3D landmarks present in C, following the strict biderectional and 2nd NN tests. If enough 3D-2D correspondences are available we use the P3P method of [53] in a RANSAC scheme to try to estimate T_{QC}. If a transformation that satisfies a mini-

mum threshold on the average reprojection error in pixels is found, C is accepted as a loop closure for Q. After looping through all the candidates in $\mathcal{S}$ for a Geometric Check, the proposed method returns the T_{QC} with the highest number of inliers (i.e. points with a reprojection error is smaller than a pre-defined threshold) and the corresponding C. For our tests we usually define this threshold to be smaller than 2 pixels, the minimum number of matches as 12 and the number of inliers to accept a loop as 8.

4 Datasets

While datasets containing outdoor visual and inertial information, such as KITTI [41] exist, they are typically unsuitable to evaluate place recognition methods on. In KITTI for example, most sequences exhibit mainly forward camera motion with a front-looking camera, rendering it very difficult to correctly label the images for ground truth. For this reason, the datasets used in this work were recorded especially for place recognition applications using both flying and hand-held setups in the city center of Zurich with a side-looking camera, permitting clear decisions on ground truth labelling. These manually labelled datasets are being made publicly available, given that there are no other public datasets suited to place recognition providing ground truth, visual and inertial data as well as posing viewpoint and situational challenges as described below.

While we use our recorded datasets to assess the quality of the proposed pipeline in deciding whether the camera's trajectory experiences a loop closure, in order to test the quality of the proposed transformation, we use the publicly available EuRoC Micro Aerial Vehicle (MAV) dataset [17] providing indoor visual and inertial data from a flying UAV, which has its poses recorded by a Vicon external tracking system, providing very accurate full pose information. All the datasets in this work were recorded with a Visual-Inertial (VI) sensor [82] providing grayscale global-shutter images at 20 Hz and synchronized inertial measurements. In our tests, we perform monocular-inertial estimation by using only the information provided by one of the cameras of the sensor.

4.1 Shopping Street 1 and 2

These two datasets were recorded in a busy shopping street in the city center of Zurich using two different configurations. Shopping street 1 uses a hand-held setup, while Shopping Street 2 was recorded months later in the same area using a $4m$-long rod held vertically in order to capture the same scene from very different viewpoints. Shopping Street 1 consists of two traverses in the same street exhibiting small viewpoint changes, perceptual aliasing and appearance changes. We combine both sequences Shopping Street 1 and 2 obtaining a challenging dataset for place recognition, with major changes in the scene appearance, challenging lighting conditions and also strong viewpoint variations. Examples are shown in Fig. 5.4. These sequences were already successfully applied in a place recognition

scenario in our previous work [68].

4.2 UAV dataset

This sequence was recorded along a residential street using the VI sensor mounted on the bottom of an AscTec Neo UAV (visible in the inset of Fig. 5.1) in a front-looking configuration, while performing lateral movements with the UAV in both directions. This sequence exhibits perceptual aliasing as well as large variance in viewpoints and difficult lighting conditions as evident in Fig. 5.7 and Fig. 5.1.

5 Results

We evaluate the proposed approach on datasets labelled with ground truth as described in Section 4 and compare to the state of the art by analyzing their precision-recall characteristics. Moreover, as the proposed pipeline does not only provide a yes-or-no decision, but goes on to suggest a transformation between the matching keyframes to be used in a subsequent optimization step to enforce loop closure, we also evaluate the quality of these estimates. We present quantitative and qualitative evaluations on both hand-held and aerial scenarios.

5.1 Precision-Recall Characteristics

We record the precision-recall characteristics of the proposed method against FAB-MAP 2.0 [28], which is considered as the most well-established place recognition pipeline designed to combat perceptual aliasing. Moreover, as the method proposed in this paper employs binary features and draws inspiration from the DBoW2 approach of [36] we also compare to its performance. These tests are conducted on the Shopping Street 1 sequence. We test the proposed approach using a vocabulary composed of outdoor images captured in Zurich different to the ones used for testing. FAB-MAP and DBoW2 are tested using their corresponding original vocabularies. As evident in Fig. 5.3 (a), the proposed approach achieves higher recall across all methods for perfect precision (i.e. equal to 1). The robustness of the proposed method is illustrated qualitatively in Fig. 5.4. FAB-MAP is particularly challenged as it employs appearance-only checks in deciding for a loop-closure, while our approach and DBoW2 incorporate also geometric checks. DBoW2 exhibits high recall for perfect precision in Shopping Street 1, however, our improved geometric checks result to improved recall, which becomes particularly evident when testing with Shopping Street 1 & 2, where the viewpoint and other challenges are far greater. FAB-MAP precision-recall rates drop drastically (both to less than 0.1) in this case and DBoW2 detects four loops only. Despite that all of them are correct, they are far fewer than the total number of loop closures. The yellow curve in Fig. 5.3(c) illustrates the recall reached by DBoW2 while varying the reprojection error.

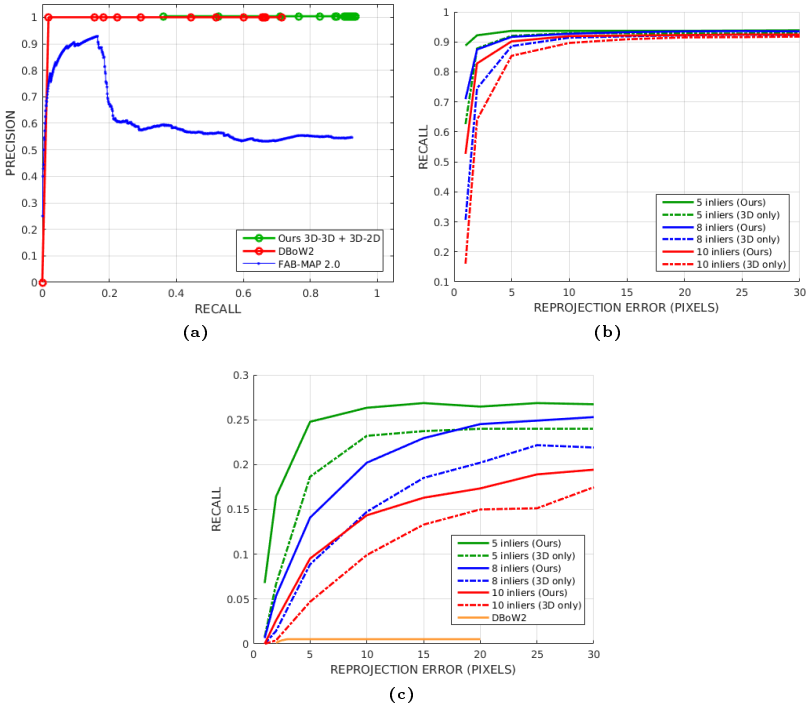


Figure 5.3: Precision & Recall analysis. Testing in Shopping Street 1, in (a) the proposed method outperforms FAB-MAP 2.0 and DBoW2. Maintaining perfect precision, in (b) and (c), recall is monitored for variable reprojection error thresholds for the proposed method in full (i.e. using all geometric checks) and using 3D checks only. Accepted inliers are varied from 10 (very restrictive) to 5 (most relaxed). Even in the more challenging dataset used in (c), the proposed method outperforms the 3D only approach and DBoW2 by a large amount.

As the proposed pipeline aims at greater robustness to viewpoint changes as well as to clean up false appearance matches, we employ both a 3D-3D geometric test similarly to ORB-SLAM [77], as well as a 3D-2D geometric test. A comparison on precision versus recall to ORB-SLAM would not be fair, however, as it was designed to conduct loop-closure tests that are well spaced in time instead of testing at every keyframe as in the proposed method. The type and quality of features used as well as the estimation processes involved in ORB-SLAM in comparison to

OKVIS have a direct impact on the quality of the performance of place recognition. So, here we isolate the effect of the 3D-3D and the 3D-2D geometric tests of the proposed pipeline to analyse the performance in both Shopping Street 1 alone and the dataset comprised of both Shopping Street 1 and 2 as shown in Fig. 5.3 (b) and (c), respectively.

Retaining perfect precision, we monitor the recall obtained for variable reprojection error dictating the number of inliers agreeing with the transformation proposed using RANSAC. While one might expect that introducing the 3D-2D geometric checks as a second chance for a candidate loop-closure following a failed 3D-3D geometric check would have a negative impact on the precision-vs-recall trade-off, Fig. 5.3(b) shows that higher recall can be achieved for the combined tests while retaining perfect precision. The added challenges in the Shopping Street 1 & 2 setup (greater changes in illumination, viewpoint and appearance as seen in Fig. 5.4), indeed causes lower overall recall in Fig. 5.3(c), but the combined 3D and 2D tests of the proposed approach still outperform the 3D only checks without compromising precision.

Traditionally, the answer to the question posed by place recognition techniques on whether we are re-visiting an already known place is binary (i.e. yes or no). Since our aim is to employ viewpoint-tolerant place recognition to indicate loop closures within SLAM, a first suggestion of the relative transformation between the loop closing frames (defined as T_{QC}) is not only very useful to a subsequent optimization step, but also an indication of the quality of the geometric checks used to decide for a loop closure in the first place. In the proposed scheme, the estimation of T_{QC} comes as a by-product of the Geometric Check step.

We use the EuRoC Vicon Room 2 03 sequence of the EuRoC MAV dataset, which provides high-precision ground-truth poses for the UAV throughout this sequence. Upon the detection of a loop closure, we evaluate the quality of T_{QC} against ground-truth testing for both the full pipeline described in Section 3 and when using the 3D-3D geometric checks only. For both variants of our pipeline, we accumulate the estimated translation error across 10 runs as illustrated in Fig. 5.5. It should be noted that due to the randomised nature of RANSAC, some loops are not detected in all runs. For completeness, we also analyse the translation error in the loop-closing transformations estimated by ORB-SLAM in the same scenario, seen on the right of Fig. 5.5. Relocalization was triggered many times due to ORB-SLAM losing track, while different keyframes are selected in each run, rendering it harder to detect the same loops across different runs than with OKVIS. Even without considering the lower recall of ORB-SLAM, Fig. 5.6 illustrates that the translation error in T_{QC} is much larger than with the proposed approach.

As evident in Fig. 5.5, the inclusion of the 3D-2D geometric tests can sometimes result to bigger translation error in the estimation of T_{QC}, as expected. In fact, loops 15 and 16 result to considerable error given the size of the room, where the dataset was recorded. However, out of the 17 loops detected by the full pipeline, only 4 have been detected when using the 3D checks only. It should be highlighted that many of the loop-closing transformations estimated by the full approach were still computed using Horn's 3D-3D method, since the covisibility consistency check

did not fail (as in the 3D-only case); given that 3 consecutive consistent keyframe matches are needed before accepting a loop closure, the additional loop detections provided by the 3D-2D checks lead to correct detection of more true-positives. The vast majority of the additional detections exhibit error of the same order as the more restrictive 3D only checks (i.e. less than 50cm), in stark contrast to the much larger error characteristics of ORB-SLAM in Fig. 5.6.

In conclusion, while the addition of inertial sensing can indeed result to better quality maps in OKVIS in comparison to ORB-SLAM, even when isolating the 3D only checks used in ORB-SLAM but using OKVIS maps, the proposed approach is evidently boosting recall and achieves better quality of loop-closing transformations T_{QC}. While T_{QC} is only a suggestion subject to further optimization in a bundle adjustment or pose-graph optimization step, the closer the estimate is to reality, the better the chances of subsequent convergence of the map to the global minimum. As a result, while the proposed use of additional 3D-2D checks can result to noisier transformations, these are still better than in ORB-SLAM and the sometimes dramatic increase in recall is evidently beneficial and can really make a difference in viewpoint-challenging scenarios.

5.2 UAV Experiments

The proposed approach was tested using the UAV dataset, exhibiting the biggest challenge for viewpoint-tolerant place recognition as visible in Fig. 5.7. Added challenges, such as in illumination can cause false negatives as feature detection is compromised. The loop-closures detected by our approach are visible (in green) in Fig. 5.8. ORB-SLAM was also tested using this sequence, but no loops were detected.

5.3 Computational Cost

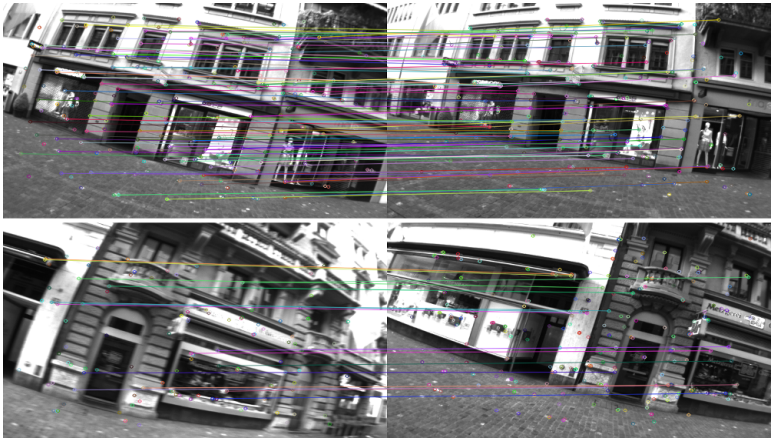
Feature extraction is usually the bottleneck in place recognition systems. With this in mind, the proposed method is re-using features extracted during the estimation of SLAM, enabling loop-closure detection at frame rate (i.e. 20Hz) in all the experiments presented in this paper. As the BRISK descriptor used within OKVIS consists of 48 bytes only, this restricts its descriptability posing bigger problems in loop detection, but makes descriptor comparisons even more efficient. Moreover, more relaxed conditions in the RANSAC scheme can be created in order to improve even more the performance, but the quality of transformations can also be affected.

6 Conclusions

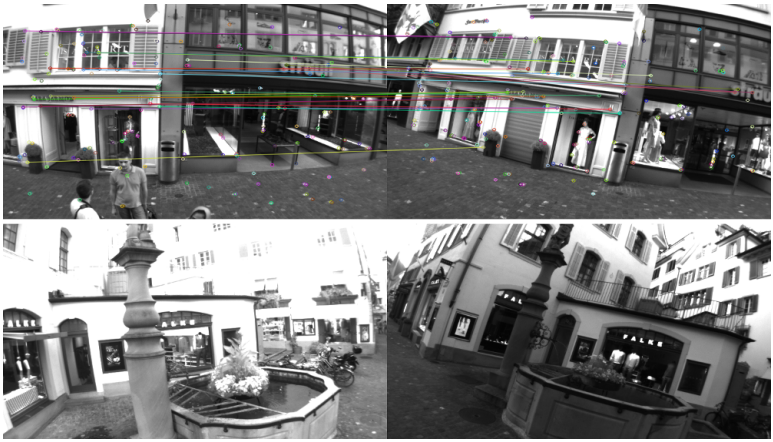
This paper proposes a novel pipeline for viewpoint-tolerant place recognition that makes use of promising leads from existing works, combining them in a way that enables unprecedented robustness to a wide range of common challenges (i.e. tolerance to viewpoint, lighting changes, occlusions, perceptual aliasing, etc). The proposed pipeline was carefully designed to support low-burden computation and

to take advantage of any scale and rotation invariance offered by BRISK using combined geometric checks that exploit not only the 2D information inherent in images but also the 3D information provided by a SLAM system.

Evaluation on newly recorded challenging outdoor datasets with both hand-held and aerial footage demonstrates that the proposed pipeline achieves better, or even drastically increased at times, recall in comparison to the state of the art, while maintaining perfect precision. Since no other such dataset appears in the literature, we make our testbed publicly available. Further evaluation on the quality of the estimated loop-closure transformation on an existing, indoor aerial dataset with pose ground truth reveals better quality of estimation than state of the art. Future work will study more extreme viewpoint changes and their impact on both similarity of appearance (e.g. consistency of word assignments) as well as geometry estimated by SLAM.

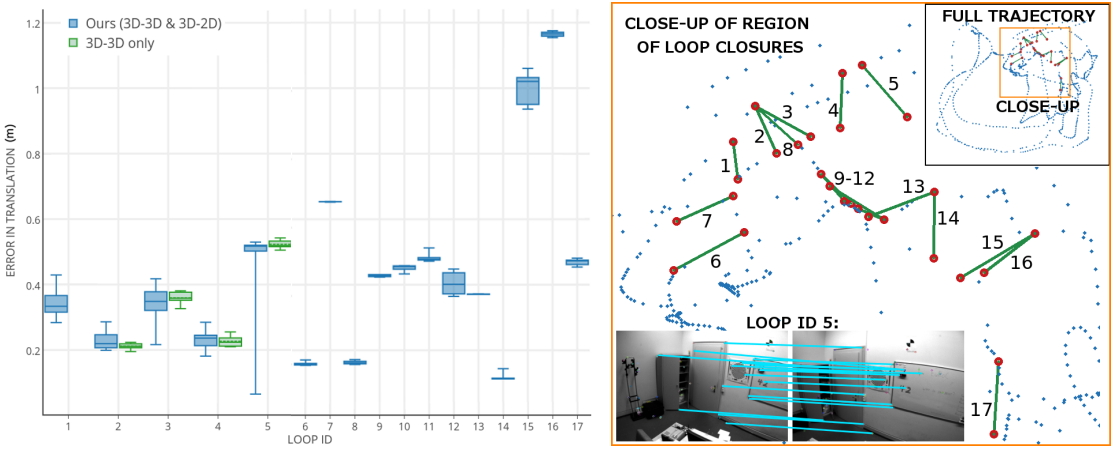


(a)



(b)

Figure 5.4: Example loop-closures from the Shopping Street dataset tested with the proposed approach. The loop-closured in (a) demonstrate robustness of the proposed approach to viewpoint changes and small motion blur (bottom left). In (b), the top image is an example of a loop detected across the Shopping Street 1 and 2 sequences exhibiting big changes in viewpoint and scene appearance, while the bottom image depicts a false negative, where the viewpoint and illumination changes proved too large for a loop-closure match.



(a) (b)

Figure 5.5: Left: Error in the translation estimates of the transformation T_{QC} between two loop-closing keyframes (each pair represented by one Loop ID) averaged over 10 runs. Right: the UAV poses obtained with OKVIS (blue dots) and the loop closures with their corresponding ID annotated in green. Loop ID 5 is shown in the inset.

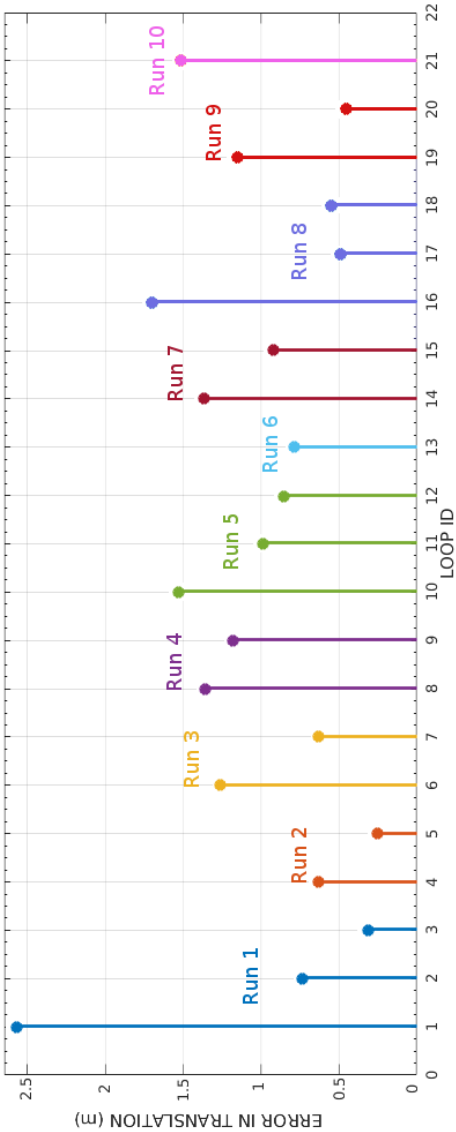


Figure 5.6: The translation error in T_{QC} as estimated by ORB-SLAM for the scenario of Fig. 5.5 (note: the loops detected here are different). Colors represent loops closed in each of the 10 runs, as different loops are detected every time.

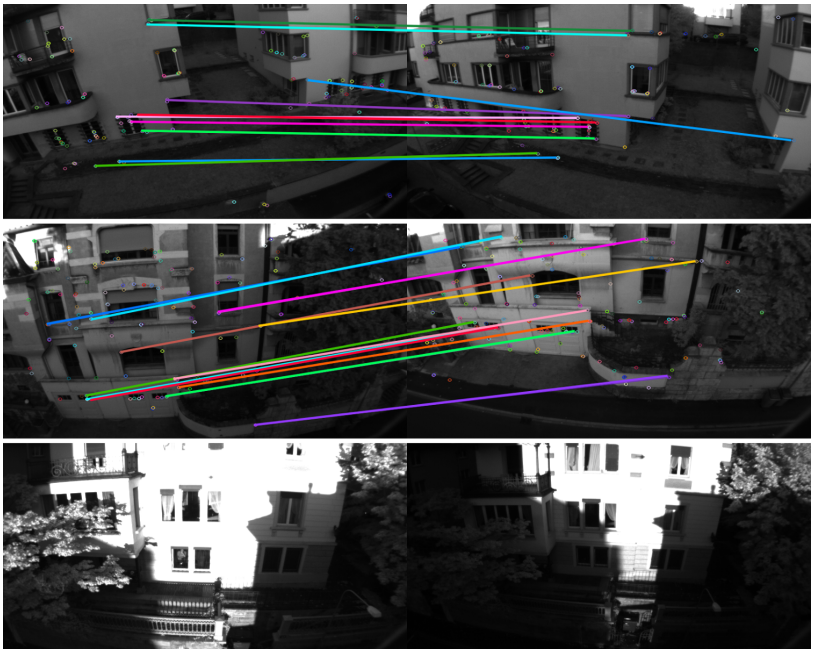


Figure 5.7: Loop-closures in the UAV dataset tested with the proposed approach. Large viewpoint changes are successfully handled (top two rows), while strong lighting can wipe crucial features out resulting to false negatives (bottom).

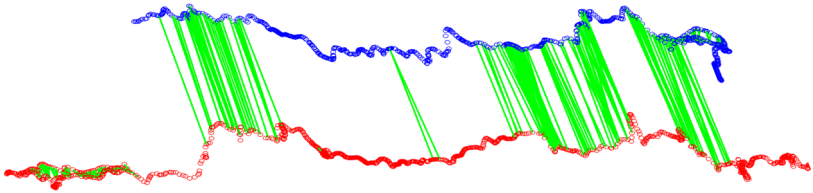


Figure 5.8: Trajectory followed by the UAV in the UAV dataset. In blue/red are the UAV trajectories when travelling in opposite directions and in green are the loop-closures detected.

Paper III

Real-time Wide-baseline Place Recognition using Depth Completion

Fabiola Maffra, Lucas Teixeira, Zetao Chen, Margarita Chli

Abstract

Place recognition is an essential capability for robotic autonomy. While ground robots observe the world from generally similar viewpoints over repeated visits, other robots, such as small aircraft, experience far more different viewpoints, requiring place recognition for images captured from very wide baselines. While traditional feature-based methods fail dramatically under extreme viewpoint changes, deep learning approaches demand heavy runtime processing. Driven by the need for cheaper alternatives able to run on computationally restricted platforms, such as small aircraft, this work proposes a novel real-time pipeline employing depth-completion on sparse feature maps that are anyway computed during robot localization and mapping, to enable place recognition at extreme viewpoint changes. The proposed approach demonstrates unprecedented precision-recall rates on challenging benchmarking and own synthetic and real datasets with up to 45° difference in viewpoints. In particular, our synthetic datasets are, to the best of our knowledge, the first to isolate the challenge of viewpoint changes for place recognition, addressing a crucial gap in the literature. All of the new datasets are publicly available to aid benchmarking.

Published in:
IEEE Robotics and Automation Letters (RA-L), 2019
DOI: 10.1109/LRA.2019.2895826

1 Introduction

Simultaneous Localization And Mapping (SLAM) refers to the process of building a map of the robot's workspace, while keeping track of its pose within it. In cases where SLAM estimation fails or drifts, it is essential to determine whether the robot has visited the current location in a previous occasion triggering relocalization. While originating from the problem of loop-closure detection, Place Recognition is also essential in multi-robot tasks, informing each robot where the others are. In scenarios, where multiple robots work in collaboration to carry out a given task, the scene is usually observed from very different viewpoints and assessing scene similarity from images captured under such wide baselines (e.g. ground to air) is known to be a very challenging task.

Place Recognition is commonly addressed using visual cues. It was the advent of real-time monocular systems for SLAM that paved the way towards the use of SLAM onboard small UAVs (Unmanned Aerial Vehicles). While many successful strategies for performing Place Recognition using range sensors have been proposed in the literature [30], these sensors are usually heavy and power greedy, severely reducing the endurance of small UAVs or even exceeding their payload capacity. For UAVs restricted to small payloads and as a result, limited computational capabilities, the employment of vision-based approaches comes as a natural choice for automating their navigation.

Motivated by the challenges of place recognition from aerial imagery, in this paper we specifically study the problem of Place Recognition under extreme changes in viewpoint. While still addressing common challenges in Place Recognition, such as illumination and situational changes, here, we push our method to the limits by testing on dramatic changes in viewpoint and showing that feature-based methods can still play a key role, enabling practical use in many common scenarios, such as 3D reconstruction of archaeological sites and collaborative multi-robot SLAM. Fig.6.1 shows a successful loop-closure detected using the proposed approach designed to address extreme changes in viewpoint.

The main contributions of this paper are:

- a novel real-time pipeline for loop-closure detection that employs depth-completion to enable feature-based matching between images captured from very different viewpoints. As such, this paper advocates and demonstrates that feature-based approaches are still useful for matching images across very wide baselines, while maintaining computation affordable for autonomous UAV navigation.

- new photo-realistic datasets exhibiting dramatic viewpoint changes in simulation, isolating for the first time the problem of viewpoint changes in Place Recognition from other challenges, such as scale variance, dynamicity of the scene, and illumination. In addition to these synthetic datasets, we also release real datasets capturing similarly large viewpoints using aerial and ground footage.

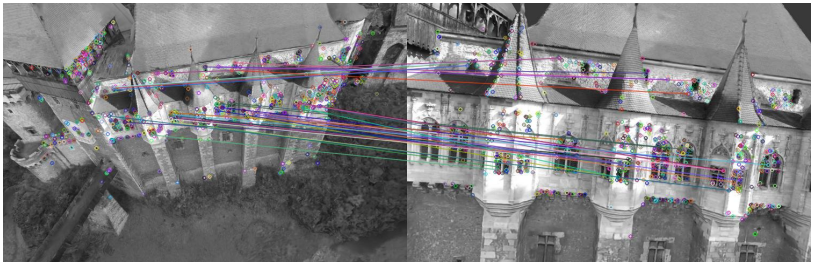


Figure 6.1: A loop in the synthetic Corvin dataset correctly detected by the proposed approach, despite the large change in viewpoint (45°).

2 Related Work

Most recent state-of-the-art SLAM systems, such as ORB-SLAM [77], employ image retrieval techniques to enable large-scale place recognition. A Bag-of-Words (BoW) approach combined with an inverted-file-index [102] or its more compact representations, such as Fisher Vectors [46] or Vectors of Locally Aggregated Descriptors (VLAD) [4] are usually applied to efficiently search for loop-closure candidates in a database of images containing all previous experiences of the robot. The widely known BoW approach relies on discretizing the feature-descriptors' space to build a dictionary of visual words that are then used to describe new images by converting locally invariant feature-descriptors into a BoW representation. Although several well-performing feature-based algorithms have been proposed for place recognition [28, 36], the extraction of unique and repeatably recognizable features has proven to be far from trivial [63]. In fact, extreme changes in appearance can pose a significant challenge for feature-based approaches. As a result, approaches using range sensors [30] or structural descriptors [25] have been proposed exploiting the fact that geometry offers better invariance to viewpoint changes when occlusion is not present.

Current feature-based BoW approaches try to circumvent major changes in appearance by using high-quality feature detectors and descriptors, such as SIFT [66] and SURF [11]. However, these features still fail when large changes in viewpoint occur, and are typically too expensive to be employed in real-time applications, for example onboard a small UAV. Affine SIFT features [76] handle large image distortions by generating multiple affine transformations of an image before applying traditional SIFT. However, their increased invariance comes at a prohibitively high computational cost of two orders of magnitude slower than SIFT. By generating a mesh of the current robot's surroundings, the work in [69], makes use of a 3D map provided by SLAM and identifies the most prominent plane in each image computing only one affine transformation, as orthophoto. This enables the creation of

a single view of the scene, while using a computationally cheap binary descriptor and avoiding the need for computing multiple transformations of the same image.

While purely 2D image-based approaches can offer the ability to localize images even if local feature matching fails, these methods are usually considered unsuitable for accurate visual localization. 3D structure-based approaches offer more precise pose estimation, becoming a natural choice for visual place recognition methods, which require the recovery of the 6-DoF camera pose. Sattler et al. [94] combine both methods by querying an image database to retrieve a set of related images depicting the same place and performing a small-scale Structure From Motion (SFM) to obtain a local 3D reconstruction around a query image. 3D structure-based techniques assume that the scene is represented by a 3D model, usually obtained from SFM [62] or SLAM [32], and the camera pose can be obtained using a PnP solver [54] in a RANSAC scheme [33]. Another widely used approach is to use LIDAR sensors to obtain the 3D structure of the environment in very fine resolution. SegMatch [30], for example, performs place recognition using 3D laser data using the concept of segment matching. Despite the reduced amount of noise, these maps are usually sparser than maps obtained using vision-based approaches, and as already mentioned, range sensors are still too heavy and often too power-consuming to be carried on a small UAV.

More recently, Convolutional Neural Networks (CNNs) have been successfully demonstrated to extract robust feature descriptors for place recognition [7] or even to regress a 6-DoF pose directly from images [51]. While shown to produce impressive results even under extreme changes in appearance, deep learning techniques, however, usually rely on powerful GPUs, rendering them too computationally expensive to run onboard a small aircraft. Besides this, they also rely on very large, annotated datasets, which are very hard to obtain.

3 Methodology

In the proposed Place Recognition pipeline, illustrated in Fig. 6.2, we assume that vision-based SLAM running onboard the robot provides, for each image entering the pipeline, a sparse 3D map of the location and, optionally, its 2D features (i.e. keypoints and descriptors). When a new image arrives, a map densification step generates a denser 3D map from the sparse 3D map provided by SLAM using a depth completion approach. New image features can be detected for Place Recognition if the user desires different features from the ones used in SLAM. All features get converted into a BoW representation in order to search for loop-closure candidates that have similar appearance to the query image. A candidate filtering step refines and removes erroneous loop-closure candidates by exploiting covisibility information captured by SLAM. Any remaining loop-closure candidates proceed to a geometric check, where geometric compatibility between the query and each candidate is evaluated by using all their 2D features and their denser 3D maps. If the geometric check succeeds, a loop-closure is deemed as detected and the pipeline returns the loop-closure match with the most keypoints in agreement with the

query.

Sections 3.1, 3.2 and 3.4 describe briefly the main steps of the pipeline already introduced in [70], while Section 3.3 focuses on the main novelty of this paper, the use of depth-completion to improve the establishment of 3D-3D and 3D-2D correspondences during geometric checks, which is the key component enabling feature-based matching across images of very different viewpoints.

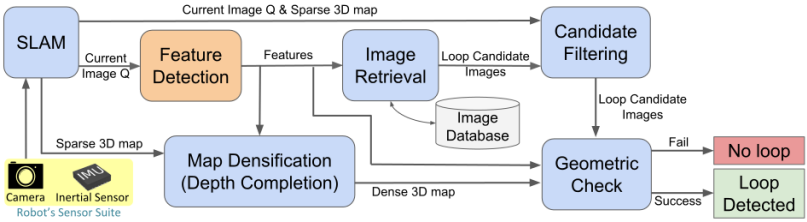


Figure 6.2: The proposed pipeline for Place Recognition employing depth completion with appearance and geometric checks to determine whether the current image Q forms a loop closure with an image in the database containing past robot experiences.

3.1 Loop-Closure Candidates Retrieval

Following the approach suggested by Galvez and Tardos [36], a hierarchical BoW visual vocabulary is formed by discretizing the feature-descriptors' space into a set of visual words. When a new query image arrives, local features, such as BRISK or SURF, are extracted and converted to a BoW representation, used to retrieve a set of database images similar to the current image. The BoW descriptor is scored based on its distance to database entries, using a 'term frequency-inverse document frequency' (tf-idf) weighting scheme [28] to suppress commonly occurring words.

The decision of the feature detector and descriptor to be used is left open in the proposed framework, as this decision affects the trade-off between precision and recall. While SIFT [66] and SURF [11] features can be used in the pipeline, for example, their bigger accuracy comes at the cost of longer run-times, when compared to binary features, such as BRISK [59] and ORB [92]. As BRISK features require low computational cost, being more suitable for UAV navigation, here we use BRISK for our experiments.

3.2 Candidate Filtering

As geometric checks are usually expensive, here covisibility information captured by SLAM is firstly used to refine and remove erroneous loop-closure candidates

suggested by the BoW descriptors when querying the image database. Following the same approach as in [77], the proposed pipeline implements a covisibility graph, where each node is a frame and an edge between two nodes exists if they share enough observations of the same 3D points in the SLAM map. As a simplification, in case of loop-closure detection the covisibility graph is not updated, keeping only covisibility information at the frames' neighbourhood. At first, the minimum score S_{min} between the query and its neighbours in the covisibility graph is recorded, and any candidate which scores lower than 75% of S_{min} is excluded from the list of candidates. While [77] removes all candidates lower than S_{min} avoiding false-positive at all costs, here we employ a more permissive filter in order to recover candidate images taken from more distinct viewpoints subject to strict checks later on. As many overlapping frames exist, when querying the database, many images will exhibit a high score when compared to the query image. These overlapping images are taken into account by summing up the scores of the images that are neighbours in the covisibility graph. Any loop-closure candidate scoring higher than 75% of the best score will proceed to the next step. A candidate loop image is accepted if three consecutive loop candidates are consistent. Two frames are defined to be covisibility-consistent if they share at least one frame among their covisibility neighbours. More details about this approach can be found in [77].

3.3 Map Densification using Depth Completion

During the geometric check, geometric consistency between the query and the candidate is evaluated by computing the query's pose in the candidate's coordinate frame. This procedure requires the establishment of 3D-3D or 3D-2D correspondences between the query-candidate pair. Assuming that the scene is represented by a 3D map, and each 3D point is associated with one or more local descriptors in the image space, 3D-3D and 3D-2D correspondences are obtained via descriptors matching in the image space. However, under extreme viewpoint changes, feature-based image matching is strongly affected by affine distortions and occlusions, resulting in a reduced number of correspondences between the query's and the candidate's keypoints. Besides this, it must be noted that only keypoints successfully tracked by SLAM have a 3D landmark associated with them. As such, only a small number of keypoints carrying 3D information arrives to the geometric check. By using a depth completion for map densification, interpolated 3D landmarks can be estimated for the 2D keypoints that have no depth-estimates yet, improving the establishment of 3D-3D and 3D-2D correspondences for images captured across very wide baselines.

Fig. 6.3 illustrates the map densification pipeline, which consists of a depth completion step, shown in Figs. 6.3a-6.3b, followed by the creation of the interpolated 3D landmarks, illustrated in Fig. 6.3c. Our map-densification algorithm, takes as input the camera pose, the 3D landmarks visible by this camera (dark green) and the 2D keypoints (red), for which we want to calculate an interpolated landmark. A dense mesh of the 3D landmarks is first computed (in purple in Fig. 6.3a) using the open-source mesh-generation pipeline of [106]. A depth image, of the same size

as the camera image is obtained by rendering this mesh into the image plane and extracting the depth-buffer of the render engine, as shown in Fig. 6.3b, illustrating the 2D keypoints in red and the projections of the 3D landmarks in the image in green. Any 2D keypoints lying over a pixel with depth information, have their corresponding 3D landmarks estimated, in camera coordinates, by using the pixel's coordinates and the depth value on that pixel, using Equation (1). Any remaining 2D keypoints cannot have a 3D landmark established. Fig. 6.3c shows, in blue, the new, interpolated 3D landmarks added to create a denser map of the scene.

$$P_c = (X, Y, Z) = \left(\frac{(u - u_0) * d}{f_u}, \frac{(v - v_0) * d}{f_v}, d \right), \tag{1}$$

where (u, v) is the position of the detected keypoint, u_0 and v_0 are the pixel coordinates of the camera's optical center, f_u and f_v are the focal length in u- and v-direction, respectively, and d is the depth provided by the mesh at the pixel (u, v).

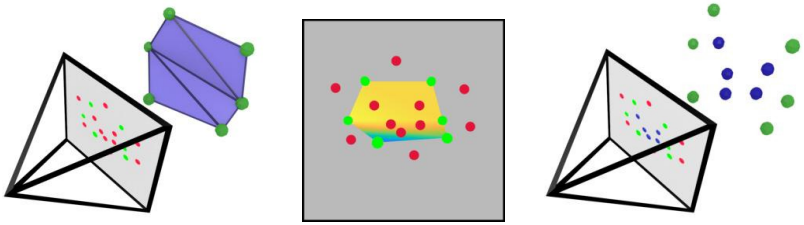


Figure 6.3: The map-densification process: the green 3D landmarks are used to estimate the depth of the red 2D keypoints by creating a mesh (in purple) in (a), and projecting it in a depth image visible in (b). This results to the additional blue 3D landmarks in (c).

While it is possible to extract all the keypoints needed for Place Recognition during SLAM, only a reduced number of them, represented in bright green in Fig. 6.3, can be tracked in order to keep its real-time performance. With OKVIS [60] (the SLAM system used in our experiments), for example, we can usually track about 400 landmarks while maintaining real-time performance, however, about 1000 keypoints were used here for Place Recognition. As such, the map-densification approach focuses in estimating the 3D landmarks for the keypoints that were ignored or not successfully tracked by SLAM. However, if the type of keypoints and descriptors used for Place Recognition is different from the one used during SLAM, new features need to be detected. In this case, the map densification will try to estimate a 3D landmark for every newly detected keypoint. One advantage of the latter case is a better decoupling between the SLAM method and Place Recognition.

Another advantage of the proposed map-densification approach is that it can handle arbitrarily sparse maps, which can contain certain amount of noise, while traditional depth-completion algorithms, such as [73], rely on good quality and not very sparse depth images as input in order to create a dense depth image. Here, we opted to use a mesh-based approach to create a dense depth image out of the 3D landmarks provided by SLAM. A higher quality representation of the scene is then obtained using the mesh generation pipeline of [106], which applies a Delaunay triangulation followed by an outlier removal to create a 3D triangle mesh out of the 3D landmarks provided by SLAM. Assuming local planarity among neighbouring vertices of the mesh, outlier removal is performed by comparing the value of a vertex with the centroid of the vertex's neighbourhood. In case of a large disagreement the vertex is eliminated. This approach prioritizes high-quality depth estimations instead of a full representation of the mesh, and holes can exist at points with a high local depth uncertainty. While very efficient in removing outliers, the use of a sparse 3D map together with the local planarity assumption create a smooth mesh of the environment, eliminating details in small areas with a large depth variation. However, as demonstrated in [69] and [106], this approach was already proven to work well in man-made environments, where locally planar structures are usually present. Besides this, this mesh generation approach takes about 7 ms per frame to create a 3D mesh out of the 3D landmarks, rendering it suitable for real-time applications.

3.4 Geometric Check

The BoW approach does not use any geometric information for image retrieval, accepting two images as a match if they present a similar collection of words. As geometry was shown to play a key role in identifying true loop-closures, here we employ the geometric checks proposed in our previous work [70]. Geometric consistency between a query-candidate pair is evaluated by computing the query's pose in the candidate's coordinate frame. If a pose P_Q^C can be successfully estimated, the candidate is accepted as a loop-closure for the query.

When testing for geometric consistency, we first search for feature correspondences between the query Q and a candidate C using only keypoints with associated 3D landmarks. If enough 3D correspondences are found, we attempt to estimate a similarity transformation (i.e. translation, rotation and scale) between the query and the candidate using Horn's method [25] in a RANSAC scheme [18]. If a transformation that satisfies a minimum threshold on the average reprojection error is found, the candidate is accepted as a loop-closure for the query. In this case, P_Q^C can be easily recovered by multiplying the candidate's pose on his own coordinate system by the similarity transformation. However, if a transformation cannot be estimated or not enough 3D-3D matches can be found, the set of correspondences to be considered is expanded by searching for feature correspondences between the candidate's keypoints with 3D landmark associated and all the keypoints in the query Q. If enough 3D-2D matches are found, we attempt to directly estimate the pose P_Q^C using the 3D-2D matches [17]. If this succeeds, a loop-closure is deemed

as detected. We repeat this process to all loop-closure candidates and select the candidate match with biggest number of inliers.

4 Datasets

In this work, two types of datasets are used to evaluate the proposed method. To isolate the problem of viewpoint changes in place recognition, while keeping full control of the test conditions, we set up a photo-realistic simulation. Finally, tests are conducted also in real conditions, using datasets recorded with hand-held cameras and aerial robots, exhibiting very different viewpoints, such as air-ground matching.

4.1 Photo-realistic Synthetic Datasets

Large scale outdoor experiments using real robots are the best way to validate a place recognition algorithm. However, such data lacks not only the ground-truth of the robot's poses but also the 3D model of the environment. Traditional methods of constructing ground-truth poses, such as with GPS or laser tracking, estimate the robot's position with an accuracy of several centimetres at best, but can also be up to a few meters inaccurate. Even more problematic is the orientation estimation of the camera that is usually unknown or only roughly estimated in post-processing. In order to guarantee good ground-truth for the loop-closures, some datasets are manually annotated, such as in [70]. By making use of synthetic datasets, ground-truth information is easily obtained, allowing quantitative evaluation of the method by automatically estimating the ground-truth.

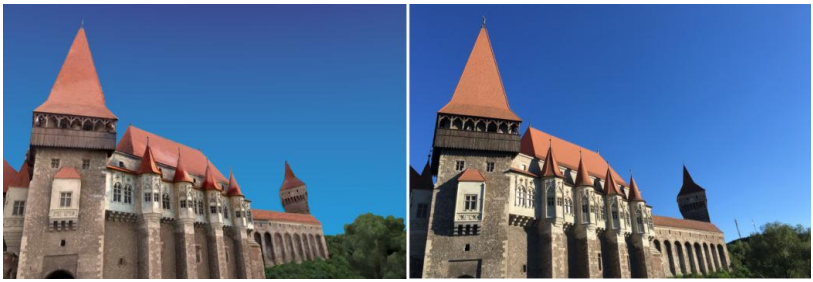


Figure 6.4: The Left image shows the result of our simulation and the right is an actual picture taken from the same place with a consumer camera.

In order to create our synthetic datasets, we use 3D models obtained by photogrammetric reconstruction. We create UAV trajectories using the Rotors UAV physical simulator [35] and the RGBD images are produced by the Blender render

(a) L'Agout in Castres **(b)** 0° view **(c)** 15° view

(d) 30° view **(e)** 45° view

Figure 6.5: L'Agout dataset: 3D photometric reconstruction of medieval houses. In (a), a picture of the location shows houses of about 15m height by 100m width in total and a depth variation of 3m among the facades. In (b)-(e) are example images from L'Agout dataset at 0°, 15°, 30° and 45°, respectively.

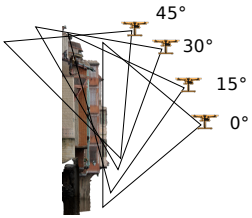
engine. Fig. 6.4 shows that our simulation produces images that are very similar to the real ones. This approach on dataset generation produces visual-inertial measurements that reproduce the Skybotix VI-Sensor with resolution of 752×480 pixels, the same resolution as in the outdoor real datasets. Defining as loop a pair of images with more than 50% of overlap and using the ground-truth poses provided by the physical simulator, we were able to easily distinguish (and annotate) the image-pairs that constitute loops.

Namely, we construct the following datasets:

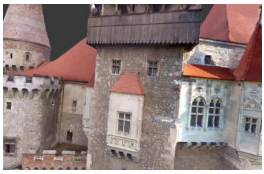
The L'Agout 0° & 15° & 30° & 45° dataset was produced using aerial pictures of "Maisons sur l'Agout" visible in Fig. 6.5, depicting medieval houses with balconies over the river Agout. We produce 4 sequences of 100 meters with a laterally moving drone carrying a camera facing the houses at 0° (i.e. pointing forwards), 15° from the horizon, 30°, and 45° as shown in Fig. 6.6b. It is important to highlight that the position of the drone was chosen in a way that the camera frustum is completely filled by the buildings in order to guarantee that the only difference between these sequences happens in the viewpoint, without any changes in scale.

The Corvin 0° & 30° & 45° dataset was produced using aerial footage of the Corvin Castle visible in Figs. 6.4 and 6.6. We produced 3 sequences at 0°, 30°, and 45°, while doing a 300-meter circular flight around the castle. These sequences

(a) Corvin Castle



(b) angles side view



(c) 0° view

(d) 30° view

(e) 45° view

Figure 6.6: Corvin dataset: 3D photometric reconstruction of Corvin Castle (a). In (b), the different viewpoints used to record the synthetic datasets, and in (c), (d) and (e), example images from Corvin dataset at 0°, 30° and 45°, respectively, are shown.

capture a scene composed of a large range of different depths.

4.2 Outdoor Real Datasets

While we focus our real world experiments on publicly available datasets, we also construct a new air-ground dataset, which we make publicly available together with the new synthetic datasets. All real datasets used in this paper were recorded using a Skybotix VI-Sensor, using only one camera and one IMU in a hand-held setup or mounted on an AscTec Hexacopter Neo for different viewpoints. The datasets are:

Shopping street 1 dataset [70] ⟼ Ground-Ground is a hand-held dataset with the camera revisiting the same location with very similar viewpoints in a busy shopping street in Zurich.

OldCity dataset [69] ⟼ Ground-Ground consists of two walking sequences of 230m in the old city of Zurich, presenting a more complex scenario due to the presence of narrow passages in this area, providing wide range of viewpoints of the same places.

Clausius street dataset [70] ⟼ Air-Air is a dataset recorded along a residential street with the camera mounted on the UAV, facing the buildings of one

street side, while performing lateral movements with the UAV in both directions. The two air sequences exhibit large viewpoint changes, perceptual aliasing and strong lighting changes.

Clausius street dataset $\longmapsto$ Air-Ground was recorded in the same street, with the air sequence taken from the previous dataset, while a new hand-held sequence was recorded on the same day. This is the most challenging real dataset because of its extreme viewpoint changes.

We benchmark the proposed pipeline against three state of the art place recognition algorithms that are suitable for UAV navigation, referred to here as BoBW [36], ORTHO [69] and VTPR, a modified version of [21] for ease of comparisons. In particular, VTPR here, corresponds to the methodology of [21], albeit using the same feature descriptors (i.e. BRISK instead of BRISK-48-bytes) as used in our method, as well as small modifications in the candidate filtering step. This strategy reveals the true power of map densification, which is also the main contribution of this work. It should be noted, however, that with these modifications VTPR achieves slightly better results than the original method of [21]. The use of BoBW with ORB [92] features in [77], was shown to provide scale and rotation invariance, while keeping real-time capabilities. ORTHO makes use of BRISK [59] features and minimizes the effect of viewpoint changes by using a mesh-based approach to create orthophotos projecting the image to the most salient plane in the scene.

Although the decision of the feature detector and descriptor to be used is left to open in the proposed pipeline, here we choose to run our experiments using BRISK features, which provide a good matching performance at a very low computational cost. To build a visual vocabulary as in [36], we discretize a BRISK descriptors' space using 6000 images, different from the ones used for testing, depicting indoor and outdoor environments. A vocabulary of 1 million words is generated by building a vocabulary tree with 10 branches and 6 depth levels. The same vocabulary is used throughout all the experiments, demonstrating the robustness of the method.

4.3 Narrow viewpoint changes

We test the proposed pipeline and the selected algorithms on narrow baselines in order to validate our algorithm on publicly available datasets against the state of the art in conditions that existing algorithms are designed for.

First, we record the precision-recall curves for all algorithms on the Shopping Street 1 dataset, which depicts a planar scene at small viewpoint changes. All the algorithms perform well in this dataset, with the proposed method presenting the highest recall (0.96) at precision 1, against 0.94 for both BoBW and VTPR, and 0.78 for ORTHO.

Precision-recall curves for the Old City dataset are visible in Fig. 6.7. This dataset exhibits both small and challenging viewpoint changes. As such, all algorithms can recover correct loops in areas with small changes in viewpoint, while maintaining perfect precision. However, the proposed method can also recover correct loops in areas with challenging viewpoint changes, achieving recall 0.79 at precision 1 and outperforming all others algorithms Example loop-closure detec-

tions using the proposed approach in the Shopping Street 1 and Old City datasets are shown in Fig. 6.8a and 6.8b, respectively.

The methods were also tested in the Air-Air Clausius Street dataset. The loop-closures detected by our approach and a correct match are illustrated in Fig. 6.9. While the proposed approach detects one false positive loop, BoBW and ORTHO detect only few correct matches and much more false positives in this dataset. VTPR detects about half of the loops detected by the proposed approach, as can be seen in comparison to the results in [70], however, without any false positive detections.

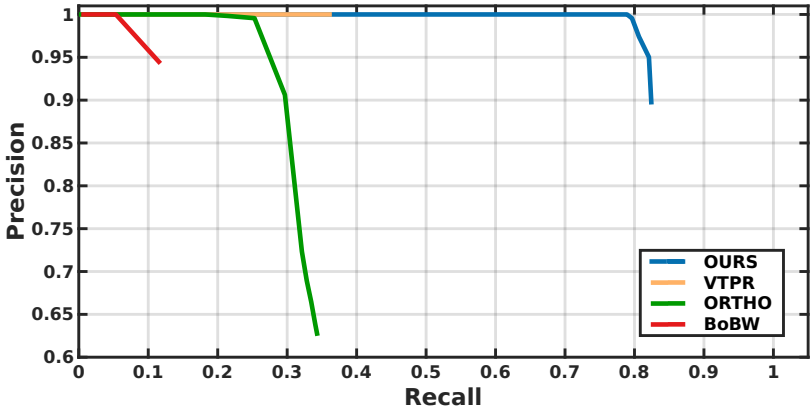
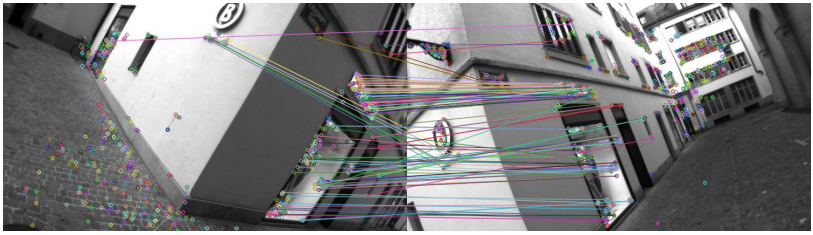


Figure 6.7: Precision-Recall Curves for the Old City dataset, showing that the proposed approach outperforms BoBW and ORTHO in scenarios where these algorithms are designed for, planar scenes (in the case of ORTHO) and narrow viewpoint changes.

4.4 Image Retrieval and Candidate Filtering in wide viewpoint changes

In our exploration towards robust loop-closure detection under large viewpoint changes, the first step was to determine whether our image retrieval algorithm works in these conditions and how many of the top candidates we need in order to guarantee a good chance of having at least one correct candidate in the set passed on to the geometric check. Fig. 6.10 shows the percentage of queries with at least one correct candidate before and after the candidate-filtering step, while varying the number of images retrieved from the image database. Note that the candidate filtering step not only removes erroneous candidates, but also filters out some

(a) Shopping Street 1: small viewpoint changes



(b) Old City: more challenging viewpoints

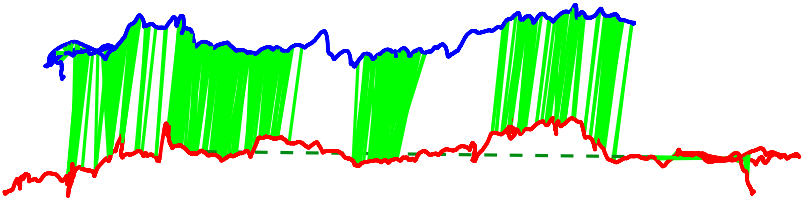
Figure 6.8: Example loop-closures from the Shopping Street 1 and Old City datasets.

correct loop-closure candidates. Empirically, retrieving the top 30 most similar images to a query from the image database is enough to provide enough correct candidates for the next steps of the pipeline. Despite the decrease from 97% (before candidate filtering) to about 88% (after candidate filtering) in the percentage of queries with at least one correct candidate in L'Agout 0°-45° sequence matching, most of the queries can still provide correct candidates for the geometric check, without much compromise in performance.

4.5 Wide viewpoint changes

In order to evaluate how the proposed method performs with increasing changes in viewpoint, we first test for loop-closures within the L'Agout dataset. We test the sequence at 0° against all others (i.e. 15°, 30° or 45°). Except for the neighbours of the current position, that depict the same place and cannot be detected, no self-loops exist along a single sequence. However, as all images entering the pipeline are tested for loop-closures before being inserted into the database of images, false positive detections are still possible inside one sequence. Fig. 6.11a shows the precision-recall curves for L'Agout for all the algorithms. Although all algorithms perform well at 15° of viewpoint changes (i.e. 0°-15°), both VTPR

(a) Loop-closure in the Air-Air Clausius Street dataset



(b) UAV trajectories (in red and blue) and detected loops (in bright green)

Figure 6.9: Loop closures in the Air-Air Clausius Street dataset: in (a), is an example loop-closure detected using the proposed method and in (b), the trajectories followed by the UAV, and the loops correctly detected between them. A false loop detection is shown in the dashed green line.

and the proposed method achieve the highest recall (0.97) for perfect precision. At 30°, both methods achieve a recall of 0.72 for perfect precision against 0.21 for ORTHO, while BoBW fails to detect loop-closures. The robustness of the proposed algorithm in viewpoint changes becomes evident at larger angles. At 45°, the proposed approach achieves recall of 0.54 for perfect precision against 0.38 for VTPR, representing an improvement of 42% with relation to the latter one, while both BoBW and ORTHO fail quickly. Fig. 6.14 shows a correct loop-closure detected in the L'Agout 0°-45° dataset, using the proposed approach.

We repeat the same experiment for the Corvin dataset, which captures a scene with strong depth variations. We record precision-recall curves for the sequence at 0° against the one at 30° and at 45°. As evident in Fig. 6.11b, these datasets present great challenges for all algorithms, with BoBW and ORTHO failing quickly. While VTPR achieves, a recall of 0.5 at 30° and 0.04 at 45° viewpoint changes for perfect precision, the proposed method achieves a recall of 0.71 at 30° and 0.14 at 45° for the same precision. This represents an improvement of 40% at 30° and 250% at 45° of viewpoint changes, when compared to VTPR. Fig. 6.13 depicts correct loop-closure detections, in the Corvin dataset, using the proposed

approach.

The methods were also tested in the Air-Ground Clausius Street dataset. While our approach detects one false positive loop and many correct loops, as shown in Fig. 6.12, BoBW, ORTHO and VTPR detect only very few correct matches (less than 5) and few more false positives.

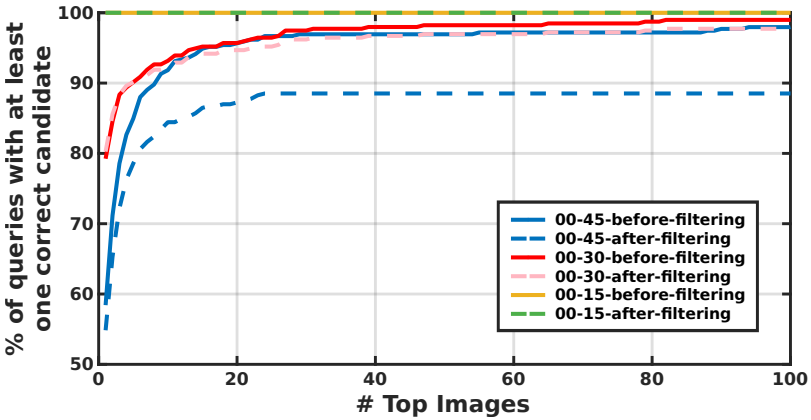
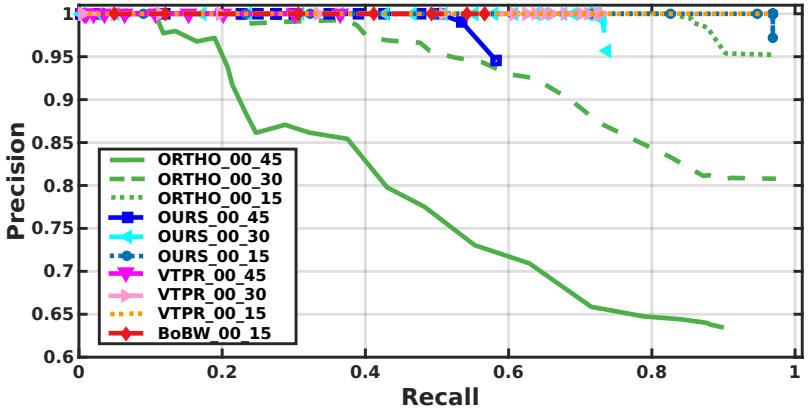


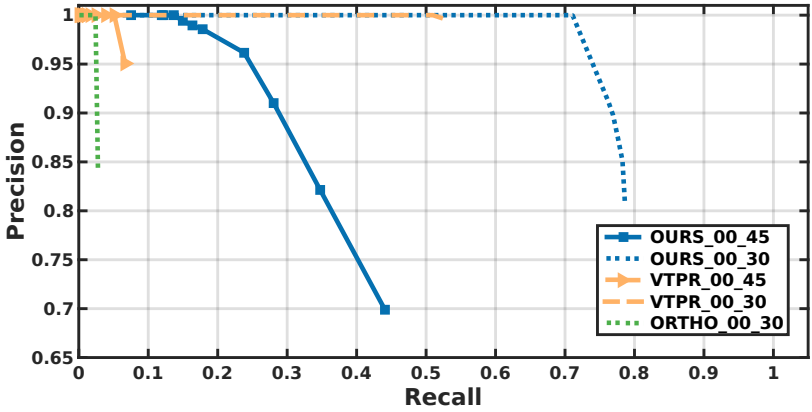
Figure 6.10: The percentage of queries with at least one correct loop-closure candidate to be passed on to the geometric check for different viewpoint changes. We provide curves both before and after the candidate-filtering step used for efficiency, while varying the number of top images retrieved from the image database. The higher the percentage achieved, the better the chance of discovering the correct loop-closure after the geometric check.

5 Timings

As consecutive frames are usually very similar, loop-closure detection does not need to be attempted at every frame, so in practice, runtime in the range of 1-5 Hz is enough for real-life applications. In the worst-case scenario, where all candidates entering the geometric check are tested for loop-closure, the proposed algorithm runs at 5Hz on average on a single core Intel i7 2.8GHz, allowing real-time place recognition within a SLAM system. Setting a maximum of 50 image-candidates at the end of the image-retrieval step, we avoid compromising the timings in cases of longer robot trajectories (resulting to larger image databases). In reality, even faster performance is expected as the geometric check can abort as soon the first suitable candidate is found.

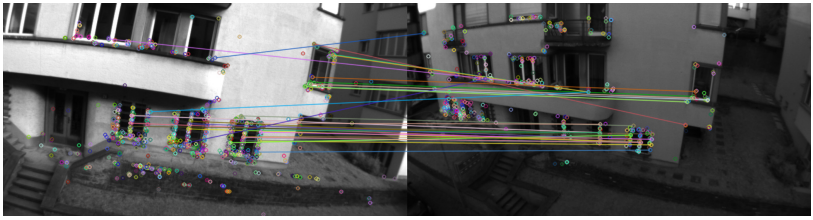


(a) L'Agout dataset

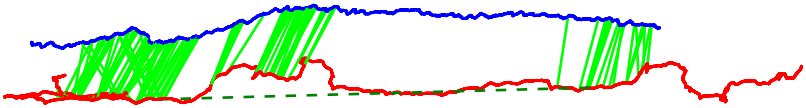


(b) Corvin dataset

Figure 6.11: Precision-Recal Curves in the L'Agout dataset in (a) using different viewpoint variations (from 0° to 15°, 30° and 45°), and in the Corvin dataset in (b) while varying the viewpoints from 0° to 30° and to 45°.



(a) Loop-Closure in the Air-Ground Clausius Street dataset



(b) UAV trajectory (in red and blue) and detected loops (in bright green)

Figure 6.12: Air-Ground Clausius Street dataset: In (a), example loop-closure detected using the proposed method and in (b) the trajectory followed by the UAV and by the hand-held setup, and the loops correctly detected between them. A false loop detection in the dashed green line.

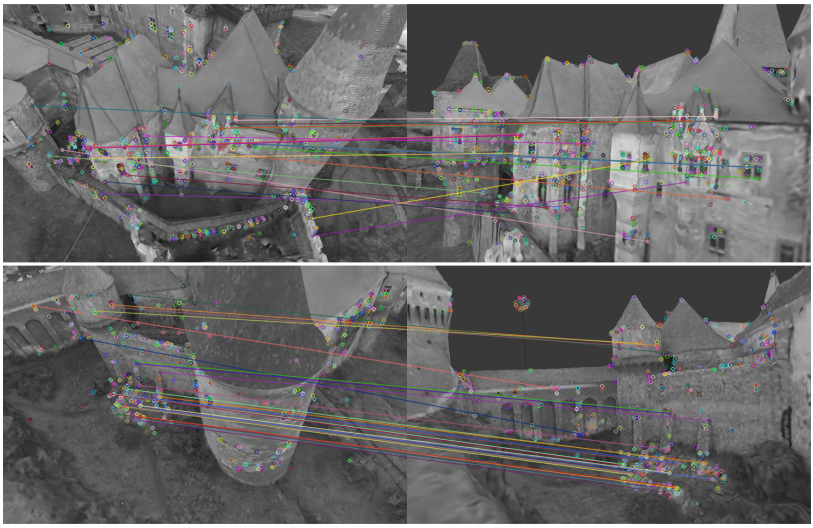


Figure 6.13: Example loop-closure detections in the Corvin dataset using the proposed approach. A viewpoint change from 0° to 45° illustrates the extent of the challenge in this dataset.

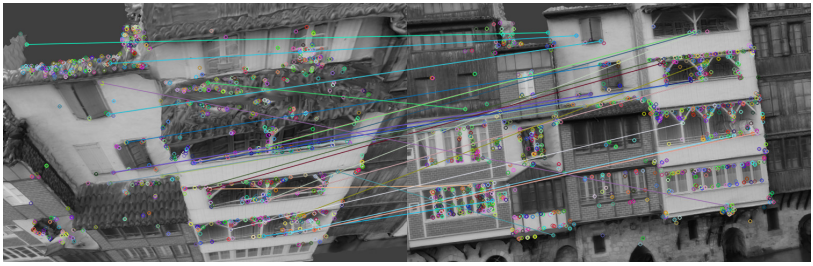


Figure 6.14: An example loop-closure detection in the L'Agout dataset using the proposed approach for a change in viewpoint from 0° to 45°.

6 Conclusion

This paper proposes a new place recognition pipeline capable of addressing dramatic changes in viewpoint (of up to 45°), while maintaining robustness at smaller angles, from narrow baselines. It relies on a depth-completion approach to improve the establishment of 3D correspondences during geometric checks, enabling feature-based matching across images captured from very wide baselines.

Evaluation on synthetic and real datasets with both hand-held and aerial footage, reveals that the proposed method achieves significant improvement in precision and recall in comparison to the state of the art, while keeping onboard computation affordable for autonomous UAV navigation, demonstrating that feature-based techniques still have a lot to offer in place recognition at extreme viewpoint changes.

To the best of our knowledge, the new synthetic datasets presented here are the first to completely isolate the problem of viewpoint changes for place recognition, closing a crucial gap in the literature. To facilitate further research on this topic, our datasets are publicly available.

Bibliography

[1] A. Alahi, R. Ortiz, and P. Vandergheynst. FREAK: Fast Retina Keypoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[2] H. Altwaijry, E. Trulls, J. Hays, P. Fua, and S. Belongie. Learning to match aerial images with deep attentive architectures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics (T-RO)*, 2008.

[4] R. Arandjelovic and A. Zisserman. All about vlad. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[5] R. Arandjelović and A. Zisserman. Dislocation: Scalable descriptor distinctiveness for location recognition. In *Asian Conference on Computer Vision*, pages 188–204. Springer, 2014.

[6] R. Arandjelović and A. Zisserman. Visual vocabulary with a semantic twist. In *Asian Conference on Computer Vision*, pages 178–195. Springer, 2014.

[7] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 2016.

[8] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera. Fusion and binarization of cnn features for robust topological localization across seasons. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4656–4663. IEEE, 2016.

[9] N. Atanasov, M. Zhu, K. Daniilidis, and G. J. Pappas. Localization from semantic observations via the matrix permanent. *The International Journal of Robotics Research*, 35(1-3):73–99, 2016.

[10] G. Baatz, K. Köser, D. Chen, R. Grzeszczuk, and M. Pollefeys. Handling urban location recognition as a 2D homothetic problem. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.

[11] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2006.

[12] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2008.

[13] B. Bescos, J. M. Fácil, J. Civera, and J. Neira. Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters*, 3(4):4076–4083, 2018.

[14] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart. Robust visual inertial odometry using a direct ekf-based approach. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 298–304. IEEE, 2015.

[15] E. Brachmann and C. Rother. Learning less is more-6d camera localization via 3d surface regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4654–4662, 2018.

[16] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2018.

[17] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 2016.

[18] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: Binary Robust Independent Elementary Features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.

[19] S. Cao and N. Snavely. Graph-based discriminative learning for location recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[20] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, et al. City-scale landmark identification on mobile devices. In *CVPR 2011*, pages 737–744. IEEE, 2011.

[21] Z. Chen, O. Lam, A. Jacobson, and M. Milford. Convolutional neural network-based place recognition. In *Australasian Conference on Robotics and Automation*, volume 2, page 4, 2014.

[22] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford. Deep learning features at scale for visual place recognition. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3223–3230. IEEE, 2017.

[23] Z. Chen, F. Maffra, I. Sa, and M. Chli. Only look once, mining distinctive landmarks from convnet for visual place recognition. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9–16. IEEE, 2017.

[24] O. Chum, A. Mikulik, M. Perdoch, and J. Matas. Total recall ii: Query expansion revisited. In *CVPR 2011*, pages 889–896. IEEE, 2011.

[25] T. Cieslewski, E. Stumm, A. Gawel, M. Bosse, S. Lynen, and R. Siegwart. Point cloud descriptors for place recognition using sparse visual information. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[26] A. Cohen, J. L. Schönberger, P. Speciale, T. Sattler, J.-M. Frahm, and M. Pollefeys. Indoor-outdoor 3d reconstruction alignment. In *European Conference on Computer Vision*, pages 285–300. Springer, 2016.

[27] M. Cummins and P. Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *International Journal of Robotics Research (IJRR)*, 2008.

[28] M. Cummins and P. Newman. Appearance-only SLAM at large scale with FAB-MAP 2.0. *International Journal of Robotics Research (IJRR)*, 2011.

[29] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005.

[30] R. Dubé, D. Dugas, E. Stumm, J. Nieto, R. Siegwart, and C. Cadena. Segmatch: Segment based place recognition in 3D point clouds. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[31] R. Dubé, A. Cramariuc, D. Dugas, J. Nieto, R. Siegwart, and C. Cadena. SegMap: 3D Segment Mapping using Data-Driven Descriptors. *Robotics: Science and Systems Online Proceedings*, 14, 2018.

[32] R. Dubé, M. G. Gollub, H. Sommer, I. Gilitschenski, R. Siegwart, C. Cadena, and J. Nieto. Incremental-segment-based localization in 3-d point clouds. *IEEE Robotics and Automation Letters*, 2018.

[33] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[34] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza. Svo: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 33(2):249–265, 2017.

[35] F. Furrer, M. Burri, M. Achtelik, and R. Siegwart. *Robot Operating System (ROS): The Complete Reference (Vol.1)*, chapter RotorS—A Modular Gazebo MAV Simulator Framework. 2016.

[36] D. Galvez-Lopez and J. D. Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics (T-RO)*, 28 (5):1188–1197, 2012.

[37] S. Garg, A. Jacobson, S. Kumar, and M. Milford. Improving condition-and environment-invariant place recognition with semantic place. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, 2017.

[38] S. Garg, N. Suenderhauf, and M. Milford. Semantic–geometric visual place recognition: a new perspective for reconciling opposing views. *The International Journal of Robotics Research*, page 0278364919839761, 2019.

[39] A. Gawel. *Mapping and Localization with Heterogeneous Robots*. PhD thesis, ETH Zurich, 2018.

[40] M. Gehrig, E. Stumm, T. Hinzmann, and R. Siegwart. Visual place recognition with probabilistic voting. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[41] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[42] P. Gronát, G. Obozinski, J. Sivic, and T. Pajdla. Learning and calibrating per-location classifiers for visual place recognition. *International Journal of Computer Vision (IJCV)*, 2016.

[43] B. K. Horn. Closed-form solution of absolute orientation using unit quaternions. *JOSA A*, 4(4):629–642, 1987.

[44] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1176. IEEE, 2009.

[45] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *International journal of computer vision*, 87(3):316–336, 2010.

[46] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, 2012.

[47] M. Kalantari, F. Jung, N. Paparoditis, and J.-P. Guédon. Robust and automatic vanishing points detection with their uncertainties from a single uncalibrated image, by planes extraction on the unit sphere. In *ISPRS2008*, pages 203–208, 2008.

[48] T. Kanji. Self-localization from images with small overlap. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4497–4504. IEEE, 2016.

[49] M. Keller, Z. Chen, F. Maffra, P. Schmuck, and M. Chli. Learning deep descriptors with scale-aware triplet networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2762–2770, 2018.

[50] A. Kendall, M. Grimes, and R. Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015.

[51] A. Kendall, R. Cipolla, et al. Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[52] G. Klein and D. W. Murray. Parallel tracking and mapping for small AR workspaces. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2007.

[53] L. Kneip, D. Scaramuzza, and R. Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[54] L. Kneip, D. Scaramuzza, and R. Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. 2011.

[55] N. Kobyshev, H. Riemenschneider, and L. Van Gool. Matching features correctly through semantic understanding. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 472–479. IEEE, 2014.

[56] A. Kristoffersson, S. Coradeschi, and A. Loutfi. A review of mobile robotic telepresence. *Advances in Human-Computer Interaction*, 2013:3, 2013.

[57] J. Ku, A. Harakeh, and S. L. Waslander. In defense of classical image processing: Fast depth completion on the cpu. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 16–22. IEEE, 2018.

[58] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In *ICCV*, volume 9, pages 2130–2137, 2009.

[59] S. Leutenegger, M. Chli, and R. Y. Siegwart. BRISK: Binary Robust Invariant Scalable Keypoints. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[60] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, and R. Siegwart. Keyframe-based Visual-Inertial SLAM using Nonlinear Optimization. In *Proceedings of Robotics: Science and Systems (RSS)*, 2013.

[61] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *International Journal of Robotics Research (IJRR)*, 2015.

[62] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

[63] S. Lieberknecht, S. Benhimane, P. Meier, and N. Navab. Benchmarking template-based tracking algorithms. *Virtual Reality*, 2011.

[64] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[65] Y. Liu and H. Zhang. Visual loop closure detection with a compact image descriptor. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1051–1056. IEEE, 2012.

[66] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 2004.

[67] R. Maffei, V. A. Jorge, V. F. Rey, M. Kolberg, and E. Prestes. Fast monte carlo localization using spatial density information. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6352–6358. IEEE, 2015.

[68] F. Maffra, Z. Chen, and M. Chli. Loop-closure detection in urban scenes for autonomous robot navigation. In *3D Vision (3DV)*, 2017.

[69] F. Maffra, L. Teixeira, Z. Chen, and M. Chli. Loop-closure detection in urban scenes for autonomous robot navigation. In *2017 International Conference on 3D Vision (3DV)*, pages 356–364. IEEE, 2017.

[70] F. Maffra, Z. Chen, and M. Chli. Viewpoint-tolerant Place Recognition combining 2D and 3D information for UAV navigation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2542–2549. IEEE, 2018.

[71] F. Maffra, L. Teixeira, Z. Chen, and M. Chli. Real-Time Wide-Baseline Place Recognition Using Depth Completion. *IEEE Robotics and Automation Letters*, 4(2):1525–1532, 2019.

[72] A. L. Majdik, D. Verda, Y. Albers-Schoenberg, and D. Scaramuzza. Air-ground matching: Appearance-based gps-denied urban localization of micro aerial vehicles. *Journal of Field Robotics (JFR)*, 2015.

[73] F. Mal and S. Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[74] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 978-0-521-86571-5.

[75] M. J. Milford. Vision-based place recognition: how low can you go? *International Journal of Robotics Research (IJRR)*, 2013.

[76] J.-M. Morel and G. Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM journal on imaging sciences*, 2009.

[77] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics (T-RO)*, 31(5):1147–1163, 2015.

[78] A. C. Murillo and J. Kosecka. Experiments in place recognition using gist panoramas. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 2196–2203. IEEE, 2009.

[79] R. R. Murphy, S. Tadokoro, D. Nardi, A. Jacoff, P. Fiorini, H. Choset, and A. M. Erkmen. *Search and Rescue Robotics*, pages 1151–1173. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[80] T. Naseer and W. Burgard. Deep regression for monocular camera-based 6-dof global localization in outdoor environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1525–1530. IEEE, 2017.

[81] T. Naseer, B. Suger, M. Ruhnke, and W. Burgard. Vision-based markov localization across large perceptual changes. In *2015 European Conference on Mobile Robots (ECMR)*, pages 1–6. IEEE, 2015.

[82] J. Nikolic, J. Rehder, M. Burri, P. Gohl, S. Leutenegger, P. T. Furgale, and R. Siegwart. A synchronized visual-inertial sensor system with fpga pre-processing for accurate real-time slam. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[83] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[84] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[85] R. Paul and P. Newman. Fab-map 3d: Topological mapping with spatial and visual appearance. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2010.

[86] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[87] T. Qin, P. Li, and S. Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

[88] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys. DeepLiDAR: Deep Surface Normal Guided Depth Prediction for Outdoor Scene from Sparse LiDAR Data and Single Color Image. 2019.

[89] F. Radenović, G. Tolias, and O. Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[90] L. Riazuelo, L. Montano, and J. Montiel. Semantic visual slam in populated environments. In *2017 European Conference on Mobile Robots (ECMR)*, pages 1–7. IEEE, 2017.

[91] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2018.

[92] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT and SURF. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[93] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys. Large-scale location recognition and the geometric burstiness problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1582–1590, 2016.

[94] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T. Pajdla. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[95] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[96] P. Schmuck and M. Chli. Multi-uav collaborative monocular slam. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3863–3870. IEEE, 2017.

[97] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler. Semantic visual localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6896–6906, 2018.

[98] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[99] Y.-S. Shin, Y. S. Park, and A. Kim. Direct visual slam using sparse depth for camera-lidar system. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[100] R. Siegwart, I. Reza Nourbakhsh, and D. Scaramuzza. *Introduction to Autonomous Mobile Robots*. MIT Press, 2nd edition, 2011.

[101] G. Singh and J. Košecká. Semantically guided geo-location and modeling in urban environments. In *Large-Scale Visual Geo-Localization*, pages 101–120. Springer, 2016.

[102] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2003.

[103] E. Stumm, C. Mei, S. Lacroix, and M. Chli. Location graphs for visual place recognition. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[104] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford. On the performance of convnet features for place recognition. In *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2015)*, pages 4297–4304. IEEE, 2015.

[105] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. *Proceedings of Robotics: Science and Systems XII*, 2015.

[106] L. Teixeira and M. Chli. Real-time mesh-based scene estimation for aerial inspection. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4863–4869. IEEE, 2016.

[107] L. Teixeira, F. Maffra, M. Moos, and M. Chli. VI-RPE: Visual-Inertial Relative Pose Estimation for Aerial Vehicles. *IEEE Robotics and Automation Letters*, 3(4):2770–2777, 2018.

[108] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl. Semantic match consistency for long-term visual localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 383–399, 2018.

[109] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool. Sparse and noisy lidar completion with rgb guidance and uncertainty. 2019.

[110] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-based localization using lstms for structured feature correlation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 627–637, 2017.

[111] S. Weiss, M. W. Achtelik, S. Lynen, M. C. Achtelik, L. Kneip, M. Chli, and R. Siegwart. Monocular Vision for Long-term MAV Navigation: A Compendium. *Journal of Field Robotics (JFR)*, 2013.

[112] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger. Mid-fusion: Octree-based object-level multi-instance dynamic slam. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[113] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, K. Alexis, and K. McDonald-Maier. Are state-of-the-art visual place recognition techniques any good for aerial robotics? In *IEEE International Conference on Robotics and Automation (ICRA), Workshop on Aerial Robotics*, 2019.

[114] B. Zeisl, T. Sattler, and M. Pollefeys. Camera pose voting for large-scale image-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2704–2712, 2015.