NATURAL SCENE CLASSIFICATION, ANNOTATION AND RETRIEVAL

Y.T.N. ALQASRAWI

PhD

UNIVERSITY OF BRADFORD

2012

NATURAL SCENE CLASSIFICATION, ANNOTATION AND RETRIEVAL

Developing different approaches for semantic scene modelling based on Bag of Visual Words

YOUSEF T. N. ALQASRAWI

Submitted for the degree of
Doctor of Philosophy

Department of Computing
School of Computing, Informatics and Media
University of Bradford

2012

Keywords: image classification, image retrieval, bag of visual words, visual vocabulary, features fusion, spatial pyramid layout, concept-based bag of visual words.

Abstract

With the availability of inexpensive hardware and software, digital imaging has become an important medium of communication in our daily lives. A huge amount of digital images are being collected and become available through the internet and stored in various fields such as personal image collections, medical imaging, digital arts etc. Therefore, it is important to make sure that images are stored, searched and accessed in an efficient manner. The use of bag of visual words (BOW) model for modelling images based on local invariant features computed at interest point locations has become a standard choice for many computer vision tasks. Based on this promising model, this thesis investigates three main problems: natural scene classification, annotation and retrieval. Given an image, the task is to design a system that can determine to which class that image belongs to (classification), what semantic concepts it contain (annotation) and what images are most similar to (retrieval).

This thesis contributes to scene classification by proposing a weighting approach, named keypoints density-based weighting method (KDW), to control the fusion of colour information and bag of visual words on spatial pyramid layout in a unified framework. Different configurations of BOW, integrated visual vocabularies and multiple image descriptors are investigated and analyzed. The proposed approaches are extensively evaluated over three well-known scene classification datasets with 6, 8 and 15 scene categories using 10-fold cross validation. The second contribution in this thesis, the scene annotation task, is to explore whether the integrated visual vocabularies generated for scene classification can be used to model the local semantic information of natural scenes. In this direction, image annotation is considered as a classification problem where images are partitioned into 10x10 fixed grid and each block, represented by BOW and different image descriptors, is classified into one of predefined semantic classes. An image is then represented by counting the percentage of every semantic concept detected in the image. Experimental results on 6 scene categories demonstrate the effectiveness of the proposed approach. Finally, this thesis further explores, with an extensive experimental work, the use of different configurations of the BOW for natural scene retrieval.

Declaration

I hereby declare that this thesis has been genuinely carried out by myself and has not been used in any previous application for a degree. The invaluable participation of others in this thesis has been acknowledged where appropriate.

Yousef Alqasrawi

Dedication

This thesis is dedicated to the memory of my grandfather, Naji, whose encouragement served as a source of inspiration, my parents, my wife Rana, daughter Rahaf, and son Mohammad for their patience, understanding, support, and love.

Acknowledgments

In the Name of Allah, the Most Gracious, the Most Merciful

All Praise is due to Allah for His Glorious Ability and Great Power for granting me the health, strength, patience and ability to complete this doctoral thesis.

I am heartily thankful to my supervisors Professor Daniel Neagu and Professor Peter Cowling, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject. They have offered me all the assistance to complete this thesis. It is an honour for me to have them both as my supervisors.

I would like to acknowledge the financial support of the Applied Science University in Jordan which offered me a scholarship for PhD studies in the University of Bradford. I also thank the School of Computing, Informatics and Media (SCIM) for offering all the support and assistance, especially the head of Computing department, Mr. Mick Ridley.

I would like to thank Dr. Paul Trundle and Dr. Sophia Alim for revising and proofreading my work. Also, I am grateful to all friends in the artificial intelligence group for their friendship and support.

I would like to thank Julia Vogel for providing me her image dataset. Also, I would like to thank all people who I have been in contact with during my doctoral study.

Last but not least, my appreciation and love goes to my parents, wife and children for their support, patience and love.

Peer Reviewed Publications and Contributions

Journal papers:

- Alqasrawi Y., Neagu D. and Cowling P.I. (2011): "Fusing Integrated Visual Vocabularies-Based Bag of Visual Words and Weighted Colour Moments on Spatial Pyramid Layout for Natural Scene Image Classification" *Signal, Image and Video Processing*, Springer.(*in print*), http://www.springerlink.com/content/b06h4h310u7t1441/

Conferences and workshops:

- Alqasrawi Y., Neagu D. and Cowling P. (2010): "Spatial Pyramid Local Keypoints Quantization for Bag of Visual Patches Image Representation", The 10th International Conference on Intelligent Design and Applications ISDA10, pp. 1270-1274, IEEE Computer Society, Cairo, Egypt.

- Alqasrawi Y., Neagu D. and Cowling P. (2009): "Natural Scene Image Recognition by Fusing Weighted Colour Moments with Bag of Visual Patches on Spatial Pyramid Layout", The 9th International Conference on Intelligent Design and Applications ISDA09 pp. 140-145, IEEE Computer Society, **Best Student Paper Award**, Pisa, Italy.

- Alqasrawi Y., Neagu D. and Cowling P. (2008): "Overview of Current Approaches Towards Better Image Retrieval Systems", Proceedings of the Ninth Informatics Workshop for Research Students, pp. 76-79, June 2008, University of Bradford. ISBN 978 1 85143 251 6.

Posters, Presentations and Training

Posters:

2009- Showcase poster: A prototype for Content-based Image Retrieval, University of Bradford.

2008- Showcase poster: Natural scene modelling and classification, University of Bradford.

Presentations:

2010- Research Seminar: Natural Scene Image Recognition by Fusing Weighted Colour Moments, University of Bradford.

2009- Presentation and Demonstration to students from Bradford School: Using a Software prototype for Content-based Image Retrieval, University of Bradford.

2009- Presentation and Demonstration to students from Bradford Grammar School: Using a Software prototype for Content-based Image Retrieval, University of Bradford.

2009- Presentation and Demonstration to students from Titus Salt Grammar School: Using a Software prototype for Content-based Image Retrieval, Titus Salt Grammar School, Shipley.

Training:

2010- International Computer Vision Summer School (ICVSS): Registration and Video Analysis. Sicily, Italy, 12-17 July, 2010.

Planed publications:

- Image Annotation based on the bag of visual words model and local semantic concepts.
- Natural scene retrieval using the bag of visual words model and local semantic concepts: A comparative study.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid development of new information technologies, use of digital cameras in our daily lives, photo sharing websites and social networks, have led to an explosion in the amount of visual media such as images and videos available. With the development of inexpensive digital camera equipment and storage, it is now not exclusively for the professional photographer to take and archive pictures. Digital photography has become affordable to amateur photographers and within the reach of our family and friends. With the growth in technological digital imaging, the problems of storing, sorting and searching images have grown too. Moreover, huge amount of images are collected everyday in various fields such as digital books, digital art, medical imaging, aerial and satellite data (Singh and Cunningham, 2008). Each of these fields poses different challenges for image analysis and understanding. The possible applications for image analysis include medicine (Müller et al., 2004), photo and news-journalism (Eakins, 2002), astronomy (Csillaghy et al., 2000), art

and fashion (Datta et al., 2008), retailing (Gangopadhyay, 2001) , military (Li et al., 2000) , education (Rui et al., 1999) etc.

Unlike digital computers, humans are able to interpret the semantic content of images, which is still beyond the capabilities of computer vision systems. For a computer system, an image is not more than a matrix of pixel values which are summarized by low-level features such as colour or texture features. For humans it is an image which contains what he/she can see, such as *sky*, *water* or *forest*. This makes visual content analysis of images a challenging problem in computer vision and other related fields, such as artificial intelligence and image data management (Quelhas, 2007).

One of the main and challenging problems in image content representation is the *semantic gap* (Smeulders et al., 2000). It is the gap between low-level image features of the image content and the human perception. For example, in content-based image retrieval (CBIR) systems, the user provides a query image and he/she wants the system to retrieve images from the database that are visually similar to the query image. However, retrieved images that are considered by the system visually relevant images to the query image may not semantically relevant (the *semantic gap*). Feature extraction is an important phase in any CBIR system to represent the image content. Much of the early work in image representation was to develop algorithms to extract different low-level features from image visual content, such as colour, texture and shape. These features are used by the system to build the index of images in the database. However, these low-level features may be unrelated to the concepts expressed in the image, such as *sky* and *grass* (semantic information).

Extracting semantic information from image content is a field that tries to mimic the way human perception works, which is still a challenging and difficult task to accomplish in a computational system. One goal of image content analysis is to narrow the semantic gap between the image's visual content and the human understanding of image content. Image classification, annotation and retrieval are challenging tasks in computer vision that are affected by this research problem when matching the semantics of images in the database. The use of image modelling based on local features has provided significant progress in terms of robustness, efficiency and quality of results. Much of the previous work on image classification was based on bag of visual words (BOW) model (Fei-Fei and Perona, 2005, Csurka et al., 2004, Quelhas et al., 2007, Battiato et al., 2010a, Liu et al., 2011, Hou et al., 2011, Jiang et al., 2010), a model that was brought from the text document retrieval field (Salton, 1983).

The bag of visual words model is based on local features extracted from image content using features detector and descriptor, such as difference of Gaussian (DoG) detector and scale-invariant feature transform (SIFT) descriptor (Lowe, 2004). These descriptors are then quantized into a number of clusters, called visual words, where an image is then represented as a histogram of the occurrence of each visual word in an image. This model has shown an impressive performance in image classification and object recognition problems. It is considered as an intermediate semantic representation of the image content. This thesis aims to progress further in this field using the bag of visual words model for representing the semantic information of image content for natural scene categorization, annotation and retrieval.

1.1 Overview of the Thesis

In this thesis, image representation forms the core body of the work which is based on the bag of visual words model. Three challenging tasks are considered in this thesis: (1) image classification (2) image annotation and (3) image retrieval. The three tasks are mainly concerned with natural scene images, whereas object recognition is out of the scope of this thesis.

The main objective of this thesis is to develop models appropriate for representing image content in the context of natural scene image classification, annotation and retrieval. This work will study the representation of images by the BOW model in its various aspects.

In image classification systems, the aim is to separate images based on their visual content into two or more disjoint classes. Image scene category can be considered as a reliable indicator of the object presence within the image which means the later depends upon the former (Oliva and Torralba, 2001). Scene classification differs from object classification, in that a scene is composed of several entities organized in an changeable layout (Quelhas et al., 2005). In many instances, images from two different scenes are visually similar and can be difficult to differentiate between them, such as scenes of beach and lake. Early efforts in scene classification targeted binary classification, such as distinguishing indoor image from outdoor (Szummer and Picard, 1998), city form landscape (Vailaya et al., 1998) etc. Recently, there has been an effort to solve the problem in larger number of scene categories (Vogel and Schiele, 2004, Quelhas et al., 2005, Quelhas et al., 2007, Fei-Fei and Perona, 2005, Lazebnik et al., 2006, Bosch et al., 2008) and a dataset of 15 categories has been used as a benchmark to compare various approaches (Lazebnik

et al., 2006). Many of these approaches aim to provide intermediate representation between low level image features and high-level abstract of images. These are obtained using "themes" or "topic" representation (Fei-Fei and Perona, 2005, Quelhas et al., 2007).

The goal of this thesis, toward this task, is to analyse the visual content of image and identifies the corresponding scene category based on BOW. The main challenge in natural scene categorization is the high variability of the scene appearance, *refer to* Figure 1-1. This problem can be addressed using visual descriptors that are developed to be invariant to the appearance changes. Representing the appearance of images as a collection of raw pixel values is not robust enough for this task. Image descriptors should be general enough to catch the similarities between images of the same category. Also, it should be invariant

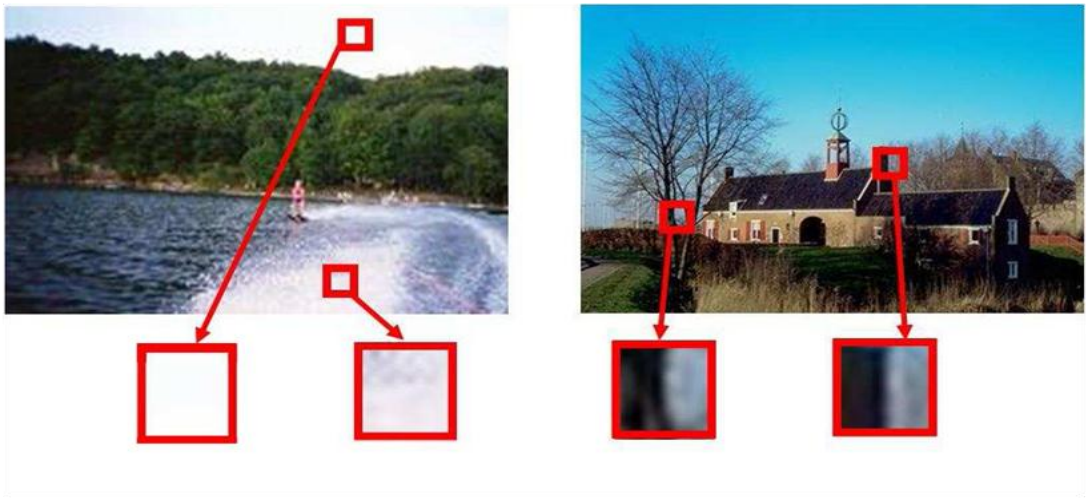


Figure 1-1: Image classification is a difficult task due to ambiguities in the appearance of image regions. In the left image, "sky" and "water" regions look similar while in the right image, "tree" and "building" look similar (Rasiwasia, 2011)

enough to differences in camera location climatic conditions etc.

1.2 Aims and Objectives

The aims and objectives of this work are as follows:

- **Image Classification Task:** Using bag of visual words model based on local descriptors is not enough to represent the visual content of natural scenes. Natural scene images depict colour information which is not modeled by the BOW approach. One goal in this task is to investigate whether densities of local keypoints located in image regions at spatial pyramid layout can be used to control the fusion of colour information and bag of visual word model in order to improve the performance of natural scene classification task. Also, the visual vocabularies required to build BOW are of an important aspect that needs to be investigated in depth to study their influence in the performance of natural scene classification task. The aim is to study the influence of using integrated visual vocabularies, generated from visual vocabularies obtained from each scene category, to improve the quality of bag of visual words model. Moreover, it is crucial to investigate the effect of using visual vocabularies generated from one dataset to build bag of visual words for another dataset on the same domain.
- **Image Annotation Task:** The goal in this task is to present a design of natural scene image annotation based on the semantic image representation. Semantic image representation refers to mapping from the space of image appearance features to a semantic space which

represents the semantic concepts, such as *water* and *grass*. Semantic scene modelling in this task refers to assigning local semantic concepts to image regions in a supervised manner. The aim in this task is to study if there is any relationship between the distribution of local semantic concepts and local keypoints located in image regions labelled with these semantic concepts. Based on this study, this thesis aims to investigate whether bag of visual words model can be used to efficiently represent the content of natural scene image regions, so images can be annotated with local semantic concepts. Another objective of this task is to study the implications of generating visual vocabularies from image halves, instead of generating them from the whole image, on the performance of annotating image regions with semantic labels.

- **Image Retrieval Task:** Currently content-based image retrieval solutions rely on visual similarities between images which are not correlated with the similarities humans use to compare images. CBIR systems use low-level features such as colour and texture to return images that are similar to the user query image. Such systems have proven to return inadequate results (Smeulders et al., 2000) because image matching was based on low-level features ignoring semantic contents. By using semantic image representation, the retrieval is performed at a higher level of semantic. The aim of this task is to investigate the plausibility of using different approaches presented in the first two tasks in order to represent the semantic information presented in images for the image retrieval task. The aim here is to

achieve higher precision accuracies for a query image while maintain the recall as high as possible.

1.3 Contributions

This thesis proposes a number of contributions to the fields of natural scene image classification, annotation and retrieval. These contributions can be summarized as follows:

- **Image Classification Task:** Based on bag of visual words model, a unified framework is developed to classify natural scene images into scene categories. The contributions in this task are threefold: (1) a new weighting method, namely *keypoints density-based weighting method*, is proposed to control the fusion of colour information of an image and its bag of visual words histogram on a spatial pyramid layout; (2) it is demonstrated that integrating visual vocabularies generated from each scene category has improved the discriminative power of bag of visual words and thus the performance of natural scene classification; (3) visual vocabularies generated for specific scene categories have shown to be appropriate to build bag of visual words for images from different dataset on the same domain. All the proposed approaches are extensively evaluated over three well-known natural scene datasets. This work has been published in (Alqasrawi et al., 2011, Alqasrawi et al., 2009)
- **Image Annotation Task:** A hypothesis is proposed to study the correlation between the distribution of local semantic concepts and local keypoints detected in image regions and annotated with these

semantic concepts. This hypothesis is justified by an in depth analysis of the distributions of both; local semantic concepts and local keypoints. Based on this hypothesis, concept-based bag of visual words is proposed to represent the visual content of image regions. Also, it is demonstrated that visual vocabularies generated from global scene categories can be used effectively as visual vocabularies to represent local semantic concept. This approach is called local from global. Finally, concept-based bag of visual words is improved by proposing to build visual vocabularies from image halves. Extensive experiments are conducted over a natural scene dataset with six categories. Part of this work has been published in (Alqasrawi et al., 2010).

- **Image Retrieval Task:** An extensive experimental work is conducted in this task to study the influence of using BOWs approaches proposed in the first task to perform natural scene retrieval. Also, the concept-occurrence vector proposed in (Vogel, 2004) is employed to represent the occurrence of local semantic concepts in natural scene images. Local semantic concepts are labels assigned to image regions represented by the concept-based BOWs histograms presented in the second task. All approaches presented in this task are compared and their performances are reported.

1.4 Thesis Structure

The rest of this thesis is organised as follows.

Chapter 2: Literature Review. This chapter describes prior work done in the research areas of computer vision in general and content-based image retrieval, image classification and image annotation in particular.

Chapter 3: Background. This chapter introduces the basic concepts and terminologies used in this thesis. It introduces image retrieval methods, image representation, machine learning and the evaluation methods used in this work.

Chapter 4: Image Classification. This chapter investigates a framework for image classification using the bag of visual words model. A new weighting method, to control the fusion of image colour information and bag of visual words histograms on a spatial pyramid layout, is proposed. The framework also investigates building visual vocabularies from image categories to build more discriminative bag of visual words histograms on spatial pyramid layout. Extensive experiments were carried out to evaluate the proposed approaches over three well-known natural scene datasets and their performances are reported and compared to a number of baseline methods.

Chapter 5: Image Annotation. This chapter presents a framework for automatic natural scene image annotation with local semantic concepts from a constrained vocabulary. This chapter is based on bag of visual words models described in Chapter 4. This chapter studies the correlation between the distribution of local semantic concepts and local keypoints located in image regions. Also, this chapter presents local from global approach which study the influence of using visual vocabularies generated from general scene categories to build bag of visual

words at region level. This chapter also investigates the plausibility of building visual vocabularies from image halves rather than from the whole image to build better quality bag of visual words histograms at image region level. The work in this chapter is extensively evaluated over a natural scene dataset with six categories and nine semantic concepts and results are reported.

Chapter 6: Image Retrieval. This chapter presents different approaches which are based on the BOW model to represent the semantic information of natural scene images for scene retrieval task. Approaches presented in Chapter 4 will be used for image retrieval. The distributions of local semantic concepts assigned to image regions, presented in Chapter 5, will be used for image retrieval. The retrieval results of all approaches presented in this chapter are reported and compared.

Chapter 7: Conclusions and Future Work. This chapter presents a summary of the research contributions contained in this thesis and suggest areas for future research.

Chapter 2

Literature Review

In the literature, much work has been done in image/object classification, annotation and retrieval based on the bag of visual words approach. This thesis will refer to scene Classification, Annotation and Retrieval as (CAR). This chapter review those approaches most strongly connected or related to the work presented in this thesis. Figure 2-1 presents a hierarchical view of the topics, with sample references, that will be covered in this chapter. While early work in scene CAR focused on building techniques to extract low-level features from images, these techniques were not efficient to represent the high-level semantic of user perception (Smeulders et al., 2000, Datta et al., 2008, Liu et al., 2007). This problem is well known as the *semantic gap*. Semantic modelling of scenes has been considered an important step towards intermediate representation of images so as to narrow the semantic gap between low-level features and high-level user understanding. Bag of visual words model and local semantic concepts are common approaches to the

semantic modelling of scenes. To this end, the proposed taxonomy presents a review of some of the early works, which are based on low-level features, and advances in scene CAR based on semantic modelling. Such review can help understand the work presented in this thesis as well as to help new researchers to understand different scene CAR approaches and the new directions in semantic scene modelling based on bag of visual words model and local semantic concepts. There are hundreds of publications about visual content representation using the BOW model as it is a promising method for visual content classification (Tirilly et al. 2008), annotation (Wu et al. 2009), and retrieval (Zheng et al. 2008).

2.1 Image Classification: A Review

Image classification is an important application in computer vision. It has been investigated in two complementary research areas, understanding human visual perception of scenes and developing computer vision techniques for automatic scene categorization. This reveals how humans are able to understand, analyze and represent scenes, which is of great research interest in many psychological studies. Knowledge from these studies can help computer vision researchers design and develop systems to close the gap between human semantic and image low-level information (Vogel et al., 2007, Ross and Oliva, 2010).

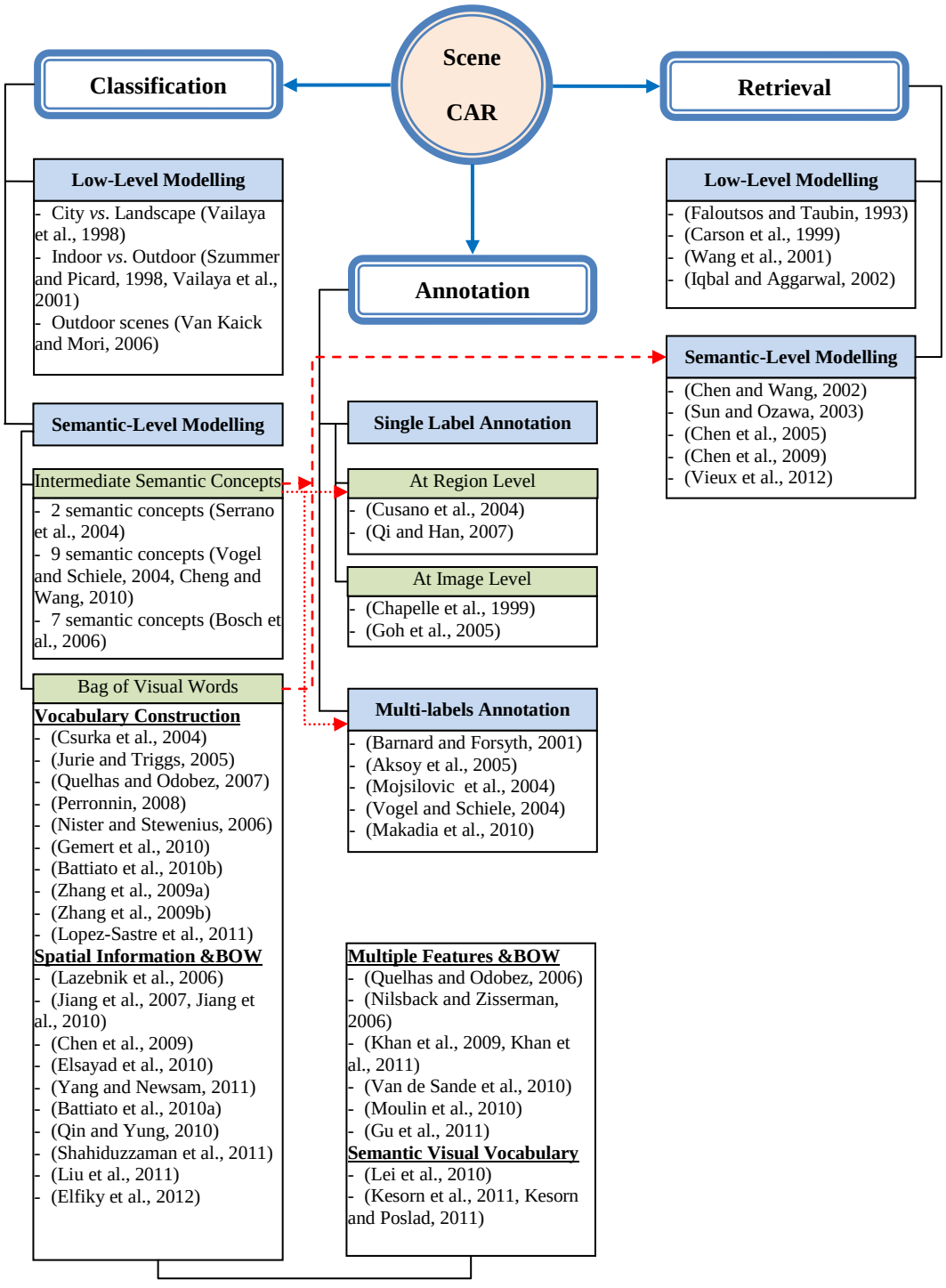


Figure 2-1: Block diagram of the topics covered in this chapter with sample references used in the scene CAR literature. Red arrows show that semantic concepts and BOW model can also be applied to scene annotation and semantic-based image retrieval.

From the computer vision viewpoint, scene classification is the task of automatically assigning an unlabelled image into one of several predefined classes (e.g., *beach*, *coast* or *forest*). It provides contextual information to help other processes such as object recognition, content-based image retrieval and image understanding (Quelhas et al., 2005, Perina et al., 2010). For instance, if images in the database are grouped into indoor and outdoor scenes, a query for an image containing a grass can be restricted to outdoor scene. However, designing and implementing algorithms that are capable of successfully recognizing image categories remains a challenging problem (Vogel and Schiele, 2004, Bosch et al., 2007a, Quelhas, 2007, Shahiduzzaman et al., 2011, Elfiky et al., 2012). This is because of illumination changes, scale variations, occlusions, large variations between images belonging to the same class and small variations between images in different classes.

As was mentioned in (Bosch et al., 2007a), there are two approaches in image classification. The first approach uses low-level features, such as global colour and texture, applied to classify small number of scene categories (Indoor vs. Outdoor). The second approach uses intermediate semantic representation to represent image content and is normally applied to larger number of scene categories.

2.1.1 Image classification based on low-level features

Early work in scene image classification (*low-level* processing) was based on low-level image features such as colour and texture, extracted automatically from the whole image or from image regions (Liu et al., 2007, Wang et al., 2001, Vailaya et al., 2001, Szummer and Picard, 1998, Oliva and Torralba, 2001), which are then

processed by machine learning classifier to infer high-level semantic label. Image classification has been first investigated for two-class classification problems (Szummer and Picard, 1998, Vailaya et al., 1998, Vailaya et al., 2001, Luo and Savakis, 2001) using low-level features, such as colour and texture. After feature extraction stage, classifiers are trained to classify images into one of two classes, such as indoor/outdoor scene.

Vailaya et al. (Vailaya et al., 1998, Vailaya et al., 2001) grouped images into semantic categories using low-level image features. Multiple classifiers were combined into a single hierarchical classifier. Images are first classified into indoor/outdoor. Outdoor images are then classified into city or landscape and then landscape is further classified into forest, sunset and mountain classes. Images are first divided into 10x10 blocks and different low-level features where examined for this task. Several classifiers have been examined to improve the indoor/outdoor classification problem. Bayesian network was proposed in (Jiebo and Savakis, 2001) to integrate low-level features with high level concepts whereas SVM classifiers are used for the same task in (Serrano et al., 2004). Also, Szummer and Picard (Szummer and Picard, 1998) addressed Indoor/Outdoor classification problem to distinguish indoor from outdoor scenes. They proposed two stages framework to extract two types of features, colour and texture, from image blocks. Each region has two classifiers, one for colour and the other for texture. Both classifiers decide whether the region belongs to indoor or outdoor scene. An image is assigned to the class using a voting approach.

In (Van Kaick and Mori, 2006), images are divided into rectangular regions to be matched. For each region, colour moments, colour histograms, edge direction

histograms and texture features are extracted. The approach computes similarity between two images based on the cost of the best pair-wise matching of regions. They show improvements in the quality of automatic image classification of outdoor scenes using images from university of Washington.

Methods based on low level features only often failed to successfully represent the high-level semantics of user perception (Liu et al., 2007, Datta et al., 2008). Semantic modelling (*mid-level* processing) uses an intermediate semantic level representation, falling between low-level image features and (*high-level*) image classification, in an attempt to narrow the semantic gap between low-level features and high-level semantic concepts (Vogel and Schiele, 2004, Fei-Fei and Perona, 2005). Next, more details about different approaches toward semantic modelling is presented.

2.1.2 Image classification based on intermediate semantic modelling

For low-level modelling, i.e. using low-level features, image content can be modeled from the whole image or from image regions. To improve the classification performance, semantic concepts found in image content can be used as cues to represent the semantic of images content. These cues are called intermediate semantic modelling which is simply named as semantic modelling (Bosch et al., 2007a). The aim of semantic modelling is to classify image regions into semantic classes. Semantic modelling requires identifying semantic concepts that appear in the image (e.g. grass, rock, and sky), and the images can be represented by the frequency of occurrence of these semantic concepts (Vogel and Schiele, 2004).

Two basic strategies can be found in literature that adopts intermediate semantic representation (Cheng and Wang, 2010): (1) use segmentation algorithm or fixed grid layout to detect semantic concepts. Image regions are then labeled with semantic concepts (2) use bag of visual words approach as an intermediate semantic representation. The first strategy identifies the semantic of image content as a set of objects or regions that appear in the image, such as *water*, *sky*, *rock* etc. Images are first segmented into regions or objects and the task is to use classifiers to assign labels to regions or objects (Vogel and Schiele, 2004, Bosch et al., 2006). The second strategy is based on the bag of visual words model. In this strategy, visual words are first generated to label local keypoints extracted using keypoint detectors in order to build the bag of visual words which further used to classify images into semantic classes.

2.1.2.1 Intermediate semantic concepts

Semantic modelling was investigated in (Bosch et al., 2006) to classify images into three different classes: *road*, *suburb* and *city*. Objects in images are first segmented, using object recognition algorithm, and manually labelled with 7 semantic concepts: *sky*, *road*, *grass*, *vegetation*, *dark house*, *white house*, and *ground*. Image regions are described using colour and texture features. Images are represented by object occurrence vectors counting the frequency of each object in an image. For image classification task, kNN classifier was used to classify new images into one of the three classes. Also, semantic information about the scene content, such as *sky* and *grass*, was incorporated with low-level features to improve the classification performance in the context of indoor/outdoor scene classification (Serrano et al., 2004).

Vogel and Schiele (Vogel and Schiele, 2004) proposed a two stages framework to categorize natural scene images based on intermediate semantic annotation. In the first stage, images are first segmented into rectangular regions using 10x10 regular grid. Nine semantic concepts, detected from 700 natural scenes, were used to annotate image regions. These concepts are [*sky*, *water*, *grass*, *trunks*, *foliage*, *field*, *rocks*, *flowers* and *sand*]. Colour and texture features are extracted from all image regions. These features are used to train support vector machines to learn the nine semantic concepts collected from the training images. Given a new image, the trained classifiers are used to annotate image regions with the nine concepts. In the second stage, intermediate semantic representation was obtained by counting the number of occurrence of each of the nine semantic concepts depicted in the image. Theses frequencies are further used for scene classification.

Cheng and Wang (Cheng and Wang, 2010) proposed a contextual Bayesian network model to integrate low-level features with semantic labels for scene classification. First, images are segmented into regions using mean-shift algorithm and low-level features are extracted from each of them. Regions are then manually labelled with the nine concepts determined in (Vogel and Schiele, 2004) then a multiple SVM classifier is trained on training image regions. For test images, segmented regions are labelled by the trained classifiers. Images are labelled with general labels resulted from two steps process: region occurrences, which is similar to concept occurrence vector in (Vogel and Schiele, 2004), and spatial relationships of the regions. For spatial arrangements of the regions, a contextual Bayesian network is proposed based on domain knowledge of the arrangements of semantic concepts depicted in images. Their model reported impressive results on natural

scene classification though some drawbacks exist. First, they employed a segmentation algorithm to segment images into regions which does not produce accurate results. Second, image regions are annotated with nine concepts though they re-annotate images again which is time consuming. Third, for new dataset, domain knowledge needs to be generated again.

An interesting work that use image classification for retrieval was introduced in (Tsai et al., 2006). They developed a system, called CLAIRE, which composed of three modules of SVMs for colour, texture and semantic concept classification for semantic label assignment. Each image is divided into 5 titles or blocks, from which colour and texture features are extracted and fed into two SVMs to learn the labels of the five tiles. Results of both classifiers are used for training semantic concept classifier.

Researchers in computer vision have recently started to make use of techniques based on text document retrieval, to represent image contents in image classification and retrieval systems (Zhu et al., 2002). The bag of words approach is one of these techniques and is very common in text-based information retrieval systems. The analogy between document and image is that both contain information. However, the main obstacle is how to extract semantic "visual words" from image content.

2.1.2.2 Semantic modelling using BOW

In recent years, local invariant features (Lowe, 2004) or local semantic concepts (Bosch et al., 2007a) and the bag of visual words (BOW) (Sivic and Zisserman, 2003) became very popular in computer vision and have shown impressive levels of performance in scene image classification and other computer

vision tasks, such as visual object recognition (Quelhas et al., 2005, Fei-Fei and Perona, 2005, Quelhas et al., 2007, Gokalp and Aksoy, 2007, Lowe, 2004, Quelhas and Odobez, 2006, Csurka et al., 2004, Ramanan and Niranjan, 2011, Lazebnik et al., 2006).

There are two main parts to build an image classification system within the BOW framework. The *first* relates to the detection and extraction of features that characterize image content at several points or patches. We refer to this part as image representation. The work described in this thesis relates to this part. The *second* part is the classifier to determine to which class an input or new image belongs to. The elements needed to build a bag of visual words are: *feature detection*, *feature description*, *visual vocabulary (codebook) construction* and *image representation*, each step is performed independently of the others (*see* Figure 2-2).

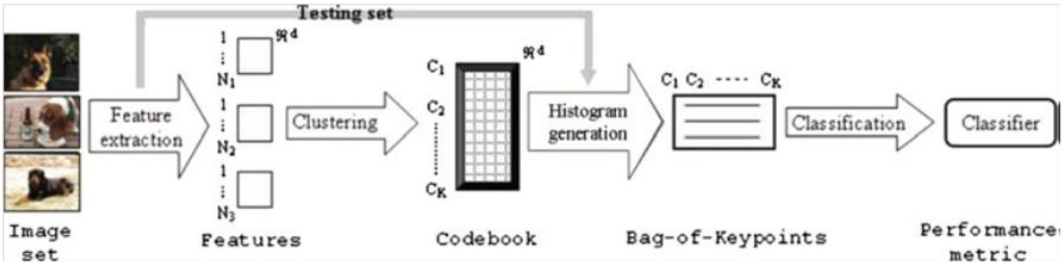


Figure 2-2: General framework for building BOW image representation (Ramanan and Niranjan, 2011)

Recently, a range of new methods have been advanced to improve the performance of the conventional BOW paradigm. We can classify these methods into three main categories:

- The *first* category attempts to improve the descriptive power of visual vocabulary (Quelhas and Odobez, 2007, Wu et al., 2009b, Nister and Stewenius, 2006, Perronnin, 2008, Wu and Rehg, 2009, Alqasrawi et al., 2011, Zhang et al., 2009b, Kesorn et al., 2011, Kesorn and Poslad, 2011, Chimlek et al., 2010, Shiliang et al., 2011, Hou et al., 2011). At this category, researchers aim is to improve the discriminative power and quality of visual vocabularies or visual words which in turn leads to a high accuracy of subsequent classification.
- The *second* category suggests using BOW, multiple features and weighting techniques to combine them (Quelhas and Odobez, 2006, Jiang et al., 2007, Alqasrawi et al., 2009, Yang et al., 2007, Khan et al., 2009, Sheng et al., 2010, Jingyan et al., 2011, Yuan and Xiaochun, 2011, Zhang et al., 2009a, Vigo et al., 2010, Khan et al., 2011, Tirilly et al., 2010).
- In the *third* category, techniques that add spatial information over the BOW have been proven to improve the performance of scene classification tasks (Lazebnik et al., 2006, Bosch et al., 2007b, Lampert et al., 2008, Battiato et al., 2010a).

Although these approaches have achieved promising results on scene image classification tasks, there is found to be no overall best approach. The complexity of natural scenes and wide variety of arrangements of entities in images means that, images with similar visual contents from two different categories are often miss-categorized, (e.g., confusion in visual appearance between *coasts* and *river/lake* classes). We believe that some of these problems could be better solved by building

a unified approach that uses knowledge from discriminative visual vocabularies, multiple image features and their spatial arrangements (*see* Chapter 4).

Next, we will list most related work which mainly used bag of visual words for scene classification/categorization task. Some of these works were applied for object categorization, which is out the scope of this theses, but worth to mention them here.

The main difference between bag of words and bag of visual words is that there is no given visual vocabulary for the image classification task and it has to be learned automatically from a training image set. Zhu et al. (Zhu et al., 2002) are perhaps the firsts who tried to represent images using an analogue of the bag of words approach from the text domain. They proposed a keyblock-based approach for content-based image retrieval. In their work, images are partitioned into equal size blocks which are then indexed using a codebook, whose entries are obtained from the block features. Each block is assigned to the index of the closest keyblock in the codebook.

The bag of visual words approach received a substantial increase in popularity and effectiveness with the development of robust salient features detectors and descriptors such as Scale Invariant Feature Transform (SIFT) (Lowe, 2004). Csurka et al. (Csurka et al., 2004) and Sivic et al. (Sivic and Zisserman, 2003) showed how to use the bag of visual words by clustering the low-level features using the K-means algorithm, where each cluster corresponds to a visual word. To build the bag of visual words, each feature vector is assigned to its closest cluster.

More informative visual words are important to bridge the semantic gap between visual information and user concepts. To add semantic information to the bag of visual word model, many algorithms are proposed targeting more discriminative visual vocabularies (Perronnin, 2008, Wu et al., 2009a, Shiliang et al., 2011).

Vocabulary Construction:

Multi-level bag of visual words was introduced in (Quelhas and Odobez, 2007) where several levels of quantization are obtained using several k-means models with different number of clusters. Their aim was to study the effect of generating visual words from coarser and finer quantization levels. SIFT descriptors are computed on regions around interest points detected by DoG detector. Images are represented by concatenating the BOW histograms associated with the corresponding visual vocabularies. They used different combinations of several levels of quantization ranging from 100 clusters to 5000 clusters. Their approach reported improvements on the classification performance evaluated on 13 scene classes.

A hybrid approach integrating unsupervised probabilistic Latent Semantic Analysis (pLSA) and discriminative classifier (SVM or *k*NN) in a unified framework was proposed by Bosch et al. (Bosch et al., 2008). Firstly, the pLSA is applied to images represented by bag of visual words to discover object categories as topics (for example, grass and houses) hence an image is modelled as a mixture of topics. Secondly, a multiclass discriminative classifiers, SVM and *k*NN, are used to learn topic distribution vectors of training images and its class label. Visual vocabularies are obtained by clustering descriptors computed from training images. Different

dense descriptors are investigated and compared to sparse descriptors. Four different datasets are used to validate their model with 8,6,13 and 15 scene class datasets.

Perronnin et al. (Perronnin, 2008) proposed the use of adapted vocabularies by combining universal vocabularies with class-specific vocabularies. In their work, a universal visual vocabulary is learned to describe the visual features of all considered image classes. Then, class-specific vocabularies are combined with the universal vocabulary to refine accuracy. Perronnin et al's work is an interesting contribution to the computation of distinctive visual vocabularies. However, their proposed adapted vocabulary does not show the differences between scene classes and it handles only one kind of image feature.

Another contribution to build discriminative visual vocabularies has been investigated by Jurie and Triggs (Jurie and Triggs, 2005), which proposes a clustering algorithm to build a visual vocabulary. Their algorithm produces an ordered list of centers. A quantization rule is used in such a way that patches are assigned to the first center in the list that lies within a fixed radius *r*, and left unlabelled if there is no such center. In Nister and Stewenius (Nister and Stewenius, 2006), local features extracted from images are hierarchically quantized in a vocabulary tree. It was shown that retrieval results are improved with a larger vocabulary.

Zhang et al. (Zhang et al., 2009b) proposed an optimization method, called category-sensitive codebook construction method, that considers the category information of keypoints as an additional term to improve visual words construction. Their approach was evaluated on PASCAL 2006 dataset.

Also, Zhang et al. (Zhang et al., 2009a) proposed concept-specific visual vocabulary construction method for object recognition. Instead of using a single descriptor, local image patches are described by different types of features. Each descriptor is clustered separately and results in different visual vocabularies. Their approach evaluated on PASCAL 2006 dataset.

To achieve high-precision and representative visual vocabulary Cluster Precision Maximization (CPM) method was proposed in (Lopez-Sastre et al., 2011) to find class representative visual words. They used a subset of Caltech 101 dataset to evaluate their approach. Ries et al. (Ries et al., 2010) proposed to build universal visual vocabulary from different domain datasets by determining their performance for scene classification task using pLSA model.

Another way of improving vocabulary constructions is to produce compact visual vocabulary. Visual vocabulary has a major impact on the efficiency of image classification using BOW models. Different approaches for building compact visual vocabularies were compared in (Gemert et al., 2010). The trade-off between vocabulary compactness and classification performance was investigated in this work. They demonstrated that compact visual vocabulary can be achieved either by reducing the size of the vocabulary or by choosing useful visual words which can be achieved by semantic vocabulary.

Four different approaches are investigated in (Gemert et al., 2010) to create compacted visual vocabulary, while retaining classification performance. The four approaches consists of (1) global visual vocabulary construction; (2) class-specific visual vocabulary construction; (3) annotating a semantic vocabulary and (4) soft-assignment of image features to visual words. These approaches were evaluated

against each other on a large dataset. The experimental results suggested that the best method depends upon application at hand. Also, in a different context, an approach to integrate different vocabularies has been introduced in (Battiato et al., 2010b) to build bag of phrases for near duplicate image detection.

Use spatial information with BOW:

There are a lot of related work for improving the discriminative power of visual word .Visual words contain limited spatial information which has been proven important for visual recognition and matching (Lazebnik et al., 2006). Many researchers have proposed algorithms to model the spatial relationships of visual words (Lazebnik et al., 2006, Battiato et al., 2009, Philbin et al., 2008).

A bag of visual words represents an image as an orderless collection of local features, without spatial information. Spatial pyramid matching was proposed by Lazebnik et al. (Lazebnik et al., 2006) as an extension to the orderless bag of visual words. The spatial pyramid divides the image into 1x1, 2x2, 4x4, etc. regions. Assuming a visual vocabulary is given; local features extracted from each region are quantized and then combined using a weighting scheme which depends on region level. Based on this approach, three different hierarchical subdivisions of image regions were recently proposed for recognizing scene categories (Battiato et al., 2010a).

To incorporate spatial dependencies between visual words, spatial pyramid co-occurrence is recently proposed by Yang and Newsam (Yang and Newsam, 2011). Spatial dependencies are characterized by two spatial predicates which consider the distance between pairs of visual words and the orientations of pairs of

visual words. Their work is inspired by two main approaches: spatial pyramid matching and grey-level co-occurrence matrix. They evaluated their spatial pyramid co-occurrence representation on 15 scene image dataset and achieved 82.51% classification accuracy.

Keypoints located in an image are normally assigned to the index of the nearest visual word based on a similarity measure. Jiang et al. (Jiang et al., 2007, Jiang et al., 2010) proposed a soft weighting method to assess the importance of a visual word to an image. They argue that nearest approach is not the optimal choice and that two keypoints assigned to the same visual word may not always equally similar to that visual word. For a visual vocabulary of size *k*, an image is represented by a vector T =[t_1, t_2,…, t_k,] of size *k* elements where the *i*th element is a soft weight for that visual word calculated by $t_i = \sum_{l=1}^{N} \sum_{j=1}^{M_l} \frac{1}{2^{l-1}} sim(j,i)$, where *N* is the top-*N* nearest visual words and M_l is the number of keypoints whose *l*th neighbour is visual word *i*. The similarity measure *sim*(*j*,*i*) corresponds to the similarity between keypoint *j* and visual word *i*. SIFT features were used based on DoG detectors to represent local features. Their approach is compared to different weighting methods. Their weighting approach has gain improvements on video concept detection problem.

A spatial weighting approach has been proposed in (Chen et al., 2009) to weight visual words based on spatial constitution of image content. In their work, DOG detector is used to find salient points in images. Each salient point is then described by a 128-dimensional SIFT feature. K-means clustering algorithm is used to created visual words from SIFT feature. To model the spatial information in

images, Gaussian mixture model (GMM) has been used to partition an image into *n* regions based on features extracted from image pixels. The contribution of each visual word to region *i* decides its weight. The Similarity between two images is measured using cosine similarity measure. Their model show better retrieval performance than the traditional BOW evaluated on part of LabelMe (Russell et al., 2008) image dataset.

Also, Elsayad et al. (Elsayad et al., 2010) proposed a weighting scheme to weight visual words according to the spatial constitution of an image content. Local interest points and edges are merged and clustered to create visual vocabulary. Colour information and spatial position of interest points are clustered using GMM. Their weighting approach is based on the contribution of salient points to the Gaussian components. For object recognition task, they used Caltech101 dataset to evaluate their weighting scheme and achieved good performance compared to the traditional bag of visual words.

Battiato et al., (Battiato et al., 2010a, Battiato et al., 2009) proposed spatial hierarchy framework that includes three different subdivisions, to build weighted BOW models. The SVM and kNN classifiers were used to learn the BOW histograms and they reported 79% classification accuracy on 15 scene classes and 67% accuracy on their previous work using the same approach but on another dataset with 13 scene classes (Battiato et al., 2008).

Shahiduzzaman et al. (Shahiduzzaman et al., 2011) proposed an improved version of spatial pyramid matching approach using scale space theory. They argue that different image appearance may have similar histograms which may affect the performance of spatial pyramid matching. Thus, they analyzed image content by

applying Gaussian filters to input images at different scales. Their approach has been validated on 15-classes dataset.

Contextual visual words has been addressed in (Qin and Yung, 2010). Their idea is to extend the traditional visual words by integrating features from a region, the features from neighbourhood regions and features from the coarser level centred around the same region. Features employed from neighbours of a patch are called context. Weighting parameters were used to control the significance of the three features. Visual vocabularies are then generated from the combined features to build binary BOW histograms. Their approach was evaluated on three datasets with 8, 13 and 15 scene classes.

Liu et al. (Liu et al., 2011) proposed regional-conditional random fields (R-CRF) to construct visual words by integrating contextual information into the bag of visual words approach for scene categorization. An image is first segmented into homogenous regions. For each region, two features are then considered: SIFT features and region contextual information. Visual words, obtained from clustering SIFT features using k-means algorithm, are used as input to the R-CRF model which model the spatial interaction between patches in the homogenous region. Their new visual words are compared to traditional visual words in scene classification task using 8 and 13 and 15 scene categories. They reported 74.5% accuracy for the 15 scene categories but accuracies for 8 and 13 scene categories were not reported.

The role of contextual information to improve the performance of BOW model has been addressed by Su and Jurie (Su and Jurie, 2011). An image is represented by a set of BOW histograms each corresponds to a semantic context. In this case, each visual word would have different frequency of occurrence for

different context. To reduce dimensionality, embedded-context BOW is proposed by selecting most discriminative context for each visual word.

Elfiky et al. (Elfiky et al., 2012) proposed an approach, called compact pyramids, to reduce the dimensionality of using spatial pyramid (Lazebnik et al., 2006) incorporated with the BOW image representation while maintain the performance on standard image classification task (15-scene dataset). Their approach is based on a clustering algorithm called divisive information theoretic feature clustering algorithm which is used for text classification.

Use multiple features with BOW:

Within the bag of visual words framework, fusion of multiple cues, such as colour and texture, still remains an active research domain (Quelhas and Odobez, 2006, Khan et al., 2011, Van de Sande et al., 2010, Vigo et al., 2010). Many studies have investigated including multiple image features within the framework of BOW. Adding colour information to the bag of visual words model can be accomplished in two ways: early fusion and late fusion. In early fusion, fused colour and shape vocabulary is constructed whereas late fusion two visual vocabularies are constructed from which histogram representation of both descriptors are concatenated (Quelhas and Odobez, 2006, Vigo et al., 2010).

Quelhas and Odobez (Quelhas and Odobez, 2006) investigated the use of colour information with traditional BOW extracted from local interest points. They addressed both early fusion and late fusion strategies (*see* Figure 2-3). Interest points are detected and described using DoG and SIFT features. Colour features are represented using the first and second colour moments computed from regions

around interest points in the LUV colour space resulted in 6-dimentional vector. For early fusion strategy, SIFT and colour features are fused before visual vocabulary construction whereas in late fusion each feature is considered independently.

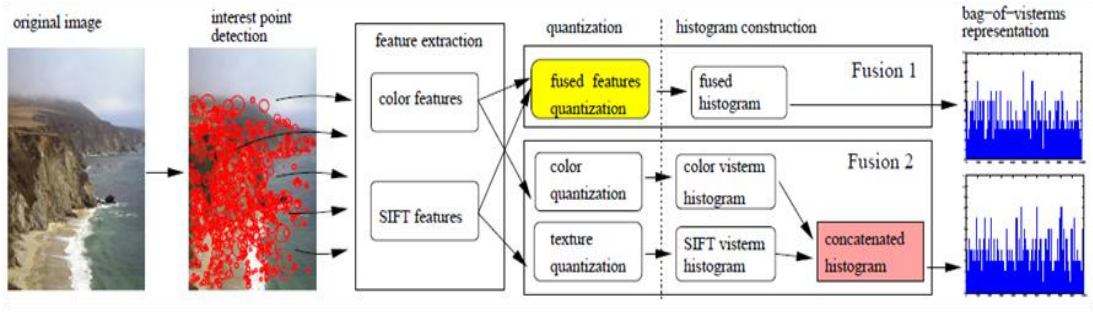


Figure 2-3: Schematic representation of the two fusion approaches. Yellow box shows fusion between features before quantization, whereas pink box shows the fusion at bag of visual words level. The diagram is obtained from Quelhas PhD thesis (Quelhas, 2007).

A weighting approach was used to merge both features. They evaluated their approach on 6 classes of natural scenes images provided by (Vogel and Schiele, 2004). They reported 66.7% classification accuracy. Although this approach has shown an improvement in classification accuracy, it has two main limitations: (1) Colour information is computed over interest regions only; and (2) No spatial information is implemented.

In (Khan et al., 2009) a novel approach is proposed to recognize object categories using multiple image features. Their model, the Top-Down Colour Attention model, considers two properties for image representation: feature binding and vocabulary compactness. The first property involves combining colour and

shape features at the local level, while the second property concerns separate visual vocabularies for each of the two features.

Recently, a new image representation approach which combines colour information and SIFT descriptors is proposed in (Khan et al., 2011). Colour information and SIFT are processed separately and combined by means of top-down and bottom-up colour attention. They used a weighting factor called attention weight to decide which features are relevant. The shape descriptor is defined as *descriptor cue* which is similar to the traditional BOW. The colour descriptor is used to modulate shape features, i.e., determines the importance of the local feature on the image representation, and is called *attention cue*. Their approach is evaluated on five standard object recognition datasets such as PASCAL VOC 2007 and 2009, Caltech-101 and Flower dataset. They have concluded that late fusion approach is favorable for classes which have colour-shape independence whereas late fusion is important for classes where shape and colour are important at local feature level.

A number of visual patch weighting methods and different configurations of BOW fused with multiple image features have been investigated by Jiang et al. (Jiang et al., 2010) for semantic concept detection in video images. The invariance properties and the distinctiveness of colour descriptors, such as rgbSIFT features, are studied recently by Sande at el. (Van de Sande et al., 2010) for object and scene classification tasks. They proposed a systematic approach to provide a set of invariance properties, such as invariance to light intensity.

In (Moulin et al., 2010), two bag of visual words generated from two different descriptors are fused in a single vector to improve the classification performance using Simplicity dataset (Wang et al., 2001). Instead of creating colour

SIFT features by concatenating SIFT features computed from the three channels of a coloured image and then quantize them into visual words (Van de Sande et al., 2010), Zhou et al. (Zhou et al., 2011) proposed to quantize SIFT features separately at each component and then concatenate them to form the final BOW. They compared the performance of both schemes on Caltech 101 dataset and PASCAL VOC 2007.

Nilsback and Zisserman (Nilsback and Zisserman, 2006) demonstrated the effect of using different visual vocabularies generated over various image descriptors in flower classification dataset. Three visual vocabularies were developed, to represent colour, shape and texture. An image is then represented by weighted concatenation of the three bag of visual words generated via the three visual vocabularies. Nearest neighbour classifier was used to classify flower images. Also, in (Nilsback and Zisserman, 2008) the same authors addressed the same problem but for larger number of flower image classes and a different classifier. A Multiple kernel classifier was used to combine four different descriptors which reported better performance over nearest neighbor classifier.

Gu et al. (Gu et al., 2011) proposed an integrated image representation based on image flatness. An image is first divided into regular grid patches and the entropy of each patch is calculated. The entropy for each patch is employed to represent its flatness, which reflects the dispersion of image pixel values, and the image flatness is obtained by summing the flatness of all image patches. For each patch, two descriptors are extracted: normalized pixels vector and SIFT features vector. Two visual vocabularies are generated, using k-means algorithm, from both features over all training images. Based on the two vocabularies, an image is represented as a

weighted integration between two histograms where image flatness is used as a weighting factor. Their image representation approach is learned using a generative model to classify natural scenes. They have reported 83.6% classification accuracy for 15-class dataset and 84.6% for 13-class dataset.

Semantic visual vocabulary:

Recently, researchers in computer vision started to consider semantic information in building visual vocabulary for BOW models. They argue that due to the traditional clustering of local features, semantic information is lost such that visual features related to the same semantics may not distribute to the same clusters.

Lei et al. (Lei et al., 2010) proposed a framework, named semantic-preserving bag of words, to learn a semantic visual vocabulary. They proposed a distance metric to measure the semantic gap between semantically similar features. Image features, e.g., SIFT features, that are located in the same region are considered relevant to each other whereas SIFT features located in regions with different semantic are considered irrelevant. All relevant SIFT features for a particular semantic label are collected from all images and then clustered to generate semantic visual words.

For semantic based image representation, using visual words are justified as their ability to group semantically similar image patches or regions such as grass or rocks thus narrowing the semantic gap. The quality of visual words can vary depending on several parameters: number of visual words, distance function used to measure the similarity between local descriptors and the clustering algorithm (Lopez-Sastre et al., 2011).

Very recently, the semantic gap has been addressed in (Kesorn and Poslad, 2011) by exploiting the bag of visual words model to represent visual content for classifying and retrieving images of athletic sports. They have focused on constructing visual words from representative keypoints to improve visual vocabulary. Also, they have detected and removed non-informative visual words that are useless to represent image content. Visual words that are semantically related are mapped to a hierarchical model, ontology model, which describes visual content using conceptual structures. Besides using representative keypoints in the clustering process, the same authors have also proposed to use spatial information of keypoints for generating semantic visual words in the vector space model (Kesorn et al., 2011). They have tested their proposed framework on the same dataset.

2.2 Image Annotation: A Review

As a result of the semantic gap between low-level image features and high-level semantics, automatic image annotation is considered a promising approach to bridge the semantic gap through extracting semantic features or concepts from images using machine learning techniques (Zhang et al., 2012). Automatic image annotation is a promising trend to understand image content. In this work, semantics refer to words used to describe an image. Having images automatically annotated with labels, images can be retrieved in the same way as text documents. Thus, annotation can facilitate image search through the use of text. It is worth to mention that in literature there are three different approaches that discuss automatic image annotation: (1) statistical approaches (2) vector space approaches and (3) classification approaches. In this thesis, approaches (2) and (3) are of our interest.

In general, automatic image annotation systems learn the semantic concepts of a set of images by associating low-level features to high-level concepts. The trained system is then used to predict a set of semantic concepts to annotate previously unseen test images. The whole process is accomplished based on visual content, the low-level features, of images and machine learning techniques. So, two major aspects are related with automatic image annotation: feature extraction and semantic concepts learning. In automatic image annotation, feature extraction can be performed at two levels:

- ***At image level:*** At this level, images are described by global features such as colour histogram which describes the visual content of the whole image. However, global features do not describe different parts of an image.
- ***At region/object level:*** First, images are partitioned into rectangular blocks or regions via fixed size grid or using a segmentation algorithm. Image regions obtained using image segmentation algorithms can be classified into four types:
    - Clustering based approaches
    - Contour based approaches
    - Statistical based approaches
    - Graph based approaches
    - Region growing based approaches

    Each block or region is represented by features extracted from it, such as colour and texture. Thus, an image with *n* blocks or regions is represented by a set of *n* feature vectors. Then, automatic image annotation is trained,

using machine learning approach such as SVM, to assign each feature vector to a pre-defined category. This is similar to the classification task.

As aforementioned, automatic image annotation at region-based requires prior image segmentation. It has been agreed among researchers that automatic image segmentation is a difficult task, computationally very expensive (Zhang et al., 2012) and leads to unsatisfied semantically heterogeneous regions (Vogel and Schiele, 2004). Thus, many automatic image annotation techniques use grid based approach to segment images into rectangular blocks (Vogel and Schiele, 2004, Mori et al., 1999, Vailaya et al., 2001, Maree et al., 2005a, Qi and Han, 2007). For this reason, this work uses grid-based image division and avoids image segmentation. However, region-based approach is not always accurate and it is difficult to decide the size of the blocks.

There are different ways to perform image annotation. According to Ja-Hwung et al. (Ja-Hwung et al., 2011), image annotation can be categorized into three types:

- *Classification-based annotation*: in this type, image annotation is treated as classification problem using multiple classifiers, such as SVM, to learn image semantic concepts.
- *Probabilistic-based annotation*: in this type, probabilistic models are developed to estimate the relation between visual image contents and semantic concepts.
- *Retrieval-based annotation*: in this type, semantic concepts of images that are semantically relevant to the query image are employed to annotate the query image.

2.2.1 Automatic image annotation techniques

For automatic image annotation, higher level semantic can be learned from image sample or image regions represented with low level features. New images are labeled with semantic labels using the trained model. Automatic image annotation can be accomplished in three ways:

- Annotate images with single label.
- Annotate images with multi labels.
- Annotate images with metadata from the web.

In the first approach, image annotation is considered as a binary classification problem. Image features are first provided to a binary classifier, such as support vector machines, artificial neural network or decision tree. The trained classifier is then used to label a new image with a semantic label. Support vector machines are the common choice for many classification problems, such as image classification, object recognition and text classification (Cusano et al., 2004, Chapelle et al., 1999, Goh et al., 2005, Qi and Han, 2007). It has an advantage over other classifiers that it achieves optimal class boundaries by finding the maximum distance between the hyperplane of classes. The hyperplane is trained to separate the samples of one class from other class. For multi-class classification and annotation problem, two common schemes are used: one-*vs*.-one and one-*vs*.-all. Thus, multiple SVMs are used to learn each class individually such that a test image is labeled with a decision fused from decisions of all classifiers.

2.2.1.1 Single-label image annotation at image level

Chapelle et al. (Chapelle et al., 1999) proposed a framework to train 14 SVM classifiers to learn 14 image semantic concepts using one-*vs.*-all paradigm. Only colour histogram is used to represent visual image content at three colour channels of HSV colour space. No image segmentation is employed in their work. Each SVM is trained on a particular semantic concept where all images belong to the same semantic concept are regarded as positive samples while the others are considered as negative. A new image is classified based on a voting approach to select the classifier with maximum probability value.

To improve the classification power of SVM, Goh et al. (Goh et al., 2005) proposed a three-level classification scheme using three different sets of SVMs, one-class, two-class, and multiclass, to annotate images with semantic classes. They developed a confidence-based approach to fuse the output of classifiers propagated at different levels based on the difference between the highest two output decisions. An image is assigned to the semantic class with highest cumulative confidence. Image annotation is performed at image level without any segmentation.

In general, most techniques that classify images into one of predefined categories can be identified as single-label image annotation techniques at image level. Thus, it is difficult to distinguish single label image annotation, at image level, from image classification discussed in (*see* Section 3.1).

2.2.1.2 Single-label image annotation at region level

Cusano et al. (Cusano et al., 2004) proposed a framework to annotate image regions with seven semantic concepts-*sky*, *skin*, *vegetation*, *snow*, *water*, *ground* and *buildings*. Their image annotation framework use multi-class SVM to assign/classify

image regions into one of the seven semantic concepts. All images used in their experiments are divided into overlapped tiles around each pixel and are labeled with the seven semantic concepts. Image regions are described by colour histogram in the HSV colour space. The classifiers are trained on random sample of tiles chosen from each semantic concept. To annotate the whole image, they proposed a threshold-based strategy to accept or reject the semantic concept based on the decision obtained from SVM classifiers.

Multilevel SVM sets are also explored in (Qi and Han, 2007) to annotate images with semantic concepts using both global features and local features extracted from the whole image and image regions, respectively. For region-based features, images are divided into 5 non-overlapping blocks. Both features are used in two different sets of SVMs. This approach also differs from the previous one in the way how predictions are fused to get the semantic label for the new image.

Single label image annotation approaches improve image retrieval by just typing the keywords related to the semantic concepts. So, there is no need to do image matching.

2.2.1.3 Multiple image annotation

Multiple image annotation refers to approaches that assign more than one label to the whole image. To make images understandable by humans it is important to represent the semantic structure of images for semantic image annotation. The choice of the semantic keywords or concepts is dependent on the domain, user needs and the application. For multiple image annotation, this work focuses on using constrained vocabulary of a small size to locally annotate natural scene images with semantic concepts.

In object recognition problems, different issues were addressed in Tousch et al. (Tousch et al., 2012). These are:

- Object detection: refers to whether an object is present in the image or not
- Object localization: refers to locating the position and scale of a specified object
- Object categorization: refers to assigning a single label, from pre-defined categories, to an image.
- Object identification: refers to who is that object is.
- Object annotation: refers to annotating an image with a list of labels selected from a controlled vocabulary.
- Region annotation: refers to annotating image regions with semantic labels.

It is important to note that categorization is much related to image annotation and semantic image retrieval. In the case of image annotation, image categorization can be used for annotation. It helps in reducing the number of possible objects that might exist in an image scene. In the case of semantic image retrieval, images that constitute or labeled with the query words are retrieved.

For multiple image annotation, it is important to decide whether to assign multiple global labels to images or to attach labels to image regions (Vogel, 2004). Automatic annotation of image regions leads to semantic image representation thus reducing the semantic gap.

As has been mentioned earlier, single label image annotation is a classification problem. In contrast, multiple labels image annotation seeks to annotate image with multiple keywords, either to image regions or part of them.

Barnard and Forsyth (Barnard and Forsyth, 2001) proposed a generative hierarchical clustering model to assign multiple labels to an image through mapping labels to regions. Their model learns the joint statistics of words and regions and build word-region co-occurrence matrix to attach words to regions. Instead of using hierarchical clustering, Duygulu et al. (Duygulu et al., 2002) proposed a translation model to map regions to words. Both models used normalized cuts algorithm (Shi and Malik, 2000) to extract regions from images.

Aksoy et al. (Aksoy et al., 2005) addressed the relations between labeled regions. They proposed a visual grammar to represent the relation between image regions to reduce the semantic gap. Bayesian classifiers are used to label image regions then the grammar model is used to classify images into scene categories.

Mojsilovic et al. (Mojsilovic et al., 2004) addressed the problem of the semantic gap by introducing semantic indicators (sky, water, skin,…etc) based on experiments with human subjects. Global and local visual features are extracted from images and quantized into regions. These regions are names by human subjects. These semantic indicators are used then for scene categorization.

Recently, Makadia et al. (Makadia et al., 2010) argue that most complicated state-of-the-art automatic image annotations techniques lack of comparisons with simple baseline measures to justify the need for such complicated models. They proposed a set of baseline methods, using colour and texture, to automatically annotate images with keywords based on nearest neighbors. For colour features, colour histograms are generated from the three components of images represented in RGB, HSV and LAB colour spaces, respectively. Gabor filters and Haar wavelets are used to represent image textures. To calculate nearest neighbors, three distance

measure are jointly combined and evaluated using *KL*-divergence, L_1-distance, and L_2-distance. For test images, predicted labels are assigned to whole image without specifying to which region the labels refer to. Many surveys on automatic image annotation are available in literature and most recently are carried out by (Zhang et al., 2012, Tousch et al., 2012).

2.3 Semantic-based image retrieval: A Review

This section provides a literature review to the previous and recent works and techniques for content-based image retrieval and particularly semantic-based image retrieval.

Since a decade, we are witnessing an incremental revolution in information technology and especially the visual information and this is due to several reasons, including: the development of Internet, wide spread of digital cameras and smart phones, image scanners, photo sharing websites, social networks. Such technologies have led to rapid increase in the number of digital images available to users. To make these images available to users, techniques for searching, organizing, indexing and retrieving images become important and a challenging problem. It is referred to this problem as image retrieval (*see* Section 3.1).

Content-based image retrieval has been introduced as a possible solution to overcome text-based image retrieval (*see* Section 3.1.1). In CBIR system, images are represented by their visual content, such as colour and texture and a query image is introduced to the system which returns all similar images. This concept of query is called *query-by-example* (QBE). But due to the semantic gap that exist between low-level image features and high-level semantic understanding of images, CBIRs

systems based on low-level features often fail to fulfill user demands (Liu et al., 2007, Smeulders et al., 2000, Datta et al., 2008, Eakins, 2002, Long et al., 2003).

In traditional CBIR, several systems have been developed for different domains and all of them have a common pipeline: extract descriptors from images and find similarities between a query image and images in the image collection. In general, Euclidean distance is used as a similarity measure between image descriptors. Another distance is the histogram intersection which compares two histograms and takes the minimum value of each two bins. Their normalized summation over all bins forms the similarity distance. A comparison study between different low-level features for content-based image retrieval has been illustrated and analyzed for small datasets in (Deselaers et al., 2008) and for web scale images in (Penatti et al., 2012). The former study showed that colour histogram can be used as a baseline for different CBIR applications. Nevertheless, they emphasized on the importance of semantic image analysis and understanding that has witnessed much work using semantic concepts (Deselaers et al., 2008). They suggested that, for better content-based image retrieval, image descriptors needs to be combined with semantic concepts or textual information.

A more recent comparative study, in the context of web images, of global colour and texture features is presented in (Penatti et al., 2012). They have evaluated 24 colour and 28 texture features based on complexities of feature extraction, distance function; and storage requirements and validation. Users are involved in their experiments to annotate the relevance of retrieved images to the query image. They only compared global colour and texture features ignoring the importance of using local features in image retrieval. They argue that local features are not

appropriate for web images due to time and complexity. Similar to (Deselaers et al., 2008), they considered global colour histogram as a baseline descriptor. Also, semantic information was not addressed in their comparisons. It is worth to mention here that web image retrieval is outside the scope of this thesis.

In this section, the focus is on recent works and techniques that involve semantic information in representing image content, which is relevant to the work presented in this thesis. Nevertheless, a short summary of literature in the early work of CBIR systems could be necessary to motivate the need for semantic image retrieval.

2.3.1 Early days of CBIR

Early work in CBIR systems has focused on extracting low-level features to represent image contents. Images are represented by their visual features as feature vectors and these vectors are compared using similarity metrics, such as Euclidean distance, similar to nearest neighbor classifier. Images with highest similarity scores are ranked first in the retrieval process. Using global image features, extracted from the entire image, is proved to be not enough to represent image contents (Liu et al., 2007). Thus, this work focuses on approaches that are based on region-based and salient-points-based image retrieval.

Many CBIR systems have used nearest neighbor approach to retrieve similar images. In QBIC system (Faloutsos and Taubin, 1993), colour histograms, moment-based shape features and texture features are used to represent image contents. Another popular image retrieval system is Blobworld (Carson et al., 1999) developed at the UC Berkeley. In Blobworld, images are segmented into regions (blobs) using

Expectation-Minimization (EM) algorithm clustering image pixel properties, colour, texture and position information. An image is represented by colour and texture features extracted from blobs rather than the entire image and the user selects a region and the system returns images composing similar regions.

In SIMPLIcity (Wang et al., 2001) image retrieval system, images are segmented into regions using wavelet-features and k-means clustering algorithm. They have developed region-matching similarity metric to match all segmented regions automatically. They claimed that pre-classification of images enhances the retrieval results. Images are first classified into two classes: graph/photograph, texture/non-texture. After classifying all images, retrieved images are returned from the same class after selecting a region from the query image, similar to the Blobworld. Their system lacked accurate image segmentation and the user has to select a representative region for the query image.

Also, Iqbal and Aggrawal (Iqbal and Aggarwal, 2002) developed another CBIR system, called CIRES, to improve image search results using image structure in combination with colour histogram and Gabor features as texture. They show that using image structure like line crossings and junctions improves the retrieval performance, particularly for man-made objects. Most image datasets used in these systems are manually annotated to measure their retrieval performance. Thus, automatic image classification could be a step to help image retrieval task.

Many approaches have been developed in the early days of CBIR. It is time consuming to list them all in this section. However, a number of excellent surveys are available in the literature, so the reader can refer to them (Rui et al., 1999, Smeulders et al., 2000, Long et al., 2003).

2.3.2 Semantic-based image retrieval

In order to derive image content features which are semantically relevant to the user's perception, researchers in computer vision have focused on developing schemes which links image regions with semantic concepts. These schemes aim at narrowing the semantic gap between low-level features and user understanding. As will be mentioned in Section 3.1.3, Liu et al. (Liu et al., 2007) categorized techniques to infer high-level semantic information into five types: (1) derive semantic concepts using Ontology, (2) derive semantic concepts using machine learning, (3) derive semantic concepts using relevance feedback, (4) derive semantic concepts using semantic templates, (5) derive semantic concepts from textual information located with images for Web retrieval. Most approaches are related to type (2) which adopt machine learning approaches to learn high-level semantic of images using visual content and textual features. Our work presented in Chapter 6 is related to this type, hence related work linked to this type will be considered in this section.

For semantic-based image retrieval using machine learning algorithms images are first segmented into regions by fixed grid size, image segmentation or by using salient points (see Section 3.2.3). The next step is to associate semantic labels with image regions or with the entire image. This step is similar to image regions annotation and also used in image classification.

Sun and Ozawa (Sun and Ozawa, 2003) proposed an region-based image retrieval approach using wavelet transform. Image regions are obtained by clustering the wavelet coefficients in the Low-Low frequency sub-band of image wavelet transform. To describe image regions, features are hierarchically extracted from

image regions from all wavelet frequency sub-bands. Finally, a weighted distance functions are used to match image regions.

Another approach for region-based image retrieval was proposed by Chen and Wang (Chen and Wang, 2002). Images are first segmented into regions, each of which is characterized by a fuzzy feature representing colour, texture and shape features. An image is then represented as a set of fuzzy sets corresponding to image regions. A new similarity measure was proposed to find the similarity between two images represented by two families of fuzzy features. Their approach was evaluated on SIMPLIcity image dataset.

To improve the performance of image retrieval, Chen et al. (Chen et al., 2005) proposed a cluster-based image retrieval which includes the similarity information between retrieved images. In cluster-based image retrieval, a clustering algorithm is applied to images retrieved that are very close the query image. The resultant clusters are displayed to the user who adjusts the model based on his feedback.

Some of the recent works on semantic-based image retrieval based on bag of visual words has been presented in section 3.1.2.2. such as the works of Nister and Stewenius (Nister and Stewenius, 2006) and Chen et al. (Chen et al., 2009). Vieux et al. (Vieux et al., 2012) proposed a similar idea to the BOW model, called bag of regions model. They proposed incremental clustering algorithm for building visual vocabularies. They have showed promising results on some of the public datasets.

2.4 Summary

This chapter has introduced an overview of the work related to scene CAR. The chapter has provided a wide range of techniques and ideas proposed in the literature, which are related to the work presented in this thesis. The chapter has focussed mainly on semantic modelling techniques that use local semantic concepts and bag of visual words model, which aims to reduce the semantic gap. An increasing number of published works uses local features with bag of visual words to represent image content which in most cases outperform global features. The chapter has presented different approaches to improve the bag of visual words model, including visual vocabulary construction, spatial information, multiple features and semantic visual vocabularies were introduced. Most of these improvements were applied to the problem of scene/object classification problem. Nevertheless, these approaches can be applied to automatic image annotation and semantic-based retrieval. In chapters 4, 5 and 6, different techniques are proposed that use bag of visual words and local semantic concepts to improve the performance of scene CAR.

Chapter 3

Background

This thesis uses techniques from different fields, covering topics from information retrieval, computer vision and machine learning. It is beyond the scope of this thesis to review all of these fields in details, so this chapter attempts to describe the techniques and basic concepts used throughout this thesis.

The chapter begins by reviewing techniques in image retrieval; firstly textual-based image retrieval, content-based image retrieval and the semantic gap. This is followed by a discussion on content-based image description, in particular global features, region-based image description and techniques based on interest points. Next, bag of visual word model and spatial layout approach is discussed. These techniques can be used in scene classification, annotation and retrieval. Finally, the chapter looks at most common machine learning approaches followed by a discussion on the evaluation methods used throughout this thesis.

3.1 Image Retrieval

Regardless the way how images are described or represented image retrieval is a computer system for browsing, organizing, searching, and retrieving images from an image collection. Image retrieval has a long history since 1970's which has been studied by two research communities: database management and computer vision. In the literature, there exist three periods of times were image retrieval have witnessed major advances.

In the *first* period, traced back to the late 1970's, image retrieval was to annotate the images with keywords and then image retrieval systems use text-based database management systems to retrieve images (Rui et al., 1999) (Section 3.1.1). In the *second* period, early 1990s, visual content of images were employed to index images for storing, browsing and retrieval. The idea is to, automatically, extract primitive features including colour, texture and shape from images that can be used in CBIR systems (Section 3.1.2). Content-based image retrieval algorithms to search images by its visual content fail to accurately relate image semantics[1] to its visual features. This problem has been well known as the *semantic gap*, the *third* period (Smeulders et al., 2000, Datta et al., 2008) (Section 3.1.3).

3.1.1 Textual-based image retrieval

Text based retrieval of images has been widely used where images are indexed by text terms and retrieved by matching terms in a query with those indexed in the database. Due to its simplicity, text based approaches can be easily scaled up to deal with billions of images. However, it has two main difficulties: (1) manual labeling of large collection of images has become time consuming and impractical,

[1] This work uses semantics and high-level semantic interchangeably

(2) the gap between image content and subjectivity of human perception, i.e. different people may annotate same image with different labels (Rui et al., 1999). Moreover, text annotations often carry little information about images' visual features or content. When users wish to retrieve images of similar visual content, a pure text based approach becomes inadequate or fail to do so without understanding the content. It is difficult for a user to give a low-level description of what image she/he if looking for (Datta et al., 2005).

For instance, consumer magazines use traditional or digital image libraries as sources for images they publish in their issues. Images in such libraries are described and annotated with keywords that intended to capture the objective and subjective of aspects of each picture. It is found that searching images with keywords is a difficult task because of synonyms, trial and error with different keywords to reach correct hits, etc (Parker, 2004). There is no guarantee that two different persons generate the same label for one image.

In some way, to search images in an image collection people may find it easier to find a picture they want by looking through the collection and making matches with the one they have in their mind, than to use keyword or textual description which fails to capture it (Datta et al., 2008).

3.1.2 Content-based image retrieval (CBIR)

Content-based image retrieval (CBIR) concerns with technologies and techniques that organize/index image collections by their visual content to allow efficient browsing, searching, and retrieval (Datta et al., 2008).

Most CBIR systems composed of three stages: *feature extraction*, *indexing* and *retrieval* design. Visual content, as a result of feature extraction process, is domain dependant. General domain visual content includes colour, texture, shape, etc. Domain specific content includes objects like faces which involve knowledge from the domain (Long et al., 2003). Many projects have been revealed to study the feasibility of retrieving images from image collections using low-level features (visual content).

IBM Query by Image Content (QBIC)(Faloutsos and Taubin, 1993) is the first commercial CBIR project which supports queries based on example image, sketches and drawing. Among many CBIR systems include *Virage* (Bach et al., 1996), *RetrievalWare* (Dowe, 1993), *Photobook* (Pentland et al., 1996), *VisualSEEK* (Smith and Chang, 1997a) and *WebSEEK* (Smith and Chang, 1997b), *Netra (Ma and Manjunath, 1997)* and *MARS* (Huang et al., 1997b). These CBIR systems are different in the way how they measure similarities between query image and images in the collection and the type of features used to represent image content.

A general framework for CBIR system is depicted in Figure 3-1. More details in visual content description will be introduced in Section 3.2. A comprehensive survey paper that review techniques and work in content based image retrieval prior 2003 was conducted by (Rui et al., 1999, Smeulders et al., 2000, Long et al., 2003).

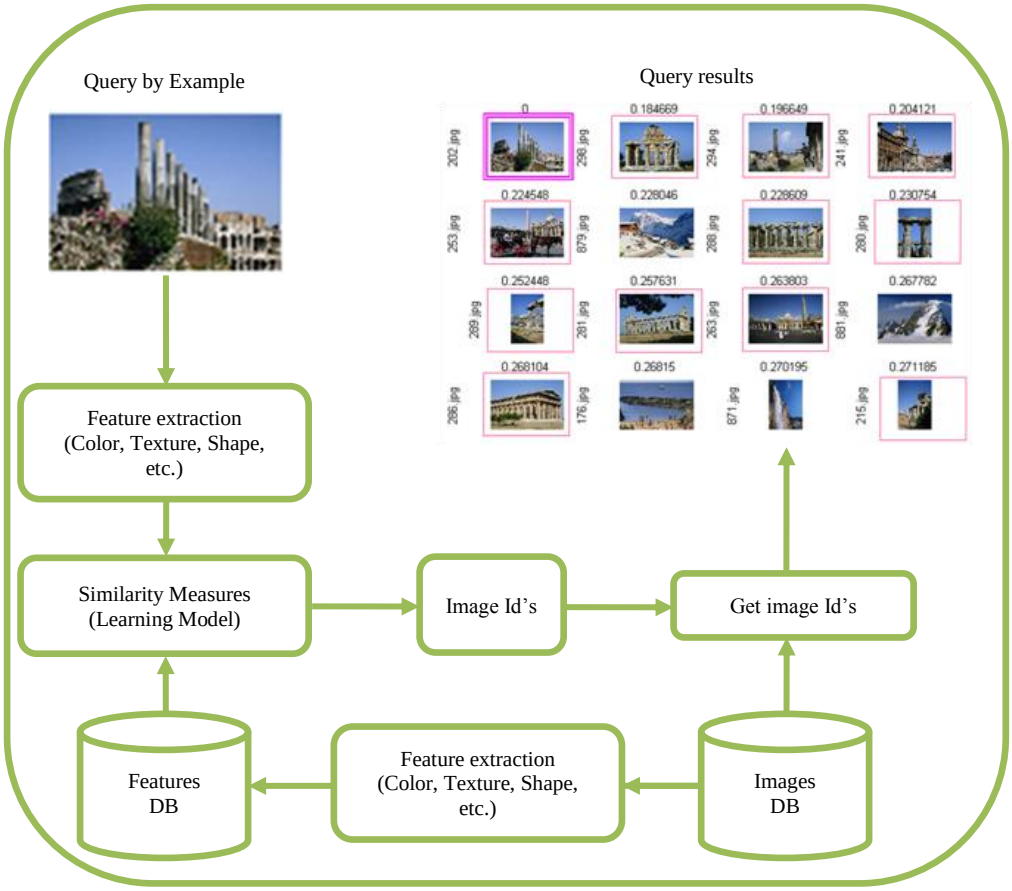


Figure 3-1: A general CBIR framework. Numbers in the query results represent the degree of similarities between the query image and images in the database. Images in the red square are images that are most similar to the query image, i.e. from same category.

3.1.3 The semantic gap

This problem of the semantic gap has been addressed in the last decade. According to Smeulders et al (Smeulders et al., 2000), semantic gap can be defined as:

"The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation"

A user searches for images containing certain objects or a certain scene. Image content, on the other hand, is characterized by data-driven features which are difficult to associate with it. With all promising techniques that are available in the literature, the semantic problem has not been solved yet, though much research has been done to link image features with high level semantics. Two images of the same semantic concept may have different visual features is shown in Figure 3-2.

Figure 3-2: An illustration of the semantic gap.Both images share the same semantic concept of beach but they have different visual features such as colour, texture etc. This figure has been adopted from (Rasiwasia, 2011).

Scene/object recognition and annotation have been active research topics to challenge the problem of semantic gap. Finding a concept pertaining in images makes the content-base image retrieval systems able to understand visual content, i.e. referred as image understanding. On the other hand, image annotation allows for image search using text. Image annotation can be thought as a set of concept

detections. Therefore, automatically annotating images with keywords makes searching images by text more semantically than by CBIR (Datta et al., 2005).

To this end, do visual features in images carry information about their semantic concepts or not. Automatic image/object classification into general types, such as indoors or beach, using machine learning approaches has been shown very useful for categorization of images into semantic classes. It can be used as a filter when searching for particular scene image or object to facilitate fast retrieval (Eakins, 2002, Wang et al., 2001). For example, SIMPLIcity is one of the CBIR systems that addressed the semantic gap by using semantic classification methods to classify images into semantic categories (Wang et al., 2001).

Liu et al (Liu et al., 2007) identified the research work in narrowing the semantic gap into five categories (1) use ontology concepts to define high-level semantics (2) use machine learning to link visual features with image concepts (3) use relevance feedback to improve retrieval results by learning user intention (4) use semantic template and (5) use HTML text next to images available in the WWW to infer their semantic.

A survey of techniques for retrieval of images by semantic content, high-level semantic and recent comprehensive survey of recent achievements in the topic of content-based image retrieval prior to 2008 has been conducted by Eakins (Eakins, 2002), Liu et al (Liu et al., 2007) and Datta et al (Datta et al., 2008), respectively.

3.2 Content-Based Image Description

In computer vision, many tasks require feature extraction as an important step. Digital images are generally stored as two dimensional matrices. A digital colour image of size 400 x 200 contains 80,000 pixels and each pixel value is specified by three 8-bit integers (or one 8-bit for grey images) between 0 and 255. If we consider this large amount of numerical values as the features of that image, memory and computation cost become expensive. Moreover, some specific tasks in computer vision require to process pixels of specific parts of an image. Therefore images commonly need a pre-processing step to extract useful information from image pixels. They need to be compacted from pictorial information (visual information) into feature values (numerical quantities).

In general, there are two main types of visual features: Global features and local features. For global features, image visual content is represented by a single global feature vector extracted from the entire image. In the case of local features, an image is partitioned into regions or objects and each region or object is represented by a local feature vector.

In this section, techniques which are related to this thesis and commonly used for extracting features from images will be presented.

In the early years of research on CBIR, global descriptors were the main choices for image description. Colour features, such as colour histogram (Swain and Ballard, 1991), colour moments (Stricker and Orengo, 1995), colour coherence vector (Pass and Zabih, 1996) , and texture features, such as Gabor filter features

(Daugman, 1988) and wavelets transform features (Daubechies, 1990) have been widely used as global features for representing images in the CBIR systems.

3.2.1 Colour features

Colours are defined on a specified colour space which is an important aspect of specifying colour features. To represent the colour information in images colour space should be selected first. The selection of the colour space depends on its uniformity. In colour vision, there are many colour spaces to represent colours in images, such as RGB, LAB, LUV, HSV (HSL), YCrCb and HMMD (Manjunath et al., 2001). RGB has been shown not efficient for image search and retrieval and not related to high-level semantics (Manjunath et al., 2001, Liu et al., 2007). The HSV colour space is widely used in image analysis and representation due to its uniformity. Therefore, this work will use HSV colour space to extract colour information.

The Hue, Saturation and Value (HSV) colour space is a non-linear transform of the RGB-cubic. Its components correspond to the categories of human colour perception making HSV colour space more suitable for analyzing visual perception. The hue (H) represents the colour in its pure, such as green, red and blue. The saturation (S) corresponds to how saturated the colour is by adding white to the pure colour. The value (V) corresponds to illumination of the colour (*see* Figure 3-3). The RGB values can be transformed to HSV values according to the following formulas (Yu et al., 2002):

$$H = \arctan \frac{\sqrt{3}(G-B)}{(R-G)+(R-B)} \tag{3-1}$$

$$S = 1 - \frac{\min\{R,G,B\}}{V} \tag{3-2}$$

$$V = \frac{(R+G+B)}{3} \tag{3-3}$$

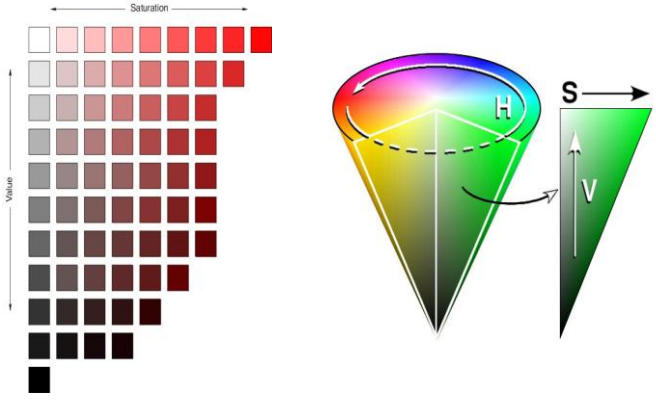


Figure 3-3: HSV colour space

Colour moments

Colour moments (Stricker and Orengo, 1995) have been proved to be successful in representing colour distribution in images. The first three moments are defined as follows:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} f_{ij} \tag{3-4}$$

$$\sigma_i = \left(\frac{1}{N} \sum_{j=1}^{N} \left(f_{ij} - \mu_i \right)^2 \right)^{\frac{1}{2}} \tag{3-5}$$

$$s_i = \left(\frac{1}{N} \sum_{j=1}^{N} \left(f_{ij} - \mu_i \right)^3 \right)^{\frac{1}{3}} \tag{3-6}$$

where f_{ij} is the value of the i-th colour channel of the image pixel j, and N is the number of pixels in the entire image or image region. So, an image region with N pixels will be represented by a feature vector of 9 values which makes the choice of colour moments preferable due to its compactness.

Colour histogram

Colour histogram has been widely used o represent colour distribution in an image. It is easy to compute and robust to translation, rotation about the view axes. For an image with three colour components, specified by a particular colour space, the histogram can be calculated for each component. To form a histogram, pixel values of a particular component are quantized into fixed bins such that the number of pixels falling into each bin is calculated. The disadvantage of colour histogram is that spatial information of pixels is ignored which makes different images to have same colour histogram. To include spatial information to the colour histogram, colour coherence vector and colour correlogram were proposed in (Pass and Zabih, 1996, Huang et al., 1997a).

3.2.2 Texture features

Texture is another important characteristic of image content since it plays an important role in human visual perception. There is no precise and agreed definition

for image texture, though there are number of different texture definitions attempted in the literature. Texture refers to the spatial distribution of grey-level pixels with properties of homogeneity of one or several primitives in an image (Abbadeni, 2011). Textural features that can be recognized by human being include coarseness, contrast, uniformity, roughness, and directionality. Tree barks, clouds, water, bricks, and fabrics are examples of texture.

There are two classes of methods in texture analysis: (1) spatial domain methods and (2) frequency domain methods. Frequency-based methods represent texture features based on the analysis of spectral density function. Frequency-based methods include discrete cosine transform (DCT), Fourier transform and Gabor and the wavelet transform. Methods in the spatial domain can be statistical methods, structural methods or hybrid methods (Abbadeni, 2011).

Texture representations include the grey-level-co-occurrence matrix (Haralick et al., 1973), Tamura representation (Tamura et al., 1978), SAR/MRSAR texture models (Mao and Jain, 1992), Gabor functions (Turner, 1986), wavelets (Daubechies, 1990), and local binary patterns LBP (Ojala et al., 1996) among many others.

However, different approaches, i.e. local descriptors, have been proposed later as researchers started to realize the limitations of global descriptors, especially for applications where a particular object in the image is of interest. We have witnessed a major shift in image representation from global features to local features and descriptors such as salient point, region-based features and spatial features. Local description approaches often choose parts from the images firstly, and then calculate descriptors for each individual part.

Region choosing can be grouped into three categories: (1) *fixed partitioning*, (2) *segmentation* and (3) *salient points*. Each category will be described in the following section. In fact, global description can be considered as a special case of region choosing, where the entire image is chosen as the region for feature extraction.

3.2.3 Region-based image representation

To allow for more efficient and precise image representation, several authors have proposed methods which use region-based representations. Region-based image representation (RBIR) has been considered as an extension to the classical content-based image retrieval: instead of extracting features from the entire image, RBIR systems divide an image into a number of regions on which individual features, such as colour and texture, are computed; theses features are called local features. RBIR aims at reducing the semantic gap between low-level features and high-level semantic concepts and adds spatial information to image representation. As such, changes in an image part affect only of the representation components which makes image representation robust to partial occlusion. To handle image classification task, image regions are classified into intermediate semantic concepts, as a result of supervised region classification step, which is used to obtain the final image classification (Vogel and Schiele, 2004)..

In literature, image regions could refer to:

- Rectangular regions resulted from partitioning an image into a fixed size blocks (Zhu et al., 2002, Vogel and Schiele, 2004, Vogel et al.,

2007, Boutell et al., 2004, Ghoshal et al., 2005, Luo et al., 2006, Wang et al., 2010).

- Segmentation of an image into homogenous regions or objects (Carson et al., 1999, Li et al., 2000, Deng and Manjunath, 2001, Liu et al., 2008, Spyrou et al., 2008, Van Kaick and Mori, 2006, Akbas and Ahuja, 2010).
- Densely, randomly or sparsely sampling an image into regions (Mikolajczyk and Schmid, 2005, Quelhas et al., 2007, Lazebnik et al., 2006, Lowe, 2004, Battiato et al., 2010a, Bosch et al., 2007a, Csurka et al., 2004, Ramanan and Niranjan, 2011, Liu et al., 2011).

Fixed partitioning

In fixed regions approach, an image is divided into blocks or regions of fixed size and features, such as colour and texture, are extracted from each region separately. The features in this approach encode spatial information about colour or textures at the cost of generating larger feature vectors. For example, an image with 25 regions will produce 25 × descriptor's length (*see* Figure 3-4(a)). The image representation is then the collection of all these local features.

Segmentation

Image segmentation can be defined as the process of partitioning an image into several "homogenous" regions based on the similarity of pixel features. There are two main methods to accomplish this task: Unsupervised image segmentation

and supervised image segmentation. In the first approach, an image is delineated into regions automatically without human intervention whereas in supervised segmentation it requires human input and intervention. The results of segmenting two images into regions are shown in Figure 3-4(b). When searching for an object, it would be most advantageous to do complete object segmentation. Object segmentation for broad domain is not likely to succeed and is still hard and application dependent (Smeulders et al., 2000, Penatti et al., 2012). Image segmentation will not be used in this thesis. It is mentioned here as one of region categories.

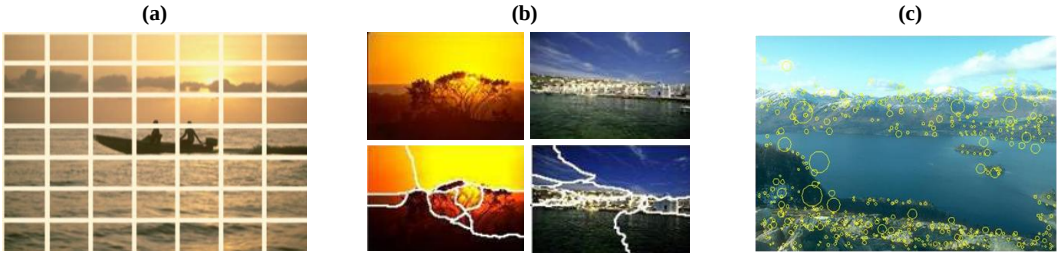


Figure 3-4: Image regions are sampled based on **(a)** fixed partitioning (Luo et al., 2006) **(b)** segmentation **(c)** salient detection using DoG detector.

Saliency

The new trend in image representation is towards the use of patch-based representation. A patch is a small region centered on a pixel and described by its local visual features. For simplicity, we can refer to image patches as regions. As it is mentioned before, image regions (patches) can be sampled densely (Fei-Fei and Perona, 2005, Jurie and Triggs, 2005), randomly (Vidal-Naquet and Ullman, 2003,

Maree et al., 2005b) or sparsely detected using various feature detectors such as Difference of Gaussian DOG (Lowe, 2004, Mikolajczyk and Schmid, 2005), *see* Figure 3-4(c).

Sampled regions are characterized using some descriptors such as colour, texture or local features which will be described next. Once image regions are sampled and described, an image is usually represented using a well known technique, the bag of visual words BOW (Csurka et al., 2004). This approach will be presented in this chapter and will be used throughout this thesis to represent image content.

3.2.4 Interest point detection and description

Image feature set extracted from an image needs to be relevant for scene/object classification while providing invariance to changes in illumination, differences in viewpoint and shift. Approaches on interest image features or descriptors can be based on points, blobs, gradients, colour, texture, or any combination of these. The interest points usually correspond to image structures that are considered important. The benefit of using local features for scene and object representation is to avoid image segmentation (Tuytelaars and Mikolajczyk, 2008). As has been mentioned before, local interest features approaches can be extracted sparsely, randomly or densely (Mikolajczyk and Schmid, 2005). In this thesis sparse representation will be used to extract local information from images due to its efficiency in image classification tasks (Quelhas et al., 2007).

Sparse representations involve two steps:

1) *Detection*: localize interest regions- usually called interest points or keypoints that contain distinctive information in their surrounding area and should be stable or invariant to geometric transformations, i.e. it ensures that given an image and its transformed version, the same image points will be extracted from both images (Quelhas et al., 2007).

2) *Description*: compute local image features on those local keypoint areas. These features should be compact and distinctive. These keypoints are assumed to be stable and more reliable and informative about local image content (DALAL, 2006, Quelhas, 2007).

Local keypoint detectors and descriptors were originally developed for point-to-point matching between two images in matching problems (Lowe, 2004) and more recently have been adopted in scene recognition (Quelhas et al., 2007, Fei-Fei and Perona, 2005, Bosch et al., 2006, Battiato et al., 2010a, Van de Sande et al., 2010, Alqasrawi et al., 2011), image annotation (Fergus et al., 2005) and video retrieval (Sivic and Zisserman, 2003).

Several salient point detectors exist in the literature. They differ by the amount of invariance they mathematically ensure and what kind of property they exploit to achieve invariance (Mikolajczyk and Schmid, 2005, Lowe, 2004). Generally speaking, the output of salient point detectors is a list of coordinates of all detected keypoints in an image. For each detected point the detector also describes some characteristics of the area around that point, such as orientation and scale.

For scale invariance, the detector deals with locating the same keypoint and associated area after image resizing or a change in camera zoom. To achieve scale

invariance, the detector determines the scale at which the local structure in an image has the highest response. This can be achieved using scale-space theory, proposed by Witkin (Witkin, 1987), by analyzing the response of the keypoint detector across scales. For digital image, the scale space representation is a set of images at different scale levels. It can be constructed by applying a smoothing kernel to the input image followed by a re-sampling of the image.

To achieve a rotation invariant representation of a local keypoint, two ways are possible: use orientation invariant descriptor or to compute a consistent orientation for the keypoint area which remains invariant when the keypoint area's image content is rotated. In this case, each local keypoint area is represented relative to this orientation and thus achieves invariance to image orientation (Quelhas, 2007).

In literature, there are several keypoint detectors which have been used in different computer vision tasks (Mikolajczyk and Schmid, 2005). Commonly used keypoint detectors include:

- Harris Corner Detector (Harris and Stephens, 1988)
- Difference of Gaussians Detector (DoG) (Lowe, 2004)
- Saliency Detector (Kadir and Brady, 2001)
- Maximally Stable Extremum Regions (MSER) (Kadir and Brady, 2001)

Regarding the computation of the descriptor over the local image regions surrounding the located keypoints, many approaches have been tried. Popular local descriptors include Scale Invariance Feature Transform (SIFT) (Lowe, 2004), GLOH (Mikolajczyk and Schmid, 2005) and SURF (Bay et al., 2006). Mikolajczyk and

Schmid (Mikolajczyk and Schmid, 2005) and Tuytelaars and Mikolajczyk (Tuytelaars and Mikolajczyk, 2008) compared several different local descriptors and showed that Scale Invariant Feature Transform (SIFT) (Lowe, 2004) descriptors perform the best when compared to other approaches, however the design of efficient image descriptors remains an open research subject. Therefore, SIFT descriptors will be used in this work without making any comparisons to other interest point detection and description approaches. Hence, SIFT descriptor will be described in more detail here.

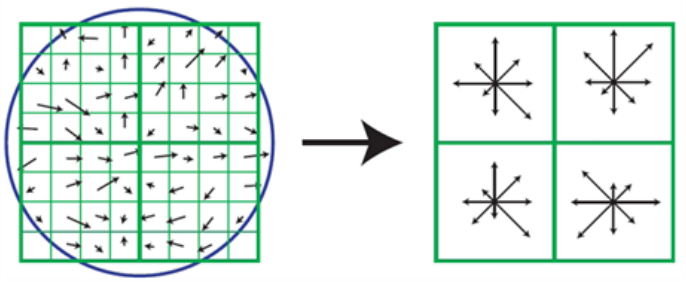


Figure 3-5: Illustration of SIFT feature detector, which consist of histograms of oriented gradients (Lowe, 2004).

SIFT computes local histograms of image gradients. It is mainly dedicated for gray-level images. Lowe (Lowe, 2004) uses the keypoints located at maxima and minima of the difference of Gaussian (DoG) keypoint detector to vote into orientation histograms with weighting based on gradient magnitude. Each SIFT descriptor is a histogram of gradient orientations computed over a Gaussian-weighted window around an interest point or keypoint. As depicted in Figure 3-5,

given a patch , SIFT descriptor is composed of a grid of 4×4 =16 histograms of 8 bins, resulting in a feature vector of length 4×4×8=128D. Each bin represents the magnitude for a particular orientation of the gradient in the cell being considered. This magnitude is weighted by a Gaussian function, indicated by the overlaid circle centered on the keypoint.

As aforementioned, SIFT features are dedicated for gray-level images. Recently, SIFT features has been extended to colour images by extracting SIFT features from each colour channel respectively and then concatenate the obtained descriptors (Van de Sande et al., 2010). For example, SIFT features are extracted over the three channels of RGB colour space and the three descriptors are merged together to obtain a final representation.

3.3 Bag of Visual Words (BOW)

The use of bag of visual words representation has become a standard choice for many computer vision tasks. The bag of words approach had proved very successful in textual analysis, and particularly for text categorization task (Joachims, 1998), where a document, d_i, is represented by a set of words, w_i, taken from predefined vocabulary. In text categorization, bag of words aims to discover the topic of a document given the words therein. This approach has been adapted to solve computer vision problems. To represent an image, Csurka et al. (Csurka et al., 2004) proposed to collect local features into unordered sets, called bag of features, for image categorization task.

The bag of visual words, illustrated in Figure 3-6, approach aims to convert sets of arbitrary elements to a fixed size feature vector. To obtain the bag of visual

words representation some keypoints in the image are first generated and descriptors are subsequently computed, as discussed in the previous section. BOW summarizes entire image based on its distribution of visual word occurrences. It turns bags of different sizes into a fixed length vector.

Figure 3-6: The bag of features approach, consists of an unordered set of local appearance descriptors (courtesy of Fei-Fei, http://vision.stanford.edu/).

The number of keypoints found in an image can be very large, thousands of keypoints in some images using difference of Gaussian approach. So, it is not appropriate for image classification, annotation and retrieval methods to process such amount of keypoints per image directly. Vector quantization is a process of grouping similar keypoints into the same class and different ones into other classes. This can be done using a clustering algorithm, such as the K-means algorithm. Each class represents the visual word and all classes constitute the visual vocabulary. Each image is represented as a histogram of the occurrence of each visual word in the

image. This is analogues to the bag of words aforementioned. More details about building BOW will be discussed in Chapter 4.

3.4 Spatial Layout

Spatial location is useful in region classification. For example, sky and sea concepts may have similar features, such as colour and texture, but their spatial locations are different as sky concept usually appears at the top of an image whereas sea at the bottom. Spatial information is important in deriving semantic features. Spatial locations are generally defined as top, middle and lower according to the location of the region in the image (Liu et al., 2007).

The bag of visual words model does not preserve any spatial information. To model spatial information, spatial pyramid scheme was proposed by Lazebnik et al. (Lazebnik et al., 2006) and works as follows. An image is divided into L different levels of a pyramid, such that $L = 0,1,..., L-1$. The level l refers to a $2^l \times 2^l$ equally spaced grid on the image. The procedure is illustrated in Figure 3-7.Thus, level 0 of the spatial pyramid is the entire image and level 1 is 2×2 cells (regions) etc. A histogram is generated for each cell in a level l, which results in 4^l histograms for this level. All histograms of each level are concatenated for a final representation resulting in a feature vector of size 4^lK, where K is the number of visual words. The similarity between two images is then measured by measuring the similarity between the concatenated histograms of both images. Another way of preserving spatial information is the work introduced by Dalal and Triggs (Dalal and Triggs, 2005). They have developed histogram of gradient Orientations (HOG) features, for pedestrian detection, by dividing an image into cells and computing a histogram of

gradient orientations for each cell. These local histograms are then concatenated into a single feature vector.

Figure 3-7: Illustration of the spatial pyramid scheme. The original image is from Vogel's dataset (Vogel and Schiele, 2004) and decomposed into two levels (middle and right). For each cell a separated histogram is computed.

3.5 Machine Learning

Image classification is typically applied for either automatic annotation, or for organizing new images into categories for the purpose of retrieval. Classification is usually categorized into two main approaches: discriminative and generative approaches. In discriminative classification methods, classification boundaries or posterior probabilities of classes are estimated, e.g. SVM and decision trees (Datta et al., 2008).

Discriminative classifiers aim to model the difference between categories. They try to find classification boundary separating object classes. Support vector machines and nearest neighbor classifiers are examples of discriminative classifiers.

3.5.1 Support Vector Machines (SVM)

The foundations of Support Vector Machines (SVM) have been introduced by Vapnik et al (Vapnik et al., 1996) for binary classification. It was mainly

developed for classification problem, and then have been extended to regression problems (Vapnik, 2000). In this work, SVM will be used for the classification task (Chapter 4) which is the standard choice in the scene classification literature (Fei-Fei and Perona, 2005, Battiato et al., 2010a, Farinella and Battiato, 2010, Liu et al., 2011, Khan et al., 2009, Vogel et al., 2007, Vogel and Schiele, 2004, Lazebnik et al., 2006). In this task, the goal is to generate a classifier that will work well, i.e. generalize well, on new data or testing data. This thesis uses LIBSVM package (Chang and Lin, 2011), dedicated for Matlab, which uses the built-in one-versus-one approach for multi-class classification.

In the simplest form, given data points represented as *p*-dimensional vectors, the SVM classifier tries to find a hyperplane which separates these points into two-class data with maximal *margin* (maximizes the distance between the margin and the nearest data point of each class). The margin is defined as the distance of the closest training point to the separating hyperplane (Gunn, 1998). There are many hyperplanes that might separate the data. The hyperplane to chose is the one that represents the largest separation. Figure 3-8 shows two-class data which can be separated by many liner classifiers, but only one is considered that maximize the margins (the green line) and the linear classifier is known as a maximum margin classifier.

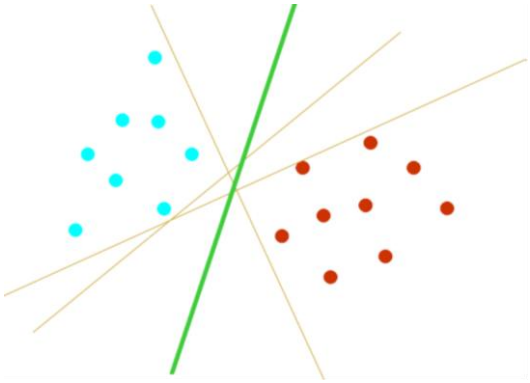


Figure 3-8: Optimal separating hyperplane (Gunn, 1998)

In other words, given n data points T of the form (Giuliodori, 2011)

$$T = \left\{ \left(x_i, y_i \right) / \, x_i \in \Re^p, y_i \in \{1,-1\} \right\}_{i=1}^{n} \tag{3-7}$$

where y_i takes the values 1 or −1, corresponding the class to which the vector x_i belongs and each vector x_i is *a* p-dimensional vector (a list of p real numbers), which can be considered as a vector that represent image i in the image collection. The goal is to find a hyperplane that divides the points of class 1 from those points of class −1. The points that lie closest to the hyperplane on each side are called support vectors. The original SVM model developed by Vapnik (Vapnik et al., 1996) was a linear classifier where a simple hyperplane is not efficient to provide discrimination. To provide non-linear decision functions in SVM, kernel functions are used. So, SVM decision function has the following form (Giuliodori, 2011):

$$f\left(x\right) = \sum_{i=1}^{N_s} \alpha_i y_i K\left(s_i, x\right) + b \tag{3-8}$$

where N_s is the number of support vectors, α_i the learned weight of the training point s_i, y_i the class label of s_i (+1,−1), $k(s_i, x)$ is the value of a kernel function for the training sample s_i and the test sample x. For those points with $\alpha_i>0$ are called support vectors. The value of f for test point x is negative if x belongs to class −1 and positive if belongs to class +1.

Many kernels are available in the machine learning literature. The common ones are linear kernel (dot product) and non-linear such as polynomial, sigmoid and radial basis function (RBF).

3.5.2 k-Nearest Neighbor (kNN) Classifier

The *k*-nearest neighbor (*k*NN) classifier is a non parametric lazy learning algorithm, also known as instance-based learning algorithm. Non parametric means it does not make any assumptions on the data distribution. It is lazy because training data is needed during the testing phase in contrast to the SVM where all non support vectors can be removed without any problem (Witten and Frank, 2005).

*k*NN determines the decision boundary locally. For *1*NN, each document is assigned to the class of its closest neighbor. For *k*NN, each document is assigned to the majority class of its *k* closest neighbors, also known as majority voting, where *k* is determined based on experience or knowledge about the classification problem at hand. The performance of the *k*NN classifier algorithm also depends on the value of *k*, the number of nearest neighbors of the test document. In this classification approach, for a test document *d*, it is expected to have the same label as the training documents located in the area surrounding *d*.

Computing the distance between two documents, represented by two feature vectors of numeric type, is a straightforward process. The standard Euclidean distance is commonly used in *k*NN classifier algorithm. The distance between the vector $a_1,a_2,\ldots,a_n$ and the vector $b_1,b_2,\ldots,b_n$ (where n is the number of features) is defined as (Witten and Frank, 2005):

$$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\ldots+(a_n-b_n)^2} \tag{3-9}$$

In this thesis, the nearest neighbor approach will be used to annotate image regions with semantic labels. The nearest neighbor approach is a KNN classifier with K=1.

3.5.3 k-Means Clustering

k-means is one of the simplest clustering algorithm, which aims to partition the points or vectors into *k* groups such that it satisfies the condition that the sum of squares from points to the assigned cluster centers is minimized. In this algorithm, the aim is generate *k* clusters or centroids. Suppose we have a pool of *n* points of *d* dimensions. First, the number of clusters should be specified in advance: this is the parameter *k*. Then *k* random points are chosen from the pool to be cluster centers or centroids. All other points are assigned to their closest cluster based on the Euclidean distance metric. Next, the mean of all points in each cluster is calculated to generate new centroids. These new centroids become new centers for their respective clusters. This process is repeated until the same points are assigned to each cluster in every round, i.e. all clusters converge. A new point is assigned to the cluster number of the centroids closest to it (Witten and Frank, 2005). *K*-means algorithm can be summarized as follows:

1) *Choose k random points from the pool of points to initialize the k_i to be the mean (centroid or centre) of each cluster.*
2) *For each point in the pool, assign it to the closest centroid (represented by k_i).*
3) *For each k_i, recalculate it based on the points that are currently assigned to it.*
4) *Repeat steps 2-3 until k_i converge.*

3.6 Evaluation Criteria

3.6.1 Confusion Matrix

For image classification task (Chapter 4), the aim is to assign each test image to one of predefined classes. The performance in most scene classification methods are measured, analyzed and visualized using the confusion matrix, also known as contingency table (Fei-Fei and Perona, 2005, Lazebnik et al., 2006, Quelhas and Odobez, 2006, Quelhas et al., 2007, Battiato et al., 2010a, Liu et al., 2011, Vogel and Schiele, 2004, Gu et al., 2011). Thus, it is possible to see if the classifier is confusing two classes and can help improving the accuracy of the system.

Table 3-1: A toy example of confusion matrix

		Classification results					
		c1	**c2**	**c3**	**c4**	**c5**	**c6**
Ground truth	**c1**	**95**	0	10	0	0	0
	c2	1	**1**	90	0	1	0
	c3	13	0	**0**	0	0	0
	c4	0	0	1	**34**	3	7
	c5	1	0	2	13	**26**	5
	c6	0	0	2	14	5	**10**

Rows correspond to the classes in the ground truth classes, i.e. actual classes, while *columns* correspond to classes in the classification result, i.e. predicted classes. Elements in the *diagonal* represent the number of correctly classified documents of each class, i.e. the number of ground truth documents with a class name that obtained the same class name during classification task. The off-diagonal row elements represent ground truth documents of a certain class which were misclassified during the classification. For example, each pair of classes $\langle c_i, c_j \rangle$ shows how many documents from c_i were incorrectly assigned to c_j where $i \neq j$. Table 3-1 shows a toy example of a confusion matrix for a dataset of six categories, c_1,c_2, ..c_6. The accuracy for each class in the confusion matrix is measured as the fraction of correctly classified documents with regard to all documents of that ground truth class. For example, the accuracy for class c_1 in Table 3-1 is 95/105=0.90 meaning that 90% of the c_1 ground truth documents are classified as c_1 by the classifier. The overall classification accuracies are measured by the average value of the diagonal values of the confusion matrix. For overall classification performance, we use Average Precision (AP) to evaluate the result of a single classifier, and mean average precision (MAP) to aggregate the performance of multiple classifiers.

3.6.2 k-Fold Cross-Validation

Cross validation is a model evaluation method. It measures how well a model generalizes to a new data. When training a learner some of the data is removed before training starts. When the learner is trained, the data that was removed can be used to test the performance of the trained model. There are different ways of doing the cross validation method. The simplest one is to use the *holdout* method. In this

method, the data is separated into two different sets, called the training set and testing set. The training set is used to train the model then the trained model is used to predict the output for the testing set in order to evaluate the model.

K-fold cross-validation is a way to improve the holdout method. The dataset is randomly divided into *K* subsets. The first subset is removed for testing the model that is trained on the remaining *K-1* subsets. The results of evaluating the trained model in the first subset are reported. Then, the second subset is removed from the dataset and the model is trained using the first and the last *K-2* subsets. This process is repeated *K* times. In each time results are reported. In this method, each data point is tested only once.

Leave-one-out cross validation is a K-fold cross validation method with *K* equal to the number of data point in the dataset. That is each data point is tested on a model that is trained on the whole dataset except the one that has been used in the testing phase.

This thesis uses 10-fold cross validation as a common approach to evaluate models in image classification tasks.

3.6.3 Precision and Recall

To evaluate the effectiveness of information retrieval systems, including CBIR systems, there are several measures to determine the performance of the system which involves counting relevant documents in the retrieved ones. In this thesis, a document refers to an image retrieved from image collection. The best-known and most widely used measures in information retrieval are *precision* and *recall* (Salton, 1968).

In the context of this thesis, precision is defined as the ratio of relevant images to the query image, which has been retrieved by the system, to the total number of retrieved images. In contrast, recall is the ratio of the relevant images to the query image, returned by the system, to the total number of relevant images in the image collection.

$$precision = \frac{number\ of\ relevant\ retrieved}{number\ of\ retrieved} \tag{3-10}$$

$$recall = \frac{number\ of\ relevant\ retrieved}{number\ of\ relevant} \tag{3-11}$$

Given classified image collection with labeled images, relevant retrieved images are those images obtained from the same class of the given query image. However, recall of the system is only possible if we know all relevant images in the image collection.

Both precision and recall describe different qualities of a retrieval result. For a query image, high precision means that most retrieved images are relevant to the query image, and a high recall means that most of relevant images in the image collection have been retrieved. The recall is one if all relevant images in the image collection are retrieved. Usually, the relationship between precision and recall is presented in a precision-recall graph, in which precision values are plotted against recall values. It shows how many relevant and irrelevant images are presented in the top ranked images. A perfect image retrieval system where only relevant images are retrieved would show a straight line. Figure 3-9 shows examples of precision-recall graphs.

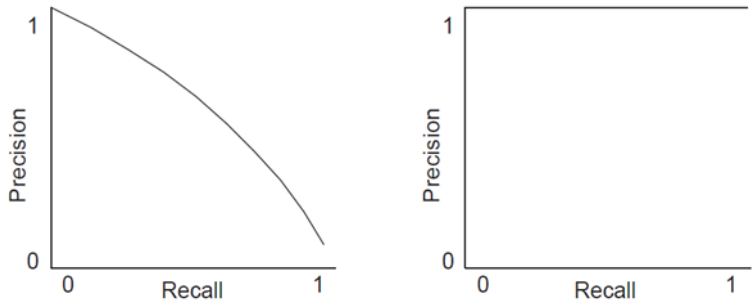


Figure 3-9: Precision-recall graphs. (right) perfect precision-recall graph (Fauzi, 2004).

3.7 Summary

This chapter has presented the fundamentals of image representations related to the fields of image classification, annotation and retrieval. The topic of image description was discussed with an emphasis on the use of local descriptors, as a result of the semantic gap, for robust image description. Image modelling using the bag of visual words and spatial layout has been discussed. Finally, machine learning models, such as support vector machines and nearest neighbor approaches, are introduced followed by the evaluation criteria for image classification and retrieval tasks.

Chapter 4

Image Classification

The bag of visual words (BOW) model is an efficient image representation technique for image classification and annotation tasks. Building good visual vocabularies, from automatically extracted image feature vectors, produces discriminative visual words which can improve the accuracy of image categorization tasks. From literature review, discussed in Chapter 3, we observe that most approaches that use the BOW model in categorizing images ignore useful information that can be obtained from image classes to build visual vocabularies. Moreover, most BOW models use intensity features extracted from local regions and disregard colour information which is an important characteristic of any natural scene image.

This chapter presents a framework to deal with aforementioned limitations. The novelty of this chapter is threefold. *First*, we propose a simple yet effective weighting method, namely *keypoints density-based weighting* (KDW) method,

which is based on the density of quantized local invariant image features over all images sub-regions, to control the fusion of image colour information (colour moments) and BOW histograms on a spatial pyramid layout. Spatial pyramid layout refers to the way we partition an image into sub-regions and it is inspired from the work of Lazebnik et al. (Lazebnik et al., 2006).

Second, we propose integrating knowledge from discriminative visual vocabularies learned from image classes, multiple image features and spatial arrangements information to improve the conventional bag of visual words, for natural scene image classification task.

To improve visual vocabulary construction, visual vocabularies extracted from the training images of each scene image category are combined into a single *integrated visual vocabulary*. It is composed of discriminative visual words from different scene categories. We show that integrating visual vocabularies generated from each image category, improves the BOW image representation and improves accuracy in natural scene image classification.

In the *third* contribution, we show that visual vocabularies generated from training images of one scene image dataset, can plausibly represent another scene image dataset on the same domain. This helps in reducing time and effort needed to build new visual vocabularies.

The proposed approaches are extensively evaluated over three well-known scene classification datasets with 6, 8 and 15 scene categories (Vogel and Schiele, 2004, Oliva and Torralba, 2001, Lazebnik et al., 2006) respectively using 10-fold cross validation. We tested our work on a fourth dataset with 4 scene classes. This

dataset is a subset of 8 scene classes (Oliva and Torralba, 2001) composing natural scene images with no man-made objects .

We will show in this work that the integrated visual vocabulary is more discriminative than the universal visual vocabulary to build BOW histograms. Moreover, we show that the Keypoint Density-based Weighting (KDW) method can be used effectively with the integrated visual vocabulary, to control the fusion of image colour information and BOW histograms on a spatial pyramid layout. We compare our approach to a number of baseline methods such as Gist features (Oliva and Torralba, 2001), rgbSIFT features (Van de Sande et al., 2010) and different configurations of conventional BOW.

This chapter is organized as follows: Section 4.1 presents the main steps to construct bag of visual word image representation. Section 4.2 describes our feature fusion approach. Section 4.3 discusses our experimental work and results. Section 4.4 summarizes this chapter with conclusions.

4.1 Location-Aware Image Semantic Representation

In this section we introduce the main steps needed to construct four forms of bag of visual words which will be used in this work to represent visual image content:

- Universal BOW (*UBOW*) based on universal visual vocabulary.
- Integrated BOW (*IBOW*) based on class-specific visual vocabularies.
- Universal Pyramid BOW (*UPBOW*) similar to *UBOW* but on spatial pyramid layout.
- Integrated Pyramid BOW (*IPBOW*) similar to *IBOW* but on spatial pyramid layout.

The following subsections details all steps required to build these four BOW image representations. In each case we will consider how we extract and describe local features from images, build universal and integrated visual vocabularies and map local features to the closest visual words on spatial pyramid layout.

4.1.1 Local invariant points detection and description

In this work we use the Difference of Gaussian (DoG) point detectors and SIFT descriptors (Lowe, 2004) to detect and describe local interest points or patches from images. Generally, these methods show good performance compared to other methods in the literature (Mikolajczyk and Schmid, 2005). The DoG detector has properties of invariance to translation, rotation, scale and constant illumination changes.

Once local invariant points are defined (*see* Figure 4-1), SIFT descriptors are used to capture the structure of the local image patches and are defined as local histograms of edge directions computed over different parts of the patch. Each patch

is partitioned into 4x4 parts and each part is represented by a histogram of 8 orientations (bins) that gives a feature vector of size 128.

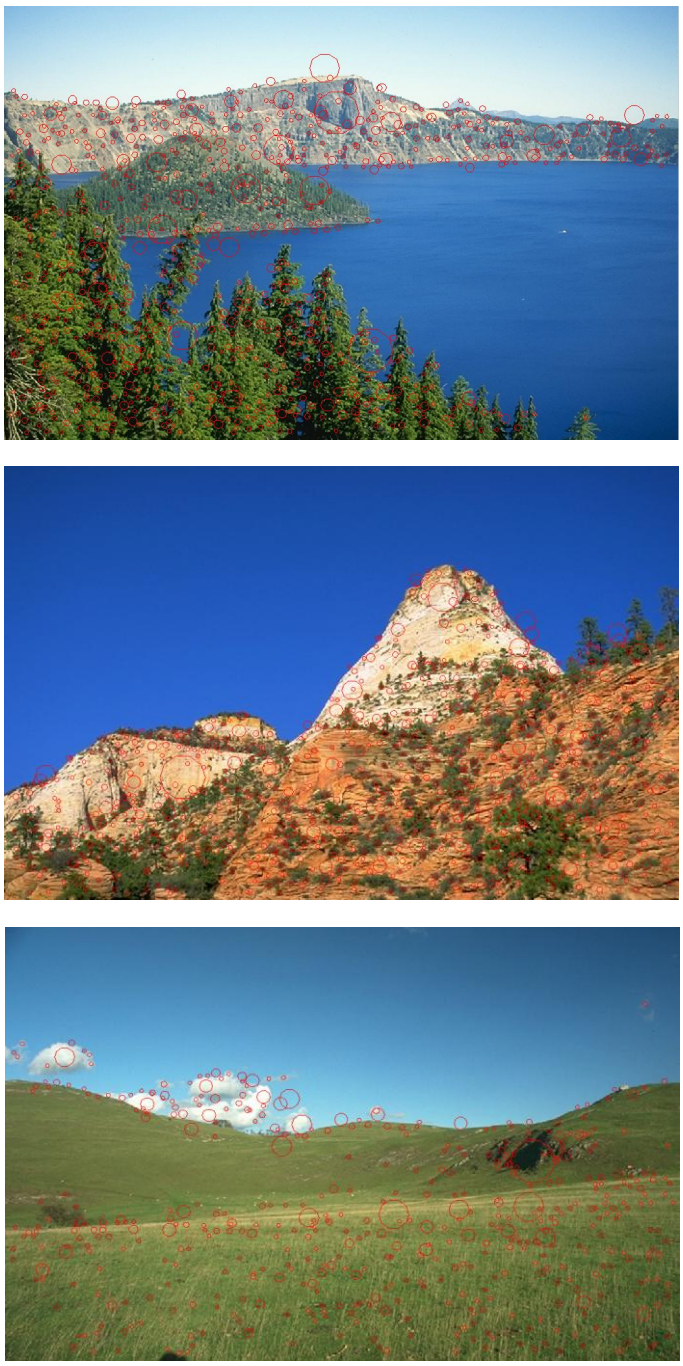


Figure 4-1: Sample images with circles around interest points detected using DoG detector.

In this work we use the binaries provided at (Mikolajczyk, 2011) to detect DoG local points and to compute the 128-D real valued SIFT descriptors from them. This process is described in Figure 4-2. Features extracted from all images are stored in Features Database. In section 4.1.2 we describe how this is used to build visual vocabularies.

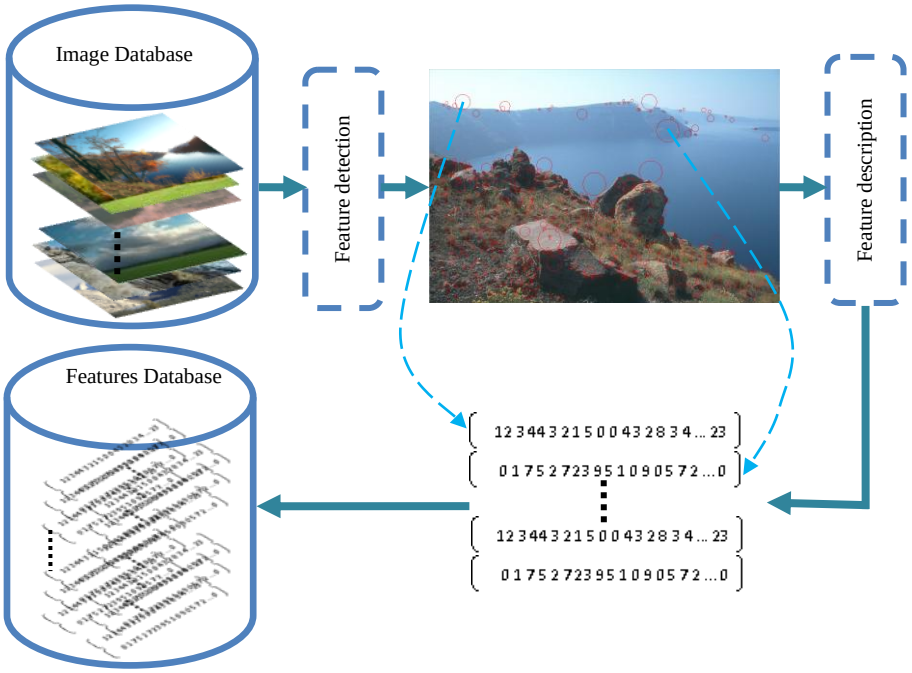


Figure 4-2 Keypoint detection and description process. The *circles* overlaid on the image indicate keypoints located using DoG feature detector. Each keypoint is described and stored in feature vector. Each feature vector contains 128 descriptive values, using SIFT descriptors.

4.1.2 Visual Vocabulary Construction

In this section, we describe how to learn both the universal visual vocabulary and the proposed integrated visual vocabulary that will be used in the rest of this work. To obtain the visual vocabulary, we use feature vectors (SIFT features) stored in image Features Database as described in Section 4.1.1. All feature vectors from all

training images on the dataset are quantized, using the *k*-means algorithm, to obtain k centroids or clusters. These centroids represent visual words. The k visual words constitute the universal visual vocabulary. This vocabulary is used to build the UBOW and the UPBOW. For the integrated visual vocabulary, SIFT features from all training images of each scene class are clustered into k visual words.

More formally: Let $C = \{C_1, C_2, .., C_M\}$ be the set of M scene classes considered. Let $V = \{V_1, V_2, .., V_M\}$ be the set of M class-specific vocabularies. Each $V_j = \{V_{j1}, V_{j2}, .., V_{jk}\}$ is a set of k visual words learned from all training images of class j. We call V the integrated visual vocabulary. This vocabulary is used later to build IBOW and IPBOW.

The rationale behind building an integrated visual vocabulary is to try to find more specific discriminative visual words from each image class in order to avoid interference with other classes. In the universal visual vocabulary, visual words that belong to a specific concept (e.g., foliage) may be assigned to a cluster or visual word of a different concept (e.g., rock). We believe that our integrated visual vocabulary may be robust enough to incorporate naturally existing intra-class variations to discriminate between different image classes.

For example, building visual vocabulary for *coasts* scene images would contain informative information about water, sand and sky, in contrast to other scene classes such as *mountains*. Figure 4-3 shows details of how to construct both kinds of visual vocabularies.

We will show later in the experimental results section how the distribution of the mean of all IBOW of training images are different and more informative and discriminative than the UBOW generated from universal visual vocabulary (*see* Figure 4-13 to Figure 4-15 for differences between universal and integrated visual vocabularies).

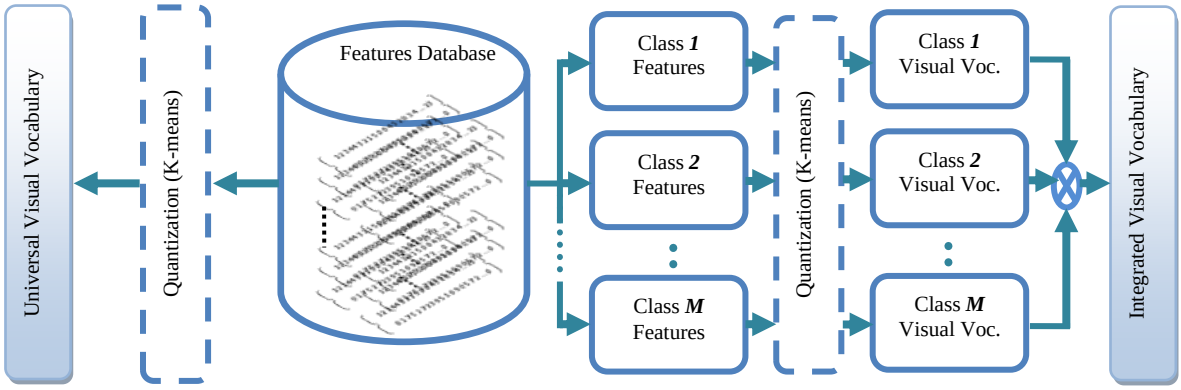


Figure 4-3: Visual vocabulary construction process. The *left* side of Features database shows universal visual vocabulary. The *right* side shows the proposed integrated visual vocabulary. Each class features (for class *1,2,…, M)* represents feature vectors of training images for a specific image category.

4.1.3 Summarizing image content using the BOW

The Bag of Visual Words provides a summary of image contents. In section 4.1.1, we discussed feature detection and description of image content. Section 4.1.2 showed how to build universal and integrated visual vocabularies from image local features.

To build the BOW histogram, each image SIFT descriptor is assigned to the index of the nearest cluster in the visual vocabulary. The visual words in the context of this thesis refer to the cluster centres (centroids) produced by the *k*-means clustering algorithm. Let V denote the set of all visual words produced from the clustering step over a set of local point descriptors $V = \{v_i | i = 1,..,|V|\}$, where v_i is the i*-th* visual word (or cluster) and $|V|$ is the size of the visual vocabulary. We use a vocabulary of size 200 for both the universal visual vocabulary and for class specific visual vocabulary. In the case of the integrated visual vocabulary, $|V|$ is 200*M* (where *M* is the number of classes). Experiments showed no improvements in performance beyond 200 (Lazebnik et al., 2006). The set of all SIFT descriptors for each image d is mapped into a histogram of visual words $h(d)$ at image-level, such that:

$$h_i(d) = \sum_{j=1}^{N_d} f_{d_j}^{(i)} \quad , i = 1,..,|V| \tag{4-1}$$

$$f_{d_j}^{(i)} = \begin{cases} 1 & , \quad \|u_j - v_i\| \le \|u_j - v_l\| \quad , \quad l = 1,..,|V| \text{ and } i \neq l \\ 0 & , \quad otherwise \end{cases} \tag{4-2}$$

where:

$h_i(d)$ is the number of descriptors in image d having the closest distance to the i-th visual word v_i and N_d is the total number of descriptors in image d. $f_{d_j}^{(i)}$ is equal to one if the j-th descriptor u_j in image d is closest to visual word v_i among other visual words in the vocabulary V.

The use of universal vocabulary to build bag of visual word histograms can help discriminate visual semantic content of an image. For example, as shown in Figure 4-4 and Figure 4-5, using integrated visual vocabulary, it is obvious that local image patches tend to be distributed in or close to clusters with same semantic content. On the other hand, this is not the case for BOW generated from universal vocabulary where many similar patches are located in different clusters which cause ambiguity when building BOW histograms.

4.1.4 Spatial pyramid Layout

Although the orderless bag of visual words approach is widely used and has made a noticeable increase of performance in object/scene image modelling, it seems likely that we can enhance it for scene recognition tasks by incorporating spatial information. Spatial pyramid matching (Lazebnik et al., 2006) works by repeatedly subdividing an image into increasingly finer sub-regions and then computing histograms of local patches found inside each image sub-region. An image is represented as a concatenation of weighted histograms at all levels of divisions. In this thesis, spatial pyramid layout refers to representing images by placing a sequence of increasingly coarser grids over an image. Here we did not penalize local histograms of BOW as described in (Lazebnik et al., 2006, Battiato et al., 2010a), since the experiments showed that to do so decreases the classification accuracy of our system.

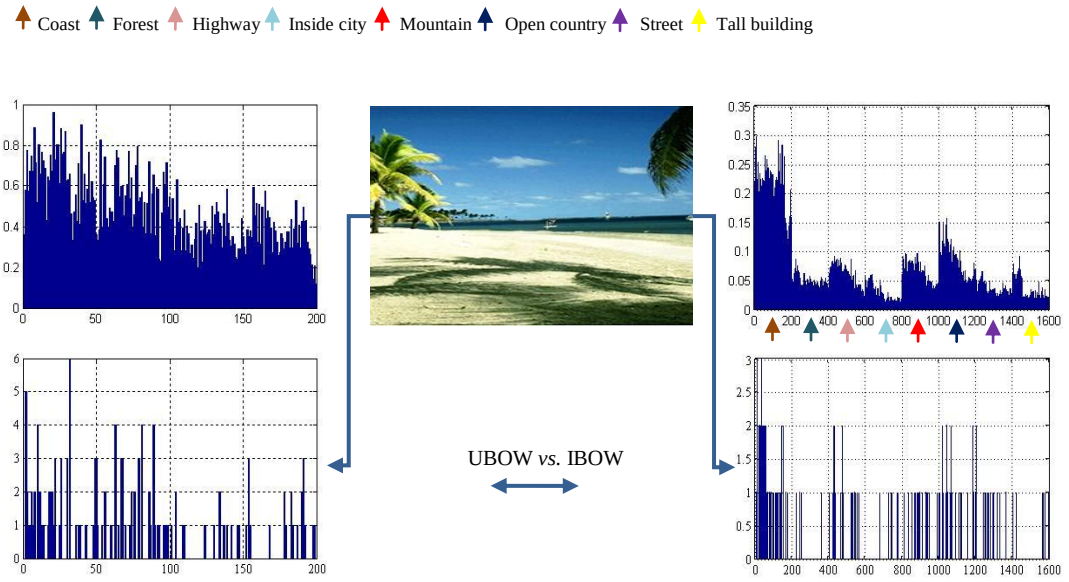


Figure 4-4: A sample image, depicted in the *middle*, of the Coast class from *Dataset 2* (*see* section 4.3.2). *First* and *third* column shows the difference between UBOW and IBOW. The first row shows the mean vector of all UBOW and IBOW histograms of training images. The second row shows BOW and IBOW of the image depicted in the *middle*.

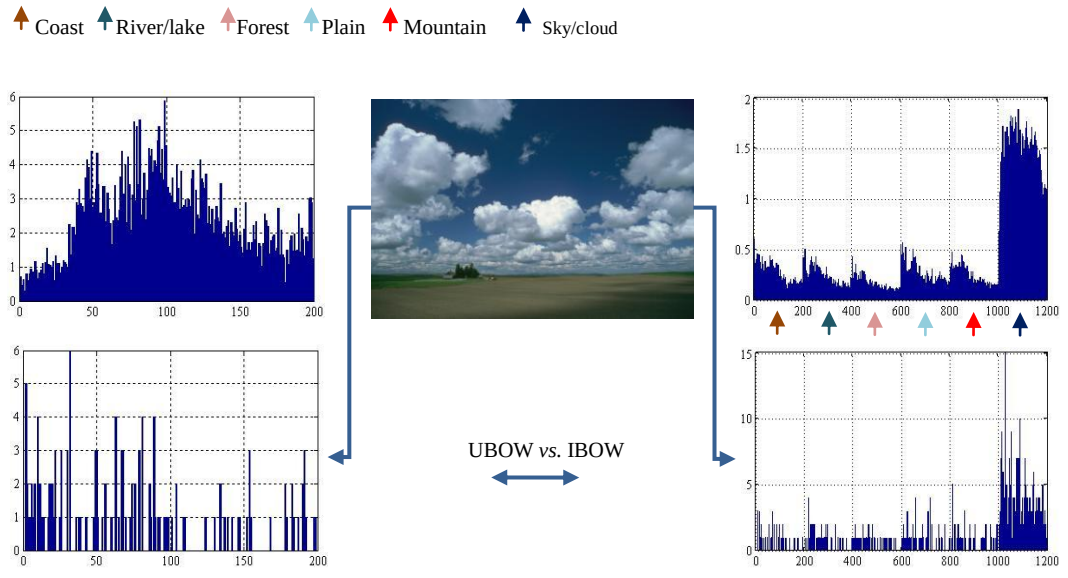


Figure 4-5: A sample image, depicted in the *middle*, of the Sky/clouds class from *Dataset 1* (*see* section 4.3.2). *First* and *third* column shows the difference between UBOW and IBOW. The first row shows the mean vector of all UBOW and IBOW histograms of training images. The second row shows BOW and IBOW of the image depicted in the *middle*.

4.2 Pyramidal fusing of BOW and image colour information

In this section, we show how to model image semantic information based on merging BOW and colour moments using a spatial pyramid layout. The motivation of our approach is that most techniques that use BOW rely only on intensity information extracted from local invariant points and neglect colour information which seems likely to help in recognition performance for scene image categories.

We can see in Figure 4-6 an image with circles around a rather dense set of interest points produced by DoG detectors. We see that interest points are not uniformly distributed across the image, but rather are clustered on salient regions in the scene. In natural scene images, colour information has a significant effect in discriminating image areas such as sky, water and sand. Hence, we believe that merging colour information and the BOW will be significant in modelling image visual semantics.

Therefore, we propose the *Keypoint Density-based Weighting* method *(*KDW) for merging colour information and BOW over image sub-regions at all granularities on the spatial pyramid layout. The KDW method aims to regulate how important colour information is in each image sub-region before fusing it with BOW. The spatial pyramid layout, as seen in Figure 4-7, works by splitting an image into increasingly coarser grids to encode spatial locations of image local points.

(a)

(b)

Figure 4-6: (a) Sample image with circles around interest points. (b) Sky and water contain little information of interest. Red borders in (b) shows important information that helps discriminate image content.

Hence an image with $L = 2$ levels, will have three different representations with $(\sum_{l=0}^{L}(2^l)^2) = 21$ image sub-regions overall. Each image sub-region is represented by a combination of the BOW and a weighted colour moments vector of size 6 on the HSV colour space (2 for Hue, 2 for Saturation and 2 for Value). Both colour moments and the BOW histogram are normalised to be unit vectors before the merging process. An image with number of levels $L = 2$ and a visual vocabulary of size 200 will produce a vector of dimension 4326.

We formulate our proposed approach below:

Let L denotes the number of levels, $l = 0,1,...,L$, needed to represent an image d on the spatial pyramid layout, i.e., an image d will have a sequence of L grids of increasingly finer granularity. Let $h^l(d_{r_i})$ and $c^l(d_{r_i})$ denote a histogram vector of BOW computed using equation (4-1) and the colour moments vector respectively. Both are computed from an image d at level l and sub-region $r_i, i = 1,..,(2^l)^2$.

The concept of Keypoint Density-based Weight (KDW)*:* Colour moment vector $c^l(d_{r_i})$ is assigned a high weight on image sub-regions that have a keypoint density below threshold $T_{r_i}^l$. Colour information will be less important in image sub-regions with high number of local interest points. The threshold T is a real valued vector. Each component represents the average density of keypoints (number of keypoints) at specific image sub-region over all training images. We propose the keypoint density-based weight as:

$$T_{r_i}^{l} = \frac{1}{m}\sum_{j=1}^{m} h^{l}(d_{r_i}^{j}) \tag{4-3}$$

where m is the number of images in the training image dataset. The components of the threshold vector, which is the average keypoint density of all images at specific sub-regions and granularity, help in making a decision about the importance of colour information at specific image sub-region. The unified feature vector $H(d)$ for image d is a concatenation of weighted colour moments and BOW at all levels and over all granularities:

$$H(d) = \left\{ \begin{array}{c} (h^{0}(d_{r_1}), w_{r_1}^{0} c^{0}(d_{r_1})), (h^{1}(d_{r_1}), w_{r_1}^{1} c^{1}(d_{r_1}), ..., h^{1}(d_{r_4}), w_{r_4}^{1} c^{1}(d_{r_4})), \\ (h^{L}(d_{r_1}), w_{r_1}^{L} c^{L}(d_{r_1}), ..., h^{L}(d_{r_{(2^L)^2}}), w_{r_{(2^L)^2}}^{L} c^{L}(d_{r_{(2^L)^2}})) \end{array} \right\} \tag{4-4}$$

$$w_{r_i}^{l} = \begin{cases} 1 & , \quad \sum_{j=1}^{|V|} h_j^{l}(d_{r_i}) \; < \; T_{r_i}^{l} \\ 0.5 & , \quad otherwise \end{cases} \tag{4-5}$$

We should notice that the values of weights w are non-negative numbers to indicate the importance of colour information. We aim to cause images from the same category to be close, and images from different categories to be far away in the new image representation. Values for the weights have been obtained empirically during learning the support vector machine SVM classifiers (Chang and Lin, 2011). We should notice that weight values are highly dependent on the threshold vector obtained from equation (4-3).

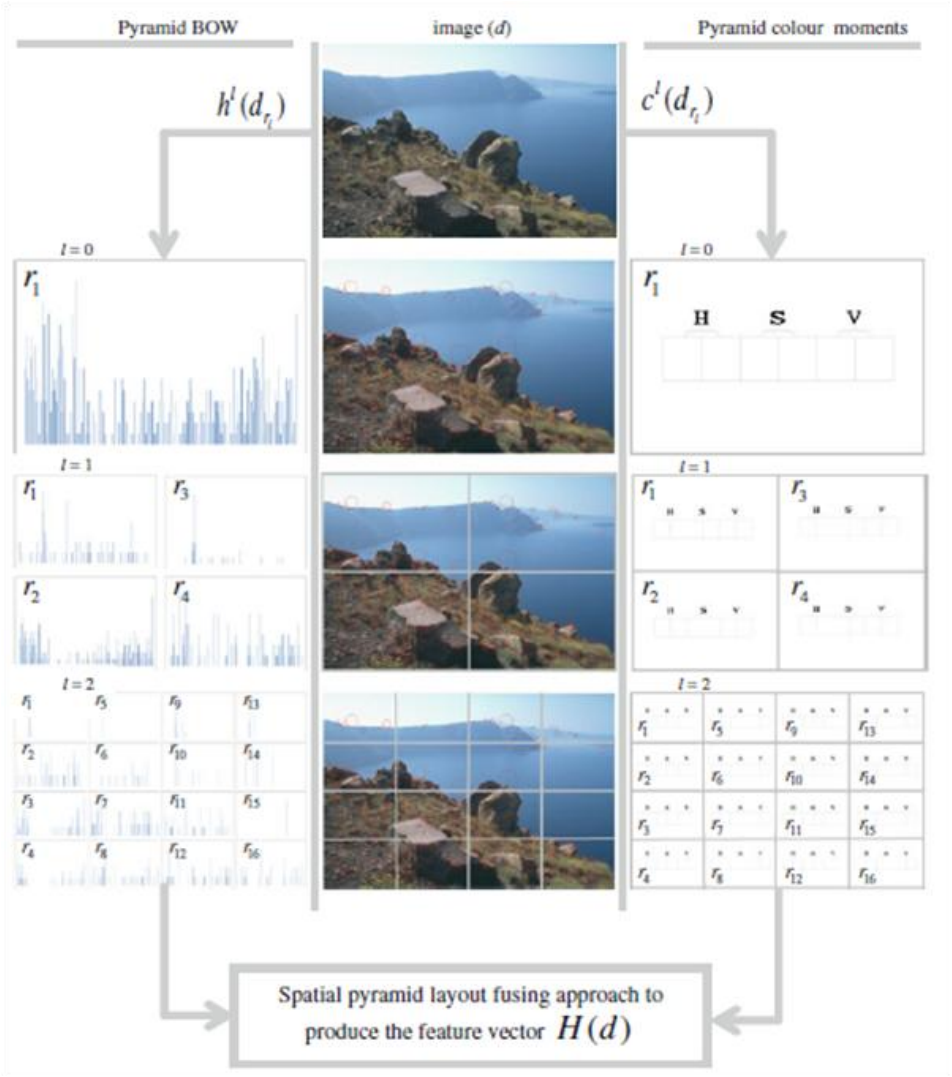


Figure 4-7: Feature fusion process on spatial pyramid layout (L=2). The *left* column represents histograms of BOW. The *right* column represents colour moments for the HSV colour space bands. The *middle* column represents an image at different levels overlaid with *circles* around interest points.

Here, we use the proposed integrated visual vocabulary described in section 4.1.2 and the spatial pyramid layout to generate IBOW and IPBOW histograms. Fusing weighted pyramidal colour moments (WPCM) with the IBOW and IPBOW histograms using equation (4-4) we obtain improved image representation (IBOW+WPCM and IPBOW+WPCM). We assume that building visual vocabularies from individual scene classes could produce more discriminative visual words than using universal vocabulary. To justify this assumption, all UBOW and IBOW histograms generated from training images of Vogel's dataset (Vogel and Schiele,

2004), described in section 4.3.2, have been averaged to see the distribution of both BOWs in each scene class. Figure 4-13 shows the difference between both averages. Some sample BOW histograms are shown and they tend to be similar or close to their average vector.

4.3 Experimental work

The first part of this section presents the support vector machine (SVM) classifier and the protocol we follow in all our experiments. Next, we describe the origin and composition of datasets we use in our experiments. Experimental results are then reported with discussion. We use the confusion matrix to assess the performance of all considered experiments.

4.3.1 Scene classifier

Multi-class classification is done using a support vector machine (SVM) with a histogram intersection kernel. We use SVM in our study as they have been empirically shown to yield higher classification accuracy in scene and text classification tasks (Quelhas and Odobez, 2006, Alqasrawi et al., 2009, Khan et al., 2009). Variations in the classification accuracy are possible due to our choice of SVM kernel function. In this work, we use the histogram intersection kernel.

Many studies in image classification observe that the histogram intersection SVM kernel is very useful. Moreover, histogram intersection has been shown more effective than the Euclidean distance in supervised learning tasks (Wu and Rehg, 2009, Lazebnik et al., 2006). Odone et al. (Odone et al., 2005) proved that histogram intersection is a Mercer kernel and thus can be used as a similarity measure in kernel

based methods. Given two BOW histograms $h(d_1)$ and $h(d_2)$, the histogram intersection kernel is:

$$K(h(d_1), h(d_2)) = \sum_i \min(h(d_{1_i}), h(d_{2_i})) \quad (4\text{-}6)$$

The protocol we follow for each of the classification experiments was as follows: All experiments have been validated, using 10-fold cross validation where 90% of all images are selected randomly for learning the SVM and the remaining 10% are used for testing. The procedure is repeated 10 times such that all images are actually tested by the SVM classifier. The average of the results over the 10 splits yields the overall classification accuracy, sometimes called mean average precision MAP.

To implement the SVM method we used the publicly available LIBSVM software (Chang and Lin, 2011), in MATLAB, where all parameters are selected based on 10-fold cross validation on each training fold. We use one-against-one multi-classification approach that results in $\frac{M(M-1)}{2}$ two-class SVMs for M scene classes.

4.3.2 Image datasets

There are many image datasets available in the computer vision literature, but most of them are dedicated to object detection and categorization tasks. Performance of the proposed scene classification approach is tested on two types of image datasets: a dataset with natural scene images only, which is our main concern, and datasets with heterogeneous images including different kind of images.

The reason for choosing natural scene images is that they generally are difficult to categorize in contrast to object-level classification because natural scenes constitute a very heterogeneous and complex stimulus class (Vogel et al., 2007). Also, we considered scene images that constitute artificial objects to allow fair and straightforward comparison with state-of-the-art scene classification methods. Four datasets were used in our experiments:

Dataset 1: This dataset, kindly provided by Vogel et al. (Vogel and Schiele, 2004), contains natural scene images only with no man-made objects. It contains a total of 700 colour images of resolution 720 × 280 and distributed over 6 categories. The categories and number of images used are: coasts, 142; rivers/lakes, 111; forests, 103; plains, 131; mountains, 179; sky/clouds, 34. One challenge in this image dataset is the ambiguity and diversity of inter-class similarities and intra-class differences which makes the classification task more challenging.

Dataset 2: This dataset is a subset of the Oliva and Torralba (Oliva and Torralba, 2001) dataset. It constitutes images of natural scene categories with no artificial objects, which are semantically similar to images in Dataset1 and is distributed as follows: coasts, 360; forest, 328; mountain, 374; open country, 410. The total number of images in this dataset is 1472.

Dataset 3: This dataset contains heterogeneous image categories. It consists of a total of 2688 colour images, 256x256 pixels, and distributed over 8 outdoor categories. The categories and number of images used are: coast, 360; forest, 328; mountain, 374; open country, 410; highway, 260; inside city, 308; tall building, 356; street, 292. This dataset is created by Oliva and Torralba (Oliva and Torralba, 2001) and is available online at *http://cvcl.mit.edu/database.htm*.

Dataset 4: This dataset, provided by Lazebnik et al. (Lazebnik et al., 2006), contains 15 heterogeneous natural scene image categories. All images in this dataset are grayscale images (i.e., no colour images). It contains different kind of images and the average image size is 300x250 pixels. Images are distributed over categories as follow: highway, 260; inside city, 308; tall building, 356; street, 292; suburb, 241; forest, 328; coast, 360; mountain, 374; open country, 410; bedroom, 216; kitchen, 210; living room, 289; office, 215; industrial, 311; store, 315. The first 8 categories were from Oliva and Torralba (Oliva and Torralba, 2001) and the first 13 were from Fei-Fei and Perona (Fei-Fei and Perona, 2005). Figure 4-8 depicts sample images from the four datasets aforementioned with different image classes.

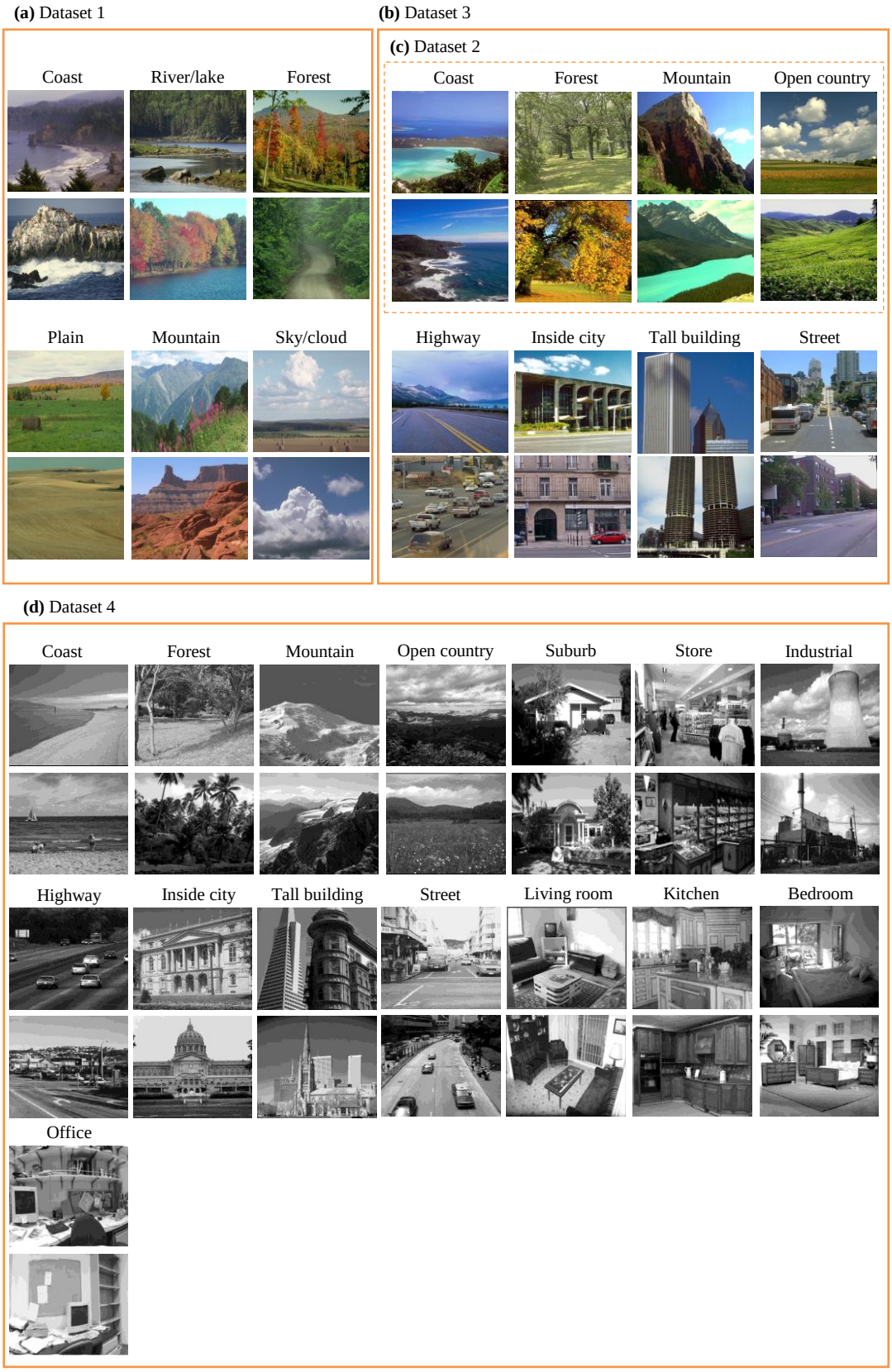


Figure 4-8: Some examples of the images used for each category from the Dataset 1, Dataset 2, Dataset3 and Dataset 4 respectively.

4.3.3 Feature extraction

In this work, we used Matlab to conduct all experiments. As we mentioned earlier, in the experiments we perform 10-fold cross validation in order to achieve more accurate performance estimation. The binaries provided by (Mikolajczyk and Schmid, 2005) are used to detect and describe local keypoints using DoG and SIFT as parameters. Further we extracted basic rgbSIFT features with local keypoint detection and standard parameters using the Color Descriptor software provided by (Van de Sande et al., 2010).

To build different visual BOW histograms, SIFT and rgbSIFT features extracted from training images are used to build 10 visual vocabularies, one for each fold. Gist features are extracted from images using the implementation provided by Oliva and Torralba at (http://people.csail.mit.edu/torralba/code/spatialenvelope/).

4.3.4 Experimental results

In this section, we conduct extensive experiments to empirically evaluate the performance of our proposed approach and compare it to the existing baseline and BOW models for natural scene image categorization tasks. We present four sets of experiments each corresponds to one of the datasets mentioned in section 4.3.2. In the first experiments, we tested the performance of our proposed approach on colour natural scene images with no artificial objects, i.e., *Dataset 1* and *Dataset 2* respectively.

The performance of our proposed approach is compared with Gist features, improved Gist features and different configurations of BOW models generated from

SIFT features and rgbSIFT features. All visual vocabularies employed in our experiments are generated separately for SIFT and rgbSIFT features.

In the second experiments, we test our proposed approach on *Dataset 3* which contains different kind of images and scene categories. We intend to find out how our approach performs on heterogeneous set of scene classes. In the third experiment, we use grayscale images, *Dataset 4*, to test the performance of our proposed approach on large number of images and scene categories. In the last experiment, we investigate the possibility of using visual vocabularies generated from *Dataset 1* to produce IBOW for *Dataset 2*.

Firstly, we present the classification performance of Gist features and improved Gist features (i.e., Gist with pyramidal colour moments) tested on *Dataset 1* and *2*. The Gist descriptor (Oliva and Torralba, 2001) uses a low dimensional representation of the scene which does not require any segmentation process. A bank of Gabor filters are employed in the frequency domain and tuned to different orientations and scales. The image is divided into a 4×4 grid for which orientation histograms are computed. The Gist features produce a vector of dimension 512. Further details can be found in (Oliva and Torralba, 2001).

The results published by Oliva and Torralba (Oliva and Torralba, 2001) are based on eight scene classes, so to compare their approach to ours we use their code to repeat their experiments for classifying the chosen four scene classes, i.e., *Dataset 2*. Table 4-1 and Table 4-2 depict the classification results of using the Gist features on *Dataset 1* and *2*. What is interesting is that although scene classes in *Dataset 2* are similar in their visual semantics to the corresponding scene classes in *Dataset 1*, the results for the former dataset are significantly better than for the latter.

It seems that the classes in *Dataset 1* do not exhibit consistent properties as detected by Gist. To improve the Gist features, we propose to integrate image colour information to Gist features by fusing pyramidal colour moments (PCM, L=2) with Gist features. This combination of image features has resulted in an improvement in the classification performance and thus supporting the significance of pyramidal colour moments approach.

Detailed results are depicted in Figure 4-9 and Figure 4-11 for class specific classification performance using Gist features compared to other approaches. Figure 4-10 report the average classification results of (Gist) and (Gist+PCM) representations on both datasets. It is clear that adding pyramidal color moments to the Gist features outperformed the classification performance of using Gist features alone.

Table 4-1: The first part of this table shows the confusion matrix of our proposed approach (IPBOW+PCM) with no weighting tested on *Dataset 2*. The diagonal bold values are the average classification rate of each image category. The overall classification accuracy is 88.7% and is clearly outperforms the Gist features shown in the second part.

	Coast	Forest	Mountain	Open country	IPBOW+PCM	Gist (Oliva and Torralba, 2001)
Coast	**0.90**	0.00	0.03	0.08	0.90	0.88
Forest	0.00	**0.96**	0.02	0.02	0.96	0.93
Mountain	0.02	0.02	**0.89**	0.07	0.89	0.84
Open country	0.08	0.04	0.06	**0.82**	0.82	0.75
				Overall accuracy rate	**88.7%**	**84.2%**

Secondly, we present the classification performance of using pyramidal colour moments fused with the proposed IBOW and IPBOW, using the KDW weighting method, to represent image contents. We used integrated visual vocabularies to build IBOW from the whole image and IPBOW from image sub-regions as discussed in section 4.1.3. The pyramidal colour moments were fused

with IBOW and IPBOW using our weighting method, to obtain two new image representations; IBOW+WPCM and IPBOW+WPCM.

We can observe from Table 4-1 that adding spatial information and colour moments to the IPBOW improves the classification performance. Table 4-1 indicates clearly that our approach, excluding the weighting technique, outperform Gist features by +4.5%. This is mainly because Gist features do not contribute colour information and spatial layout which provides informative features for scene classification task.

In Table 4-2, our approach to represent image content (IPBOW+WPCM) outperforms others' work (Vogel and Schiele, 2004, Oliva and Torralba, 2001, Quelhas and Odobez, 2006). This provides empirical evidence that the integrated visual vocabulary provides more informative visual words than the universal visual vocabulary. Also, the results show how the weighting influences the performance of the IBOW and thus improves the classification results. Despite this, Gist features still performs very well in some classes such as *'river/lakes'* and '*sky/clouds'* classes which are most difficult for our approach to recognize.

Refer to Figure 4-9, it can be seen that our approach works very well in the first three classes, but the performance degrades for the '*open country'* scene class against (Gist+PCM) image representation. Furthermore, in Figure 4-11, the performance of our approach on '*river/lakes*' and '*sky/clouds*' scene classes have gained comparable results against other approaches and outperformed them on the other four classes. The overall performance results of our approach against other methods are shown in Figure 4-10.

Table 4-2: The first part of this table shows the confusion matrix of our proposed approach (IPBOW+WPCM) tested on *Dataset 1*. The diagonal bold values are the average classification rate of each image category. The overall classification accuracy is 73.7%. The second part of this table reports results of other approaches on the same dataset. It is obvious that our approach outperforms other approaches reported in the literature.

	Coast	River/ lake	Forest	Plain	Mountain	Sky/ cloud	IPBOW+WPCM	Gist (Oliva and Torralba, 2001)	Vogel (Vogel and Schiele, 2004)	Quelhas (Quelhas and Odobez, 2006)
Coast	**72.54**	8.45	2.11	5.63	11.27	0.00	72.54	54.93	59.9	69.0
River/lake	18.02	**49.55**	10.81	5.41	15.32	0.90	49.55	49.55	41.6	28.8
Forest	1.94	3.88	**90.29**	1.94	1.94	0.00	90.29	83.50	94.1	85.4
Plain	9.16	4.58	6.11	**64.89**	14.50	0.76	64.89	58.02	43.8	62.6
Mountain	6.15	2.79	1.68	3.91	**84.36**	1.12	84.36	74.30	84.3	77.7
Sky/cloud	5.88	2.94	0.00	5.88	0.00	**85.29**	85.29	85.29	100	76.5
Overall accuracy rate							**73.7%**	**65.3%**	**67.2%**	**66.7%**

Our proposed approach is also compared with colour by design methods. We used rgbSIFT (Van de Sande et al., 2010) features, extracted from training images, to generate integrated visual vocabularies. Each rgbSIFT feature is a vector of 384-D (SIFT features of 128-D are extracted from *RGB* image bands respectively). An image is then represented as a histogram counting the number of keypoints characterized by rgbSIFT that belongs to a specific vocabulary index. The average of the 10 accuracy rates using 10-fold cross validation is used to measure the performance of all experiments as mentioned in section 4.3.3.

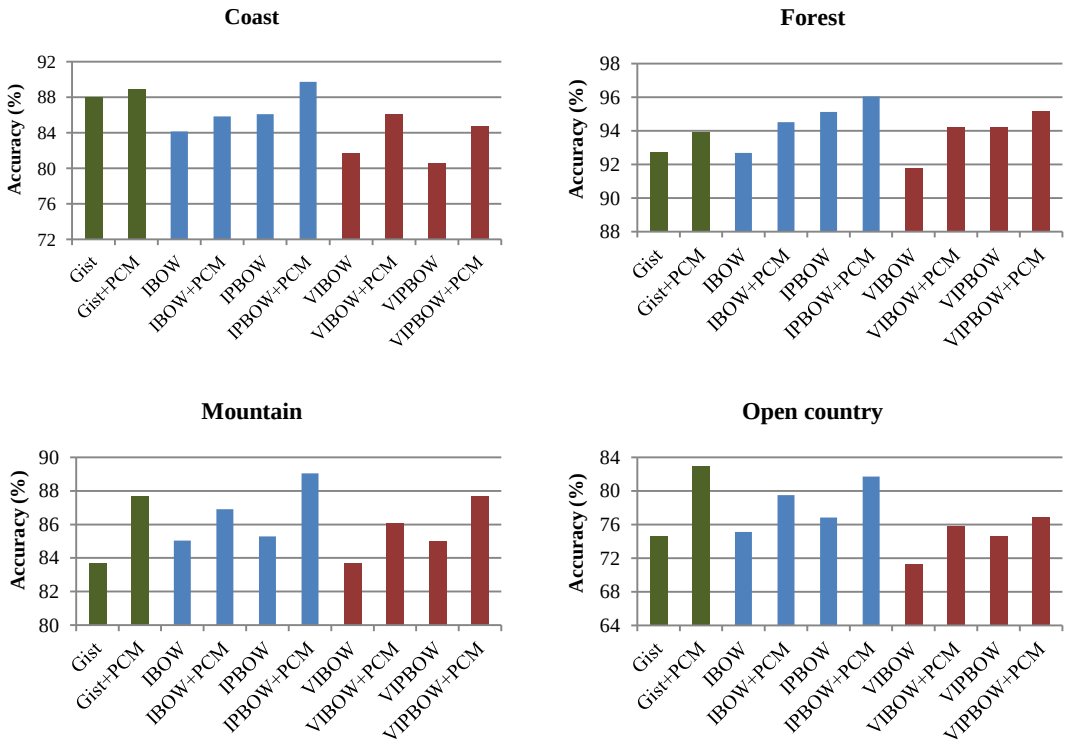


Figure 4-9: The classification performance of IPBOW+PCM compared with different baseline methods for each scene class of *Dataset 2*. It is clear that in most scene classes IPBOW+PCM outperforms other methods. Gist+PCM features perform best for the open country scene class.

Figure 4-12(a) compares the average classification rate of our proposed approach and rgbSIFT features tested on *Dataset 1*. The results confirm the effectiveness of our approach compared with rgbSIFT in image classification task. Our proposed model achieved better results on five image categories out of 6 while rgbSIFT features performed better in recognizing "*plains"* category.

Moreover, this work investigates the influence of applying visual vocabularies generated from one image dataset to generate BOW from another image dataset. We hypothesize that visual words that exist in a specific image class are similar to those in another class with same visual semantic features.

For example, '*coasts'* class in *Dataset 1* contains visual words that are similar in semantic to visual words in '*coasts'* class of *Dataset 2*. We used integrated visual vocabularies generated from *Dataset 1* to index visual patches of images in *Dataset 2*. Figure 4-9 and Figure 4-10(a) show our findings, where VIBOW stands for BOW produced by applying integrated visual vocabularies generated from *Dataset 1*. Although classification results are lower than our approach it is found to be better than Gist features. The results show the plausibility of using visual vocabularies of one dataset to generate BOW for another dataset within the same domain. On the other hand, this is not applicable if images in datasets are different in their visual appearance and semantic content.

In order to test the performance of our proposed approaches on more heterogeneous image categories we conducted extensive experiments on *Datasets 3* and *4*. Table 4-3 depicts the confusion matrix of our proposed model tested on 8-scene categories, i.e., *Dataset 3*. In terms of average classification accuracy rate, we achieved 88.28% which is comparable to other baseline approaches as depicted in Table 4-4. The results show an improvement on the classification performance of IBOW over UBOW and how our weighting method influences the overall classification rate. Figure 4-12(b) compares the average classification rate of our approach with rgbSIFT features.

Table 4-3: Confusion matrix of eight class dataset, *Dataset 3*, based on our proposed approach. Rows and columns corresponds to correct and predicted classes respectively. The diagonal bold values are the average classification rate of each image category. The overall system accuracy is 88.28% and is comparable to other state-of-the-art image classification approaches.

	Coast	Forest	Highway	Inside city	Mountain	Open country	Street	Tall building
Coast	**90.83**	0.28	1.11	0.00	2.22	5.00	0.00	0.56
Forest	0.00	**95.73**	0.00	0.00	2.13	2.13	0.00	0.00
Highway	6.15	0.00	**81.54**	3.46	1.15	2.31	2.31	3.08
Inside city	0.00	0.32	0.00	**88.96**	0.00	0.32	3.25	7.14
Mountain	2.14	0.80	0.00	0.00	**90.91**	5.61	0.53	0.00
Open country	7.07	5.61	0.73	0.00	6.10	**80.49**	0.00	0.00
Street	0.00	0.00	1.03	6.16	0.68	0.00	**86.64**	5.48
Tall building	2.53	0.28	0.00	3.93	1.40	0.56	0.56	**90.73**

Table 4-4: Average classification accuracy rate (%) on *Dataset 3* using universal and integrated visual vocabularies with different BOW configurations with/out pyramid colour moments.

UBOW	UBOW+PCM	IBOW	IBOW+PCM	IPBOW	IPBOW+PCM	IPBOW+WPCM	Gist	rgbSIFT
74.74	**81.10**	**79.50**	**84.90**	**82.25**	**85.23**	**88.28**	**87.31**	**87.54**

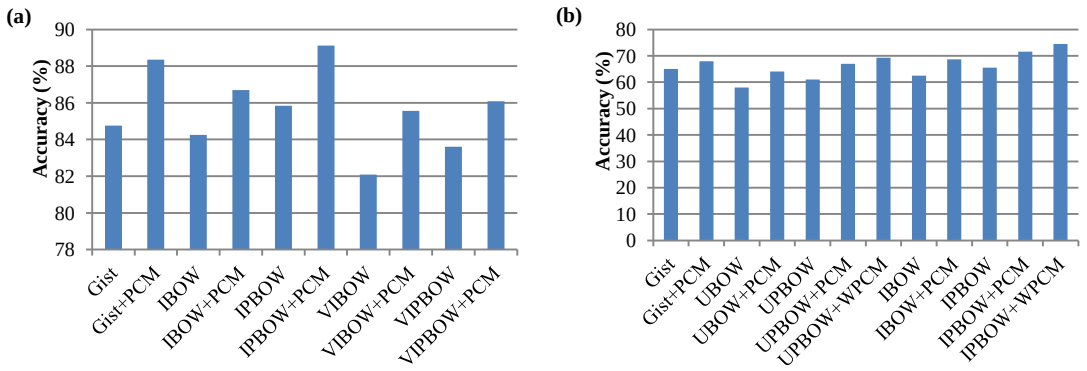


Figure 4-10: A comparison of the average classification performance accuracy of different image representation methods on Dataset 2 (a) and Dataset 1 (b).

For *Dataset 4*, we tested our proposed approach on grayscale images for both training and testing. In this case, our pyramidal colour moments represents only the information that are available in single image band i.e., no colour information. First and second moments are computed from all image sub-regions at pyramidal layout with L=2 and resulted in a vector of size 42-D.

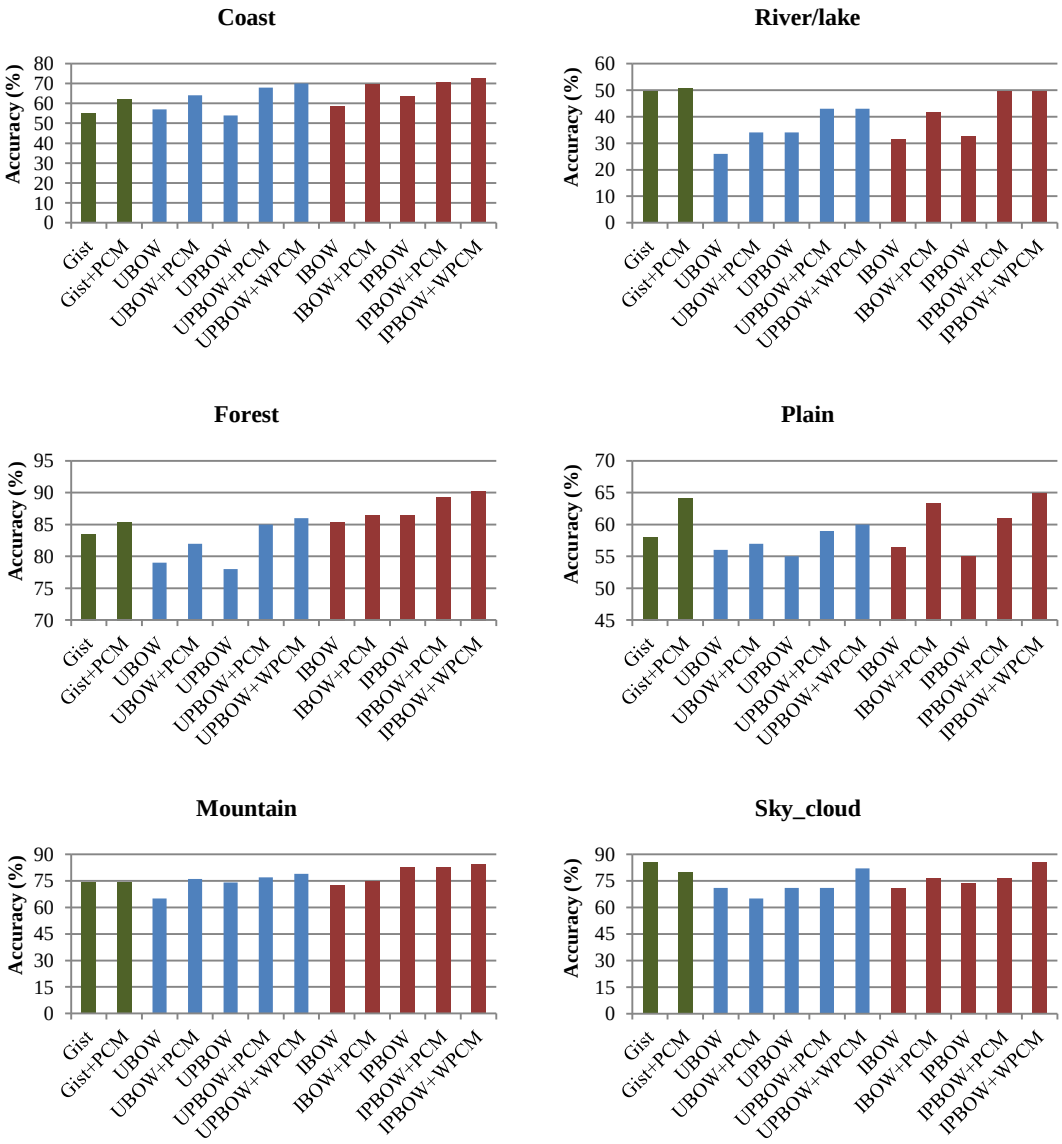


Figure 4-11: The classification performance of our approach compared with different methods for each scene class of *Dataset 1*. It is clear that in most scene classes our approach outperforms other methods.

The confusion matrix, depicted in Table 4-5, illustrate the performance of our proposed approach. We achieved 81.03% overall classification rate which is higher than traditional BOW with universal vocabularies. We compared the performance of universal BOW and integrated BOW with different configurations. Results are reported in Table 4-6. Also, our approach is comparable to the results obtained by

Battiato et al. (Battiato et al., 2010a) where they achieved 79.43% classification rate on the same dataset.

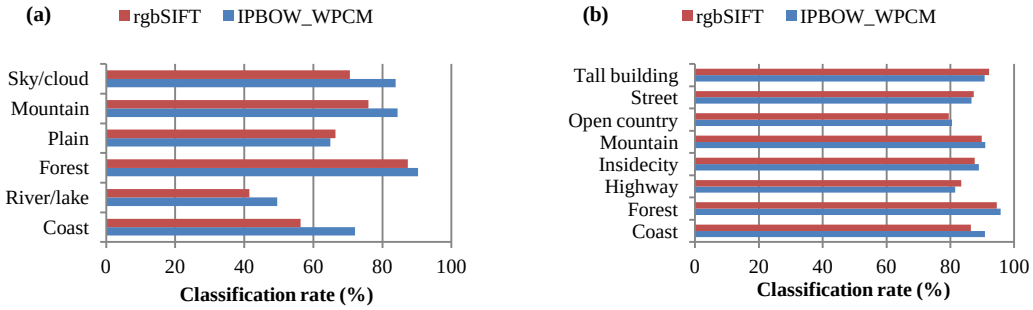


Figure 4-12: Performance comparisons between our proposed approach (IPBOW+WPCM) based on SIFT features and IBOW image representation based on rgbSIFT features (Van de Sande et al., 2010) both tested on *Dataset 1* (a) and *Dataset 3* (b).

Table 4-5: Confusion matrix of *Dataset 4* based on our proposed approach. Rows and columns corresponds to correct and predicted classes respectively. The diagonal bold values are the average classification rate of each image category. The overall system accuracy is 81.03% and is comparable to other state-of-the-art image classification approaches.

	Suburban	Coast	Forest	Highway	Inside City	Mountain	Open Country	Street	Tall building	Office	Bedroom	Industrial	Kitchen	Living room	Store
Suburban	**92.95**	0.00	0.41	0.00	0.41	0.00	0.41	0.00	1.66	0.00	0.00	1.24	0.00	2.90	0.00
Coast	0.00	**91.39**	0.28	1.94	0.00	1.39	4.44	0.00	0.28	0.00	0.00	0.28	0.00	0.00	0.00
Forest	0.00	0.00	**96.95**	0.00	0.00	0.91	2.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Highway	0.00	5.77	0.00	**82.69**	1.54	0.77	2.69	1.54	1.92	0.00	0.00	1.92	0.00	0.00	1.15
Inside City	0.00	0.32	0.32	0.00	**73.70**	0.00	0.00	2.92	5.84	0.32	0.32	4.87	1.95	1.30	8.12
Mountain	0.00	3.21	2.14	0.00	0.00	**90.11**	4.01	0.00	0.27	0.00	0.27	0.00	0.00	0.00	0.00
Open Country	0.24	6.10	2.93	1.22	0.00	4.15	**84.88**	0.00	0.00	0.00	0.00	0.49	0.00	0.00	0.00
Street	0.00	0.00	0.00	1.71	6.16	0.34	0.00	**81.16**	4.79	0.00	0.00	3.77	0.00	0.34	1.71
Tall building	0.00	1.69	0.56	1.12	2.25	1.40	0.56	1.12	**84.83**	0.28	0.84	3.37	0.56	0.28	1.12
Office	0.00	0.47	0.00	0.00	0.93	0.00	0.00	0.00	0.47	**78.14**	6.05	0.93	4.19	7.44	1.40
Bedroom	0.00	1.85	0.00	0.00	0.93	1.39	0.46	0.00	3.24	4.63	**57.87**	5.56	3.24	18.06	2.78
Industrial	0.64	2.57	0.00	1.93	5.14	3.86	0.64	1.29	6.43	0.64	2.25	**64.63**	0.64	2.89	6.43
Kitchen	0.00	0.48	0.00	0.48	6.19	0.00	0.00	0.00	2.38	5.24	4.29	2.86	**65.71**	9.52	2.86
Living room	0.00	0.69	0.00	0.00	1.04	0.00	0.35	0.35	1.38	2.77	6.23	3.46	6.23	**72.32**	5.19
Store	0.00	0.00	0.95	0.00	5.40	0.95	0.00	0.95	1.27	0.95	0.95	1.90	1.59	3.81	**81.27**

Table 4-6: Classification results on *Dataset 4* using universal and integrated visual vocabularies with different configurations of BOW to represent visual content.

	UBOW	UBOW+PCM	IBOW	IBOW+PCM	IPBOW	IPBOW+PCM	IPBOW+WPCM
Suburban	87.97	88.38	92.95	95.02	91.70	92.95	92.95
Coast	76.11	78.33	84.72	86.39	86.39	88.61	91.39
Forest	91.46	94.21	91.16	93.29	92.68	95.43	96.95
Highway	64.23	70.77	79.62	81.54	78.85	80.77	82.69
Inside city	57.47	62.34	70.45	72.73	70.45	72.08	73.70
Mountain	75.13	78.88	87.17	88.77	87.43	87.17	90.11
Open country	60.00	64.88	76.83	79.76	78.05	79.27	84.88
Street	63.36	71.58	75.34	80.14	76.37	78.08	81.16
Tall building	60.11	65.17	77.25	80.06	80.62	83.43	84.83
Office	68.84	73.95	77.67	80.47	79.07	78.14	78.14
Bedroom	31.02	35.19	55.56	56.48	55.56	57.87	57.87
Industrial	39.87	44.05	52.41	56.59	55.63	57.23	64.63
Kitchen	46.19	50.00	55.71	61.43	62.38	64.29	65.71
Living room	50.52	54.33	57.09	61.25	62.98	67.82	72.32
Store	60.63	67.94	67.94	74.92	69.84	78.10	81.27
Accuracy (%)	**63.08**	**67.56**	**74.34**	**77.44**	**76.05**	**78.31**	**81.03**

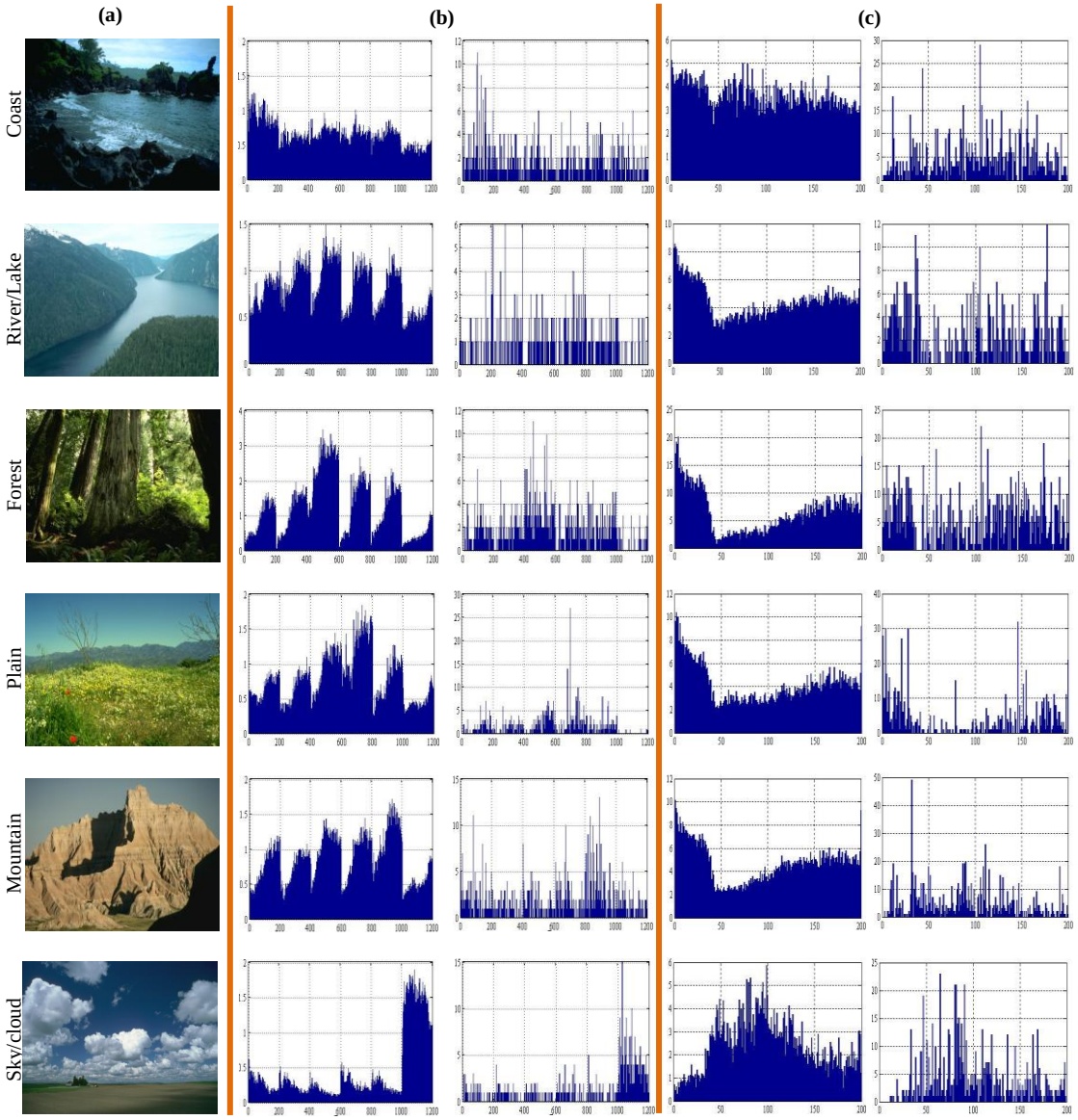


Figure 4-13: UBOW vs. IBOW for *Dataset 1* (6-classes). For each scene concept (rows), **(a)** shows sample images from the dataset **(b)** the average of IBOW histograms of all training images and the IBOW histogram for the corresponding image category and sample image, respectively **(c)** the average of UBOW histograms of all training images and the UBOW histogram for the corresponding image category and sample image, respectively. From **(b)** we can see that most image histograms tend to belong to their average histograms. Though, some classes get confused with other classes such as "river/lakes" and "mountains" since many "mountain" images contain "water" and vice versa.

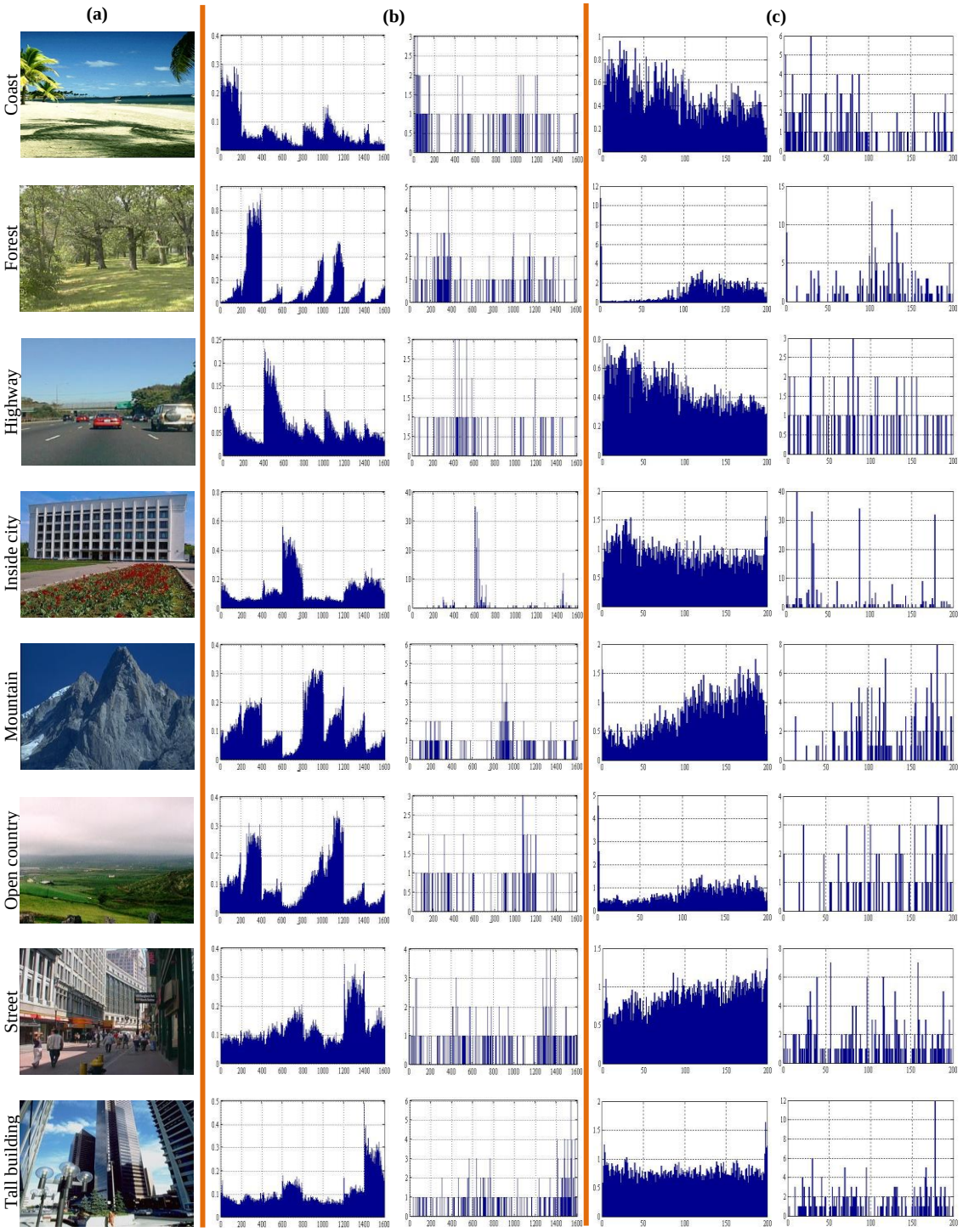
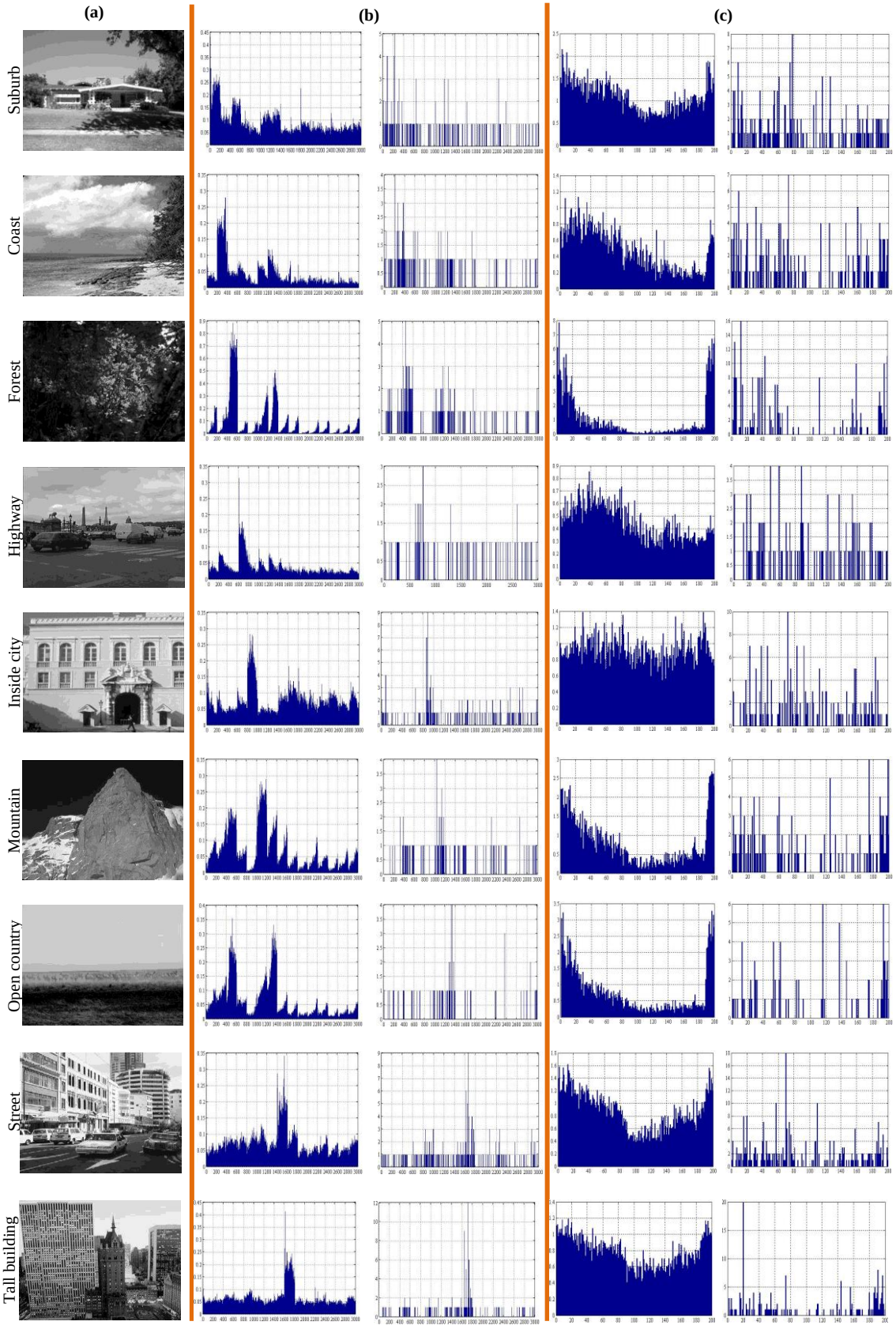


Figure 4-14: BOW vs. IBOW for *Dataset 3* (8 classes). For each scene concept (rows), (a) shows sample images from the dataset (b) the average of IBOW histograms of all training images and the IBOW histogram for the corresponding image category and sample image, respectively (c) the average of UBOW histograms of all training images and the UBOW histogram for the corresponding image category and sample image, respectively. From (b) we can see that most image histograms tend to belong to their average histograms. Though, some classes get confused with other classes such as "Open country" and "Forest" since many "Open country" images contain "trees" and vice versa.


(a)
(b)
(c)
Suburb
Coast
Forest
Highway
Inside city
Mountain
Open country
Street
Tall building

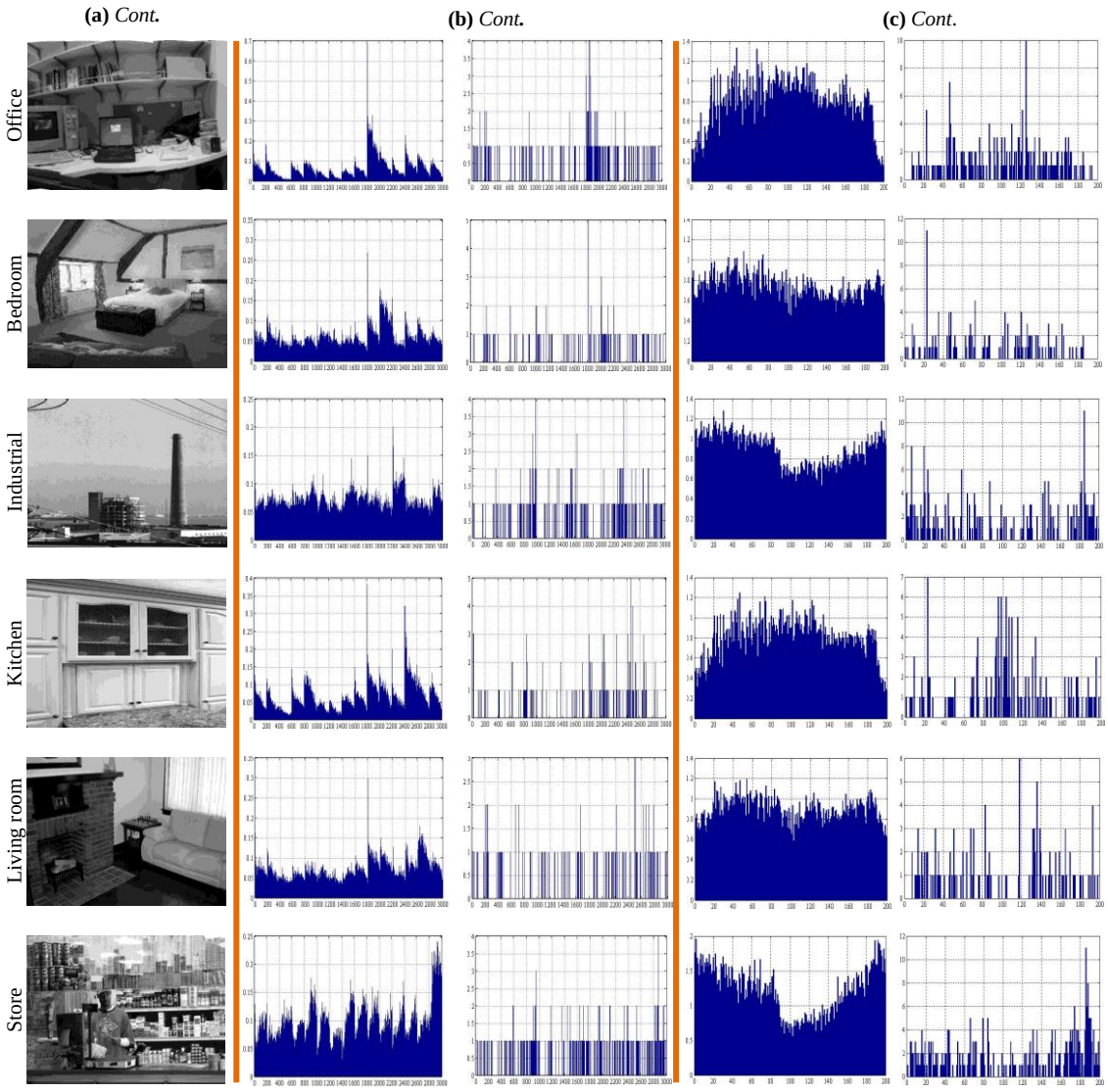


Figure 4-15: BOW vs. IBOW for *Dataset 4* (15 classes). For each scene concept (rows), (a) shows sample images from the dataset (b) the average of IBOW histograms of all training images and the IBOW histogram for the corresponding image category and sample image, respectively (c) the average of UBOW histograms of all training images and the UBOW histogram for the corresponding image category and sample image, respectively. From (b) we can see that most image histograms tend to belong to their average histograms. Though, some classes get confused with other classes such as "Living room" and "Kitchen" since both classes are indoor images and contain similar furniture. "Open country" and "Forest" is another example of confusion in their visual contents.

4.4 Summary

In this chapter, we have presented a unified framework to classify natural scene images into one of a number of predefined scene classes. Our work is based on the bag of visual words (BOW) image representation scheme. The proposed framework improved BOW image representation model in two ways: (1) It generates discriminative visual vocabularies by integrating visual vocabularies learned from class-specific data; (2) It fuses image colour information with intensity-based BOW using a spatial pyramid layout. The fusion has been done using the proposed keypoints density-based weighting (KDW) method. One of the drawbacks of using a universal visual vocabulary is that similar visual patches may be clustered into different clusters and thus loses their information. We investigated different configurations of BOW and compared their performance on three natural scene datasets. Also, we made an improvement to the well-known intensity-based Gist features by adding pyramidal colour moments in an early fusion approach. We have shown that integrated BOW (IBOW) and pyramidal colour moments (PCM) weighted on spatial pyramid layout (IPBOW+WPCM) outperformed other baseline approaches. Experimental results showed that building integrated visual vocabulary provides better performance than the conventional universal visual vocabulary. Moreover, it is obvious that building integrated visual vocabulary is faster than universal visual vocabulary, since the clustering algorithm will deal with less feature vectors and it will probably converge faster. We have also shown that visual vocabularies of one dataset could be used to generate BOW for another dataset with acceptable classification performance.

Chapter 5

Image Annotation

This chapter investigates using bag of visual words model for semantic-based image annotations at region level. Due to recent advances in developing robust local invariant detectors and descriptors, bag of visual words model has been a common choice, as an intermediate representation, to represent visual content of images for scene classification (*see* Chapter 4). Few works have addressed the use of bag of visual words for scene annotation at region level.

As discussed in Section 2.2, images are annotated at either image level or region level. In image annotation at image level, images are automatically annotated with some semantic labels which can be useful for image retrieval to search images using keywords. Such semantic labels are linked to the whole image or to image regions. Image regions can be obtained using two methods: (1) image segmentation (2) fixed grid. Due to inaccurate segmentation algorithms, fixed grid methods are preferable. In this chapter, image regions are obtained using fixed grid method. The

task of automatic image annotation at region level is to annotate image regions or blocks with semantic labels, such as *sky*, *water* and *grass*, and the image is then annotated with a fixed size semantic feature vector of semantic labels.

To this point, this chapter presents a framework for automatic natural scene image annotation with local semantic labels from a constrained vocabulary. The framework is based on a hypothesis that, in natural scenes, intermediate semantic concepts are correlated with the local keypoints. The hypothesis is justified by analyzing the distribution of local semantic concepts in images and the distribution of local keypoints detected in the regions labelled with these semantic concepts. Based on this hypothesis, image regions can be efficiently represented by BOW model and using a discriminative learning approach, such as SVM, to annotate image regions with semantic labels. The contributions related to this chapter can be summarized as follows:

1. A hypothesis is proposed in this chapter, which studies the correlation between the distribution of semantic concepts and local keypoints located in image regions labeled with these semantic concepts.
2. In Chapter 4, integrated visual vocabularies have shown to be more discriminative to build bag of visual words histograms than universal visual vocabularies, for natural scene classification task. For *m* scene categories, visual vocabularies were generated from images in each scene category and then integrated into a single visual vocabulary called integrated visual vocabulary. Therefore, in image annotation, representing image regions that belong to *n* semantic concepts, where $n>m$, requires *n* visual vocabularies generated from local features of

all training image regions of each semantic concept. To address this problem, and based on the hypothesis aforementioned, this chapter investigates the plausibility of using visual vocabularies generated from scene categories, presented in Chapter 4, to build BOW for image regions without the need to re-build visual vocabularies again from image regions. This is called *Local from Global* approach.

3. Investigate the performance of using multiple features with BOW to represent image regions and then use these representations to train a classifier to label image regions with semantic concepts. In this chapter, two classifiers are used and their performances are compared: support vector machines and the K-NN classifiers.
4. Finally, to study the influence of generating visual vocabularies from image halves on the performance of image annotation at region level. The assumption here is that the top half of natural scene images may share relevant information or local features that rarely exist at the bottom half. For example, semantic concepts such as *rocks* and *sky* are usually located at the top half of natural scenes whereas *grass* and *sand* are usually located at the bottom half. To this end, building visual vocabularies from each half of all training images can lead to more discriminative visual vocabularies and, at the same time, it reduces the time of clustering process.

The rest of this chapter is organized as follows. In Section 5.1, a description of the image dataset used in this chapter is first presented. The correlation between the distribution of local semantic concepts and local keypoints is presented in

Section 5.2. Next, the proposed framework for automatic image annotation with local semantic concepts is demonstrated in Section 5.3. Experimental work and results are presented in Section 5.4. The chapter ends with a short summary in Section 5.5.

5.1 Natural Scene Dataset

This chapter addresses the problem of annotating image regions with semantic concepts, named local semantic concepts. To perform this task, this chapter uses a dataset of 700 images of natural scenes grouped into six natural scene categories (Vogel and Schiele, 2004). These categories are *coasts*, *river*/*lakes*, *forests*, *plains*, *mountains,* and *sky*/*clouds* (*see* Chapter 4, Section 4.3.2). The dataset has been annotated manually with nine local semantic concepts: *sky*, *water*, *grass*, *trunks*, *foliage*, *field*, *rocks*, *flowers* and *sand*. To obtain a ground truth for local concept annotation, images were divided into a grid of (10x10) regions which results in 70000 local regions (*see* Figure 5-1). In the original work of Vogel and Schiele (Vogel and Schiele, 2004), image regions with more than one semantic concepts were disregarded in training and testing local semantic classifiers For example, in Figure 5-1, the image region at the fourth row and fifth column shares two semantic concepts (*rocks* and *sky*) and thus this region does not contribute in learning the classifier. The natural scene dataset used in this chapter has two folders. The first folder contains 700 natural scene images. The second folder contains 700 text files, where each file corresponds to one of the natural scene images in the first folder. Let $I=\{i_1,i_2,\ldots,i_N\}$ be the set of images in the first folder and $T=\{t_1,t_2,\ldots,t_N\}$, where N=700. A text file t_j contains a list of annotations for an image i_j, where $j=1,2,\ldots,N$.

Table 5-1 shows the distribution of local semantic concepts over the six scene categories. Exemplary annotated images of each scene category are shown in Figure 5-2.

Table 5-1: Sizes of the nine local concept classes located in each scene category. For example, scene category '*Coasts*' contains 2960 regions labeled with semantic concept '*Sky*'.

	Sky	**Water**	**Grass**	**Trunks**	**Foliage**	**Field**	**Rocks**	**Flowers**	**Sand**
Coasts	2960	4326	430	32	1284	194	1922	46	825
Rivers/lakes	1728	2826	273	82	2629	204	1553	12	0
Forests	335	39	465	1419	6464	309	47	31	0
Plains	2879	14	1608	36	964	2649	330	1898	897
Mountains	4416	54	649	56	2335	735	7401	62	23
Sky/clouds	2978	34	78	0	33	97	57	0	0
# of image regions	**15296**	**7293**	**3503**	**1625**	**13709**	**4188**	**11310**	**2049**	**1745**
OVERALL					**60718**				

5.2 Local Semantic Concepts and Local Keypoints

This section investigates the correlation between local semantic concepts and local keypoints, based on their distributions over all image regions. Local semantic concepts are labels assigned to image regions. Using (10x10) fixed grid layout, an image is divided into 100 image regions. Low-level features, such as colour and texture, are normally used to represent the visual content of image regions. These features are not invariant to different changes of the visual contents. Thus, to provide invariance to changes in illumination, rotation, etc, interest points can be used (Mikolajczyk and Schmid, 2005). Interest points or local keypoints correspond to image structures that are considered important.

Using sparse representation approach, a keypoint detector, such as DoG detector (Lowe, 2004), locates local keypoints that contain distinctive information in their surrounding area and should be invariant to geometric transformations. These keypoints are then described using robust and informative features such as SIFT. These features are normally used in the BOW model for image classification.

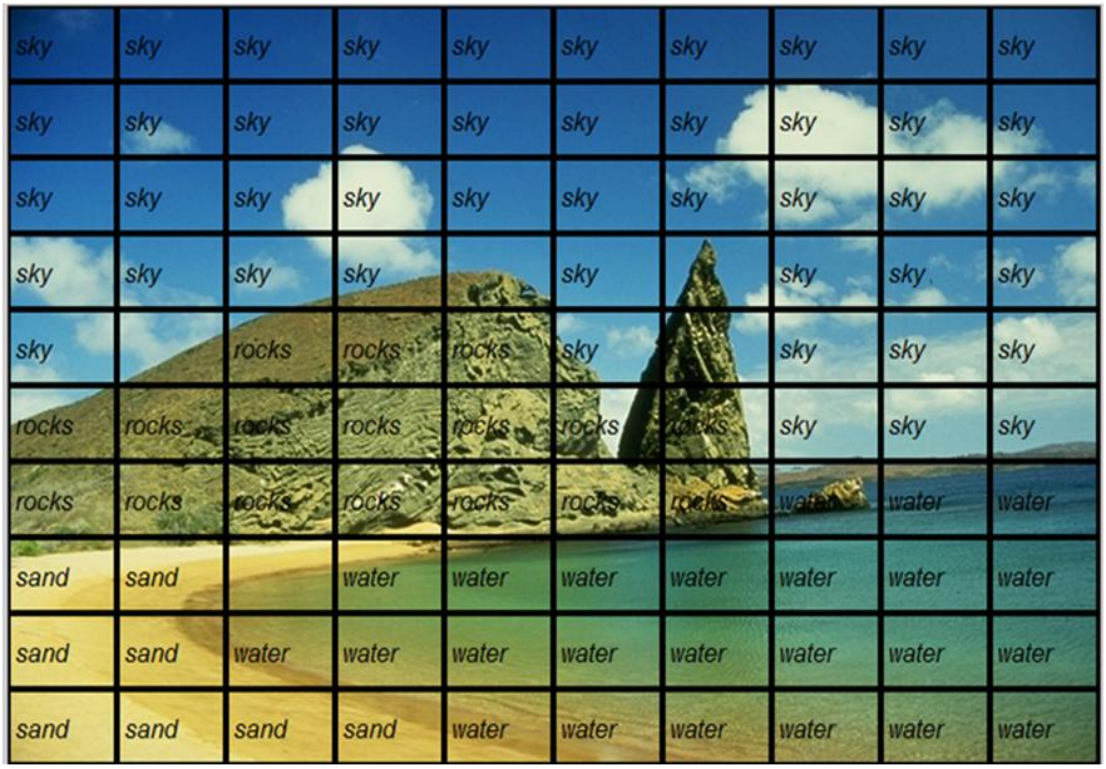
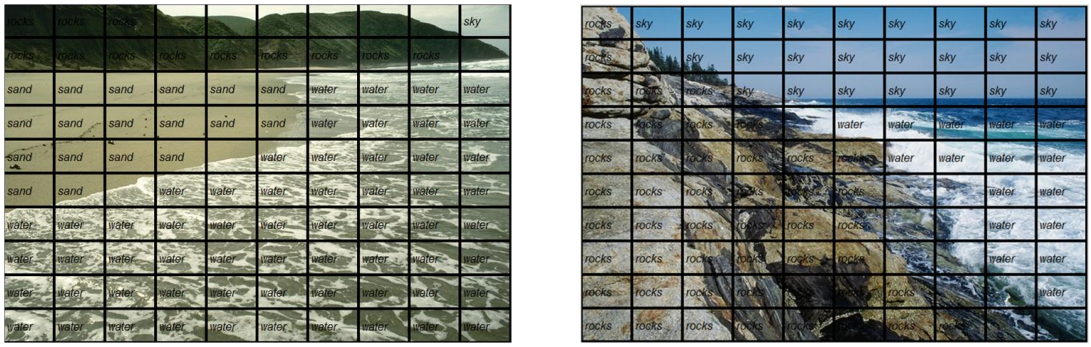
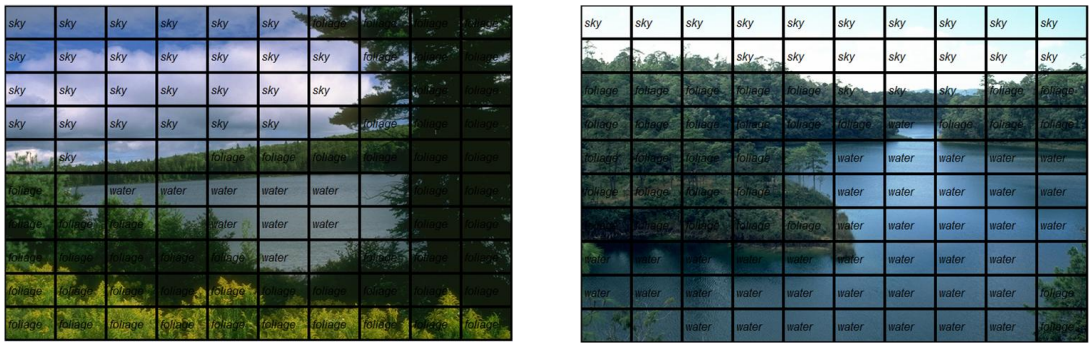


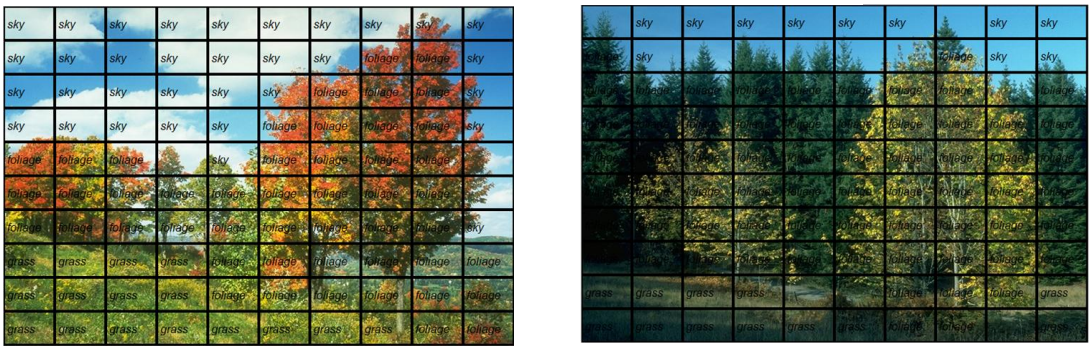
Figure 5-1: A sample image from the 'coast' scene category. Image regions are manullay annotated with semantic concepts. Image regions that contain more than once sematnic concepts are discarded in the annotation process.



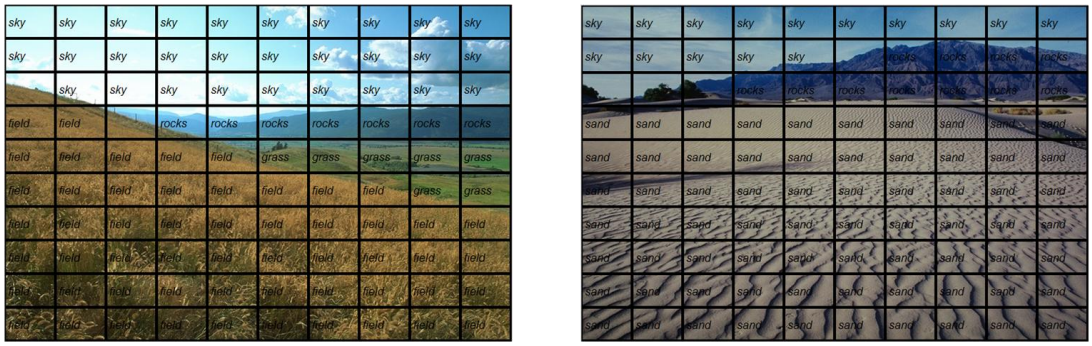
(a) Coasts



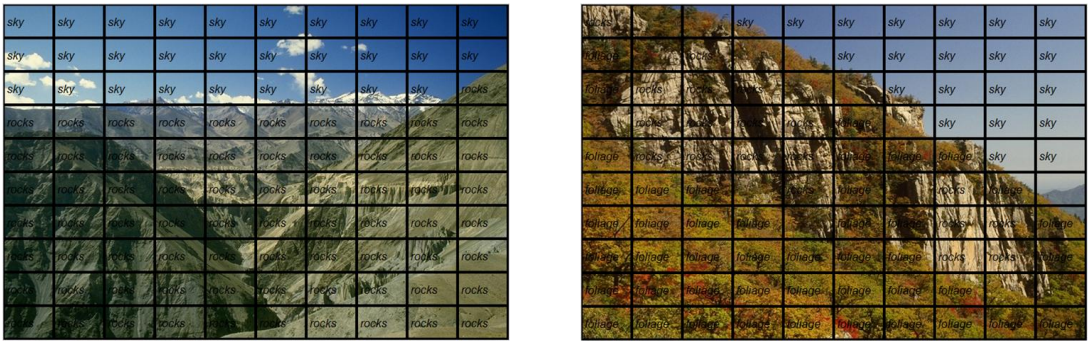
(b) Rivers/lakes



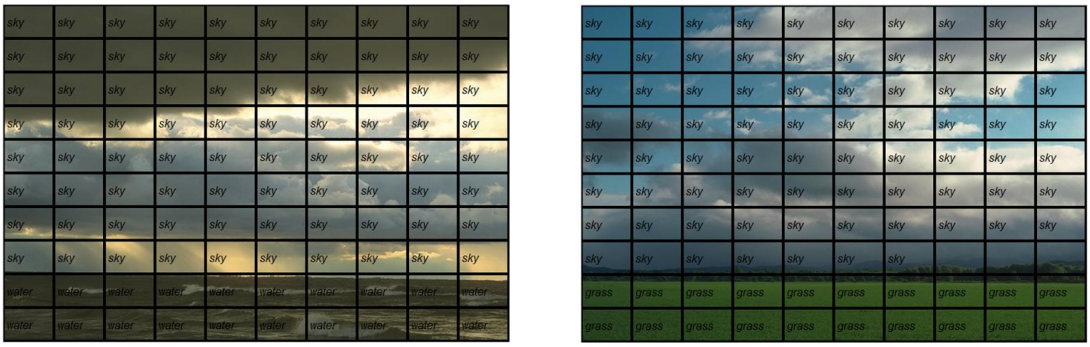
(c) Forests



(d) Plains



(e) Mountains



(f) Sky/clouds

Figure 5-2: Exemple of images from each scene category. Each row contains two images selected from the same category. Image regions are manually annotated with local semantic concepts.

Having the ground truth semantic labels provided with the natural scene dataset, and the local keypoints provided by an interest point detector, two facts can be drawn:

1. Each region has a particular location with coordinates (x_1,y_1) for the top left corner and (x_2,y_2) for the bottom right corner, such that (x_2-x_1)x(y_2-y_1) is the dimension of the given region.
2. The output of keypoint detector is a list of coordinates of all keypoints in an image. The detector also describes some other

characteristics of the area around each point, such as scale and orientation. Coordinates of keypoints are our interest here.

Given the coordinates of all image regions and coordinates of all detected keypoints, it is possible to assume that the distribution of semantic labels, used to annotate images regions, correlates with the distribution of local keypoints for each natural scene category. This hypothesis can be justified by counting, for each natural scene category and semantic concept, the number of keypoints with coordinates located in the areas of all image regions labelled with the same semantic concept. If the distribution of semantic concepts is similar or close to the distribution of local keypoints then there is a possible correlation between them.

Figure 5-3 shows the distribution of local keypoints located in image regions of each semantic concept and over all scene categories. For example, there are 179131 keypoints with coordinates located in image regions labelled with the semantic label *foliage*. The correlation between semantic concepts and local keypoints is shown in Figure 5-4. The two lines (Red and Blue) shown in this figure represents the percentage (%) of occurrence for each semantic concepts and local keypoints over all scene categories. Also, the distributions of local semantic concepts and local keypoints over each of the six natural scenes are depicted in Figure 5-5 and Figure 5-6. The distribution of local semantic concepts at the upper and lower halves of all images is shown in Figure 5-7 whereas the distribution of local keypoints in image regions labelled with the semantic concepts is shown in Figure 5-8.

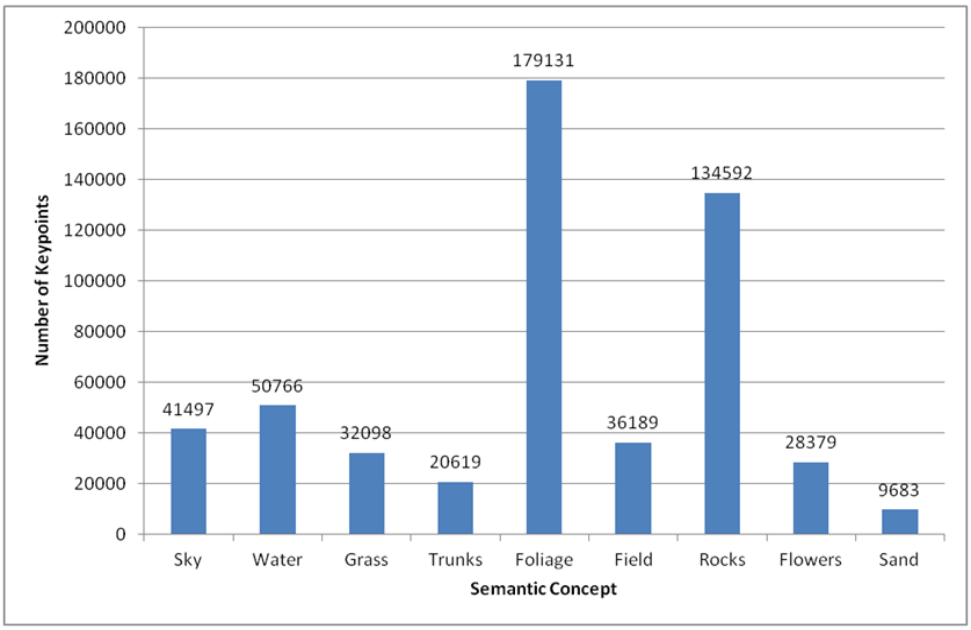


Figure 5-3: Distribution of local keypoints detected in image regions of each semantic concept over all scene categories.

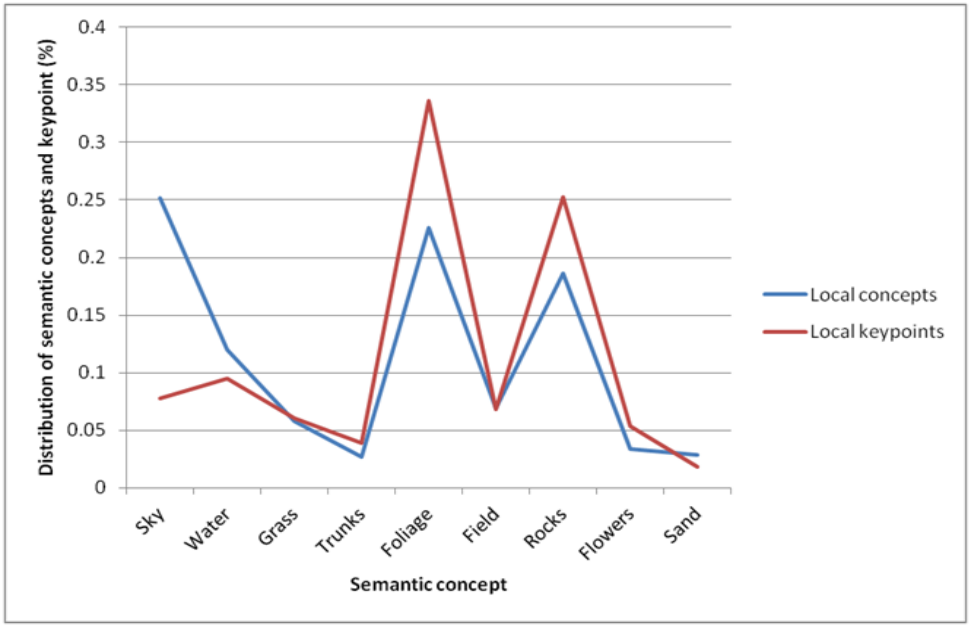
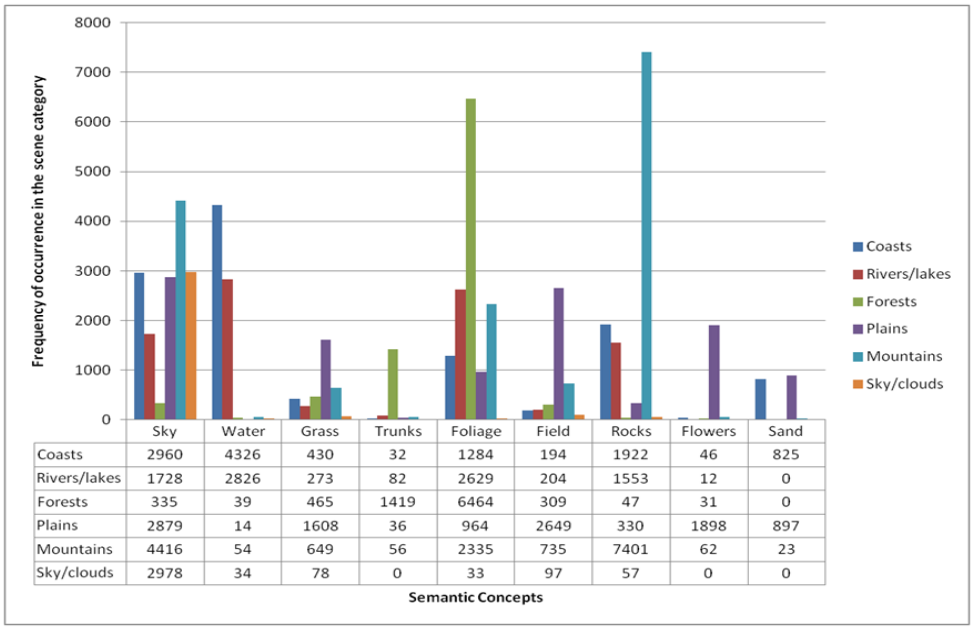
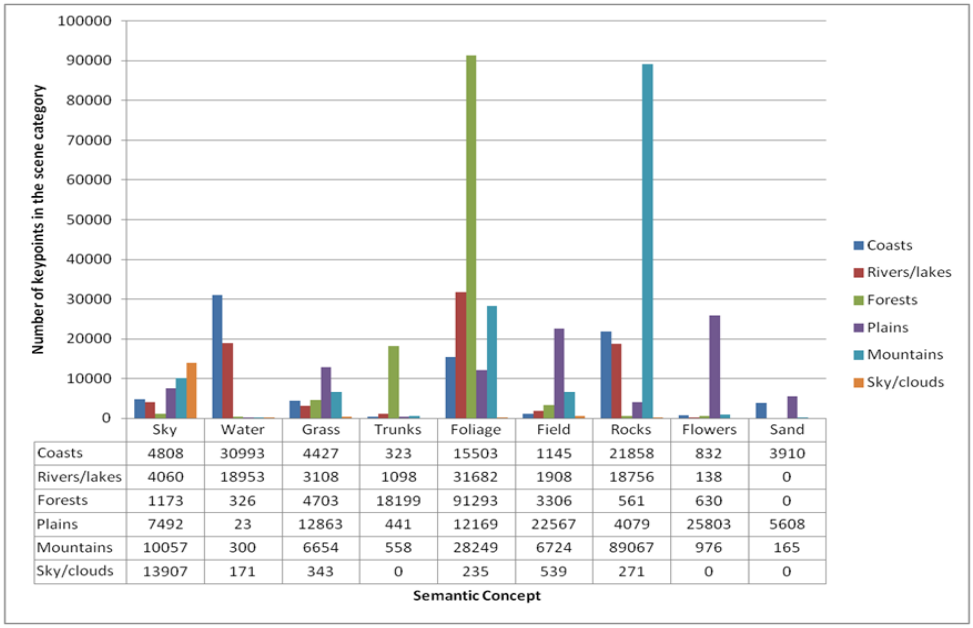


Figure 5-4: the correlation between the distributions (%) of local semantic concepts and local keypoints



	Sky	Water	Grass	Trunks	Foliage	Field	Rocks	Flowers	Sand
Coasts	2960	4326	430	32	1284	194	1922	46	825
Rivers/lakes	1728	2826	273	82	2629	204	1553	12	0
Forests	335	39	465	1419	6464	309	47	31	0
Plains	2879	14	1608	36	964	2649	330	1898	897
Mountains	4416	54	649	56	2335	735	7401	62	23
Sky/clouds	2978	34	78	0	33	97	57	0	0

Figure 5-5: Distribution of each semantic concept over each scene category.



	Sky	Water	Grass	Trunks	Foliage	Field	Rocks	Flowers	Sand
Coasts	4808	30993	4427	323	15503	1145	21858	832	3910
Rivers/lakes	4060	18953	3108	1098	31682	1908	18756	138	0
Forests	1173	326	4703	18199	91293	3306	561	630	0
Plains	7492	23	12863	441	12169	22567	4079	25803	5608
Mountains	10057	300	6654	558	28249	6724	89067	976	165
Sky/clouds	13907	171	343	0	235	539	271	0	0

Figure 5-6: Distribution of keypoints located in regions of each semantic concept and over each scene category.

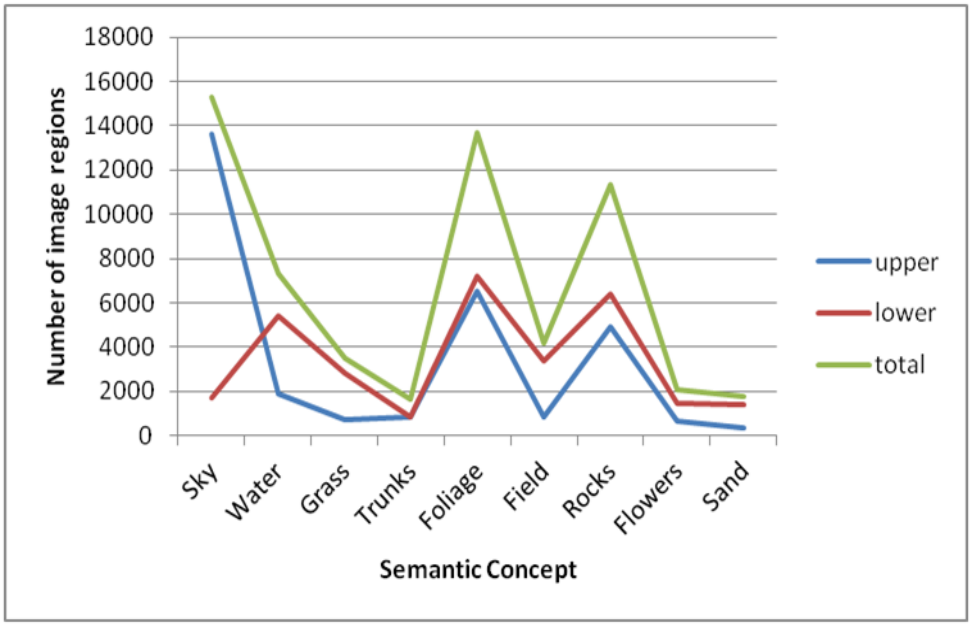


Figure 5-7: Distribution of image regions located in the upper and lower halves of images.

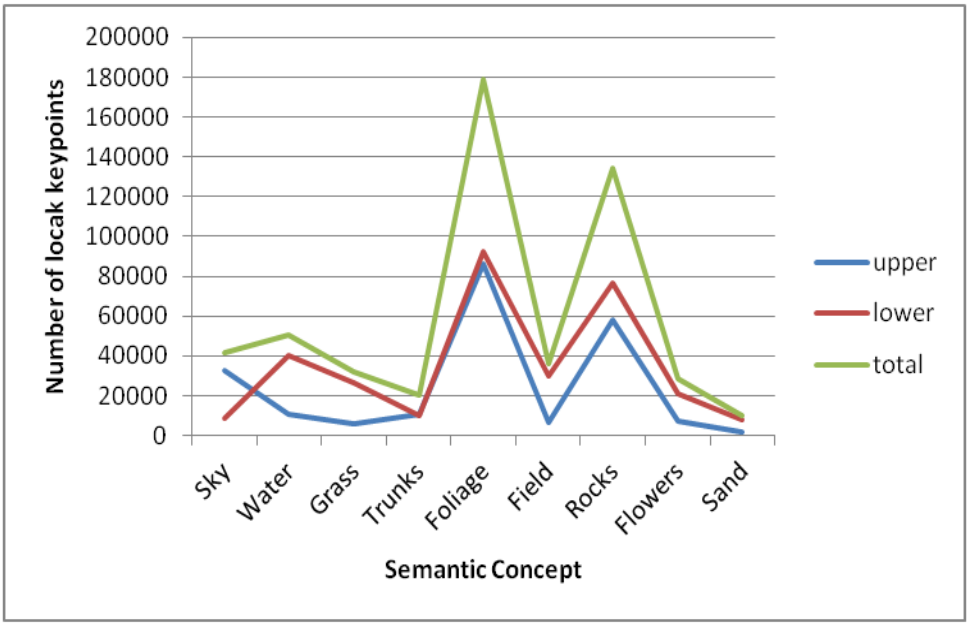


Figure 5-8: Distribution of local keypoints found in image regions in the upper and lower halves of images.

5.3 Image Annotation Framework

In this section, a framework for annotating image regions with local semantic concepts based on bag of visual words is presented. The framework consists of two parts as depicted in Figure 5-9. In the *first* part, DoG detector and SIFT descriptor are employed to find and represent local keypoints in all images of the dataset. Each SIFT descriptor is a vector of size 128-D. For each natural scene category, k-means algorithm is applied to all descriptors to build visual vocabularies. These vocabularies are then aggregated to form an integrated visual vocabulary of size (K×M) where K is the vocabulary size and M is the number of scene categories.

Visual vocabularies generated in Chapter 4, and in particular for the six scene categories dataset mentioned in Section 5.1, are used in this chapter. Instead of building visual vocabularies from local features located in image regions of each semantic concept, this chapter study the discriminative power of visual vocabularies generated from scene categories to represent local semantic concepts through the use of bag of visual words. This approach is referred to as L*ocal* from *Global*.

In the *second* part, bag of visual words is used to represent image regions. A frequency histogram is generated from each image region where the number at each bin corresponds to the frequency of occurrence of each visual word in that image region. Bag of visual words generated from image regions are call Concept-based Bag of Visual Words (CBOWs). Having all image regions represented by bag of visual words, and given that image regions are annotated with local semantic concepts, these CBOW histograms are used to train classifiers, such as SVM and kNN, to annotate image regions with semantic concepts in the test dataset.

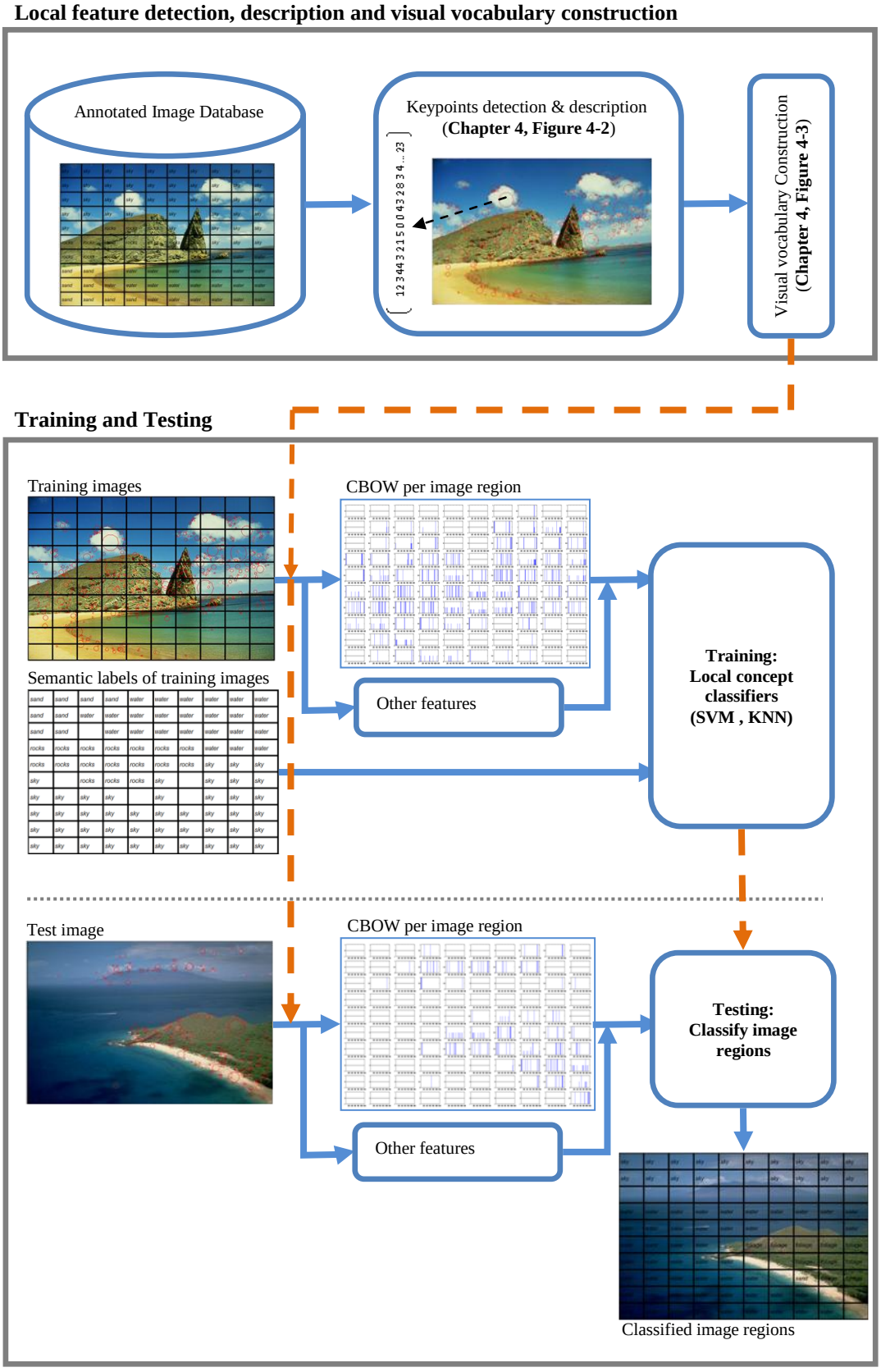


Figure 5-9: Flow diagram of the proposed framework for local semantic annotation

5.3.1 Scene visual vocabulary construction

The visual vocabulary on which CBOWs rely on is of great importance. Building visual vocabularies from each scene category have shown improvements in the performance of natural scene classification task. To apply the same approach in building visual vocabularies to local semantic concepts, the number of visual vocabularies becomes larger. This is because of the fact that the number of local semantic concepts are usually larger than the number of scene categories these semantic concepts belong to. Therefore, it is worth to explore the influence of applying visual vocabularies obtained from natural scene categories (globally obtained from natural categories) to map SIFT features located in image regions (locally used to represent regions) to the indices of visual words and thus build CBOWs histograms.

Another choice of improving the power of visual vocabulary is to incorporate spatial information about local keypoints when clustering their features. Natural scene images, such as *coasts*, contain semantic concepts that usually appear in common places. For example, the *sand* concept can be found at the bottom of an image whereas the *sky* and *water* concepts appear at the top. Also, building visual vocabularies using k-means algorithm may lead to group dissimilar local keypoints in the same cluster. To avoid this problem, images are partitioned into two halves of equal size: upper and lower. An image of size W×H is divided into two halves, each has a dimension of (W/2)×(H/2). In this case, two integrated visual vocabularies are generated: Upper integrated visual vocabulary and Lower integrated visual vocabulary. Upper integrated visual vocabulary is generated by clustering SIFT features located at the upper half of all training images. Lower integrated visual

vocabulary is generated by clustering SIFT features located at the lower half of all training images. This approach has two main advantages:

1. Create more representative visual vocabularies about natural scenes that benefits from the spatial information of the keypoints and at the same time reduce ambiguities between clusters resulted from the clustering step.
2. Reduces the clustering time while maintaining or improving the discriminative power of the clusters. Preliminary experimental work has shown that quantizing local features at different spatial levels reduce the clustering time as well as improving the discriminative power of visual words or clusters. But, it increases the dimensionality of the BOW.

Comparing the annotation performance of the proposed framework using universal visual vocabulary, integrated visual vocabulary and upper and lower visual vocabularies allows us to analyse indirectly the dependence of bag of visual words representation on spatial locations of local keypoints as well as the influence of using global visual vocabularies to represent image regions.

5.3.2 Image region representation

In this section, two types of features are used to represent the visual content of image regions. In the first type, image regions are represented using concept-based bag of visual words (CBOWs). These CBOWs are generated using different visual vocabularies, as described in the previous section. Multiple low-level features

are used as baseline methods and also to expand the discriminative power of CBOWs.

5.3.2.1 Concept-based bag of visual words (CBOWs)

This section is dedicated to representing visual contents of image regions using the bag of visual words model. In this chapter, bag of visual words generated at region level are called concept-based bag of visual words (CBOWs). Images in the dataset are firstly divided into 10×10 sub-regions. To construct a CBOW histogram from an image region four steps are required. First, local keypoints are automatically located in the image. Second, local descriptors are extracted from regions defined around those local keypoints. Third, build visual vocabularies from local descriptors, as mentioned in Section 5.3.1. Fourth, local keypoint are mapped to the index of the closest visual word and their occurrences are counted to build the CBOW for the image sub-region *r*. Four types of visual vocabularies are used in this chapter to build the CBOWs from image region. Their performances to annotate image regions are then compared and analyzed in the experimental work section. These visual vocabularies are:

- *Universal visual vocabulary*: this vocabulary was obtained in chapter 4. This visual vocabulary contains 200 visual words.
- *Integrated visual vocabulary*: this vocabulary was obtained in chapter 4. This visual vocabulary contains 1200 visual words.
- *Universal Upper and Lower visual vocabulary*: the upper visual vocabulary contains 200 visual words and the lower visual vocabulary contains 200 visual words.

- *Integrated Upper and Lower visual vocabulary*: the upper visual vocabulary contains 1200 visual words and the lower visual vocabulary contains 1200 visual words.

Let $I = \{i_1, i_2, ..., i_N\}$ be the set of all images in the dataset and that each image i_k is divided into 10×10 sub-images such that $i_k = \{i_{k_1}, i_{k_2}, ..., i_{k_{100}}\}$, where i_{k_r} is the *r*-th region in image i_k. In Algorithm 5.1, universal visual vocabulary generated for the natural scene dataset with 6 scene categories is used to build CBOWs from image regions. This vocabulary has been used in Chapter 4 for image classification. This vocabulary contains 200 visual words generated by clustering all SIFT features extracted from training images.

Algorithm 5.1: An algorithm to build CBOWs from images regions using universal visual vocabulary.

Input: Use the *universal visual vocabulary V*, generated in Chapter 4, to construct CBOWs from local image regions as follows:

For all images $I = \{i_1, i_2, ..., i_N\}$ in the dataset **Do**:

a. Apply DoG feature detection technique to locate interest points in image i_k.
b. Extract SIFT features (128-D) from each located keypoint.
c. **For** each image region i_{k_r} in the image i_k **Do**:

Quantize all SIFT descriptors located in region i_{k_r} into one of the *M* visual words, where *M* is the size of the visual vocabulary *V*. An image region i_{k_r} is then represented as a histogram h_{k_r} of the frequencies of visual words.

end

end

Output: a collection of h_{k_r} histograms, where each h_{k_r} is the CBOW for image *k* at region *r*.

Algorithm 5.2: An algorithm to build CBOWs from images regions using integrated visual vocabulary.

Input: Use the *integrated visual vocabulary V*, generated in Chapter 4, to construct CBOWs from local image regions as follows:

For all images $I = \{i_1, i_2, ..., i_N\}$ in the dataset **Do**:

a. Apply DoG feature detection technique to locate interest points in image i_k.
b. Extract SIFT features (128-D) from each located keypoint.
c. **For** each image region i_{k_r} in the image i_k **Do**:

Quantize all SIFT descriptors located in region i_{k_r} into one of the *M* visual words, where *M* is the size of the visual vocabulary *V*. An image region i_{k_r} is then represented as a histogram h_{k_r} of the frequencies of visual words.

end

end

Output: a collection of h_{k_r} histograms, where each h_{k_r} is the CBOW for image *k* at region *r*.

In Algorithm 5.2, the integrated visual vocabulary generated by clustering all SIFT features of training images over each scene category is used to build CBOWs. This vocabulary has been used in Chapter 4 for the same dataset. It is important to study the influence of building visual vocabularies from local features located in parts of an image rather the whole image and compare their performances with the traditional visual vocabularies. As mentioned in Section 5.3.1, clustering local descriptors located at the upper halves of images may generate a better quality clusters, i.e., their members are more semantically similar. It aims to reduce inter-class similarity and increase intra-class similarity throughout clustering features at image parts. Although images can be divided to any number of tiles, only two parts from images are considered in this chapter; the upper half and the lower half. In

other words, 50% of the image content, the upper part, is used to build the upper visual vocabulary whereas the other 50%, the lower part, is used to build the lower visual vocabulary. Upper and lower visual vocabularies are generated at two levels. The first level considers building upper and lower visual vocabularies from SIFT descriptors of all scene categories whereas in the second level visual vocabularies are generated from SIFT descriptors of each scene category, as shown in Figure 5-10.

Having all visual vocabularies generated from upper and lower halves of images, the next step is to use them to build CBOWs from image regions. Algorithms 5-3 and 5-4 illustrate the required steps to build CBOWs from images regions at upper halves of images. To avoid repeating the same algorithms for the lower halves, lower visual vocabularies and image regions at the lower halves (*r=51,51,…,100)* are replaced with the corresponding ones used in Algorithms 5-3 and 5-4.

Algorithm 5.3: An algorithm to build CBOWs from images regions, at *upper halve*, using universal visual vocabulary.

Input: Use the *upper universal visual vocabulary V*, to construct CBOWs from local image regions as follows:

For all images $I = \{i_1, i_2, ..., i_N\}$ in the dataset **Do**:

a. Apply DoG feature detection technique to locate interest points in image i_k.
b. Extract SIFT features (128-D) from each located keypoint.
c. **For** each image region i_{k_r} in the upper halve of image i_k**,** where *r=1,2,…,50* **Do**:

Quantize all SIFT descriptors located in region i_{k_r} into one of the *M* visual words, where *M* is the size of the visual vocabulary *V*. An image region i_{k_r} is then represented as a histogram h_{k_r} of the frequencies of visual words.

end

end

Output: a collection of h_{k_r} histograms, where each h_{k_r} is the CBOW for image *k* at region *r*.

Algorithm 5.4: An algorithm to build CBOWs from images regions, at *upper halve*, using integrated visual vocabulary.

Input: Use the *upper integrated visual vocabulary V*, to construct CBOWs from local image regions as follows:

For all images $I = \{i_1, i_2, ..., i_N\}$ in the dataset **Do**:

a. Apply DoG feature detection technique to locate interest points in image i_k.
b. Extract SIFT features (128-D) from each located keypoint.
c. **For** each image region i_{k_r} in the upper halve of image i_k**,** where *r=1,2,…,50* **Do**:

Quantize all SIFT descriptors located in region i_{k_r} into one of the *M* visual words, where *M* is the size of the visual vocabulary *V*. An image region i_{k_r} is then represented as a histogram h_{k_r} of the frequencies of visual words.

end

end

Output: a collection of h_{k_r} histograms, where each h_{k_r} is the CBOW for image *k* at region *r*.

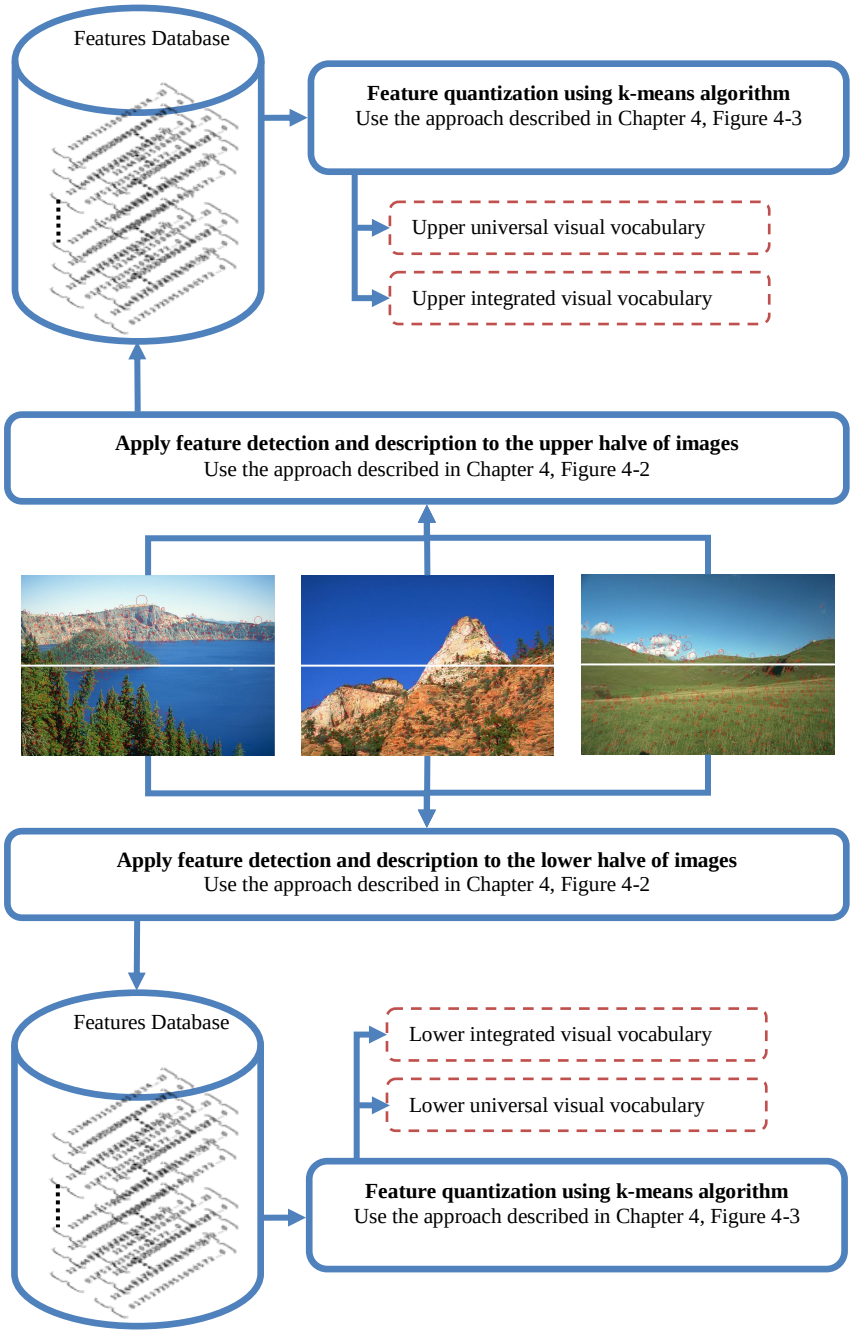


Figure 5-10: Upper and lower visual vocabularies construction. Images in the middle are samples from all training images used in the construction process.

5.3.2.2 Local from Global CBOWs

Constructing CBOWs from visual vocabularies generated from each scene category allows us to study the relationship between natural scene categories and local semantic concepts. The principle of building integrated visual vocabulary is to quantize local features from each scene category. Thus, in the case of local semantic concepts, integrated visual vocabularies should be generated from local features located in image regions labeled with theses semantic concepts.

The main aim of this section is to analyze the relationship between visual vocabularies, generated from scene categories, and CBOWs histograms generated from local image regions. The analysis is based on the distribution of all CBOWs histograms generated from image regions labeled with each of the semantic concepts. Also, we show that integrated visual vocabularies are more suitable to represent local features in image regions rather than the universal visual vocabulary. Furthermore, the distributions of CBOWs generated from upper and lower integrated visual vocabularies are analyzed.

As mentioned in Section 5.1, the natural scene dataset used in this chapter contains six scene categories and nine semantic concepts. To analyze the relationship between a visual vocabulary and the semantic concepts, all CBOWs histograms generated from a visual vocabulary are summed up. Given that all image regions are annotated with semantic concepts and that each image region is represented by a CBOW then it is possible to sum up all CBOWs histograms for each semantic concept. For example, the distribution of all CBOWs generated from image regions labeled with the semantic concept *sky*, using universal visual vocabulary, is shown in Figure 5-11. The same figure shows the distributions of all CBOWs for *water*,

grass, *trunks*, *foliage*, *field*, *rocks*, *flowers*, and *sand*. These distributions do not show any relation of the natural scene categories with the local semantic concepts.

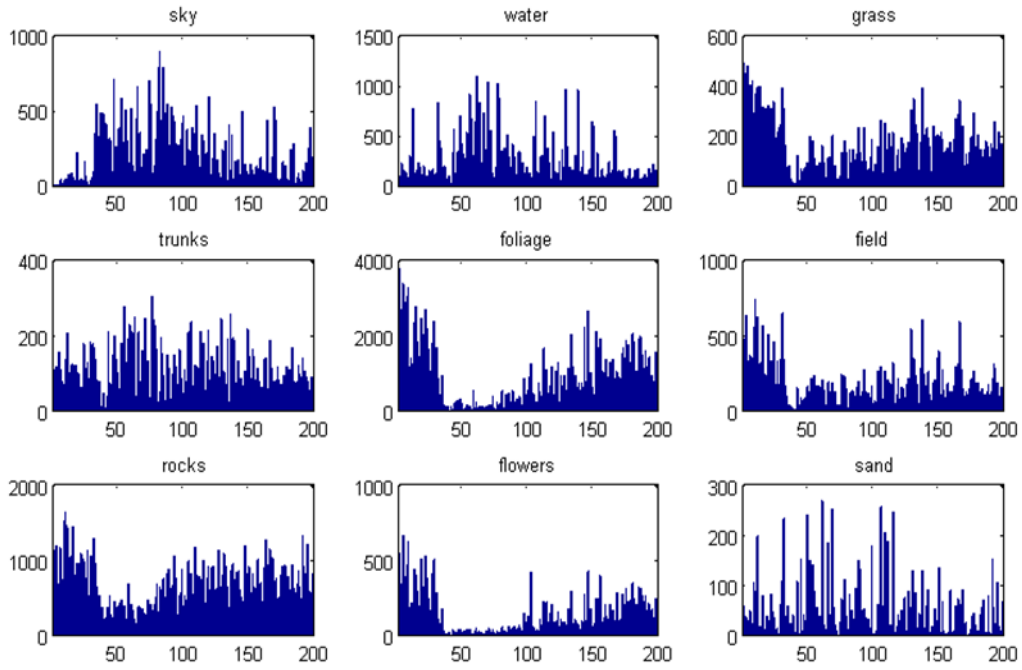


Figure 5-11: Sum of the CBOW histograms obtained using *universal* visual vocabulary at image level.

In contrast, Figure 5-12 shows CBOWs distributions of image regions using integrated visual vocabulary. The integrated visual vocabulary contains 1200 visual words in which each 200 visual words represent 200 clusters generated from clustering all local descriptors located in images of specific scene category. The first 200 visual words represent clusters for the natural scene *coasts*, and the other 1000 visual words are clusters for the natural scene categories *river*/*lakes*, *forests*, *plains*, *mountains* and *sky/clouds*, respectively.

This figure depicts interesting relationship between local semantic concepts and the natural scene categories. For example, large numbers of the local keypoints, found in image regions labeled with the semantic concept *sky*, are assigned to the last 200 visual words of the CBOWs histograms. The last 200 visual words are clusters belonging to the scene category *sky/clouds*. And it is usual to see *sky* areas in *sky/clouds* natural scenes. Another interesting indication in the figure is the plot for the semantic concept *water*. For this concept, large numbers of the keypoints are assigned to the first 200 visual words which actually belong to the natural scene *coasts*, and this is natural to have water in natural scene *coasts*. The concept *water* is also available in the *river/lakes* scene category depicted in the same plot. The same figure contains the distributions for the semantic concept *sand*. Many of the keypoints are located in the first 200 visual words and visual words (600-799).

To this end, there is a relationship between natural scene categories and the semantic concepts they contain which are analyzed using the distributions of CBOWs obtained using the integrated visual vocabulary (*see* Figure 5-12) and as has been discussed above. Moreover, these relationships confirm the hypothesis, presented in Section 5.2, in a way that image regions labeled with a semantic concept contain keypoints that are more likely to appear in the distributions of the CBOWs histograms for that semantic concept. Furthermore, Figure 5-12 shows the plausibility of using visual vocabularies, generated from local descriptors at image level, to build CBOWs. This will be confirmed in the experimental results, section 5.4.2.

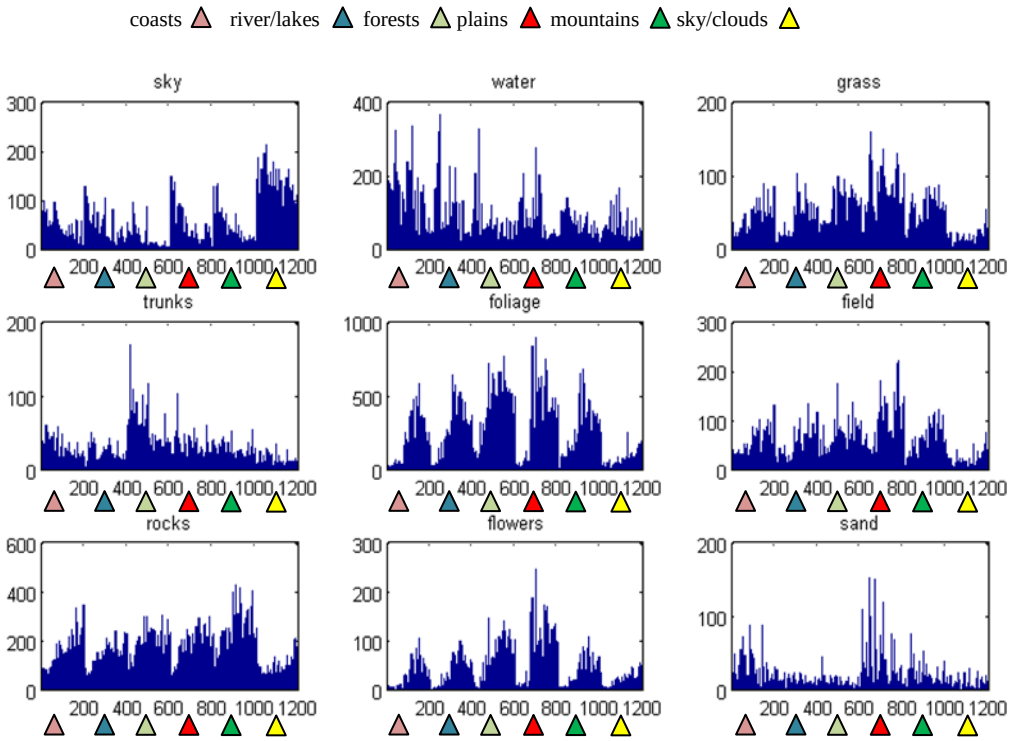


Figure 5-12: Sum of the CBOW histograms obtained using *integrated* visual vocabulary at image level.

In the case of building visual vocabularies from local keypoints descriptors located at the upper and lower halves of images, the same approach is followed to analyze their relationship with the local semantic concepts. For the upper halves of images, two visual vocabularies are generated: universal and integrated visual vocabularies from which CBOWs histograms are generated. For the lower halves of images, two visual vocabularies are also constructed: universal and integrated visual vocabularies.

Similar to using universal visual vocabulary in Figure 5-11, Figure 5-13 and Figure 5-14 show that the distributions of the CBOWs over the nine semantic concepts do not show any indication from the distributions of the local semantic concepts on the upper and lower halves of images. Therefore no relationship can be inferred from both halves, though for image annotation, the performance of upper and lower CBOWs may still outperform CBOWs at image level. This will be discussed in the experimental work, section 5.4.

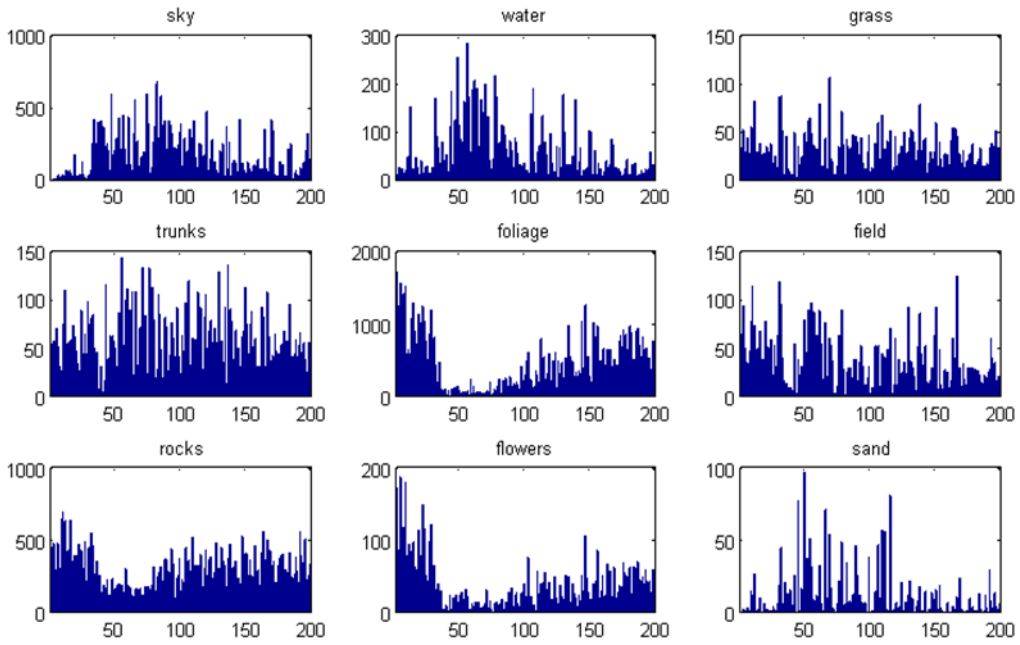


Figure 5-13: Sum of the CBOW histograms obtained using *universal* visual vocabulary at the *Upper* half of the images.

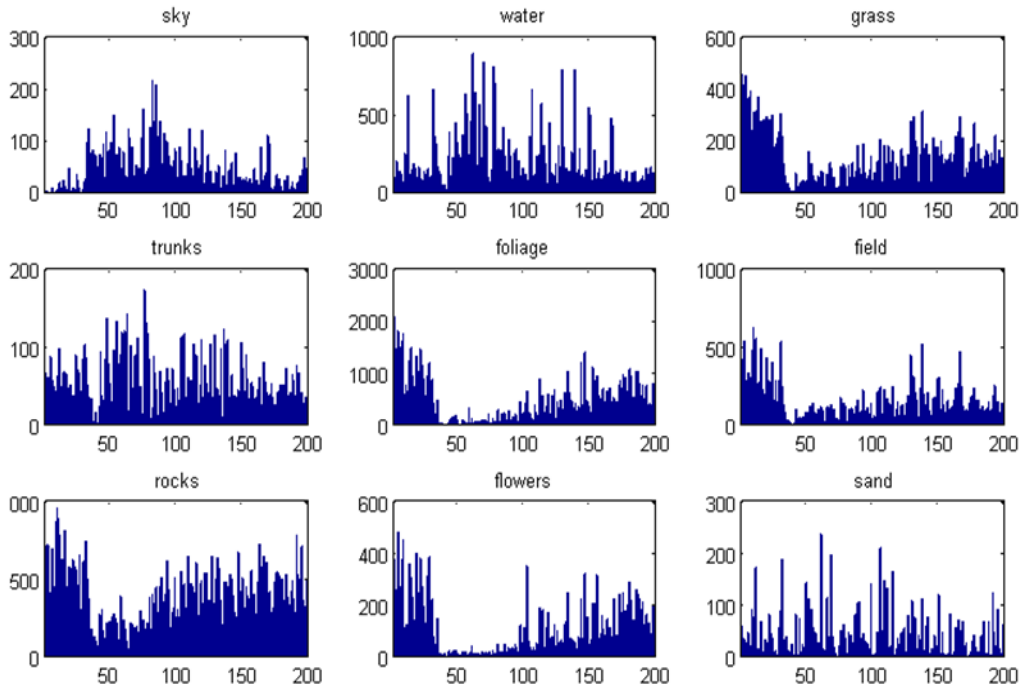


Figure 5-14: Sum of the CBOW histograms obtained using *universal* visual vocabulary at the *Lower* half of the images.

To reduce the ambiguity between visual words and improve their discriminative power, integrated visual vocabularies are generated from image halves. This is confirmed by analyzing the distributions of local keypoints in the CBOWs over the nine semantic concepts and at both image halves. In Figure 5-15, the semantic concept *grass* is usually appear in the lower halves of images but could also appear at the upper halve of the image, such as images of *plains* and *mountains* scenes. This is shown in the *grass* plot, where many of the local keypoints are indexed to the visual words (600-799) and (800-1000). These visual words represent clusters of *plains* and *mountains* scene categories.

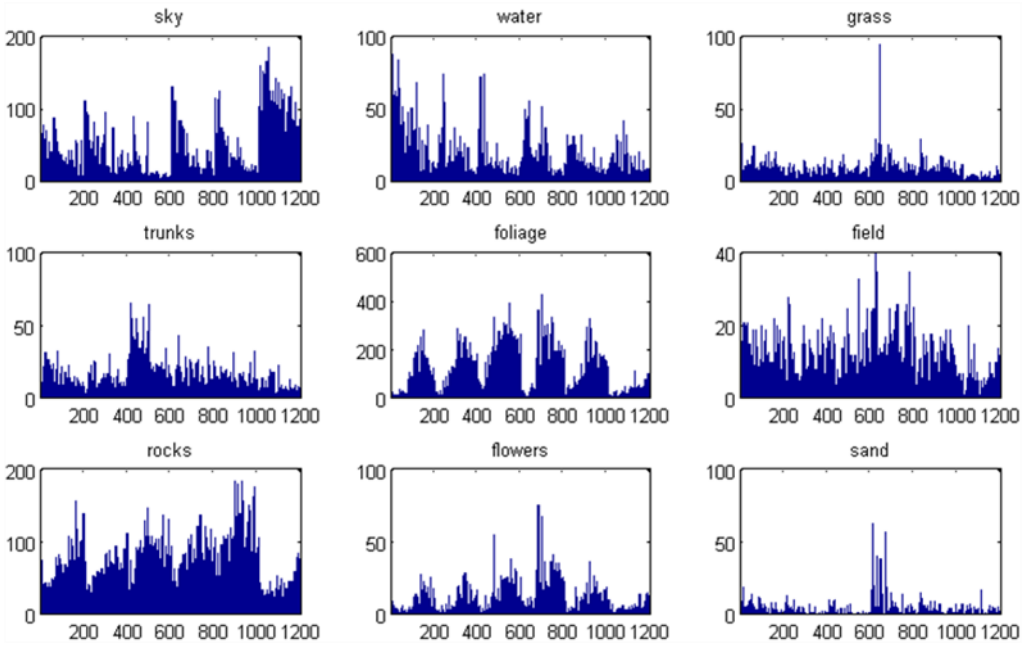


Figure 5-15: Sum of the CBOW histograms obtained using *integrated* visual vocabulary at the *Upper* half of the images.

For the *water* semantic concept, many of the keypoints are assigned to the visual words of the natural scene *coasts*. But, for the same semantic concept, *water* appears more in the lower halves of *coasts* and *river/lakes* scenes categories, as shown in Figure 5-16. It is interesting to use Figure 5-15 and Figure 5-16 to analyze the differences between the distributions of the CBOWs at upper and lower halves. For example, local keypoints located in image regions of the *sand* semantic concept appear in the lower halve more than the upper halve which gives an indication of the importance of building multi-level visual vocabularies from image halves.

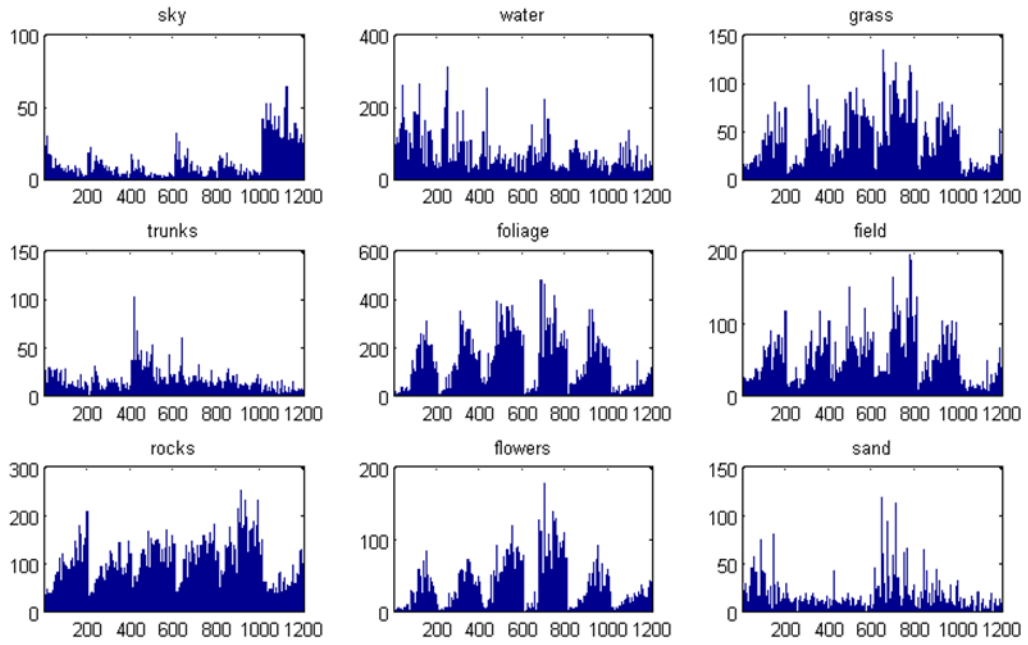


Figure 5-16: Sum of the CBOW histograms obtained using *integrated* visual vocabulary at the *Lower* half of the images.

5.3.2.3 Multiple features

Motivated by the importance of using colour information and textural featuers to describe the visual contents of natural scene images, this section presents a number of features that are used to improve the performance of CBOWs image region representation and therefore will improve the performance of natural scene annotation task. Beside the CBOWs, three types of features are chosen in this work. The first two features are devoted to represent colour whereas the third feature is devoted to represent texture. Colour histogram and colour moments are used to extract colour information from image regions. Discrete Wavelet Transform (DWT) is used to extract the textural features from image regions. To represent the visual contents of image regions, colour moments, colour histogram and DWT are used.

Different combinations of these features are used without any weighting approach. Features are directly propagated to form a single feature vector. Next, the principle of DWT is briefly illustrated.

Discrete Wavelet Transform (DWT)

Discrete wavelet transform is a multi-resolution approach for texture analysis. It has been widely used in image processing and computer vision applications including CBIR, texture image classification, compression, image analysis, etc (Wang et al., 2001, Kokare et al., 2007, Serrano et al., 2004).

Wavelet transforms detect details from an image at horizontal and vertical directions and at different scales. In DWT, an image is decomposed into four sub-bands (1) LL (2) LH (3) HL (4) HH. The first sub-band LL is called approximation coefficients. It represents the horizontal and vertical low frequency components of the image. The sub-band LH is called the vertical coefficients. It represents the horizontal low and vertical high frequency components. The sub-band HL is called horizontal coefficients. It represents the horizontal high and vertical low frequency components. The sub-band HH is called the diagonal coefficients. It represents the horizontal and vertical high frequency components. The sub-band LL can be further decomposed into another four sub-bands and this can be repeated according to the number of levels wanted (*see* Figure 5-17).

The common approach to represent these details is to extract energy measures of the wavelet coefficients from each sub-band as texture features. Energy measures are chosen to represent the texture features because the energy distribution

in the frequency domain recognizes a texture (Kokare et al., 2007). The energy of a wavelet sub-band is computed as follows:

$$Energy = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left| X_{ij} \right| \tag{5-1}$$

where $M \times N$ is the size of the wavelet sub-band, X is the wavelet coefficient.

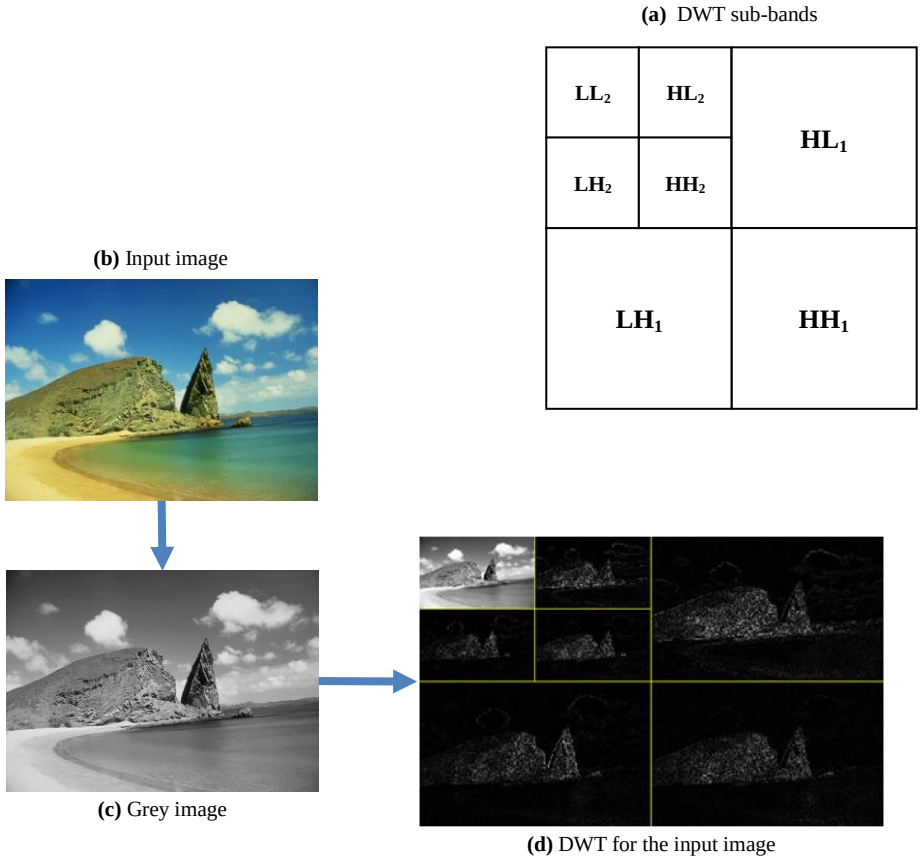


Figure 5-17**:** DWT decomposition using two levels. (a) DWT sub-bands. (b-d) shows an example using DWT decomposition.

In this chapter, the wavelet decomposition is performed at image region level using 2 levels of decomposition using two dimensional DWT functions available in MATLAB[2]. Since images are converted into HSV colour space, each image thus has three components: H, S and V. So, DWT is applied to each component of an image region separately. A feature vector of length 18 (3components × 6 energy measure) is therefore constructed from the three sub-bands (LH, HL, HH) at all resolutions and for each image region. The following code has been used to extract texture features using DWT from the Hue component (H) of the image. The same code is used to extract DWT features from the other two components (S and V):

```
[c,s] = wavedec2(H,2,'haar'); % Wavelet decomposition
[chd2,cvd2,cdd2] = detcoef2('all',c,s,2); % Get details Coeficient
at level 2
[chd1,cvd1,cdd1] = detcoef2('all',c,s,1); % Get details Coeficient
at level 1
D11=Energy2(chd1);D12=Energy2(cvd1);D13=Energy2(cdd1);
D21=Energy2(chd2);D22=Energy2(cvd2);D23=Energy2(cdd2);
```

5.3.2.4 Prototypical local semantic concept representation

The idea of prototypical semantic concept representation is to associate local semantic concepts with instances that are considered to be prototypical with regards to their visual information (Vogel and Schiele, 2007). Based on this idea, prototypical local semantic concepts are learned from visual information extracted from image regions. For the natural scene dataset used in this chapter, there are nine semantic concepts. For each semantic concept, visual features extracted from image regions labelled with this semantic concept are averaged to obtain a local semantic prototype that represents instances of this semantic concept, hence this prototype does not necessarily to be member of the corresponding semantic concept. It is only

[2] http://www.mathworks.co.uk/help/toolbox/wavelet/ref/dwt2.html

a summary representation of each semantic concept. This result in generating nine local semantic prototypes learned from instances (visual features) of the nine semantic concepts. These prototypes will be used in the next section to annotate image regions using KNN approach.

5.4 Experimental Work

The first part of this section presents local semantic concept annotators, to annotate image regions, using KNN and SVM classifiers. It also presents the protocol that is followed to conduct the experiments. Experimental results are then reported with some discussion. The performance of local image annotators is assessed using the average precision produced from the confusion matrix.

5.4.1 Local Semantic Annotators

Image annotation at region level can be considered as a supervised classification problem. Each image region needs to be classified into one of predefined classes. In this chapter, two classifiers are employed for the annotation task: the nearest neighbour approach (KNN) and support vector machines (SVM). The protocol used for all experimental work is as follows: Using the KNN classifier, all the experiments have been validated using 10-folds cross validation where 90% of all image regions are selected randomly for generating the local semantic prototypes and the remaining 10% are used for testing. Using SVM classifier, the same approach has been used where the 90% of all image regions used to generate the semantic concept prototypes are also used to train the SVM classifier whereas the 10% of image regions used in KNN are used for testing. The procedure is repeated 10 times for both classifiers such that an image region appears once in the testing

part over the ten folds. The publicly available LIBSVM tool is used to implement the SVM classifier with all parameters chosen based on 10-fold cross validation performed in each training set, also the Histogram intersection kernel is used to train the SVM as explained in Section 4.3.1.

The principles of both classifiers were introduced in Sections 3.5.1 and 3.5.2. For the KNN approach, it is a pre-requisite to decide the similarity metric and the number of neighbours (K) that need to be visited to decide to which class an input instance belongs to, using a voting technique. Using the nine local semantic prototypes presented in Section 5.3.2.4, the KNN classifier needs to assign an input feature vector, which represents the visual content of input image region, to one of the local semantic concepts, thus K=1. For simplicity, in this section the bag of visual words BOW and IBOW histograms are used to refer to the concept-based bag of visual words CBOW generated using the universal visual vocabulary and the integrated visual vocabulary, respectively. In other words, instead of using CBOW and CIBOW, the author simply refers to them by BOW and IBOW. This section explicitly refers to the BOWs and IBOWs generated from image regions at the upper and lower halves in the experimental results. Algorithm 5.5 presents the work flow of annotating image regions with local semantic concepts using KNN and local semantic prototypes. The same algorithm is used to conduct different experiments using the local semantic prototypes approach.

Algorithm 5.5: A step-by-step algorithm for generating semantic concept prototypes and how they are used to annotate image regions with local semantic concepts using KNN approach.

Input: (1) A collection of BOW histograms representing all image regions in the image dataset.

(2) A confusion matrix **mat** of size $N \times N$, where N is the number of semantic concepts. Initialize **mat** to zeros.

For i=1 **to** 10 **Do**:

a. Randomly select 90% of all BOWs histograms for training and 10% for testing.
b. For each semantic concept $\bar{c}_j$, j=1,2,…, N, generate local semantic concept prototype P_j by averaging all BOWs from the training set.
c. Use KNN classifier, using the Euclidean distance and K=1, to find similarities between the BOWs in the test set and the local semantic prototypes P. An image region, represented by a BOW histogram, is assigned to the semantic concept j such that the semantic prototype P_j is the most similar prototype to the BOW of this image region.
d. Compare the labels assigned to image regions from the test set with the ground truth labels. Report the results in the confusion matrix **mat.**

end

Output: Confusion matrix **mat**.

Algorithm 5.6 presents the work flow of annotating image regions with local semantic concepts using KNN and local semantic prototypes generated from the upper and lower halves of images.

Algorithm 5.6: A step-by-step algorithm for generating semantic concept prototypes at the upper and lower halves of images and how they are used to annotate image regions with local semantic concepts using KNN approach.

Input: (1) A collection of BOW histograms representing all image regions at the *upper* halve of images in the dataset.

(2) A collection of BOW histograms representing all image regions at the *lower* halve of images in the dataset.

(3) Two confusion matrices **mat_upper** and **mat_lower**, both of size $N \times N$, where N is the number of semantic concepts. Initialize **mat_upper** and **mat_lower** to zeros.

For i=1 **to** 10 **Do**:

a. Randomly select 90% of all BOWs histograms, generated from the upper halves, for training and 10% for testing.
b. For each semantic concept $\bar{c}_j$, j=1,2,…, N, generate local semantic concept prototype P_j_upper by averaging all BOWs from the training set.
c. Use KNN classifier, using the Euclidean distance and K=1, to find similarities between the BOWs in the test set and the local semantic prototypes *P_upper*. An image region at the upper halve, represented by a BOW histogram, is assigned to the semantic concept j such that the semantic prototype P_j_upper is the most similar prototype to the BOW of this image region.
d. Compare the labels assigned to image regions from the test set with the ground truth labels. Report the results in the confusion matrix **mat_upper.**

end

For i=1 **to** 10 **Do**:

a. Randomly select 90% of all BOWs histograms, generated from the lower halves, for training and 10% for testing.
b. For each semantic concept $\bar{c}_j$, j=1,2,…, N, generate local semantic concept prototype P_j_lower by averaging all BOWs from the training set.
c. Use KNN classifier, using the Euclidean distance and K=1, to find similarities between the BOWs in the test set and the local semantic prototypes *P*_lower. An image region at the lower halve, represented by a BOW histogram, is assigned to the semantic concept j such that the semantic prototype P_j_lower is the most similar prototype to the BOW of this image region.
d. Compare the labels assigned to image regions from the test set with the ground truth labels. Report the results in the confusion matrix **mat_lower.**

end

Output: Confusion matrix **mat**, where **mat** = **(mat_upper+mat_lower)**

5.4.2 Experimental Results

In this chapter, four sets of experiments are conducted. The *first two sets* consider using universal and integrated visual vocabularies, obtained in Chapter 4 and discussed in Section 5.3, to build BOW and IBOW histograms from image regions. As mentioned at the beginning of the previous section, BOW and IBOW refer to concept-based BOWs generated using both visual vocabularies. The *last two sets* of experiments investigate generating visual vocabularies from image halves to build BOW and IBOW for image regions at each halve separately (*see* Section 5.3.2.1). For fair comparisons, different types of features are included in the experiments and their annotation performances are reported. Three types of features are included: colour histogram, colour moments and DWT. These features are used separately to represent the visual content of image regions. Also, these features are linearly integrated with the BOW and IBOW histograms to study their influence on the discriminative power of BOWs by including colour and textural features.

The first set of experiments investigates image region annotation using local semantic prototypes and the KNN classifier. Algorithm 5.5 is used to generate local semantic prototypes from BOW histograms and to annotate new image regions with semantic concepts using the KNN classifier. The same algorithm can be applied to generate local prototypes for different features. The only thing that needs to be changed in the algorithm is to replace BOW with the features extracted from image regions. For example, replacing BOW with IBOW histograms will generate local semantic prototypes from IBOWs and then the KNN classifier is used to annotate new image region, represented by an IBOW histogram, with one of the predefined semantic labels. Another example is to use colour histogram to represent the visual

content of image regions. In this work and similar to (Vogel and Schiele, 2004), three histograms are obtained from an image region represented in the HSV colour space. The first histogram (36-bins) is obtained from the Hue component of the image region, the second histogram (32-bins) is obtained from the S component while the third histogram (16-bins) is obtained from the V component. These three histograms are concatenated to form a single HSV colour histogram of 84-bins. Algorithm 5.5 is used again but BOW histograms are now replaced with the HSV colour histograms from which local semantic prototypes are generated and used by the KNN classifier to annotate new image regions.

For multiple features, local semantic prototypes are first generated for each feature separately and then for each semantic concept the prototypes of the different feature types are aggregated to form a single local semantic prototype for the corresponding local semantic concepts. For example, the natural scene dataset used in this chapter is composed of nine semantic concepts. Suppose that colour histogram (84-D feature vector) and BOW (200-D feature vector) are the features to be concatenated. Firstly, Algorithm 5.5 is used to generate nine local semantic prototypes for the colour histogram and the same algorithm is used to generate another nine local semantic prototypes for the BOW histograms. For each semantic concept, the local semantic prototype produced from the colour histograms is concatenated with the local semantic prototype produced from the BOW histograms. This will generate nine local semantic prototypes each of length (84+200 = 284-D). The same approach can be applied for more than two feature types.

To this end, in the first experimental set, 14 experiments are conducted using the following features and their combinations for image region annotation: *IBOW*,

UBOW[3], *Colour histogram* (*ColHist*), *colour moments* (*Mom*), *DWT* (*Wav*), *IBOW+Mom*, *UBOW+Mom*, *IBOW+ColHist*, *UBOW+ColHist*, *IBOW+Wav*, *UBOW+Wav*, *IBOW+ColHist+Wav*, *UBOW+ColHist+Wav*, and *ColHist+Wav*.

In the second set of experiments, SVM is used to annotate image regions. Here, local semantic prototypes are not used. For IBOW histograms, SVMs are trained on 90% of all image regions in the dataset, represented by their IBOW histograms. The trained SVMs are then used to assign labels to the remaining image regions in the dataset. For multiple features, each feature vector is first normalized to a unit length vector and then they are aggregated into a single feature vector. The SVMs are then used for training and testing. Similar to the first set of experiments, 14 experiments are conducted using the same features and their combination as aforementioned.

Figure 5-18 presents the performance of the first two sets of experiments. Bars in blue show the performance of image region annotation using local semantic prototypes with the KNN classifier. Bars in red show the performance of image region annotations using SVMS. It is worth to remind that IBOW and UBOW histograms are generated from visual vocabularies constructed at scene level and not concept level, the local from global approach. Thus, it is interesting to see their performances for image region annotation. It is obvious that SVMs outperforms the local semantic prototypes in all types of features and their combinations. Also, IBOW histograms outperform UBOW in both sets of experiments. Adding colour and textural features to the UBOW and IBOW histograms improved the accuracy of image region annotation. The performance of IBOW histograms has gained the best

[3] UBOW refers to the BOW generated using universal visual vocabulary.

annotation accuracy when colour histograms and DWT are added to them. Nevertheless, integrating low-level features such as colour histogram with the DWT has also gained a good annotation performance compared with more complicated features such as IBOW and UBOW histograms. However, they have not got the best annotation results but they can be used to improve the performances of the IBOW and UBOW histograms. The annotation performance for each semantic concept resulted from the 14 experiments using both SVM and local semantic prototypes are shown in Table 5-2 and

Table 5-3. Each row contains the accuracies of each semantic concept. These values are the diagonal elements of the confusion matrix generated from each experiment. The last columns from both table is used to generate the results shown in Figure 5-18.

Table 5-2: (*KNN and semantic prototypes*) Accuracies of each experiment (row), where elements in each row are the diagonal elements of the confusion matrix resulted from each experiment. Accuracy is generated based on 10-folds CV.

	Sky	**Water**	**Grass**	**Trunks**	**Foliage**	**Field**	**Rocks**	**Flowers**	**Sand**	**Acc.**
IBOW	88.54	34.50	9.93	37.35	45.56	25.93	32.56	44.22	26.42	48.41
UBOW	86.61	29.52	6.45	24.12	39.85	25.88	24.15	42.61	26.88	43.87
ColHist	58.91	49.69	61.00	24.25	26.79	43.46	22.10	49.15	48.19	41.19
Mom	36.24	29.12	3.94	35.75	15.22	32.95	10.68	42.80	43.21	24.20
Wav	84.68	29.51	22.44	38.65	47.06	16.05	16.45	48.22	29.34	44.47
IBOW+Mom	44.61	36.38	9.79	44.06	29.76	37.01	16.60	49.00	49.17	32.78
UBOW+Mom	48.76	35.69	13.90	42.89	35.43	35.72	18.29	49.19	50.26	35.50
IBOW+ColHist	65.29	54.93	63.49	36.62	36.47	46.08	29.55	54.32	51.75	47.92
UBOW+ColHist	66.63	54.48	59.12	38.40	40.53	46.20	29.31	55.30	51.69	48.92
IBOW+Wav	88.81	38.26	13.05	43.51	49.39	26.12	36.37	47.58	30.83	51.11
UBOW+Wav	86.47	33.85	8.02	30.34	43.23	26.48	26.50	44.75	30.83	46.05
IBOW+ColHist+Wav	66.32	55.44	64.37	37.48	37.59	46.37	30.61	54.86	52.84	48.84
UBOW+ColHist+Wav	67.62	55.00	60.38	39.08	41.59	46.63	29.98	55.78	52.95	49.77
ColHist+Wav	60.09	50.19	61.95	25.66	28.10	44.46	23.74	50.61	49.63	42.39

Table 5-3: (*SVM*) Accuracies of each experiment (row), where elements in each row are the diagonal elements of the confusion matrix resulted from each experiment. Accuracy is generated based on 10-folds CV.

	Sky	Water	Grass	Trunks	Foliage	Field	Rocks	Flowers	Sand	Acc.
IBOW	92.10	45.00	43.53	32.06	77.76	41.55	71.58	42.31	35.59	68.12
UBOW	91.59	45.30	24.12	37.35	76.98	13.04	67.40	24.40	3.50	61.65
ColHist	66.85	41.16	45.45	9.42	55.55	30.68	70.53	43.92	45.90	55.26
Mom	90.47	26.44	18.84	0.00	77.17	18.74	59.70	18.89	10.20	57.82
Wav	90.76	63.77	4.65	24.06	81.90	12.23	54.55	17.03	6.53	61.70
IBOW+Mom	94.70	47.44	42.62	37.78	81.22	45.37	72.86	27.38	42.29	70.20
UBOW+Mom	91.59	45.30	24.12	37.35	76.98	13.04	67.40	24.40	3.50	65.54
IBOW+ColHist	95.56	64.45	55.27	21.42	85.56	55.75	79.50	64.52	57.08	77.37
UBOW+ColHist	91.89	63.99	51.81	18.65	80.68	52.27	78.77	67.69	48.88	74.51
IBOW+Wav	93.89	64.80	38.28	32.86	80.64	40.19	71.56	50.37	20.57	71.12
UBOW+Wav	94.12	64.66	15.13	32.80	80.53	30.56	61.65	39.73	26.36	67.10
IBOW+ColHist+Wav	97.74	77.14	59.32	41.78	86.92	61.37	81.54	69.64	63.15	81.64
UBOW+ColHist+Wav	94.36	72.45	52.87	33.91	72.52	59.57	79.73	73.01	66.19	76.13
ColHist+Wav	95.93	68.57	55.75	38.77	75.00	67.93	82.14	71.79	56.05	77.61

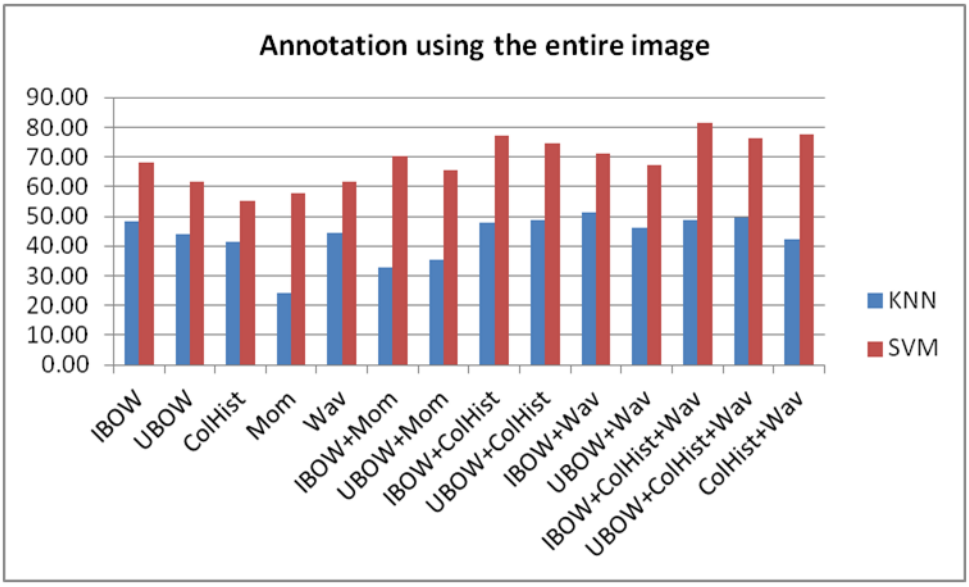


Figure 5-18: Accuracies of annotating images with the nine semantic concepts using KNN and SVM classifiers. BOWs and IBOWs are generated from image regions using visual vocabularies constructed in Chapter 4.

The last two sets of experiments are dedicated to study the influence of building visual vocabularies from local keypoints located at the upper and lower halves of images. It aims to improve the quality of visual words generated from clustering features of local keypoints at upper and lower halves of images. Two visual vocabularies are generated from the upper halves of images; universal visual vocabulary and integrated visual vocabulary. Image regions located at the upper halves of images are represented by UBOW and IBOW histograms generated using the upper universal and integrated visual vocabularies, respectively. Another two visual vocabularies are generated from the lower halves of images. Image regions located at the lower halves of images are represented by UBOW and IBOW histograms generated using the lower universal and integrated visual vocabularies. To use multiple features with the UBOW and IBOW, the same low-level features used in the first two sets are employed here.

In the third set of experiments Algorithm 5.6 is used to generate local semantic prototypes for the upper halve and local semantic prototypes for the lower halve. Also, 14 experiments are conducted using this algorithm. Local semantic prototypes at the upper halve are used by KNN to annotate test image regions located at the upper halve of images. Results of annotations are compared with the ground truth and then reported in **mat_upper** confusion matrix. The same procedure is applied to the lower halve with results reported in another confusion matrix, **mat_lower**. Both matrices are added together to get a single confusion matrix. Results of annotating images at the upper and lower halves, using upper and lower local semantic prototypes and the KNN classifier, is shown in blue bars in Figure 5-19. It is obvious from the figure that generating local semantic prototypes from

image halves improved the annotation accuracies in all experiments and for most features. The IBOW histograms are still the best in representing image regions. Adding textural features to IBOW has slightly improved the annotation and works better than adding colour information. More details about the annotation accuracies of each of the experiments are shown in Table 5-4.

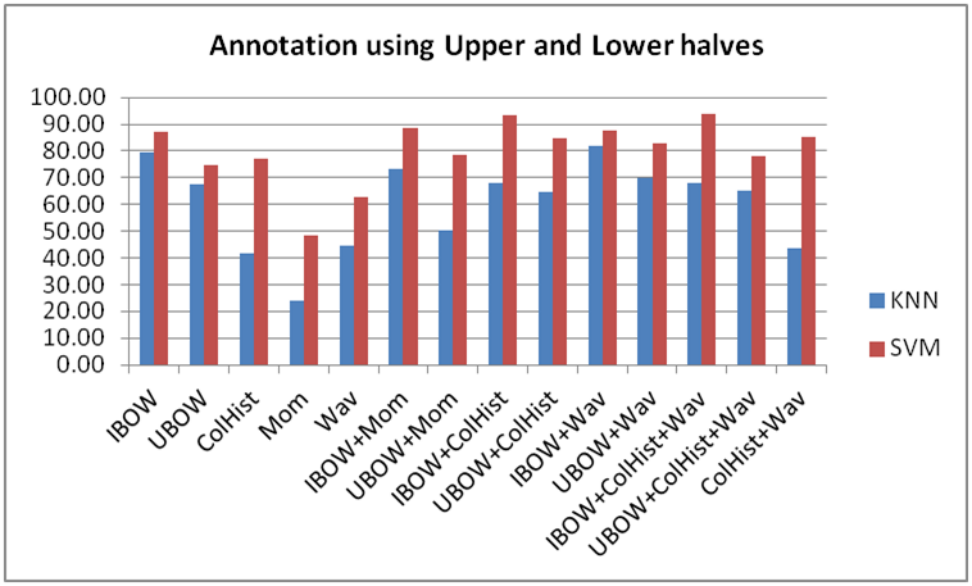


Figure 5-19: Accuracies of annotating images with the nine semantic concepts using KNN and SVM classifiers. BOWs and IBOWs are generated from image regions using visual vocabularies generated from the upper and lower halves of images.

.

The last set of experiments (14 experiments) is conducted on different features using SVM classifier. SVMs are trained and tested on image regions at the upper and lower halves of images. At the upper halve, image regions represented by the IBOW histograms or other features are used to train SVM which in turn used later to annotate test image regions at the upper halve of images. Results are reported

in a confusion matrix mat_upper. The same approach is applied for image region at the lower halve of images. Results are reported in another confusion matrix mat_lower. Both matrices are add to each other to get final confusion matrix. The red bars in Figure 5-19 show annotations accuracies using different features. The best result is achieved using IBOW combined with ColHist and Wav features. It reports 94.02% annotation accuracy. More details are shown in Table 5-5.

Table 5-4: (*KNN and semantic prototypes at upper and lower halves*) Accuracies of each experiment (row), where elements in each row are the diagonal elements of the confusion matrix (upper+lower) resulted from each experiment. Accuracy is generated based on 10-folds CV at each halves.

	Sky	**Water**	**Grass**	**Trunks**	**Foliage**	**Field**	**Rocks**	**Flowers**	**Sand**	**Acc.**
IBOW	99.92	51.80	84.04	77.72	69.98	56.97	89.06	93.22	67.74	79.73
UBOW	98.09	29.77	68.28	55.88	55.39	36.75	81.41	67.40	47.45	67.56
ColHist	57.30	53.13	60.29	29.54	29.11	45.61	22.24	48.71	43.78	41.85
Mom	34.20	26.22	5.48	38.58	15.54	31.06	13.60	46.17	44.47	24.13
Wav	83.25	28.34	22.24	41.35	48.34	16.79	16.98	50.17	28.14	44.50
IBOW+Mom	60.13	60.77	87.70	80.12	74.63	61.77	91.60	93.51	75.99	73.16
UBOW+Mom	44.23	40.61	48.07	64.55	48.21	50.24	61.85	66.18	61.26	50.39
IBOW+ColHist	64.91	68.83	85.95	68.55	61.65	67.88	72.07	76.87	72.44	68.11
UBOW+ColHist	64.08	60.40	80.53	62.58	59.49	61.91	70.21	71.40	64.01	64.75
IBOW+Wav	86.71	63.57	85.90	83.75	82.23	62.63	90.19	93.85	75.42	81.70
UBOW+Wav	85.00	40.44	69.57	62.65	69.39	42.74	81.96	69.89	56.91	69.84
IBOW+ColHist+Wav	62.87	70.22	85.98	72.98	63.29	71.06	68.69	76.77	70.83	67.80
UBOW+ColHist+Wav	62.07	63.83	81.67	66.15	60.94	65.47	68.09	72.86	64.18	65.05
ColHist+Wav	55.08	56.56	61.55	37.66	34.13	51.03	22.88	46.80	50.72	43.75

Table 5-5: (*SVM at upper and lower halves*) Accuracies of each experiment (row), where elements in each row are the diagonal elements of the confusion matrix (upper+lower) resulted from each experiment. Accuracy is generated based on 10-folds CV at each halves.

	Sky	**Water**	**Grass**	**Trunks**	**Foliage**	**Field**	**Rocks**	**Flowers**	**Sand**	**Acc.**
IBOW	90.44	82.96	99.46	89.42	76.53	68.55	99.55	99.90	82.35	87.17
UBOW	86.80	33.46	44.48	80.12	75.07	67.26	95.49	83.94	74.15	74.94
ColHist	94.36	65.94	58.15	10.77	84.85	59.15	75.05	71.45	66.65	76.88
Mom	85.28	0.40	1.68	11.63	47.38	4.92	77.30	20.35	24.30	48.45
Wav	91.72	66.02	9.31	24.25	81.88	16.95	52.35	18.01	12.84	62.61
IBOW+Mom	85.39	82.83	88.07	75.45	92.57	90.76	93.36	81.16	90.26	88.44
UBOW+Mom	86.17	41.82	50.56	57.29	82.23	67.86	99.52	99.66	80.80	78.65
IBOW+ColHist	95.86	85.58	99.49	87.75	96.21	89.57	94.24	85.31	95.99	93.61
UBOW+ColHist	86.80	86.25	93.01	73.97	76.53	72.04	94.61	83.94	79.89	84.57
IBOW+Wav	91.09	84.33	99.46	89.42	76.53	68.55	99.55	99.90	88.08	87.66
UBOW+Wav	86.15	85.56	90.15	73.97	75.07	67.26	94.16	79.06	74.15	83.09
IBOW+ColHist+Wav	96.51	85.58	99.49	87.75	96.21	89.57	94.24	92.63	95.99	94.02
UBOW+ColHist+Wav	88.08	41.23	59.12	58.09	75.84	63.49	99.72	99.66	86.82	78.04
ColHist+Wav	98.08	84.82	67.88	50.46	87.73	73.35	81.97	77.89	74.50	85.07

5.5 Summary

This chapter has presented the problem of image annotation at image region level. The task was to assign labels to image regions generated from a regular grid. The chapter has addressed using BOW model to represent image regions. A framework for image region annotation has been proposed. The relationship between the distributions of local semantic concepts and local keypoints located in image regions labelled with these semantic concepts are studied in detail. Also, this chapter has investigated using visual vocabularies generated from natural scene classes to represent local semantic concepts, the local from global approach. Generating visual vocabularies from image halves were also investigated to generate BOW histograms. An extensive experimental work has been conducted using different features and classifiers to annotate image regions with semantic concepts. Our experimental

results shows the plausibility of local from global approach for image region annotation as well as the discriminative power of using visual vocabularies from image halves. It showed an improved annotation results using IBOW combined with low-level features.

Chapter 6

Image Retrieval

This chapter addresses the problem of semantic-based image retrieval of natural scenes. A typical content-based image retrieval system deals with the query image and images in the dataset as a collection of low-level features and retrieves a ranked list of images based on the similarities between features of the query image and features of images in the image dataset. However, top ranked images in the retrieved list, which have high similarities to the query image, may be different from the query image in terms of the semantic interpretation of the user (Chen et al., 2005). This has been referred to as the semantic gap (Smeulders et al., 2000). In Section 3.1.3, different approaches have been introduced that deal with the semantic gap, such as image classification, annotation, ontology, etc. Moreover, efficient representation of the visual content of images plays an important role in reducing the semantic gap. Based on the works presented in Chapters 4 and 5, the semantic- based image representation can be obtained using two different scenarios, as follows:

The *first* scenario considers an image as a collection of local semantic concepts. These semantic concepts are semantic labels, such as *sky* and *grass*, used to describe the visual content of image regions. In Chapter 5, a natural scene dataset was used for region-based image annotation task. In this dataset, images were divided into 10×10 regular grid and each region was manually labeled with one of predefined semantic labels. Representing the image content as a collection of local semantic concepts would make the image representation more semantically meaningful. However, every image would have different number of local semantic concepts, i.e. not all semantic concepts appear in every image. Similar to the work of Vogel and Schiele (Vogel and Schiele, 2007), the local semantic concepts appear in an image can be summarized as a histogram of their occurrence in the image. This histogram is called the concept-occurrence vector (COV).

The *second* scenario considers using different configurations of the bag of visual words model, presented in Chapter 4, to represent the invariance characteristics of local image regions detected and described using the DoG and SIFT features. The bag of visual words model can be considered as an intermediate semantic representation of the image content. Although no semantic labels are contained in the bag of visual word histograms, the visual words generated from the clustering process can be regarded as visual words of semantic information. This can be justified in the sense that similar local keypoints may be allocated to the same cluster (visual word), particularly in the case of using integrated visual vocabularies.

To this end, this chapter investigates how natural scene retrieval can be performed using the bag of visual word model and the distribution of local semantic concepts. The aim of this chapter is to study the efficiency of using the two

aforementioned scenarios for representing the semantic information, depicted in natural scene images, for image retrieval. To achieve this aim, this chapter presents an extensive comparative study between both scenarios as well as other baseline methods, such as colour and texture which represent the visual content of images without any semantic information.

The current chapter proceeds as follows. The *first* section (Section 6.1) introduces the evaluation methodology employed to evaluate the performance of different image retrieval approaches presented in this chapter. The *second* section (Section 6.2) introduces the concept-occurrence vector (COV) approach adopted from the work of Vogel and Schiele (Vogel and Schiele, 2007) to summarize the amount of different local semantic concepts, depicted in an image, into a global image representation. The *third* section (Section 6.3) introduces the different approaches, presented in Chapter 4, to be used for the task of natural scene retrieval. Finally, the *fourth* section (Section 6.4) presents an evaluation of the aforementioned approaches by carrying out an extensive experimental work to study the efficiency of using COVs as well as the bag of visual words model for natural scene retrieval.

6.1 Evaluation Methodology

There are different measures for retrieval performance proposed in the information retrieval and pattern recognition literature to evaluate the results of the experiments. The most common and widely used performance measures in information retrieval are the *precision P* and *recall R* (*see* Section 3.6.3). These measures evaluate how well an information retrieval system performs on the ground truth data. It gives a good indication of the image retrieval system performance (Muller et al., 2001).

For a given query *Q*, let *X* be the number of relevant images that belongs to the same category in the image dataset and *Y* be the set of all images retrieved. Assume that *Z* be the number of retrieved images coming from the same category (i.e., correctly retrieved images) which are among the *Y* retrieved images, then precision *P* and recall *R* care defined as:

$$P = \frac{Z}{Y} \tag{6-1}$$

$$R = \frac{Z}{X} \tag{6-2}$$

Usually, precision decreases as the number of images retrieved increase, whereas recall increases as the number of retrieved images increases; recall is a non-decreasing function of the number of images retrieved. These two values are commonly combined into a so called *recall-precision* graph where precision values (y-axis) are plotted against recall values (x-axis) and each dot in the graph represents a retrieved image of the ordered result list. It shows how many retrieved images retrieved are relevant or irrelevant among the top ranked images. However, interpreting recall-precision graphs is not an easy task. Thus, it is possible to summarize precision and recall in a single value by calculating the average precision at each point when a relevant image is found and then calculate the mathematical mean of these precisions. This measure is called the Mean Average Precision (*MAP*). For every query image, precision and recall measures are computed over all images retrieved which are ranked, in a descending order, according to a similarity measure. Then the measures are averaged over all the queries in the test dataset.

In (Baeza-Yates and Ribeiro-Neto, 1999), the average precision is measured as the arithmetic mean of all precisions at the 11 recall cut-off values 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. Table 6-1 shows an example of how to calculate the average precision for a particular query image. The mean value of the average precision over all query images is called MAP and is defined as follows:

$$MAP(Q) = \frac{1}{Q}\sum_{j=1}^{|Q|}\frac{1}{11}\sum_{k=1}^{11}P_{jk} \tag{6-3}$$

where Q is the set of query images (test images) from a particular scene category and P_{jk} is the precision at recall level k.

Table 6-1: An example of recall vs. precision: for a query image, the first column shows the 11 recall cut-off values whereas the second column shows the precision of retrieved images at every recall value.

Recall	Precision
0.00	0.78
0.10	0.67
0.20	0.53
0.30	0.48
0.40	0.37
0.50	0.32
0.60	0.24
0.70	0.18
0.80	0.13
0.90	0.08
1.00	0.04
Average precision	0.3472

6.2 Image Retrieval Based on Annotated Image Regions

This section is related to semantic-based image representation using the first scenario mentioned at the beginning of this chapter. Based on this scenario, a query image can be considered as a collection of local semantic concepts, each of which

describes a particular region in the image. In this section, image regions in the database are assumed to be annotated with semantic concepts. For this reason, the natural scene dataset presented in Section 5.1 which provides ground truth data about annotations of image regions is employed in this section. Images in the dataset are divided into 10×10 regular grid which yields 100 regions per image. All these regions were manually annotated with semantic concepts. These semantic concepts are *sky*, *water*, *grass*, *trunks*, *foliage*, *field*, *rocks*, *flowers* and *sand*. The bag of visual words was used to represent the visual content of these regions (*see* Section 5.3.2.1). To annotate the regions of test images, SVM and KNN classifiers were trained on different configurations of the bag of visual words model. The experimental work demonstrated that the SVM classifier achieved the best results to annotate image regions at image halves.

To this end and having that test images are annotated with local semantic concepts, an image is described as a collection of local semantic concepts. These semantic concepts provide semantic interpretation of the image content which can reduce the semantic gap between the user perception and the image content. However, different images would have different number of local semantic concepts. Thus, it is important to summarize the amount of local semantic concepts found in an image into a global feature vector. To do so, the concept-occurrence vector (COV) proposed by Vogel and Schiele (Vogel and Schiele, 2007) is adopted. To represent an image, the frequency of occurrence of each semantic concept is determined. Since there are nine semantic concepts available in the natural scene dataset, each image can be represented as a feature vector of size nine. Each component corresponds to a

semantic concept and it contains an integer number of the frequency of occurrence of this semantic concept in the image.

Figure 6-1 shows image representation using the COV adopted from (Vogel and Schiele, 2007). This figure shows an image divided into 10×10 regular grid and each region is manually annotated with one of the nine semantic concepts. The COV is then generated by counting how many times a particular semantic concept appears in the image. For example, the semantic concept *sky* appears 47 times and a half. By dividing this number by 100 yields a normalized occurrence value of 47.5%. This is replicated for all semantic concepts. All natural scenes in the database are indexed using their COV generated from the ground truth annotations.

To retrieve images from the database, the COV is first constructed from the query image. This feature vector is then compared to all COVs of images available in the database using a similarity measure, such as the Euclidean distance, and a ranked list of the relevant images are retrieved as a response to the query image. The performance of using COVs generated from ground truth annotations for natural scene retrieval will be presented in the experimental work section. Their retrieval results will be considered as benchmark which gives the best retrieval results to expect.

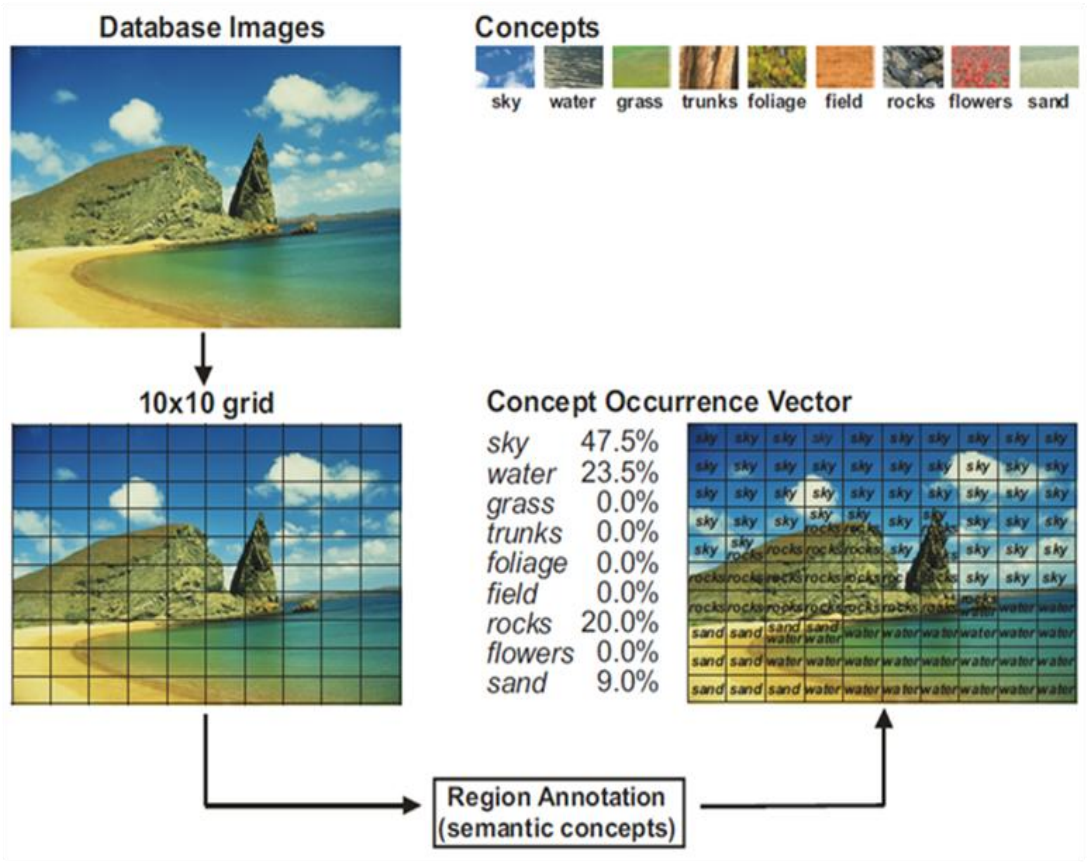


Figure 6-1: Image representation using concept-occurrence vector (COV) (Vogel and Schiele, 2007)

To annotate image regions with local semantic concepts, different approaches were proposed in Section 5.3.2.1 to represent image regions content. Different configurations of the bag of visual words model were employed to represent the visual characteristics of image regions. The task was to annotate test images with local semantic concepts assigned to image regions. In Section 5.4.2, four sets of experiments have been conducted for image region annotation using the SVM and KNN classifiers. The best annotation results are obtained using the last set of experiments. In the last set of experiments, 14 experiments are conducted using the SVM classifier and 14 different image region representation approaches.

Figure 6-2 shows the main steps needed to construct the COV for a new image. As mentioned above, there are 14 different approaches to represent the visual content of image regions (*see* * note in Figure 6-2 for the list of approaches used in this chapter). The COVs resulted from each approach will be compared in the experimental work section to see which of these approaches are suitable to represent image regions for the task of natural scene image retrieval. Their retrieval performances will be compared to the COVs benchmark.

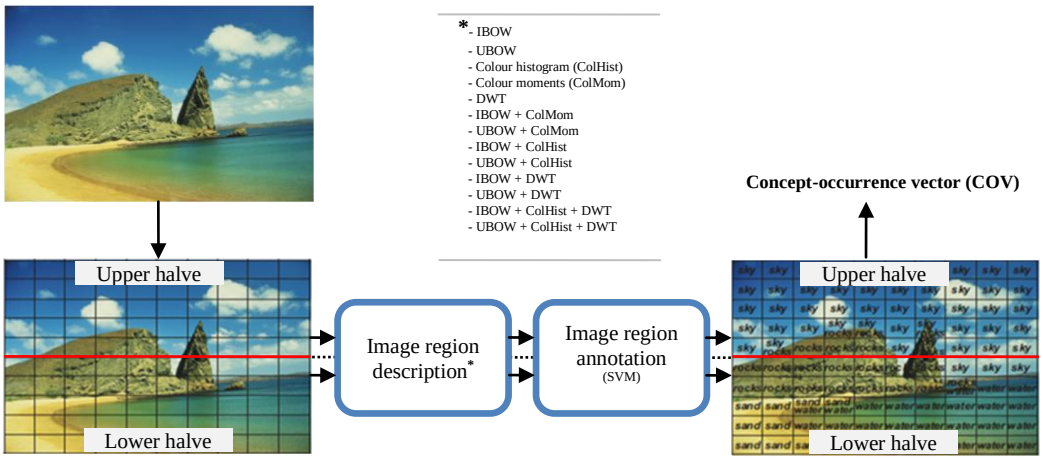


Figure 6-2: Image representation using COV of local semantic concepts assigned to image regions at the upper and lower halves of the image.

6.3 Image Retrieval Using Bag of Visual Words

The problem of using global descriptors, such as colour histograms, to represent the visual content of an image is that they cannot describe different characteristics of the image content. To overcome this problem, descriptors of local keypoints aim to capture characteristics of different parts in an image. Grouping

these descriptors into visual words, using a clustering algorithm, may result in allocating these descriptors to visually similar visual words.

The bag of visual word model has shown to be effective to represent the distribution of local keypoints, detected in the image, for the natural scene classification task. Descriptors of local keypoints mainly represent the intensity information extracted from regions around interest points while discarded colour information (*See* Section 4.2). Also, bag of visual word model ignores the order of local keypoints found in the image. Moreover, clusters (visual words) of the traditional visual vocabulary do not capture the differences between local keypoints found in different scene categories. Thus, in Chapter 4, different approaches were introduced to overcome these limitations and have been applied to the natural scene classification task. Some of these approaches will be used in this chapter to investigate how well they perform for the natural scene retrieval task. Also, this chapter aims to compare the performances of using the COV and the BOW for representing the semantic information of images to perform the retrieval task.

The difference between using the COV and the bag of visual words model for natural scene retrieval is that the former approach requires images to be manually annotated with local semantic concepts in order to train semantic concept annotators which are used to annotate new image, while the later approach does not. However, the dimensionality if COV is too small compared to the high dimensional BOW histograms.

In this section, no annotations are required to represent the semantic information in the image. Images in the database are assumed to be indexed by their BOW histograms and its different configurations. Table 6-2 shows details of the

different approaches used in this chapter to represent the visual content of images. Most of these image representations are obtained from the work presented in Chapter 4 for three natural scene datasets with 6, 8, and 15 scene categories, respectively. The next section introduces the experimental results of natural scene retrieval using all approaches presented in this section and the previous one.

Table 6-2: Image representation using different approaches. The second column describes each approach and specifies the size needed for each representation.

Image representation approach	Description of the approach and the dimensionality of feature vectors produced
ColHist	HSV colour histogram (H: 36, S: 32, V: 16 = 84-D), for grey images (36-D)
PColMom_L0	HSV pyramidal colour moment at level 0 (firs & second moments = 6-D)
DWT	Discrete Wavelet transform (H: 6-D, S: 6-D, V:6-D = 18-D)
ColHist+DWT	Colour histograms integrated with DWT (102-D)
PColMom_L2	HSV pyramidal colour moments at level 2 (126-D)
UBOW	BOW histogram using universal visual vocabulary (200-D)
IBOW	BOW histogram using integrated visual vocabulary (# scene categories ×200)
PUBOW_L1	Pyramidal UBOW at level 1, (200×5 = 1000-D)
PUBOW_L2	Pyramidal UBOW at level 2, (200×21 = 4200-D)
PUBOW_L2+PColMom_L2	Pyramidal UBOW at level 2 + Pyramidal colour moments at level 2 ((200-D+6-D)×21 = 4326-D)
PIBOW_L1	Pyramidal IBOW at level 1 (# scene categories × 200-D × 5)
PIBOW_L2	Pyramidal IBOW at level 2 (# scene categories × 200-D × 21)
PIBOW_L2+PColMom_L2	Pyramidal IBOW at level 2 + Pyramidal colour moments at level 2 ((# scene categories × 200-D × 21 + (6-D×21))
PIBOW_L2+WPColMom_L2	Pyramidal IBOW at level 2 + Weighted Pyramidal colour moments at level 2 ((# scene categories × 200-D × 21 + (6-D×21))

6.4 Experimental Work

6.4.1 Experimental setup

To be able to compare the performance of different image retrieval algorithms, ground truth images are used, i.e. images in the dataset should be grouped into categories, thus every image in the dataset belongs to one of the predefined scene categories. All experimental works presented in this chapter are evaluated based on 10-folds cross-validation. For a particular image dataset, 10% of

the images are randomly selected from each scene category. These images are used as queries in the image retrieval experiments. The other remaining 90% of the images from each scene category form the ground truth images from which images are retrieved in response to the query images.

Different measures are used to evaluate the performance of the different image retrieval implementations. Firstly, the recall-precision graphs are used for each approach. For each scene category, ranked precisions and recalls of all query images (test images) are averaged and plotted on the recall-precision graph. Another performance measure is the mean average precision (MAP) (*see* Section 6.1). This measure is calculated for each scene category. The third measure is the retrieval accuracy of each approach which is the arithmetic mean of the MAPs over all scene categories. For all image representation approaches, each image is represented as a vector of values which is normalized to a unit length. For multiple feature image representation, to be concatenated, each feature type is first normalized. The Euclidean distance is used to find similarities between the query image and images in the database.

Three natural scene datasets are used in this chapter to evaluate the retrieval performances of different image representation approaches. These datasets were introduced in Section 4.3.2 for the scene categorization task. The first dataset, referred to as *Vogel_6DS*, consists of 700 colour natural scene images distributed over six scene categories (Vogel and Schiele, 2004). The second dataset, referred to as *Oliva_8DS*, consists of 2688 colour images distributed over 8 scene categories (Oliva and Torralba, 2001). The third dataset, referred to as *Lazebnik_15DS*,

contains 4485 gray images distributed over 15 scene categories (Lazebnik et al., 2006).

6.4.2 Experiments on image retrieval using COV

This section presents the experimental work for image retrieval using the concept-occurrence vector evaluated on the dataset *Vogel_6DS*. In this section a set of experiments are carried out. The first experiment is carried out to evaluate the performance of using COV, constructed from the ground truth annotations of image regions, for image retrieval. The results obtained from this experiment serves as a benchmark to evaluate how discriminative are the bag of visual word model and other baseline methods in describing image regions, which in turn used by concept annotator to generate local semantic concept needed to construct the COV.

The MAP results of each scene category using the COV benchmark is depicted in the first row of Table 6-4. The MAPs for scene categories *Coasts* and *River/lakes* are the most difficult categories to retrieve. Images from both scene categories are visually ambiguous. The retrieval accuracy of using the COV benchmark is 83%. The same experiment was carried out in (Vogel and Schiele, 2004) and they achieved 80.6% retrieval accuracy. The difference between their experiment and our experiment is that they used the SVM classifier to rank the retrieved images while in this work the retrieved images are ranked using the Euclidean distance. Thus, to make fair comparisons, the results obtained in this section are used as a benchmark to compare the retrieval performances of using the COV obtained using other approaches. The recall-precision graph of the COV benchmark is also depicted in Figure 6-3 (a).

In the next experiments, the COVs are generated from image regions labelled with local semantic concepts using the different image representation approaches presented in Chapter 5 that gave the best annotation results. These approaches are listed in Table 6-3. Each of these approaches is used to annotate images with local semantic concepts. From these local semantic concepts the COVs are generated as global image representation. These COVs are then used in the retrieval task. It worth to remind that the UBOW and IBOW listed in Table 6-3 are concept-based bag of visual words (CBOW) generated at the upper and lower halves of images using the universal and integrated visual vocabularies.

For simplicity reasons, the UBOW and IBOW are used in this chapter to refer to the CBOW generated either by the universal visual vocabulary or the integrated visual vocabulary.

Table 6-3: Different approaches used to represent image regions. These approaches are used in Chapter 5 to annotate image regions with local semantic concepts.

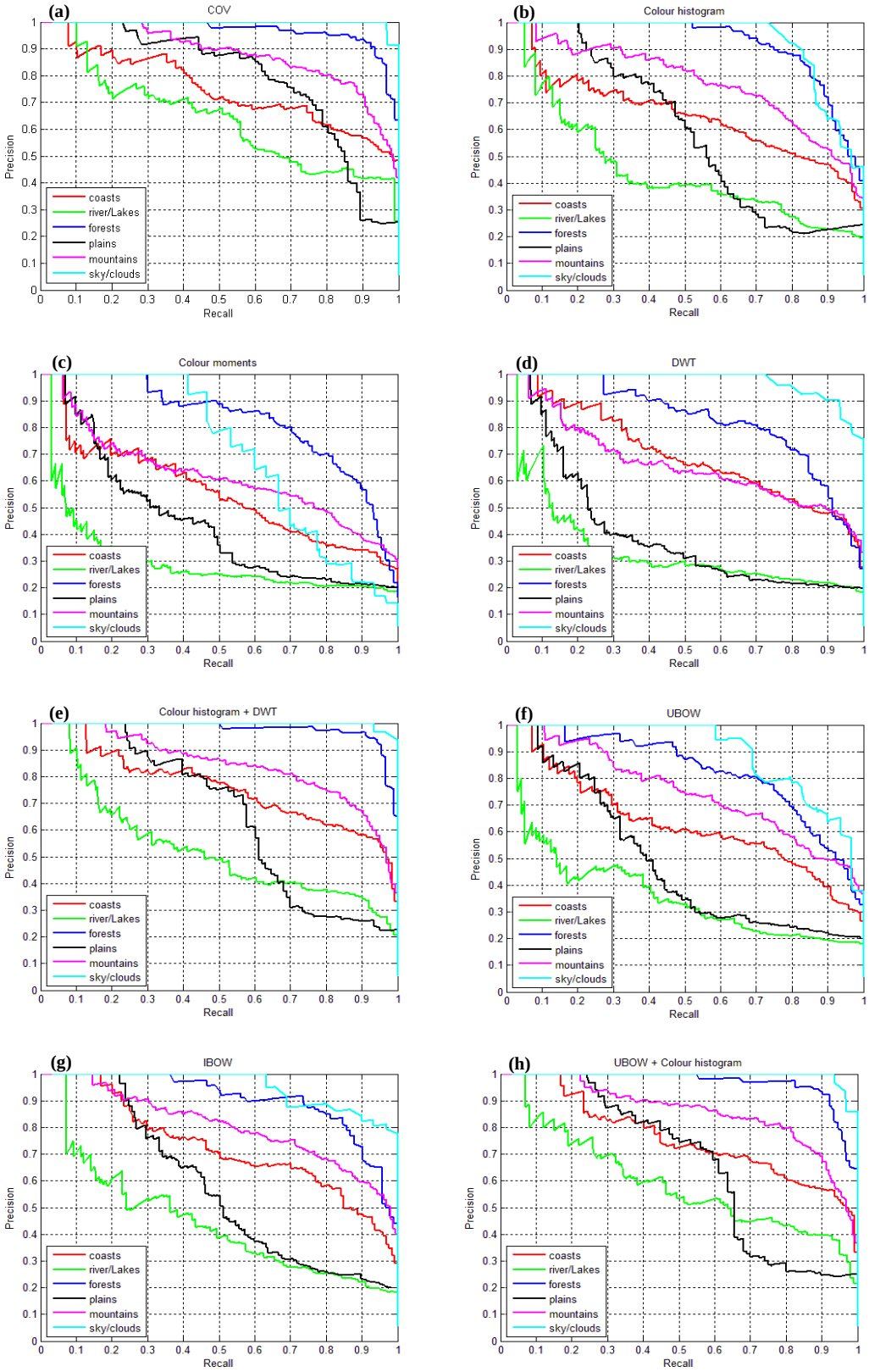
(1) Colour histogram (ColHist)	(2) Colour moments (ColMom)
(3) DWT	(4) Colour histogram + DWT
(5) UBOW	(6) IBOW
(7) UBOW + Colour histogram	(8) UBOW + Colour moments
(9) UBOW + DWT	(10) UBOW + Colour histogram + DWT
(11) IBOW + Colour histogram	(12) IBOW + Colour moments
(13) IBOW + DWT	(14) IBOW + Colour histogram + DWT

The MAPs results of each scene category using the 14 approaches listed in Table 6-3 are shown Table 6-4. The results can be compared directly to the retrieval results obtained using the COV benchmark. The colour histogram shows a good retrieval performance (72%) compared with the colour moments (58%) and DWT

(65%) with a slight improvement when concatenated with the DWT. Interestingly; the colour histogram achieved better retrieval accuracy than the UBOW. The reason for the worse retrieval performance of the UBOW approach is due to that the universal visual vocabulary used to build the UBOW is not discriminative enough. Also, colour information is not included with UBOW.

The UBOW has gained better performance (80%) when it is concatenated with the colour histogram, UBOW + ColHist. The IBOW approach shows better performance than the UBOW and colour histogram and also has improved when combined with the colour histogram. The retrieval results of the 14 experiments revealed that the BOW model combined with the colour information gained very good retrieval results (80%) compared to the retrieval results of the COV benchmark (83%). The recall-precision graphs of the 14 experiments are shown in Figure 6-3 (b-o). From this figure, it is possible to see the differences in the retrieval performance of each scene category and using the 14 different approaches.

The recall-precision plots of each scene category are averaged such that the performance of each approach can be visualized as a recall-precision graph of all approaches. This can be seen in Figure 6-4. It is obvious that the COVs based on BOW models can perform closer to the COVs benchmark.


(a)
COV
(b)
Colour histogram
(c)
Colour moments
(d)
DWT
(e)
Colour histogram + DWT
(f)
UBOW
(g)
IBOW
(h)
UBOW + Colour histogram
Precision
Recall
coasts
river/Lakes
forests
plains
mountains
sky/clouds

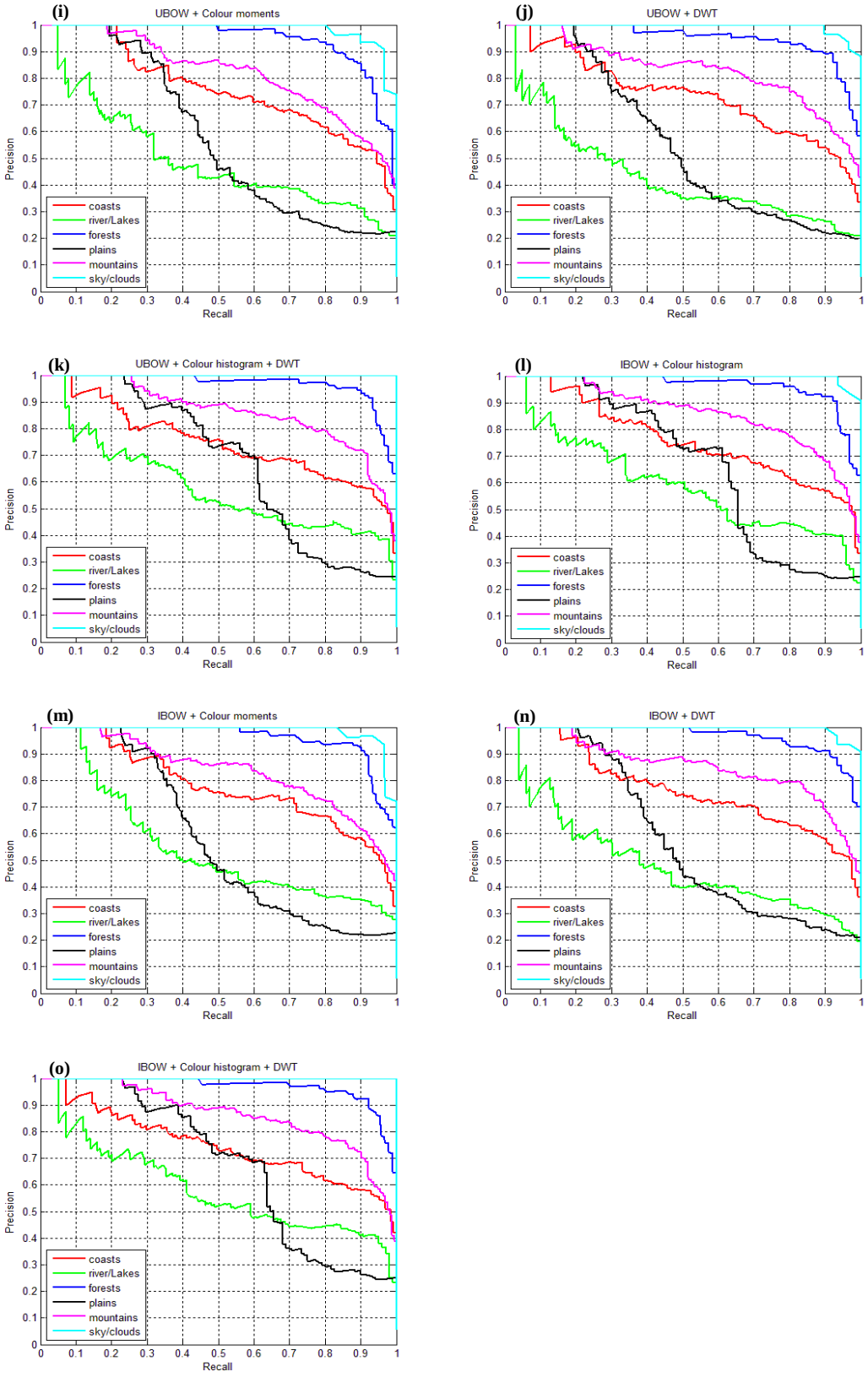


Figure 6-3: Precision-recall graphs, for *Vogel_6DS*, using COV image representation implemented using the ground truth annotations (**a**) and annotations obtained by different region representation approaches (**b**-**o**).

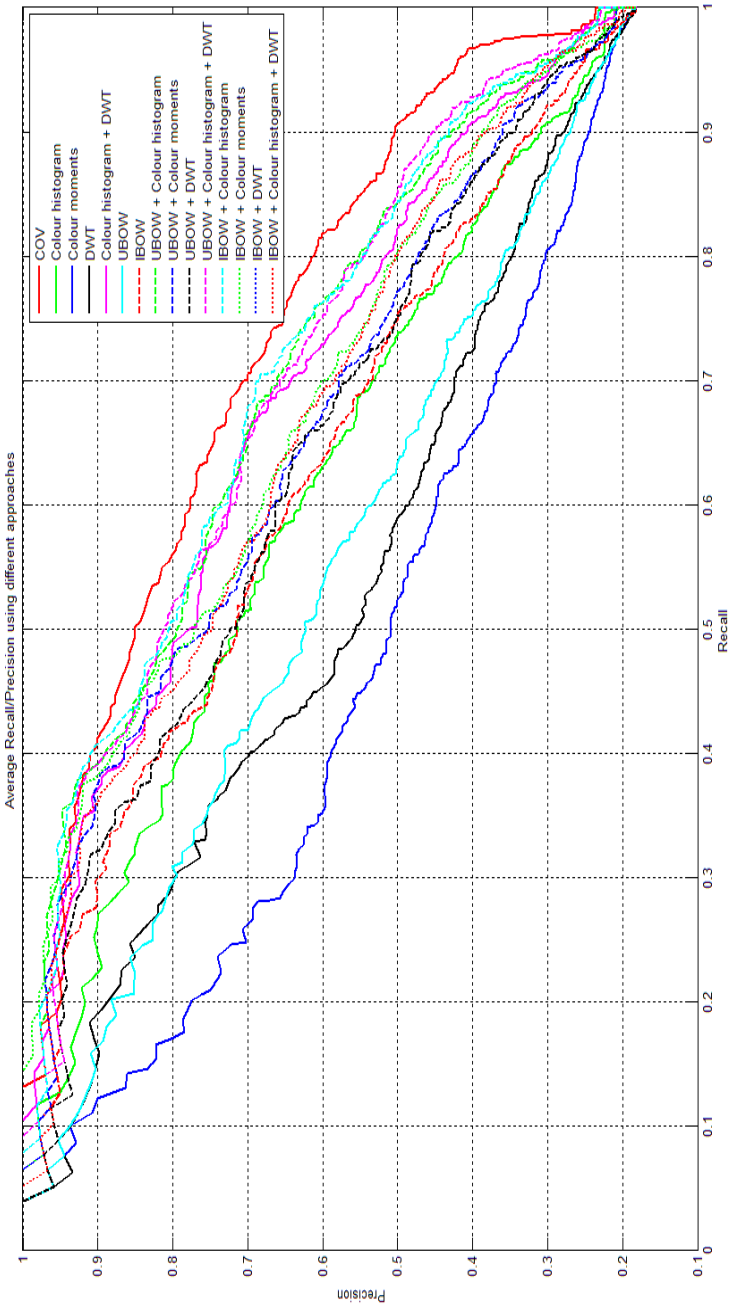


Figure 6-4: Recall-precision graph, for *Vogel_6DS*, of the performance of the COV benchmark and the 14 different approaches.

Table 6-4: The MAPs of each scene category using the COV benchmark and the other 14 different approaches. The last column shows the retrieval accuracy of each of the corresponding approach.

	MAP per scene category						
	Coasts	River/lakes	Forests	Plains	Mountains	Sky/clouds	Acc.
COV	**0.75**	**0.63**	**0.95**	**0.77**	**0.86**	**0.99**	**0.83**
ColHist	0.65	0.46	0.90	0.60	0.76	0.91	0.72
ColMom	0.56	0.33	0.81	0.45	0.62	0.69	0.58
DWT	0.69	0.38	0.81	0.42	0.66	0.96	0.65
ColHist+DWT	0.68	0.49	0.93	0.65	0.75	0.90	0.73
UBOW	0.64	0.39	0.81	0.53	0.74	0.87	0.66
IBOW	0.72	0.45	0.89	0.61	0.79	0.97	0.74
UBOW+ColHist	0.75	0.59	0.95	0.66	0.84	0.99	0.80
UBOW+ColMom	0.75	0.50	0.92	0.58	0.81	0.97	0.75
UBOW+DWT	0.73	0.46	0.93	0.59	0.82	0.99	0.75
UBOW+ColHist+DWT	0.73	0.57	0.93	0.67	0.85	1.00	0.79
IBOW+ColHist	0.75	0.60	0.95	0.67	0.84	0.99	0.80
IBOW+ColMom	0.77	0.56	0.95	0.58	0.82	0.97	0.78
IBOW+DWT	0.76	0.49	0.95	0.58	0.84	0.99	0.77
IBOW+ColHist+DWT	0.74	0.58	0.95	0.67	0.85	1.00	0.80

The retrieval accuracy of all approaches is depicted in Figure 6-5. Another way to analyse the performance of using the 14 approaches to annotate image regions and thus building the COVs is to compare their distribution of the nine semantic concepts in each scene category against the distributions of the nine semantic concepts of the COVs benchmark.

Figure 6-6 shows a scatter plot of the distribution of the nine semantic concepts in each scene category. The x-axis corresponds to the nine semantic concepts while the y-axis is the number of local semantic concepts labeled by a particular approach. Each approach is labeled with a unique colour and shape as shown the legend of the figure. The COV benchmark is labeled with blue diamond shape. It is clear that the distribution of the nine semantic concepts generated by the IBOW+ColHist+DWT approach, labeled with a light green circle shape, is close to the COVs benchmark in most of the scene categories. It gives a closer look on how the different approaches can perform in the retrieval task. Moreover, the distribution

of the nine semantic concepts averaged over all scene categories and for all approaches is shown in Figure 6-7.

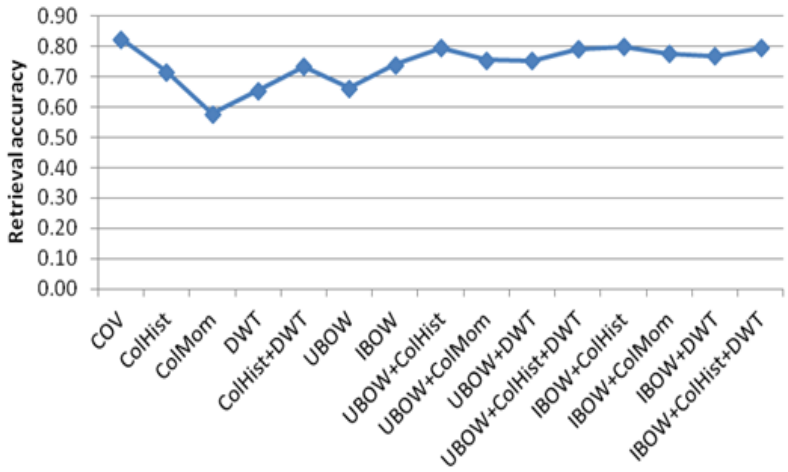
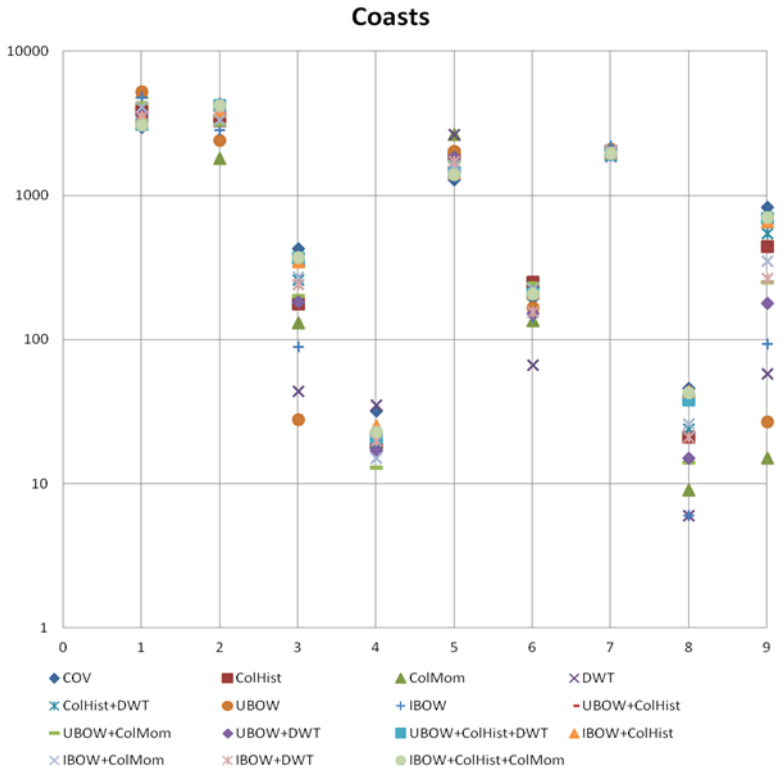
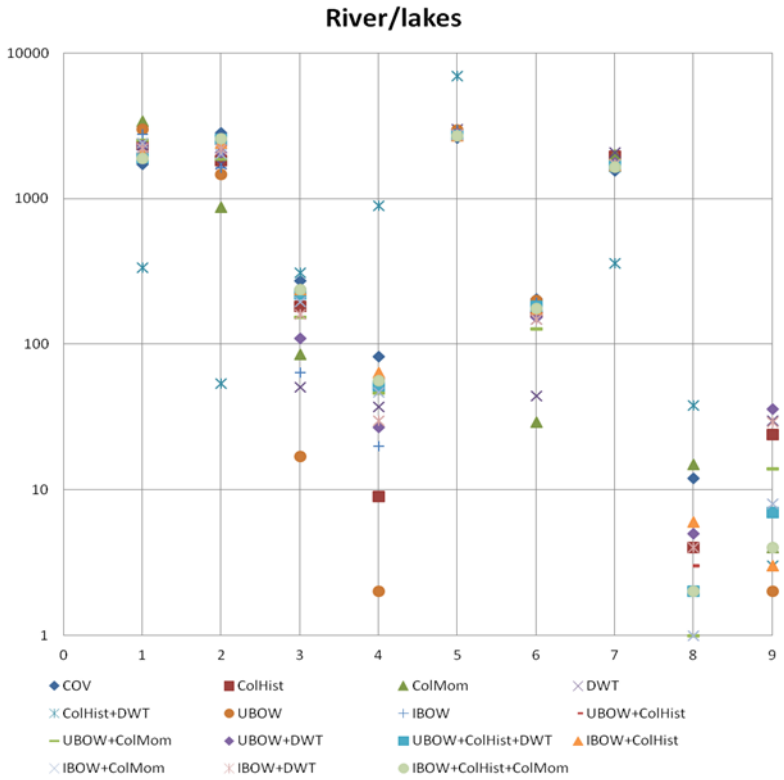
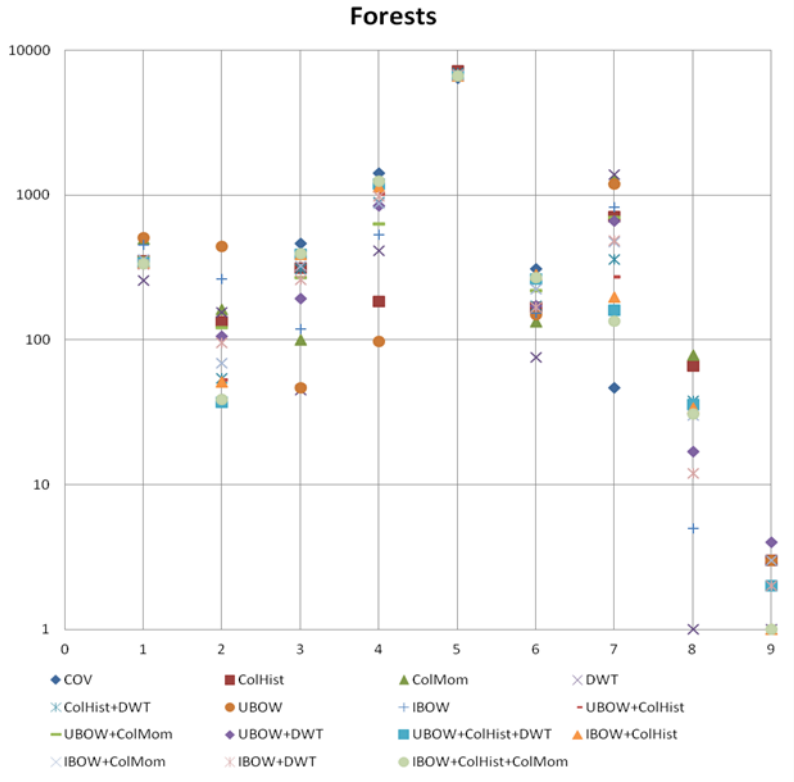
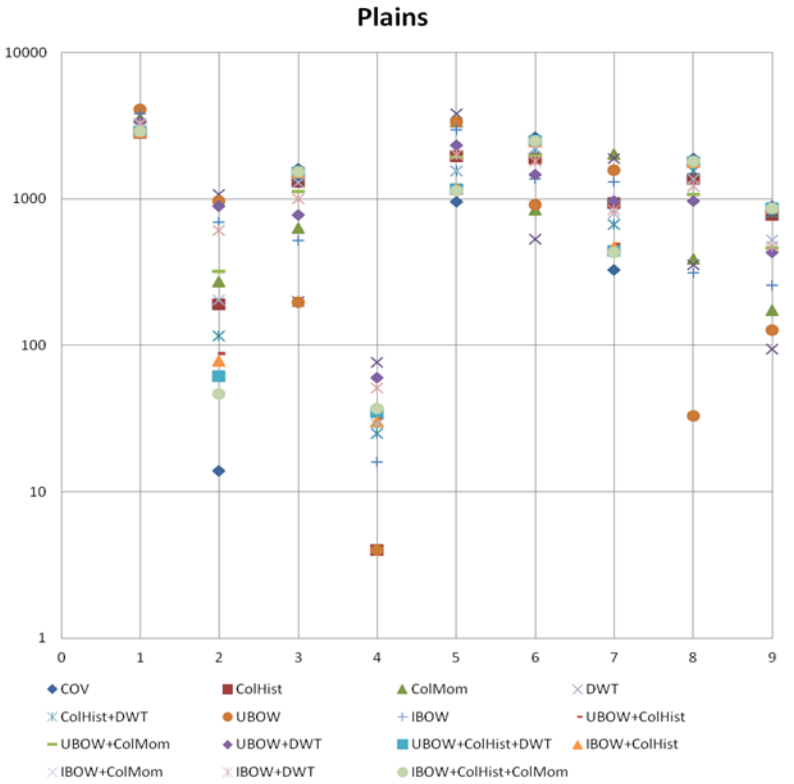


Figure 6-5: Retrieval performance, for *Vogel_6DS*, in terms of the average of MAPs over all scene categories. The x-axis represents different approaches used for image retrieval.


Coasts
10000
1000
100
10
1
0
1
2
3
4
5
6
7
8
9
COV
ColHist
ColMom
DWT
ColHist+DWT
UBOW
IBOW
UBOW+ColHist
UBOW+ColMom
UBOW+DWT
UBOW+ColHist+DWT
IBOW+ColHist
IBOW+ColMom
IBOW+DWT
IBOW+ColHist+ColMom


River/lakes
10000
1000
100
10
1
0
1
2
3
4
5
6
7
8
9
COV
ColHist
ColMom
DWT
ColHist+DWT
UBOW
IBOW
UBOW+ColHist
UBOW+ColMom
UBOW+DWT
UBOW+ColHist+DWT
IBOW+ColHist
IBOW+ColMom
IBOW+DWT
IBOW+ColHist+ColMom


Forests
10000
1000
100
10
1
0
1
2
3
4
5
6
7
8
9
COV
ColHist
ColMom
DWT
ColHist+DWT
UBOW
IBOW
UBOW+ColHist
UBOW+ColMom
UBOW+DWT
UBOW+ColHist+DWT
IBOW+ColHist
IBOW+ColMom
IBOW+DWT
IBOW+ColHist+ColMom


Plains
10000
1000
100
10
1
0
1
2
3
4
5
6
7
8
9
COV
ColHist
ColMom
DWT
ColHist+DWT
UBOW
IBOW
UBOW+ColHist
UBOW+ColMom
UBOW+DWT
UBOW+ColHist+DWT
IBOW+ColHist
IBOW+ColMom
IBOW+DWT
IBOW+ColHist+ColMom

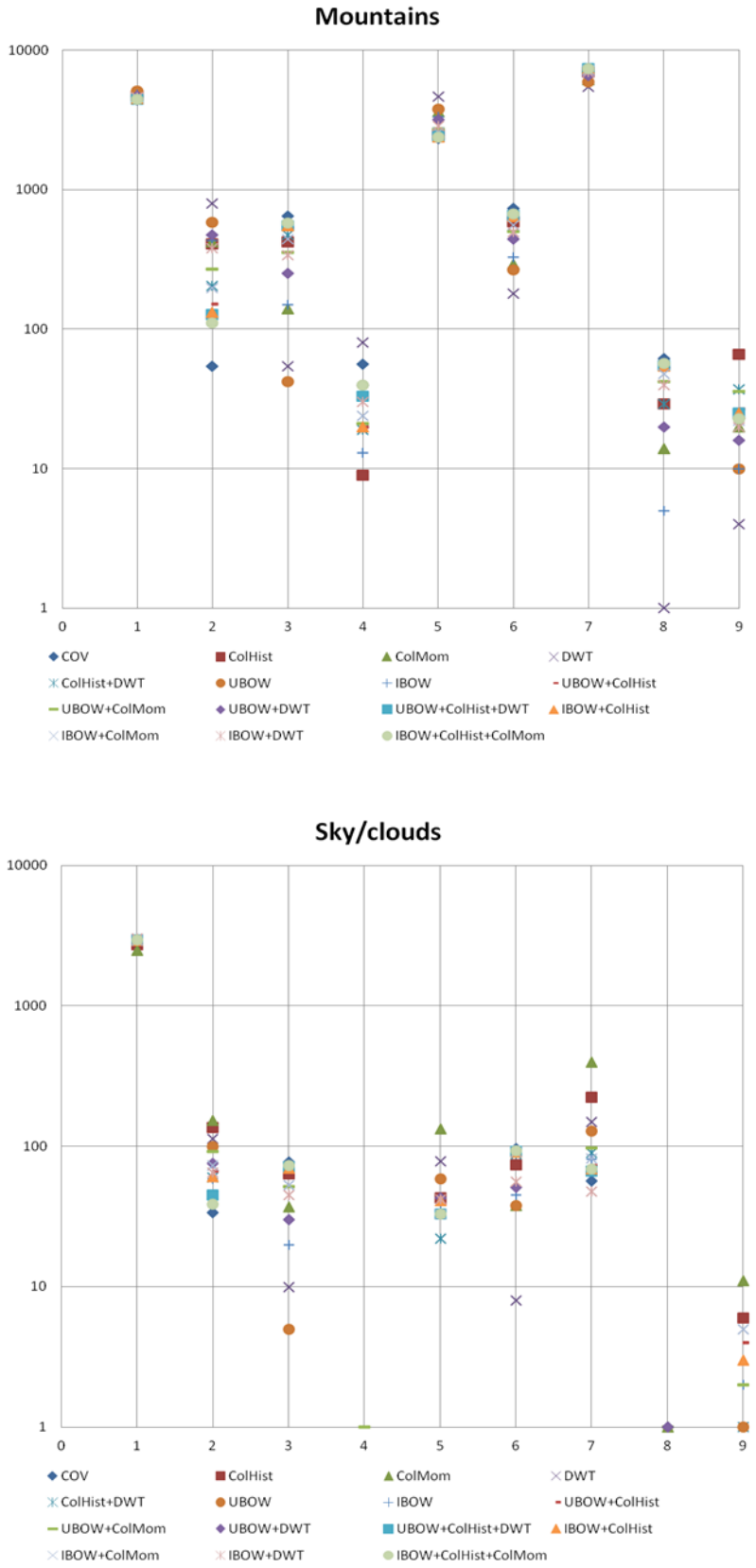


Figure 6-6: Scatter plot of the retrieval accuracy of the COV benchmark and the 14 approaches per scene category. The x-axis represents the nine semantic concepts: *sky*, *water*, *grass*, *trunks*, *foliage*, *field*, *rocks*, *flowers* and *sand*, respectively. Y-axis is in Logarithmic scale.

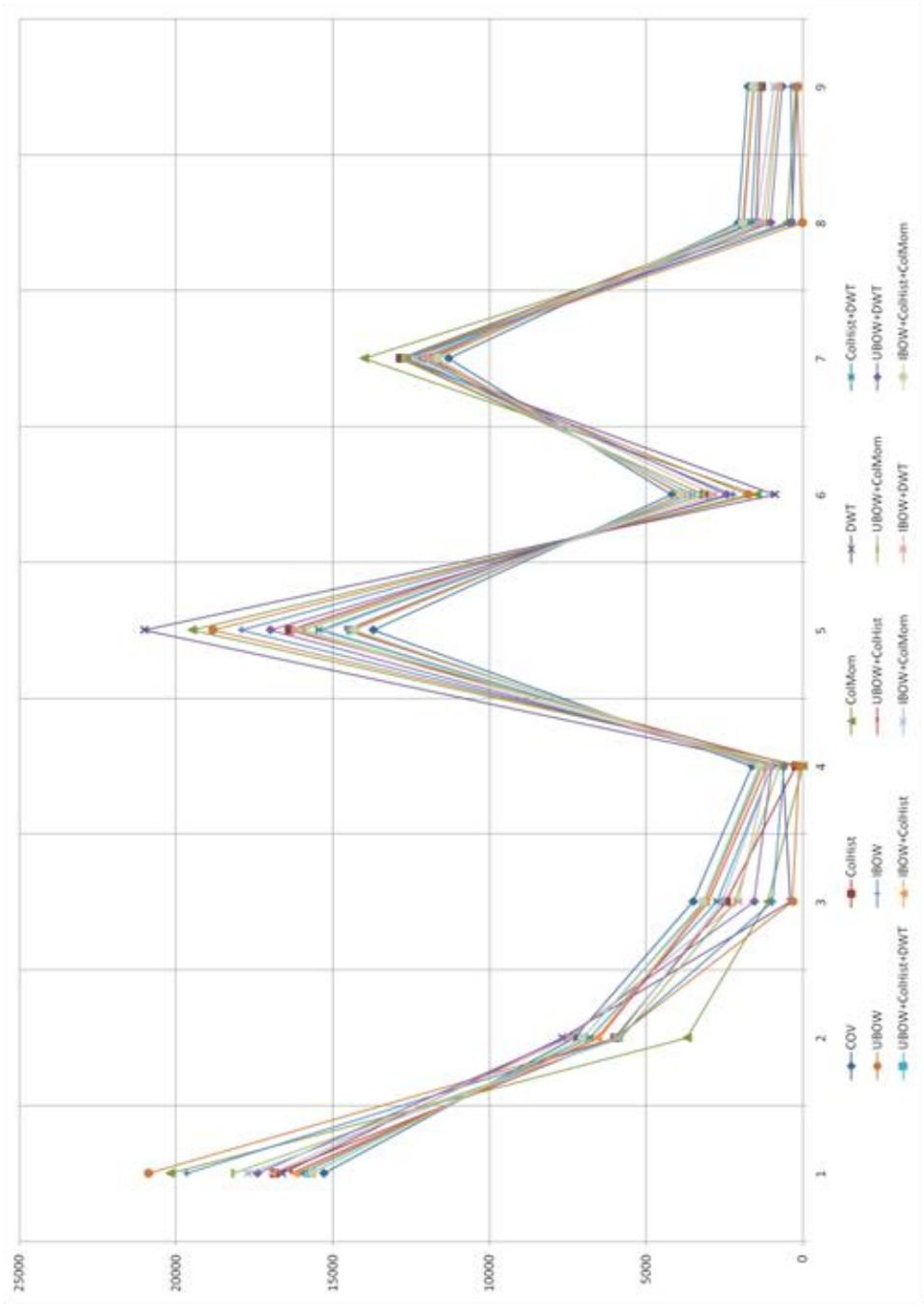


Figure 6-7: The distribution of the nine semantic concepts averaged over all scene categories and for all approaches. The x-axis represents the nine semantic concepts: *sky*, *water*, *grass*, *trunks*, *foliage*, *field*, *rocks*, *flowers* and *sand*, respectively.

6.4.3. Experiments on image retrieval using BOW

This section presents the experimental results of using the different approaches listed in Table 6-2 evaluated on the three natural scene datasets *Vogel_6DS*, *Oliva_8DS* and *Lazebnik_15DS* for natural scene retrieval. The experimental work presented in this section assumes no annotations are available for the natural scenes. The only information available about images is their scene category.

This section tries to answer the following questions. What is the performance of using bag of visual words model for natural scene retrieval? How good is the spatial pyramid bag of visual words model for the natural scene image retrieval? What is the effect of using the proposed weighting approaches presented in Section 4.2 on the performance of scene retrieval? How good these approaches are compared to the baseline methods? These questions can be answered by evaluating the performances of all approaches presented Table 6-2 for natural scene retrieval.

Next, three sets of experiments are presented in the following subsections each of which corresponds to experiments carried out using a particular natural scene dataset.

6.4.3.1 Experimental results: *Vogel_6DS* dataset

This section presents the experimental results of using different configurations of the bag of visual words model to represent the semantic information contained in natural scene images for the natural scene retrieval task. Other baseline methods are also used for comparisons such as colour histogram and DWT. Colour histogram and DWT are extracted from the entire image as a global

representation for that image. The pyramidal colour moments and all configurations of the bag of visual word models were illustrated in Chapter 4. Thus, these approaches will not be explained again.

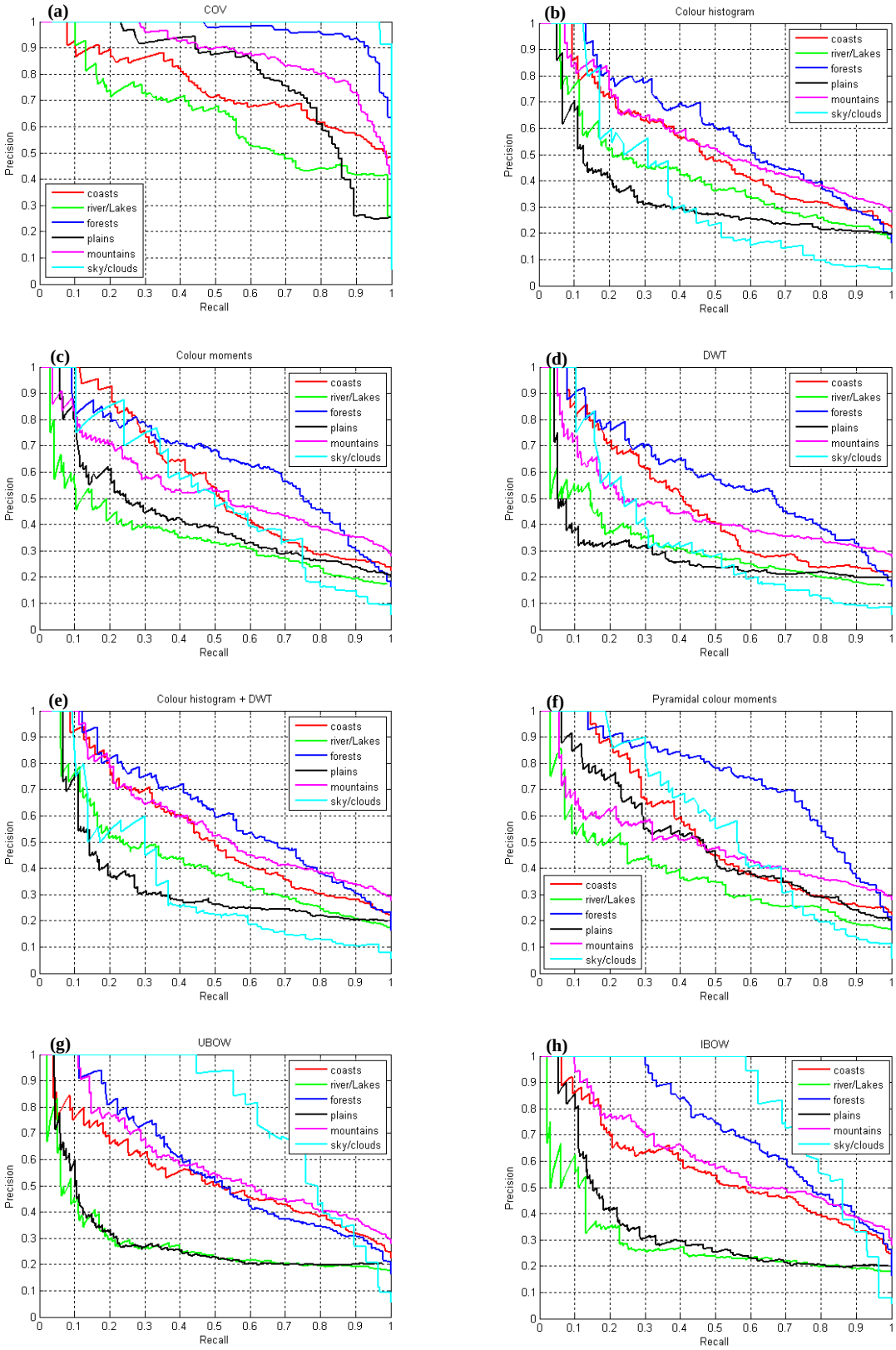
Figure 6-8 shows the recall-precision graphs of using baseline methods, such as colour histogram, and different configurations of the BOW model for natural scene retrieval. The recall-precision graph in Figure 6-8 (a) shows the retrieval performance of using COV benchmark evaluated on the *Vogel_6DS* dataset. The same graph has been shown in the previous section and listed again in this chapter for the sake of comparison. The performance of using bag of visual words to represent image content has gained better retrieval performance over the baseline methods. The IBOW shows better performance than the UBOW. Also, using the spatial pyramid layout has gained another improvement in the retrieval performance over using the UBOW and IBOW without any spatial information. Some scene classes has gained better retrieval performances when the weighting approach is used to add colour to the pyramidal IBOW model.

The performances of all experiments presented in this section are shown in Table 6-5. The pyramidal integrated bag of visual words integrated with the weighed colour moments both implemented at level 2 achieved the best retrieval performance with an increase of (+16%) over the colour histogram and (+13%) over the UBOW. The colour moments perform surprisingly well compared with the colour histogram and DWT. The retrieval accuracy of scene categories Sky/clouds, Plains and Forests have gained batter performance compared to using only baseline methods. To compare the behaviour of the different image representation approaches in the retrieval task, the recall-precision plots of all scene categories presented in each

recall-precision graph are averaged into a single recall-precision plot. The recall-precision plots of the retrieval performance of all image representation approaches are shown in Figure 6-9. The retrieval accuracy of all approaches is depicted in Figure 6-10. It is obvious that all approaches work worse than the COV benchmark. This is due to the fact that the COV approaches rely on the local semantic concepts which require image regions to be annotated by the user. If images in the database are not annotated at region level, then the bag of visual words model becomes a good choice for natural scene retrieval since it demonstrated better retrieval accuracy than the baseline methods.

Table 6-5: The MAPs of each scene category, for *Vogel_6DS*, using the COV benchmark and the other 14 different approaches presented in Table 6-2. The last column shows the retrieval accuracy of each of the corresponding approach.

	MAP per scene category						
	Coasts	River/lakes	Forests	Plains	Mountains	Sky/clouds	Acc.
COV	**0.75**	**0.63**	**0.95**	**0.77**	**0.86**	**0.99**	**0.83**
ColHist	0.53	0.44	0.61	0.38	0.57	0.39	0.49
PColMom_L0	0.58	0.41	0.63	0.46	0.56	0.54	0.53
DWT	0.79	0.37	0.59	0.34	0.48	0.39	0.44
ColHist+DWT	0.55	0.44	0.62	0.39	0.59	0.35	0.49
PColMom_L2	0.55	0.40	0.73	0.51	0.52	0.55	0.54
UBOW	0.54	0.32	0.58	0.35	0.60	0.75	0.52
IBOW	0.57	0.34	0.72	0.40	0.62	0.78	0.57
PUBOW_L1	0.52	0.34	0.65	0.36	0.60	0.76	0.54
PUBOW_L2	0.51	0.34	0.69	0.37	0.61	0.75	0.54
PUBOW_L2+PColMom_L2	0.59	0.36	0.76	0.42	0.63	0.79	0.59
PIBOW_L1	0.56	0.33	0.77	0.43	0.64	0.78	0.58
PIBOW_L2	0.55	0.32	0.78	0.44	0.63	0.78	0.58
PIBOW_L2+PColMom_L2	0.62	0.36	0.82	0.46	0.63	0.81	0.62
PIBOW_L2+WPColMom_L2	0.64	0.45	0.84	0.48	0.63	0.83	0.65


(a)
COV
(b)
Colour histogram
(c)
Colour moments
(d)
DWT
(e)
Colour histogram + DWT
(f)
Pyramidal colour moments
(g)
UBOW
(h)
IBOW
Precision
Recall
coasts
river/Lakes
forests
plains
mountains
sky/clouds

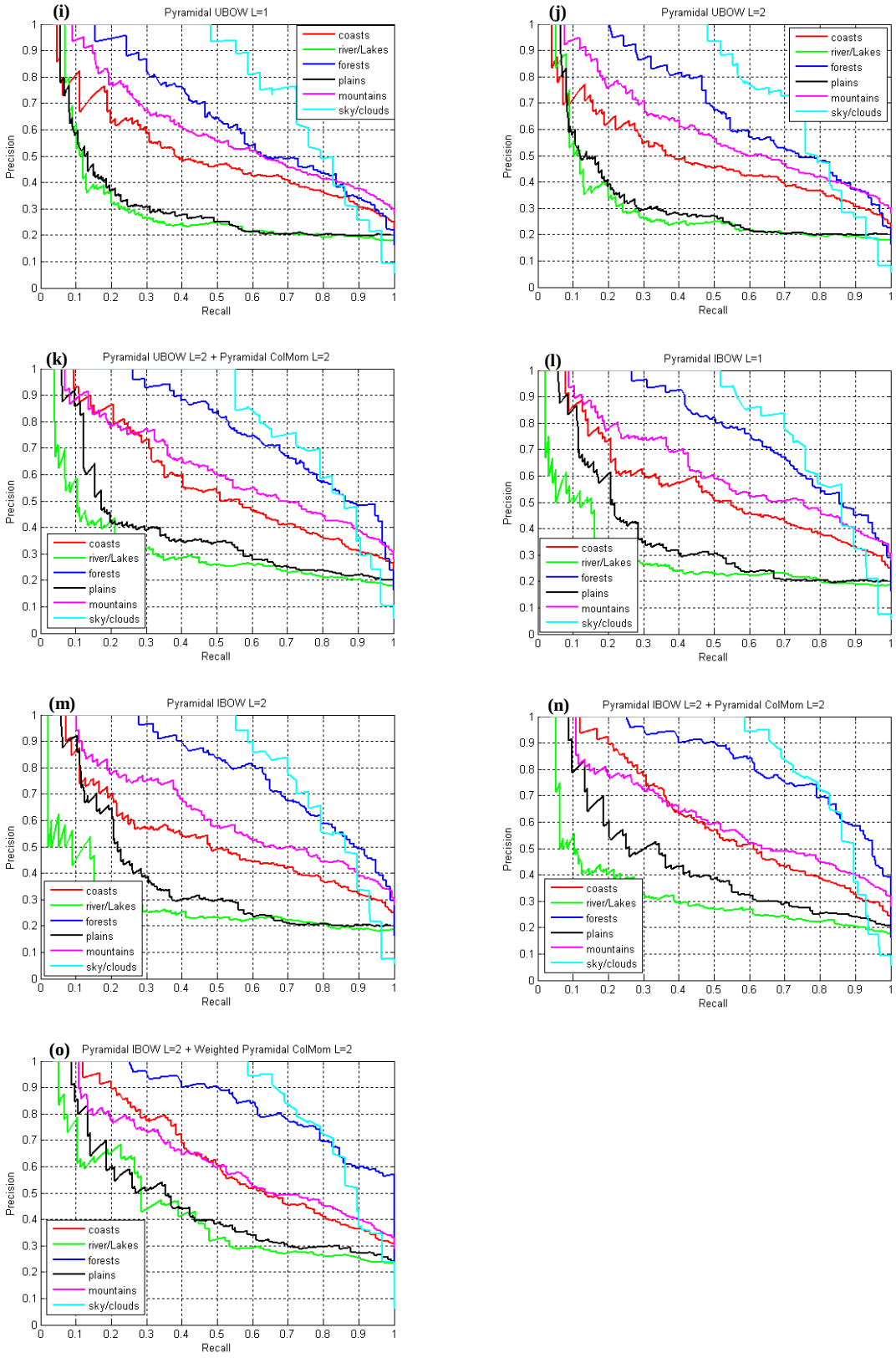


Figure 6-8: Precision-recall graphs, for *Vogel_6DS*, using COV image representation implemented using the ground truth annotations (**a**) and other approaches (**b-o**) presented in Table 6-2.

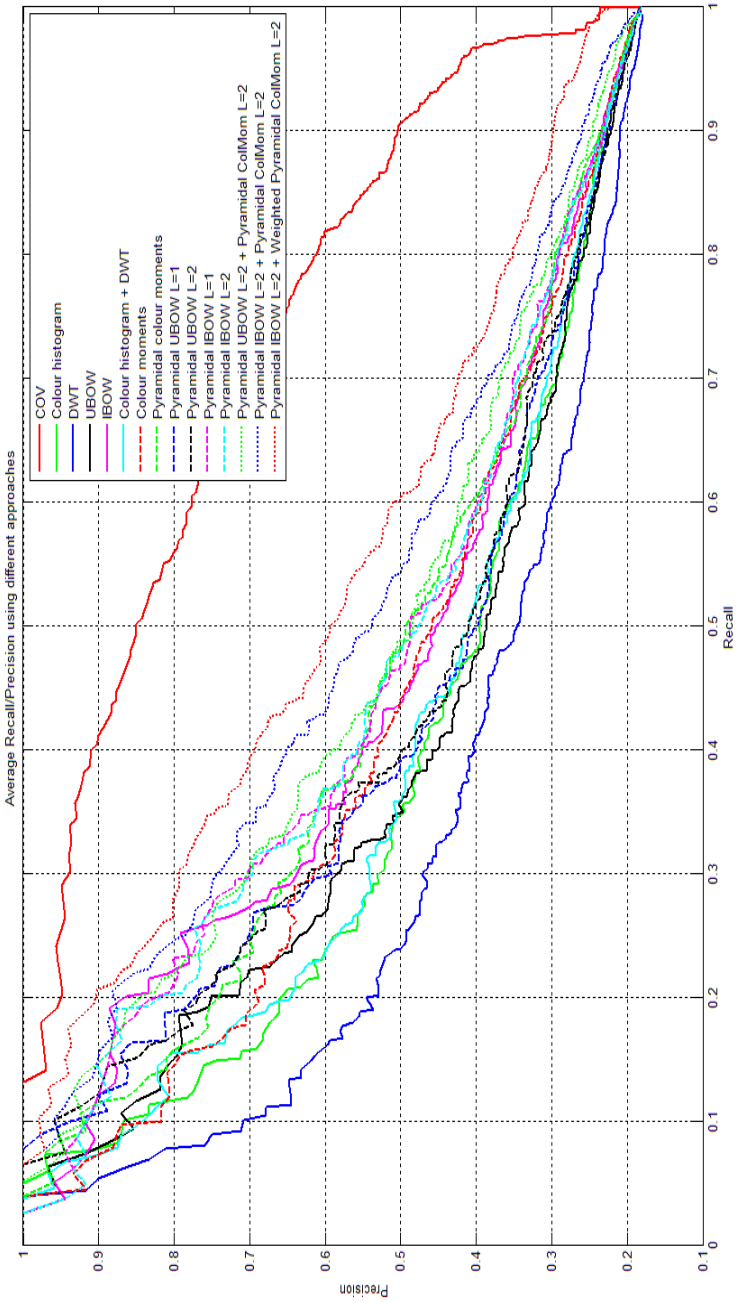


Figure 6-9: Recall-precision graph, for *Vogel_6DS*, of the 14 different approaches presented in Table 6-2 compared against the COV benchmark.

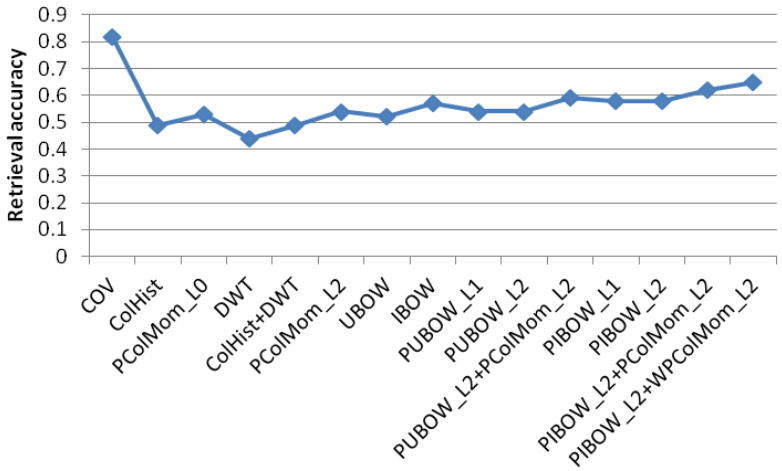


Figure 6-10: Retrieval performance, for *Vogel_6DS*, in terms of the average of MAPs over all scene categories. The x-axis represents different image representation approaches use for image retrieval.

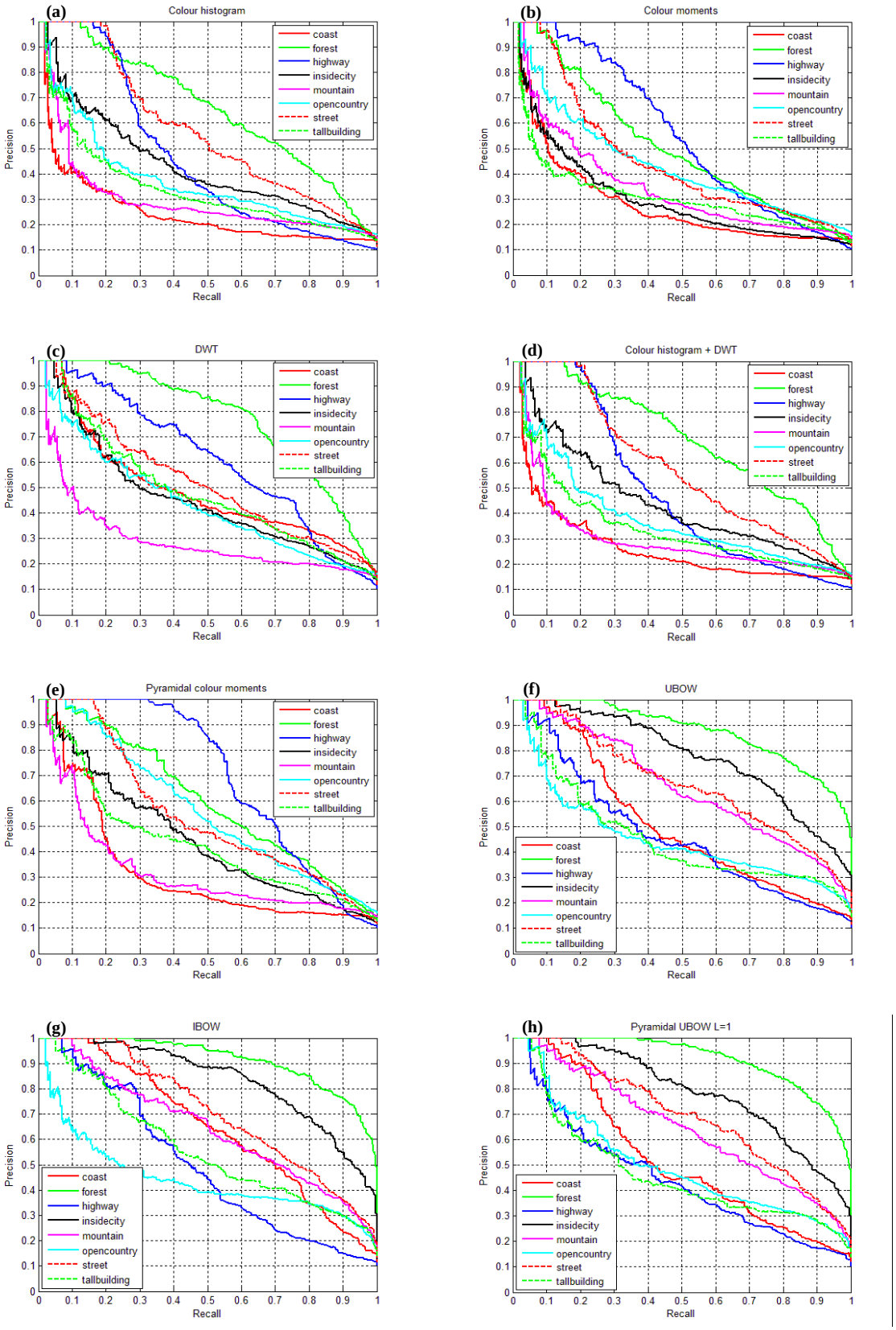
6.4.3.2 Experimental results: *Oliva_8DS*

This section presents the experimental results carried out on the *Oliva_8DS* dataset. The same approaches used in the previous section are also employed in this section but for different dataset. The recall-precision graphs of the retrieval performances are depicted in Figure 6-11. The images in the dataset used in this section are not annotated at image regions. Thus, it is not possible to compare the performance of BOW-based approaches against the COV approach. As mentioned at the beginning of this chapter, the BOW approach can be considered as an intermediate semantic representation of the visual content of images. Compared to baseline methods, the BOW-based approaches achieved better retrieval performances in all scene categories. Using the proposed BOW-based approaches, most scene categories achieved better retrieval results than the baseline methods. It is worth to note that natural scenes with man-made objects, such as Street and Inside city achieved very good performances compared to scene categories without man-

made objects. This can be justified by the ability of the SIFT features to capture the structure if buildings and other man-made objects. For natural scene categories without man-made objects the task of image retrieval becomes harder. The recall-precision graph obtained by averaging the recall-precision plots of all scene categories for each approach is depicted in Figure 6-12. The precision plot of our proposed approach (approach number 14 in the figure) shows always better retrieval accuracy than other approaches. The MAP results of each scene category can be shown in Table 6-6. This table shows that the PIBOW_L2+WPColMom_L2 approach reports (69%) compared with traditional UBOW which has achieved (61%) retrieval rate. The retrieval accuracy of the approaches is also depicted in Figure 6-13.

Table 6-6: The MAPs of each scene category, for *Oliva_8DS*, using the 14 different approaches presented in Table 6-2. The last column shows the retrieval accuracy of each of the corresponding approach.

	MAP per category								
	Coast	Forest	Highway	Inside city	Mountain	Open country	Street	Tall building	Acc.
ColHist	0.29	0.65	0.47	0.44	0.32	0.40	0.57	0.37	0.44
PColMom_L0	0.32	0.52	0.56	0.33	0.37	0.45	0.46	0.34	0.42
DWT	0.50	0.75	0.61	0.46	0.33	0.45	0.53	0.49	0.52
ColHist+DWT	0.30	0.68	0.49	0.45	0.33	0.40	0.58	0.37	0.45
PColMom_L2	0.34	0.60	0.68	0.47	0.36	0.56	0.54	0.45	0.50
UBOW	0.52	0.86	0.48	0.76	0.65	0.47	0.67	0.47	0.61
IBOW	0.64	0.89	0.51	0.81	0.64	0.46	0.70	0.56	0.65
PUBOW_L1	0.53	0.90	0.46	0.77	0.64	0.51	0.69	0.48	0.62
PUBOW_L2	0.53	0.91	0.45	0.78	0.63	0.50	0.69	0.49	0.62
PUBOW_L2+PColMom_L2	0.52	0.90	0.63	0.74	0.62	0.62	0.69	0.56	0.66
PIBOW_L1	0.65	0.91	0.50	0.82	0.62	0.48	0.72	0.56	0.65
PIBOW_L2	0.65	0.91	0.50	0.82	0.61	0.48	0.71	0.56	0.65
PIBOW_L2+PColMom_L2	0.52	0.87	0.69	0.75	0.55	0.62	0.69	0.58	0.66
PIBOW_L2+WPColMom_L2	0.57	0.87	0.69	0.77	0.59	0.64	0.77	0.62	0.69


(a)
Colour histogram
(b)
Colour moments
(c)
DWT
(d)
Colour histogram + DWT
(e)
Pyramidal colour moments
(f)
UBOW
(g)
IBOW
(h)
Pyramidal UBOW L=1
Precision
Recall
coast
forest
highway
insidecity
mountain
opencountry
street
tallbuilding

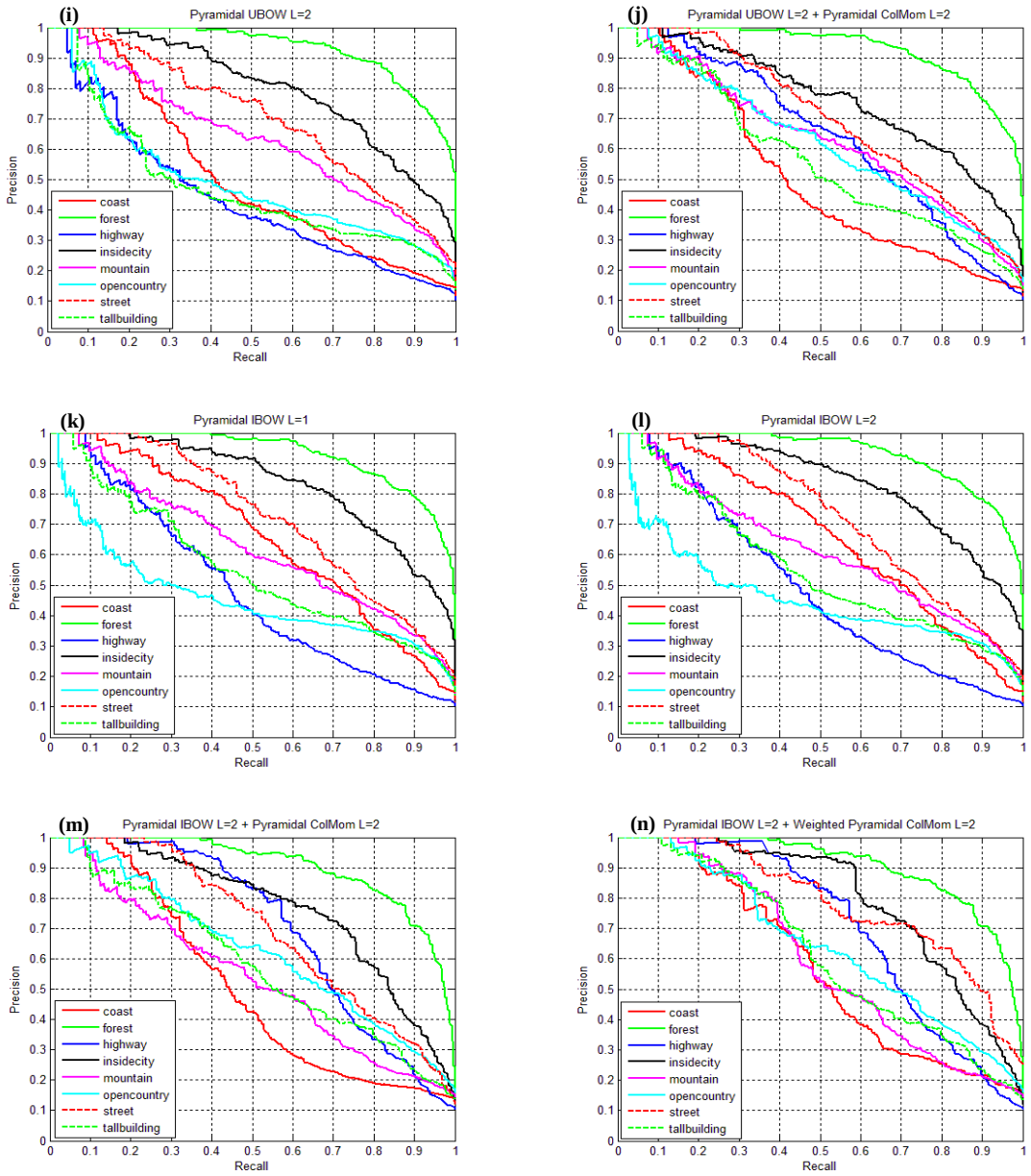


Figure 6-11: Precision-recall graphs, for *Oliva_8DS*, using different approaches presented in Table 6-2.

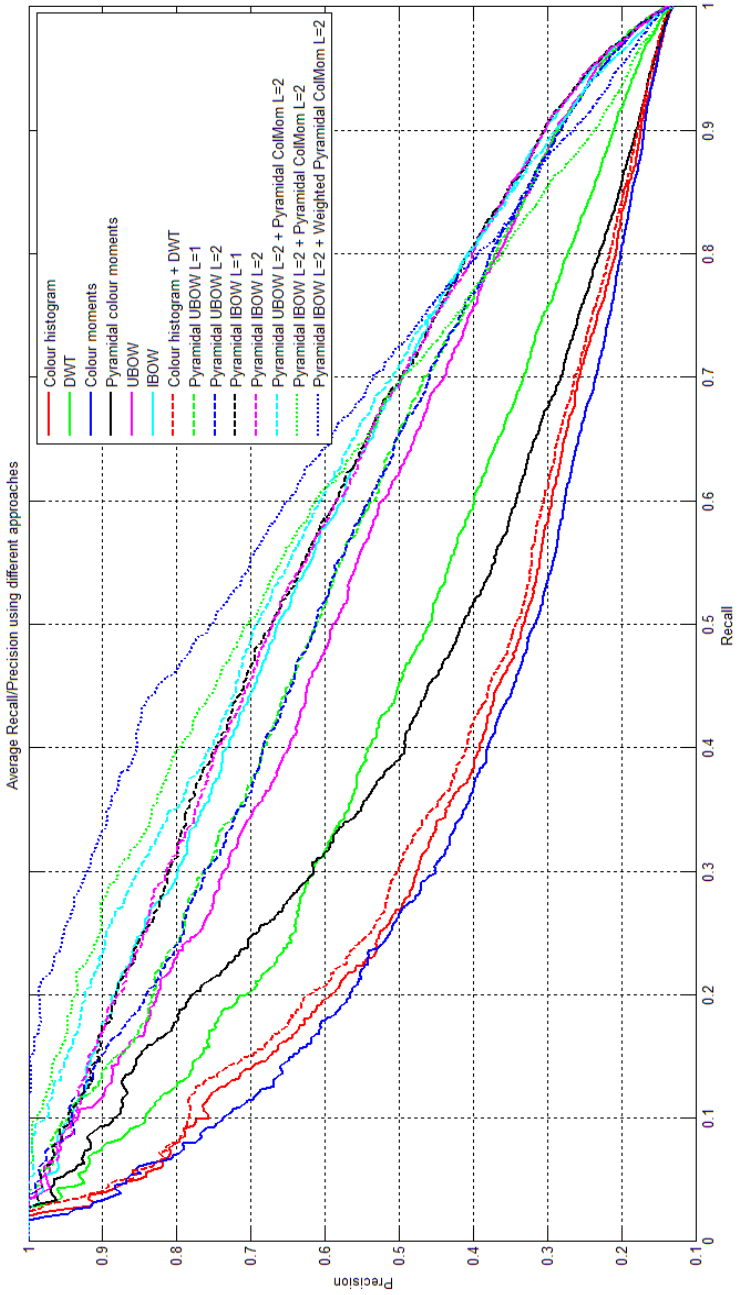


Figure 6-12: Recall-precision graph, for *Oliva_8DS*, of the 14 different approaches presented in Table 6-2.

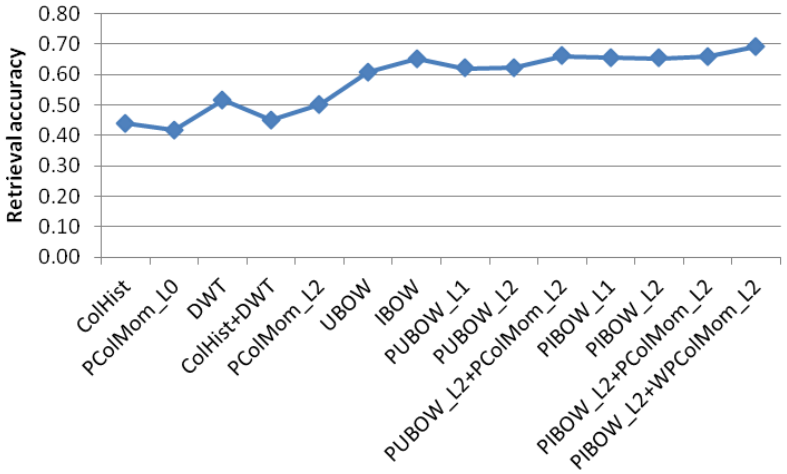


Figure 6-13: Retrieval performance, for *Oliva_8DS*, in terms of the average of MAPs over all scene categories. The x-axis represents different approaches used for image retrieval.

6.4.3.3 Experimental results: *Lazebnik_15DS*

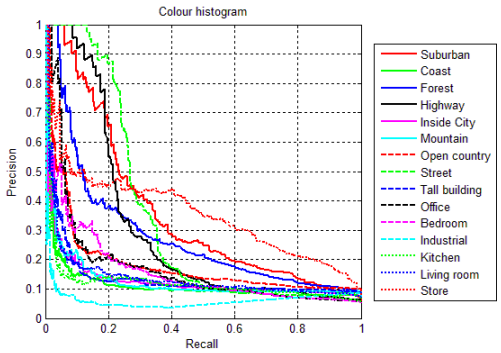
This section demonstrates the retrieval performance of using BOW-based approaches and other baseline methods evaluated on a larger dataset of 15 scene categories. The dataset contains gray images of indoor and outdoor scenes. Some of the baseline approaches, presented in the previous two sections, are extracted from colour images, i.e. from the three colour components of HSV. For gray images, these approaches are only available from one component, i.e. the gray component.

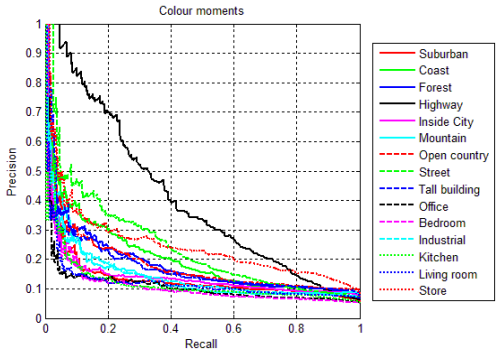
The recall-precision graphs of the 14 different experiments are depicted in Figure 6-14. It shows that the baseline methods failed to retrieve images for most of the scene categories. The colour moments shows good performance for the scene category Highway but failed for other scene categories. In contrast, the BOW-based approaches shown good retrieval performances compared with the baseline methods.

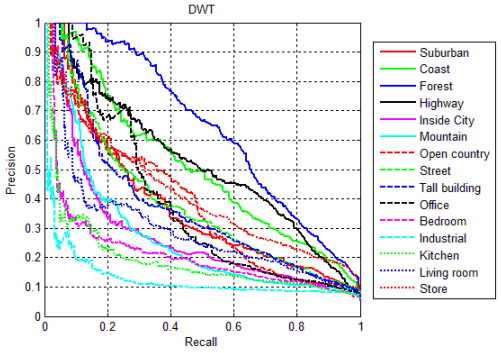
The MAP for each scene category as well as the overall scene retrieval rate using different image representation approaches are listed in Table 6-7. The best retrieval rate is achieved using PIBOW_L2+WPColMom_L2. It indicates that the BOW-based approaches are appropriate for indoor/outdoor scene categories. The approach has reported (49%) retrieval rate with (+10%) increase in the retrieval rate over the best baseline methods. Also, the performance of the UBOW has achieved good results when the spatial pyramid layout is employed. The recall-precision graph of the different approaches is shown in Figure 6-15, where the recall-precision plots of each scene category are averaged for all different approaches. The retrieval performance of the different approaches is also shown in Figure 6-16.

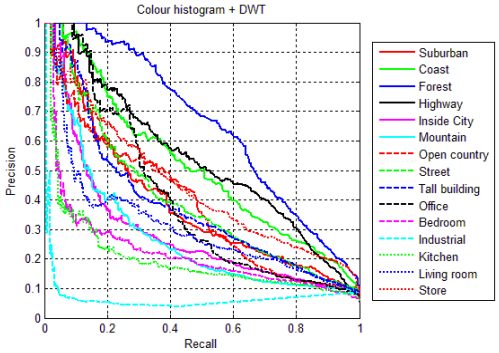
Table 6-7: The MAPs of each scene category, for *Lazebnik_15DS*, using the 14 different approaches presented in Table 6-2. The last column shows the retrieval accuracy of each of the corresponding approach.

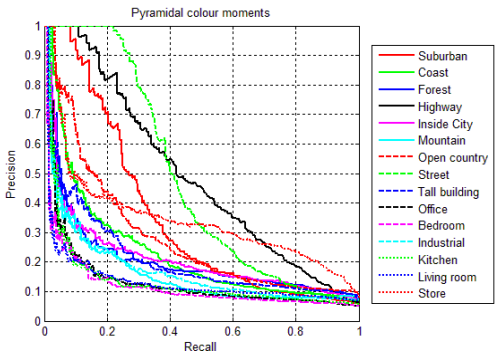
	MAP per category															
	Suburban	Coast	Forest	Highway	Inside City	Mountain	Open country	Street	Tall building	Office	Bedroom	Industrial	Kitchen	Living room	Store	Accuracy
ColHist	0.38	0.18	0.31	0.32	0.19	0.19	0.23	0.36	0.19	0.21	0.21	0.14	0.18	0.20	0.38	0.24
PColMom_L0	0.20	0.25	0.23	0.42	0.20	0.20	0.24	0.28	0.23	0.18	0.18	0.21	0.18	0.18	0.29	0.23
DWT	0.39	0.51	0.63	0.52	0.30	0.30	0.40	0.40	0.39	0.40	0.25	0.19	0.23	0.32	0.43	0.38
ColHist+DWT	0.40	0.52	0.64	0.54	0.31	0.31	0.40	0.41	0.40	0.41	0.26	0.15	0.24	0.33	0.46	0.39
PColMom_L2	0.37	0.28	0.25	0.50	0.25	0.23	0.32	0.48	0.26	0.19	0.18	0.22	0.19	0.19	0.36	0.28
UBOW	0.66	0.44	0.83	0.34	0.40	0.55	0.44	0.53	0.24	0.48	0.22	0.24	0.39	0.33	0.53	0.44
IBOW	0.63	0.57	0.88	0.43	0.49	0.58	0.41	0.54	0.36	0.53	0.22	0.27	0.38	0.29	0.51	0.47
PUBOW_L1	0.69	0.45	0.88	0.34	0.40	0.53	0.49	0.53	0.24	0.46	0.22	0.26	0.37	0.32	0.55	0.45
PUBOW_L2	0.69	0.45	0.90	0.34	0.41	0.52	0.49	0.53	0.24	0.45	0.23	0.26	0.37	0.32	0.57	0.45
PUBOW_L2+PColMom_L2	0.70	0.43	0.88	0.46	0.42	0.47	0.54	0.57	0.24	0.43	0.23	0.25	0.37	0.31	0.57	0.46
PIBOW_L1	0.66	0.58	0.89	0.42	0.50	0.56	0.43	0.58	0.36	0.54	0.22	0.27	0.36	0.29	0.49	0.48
PIBOW_L2	0.66	0.57	0.89	0.43	0.50	0.55	0.43	0.59	0.37	0.54	0.22	0.26	0.34	0.29	0.49	0.48
PIBOW_L2+PColMom_L2	0.54	0.52	0.82	0.58	0.45	0.38	0.50	0.60	0.33	0.42	0.22	0.25	0.31	0.26	0.50	0.44
PIBOW_L2+WPColMom_L2	0.74	0.64	0.91	0.59	0.45	0.58	0.53	0.60	0.34	0.42	0.22	0.25	0.31	0.26	0.50	0.49

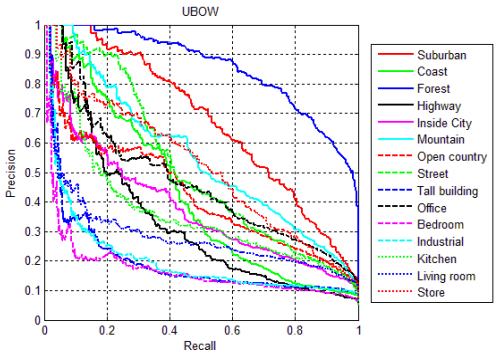

Colour histogram
Precision
Recall
Suburban
Coast
Forest
Highway
Inside City
Mountain
Open country
Street
Tall building
Office
Bedroom
Industrial
Kitchen
Living room
Store

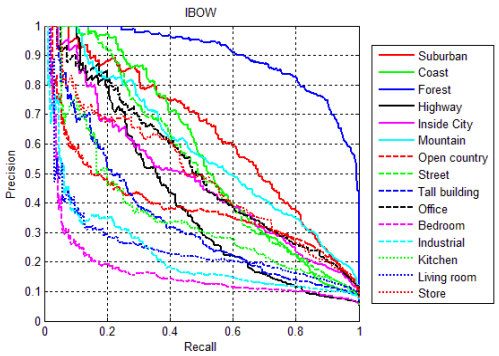

Colour moments
Precision
Recall
Suburban
Coast
Forest
Highway
Inside City
Mountain
Open country
Street
Tall building
Office
Bedroom
Industrial
Kitchen
Living room
Store

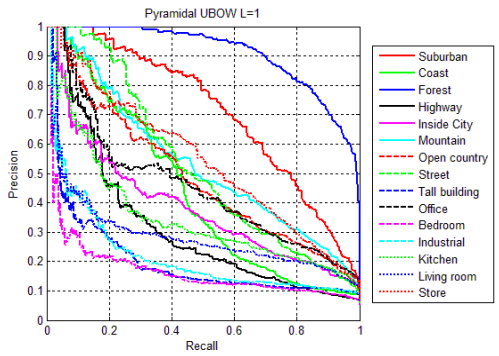

DWT
Precision
Recall
Suburban
Coast
Forest
Highway
Inside City
Mountain
Open country
Street
Tall building
Office
Bedroom
Industrial
Kitchen
Living room
Store


Colour histogram + DWT
Precision
Recall
Suburban
Coast
Forest
Highway
Inside City
Mountain
Open country
Street
Tall building
Office
Bedroom
Industrial
Kitchen
Living room
Store


Pyramidal colour moments
Precision
Recall
Suburban
Coast
Forest
Highway
Inside City
Mountain
Open country
Street
Tall building
Office
Bedroom
Industrial
Kitchen
Living room
Store


UBOW
Precision
Recall
Suburban
Coast
Forest
Highway
Inside City
Mountain
Open country
Street
Tall building
Office
Bedroom
Industrial
Kitchen
Living room
Store


IBOW
Precision
Recall
Suburban
Coast
Forest
Highway
Inside City
Mountain
Open country
Street
Tall building
Office
Bedroom
Industrial
Kitchen
Living room
Store


Pyramidal UBOW L=1
Precision
Recall
Suburban
Coast
Forest
Highway
Inside City
Mountain
Open country
Street
Tall building
Office
Bedroom
Industrial
Kitchen
Living room
Store

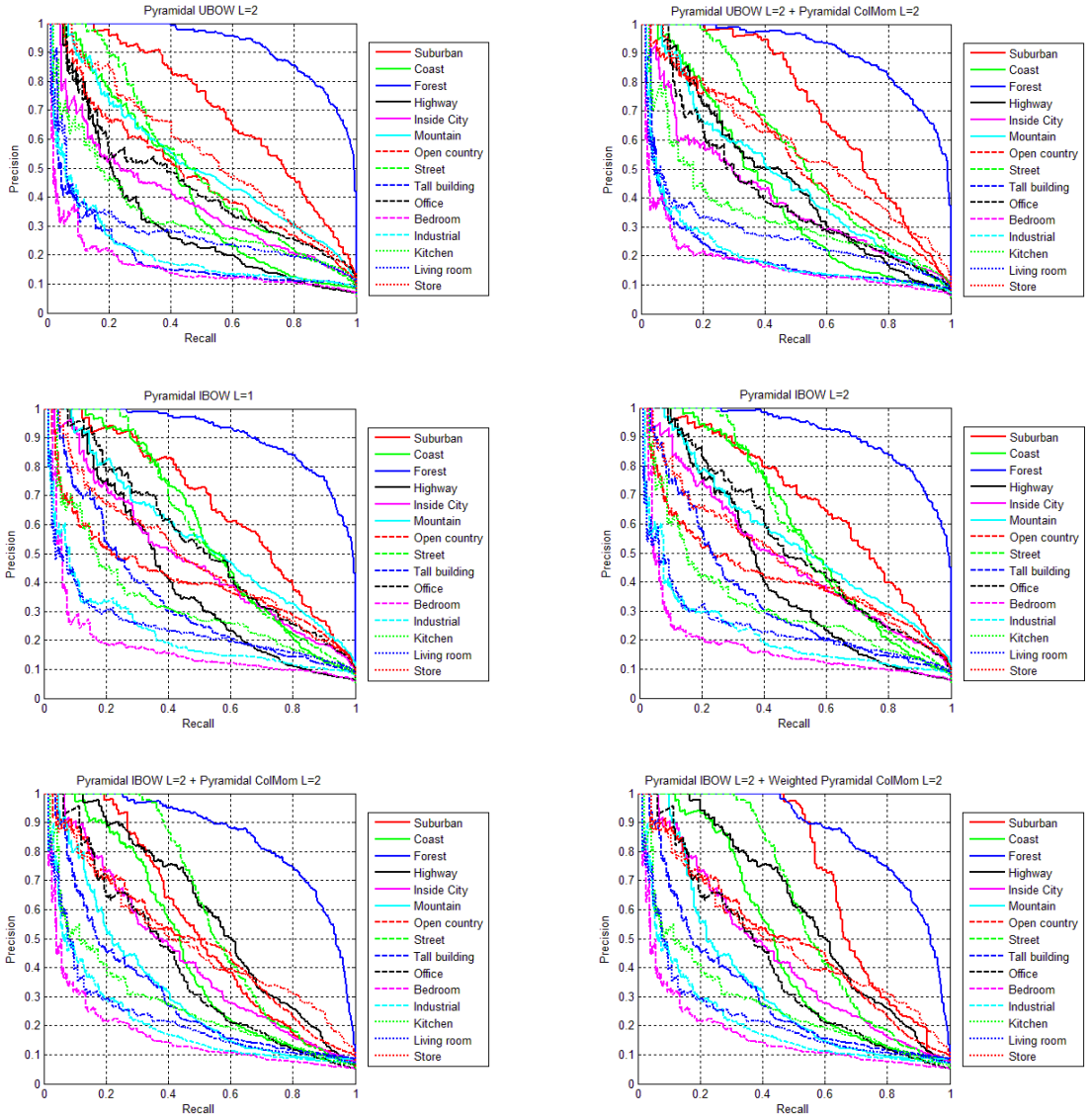


Figure 6-14: Precision-recall graphs, for *Lazebnik_15DS*, using different approaches presented in Table 6-2.

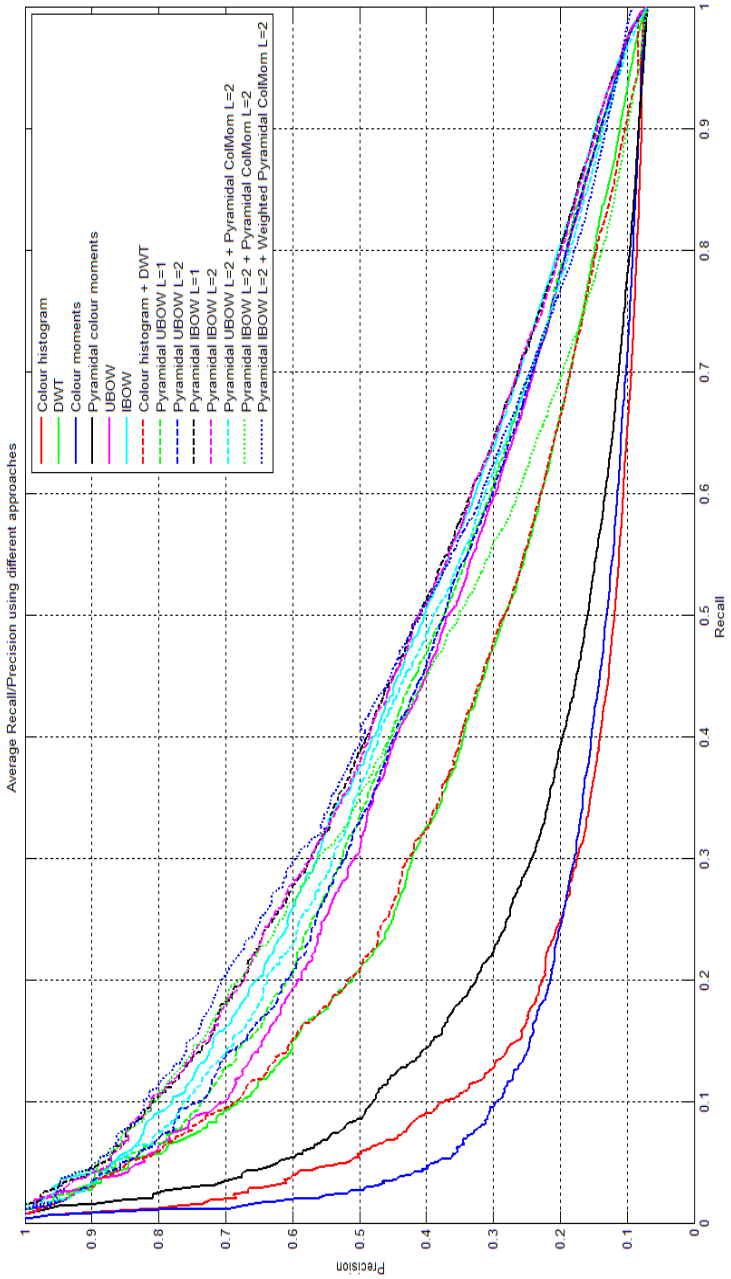


Figure 6-15: Recall-precision graph, for *Lazebnik_15DS*, of the 14 different approaches presented in Table 6-2.

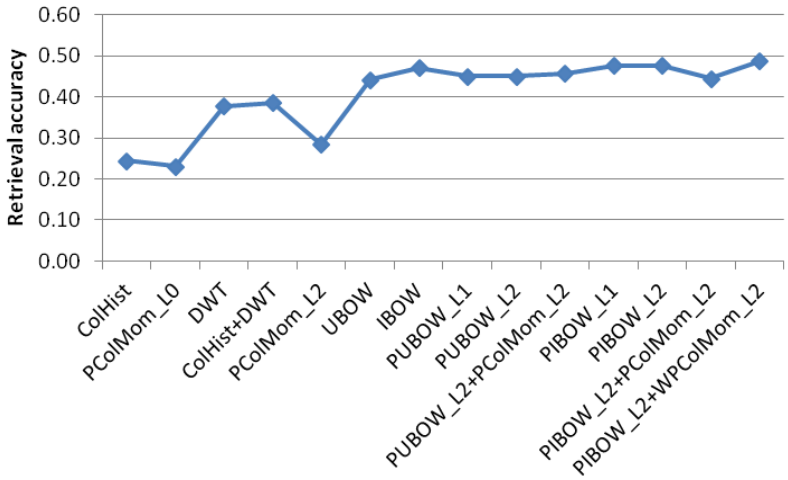


Figure 6-16: Retrieval performance, for *Lazebnik_15DS*, in terms of the average of MAPs over all scene categories. The x-axis represents different approaches used for image retrieval.

6.5. Summary

In this chapter, several experiments concerning the semantic retrieval of natural scenes have been carried out. The semantic representation of the natural scene images implemented using the two scenarios presented at the beginning of this chapter. Firstly, the retrieval performance when employing the COV to summarize the amount of local semantic concepts depicted in an image have reported an encouraging results. The COV constructed from the labels of image regions represented by the BOW model have shown better performance compared with the baseline methods, such as colour histogram, and also comparable with the COV benchmark (*see* Table 6-4).

Secondly, the retrieval performance of using different configuration of the bag of visual word model have been studied and evaluated experimentally using three natural scene datasets. The experimental results obtained using the *Vogel_6DS*

dataset have shown that the COV approaches achieved better retrieval accuracy compared to the BOW-based approaches and baseline methods (*see* Table 6-5). Also, the COV, as a global image representation, has lower dimensionality (9-D) than all other approaches. However, using the COV requires all image regions to be annotated manually.

In the case of representing the semantic information of image content without using the COV approach, the proposed approaches presented in Chapter 4 for the image classification task have achieved better retrieval performance compared to other baseline methods on the datasets *Oliva_8DS* and *Lazebnik_15DS*. The retrieval accuracies of the two datasets using the different image representation approaches are reported in Table 6-6 and Table 6-7.

Chapter 7

Conclusions and Future Work

This thesis has proposed a number of techniques for semantic-based image representation which are based on the bag of visual words (BOW) model. These techniques have been applied to three challenging problems in the computer vision community: natural scene categorization, annotation and retrieval. This chapter presents a summary of the work presented in this thesis highlighting the main contributions and conclusions and suggesting some recommendations for the future work.

7.1 Summary of Contributions and Conclusions

This thesis has presented different approaches for semantic-based image representation based on the well-known bag of visual words model. These approaches have been investigated and evaluated on three challenging tasks in the computer vision community: natural scene image classification, annotation and retrieval. The performance of these tasks is typically influenced by the discriminative power of the approaches utilized to represent the visual content of images. In content-based image retrieval the focus was on developing techniques for extracting low-level features from images to represent their visual content, which ignored the semantic gap between the visual content of image features and the user perception.

Advances in image analysis have led to development of features that are invariant to geometric transformations. Based on such features, the bag of visual word model has become a standard choice for many computer vision tasks. This model has shown an impressive performance in image classification and object recognition problems. Nevertheless, the BOW model still needs further investigation of its ability to represent the semantic information of the image content and how it performs in representing natural scene images for classification, annotations and retrieval tasks.

Despite that much progress and many efforts have been made during the past few years, investigating the robustness and improving the quality of the BOW model for representing the semantic information of the image content are an open and very challenging tasks for natural scene classification, annotation and retrieval.

To this point, this thesis investigated different approaches introduced by the author for representing the semantic information of images based on the bag of visual words model. The proposed approaches have addressed different methods for improving the discriminative power of the BOW model. These contributions are evaluated on natural scene images for three different tasks: image classification, annotation and retrieval.

In Chapter 2, an extensive and structured literature review has been conducted focusing on research progress, advances and techniques that are mostly related to the work presented in this thesis. In Chapter 3, the author has introduced basic concepts related to different topics used throughout the thesis. A summary and conclusions of the original contributions that have been presented for each task are as follows:

Chapter 4 (Image classification) has focused on the problem of classifying images into one of predefined classes, based on the BOW model. The work presented in this chapter has addressed different issues related to improving the discriminative power of the BOW model. For natural scene images, colour information is an important characteristic of the image content which is normally ignored by the BOW model. Including colour information with the BOW is not an easy task. In this chapter, a new weighting approach has been proposed to integrate the colour information with the BOW model in a spatial pyramid layout. It is based on the densities of local keypoints at spatial pyramid layout. The spatial pyramid layout employed overcomes the problem of orderless nature inherited with the BOW model. Also, the chapter has addressed the influence of using visual vocabularies obtained from each scene category to build integrated bag of visual words (IBOW)

model. The framework proposed in Section 4.2 has employed all these issues to represent the semantic information of natural scene images. Also, different configurations of building the BOW model as well as some baseline methods for representing image content, such as GIST features, have been considered and used for comparisons. The baseline methods have shown lower performance than the approaches proposed by the author.

The experimental work carried out in Section 4.3.4 indicates the feasibility of our approaches in representing the semantic information of image content for natural scene classification task. This has been evaluated on three well-known natural scene datasets. The experimental work revealed that the results of using the proposed approaches are comparable to or better than the results reported in similar work in the literature as shown in Section 4.3.4. Moreover, many ideas have been investigated in this chapter. The GIST features have shown better classification performance when integrated with the pyramidal colour moments. Also, an experimental work has been done to study the influence of using visual vocabularies generated from one image dataset to build BOW histograms for another dataset from the same domain. The results of the experimental work have shown good classification results on a small dataset.

Chapter 5 (Image Annotation) has addressed the problem of annotating images with local semantic concepts at region level. The aim was to investigate the feasibility of using the BOW model to represent the visual content of image regions for the annotation task. Before employing the BOW model to represent image regions, a hypothesis has been introduced that assumed that there is a relationship between the distribution of local semantic concepts and local keypoints located in

image regions labelled with these semantic concepts. An in-depth analysis of both distributions has provided strong evidence in support of this hypothesis. It is concluded that BOW can be a good choice for representing the visual content of image regions.

Moreover, this chapter has investigated using visual vocabularies generated from the entire images to build BOW histograms for image regions. This was called local from global approach. Also, this chapter has investigated building visual vocabularies from image halves to improve the discriminative power of the BOW model. This adds more semantic information to the visual vocabularies which in turn affects the intermediate semantic representation in the BOW model. All BOW-based approaches as well as baseline methods have been extensively evaluated on 6-categories dataset of natural scenes using the SVM and KNN classifiers. The reported results have shown the plausibility of using the BOW model to represent the semantic information of image regions. From the experimental results reported in (Table 5.4), it can be concluded that Local from Global approach (presented in Section 5.3.2.2) is efficient to build BOW histograms for image regions at the upper and lower halves of images. The SVM classifiers performed better than the KNN classifier. Also, it is shown that the integrated bag of visual words outperformed the universal bag of visual words. This confirms the results obtained in the experimental work of Chapter 4.

Chapter 6 (Image Retrieval) has addressed the problem of semantic-based retrieval of natural scenes. This chapter has presented a comparative study between using the concept-occurrence vector, presented in Chapter 5, and the BOW-based

approaches, presented in Chapter 4, to represent the semantic information of images for natural scene retrieval.

The concept-occurrence vector has been used in this chapter to summarize the amount of local semantic concepts used to annotate regions of an image, i.e. it is a histogram of the local semantic concepts. Using 6-scene categories dataset, the BOW-based approaches have achieved good retrieval accuracy and outperformed baseline methods, such as colour histograms. The COV based on the use of BOW-based approaches have reported retrieval accuracy close to the COV benchmark. It is concluded that the BOW-based approaches, used in Chapter 5 to represent image regions, are also useful when the COV is employed to represent the semantic of the image content. An analysis of the distribution of the local semantic concepts, represented by the COVs benchmark, and the distribution of the local semantic concepts represented by BOW-based COVs, has been carried out. It showed how close the BOW-based COVs is to the COVs benchmark. However, the COV is only applicable when image regions are annotated with labels.

This chapter has also investigated the use of BOW-based approaches presented in Chapter 4 for semantic-based image retrieval. In contrast to the COV approach, the BOW-based approaches have emphasized on representing the semantic information of the image content using the UBOW, IBOW and the pyramidal integrated BOW fused with the pyramidal weighted colour moment approaches. No local concepts are employed with these approaches. The experimental results revealed that the BOW-based approaches perform worse than the COV approaches. However, the BOW-based approaches outperform baseline methods such as colour histogram and DWT. It is worth to mention that the COV can only be employed if

image regions are annotated with local semantic concepts, which is not the case when using BOW-based approaches.

7.2 Future Work

The following summarizes some ideas for the future work, as resulted from the contributions across the thesis; these future research topics are a natural continuation of the work presented hereby:-

- Generating visual vocabularies from each scene category has shown better performance than the universal visual vocabulary for semantic-based image representation. However, it would be interesting to refine each visual word of the vocabulary by defining criteria to choose an informative subset of the SIFT features allocated to this visual word. Calculating the mathematical mean of the chosen SIFT subset will result in a refined visual word.
- It would be interesting to research further into finding criteria to select a subset of visual words from the visual vocabulary to generate a smaller and more informative visual vocabulary which in turn should improve the discriminative power of the BOW word model. This point and the previous point have shown a great interest from many researchers. They are interested to build compact and more discriminative visual vocabularies (Elfiky et al., 2012, Su and Jurie, 2011, Shiliang et al., 2011, Ramanan and Niranjan, 2011).

- It would be interesting to investigate the use of multiple BOWs to represent the semantic information of the image content as shown in Figure 7-1.

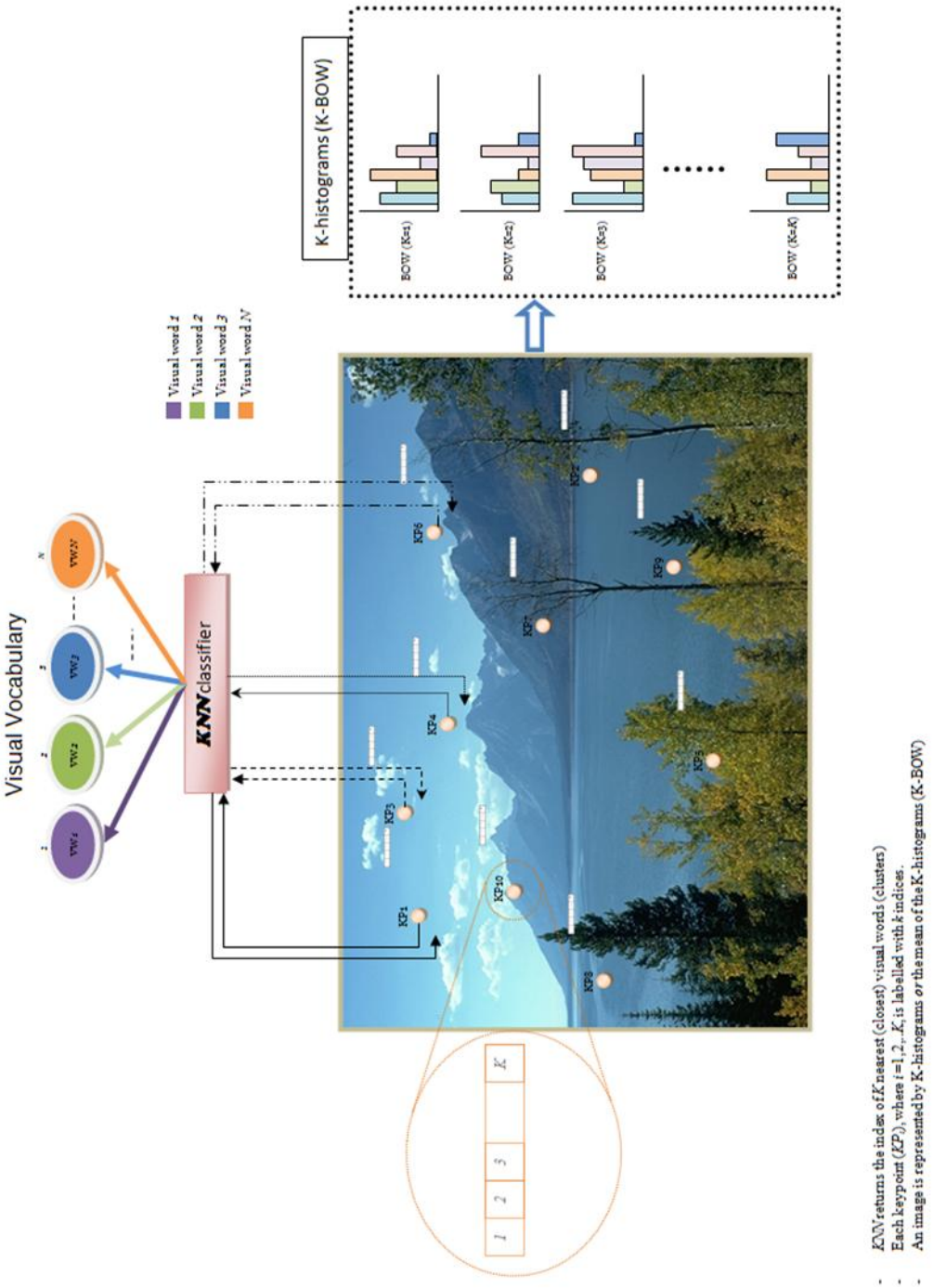


Figure 7-1: Proposed model for building multiple BOWs histogram to represent the semantic information of the image content.

Using a single BOW histogram enforce local keypoints, detected in the image, to be assigned to only one visual word ignoring other visual words which may have similarities to them. For example, fuzzy logic approach can be used to find similarities between descriptors of the keypoints and the visual words such that each keypoint with have a set of membership values to each visual word of the visual vocabulary. The visual vocabulary can also be generated using the fuzzy c-mean clustering approach.

- It would be interesting to extend the proposed approaches to work with datasets at the object level. There are many public datasets available in the literature, such as Caltech-101[4], Caltech-256[5] and the Pascal Visual Object Classes (VOC)[6], for object categorization, annotation and retrieval.

[4] http://www.vision.caltech.edu/Image_Datasets/Caltech101/
[5] http://www.vision.caltech.edu/Image_Datasets/Caltech256/
[6] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

References

ABBADENI, N. (2011) Computational Perceptual Features for Texture Representation and Retrieval. *IEEE Transactions on Image Processing,* 20**,** 236-246.

AKBAS, E. & AHUJA, N., 2010. Low-level Image Segmentation Based Scene Classification. 20th International Conference on Pattern Recognition, ICPR, Istanbul, 23-26 Aug, 3623-3626

AKSOY, S., KOPERSKI, K., TUSK, C., MARCHISIO, G. & TILTON, J. C. (2005) Learning Bayesian classifiers for scene classification with a visual grammar. *IEEE Transactions on Geoscience and Remote Sensing,* 43**,** 581-589.

ALQASRAWI, Y., NEAGU, D. & COWLING, P., 2009. Natural Scene Image Recognition by Fusing Weighted Colour Moments with Bag of Visual Patches on Spatial Pyramid Layout. 9th International Conference on Intelligent Systems Design and Applications, ISDA, Pisa, Italy, 30 November-2 December, 140-145

ALQASRAWI, Y., NEAGU, D. & COWLING, P., 2010. Spatial pyramid local keypoints quantization for bag of visual patches image representation. The10th International Conference on Intelligent Systems Design and Applications (ISDA), 1270-1274

ALQASRAWI, Y., NEAGU, D. & COWLING, P. I. (2011) Fusing integrated visual vocabularies-based bag of visual words and weighted colour moments on spatial pyramid layout for natural scene image classification. *Signal, Image and Video Processing***,** 1-17.

BACH, J. R., FULLER, C., GUPTA, A., HAMPAPUR, A., HOROWITZ, B., HUMPHREY, R., JAIN, R. C. & SHU, C. F., 1996. Virage image search engine: an open framework for image management. 76

BAEZA-YATES, R. & RIBEIRO-NETO, B. (1999) *Modern Information Retrieval*, ACM Press.

BARNARD, K. & FORSYTH, D., 2001. Learning the semantics of words and pictures. Eighth IEEE International Conference on Computer Vision, ICCV, 408-415 vol. 2

BATTIATO, S., FARINELLA, G., GALLO, G. & RAVI, D. (2009) Spatial Hierarchy of Textons Distributions for Scene Classification. *Proc. Eurocom Multimedia Modeling***,** 333-342.

BATTIATO, S., FARINELLA, G., GALLO, G. & RAVÌ, D., 2008. Scene categorization using bag of textons on spatial hierarchy. 15th IEEE International Conference on Image Processing, 2008, ICIP 12–15

BATTIATO, S., FARINELLA, G. M., GALLO, G. & RAVI, D. (2010a) Exploiting Textons Distributions on Spatial Hierarchy for Scene Classification. *EURASIP Journal on Image and Video Processing,* 2010**,** 1-13.

BATTIATO, S., FARINELLA, G. M., GUARNERA, G. C., MECCIO, T., PUGLISI, G., RAVI, D. & RIZZO, R. (2010b) Bags of phrases with

codebooks alignment for near duplicate image detection. *2nd ACM workshop on Multimedia in forensics, security and intelligence.* Firenze, Italy.

BAY, H., TUYTELAARS, T. & VAN GOOL, L. (2006) Surf: Speeded up robust features. *ninth European Conference on Computer Vision, ECCV***,** 404-417.

BOSCH, A., MUNOZ, X. & MARTÍ, R. (2007a) Which is the best way to organize/classify images by content? *Image and vision computing,* 25**,** 778-791.

BOSCH, A., MUNOZ, X., OLIVER, A. & MARTI, R., 2006. Object and scene classification: what does a supervised approach provide us. 18th International Conference on Pattern Recognition, ICPR, Hong Kong, China, 20-24 August, 773-777

BOSCH, A., ZISSERMAN, A. & MUNOZ, X., 2007b. Representing shape with a spatial pyramid kernel. 6th ACM International Conference on Image and Video Retrieval, CIVR, Amsterdam, The Netherlands, 9-11 July, 401-408

BOSCH, A., ZISSERMAN, A. & MUOZ, X. (2008) Scene Classification Using a Hybrid Generative/Discriminative Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30**,** 712-727.

BOUTELL, M. R., LUO, J., SHEN, X. & BROWN, C. M. (2004) Learning multi-label scene classification. *Pattern Recognition,* 37**,** 1757-1771.

CARSON, C., THOMAS, M., BELONGIE, S., HELLERSTEIN, J. & MALIK, J., 1999. Blobworld: A system for region-based image indexing and retrieval. 3rd International Conference on Visual Information and Information Systems,

CHANG, C.-C. & LIN, C.-J. (2011) LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST),* 2**,** 1-27. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm

CHAPELLE, O., HAFFNER, P. & VAPNIK, V. N. (1999) Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks,* 10**,** 1055-1064.

CHEN, X., HU, X. & SHEN, X., 2009. Spatial Weighting for Bag-of-Visual-Words and Its Application in Content-Based Image Retrieval. Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Bangkok, Thailand, 867-874

CHEN, Y. & WANG, J. Z. (2002) A region-based fuzzy feature matching approach to content-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 24**,** 1252-1267.

CHEN, Y., WANG, J. Z. & KROVETZ, R. (2005) Clue: Cluster-based retrieval of images by unsupervised learning. *IEEE Transactions on Image Processing,* 14**,** 1187-1201.

CHENG, H. & WANG, R. (2010) Semantic modeling of natural scenes based on contextual Bayesian networks. *Pattern Recognition,* 43**,** 4042-4054.

CHIMLEK, S., KESORN, K., PIAMSA-NGA, P. & POSLAD, S., 2010. Semantically similar visual words discovery to facilitate visual invariance. 2010 IEEE International Conference on Multimedia and Expo (ICME), Singapore, 19-23 July, 1242-1247

CSILLAGHY, A., HINTERBERGER, H. & BENZ, A. (2000) Content-based image retrieval in astronomy. *Information Retrieval,* 3**,** 229-241.

CSURKA, G., DANCE, C., FAN, L., WILLAMOWSKI, J. & BRAY, C., 2004. Visual categorization with bags of keypoints. ECCV Workshop on Statistical Learning in Computer Vision, Czech Republic, 11-14 May, 59-74

CUSANO, C., CIOCCA, G. & SCHETTINI, R., 2004. Image annotation using SVM. Proceedings of Internet imaging IV, SPIE, 330-338

DALAL, N. (2006) Finding People in Images and Videos, PhD thesis, INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE,

DALAL, N. & TRIGGS, B., 2005. Histograms of oriented gradients for human detection. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 886-893

DATTA, R., JOSHI, D., LI, J. & WANG, J. (2008) Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40**,** 1-60.

DATTA, R., LI, J. & WANG, J. Z., 2005. Content-based image retrieval: approaches and trends of the new age. Proceedings of the 7th International Workshop on Multimedia Information Retrieval, in conjunction with ACM International Conference on Multimedia, Singapore, November 253-262

DAUBECHIES, I. (1990) The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36**,** 961-1005.

DAUGMAN, J. G. (1988) Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36**,** 1169-1179.

DENG, Y. & MANJUNATH, B. (2001) Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23**,** 800-810.

DESELAERS, T., KEYSERS, D. & NEY, H. (2008) Features for image retrieval: an experimental comparison. *Information Retrieval*, 11**,** 77-107.

DOWE, J., 1993. Content-based retrieval in multimedia imaging. In Proceeding of SPEI Storage and Retrieval for Image and Video Databases,

DUYGULU, P., BARNARD, K., FREITAS, J. F. G. D. & FORSYTH, D. A., 2002. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. Proceedings of the 7th European Conference on Computer Vision-Part IV, ECCV, 97-112

EAKINS, J. P. (2002) Towards intelligent image retrieval. *Pattern Recognition*, 35**,** 3-14.

ELFIKY, N. M., SHAHBAZ KHAN, F., VAN DE WEIJER, J. & GONZALEZ, J. (2012) Discriminative compact pyramids for object and scene recognition. *Pattern Recognition*, 45**,** 1627–1636.

ELSAYAD, I., MARTINET, J., URRUTY, T. & DJERABA, C., 2010. A new spatial weighting scheme for bag-of-visual-words. 2010 International Workshop on Content-Based Multimedia Indexing (CBMI), Grenoble , France 23-25 June 1-6

FALOUTSOS, C. & TAUBIN, G., 1993. The QBIC project: Querying images by content using color, texture, and shape. In Proceeding of SPIE Storage and retrieval for Image and Video Databases, 173-187

FARINELLA, G. M. & BATTIATO, S. (2010) representation models and machine learning techniques for scene classification. IN WANG, P. S. P. (Ed.) *Pattern Recognition, Machine Vision, Principles and Applications.* Denmark, River publisher.199-214

FAUZI, M. F. A. (2004) Content-Based Image Retrieval of Museum Images, PhD thesis, UNIVERSITY OF SOUTHAMPTON,

FEI-FEI, L. & PERONA, P., 2005. A bayesian hierarchical model for learning natural scene categories. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, San Diego, CA, USA, 20-26 June, 524-531

FERGUS, R., FEI-FEI, L., PERONA, P. & ZISSERMAN, A., 2005. Learning object categories from Google's image search. IEEE International Conference on Computer Vision , ICCV Beijing 1816-1823

GANGOPADHYAY, A. (2001) An image-based system for electronic retailing. *Decision Support Systems,* 32**,** 107-116.

GEMERT, J. C. V., SNOEK, C. G. M., VEENMAN, C. J., SMEULDERS, A. W. M. & GEUSEBROEK, J.-M. (2010) Comparing compact codebooks for visual categorization. *Comput. Vis. Image Underst.,* 114**,** 450-462.

GHOSHAL, A., IRCING, P. & KHUDANPUR, S., 2005. Hidden Markov models for automatic annotation and content-based retrieval of images and video. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil, 544-551

GIULIODORI, M. A. (2011) Statistical classification of images, PhD theses, Universidad Carlos III de Madrid, 118

GOH, K. S., CHANG, E. Y. & LI, B. (2005) Using one-class and two-class SVMs for multiclass image annotation. *IEEE Transactions on Knowledge and Data Engineering,* 17**,** 1333-1346.

GOKALP, D. & AKSOY, S., 2007. Scene classification using bag-of-regions representations. IEEE Conference on Computer Vision and Pattern Recognition, Beyond Patches Workshop, CVPR, Minneapolis, Minnesota, USA, 18-23 June, 1-8

GU, G., ZHAO, Y. & ZHU, Z. (2011) Integrated image representation based natural scene classification. *Expert Systems with Applications,* 38**,** 11273-11279.

GUNN, S. R. (1998) Support vector machines for classification and regression. *ISIS technical report.*

HARALICK, R. M., SHANMUGAM, K. & DINSTEIN, I. H. (1973) Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics,* 3**,** 610-621.

HARRIS, C. & STEPHENS, M., 1988. A combined corner and edge detector. Proceedings of the Alvey Vision Conference, 147–151

HOU, J., FENG, Z.-S., YANG, Y. & QI, N.-M. (2011) Towards a Universal and Limited Visual Vocabulary. *Advances in Visual Computing.* Springer Berlin.398-407

HUANG, J., KUMAR, S. R., MITRA, M., ZHU, W. J. & ZABIH, R., 1997a. Image indexing using color correlograms. IEEE International Conference on Computer Vision and Pattern Recognition, CVPR, Puerto Rico, June, 762-768

HUANG, T., MEHROTRA, S. & RAMCHANDRAN, K., 1997b. Multimedia analysis and retrieval system (MARS) project. In Proceeding of 33rd Annual Clinic on Library Application of Data Processing-Digital Image Access and Retrieval,

IQBAL, Q. & AGGARWAL, J., 2002. CIRES: A system for content-based retrieval in digital image libraries. 7th International Conference on Control, Automation, Robotics and Vision,ICARCV
205-210

JA-HWUNG, S., CHIEN-LI, C., CHING-YUNG, L. & TSENG, V. S. (2011) Effective Semantic Annotation by Image-to-Concept Distribution Model. *IEEE Transactions on Multimedia,* 13**,** 530-538.

JIANG, Y., NGO, C. & YANG, J., 2007. Towards optimal bag-of-features for object categorization and semantic video retrieval. 6th ACM International Conference on Image and Video Retrieval, CIVR, Amsterdam, The Netherlands, 9-11 July, 494-501

JIANG, Y., YANG, J., NGO, C. & HAUPTMANN, A. (2010) Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transaction on Multimedia,* 12**,** 42-53.

JIEBO, L. & SAVAKIS, A., 2001. Indoor vs outdoor classification of consumer photographs using low-level and semantic features. 2001 International Conference on Image Processing, 7-10 Oct, 745-748 vol.2

JINGYAN, W., YONGPING, L., YING, Z., HONGLAN, X. & CHAO, W., 2011. Bag-of-Features Based Classification of Breast Parenchymal Tissue in the Mammogram via Jointly Selecting and Weighting Visual Words. 2011 Sixth International Conference on Image and Graphics (ICIG), Hefei, Anhui, China, 12-15 Aug. 2011, 622-627

JOACHIMS, T., 1998. Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the European Conference on Machine Learning, ECML,* 137-142

JURIE, F. & TRIGGS, B., 2005. Creating efficient codebooks for visual recognition. IEEE International Conference on Computer Vision, Beijing, China, 17-20 October, 604-610

KADIR, T. & BRADY, M. (2001) Saliency, scale and image description. *International Journal of Computer Vision,* 45**,** 83-105.

KESORN, K., CHIMLEK, S., POSLAD, S. & PIAMSA-NGA, P. (2011) Visual content representation using semantically similar visual words. *Expert Systems with Applications,* 38**,** 11472-11481.

KESORN, K. & POSLAD, S. (2011) An Enhanced Bag of Visual Word Vector Space Model to Represent Visual Content in Athletics Images. *IEEE Transactions on Multimedia,* PP**,** 1-1.

KHAN, F., VAN DE WEIJER, J. & VANRELL, M., 2009. Top-Down Color Attention for Object Recognition. 12th IEEE International Conference on Computer Vision, ICCV, Kyoto, Japan, 27 September- 4 October, 979-986

KHAN, F., VANÂ DE WEIJER, J. & VANRELL, M. (2011) Modulating Shape Features by Color Attention for Object Recognition. *International Journal of Computer Vision***,** 1-16.

KOKARE, M., BISWAS, P. K. & CHATTERJI, B. N. (2007) Texture image retrieval using rotated wavelet filters. *Pattern Recognition Letters,* 28**,** 1240-1249.

LAMPERT, C., BLASCHKO, M., HOFMANN, T. & ZURICH, S., 2008. Beyond sliding windows: Object localization by efficient subwindow search. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Alaska, USA, 24-26 June, 1-8

LAZEBNIK, S., SCHMID, C. & PONCE, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, New York, USA, 17-22 June, 2169-2178

LEI, W., HOI, S. C. H. & NENGHAI, Y. (2010) Semantics-Preserving Bag-of-Words Models and Applications. *IEEE Transactions on Image Processing* 19**,** 1908-1920.

LI, J., WANG, J. Z. & WIEDERHOLD, G., 2000. IRM: integrated region matching for image retrieval. Proceedings of the eighth ACM international conference on Multimedia, Los Angeles, CA, USA, October 30 - November 03, 147-156

LIU, S., XU, D. & FENG, S. (2011) Region Contextual Visual Words for scene categorization. *Expert Systems with Applications,* 38**,** 11591-11597. doi: 10.1016/j.eswa.2011.03.037

LIU, Y., ZHANG, D. & LU, G. (2008) Region-based image retrieval with high-level semantics using decision tree learning. *Pattern Recognition,* 41**,** 2554-2570.

LIU, Y., ZHANG, D., LU, G. & MA, W. (2007) A survey of content-based image retrieval with high-level semantics. *Pattern Recognition,* 40**,** 262-282.

LONG, F., ZHANG, H. & FENG, D. D. (2003) Fundamentals of content-based image retrieval. *Multimedia Information Retrieval and Management.*

LOPEZ-SASTRE, R., TUYTELAARS, T., ACEVEDO-RODRIGUEZ, F. & MALDONADO-BASCON, S. (2011) Towards a more discriminative and semantic visual vocabulary. *Computer Vision and Image Understanding,* 115**,** 415-425.

LOWE, D. (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision,* 60**,** 91-110.

LUO, J., BOUTELL, M. & BROWN, C. (2006) Pictures are not taken in a vacuum-an overview of exploiting context for semantic scene content understanding. *Signal Processing Magazine, IEEE,* 23**,** 101-114.

LUO, J. & SAVAKIS, A., 2001. Indoor vs outdoor classification of consumer photographs using low-level and semantic features. IEEE International Conference on Image Processing, ICIP, 745-748 vol. 2

MA, W. Y. & MANJUNATH, B., 1997. Netra: A toolbox for navigating large image databases. IEEE International Conference in Image Processing, ICIP, 568-571

MAKADIA, A., PAVLOVIC, V. & KUMAR, S. (2010) Baselines for image annotation. *International Journal of Computer Vision,* 90**,** 88-105.

MANJUNATH, B. S., OHM, J. R., VASUDEVAN, V. V. & YAMADA, A. (2001) Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology,* 11**,** 703-715.

MAO, J. & JAIN, A. K. (1992) Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition,* 25**,** 173-188.

MAREE, R., GEURTS, P., PIATER, J. & WEHENKEL, L., 2005a. Random subwindows for robust image classification. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 34-40

MAREE, R., GEURTS, P., PIATER, J. & WEHENKEL, L., 2005b. Random subwindows for robust image classification. International Conference on Computer Vision and Pattern Recognition, CVPR, 34-40

MIKOLAJCZYK, K. (2011), Binaries for Interest point detectors and descriptors, *last accessed*: June 2011, <http://lear.inrialpes.fr/people/mikolajczyk/>

MIKOLAJCZYK, K. & SCHMID, C. (2005) A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 27**,** 1615-1630.

MOJSILOVIC , A., GOMES, J. & ROGOWITZ, B. (2004) Semantic-friendly indexing and quering of images based on the extraction of the objective semantic cues. *International Journal of Computer Vision,* 56**,** 79-107.

MORI, Y., TAKAHASHI, H. & OKA, R., 1999. Image-to-word transformation based on dividing and vector quantizing images with words. First International Workshop on Multimedia Intelligent Storage and Retrieval Management,

MOULIN, C., BARAT, C. & DUCOTTET, C., 2010. Fusion of tf.idf weighted bag of visual features for image classification. 2010 International Workshop on Content-Based Multimedia Indexing (CBMI), Grenoble , France, 23-25 June 1-6

MÜLLER, H., MICHOUX, N., BANDON, D. & GEISSBUHLER, A. (2004) A review of content-based image retrieval systems in medical applications—clinical benefits and future directions *International Journal of Medical Informatics,* 73**,** 1-23.

MULLER, H., MULLER, W., SQUIRE, D. M. G., MARCHAND-MAILLET, S. & PUN, T. (2001) Performance evaluation in content-based image retrieval: Overview and proposals. *Pattern Recognition Letters,* 22**,** 593-601.

NILSBACK, M. & ZISSERMAN, A., 2006. A visual vocabulary for flower classification. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, New York, USA, 17-22 June, 1447-1454

NILSBACK, M. E. & ZISSERMAN, A., 2008. Automated flower classification over a large number of classes. Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP '08. , 722-729

NISTER, D. & STEWENIUS, H., 2006. Scalable recognition with a vocabulary tree. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, New York, USA, 17-22 June, 2161-2168

ODONE, F., BARLA, A. & VERRI, A. (2005) Building kernels from binary strings for image matching. *IEEE Transactions on Image Processing,* 14**,** 169-180.

OJALA, T., PIETIKÃ¤INEN, M. & HARWOOD, D. (1996) A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition,* 29**,** 51-59.

OLIVA, A. & TORRALBA, A. (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision,* 42**,** 145-175.

PARKER, J. J. S. (2004) Commercial digital image libraries, digital images and digital discontent. *International Journal on Digital Libraries,* 4**,** 124-136.

PASS, G. & ZABIH, R., 1996. Histogram refinement for content-based image retrieval. Proceedings 3rd IEEE Workshop on Applications of Computer Vision, 96-102

PENATTI, O. V. A. B., VALLE, E. & TORRES, R. D. S. (2012) Comparative study of global color and texture descriptors for web image retrieval. *Journal of visual communication and image representation,* 23**,** 359-380.

PENTLAND, A., PICARD, R. W. & SCLAROFF, S. (1996) Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision,* 18**,** 233-254.

PERINA, A., CRISTANI, M. & MURINO, V. (2010) Learning natural scene categories by selective multi-scale feature extraction. *Image and vision computing,* 28**,** 927-939.

PERRONNIN, F. (2008) Universal and adapted vocabularies for generic visual categorization. *IEEE Transactions on pattern analysis and machine intelligence,* 30**,** 1243-1256.

PHILBIN, J., CHUM, O., ISARD, M., SIVIC, J. & ZISSERMAN, A., 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Anchorage, AK, 23-28 June, 1-8

QI, X. & HAN, Y. (2007) Incorporating multiple SVMs for automatic image annotation. *Pattern Recognition,* 40**,** 728-741.

QIN, J. & YUNG, N. H. C. (2010) Scene categorization via contextual visual words. *Pattern Recognition,* 43**,** 1874-1888.

QUELHAS, P. (2007) Scene Image Classification and Segmentation with Quantized Local Descriptors and Latent Aspect Modeling, PhD thesis, Ecole Polythechnique F´ed´erale de Lausanne (EPFL), Lausanne, Switzerland,

QUELHAS, P., MONAY, F., ODOBEZ, J., GATICA-PEREZ, D. & TUYTELAARS, T. (2007) A thousand words in a scene. *IEEE Transactions on pattern analysis and machine intelligence,* 29**,** 1575-1589.

QUELHAS, P., MONAY, F., ODOBEZ, J., GATICA-PEREZ, D., TUYTELAARS, T. & VAN GOOL, L., 2005. Modeling scenes with local descriptors and latent aspects. IEEE International Conference on Computer Vision ICCV, Beijing, China, 17-21 October, 883–890

QUELHAS, P. & ODOBEZ, J., 2006. Natural scene image modeling using color and texture visterms. International Conference on Image and Video Retrieval, CIVR, Tempe, AZ, USA, 13-15 July 411-421

QUELHAS, P. & ODOBEZ, J., 2007. Multi-level local descriptor quantization for bag-of-visterms image representation. Then 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands 9-11 July, 242-249

RAMANAN, A. & NIRANJAN, M. (2011) A Review of Codebook Models in Patch-Based Visual Object Recognition. *Journal of Signal Processing Systems***,** 1-20.

RASIWASIA, N. (2011) Semantic Image Representation for Visual Recognition, PhD thesis, UNIVERSITY OF CALIFORNIA, SAN DIEGO, 232

RIES, C. X., ROMBERG, S. & LIENHART, R., 2010. Towards universal visual vocabularies. 2010 IEEE International Conference on Multimedia and Expo (ICME), Augsburg, Germany 19-23 July, 1067-1072

ROSS, M. & OLIVA, A. (2010) Estimating perception of scene layout properties from global image features. *Journal of Vision,* 10**,** 1-25.

RUI, Y., HUANG, T. & CHANG, S. (1999) Image retrieval: Current techniques, promising directions, and open issues. *Journal of visual communication and image representation,* 10**,** 39-62.

RUSSELL, B. C., TORRALBA, A., MURPHY, K. P. & FREEMAN, W. T. (2008) LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision,* 77**,** 157-173.

SALTON, G. (1968) *Automatic Information Organization and Retrieval,* McGraw-Hill.

SALTON, G. (1983) *Introduction to Modern Information Retrieval,* McGraw-Hill, New York.

SERRANO, N., SAVAKIS, A. E. & LUO, J. (2004) Improved scene classification using efficient low-level features and semantic cues. *Pattern Recognition,* 37**,** 1773-1784.

SHAHIDUZZAMAN, M., ZHANG, D. & LU, G. (2011) Improved spatial pyramid matching for image classification. *Computer Vision–ACCV 2010.* LNCS,Springer.449-459

SHENG, X., TAO, F., DEREN, L. & SHIWEI, W. (2010) Object Classification of Aerial Images With Bag-of-Visual Words. *Geoscience and Remote Sensing Letters, IEEE,* 7**,** 366-370.

SHI, J. & MALIK, J. (2000) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI,* 22**,** 888-905.

SHILIANG, Z., QI, T., GANG, H., QINGMING, H. & WEN, G. (2011) Generating Descriptive Visual Words and Visual Phrases for Large-Scale Image Applications. *IEEE Transactions on Image Processing,* 20**,** 2664-2677.

SINGH, M. & CUNNINGHAM, P. (2008) Active Learning for Image Analysis. Dublin, University College Dublin,.

SIVIC, J. & ZISSERMAN, A., 2003. Video Google: A text retrieval approach to object matching in videos. 9th IEEE International Conference on Computer Vision, ICCV, Nice, France, 14-17 October, 1470–1477

SMEULDERS, A. W. M., WORRING, M., SANTINI, S., GUPTA, A. & JAIN, R. (2000) Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22**,** 1349-1380.

SMITH, J. R. & CHANG, S. (1997a) Querying by color regions using the VisualSEEk content-based visual query system. *Intelligent multimedia information retrieval,* 7**,** 23-41.

SMITH, J. R. & CHANG, S. F. (1997b) Visually searching the web for content. *Multimedia, IEEE,* 4**,** 12-20.

SPYROU, E., MYLONAS, P. & AVRITHIS, Y., 2008. Using region semantics and visual context for scene classification. 15th IEEE International Conference on Image Processing, ICIP, San Diego, CA, USA, 12-15 Oct. 2008 53-56

STRICKER, M. & ORENGO, M., 1995. Similarity of color images. Proceeding of SPIE Storage and Retrieval for Image and Video Databases, 381-392

SU, Y. & JURIE, F., 2011. Visual word disambiguation by semantic contexts. IIEEE International Conference on Computer Vision, ICCV, 6-13 Nov, 311-318

SUN, Y. & OZAWA, S., 2003. Semantic-meaningful content-based image retrieval in wavelet domain. Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval, 122-129

SWAIN, M. & BALLARD, D. (1991) Color indexing. *International Journal of Computer Vision,* 7**,** 11-32.

SZUMMER, M. & PICARD, R., 1998. Indoor-outdoor image classification. IEEE International Workshop on Content-Based Access of Image and Video Databases, CAIVD, Bombay, India, January, 42-51

TAMURA, H., MORI, S. & YAMAWAKI, T. (1978) Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics,* 8**,** 460-473.

TIRILLY, P., CLAVEAU, V. & GROS, P., 2010. Distances and weighting schemes for bag of visual words image retrieval. Proceedings of the international conference on Multimedia information retrieval, Philadelphia, Pennsylvania, USA, 29-31 March, 323-332

TOUSCH, A.-M., HERBIN, S. & AUDIBERT, J.-Y. (2012) Semantic hierarchies for image annotation: A survey. *Pattern Recognition,* 45**,** 333-345.

TSAI, C. F., MCGARRY, K. & TAIT, J. (2006) CLAIRE: A modular support vector image indexing and classification system. *ACM Transactions on Information Systems (TOIS),* 24**,** 353-379.

TURNER, M. R. (1986) Texture discrimination by Gabor functions. *Biological Cybernetics,* 55**,** 71-82.

TUYTELAARS, T. & MIKOLAJCZYK, K. (2008) Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision,* 3**,** 177-280.

VAILAYA, A., FIGUEIREDO, M., JAIN, A. & ZHANG, H. (2001) Image classification for content-based indexing. *IEEE Transactions on Image Processing,* 10**,** 117-130.

VAILAYA, A., JAIN, A. & ZHANG, H. J. (1998) On image classification: City images vs. landscapes. *Pattern Recognition,* 31**,** 1921-1935.

VAN DE SANDE, K. E. A., GEVERS, T. & SNOEK, C. G. M. (2010) Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 32**,** 1582-1596.

VAN KAICK, O. & MORI, G., 2006. Automatic classification of outdoor images by region matching. The 3rd Canadian Conference on Computer and Robot Vision, 7-9 June, 9-9

VAPNIK, V., GOLOWICH, S. E. & SMOLA, A., 1996. Support vector method for function approximation, regression estimation, and signal processing. Advances in Neural Information Processing Systems, NIPS, 281-287

VAPNIK, V. N. (2000) *The nature of statistical learning theory*, Springer Verlag.

VIDAL-NAQUET, M. & ULLMAN, S., 2003. Object recognition with informative features and linear classification. Ninth IEEE International Conference on Computer Vision, ICCV, Washington, DC, USA, 13-16 Oct, 281-288

VIEUX, R., BENOIS-PINEAU, J. & DOMENGER, J.-P. (2012) Content Based Image Retrieval Using Bag-Of-Regions *Advances in Multimedia Modeling, LNCS.* Springer.507-517

VIGO, D. A. R., KHAN, F. S., VAN DE WEIJER, J. & GEVERS, T., 2010. The Impact of Color on Bag-of-Words Based Object Recognition. 20th International Conference on Pattern Recognition (ICPR), Barcelona, Spain, 23-26 Aug., 1549-1553

VOGEL, J. (2004) Semantic Scene Modeling and Retrieval, PhD thesis, SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH, 158

VOGEL, J. & SCHIELE, B. (2004) A semantic typicality measure for natural scene categorization. *Lecture notes in computer science.* Springer.195-203

VOGEL, J. & SCHIELE, B. (2007) Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision,* 72**,** 133-157.

VOGEL, J., SCHWANINGER, A., WALLRAVEN, C. & BÜLTHOFF, H. (2007) Categorization of natural scenes: Local versus global information and the role of color. *ACM Transactions on Applied Perception,* 4**,** Article 19.

WANG, J., LI, J. & WIEDERHOLD, G. (2001) SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 23**,** 947-963.

WANG, Y., LIU, X. & JIA, Y. (2010) Automatic Image Annotation with Cooperation of Concept-Specific and Universal Visual Vocabularies. *Advances in Multimedia Modeling.* Lecture Notes in Computer Science 5916 Springer.262-272

WITKIN, A. P. (1987) Scale-space filtering. *Readings in computer vision: issues, problems, principles, and paradigms***,** 329-332.

WITTEN, I. H. & FRANK, E. (2005) *Data Mining: Practical machine learning tools and techniques,* Morgan Kaufmann.

WU, J. & REHG, J., 2009. Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. 12th IEEE International Conference on Computer Vision, ICCV, Kyoto, Japan, 27 September-4 October, 630-637

WU, L., HOI, S. C. H. & YU, N., 2009a. Semantics-preserving bag-of-words models for efficient image annotation. Proceedings of the First ACM workshop on Large-scale multimedia retrieval and mining, 19-26

WU, Z., KE, Q., SUN, J. & SHUM, H.-Y., 2009b. A Multi-Sample, Multi-Tree Approach to Bag-of-Words Image Representation for Image Retrieval. 12th IEEE International Conference on Computer Vision, ICCV, Kyoto, Japan, 127 September- 4 October, 1992-1999

YANG, J., JIANG, Y., HAUPTMANN, A. & NGO, C., 2007. Evaluating bag-of-visual-words representations in scene classification. 9th ACM International Workshop on Multimedia Information Retrieval, MIR, University of Augsburg, Germany, 28-29 September, 197-206

YANG, Y. & NEWSAM, S., 2011. Spatial Pyramid Co-occurrence for Image Classification. IEEE International Conference on Computer Vision, ICCV, 6-13 Nov, 1465-1472

YU, H., LI, M., ZHANG, H. J. & FENG, J., 2002. Color texture moments for content-based image retrieval. IEEE International Conference on Image Processing, ICIP, Beijing, China, 24-28 June, 929-932

YUAN, L. & XIAOCHUN, C., 2011. Visual Word Pairs for Similar Image Search. 2011 Sixth International Conference on Image and Graphics (ICIG), Hefei, Anhui, China, 12-15 Aug. 2011, 987-992

ZHANG, C., LIU, J., OUYANG, Y., LU, H. & MA, S. (2009a) Concept-Specific Visual Vocabulary Construction for Object Categorization. *Advances in Multimedia Information Processing-PCM 2009.* Springer Berlin.936-942

ZHANG, C., LIU, J., OUYANG, Y., TIAN, Q., LU, H. & MA, S., 2009b. Category sensitive codebook construction for object category recognition. 16th IEEE International Conference on Image Processing, ICIP, Cairo, 7-10 Nov., 329-332

ZHANG, D., ISLAM, M. M. & LU, G. (2012) A review on automatic image annotation techniques. *Pattern Recognition,* 45**,** 346-362.

ZHOU, X., ZHU, C.-Z., SATOH, S. I. & GUO, Y.-T., 2011. Efficient quantization of color sift for image classification. IEEE International Conference on Image Processing, ICIP, Brussels, Belgium 1073-1076

ZHU, L., RAO, A. B. & ZHANG, A. (2002) Theory of keyblock-based image retrieval. *ACM Transactions on Information Systems (TOIS),* 20**,** 224-257.