MULTIPLE VISUAL DESCRIPTOR COMBINATION FOR LOOP CLOSURE DETECTION AND VISUAL ODOMETER TRAJECTORY ESTIMATION

MOHAMMED OMAR MOH’D SALAMEH

UNIVERSITI KEBANGSAAN MALAYSIA

MULTIPLE VISUAL DESCRIPTOR COMBINATION FOR LOOP CLOSURE DETECTION AND VISUAL ODOMETER TRAJECTORY ESTIMATION

MOHAMMED OMAR MOH’D SALAMEH

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY,
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

HIMPUNAN PELBAGAI DESKRIPTOR VISUAL UNTUK PENGESANAN GELUNG DAN ANGGARAN TRAJEKTORI ODOMETER VISUAL

MOHAMMED OMAR MOH'D SALAMEH

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH
IJAZAH DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

2 February 2018

MOHAMMED OMAR MOH’D SALAMEH
P68462

ACKNOWLEDGEMENTS

In the Name of Allah, Most Merciful and Most Compassionate

Throughout the years of my research for this PhD theses, I have encountered a lot of unusual hardships and challenges which usually face most foreign students. Therefore, I would first of all like to thank my supervisor Associate Professor Dr Azizi Abdullah and co-supervisor Associate Professor Dr Shahnorbanun Sahran for their great support and academic advice.

My deepest gratitude goes to my dear mother, without having to mention reasons for the thanks because she means everything to me. I also would like to thank my brothers and sister who never stopped their support and prayers for my success.

Special thanks go to my beloved wife, Razan, who spared no efforts to facilitate my daily life, and for her continued support throughout my study years, before and beyond, without whose steadfastness and patience, this thesis wouldn't have seen the light. I should not exclude my little daughter, Larin, whose smiles and tears did make my life a real joy.

And finally, I am deeply indebted to my dear father-in-law, Mr Omar Hammad, for his incessant fatherly advice, and for his linguistic support and the proofreading of the research.

ABSTRACT

Memory management is one of the crucial elements in Visual Simultaneous Localization and Mapping (VSLAM) for long-time autonomous navigation in a real environment. The real environment involves landmark variations which create challenges to VSLAM for a robot to recognise the visited landmark locations and estimate its trajectory. In VSLAM, many algorithms utilised a single descriptor for describing landmarks for loop closure detection (LCD). However, LCD using single descriptors may hinder the recognition of landmarks and may worsen LCD performance as a database of image maps grows. Thus, this work proposes an Ensemble Bayesian Filter for Active Locations (EBF-AL) for LCD. The EBF-AL is based on the Real-Time Appearance-Based Mapping (RTAB-Map) model to deal with the growing of image maps. In RTAB-Map, it manages the active locations in Working Memory (WM) for LCD, and stores the passive locations in Long Term Memory (LTM). However, some relevant of passive locations that are transferred to LTM may be undetected by LCD. Thus, a proposed algorithm, namely Ensemble Bayesian Filter for Active and Passive Locations (EBF-APL) is used to manipulate information from LTM and WM for LCD. Besides LCD, one important element that utilises visual descriptor for VSALM is visual odometry trajectory estimation (VOTE). Similar to LCD, the most widely used solution for VOTE is arguably using a single keypoints descriptor. However, VOTE using a single keypoints detector seems to be unreliable due to image variation problems in finding the best corresponding features in other target images. Therefore, this research proposes a method that used a random sampling scheme. The scheme extracts the best keypoints from a different type of keypoints detector to reduce the trajectory estimation errors. The proposed algorithms then evaluated on several benchmark datasets namely Lip6 Indoor, Lip6 Outdoor and City Centre for LCD and KITTI for VOTE. The results show that EBF-AL and EBF-APL outperformed the standard RTAB-Map that uses WM for LCD. EBF-AL achieves a recall of 80%, 97% and 86% and EBF-AL score a recall of 91%, 98% and 88% respectively on the Lip6 Indoor, Lip6 Outdoor and City Centre data sets respectively. In trajectory estimation experiment, the proposed algorithm can eliminate the trajectory error of 44%, 8% and 13% on KITTI dataset for the sequence 00, 02 and 05 respectively.

ABSTRAK

Pengurusan ingatan adalah salah satu unsur penting dalam Penyetempatan dan Pemetaan Serentak Visual (VSLAM) untuk navigasi berautonomi masa lama dalam persekitaran sebenar. Persekitaran sebenar melibatkan variasi mercu tanda yang menimbulkan cabaran VSLAM terhadap robot untuk mengenali lokasi yang dikunjungi dan menganggarkan trajektori berkaitan. Dalam VSLAM, banyak algoritma menggunakan penerang tunggal untuk menggambarkan mercu tanda bagi Pengesanan Penutupan Gelung (LCD). Walau bagaimanapun, LCD menggunakan deskriptor tunggal boleh menghalang pengecaman mercu tanda dan boleh memburukkan prestasi LCD apabila peta pangkalan data imej semakin bertambah. Oleh itu, kajian ini mencadangkan suatu Penggabungan Penapis Bayesian untuk Lokasi Aktif (EBF-AL) untuk LCD. EBF-AL didasarkan pada model Pemetaan Berdasarkan-Penampakan Masa-Sebenar (RTAB-Map) untuk menangani pertambahan peta imej. Peta RTAB menguruskan lokasi aktif dalam Ingatan Kerja (WM) untuk LCD, dan menyimpan lokasi pasif dalam Ingatan Jangka Panjang (LTM). Walau bagaimanapun, beberapa lokasi pasif yang dipindah ke LTM mungkin tidak dapat dikesan oleh LCD. Jadi suatu algoritma yang dicadangkan, iaitu Penggabungan Penapis Bayesian untuk Lokasi Aktif dan Pasif (EBF-APL) digunakan untuk memanipulasi maklumat dari LTM dan WM untuk LCD. Selain daripada LCD, satu perkara penting lain yang menggunakan penerang tunggal visual untuk VSALM ialah Anggaran Trajektori Odometri Visual (VOTE). Sama seperti LCD, penyelesaian yang paling banyak digunakan ialah penerang titik-titik kunci tunggal. Walau bagaimanapun, VOTE menggunakan satu pengesan titik-titik kunci tunggal tidak boleh dipercayai kerana masalah variasi imej dalam mencari ciri-ciri yang paling sesuai dengan imej sasaran yang lain. Oleh itu, kajian ini mencadangkan suatu kaedah yang menggunakan skim pensampelan rawak. Skim ini mengekstrak jenis titik-titik kunci terbaik yang berbeza untuk mengurangkan ralat anggaran trajektori. Algoritma yang dicadangkan kemudiannya diuji pada beberapa dataset penanda aras iaitu Lip6 Indoor, Lip6 Outdoor dan City center bagi LCD dan KITTI bagi VOTE. Keputusan menunjukkan bahawa EBF-AL dan EBF-APL mengatasi RTAB-Map piawai yang menggunakan WM untuk LCD. EBF-AL mencatat nilai perolehan 80%, 97% dan 86% dan EBF-AL masing-masing mencatat nilai perolehan sebanyak 91%, 98% dan 88% masing-masing di Lip6 Indoor, Lip6 Outdoor dan City Centre. Dalam eksperimen anggaran trajektori, algoritma yang dicadangkan boleh menghapuskan ralat trajektori sebanyak 44%, 8% dan 13% pada dataset KITTI untuk urutan masing-masing 00, 02 dan 05.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

Figure No. Page

LIST OF ABBREVIATIONS

ANN	Approximate Nearest Neighbours.
BLOB	Binary Large OBject.
BoRF	Bag-of-Raw-Features.
BoW	Bag-of-Words.
BoWP	Bag of Word Pairs.
BRIEF	Binary Robust Independent Elementary Features.
DBF	Discrete Bayesian Filters.
DLT	Direct Linear Transform.
DoG	Differences of Gaussian.
DoH	Determinant of Hessian.
EBF-AL	Ensemble Bayesian Filters for Active Locations.
EBF-APL	Ensemble Bayesian Filters for Active and Passive Locations.
FAB-MAP	Fast Appearance-Based Mapping.
FAST	Features for Accelerated Segment Test.
FIFO	First-In-First-Out.
FLANN	Fast Library for Approximate Nearest Neighbours.
FN	False Negatives.
FP	False Positives.
FSH	Feature Stability Histogram.
kd-tree	k-dimensional tree.
KITTI	Vision Benchmark Suite from Karlsruhe Institute of Technology and Toyota Technological Institute.
LCD	Loop Closure Detection.
LoG	Laplacian of Gaussian.
LTM	Long Term Memory.
MD-VOTE	Multiple Descriptors for Visual Odometry Trajectory Estimation.
MIH	Multi-Index Hashing.
NNDR	Nearest Neighbour Distance Ratio.
OACH	Orientation Adjacency Coherence Histogram.
OFAST	Orientation FAST Keypoint.
ORB	Oriented FAST and Rotated BRIEF.

PHOG	Pyramid Histogram of Oriented Gradients.
PIRF	Position Invariant Robust Features.
PIRFs	Position-Invariant Robust Features.
PnP	Perspective-n-Point.
PnP-RANSAC	Perspective-n-Point using RANSAC.
PR	Precision-Recall.
RANSAC	RANdom SAmple Consensus.
RatSLAM	Rat SLAM: a hippocampal model for simultaneous localization and mapping.
RTAB-Map	Real-Time Appearance-Based Mapping.
SE	Sensitivity.
SFM	Structure From Motion.
SIFT	Scale-Invariant Feature Transform.
SLAM	Simultaneous Localization And Mapping.
SM	Sensory Memory.
SP	Specificity.
STM	Short Term Memory.
SURF	Speeded-Up Robust Features.
TE	Trajectory Estimation.
TF-IDF	Term Frequency-Inverse Document Frequency.
TN	True Negatives.
TP	True Positives.
VO	Visual Odometry.
VOTE	Visual Odometry Trajectory Estimation.
VSLAM	Visual Simultaneous Localization and Mapping.
WGII	Weighted Grid Integral Invariant.
WGOH	Weighted Gradient Orientation Histogram.
WM	Working Memory.

CHAPTER I

INTRODUCTION

1.1 RESEARCH BACKGROUND

In our modern life, service robots are needed in a variety of fields to successfully navigate and perform their tasks by generating a map useful for autonomous navigation throughout their lives.

For a mobile robot to navigate successfully, it must be capable of constructing a map where it can detect its location using built-in sensors and maintaining the map so that it will remain useful for navigation over the entire service life of the robot. This capability is referred to as persistent localisation, mapping and navigation, and is closely related to the well-studied Simultaneous Localization And Mapping (SLAM) problem, a widely researched topic in the realm of robotics (Thrun & Leonard 2008) .

The SLAM algorithm can construct a map in two different approaches. In the first approach, the sensors' data are processed online where the map registers observations as they occur. In the second approach, the sensors' data are recorded in the memory which will be used offline for constructing the map (Cummins & Newman 2010b; Rehder & Albrecht 2015). The online mapping SLAM approach is the target system which is capable of operating the autonomous mobile robot in real-time.

Decades of research produced great SLAM algorithms focusing on one or more aspects of SLAM solutions. In the literature, SLAM based on laser rangefinder was used for constructing a grid-metric map with particle filtering (Grisetti et al. 2005; Montemerlo et al. 2002) or batch optimisation (Kümmerle et al. 2011; Thrun et al. 2006) approaches. The grid-metric map is an occupancy grid which is constructed around the geometry of a mapped area which is also known as geometric map (Elfes 1987; Taketomi et al. 2017). The geometric map is hard to expand and fit into areas representing larger environments without high computational costs (Thrun, et al. 2008). Likewise, the geometric map strives to detect the loop closure locations as a result of odometry error accumulation.

On the other hand, the visual sensors (e.g. cameras) become widely used in SLAM because they are light, less power-consuming, affordable and provide comprehensive perspectives. These approaches are known as Visual Simultaneous Localization and Mapping (VSLAM). VSLAM depends on appearance-based representation and is used to construct a topological map. This topological map is constructed by tracking feature points with probability filters (Angeli et al. 2008a, 2009; Chli & Davison 2009; Cummins & Newman 2008b; Davison 2003; Milford & Wyeth 2008) or implementing 3D map construction based on bundle adjustment (Klein et al. 2007; Mei et al. 2011). VSLAM constructs a topological map which is a graph structure based on describing the landmark connectivity of a surrounding environment. The landmark locations are saved in the graph nodes, and the graph edges record the transformation between neighbourhood nodes in time and/or location (Salameh et al. 2017; Taketomi et al. 2017).

The comprehensive solution to autonomous visual robot navigation is built on three main pillars (Nguyen et al. 2013; Stachniss 2006) as shown in Figure 1.1: (1) Mapping, which is a method of representing the visual features environment captured by a robot's sensors. (2) Localisation, which is the ability of a robot to detect its location within the map referred to in this research as LCD. (3) Visual path planning is a subset of a local map which constructs a correlated sequence of image locations connected between an initial robot's pose and target locations, whereby a robot can navigate safely, collision-free, through the environment using these paths. Furthermore, memory management is the base which makes all these pillars function harmoniously and efficiently.

As regards VSLAM, the crucial aspects of mapping and localisation, it provides the robot with the capability of perceiving the data captured by the visual sensors from

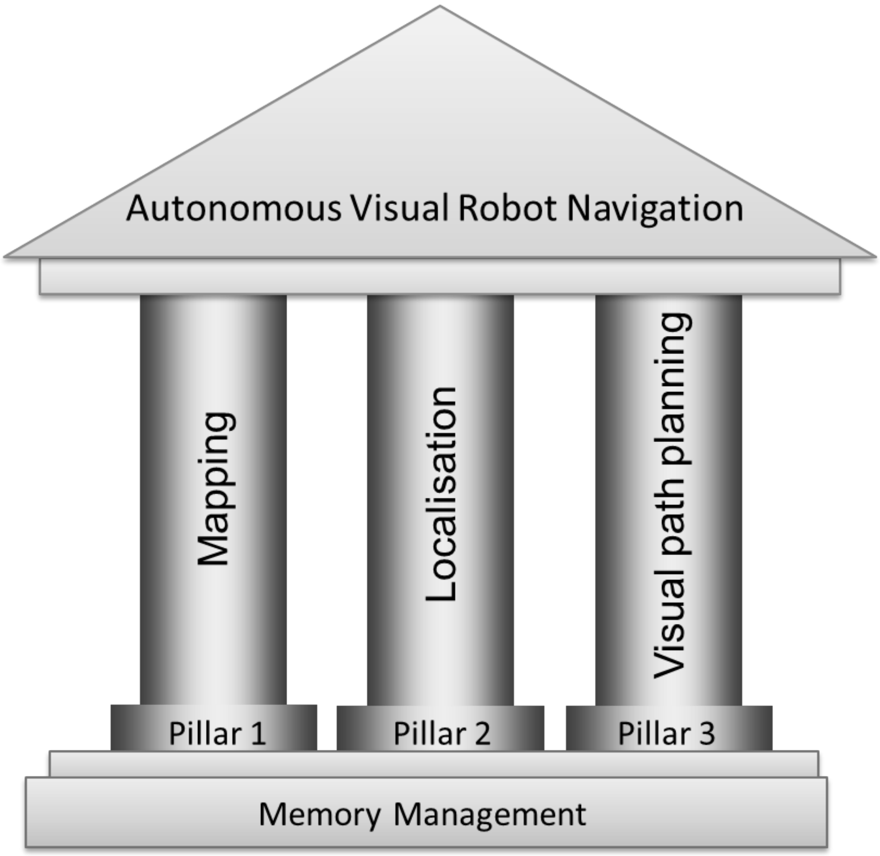


Figure 1.1: A comprehensive solution to autonomous visual robot navigation

the surrounding environment used in constructing the map with image features of the visited locations. In spite of the advantages provided by the visual sensors, they are still not devoid of noise. This associated noise prevents the robot from creating accurate mapping and localisation. And here the need arises for obtaining an accurate LCD method that can recognise previous locations once the robot visits the location again.

LCD makes the robot construct a consistent map with more accurate representation of the environment and uncertain locations. In addition to the sensor's noise, the real environment suffers from high fluctuating scenes, such as scaling, perspectives, illumination and scattering objects. Therefore, new challenges emerge (Garcia-Fidalgo & Ortiz 2015b), such as: growing mapping which makes LCD process a linear problem because the more images are captured, the more complex processes are required to detect the loop closure locations. Additionally, LCD is a repetitive process which is necessary for a robot to evaluate each new image under real-time constraints (García Fidalgo et al. 2016; Labbe & Michaud 2013). An indoor environment contains a lot of locations which are similar to each other and share a large number of features (e.g. long corridor) in contrast with the outdoor environment. It is influenced by the movement of sun rays (i.e. change in the illumination and shadows). Because of these reasons, different locations are recognised as loop closure locations whereas, in reality, they are false loop closure detections. Most recent researches focus on the proposed solutions to tackle the these challenges.

One of the popular approaches for loop closure detection is Fast Appearance-Based Mapping (FAB-MAP) (Cummins & Newman 2008b). FAB-MAP is proposed to use a single descriptor (e.g. Scale-Invariant Feature Transform (SIFT)) to generate location landmarks using offline Bag-of-Words (BoW) and a Bayesian filter used as a prediction model to predict the loop closure candidates.

The visual feature combination approaches are introduced to recover the lack of single descriptor landmarks, like a combination of global features with local features (Goedemé et al. 2004, 2007), a hierarchical descriptor and multi-layer descriptors (Murillo et al. 2007a; Wang & Yagi 2013, 2012). Nonetheless, these solutions have the same error rate of a single detection system and face the overfitting of combining multiple features into a single large feature vector.

Examples of the bio-inspired VSLAM approaches are Rat SLAM: a hippocampal model for simultaneous localization and mapping (RatSLAM) (Milford et al. 2004) and RTAB-Map (Labbe & Michaud 2013). In RatSLAM, the association between environment, scene and pose of the robot's model is the rodent hippocampus, which is not a probabilistic method for visual localisation and mapping. However, it has some limitations, for example, it strongly depends on the lighting conditions for data associations and become inefficient for long time navigation.

RTAB-Map uses the single visual feature Speeded-Up Robust Features (SURF) to construct an online BoW. RTAB-Map adopts the human memory model (Labbe & Michaud 2013) to successfully run the loop closure detection in real-time for a large area. RTAB-Map splits the memory into two main parts, the first one is the WM with pre-determined temporal criteria which determine the WM size to keep the operation of a robot under real-time constraints. This part of memory contains the active locations which are directly used by a Bayesian filter for estimation loop closure hypothesis. An active location is either a new location or a frequently visited one. The second part of memory is Long Term Memory (LTM) which is the enlarged memory of passive locations having been transferred from WM to LTM based on preset conditional criteria. This solution with a single descriptor leads to the lack of discriminative image feature, and it affects the ability of a robot to recognise some previously visited locations. Regarding the memory model adopted by RTAB-Map, some relevant locations or data are transferred to LTM, and are based on preset criteria, which cause the neighbourhood locations to split into two different memory areas. In this case, the prediction model "Bayesian filter" becomes unable to correlate the split neighbourhood locations at the same time to accurately estimate the loop closure hypotheses.

The visual path planners require a map capable of identifying locations attached to their poses. A pose is a combination of position and orientation of the visual sensor "camera" which represents the translational and rotational matrix between a corresponding sequence of image locations. The Visual Odometry (VO) approaches estimate the poses under the VSLAM. The Trajectory Estimation (TE) is a backward-tracking method for evaluating the path that a camera moves through the poses of locations which have been visited by a robot during a certain period of time. As for VO, it focuses on constructing consistent trajectory using a local map to obtain highly accurate local trajectory estimation, whereas VSLAM tackles global map consistency.

In the literature, the Perspective-n-Point using RANSAC (PnP-RANSAC) has been successfully adopted to estimate the location poses used for assembling the trajectory of a robot. The progress of PnP-RANSAC has earned a considerable attention from researchers, since this method was introduced by Fischler and Bolles (Fischler & Bolles 1981), regarding convergence speed and performance (Bhattacharya & Gavrilova 2013; Guo et al. 2015; Liu et al. 2017; Wang et al. 2016; Xing & Huang 2010; Zhang et al. 2011). Pertinent researchers contributed to the development of PnP-RANSAC which focuses on the matching methods (Bhattacharya & Gavrilova 2012; Li et al. 2014), the keypoint selection methods (Shi et al. 2013), enhance the processing time (Zhao et al. 2016). Notably, in spite of the importance of the keypoint detection process, not too many works tackle this topic. The importance of this process lies in the capability of a visual descriptor to extract the distinctive keypoints back-trackable among image locations having been visited by a robot.

Furthermore, the keypoints selection process is a crucial part of VOTE because it provides the matching pair keypoints as the main input data for trajectory estimation. It is noted that the main source of outliers points and the noise in the pair keypoints are the feature matching methods. However, the real environment suffers from high fluctuating scenes, such as scaling, perspectives, illumination and scattering objects, where current approaches using a single visual descriptor cannot distinguish the keypoints to find their corresponding features in other image frames.

1.2 PROBLEM STATEMENT

VSLAM is an approach reliable for autonomous navigation which is used by robots for constructing a map of the surrounding environment and detecting its location within the map. These two processes are heavily dependent on each other, where the visual sensors are the primary source of visual information captured from the service area to be used in mapping and localisation processes. VSLAM is a target research domain because of its effective applicability in autonomous navigation.

The development of autonomous visual robot navigation is an ongoing process. The performance efficiency of the autonomous navigation is built on three main pillars as shown in Figure 1.1: Mapping, localisation and visual path planning, and here the need arise for efficient memory management to make the three pillars function harmoniously.

Several types of research have been made on LCD to improve VSLAM performance. The improvement resulted from these researches was obtained from different visual descriptors or by developing a loop closure hypothesis estimation. Other researches focus on improving a memory management model that is capable of running VSLAM for a long time operation in a large area. Some other studies also focus on VOTE approaches whereby it introduces different methods to estimate the camera poses derived from camera calibration.

A primary goal of VSLAM is to construct a map which identifies each location and recognises them in the next visit. The LCD is a continuous process whereby it verifies whether a new captured image feature corresponds to a previously visited location that is available on the map. However, LCD performance relying on a single feature descriptor worsens as a map of images grows (Angeli et al. 2008c; Polikar 2012; Rebai et al. 2014). Besides, multiple descriptors are used and combined to determine discriminative landmarks. Yet, it generates a dense feature vector; in that, it will increase the problem of overfitting and will hinder the generalisation performance (Abdullah et al. 2010). Therefore, effective indexing and searching in the large image map is required and remains a challenge for LCD.

One of the crucial functions of VSLAM is the memory management which ensures the continuity of a robot's operation under real-time constraints as long as possible within a growing map. RTAB-Map is the state-of-the-art VSLAM algorithm which mimics the human memory model (Labbe & Michaud 2013). RTAB-Map splits the memory into its two main parts Working Memory (WM) and Long Term Memory (LTM). The WM is used for loop closure detection with a constrained number of active locations contained in the WM. The LTM with enlarged memory of passive locations being transferred from WM based on pre-set conditional criteria. However, some relevant locations or data having been transferred to LTM, and based on pre-set criteria, which cause the neighbourhood locations to split into two different memory areas. In this case, the prediction model becomes unable to correlate the split neighbourhood locations at the same time to accurately estimate the previous visited location (Hua & Tan 2017).

Trajectory estimation is a part of localisation task in VSLAM where a robot estimates the camera pose in each image location visited. The poses are estimated using Visual Odometry Trajectory Estimation (VOTE) by extracting distinctive and trackable keypoints from sequence image locations having been visited by a robot. However, the sensor noise and the high fluctuating scenes constitute an inevitable shortcoming that reduces the single visual descriptor performance in extracting the distinctive and trackable keypoints.

The poses are the main data used for visual path planning. VO estimates the pose location by extracting distinctive and trackable keypoints from sequential image locations having been visited by a robot. PnP-RANSAC is a common approach used for estimating the VOTE which uses a feature descriptor such as SURF to extract keypoints and match them by using their corresponding descriptors. However, the sensor noise and the high fluctuating scenes constitute an inevitable shortcoming that reduces the descriptor performance in extracting the distinctive and trackable keypoints.

1.3 HYPOTHESES

The hypotheses that were set in this research to improve the efficiency of the VSLAM regarding the improvement of the LCD performance enhance the memory management and reduce the trajectory error of the robot.

Hypothesis 1 *Ensemble of visual descriptors in WM that can be used to improve the LCD performance.*

Hypothesis 2 *Combination of WM and LTM. can improve the LCD performance.*

Hypothesis 3 *Multiple keypoint detectors can extract a set of distinctive keypoints which can improve the trajectory estimation.*

1.4 RESEARCH QUESTIONS

The following research questions are the direct motivation behind this research:

RQ1 What is the appropriate model rich enough to represent the landmarks locations to improve the performance of Loop Closure Detection?

RQ2 How can the memory management scheme be improved in Loop Closure Detection in order to have prominent solutions?

RQ3 How can the multiple models combinations address the problem of extracting distinctive keypoints to improve the Visual Odometry Trajectory Estimation?

1.5 OBJECTIVE OF RESEARCH

The aim of this research is to propose a variant of VSLAM algorithm that enhances the LCD and the VOTE by adopting multiple descriptors and ensemble learning approach. The proposed algorithms will improve the performance of the LCD for both indoor and outdoor environments, and reduce the trajectory estimation error for long journeys in an outdoor environment. This research aims at realising the following objectives:

1. To propose a method that combines multiple models in working memory for loop closure detection.
2. To enhance the loop closure detection performance by manipulate information from long term memory and working memory in memory management.
3. To propose keypoint selection for multiple models combination for visual odometry trajectory estimation.

1.6 SCOPE OF RESEARCH

This research focuses on the improvement of LCD and VOTE under VSLAM. The improvement of the LCD is achieved by using ensemble learning method for combining multiple models based on multiple visual descriptors that accurately estimates the loop closure hypothesis and by enhancing the memory management model which can retrieve relevant locations. The improvement of VOTE is also achieved by introducing a point selection scheme to facilitate the VO approach for poses estimation.

First, the loop closure hypothesis estimation is improved, and the result of this improvement is evaluated on the widely used public datasets Lip6 Indoor, Lip6 Outdoor and City Center (Angeli et al. 2008c; Cummins & Newman 2008b). Also, the evaluation of the enhancement of the loop closure hypothesis estimation is measured by precision-recall criteria (Cummins & Newman 2008b; Fuentes-Pacheco et al. 2015; Garcia-Fidalgo & Ortiz 2014; Labbe & Michaud 2013; Scherer et al. 2013) and the significant improvement is evaluated using the statistical t-test, as explained in Section 3.2.4.

Second, the proposed enhance of the memory management has improved the performance of LCD to work beside the enhanced loop closure hypothesis estimation to solve the challenging in growing map. Precision-recall criteria measure the improvement of the proposed algorithm, and the significant improvement is evaluated by using the statistical t-test as explained in Section 3.2.4.

Third, the keypoint selection scheme for VO approach is benchmarked with the three longest sequence of the Vision Benchmark Suite from Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) dataset. The trajectory error is calculated based on the translational and rotational errors. The translational error is measured in percentages, and the rotational error is measured by degrees per metre where the translational error is calculated using segments of a trajectory at 100, 200,. . . , 800 m lengths (Geiger et al. 2013).

1.7 THESIS ORGANISATION

This thesis consists of seven chapters including the current one, and is organised as follows:

Chapter II Literature Review overviews VSLAM approaches and focuses on LCD and VOTE approaches. The chapter includes analytical details and discussion on recent and relevant researches carried out on VSLAM and illustrates the main problems and challenges correlated with the subject. The main properties of the most relative LCD and VOTE approaches are listed herein. Special attention is paid to RTAB-Map because of its importance in representing the state-of-the-art approach in VSLAM. This chapter is so designed as to cover the essential information required to develop the proposed algorithms in this research.

Chapter III Methodology displays the methodology followed in this research. Furthermore, Chapter III illustrates the evaluation criteria which is used to assess the proposed algorithms.

Chapter IV Ensemble of Bayesian Filter for Loop Closure Detection presents and discusses the proposed EBF-AL algorithm for LCD, and elaborates on its processes in constructing multiple Bayesian models visual descriptors and combine them at decision level. The evaluation of the proposed EBF-AL algorithm is analysed and benchmarked against the state-of-the-art VSLAM algorithms.

Chapter V Ensemble Bayesian Filters for Active and Passive Locations for Loop Closure Detection focuses on the memory management EBF-APL algorithm for LCD. It aims at developing a new memory management approach using three different visual descriptors to identify, select and retain the prominent relative locations to construct valid nearest neighbourhood indexing for more accurate loop closure hypothesis estimation. The experiments carried out by the proposed EBF-APL algorithm are demonstrated in details.

Chapter VI Visual Odometry Trajectory Estimation illustrates the MD-VOTE approach which is introduced for improving the VOTE performance. The MD-VOTE is based on the extracted distinctive keypoints which are combined from three different keypoint descriptors. Also, the new filter is set to retain the most prominent ones. The MD-VOTE has experimented on the longest three journeys of the KITTI dataset. The trajectory estimation errors are calculated by estimating the translational and rotational errors.

Finally, Chapter VII Conclusion and Future Work summarises the research, outlines the main contributions and concluding the research. This Chapter also submits recommendations and suggestions for the future research in the VSLAM domain.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter discusses the techniques used for Loop Closure Detection (LCD) and Visual Odometry Trajectory Estimation (VOTE) applied in VSLAM systems. VSLAM is a crucial system that a robot uses to construct a map and localise its current location on the map which keeps growing while new locations are explored, as discussed in Section 2.2. These locations, whether new or previously visited, are distinguished by applying the LCD to each captured image.

The LCD depends on the features extracted from the image to identify the corresponding locations. However, the performance of the LCD algorithms, which relies on a single feature descriptor, worsens as a map of images grows. Section 2.3 shows the LCD challenges and discusses the LCD approaches according to different feature descriptors.

A growing map is a challenge facing the VSLAM, where the excessive and increased number of image features being stored in the map make a robot not to be able to recognise the obtained image features in a real-time operation. Section 2.4 discusses the memory management approach regarding the growing map and LCD.

Regarding VSLAM that serves autonomous navigation where it requires to extract the metric data of the surrounding environment. The metric data in VSLAM is represented by the robot's pose which is estimated by the VOTE. Section 2.5 discusses the VOTE approaches.

Section 2.6 discusses the related VSLAM systems: FAB-MAP, Angeli et al. works, PIRF-Nav2.0 and RTAB-Map.

Section 2.7 shows the public datasets which are used in evaluating the performance of the proposed algorithm in this research. The last section is the Summary.

2.2 VISUAL SIMULTANEOUS LOCALIZATION AND MAPPING (VSLAM)

SLAM is an approach for a mobile robot to construct a map of the surrounding environment while at the same time it recognises the robot's location on the map (Ajay & Venkataraman 2013; Fuentes-Pacheco et al. 2015; Korrapati 2013).

The use of robotic visual sensors in SLAM is an essential element in perceiving and obtaining measurements of the features from the environment. They also serve as a crucial source for constructing a map reliable enough to be used for autonomous navigation in unknown indoor or outdoor environments (Aulinas et al. 2008).

The visual sensors have different types which produce various features of the surrounding environment. For an efficient VSLAM approach, the type of sensor must be carefully chosen according to the nature of environment so as to produce reliable visual information on which the mapping method depends.

2.2.1 Mapping

Mapping is a registration process for observing landmark features captured by a robot from the surrounding environment and saving them in a structural manner that mimics the environment's landmark positions in relation to each other, and/or according to a reference position in correlation with other landmarks to be used for a robot's diverse functions (Bonin-Font et al. 2008; Fuentes-Pacheco et al. 2015; Taketomi et al. 2017; Thrun et al. 2002).

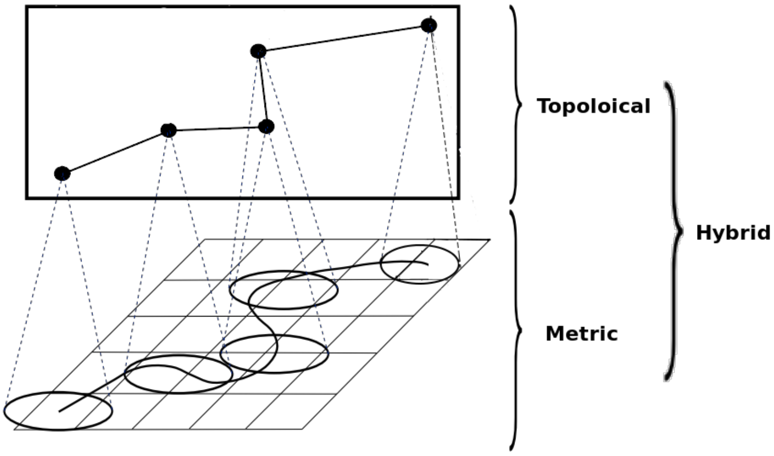


Figure 2.1: Types of map representation
Source: (Chapoulie et al. 2013)

Mapping process is a mimic representation of an environment conducted by a robot to be used for a robot's diverse functions if it is provided with a map in advance.

Figure 2.1 shows common types of map representation: metric map, topological map and hybrid map. These types of map are discussed as follows:

a. Metric Representation

Metric maps represent the surrounding environment regarding the geometric relations between landmarks and a fixed reference frame (Yousif et al. 2015). Maps of this type are capable of maintaining detailed information about the surrounding environment regarding physical features as measurement, size and distance. This type of map has multiple forms and the most common form is the occupancy grids.

The occupancy grids map is a metric map which is developed from the geometric structure as represented by the environment (Elfes 1989). The general scheme for constructing the occupancy grids map starts by dividing the targeted area into equally spaced cells as grids. Second, the grids cells matrix represents the space, whereby each grid registers the estimation of the probable occupancy of the space grid. Finally, the probable occupancy is limited by (0,1), where "0" represents a vacant space, and "1" represents an occupied space, and also "0.5" represents an unregistered space. Grid state representation requires the modelling of the measurements of a robot's sensors associated with the stored map because it cannot be obtained directly from the sensors data.

One of the limitations of the occupancy grids map is that it needs modelling the sensor’s data into geometric forms, which restrict this map to environments only fitted to such constraints. Other limitations appear in scaling the grid map to a large expandable environment. Additionally, an environment with more precise features requires a finer grid "grid granularity" which leads to an increased computational effort and storage capacity (Gokhool 2015; Yi et al. 2012).

Feature map is another branch of metric maps which describes the environment as a set of sparse geometric patterns "features", like points and lines, which are associated with a position in the metric space (Thrun et al. 2005). With a limited number of sparse features in representing the environment, it gives some improvement in operating for large environments. However, it is highly sensitive to false measurements of the features associated with that of the saved map. Besides, feature maps do not represent the free space; these maps cannot be used for obstacle avoidance and trajectory estimation.

b. Topological Representation

A topological map is a method of constructing a map describing the surrounding environment in an abstract representation which is contrary to the metric map that depends on the geometric information of landmarks.

Figure 2.2 shows an example of a topological map which represents a part of Kula Lumpur train stations which is easy to understand even without any geometrical information.

The topological map uses a graph structure where the nodes include similar features of locations and the edge represents the relationship between two nodes including actions, distance or position. Based on a graph structure, the topological maps have compact information and more efficient in memory space storage than metric maps. The graph search algorithms can be applied directly to solve the navigational problems and

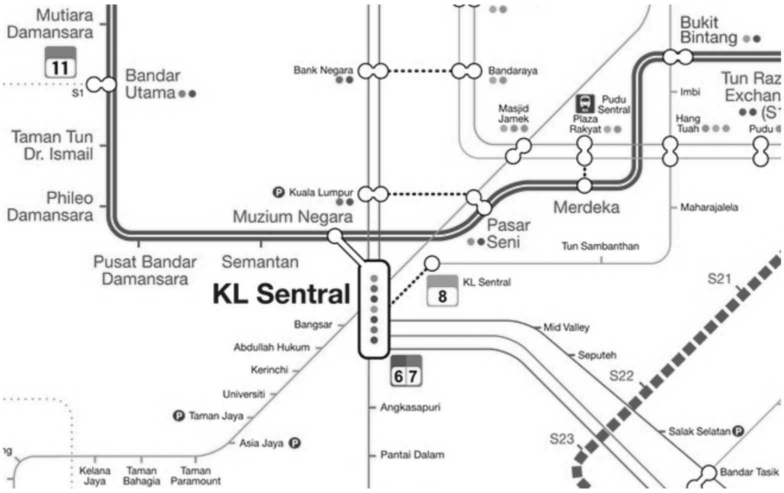


Figure 2.2: An example of topological map from daily life: Kuala Lumpur train stations
Source: (MRT 2017)

to improve the data association performance. Additionally, the topological maps can handle the growing maps of a large environment.

A VSLAM approach uses a topological map based on the assumption that locations contain distinctive features through which the locations can be recognised from their neighbourhood. These features are grouped into a single node to represent the location on the map. Furthermore, edges represent the relationship among the locations; such as neighbourhood, loop closure or even action information (Garcia-Fidalgo & Ortiz 2014; Saeedi et al. 2014; Wang & Yagi 2013). Feature extraction, description and the matching approaches will be discussed in Section 2.3.

c. Hybrid Representation

Metric maps and topological maps have different characteristics where metric maps are more accurate but computationally more expensive. On the other hand, topological maps are more scalable to the large environment but cannot provide any metrical data about the environment. Hybrid maps combine both types of map to get the best of their capabilities. An example of the hybrid map is the topo-metric map which is a topological map containing metric data.

The topo-metric map is a topological map merging the metric data with its own structure. This kind of merging can produce two forms of topo-metric maps:

1. When the local metric data are merged with the nodes which are linked topologically to each other; the topo-metric map becomes a global map. This type of map is known as hierarchical maps (Bazeille et al. 2015; Blanco et al. 2008; Bosse et al. 2004; Estrada et al. 2005; Lim et al. 2012).
2. When the pose metric data are merged with the edges, the edges will contain the translational and rotational matrix between the corresponding nodes. The SLAM approaches using this type of maps are known as pose-graph SLAM algorithms (Angeli et al. 2008c; Cummins & Newman 2011; Labbe & Michaud 2013; Mur-Artal & Tardós 2017; Sünderhauf & Protzel 2012).

Based on the advantages obtained from the hybrid map as previously explained, this research adopts the topo-metric map structure with edges containing the pose metric data (Angeli et al. 2008c; Cummins & Newman 2011; Labbe & Michaud 2013; Mur-Artal & Tardós 2017) which is a graph-based map expressed in Equation 2.1

$$G = (V, E) \tag{2.1}$$

where V, E represent a set of nodes and edges respectively. The nodes have a constant unique ID and each node v_i represents a location L_i where $ID \neq i$. The set of n nodes are expressed in Equation 2.2

$$V = (v_1, ..., v_n). \tag{2.2}$$

Also, each node v_i contains the associated-feature signature z_i which is extracted from the image location I_i. The set of m edges link the n nodes as expressed in Equation 2.3

$$E = (e_{1,1}, ..., e_{n,n}) \tag{2.3}$$

where edge $e_{i,j}$ represents a relationship between node v_i and v_j which can be expressed in Equation 2.4

$$e_{ij} = (v_i, v_j). \tag{2.4}$$

As mentioned earlier, one of the advantages of the topological map lies in its ability to contain different data that enrich the map which makes it represent the surrounding environment accurately. Such data which are the location poses are required for a robot's navigation. In the topological map, edges contain the poses which represent a transformation model between two locations. The transformation $T_{e_{ij}} \in \mathbb{R}^{4\times4}$ is expressed in Equation 2.5 (Scaramuzza & Fraundorfer 2011)

$$T_{e_{ij}} = \begin{bmatrix} R_{e_{ij}} & T_{e_{ij}} \\ 0 & 1 \end{bmatrix} \tag{2.5}$$

where $R_{e_{ij}} \in SO(3)$ is the rotation matrix $\alpha^i, \beta^i, \gamma^i$ for the Euler Angles. SO(3) is the Special Orthogonal Group that represents rotation matrices which its elements rotations around a line in 3D and defined as Equation 2.6 shows. $T_{e_{ij}} \in \mathbb{R}^{3x1}$ is the translation vector x^i, y^i, z^i. Approaches for estimation $transformation$ will be discussed in Section 2.5.

$$SO(3) \doteq \{R \in \mathbb{R}^{3x3} : R^T R = I,\ det(R) = 1\} \tag{2.6}$$

2.2.2 Localisation

Localisation is the process of using the observed sensor readings of a robot to identify the location of the robot on the map which is instantly built in a timely manner according to the movement of the robot (Garcia-Fidalgo & Ortiz 2015a; Payá et al. 2017; Taketomi et al. 2017). The building of the map depends on the sensor data association in detecting whether the observed features belong to a new location or to the location already saved on the map. This process is known as Loop Closure Detection (LCD) which detect the robot's position and reduces the location uncertainty. The following Section discusses the LCD approaches and their relevance to the topological map.

2.3 LOOP CLOSURE DETECTION (LCD)

The correct perception of the surrounding environment is one of the essential aspects of VSLAM as regards mapping and localisation. Although different types of visual sensors are used to carry out this task, all of them have the problem of noise. This associated noise of the sensors directly affects the efficiency of mapping and localisation as VSLAM relies only on raw sensor data. Therefore, the LCD is highly required for the accurate mapping and localisation. LCD can be defined as a process that ensures whether the current location has been visited for the first time or it had been visited before, and is stored on a map. Accordingly, the robot will be able to reduce the uncertainty locations and will construct a consistent representation of the environment (Angeli et al. 2008a; Labbe & Michaud 2013).

The LCD is a really challenging task because of the following reasons (Angeli et al. 2008c; Glover et al. 2010; Hajebi 2015; Kejriwal et al. 2016):

1. Sensor noise: As mentioned earlier, the associated noise which is caused by the high fluctuating scenes.
2. Perceptual aliasing: this challenge appears when two different locations have similar features but wrongly identify them as loop closure locations.
3. Environment variations: day and night, light and shade and moving objects; all of these factors increase the difficulty in identifying the same location.
4. Scalability: it will become harder for a robot to identify the current location and compare it with the increasing number of previously visited locations on a growing map.

Figure 2.3 shows a general form of the LCD processes. The following subsections discuss the LCD approaches which use the global and local descriptor to generate the

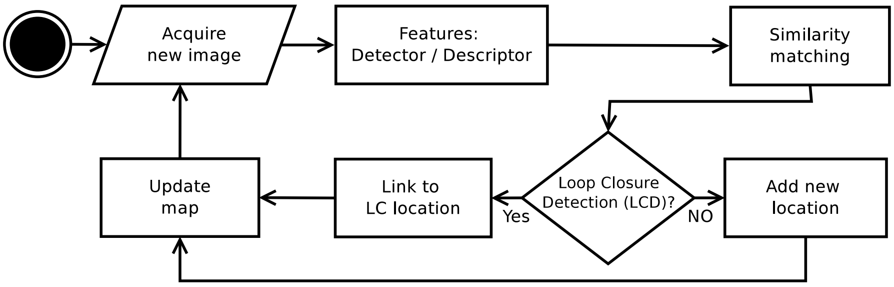


Figure 2.3: A general form of loop closure detection

Source: (Labbe & Michaud 2013)

landmarks and their similarity matching methods.

2.3.1 LCD Approaches based on Global Descriptors

Global descriptors; such as Histograms Descriptor and GIST Descriptor (Oliva & Torralba 2001) are utilised primarily to define an image in a whole manner taking into account the full image in the process. These descriptors are usually fast to compute which makes it suitable for the process of matching images for localisation and mapping tasks. They have been used in many applications characterised by place recognition with appropriate results.

However, the global descriptor approaches lack the ability to differentiate between the background and the foreground of image locations, and the information obtained from each part is mixed together which increases the problem of the perceptual aliasing (Hassaballah et al. 2016). Additionally, a global description is a compact representation of the entire image which makes it practically inefficient in identifying the locations containing perspective transformation, change in illumination or a partial occlusion (Chen et al. 2014; Liu et al. 2016).

Accordingly, local features use the individual pixels data or discrete regions. In fact, the salient features of objects which are highly textured are obtained from the feature detectors. The description of these features is calculated by feature descriptors. This description represents the image location which is more robust to the perspective transformation and change in illumination (Liu et al. 2016). The following subsection discusses the LCD approaches which use the local features to identify the image locations.

2.3.2 LCD Approaches based on Local Features

Local features are used to identify distinctive keypoints in an image location. The keypoint features can be extracted from a neighbourhood operation or by detecting prominent structures in the image, such as Binary Large OBject (BLOB)s or corners. After that, a description of an image location is calculated based on the features. In general, local features can handle the change in image scale and the variance in illumination and deal with issues that the global descriptor fails to handle; such as partial occlusion. And these characteristics make the local feature more suitable to be used in LCD approaches. Tuytelaars et al. (2008) summarised the characteristics of the effective feature detector as regards accuracy, quantity, locality, distinctiveness, efficiency and repeatability. The repeatability is an important feature which can be obtained, either by invariance where big distortion appeared with the change of the relevant view, or by robustness when a little distortion happens.

The Scale-Invariant Feature Transform (SIFT) descriptor is one of the most popular descriptors used in VSLAM approaches. David Lowe introduced the SIFT (Lowe 2004) which is a feature detection and a descriptor. SIFT extracts the local features from the remarkable keypoints based on the object appearance where the feature of the keypoints are invariant to image scale and rotation and can handle the variation of illumination and tilt. The following points summarise the main process of the SIFT:

1. Scale-space: Gaussian blurred approach used to generate multiple scale images from the input image. The blurred images are used for calculating the Differences of Gaussian (DoG) from the neighbours in the scale space.
2. Keypoint detection: The previous step generates a lot of keypoints which are locally extremal in the DoG images for both space and scale. The keypoints selection is based on Hessian matrix thresholding which eliminates the point of the edge with a high ratio and corner points with a low ratio. Additionally, the keypoints with the low contrast are eliminated and only the keypoints which are

interpolating through the DoG images remain.

3. Keypoint orientation: The local image gradient directions are used to register the orientation for each keypoint. The keypoint orientation is an essential characteristic for the keypoint descriptor regarding the image rotation problem.
4. Keypoint descriptor: The final step is a highly distinctive descriptor based on the 128-bins histogram of local oriented gradients which calculates each keypoint, and stores the bins in a 128-dimensional vector to obtain highly distinctive descriptors for the keypoints.

A local descriptor is a floating-point data type in a multi-dimensional vectors format. Recently, binary descriptors started to attracted more attention to researchers because of their bit representation characteristics which reduce the storage and processing time. Garcia and Ortiz (2015b) collected the most common local features and summarise their features regarding their capability of handling the transformation: rotation, scale and affine. Additionally, the local features are categorised according to their data types: floating-point or binary.

Many VSLAM approaches use local features to construct topological mapping and perform localisation tasks.

Kosecka and Yang (2004) adopted the SIFT descriptor for representing the image location of indoor environments, and a simple voting method for a global localization. The dynamic environment issues have been solved by including the neighbourhood information in a Hidden Markov Model with the likelihood function of measuring the correspondence matching between the current image location and the locations stored in a map. This work is improved in Li & Kosecka (2006) by selecting the distinctive feature order to reduce the number of keypoints per image location. The selection criteria are based on computing the information entropy which uses posterior probabilities of all features.

Zhang (2011), in addition, selected a subset of SIFT features from an image to construct the Bag-of-Raw-Features (BoRF). The selected criterion is based on the feature's scale level where they are located. The set of features which are matched in several sequential images represent the image location. Additionally, a feature indexing method based on k-dimensional tree (kd-tree) structures is used in the work of Liu & Zhang (2012). kd-tree is a binary search tree useful to overcome the limitation of the linear search for BoRF with the incremental number of features. He et al. (2006) constructed a growing topological map based on the idea of using the persistent features in representing the image locations. They used the manifold constraints approach to select the SIFT features which are persistent in the image locations. Additionally, Sabatta (2008) used the idea of persistence of the SIFT features within a sequence of omnidirectional images which improve the LCD and enhance the SIFT descriptor by adding the colour information.

In the recent years, a new approach has been introduced whereby a map is built with distinctive landmarks which are extracted from the multiple images in the area between the nodes (Johns & Yang 2011). With a query image, the resulting matches will be transformed into landmarks instead of normal images, which will eventually lead to a continuous dense topological map without too much speed.

This probabilistic localisation approach has been produced by using the selected distinctive features of every landmark. Also introduced was the Position Invariant Robust Features (PIRF) method whereby average features from SIFT is used to create matchability within a span of sequential image frames (Kawewong et al. 2010, 2011b). These image frames are represented by PIRFs whose appearance variations are considered small in relation to the robot's movement. Such image features are used to construct the growing appearance-based SLAM algorithm, referred to as PIRF-Nav.

Likewise, a new form of the SIFT descriptor was modified by Andreasson and Duckett (2004) to capture omnidirectional image features, where keypoints are detected according to a Sobel filter and the descriptor computed using a Modified Scale Invariant Feature Transform (M-SIFT). The M-SIFT approach was used by Valgren et al. (2006), who reflected the surrounding environment through the image similarity matrix.

Bay et al. (2006) present Speeded-Up Robust Features (SURF) which is relatively similar to SIFT. SURF uses integral images, and an efficient scale space method in order to detect the keypoints and generate efficient descriptors. The following points summarise the main processes of the SURF:

1. Keypoint detection: Bay et al., use integral images instead of using DoG in order to compute the Laplacian of Gaussian (LoG) images by applying a 2D box filters, and the cost of the computation does not rely on the size of the box filter. Next, the Hessian matrix is applied to detect the keypoints which are invariance to the location and scale because the size of the image is the same despite the size difference of the box filter. After that, the keypoint orientation is calculated according to Haar wavelet which applies to each scale where the keypoint is detected.
2. Keypoint descriptor: a square window is built with the corresponding keypoint in the middle. The oriented window is set according to the keypoint orientation and is divided into 4×4 square sub-windows. Haar wavelets are calculated in each sub-window to produce 4 values. The combination of these values from the 4×4 sub-windows constructs the vector descriptor of length 64 for each window. These descriptors are invariant to scale, rotation and contrast.

The SURF descriptor is one of the commonly used approaches in VSLAM because of its robustness to the photometric and geometric distortions.

Kawewong et al. (2011a) extend PIRF and present the PIRF 2.0 with a dictionary management approach to extract the SURF features from the image location. The PIRF 2.0 is discussed in subsection 2.6.3. Also, Cummins & Newman (2011) extend FAB-MAP and present the FAB-MAP 2.0 with an inverted index method and extract the SURF features to handle a large area. The FAB-MAP 2.0 is discussed in subsection 2.6.1. And, Labbé & Michaud (2011); Labbe & Michaud (2013) present RTAB-Map which extracts the SURF to construct an online-BoW. The RTAB-Map is presented in subsection 2.6.4.

SIFT and SURF are the most common vector descriptors used in VSLAM.

Recently, the binary descriptor has been introduced in VSLAM as an alternative to the vector descriptor. The binary descriptor is a bit vector representation which is very efficient in storing and matching (Muja & Lowe 2012; Rublee et al. 2011a). The binary descriptor requires a keypoint detection before computing the descriptor, and each bit of the descriptor representing the result of the pairwise pixels intensity comparison of the sampling pattern around a keypoint.

Calonder et al. (2010, 2012) present the Binary Robust Independent Elementary Features (BRIEF) descriptor. This descriptor is constructed according to the pixels intensity comparisons of a subregion centred by the keypoint. And the descriptor matching uses the Hamming distance according to the XOR binary operation. The BRIEF descriptor is not invariant to the scale and rotation.

Rublee et al. (2011b) present the Oriented FAST and Rotated BRIEF (ORB). ORB enhances the Features for Accelerated Segment Test (FAST) keypoint detector with an orientation estimated for each keypoint according to the linked vector between the keypoints and the means centroids of the image moments, and this keypoint detection method is known as Orientation FAST Keypoint (OFAST). Similarly, ORB uses and enhances the BRIEF descriptor. After that, the OFAST's keypoints are detected, and the binary test pattern is rotated according to the keypoints orientation for extracting the BRIEF descriptor. ORB is the fastest and most resistant descriptor to the Gaussian image noise, unlike SIFT (Kumar & Sreekumar 2014).

Han & Fang (2017) present a VSLAM approach and extract the ORB features image location to construct Multi-Index Hashing (MIH) according to Approximate Nearest Neighbours (ANN) instead of BoW in order to handle the perceptual aliasing problem. They use a Bayesian filter for estimating the loop closure hypothesis and calculate the similarity matches between the current image and all candidate locations using MIH according to Hamming distance between the features. Han et al. (2017) extend their work in order to achieve a higher accuracy in loop closure detection. They handle the burstiness problem of the binary features used in MIH. The burstiness problem occurs when a statistically independent model cannot predict visual features which appear in an image with high frequency. Also, they use a feature selection method as an early process to eliminate the features that have a large Hamming distance in order to maintain the storage size.

2.3.3 LCD Approaches based on Bag-of-Words Schemes

The Bag-of-Words (BoW) approach is commonly used in VSLAM because BoW can quickly obtain similar image location candidates in a large map. The BoW method is originally developed for text retrieval, and in VSLAM it is used to improve the performance of finding the similar candidates of the query image location. Each feature in the image location is represented as a visual word, and the quantity of these visual words construct the visual vocabulary according to a set of representative features.

Usually, the visual vocabulary used in the BoW approach is constructed offline based on a train data. The offline-BoW has limitations; such as the visual vocabulary that cannot be used in a different environment like the one used in the training phase. One of the most well-known VSLAM approach which constructing an offline-BoW is FAB-MAP which is discussed in subsection 2.6.1.

Online-BoW approaches are presented to overcome the problem in the offline-BoW. The online-BoW approaches construct the visual vocabulary incrementally based on the obtained data.

a. On-line Approaches

The online-BoW approaches constructing a visual vocabulary synchronise with the robot's movement in discovering the surrounding environment. This visual vocabulary is adapted and maintained according to the images obtained from the operating environment.

Filliat (2007) constructs a visual vocabulary dynamically. A linear search method is used to select similar features which are grouped, and the mismatched feature based on the similarity measure is used to create a new visual word in the vocabulary. This approach is used in the VSLAM system in a small area due to the inability of the linear search method.

Angeli et al. (2008b) extend Filliat's approach, where the incremental vocabulary is constructed in a tree structure. The Bayesian filter is used for estimating the loop closure hypothesis where the likelihood is estimated according to the Term Frequency-Inverse Document Frequency (TF-IDF) coefficients of the visual word found in the current image.

Labbé & Michaud (2011); Labbe & Michaud (2013) introduce an online-BoW approach which is constructed according to a limited number of image locations. These image locations are selected by memory management which keeps the recent and most frequently visited locations to be used in constructing the incremental vocabulary. The vocabulary is a tree structure constructed according to the Nearest Neighbour Distance Ratio (NNDR).

2.3.4 LCD Approaches based on Combined Schemes

The combined approaches have been proposed in many different researches in order to improve the robustness of the VSLAM approaches. The multi-modal descriptors can be adopted to increase the reliability of LCD. The idea is to compute a different type of descriptor for the same image location which can maximise the advantages of each descriptor. This combination can achieve a highly discriminative descriptor capable of handling the noise and the challenges in the real environment that faces a robot in distinguishing the locations.

Researchers commonly utilise the global descriptors as an approach to ensure a fast search of similar images. After that, the local features are used to accurately confirm the associated process.

Goedemé et al. (2004) extract vertical column segments from omnidirectional images. The descriptor is constructed by combining ten different descriptors (three colour invariants and seven intensity invariants) into a single descriptor vector. These descriptor vectors are clustered and are stored in a kd-tree structure. The loop closure candidates are selected according to the column segments distance supported by the Bayesian filter. Goedemé et al. (2007) added the SIFT descriptor to the combination for the localisation process and constructing a topological map which is used in a navigation system.

Murillo et al. (2007a) introduce a hierarchical localisation approach. This approach starts by selecting the suitable candidate locations according to the global colour descriptor matching. Then, the line features are applied to these candidates to select the loop closure location based on pyramidal matching. After that, the metric data are extracted by using the 1D radial trifocal tensor. Murillo et al. (2007b) expand their work by replacing the line features with the SURF feature to achieve more robustness and accurate matching.

Wang & Yagi (2013) use their global descriptor; namely Orientation Adjacency Coherence Histogram (OACH) to select the loop closure location candidates, and this process which is called coarse localisation. After that, the Harris-Laplace method is used for keypoint detection from the image locations which resulted from the coarse localisation. The SIFT descriptor describes these keypoints, and this process is called fine localisation. Finally, a RANdom SAmple Consensus (RANSAC) based verification is applied to evaluate the image association.

Weiss et al. (2007a) compute the Weighted Gradient Orientation Histogram (WGOH) and Weighted Grid Integral Invariant (WGII) global descriptors independently. Then, the normalised histogram intersection matching between the two images according to each descriptor separately. The final output is calculated according to the product of each result. This similarity matching process is used to update the weights particle for the localisation task. Weiss et al. (2007b) expand their previous work by estimation the robot's pose according to SIFT descriptor when the global descriptors approach fails in detecting the loop closure.

Angeli et al. (2008c) extract the SIFT and the local colour histograms independently from the image location to construct two incremental BoWs. Each image location has two descriptors to estimate the likelihood separately. The two likelihood outputs are combined by using the multiplication function, and the result is used in the Bayesian filter for estimating the loop closure detection.

Inspired by Angeli et al. work, Chapoulie et al. (2011) extract the SIFT descriptor and compute the histogram of the SIFT's keypoints distribution which is used as global features. And the Bayesian filter combines the two descriptors to estimate the loop closure detection.

Garcia-Fidalgo & Ortiz (2017) introduce a hierarchical loop closure approach. The Pyramid Histogram of Oriented Gradients (PHOG) descriptors are extracted and used to select the loop closure location candidates. Next, the ORB features are extracted from the candidate's locations. The likelihoods of both descriptors are combined and are used with the Bayesian filter for estimating the loop closure detection.

2.4 MEMORY MANAGEMENT METHODS FOR GROWING MAP AND LCD

One of the primary challenges of a robot in VSLAM is to operate for a long time and serving in a large area where VSLAM constructs a growing map. The growing map requires an increase in the amount of time to process the LCD for new observations. However, when the LCD processing time becomes longer than the capture image time, a delay appears, which makes the updating and processing of the map difficult to achieve online.

The classical LCD approach required to match between the observed image and each image location stored in a map and this matching process is a linear search. Some researchers used the tree structure for storing the location features which speed the matching process, but with the tree structure it requires rebuilding the tree every time a new location feature is added; i.e. to overcome the problem of the unbalanced tree, which consumes time too (Suger et al. 2014).

Researchers in the VSLAM domain start to focus on designing VSLAM approaches capable of serving in a large area and operating for a long time under real-time constraint.

The most common challenges that are associated with the growing map for VSLAM are summarised as follow:

1. Maintaining the stability of the time required for the search process while the map is growing. The regular VSLAM approach matches the observed image to all previously visited locations, to estimate the likelihood of detecting a loop closure location. However, this action increases the LCD processing time linearly with the number of stored locations. Furthermore, the VOTE approach requires a back-ends optimisation process to estimate a trajectory of all locations visited by a robot. However, this process requires an operating time exponential to the number of locations used in trajectory estimation.
2. Keeping the storage size required for mapping synchronized with the explored area of the environment, but not being attached to the operational time. In case of a robot stops, it is not reasonable to store all the data obtained from the same location. This point is parallel to the previous one, where the LCD process uses a limited number of locations, and the processing time will also be limited.

In the VSLAM literature, some of the researches proposed solutions to these challenges by eliminating repetitive data trying to control the map size. Konolige & Bowman (2009) clustered the similar image locations at nodes in the topological map. The current captured image is clustered according to the score of the inlier similar match. This score is a percentage of the number of matched features to the average number of the features in the corresponding locations. After that, the observed features are deleted according to the least-recently-used constrain to eliminate the repetitive features and keep at least one exemplary view from each cluster.

Nature-inspired approaches are proposed by researchers to overcome the challenges on the VSLAM with growing map and operation for a long time. Atkinson & Shiffrin (1968a) have proposed that model for the human memory which is divided into multiple memory known as "multi-store model" as shown in Figure 2.4. This model has three parts: Sensory Memory (SM): where input information is stored.

Short Term Memory (STM): where selective criteria are applied to move specific information from Sensory Memory (SM) to Short Term Memory (STM), and the unattended information can be forgotten from the STM. The last part is Long Term Memory (LTM): which keeps the information for long time period; this information moves from STM to LTM based on a rehearsal process.

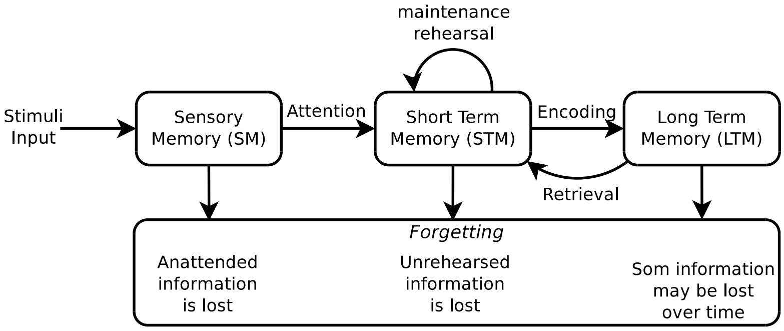
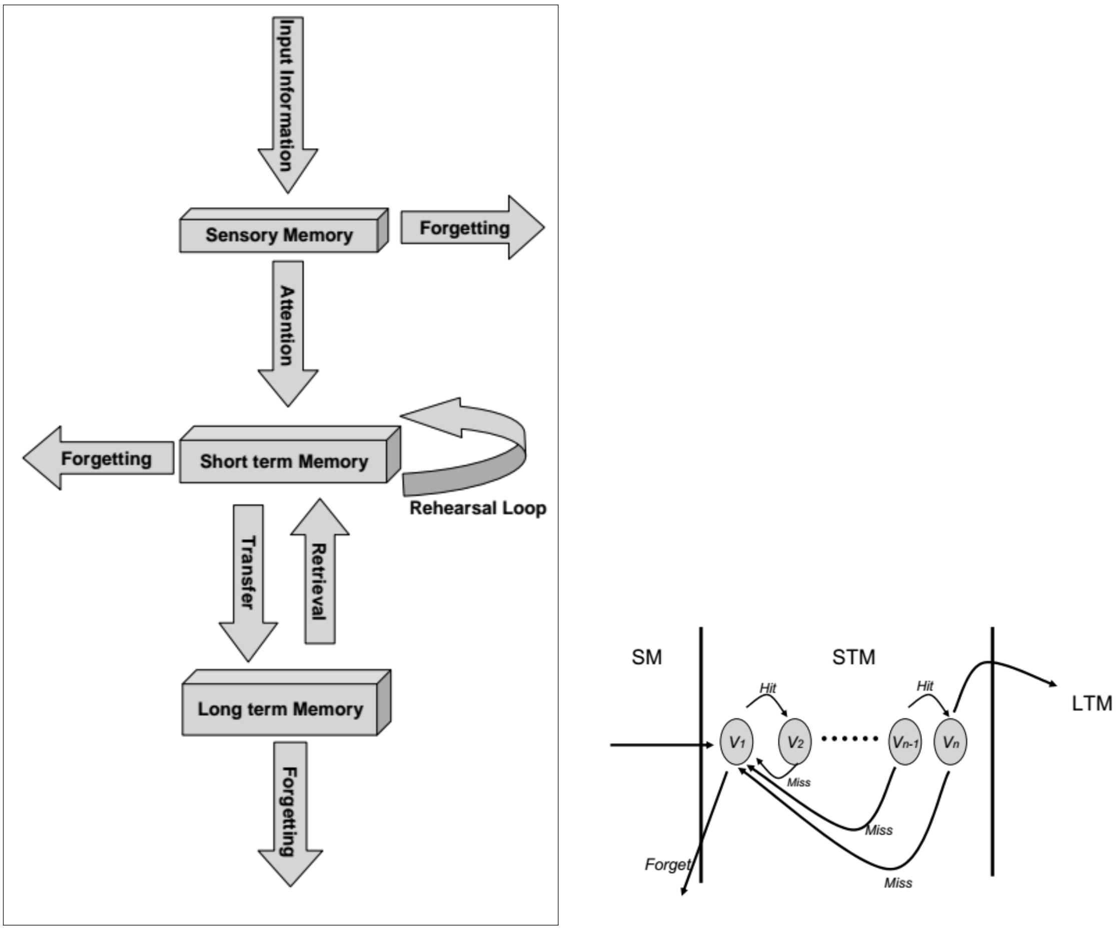


Figure 2.4: Atkinson and Shiffrin human memory model
Source: (Atkinson & Shiffrin 1968b)

Dayoub et al. (2008, 2010) used the concepts of the multi-store model to control a map size and keep it up to date to the environmental changes of the appearance location, by updating the features of this location. Dayoub et al. designed a memory management which split the memory into three parts SM, STM and LTM as shown in Figure 2.5 (a). The SM which stores the extracted SURF features of a panoramic image, whereas STM is used as an intermediate storage which tracks stable features during the localisation. Figure 2.5 (b) shows the rehearsal process which is based on a finite state machine for tracking the stable features. After that, the stable features are transferred to LTM. The stored features in LTM are used in LCD and any feature out-of-date is deleted from the map. Dayoub et al. (2011) change the map to a hybrid topo-metric map which uses the spherical representation of the feature points extracted from the panoramic images.

However, this approach is consuming a lot of time where the features stored in the LTM are continuously recalled for the rehearsal process in a lifetime. Additionally, the used finite state machine is very sensitive; once the current image features are mismatched with the current node, the remaining features in the first state will be deleted.

Bacca et al. (2010, 2011) proposed a topological VSLAM approach using a



(a) Memory management structure (b) Feature tracking finite state machine

Figure 2.5: Dayoub et al. memory management structure
Source: (Dayoub et al. 2010)

memory management similar to that used by Dayoub & Duckett (2008) as Figure 2.6 shows. In contrast with the rehearsal finite state machine, Bacca built a Feature Stability Histogram (FSH) using a voting scheme to evaluate the features that are transfer to the LTM or to delete them from the map, and with FSH the mismatched features have three-time chances before they are removed from the map. Also, both features in STM and LTM are used for localization. Bacca et al. (2013) updated his approach and used the k-means to overcome the constant threshold which is used to clustered the features between STM and LTM.

Regarding the multi-store model proposed by Atkinson & Shiffrin (1968a), another famous nature-inspired memory model is proposed by Baddeley & Hitch (1974). Baddeley and Hitch memory model is designed based on the argument that the

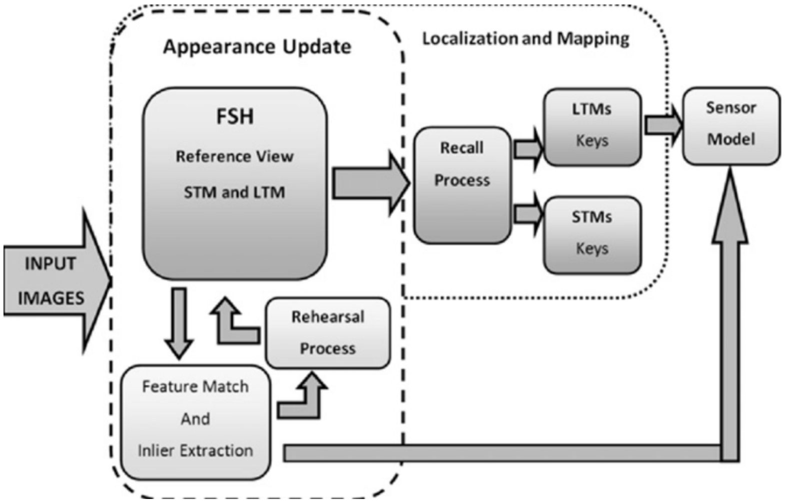


Figure 2.6: Bacca memory management with the FSH rehearsal
Source: (Bacca et al. 2010)

human mind can remember the information even if it is not rehearsed.

The models of Atkinson & Shiffrin (1968a) and Baddeley & Hitch (1974) have inspired Mathieu Labbé to introduce his VSLAM approach; namely Real-Time Appearance-Based Mapping (RTAB-Map) (Labbé & Michaud 2011; Labbe & Michaud 2013). RTAB-Map is an integration of the loop closure detection with the multi-store memory model for an autonomous navigation robot. RTAB-Map constructs a topological map to represent the surrounding environment. The mapping starts with an empty memory, and the loop closure detection tries to link a captured location image with a previously visited location. RTAB-Map uses SURF for a feature descriptor and evaluates the loop closure using the Bayesian filter. The RTAB-Map memory model scheme is inspired by human memory model where a human mind can quickly remember the recent and often visited locations.

Figure 2.7 shows the RTAB-Map processes of handling memory management for LCD. The multiple storage memory model which is used in RTAB-Map can actually divide the memory into four parts:

1. Sensory Memory (SM): After the robot captures the image, the SM extracts the visual features and constructs the visual location descriptor "signature", and examines the signature according to the pre-set threshold T_{bad}; otherwise, the

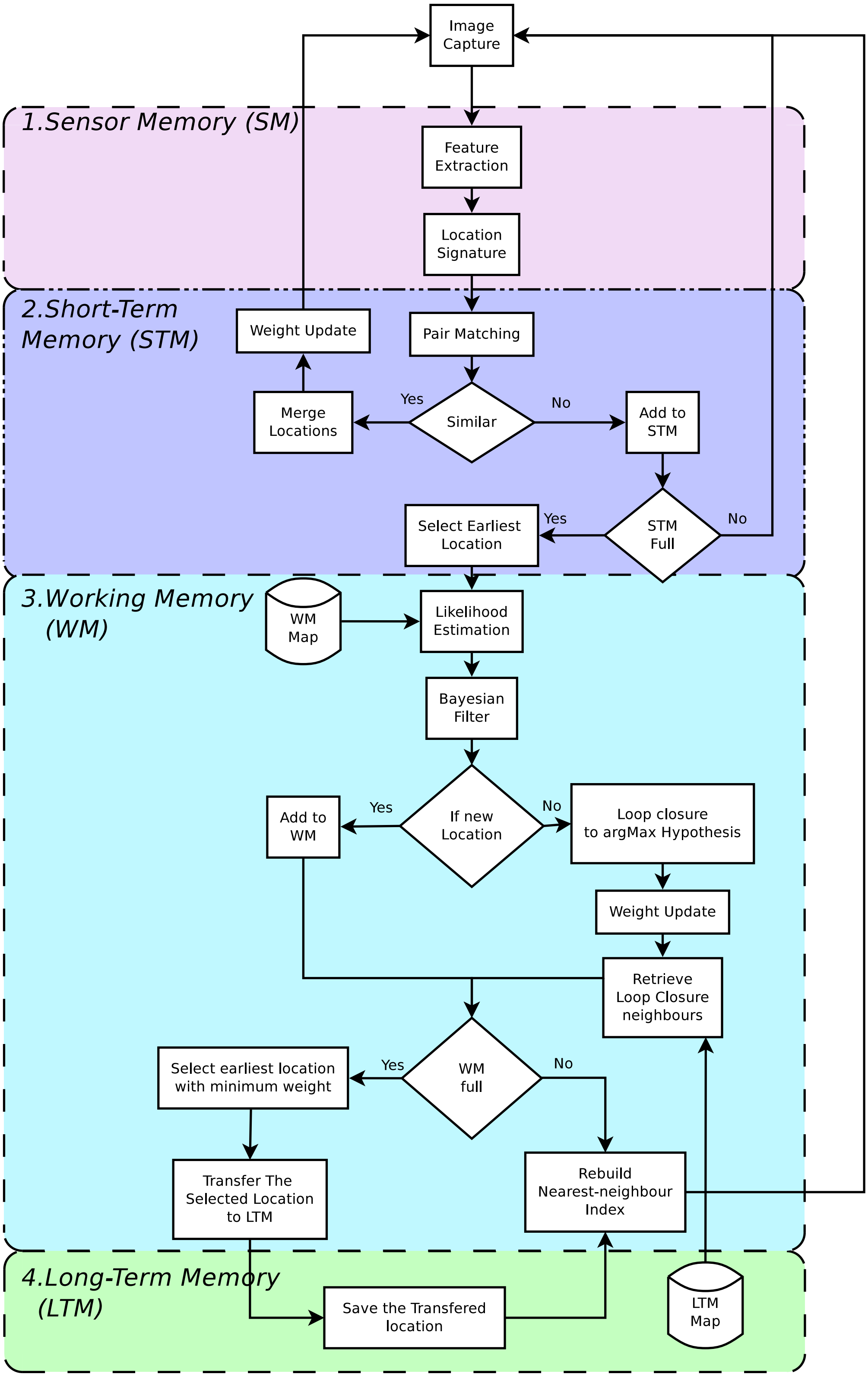


Figure 2.7: Real-Time Appearance-Based Mapping (RTAB-Map) flowchart combining visual loop closure with multiple storage memory model.

Source: (Labbé & Michaud 2011; Labbe & Michaud 2013) image will be ignored because the signature is considered unacceptable. Then, the location with the signature will be ready to pass to the next memory part.

2. Short Term Memory (STM): The STM is a fixed sized stack memory that functions on First-In-First-Out (FIFO). The main task for the STM is to identify the captured image location if it is similar to any of the previous locations in time sequence. In case the robot stops moving and is still capturing images, these images will represent the same location which are unnecessary. The criteria of similarity are according to the signature of the pair matching function. If two locations highly match, they will merge into one location, and will increment the weight for the reference location, i.e. the weight update is counting how many times the robot visited the location; otherwise, the current location will be added to the STM. When STM reaches the maximum storage size, the earliest location in the stack will automatically move to the next memory.
3. Working Memory (WM): The WM is the basic memory part of the LCD. WM will save the recent and most frequently visited locations which will be used for detecting the loop closures. WM starts by building the hypotheses based on the likelihood estimation between the current image location and the locations available in WM. Discrete Bayesian Filters (DBF) is used to calculate the posterior probability of the hypotheses. The current location can be identified based on either of the following conditions: (1) if the probability of the hypothesis representing the new location agrees with the threshold T_{loop} the current location will be identified as a new location and will, therefore, be added to the WM. Or, (2) the current image location will be identified as a loop closure to the location which gets the highest hypothesis probability. After that, the loop closure relation is created, the weight of the reference loop closure location will be incremented accordingly. Additionally, the loop closure location neighbours will be retrieved from LTM to WM and there will be a strong likelihood for the retrieved locations to be the next loop closure. To keep a consistent LCD operation time that agrees with the real time operation, the earliest location having the minimum weight in WM will be transferred to LTM so as to maintain the required maximum size of the WM functioning properly.

The last function of the WM is to rebuild the nearest neighbourhood indexing which will update the visual location signatures available in the WM.

4. Long Term Memory (LTM): That last memory part in RTAB-Map is the LTM. LTM stores the transferred locations from the WM according to preset criteria and these locations will be in the passive memory which will not be used for the loop closure detection.

To sum up, a lot of VSLAM approaches are put forward in order to overcome the challenges which face VSLAM with a growing map and operate for a long time. In Dayoub work (Dayoub & Duckett 2008; Dayoub et al. 2010, 2011) and Bacca et al. work (Bacca et al. 2010, 2011; Bacca Cortés 2012; Bacca et al. 2013) did not show the influence of the using a memory management model on the LCD performance which influence on the performance of VSLAM specials with a growing map. In contrast, Labbe and Michaud work (Labbé & Michaud 2011; Labbe & Michaud 2013) introduce the RTAB-Map which improve the LCD performance by limiting the number of matching exclusively to the locations available in WM. However, some relevant locations or data are transferred to LTM based on the pre-set criteria which makes the neighbourhood locations split into two different memory areas. The Bayesian filter will not be able to correlate them at the same time to estimate the loop closure hypotheses probability. For further information about Bayesian filter, see Section 2.6.4. Memory management have a significant improve on the performance of the LCD as it handles the growing map where LCD needs to compare the current location features with all features available on a map.

In contrast with the VO which requires matching the current location features with the previous location features to estimate the camera pose. The pose gives the metric data which are essential for autonomous navigation, and the Trajectory Estimation (TE) is the process of assembling the poses to construct the robot's trajectory. The following section discusses the common approaches of the VOTE.

2.5 VISUAL ODOMETRY TRAJECTORY ESTIMATION

In autonomous navigation, Visual Odometry (VO) is the process of estimating the pose of a robot depending on a visual sensor attached to it. The pose is the position and orientation of a robot at a certain location which is determined by analysing the corresponding image location. The combination of a sequence of poses over a period of time produces a trajectory which represents a robot's movement between the visited locations at that period of time (Okatani et al. 2014; Scaramuzza & Fraundorfer 2011). VO is the main source of the metric data which is used to construct a topo-metric map in VSLAM approaches. In the VOTE literature, VOTE is strongly related to Structure From Motion (SFM). The SFM constructs a 3D structure of image locations and their corresponding poses which could be disorderly and uncalibrated. In contrast, the sequentially ordered image locations is prerequisite for VOTE (Okatani et al. 2014; Özyeşil et al. 2017).

2.5.1 Visual Odometry Approaches

VO approaches can use different types of visual sensor and can be categorised into two main types: appearance-based and feature-based.

a. Visual Odometry Appearance-based Approaches

Appearance-based VO approaches use the entire appearance of the image locations or parts of it; such as the pixel intensity information, without any process depending on extracting features.

Goecke et al. (2007) registered the ground plane from the image locations used the Fourier–Mellin transform. Milford & Wyeth (2008) introduced an approach to derive the rotation and translation of a car on which a single camera is fixed. The rotation is measured first by finding the pixel intensity differences between the two image locations and then calculating the average absolute value of the difference. The translational velocity represents the car's movement speed in a perceptual space based on the rate of the image change. This rate is calculated according to the average intensity differences between the current and last image locations. This approach is applied in the Rat SLAM: a hippocampal model for simultaneous localization and mapping (RatSLAM) (Milford et al. 2004).

In the literature, most appearance-based VO approaches have some limitations in handling partial occlusion and fewer accuracy features where it uses the entire image for estimating the pose, compared to the feature-based VO approaches which use the local features to estimate the pose (Aqel et al. 2016).

b. Visual Odometry Feature-based Approaches

Contrary to the appearance-based approaches which use the pixels intensity information of the whole image or parts of it, the feature-based VO approaches use the salient and trackable extracted features to estimate the poses over a sequence of image locations, where the distinctive and repeated features are essential characteristics for accurate pose estimation.

One of the well-known VO approaches is the work of Nister et al. (2004) who presented a feature-based VO approach which estimates the pose based on 3D-to-2D PnP-RANSAC method with a minimum five-point solver to estimate the RANSAC solution (Nistér 2004). They used RANSAC in order to reject the outlier keypoint which is a false feature matched according to the re-projection error of the 3D-to-2D.

Feature-based VO approaches commonly include three main stages: 1) Keypoint detection, 2) Keypoint tracking and 3) Motion estimation. Subsection 2.5.2 discusses these stages, and the main issues can be summaries as follows (Fraundorfer & Scaramuzza 2012; Rais et al. 2017; Shi et al. 2013):

i. Keypoints Distribution

The distribution of keypoints is a significant factor for efficient image matching. However, dividing the image locations into sections is an undesirable way in configurations (Kersten & Rodehorst 2016; Yu et al. 2015). In this context, different approaches tackle the problem of keypoint distribution. Nister et al. (2006) used Harris corner with non-maxima suppression which is applied to every 5 × 5 neighbourhood pixel to determine the actual keypoints. The number of keypoint detection is limited according to local density rather than using the threshold of the global corner response. Additionally, keypoints are detected in buckets of 10 × 10.

Achtelik et al. (2009) present a refining keypoint method in order to minimise the cost of computation and improve the keypoint tracking. This method computes the distance among all potential pair-keypoints. Then, it eliminates the keypoint with the minimum Harris corner response with a distance value below the pre-set threshold.

Chen & Chiang (2015) divided the image location into non-overlapping rectangles in order to decrease the number of keypoints and distribute these keypoints in uniform order.

ii. Number of keypoints

The second factor that can influence the VO performance is the number of keypoint (Strasdat et al. 2012). The number of keypoints must be carefully selected, where RANSAC picks up the sample randomly. The degree of variation of outputs to the same inputs is affected by the number of keypoint.

Shi et al. (2013) present a method to reduce the number of keypoints which are used by RANSAC algorithm. They eliminate the keypoints that do not belong to the target area and the keypoints which have a cross match points.

It can be concluded that the number of points and their distribution associated with the correct matched keypoint can improve the VO performance.

2.5.2 The VOTE processes flow

Figure 2.8 shows a general flowchart of the VOTE approach designed and categorised by the researcher according to Scaramuzza's description of the VOTE processes (Scaramuzza & Fraundorfer 2011) as follows:

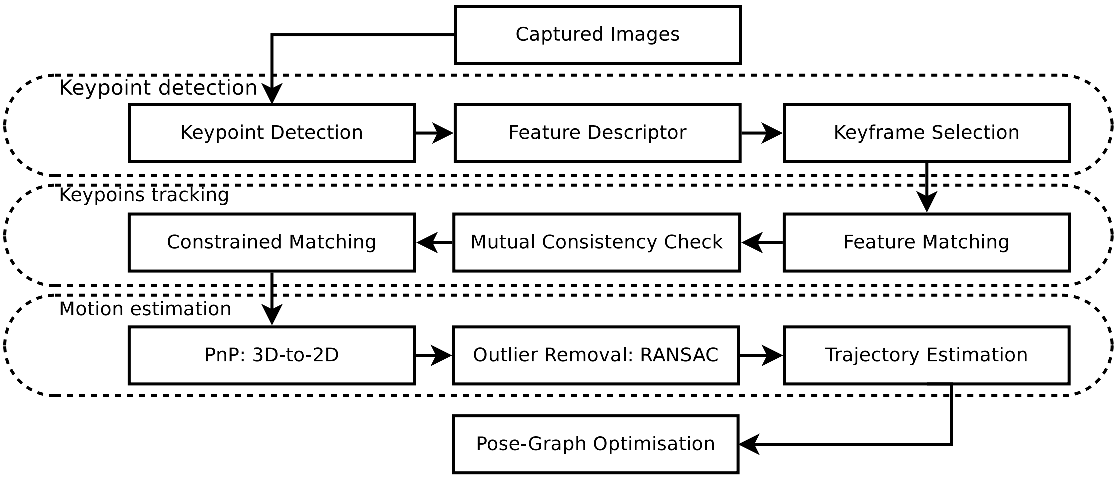


Figure 2.8: The VOTE processes flow

Source: (Scaramuzza & Fraundorfer 2011)

a. Keypoint detection

This stage focuses on the detection of keypoints and computes the feature descriptors for these keypoints. Additionally, it selects the keyframe which is used as a reference for estimating the pose of the image location. The main phases of this stage are described as follows:

i. Keypoint detections

The first phase is to detect keypoints from the current image location. These keypoints are expected to pair with their corresponding keypoints extracted from the previous image locations. This phase creates a substantial reason for the success of the VOTE, where the keypoint features need to be robust and be able to find the same keypoint features in the previous images regardless of any differences in the image scaling, rotation, or variety in illumination. Keypoint detection has a primary influence on the accuracy of estimation, the calibration matrix and the fundamental matrix which is used for estimating the camera poses (Govender 2009; SHI et al. 2016).

ii. Feature Descriptor

The second phase is the feature description of the region around each keypoint to distinguish between other keypoints and match with their corresponding features in the other images.

iii. Keyframe Selection

The third phase, the trajectory estimation depends on selecting keyframes which serve as a reference for extracting the 3D keypoints.

In the literature, the common rules for retaining or discarding image locations rely on the process of selecting keyframes based on one of the following methods:

1. the keyframe is selected according to a pre-set threshold for the average uncertainty of the 3D points. The uncertainty of the 3D points measures the standard deviation of the distances of the triangulated 3D point from all keypoints extracted from the frame (Scaramuzza & Fraundorfer 2011).
2. Alternatively, the keyframe selected is based on the estimation of the average temporal disparity (Lee et al. 2011) or a regular interval keyframe (Nister et al. 2004).

b. Keypoint tracking

This stage focuses on matching the correspond keypoints between the current image location and the keyframe. The main phases of this stage are described as follows:

i. Feature Matching

The fourth phase, where matching occurs between the current image feature "keypoints" and their similar features in the previously visited image locations. The set of feature matching with a single feature is known as feature track.

ii. Mutual Consistency Check

The fifth phase is when a comparison is made between feature descriptors in both images, i.e., current and previous. A set of features is selected from the previous images which score the highest similarity. Nevertheless, there are some features, extracted from the current image, that exactly match more than single feature in a previous image. To clarify this ambiguity as to which match should be selected, the mutual consistency check should be used. The mutual consistency check is finding a highly similar pair of feature matching in both ways (Krešo 2014; Nister et al. 2004).

iii. Constrained Matching

The sixth phase, the system suffers from the exhaustive matching where the number of matches is quadratic in the number of features, which leads to poor performance.

A common solution is to use a rapid search approach; such as a hash table or a multi-dimensional search tree, which can rapidly find the nearest features for a query feature.

c. Motion estimation

In this stage, the pose of the current location is estimated depending on the matched 2D keypoint with their corresponding 3D keypoints. The main phases of this stage are described as follows:

i. Perspective-n-Point (PnP)

The seventh phase, the Perspective-n-Point (PnP) problem which estimates the transformation camera pose between a sequence of image locations. This research uses the 3D-to-2D PnP method which is used for estimating the absolute pose based on a set of 3D-to-2D matching keypoints extracted from the corresponding image locations (Aufderheide et al. 2012; Davide & Friedrich 2011).

In motion estimation there are three approaches (Davide & Friedrich 2011; Nister et al. 2004): 2D-to-2D, 3D-to-2D and 3D-to-3D. According to Nister et al. comparative study (Nister et al. 2004), 2D-to-2D and 3D-to-2D approaches are more accurate in estimating the pose compared to 3D-to-3D approach because the reprojection error which is used in 3D-to-2D approach is more reliable than the 3D position error which is used in 3D-to-3D approach (Yousif et al. 2015).

Let the 3D scene point be $X = [x, y, z]^T$ in the camera reference frame and its projections 2D point in the image is $p = [u, v]^T$ measured in pixels as in Figure 2.9 shows.

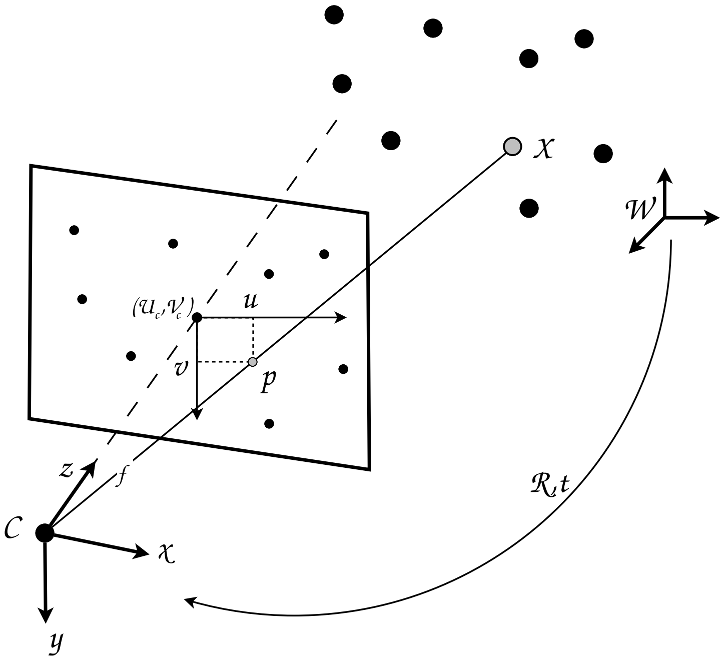


Figure 2.9: Perspective projection, C is Camera coordinate system and W is World coordinate system

Source: (Scaramuzza & Fraundorfer 2011)

The perspective projection is calculated using Equation 2.7

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K[R|t] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} f_u & 0 & u_c \\ 0 & f_v & v_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{22} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{2.7}$$

where λ is the scale "depth" factor and the camera matrix is K which are the intrinsic parameters of the camera model for the focal lengths f_u and f_v, and the principal point "image centre" (u_c, v_c), $R \in SO(3)$ is the rotation matrix and the $t \in \mathbb{R}^{3x1}$ is the translation vector. R and t are the Euclidean transformations presented by the motion of the camera.

The VO estimation pose is based on the transformation matrix $[R|t]$ which is unknown and need to be found.

The projection matrix $P = K[R|t]$ can be handled as a single unit using the Direct Linear Transform (DLT) method (Hartley & Zisserman 2003). However, this method might produce a rotation matrix which is not a valid element in $SO(3)$ due to over-parametrisation. In the literature, the outlier rejection RANSAC is used to avoid exhaustive search in the estimation step.

ii. RANSAC Outlier Removal

The eighth phase, a real environment contains fluctuating scenes where captured images are influenced by blur, noise and changes in illumination, scaling and perspectives. The extracted matching points are usually corrupted by outliers points. To the best of the researcher's knowledge, outlier points can be defined as incorrect matching of pair-keypoints or a set of pair-keypoints which are deviating from and beyond the alignment of the camera motion model.

The VOTE method is sensitive to the existence of outliers points which make it necessary to remove such outliers. The well-known outlier rejection algorithm is RANSAC (Fischler & Bolles 1981). RANSAC is used to assess camera motion parameters model by a random selection of pair keypoints sample to estimate a hypothesis of the corresponding model parameters. After that, RANSAC verifies the accuracy of this hypothesis by testing on the other pair keypoints. The hypothesis getting the highest consensus with the other pair keypoints is selected as a solution. RANSAC is repeated until it is solved with a sufficient number of inliers. Algorithm 1 clarifies the RANSAC processes.

Algorithm 1 RANSAC using 3D-to-2D method for VO

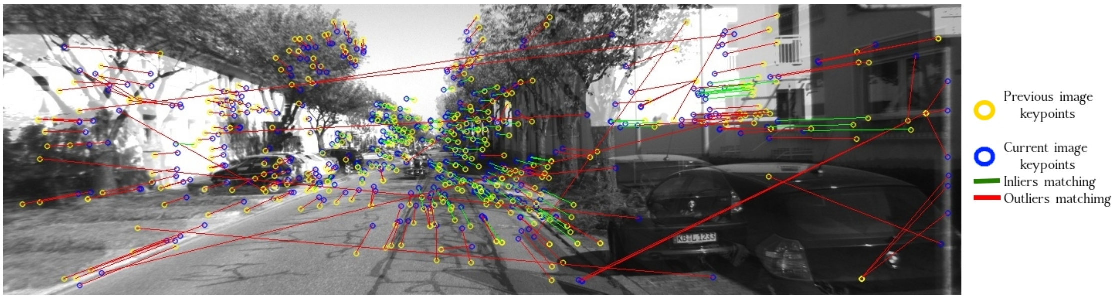
```
Initial: let P be a set of pair-matched keypoints extracted from image location $L_k$ and $L_{k-1}$
for $i = 0 <$ maximum number of iterations do
    Randomly select a sample of s points from P
    Adapt the camera model $C_i$ to these points which estimate the extrinsic parameter ($R$otation and $T$ranslation).
    Compute the reprojection error of all other points to this camera model
    let $A_i$ be a set of inlier pair-matched keypoints which their corresponding error $< d$
The camera model $C_j$ which scores the maximum inlier keypoints is chosen.
Estimate the model using all the inliers.
```

Figure 2.10 shows an example how RANSAC detect the inliers and the outliers keypoints matching which is used for the estimation of the camera pose. In this case, SURF is used for keypoint detection where 3176 keypoints were identified from the previous image and 3415 keypoints from the current image. Fast Library for Approximate Nearest Neighbours (FLANN) succeeds in matching 335 keypoints between the two images. After applying RANSAC, 46% of the keypoints are detected as inliers and are used for the pose estimation.

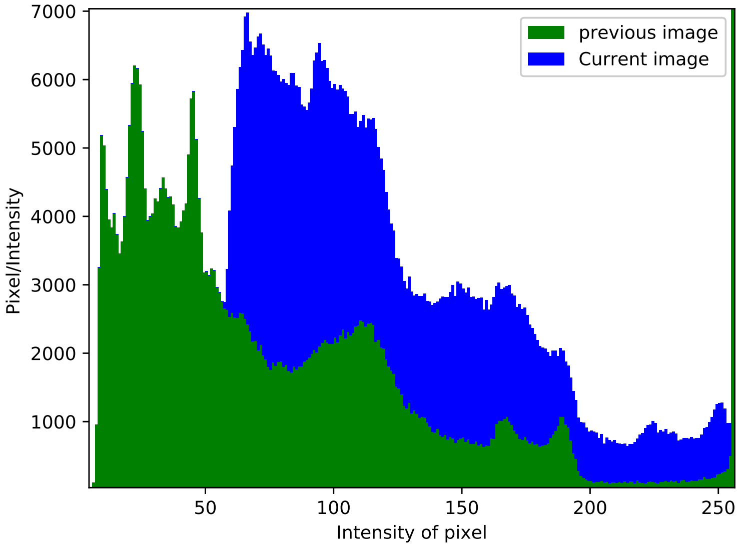
iii. Trajectory Estimation

The ninth phase, the trajectory is a concatenation of transformations between the current image location and any previously visited locations. Even the trajectory can be estimated for any sub-sequence of the visited locations.

For example, two camera poses at time $k-1$ and k are correlated by the rigid body transformation $T_{k,k-1} \in \mathbb{R}^{4\times4}$ is calculated using Equation 2.5. These transformations are used in the pose-graph optimisation to improve the trajectory estimation as additional constraints.



(a) Two images overlap with matching keypoints



(b) The histogram for the two images shows the fluctuation between the images.

Figure 2.10: Example for RANSAC in detecting the correct matching between two image locations.

d. Pose-Graph Optimisation

The tenth phase, VSLAM with VO constructs a pose graph map which is a topological map containing the transformational information lying between each location on the map and the camera poses of each location. On a topological map, nodes save the camera poses, where edges represent the constraints joining the node's poses. Each node's pose L_k receives an error from the combination of the previous node's L_{k-1} error and the transformation $T_{k,k-1}$ error, where the error accumulates.

Pose graph optimisation methods target to find the optimal camera pose parameters to keep the errors of individual transformations as small as possible which will reduce the trajectory drift. A general cost function based on the transformational constraints $e_{k,k-1}$ (Scaramuzza & Fraundorfer 2011):

$$\sum_{e_{k,k-1}} ||L_k - T_{e_{k,k-1}} L_{k-1}||^2 \tag{2.8}$$

It is noticed that the transformational constraints cost function is a linear function where the rotational cost function is nonlinear. Pose-Graph Optimization is beyond the scope of this research. Therefore, readers are advised to refer to (Scaramuzza & Fraundorfer 2011).

2.6 RELATED WORK ON VSLAM SYSTEM

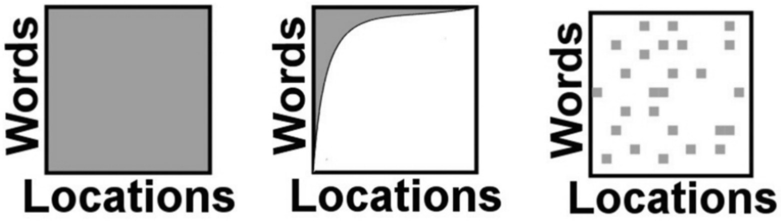
2.6.1 FAB-MAP

Fast Appearance-Based Mapping (FAB-MAP) is one of the known VSLAM approaches presented by Cummins & Newman (2007, 2008b). FAB-MAP also known as appearance-space VSLAM which uses offline BoW to a model of image locations and uses a recognition method based on a probability location calculated for each early visited locations. This approach reduces the influence of the perceptual aliasing problem by giving less probability to indistinctive images with highly similar features than the images observed from the same location.

The FAB-MAP takes up a model that is generated from the BoW data that represent a combination of visual words most likely to appear together because they come out from common features. Based on this model, the image locations are easily identified even with the least possible feature matching. At the same time, the false matching of images is declined because they represent perceptual aliasing.

FAB-MAP handles a new image location by estimating the probability of the image being matched with a previously stored location on the map, and also with the probability that the current image location is a new location. The image locations in the surrounding environment can be associated with each other depending on their probability matches. The probability matching is estimated by using the Bayesian filter which utilises a Chow Liu approximation method for computing the likelihood between the current image and all past observations.

Many developments on FAB-MAP algorithm were made to improve the efficiency of the map storage and speed up the matching process. FAB-MAP 2.0 is an improved algorithm which makes it highly capable of handling a large environment by applying a lookup scheme based on an inverted index which allowed a sparse approximation to minimise the amount of computation needed. Figure 2.11 shows how the performance influenced by application inverted index, given a new image: (a) the original FAB-MAP evaluated the likelihood method for all words and locations. (b) FAB-MAP 1.0 used a probabilistic bail-out strategy based on Bennett's Inequality to remain promising locations which are used in evaluating the likelihood (Cummins & Newman 2008a, 2010a). (c), FAB-MAP 2.0 used the fully sparse evaluation which minimises the computation cost (Cummins & Newman 2011).



(a) original FAB-MAP (b) FAB-MAP 1.0 (c) FAB-MAP 2.0

Figure 2.11: Comparison between the FAB-MAP versions: a) original FAB-MAP performance, b) FAB-MAP 1.0 performance with the bail-out strategy, and c) FAB-MAP 2.0 performance using an inverted index

Source: (Cummins & Newman 2011)

2.6.2 Incremental Topological VSLAM System

Angeli et al. (Angeli et al. 2008b) proposed VSLAM approach using single camera for constructing topological map. In the topological map, locations represented by node included the image location descriptor. The image location descriptor is generated by online-BoW which is online constructed while a robot is exploring the environment. The SIFT extracted from image locations with corresponding 128 dimension descriptor which represents the visual word. The words are compared to each other using L2 distance to construct the BoW in a tree structure. In order to reduce the operating time, Angeli et al. used the inverted index construct the BoW, where each image location in the map contains a list of its visual word and the corresponding word frequency counter. This counter tracking the frequency appeared of each word in a map and named as a score. The score is used to estimate the likelihood of the current image location and locations in a map in a simple voting scheme as Figure 2.12 shows. The likelihood one of the elements of the Bayesian filet for estimating the loop closure detection hypothesis.

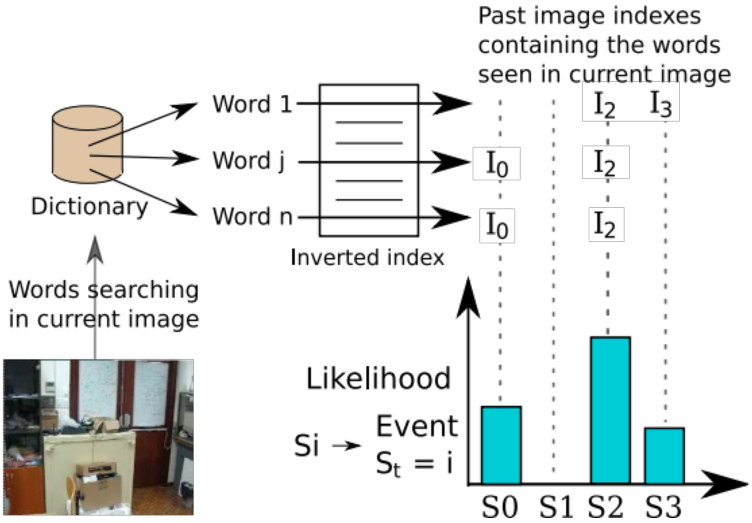


Figure 2.12: The likelihood estimation in a voting scheme proposed by Angeli et al
Source: (Angeli et al. 2008b)

This work was extended in Angeli et al. (2008c) by using two independent online-BoW which are constructed according to two different visual features SIFT and local colour histograms to achieved more distinctive image location descriptor. These BoW are used together as input for the Bayesian filter, where the likelihood output for each BoW are combined using the multiplication function as Figure 2.13 show.

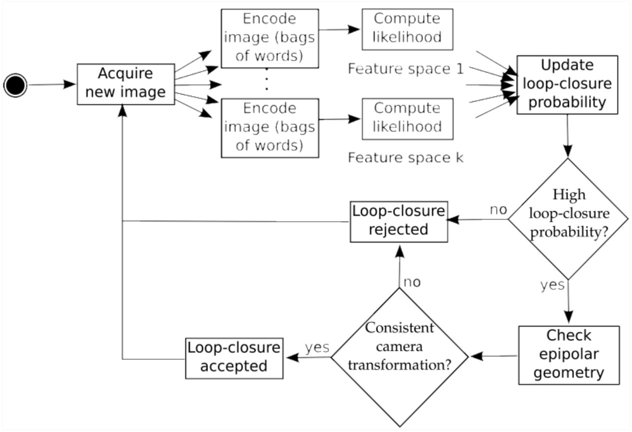


Figure 2.13: Flowchart of Angeli et al. VSLAM approach

Source: (Angeli et al. 2008c)

Additionally, Angeli et al. (2008a) present the incremental topological VSLAM system. Figure 2.14 shows the flowchart of the system.

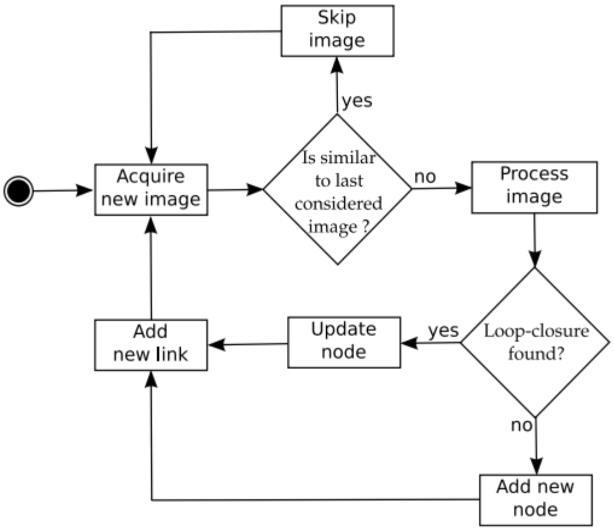


Figure 2.14: Flowchart of Angeli et al. incremental topological VSLAM

Source: (Angeli et al. 2008a)

2.6.3 PIRF-Nav2.0

Kawewong et al. (2010, 2011b) presented Position-Invariant Robust Features (PIRFs). The PIRFs generates averaged features of the SIFT descriptors which are matched with a sequence of image locations over a certain period of time. The PIRFs-dictionary represents each image location according to their PIRFs which have smaller appearance variation with respect to a robot's motion. PIRF-Nav is in an incremental appearance-based SLAM approach extract PIRFs from omnidirectional images in order to construct an incremental topological map.

These features were then used in an incremental appearance-based SLAM algorithm called PIRF-Nav, which was based on a majority voting scheme. Kawewong et al. (2011a) propose PIRF-Nav 2.0 which is an extension to his PIRF-Nav. PIRF-Nav 2.0 used the SURF feature descriptor and modified the PIRF to handled a less frequently feature indoor environment. Moreover, Kawewong et al. also introduced dictionary management which reduces the repetitive searching and maintains the memory size.

2.6.4 RTAB-Map

Labbé & Michaud (2011); Labbe & Michaud (2013) presented Real-Time Appearance-Based Mapping (RTAB-Map) which is inspired by Angeli et al. VSLAM approach (Angeli et al. 2008c). RTAB-Map combined a loop closure detection method with the multi-store memory management which maintains the memory size and keeps the LCD processing time under the real-time constraints. RTAB-Map constructs a growing topological map using a single camera. Each node in a map representing landmarks extracted from the corresponding location in the surrounding environment. The node contains a location signature which is a set of visual words that represents the extracted SURF feature. The visual words are assigned by the incremental online-BoW which have a tree structure according to Nearest Neighbour Distance Ratio (NNDR). RTAB-Map limiting number of image location which is used to construct the BoW and to keep the processing time consistent. These locations are selected according to the memory management approach which keeps the recent and most frequently visited location in the WM. RTAB-Map keeps tracking the frequently visited location and when the process time reaches the pre-set threshold, the less frequently visited location will be transferred to the LTM. RTAB-Map uses The Bayesian filter to estimate the loop closure hypothesis for each newly captured images which is compared to the

WM's locations. Subsection a. discusses the Bayesian filter.

RTAB-Map handles the problem of having a sequence of images representing the same location, these images will be detected as a loop closure because it will have a high likelihood matching and will influence the detection of the correct loop closure location. Such a situation occurs when a robot stops at a location or when the speed of the robot's movement is inconsistent with the processing speed. RTAB-Map uses the STM to verify such a situation and merge the images into a single corresponding node. Contrary to Angeli et al., who used a fixed number of al images to be excluded from the LCD.

Labbe & Michaud (2014) extended the RTAB-Map to be able to estimate the robot's trajectory and LCD over multiple mapping sessions. VOTE approach in Labbe & Michaud (2014) follows the general structure of the VOTE, which is discussed in subsection 2.5.2, where the keyframe method selects the last frame in the STM to be considered as a keyframe. This method combines the two rules of selecting the keyframes to be made as a reference for pose estimation, as discussed in subsection 2.5.2. Additionally, the feature matching between the current image location and the keyframe is estimated according to the Nearest Neighbour Distance Ratio (NNDR) method, which considers the match between the two features as a real match if the distance to the nearest neighbour is less than T_{NNDR} times the distance to the second-nearest neighbour (Labbe & Michaud 2013).

Regarding the constrained matching, RTAB-Map uses a randomized forest of $4k - d\ trees$ with FLANN (Muja & Lowe 2009). This tree structure contains the previous features, which improve the performance of the nearest neighbour search. When a new feature is extracted from a current image, the FLANN method finds the two nearest neighbours in the $k - d\ trees$ and verifies the NNDR criteria to select the corresponding features.

a. Bayesian Filter

A Bayesian filter is used in loop closure detection to estimate the similarities between the current location and the locations that have been visited. A decision is then taken to register the current location as a new location or as a loop closure to a previous location.

The Bayesian filter tries to estimate the full posterior probability $p(S_t|L^t)$ Equation 2.9 for the current location L_t at time t, where S_t is a set of all the loop closure candidates for the location L_t. If the locations L_t and L_i are representing the same location, where $i \in [0, ..., n]$ and n is the number of locations in the map, then the probability of the loop closure between L_t and L_i will be $S_t = i$. On the other hand, if the location L_t is a new location, then $S_t = -1$.

$$p(S_t|L^t) = \eta \underbrace{p(L_t|S_t)}_{Observation} \underbrace{\sum_{i=-1}^{t_n} \underbrace{p(S_t|S_{t-1} = i)}_{Transition} p(S_{t-1} = i|L^{t-1})}_{Belief} \tag{2.9}$$

where η is a normalization term and L^t is for locations in WM (the search space of the Bayesian filter), then $L^t = L_{-1}, ..., L_t$.

The observation is a likelihood function, $\mathfrak{L}(S_t|L_t)$ that is used to evaluate $p(L_t|S_t)$ using Equation 2.10

$$p(L_t|S_t = j) = \mathfrak{L}(S_t = j|L_t) = \begin{cases} \frac{\mu}{\sigma} + 1 & , j = -1 \\ \frac{\mathfrak{s}_j - \sigma}{\mu} & , j > -1 \ \& \ \mathfrak{s}_j \geq \mu + \sigma \\ 1 & , otherwise \end{cases} \tag{2.10}$$

where $\mathfrak{s}_j$ is the likelihood matches using the pair matching Equation (2.11) for L_t to all locations, $S_t = j$, where $j = -1, ..., t_n$, and the standard deviation σ is normalized by the mean μ for the non-null $\mathfrak{s}_j$.

The second part is *the belief* part, where *the transition* model is combined with the recursive part of the filter. The transition model, $p(S_t|S_{t-1} = i)$ is measured by the probability of the transition from one state $S_{t-1} = i$ to every possible state S_t as shown in Table 2.1 (Angeli et al. 2008c).

Table 2.1: The transition model declaration

Current location L_t	**Previous location** L_{t-1}	**probability** $p(S_t|S_{t-1} = i)$
New location	New location	$p(S_t = -1|S_{t-1} = -1) = 0.9$
New location	Loop closure to L_j	$p(S_t = -1|S_{t-1} = j) = 0.1$
Loop closure to L_i	New location	$p(S_t = i|S_{t-1} = -1) = 0.1/No.locations$
Loop closure to L_i	Loop closure to L_j	$p(S_t = i|S_{t-1} = j)$ Probability is defined as a discretized Gaussian curve ($\sigma = 1.6$) centered on j and the value is set recursively by starting from $i = j$ to the end of the neighborhood range

Source: (Labbe & Michaud 2013)

i. Location Matching:

The similarity measurement, $\mathfrak{s}$ is used to find the match between locations in the features level using Equation 2.11

$$\mathfrak{s}(z_a, z_b) = N_{pair}/max(N_{z_a}, N_{z_b}) \tag{2.11}$$

where N_{pair} is the number of matched word pairs between the signatures z_a and z_b, $a, b \in [0, .., n]$, N_{z_a} and N_{z_b} are the total number of words of the signatures z_a and z_b, respectively. If $\mathfrak{s}(z_a, z_b)$ passes the similarity threshold, then the two signatures z_a and z_b are matching.

Kejriwal et al. (2016) present a topological VSLAM approach. The SURF features are extracted from image locations, and two BoWs are constructed. The first BoW is constructed according to a set of visual words as in Labbe & Michaud (2013). The second BoW is a multimap data structure (Bergin 1998) constructed according to a set of visual word pairs based on their spatial proximity direction, and this BoW method is named as Bag of Word Pairs (BoWP). Accordingly, each image location is represented by the two BoWs, and the likelihood is estimated individually for each BoW similar to

Labbe's likelihood method (Labbe & Michaud 2013). The results of the likelihood are combined using the multiplication operation as work of Angeli et al. (2008c).

Hua & Tan (2017) present a VSLAM approach adopted on the RTAB-Map, where they eliminate some visual words to improve the likelihood estimation. These eliminated words do not apply to the geometric constraints according to the affine-invariant. Also, they estimate the affine-invariant hypothesis and combine it with the dispersion of the words in order to estimate the likelihood accurately.

2.7 DATASETS

Public datasets are used to evaluate the capabilities and robustness of the proposed algorithms under various conditions, including indoor and outdoor environments.

The four public datasets: Lip6 Indoor, Lip6 Outdoor, City Centre and KITTI have been utilised in this research. These datasets are widely used to evaluate the VSLAM algorithms which are in themselves considered challenges since they involve highly similar image features in the indoor environment, and the outdoor environment contains the effect of fluctuating scenes. Another reason to choosing these datasets is the frame rate of the image capture which is compatible with the proposed algorithms. Table 2.2 summarises these datasets, and the following subsections describe the main features and the challenges of each dataset.

2.7.1 Lip6 Dataset

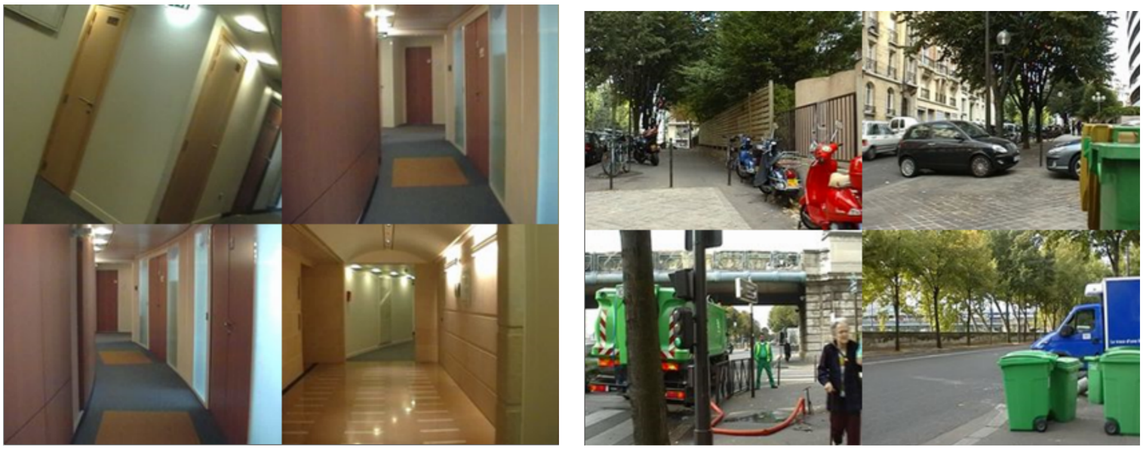
Lip6 Indoor is the first part of the dataset that is collected by Angeli et al. (2008c), and is used to validate his work in LCD. Lip6 Indoor dataset contains 388 images of two loops in medium-sized corridors with corners. The image size of 240 × 192 which is captured at 1 Hz using a single-monocular hand-held camera with a 60° field of view. The illumination is constant under artificial lighting conditions, Figure 2.15 (a) shows sample images from this dataset. Some of the challenges in the dataset are that they

Table 2.2: The public datasets used in this research. Env. = environment and the distance is unknown for Lip6 Indoor dataset

Dataset (ref)	#Images	Image size (px)	Dist(Km)	Env.
Lip6 Indoor (Angeli et al. 2008c)	388	240 × 192	-	Indoor
Lip6 Outdoor (Angeli et al. 2008c)	1063	240 × 192	1.41	Outdoor
City Center (Cummins & Newman 2008b)	1237	1280 × 480	2.01	Outdoor
KITT 00 (Geiger et al. 2012)	4541	1241 × 370	3.73	Outdoor
KITT 02 (Geiger et al. 2012)	4661	1226 × 370	3.89	Outdoor
KITT 05 (Geiger et al. 2012)	2761	1226 × 370	2.22	Outdoor

contain high perceptual aliasing and perspective transformation; such as the changes in scale and the point of view.

The second part of the Lip6 dataset of Angeli et al. (2008c) is the Lip6 Outdoor. The dataset contains 1063 images that complete a large loop outdoors about 1.4 km in Paris city street. The images are captured at 1 Hz of image size 240 × 192. The images are captured in sunny weather conditions, and the trip was short taking around 20 minutes to experience changes in lighting (Angeli et al. 2008c), Figure 2.15 (b) shows sample images from this dataset. Angeli et al. provide the ground truth maps of the loop closure locations of each part of the dataset, which were manually signed by themselves.



(a) Indoor

(b) Outdoor

Figure 2.15: Samples of Lip6 Indoor Outdoor dataset

Source: (Angeli et al. 2008c)

2.7.2 City Centre Dataset

City Centre dataset is a subset of the Oxford dataset. This dataset is used to evaluate the FAB-MAP (Cummins & Newman 2008b). The City Centre dataset contains 2474 images with a size of 640×480 acquired from two cameras (left and right). The images are captured outdoors along 2 km in public roads at a rate of 0.5 Hz. The images include dynamic objects which are taken on a windy day in bright sunshine that show the foliage and shadows in dynamic motion. To evaluate the performance of the proposed LCD algorithms which requires a single image; the two images, left and right, have been concatenated into a single image of a new size of 1280×480, and the total number of sequential images is 1237 of two rounds. The authors provide the ground truth map of the loop closure locations which was manually signed by themselves (Cummins & Newman 2008b). Figure 2.16 shows samples of images taken from the dataset.

Figure 2.16: Samples of City Centre dataset

Source: (Cummins & Newman 2008b)

2.7.3 KITTI Dataset

Vision Benchmark Suite from Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) contains different types of datasets for an outdoor environment like stereo images dataset, optical flow and visual odometry. KITTI uses a vehicle that is set up with a localisation system, 360° Velodyne laser scanner and stereo cameras rig as shown in Figure 2.17. KITTI recorded 22 outdoor sequences, 11 sequences provided with ground truth to evaluate the performance of the trajectory estimation.

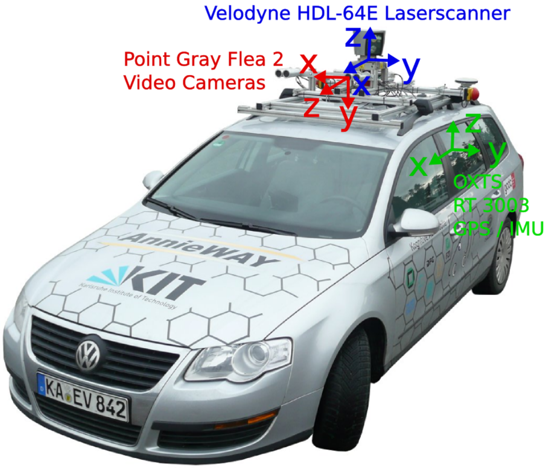


Figure 2.17: The car which is used in collecting the KITTI dataset

Source: (Geiger et al. 2013)

In this research, the visual odometry dataset of sequences 00, 02, 05 is used (Geiger et al. 2013). The KITTI dataset provides a tool which is used to evaluate the accuracy of the estimated trajectory. This tool is also used in this research to measure the performance of the proposed algorithm. Figure 2.18 shows samples of KITTI dataset.

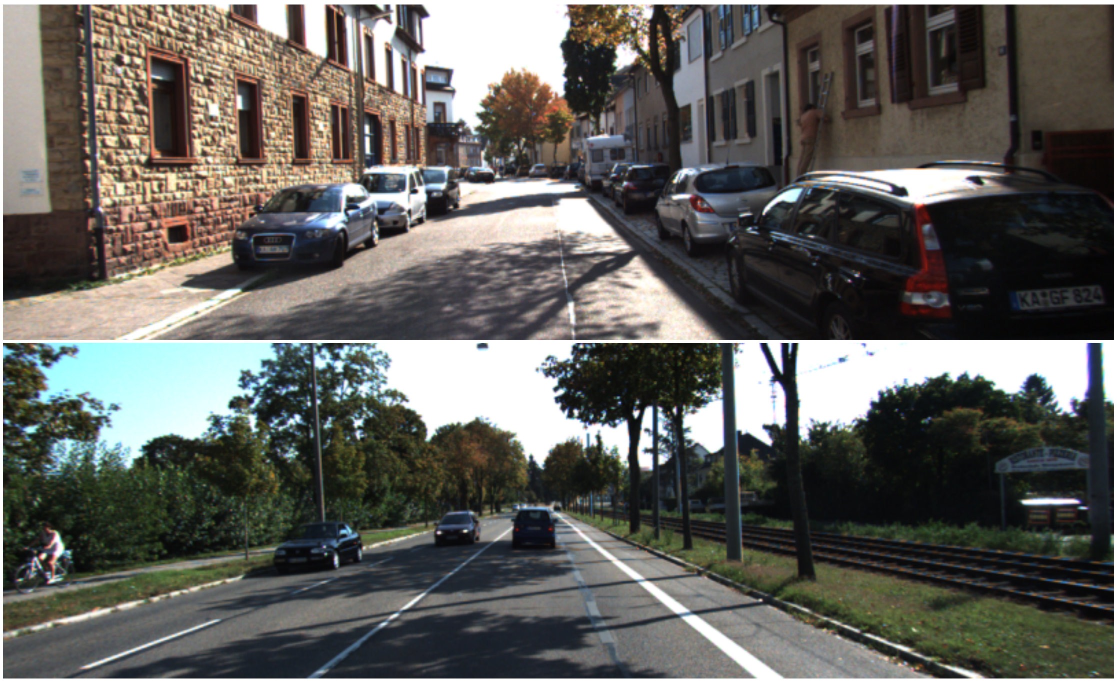


Figure 2.18: Samples of KITTI dataset

Source: (Geiger et al. 2013)

While KITTI dataset does not provide a ground truth map for loop closure locations, and the ground truth which is provided by Arroyo et al. (2014) for the KITTI dataset is not consistent.

2.8 SUMMARY

This chapter demonstrates the literature review related to the subjects of LCD and VOTE, and the methods adopted in connection with this subject. This review focusses on the well-known authors in the VSLAM domain who tackled this subject and their contribution to handling the challenges related to VSLAM.

First, it starts with discussing the VSLAM components and the challenges relating to VSLAM. The LCD and its different approaches have been discussed and analysed according to the visual feature descriptors. The literature review brings up the memory management approaches adopted in VSLAM to handle the growing map challenges.

Second, the VOTE approaches and their processes are presented and explained the difference between these approaches, which are the appearance-based and feature-based approaches. Additionally, the process of the feature-based approach is discussed in some detail.

Third, the VSLAM-related systems are discussed and followed by a description of the three public datasets; namely, Lip6 Indoor, Lip6 Outdoor and City Centre which are used to evaluate the LCD proposed algorithms and the KITTI dataset to evaluate the VOTE proposed algorithm.

CHAPTER III

METHODOLOGY

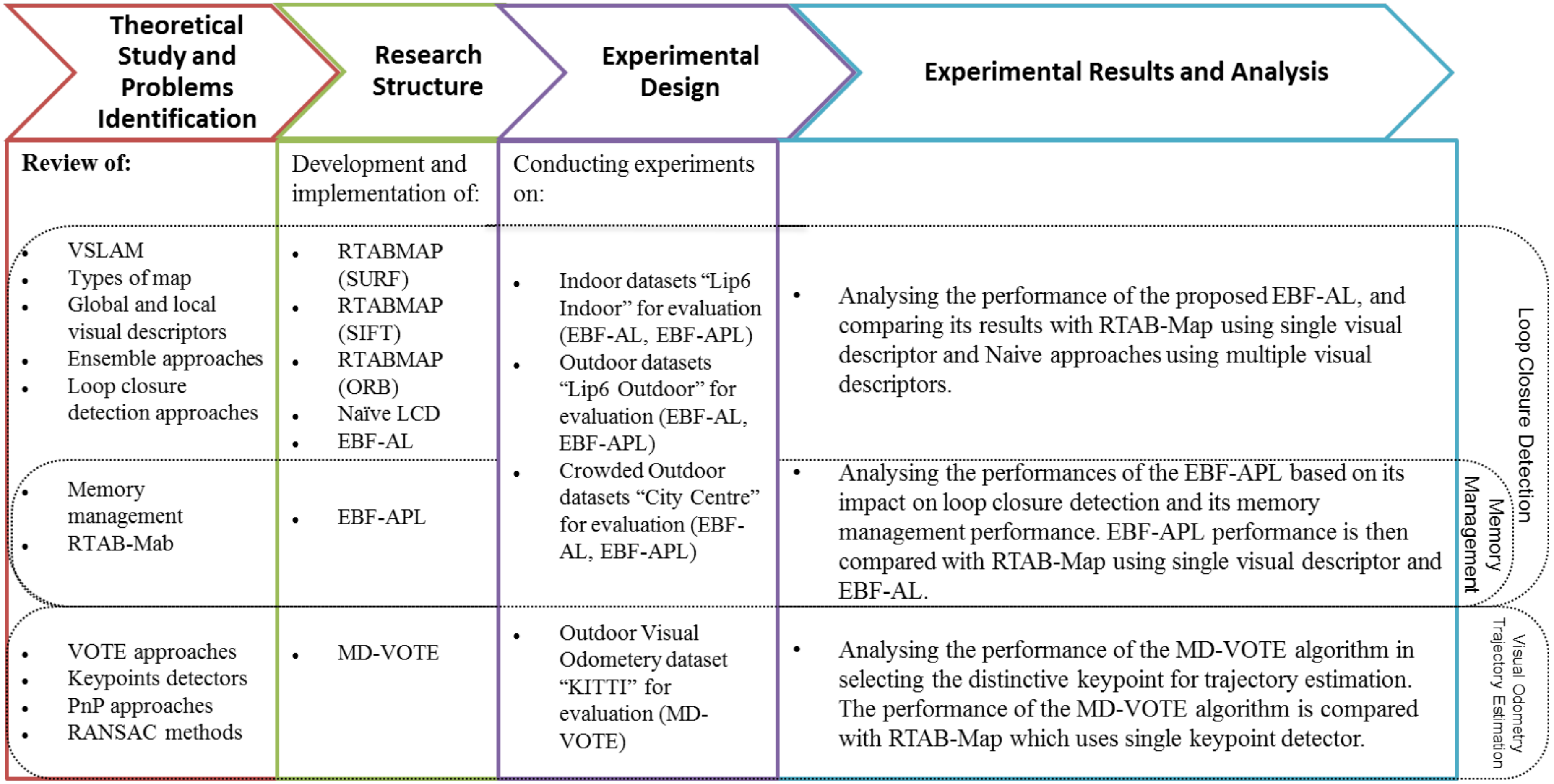
3.1 INTRODUCTION

This chapter summarises the research methodology followed in this research. The methodology is composed of four main phases and Figure 3.1 shows the summary of the four phases of the research methodology. First, the theoretical phase which reviews the problems and challenges relating to this research from the points of view of the literature. Second, the research structure illustrates the design and implementation of the proposed methods. Third, the experimental design phase discusses the several experiments carried out as part of the study. The experiments are conducted on two types of dataset: (1) the loop closure detection datasets, namely Lip6 Indoor, Lip6 Outdoor and City Center, and (2) the trajectory estimation dataset, namely KITTI. Fourth, the results of the evaluation and analysis of the proposed algorithms are provided herein and are compared with state-of-the-art methods. A summary of this chapter is provided at the end.

3.2 RESEARCH METHODOLOGY

The principal objective of this research is to improve the accuracy of the VSLAM approach to be capable of handling long time operation and producing high-quality maps including distinctive landmarks that are capable of accurately localising the robot's locations "loop closure detection" with enhanced memory management scheme and estimating a robot's trajectory in unknown environments.

To achieve the objectives of this research, and to answer the research questions,



Theoretical Study and Problems Identification	Research Structure	Experimental Design	Experimental Results and Analysis	
Review of:	Development and implementation of:	Conducting experiments on:		
• VSLAM • Types of map • Global and local visual descriptors • Ensemble approaches • Loop closure detection approaches	• RTABMAP (SURF) • RTABMAP (SIFT) • RTABMAP (ORB) • Naïve LCD • EBF-AL	• Indoor datasets “Lip6 Indoor” for evaluation (EBF-AL, EBF-APL) • Outdoor datasets “Lip6 Outdoor” for evaluation (EBF-AL, EBF-APL) • Crowded Outdoor datasets “City Centre” for evaluation (EBF-AL, EBF-APL)	• Analysing the performance of the proposed EBF-AL, and comparing its results with RTAB-Map using single visual descriptor and Naive approaches using multiple visual descriptors.	Loop Closure Detection
• Memory management • RTAB-Mab	• EBF-APL		• Analysing the performances of the EBF-APL based on its impact on loop closure detection and its memory management performance. EBF-APL performance is then compared with RTAB-Map using single visual descriptor and EBF-AL.	Memory Management
• VOTE approaches • Keypoints detectors • PnP approaches • RANSAC methods	• MD-VOTE	• Outdoor Visual Odometery dataset “KITTI” for evaluation (MD-VOTE)	• Analysing the performance of the MD-VOTE algorithm in selecting the distinctive keypoint for trajectory estimation. The performance of the MD-VOTE algorithm is compared with RTAB-Map which uses single keypoint detector.	Visual Odometry Trajectory Estimation

Figure 3.1: The Research Methodology

an experimental methodology has been adopted with four phases. Figure 3.2 shows the flowchart of the research methodology.

3.2.1 Theoretical Study and Problems Identification

The first phase provides a review of the previous research in the VSLAM where it focuses on two subjects of interest: LCD and VOTE.

For the LCD, this phase includes reviewing the visual landmark descriptors used in VSLAM which is based on the existing literature. The review of the previous works focuses on the descriptor types and methods. Furthermore, a review is conducted on the visual descriptor combination approach for LCD and the aggregation strategies.

Also, a review of recent works was carried out and focused on the appearance methods of loop closure detection and how those methods handled their memory. In addition to that, several experiments were made for monitoring the behaviour of the current memory management scheme for LCD. These reviews supported by experiments help to identify the challenges and limitations of the current memory management scheme for LCD. Also, RTAB-Map is carefully studied and reviewed because it is the state-of-the-art memory management algorithm for appearance-based loop closure detection. The following subjects have been focused upon while reviewing RTAB-Map: memory models, memory partitions, the pre-set conditions and thresholds, and the data-flow structure between the memory partitions. Practically, RTAB-Map is selected and integrated with the proposed algorithms.

The second subject of interest is Visual Odometry Trajectory Estimation. The recent literature on Visual Odometry methods was reviewed as it focused on the Trajectory Estimation processes and techniques in VSLAM. The review of the VOTE illustrates the general structure and the main improvements in the methods.

This phase clarifies the literature review of LCD in the memory management scheme and the VOTE. The relating problems and challenges in this matter are identified to form the theoretical and experimental observations that encourage the research for