INCREASING ROBUSTNESS IN APPEARANCE-BASED SPATIAL COGNITION

by

Berkan Höke

B.S., Electrical and Electronics Engineering, Bilkent University, 2014

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Electrical and Electronics Engineering
Boğaziçi University

2019

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my supervisor Prof. Işıl Bozma for her guidance, enthusiasm and patience. Her guidance helped me to keep my motivation high at all times. Her suggestions and ideas helped me to find my way through at difficult times and successfully complete my thesis.

I would like to thank Prof. Yağmur Denizhan and Prof. Antonios Gasteratos for being a member of my thesis committee as well as giving valuable feedback through out my research.

I would like to express my sincerest thanks and appreciation to all ISL lab members but especially Kadir Cumali, Mahmut Demir, Hakan Karaoğuz, Mustafa Mete, Çağatay Odabaşı, Doğan Patar, Kadir Türksoy and Esen Yel for their countless support during software development.

I am also grateful to people in Migros, particularly Zeynep Zerrin Turgay for providing me an opportunity to complete my thesis.

Finally, I am deeply indebted to all my friends for their friendship and support during my hard times. I would like to offer my thanks to Kübra Toka, who always tried to be there for motivating and helping me from beginning to end of this work. My deepest thanks are to my family for their patience and unfailing support through my whole life.

This thesis was supported in part by the Scientific and Technological Research Council of Turkey (EEEAG 115E380).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

A_{AG}	Adaptive Gamma Corrected Image
A_{GD}	Gamma-DCT corrected image
A_{HE}	Histogram equalized image
A_{SM}	Sensor model estimated image
A_{WF}	Wiener filtered image
$c \in R^2$	Robot's position
$d_N(I(x_k)$	One-class SVM function between node N and descriptor $I(x_k)$
D_j	j^{th} subplace
h	Height in SLINK
$I(x_k)$	Associated descriptor with k^{th} basepoint
n_b	Number of basepoints in a place
n_s	Number of subplaces in a place
N	Node in the place memory
$N_\downarrow$	Children nodes of N
$N_\uparrow$	Parent nodes of N
$g_N(D_j)$	Reward function of associating D_j with node N
s	Subplace
U	Informativeness function
W	Wiener Filtering in spatial domain
$\mathcal{W}$	Wiener Filtering in frequency domain
x_k	k^{th} basepoint
$\alpha \in S^1$	Robot's heading
$\eta_N(D_j)$	One-class SVM function between node N and subplace D_j
κ	Incoherency function
λ	Closest distance representation in SLINK
$\nu_N(j)$	Binary function of resulting one-class SVM
π	Node with the closest in SLINK

ρ	Pearson correlation function
τ_ω	Temporal window size
τ_n	Temporal window extension threshold
τ_p	Place size threshold
τ_r	Recognition threshold
τ_μ	Intensity Mean Threshold
τ_σ	Intensity Variance Threshold
ζ	Hierarchical Clustering Set

LIST OF ACRONYMS/ABBREVIATIONS

AG	Adaptive Gamma
BD	Bubble Descriptor
DCT	Discrete Cosine Transform
FFT	Fast Fourier Transform
Fr	Freiburg Site
GD	Gamma Corrected Discrete Cosine Transform
HE	Histogram Equalization
Lj	Ljubljana Site
NC	New College Site
ROS	Robot Operating System
Sa	Saarbrücken Site
SIFT	Scale Invariant Feature Transform
SLAM	Simultaneous Localization and Mapping
SLINK	Single Linkage Clustering
SM	Sensor Modeling
SURF	Speeded up Robust Features
SVM	Support Vector Machine
TSC	Topological Spatial Cognition
WF	Wiener Filter

ABSTRACT

INCREASING ROBUSTNESS IN APPEARANCE-BASED SPATIAL COGNITION

This thesis is concerned with the robustness of robots' spatial cognition. Two separate issues are considered: Illumination invariance and place recognition. Illumination invariance ensures that appearances of the same scene under different illumination conditions do not change. Four existing approaches are considered as well as a new approach is proposed in this framework. Aforementioned two issues are addressed in two stages individually. In the first stage, appearances collected from robot's image sensor are reconstructed in order to make robot interpret surrounding environment robust to illumination variations by using any descriptor. Illumination invariant appearances are employed so that robot is able to both distinguish one place from another in a single environment and recognize any place regardless of the change in the illumination conditions. An extensive comparative experimental evaluation serves to demonstrate how each method performs with respect to appearance similarity, place detection and place recognition. On the other hand, the second stage focuses on robot's spatial memory. Simple yet efficient method is proposed in order to address the place learning and recognition performance issues. Detected places are clustered into smaller place clusters called "subplaces" where each subplace includes canonical appearances associated with the corresponding place. Thus, canonical appearances are evaluated for learning and recognition instead of employing the whole place. It guarantees that only dominant appearances in a place are recognized and thus achieves a better performance by the help of these canonical views. Likewise in the first stage, comparative experiments are conducted on various datasets and on the robot in real time with and without subplace approach. Learning and recognition performance alterations are discussed in detail.

ÖZET

GÖRÜNÜŞE DAYALI UZAMSAL ANLAMLANDIRMANIN DAYANIKLI HALE GETİRİLMESİ

Bu tezde, robotlarda görünüşe dayalı uzamsal anlamlandırmanın dayanıklı bir hale getirilmesi üzerinde çalışılmıştır. Çalışmada, görünüşlerde ışık değişmezliği ve yer tanıma olmak üzere iki farklı konu göz önünde bulundurulmuştur. Işık değişmezliğinin hedefi, görünüşün farklı ışık koşulları altında değişmemesidir. Bu çerçevede, daha önce geliştirilmiş olan dört yöntemin yanı sıra yeni bir yöntem önerilmiştir. Bu konular için üretilen çözüm iki aşamada gösterilmiştir. Birinci aşamada, robotun kamerasından elde edilen görüntülerin tekrar yapılandırması ile, robotun çevresindeki ışık değişikliklerine dayanıklı bir şekilde herhangi bir tanımlayıcı ile yorumlayabilmesi sağlanmıştır. Işık değişmezliği e robotun bir ortamdaki yerleri birbirinden ayırt edebilmesi ve tekrar ziyaret ettiği bir yeri ışık koşullarından bağımsız olarak tanıyabilmesi hedeflenmiştir. Geniş kapsamlı karşılaştırmalı deneyler yapılarak uygulanan yöntemlerin görünüş benzerliği, yer sezinleme ve yer tanıma başarımları ölçülmüştür. İkinci aşama ise robotun uzamsal hafızası üzerine yoğunlaşmıştır. Bu aşamada sade fakat etkin bir yöntem uygulanarak yer öğrenme ve yer tanıma performansının geliştirilmesi üzerine çalışılmıştır. Sezinlenen yerler küçük parçalara ayrıştırılarak o yere ait yalnızca temel görünümler içeren "alt yerler" oluşturulmuştur. Böylece, yer öğrenme ve tanıma için o yere ait bütün görünümleri kullanmak yerine, o yerin içerdiği temel görünümler değerlendirilmiştir. Temel görünümler sayesinde, bir yerin içerisindeki dominant *alt yerlerin* kullanılarak tanınması ve böylece daha yüksek bir başarım göstermesi kesinleştirilmiştir. Birinci aşamaya benzer olarak, çeşitli veri kümeleri üzerinde ve ayrıca robot üzerinde gerçek zamanlı "alt yerler" yaklaşımının kullanıldığı ve kullanılmadığı deneyler karşılaştırmalı olarak gerçekleştirilmiştir. Deneylerde yer öğrenme ve tanıma başarımlarındaki değişim ayrıntılı şekilde ele alınmıştır.

1. INTRODUCTION

This thesis is concerned with improving the robustness of spatial cognition in robots. Spatial cognition is concerned with the acquisition, improvement and revision of knowledge of places [1–3]. If robots are ever to have lifelong operation and thus play a bigger role in daily life, it is crucial for them to have spatial cognition. Thus, the robots should be able to have successful spatial knowledge and interpretation of the environment.

The proposed models are motivated by findings in humans' spatial cognitive abilities [4]. Similar to human perception, the term 'place' refers to human-like definition such as A's room or B laboratory [5].

Hence, this definition differs from most appearance based SLAM or some topological approaches that define each location as a place. Appearances play a key role in defining places - in cases of unavailable or unreliable position information. In this case, each place is defined by a collection of resembling appearances sharing common perceptual signatures. A place is constructed by considering the features extracted that are spatially relevant, as localization data may not be always available or reliable [6, 7].

1.1. Topological Spatial Cognition (TSC)

In this thesis, we consider the topological spatial cognition (TSC) model as proposed in [8]. In this model, the flow of processing is as shown in Figure 1.1. It takes consecutive appearances as input and checks if there exist any transition regions among them. As long as a transition region is detected, the place detection decision is made based on the criteria such as place extent and similarity of descriptors. Each detected place is kept in database. If there are more than three detected places, then TSC continues with the recognition with criteria explained in 3. If a place could be recognized with a previously visited place, these two places are merged. Otherwise, the latter learned as a new place and saved to another database.

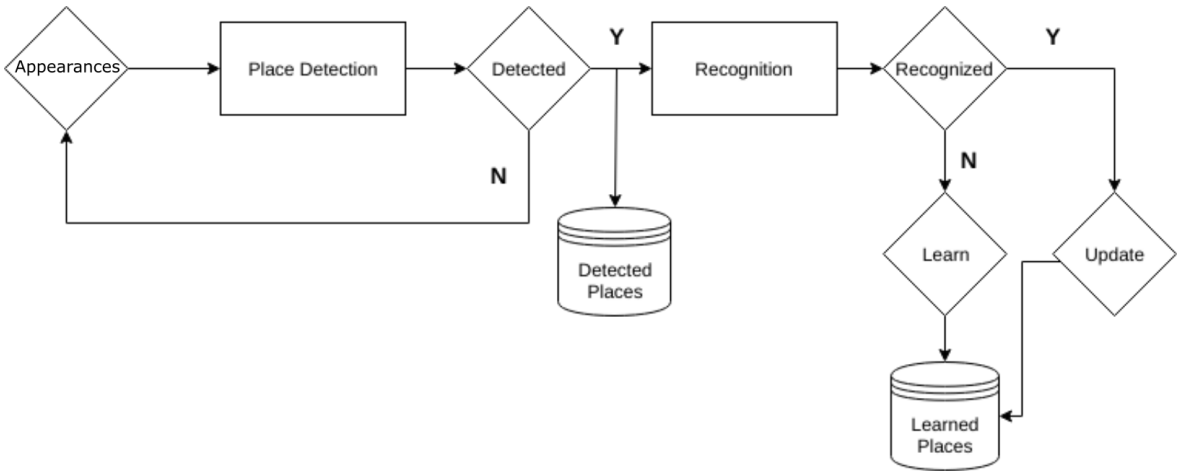


Figure 1.1. Flowchart of TSC model.

1.2. General Approach

Let $x = [c^T a]^T$ be the 'base point', where position defined as $c \in R^2$ and heading is represented as $a \in S^1$. The set $\mathcal{X} \subseteq R^2 \times S^1$ is the base space (all possible locations and headings). If odometric data is not known or is inconsistent due to some technical reasons, we will not have the knowledge of the coordinates of the base point x as which is the case in this study. The robot internally encodes each appearance by a visual descriptor. This model uses *bubble descriptors* which is explained in Appendix A [9]. As the robot moves along the path that consists of base points x_k where $k \in K$, the collected appearances result in a sequence of descriptors $I(x_k) \in R^{N_I}$.

The goal of place detection is to delineate distinct places as the robot navigates around via processing the incoming visual stream without any prior knowledge regarding the extent of distinct places. It thus provides a basis for place recognition[1] , topological mapping, visual navigation and higher level spatial reasoning such as semantic scene understanding [12]. As such, detecting places is an integral capability - if robots are to become spatially aware [13]. Note that as place detection is done incrementally as the robot traverses its environment, even though illumination conditions such as sunny or cloudy weather will not change in general, illumination still varies across the scene depending on the relative position of the robot with respect to the illumination sources.

[1]In some approaches such as [10, 11], place detection and recognition are solved together.

Place detection is based on iterative partitioning where partitioning is implemented via the coherency between consecutive frames. Latest partition is continuously updated as robot visits new base points. Let us define the partitions as $D_0, \ldots, D_m^*$ where $D_m \subset \mathcal{M}$ indexed by $\mathcal{M} = 1, \ldots, m^*$. Each D_m corresponds to 'detected place'. Partitioning process is done in real-time as robot navigates around. Separation of each place is decided via informativeness and coherency constraints among the descriptors. These constraints help us to determine maximal neighborhoods and temporal windows. Each maximal neighborhood consists of set of base points which are affiliated with one detected place. On the other hand, each temporal window consists of set of basepoints which belong to a transition region but not place. During the processing of each frame, each base point is appended to the maximal neighborhood or temporal window considering the informativeness, coherency and plenitude of the associated descriptors $I(x_k), k \in K$. Previously mentioned illumination invariance methods corrects the images where value of the pixels are reversible such that appearance between same scene in different illumination conditions converges to each other. Nevertheless, there are exceptions where it is impossible to regain a pixel's value since there is no valuable information in the corresponding frame as robot moves towards a wall or completely dark region. Examples regarding this problem are demonstrated in Figure 1.2 and these frames are labeled as *uninformative frames* which are added to temporal window and not considered in any place. In such circumstances, uninformativeness of the frames is trivially identified by using image statistics. Binary valued function $U : X \to [0, 1]$ using mean and variance thresholds $\tau_\mu > 0$ and $\tau_\sigma > 0$, respectively.

$$U(X) = \begin{cases} 1, & \text{if } mean(X) > \tau_\mu \quad \text{and} \quad var(X) > \tau_\sigma \\ 0, & otherwise \end{cases} \tag{1.1}$$

The coherency check between frames is the χ^2 distance between the corresponding descriptors which is defined as

$$\chi^2(I(x_k), I(x_{k+1})) = \frac{(I(x_k) - I(x_{k+1}))^2}{I(x_k) + I(x_{k+1})} \tag{1.2}$$

Figure 1.2. Sample Uninformative Images

Once the coherency constraint is not satisfied, newly visited frame is appended to current temporal window.

Place recognition approach is based on matching detected places as mentioned in the previous section. Additionally, single linkage hierarchical clustering approach is employed in order to construct place memory where each visited place is represented as leaves of binary tree in the recognition part [14]. Hierarchical approach provides that the places does not need to be compared with the all nodes in the memory and thus a significant time save. The construction of place memory is explained in detail in Chapter 3. Distinctively, correlation is employed instead of sum of squared distance, since it comes up with the result $d(D_m, N) \in [-1, 1]$ and thus in same interval with SVM vote percentage.

All the newly detected places are checked whether $g_n(D_m) > \tau_r$ condition is satisfied at each level of the tree.

Figure 1.3. Sample Incoherent Images

.

1.3. Problem Statement

This thesis considers this problem in two stages:

- First, each place should consists of images that are insensitive to environmental changes, especially illumination level as it is the most unstable factor for both indoor and outdoor. Robot visits and detects the places sequentially. As it discovers the environment, it must relate all the appearances within a place without interruption in case it does not enter any new region. Since we use the previously developed bubble descriptors [15] due to its rotational invariance property, the only alternative is converting the images into illumination invariant ones beforehand. The robot recognizes or learns the places it has just detected. Illumination invariance also consider to make the robot robust against illumination changes when it visits same place more than one time. This concept is different from the previous problem, since previous one covers the illumination changes for the detection phase in the sequential manner.
- Secondly, any place may consist of more than one distinctive appearances. These distinctive appearances play a key role at the recognition stage, since the robot may fail to recognize places even though these appearances seem alike. Alternatively, a learned place in the robot's memory may contain outlier descriptors that is present in the place, although corresponding appearances are unrelated to general appearance of that place. These outliers may confuse the robot while performing recognition and they must be eliminated.

1.3.1. Sensitivity to Illumination

Dealing with varying illumination conditions is one of the challenges in tasks such as place detection and place recognition. This is because appearances play a key role in having the robot relate its surroundings to its stored knowledge that has been learned in past experiences. Note that illumination changes involve different cases for indoor place detection applications:

- Weather conditions such as cloudy, sunny, night
- Time of the day such as morning, mid-afternoon and night that affects overall lighting
- The position of the light source with respect to the robot
- Strong light source from a single direction which causes camera to darken the rest of the image

1.3.2. Canonical Views

The second issue is regarding how each place should be learned and recognized.

Another concern of this thesis is providing a way how a place is learned and recognized. Clustering of detected places in hierarchical manner using single linkage clustering and extraction of canonical appearances, with bubble descriptors as the place representatives employed to address this problem. At the end of the process, main goal is to gather similar appearances in a single cluster. Therefore, this approach aims to increase precision and recall rates of the recognition with post pruning of outliers and divides the places into clusters that have more extent than it should have. Clustering process should be also fast and efficient method as well, since real - time mobile robotics is aimed with this work. Hence, we proposed a fast and online approach.

1.4. Contribution

The major contributions of this thesis are as follows:

(i) Recent studies about the illumination invariance are investigated and implemented. On the other hand, a novel approach is proposed in order to make robot's detection and recognition more robust against illumination changes.

(ii) Places are clustered into canonical views in order to achieve higher recognition results with a novel method presented in this study. Each cluster is called as subplace which enables that the place is able to be recognized with its dominant cluster.

(iii) In addition, TSC software is revised so that software works more consistently and provides more accurate results. Details of the revisions are indicated in Appendix B.

1.5. Organization of Thesis

The rest of this thesis is organized as follows:

The problem of illumination invariance is considered in Chapter 2. General flow containing such as place detection, place memory and place recognition based on bubble descriptors are explained. For the interested reader, a brief overview of bubble space is given in Appendix A. The performance of recent algorithms are measured with various datasets by conducting experiments on very different lighting conditions in terms of detection, memory and recognition. Lastly, the performance of these methods is discussed at the end of the chapter.

In Chapter 3, learning places based on canonical view from detected places is presented. Subplace term is introduced in this section. It is observed that canonical views provide consistent recognition matches as well as they enable the robot to determine outlier appearances within a place.

The thesis concludes with a brief summary and a discussion regarding future work as presented in Chapter 4.

2. ILLUMINATION INVARIANT SPATIAL COGNITION

2.1. Introduction

Different illumination conditions lead difficulties, while detecting and recognizing places. The main reason behind this challenge is the role of appearances during the matching process of robot that relates new appearances with its stored knowledge has been learned in past experiences. Nonetheless, illumination conditions affect the appearance of the same scene. For example, consider the different appearances of one scene as seen in Figure 2.1. Human observers can easily decide that these appearances belong to same scene even if they are different in terms of appearance. On the other hand, the robot can classify them as different scenes when there is no illumination invariance, even if this is not the case.

There may be situation where captured image is impossible to reverse e.g – there is no light source. If this is the case, optical sensors is not sufficient and another type of sensor must be employed in order to overcome this problem. However, another sensor outputs are not evaluated in this study, since appearances of the place are the only concern.

(a) Cloudy (b) Sunny (c) Night

Figure 2.1. Appearance of a Scene Under Different Illumination Conditions

The main aim of this chapter is to reveal the method which minimize the impacts of illumination. The outline of the this chapter is as follows:

First, a brief summary of proposed approaches is given in Section 2.2. Histogram equalization on color images is explained briefly in Section 2.3. This is followed by the discussion of gamma normalized discrete cosine transform in Section 2.4. Camera modeling is presented in Section 2.5. Similar method is the Wiener filtering as summarized in Section 2.6. The last method is the adaptive gamma correction that is a novel method proposed in this thesis 2.7. We then compare their the details of the detection phase and conducted experiments on COLD [16] and New College [17] datasets in 2.8.2. Construction of place memory and its consistency is included in 3.3 Our experimental results demonstrate the advantages and disadvantages of each method in getting illumination-invariant appearances.

2.2. Related Literature

Many methods have been proposed to deal with illumination changes. These methods can be classified in two groups:

First category includes illumination invariant features obtained from the image descriptors such as SIFT (scale invariant feature transform) and SURF (speeded up robust features) which are tested under illumination changes [18]. Although, U-SURF outperforms other methods, none of them achieve robust results in case of wide illumination interval [6]. On the other hand, second category includes illumination invariant imaging which constitutes another image by suspending the effects of light. It is defined as "base image", which focuses on removing illumination effects on color and intensity [19]. On the other hand, any descriptor can be utilized for place detection and recognition with generated illumination invariant images.

Histogram equalization method is a common approach which process the image through CDF (cumulative distribution function) of pixel values and creating a special function which makes CDF as linear as possible by remapping the pixel values. It is a simple yet effective method for balancing image contrast.

Finlayson et al. [20] proposed an approach which transforms RGB color space to 2-dimensional log-chromaticity space, and seeks for a direction which minimizes energy in that space in order to retrieve illumination-invariant images. Maddern et al. [21] proposed a method based on spectral response of the camera sensor cells to red, green and blue light. Peak spectral response for each channel is assumed to be dirac delta function and illumination effects in the image is obtained by removing the effect of luminance and retrieving the reflectance of the scene. Shakeri and Zhang [22] proposed a similar approach where Wiener filter is applied to the image in order to separate reflectance and luminance components of the image. Illumination invariant images obtained with this method is based on reflectance, since it does not change with the external light sources but depends on the surface properties.

GammaDCT method is based on the method proposed in [23] where gammaDCT method uses exponential values instead of logarithm approach. Applying gamma correction gains an advantage over logarithm transform where logarithm transform causes noise in dark areas of the image [24, 25]. In this method, pixel values are mapped to another value with a constant γ value and hence an pixel values could be stretched or compressed depending on γ. Besides that, low frequency components obtained via DCT (discrete cosine transform) are down-sampled in order to reduce the effect of illumination on the image.

2.3. Histogram Equalization

Histogram equalization is a method that is used for contrast adjustment on images based on the cumulative distribution function of the intensity levels. Originally, it is used for gray-scale images. Then, the extension of histogram equalization to color images has been considered. Let $A_{HE} : U \rightarrow S^3$ denote the corresponding transformation map. Histogram equalization could not be applied on each channel separately, since this affects the red, green and blue depending on pixel distribution in corresponding channel. In order to avoid this, RGB image is converted to another color space. In particular, YCbCr color space is preferred because its common usage in digital images such as JPEG. In this case, the histogram of the luminance channel (Y) is equalized, so

that color balance does not change [26, 27]. This is followed by converting the YCbCr image back to RGB. For the sample scene, histogram-equalized appearances demonstrated in Figure 2.2. It is observed that the darkness observed in the cloudy and sunny images are removed. The resemblance of the scene under night illumination to them has also increased - even though there is still a difference.

(a) A^C_{HE} (b) A^S_{HE} (c) A^N_{HE}

Figure 2.2. Histogram equalization.

2.4. Gamma-Normalized DCT

Gamma-normalized DCT method is based on compressing the pixel values and suppressing the effect of illumination by down scaling low frequencies [24]. It is a modified version of [23] where the logarithm operation is replaced with γ-correction. Gamma correction is a nonlinear operation that compresses or stretches the intensity values with respect to $\gamma > 0$ value. Let V_γ denote the resulting intensity image:

$$V_\gamma(u) = V^\gamma(u)$$

If $\gamma > 1$, the effect is to expand the intensity values. On the other hand, if $\gamma < 1$, intensity values are compressed. As such, it can transform a high contrast image into a balanced one with simple operation.

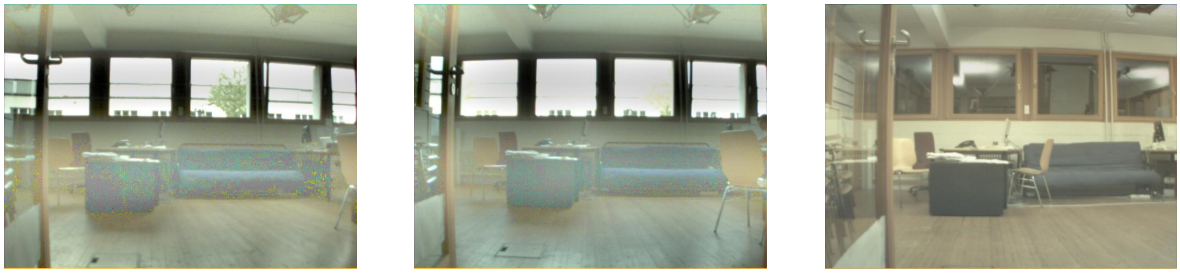
This is followed by discrete cosine transform (DCT) that is used to express V_γ in frequency domain $\mathcal{A}(\omega)$ where $\omega = [\omega_1, \omega_2]^T$ represents the spatial frequencies.

$$\mathcal{A}(\omega) = \sum_{u_1=0}^{N_1-1} \sum_{u_2=0}^{N_2-1} V_\gamma(u) \cos\left(\frac{\pi}{N_1}(u_1 + \frac{1}{2})\omega_1\right) \cos\left(\frac{\pi}{N_2}(u_2 + \frac{1}{2})\omega_2\right) \tag{2.1}$$

As illumination variations occur in the low frequencies, while the details are encoded by the high frequencies, first N_T low frequency elements are downsampled in order to remove illumination effects [24]. In addition, the average intensity of the image can be adjusted to a mid value by setting the first term of $\mathcal{A}$ [23]:

$$\mathcal{A}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \mu^\gamma \sqrt{N_1 N_2} \tag{2.2}$$

The inverse DCT of $\mathcal{A}$ after it has been modified in the spatial frequency domain followed by contrast stretching to the $[0, 2^b]$ range are used to obtain the illumination invariant intensity image. The parameter $\mu > 0$ is an arbitrary value for the desired average intensity. Finally, $A_{GD} : U \to S^3$ is obtained by merging the resulting intensity image with hue and saturation channels. The parameters are determined empirically as $\gamma = 0.45$, $N_T = 5$ and $\mu = 0.3$ and the low frequencies downscaled to the half of their actual values. For the sample scene, gamma-normalized DCT method yields illumination-invariant appearances as seen in Figure 2.3. It is observed that this method has quite good performance visually – especially in low illumination conditions a slight wavy effect is seen – due to lack of low frequencies. However, while the night scene better resembles the cloudy and sunny scenes, discrepancies still exist.



(a) A^C_{GD} (b) A^S_{GD} (c) A^N_{GD}

Figure 2.3. Gamma Corrected DCT

2.5. Camera Sensor Modeling

Camera modeling is based on mapping peak responses of red (R), green (G) and blue (B) image sensor A_R, A_G and A_G respectively to an illumination-invariant chromaticity space [21], the origin of which can be traced in [28]. In this approach, the spectral sensitivity of the image sensor is assumed to be infinitely narrow. Thus, the daylight spectrum is approximated by a black body. If the illuminant spectrum is approximated as Planckian source, it can be shown that the log of each color image consists of three components: a wavelength-independent component encoding illumination geometry, a component that depends on the reflectance of the surfaces being imaged and a component that depends on the correlated color temperature of the illuminant. The transformation map $A_{CM} : U \rightarrow R$ is obtained by eliminating the first and third components. This is achieved by taking the log difference between the response of one color sensor A_G and the weighted sum of the responses of two other sensors A_B, and A_R - as proposed in [29]:

$$A_{CM}(u) = \log\big(A_G(u)\big) - \alpha \log\big(A_B(u)\big) - (1-\alpha)\log\big(A_R(u)\big)$$

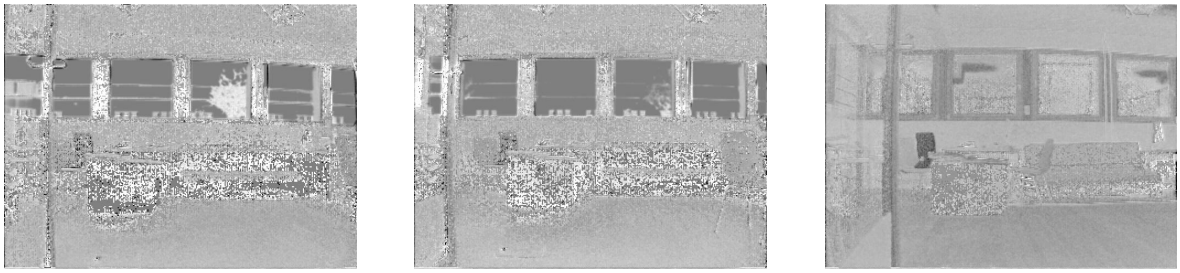
The resulting image is a single channel image. It is argued that a single channel is sufficient to differentiate between most materials in natural scenes.

The parameter α is uniquely determined by peak response wavelengths $\lambda_B < \lambda_G < \lambda_R$ of the three color sensors :

$$\frac{1}{\lambda_G} = \frac{\alpha}{\lambda_B} + \frac{1-\alpha}{\lambda_R}$$

Thus, it can be uniquely determined for a given RGB camera - based on the knowledge of the peak spectral responses of each color sensor. This information is usually available in the datasheet of the camera. For example, the Videre Design MDCS2 digital color cameras that are used in the COLD dataset [16], have Micron MT9M001 image sensors with $\alpha = 0.3975$. For our sample scene, the results of camera modeling method are

as shown in Fig. 2.4. It can be observed that there are a lot of erroneous artificial highlights exist, owed to the high contrast nature of the images. Nonetheless, as far as shadows and low contrast images are concerned, the method seems to perform adequately.



(a) A^C_{CM} (b) A^S_{CM} (c) A^N_{CM}

Figure 2.4. Using Camera Modeling

2.6. Wiener Filtering

The fourth method is based on decomposing the intensity image V into its reflectance R and luminance L components - assuming the following formation model of the intensity image:

$$V(u) = R(u)L(u) \tag{2.3}$$

This is achieved by applying Wiener filter [22]. Differing from previous method, this method does not assume the ideal narrow-band color camera or a calibration step for each environment. By taking logarithm of Eq. 2.3:

$$f(u) = \nu(u) + \eta(u) \tag{2.4}$$

where $f = \log V$, $\nu = \log R$ and $\eta = \log L$. The reflectance component is then used as the illumination invariant image:

$$\nu(u) = f(u) - \eta(u) \tag{2.5}$$

The map η is estimated from a single image by applying Wiener filter with optimal parameters. It is based on work as presented in [30]. Similarly, power law spectrum is assumed - namely if P_μ and P_η are power spectral densities of μ and η respectively:

$$P_\mu(\omega) \propto \omega^{-\alpha_\mu} \tag{2.6}$$

$$P_\eta(\omega) \propto \omega^{-\alpha_\eta} \tag{2.7}$$

where $\alpha_\mu, \alpha_\eta > 0$ are positive scalars. Wiener filter in frequency domain can be shown as:

$$W(\omega) = \frac{P_\eta(\omega)}{P_\eta(\omega) + P_v(\omega)} = \frac{\zeta}{\zeta + \omega^{\alpha_\eta - \alpha_\nu}} \tag{2.8}$$

where $\omega \in R$ is the frequency in one direction, ζ is the ratio of power spectra P_η and P_ν at the frequency $\omega = 1$. For most of the natural images, $\alpha_\eta - \alpha_\nu \approx 2$ [31]. The value ζ is chosen empirically and it is set to 0.05, since higher values might cause to loss of details. The resulting discrete equation in the spatial domain is as follows:

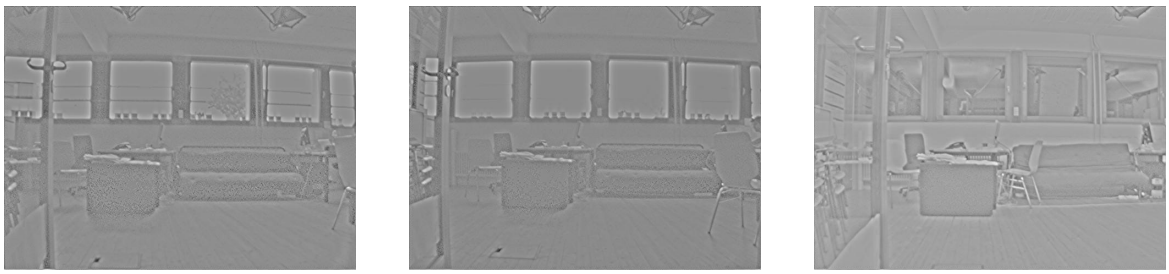
$$(0.05I + D^T D)\eta_j = 0.50 f_j \tag{2.9}$$

Here, I is the identity matrix with appropriate dimension, f_j and η_j correspond to the j-th row (column) of f and η respectively and D is $(N_1 - 1) \times N_2$ ($(N_2 - 1) \times N_1$) difference matrix with entries:

$$D_{ij} = \begin{cases} -1 & \text{if } i = j \\ 1 & \text{if } i = j - 1 \\ 0 & \text{otherwise} \end{cases}$$

The transformation map $A_{WF} : U \to R$ is obtained by taking the average of $f - \eta$ computed using rows (columns).

Illumination-invariant appearances that are obtained using this method are shown in Fig. 2.5. The generated image appears to be similar to that of the previous method, yet with absence of artificial highlights. It should be noted that, as overall intensity may differ dramatically between consecutive frames, it does not seem promising in detecting place extents and transition regions correctly. Also, note that camera sensor modeling and wiener filtering approaches only provide single channel images, since their outputs use and it is not replaceable with intensity of the image.



(a) A^C_{WF} (b) A^S_{WF} (c) A^N_{WF}

Figure 2.5. Using Wiener Filtering

2.7. Adaptive Gamma Correction

The last method is motivated by gamma correction, but it differs by using a varying gamma value that adapts depending on the intensity V of the appearance A [32].

$$V^p(u) = V(u)^{\gamma(u)} \tag{2.10}$$

This is because correction with a constant gamma value may fail if the illumination varies across the scene and thus if there are both dark and bright areas in the image. The resulting intensity map V^p is subjected to contrast stretching in the $\left[0, 2^b\right]$ range and is then used to obtain the transformation map $A_{AG} : U \rightarrow S^3$ that maps back to the HSI color space.

The γ function is based on the center-surround relationship at each pixel u considering the mean value $V'(u)$ in the ϵ-neighborhood $N_\epsilon(u)$ of the pixel [33]:

$$\gamma(u) = \gamma_0 + \Delta\gamma V'(u) \tag{2.11}$$

where

$$V'(u) = \frac{1}{|N_\epsilon(u)|} \sum_{u' \in N_\epsilon(u)} V(u') \tag{2.12}$$

Here, the parameter γ_0 denotes the minimal gamma value. On the other hand, the parameter $\Delta\gamma$ denotes the slope of the linear map and thus determines the increase in $\gamma(u)$.

Commonly, fixed γ values are chosen inside the interval $[0.45, 2.2]$ [34] [35]. However, with adaptive $\gamma(u)$ values these are stretched slightly to $\gamma(u) \in [0.4, 2.5]$. Note that, larger intervals may cause inconsistency between consecutive images during the robot navigation. Hence, $\Delta\gamma$ is constrained to $(2.5 - 0.4)/(2^b - 1)$, thus $\Delta\gamma = 0.0082$ for an 8 bit image. In Figure 2.6 pixel values of output image are shown as a look-up table depending on the relation of center-surround pixels. Obviously, pixel mapping leans towards a logarithmic function instead of an exponential one, as the average of surround pixels increases. Through this method, the details in darker areas are revealed as well as illumination effects from the light source are reduced.

Since $\gamma(u)$ is tempered to center-surround pixel relation, adaptive gamma correction is able to both compress or stretch pixel values depending on the corresponding illumination levels. As such, the values are $\gamma(u) < 1$ for dark regions of the image and $\gamma(u) > 1$ for the illuminated regions. By this way, brighter pixel values are pulled down with the amount of indicated with average surround pixel values non-linearly. Sample results are represented in Figure 2.7.

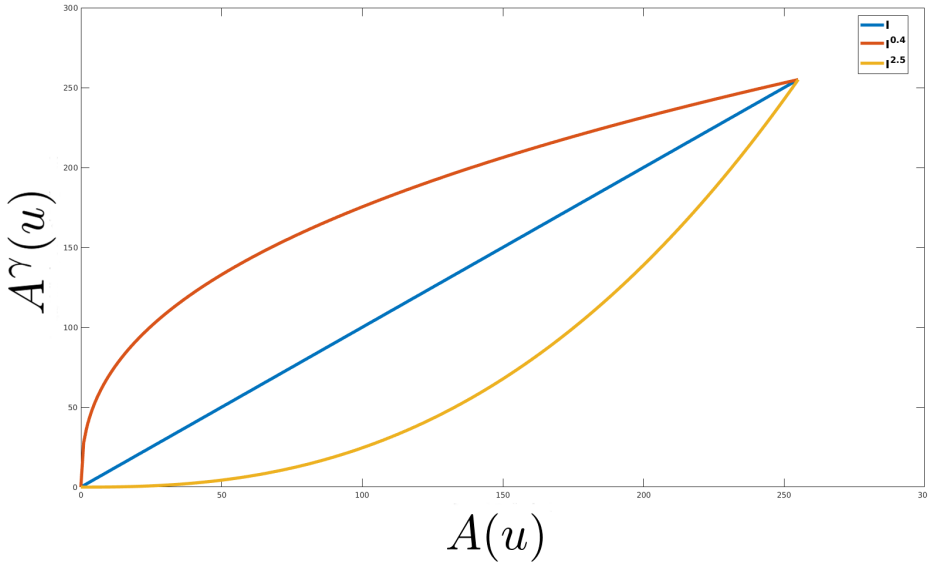


Figure 2.6. Pixel values of output image depending on gamma value

(a) A^C_{AG} (b) A^S_{AG} (c) A^N_{AG}

Figure 2.7. Images Enhanced with Adaptive Gamma Correction

2.8. Experimental Results

In this section, experimental results which consists of three set of measurements are presented. First, the distance of visual descriptors has been examined before and after each illumination invariance approach. Second, place detection is performed for each illumination compensation method covering four datasets together with real time robot experiment. Lastly, each place is revisited in different illumination conditions in order to measure place recognition performance of each method.

2.8.1. Visual Descriptor Invariance

In this section, the similarity of appearances under different illumination conditions is compared. For this, 19 different scenes and two different measures are considered. The measures are computed based on the χ^2-distance between the descriptors with corresponding appearances. In the first measure, appearances of a scene under different illumination conditions are considered. If the appearances have become truly illumination-invariant, their χ^2-distance should be low which means that the appearances are similar. The measure $M_1(T_1, T_2)$, where $T_1, T_2 \in \{S, C, N\}$ is defined by the average similarity as:

$$M_1(T_1, T_2) = \frac{1}{N} \sum_{j=1}^{N} \chi^2(I_j^{T_1}, I_j^{T_2})$$

where $I_j^T \in R^d$ is the bubble descriptor associated with the appearance of scene j under illumination condition $T \in \{S, C, N\}$. Appearances are more similar as seen from the respective illumination conditions, as $M_1(T_1, T_2)$ approaches 0.

For control purposes, we also consider the dissimilarity of appearances from different places under the same illumination conditions [36]. As such, we would like to assess how much dissimilarity is accounted for by only scene changes in comparison to only illumination changes. In this case, differing from the first measure, the illumination invariance method should be such that the similarity among appearances from different places is low. This is measured by $M_2(T)$, $T \in \{S, C, N\}$ defined as:

$$M_2(T) = \frac{2}{N(N-1)} \sum_{\substack{j=1 \\ j \neq k}}^{N} \sum_{k=1}^{N} \chi^2(I_j^C, I_k^C)$$

Similarly, dissimilarity becomes higher, as $M_2(T)$ increases.

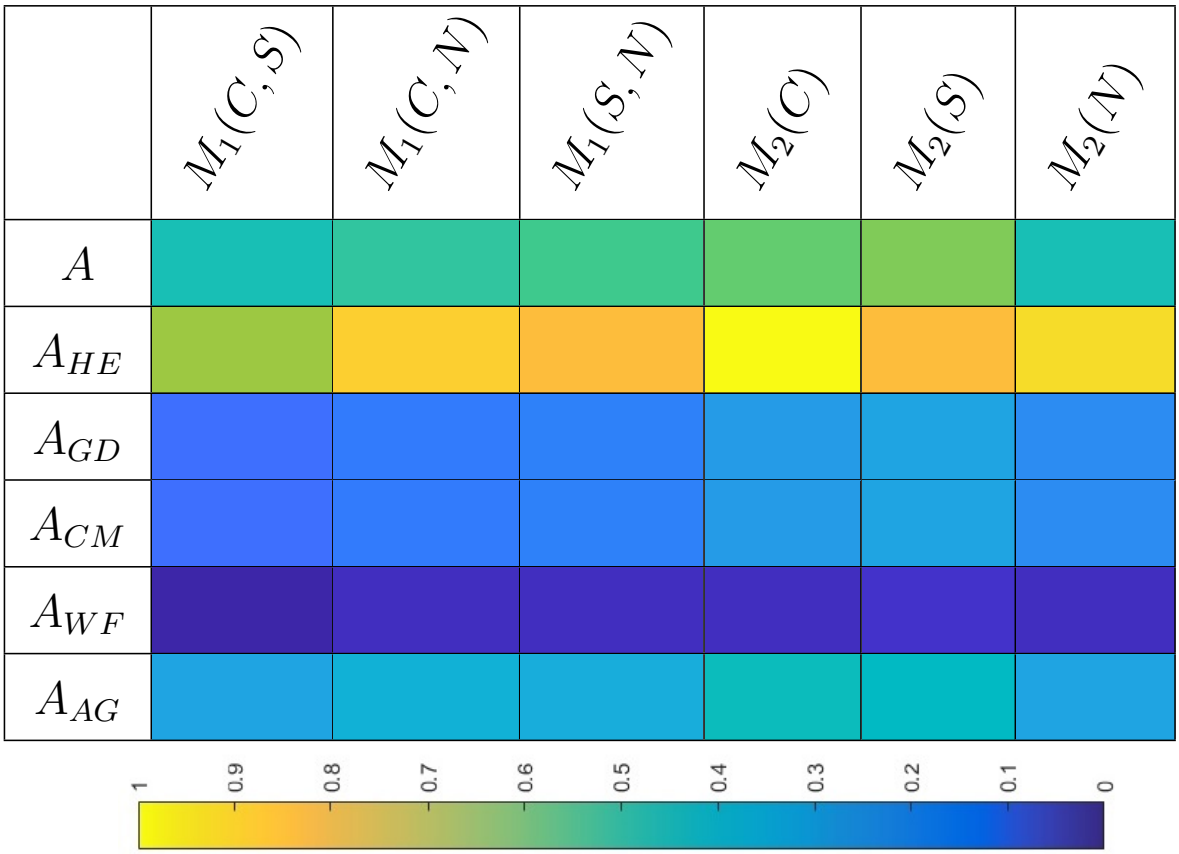


Figure 2.8. Appearance similarity vs varied illumination invariance methods. The values range from 0 (dark blue) to 1 (light yellow). The first three columns consider appearances under varied illumination while the remaining three columns consider different appearances under the same illumination. In the former case, their differences should be as low as possible while in the latter, the differences should be as high as possible.

The results are presented in Figure 2.8. For ease of visualization, they are color coded in interval [0,1]. It is observed that without applying any illumination-invariant methods, the values $M_1(C,S)$, $M_1(C,N)$ and $M_1(S,N)$ are slightly lower than $M_2(C)$, $M_2(S)$ and $M_2(N)$ values. This suggests that without applying any illumination invariance, the robot is highly likely to get confused whether appearances are from a previously seen place or not, should their illumination conditions are different. As expected, appearances from cloudy and sunny illumination seem to be most similar, as the measure $M_1(C,S)$ is lowest among them. With illumination invariance methods applied, these appearances become more similar. However, performance varies. With histogram equalization, while the similarities of differently illuminated scenes are low, the dissimilarities of appearances of different scenes are more pronounced compared to the original appearances. With the gamma-normalized DCT method, the similarities are increased for each of the different illumination-pairs around 10-15%. Not surprisingly, this method also has low $M_2(T)$ values which suggests that it tends to increase the similarity between different scenes. In camera-modeling and Wiener filtering methods, the similarities of appearances under different illumination conditions improve since $M_1(C,S)$, $M_1(C,N)$ and $M_1(S,N)$ are lower. However, $M_2(T)$ are not as high as they need to be, which suggests that discrimination among different scenes will be harder. With adaptive gamma correction, while the similarities of appearances under different illumination conditions are slightly less compared to last two methods, $M_1(C,S)$, $M_1(C,N)$ and $M_1(S,N)$ both remain unchanged relatively with each other, which suggests that appearances from different scenes are less affected.Another observation is that scenes under night illumination tend to be less similar than those under cloudy and sunny illumination with histogram equalization or gamma-normalized DCT. Nevertheless, this is not a major concern as their dissimilarities are greater than those under the same illumination conditions. With camera modeling and Wiener filtering methods, different scenes under night illumination are deemed to be more similar than the same scenes under different illumination conditions which implies that these methods are not adaptive enough for night imagery.

2.8.2. Place Detection

Next, a comparative study of illumination invariant methods is conducted as they perform in automated place detection. The goal of place detection is to delineate distinct places without any prior knowledge regarding the extent of distinct places. It thus provides a basis for place recognition, topological mapping [8], visual navigation [37] and higher level spatial reasoning such as semantic scene understanding [12,38]. Nevertheless, in some approaches, place detection and recognition are solved together [10,11]. As such, detecting places is an integral capability, should robots are to become spatially [13] and socially aware [39].

The method used for place detection is presented in [40]. Consider a robot that navigates through a sequence of base points as indexed by the set $\mathcal{K}$. Here, places are detected through partitioning the index set $\mathcal{K}$ so that appearances belonging to each distinct place are grouped together as $\mathcal{D} \subset \mathcal{K}$. This is achieved by the iterative clustering of the index set $\mathcal{K}$ - considering the informativeness, coherency and plenitude of the associated visual descriptors that are obtained from the incoming sequence of appearances. The clustering is done by identifying maximally coherent descriptor neighborhoods that correspond to distinct places and temporal windows that correspond to in-between transitions. Uninformative descriptors that arise in problematic environmental conditions such as low illumination or viewpoint (robot looking at a large object or being very close to one) are not taken into consideration. During place detection, even though illumination conditions such as sunny or cloudy weather will not change in general, appearances are nevertheless subject to abrupt illumination changes due to relative position of illumination sources and scene contents.

Place detection performance is evaluated in comparison to the ground truth, which is obtained manually by visual annotation. Note that the number of places is not necessarily the same across all subsets of COLD dataset, since robot's path is not exactly the same across different illumination conditions in a single site. This is due to the robot's path being slightly altered from on shot to the other, so that it appears that more (or less) places might exist on the same trajectory. Moreover, there exist

frames in the night dataset where the illumination conditions is impossible to reverse, e.g no light source . For each site, a summary of ground truth is provided in Table 2.1.

Table 2.1. Ground Truth for the Number of Detected Places in each Site

Site	Cloudy	Sunny	Night/Evening
Fr	21	23	24
Lj	15	13	13
Sa	12	16	15
NC	10	N/A	10

Performance results are represented in Table 2.2. A detected place D is determined to be correct if there is at least 50% overlap with the corresponding ground truth $G \subset \mathcal{K}$, namely $\frac{|D \bigcap G|}{|G|} \geq 0.5$. Furthermore, they are constrained to have compatible extents. This is checked as $\frac{|D|}{|G|} \leq \tau^+$. τ^+ is the upper bound of extent ratio and is set to $\tau^+ = 2$. It guarantees that the number of appearances associated with a detected place can be at most two times of those corresponding to the ground truth. Hence, detected places having an extremely long extents are not considered to be correct detections.

In Fr site, gamma normalized DCT and adaptive gamma correction methods outperform the other methods with respect to both precision and recall rates. Wiener filtering fails to detect transition points between places and performs worse as compared to applying no illumination invariance method. With camera sensor modeling method, performance is slightly better for cloudy and sunny conditions, but there is complete failure at night. In Lj site, robot's zigzag movement affects detection rates of all methods dramatically. Here, histogram equalization has best performance with adaptive gamma correction being the second. Gamma normalized DCT yields significant improvement only on precision rates, but not on recall ones, which means it detects much more places than it should. Camera modeling method does not boost the performance effectively. Wiener filtering exhibits lesser performance degradation since adjacent frames are likely to have similar appearance with each other. Similar to the Fr site, adaptive gamma correction demonstrates a superior performance also in Sa

Table 2.2. Comparative Place Detection Performance

(a) Precision Rates at Different Sites under Varying Illumination Conditions

		A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
Fr	Cloudy	0.68	0.86	0.6	0.77	0.44	0.82
	Sunny	0.65	0.7	0.75	0.72	0.42	0.77
	Night	0.71	0.67	0.68	0.33	0.2	0.76
Lj	Cloudy	0.28	0.25	0.42	0.23	0.44	0.26
	Sunny	0.41	0.57	0.56	0.5	0.67	0.37
	Night	0.29	0.36	0.50	0.45	0.4	0.45
Sa	Cloudy	0.53	0.65	0.53	0.4	0.43	0.63
	Sunny	0.65	0.72	0.61	0.33	0.4	0.78
	Night	0.71	0.79	0.73	0.63	0	0.86
NC	Cloudy	0.6	0.5	0.6	0.9	0.7	0.6
	Evening	0.5	0.7	0.6	1	0.8	0.7

(b) Precision Rates at Different Sites under Varying Illumination Conditions

		A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
Fr	Cloudy	0.79	0.95	0.95	0.89	0.42	0.95
	Sunny	0.63	0.58	1.00	0.75	0.33	0.83
	Night	0.71	0.76	0.71	0.19	0.1	0.90
Lj	Cloudy	0.47	0.60	0.53	0.33	0.27	0.60
	Sunny	0.54	0.92	0.69	0.46	0.31	0.77
	Night	0.38	0.69	0.38	0.38	0.15	0.38
Sa	Cloudy	0.67	0.92	0.75	0.33	0.25	0.83
	Sunny	0.81	0.81	0.69	0.19	0.25	0.88
	Night	0.67	0.73	0.73	0.33	0	0.80
NC	Cloudy	0.46	0.42	0.55	0.9	0.64	0.55
	Evening	0.42	0.58	0.5	1	0.73	0.58

site. Histogram equalization outperforms in cloudy dataset, and produces satisfactory results for other conditions. The performance of Wiener filtering and camera modeling methods worsen. This is attributed to the fact since appearances in Sa site are similar to each other, their similarity increases through employing these methods. As such, it becomes easier for the robot to miss transition points. Interestingly, performance changes in the outdoors NC site. Here, camera modeling has best performance followed by Wiener filtering. Comparing to indoor datasets, NC site has smoother image regions such as sky and ground and thus highly consists of low frequency elements. Other methods perform closely to unmodified images.

2.8.3. Place Recognition

2.8.3.1. Dataset Experiments. The experiments are done in each site by having the robot visit the different sites multiple times; first under cloudy illumination followed by sunny and night illumination for the COLD sites, and early morning followed by evening in the NC site. Thus, we expect the robot to learn most of the places that exists in first visit. Furthermore, the robot does not necessarily revisit all the learned places. Furthermore, the routes of revisits go through some places that are visited for the first-time. If the robot detects these places are detected as distinct places, we expect the robot to learn them. The places in the COLD sites are challenging, since the scenes are likely to have dynamic entities such as chairs or doors with changed positions in the revisits. New College dataset has lesser difference between cloudy and evening illumination conditions and has lesser dynamic entities.

A summary of recognition and learning performance is given in Table 2.3.In order to maximize precision and thus minimize wrong learning, $\tau_r = 1.8$ value is used. In the first-time visits, the number of detected places is given by the sum of recognized and learned places. As explained, any recognition during first visit is considered to be false. For the revisits, the number of the detected places is provided. The number of detected places that are visited for the first-time along the revisit routes are indicated in parentheses as determined by an external user. Two remarks are noteworthy here. First, the robot does not have any a priori information regarding whether a detected

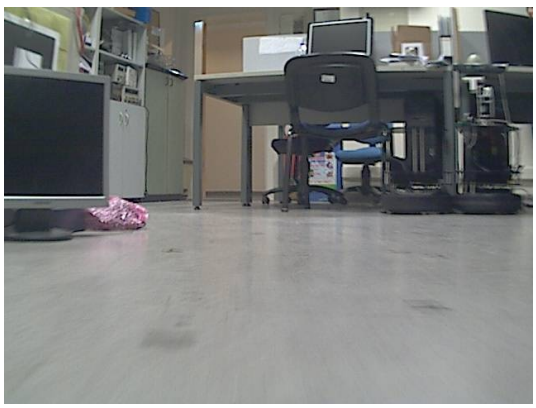
place has been previously visited or not. Rather, robot is supposed to detect these places and should determine this on its own through recognition - namely it should not be recognized as any of these places. Second, different methods vary in the number of detected places that are visited for the first-time. This is because, in some cases, these places are detected as being part of a place that has been previously visited - through external inspection. NC site does not contain any new place during revisit. Nevertheless, Fr site does include two new places during the second visit and one more new place during third visit. Therefore, number of newly detected places – namely the number indicated in parentheses – could be at most 2 for second visit and 1 for third visit for Fr site. Number of new places visited which appears on revisits should be equal to the number of new detected places mentioned in 2.3. All the methods except Wiener filtering approach and histogram equalization in second visit, while all methods except Wiener filtering and camera sensor modeling approaches in third visit successfully detect these new places.

The first revisit is done under sunny illumination. We expect the number of recognized places to be equal the number of detected places that are first-time visited. Recognition performance is given by the number of correct and wrong recognition decisions. It is observed that with no illumination invariance, only about one thirds of the places are recognized in revisits which means the robot learns the rest as new places. Furthermore, seven places out of nine are correctly recognized. In outdoors, recognition performance is better, so the robot does not learn the revisited places. However, accuracy is worse since the robot correctly recognizes six places out of the eleven. With histogram equalization, recognition performance is better, since the robot learns a smaller number of previously visited places as new places. However, the accuracy of recognition is around fifty percent. Gamma normalized DCT performs similar to histogram equalization. The performance of both camera sensor modeling and Wiener filtering are inadequate both in indoor locations and night condition as indicated by high false recognition rates. With adaptive gamma correction, indoors, true recognition rate is much higher with lower false recognition rate. Outdoors, false positive rate is low, while true positives remain constant. Considering combined datasets, adaptive gamma correction is the best among all methods.

We also study precision and recall performance. In Fr site, the resulting precision and recall tables are given in Table 2.4. We have satisfactory precision rates, but the algorithm comes up with a low recall rate. On the other hand, we observe that camera modeling and Wiener filtering approaches have both poor recognition for indoor datasets, since images produced with these methods are very close to each other in terms of appearance as seen in Figure 2.4 and Figure 2.5. Adaptive gamma correction clearly outperforms other methods' precision - recall rates for indoor sites. We can achieve perfect precision rate with a very low recall rates by using original images. The results for Lj site are given in Table2.5. Here, differing from Fr site, recall rates are high while precision is low. The zigzag movement of the robot confuses robot and it results with false positives in Lj site. Similar to Fr site, adaptive gamma correction method is the best approach to choose in this site. Histogram equalization and gamma normalized DCT methods has similar results as the original images while camera modeling and Wiener filtering methods fail. In the Sa site, there are less place than other sites and these are similar to each other in terms of appearance. Therefore, Wiener filtering method and original images fail. Surprisingly, camera sensor modeling approach performs better than gamma normalized method. Adaptive gamma correction method comes the first place for Sa site as seen in Table 2.6. Outdoors appearances have different characteristics as compared to those from indoors, hence NC site comes up with quite different performance order. This is attributed to the smoothness of image texture. Both gamma normalized DCT and camera modeling methods generate quite similar images under different illumination conditions and thus demonstrate comparatively high recall rates, as seen in Table 2.7. However, both of their precision rates decrease. Interestingly, while adaptive gamma correction method does not have as high recall rates, its precision rates seems to be persistent as it is maintained around 90% in all the cases.

2.8.3.2. Real Time Experiments. The second set of experiments is done with Jaguar robot. Place detection and recognition modules with/without adaptive gamma operates on the captures frames while navigating on real time. The robot follows approximately 100 meter path and captured 2419 base points. The robot has a threaded

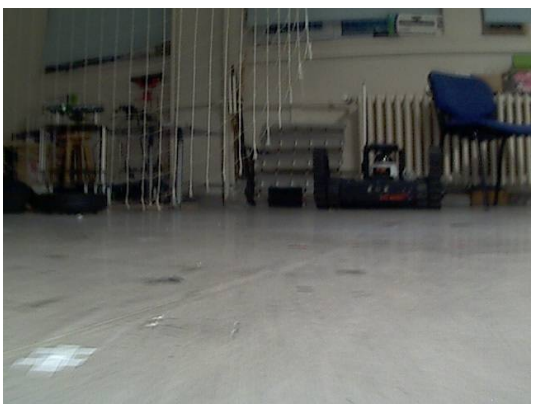
motion system with a maximum speed of 1.4 meters/second. The robot has standard front camera, PTZ camera. In this implementation, only the front camera of the robot is used as sensory input. Tele-operation is done with a remote mobile computer with Intel-i7 4770HQ processor. Notice that there is a wired unidirectional communication between the sensors and the robot while the remote computer has wireless bi-directional communication. The experiments are done in North Campus of Boğaziçi University. Experiments consists of 10 places for both visits where 2 of these places are revisited within the same visit. In the second visit, all of the 10 places visited in the first visit is revisited again. Hence, it is expected that the robot is able to recognize 2 places in its first visit and recognize all 10 places in its second visit for such an ideal case. Sample place appearances are shown in 2.9. The acquisition of the appearances has better condition with respect to frames in the dataset, since avoided zigzag movements and very close range uninformative frames. Results of the conducted experiments are shown in 2.8. First visit without illumination invariance has one true recognition which corresponds to %100 precision and %50 recall, while second visit results with 6 true recognitions and 2 false ones which corresponds to %75 precision as well as %46 recall rate. Precision and recall rates with adaptive gamma illumination invariant approach performs perfectly for the first visit with %100 precision and recall rates. Also, experiment conducted with adaptive gamma produces more precise detection results, hence it results with higher recall rates with respect to former case. %78 precision and %64 recall rates are acquired in the second visit with adaptive gamma correction. These results reveal that illumination invariance has significant impact on the place detection, and hence the precision rates though it barely affects recall rate.



(a) Place 1 Appearance

(b) Place 4 Appearance

(c) Place 7 Appearance



(d) Place 10 Appearance

Figure 2.9. Sample Appearances from Real time Jaguar Visits

Table 2.3. Comparative Learning & Recognition Performance(T: True, F: False). In Cases of Revisits, Detected Places that are First-time Visited are Indicated by parentheses

(a) Without illumination invariance

Site	Illumination							
	Cloudy/Morning (1^{st} visit)		Sunny (Next visit)			Night/Evening (Last visit)		
	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized
Fr	22	0	25(2)	16	7T + 4F	25(1)	16	7T + 4F
NC	11	0	-			11(0)	0	6T + 5F

(b) Histogram equalization

Site	Illumination							
	Cloudy/Morning (1^{st} visit)		Sunny (Next visit)			Night/Evening (Last visit)		
	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized
Fr	21	0	20(1)	9	6T + 5F	24(1)	9	8T + 7F
NC	11	0	-			12(0)	0	6T + 6F

(c) GammaDCT

Site	Illumination							
	Cloudy/Morning (1^{st} visit)		Sunny (Next visit)			Night/Evening (Last visit)		
	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized
Fr	21	0	26(2)	8	10T + 8F	23(1)	13	6T + 4F
NC	14	0	-			16(0)	0	10T + 6F

(d) Camera sensor modeling

Site	Illumination							
	Cloudy/Morning (1^{st} visit)		Sunny (Next visit)			Night/Evening (Last visit)		
	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized
Fr	20	2F	25(2)	15	5T + 5F	12(0)	6	2T + 4F
NC	13	0	-			14(0)	0	9T + 5F

(e) Wiener Filtering

Site	Illumination							
	Cloudy/Morning (1^{st} visit)		Sunny (Next visit)			Night/Evening (Last visit)		
	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized
Fr	18	3F	19(1)	10	5T + 4F	10(0)	5	2T + 3F
NC	14	0	-			13(0)	0	9T + 4F

(f) Adaptive Gamma Correction

Site	Illumination							
	Cloudy/Morning (1^{st} visit)		Sunny (Next visit)			Night/Evening (Last visit)		
	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized	# Detected Places	# Places Learned	# Places Recognized
Fr	22	0	23(2)	12	10T + 1F	21(1)	10	10T + 1F
NC	8	0	-			8(0)	0	6T + 1F

Table 2.4. Performance of Illumination Invariance Methods at Freiburg Site

(a) Precision Rates with Varying τ_r

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.78	0.54	0.57	0.37	0.39	0.91
1.2	0.78	0.56	0.55	0.33	0.39	0.90
1.4	0.83	0.67	0.60	0.36	0.40	0.89
1.6	0.80	0.67	0.50	0.30	0.36	0.86
1.8	1.00	1.00	0.40	0.33	0.40	0.86

(b) Recall Rates with Varying τ_r

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.33	0.33	0.38	0.33	0.33	0.48
1.2	0.33	0.24	0.29	0.24	0.33	0.43
1.4	0.24	0.19	0.29	0.24	0.29	0.38
1.6	0.19	0.19	0.19	0.14	0.19	0.29
1.8	0.19	0.14	0.10	0.14	0.19	0.29

Table 2.5. Performance of Illumination Invariance Methods at Ljubljana Site

(a) Precision Rates with Varying τ_r

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.37	0.50	0.50	0.38	0.42	0.55
1.2	0.39	0.50	0.50	0.33	0.36	0.57
1.4	0.38	0.44	0.50	0.33	0.36	0.59
1.6	0.43	0.41	0.50	0.31	0.29	0.62
1.8	0.45	0.44	0.45	0.33	0.29	0.62

(b) Recall Rates with Varying τ_r

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.50	0.71	0.71	0.64	0.36	0.86
1.2	0.50	0.71	0.57	0.50	0.29	0.86
1.4	0.43	0.57	0.57	0.43	0.29	0.71
1.6	0.43	0.50	0.50	0.36	0.14	0.57
1.8	0.36	0.50	0.36	0.36	0.14	0.57

Table 2.6. Performance of Illumination Invariance Methods at Saarbrücken Site

(a) Precision Rates with Varying τ_r

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.27	0.50	0.42	0.50	0.07	0.53
1.2	0.33	0.60	0.45	0.55	0	0.62
1.4	0.33	0.63	0.45	0.50	0	0.70
1.6	0.43	0.60	0.44	0.50	0	0.67
1.8	0.40	0.75	0.50	0.50	0	0.80

(b) Recall Rates with Varying τ_r

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.23	0.46	0.38	0.46	0.08	0.62
1.2	0.23	0.46	0.38	0.46	0	0.62
1.4	0.23	0.38	0.38	0.31	0	0.54
1.6	0.23	0.23	0.31	0.31	0	0.46
1.8	0.15	0.23	0.23	0.23	0	0.31

Table 2.7. Performance of Illumination Invariance Methods at New College Site

(a) Precision Rates with Varying τ_r

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.55	0.50	0.63	0.64	0.69	0.86
1.2	0.44	0.45	0.71	0.67	0.73	0.86
1.4	0.50	0.43	0.75	0.73	0.70	0.83
1.6	0.60	0.60	0.88	0.89	0.71	0.83
1.8	1.00	0.60	1.00	0.86	0.80	1.00

(b) Recall Rates with Varying τ_r

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.60	0.60	1.00	0.90	0.90	0.60
1.2	0.40	0.50	1.00	0.80	0.80	0.60
1.4	0.30	0.30	0.90	0.80	0.70	0.50
1.6	0.30	0.30	0.70	0.80	0.50	0.50
1.8	0.10	0.30	0.50	0.60	0.40	0.40

Table 2.8. Number of Detected, Learned and Recognized Places for Real Time Jaguar Experiments with Unmodified Appearances and Adaptive Gamma Approach

	A		A_{AG}	
	First Visit	Second Visit	First Visit	Second Visit
# Detected Places	12	13	10	11
# Learned Places	11	5	8	2
# Recognized Places	1T	6T+2F	2T	7T + 2F

3. RECONSTRUCTING LONG TERM SPATIAL MEMORY

3.1. Introduction

The main focus of this chapter is to represent places in such a way that the robot can flexibly either recognize or learn them as it should. Differing from the recognition and learning method mentioned in the previous chapter, each place is viewed as possibly being comprised of a set of characterizing subplaces. The set of appearances having the common visual features are defined as a subplace which coincides with one distinct scene in the place. The robot determines the subplaces based on hierarchical clustering of the appearances that are collected during place detection. Because of the size and redundancy of the collected set of appearances, this study address how to use this set as to enable efficient and flexible spatial reasoning.

The definition of a place as a spatial region is adopted in this study and it is sought that how to better represent appearance-based place knowledge as to overcome the problem of partial overlap as to have more flexible recognition. The robot is assumed to be completely autonomous in its spatial cognition - namely it detects, recognizes or learns places completely on its own through relating to its place memory using TSC model [8]. In this model, the robot retains its knowledge of places in its place memory with an hierarchical organization.

Additionally, proposed approach brings two fundamental extensions by considering each place as being comprised of a set of subplaces:

- First, once a place is detected, the robot also determines the characterizing subplaces. Each subplace corresponds to one distinct scene in the place as defined by a set of appearances sharing common visual signatures.
- Second, both recognition and learning are modified as to take the knowledge of characterizing subplaces into account.

The robot determines them through hierarchical clustering in the appearance space [14]. The splitting level with the maximal change in the resulting hierarchy naturally defines the candidate subplaces in the detected place. This method is preferred as it is both incremental and order-invariant. Both recognition and learning methods are then modified in order to accommodate the new definition of a place. In recognition, the robot uses the subplaces in relating to its place memory. In learning, the robot learns each place as a union of subplaces by incorporating this knowledge into its place memory accordingly. For example, subplace knowledge is added to the place memory as shown in Figure 3.9. As such, the robot gains the ability to recognize a place even if the current appearances are only partially overlapping with those that were used in learning.

3.2. Related Literature

In this part, a solution to extract canonical views from a place is sought. Thus, a place could be described as the collection of visual appearances that shares common perceptual signatures [41]. In most applications, the robot traverses along a previously specified path [8, 42–44]. Alternatively, robot explores the places autonomously with the help of random walk [45], greedy mapping [46–48], frontier based approaches [49–52] or topological approaches [53]. However, in all of these approaches robot arbitrarily collects visual data based on distance or time. However, this causes a place to be large and redundant set of appearances. There are many approaches have been proposed to solve this problem. Singular value decomposition is employed in order to summarize videos and obtain keyframes, where the approach only works offline and needs entire video stream as a whole [54]. In mobile robotics, Girdhar and Dudek [55], proposed an online method that extracts salient images from an image set using set-theoretic surprise and measures the suitability of an image. In another method, Bayesian surprise is utilized in order to indicate salient landmarks via vision and laser features for topological mapping [56]. Konolige et al. [57] proposed an online large-scale mapping technique for constructing topological maps with stereo data. Another solution set for extracting salient images is topic modeling, that is proposed in several papers. In [58],

topics are learned online, but salient images are culled out after updating the topic model by associating documents with the updated topics. In a similar approach, each image is characterized as a mixture of visual topics, maps to a point in topic vector space. Streamed images are incrementally organized with an online graph clustering method [59, 60]. Former approach summarizes set of images into a single image where it is not able to represent a place as a spatial area and latter approaches aim to obtain salient images in order to reduce computational complexity and achieve a compact memory but not to better recognition performance.

In this part of the thesis, a method is proposed that differs from the previously proposed methods. Given the visual stream, robot needs to extract set of canonical images from the for better memory management and recognition performance. More particularly, our goal is summarizing continuous image streams in order to gather images that have similar appearances. This scenario is common for robot applications requiring high accuracy and large scale appearance-based detection, recognition and mapping applications. Moreover, we seek an incremental online approach that keeps computation cost as low as possible while salient set of images, since real time robotic applications is aimed by this work. In this study, we proposed an online, hierarchical and incremental algorithm that clusters detected places with a tree structure algorithm for generating canonical visual appearances. Additionally, we combined the clustering approach with several illumination-invariant algorithms for comparative results in both indoor and outdoor datasets. The main idea is to extract semantically related images to construct representatives from the image stream, called as subplaces, such that every image belongs to subplace has a similar appearance. We proposed a fast and stable algorithm for determining subplaces. The results obtained by clustering places into subplaces yields significant improvement as we compare the results with original stream without clustering. In summary, the main contributions of this part are:

- A fast algorithm where computation of subplaces takes approximately linear time for large datasets and data streams.
- Evaluation of the approach on publicly available datasets from a mobile platform demonstrating significant performance increment.

3.3. Place Memory

Once places are detected as mentioned in 2.8.2, each place $D \in \mathcal{D}$ is kept in place memory in hierarchical manner. Single linkage clustering algorithm is used in order to constitute this hierarchical structure [14]. This structure helps to recognize places much more faster as it decrease computation time during place recognition to $\mathcal{O}(NlogN)$ instead of $\mathcal{O}(N^2)$ which is explained in next section in details. Another advantage of the SLINK algorithm is that tree can be updated incrementally as new places detected and thus place detection and recognition can be performed simultaneously. The system was designed to construct whole tree at once as new place detected, but incremental SLINK is replaced instead which decreases computation time from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ which saves significant time for large set of places. Pseudo code of incremental SLINK algorithm is shown below. $\pi(i)$ denotes the maximum index where i^{th} node linked to $\pi(i) > i$, and $\lambda(i)$ denotes the distance between these nodes.

Place memory is constructed for Freiburg, Ljubljana, Saarbrücken and New College sites. Places in the cloudy, sunny and night sequences are learned cumulatively for each site, respectively. Hence, total number of dendrograms for Freiburg,Ljubljana, Saarbrücken and New College sites are obtained and demonstrated in Figure 3.1, 3.2, 3.3 respectively. It is observed that as distances get smaller, corresponding appearances get similar.

3.4. Determining Subplaces

The robot considers each place p as possibly being characterized by a set of n_s subplaces. The number n_s can vary between one to many depending on the spatial extent of the place relative to the robot as well as obstructions contained therein. For example, in the home site shown in Figure 3.9, place 2 and place 3 contain only one subplace while the remaining places have multiple subplaces.

The robot detects subplaces through the hierarchical clustering of appearances collected during place detection. Each subplace j is defined by a subset $D_j \subseteq D$ of

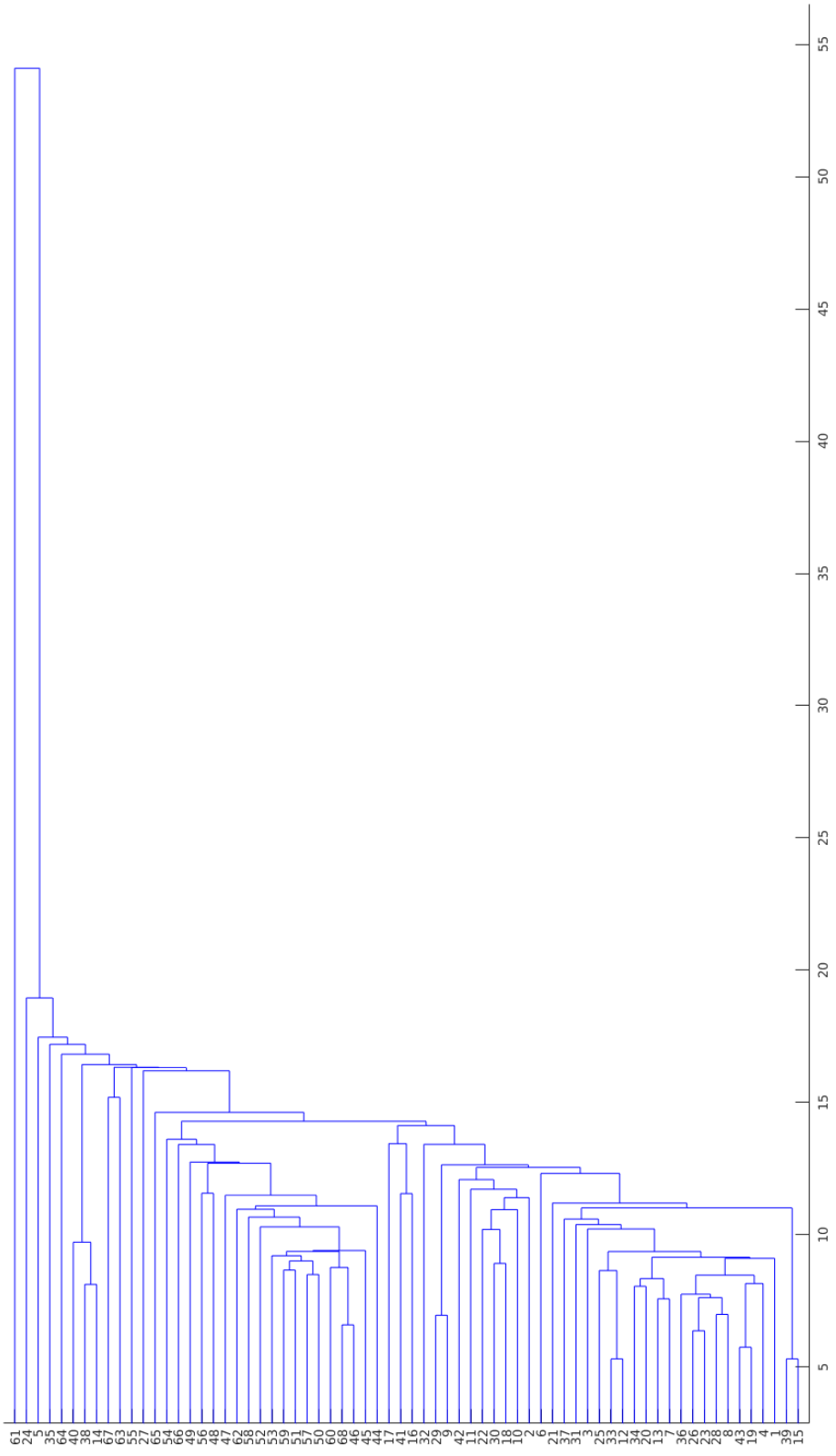


Figure 3.1. Place Memory Constructed for Freiburg Site

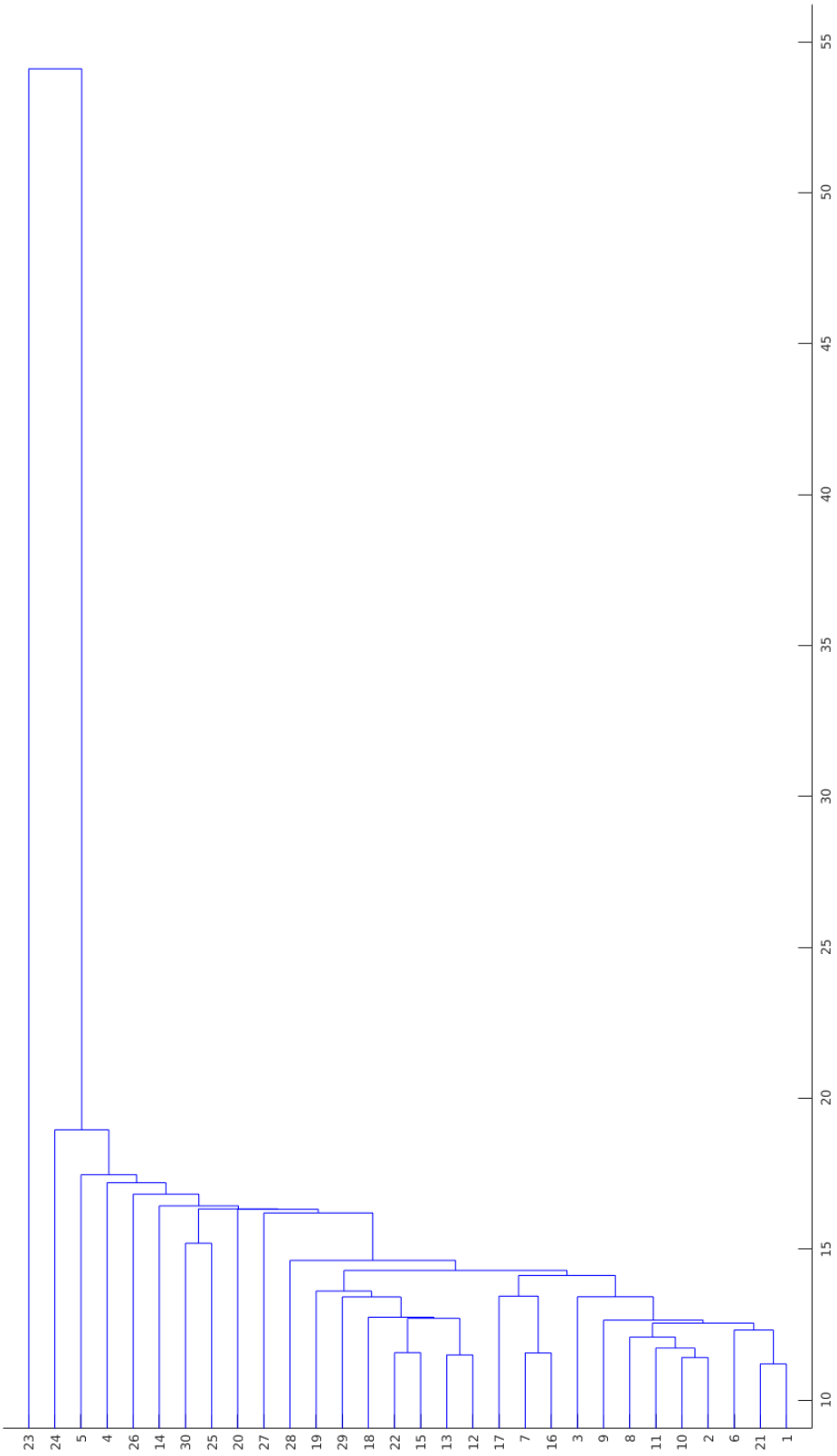


Figure 3.2. Place Memory Constructed for Ljubljana Site

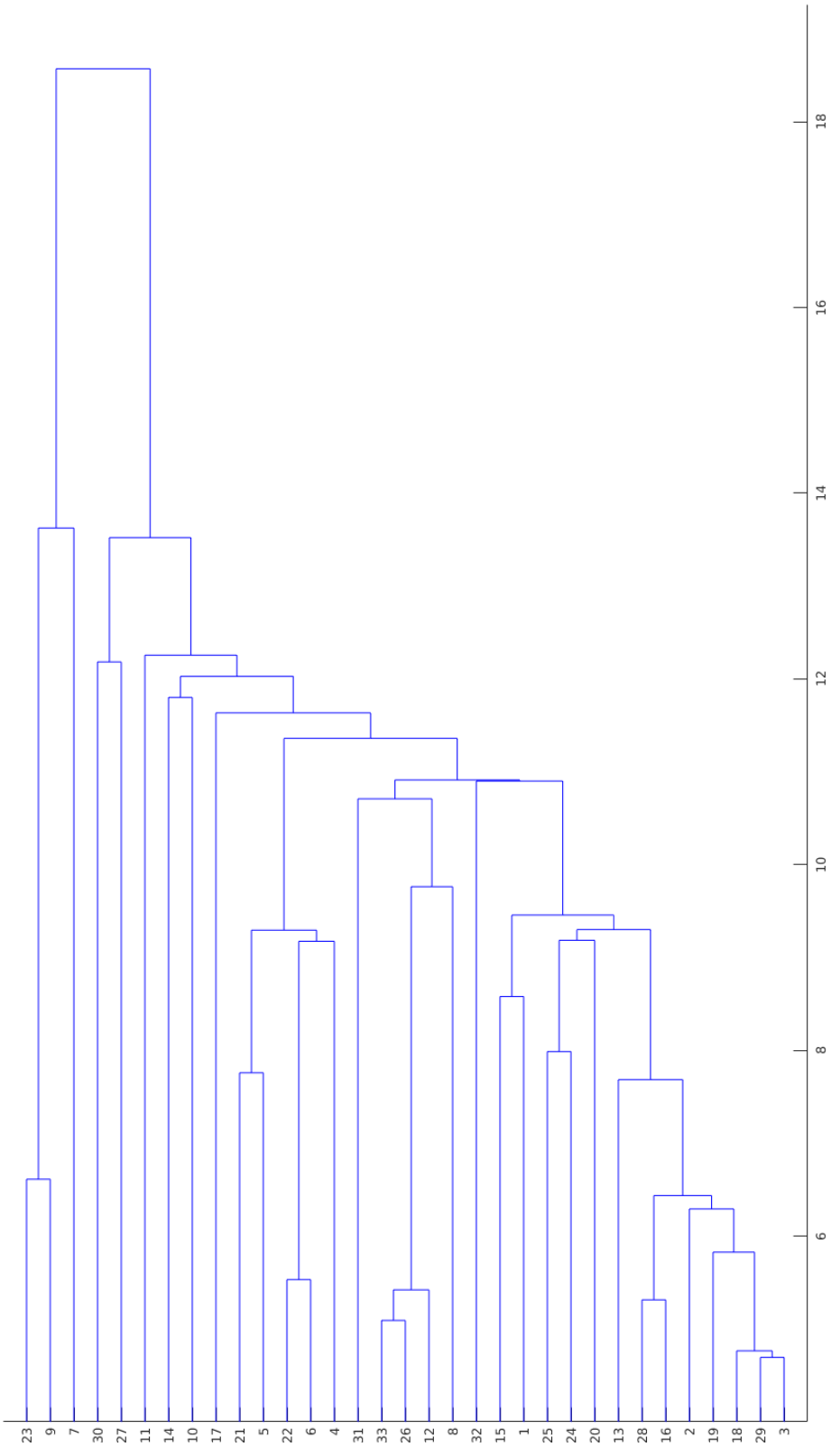


Figure 3.3. Place Memory Constructed for Saarbrücken Site

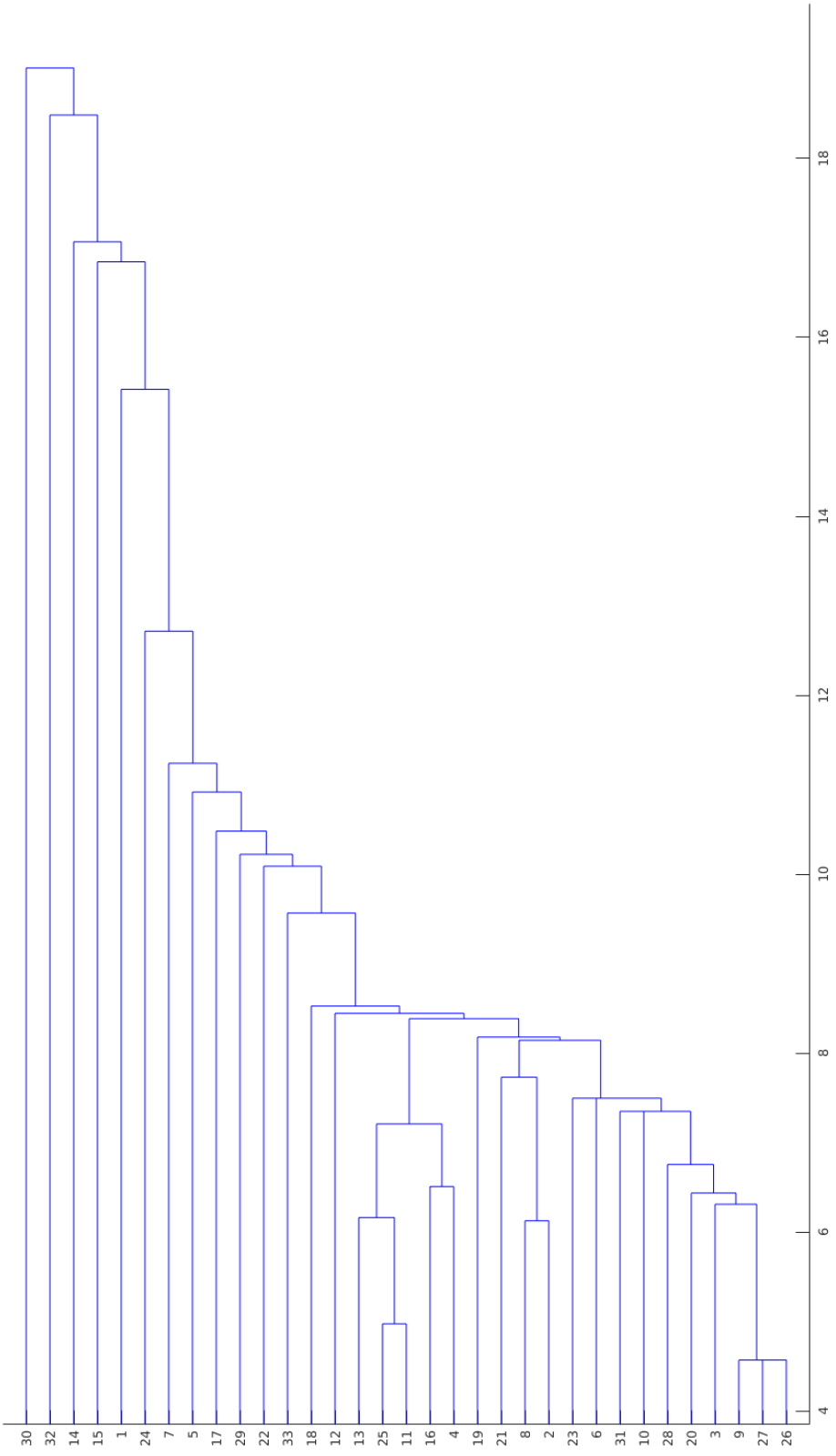


Figure 3.4. Place Memory Constructed for New College Site

these appearances. Thus, if the robot determines n_s subplaces, then $\cup_{j=1}^{n_s} D_j \subseteq D$. The hierarchy is obtained by a nested sequence of partitions of D in the appearance space based on SLINK [14]. The partitions are affiliated with numerical levels as measured by a similarity metric in the appearance space. The partition at the highest level corresponds to the whole D. Let $E(D)$ denote the set of equivalence relations on D. The partitioning function $\zeta : R^{\geq 0} \rightarrow E(D)$ satisfies the following constraints:

$$\begin{aligned} &1. \quad 0 \leq h \leq h' \rightarrow \zeta(h) \subseteq \zeta(h') \\ &2. \quad \zeta(h + \delta h) = \zeta(h) \ \forall \delta h \approx 0 \\ &3. \quad \exists h^r \text{ s.t. } \zeta(h^r) = D \times D \end{aligned} \tag{3.1}$$

According to the first constraint, as h increases, clusters get larger. The second constraint states that clusters are formed based on discrete increments in h value. Thus, the resulting partitions are different for sufficiently different h values. The final constraint ensures that the top level is a single cluster.

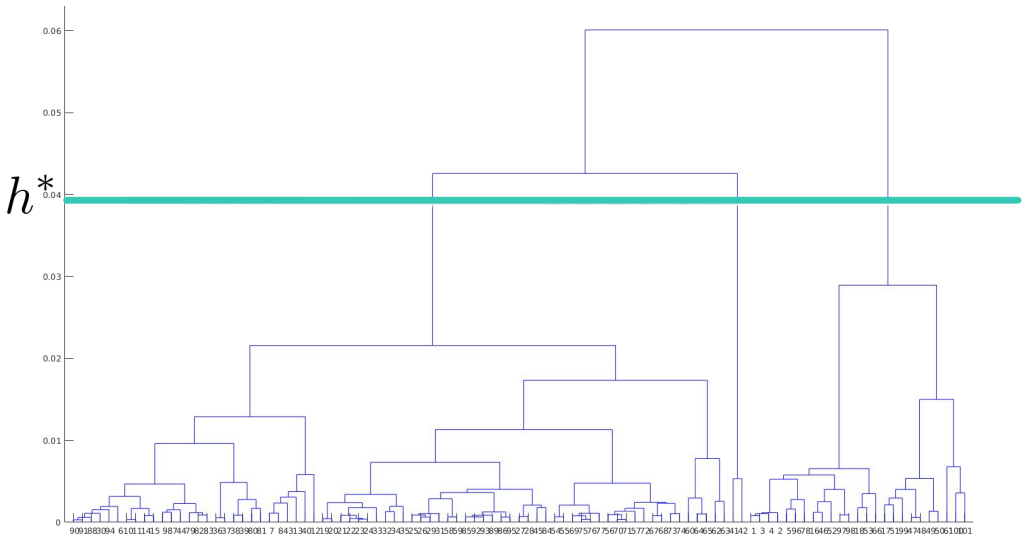


Figure 3.5. Hierarchical Clustering base on Appearances in Place 7 in Fr Site

The resulting nested sequence of partitions can be shown to be equivalent to tree hierarchy consisting of n_L levels. In the hierarchy, each node N corresponds to a particular cluster of indexes $D(N) \subseteq D$. In particular, for a fixed distance $h \in R^{\geq 0}$, consider a graph whose terminal nodes are in D, but whose edges link those pairs of

Initially there are n appearances. $n+1$ *is the new appearance which will be added.*

Initialize: $\pi(n+1) \leftarrow n+1$ and $\lambda(n+1) \leftarrow \infty$

Let $d(i)$ denotes the distance between the i^{th} node and $n+1$

for $i = 0$ to n **do**

  if $\lambda(i) \geq d(i)$ **then**

    $d(\pi(i)) \leftarrow min(d(\pi(i)), \lambda(i))$

    $\lambda(i) \leftarrow d(i)$

    $\pi(i) \leftarrow n+1$

  else

    $d(i) \leftarrow min(d(\pi(i)), d(i))$

  end if

end for

for $i = 0$ to n **do**

  if $\lambda(i) \geq \lambda(\pi(i))$ **then**

    $\lambda(i) \leftarrow n+1$

  end if

end for

Figure 3.6. Incremental SLINK Algorithm

appearances $D(N)$ and $D(N')$ having distance of at most h as measured by a distance function γ:

$$\gamma(N, N') = \|\bar{I}_N - \bar{I}_{N'}\|^2 \tag{3.2}$$

where

$$\bar{I}_N = \frac{1}{|D(N)|} \sum_{i \in D(N)} I(x_i) \tag{3.3}$$

Then $\zeta(h)$ is the equivalence relation corresponding to the partition of D defined by the connected components of this graph. A cluster is represented by the member appearance having the largest index. This is defined by the function $\pi : D \rightarrow D$ such

that $\forall i \in D$, $\pi(i)$ denotes the first cluster (as represented by the base point with the largest index) that base point i joins:

$$\pi(i) \quad = \quad \max\{j \mid \exists h > 0 \text{ s.t.}(i,j) \in \zeta(h)\}$$

This implies that if $i \neq n$, $\pi(i) > i$ while $\pi(n) = n$. The corresponding distance is encoded by the function $\lambda : D \to R^{\geq 0}$ such that $\forall i \in \mathcal{D}$, $\lambda(i)$ denotes the distance between cluster i and cluster $\pi(i)$ when they join. In particular,

$$\lambda(i) \quad = \quad \inf\{h \mid \exists j > i \text{ s.t. } (i,j) \in \zeta(h)\}$$

As such, $\lambda(n) = \infty$ while for $i < n$, $\lambda(\pi(i)) > \lambda(i)$. It can be shown that:

$$\pi(i) \quad = \quad \max\{j \mid (i,j) \in \zeta(\lambda(i))\}$$

The pointer representation also yields a tree structure. This can be verified via defining $\sigma(i,h)$ to be the first element k in the sequence $\{i, \pi(i), \pi^2(i), \ldots\}$ for which $\pi(k) > h$. Then, the function ζ can be expressed as:

$$\zeta(h) = \{(i,j) : \sigma(i,h) = \sigma(j,h)\} \tag{3.4}$$

It results with a $n-1$ non-leaf nodes for n base points. There is a 1-1 correspondence between the hierarchy and pointer representation.

The SLINK algorithm is given in Algorithm 3.6. As seen, there are three arrays where the first two correspond to λ_n and π_n in the first n locations. The third array d stores the row of a subset of pairwise distances. The arrays are updated to incorporate each new base point. Suppose that the index $n+1$ of a new base point x_{n+1} is inserted into the detected place D. The tree hierarchy changes accordingly. The iterative

functions λ_{n+1} and π_{n+1} are defined as follows:

$$\begin{aligned}
\lambda_{n+1}(i) &= \begin{cases} \infty & i = n+1 \\ \min\{\mu_n(i), \lambda_n(i)\} & i < n+1 \end{cases} \\
\pi_{n+1}(i) &= \begin{cases} n+1 & i = n+1 \\ n+1 & \mu_n(i) \leq \lambda_n(i) \\ n+1 & \mu_n(\pi_n(i)) \leq \lambda_n(i) \\ \pi_n(i) & \text{otherwise} \end{cases}
\end{aligned} \tag{3.5}$$

The function $\mu_n : D \to D$ is defined iteratively as:

$$\mu_n(i) = \min\left\{\gamma(N_i, N_{n+1}), \min_{\pi_n(j)=i} \max\{\mu_n(j), \lambda_n(j)\}\right\}$$

where N_i and N_{n+1} are the terminal nodes associated with appearances $I(x_i)$ and $I_{x_{n+1}}$ respectively. Since $\mu_n(i) \leq \gamma(N_i, N_{n+1})$, it is finite for all i. SLINK algorithm is known to have four advantages:

- SLINK has been shown to achieve theoretical order-of-magnitude bounds for both efficiency of storage and construction. The required storage is $3O(n)$ while the number of operations required to find λ_n and π_n are of order $O(n)$.
- The hierarchy can be updated incrementally as the robot is detecting a place.
- The resulting hierarchy is invariant with respect to the order of appearances.

A sample hierarchy from the place 7 is given in Figure 3.5 . In this case, the detected place is defined by appearances appearances from 101 base points as indexed by the set D .

Subplaces are determined from the tree hierarchy considering the ordering of the inner nodes with respect to the height $h \in (0, h^r]$. Note that $h = 0$ corresponds to the terminal nodes while $h = h^r$ corresponds to the root node. As such, the

heights of the inner nodes (clusters) are in between these values. With the second constraint of Eq. 3.1, the h values range in discrete increments. Let these be denoted as $\{0, h^1, h^2, \ldots, h^r\}$. Now, consider the height $h^* \in (0, h^r)$ with maximal seperation $\delta h_l = h^{l+1} - h^l$ namely:

$$h^* \in \arg\max_l \delta h_l \tag{3.6}$$

The respective equivalence relation $\zeta(h^*) \subset E(D)$ defines a partition of D as $D = \bigcup_j D_j$ where $D_j \subset D$ and $D_j \bigcap D_m = \emptyset$ for $j \neq m$. Each subplace is defined by the elements of this partition that have sufficiently large cardinality - namely

$$|D_j| > \tau_d$$

With the example of Figure 3.5, h^* is as shown. The resulting equivalence relation corresponds to a partition of three clusters. As one of the clusters does not have sufficiently large cardinality, it is discarded. Thus, the place is viewed as consisting of two subplaces. Each subplace is defined by characterizing appearances as shown in Figure 3.8a-3.8b. Note that the cluster with low cardinality contains outlier appearances as seen in Figure 3.8c. Also, the pseudocode for clustering places into subplaces is given in Algorithm 3.7.

3.5. Recognition

The robot traverses down the hierarchy of its place memory level by level. At each level, the child node $N' \in N^{\downarrow}$ with maximal appearance similarity is selected while ensuring that similarity is above a recognition threshold τ_r: However, differing from the TSC model, at each level, the appearances belonging to a detected place are not all considered altogether. Instead, appearance similarity is determined by considering each subplace D_j separately and then combining the results to make a decision:

$$N^* = \arg\max_{N' \in N^{\downarrow}} g_{N'}(D) \text{ subject to } g_{N^*}(D) > \tau_r \tag{3.7}$$

```
1: icluster ← 0
2: n ← nelements − nclusters
3: for i = 0, . . . , nclusters do
4:     count[i] ← 0
5: end for
6: for i = nelements − 2 to n do
7:     k ← tree[i].left
8:     if k >= 0 then
9:         icluster ← icluster + 1
10:    end if
11:    k ← tree[i].right
12:    if k >= 0 then
13:        icluster ← icluster + 1
14:    end if
15: end for
16: for i = 0 to n do
17:    NodeID[i] ← −1
18: end for
```

Figure 3.7. Pseudocode for Clustering Places into Canonical Views

```
19: for i = n − 1, . . . , 0 do
20:     if NodeID[i] < 0 then
21:         j ← icluster
22:         NodeID[i] ← j
23:         icluster ← icluster + 1
24:     else
25:         j ← NodeID[i]
26:     end if
27:     k ← tree[i].left
28:     if k < 0 then
29:         NodeID[−k − 1] ← j
30:     else
31:         count[j] ← count[j] + 1
32:         subPlaces[j][count[j] − 1] ← k
33:     end if
34:     k ← tree[i].right
35:     if k < 0 then
36:         NodeID(−k − 1) ← j
37:     else
38:         clusterid[k] ← j
39:         subPlaces[j][count[j] − 1] ← k
40:     end if
41: end for
```

Figure 3.7. Algorithm for Clustering Places into Canonical Views(continued)

(a) Subplace 1 (b) Subplace 2 (c) Low cardinality

Figure 3.8. Sample appearances from the two subplaces of Figure 3.5.

Appearance similarity is measured by $g_N(D)$ defined as:

$$g_{N'}(D) = \sum_{j=1}^{n_s} \frac{|D_j|}{|D|} g_{N'}(D_j) \tag{3.8}$$

The function $g_N : R^d \rightarrow R^{\geq 0}$ is defined as consisting of two terms:

$$g_N(D_j) = \rho(\bar{I}_j, \bar{I}_N) + \eta_N(D_j) \tag{3.9}$$

The first term $\rho : R^d \times R^d \rightarrow R^{\geq 0}$ is the Pearson correlation that evaluates how well overall the appearances of the subplace match those associated with the current node N:

$$\rho(\bar{I}_j, \bar{I}_N) = \frac{\sum_{l=1}^{d}(\bar{I}_{jl} - \mu(\bar{I}_j))(\bar{I}_l(N) - \mu(\bar{I}(N)))}{\sqrt{\sum_{l=1}^{d}(\bar{I}_{jl} - \mu(\bar{I}_j))^2 \sum_{l=1}^{d}(\bar{I}_l(N) - \mu(\bar{I}(N)))^2}} \tag{3.10}$$

Here, $\bar{I}_j$ is the mean descriptor associated with the subplace j:

$$\bar{I}_j = \frac{1}{|D_j|} \sum_{k \in D_j} I(x_k) \tag{3.11}$$

with entries $\bar{I}_{jl} \in R$. The term $\mu(I)$ is the mean value of the descriptor I defined as:

$$\mu(I) = \frac{1}{d} \sum_{l=1}^{d} I_l \tag{3.12}$$

The second term $\eta_N : R^d \rightarrow R^{\geq 0}$ evaluates each appearance one-by-one:

$$\eta_N(D_j) = \frac{1}{|D_j|} \sum_{k \in D_j} \nu_N(I(x_k)) \tag{3.13}$$

Here, ν_N is the sum of the result of discriminant function d_N that is based on one-class SVM [61].

$$\nu_N(j) = \begin{cases} 1 & \text{if} d_N(I(x_k)) > 0, \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

The decision-making at each level is repeated until either similarity is below the recognition threshold or the terminal level of places is reached. The former case indicates that the detected place cannot be recognized and hence the robot needs to learn it. In the latter case, suppose the recognized place decision is p. Then, the robot proceeds to compare the current subplaces with those $l \in \{1, \ldots, N_{p_s}\}$ that are associated with p. This is done by comparing their respective appearances and chooses the subplace s^* with maximal similarity:

$$s^* = \underset{l \in \{1,\ldots,N_{p_s}\}}{\arg\max} f_l(D) \tag{3.15}$$

where

$$f_l(D) = \sum_{j=1}^{n_s} \frac{|D_j|}{|D|} \rho(\bar{I}_j, \bar{I}_l) + \eta_l(D_j) \tag{3.16}$$

where I_m is the mean descriptor associated with the learned subplace l. Once a place is recognized, both the place and the subplace knowledge is augmented to incorporate the new information.

On the other hand, it brings complexity to TSC model. There are two factors that affects the complexity: first one is the number of basepoints n_b which belongs to a subplace the second one is the number of subplaces n_s which belong to a place. We build a single linkage hierarchical clustering tree for the former situation which adds a complexity $\mathcal{O}(N)$. In the latter situation, reward is calculated for each derived subplace along the path through top-to-bottom, where the subplaces below the threshold is not evaluated. Therefore, the complexity with clustering approach is $\mathcal{O}(N^2 log(N^2))$ for worst case scenario.

3.6. Learning

Learning consists of three stages: i) Incorporating the detected place D; ii) Updating the one-SVM discriminant functions at the changed nodes of the hierarchy; iii) Adding the n_s subplaces into the place memory hierarchy;

The first two stages are done exactly as in the TSC model. Namely, first, the place memory hierarchy is incrementally modified to accommodate the new place. Next, one-SVM discriminant functions at the affected nodes are re-learned. Finally, differing from the original model, each learned place p is now associated with with a set of subplaces.

3.7. Experiments

Two sets of experiments have been conducted. The first set of experiments is done using benchmark datasets while the second set is with a tele-operated robot in our lab.

3.7.1. Dataset Experiments

The experiments are done using the indoors benchmark COLD dataset and outdoors New College (NC) dataset as previously used in Chapter 2. The COLD dataset consists of visual data taken from three different sites - Fr, Lj and Sa under cloudy conditions. NC site is from a perspective camera along a route of approximately two kilometers coverage. The data is taken in the early hours of the day and evening. Maps of the datasets with the associated subplaces are represented in 3.10,3.11,3.12 and 3.13 respectively. Odometry data given with dataset is not accurate, thus paths are edited accordingly to fit outline of corresponding site. In most of the cases, subplaces are precise in terms of separating canonical appearances shown in figures. These appearances belongs to the same place, but there is no common signatures between them. There are also exceptions where subplace approach dividing subplaces that consist of similar appearances such as Lj Place 20, Sa Place 15 and NC Place 4. Consider these cases one by one: As it stated before, the robot Lj site makes zigzag movement, these moves confuses robot and make it to cluster similar appearances. Place 15 in Sa site contains a wall on the left hand side which suddenly moves to out of frame. This dramatic change affects clustering as though there are 2 appearances. There are 2 people walking along the path in Place 4 of NC site, hence the robot finds two distinct set of descriptors. Other than these exceptions, subplaces successfully find the salient appearances in the image.

Site	# Places	Mean n_s	Variance of n_s
Fr	21	1.43	0.46
Lj	20	1.60	1.31
Sa	16	1.31	0.23
NC	10	1.10	0.10

Table 3.1. Subplace Statistics

First, subplaces that are determined in each site are evaluated. A sample place memory after the first visit Saarbrucken site with subplaces is demonstrated in Figure 3.9. The corresponding areas in the associated site map are as shown in Figure 3.12. Five of the sixteen learned places are learned as consisting of two subplaces while the

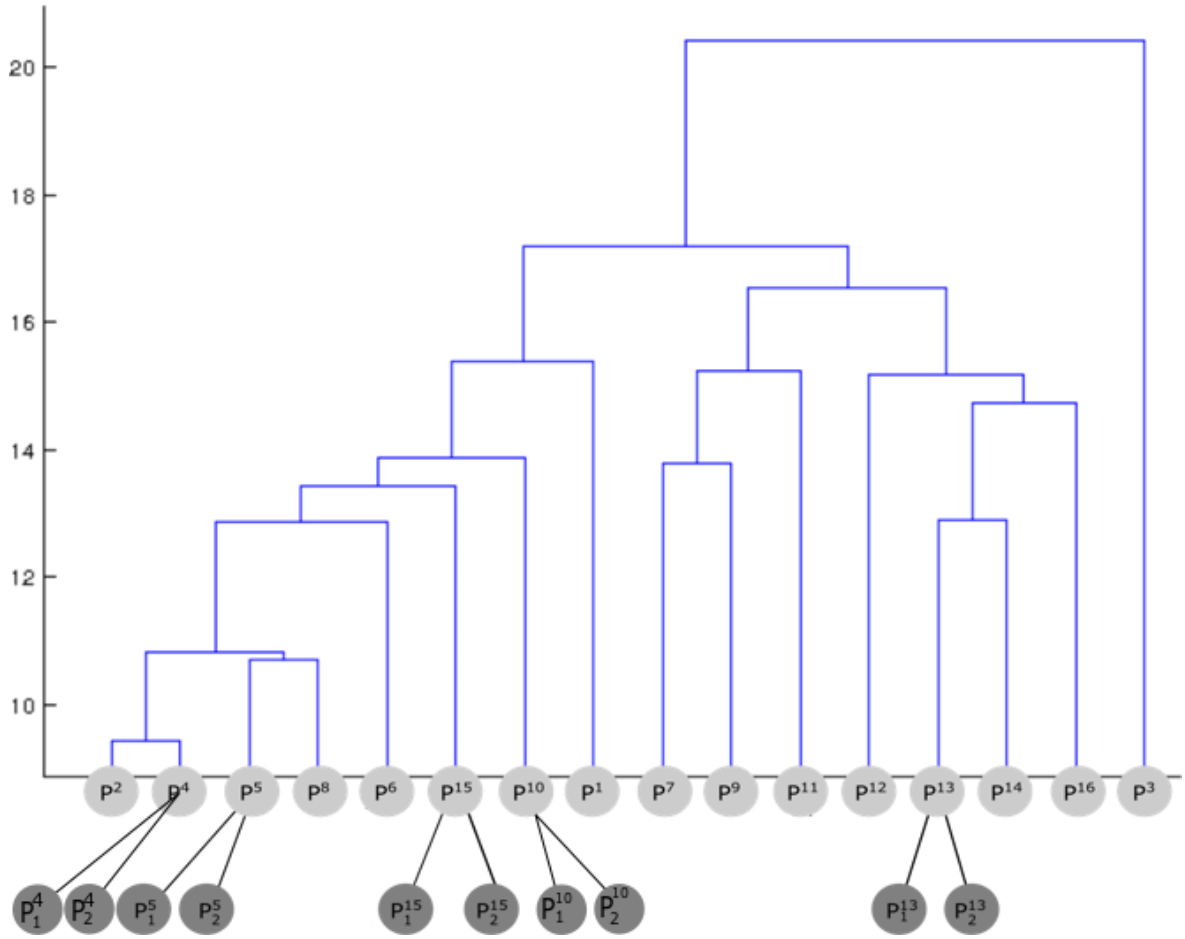


Figure 3.9. Place Memory with Subplaces of First Visit Saarbrucken Site

Figure 3.10. Freiburg Place Map

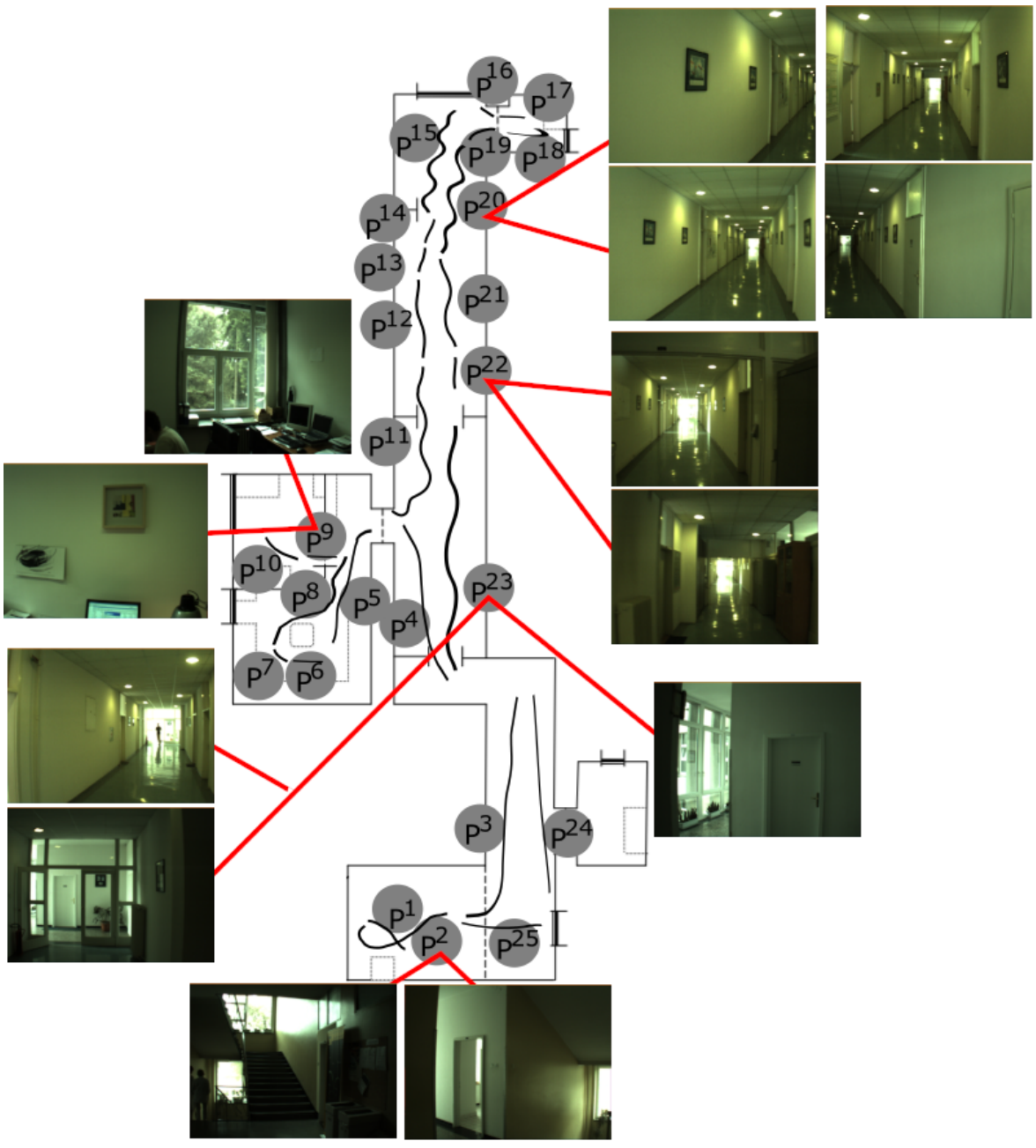


Figure 3.11. Ljubljana Place Map

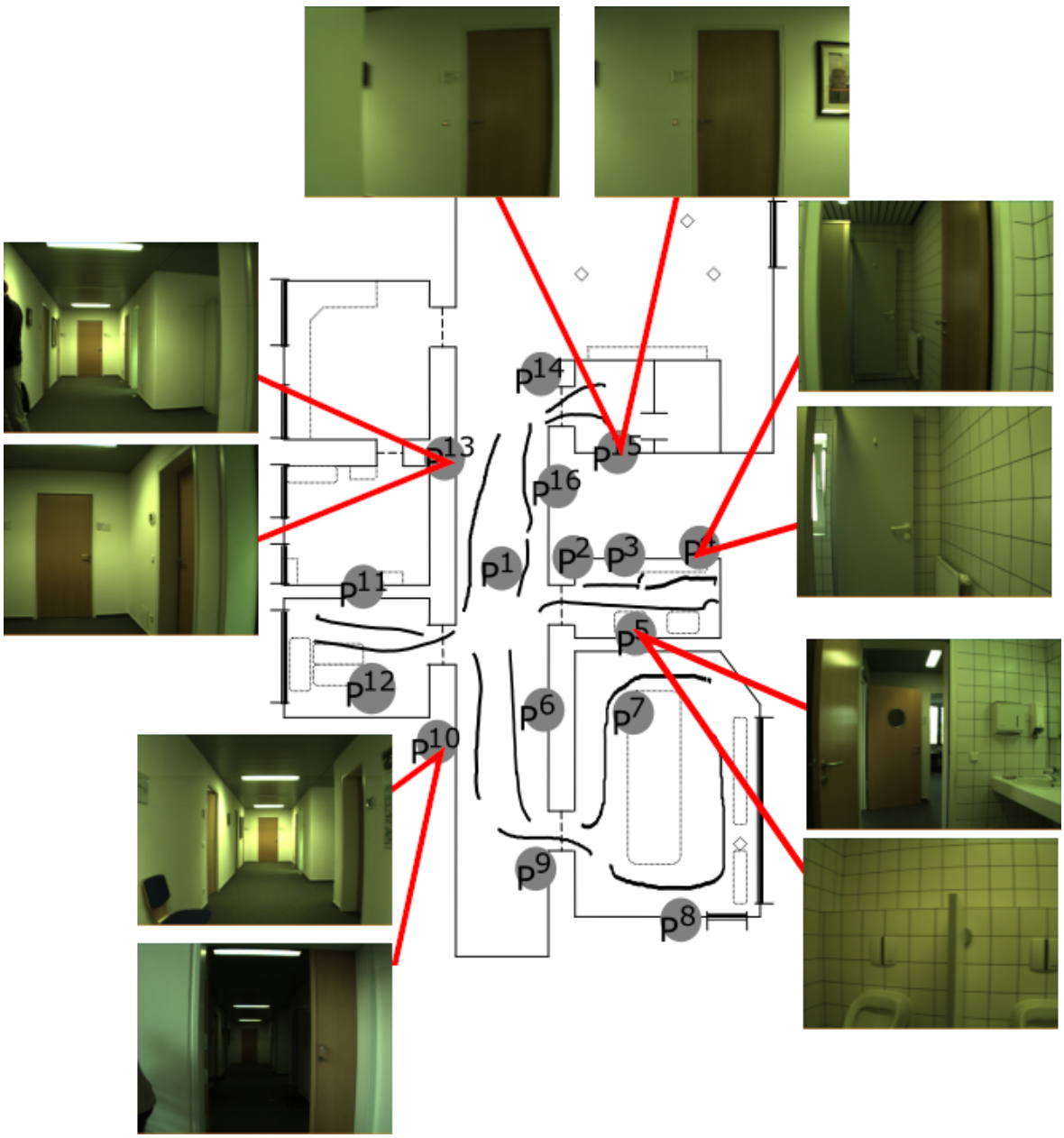


Figure 3.12. Saarbrücken Place Map

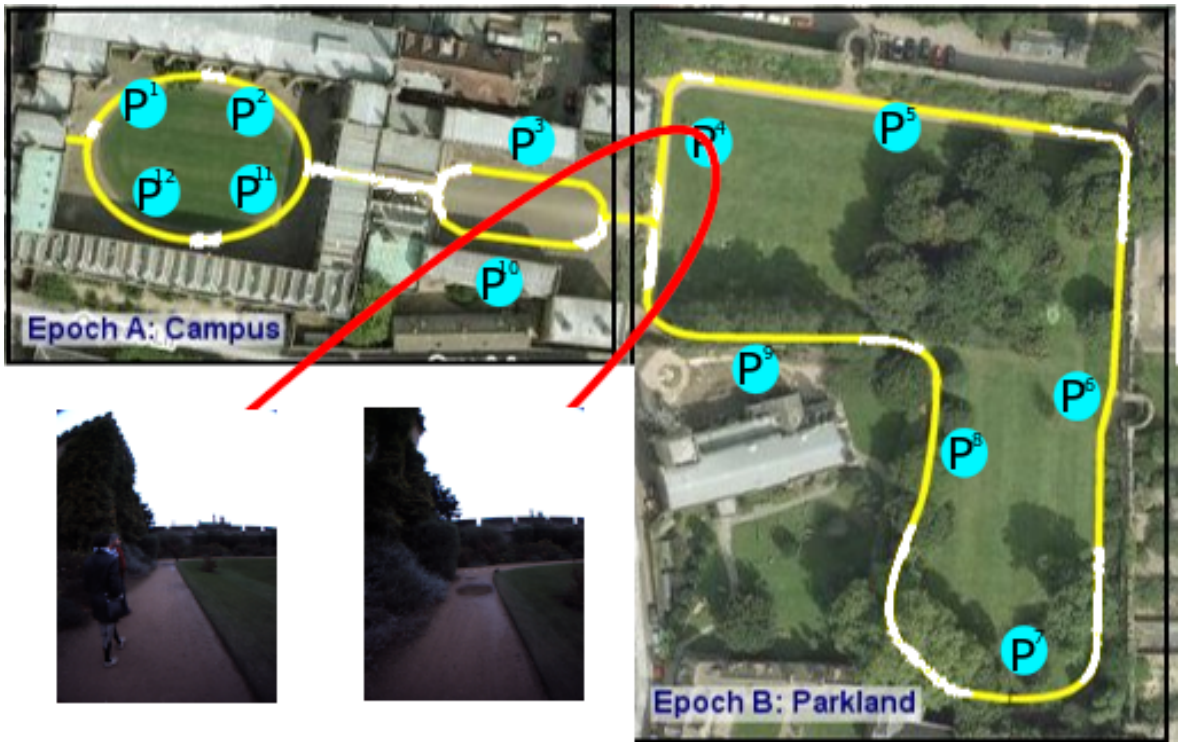


Figure 3.13. New College Place Map

remaining are learned as single places. For the remaining sites, places and their respective subplaces are as shown in Figure 3.10, Figure 3.11 and Figure 3.13 respectively. Statistical results including the mean and standard deviation of number of subplaces n_s are as given in Table 3.1. It is observed that the mean varies between 1.1 and 1.6. The places in the Fr and Sa sites have similar appearances as they are mostly rooms and corridors. Hence, subplaces depend on abrupt transitions in appearances in these places and their mean n_s values are close to each other. In the Lj site, robot moves along a zigzag path which causes many different appearances in a single run. As a result, it has the greatest mean and variance overall as expected. The places in the NC site has too few occlusions compared to other datasets, so the number of subplaces turns out to be one on average. It is obvious that there is no dramatic change in complexity based on these statistics.

Next, recognition performance is studied. As the concept of subplaces enables the partitioning of the appearances into canonical views in the place, both true misses and false recognition rates are reduced. One example is Place 6 and Place 22 from Sa site both include 2 subplaces. While they cannot be recognized with the original

TSC model, the robot can recognize them after representing them as consisting of two subplaces. Precision and recall rates are presented in Table 3.2,3.3,3.4,3.5 for Fr, Lj, Sa and NC sites respectively with varying τ_r values. Results are also provided for the original TSC model where the learning of each place is based on all the appearances collected in this place - in order to compare the two approaches. In the figures, the precision and recall rates are shown with and without subplace approach. Significant improvement on recognition performance for indoor datasets is observed. In the COLD dataset sites, there is an average of 15% improvement for both precision and recall rates. Interestingly, such improvement is not observed with the places in the NC site. This is expected since most of the places are characterized by one subplace. This may be attributed to the fact that these places have open space and thus the respective collected appearances do not have much variance.

Table 3.2. Precision and Recall Rates with and without Subplace Approach at Freiburg Site

Method / τ_r	Without Subplace Approach		With Subplace Approach	
	Precision	Recall	Precision	Recall
1	0.64	0.37	0.78	0.37
1.2	0.67	0.32	0.78	0.37
1.4	0.75	0.32	0.83	0.26
1.6	0.80	0.21	0.83	0.26
1.8	1.00	0.21	1.00	0.21

Table 3.3. Precision and Recall Rates with and without Subplace Approach at Ljubljana Site

Method / τ_r	Without Subplace Approach		With Subplace Approach	
	Precision	Recall	Precision	Recall
1	0.37	0.50	0.50	0.64
1.2	0.39	0.50	0.50	0.57
1.4	0.38	0.43	0.53	0.57
1.6	0.43	0.43	0.57	0.57
1.8	0.45	0.36	0.62	0.57

Finally, we consider combined illumination invariance and subplaces are studied.

Table 3.4. Precision and Recall Rates with and without Subplace Approach at Saarbrücken Site

Method / τ_r	Without Subplace Approach		With Subplace Approach	
	Precision	Recall	Precision	Recall
1	0.27	0.23	0.50	0.54
1.2	0.33	0.23	0.50	0.46
1.4	0.33	0.23	0.55	0.46
1.6	0.43	0.23	0.50	0.31
1.8	0.40	0.15	0.50	0.23

Table 3.5. Precision and Recall Rates with and without Subplace Approach at New College Site

Method / τ_r	Without Subplace Approach		With Subplace Approach	
	Precision	Recall	Precision	Recall
1	0.55	0.60	0.44	0.40
1.2	0.44	0.40	0.50	0.40
1.4	0.50	0.30	0.50	0.30
1.6	0.60	0.30	0.60	0.30
1.8	1.00	0.10	0.67	0.20

With the illumination invariance, extent of each single place changes. Therefore, place and subplaces do not corresponds the same index for for every illuminaion invariance approach. Precision and recall rates are presented in Figure 3.6,3.7,3.8,3.9 respectively. It is observed that adaptive gamma correction with subplace approach obtains the best results amongst illumination invariance and memory approaches in terms of both precision and recall for indoor datasets. Performance of place recognition boosts significantly with the help of subplace approach. Also, adaptive gamma approach make robot to detect start and end points of places truly. Subplace approach does not affect New College dataset dramatically, since this dataset generally consist of smooth regions and lack of transition points. Camera sensor modeling performs best in this case, where gammaDCT method comes off the second best with a close performance. Overall, place recognition boosts about 20-25% for indoor datasets.

Table 3.6. Performance of Illumination Invariance Methods Combined with Subplace Approach at Freiburg Site

(a) Precision Rates

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.78	0.62	0.42	0.41	0.35	0.73
1.2	0.78	0.67	0.45	0.44	0.37	0.79
1.4	0.83	0.67	0.50	0.38	0.31	0.82
1.6	0.83	0.71	0.56	0.36	0.33	0.88
1.8	1.00	0.83	0.63	0.33	0.25	1.00

(b) Recall Rates

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.37	0.42	0.26	0.37	0.37	0.58
1.2	0.37	0.42	0.26	0.37	0.37	0.58
1.4	0.26	0.32	0.26	0.26	0.26	0.47
1.6	0.26	0.26	0.26	0.21	0.26	0.37
1.8	0.21	0.26	0.26	0.16	0.16	0.37

3.7.2. Real Time Experiments

Real time experiments are conducted with Jaguar robot. As a conclusion of the detection results, quite successful recognition results are obtained, since start and endpoints of a place represents this places truly. The results are given in Table 3.10. Note that number of detected places does not change, since clustering only affects the recognition performance. As the table compared with the table 2.8, it is observed that two places in the ground truth recognized truly for the case without illumination invariance method in the first visit. In the second visit of the same condition, a false recognition is replaced with true one. There is no change observed in the first visit of the setup with illumination invariance, while 2 true recognized places hold instead of 2 false recognized places in the second visit with clustering approach. Thereupon, 100%

Table 3.7. Performance of Illumination Invariance Methods Combined with Subplace Approach at Ljubljana Site

(a) Precision Rates

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.50	0.50	0.45	0.52	0.46	0.65
1.2	0.50	0.52	0.41	0.50	0.45	0.67
1.4	0.53	0.53	0.44	0.47	0.44	0.67
1.6	0.57	0.53	0.47	0.47	0.50	0.69
1.8	0.62	0.56	0.46	0.42	0.50	0.73

(b) Recall Rates

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.64	0.79	0.64	0.86	0.43	0.93
1.2	0.57	0.79	0.50	0.71	0.36	0.86
1.4	0.57	0.71	0.50	0.57	0.29	0.71
1.6	0.57	0.64	0.50	0.50	0.29	0.64
1.8	0.57	0.64	0.43	0.36	0.21	0.57

precision is achieved in both visits and the second visit with illumination invariance method and 88% precision in the second visit without illumination invariance. Recall rates for the both of the first visits are 100%. Recall rate of the second visit without illumination invariance is increased to %54, while the case with illumination invariance method boosted up to 82%. Consequently, obtaining subplaces by clustering places affects both precision and recall rates significantly as it decreases false positives as well as increases true positives.

Table 3.8. Performance of Illumination Invariance Methods Combined with Subplace Approach at Saarbrücken Site

(a) Precision Rates

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.50	0.62	0.47	0.42	0.33	0.65
1.2	0.50	0.60	0.46	0.45	0.25	0.67
1.4	0.55	0.67	0.50	0.50	0.11	0.69
1.6	0.50	0.67	0.50	0.56	0.13	0.70
1.8	0.50	0.80	0.43	0.63	0.00	0.78

(b) Recall Rates

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.54	0.62	0.54	0.38	0.38	0.85
1.2	0.46	0.46	0.46	0.38	0.23	0.77
1.4	0.46	0.46	0.46	0.38	0.08	0.69
1.6	0.31	0.31	0.38	0.38	0.08	0.54
1.8	0.23	0.31	0.23	0.38	0.00	0.54

Table 3.9. Performance of Illumination Invariance Methods Combined with Subplace Approach at New College Site

(a) Preicison Rates

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.44	0.54	0.90	1.00	0.82	0.82
1.2	0.50	0.55	0.90	1.00	0.73	0.73
1.4	0.50	0.60	0.82	0.91	0.75	0.75
1.6	0.60	0.57	0.89	1.00	0.86	0.83
1.8	0.67	0.60	1.00	1.00	1.00	1.00

(b) Recall Rates

Method / τ_r	A	A_{HE}	A_{GD}	A_{CM}	A_{WF}	A_{AG}
1	0.40	0.70	0.90	1.00	0.90	0.90
1.2	0.40	0.60	0.90	1.00	0.80	0.80
1.4	0.30	0.60	0.90	1.00	0.60	0.60
1.6	0.30	0.40	0.80	1.00	0.60	0.50
1.8	0.20	0.30	0.60	1.00	0.40	0.30

Table 3.10. Number of Detected, Learned and Recognized Places for Real Time Jaguar Experiments with and without Subplace Approach

	A		A_{AG}	
	First Visit	Second Visit	First Visit	Second Visit
# Detected Places	12	13	10	11
# Learned Places	10	5	8	2
# Recognized Places	2T	7T + 1F	2T	9T

4. CONCLUSION

This thesis has focused on making topological spatial cognition more robust in robots. Appearances play a key role in this - as place detection, recognition and learning all are based on the appearances collected from the respective places.

The first contribution of this study is a novel approach to enhance the robustness of place detection and recognition against illumination changes. As appearances change depending on illumination, performances of both are negatively affected if illumination conditions differ from that of learning. In the literature, a variety of different methods has been proposed in order to alleviate this problem. In this thesis, four such methods are considered: histogram equalization, gamma normalized discrete cosine transform, camera modeling and Wiener filtering. All these methods propose to transform the incoming images into illumination invariant images and differ in how this is done. Finally, a new method referred to as *adaptive gamma correction* is proposed. A comparative study is conducted as to evaluate performance in both place detection and place recognition. It is observed that illumination invariance improves the performance of the robot for both detection and recognition.

The second contribution concerns place representation. In the current model, the appearances collected in a place are considered altogether in both place recognition and learning. As such, recognition performance is affected if there is a partial overlap of input appearances as compared to those of learning. For this, the concept of 'subplaces', is proposed. First, place detection is modified so that the appearances collected in a place are partitioned into distinct clusters - each of which denoting a canonical view in the place. Thus, each place is represented by a set of subplaces. Both recognition and learning are then revised as to accommodate such a representation.

Both of the contributions boosts the performance significantly. Nevertheless, every improvement brings a cost associated with it. The approaches used in this study brings nothing but the computational complexity. Although, computational complexity

increases with the improvements, algorithms can be still run in real-time thanks to the computational power we reached today.

Finally, miscalculations in the topological spatial cognition software have been corrected. Coding errors pertaining to spatial filter calculations and 8-bit overflows are corrected. Coherency between hue and intensity descriptors is distinguished for more accurate results. These corrections are held at Appendix B.

Two extensions are planned as a future work. First one is to consider other challenging factors that lead to changes in appearances such as seasonal changes and dynamic entities such as moving people. The second is to be able to recognize similar appearances even if they are shot from different viewpoints.

REFERENCES

1. Lloyd, R., *Spatial cognition: Geographic environments*, Vol. 39, Springer Science & Business Media, 1997.

2. Denis, M. and J. M. Loomis, "Perspectives on human spatial cognition: memory, navigation, and environmental learning", *Psychological Research*, Vol. 71, pp. 235–239, Springer, 2007.

3. Dolins, F. L. and R. W. Mitchell, *Spatial cognition, spatial perception: mapping the self and space*, Cambridge University Press, 2010.

4. Tversky, B., "Functional significance of visuospatial representations", *Handbook of higher-level visuospatial thinking*, pp. 1–34, 2005.

5. Miller, S., *Space and Sense*, Psychology Press, 2008.

6. Lowry, S., N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke and M. J. Milford, "Visual Place Recognition: A Survey", *IEEE Transactions on Robotics*, Vol. 32, No. 1, pp. 1–19, 2016.

7. Kostavelis, I. and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey", *Robotics and Autonomous Systems*, Vol. 66, pp. 86 – 103, 2015.

8. Karaoguz, H. and H. I. Bozma, "An Integrated Model of Autonomous Topological Spatial Cognition", *Aut. Rob.*, Vol. 40, No. 8, pp. 1379–1402, 2016.

9. Erkent, O. and H. I. Bozma, "Bubble space and place representation in topological maps", *The Int. J. of Rob. Res.*, Vol. 32, No. 6, pp. 672–689, 2013.

10. Chella, A., I. Macaluso and L. Riano, "Automatic place detection and localization in autonomous Rob.", *Int. Conf. on Intel. Rob. Sys.*, pp. 741–746, 2007.

11. Ranganathan, A., "PLISS: labeling places using online changepoint detection", *Autonomous Robots*, Vol. 32, No. 4, pp. 351–368, 2012.

12. Vasudevan, S. and R. Siegwart, "Bayesian Space Conceptualization and Place Classification for Semantic Maps in Mobile Robotics", *Rob. Auto. Sys.*, Vol. 56, No. 6, pp. 522–537, Jun. 2008.

13. Beeson, P., N. K. Jong and B. Kuipers, "Towards autonomous topological place detection using the extended voronoi graph", *IEEE Int. Conf. on Rob. and Aut.*, pp. 4373–4379, 2005.

14. Sibson, R., "SLINK: an optimally efficient algorithm for the single-link cluster method", *The computer journal*, Vol. 16, No. 1, pp. 30–34, 1973.

15. Erkent, Ö. and H. I. Bozma, "Bubble space and place representation in topological maps", *The International Journal of Robotics Research*, Vol. 32, No. 6, pp. 672–689, 2013.

16. Pronobis, A. and B. Caputo, "COLD: The CoSy localization database", *The International Journal of Robotics Research*, Vol. 28, No. 5, pp. 588–594, 2009.

17. Smith, M., I. Baldwin, W. Churchill, R. Paul and P. Newman, "The New College vision and laser data set", *The Int. J. Robot. Res.*, Vol. 28, No. 5, pp. 595–599, 2009.

18. Valgren, C. and A. J. Lilienthal, "SIFT, SURF & Seasons: Appearance-based Long-term Localization in Outdoor Environments", *Robot. Auton. Syst.*, Vol. 58, No. 2, pp. 149–156, Feb. 2010.

19. Barrow, H. G. and J. M. Tenenbaum, *Recovering Intrinsic Scene Characteristics from Images*, Academic Press, 1978.

20. Finlayson, G. D., M. S. Drew and C. Lu, "Intrinsic Images by Entropy Minimiza-

tion", *European Conference on Computer Vision*, Vol. 3, pp. 582–595, 2004.

21. Maddern, W., A. Stewart, C. McManus, B. Upcroft, W. Churchill and P. Newman, "Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles", *Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China*, Vol. 2, p. 3, 2014.

22. Shakeri, M. and H. Zhang, "Illumination invariant representation of natural images for visual place recognition", *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pp. 466–472, IEEE, 2016.

23. Chen, W., M. J. Er and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, Vol. 36, No. 2, pp. 458–466, 2006.

24. Bormann, R., T. Zwölfer, J. Fischer, J. Hampp and M. Hägele, "Person recognition for service robotics applications", *Humanoid Robots (Humanoids), 2013 13th IEEE-RAS International Conference on*, pp. 260–267, IEEE, 2013.

25. Tan, X. and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions", *IEEE transactions on image processing*, Vol. 19, No. 6, pp. 1635–1650, 2010.

26. Naik, S. K. and C. Murthy, "Hue-preserving color image enhancement without gamut problem", *IEEE Transactions on Image Processing*, Vol. 12, No. 12, pp. 1591–1598, 2003.

27. Song, K. S., H. Kang and M. G. Kang, "Hue-preserving and saturation-improved color histogram equalization algorithm", *JOSA A*, Vol. 33, No. 6, pp. 1076–1088, 2016.

28. Ratnasingam, S. and T. M. McGinnity, "Chromaticity Space for Illuminant Invariant Recognition", *IEEE Transactions on Image Processing*, Vol. 21, No. 8, pp. 3612–3623, 2012.

29. Marchant, J. A. and C. M. Onyango, "Shadow-invariant classification for scenes illuminated by daylight", *Journal of the Optical Society of America A*, Vol. 17, pp. 1952–1961, 2000.

30. Chen, L.-H., Y.-H. Yang, C.-S. Chen and M.-Y. Cheng, "Illumination invariant feature extraction based on natural images statistics—Taking face images as an example", *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 681–688, IEEE, 2011.

31. Torralba, A. and A. Oliva, "Statistics of natural image categories", *Network: computation in neural systems*, Vol. 14, No. 3, pp. 391–412, 2003.

32. Hoke, B. and H. I. Bozma, "Uyarlamalı Gama Düzeltmesiyle Görünüşlerde ışık Değişmezliği (in Turkish)", *Türkiye Robotbilim Konferansı*, 2018.

33. Vonikakis, V., I. Andreadis and A. Gasteratos, "Fast centre–surround contrast modification", *IET Image processing*, Vol. 2, No. 1, pp. 19–34, 2008.

34. Poynton, C., *Digital video and HD: Algorithms and Interfaces*, Elsevier, 2012.

35. Ebner, F. and M. D. Fairchild, "Development and testing of a color space (IPT) with improved hue uniformity", *Color and Imaging Conference*, pp. 8–13, Society for Imaging Science and Technology, 1998.

36. Ross, P., A. English, D. Ball and P. Corke, "A method to quantify a descriptor's illumination variance", *Australian Conf. on Rob. and Aut.*, 2014.

37. Kostavelis, I. and A. Gasteratos, "Learning spatially semantic representations for cognitive robot navigation", *Robotics and Autonomous Systems*, Vol. 61, No. 12,

pp. 1460–1475, 2013.

38. Kostavelis, I., K. Charalampous, A. Gasteratos and J. K. Tsotsos, "Robot navigation via spatial and temporal coherent semantic maps", *Engineering Applications of Artificial Intelligence*, Vol. 48, pp. 173–187, 2016.

39. Charalampous, K., I. Kostavelis and A. Gasteratos, "Recent trends in social aware robot navigation: A survey", *Robotics and Autonomous Systems*, Vol. 93, pp. 85–104, 2017.

40. Karaoguz, H. and H. I. Bozma, "Reliable topological place detection in bubble space", *IEEE Int. Conf. on Rob. Aut.*, pp. 697–702, 2014.

41. Zivkovic, Z., O. Booij and B. Kröse, "From images to rooms", *Rob. and Auto. Sys.*, Vol. 55, No. 5, pp. 411–418, 2007.

42. Remolina, E. and B. Kuipers, "Towards a general theory of topological maps", *Artificial Intelligence*, Vol. 152, No. 1, pp. 47–104, 2004.

43. Tapus, A. and R. Siegwart, "Incremental robot mapping with fingerprints of places.", *IROS*, Vol. 1, pp. 2429–2434, 2005.

44. Cummins, M. and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0", *The International Journal of Robotics Research*, Vol. 30, No. 9, pp. 1100–1123, 2011.

45. Thrun, S., "Exploration in active learning", *Handbook of Brain Science and Neural Networks*, pp. 381–384, 1995.

46. Deng, X., T. Kameda and C. Papadimitriou, "How to learn an unknown environment. I: the rectilinear case", *Journal of the ACM (JACM)*, Vol. 45, No. 2, pp. 215–245, 1998.

47. Moorehead, S. J., R. Simmons and W. L. Whittaker, "Autonomous exploration

using multiple sources of information", *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, Vol. 3, pp. 3098–3103, IEEE, 2001.

48. Tovey, C. and S. Koenig, "Improved analysis of greedy mapping", *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, Vol. 4, pp. 3251–3257, IEEE, 2003.

49. Yamauchi, B., "A frontier-based approach for autonomous exploration", *Computational Intelligence in Robotics and Automation, 1997. CIRA'97., Proceedings., 1997 IEEE International Symposium on*, pp. 146–151, IEEE, 1997.

50. Oriolo, G., G. Ulivi and M. Vendittelli, "Real-time map building and navigation for autonomous robots in unknown environments", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, Vol. 28, No. 3, pp. 316–333, 1998.

51. González-Banos, H. H. and J.-C. Latombe, "Navigation strategies for exploring indoor environments", *The International Journal of Robotics Research*, Vol. 21, No. 10-11, pp. 829–848, 2002.

52. Bourgault, F., A. A. Makarenko, S. B. Williams, B. Grocholsky and H. F. Durrant-Whyte, "Information based adaptive robotic exploration", *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*, Vol. 1, pp. 540–545, IEEE, 2002.

53. Akdeniz, B. C. and H. I. Bozma, "Exploration and topological map building in unknown environments", *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pp. 1079–1084, IEEE, 2015.

54. Gong, Y. and X. Liu, "Video summarization and retrieval using singular value decomposition", *Multimedia Systems*, Vol. 9, No. 2, pp. 157–168, 2003.

55. Girdhar, Y. and G. Dudek, "Online navigation summaries", *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 5035–5040, IEEE, 2010.

56. Ranganathan, A. and F. Dellaert, "Bayesian surprise and landmark detection", *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pp. 2017–2023, IEEE, 2009.

57. Konolige, K., J. Bowman, J. Chen, P. Mihelich, M. Calonder, V. Lepetit and P. Fua, "View-based maps", *The International Journal of Robotics Research*, Vol. 29, No. 8, pp. 941–957, 2010.

58. Murphy, L. and G. Sibley, "Incremental Unsupervised Topological Place Discovery", *IEEE Int. Conf. Robot. Aut.*, pp. 1312 – 1318, June 2014.

59. Paul, R., D. Rus and P. Newman, "How was your day? online visual workspace summaries using incremental clustering in topic space", *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pp. 4058–4065, IEEE, 2012.

60. Paul, R., D. Feldman, D. Rus and P. Newman, "Visual precis generation using coresets", *IEEE Int. Conf. on Rob. and Aut.*, pp. 1304–1311, 2014.

61. Schölkopf, B., J. C. Platt, J. Shawe-Taylor, A. J. Smola and R. C. Williamson, "Estimating the support of a high-dimensional distribution", *Neural computation*, Vol. 13, No. 7, pp. 1443–1471, 2001.

62. Quigley, M., K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler and A. Y. Ng, "ROS: an open-source Robot Operating System", *ICRA workshop on open source software*, p. 5, Kobe, Japan, 2009.

63. Bradski, G., "The OpenCV Library", *Dr. Dobb's Journal of Software Tools*, 2000.

APPENDIX A: BUBBLE SURFACE AND DESCRIPTORS

This section presents a brief summary of bubble space representation for completeness. The bubble space $\mathcal{B} = \mathcal{X} \times \mathcal{F}$ is an abstract representation of the robot's base $\mathcal{X}$ along with its viewing directions (pan and tilt) $\mathcal{F} \subset S^2$ with $b \in \mathcal{B}$ defined as $b = [x\, f]^T$ where $x \in \mathcal{X}$ and $f \in \mathcal{F}$. As the robot looks around, for each viewing direction $f \in \mathcal{F}$, it computes a set of visual feature values $q_i(b,t) \geq 0$, $i = 1, \ldots, N$. In the experiments, the robot computes $N = 6$ bubble surfaces corresponding to seven visual features (hue, Cartesian and non-Cartesian filters). For each feature, a bubble surface $B_i(x,t) : Im(h(x)) \times R^{\geq 0} \rightarrow R^{\geq 0}$ is generated. This is an hypothetical spherical surface surrounding the robot that encodes the observed values of i^{th} sensory feature.

$$B_i(x,t) = \left\{ \begin{bmatrix} f \\ \rho_i(b,t) \end{bmatrix} \mid \forall f \in \mathcal{F} \text{ and } b = [x\, f]^T \right\} \tag{A.1}$$

Here, $Im(h(x))$ denotes the set of viewing directions from a given base x, $h : \mathcal{X} \rightarrow \mathcal{B}$ is defined as a continuous map such that $\forall x \in \mathcal{X}$, $\pi(h(x)) = x$, $\pi : \mathcal{B} \rightarrow \mathcal{X}$ is defined as the projection of b onto $\mathcal{X}$ as $\pi(b) = x$ and observed values of v_i^{th} sensory feature are encoded by a Riemannian metric $\rho_i : \mathcal{B} \times R^{\geq 0} \rightarrow R^{\geq 0}$. It is initialized to be a S^2 sphere with radius $\rho_0 \in R^{\geq 0}$ – namely $\rho_i(b,0) = \rho_0$. For each viewing direction f, each bubble surface $B_i(x,t)$ is deformed at f by an amount that depends on $q_i(b,t)$ as:

$$\rho_i(b,t^+) = q_i(b,t)$$

where the superscript t^+ denotes time just after t.

Bubble descriptors are holistic (vector) representations of bubble surfaces. They are constructed using the double Fourier series representation of bubble surfaces as:

$$\rho_i(b,t) = \sum_{h_1=0}^{H_1} \sum_{h_2=0}^{H_2} \lambda_{h_1 h_2} z_{xi,h_1 h_2}^T(t) e_{h_1 h_2}(f)$$

If $f \in \mathcal{F}$ is defined as $f = [f_1\ f_2]^T$, for each (h_1, h_2), the vector $e_{h_1h_2}(f) \in R^4$ consists of an orthonormal set of trigonometric basis functions as:

$$e_{h_1h_2}(f) = \begin{bmatrix} cos(h_1f_1)cos(h_2f_2) \\ sin(h_1f_1)cos(h_2f_2) \\ cos(h_1f_1)sin(h_2f_2) \\ sin(h_1f_1)sin(h_2f_2) \end{bmatrix}$$

The corresponding vector $z_{xi,h_1h_2}(t) \in R^4$ is defined as:

$$z_{xi,h_1h_2}(t) = \frac{1}{\pi^2} \begin{bmatrix} \int_0^{2\pi}\int_0^{\pi} \rho_i(b,t)cos(h_1f_1)cos(h_2f_2)df_1df_2 \\ \int_0^{2\pi}\int_0^{\pi} \rho_i(b,t)sin(h_1f_1)cos(h_2f_2))df_1df_2 \\ \int_0^{2\pi}\int_0^{\pi} \rho_i(b,t)cos(h_1f_1)sin(h_2f_2))df_1df_2 \\ \int_0^{2\pi}\int_0^{\pi} \rho_i(b,t)sin(h_1f_1)sin(h_2f_2)df_1df_2 \end{bmatrix}$$

The parameters $\lambda_{h_1h_2}$ are defined as:

$$\lambda_{h_1h_2} = \begin{cases} \frac{1}{4} & \text{if } h_1 = 0, h_2 = 0 \\ \frac{1}{2} & \text{if } h_1 > 0, h_2 = 0 \text{ or } h_1 = 0, h_2 > 0 \\ 1 & \text{if } h_1 > 0, h_2 > 0 \end{cases}$$

A bubble descriptor $I(x,t) \in R^d$ is a $d-$dimensional vector with $d = N(H_1+1)(H_2+1)$ defined as:

$$I(x,t) = [I_{1,00}(x,t), \ldots, I_{N_v,H_1H_2}(x,t)]^T$$

where

$$I_{i,h_1h_2}(x,t) = z^T_{xi,h_1h_2}(t)z_{xi,h_1h_2}(t)$$

Bubble descriptors have been shown to be rotationally invariant with respect to heading changes while being computable in an incremental manner- as new observations

are made. Furthermore, they have flexibility in integrating visual features since their dimensionality is independent of the number of observations and no data association is required for finding correspondences among the different observations. In the experiments, the bubble descriptors are constructed with $H_1 = H_2 = 9$. Thus, the lengths of bubble descriptors are $d = 600$.

APPENDIX B: SOFTWARE

B.1. Required Software and Libraries

Whole system included in this thesis is implemented in C++ using ROS [62] and Qt frameworks. OpenCV [63] computer vision library is also used in order to implement image processing and computer vision algorithms. Besides that, MATLAB scripts are used for prototype implementations and data display due to its ease to use. ROS enables us to easily migrate the code to the robot. Our system consists of two ROS nodes:

- *place_ detection_ isl* – Available at: https://github.com/islboun/place_detection_isl
- *create_ bdst_ isl* – Available at: https://github.com/islboun/create_bdst_isl

whose tasks are detection and recognition/place memory, respectively. The code was originally written with following versions which are outdated and it gets more difficult to seek support from documentations and community. Thus, all used software packages are updated to newest versions.

Software	**Old Versions**	**New Versions**
Ubuntu	12.04	16.04
ROS	Fuerte	Kinetic
Qt	4	5.10.1
OpenCV	2.4.9	3.3.1
Build System	rosbuild	catkin

B.2. Issues

There were several issues in the software that is revealed through this study. These issues generally affect the results by leading miscalculations or causes software to run inefficiently. The software is revised in order to solve these issues and they are provided with their solutions as follows.

B.2.1. Overflow Issue of RGB Images after Applying Spatial Filters

RGB images are convolved with six spatial filters in order to simulate the cognition of macaque monkeys as mentioned in 2. Most of the RGB images are represented as 8 bits digitally and so in the interval [0,255]. However, if a 8 - bit type image is convolved with one of these filters, most of the pixels tends to have higher values than 255 or lower than 0. This situation causes overvalued pixels values to become 255 and undervalued ones become 0. Finally, bubble descriptors were miscalculated. This bug does not overcome with completely wrong but deficient results. The images are converted to OpenCV CV_32FC3 type which represents images as 32 - bit floating point and supports majority of OpenCV methods.

B.2.2. Shift of Spatial Filters

Bug related to spatial filters is also observed such as previous example. The calculated values were shifted 2 pixels to left than it should be. Center of the concentric filter have to be in the center of the filter image for instance. The problem was caused by a phase shift while creating filters. Bug is solved by changing initial and final frequencies and normalizing the filter responses. Corrected filters can be generated with a MATLAB script which is available at GitHub repository: *https://github.com/islboun/filterGenerator*

B.2.3. Recalculation of Pan and Tilt Angles for each Bubble Surface

Bubble descriptor values are calculated by using the pan and tilt angles as shown in A. Pan and tilt angles are calculated using image width and height, respectively. Therefore, pan and tilt angles are same for the images with same width and height. Once it was computed for each image individually, and it needs much more computations compared to single computation even though single computation is enough. This modification speeds up the bubble surface computation dramatically.

B.3. Running the Software

B.3.1. Prerequisites

(i) Ubuntu 12.04 or higher

(ii) ROS version with catkin build system

(iii) OpenCV 2.4 or higher

(iv) Qt 4 or higher

(v) Specified Catkin Workspace Path(or create it with the command: *mkdir -p /<Catkin Worksapce Path>/src*)

(vi) emptydb folder which includes empty database files: detected_ places.db and knowledge.db under home directory

(vii) visual_filters folder including related filters within the home directory

B.3.2. Configuring .bashrc file

(i) Open terminal and enter the following command: *sudo gedit ∼/.bashrc* – it will open the terminal startup configurations

(ii) Add following lines to the end of .bashrc file:

- *source /opt/ros/<rosversion>/setup.bash*
- *source <Catkin Worksapce Path>/devel/setup.bash*

B.3.3. Configuring and Compiling

(i) Open terminal and type:
cd <Catkin Workspace Path>/src
git clone -b kinetic https://github.com/islboun/place_detection_isl.git

(ii) Open different terminal and type:
git clone -b Verison2 https://github.com/islboun/create_bdst_isl.git

(iii) Open different terminal and type:
catkin_make -DCMAKE_BUILD_TYPE=Release

B.3.4. Configuring launch file

Launch file is basically a xml file where user can enter parameters without compiling the code. The parameters below must be set in the launch file to run software.

- tau_w : Threshold for temporal window size
- tau_n : Upper bound of temporal window extension
- tau_p : Threshold for the size of a place
- tau_kappa : Coherency threshold for intensity channel
- tau_kappaHue : Coherence threshold for hue channel
- tau_val_mean : Threshold for the mean value of invariants (Checks informativeness)
- tau_val_var : Threshold for the variance value of invariants (Checks informativeness)
- image_width : Width of the images that will be processed
- image_height : Height of the images that will be processed
- focal_length_pixels : Focal length in pixels
- sat_lower : Lower threshold for saturation
- val_lower : Lower threshold for intensity
- val_upper : Upper threshold for intensity
- debug_mode : Set 'false' to run on robot, 'true' to run on dataset
- file_path : Path of the dataset if debug mode is 'true'
- previous_memory_path : Path of the previous run in order to make

B.3.5. Running Software

(i) Open a terminal window and type:
roslaunch place_detection_isl <filename>.launch – This will run the place_detection_isl node with given parameters from <filename>.launch file

(ii) Open another terminal window and type:
rosrun create_bdst_isl create_bdst_isl_node – This will run create_bdst_isl node

(iii) Open another terminal window and type:
rostopic pub /place_detection_isl/nodecontrol std_msgs/Int16 "data: 1" – This will publish start command to place_detection_isl node

If all of the steps are done successfully, then the terminal which called *roslaunch* command displays the informativeness of each frame and dissimilarity between the frames. Program warns the user if it could not find one of the necessary files such as ∼/detected_places, ∼/knowledge.db, visual_filters/filtre#.txt or the dataset folder specified in launch file. The terminal which called *rosrun* command displays the learned place counts and recognition scores for new coming places.