Place Recognition for Indoor Blind Navigation

Senior Honors Thesis 2010-2011
Computer Science

Samreen Anjum
School of Computer Science
Carnegie Mellon University in Qatar

Advisors:
Dr. Brett Browning Dr. M. Bernardine Dias
Carnegie Mellon University/Carnegie Mellon Qatar

Abstract

Navigating inside buildings is a challenging task for the blind which often makes them dependent on others for guidance. This thesis explores place recognition and navigation algorithms accessible via hand held devices to guide the visually impaired inside unfamiliar buildings. In computer vision, place recognition addresses the problem of identifying places based on appearance. The majority of this work focuses on implementing, evaluating, and enhancing place recognition algorithms for this application domain. The outcome of this work is a camera-based indoor blind-navigation model that can autonomously recognize previously mapped indoor locations. In the mapping or calibration phase, images of an indoor setting are collected, labeled, and mapped to specific locations in the indoor environment. The effectiveness of the system is evaluated by measuring the recall rate of the place recognition algorithm.

Acknowledgements

I would like to thank my advisors for their time, and constant help and guidance throughout the year to accomplish this thesis work. I would also like to thank my family and friends for their support while I worked on this thesis. I would like to appreciate Hatem Alismail, Dr. Peter Hansen, Ermine Teves and Samira Islam for their help in this project.

Contents

1. Introduction

Indoor navigation in unfamiliar places is a challenging task both for the sighted and the visually impaired. Sighted individuals often depend on visual cues such as the direction boards and hand gestures to find their way inside a building. The visually impaired, however, are generally dependent on additional aids such as seeing-eye dogs and other sighted individuals for directions. As a result, they suffer from loss of independence as well as privacy. There has been a growing interest in recent years to solve this problem using technology and various approaches have been taken. The two main components of indoor navigation can be divided into user localization and path planning. User localization, as the names suggests, refers to the process that identifies the location of the user in the building, and path planning is process where the path taken by the user to reach his/her destination is planned. A successful navigation system must first ensure that the user has been localized effectively. Hence, user localization is a vital component of indoor blind navigation and is the primary focus of this thesis work. The navigation component, in this thesis, is a simple implementation of a relevant path planning algorithm.

Recently, there has been growing interest to solve the problem of user localization using place recognition techniques [1-4]. As mentioned earlier, the fundamental goal of place recognition is to recognize places based on the cues provided. There are several different ways to perform place recognition using Bar codes, RFID Information Grids, sensors and WiFi[5],[6],[1]. However, in most of these approaches there exist challenges such as high cost of deployment, scalability, and lack of user orientation information. In this thesis, we aim to use computer vision techniques for place recognition to approach the problem of user localization. Computer vision has been rapidly developing in the fields of scene understanding and recognition. Derived

from these techniques, place recognition in the field computer vision aims to recognize the place using its visual appearance. This has applications to topological map building, navigation and loop closure in a mapping system [7]and is often deeply related to the task of image retrieval.

Although there has been extensive ongoing research in place recognition inside buildings, it is still considered to be a challenging research area. This thesis seeks to build upon and enhance state-of-the-art place recognition techniques that can be used to create a portable indoor blind navigation system that will allow the visually impaired to obtain directions to their destinations. The specific research topic for this thesis is:

> **To develop and evaluate relevant place recognition algorithms that can be combined with an intelligent path planning algorithm on a portable device to enable independent navigation for the visually impaired in GPS-denied indoor environments.**

1.1 Overview of Approach and Contributions

Our approach focuses on separating the problem into two parts: mapping and localization. The mapping, or training phase, takes place offline and consists of building a map of the environment by a sighted person taking images that are stored for use in the localization phase. In operation, as shown in fig 1, the user takes images, for instance, using a smart phone, and these images are used to register, or localize the visually impaired user in the map of the building. After localization, the system will find the shortest path to the named destination (e.g. the "cafeteria"), and guide the user along the path.

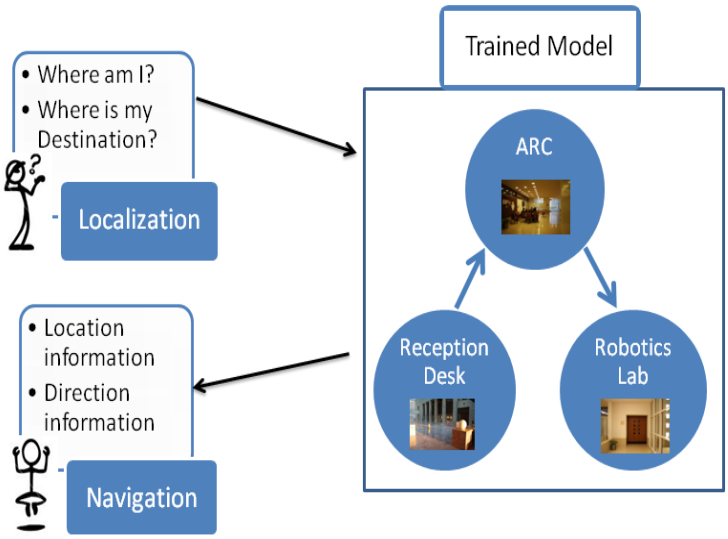


Figure 1 System Overview

The specific contributions of this thesis work include:

- A camera-based place recognition model to enable user localization in indoor environments
- Analysis of five different place recognition techniques/enhancements
- Introduction of two enhancements to the place recognition algorithm
- Evaluation of the tests in the Carnegie Mellon University in Qatar (CMUQ) building

The following sections are organized as followed: Chapter 2 discusses some of the related work in both blind navigation and place recognition. In chapter 3, we explain the approach taken in this thesis, and chapter 4 and 5 describe the two components of the approach, mapping and localization, in detail respectively. Chapter 6 illustrates the experiments conducted in the Carnegie Mellon University in Qatar building and the results obtained. Finally, the conclusions and future work is described in chapter 7.

2 Related Work

Significant research efforts have resulted in the development of several place recognition algorithms. Variations of these algorithms work on images collected using different cameras such as monocular cameras, omni-directional cameras, laser cameras, etc. In contrast to collecting information from a single camera, in some cases multiple cues are collected from different cameras and integrated together to gather more information. Several approaches are then taken to analyze the images to extract the significant information that facilitates identification of the place. Examples include local and global feature extraction (such as SIFT [8], SURF, PIRF [9]), color histograms, and PCA. A lot of research has been done to compare the performance of these descriptors [10].

In order to classify the images, different algorithms such as nearest neighbor search [11], SVM [12], barcode reading [1], RFID tags [5],[13], Most of these approaches have been used with robot localization and topological mapping. Amongst these approaches, Bag of Words (BoW) has been very popular and used in many approaches. In this approach, a visual vocabulary is generated using several different ways. One of the most common ways in image classification is to build similarity matrices where pairs of images are compared and the similarity scores are recorded [14]. Methods based on probabilistic approaches have also been intensively explored. One such approach is known FABMAP [7]. The main goal of this approach is to detect loop closures by addressing the problem of perceptual aliasing - which refers to the problem of identifying similar locations distinctly.

Apart from employing these place recognition techniques, several other approaches have been

taken to develop indoor blind navigation aids. One such effort reads bar color codes using mobile phone cameras to identify the current location [1]. The phone scans the bar code, identifies the location and reads it aloud to the user. However, there exist certain issues with this approach. Firstly, it requires the user to aim the camera correctly in order to locate the bar code. Furthermore, it works only for short distances, that is, the user is expected to be close to the code in order to identify and scan it. Sonnenblick implemented an indoor navigation aid for the blind using built-in infrared beacons for orienting the users [6]. Furthermore, there are several approaches that rely on WLAN and Wi-Fi to send and receive information, and position of the blind individual [3],[4]. Drishti [3] is one such aid that utilizes a wireless network, and switches between indoors and outdoors.

In addition to these approaches, RFID tags and information grids have also been popularly used to facilitate blind navigation. [5],[13] These tags are integrated in the floor of the building. The user generally carries a portable device that contacts these grids and uses the information in the closest grid to find the position of the user in the building. Although it is argued to be cheap to install, it is challenging to integrate these grids on concrete floors. In addition, since the person is required to be close to these grids, it is vital to have these grids covering every small part of the building, which results in scalability issues in huge buildings.

In this thesis, we explore computer vision based techniques to approach the problem of indoor blind navigation. Our main contribution is a camera-based place recognition system that can be is used to identify the user's location in the building. This system is then integrated with an intelligent path planning algorithm to find the directions to the user's destination. In this work, we develop a basic image retrieval system for indoors and then enhance its performance by exploring different image verification techniques.

3 Approach

The overall goal of the system developed in this thesis, is to be able to identify the user's current location and provide direction advice to his/her destination. The user sends an image of his/her current location to the system, where it is processed. Our approach for recognizing the user's location is based on image retrieval methods. In order to perform image retrieval, we train the system with the images of the building and tag them to their respective locations in the building. The image sent by the user is compared with all these images in the system and the ones similar to the query image are retrieved. The tags of these similar images are then used to identify the location of the image in the building.

Our approach for place recognition is divided into two main phases: mapping and localization. In the mapping phase, we train a model with the images of the building to create a dictionary of visual words, and the map of the building. In the localization phase, we build a baseline system that generates a similarity vector by comparing the query image with the images in the trained system. In addition to the baseline system, we have performed four different validation tests to reject the false positives. In this chapter, we describe our approach and then describe the details and experimental results in the following chapters.

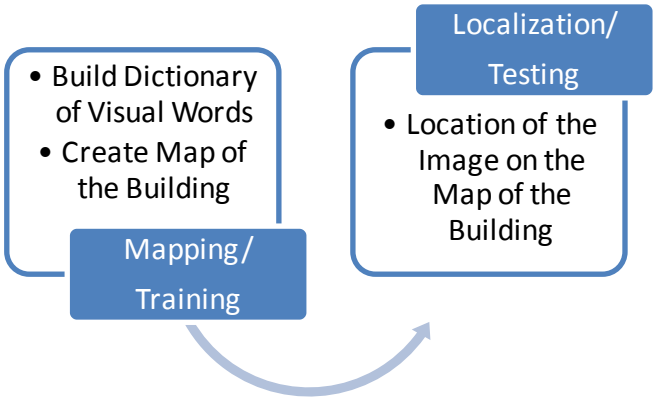


Figure 2 Overview of the approach

3.1 Mapping

The place recognition approach requires a model that is trained with the testing environment prior to localization phase. This process occurs once and is performed offline. Essentially, this model is trained with images of various locations inside the building and the physical map of the building. The primary purpose of these images is to perform image retrieval or place recognition and the physical map is utilized in the navigation process. In this work, we use a vector space model to represent these images in the system. In a vector space model, an image is represented as a vector or an image descriptor. To generate these image descriptors, we use the Bag of Words model [11],[14]. In this model, the images are used to build a dictionary of visual words, which is one of the primary products of the mapping phase, and is essential for representing these images in a vector space model. Next, the model is trained with the physical map of the building using a graphical representation. The different locations on the map correspond to the nodes of the graph and the edge between two nodes shows that they are neighbors in the building. The image descriptors are then tied to their respective locations in the graph, and are stored for use in the localization phase.

3.2 Localization

The fundamental purpose of this phase is to be able to localize an image against the physical map of the building. This image is taken by the user using a hand-held camera, such as phone cameras, web cameras and digital cameras, and queried to the trained model to identify his/her current location. In this phase, we use the trained model to perform image retrieval for place

recognition. When the user queries the image, the trained model uses the dictionary of visual words to process it and find the similar images in the system.

The dictionary is used to generate an image descriptor for this query image, using the same technique as described in the mapping phase. This query image descriptor is then compared with each of the trained image descriptors using the cosine distance metric to generate a ranking vector. The higher the value, greater is the similarity between the images. The top images retrieved using this ranking vector assist us in determining the location of the image in the building.

Although, the similarity ranking for image localization works in most cases, there are various situations where they fail to provide accurate information or do not consider significant factors in similarity such as common places and image geometry. To improve the accuracy of the images retrieved from similarity ranking, we explore four enhancements to the algorithm. These are homography test [15], query expansion test[16], physical map test and sequential images test. The homography and query expansion tests have been implemented and tested in several approaches. [15], [7]We propose the physical map and sequential images tests in this thesis as enhancements to the baseline system.

- Homography test [15]

  In this test, we perform image verification by using the 4-point RANSAC algorithm on the initial feature correspondences between the query image and the results obtained from the ranking vector. We extract these correspondences using a feature matching algorithm. Then an optimal homography is obtained for the correspondences and the number of inliers are counted. The results that are below a certain threshold of inliers are rejected.

- Query Expansion Test [16]

  This is an image retrieval algorithm where an image is queried, and the results are generated. Of these results, the top 2-5 choices are picked, and queried recursively as single images. The results are stored, and the most common results in the cumulative set of results are chosen as the final output.

- Physical Map Test

In this test, we consider the location tags of all the results retrieved, and reject those images with tags that are in the minority. This test helps us to verify the tagging process for certain areas of the building that look similar to each other.

- Sequential Images Test

This approach helps tackle the problem of tagging common areas, such as notice boards, plain walls, etc. In this process, we consider a sequence of images instead of just one query image. The user is requested to provide a query image as well as 2-3 neighboring images in the physical map. Finally, the system assigns a tag to the initial query image based on the tags obtained from the query image and these additional images.

In this thesis, we test and analyze these four different approaches independently and in different combinations.

3.3 Navigation

After successful localization, the next task is to direct the user to his/her destination. As mentioned earlier, the user will be prompted for a destination after localization. Once the user enters the destination into the system, an intelligent path planning algorithm will compute the path from the current location to the destination in the physical map. In this work, we use the Dijkstra's graph search algorithm [17] to generate a path from the user location to the destination. This path is then narrated back to the user as approximate number of steps to facilitate navigation.

4 Mapping

In the mapping phase, we develop a trained model for a specific testing environment. Since we are using an image retrieval approach to perform place recognition, it is crucial to have an efficient and intelligent trained model. In the system, this model is used to retrieve similar images for user localization and to request direction advice.

There are various ways to perform image retrieval. [18] In this thesis, we use the Bag of Words approach which has been developed and used in several approaches. [11],[14],[7]The main idea of this model is to represent an image in the vector space model, where each image corresponds to a vector or an image descriptor. In this model, the important features of all the training images are extracted and used to build a dictionary of visual words. To generate an image descriptor, the important features of an image are extracted and matched to the visual words in the dictionary. This image is then represented as "bags" of significant visual words, where each bag consists of the number of features that match the significant visual word. Therefore, building a dictionary is essential for representing an image in a vector space model.

This phase consists of two significant components:
1. Building a dictionary of visual words based on the images,
2. Creating a map of the building and tagging the images with its appropriate locations.
The localization phase uses this trained model to localize the user correctly and to navigate the user inside the building.

In this section, we will describe the process of developing the dictionary in detail.

4.1 Dictionary of Visual Words

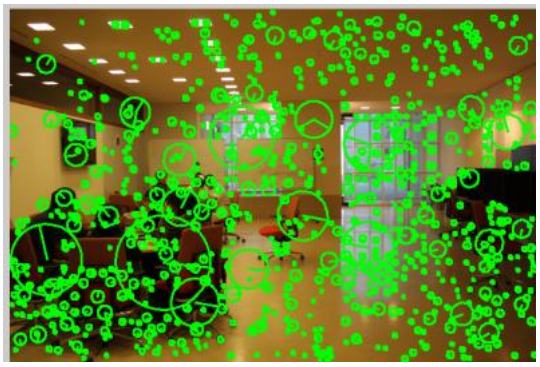


Figure 3 SIFT features

Building a dictionary of visual words, which is shown as step 3 in fig 4, involves data collection, feature extraction and clustering of the entire set of features based on their similarity. For data collection, several images of the testing environment are collected to train the model. To train the system with these images, we first extract significant features from these images to facilitate identification. There are several approaches to perform feature extraction such as SIFT, SURF, and MSER [19] .We use 128-dimensional SIFT (Scale-Invariant Feature Transform) descriptors [8]for feature extraction as these descriptors are invariant to rotation, luminosity and scale. Finally, we cluster these descriptors to generate the central points of the clusters. There are several clustering algorithms that can be used such as flat k-means, hierarchical k-means and random forest. In this thesis, we use the flat k-means clustering algorithm [18], however, it is also advisable to use the other clustering algorithms. We use the flat k-means clustering algorithm to cluster these features of all the images into k different clusters, and the cluster centers form the dictionary of k visual words, where k = {5000, 10,000}.

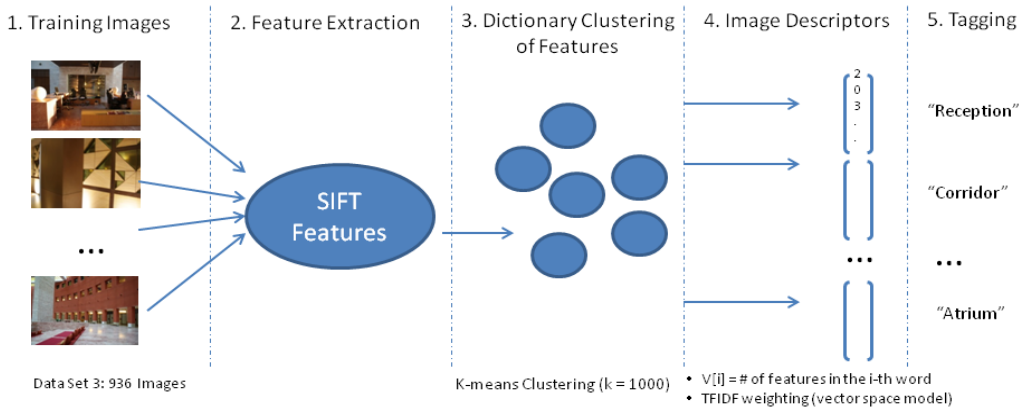


Figure 4 Graphical Representation of the Mapping Phase

Our next task is to represent these images in a vector space model using the bag of words approach. The k-sized dictionary generated is used to represent an image as a vector. For each image I in the training data set, a k-sized histogram vector V is generated, such that

$$V_i = |\#\ of\ features\ matching\ the\ i^{th}\ K-mean\ |$$

that is the number of SIFT features of image I that are closest to the i^{th} visual words of the dictionary. To down weight visual words that might appear more frequently and therefore be less informative, we use the Term Frequency – Inverse Document Frequency (tf-idf) weighting [18]. The term frequency measures the number of times a visual word occurs in the image with respect to the total number of occurrences of all the words that appears in that images to account for possible bias. Formally, it is calculated as

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where w_i is the i^{th} visual word, I_j is j^{th} the image, and $n_{i,j}$ is the number of features in the j^{th} image that are close to the i^{th} visual word. The inverse document frequency measures the importance of a visual word in all the images. It is calculated as

$$idf_i = log\ \frac{|I|}{|j: w_i \in I_j|}$$

Where idf_i is the inverse document frequency measure for w_i, $|I|$ is the size of all images in the set and $|j: w_i \in I_j|$ is the size of all the images in the set that contain the visual word, w_i. The tf-idf vector for the image I_j is calculated as follows:

$$tfidf_{i,j} = tf_{i,j} * idf_i$$

These tf-idf vectors for each image form the respective image descriptors.

In addition to collecting and storing images, we must train the model with the physical map of the building. In this stage, the official map of the building is represented as a graph where each node corresponds to a location in the map, and an edge between two nodes signifies that they are connected in the physical map. Each image in the training data set is then manually

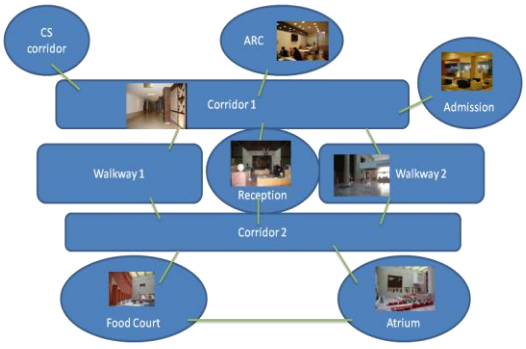


Figure 5 A graphical representation of CMU-Q's first floor

tagged to its appropriate location in the map.

At the end of this phase, we have a trained model that consists of various tagged images, and the map of the building. We now proceed to Phase 2 where we use this trained model to test the query images for image retrieval as well as image labeling.

5 Localization

As described earlier, the main purpose of the localization phase is to localize an image against the physical map of the building. To accomplish this, the user collects an image of the surrounding and sends it to the trained model for processing. This trained model is a product of the mapping phase and consists of the dictionary of visual words, the trained image descriptors and the map of the building. In this phase, we use these results in order to perform image localization. First, we create a basic image retrieval system that compares the query image with the images in the trained system and return a ranked list of images based on the similarity. We then improve the accuracy of the ranking by enhancing the algorithm using additional validation assessments.

5.1 Basic Image Retrieval System

The baseline system uses the cosine distance metric to return a ranked list of similar images for a query image. [14] The query image is collected by the user using a camera device and sent to the system. This system reads the query image and uses the dictionary of visual words that was created in the mapping phase to represent it as an image descriptor. Similar to the mapping phase, to generate the image descriptor, the SIFT descriptors of the query image are first extracted and compared with the visual words of the dictionary. The image descriptor is the size of the visual words in the dictionary and each element i in the image descriptor represents the number of raw features of the image in the i^{th} cluster.

This image descriptor is then compared with each image descriptors of the trained images using the cosine distance metric,

$$similarity = \frac{Q.A_i}{|Q||A_i|}$$

where Q is the query descriptor and A_i is a trained image descriptor. This generates a ranking or similarity vector that indicates the similarity of the query image with each image in the trained model. The higher the ranking, the closer the image is to the query image. When the results have been retrieved, the next step is to tag the query image with the location on the map of the building. To tag the query image, we consider the majority of the location tags of the top results generated by the ranked list. These labels had been manually assigned in the training phase.

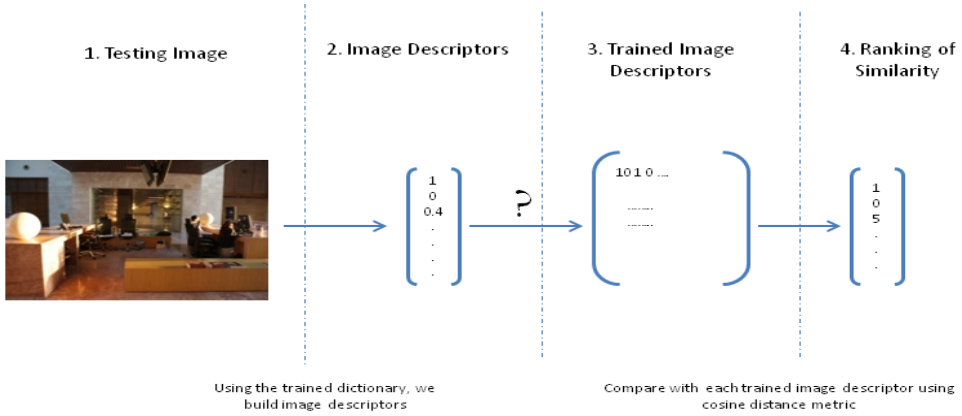


Figure 6 Graphical Representation of the Basic Image Retrieval System

5.2 Algorithm Enhancements:

Although the baseline system provides a decent indication of the similar images, there exist several cases where the ranking vector provides false indications or requires refinement. This is primarily due the fact that the image descriptors are merely a count of common features and not the order of their appearance on the image. These descriptors do not account for their location on the image or the geometry of the images. Furthermore, there are several areas inside buildings which look very similar and appear frequently. Examples include plain walls, notice boards, and common patterns.

Figure 7 Examples where the basic algorithm fails

To address these issues, we explore four enhancements to the performance of the baseline system that take into account additional factors for image retrieval. These factors include the geometry of the image, surrounding area of the image and its location on the physical map of the building. The four tests are homography, query expansion, physical map and sequential images tests.

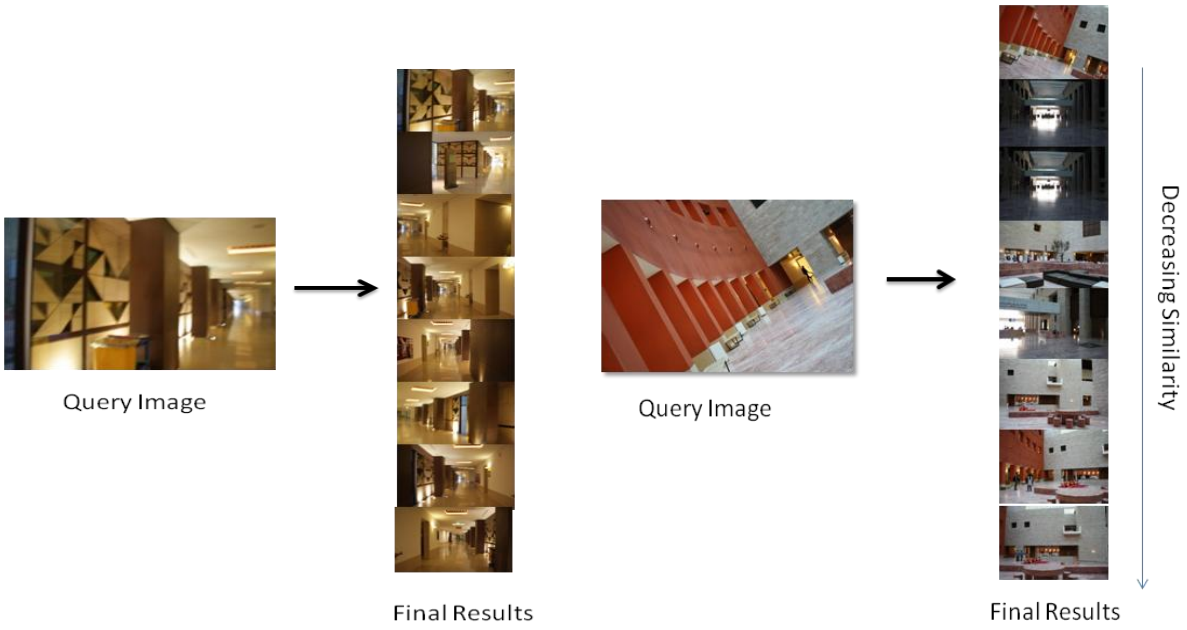


Figure 8 Examples where the basic retrieval system does not provide good results

5.2.1 Homography Test

Our main goal in this test is to perform spatial verification or test the spatial geometry of the resulting images, obtained from the ranking vector, with the original query image. This helps in rejecting those images that have been ranked similar in the basic image retrieval system but may not be geometrically similar such as flipped images. These tests use raw features of the images and perform feature matching to create a matrix that relates the two sets of features. There are several ways to perform spatial test such as 5- point algorithm and 8 point algorithm. However, these algorithms require several matches and hence perform slower on large datasets. In

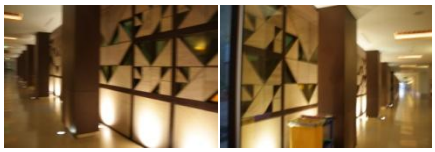


Figure 9 Flipped Images in CMUQ

this thesis, we use the homography test to test for spatial geometry. This test is relatively fast and performs well given loose threshold. The rest of this subsection will describe the homography test in detail.

To perform this test, we use a 4-point RANSAC algorithm[15] to fit the best homograph to a set of feature correspondences. We first extract and match the SIFT features from the two images using the feature matching algorithm [8]. These are 2D points on the two images that correspond to each other and are called feature correspondences. Our goal is to then fit a homography matrix such that x = Hx', where x is the list of 2D points on image 1, x' is the list of 2D points on image 2, and H is a 3X3 Matrix. This gives us the indication of the alignment of the two images. Since it is not possible to find a perfect homography matrix in all situations, we use a 4-point RANSAC algorithm to address this issue. This algorithm estimates the homography matrix with different sets of feature correspondences and eventually provides the homography matrix that fits best with the maximum number of inliers. If the number of inliers is above a certain threshold, we mark it is a valid result and reject the ones below the threshold. Homography test primarily helps in eliminating those images which have certain common features with the query image but are an image of the same area. By verifying the geometry of the images, it also helps in eliminating those images that are of the same location but taken with opposite directions. This helps in maintaining the information regarding the orientation of the user.

5.2.2 Query Expansion Test:

Query expansion test is a recursive querying approach that is inspired by the image retrieval work done by Philbin et. al. [16] As described in fig , the query image is first sent to the basic image retrieval system to generate a ranked list of images based on its similarity. After we obtain the ranked list of images for a query image, we pick the top 2-5 images, and query them recursively as new query images. We repeat the same process on the new query images, and store all these different sets of ranked results. Finally, we pick the common results from the cumulative set of ranked results and mark them as the final retrieved results.

However, the results obtained by the basic retrieval system are not spatially verified, i.e. they may have certain similarity in the images but may not have the spatial geometry. It is highly essential that the images that are being queried recursively are reliable ones. Therefore, we perform two different kinds of tests. In the first test, we use a combination of basic system as well as the homography test. We obtain the top results from the basic system and validate them

using the homography test. Then, we query on these spatially verified images. In the second test, we query the image using the homography test alone and rank the results based on the number of inliers. Since these images are spatially verified, we pick the top results and query on them recursively.

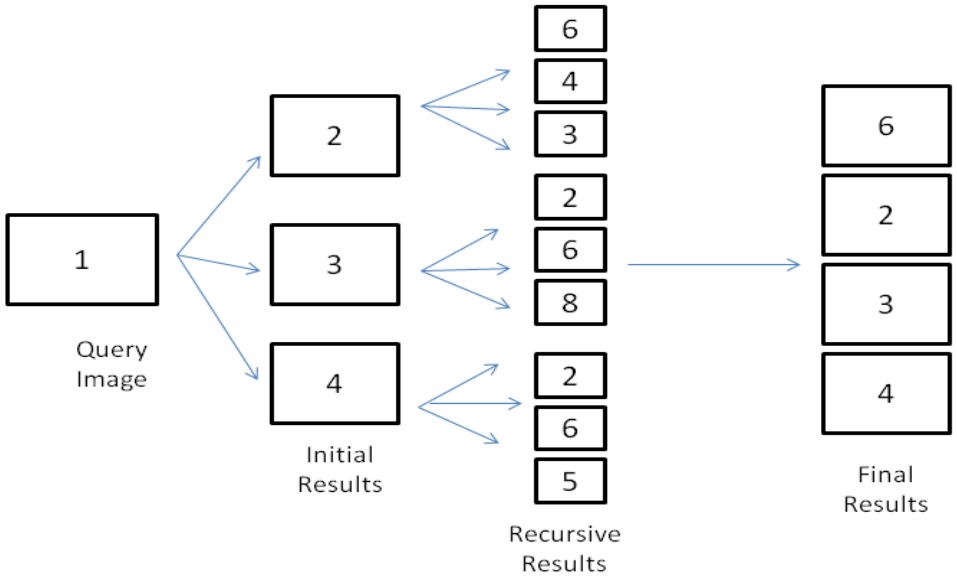


Figure 10 Graphical Representation of QET

5.2.3 Physical Map Test

Some areas of the building may have slight resemblances with different parts of the building. In such situations, when an image from these locations is queried, the basic retrieval system may give a high similarity score to those images that belong to different areas but have small resemblances. To resolve this issue, we utilize the underlying map of the building that was created in the mapping phase. This is relevant to this work as our application domain is navigation and we can safely ignore those images that might look similar but do not belong to the actual location of the image.

In this technique, we query the image in the basic image retrieval system and retrieve the ranked list of images. We pick the top 10 images in the ranked list and extract their location tags that had been assigned to each image in the mapping process. We choose the tag that is in majority and eliminate all those images that are in minority or do not belong to the same region. This test helps in rejecting those images that might appear to be slightly similar to each other but do not belong to the same region.

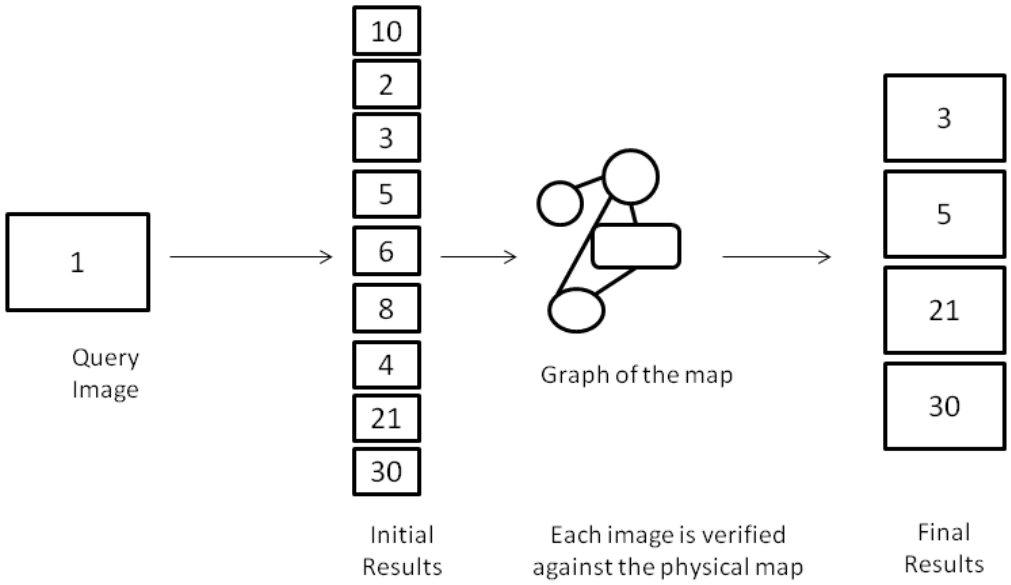


Figure 11 Graphical Representation of Physical Map Test

5.2.4 Sequential Images Tests

In some cases, the query image may be challenging for the system and humans themselves to recognize the original location of the image. These images may not have any significant features that can be extracted for generating a good image descriptor. Examples of such areas include plain walls, doors, white boards, etc. An example of this kind is shown in (include an image descriptor) Moreover, some buildings often have certain areas that may have significant features but not unique to one location. As shown in fig 12, examples of such instances are images of notice boards, common patterns inside the building. These areas although belong to different regions may have a similar image descriptor due to the similarity in their significant features. When such images are queried for localization, the system may retrieve similar images but these images may not belong to the correct region, and hence cannot help in localizing the query image correctly. In such cases, the system requires additional

Figure 12 Examples of images that do not contain unique features

information of the area. In this thesis, we address this issue by collecting additional images of the surrounding.

We propose this test to account for those regions that do not have significant features for effective comparison. In this test, we request the user to collect additional images of his/her surrounding area. These images can be taken by moving a step further or behind and/or turning around at the same location. We then perform a basic image retrieval test on each of these images and generate a ranked list of results. We can also choose to perform a homography inliers test and generate a ranked list of results based on the number of inliers. This gives us different ranked similarity scores for each image. We then compute an average over these scores and return the new ranked list of similarity scores. This increases the score of the images belonging to different directions in the same region but is not similar to the original query image, while maintaining the score of those that have a similar score in all the different images. We then pick the top 10 – 15 results obtained from the ranked list and these form the final similar images. Finally, the system assigns a location tag to the original query image by considering the majority of the tags of these similar images.

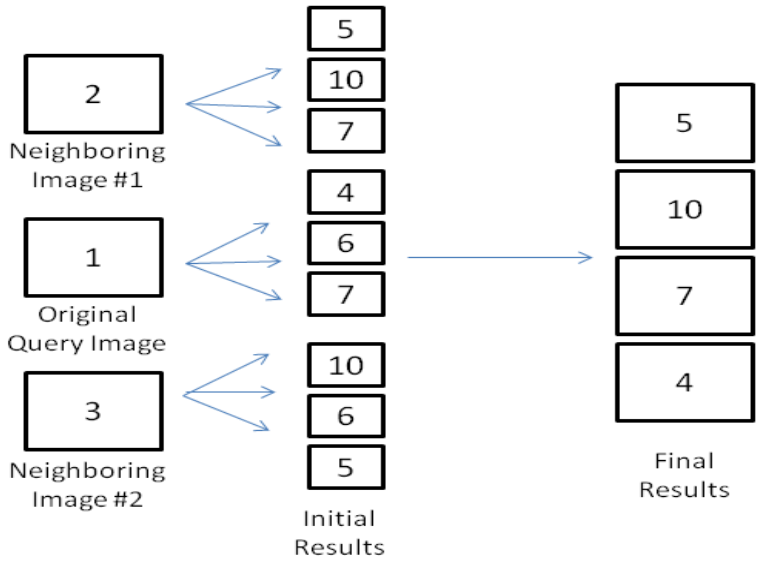


Figure 13 Graphical Representation of Sequential Images Test

6 Experiments and Results

To perform experiments with our place recognition system, we first trained the model with the images of the building and the map of the building. We evaluated our system on the first floor of the Carnegie Mellon University in Qatar building. We collected our data with an SLR Camera and the images were resized to 640 X 480. We then performed the bag of words approach to represent these images in the vector space model. To perform the bag of words approach, we first built a dictionary of visual words. In order to build the dictionary, we extracted the SIFT descriptors and clustered them using the flat k-means clustering algorithm where k = {5000, 10,000}. The 5000 k-means (centroids) generated in this algorithm formed the dictionary of visual words. Finally, we used this dictionary and the features of an image to generate the k-sized image descriptor V, which was calculated using the formula mentioned in Chapter 3 and down-weighted using tf-idf.

Next, we performed image retrieval on the query image. This query image is also taken using the SLR Camera manually, resized to 640 X 480 and sent to the system for image retrieval. In order to perform image retrieval, we need to represent the query image as an image descriptor. Then we compare this query image descriptor with all the trained image descriptor using the cosine distance metric and generate a similarity score vector.

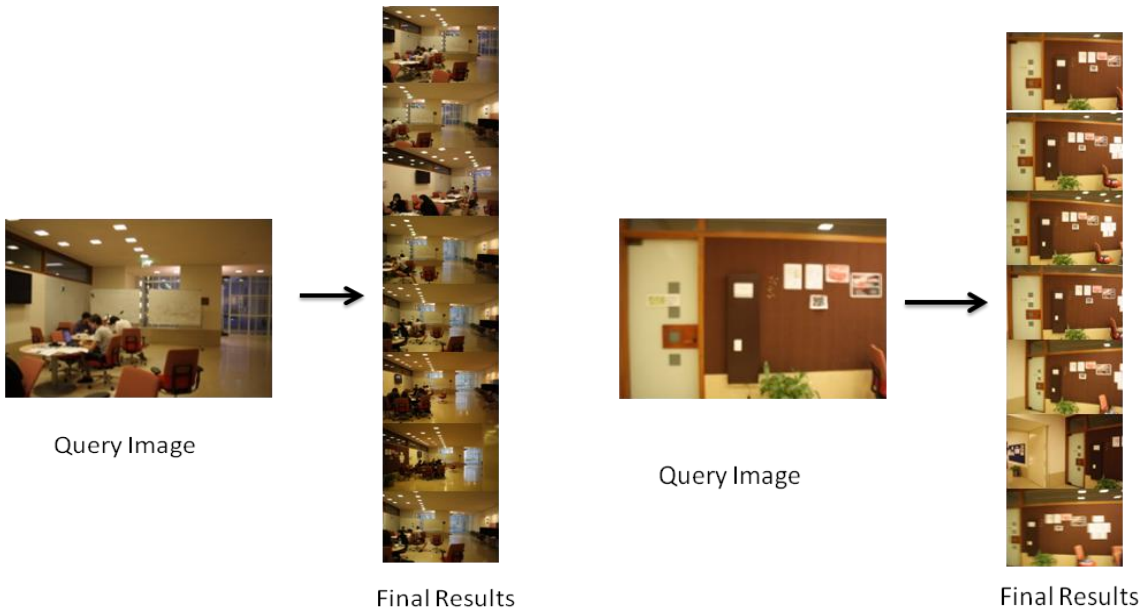


Figure 14 Examples where the basic image retrieval system retrieves good results

The images represented in the fig:14 are the results obtained for two query images. Two images were queried with the basic image retrieval system and the top results were recorded. These images were used to test the performance of the basic image retrieval system and in both these cases, the basic image retrieval system retrieved good results. However, this was not the same case with many other images. In fig 15, we can see observe that the initial results obtained by querying the image in the basic image retrieval system performed fairly. The results included images that are similar to the query image but are not in the same orientation. These cases require additional testing for spatial geometry. Therefore, we can notice that when these images are sent for spatial verification, the final results are all geometrically valid.

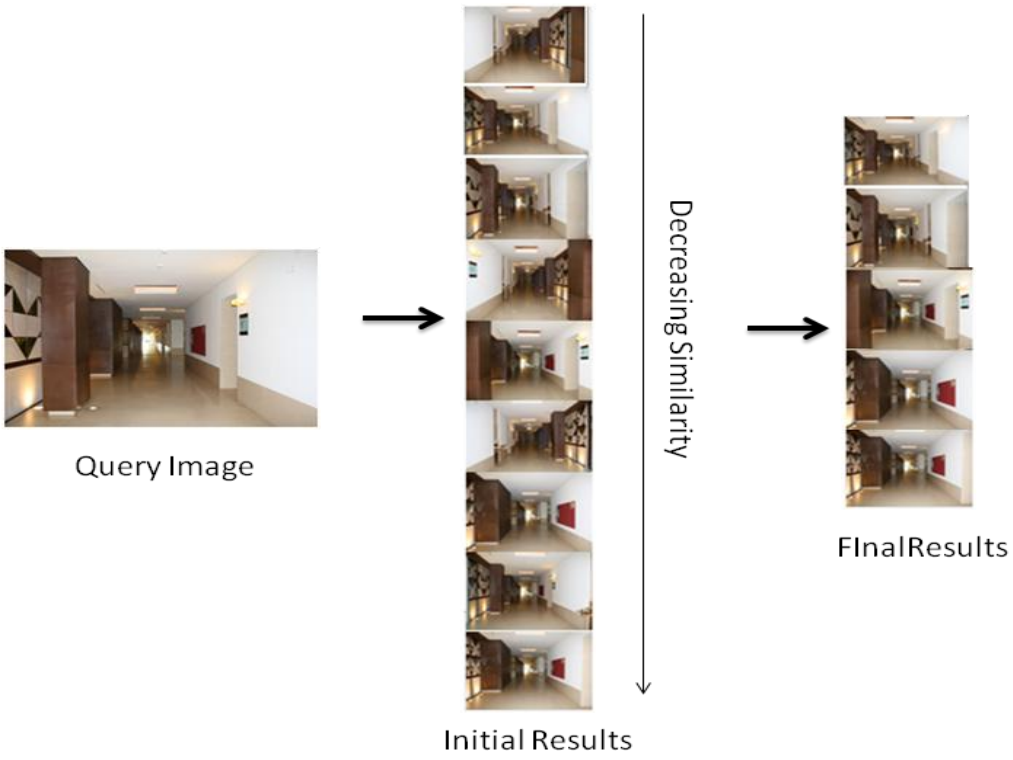


Figure 15 This image shows the initial results given by the basic system and then the final results after spatial verification

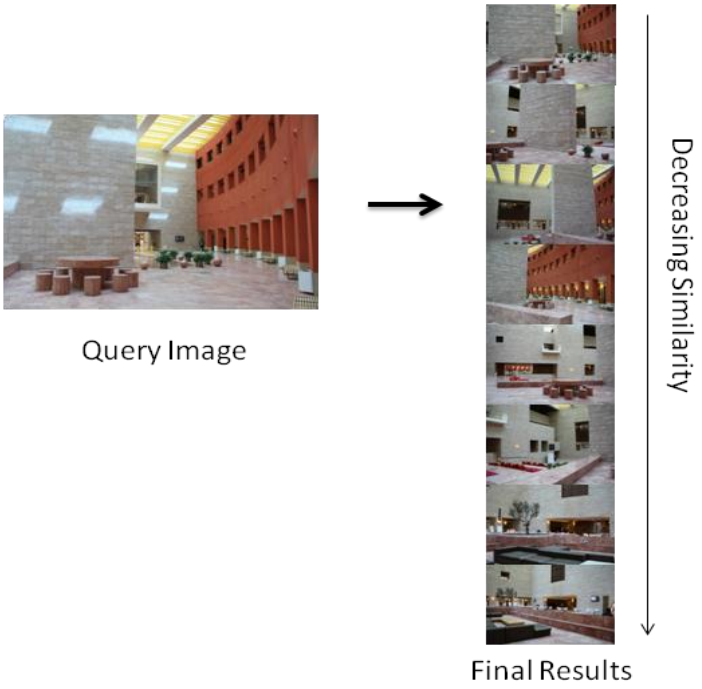


Figure 16 This image shows results generated using only homography test

However, there are certain cases, where even homography test fails to reject false positives. An example of this case is shown in fig 16, where the 7th and the 8th images do not belong to region. However, we notice that the top 6 images have been retrieved correctly. In this case, we can apply the physical map test and since the tags of the last two images are in minority, and away from the tags of the majority results, we can reject these images to achieve true positives.

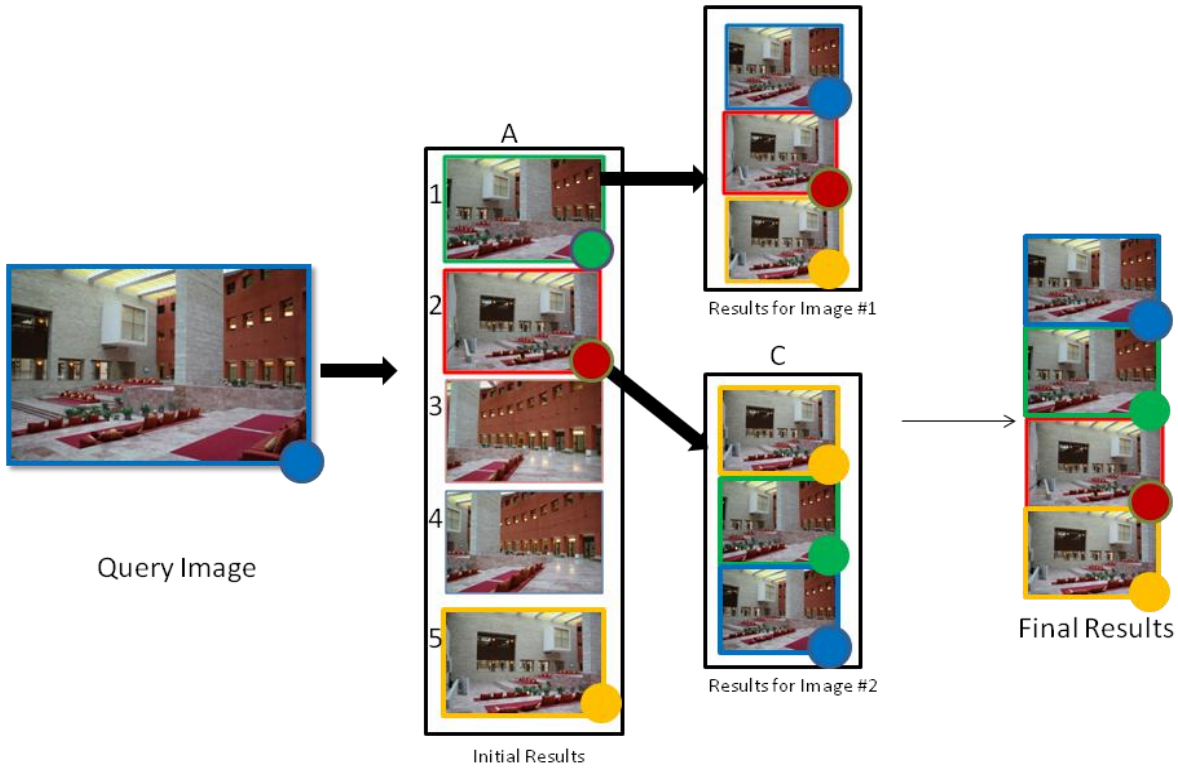


Figure 17 An example of Query Expansion Test

As example of the query test in shown in fig. In the query test, we query for an image (with the blue icon) with the basic image retrieval system and then spatially verify them. As mentioned earlier, spatial verification is a necessarily as the query expansion test is highly dependent on the accuracy of its initial results. We then pick the top 5 results which are marked as initial results in the fig 17. For each of the top 2 results, we recursively query them, that is repeat the process of image retrieval followed by spatial verification. We can notice that four images have appeared in the most number of sets and therefore, marked as final results.

Sequential test

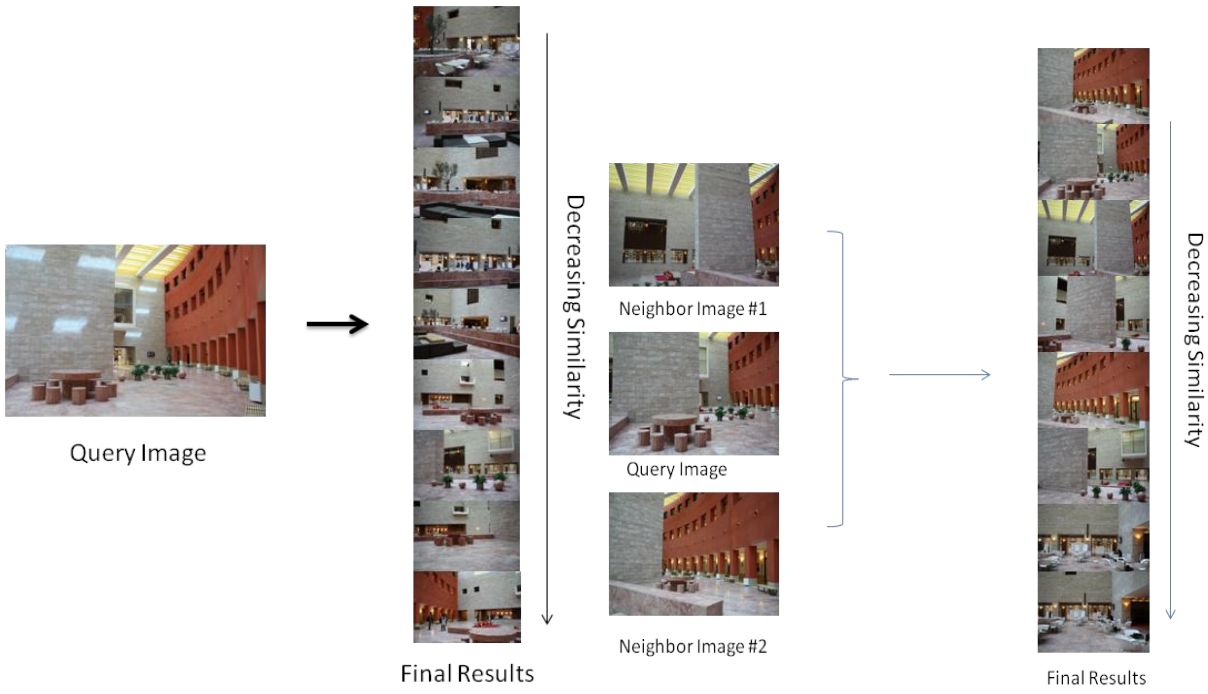


Figure 18 Example of an image with basic system results on the left and sequential test results on the right

Figure 18 is an example of the sequential images test and shows a comparison of results obtained by the basic image retrieval system and the sequential images test. The left side of the figure contains the results obtained on querying the image with the basic image retrieval system and right one illustrated the results obtained on testing them with sequential images. We can notice that the images retrieved by the basic system contained many false positives. The top 5 results are slightly similar to the query image in appearance but do not belong to the correct location. When we perform sequential test along with spatial verification, we can notice that the top 6 images belong to the same location as the query image.

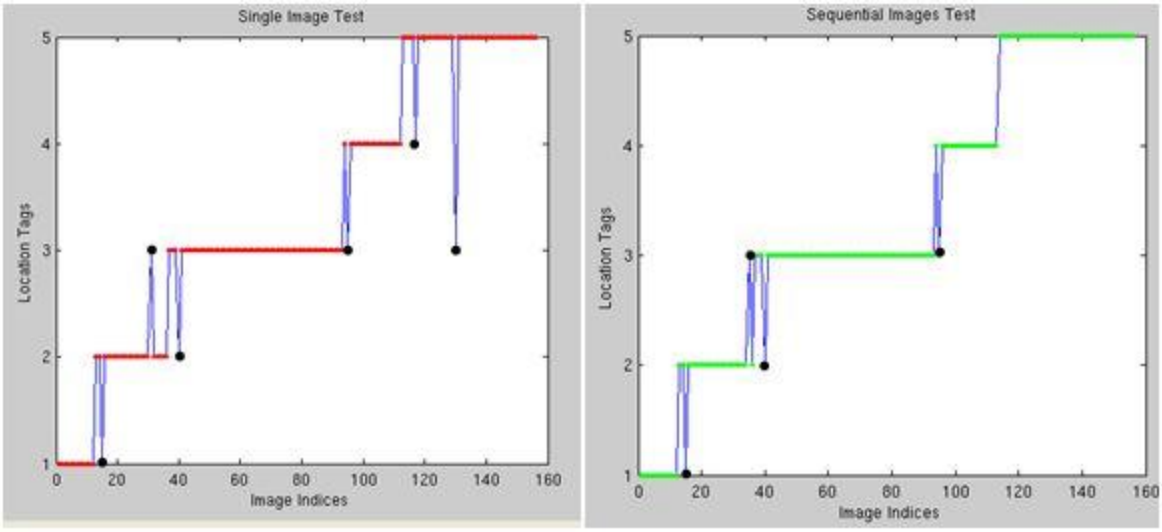
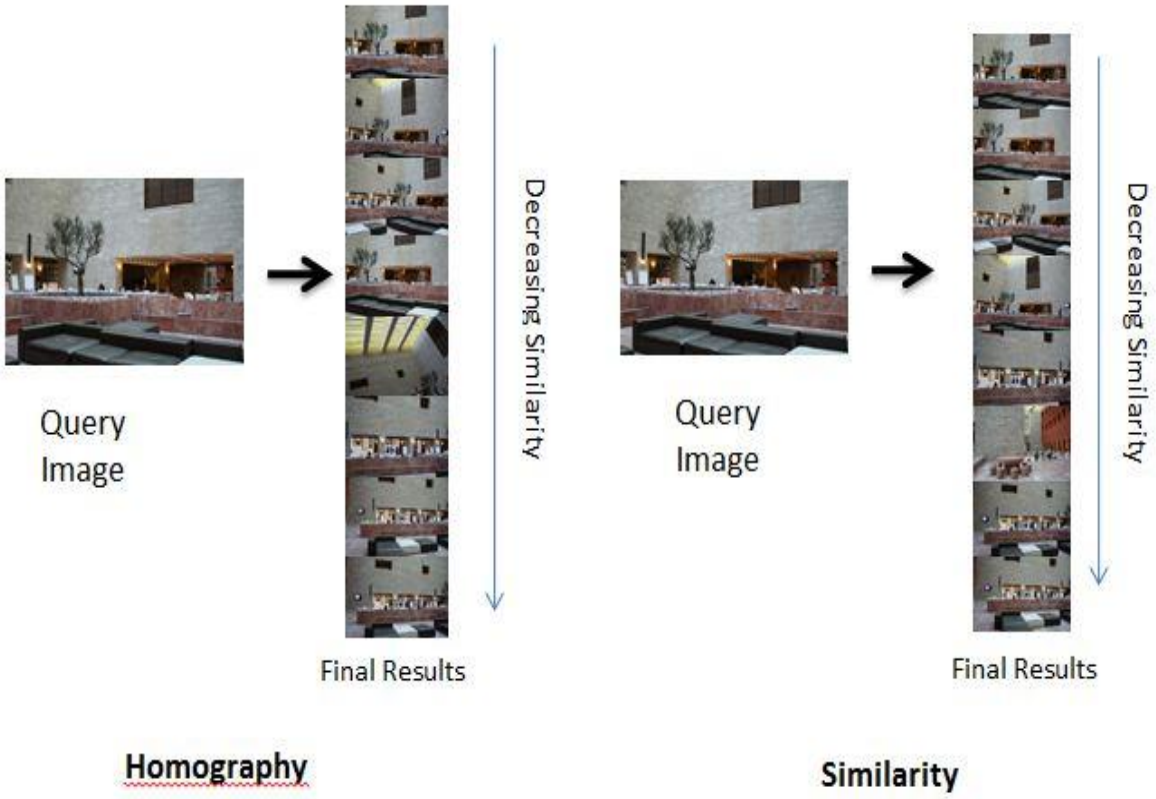


Figure 19 Sequential Image Test

Sequential Image Test can be better understood by referring to fig 19. This figure illustrates the tagging performance of the images based on the basic image retrieval system and the sequential images test. This was tested on 156 sequential images from the Reception Desk at CMU-Q to long corridor and finally the Academic Resource Centre (ARC). All the images at the Reception Desk are labeled as 1, the images at the corridor are labeled 3, and the ones that belong to the ARC are tagged 3. The images that lie at the intersection of the reception desk and the corridor are labeled 2 while the ones at the junction of the corridor and the ARC are marked as 4. When we tag each query image based on the top results obtained by the basic image retrieval system, we notice that while most of them are tagged correctly (red dots), there are quite many that are localized incorrectly (black dots). On the other hand, when we consider their neighboring images in the left graph, we notice a reduction in the number of incorrectly tagged images reduce. The ones that remain to be incorrectly localized are those images that belong to the intersection and cannot be easily localized as they may be considered as a part of both common areas.

6.1 Analysis


Query Image
Decreasing Similarity
Final Results
Homography
Query Image
Decreasing Similarity
Final Results
Similarity
Query Image
Decreasing Similarity
Final Results
Query Expansion Test + Homography

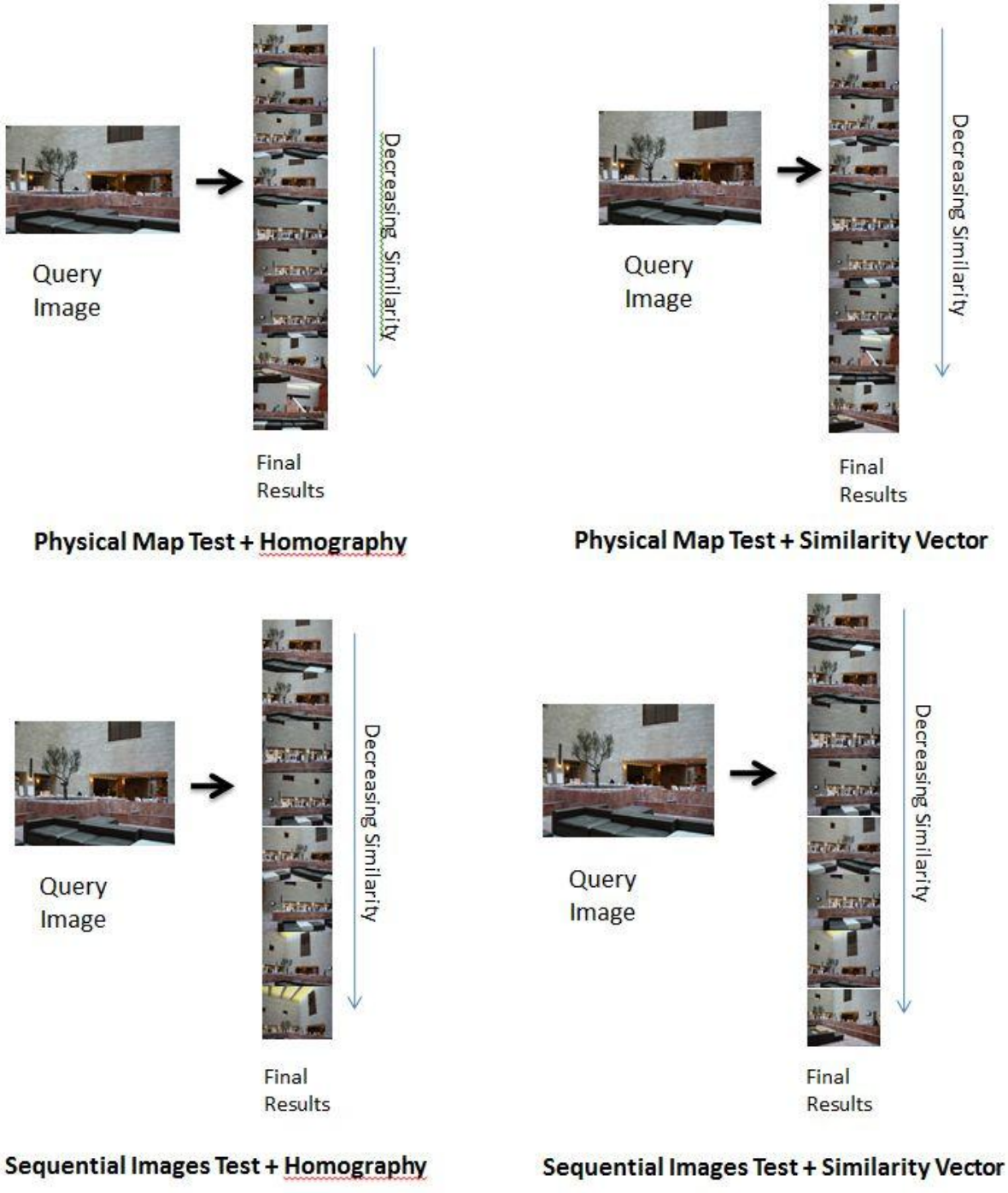


Figure 20 Results obtained for the same query image using different combinations of algorithms

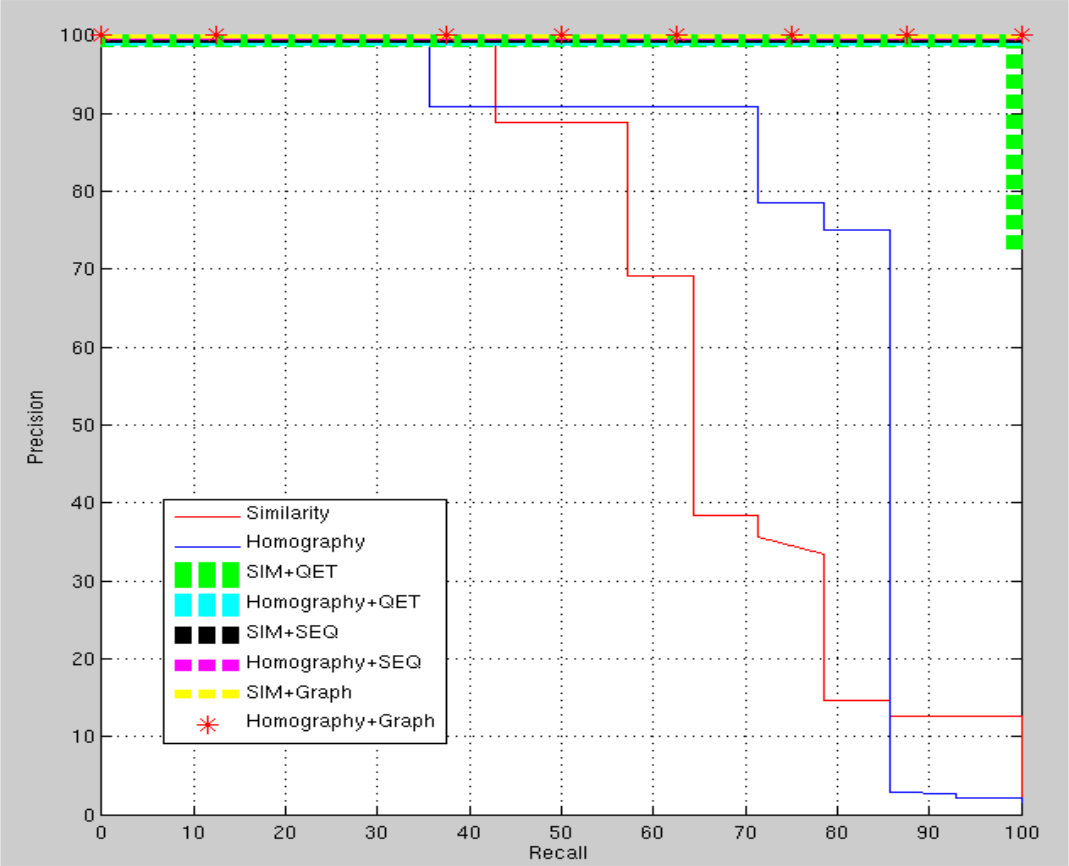


Figure 21 Precision Recall Curve for the query image and results in fig 20

Figure 21 shows the precision recall curve for a query image. The query image and the results retrieved by each test are illustrated in fig 20. This precision recall curve has been plotted for the top 20 images retrieved from the different place recognition tests. We can observe that the basic image retrieval system labeled as similarity in the graph performs the worst amongst all the tests. The homography test performs slightly better than the basic system as it tests for spatial geometry of the image as well. However, we notice that the combination of these results with the additional tests improves the retrieval performance of the system. The results obtained from the basic image retrieval system are validated by the additional tests and the false positives are rejected.

7 Conclusion & Future Work

In this thesis, we have addressed the problem of indoor blind navigation using computer vision techniques for place recognition. The visually impaired and blind are generally dependent on other aids for navigation assistance. The specific aim of this thesis was to develop and evaluate different place recognition techniques that can be used with an intelligent path planning algorithm on a portable device to enable the visually impaired to navigate independently. The main contributions of these include:

- A camera-based place recognition model to enable user localization in indoor environments
- Analysis of five different place recognition techniques/enhancements
- Introduction of two enhancements to the place recognition algorithm
- Evaluation of the tests in the Carnegie Mellon University in Qatar (CMUQ) building

The overall aim of this system is to be able to localize the visually impaired user and provide direction advice to his/her destination. The user collects an image of the current locations and sends it to the trained system for localization. This model has been trained with the images and the map of the building. Upon receiving a query request, this model uses an image retrieval to retrieve similar images in the system and localize the user based on the location of these images in the map of the building. Upon localization, the system computes the path to the destination. In this thesis, we divided our approach into two phases: mapping and localization. In the

mapping phase, we use a Bag of Words approach to build a dictionary of visual words and create a map of the building. In the localization phase, we compare the query image with each image in the system and perform a basic image retrieval test.

In addition to the basic image retrieval test, we enhance the algorithm by considering various factors that are associated with indoor localization. Four tests are explored to localize challenging images such as notice board, plan walls, common areas in the building. These include the homography test, query expansion test, sequential images test and the physical map test. In this thesis, we introduce the latter two tests to address to perform better localization. The sequential images test helps localizing images that belong to areas that do not that any significant or unique features for effective comparison. The physical map test utilizes the original physical map of the building to locate neighboring images from the results retrieved and eliminates those results that may look similar to the query image but do not belong to the region.

In this thesis, we tested our system inside Carnegie Mellon University in Qatar building. The results have been illustrated and evaluated using precision recall curves. From the results obtained, the combination of homography test along with the graph test performs better in validating all retrieved images.

There are several ways this thesis work can be enhanced. The place recognition technique currently does not account for the resolution of objects in the images. This factor has been tested in the work done by Philbin et. all and has proved to improve results significantly. Adding this enhancement to this work, may improve the retrieval results and eventually the localization results. Furthermore, this system can be tested with visually impaired users to evaluate their experience with indoor navigation.

8 References

[1] J. Coughlan, R. Manduchi, and H. Shen, "Cell phone-based wayfinding for the visually impaired," *Proc. IMV 2006*, 2006.

[2] J.A. Hesch and S.I. Roumeliotis, "Design and Analysis of a Portable Indoor Localization Aid for the Visually Impaired," Jun. 2010.

[3] L. Ran, S. Helal, and S. Moore, "Drishti: an integrated indoor/outdoor blind navigation system and service," 2004.

[4] A. Hub, J. Diepstraten, and T. Ertl, "Design and development of an indoor navigation and object identification system for the blind," *ACM SIGACCESS Accessibility and Computing*, 2003, pp. 147–152.

[5] S. Willis and S. Helal, "A passive RFID information grid for location and proximity sensing for the blind user," *University of Florida Technical Report number TR04-009*, 2004.

[6] Y. Sonnenblick, "An indoor navigation system for blind individuals," *Proceedings of the 13th annual Conference on Technology and Persons with Disabilities*, 1998.

[7] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *The International Journal of Robotics Research*, vol. 27, Jun. 2008, pp. 647 -665.

[8] D.G. Lowe, "Object recognition from local scale-invariant features," *iccv*, 1999, p. 1150.

[9] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: speeded-up robust features," *9th European Conference on Computer vision*, 2008, pp. 346–359.

[10] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, pp. 1615–1630.

[11] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," *IEEE International Conference on Robotics and Automation*, 2000, pp. 1023–1029.

[12] A. Pronobis, O. Martinez Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *The International Journal of Robotics Research*, vol. 29, 2010, p. 298.

[13] S. Willis and S. Helal, "RFID information grid for blind navigation and wayfinding," *Ninth IEEE International Symposium on Wearable Computers, 2005. Proceedings*, 2005, pp. 34–37.

[14] I. Posner, D. Schroeter, and P. Newman, "Using scene similarity for place labelling," *Experimental Robotics*, 2008, pp. 85–98.

[15] R. Hartley and A. Zisserman, *Multiple view geometry*, Cambridge university press, 2000.

[16] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," 2007.

[17] E.W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, 1959, pp. 269–271.

[18] C.D. Manning, P. Raghavan, H. Sch\ütze, and E. Corporation, *Introduction to information retrieval*, Cambridge University Press, 2008.

[19] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, 2004, pp. 761–767.