6-DoF Semantics-Aware Condition- and Viewpoint-Invariant Visual SLAM

Sourav Garg

Confirmation of Candidature

***Principal Supervisor*:**
Associate Professor Michael Milford

***Associate Supervisor*:**
Associate Professor Ben Upcroft

Abstract

Visual SLAM is a robotic competency to simultaneously localize and map the environment using only visual cues. Although it has been vastly explored by researchers in past decades, it still lacks the required robustness to be put to use in any practical application. A visual SLAM system has three primary components: visual place recognition for localization, visual odometry for egomotion estimation and a representation map of the environment. In this report and later in the thesis, we will explore each of these components in detail and then comprehend it with another dimension of semantics.

The use of visual cues implies using robust representations of raw images to be able to correctly recognize places in the map that a robot revisits as well as to accurately estimate egomotion in order to create a map of environment. Such image representations need to be robust against variations in the viewpoint and environmental conditions or alternatively, be able to accordingly adapt to such variations. Further, the representation maps created by state-of-the-art methods are generally only geometric and researchers have only recently started adding semantics to the maps for obtaining interactive 3D reconstructions of environment. However, a semantic SLAM system would not just require semantic mapping, it should also be able to exploit semantic information for effective localization, and finally, an integrated SLAM framework which is able to generate better semantic labels as well.

We will particularly address the following research questions:

- How can we characterize a visual SLAM system against variations in environment and egomotion, and make it more robust for both visual place recognition and visual odometry?
- How can semantic information related to images add value to a visual SLAM system?
- Can we develop a general purpose 6-DoF semantics-aware SLAM system robust to condition and viewpoint variations.

The answers to these questions will finally allow us to build a 3D reconstruction of environment which is semantically interactive and can be used to demonstrate its use in plethora of practical applications that use mobile robots, like indoor monitoring, grasping, navigation etc.

Contents

1 Introduction **4**
- 1.1 Robotics . . . 4
- 1.2 SLAM: *Is it solved?* . . . 4
- 1.3 Visual SLAM . . . 4

2 Literature Review **5**
- 2.1 SLAM Overview . . . 5
- 2.2 Visual Place Recognition . . . 5
- 2.3 Visual Odometry . . . 6
- 2.4 Representation Map and 3D Reconstruction . . . 7
- 2.5 Semantics in visual SLAM . . . 8

3 Research Problem **9**
- 3.1 Research Gap . . . 9
- 3.2 Problem Statement . . . 9
- 3.3 Research Questions . . . 9

4 Research Plan **10**
- 4.1 Characterization of visual SLAM system . . . 10
  - 4.1.1 Visual Odometry and Place Recognition . . . 10
  - 4.1.2 Adapting to Environment . . . 11
  - 4.1.3 Egomotion . . . 11
- 4.2 Utilizing Semantic Information . . . 12
  - 4.2.1 Environment Semantics . . . 12
  - 4.2.2 Semantic Segmentation *Across* Places . . . 12
  - 4.2.3 Semantic Segmentation *Within* Places . . . 12
- 4.3 6-DoF Semantics-Aware SLAM . . . 13
  - 4.3.1 Condition-Invariant Relative Pose Estimation . . . 13
  - 4.3.2 Semantics for SLAM . . . 13
  - 4.3.3 SLAM for Semantics . . . 13
- 4.4 Demo: Semantically Interactive 3D Reconstruction . . . 13
- 4.5 Timeline . . . 14

5 Work Progress **14**
- 5.1 Performance Evaluation Using High Fidelity Simulation . . . 14
  - 5.1.1 Place Recognition - SeqSLAM . . . 15
  - 5.1.2 Visual SLAM - ORB-SLAM . . . 16
  - 5.1.3 Paper Accepted at IROS 2016 . . . 18
- 5.2 Place Recognition Using Semantics . . . 18
  - 5.2.1 Semantic Segmentation of Environment . . . 19
  - 5.2.2 Paper Submitted to IROS 2017 . . . 19
- 5.3 Motion Estimation under Unfavorable Conditions . . . 19
  - 5.3.1 Low Light and High Speed . . . 20
  - 5.3.2 Speed-Normalized Data Sampling . . . 20
  - 5.3.3 Planned Submission for ICRA 2018 . . . 22

6 Conclusion **22**

1 Introduction

1.1 Robotics

Robotics is a rapidly growing field with an increasing number of applications in industry, household and military. The ultimate goal of creating human-like, rather advanced, artificially intelligent, completely autonomous robots, capable of performing multiple complex tasks is yet a long way to go. However, gradual research and development within different aspects of robotics has enabled the use of robots in specific practical applications.

An interesting robotic competency, which will be explored in this report is *SLAM*, that is, Simultaneous Localization and Mapping. It is the capability of a mobile robot to create a map of its environment while also keeping track of its location within that map. Such an ability is really important for a mobile robot because the foremost requirement for accomplishing any complex task is *awareness*, that is, a robot being aware of its operating environment as well as itself within that environment. While a number of practical applications may use pre-built map of the environment, identifying changes within that map, which is often the case, would still require SLAM. Autonomous vehicles or ADAS systems, retail store and warehouse stock management, deep sea bed modeling, domestic services etc. are all the practical use cases of a SLAM based robotic solution.

1.2 SLAM: *Is it solved?*

This question here is not just asked generally by those who have heard of SLAM as a research term before, but many roboticists as well. Sebastian Thrun and Jose Neira in [1] answered to this in 2010, confirming that it is solved for static environments and the pertaining problems are: dynamic objects within the environment and semantic understanding of the environment. The interview also discussed the community's shift towards visual SLAM, given that the digital cameras are an affordable source of information-rich data. This defines the locus of our research which will deal with visual SLAM, its challenges related to dynamics of the environment as well as its semantic understanding.

1.3 Visual SLAM

Visual SLAM is a robotic competency to simultaneously localize and map the environment using only visual cues. These visual cues can come from a monocular camera, a stereo-rig or RGB-D sensors. The key components of a visual SLAM system are: *Visual Odometry*, *Place Recognition* and *Representation Map*. *Visual Odometry* is the capability of a mobile robot to estimate egomotion using only vision as input. *Place Reconstruction* means the ability to recognize a place when the robot revisits that place. *Representation Map* mainly implies the way of storing the visual information in form of a map of places or visual landmarks 1. We will explore visual SLAM and its components further in subsequent sections along with the corresponding literature in respective fields.

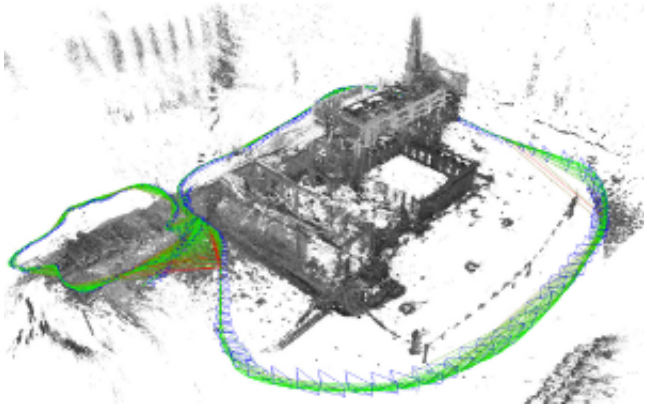


Figure 1: 3D reconstruction of environment using LSD-SLAM [2]

2 Literature Review

2.1 SLAM Overview

The most common and traditional approaches for SLAM use probabilistic methods to maintain both the pose of the robot and the location of the landmarks. This includes Extended Kalman Filter [3], Rao-Blackwellized Particle Filter [4], Expectation Maximization [5] etc. The second set of methods uses sparse graph of constraints and non-linear optimization for recovering map and camera poses as in [6, 7, 8]. The third and relatively newer paradigm, with respect to visual SLAM, includes Structure from Motion (SfM) [9] techniques with help of Bundle Adjustment [10], where the scene structure and camera motion, both are jointly recovered using optimization techniques [11]. These methods have been initially used for visual odometry derived from stereo [12, 13] or monocular vision [14, 15, 12, 11] and recently for visual SLAM [16, 17]. The fourth and the last category of SLAM solutions is inspired from biology [18] and builds on the cognitive abilities of animals to localize and map their environment, for example, rodents as shown in [19] and ants in [20].

2.2 Visual Place Recognition

A place is defined as a distinct location in the representation map of an environment. Visual places are described using the image features which can be broadly classified into two categories: **local** and **global** image features.

Local Image Features: The local image features have a corresponding location within the image associated with each feature and are generally described using a local image descriptor. The most common and efficient approaches for place recognition using point-based local features employ Bag of Words (BoW) approach [21, 22]. The image features like SIFT [23], SURF [24], BRIEF [25], ORB [26] etc. have been used in a BoW framework in FAB-MAP 2.0 [27], LSD-SLAM [2], ORB-SLAM [16] and others [28, 29]. There exist further extensions to these methods for high performance, for example, building an incremental visual words vocabulary on-line [30], using BoW-Pairs [31], adding more geometrical constraints between the words [32] etc.. These methods and in general, point-feature based methods, work remarkably well with viewpoint invariance and can be used for wide-baseline stereo matching efficiently. However, they are susceptible to extreme variations in the appearance of the environment as well as feature redundancy that arises due to repeated, bland or texture-less environment. Such conditions lead to, what is know as, **perceptual aliasing** as shown in Figure:2, that is, two different places (or features) appear to be similar. The variations in appearance, as mentioned, can also arise due to environmental conditions like different time of day, weather or season, or due to shadow, motion blur or varying illumination etc. for example in scenarios as shown in Figure:3.

Figure 2: *Perceptual Aliasing* between images in first and second row shows perceptual similarity between these places, though belonging to different physical locations in the map. [33]

(a) (b)

Figure 3: Vast appearance changes in the environmental conditions can cause same place to appear very different due to variation in (a) seasons [34], (b) weather and time of day [35].

Global Image Features: The second method of describing places uses global image descriptors. Such features include HoG [36], PCA [37], WI-SURF [38], BRIEF-GIST [39] etc. that describe the whole image with a single feature. The methods like SeqSLAM [35], LSD-SLAM [2], Photometric Bundle Adjustment [40] and others [36, 41] have proven the successful use of whole-image matching for place recognition. These image matching approaches fall in the category of **direct** or **featureless** methods. Unlike local feature matching, these methods can be used efficiently for only narrow-baseline matching and therefore, are susceptible to large variations in the viewpoint of the scene. Though, they can guarantee sub-pixel accuracy in matching, the computation complexity goes upward, especially when an image pyramid is used to tackle viewpoint variations. As opposed to local point features, these techniques work well with vast changes in appearance and conditions of the environment.

Hybrid and Deep-Learned: There are some of the place recognition methods that make use of a combination of local and global image descriptors. Such approaches have been shown to work well with both condition and viewpoint variations as in [42], [43] and [44], but they either require site-specific training, are not scalable or lack generality. Also, the use of deep-learned features as in [45, 46] has not been shown to be helpful for visual SLAM beyond 1D route traversals.

Summary: The local image features suffer from appearance and condition variations and global image features are susceptible to viewpoint variations. The condition- and viewpoint-invariant place representation using deep-learned features looks promising, but none of such representations have been ever used to estimate relative 3D pose between the pair of matching places, also termed as **loop closures**. Hence, it remains a challenge to develop a better place representation or use the existing ones in such a way that they can be integrated in a 6-DoF visual SLAM system. Such a system can then be used to create 3D maps, therefore, leading to numerous opportunities of interacting with such a map in order to develop practical SLAM applications.

2.3 Visual Odometry

Motion Estimation: A visual SLAM system needs a source of motion information in order to build a map with metric relationship among places that it represents. There are several methods that use different types of sensors to get this motion information, for example, IMU (Inertial Measurement Unit) in [47, 48], GPS (Global Positioning System) in [49, 50], LASER in [51], [52], robot-wheel odometry in [53], OBD (On-Board Diagnostics) data in [54, 55], Hall-Effect sensor in [55], DVL (Doppler Velocity Log) dead reckoning in [56] etc. On the other hand, a purely vision based SLAM system only uses visual cues to estimate egomotion. This is termed as *Visual Odometry* and is used quite often in visual SLAM systems. Although the use of dedicated sensors provides an accurate and regular estimation of motion as compared to visual odometry, it is still preferable because it eliminates the use of an extra sensor and the overhead of interfacing and sensor fusion.

Feature Tracking or Direct Matching: Most of the methods used for visual odometry are based on local features which are either detected as *corners* like Harris [57], FAST [58] etc., *blobs* like SIFT [23],

SURF [24], CenSurE [59] etc. or *edgelets* [60]. All these methods are based on local feature extraction and matching, and are therefore efficient in wide-baseline matching leading to its high viewpoint-invariance. On the other hand, there are methods that use direct whole image matching for semi-dense [61, 62] or dense [63, 64] 3D reconstruction of environment. These methods are generally more suitable for narrow-baseline matching and are capable of sub-pixel accuracy. Though these methods cannot handle large viewpoint variations, they certainly perform better than local features based methods when it comes to robustness towards appearance variations or repeated and texture-less environments. The authors in [65] have also shown robustness towards low-light environment using direct visual odometry.

Deep VO: The authors in [66] explicitly model visual odometry problem in a deep learning framework using stereo image pairs. The other deep-learned models that estimate pose between images and are closely related to visual odometry are Posenet [67], Deep Tracking [68] and Sfm-Net [69]. The use of deep learning for motion estimation using visual cues is in its stage of infancy as all these methods are very recent and either require range information, environment-specific training or extensive engineering and have not be shown to be scalable similar to what a visual SLAM system would require. Moreover, all these systems use images under ideal environmental conditions and are likely to fail in adverse conditions because lack of generality in image data while training.

Summary: The visual odometry approaches suffer from similar issues like visual place recognition. The choice between local and global image features has trade-off between viewpoint- and condition-invariance of image representation. Moreover, most of the visual odometry methods require parameter tuning according to the environment and are prone to failures due to rapid camera motion, low-light environment and radiometric variations. The deep-learning based solutions haven't yet been convincingly able to offer better solution either. The failure of visual odometry is catastrophic for a visual SLAM system if it solely relies on it for motion estimation, hence it remains a challenge to develop robust motion estimation solution using visual odometry or some hybrid approach.

2.4 Representation Map and 3D Reconstruction

Types of Maps: Mapping is an integral part of a SLAM system. An environment representation map acts as the reference database for a place recognition system, unless it is based on pure image retrieval, where actual location of a place does not really matter as in [27] and [21]. For all other cases, these representation maps can be of very different nature in terms of how they represent the environment. They can be classified into three categories: **topological**, **metric** and **topometric**. A *topological* map has only a relative arrangement of places with respect to each other, for example, in [70, 35]. The places in such a map are merely set of points on a 1D route traversal and are connected with the logic of their order of occurrence. On the other hand, a *metric* map is built using physical location information of a place as in [71, 15]. Here, the places are connected using metric distance (up to scale) and closely represent the real map of the environment. The third category of representation is a combination of both these approaches that uses topological-metric or *topometric* map as in [72, 53, 73]. These representations help in handling large-scale maps or defining a logical layer over the metric maps for an easier interface. For example, semantic maps of environment [74, 75] can enable higher-level reasoning and planning.

3D Maps: The representation maps often describe places in 3D using range information either from LASER [52, 76], RGBD sensors [77, 78] or stereo images [32, 65, 79]. Though explicitly sensing 3D information makes the problem simpler, it comes with other overheads of using more sensors and complex sensor fusion techniques. The 3D structure information of places in a map can also be extracted from moving cameras using Structure from Motion (SfM) [9] and local Bundle Adjustment [10, 80] techniques. Most of these methods use local image feature for sparse [79, 16], semi-dense [2, 81] or dense [63] reconstruction of environment. Some preliminary work has been done in [40] that performs featureless, and therefore direct photometric bundle adjustment.

Summary: The state-of-the-art visual SLAM methods use 3D maps demonstrating dense reconstruction of the environment. Such a representation is indeed important for enabling robot's interaction with the environment similar to humans. However, it still lacks semantics which are important to impart meaning to the pixels or segments in a 3D reconstruction for an effective interaction. The use of semantics in visual SLAM has been explored in next section.

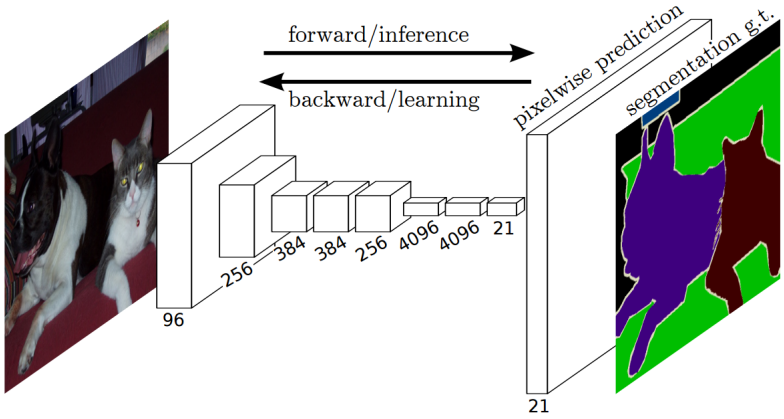


Figure 4: Dense Semantic Object Segmentation as proposed in [96] using Fully Convolutional Networks.

2.5 Semantics in visual SLAM

Semantics Within and Across Places: The use of semantics in a visual SLAM system has been explored from two different perspectives. The first one explores the meaning **within a place** by semantically labeling the objects [82], patches [83], pixels [75], superpixels [84] etc. within the image, often in conjugation with range information. The second perspective is a broader view of these places that performs categorization **across the places** within a map, usually on a larger scale, for example, in [85, 74, 86]. However, in both the scenarios, use of semantics enables higher-level reasoning and interactive modeling of the environment.

Object Semantics: There have been several attempts towards developing a semantic SLAM system based on objects encountered in the environment. The use of pre-trained 3D object models in [87] and semantic object-class segmentation in [88], is one way to incorporate object semantics in an off-the-shelf visual SLAM system. Another set of approaches include SLAM frameworks where objects are integral part of the optimization equation as in [89, 90, 91, 92]. The authors in [93] combine object recognition and semantic image segmentation for dense semantic SLAM. While all these approaches are **object-centric**, the authors in [75, 94] also consider semantics of structural elements like wall, floor, ceiling etc. to semantically characterize the visual SLAM system. Furthermore, the use of deep learning frameworks for more powerful semantic object recognition has also helped in sparse [95] or dense [96, 97] object-centric semantic segmentation as shown in 4.

Place Semantics: Although, *object-centric* approaches look very promising, they usually neglect the importance of visual scenes that do not really have any particular object in focus, for example, a mountain range, crop-field, train-station etc. The need of semantics for such a **place-centric** scene has been emphasized in [98], where the authors train a deep-convolutional network on *place-centric* images to obtain semantic place categories. Similarly, authors in [99] use a plethora of semantic scene attributes to label different places. Furthermore, authors in [100] use transient semantic attributes, like time of day, weather, season etc. to characterize places.

Summary: The use of semantics in most of the research work is limited by the type of semantics and the extent to which they are integrated within a visual SLAM framework. The semantics related to objects as opposed to places and those focusing on structural elements as opposed to transient attributes, limit the generality and scalability of their use. Also, very few of the proposed methods utilize semantics to an extent where both localization and mapping can benefit from it. Further, there are even fewer such methods that have demonstrated use of SLAM framework to improve semantic labeling of either objects or places. Hence, a holistic approach taking these challenges and limitations into account is currently needed to develop a **Semantic SLAM** system.

3 Research Problem

3.1 Research Gap

The research gap following the literature review can be summarized in following points:

- It remains a challenge to develop a robust place recognition system that can be seamlessly integrated into a 6-DoF visual SLAM system.
- Visual Odometry solutions are brittle to variations in environmental conditions, rapid camera motion and radiometric variations.
- The failure of visual odometry is catastrophic for visual SLAM system, hence a hybrid approach towards motion estimation is a must.
- The geometric maps created by state-of-the-art methods lack semantics and therefore hinder effective interaction with the reconstructed environment.
- The use of semantics in visual SLAM is limited to either localization or mapping and lacks a holistic approach from which potentially both, semantic labeling and visual SLAM, can benefit.

3.2 Problem Statement

The overarching research problem following the literature review and the identified research gap can be stated as:
How can we develop a general purpose 6-DoF semantics-aware visual SLAM system which is condition- and viewpoint-invariant and adapts to changes in environment as well as egomotion?

3.3 Research Questions

In order to address the problem as described above and to enable such a SLAM system, we need to answer the following questions:

- How can we characterize a visual SLAM system against variations in environment and egomotion, and make it more robust for both visual place recognition and visual odometry?
  - How can we characterize the place recognition, visual odometry and visual SLAM algorithms for robustness against variations in camera viewpoint and environmental conditions?
  - How can we characterize the environment so as to enable adaptive behavior in robots by dynamically determining the changes in the environment?
  - How can we estimate the camera motion under extreme environmental conditions and rapid camera movements?
- How can we use semantic information related to images in order to increase the robustness of visual place recognition and SLAM methods?
  - How can we encode environmental conditions information into semantic labels within and across the images in an image sequence?
  - Can we use semantic image labels to categorize the environment into meaningful segments for adaptive robot behavior?
  - Can we learn image pixel-level semantic mask for robust place recognition under extreme environmental conditions?
- How can we develop a 6-DoF *semantics-aware* SLAM system which is robust towards condition and viewpoint variations?
  - How can we estimate 3D relative pose between images with vast appearance variations?
  - What are the different ways in which semantics can help 6-DoF visual SLAM system for both localization and semantic mapping?
  - How can we use the visual SLAM system to perform better semantic labeling and learn new semantic classes online?
  - Can we use our system to demonstrate an application that semantically interacts with the 3D reconstruction of environment?

4 Research Plan

4.1 Characterization of visual SLAM system

The visual SLAM system or any other similar competency developed as a robotic application can broadly be understood to comprise these three general components:

1. *Environment* where the robot operates or learns about its surroundings.
2. *Robot and Sensors* for interacting with the environment and enabling sensing and responding in some form.
3. *Algorithm Interface* is what enables the interaction between the robot and the environment and is designed specific to the application.

In order to develop a well-performing robotic application, it is important to understand the aforementioned components and their characteristics that significantly impact the performance. Therefore, it remains one of the research questions to identify such characteristics specific to the application and appropriately model them.

4.1.1 Visual Odometry and Place Recognition

Visual Simultaneous Localization and Mapping constitutes: *Visual Odometry*, *Visual Place Recognition* and *Mapping* of environment. The state-of-the-art visual SLAM systems or any of its individual constituents are often developed based on certain assumptions with respect to its environment, its sensors or the operating characteristics. Though it is not possible to estimate all the possibilities of operating characteristics beforehand, it is vital to understand those significantly impacting the performance.

SLAM systems at disposal: We plan to explore the state-of-the-art methods like SeqSLAM [35], ORB-SLAM [16], LSD-SLAM [2] etc. in order to understand their certain characteristics, for example, invariance to camera viewpoint and environmental conditions - both lie at the core of a general purpose visual SLAM system.

Data Gathering: The simultaneous condition- and viewpoint-invariance of a 3D visual SLAM system has yet not been achieved due to challenges involved. Therefore, it is important to understand the performance sensitivity of the existing algorithms towards these variations. Such an analysis will also require the appropriate data to conduct tests. However, the availability of real or simulated data corresponding to condition and viewpoint variations is limited. This is mainly because it is difficult, time-consuming and expensive to collect real data with a sound ground-truth generation system especially for condition variations - which essentially means traversing places at different times of day, in different weather conditions and across seasons. Similarly, capturing images of a place for all possible camera viewpoints is not always feasible. It seems relatively easier to do the same in simulation, given there exists a 3D model of a city or alike with sufficient rendering, so that condition- and viewpoint-varied traverses could be infinitely generated, but such a ready-to-use model does not exist and needs a lots of effort.

Simulation: We plan to collaborate with peers to simulate a city-like environment in 3D and obtain repeated traverses with varying viewpoint and conditions of the environment. Along with that, we will also make use of some of the existing and newly collected real world data which may not be sufficiently large as compared to the simulated one, but be rich enough with respect to the variations in the environment such that the performance curves so obtained will look similar to those using simulated world. The choice of relying on simulation than real world for characterizing the visual SLAM systems is merely due to large amount of effort required in the latter and collaboration opportunities at disposal for the former.

Conclusion: The characterization in this form will immensely help in performance benchmarking using simulated data for different robotic applications. This will also enable the enhanced understanding of parameters of the system that particularly drive its performance for the variations induced in the test data.

4.1.2 Adapting to Environment

Motivation: The environment plays a key role in determining the performance of any robotic application because it is often the case that many of the applications are designed specific to certain environment types and therefore are brittle towards significant variations in the environmental settings. In context of visual place recognition and visual odometry, the variations in environment are induced mainly due to *scene structure* and *transient conditions*. The scene structure is actually the basis for differentiating places from each other, but frequent transitions within different environments call for different parameter settings for improved performance, for example, urban canyons versus highways or forests, or outdoor vs indoor etc. On the other hand, transient conditions like time of day, weather and season, though representing the same place, remarkably change the visual appearance of the scene. Therefore, it is necessary to understand the environment and allow possible adaptive robot behavior for variations in the environment.

Recognizing places in varying environment: The condition-invariant place recognition methods such as SeqSLAM [35], SMART [55], and others [36, 44, 101] have been shown to able to recognize places under the influence of extreme variations in conditions of the environment. These variations are mainly *global* which means that they are static with respect to space but not time, for example, change in season, weather or time of day will only affect the appearance of a particular place in the environment when it is visited after a certain interval of time. On the other hand, the spatial neighborhood of that place will be similarly affected by the changes in the environmental conditions.

We are rather interested in local variations in the environment where global conditions changes may or may not occur. Examples of such local variations are mainly related to transitioning from outdoor to indoor environments, texture-less to cluttered scenes, urban canyons to forest roads etc. The seamless transition between such environmental settings is only possible by segmenting the environment into different chunks based on their appearance attributes.

Segmenting the environment: The place recognition methods generally employ a measure of matching places which can be either whole-image based as in SeqSLAM [35] or point features based collective score as in FAB-MAP [102]. These place matching scores effectively discriminate places from each other and hence could also be used to create segments based on these scores. The idea is to build a self-similarity matrix for a given dataset and then statistically perform segmentation or data clustering based on the affinity scores. The segments so formed can then be used to define neighborhood region for a place such that place recognition performance may be improved based on collective matching of different segments. This would also require us to collect datasets which possess transitions from one type of environment into the other quite often.

Conclusion: The characterization of environment with respect to local variations in overall appearance of environment will help in dynamically adapting the robot behavior accordingly. For example, place matching and motion estimation can be improved by tuning the system parameters in order to accommodate the variations in environment.

4.1.3 Egomotion

Motivation: The egomotion estimation is an important competency for robotic tasks that involve movement. Visual Odometry is the means of estimating egomotion using visual cues. The state-of-the-art visual odometry solutions are often prone to failures because of fast camera motion, motion blur, photometric and radiometric changes in appearance of environment etc. The failure of visual odometry is catastrophic for a visual SLAM system as the robot immediately loses its localization information and cannot relate current motion estimates with those collected earlier. The camera motion also plays a significant role in place recognition especially when repeated traverses of an environment exhibit different motion pattern, which makes it harder to recognize places. This is because a fixed-size temporal neighborhood window centered at a place will no longer contain similar places in different traverses.

Speed Normalization: We will look into the scope of improving place recognition using visual odometry. Most of the practical visual place recognition applications for example, an autonomous vehicle, do not exhibit uniform motion during the journey. The motion estimation can help in speed normalizing the collected imagery, so that the places are separated by a constant physical distance instead of constant number of frames.

Motion Estimation: We also plan to develop a motion estimator or predictor as a backup switch for state-of-the-art visual odometry methods such that it provides some source of information to roughly localize the robot even if such an estimate is less accurate or less precise. It is important because such estimates can be improved at a later stage by either using visual place recognition or improved visual odometry.

4.2 Utilizing Semantic Information

The use of semantic information has recently become easier because of deep-learned classifiers and regressors [95, 98] that can be trained on very large image data to precisely predict the semantics of the test image. These semantics can be used to achieve a meaningful interpretation of places and maps used in a visual SLAM system. We plan to make use of this semantic information to improve visual place recognition and visual SLAM system in different ways as described in following subsections.

4.2.1 Environment Semantics

One of the simplest use of semantics seems to be imparting meaning to the environment where robot operates in form of a semantic map. The place categorization [98] and scene attributes detection [99] using semantics has not been explored within a place recognition framework so far. Unfortunately, these semantic classifiers do not provide sufficient information regarding environmental conditions like time of day, season or weather etc., but mostly characterize the structure of the scene leading to particular place category labels. These transient conditions of environment have been semantically explored separately in [100].

We plan to use the semantic attributes based on scene structure as well as transient conditions from different classifiers to create ground truth semantic labels of both types for a huge database. These labeled images will then be used to train (or retrain) a deep neural network to learn the semantics of images for both scene structure as well as transient conditions. This will make place categorization more robust and easy to incorporate in a place recognition system.

4.2.2 Semantic Segmentation *Across* Places

The segmentation of places in an environment has been explored in earlier sections based on place matching scores. The use of semantics for segmenting the environment can be advantageous because firstly, deep-learned features are more discriminative [45] as compared to hand-crafted ones and secondly, it can help in filtering or shortlisting of the place matching candidates by performing a semantic matching first.

The semantic segmentation within an environment will be explored in context of place recognition. The filtering or shortlisting of place matching candidates leading to reduction in computation time for searching places will be benchmarked for large databases. The performance improvement in place recognition using place matching score as described earlier and using semantics will be also be compared.

4.2.3 Semantic Segmentation *Within* Places

The semantic segmentation of an image has been vastly explored in literature, but depending on the target application for such segmentation, most of them are not relevant for a visual SLAM system. The advances in semantic segmentation based on object recognition, whether sparse [95] or dense [96], are useful in visual SLAM because it enables resolving of each individual entity within an image. Most of these methods use deep convolutional networks and will be a suitable choice for this thesis as well. The bottleneck that can be foreseen is the lack of generality with respect to the objects that are recognized.

First, it does not cover each and every object in the world, which is obviously not an easy task, hence we will need a hybrid approach that also learns new object classes on-line.

Second, object recognition focuses more on indoor environments where most of these objects are found. It is quite common to encounter images that don't have any "object", especially in outdoor scenarios. The way objects are defined usually for object recognition tasks do not always consider wall, ceiling or floor like entities as a separate object, rather it is considered as background and the focus is on the object categories defined for the problem. This shortcoming has been fulfilled separately by some authors by proposing semantic place categorization that emphasizes the concept of a *place* as opposed to object-centric images.

Third, object recognition doesn't take into account the changing conditions of the environment. Even the semantic place categorization methods do not wholly represent a place with attributes that

correspond to environmental conditions. In order to incorporate semantics *within* an image into a visual SLAM system, it is important to have a semantic representation of an image that covers objects, place categories as well as the attributes corresponding to environmental conditions.

We plan to use different pre-trained networks that individually semantically characterize objects, places and environmental conditions, to cross-label the training images of each other. Then we will use one of the more robust networks to retrain all the re-labeled images. The trained network will hence have all the semantic labeling properties as desired for a robust visual SLAM system.

4.3 6-DoF Semantics-Aware SLAM

4.3.1 Condition-Invariant Relative Pose Estimation

We have explored visual place recognition systems that are condition-invariant [35, 44, 45], but they lack the ability to generate a 3D relative pose between the matching places, also termed as *loop closures*. In order to obtain a 6-DoF visual SLAM system, it is important to have the 3D pose estimation for condition-invariant representation of places.

We plan to first explore the possibility of using a condition- and viewpoint-invariant robust place representation and then enable it in some way to estimate relative pose. The deep-learned features are the most likely option for this case. The second option is to use a place representation which is known to efficiently work in the pose estimation framework of visual SLAM and then make it robust towards condition variations. The best choice for this seems to be the whole image feature descriptors which can be used in direct image matching framework for relative pose estimation with a descriptor-constancy assumption instead of brightness-constancy.

4.3.2 Semantics for SLAM

The use of semantics in visual SLAM framework has mainly focused on individual components of SLAM, that is, either localization or mapping. As also discussed in previous section, a holistic approach for semantic segmentation *within* places is required which will help in semantics based localization within the map. However, at the same time, we also want a semantic map that will cover all the semantics whether *within* or *across* the places. Hence, it is important to understand that role of semantics is with respect to both *space* and *time*, that actually echoes our concept of semantics *within* and *across* the places respectively as discussed earlier as well.

Our approach towards fusing semantic information in visual SLAM system would be to use pixel- or segment- (or patch-) level semantic descriptor for representation of a place. Each semantic segment can then be jointly optimized with camera poses in either graph optimization or local bundle adjustment framework. The semantic map can then be hierarchical, that is, a broader view of the map would be a topological semantic map containing different chunks of environment, and each such chunk can then be a metric map of semantic entities.

4.3.3 SLAM for Semantics

We have so far explored in lots of detail the use of semantics for visual SLAM. In order to have a complete semantic SLAM system, we also need to close the loop by creating the flow of information from SLAM towards semantics. For example, visual place recognition applied on conditionally varied places can help in applying same structural semantic labels to those places, but different corresponding label for environmental conditions.

We plan to use a set of semantic labels under the category of *unclassified* which will be applied to image segments that couldn't be labeled with sufficient confidence. This could happen either because of varying appearance of the segment or because of occlusion due to dynamic objects in environment. A new *unclassified* label will be generated whenever the unlabeled image segment is sufficiently different from the previous *unclassified* image segments. Now, these *unclassified* labels can be classified correctly whenever the place is revisited with the help of place recognition by helping in deciding which part of the place should actually become part of the map, and what part is actually occluded and should not be considered in map.

4.4 Demo: Semantically Interactive 3D Reconstruction

The final stage of this thesis will be a demonstration of semantically interactive 3D reconstruction of an environment using the proposed system. This setup can then be interacted for different tasks like

navigation, for example, “go to the kitchen”, “lead me to exit gate” etc. or inventory management, for example, number of chairs and tables in restaurant, stock inquiry in a retail store etc., or use it as a semantic interface for other robotic applications.

4.5 Timeline

The proposed timeline for the thesis is attached below:

Time Elapsed (in months for 3 yr study)	3	6	9	12	15	18	21	24	27	30	33	36
PhD Milestones												
Stage 2												
Confirmation												
Annual Progress												
Final Seminar												
Lodgement												
Coursework												
Advanced Information Retrieval Skills												
Research Ethics, Integrity and Safety	Quiz-1	Quiz-2										
Thesis Writing												
Title & Abstract												
Introduction												
Literature Review												
Characterizing Visual SLAM System With Egomotion, Camera Viewpoint and Environmental Conditions												
Utililizing Semantic Information												
3D Condition- and Viewpoint-Invariant Visual SLAM system												
Discussion												
Conclusion												
Research Process												
Accessing Literature												
Characterizing Visual SLAM Using High Fidelity Simulation												
Segmentation of Environment for Improved Place Recognition												
High-FPS Visual Odometry for Adverse Environmental Conditions												
Encoding Environmental Conditions in Image Semantics												
Semantic Segmentation of Environment												
Semantic Object Saliency for Condition-Invariant Place Recognition												
Condition-Invariant 3D Relative Pose Estimation for Loop Closures												
Use of Semantics Beyond 1D Route Traversals												
Demonstration												
Approvals and Applications												
Intellectual Property												
Ethics												
Health & safety												
Scholarships												
Write Up Scholarship												
Outputs												
Conference Papers		IROS'16	IROS'17		ICRA'18		IROS'18		ICRA'19		IROS'19	
Journals/Transactions						IJRR		IJRR				TRO
Deployable Solution												

5 Work Progress

5.1 Performance Evaluation Using High Fidelity Simulation

The use of high fidelity simulation is an attractive option for most of the researchers primarily because it gives access to infinite data generation which is very close to real world. A simulated environment enables detailed performance evaluation of algorithms which helps in understanding the data-dependent characteristics of the system and optimizing its parameters. We evaluated various robotic vision algorithms like place recognition, visual odometry, visual SLAM and object recognition to study their performance variations with respect to variations in the input data. The high fidelity simulation of a city-like environment was initially developed and the required datasets were then generated which were finally fed to different algorithms for performance evaluation.. Figure:5 shows a grid of sample images from a set of datasets used for evaluating place recognition performance. The images are taken from same place but

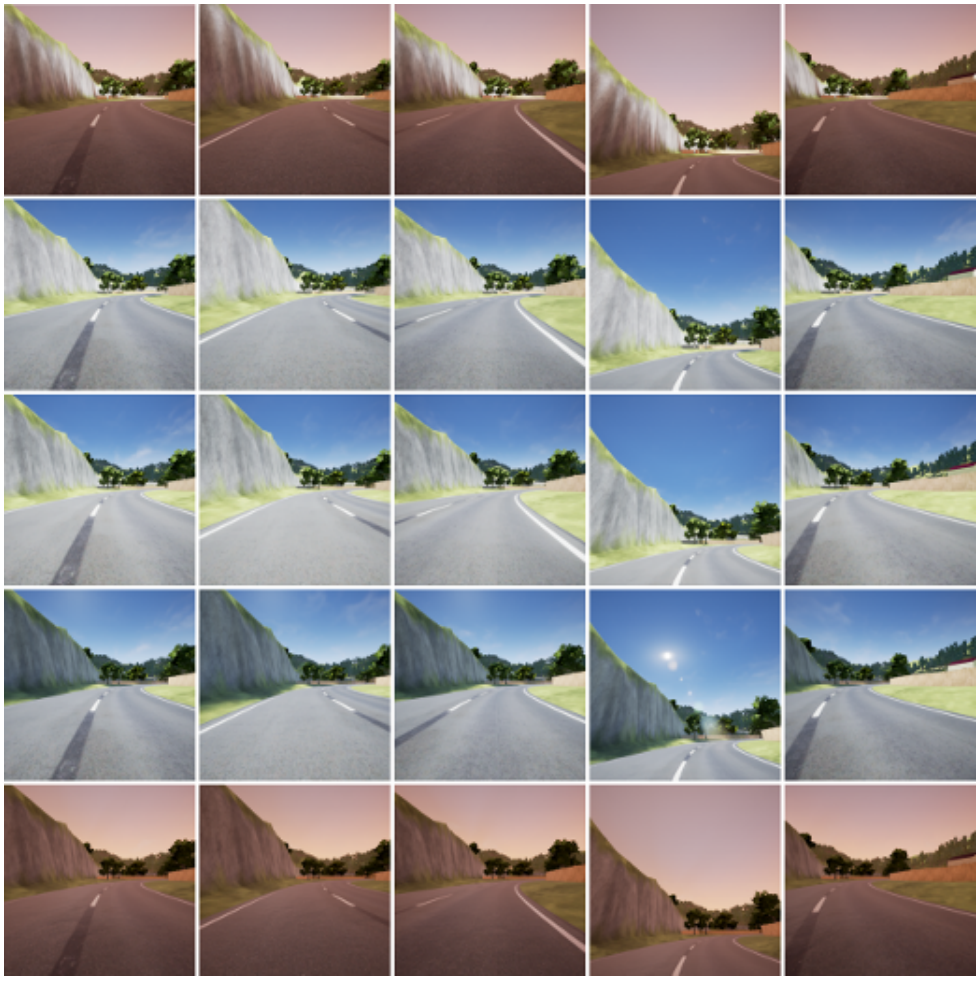


Figure 5: Sample images from the datasets used, all images are taken from the same place at different times of day and viewpoint variations. From left to right, viewpoints are the base unmodified path, offset left 2m, offset right 2m, angled up 30°, and angled right 30°.

at different time of day and varying camera viewpoints. The work was done with collaborative efforts from peers and the research components that are relevant to this report are described in subsequent subsections.

5.1.1 Place Recognition - SeqSLAM

The visual place recognition is the capability of a mobile robot to recognize a place during a revisit solely using visual cues. Thus, it requires repeated traverses of the environment with no restriction on viewpoint or environmental conditions like time of day, weather or season etc. during the subsequent visit. We evaluated performance of SeqSLAM - a condition-invariant place recognition algorithm using different datasets generated from the simulated environment as shown in Figure:5. These datasets vary with respect to 5 different times of day, that is, Dawn, Morning, Noon, Afternoon and Sunset as well as different camera viewpoints with variations in lateral offset, vertical orientation (pitch) and horizontal orientation (yaw).

The performance evaluation curves for SeqSLAM as shown in Figure:8(a) show a decrease in performance with extreme viewpoint variations as per the expectations and a constant high performance with varying conditions of environment.

We also performed some experiments on real world data along with its simulated version to observe the performance trends which were found to be similar. However, the results on simulated data were consistently better than the real world data which is also often the case with any simulation. Figure:6

Figure 6: Images for five different traversals of a street with different lateral offsets. This real world data (top) is used to compare the trend of performance change with change in lateral shifts as compared to the simulated data (bottom).

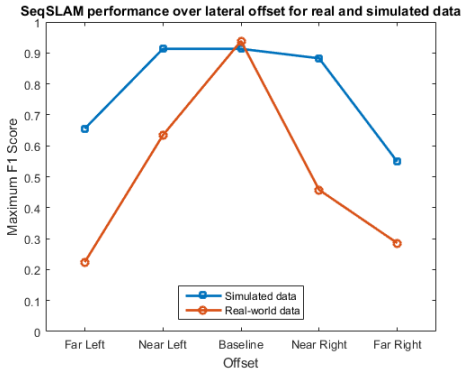


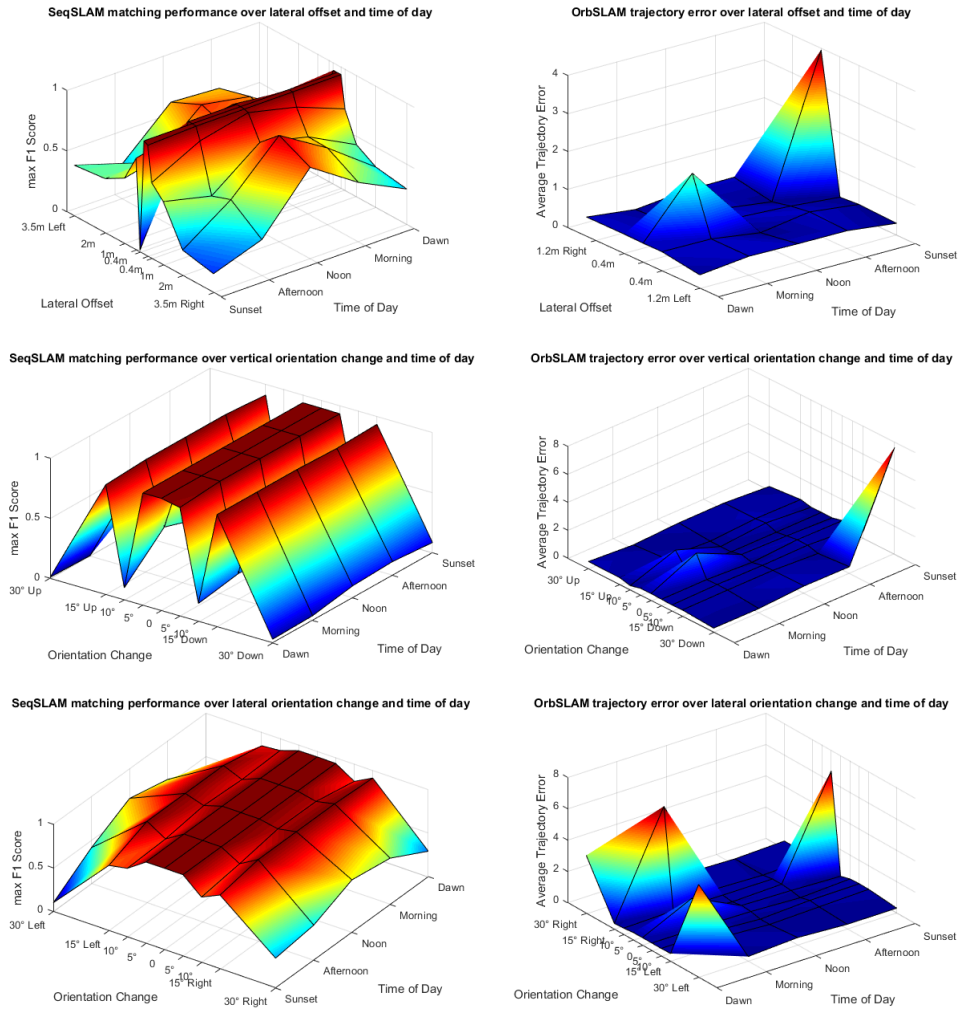
Figure 7: Comparison of SeqSLAM performance falloff between simulated and real-world environments as lateral offset increases. As is often the case, simulated performance is better in an absolute sense, but the trend is the same in both cases.

shows the real and simulated version of a street dataset and Figure:7 shows the corresponding performance comparison.

5.1.2 Visual SLAM - ORB-SLAM

ORB-SLAM is a state-of-the-art visual SLAM system proven to work well for both monocular and stereo or depth-based input. Similar to performance evaluation of visual place recognition, evaluating a visual SLAM system also requires repetitive traversing of an environment, but in a continuous loop. This is necessary to make sure that all the components of a SLAM system that is visual odometry, place recognition and mapping are effectively tested. Therefore, different traverses of the environment at different times of day and varying viewpoints as also described in previous subsection were always appended by a noon baseline dataset. Such an arrangement made sure that visual odometry component gets tested in the first part of the data along with its capability of continuity across the appended data in terms of local changes. The variation of first part across different datasets and second part within the dataset made sure that the place recognition component is effectively tested.

We used average trajectory error as a performance measure for generating the performance curves. It was observed that the performance of ORB-SLAM did not show any consistent pattern with variations to either viewpoint or environmental conditions as shown in Figure:8(b). In general, in a visual SLAM system, occurrence of even a single false loop closure or an instance of failure in visual odometry is catastrophic for the system. The former leads to an absurd trajectory with high error and the latter imlies an incomplete trajectory. Moreover, most of the visual SLAM algorithms use initialization methods that use random numbers, therefore, expecting a consistent pattern in a performance curve is probably not viable.



(a) SeqSLAM (b) ORB-SLAM

Figure 8: (a) SeqSLAM and (b) ORB-SLAM's performance for datasets varying with respect to lateral camera offset, vertical orientation and lateral orientation from top to bottom as well time of day. SeqSLAM shows expected behaviour mostly apart from few abrupt troughs. ORB-SLAM doesn't show any consistent performance change with respect to the imposed variations.

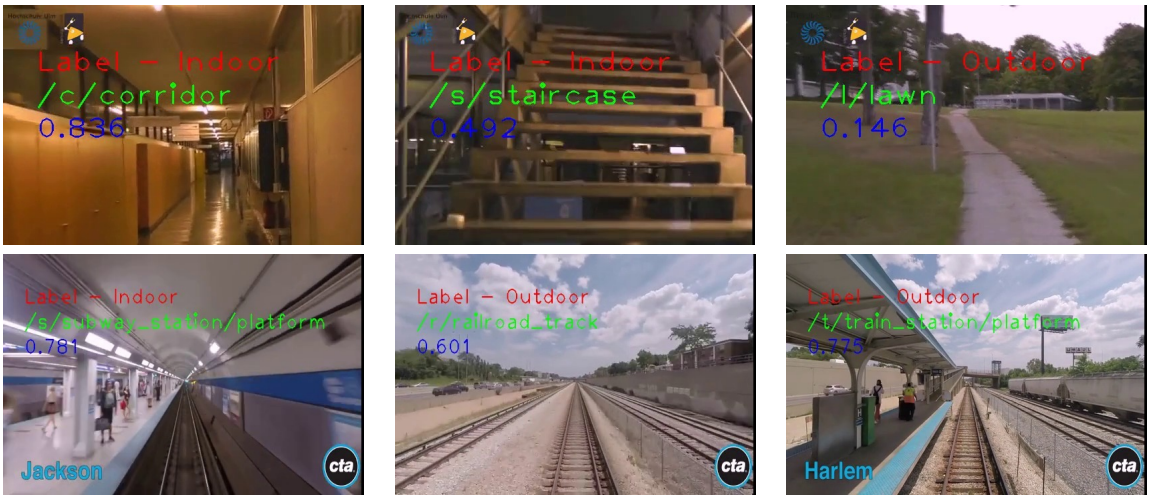


Figure 9: Semantically labelled images from Campus Indoor-Outdoor and CTA-Rail Dataset.

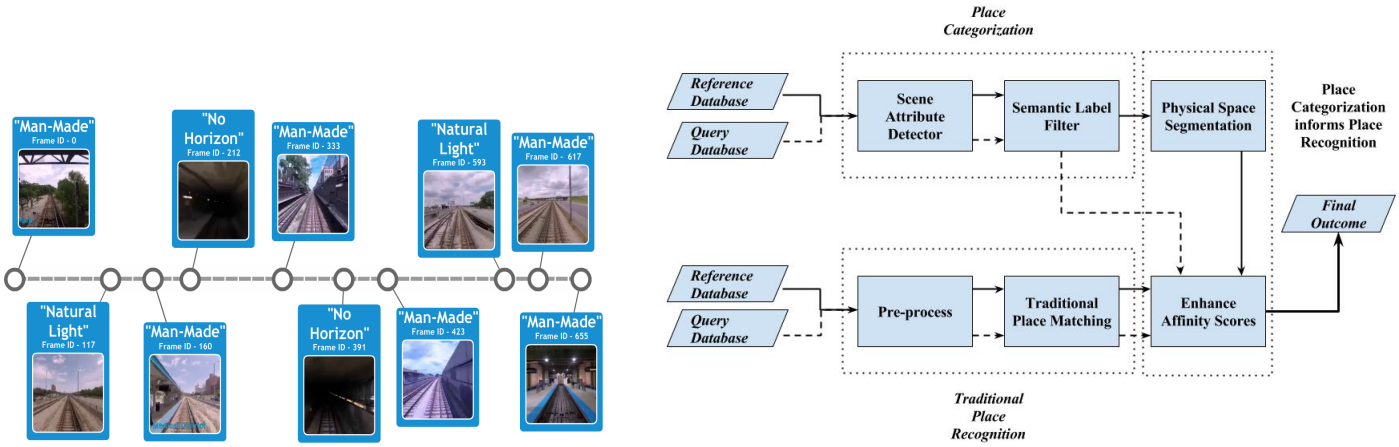


Figure 10: (a) shows semantic labels corresponding to different segments of the environment in CTA-Rail dataset. (b) shows a block diagram representing the flow of semantic information from the place categorization module to the place recognition module for improving place matching scores.

5.1.3 Paper Accepted at IROS 2016

The research work was accepted as a conference paper at IROS 2016 [103].

5.2 Place Recognition Using Semantics

The use of semantics with respect to places has become popular only recently with deep-learned CNN models after its successful experiments with object recognition. The semantics for places are generally described with respect to a single image, and provide labels about the scene structure, texture, environmental conditions etc as shown in Figure:9. Though, it enables place recognition in a course manner, recognizing *specific* places in an environment which generally have similar semantic labels in their neighborhood, is the *traditional* place recognition problem usually dealt in robotics. We explored the use of semantic labels for individual places within an image dataset in improving place recognition performance as described in subsections below.

5.2.1 Semantic Segmentation of Environment

One of the apparent ways of using semantic labels in place recognition is to coarsely filter the places that may match the query place. This will definitely reduce the computation time for place search within the database, and may also improve performance depending on the robustness of features used to semantically categorize places. Another way of utilizing the semantic information in context of place recognition is to exploit the temporal nature of place data encountered in both reference and query databases. A mobile robot traversing an environment is most likely to witness variations in its environment, both minor and major, especially in the applications of a visual SLAM system.

We developed a system that performed semantic place categorization on an incoming stream of images to form temporal semantic segments of places which are appear similar. Figure:10(a) shows a timeline of labels for different segments within one of the experimental datasets. The segmentation was then followed by a more *specific* (or *traditional*) place recognition system that utilized these temporal segments to define the neighborhood region around reference query places to bias their matching scores accordingly. Figure:10(b) shows a block diagram reprsenting the flow of semantic information from place categorization to place recognition.

The effectiveness of such a system is more pronounced for the image datasets having significant variations in the environment, for example, transiting between indoor and outdoor environments like a train running though tunnels as well as open areas (Figure-10(a)) or a mobile robot within a university campus (Figure:9); transiting between regions of varying illumination or texture like an urban canyon versus highway within a forest etc. We tested our system on a wide variety of datasets ranging from a 23 km train journey to short campus traverses, all exhibiting variations in environment with respect to scene structure, illumination, environmental conditions, texture etc. Figure:11 shows transition from outdoor to indoor of a house with varying illumnation. We observed a considerable performance gain using the proposed system for datasets with medium to extreme variations in their environment. Figure:12 shows comparative results between vanilla method and proposed approach for two of the datasets we experimented with.

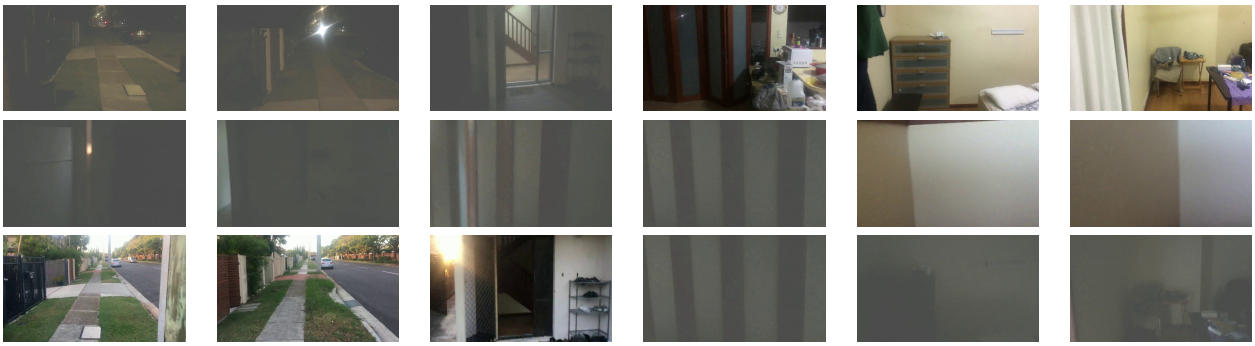


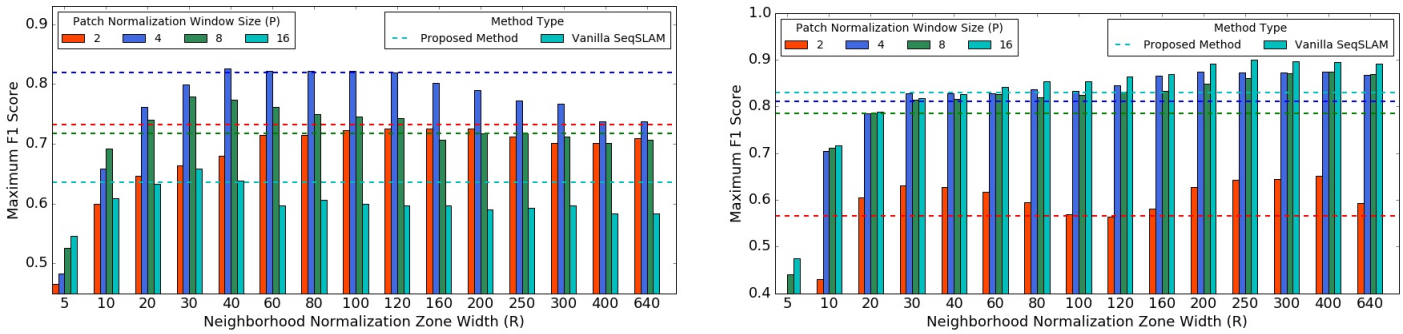
Figure 11: **Top Row:** Ground truth images from the Residence Indoor-Outdoor dataset in a temporal order showing transition from nightly outdoor environment into bright indoor areas. **Middle Row:** Matched places using Vanilla SeqSLAM mostly showing false matching of images having similar lighting conditions. **Bottom Row:** Matched places using proposed method showing correct place recognition. It also shows the transition in environment from broad daylight to dark indoor areas. *Note:* The images captured at night or in dark are shown here after manually brightening them only for sake of visualization.

5.2.2 Paper Submitted to IROS 2017

The proposed research work has been submitted as a conference paper to IROS 2017.

5.3 Motion Estimation under Unfavorable Conditions

Visual Odometry being one of the important competencies for mobile robotics is also an integral part of a visual SLAM system. The correct estimation of egomotion depends on the characteristics of both the camera motion and the environment. The state-of-the-art visual odometry methods, though able to generate a 6-DoF pose, are sensitive to extreme changes in the camera motion or the operating



(a) Parking Lot

(b) CTA-Rail

Figure 12: Performance charts showing maximum F1 Score with respect to different R (Neighborhood Normalization Zone Width) values. R is varied to the maximum value, that is, the size of reference database, after which maximum F1 score becomes constant. The patch normalization window size (P) parameter with value 4 happens to perform better as compared to others most of the times.

environment. We explored the applicability of these methods on some of the real world datasets that exhibit variations in camera speed, for example, a vehicle on a heavy traffic route with frequent halts, and variations in illumination, for example, transiting from an artificially-lit to unlit environment at night etc.

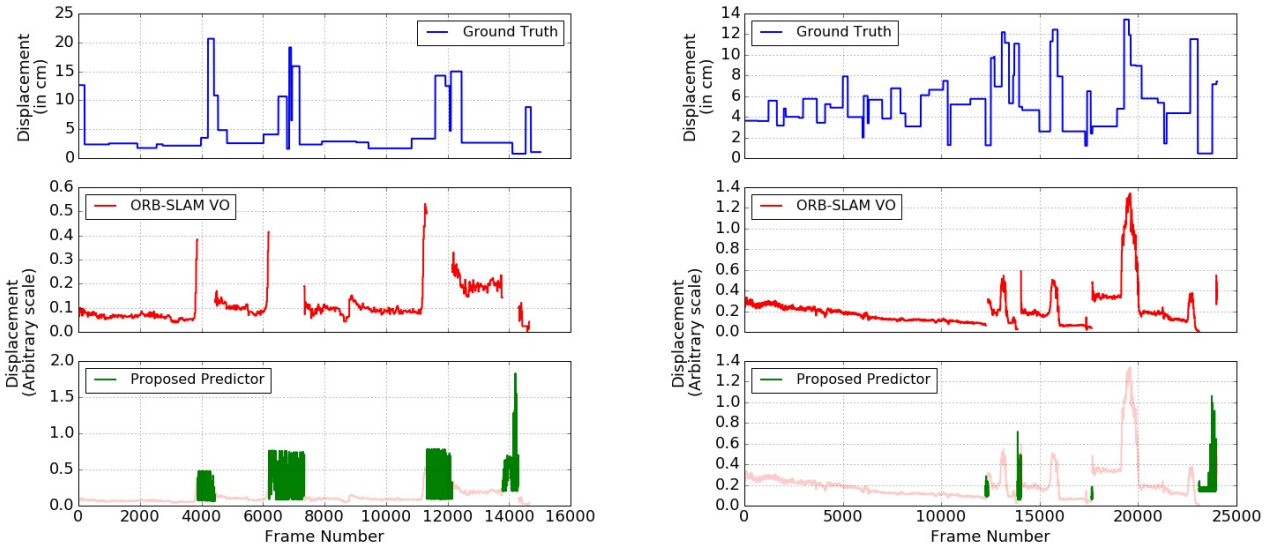
5.3.1 Low Light and High Speed

The low light environment means less visual features as compared to the scene captured in broad daylight. A gradual drop of visual features, for example, while moving from an artificially-lit, cluttered scene to an unlit bland environment at night, leads to failure of visual odometry. We performed experiments on different datasets exhibiting such variations in environment using visual odometry component of ORB-SLAM and LSD-SLAM. The test datasets also had varying camera speeds in different parts of environment which further led to failures because capturing large and small motion signals requires different parameters settings for the system.

We developed a method to generate a motion signal for unfavorable conditions as described above, using sum of absolute difference (SAD) score between down-sampled and patch-normalized consecutive images. This is equivalent to *normalized photometric error* and closely relates to place recognition method used in SeqSLAM. The patch-normalization of images is able to handle the variations due to environmental conditions. It was also observed that SAD scores between images drops gradually with increasing frame separation between the images in an image sequence. Hence, we used a polynomial fitting method to estimate the camera speed according to the SAD score patterns. We were able to estimate the high-speed instances of camera motion within the test dataset where stat-of-the-art methods failed. This is shown in Figure:13 with ground truth trajectory. The point of failures are mainly related to transition from well-lit to unlit areas and sudden increase in camera speed. We plan to use a hybrid approach that can make use of the proposed motion estimator to either tune the state-of-the-art method parameters to correctly estimate a 6-DoF pose or to club the output of both the systems to prevent failure and generate a consistent odometry information.

5.3.2 Speed-Normalized Data Sampling

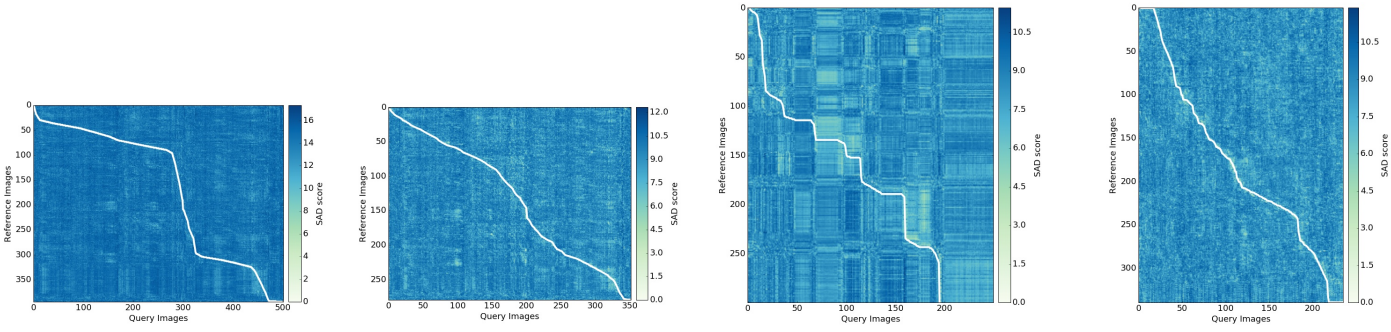
We explored the scope of improving place recognition performance using state-of-the-art visual odometry methods. The motion information is often required in the place recognition framework in order to sample the images at a constant distance as opposed to constant frame separation. A robot moving uniformly will not be affected by such a choice, but in practice, the place recognition and visual SLAM applications contain reference and query imagery captured at varying camera speeds. A perfect example of such a scenario is an on-road vehicle moving at variable pace, due to varying traffic conditions on a particular route at different times of day. Figure:14 shows similarity matrices between reference and query image databases for two different datasets with different methods of frame sampling.



(a) Night Traverse

(b) Day Traverse

Figure 13: The conventional state-of-the-art visual odometry suffers from failures due to high camera speed, sharp turns and low-light or featureless environment. It is shown here how the proposed motion estimator can be used to correctly estimate the displacement under such circumstances.



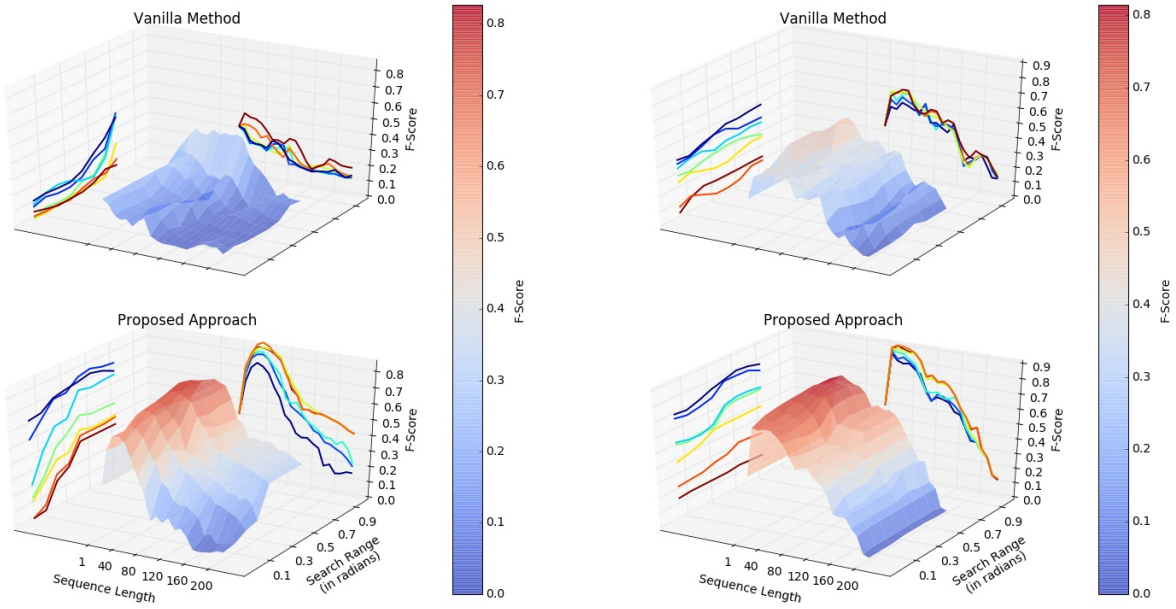
Vanilla Proposed

KG-Footpath

Vanilla Proposed

Surfers Paradise

Figure 14: (a) Ground Truth Matches plotted over the SAD score difference matrix between query and reference images for both vanilla and proposed method.



(a) Kelvin Grove Footpath (b) Surfers Paradise

Figure 15: Performance curves for vanilla and proposed approach for Kelvin Grove Footpath and Surfers Paradise datasets. Though the trends across the two important parameters - *Sequence Length* and *Search Range*, remain same as shown in contours; the overall performance gain is huge using proposed approach.

The speed-normalized imagery can either be obtained using a separate sensor or a visual odometry solution. The challenges involved in using a state-of-the-art visual odometry solution are described in previous subsection. Hence, we resort to the use of proposed motion estimator, also described above, to speed-normalize the image databases for both query and reference. The motion estimator determines the ideal frame separation between the images based on the SAD score between them. This dynamic frame separation corresponds to the physical camera motion and is estimated high when camera moves slowly and low when camera moves fast. We observed a significant improvement in SeqSLAM's place recognition performance using the proposed speed-normalized data sampling before starting to match the places. Figure:15 shows performance comparison between vanilla method and proposed approach with speed-normalized frame sampling.

5.3.3 Planned Submission for ICRA 2018

We plan to submit the proposed research work to ICRA 2018 with some additional research work on top of it.

6 Conclusion

We described the visual SLAM system and its individual components, that is, visual place recognition, visual odometry and representation map. We also explored the possibilities of use of semantics in SLAM framework. We reviewed the relevant literature for all the related components and identified the challenges and viable research gaps. Then we formulated our research problem stating the need of a *6-Dof Semantics-Aware Condition- and Viewpoint-Invariant Visual SLAM* system along with a set of research questions. The possible solutions to these research questions, addressing all the challenges are then detailed within a research plan for the thesis. Finally, the research work progress of last year in line with the research plan is described. The challenges that still remain to solve are summarized as following:

- Developing a robust visual odometry solution for rapid camera motion and adverse environmental conditions.
- Estimating 3D relative pose for loop closures influenced by vast appearance changes.

- Exploring the use of semantics with an integrated approach towards a semantics-aware visual SLAM system.

With reference to the attached timeline, we plan to address all the challenges and fill these research gaps by the end of PhD with research publications.

References

[1] Udo Frese. Interview: Is slam solved? *KI-K{ü}nstliche Intelligenz*, 24(3):255–257, 2010.

[2] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large- Scale Direct Monocular SLAM,. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision (ECCV)*, volume 8690 of *Lecture Notes in Computer Science*, pages 834–849, Cham, 2014. Springer International Publishing.

[3] R. Chatila and J. Laumond. Position referencing and consistent world modeling for mobile robots. In *Proceedings. 1985 IEEE International Conference on Robotics and Automation*, volume 2, pages 138–145. Institute of Electrical and Electronics Engineers, 1985.

[4] Michael Montemerlo and Sebastian Thrun. *FastSLAM: A scalable method for the simultaneous localization and mapping problem in robotics*, volume 27. Springer, 2007.

[5] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. 2005.

[6] E. Olson, J. Leonard, and S. Teller. Fast iterative alignment of pose graphs with poor initial estimates. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 2262–2269. IEEE, 2006.

[7] S. Thrun. The Graph SLAM Algorithm with Applications to Large-Scale Mapping of Urban Structures. *The International Journal of Robotics Research*, 25(5-6):403–429, may 2006.

[8] Niko Sunderhauf and Peter Protzel. Switchable constraints for robust pose graph SLAM. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1879–1884. IEEE, oct 2012.

[9] Paul A Beardsley, Andrew Zisserman, and David W Murray. Sequential updating of projective and affine structure from motion. *International journal of computer vision*, 23(3):235–259, 1997.

[10] Andrew Fitzgibbon Bill Triggs, Philip McLauchlan, Richard Hartley. *Bundle Adjustment A Modern Synthesis*, volume 1883 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, apr 2000.

[11] Hauke Strasdat, J. M M Montiel, and Andrew J. Davison. Real-time monocular SLAM: Why filter? *Proceedings - IEEE International Conference on Robotics and Automation*, 30:2657–2664, may 2010.

[12] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages 652–659. IEEE, 2004.

[13] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual Modeling with a Hand-Held Camera. *International Journal of Computer Vision*, 59(3):207–232, sep 2004.

[14] Johannes Grater, Tobias Schwarze, and Martin Lauer. Robust scale estimation for monocular visual odometry using structure from motion and vanishing points. In *Intelligent Vehicles Symposium (IV), 2015 IEEE*, pages 475–480. IEEE, 2015.

[15] Georg Klein and David Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 1–10. IEEE, nov 2007.

[16] Raul Mur-Artal, J M M Montiel, and Juan D Tardos. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, pages 1–17, feb 2015.

[17] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Monocular Vision Based SLAM for Mobile Robots. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 1027–1031. IEEE, 2006.

[18] Michael John Milford. *Robot navigation from nature: Simultaneous localisation, mapping, and path planning based on hippocampal models*, volume 41. Springer Science & Business Media, 2008.

[19] M.J. Milford, G.F. Wyeth, and D. Prasser. RatSLAM: a hippocampal model for simultaneous localization and mapping. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, volume 1, pages 403–408 Vol.1. IEEE, 2004.

[20] Matthew Collett. How desert ants use a visual landmark for guidance along a habitual route. *Proceedings of the National Academy of Sciences of the United States of America*, 107(25):11638–11643, jun 2010.

[21] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, volume 2, pages 2161–2168. IEEE, 2006.

[22] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2. IEEE, 2003.

[23] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, nov 2004.

[24] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, jun 2008.

[25] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: binary robust independent elementary features. In *European Conference on Computer Vision (ECCV)*, pages 778–792. Springer-Verlag, sep 2010.

[26] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, pages 2564–2571. IEEE, nov 2011.

[27] M. Cummins and P. Newman. Appearance-only SLAM at large scale with FAB-MAP 2.0. *The International Journal of Robotics Research*, 30(9):1100–1123, nov 2010.

[28] D. Galvez-Lopez and J. D. Tardos. Bags of Binary Words for Fast Place Recognition in Image Sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, oct 2012.

[29] Raul Mur-Artal and Juan D. Tardos. Fast relocalisation and loop closing in keyframe-based SLAM. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 846–853. IEEE, may 2014.

[30] Adrien Angeli, Stephane Doncieux, Jean-Arcady Meyer, and David Filliat. Real-time visual loop-closure detection. In *2008 IEEE International Conference on Robotics and Automation*, pages 1842–1847. IEEE, may 2008.

[31] Nishant Kejriwal, Swagat Kumar, and Tomohiro Shibata. High performance loop closure detection using bag of word pairs. *Robotics and Autonomous Systems*, 77:55–65, 2016.

[32] César CADENA, Dorian GALVEZ-LOPEZ, Juan D. TARDOS, and José NEIRA. Robust Place Recognition With Stereo Sequences. *IEEE transactions on robotics*, 28(4):871–885, 2012.

[33] Mark Cummins and Paul Newman. Probabilistic appearance based navigation and loop closing. In *Robotics and automation, 2007 IEEE international conference on*, pages 2042–2048. IEEE, 2007.

[34] Manuel Lopez-Antequera, Ruben Gomez-Ojeda, Nicolai Petkov, and Javier Gonzalez-Jimenez. Appearance-Invariant Place Recognition by Discriminatively Training a Convolutional Neural Network. *Pattern Recognition Letters*, 2017.

[35] Michael J. Milford and Gordon F. Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 1643–1649, 2012.

[36] Tayyab Naseer, Luciano Spinello, and Cyrill Stachniss. Robust Visual Robot Localization Across Seasons using Network Flows. *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2014.

[37] B.J.A Kröse, N Vlassis, R Bunschoten, and Y Motomura. A probabilistic model for appearance-based robot localization. *Image and Vision Computing*, 19(6):381–391, apr 2001.

[38] Hernan Badino, Daniel Huber, and Takeo Kanade. Real-time topometric localization. In *2012 IEEE International Conference on Robotics and Automation*, pages 1635–1642. IEEE, may 2012.

[39] N. Sunderhauf and P. Protzel. BRIEF-Gist - Closing the loop by simple means. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1234–1241. IEEE, sep 2011.

[40] Hatem Alismail, Brett Browning, and Simon Lucey. Photometric Bundle Adjustment for Vision-Based SLAM. *arXiv preprint arXiv:1608.02026*, 2016.

[41] Colin McManus, Ben Upcroft, and Paul Newman. Learning place-dependant features for long-term vision-based localisation. *Autonomous Robots*, 39(3):363–387, 2015.

[42] Colin McManus, Ben Upcroft, and Paul Newmann. Scene signatures : localised and point-less features for localisation, jul 2014.

[43] Michael J Milford and Gordon F Wyeth. Mapping a Suburb With a Single Camera Using a Biologically Inspired SLAM System. *IEEE Transactions on Robotics*, 24(5):1038–1053, 2008.

[44] S Niko, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free. *Robotics Science and Systems*, 2015.

[45] Zetao Chen, Adam Jacobson, Niko Sunderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. Deep Learning Features at Scale for Visual Place Recognition. *arXiv preprint arXiv:1701.05105*, 2017.

[46] Zetao Chen, Obadiah Lam, Adam Jacobson, and Michael Milford. Convolutional neural network-based place recognition. *arXiv preprint arXiv:1411.1509*, 2014.

[47] Laurent Kneip, Margarita Chli, and Roland Yves Siegwart. Robust Real-Time Visual Odometry with a Single Camera and an IMU. In *British Machine Vision Conference*. Eidgenössische Technische Hochschule Zürich, Autonomous Systems Lab, 2011.

[48] Michael Milford, Jennifer Firn, James Beattie, Adam Jacobson, Edward Pepperell, Eugene Mason, Michael Kimlin, and Matthew Dunbabin. Automated sensory data alignment for environmental and epidermal change monitoring. In *Australasian Conference on Robotics and Automation*, pages 1–10, 2014.

[49] M. Agrawal and K. Konolige. Real-time Localization in Outdoor Environments using Stereo Vision and Inexpensive GPS. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 1063–1068. IEEE, 2006.

[50] Georgios Floros, Benito van der Zander, and Bastian Leibe. OpenStreetSLAM: Global vehicle localization using OpenStreetMaps. In *2013 IEEE International Conference on Robotics and Automation*, pages 1054–1059. IEEE, may 2013.

[51] Stefan Kohlbrecher, Oskar von Stryk, Johannes Meyer, and Uwe Klingauf. A flexible and scalable SLAM system with full 3D motion estimation. In *2011 IEEE International Symposium on Safety, Security, and Rescue Robotics*, pages 155–160. IEEE, nov 2011.

[52] Rohan Paul and Paul Newman. FAB-MAP 3D: Topological mapping with spatial and visual appearance. In *2010 IEEE International Conference on Robotics and Automation*, pages 2649–2656. IEEE, may 2010.

[53] Stephane Bazeille and David Filliat. Incremental topo-metric SLAM using vision and robot odometry. In *2011 IEEE International Conference on Robotics and Automation*, pages 4067–4073. IEEE, may 2011.

[54] Edward Pepperell, Peter I Corke, and Michael J Milford. Automatic image scaling for place recognition in changing environments. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 1118–1124. IEEE, 2015.

[55] Edward Pepperell, Peter Corke, and Michael Milford. All-environment visual place recognition with SMART, jun 2014.

[56] I. Mahon, S.B. Williams, O. Pizarro, and M. Johnson-Roberson. Efficient View-Based SLAM Using Visual Loop Closures. *IEEE Transactions on Robotics*, 24(5):1002–1014, oct 2008.

[57] C.G. Harris and J.M. Pike. 3D positional integration from image sequences. *Image and Vision Computing*, 6(2):87–90, may 1988.

[58] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision (ECCV)*, pages 430–443, 2006.

[59] Motilal Agrawal, Konolige Kurt, and Blas Morten Rufus. *CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching*, volume 5305 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[60] Georg Klein and David Murray. Improving the agility of keyframe-based SLAM. In *Computer Vision–ECCV 2008*, pages 802–815. Springer, 2008.

[61] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO : Fast Semi-Direct Monocular Visual Odometry. *IEEE International Conference on Robotics and Automation*, 2014.

[62] Jakob Engel, Jurgen Sturm, and Daniel Cremers. Semi-dense Visual Odometry for a Monocular Camera. In *2013 IEEE International Conference on Computer Vision*, pages 1449–1456. IEEE, dec 2013.

[63] Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. DTAM: Dense tracking and mapping in real-time. In *2011 International Conference on Computer Vision*, pages 2320–2327. IEEE, nov 2011.

[64] Tommi Tykkälä, Cédric Audras, and Andrew I. Comport. Direct iterative closest point for real-time visual odometry. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2050–2056, 2011.

[65] Hatem Alismail, Brett Browning, and Simon Lucey. Direct visual odometry using bit-planes. *arXiv preprint arXiv:1604.00990*, 2016.

[66] Kishore Reddy Konda and Roland Memisevic. Learning Visual Odometry with a Convolutional Network. In *VISAPP (1)*, pages 486–490, 2015.

[67] Alex Kendall, Matthew Grimes, and Roberto 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.

[68] Julie Dequaire, Dushyant Rao, Peter Ondruska, Dominic Wang, and Ingmar Posner. Deep Tracking on the Move: Learning to Track the World from a Moving Vehicle using Recurrent Neural Networks. *arXiv preprint arXiv:1609.09365*, 2016.

[69] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. SfM-Net: Learning of Structure and Motion from Video. *arXiv preprint arXiv:1704.07804*, 2017.

[70] W. L. D. Lui and R. Jarvis. A pure vision-based topological SLAM system. *The International Journal of Robotics Research*, 31(4):403–428, feb 2012.

[71] J.-S. Gutmann, M. Fukuchi, and M. Fujita. 3D Perception and Environment Map Generation for Humanoid Robot Navigation. *The International Journal of Robotics Research*, 27(10):1117–1134, oct 2008.

[72] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat. Visual topological SLAM and global localization. In *2009 IEEE International Conference on Robotics and Automation*, pages 4300–4305. IEEE, may 2009.

[73] Kurt Konolige, Eitan Marder-Eppstein, and Bhaskara Marthi. Navigation in hybrid metric-topological maps. In *2011 IEEE International Conference on Robotics and Automation*, pages 3041–3047. IEEE, may 2011.

[74] Niko Sunderhauf, Feras Dayoub, Sean McMahon, Ben Talbot, Ruth Schulz, Peter Corke, Gordon Wyeth, Ben Upcroft, and Michael Milford. Place categorization and semantic mapping on a mobile robot. In *Proceedings of the International Conference on Robotics and Automation*. IEEE, 2016.

[75] Alex Flint, Christopher Mei, Ian Reid, and David Murray. Growing semantically meaningful models for visual slam. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 467–474. IEEE, 2010.

[76] Alexander D. Stewart and Paul Newman. LAPS - localisation using appearance of prior structure: 6-DoF monocular camera localisation using prior pointclouds. In *2012 IEEE International Conference on Robotics and Automation*, pages 2625–2632. IEEE, may 2012.

[77] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *The International Journal of Robotics Research*, 31(5):647–663, feb 2012.

[78] Thomas Whelan, Hordur Johannsson, Michael Kaess, John J. Leonard, and John Mcdonald. Robust Real-Time Visual Odometry for Dense RGB-D Mapping. *IEEE International Conference on Robotics and Automation*, (i):5724–5731, 2013.

[79] Andrew J. Davison, Ian D. Reid, Nicholas D. Molton, and Olivier Stasse. MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[80] Chris Engels, H Stewénius, and D Nistér. Bundle adjustment rules. *Photogrammetric computer vision*, pages 266–271, 2006.

[81] Raul Mur-Artal and Juan D. Tardos. Probabilistic Semi-Dense Mapping from Highly Accurate Feature-Based Monocular SLAM. *Proceedings of Robotics: Science and Systems, Rome, Italy*, 2015.

[82] Ananth Ranganathan and Frank Dellaert. Semantic modeling of places using objects. Georgia Institute of Technology, 2007.

[83] Ingmar Posner, Derik Schroeter, and Paul Newman. Online generation of scene descriptions in urban environments. *Robotics and Autonomous Systems*, 56(11):901–914, 2008.

[84] Jianxiong Xiao and Long Quan. Multiple view semantic segmentation for street view images. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 686–693. IEEE, 2009.

[85] Cyrill Stachniss, Oscar Martinez Mozos, Axel Rottmann, Wolfram Burgard, and Others. Semantic labeling of places. 2005.

[86] Andrzej Pronobis. *Semantic mapping with mobile robots*. PhD thesis, KTH Royal Institute of Technology, 2011.

[87] Javier Civera, Dorian Gálvez-López, Luis Riazuelo, Juan D Tardós, and J M M Montiel. Towards semantic SLAM using a monocular camera. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 1277–1284. IEEE, 2011.

[88] Jörg Stückler, Nenad Biresev, and Sven Behnke. Semantic mapping using object-class segmentation of RGB-D images. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 3005–3010. IEEE, 2012.

[89] Nicola Fioraio and Luigi Di Stefano. Joint detection, tracking and mapping by semantic bundle adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1538–1545, 2013.

[90] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul H J Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1352–1359, 2013.

[91] Shrihari Vasudevan, Stefan Gächter, Viet Nguyen, and Roland Siegwart. Cognitive maps for mobile robotsan object based approach. *Robotics and Autonomous Systems*, 55(5):359–371, 2007.

[92] Dorian Gálvez-López, Marta Salas, Juan D Tardós, and J M M Montiel. Real-time monocular object slam. *Robotics and Autonomous Systems*, 75:435–449, 2016.

[93] Renato F Salas-Moreno. *Dense semantic SLAM*. PhD thesis, Citeseer, 2014.

[94] Andreas Nüchter and Joachim Hertzberg. Towards semantic maps for mobile robots. *Robotics and Autonomous Systems*, 56(11):915–926, 2008.

[95] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[96] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[97] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[98] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

[99] Genevieve Patterson and James Hays. SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In *Proceeding of the 25th Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[100] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (TOG)*, 33(4):149, 2014.

[101] Will Maddern, Alexander D Stewart, and Paul Newman. LAPS-II: 6-DoF Day and Night Visual Localisation with Prior 3D Structure for Autonomous Road Vehicles.

[102] Mark Cummins and Paul Newman. Highly scalable appearance-only SLAM - FAB-MAP 2.0. In *Robotics: Science and Systems*, Seattle, United States, 2009.

[103] John Skinner, Sourav Garg, Niko Sünderhauf, Peter Corke, Ben Upcroft, and Michael Milford. High-fidelity simulation for evaluating robotic vision performance. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 2737–2744. IEEE, 2016.