Universidad de Alcalá

PhD. Program in Electronics: Advanced Electronic Systems. Intelligent Systems

Topological Place Recognition for Life-Long Visual Localization

PhD. Thesis Presented by

Roberto Arroyo Contera

2017

Universidad de Alcalá

PhD. Program in Electronics: Advanced Electronic Systems. Intelligent Systems

Topological Place Recognition for Life-Long Visual Localization

PhD. Thesis Presented by

Roberto Arroyo Contera

Advisor

Dr. Luis Miguel Bergasa Pascual

Co-Advisor

Dr. Pablo Fernández Alcantarilla

Alcalá de Henares, March 31st, 2017

Agradecimientos

El largo camino recorrido durante estos últimos cuatro años de doctorado para llegar a este momento en el que el viaje toca a su fin no habría sido posible sin todas las personas que me han apoyado cada día y me han regalado su afecto para poder lograr alcanzar esta meta.

En primer lugar quería agradecer a mis tutores su inestimable ayuda. A Luis Miguel, por darme la oportunidad de realizar esta tesis, por supervisar diariamente mi trabajo de una forma tan cuidadosa y por su apoyo constante a lo largo de este tiempo. A Pablo, por todo el esfuerzo que siempre ha puesto conmigo y por darme el privilegio de trabajar con él codo con codo durante mi estancia en Cambridge, además de brindarme incluso cobijo en su casa y, lo más importante de todo, por su amistad durante estos años. Espero haber estado a la altura de ambos, es lo mínimo que os merecéis.

De aquellos días que pasé por Cambridge guardo un muy buen recuerdo gracias a todos los buenos compañeros que tuve en Toshiba Research Europe: Björn, Minh-Tri, Riccardo, Frank, Chris, Ujwal, Simon, Adrián y Germán. Sin olvidarme de la comunidad española que había por allí y con la que hice tan buenas migas: Iván, Marcos, Davides y Tania. ¡Volveremos a vernos por el Regal algún dia!

El último año también realicé una estancia de tres meses en Sydney que sirvió como colofón de la tesis y que me permitió trabajar con excelentes investigadores y sobre todo grandes personas. En primer lugar, Eduardo, mi supervisor allí y director del Australian Centre for Field Robotics, un hombre hospitalario y amable donde los haya. También pasé muy buenos momentos con todos los miembros de su equipo y siempre me acordaré de ellos, en especial de Wei, Stewart, Alex y James. ¡Tenéis un amigo en España!

Me paro a pensar y gracias a la tesis he recorrido medio mundo y creo que todas esas aventuras me han enriquecido como persona. Entre esas experiencias están los congresos en Chicago, Seattle, Sicilia o Daejeon, en los que conocí a otros investigadores españoles que hicieron que aquellos días de conferencias fuesen mucho más agradables: Rafa, Llanos, Edu, Mónica, Estefania, Cristina, Macarena, Jesús... espero que algún día la vida nos vuelva a juntar en algún sitio de este pequeño mundo.

A los que sí que les tengo que hacer la ola sin duda es a mis compañeros y amigos durante estos años en los grupos de investigación RobeSafe e INVETT. Sois todos personas fantásticas y sin vosotros todo este tiempo no habría sido lo mismo ni habría conseguido las cosas que he conseguido gracias a vuestro buen rollo y ayuda constante. Hemos pasado momentos geniales, desde cenas, hasta noches de fiesta, cabritadas y algún que otro viajecillo. Me gustaría tener unas palabras para todos, pero habéis sido muchos, y todos de una forma u otra me habéis aportado algo, ya lo sabéis, espero no dejarme a nadie: Alberto, Alex, Almazán, Álvaro, Amaia, Andrés, Augusto, Balky, Carlos, Carlota, Chris, Dani, David, Edus, Estefanía, Fer, Fran, Gavi, Iván, Llama, Javis, Lidia, Mario, Nacho, Noe, Oscar, Raúl, Rubén, Sebas y Sergios.

No puedo evitar hacer una mención especial a los dos compañeros de laboratorio con los que más horas de trabajo y conversaciones he compartido. A Javi Yebes, porque sin toda la ayuda que me prestó desde el principio posiblemente no habría llegado ni a arrancar el ordenador y porque es una de las personas con más calidad humana a las que he tenido la suerte de conocer en los últimos años. A Edu Romera, por su enorme compañerismo, su gran sentido del humor, sus visitas por Sydney y por estar siempre dispuesto a echar una mano, incluso escalando Mordor (sí, eso es real y sucedió en Nueva Zelanda, ¡somos los mejores alpinistas de todos los tiempos, sin duda!).

Por supuesto, también agradecer todo su apoyo a los profesores del RobeSafe, que siempre han estado dispuestos a dedicar su tiempo para contribuir en mi investigación: Miguel Ángel, Manuel, Rafa, Elena y Pedro. Especial agradecimiento a Llorca, sin él nunca me habría picado el gusanillo de la visión artificial y todavía a día de hoy las charlas con él siempre me hacen aprender algo y pasar un rato agradable.

También quiero dar las gracias a todos los alumnos a los que he tenido el gusto de dar clase durante estos años en las asignaturas de visión artificial y sistemas de percepción, porque yo también he aprendido mucho de ellos y por saludarme siempre con una sonrisa cuando me les encuentro por los pasillos de la universidad.

No me olvido de que hace no tanto yo también fui alumno. En aquellas aulas empezó realmente este viaje junto a mis buenos compañeros de la carrera de Ingeniería Técnica en Informática de Gestión: Jose, Ricardo, Arturo, Dani, Martín, Pili, Gema y Sonia. Sin olvidarme tampoco de los de la Ingeniería Superior en Informática: Nacho, Dani y Marta.

En estos agradecimientos no pueden faltar tampoco mis grandes amistades de Guadalajara. En primer lugar, tengo que agradecer a Adri, Iván y Clavo su visita mientras estaba en Inglaterra y los ratos de diversión por allí, que no tuvieron precio. Y como no, a todos los demás por estar siempre ahí en las buenas y en las malas: Baeza, Chechu, Darío, Fran, Galán, Héctor, Javi, Joseda y Urban. Y, por qué no decirlo, gracias a sus novias por aguantar a tales elementos y sobre todo porque también las considero ya mis amigas después de tanto tiempo: Bela, Debo, Laura, Lucia, Mari, Marina y Rocío.

Tampoco pueden faltar mis amigos de toda la vida de Gajanejos, gente auténtica donde las haya y que nunca han dejado de estar a mi lado pase lo que pase. Primero, a mi buena amiga Araceli y a mi compinche Pedro, que también estuvieron visitándome por las islas británicas (si, con vosotros también volveré al Regal algún día, ¡sé que queréis volver a ver a gente bailando sin zapatos!). A mi gran amigo Nacho por acompañarme en la aventura australiana y por la dicción perfecta de la lengua inglesa que allí adquirió (sé que ningún gajanejero se creerá esto último, pero es verdad, ¡por momentos parecía una versión alcarreña de Shakespeare!). También a Álvaro por las tres semanas que compartió con nosotros por Sydney. Y por supuestísimo, a todos los demás buenos amigos que tengo por allí: Ana, Jose y Gema, Nuria y Diego, Carlos y, cómo no, el ídolo local, Jaime. Sin olvidarme de hijos adoptivos del pueblo como Sergio Loqueras o el señor Verbenas. Tampoco hay que olvidar nunca que la Alcarria es una tierra grande, y supongo que por eso en ella también hay personas igual de grandes, como Laura: gracias por todos los ánimos y apoyo durante la escritura de este libro y por tener siempre un paraguas a mano para mí cuando llueve.

El agradecimiento más infinito va para toda mi familia. En especial a mis padres, porque sin ellos nunca habría llegado hasta aquí ni sería lo que soy, gracias por todo. Y a mi hermano Álvaro, que ha sido siempre un referente para mí y me ha ayudado en todo lo que ha hecho falta desde que tengo uso de razón. Como no, también a Marta y a Adrián: espero que pronto puedas leer estas líneas que te dedica tu tío Rober.

Por último, siempre he tenido un fiel compañero de viaje que me ha dado mucho sin pedir nada a cambio: el rock'n'roll. Las horas de trabajo no habrían sido lo mismo sin la música que volaba a través de mis oídos mientras programaba, procesaba resultados o leía papers. De hecho, mientras escribo estas líneas la música me sigue acompañando. Por ello, quería terminar estos agradecimientos con unas palabras de la canción que estoy escuchando al concluir este viaje. Unas palabras de un músico humilde, que emanan la misma humildad que siempre se ha tratado de tener a la hora de realizar esta tesis con esfuerzo y constancia.

No pretendas engañarte, ni dejarte arrastrar,
por instintos infundados que te harán fracasar.
Lo que quieres tendrás que ganarlo, nadie te lo viene a dar.
Solo intenta ser tú mismo, aprendiendo a escuchar.

Rosendo Mercado.

Resumen

La navegación de vehículos inteligentes o robots móviles en períodos largos de tiempo ha experimentado un gran interés por parte de la comunidad investigadora en los últimos años. Los sistemas basados en cámaras se han extendido ampliamente en el pasado reciente gracias a las mejoras en sus características, precio y reducción de tamaño, añadidos a los progresos en técnicas de visión artificial. Por ello, la localización basada en visión es una aspecto clave para desarrollar una navegación autónoma robusta en situaciones a largo plazo. Teniendo en cuenta esto, la identificación de localizaciones por medio de técnicas de reconocimiento de lugar topológicas puede ser complementaria a otros enfoques como son las soluciones basadas en el *Global Positioning System (GPS)*, o incluso suplementaria cuando la señal *GPS* no está disponible o es denegada.

El estado del arte en reconocimiento de lugar topológico ha mostrado un funcionamiento satisfactorio en el corto plazo. Sin embargo, la localización visual a largo plazo es problemática debido a los grandes cambios de apariencia que un lugar sufre como consecuencia de elementos dinámicos, la iluminación o la climatología, entre otros. Por ello, una de las áreas de investigación más popularizadas en los últimos tiempos es la identificación de lugares durante las cuatro estaciones del año. El objetivo de esta tesis es enfrentarse a las dificultades de llevar a cabo una localización topológica eficiente y robusta a lo largo del tiempo. En consecuencia, se van a contribuir dos nuevos enfoques basados en reconocimiento visual de lugar para resolver los diferentes problemas asociados a una localización visual a largo plazo.

Por un lado, un método de reconocimiento de lugar visual basado en descriptores binarios es propuesto. La innovación de este enfoque reside en la descripción global de secuencias de imágenes como códigos binarios, que son extraídos mediante un descriptor basado en la técnica denominada *Local Difference Binary (LDB)*. Los descriptores son eficientemente asociados usando la distancia de *Hamming* y un método de búsqueda conocido como *Approximate Nearest Neighbors (ANN)*. Además, una técnica de iluminación invariante es aplicada para mejorar el funcionamiento en condiciones luminosas cambiantes. El empleo de la descripción binaria previamente introducida proporciona una reducción de los costes computacionales y de memoria, lo cual es necesario para un funcionamiento a largo plazo. Adicionalmente, tres versiones de esta propuesta son diseñadas para explotar las ventajas de diferentes cámaras: monoculares, estéreo y panorámicas.

Por otro lado, también se presenta un método de reconocimiento de lugar visual basado en *deep learning*, en el cual los descriptores aplicados son procesados por una *Convolutional Neural Network (CNN)*. Este es un concepto recientemente popularizado en visión artificial que ha obtenido resultados impresionantes en problemas de clasificación de imagen. En este trabajo, las *CNNs* son usadas para mejorar la precisión de la localización topológica contra cambios de apariencia relacionados con las estaciones del año, porque los métodos tradicionales tienen más dificultades en estas condiciones. La novedad de nuestro enfoque reside en la fusión de la información de imagen de múltiples capas convolucionales a varios niveles y granularidades. Además, los datos redundantes de los descriptores basados en *CNNs* son comprimidos en un número reducido de bits para una localización más eficiente. El descriptor final es condensado aplicando técnicas de compresión y binarización para realizar una asociación usando de nuevo la distancia de *Hamming*. A lo largo de la tesis, se discuten los pros y los contras de esta propuesta de reconocimiento visual con respecto a la basada en descriptores tradicionales. En términos generales, los métodos centrados en *CNNs* mejoran la precisión generando representaciones visuales de las localizaciones más detalladas. Sin embargo, la desventaja es que los costes computacionales también son incrementados comparados con los requeridos para procesar descriptores tradicionales.

Ambos enfoques de reconocimiento de lugar visual son extensamente evaluados sobre varios *datasets* públicos. Estas pruebas arrojan una precisión satisfactoria en situaciones a largo plazo, como es corroborado por los resultados mostrados, que comparan nuestros métodos contra los principales algoritmos del estado del arte, mostrando mejores resultados para todos los casos. En estos experimentos, nuestras propuestas son validadas en localización visual a largo plazo, especialmente si se considera que una distancia mayor a 3000 km es atravesada en las pruebas realizadas a lo largo de las estaciones del año.

Además, también se ha analizado la aplicabilidad de nuestro reconocimiento de lugar topológico en diferentes problemas de localización. Estas aplicaciones incluyen la detección de cierres de lazo basada en los lugares reconocidos o la corrección de la deriva acumulada en odometría visual usando la información proporcionada por los cierres de lazo. El problema de *Simultaneous Localization And Mapping (SLAM)* también es estudiado teniendo en cuenta las medidas corregidas basadas en cámaras y la información proporcionada por otros sensores, como el *GPS* o el *Light Detection And Ranging (LiDAR)*. Asimismo, también se consideran las aplicaciones de la detección de cambios geométricos a lo largo de las estaciones del año, que son esenciales para las actualizaciones de los mapas en sistemas de conducción autónomos centrados en una operación a largo plazo. Todas estas contribuciones son discutidas al final de la tesis, incluyendo varias conclusiones sobre el trabajo presentado y algunas líneas de investigación futuras.

Palabras clave: navegación autónoma, localización visual a largo plazo, reconocimiento de lugar topológico, descriptores binarios, descriptores basados en *CNNs*, odometría visual, *SLAM*.

Abstract

The navigation of intelligent vehicles or mobile robots in long periods of time has experienced a great interest by the research community in the last years. In this sense, visual information has become a valuable asset in any perception scheme designed to improve the scene understanding for autonomous driving. Camera-based systems have been broadly extended within the recent past due to the improvements in camera features, price and size reduction, added to the progress in computer vision. For this reason, vision-based localization is a key aspect to develop a robust automated navigation approach in long-term situations. According to this, the identification of locations by means of topological place recognition techniques can be complementary to other sensing technologies such as solutions based on Global Positioning System (GPS), or even supplementary when GPS signal is not completely available or denied.

The state of the art in topological place recognition has shown satisfactory performance in short-term problems. However, life-long visual localization is a challenging topic that is more problematic because of the strong appearance changes that a place suffers due to dynamic elements, illumination or weather, among others. In this regard, one of the most popularized research areas appeared in recent times is related to the identification of places across the four seasons of the year. The goal of this dissertation is to cope with the main difficulties of carrying out an efficient and robust topological localization along the time course. Consequently, we contribute two novel approaches based on visual place recognition in order to solve the different problems associated with life-long visual localization.

On the one hand, a visual place recognition method based on hand-crafted binary descriptors is proposed. The innovation of this approach resides in the global description of sequences of images as binary codes, which are extracted from a Local Difference Binary (LDB) descriptor and efficiently matched using the Hamming distance in an Approximate Nearest Neighbors (ANN) search. Besides, an illumination invariant technique is applied for improving the performance in changing lighting conditions. The usage of the introduced binary description and matching method provides a reduction of memory and computational costs, which is necessary for a long-term operation. In addition, three versions of this proposal are designed with the aim of exploiting the advantages of different types of cameras: monocular, stereo and panoramic.

On the other hand, we also present a visual place recognition method based on deep learning, in which the applied features are processed by a Convolutional Neural Network (CNN). This is a concept recently popularized in the computer vision community that has obtained impressive results in image classification problems. Here, we take advantage of CNNs in order to improve the accuracy of topological localization against seasonal changes, because traditional solutions focused on hand-crafted descriptors have more difficulties in these conditions. The novelty in our approach relies on fusing the image information from multiple convolutional layers at several levels and granularities. In addition, the redundant data of CNN features is compressed into a tractable number of bits for a more efficient and robust life-long localization. The final descriptor is reduced by applying simple compression and binarization techniques for fast matching using again the Hamming distance. Along the dissertation, we discuss the pros and cons of this place recognition proposal with respect to the one based on hand-crafted features. In general terms, methods focused on CNNs improve the precision by generating more detailed visual representations of locations. However, the disadvantage is that computational costs are also incremented compared to the ones required for processing hand-crafted descriptors.

Both topological place recognition approaches are extensively evaluated over several publicly available datasets. These tests yield a satisfactory precision in long-term conditions, as corroborated by the exhibited results, which compare our methods against the main state-of-the-art algorithms, showing better results for all the cases. In these experiments, our proposals are validated in life-long visual localization, especially if we consider that a distance higher than 3000 km is traversed in the performed tests across the seasons.

Additionally, we also analyze the applicability of our topological place recognition in different localization problems. These applications include the detection of loop closures based on the recognized places or the correction of the accumulated drift in visual odometry estimations using the loop closure information. The Simultaneous Localization And Mapping (SLAM) problem is also studied by taking into account the corrected measurements obtained by means of camera-based localization and the information provided by other sensing technologies, such as GPS or Light Detection And Ranging (LiDAR). Besides, we also consider the applications of geometric change detection across seasons, which is essential for mapping updates in autonomous driving systems focused on a long-term operation. All these contributions are discussed at the end of the dissertation, including several conclusions about the presented work and some future research lines.

Keywords: autonomous navigation, life-long visual localization, topological place recognition, binary descriptors, CNN features, visual odometry, SLAM.

Contents

List of Figures

List of Tables

List of Acronyms

ABLE	Able for Binary-appearance Loop-closure Evaluation.
ADAS	Advanced Driver Assistance Systems.
ANN	Approximate Nearest Neighbors.
BoVW	Bags of Visual Words.
BRIEF	Binary Robust Independent Elementary Features.
BRISK	Binary Robust Invariant Scalable Keypoints.
CenSurE	Center Surround Extrema.
CNN-VTL	Convolutional Neural Network for Visual Topological Localization.
CNNs	Convolutional Neural Networks.
D-LDB	Disparity Local Difference Binary.
ELAS	Efficient LArge-scale Stereo matching.
FAB-MAP	Fast Appearance-Based MAPping.
FAST	Features from Accelerated Segment Test.
FLANN	Fast Library for Approximate Nearest Neighbors.
FPFH	Fast Point Feature Histograms.
FREAK	Fast REtinA Keypoint.
GCH	Global Color Histogram.
GPGPU	General-Purpose computing on Graphics Processing Units.
GPS	Global Positioning System.
HOG	Histogram of Oriented Gradient.

LBP	Local Binary Patterns.
LDB	Local Difference Binary.
LiDAR	Light Detection And Ranging.
LSH	Local Sensitive Hashing.
MSER	Maximally Stable Extremal Regions.
NARF	Normal Aligned Radial Feature.
OGRE 3D	Object-oriented Graphics Rendering Engine 3D.
ORB	Oriented fast and Rotated BRIEF.
PCA	Principal Component Analysis.
PCL	Point Cloud Library.
PFH	Point Feature Histograms.
PIRF	Position-Invariant Robust Features.
PTAM	Parallel Tracking And Mapping.
RaDAR	Radio Detection And Ranging.
RBS	Random Bit Selection.
ReLU	Rectified Linear Unit.
SAD	Sum of Absolute Differences.
SfM	Structure from Motion.
SGBM	Semi-Global Block Matching.
SIFT	Scale Invariant Feature Transform.
SLAM	Simultaneous Localization And Mapping.
SURE	SURface Entropy for 3D features.
SURF	Speeded Up Robust Features.
UAVs	Unmanned Aerial Vehicles.
UGVs	Unmanned Ground Vehicles.
UUVs	Unmanned Underwater Vehicles.
VFH	Viewpoint Feature Histogram.
VGG-F	Visual Geometry Group Fast network.

VLAD	Vector of Locally Aggregated Descriptors.
VO	Visual Odometry.
WI-SIFT	Whole Image SIFT.
WI-SURF	Whole Image SURF.

Chapter 1

Introduction

Where am I? Probably, all of us have asked ourselves this question during our life in varied situations and contexts. However, not only people need to answer it, but also mobile robots and autonomous vehicles. A simple question, but with a difficult solution that experts in robotics have been trying to find for several years ago, with the aim of designing robust navigation systems.

The localization problem can be dealt with by means of different approaches depending on the way of perceiving the environment. This situational awareness is conditioned by the sensors used for recognizing a place and identifying where is located. On the one hand, there are several well-established sensing technologies commonly applied in autonomous navigation, such as Global Positioning System (GPS) or range-based. On the other hand, the community is also considering camera-based solutions as an interesting alternative, because of the useful information that they can provide for scene understanding.

Unfortunately, visual localization in a long-term context is not an easy task due to the drastic appearance changes that places suffer along the time course, as exemplified in Fig. 1.1. Nowadays, research efforts in this line are focused on solving the life-long visual localization problem. This dissertation studies this challenging topic and contributes novel solutions for long periods of time based on topological place recognition methods.

(1.1.1) Visual appearance of the place in winter.

(1.1.2) Visual appearance of the place in spring.

Figure 1.1: An example of the appearance changes suffered by a place across the seasons. The aspect of the environment is drastically modified, which is a very difficult problem for life-long visual localization. The depicted frames correspond to sequences recorded in the Nordland dataset [Sünderhauf et al., 2013a].

1.1 General Overview and Problem Description

Vision-based localization systems have been broadly extended within the recent past due to the improvements in camera features, price and size reduction, added to the progress in computer vision algorithms for place recognition. Moreover, Simultaneous Localization And Mapping (SLAM) [Durrant-Whyte and Bailey, 2006, Bailey and Durrant-Whyte, 2006] systems have evolved in parallel to these improvements in order to achieve a more detailed representation of locations. In fact, visual SLAM methods [Ros et al., 2012, Fuentes-Pacheco et al., 2012] have become one of the main tendencies in visual localization jointly with other related technologies, such as Visual Odometry (VO) [Nistér et al., 2006, Scaramuzza and Fraundorfer, 2011, Fraundorfer and Scaramuzza, 2012].

Although visual SLAM and VO have represented an important attainment that has supposed a great advance in visual localization, still there are some important related questions where great research efforts are being carried out, such as the recognition of places in long-term operation [Lowry et al., 2016] or the identification of revisited locations. In this sense, loop closure detection algorithms have tried to solve this problem with the aim of correcting the accumulated drift in vision-based navigation systems by applying different place recognition techniques. As shown in Fig 1.2, loop closure detection and place recognition methods based on visual information can be classified into three main groups, as stated in [Williams et al., 2009]:

- ***Map-to-map:*** A correlation between the features in two maps is carried out by considering their appearance and their relative positions. This approach is applied in visual metric localization. One of the most known references for loop closure detection using a map-to-map strategy was proposed in [Clemente et al., 2007].
- ***Image-to-map:*** The features extracted from the latest image captured by the camera are matched to the previous features, which are represented in a stored map. This approach is applied in visual topometric localization. One of the most known references for loop closure detection using an image-to-map strategy was proposed in [Williams et al., 2008].
- ***Image-to-image:*** A correspondence between the latest image captured by the camera and the previous ones is achieved, where only visual appearance is necessary (without any metric map or relative positions information). This approach is applied in visual topological localization. One of the most known references for loop closure detection using an image-to-image strategy was proposed in [Cummins and Newman, 2008b], who named the method Fast Appearance-Based MAPping (FAB-MAP). The idea associated with image-to-image place recognition jointly with the definition of the algorithm of FAB-MAP meant a revolution in the robotics community for visual topological localization and became the most popularized methodology for loop closure detection by analyzing the space of visual appearance.

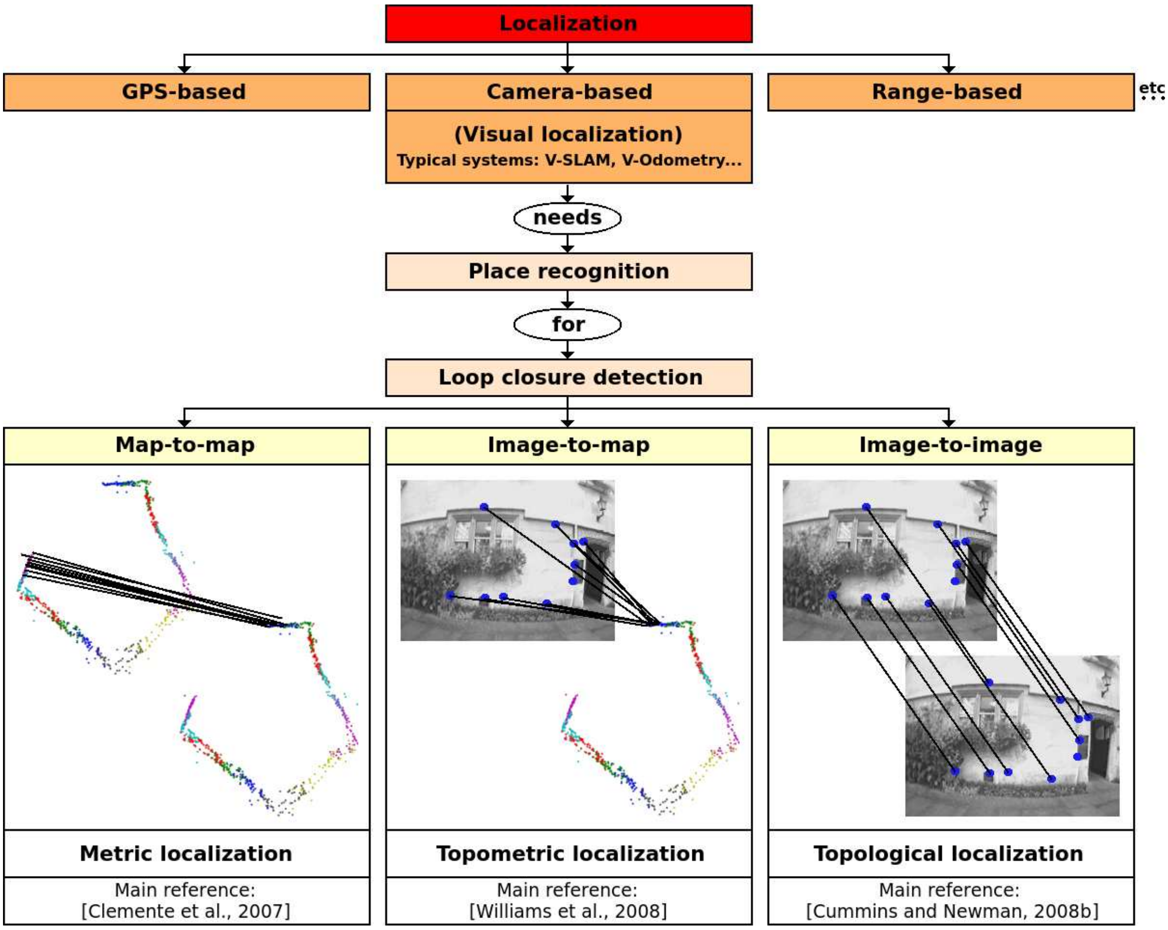


Figure 1.2: Classification of typical place recognition and loop closure detection methods.

Topological place recognition methods have continued evolving since the publication of FAB-MAP and it has become a very active research area, as studied in surveys such as the presented in [Fidalgo and Ortiz, 2015]. The proliferation of these topological techniques during the last years is due to their benefits in the abstract representation of the environment, which is simple, compact, scalable and requires less space to be stored than metric maps. Topological maps represent locations by means of a graph, where places are defined as nodes related among them. For these reasons, topological place recognition is an alternative to metric models in localization problems, which can be also a complement in hybrid approaches where topological information is useful for the correction of metric measurements and for achieving a more detailed scene understanding.

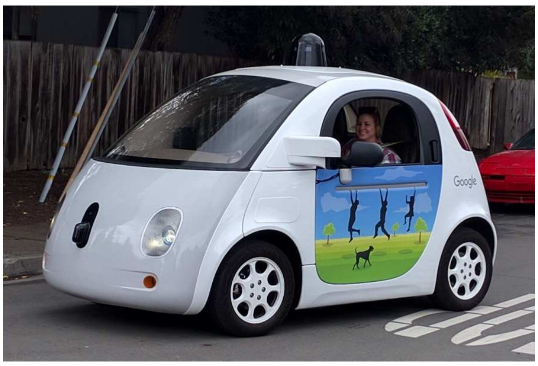
Inspired by the success of topological approaches in short-term situations, the present thesis is focused on the definition of new place recognition methods that suppose an improvement in effectiveness and efficiency with respect to the current state-of-the-art proposals in a life-long localization context. Traditional solutions as the proposed by FAB-MAP decrease their accuracy in long-term conditions due to the changes in visual appearance suffered by places, which is demonstrated in works such as [Milford and Wyeth, 2012]. The main objective of this dissertation is to study how to solve these problems with the aim of procuring a robust life-long visual localization for autonomous navigation.

1.2 Motivation and Goals

The dream of obtaining fully autonomous mobile robots and self-driving cars is one of the main motivations of this thesis. Intelligent vehicles are currently being tested in controlled conditions with promising results, such as the Google Self-Driving Car [Google, 2017] depicted in Fig. 1.3.2. Nevertheless, it is still hard to predict a date when the provided solutions will be safe enough to be used without any human supervision and under any circumstance, because several difficulties and dilemmas must be solved before in order to guarantee the physical integrity of all the people involved in a driving situation, apart from the adaptation of the society to these new technologies. Autonomous driving implies a high complexity due to the influence of a big amount of external variables, so the correct perception of this information is a key aspect to achieve the goal of a driverless car.

Perception systems for intelligent vehicles and robots consist of different sensors that provide them the situational awareness of the environment. The evolution of solutions based on GPS is probably one of the most relevant in the last years and it supplies a valuable information for localizing the position of a vehicle [Skog and Händel, 2009]. Besides, perception approaches using Radio Detection And Ranging (RaDAR) or Light Detection And Ranging (LiDAR) can help to define detailed 3D maps of the scene [Wolcott and Eustice, 2014]. Unfortunately, the acquisition of huge mapping data and the application of GPS information is not always enough to identify the place where an autonomous vehicle or a mobile robot is navigating, because of the constant need of precise map updates, the loss and limited accuracy of GPS signal in some cases and the dynamic changes on the scene. According to this, the visual information processed by camera-based solutions can be decisive in order to enhance the recognition of the features that characterize a place by means of visual localization techniques, which can be complementary to GPS-based proposals, or even supplementary in environments where GPS signal is not completely available or denied [Lategahn and Stiller, 2014]. In fact, novel intelligent vehicles are trusting in vision-based systems using deep learning for navigation tasks, such as the Nvidia Autonomous Car [Bojarski et al., 2016] showed in Fig. 1.3.1.

(1.3.1) Nvidia Auton. Car [Bojarski et al., 2016].



(1.3.2) Google Self-Driving Car [Google, 2017].

Figure 1.3: Some recent examples of intelligent vehicles using camera-based approaches.

Computer vision advances can be applied for varied tasks in camera-based technologies for mobile robots and intelligent transportation systems. In this regard, autonomous vehicles are using visual information in recent research for diverse topics, such as Advanced Driver Assistance Systems (ADAS) [Daza et al., 2014], smartphone applications for safe driving [Almazán et al., 2013, Bergasa et al., 2014, Romera et al., 2015], data mining and analytics for driver behavior analysis [Romera et al., 2016], vehicle detection [Yebes et al., 2014], identification of vehicle's characteristics [Llorca et al., 2013] or teleoperated robots and intelligent vehicles for rescue missions [Cela et al., 2013b, Cela et al., 2013a, Molinos et al., 2013], among others. In the case of the present dissertation, the goal is focused on the use of computer vision techniques in the topic of visual localization for robots and vehicles, with the aim of contributing new proposals which can be useful for long-term navigation and help to make the dream of autonomous driving a reality in a near future. Motivated by these purposes, this thesis defines some specific objectives that will be studied and described in detail along this document:

- Defining a new topological place recognition method based on binary hand-crafted image features. The proposal aims to improve the performance achieved by the current state-of-the-art algorithms in long-term localization and to reduce the computational complexity needed to identify locations by exploiting the advantages of the efficient matching and description of images provided by binary features. This model can be benefited by the application of different cameras such as monocular, stereo or panoramic. In this sense, the objective is to take advantage of the specific information provided by each type of camera, with the aim of adapting the algorithm to the best conditions for each case.

- To develop a new topological place recognition method based on features extracted by means of supervised deep learning techniques. The goal of this approach is to increase the distinctiveness of the features with respect to traditional descriptors for life-long visual localization problems. Deep learning helps to improve the precision in these cases by generating more detailed visual representations of places, which enhance the accuracy in locations matching. However, the disadvantage is that computational costs are also incremented, so feature compression and fast matching solutions must be also designed in order to reduce the impact in memory consumption and processing times.

- To present some applications for our topological place recognition proposals in specific problems typically faced in a life-long visual localization, such as loop closure detection, the correction of visual SLAM and VO measurements, the representation of large-scale 3D models of the environment or the detection of changes on scene along the time course.

1.3 Challenges in Life-Long Visual Localization

In order to completely understand the significance of this dissertation, it is necessary to introduce the main challenges associated with the idea of accomplishing a life-long visual localization, which is one of the principal topics currently studied by the research community in the area of mobile robots and intelligent vehicles navigation.

Some recent works have analyzed the impact of long-term problems in visual localization, where camera configurations and environmental conditions are critical to obtain a robust performance [Bansal et al., 2014]. Besides, other important issues that must be taken into account are the consumption of computational resources and memory costs [Labbé and Michaud, 2011, Nguyen et al., 2013], which should be reduced in order to make possible a long-term operation [Labbé and Michaud, 2013].

However, the most challenging problem in life-long visual localization is related to the extreme changes that the appearance of places experiences throughout the day [Johns and Yang, 2013b], the months [Johns and Yang, 2014] or even across the seasons [Sünderhauf et al., 2013a]. In this sense, the recognition of places over seasonal changes is probably the open problem more studied within the recent past [Lowry et al., 2016], where illumination [McManus et al., 2014], weather [Milford and Wyeth, 2012], dynamic elements on scene [Johns and Yang, 2013a] or changes in the field of view [Lowry and Milford, 2015] also have a great influence in the difficulty to find a robust solution.

All these challenges are faced in this thesis, where the proposed methods are designed to be efficient in a life-long operation and effective in the problematic cases typically derived from the changing appearance of places in long-term, whose most representative examples are shown in Fig. 1.4 and Fig. 1.5.

(1.4.1) At different times of day.

(1.4.2) At different times of day and night.

(1.4.3) Along the months.

(1.4.4) Across the seasons.

Figure 1.4: Temporal examples of challenging cases of changing appearance in locations. The image samples correspond to several tested datasets, which are described in detail in Section 2.5.

(1.5.1) Perceptual aliasing between different places.

(1.5.2) Situations with reduced visibility.

(1.5.3) Partial camera occlusions.

(1.5.4) Changes in the field of view.

(1.5.5) Dynamic elements such as vehicles.

(1.5.6) Constructions.

Figure 1.5: Other specific challenging cases of life-long place recognition. The image samples correspond to several tested datasets, which are described in detail in Section 2.5.

1.4 Organization of the Dissertation

The remainder of this dissertation is organized as follows: The state of the art is initially discussed in Chapter 2. Then, the methods contributed for visual place recognition are described. In particular, Chapter 3 explains our multi-camera proposal based on hand-crafted binary features suitable for a localization with both low computational costs and memory resources, with the aim of facilitating an efficient life-long operation. Chapter 4 presents our method for recognizing locations based on features extracted by means of deep learning, in order to obtain a better performance in long-term situations that are problematic for solutions relying on traditional image descriptors. Both chapters contain a wide set of experiments and results validating the different characteristics of the introduced approaches and comparing them against the main state-of-the-art algorithms in varied challenging datasets, apart from some conclusions about the advantages and disadvantages of both methods depending on the required effectiveness or the needed efficiency in cases where computational and memory resources are limited. After that, Chapter 5 analyzes the applicability of the defined topological place recognition proposals in some typical life-long visual localization problems, such as loop closure detection, VO and SLAM correction, large-scale 3D reconstructions or the detection of changes on scene. Finally, Chapter 6 reports the main conclusions and contributions of this thesis, jointly with the future work directions in the studied research area.

Chapter 2

Related Work

The knowledge about the related work carried out in the research line followed by this thesis is crucial. It allows to comprehend the contributions presented to the community by the techniques proposed in this dissertation to achieve an efficient and effective life-long visual localization based on topological place recognition. According to this, the evolution of computer vision algorithms applied to autonomous vehicles and robots navigation in the state of the art needs to be understood to assimilate the possibilities of camera-based systems in these tasks.

More specifically, the goal of this chapter about state-of-the-art proposals is focused on providing a detailed overview about the classic approaches and current tendencies in the recognition of locations by means of camera sensors. Traditionally, the designed solutions have trusted on the description of the places represented in the images using hand-crafted features, which were usually characterized by researchers attending to different pre-defined image operations that extract some key visual information: points of interest [Mikolajczyk and Schmid, 2004, Aanaes et al., 2011], global image descriptions [Oliva and Torralba, 2001, Oliva and Torralba, 2006] and specific features related to gradients [Dalal and Triggs, 2005], textures [Ojala et al., 1996] or image intensities [Calonder et al., 2010], among others. In fact, the application of these kinds of hand-crafted descriptors for topological place recognition was broadly extended after the seminal publication of FAB-MAP [Cummins and Newman, 2008b], which combines Bags of Visual Words (BoVW) [Sivic and Zisserman, 2003, Lazebnik et al., 2006] jointly with scale-invariant features [Bay et al., 2006, Bay et al., 2008]. During the subsequent years, the research in this field followed similar models commonly based on varied hand-crafted features, as deduced from recent reviews [Fidalgo and Ortiz, 2015, Lowry et al., 2016]. However, works such as [Sünderhauf et al., 2013a] have evidenced some of the limitations of hand-crafted descriptors in life-long visual localization over datasets of more than 3000 km.

For these reasons, the rise of deep learning techniques in computer vision within the last years has become an alternative with respect to traditional image descriptors. This proliferation is mainly due to the extension of the concept of Convolutional Neural Networks (CNNs), which is currently one of the main trending topics in the computer vision community. In this case, the processed features are automatically learned by the network, which typically provides more accurate image descriptions than using hand-crafted features. Initially, CNNs were used for image description in object classification problems with a remarkable precision [Krizhevsky et al., 2012]. After their success, CNNs have been started to be employed in a wide range of camera-based problems. In this regard, visual localization is also a topic where they can be now extensively exploited and improved to obtain enhanced results in long-term navigation.

Nevertheless, deep learning can not be always considered as the most adequate solution for any case or situation. Apart from the required previous training, the computational costs can be also a barrier in specific problems or systems where resources are more limited (i.e., smartphones). Due to this, current dilemmas in this topic are in part focused on reducing memory and processing costs related to CNNs [Ros et al., 2016], without discarding the use of hand-crafted descriptors in embedded systems where localization must be implemented in restricted or low-cost platforms. These dilemmas will be discussed in detail along this state-of-the-art study within a life-long visual localization context.

2.1 Computer Vision applied to Autonomous Vehicles and Robots Localization

Computer vision is one of the most powerful tools for helping autonomous vehicles and robots to enhance their situational awareness, with the aim of facilitating localization and navigation tasks. In this sense, vision-based systems for Unmanned Ground Vehicles (UGVs) allow to perceive characteristics of the environment that other sensors can not notice. At this point, it must be noted that although this dissertation is focused on UGVs, researchers are also working in similar visual localization models for other types of vehicles, such as Unmanned Aerial Vehicles (UAVs) [Pérez-Grau et al., 2016] or Unmanned Underwater Vehicles (UUVs) [Hong et al., 2016], as shown in the examples of Fig. 2.1.

(2.1.1) Example of UAV [Pérez-Grau et al., 2016].

(2.1.2) Example of UUV [Hong et al., 2016].

Figure 2.1: Unmanned aerial and underwater vehicles using camera-based localization.

In any case, the progress of computer vision algorithms for visual SLAM is probably one of the most important achievements in the state of the art in order to understand the evolution of camera-based localization for autonomous navigation. Visual SLAM allows to establish the position of a moving camera and develop a representation of the explored zone at the same time. Some popular state-of-the-art methods were presented during the late 2000s decade for trying to solve this problem by using visual information: MiniSLAM [Andreasson et al., 2007], MonoSLAM [Davison et al., 2007], StereoSLAM [Paz et al., 2008], FrameSLAM [Konolige and Agrawal, 2008, Konolige and Bowman, 2009], RatSLAM [Milford and Wyeth, 2008], GraphSLAM [Eade and Drummond, 2008], TopologicalSLAM [Angeli et al., 2008b], etc. More recently, ORB-SLAM [Mur-Artal et al., 2015] has supposed a new attainment in this field, providing an open-source SLAM library for monocular and stereo cameras with loop closure detection and relocalization capabilities in real-time. Some of these visual SLAM proposals typically use for map optimization a technique named Parallel Tracking And Mapping (PTAM) [Klein and Murray, 2007], which is also commonly applied in other similar problems such as Structure from Motion (SfM) [Bronte et al., 2014]. The movement of the camera along the time is also useful in localization based on VO, where the analysis of the ego-motion between consecutive images can determine the current position and orientation, which is facilitated by public libraries such as LIBVISO [Kitt et al., 2010, Geiger et al., 2011].

Several computer vision techniques can also collaborate to improve the scene understanding required by intelligent vehicles and robots for visual localization using 3D representations. According to this, some works have developed interesting proposals for large-scale dense 3D reconstructions in driving environments by using only cameras [Alcantarilla et al., 2013a] or a combination of cameras and LiDAR [Miksik et al., 2015]. In addition, this 3D data can be enriched by means of semantic segmentation approaches to identify specific elements on scene [Vineet et al., 2015, Sengupta and Sturgess, 2015, Wolf et al., 2015]. Some of these methods combine the visual information with 3D data obtained by other sensing technologies to enhance the semantic segmentation [Zhang et al., 2015]. In some recent cases, deep learning is applied to acquire a more precise segmentation in 2D models based on convolutional architectures [Badrinarayanan et al., 2017]. Moreover, object detection is a related topic in road perception which has also been studied within the last years [Cadena et al., 2015]. All these computer vision algorithms described for scene understanding can be used in order to design a robust visual perception paradigm for autonomous driving [Ros et al., 2015].

As deduced from some of the references previously described, it is important to remark that all the visual information provided by these computer vision methods for localization can be enhanced by fusing it with the data acquired by other sensors. In fact, although the major part of the state-of-the-art works for visual place recognition only use vision-based perception, some recent proposals have fused 3D LiDAR and camera imagery to strengthen the performance in the identification of locations [Pandey et al., 2014].

2.2 Visual Place Recognition based on Hand-Crafted Features

Methods based on hand-crafted features have been traditionally the most commonly chosen option in visual place recognition literature. In this section, the main state-of-the-art works related to this research line are presented, jointly with a review of the descriptors more typically used in these cases.

2.2.1 Hand-Crafted Descriptors

Hand-crafted features are acquired by means of image operations that are well-engineered to particular problems in image matching. On the one hand, local approaches are focused on detecting some points or regions of interest in the image for extracting only the features related to these specific zones. On the other hand, global descriptors directly obtain the features for the whole image in a more simple way, without any previous detection process.

Moreover, hand-crafted descriptors are usually classified depending on the format of their features. Typically, they were stored as float values in vectors that compose the final descriptor. However, binary features have been recently popularized because of their simplicity and efficiency, which are based on a bit-level computation.

According to the previous considerations, descriptors can be generally grouped as local or global and vector-based or binary, as graphically depicted in Fig. 2.2.

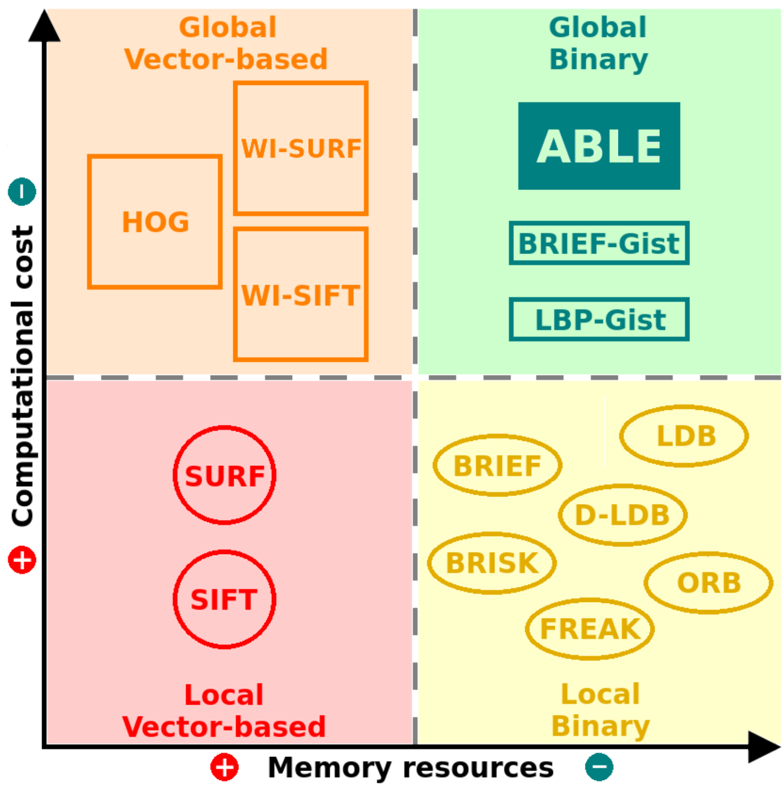


Figure 2.2: Qualitative classification of typical hand-crafted descriptors. It includes some of the most popular features commonly used in visual place recognition, which are grouped in different categories: Vector-based vs Binary / Local vs Global.

Recent works have compared some of these features for visual SLAM problems [Hartmann et al., 2013] and life-long localization [Krajník et al., 2013]. In addition, whole-image or global descriptors are also lately studied to evaluate their performance in visual loop closure detection [Liu and Zhang, 2013]. We are going to introduce the most relevant descriptors presented in Fig. 2.2, because some of them are used in different analyses within the visual place recognition methods implemented in this dissertation.

2.2.1.1 Local Vector-based Descriptors

Commonly, vector-based descriptors have applied different local feature detectors for the subsequent extraction of features. State-of-the-art algorithms for visual place recognition based on local descriptors typically detect points or regions of interest in the image by means of techniques such as Maximally Stable Extremal Regions (MSER) [Matas et al., 2002], Features from Accelerated Segment Test (FAST) [Rosten and Drummond, 2006] or the STAR detector, derived from Center Surround Extrema (CenSurE) [Agrawal et al., 2008]. After detecting these zones, a feature extractor must be computed to obtain the final vector containing the descriptor. In this regard, there are two popularized methods that have been extensively used for local vector-based description, including the detection and extraction of the features:

- ***Scale Invariant Feature Transform (SIFT) [Lowe, 1999, Lowe, 2004]:*** This approach detects keypoints by searching for extremas of a function based on a difference of Gaussians at different scales. The orientation of the detected keypoints is estimated by means of the local image gradient. A feature descriptor is then derived from the local image gradients in a region defined by the scale and orientation of the keypoint. The descriptor is normalized and some thresholds are taken into account to limit the effects of image saturation. This configuration allows to obtain an invariance to scale, orientation and basic illumination changes. Some refined versions of the algorithm have been presented in the subsequent years, such as PCA-SIFT [Ke and Sukthankar, 2004]. In addition, 3D SIFT has been proposed for tridimensional computer vision problems [Scovanner et al., 2007].

- ***Speeded Up Robust Features (SURF) [Bay et al., 2006, Bay et al., 2008]:*** This method is in part inspired by SIFT, but with the goal of improving its processing costs. In this case, the detection of keypoints is based on an approximation of the determinant of a Hessian blob detector at different scales by means of integral images [Viola and Jones, 2004]. Features are extracted by pooling first order Haar wavelets around the points of interest. This proposal is also invariant to scale and rotation. Moreover, a SURF version for 3D descriptions was also suggested in [Knopp et al., 2010].

Although SIFT and SURF are the most well-known examples of local vector-based descriptors, computer vision libraries such as OpenCV [Bradski, 2000] have recently extended the usage of other similar 2D features, such as KAZE [Alcantarilla et al., 2012] or A-KAZE [Alcantarilla et al., 2013b], that operate in a nonlinear scale space increasing detector repeatability and extractor distinctiveness. Additionally, it must be noted that there are also more options for 3D description derived from the Point Cloud Library (PCL) [Rusu and Cousins, 2011] that have been used in some specific place recognition problems, including Point Feature Histograms (PFH) [Rusu et al., 2008b, Rusu et al., 2008a], Fast Point Feature Histograms (FPFH) [Rusu et al., 2009], Viewpoint Feature Histogram (VFH) [Rusu et al., 2010], Normal Aligned Radial Feature (NARF) [Steder et al., 2010] or SURface Entropy for 3D features (SURE) [Fiolka et al., 2012].

2.2.1.2 Local Binary Descriptors

Binary descriptors have become a great alternative with respect to vector-based approaches in the recent past. This is due to the efficient properties of these kinds of binary features, which need less memory resources to be stored and can be fast matched by using the Hamming distance [Muja and Lowe, 2012].

The computation of local binary descriptors requires a previous detection process, but the extraction of the features is usually easily computed. Normally, they are calculated using a set of pairwise intensity comparisons from a sampling pattern centered in a point of interest, where each bit in the descriptor is the result of exactly one comparison. In this thesis, some binary descriptors are applied in different contexts related to our topological place recognition techniques. The most important are mainly the following:

- ***Binary Robust Independent Elementary Features (BRIEF) [Calonder et al., 2010, Calonder et al., 2012]:*** This is the initial proposal for local binary description derived from intensity difference tests. The descriptor is formed by computing these tests on a number of randomly selected but fixed location pairs in a patch around a keypoint. These features are not invariant to rotation and scale.

- ***Oriented fast and Rotated BRIEF (ORB) [Rublee et al., 2011]:*** This approach is also based on intensity comparisons, but it includes a method of orientation estimation focused on the FAST detector that provides independence to rotation. Thanks to these improvements, ORB solves the lack of rotation invariance associated with BRIEF and improves the resistance to noise.

- ***Binary Robust Invariant Scalable Keypoints (BRISK) [Leutenegger et al., 2011]:*** An improved method for detecting the keypoints at different scales is designed. The binary tests are computed on a circular pattern around each keypoint. Intensity values are smoothed and the keypoint orientation is estimated using pairs separated by a long distance. This gives invariance to scale and rotation.

- ***Fast REtinA Keypoint (FREAK) [Alahi et al., 2012]:*** This binary descriptor imitates the organization of the retina, employing a circular grid where receptive fields of different sizes are considered. The difference in intensity between the pairs of the receptive fields is calculated. Besides, the sampling regions are overlapped in order to add redundancy with the aim of enhancing the discriminative power.

- ***Local Difference Binary (LDB) [Yang and Cheng, 2012, Yang and Cheng, 2014]:*** These features use not only intensity information, but also gradient difference tests at several image granularities. In this regard, the descriptor calculates the features in multiple resolutions using varied grids (2x2, 3x3, 4x4, etc), which alleviates the dependence on the field of view. In the work derived from this thesis, an improved version named Disparity Local Difference Binary (D-LDB) [Arroyo et al., 2014a] is developed to add disparity information when stereo cameras are available, with the aim of increasing the performance by obtaining geometric information of the environment.

2.2.1.3 Global Vector-based Descriptors

This family of descriptors does not require a previous keypoint detection process, so the features are directly computed using the whole image and stored in a vector of float values. The most representative hand-crafted descriptors in the state of the art based on global vector-based features that have been tested along this dissertation are the following:

- ***Histogram of Oriented Gradient (HOG) [Dalal and Triggs, 2005]:*** In this descriptor, local gradients are binned in accordance with orientation and weighted depending on their magnitude, within a spatial grids of cells with overlapping blockwise contrast normalization. For each overlapping block of cells, a feature vector is obtained by sampling the histograms from the contributing spatial cells. The feature vectors for all overlapping blocks are concatenated to produce the final descriptor. Improved versions of the algorithm such as CHOG [Chandrasekhar et al., 2012] have been recently studied, as well as 3D HOG [Kläser et al., 2008]. In some problems, the fusion of global vector-based descriptors can greatly improve the performance, as proposed in [Arroyo et al., 2015c], where HOG is combined with other features such as Local Binary Patterns (LBP) [Ojala et al., 1996] and Global Color Histogram (GCH) [Novak and Shafer, 1992]).

- ***Whole Image SIFT (WI-SIFT) [Badino et al., 2012]:*** In this case, a standard SIFT descriptor is calculated over the whole image, without detecting any previous keypoint. Due to this, this descriptor only requires one feature per image, reducing the memory and processing costs with respect to the original SIFT.

- ***Whole Image SURF (WI-SURF) [Badino et al., 2012]:*** Similar to WI-SIFT but using SURF features.

2.2.1.4 Global Binary Descriptors

Global binary descriptors are a type of features that has a great computational efficiency, because it combines the advantages of whole image description techniques with the simplicity of binary representations. Furthermore, some state-of-the-art works for place recognition have obtained remarkable results using these kinds of descriptors, such as BRIEF-Gist [Sünderhauf and Protzel, 2011], which calculates a global BRIEF descriptor based on the Gist of scenes [Oliva and Torralba, 2001, Oliva and Torralba, 2006]. Other similar methods are derived from the Gist of different local features, such as LBP-Gist [Campos et al., 2013] (based on LBP descriptors) or the Gabor-Gist algorithm [Liu and Zhang, 2012]. In addition, it must be noted that our approach for visual place recognition based on hand-crafted features trust in a global binary descriptor that uses LDB as core, as will be explained in detail in Chapter 3.

2.2.2 Related Works based on Hand-Crafted Features

Although the concept of topological localization based on appearance had been previously proposed by other researchers in the past [Ulrich and Nourbakhsh, 2000], the influence of FAB-MAP [Cummins and Newman, 2008a, Cummins and Newman, 2008b] can be considered as the point of reference in the state of the art of visual place recognition and loop closure detection based on hand-crafted features since the moment of its release until today. In addition, the authors of FAB-MAP also tested an improved version over 1000 km in more recent papers [Cummins and Newman, 2010a, Cummins and Newman, 2010b], including a subsequent research where it is combined with the RatSLAM algorithm [Glover et al., 2010]. A 3D implementation of FAB-MAP [Paul and Newman, 2010] was also developed in order to incorporate geometric information, but in this case it was only evaluated for short-term localization. Besides, an open-source toolbox named OpenFABMAP [Glover et al., 2012] was contributed to the research community for testing the method and using it in a free way.

Nevertheless, FAB-MAP needs a prior training phase and applies a computationally expensive approach that requires feature extraction followed by probabilistic inference, which can make the proposal not suitable for real-time applications. Due to these issues, a great number of novel methods based on hand-crafted features have been lately presented by following the research line started by FAB-MAP, with the aim of improving the state-of-the-art results. A great part of them are also based on similar BoVW models combined with SURF [Mei et al., 2010, Maddern et al., 2011, Maddern et al., 2012, Nicosevici and Garcia, 2012, Labbé and Michaud, 2011, Labbé and Michaud, 2013] or SIFT features [Angeli et al., 2008b, Angeli et al., 2008a, Angeli et al., 2009, Zhang et al., 2010, Zhang, 2011, Mei et al., 2010, Mei et al., 2011]. In other cases, BoVW are based on typical binary features such as BRIEF [Gálvez-López and Tardós, 2011, Gálvez-López and Tardós, 2012] or in novel description techniques such as Position-Invariant

Robust Features (PIRF) [Kawewong et al., 2011, Khan et al., 2012]. Besides, relocalization techniques focused on the classification of features have also been proposed for SLAM problems [Williams et al., 2011].

Another approach commonly used in the last years is the application of global image descriptors, in order to achieve an efficient long-term performance and try to obtain a real-time visual localization. Solutions similar to the proposed by BRIEF-Gist [Sünderhauf and Protzel, 2011] are the most representative of this tendency. In this sense, other works directly apply Gist descriptors in epitomic image analysis for indoor localization [Ni et al., 2009]. More recently, a method based on global image signatures has also been published for visual loop closure detection [Negre-Carrasco et al., 2016], which demonstrates the proliferation of these kinds of techniques.

One of the most relevant proposals contributed for topological localization within the recent past is SeqSLAM [Milford and Wyeth, 2012], that introduced the idea of recognizing places as sequences of images instead of single images, in contrast to previous proposals such as FAB-MAP and similar works. It uses a global description method over patch-normalized sequences of images that is named Sum of Absolute Differences (SAD). SeqSLAM was satisfactorily evaluated in long-term conditions, where a same route was traversed in a sunny summer day and a stormy winter night. However, in [Sünderhauf et al., 2013a] some drawbacks of SeqSLAM were revealed, such as dependence in the field of view and the influence of parameters configuration. These problems have been ameliorated in newer papers that study the difficulties of a changing viewpoint [Milford, 2013, Milford et al., 2014, Pepperell et al., 2014, Lowry and Milford, 2015].

Finally, although a great part of the algorithms based on hand-crafted descriptors in the state of the art are designed for monocular cameras, there are some specific approaches that are focused on stereo and panoramic images. On the one hand, stereo information allows to acquire a more complete description of the geometry of an environment, which is exploited in works such as [Cadena et al., 2010, Cadena and Neira, 2011, Cadena et al., 2012], where a BoVW model is combined with the usage of stereo pairs to check a valid spatial transformation in place matching. Besides, the 3D data that can be obtained by means of stereo cameras is also applied in different techniques used for identifying places in indoor environments [Fiolka et al., 2013, Scherer et al., 2013]. On the other hand, some state-of-the-art algorithms trust in panoramic images for a more robust topological localization [Valgren and Lilienthal, 2010, Murillo and Kosecká, 2009, Singh and Kosecká, 2010, Murillo et al., 2013, Korrapati et al., 2013, Korrapati and Mezouar, 2016]. The main advantage of panoramas is that they allow a visual perception of the environment in all the possible orientations, which can be used for detecting places revisited in other direction.

In order to facilitate the understanding of the evolution in the state of the art related to visual place recognition using hand-crafted features, Tables 2.1 and 2.2 can be revised to have a more detailed idea about the main works and their specific characteristics.

2.3 Visual Place Recognition based on Learned Features

Nowadays, features obtained by means of deep learning techniques are being started to be studied for solving visual place recognition approaches. Here, the most important proposals corresponding to the state of the art in this research field are explained and discussed, jointly with a brief review about CNNs and the features derived from them.

2.3.1 Convolutional Neural Networks

The main reasons why deep learning is very successful these days are the availability of powerful General-Purpose computing on Graphics Processing Units (GPGPU), larger datasets with annotations and better training strategies. In addition, CNNs have revolutionized the computer vision community, mainly due to the innovative work presented by [Krizhevsky et al., 2012], which defined one of the most relevant CNN architectures: AlexNet. This network obtained impressive results in image classification over the challenging ImageNet dataset [Deng et al., 2009], which contains millions of image samples.

After that, other works tested the power of CNNs by designing new refined architectures based on AlexNet and widely comparing them against other visual recognition methods [Chatfield et al., 2014]. Apart from this, the benefits of CNNs have been exhibited in a different range of typical computer vision problems such as semantic segmentation [Long et al., 2015], change detection [Alcantarilla et al., 2016, Kataoka et al., 2016] or optical flow [Dosovitskiy et al., 2015], among others. Currently, the most recent works in the community are also carrying out a great research in different deep learning dilemmas: networks compression [Han et al., 2016], unsupervised learning [Isola et al., 2016], training-free methods [Tolias et al., 2016], CNNs initialization [Krahenbuhl et al., 2016] and questions such as how deep should be the networks [Urban et al., 2016].

The popularization and extension of these deep learning algorithms in varied contexts has also been possible thanks to the useful open toolboxes provided by some authors to the community, such as Caffe [Jia et al., 2014], Torch [Collobert et al., 2011] or MatConvNet [Vedaldi and Lenc, 2015]. Besides, other contributions helped researchers to better understand the complex division into layers behind CNNs and to visualize their features [Zeiler and Fergus, 2014].

2.3.2 CNN-based Descriptors

The application of CNNs to learn robust visual descriptors has been studied in works such as [Razavian et al., 2014], where features are processed for global image description in a diverse range of recognition tasks. Moreover, other proposals extract CNN features over a set of detected points [Simo-Serra et al., 2015], like traditional hand-crafted local descriptors such as SURF.

Apart from general image recognition, there are also approaches where deep descriptors are trained over a specific database of places [Zhou et al., 2014], with the aim of categorizing concrete place scenes such as forests, coasts, rooms, etc. On the other hand, some works analyze how transferable is the knowledge acquired in training by the features derived from deep neural networks [Oquab et al., 2014, Yosinski et al., 2014], which demonstrates that learned features appear not to be specific to a particular dataset or problem, they can be generalizable to several datasets and problems. In this regard, features pre-trained in image classification problems are an interesting option for generalization in vision-based localization problems, as will be demonstrated in our architecture described in detail in Chapter 4.

2.3.3 Related Works based on Learned Features

Methods based on CNN features can be a promising alternative for a more precise place recognition in life-long localization. In fact, recent research is starting to analyze the application of these techniques in problems related to the recognition of locations. One of the first approaches in the state of the art is the work presented by [Chen et al., 2014], where several experiments about the performance of the features from different layers are conducted for place recognition in environments with significant viewpoint changes.

Recent papers exhibited more studies about possible utilities of pre-trained CNN descriptors in visual topological localization [Sünderhauf et al., 2015b, Sünderhauf et al., 2015a]. These works are based on features derived from AlexNet, which have been previously trained in the ImageNet dataset for image classification tasks. They evidenced that the descriptors related to intermediate layers (especially from the third convolution) are more robust against appearance changes than the features from any other layer.

Moreover, the approach defined in [Arandjelovic et al., 2016] performs end-to-end learning of a CNN for identifying locations by introducing a novel differentiable layer based on the Vector of Locally Aggregated Descriptors (VLAD) [Korrapati and Mezouar, 2016]. The proposed architecture is trainable in a direct manner for the place recognition task.

Other interesting models based on deep learning are focused on pose regression for relocalization in small-scale environments [Kendall et al., 2015], in contrast to works where images are matched under substantial appearance changes exploiting GPS metric priors [Vysotska et al., 2015].

Obviously, the amount of publications associated with the topic of visual place recognition based on CNN-features is lower than the described for methods focused on hand-crafted descriptors, because this is a much more recent research field and the proposals presented until today are currently introducing these deep learning techniques in the visual localization context. However, in the chronological evolution listed in Table 2.2 it can be seen how CNN-based approaches are experiencing a great rise in the state of the art within the last two years.

2.4 Life-Long Visual Localization using Topological Place Recognition

Nowadays, the topic of visual localization in long-term scenarios is one of the most broadly studied by the research community in this area. The recognition of previously visited places along the different seasons of the year is currently the most challenging topic in life-long topological localization due to the difficulties associated with this task: changes in vegetation, illumination, weather, dynamic elements, etc. For this reason, several proposals have been presented within the recent past to face these problems related to seasonal changes. To the best of our knowledge, the first researches in visual place recognition across seasons were performed in [Valgren and Lilienthal, 2008, Valgren and Lilienthal, 2010], where panoramic images were used to carry out a long-term localization in outdoor environments by means of SIFT and SURF descriptors. In similar works where topometric localization is applied, WI-SURF features were employed for global image description in tests carried out along different months [Badino et al., 2011, Badino et al., 2012].

In papers such as [Sünderhauf et al., 2013b, Neubert et al., 2013, Neubert et al., 2015], a novel algorithm for identifying locations based on appearance change prediction was evaluated in the Nordland dataset [Sünderhauf et al., 2013a], where a same route of more than 750 km is traversed by a train four times (one in each season). This dataset has also been employed in other works, such as [Mohan et al., 2015], where co-occurrence matrices are computed to improve the precision for matching places in long-term scenarios. In fact, the effectiveness of co-occurrence for place recognition in dynamic scenes had been previously demonstrated by [Johns and Yang, 2013b, Johns and Yang, 2013a, Johns and Yang, 2014]. Other algorithms focused on seasonal changes are based on visual experiences [Churchill and Newman, 2012b, Churchill and Newman, 2012a, Churchill and Newman, 2013, Linegar et al., 2015, Dymczyk et al., 2015], which are defined as appearance representations of an environment under certain conditions to obtain a visual memory.

Besides, although new tendencies also propose to use CNNs in life-long visual localization, it must be noted that supervised deep learning techniques require a large amount of manually annotated data for a specific problem at hand and they usually are computationally expensive, as studied in very recent methods focused on topological place learning [Erkent and Bozma, 2015] or loop closure detection for visual SLAM based on deep neural networks [Gao and Zhang, 2017]. In this sense, these considerations must be taken into account, because they can be restrictive for some specific life-long visual place recognition problems or applications. Fortunately, these systems can reduce memory resources and computational costs using less complex solutions, such as the computation of simplified image representations. In this line, automatic image scaling can be an interesting idea for achieving more efficiency in place recognition for changing environments, as exposed in [Pepperell et al., 2015]. Moreover, compact scene descriptors have obtained remarkable results in cross-season place recognition, such as [Masatoshi et al., 2015].

With the aim of organizing the most important state-of-the-art publications for life-long visual localization, Tables 2.1 and 2.2 are presented. The main references in visual place recognition described along this chapter are included to justify the evolution of the approaches previously introduced in this area and the motivation of the contributions developed in this thesis. The list is divided into a group corresponding to the proposals appeared between the seminal publication of FAB-MAP and the year before this dissertation was started (Table 2.1: 2008-2012) and a group that incorporates the remaining works proposed until the end of the dissertation (Table 2.2: 2013-2017). The publications associated with the contributions presented in this thesis are also listed to comprehend their impact in the state of the art, including versions recently provided by other authors that are directly derived from our work, such as the described in [Nowicki et al., 2016].

Table 2.1: Chronological evolution of visual place recognition methods (2008-2012). The main references in the state of the art are listed with the aim of classifying them according to their characteristics.

Work		**Image description**				**Camera type**	**Tests scenarios**
Main reference	**Method name**	**BoVW**	**Type**	**Detector**	**Extractor**		
[Cummins and Newman, 2008b]	FAB-MAP	✓	Local Vector	SURF	SURF	Monocular	Outdoors
[Angeli et al., 2008b]	TopologicalSLAM	✓	Local Vector	SIFT	SIFT	Monocular	Indoors
[Angeli et al., 2008a]	IAB-MAP	✓	Local Vector	SIFT	SIFT + LCH	Monocular	Indoors, Outdoors
[Angeli et al., 2009]	Bayesian LCD	✓	Local Vector	SIFT	SIFT + LCH	Monocular	Indoors, Outdoors
[Ni et al., 2009]	Epitome model	×	Global Vector	-	Gist	Stereo, Panoramic	Indoors
[Valgren and Lilienthal, 2010]	-	×	Local Vector	SIFT, SURF	SIFT, SURF	Panoramic	Outdoors
[Paul and Newman, 2010]	FAB-MAP 3D	✓	Local Vector	SURF	3D SURF	Monocular	Outdoors
[Glover et al., 2010]	FAB-MAP + RatSLAM	✓	Local Vector	SURF	SURF	Monocular	Outdoors
[Mei et al., 2010]	DBW	✓	Local Vector	SURF	SURF	Stereo	Outdoors
[Cummins and Newman, 2010b]	FAB-MAP 2.0	✓	Local Vector	SURF	SURF	Monocular, Panoramic	Outdoors
[Kawewong et al., 2011]	PIRF-nav	×	Local Vector	SIFT	PIRF	Monocular, Panoramic	Outdoors
[Zhang, 2011]	BoRF	✓	Local Vector	SIFT	SIFT	Monocular	Indoors
[Sünderhauf and Protzel, 2011]	BRIEF-Gist	×	Global Binary	-	BRIEF	Monocular, Panoramic	Outdoors
[Williams et al., 2011]	RLC	×	Local Vector	FAST	SIFT	Monocular	Indoors, Outdoors
[Mei et al., 2011]	RSlam	✓	Local Vector	FAST	SIFT + SAD	Monocular, Stereo	Outdoors
[Maddern et al., 2012]	CAT-SLAM	✓	Local Vector	SURF	SURF	Monocular, Panoramic	Outdoors
[Badino et al., 2012]	WI-SURF	×	Global Vector	-	SURF	Monocular	Outdoors
[Glover et al., 2012]	OpenFABMAP	✓	Local Vector	STAR, FAST, MSER	SIFT, SURF	Monocular, Panoramic	Indoors, Outdoors
[Milford and Wyeth, 2012]	SeqSLAM	×	Global Vector	-	SAD	Monocular	Outdoors
[Cadena et al., 2012]	BoW-CRF	✓	Local Vector	SURF	SURF, 3D SURF	Stereo	Indoors, Outdoors
[Nicosevici and Garcia, 2012]	OVV	✓	Local Vector	MSER	SIFT, SURF	Monocular	Outdoors
[Gálvez-López and Tardós, 2012]	DBoW	✓	Local Binary	FAST	BRIEF	Monocular	Indoors, Outdoors
[Liu and Zhang, 2012]	Gabor-Gist	×	Global Vector	-	Gabor	Monocular	Outdoors
[Khan et al., 2012]	PIRF-nav 3D	✓	Local Vector	SURF	3D PIRF	Stereo	Outdoors

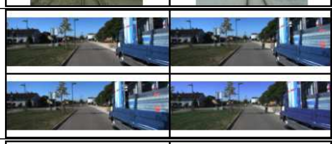
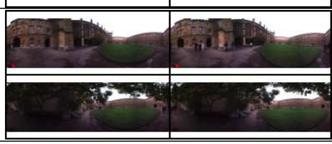
Table 2.2: Chronological evolution of visual place recognition methods (2013-2017). The main references in the state of the art are listed with the aim of classifying them according to their characteristics.

Work		Image description				Camera type	Tests scenarios
Main reference	**Method name**	**BoVW**	**Type**	**Detector**	**Extractor**		
[Murillo et al., 2013]	Gist Panoramas	✓	Global Vector	-	Gist	Panoramic	Outdoors
[Johns and Yang, 2013b]	Cooc-Map	✓	Local Vector	-	SIFT	Monocular	Outdoors
[Sünderhauf et al., 2013a]	OpenSeqSLAM	×	Global Vector	-	SAD	Monocular	Outdoors
[Labbé and Michaud, 2013]	RTAB-Map	✓	Local Vector	SURF	SURF	Monocular	Indoors, Outdoors
[Campos et al., 2013]	LBP-Gist	×	Global Binary	-	LBP	Monocular	Outdoors
[Fiolka et al., 2013]	-	✓	Local Vector	NARF, SURE	NARF, SURE	Stereo	Indoors
[Scherer et al., 2013]	-	✓	Local Vector	FAST, NARF	SIFT, BRIEF, NARF	Stereo	Indoors
[Churchill and Newman, 2013]	-	×	Local Binary	FAST	BRIEF	Monocular	Outdoors
[Johns and Yang, 2014]	STL Geo	✓	Local Vector	SIFT	SIFT	Monocular	Outdoors
[Arroyo et al., 2014b]	ABLE-P	×	Global Binary	-	LDB	Panoramic	Outdoors
[Pepperell et al., 2014]	SMART	×	Global Vector	-	SAD	Monocular	Outdoors
[Arroyo et al., 2014a]	ABLE-S	×	Global Binary	-	D-LDB	Stereo	Outdoors
[Chen et al., 2014]	-	×	CNN-based	-	-	Monocular, Panoramic	Indoors, Outdoors
[Neubert et al., 2015]	SP-ACP	✓	Global Vector	-	SIFT + LCH	Monocular	Outdoors
[Arroyo et al., 2015b]	ABLE-M	×	Global Binary	-	LDB	Monocular	Outdoors
[Lowry and Milford, 2015]	Change removal	×	CNN-based	-	-	Monocular	Outdoors
[Mohan et al., 2015]	WWC	✓	Local Vector	FAST	ORB	Monocular, Panoramic	Indoors, Outdoors
[Neubert et al., 2015]	SP-ACP	✓	Global Vector	-	SIFT + LCH	Monocular	Outdoors
[Sünderhauf et al., 2015b]	Landmarks (conv3)	×	CNN-based	-	-	Monocular	Outdoors
[Sünderhauf et al., 2015a]	AlexNet (conv3)	×	CNN-based	-	-	Monocular	Outdoors
[Mur-Artal et al., 2015]	ORB-SLAM	✓	Local Vector	FAST	ORB	Monocular, Stereo	Indoors, Outdoors
[Kendall et al., 2015]	PoseNet	×	CNN-based	-	-	Monocular	Outdoors
[Korrapati and Mezouar, 2016]	VLAD	✓	Global Vector	-	SURF	Panoramic	Outdoors
[Arandjelovic et al., 2016]	NetVLAD	×	CNN-based	-	-	Monocular	Outdoors
[Arroyo et al., 2016a]	CNN-VTL	×	CNN-based	-	-	Monocular	Outdoors
[Nowicki et al., 2016]	FastABLE	×	Global Binary	-	LDB	Monocular	Indoors
[Arroyo et al., 2016b]	OpenABLE	×	Global Vector, Binary	-	SIFT, SURF, HOG, BRIEF, ORB, BRISK, FREAK, LDB, D-LDB	Monocular, Stereo, Panoramic	Outdoors
[Negre-Carrasco et al., 2016]	HALOC	×	Global Vector	-	SIFT, SURF	Monocular	Indoors, Outdoors
[Arroyo et al., 2017]	ABLE	×	Global Binary	-	LDB, D-LDB	Monocular, Stereo, Panoramic	Outdoors

2.5 Datasets used in Experimentation

Before concluding this chapter, some of the most important public datasets typically applied for experimentation in the state of the art for topological place recognition are described. They are used along this dissertation for varied tests in order to obtain results in several life-long visual localization problems. In this regard, the following datasets have been employed: the St Lucia dataset [Glover et al., 2010], the Alderley dataset [Milford and Wyeth, 2012], the CMU-CVG Visual Localization dataset [Badino et al., 2011,Badino et al., 2012], the Nordland dataset [Sünderhauf et al., 2013a], the KITTI Odometry dataset [Geiger et al., 2012,Geiger et al., 2013] and the Oxford New College dataset [Smith et al., 2009]. For each dataset, Table 2.3 shows the number and type of images which compose the dataset jointly with the number of recorded sequences and their length, a brief description of the dataset with general comments about it, some image samples of revisited places and the map route followed in the sequences. The depicted routes are represented by processing the available GPS measurements, which can be also used as ground-truth.

Table 2.3: Characteristics of the state-of-the-art datasets. It includes the main properties of all the public datasets used in the evaluations carried out along this thesis.

Dataset	No. images	General comments	Samples of revisited places	Map route
St Lucia	218140 in 10 sequences (10x12 km) (640x480 px) (15 fps) (Monocular)	It is collected in the Brisbane suburb of St Lucia, Australia. A route is traversed by a car 10 times at different day hours. GPS positions are logged during each journey.		
Alderley	31567 in 2 sequences (2x8 km) (640x256 px) (25 fps) (Monocular)	Two car rides recorded in a route during a sunny summer day and a stormy winter night in the Brisbane suburb of Alderley, Australia. Matches are manually labeled.		
CMU-CVG Visual Loc.	245792 in 16 sequences (16x8 km) (1024x768 px) (15 fps) (Monocular)	A route is traversed 16 times in Pittsburgh, PA, USA. The sequences are recorded in different months under varying environmental conditions. GPS information is registered.		
Nordland	3576688 in 4 sequences 4x728 km (1920x1080 px) (25 fps) (Monocular)	A train ride recorded in Norway four times, once in every season. Sequences are synchronized and the camera position and field of view are always the same. Available GPS readings.		
KITTI Odometry	44182 in 22 sequences 39,2 km (1226x370 px) (10 fps) (Stereo)	It contains 22 sequences recorded across different car routes around the city of Karlsruhe, Germany. GPS readings are available for 11 of the registered sequences.		There is more than one map. See Appendix B for more information about them and the ground-truth designed for loop closure detection.
Oxford New College	8127 in 1 sequence 2,2 km (2048x618 px) (5 fps) (Panoramic)	It is one sequence captured by a robot at the University of Oxford, UK. Stereo images are also available and a ground-truth matrix can be used for evaluation purposes.		

Chapter 3

Visual Place Recognition based on Hand-Crafted Binary Features

The contribution of a novel solution to the open problem of visual place recognition in long-term localization is the main motivation of the following chapter. Initially, a method based on traditional hand-crafted features is described in order to enhance the performance currently achieved in the state of the art related to this research area, especially in situations where extreme appearance changes over long periods of time are registered in locations.

According to this, we consider to develop a solution that builds upon the concept of binary description of images, due to the beneficial properties that it provides in image recognition. Binary features were popularized in computer vision during the last years because of their simplicity and favorable conditions, such as the low memory requirements needed to store them or the possibility of carrying out a very fast matching by using the Hamming distance, as exposed in [Muja and Lowe, 2012]. For this reason, one of the main advantages of this proposal with respect to vector-based descriptors is the reduction of computational and memory costs without a significant loss of description power. In this sense, works such as [Milford, 2012] have demonstrated that a handful of bits is sufficient for conducting a robust visual route recognition. The goal of the method presented in this chapter is to design an improved approach in this line which can satisfactorily and efficiently operate in a life-long visual localization context.

Along the following pages, we describe our proposed method called Able for Binary-appearance Loop-closure Evaluation (ABLE). Its main characteristics are exposed jointly with the main contributions that are provided to the state of the art and the scientific community. Additionally, the experimental evaluation of our approach is presented in order to validate the properties of ABLE and compare it against the principal state-of-the-art methods in several challenging public datasets typically used for testing in long-term localization problems. A final discussion about the conclusions obtained in this research is given.

3.1 Overview: The ABLE Method

ABLE is a novel method for performing a life-long visual topological localization in a robust manner and trying to maintain the maximum efficiency along the time course. The evolution achieved during the development of our proposal can be seen in the previous work published in different international conferences in computer vision, robotics and autonomous vehicles [Arroyo et al., 2014b, Arroyo et al., 2014a, Arroyo et al., 2015b, Arroyo et al., 2015a, Arroyo et al., 2016b, Arroyo et al., 2017].

The diagram presented in Fig. 3.1 illustrates the main characteristics of the topological place recognition methodology contributed by ABLE, whose main properties are the following:

- Sequences of images are used instead of single images with the aim of carrying out a better recognition of places in long-term scenarios, as introduced in works such as [Milford and Wyeth, 2012].
- An illumination invariant transformation is applied to the input images in a pre-processing step in order to minimize the problems related to changing lighting conditions and shadows in a visual place recognition context, inspired by proposals such as [Upcroft et al., 2014, McManus et al., 2014].
- Global binary features based on LDB [Yang and Cheng, 2012, Yang and Cheng, 2014] are applied in image description jointly with a matching focused on the Hamming distance and an Approximate Nearest Neighbors (ANN) search, that provides both low processing times and high precision rates.
- Different versions of ABLE are proposed depending on the type of camera for providing a higher adaptability and taking advantage of the additional image information that can be obtained in each case: monocular (ABLE-M), stereo (ABLE-S) or panoramic (ABLE-P).

3.2 Image Description of Locations in ABLE

To define a robust methodology for describing the images corresponding to the places analyzed by our system is one of the most important tasks to achieve an effective visual localization in long-term scenarios, with the aim of facilitating the subsequent matching of the processed images. The core of our image description approach is focused on hand-crafted binary features. However, we also include other interesting techniques that help to improve the descriptive power of ABLE in its three versions. In this regard, all these techniques and their interaction in our method are explained in the present section.

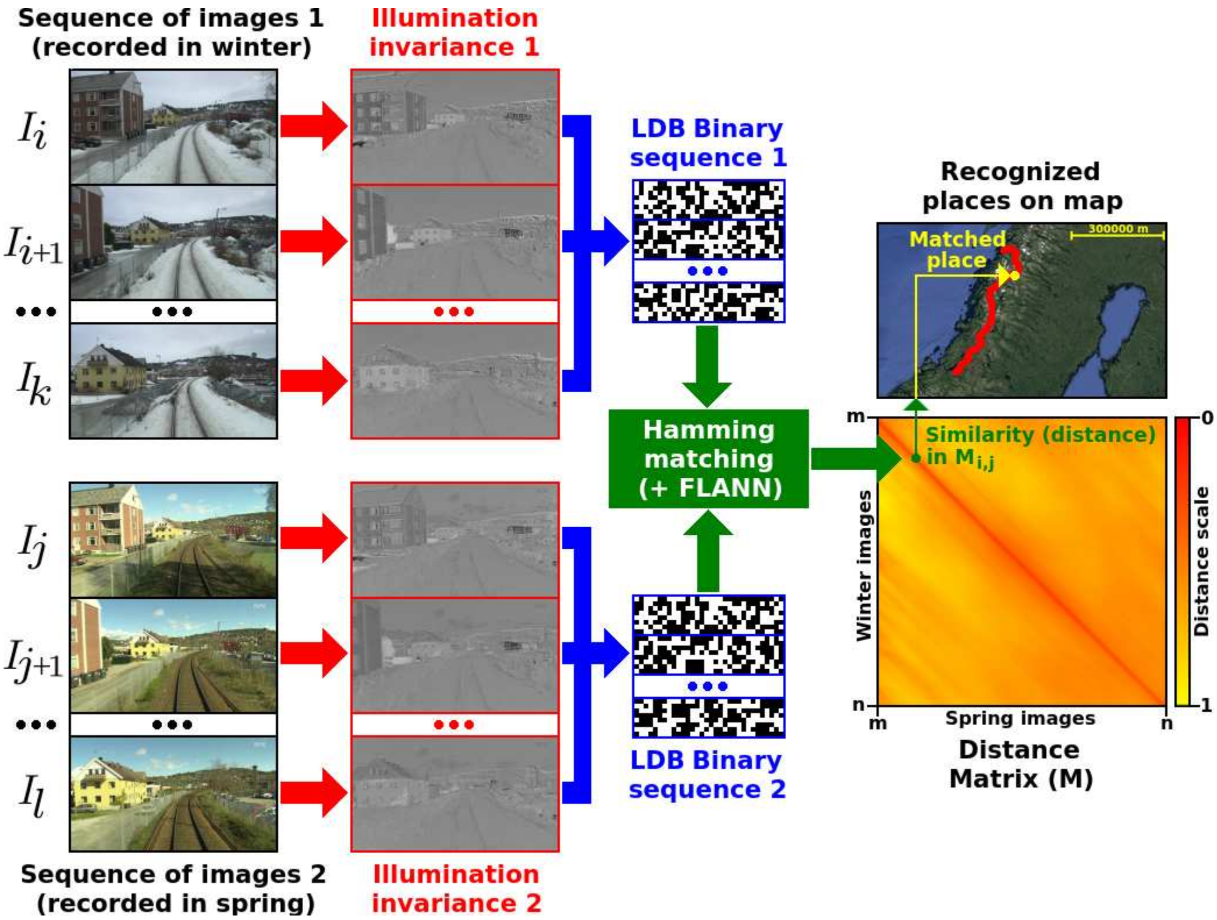


Figure 3.1: General diagram about the ABLE method. The image samples correspond to the Nordland dataset. This representation shows monocular images (ABLE-M), but along this chapter we also describe more detailed diagrams about our versions for stereo (ABLE-S) and panoramic images (ABLE-P).

3.2.1 Sequences of Images instead of Single Images

The major part of the traditional state-of-the-art algorithms proposed for visual topological localization are based on the typical assumption that places are defined by a single image. Actually, some of the most popular methods applied in visual place recognition follow this philosophy, such as WI-SURF, BRIEF-Gist or FAB-MAP.

Nevertheless, more recent algorithms such as SeqSLAM changed this assumption and considered the idea of identifying places as sequences of images, with the aim of enhancing the situational awareness in long-term conditions. For this reason, our current proposal also follows a similar methodology, because a priori it is more accurate than analyzing places as single images.

In this case, ABLE extracts binary codes as descriptors of each individual image, but they are concatenated ($+\!\!+$) to build the final binary sequence ($\mathbf{d}$), which corresponds to a sequence of images. This is formulated in Eq. 3.1, where $k - i$ is equal to $\mathbf{d}_{length}$, which is the length of the sequence of images considered by the algorithm:

$$\mathbf{d} = \mathbf{d}_{\mathcal{I}_i} +\!\!+ \mathbf{d}_{\mathcal{I}_{i+1}} +\!\!+ \mathbf{d}_{\mathcal{I}_{i+2}} +\!\!+ ... +\!\!+ \mathbf{d}_{\mathcal{I}_{k-2}} +\!\!+ \mathbf{d}_{\mathcal{I}_{k-1}} +\!\!+ \mathbf{d}_{\mathcal{I}_k}. \tag{3.1}$$

According to this, $\mathbf{d}_{length}$ is an adaptable parameter that mainly depends on the camera frame rate, as will be explained in detail in the experiments carried out along Section 3.5.2.3.

3.2.2 Illumination Invariant Transformation of Images

One of the main problems in life-long visual place recognition is the identification of places when there are important illumination changes on scene. This question has acquired a great interest in some works that try to solve these issues in several difficult situations: zones with shadows [Corke et al., 2013], dynamic lighting environments [Carlevaris-Bianco and Eustice, 2014] or night conditions [Nelson et al., 2015].

We also consider these lighting problems in the solution defined by our method. For this reason, ABLE transforms images into an illumination invariant color space, with the aim of refining the description process in these troublesome situations, as it was introduced in other previous approaches such as [Upcroft et al., 2014, McManus et al., 2014], where place recognition is enhanced using this kind of transformation. According to this, our proposal includes an initial stage to obtain illumination invariance, as exposed in Eq. 3.2:

$$\mathcal{I} = \log(G) - \alpha \cdot \log(B) - (1-\alpha) \cdot \log(R), \tag{3.2}$$

where R, G, B represent the color channels of the computed image and $\mathcal{I}$ is the processed illumination invariant image. As presented in Eq. 3.3, α is a parameter that is conditioned by the peak spectral responses of each color channel (λ_R, λ_G, λ_B), which are typically available in camera specifications:

$$\frac{1}{\lambda_G} = \frac{\alpha}{\lambda_B} + \frac{(1-\alpha)}{\lambda_R}, \tag{3.3}$$

where α is a parameter which can be simply determined if these peak spectral responses are taken into account, as explained in Eq. 3.4:

$$\alpha = \frac{(\frac{\lambda_B}{\lambda_G} - \frac{\lambda_B}{\lambda_R})}{(1 - \frac{\lambda_B}{\lambda_R})}. \tag{3.4}$$

For example, the PointGrey Flea2 camera used in datasets such as the KITTI Odometry has $\lambda_R = 610nm$, $\lambda_G = 535nm$, $\lambda_B = 470nm$, so in this case $\alpha = 0.47$. The calculation is analogue for other datasets.

As deduced from all the previous equations, illumination invariant transformation is not an arduous or computationally expensive process, but its application contributes an extra robustness to our method when lighting problems appear, as shown in the qualitative examples depicted in Figs. 3.2 and 3.3, jointly with the quantitative results presented in Sections 3.5.2.1 and 3.5.2.2.

(3.2.1) 8:45 am (initial image).

(3.2.2) 8:45 am (illumination invariance).

(3.2.3) 2:10 pm (initial image).

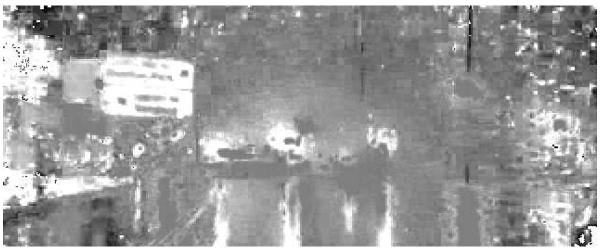
(3.2.4) 2:10 pm (illumination invariance).

Figure 3.2: An example of illumination invariance in the St Lucia dataset. Two places are depicted at different hours of the day. The images on right are the illumination invariant transformation of the left images. Here, it can be seen how this approach reduces the effects produced by sunlight, shadows and variable lighting conditions.

(3.3.1) Day (initial image).

(3.3.2) Day (illumination invariance).

(3.3.3) Night (initial image).



(3.3.4) Night (illumination invariance).

Figure 3.3: An example of illumination invariance in the Alderley dataset. Two places are depicted in a sunny summer day and in a stormy winter night. The images on right are the illumination invariant transformation of the left images. In this case, it is slightly more difficult to obtain a reduction in the effects produced by the extreme illumination changes between day and night.

3.2.3 Definition and Construction of Binary Descriptors

The application of binary features for describing places is one of the main characteristics of our proposal based on hand-crafted descriptors. Before starting the explanation about the extraction of features in ABLE, it is necessary to introduce the main properties of these kinds of descriptors and how they work, with the aim of understanding the main benefits of using them in our approach.

Binary descriptors are typically constructed from a set of pairwise comparisons from a sampling pattern which is normally centered in a point of interest of the image. The sampling pattern differs depending on the specific binary descriptor and it can be adapted for obtaining invariance to scale and rotation. When the descriptor is computed, each bit in the binary feature is the result of precisely one pairwise comparison.

Apart from the previous considerations, it must be explained how these binary features are formulated and built. If we define a smoothed image patch ($\mathbf{p}$) centered in the point of interest $\mathbf{x} = (x, y)$, a binary test (τ) is characterized as:

$$\tau\left(\mathbf{p}; f(i), f(j)\right) \;=\; \begin{cases} 1 & f(i) < f(j) \\ 0 & f(i) >= f(j) \end{cases} \;,\; i \neq j, \tag{3.5}$$

where $f(i)$ is a function that returns an image feature response for the point of interest, which is compared to other $f(j)$ for a certain pixel or cell in $\mathbf{p}$. According to this, $f(i)$ can simply be the smoothed image intensity (I) at one pixel location $\mathbf{x_i} = (x_i, y_i)$, as proposed by binary descriptors such as BRIEF, which is probably the most popular approach:

$$f(i) \;=\; I(\mathbf{x_i}). \tag{3.6}$$

However, $f(i)$ can also be the concatenation of other different binary comparisons, such as averaged image intensities (I_{avg}) and image gradients (G_x, G_y) on a specific cell ($\mathbf{c_i}$) in $\mathbf{p}$, as proposed by other binary features such as LDB, which is the descriptor used as core by ABLE:

$$f(i) \;=\; \left\{I_{avg}(\mathbf{c_i}), G_x(\mathbf{c_i}), G_y(\mathbf{c_i})\right\}. \tag{3.7}$$

Furthermore, we have defined a new binary descriptor called D-LDB, which also computes features based on geometric characteristics of the environment in the binary description process. These features are based on the extra information provided by stereo cameras, so D-LDB is applied in ABLE-S. This novel strategy is designed in order to reduce the effect of different place recognition problems such as perceptual aliasing and to obtain better results in long-term situations. The initial proposal of LDB is improved by our D-LDB descriptor, where several binary comparisons are also applied for averaged disparity information (D_{avg}):

$$f(i) \;=\; \left\{I_{avg}(\mathbf{c_i}), G_x(\mathbf{c_i}), G_y(\mathbf{c_i}), D_{avg}(\mathbf{c_i})\right\}. \tag{3.8}$$

As a last step in the procedure for constructing the binary feature, the resulting descriptor ($\mathbf{d}$) is processed as a sequence of n binary tests, where n is also the final dimension of the resultant descriptor, which can be empirically adjusted depending on the system requirements or other constraints:

$$\mathbf{d}(\mathbf{p}) = \sum_{1 \leq i \leq n} 2^{i-1} \tau \left(\mathbf{p}; f(i), f(j) \right). \tag{3.9}$$

The definitions previously contributed about the construction of binary features give an idea of their advantages for describing images in an efficient way. Firstly, these descriptors consist of a simple concatenation of bits, which involves a minor memory consumption in general terms, especially if it is compared to descriptors based on vectors of features, such as SIFT or SURF. In addition, binary features can be matched using a basic Hamming distance, which is much more efficient than the traditional way of matching descriptors with the L_2-norm. This efficiency provided by the Hamming distance ($dist_H$) is due to the simplicity of the calculation needed to compute it, which is based on an elementary XOR operation ($\oplus$) and a basic sum of bits:

$$dist_H = \mathrm{bitsum}(\mathbf{d}(\mathbf{p}_i) \oplus \mathbf{d}(\mathbf{p}_j)). \tag{3.10}$$

Finally, it must be noted that the different ABLE versions do not apply a local description model, LDB and D-LDB in each case are computed as global binary descriptors. This approach is computationally more efficient than keypoint-based methods that extract several local image descriptors over points of interest in the image.

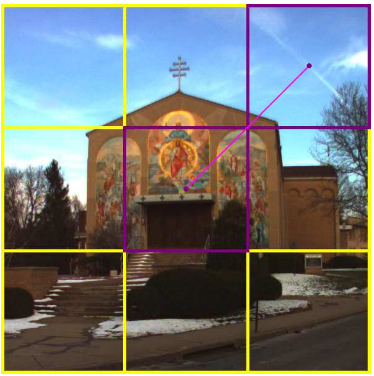
3.2.4 Extraction of Binary Features

The binary codes that form the final binary sequence (previously introduced in Eq. 3.1) are extracted from each image using a global LDB descriptor (D-LDB in the case of the stereo images processed by ABLE-S).

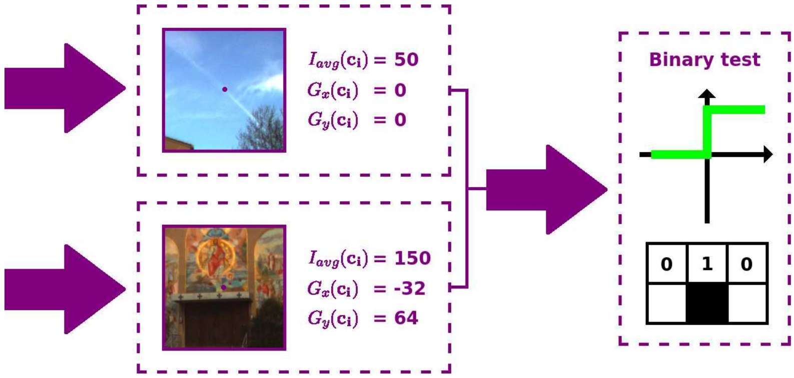
The first step carried out by the ABLE method before starting the extraction of binary codes is to downsample the acquired images to 64x64 pixels. This size is also applied in the patch ($\mathbf{p}$) considered in the binary description process. The reduction of the image size is performed because high resolution images are not required to carry out a robust visual topological localization, as evidenced in works such as [Milford, 2012]. Besides, this strategy followed by ABLE allows to reduce memory and computational costs without decreasing precision. Additionally, downsampling the initial images implicates smoothing and interpolation over neighboring regions that attenuate the negative influence of rotation and scale in place recognition, as stated in [Sünderhauf and Protzel, 2011].

After the computation of each image patch, the global binary descriptor is extracted by processing the center of the resized image patch as a keypoint without dominant rotation or scale. The resultant binary code is adjusted to a dimension (n) of 32 bytes, which is supported by some previous experiments carried out in [Arroyo et al., 2014b]. This value of n is fixed using a Random Bit Selection (RBS), as proposed in LDB works [Yang and Cheng, 2012, Yang and Cheng, 2014]. In this sense, the evaluation presented in [Yang and Cheng, 2012] demonstrated that the precision of RBS is close to the one achieved by more refined methods, such as entropy-based.

LDB and its derivatives are chosen as the core of our global binary description method because these features provide several advantages with respect to other descriptors. First of all, LDB is not only based on intensity comparisons, like other popular approaches such as BRIEF [Calonder et al., 2010]. This technique gives more robustness to the description process thanks to the extra gradient information. Apart from this, one of the main benefits provided by LDB is that it computes the features using a multi-resolution scheme, where different grids (2x2, 3x3, 4x4...) are applied to capture information at different granularities, as depicted in Fig 3.4. The application of this multi-resolution approach alleviates the dependence on the field of view suffered in visual place recognition by proposals such as SeqSLAM, as it was exhibited in [Sünderhauf et al., 2013a] and will be confirmed by some of the tests presented in this chapter in Section 3.5.2.3.



(3.4.1) Example of binary test.



(3.4.2) Computation of the binary test on a pair of grids.

(3.4.3) Patch with 2x2 gridding.

(3.4.4) Patch with 3x3 gridding.

(3.4.5) Patch with 4x4 gridding.

Figure 3.4: Extraction of binary features based on LDB. As shown above, an image patch is divided into equal-sized grids to compute intensity and gradients of each cell, which are compared between pairs in a binary test to obtain the final result (0 or 1, depending on the result of the evaluation). As shown below, a multiple gridding strategy (2x2, 3x3, 4x4) is used to capture information at different image levels.

3.3 Image Matching of Locations in ABLE

The bottleneck in efficiency of visual place recognition methods is normally related to the matching of the extracted features. This is because the number of images to be matched is progressively increased in each iteration. However, the properties of the binary description approach designed in ABLE allow to reduce the costs of matching, jointly with a proposed search method based on ANN, as described in detail in the following pages.

3.3.1 Fast Matching of Binary Features using the Hamming Distance

The similarity or distance between the binary sequences extracted from the images processed by ABLE is efficiently calculated using the Hamming distance. The obtained values can be included in a distance matrix (M) for analyzing them in loop closure detection or for evaluation purposes, as explained in Eq. 3.11:

$$M_{i,j} = M_{j,i} = \text{bitsum}(\mathbf{d}_i \oplus \mathbf{d}_j). \tag{3.11}$$

In addition, POPCNT is a machine SSE4.2 instruction which can be used for a faster matching of the binary sequences, since it allows to count the total number of bits that are set to one in a more efficient way, as exposed in [Muja and Lowe, 2012]. We take advantage of this instruction for increasing the speed in the calculation of similarity, as shown in Eq. 3.12:

$$M_{i,j} = M_{j,i} = \text{POPCNT}(\mathbf{d}_i \oplus \mathbf{d}_j). \tag{3.12}$$

It must be noted that ABLE-M and ABLE-S apply a simple correlation in image matching, but ABLE-P uses a cross-correlation of panoramas in order to exploit the valuable visual information in all the possible directions provided by panoramic views, as will be broadly described in Section 3.4.3.

3.3.2 Approximated Nearest Neighbors for Reducing Matching Costs

As a last contribution in our matching method, we implement an ANN search based on the functionalities given by the Fast Library for Approximate Nearest Neighbors (FLANN) [Muja and Lowe, 2012, Muja and Lowe, 2014]. The index used in this search consists of a multi-probe Local Sensitive Hashing (LSH).

This hashing technique is used due to the characteristics of our binary codes, which are keys compatible with this method and give a fast performance for our Hamming matcher. The main idea behind the multi-probe LSH index used in our ANN search is focused on systematically testing multiple binary codes for the image queries in a hash table, whose hash keys may not necessarily be completely identical to the hash value of the query vector. If we consider an image query ($\mathbf{q}$), the applied hash function ($g(\mathbf{q})$) is denoted by the different hash slots (h) which are involved on it:

$$g(\mathbf{q}) = \{h_1(\mathbf{q}), h_2(\mathbf{q}), h_3(\mathbf{q}), ..., h_{k-2}(\mathbf{q}), h_{k-1}(\mathbf{q}), h_k(\mathbf{q})\}, \tag{3.13}$$

where multi-probe LSH searches for a sequence of a hash perturbation vector (δ_i), which is formulated as follows:

$$\delta_i = \{\delta_{i_1}, \delta_{i_2}, \delta_{i_3}, ..., \delta_{i_{k-2}}, \delta_{i_{k-1}}, \delta_{i_k}\}. \tag{3.14}$$

Attending to Eq. 3.13 and 3.14, the algorithm sequentially probes the different hash buckets $\{g(\mathbf{q}) + \delta_i\}$. Finally, a score ($s_i(\mathbf{q})$) is computed to sort the perturbation vectors with the aim of accessing the buckets in order of increasing scores and easily obtaining the searched hash codes. The score is calculated as shown in Eq. 3.15, where $x_j(\delta_{ij})$ is the distance of $\mathbf{q}$ from the boundary of the slot $h_j(\mathbf{q}) + \delta_j$:

$$s_i(\mathbf{q}) = \sum_{j=1}^{k} x_j^2(\delta_{ij}). \tag{3.15}$$

As demonstrated in [Lv et al., 2007], multi-probe LSH achieves the same search quality with a similar time consumption if it is compared to the conventional LSH. However, the difference resides in the number of hash tables, which is reduced in an order of magnitude by multi-probe LSH, and for this reason we chose this method as core of our ANN search.

3.4 ABLE for Monocular, Stereo and Panoramic Cameras

As briefly introduced in Chapter 2, the visual place recognition methods in the literature employ different types of cameras for perceiving the environment: monocular, stereo or panoramic. Each approach has its pros and cons, but the usage of one or another camera usually depends on the constraints of the specific application. For this reason, we have developed a solution which can take advantage of the varied information provided by the images of the different types of cameras, with the aim of adapting our algorithm to the best conditions for each case: ABLE-M for monocular images, ABLE-S for stereo images and ABLE-P for panoramic images. The main characteristics and differences between the three ABLE versions are detailed in Table 3.1.

Apart from the technical differences that will be explained in detail along this section, there are other parameters related to the performance of each ABLE version that are also qualitatively described in Table 3.1. Obviously, the effectiveness of each version in topological place recognition has a great dependence on the amount of information provided by each type of camera: ABLE-P obtains the best precision because it exploits the visual data acquired in all the possible directions, ABLE-S also has a remarkable effectiveness because geometrical information given by stereo cameras is used, but ABLE-M is not as precise as the other two versions because it only employs monocular cameras, which provide a worse visual awareness of the environment. However, ABLE-M has a better efficiency in memory and computational costs due to the processing of a lower amount of data with respect to ABLE-S or ABLE-P, which can be an advantage in systems that require a localization for long operating periods, because in these cases a moderate consumption of resources is critical. All these qualitative assessments about the performance of each ABLE version will be quantitative demonstrated in the experiments explained along Section 3.5.2.6.

Table 3.1: Differences among the properties of each ABLE version. Qualitative assessments are represented by $\star\star\star$ for the best effectiveness and the lowest memory consumption and computational costs.

	ABLE-M	**ABLE-S**	**ABLE-P**
Camera	Monocular	Stereo	Panoramic
Description	Global LDB	Global D-LDB	Global LDB
Matching	Simple correlation	Simple correlation	Cross-correlation
Loop closure	Unidirectional	Unidirectional	Bidirectional
Effectiveness	$\star$	$\star\star$	$\star\star\star$
Memory	$\star\star\star$	$\star\star$	$\star$
Computation	$\star\star\star$	$\star\star$	$\star$

3.4.1 Monocular Images: ABLE-M

ABLE-M can be considered as the standard version of ABLE. As explained in previous sections, it uses a global LDB descriptor as core for extracting image features and a Hamming matching based on a simple correlation of the processed images. In the following sections, we are going to explain the main differences between ABLE-S and ABLE-P with respect to the standard monocular version.

3.4.2 Stereo Images: ABLE-S

ABLE-S has a notable difference compared to the other versions, which is the application of an improved method for extracting features. Although the three versions use a global binary description, ABLE-M and ABLE-P apply a LDB descriptor, while ABLE-S computes D-LDB, with the aim of adding the valuable disparity information provided by stereo cameras to the description process.

3.4.2.1 The D-LDB Descriptor

As briefly introduced in Section 3.2.3, D-LDB is a descriptor derived from our thesis work [Arroyo et al., 2014a] that improves the initial proposal presented by LDB. In this regard, binary comparisons are also applied for averaged disparity information, which provides a higher resistance to geometrical changes in locations. The computation of D-LDB is graphically described in Fig. 3.5, where it is also shown how the different approaches add not only intensity comparisons to the description process, but also gradients and disparity information.

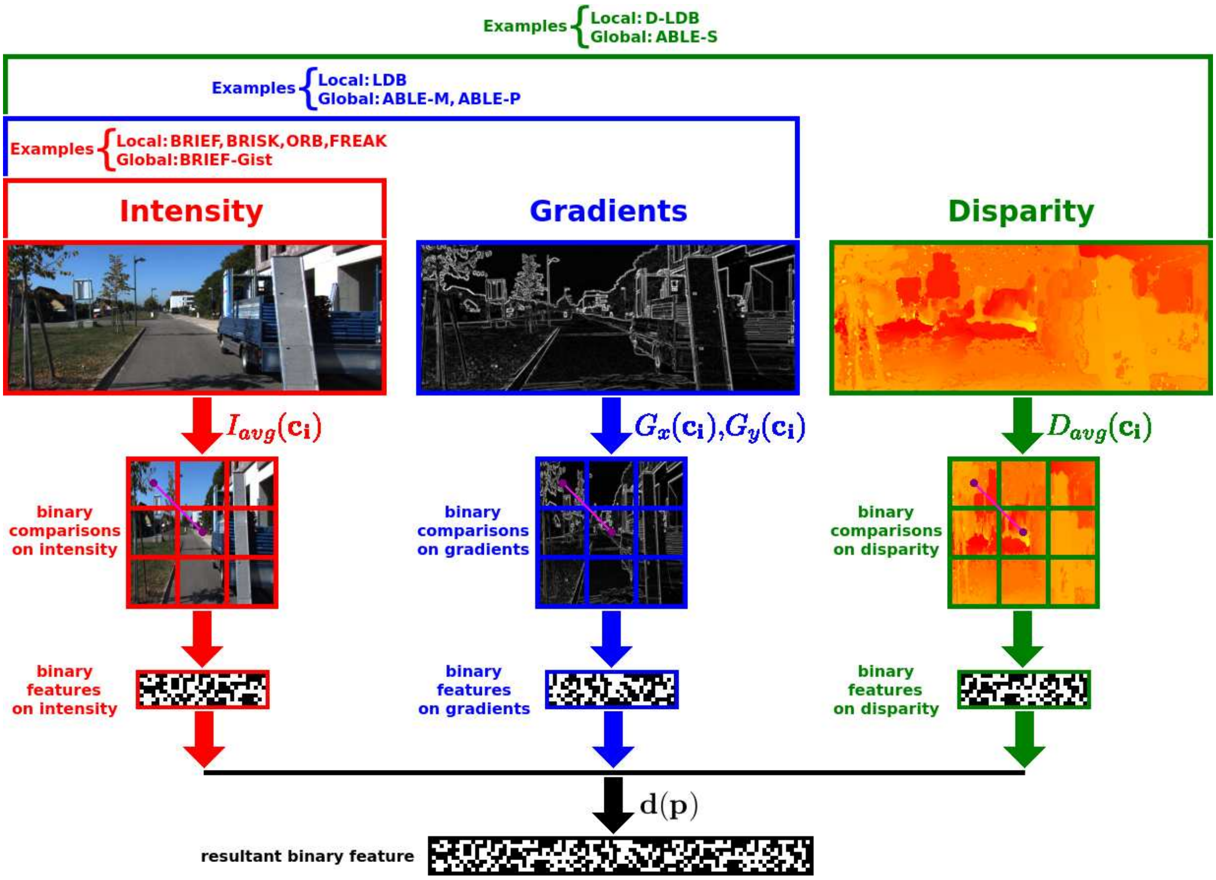
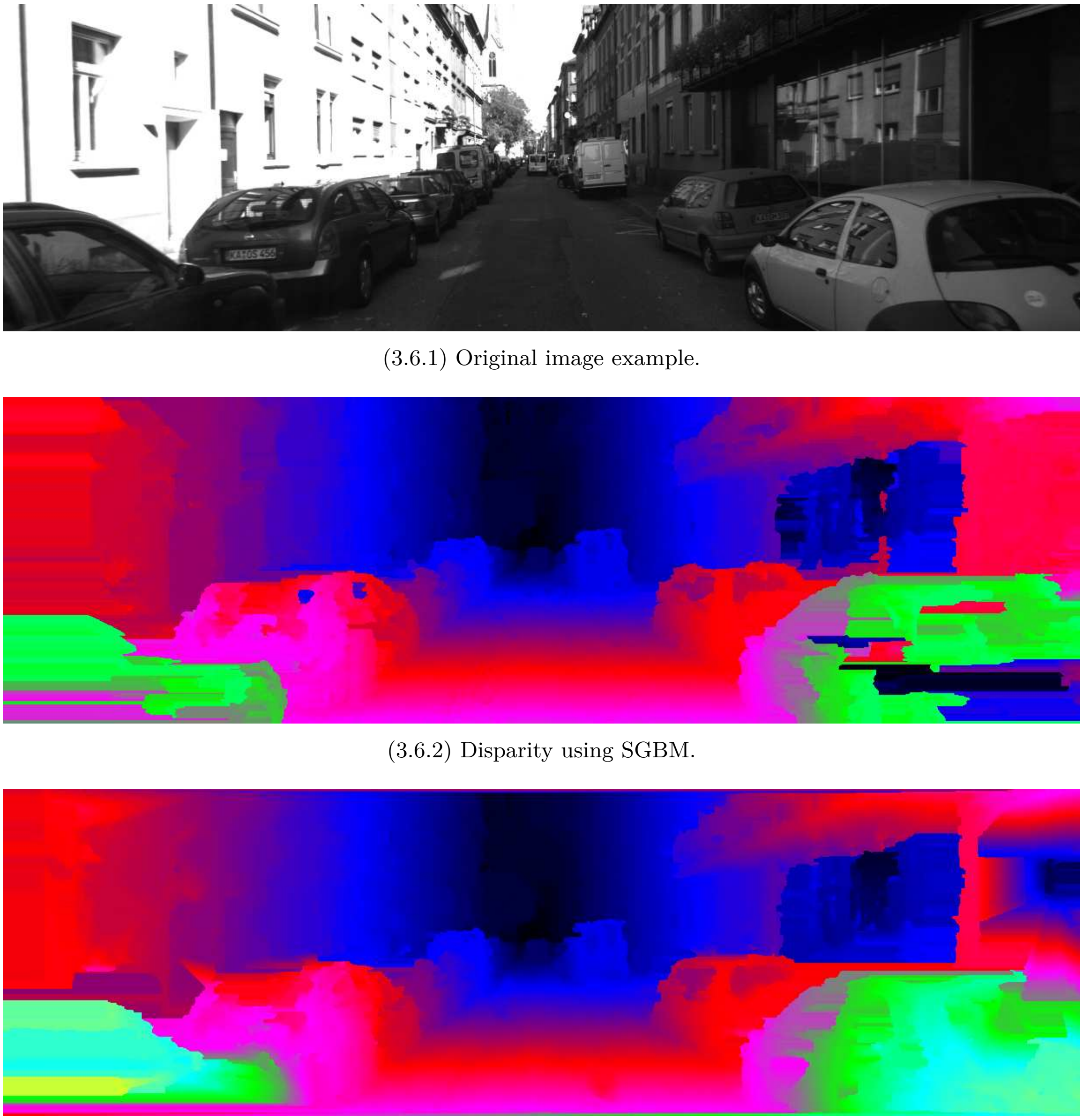


Figure 3.5: Features used by the different versions of ABLE. Intensity and gradient information are computed by ABLE-M and ABLE-P. Besides, disparity is also included in ABLE-S by means of the D-LDB descriptor.

3.4.2.2 Stereo Matching for Disparity Calculation in D-LDB

The performance of the D-LDB descriptor in visual place recognition depends on the stereo matching method applied for disparity computation. Stereo matching was typically performed in the D-LDB descriptor using an approach based on a standard Semi-Global Block Matching (SGBM) [Hirschmuller, 2008], which is available in the OpenCV libraries. However, we also implement Efficient LArge-scale Stereo matching (ELAS) [Geiger et al., 2010] to obtain more precise disparity maps, which improve the effectiveness of ABLE-S in place recognition, as will be demonstrated in the tests presented in Section 3.5.2.5. The differences between the disparity maps provided by SGBM and ELAS can be also noticed in the examples depicted in Fig. 3.6.



(3.6.1) Original image example.

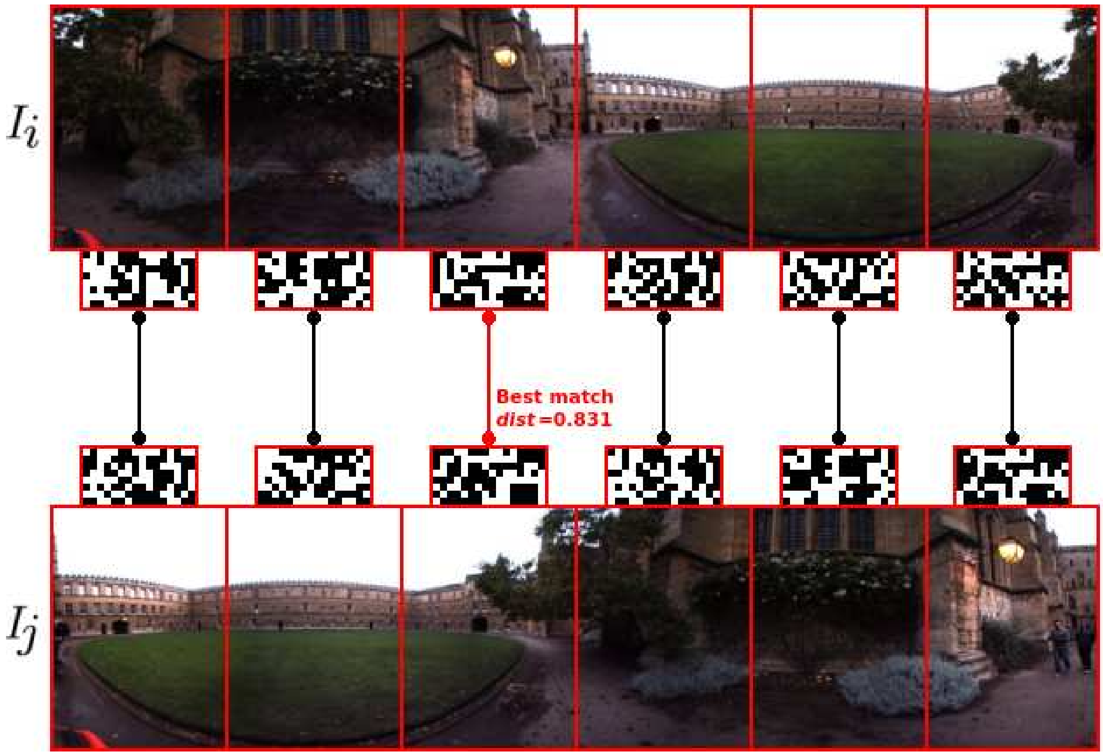
(3.6.2) Disparity using SGBM.

(3.6.3) Disparity using ELAS.

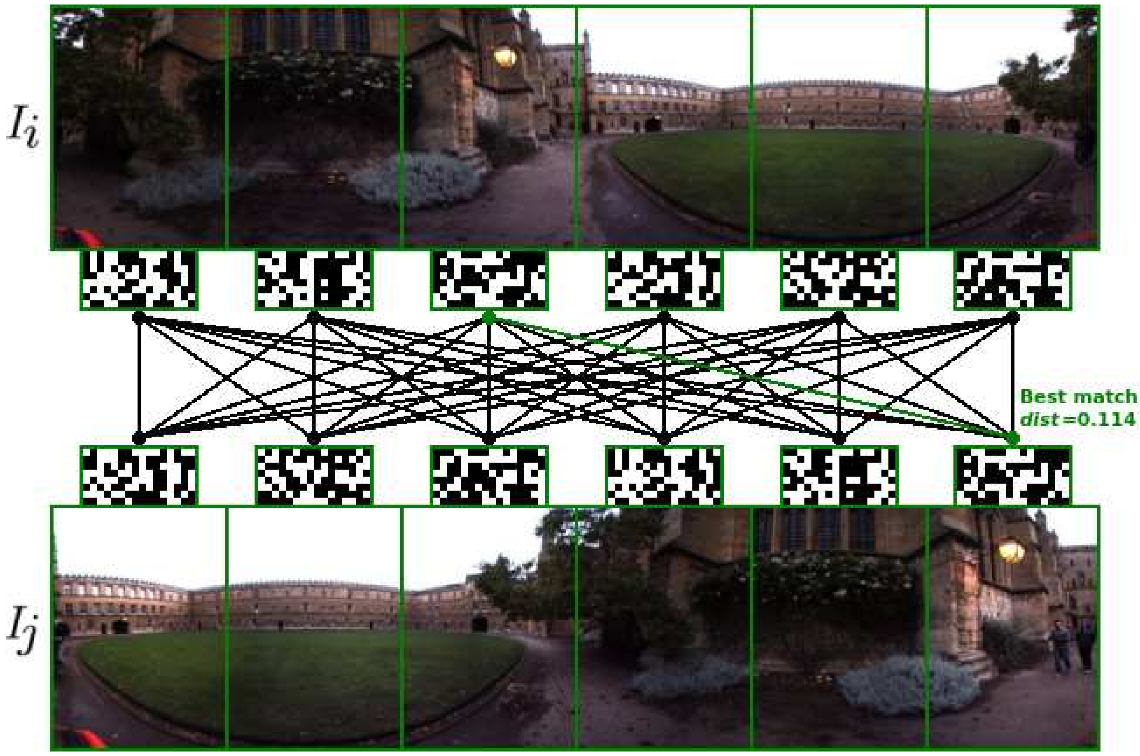
Figure 3.6: Disparity calculation using SGBM and ELAS in the KITTI dataset. This qualitative example corresponds to the stereo matching benchmark.

3.4.3 Panoramic Images: ABLE-P

ABLE-P has a particular way of processing the similarity between places with respect to the other ABLE versions. More specifically, in ABLE-P each panoramic image is divided into subpanoramas, which are matched by using a cross-correlation technique, achieving more accurate similarity distances between a pair of panoramas. This approach is very useful when a place is revisited in an opposite direction. In these cases, when methods such as BRIEF-Gist compute a simple correlation of subpanoramas, they can not identify the revisited location, because they do not exploit the visual perception in several directions provided by the panoramic images. This is solved using our cross-correlation of subpanoramas, as illustrated in Fig. 3.7.



(3.7.1) Simple correlation in panoramic images (Example: BRIEF-Gist).



(3.7.2) Cross-correlation in panoramic images (Example: ABLE-P).

Figure 3.7: Differences between a simple correlation and a cross-correlation of panoramas. The two panoramic images to be matched in these examples correspond to a place revisited in an opposite direction in the Oxford New College dataset. It can be seen how the best match between subpanoramas obtained in the cross-correlation applied by ABLE-P has a much lower distance and is much more similar than using a simple correlation.

3.4.3.1 Cross-Correlation of Panoramas in Matching

The cross-correlation of panoramas applied by ABLE-P is formally defined in Eq. 3.16 and Eq. 3.17, where similarity is computed for each pair of sub-panoramas (m, n) corresponding respectively to the panoramic images (i, j). The distances between all the sub-panoramas are saved in a preliminary cross-correlation matrix (C), where a minimum is calculated to obtain the final value stored in M:

$$C_{k,l} = C_{l,k} = \mathrm{POPCNT}(\mathbf{d}_{i,m} \oplus \mathbf{d}_{j,n}). \tag{3.16}$$

$$M_{i,j} = M_{j,i} = \min(C). \tag{3.17}$$

It must be remarked that the formulated cross-correlation allows to identify bidirectional loop closures, which are considered when the analyzed location is revisited in a different direction. This detection of loop closures is one of the main applications of topological place recognition methods in life-long localization, which will be described in Chapter 5.

3.5 Experiments and Results

The evaluation of the performance for the different ABLE versions is mainly based on analyzing their benefits and results in life-long localization in several datasets where topological place recognition can be tested for monocular, stereo and panoramic images. We also compare our approach against the main state-of-the-art algorithms in challenging long-term conditions, with the aim of validating the effectiveness and efficiency of ABLE in an objective way and using a fair evaluation methodology.

3.5.1 Experimental Setup

Before presenting the results obtained by means of the research described in this chapter, it is important to define the methodology used for evaluating the ABLE method and the main characteristics of our experimental setup, which are described in the following sections.

3.5.1.1 Evaluation Methodology

The designed methodology for testing the performance of ABLE is principally based on precision-recall curves, which are calculated from the distance matrices (M) obtained in each test performed by our method. Before starting evaluation, the distance values contained in M must be normalized by following Eq. 3.18:

$$M_{i,j} = \frac{M_{i,j}}{\max(M)}. \tag{3.18}$$

After the previous step, M is thresholded for comparing it against the ground-truth matrix (G) associated with a specific dataset. In this respect, true positives (tp) are contemplated if a positive of the thresholded M matches with a positive of G in a temporal vicinity according to the frame rate. False positives (fp) are considered in the inverse situation, and false negatives (fn) if a negative is found in the thresholded M when a positive should appear. According to these considerations, the values of precision and recall can be calculated as shown in Eq. 3.19 and Eq. 3.20:

$$precision = \frac{tp}{tp + fp}. \tag{3.19}$$

$$recall = \frac{tp}{tp + fn}. \tag{3.20}$$

The final precision-recall curve is processed by varying the threshold value (θ) in a uniform distribution between 0 and 1 and computing the associated values of precision and recall in each iteration. In our tests, 100 values of θ are taken into account in order to obtain well-defined curves.

3.5.1.2 State-of-the-art Methods evaluated in Comparisons

The provided tests compare the different ABLE versions against the main state-of-the-art algorithms that are available for evaluation: WI-SURF, BRIEF-Gist, FAB-MAP and SeqSLAM. For testing WI-SURF and BRIEF-Gist, we developed implementations based on the SURF and BRIEF descriptors provided by the OpenCV libraries. OpenFABMAP [Glover et al., 2012] is the toolbox chosen for testing FAB-MAP, which is applied in a standard configuration and properly trained. The experiments for SeqSLAM are performed with the source code provided by OpenSeqSLAM [Sünderhauf et al., 2013a].

3.5.1.3 Tested Datasets

The validation of the three ABLE versions is carried out over several publicly available datasets. Concretely, six datasets are used in these tests, where experiments are focused on a specific ABLE version in each case depending on the type of available camera. ABLE-M is mainly tested in the St Lucia dataset, the Alderley dataset, the Nordland dataset and the CMU-CVG VL dataset. ABLE-S is tested in the KITTI Odometry dataset. Finally, ABLE-P is tested in the Oxford New College dataset . The characteristics of these datasets were described along Section 2.5 and summarized in Table 2.3.

In these experiments, ABLE is clearly validated in long-term conditions, especially if we consider that a distance higher than 3000 km is traversed over all the performed tests. Furthermore, several challenging situations appear in the different datasets that are used in the evaluations: seasonal changes, illumination problems, perceptual aliasing, dynamic objects on scene or loop closures.

3.5.2 Main Results

The wide set of results presented along this chapter is divided into six sections depending on the tested dataset, with the aim of evaluating the precision of the three ABLE versions under different conditions and environments. Moreover, the computational costs are also discussed. These evaluations focused on life-long topological localization will corroborate the satisfactory performance and efficiency of our visual place recognition based on hand-crafted binary descriptors, especially if it is compared to the state of the art in this research line.

3.5.2.1 ABLE-M in the St Lucia Dataset

This dataset allows to evaluate the improvements provided by the illumination invariant proposal defined by ABLE in a visual place recognition along the day using monocular cameras. This is because the St Lucia dataset contains several video sequences recorded for a same route where varied illumination changes and shadowing effects appear when a location is traversed at different times of day.

Apart from the qualitative examples previously introduced for illumination invariance in Fig. 3.2, now we present precision-recall curves where the advantages of applying the illumination invariant transformation in ABLE-M are evidenced by comparing two sequences of the St Lucia dataset recorded at two different hours of a same day, as depicted in Fig. 3.8.

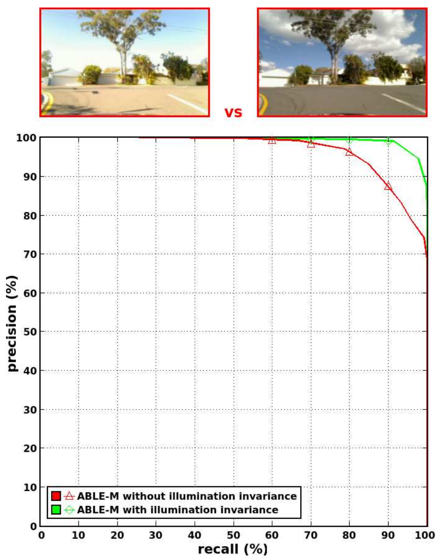


Figure 3.8: Results about ABLE-M in the St Lucia dataset. Precision-recall curves are depicted with the aim of supporting the performance comparison obtained by ABLE-M with and without using the illumination invariant technique between the sequences corresponding to the car rides recorded on 10/09/09 at 8:45 am and at 2:10 pm

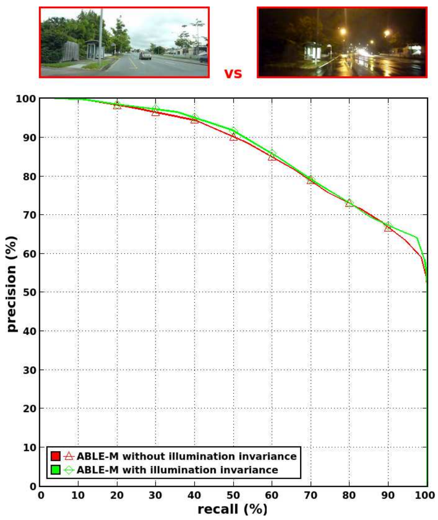
In Fig. 3.8, it is demonstrated that the usage of illumination invariance improves the general performance of ABLE-M when a route is traversed along the day. The description of the sequences is more robust in this case due to the reduction of the effects produced by sunlight and shadows when illumination invariance is applied by our method.

3.5.2.2 ABLE-M in the Alderley Dataset

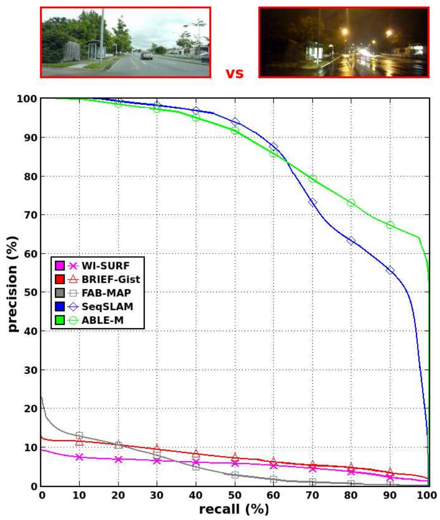
The Alderley dataset contains two video sequences which are very challenging, because they are recorded in a sunny summer day and a stormy winter night. Obviously, in these conditions is much more difficult to match places and ABLE-M achieves worse results than for the tests carried out in the St Lucia dataset, as corroborated if the precision-recall curves presented in Fig. 3.9.1 are compared to the previously depicted in Fig. 3.8.

The application of the illumination invariant technique does not substantially improve the results at night because the source illuminant cannot be modeled as a black-body radiator and the obtained appearance does not completely match the illumination invariance captured during the day. In addition, in the specific circumstances of the Alderley dataset, raindrops and humidity in camera also negatively affect to the effectiveness of this method, as can be observed in the qualitative examples previously presented in Fig. 3.3.

Similar problems due to night illumination conditions were reported in [Maddern et al., 2014], where their approach combined with an illumination invariant transformation also has worse results for a route traversed between 8:00 pm and 7:00 am. However, the precision-recall curves showed in Fig. 3.9.2 evidence that ABLE-M is very competitive in these extreme situations compared to other successful state-of-the-art methods based on hand-crafted features, as will be exhaustively demonstrated in the following experiments carried out in the Nordland and the CMU-CVG VL datasets.



(3.9.1) Illumination invariance study.

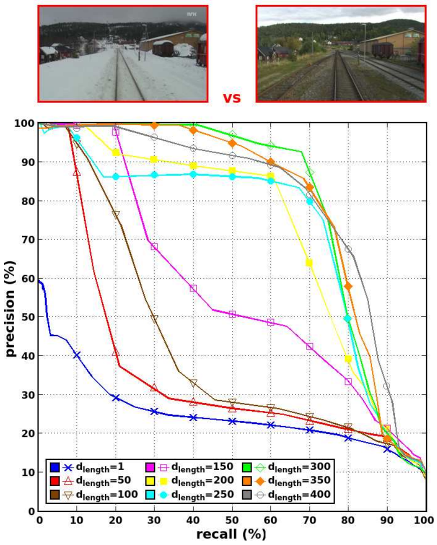


(3.9.2) State-of-the-art methods.

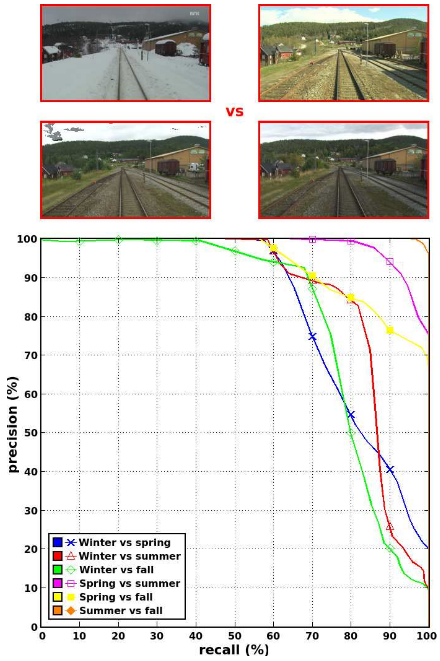
Figure 3.9: Results about ABLE-M in the Alderley dataset. Precision-recall curves are depicted with the aim of supporting the performance comparison obtained by ABLE-M against the main state-of-the-art methods. Besides, a study about the precision of the method depending on the application of the illumination invariant technique is showed.

3.5.2.3 ABLE-M in the Nordland Dataset

The Nordland dataset is probably the longest (≈ 3000 km) and one of the most challenging datasets that can be currently used for life-long topological localization evaluation. It contains four videos with very strong seasonal appearance changes in places for a same train ride. The first results presented for this dataset are focused on comparing the performance of our ABLE-M method depending on the length of the sequence ($\mathbf{d}_{length}$) used in the experiments, as explained in Section 3.2.1. These results show some tests performed between the sequences of winter and fall using different values of $\mathbf{d}_{length}$, as depicted in Fig. 3.10.1. In addition, Fig. 3.10.2 provides an initial study about the performance of ABLE-M for the six possible combinations between the four seasons.



(3.10.1) Sequence length comparison.



(3.10.2) Performance across seasons.

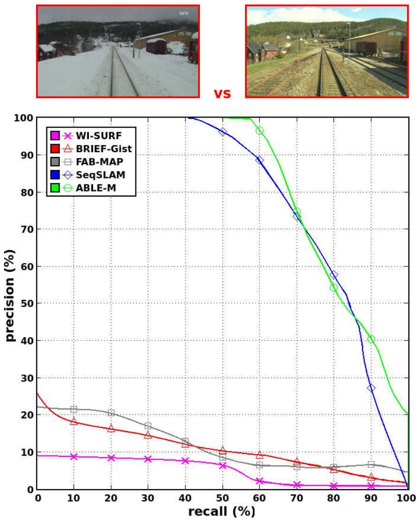
Figure 3.10: Results about ABLE-M in the Nordland dataset. Precision-recall curves are depicted with the aim of supporting the performance comparison obtained by ABLE-M for the six possible combinations between the four seasons. In addition, a study about the precision of the method depending on the image sequence length ($\mathbf{d}_{length}$) is provided between the winter and fall sequences.

Attending to the precision-recall curves showed in Fig. 3.10.1, the influence of $\mathbf{d}_{length}$ is decisive for improving the performance of visual place recognition in life-long localization. Furthermore, there is a limit near to a length of 300 where results are not greatly enhanced. For this reason, we will apply a $\mathbf{d}_{length} = 300$ in the rest of the experiments. This parameter is proportionally adaptable to the frame rate for other datasets.

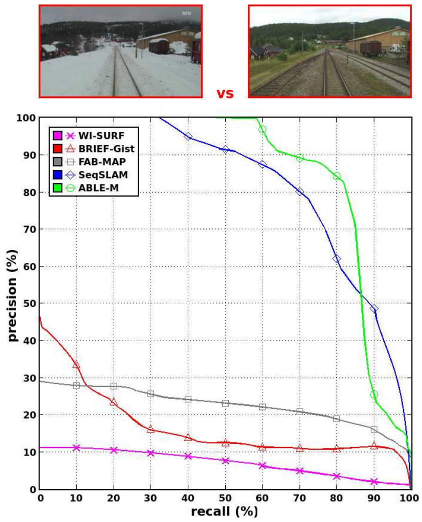
Apart from the previously introduced evaluations, a comparison about the accuracy of ABLE-M with respect to the main state-of-the-art approaches based on hand-crafted features is provided for the different seasonal changes that the Nordland dataset contains in all its video sequences. In Fig. 3.11, it is confirmed how the methods based on single images instead of sequences have a much lower precision in long-term conditions. For this reason, WI-SURF, BRIEF-Gist and FAB-MAP considerably reduce their performance in these cases, especially if their effectiveness is compared to the obtained by SeqSLAM and ABLE-M in the six tests among all the seasons. In addition, the differences between the precision of ABLE-M and SeqSLAM are not extremely significant, but in general terms, it can be seen that our method has a slightly higher performance in the major part of the cases. Besides, it is also remarkable that the winter sequence is the most troublesome in the tests. The accuracy of the algorithms is clearly reduced due to the extreme changes that the appearance of a place suffers when snow covers the scene and illumination is also substantially variable. This is confirmed by the distance matrices generated by ABLE-M for the winter sequence, which are presented in Fig. 3.12.

Nevertheless, the Nordland dataset has a characteristic that can be advantageous for the precision of some visual place recognition algorithms: this dataset is recorded with a static monocular camera mounted on a train, which always offers the same point of view of the scene. This particular property makes the Nordland dataset propitious for the adequate performance of methods such as SeqSLAM, which has a great sensitivity with respect to changes on the field of view. In fact, this sensitivity has been initially proven in [Sünderhauf et al., 2013a], where some artificial changes on the field of view of the frames registered in the Nordland dataset are introduced in some of the tests described in that paper. We decide to process more experiments supporting this assumption, with the aim of conducting an evaluation as fair as possible of the ABLE method. The associated results are presented in Fig. 3.13, where the following changes on the field of view are tested: a translation of a 10%, a rotation of a 10% and a combination of both. We perform all these new evaluations between the sequences of winter and spring, because this is the most difficult combination and the worst case for all the algorithms, according to the results shown in Fig. 3.11.1.

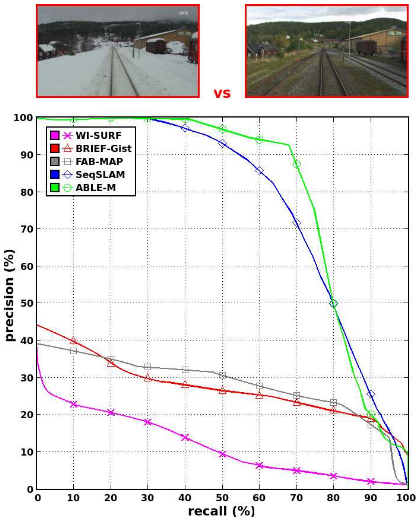
Attending to the results contributed in Fig. 3.13, the changes on the field of view (especially rotations) negatively affect the effectiveness of SeqSLAM and ABLE-M, mainly if their curves are compared to the previously obtained in Fig. 3.11.1. However, it is also evident that the performance of SeqSLAM decreases much more than the achieved by our approach in these particular conditions. The best behavior of ABLE-M to the changes on the field of view is associated with two of the properties about its description method, introduced in Section 3.2.4. On the one hand, the multi-resolution scheme applied by the global LDB features used as core in ABLE-M mitigates the dependence on the field of view, with respect to the approach considered by SeqSLAM, based on simple image difference vectors. On the other hand, the negative effects produced by scale and rotation are also slightly alleviated with the initial image downsampling previously computed by ABLE-M, due to the benefits provided by smoothing and interpolation over neighboring zones. Finally, it must be noted that the approaches based on single images (WI-SURF, BRIEF-Gist, FAB-MAP) are not represented in the precision-recall curves of Fig. 3.13, because their low accuracy in long-term conditions was sufficiently evidenced in Fig. 3.11.



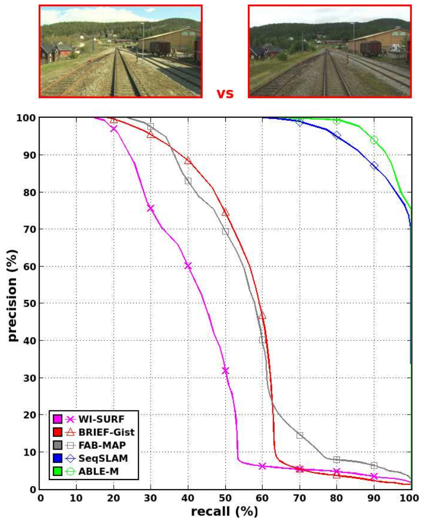
(3.11.1) Winter vs Spring.



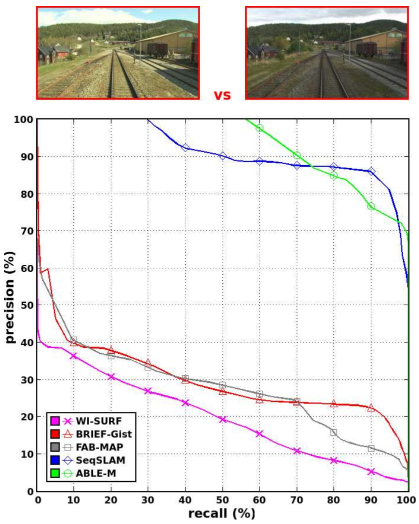
(3.11.2) Winter vs Summer.



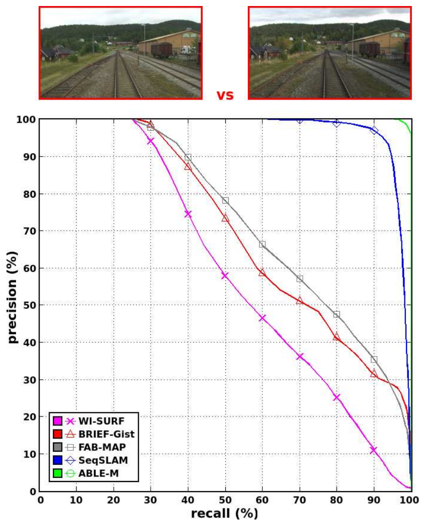
(3.11.3) Winter vs Fall.



(3.11.4) Spring vs Summer.

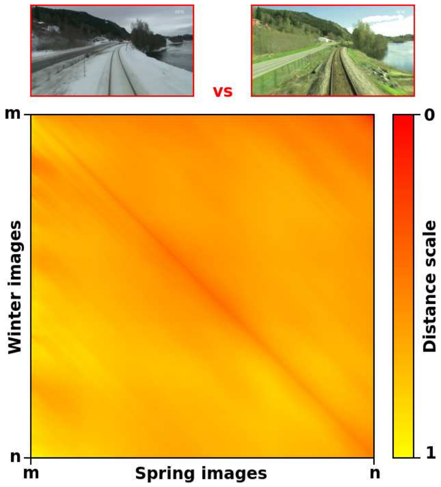


(3.11.5) Spring vs Fall.

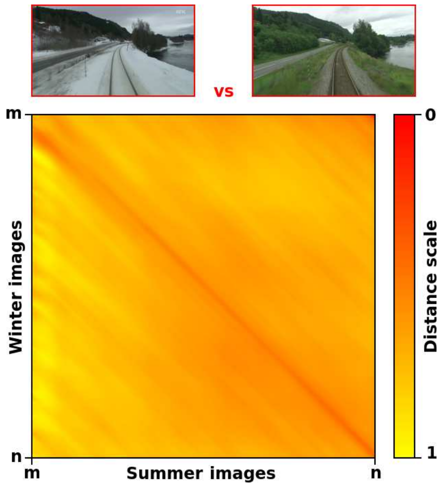


(3.11.6) Summer vs Fall.

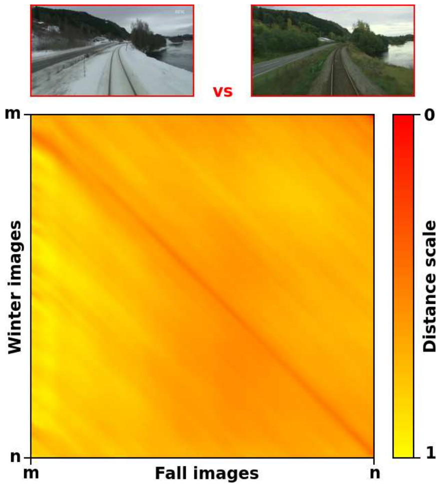
Figure 3.11: ABLE-M vs state-of-the-art methods in the Nordland dataset. Precision-recall curves are depicted with the aim of supporting the performance comparison. The four video sequences corresponding to each season are evaluated between them for all the algorithms. An illustrative frame from the sequences matched is shown in order to visually understand the complexity of place recognition in each case.



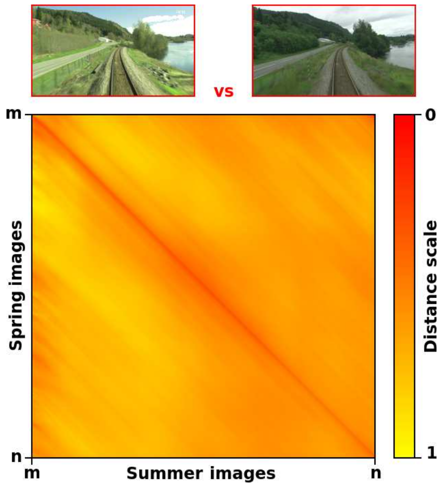
(3.12.1) Winter vs Spring.



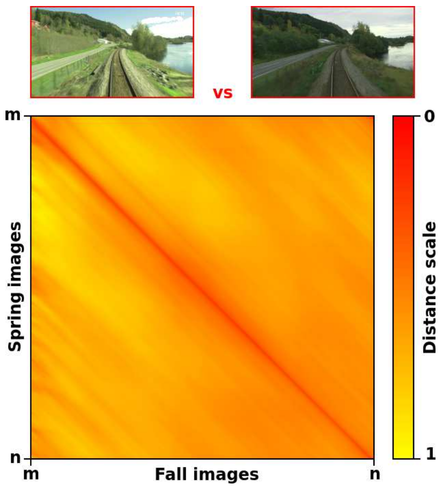
(3.12.2) Winter vs Summer.



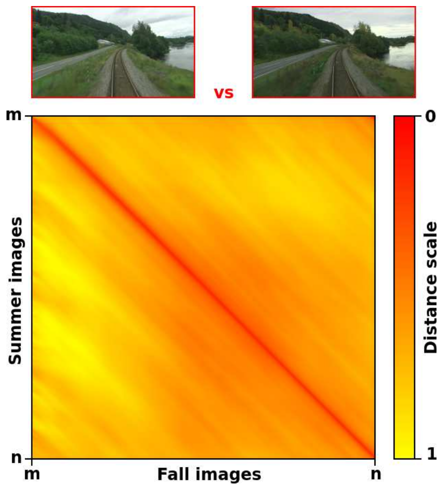
(3.12.3) Winter vs Fall.



(3.12.4) Spring vs Summer.

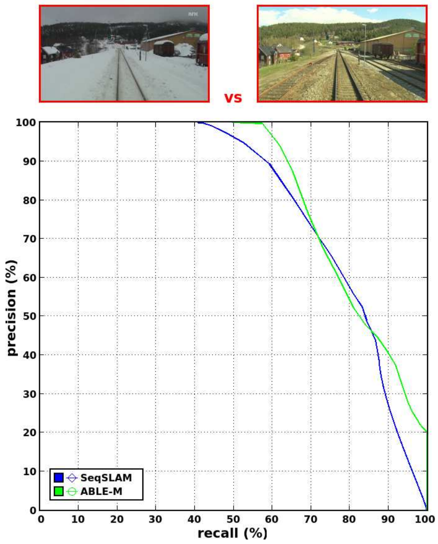


(3.12.5) Spring vs Fall.

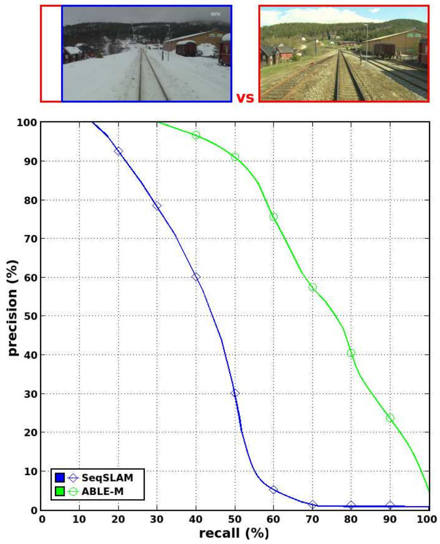


(3.12.6) Summer vs Fall.

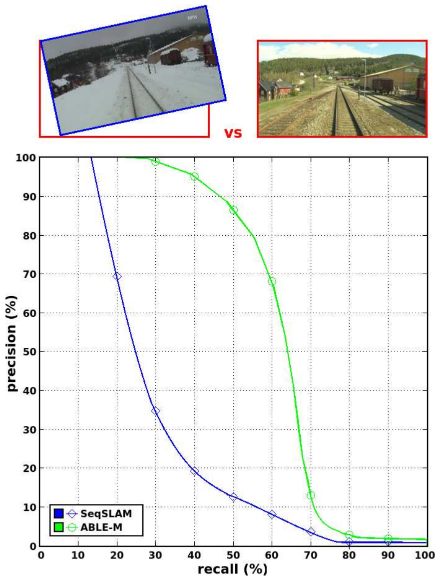
Figure 3.12: Distance matrices obtained by ABLE-M in the Nordland dataset. Comparisons between the four seasons are depicted, jointly with some image samples of recognized locations. It must be noted that the results presented in the distance matrices correspond to the representative frames between m=200000 and n=201000, because we can not conveniently show the matrices for the full dataset due to the limitations of the document format. As can be seen, the winter sequence is the most problematic because of the strong appearance changes related to this season, such as snow or low illumination. For this reason, the diagonal which appears in the matrices is not so clear when the winter sequence is evaluated.



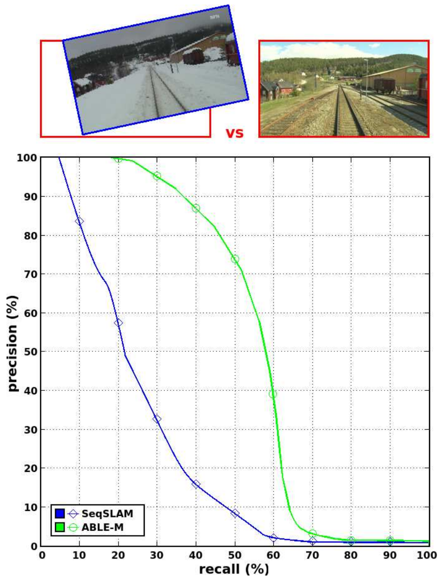
(3.13.1) Without rotation and trans.



(3.13.2) With translation.



(3.13.3) With rotation.

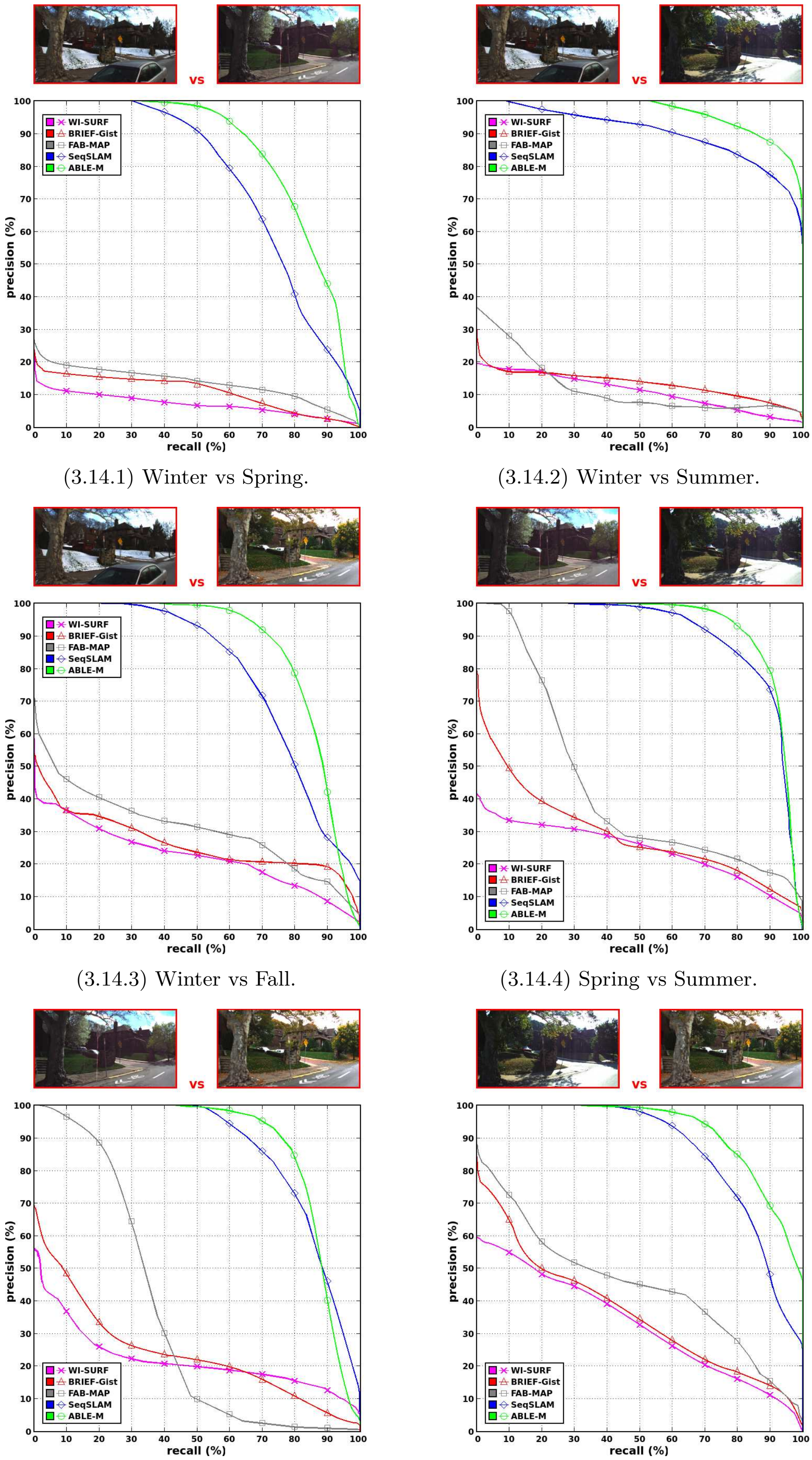


(3.13.4) With rotation and translation.

Figure 3.13: ABLE-M vs SeqSLAM in the Nordland dataset with changing field of view. Precision-recall curves are depicted with the aim of supporting the performance comparison when translation and rotation effects are introduced in the images of the dataset (between the sequences of winter and spring). These new tests demonstrate that the proposal presented by ABLE-M outperforms other algorithms based on sequences such as SeqSLAM, especially when the field of view is changing. It must be also noted that both methods decrease its precision with respect to the case of an invariant field of view, but this is much more accentuated in the SeqSLAM results than in the obtained by our final approach.

3.5.2.4 ABLE-M in the CMU-CVG VL Dataset

The CMU-CVG VL dataset contains sequences acquired by a car in different months of the year around a same route. In this case, apart from seasonal variations, there are changes on the camera field of view between the images recorded for a same place, which allows to confirm the satisfactory performance of ABLE-M in these conditions compared to other methods based on hand-crafted features, as shown in Fig. 3.14.



(3.14.1) Winter vs Spring.

(3.14.2) Winter vs Summer.

(3.14.3) Winter vs Fall.

(3.14.4) Spring vs Summer.

(3.14.5) Spring vs Fall.

(3.14.6) Summer vs Fall.

Figure 3.14: ABLE-M vs state-of-the-art methods in the CMU-CVG VL dataset. Precision-recall curves are depicted with the aim of supporting the performance comparison. The four sequences corresponding to the seasons are evaluated among them for all the algorithms. An illustrative frame from the sequences matched is shown in order to visually understand the complexity of place recognition in each case.

The precision-recall curves presented in Fig. 3.14 demonstrate again the importance of using sequences instead of single images in visual place recognition based on hand-crafted descriptors. For this reason, SeqSLAM and ABLE-M also obtain the most successful performances in this dataset. However, the changes on the field of view produce a higher difference between the results obtained by SeqSLAM and ABLE-M in this dataset. Now, ABLE-M is clearly the best option in the tests carried out across the different seasons. It must be noted that the sequences associated with each season in Fig. 3.14 correspond to a specific route of the dataset: winter (21/12/10), spring (21/04/11), summer (01/09/10) and fall (28/10/10).

All these results corroborate the benefits of using ABLE for life-long visual localization with monocular cameras. In the following sections, the precision of our proposal will be evaluated for the stereo and panoramic versions.

3.5.2.5 ABLE-S in the KITTI Dataset

The experiments presented for stereo cameras are focused on the KITTI Odometry dataset, which is selected for these tests because it is a consolidated benchmark commonly used in autonomous driving, robotics and computer vision. In this case, this dataset is not as long as some of the datasets chosen for monocular tests, but it contains several challenging situations for visual place recognition in long-term conditions, such as perceptual aliasing between scenes, dynamic objects in places and a considerable amount of loop closures in the different recorded sequences (defined in the ground-truth presented in Appendix B).

The first test carried out in the sequence 06 of the KITTI Odometry dataset is related to the performance of our D-LDB features applied in the global description approach proposed by ABLE-S. We check out the precision of some descriptors as core of our method compared to D-LDB, as shown in Fig. 3.15. The descriptors used in the presented comparison were described in Chapter 2. They can be grouped into two main categories: vector-based (SIFT, SURF, HOG) and binary (BRIEF, BRISK, ORB, FREAK, LDB, D-LDB).

As deduced from the precision-recall curves presented in Fig. 3.15, we decided to use LDB and D-LDB as core of our description approach because they achieve a higher accuracy for solving localization problems, especially if they are compared to other state-of-the-art descriptors based on hand-crafted features. In these results, it is also remarked that our D-LDB features are more effective than LDB for stereo place recognition. This is due to the addition of disparity information in the global description process provided by D-LDB. In this way, spatial information about scene can be better captured, with the aim of solving common life-long localization difficulties, such as perceptual aliasing.

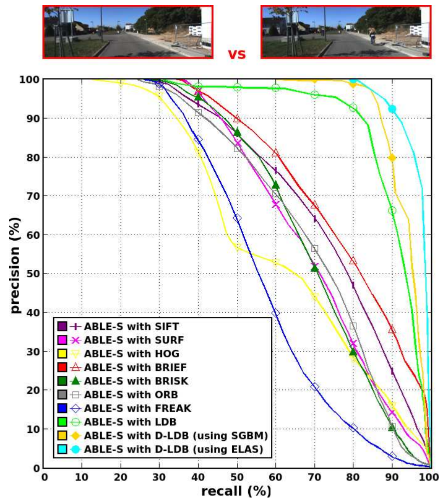


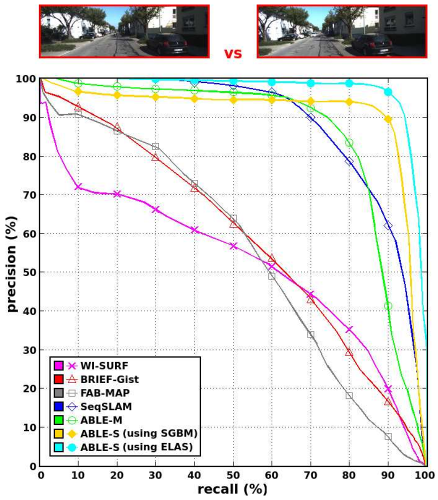
Figure 3.15: ABLE-S using different features as core in the KITTI Odometry dataset. Precision-recall curves are depicted with the aim of supporting the performance comparison in the sequence 06 of the dataset. D-LDB precision is evaluated using the different stereo matching methods implemented for disparity calculation.

In the evaluations depicted in Fig. 3.16, the performance of ABLE-S is compared against the main state-of-the-art algorithms for topological place recognition based on hand-crafted descriptors. In this case, we use four specific sequences from the KITTI Odometry dataset that are very representative and contain a high number of loop closures included in the traversed route, which are the sequences 00, 05, 06 and 13. The precision-recall curves represented in Fig. 3.16 also confirm the satisfactory performance of ABLE for stereo images with respect to the other methods based on hand-crafted features.

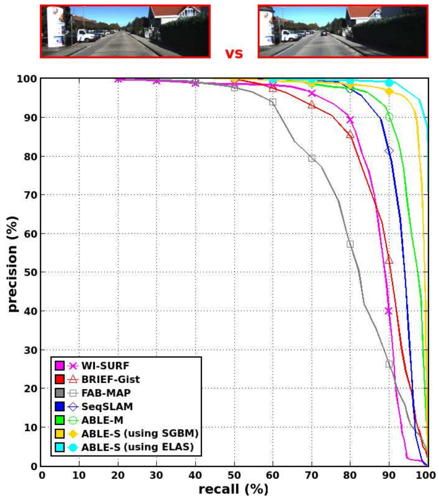
It must be noted that the methods based on single images (BRIEF-Gist, WI-SURF, FAB-MAP) obtain better results in Fig. 3.16 than in the monocular evaluations showed in Fig. 3.11 or Fig. 3.14, which is due to the minor distance traveled in the sequences of the KITTI Odometry dataset with respect to the datasets used in the monocular tests. The influence of using sequences instead of single images is much more evident when a very large amount of kilometers is processed, such as in the cases of the Nordland or the CMU-CVG VL datasets.

Apart from this, ABLE also outperforms the results of SeqSLAM in all the cases presented in Fig. 3.16. More specifically, the stereo version (ABLE-S) is more precise than the monocular version (ABLE-M), because of the use of disparity. As said in Section 3.4.2.2, a stereo matcher based on ELAS is implemented in our final approach for calculating disparity, which slightly improves the effectiveness of ABLE-S with respect to the traditional SGBM stereo matcher. These tests demonstrate the importance of the chosen algorithm for disparity computation in the D-LDB descriptor, because the spatial information captured by the features has a great dependence on the quality of the applied stereo matching method.

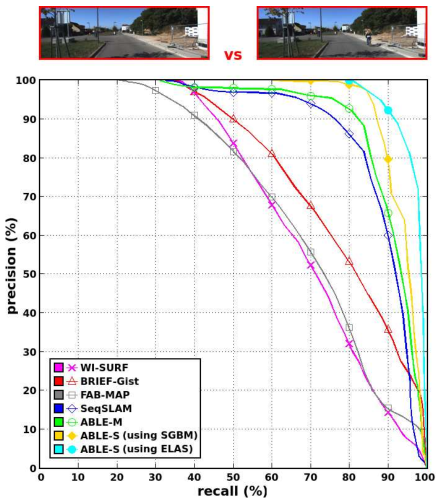
As a last contribution to the stereo tests explained in this section, Fig. 3.17 presents the revisited places detected by ABLE-S over the metric maps of the KITTI Odometry sequences analyzed in Fig. 3.16. These results support one of the practical applications of our visual place recognition method, where the identification of loop closures can help to detect and correct errors in several visual localization tasks. In Fig. 3.17, revisited places are depicted when the matched scenes exceed a certain similarity value in the distance matrices, as deduced from some of the explanations about thresholding given in Section 3.5.1.1. It can be seen how all the loop closures are correctly identified according to the ground-truth defined in Appendix B. More information about the application of our visual place recognition methods in loop closure detection for life-long visual localization will be given in Chapter 5.



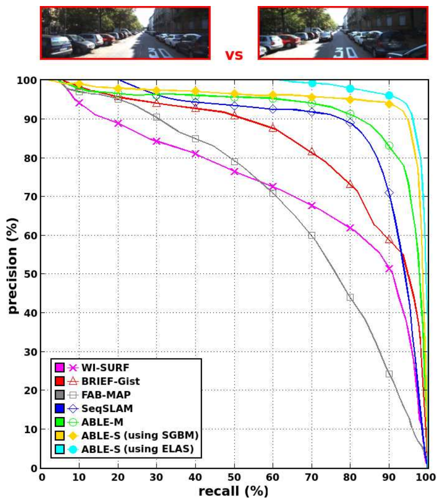
(3.16.1) Sequence 00.



(3.16.2) Sequence 05.

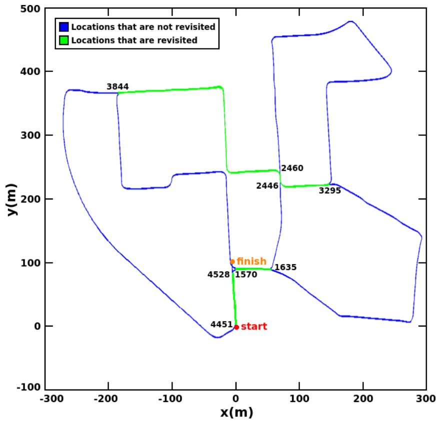


(3.16.3) Sequence 06.

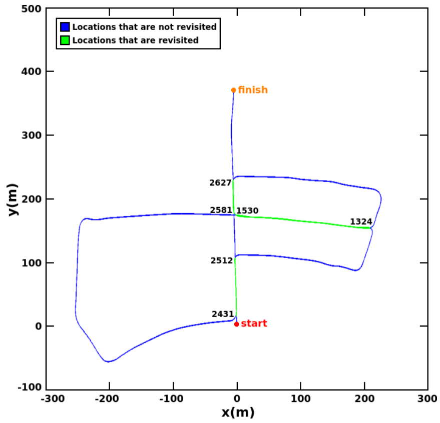


(3.16.4) Sequence 13.

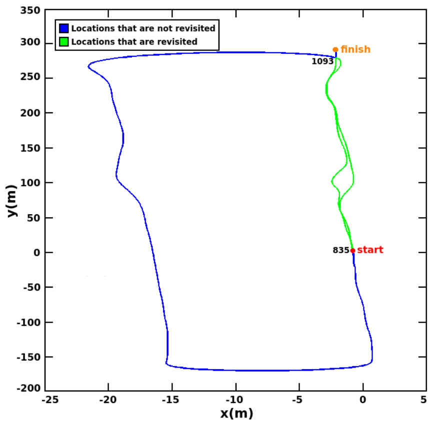
Figure 3.16: ABLE-S vs state-of-the-art methods in the KITTI Odometry dataset. Precision-recall curves are depicted with the aim of supporting the performance comparison over some of the most representative sequences in the dataset, jointly with some image samples of recognized locations. ABLE-S is tested using the different implementations of D-LBD depending on the stereo matching method: SGBM or ELAS.



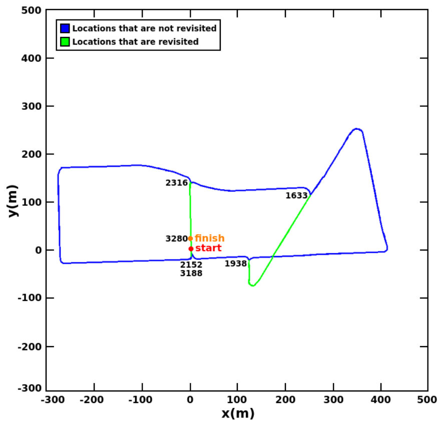
(3.17.1) Sequence 00.



(3.17.2) Sequence 05.



(3.17.3) Sequence 06.



(3.17.4) Sequence 13.

Figure 3.17: Recognized locations over the metric maps in the KITTI Odometry dataset. These results are obtained using ABLE-S over some of the most representative sequences in the dataset. Image indexes are shown in the zones where places start and finish to be revisited, with the aim of allowing a comparison to the ground-truth available in Appendix B.

3.5.2.6 ABLE-P in the Oxford New College Dataset

The Oxford New College dataset is selected for the last experiments carried out in this chapter, which are focused on the ABLE version designed for panoramic cameras: ABLE-P. This dataset is not very long with respect to the used in the previous tests, but it has been chosen because it allows to compare ABLE-P against the other two versions (ABLE-M and ABLE-S). This is possible because the dataset is not only recorded with panoramic cameras, but also with stereo.

In the precision-recall curves presented in Fig. 3.18, results about the effectiveness of each ABLE version are faced against the obtained by the main state-of-the-art proposals based on hand-crafted features in the Oxford New College dataset. Again, ABLE clearly outperforms the precision yielded by algorithms such as WI-SURF, BRIEF-Gist, FAB-MAP or SeqSLAM.

However, the most important conclusions about Fig. 3.18 are focused on the comparison between the three ABLE versions. First of all, the difference between the precision-recall curves obtained by ABLE-M and ABLE-S is appreciable. Similarly to the results presented in Section 3.5.2.5, in this case the application of stereo information is also decisive to improve the performance achieved by ABLE-M, due to the exploitation of the disparity integrated in D-LDB, which in these tests is definitively calculated using ELAS because of its demonstrated better accuracy. Even so, the most significant results are the provided by ABLE-P. The main reason of its improved performance with respect to the rest of the methods is that ABLE-P is the only algorithm that can detect the locations traversed in an opposite direction along the route (bidirectional loop closures), which are one of the most challenging characteristics of the Oxford New College dataset. This is due to the usage of cross-correlation for matching the subpanoramas contained in the panoramic images, which allows to clearly identify the locations revisited in a different direction in these cases, as justified in some of the explanations given along Section 3.4.3 and in Fig. 3.7. Besides, the applications of the visual place recognition provided by ABLE-P for detecting unidirectional and bidirectional loop closures will be explained in detail in Chapter 5.

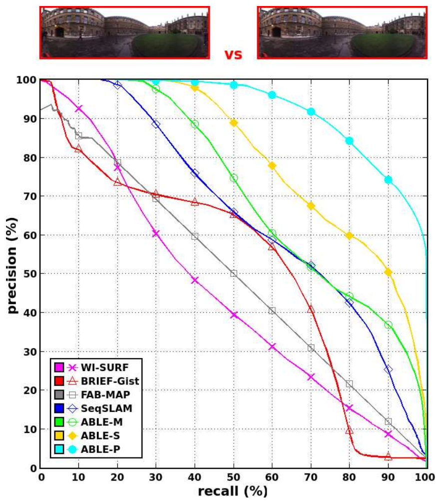


Figure 3.18: ABLE versions vs state-of-the-art methods in the Oxford New College dataset. Precision-recall curves are depicted with the aim of supporting the performance comparison, jointly with some image samples of recognized locations. The results obtained using ABLE-P are the best due to the application of the extra information provided by panoramic images.

3.5.2.7 Results about the Efficiency of ABLE

Apart from the advantages related to the effectiveness of ABLE in life-long visual topological localization, the efficiency yielded by our method is also very important in order to achieve a long-term operation. For this reason, in Fig. 3.19 we provide a graph where the evolution of the processing times consumed by ABLE-M to match a determined number of images is compared to the achieved by some of the state-of-the-art algorithms based on hand-crafted features. These times have been obtained in tests over images of the Nordland dataset using a standard computer with an Intel Core i7 2,40 GHz processor and a 8 GB RAM.

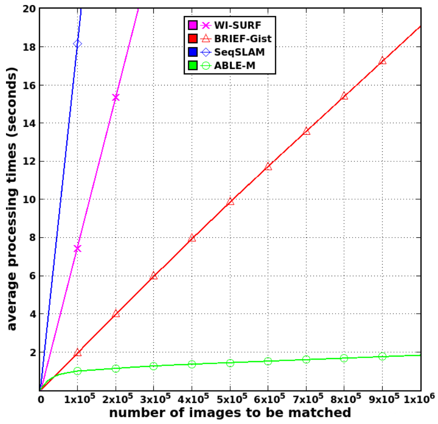


Figure 3.19: Average processing times in matching: ABLE-M vs state-of-the-art methods. These results are computed in the Nordland dataset. It demonstrates how the application of the ANN search implemented by ABLE reduces the matching costs to a sublinear time.

In Fig. 3.19, the evaluation shows a large-scale matching of places until a million of images. Due to the graph scale requirements, the average processing times are only shown until a maximum of 20 seconds. The curves obtained by SeqSLAM and WI-SURF exceed this time limitation before arriving to the million of images, because these methods require a much higher memory resources and computational costs in image matching with respect to the other methods, which are based on binary descriptors that can be much faster matched by means of the Hamming distance. In the case of the approaches that apply a binary matching, BRIEF-Gist has a progressive increment of the average processing times, because it is based on a linear search. However, ABLE-M applies an ANN search using a multi-probe LSH index, which decreases the accumulated computational cost of matching the binary sequences with the previously processed to a sublinear time in a large-scale context. For instance, when 100000 images are matched, the average processing time for BRIEF-Gist is 1.98 s, and ABLE-M obtains 0.93 s. Nevertheless, in the case of 1000000 images, BRIEF-Gist achieves about 19 s, while ABLE-M only needs 1.87 s, which clearly evidences the high influence in the efficiency of our sublinear search.

ANN is used in the three ABLE versions, so the average processing times presented by ABLE-M will be also sublinear for the cases of ABLE-S and ABLE-P. Even so, there are slight differences in the individual processing times depending on the extra information that must be computed by each method. In order to understand these differences, we show Table 3.2, where the average times per individual image description and matching in each version are included. The description process has a higher computational cost in ABLE-S because of the extra effort that requires the calculation of the disparity. However, the matching costs are the bottleneck of our system. In this sense, the individual costs of a matching between two images are more critical in ABLE-P, because the cross-correlation of panoramas adds an extra computation.

Table 3.2: Comparison between the average processing times of each ABLE version in milliseconds (ms). The costs for describing and matching an individual image can be observed in these results about the efficiency of our method in the Oxford New College dataset.

	ABLE-M	ABLE-S	ABLE-P
Description	$0.11\ ms$	$7.11\ ms$	$0.54\ ms$
Matching	$2.29 \cdot 10^{-5}\ ms$	$4.51 \cdot 10^{-5}\ ms$	$5.71 \cdot 10^{-4}\ ms$

3.6 Conclusions and Contributions

Along this chapter, our proposal for visual place recognition based on hand-crafted binary features has been extensively justified, jointly with the description of the main contributions of ABLE, which have been validated by a wide set of experiments. The different versions (ABLE-M, ABLE-S, ABLE-P) constitute a relevant innovation in different life-long topological localization fields. These versions are completely adaptable to several types of cameras and can take advantage of the information acquired by monocular, stereo and panoramic images in each case. Besides, the three versions present novel characteristics that have enhanced the performance of our system: the representation of places as sequences of images instead of single images, the illumination invariant transformation, the application of efficient LDB and D-LDB binary features and the matching based on an ANN search jointly with LSH.

Due to the described contributions, ABLE achieves a satisfactory precision in long-term conditions, which is corroborated by the exhibited results, especially if our method is compared to the main state-of-the-art algorithms based on hand-crafted features, such as WI-SURF, BRIEF-Gist, FAB-MAP or SeqSLAM. The efficiency of our approach is also significant, mainly because of the application of global binary features, which supply an image description methodology with a low computational cost and a fast matching capacity.

Additionally, it must be remarked that we also contribute an open version of our method to the community named OpenABLE, which is publicly available[1] and described in detail in Appendix A.

In future works, our research is extensible to new concepts that could improve even more the accuracy of these kinds of techniques in long-term situations, such as localization across seasons. New alternatives recently followed in visual place recognition could be employed, such as the usage of semantic information [Drouilly et al., 2015, Mousavian et al., 2015] or CNNs [Sünderhauf et al., 2015b, Sünderhauf et al., 2015a]. In fact, the application of features based on CNNs is going to be studied in our method presented in Chapter 4.

[1] OpenABLE is available from: http://www.robesafe.com/personal/roberto.arroyo/openable.html

Chapter 4

Visual Place Recognition based on Learned CNN Features

In this chapter, a novel visual place recognition method based on CNN features is described in order to improve the precision with respect to the achieved in Chapter 3 in complex long-term situations. In this sense, the major part of the state-of-the-art proposals for topological localization are focused on description methods that compute conventional hand-crafted features, as exposed in Chapter 2. However, approaches based on deep learning can be a promising alternative for a more accurate identification of locations, especially in the particular topic of localization across seasons.

Motivated by the achievements recently accomplished in the computer vision community by means of deep learning, we consider to develop other solution for visual place recognition that builds upon the concept of CNN features, due to the success of proposals based on CNNs in complex problems such as large-scale image classification [Krizhevsky et al., 2012]. In fact, recent researches have started to study the application of CNN-based features in topological localization [Sünderhauf et al., 2015b, Sünderhauf et al., 2015a]. In this regard, our approach is focused on transferring the knowledge learned in image classification problems by CNNs to our own interests for visual place recognition in a long-term context.

Along the following pages, we describe our proposed method called Convolutional Neural Network for Visual Topological Localization (CNN-VTL). Its main properties are explained jointly with the contributions provided to the state of the art. Besides, a wide set of results is presented in order to validate the specific characteristics of CNN-VTL and compare it against the main state-of-the-art methods in several challenging datasets commonly used for evaluation in visual localization across seasons. In addition, some final conclusions about the principal ideas derived from this chapter are given. According to this, we also discuss about the main pros and cons of using CNN-based features instead of hand-crafted descriptors in problems related to life-long visual localization.

4.1 Overview: The CNN-VTL Method

CNN-VTL exploits the advantages of powerful feature representations via CNNs in order to perform a robust topological vision-based localization across the seasons of the year, as introduced in the graphical explanation of the approach given in Fig. 4.1. Our proposal contributes several new concepts for improving the effectiveness and efficiency of CNN features in visual place recognition. It must be noted that previous work about our CNN-VTL was published in [Arroyo et al., 2016a]. Here, we present the complete method, including extensive tests and results.

Inspired by the success of how image representations learned with CNNs on large-scale annotated datasets can be transferred to other recognition tasks [Oquab et al., 2014], we consider the possibility of using pre-trained CNN features for the identification of locations in our CNN-VTL method. The goal of this strategy is that the learned features may not be specific to a particular dataset. Our objective is that they can be generalizable to different environments.

Nevertheless, the main inconvenience of using CNNs is that they are usually expensive in terms of computational costs and memory resources, which sometimes is a problem for real-time processing in embedded systems. For this reason, we present an efficient CNN model for computing our features that provides not only a high precision in life-long visual localization, but also a reduced consumption of memory and processing costs. With the aim of achieving this efficiency maintaining the effectiveness of our approach, we provide several innovative proposals regarding to the current state of the art in visual place recognition based on CNN features. These proposals are mainly the following:

- An improved CNN architecture based on some of the ideas of [Krizhevsky et al., 2012] and [Chatfield et al., 2014]. Our model is adapted and reduced to the requirements of our visual place recognition system. The CNN is pre-trained in a different dataset (ImageNet [Deng et al., 2009]) with respect to the used in tests to demonstrate how transferable the learned features are [Yosinski et al., 2014].

- A novel fusion of the features obtained by the convolutional layers that improves the performance. The redundancy of these fused features is subsequently decreased by applying feature compression techniques, such as Random Bit Selection (RBS) [Yang and Cheng, 2014] or Principal Component Analysis (PCA) [Wold et al., 1987].

- A binarization of the final reduced features with the aim of improving the matching of locations by computing an efficient Hamming distance instead of a traditional L_2-norm.

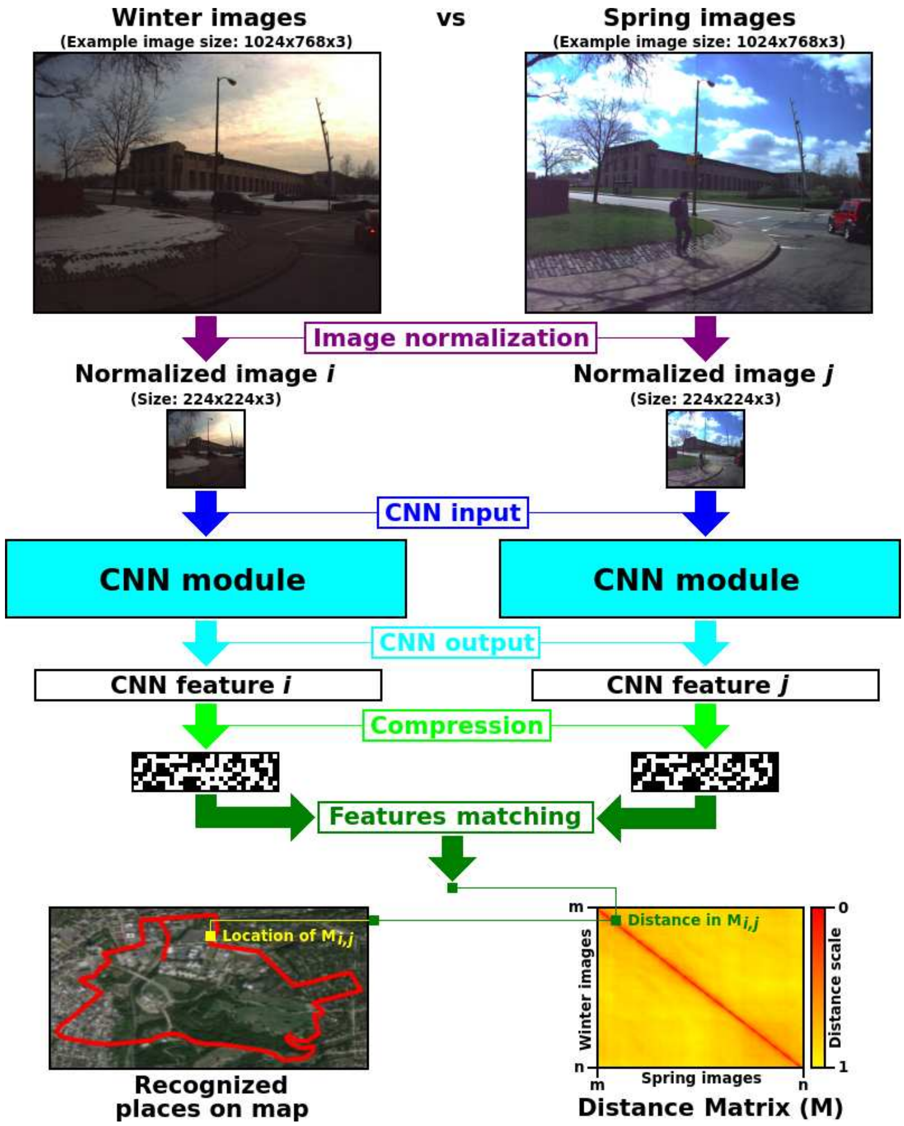


Figure 4.1: General diagram about the CNN-VTL method. The image samples correspond to the CMU-CVG VL dataset. It must be noted that the internal architecture of the CNN module is extensively described in Fig. 4.2.

4.2 Network Architecture

The architecture designed for our CNN model follows some concepts of the Visual Geometry Group Fast network (VGG-F) presented in [Chatfield et al., 2014], which takes as reference a similar structure to the one suggested by AlexNet [Krizhevsky et al., 2012], but including some improvements for a faster processing. To achieve a more efficient performance, VGG-F reduced the number of convolutional layers originally proposed by AlexNet to eight learnable layers, five of which are convolutional, and the last three are fully-connected. In the results provided by [Chatfield et al., 2014], it is corroborated how this simplification does not have a significant impact in the effectiveness for image recognition, but it greatly decrease the computational costs. Inspired by these experiments, we implement a much more simplified approach. We also take into account the study about CNNs performance in place recognition carried out by [Sünderhauf et al., 2015a], which demonstrates that fully-connected layers are not so effective as the convolutional ones in this task, as certified by our own observations. For this reason, our final model eliminates fully-connected layers and is mainly based on five convolutions. According to all the previous considerations, our architecture is graphically described in Fig. 4.2.

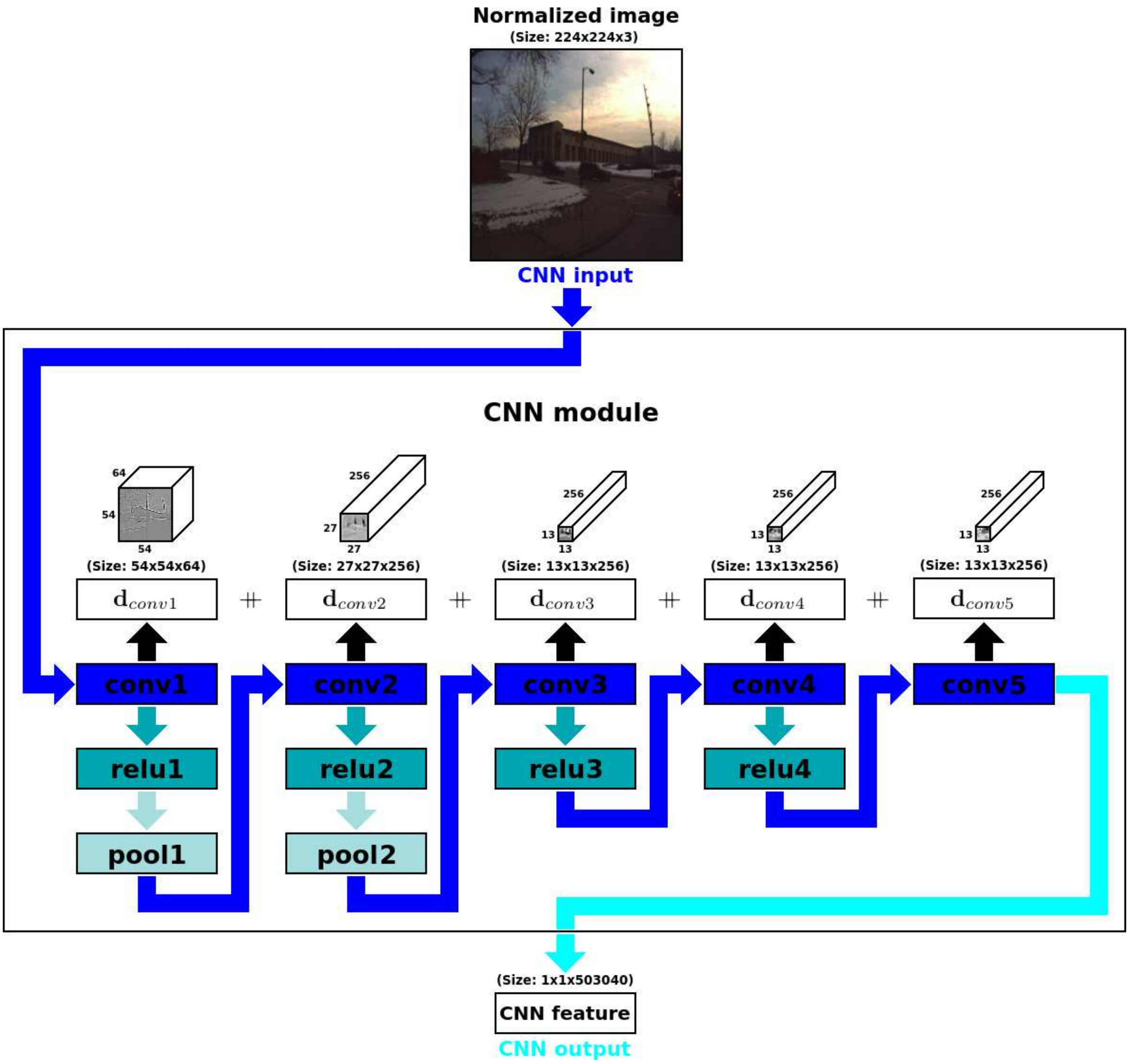


Figure 4.2: CNN-VTL internal architecture. The description process takes as input an image normalized to a size of 224x224x3 and returns as output a CNN feature. Internally, our CNN is divided into different layers that capture image information at several levels and granularities, where five convolutions are computed. The features obtained for each convolution are fused to form the final feature, which will be compressed in subsequent stages of our system.

Following the ideas exposed in works such as [Oquab et al., 2014, Yosinski et al., 2014], our CNN-VTL architecture is based on a pre-trained model over the ImageNet dataset, in order to confirm the generalization of the automatically learned features. This will demonstrate that the description power acquired by the proposed CNN features is transferable to the specific datasets used in the tests of our visual topological localization across seasons.

The architecture designed in CNN-VTL is modeled thanks to some of the functionalities provided by the MatConvNet toolbox defined in [Vedaldi and Lenc, 2015], which allows to create a great variety of different wraps to define the layers in the network. As can be seen in Fig. 4.2, our architecture is formed by three main types of layers: convolutional layers, Rectified Linear Unit (ReLU) layers and spatial pooling layers. The mechanisms for computing the information of each layer inside of our CNN-VTL are not trivial. For this reason, now we must provide a more detailed explanation about how the different layers work in our specific model.

4.2.1 Convolutional Layers

Convolutions are the basic layer in any CNN and they are usually the main level in all the blocks. In the case of our CNN-VTL, five convolutional layers are on the top of the five blocks that form the proposed architecture. The derivatives associated with convolutions are solved with techniques of backpropagation. Each convolutional layer receives an input map ($\mathbf{x} \in \mathbb{R}^{H \times W \times D}$) and a bank of filters with multiple dimensions ($\mathbf{f} \in \mathbb{R}^{H' \times W' \times D \times D''}$), returning the subsequent output ($\mathbf{y} \in \mathbb{R}^{H'' \times W'' \times D''}$). It must be noted that in our implementation any bias input is processed. Taking into account the inputs and the output, the internal computation of our convolutions is represented by Eq. 4.1:

$$y_{i''j''k''} = \sum_{i'=1}^{H'} \sum_{j'=1}^{W'} \sum_{k'=1}^{D} f_{i'j'k} \times x_{i''+i'-1,j''+j'-1,k',k''}. \tag{4.1}$$

4.2.2 ReLU Layers

ReLU is an activation function used by our CNN model. In CNN-VTL, a ReLU layer is located after all the convolutional layers, except in the case of the last convolution ($conv_5$), where it is unnecessary according to our feature generation proposal. We choose the ReLU activation function because it is very easily and efficiently computed, in contrast to other more complex functions, such as sigmoidals. The activation in ReLU is simply thresholded at zero, as exposed in Eq. 4.2:

$$y_{ijk} = \max(0, x_{ijk}). \tag{4.2}$$

4.2.3 Spatial Pooling Layers

Pools are a very important layer in the proposed architecture. Spatial pooling basically decreases the amount of parameters and computational costs over our CNN using a non-linear downsampling, which also allows to control problems derived from overfitting. CNN-VTL contains two pools in the lower part of its two first blocks connected to $conv_2$ and $conv_3$, which are sufficient to achieve the desired reduction of the processed data. In our case, a max pooling operator is implemented. It computes the maximum response of each feature channel in a $H' \times W'$ patch. The internal application of this layer is formulated in Eq. 4.3:

$$y_{i''j''k} = \max_{1 \leq i' \leq H', 1 \leq j' \leq W'} x_{i''+i'-1, j''+j'-1, k}. \tag{4.3}$$

4.3 Image Description of Locations in CNN-VTL

The model defined by the layers of CNN-VTL allows to carry out a strong visual description of places at different image levels and granularities for the proposed topological localization across seasons. The visual features automatically learned by our network can be used now by applying the designed convolutions.

4.3.1 Fusion of Convolutional Features at Different Levels

With the aim of forming a more robust final descriptor ($\mathbf{d}_{cnn}$) from our CNN-VTL architecture, we concatenate ($+\!\!+$) the vectorized features obtained by the n different convolutional layers ($\mathbf{d}_{conv_n}$):

$$\mathbf{d}_{cnn} = \mathbf{d}_{conv_1} +\!\!+ \mathbf{d}_{conv_2} +\!\!+ \mathbf{d}_{conv_3} +\!\!+ \mathbf{d}_{conv_4} +\!\!+ \mathbf{d}_{conv_5}. \tag{4.4}$$

The goal of the strategy formulated in Eq. 4.4 is to conserve the multi-resolution information provided by each convolution, which acts as a local and translation invariant operator, as stated in [Vedaldi and Lenc, 2015]. In works such as [Sünderhauf et al., 2015b], only the features generated by the third convolutional layer ($\mathbf{d}_{conv_3}$) are considered in the place description process, which produces a loss of invariance. Due to this, [Sünderhauf et al., 2015b] needs to use a complex and expensive algorithm to obtain several region landmarks per image and compute their respective CNN features in order to maintain the robustness when revisited locations have important changes on the field of view. On the other hand, our CNN-VTL can directly use whole images thanks to the invariance procured by the fusion of the convolutional outputs, which is a more efficient approach.

The different features contained in $\mathbf{d}_{cnn}$ are initially returned by the CNN in a float format. For this reason, with the aim of facilitating a subsequent binarization, we cast these features into a normalized 8-bit integer format ($\mathbf{d}_{cnn}^{int}$) by following Eq. 4.5, where $min = 0$ and $max = 255$:

$$\mathbf{d}_{cnn}^{int} = (\mathbf{d}_{cnn} - \min(\mathbf{d}_{cnn})) \frac{max - min}{\max(\mathbf{d}_{cnn}) - \min(\mathbf{d}_{cnn})} + min. \tag{4.5}$$

4.3.2 Techniques for Features Compression

The length of the CNN-based descriptor (l_{cnn}) acquired after the fusion of the five convolutional outputs can be calculated as presented in Eq. 4.6, where h_{conv_n}, w_{conv_n}, d_{conv_n} are the height, width and dimensions of each convolution, respectively:

$$l_{cnn} = \sum_{n=1}^{5} h_{conv_n} \times w_{conv_n} \times d_{conv_n}. \tag{4.6}$$

If we solve Eq. 4.6 using the output sizes of the convolutions applied in our CNN-VTL architecture (see Fig. 4.2), we obtain $l_{cnn} = 503040$ bytes. This length can be excessive for efficiently performing the subsequent features matching. Due to this, we apply reductions to this size in order to analyze how they affect to the accuracy of the place recognition method. This is motivated by works such as [Milford, 2012], which evidences that a handful of bits is sufficient for conducting an effective vision-based navigation. This is also corroborated by ABLE in some of the experiments presented in Chapter 3. Besides, in recent studies [Alcantarilla and Stenger, 2016], several traditional hand-crafted binary descriptors are tested to conclude that a remarkable precision can be achieved using a small fraction of the total number of bits from the whole descriptor. We demonstrate in the experiments presented in Section 4.5.2.1 that the features extracted from CNNs have a similar behavior to the observed in [Alcantarilla and Stenger, 2016].

In this regard, we propose the application of two typical compression techniques with the aim of reducing the dimensions of the final descriptor: RBS and PCA. Both methods have some pros and cons, which will be studied and compared in the results presented in Section 4.5.2.1 and more specifically in Table 4.1.

4.3.2.1 Random Bit Selection (RBS)

In order to efficiently decrease the size of our CNN descriptors without losing a great accuracy, the redundant features can be omitted to compress the final length. In works such as [Calonder et al., 2010] or [Yang and Cheng, 2014], methods based on RBS have demonstrated to be an efficient and effective alternative with respect to more complex algorithms. In fact, the evaluation presented in the binary description performed by LDB in [Yang and Cheng, 2014] yielded surprisingly favorable results, where the precision of RBS is close to the one achieved using more refined methods, such as entropy-based.

We also implement a similar random selection of features in order to compress our CNN descriptor in an easy and efficient way. This technique randomly chooses a specific set of features and applies the same selection in all the following descriptions to match the same correlative features. A proportional number of features is randomly selected for each layer to preserve as possible the multi-resolution provided by our fusion of convolutional features at different granularities.

4.3.2.2 Principal Component Analysis (PCA)

PCA is a well-known procedure for identifying a smaller number of uncorrelated variables from a large set of data. The goal of this compression technique is to define the maximum amount of variance with the fewest number of principal components, as explained in works such as [Wold et al., 1987].

According to this, we take advantage of the PCA method for reducing the dimensionality of the descriptor processed by means of our CNN-VTL. As will be shown in the results presented in Table 4.1 of Section 4.5.2.1, the compression carried out using PCA has a slightly better performance than the obtained by RBS. However, it must be also noted that PCA also requires a higher computational cost to compute the compression than RBS, which is based on a much more simple reduction.

4.3.3 Binarization of Features

Taking into account the normalization of the CNN descriptors into integer values that was shown in Eq. 4.5, it is trivial to apply a binarization to the features. In this regard, each obtained value can be easily represented as an 8-bit binary feature. The objective of this conversion is to make possible the application of the efficient Hamming distance in matching, which can be used when binary features are processed.

4.4 Image Matching of Locations in CNN-VTL

The bottleneck of our system for visual place recognition based on CNN features resides in the matching of descriptors for identifying locations, because the number of images to be matched is increased in each iteration, while the description costs are constant along the time. Apart from features compression, other techniques can be applied for reducing the computational costs of matching tasks. One of them is the usage of the Hamming distance for obtaining the similarity between features, which is more efficient than the cosine distance used in other works based on CNNs, such as [Sünderhauf et al., 2015b]. This efficiency is due to the simplicity of its calculation, which consists of an elementary XOR operation ($\oplus$) and a basic sum of bits, as formulated in Eq. 4.7. According to this, our CNN descriptors are previously binarized as explained in Section 4.3.3, because this is the main condition to correctly use the Hamming distance. As a final step in the matching process, a distance matrix (M) is computed by calculating the similarity between all the binary features ($\mathbf{d}_{cnn}^{bin}$):

$$M_{i,j} = \text{bitsum}(\mathbf{d}_{cnn_i}^{bin} \oplus \mathbf{d}_{cnn_j}^{bin}). \tag{4.7}$$

4.5 Experiments and Results

The experiments presented to evaluate the performance of CNN-VTL are based on analyzing their benefits and results in life-long localization along several datasets where topological place recognition can be tested in routes traversed across the seasons of the year. We also compare our CNN-based approach against the main state-of-the-art algorithms in these challenging long-term conditions.

4.5.1 Experimental Setup

Before showing the results obtained by means of the research described in this chapter, it is important to define the methodology used for testing the CNN-VTL method and the main characteristics of our experimental setup, which are described in the following sections.

4.5.1.1 Evaluation Methodology

The methodology applied for evaluating the performance of CNN-VTL is mainly based on precision-recall curves, in a similar way to the detailed in Section 3.5.1.1 for the tests computed in Chapter 3.

4.5.1.2 State-of-the-art Methods evaluated in Comparisons

We compare the accuracy of our solution against some of the main state-of-the-art works based on CNN features and hand-crafted descriptors. For evaluating WI-SURF and BRIEF-Gist, we use implementations of them based on the SURF and BRIEF descriptors provided by the OpenCV library. FAB-MAP is tested using the OpenFABMAP toolbox. The experiments for SeqSLAM are performed with OpenSeqSLAM. ABLE-M evaluations are computed thanks to the source code developed in OpenABLE (see Appendix A). Additionally, we implement the approach defined in [Sünderhauf et al., 2015a] based on the $conv_3$ features obtained from an AlexNet pre-trained in MatConvNet over the ImageNet dataset.

4.5.1.3 Tested Datasets

With the aim of demonstrating the capability of our CNN-VTL method, we carry out several tests in long-term situations using three datasets that contain several image sequences recorded for a same route across the seasons of the year: the Nordland dataset, the CMU-CVG VL dataset and the Alderley dataset. These tests allow us to analyze the behavior of our proposal over more than 3000 km and in the different conditions associated with each dataset.

4.5.2 Main Results

The results presented along this chapter are divided into three sections depending on the tested dataset. All the experiments are focused on monocular cameras because of the properties of our CNN-VTL architecture. Besides, the computational costs are also discussed. These evaluations corroborate the satisfactory performance in long-term conditions provided by our visual place recognition based on CNN features, especially if it is compared to the state of the art in this research line.

4.5.2.1 CNN-VTL in the Nordland Dataset

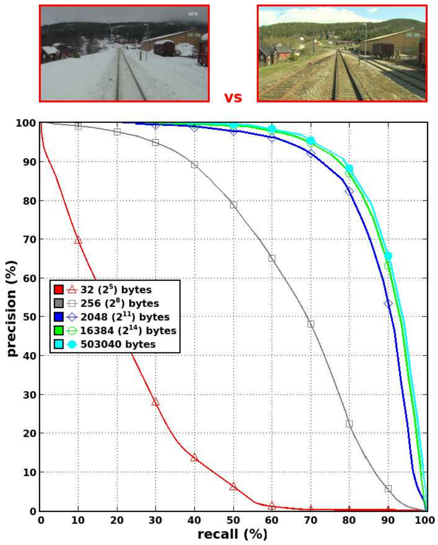
We present several precision-recall curves for the Nordland dataset obtained between the routes recorded in winter and spring, which is one of the most challenging evaluations because of the extreme variability in the visual appearance between both seasons.

In Fig. 4.3.3, it is corroborated how our approach focused on using a fusion of features from convolutional layers works much better than the features extracted from individual layers, which are proposed in works such as [Sünderhauf et al., 2015b] or [Sünderhauf et al., 2015a]. In this case, we also include a test for an added fully-connected layer (fc_6) to confirm its worse behavior in our problem.

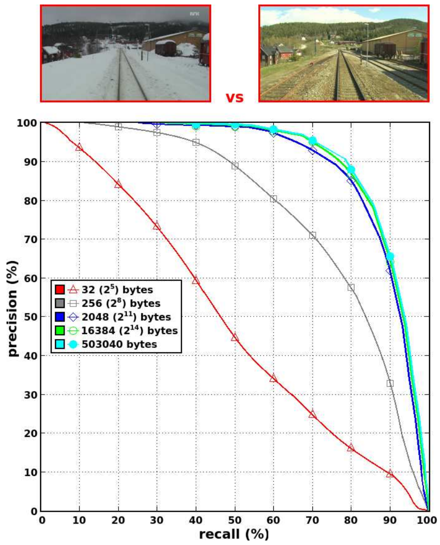
Precision-recall curves depicted in Fig. 4.3.1 and Fig. 4.3.2 validate our approach for reducing the amount of redundant information in our CNN features. Here, it is demonstrated that features can be highly compressed maintaining a remarkable performance, obtaining a slightly better accuracy using PCA than by means of RBS. Table 4.1 presents a more detailed study, where it can be seen how our initial CNN features can be reduced to 2048 bytes (a compression of 99.59%), losing only about a 2% of precision. In higher reductions, (256 or 32 bytes) the loss of accuracy is much more critical, especially in RBS. Table 4.1 also details the average speedups achieved in matching when features are compressed, which is proportional to the magnitude of the reduction.

The results in Fig. 4.3.4 show that our CNN-VTL proposal obtains a successful performance compared to the state-of-the-art algorithms based on traditional hand-crafted features. It must be also noted that the precision yielded by CNN-VTL is superior to the achieved in the curve computed for the CNN-based method for place recognition defined in [Sünderhauf et al., 2015a] (AlexNet $conv_3$).

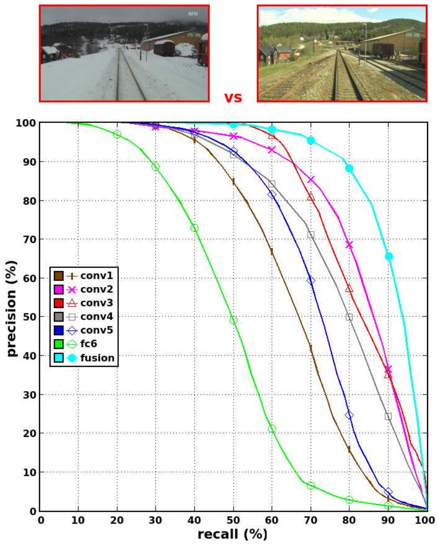
Additionally, in Fig. 4.4 we depict several image examples of challenging locations correctly detected by CNN-VTL across the seasons in the four sequences of Nordland dataset.



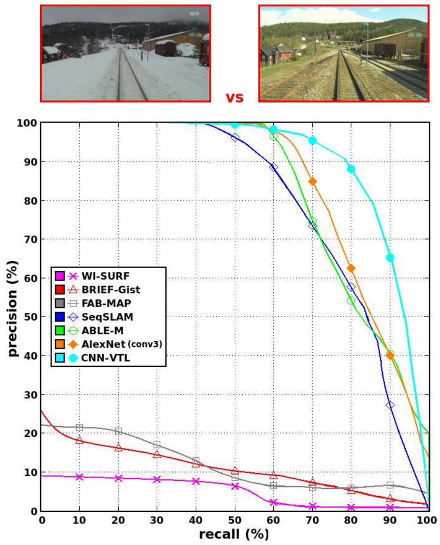
(4.3.1) Compression using RBS.



(4.3.2) Compression using PCA.



(4.3.3) CNN layers comparison.



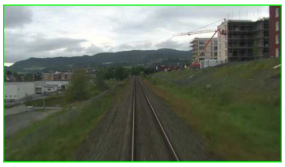
(4.3.4) State of the art comparison.

Figure 4.3: Results about CNN-VTL in the Nordland dataset (Winter vs Spring). Precision-recall curves are depicted with the aim of supporting the performance comparison, jointly with some image samples of recognized locations.

Table 4.1: Study about the performance of compressed features in CNN-VTL.

Size in bytes	**F_1-score using RBS**	**F_1-score using PCA**	**Percentage of compression**	**Average speedup for matching**
503040	0.899	0.899	0 %	None
16384	0.894	0.896	96.74 %	30x
2048	0.872	0.881	99.59 %	245x
256	0.651	0.768	99.94 %	1965x
32	0.216	0.484	99.99 %	15720x


(4.4.1) Winter (a).

(4.4.2) Spring (a).

(4.4.3) Summer (a).

(4.4.4) Fall (a).

(4.4.5) Winter (b).

(4.4.6) Spring (b).

(4.4.7) Summer (b).

(4.4.8) Fall (b).

(4.4.9) Winter (c).

(4.4.10) Spring (c).

(4.4.11) Summer (c).

(4.4.12) Fall (c).

(4.4.13) Winter (d).

(4.4.14) Spring (d).

(4.4.15) Summer (d).

(4.4.16) Fall (d).

(4.4.17) Winter (e).

(4.4.18) Spring (e).

(4.4.19) Summer (e).

(4.4.20) Fall (e).

(4.4.21) Winter (f).

(4.4.22) Spring (f).

(4.4.23) Summer (f).

(4.4.24) Fall (f).

(4.4.25) Winter (g).

(4.4.26) Spring (g).

(4.4.27) Summer (g).

(4.4.28) Fall (g).

(4.4.29) Winter (h).

(4.4.30) Spring (h).

(4.4.31) Summer (h).

(4.4.32) Fall (h).

Figure 4.4: Places detected by CNN-VTL across the seasons in the Nordland dataset. Some challenging examples of places correctly matched by our method in the four sequences are depicted.

4.5.2.2 CNN-VTL in the CMU-CVG VL Dataset

Apart from seasonal variations, the CMU-CVG VL dataset has changes on the camera field of view between the images recorded for a same place, which allows to test CNN-VTL in these challenging conditions with respect to the evaluations in the Nordland dataset, that provides a static field of view. In addition, this dataset includes 16 sequences captured along the months of the year, where our method can be exhaustively evaluated. This is represented by the example of Fig. 4.5, where a location correctly matched by CNN-VTL in all the sequences is shown.

We process precision-recall curves for sequences corresponding to the four seasons of the year in the six possible combinations. These evaluations are depicted in Fig. 4.6, where it can be observed how the different seasons affect to the performance. In general terms, the sequence captured in winter is again the most problematic in these tests, due to the extreme changes that a place suffers in this season: snow, less vegetation or different illumination, among others.

(4.5.1) 01/09/10.

(4.5.2) 15/09/10.

(4.5.3) 23/09/10.

(4.5.4) 01/10/10.

(4.5.5) 08/10/10.

(4.5.6) 19/10/10.

(4.5.7) 28/10/10.

(4.5.8) 03/11/10.

(4.5.9) 12/11/10.

(4.5.10) 22/11/10.

(4.5.11) 21/12/10.

(4.5.12) 02/02/11.

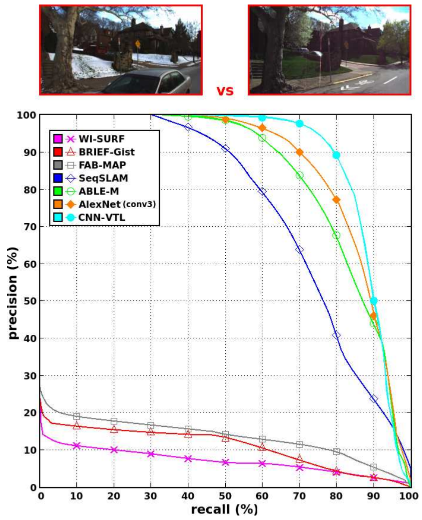
(4.5.13) 04/03/11.

(4.5.14) 21/04/11.

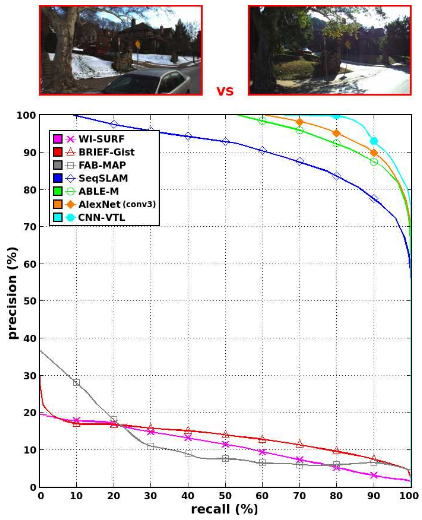
(4.5.15) 28/07/11.

(4.5.16) 02/09/11.

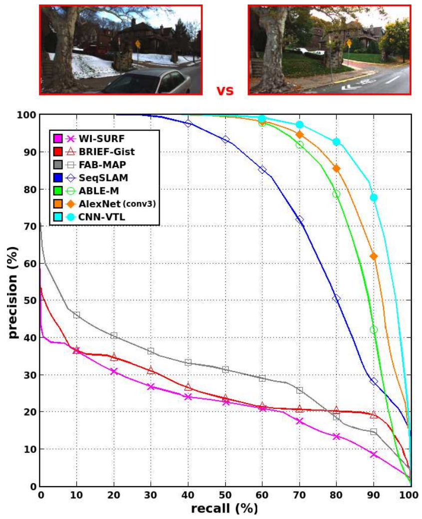
Figure 4.5: A place detected by CNN-VTL along the year in the CMU-CVG VL dataset.



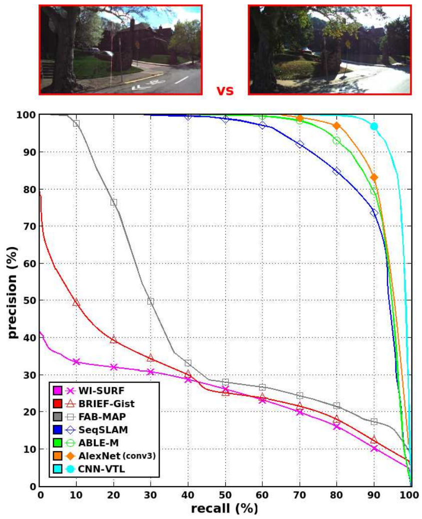
(4.6.1) Winter vs Spring.



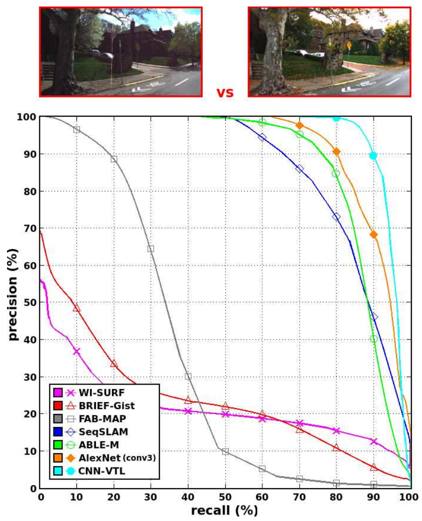
(4.6.2) Winter vs Summer.



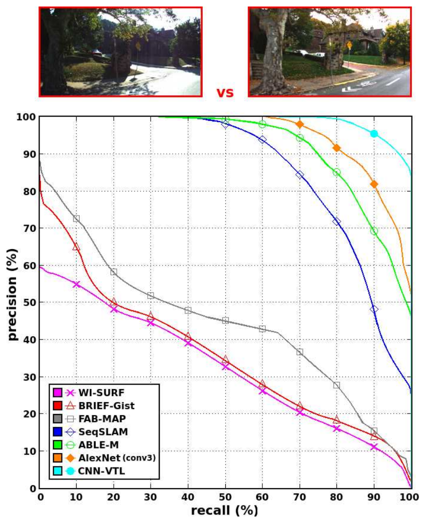
(4.6.3) Winter vs Fall.



(4.6.4) Spring vs Summer.



(4.6.5) Spring vs Fall.



(4.6.6) Summer vs Fall.

Figure 4.6: CNN-VTL vs state-of-the-art methods in the CMU-CVG VL dataset. Precision-recall curves are depicted with the aim of supporting the performance comparison. The four sequences corresponding to the seasons are evaluated among them for all the algorithms. An illustrative frame from the sequences matched is shown in order to visually understand the complexity of place recognition in each case.

The results presented in Fig. 4.6 compare the precision of CNN-VTL against the main state-of-the-art algorithms, but now in an environment where changes on the field of view have a negative effect in the accuracy of the major part of the methods. Our proposal obtains a better performance in this situation, because our fusion of CNN convolutional features at different levels and granularities provides a higher local and translation invariance with respect to approaches based on individual layers, such as [Sünderhauf et al., 2015a]. Apart from this, the precision-recall curves computed in Fig. 4.6 yield much worse results for algorithms based on traditional hand-crafted features using single images (WI-SURF, BRIEF-Gist, FAB-MAP) than using sequences (SeqSLAM, ABLE-M).

Moreover, Fig. 4.7 depicts a distance matrix computed over two sequences of the dataset by our CNN-VTL. This example evidences the reliable performance of our method, which is able to match almost all the images (except some few cases when a truck occludes the camera field of view). Additionally, Fig. 4.8 presents other complex situations where our method correctly detects a revisited place. In these cases, geometric change detection [Alcantarilla et al., 2016] could be applied to detect the specific variations in the structure of the matched place.

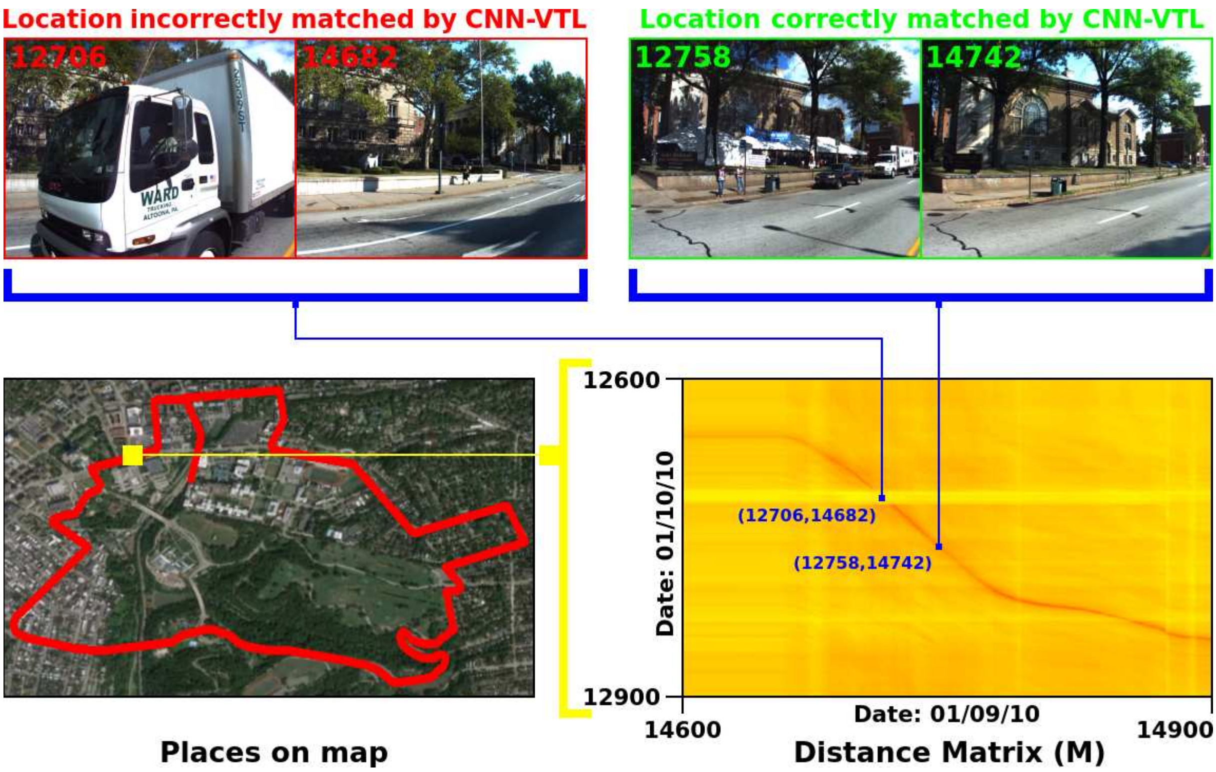


Figure 4.7: Distance matrix obtained by CNN-VTL in the CMU-CVG VL dataset. This representative example is processed in a part of the map between sequences recorded at 01/09/10 and 01/10/10. In almost all the cases, it can be observed that locations are correctly matched (see red line in distance matrix), except in a low amount of frames where a truck completely occludes the camera view (see frames 12706 and 14682). Other complex situations are correctly matched, such as locations with new buildings that change the initial appearance of a place (see frames 12758 and 14742).

(4.8.1) Challenging place matching (a).

(4.8.2) Challenging place matching (b).

(4.8.3) Challenging place matching (c).

(4.8.4) Challenging place matching (d).

(4.8.5) Challenging place matching (e).

(4.8.6) Challenging place matching (f).

(4.8.7) Challenging place matching (g).

(4.8.8) Challenging place matching (h).

(4.8.9) Challenging place matching (i).

(4.8.10) Challenging place matching (j).

(4.8.11) Challenging place matching (k).

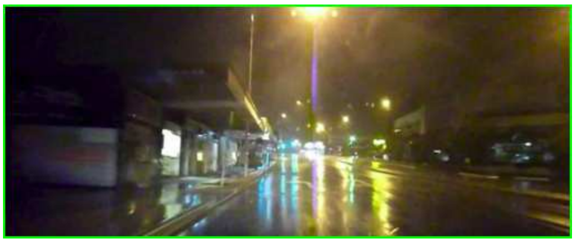
(4.8.12) Challenging place matching (l).

Figure 4.8: Complex locations matched by CNN-VTL in the CMU-CVG VL dataset. In this case, several challenging examples correctly identified in different pairs of sequences are depicted. The images show difficult cases with extreme changes over the past appearance of a place: new buildings, constructions, dynamic elements, important changes on the field of view or partial occlusions, among others.

(4.9.1) Day (a).

(4.9.2) Night (a).

(4.9.3) Day (b).



(4.9.4) Night (b).

(4.9.5) Day (c).

(4.9.6) Night (c).

(4.9.7) Day (d).

(4.9.8) Night (d).

(4.9.9) Day (e).

(4.9.10) Night (e).

(4.9.11) Day (f).

(4.9.12) Night (f).

Figure 4.9: Places detected by CNN-VTL at day and night in the Alderley dataset. Some challenging examples of places correctly matched by our method between the two recorded sequences are depicted.

4.5.2.3 CNN-VTL in the Alderley Dataset

The Alderley dataset comprises two sequences of images acquired in a stormy winter night and a sunny summer day. For this reason, apart from the typical seasonal changes previously studied, we can now perform evaluations under extremely variable illumination conditions in this dataset.

The precision-recall curves exposed in Fig. 4.10 present an acceptable accuracy for our CNN-VTL method in this challenging case. In fact, our approach obtains better results than the evaluated state-of-the-art proposals.

Moreover, in Fig. 4.9 are shown several image examples of challenging locations correctly detected by CNN-VTL along the two sequences of the Alderley dataset.

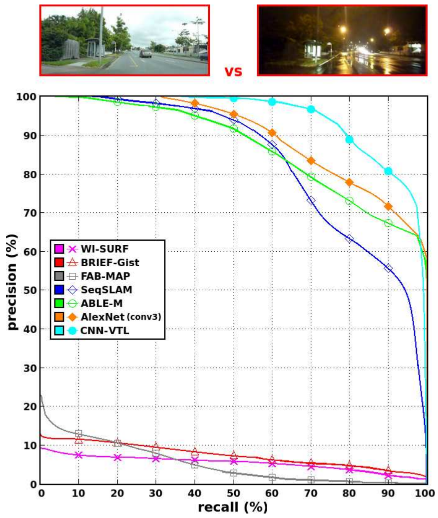


Figure 4.10: CNN-VTL vs state-of-the-art methods in the Alderley dataset. Precision-recall curves are depicted with the aim of supporting the performance comparison, jointly with some image samples of recognized locations.

4.6 Conclusions and Contributions

Along this chapter, our novel approach for life-long visual topological localization using CNN features (CNN-VTL) has extensively demonstrated its contributions to the robotics and computer vision communities in different areas. The proposed method has a valuable applicability in several fields related to tasks such as visual place recognition for loop closure detection based on distance matrices, which are usually indispensable in any SLAM or VO system. In fact, some of these applications are going to be studied in Chapter 5.

Our proposal is validated in challenging conditions derived from the extreme changes that the visual appearance of a place suffers across the four seasons of the year. A wide set of results in varied long-term scenarios corroborates the remarkable performance of our CNN-VTL compared against the main state-of-the-art algorithms based on hand-crafted descriptors, such as WI-SURF, BRIEF-Gist, FAB-MAP, SeqSLAM or ABLE-M, which was previously described in Chapter 3. Moreover, we have also evidenced that our method reports a better precision than other recent approaches which have studied the application of CNNs for place recognition in robotics [Sünderhauf et al., 2015a]. This is mainly due to our improved CNN architecture and to the fusion of the features acquired in several convolutional layers, that provides an enhanced local and translation invariance with respect to [Sünderhauf et al., 2015a], which is mainly based on CNN features from a $conv_3$ layer computed by a pre-trained AlexNet.

In addition, we have contributed an efficient model with the aim of decreasing the costs associated with CNN descriptors. We shown how our compression of features can reduce the redundancy of our descriptors in a 99.59%, while precision is only decreased in about a 2%, achieving a speedup in matching near to 245x in this case. Besides, our binarization of features allows to use the Hamming distance, that also represents a speedup to match locations. However, it must be noted that in spite of these reductions in processing times, CNN-VTL is in general terms more expensive than methods based on hand-crafted features. For example, ABLE-M reported a description time per individual image of 0.11 ms (see Table 3.2), while CNN-VTL requires around 50 ms in our tests with the same computer. Due to this, solutions based on hand-crafted features should not be discarded in embedded systems where computational resources are limited.

Moreover, although the accuracy of our CNN-VTL method can be considered quite satisfactory with respect to the main works in visual topological localization, there are some interesting future directions to follow such as:

- Test the application of sequences instead of single images in CNN-VTL, similarly to the proposed by ABLE in Section 3.2.1.
- Study the effect of compressing CNN features by means of other techniques, such as hashing methods.
- Perform an end-to-end training of a CNN architecture such as the one described in [Arandjelovic et al., 2016] and analyze its generalization properties to different domains.

Chapter 5

Life-Long Visual Localization using Topological Place Recognition

The application of topological place recognition provides relevant information about the situational awareness of the environment, which can be used in a varied range of life-long visual localization systems. In this regard, the methods presented in Chapters 3 and 4 are not only a valuable tool for using them in specific problems related to the identification of locations, but also for application in other tasks such as loop closure detection, the correction of localization measurements or the perception of geometric changes on scene.

The goal of this chapter is to present some of these applications for localization in a long-term context that have been developed in different researches along this thesis. It demonstrates the multiple utilities of topological place recognition and introduces several techniques that are very useful in order to achieve a robust life-long navigation for autonomous vehicles and mobile robots.

According to the previous considerations, the main applications described for life-long visual localization problems along this chapter are the following:

- The detection of loop closures as the most intuitive application of topological place recognition proposals. Similarity between the different locations can be analyzed and thresholded with the aim of identifying that a place has been revisited. We will show examples for several situations, where unidirectional and bidirectional loop closures are satisfactorily detected by our methods.
- The loop closure application to correct VO or visual SLAM measurements. In this sense, the amelioration of the estimated poses is useful for correctly projecting 3D point clouds in large-scale environments, as will be illustrated by several examples of 3D reconstructions.
- The recognition of geometric changes over places with the aim of facilitating the update of predefined maps, which is essential for autonomous driving systems in a long-term navigation.

5.1 Long-Term Loop Closure Detection based on Topological Place Recognition

The most direct application of topological place recognition is related to the detection of loop closures in localization and navigation problems. Loop closures are identified when a place is revisited, which causes that the topological localization methods yield a high similarity between the images captured for the same place in different instants of time, as shown in Fig. 5.1. This information is very helpful for reducing the negative effects of the accumulated drift in life-long visual localization.

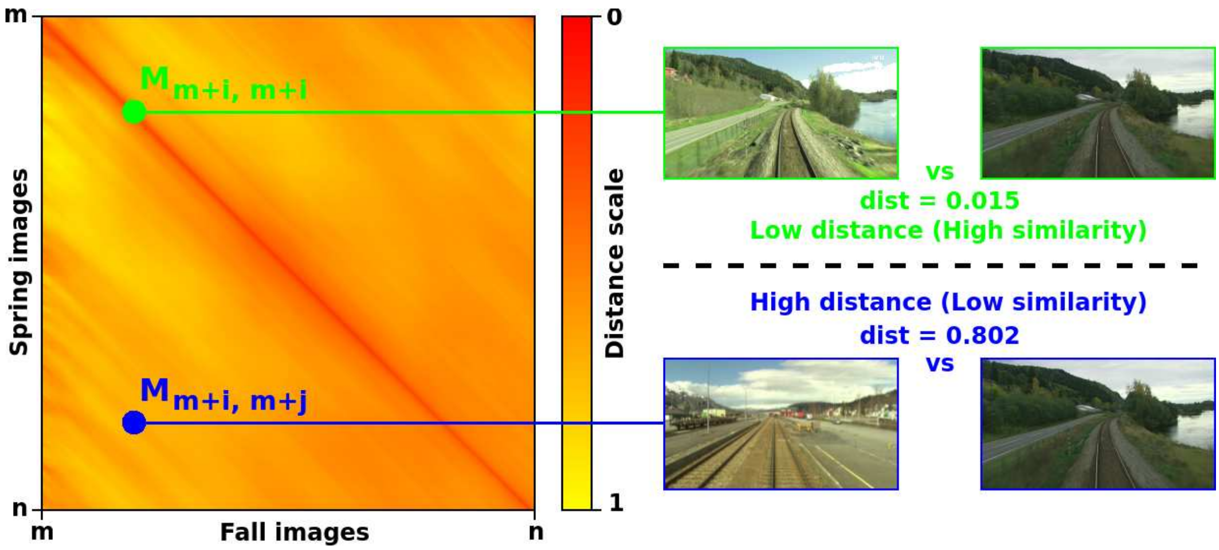


Figure 5.1: An example of similarities between places in long-term loop closure detection. The distance matrix depicted in this example is calculated using our visual place recognition in the Nordland dataset between the sequences of spring and fall. It can be seen how comparisons between a same place in different seasons have low distance values (high similarity), while different places have high distance values (low similarity). According to this, high similarities are typically associated with loop closures and can be detected by means of thresholding techniques.

The similarity values registered in a distance matrix (M) can be used for correcting localization measurements by means of loop closure detection. A threshold (θ) must be applied to discern if the similarity is sufficient to consider a loop closure between two places, as stated in Eq. 5.1:

$$loop\ \ closure = \begin{cases} true & \text{if } M_{i,j} < \theta \\ false & \text{otherwise} \end{cases} \tag{5.1}$$

It must be noted that adaptive thresholds can be also an interesting option for adjusting them according to the evolution of the environment conditions, as studied in some approaches such as [Lee and Pollefeys, 2014].

In any case, the inherent difficulties of correctly detecting a revisited location vary depending on the analyzed situation: unidirectional or bidirectional loop closures.

5.1.1 Unidirectional Loop Closures

Unidirectional loop closures are the standard case when a place is revisited. These loops appear if a location is traversed again in the same direction. Commonly, all topological place recognition algorithms are able to detect this type of loop closure.

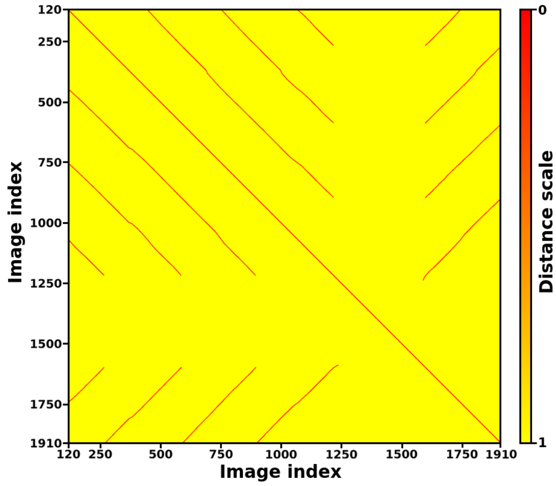
5.1.2 Bidirectional Loop Closures

Bidirectional loop closures are considered when a place is traversed by an autonomous vehicle or a mobile robot in a different direction, which adds an extra difficulty to detect them due to the important changes in the perspective of the scene. In these cases, the usage of panoramic cameras is usually crucial to correctly identify bidirectional loops, because of the multi-directional perception of the environment that they provide, as explained in Section 3.4.3. In fact, we designed the cross-correlation matching of panoramas presented in ABLE-P with the aim of detecting these situations, as was shown in Fig. 3.7.

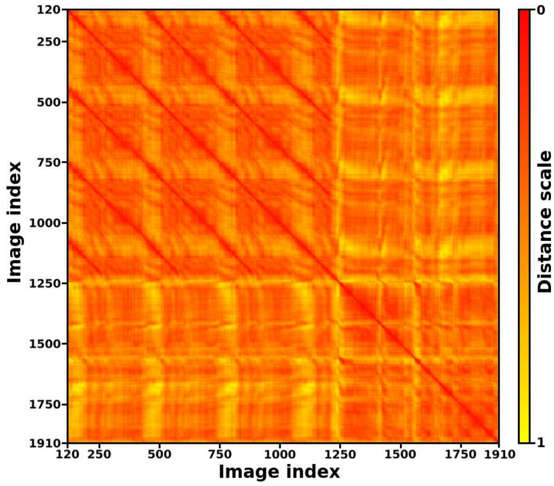
5.1.3 Examples of Application in the Oxford New College Dataset

In the results previously described along Section 3.5.2.6, several precision-recall curves evidenced the better performance of ABLE-P in the Oxford New College dataset with respect to other state-of-the-art approaches for visual place recognition based on hand-crafted features. The main reason of the improved performance of ABLE-P is that it can recognize the multiple bidirectional loop closures appeared in this dataset, while the other methods are not capable of identifying them. In order to demonstrate how ABLE-P detects these bidirectional loop closures, Fig. 5.2 shows a representative part of the distance matrices obtained by the different ABLE versions and the two state-of-the-art proposals that yielded the best results in this case (BRIEF-Gist and SeqSLAM), jointly with the ground-truth matrix. In the different matrices, unidirectional loop closures appear emphasized as right-side diagonals ($\searrow$) and the bidirectional ones as left-side diagonals ($\swarrow$). According to this, it can be seen that ABLE-P is the only method that represents the bidirectional loop closures in its distance matrix (see the inferior zone depicted in Fig. 5.2.6 to check it out).

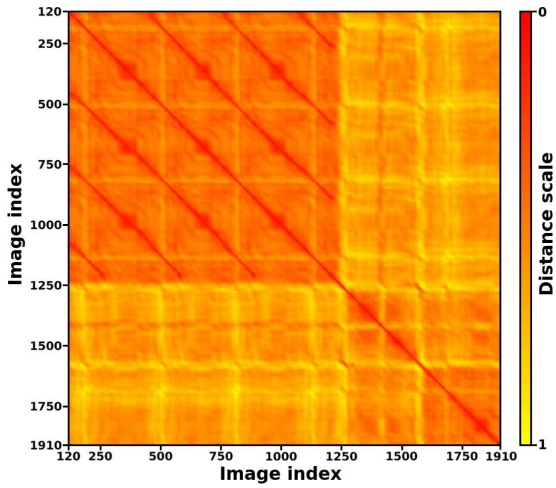
In addition, the loop closures identified in Fig. 5.2.6 by ABLE-P are also marked over the corresponding part of the map generated by means of the metric poses associated with the Oxford New College dataset, as shown in Fig. 5.3. The distance matrix is thresholded by following Eq. 5.1 to clearly distinguish the similarities associated with a loop closure (see Fig. 5.3.1). Progressive representations of the map are depicted each time that one of these loop closures is completed in the route. This valuable information can be applied for correcting the accumulated drift that appears in the represented metric measurements, as will be described in the following section.



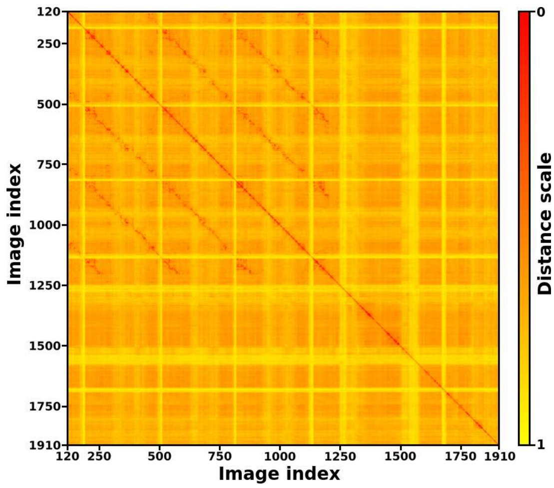
(5.2.1) Ground-truth matrix.



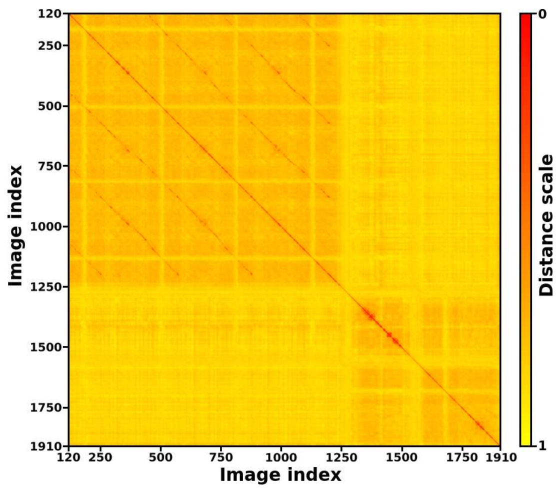
(5.2.2) Distance matrix using BRIEF-Gist.



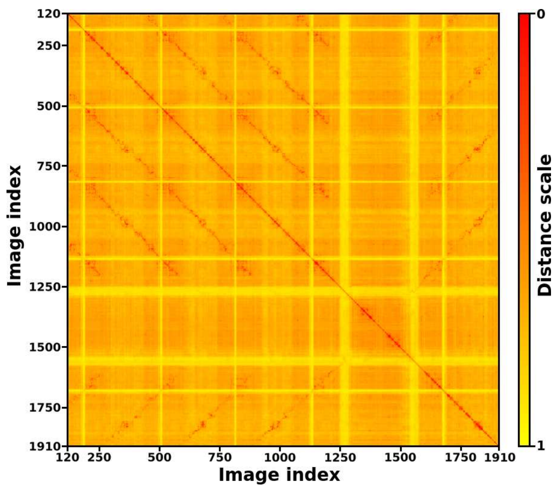
(5.2.3) Distance matrix using SeqSLAM.



(5.2.4) Distance matrix using ABLE-M.

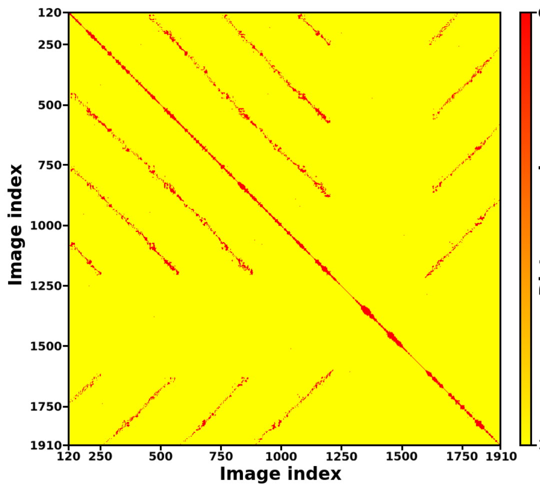


(5.2.5) Distance matrix using ABLE-S.

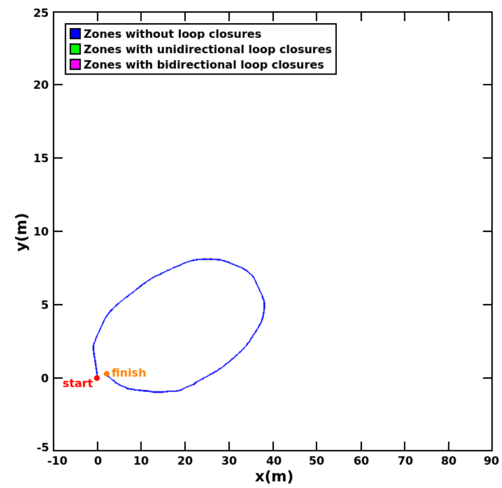


(5.2.6) Distance matrix using ABLE-P.

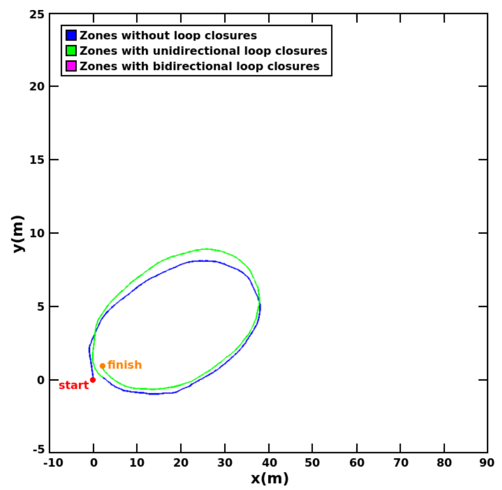
Figure 5.2: Distance matrices for loops detection in the Oxford New College dataset. The different ABLE versions are compared against some of the main visual place recognition methods in the state of the art of visual place recognition based on hand-crafted features. We only show a part of the distance matrices for a representative subset between images 120 and 1910, because of the limitations of document format. It can be seen how ABLE-P can identify the bidirectional loop closures.



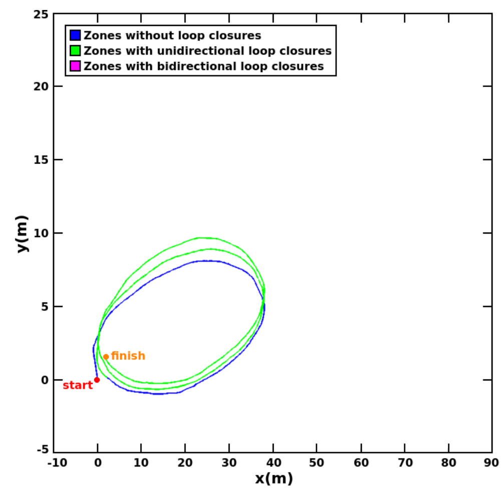
(5.3.1) Thresholded matrix using ABLE-P.



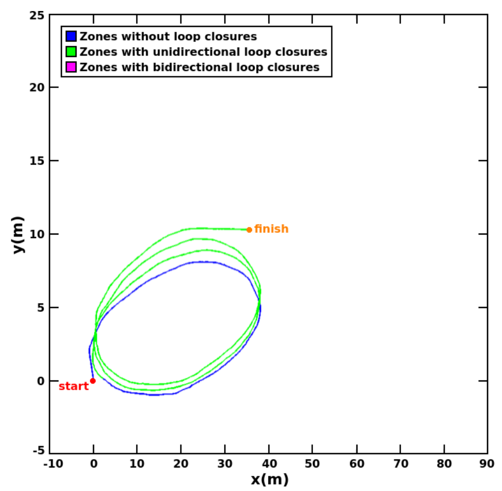
(5.3.2) Map after 1st lap (index = 449).



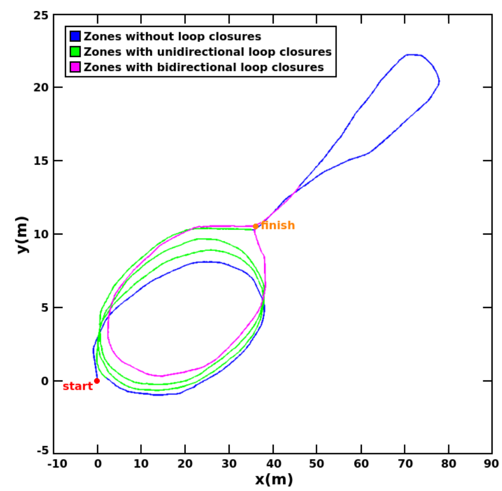
(5.3.3) Map after 1st loop (index = 752).



(5.3.4) Map after 2nd loop (index = 1066).



(5.3.5) Map after 3rd loop (index = 1219).



(5.3.6) Map after 4th loop (index = 1910).

Figure 5.3: Loops detected over a part of the map in the Oxford New College dataset. The unidirectional and bidirectional loop closures are detected by means of ABLE-P. We only show a part of the thresholded distance matrix for a representative subset between images 120 and 1910, because of the limitations of document format.

5.2 Correction of Localization Measurements based on Loop Closures

In this section, the recognition of revisited places provided by our loop closure detection techniques is applied in order to carry out a correction of localization measurements affected by the drift appeared in life-long visual localization. In this regard, VO proposals can use the identified loop closures to ameliorate the precision of the poses estimated by them.

Additionally, we also present results in which the corrected metric measurements obtained by means of camera-based approaches can be fused with other sensors such as GPS or LiDAR in order to solve the SLAM problem as a factor graph. According to this, we also represent city-scale 3D reconstructions based on point clouds, with the aim of depicting results where the application of these techniques can be seen.

5.2.1 VO Correction

VO has the goal of estimating the position and orientation of an autonomous vehicle or mobile robot by analyzing an image sequence acquired by cameras without any previous information about locations. Pairs of images are typically considered to match their keypoints extracted by means of local description methods and calculate the translation and rotation between two poses.

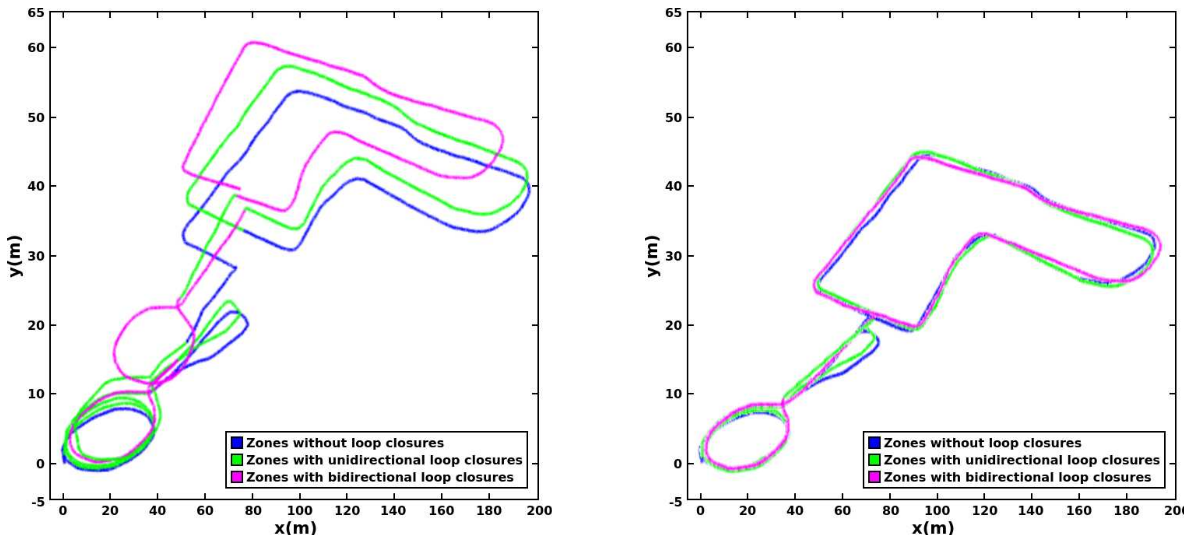
Unfortunately, VO commonly accumulates a drift when long periods of time are taken into account. This problem makes that localization tasks could not be completely reliable in these cases. For this reason, the information provided by standard VO algorithms usually gives errors in extended trajectories.

For this reason, we proposed a novel approach based on loop closure detection using ABLE for correcting the drift in VO, which is described in detail in [Caramazana et al., 2016, Arroyo et al., 2016b]. Here, the VO estimations are initially processed by means of the LIBVISO library [Kitt et al., 2010, Geiger et al., 2011]. The designed approach recognizes the revisited places and recalculates a corrected pose using a graph relaxation similar to the defined in [Bazeille and Filliat, 2010]. After a loop closure is detected in a specific frame, the drift of the pose currently estimated is compensated by taking into account the pose obtained when the place was previously traversed. Besides, an average deviation is subsequently computed after detecting the first pose corresponding to a loop closure. This information is employed to ameliorate the precision of the poses in the rest of the trajectory. The usage of these corrections in poses improves the accuracy initially obtained by only using VO without consider the progressive drift. This is corroborated in the following pages, where several examples of application carried out along different researches in this thesis are presented.

5.2.1.1 Examples of Application in the Oxford New College Dataset

In Fig. 5.3, we previously introduced some representations of the loop closures detected by applying our visual place recognition techniques over a part of the metric map corresponding to the Oxford New College dataset, where a drift can be clearly perceived in the registered trajectory.

Here, we present the whole map in Fig. 5.4.1, with the aim of evidencing how the drift is increased in the estimated trajectory in longer periods of time. However, the application of loop closure information for correcting the registered poses can be used in order to improve the accuracy of these measurements, as depicted in Fig. 5.4.2.



(5.4.1) Map after loop closure correction. (5.4.2) Map before loop closure correction

Figure 5.4: Loop closures used for VO correction in the Oxford New College dataset. It can be seen how the uncorrected measurements are improved by means of the loop closures identified by our topological place recognition techniques.

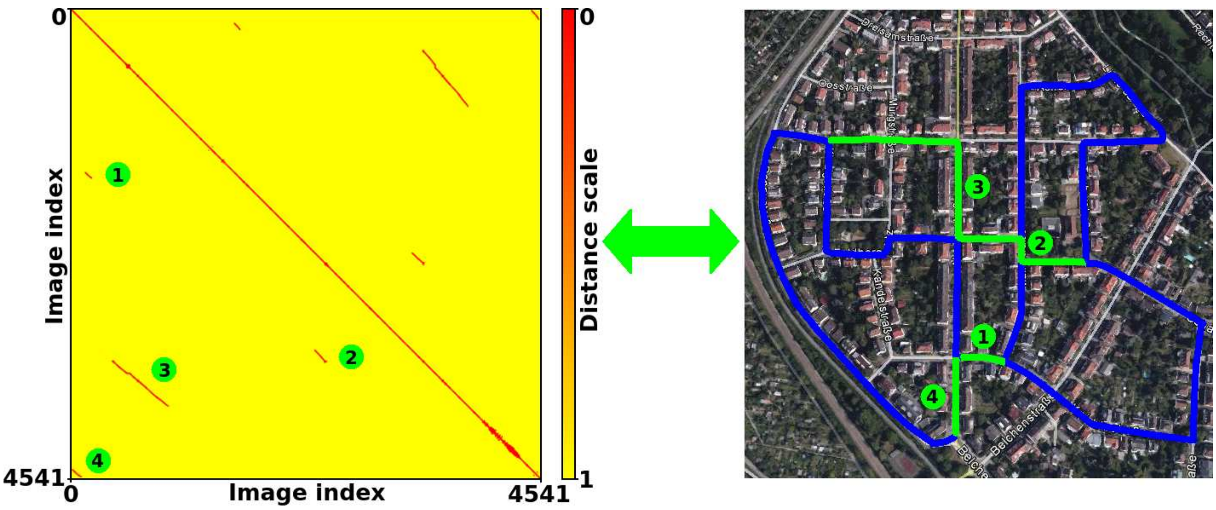
5.2.1.2 Examples of Application in the KITTI Odometry Dataset

We also conduct some experiments about the application of VO correction in the KITTI Odometry dataset, because it allows to test the usage of loop closure detection for ameliorating the accuracy of the poses in more varied environments focused on autonomous driving. In this case, stereo images are employed, which are also supported by the LIBVISO libraries and ABLE-S.

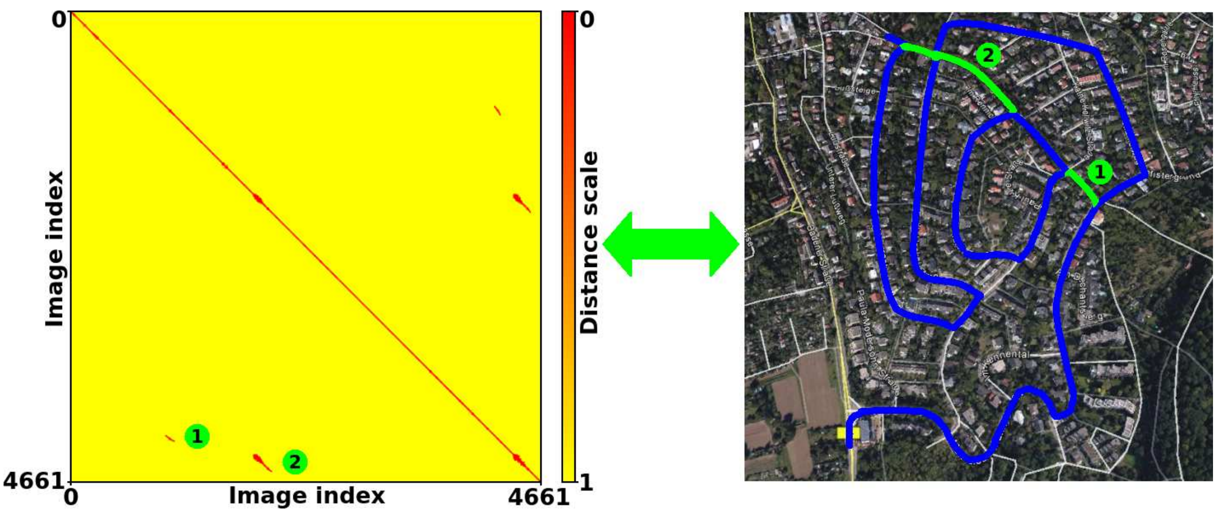
According to this, Figs. 5.5 and 5.6 show how the thresholded distance matrices generated by using our topological place recognition are directly applicable to loop closure detection for VO correction. We present several mapping results, where loop closures are annotated while they are identified by our approach. The maps are constructed using KML files for saving the registered poses, which are represented over satellite images by means of Google Earth®[1].

[1]Google Earth is currently available from: https://www.google.com/earth/

The results depicted in Figs. 5.5 and 5.6 correspond to some of the most representative sequences of the KITTI Odometry dataset, where several loop closures can be identified. More specifically, these experiments are performed over sequences 00, 02, 05 and 13. The ground-truth maps acquired for these sequences by means of a precise GPS sensor can be seen in Appendix B. If these ground-truth maps are compared to the trajectories obtained by considering our corrected VO measurements, it can be observed how the application of loop closure detection for reducing the problems derived from drift is also satisfactory in these cases. In fact, the corrected trajectories correctly fit in the streets represented over the satellite maps. Besides, the relationships between the identified loop closures and their positions in the map are clearly illustrated in the provided examples. The diagonals appeared in the distance matrices derived from a loop closure are related by a numerical identifier with their positions in the maps.

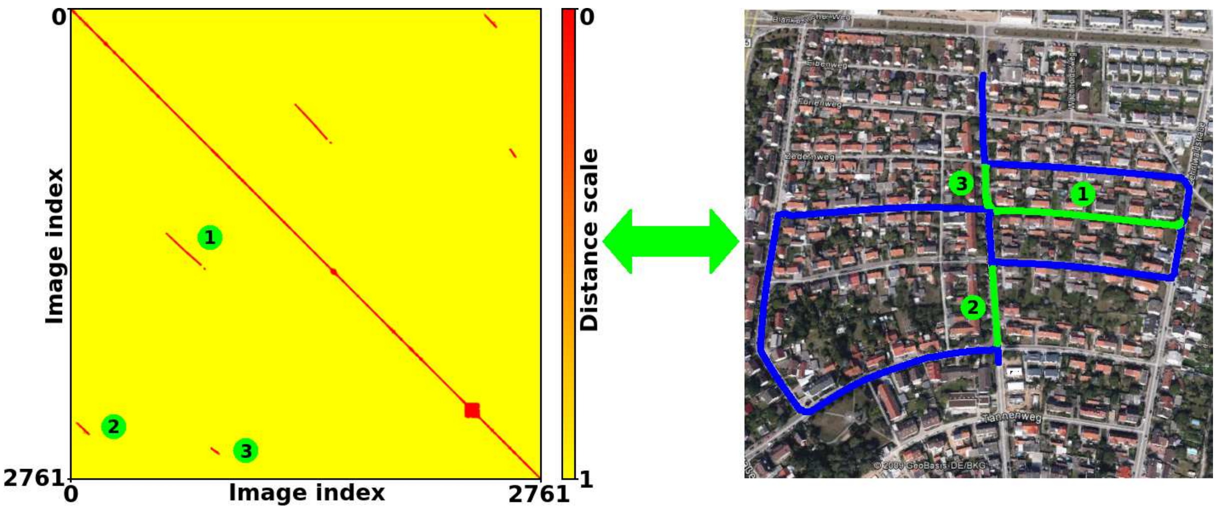


(5.5.1) Sequence 00.

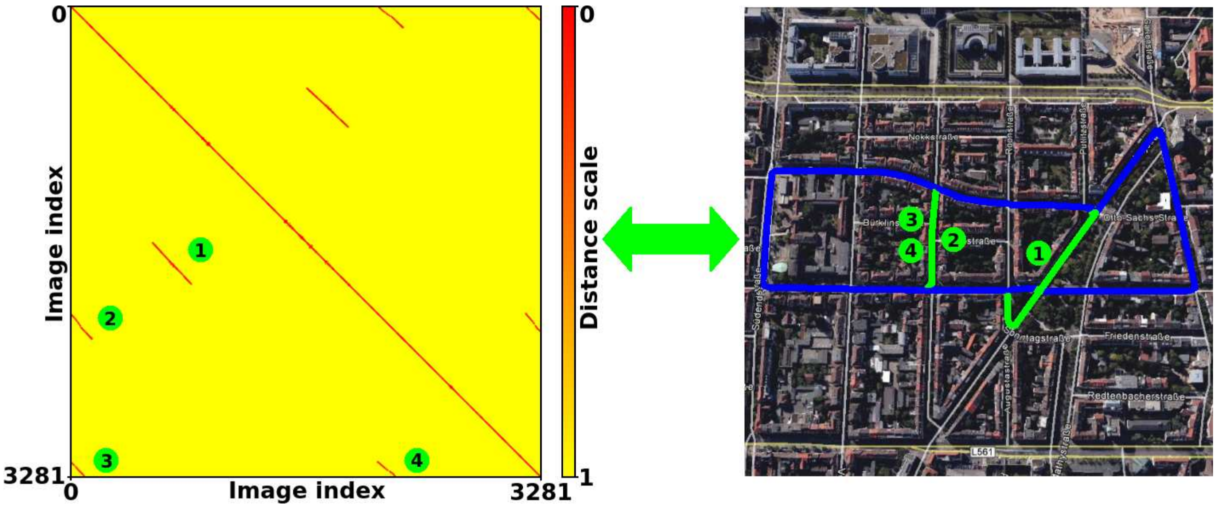


(5.5.2) Sequence 02.

Figure 5.5: Loop closures used for VO correction in the KITTI Odometry dataset (I). Distance matrices are depicted jointly with the corresponding satellite maps, where loop closures are related between them in both cases. It can be seen how the processed trajectories fit in the streets over the satellite maps.



(5.6.1) Sequence 05.



(5.6.2) Sequence 13.

Figure 5.6: Loop closures used for VO correction in the KITTI Odometry dataset (II). Distance matrices are depicted jointly with the corresponding satellite maps, where loop closures are related between them in both cases. It can be seen how the processed trajectories fit in the streets over the satellite maps.

5.2.2 Fusion of Corrected VO and GPS for 3D LiDAR reconstruction

The information provided by corrected VO poses can be fused with the obtained by means of other sensing technologies such as GPS or LiDAR, with the aim of generating 3D representations that improve the situational awareness of an autonomous vehicle or mobile robot in life-long visual localization. In this sense, a multi-sensor fusion SLAM problem can be defined as a factor graph, where each factor encodes the connectivity between the unknown variable nodes and the sensor measurements, as proposed in works such as the presented in [Indelman et al., 2013]. The objective is to estimate a set of n camera poses ($X = \{x_i\}_{i=1:n}$) and m reconstructed 3D points ($Y = \{y_j\}_{j=1:m}$), given a set of sensor measurements (Z). Here, we assume Z is composed of 3D LiDAR measurements (z_i^{proj}), a GPS measurement per camera pose (z_i^{gps}) and the corrected VO measurements between any two camera frames (z_{i_1,i_2}^{odo}).

Given the measurements of Z, the joint probability distribution of the navigation variables ($\theta = \{X^*, Y^*\}$) can be factorized as the product of the contribution of each individual factor in the graph:

$$P(X, Y; Z) \propto \prod_{k=1}^{K} f_k \left(\theta_v^k\right), \tag{5.2}$$

where θ_v^k represents a subset of the variable nodes and K is the total number of factors in the graph. Each factor f_k represents an error function that connects variables ($f^{proj}, f^{gps}, f^{odo}$) and measurements ($z_i^{proj}, z_i^{gps}, z_{i_1,i_2}^{odo}$), where Gaussian noise distributions are assumed for all the factors. Fig. 5.7 depicts a simplified example of this factor graph.

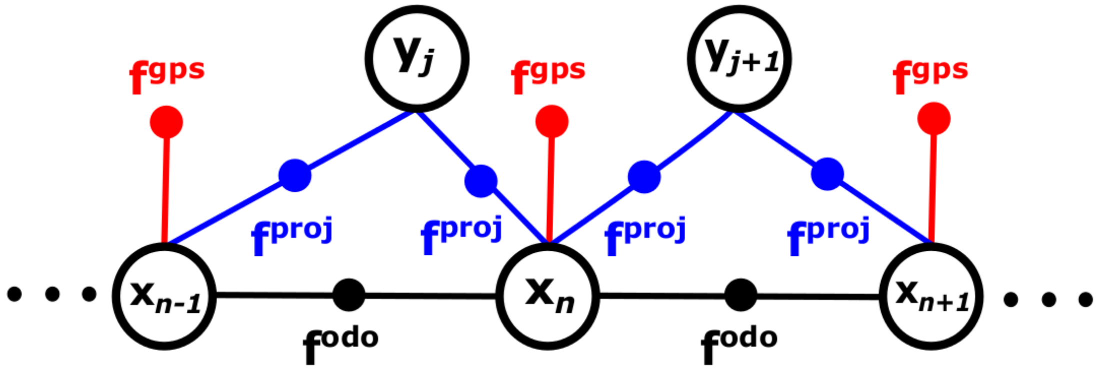


Figure 5.7: Factor graph encoding the multi-sensor fusion SLAM. The relationships between camera poses and scene structure are given in this representation.

Following the previous definitions, the corrected VO measurements are fused with GPS in order to obtain the final poses, which are used for the projection of the 3D points captured by the LiDAR. This is illustrated in the following pages, where several examples of application for 3D LiDAR reconstruction are presented.

5.2.2.1 Examples of Application in the CMU-CVG VL Dataset

The CMU-CVG VL dataset was evaluated in several results presented for our topological place recognition methods along Chapters 3 and 4. In those cases, images were only needed to perform the tests. However, this dataset has also registered data about GPS and LiDAR measurements associated with the acquired images.

As exposed in [Badino et al., 2012], the vehicle used for acquiring the dataset is equipped with two monocular cameras aimed 45° to the left and right of forward. Besides, two vertically scanning LiDARs provide range information, as described in the diagrams presented in Fig. 5.8. In addition, a GPS is also mounted in the vehicle, but the errors of this sensor can often be up to 10 meters in urban areas, as corroborated by the noisy measurements registered in the the CMU-CVG VL dataset. For this reason, the GPS poses employed for 3D reconstruction are interpolated with the information obtained by means of corrected VO measurements, with the aim of obtaining very precise trajectories for correctly representing the 3D points captured by the LiDARs.

(5.8.1) Vehicle and sensors.

(5.8.2) Diagram about cameras and LiDARs positions.

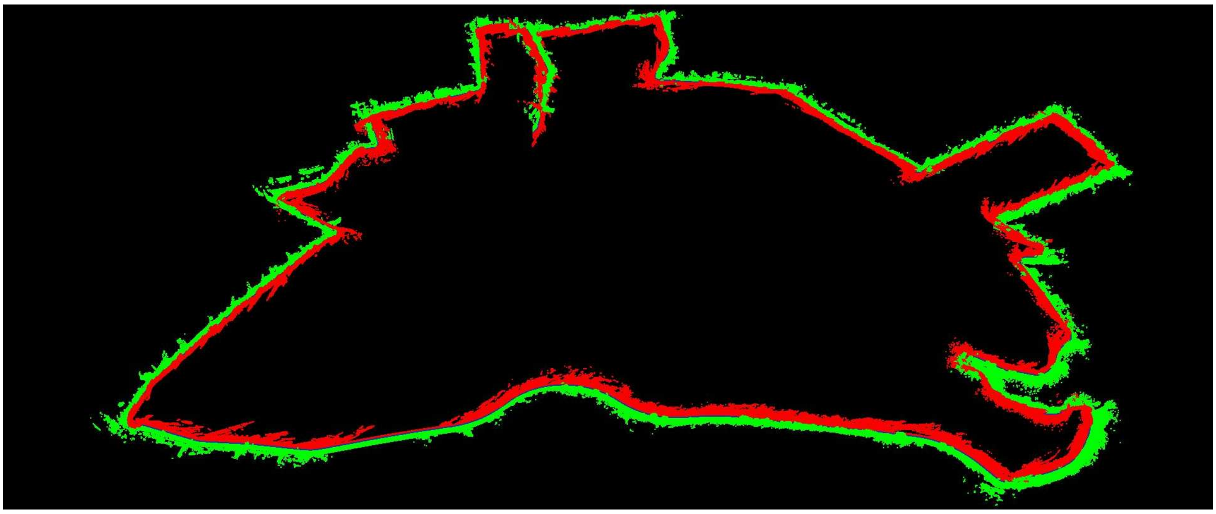
Figure 5.8: Sensors mounted in the car used for the CMU-CVG VL dataset.

Taking into account the configuration of the dataset, we present some examples of our city-scale 3D LiDAR reconstructions in Figs. 5.9 and 5.10. The poses of the car are shown in blue, the 3D points captured by the left LiDAR in green and the 3D points captured by the right LiDAR in red. The depicted representations are processed by applying some of the functionalities contained in the libraries provided by the Object-oriented Graphics Rendering Engine 3D (OGRE 3D)[2]. These libraries are a scene-oriented, flexible 3D software written in C++. They are designed to make it easier and more intuitive for developers to produce applications based on 3D graphics.

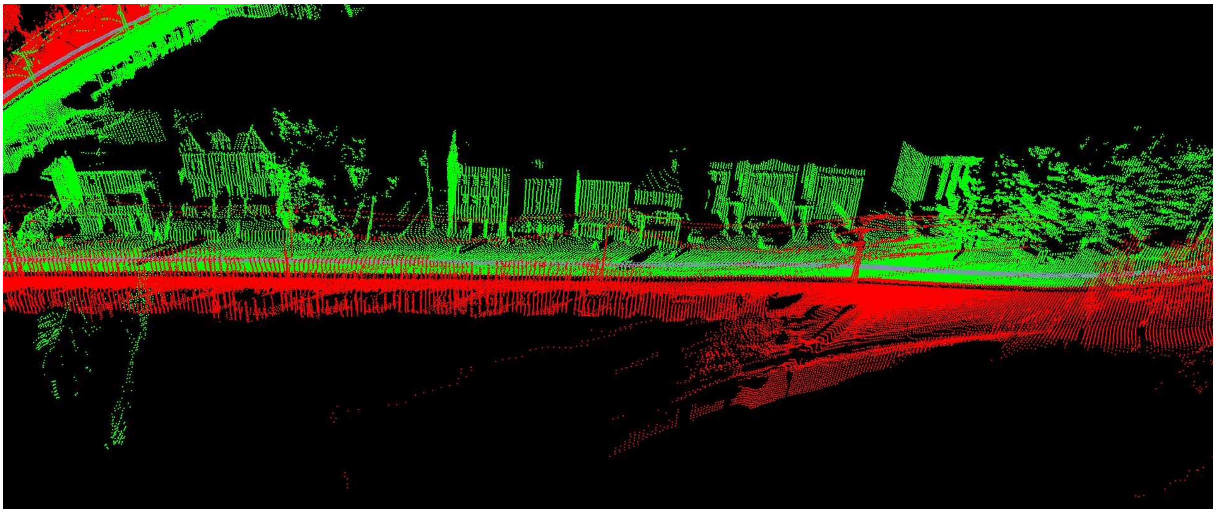
Moreover, we also use OGRE 3D to represent the street-view 3D LiDAR reconstructions shown in Figs. 5.11 and 5.12. Here, a color scale is used to depict the 3D points depending on their distances to the camera pose. The two camera images that are associated with the reconstructed location are also presented in each case for a better understanding of the scene. These examples demonstrate the possibilities of fusing visual information with the data obtained by other sensors for improving the situational awareness in life-long localization. The discussed 3D representations can be very helpful for autonomous navigation in intelligent vehicles and mobile robots, because they provide a more detailed perception of the environment.

In addition, probabilistic models are an interesting option for improving the performance in long-term conditions [Arroyo et al., 2015a]. In this regard, 3D representations based on Octomap [Hornung et al., 2013] can be applied, as shown in the example processed for the CMU-CVG VL dataset in Fig. 5.13. The data structure of OctoMap is focused on octrees, which enable a compact memory representation and multi-resolution queries for 3D occupancy grid mapping. Besides, OctoMap models occupied areas as well as free space, where unknown locations are implicit. While the distinction between free and occupied space is essential for a safe navigation, the information about unknown areas is important for tasks such as the autonomous exploration of an environment.

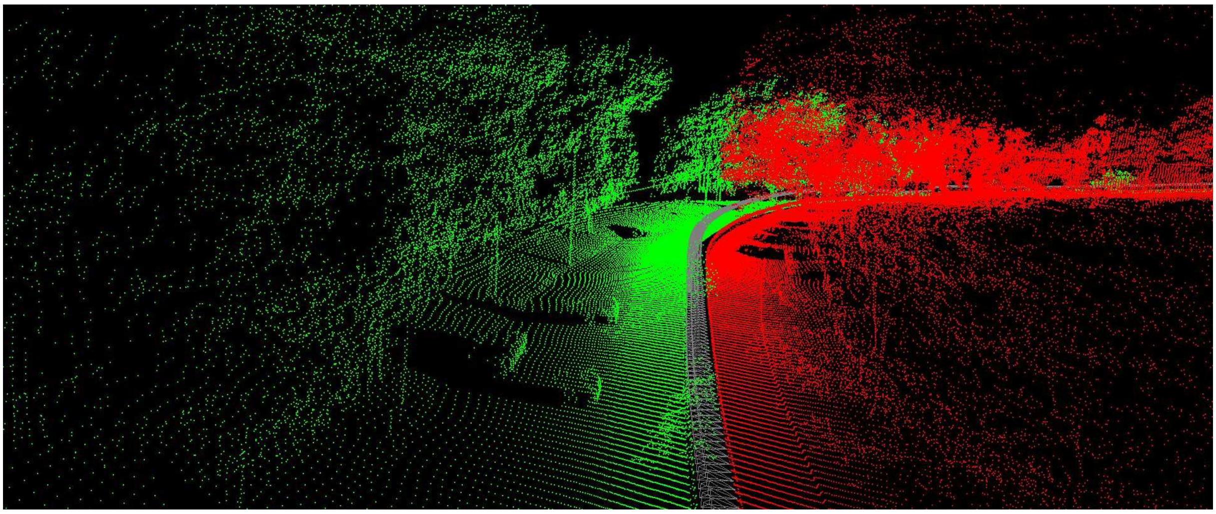
[2]OGRE 3D is currently available from: http://www.ogre3d.org/



(5.9.1) City-scale 3D reconstruction.

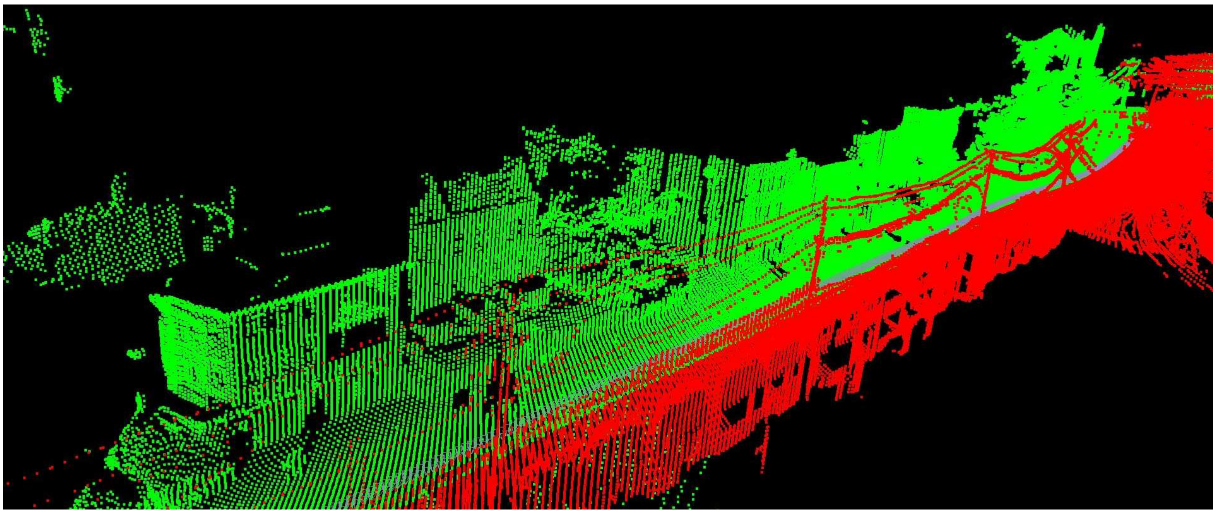


(5.9.2) Detail of the 3D reconstruction (a).

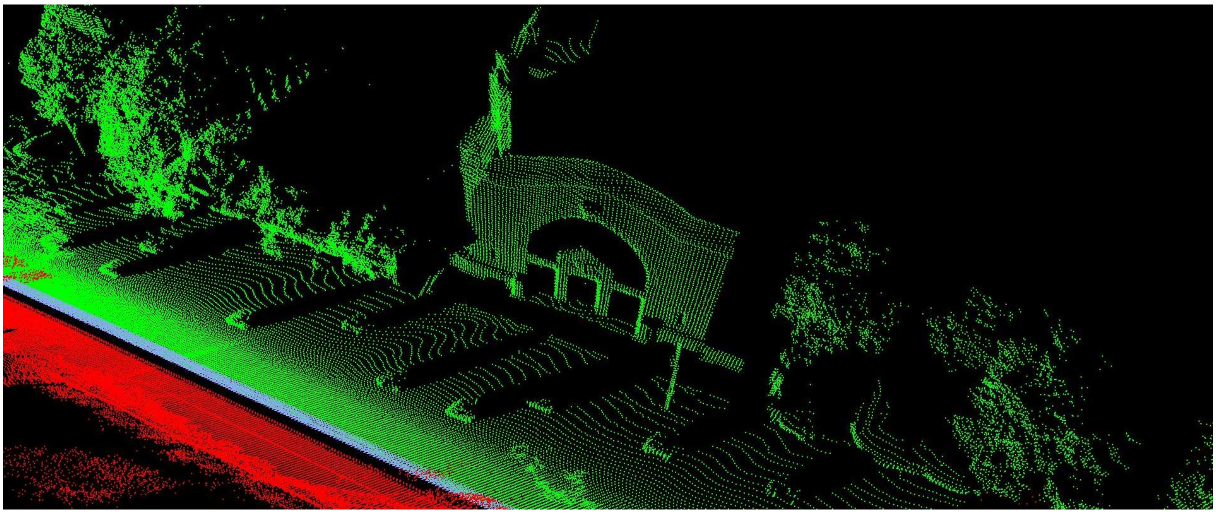


(5.9.3) Detail of the 3D reconstruction (b).

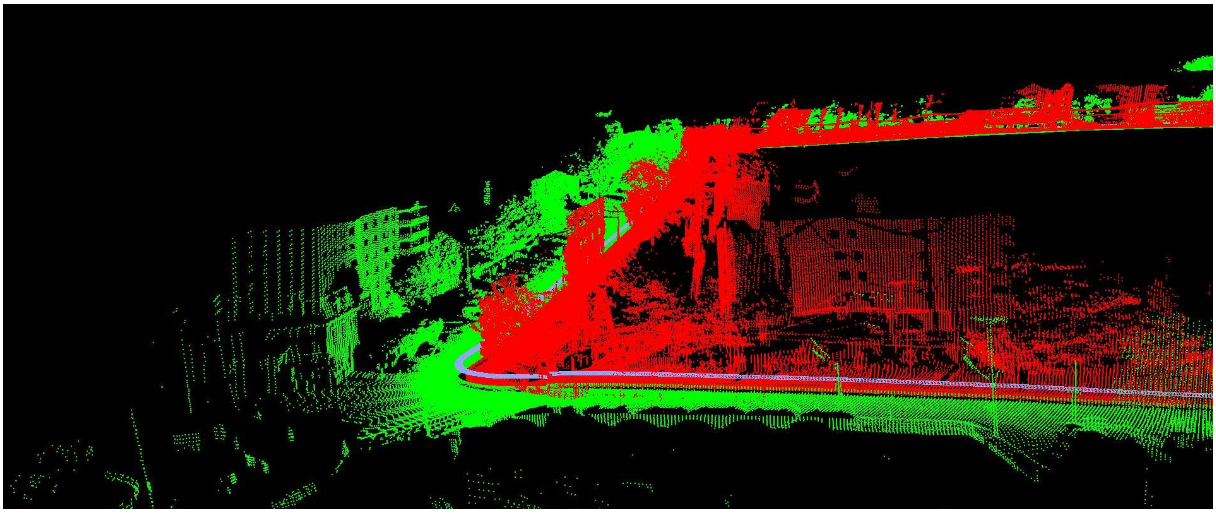
Figure 5.9: 3D reconstructions using corrected poses in the CMU-CVG VL dataset (I). The poses of the car are represented in blue, the 3D points captured by the left LiDAR in green and the 3D points captured by the right LiDAR in red.



(5.10.1) Detail of the 3D reconstruction (c).

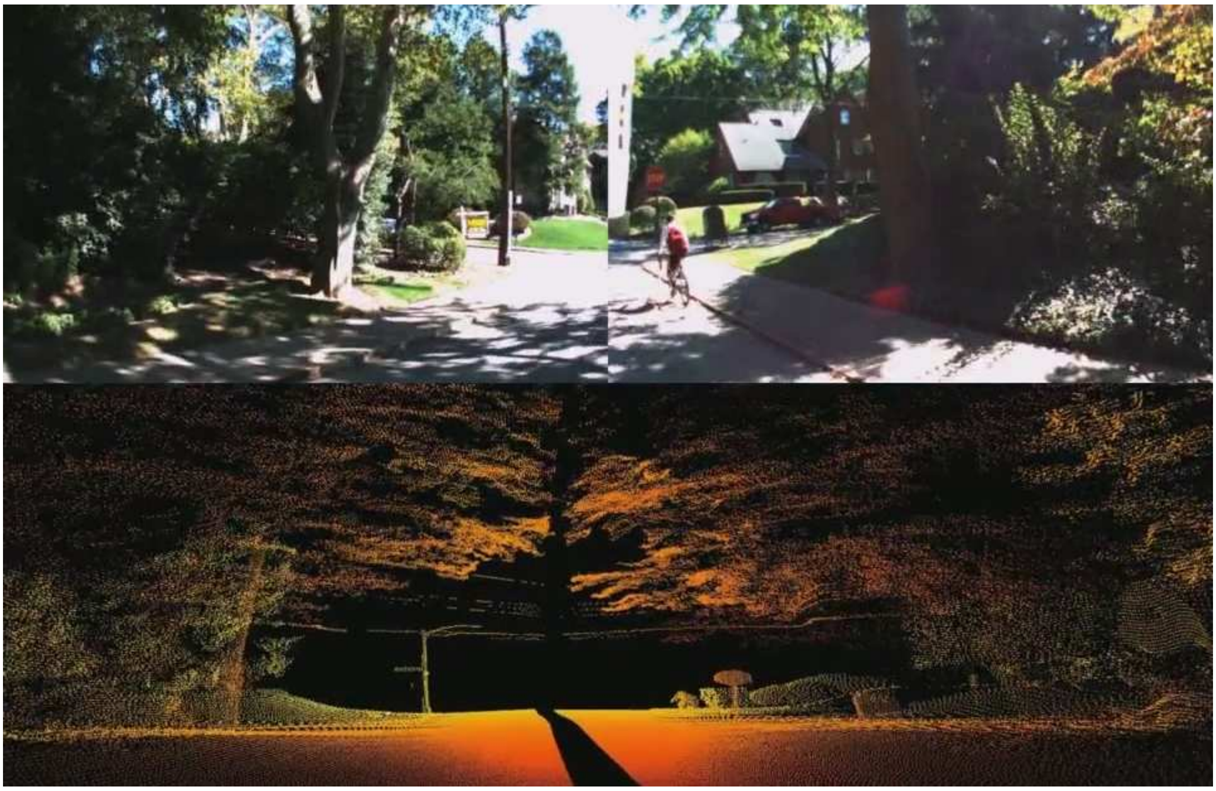


(5.10.2) Detail of the 3D reconstruction (d).

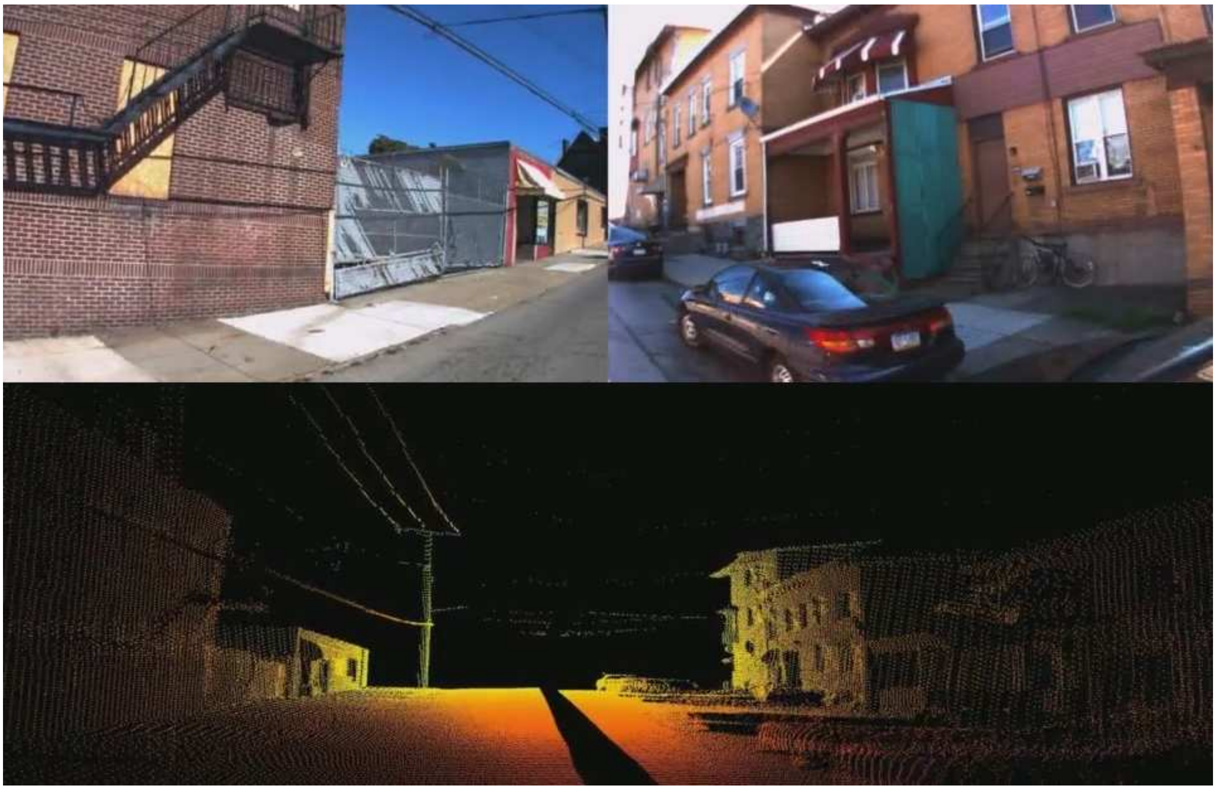


(5.10.3) Detail of the 3D reconstruction (e).

Figure 5.10: 3D reconstructions using corrected poses in the CMU-CVG VL dataset (II). The poses of the car are represented in blue, the 3D points captured by the left LiDAR in green and the 3D points captured by the right LiDAR in red.

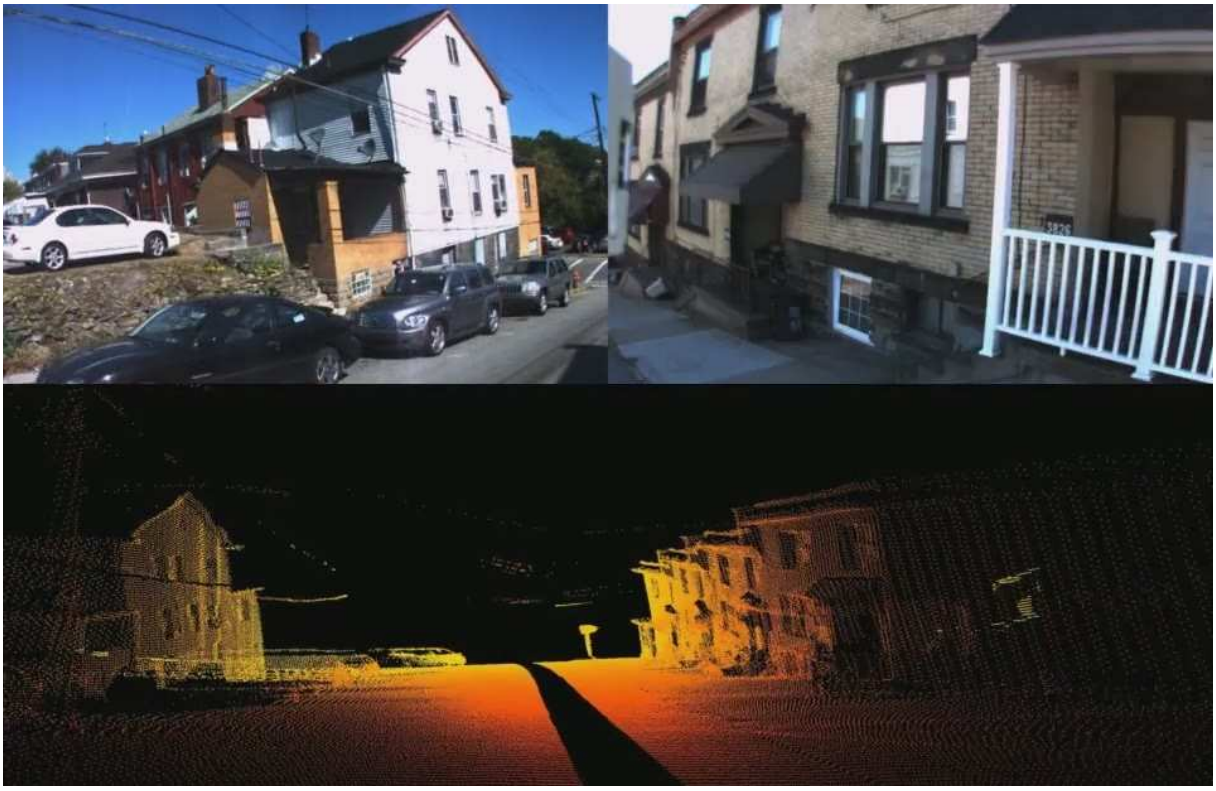


(5.11.1) Example of street-view 3D reconstruction (a).



(5.11.2) Example of street-view 3D reconstruction (b).

Figure 5.11: Street-view 3D reconstructions in the CMU-CVG VL dataset (I). The two camera images that are associated with the reconstructed location are depicted in each example.

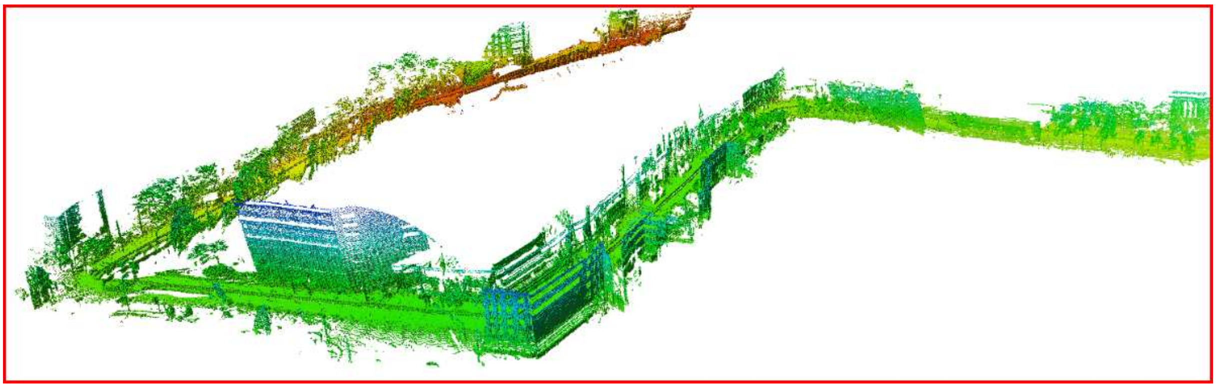


(5.12.1) Example of street-view 3D reconstruction (c).

(5.12.2) Example of street-view 3D reconstruction (d).

Figure 5.12: Street-view 3D reconstructions in the CMU-CVG VL dataset (II). The two camera images that are associated with the reconstructed location are depicted in each example.

(5.13.1) City-scale OctoMap reconstruction.



(5.13.2) Zoomed region in the OctoMap reconstruction.

Figure 5.13: 3D reconstruction using OctoMap in the CMU-CVG VL dataset. The area where a loop closure is identified in the dataset is zoomed.

5.2.2.2 Examples of Application in the Oxford New College Dataset

The last experiments carried out in this section are focused on processing a 3D reconstruction in the Oxford New College dataset. Here, 3D measurements are also acquired by means of two vertically scanning LiDARs. For this reason, we have all the required elements for solving our multi-sensor fusion SLAM problem using the information provided in the dataset.

In this case, we directly show the 3D representation of the environment based on OctoMap, as depicted in Fig. 5.14. This 3D reconstruction of the Oxford New College dataset can be compared with the corrected VO poses previously presented in Fig. 5.4.2. According to this, the importance of using loop closures for reducing the effects of the accumulated drift can be observed. In addition, the fusion of the data computed by varied sensors is essential in order to obtain a more robust and accurate 3D reconstructions.

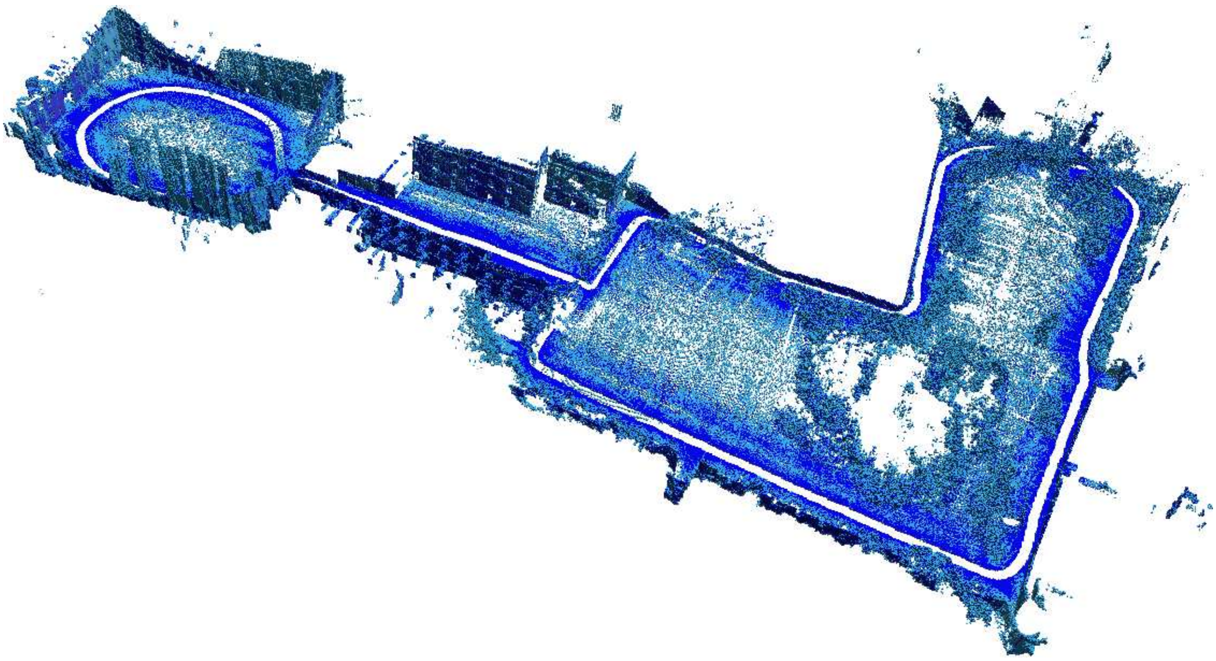


Figure 5.14: 3D reconstruction using OctoMap in the Oxford New College dataset.

5.3 Change Detection in Locations across the Seasons

Autonomous vehicles need frequent and efficient updates in the large-scale maps required by them for life-long navigation. In this sense, the detection of changes in the appearance of the places is an important localization task in long-term situations that is having a growing interest in the community.

In this section, we introduce some examples of change detection related to a collaborative research carried out in the ambit of this thesis and published in [Alcantarilla et al., 2016]. Here, it is proposed a system for performing structural change detection in street-view videos captured by a vehicle-mounted monocular camera over time. The described method chains a multi-sensor fusion SLAM and a dense 3D reconstruction pipeline, which provide coarsely registered image pairs to a deep deconvolutional network for pixel-wise change detection. The multi-sensor approach is similar to the previously explained in Section 5.2.2, but in this case the 3D points are obtained by a camera-based reconstruction instead of using LiDAR measurements, as depicted in the example presented in Fig. 5.15. The combination of dense geometry and accurate registration allows images from different times to be warped into alignment with one another for change detection comparison.

Although topological place recognition is not directly applied in the previously introduced scheme, it can be used in future research in order to obtain the aligned images needed for change detection. This approach would be more efficient than the application of 3D reconstructions in order to obtain this alignment, because topological place recognition normally requires a lower consumption of computational costs to be processed.

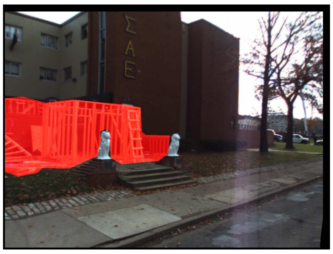
Figure 5.15: Example of camera-based 3D reconstruction in the CMU-CVG VL dataset. These 3D representations are used for change detection in the system presented in [Alcantarilla et al., 2016].

5.3.1 Examples of Application in the CMU-CVG VL Dataset

With the aim of understanding the complexity of change detection across the seasons, Fig. 5.16 shows several examples corresponding to the ground-truth presented in [Alcantarilla et al., 2016] for the CMU-CVG VL dataset.

(5.16.1) First image (a).

(5.16.2) Second image (a).

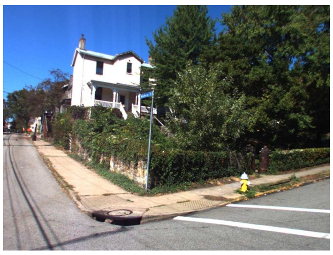


(5.16.3) Change detected (a).

(5.16.4) First image (b).

(5.16.5) Second image (b).

(5.16.6) Change detected (b).



(5.16.7) First image (c).

(5.16.8) Second image (c).

(5.16.9) Change detected (c).

Figure 5.16: Examples of change detection in the CMU-CVG VL dataset. Left and middle columns show registered images from different instants of time, while right column depicts ground-truth structural changes highlighted in red.

5.4 Conclusions and Contributions

Along this chapter, we have described several techniques that represent an important attainment to achieve a life-long localization in different research areas. The application examples presented for the contributed approaches use topological place recognition with the aim of improving the accuracy of the systems in long-term conditions.

In fact, the application of the methods explained in Chapters 3 and 4 is demonstrated in varied cases. Firstly, visual place recognition is employed to detect the loop closures appeared in the trajectory described by an autonomous vehicle or mobile robot. Moreover, the information about loop closures is very useful to correct the drift appeared in localization measurements when long periods of time are taken into account. In this sense, the pose estimations computed in VO or SLAM systems can be improved by means of these techniques. The fusion of the corrected poses with other sensing technologies such as GPS or LiDAR provides a precise perception of the environment, which is applied for carrying out large-scale 3D reconstructions that are extremely helpful for increasing the situational awareness required for a robust autonomous navigation.

In future works, applications derived from the use of topological localization in change detection across seasons could be proposed, with the aim of reducing the computational costs needed for the camera-based 3D reconstructions used in [Alcantarilla et al., 2016] for the previous image alignment. Additionally, the application of semantic information can be also studied for improving the performance of change detection, as exposed in recent works such as [Kataoka et al., 2016].

Chapter 6

Final Conclusions and Future Works

This thesis has contributed to the state of the art within the area of life-long visual localization using topological place recognition. The goal of this chapter is to describe the main conclusions derived from this work and to discuss about the obtained results. Besides, we also enumerate and explain the main contributions provided by the research carried out along this dissertation. Finally, several future works are proposed in order to continue the research line studied in this thesis.

6.1 Main Conclusions

The main conclusions associated with the work presented along this dissertation are the following:

- The studied methods for life-long visual localization are a powerful tool for autonomous navigation. However, driverless cars require a big amount of information about the situational awareness of the environment to operate in long-term conditions. Although camera-based systems provide important descriptions related to the appearance of the scene, it is important to combine this information with the acquired by other sensors, with the aim of achieving the dream of obtaining fully autonomous vehicles available for any user.

- We have performed evaluations of our methods in several datasets focused on long-term localization. Nevertheless, the acquisition and annotation of larger amounts of data is required for carrying out tests as reliable as possible, because autonomous driving needs a high safety. The research community demands publicly available datasets focused on data mining, such as [Romera et al., 2016]. Besides, the registration of synchronized measurements using varied sensors is important in order to improve the scene understanding, as proposed in some datasets very recently appeared [Carlevaris-Bianco et al., 2016, Maddern et al., 2016].

- The specific topic of topological place recognition across the seasons is one of the most important challenges currently studied in the life-long visual localization field. In this sense, our work has presented different approaches that yield an acceptable performance in this difficult situation. However, the research in this area should continue proposing new solutions for improving the performance even more, especially in extreme conditions associated with occlusions derived from snow or vegetation, environments traversed at night, drastic changes on the camera field of view and dynamic objects on scene, which could be identified by means of change detection algorithms.

- During the development of this thesis, the evolution of deep learning techniques for image description has experienced an important growing and their characteristics have been constantly improved. Although in this dissertation some approaches based on CNNs have been presented, the possibilities of these methods are still being broadly studied by the research community in computer vision. In fact, this research line is currently the most promising in order to define accurate place recognition methods.

- Life-long localization for embedded systems on vehicles or robots requires algorithms that provide efficient solutions for long-term operation. According to this, the consumption of computational resources and memory costs must be always considered. In spite of the fact that GPGPU for embedded systems has been growing up in the last years to reach real-time applications with CNNs, so far it is not possible in some of the cases. Then, traditional hand-crafted descriptors should not be discarded at the moment in embedded systems for autonomous driving or in other platforms with limited resources (i.e., smartphones), because they can be more efficient in these cases.

- This thesis is focused on camera-based systems, but we have also shown applications in which other sensors such as GPS or LiDAR are employed. In these cases, the accuracy in localization is clearly enhanced due to the fusion of the different sensors. Furthermore, the usage of 3D reconstructions by means of fused information is interesting in order to have a better perception of the environment and to improve the scene understanding.

- A great part of the material described in this thesis is publicly available online for the benefit of the research community (see Appendix A and B). We think that open access is important in order to help future researchers in topics related to life-long visual localization. Many state-of-the-art algorithms do not provide open implementations and their evaluation is not always possible, so an open-source philosophy facilitates the research progress in this field.

6.2 Main Contributions

From the results obtained in previous chapters, we conclude that the main contributions of this thesis are the following:

- ***A novel visual place recognition method based on hand-crafted features:*** We have supplied our ABLE proposal [Arroyo et al., 2017], as it was described in Chapter 3. The innovation contributed by this approach resides in the global description of sequences of images as binary codes, which are extracted from a LDB or D-LDB descriptor and efficiently matched using the Hamming distance in an ANN search. Besides, an illumination invariant technique is proposed for improving the performance in changing lighting conditions. The application of the introduced binary description and matching method provides a reduction of memory and computational costs, which is necessary for a long-term operation. In addition, three versions of this proposal have been designed with the aim of exploiting the advantages of different types of cameras: monocular (ABLE-M) [Arroyo et al., 2015b], stereo (ABLE-S) [Arroyo et al., 2014a] and panoramic (ABLE-P) [Arroyo et al., 2014b]. The characteristics provided by the versions of ABLE allow a satisfactory and efficient performance in life-long topological localization.

- ***A novel visual place recognition method based on CNN features:*** We have supplied our CNN-VTL proposal [Arroyo et al., 2016a], as it was described in Chapter 4. The novelties contributed by our approach are focused on the fusion of the image information from multiple convolutional layers at several levels and granularities, which is not considered in other state-of-the-art systems that typically trust in individual layers [Sünderhauf et al., 2015b,Sünderhauf et al., 2015a]. In addition, the redundant data of CNN features is compressed into a tractable number of bits for a more efficient and robust life-long visual localization. The final descriptor is reduced by applying simple compression and binarization techniques for fast matching using the Hamming distance. The described techniques provide a high accuracy to the method even in difficult situations derived from seasonal changes.

- ***An exhaustive study of results for topological place recognition:*** We have contributed a wide set of results in long-term conditions in multiple tests carried out over the St Lucia dataset, the Alderley dataset, the Nordland dataset, the CMU-CVG VL dataset, the KITTI Odometry dataset and the Oxford New College dataset. In these experiments presented along the results sections of Chapters 3 and 4, we have demonstrated the satisfactory performance of ABLE and CNN-VTL against the main state-of-the-art algorithms (WI-SURF, BRIEF-Gist, FAB-MAP, SeqSLAM or the CNN-based approach presented in [Sünderhauf et al., 2015a]), showing better results for all the cases. Besides, we have also evidenced the concerns about efficiency for the different proposals. All these evaluations represent a broad study about multiple characteristics of varied problems in the state of the art related to life-long visual localization, because it must be taken into account that more than 3000 km are traversed in the tests between day and night, along the months and across the four seasons of the year.

- ***Applications of topological place recognition in life-long localization:*** We have contributed several applications of our methods over varied visual localization problems, as it was shown in Chapter 5. Topological place recognition has been employed for detecting loop closures and correcting SLAM and VO measurements [Caramazana et al., 2016, Arroyo et al., 2016b]. In addition, we have used this visual information for fusing it with GPS and LiDAR sensors in order to obtain city-scale 3D representations of the environment, which are very useful for scene understanding in probabilistic algorithms for long-term navigation [Arroyo et al., 2015a]. Moreover, geometric change detection is studied with the aim of updating the large-scale maps required by driverless cars [Alcantarilla et al., 2016]. All these applications represent a wide range of contributions in multiple areas associated with life-long visual localization.

- ***Open code provided to the community for life-long visual localization:*** We trust in an open-source philosophy in order to give support to researchers in the specific area of life-long visual localization for autonomous vehicles or mobile robots. In this sense, the publication of open code is a clear advantage for the research progress. For this reason, we have contributed the OpenABLE toolbox [Arroyo et al., 2016b] in order to share with the research community some of the innovative proposals derived from this thesis, as it is described in Appendix A. In addition, we have also contributed a downloadable ground-truth for loop closure detection in the KITTI Odometry dataset [Arroyo et al., 2014a], as it is presented in Appendix B.

6.3 Future Works

Here, we identify some of the future works that can be derived from the experiments, contributions and conclusions of this dissertation, which are mainly the following:

- ***Full scene understanding for driverless cars:*** The study of techniques for enhancing the perception of the environment is crucial for the evolution of intelligent vehicles and mobile robots. In future works, vision-based paradigms for autonomous driving similar to the proposed in [Ros et al., 2015] can be an interesting research area. 3D representations in city-scale urban scenarios can also help in this line [Alcantarilla et al., 2013a]. In addition, the adequate fusion of multiple sensors information is a topic that must be explored in further approaches in order to continue the progression of current driverless cars [Bojarski et al., 2016, Google, 2017].

- ***Semantic data for a better situational awareness in localization:*** In the last years, algorithms for semantic segmentation have become a useful tool for identifying different elements on the scene. Some datasets for the benchmarking of the proposed solutions (i.e., CityScapes [Cordts et al., 2016]) are promoting the evolution of these techniques. In this regard, new approaches could be studied for 3D semantic segmentation, as proposed in recent works such as [Vineet et al., 2015, Sengupta and Sturgess, 2015, Wolf et al., 2015]. Some of these methods combine the visual information with 3D data obtained by other sensors to enhance the semantic segmentation [Zhang et al., 2015]. Besides, the information provided by means of this semantic data can be used to enhance the description stage in future works related to topological place recognition.

- ***More tests for topological localization based on end-to-end CNNs:*** In this thesis, we design a solution for visual place recognition based on pre-trained CNN descriptors, with the aim of demonstrating the transferability of this type of features [Oquab et al., 2014, Yosinski et al., 2014]. However, methods focused on end-to-end learning are also interesting, especially for optimizing the performance in specific scenarios. According to this, proposals similar to the introduced in [Arandjelovic et al., 2016] can be considered in further research.

- ***Place recognition for aiding change detection in image alignment:*** The detection of changes in the appearance of locations is having a growing interest in the research community. This is due to its helpful application in map updating for autonomous vehicles. Current solutions have obtained remarkable results using 3D reconstructions for obtaining the aligned images where the changes are evaluated [Alcantarilla et al., 2016]. Nevertheless, approaches based on topological place recognition can be studied to perform this task in a more efficient way. In addition, it must be noted that images from Google Street View® could be of interest for change detection in order to carry out extensive tests, as deduced from works such as [Arandjelovic et al., 2016].

Bibliography

[Aanaes et al., 2011] H. Aanaes, A. L. Dahl and K. S. Pedersen, 2011, 'Interesting interest points - A comparative study of interest point performance on a unique data set'. *International Journal of Computer Vision (IJCV)*, 97(1):18–35.

[Agrawal et al., 2008] M. Agrawal, K. Konolige and M. R. Blas, 2008, 'CenSurE: Center surround extremas for realtime feature detection and matching'. In *European Conference on Computer Vision (ECCV)*, vol. 5305, 102–115.

[Alahi et al., 2012] A. Alahi, R. Ortiz and P. Vandergheynst, 2012, 'FREAK: Fast retina keypoint'. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 510–517.

[Alcantarilla et al., 2012] P. F. Alcantarilla, A. Bartoli and A. J. Davison, 2012, 'KAZE features'. In *European Conference on Computer Vision (ECCV)*, vol. 7577, 214–227.

[Alcantarilla et al., 2013a] P. F. Alcantarilla, C. Beall and F. Dellaert, 2013a, 'Large-scale dense 3D reconstruction from stereo imagery'. In *Workshop on Planning, Perception and Navigation for Intelligent Vehicles at the IEEE/RSJ International Conference on Intelligent Robots and Systems (W-IROS)*.

[Alcantarilla et al., 2013b] P. F. Alcantarilla, J. Nuevo and A. Bartoli, 2013b, 'Fast explicit diffusion for accelerated features in nonlinear scale spaces'. In *British Machine Vision Conference (BMVC)*.

[Alcantarilla and Stenger, 2016] P. F. Alcantarilla and B. Stenger, 2016, 'How many bits do I need for matching local binary descriptors?' In *IEEE International Conference on Robotics and Automation (ICRA)*, 2190–2197.

[Alcantarilla et al., 2016] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo and R. Gherardi, 2016, 'Street-view change detection with deconvolutional networks'. In *Robotics Science and Systems Conference (RSS)*.

[Almazán et al., 2013] J. Almazán, L. M. Bergasa, J. J. Yebes, R. Barea and R. Arroyo, 2013, 'Full auto-calibration of a smartphone on board a vehicle using IMU and GPS embedded sensors'. In *IEEE Intelligent Vehicles Symposium (IV)*, 1374–1380.

[Andreasson et al., 2007] H. Andreasson, T. Duckett and A. Lilienthal, 2007, 'Mini-SLAM: Minimalistic visual SLAM in large-scale environments based on a new interpretation of image similarity'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 4096–4101.

[Angeli et al., 2008a] A. Angeli, D. Filliat, S. Doncieux and J. Meyer, 2008a, 'Fast and incremental method for loop-closure detection using bags of visual words'. *IEEE Transactions on Robotics (TRO)*, 24(5):1027–1037.

[Angeli et al., 2008b] A. Angeli, D. Filliat, S. Doncieux and J. Meyer, 2008b, 'Incremental vision-based topological SLAM'. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1031–1036.

[Angeli et al., 2009] A. Angeli, D. Filliat, S. Doncieux and J. Meyer, 2009, 'Visual topological SLAM and global localization'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 4300–4305.

[Arandjelovic et al., 2016] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla and J. Sivic, 2016, 'NetVLAD: CNN architecture for weakly supervised place recognition'. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5297–5307.

[Arroyo et al., 2015a] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa and E. Romera, 2015a, 'Life-long visual localization using probabilistic temporal inference'. In *International Computer Vision Summer School (ICVSS)*.

[Arroyo et al., 2015b] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa and E. Romera, 2015b, 'Towards life-long visual localization using an efficient matching of binary sequences from images'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 6328–6335.

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

[Arroyo et al., 2016b] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa and E. Romera, 2016b, 'OpenABLE: An open-source toolbox for application in life-long visual localization of autonomous vehicles'. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 965–970.

[Arroyo et al., 2017] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa and E. Romera, 2017, 'Are you ABLE to perform a life-long visual topological localization?' *submitted to Autonomous Robots*.

[Arroyo et al., 2014a] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, J. J. Yebes and S. Bronte, 2014a, 'Fast and effective visual place recognition using binary codes and disparity information'. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3089–3094.

[Arroyo et al., 2014b] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, J. J. Yebes and S. Gámez, 2014b, 'Bidirectional loop closure detection on panoramas for visual navigation'. In *IEEE Intelligent Vehicles Symposium (IV)*, 1378–1383.

[Arroyo et al., 2015c] R. Arroyo, J. J. Yebes, L. M. Bergasa, I. G. Daza and J. Almazán, 2015c, 'Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls'. *Expert Systems With Applications (ESWA)*, 42(21):7991–8005.

[Badino et al., 2011] H. Badino, D. F. Huber and T. Kanade, 2011, 'Visual topometric localization'. In *IEEE Intelligent Vehicles Symposium (IV)*, 794–799.

[Badino et al., 2012] H. Badino, D. F. Huber and T. Kanade, 2012, 'Real-time topometric localization'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1635–1642.

[Badrinarayanan et al., 2017] V. Badrinarayanan, A. Kendall and R. Cipolla, 2017, 'SegNet: A deep convolutional encoder-decoder architecture for scene segmentation'. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, PP(99):1–14.

[Bailey and Durrant-Whyte, 2006] T. Bailey and H. Durrant-Whyte, 2006, 'Simultaneous localisation and mapping (SLAM): Part II State of the art'. *IEEE Robotics and Automation Magazine (RAM)*, 13(3):108–117.

[Bansal et al., 2014] A. Bansal, H. Badino and D. Huber, 2014, 'Understanding how camera configuration and environmental conditions affect appearance-based localization'. In *IEEE Intelligent Vehicles Symposium (IV)*, 800–807.

[Bay et al., 2008] H. Bay, A. Ess, T. Tuytelaars and L. van Gool, 2008, 'Speeded-up robust features (SURF)'. *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359.

[Bay et al., 2006] H. Bay, T. Tuytelaars and L. van Gool, 2006, 'SURF: Speeded up robust features'. In *European Conference on Computer Vision (ECCV)*, vol. 3951, 404–417.

[Bazeille and Filliat, 2010] S. Bazeille and D. Filliat, 2010, 'Combining odometry and visual loop-closure detection for consistent topometrical mapping'. *RAIRO Operations Research (COGIS'09)*, 44:365–377.

[Bergasa et al., 2014] L. M. Bergasa, D. Almería, J. Almazán, J. J. Yebes and R. Arroyo, 2014, 'Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors'. In *IEEE Intelligent Vehicles Symposium (IV)*, 240–245.

[Bojarski et al., 2016] M. Bojarski, D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao and K. Zieba, 2016, 'End to end learning for self-driving cars'. *Computing Research Repository (CoRR)*, arXiv:1604.07316:1–9.

[Bradski, 2000] G. Bradski, 2000, 'The OpenCV library'. *Dr. Dobb's Journal of Software Tools (DDJ)*, 25(11):122–125.

[Bronte et al., 2014] S. Bronte, M. Paladini, L. M. Bergasa, L. Agapito and R. Arroyo, 2014, 'Real-time sequential model-based non-rigid SFM'. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1026–1031.

[Cadena et al., 2015] C. Cadena, A. Dick and I. D. Reid, 2015, 'A fast, modular scene understanding system using context-aware object detection'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 4859–4866.

[Cadena et al., 2010] C. Cadena, D. Gálvez-López, F. Ramos, J. D. Tardós and J. Neira, 2010, 'Robust place recognition with stereo cameras'. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5182–5189.

[Cadena et al., 2012] C. Cadena, D. Gálvez-López, J. D. Tardós and J. Neira, 2012, 'Robust place recognition with stereo sequences'. *IEEE Transactions on Robotics (TRO)*, 28(4):871–885.

[Cadena and Neira, 2011] C. Cadena and J. Neira, 2011, 'A learning algorithm for place recognition'. In *Workshop on Long-Term Autonomy at the IEEE International Conference on Robotics and Automation (W-ICRA)*.

[Calonder et al., 2012] M. Calonder, V. Lepetit, M. Özuysal, T. Trzcinski, C. Strecha and P. Fua, 2012, 'BRIEF: Computing a local binary descriptor very fast'. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(7):1281–1298.

[Calonder et al., 2010] M. Calonder, V. Lepetit, C. Strecha and P. Fua, 2010, 'BRIEF: Binary robust independent elementary features'. In *European Conference on Computer Vision (ECCV)*, 778–792.

[Campos et al., 2013] F. M. Campos, L. Correia and J. M. F. Calado, 2013, 'Loop closure detection with a holistic image feature'. In *Portuguese Conference on Artificial Intelligence (EPIA)*, vol. 8154, 247–258.

[Caramazana et al., 2016] L. Caramazana, R. Arroyo and L. M. Bergasa, 2016, 'Visual odometry correction based on loop closure detection'. In *Open Conference on Future Trends in Robotics (RoboCity16)*, 97–104.

[Carlevaris-Bianco and Eustice, 2014] N. Carlevaris-Bianco and R. M. Eustice, 2014, 'Learning visual feature descriptors for dynamic lighting conditions'. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2769–2776.

[Carlevaris-Bianco et al., 2016] N. Carlevaris-Bianco, A. K. Ushani and R. M. Eustice, 2016, 'University of Michigan North Campus long-term vision and lidar dataset'. *International Journal of Robotics Research (IJRR)*, 35(9):1023–1035.

[Cela et al., 2013a] A. Cela, L. M. Bergasa and R. Arroyo, 2013a, 'Gate recognition and reconstruction for DARPA Robotics Challenge using Bayesian classifier optimized by Mahalanobis distance"'. In *IEEE European Modelling Symposium (EMS)*, 255–260.

[Cela et al., 2013b] A. Cela, J. J. Yebes, R. Arroyo, L. M. Bergasa, R. Barea and E. López, 2013b, 'Complete low-cost implementation of a teleoperated control system for a humanoid robot'. *Sensors*, 13(2):1385–1401.

[Chandrasekhar et al., 2012] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, Y. Reznik, R. Grzeszczuk and B. Girod, 2012, 'Compressed histogram of gradients: A low-bitrate descriptor'. *International Journal of Computer Vision (IJCV)*, 96(3):384–399.

[Chatfield et al., 2014] K. Chatfield, K. Simonyan, A. Vedaldi and A. Zisserman, 2014, 'Return of the devil in the details: Delving deep into convolutional nets'. In *British Machine Vision Conference (BMVC)*, 1–12.

[Chen et al., 2014] Z. Chen, O. Lam, A. Jacobson and M. Milford, 2014, 'Convolutional neural network-based place recognition'. In *Australasian Conference on Robotics and Automation (ACRA)*, 1–8.

[Churchill and Newman, 2012a] W. Churchill and P. Newman, 2012a, 'Continually improving large scale long term visual navigation of a vehicle in dynamic urban environments'. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 1371–1376.

[Churchill and Newman, 2012b] W. Churchill and P. Newman, 2012b, 'Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 4525–4532.

[Churchill and Newman, 2013] W. Churchill and P. Newman, 2013, 'Experience-based navigation for long-term localisation'. *International Journal of Robotics Research (IJRR)*, 32(14):1645–1661.

[Clemente et al., 2007] L. A. Clemente, A. J. Davison, I. D. Reid, J. Neira and J. D. Tardós, 2007, 'Mapping large loops with a single hand-held camera'. In *Robotics Science and Systems Conference (RSS)*, 297–304.

[Collobert et al., 2011] R. Collobert, K. Kavukcuoglu and C. Farabet, 2011, 'Torch7: A Matlab-like environment for machine learning'. In *Workshops in Advances in Neural Information Processing Systems (W-NIPS)*, 1–6.

[Cordts et al., 2016] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele, 2016, 'The Cityscapes dataset for semantic urban scene understanding'. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223.

[Corke et al., 2013] P. Corke, R. Paul, W. Churchill and P. Newman, 2013, 'Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation'. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2085–2092.

[Cummins and Newman, 2008a] M. Cummins and P. Newman, 2008a, 'Accelerated appearance-only SLAM'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1828–1833.

[Cummins and Newman, 2008b] M. Cummins and P. Newman, 2008b, 'FAB-MAP: Probabilistic localization and mapping in the space of appearance'. *International Journal of Robotics Research (IJRR)*, 27(6):647–665.

[Cummins and Newman, 2010a] M. Cummins and P. Newman, 2010a, 'Accelerating FAB-MAP with concentration inequalities'. *IEEE Transactions on Robotics (TRO)*, 26(6):1042–1050.

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

[Dalal and Triggs, 2005] N. Dalal and B. Triggs, 2005, 'Histograms of oriented gradients for human detection'. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 886–893.

[Davison et al., 2007] A. J. Davison, I. D. Reid, N. D. Molton and O. Stasse, 2007, 'MonoSLAM: Real-time single camera SLAM'. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(6):1052–1067.

[Daza et al., 2014] I. G. Daza, L. M. Bergasa, S. Bronte, J. J. Yebes, J. Almazán and R. Arroyo, 2014, 'Fusion of optimized indicators from advanced driver assistance systems (adas) for driver drowsiness detection'. *Sensors*, 14(1):1106–1131.

[Deng et al., 2009] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, 2009, 'ImageNet: A large-scale hierarchical image database'. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

[Dosovitskiy et al., 2015] A. Dosovitskiy, P. Fischery, E. Ilg, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers and T. Brox, 2015, 'FlowNet: Learning optical flow with convolutional networks'. In *International Conference on Computer Vision (ICCV)*, 2758–2766.

[Drouilly et al., 2015] R. Drouilly, P. Rives and B. Morisset, 2015, 'Semantic representation for navigation in large-scale environments'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1106–1111.

[Durrant-Whyte and Bailey, 2006] H. Durrant-Whyte and T. Bailey, 2006, 'Simultaneous localisation and mapping (SLAM): Part I The essential algorithms'. *IEEE Robotics and Automation Magazine (RAM)*, 13(2):99–110.

[Dymczyk et al., 2015] M. Dymczyk, S. Lynen, T. Cieslewski, M. Bosse, R. Siegwart and P. Furgale, 2015, 'The gist of maps - Summarizing experience for lifelong localization'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2767–2773.

[Eade and Drummond, 2008] E. Eade and T. Drummond, 2008, 'Unified loop closing and recovery for real time monocular SLAM'. In *British Machine Vision Conference (BMVC)*, vol. 6, 1–10.

[Erkent and Bozma, 2015] O. Erkent and H. I. Bozma, 2015, 'Long-term topological place learning'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 5462–5467.

[Fidalgo and Ortiz, 2015] E. Fidalgo and A. Ortiz, 2015, 'Vision-based topological mapping and localization methods: A survey'. *Robotics and Autonomous Systems (RAS)*, 64:1–20.

[Fiolka et al., 2012] T. Fiolka, J. Stückler, D. A. Klein, D. Schulz and S. Behnke, 2012, 'Place recognition using surface entropy features'. In *Workshop on Semantic Perception, Mapping, and Exploration at the IEEE International Conference on Robotics and Automation (W-ICRA)*.

[Fiolka et al., 2013] T. Fiolka, J. Stückler, D. A. Klein, D. Schulz and S. Behnke, 2013, 'Distinctive 3D surface entropy features for place recognition'. In *European Conference on Mobile Robotics (ECMR)*, 204–209.

[Fraundorfer and Scaramuzza, 2012] F. Fraundorfer and D. Scaramuzza, 2012, 'Visual Odometry - Part II: Matching, robustness, and applications'. *IEEE Robotics and Automation Magazine (RAM)*, 19(2):78–90.

[Fuentes-Pacheco et al., 2012] J. Fuentes-Pacheco, J. Ruiz-Ascencio and J. M. Rendón-Mancha, 2012, 'Visual simultaneous localization and mapping: A survey'. *Artificial Intelligence Review (AIR)*, 1–27.

[Gálvez-López and Tardós, 2011] D. Gálvez-López and J. D. Tardós, 2011, 'Real-time loop detection with bags of binary words'. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 51–58.

[Gálvez-López and Tardós, 2012] D. Gálvez-López and J. D. Tardós, 2012, 'Bags of binary words for fast place recognition in image sequences'. *IEEE Transactions on Robotics (TRO)*, 28(5):1188–1197.

[Gao and Zhang, 2017] X. Gao and T. Zhang, 2017, 'Unsupervised learning to detect loops using deep neural networks for visual SLAM system'. *Autonomous Robots*, 41(1):1â–18.

[Geiger et al., 2013] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, 2013, 'Vision meets robotics: The KITTI dataset'. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237.

[Geiger et al., 2012] A. Geiger, P. Lenz and R. Urtasun, 2012, 'Are we ready for autonomous driving? the KITTI vision benchmark suite'. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3354–3361.

[Geiger et al., 2010] A. Geiger, M. Roser and R. Urtasun, 2010, 'Efficient large-scale stereo matching'. In *Asian Conference on Computer Vision (ACCV)*, vol. 6492, 25–38.

[Geiger et al., 2011] A. Geiger, J. Ziegler and C. Stiller, 2011, 'StereoScan: Dense 3D reconstruction in real-time'. In *IEEE IV*, 963–968.

[Glover et al., 2010] A. J. Glover, W. Maddern, M. Milford and G. F. Wyeth, 2010, 'FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 3507–3512.

[Glover et al., 2012] A. J. Glover, W. Maddern, M. Warren, S. Reid, M. Milford and G. F. Wyeth, 2012, 'OpenFABMAP: An open source toolbox for appearance-based loop closure detection'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 4730–4735.

[Google, 2017] Google, 2017, 'Google Self-Driving Car project'. [Web] https://www.google.com/selfdrivingcar/.

[Han et al., 2016] S. Han, H. Mao and B. Dally, 2016, 'Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding'. In *International Conference on Learning Representations (ICLR)*.

[Hartmann et al., 2013] J. Hartmann, J. H. Klüssendorff and E. Maehle, 2013, 'A comparison of feature descriptors for visual SLAM'. In *European Conference on Mobile Robotics (ECMR)*, 56–61.

[Hirschmuller, 2008] H. Hirschmuller, 2008, 'Stereo processing by semiglobal matching and mutual information'. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(2):328–341.

[Hong et al., 2016] S. Hong, J. Kim, J. Pyo and S. Yu, 2016, 'A robust loop-closure method for visual SLAM in unstructured seafloor environments'. *Autonomous Robots*, 40(6):1095–1109.

[Hornung et al., 2013] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss and W. Burgard, 2013, 'OctoMap: an efficient probabilistic 3D mapping framework based on octrees'. *Autonomous Robots*, 34(3):189–206.

[Indelman et al., 2013] V. Indelman, S. Williams, M. Kaess and F. Dellaert, 2013, 'Information fusion in navigation systems via factor graph based incremental smoothing'. *Robotics and Autonomous Systems (RAS)*, 61(8).

[Isola et al., 2016] P. Isola, D. Zoran, D. Krishnan and E. H. Adelson, 2016, 'Learning visual groups from co-occurrences in space and time'. In *International Conference on Learning Representations (ICLR)*.

[Jia et al., 2014] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell, 2014, 'Caffe: Convolutional architecture for fast feature embedding'. In *ACM International Conference on Multimedia (ACMMM)*, 675–678.

[Johns and Yang, 2013a] E. Johns and G. Yang, 2013a, 'Dynamic scene models for incremental, long-term, appearance-based localisation'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2731–2736.

[Johns and Yang, 2013b] E. Johns and G. Yang, 2013b, 'Feature co-occurrence maps: Appearance-based localisation throughout the day'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 3212–3218.

[Johns and Yang, 2014] E. Johns and G. Yang, 2014, 'Generative methods for long-term place recognition in dynamic scenes'. *International Journal of Computer Vision (IJCV)*, 106(3):297–314.

[Kataoka et al., 2016] H. Kataoka, S. Shirakabe, Y. Miyashita, A. Nakamura, K. Iwata and Y. Satoh, 2016, 'Semantic change detection with hypermaps'. *Computing Research Repository (CoRR)*, arXiv:1604.07513:1–6.

[Kawewong et al., 2011] A. Kawewong, N. Tongprasit, S. Tangruamsub and O. Hasegawa, 2011, ‘Online and incremental appearance-based SLAM in highly dynamic environments’. *International Journal of Robotics Research (IJRR)*, 30(1):33–55.

[Ke and Sukthankar, 2004] Y. Ke and R. Sukthankar, 2004, ‘PCA-SIFT: A more distinctive representation for local image descriptors’. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 506–513.

[Kendall et al., 2015] A. Kendall, M. Grimes and R. Cipolla, 2015, ‘PoseNet: A convolutional network for real-time 6-DOF camera relocalization’. In *International Conference on Computer Vision (ICCV)*, 2938–2946.

[Khan et al., 2012] S. Khan, D. Wollherr and M. Buss, 2012, ‘PIRF 3D: Online spatial and appearance based loop closure’. In *International Conference on Control, Automation, Robotics & Vision (ICARCV)*, 335–340.

[Kitt et al., 2010] B. Kitt, A. Geiger and H. Lategahn, 2010, ‘Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme’. In *IEEE IV*, 486–492.

[Kläser et al., 2008] A. Kläser, M. Marszalek and C. Schmid, 2008, ‘A spatio-temporal descriptor based on 3D-gradients’. In *British Machine Vision Conference (BMVC)*, 995–1004.

[Klein and Murray, 2007] G. Klein and D. Murray, 2007, ‘Parallel Tracking And Mapping for small AR workspaces’. In *International Symposium on Mixed and Augmented Reality (ISMAR)*, 1–10.

[Knopp et al., 2010] J. Knopp, M. P. G. Willems, R. Timofte and L. van Gool, 2010, ‘Hough transform and 3D SURF for robust three dimensional classification’. In *European Conference on Computer Vision (ECCV)*, vol. 6316, 589–602.

[Konolige and Agrawal, 2008] K. Konolige and M. Agrawal, 2008, ‘FrameSLAM: From bundle adjustment to real-time visual mapping’. *IEEE Transactions on Robotics (TRO)*, 24(5):1066–1077.

[Konolige and Bowman, 2009] K. Konolige and J. Bowman, 2009, ‘Towards lifelong visual maps’. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1156–1163.

[Korrapati and Mezouar, 2016] H. Korrapati and Y. Mezouar, 2016, ‘Multi-resolution map building and loop closure with omnidirectional images’. *Autonomous Robots*, (Published online).

[Korrapati et al., 2013] H. Korrapati, F. Uzer and Y. Mezouar, 2013, ‘Hierarchical visual mapping with omnidirectional images’. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3684–3690.

[Krahenbuhl et al., 2016] P. Krahenbuhl, C. Doersch, J. Donahue and T. Darrell, 2016, 'Data-dependent initializations of convolutional neural networks'. In *International Conference on Learning Representations (ICLR).*

[Krajník et al., 2013] T. Krajník, P. Cristóforis, J. Faigl, H. Szucsová, L. Preucil, M. Nitsche and M. Mejail, 2013, 'Image features for long-term mobile robot autonomy'. In *Workshop on Long-Term Autonomy at the IEEE International Conference on Robotics and Automation (W-ICRA).*

[Krizhevsky et al., 2012] A. Krizhevsky, I. Sutskever and G. E. Hinton, 2012, 'ImageNet classification with deep convolutional neural networks'. In *Advances in Neural Information Processing Systems (NIPS)*, 1106–1114.

[Labbé and Michaud, 2011] M. Labbé and F. Michaud, 2011, 'Memory management for real-time appearance-based loop closure detection'. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1271–1276.

[Labbé and Michaud, 2013] M. Labbé and F. Michaud, 2013, 'Appearance-based loop closure detection for online large-scale and long-term operation'. *IEEE Transactions on Robotics (TRO)*, 29(3):734–745.

[Lategahn and Stiller, 2014] H. Lategahn and C. Stiller, 2014, 'Vision-only localization'. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 15(3):1246–1257.

[Lazebnik et al., 2006] S. Lazebnik, C. Schmid and J. Ponce, 2006, 'Beyond bags of features: spatial pyramid matching for recognizing natural scene categories'. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2169–2178.

[Lee and Pollefeys, 2014] G. H. Lee and M. Pollefeys, 2014, 'Unsupervised learning of threshold for geometric verification in visual-based loop-closure'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1510–1516.

[Leutenegger et al., 2011] S. Leutenegger, M. Chli and R. Y. Siegwart, 2011, 'BRISK: Binary robust invariant scalable keypoints'. In *International Conference on Computer Vision (ICCV)*, 2548–2555.

[Linegar et al., 2015] C. Linegar, W. Churchill and P. Newman, 2015, 'Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 90–97.

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

[Liu and Zhang, 2013] Y. Liu and H. Zhang, 2013, 'Performance evaluation of whole-image descriptors in visual loop closure detection'. In *IEEE International Conference on Information and Automation (ICIA)*, 716–722.

[Llorca et al., 2013] D. F. Llorca, R. Arroyo and M. A. Sotelo, 2013, ‘Vehicle logo recognition in traffic images using HOG features and SVM’. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2229–2234.

[Long et al., 2015] J. Long, E. Shelhamer and T. Darrell, 2015, ‘Fully convolutional networks for semantic segmentation’. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

[Lowe, 1999] D. G. Lowe, 1999, ‘Object recognition from local scale-invariant features’. In *International Conference on Computer Vision (ICCV)*, 1150–1157.

[Lowe, 2004] D. G. Lowe, 2004, ‘Distinctive image features from scale-invariant keypoints’. *International Journal of Computer Vision (IJCV)*, 60(2):91–110.

[Lowry and Milford, 2015] S. Lowry and M. Milford, 2015, ‘Change removal: Robust online learning for changing appearance and changing viewpoint’. In *Workshop on Visual Place Recognition in Changing Environments at the IEEE International Conference on Robotics and Automation (W-ICRA).*

[Lowry et al., 2016] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke and M. Milford, 2016, ‘Visual place recognition: A survey’. *IEEE Transactions on Robotics (TRO)*, 32(1):1–19.

[Lv et al., 2007] Q. Lv, W. Josephson, Z. Wang, M. Charikar and K. Li, 2007, ‘Multi-probe LSH: Efficient indexing for high-dimensional similarity search’. In *International Conference on Very Large Data Bases (VLDB)*, 950–961.

[Maddern et al., 2011] W. Maddern, M. Milford and G. F. Wyeth, 2011, ‘Continuous appearance-based trajectory SLAM’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 3595–3600.

[Maddern et al., 2012] W. Maddern, M. Milford and G. F. Wyeth, 2012, ‘CAT-SLAM: Probabilistic localisation and mapping using a continuous appearance-based trajectory’. *International Journal of Robotics Research (IJRR)*, 31(4):429–451.

[Maddern et al., 2016] W. Maddern, G. Pascoe, C. Linegar and P. Newman, 2016, ‘1 year, 1000 km: The oxford RobotCar dataset’. *International Journal of Robotics Research (IJRR)*, Published online.

[Maddern et al., 2014] W. Maddern, A. Stewart and P. Newman, 2014, ‘LAPS-II: 6-DoF day and night visual localisation with prior 3D structure for autonomous road vehicles’. In *IEEE Intelligent Vehicles Symposium (IV)*, 330–337.

[Masatoshi et al., 2015] A. Masatoshi, C. Yuuto, T. Kanji and Y. Kentaro, 2015, ‘Leveraging image-based prior in cross-season place recognition’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 5455–5461.

[Matas et al., 2002] J. Matas, O. Chum, M. Urban and T. Pajdla, 2002, ‘Robust wide baseline stereo from maximally stable extremal regions’. In *British Machine Vision Conference (BMVC)*, 384–393.

[McManus et al., 2014] C. McManus, W. Churchill, W. Maddern, A. Stewart and P. Newman, 2014, ‘Shady dealings: Robust, long-term visual localisation using illumination invariance’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 901–906.

[Mei et al., 2011] C. Mei, G. Sibley, M. Cummins, P. Newman and I. D. Reid, 2011, ‘RSLAM: A system for large-scale mapping in constant-time using stereo’. *International Journal of Computer Vision (IJCV)*, 94(2):198–214.

[Mei et al., 2010] C. Mei, G. Sibley and P. Newman, 2010, ‘Closing loops without places’. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3738–3744.

[Mikolajczyk and Schmid, 2004] K. Mikolajczyk and C. Schmid, 2004, ‘Scale & affine invariant interest point detectors’. *International Journal of Computer Vision (IJCV)*, 60(1):63–86.

[Miksik et al., 2015] O. Miksik, Y. Amar, V. Vineet, P. Pérez and P. H. S. Torr, 2015, ‘Incremental dense multi-modal 3D scene reconstruction’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 908–915.

[Milford, 2012] M. Milford, 2012, ‘Visual route recognition with a handful of bits’. In *Robotics Science and Systems Conference (RSS)*, 297–304.

[Milford, 2013] M. Milford, 2013, ‘Vision-based place recognition: how low can you go?’ *International Journal of Robotics Research (IJRR)*, 32(7):766–789.

[Milford et al., 2014] M. Milford, W. J. Scheirer, E. Vig, A. J. Glover, O. Baumann, J. Mattingley and D. D. Cox, 2014, ‘Condition-invariant, top-down visual place recognition’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 5571 – 5577.

[Milford and Wyeth, 2008] M. Milford and G. F. Wyeth, 2008, ‘Single camera vision-only SLAM on a suburban road network’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 3684–3689.

[Milford and Wyeth, 2012] M. Milford and G. F. Wyeth, 2012, ‘SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1643–1649.

[Mohan et al., 2015] M. Mohan, D. Gálvez-López, C. Monteleoni and G. Sibley, 2015, ‘Environment selection and hierarchical place recognition’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 5487–5494.

[Molinos et al., 2013] E. J. Molinos, A. Llamazares, N. Hernández, R. Arroyo, A. Cela, J. J. Yebes, M. Ocaña and L. M. Bergasa, 2013, 'Perception and navigation in unknown environments: The DARPA Robotics Challenge'. In *First Iberian Robotics Conference (ROBOT)*, 321–329.

[Mousavian et al., 2015] A. Mousavian, J. Kosecká and J. Lien, 2015, 'Semantically guided location recognition for outdoors scenes'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 4882–4889.

[Muja and Lowe, 2012] M. Muja and D. G. Lowe, 2012, 'Fast matching of binary features'. In *Canadian Conference on Computer and Robot Vision (CRV)*, 404–410.

[Muja and Lowe, 2014] M. Muja and D. G. Lowe, 2014, 'Scalable nearest neighbor algorithms for high dimensional data'. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(11).

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

[Murillo and Kosecká, 2009] A. C. Murillo and J. Kosecká, 2009, 'Experiments in place recognition using gist panoramas'. In *Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras on International Conference on Computer Vision (W-ICCV)*, 2196–2203.

[Murillo et al., 2013] A. C. Murillo, G. Singh, J. Kosecká and J. J. Guerrero, 2013, 'Localization in urban environments using a panoramic gist descriptor'. *IEEE Transactions on Robotics (TRO)*, 29(1):146–160.

[Negre-Carrasco et al., 2016] P. L. Negre-Carrasco, F. Bonin-Font and G. Oliver-Codina, 2016, 'Global image signature for visual loop-closure detection'. *Autonomous Robots*, 40(8):1403–â1417.

[Nelson et al., 2015] P. Nelson, W. Churchill, I. Posner and P. Newman, 2015, 'From dusk till dawn: Localisation at night using artificial light sources'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 5245–5252.

[Neubert et al., 2013] P. Neubert, N. Sünderhauf and P. Protzel, 2013, 'Appearance change prediction for long-term navigation across seasons'. In *European Conference on Mobile Robotics (ECMR)*, 198–203.

[Neubert et al., 2015] P. Neubert, N. Sünderhauf and P. Protzel, 2015, 'Superpixel-based appearance change prediction for long-term navigation across seasons'. *Robotics and Autonomous Systems (RAS)*, 69(7):15–27.

[Nguyen et al., 2013] V. A. Nguyen, J. A. Starzyk and W. Goh, 2013, 'A spatio-temporal long-term memory approach for visual place recognition in mobile robotic navigation'. *Robotics and Autonomous Systems (RAS)*, 61(12):1744–1758.

[Ni et al., 2009] K. Ni, A. Kannan, A. Criminisi and J. Winn, 2009, 'Epitomic location recognition'. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(12):2158–2167.

[Nicosevici and Garcia, 2012] T. Nicosevici and R. Garcia, 2012, 'Automatic visual bag-of-words for online robot navigation and mapping'. *IEEE Transactions on Robotics (TRO)*, 28(4):886–898.

[Nistér et al., 2006] D. Nistér, O. Naroditsky and J. R. Bergen, 2006, 'Visual odometry for ground vehicle applications'. *Journal of Field Robotics (JFR)*, 23(1):3–20.

[Novak and Shafer, 1992] C. L. Novak and S. A. Shafer, 1992, 'Anatomy of a color histogram'. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 599–605.

[Nowicki et al., 2016] M. Nowicki, J. Wietrzykowski and P. Skrzypczynski, 2016, 'Experimental evaluation of visual place recognition algorithms for personal indoor localization'. In *IEEE International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 1–8.

[Ojala et al., 1996] T. Ojala, M. Pietikäinen and D. Harwood, 1996, 'A comparative study of texture measures with classification based on featured distributions'. *Pattern Recognition (PR)*, 29(1):51–59.

[Oliva and Torralba, 2001] A. Oliva and A. Torralba, 2001, 'Modeling the shape of the scene: A holistic representation of the spatial envelope'. *International Journal of Computer Vision (IJCV)*, 42(3):145–175.

[Oliva and Torralba, 2006] A. Oliva and A. Torralba, 2006, 'Building the gist of a scene: The role of global image features in recognition'. *Visual Perception, Progress in Brain Research (PBR)*, 155(B):23–36.

[Oquab et al., 2014] M. Oquab, L. Bottou, I. Laptev and J. Sivic, 2014, 'Learning and transferring mid-level image representations using convolutional neural networks'. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1717–1724.

[Pandey et al., 2014] G. Pandey, J. R. McBride, S. Savarese and R. M. Eustice, 2014, 'Toward mutual information based place recognition'. In *IEEE International Conference on Robotics and Automation (ICRA)*.

[Paul and Newman, 2010] R. Paul and P. Newman, 2010, 'FAB-MAP 3D: Topological mapping with spatial and visual appearance'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2649–2656.

[Paz et al., 2008] L. M. Paz, P. Pinies, J. D. Tardós and J. Neira, 2008, 'Large-scale 6-DOF SLAM with stereo-in-hand'. *IEEE Transactions on Robotics (TRO)*, 24(5):946–957.

[Pepperell et al., 2014] E. Pepperell, P. Corke and M. Milford, 2014, 'All-environment visual place recognition with SMART'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1612–1618.

[Pepperell et al., 2015] E. Pepperell, P. Corke and M. Milford, 2015, 'Automatic image scaling for place recognition in changing environments'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1118–1124.

[Pérez-Grau et al., 2016] F. Pérez-Grau, F. Fabresse, F. Caballero, A. Viguria and A. Ollero, 2016, 'Long-term aerial robot localization based on visual odometry and radio-based ranging'. In *IEEE International Conference on Unmanned Aircraft Systems (ICUAS)*, 608–614.

[Razavian et al., 2014] A. S. Razavian, H. Azizpour, J. Sullivan and S. Carlsson, 2014, 'CNN features off-the-shelf: An astounding baseline for recognition'. In *Workshops at the IEEE Conference on Computer Vision and Pattern Recognition (W-CVPR)*.

[Romera et al., 2015] E. Romera, L. M. Bergasa and R. Arroyo, 2015, 'A real-time multi-scale vehicle detection and tracking approach for smartphones'. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 1298–1303.

[Romera et al., 2016] E. Romera, L. M. Bergasa and R. Arroyo, 2016, 'Need data for driver behaviour analysis? Presenting the public UAH-Driveset'. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 387–392.

[Ros et al., 2015] G. Ros, S. Ramos, M. Granados, D. Vázquez and A. López, 2015, 'Vision-based offline-online perception paradigm for autonomous driving'. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 4321–4328.

[Ros et al., 2012] G. Ros, A. Sappa, D. Ponsa and A. López, 2012, 'Visual SLAM for driverless cars: A brief survey'. In *Workshop on Navigation, Perception, Accurate Positioning and Mapping at the IEEE Intelligent Vehicles Symposium (W-IV)*.

[Ros et al., 2016] G. Ros, S. Stent, P. F. Alcantarilla and T. Watanabe, 2016, 'Training constrained deconvolutional networks for road scene semantic segmentation'. *Computing Research Repository (CoRR)*, arXiv:1604.01545:1–6.

[Rosten and Drummond, 2006] E. Rosten and T. Drummond, 2006, ‘Machine learning for high-speed corner detection’. In *European Conference on Computer Vision (ECCV)*, vol. 3951, 430–443.

[Rublee et al., 2011] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, 2011, ‘ORB: An efficient alternative to SIFT or SURF’. In *International Conference on Computer Vision (ICCV)*, 2564–2571.

[Rusu et al., 2009] R. B. Rusu, N. Blodow and M. Beetz, 2009, ‘Fast point feature histograms (FPFH) for 3D registration’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1848–1853.

[Rusu et al., 2008a] R. B. Rusu, N. Blodow, Z. C. Marton and M. Beetz, 2008a, ‘Aligning point cloud views using persistent feature histograms’. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3384–3391.

[Rusu et al., 2010] R. B. Rusu, G. Bradski, R. Thibaux and J. Hsu, 2010, ‘Fast 3D recognition and pose using the viewpoint feature histogram’. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2155–2162.

[Rusu and Cousins, 2011] R. B. Rusu and S. Cousins, 2011, ‘3D is here: Point Cloud Library (PCL)’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1–4.

[Rusu et al., 2008b] R. B. Rusu, Z. C. Marton, N. Blodow and M. Beetz, 2008b, ‘Persistent point feature histograms for 3D point clouds’. In *International Conference on Intelligent Autonomous Systems (IAS)*, vol. 154, 119–128.

[Scaramuzza and Fraundorfer, 2011] D. Scaramuzza and F. Fraundorfer, 2011, ‘Visual Odometry - Part I: The first 30 years and fundamentals’. *IEEE Robotics and Automation Magazine (RAM)*, 18(4):80–92.

[Scherer et al., 2013] S. A. Scherer, A. Kloss and A. Zell, 2013, ‘Loop closure detection using depth images’. In *European Conference on Mobile Robotics (ECMR)*, 100–106.

[Scovanner et al., 2007] P. Scovanner, S. Ali and M. Shah, 2007, ‘A 3-dimensional SIFT descriptor and its application to action recognition’. In *International Workshop on Multimedia Signal Processing (MMSP)*, 357–360.

[Sengupta and Sturgess, 2015] S. Sengupta and P. Sturgess, 2015, ‘Semantic Octree: Unifying recognition, reconstruction and representation via an Octree constrained higher order MRF’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1874–1879.

[Simo-Serra et al., 2015] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua and F. Moreno-Noguer, 2015, ‘Discriminative learning of deep convolutional feature point descriptors'. In *International Conference on Computer Vision (ICCV)*, 118–126.

[Singh and Kosecká, 2010] G. Singh and J. Kosecká, 2010, 'Visual loop closing using gist descriptors in Manhattan world'. In *Workshop on Omnidirectional Robot Vision at the IEEE International Conference on Robotics and Automation (W-ICRA)*.

[Sivic and Zisserman, 2003] J. Sivic and A. Zisserman, 2003, 'Video Google: A text retrieval approach to object matching in videos'. In *International Conference on Computer Vision (ICCV)*, 1470–1477.

[Skog and Händel, 2009] I. Skog and P. Händel, 2009, 'In-car positioning and navigation technologies - A survey'. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 10(1):4–21.

[Smith et al., 2009] M. Smith, I. Baldwin, W. Churchill, R. Paul and P. Newman, 2009, 'The New College vision and laser data set'. *International Journal of Robotics Research (IJRR)*, 28(5):595–599.

[Steder et al., 2010] B. Steder, R. B. Rusu, K. Konolige and W. Burgard, 2010, 'NARF: 3D range image features for object recognition'. In *Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ International Conference on Intelligent Robots and Systems (W-IROS)*.

[Sünderhauf et al., 2013a] N. Sünderhauf, P. Neubert and P. Protzel, 2013a, 'Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons'. In *Workshop on Long-Term Autonomy at the IEEE International Conference on Robotics and Automation (W-ICRA)*.

[Sünderhauf et al., 2013b] N. Sünderhauf, P. Neubert and P. Protzel, 2013b, 'Predicting the change - A step towards life-long operation in everyday environments'. In *Workshop on Robotics Challenges and Vision at the Robotics Science and Systems Conference (W-RSS)*.

[Sünderhauf and Protzel, 2011] N. Sünderhauf and P. Protzel, 2011, 'BRIEF-Gist - Closing the loop by simple means'. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1234–1241.

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

[Sünderhauf et al., 2015b] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft and M. Milford, 2015b, 'Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free'. In *Robotics Science and Systems Conference (RSS)*.

[Tolias et al., 2016] G. Tolias, R. Sicre and H. Jegou, 2016, ‘Particular object retrieval with integral max-pooling of CNN activations’. In *International Conference on Learning Representations (ICLR)*.

[Ulrich and Nourbakhsh, 2000] I. Ulrich and I. R. Nourbakhsh, 2000, ‘Appearance-based place recognition for topological localization’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1023–1029.

[Upcroft et al., 2014] B. Upcroft, C. McManus, W. Churchill, W. Maddern and P. Newman, 2014, ‘Lighting invariant urban street classification’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1712–1718.

[Urban et al., 2016] G. Urban, K. J. Geras, S. E. Kahou, O. Aslan, S. Wang, R. Caruana, A. Mohamed, M. Philipose and M. Richardson, 2016, ‘Do deep convolutional nets really need to be deep (or even convolutional)?’ In *International Conference on Learning Representations (ICLR)*.

[Valgren and Lilienthal, 2008] C. Valgren and A. J. Lilienthal, 2008, ‘Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1856–1861.

[Valgren and Lilienthal, 2010] C. Valgren and A. J. Lilienthal, 2010, ‘SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments’. *Robotics and Autonomous Systems (RAS)*, 58(2):149–156.

[Vedaldi and Lenc, 2015] A. Vedaldi and K. Lenc, 2015, ‘MatConvNet: Convolutional neural networks for MATLAB’. In *ACM International Conference on Multimedia (ACMMM)*, 689–692.

[Vineet et al., 2015] V. Vineet, O. Miksik, M. Lidegaard, M. Niebner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez and P. H. S. Torr, 2015, ‘Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 75–82.

[Viola and Jones, 2004] P. Viola and M. J. Jones, 2004, ‘Robust real-time face detection’. *International Journal of Computer Vision (IJCV)*, 57(3):137–154.

[Vysotska et al., 2015] O. Vysotska, T. Naseer, L. Spinello, W. Burgard and C. Stachniss, 2015, ‘Efficient and effective matching of image sequences under substantial appearance changes exploiting GPS priors’. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2774–2779.

[Williams et al., 2008] B. Williams, M. Cummins, J. Neira, P. Newman, I. D. Reid and J. D. Tardós, 2008, ‘An image-to-map loop closing method for monocular SLAM’. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2053–2059.

[Williams et al., 2009] B. Williams, M. Cummins, J. Neira, P. Newman, I. D. Reid and J. D. Tardós, 2009, 'A comparison of loop closing techniques in monocular SLAM'. *Robotics and Autonomous Systems (RAS)*, 57(12):1188–1197.

[Williams et al., 2011] B. Williams, G. Klein and I. D. Reid, 2011, 'Automatic relocalization and loop closing for real-time monocular SLAM'. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(9):1699–1712.

[Wolcott and Eustice, 2014] R. W. Wolcott and R. M. Eustice, 2014, 'Visual localization within LIDAR maps for automated urban driving'. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 176–183.

[Wold et al., 1987] S. Wold, K. Esbensen and P. Geladi, 1987, 'Principal Component Analysis'. *Chemometrics and Intelligent Laboratory Systems (CILS)*, 2(1):37–52.

[Wolf et al., 2015] D. Wolf, J. Prankl and M. Vincze, 2015, 'Fast semantic segmentation of 3D point clouds using a dense CRF with learned parameters'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 4867–4873.

[Yang and Cheng, 2012] X. Yang and K. T. Cheng, 2012, 'LDB: An ultra-fast feature for scalable augmented reality on mobile devices'. In *International Symposium on Mixed and Augmented Reality (ISMAR)*, 49–57.

[Yang and Cheng, 2014] X. Yang and K. T. Cheng, 2014, 'Local difference binary for ultrafast and distinctive feature description'. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(1):188–194.

[Yebes et al., 2014] J. J. Yebes, L. M. Bergasa, R. Arroyo and A. Lázaro, 2014, 'Supervised learning and evaluation of KITTI's cars detector with DPM'. In *IEEE Intelligent Vehicles Symposium (IV)*, 768–773.

[Yosinski et al., 2014] J. Yosinski, J. Clune, Y. Bengio and H. Lipson, 2014, 'How transferable are features in deep neural networks?' In *Advances in Neural Information Processing Systems (NIPS)*, 3320–3328.

[Zeiler and Fergus, 2014] M. D. Zeiler and R. Fergus, 2014, 'Visualizing and understanding convolutional networks'. In *European Conference on Computer Vision (ECCV)*, vol. 8689, 818–833.

[Zhang, 2011] H. Zhang, 2011, 'BoRF: Loop-closure detection with scale invariant visual features'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 3125–3130.

[Zhang et al., 2010] H. Zhang, B. Li and D. Yang, 2010, 'Keyframe detection for appearance-based visual SLAM'. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2071–2076.

[Zhang et al., 2015] R. Zhang, S. A. Candra, K. Vetter and A. Zakhor, 2015, 'Sensor fusion for semantic segmentation of urban scenes'. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1850–1857.

[Zhou et al., 2014] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba and A. Oliva, 2014, 'Learning deep features for scene recognition using places database'. In *Advances in Neural Information Processing Systems (NIPS)*, 487–495.

Appendix A

OpenABLE: An Open-source Toolbox for Application in Life-Long Visual Localization of Autonomous Vehicles

Visual information is a valuable asset in any perception scheme designed for an intelligent transportation system. In this regard, the camera-based recognition of locations provides a higher situational awareness of the environment, which is very useful for varied localization solutions typically required by the research community in long-term autonomous navigation, such as loop closure detection and VO or SLAM correction.

In this appendix, we present OpenABLE, an open-source toolbox contributed to the community with the aim of helping researchers in the application of these kinds of life-long localization algorithms. The implementation follows the philosophy of our topological place recognition method named ABLE, which was described in detail in the Chapter 3 of this dissertation. In addition, this open implementation includes several new features and improvements. These functionalities allow to match locations using different global image description methods and several configuration options, which enable the users to control varied parameters in order to improve the performance of place recognition depending on their specific problem requisites. The applicability of our toolbox in visual localization purposes for intelligent vehicles and autonomous robots is validated by varied results, jointly with comparisons to the main state-of-the-art methods.

The motivation of our toolbox is inspired by recent open projects for intelligent transpotation systems that have demonstrated to be a great public contribution to the community in their respective areas, such as the KITTI benchmark suite or the LIBVISO libraries for VO. Our goal is that OpenABLE can also give support to researchers in the specific area of life-long visual localization for autonomous vehicles or mobile robots.

A.1 Overview: The OpenABLE toolbox

As discussed along this dissertation, vision-based methods for identifying locations based on hand-crafted descriptors have been broadly studied in the recent past, especially after the definition of proposals such as FAB-MAP or SeqSLAM. Some time after their formulation, open implementations of FAB-MAP and SeqSLAM were released for the benefit of the research community: OpenFABMAP and OpenSeqSLAM. Motivated by these contributions, the OpenABLE toolbox described in this appendix also expects to assist in future researches related to the new challenges associated with the novel trends appeared in life-long visual localization for autonomous vehicles [Badino et al., 2011, Churchill and Newman, 2012a, Bansal et al., 2014].

OpenABLE is a project that arises from the growing interest in an open-source philosophy in research. Our main objective is to publicly share our toolbox with the intelligent vehicles, robotics and computer vision communities, because it can be helpful for the common progress of researchers in life-long visual localization topics. Due to this, the code provided in OpenABLE[1] can be freely employed and modified by the users, which can also include our algorithms into a larger system or directly apply them over their localization problems. In this sense, the toolbox contains varied novel configuration options to facilitate its utilization and it is easily adaptable to different applications and specifications. In Fig. A.1, a general diagram about our toolbox is presented.

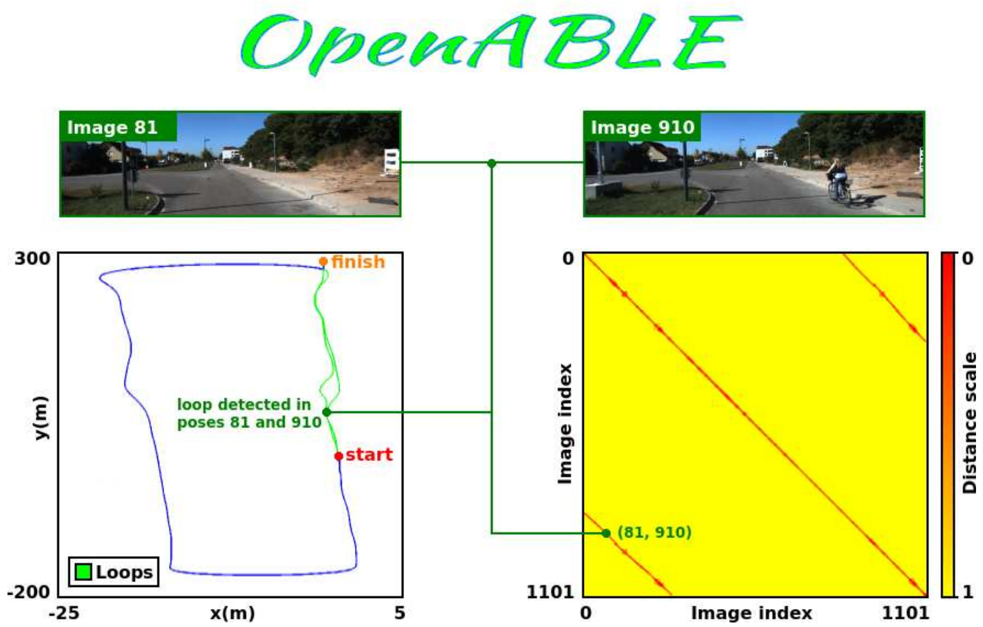


Figure A.1: OpenABLE logo and a general diagram. It graphically depicts how the application of our toolbox can help in varied localization problems, such as loop closure detection and metric measurements correction. The example showed in this case correspond to the recognition of a revisited location in the sequence 06 of the KITTI Odometry dataset.

[1]More information, extra material, videos and source code about the OpenABLE toolbox are available from the website of the project: http://www.robesafe.com/personal/roberto.arroyo/openable.html

A.2 Novelties in OpenABLE

OpenABLE is something more than a simple open implementation of the original ABLE method. The described toolbox contributes several new characteristics and varied functionalities with the aim of providing a wide range of possibilities to the users of the implemented code:

- Apart from the descriptors originally used by ABLE (LDB and D-LDB), OpenABLE supplies implementations based on other hand-crafted features such as BRIEF, BRISK, ORB, FREAK, SIFT, SURF and HOG.
- The typical Hamming matching commonly used for binary features is substituted by a L_2-norm when vector-based descriptors are chosen in the configuration options of the toolbox (SIFT, SURF, HOG).
- An image description method based on multiple grids is now implemented, as well as the common global description normally performed by ABLE.
- A thresholding functionality is provided for improving the filtering of loop closures in the similarity matrix.
- Multi-camera approaches are also available. OpenABLE exploits the extra image information procured in any case: monocular, stereo or panoramic.

In addition, OpenABLE has some other advantages with respect to similar toolboxes. For instance, OpenFABMAP requires a previous training, while our toolbox is completely training-free. Besides, our method has a better behavior for the changing fields of view typically appeared in life-long visual localization for autonomous vehicles, which is one of the weaknesses of OpenSeqSLAM, as evaluated in [Sünderhauf et al., 2013a] and in some results presented in this thesis in Section 3.5.2.6.

A.3 Main Characteristics of the Toolbox

OpenABLE is developed in C++ because it is a standard programming language adaptable to varied system requirements. Although the code has been designed under a Linux operating system, it is easily portable to other platforms such as Windows or Mac OS. OpenCV 3.0[2] is required because some of the computer vision algorithms programmed in OpenABLE apply functionalities of these libraries. Furthermore, a file named *CMakeLists.txt* is provided jointly with the code to facilitate the compilation of the toolbox by using CMake[3].

[2]OpenCV 3.0 is currently available from: http://opencv.org/
[3]CMake is currently available from: https://cmake.org/

The core and main functions of the toolbox are contained in the source code files named *OpenABLE.h* and *OpenABLE.cpp*. An evaluation program (*Test_OpenABLE.h*) is supplied, where an image sequence or a video is processed to return the final distance matrix, which is normalized between 0 and 1. The open code of the LDB descriptor provided by its authors in [Yang and Cheng, 2014] is also included jointly with our own implementation of D-LDB. In addition, a file named *Config.txt* is incorporated to easily adapt the properties and functionalities of OpenABLE to the different interests of the users.

A.4 Configuration Options

The configuration options provided by OpenABLE are explained in the following sections because of their importance for the application of the toolbox in a varied range of possibilities. Processing times are reported after execution to know the efficiency of an specific configuration.

A.4.1 Configuration Parameters for Datasets

Some basic parameters are configurable to select the paths to the input recordings or datasets that contain the images or videos used in the required visual localization tasks.

A.4.2 Configuration Parameters for Representation

The configuration options included in this group allow to define the paths where the results generated by OpenABLE will be stored and some other representation features, such as 12 ranges of colors to visualize the distance matrices.

A.4.3 Configuration Parameters for Description and Matching

Description and matching parameters are the most important for adapting the performance of OpenABLE to the users' priorities and they must be individually described in detail in the following pages.

A.4.3.1 Camera_type

This option allows to select the type of camera used for acquiring the images applied in the tests performed with OpenABLE: monocular (ABLE-M), stereo (ABLE-S) and panoramic (ABLE-P).

A.4.3.2 Description_type

This parameter lets to choose between a computation of features using a global or a grid-based image description. In [Arroyo et al., 2014a], some results proved that the application of grids can slightly improve the precision, but it also progressively increases the computational cost.

A.4.3.3 Patch_size

The length of the square defined as patch for the image description process can be adjusted in OpenABLE with this parameter. If a global description is used, the images are also downsampled to this size. In [Arroyo et al., 2014b], it is evidenced how a patch of 64x64 can be enough for an effective and efficient visual place recognition based on hand-crafted features.

A.4.3.4 Grid_x

It defines the number of horizontal grids applied over an image if grid-based description is enabled.

A.4.3.5 Grid_y

It defines the number of vertical grids applied over an image if grid-based description is enabled.

A.4.3.6 Panoramas

If the toolbox is configured for computing panoramic images, the number of subpanoramas considered in image processing can be selected using this parameter. This concept was explained in detail in Section 3.4.3, where it is exposed how a cross-correlation of subpanoramas can detect places revisited in an opposite direction (bidirectional loop closures).

A.4.3.7 Illumination_invariance

If this option is enabled, the illumination invariant technique is applied to improve the robustness in changing lighting conditions. Its implementation is based on the formulation introduced in Section 3.2.2.

A.4.3.8 Alpha

When illumination invariance is enabled, this parameter represents the value of α defined in Eq. 3.2 of Section 3.2.2.

A.4.3.9 Image_descriptor

Although ABLE originally applies LDB as core image descriptor (or D-LDB for stereo images), OpenABLE allows to choose other binary features, such as BRIEF, BRISK, ORB and FREAK. In addition, vector-based descriptors are also available, such as SIFT, SURF and HOG. In these cases, a matching based on the L_2-norm is internally computed by default (see Eq. A.1), because vector-based features are not compatible with the Hamming distance.

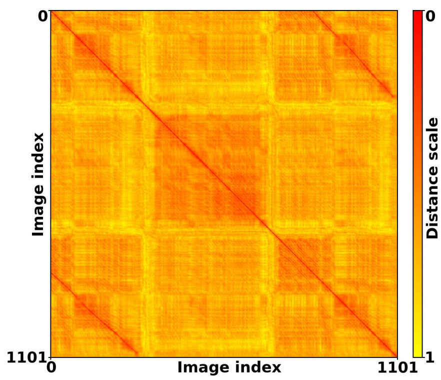
$$M_{i,j} = M_{j,i} = \sqrt{\sum_{k=1}^{n}(\mathbf{d}_{i_k} - \mathbf{d}_{j_k})^2}. \tag{A.1}$$

A.4.3.10 Image_sequences

The length of the sequence of images used by OpenABLE is configurable. If the value of this option is 1, single images are employed. A higher length gives better precision in life-long localization. A detailed study of this was provided in Section 3.5.2.3. For instance, a length of 300 is defined as proper for the Nordland dataset.

A.4.3.11 Threshold

This parameter allows to apply a threshold (θ) over the distance matrix calculated by OpenABLE. The advantage of using this option is that loop closures can be better detected, as shown in Fig. A.2.



(A.2.1) Before thresholding.

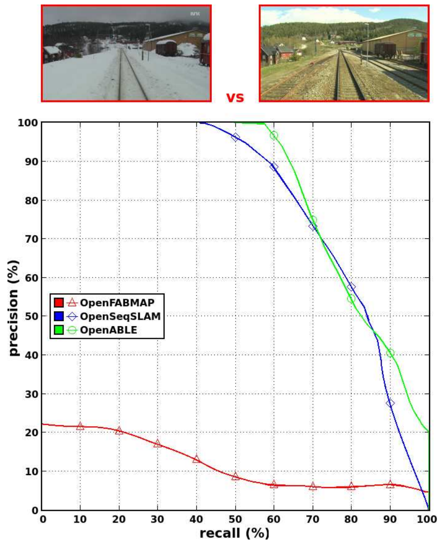
(A.2.2) After thresholding.

Figure A.2: Distance matrices generated by OpenABLE before and after thresholding. The example corresponds to the sequence 06 of the KITTI Odometry dataset. A standard stereo configuration is used and a thresholding parameter of $\theta = 0.2$.

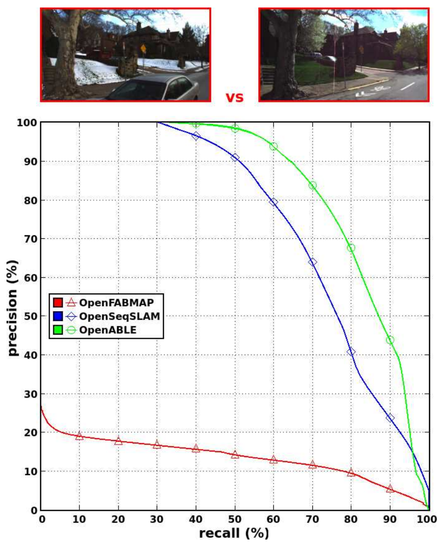
A.5 Experiments and Results

ABLE was widely validated over several datasets and under varied long-term conditions in the results presented in Chapter 3 for monocular, stereo and panoramic cameras. For this reason, in this section we only show some experiments that demonstrate that the code implemented in OpenABLE is equally robust for life-long visual localization.

We have performed some tests using the Nordland dataset and the CMU-CVG VL dataset, which are recorded in the most challenging long-term situations. Precision-recall results are shown for one of the most difficult cases in both datasets: sequences acquired in winter vs spring. These tests are presented in Fig. A.3, where OpenABLE is compared against the other state-of-the-art toolboxes. OpenFABMAP and OpenSeqSLAM are applied using their standard configurations defined in [Glover et al., 2012] and [Sünderhauf et al., 2013a], while OpenABLE uses the original parameters of the ABLE method in this case, which were exposed along Chapter 3. The results evidence the better performance of the proposals based on sequences of images (OpenSeqSLAM and OpenABLE). Besides, our toolbox has a more favorable behavior in datasets with changes on the field of view, such as the CMU-CVG VL dataset. These results corroborate the satisfactory performance of OpenABLE for life-long localization.



(A.3.1) Norland dataset.



(A.3.2) CMU-CVG VL dataset.

Figure A.3: OpenABLE vs state-of-the-art toolboxes. Precision-recall curves are depicted with the aim of supporting the performance comparison in the Nordland and CMU-CVG VL datasets. The sequences employed in the tests were recorded in winter and spring to validate the toolbox for life-long visual localization. An illustrative frame from the sequences matched is shown in order to visually understand the complexity of place recognition in each case.

A.6 Contributions and Conclusions

This appendix has formally presented OpenABLE, whose main characteristics and functionalities have been explained in detail for facilitating the usage of this open-source toolbox contributed to the research community. The satisfactory results have validated its performance compared to other open toolboxes in the state of the art, such as OpenFABMAP or OpenSeqSLAM. In addition, the practical application of our toolbox can be helpful in different life-long visual localization problems for intelligent vehicles or mobile robots, which confirms that OpenABLE can be a useful resource for researchers in this topic.

The source code of OpenABLE is publicly available and it is completely customizable by its users in order to adapt it to their own requirements. This is a clear advantage for the research progress with respect to other localization methods that do not follow this kind of open-source philosophy. In fact, some researchers have used our toolbox in their own research during the last months. More specifically, some authors have designed a method named FastABLE [Nowicki et al., 2016], which is derived from our research. In this case, FastABLE is an adaptation of our method that is focused on indoor localization using smartphones.

In the future, the code of OpenABLE will be maintained and updated when needed in the Github[4] repository of the project. Besides, other future improvements could be focused on including powerful descriptors derived from CNNs, such as the described in Chapter 4.

[4]The code repository of OpenABLE in GitHub is available from: https://github.com/roberto-arroyo/OpenABLE

Appendix B

Ground-truth Designed for Loop Closure Detection in the KITTI Odometry Dataset

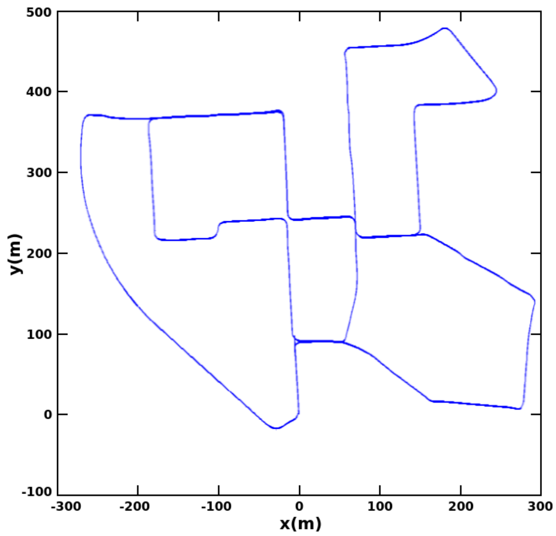
The KITTI Odometry dataset has 22 sequences of stereo images that have been commonly used along this thesis. These sequences include environments with different characteristics and challenging situations such as perceptual aliasing, changes on scene and a considerable amount of loop closures.

Nevertheless, there is not any specific ground-truth for loop closure detection in this dataset. Besides, the KITTI Odometry dataset has GPS measurements that are only available for 11 of the sequences.

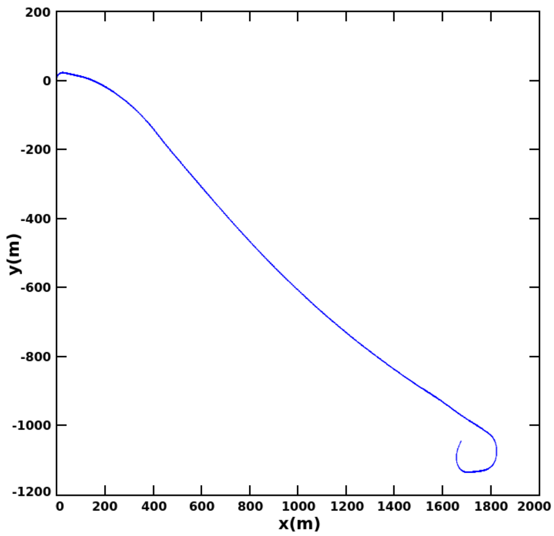
According to the previous considerations, we have created this ground-truth for all the sequences, which is described in Table B.1. There are 12 sequences that contain loop closures, 21 of them unidirectional and 4 bidirectional in total. The ground-truth matrices that we have created for this dataset are publicly available[1].

Additionally, we depict the ground-truth maps derived from the available GPS data in Fig. B.1 and Fig. B.2. As can be seen, these maps are not represented from sequences 11 to 21, because in these cases GPS measurements are not given. For this reason, these sequences were manually annotated in the loop closure ground-truth presented in Table B.1, with the aim of making possible their evaluation in our tests. In fact, the results obtained along the different chapters of this dissertation for the KITTI Odometry dataset can be compared to all the information presented in this appendix for a better understanding of them.

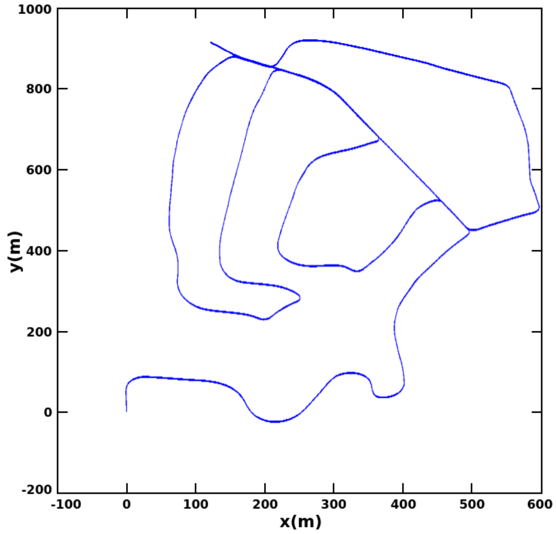
[1]The ground-truth is available from: http://www.robesafe.com/personal/roberto.arroyo/downloads.html



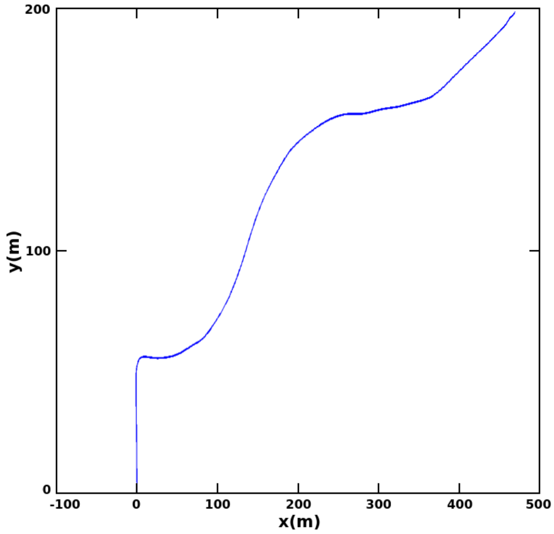
(B.1.1) Sequence 00.



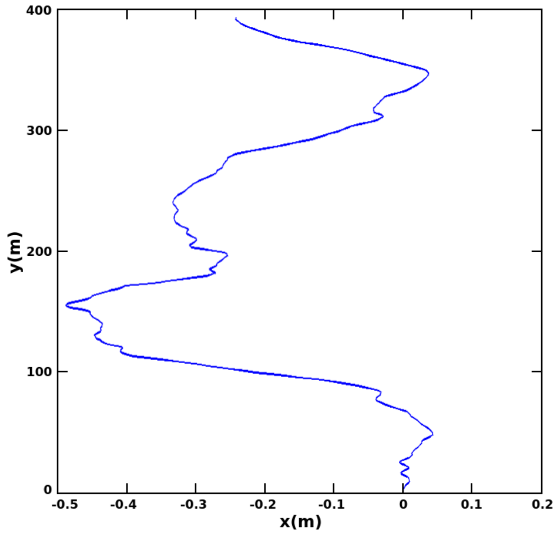
(B.1.2) Sequence 01.



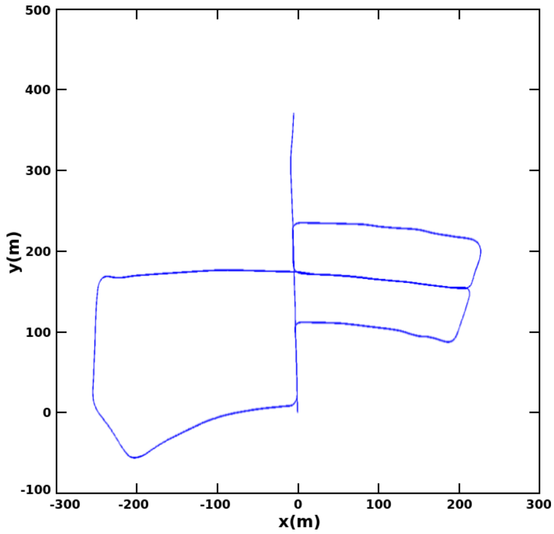
(B.1.3) Sequence 02.



(B.1.4) Sequence 03.

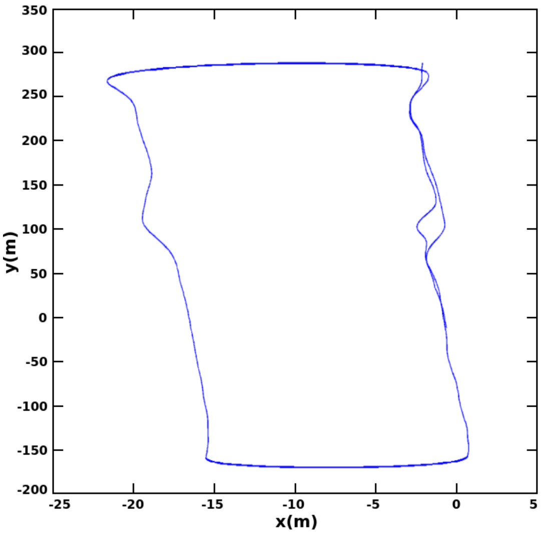


(B.1.5) Sequence 04.

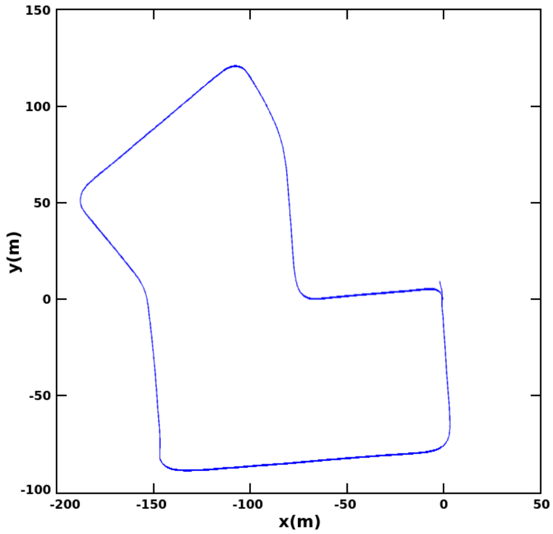


(B.1.6) Sequence 05.

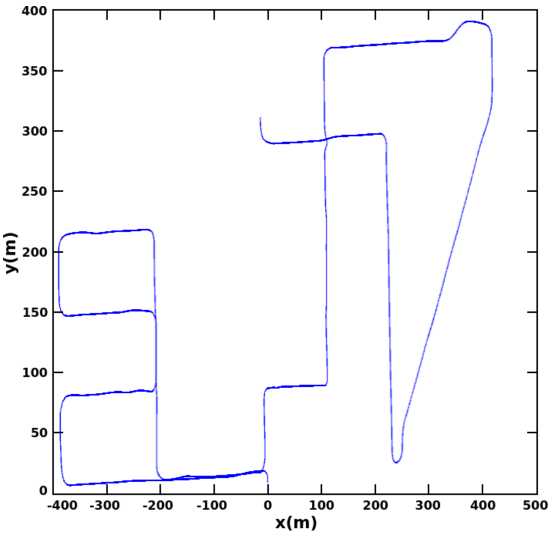
Figure B.1: Ground-truth maps based on GPS for the KITTI Odometry dataset (I).



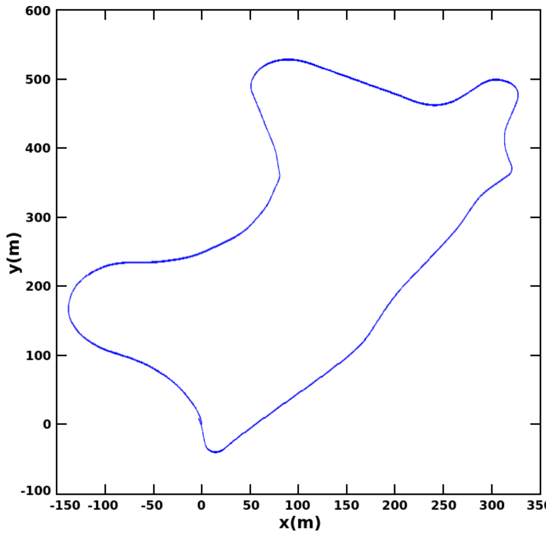
(B.2.1) Sequence 06.



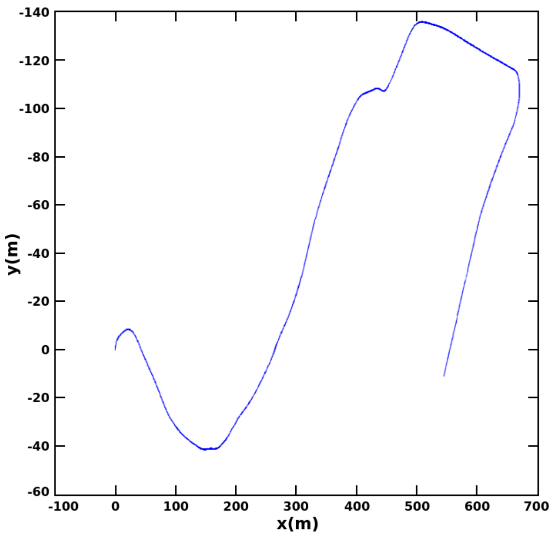
(B.2.2) Sequence 07.



(B.2.3) Sequence 08.



(B.2.4) Sequence 09.



(B.2.5) Sequence 10.

Figure B.2: Ground-truth maps based on GPS for the KITTI Odometry dataset (II).

Table B.1: Ground-truth for loop closure detection in the KITTI Odometry dataset.

Sequence	No. of frames	Unidirectional loop closures			Bidirectional loop closures		
		No.	Initial frames	Loop frames	No.	Initial frames	Loop frames
00	4541	5	0000 - 0099	4451 - 4528	0		
			0122 - 0196	1570 - 1635			
			0392 - 0412	2446 - 2460			
			0392 - 0941	3398 - 3844			
			2354 - 2460	3295 - 3418			
02	4661	2	0933 - 1026	4205 - 4266	1	3332 - 3397	4566 - 4620
			1810 - 1997	4404 - 4569			
05	2761	3	0031 - 0121	2431 - 2512	0		
			0565 - 0787	1324 - 1530			
			0819 - 0885	2581 - 2627			
06	1101	1	0000 - 0280	0835 - 1093	0		
07	1101	1	0000 - 0013	1060 - 1067	0		
08	4071	0			2	0075 - 0227	1640 - 1796
						0726 - 0765	1422 - 1464
09	1591	1	0000 - 0023	1578 - 1590	0		
13	3281	4	0000 - 0138	2152 - 2316	0		
			0000 - 0089	3188 - 3280			
			0553 - 0839	1633 - 1938			
			2152 - 2264	3188 - 3280			
15	1901	1	0000 - 0086	1808 - 1900	0		
16	1731	1	0000 - 0146	1614 - 1730	1	0023 - 0079	0798 - 0843
18	1801	1	0322 - 0493	1616 - 1800	0		
19	4981	1	4246 - 4390	4812 - 4943	0		