ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

EFFICIENT VISUAL LOOP CLOSURE DETECTION VIA LOCALIZED MOMENT DESCRIPTORS

M.Sc. THESIS

Can ERHAN

Department of Mechatronics Engineering

Mechatronics Engineering Programme

JULY 2016

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

EFFICIENT VISUAL LOOP CLOSURE DETECTION VIA LOCALIZED MOMENT DESCRIPTORS

M.Sc. THESIS

Can ERHAN
(518121005)

Department of Mechatronics Engineering

Mechatronics Engineering Programme

Thesis Advisor: Prof. Dr. Hakan TEMELTAŞ

JULY 2016

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

HIZLI VE VERİMLİ ÇALIŞAN YERELLEŞTİRİLMİŞ GÖRSEL MOMENT TANIMLAYICILARIYLA ÇEVRİM KAPAMALARIN SAPTANMASI

YÜKSEK LİSANS TEZİ

Can ERHAN
(518121005)

Mekatronik Mühendisliği Anabilim Dalı

Mekatronik Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Hakan TEMELTAŞ

HAZİRAN 2016

Can ERHAN, a M.Sc. student of ITU Graduate School of Science Engineering and Technology 518121005 successfully defended the thesis entitled "EFFICIENT VISUAL LOOP CLOSURE DETECTION VIA LOCALIZED MOMENT DESCRIPTORS", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Hakan TEMELTAŞ** Istanbul Technical University	.....
Jury Members :	**Assoc. Prof. Hatice KÖSE** Istanbul Technical University	.....
	Prof. Dr. Işıl BOZMA Boğaziçi University	.....

Date of Submission : **2 May 2016**
Date of Defense : **9 July 2016**

To my grandfather,

FOREWORD

This thesis has been completed also thanks to many persons which contributed with suggestions, thoughts, and constructive criticisms. I take therefore the occasion to briefly mention them here.

I am greatly indebted to my thesis supervisor Prof. Dr. Hakan Temeltaş, who has provided me with the opportunity to carry out research in the robotics laboratory for more than two years. His professional knowledge and constant support helped me proceed throughout my studies. I am also grateful to my colleagues and friends at the Robotics Laboratory.

Thanks very much to Evangelos Sariyanidi, who always given me a helping hand whenever needed during this project in spite of their busy timetable.

Thanks to Borda Technology, the company that I work on, for their understanding and supporting me to accomplish my thesis.

My deepest thanks, and apologies, to my flatmate Çaglar Kutlu for his support, patience and understanding during the writing of my thesis.

Last but not least, thanks a lot to my family, for always supporting me during my studies.

July 2016 Can ERHAN

TABLE OF CONTENTS

Page

ABBREVIATIONS

AUC : Area Under Curve
BoRF : Bag-of-Raw-Features
BoW : Bag-of-Words
BRIEF : Binary Robust Independent Elementary Features
CPU : Central Processing Unit
CRF : Conditional Random Fields
CZM : Complex Zernike Moments
FAB-MAP : Fast Appearance Based Mapping
HoG : Histogram of Oriented Gradients
LIDAR : Light Detection And Ranging
LDB : Local Difference Binary
LZM : Local Zernike Moments
NNS : Nearest Neighbor Search
PALM : Patterns of Approximated Localized Moments
QLZM : Quantized Local Zernike Moments
SIFT : Scale-Invariant Feature Transform
SLAM : Simultaneous Localization and Mapping
SONAR : Sound Navigation And Ranging
SURF : Speeded Up Robust Features

LIST OF TABLES

Page

LIST OF FIGURES

Page

EFFICIENT VISUAL LOOP CLOSURE DETECTION VIA LOCALIZED MOMENT DESCRIPTORS

SUMMARY

In the context of autonomous mobile robotics, constructing a map of an unknown environment and localizing the robot itself in it are essential tasks to accomplish the missions that are programmed to the robot. Although both tasks initially appear to be independent, they are closely related and considered as a single problem, known as Simultaneous Localization and Mapping (SLAM). Loop closing is defined as the correct identification of previously visited location in terms of SLAM. This ability is crucial for not only accurate localization, but also creating consistent maps by minimizing the accumulated errors arising from the sensory information. Range sensors such as LIDARs and SONARs have been utilized for many years in order to solve this problem. On the other hand, the usage of vision-based sensors have been quite popular in recent years due to their competitive prices and compact structure being able to provide rich information. When camera is the only sensor, loop closure detection is performed by comparing the images directly. Working with images brings new complications to be dealt with such as perceiving the images from different places as the same, known as perceptual aliasing.

Visual loop closure detection is still an active and challenging problem that must be handled in real-world SLAM applications. Avoiding false detections is a crucial factor, since they may cause catastrophic consequences for the general SLAM process. Hence, it is essential to use a fast and efficient algorithm with high discrimination power in order to cope with the problems arising from the visual sensory.

In this thesis, a novel visual loop closure detection method has been presented. This method relies on computing localized moment descriptors all along the image to achieve place recognition. The places are represented with their images. Therefore, the loop closure detection is performed by comparing the image descriptor of the most recent place with all the descriptors that have been extracted throughout the trajectory.

The major contribution in this thesis is the description technique which is the key point to achieve good matching results in terms of place recognition and ultimately loop closure detection. The scene descriptor is computed by extracting Zernike moment patterns in an overlapping manner. Briefly, this technique benefits from the discrimination power and robustness to perceptual aliasing of local features, and then combines them into a global or whole-image descriptor with a low computational complexity. As low-level shape features, local Zernike moments have outstanding representation capability for a scene image containing discontinuities spread onto different locations. Quantizing the moments computed all along the image enables a reliable representation by reducing the effect of image noise and background illumination. In order to achieve pose recovery, the resulting descriptor vector is constructed by concatenating the histograms that are built from the patterns at different

regions as a global image descriptor. As a further improvement, the time complexity is decreased tremendously by using integral images to operate the method in real-time without introducing any information loss at all.

The proposed loop closure detection method has been evaluated on the challenging real-world datasets publicly available. It is reported that the proposed method gives promising results with high true positive rates while avoiding false detections. In this sense, a comparison study available in this thesis. It shows that the proposed method outperforms some other state-of-the-art methods in the literature. Also, a straightforward implementation of the method is shown to perform real-time even for long sequences containing more than thousand images.

HIZLI VE VERİMLİ ÇALIŞAN YERELLEŞTİRİLMİŞ GÖRSEL MOMENT TANIMLAYICILARIYLA ÇEVRİM KAPAMALARIN SAPTANMASI

ÖZET

Otonom navigasyon, mobil robotik alanında üzerinde en çok çalışılan konulardan biridir. Eşzamanlı Konum Belirleme ve Haritalama'da (EZKH), otonom navigasyon konusu içinde en çok araştırılmış ve hala araştırılmakta olan problemlerden biridir. Ancak uzun soluklu çalışmalara rağmen, özellikle geniş ölçekli dış ortamlar baz alındığında EZKH kapsamında çözülmesi gereken birçok problem bugün hala tam olarak çözülememiş durumdadır. EZKH bağlamında çevrim kapama problemi, otonom bir robotun daha önce bulunmuş olduğu bir yeri başarıyla tanıyabilmesi olarak açıklanabilir. Çevrim kapama çalışmalarının EZKH kapsamında ayrı bir önemi vardır, çünkü başarıyla gerçekleştirilen çevrim kapamalar robotun en güncel konumunu çok daha yüksek bir hassasiyetle belirleyip, geçmiş yörüngesindeki konumları üzerindeki kestirimlerini iyileştirmesine olanak sağlar. Konum kestirmede sağlanan bu iyileştirme, haritalama başarımını da önemli ölçüde artırır. Öte yandan hatalı gerçekleştirilen çevrim kapamalar, EZKH kestirimlerindeki konum ve haritalama süreçlerinin hatalı biçimde güncellenmesine yol açacağı için, hatalı çevrim kapamaların genel EZKH sistemi üzerindeki etkisi yıkıcı boyutlara ulaşabilmektedir. Dolayısıyla hassasiyet, geliştirilen çevrim kapama sisteminde can alıcı bir öneme sahiptir.

Bir çevrim kapama sistemi tasarlanırken, dikkate alınması gereken kriterler yalnızca hassasiyet ve yüksek başarım değildir. En az bu iki kriter kadar önemli olan diğer bir kriter de sistemin hızı, ve dolayısıyla etkinliğidir. Bunun en önemli nedeni, EZKH sürecinin genellikle çevrimiçi bir süreç olması ve gerçek zamanlı işleyişin bir EZKH uygulamasında ayrı bir öneminin olmasıdır. Görüntü işleme tekniklerinin genel olarak yoğun işlem gerektiriyor olması da, etkin bir sistem tasarımını daha da güçleştirmektedir.

Çevrim kapama problemi, bu tez çalışmasında görsel algılayıcılar kullanılarak görüntü işleme teknikleriyle çözülmüştür. Görüntü işlemeye dayanan çevrim kapama problemi, temele indirgendiğinde bir görüntü eşleştirme, diğer bir deyişle görüntüler arasındaki benzerliği ölçme problemidir. Bu problem, birçok açıdan çözülmesi zor bir problemdir. Problemi zor kılan etmenler arasında en öne çıkanı, eşleştirilmeye aday görüntülerin çoğu durumda birbirine oldukça benziyor olmasıdır. EZKH probleminin dış ortamdaki olası uygulama alanları arasında çöl veya ormanlık alan gibi doğal ortamlar, veya sokak ve otoyol gibi kentsel ortamlar vardır. Bütün bu ortamlarda, birbirine benzeyen görüntülere sıklıkla rastlanabileceği için sistem kolayca yanılabilir. Diğer bir etmen ise görsel sensörlerden kaynaklanan aydınlanma etkisidir. Yoğun aydınlanmaya maruz kalan görüntülerin istenmeyen bir şekilde algılanması kaçınılmazdır. Hatalı çevrim kapamaların genel EZKH sistemindeki yıkıcı etkisi gözönüne alınırsa, bu tip benzer görüntülerde yapılabilecek olası yanlış eşleştirmelere karşı özel önlemler alınması

gerekmekte olup, çevrim kapama hipotezleri yeterince güvenilir olmadıkları sürece kesinlikle kabul edilmemelidir.

Bilgisayarla görüye dayanan teknilerin çevrim kapama probleminde kullanımı, son on yılda kayda değer ölçüde yaygınlaşmıştır. Bunun en önemli nedenlerinden biri, bilgisayar donanımı ve özellikle işlemci teknolojisindeki gelişmelerin, yoğun işlem gerektiren görüntü işleme yöntemlerinin kulanımını mümkün kılmasıdır. Diğer bir önemli etken de, çevrim kapama problemine uyarlanabilecek birçok bilgisayarla görü ve görüntü işleme tekniğinin önerilmiş olmasıdır. Kameradan önce kullanılan LIDAR gibi algılayıcılar, sözkonusu çevrim kapama problemini çözmekte kısıtlı olanaklar sunabilmişlerdir.

Bu tez çalışmasında, özgün bir çevrim kapama yöntemi sunulmaktadır. Önerilen yöntem, görüntülerin üzerinde yerelleştirilmiş momentlerin hesaplanmasına dayanmaktadır. Robotun bulunduğu her bir konum, o konuma denk gelen görüntülerle temsil edilmekte ve bu sayede, çevrim kapamalar robotun son konumundan alınan görünütü ile o zamana kadar toplanan görüntülerin birebir karşılaştırılmasıyla tespit edilmektedir. Bu tez çalışmasının en önemli katkısı, mekanların ve en nihayetinde çevrim kapamaların tespit edilmesinde en büyük etkiye sahip olan görüntülerin temsil edilme yöntemidir. Görüntü betimleyicilerinin hesaplanması üstüste binmiş şekilde hesaplanan kabaca nicelendirilmiş yerel Zernike momentlerini baz almaktadır. Geliştirilen bu teknik yerel ayrırıcı niteliklerin aydınlanma değişimlerine daha az hassas olma ve ayırıcı gücü yüksek olma özelliklerinden faydalanarak, düşük bilgi işlem yüküyle hesaplanabilen bütünsel tanımlayıcılar oluşturmaktadır. Düşük seviye nitelik belirleyiciler olarak, yerel Zernike momentleri oldukça başarılı bir temsil etme yeteneğine sahiptir. Hesaplanan yerelleştirilmiş momentlerin gelişi güzel bir şekilde nicelendirilmesi, görüntü üzerinde değişken bir şekilde dağılmış olan aydınlanma etkisini ortadan kaldırmaktadır. Robotun görüş açısının sabit olmamasından kayaklanan göreceli açı değişiklikleri, hesaplanan yerel momentlerin bölgesel olarak histogramlarla temsil edilmesiyle daha az etkili hale getirilmiştir.

Robotun üzerinde bulunan görsel sensörler yardımıyla alınan görüntüler çevresel faktörlere karşı duyarlı olan bir betimleyici olarak hesaplanmaktadır. Fakat, çevrim kapama hipotezinin gerçekleştirilebilmesi için bir görüntünün daha önce görülüp görülmediğini ortaya çıkaracak bir yöntem gerekmektedir. Bu tez çalışmasında, gelen bir görüntünün daha önce görüntülenmiş bir alanı tesmil edip etmediğini ortaya çıkarmak için, en basit sınıflandırıcılardan biri olan en yakın komşu algoritması kullanılmaktadır. Son alınan görüntü daha önceden kaydedilmiş görüntüler ile karşılaştırılarak birbirlerine en yakın görüntü çevrim kapayacak görüntü adayı olarak seçilmektedir. O görüntünün çevrim kapayıp kapamadığı ise, önceden belirlenmiş bir eşik değeri ile kıyaslanarak belirlenmektedir. Görüntülerin betimleyicileri arasındaki uzaklığı ölçmek için hali hazırda var olan ve çok hızlı çalışan bir uzaklık metriği kullanılmaktadır.

Bu çalışmada önerilen görsel yollarla çevrim kapatma yöntemi, bilinen diğer yöntemlerden kabul görmüş veritabanlarından üçü üzerinde karşılaştırılmıştır. Bu veritabanlarından biri çok sayıda ağaç, duvar ve çalıların bulunduğu yerlerde kaydedilmiş, test edilecek metodun farklı yerleri karıştırma hassasiyetini değerlendirmeye yönelik bir şekilde hazırlanmıştır. Diğeri ise kısmen güneşli ve rüzgarlı bir bölgede oluşturularak test edilecek metodun aydıntlanmaya ve hareketli cisimlerle karşı olan gürbüzlüğünü ölçmeye yöneliktir. Bu veritabanlarından sonuncusu ise, aracın hızından

dolayı oluşan geniş yer değiştirmeler ve göreceli poz değişiklikler bakımından zorlu bir veritabanıdır. Elde edilen sonuçlar, çalışmadaki yaklaşımın ve genel olarak önerilen yöntemin literatürce kabul görmüş diğer bütünsel betimleyicilerle karşılaştırılarak, önerilen metodun diğer metodlara kıyasla oldukça iyi çalıştığını göstermektedir. Ayrıca, tüm bu başarının yanında geliştirilen yöntem çok hızlı ve verimli bir şekilde gerçek zamanlı çalışabilmektedir.

1. INTRODUCTION

In the context of autonomous mobile robotics, constructing a map of an unknown environment and localizing the robot itself in it are essential tasks to accomplish the missions that are programmed to the robot. Although both tasks initially appear to be independent, they are closely related and considered as a single problem, known as SLAM. Despite the extensive studies that have been investigated for many years, SLAM still remains in an active research field, which becomes even more challenging when it is solved for outdoor environments.

1.1 Problem Statement

Loop closure detection is one of the most essential problems of the SLAM, and it is defined as the correct identification of previously visited places. Detecting loop closure events is an extremely important ability for a mobile robot in order to perform SLAM, since this ability allows generating consistent maps and reducing their uncertainty.

Avoiding false detections is a key factor must be taken into consideration for a loop closure detection system. The location estimations provided by the outcome of the SLAM process are always erroneous. Even the slightest errors arising from the sensory information grow incrementally and accumulated to the point that they cannot be handled. The wrong estimations caused by these errors are minimized or reset with correctly detecting loop closure events. In contrast, any wrong detections of loop closures may effect the SLAM system in a catastrophic way.

Traditional SLAM approaches have relied on range sensors such as LIDARs in terrestrial environments or SONARs in underwater. However, the usage of vision-based sensors have been quite popular in recent years due to their competitive prices and compact structure being able to provide rich information. When vision is the source, loop closure detection is performed by comparing the images directly.

Working with images captured from the camera brings new complications to be dealt with. Perceiving different places as the same, known as perceptual aliasing, is one of the main problems of vision-based loop closure detection systems. It is nearly impossible tackle this problem especially when the system uses only a single camera without depth information. The other problem caused by visual sensory is the effect of illumination which can be very preventative to recognize the places correctly. In addition to this, relative pose variations at the same place caused by the particular view point of the camera is another problem which prevents performing reliable matching and increases miss rate. In dynamic environments where partial occlusions from moving objects are taking place, loop closure detection becomes a hard task to be succeeded. Thus, any vision-based loop closure system must be robust to these conditions presented on the environment.

Overcoming the problems revealed by visual sensory is not the only factor for developing a successful loop closure detection system. In real-time operations, detecting loop closures events can be computationally expensive when the limited computational architecture of mobile robots is considered. Storage capacity is another concern especially in large-scale environments. As the length of the image sequence increases throughout the trajectory, the process time for detecting loop closure events can grow boundless. Therefore, it is important that hardware limitations must be managed in an efficient way to operate the system in real-time, even for long sequences.

In a nutshell, loop closure detection is still an active and challenging issue for real-world SLAM applications. Avoiding false detections is a crucial factor, since they may cause catastrophic consequences for the general SLAM process. In addition to this, it is essential to use a fast and efficient algorithm with high discriminative power in order to cope with the problems arising from sensory information. In this thesis, a novel image description method considering all of the challenges for the use of visual loop closure detection is proposed. The literature review is presented in the next section, furthermore, the proposed method in this thesis is summarized in the subsequent section.

1.2 Literature Review

In vision-based loop closure detection systems, the description methodology is the primary point of focus when aiming to achieve good matching results without false detections. In [1], Garcia-Fidalgo and Ortiz classifies the loop closure detection problem according to description methodology as follows: Based on Bag-Of-Words (BoW) schemes [2], based on local features, based on global descriptors and the approaches based on combination of the other approaches. Figure 1.1 shows the graphical illustration of this classification.

The first approach for visual loop closure detection is BoW schemes which are the most common way to solve loop closure detection problem. BoW schemes are suitable for large-scale operations due to their well-indexed structure, termed as visual dictionary in which a set of similar features are grouped as a visual word. This approach is employed in the well known FAB-MAP [3] in which Cummins et al. proposed that modeling the probabilities that the visual words appear simultaneously can help in the localization process. By utilizing a Chow Liu tree computed from a set of training data, these probabilities were approximated to compute an observation likelihood. Predicting the loop closure events is performed by using Bayes filter. In [4], Sariyanidi et al. introduced an energy maximization based saliency detection method that has been used for unsupervised landmark extraction. They show that their method achieves

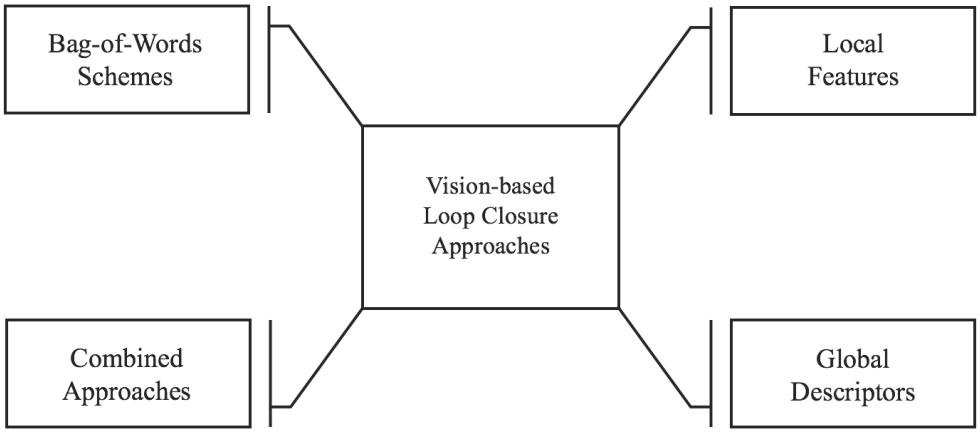


Figure 1.1 : Classification of vision-based loop closure approaches based on their description methodology.

high recall rates without false detections either outdoor or indoor dataset. The main drawback for these methods are that they come with high computational costs.

In BoW schemes, the size of a dictionary can vary within a large range which increases the performance of the retrieval process. Although the larger size of the dictionary provides more discrimination power, performing them in real-time requires too much CPU. To decrease the computational cost, Galvez-Lopez and Tardos attempted to create a visual dictionary from binary features in [5]. They adapted the hierarchical BoW approach to Bag-of-Binary-Words using BRIEF [6] descriptors instead of the gradient-based descriptors such as SIFT [7] and SURF [8]. They also reported that the proposed algorithm is suitable for large-scale operations. However, the hierarchical grouping process of BoW schemes engenders perceptual aliasing by using the same index for the other different words. In order to solve this problem, Cadena et al. [9] proposed a place recognition framework based on stereo vision which combined a BoW model for detecting loop closure candidates and an matching algorithm based on Conditional Random Fields (CRF-Matching). By using this matching technique, geometrically inconsistent images are discarded among the loop closure candidates that are detected with BoW schemes.

The other approach is based on local features consisting of two phases: A detection phase where interest points (such as corners or saliency regions) are detected over the entire image, and a description phase in which all of these points are represented with descriptor vectors. Zhang [10] presented a method for selecting a set of scale-invariant visual features (e.g. SIFT) from an image called Bag-of-Raw-Features (BoRF). A location was represented by these features that can be matched consecutively in several images. The techniques based on local features avoids the off-line process of vocabulary construction, and does not suffer from the perceptual aliasing unlike BoW schemes.

Local features is good at avoiding occlusions and decreasing the effect of illumination variations. Motivated by the argument, Carlevaris-Bianco and Eustice [11] introduced a new method for learning illumination changes by tracking feature points in time-lapse videos. The Euclidean distance is used to match the interest points in feature space. However, the matching complexity of local features is considerably high.

The last approach is describing the image in a holistic manner as a global image descriptor that is usually used for scene classification. Global descriptors recently gained increasing popularity [12–21]. An important advantage of those methods is that they do not require keypoint detection, which can be time consuming. The Gist descriptor [22, 23] is one of the early studies that proved the success of global descriptors [12]. Gist describes an image by convolving it with Gabor filters of various orientations and frequencies, and was shown to be suitable for large-scale operations (e.g. 12K images), owing to its compactness and computational simplicity. Applying PCA to Gist descriptors proved also useful in reducing the its size and improving its discrimination power [14].

Another approach to global image description is to first down-sample an image by a large factor, and then apply a well-established describe the image by computing a (single) local descriptor. The whole-image SURF (i.e. WI-SURF) was used in [15] as a main image descriptor. In this regard, in [24], it was presented that low resolution images are enough to achieve high accuracy loop closure detection and it is not needed to use complex algorithms. Following this work, several local descriptors have been compared [17], and the SIFT descriptor proved efficient and stability in describing a whole image (i.e. WI-SIFT).

Inspired by the Gist concept, BRIEF-Gist [13] has been introduced by using BRIEF as a global image descriptor computed around the center of the image. Binary descriptors require less memory and computational resources then real-valued descriptors [19]. Another approach to global description was to compute local descriptors in a predefined tiled structure (e.g. the image is partitioned to $m \times m$ tiles and each keypoint is defined at the center of the tiles) and concatenate them to a single descriptor [13]. Moreover, it was reported that BRIEF-Gist is not sensitive to global illumination changes to a certain extend when non-tiled version is used.

In more recent works [19, 20], LDB descriptors [25] have used for visual place recognition as global descriptors. Unlike BRIEF, LDB uses first order gradient information along with the intensity. In the presence of strong perceptual aliasing, global LDB requires further validation in spite of its efficiency. In [19], LDB was extended with additional disparity information (named as D-LDB) in order to deal with perceptual aliasing. All those recent studies focused on computational efficiency,

Table 1.1 : The pros and cons of each loop closure detection approach. More '⋆' indicates better performance regarding the corresponding attribute.

Feature	Global Descriptors	Local Features	BoW Schemes
CPU requirements	⋆⋆⋆	⋆	⋆⋆
Storage requirements	⋆⋆⋆	⋆	⋆⋆
Discrimination power	⋆	⋆⋆⋆	⋆⋆
Perceptual aliasing effect	⋆⋆	⋆⋆⋆	⋆
Large-scale operation	⋆⋆	⋆	⋆⋆⋆
Pose recovery complexity	⋆	⋆⋆⋆	⋆⋆

but, as we also experiment in this paper, their accuracy is limited particularly in the presence of strong perceptual aliasing.

The loop closure detection method presented in this thesis has been developed by considering the pros and cons of several visual loop closure detection techniques. It is obvious that, applying the only one approach is not sufficient for several reasons summarized in Table 1.1. However, in contrast to most studies, the study carried out in this thesis focuses on extracting local features on the image and combine them into a global image descriptor. The method proposed in this thesis is presented in the following section.

1.3 Hypothesis

The visual loop closure detection method presented in this thesis benefits from the discrimination power and robustness to perceptual aliasing of local features by combining them into a global image descriptor with a low computational complexity. This method relies on computing localized moment descriptors to achieve place recognition and ultimately loop closure detection.

Loop closure detection via localized moment descriptors involves two major components: 1) computing the image descriptor to represent the regarding location, 2) loop closure detection by matching the descriptors belonging to the locations.

Within the scope of this thesis, a global image descriptor that enables reliable and real-time loop closure detection in the presence of strong perceptual aliasing, is proposed. This technique relies on extracting QLZM patterns [26, 27] over the whole-image. Using local approach provides increasing robustness to background

illumination presented over the image since CZMs are computed all along the image in a local manner. As low-level shape features, LZMs have outstanding representation capability for an image containing discontinuities spread into different locations [26]. The descriptor representation, which is the major contribution of this thesis, is twofold: 1) Pattern image extraction and 2) histogram representation. A very straightforward quantization is applied to locally computed features with a simple binary test. Quantizing enables a reliable representation by reducing the effect of image noise. The pattern image consisting of non-linearly encoded quantized data is partitioned into cascaded subregions in order to achieve pose recovery. The final descriptor vector is constructed by concatenating the histograms that are built from each subregion as a global image descriptor.

Unlike the original QLZM representation [28], LZMs are computed in an overlapping way, and this approach has a significant effect on loop closure detection accuracy. As a further improvement, ZM bases are approximated to their minimal structures to be able to be computed with integral images (i.e. summed area tables). The integral images improve computation speed by a large margin without compromising loop closure accuracy, which is particularly useful for canceling out the computational caused by overlapping approach. With all these improvements, the descriptor is named as PALM (Patterns of Approximated Localized Moments) to clarify.

According to the loop closure methodology described in this thesis, the places are represented with their images. Therefore, the loop closure detection is performed by comparing the image of the most recent place with all the images that have been collected throughout the trajectory except itself and its neighbors to a certain extent. The nearest neighbor classifier is used for searching and retrieving the best candidate to close a loop with the most recent image. In this thesis, a similarity measure indicating the detection confidence is employed for this purpose. Consequently, a threshold is defined over the similarity measure to determine if the most recent location closes the loop.

The proposed loop closure detection technique has been evaluated on the public real-world datasets as follows: City Center and New College datasets [3] as well as the sequences number 5 and 13 of the KITTI Vision dataset [29]. These datasets are chosen by their aspects of containing diverse set of challenges such as similar

images for different places, dynamic environments where vehicles and pedestrians exist, large displacements in robot's positions due to its velocity. The results show that the proposed loop closure detection method performs in real-time with high recall rates avoiding false detections even for long sequences, and outperforms several state-of-the-art methods in the literature, which are FAB-MAP [3], BRIEF-Gist [13], D-LDB [19] and WI-SIFT [17].

This thesis involves a publication [30] which emphasis a brief introduction about the proposed loop closure detection method without the recent improvements.

2. DESCRIPTION METHODOLOGY

The description methodology plays a significant role for achieving good matching results. In order to decrease the effect of perceptual aliasing in description level, the image must be represented carefully. This chapter focuses on to explain computing PALM (Patterns of Approximated Localized Moments) descriptor that relies on extracting QLZM patterns [26] across the image.

2.1 Complex Zernike Moments

By nature, CZMs have a superior capability than the other image moments in terms of image representation since not only their noise resilience but also no redundancy or overlap of information between the moments as stated in [31]. Computing CMZs of an image are used to represent the image on the 2-dimensional complex subspace with a set of Zernike polynomials that are orthogonal to each other. These polynomials are defined as:

$$V_n^m(\rho, \phi) = R_n^m(\rho) e^{im\phi}, \tag{2.1}$$

where ρ and ϕ are the polar coordinates, n (≥ 0) is the order of the polynomials and m is the number of iterations. The constraint between n and m is that $|m| \leq n$ and $n - |m|$ must be even. The radial polynomials $R_{nm}(\rho)$ are given as:

$$R_n^m(\rho) = \sum_{s=0}^{\frac{n-|m|}{2}} \frac{(-1)^s \rho^{n-2s}(n-s)!}{s!(\frac{n+|m|}{2}-s)!(\frac{n-|m|}{2}-s)!}. \tag{2.2}$$

Let $\hat{x}$ and $\hat{y}$ be the pixel coordinates on a digital image $I(x, y)$ mapped to the range of $[-1, +1]$, $\rho_{xy} = \sqrt{\hat{x}^2 + \hat{y}^2}$ and $\phi_{xy} = \tan^{-1} \frac{\hat{y}}{\hat{x}}$. The CZMs of the image $\mathcal{Z}_n^m$, consists of a real (i.e. magnitude) and an imaginary (i.e. phase) components and can be computed as follows:

$$\mathcal{Z}_n^m(I) = \frac{n+1}{\pi} \sum_x \sum_y I(x, y)[V_n^m(\rho_{xy}, \phi_{xy})]^\star, \tag{2.3}$$

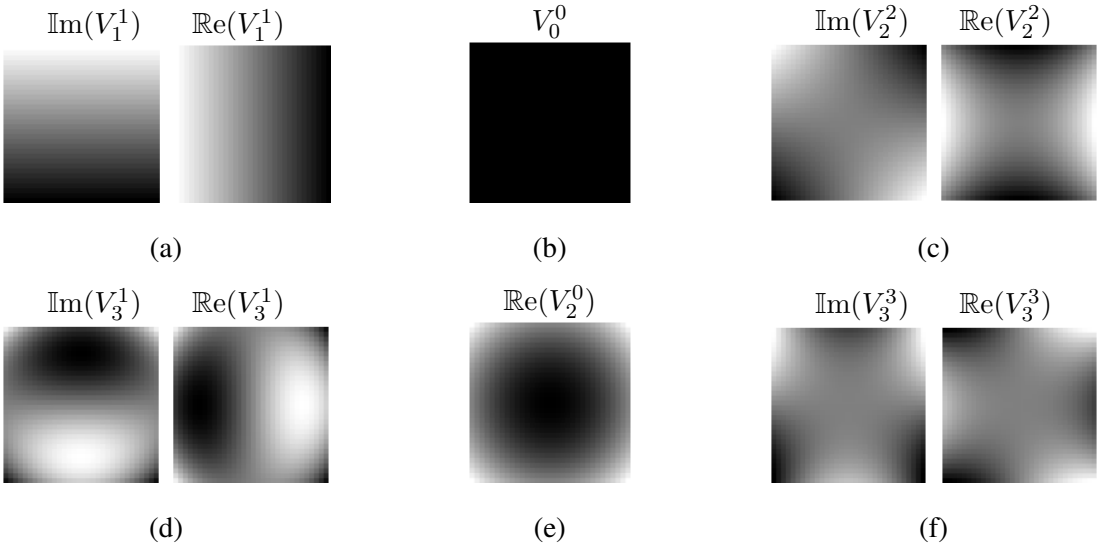


Figure 2.1 : Plots of the real and imaginary components of normalized Zernike bases, V_n^m, for $n = 1, 2, 3$ with all satisfying m. Note that the dark shaded areas indicate the higher values.

where * denotes the hermitian transpose of a complex matrix. CZMs can be viewed as the responses of the image function to a set of quadrature-pair filters $\{V_n^m(\rho_{xy}, \phi_{xy})\}$. Exemplary Zernike polynomials (i.e. filters) are depicted in Figure 2.1.

The completeness and orthogonality of Zernike polynomials enables representing any digital image function with a set of complex numbers obtained from the equation 2.3. If all the moments $\mathcal{Z}_n^m$ of the image function $I(x, y)$ with a max order N are known, it is possible to reconstruct this function by using the moments up to N. For a larger value of N, the reconstructed image has more resemblance to the original one. The reconstruction is performed as:

$$I'(x, y) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \mathcal{Z}_n^m(I) V_n^m(\rho_{xy}, \phi_{xy}), \tag{2.4}$$

where $n - |m|$ is even.

Determining the maximum order is important to prevent the presence of redundant information in terms of image representation. While higher order moments store the detailed information like edges and repetitions (i.e. higher frequencies), lower ones preserves holistic components (i.e. lower frequencies) spread over the image. Therefore, higher order moments can be ignored in a certain extent to represent the general structure of the image. The moment coefficients of an image for all possible n

and m values can be formed as:

$$\mathcal{Z}_N(I) = \{\mathcal{Z}_0^0, \mathcal{Z}_1^1, \mathcal{Z}_2^0, \mathcal{Z}_2^2, \ldots, \mathcal{Z}_N^m\}. \tag{2.5}$$

The global form of CZMs is expressed in the equation 2.3. By definition, it takes an image and computes the CZMs with respect to the moment order and corresponding iteration number. In other words, CZMs can simply be thought as a large set of coefficients that represents the holistic shape characteristics present on the entire image. By applying several position normalization techniques, these moments can be easily made invariant to translation, since they lack localization information. However, they are not capable of representing a scene image containing discontinues spread into different locations. Yet, displaying image continuities effectively is critical, particularly in the presence of perceptual aliasing. In order to cope with this problem, CZMs are computed all along the input image in a local manner. The next section explains the local approach of computing CZMs, which are called LZMs.

2.2 Local Zernike Moments

LZMs rely on computing the moment coefficients around predefined neighborhoods throughout the entire image. Let I be the $p \times q$-sized input image [26, 28]. In order to extract localized moments, firstly the input image $I_{p\times q}$ is divided to non-overlapping $k \times k$-sized image patches I_{ij}. Note that if the image is not divided by k perfectly, the pixels falling into the remaining area are simply discarded. Thereby, a set of image patches is obtained by dividing the input image as follows:

$$I_{p\times q} = \begin{bmatrix} I_{11} & \ldots & I_{1Q} \\ \vdots & \ddots & \vdots \\ I_{P1} & \ldots & I_{PQ} \end{bmatrix} \tag{2.6}$$

where $P = \lfloor {}^{p}/_{k} \rfloor$ and $Q = \lfloor {}^{q}/_{k} \rfloor$. An exemplar input image divided to local patches is shown in Figure 2.2.

The next step is computing the CZMs of each local image patch using the equation 2.3 with all possible n and m values with respect to maximum moment order N. Therefore, LZM transformation of the input image $\mathcal{L}_N(I)$ is given as:

$$\mathcal{L}_N(I) = \begin{bmatrix} \mathcal{Z}_N(I_{11}) & \ldots & \mathcal{Z}_N(I_{1Q}) \\ \vdots & \ddots & \vdots \\ \mathcal{Z}_N(I_{P1}) & \ldots & \mathcal{Z}_N(I_{PQ}) \end{bmatrix} \tag{2.7}$$

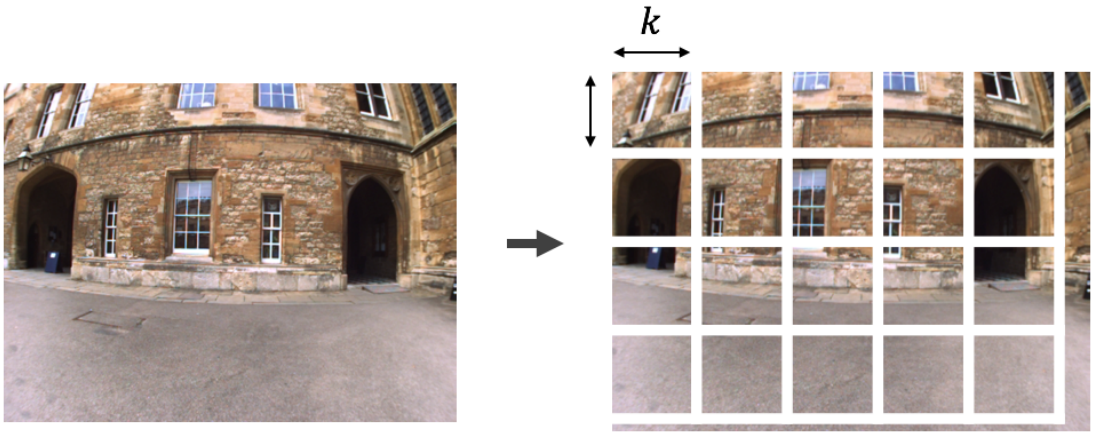


Figure 2.2 : Input image divided to $k \times k$ sized non-overlapping local patches. The remaining pixels falling into the outside of local patches are discarded.

Hence, the moment coefficients for each local image patch is formed as shown in the equation 2.5:

$$\mathcal{L}_N(I) = \begin{bmatrix} \{\mathcal{Z}_1^1, \mathcal{Z}_2^2, \ldots, \mathcal{Z}_N^m\}_{00} & \cdots & \{\mathcal{Z}_1^1, \mathcal{Z}_2^2, \ldots, \mathcal{Z}_N^m\}_{1Q} \\ \vdots & \ddots & \vdots \\ \{\mathcal{Z}_1^1, \mathcal{Z}_2^2, \ldots, \mathcal{Z}_N^m\}_{P1} & \cdots & \{\mathcal{Z}_1^1, \mathcal{Z}_2^2, \ldots, \mathcal{Z}_N^m\}_{PQ} \end{bmatrix} \tag{2.8}$$

where the moment components that have their iteration numbers equal to zero ($m = 0$) are ignored. Although these components can be useful for reconstruction tasks, they do not provide any information regarding image recognition and matching [32, 33]. Also, their structure can be viewed as a moving average filter in which the sum of the values is not zero. Therefore, the number of effective moment components can be calculated through the following expression:

$$K(N) = \begin{cases} \frac{N(N+2)}{4} & \text{if } N \text{ is even} \\ \frac{(N+1)^2}{4} & \text{if } N \text{ is odd.} \end{cases} \tag{2.9}$$

In other words, $K(N)$ indicates the total number of complex coefficients for each local image patch. However, either magnitude component or the phase component of each complex coefficient contains diverse set of information [34]. To this end, in order to exploit them separately, a new set is constructed from $\mathcal{Z}_N(I_{ij})$ by decoupling each complex coefficient as:

$$\bar{\mathcal{Z}_N}(I_{ij}) = \{\mathbb{R}\text{e}(\mathcal{Z}_1^1), \mathbb{I}\text{m}(\mathcal{Z}_1^1), \ldots, \mathbb{R}\text{e}(\mathcal{Z}_N^m), \mathbb{I}\text{m}(\mathcal{Z}_N^m)\}_{ij} \tag{2.10}$$

where the length of each set $|\bar{\mathcal{Z}}_N(I)| = 2K(N)$. Hence, the final form of LZM transformation becomes as follows:

$$\bar{\mathcal{L}}_N(I) = \begin{bmatrix} \bar{\mathcal{Z}}_N(I_{11}) & \dots & \bar{\mathcal{Z}}_N(I_{1Q}) \\ \vdots & \ddots & \vdots \\ \bar{\mathcal{Z}}_N(I_{P1}) & \dots & \bar{\mathcal{Z}}_N(I_{PQ}) \end{bmatrix}. \tag{2.11}$$

LZMs have better representation capability than the global form of CZMs with only a small set of coefficients (i.e. with low order moments) [26]. In addition to this, using local approach provides increasing robustness to illumination variations presented over the image, since the assumption suggests that the illumination effects can be ignored when the a small piece of the image is considered. The next section explains how to efficiently utilize the information that is extracted from the input image in local manner.

2.3 Pattern Image Extraction

The pattern image P consists of roughly quantized overlapping LZM patterns in an image form. This process is the first phase of constructing the PALM descriptor. In the practice of extracting pattern image; 1) overlapping LZMs are computed all along the input image, 2) a straightforward quantization is applied in order to make a reliable representation.

2.3.1 Overlapping Approach

A simple way to improve the ability of LZMs to represent image discontinuities is to compute them in overlapping image patches. In regular LZM transformation, the input image is divided to non-overlapping $k \times k$-sized local patches, and compute CZMs independently for each patch. In other words, to start from the top-left of the image, compute the CZMs for across $k \times k$ pixels, then move rightwards by k pixels, and compute the CZMs for this new patch, and so on. In order to represent discontinuities even better, LZMs are computed from the local patches just like the regular one, but more densely. It is performed by starting from the top-left of the image, and then moving rightwards by a step of s pixels (instead of k) with the constraint of $k = cs$ where c specifies overlapping density ($c > 0$). This process is illustrated in Figure 2.3.

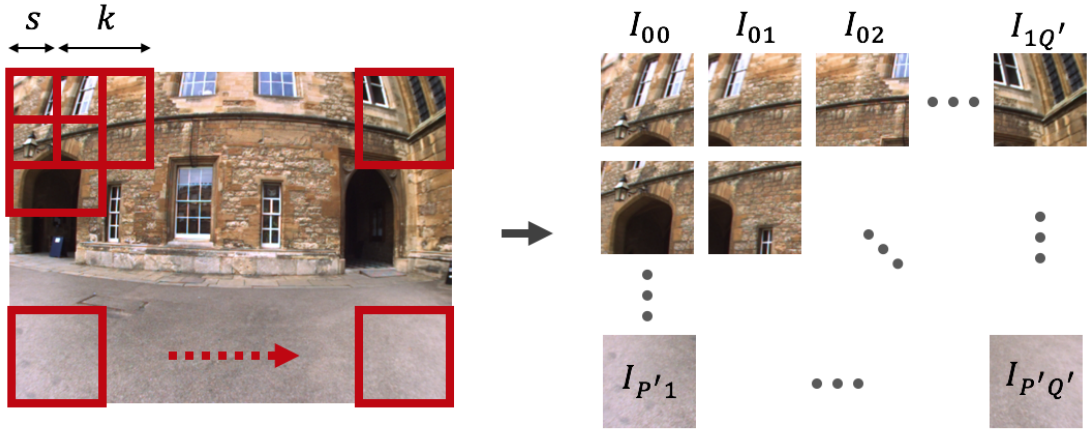


Figure 2.3 : Input image divided to overlapping $k \times k$ sized local image patches by a step of s pixels. Notice that every local patch, except the patches next to the borders, overlaps with its adjacent patches.

With the overlapping approach, the LZM transformation of an input image $\mathcal{L}_N(I_{p\times q})$, through densely computed CZMs, becomes as follows:

$$\bar{\mathcal{L}_N}^c(I_{p\times q}) = \begin{bmatrix} \bar{\mathcal{Z}}_N(I_{11}) & \dots & \bar{\mathcal{Z}}_N(I_{1Q'}) \\ \vdots & \ddots & \vdots \\ \bar{\mathcal{Z}}_N(I_{P'1}) & \dots & \bar{\mathcal{Z}}_N(I_{P'Q'}) \end{bmatrix} \tag{2.12}$$

where $P' = \lfloor \frac{p}{s} - c \rfloor + 1$ and $Q' = \lfloor \frac{q}{s} - c \rfloor + 1$. It is clear that if $c = 1$ (i.e. $k = s$), there will be no overlaps between successive local patches and it can be considered as the regular LZM computation [28]. Although this improvement enhances the processing time, it has direct effect on the discrimination power of the descriptor which allows to decrease the effect of perceptual aliasing. Furthermore, the problem of increase in computational time consumption is minimized by using integral images which is explained in Section 2.5.

2.3.2 Quantization

A large number of coefficients are obtained when Zernike moments are computed separately for each local patch I_{ij}. To reduce computational storage and increase computational speed, the data must be compressed somehow. To this end, a very straightforward quantization is applied to not only compress the data, but also reduce the effect of background illumination that slightly varies across the scene image. Quantization is performed by taking the sign of each Zernike moment with a piecewise

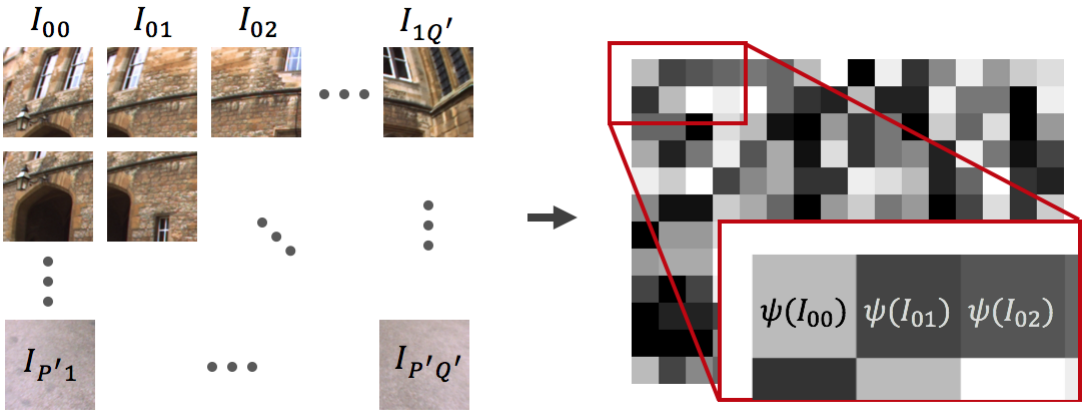


Figure 2.4 : Composition of the pattern image that contains overlapping QLZM patterns.

function $\theta(.)$ (i.e. unit step function) defined as:

$$\theta(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0. \end{cases} \tag{2.13}$$

By applying $\theta(.)$ to each element of $\bar{\mathcal{Z}_N}(I_{ij})$ in the equation 2.6, $\bar{\mathcal{Z}_N}(I_{ij})$ becomes a set of binary numbers:

$$\mathcal{Q}_N(I_{ij}) = \left\{ \theta\Big(\mathbb{R}\mathrm{e}(\mathcal{Z}_1^1) \Big), \ldots, \theta\Big(\mathbb{I}\mathrm{m}(\mathcal{Z}_n^m) \Big) \right\}_{ij} \tag{2.14}$$

where $|\mathcal{Q}_N(I_{ij})| = 2K(N)$. Knowing the length of $\mathcal{Q}_N(I_{ij})$ is important, since this number is used to determine the size of PALM descriptor which stands for the scene image. Also, the sign of the moments provides a discriminative ability comparable to that of the raw Zernike moments, while improving compactness greatly [26].

Using $\mathcal{Q}_n(I_{ij})$ as a raw is not a practical solution since it can not be represented with histograms which enable robustness to relative pose variations of the robot. The quantized information provided by different CZM coefficients does not overlap and they describe the variation within the local patches at a unique scale and orientation as stressed in Section 2.1. Therefore, to make a reliable representation for each local image patch I_{ij}, $\mathcal{Q}_N(I_{ij})$ is encoded by a non-linear function which excels at enhancing the relevance of low-level features (e.g. moments coefficients) by increasing their robustness against image noise [35]. Therefore, a single pixel of the pattern image $P(i, j)$ containing the QLZM patterns of I_{ij} is computed through a function $\psi(.)$,

illustrated in Figure 2.4, as follows:

$$P(i,j) = \psi(I_{ij}) = \sum_{q=0}^{2K(N)-1} 2^q \Big[\mathcal{Q}_n(I_{ij})\Big]_q. \tag{2.15}$$

It is clear that all elements of the pattern image are decimal integers ranging between 0 and $2^{2K(N)} - 1$.

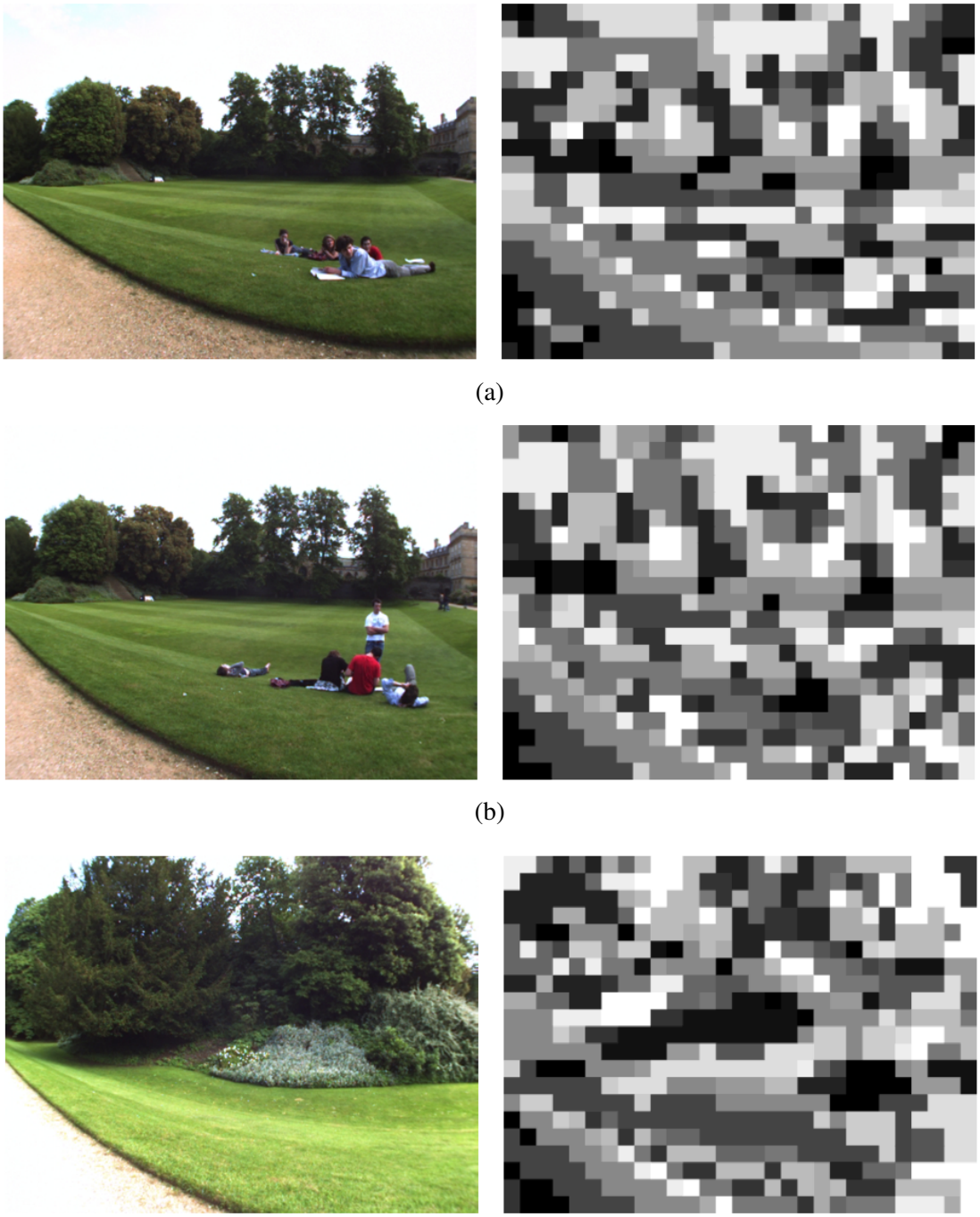


Figure 2.5 : Examples of scene images and their corresponding pattern images.

By definition, $\psi(.)$ is the core process of the proposed description method which has robustness to illumination variations and noise arising from visual sensory data. Figure 2.5 presents several examples of scene images and their corresponding pattern images. As it can be seen, Figure 2.5a and 2.5b are taken from the same location with a slightly slipped view point of the robot. Besides, it is clearly noticed that the pattern images in Figure 2.5a and 2.5b have more common features than the one in Figure 2.5c which is actually taken in a completely different place. The people sitting on the grass can be viewed as image noise that is successfully eliminated. In addition to this, the part of road located on bottom-left corner of the scene images is processed as the same in the pattern images despite the intense illumination. The next section explains how these information contained in the pattern images are used to build a descriptor vector in terms of visual place recognition, ultimately loop closure detection.

2.4 Histogram Representation

The methodology described so far, takes an input image I and extracts a pattern image P consisting of decimal integers encoded with overlapping QLZM patterns of corresponding local patches of the input image. Reshaping pattern image to a descriptor vector is not a good solution to compensate relative pose variations of the robot at the same location (See the pose variation between Figure 2.5a and 2.5b). To this end, the descriptive features contained in the pattern image are represented with regional histograms to decrease the effect of not only small relative pose variations but also partial occlusions from moving objects on the image.

Initially, the pattern image is partitioned into $g \times g$ (i.e. grid size) subregions uniformly, illustrated in Figure 2.6. It is reported that using overlapping subregions increases the descriptive performance in terms of global image description in [36]. Therefore, an inner partitioning is applied to the same pattern image with another $(g-1) \times (g-1)$ sized gridded subregions of which centers fall into the intersection points of the outer partitions in an overlapping way.

For each subregion, a histogram that consists of $b = 2^{2K(N)} - 1$ bins is defined since the pattern image contains decimal integers ranging between 0 and $2^{2K(N)} - 1$ as stressed in the previous section. In these histograms, each bin simply counts the number of occurrences of the pattern features in relevant subregion. However, a problem with

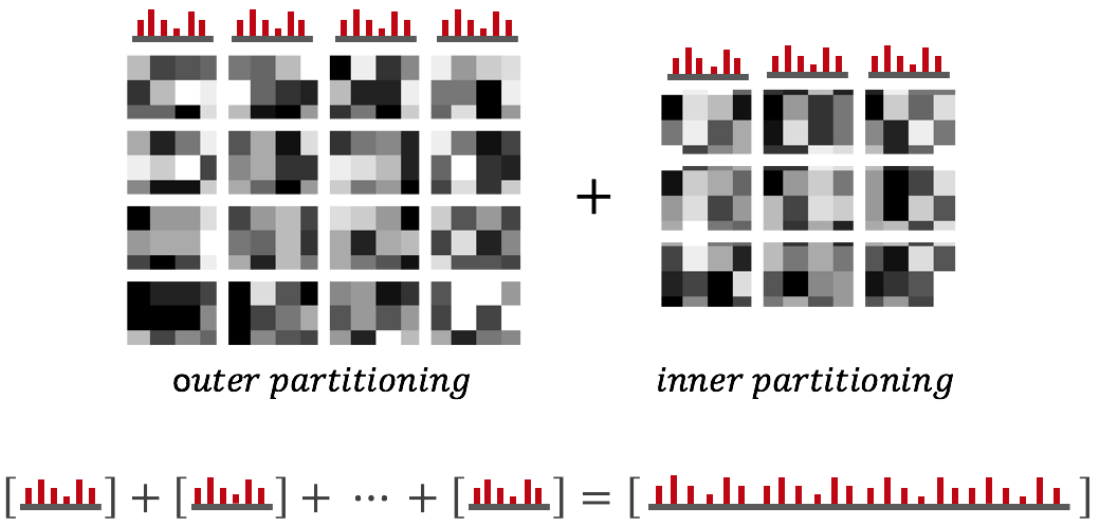


Figure 2.6 : Outer and inner partitioning applied to the pattern image to construct the final descriptor vector.

regional histograms is that local shape features along the borders may fall out of the histogram in the presence of small geometric variations. To deal with this, the features along the borders are down-weighted by applying a Gaussian mask peaked at the center of each subregion. In addition to this, the histograms regarding different subregions are normalized in itself. Similar approaches are employed in the well-known SIFT [7] and HoG [36], which was inspired from the SIFT descriptor by spreading them all along the image.

Consequently, the final descriptor vector is constructed by concatenating the regional histograms as a global image descriptor. The size of the descriptor d can be calculated as follows:

$$size(d) = b[g^2 + (g-1)^2] \tag{2.16}$$

where b directly depends on the maximum moment order N through $b = 2^{2K(N)} - 1$. Exemplar descriptor sizes is calculated with a particular moment order with respect to different grid sizes in Table 2.1.

Table 2.1 : Descriptor sizes for the moment order $N = 2$ (accordingly the bin size $b = 16$ for each histogram) with respect to different grid sizes.

Grid size (g)	2	3	4	5	6	7	8
Descriptor size (d)	80	208	400	656	976	1360	1808

2.5 Integral Image Approximation

One particular challenge is that a loop closure algorithm must be operated in real-time even if a robot has limited computational architecture. Overlapping LZM approach can be computationally expensive in case overlapping density c is high. In order to deal with this problem, the integral images (i.e. summed area tables) [37, 38] are employed. Therefore, the computation of the LZM patterns is tremendously accelerated by representing Zernike bases with the approximated ones without almost no loss of information. In this section, the usage of integral images in the context of pattern image extraction is described.

Integral images have been used to accelerate various computer vision algorithms by quickly and efficiently generating the sum of values in a rectangular subset of an image. Zernike bases are roughly approximated with a group of rectangular areas preserving the main structures of them. Figure 2.7 depicts several examples of approximated Zernike bases obtained by the approximation process. Thanks to integral images, the sum of any rectangular area can be computed in a constant time for a single image patch p, illustrated in Figure 2.8, as:

$$\beta_j \sum_{x=x_0}^{x_1} \sum_{y=y_0}^{y_1} p(x,y) = \beta_j \Big[i(x_1,y_1) + i(x_0,y_0) - i(x_0,y_1) - i(x_1,y_0) \Big] \qquad (2.17)$$

where $i(x,y)$ denotes the integral image of p and β_j indicates the corresponding coefficient for each rectangular area of which top-left and bottom-right coordinates are (x_0,y_0) and (x_1,y_1) respectively. Similar strategy was applied to Gaussian filters in [8] to speed up the computation of 2-dimensional filters.

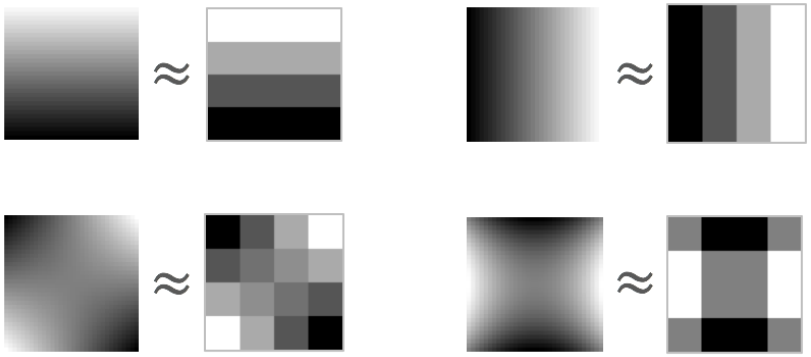
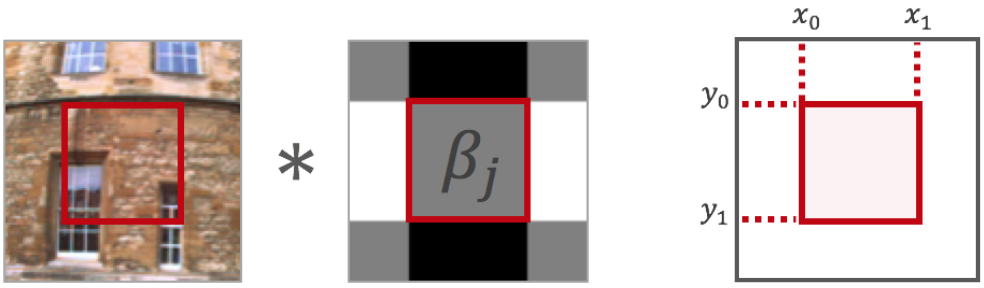


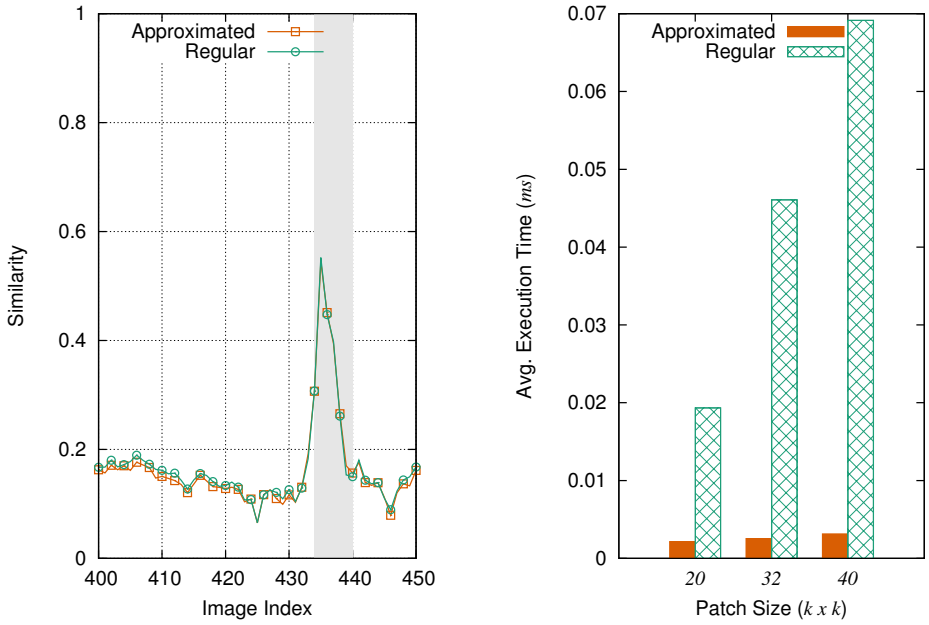
Figure 2.7 : Several examples of approximated Zernike bases.



(a) (b)

Figure 2.8 : Efficient implementation via integral images. (a) CZM computation of a local patch. (b) Integral image of the local patch.

Figure 2.9a demonstrates a comparison of the outcome of regular and approximated bases in terms of the normalized similarities between an image and the other images across the sequence. As it is seen, there is almost no difference between the outcome of regular and approximated Zernike bases. In Figure 2.9b, the average execution time of using the regular and approximated Zernike bases are compared with respect to different patch sizes. By using approximated bases, a single image patch, in other words, a single pixel of the pattern image, can be computed in an almost constant time regardless of what the patch size k is.



(a) (b)

Figure 2.9 : Comparison of the regular and approximated Zernike bases. (a) Average execution time with respect to different patch sizes $k \times k$. (b) Similarity between an image and the other images across the sequence. Note that gray shaded area indicates ground truth locations.

3. LOOP CLOSURE METHODOLOGY

In the previous chapter, the description method that is capable of preventing most of the challenges arising from the visual sensory is explained. Accordingly, this chapter focuses on explaining how the loop closures are detected. There are two major issues that must be tackled in order to detect loop closure events: 1) similarity between two images must be measured, 2) previously visited locations must be determined whether a location is seen before or not.

3.1 Similarity Measure

Measuring the similarity between two locations, ultimately their descriptor vectors, is an important factor to distinguish them apparently. In addition to this, the computational overhead must be taken into consideration to operate the system in large-scale environments. The similarity measure for the loop closure detection method presented in this thesis is established on computing the dissimilarity between the descriptor vectors of corresponding locations.

The dissimilarity function simply indicates that the less similar descriptor vectors yield the larger function values in contrast to the similarity between them. By this way, the dissimilarity is computed by using the regular L_1 distance [1] which is one of the most simple distance metric available in terms of computational efficiency. L_1 distance is also suitable for comparing histogram based descriptors. The dissimilarity function is defined as:

$$\textit{dissimilarity}(d_i, d_j) = L_1(d_i, d_j) = ||d_i - d_j||_1 = \sum_t |d_i^t - d_j^t| \tag{3.1}$$

where d_i and d_j denote two different descriptor vectors which sizes are equal. The next section, explains how the dissimilarity measure is used in the proposed loop closure detection system to find out the most similar locations.

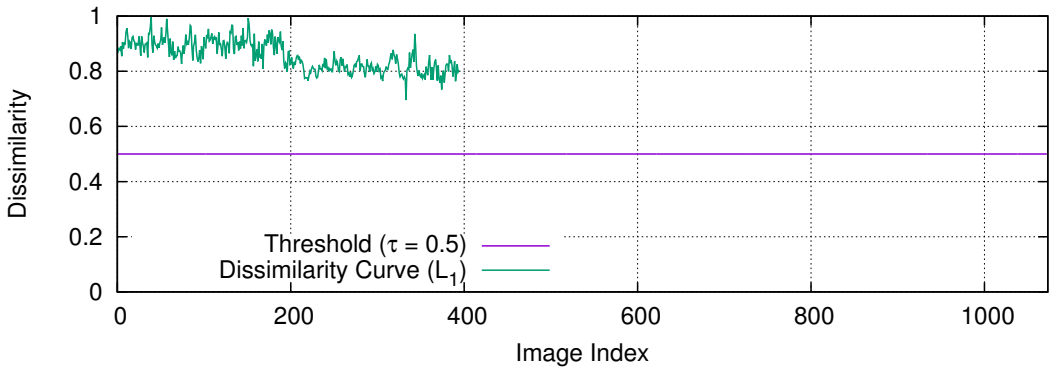
[1] The L_1 distance is also known as Manhattan or City Block distance defined in the Taxicab Geometry.

3.2 Loop Closure Validation via Nearest Neighbor Search

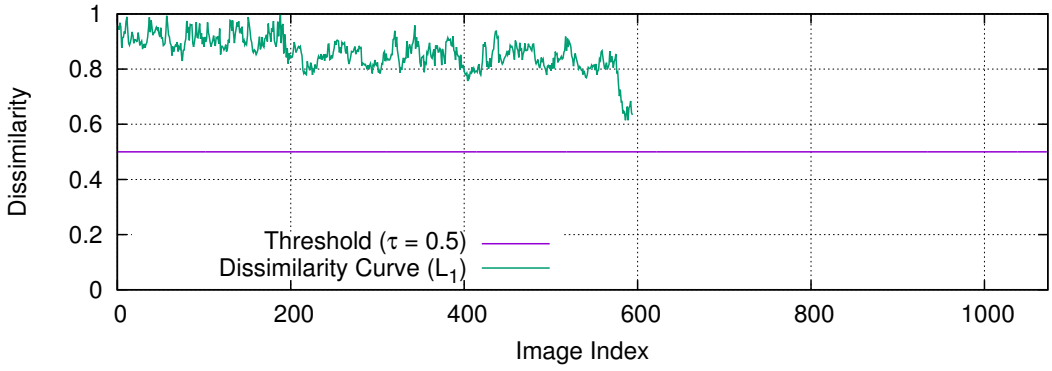
In the proposed loop closure methodology, NNS algorithm is adopted to recognize the locations that are previously visited. Loop closure detection is performed by simple brute force searching and retrieving the best candidate to close a loop with the most recent image by comparing with the other images collected throughout the trajectory.

In this context, every single image captured from the camera is considered as a location. For every new location in the trajectory, the PALM descriptor of the corresponding location is computed. The next step is to calculate the dissimilarity between the other locations with the equation 3.1 except itself and its adjacent locations (i.e. 5 images around the most recent location [21]). The loop closure hypothesis is accepted if the distance between candidate and matched locations is below a pre-defined threshold τ. Figure 3.1 presents several exemplar dissimilarity curves for the most recent image in the case of being at different indexes. As it can be seen, there are no detected loop closure event by the system in Figures 3.1a and 3.1b since any part of the curve is not below the threshold. In contrast to this, Figures 3.1c and 3.1d demonstrates successfully detected loop closures, because the thin rectangles that are perpendicular to the line indicating the threshold, denotes the indexes of corresponding ground truth locations.

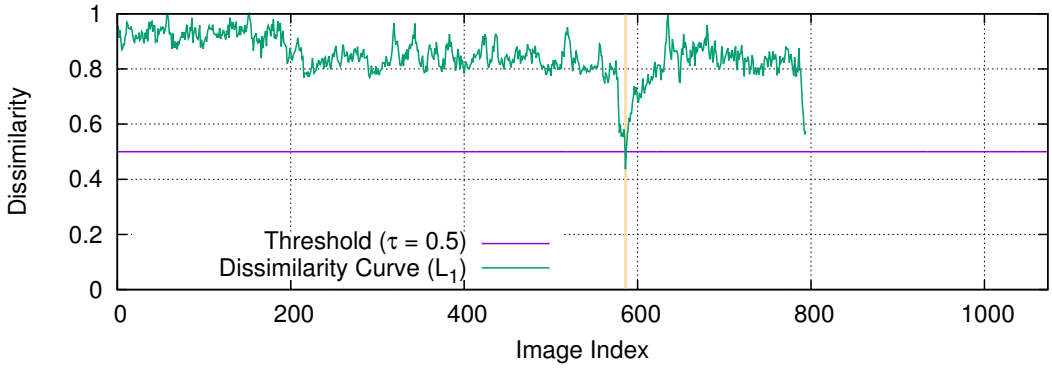
The main drawback of this approach is the linear time complexity $\mathcal{O}(dn)$ over-headed by NNS algorithm. For each recently queried image, the overall process time will be increased linearly. However, the computation of the dissimilarity using L_1 between the locations is performed quite fast, therefore the slope of incrementing processing time is slightly low. The time consumption of the proposed loop closure detection method has been analyzed in the subsequent chapter.



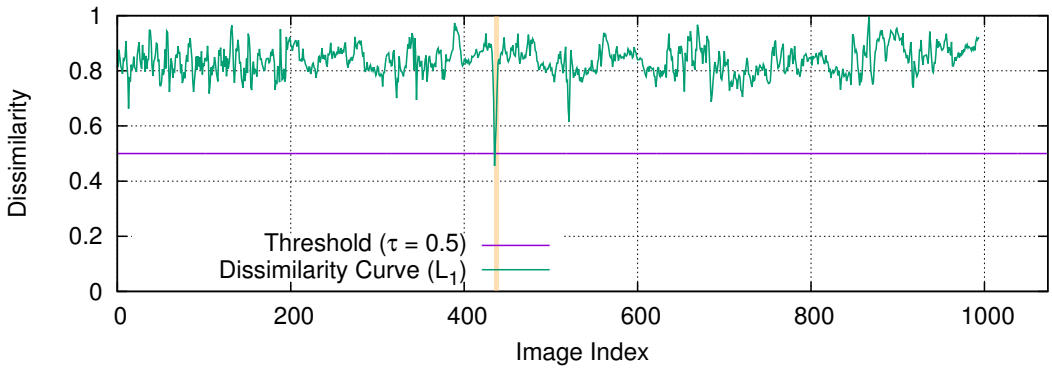
(a) The candidate image index is at 400.



(b) The candidate image index is at 600.



(c) The candidate image index is at 800.



(d) The candidate image index is at 1000.

Figure 3.1 : Dissimilarity curves for the candidate image index is at (a) 400, (b) 600, (c) 800 and (d) 1000. Note that the yellowish rectangles denotes the indexes of corresponding ground truth locations.

4. EXPERIMENTAL RESULTS

In this chapter, a set of experiments has been conducted on several well-known real-world datasets to extensively validate the proposed loop closure detection method. Additionally, the method is compared with the other state-of-the-art loop closure detection methods in the literature.

4.1 Datasets

New College and the City Center datasets were first employed for evaluating the well-known FAB-MAP [3]. The images are collected at every 1.5 m from the two cameras that are pointing to the opposite directions from the point of view of the robot. The ground truth informations are obtained with a hand-corrected GPS and published with the datasets for convenient usage. These datasets are chosen since each of them contains diverse set of challenges in order to evaluate the proposed loop closure detection system. Also, they are often employed in the literature [1, 3, 13].

The first one, New College dataset is gathered in 1.9 km long trajectory consisting of 640×480-sized 1073 image pairs, including left and right ones, and several large areas of identically repeating structures like stone walls, trees and bushes. In particular, it is constructed to test the system's robustness to perceptual aliasing. Figure 4.1a presents example images taken from the New College dataset.

The latter, City Center dataset is collected along public roads near the city center where various dynamic objects such as vehicles and pedestrians exist. There are 1237 pairs of images in this dataset which trajectory length is approximately 2 km. Specifically, it was created to examine matching ability in the presence of scene change. The abundant foliage and shadow features are unstable since it was recorded on a windy day with bright sunshine. This dataset also contains bidirectional loop closures that the proposed method cannot handle. Several example images are shown in Figure 4.1b.

(a) Example images in the New College dataset.

(b) Example images in the City Center dataset.

(c) Example images in the KITTI Vision dataset.

Figure 4.1 : Example images taken from the (a) New College, (b) City Center and (c) KITTI Vision datasets.

The KITTI Vision dataset [29] has 22 sequences containing a total of 44182 stereo images, each size of 1240×376, with 39.2 km long trajectory. However, only two of them (i.e. sequences number 5 and 13) are used to validate the proposed loop closure detection method since these sequences contain a lot of loop closure events. The difficulties in these sequences include intense illumination variations as well as a large variety of displacements in the agent's positions due to its velocity. Also, the sequence number 5 contains frames in which the agent is kept stationary. Figure 4.1c presents several example images taken from these datasets. Sequences that is utilized in this thesis consist of 2761 and 3281 stereo images respectively. The proposed method is not specialized for stereo vision, therefore only the images collected from the left camera is utilized. The ground truth informations are provided by [19] with the form of similarity mask.

4.2 Loop Closure Detection Performance

This section demonstrates the loop closure performance of the proposed method by using these datasets stated in the previous section. The design parameters of the proposed description method highly depend on the input image size. Thereby, all the images queried from the datasets are downsized to a size of 320×320 for test purposes to measure the performance independent from the image size. The maximum moment order n and the grid size g are held constant (i.e. $n = 2$ and $g = 5$) to stabilize the descriptor size which is calculated as it is expressed in the equation 2.16. By this way, the descriptor size is calculated as 656 that is an appropriate value for a single image. Also, $n = 2$ suffices to create an adequate description while also maintaining a low dimensionality [28]. The overlapping density c is set to 4 through which using approximated Zernike bases decreases the computational time tremendously even for higher c values as it is emphasized in Section 2.5. Furthermore, the major design parameter that can make a difference in terms of results is the patch size k, which is the same value as the Zernike bases sizes. The proposed loop closure detection system is evaluated with respect to different patch sizes and the settings mentioned in this paragraph will be used in the subsequent sections.

In Section 1.1, it was highlighted that the false detections might turn out catastrophic for the overall SLAM system. Therefore, the loop closure detection performance is measured by counting the true positives, true negatives, false positives and false negatives. A quick summary of these terms is as follows:

- *True Positive (TP):* A loop closure is detected and there is indeed a closure.
- *False Positive (FP):* A loop closure is detected, but there isn't any closure.
- *True Negative (TN):* No loop closure is detected and there is no closure.
- *False Negative (FN):* No loop closure is detected, but there is actually a closure.

By using these terms, two significant metric to evaluate the performance of the loop closure detection system can be calculated as follows:

Table 4.1 : Loop closure detection performances of the proposed method on the datasets with respect to different patch sizes.

	Recall @100% Precision		
Dataset	$(k = 20)$	$(k = 32)$	$(k = 40)$
New College	0.441	**0.552**	0.521
City Center	0.446	**0.565**	0.551
KITTI Vision 05	**0.744**	0.652	0.585
KITTI Vision 13	0.169	**0.279**	0.222

- *Precision:* Ratio between true positives and the total amount of loop closures detected $\left(\frac{TP}{TP+FP}\right)$.
- *Recall:* Ratio between true positives and the total amount of loop closures available on the dataset $\left(\frac{TP}{TP+FN}\right)$.

Since precision is more valuable measure in this context, the recall rates at high precision are much more significant for the system (i.e. a system detecting less loop closures with high precision is more valuable than a system detecting more loop closures with less precision). The major performance metric emphasized in the experiments is the recall rate at 100% precision. Table 4.1 lists the loop closure detection performances of the proposed method on the datasets with respect to different patch sizes k.

In New College and City Center datasets, the proposed method that $k = 32$ is able to detect loop closure events with 55.2% and 56.5% recall rates at 100% precision respectively. When same patch sizes are used, the performance results are quite close to each other for both datasets. In addition to this, the results at $k = 40$ is lower than the ones at $k = 32$, even though the increase in patch size from 20 to 32 provides higher detection performance. In KITTI Vision sequences number 5 and 13, the results are highly different from each other except for the sequence number 13 in which the change of the results with respect to increasing patch sizes is similar to the result behavior on New College and City Center datasets. In Kitti Vision 05 dataset, the proposed system yields maximum 74.4% recall rate at 100% precision by using 20×20-sized patches and the values slightly decrease while k is increasing. In contrast, the settings with $k = 32$ gives 27.9% recall rates which is the best results for the sequence number 13.

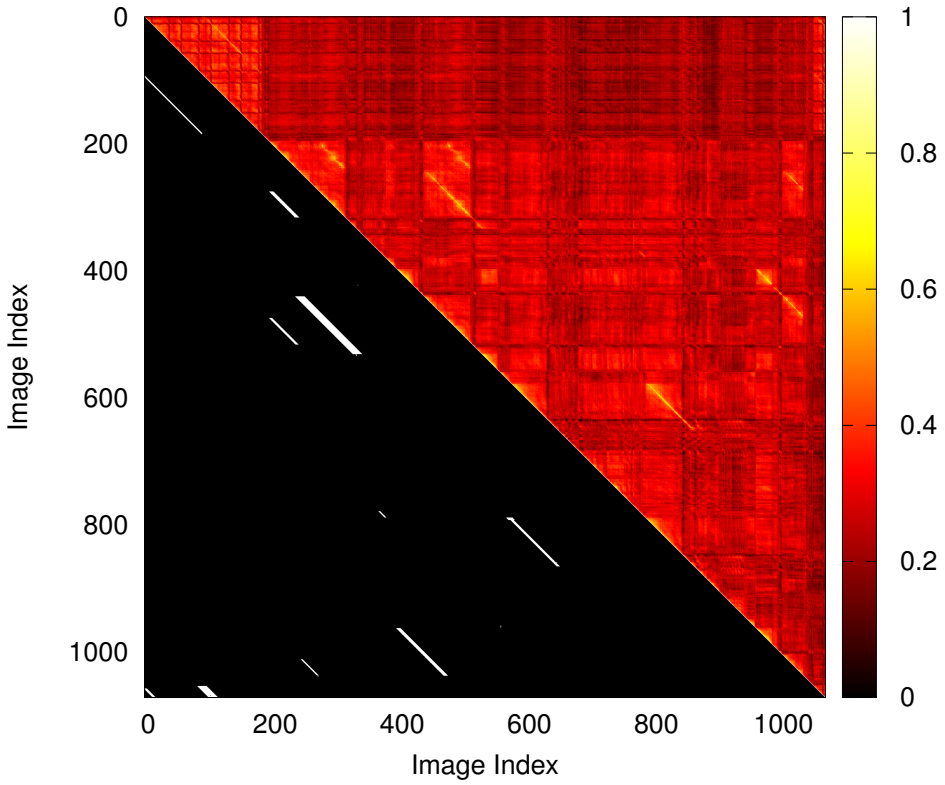


Figure 4.2 : Similarity matrix extracted by using New College dataset with the setting $k = 32$.

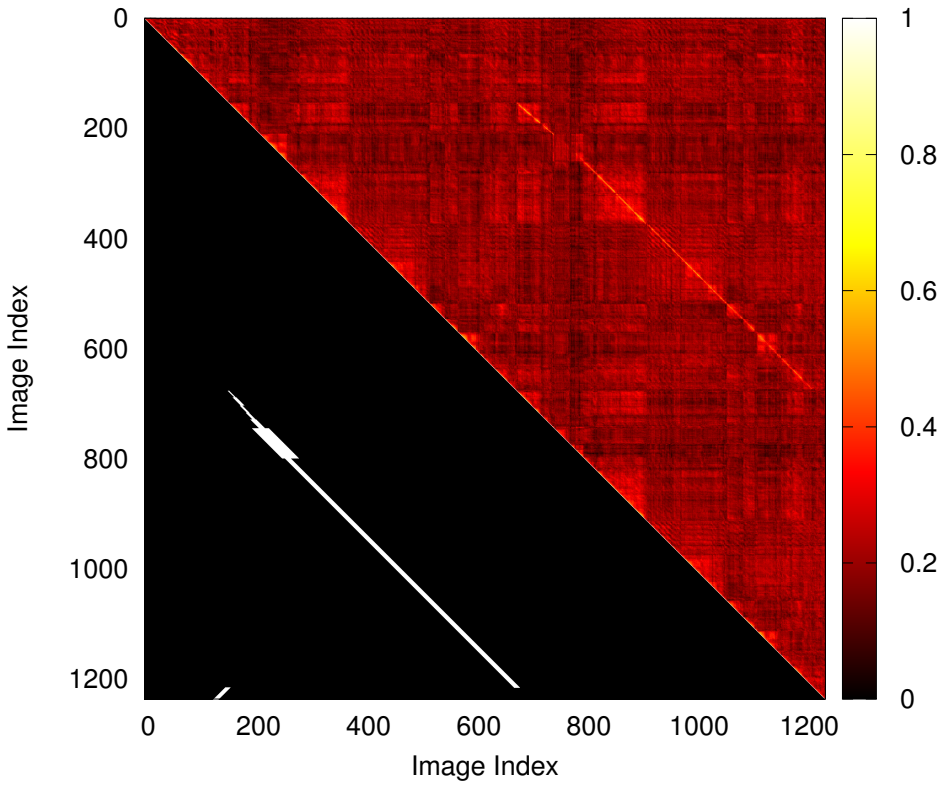


Figure 4.3 : Similarity matrix extracted by using City Center dataset with the setting $k = 32$.

According to the overall performances in the Table 4.1, the best results are obtained with the setting $k = 32$ except for the sequence number 5. However, the result at $k = 32$ in this sequence is considerably high as it is in the other datasets. For this reason, it is easily noticed that the setting with $k = 32$ can be selected for any other datasets, if there is only one setting that must be chosen. Additionally, these settings are going to be used in the subsequent section to compare the method with the other state-of-the-art methods in the literature.

An interesting way of visualizing the locations where loop closures are detected is in similarity matrices as illustrated in Figures 4.2, 4.3, 4.4 and 4.5 which are extracted for the New College, City Center, the sequences number 5 and 13 of KITTI Vision dataset respectively. In the upper triangle of the similarity matrices, the similarities between all images in the sequences are visualized by a color map in which the values close to white indicate high similarity in contrast to the ones close to black. Otherwise, the lower triangle present the actual loop closure locations (i.e. ground truth locations). These matrices are also symmetric with respect to their diagonal on which the values are always the maximum similarity since the image indexes are the same. Therefore,

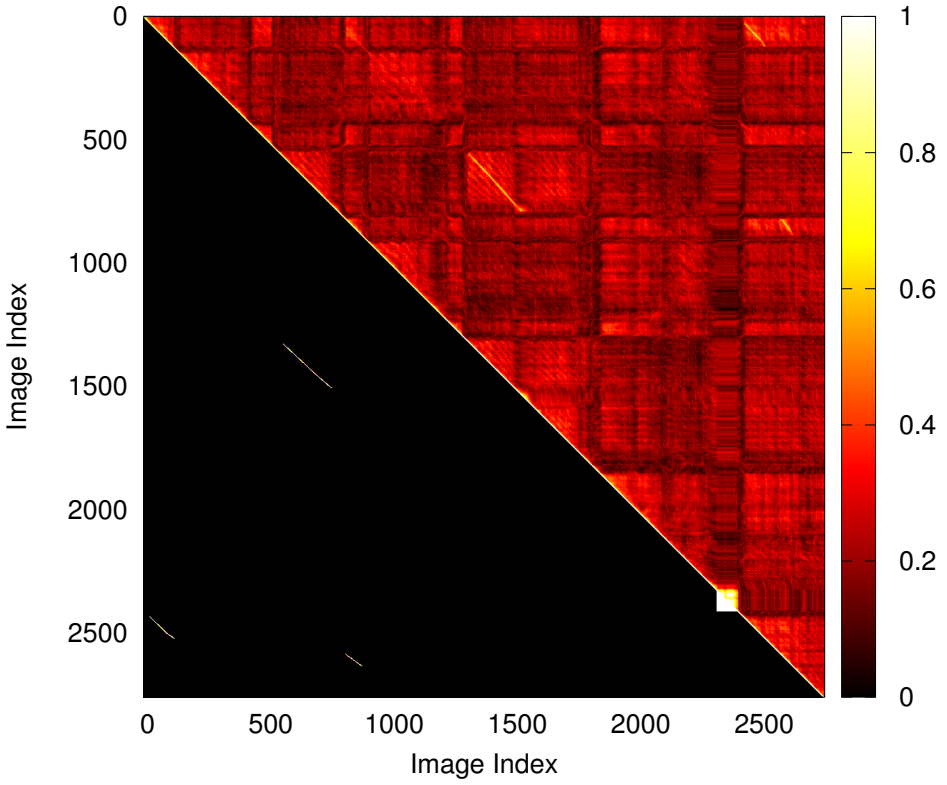


Figure 4.4 : Similarity matrix extracted by using KITTI Vision 05 dataset with the setting $k = 32$.

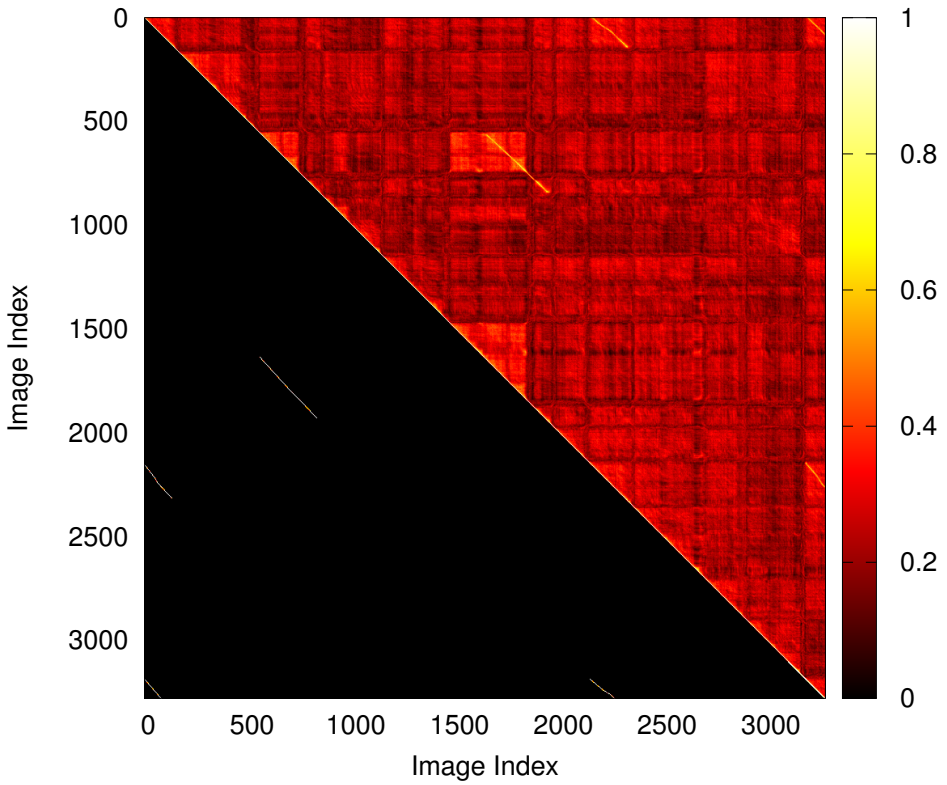


Figure 4.5 : Similarity matrix extracted by using KITTI Vision 13 dataset with the setting $k = 32$.

the loop closure performance can be easily observed by comparing the locations that have high similarity with the actual ones presented in the lower triangle of the matrices.

In Figure 4.2, the similarity matrix extracted from New College dataset is presented. As it can be seen, there exist a lot of white lines in different thicks that are parallel to the diagonal. The thicker lines indicate that a location has more loop closure events than the thinner ones. In the upper triangle, there are yellow traces that corresponds to the white lines in the lower triangle. Figure 4.3 illustrates the similarity matrix extracted from City Center dataset. On the bottom-left corner of the matrix there are bidirectional loop closure events (i.e. the short line that is perpendicular to the diagonal) which the proposed method cannot handle. In addition to this, the tones of loop closing traces on the upper triangle are close to the dominant tone of the area. This means, the loop closure events are hardly distinguished by the proposed method. As in the Figure 4.4, which depicts the similarity matrix of KITTI Vision 05 dataset, the shiny region on the bottom-right corner indicates the loop closure locations where the agent does not move. Additionally, in both KITTI Vision datasets that are used in these experiments the ground truth lines are thinner than the other datasets as it is also seen in Figure 4.5,

which presents the similarity matrix of the sequence number 13. Although the most of the yellow traces cover the actual loop closure locations, the results listed in the Table 4.1 are a bit low because of the thin ground truth lines.

4.3 Comparison with Other Methods

In this section, the proposed method (PALM) is compared with three global descriptors, namely, D-LDB [19], BRIEF-Gist [13] and WI-SIFT [17]. These methods are selected for comparison as (i) they achieve state-of-the-art results and (ii) they have available software. For a fair comparison, same loop closure methodology is used. Also, FAB-MAP [3] is included in the comparison study since it can be the one of the most significant works in visual loop closure detection.

4.3.1 Implementation Details

The methods that are compared with the proposed method are implemented by following the guidelines stated in the corresponding papers [3, 13, 17, 19]. In addition to this, the OpenCV [39] implementations are used to compute SIFT [7] and BRIEF [6] descriptors.

For the first approach [19], the code provided by [40] is used to compute the LDB descriptor which is the base structure of this approach. More precisely, for implementing D-LDB, the input image is resized to 64×64 pixels and the patch size of the LDB descriptor is set to 48 as it is equal to the default value. Then the resulting descriptor vector is computed around the center of the input image. Note that the disparity information involved by D-LDB is not considered due to the datasets not providing this kind of information directly. Furthermore, Hamming distance is used to measure the similarity between the descriptors since LDB is constructed with a set of binary features.

The second technique [13] is called BRIEF-Gist which is based on computing the BRIEF descriptor on the image in a holistic manner. BRIEF-Gist has been implemented by dividing the image into $m \times m$ tiles. Specifically, the image is downsampled to a size of $ms \times ms$ with the patch size $s = 48$ which is the standard patch size of BRIEF descriptor. A keypoint is defined at the center of each patch and then BRIEF descriptors are computed around these keypoints. The final

descriptor vector is constructed by concatenating the m^2 bit-vectors extracted from their corresponding tiles. In experiments, m is set to 7 and BRIEF-32 is used as it is reported in the corresponding paper. The similarity measure for BRIEF-Gist is the conventional Hamming distance as used in D-LDB.

For the other approach [17], WI-SIFT is implemented in a similar way to BRIEF-Gist except for the tiles and descriptors itself which is SIFT. The image is downsized to 128×128 pixels and the Euclidean distance is used for similarity calculation.

Finally, for each method, including the proposed one, the resulting descriptor vector is computed for each image provided by the datasets. In New College and City Center datasets, two separate descriptor vectors are computed and concatenated since these datasets provide two images, which are left and right, for each location. Moreover, only left image is used for the sequences number 5 and 13 of KITTI Vision dataset that provides stereo images.

Additionally, for the state-of-the-art FAB-MAP [3], its open-source implementation OpenFABMAP [41] is used. For KITTI Vision datasets, it is trained with 2800 images that are randomly selected from all the sequences, including the ones that are used in the experiments. At the end, 9691-words vocabulary is created. The default settings file of OpenFABMAP is used for both training and evaluation. For New College and City Center datasets, precision-recall curves are roughly taken from the original paper [3], since evaluation results that are obtained with OpenFABMAP are not similar to the results in the original work.

4.3.2 Results

The metric that is mostly used to evaluate the loop closure detection systems is the precision-recall curve which implies the operating characteristics of the system. A single data point of these curves is obtained by setting a dissimilarity threshold τ as discussed in Section 3.2, in order to determine whether a location closes the loop or not. Consequently, all the points in these curves are obtained by sweeping τ from 0 to the value that makes the recall 100%. The lower values of τ may result in rejection of the false detections and the loop closure events which can be hardly distinguished since their similarities are close to the non-loop closures.

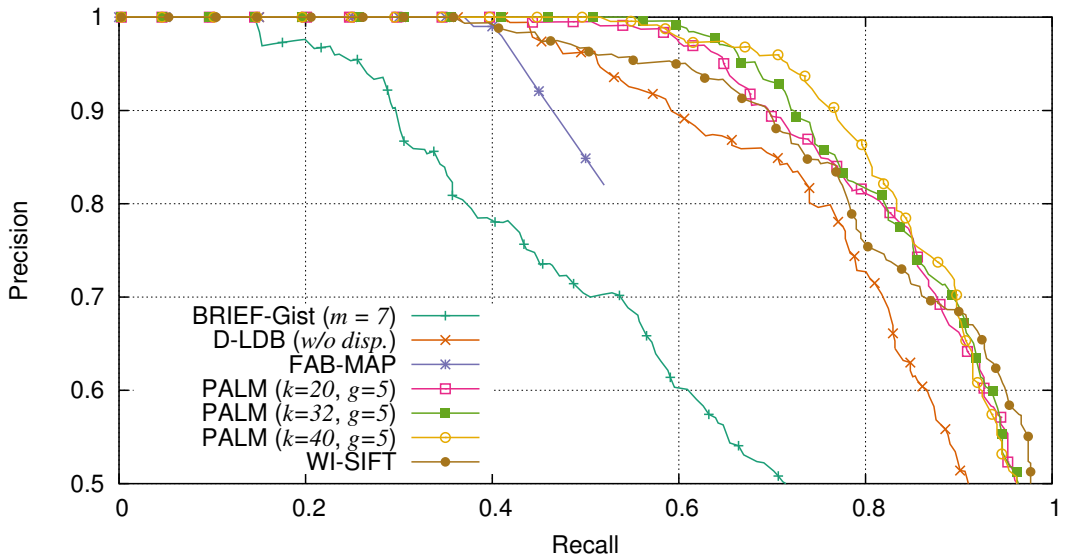


Figure 4.6 : Precision-recall curves of the proposed method against the other state-of-the-art methods for New College.

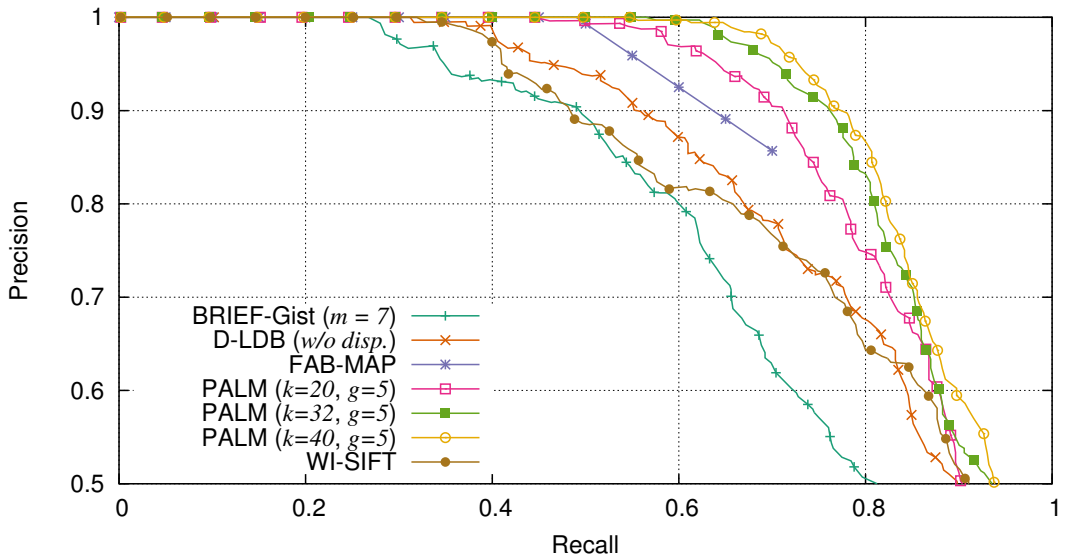


Figure 4.7 : Precision-recall curves of the proposed method against the other state-of-the-art methods for City Center.

In Figure 4.6, precision-recall curves of the proposed method against the other state-of-the-art methods are demonstrated for New College dataset. The proposed method (PALM) with all the settings that are specified, gives the best results. WI-SIFT and FAB-MAP are the second bests and the others especially BRIEF-Gist clearly perform under the average. In contrast to New College dataset, the performance of WI-SIFT decreases to the level of BRIEF-Gist and D-LDB for City Center dataset as it is seen in the Figure 4.7. Despite that, it is obvious that the proposed method outperforms all the other methods and remains nearly the same as the results for New College dataset.

Table 4.2 : Comparison results for the proposed method against the other methods.

	New College		City Center	
Method	Recall @100% Precision	Area Under Curve	Recall @100% Precision	Area Under Curve
FAB-MAP	0.370	–	0.480	–
WI-SIFT	0.358	0.896	0.324	0.831
D-LDB (Without Disparity)	0.240	0.825	0.357	0.808
BRIEF-Gist ($m = 7$)	0.148	0.695	0.273	0.780
PALM ($k = 20$)	0.441	0.904	0.446	0.884
PALM ($k = 32$)	**0.552**	0.911	**0.565**	0.902
PALM ($k = 40$)	0.521	**0.916**	0.551	**0.911**

There is a new metric called Area Under Curve (AUC) that appears with the precision-recall curves. This metric is used to present the overall performance of the system. Higher values of AUC indicate better performance in terms of the amount of the loop closures that are covered by the system. Using this metric the final comparison results are listed in Table 4.2 for the New College and City Center datasets. Each comparison made for a single dataset (i.e. a major column in the table) includes two result entries which are the recall rate at 100% precision and AUC respectively. As it can be seen, the proposed method notably outperforms the other methods in terms of either recall rates at 100% precision or AUCs for all parameter settings in both dataset comparisons.

The other comparisons are experimented by using the sequences number 5 and 13 of the KITTI Vision dataset as they are illustrated in Figures 4.8 and 4.9 respectively. As it can be seen, the difference between precision-recall curves are not that obvious as the ones obtained by using New College and City Center datasets. The reason is that the other methods are not good at dealing with perceptual aliasing and dynamic objects in comparison to the proposed method since New College and City Center dataset contain more images presenting these kind of challenges than the KITTI Vision sequences. Despite the fact, the proposed method slightly better than the other state-of-the-art methods as shown in the Table 4.3. BRIEF-Gist gives considerably high recall rates at 100% precision which are slightly lower than the results of the proposed method with the setting $k = 20$ for the KITTI Vision 05 dataset.

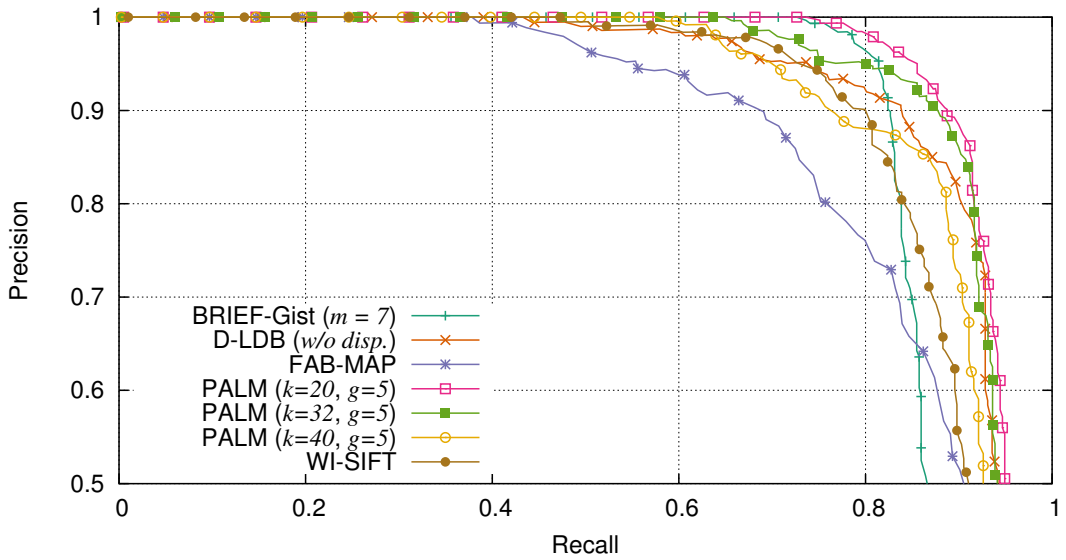


Figure 4.8 : Precision-recall curves of the proposed method against the other state-of-the-art methods for KITTI Vision 05.

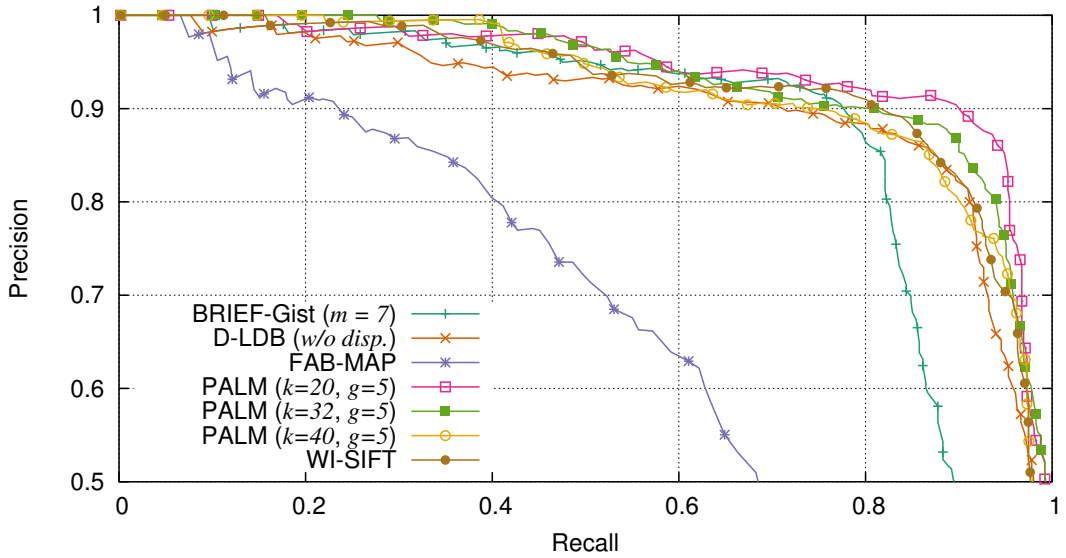


Figure 4.9 : Precision-recall curves of the proposed method against the other state-of-the-art methods for KITTI Vision 13.

As it is stressed earlier, the patch size k is the most significant parameter that effects the results. It must be selected carefully for different environments. There are two types of datasets that are used in the experiments. New College and City Center are recorded similarly and the images are taken from the complete rear views of the robot. On the other hand, KITTI Vision datasets contain images that are recorded from the direction of displacement. This means, the speed of motion in the images in New College and City Center is greater than the one in KITTI Vision. This situation might be the reason of obtaining different results with different k values. In addition to this, small differences in the performances is important for a visual loop closure detection system

Table 4.3 : Comparison results for the proposed method against the other methods.

Method	KITTI Vision 05		KITTI Vision 13	
	Recall @100% Precision	Area Under Curve	Recall @100% Precision	Area Under Curve
FAB-MAP	0.378	0.863	0.066	0.652
WI-SIFT	0.473	0.884	0.157	0.921
D-LDB (Without Disparity)	0.371	0.806	0.052	0.865
BRIEF-Gist ($m = 7$)	0.726	0.890	0.097	0.869
PALM ($k = 20$)	**0.744**	**0.939**	0.169	**0.941**
PALM ($k = 32$)	0.652	0.931	**0.279**	0.934
PALM ($k = 40$)	0.585	0.907	0.222	0.920

since a single false detection may effect the general SLAM system in a catastrophic way [3, 13]. Therefore, the parameters must be selected carefully to be the best in the environment. However, the setting with $k = 32$ can be adapted to any environment, if false detections are not that important according to the application.

To sum up, according to the evaluation results that have been experimented by using well-known outdoor datasets, the proposed loop closure detection method outperforms the other state-of-the-art global description methods — namely, BRIEF-Gist, D-LDB, WI-SIFT — plus FAB-MAP, in terms of either recall rates at 100% precision or AUC values. The next section presents the speed performance of the proposed method and gives informations about the usage in large-scale operations.

4.4 Speed Performance

It is important to see how long the computation of the loop closure method takes with respect to the number of frames growing incrementally. The tests are done by using a MacBook Pro 12.1 that has Intel Core i5 2.7 GHz CPU.

There are two main processes in this loop closure detection system: 1) Descriptor extraction process involving the computation of PALM descriptor for the most recent image, 2) the matching process in which the loop closing hypotheses is cast. Figure 4.10 demonstrates the speed performance of the proposed method in the New College dataset which was collected in 1.9 km long trajectory. As it can be seen, most of the time was spent to extraction of the descriptor vector consuming 1.6 ms on average. However, the computational overhead of the overall loop closure detection process

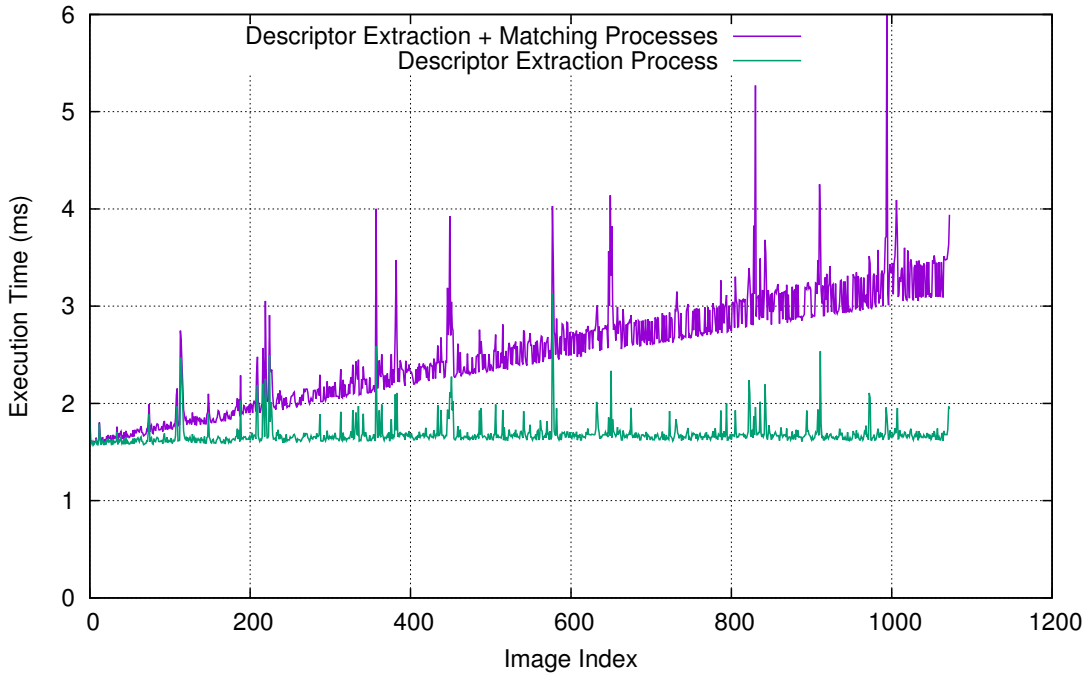


Figure 4.10 : Performance of the proposed method as the number of images increase while evaluating New College dataset.

containing both processes increases linearly since the time complexity of the NN classifier is $\mathcal{O}(dm)$ — d is the descriptors size and m indicates the total images in a sequence. Despite the linear time complexity, this method consumes approximately 3.5 ms in a 1.9 km (1073 frames) long trajectory. The performance results show that the proposed loop closure detection method is capable of operating in real-time[1] up to 15 km ($\sim$8000 frames) long trajectories.

[1]Execution in real-time is considered that the processing of a single frame should be completed within 30-40 milliseconds.

5. CONCLUSIONS AND FUTURE WORKS

In this thesis, a novel visual loop closure detection method was presented. The problem of loop closure detection is considered a part of the SLAM problem, and the proposed system was developed within this context. However, the developed system does not depend on the SLAM estimations.

5.1 Conclusions

As it was explained in the introduction chapter, loop closure problem is a very complex problem which is quite challenging in various aspects. All of the following criteria are major issues of concern in the context of vision-based loop closure detection problem: efficiency, avoiding false detections and the issues arising from the visual sensory. Developing a descriptor that represents the places with a high discrimination power is the most essential component to achieve reliable detections. Presence of perceptual aliasing and illumination variations on the environment makes the loop closing problem even more challenging.

The main idea of the proposed system is extracting shape features in a local manner and representing them in an efficient way as a global image descriptor. By using the descriptors computed from the images, loop closure detection is cast by a simple matching via brute-force image search. The methodology was described in the previous chapters.

The main contribution of this thesis is the fast and efficient description method with high discrimination power. A visual loop closure system is developed using the description method. It was demonstrated that the performance of the presented system outperforms the other state-of-the-art methods in terms of image description. The system can be operated in real-time even for long sequences. On the other hand, having linear time complexity is the main drawback of this loop closure detection method.

5.2 Future Work

The method presented in this thesis describes a novel loop closure detection system, which is subject to be implemented into a real-world SLAM application. However, before it can be implemented on an actual SLAM system, it requires several improvements.

One of the most crucial improvements that should be carried out is developing a more efficient image matching technique. This can be achieved by replacing the brute-force NN classifier with a more intelligent one to detect loop closure candidates efficiently. It is obvious that any improvement on the loop closure detection methodology is going to increase the overall performance of the method.

The second most significant improvement would be parallelizing the computation of the descriptor vector to be processed on a GPU or on a multi-processor system. Since the extraction of localized moments is independent from each other, this parallelization should be fairly straightforward when the right tools are used.

REFERENCES

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

[2] **Csurka, G.**, **Dance, C.**, **Fan, L. and Willamowski, J.** (2004). Visual categorization with bags of keypoints, *In Workshop on Statistical Learning in Computer Vision, ECCV*.

[3] **Cummins, M. and Newman, P.** (2008). FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance, *The International Journal of Robotics Research*, *27*(6), 647–665.

[4] **Sariyanidi, E.**, **Sencan, O. and Temeltas, H.** (2012). An image-to-image loop-closure detection method based on unsupervised landmark extraction, *2012 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, pp.420–425.

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

[6] **Calonder, M.**, **Lepetit, V.**, **Strecha, C. and Fua, P.**, (2010). BRIEF: Binary Robust Independent Elementary Features, Computer Vision – ECCV 2010, Springer Berlin Heidelberg, Berlin, Heidelberg, pp.778–792.

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

[8] **Bay, H.**, **Tuytelaars, T. and Van Gool, L.**, (2006). SURF: Speeded Up Robust Features, Computer Vision – ECCV 2006, Springer Berlin Heidelberg, Berlin, Heidelberg, pp.404–417.

[9] **Cadena, C.**, **Galvez-Lopez, D.**, **Tardós, J.D. and Neira, J.** (2012). Robust Place Recognition With Stereo Sequences, *IEEE Transactions on Robotics*, *28*(4), 871–885.

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

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

[12] **Singh, G. and Kosecka, J.** (2010). Visual loop closing using gist descriptors in manhattan world, *ICRA Omnidirectional Vision Workshop*.

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

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

[15] **Badino, H.**, **Huber, D. and Kanade, T.** (2012). Real-time topometric localization, *2012 IEEE International Conference on Robotics and Automation (ICRA)*, 1635–1642.

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

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

[18] **Lategahn, H.**, **Beck, J.**, **Kitt, B. and Stiller, C.** (2013). How to learn an illumination robust image feature for place recognition, *IEEE Intelligent Vehicles Symposium*, IEEE, pp.285–291.

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

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

[21] **Negre Carrasco, P.L.**, **Bonin-Font, F. and Oliver-Codina, G.** (2015). Global image signature for visual loop-closure detection, *Autonomous Robots*, 1–15.

[22] **Oliva, A. and Torralba, A.** (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, *International Journal of Computer Vision*, *42*(3), 145–175.

[23] **Oliva, A. and Torralba, A.**, (2006). Chapter 2 Building the gist of a scene: the role of global image features in recognition, Visual Perception - Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception, Elsevier, pp.23–36.

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

[25] **Yang, X. and Cheng, K.T.T.** (2014). Local Difference Binary for Ultrafast and Distinctive Feature Description, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(1), 188–194.

[26] **Sariyanidi, E.**, **Gunes, H.**, **Gokmen, M. and Cavallaro, A.** (2013). Local Zernike Moment Representation for Facial Affect Recognition, *British Machine Vision Conference 2013*, British Machine Vision Association, pp.108.1–108.11.

[27] **Sariyanidi, E.**, **Dagli, V.**, **Tek, S.C.**, **Tunc, B. and Gokmen, M.** (2012). Local Zernike Moments: A new representation for face recognition, *2012 19th IEEE International Conference on Image Processing (ICIP 2012)*, IEEE, pp.585–588.

[28] **Sariyanidi, E.**, **Sencan, O. and Temeltas, H.** (2013). Loop closure detection using local Zernike moment patterns, ***J. Röning and D. Casasent**, editors, IS&T/SPIE Electronic Imaging*, SPIE, p.866207.

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

[30] **Erhan, C.**, **Sariyanidi, E.**, **Sencan, O. and Temeltas, H.** (2015). An online visual loop closure detection method for indoor robotic navigation, *Proc. SPIE 9406, Intelligent Robots and Computer Vision XXXII: Algorithms and Techniques*, SPIE, p.940607.

[31] **Teh, C.H. and Chin, R.T.** (1988). On image analysis by the methods of moments, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10*(4), 496–513.

[32] **Chen, Z. and Sun, S.** (2010). A Zernike Moment Phase-Based Descriptor for Local Image Representation and Matching, *IEEE Transactions on Image Processing*, *19*(1), 205–219.

[33] **Singh, C.**, **Walia, E. and Mittal, N.** (2011). Rotation invariant complex Zernike moments features and their applications to human face and character recognition, *IET Computer Vision*, *5*(5), 255–266.

[34] **Li, S.**, **Lee, M.C. and Pun, C.M.** (2009). Complex Zernike Moments Features for Shape-Based Image Retrieval, *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *39*(1), 227–237.

[35] **Jarrett, K.**, **Kavukcuoglu, K.**, **Ranzato, M.A. and LeCun, Y.** (2009). What is the best multi-stage architecture for object recognition?, *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, IEEE, pp.2146–2153.

[36] **Dalal, N. and Triggs, B.** (2005). Histograms of oriented gradients for human detection, *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *1*, 886–893 vol. 1.

[37] **Viola, P. and Jones, M.** (2001). Rapid object detection using a boosted cascade of simple features, *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, IEEE Comput. Soc, pp.I–511–I–518.

[38] **Crow, F.C.** (1984). Summed-area tables for texture mapping, *The 11th Annual Conference*, ACM Press, New York, USA, pp.207–212.

[39] **Bradski, G.** (2000). *OpenCV computer vision library*, Dr. Dobb's Journal of Software Tools.

[40] **Yang, X. and Cheng, K.T.** (2012). LDB: An ultra-fast feature for scalable Augmented Reality on mobile devices, *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, IEEE, pp.49–57.

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

CURRICULUM VITAE

Name Surname: Can Erhan

Place and Date of Birth: Gölcük, 1989

E-Mail: erhanc@itu.edu.tr

B.Sc.: Istanbul Technical University, Faculty of Science and Letters, Physics Engineering, 2012

Professional Experience:

- R&D Software Engineer in Borda Technology (2013 -)

Rewards:

- 1^{st} Rank Student in B.Sc.
- Scholarship for M.Sc. Students, Ministry of Industry and Technology of Turkey and Temsa Global A.S. (SAN-TEZ)
- 3^{rd} Best Application Award on SIU 2013

List of Publications:

- **Erhan, C.**, Tazehkandi, A. A., Yalcin, H., and Bayram, I. (2013). Traffic sign detection and recognition fusing feature descriptors. *IEEE 21st Signal Processing and Communications Applications Conference (SIU 2013)*, April, Cyprus, Turkey.

PUBLICATIONS/PRESENTATIONS ON THE THESIS

- **Erhan, C.**, Sariyanidi, E., Sencan, O. and Temeltas, H. (2015). An online visual loop closure detection method for indoor robotic navigation, *Proc. SPIE 9406, Intelligent Robots and Computer Vision XXXII: Algorithms and Techniques, SPIE*, p.940607.