Exploiting Pre-trained Deep Convolutional Neural Networks for Robust Visual Indoor Place Recognition

By

Ujala Razaq

2011-NUST-MS-CS-05

Supervisor

Dr. Muhammad Muneeb Ullah

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science (MS-CS)

In

NUST School of Electrical Engineering and Computer Science (SEECS)

National University of Science and Technology (NUST), Islamabad, Pakistan.

(2015)

Certificate

Certified that the contents of thesis document titled "Exploiting Pre-trained Deep Convolutional Neural Networks for Robust Visual Indoor Place Recognition" submitted by Miss Ujala Razaq have been found satisfactory for the requirement of degree.

Advisor: _____

Dr. Muhammad Muneeb Ullah

Committee Member1: _____

Dr. Asad Anwar Butt

Committee Member2: _____

Dr. Anis ur Rehman

Committee Member3: _____

Dr. Omer Arif

Abstract

This thesis addresses the problem of visual indoor place recognition (e.g., in an office setting, automatically recognizing different places, such as offices, corridor, wash room, etc.). The potential applications include robot navigation, augmented reality, and image retrieval, etc. However, the task is highly challenging due to the large appearance variations in such dynamic setups (e.g., view-point, occlusion, illumination, scale, etc.). While local feature based methods (e.g., bag-of-features) have been promising, they are still limited in their capability to tackle severe visual variations. Recently, Convolutional Neural Network (CNN) has emerged as a powerful learning mechanism, able to learn higher-level deep features when provided with a relatively large amount of labeled training data. Such networks have shown state-of-the-art object and scene recognition results on the ImageNet and Places dataset. Here, we exploit the generic nature of CNN features by employing the pre-trained CNNs (on objects and scenes) for deep feature extraction on the challenging COLD dataset. We demonstrate that these off-the-shelf deep features when combined with a simple linear SVM classifier, outperform their bag-of-features counterpart. Moreover, a simple combination scheme, combining the local bag-of-features and higher-level deep CNN features, highlights their complementary nature. We benchmark our results with two other methods, and present superior results on the COLD dataset.

Certificate of Originality

I hereby declare that the research paper titled “Exploiting Pre-trained Deep Convolutional Neural Networks for Robust Visual Indoor Place Recognition” is my own work and to the best of my knowledge. It contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NIIT or any other education institute, except where due acknowledgment, is made in the thesis. Any contribution made to the research by others, with whom I have worked at SEECS-NUST or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project’s design and conception or in style, presentation and linguistic is acknowledged. I also verified the originality of contents through plagiarism software.

Signature: _____

Author Name: Ujala Razaq

Acknowledgements

I bow before Almighty Allah to express my gratitude for He is the only one who brightens my mind and heart when I feel standing alone in the darkness. He is the only one who helps me when I fall and who is the only hope when I feel broken. I am nothing but His benevolence makes me what I am today. He blessed me even more than I deserve. Thank you Allah!

My words are not enough to pay special credit and appreciation to my supervisor, Dr. Muhammad Muneeb Ullah for their continued support and encouragement. I wish to convey my sincere thanks to him for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor than him, for my research.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Asad Anwar Butt, Dr Anis ur Rehman and Dr. Omer Arif for their insightful comments and encouragement. I offer my sincere appreciation for the learning opportunities provided by my committee.

I place on record, my sincere thank you to my Institution and the Department for providing me a platform where I can learn not only about course books but also about how to excel in every field of life. My Institution makes me feel proud.

I wish to express my special thanks to my friends Mehreen Shakoor for being there as my best friend when I was facing the most difficult time of my life, she remained with me and endowed me with the best she could, Anjum Afzal and Asim Afzal for lending their helping hand in technical matters. Thank you so much!

I would also like to thank my teachers, my classmates, my friends and my colleagues for being the best human beings I have ever met and to those who forwarded positive criticism to me.

I also place on record, my sense of gratitude to one and all, who directly or indirectly, have lent their hand in this venture.

I express my heartfelt thanks to my sister Honey for all she had done for me and to my brothers Usman and Subhan for they are the angels of my life. This whole success is due to my siblings. I couldn't have come to this milestone without their help and support, and above all their love and care. Their encouragement when the times got rough is much appreciated and duly noted. It was a great comfort and relief to know that they are with me.

Last but not the least, no words in this world are enough to show gratitude and say thanks to my Parents for supporting me spiritually throughout and for all of the sacrifices they have made for me. Their prayers for me were the only thing that sustained me thus far. They were always there for me to support in the moments when there was no one to answer my queries.

My deepest gratitude!

I truly dedicate my endeavor to my Parents

For their endless love, support, encouragement and sacrifices for me.

Table of Contents

List of Figures

List of Tables

Chapter 1

1. Introduction

Scientists are striving to build artificial intelligent systems in a bid to replicate human intelligence. Intelligence comprises of different capabilities, such as learning, understanding, thinking logically, problem solving, and planning etc. Research is being conducted on these aspects of intelligence to make intelligent systems that can support human beings. One aspect of intelligence involves recognition of different objects, scenes, persons, and their activities, based upon the previously learned knowledge. In other words, scientists are working to develop machines which can see and interpret, thereby mimicking the human visual system. Google's reverse image search which allows its user to use a picture as query to find related images from around the web is one such practical example.

1.1 Automated Visual Recognition

An automated visual recognition system is trained on representative images of the target classes (e.g., in home setting, the target classes could be bathroom, bedroom, kitchen, corridor, etc.). Once trained, the system is expected to correctly classify similar unseen images, which are not used during the training phase. In other words, the trained system generalizes its knowledge to recognize unseen images.

Several diverse applications of the recognition automated visual recognition system are:

- *Object recognition* (also called *object classification*) – one or several pre-defined or learned objects and object classes can be predicted, usually together with their 2D positions in the image or 3D poses in the scene.
- *Identification* – an individual case of an object is predicted. For example, identification of specific persons faces or fingerprints, identification of handwritten digits, and identification of a specific vehicle.
- *Detection* – localization of regions of interest in images or videos. For example, detection of possible abnormal cells or tissues in medical images or detection of a vehicle in an automatic road toll system.

1.1.1 Visual recognition of indoor scenes

In this thesis, we follow a fully-supervised approach and classify indoor places (see Figure 1.1). We use the challenging COLD dataset [14], which is comprised of indoor images, categorized into different classes. Images are provided to a feature extraction module. Subsequently, a model is learnt using the extracted features and class labels. This learned model is used for classifying an unseen test image. And the accuracy is calculated by the number of test images correctly classified by the model classifier.

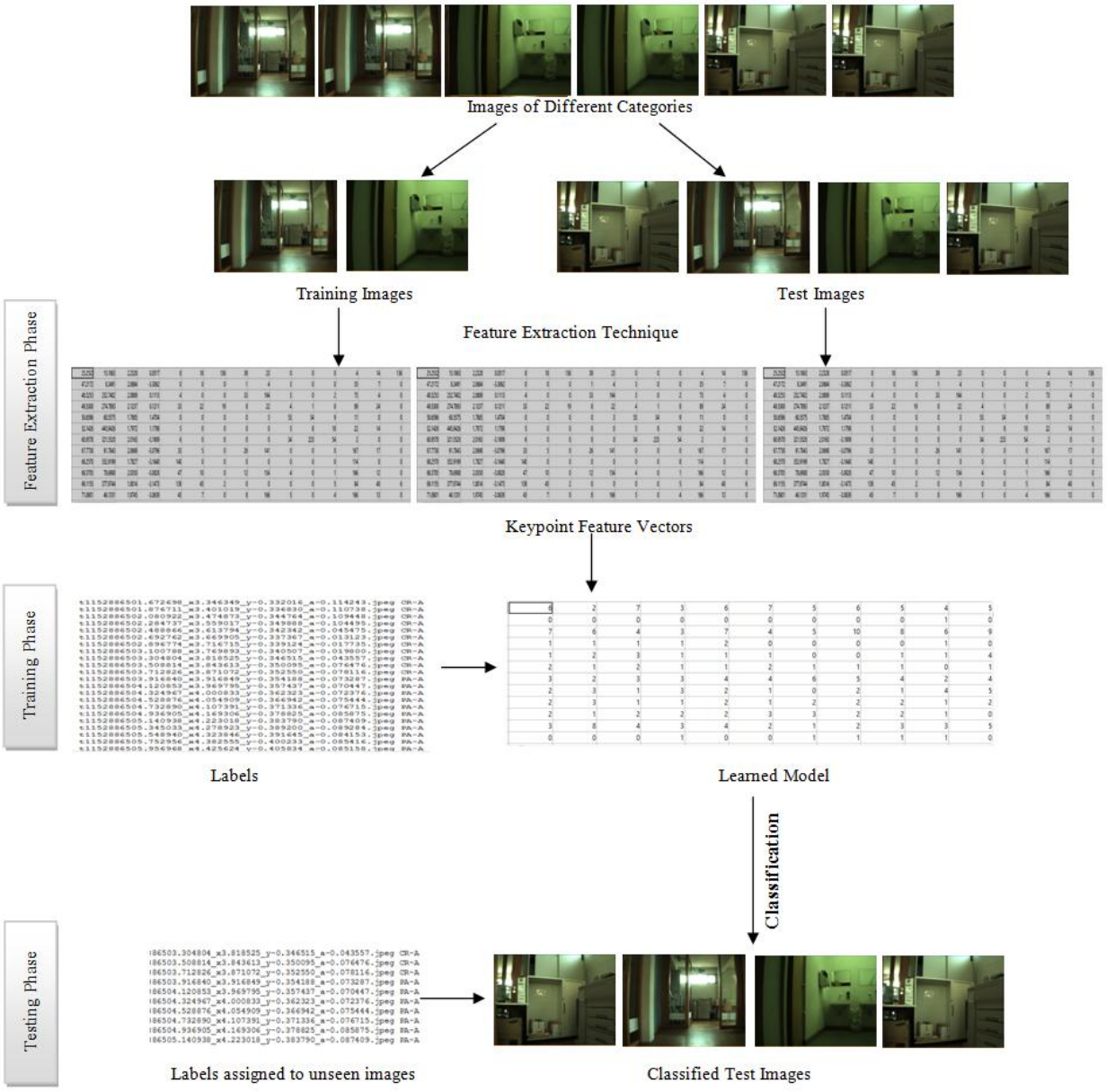


Figure 1.1 A Flowchart depicting the overall process of visual recognition

1.3 Problem Statement

Automated visual indoor place recognition is a challenging task due to large appearance variations in realistic setups. Image recognition is a process that aims to assign class labels to unseen images supported by the information learned previously on the trained model. This whole process should be automatic and less human intervention is desirable.

At this stage, it is important to note, that the recognition process can be divided into a number of steps. Before the classification can be done, keypoint features from the images need to be extracted. As the recognition problem is an intricate one, sometimes it becomes very hard to

speculate the parameters for the training model. Therefore, we may attempt to get a set of images (training set) that represents all classes and use it in order to train the classifier. Typically, in case of image recognition, the training images are labeled, that is their distinctive membership to one of the classes is known before training. During training phase, special attention is paid to the fact that parameters of the model are selected such that allowing it to reduce error on the training set. This is for the reason that we hope after training, system will be able to not only recognize training instances properly but also unseen test images encountered in the future. This ability is known as generalization, as the system can now generalize its previously leaned knowledge to recognize unseen images. But it should also be kept under consideration that the system should not overfit on training samples that it loses it generalization ability. The generalization is a very essential property for place recognition systems, because of the fact that the environment and the conditions under which the system is going to perform alter continuously. Moreover, images of places can be captured from different viewpoints. Therefore, it is important to use such classification and learning techniques that are able to generalize their knowledge. The training set should be also carefully chosen in order to instill the variability of the environment. Different challenges that come across during this task are mentioned under sub-heading.

1.3.1 Challenges

Indoor scenes are different and challenging than outdoor scenes when it comes to recognize those using AI techniques. The reason for this difference is that some indoor places are characterized by the their spatial properties, for example, corridors, stairs etc, while other indoor places are well characterized by the objects they contain, for example a room can be a bedroom, office, dining room etc, the difference is created by the objects that are placed in that room. Another difficulty is that indoor environments are usually imaged at a much closer distance than outdoor scenes; therefore they present a much higher variability in their visual appearance as the imaging viewpoint changes. Figure 1.2 [14] shows different images with the aforementioned variations, (a) shows people come in and go out of a place and cause variation, (b) shows variations caused by illumination conditions.

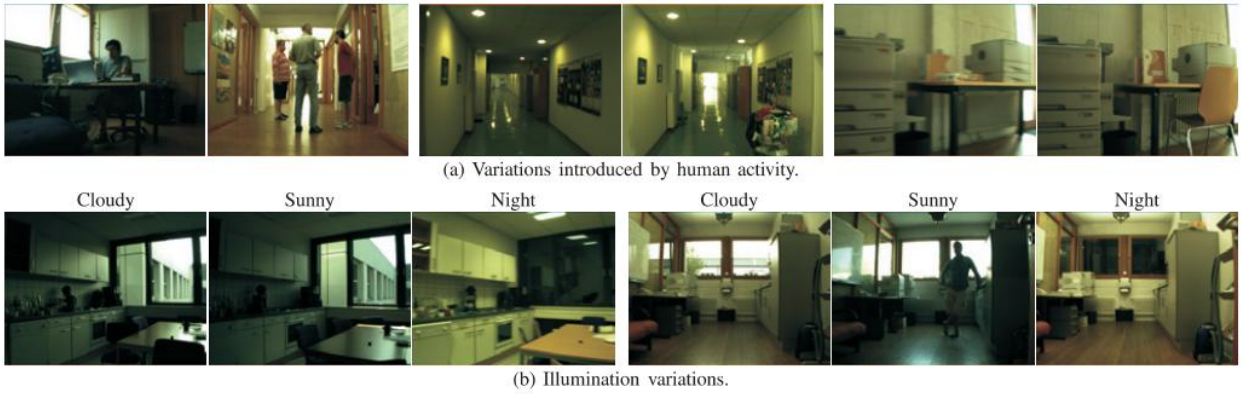


Figure 1.2 Images with different variation

1.3.2 Proposed Solution

To the challenging problem of Indoor Place Recognition, a possible solution would be to study deep and high level features in combination with local features. Deep features provide more details about objects and variations occurred due to various reasons. System can learn finest aspects of images when trained on deep and high level features. Therefore we can say that deep features can be very helpful in designing a place recognition system that is more accurate and precise.

1.4 Outline

The thesis is organized as follows. Chapter 2 presents the literature review, background of Bag of Visual Words Technique and SVM Classification and gives a description of the COLD database that was used during all experiments with the place recognition. Chapter 3 presents the methods used to extract characteristic image features, based on which the images are classified using a linear SVM classifier. Experimental evaluation and results are presented in Chapter 4. Thesis concludes with a summary and suggestions for further research in Chapter 5.

Chapter 2

2. Background

Real world environments are not stationary; there are many kinds of variation in real world settings. To build systems with vision based solutions, it is essential to make sure that the algorithms for such tasks provide robustness and invariance against dynamic changes influencing the appearance of places observed over time. There are mainly illumination and light variations that are frequently observed in real setups, and changes that are introduced by human activity (like people coming in and going out of the places, objects and furniture being repositioned or removed from the room etc) as well. In vision based systems to mimic human vision system, handling categorical changes properly is also very important. This can be thought of "Aliasing", where two different places look similar. Humans are able to recognize different places and categorize them. For example, our brain can classify a room as "an office", "a bedroom", "a kitchen", "a Conference room", or "a classroom", even if we see these rooms for the first time, in a very first glimpse, we can recognize and categorize them. Similarly in case of objects, our brain can recognize all objects and categorize them even if they look different in shape. This is because humans are able to build robust categorical models of places. It is an exceptionally complicated task to provide similar capability for an artificial place categorization system due to great within-category variability. Finally, a universal localization system should deliver constant performance independently of the environment it is applied in.

In this chapter, methods and techniques employed for the task of indoor place recognition in most recent and closely related literature are reviewed. We also discuss Bag of Visual Words and SVM Classification to revitalize the concepts for our task. In the end of this chapter, a brief description of the COLD Database and detailed picture of the sub-dataset COLD Freiburg is provided.

2.1 Literature Review

Computer Vision has become one of the most active research area in Artificial Intelligence, thus, a lot of work is continuously being done in this field. Here we will report few research findings about visual indoor place recognition. It is important to note that techniques good at indoor place recognition are not necessarily good at outdoor place recognition. M. M. Ullah et al. [3] presented a new database consisting of image sequences of several rooms under dynamic changes, acquired in three different labs in Europe. They assessed the new database with an appearance-based algorithm that combines local features with SVM through an ad-hoc kernel. Toori et.al. [4] in their study, demonstrated that matching of test image with larger changes in appearance and the database image is much easier when both viewed at approximately the same viewpoint. They implemented their idea by synthesizing virtual views on a densely sampled grid

on the map. To handle the large amount of synthesized data – as much as nine times more images than in the original street-view – they used the compact VLAD encoding of local image descriptors and they represented their training images using densely sampled local gradient based descriptors SIFT across multiple scales. They also introduced a new challenging dataset of 1,125 images of Tokyo that contain major changes in illumination (day, sunset, night) as well as structural changes in the scene. They showed that the proposed approach appreciably outperforms other large-scale place recognition techniques on the challenging data. Guillaume et. al. [5] in their research presented an algorithm for integrating the answers from different images. In this perspective, scenes are encoded with the help of and an unsupervised technique called Self-Organizing Map and then classified. They developed prototypes to form a visual dictionary which can roughly describe the environment. They used frequency of prototypes to train the system. This approach is a variant of Bag-of-Words with the major difference that they extracted different "words" taken from temporally ordered images and not from the same image. They evaluated their system with the COLD database. In a study conducted by Iwan Ulrich et al. [6], the authors proposed an appearance based topological localization method. They used nearest neighbor learning to classify color images and reported that their technique correctly classified 87% to 98% of the input color images. In another study, Benjamin et al. [7] proposed a method which recognized places with high accuracy in spite of perceptual aliasing and image variability. They proposed a bootstrap learning technique that used clustering for providing the pre requisites for map building, which in turn provided the labels required by a stronger supervised learning method (nearest neighbor) that can finally achieve greater performance. Axel Rotmann et al. [8] incorporated semantic information about the types of different places and used supervised learning approaches to label different locations using boosting. Their experimental results showed 87% recognition rate for different type of places. Svetlana et al. [9] introduced "spatial pyramids", which is partitioning the image into increasingly fine sub–regions and computing histograms of local features found inside each region. They found this technique better than "bags of features" approach. They performed experiments with their technique on different databases and reported higher accuracy. Francesco Orabona et al. [10] proposed an improvement to SVMs and called it as Online Independent SVM. Their technique exploited linear independence in the image feature space to incrementally keep the size of the learning machine small while maintaining the accuracy of a standard SVM. A. Pronobis et al. [11] in their study integrated multiple cues coming from different sensors and considered it as a fundamental capability of autonomous robots. For each cue, they trained a large margin classifier which generated a set of scores, indicating confidence of the decision. These scores were then used as input to SVM, which learnt weights for each cue. In another study conducted by Mayank Juneja et al. [12], a novel method for automatic learning of distinctive parts of objects or scenes has been proposed. They used exemplar SVMs for simultaneously learning a part model and detecting its occurrences in the training data. Entropy rank was used to measure the distinctiveness of parts. The outcome of their work was blocks that correspond to semantically meaningful parts. Their learned parts have shown to perform very well.

2.2 SIFT Feature Descriptor

Scale Invariant Feature Transform (SIFT) [20] as its name suggests, is a well known feature descriptor due to its invariant nature. For images that have lots of variations (such as scale, view point, light, aliasing), it accommodates all variations. It covers scale variations, and provides strong interest points. An orientation is allotted to each key point to achieve invariance to image rotation. It generates key points with same location and scale, but different directions that contribute to stability of matching. A 16x16 neighborhood around the interest points is taken. It is then divided into 16 sub-blocks of 4x4 size. For each sub-block, 8-bin orientation histogram is generated. So a total of 128 bin values are available. It is represented as a vector to form key point descriptor. Figure 2.1 shows a graphical view of an "interest point" and its neighborhood and 128 bins histogram [20].

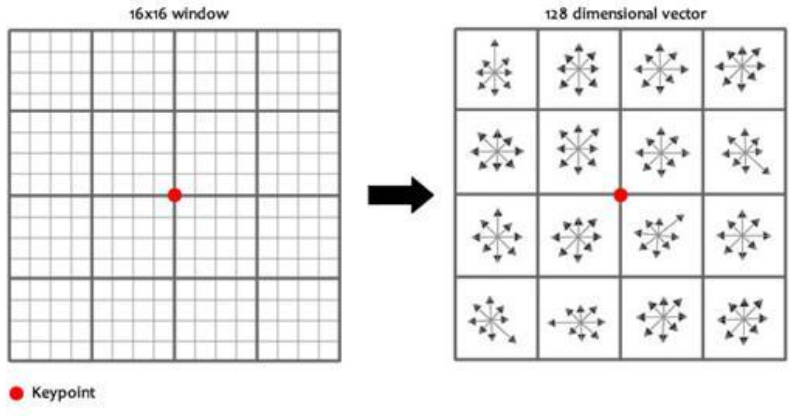


Figure 2.1 A graphical depiction of interest point and 128 bins

SIFT can robustly identify objects even among clutter and under partial occlusion, because the SIFT feature descriptor is invariant to uniform scaling, orientation, and partially invariant to affine distortion and illumination changes. These features are also robust to illumination, noise, and minor variations in viewpoint. In addition to these properties, they are highly distinguishing, comparatively simple to extract and allow correct object identification with less likelihood of mismatch. They are relatively easy to match against a large database of local features.

Since the images in our data i.e., COLD Dataset have lots of variations (such as scale, view point, light, aliasing) in them, therefore SIFT was more appropriate feature descriptor due to its invariant nature.

2.3 Bag of Features Representation

Bag of Visual Words (BoVW) analogy comes from the Bag of Features (BoF) technique in textual information retrieval, where an image is a sort of a document. BoVW is a popular

technique for image classification inspired by models used in natural language processing. However, in the context of visual classification, Bag of Features and Bag of Visual Words represent the same technique, and from now onwards, we will use Bag of Visual Words instead of Bag of Features. The following section briefly describes the technique, in general, for images.

2.3.1 Bag of Visual Words

This technique represents an image as an order less collection of local features, as shown in Figure 2.2. Features (e.g., SIFT) are extracted from images. A sample of features from the training images is then clustered to build a visual vocabulary. As the terminology suggests, "visual vocabulary" is a dictionary of patches of images. Different local patches of images are represented by feature descriptors. The purpose of clustering is to collect centroids (being the most central points) to represent millions (or billions) of local features, just as in a document a list of those words is prepared that are central to that document. Each centroid is a visual word. K-means is the mostly used clustering technique used in BOVW representation. Given an unseen test image, features are detected and assigned to their nearest visual word (cluster center) from the visual vocabulary. When determining the nearest visual word from visual vocabulary, common choices are the Manhattan (L1), Euclidean (L2), or Mahalanobis distances. After matching, occurrences of each visual word are counted to make a histogram for that image. Consequently, an image is reduced to the set of visual words it contains, represented as a histogram. The histogram is subsequently normalized to unit length.

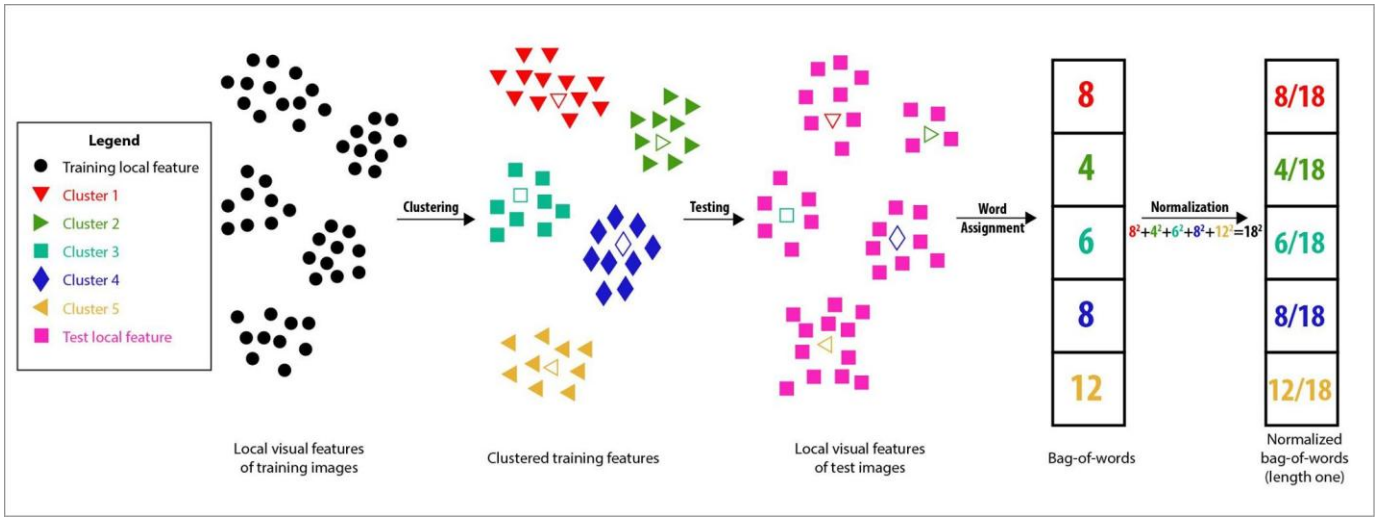


Figure 2.2 Graphical summary of BOVW

2.4 Convolution Neural Networks

Convolutional Neural Networks consists of multiple layers of small neurons that inspect smaller portions of the input images, which are then tiled so that they overlap to get a better

representation of the original image. This process is repeated for every layer of neurons. CNNs include local or global pooling layers, which merge the outputs of neurons. They also consist of various combinations of convolutional layers and fully connected layers. One major lead of CNNs is that they use shared weight in convolutional layers, which means that the same filter (weights bank) is used for each pixel in the layer; that is how it reduces required memory size and also it improves performance as compare to other image classification techniques, convolutional neural networks requires less pre-processing. The network itself is responsible for learning the filters instead of hand-engineered filters. The less dependency on prior-knowledge and the existence of difficult to design hand-engineered features is a major advantage for CNNs. Figure 2.3 [21] shows a very simple diagram of a convolutional neural network. The diagram describes well the concept of convolutional and fully connected layers.

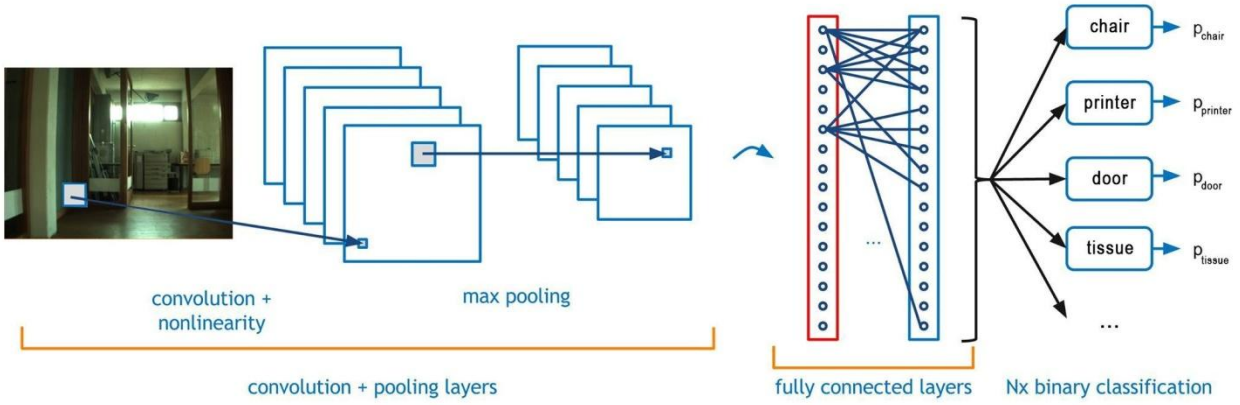


Figure 2.3 A Convolution Neural Network

2.5 SVM Classification

Support Vector Machines (SVM) is a supervised learning model that analyzes data and learns patterns to recognize unseen data. Given a set of training labeled data instances, an SVM training algorithm constructs a model that classifies new examples label by using the learned model. Initially SVMs were used for only linear classification, but later on, kernels were defined to map the non-linear data into high dimensional space so that non-separable data become separable, so that a clear boundary can be drawn between the data. An SVM model defines a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class, since in general the larger the margin the lower the generalization error of the classifier. Classification process completes in two phases, i.e. training and testing

- **Training**

Training is the process of learning a model based on seen data instances that are known to belong to specified classes. A trained linear SVM model is based on two major values of w and b, where w represents “weights” and b represents “biasness”. There values are

used to compute a hyperplane that separates positive and negative examples creating maximum margin between them.
Equation of Hyperplane is: $f(x) = w^T x + b$
such that $w^T x + b = +1$ and $w^T x + b = -1$ for the positive and negative support vectors respectively.

- **Testing**

Testing is the process of taking a classifier built with seen and labeled training examples and running it on unknown/unseen test examples to determine their category/class. Classifier computes "test scores" of test images and then verifies the value of test scores as:

$$w^T x + b > 0 \text{ and } w^T x + b < 0$$

If the value is greater than zero for an image, it assigns +1 to that image. And if the value is less than zero, it assigns -1 to that image. That is how it classifies images into positive and negative respectively.

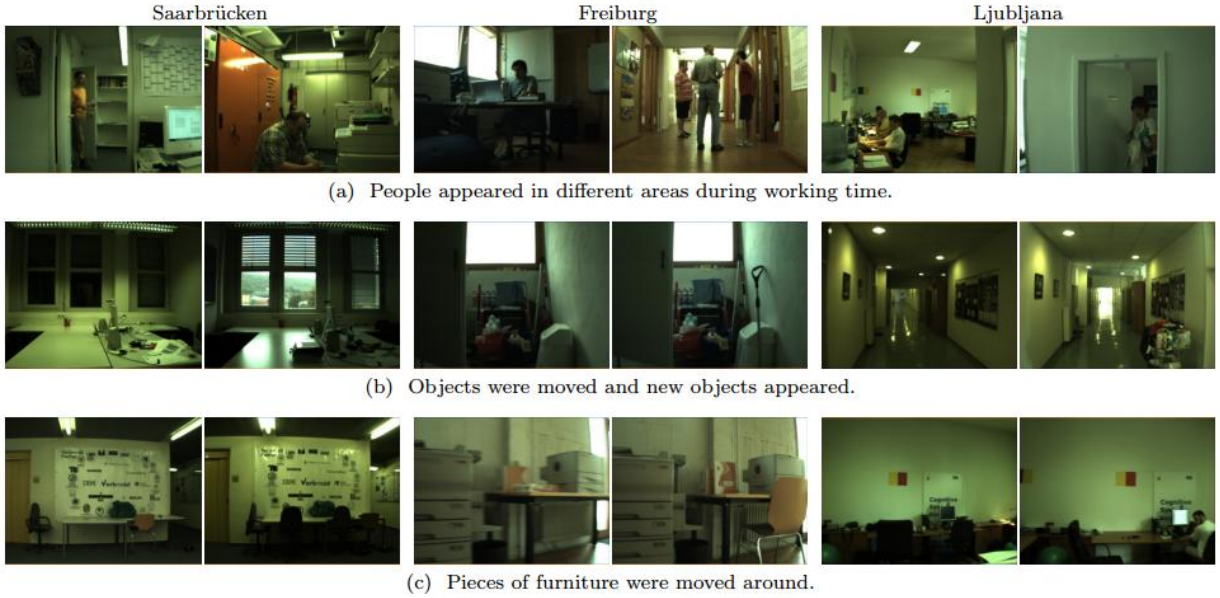
Support Vector Machine Classifier offers many advantages over the others. A prominent advantage of SVMs is that the solution to an SVM is global and unique. They have simple geometric interpretation. The computational complexity of SVMs does not depend on the dimensionality of the input space. And also SVMs are less prone to over fitting.

2.6 COLD Dataset

The acronym COLD [14] stands for COsy Localization Database. It provides a large-scale, flexible testing environment to evaluate vision-based localization systems that seek to work on mobile platforms in realistic settings. There are three separate datasets in COLD Database, that are acquired at three different indoor laboratory environments located in three different European cities: the Visual Cognitive Systems Laboratory at the University of Ljubljana, Slovenia; the Autonomous Intelligent Systems Laboratory at the University of Freiburg, Germany; and the Language Technology Laboratory at the German Research Center for Artificial Intelligence in Saarbrucken, Germany.

COLD database is a comprehensive set of visual data that can be employed to benchmark localization as well as place recognition and categorization algorithms. This database consists of images in which variations of light (day and night, artificial light on and off) and human activity (furniture moved around, objects being taken in/out of drawers and etc.) has been captured. These changes are known as dynamic changes because these can be observed only when considering the indoor environments over a period of time of at least several hours. Illumination variations have a very significant influence on the visual appearance of the indoor environment. In order to capture this variability, image sequences were acquired under three different illumination and weather conditions: cloudy weather, sunny weather and night. • People

appeared in the room (Figure 2.3a) • Objects were moved and taken in/out of the cupboards/drawers (Figure 2.3b). • Pieces of furniture, such as chairs, were moved around (Figure 2.3c). The COLD database was acquired in several environments consisting of rooms that can be assigned to the same functional category. As a result, large within-category variability can be observed in the images.



(a) People appeared in different areas during working time.

(b) Objects were moved and new objects appeared.

(c) Pieces of furniture were moved around.

Figure 2.4 Different type of variations introduced in COLD database [11]

2.6.1 COLD Freiburg

COLD Freiburg is sub data set of COLD database. And it is named after the laboratory where the images have been captured, the Autonomous Intelligent Systems Laboratory at the University of Freiburg, Germany. This sub-dataset consists of two parts, Part A and Part B .These parts are then sub-divided into Standard and Extended Sequences. Note that standard path refers to the path where the robot was driven to the rooms that are found in most labs, whereas extended path refers to the path where the robot was driven to all the available rooms. Part A Path 1 is Standard path and consists of 10 different sequences captured under different illumination conditions, similar is true for Part A Path 2 except that it is Extended path. Part B Path 3 consists of 6 different sequences. Images for these sequences are captured on standard path. Night sequences are not captured for Part B Path 3. The detailed description of sequences of COLD Freiburg under various illumination condition, and number of images in each sequence is given in Table 2.1 and Figure 2.4 shows hierarchy of COLD Freiburg Dataset to get a quick view of different parts and sequences. Note that the sequences in the database consist of both perspective and omni-directional images. This research and all the experiments are based only on perspective images, and not omni-directional images. In Table 2.2, detailed picture of categories in each sequence of Part A Path 1(Standard) is given. There are five categories

in each sequence of this path. Number of negative and positive images is shown for each category. Negative and positive images are based on the fact that we adopted one-versus-all strategy for multi-class classification. That's how we converted it into binary-class classification. For example in Cloudy 1 sequence of Part A Path 1, there are 1459 total images. In total images, there are 741 images of corridor, so for classification corridor is positive and rest of the images would be considered as negative. Same is true for all categories. Table 2.3 provides the same full description of categories in each sequence of Part A Path 2(Extended) and Table 2.4 have the details of Part B Path 3 (Standard). In Figure 2.5 examples of perspective camera images in the database are shown that present the interior of some of the rooms in Freiburg Lab [14].

	COLD Freiburg		
	Standard		**Extended**
	Part A	**Part B**	**Part A**
	Path 1	Path 3	Path 2
Cloudy	1 (1459)	1 (2236)	1 (2146)
	2 (1832)	2 (2008)	2 (2595)
	3 (1672)	3 (1898)	3 (2778)
Night	1 (1911)		1 (2876)
	2 (1582)		2 (2707)
	3 (1703)		3 (2896)
Sunny	1 (1598)	1 (1754)	1 (2319)
	2 (1514)	2 (1797)	2 (2114)
	3 (1551)	3 (2225)	3 (2231)
	4 (1777)		4 (2807)
Total Sequences	10	6	10

Table 2.1 Detailed description of sequences in COLD Freiburg

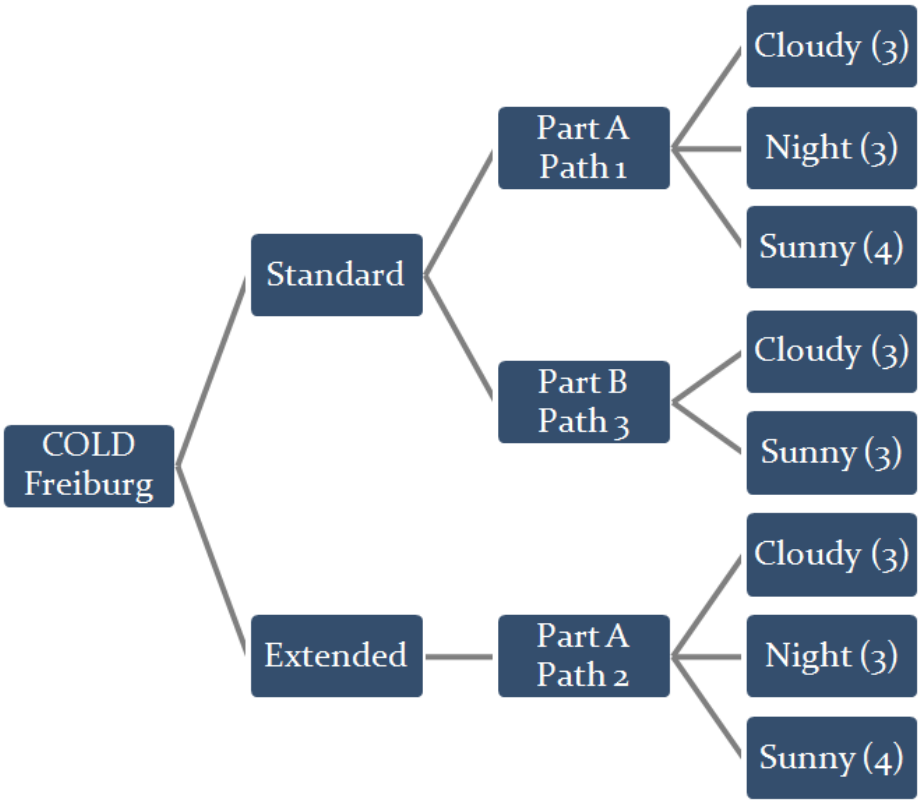


Figure 2.5 Hierarchy of COLD Freiburg

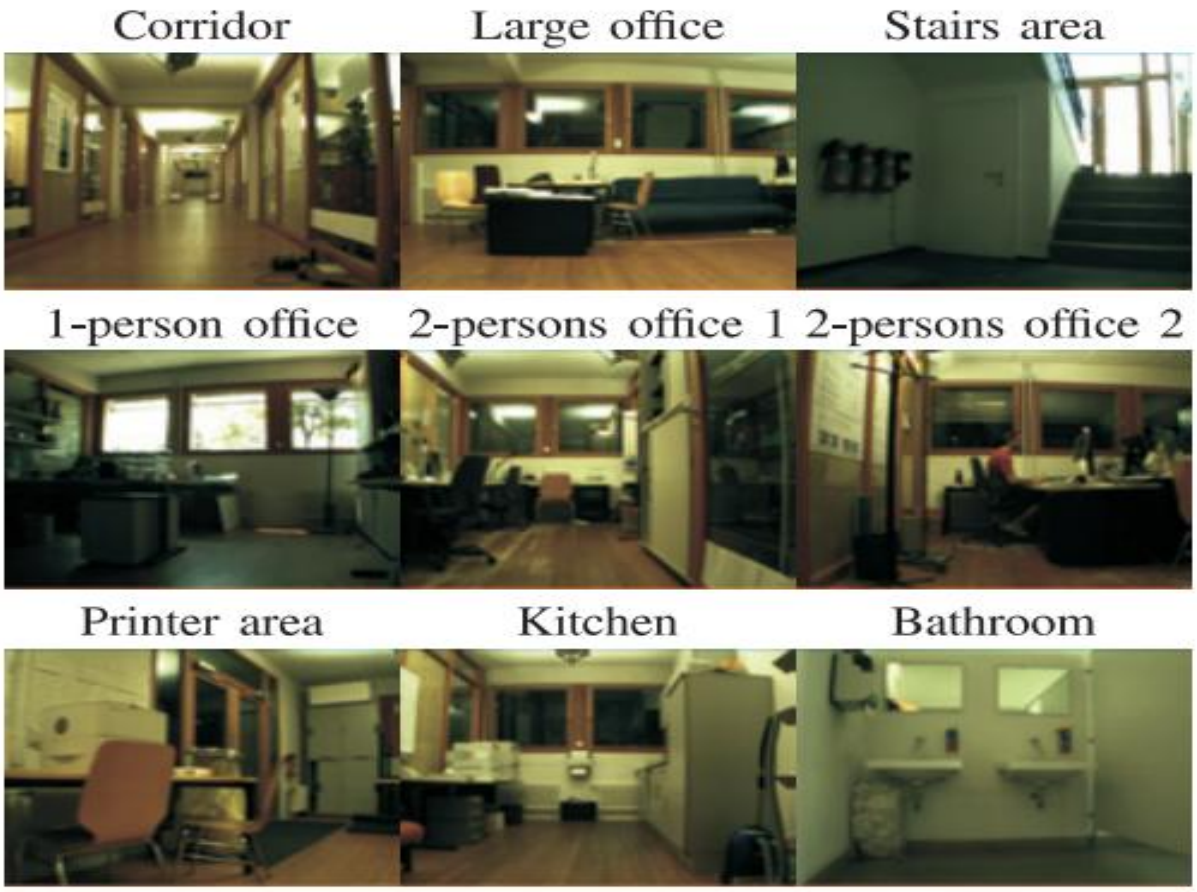


Figure 2.6 Images in the COLD Freiburg presenting the interior of rooms

Part A Path 1

	Cloudy 1 (1459)		Cloudy 2 (1832)		Cloudy 3 (1672)	
Classes	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Corridor	741	718	974	858	831	841
Printer's Area	202	1257	247	1585	243	1429
2 Person's Office	157	1302	218	1614	234	1438
Stair Area	153	1306	135	1697	145	1527
Toilet	206	1253	258	1574	219	1453

	Night 1 (1911)		Night 2 (1582)		Night 3 (1703)	
Classes	**Positive**	**Negative**	**Positive**	**Negative**	**Negative**	**Negative**
Corridor	949	962	786	796	802	901
Printer's Area	275	1636	219	1363	222	1481
2 Person's Office	227	1684	192	1390	230	1473
Stair Area	208	1703	201	1381	240	1463
Toilet	252	1659	184	1398	209	1494

	Sunny 1 (1598)		Sunny 2 (1514)		Sunny 3 (1551)		Sunny 4 (1777)	
Classes	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Corridor	732	866	791	723	852	699	829	948
Printer's Area	238	1360	203	1311	187	1364	357	1420
2 Person's Office	220	1378	154	1360	141	1410	230	1547
Stair Area	170	1428	177	1337	153	1398	152	1625
Toilet	238	1360	189	1325	218	1333	209	1568

Table 2.2 Categories and number of Positive/Negative images in each sequence of Part A Path 1

Part A Path 2

	Cloudy 1 (2146)		Cloudy 2 (2595)		Cloudy 3 (2778)	
Classes	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Corridor	1017	1129	1040	1555	1183	1595
Printer's Area	227	1919	222	2373	284	2494
Large Office	146	2000	177	2418	132	2646
2 Person's Office 1	187	1959	230	2365	233	2545
2 Person's Office 2	115	2031	135	2460	158	2620
1 Person's Office	138	2008	155	2440	218	2560
KT	---		254	2341	229	2549
Stair Area	138	2008	133	2462	151	2627
Toilet	178	1968	249	2346	190	2588

	Night 1 (2876)		Night 2 (2707)		Night 3 (2896)	
Classes	**Positive**	**Negative**	**Positive**	**Negative**	**Negative**	**Negative**
Corridor	1110	1766	1114	1593	1005	1891
Printer's Area	313	2563	241	2466	487	2409
Large Offoce	143	2733	121	2586	115	2781
2 Person's Office 1	296	2580	215	2492	218	2678
2 Person's Office 2	127	2749	168	2539	135	2761
1 Person's Office	138	2738	168	2539	177	2719
KT	272	2604	270	2437	272	2624
Stair Area	220	2656	198	2509	198	2698
Toilet	257	2619	212	2495	289	2607

	Sunny 1 (2319)		Sunny 2 (2114)		Sunny 3 (2231)		Sunny 4 (2807)	
Classes	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Corridor	957	1362	793	1321	965	1266	1139	1668
Printer's Area	178	2141	191	1923	207	2024	309	2498
Large Office	170	2149	102	2012	120	2111	182	2625
2 Person's Office 1	151	2168	187	1927	165	2065	222	2585
2 Person's Office 2	120	2199	109	2005	143	2088	116	2691
1 Person's Office	156	2163	123	1991	162	2069	171	2636
KT	196	2123	213	1901	173	2058	244	2563
Stair Area	173	2146	180	1934	124	2107	172	2635
Toilet	218	2101	216	1898	172	2059	252	2555

Table 2.3 Categories and number of Positive/Negative images in each sequence of Part A Path 2

Part B Path 3

	Cloudy 1 (2236)		Cloudy 2 (2008)		Cloudy 3 (1898)	
Classes	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Corridor	760	1476	733	1275	731	1167
2 Person's Office	549	1687	437	1571	372	1526
1 Person's Office	400	1836	395	1613	342	1556
Stair Area	256	1980	197	1811	195	1703
Toilet	271	1965	246	1762	258	1640

	Sunny 1 (1754)		Sunny 2 (1797)		Sunny 3 (2225)	
Classes	**Positive**	**Negative**	**Positive**	**Negative**	**Negative**	**Negative**
Corridor	711	1043	670	1127	822	1403
2 Person's Office	306	1448	376	1421	322	1903
1 Person's Office	384	1370	342	1455	507	1718
Stair Area	180	1574	164	1633	249	1976
Toilet	173	1581	245	1552	325	1900

Table 2.4 Categories and number of Positive/Negative images in each sequence of Part B Path 3

Chapter 3

3. Our Approach

We discussed general concepts of Bag of Visual Words Technique and Convolution Neural Networks in chapter 2. We gave a brief description of how these techniques work and what are the parameters that we need to set. Now, in this chapter we will explain that how we applied these concepts to our task. All the technical details of method and whole procedure will be discussed in detail.

3.1 Bag of Visual Words Representation

Let's explain how we represented images as Bag of Visual Words. Step by step detail with figures is given in this section.

3.1.1 SIFT Feature Extraction

As discussed in chapter 2, SIFT is invariant in its nature therefore, this made SIFT best suitable for our data. Fig 3.1 shows an image from COLD Freiburg with random 50 SIFT keypoints. The figure shows only 50 random SIFT keypoints, but there are varying number of SIFT keypoints in each image and the number is always greater than 1000, as we computed descriptors for densely sampled keypoints..

VLFeat library in MATLAB was used that provides built in function for extraction of SIFT Features. A built in function that generates SIFT features is given as under:

```
[F,D] = vl_sift(I) ;
```

computes the SIFT frames (keypoints) F of the image I. The matrix F has a column for each frame. Each column of F is a feature frame and has the format [X; Y; S; TH], where X, Y the (fractional) center of the frame, S is is the scale and TH is the orientation (in radians). And matrix D represents 128 bin values. So that's how every SIFT feature is a vector of dimension 132. Each image has a varying number of SIFT features.

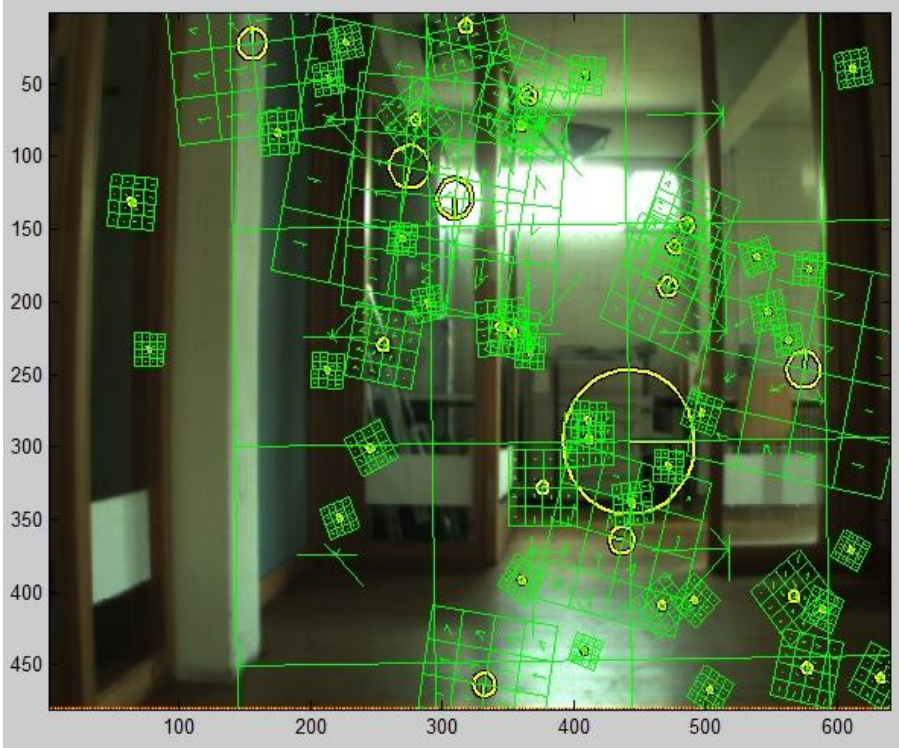


Figure 3.1 Random SIFT features of an image

3.1.2 Visual Vocabulary Generation

We got a huge number of images in our dataset and there were varying number of SIFT features for each image. As explained above we got dense interest points so we got on average more than 1000 SIFT features for each image. That sums up a big number for one sequence. We decided to pick randomly one hundred thousand features from each sequence (detail of number of images in each sequence in COLD Freiburg is given in chapter 2.), for purpose of clustering. Clustering served for generating ***Visual Words*** that collectively made a ***Visual Vocabulary***.

When we came to the step of clustering, it was a big deal to come up with an appropriate value of "k" as in k-means clustering algorithm. A small value of "k" can't represent all the points fairly and a large value of "k" can generate a risk of over fitting. Hence we ended up with a value k = 500, as it is neither too small nor too large for hundred thousand points to cluster. We generated one Visual Vocabulary for each sequence, as we wanted to test all sequences against every one sequence. The number of the clusters is the visual vocabulary size. Dimension of Visual vocabulary is 500-by-128; as there were 500 cluster centers i.e., Visual words. Next step was quantization of features with visual words.

3.1.3 Quantization Process

SIFT features of an image was then mapped to the index of the nearest visual word in the visual vocabulary of training sequence. Mapping referred to calculating distance using Square Euclidean method between each feature vector and every visual word and then selecting one visual word with minimum distance. That was how all features for an image were mapped to nearest visual word. That provided a representation of an image in form of visual words. Next step in quantization process is to count the number of each visual word i.e. the frequency of each visual word. Noted frequency of each visual word in an image gave us frequency histogram of that image. And same process went on for all images in one sequence. A graphical representation of the feature vectors and their dimensions can be viewed in Figure 3.2. In this way, we computed bag of visual words for each image in a given sequence. Next step was the classification task.

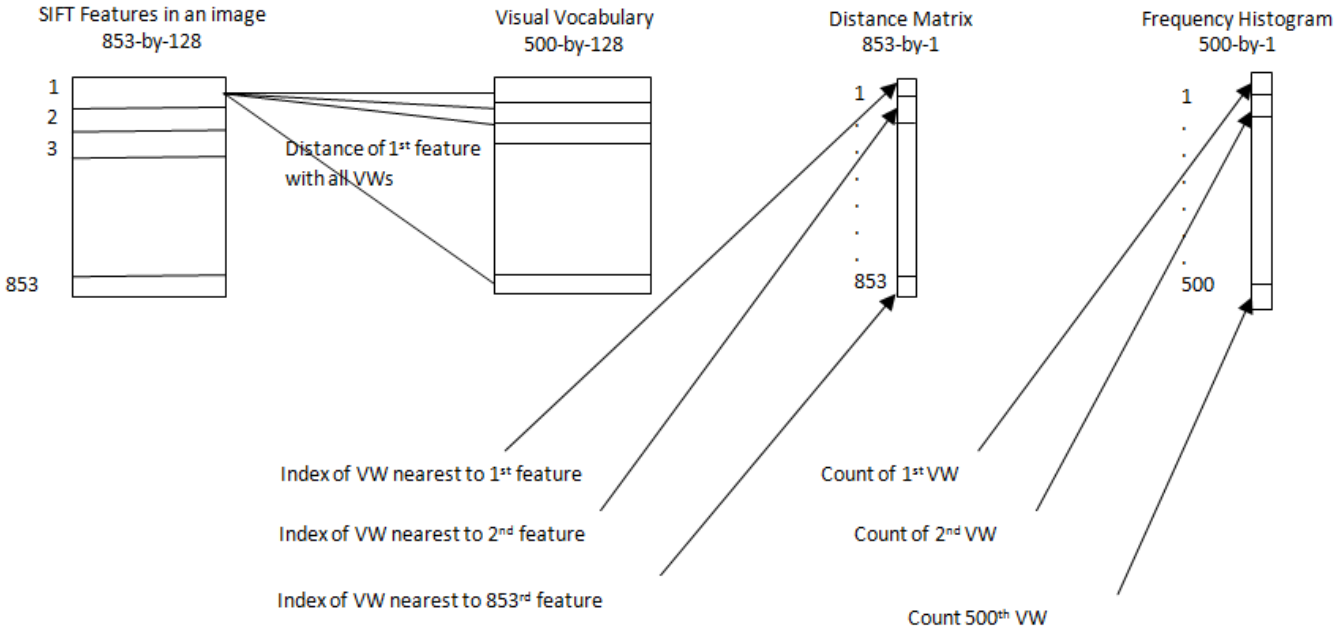


Figure 3.2 Feature vectors in BOVW

3.2 CNN Features

The convolution neural network [15] that we used in our task has eight layers with weights; out of which the first five are convolutional and the remaining three are fully connected. The output of the last fully-connected layer is then provided to a softmax layer which produces a probability distribution over the class labels. Fig 3.3 shows a convolution neural network FC-7 Layer and softmax layer. This CNN is basically designed by Alex et.al in their research [15]; they trained this architecture on ImageNet dataset. We used these pre-trained CNNs to extract FC-7 layer features and SoftMax layer features from images of COLD dataset.

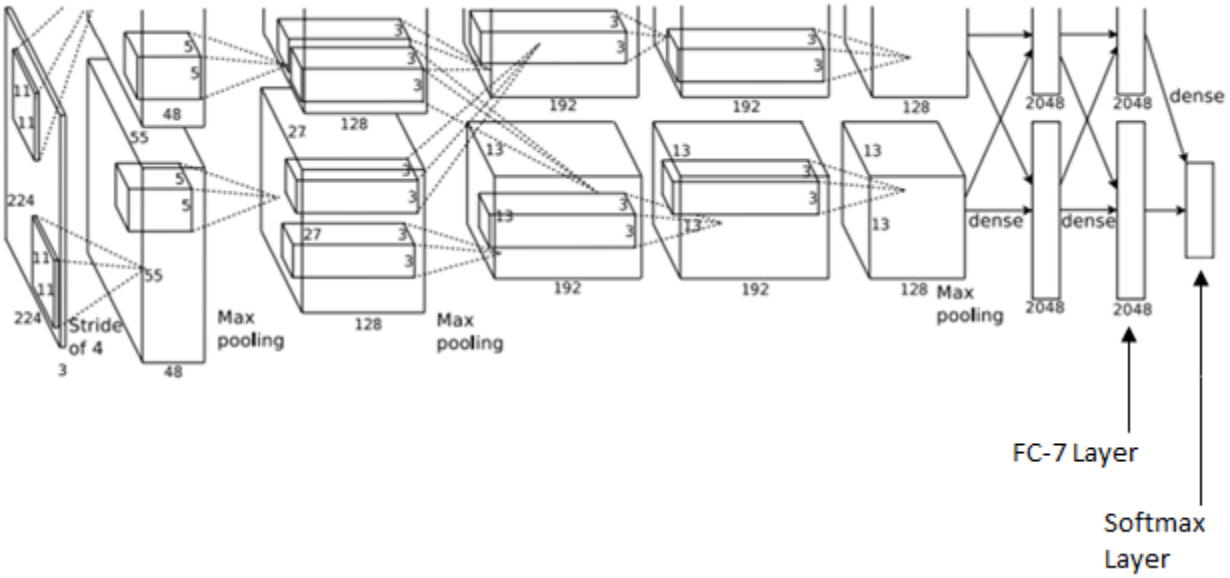


Figure 3.3 A Convolution Neural Network

3.2.1 ImageNET-CNN

ImageNet [16] is an image dataset of over 15 million labeled high-resolution images belonging to roughly 22,000 categories. It's an object centric database.

Figure 3.4 Example of images in ImageNet Dataset

Figure 3.4 shows some example images from ImageNet Dataset. We can see that ImageNet dataset consists of different categories of objects that are also found in COLD dataset, like tables, chairs, computers, printers, toiletries, and kitchen utensils etc.

ImageNet-CNN: Convolution neural networks trained on ImageNet dataset is termed as ImageNet-CNN.

ImageNet-CNN-FC-7 Features: 4096-by-1

ImageNet-CNN-Softmax Features: 1000-by-1

3.2.2 Places-CNN

Places205 [17] is an image dataset which contains 2,448,873 images from 205 scene categories.

Figure 3.5 Example of images in Places205 Dataset

Figure 3.5 shows some example images from Places Dataset. We can see that Places dataset consists of different categories of places that are also found in COLD dataset, like office, kitchen, bathroom, stairs, corridors etc.

Places-CNNs: Convolution neural networks trained on Places dataset is termed as Places-CNNs

Places- CNN-FC-7 Features: 4096

Places- CNN-Softmax Features: 205

3.2.3 Hybrid-CNN

Hybrid refers to the combination of ImageNet and Places dataset. It combines the images from these two datasets.

Hybrid-CNNs: Convolution neural networks trained on Hybrid dataset is termed as Hybrid-CNN.

Hybrid-CNN-FC-7 Features: 4096

Hybrid-CNN-Softmax Features: 1183

As explained above, these datasets are huge and contains objects and scenes from almost all categories found in real world. We were motivated to use pre-trained CNNs because of the fact that Places, ImageNet dataset and COLD Database have few categories that are in common. It inspired us to use pre- trained CNNs to experiment with the FC-7 layer Features and SoftMax Features extracted from COLD images.

3.3 Feature Combination

Our approach was to simply concatenate the deep features extracted from COLD Freiburg images using pre-trained CNNs and to experiment with these concatenated channels. Let's see how these deep features were concatenated and what there dimensions were after concatenation.

3.3.1 CNN-FC-7 Channel

The name "FC-7 channel" was adopted due to the fact that this channel was constructed using deep features extracted from the 7th layer of pre-trained-CNNs. This layer gives a vector of deep features of dimension 4096. As we have three types of pre-trained CNNs, we used all of these to extract FC-7 layer deep features and concatenated them to get a single vector of dimension 12288. Figure 3.6 shows how we concatenated and what the dimensions were after concatenation.

- Hybrid-CNN- FC-7 Features 4096-by-1
- ImageNet-CNN- FC-7 Features 4096-by-1
- Places-CNN- FC-7 Features 4096-by-1

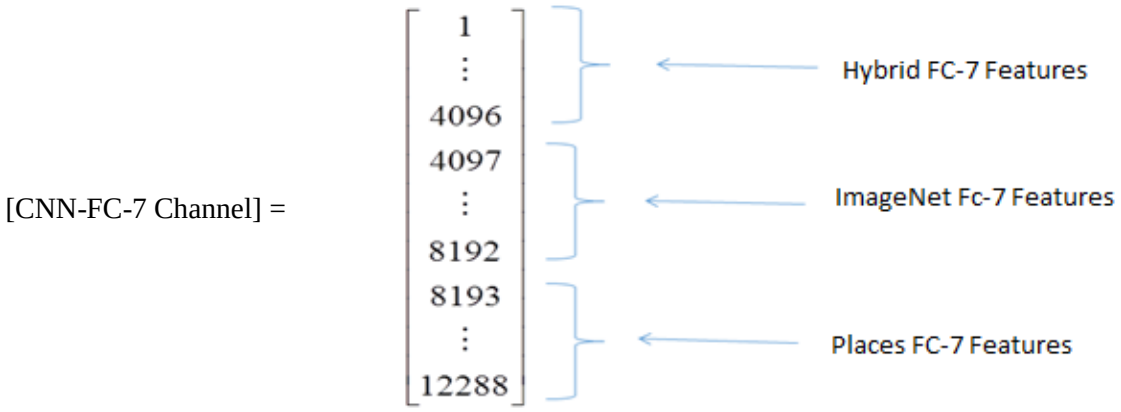


Figure 3.6 CNN-FC-7 Channel

So that is how FC-7 Channel was constructed by concatenating deep features of COLD Freiburg images obtained from three different pre-trained CNNs.

3.3.2 CNN-SoftMax Channel

The name "CNN-SoftMax channel" was adopted due to the fact that this channel was constructed using deep features extracted from the softmax layer of pre-trained-CNNs. This layer gives a vector of class probabilities of different objects. As we have three types of pre-trained CNNs, we used all of these to extract softmax layer deep features and concatenated them to get a single vector. Note that in the case of softmax layer deep features, the dimension of vectors, obtained from three different pre-trained CNNs, is different. Dimensions of these vectors are given as under:

- Hybrid-CNN- Features 1183-by-1
- ImageNet-CNN- Features 1000-by-1
- Places-CNN- Features 205-by-1

Figure 3.7 shows how we concatenated and what the dimensions were after concatenation.

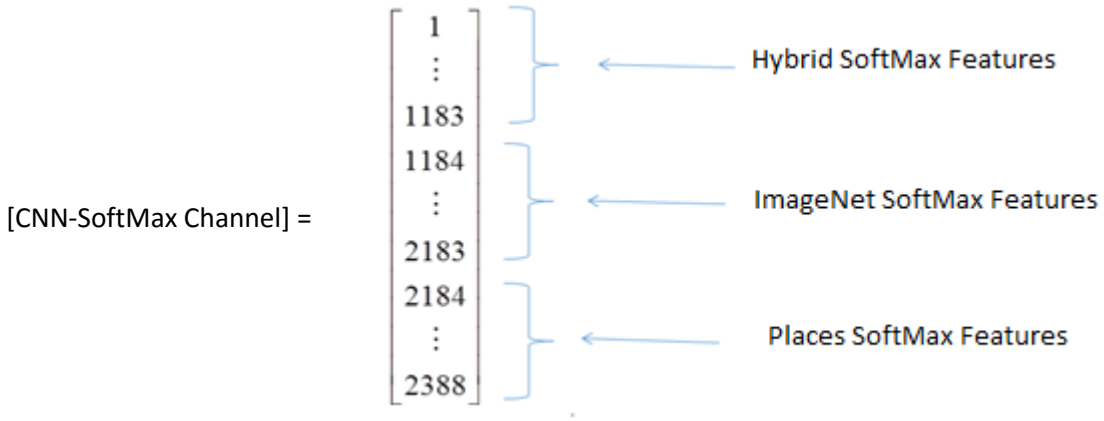


Figure 3.7 CNN-SoftMax Channel

So that is how SoftMax Channel was constructed by concatenating deep features of COLD Freiburg images obtained from three different pre-trained CNNs.

3.3.3 Final Combination

Our last channel is "BOVW+ Deep Features" (BoVW+CNN-FC-7+CNN-Softmax), as the name suggests, there must be three channels that were concatenated to construct a channel. Yes, that's true! Three different channels were concatenated to get a single channel. These three channels were "BOVW Channel", "FC-7 Channel" and "SoftMax Channel". Let's see in Fig 5 the concatenation of these channels and the dimension of final vector.

- BOVW Channel 500-by-1
- CNN- FC-7 Channel 12288-by-1

- CNN- SoftMax Channel 2388-by-1

Figure 3.8 show we concatenated and what the dimensions were after concatenation.

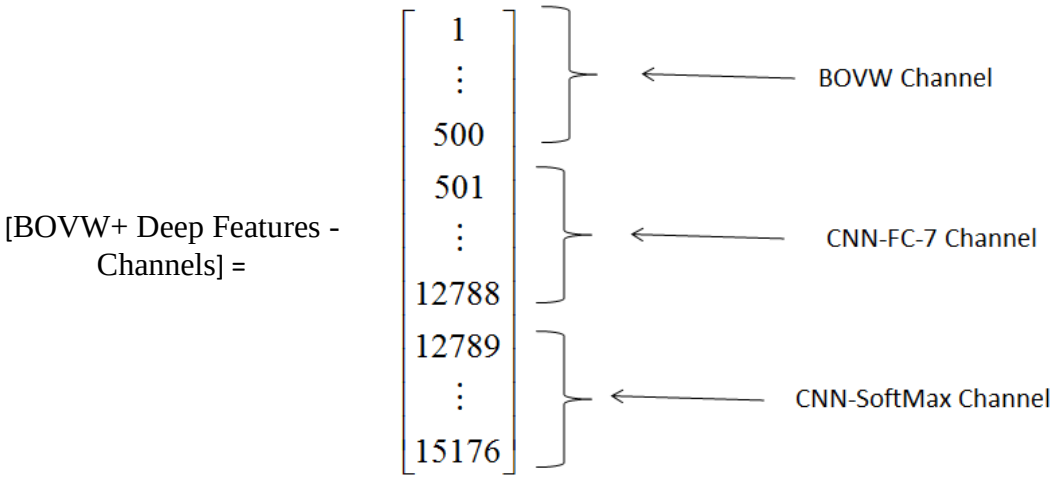


Figure 3.8 Final Combination "BOVW+ Deep Features"

So that is how "BOVW+ Deep Features -Channels" was constructed by concatenating three different channels. After getting all these concatenated channels, positive and negative vectors according to every category were separated to be used in classification process.

3.4 SVM Classification

This step involves identification of unseen test data on the basis of previously learned knowledge from the training data. We considered one sequence as a training sequence and remaining other sequences as test sequences every time. In that way, training was performed on each sequence and testing against all the others. Our data is multi-class labeled high dimensional data, so we needed to select a supervised learning classifier. Tuning of its parameters is easy, i.e., the selection of kernel, the kernel's parameters, and soft margin parameter C. It is the regularization parameter of the SVM. We experimented with different values of C and checked using cross validation set of sequences, and the value with best cross-validation accuracy was picked, where C =10. Since our data has multiple classes, therefore we adopted strategy of one-versus-all to reduce the single multiclass problem into multiple binary classification problems. That means we considered one class positive at one time and all other classes negative, that's how we did it with all classes and ended up doing multiple binary classifications. Suppose, in a sequence we have 1459 images in total, 202 images belong to category "PA" i.e. Printer's Area will be considered as positive, and remaining all images belonging to any category will be considered as negative. We measured performance of the classifier by computing its accuracy by the formula given as under:

Accuracy = (TP + TN) / (TP + TN + FP + FN)
Where TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

These four values were computed using test scores obtained from the trained model. Few parameters of SVM classifier are:

- **Classifier – Linear Support Vector Machines**

In practice, the linear kernel exhibits good performance when the number of observations is larger than number of samples (e.g. there is no need to map to an even higher dimensional feature space), since we have larger feature space so we used linear classifier.

- **Kernel – Hellinger**

We didn't compute kernel values but we explicitly computed the feature map, so that the classifier remains linear (in the new feature space).

$$k(h, h') = \sum_i \sqrt{h(i)h'(i)} \text{ (compared to a linear kernel } \sum_i h(i)h'(i))$$

Where h and h' are histograms in case of BOVW, FC-7 layer feature vectors in case of CNN-FC-Channel, and Softmax layer feature vectors in case of CNN-SoftMax-Channel.

- **Encoding Technique – Bag of Visual Words/ Pre-trained CNNs**

- **Normalization – L2**

L2-norm is also known as least squares. It is basically minimizing the sum of the square of the differences between the target value and the estimated values.

That was our approach to experiment with. In the next chapter we will present all the experiments that we did using Bag of Visual Words Channel, CNN-FC-7 Channel, CNN-SoftMax Channel and the combination of these three channels, BOVW+ Deep Features - Channels. We will show, in the next chapter, all the results and comparison of accuracy that we obtained after experimentation.

Chapter 4

4. Experimental Evaluation

In this chapter, we will describe in detail the experiments we have carried on COLD Freiburg (Standard and Extended) image sequences. We assessed the COLD database with two series of experiments. In the first series of experiments, we used Bag of features technique on all sequences to generate baseline results on COLD Freiburg Data. For each set, training and testing was always performed on different sequences captured in the same lab. Training was done on one illumination condition, and then tested on sequences captured under different illumination conditions, and after some time. With these experiments we were able to compare accuracy of classification between baseline results and BOVW+ Deep Features -Channels approach results. Also we compared our results presented in [3] and [5]

In the next series of experiments we used pre-trained CNNs and got deep features. We had two types of deep features as explained in Chapter 3. Deep features obtained from 7th layer of CNNs that gave 4096 dimension vector for each image are known as FC-7 features, and the other type of deep features are those obtained from softmax layer of CNNs that provided us with class probabilities of different objects, we call these features, SoftMax Features. Dimension of these features varies (for details read Chapter 3).

4.1 Experimental Protocol

Experiments were repeated on all possible combinations/permutations of training and testing sequences as we know that COLD Freiburg is such a challenging dataset because of dynamic, geographic and illumination variations across different images/sequences. Different permutations of sequences of Part A Path 1 (Standard) and Part A Path 2 (Extended) are shown in Table 4.1 and different permutations of sequences of Part B Path 3 (Standard) are shown in Table 4.2. For all the experiments, we used vlfeat [19] library, and we determined the SVM and kernel parameters via cross-validation. The cross-validation was performed using four sequences, i.e., Cloudy 1 was used for training and Cloudy2, Night 1 and Sunny 1 were used for testing. The values we obtained from cross validation set experiments were then used for the experiments on whole COLD Freiburg dataset. Using all the possible permutations of the training and test sequences, the experiments were conducted several times and the average results are stated.

Train Sequence	Test Sequence
Cloudy 1	Cloudy 2, Cloudy 3, Night 1, Night 2, Night 3, Sunny 1, Sunny 2, Sunny 3, Sunny 4
Cloudy 2	Cloudy 1, Cloudy 3, Night 1, Night 2, Night 3, Sunny 1, Sunny 2, Sunny 3, Sunny 4
Cloudy 3	Cloudy 1, Cloudy 2, Night 1, Night 2, Night 3, Sunny 1, Sunny 2, Sunny 3, Sunny 4
Night 1	Cloudy 1, Cloudy 2, Cloudy 3, Night 2, Night 3, Sunny 1, Sunny 2, Sunny 3, Sunny 4
Night 2	Cloudy 1, Cloudy 2, Cloudy 3, Night 1,Night 3, Sunny 1, Sunny 2, Sunny 3, Sunny 4
Night 3	Cloudy 1, Cloudy 2, Cloudy 3, Night 1, Night 2, Sunny 1, Sunny 2, Sunny 3, Sunny 4
Sunny 1	Cloudy 1, Cloudy 2, Cloudy 3, Night 1, Night 2, Night 3, Sunny 2, Sunny 3, Sunny 4
Sunny 2	Cloudy 1, Cloudy 2, Cloudy 3, Night 1, Night 2, Night 3, Sunny 1, Sunny 3, Sunny 4
Sunny 3	Cloudy 1, Cloudy 2, Cloudy 3, Night 1, Night 2, Night 3, Sunny 1, Sunny 2, Sunny 4
Sunny 4	Cloudy 1, Cloudy 2, Cloudy 3, Night 1, Night 2, Night 3, Sunny 1, Sunny 2, Sunny 3

Table 4.1 Permutations of Sequences for training/testing for Part A Path 1(Standard) and Path 2 (Extended)

Train Sequence	Test Sequence
Cloudy 1	Cloudy 2, Cloudy 3, Sunny 1, Sunny 2, Sunny 3
Cloudy 2	Cloudy 1, Cloudy 3, Sunny 1, Sunny 2, Sunny 3
Cloudy 3	Cloudy 1, Cloudy 2, Sunny 1, Sunny 2, Sunny 3
Sunny 1	Cloudy 1, Cloudy 2, Cloudy 3, Sunny 2, Sunny 3
Sunny 2	Cloudy 1, Cloudy 2, Cloudy 3, Sunny 1, Sunny 3
Sunny 3	Cloudy 1, Cloudy 2, Cloudy 3, Sunny 1, Sunny 2

Table 4.2 Permutations of Sequences for training/testing for Part B Path 3(Standard)

4.1.1 Cross Validation Set

Cross- Validation set, on which we conducted our initial experiments to come up with appropriate values of different parameters, consists of four image sequences as following:

- **Train Data – Cloudy 1**
- **Test Data – Cloudy 2, Night 1, Sunny 1**

4.2 Results on Cross-Validation Set

This section presents all results that we executed on our cross validation set. In Table 4.3, results of classification for five different techniques are shown. Bag of Visual Words technique which is our baseline, results of deep features extracted from COLD Freiburg images obtained from 7th layer of pre-trained CNNs (pre-trained on ImageNet, Places, Hybrid Datasets) are shown. Last column of the table shows results of classification that were obtained by the concatenation of 7th layer's deep features obtained from three different pre-trained CNNs. Note that the concatenated channel is known as FC-7 Channel It is apparent from the results that concatenation of deep features provided us with accuracy that is best among all techniques.

		BoVW	Hybrid_CNN	ImageNet _CNN	Places_CNN	FC-7 Channel
			FC-7 Layer	FC-7 Layer	FC-7 Layer	
			4096-by-1	4096-by-1	4096-by-1	12288-by-1
Cloudy 2	**CR**	0.9247	0.9607	0.9634	0.9563	0.9689
	PA	0.9258	0.9629	0.9716	0.9531	0.9754
	2PO1	0.9394	0.9694	0.9623	0.9662	0.9743
	ST	0.9689	0.9885	0.9885	0.9858	0.9940
	TL	0.9814	0.9842	0.9913	0.9902	0.9918
Average		***0.9480***	***0.9731***	***0.9754***	***0.9703***	***0.9809***
Night 1	**CR**	0.8687	0.9178	0.9201	0.9095	0.9325
	PA	0.8765	0.8650	0.8765	0.9158	0.8917
	2PO1	0.8849	0.9335	0.9302	0.9477	0.9560
	ST	0.9058	0.9246	0.9126	0.9241	0.9362
	TL	0.9440	0.9440	0.9907	0.9717	0.9853
Average		***0.8960***	***0.9170***	***0.9260***	***0.9338***	***0.9403***
Sunny 1	**CR**	0.8842	0.9124	0.9080	0.9011	0.9099
	PA	0.9180	0.9506	0.9543	0.9424	0.9549
	2PO1	0.9368	0.9537	0.9506	0.9612	0.9706
	ST	0.9643	0.9781	0.9869	0.9681	0.9862
	TL	0.9668	0.9900	0.99	0.9831	0.9944
Average		***0.9340***	***0.9570***	***0.9580***	***0.9512***	***0.9632***

Table 4.3 Results and comparison between BOVW, pre-trained CNNs (FC-7 layer) and FC-7 Channel on cross validation sequences

In Table 4.4, results and comparison of BOVW and deep features extracted from COLD Freiburg obtained from softmax layer of three different pre-trained CNNs has been shown. As in Table 4.3, last column presents results of concatenation of deep features, here also in Table 4.4, last column shows results that we acquired when we concatenated deep features of softmax layer. We call it SoftMax Channel. The trend of accuracy is same here, and we can notice that SoftMax Channel gave us results better than all techniques.

		BOVW	Hybrid_CNN	ImageNet _CNN	Places_CNN	SoftMax Channel
			Softmax Layer	Softmax Layer	Softmax Layer	
			1183-by-1	1000-by-1	205-by-1	2388-by-1
Cloudy 2	**CR**	0.9247	0.9067	0.9148	0.8799	0.9072
	PA	0.9258	0.9569	0.9138	0.9088	0.9241
	2PO1	0.9394	0.9492	0.9285	0.9274	0.9514
	ST	0.9689	0.9662	0.9727	0.9607	0.9743
	TL	0.9814	0.9711	0.9749	0.9612	0.9880
Average		***0.9480***	***0.9500***	***0.9409***	***0.9276***	***0.95***
Night 1	**CR**	0.8687	0.8948	0.8948	0.8425	0.8870
	PA	0.8765	0.8938	0.8885	0.8488	0.8645
	2PO1	0.8849	0.9210	0.9131	0.9435	0.9670
	ST	0.9058	0.8844	0.9178	0.8655	0.9079
	TL	0.9440	0.9356	0.9702	0.8990	0.9587
Average		***0.8960***	***0.9059***	***0.9169***	***0.8799***	***0.9170***
Sunny 1	**CR**	0.8842	0.8655	0.8717	0.8335	0.8861
	PA	0.9180	0.9205	0.9255	0.8992	09462
	2PO1	0.9368	0.9337	0.9274	0.9243	0.9324
	ST	0.9643	0.9437	0.9725	0.9412	0.9787
	TL	0.9668	0.9593	0.9581	0.9368	0.9869
Average		***0.9340***	***0.9245***	***0.9310***	***0.907***	***0.9461***

Table 4.4 Results and comparison between BOVW, pre-trained CNNs (Softmax layer) and SoftMax-Channel on cross validation sequences

Although it was also noticed that generally, working with FC-7 deep features of all the sets gave results that were better than working with softmax layer features. In each case, Accuracy is higher for FC-7 features than Accuracy of softmax layer features.

The successful results obtained from concatenation of different type of deep features motivated us to concatenate three different channels, i.e., BOVW, FC-7 Channel, SoftMax Channel. So we applied the same idea and concatenated these three channels and what we got as results is shown in Table 4.5 Our experiments of concatenating three channels became successful on our cross-validation set as we can see the accuracy in each column of Table 4.5. Although it is noticeable that there wasn't a big increase in accuracy, especially the difference between FC-7 Channel and BOVW+ Deep Features -Channels is not really much, but generally the increasing trend of accuracy was observed in all test sequences. With that we concluded our experiments on cross validation set, and moved on to execute experiments on the whole COLD Freiburg Dataset.

		BoF	CNN (FC-7)	CNN (SoftMax)	BOVW+ Deep Features -Channels
		500-by-1	12288-by-1	2388-by-1	15176by-1
Cloudy 2	CR	0.9247	0.9689	0.9072	0.9651
	PA	0.9258	0.9754	0.9241	0.9776
	2PO1	0.9394	0.9743	0.9514	0.9733
	ST	0.9689	0.9940	0.9743	0.9918
	TL	0.9814	0.9918	0.9880	0.9924
Average		**0.9480**	**0.9809**	***0.95***	**0.98004**
Night 1	CR	0.8687	0.9325	0.8870	0.9383
	PA	0.8765	0.8917	0.8645	0.9016
	2PO1	0.8849	0.9560	0.9670	0.9618
	ST	0.9058	0.9362	0.9079	0.9471
	TL	0.9440	0.9853	0.9587	0.9801
Average		**0.8960**	**0.9403**	***0.9170***	**0.94578**
Sunny 1	CR	0.8842	0.9099	0.8861	0.9111
	PA	0.9180	0.9549	09462	0.9543
	2PO1	0.9368	0.9706	0.9324	0.9762
	ST	0.9643	0.9862	0.9787	0.9912
	TL	0.9668	0.9944	0.9869	0.9950
Average		**0.9340**	**0.9632**	***0.9461***	**0.96556**

Table 4.5 Results and comparison between BOVW, FC-7-Channel, SoftMax-Channel and BOVW+ Deep Features -Channels on cross validation sequences

4.3 Results on COLD Freiburg

The experiments were meant for the purpose of recognizing a room, observed during training, when captured under different conditions, i.e. at a different time and/or under different illumination condition. We performed experiments for all sequences. For each experiment, training set consisted of one sequence, and testing was done on all other sequences acquired in the same laboratory, under various conditions. The parameters of the algorithms were always the same.

4.3.1 Baseline: BOVW Results

The obtained results of Bag of Visual Words Technique are shown in Figure 4.1 and Figure 4.2. We considered these results as our baseline. Our baseline results were even better than [3] and [5]. Extended sequences performed better than standard sequences.

The bar chart on left side in Figure 4.1 shows results of BOVW for the standard sequences i.e., Part A Path 1 and Part B Path 3 i.e., Standard Path; and Figure 4.2 shown on right side presents results for the extended sequences Part A Path 2, Extended Path. For each training illumination condition (marked on top of the charts), the bars present

the average accuracy i.e. correctly classified instances over the corresponding testing sequences under the illumination condition labeled on the x- axis.

It can be observed from the results that recognition rate is higher when training and testing was done on sequences with same illumination condition. Also we can see that there is less difference between accuracy when trained and tested using cloudy and sunny sequences respectively and vice versa. But the difference is considerable when trained and tested using cloudy/sunny and night respectively and vice versa. One thing more that is obvious is that extended sequences gave better results than standard sequences. Note that training and testing sequences belong to the same path, that is, training/testing on all possible permutations of Part A Path 1, and then training/testing on all possible permutations of Part A Path 2, and Part B Path 3.

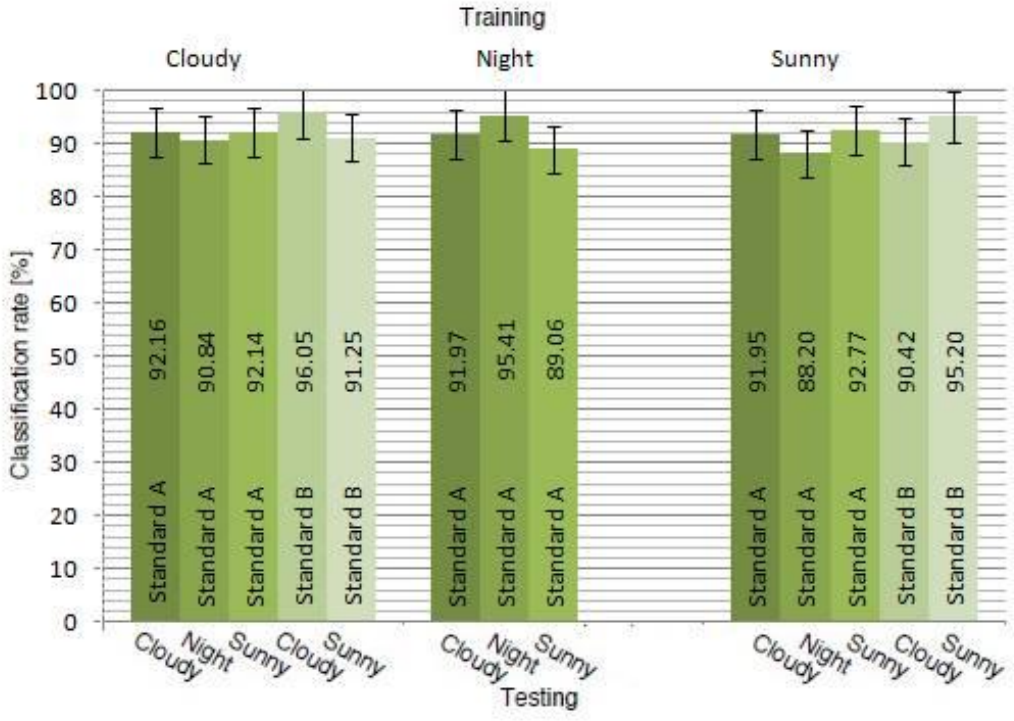


Figure 4.1 BOVW on COLD Freiburg (Standard)

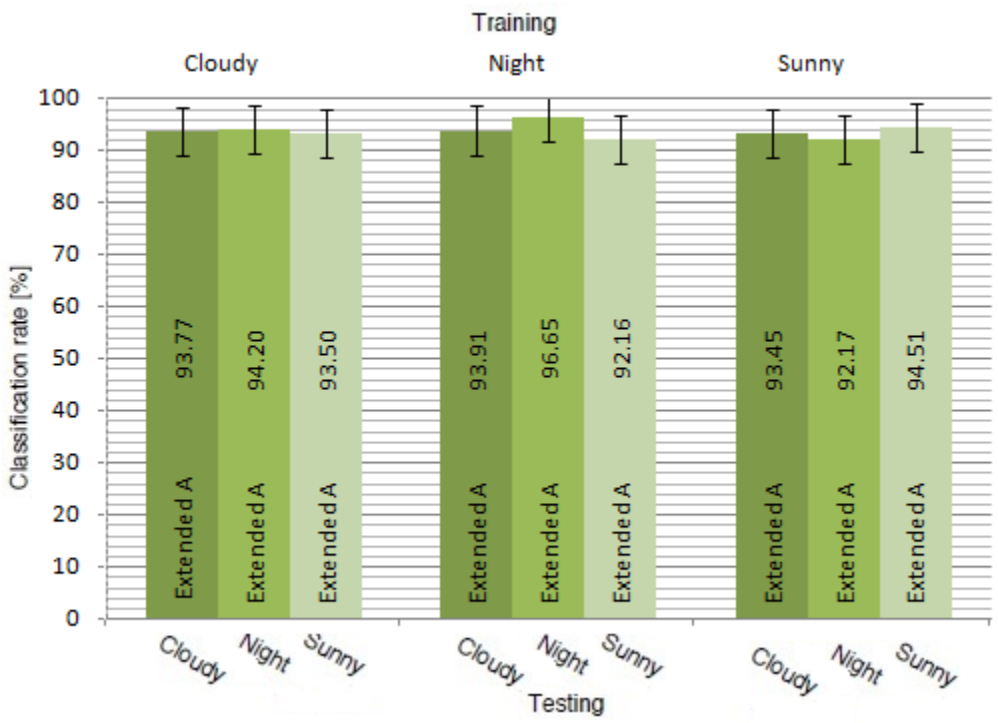


Figure 4.2 BOVW on COLD Freiburg (Extended)

4.3.2 Our Approach: BOVW+ Deep Features -Channels Results

In Figures 4.3 and 4.4 we have shown results of classification on "BOVW+ Deep Features –Channels" (obtained by concatenating BOVW, CNN-FC-7 Channel, and CNN-SoftMax Channel). These results outperform our baseline results but the trend of accuracy is same as was noticed previously in our baseline results. Under same illumination conditions, training and testing gave better results. And as mentioned earlier, here also, extended sequences performed better than standard sequences.

The bar chart on left side in Fig. 4.3 shows results of BOVW+ Deep Features -Channels for the standard sequences i.e., Part A Path 1 and Part B Path 3 i.e., Standard Path; and Fig. 4.4 shown on right side presents results for the extended sequences Part A Path 2, Extended Path. For each training illumination condition (marked on top of the charts), the bars present the average accuracy i.e. correctly classified instances over the corresponding testing sequences under the illumination condition labeled on the x- axis.

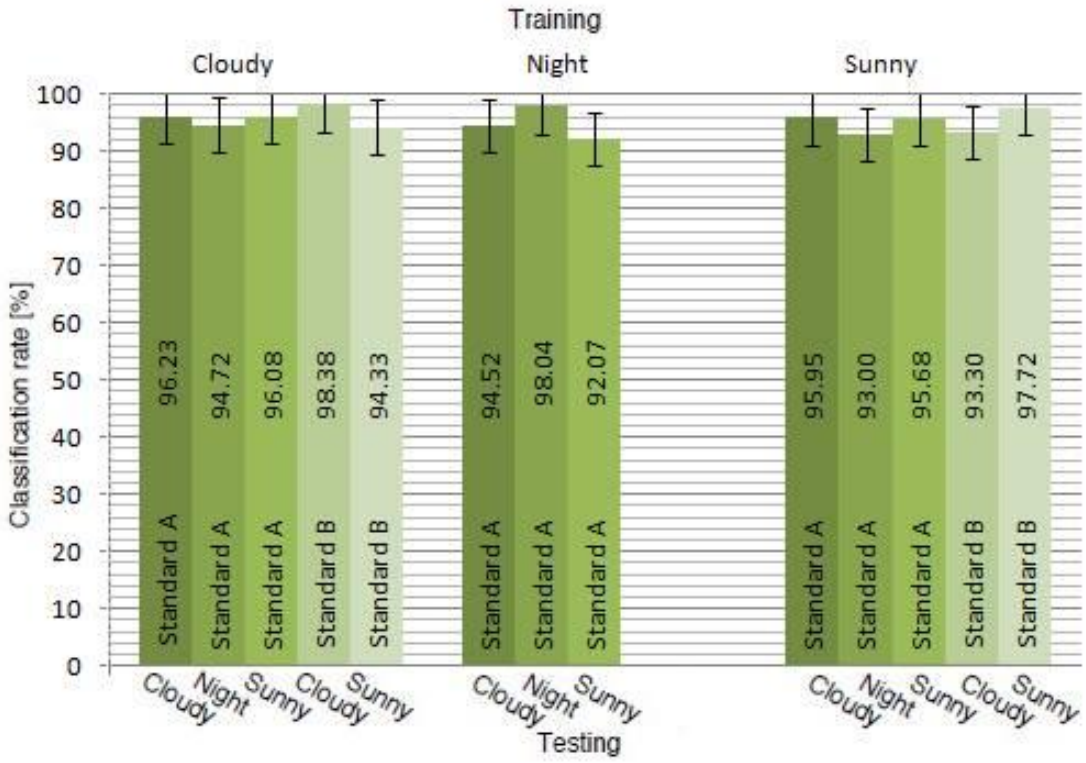


Figure 4.3 BOVW+ Deep Features -Channels on COLD Freiburg (Standard)

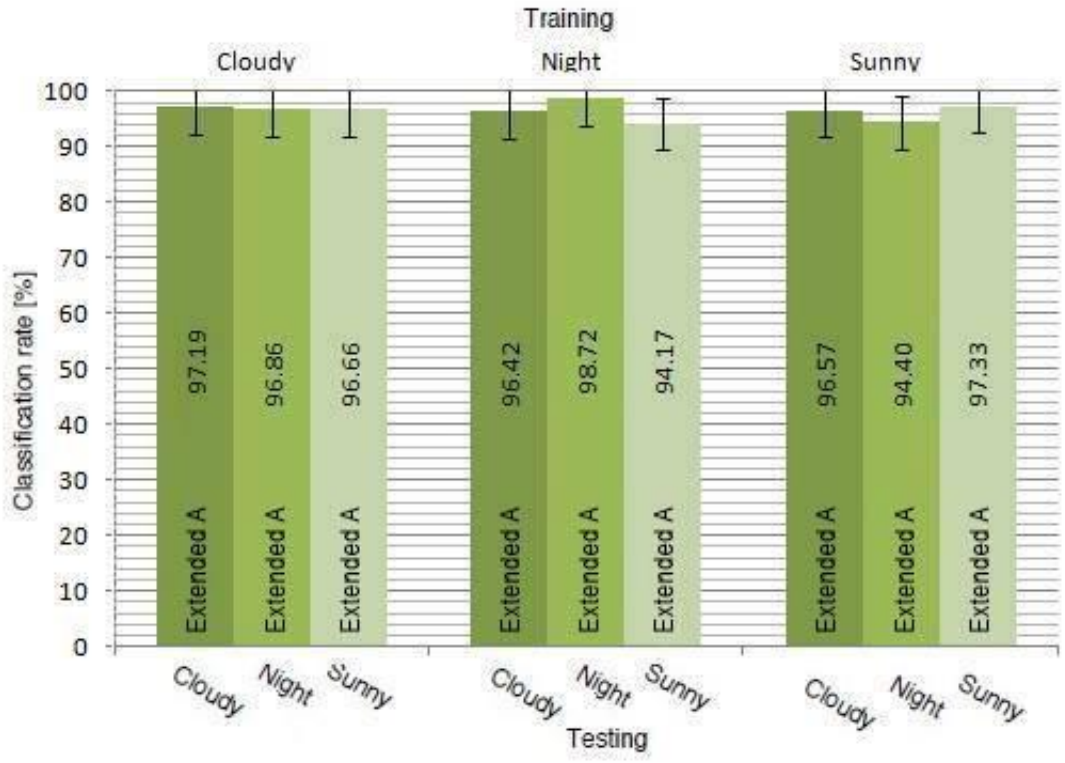


Figure 4.4 BOVW+ Deep Features -Channels on COLD Freiburg (Extended)

It can be noted that the overall performance of the method is better when trained and tested under stable illumination conditions. But it is very important to note here that even if the illumination conditions for training and testing were the same, the method had to tackle with other kinds of variations introduced e.g. by human activity or viewpoint changes due to the manual control of the robot. The errors usually occur in the transition areas between the rooms. We could also see that the classification rates obtained for the extended image sequences are generally better than those obtained for the standard sequences.

4.4 Comparison with other Methods

This section now compares our baseline results and BOVW+ Deep Features -Channels results with the results presented in [3] and [5]. Graph in Figure 4.5 shows the average accuracy of whole dataset when trained on Standard Cloudy, Night and Sunny sequences and tested against all other sequences. It is very clear from the graph that BOVW+ Deep Features-Channels approach outperforms all other techniques, including our baseline method. Graph in Figure 4.6 compares the performance of all approaches on extended sequence of COLD Freiburg. We can see that our BOVW+ Deep Features-Channels approach gave the best results. Difference of accuracy is quite noticeable among all the methods.

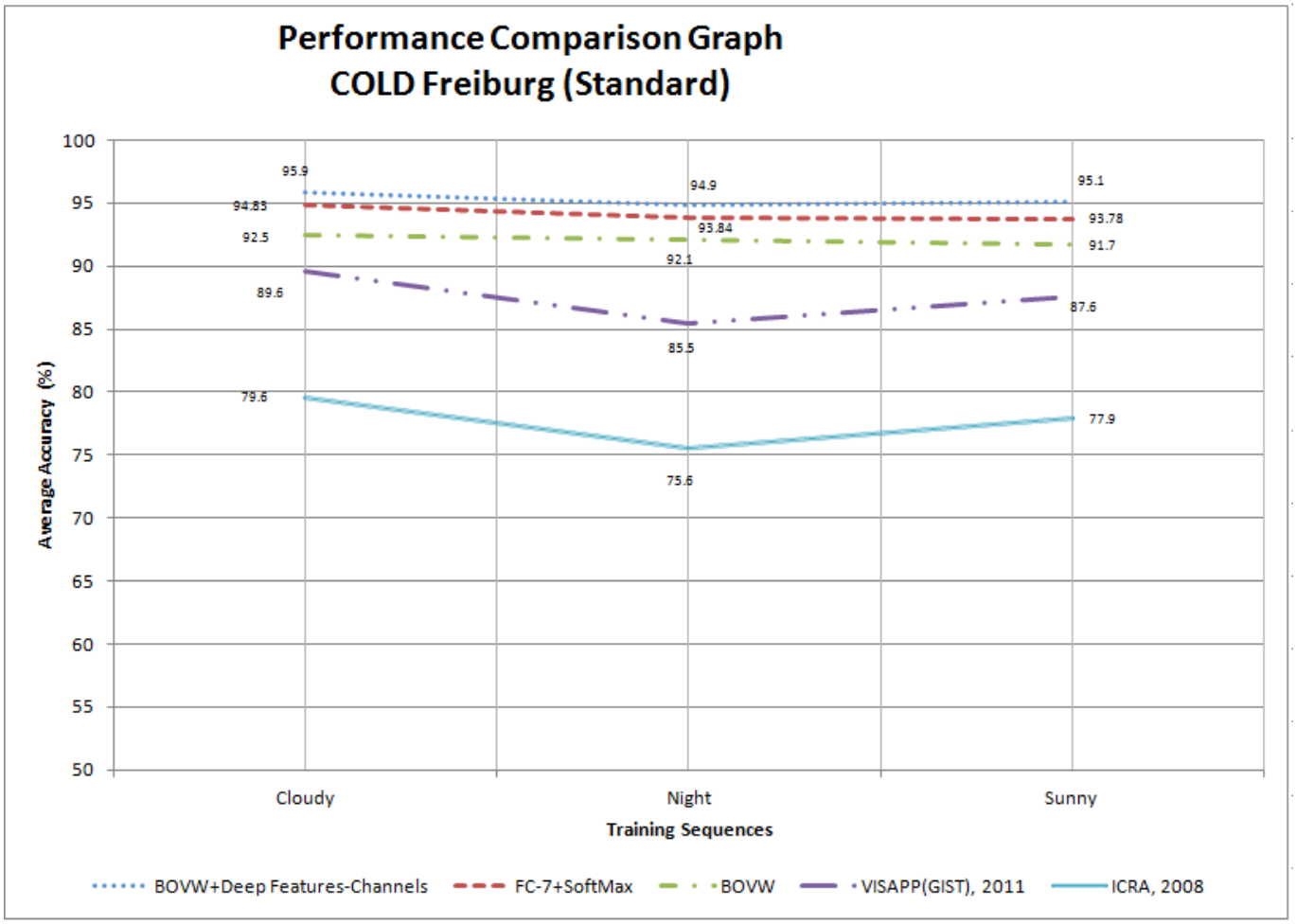


Figure 4.5 Performance Comparison COLD Freiburg (Standard)

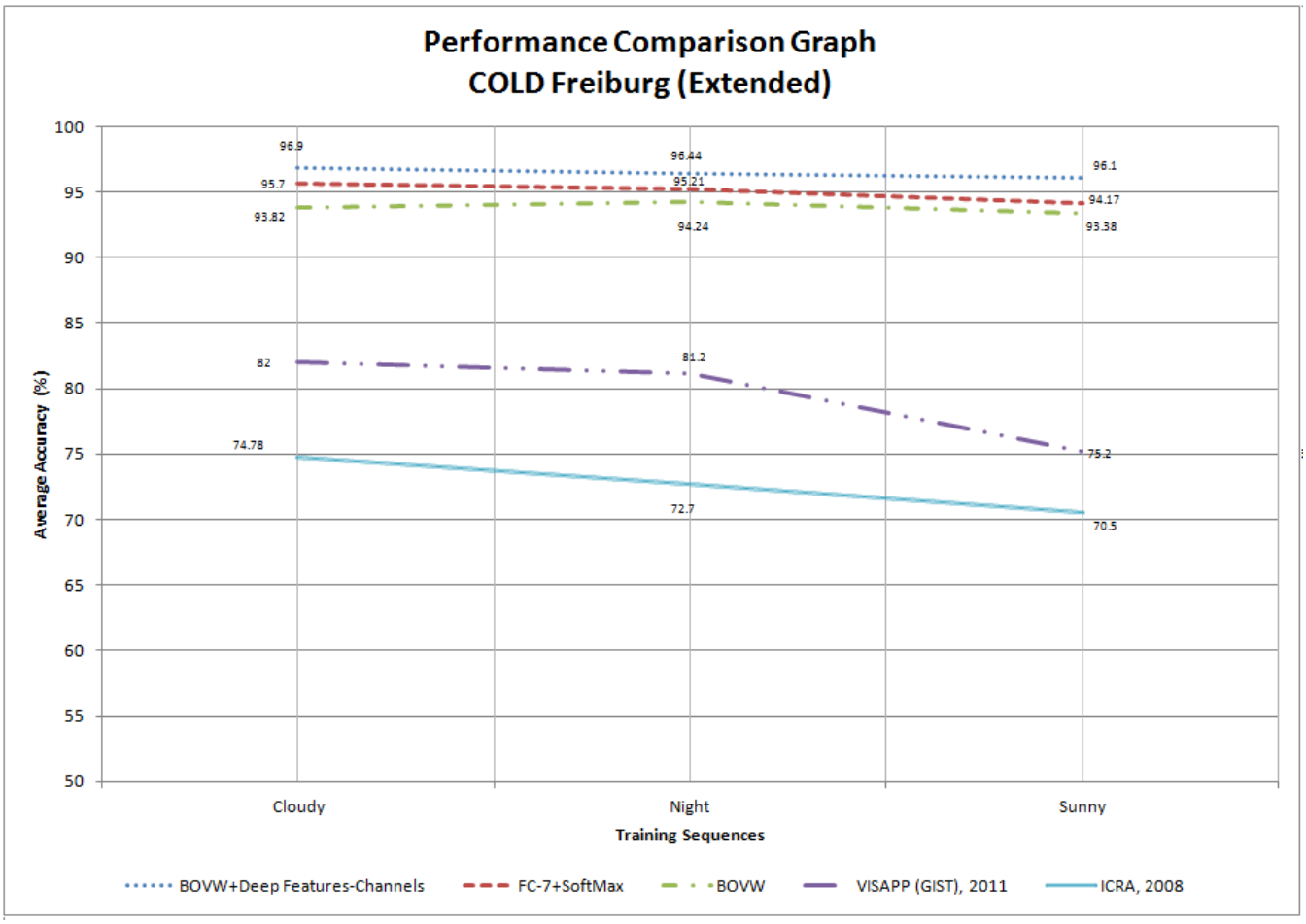


Figure 4.6 Performance Comparison COLD Freiburg (Extended)

Performance graphs in Figure 4.5 and Figure 4.6 not only compare our experimental results with other research methods, but in these graphs, there is an additional line that is plotted against the average accuracy of classifying a combination of CNN-FC-7 Channel and CNN-SoftMax-Channel. We concatenated these two channels and performed classification to know that how much these deep features contributed to BOVW performance. Deep features contributed to a greater extent in the performance of BOVW and gave promising results. We showed Individual curves of BOVW and 2-channels, and then we showed a curve representing the results of BOVW+ Deep Features -Channels i.e., concatenation of BOVW, CNN-FC-7 Channel and CNN-SoftMax Channel. These results show that in Standard sequences, average accuracy of BOVW was improved by 1%-2% when combined with deep feature channels, and in extended sequences, average accuracy of BOVW was improved by 3%-4% when combined with deep feature channels. It is also a considerable point that our baseline results on BOVW, results of deep features and results of BOVW+ Deep Features-channels remained consistent for all sequences under different illumination conditions. There is not a big difference of accuracy when trained and tested under different illumination conditions. Our approach of combining deep features with BOVW was quite significant, although very simple.

Chapter 5

5. Conclusion

We know that real life images have many variations in them, such as illumination variations, aliasing, weather conditions, moving and re-arrangement of objects over a certain period of time, variations introduced by people coming in and out of a certain place etc. These variations make this task a challenging one. In this research, we used a sub-dataset COLD Freiburg of COLD dataset. This dataset is huge dataset of indoor perspective and omni-directional annotated images, captured at three labs of Europe. In this image dataset, images are captured over a period of six months, under different illumination conditions. Different variations introduced in this dataset make it a perfect dataset to test and verify the robustness and efficiency of technique used for indoor place recognition. A detailed description of COLD Freiburg sequences and positive/negative images has been given in chapter 2. First of all, we built our baseline using *Bag of Visual Words* Technique and computed accuracy by training and testing on all possible permutations of image sequences. Average accuracy of BOVW was 92.1% for Standard sequences and 93.81% for Extended sequences. After this we experimented with pre-trained CNNs to extract deep features from COLD Freiburg images and combine these features with BOVW. These CNNs are pre-trained on three different datasets, i.e., ImageNet; a huge object-centric dataset comprising of millions of images of objects belonging to hundreds of different categories, Places dataset; a huge scene-centric dataset that consists of thousands of images of places of many different categories, a hybrid of these two datasets. We used, in our, experiments, deep features from two different layers of CNNs, FC-7 layer of CNN and SoftMax layer of CNN. Intuition of experimenting with these pre-trained CNNs was that COLD Freiburg images consist of objects and places whose category is also found in ImageNet and Places datasets. So we pondered over the fact that these pre-trained CNNs can give us promising results. We combined features from FC-7 layer to have CNN-FC-7 Channel, features from SoftMax layer to come up with CNN-SoftMax Channel and BOVW channel. And finally to concatenate these three channels to construct a final combination. The classifier we used was a linear SVM. Average accuracy of BOVW+ Deep Features -Channels was 95.3% for Standard sequences and 96.48% for Extended sequences which was quite better than our baseline results. All the results have been reported in Chapter 4. We deduced that deep features contributed 2% to 3% in the performance of BOVW, and improved the accuracy. We compared our results with two methods reported in [3] where accuracy was 77.7% for Standard sequences and 72.66% for Extended sequences, and the accuracy reported in [5] using GIST was 87.57% for Standard sequences and 79.47% for Extended sequences, they used COLD dataset in their research. These figures clearly show that a simple concatenation of BOVW with deep features extracted from FC-7 layer and SoftMax layer of pre-trained CNNs proved to be a positive idea for improving accuracy of Indoor Place Recognition. We come up with results that out-performed both methods and showed accuracy that is much better than these two research methods.

5.1 Future Directions

In this research we proposed a simple technique of combining deep features with local features extracted from images, and came up with better accuracy rate. This research idea can be further extended to study other combination schemes to make place recognition more robust. There are two directions that can lead us in future:

5.1.2 Combination Schemes

- Pre-computed kernel matrices can be used to come up with more sophisticated feature combination technique. Sum/product of different pre-computed kernels can be a good step.

- 'Late Fusion' of features, i.e., train different classifiers and average their scores, is another future directive.

5.1.3 Training/Re-training of CNN

- Instead of using pre-trained CNNs, training of CNNs on COLD dataset and use these CNNs for classification.

- Another future plan is to apply a "transfer learning" technique that aims to transfer knowledge between related source and target domains. We can use pre-trained internal layers of the CNN (e.g., on ImageNet) and fine tune the fully connected layers on the COLD database (e.g., [22]).

Bibliography

[1] Image Analysis by Denise Hattwig, University of Washington

[2] Bernd Jähne and Horst Haußecker (2000). Computer Vision and Applications, A Guide for Students and Practitioners. Academic Press. ISBN 0-13-085198-1

[3] M. M. Ullah, *"Towards Robust Place Recognition for Robot Localization"*, IEEE International Conference on Robotics and Automation; Pasadena, CA, USA; May 19-23, 2008.

[4] Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, Tomas Pajdla, *24/7 place recognition by view synthesis,* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[5] Hervé Guillaume, Mathieu Dubois, Frenoux Emmanuelle, Philippe Tarroux. *Temporal Bag-of-Words - A Generative Model for Visual Place Recognition using Temporal Integration*. VISAPP - International Conference on Computer Vision Theory and Applications - 2011, Mar 2011, Vilamoura, Portugal. 201

[6] Iwan Ulrich, Illah Nourbakhsh, *"Appearance-Based Place Recognition for Topological Localization"*, Proceedings of the 2000 IEEE International Conference on Robotics & Automation San Francisco, CA April 2000

[7] Benjamin Kuipers and Patrick Beeson,*"Bootstrap Learning for Place Recognition",* AAAI-02 Proceedings. Copyright © 2002

[8] Axel Rottmann , Óscar Martínez Mozos, Cyrill Stachniss, Wolfram Burgar, *"Semantic Place Classification of Indoor Environments with Mobile Robots using Boosting"*, AAAI'05 Proceedings of the 20th national conference on Artificial intelligence - Volume 3 Pages 1306-1311

[9] S Lazebnik, C Schmid, J Ponce, "*Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories"*. In Proc. Computer Vision and Pattern Recognition (CVPR), 2006.

[10] F Orabona, C Castellini, B Caputo, J Luo, G Sandini, "*Indoor Place Recognition using Online Independent Support VectorMachines"*, 18th British Machine Vision Conference (BMVC07), 2007

[11] A. Pronobis , O. Mart´ınez Mozos , B. Caputo, *"SVM-based Discriminative Accumulation Scheme for Place Recognition"*, 2008 IEEE International Conference on Robotics and Automation Pasadena, CA, USA, May 19-23, 2008

[12] Mayank Juneja, Andrea Vedaldi, C. V. Jawahar, Andrew Zisserman, *"Blocks that Shout: Distinctive Parts for Scene Classification*", IEEE Conference on Computer Vision and Pattern Recognition, 2013

[13] M. M. Ullah, A. Pronobis, B. Caputo, J. Luo, and P. Jensfelt. The COLD database. Technical Report TRITA-CSC-CV 2007:1, Kungliga Tekniska Hoegskolan, CVAP/CAS, October 2007.

[14] A. Pronobis, B. Caputo, "*COLD: COsy Localization Database*", The International Journal of Robotics Research (IJRR), 28(5), 2009.

[15] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "*ImageNet Classification with Deep Convolutional Neural Networks"*, Advances in Neural Information Processing Systems 25 (NIPS 2012)

[16] http://image-net.org/index

[17] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. *"Learning Deep Features for Scene Recognition using Places Database"*. Advances in Neural Information Processing Systems 27 (NIPS), 2014.

[18] http://www.aishack.in/tutorials/sift-scale-invariant-feature-transform-features/

[19] http://www.vlfeat.org/matlab/matlab.html

[20] Lowe, David G. (1999), *"Object recognition from local scale-invariant features"*. Proceedings of the International Conference on Computer Vision 2. pp. 1150–1157.doi:10.1109/ICCV.1999.790410

[21] http://code.flickr.net/2014/10/20/introducing-flickr-park-or-bird/

[22] M. Oquab, L. Bottou, I. Laptev and J. Sivic. *"Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks"*. In Proc. Computer Vision and Pattern Recognition (CVPR), Columbus, Ohio, USA, 2014.