SOFT ASSIGNMENT VISUAL DESCRIPTORS FOR VISUAL PLACE RECOGNITION

ABBAS M. ALI

THESIS SUBMITTED IN FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2013

UMPUKAN LEMBUT PEMERIHAL VISUAL UNTUK PENGECAMAN TEMPAT SECARA VISUAL

ABBAS M. ALI

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH
DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2013

DECLARATION

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

27th June 2013

ABBAS M. ALI
P47845

ACKNOWLEDGMENT

In the name of Allah, Most Greatest, Most Gracious, Most Merciful and
Alhamdulillah with all praises to Allah.

First and foremost, I would like to thank Allah in this day for giving me the prudence, strength and inspiration along with my research journey. For everything I have, I feel blessed, Alhamdulillah.

I would like to offer my sincerest gratitude with appreciation to my supervisor Assoc. Prof. Dr. Md. Jan Nordin for his supervision, guidance and the invaluable help with all encouragements. Special thanks with indebted goes to my co-supervisor Assoc. Prof. Dr. Abdullah Azizi for his unflinching support, advising and the valuable help of constructive suggestions throughout the experimental and thesis works have contributed to the success of this research. Not forgotten, my thanks to Dr. Siti Nuralhuda for providing me some valuable notes and supports through some seminars and workshops held in UKM. Highly appreciation to all lecturers, friends, staffs and workers in UKM, especially FTSM for providing me everything that I needed much greater in all the years I have been here. I am grateful for being in this amazing country experiencing life with those friendly Malaysian people.

My special thanks and gratitude are due to my great parents for their morale support and encouragement. Finally, I would like to thank my wife for her continued patience and support without which I would not have been able to complete this research.

ABSTRACT

Upon increasing the popularity of using the Hard Bag of Features (HBOF) for accurate object and place categorization problem, there are some issues which are still being scrutinized. In fact, most of the previous researches in place recognition area are based on using Histogram descriptors. Based on the literature, these methods have several issues such as the inability to include spatial relation among the local appearance features for representing the scene image in a more informative way. Therefore, the main objective of this research is to improve the performance of the HBOF in visual place recognition by developing spatial relations for Soft assignment features. These features extracted by measuring the distances of patches from the centroids of codebook constructed by clustering SIFT features by K-means. The covariance of minimum distance (CMD) with whitening filters and some normalization parameters are used to increase the accuracy performance. The visual place confusion has been decreased by implementing Entropy of covariance feature vectors (ECV) which is investigated alone and combined with the edge histogram descriptors (EHD) using the conceptual semantic representation. To demonstrate the effectiveness of the proposed approaches in visual place recognition, several experiments have been setup such as CMD, ECV, and Semantic in order to evaluate the accuracy rate. Practically, several comparative studies were conducted with other related approaches namely a conventional BoW or HBOF, Minimum distance table, and Covariance of Distance table. The proposed methods have been evaluated based on different datasets such as IDOL and on real images from a handheld camera taken for some places in FTSM-UKM. Based on the obtained results, the combined features of EHD and ECV bring a significant improvement in the classification accuracy rates up to 98.6% and 93.423% for IDOL and FTSM-UKM dataset respectively.

ABSTRAK

Penambahan populariti menggunakan Hard Bag of Features (HBOF) untuk masalah pengkategorian objek dan tempat dengan tepat mempunyai beberapa isu yang masih dikaji. Kebanyakan daripada penyelidikan sebelum ini dalam bidang pengecaman tempat berdasarkan penggunaan pemerihal histogram. Berdasarkan kesusasteraan, kaedah ini mempunyai beberapa isu seperti ketakdayaan untuk merangkum hubungan ruang antara sifat rupa tempatan bagi mewakili imej pemandangan dalam bentuk yang lebih informatif. Oleh itu, objektif utama penyelidikan ini ialah untuk meningkatkan prestasi HBOF dalam pengecaman tempat visual dengan membangunkan hubungan ruang bagi sifat umpukan lembut. Sifat ini disari dengan mengukur jarak tompok dari sentroid kod buku yang dibina dengan pengelompokan sifat SIFT oleh pendekatan cara K. Kovarians jarak minimum (KJM) dengan penapis pemutih dan beberapa parameter penormalan digunakan untuk meningkatkan ketepatan prestasi. Kekeliruan tempat visual telah dikurangkan dengan melaksanakan entropi vektor kovarians sifat (EVK) yang dikaji tersendiri dan menggabungkan dengan pemerihal histogram pinggir (PHP) menggunakan perwakilan semantik konsepsual. Untuk menunjukkan keberkesanan pendekatan yang dicadangkan dalam pengecaman tempat visual, beberapa eksperimen telah dilaksanakan seperti KJM, EVK, dan Semantik bagi menilai kadar ketepatan. Beberapa kajian perbandingan telah dijalankan dengan beberapa kaedah yang berkaitan seperti BoW atau HBOF konvensional, jadual jarak minimum dan jadual kovarians jarak. Kaedah yang dicadangkan telah dinilai menggunakan set data berbeza seperti IDOL dan imej sebenar dari satu kamera pegang di beberapa tempat dalam FTSM-UKM. Berdasarkan keputusan yang diperolehi, sifat gabungan daripada PHP dan EVK memberi peningkatan yang ketara bagi kadar ketepatan pengelasan, iaitu sebanyak 98.6% dan 93.423% apabila diuji terhadap set data IDOL dan FTSM-UKM secara tertib.

CONTENTS

CHAPTER VII VARIOUS APPROACH TECHNIQUES FOR VISUAL PLACE RECOGNITION

CHAPTER VIII CONCLUSION AND FUTURE WORKS

LIST OF TABLES

Table No. **Page**

LIST OF FIGURES

Figure No. **Page**

LIST OF ABBREVIATIONS

BoW	Bag of Words
CALTECH	CALifornia TECHnology dataset
CBIR	Content Base Image Retrieval
CMD	Covariance of Minimum Distance
CENTRIST	CENsus TRansform hISTogram
COLD	COsy Localization Database
COD	Covariance Of Distance
COB	Combined features
CRFH	Composed Receptive Field Histogram
CT	Census Transform
DCM:	Dirichlet Compound Multinomial
DOG	Different Of Gaussian
Dt	Distance Table
ECV	Entropy of Covariance feature Vector
EHD	Edge Histogram Descriptor
EOH	Edge Orientation Histogram
Er	Eigen Values
Ev	Eigen Vectors
GIST	Global Image Texture Features.
GPS	Global Positioning System
FTSM	Fakulti Teknologi dan Sains Maklumat
HBOF	Hard Bag Of Features
HMM	Hidden Marcov Model
HOG	Histogram Of Gradient
IDOL	Image Database for rObot Localization

k-NN	k Nearest Neighbor
LBP	Local Binary Pattern
Lm	Land Mark
MDT	Minimum Distance Table
MSER	Maximally Stable Extremal Regions
PCA	Principle Component Analysis
PLISS	Place Labelling through Image Sequence Segmentation
RBF	Radial Basis Function
RFID	Radio-Frequency Identification
SIFT	Scale Invariant Features Transformation
SLAM	Simultaneous Localization and Mapping
STDEV	Standard Deviation
SURF	Speed Up Robust Features
SVM	Support Vector Machine
TF-IDF	Term Frequency Inverse Document Frequency
UNC	Uncertainty Codeword
VSLAM	Visual Simultaneous Localization and Mapping
WEKA	Waikato Environment for Knowledge Analysis

CHAPTER I

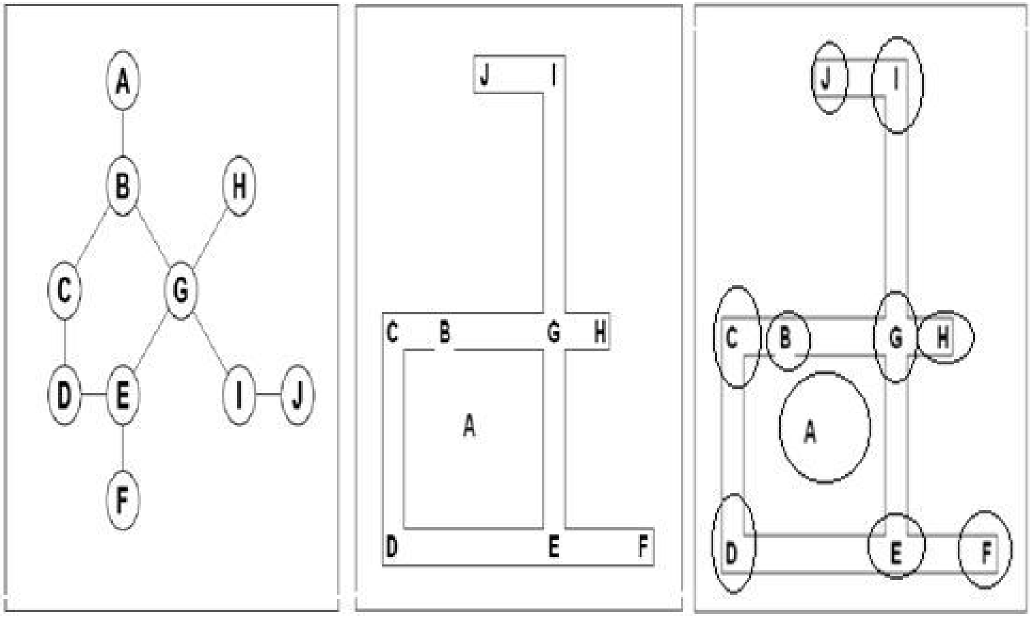
INTRODUCTION

1.1 INTRODUCTION

We, human beings, always ask ourselves or others a frequently used question. It is a very important question: “Where am I?" or “where are you?” The answer is directly responded by our brain when we just see the scene in the place. This answer may depend on the recognized image and its content features. If the same question asked for the mobile robot, the situation will be different. The question needs many processes for the query image which is captured by the camera. And the answer is often needs category of places such as “bedroom, corridor or kitchen". It is clear that the way of processing the query image by the two sides (human and robot) is different. Robots cannot process the visual information quickly.

The problem is still unsolved when dealing with robotic visual place recognition and robot localization. With the development of the technology and robotics, the robot found its way to enter into our daily home lives for example they do some human jobs in a limited way such as cleaning and washing. One of the human ambitious aims is that to create a robot to help at home and to do some jobs as a family member. The manufacturers of this device to try to include the latest inventions and accessories to serve the human in many fields. One of the interesting areas by the researcher nowadays is how to make the mobile robots or intelligent agents to recognize or categorize visual places in order to work with the human in daily home works (Kruijff et al., 2007; Topp et al., 2006). This would need techniques to facilitate robot-human interaction to overcome some difficulties that faces the robot in some cases.

There are several techniques and approaches such as Simultaneous Localization and Mapping (SLAM) algorithms that can be used for robot localizations and place recognition. The accurate place recognition is highly important for the SLAM algorithm. The landmarks are always used to increase the accuracy of the localization, where sometimes a landmark represents a node within the topological map which contains many nodes and edges to represent the environment by a graph-based structure. The edge shows the connectivity between these nodes which are representing places. Each place has some landmarks which are slightly different from other places and the robots recognize them as distinctive places during place recognition process. This is a useful way to avoid any confusion with the places during the navigation process for the robot. However there are some other ways such as metric mapping, semantic mapping and combined of metric with topology mapping (hybrid mapping). Where metric mapping is scaling the space representing the real environment (Ranganathan & Dellaert, 2011), while semantic mapping is using a conceptual graph representation of the environment, this will be explained in Chapter II. Figure 1.1 shows types of mapping.



A: Topological mapping B: Metric mapping C: Semantic mapping

Figure 1.1 Types of mapping

1.2 PLACE RECOGNITION AND CLASSIFICATION

Place classification is a pattern recognition which assigns local features in an environment to one of the predefined categories based on discriminating properties. In other words, the key element of the place recognition is looking for similar images for the places in the database that look like in the input queried image. Place recognition depends on the descriptors which represent the image scene. Due to the similarity of these descriptors for different categories of place images, these descriptors need to be discriminating from each other to retrieve images accurately. The highly discriminating descriptors give high accuracy to place recognition process. On the other hand the heart of the problem with the place recognition is that the image query for the environmental place needs to be compared with the millions of stored images in the database and thus it will take a long time to retrieve the similar images (Botterill, 2010). This time is very important for the mobile robot or the visual impaired. The system of place recognition checks the query image against current location images and its immediate neighbor images. This helps the procedure in the increasing matching process, and also the localization task can be achieved by avoiding wrong matches. Bag of Words (BoW) techniques are popular image retrieval systems based on histograms of features patches for the matching process. The histogram gives some important properties for BoW, characterized by (Takeuchi & Hebert, 1998):

1- Image histogram requires less memory space and provides small representation of an image, which leads to faster comparison than two raw image comparison.

2- Image histogram is a property that greatly reduces the number of reference images, and the speed of place recognition system improves, because the comparisons of the input images need fewer reference images. Furthermore image histograms are invariant to rotation of the images around the optical axis.

3- Image histograms are invariant to image translation which is very useful in reduction of reference images.

The scene recognition sometimes is identified as place classification or categorization, where it is the process of recognizing the semantic label of a scene when a class of places requested. The scene class recognition is obtained by analyzing and understanding the scene. The label of a scene denotes the types of objects in that scene. For instance, in the kitchen it is much more likely to find refrigerator, cooker, or cylinder of gas and so on. Mostly place classification algorithms, using a finite set of place classifications, require a lot of learning. Almost the labels are learned offline in a supervised mode based on a set of training data with manually labels. According to these labels the system can learn how to group the places. The classifier like Support vector machine (SVM) or k-Nearest Neighbour (k-NN) separates and categorizes input into their corresponding labels based on the previously learned data.

Place recognition is a process of consistent labeling a place as the same when a particular place is being revisited (Ranganathan, 2010). The process does not need semantic labeling of the place so it does not need to understand the whole scene. Almost all of the existing place recognition approaches need training. This can be established by choosing some measurements from a specific environment and manually labelling them based on their contents, and then the system needs to be evaluated by testing different scenes related to the same environment.

In general, place recognition is closely related to both topological place recognition and global localization which help to find the rough location (e.g., Kitchen for instance, that is, the robot knows the place, but it doesn't know it's accurate positions in the kitchen). In addition, the robot's exact pose is determined through topological place recognition and global localization (Jianxin & Jim, 2010). Also the supervised place recognition is similar to placing categorization; it is an easy task because the learning and testing measurements of the images from the same places are quite similar.

However the supervised systems for recognition and classification have simplicity property which is an advantage, they contains some drawbacks that will be listed in this section. The main drawbacks of the existing approaches are related to the labeling, where no interpolation for labelling create new labels.

The best algorithm for place recognition needs to consider the dynamic changes in the environment. The algorithm should be designed and created in such a way to process the different images of the same place as one label, there differ images of the same place can be resulted from illumination changes, pose for the camera, removed or shifted obstacles, and so on, as shown in Figure 1.2.

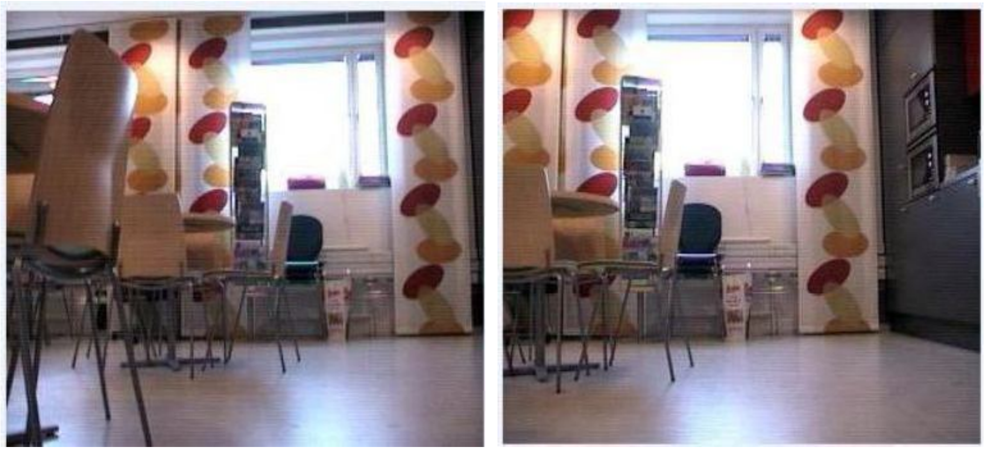


Figure 1.2 Two different views of the same scene.

Other drawbacks of recognition and classification of the places (Ranganathan, 2010; Fei-Fei & Perona, 2005) are listed below:

1. Since there is a large variation in the measurements as each labeled scene has its own different representations, so the classifier needs to learn a large amount of labeled training data. This is time consuming, difficult and expensive, since it is a manual work where labels are assigned to scenes and images by hand.

2. The labels are defined by an expert are somewhat arbitrary so they are suboptimal.

3. To learn the different labels by the classifier in the best possible manner, the main characteristics of the underlying scene need to be included in each training data set. The new data which are used for testing the system must also have the same main characteristics. This needs supervision from the human in the process by recording the data, and making the use of continuous measurements which is almost impossible.

4. Recognizing and adding new categories is impossible, since a fixed number of different labels for the training data are used and a system will classify new measurements according to these fixed labels.

5. Each measurement which is individually classified by the system does not make decisions based on recently seen measurements. With regard to supervised classification, there is an important issue. When using supervised classification, the classification should be applicable across a wide range of spatial environments. Otherwise it is impossible to do the accurate semantic labeling for places which are not visited before.

As mentioned earlier, there has been a growing body of research focusing on robot localization and place recognition. Research in this area has yielded different techniques for place recognition (Coughlan et al., 2006; Hesch & Roumeliotis, 2010; Ran et al., 2004; Hub et al., 2004). The objective of place recognition algorithm is to identify places according to some determined cues. There are several techniques used to establish the mission of place recognition by using RFID information grids, bar codes, sensors and Wi-Fi (Willis & Helal, 2005; Sonnenblick, 1998; Coughlan et al., 2006).

The challenges face most of these techniques presumably are the same such as scalability, the cost, and the lack of orientation information for the user. In this thesis work, the researcher aims to use techniques of computer vision to propose an approach in visual place recognition to solve some problems faced in this field. Computer vision techniques are increasingly researched in the development of scene understanding. The major concept of computer vision technique in this field has one goal which is to identify the visual appearance for places. Practically, this goal has compromised many applications like topological mapping, indoor and outdoor navigation and loop closure issue (Cummins & Newman, 2008) and it is usually related to the image retrieval task. Although the indoor place recognition has been heavily searched, it is still facing some challenges which are not easy to be solved in this area. This thesis work is looking to propose and enhance the current state of indoor place recognition techniques which may be used in place recognition for indoor navigation system that help the visually impaired or mobile robotics to know directions to their desired destinations in GPS-denied indoor environments.

1.3 MOTIVATION OF THE RESEARCH

Visual place recognition is very important for mobile robotics in indoor areas for many purposes, either it is the interaction between the human for the homeworks or it is for helping the visually impaired to overcome their difficulties to go to any place needed. Besides, the developments of our knowledge in many fields especially in technology like robotic, it is essential and important to introduce these robotic devices to home lives instead of the animals, and they can be exploited in some daily home works (Stachniss et al., 2006; Zender et al., 2007, Galindo et al., 2008). The development of robotic science and systems plays an important role in technology industries as the researchers and the companies try to focus on users' requirements for indoor applications. In addition, this study gives some solutions for challenging cases that have not been solved completely, these cases are : (Orabona et al., 2007).

1. The image of the same place changes because illumination has been changed because of changing the light in the room and so on.

2. The image of the same place changes because of moving some objects or (obstacles) get in the way.

3. The image of the same place changes because of change the side view of the camera so the input space is huge, and lots of images need to be analyzed, and recognized online to be suitable for the robot for their navigation or localization process.

1.4 PROBLEM STATEMENT

Place recognition is one of the highly challenging problems in computer vision. The problem can be more clear and tangible when it concerns the intelligent mobile robots in indoor environments. The indoor environment has its own characteristics, for example the type of room (kitchen, bedroom, living room, corridor), the position and orientation of the camera (camera pose), and the number of cameras (mono, stereo). The accurate place recognition is a challenging task, where it is extremely important that self-localization be enforced precisely).

a) One of the known disadvantages of BoW is that it ignores the spatial relationships among the patches, which is very important in image representation.
b) Soft assignment features also ignore the spatial relationships among the patches since the patches are unordered detected in the images
c) The arrangement of the distances inside the soft assignment features are also not based on any rules or spatial relation between these distances.
d) Using Single descriptors like soft assignment features might not be enough in describing the scene image to consider all the required conditions.

The topic is widely researched, and many algorithms have been proposed, among which the most widely used is the environment representation (Brunskill & Roy, 2005). The main and essential approach used for the place recognition task is that to give some encoding for regions of the place scene as a descriptor that can be stored and used to recognize such places.

This study concerns with investigating visual place recognition approaches based on modern techniques for describing features in order to represent an image's place and state-of-the-art classification methods for learning. In more details, the visual place classification involves a process of determining which class a specific visual place belongs to. The research consists of applying various machine vision techniques and a machine learning algorithm to extract informative descriptions for accurately classifying visual places. The proposed algorithms will include the following parts:

- Spatial feature description: The image content will be characterized using spatial information signatures, not only by local features. The spatial approach represented by the covariance of soft assignment instead of using hard assignment. This approach will offer several advantages over the local feature basis. One of the advantages is reducing confusion matching of the visual places.

- Visual Place categorization: The algorithms will be able to give a certain degree of generalization by omitting certain details for descriptions of labeled training image places for each test group.

- New spatial image representation: The image content will be represented by stable spatial information and combined with other features to represent the scene image. This will be tried by entropy and edge orientation histogram.

- The semantic of visual place representation of the environment to increase the accuracy performance of the system.

1.5 RESEARCH OBJECTIVES

The primary aim of this work is to enhance the performance of Hard Bag of Features (HBOF) for visual place recognition by utilizing the soft assignment model. The soft assignment model has been utilized to include spatial relation among the entire content of the visual words. In order to achieve this aim, several specific objectives had to be established and these objectives are stated as follows:

1. To improve the performance of Hard Bag Of Features (HBOF) for visual place recognition by utilizing the Soft Assignment approach through introducing spatial relations for the visual words

2. To construct stable spatial features for the visual place recognition because using these features may improve the accuracy performance.

3. To improve the accuracy performance of recognition by using several descriptors instead of a single descriptor in representation of the visual places.

1.6 SCOPE OF THE RESEARCH

This research focuses on implementing and proposing approaches for visual place recognition and categorization using spatial relations among the visual words for the Soft assignment model to increase the accuracy performance of HBOF model. The nature of the visual place recognition for robotics is considered as indoor visual place recognition, since the work focusing on some dataset for indoor navigation, therefore outdoor visual places are not considered in this study. The following is the scope of the study:

1. Extraction of discriminate local features for visual places by using SIFT grid, where a grid of 30×30 blocks used for each image scene.

2. Clustering of the local features using k-means algorithm for all images entered into the query process and stored set of images.

3. Calculating Soft assignment features then evaluate some extra operations to extract feature vectors.

4. Calculating the performance accuracy of the resulted classification and recognition.

5. Datasets: The set of images of the indoor places and objects is obtained from more than three sources, which are:

- CALTECH101: This database, was introduced at the California Institute of Technology by Fei-Fei et al. (2003).
- IDOL : This database, introduced by Pronobis et al. (2006), is available for download. All parts has been taken, including (sunny, cloudy, night).
- COLD : this database also introduced by Pronobis and Caputo 2009, Freiburg part has been taken , including (sunny, cloudy, night).
- UKM-FTSM : a set of real images set for the building of FTSM. These sets of images have been used for the localization process.

1.7 CONTRIBUTIONS

The major contributions of this thesis work include:

- The proposed methods introduce spatial relations for soft assignment features to improve the HBOF performance in visual place recognition.
- The proposed methods decrease confusion in visual place recognition by analyzing images from the environment in a robust manner, promptly and with less constraint, using a stable spatial relation of visual words.
- The proposed methods improve the accuracy performance of recognition by using several descriptors instead of a single descriptor in representation of the visual places.

1.8 THESIS OUTLINE

In this thesis, a computer vision technique is implemented in order to give some approaches which are useful for mobile robots to make visual place recognition in a way that give the mobile robot some independence and interact with the humans in indoor localization jobs. This thesis is organized according to the following plan; after the general introduction presented in this chapter. Chapter II introduces local features for describing the scene image, focuses on clustering algorithm in this work for place recognition. This chapter is a background review of the literatures and systems used SIFT with descriptions for some techniques that used for extracting features. And it gives a good survey on place recognition using local features, and place classification and categorization using classification techniques like SVM, K-NN.

Chapter III presents the general detailed methodology that has been adopted to achieve the objectives of the thesis. It provides an introduction for the spatial feature algorithm, which has been used in this work; the chapter describes in detail the covariance for the minimum distance approach, the entropy of covariance features proposed method which has been used in this work, and Entropy of covariance features has been used with the other features to construct multi descriptors for representing the visual place.

Chapter IV presents the results of the proposed algorithm covariance of minimum distance. In Chapter V, Entropy of covariance features has been tested with some datasets to take the stability of these features in visual place recognition. In Chapter VI, Entropy of covariance features has been tested with the other features to construct multi descriptors for representing the visual place. In Chapter VII, further and additional methods were used to improve the performance results for the adopted approach using some semantic rules place recognition. WEKA software has been used to verify the classification with some Datasets like IDOL, COLD, CALTECH101 and local data set for UKM-FTSM University. Chapter VIII is the conclusion and recommendation for the future work; addition outlines, some important points that can be considered for improving the system have been suggested.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

The accurate place recognition can be a challenging task in situations where it is particularly important to enforce self-localization precisely. Information is running from the environmental navigation for the robots through a set of sensors which could be considered as the perceptual organs of the system. For visual perception, digital cameras are exploited to capture images from the environment. These images form a matrix of pixels which contain a huge amount of information to be analyzed and represented as visual information. Visual information is an efficient symbolic description of the images scene; these descriptions are descriptors for visual features of the image's contents. The type of the cameras determines the quality and characteristics of image input; hence, the types and characteristics of images vary based on the type of the cameras. The place is a space or region whose structure contains decision points; these places are made up of distinctive states. Two types of cameras are popular to be used for a visual place recognition system: a regular perspective camera and an omni-directional camera. For mobile robots, the omni-directional cameras are commonly applied to localization processes, whereby a horizontal view of approximately 360^0 can be captured and this huge amount of visual information helps the recognition process. The most popular visual sensor applied in visual place recognition, is the perspective camera which has many functions in various recognition systems. In this research study, all the visual data sets have been captured by perspective camera, in other words, the only camera sensor has been employed rather than other types of sensors such as laser and sonar.

The main task for the place recognition refers to discriminating coding of the placement scene as descriptors, which can be stored and retrieved for identification purposes. This chapter will present a detailed description of some important descriptors that have been used in place recognition techniques as this study has used them.

2.2 PLACE RECOGNITION USING FEATURES

Features are distinct properties or pieces selected from the visual scene and they are used as a basis for the place recognition. These features as interest points can be detected or extracted in many ways; each of these points represents a type of feature differing from the others. In place recognition, many types of feature detectors have been used to recognize the visual place scene including objects in front of the camera sensor. The best type of feature detector is the one that is not affected by the diversity of the environmental impacts such as illumination, rotation, scaling, transformation, etc. Furthermore, some features are very important for recognition of landmarks.

Features can be divided into two types namely the local and global. The global features are useful for recognizing objects' shapes or general form of the object scene; for example, by using moment function, they get to know the orientation of the object as a global feature. This type of feature is sometimes used with the local features in content based image retrieval (CBIR) and it is difficult to use such a feature in localization for mobile robotics. For this reason, the uses of local features are very common in navigation because their merits to make them more popular and useful than the global ones.

The reasons why the uses of local features are required more than the global ones in this work and other research studies can be summarized as follows (Lowe, 1999):

- Locality: features are local, so they are robust to be used for occlusion and clutter.
- Distinctiveness: It can differentiate a large database of objects.

- Quantity: hundreds or thousands of local features are in a single image.
- Efficiency: real-time performance is achievable.
- Generality: local features exploit more than one type of features in different situations.

The main issue when using VSLAM is how to select suitable features of the images to be used as reliable landmarks (Oscar et al., 2007). Different features extraction techniques have been used for this purpose; features like lines (Lemaire & Lacroix, 2007), region of interest (Frintrop et al., 2006), and interest points like SIFT (Zivkovic et al., 2007; Gil et al., 2006; Valls et al., 2006; Little et al, 2001), a Harri's corner detector (Davison & Murray, 2002; Hygounenc et al., 2004), and SURF (Murillo et al., 2007), are also used in localization and mapping.

In this study, SIFT descriptors are used for visual place recognition for the mobile robotics needs a lot of efforts, where the same image may be taken in many viewpoints, so the algorithms selected for feature extraction should be invariant under many conditions of illuminations. SIFT features proofed typically invariant under rotation, translation, scaling and are partially invariant under changes in illumination (Oscar et al., 2007). The next sections are feature descriptors used for place recognition; these descriptors are discussed in some details.

2.2.1 SIFT Descriptor

Scale Invariant Feature Transforms (SIFT) algorithm, developed by Lowe (1999), is an algorithm to detect and describe local features in images. These features (referred as key points in the algorithm) are partially invariant to illumination changes but are totally invariant to scaling, image translation, and rotation (see Figure 2.1, SIFT Keypoints).

SIFT is widely used in navigation and other computer vision applications. It generates distinctive features of the image scene in a reliable, robust way, and in a relatively fast and suitable way for practical uses. It is the most widely used algorithms nowadays (Bay et al., 2008).

The SIFT algorithm compromises four steps (Lowe, 2004) which can be described as follows:

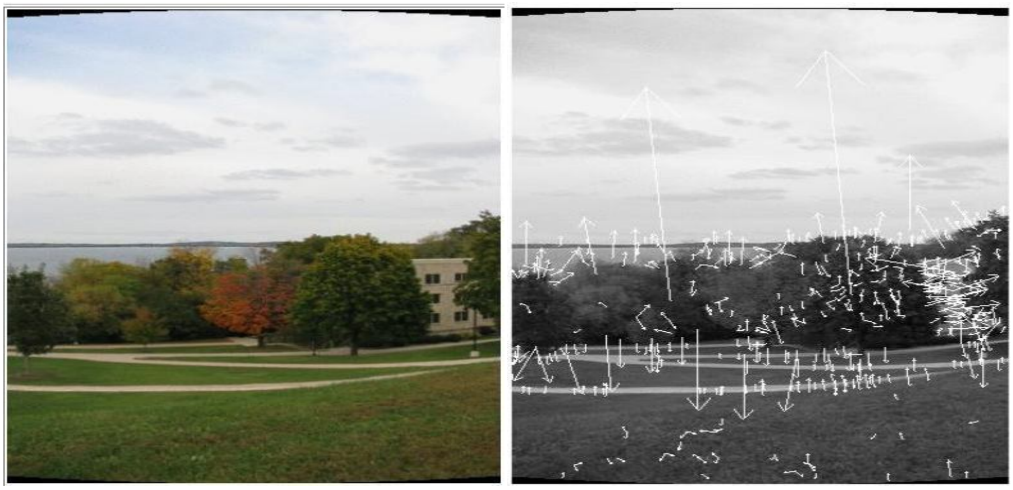


Figure 2.1 SIFT Keypoints

i. Scale-Space Extrema Detection

In this stage, the interest points (referred key points) are detected. These key points are generated when the image is convolved with Gaussian filters at different scales, and then the differences of successive Gaussian-blurred images (convolved images) are taken and grouped by an octave (an octave corresponds to double the value of σ).

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/_{2\sigma^2}} \tag{2.1}$$

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \tag{2.2}$$

$$\begin{aligned} D(x,y,\sigma) &= (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y) \\ &= L(x,y,k\sigma) - L(x,y,\sigma), \end{aligned} \tag{2.3}$$

Where *G* is Gaussian blur with scale $k\sigma$ for the original image *I(x,y)* and *L* is resulted convolution of the original image *I* and *G.*

The value of k_i is selected so that a fixed number of convolved images will be obtained per octave; it is not necessarily very small in practice. Key points are then taken by locating the extrema (maxima /minima) of the difference of Gaussian convoluted images (DoG), between adjacent scales $k_i\sigma$ and $k_j\sigma$ per octave. This will be established by comparison operation which applied to each pixel in the DoG images with its eight neighbors at the same level (scale) and nine corresponding neighboring pixels in each of the neighboring levels (scales). The candidate key point will be selected in case if the pixel value is maximum or minimum among all the compared pixels, as shown in Figure 2.2, the scale-invariant property achieved by SIFT method.

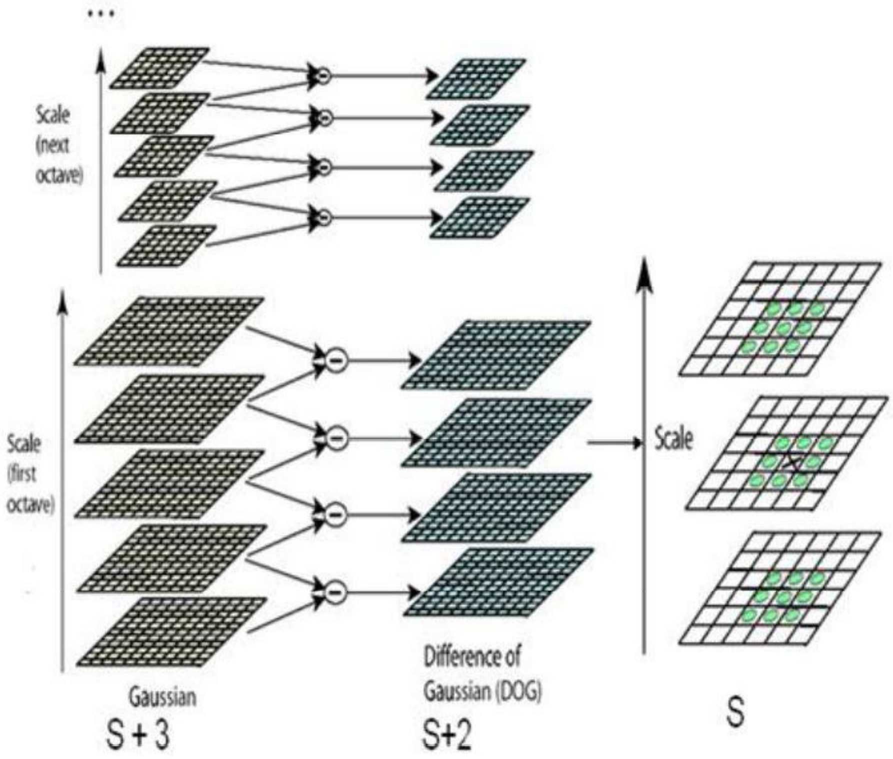


Figure 2.2 Scale space extrema

Source : Lowe, 2004

This means that the texture for the scene image detected by SIFT, at different scales (at different relative distances) will be recognized as the same feature points (Chwan & Yung-Pyng, 2007). The number of scales per one octave affects the number of key points generated per image, and their stability. The more scales are evaluated; the more key points will be found (Lowe, 2004).

ii. Key Points Localization

The first step of the algorithm produces many key point candidates; some of them are unstable. This stage of the algorithm is to perform a detailed fit to sub-pixel, and sub-scale location scale determination. Ratio of principal curvature is to reject edges and flats (i.e. to detect corners). This process improves the accuracy of the extrema location in the scale space to the sub-pixel level (Brown & Lowe, 2007), using the Taylor series for approximating the DoG function at the extrema point $(x, y, \sigma\ )$ (Chwan & Yung-Pyng, 2007):

$$\mathrm{D}\,(\mathrm{x} + \acute{\mathrm{x}}) \ = \mathrm{D}(\mathrm{x}) + \frac{\partial \mathrm{D}^{\mathrm{T}}}{\partial \mathrm{x}} \acute{\mathrm{x}} + \frac{1}{2} \acute{\mathrm{x}}^{\mathrm{T}} \frac{\partial^2 \mathrm{D}}{\partial \mathrm{x}^2} \acute{\mathrm{x}} \qquad (2.4)$$

$$\frac{\partial \mathrm{D}}{\partial \mathrm{x}} = 0 \qquad (2.5)$$

$$\acute{\mathrm{x}} = -\left(\frac{\partial^2 \mathrm{D}}{\partial \mathrm{x}^2}\right)^{-1} \frac{\partial \mathrm{D}}{\partial \mathrm{x}} \qquad (2.6)$$

The offset vector $\acute{\mathrm{x}}$ is the real feature point offset from the found position x, and the associated extrema DoG value is:

$$D(x + \acute{x}) = D(x) + D(x + \acute{x}) = D(x) + \frac{1}{2} \frac{\partial D^T}{\partial x} \acute{x} \qquad (2.7)$$

The second-order Taylor expansion $\mathrm{D}(\mathrm{x})$ value is calculated at the offset $\acute{x}$ in order to discard the key points with low contrast.

The threshold is fixed at value 0.03. If the key point value is less than this value, the key point candidate is discarded. Otherwise it will be kept, with final location $x + \acute{x}$ and scale σ, where x is the original location of the key point at scale σ. The DoG function will have strong responses along edges. Then edge responses will be removed, and the features within the object profile will be retained in order to improve the repeatability in the subsequent images (Chwan & Yung-Pyng, 2007).

Pixels which have strong asymmetry in the local curvature of the indicator function (in this case DoG) are rejected. To do this, the ratio r between the Eigen values (λ1, λ2) of the local Hessian matrix of the difference image D is evaluated (Szeliski, 2010).

$$H = \begin{bmatrix} Dxx & Dxy \\ Dxy & Dyy \end{bmatrix} \tag{2.8}$$

One makes use of the following property:

$$\text{Trace (H)} = \lambda 1 + \lambda 2, \quad \text{Det(H)} = \lambda 1\ \lambda 2 \text{ and ratio } R = \frac{\text{Trace(H)}^2}{\text{Det(H)}} \tag{ 2.9}$$

R is minimum when the eigen values equal to each other. Therefore, the higher is the absolute difference between the two eigen values, the higher will be the absolute difference between the two principal curvatures of D, and also the higher will be the value of R. According to Lowe's (1999) experiments, when the value of R is more than 10, the key point will be rejected as a part of the edge.

iii. Orientation Assignment

In this step, the orientation assignment for each key point will depend on the gradient direction for the local image. To do this, the gradient vector at each pixel within a 16x16 pixel region around key point in the scale space is calculated by finite differences for an image sample *L(x,y)* scale σ, the gradient magnitude *m(x,y)*, and orientation *θ(x,y)*.

They are pre-computed using pixel differences:

$$m_{x,y} = \sqrt{\left(L_{x+1,y} - L_{x-1,y}\right)^2 + \left(L_{x,y+1} - L_{x,y-1}\right)^2} \tag{2.10}$$

$$\theta_{x,y} = \tan^{-1}\frac{L_{x,y+1} - L_{x,y-1}}{L_{x+1,y} - L_{x-1,y}} \tag{2.11}$$

The magnitude and direction calculations for the gradient are done for every pixel in a neighboring region around the key point in the Gaussian-blurred image L. An orientation histogram with 36 bins is formed, and each bin covers 10 degrees. Each sample in the neighboring window added to a histogram bin is weighed by its gradient magnitude and by a Gaussian-weighted circular window with σ that is 1.5 times that of the scale of the key point (Lowe, 2004). The peaks in this histogram correspond to the dominant orientations. Once the histogram is filled, the orientations corresponding to the highest peaks and local peaks that are within 80% of the highest peaks will be assigned to the key points. In the case of multiple orientations being assigned, an additional key point is created having the same location and scale as the original key point for each additional orientation, as shown in Figure 2.3.

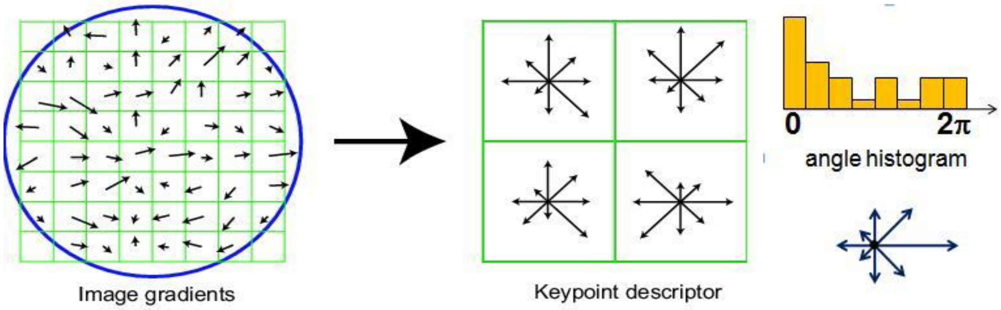


Figure 2.3 SIFT descriptors

Source: Lowe, 2004

iv. Key Point Descriptor

For region description, SIFT uses gradients in subspaces and each subspace contains 4 x 4 gradients with 8 orientations. Each feature is, therefore, described using a 4 x 4 x 8 = 128 byte vector. The SIFT features are known to be highly robust in some applications but the only trade off is that their computation is often slow. Generally, the high dimensionality of the descriptor is a drawback of SIFT at the matching step (Bay et al., 2008).

The SIFT algorithm has been frequently researched so far. The results of substantial research gave proof that SIFT features are robust and invariant to changes in size, orientation, and partially invariant to illumination (Lowe, 2004). Moreover, the algorithm uses multiple key points for matching; consequently, it is also robust to partial occlusion. The characteristics of these features let SIFT to be selected for many implementations and different research areas including this current research work. Comparing SIFT with the other good algorithms for extracting features like SURF and Harris, SIFT is the most robust and distinctive one, and it is best suited for feature matching (Oyallon & Rabin, 2013). However, sometimes SIFT features need to be optimized for more accurate robots. These features are extracted according to the principal curvatures, meaning that they are not extracted semantic. In some cases, we have more than 500 features extracted from the scene, while some of the scenes do not provide any features.

In general, the number of extracting features is in proportion to the texture of the scene; where more features are extracted for the more complicated texture. For this reason several factors are considered in the implementation in order to obtain the optimal number of images needed in the SIFT database. On the other hand, a large number of images in the database give a large number of features to be compared with the image scene during the application process. The optimum number of images is normally determined by storing the most important images that affect the navigation process. This will make the system operate in real-time while no processing time is wasted because the matching of the images overlaps.

Furthermore, the match between features themselves is high since the SIFT features are highly distinctive. As a result, false positives for an entire image scene are less. This can be reduced further by adjusting the threshold used. This makes SIFT to be clear enough to be chosen as compared with the other descriptors of the features.

2.2.2 Spatial Pyramid Layout

The spatial pyramid approach is a process through which features are taken from multi level partition, and then are combined together to form a single feature vector. Lazebnik et al. (2006) and Bosch et al. (2007) use the fixed partitioning method to construct levels of HOG histograms then combine these levels to form a feature descriptor for the image, as illustrated in Figure 2.4. This multi-resolution approach has been proven to give better performance result compared to the single level resolution; this comparison has been shown by many researchers (Grauman & Darrell, 2005; Lazebnik et al., 2006; Bosch et al., 2007; Hadjidemetriou et al., 2001).

The spatial pyramid approach or multi-resolution approach has been used to obtain the spatial correspondence of histograms (Hadjidemetriou et al., 2001). In several research studies, global features and multiple local histograms generated for the regions of the partitioned image have been employed to describe the image in a more precise way. The histogram representation is commonly used for global features to give a description for simple images however it is not so robust for describing clutter and occlusion of the images. To solve this problem, a more robust way is needed. The multi spatial resolution level is a robust method which is based on the histogram for the regions generated at each level of resolution, by this method; kernel features for the image will be created through combining the multilevel features (Grauman & Darrell, 2005).

Lazebnik et al. (2006) has introduced some spatial information for the words which have been constructed by using the bag-of-words approach. While the words are assigned for the SIFT descriptors, kernel features are extracted on a grid over the image. Then the spatial pyramid algorithm will use further processing of these words.

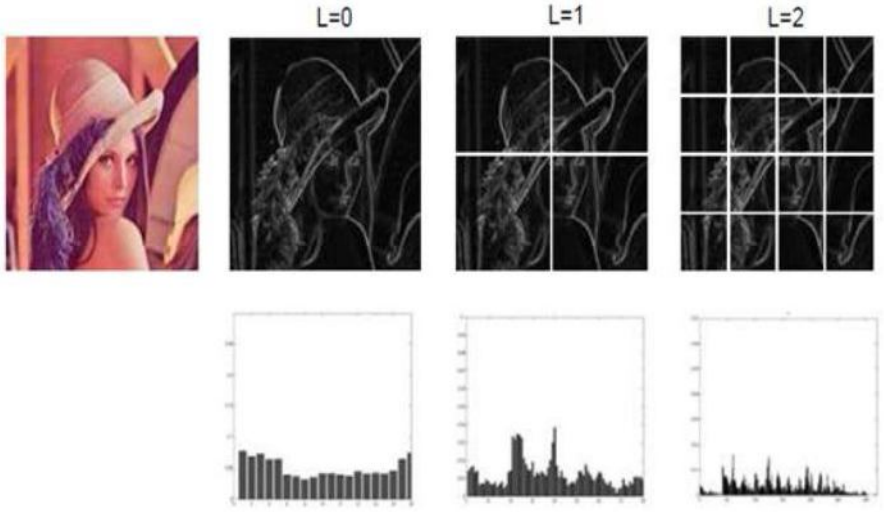


Figure 2.4 A spatial pyramid representation with HOG histogram of local regions
Source: Abdullah, 2010

The Spatial pyramids are established by partitioning the image into a grid as shown in Figure 2.5. Then, the image descriptor will be formed by weighing and concatenating the histograms of words for each cell of the grid in the image.

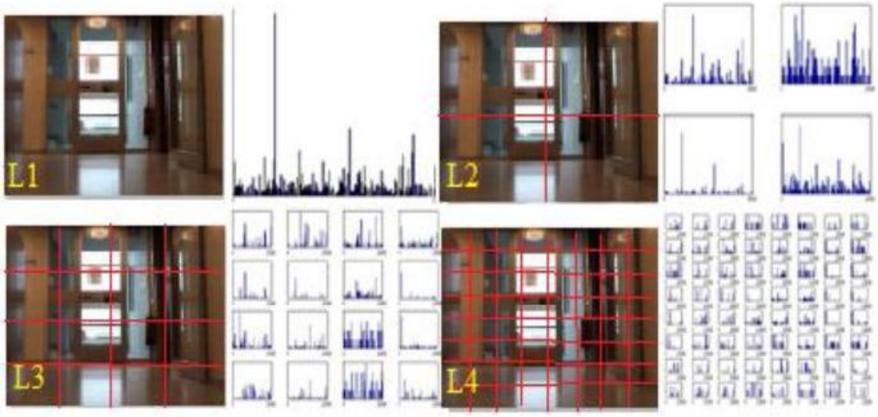


Figure 2.5 Spatial Pyramid histogram according to Lazebnick et al. (2006)