Improving Bags-of-Words Model

For

Object Categorization

A thesis

submitted in fulfilment

of the requirements for the degree

of

Doctor of Philosophy

at

The University of Waikato

by

Edmond Yiwen Zhang

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2013

Abstract

In the past decade, Bags-of-Words (BOW) models have become popular for the task of object recognition, owing to their good performance and simplicity. Some of the most effective recent methods for computer-based object recognition work by detecting and extracting local image features, before quantizing them according to a codebook rule such as k-means clustering, and classifying these with conventional classifiers such as Support Vector Machines and Naive Bayes.

In this thesis, a Spatial Object Recognition Framework is presented that consists of the four main contributions of the research.

The first contribution, frequent keypoint pattern discovery, works by combining pairs and triplets of frequent keypoints in order to discover intermediate representations for object classes. Based on the same frequent keypoints principle, algorithms for locating the region-of-interest in training images is then discussed.

Extensions to the successful Spatial Pyramid Matching scheme, in order to better capture spatial relationships, are then proposed. The pairs frequency histogram and shapes frequency histogram work by capturing more refined spatial information between local image features.

Finally, alternative techniques to Spatial Pyramid Matching for capturing spatial information are presented. The proposed techniques, variations of binned log-polar histograms, divides the image into grids of different scale and different orientation. Thus captures the distribution of image features both in distance and orientation explicitly.

Evaluations on the framework are focused on several recent and popular datasets, including image retrieval, object recognition, and object categorization. Overall, while the effectiveness of the framework is limited in some of the datasets, the proposed contributions are nevertheless powerful improvements of the BOW model.

Acknowledgements

First and foremost, I would like to offer my sincerest gratitude to my supervisors, Michael Mayo and Bernhard Pfahringar, because without them, this PhD research would not have been possible. I want to thank Michael for supported me throughout my thesis with his patience and knowledge whilst allowing me to do my own research. I appreciate the fact that you have always kept your doors open to my questions and problems, and always happy to help whenever and wherever you can. Both Diana and I look forward to visit your family in the future.

Second, I would like to thank The University of Waikato and especially the Computer Science Department, which essentially was my home for a good deal of the last decade. I would not be able to carry out my research without the financial support from them. I am forever indebted to all of the lecturers and tutors that have taught me over the years, you opened my eyes to a world of possibilities.

Third, I would like to thank my wife, Diana. Thank you for accompanying me through this journey and putting up with me when things were not going well. You have never doubted me in completing this thesis and have always encouraged me to challenge myself. Without your support, and more than your share around the house as I sat at the computer, this thesis would not have been possible. I dedicate this thesis to you.

Fourth, I must thank my family, Mum, Dad, Daniel, and Aaron. You have always been there for me and words simply cannot express my gratefulness. Thanks to my parents who have always provided me with everything I needed for my education.

Finally, I need to thank my friends that are always been there for me. Especially Luke Hogan, Leo Li, Tony Li, Gareth Judson, Matt Jervis, Albert Bifet, Sam Bartels, and Jesse Read. I will forever treasure your friendships.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Computer vision is the study of designing machines that can see and interact with the world through visual information. Over the past 50 years, as technology has developed, more and more applications of computer vision became integrated into our lives. These applications vary greatly in terms of complexity, from simple tasks such as searching an image database, to more difficult robotics tasks such as autonomously navigating a car through the desert.

Although these applications are designed for numerous different tasks, they share many similarities. For many of them, the foundation is about extracting abstract information of what is happening in the scene and making sense of it. Indeed, in computer vision research, the ultimate goal is to develop algorithms and tools that will allow a computer to analyze the visual world automatically. Human understanding of images comes in very abstract terms, while the raw data received from imaging sensors are concrete, quantitative measurements of light. This challenge of image understanding has been and still is a shared motivation in the field of computer vision since the beginning.

This thesis investigates traditional and modern approaches to object recognition and develops new algorithms that improve on the previous work. The proposed algorithms were tested on diverse and challenging datasets, and the results improved the state-of-the-art in some cases.

1.1 Motivation

Over the past few decades, tremendous progress has been made in the field of object recognition with computer algorithms. One of the primary contributors to this progress has been the widespread use of machine learning techniques to make sense of large quantities of noisy data. Object recognition is a difficult computational problem because there are two conflicting requirements. The first requirement is that the recognition system needs to be selective and specific for different objects, as different objects may look similar and share certain features. The second requirement is that the recognition system needs to be invariant or tolerant to various transformations of an object, as often objects belonging to the same category differ immensely. Figure 1.1 illustrates different object appearances belonging to the same category.

The fundamental problem addressed in this thesis is how to improve the accuracy of automatic object category detection in images.

For object categorization, the emphasis is on predicting whether or not a novel test image belongs to a known category. The focus of the problem shifts from identifying concrete shapes to making sense of shape concepts. Strategies used for object recognition are somewhat inadequate – not because flexible template matching [154] cannot keep up with the demands of the task, but rather because

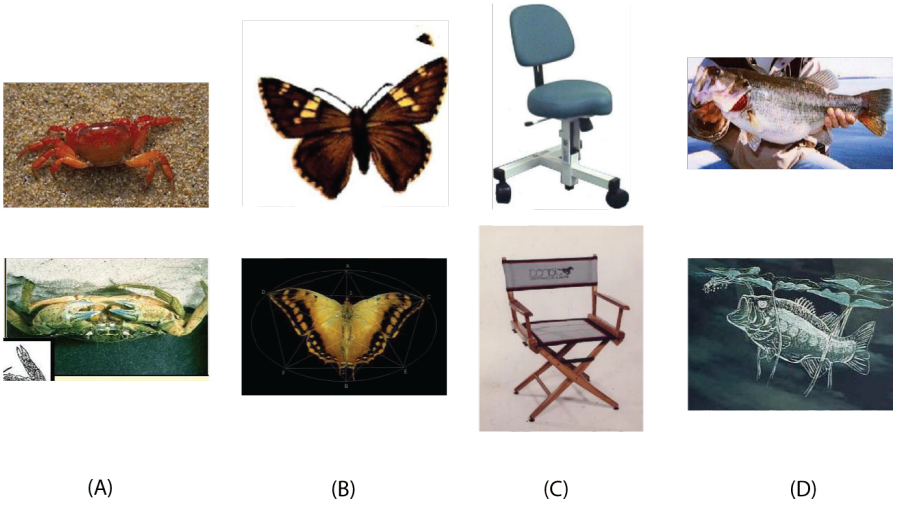

Figure 1.1: Objects belonging to the same object category do not always look alike. For example, (A) crab, (B) butterfly, (C) chair, and (D) bass.

the template library is no longer well-defined at the level of abstraction on which the system must operate. The Achilles heel of both object recognition and categorization is the assumption that the input is fully interpretable in terms of a finite set of well-defined visual concepts.

One notable problem is that raw data from images come in the form of matrices of numbers. However, the inferences that enable humans to recognize occur at a much higher level of abstraction. As a result of this observation, many researchers have proposed algorithms based on the concept of dividing the recognition task into three separate stages: early vision, intermediate vision, and high-level vision [31][110]. Early vision mainly concerns the detection of raw image pixels in order to produce useful features. Intermediate vision involves making connections between groups in the detected image in determining the composition of the image, and high-level vision involves producing a semantically meaningful representation

of the image.

The bulk of this thesis focuses on models that fall in the stage of intermediate vision. It builds on many state-of-the-art methods developed over the past several years to introduce models that produce improved categorization of objects. More specifically, SIFT descriptors [90] were used in describing image features and the proposed models are based on the successful BOW model [24]. Intermediate vision is concerned with capturing the spatial relationship between image features, which often includes reasoning about entities at a higher level of abstraction. Furthermore, the Spatial Pyramid Matching (SPM) scheme [81] is also investigated in this thesis.

1.2 Hypothesis

This thesis argues that:

> *the proposed region-of-interest detection and spatial sampling framework is more accurate than the existing state-of-the-art spatial pyramid matching approach.*

While there are many different approaches and models for computer vision systems, this thesis focuses on improving the BOW model. The goal of this research is to demonstrate that improved recognition accuracy can be obtained with better techniques for capturing spatial information between image features. In order to achieve this goal, the objectives of this thesis are to investigate the BOW model; investigate existing spatial information capturing techniques; improve spatial information capturing techniques; and create a competitive object recognition

framework.

- **Investigate the Bags-of-Words model.** Broadly speaking, the BOW model can be divided into four steps: detection of image features, description of image features, capture of spatial relationships between image features and classification (See Section 2.1 for an in-depth analysis of the model). In this research, each of the those steps are examined, to better understand the strengths and shortcomings of the model and to determine whether any improvements can be made to any of the steps.

- **Investigate existing spatial information capturing techniques.** There are a number of approaches in incorporating spatial information into the BOW model. Some of the most notable approaches are investigated, including the SPM scheme. The objective is to show that such spatial capturing techniques can be further improved and extended to better capture relevant spatial information between image features.

- **Improve spatial information capturing techniques.** The primary goal of this thesis is to develop novel spatial information capturing techniques for the task of object recognition.

- **Create a competitive object recognition framework.** Finally, in order to determine the effectiveness of the combined techniques, an object recognition framework is built, based on the proposed techniques. Experiments are carried out on some of the most popular benchmark datasets.

1.3 Contribution

The main contribution of this thesis is a set of methods and techniques that are combined together into forming a recognition framework for images. The pipeline approach, consisting of the four main contributions of this research, which includes frequent keypoint discovery [172], region of interest detection [173], and spatial sampling techniques [174][175]. Figure 1.2 details the proposed framework. The framework works by first detecting the frequently occurring local image features in the form of keypoints, in order to determine the ROI based on the detected keypoints and keypoint patterns. Various spatial sampling techniques are then applied to the ROI of the training images, before the classification stage.

The proposed framework is similar, but ultimately different, to the popular and successful SPM method. Namely, the SPM method does not use ROI detection and has a simplistic method for spatial pattern sampling.

Contribution 1: Frequent keypoint discovery. The first contribution attempts to improve recognition accuracy of the BOW model by combining pairs and triplets of frequent keypoints. The key idea behind this approach is to discover intermediate spatial representations for each object category, in an attempt to provide more descriptive power to the BOW model. Leveraging frequent keypoints, the technique works by computing spatial relationships between co-occurring image features. Experimental results show that the inclusion of such explicit co-located image features leads to improved BOW model performance.

Contribution 2: Region of interest detection. Many state-of-the-art object recognition systems rely on identifying the location of objects in images, in order to better learn their visual attributes. Based on the frequent keypoint discovery

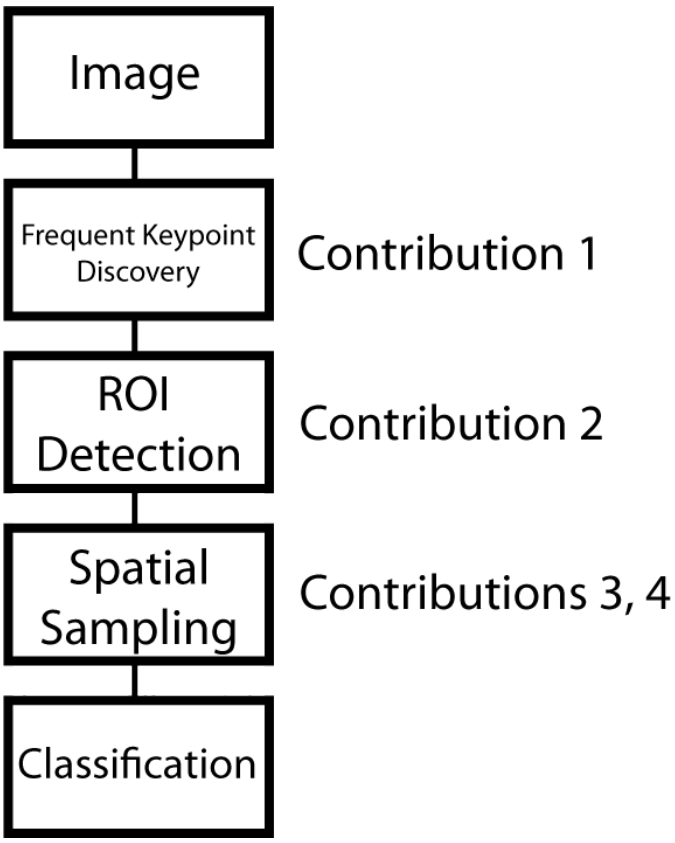

Figure 1.2: The proposed recognition framework consisting of the four contributions of this research.

technique from the prior contribution, the second contribution attempts to discover the region of interest from images. We show that the region of interest can be efficiently detected using both singles and pairs of frequent keypoints. The benefit of the proposed detection technique has been validated in two different types of datasets.

Contribution 3: Pairs and shapes frequency histogram. One of the disadvantages of the BOW model is that it discards the spatial relationships of local descriptors, which severely limits its descriptive power. In recent years, the SPM

model has been popular in solving this challenge. The third contribution is built on the SPM scheme. In that, proposed techniques are performed to capture more refined spatial information between image features. The techniques are pairs frequency histograms and shapes frequency histograms. Furthermore, various combinations of spatial and feature frequency information were also experimented with. Experimental results were encouraging.

Contribution 4: Binned log-polar representation. For the last contribution, new spatial information capturing techniques are presented. The techniques, variations of binned log-polar histograms, are based on the bin log-polar representation, which was initially developed for shape matching. Unlike the spatial pyramid model, where a sequence of increasingly coarser grids are placed over the image, this approach divides the image into grids of different scales and different orientations. This explicitly captures the distribution of image features both in distance and orientations.

1.4 Road Map

The thesis has ten chapters. Following this chapter, Chapter 2 provides the background of existing work and theories in the field of object recognition. Chapter 3 reviews some of the recent works and existing approaches for automatic object categorization. More specifically, the focus is on the successful BOW model and other schemes based on this model.

Chapter 4 introduces the datasets used in this thesis, which includes the Caltech101, Graz-02, Moths, Galaxies, MIT 15 Scenes, and VOC2008 datasets. Chapter 5 introduces the initial frequent keypoints discovery technique, and shows how

it builds on the existing work. Experiments are performed on some of the challenging datasets to show the operation of the model. Chapter 6 builds on the frequent keypoint discovery technique for the task of detecting the region-of-interest from images.

Chapter 7 introduce the pairs and shapes frequency techniques which built on the SPM scheme. An overview on the two techniques is provided first, followed by implementation details and experimental results. In Chapter 8, two more novel feature extraction techniques are presented – binned log-polar shapes and log-polar distributions.

In Chapter 9, a final and grand evaluation of the entire framework is performed on a particularly challenging datasets, namely the VOC2008 [35] dataset. Finally, in Chapter 10 the thesis statement is revisited and discusses the main hypothesis of this thesis and offers final conclusions.

1.5 Publications

The following is a list of publications that have arisen out of this Ph.D. research.

- E. Zhang, M. Mayo. 2010. Improving Bag-of-Words Model with Spatial Information. Paper accepted for presentation at the IVCNZ 2010. 8-9 November 2010, Queenstown, New Zealand.

- E. Zhang, M. Mayo. 2010. Enhanced Spatial Pyramid Matching Using Log-Polar-Based Image Subdivision and Representation. Paper accepted for presentation at the International Conference on Digital Image Computing: Techniques and Applications (DICTA). 1-3 December 2010, Sydney, Aus-

tralia.

- E. Zhang, M. Mayo. 2009. SIFTing the relevant from the irrelevant: Automatically detecting objects in training images. Paper accepted at the International Conference on Digital Image Computing: Techniques and Applications (DICTA). 1-3 December 2009, Melbourne, Australia.

- M. Mayo, E. Zhang. 2009. 3D face recognition using multiview keypoint. To Appear, Proc. 6th IEEE International Conference of Advanced Video and Signal Surveillance. 2-4 September 2009, Genoa, Italy.

- E. Zhang, M. Mayo. 2008. Pattern discovery for object recognition. Pro. Int. Vision and Computing NZ, IVCNZ.

- M. Mayo, E. Zhang. 2008. Improving face gender classification by adding deliberately misaligned faces to the training data. Proc. Int. Vision and Computing NZ, IVCNZ.

- E. Zhang, M. Mayo. 2008. Mining spatially related features for object recognition. Proc. New Zealand Computer Science Research Student Conference, Christchurch, New Zealand.

It is important to note that because this Thesis is fundamentally based on the above publications, many of the following chapters would naturally share content in a literal way in varying amounts.

Chapter 2

Background

Object recognition in computer vision began in the 1960s, when computers first became powerful enough to obtain and process digital images: a task that was far more complex than first imagined, when scientists proclaimed that:

...building perceiving machines would take about a decade.

Although what was claimed is still far from reality, progress in the field of computer vision over the last half of the century has been remarkable. One can measure this progress in terms of the problem that we are trying to solve: early works mainly focus on finding a specific object instance, with the more challenging category level recognition addressed later. It was not until the 1980s that natural images were used directly but under favourable laboratory conditions – uniform background, simple segmentation of the object, and with little or no background noise. Due to the boom in personal computers and digital cameras in the last 10 to 20 years, the field of computer vision has attracted much attention and popularity. Consequently, problems became more realistic and challenging; we are

now dealing with real life images with significant variations in scale, viewpoint, and most importantly, wider variety of object classes.

2.1 Appearance Matching

In the 1990s, when the geometric matching methods were reaching the end of their active period, a new type of recognition technique started to emerge – the global appearance matching model. In essence, an object is represented by a vector of feature values; the features can be shapes, intensity appearance, or colour. Recognition is achieved by finding the feature space closest to the input image [134, 107, 105]. More importantly, this new approach facilitates the representation of a novel object by its membership in a number of categories simultaneously. Moreover, through the use of histograms computed over the entire image, the resulting multidimensional vectors eliminate the need for precise matching – something that the geometric matching model was unable to cope with.

In this section, some of the most notable global appearance matching models are reviewed in turn: colour histograms, texture and shapes. In Appendix A, a summary of early object instance recognition is presented. In addition, early category recognition work is discussed in Appendix B.

2.2 Colour Histograms

Colour features are one of the most powerful and widely used statistics for object recognition. Among the earliest use of colour features was that of Swain *et al.* [147] for the purpose of image retrieval. More specifically, an L_1 metric, Histogram

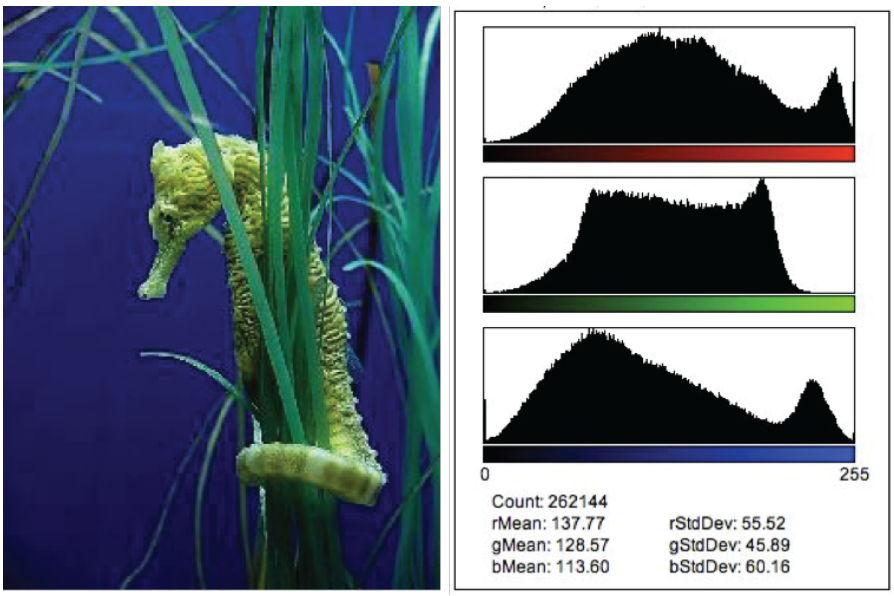

Figure 2.1: A colour histogram representation of a seahorse image.

Intersection, was proposed as similarity measure for colour histograms. Histograms are one of the most commonly used feature representations.

Statistically, colour histograms denote the distribution of the three colour channels (RGB) and are well suited to the task of object recognition because of their ability to implicitly capture complex, multi-modal patterns of colour [29]. And because they disregard all geometric information, they remain relatively invariant to scale, occlusion and noise. Moreover, it is important to appreciate that Swain *et al.'s* [147] real contribution is more than just colour histograms. Their work essentially popularized the concept of feature extraction – the means of coming up with a set of specific visual features, and techniques for extracting them, both globally and locally. Figure 2.1 illustrates this property.

Over the years, numerous efforts were made to improve the original colour

histogram model. These efforts can be summarized into two groups: distance functions and colour indexing techniques. The distance functions group mainly concerns about finding better similarity distance functions for colour histograms. For example, Ioka [65] and Niblack et al. [112] proposed the use of L2-related metrics for comparing histograms that have similar colours. The second group, colour indexing techniques, concerns with how colour is extracted and indexed. Stricker and Orengo works [146] is a good example of this. They argue that because the histograms were sparse, the original technique is extremely sensitive to noise, they advocated the use of the cumulated colour histograms to overcome this problem.

2.3 Texture

The previous section reviewed how colour can be used for object recognition. Here alternative approaches are examined which instead extract textures from images and use these as features for object recognition.

Texture is different to colour. The latter is a purely a pointwise property of the image and has no texture, whereas the former carries a notion of geometrical relationship. According to Smith *et al.* in [144], texture is not accidental or random – it is not the end product of a single colour or intensity, but rather the presence of a uniform visual patterns. Texture features, argued Datta [29], are intended to capture the granularity and repetitive patterns, as well as the spatial configurations of surfaces within an image. For example, pictures of grassland, brick walls, trees, clouds, and hair are all texture features, but with different properties of homogeneity.

The study of texture features has been popular in the field of computer vision, image processing, and computer graphics [52]. Examples of texture detectors are multi-orientation filter banks [94] and wavelet transforms [159]. The research development of using texture features can be categorized into three stages: initial approaches, psychological influenced approaches, and wavelet transform-based approaches. Each of these approaches will be reviewed in turn.

2.3.1 Initial Approaches

One of the notable earlier texture feature representations, proposed by Haralick *et al.* [52] was based on the co-occurrence matrix. The key idea was to first represent the image as a co-occurrence matrix, and because the matrix essentially captures the rough spatial configurations between texture patterns, this representative information is then extracted. Building on the foundation of this work, many researchers [45, 52] further proposed enhanced versions, after experiments showed that contrast, moments and entropy had the biggest discriminatory power.

2.3.2 Psychological Influenced Approaches

Because human vision is much superior to machine vision, it is natural for researchers to try to understand how human vision works, and to implement machine vision algorithms as close to human vision as possible. This is what the second approach attempts to do, in the work by Tamura *et al.*, in [148]. They proposed six visual texture properties (coarseness, contrast, directionality, linelikeness, regularity, and roughness) that, they argued, are the machine equivalent to the visual texture in human vision. Tamura *et al.* argued that their features were funda-

mentally different to those of the original co-occurrence matrix features on the basis that their features are visually meaningful, and that the original features (for example, entropies) may not be. Because these visual texture properties are more human friendly, in that visual textures are measured in properties that are visually meaning to humans, the Tamura texture were often successfully employed in image retrieval platforms.

2.3.3 Wavelet-Based Approaches

In the early 1990s, two decades after the original co-occurrence matrix by Haralick, the research community discovered the benefits of using wavelet transforms to represent texture features. The fundamental difference between the previous co-occurrence-based and wavelet-based methods, lies in the scale space in which features are extracted. Co-occurrence-based methods compute texture features based on the original single scale space; while wavelet-based methods, first transforms the image into a higher scale before features are extracted [79]. Due to the good performance of this simple scheme, it was later extended by many researchers [23, 57, 58, 157, 158]. These simple schemes often outperform traditional co-occurrence methods that are based on second-order statistics. Because of the lack of adequate tools, Chang [18] argues that the ineffectiveness of characterizing different scales of texture, was the biggest challenge for traditional texture analysis. This difficulty is better handled by the spatial/frequency characteristic of wavelet transformation that provides good multi-resolution analytical tools. Improved recognition performance were achieved by combining wavelet transform to other techniques. For example, Gross *et al.* [47] performed texture analysis using

wavelet transforms with KL expansion and Kohonen maps. Others [47, 76] also propose combining wavelet transforms with the co-occurrence matrix to combine the strengths of the both models.

2.4 Shape

The third appearance matching model reviewed is the shape representation. Shape features is one of the key components of image representation. Because it is relatively insensitive to noise and is inexpensive to compute, it is widely used for both object recognition and image retrieval [29]. Reflecting the historical development of appearance based object recognition approaches, there has been a shift from global representations [41, 52, 53] to more local descriptors [119, 97, 7] due to the typical modeling limitations of the global representations. In essence, there are two major approaches to shape representation: outline-based and region-based. The outline-based approaches focus on using the boundary of the shape, while the region-based approaches are more focused on the entire shape region.

One of the most notable pieces of work in the outline-based domain is by Belongie *et al.*, in [5]. Figure 2.2 demonstrates how shape context is extracted from shapes. This technique effectively works by computing the relative spatial configurations on a set of points sampled from the outline of a shape. The first step usually involves detecting the outline of the shape (for example, Canny edge detection [16]), from both internal and external contours of the shape. The second step converts these shape outlines into sample points, while preserving the geometrical configurations of the original shape. The third step concerns with selecting a point of reference, usually in the middle of the sample points, to express the configura-

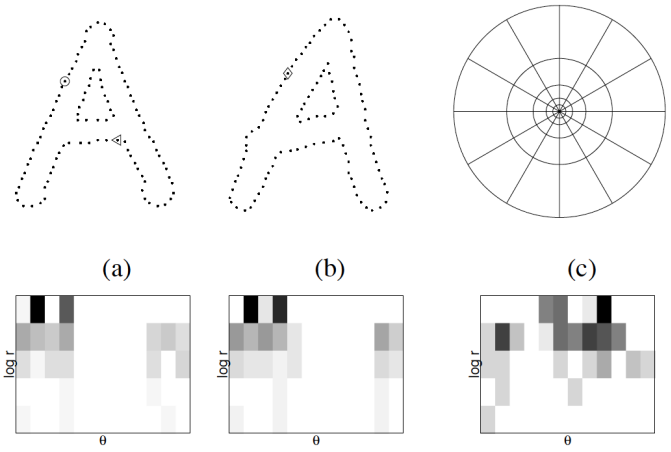

Figure 2.2: Shape contexts: (a) and (b) illustrate two similar shapes represented by edge points; (c) A diagram of log-polar histogram representation. [6]

tion of all the sample points, using the reference point. Lastly, the final step bins all the sample points in a log-polar manner – sample points closer to the reference point are treated more *carefully* as oppose to the sample points far away. Mori *et al.*, in [104] further developed this method to make it more robust and compact by constructing a bipartite graph to represent sample points on the shapes. Berretti *et al.* in [7] proposed a similar method, however, a metric tree structure is used to represent the set of curve segments or tokens.

For the region-based approaches, the main idea has been the moment invariants, proposed by Hu [59] (he proposed a set of seven moments), which are invariant to reasonable shape transformations, such as shift, scaling, and rotation. Holm [56] further improved this model by treating closed boundary regions differently. That is, these regions are represented by properties such compactness, size, perimeter, moments and moment invariants. This body of work has been successful in many research areas such as aircraft recognition [32].

2.5 Texture Region Matching

The previous sections discussed how colour, texture and edge information from the image can be used for object recognition. This section will review some of the alternative methods which extract a set of local texture regions and use these for recognition instead. These local region-based recognition methods involve three stages: detection, description and matching. In the detection stage, a so-called 'interest point detector' is used to find a set of salient parts of the image. Then, in the description stage, the detected salient parts of the image are described by some kind of descriptors. Finally, in the matching stage, recognition proceeds by matching the descriptors from the test image to those in a database of descriptors from previously 'learned' objects, an object being found if sufficient matches are found.

The significance of such local texture descriptor lies in their ability to accurately represent a local region of an image, that is searchable, comparable and have good discriminative power. Especially suitable for the task of instance-based object recognition. If the method is to be invariant to certain kinds of transformation, then both the detection and representation of salient regions must also be invariant to such transformations. The success of this method ultimately depends on the type of regions being detected, how these regions are represented, and how the regions combine to perform recognition.

The credit for the pioneering work in this area must go to Schmid and Mohr [135]. In their approach, the Harris corner detector was first used to identify interest points, then local descriptors are created for each of the interest points detected. These interest points are also invariant to scale since they are based on the mea-

Figure 2.3: Model images of planar object are shown in A: Testing image is shown in B: Recognition results are shown in C with keypoint matches. [90]

surement over a variety of scales. In training, the vectors from each interest point in the image are stored in a hash table. Recognition is done in two stages. Firstly, by looking for matching descriptors that matching both orientation and scale. Secondly, a voting scheme using the positions of interest points, is used to make the final recognition decision. Since most of the interest points from the image are used in the voting scheme, the system is not only able to tolerate occlusion and clutter, but also has an impressive speed of recognition in a large dataset [111].

Not long after the initial work by Schmid and Mohr in this field, a robust and flexible affine-invariant real time recognition system was introduced by Lowe [91], which will be discussed in the next chapter.

Local region texture has been shown to be well adapted to the the task of object recognition and matching, as it allows robustness to partial occlusion and background clutter. Recent work has concentrated on making these descriptors invariant to image transformation. One major disadvantage of the local texture

approach is that the high dimensionality of a descriptor results in a high computational complexity for recognition, while the most important breakthrough is the concept that objects can be modeled as a collection of views found by low-level texture descriptors – an idea that is responsible for a new era of research in computer vision and is the foundation of this thesis.

2.6 Summary of Appearance Matching

Overall, appearance methods, while having the advantage of being simple, are extremely sensitive to background noise and clutter. Although various improvements have been proposed, such as dividing the image into smaller sub-blocks, these types of method ultimately failed to deliver acceptable performances for recognition. Texture region matching methods, which were reviewed in the previous section, have proved to give superior performance and are more robust to background clutter and image noise.

Chapter 3

Recent Works

Despite some success in the early category level recognition methods for pedestrians, handwritten digits, and faces, the solutions are tightly tuned to these particular domains. In comparison, humans are able to recognize thousands of object categories with relative ease, despite many of the categories containing significant differences from instance to instance. For example, chairs and trees are both valid object categories but are difficult to model: one has distinctive shapes and contours and the other one is an amorphous form. The missing ingredient, it is argued, is the higher level semantic knowledge about what chairs and trees are and look like.

Recently, there has been much interest in designing recognition models that are capable of recognizing many categories, with no category specific emphasis required. For example, the Caltech101 dataset [37], where the emphasis is on identification of the intra-class variability in order to capture the essence of each class. Indeed, since purely geometric based methods are currently out of fashion, most of the recent work represents the object as a collection of statistical features.

There are essentially two categories of approaches in machine learning: gen-

erative and discriminative models (there are also learning methods that combine both generative and discriminative models). Generative or non-parametric model usually deals with high dimensional data, as they are capable of handling values of any variable in the model, in that it observes the data directly to produce the joint probability distribution. Discriminative or parametric model on the other hand, is based on the so called conditional probability distribution. That is, the values of the unobserved parameters are formed from a set of observed data [156]. However, as described above, the fundamental problem is that it is not obvious which features of the object are unique and distinctive so as to distinguish categories. Since the task of selecting appropriate features is not straightforward, many of the algorithms use some kind of feature selection techniques to make this choice automatically.

This section will first review some of the popular and recent appearance based models for the task of object category recognition. This is followed with a discussion on some of the work in incorporating spatial relationships into this model.

3.1 Bags-of-Words Model

The goal of this thesis is to improve the BOW method for object categorization. An extensive review of the original algorithm is made. Csurka *et al.* first proposed the simple yet powerful BOW technique in [24]. The approach is borrowed directly from the bags-of-words framework for text documents, with the only difference in that text words are substituted with local image descriptors. The original BOW model has shown remarkable performance in a wide range of object recognition tasks, in spite of its simplicity. In essence, the model works by representing an

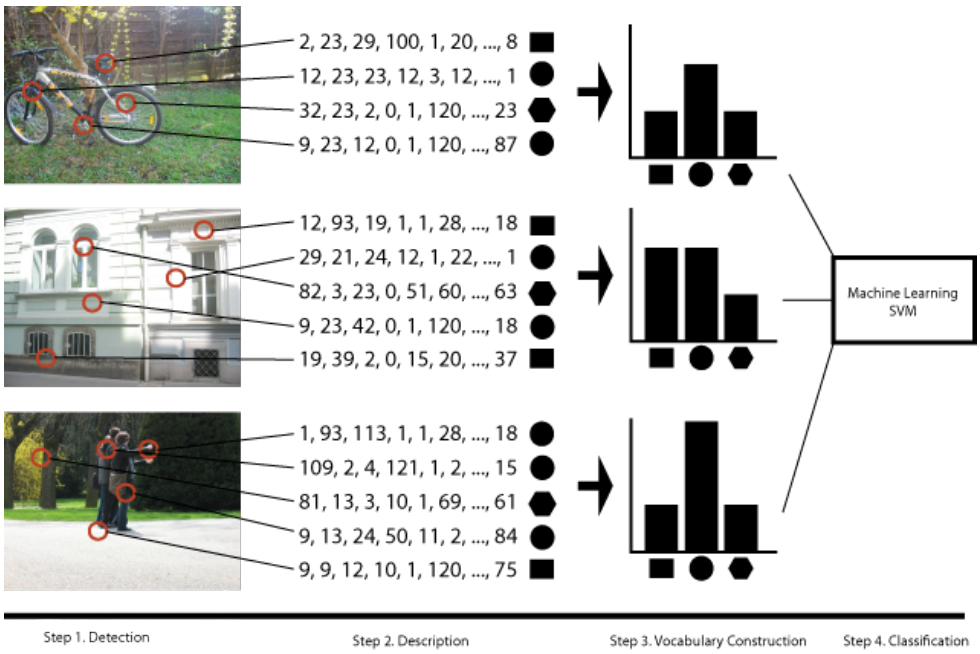

Figure 3.1: Bags-of-words (BOW) model.

image as a collection of orderless local features without any inter-mediate representation. Intermediate representations are seen as a 'bridge' to reduce the semantic gap between low-level and high-level image processing, therefore better matching computational object models with human perception.The key idea is that images can be represented by different distributions of visual words (usually SIFT keypoints [90]). A BOW model is then built as a histogram over visual word occurrences. Broadly speaking, the BOW representation is comprised of the following steps: feature detection, feature description, codebook construction, and classification. Each of the steps is reviewed in turn. Figure 3.1 illustrators the steps for the BOW model.

3.1.1 Feature Detection

The first step in the BOW model is feature detection. Recently there has been a trend towards using image keypoints in image retrieval and recognition, for example [86, 68, 178]. Image keypoints are descriptive local image patches that store distinct information about that specific region of the image. They are normally detected using a number of different feature detectors [101], then represented by various of descriptors [99]. Image patches can be defined as image patterns that differ from their immediate neighbourhood. They are usually associated with changes in image intensity, colour and texture (e.g. corners). Typically, local features are points, edges or even small image patches. (See Figure 3.2 for an example of SIFT keypoints detected from an image)

Local feature detection can be simply seen as the process of locating salient points and/or regions from images, in order to produce useful local image descriptors. Keypoints must be able to describe these salient image region so that they survive longest when the image is gradually blurred in scale scape, parametrized by the size of the smoothing kernel. Rosin [128] also pointed out that the image lifetime of edge saliency is an important selection criterion for interest points as well as local image features such as wiggliness, spatial width, and phase congruency of edges. Other important attributes of a reliable and meaningful interest point are that it must be invariant to image transformation, such as scale, rotation, and affine transform, in addition to perspective transformation, illumination and brightness variations.

Local features are sometimes referred to as invariant features as they are, to a degree, invariant to all the transformations mentioned above. Moreover, what the

Figure 3.2: Keypoints (green eclipses) are detected on the image. The scale of the keypoints is not fixed.

local features actually represent is not really relevant, as long as their placement on the image can be determined accurately and in a stable manner. Indeed, especially for the task of object recognition, local features do not even have to be localized accurately, since the goal of the task is not to match them on an individual basis, but instead, to analyze their occurrence statistics over the entire image. According to [155], good local features should have the following properties:

- Repeatability. Given two images of the same object, regardless of changes in viewpoint and scale, a high percentage of features detected on the two images should match.

- Distinctiveness. Local image features need to be unique enough to be distinguishable between similar regions.

- Locality. In order to reduce the effects of occlusion and allow for simple model approximation, image features should be local.

- Quantity. The number of detected features should be large, even for a small object in the image. This quality is important in providing a good foundation for extracting 'higher level' features in the future.

- Accuracy. The placement of the detected features should be accurately described.

- Efficiency. The process of feature detection should be fast, in order to work with real world applications.

Initially, researchers relied on using interest point detectors in determining local features. For example, the Difference of Gaussian (DoG) detector in the original SIFT keypoints [90], the Harris-Laplace (HL) detector [80], and the Laplacian-of-Gaussian (LoG) detector [88]. For the DoG algorithm, the input image is first successively smoothed with a Gaussian kernel before being sampled, where the DoG representation is obtained by subtracting two successive images. This means that all the DoG levels are constructed by combining smoothing and sub-sampling.

The Harris-Laplace detector, on the other hand, responds to localized points in scale-space, and then selects points for which the Laplacian of Gaussian attains maximum values over a scale. The Harris Affine algorithm is then used to estimate the affine neighbourhood by the affine adaptation based on the second moment matrix. The Laplacian-Gaussian detector is similar to the DoG detector in that its scale-space is built by smoothing of high resolution images with the Gaussian kernel. It is circularly symmetric and it detects blob-like structures.

It is interesting to note that recently, separate work by Lazebnik [81] and Nowak [113] have questioned whether interest point detectors are necessary at all. In their research, both researchers have used alternative local feature detection strategies to achieve good performance. Lazebnik computed 16 by 16 pixel patches over a grid with spacing of 8 pixels. Intuitively, the researcher argues that densely sampled image patches are necessary to capture uniform and low contrast regions of the image. Nowak, on the other hand, proposed random sampling of local features over the entire image. Experimental results proved that random sampling works just as well as other state-of-the-art detectors on many datasets.

3.1.2 Feature Description

The second step of the BOW model concerns describing each local feature with suitable visual descriptors. Because local region descriptors essentially *identifies* a specific region of an image, they need to be indexible and searchable, it is vital that they are robust to occlusion and yet distinctive. The proposed methods vary in complexity from a simple vector of image pixel intensities or colours, to feature similarity cross-correlation descriptors. However, despite improved performances, high dimensional descriptors result in high computational complexity for recognition and, therefore, maybe unsuitable for large datasets with millions of images. Thus, at present, local feature descriptors are mainly used for determining correspondences between images.

Local descriptors are extracted around the region around each keypoint detected by the keypoint detection process. There are several sources of information at the local level that local descriptors can be based on, such as gradient orienta-

tion, size, and point of origin. In this thesis, local invariant feature descriptors are defined simply as the fingerprint of images, where these fingerprints are essential for the task of object recognition. Generally, the primary role of descriptors is to identify and extract local features around the keypoints within the image. The extracted features must be highly distinctive and should be invariant to image noise, slight change in viewing direction, scale, rotation, and changes in illumination.

The three main categories of local feature descriptors are distribution-based, spatial frequency-based, and differential-based descriptors.

3.1.2.1 Distribution-based descriptors

This type of descriptor relies on using histograms to represent various characteristics of appearance or shape. According to [99], distribution-based image descriptors use histogram to represent the distribution of pixel intensities within the local region. A classic example of this approach is the popular SIFT descriptor proposed by Lowe [90]. The SIFT (Scale Invariant Feature Transform) descriptor builds on a grey scale image representation. SIFT features are essentially histograms of local gradient direction distribution over the entire keypoint region. Moreover, this type of descriptors are, to certain extend, are invariant to small changes in orientation and illumination. They comprise gradient vectors for each pixel in the keypoint's 4×4 pixel neighbourhood and a normalised orientation histogram of gradient directions. The size of the feature vector is thus 128 (that being $4 \times 4 \times 8$) attributes. Figure 3.3 demonstrates how SIFT is computed from local gradient directions.

Other notable distribution-based descriptors include Geometric histograms [3] and shape context [6]. Both of the these descriptors are based on the same idea, and are very similar to the SIFT descriptor. Unlike the SIFT descriptor, both methods

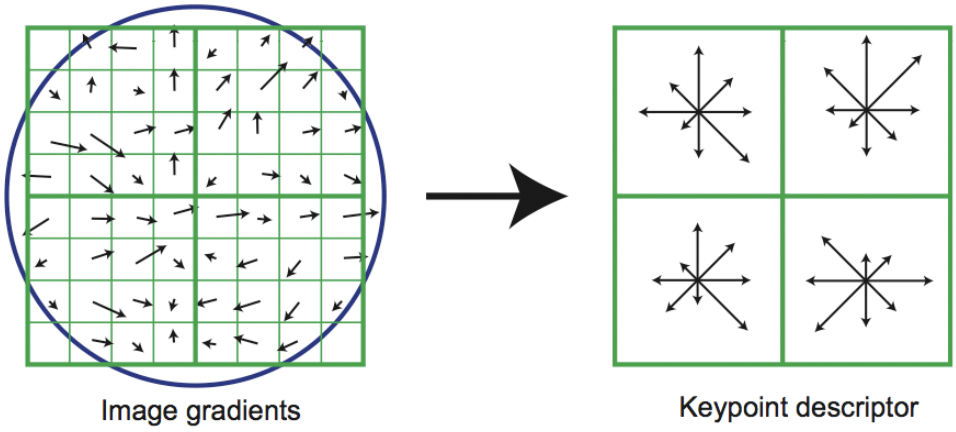

Figure 3.3: Thresholded image gradients are sampled over a 16 × 16 array of locations in scale space. Each SIFT keypoint consists of 4 orientation histograms, each histogram of size 8, which is 128 dimensions [90].

handle only edge points that have equal weight in the histogram. Moreover, both of these descriptors are mainly used for the purpose of shape matching.

3.1.2.2 Spatial-based descriptors

Spatial-based descriptors, on the other hand, transforms the image into a different frequency before describes the content of the image in the new frequency. Broadly speaking, the image is first decomposed into basis functions [99] using the Fourier Transformation. However, unlike distribution-based descriptors, spatial relationships between image features are not explicitly mapped, and because the basis function is finite, it is difficult to use this type of descriptors for localized correspondence matching. The Gabor transformation was later introduced to overcome this problem. Nevertheless, in order to capture all minor variations in the image, a large number of Gabor filters is needed, rendering this approach rather expensive.

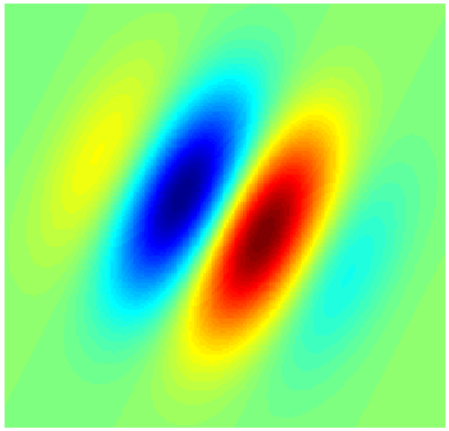

Figure 3.4: Example of a 2D Gabor filter.

Vetterli, in [160], explored the possibilities of using Gabor filters and wavelets in the context of pattern classification. Figure 3.4 demonstrates an example of a 2D Gabor filter.

3.1.2.3 Differential based descriptors

Given an order for the approximation of a point neighbourhood, differential-based image descriptors works by computing the set of image derivatives [99]. Freeman *et al.* [43] proposed steerable filters, which essentially works by steering derivatives in a particular direction, determined by the local 'jet' component. Moreover, the ability to predefine the derivative steering direction make these types of descriptors invariant to rotation. Baumberg [4] further improved this approach with the use of complex filters based on Gaussian derivatives.

3.1.3 Codebook generation

The third step in the BOW model is codebook generation. A visual codebook is essentially a rough representation of the different features in a dataset, characterizing the distribution of features within each image with a histogram. In contrast to traditional text clustering, the size of the codebook is determined by the pre-specified number of keypoint clusters.

If the size of the codebook is too small, then the model will lack discriminative power since two significantly different keypoints might be assigned to the same cluster. However, if the size of the codebook is too large, it becomes less generalized, less tolerant to noise and clutter, and requires more computational power. Previous work has used a wide range of codebook sizes for the respective models. For example, Lazebnik adopted a codebook size of 200 in [81], Zhang *et al.* [178] adopted 1000 clusters, and Sivic *et al* [142] adopted 10000 clusters for their model.

There are many approaches in creating codebooks. K-means clustering [89] is currently the most common generative method for codebook construction. K-means clustering gives impressive results but is computationally expensive owing to the cost of comparing and assigning keypoint descriptors during clustering and use. This is mainly due to the nearest neighbour search in the high dimensionality of keypoints. Moreover, Beyer [8] argues that nearest neighbour based assignments can sometimes be unstable – in high dimensions, concentration of measures tend to ensure that there are many centres with similar distances to any given point.

Once the keypoints are clustered, existing approaches with BOW models mostly adopt conventional term frequency [81] and inverse document frequency [142] approach to extract features. In the domain of text information retrieval, term

weighting is known to have a critical impact on performance. Much work has been proposed on adopting keypoint weighting into the BOW model. However, the effectiveness of such approaches remains an interesting open question. A fundamental difference between visual words and text words is that the former carries statistical information and is the product of clustering; while the latter carries semantic meaning and are sampled explicitly. Nowak [113] proposed binary weighting, which indicates the presence and absence of a image feature with values 1 and 0 respectively. Broadly speaking, these methods rely on nearest neighbour search in the codebook, and each keypoint is assigned to the closes cluster centroid.

However, Jiang *et al.* [68] argued that if the codebook size is large, it is likely for two similar keypoints to be assigned to different clusters, which is not the optimal choice. Additionally, they also argue that two keypoints belonging to the same cluster are not necessarily the same, as their distance to the centroid is different.

Several extensions were proposed to overcome these problems, Agarwal *et al.* proposed fitting a probabilistic mixture model to the distribution of a set of training image features in descriptor space. New image features can then be assigned according to their vectors of posterior mixture-component membership probabilities [1]. Jiang *et al.* [68] proposed a technique for measuring the significance of visual features called *soft-weighting*. In their work, for each keypoint in an image, instead of assigning the keypoint to the nearest clustering, determined by the centroid, they select and assign the keypoints to the top-N nearest clusters. They argue that this approach will address the problem of the aforementioned weighing schemes.

3.1.4 Classification

The final step of the BOW model concerns classification. After the image is represented by a frequency histogram of each visual word, a wide range of classifiers can then be used for the classification of this representation. Kernel-based Support Vector Machines (SVMs) are one of the most promising classifiers for such a task. However, the choice of a good kernel function is vital for the BOW model. Traditionally, the linear kernel or the RBF kernel (shown in Equation 3.1) is used.

$$K_{d-RBF}(H, P) = e^{-pd(H,P)} \tag{3.1}$$

In Equation 3.1, $d(H, P)$ can be any distance function in the feature space (for example, Chi-Square distance, Euclidean distance, and KL distance). Zhang *et al.* argued that since BOW is a histogram of visual words with discrete densities, functions such as χ^2 distance (See Equation 3.2) are more appropriate [178].

$$d_{\chi^2}(H, P) = \sum_i \frac{(h_i - m_i)^2}{m_i} \tag{3.2}$$

where:

$$m_i = \frac{h_i + p_i}{2}$$

In Equation 3.2, H and P represent keypoint 1 and 2, while h_i and p_i are the bin indexes for each of the keypoints respectively. If the two features are identical, then the χ^2 distance between them is 0. However, the chance of finding two identical keypoint matches is extremely low.

Chapelle *et al.* introduced a new type of kernel for colour histogram-based image classification [19], with the distance function defined in Equation 3.3. These

kernels are commonly used in colour image retrieval, and have demonstrated superior performance to the traditional linear and RBF based kernels [19].

$$d_b(H, P) = \sum_i |h_i - p_i|^b \tag{3.3}$$

Perronnin *et al.* in [118] proposed to apply the Fisher kernel to image object categorization. The Fisher kernel is a hybrid framework that attempts to combine both the generative and discriminative models for the task of pattern classification [66]. In essence, the kernel characterises a signal function which models the generative process of the signal. Consequently, a discriminative classifier then uses this signal as the input.

Figure 3.1 shows the key steps for a typical BOW-based method. In its basic form, the BOW approach disregards all information about the geometrical layout of the image features, which limits its descriptive abilities. In particular, the BOW models are incapable of separating the object of interest from background clutter and image noise.

However, simplicity is the ultimate strength of the BOW model. Image features tend to be very easy to detect and extract, and the statistical approach provides a framework that is well suited for both object recognition and categorization.

There are two main inherent weaknesses of current BOW approaches. First, the number of features (i.e. the codebook size) extracted from images is often very large. For example, thousands of high dimensional SIFT keypoints may be extracted from a single 640 by 480 pixel image. The BOW model also requires a large number of features because they are the parts that make up the object. Too few features will not be sufficient to represent the object, while too many

features will introduce too much background noise and image clutter. It is also computationally expensive to compare large numbers of high dimensional features.

The second weakness of the BOW model is that it does not take into account object parts that are dependent on each other. Physical objects are more than just the sum of their parts. For example, the current BOW approach might wrongly classify a motorbike object as a bike object because both objects contain wheel parts that are coarsely similar. However, in theory, this problem might be managed better if subtle geometrical information that describes how features are related were computed and learned.

3.2 Incorporating Spatial Information

In the last few years, a variety of papers have focused on incorporating spatial information into the BOW model. The challenge is how to get sufficient location information into the model to be useful for recognition without introducing computational complexities that limit performance. This section will review some of the most notable work in this research area.

3.2.1 Probabilistic Latent Semantic Analysis

Quelhas *et al.* argued that the BOW model suffers from two fundamental problems: *polysemy* – that a single cluster feature may represent a variety of image content, and *synonymy* – that the same image content might be represented by several cluster features.

In order to overcome these issues and to disambiguate the BOW representation, they proposed a generative method, probabilistic latent semantic analysis

(PLSA), for modeling the co-occurrence of image features in [123], which effectively complemented invariant local image features with the probabilistic latent space models.

Recall that generative models rely on the concept of joint probability distribution over observed data for modelling data directly. PLSA works by first assigning each co-occurrence local image features with a latent variable [1], then using these variables to construct the joint probability distribution on both the training images and codebook (this is defined as a mixture). Early experimental results were promising, and it is interesting to note that this technique works better with decreasing number of training samples.

In parallel to this work, Sivic *et al.* also proposed the joint use of local feature descriptors and probabilistic latent aspect models [141]. More specifically, Latent Dirichlet Allocation (LDA) and PLSA were used for grouping images into categories (i.e. equivalent to the number of object categories in training data). They showed that the learned (unsupervised) latent aspect has a strong correlation to the actual object categories in their experiments.

The use of LDA was further explored by Fei-Fei and Perona in [86], where they proposed to model objects as a mixture of aspects. They defined each of the aspects by a multi-dimensional distribution over the quantized local descriptors. The main difference between their version of LDA for images and the original LDA for text, is the injection of an observed class node in the model, for each of the training images – this effectively changed the unsupervised nature of LDA into a semi-supervised problem.

[1] In statistics, latent variables are not directly observed but rather inferred from other observed variables.

3.2.2 Local Spatial Information

Belongie *et al.* first proposed the shape context descriptor in [5]. Essentially, shape context is represented by the binned log-polar scheme as a descriptor for the purpose of shape matching. The idea is that relative to the reference point, sample points can represent the configuration of the entire shape. A histogram of the distribution of points over relative positions was used as a highly discriminative and compact descriptor. In order for the descriptor to be more sensitive to the sample points that are close, bins are uniform in log_j polar space. While the descriptor can be applied to greyscale images, it is very dependent on brightness values. Hence, it is more applicable for line drawings.

Broadly speaking, the first step of extraction a shape context descriptor is to identify a set of sample points based on both the external and internal contour of an object (the contour of an object is normally detected using an edge detector). The rationale is that objects belonging to the same category should exhibit similar contours and thus producing similar shape context descriptors, for both training and matching. (Figure 8.1 illustrates an example of the log-polar representation.) The shape context scheme is based on deformable shape matching where the correspondence between the model and the features in the image is posed as as integer quadratic programming problem. While such problems are typically NP-complete, they use some approximation which enables correspondence to be made in reasonable time with 50 model points and 50 possible matches for each [5]. Matching is done by iteratively deforming one contour using thin plate splines.

The shape context descriptor-like feature extraction techniques were used in this work, but the main difference was that instead of edge points, where they

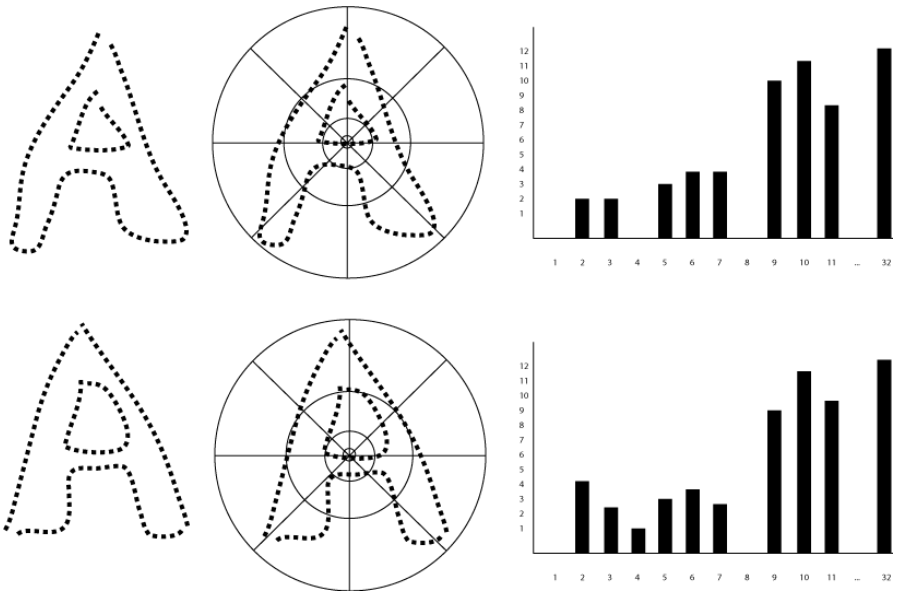

Figure 3.5: Shape matching with log-polar representation.

are all treated equally, this work focuses on capturing the shape of individual high-dimensional keypoints. See Section 8.2.1 for an in-depth description.

3.2.3 Topological Information

Sivic *et al.* were one of the first to attempt incorporating topological information by merging features into pairs [141]. Zhang *et al.* utilized proximity between image features, measured by distance between keypoint coordinates [178].

However, these approaches exploit the weakness of the dataset, where the objects of interest are almost always located in the middle of their image, and are also roughly aligned. Thureson *et al.* proposed extending the pairs-of-features approach by organizing features into triplets [153]. In their work, qualitative statistical descriptors based on local image gradient direction was investigated. Leveraging multiple gradient directions and locations, the descriptors encode qualitative rela-

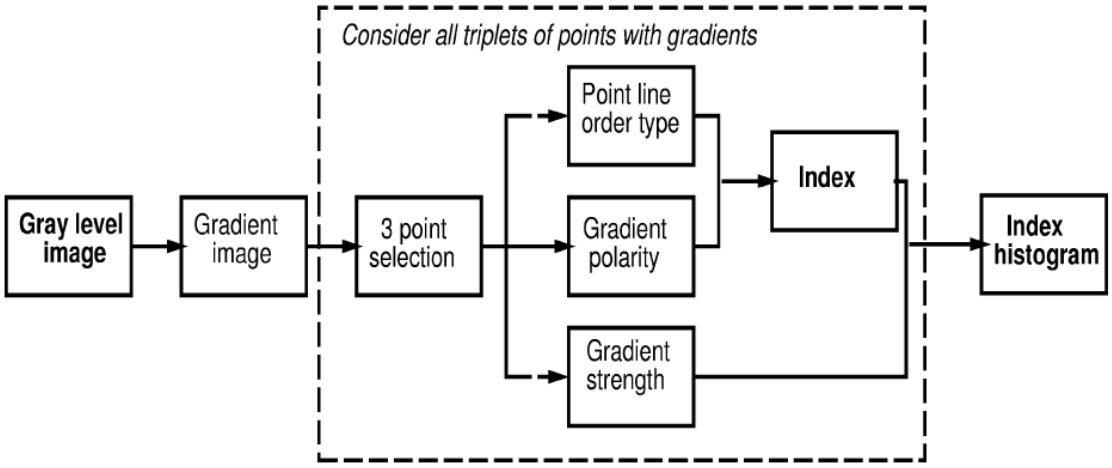

Figure 3.6: First, a gradient descriptor is applied to the images to obtain the gradient image. Then gradient information are extracted for all sampled triplets. Finally, leveraging order types and polarity, the joint qualitative structure of the three gradients are represented in an index.

tionships utilizing the order type. The descriptor for the image patch is essentially the histogram of order types. The advantage of this type of descriptor is that statistical information about the qualitative image structure are encoded, thus likely to capture the variations between images belonging to the same object category. Figure 3.6 demonstrates the procedures for the construction of the weighted index histogram.

The theory behind this model is that considering an image collection of various order types of deformations, the relative spatial configuration of the three points is key for order type invariance. They argue that provided the object has interesting invariant properties in the properties in the presence of smooth deformation (i.e. deformation that does not alter the shape significantly), the object can be retrieved if the collection of order types have well defined gradient directions.

One advantage of such an approach is that with a single shape descriptor histogram, the complexity of the model can be reduced dramatically. Moreover, higher order statics of image gradients can easily be captured using histogram – a more intermediate representation of the image.

3.2.4 Edge Fragments

Leordeanu *et al.* proposed the use of edge fragments. They argued that shape information are more distinct and representative for a range of object categories. For example, they argued that it is the shape of the object that enables *"a plane to fly, an animal to run, or a human hand to manipulate objects"* [84]. They argue that it is often the object's functionalities that defines the look and shape of the object and not the high-level surface appearance. Moreover, research in cognitive science also supports that idea that pairwise relationships between object parts are fundamental to human object recognition [63].

In their work, Leordeanu *et al.* attempt to explicitly capture the pairwise relationship between contour features, leveraging a large set of parameters. Good performance was obtained without using local appearance features, and proved that most matches are possible leveraging geometric constraints. Based on Conditional Random Fields, the proposed category shape model utilizes potential pairwise features to represent objects as a graph of fully connected parts of various spatial configurations. The resulting representation is simple, consisting mainly of sparse and abstract connected points [84].

Pairwise relationships are represented by an over-complete set of parameters. During training, boundary fragments are first extracted from images before their

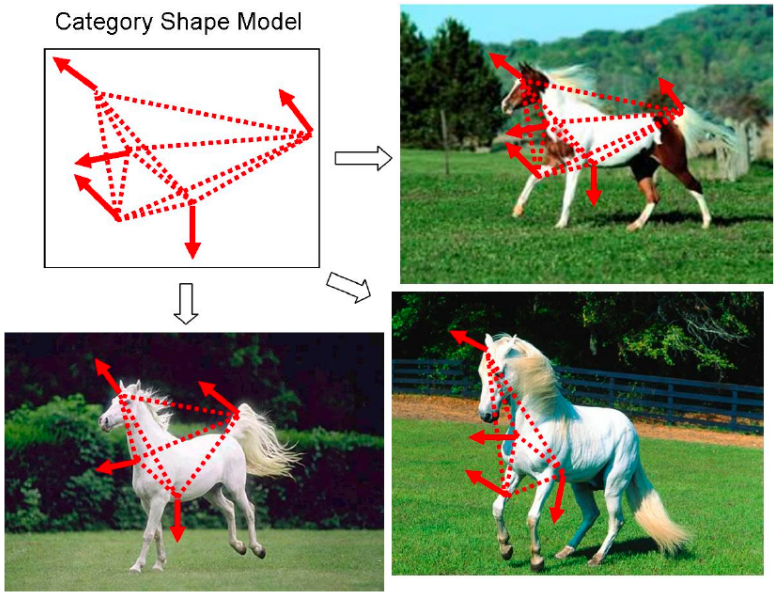

Figure 3.7: Pairwise relationships between edges are used to form the model, as the configuration of the edges capture different pose and aspects.

geometric relationships are learned. At a higher level, the model is effectively a collection of abstract pairwise spatial relationships. Figure 3.7 illustrates the model capturing different aspects or poses. The parameters for the model are learned sequentially, and the task of image recognition is essentially a quadratic assignment problem.

Opelt *et al.* also investigated using edge fragments in [116] for the task of object detection. They propose to represent object categories with a unique set of contour fragments. The rationale is that each object category contains its own set of frequently occurring contour fragments. Although they can be discriminative for learned classes, the fragments are not scale-invariant, which limits their applicability.

Dalal *et al.*, in their landmark paper on human detection proposed the his-

togram of oriented gradients descriptor (HoG), that took advantage of edge fragments [28]. In their system, an image is first divided into tiles and each one is represented by a HoG. A sliding window approach is then used to locate objects. However, it is important to note that HoG descriptors only describe the given image patch – they do not have the concepts of locality and scale.

3.2.5 Spatial Pyramid Matcing

Most recently, the SPM by Lazebnik *et al.* [81] has demonstrated promising results. Figure 3.8 illustrates the basic idea behind the SPM model. It is important to differentiate the SPM method with multi-resolution histograms, which in essence is a histogram of pixel values from different level of image resolution.

The SPM is one of the most successful extensions of the BOW model. The model builds on the pyramid matching kernel by Grauman and Darrell [46]. Broadly speaking, pyramid matching works by first divide the image into increasingly coarser scales, and feature matches are weighted accordingly. Feature matches from finer scales are given more weight. Images are perceived as been belonging to the same class if their weighted sum of feature matches are similar.

Given an image and a predetermined scale L, the first set of features are always based on the entire image. Then SPM subdivides the image progressively and extracts features again, until the L is satisfied. This means that when $L = 0$, the feature vector size is the size of the codebook, M.

Bosch *et al.* in [15] proposed selecting different fusion weights in a non-heuristic fashion, based on a cross-validation strategy. However, some researchers argue that weights are not spatially adaptive, meaning that the same image feature will

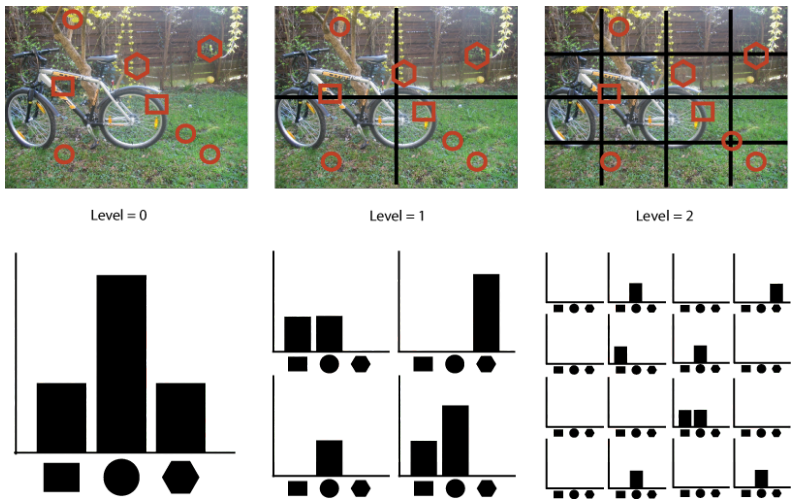

Figure 3.8: Spatial pyramid matching. It is important to note that matches found in scale L also include all the matching features found at the finer scale L - 1.

produce the same unique descriptor, even in different scale and thus unable to capture and emphasis each object classes' unique set of spatial features [54].

3.3 Summary of Recent Work

The BOW model has shown remarkable performance in a wide range of object recognition tasks, in spite of its simplicity. The key idea is that images can be represented by different distributions of visual words (usually SIFT keypoints). A BOW is then built as a histogram over visual word occurrences. More precisely, the construction of a BOW representation consist of the following steps: feature detection, feature description, codebook construction and classification. In its basic form, the BOW method discards all spatial information about how features are related and distributed across images.

The BOW model, however, discards the spatial relationships of local descrip-

tors, which severely limits its descriptive power. One of the most successful solutions to this problem, described in the seminal work by Lazebnik *et al.*, is called spatial pyramid matching (SPM). Spatial relationships between image features are important in the sense that they provide a kind of linkage information between independent image features. We believe that this information will help us better understand how object parts are related to each other, and in theory, enable classifiers to better discriminate object categories from each other. This research proposes three extensions to the SPM model. It is argued that objects belonging to the same category exhibit significant spatial regularity, and that this information should and can be incorporated into object recognition systems.

Chapter 4

Datasets

In order to provide an in-depth evaluation of the algorithms described later in this thesis, a wide range of different datasets were used. This chapter details each of the datasets and discusses their primary points of difference. Broadly speaking, the datasets vary along the following dimensions: intra-class variability, occlusion, viewpoint variation, background clutter, and image quality. See Table 4.1 for the characteristics of the datasets.

- Intra-class variability – The degree to which object instances from the same category differ from each other.
- Occlusion – The extent to which the object of interest is occluded. Detection and recognition will become more challenging if large parts of the object are not visible.
- Viewpoint variation – The viewing angle of different object instances.
- Background clutter – The amount of background noise and clutter included in the image.

- Image quality – The quality and clarity of the images in the datasets.

Table 4.1: Characteristics of the datasets.

	Caltech101	Graz-02	Moths	VOC2008	Galaxies	MIT 15
Intra-class variability	Y	Y	N	Y	N	Y
Occlusion	N	Y	N	Y	N	N
Viewpoint variation	Y	Y	N	Y	N	N/A
Background clutter	Y	Y	N	Y	N	N/A
Image quality	Average	Good	Good	Good	Average	Good

The datasets chosen focus on different aspects of these challenges, which means that the strengths and weakness of the proposed algorithms can be evaluated rigorously. Moreover, it is important to note that, in order to compare our results directly with those in the literature, different experimental setups are used for each of the datasets.

4.1 Caltech101

The Caltech101 dataset, compiled by Fei-Fei Li *et al.*, in [37]., is arguably one of the most diverse datasets in the research community. There are, in total, 101 object categories, each category containing between 31 and 800 images. The resolution for most of the images is about 300 by 300 pixels. Figure 4.1 illustrates some examples of the Caltech101 dataset.

The main difference between Caltech101 and other datasets lies in the fact that the Caltech 101 dataset is an attempt to represent objects using real world

images. This is because traditional datasets are tailored for specific problems that are being worked on, such as the SOIL47 [75], and too often methods that worked on one dataset may not work on other datasets. Moreover, this dataset also has the advantage that almost all images are uniform in pixel size and object position, and background clutter is relatively low.

However, a recent study [121] demonstrated that tests based on uncontrolled real world images can be misleading, and potentially lead research in the wrong direction. Other weaknesses of the dataset include a limited number of categories, some object categories containing only a small number of images, and some categories are simply too easy because images are highly uniform in presentation, having objects with same basic shape, viewpoint and scale.

For this dataset, 20 images per class are used for training and the rest are tagged as test images. Experiments are repeated 10 times with randomly selected training and test splits.

4.2 Graz-02

The Graz-02 dataset, compiled by Oplet *et al.* [115], is a database for both object recognition and object detection. Images from this dataset contain objects with a high levels intra-class variability and significant background clutter.

The Graz-02 dataset contains four categories: bike, person, cars and background. In total, 365 images contain bikes, 311 images contain persons, 420 images contain cars, and the final 380 images containing none of the three object classes. Figure 4.2 illustrates some examples of the Graz-02 dataset.

For this dataset, the experimental setup of Opelt *et al.* [115] was followed.

Figure 4.1: Caltech101 dataset.

Specifically, 150 example images of a positive class (bike, person, or car) were selected along with 150 random images from the other classes to serve as training data. A further 150 positive and 150 negative images were also selected for testing. Experiments were repeated ten times with randomly selected training and testing images.

Figure 4.2: Graz-02 dataset.

4.3 MIT 15 Scenes

The MIT 15 Scenes dataset is comprised of fifteen different scenes. For many years, this dataset has been the benchmarking scene category dataset in the research field. Each scene category contains roughly 200 to 400 images and the average size of images is roughly 300×250 pixels. The images were compiled from the COREL collection, person photographs, and Google images search results. Lazebnik *et al.* [81] created this dataset, where thirteen of the object categories were provided by Fei-Fei and Perona [86] and the remaining two categories were collected by the authors themselves. The fifteen categories are office, kitchen, living room, bedroom, store, industrial, tall building, inside city, street, highway, coast, open country, mountain, forest, and suburb. Figure 4.3 illustrates some examples of the

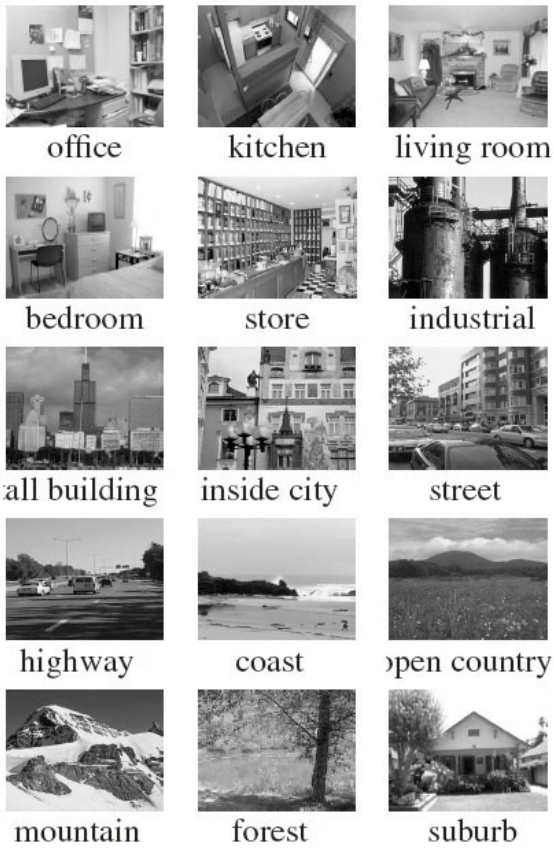

Figure 4.3: MIT 15 Scene dataset.

dataset.

The experimental setup of Lazebnik *et al.* [81] was followed for evaluating this dataset. That is, 100 images are randomly selected from each categories for training, and the remaining images are tagged as test images. The experiment is repeated ten time with randomly selected images for both training and testing.

Figure 4.4: Moths dataset.

4.4 Moths

The moths dataset is a collection of live moth images compiled by Watson *et al.* [165]. The moth trap was placed in Treborth Botanical Garden, Gwynedd, UK, and was sampled every morning. Moths were released after been photographed.

All images are 1024×960 pixels in resolution, and are displayed in full 24-bit RGB colours. In total, 35 classes of moths, or a total of 774 images, were actually used. In all of the images, the moth is always located in the middle of the image and photographed against a flat and constant background. Figure 4.4 depicts some examples of the moths dataset.

For this dataset, the experimental setup of Mayo *et al.* [96] is followed, that is, ten-fold cross validation is applied to all of the images.

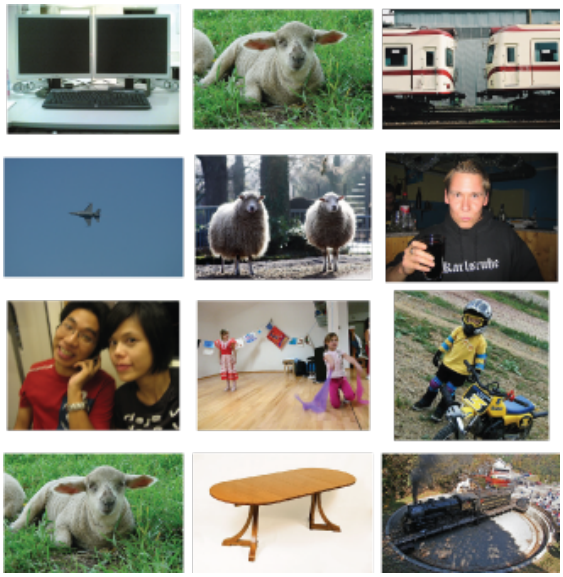

Figure 4.5: VOC 2008 dataset.

4.5 PASCAL Visual Object Classes Challenge 2008 (VOC2008)

The VOC2008 dataset [35] is a yearly benchmarking dataset for object recognition and detection algorithms. The dataset consists of 8465 images of 20 different object categories such as bike, train, human, cat, etc. The dataset is divided into a predefined training set of 4332 images and testing set of 4133 images. Figure 4.5 depicts some examples of the dataset.

For this dataset, the experimental setup is 50 images per class used for training and 50 images per class tagged as test images. Experiments are repeated ten times with randomly selected training and testing images.

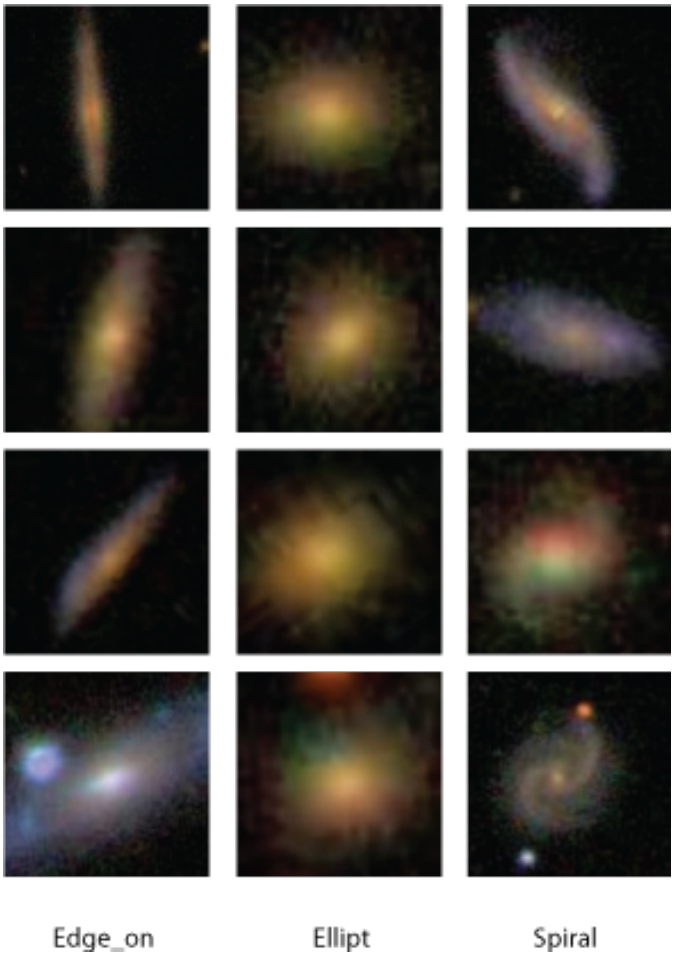

Figure 4.6: Galaxies dataset.

4.6 Galaxies

The final dataset used for evaluating the proposed algorithms is the Galaxies dataset [137]. The dataset consists of three categories of galaxy images – edge, elliptical, and spiral. The dataset contains, in total, 567 colour images with a relatively small resolution at 120×120 pixels. Figure 4.6 depicts some examples of the dataset.

For this dataset, the experimental setup is 80 images per class used for training

and 20 images per class tagged as test images. Experiments are repeated ten times with randomly selected training and testing images.

Chapter 5

Frequent Keypoint Discovery

In this chapter, a new approach is presented: the frequent keypoints discovery model. This is an extension of the original BOW model proposed by Csurka *et al.* [24].

In this approach, spatial information is extracted between image features by discovering the frequently occurring keypoints in each object category. This model is inspired by Hummel *et al.*'s [63] research into cognitive science which argues that pairwise relationships between object parts play an essential role in human object recognition. Indeed, spatial information between object parts (image features) are important in the sense that they provide a kind of linkage information between otherwise independent object parts. This information informs us how object parts are related to each other and thus enables classifiers to better discriminate object categories from each other.

The model differs from previous works into incorporating spatial relationships between local features in that it models these relationships explicitly. The focus is on the problem of finding frequently occurring keypoints and keypoint patterns

from images. There are two main contributions of this model. Firstly, it is claimed that pairs or triplets of keypoints are worth analysing because their inclusion increases accuracy, compared to when they are not used. Secondly, the proposed model enables frequent keypoint patterns to be visualised and interpreted.

The next section gives an overview of the problem and some of the existing solutions. Section 5.2 describes the approach and its rationale. The experimental results of the two standard datasets are then presented in Section 5.4.2. Results from the experiments will be explained in Section 5.4.3, before the chapter is concluded in Section 5.5.

5.1 Overview

The pioneering BOW approach [24] works by representing an image as an orderless collection of local features, without any intermediate representation. Because of this orderlessness, the BOW model disregards all information about the geometrical layout of image features; It therefore has limited descriptive abilities.

According to Lazebnik *et al.* [81], the four main shortcomings of BOW models is their inability to separate the object of interest from background clutter and image noise, viewpoint variation, occlusion and scale changes. Recently, in [86] and [142], researchers have shown that the inclusion of intermediate representation or themes could improve recognition accuracy significantly. Section 3.2 gives an in-depth review of existing spatial information capturing techniques.

It is important to note that BOW models do not take into account the fact that object parts are dependent on each other; that is, that physical objects are more than just the sum of their parts.

For example, the current BOW model might wrongly classify a motorbike object as to a bike object because both objects contain wheel parts. However, in theory, this problem might be managed better if the geometrical information that describes how features are related is computed and learned. It is argued that objects belonging to the same category exhibit significant similarity in their spatial configuration, and that this information can and should be incorporated into object recognition systems.

It is this notion that forms the basis of this model. Is it possible to discover patterns based on frequently occurring image features? Furthermore, how can this information be utilised to achieve better recognition performance? The methodology of this preliminary prototype is explained in the following section.

5.2 Methodology

Object models consist of a number of parts. Each part has an unique feature appearance, relative scale and the possibility of being occluded. Given a set of images containing an object, the features of the object will not only be found on the object, but also on the background of the images. Indeed, a large portion of keypoints extracted from images are not useful, as these mainly come from the background and other irrelevant parts of the image. In order to filter out these keypoints, a frequent keypoint selection technique for discovering frequent and informative image features for each image category is first developed. Then, in order to discover spatially related keypoints to make objects more discriminative, a pattern discovery technique that works by discovering patterns between the frequently occurring keypoints is proposed. Finally, a fast and efficient method for

generating low dimensional feature vectors from high dimensional keypoints and keypoint patterns is also presented.

5.2.1 Frequent keypoint selection

In the original BOW approach, local image features are first detected and identified by descriptors such as the SIFT descriptor. The k-means clustering technique is then used to identify the most informative image features. Essentially, image descriptors from all training images are first grouped into different clusters using the k-means technique. The mean of "centroid" of each cluster is then extracted as codebook words. Normally the Euclidean distance is used as the distance function, and the number of clusters is dependent on the desired codebook size.

One major disadvantage of the k-means clustering approach is that the performance and accuracy tradeoff must be fine tuned [113]. For example, if the number of clusters (C) is small (e.g. $C = 100$), then too many unrelated features are grouped together. This means that many visually unrelated features are grouped into the same cluster. However, if the number of clusters is large (e.g. $C = 10,000$), the amount of time this technique requires for computing the centroid of the clusters is impractical. This is because, at every iteration, the centroid of each cluster must be recomputed. This is computationally expensive when working with a large number of high dimensional SIFT keypoints. Moreover, using a large number of clusters is not necessarily an advantage for recognition systems as the codebook will overfit the training images and therefore be unable to cope with slight normal variations in the features.

Instead of using k-means clustering, a new frequent keypoint selection method

is presented which is significantly faster because there is no need to recompute the keypoint clusters at each iteration. Algorithm 5.1 describes the proposed approach. Essentially, the proposed method attempts to determine object features that are the most informative in describing a specific category. To this end, keypoints are first grouped and then ranked in terms of frequency, from most frequent to least frequent.

Algorithm 5.1 The frequent keypoint selection method

Input: Labelled training images and N.

1: For each object category, detect and extract keypoints from all training images in that category.
2: Keypoints from the first image are used to populate a frequent keypoints list.
3: Traverse through one keypoint at a time, using the χ^2 distance function to determine the distance between the current keypoint and all the existing frequent keypoints.
4: **if** the distance between the current keypoint and the closest frequent keypoint is too large, **then**
5: Add the current keypoint to the frequent keypoint list.
6: **else**
7: Increase the weight counter of the matching frequent keypoint by 1.
8: **end if**
9: After traversing through all of the keypoints, rank all of the frequent keypoints based on their weight and select only the top N number of frequent keypoints per object category.

Output: A list of frequently occurring keypoints for each object category.

Figure 5.1: All keypoints are selected for the image at the top. Only frequent keypoints from the *Bike* class are selected for the image at the bottom.

Algorithm 5.1 selects only the top N (where N is low, e.g. $N = 50$) of frequent keypoints because they, in theory, are the most distinctive and informative of a particular object category. Lower ranked keypoints are ignored as they are more likely to be background clutter or image noise. This is because they do not frequently appear through all of the training images belonging to that specific class. Figure 5.1 gives an example. In Figure 5.1, it is clear that keypoints are not only selected on the object of interest, but also significant numbers of keypoints are detected on the background as well. After selecting only the top N frequent keypoints from the object class, a more accurate representation of keypoints are detected on the bike.

Because SIFT keypoints are high dimensional, several different high dimensional distance measuring functions were tested in determining the distance between keypoints. Experimental results suggest that the χ^2 distance measure [31] is the fastest and also has the highest accuracy. Section 5.3 has an in-depth review of the different distance measure algorithms evaluated.

5.2.2 Spatially related feature discovery

Once frequent keypoints are found for each of the object categories, the next step is to use these frequent keypoints to discover keypoint patterns from each of the images. It is important to note that these patterns are based on frequent keypoints only. In Figure 5.2, frequent keypoints are first depicted in relative positions before patterns are discovered based on the proximity between all of the frequent keypoints.

It is argued that single independent keypoints may not always be unique to any

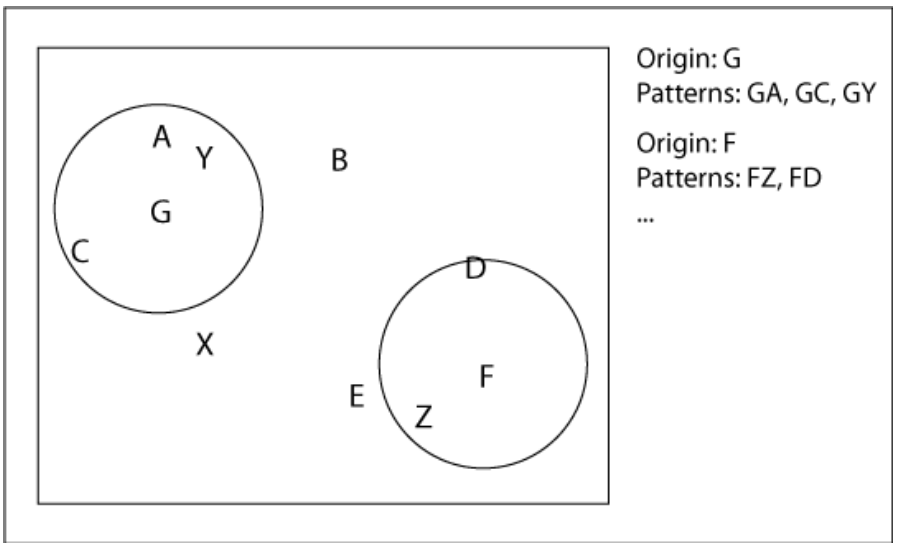

Figure 5.2: Keypoint patterns are discovered based on the frequent keypoints. Note that in practice, a circle or a region-of-interest is mapped for all frequent keypoints in determining patterns. Here, only two regions-of-interest are displayed for clearer demonstration of the technique.

object classes. However, pairs ($K = 2$) or triplets ($K = 3$) of keypoint patterns are more distinctive and informative to individual object classes. In this work, K is limited to 1, 2, or 3 only. The $K = 1$ situation was described previously (Algorithm 5.1), while $K = 2$ and $K = 3$ are described here. Algorithm 5.2 illustrates the pattern recognition approach. Broadly speaking, this approach works by looking for pairs or triplets of frequent keypoints on a predefined area. The predefined area is determined by the location of the frequent keypoints. For example, if there are ten frequent keypoints detected on the image, then the regions around those keypoints will be used to generate pairs or triplets of patterns.

Once again, after all possible patterns are extracted from an image, the algorithm selects only the top N ($N = 50$) most frequent patterns from each object

Algorithm 5.2 The pattern discovery method

Input: An image and its detected frequent keypoints, N, radius and K.

1: Traverse through all frequent keypoints extracted from each training image.

2: For each frequent keypoint, determine all possible pattern combinations within a predefined radius.

3: Generate all unique pairs or triplets (depending on the choice of K) of keypoint patterns from the set of selected keypoints from step 2.

4: Keypoint patterns from the first image are used to populate a patterns list.

5: New patterns are compared to existing patterns list and ranked similarly to the single keypoint approach described in Algorithm 5.1.

6: Select only the top N number of patterns from each object class.

Output: A list of frequent patterns.

class. Various region radius sizes ranging from ten pixels to the entire image were tested; based on cross validation information, it is determined that the radius of 50 pixels is the best for the datasets used in this work.

5.2.3 Binary feature vector generation

The final step of the process is to generate a feature vector based not only on the single frequent keypoints, but also on the frequent patterns. In the model, for each of the object classes, only the top N single keypoints as well as top N number of keypoint patterns are taken, as they are, in theory, the most informative features that describe that object's class (see Table 5.1). A table is then formed by combining all the single keypoints and patterns from all the object classes. Essentially, the table contains all the frequent keypoints and keypoint patterns for

every object class in the dataset.

Table 5.1: Example of binary feature vector generation. Q_i represents keypoints, P_i represents keypoint patterns.

Image	Q_1	Q_2	Q_3	Q_4	Q_5	...	P_1	P_2	P_3	P_4	...	Class
Image 1	45	2	0	1	2	...	2	0	1	3	...	Bike
Image 2	1	0	12	6	7	...	1	20	11	2	...	Car
Image 3	0	0	14	3	7	...	0	13	8	11	...	Car

Once the table is constructed, all of the keypoints from the training images are compared to the table. If a match is found, then the counter for the keypoint of that image on the table is incremented by 1. The same applies to patterns. This is done by first finding a match for one of the keypoint patterns, and then within the predefined radius the system will try to find the remaining matches for the other keypoints in the pattern. The counter of that pattern of the image will also be incremented if a match for that pattern is found.

By representing high dimensional keypoints and spatially distributed keypoint patterns with a single number, this binary feature vector generation approach not only reduces the size of feature vectors, but most importantly, provides a simple and intuitive way in representing spatial patterns in images.

5.3 Similarity Measuring Techniques

In Section 5.2.1, a new image feature selection method for producing a set of frequently occurring keypoints for each object class is presented. This new method not only replaces k-means for building BOW-like model, but additionally extended

to spatial relationships. The rationale for the keypoint selection method is that for all the images belonging to the same object class, it is claimed that a spatially localized subset of visual features will occur in most of the images. For example, features from wheels and tyres from images belonging to a car object class will tend to occur in one region of an image and this can help.

Essentially, the high dimensional SIFT keypoints used for describing image features are histograms. They require high dimensional metrics for comparison. Several popular high dimensional distance functions were evaluated, specifically Euclidean distance (L^2), Kullback-Leibler divergence, and χ^2. It is determined that the choice of distance function is vital for good performance. Since the distribution of data can be easily be captured (although roughly) using histograms, closeness measures can easily be obtained using both statistical-based tests and vector norms. This section reports the performance of the three distance functions.

5.3.1 Euclidean distance

Euclidean distance is one of the most commonly used distance functions. The distance function examines the root of squared differences between various aspects of a pair of objects. In computer vision, Euclidean distance is often used with great success for computing dissimilarity between colour images. Equation 5.1 illustrates the Euclidean distance function.

$$d_{L^2}(H, P) = \left(\sum_i |h_i - p_i|^2 \right)^{\frac{1}{2}} \tag{5.1}$$

In Eq. 5.1, H and P represent keypoints 1 and 2, while h_i and p_i are the bin index for each of the keypoints, respectively.

5.3.2 Kullback-Leibler divergence

The second method evaluated is the Kullback-Leibler divergence distance function. The Kullback-Leibler divergence measures the difference between two probability distributions, in a non-symmetric fashion. From an information theory point of view, the KL divergence has the advantage that it measures the average efficiency, it would need to construct a histogram using the other as the codebook. However, one drawback of the K-L divergence is that it is non-symmetric and is thus sensitive to the size of histogram bins. Equation 5.2 illustrates the Kullback-Leibler divergence.

$$d_{KL}(H, P) = \sum_i h_i \log \frac{h_i}{p_i} \tag{5.2}$$

5.3.3 χ^2 distance

Finally, the χ^2 distance function was also evaluated. Unlike Euclidean distance, the χ^2 distance attempts to provide a more "balanced" measure weighting each square by taking the inverse frequency for each corresponding terms. According to [130], the main reason for dividing each squared term by the expected frequency is to normalize the effect of variance between the high and low frequencies. Moreover, the normalization process guarantees that the differences between larger and small proportions are equalized; otherwise differences from larger proportions will dominate and overshadow smaller proportions. Equation 5.3 describes the χ^2 distance.

$$d_{\chi^2}(H,P) = \sum_i \frac{(h_i - m_i)^2}{m_i} \tag{5.3}$$

In Eq 5.3, $m_i = \frac{h_i + p_i}{2}$ This distance measures how unlikely it is that one distribution was drawn from the population represented by the other [130].

5.3.3.1 Performance comparison

The three different distance functions are appropriate in their own areas. For example, the KL divergence is justified from an information theory point of view, while χ^2 distance is based on statistical methods. Overall, χ^2 achieved the best results in both the datasets and was subsequently used as the distance measure in the proposed model. Euclidean distance, on the other hand, did not perform well. See Table 5.2 for results on the MIT 15 Scenes dataset and Table 5.3 for results on the Caltech101 dataset.

Table 5.2: MIT 15 Scenes dataset.

Method	K = 1	K = 2	K = 3
L^2	49.3%±2.3	57.1%±2.3	55.9%±2.4
KL	51.8%±2.4	58.3%±3.0	58.1%±2.6
χ^2	52.5%±3.0	59.1%±2.9	58.7%±2.9

5.4 Evaluation

This section reports the experiment setup and results. All experiments are repeated ten times with randomly selected training and test images. Multi-class

Table 5.3: Caltech101 dataset.

Method	K = 1	K = 2	K = 3
L^2	37.8%±1.2	41.3%±1.2	39.8%±1.5
KL	40.1%±1.8	44.6%±1.6	41.4%±1.7
χ^2	40.9%±1.6	44.5%±1.7	42.3%±1.6

classification is done with a polynomial Support Vector Machine classifier with default parameters as specified in WEKA V. 3.5.5 [167] and an exponent value of 0.5. For the number of frequent keypoints and keypoint patterns, only the top 50 ranked keypoints and patterns from each class were used. The radius for generating keypoint patterns is set at 50 pixels.

This section is divided into three parts: datasets, experimental results and discussion. In the datasets section, two datasets used for the evaluations are presented briefly. In the experimental result section, the experimental setup are described. Finally, this chapter is concluded with a discussion section.

5.4.1 Datasets

The proposed model was evaluated on two of the most popular datasets: Caltech101 (Section 4.1) and MIT 15 Scenes (Section 4.3). Results of the evaluations are now described in turn.

5.4.2 Experimental Results

The proposed model was evaluated on the two datasets. For the Caltech101 dataset, the experimental setup of Zhang *et al.* [177] was used. Specifically, 30

Table 5.4: Results for the Caltech101 dataset

Zhang *et al.*	K = 1	K = 2	K = 3
66.2%	40.9%±1.6	44.5%±1.7	42.3%±1.6

Table 5.5: Results for the MIT 15 Scene dataset

Lazebnik *et al.*	K = 1	K = 2	K = 3
81.4%	52.5%±3.0	59.1%±2.9	58.7%±2.9

images per class are used for training and the rest are tagged as test images. Experiments are repeated ten times with randomly selected training and testing images. Zhang *et al.* achieved an accuracy of 66.2% for this dataset, while the proposed model achieved 44.5%. (See Table 5.4 for results produced from the proposed methods.) Essentially, their model is based on the BOW model, but instead of using SIFT keypoints, they use a combination of shape and context descriptors such as geometric blur and shape context descriptors. Moreover, they have also proposed a new kNN kernel for their SVM-kNN classifier.

For the MIT 15 Scenes dataset, the experimental setup of Lazebnik *et al.* [81] is followed. That is, for each of the categories, 100 images are randomly selected for training and the remaining images are flagged as test images. Lazebnik *et al.* achieved 81.4% for this dataset. The best result obtained using K = 2 is 59.4%. (See Table 5.5.) Their approach is based on the spatial pyramid matching scheme. (For more details on the SPM method, see Section 3.2.5).

5.4.3 Discussion

Both Tables 5.4 and 5.5 show the benefits of combining keypoint patterns (i.e. $K = 2, 3$) compared to using a standard single keypoint (i.e. $K = 1$) approach. Results obtained from the current approach are no better than the state-of-the-art results for both datasets. However, this was, nevertheless a good attempt with the proposed spatially related keypoint approach, especially for the Caltech101 dataset. There are two main reasons why the proposed model did not perform as well as some of the state-of-the-art method: feature descriptors were not optimized and, more importantly, kernel selection was not performed.

For the purpose of evaluating the spatial keypoint patterns concept, only the SIFT keypoints were used for extracting features from images. Although the SIFT keypoints are robust and invariant to various transformations, it is not an optimized descriptor for all images. Zhang *et al.* in [177], have shown that the shape context [6] descriptor is more suitable for the capturing of scene context, while object shapes are better captured with geometric blur [6] descriptors. For the proposed model, instead of using SIFT keypoints as the only feature descriptor, different local feature descriptors should be evaluated in determining the most effective descriptor for the dataset.

The second reason the proposed model did not perform well is due to kernel selection of the SVM classifier. In the proposed model, a SVM classifier with a polynomial kernel with default parameters is trained for classification. In [177], Zhang *et al.* proposed SVM-kNN, a nearest neighbour based kernel with various distance measures for their SVM classifier, which clearly outperformed the polynomial kernel.

5.5 Summary

In this section, a new BOW approach for image categorization is proposed. The proposed approach is different to the original model in the sense that a faster alternative to finding the most informative patches from the image is introduced. Moreover, a spatially related keypoint pattern discovery and matching technique is also developed. Finally, an efficient method in constructing the feature vector that also supports spatial keypoint patterns is suggested.

This work focused on the problem of finding frequently occurring keypoints and keypoint patterns from images. The two main contributions are: 1) it is argued that spatial relationships between keypoints are worth analysing because they increase accuracy, compared to when they are not used; and 2) the proposed model enables frequent keypoints and keypoint patterns to be visualised and interpreted. Results obtained so far are promising, especially for the challenging Caltech101 dataset.

Chapter 6

Automatic Region of Interest Detection for Improved Training Images

The previous chapter showed how frequently occurring keypoints and keypoint patterns can be utilized to help improve object recognition accuracy. In this chapter, several methods of using these frequent keypoints and frequent keypoint patterns for the task of object region-of-interest detection are presented.

Many state-of-the-art object recognition systems rely on identifying the location of objects in images, in order to better learn visual attributes. In this chapter, four simple yet powerful hybrid region-of-interest (ROI) detection methods are proposed. The methods combine both local and global features, and are based on the frequently occurring keypoints introduced in the previous chapter. The proposed ROI detection methods demonstrate competitive performance using the popular benchmarking dataset, the VOC2008 dataset.

6.1 Overview

Generic object categorization is a challenging problem in computer vision. Given an arbitrary image, the goal is to classify it according to the objects that can be detected and recognized – a task that is natural and effortless for the human visual system, but has proven to be difficult for current computer vision algorithms. One of the main challenges is variability and the need to generalize across variations in the appearance of objects belonging to the same class. Specifically, this chapter focuses on the problem of determining the ROI from images. It is argued that by homing in on the object of interest, visual attributes will be better learned while eliminating background noise from the images in which the objects are detected. It is vital to note that when dealing with large datasets, it is especially important to eliminate as much background as possible. Figure 6.1 depicts an example of ROI detection.

The reasoning for this approach is two-fold. Firstly, it is argued that the category of an image can be described reliably by low-dimensional global features, as demonstrated in [114], where spectral and coarsely localized information is used to provide a meaningful description of the image and its semantic category. Secondly, unlike local features, global features are inexpensive to compute, which is essential for large datasets (e.g. the Caltech101 dataset).

To this end, a hybrid approach that is based on both local and global features is proposed. More specifically, local features are used to determine the frequently occurring visual attributes for object classes, before aggregating statistical information over not the entire image but rather a specific subregion that is detected the region-of-interest.

Figure 6.1: The entire image is used to extract visual features in A. In B, only the ROI is used to extract visual features.

The motivation for this chapter is that detecting ROIs in training images should increase the accuracy of classifiers built with those ROIs rather than the entire image. The rationale is that for all the images belonging to the same object class, a spatially localized subset of visual features will occur in most of the images. For example, features from keyboard and monitor from images belonging to a computer class will tend to occur in one region of an image. Background clutter and image noise will also occur in all images, but they will be different across different images.

In order to locate informative image features, three methods for automatically detecting ROIs from training images were developed. Unlike the popular sliding window object localization method, the proposed methods only need to deal with a small number of frequent keypoints in each of the images. The method builds on the theory that frequently occurring keypoints are unique and informative in representing specific object classes.

The rest of this chapter is organized in the following order. In the next section,

current ROI detection approaches are discussed, both their strengths and weaknesses. Then in Section 6.3, methods for detecting ROI are presented. Evaluation results of the two datasets are presented in Section 6.4.2. In Section 6.4.3, the results of the experiments are discussed, and this is followed by the conclusion section.

6.2 Background

This section consists of two parts. In the first part, some of the popular sliding windows approaches for detecting ROI in images are discussed. In the second part, the PHoG (pyramid representation of histogram of gradients) descriptor is presented; this was the feature descriptor used to extract visual features from both training and testing images.

6.2.1 Sliding Window for Object Localization

Object localization with bounding boxes, based on the sliding window technique, has been popular recently [22][27][39][15]. Broadly speaking, the sliding window method works by first dividing an image into smaller patches, then transforming the object localization problem as localized feature detection. In other words, a classifier is applied to all sub-images generated from the sliding windows, the regions with the highest classification scores indicating possible object presence [78].

One inherent disadvantage of this approach is the significant increase in computational cost, because of the large number of candidate sub-images. Moreover, according to Lampert *et al.*, in [78], the number of sub-images with order n^4 for images of size $n \times n$ is computationally too expensive when dealing with large

datasets.

It is also argued that in order to speed up the search process, several heuristic methods have been proposed. These approaches can be grouped into two categories. The first category [27][39] is based on reducing the number of function evaluations by enforcing a coarser grid for all possible sliding windows, and by allowing only limited shapes and sizes as candidates. The second category [14][22] utilizes local optimization methods by applying them locally, where the shape of the region is optimized by making small changes to it in a gradient ascent procedure. However, these sped-up approaches severely limit localization robustness in order to achieve acceptable speed. The two approaches will be described in turn.

6.2.2 The PHoG Descriptor

The PHoG descriptor [15] is a version of the HoG descriptor [27] with spatial pyramid extensions proposed by Lazebnik *et al.* [81]. The HoG descriptor, in essence, represents local shapes by histograms of edge orientation gradients within an image sub-region quantized into M bins. Each bin in the histogram represents a similar group of image gradients that share certain angular properties. The PHoG descriptor can be seen as a multi-level BOW model where each visual word represents a particular gradient orientations.

Based on the shape presentation of the HoG descriptor, spatial properties of images can be better captured by combining the spatial pyramid scheme. In other words, the image is repeatedly subdivided into smaller sub-regions and features are extracted from these progressively smaller sub-images.

6.3 Algorithms for ROI Detection

This section discusses the algorithms developed for detecting region-of-interest (ROI), from training images, in order to improve recognition performance. The four algorithms, single keypoint patch selection (Algorithm A), single keypoint bounding box (Algorithm B), pairs of keypoint patch selection (Algorithm C), and pairs of keypoint bounding box (Algorithm D), rely on the use of frequent keypoints for locating the ROI from images. In the remainder of this section, frequent keypoints selection is first discussed, then explanations are given for each of the proposed detection methods.

6.3.1 Frequent Keypoint Selection

A large portion of features (keypoints) detected and described by descriptors such as SIFT descriptors are not useful, as they consist mainly of background clutter and image noise. Using all the keypoints of the whole image for classification can lead to a very high computational complexity which severely hampers recognition performance. In order to select only the most informative image features for classification, k-means clustering [89] is one of the popular approaches used to determine the most frequent image features.

Instead of using k-means clustering to discover the representative features of the object category, an alternative technique described in the previous chapter is proposed – frequent keypoint selection. Since the frequent keypoints selection method is already described in detail in section 5.2.1, it will not be discussed here in depth.

One advantage of the proposed method is that it is able to visualize frequently

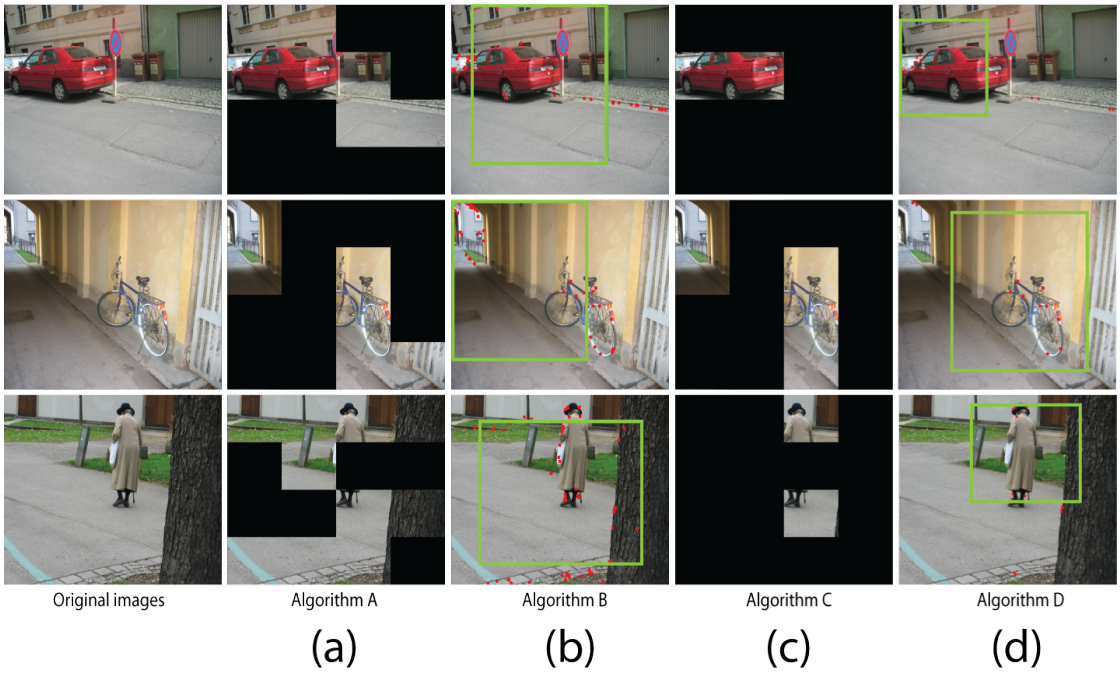

Figure 6.2: Example of single keypoint patch selection (Algorithm A), single keypoint bounding box (Algorithm B), pairs of keypoint patch selection (Algorithm C), and pairs of keypoint bounding box (Algorithm D).

occurring object parts in training images. It is this property that forms the basis of the ROI detection technique. Once the ROI is located from images, the PHoG is used to describe that region of the image. Attributes are then extracted from the descriptors before being concatenated to form a feature vector. This process takes advantage of the abstraction provided by the feature vector representation of input data that enables the use of numerous domain-independent classifiers.

6.3.2 Algorithm A – Single Frequent Keypoint Image Patch Selection

For the first method, the image is divided into smaller patches. Each small patch is then tested for the number of single frequent keypoints existing within it. If the number of frequent keypoints found is greater than the parameter X ($X = 3$), then the patch is considered to be informative and will be kept. Otherwise, the image patch will be discarded by blanking it out. Figure 6.2(a) illustrates an example of image patch selection method. The size of the image patches is determined in the same way as the spatial pyramids scheme [81], where the image is divided into 4, 16, and 64 blocks. Figure 3.8 illustrates an example of the spatial pyramid scheme. Based on cross validation evaluation on the two datasets, it is determined that the image should be divided into 16 blocks and $X = 3$, as these parameters consistently give better recognition performance.

6.3.3 Algorithm B – Single Frequent Keypoint Bounding Box

The second method creates a bounding box around the ROI. Everything inside the box is kept for feature extraction, while everything else will be discarded. Figure 6.2(b) depicts examples of the single frequent keypoint bounding box method.

The placement of the bounding box is defined by the centre-of-mass of all the frequent keypoints found in the image. Essentially, the bounding box is the smallest possible box determined by the occurrence of the frequent keypoints. This approach is made possible because the majority of the frequent keypoints are either on or near the object of interest. Figure 5.1 illustrates an example

comparison between frequent keypoints and all keypoints.

Initially, the size of the bounding box is fixed at 60% of the original image size. However, because the size of objects varies greatly across different images, a simple method is developed, depicted in Algorithm 6.1, to automate the bounding box size selection. The available sizes are 15%, 30% and 60% of the original image size. See Figure 6.3 for an example comparison between fixed and variable bounding boxes.

Algorithm 6.1 The single keypoint bounding box method

Input: An image and its set of frequent keypoints.

1: Start with the smallest bounding box, which is at 15% of the original image size. Place the box around the centre-of-mass of the frequent keypoints.
2: **if** the number of keypoints found inside of this area is greater than or equal to 90% of the total number of keypoints found in the image **then**
3: The current bounding box will be the final bounding box. Exit Algorithm.
4: **else**
5: Increase the size of bounding box.
6: **end if**
7: Repeat step 2 until no more bigger bounding boxes to test.

Output: An image with a bounding box

6.3.4 Algorithm C – Pairs of Frequent Keypoint Patch Selection

The third method is similar to the first, namely, the image is divided into smaller patches and the total number of frequent keypoint pairs existing in the patch

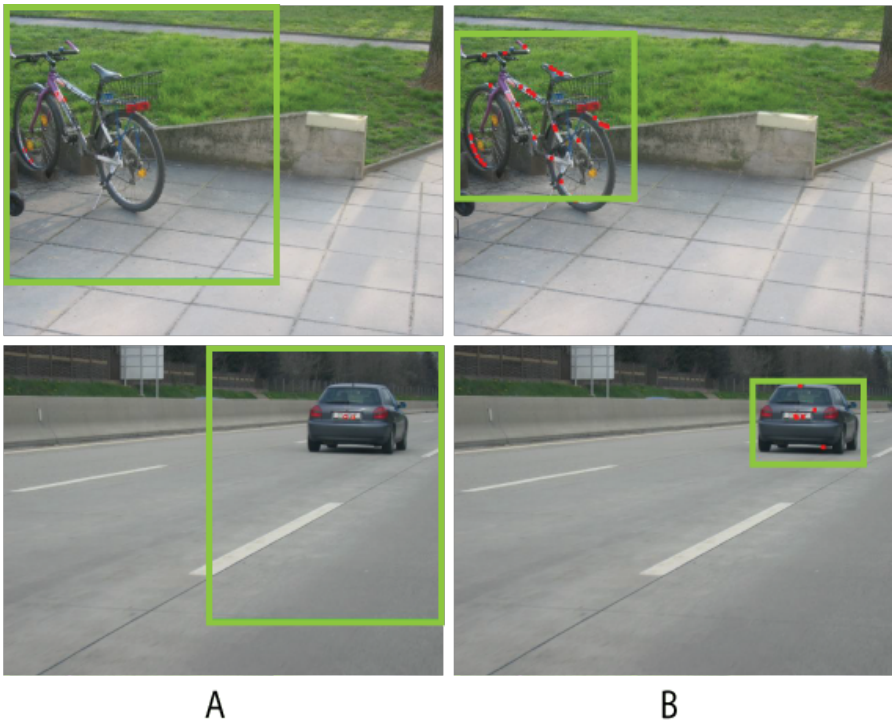

Figure 6.3: Difference between manual (A, 60%) and automatic(B) bounding box size selection.

determines whether that patch is kept or discarded. The only difference to the first method is that pairs of frequent keypoints are used, instead of single frequent keypoints.

It is argued that single frequent keypoints are not unique enough to independently represent object classes by themselves. However, pairs of frequently occurring keypoints are more distinctive and efficient in representing object classes, due to the spatial properties they carry. Instead of computing the pairs of frequent keypoints from all available keypoints, only frequent keypoints are used. This approach significantly reduces background clutter and image noise. Algorithm 6.2 shows the steps in frequent keypoint pairs generation.

Algorithm 6.2 The pairs of frequent keypoints discovery method.

Input: An image and its set of frequent keypoints, pre-defined radius (R), number of pair (X)

1: Traverse through all frequent keypoints extracted from each training image.
2: Determine all the frequent keypoints that are within a pre-defined radius (R) of the currently selected frequent keypoint.
3: Generate unique pairs of frequent keypoint patterns from the set of selected frequent keypoints from step 2. Algorithm 5.2 details how frequent keypoint patterns are generated.
4: Frequent keypoint pairs are ranked from the most frequent to less frequent.
5: Only the top X pairs are selected for each object class.

Output: An image with a bounding box

6.3.5 Algorithm D – Pairs of Frequent Keypoint Bounding Box

The final method is similar to the second method, in that a bounding box of variable size is placed on the image. The only difference is that pairs of frequent keypoints are used to determine the placement of the bounding box, instead of the single frequent keypoints.

6.4 Evaluation

The experiment setup and results are reported in this section. Multi-class classification is done with quadratic SVM classifier and the SMO learning algorithm, with a default parameters as specified in WEKA V. 3.5.5 [167] with the χ^2 kernel.

Using the frequent keypoints selection method proposed in Algorithm 5.1, the top five clusters in each class are tagged as frequently occurring features. Based on the selected frequent keypoints, frequent keypoint patterns are formed and the top 50 patterns from each object class are selected. Various different radius sizes ranging from ten pixels to the entire image were evaluated for discovering patterns, and the radius of 50 pixels was found to be the optimal size for both the datasets. Figure 6.2(d) illustrates the example output produced by this method. For Algorithms 6.1 and 6.2, the parameter X, which is the threshold for the number of keypoints per image patch, is set at 3.

In order to see the effects of ROI detection in terms of overall performance for each of the datasets, we first experimented with the PHoG descriptor on the whole image, before applying the PHoG descriptor on ROI only. Moreover, since frequent keypoints that determine the ROI are class specific, ROI detection is applied on training images only and test images are not altered in any manner. This is because frequent keypoints discovered in, for example, the Bike class, are only used to locate ROI for images belonging to the Bike class, taking advantage of the object class information. However, for test images, where the object class is unknown, frequent keypoints cannot be used to determine the ROI of test images.

This section is divided into three parts: datasets, experiments and discussion. In the datasets section, the two datasets used for the evaluations are presented. In the experiments section, the experimental setup is explained. Finally, results from experiments are discussed in the final section.

Table 6.1: Results for the Airplane class.

No ROI	Algorithm A	Algorithm B	Algorithm C	Algorithm D
77.23%±1.1	77.45%±1.6	79.45%±1.3	76.45%±1.7	**83.23%**±1.3

Table 6.2: Results for the Boat class.

No ROI	Algorithm A	Algorithm B	Algorithm C	Algorithm D
55.94%±1.2	56.13%±1.3	56.69%±1.3	54.23%±1.7	**61.87%**±1.4

6.4.1 Datasets

The proposed ROI detection methods are evaluated on the Airplane, Boat, and Bike classes of the VOC2008 [35] dataset. See Section 4.5 for image samples of this dataset.

6.4.2 Experiments

For this dataset, the experimental setup is 50 images per class used for training and 50 images per class tagged as test images. Experiments are repeated ten times with randomly selected training and testing images. See Table 6.1 for results produced from the proposed methods for Airplane class, Table 6.2 for the Boat class, and Table 6.3 for the Bike class.

Table 6.3: Results for the Bike class.

No ROI	Algorithm A	Algorithm B	Algorithm C	Algorithm D
66.33%±1.1	66.62%±1.2	68.17%±1.4	63.44%±1.0	**68.43%**±1.4

6.4.3 Discussion

Results obtained from experiments on the three classes were comparable to those of the original authors. Consistently, the bounding box methods obtained better results than the patch selection method. It is hypothesized that the main reason for the difference in the two performances, is that with the patch selection method, often only parts of the objects are selected because some parts of the object do not contain any keypoints. However, for the bounding box methods, the algorithms were able to capture the object as a whole more frequently. It has also been shown that the pairs of keypoints bounding box method was able to home in on the object of interest more accurately than the single keypoint bounding box method, resulting in better overall performances.

Specifically, the Airplane and Boat classes realized a performance increase of about 5 to 6%, compared to an increase of about 2% of the Bike class. One possible reason for the difference in performance, could be attributed to background clutter of the images. For both Airplane and Boat classes, the majority of the images have a uniform background texture, for example the sky and the sea, which make ROI detection easier, whereas for the Bike class, the difference between the object of interest and background are not as significant, resulting in smaller performance gains.

It is important to note that the proposed ROI detection techniques are limited only to images containing only single objects. However, it is possible to detect multiple objects provided that the frequent keypoints or patterns are discovered for the interested object classes. The detection of multiple objects can then be done by identifying the ROI for each of the known objects first, before combining

all the ROIs.

6.5 Conclusion

In this chapter, four algorithms for detecting ROI for better object categorization performance are proposed. The proposed methods were evaluated on one of the most popular datasets, with promising results. Unlike the popular sliding window approaches, where classifiers have to be evaluated over a large set of candidate sub-images, the methods described here rely only on detecting frequently occurring keypoints for locating the ROI.

Chapter 7

Capturing Spatial Information with Pairs and Shapes Frequency

In spite of the simplicity and good performance of the BOW model [24], one of its weaknesses lies in the fact that spatial information between image features is not explicitly represented. Much work has been proposed over the years to improve the BOW model, where the spatial pyramid matching [81] technique is the most notable. In this work, two novel techniques are proposed to capture more refined spatial information between image features than that provided by the spatial pyramids. The proposed techniques demonstrate a performance gain over the Spatial Pyramid representation of the BOW model.

7.1 Overview

It is argued that this information will help in better understanding how object parts are related to each other, and in theory enable classifiers to better discriminate

object categories from each other. It is assumed that objects belonging to the same category exhibit significant regularity in their geometry, and that this information can and should be incorporated into object recognition systems.

In this chapter, two novel extensions to the SPM approach are proposed. More precisely, two techniques for capturing spatial information based on the BOW model are introduced: *pairs frequency histograms* and *shapes frequency histograms* of image features. Furthermore, various combinations of spatial and feature frequency information are experimented with. The reasoning is that because the captured spatial information is based on image labels, it should be complementary to the original frequency histogram of words, as it captures different types of dependencies [170].

Since the descriptions of the BOW model and the spatial pyramid matching scheme are already given in detail in Sections 3.1 and 3.2.5, respectively, the rest of the chapter is organized as follows. The proposed algorithms are explained in the following section, followed by a section presenting the datasets and experimental results. Finally, the chapter is concluded with a discussion and conclusion section.

7.2 New Methods for Capturing Geometrical Information

In this section, methods for exploiting and capturing geometrical information between image features are first described. Because the proposed algorithms are built on the visual words from the BOW model, it is important to first explain how these visual words are obtained in detail. To this end, before introducing

the proposed algorithms, the required preprocessing steps in order to produce the codebook are explained.

7.2.1 Preprocessing from SIFT Keypoints to Visual Dictionary

Recall that there are two categories of approaches in sampling areas of interest from images: scale invariant detectors and dense sampling. For this work, the second approach is used. The reasoning for this is two-fold. Firstly, scale invariant detectors are not known to be good at capturing uniform information such as sea, sky or flat surfaces – information that is essential for this work. Secondly, research by Fei-Fei *et al.* [86] found that dense features work better for scene classification and that random sampling of keypoints work nearly as well as keypoints selected by detectors [113]. We therefore take the dense sampling approach here. Moreover, while the proposed frequent keypoint discovery is effective for discovering frequently occurring keypoints in a single object class, for the task of grouping densely sampled image patches from all training images, k-means clustering is still preferred.

A visual codebook is constructed, based on a dense and overlapped grid of 16×16 pixels over the entire image, with a spacing of 8 pixels per grid, is first computed. Lowe's high dimensional SIFT descriptor is then used to describe each of the 16×16 patches, where each descriptor consists of 128-dimensions. K-means clustering is then utilized to group similar image patches (now in SIFT descriptor format) into M bins, where M is the vocabulary size and is set to either 200, 400 or 600.

Figure 7.1: Dense sampling of an image before representing features with an image label grid.

In order to simplify the problem into more intuitive and describable terms, each descriptor is visualized as a label. The label corresponds to the cluster number that the descriptor most closely matches in $L2$ distance. For example, if a patch descriptor most closely matches cluster centre 202, then that patch is replaced with 202. Figure 7.1 is an example of this representation. This is the image representation used by all of the proposed approaches in this chapter.

It is important to note that the process of generating a codebook, is based on k-means clustering, instead of the frequent keypoints approach introduced in earlier chapters. The proposed codebook generation technique, while is fast in finding frequently occurring features to form a codebook, does not scale well for the task of converting SIFT keypoints into a simple number-based image features. Unlike k-means clustering, where the number of "words" for the codebook can be predefined, the frequent keypoint approach allows the codebook size to increase, provided the new feature is not deemed "close" to the previous features.

Due to the way in which spatial features are captured for both the pairs and shapes frequency approaches, a large codebook size results in a huge number of spatial features. For example, with k-means clustering, when k = 200, the size of the final feature vector including spatial features is 8400, independent of the number of object categories. On the other hand, with the frequent keypoint approach, with 50 most frequent keypoints selected from each of 101 object classes, the final feature vector size will be 212,000 (50^{1}, $\times 101^{2} \times 2^{3} \times 21^{4} = 212,000$). The frequent keypoint approach clearly does not scale well for datasets with large number of

[1]Number of frequent keypoints

[2]Number of classes

[3]Pairs/shape spatial information

[4]Number of sub-regions

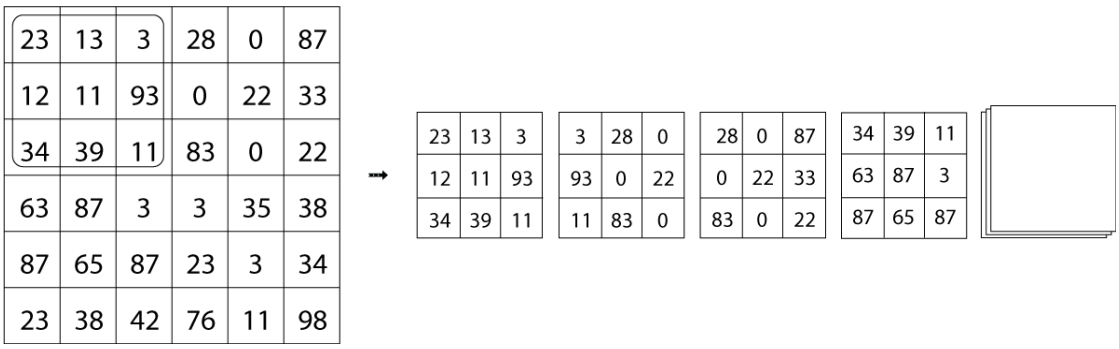

Figure 7.2: A set of overlapped, predefined grids over entire image.

object classes. Because of this, the classic k-means clustering technique is used in this (and the following) chapter for converting SIFT image features into the simple number-based representation.

7.2.2 Approach 1: Pairs Frequency Histogram

The first approach is inspired by the vector space BOW model frequently used in text document representation. After the image is represented by a simple vocabulary of labels of size M (see Figure 7.1), it is possible to apply many successful text mining techniques such as *tf-idf* weighting (term frequency – inverse document frequency) and feature selection [70]. In many aspects, the proposed image representation is semantically similar to text document representation. That is, words convey the meaning of the document just as visual words carry visual characteristics of the image.

To this end, discovering pairs of frequent labels is proposed. Unlike [123], where probabilistic latent space models were used to capture spatial information, the proposed model works by looking for matching labels within a predefined area. This is achieved by first computing predefined grids (overlapped) over the entire

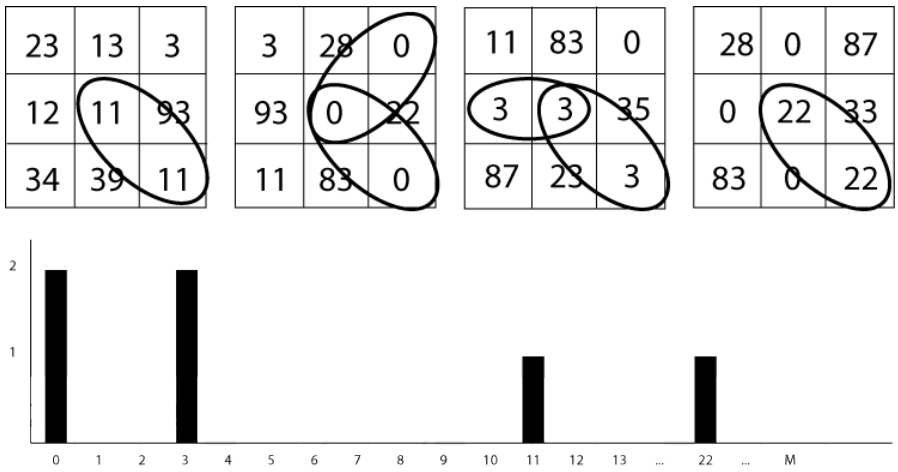

Figure 7.3: Discovering pairs of labels.

image label grid, where the grid size used is 3×3. Figure 7.2 demonstrates how this grid is formed.

For each of the grids, the middle label is used as the *reference label* to compare its neighbouring labels for matches. It was decided to search for pairs of the same label only, because it is simply not feasible to include all possible label combinations. For example, if $M = 200$, then the number of possible combinations of all 200 labels is a $200 \times 200 = 40,000$. However, once SPM is applied, the size of the feature vector will quickly jump to $21 \times 40,000 = 840,000$ dimensions when $L = 2$ (Section 3.2.5 provides a detailed explanation on how SPMs are constructed). Therefore, this work focuses only on matching the occurrences of *reference label* with respect to its neighbours.

Once all pairs are accounted for, a frequency histogram is built on the number of pairs, where the size of the feature vector is the same as M. (See Figure 7.3 for example on how pairs of labels are discovered.)

7.2.3 Approach 2: Shapes Frequency Histogram

The second model focuses on capturing the shapes of image features, leveraging the Local Binary Patterns (LBP) technique [109]. Although LBP was originally adapted for the task of text classification [122], the technique has been proven to be effective for face recognition [2][168][179].

Briefly speaking, in its original form (but not in the proposed approach), an LBP is a property of a pixel. All surrounding pixels in an equally sampled, circular neighbourhood with a certain radius value are examined and a string of binary numbers is constructed such that 1 is given if the neighbour pixel's intensity is greater than the middle, and 0 if the intensity value is equal or less. Only "uniform" bit-strings are considered and assigned to a category specified by the number of 1s in the string. Uniform bit-strings are binary strings with two or less 0 to 1 or 1 to 0 transitions. Consequently, LBP tend to capture curves, peaks, edges and troughs in images [109]. In this approach, it is the LBP shapes formed by the labels and not the pixels that we are most interested in.

Here, the standard LBP approach is modified to treat neighbouring labels as pixels, as well as converting the 8 bit binary string into a decimal number. For example, $00000011 = 3$. This enables the assignment of all possible shapes within the 3×3 grid to only 256 different bins, which can then be turned into a frequency histogram when this is applied to the entire image label grid. Figure 7.4 depicts an example.

One advantage of this approach is that it is not limited to shapes formed by any particular features. Instead, because only the middle label is used as the *reference label* to compare its neighbouring labels, this enables the capturing of

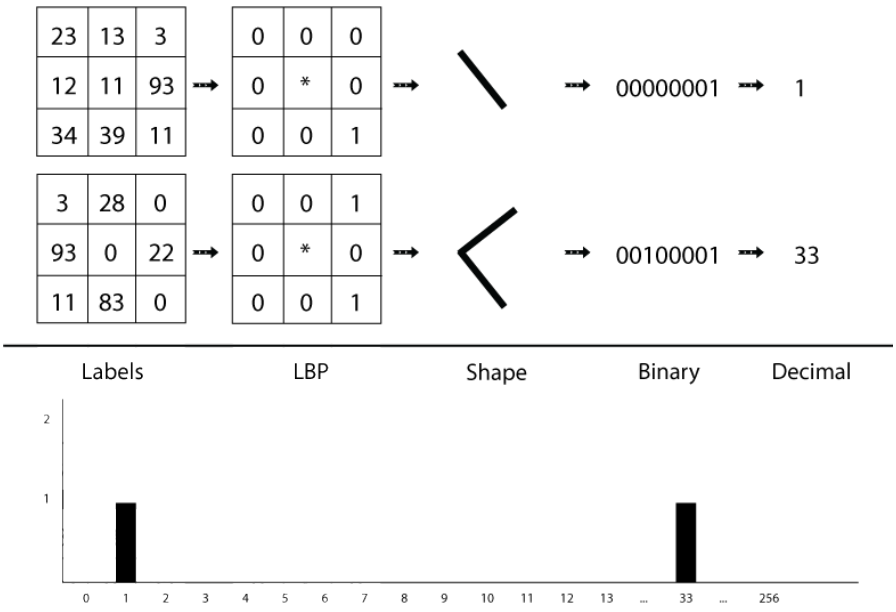

Figure 7.4: Representing shapes with LBP-like approach. A 256 dimension histogram is used to capture the frequency of all possible shapes from image labels.

shapes formed by any labels. In addition, with the original LBP approach, if the neighbour's intensity exceeds that of the middle pixel, 1 will be assigned to the neighbour, and 0 otherwise. This convention was not adopted because even though labels are represented by a number, they are not related in any form. For example, label 100 is not greater than label 2, as the labels represent different types of features rather than pixel intensities. Instead, only labels around the *reference label* are searched for matches, since only matching labels are related meaningfully.

7.3 Evaluation

In this section, datasets used to evaluate the proposed algorithms are first described, followed by a section describing the performed experiments, and the re-

sults are given.

7.3.1 Datasets

The proposed algorithms were evaluated on three popular datasets: Caltech101 (Section 4.1), Graz02 (Section 4.2) and MIT 15 Scenes (Section 4.3).

7.3.2 Methods

The experiment setup and results are reported in this section. Multi-class classification is done with the SVM classifier and the SMO learning algorithm, with default parameters as specified in WEKA V.3.5.5 [167]. All experiments are repeated ten times with different randomly selected training and testing images.

7.3.3 Experimental Results

The final result is reported as the mean and standard deviation accuracy of the individual runs. Experimental results using only the proposed models are first shown. Following this, results from combining the proposed models with the original frequency histogram and SPM are given.

7.3.4 Discussion

For such an elegant and simple attempt at capturing spatial information, the pairs frequency method is fairly effective across all three datasets. When combined with BOW frequency histogram, considerable improvements over the original BOW work were repeatedly achieved. This method is fundamentally the same as the BOW frequency histogram; however, it differs in what it tries to capture. Instead

Table 7.1: Results for Caltech101, the proposed methods combined with original SPM.

Spatial Pyramid Matching (SPM)	L = 2
BOW Baseline, M = 200	54.90%
Pairs Frequency + SPM, M = 200	52.49% ±0.9
Pairs Frequency + SPM, M = 400	**54.44%** ±0.8
Pairs Frequency + SPM, M = 600	54.36% ±1.1
Shapes Frequency + SPM, M = 200	53.68% ±0.9
Shapes Frequency + SPM, M = 400	**53.82%** ±0.9
Shapes Frequency + SPM, M = 600	53.55% ±1.0

of single features, this method counts the frequency of pairs of features occurring at a close proximity.

The shapes frequency method, on the other hand, did not perform as well as the pairs methods, usually underperforming BOW by a few percent. The motivation behind this approach was to capture the shape of features, utilizing the LBP scheme. The main reason for the poor performance may be because there are only 256 bins, not $M \times 256$ bins, used to represent all possible shapes, so there is no information about what the pattern is, specific to M. Another reason for the poor performance may be the size of image patches and codebook. The image patch size is 16×16 for this work, which may be too large to capture unique image features for the LBP method to take advantage of. The other issue is the codebook size. Since M is relatively small, too many dissimilar image features might have been treated as the same. This is a major disadvantage for 'strict' edge-capturing methods like LBP.

Table 7.2: Results for MIT 15 Scenes, the proposed methods combined with original SPM.

Spatial Pyramid Matching (SPM)	L = 2
BOW Baseline	79.4%
Pairs Frequency + SPM, M = 200	80.93% ±1.1
Pairs Frequency + SPM, M = 400	**81.54%** ±1.3
Pairs Frequency + SPM, M = 600	80.56% ±1.1
Shapes Frequency + SPM, M = 200	77.3%±0.9
Shapes Frequency + SPM, M = 400	**78.23%** ±1.5
Shapes Frequency + SPM, M = 600	77.45%±1.1

Table 7.3: Results for GRAZ-02 (Bike), the proposed methods combined with original SPM.

Spatial Pyramid Matching (SPM)	L = 2
BOW Baseline	66.34%
Pairs Frequency + SPM, M = 200	70.11% ±1.8
Pairs Frequency + SPM, M = 400	**71.49%** ±1.6
Pairs Frequency + SPM, M = 600	70.1% ±1.7
Shapes Frequency + BOW, M = 200	65.92% ±1.9
Shapes Frequency + BOW, M = 400	**65.11%** ±1.7
Shapes Frequency + BOW, M = 600	64.12% ±1.6

7.4 Conclusion

The goal in this work is to capture geometric information between image features, thus improving the bag-of-words model for object recognition. To this end, two novel spatial information capturing approaches were proposed: pairs frequency and shapes frequency.

Both the pairs frequency models, when combined with the BOW model, have outperformed the original BOW method by approximately 2 to 3% across three diverse datasets. The LBP representation of the shapes frequency, however, did not perform as well.

In [81], Lazebnik *et al.* found that their spatial pyramid matching scheme is most effective when $M = 200$. Although they tried different codebook sizes, they did not report any performance gains.

For the proposed methods in this chapter, across all three datasets, experimental results consistently found that the proposed methods work best when $M = 400$. Perhaps the main reason is that if the codebook size is small, too many unrelated patches will be grouped together, and if the codebook size is large, then similar features will not be seen as the same.

The proposed approaches, though similar to earlier work by Saverse *et al.* [132] and Wang *et al.* [164], are different in many respects. The correlograms approach proposed by Savarese *et al.* captures the distribution of distances between all pairs of image features. These measurements are then used for classification tasks. Interestingly, in their paper, correlograms perform much worse than the standard BOW model. In comparison, the pairs-of-feature approach captures spatial information between image features, which is more reliable and efficient.

In Wang *et al.* [164], quite a different approach to that described in this chapter is used, where they represent objects using histograms of oriented gradients. These histograms incorporate detailed spatial distributions of object colour across different parts of the object. However, this method relies on objects having similar poses and the images being of good quality. It is evident that under more realistic conditions, texture and shape information are either non-existent or unreliable due to low image quality. Moreover, the authors in that paper use low level oriented gradients whereas the proposed approaches use higher image features, in the form of SIFT keypoints. Thus, the effectiveness of this method is unclear for the real world object categorization problem.

Chapter 8

Log-Polar-Based Image Subdivision and Representation

In the previous chapter, two novel approaches for capturing spatial information for the BOW model were presented. It was shown that the pairs frequency approach showed improvements over the popular spatial pyramid matching scheme.

In this chapter, new methods to exploit spatial relationships between image features, based on binned log-polar grids are presented. These new methods work by partitioning the image into grids of different scales and orientations, and computing histograms of local features within each grid. Experimental results show that the proposed approaches lead to performance improvements on three diverse datasets over the SPM scheme.

8.1 Overview

In this chapter, a novel approach for capturing spatial information for the BOW model is proposed. The proposed technique, *binned log-polar histograms* (BLPHs), is based on the binned log-polar representation (BLPR), which was initially developed for shape matching. Unlike the SPM scheme, where a sequence of increasingly coarser grids are placed over the image, BLPHs divide the image into grids of different scales and different orientations. This explicitly captures the distribution of image features both in distance and orientation, so the pairs frequency is extended by adding orientation information. Variations of the proposed model have been evaluated on three diverse datasets: Caltech101, Graz-02 and MIT 15 Scenes. The initial experiments in this chapter lead to the observation that the proposed model outperforms SPM in capturing spatial information. See Chapter 9 for a detailed comparison between all proposed methods.

The BLPH is based on a binned log-polar representation. Belongie *et al.* in [5] first proposed the binned log-polar scheme as a descriptor for the purpose of shape matching. In the original work, a histogram of the distribution of points over relative positions was used as a compact, yet highly discriminative descriptor. Sensitivity to nearby sample points is achieved by binning feature descriptors in log_j polar space, which means that relative to a reference point, the spatial configuration of the entire object can be captured. The descriptor can be applied to greyscale images, but it is very dependent on brightness. Hence, it is more applicable for line drawings.

Broadly speaking, a set of sample points is extracted from the object's inner and outer contours (normally detected using edge detectors), to represent the object's

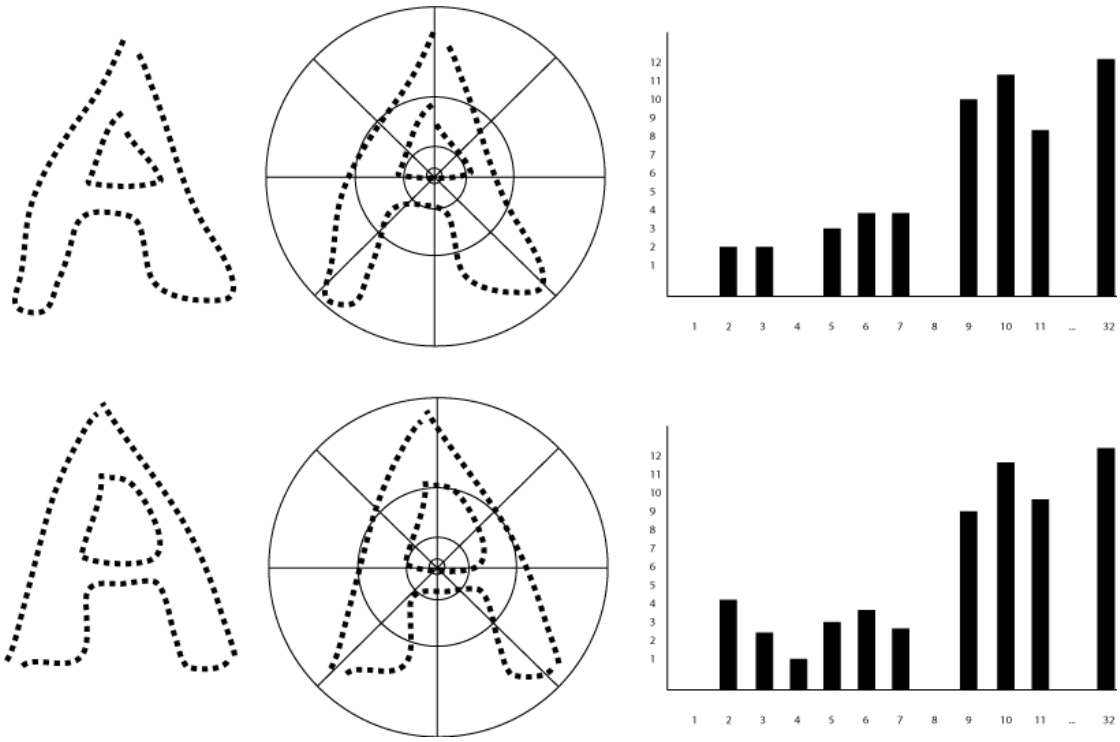

Figure 8.1: Shape matching with log-polar representation.

shape. The technique assumes that objects are sufficiently represented in various sample point configurations and pose, the matching process should be possible. In Figure 8.1, a histogram is used to represent the distributions of sample points in all 32 regions (4 scales and 8 orientations).

The rest of the chapter is organized as follows. The proposed algorithms are explained in the following section, followed immediately by a section presenting the datasets and experimental results. The chapter is concluded with a discussion and conclusion Section.

8.2 Two Methods for Capturing Spatial Information

In this section, methods for exploiting and capturing geometrical information between image features is presented. Similar to the work described in the previous chapter, the proposed algorithms are also built using the BOW model. Section 7.2.1 previously described the image pixel to image label conversion step, so only the proposed algorithms will the presented next.

8.2.1 Method 1: Log-Polar Shapes

Once the image is converted and represented by labels, the proposed algorithms are applied directly on top of this new representation. The first type of method focuses on capturing the distribution of image features using the BLPR.

In the original shape-matching binned log-polar representation, edges are first detected from objects, and these edges are then converted into points. A binned log-polar descriptor is used to describe the distribution of these points in 2D space. For this work, image feature labels are treated as points and the binned log-polar representation is utilized to capture the spatial relationships between all labels. However, the proposed algorithm distinguishs the types of points since each label represents a different visual pattern.

To this end, for every label in the codebook, the algorithm searches for the same label from all of the regions within the log-polar representation, where the distribution of labels is characterized with a histogram. Figure 8.2 shows an example of the proposed log-polar representation. After computing the distribution

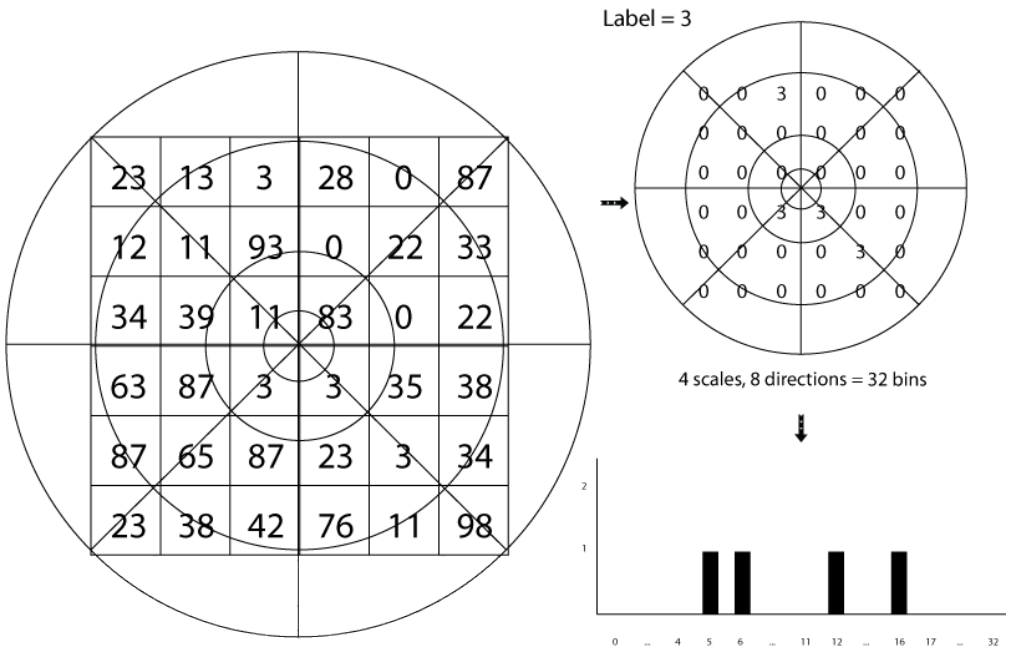

Figure 8.2: Log-polar label histogram representation.

for every label, all histograms are simply concatenated (one histogram per label) to form a large single feature vector, where the histogram size is $M \times 32$, where the size of the codebook M, is 200, 400 or 600.

The rationale for this approach is that, for example, if *label 3* represents an image patch depicting the wheel of a car, by looking for the same label across the entire image, it will be possible to see other occurrences of the same image patch, in this case, the wheel of a car. It is important to note that this is applied to all of the labels in the codebook.

The benefits of this representation are twofold. First, it results in a compact, yet discriminative descriptor for each image feature (label). And second, the representation accounts for increasing positional uncertainty with distance from the point of origin, which is an important component for capturing spatial information.

One limitation of the proposed single log-polar representation is that the centre of the log-polar grid is always located in the middle of the image. However, in many instances, the object of interest is not always located in the middle of the image, therefore, the object might not be represented properly. In order to improve on this, the single log-polar approach is further extended by having multiple log-polar grids (five in total) in the image. They are located in the middle and also the four corners of the image, to better capture the distribution of image features of objects. Finally, histograms from each log-polar grid are simply concatenated to form a large feature vector of size $5 \times 32 \times M$.

Lastly, objects can be of different sizes when depicted in images, which means that the fixed size log-polar approach will not be sufficient in representing all objects. To solve this problem, the multiple log-polar grids approach was modified by including multiple multi-scaled log-polar grids over the image, in order to account for objects of different sizes. This extension is similar to the SPM approach, Figure 8.3 illustrates an example of our multi-scaled approach.

8.2.2 Method 2: Log-Polar Histogram

The second proposed method focuses on characterizing the distribution of all image feature labels within each of the cells. Similar to the previous approach described in section 8.2.1, a binned log-polar representation is mapped onto the label representation of the image, then for each of the grids, a histogram with size M is computed. This approach is similar to the SPM scheme, where the image is divided into smaller sub-regions and the distribution of image features is then characterized with a histogram. Figure 8.4 illustrates an example of this approach.

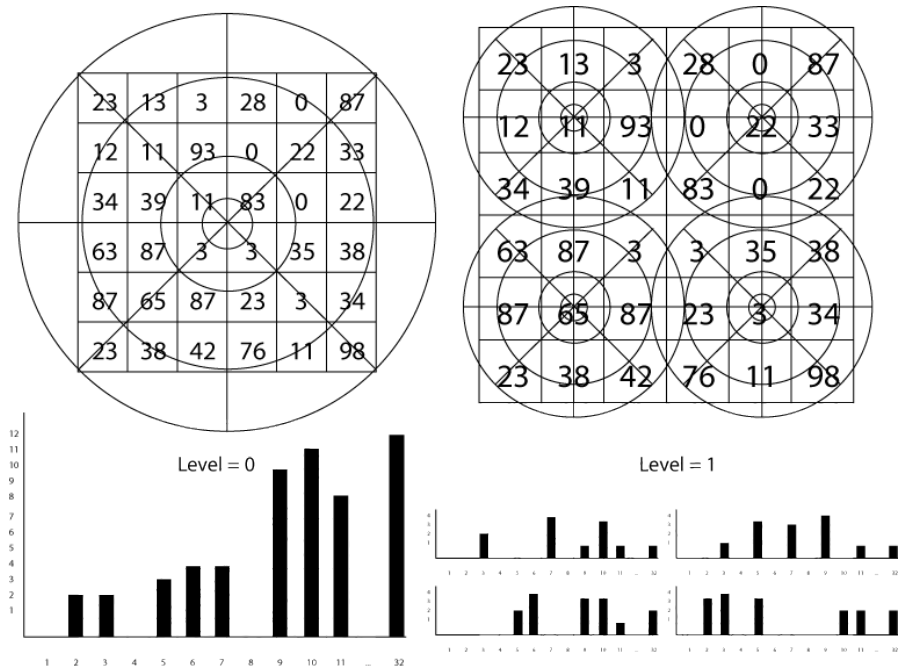

Figure 8.3: Multiple multi-scaled log-polar grids.

The difference of this approach, compared to SPM, lies in the way sub-regions are defined. Unlike the original SPM scheme, the size of sub-regions can vary greatly, depending on their distance from the the centre point. Regions that are closer to the centre contains fewer labels, while regions further away contain significantly more labels. This implicitly accounts for increasing positional uncertainty with distance from the point of origin, and hence captures uncertain spatial relationships.

Similar to the previous proposed methods, this approach is also further extended to include both multiple log-polar and multiple multi-scaled log-polar representation to account for variation in object location and size.

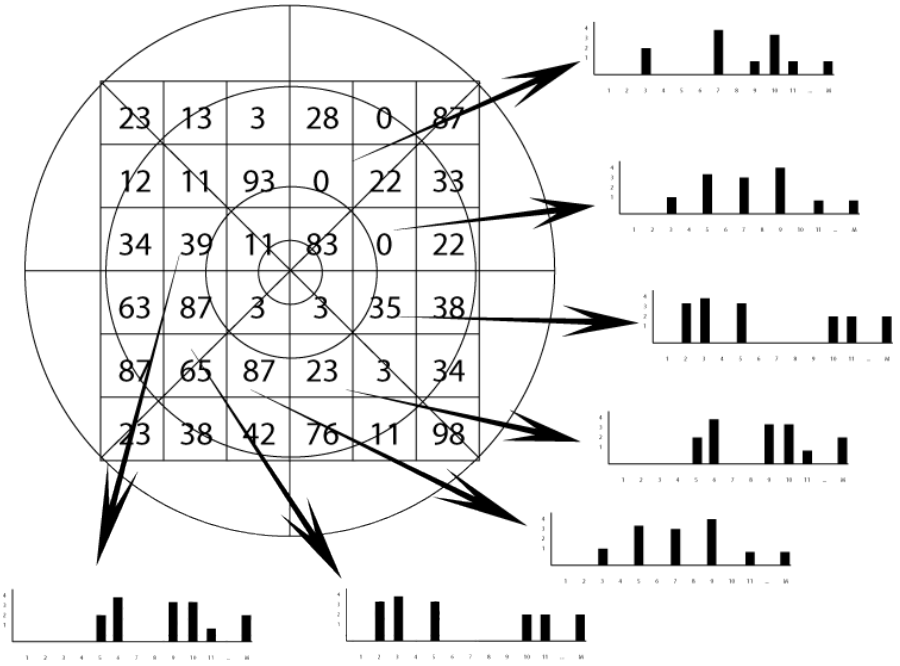

Figure 8.4: Binned log-polar histogram representation.

8.3 Evaluation

In this section, datasets used to evaluate the proposed algorithms are first described, followed by a section describing the experiments performed and the results are given.

8.3.1 Datasets

The proposed algorithms were evaluated on three popular datasets: Caltech101 (Section 4.1), Graz02 (Section 4.2) and MIT 15 Scenes (Section 4.3).

8.3.2 Methods

The experiment setup and results are reported in this section. Multi-class classification is done with an SVM classifier and the SMO learning algorithm, with default parameters as specified in WEKA V.3.5.5 [167]. All experiments are repeated 10 times with different randomly selected training and testing splits. The final performance is reported as the mean and standard deviation accuracy of the individual runs. Experiment results using only the proposed models is first shown, then followed with results from combining the proposed models with the original frequency histogram and SPM.

8.3.3 Experimental Results

In Table 8.1, the performance of the SPM scheme is 54.90% for the Caltech101 dataset. Both the proposed single log-polar shapes and histogram representations performed well on this dataset, with accuracy of 54.27% and 57.32% respectively. One of the main reasons why the single log-polar representations worked so well on this dataset is due to the placement of the objects in images – nearly all objects of interest are located in the middle of the image, which is completely covered by log-polar grids. For the shapes approaches, performance is increased by 2 to 3% after either multiple log-polar grids were included, both fixed and different scales. However, such increase in performance did not occur for the histogram-based approaches, instead, it is observed that a performance decrease of about 2%, mainly due to over-fitting.

For the Graz-02 dataset, in Table 8.2, the performance of the SPM model is 69.34% for the bike class, which is fairly poor considering there are only two

Table 8.1: Results for Caltech101, our methods compared with the original SPM.

Spatial Pyramid Matching (SPM)	54.90% ±1.5
Single Log-Polar Shapes	54.27% ±1.3
Multiple Log-Polar Shapes	55.71% ±1.4
Multi-Scaled Log-Polar Shapes	57.28% ±1.4
Single Log-Polar Histogram	**57.32%** ±1.5
Multiple Log-Polar Histogram	56.81% ±1.1
Multi-Scaled Log-Polar Histogram	57.13% ±1.2

classes – bike and background. In this dataset, the object of interest (bikes), is not always located in the middle of the image and it varies greatly in terms of size and appearance. The single log-polar representations performed about the same as the SPM model. However, once multiple log-polar grids were included, a performance increase of 3 to 4% was observed, especially the multi-scaled log-polar representation.

Finally, for the MIT 15 Scenes dataset, the performance of the SPM model is 79.4%, see Table 8.3. All of the proposed approaches yield similar results to the SPM model. The main reason, it is argued, is that unlike objects, there are no repeating shapes to capture in a scene. Since there are no shapes to capture, the proposed log-polar representation is reduced to a normal SPM-like model, in capturing the distribution of image features only.

Table 8.2: Results for the Bike class in Graz-02, our methods compared with the original SPM.

Spatial Pyramid Matching (SPM)	69.34%±1.7
Single Log-Polar Shapes	68.76% ±1.4
Multiple Log-Polar Shapes	72.98% ±1.3
Multi-Scaled Log-Polar Shapes	**73.18**% ±1.3
Single Log-Polar Histogram	67.11% ±1.4
Multiple Log-Polar Histogram	72.78% ±1.5
Multi-Scaled Log-Polar Histogram	73.11% ±1.2

Table 8.3: Results for MIT 15 Scenes, our methods compared with the original SPM.

Spatial Pyramid Matching (SPM)	79.4% ±0.3
Single Log-Polar Shapes	74.5% ±0.8
Multiple Log-Polar Shapes	**79.9%** ±0.5
Multi-Scaled Log-Polar Shapes	79.5% ±0.4
Single Log-Polar Histogram	75.5% ±0.4
Multiple Log-Polar Histogram	79.8% ±0.4
Multi-Scaled Log-Polar Histogram	79.8% ±0.5

8.4 Discussion and Conclusion

Appearance-based methods have been successfully applied recently for the task of object recognition, due to their simplicity and good performance. One of the popular strategies is the BOW model, which represents an image as a collection of orderless local features. In order to incorporate spatial information, one of the most notable models is the spatial pyramid matching scheme.

Ever since the scheme was introduced in 2006, it has been the cornerstone of many successful object recognition models. Over the years, various improvements have been proposed for the SPM scheme, for example, some focus on alternative ways of codebook construction in order to produce a more representative codebook; others focus on new kernels and classification techniques; and others focus on using different or multiple descriptors. However, not much work has been done on how to more effectively capture spatial information directly from images.

Despite the good performance of the SPM model, it is hypothesized that taking the weighted sum of the number of matches that occur in each coarser grid, is not the most effective way of capturing spatial information. This weakness was demonstrated by the low performance of the GRAZ-02 dataset using the SPM model.

In this chapter, two new types of approaches for capturing spatial information based on the binned log-polar representation are proposed. Unlike the SPM model, the proposed models work by partitioning the image into grids of different scales and orientations. Experimental results from three popular datasets using the proposed methods showed significant improvements over the original SPM model.

Chapter 9

Evaluation

This chapter details a thorough evaluation of the Spatial Object Recognition Framework proposed in Chapter 1. The purpose is to determine, whether the entire framework will actually improve on the performances from individual methods alone.

It is often argued that one of the main reasons for the good performance of many state-of-the-art object recognition systems is the reliance on identifying the location of the objects of interest in images, in order to better the visual attributes. In this evaluation, the ROI detection method described in Chapter 6 is first used to locate the ROI in images, before some of the well-established object recognition techniques, such as BOW, SPM, PHoG, as well as techniques proposed in this thesis are evaluated.

This chapter first gives an overview of the framework for combining the propose techniques, followed by experimental results and discussion for each of the datasets.

9.1 Overview

A novel approach for identifying both frequent keypoints and patterns, based on the spatial relationships between SIFT keypoints was introduced in Chapter 5. Based on this technique, four different hybrid ROI detection methods were developed and were detailed in Chapter 6. Because previous experimental results suggest that the pairs of keypoints bounding box selection technique was most effective at detecting ROI from images, the technique was chosen as the sole ROI detection method for this framework.

Once the ROI is identified, various feature extraction methods are then applied to both the ROI and the original image for comparisons. Feature extraction methods include some of the well-established techniques such as BOW, SPM, and the PHoG descriptors, as well as the techniques proposed in this thesis, which are SPM + pairs, SPM + shapes, multi-scaled log-polar shapes, and multi-scaled log-polar histogram.

For the ROI detection technique, the top 20 frequent keypoints from each object class are flagged as frequently occurring features. Based on the selected frequent keypoints, frequent keypoint patterns are formed and the top 20 patterns from each object class are selected. The radius size was set at 50 pixels.

Multi-class classification is done with SVM classifier and the SMO learning algorithm, with default parameters as specified in WEKA V.3.5.5 [167]. All experiments are repeated ten times with different randomly selected training and testing images.

In order to see the difference in performance between ROI and normal images, for each of the datasets, feature extraction is performed on both the ROI and the

original images. ROI detection is applied on training images only and test images are not altered in any manner.

9.2 Caltech101

For the Caltech101 dataset, 20 images per object class are randomly selected for training and the rest are tagged as test images. The final performance is reported as the means and standard deviation accuracy of the individual runs. Experiment results from the original images are shown first, then followed by results from the ROI images. Table 9.1 illustrates the results produced from the methods described.

Table 9.1: Results for the Caltech101 dataset.

Methods	Original Image	ROI Image
BOW	36.45%±1.6	35.46%±1.7
SPM	54.90%±1.5	53.94%±1.5
PHoG (χ^2 kernel)	40.43%±1.2	47.82%±1.3
SPM + Pairs	54.44%±1.2	53.32%±1.3
SPM + Shapes	53.82%±1.4	51.21%±1.6
Multi-scaled Log-polar Shapes	57.28%±1.6	54.49%±1.8
Multi-scaled Log-polar Histogram	**57.32%**±1.5	54.12%±1.6

Both the multi-scaled log-polar shapes, and histogram feature extraction methods performed the best for this dataset, followed closely by the SPM-based methods. All the methods, with the exception of the PHoG descriptors, performed poorly on the ROI images. The PHoG descriptor achieved an increase of about 7% on the ROI images over the original images.

9.3 Graz-02

For the Graz-02 dataset, a training set for each object category consist of 150 images as belonging to the category as positive images, and 150 images of the counter-class as negative images. The same number of images, belonging to the category and half not, were used for evaluation. Table 9.2 illustrates the results produced from the methods described.

Table 9.2: Results for the Bike class in the Graz-02 dataset.

Methods	Original Image	ROI Image
BOW	63.45%±2.3	62.25%±2.1
SPM	66.34%±2.1	66.45%±1.9
PHoG (χ^2 kernel)	75.53%±1.9	**79.42%**±1.9
SPM + Pairs	71.49%±2.5	72.12%±2.5
SPM + Shapes	65.11%±2.1	64.23%±1.9
Multi-scaled Log-polar Shapes	73.18%±2.1	73.56%±2.5
Multi-scaled Log-polar Histogram	73.11%±2.0	72.89%±2.4

The PHoG descriptor in conjunction with the ROI technique from Chapter 6 performed the best for this dataset with an accuracy of close to 80% on ROI images. Unlike the previous Caltech101 dataset, only a small increase in performance was observed from the BOW to SPM approach. The top-performing proposed method is the multi-scaled log-polar shapes technique with an accuracy of 73.18%. Overall, the BOW-based methods did not perform well for this dataset, considering there are only two classes – positive and negative.

9.4 MIT 15 Scenes

The experimental setup of Lazebnik *et al.* [81] is followed. That is, for each of the categories, 100 images are randomly selected for training and the remaining images are flagged as test images. Table 9.3 illustrates the results produced from the methods described.

Table 9.3: Results for the MIT 15 Scenes dataset.

Methods	Original Image	ROI Image
BOW	72.23%	67.45%
SPM	79.49%	74.19%
PHoG (χ^2 kernel)	60.21%	58.90%
SPM + Pairs	**81.54%**	75.33%
SPM + Shapes	78.23%	74.65%
Multi-scaled Log-polar Shapes	79.92%	75.76%
Multi-scaled Log-polar Histogram	79.85%	74.18%

The original BOW and PHoG methods performed poorly on this dataset with recognition accuracy of 67.45% and 58.90% for ROI images, respectively. The spatially-aware methods on the other hand, have achieved similar accuracies. For this dataset, original images in general have outperformed ROI images.

9.5 Moths

For the moths dataset, ten $\times$ ten cross validation is applied to all images from all moth classes. Table 9.4 illustrates the results produced from the methods

described.

Table 9.4: Results for the Moths dataset.

Methods	Original Image	ROI Image
BOW	75.33%	80.22%
SPM	81.10%	**85.40%**
PHoG (χ^2 kernel)	59.56%	68.29%
SPM + Pairs	82.34%	85.30%
SPM + Shapes	80.12%	83.12%
Multi-scaled Log-polar Shapes	77.76%	77.79%
Multi-scaled Log-polar Histogram	84.11%	83.86%

With this dataset, results from the ROI images have nearly all outperformed methods applied on original images. The SPM method, achieving 85.40% on ROI images and 81.10% on original images, was the top performing method for this dataset. Surprisingly, the PHoG method performed poorly and is about 20% behind the SPM method.

9.6 Galaxies

A ten $\times$ ten cross validation was applied to all images from all the galaxy classes. Table 9.5 illustrates the results produced from the methods described.

Due to the unique characteristic of the dataset, all the methods performed fairly similarly. Once again, all the BOW-based methods performed much better than the PHoG method. The top performing method was the SPM method with

Table 9.5: Results for the Galaxies dataset.

Methods	Original Image	ROI Image
BOW	88.37%	87.89%
SPM	**90.30%**	89.23%
PHoG (χ^2 kernel)	81.43%	81.53%
SPM + Pairs	90.12%	88.43%
SPM + Shapes	88.32%	87.71%
Multi-scaled Log-polar Shapes	89.41%	89.12%
Multi-scaled Log-polar Histogram	88.37%	88.78%

an accuracy of 90.30%. Interestingly, there was only a 2% difference between BOW and SPM.

9.7 VOC 2008

Seven classes from the VOC 2008 dataset were selected for the evaluation. For each of the classes, a training set consisting of 150 images of the object category as positive images and 150 of the counter-class (random selection of images from other classes) as negative images were selected. A ten $\times$ ten cross validation was applied to all images from both the positive and negative class. Table 9.6 – 9.12 illustrate the results produced from the methods described for the seven classes.

For the Airplane class (Table 9.6), the top performing methods were both multi-scaled log-polar shapes and histogram, in which both achieved 86.34%. There was no significant improvement in performance on using the ROI images for the BOW-based methods. However, a 5% increase in performance was realised with

Table 9.6: Results for the Airplane class in the VOC2008 dataset.

Methods	Original Image	ROI Image
BOW	76.31%±1.5	78.39%±1.2
SPM	84.31%±1.3	84.37%±1.2
PHoG (χ^2 kernel)	77.23%±1.1	83.23%±1.3
SPM + Pairs	83.29%±1.4	85.23%±1.2
SPM + Shapes	82.12%±1.6	82.87%±1.6
Multi-scaled Log-polar Shapes	86.34%±1.1	85.43%±1.3
Multi-scaled Log-polar Histogram	**86.34%**±1.2	85.21%±1.4

Table 9.7: Results for the Boat class in the VOC2008 dataset.

Methods	Original Image	ROI Image
BOW	77.45%±0.8	77.81%±0.9
SPM	78.95%±0.9	77.23%±1.1
PHoG (χ^2 kernel)	55.94%±1.2	61.87%±1.4
SPM + Pairs	**79.13%**±0.9	78.63%±0.9
SPM + Shapes	75.98%±1.1	74.34%±1.2
Multi-scaled Log-polar Shapes	76.45%±1.2	75.88%±1.1
Multi-scaled Log-polar Histogram	76.77%±1.2	76.02%±1.3

Table 9.8: Results for the Bus class in the VOC2008 dataset.

Methods	Original Image	ROI Image
BOW	76.17%±1.8	76.12%±1.9
SPM	74.91%±1.7	73.54%±1.9
PHoG (χ^2 kernel)	66.63%±2.1	72.43%±2.3
SPM + Pairs	74.23%±1.5	74.56%±1.7
SPM + Shapes	73.91%±1.7	72.12%±1.6
Multi-scaled Log-polar Shapes	**76.89%**±1.7	75.19%±1.8
Multi-scaled Log-polar Histogram	75.25%±1.7	75.34%±1.8

Table 9.9: Results for the Bird class in the VOC2008 dataset.

Methods	Original Image	ROI Image
BOW	60.84%±0.8	58.76%±1.1
SPM	64.12%±1.0	64.32%±1.2
PHoG (χ^2 kernel)	63.86%±1.6	**65.98%**±1.4
SPM + Pairs	64.65%±1.1	63.10%±0.8
SPM + Shapes	61.12%±1.2	59.34%±1.3
Multi-scaled Log-polar Shapes	63.97%±1.0	61.44%±0.8
Multi-scaled Log-polar Histogram	63.92%±1.1	62.56%±0.9

Table 9.10: Results for the Bike class in the VOC2008 dataset.

Methods	Original Image	ROI Image
BOW	67.82%±0.5	65.71%±0.7
SPM	64.14%±0.5	63.23%±0.8
PHoG (χ^2 kernel)	66.33%±1.1	**68.43%**±1.4
SPM + Pairs	63.81%±0.7	63.47%±0.5
SPM + Shapes	62.55%±0.6	60.77%±0.9
Multi-scaled Log-polar Shapes	66.85%±1.0	64.22%±0.9
Multi-scaled Log-polar Histogram	65.61%±0.8	63.79%±1.1

Table 9.11: Results for the Bottle class in the VOC2008 dataset.

Methods	Original Image	ROI Image
BOW	69.78%±1.6	65.12%±1.3
SPM	75.16%±1.4	73.90%±1.2
PHoG (χ^2 kernel)	63.36%±1.7	68.34%±1.5
SPM + Pairs	**75.78%**±1.5	73.90%±1.5
SPM + Shapes	72.75%±1.6	71.59%±1.5
Multi-scaled Log-polar Shapes	72.11%±1.4	71.42%±1.5
Multi-scaled Log-polar Histogram	71.82%±1.3	69.34%±1.4

Table 9.12: Results for the Car class in the VOC2008 dataset.

Methods	Original Image	ROI Image
BOW	72.14%±0.8	72.34%±1.0
SPM	**77.58%**±1.0	76.83%±1.1
PHoG (χ^2 kernel)	71.28%±1.4	75.62%±1.2
SPM + Pairs	76.02%±1.7	75.34%±1.6
SPM + Shapes	74.34%±1.5	74.91%±1.5
Multi-scaled Log-polar Shapes	72.91%±1.4	72.49%±1.3
Multi-scaled Log-polar Histogram	71.82%±1.3	70.12%±1.4

the PHoG method.

For the Boat class (Table 9.7), the SPM + Pairs was the top performing method with an accuracy of 79.15% on the original images. All BOW-based methods achieved similar performances.

For the Bus class (Table 9.8), the multi-scaled log-polar shapes method achieved the top score of 76.89% with the original images and only 75.19% on the ROI images.

All of the methods performed poorly with the Bird class (Table 9.9), where the top performance is only at 65.98% with the PHoG method. There is little difference in performance between ROI and original images.

For the Bike class (Table 9.10), the PHoG technique, as expected, was the top performing method. The PHoG method also was the top performing method for the Bike class in the Graz-02 dataset. A top score of 68.43% was achieved on the ROI with the PHoG descriptor. Surprisingly, the BOW method is actually the

second placed method, ahead of SPM and other spatial-capturing methods.

For the Bottle class (Table 9.11), neither of the log-polar methods performed as well as the SPM-based methods. The top performing method was the SPM + Pairs method which achieved 75.78% on the original images. However, a 2% decrease in performance was recorded on the ROI images.

The last class for the VOC 2008 dataset is the Car (Table 9.12) class. For this class, the SPM method achieved the best performance of 77.58%. It is interesting to note that, again, all of the methods, with the exception of the PHoG method, performed poorly on the ROI images.

9.8 Evaluation Summary

Two tables are presented in this section. Table 9.13 shows the best performing methods for each of the datasets using original images. Table 9.14 shows the best performing methods for each of the datasets using the ROI images.

9.9 Discussion

The evaluation results show that the ROI and Pairs methods are a good techniques for boosting the performance of standard SPM or PHoG. However, these methods seldom boost the performance of the log-polar and other proposed methods – this is simply because the new methods are already outperformed SPM and PHoG methods when they do work. Interestingly, only limited improvement in recognition accuracy was observed when ROI is combined with the proposed methods. The core reason for this, this thesis claims, is not because of the failure of the

ROI detection method, but rather, is due to the characteristic of the BOW-based methods.

For the BOW-based methods, the codebook is constructed from densely sampled image patches using k-means clustering. This means that the background image patches contribute heavily (as much as the object of interest) in the forming the bag of features for that image and class. However, for the PHoG method, which is based on extracting gradient directions from image lines and edges, the background of the image will contribute significantly less because, generally speaking, the background of the image will contain far fewer lines and edges compared to the object of interest. In other words, the background of images will have less impact on the overall distribution of features for both the training and testing images with the PHoG method.

All of the experimented methods, with the exception of the PHoG method, were based on the BOW method. In all of the experiments, the performance of the PHoG method on ROI images was constantly better than the PHoG method on original images, despite the fact that in most datasets, the top performance from the PHoG method still lags far behind the top performing BOW-based methods.

Overall, both the log-polar based methods performed well in many of the datasets where the object of interest is clearly defined, such as the Caltech101, Graz-02, and the Airplane class of the VOC2008 dataset. Where these two methods performed poorly is in objects without clear edges and shapes, such as the Boat class, Bird class, and the Bottle class of the VOC2008 dataset.

Another interesting observation is that the PHoG method performed well for the Bike class in both the Graz-02 and VOC2008 dataset – it outperformed all other methods. One reason for this, this thesis argues, is because the PHoG method is

effective at detecting and extracting lines and edges from the bike parts such as the wheel, frame and handle bar.

Table 9.13: A summary table for evaluation results on original images.

	BOW	SPM	PHoG	SPM + Pairs	SPM + Shapes	LP Shapes	LP Histogram
Caltech101							√
Graz-02			√				
MIT15				√			
Moths							√
Galaxies		√					
VOC-Airplane						√	√
VOC-Boat				√			
VOC-Bus						√	
VOC-Bird		√					
VOC-Bike						√	
VOC-Bottle				√			
VOC-Car		√					

Table 9.14: A summary table of evaluation results on ROI images.

	BOW	SPM	PHoG	SPM + Pairs	SPM + Shapes	LP Shapes	LP Histogram
Caltech101						√	
Graz-02			√				
MIT15						√	
Moths		√					
Galaxies		√					
VOC-Airplane						√	
VOC-Boat				√			
VOC-Bus	√						
VOC-Bird		√					
VOC-Bike			√				
VOC-Bottle		√		√			
VOC-Car		√					

Chapter 10

Conclusions

Computer vision is the study of designing machines that can see and interact with the world through visual information. The foundation of computer vision is extracting abstract information of what is happening in a scene and making sense of it. Indeed, in computer vision research, the ultimate goal is to develop algorithms and tools that will allow a computer to analyze the visual world automatically.

This thesis has dealt with the challenging problem of object recognition using machine learning techniques. While the effectiveness of the Spatial Object Recognition Framework is somewhat limited in some datasets, the proposed components are powerful improvements for the BOW model. Section 10.1 summarizes the main contribution of this thesis, that being the frequent keypoint discovery method, ROI detection, pairs and shapes frequency histogram and binned log-polar representation. Finally, Section 10.2 describes the final framework.

10.1 Summary of the Spatial Object Recognition Framework

This thesis has proposed methods for improving the BOW model, and a Spatial Object Recognition Framework for the object recognition problem. The original BOW model has shown remarkable performance in a wide range of object recognition tasks, in spite of its simplicity. In essence, the model works by representing an image as an orderless collection of local features without any intermediate representation. Intermediate representations are seen as a bridge to reduce the gap between low-level and high-level image processing, therefore better matching computational object models with human perception. The key idea is that images can be represented by different distributions of visual words. A BOW model is then built as a histogram over visual word occurrences.

The main hypothesis of this thesis is that with better spatial information capturing techniques, the BOW model for object recognition can be further improved. Section 1.2 details the steps to test this hypothesis, leading to its validation through the development and evaluation of the proposed contributions (Section 1.3). The four central contributions of this work are: frequent keypoint discovery, region of interest detection, pairs and shapes frequency histogram, and binned log-polar representation.

- **Frequent keypoint discovery.** The first contribution improve recognition accuracy of the BOW model by combining pairs and triplets of frequent keypoints. The key idea behind this approach is to discover intermediate representations for each object class. Broadly speaking, this approach works

by partitioning the image into smaller regions, then computing the spatial relationships between all of the informative keypoints in the region. Experimental results show that the inclusion of these explicit spatial relationships leads to a measurable increase in performance compared to the standard BOW model.

- **Region of interest detection.** Based on the frequent keypoint discovery technique from our first contribution, the second contribution discover the region of interest from images. Many state-of-the-art object recognition systems rely on identifying the location of an object in images, in order to better learn its visual attributes. We show that the region of interest can be efficiently detected using both single and pairs of frequent keypoints. The benefit of our detection technique is validated in two different types of datasets.

- **Pairs and shapes frequency histogram.** One of the disadvantages of the BOW model is that it discards the spatial relationships of local descriptors, which severely limits its descriptive power. One of the most successful solutions to this problem is the spatial pyramid matching scheme. Our third contribution is built on this. In that, we propose techniques to capture more refined spatial information between image features. The techniques are pairs frequency histograms, shapes frequency histograms, and the binned log-polar representation of image features. Furthermore, we also experiment with various combinations of spatial and feature frequency information. Experimental results were encouraging, we argue that this is because the captured spatial information is based on image feature descriptors, they should be comple-

mentary to the original frequency histogram of words, as they capture different types of dependencies.

- **Binned log-polar representation.** For our last contribution, a spatial pyramid matching alternative was proposed, in capturing spatial information for the BOW model. Our proposed techniques, variations of binned log-polar histograms, are based on the bin log-polar representation, which was initially developed for shape matching. Unlike the spatial pyramid model, where a sequence of increasingly coarser grids are placed over the image, our approach divides the image into grids of different scales and different orientations, thus explicitly capturing the distribution of image features both in distance and orientations.

10.2 Spatial Object Recognition Framework

The Spatial Object Recognition Framework consists of the four contributions of this research. Frequent keypoint patterns are first identified for each of the object class (contribution 1), followed by detecting the region-of-interest based on the keypoint patterns (contribution 2). Once the region-of-interested is extracted from the image, and the background discarded, several feature extraction methods are then applied, including PHoG, BOW, SPM, and our proposed methods (contributions 3, 4).

The framework was evaluated on several datasets with mixed results. It generally performed well on datasets where there is not much viewpoint variation between object to object within the same class. For example, the Graz-02 and Moths dataset.

10.3 Conclusion

One of the underlying assumptions of the BOW model is that images can be, based on the codebook, represented by a distribution of viewpoint-dependent image features. This assumption holds because all of the difficult work, such as determining how closely two sets of features relate, is handled by powerful machine learning techniques such as SVM. Indeed, we argue that good performances achieved in object recognition in the past decade owe more to the advancement in machine learning algorithms and powerful CPUs, than improved object recognition theories.

In this thesis, our work is based on this assumption – we ignore the actual physical structure of objects and disregard how object parts related to each other. We leverage the power of modern machine learning techniques by training classifiers on a large number of viewpoint-dependent image features and hope that during testing, the classifier has already seen/learned the extracted features from test images, in order to make an informed prediction. This is precisely the reason why the BOW method worked so well on datasets such as the Caltech101 dataset where images belonging to the same object class share so many similarities, and not as effective on real world datasets such as the Graz-02 and VOC 2008 where images belonging to the same object class exhibit significant differences in scale, orientation and appearance.

Moreover, we claim that view-dependent approaches are nearing the end of their current cycle. There is no doubt that they will come back in the future, in a different form; however, given the limitations of current hardware and learning algorithms, a different type of object recognition approach must be conceived, in

order to tackle real world problems.

Intuitions and insights from this research suggest that in order to take object recognition to the next level, two immediate problems need to be addressed – visual representation and context-awareness.

- **Visual representation.** The fundamental assumption of the viewpoint-dependent approach (such as the BOW model), is that an object class can be represented by a large number of 2D exemplars of objects belonging to the same class. This assumption has served the object recognition community successfully in the past decade, contributing significantly to the overall improvement in the research field. The weakness of this assumption is that when a test object is completely different to any of the learned/trained object exemplars, either in scale, orientation, or shape, the classifier will not be able to make a confident prediction. A more ideal and intuitive approach would be to first build a 3D model of the object that is based on the 2D images, before training a classifier on the 3D object representation.

- **Context-awareness.** After the object is represented by 3D parts, it is important to introduce a context-awareness concept that is able to link all of the parts together, forming the object as a whole. For example, a bike and a chair might share similar part; however, it is their spatial configuration or the context that differentiate the two objects.

Once the challenges above are met, object recognition can then be truly invariant to viewpoint changes because the projected shape of an object can be accurately predicted based on the known perspective projection. This ability is

an important stepping stone in building machine vision that will one day, rival human vision.

References

[1] Ankur Agarwal and Bill Triggs. Hyperfeatures - multilevel local coding for visual recognition. In *In ECCV*, pages 30–43. Springer, 2006.

[2] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:2037–2041, December 2006.

[3] A. P. Ashbrook, N. A. Thacker, P. I. Rockett, and C. I. Brown. Robust recognition of scaled shapes using pairwise geometric histograms. In *BMVC '95: Proceedings of the 6th British conference on Machine vision (Vol. 2)*, pages 503–512, Surrey, UK, UK, 1995. BMVA Press.

[4] A. Baumberg. Reliable feature matching across widely separated views. *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 1:774–781 vol.1, 2000.

[5] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002.

[6] Serge Belongie, Chad Carson, Hayit Greenspan, and Jitendra Malik. Color- and texture-based image segmentation using em and its application to content-based image retrieval. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, page 675, Washington, DC, USA, 1998. IEEE Computer Society.

[7] Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. Retrieval by shape

similarity with perceptual distance and effective indexing. *IEEE TRANSACTIONS ON MULTIMEDIA*, 2(4):225–239, 2000.

[8] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *ICDT '99: Proceedings of the 7th International Conference on Database Theory*, pages 217–235, London, UK, 1999. Springer-Verlag.

[9] I. Biederman, E. E. Cooper, J. E. Hummel, and J. Fiser. Geon theory as an account of shape recognition in mind, brain, and machine. In *J. Illingworth (Ed.) Proceedings of the Fourth British Machine Vision Conference, 1*, pages 175–186. Surrey, U.K.: BMVA Press, 1993.

[10] I. Biederman and P. C. Gerhardstein. Recognizing depth-rotated objects: Evidence and conditions for 3d viewpoint invariance. In *Journal of Experimental Psychology: Human Perception and Performance, 19*, pages 1162–1182, 1993.

[11] Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[12] Elie Bienenstock, Stuart Geman, and Daniel Potter. Compositionality, mdl priors, and object recognition. In *Neural Information Processing Systems*, pages 838–844. MIT Press, 1997.

[13] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1063–1074, 2003.

[14] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Image classification using random forests and ferns. pages 1–8, 2007.

[15] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408, New York, NY, USA, 2007. ACM.

[16] J Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.

[17] Glenn L. Cash and Mehdi Hatamian. Optical character recognition by the method of moments. *Comput. Vision Graph. Image Process.*, 39(3):291–310, 1987.

[18] T. Chang and C. Jay Kuo. Texture analysis and classification with tree-structured wavelet transform. *IEEE trans. Image Proc.*, 2(4):429–441, 1993.

[19] O. Chapelle, P. Haffner, and V. N. Vapnik. Support vector machines for histogram-based image classification. volume 10, pages 1055–1064, 1999.

[20] Chaur-Chin Chen. Improved moment invariants for shape discrimination. *Pattern Recognition*, 26(5):683 – 686, 1993.

[21] Tat Seng Chua, Kian-Lee Tan, and Beng Chin Ooi. Fast signature-based color-spatial image retrieval. In *ICMCS '97: Proceedings of the 1997 International Conference on Multimedia Computing and Systems*, page 362, Washington, DC, USA, 1997. IEEE Computer Society.

[22] O. Chum and A. Zisserman. An exemplar model for learning object classes. pages 1–8, 2007.

[23] P. Cohen, C.T. LeDinh, and V. Lacasse. Classification of natural textures by means of two-dimensional orthogonal masks. In *Acoustics, Speech and Signal Processing, IEEE Transactions on*, volume 37(1), pages 125–128, 1989.

[24] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cdric Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[25] Y. Le Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson. Handwritten digit recognition with a back-propagation network. pages 396–404, 1990.

[26] C. Cyr and B. Kimia. 3d object recognition using shape similiarity-based aspect graph. In *Proceedings of the International Conference on Computer Vision*, pages 254–261. Vancouver, Canada, 2001.

[27] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

[28] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

[29] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008.

[30] Sven Dickinson, Ales Leonardis, Bernt Schiele, Michael Tarr, Shimon Edelman, and Paul Valry. Object categorization: Computer and human vision perspectives edited by, 2009.

[31] O. Duda, P.E. Hart, and D.G. Stock. Pattern classification. In *John Wiley and Sons*, 2000.

[32] Sahibsingh A. Dudani, Kenneth J. Breeding, and Robert B. McGhee. Aircraft identification by moment invariants. *Computers, IEEE Transactions on*, C-26(1):39–46, 1977.

[33] Shimon Edelman. Computational theories of object recognition. In *Trends in Cognitive Science*, pages 296–304, 1997.

[34] W. Eric, W. Eric L. Grimson, and Daniel P. Huttenlocher. On the sensitivity of the hough transform for object recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12, 1990.

[35] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[36] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *J. Intell. Inf. Syst.*, 3(3-4):231–262, 1994.

[37] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, 2007.

[38] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1):55–79, 2005.

[39] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(1):36–51, 2008.

[40] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Comput.*, 22(1):67–92, 1973.

[41] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content: The qbic system. *Computer*, 28(9):23–32, 1995.

[42] D. Forsyth and A. Zisserman. Shape from shading in the light of mutual illumination. *Image Vision Comput.*, 8(1):42–49, 1990.

[43] William T. Freeman and Edward H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(9):891–906, 1991.

[44] A. J. Goldstein, L. D. Harmon, and A. B. Lesk. Identification of human faces. *Proceedings of the IEEE*, 59(5):748–760, 1971.

[45] Calvin C. Gotlieb and Herbert E. Kreyszig. Texture descriptors based on co-occurrence matrices. *Comput. Vision Graph. Image Process.*, 51(1):70–86, 1990.

[46] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1458–1465 Vol. 2, October 2005.

[47] M.H. Gross, R. Koch, L. Lippert, and A. Dreger. Interpreting pictures of polyhedral scenes. In *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, volume 3, pages 412–416. Austin, TX, 1994.

[48] A. Guzman. Decomposition of a visual scene into three-dimensional bodies. In *Proceedings Fall Joint Computer Conference, volume 33*, pages 291–304, 1968.

[49] A. Guzman. Analysis of curved line drawings using context and global information. In *B. Meltzer and D. Michie, editors, Machine Intelligence 6*, pages 325–375. John Wiley and Sons, Inc., New York, NY, 1979.

[50] James Hafner, Harpreet S. Sawhney, Will Equitz, Myron Flickner, and Wayne Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(7):729–736, 1995.

[51] Peter W. Hallinan, Gaile G. Gordon, A. L. Yuille, Peter Giblin, and David Mumford. *Two- and three-dimensional patterns of the face.* A. K. Peters, Ltd., Natick, MA, USA, 1999.

[52] R. M. Haralick. Statistical and structural approaches to texture. 67(5):786–804, 1979.

[53] R. M. Haralick, Dinstein, and K. Shanmugam. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3:610–621, November 1973.

[54] Junfeng He, Shih-Fu Chang, and Lexing Xie. Fast kernel learning for spatial pyramid matching. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, USA, June 2008.

[55] Geoffrey E. Hinton, Geo Rey E. Hinton, Peter Dayan, and Michael Revow. Modelling the manifolds of images of handwritten digits. *IEEE Transactions on Neural Networks*, 8:65–74, 1997.

[56] M. Holm. Towards automatic rectification of satellite images using feature based matching. *Proceedings of the International Geoscience and Remote Sensing Symposium*, pages 2439–2442, 1991.

[57] J. Y. Hsiao and A. A. Sawchuk. Supervised textured image segmentation using feature smoothing and probabilistic relaxation techniques. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(12):1279–1292, 1989.

[58] John Y. Hsiao and Alexander Sawchuk. Unsupervised textured image segmentation using feature smoothing probabilistic relaxation techniques. *Comput. Vision Graph. Image Process.*, 48(1):1–21, 1989.

[59] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *Information Theory, IEEE Transactions on*, 8(2):179–187, 1962.

[60] Jing Huang, S Ravi Kumar, Mandar Mitra, and Wei-Jing Zhu. Spatial color indexing and applications. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, page 602, Washington, DC, USA, 1998. IEEE Computer Society.

[61] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 762, Washington, DC, USA, 1997. IEEE Computer Society.

[62] D. A. Huffman. Impossible objects as nonsense sentences. In *B. Meltzer and D. Michie, editors, Machine Intelligence 6*, pages 295–324. Edinburgh University Press, 1971.

[63] J. E. Hummel and I. Biederman. Dynamic binding in a neural network for shape recognition. In *Psychol Rev. 99*, pages 480–517, 1992.

[64] D. P. Huttenlocher, G. A. Klanderman, and W. A. Rucklidge. Comparing images using the hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(9):850–863, 1993.

[65] M. Ioka. A method of defining the similarity of images on the basis of color information. In *Technical Report RT-0030 IBM Research*. Tokyo Research Laboratory, 1989.

[66] Tommi S. Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 487–493, Cambridge, MA, USA, 1999. MIT Press.

[67] Anil K. Jain and Douglas Zongker. Representation and recognition of handwritten digits using deformable templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(12):1386–1391, 1997.

[68] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 494–501, New York, NY, USA, 2007. ACM.

[69] Michael Jones, Paul Viola, Paul Viola, Michael J. Jones, Daniel Snow, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. In *In ICCV*, pages 734–741, 2003.

[70] Yang Jun, Jiang Yu-Gang, Hauptmann Alexander G., and Ngo Chong-Wah. Evaluating bag-of-visual-words representations in scene classification. In *MIR '07*, pages 197–206, NY, USA, 2007. ACM.

[71] Takeo Kanade. Picture processing system by computer complex and recognition of human faces. In *doctoral dissertation, Kyoto University*. November 1973.

[72] Takeo Kanade. Computer recognition of human faces. *Interdisciplinary Systems Research*, 47, 1977.

[73] Gerald J. Kaufman and Kenneth J. Breeding. The automatic recognition of human faces from profile silhouettes. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-6(2):113 –121, feb. 1976.

[74] S. M. Kosslyn and P. Jolicoeur. A theory-based approach to the study of individual differences in mental imagery. In *R. E. Snow, P-A. Federico, and W. E. Montague (Eds.), Aptitude, learning, and instruction, vol. 2: Cognitive process analyses of learning and problem-solving.* Hillsdale, NJ: Erlbaum Associates, 1980.

[75] D. Koubaroulis, J. Matas, and J. Kittler. Evaluating colour-based object recognition algorithms using the soil-47 database. In *in Asian Conference on Computer Vision*, pages 840–845, 2002.

[76] Amlan Kundu and Jia-Lin Chen. Texture classification using qmf bank-based subband decomposition. *CVGIP: Graph. Models Image Process.*, 54(5):369–384, 1992.

[77] L. Lam and Ching Y. Suen. Structural classification and relaxation matching of totally unconstrained handwritten zip-code numbers. *Pattern Recogn.*, 21(1):19–32, 1988.

[78] Christoph H. Lampert, Matthew B. Blaschko, and Thomas Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR*, pages 1–8, 2008.

[79] K. I. Laws. Texture image segmentation. In *Ph.D. disseration*, volume Image Processing Inst. Univ. of Southern California, 1980.

[80] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 649, Washington, DC, USA, 2003. IEEE Computer Society.

[81] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

[82] Yann Lecun, Lon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[83] Bastian Leibe, Edgar Seemann, and Bernt Schiele. Pedestrian detection in crowded scenes. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 878–885, Washington, DC, USA, 2005. IEEE Computer Society.

[84] Marius Leordeanu, Martial Hebert, and Rahul Sukthankar. Beyond local appearance: Category recognition from pairwise interactions of simple features. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2007.

[85] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, 43(1):29–44, 2001.

[86] Fei-Fei Li and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 524–531, Washington, DC, USA, 2005. IEEE Computer Society.

[87] Wen-Gou Lin and Shuenn-Shyang Wang. A note on the calculation of moments. *Pattern Recogn. Lett.*, 15(11):1065–1070, 1994.

[88] Tony Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision*, 11:283–318, 1993.

[89] S. Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982.

[90] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*, page 1150, Washington, DC, USA, 1999. IEEE Computer Society.

[91] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*, page 1150, Washington, DC, USA, 1999. IEEE Computer Society.

[92] H. Lu, B. Ooi, and K. Tan. Efficient image retrieval by color contents. In *Pro. of the 1994 Int. Conf. on Applications of Databases*, 1994.

[93] A. K. Mackworth. Interpreting pictures of polyhedral scenes. In *Artificial Intelligence Journal, 4*, pages 99–118. Edinburgh University Press, 1973.

[94] Jitendra Malik and Pietro Perona. Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A*, 7:923–932, 1990.

[95] D. Marr and H. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. In *Proceedings of the Royal Society of London*, pages 269–294, 1978.

[96] Michael Mayo and Anna T. Watson. Automatic species identification of live moths. *Know.-Based Syst.*, 20:195–202, March 2007.

[97] Rajiv Mehrotra and James E. Gary. Similar-shape retrieval in shape data management. *Computer*, 28(9):57–62, 1995.

[98] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004.

[99] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005.

[100] Krystian Mikolajczyk, Bastian Leibe, and Bernt Schiele. Multiple object class detection with a generative model. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 26–36, Washington, DC, USA, 2006. IEEE Computer Society.

[101] Krystian Mikolajczyk and Cordelia Schmid. An affine invariant interest point detector. In *Proc. European Conf. Computer Vision*, pages 128–142. Springer Verlag, 2002.

[102] Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman. Human detection based on a probabilistic assembly of robust part detectors. pages 69–82, 2004.

[103] Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio. Example-based object detection in images by components. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(4):349–361, 2001.

[104] Greg Mori and Jitendra Malik. Recovering 3d human body configurations using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(7):1052–1062, 2006.

[105] Joseph L. Mundy. Object recognition in the geometric era: A retrospective. In *Toward CategoryLevel Object Recognition, volume 4170 of Lecture Notes in Computer Science*, pages 3–29. Springer, 2006.

[106] Joseph L. Mundy and Andrew Zisserman, editors. *Geometric invariance in computer vision*. MIT Press, Cambridge, MA, USA, 1992.

[107] Hiroshi Murase and Shree K. Nayar. Visual learning and recognition of 3-d objects from appearance. *Int. J. Comput. Vision*, 14(1):5–24, 1995.

[108] Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, 2002. Chair-Russell, Stuart.

[109] T. Menp and M. Pietikinen. Texture analysis with local binary patterns. In *Handbook of Pattern Recognition and Computer Vision*, 2005.

[110] V.S. Nalwa. A guided tour of computer vision. Addison-Wesley, Reading, MA, 1993.

[111] Sameer Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (coil-100). Technical report, 1996.

[112] W. Niblack, R. Barber, and et al. The qbic project: Querying images by content using color, texture and shape. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1994.

[113] Eric Nowak, Frdric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *In Proc. ECCV*, pages 490–503. Springer, 2006.

[114] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001.

[115] Andreas Opelt, Axel Pinz, Michael Fussenegger, and Peter Auer. Generic object recognition with boosting. *PAMI*, 28:2006, 2004.

[116] Andreas Opelt and Andrew Zisserman. A boundary-fragment-model for object detection. In *In ECCV*, pages 575–588, 2006.

[117] Constantine Papageorgiou and Tomaso Poggio. A trainable system for object detection. *Int. J. Comput. Vision*, 38(1):15–33, 2000.

[118] Florent Perronnin and Christopher Dance. *Fisher Kernels on Visual Vocabularies for Image Categorization*, pages 1–8. IEEE, 2007.

[119] Euripides G. M. Petrakis, Christos Faloutsos, and King-Ip (David) Lin. Imagemap: An image indexing method based on spatial similarity. *IEEE Trans. on Knowl. and Data Eng.*, 14(5):979–987, 2002.

[120] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The feret evaluation methodology for face recognition algorithms. In *Proc. Office of Nat'l Drug Control Policy, CTAC Int'l Technology Symp*, 1997.

[121] Nicolas Pinto, David D. Cox, and James J. DiCarlo. Why is real-world visual object recognition hard? *PLoS computational biology*, 4(1):e27+, January 2008.

[122] Jan Puzicha, Joachim M. Buhmann, Yossi Rubner, and Carlo Tomasi. Empirical evaluation of dissimilarity measures for color and texture. pages 1165–1173, 1999.

[123] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 883–890, Washington, DC, USA, 2005. IEEE Computer Society.

[124] Pedro Quelhas, Florent Monay, Jean marc Odobez, Daniel Gatica-perez, and Tinne Tuytelaars. A thousand words in a scene. In *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[125] T. H. Reiss. The revised fundamental theorem of moment invariants. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(8):830–834, 1991.

[126] Richard M. Rickman and T. John Stonham. Content-based image retrieval using color tuple histograms. In *Proc. SPIE Vol. 2670, p. 2-7, Storage and Retrieval for Still Image and Video Databases IV*. Ishwar K. Sethi; Ramesh C. Jain; Eds., 1996.

[127] L. G. Roberts. Machine perception of three-dimensional solids. In *Tippett, J. and Berkowitz, D. and Clapp, L. and Koester, C. and Vanderburgh, A., editor, Optical and Electrooptical Information processing*, pages 159–197. MIT Press, 1965.

[128] Paul L. Rosin. Edges: saliency measures and automatic thresholding. *Mach. Vision Appl.*, 9(4):139–159, 1997.

[129] Fredrick Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *In Proc. CVPR*, pages 272–277, 2003.

[130] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover"s distance as a metric for image retrieval. Technical report, Stanford, CA, USA, 1998.

[131] Yong Rui, Thomas S. Huang, and Shih-fu Chang. Image retrieval: Past, present, and future. In *Journal of Visual Communication and Image Representation*, volume 10, pages 1–23, 1997.

[132] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *In IEEE Computer Vision and Pattern Recognition*, pages 2033–2040, 2006.

[133] Haline E. Schendan and Marta Kutas. Neurophysiological evidence for the time course of activation of global shape, part, and local contour representations during visual object categorization and memory. *J. Cognitive Neuroscience*, 19(5):734–749, 2007.

[134] Cordelia Schmid, Philippe Bobet, Bart Lamiroy, and Roger Mohr. An image oriented cad approach. In Jean Ponce, Andrew Zisserman, and M. Hébert, editors, *Object Representation in Computer Vision II*, number 1144 in Lecture Notes in Computer Science, pages 221–245. Sprin-ger-Ver-lag, April 1996.

[135] Cordelia Schmid and Roger Mohr. Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:530–535, 1997.

[136] Henry Will Schneiderman. *A statistical approach to 3d object detection applied to faces and cars.* PhD thesis, Pittsburgh, PA, USA, 2000. Chair-Kanade, Takeo.

[137] Lior Shamir. Automatic morphological classification of galaxy images. *Monthly Notices of the Royal Astronomical Society*, page 399: 13671372, 2009.

[138] R. M. Shepard. Toward a universal law of generalization for psychological science. In *Science*, pages 1317–1323, 1987.

[139] M. Shridhar and A. Badreldin. Recognition of isolated and simply connected hand-written numerals. *Pattern Recogn.*, 19(1):1–12, 1986.

[140] L. Sirovich and M. Kirby. *Low-dimensional procedure for the characterization of human faces*, volume 4. OSA, March 1987.

[141] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2005.

[142] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1470, Washington, DC, USA, 2003. IEEE Computer Society.

[143] J. R. Smith and Shih-Fu Chang. Single color extraction and image query. In *ICIP '95: Proceedings of the 1995 International Conference on Image Processing (Vol. 3)-Volume 3*, page 3528, Washington, DC, USA, 1995. IEEE Computer Society.

[144] J. R. Smith and Shih-Fu Chang. Automated binary texture feature sets for image retrieval. In *ICASSP '96: Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference*, pages 2239–2242, Washington, DC, USA, 1996. IEEE Computer Society.

[145] Markus Stricker and Alexander Dimai. Color indexing with weak spatial constraints. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1996.

[146] Markus Stricker and Markus Orengo. Similarity of color images. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1995.

[147] Michael J. Swain and Dana H. Ballard. Color indexing. *Int. J. Comput. Vision*, 7(1):11–32, 1991.

[148] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Texture features correspond to visual perception. In *Systems, Man and Cybernetics, IEEE Transactions on, 8(6)*, 1978.

[149] Michael J. Tarr, Heinrich H. Blthoff, Marion Zabinski, and Volker Blanz. To what extent do unique parts influence recognition across changes in viewpoint? *Psychological Science*, 8:282–289, 1997.

[150] Michael J. Tarr and Heinrich H. Blthoff. Image-based object recognition in man, monkey and machine. *Cognition*, 67(1-2):1 – 20, 1998.

[151] M.J. Tarr and H.H. Blthoff. Is human object recognition better described by geon-structural-descriptions or by multiple-views? In *Journal of Experimental Psychology: Human Perception and Performance, 21*, pages 1494–1505, 1995.

[152] M. R. Teague. Image analysis via the general theory of moments. *Journal of the Optical Society of America (1917-1983)*, 70:920–930, August 1980.

[153] Johan Thureson and Stefan Carlsson. Appearance based qualitative image description for object class recognition. In *In Proc. ECCV*, pages 518–529, 2004.

[154] D. Trier, Anil K Jain, and Torfinn Taxt. Feature extraction methods for character recognition - a survey, 1995.

[155] Tinne Tuytelaars and Krystian Mikolajczyk. Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3(3):177–280, 2008.

[156] Ilkay Ulusoy and Christopher M. Bishop. Generative versus discriminative methods for object recognition. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 258–265, Washington, DC, USA, 2005. IEEE Computer Society.

[157] M Unser. Local linear transforms for texture measurements. *Signal Process.*, 11(1):61–79, 1986.

[158] M Unser. Sum and difference histograms for texture classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(1):118–125, 1986.

[159] M. Unser. Texture classification and segmentation using wavelet frames. *IEEE Transactions on Image Processing*, 4(11), 1995.

[160] Martin Vetterli and Jelena Kovačevic. *Wavelets and subband coding*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1995.

[161] Paul Viola, Paul Viola, Michael Jones, and Michael Jones. Rapid object detection using a boosted cascade of simple features. pages 511–518, 2004.

[162] D. Waltz. Understanding line drawings of scenes with shadows. In *Patrick H. Winston, editor, The Psychology of Computer Vision*, pages 19–91. McGraw-Hill, 1975.

[163] Gang Wang and Ye Zhang Li Fei-fei. Using dependent regions for object categorization in a generative framework. In *In CVPR*, pages 1597–1604. IEEE Computer Society, 2006.

[164] X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *ICCV*, pages 1–8, 2007.

[165] Anna T. Watson, Mark A. O'Neill, and Ian J. Kitching. Automated identification of live moths (macrolepidoptera) using digital automated identification system (daisy). *Systematics and Biodiversity*, 1(03):287–300, 2003.

[166] Isaac Weiss. Geometric invariants and object recognition. *Int. J. Comput. Vision*, 10(3):207–231, 1993.

[167] Ian H. Witten and E. Franks. *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[168] John Wright, Student Member, Allen Y. Yang, Arvind Ganesh, Student Member, S. Shankar Sastry, Yi Ma, and Senior Member. Robust face recognition via sparse representation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI*, 2008.

[169] Chyuan Jy Wu and Jun S. Huang. Human face profile recognition by computer. *Pattern Recogn.*, 23(3-4):255–259, 1990.

[170] Jianxin Wu and James M. Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *ICCV*, 2009.

[171] Ming-Hsuan Yang, David J. Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(1):34–58, 2002.

[172] Edmond Zhang and Michael Mayo. Pattern discovery for object classification. In *in Pro. Int. Vision and Computing NZ, IVCNZ08*, 2008.

[173] Edmond Zhang and Michael Mayo. Sifting the relevant from the irrelevant: Automatically detecting objects in training images. In *in Pro. DICTA2009 (Digital Image Computing: Techniques and Applications), 1-3 December 2009, Melborne, Australia*, 2009.

[174] Edmond Zhang and Michael Mayo. Enhanced spatial pyramid matching using log-polar-based image subdivision and representation. In *in Pro. the International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, Australia*, 2010.

[175] Edmond Zhang and Michael Mayo. Improving bag-of-words model with spatial information. In *in Pro. Int. Vision and Computing NZ, IVCNZ10*, 2010.

[176] G. Zhang and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. *In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (2), 2006.

[177] Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *In CVPR*, pages 2126–2136, 2006.

[178] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. Comput. Vision*, 73(2):213–238, 2007.

[179] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29:915–928, June 2007.

[180] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. volume 35, pages 399–458. ACM, December 2003.

[181] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1491–1498, Washington, DC, USA, 2006. IEEE Computer Society.

Appendix A

Object Instance Recognition

This thesis focuses on the problem of learning and recognizing object categories. While single object recognition is not directly relevant to this work, useful insights can be drawn from the years of research and progress [129, 90, 42, 34] in this area. By not having to consider intra-class variability, object instance recognition is much easier compared to category level recognition. Much progress has been made on efficient recognition [90], illumination-invariant [101, 90], and view-point invariant [129, 64] representation and recognition. In this section, the three notable approaches in single instance recognition are reviewed in turn: geometric matching, global appearance matching, and texture region matching.

A.1 Geometric Matching

Ever since the beginning of object recognition by computers, the main approach has been dominated by the drive to discover a model representation of objects [105]. This model is then used to predict the appearance of an object, under any view-

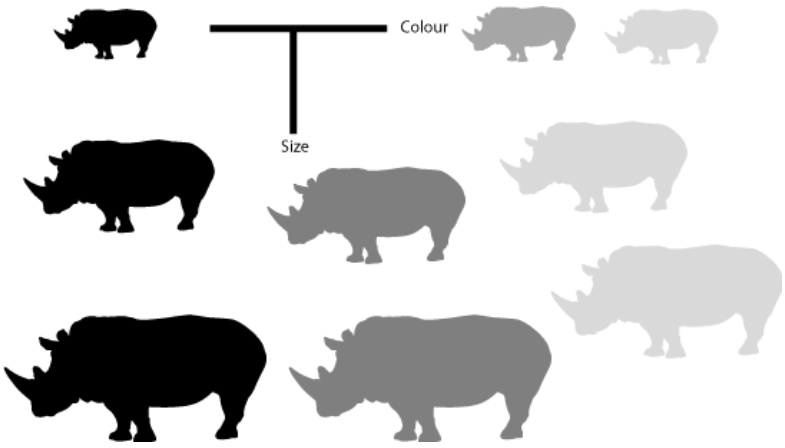

Figure A.1: This set of rhinoceros is arranged in a similarity space of two dimensions, size and colour.

point, scale, background clutter and occlusion. The geometric approach stresses the representation of similarity of relationships among the members of a set of objects as is demonstrated in Figure A.1. Much attention was given to extracting geometric primitives (e.g. lines, curves, etc.) that are invariant to viewpoint change [106].

A.2 Blocks World

One of the initial theories of geometric representation was the blocks world concept [127] where the emphasis is on establishing a theoretical framework for cognitive tasks. The motivation is that computers could carry out the necessary reasoning using formal logic and other mathematical tools. Figure A.2 illustrates an example of this approach.

The plan was to start with a simplification of the image, so that the mathematical models can be applied rigorously to solve most of the resulting recognition,

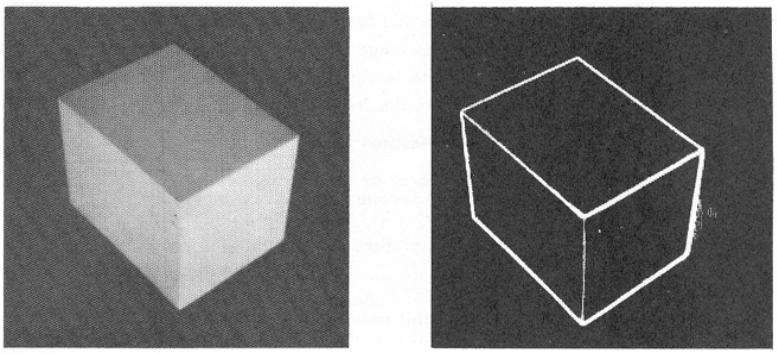

Figure A.2: The image on the left is the original image and the image on the right is the differentiated image.

before proceeding to more difficult tasks [105]. For the problem of object recognition, this theory works by restricting objects to polyhedral shapes on a uniform background. The goal of this theory is to be able to recognize generalized shapes in an arbitrary spatial arrangement including significant occlusion of one object by itself or others [105].

Despite the popularity of this approach, the blocks world model fails to address many of the difficulties in object recognition, such as curved surfaces and boundaries, moving objects, multiple light sources, and remote shadowing [48, 26, 93, 62, 162]. This model was later extended in trying to handle these conditions, where the most notable work was done by Guzman in [49]. Guzman tackled these problems with a different approach, by restricting the problem to only line drawings. Many of the difficult scene rendering issues were avoided. However, the restriction to line drawings is far from a natural image problem, which was the main focus for computer vision community.

A new wave of psychologically sound viewpoint invariant theory for object recognition started to gain momentum in the 1980s. This theory is based on the

commonly held belief that the goal of vision is to reconstruct the 3D scene [63, 12, 108, 150, 33, 133]. According to the view-independent model, objects are represented in the forms of structural description of their component parts, by the visual system, and the relationships between those parts are independent of extrinsic factors.

A.3 Generalized 3D Cones

In Marr and Nishihara's [95] work, it was argued that the basic descriptor for all object parts is 3D generalised cones. Generalised cones can form a range of shapes, and component parts could themselves be decomposed into smaller parts. Hence, the model is hierarchical. Recognition is achieved by matching a model description derived from the image with stored 3D descriptions, and is described in the following three levels:

- **Single-model axis.** The identification of the main axis of the object;
- **Component axes.** The identification of other smaller sub-components of the objects; and
- **3D model matching.** The arrangements of components are matched against a stored 3D model description in order to identify the object.

Marr and Nishihara assumed that viewer-centred descriptions are remapped into 3D object-centred representations. They suggested that object representations should be relatively stable, that is, they should generalize or be invariant over changes in the retinal image. Moreover, they argued that new distinct rep-

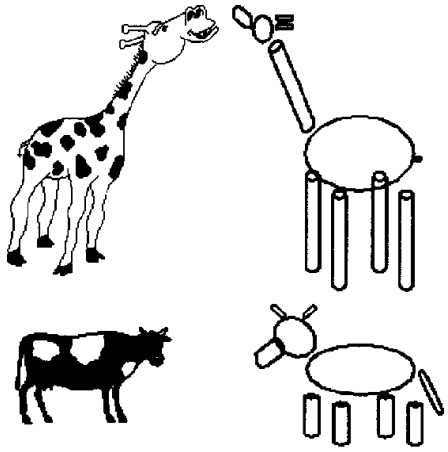

Figure A.3: Two objects and their structural descriptions (image taken from [33]).

resentations would be required for each small variation in the image of a given object. Figure A.3 is an example of Marr and Nishihara's 3D cones theory.

Marr and Nishihara's model, argues Tarr [150], meant that their object representation should be object-centred rather than viewer-centred, as objects are represented as configuration of 3D components.

In evaluating the theory of machine perception, Marr and Nishihara also proposed five criteria against which models of object perception can be evaluated:

- **Accessibility** How easily object descriptions can be derived from images.
- **Scope** The range of objects which the model applies.
- **Uniqueness** The same object should always result in the same description.
- **Stability** How well the description of the object remains stable under small variations in viewpoint, illumination and occlusion.
- **Sensitivity** The description should allow discrimination between objects.

The fundamental strength of Marr and Nishihara's model lies in the principles underlying the theory – it was the first meaningful psychologically-inspired attempt at object recognition. At its core, the properties of the theory are elegant and intuitive. It also had a tremendous impact on the study of machine vision. In particular, it helped shift the focus of high-level vision research from visual imagery [74] to visual object recognition. It also influenced and provided blueprints for many important theories to emerge in that era, including the Recognition by Component [11] theory.

Critics of this model have maintained that there are actually very limited experimental data to support this psychological model [149], despite the theoretical elegance of this approach. Tarr pointed out that one of the problems with this approach is that it has never been obvious that recovering descriptions of 3D parts from 2D images is generally possible. This argument was supported by the numerous machine vision experiments based on this theory which resulted in only limited success [110].

There are also a number of computational shortcomings for this model, including difficulties with recovery of 3D cones, instability of descriptions and the need for metric information.

- Difficulties with the recovery of 3D cones. This is the Achilles heel of any structural approach in object recognition. The construction of generalised 3D cones from 2D images are always going to be a challenge, mainly due to the detection of meaningful lines and junctions. It is interesting to note that this model worked well for hand-labelled line drawings [63]; however, results did not transfer to real world images [110].

- Instability of descriptions. Generalised 3D cones formed from 2D lines are not stable across different images. Edelman [33] argued that this is because of the inherent instability of structural interpretation that affects all structural approaches. He gave the example that the letter A can be decomposed into either three or five lines, depending on whether the sides of the letter A are represented by one or two straight lines.

- The need for metric information. In principle, Marr and Nishihara's model is capable of representing shapes by object decomposition; however, this ability has severally limited the uniqueness and representational and fine distinctions between similar shapes. For example, a 3D cylinder shape can represent an enormous number of object parts.

A.4 Recognition by Components – Geons

Biederman's recognition by component (RBC) [11] model for object recognition is also based on the structural description principle. However, one fundamental difference between the two models is that the RBC model limits the part descriptors to a set of 30-odd geometric shapes or Geons [9]. Geons come from a 2D image representation rather than the 3D representation of Marr and Nishihara. According to the authors, geons are detected on the basis of certain non-accidental properties of the contour in the image such as linearity, parallelism, curvilinearity, and symmetry.

The RBC theory holds that the resemble shape of objects is represented by constellations of recovered 3D parts (geons). The innovation of RBC theory is its use of combinations of non-accidental properties (e.g. parallel lines or collinear

line segments) as the basis for the recovery of parts. Because combinations of non-accidental properties are presumed to be viewpoint invariant, recovered geons exhibit restricted viewpoint invariance – as long as the same combinations are visible, leading to the same part description. Such viewpoint-invariance is the fundamental prediction of RBC theory tested by the authors [63].

RBC assumes that a representation of the object is either segmented, or parsed, into separate regions at points of deep concavity, particularly at corners where there are discontinuities in curvature. Biederman argues that such subdivision of parts assimilates well with human intuitions about the boundaries of object parts, where familiarity with objects are not required. Each segmented region is then approximated by one of a possible set of simple geons (Biederman suggests 36 geons).

The geons (primitive components) are theorised to be simple, such as blocks, cylinders, spheres, and wedges, typically symmetrical volumes without significant concavities. The fundamental perceptual assumption of RBC is that the geons can be differentiated on the basis of perceptual properties in the 2D image that are readily detectable and relatively independent of viewing position and degradation [11].

Moreover, Biedarman argues that RBC correlates well with human perceptual organization and computer-based pattern recognition: objects can be complex and irregular, but the components in which objects are constructed from are simple and regular. This means that the object's components and the structural description determine the characteristic of the object.

He suggests that a complete object, such as a chair, can be highly complex and asymmetrical, but the components will be simple volumes. A consequence of

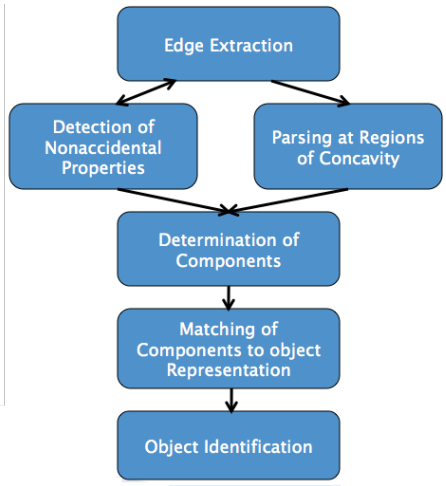

Figure A.4: Recognition by components starts with edge extraction in order to parse regions of concavity. This is done by the detection of non-accidental properties. Once all components are determined, Biederman argues that objects can then be readily identified.

this implementation is that it is the components that will be stable under noise or perturbation. If the components can be recovered and object perception is based on the component, then the object will be recognizable. RBC holds that the loci of parsing is a the corners; the geons are organized from the contours between corners.

The authors also suggested that the RBC model is an account for entry-level recognition performance, this is, the particular level of categorical abstraction assigned to objects at the time of initial identification [151]. Other strengths of the RBC theory suggested by Edelman [33] are: conceptual parsimony, invariance to change in viewing conditions, and good support for organization.

- **Conceptual parsimony.** A handful of primitives can allow a very large number of object classes to be represented.

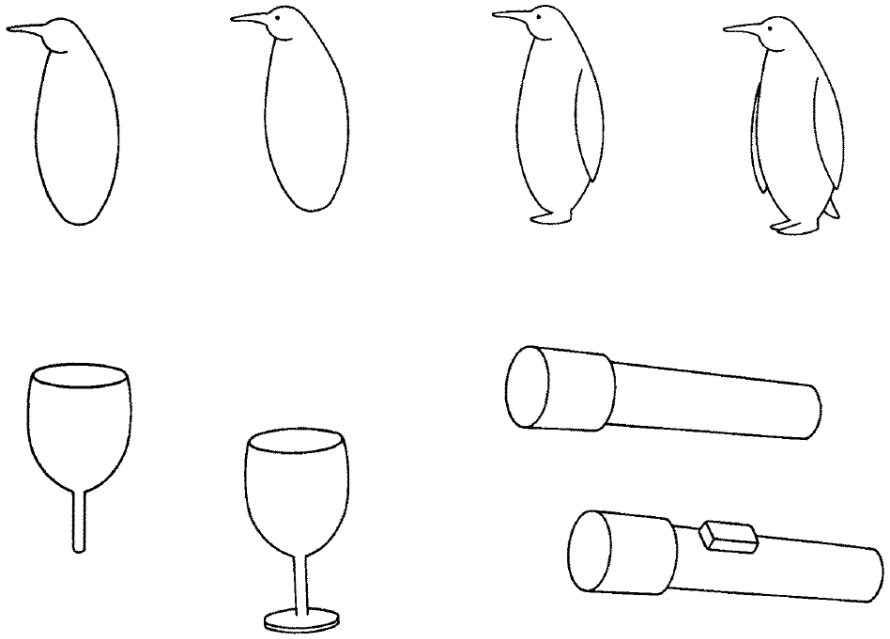

Figure A.5: Objects are decomposed into goens. Image taken from [11]

- **Invariance to changes in viewing conditions.** As long as the parts and their relationships can be identified.
- **Good support for organization.** Stemming from the possibility of repressing novel objects in terms of the same primitives as the familiar ones.

Opponents of this model have always argued that while the RBC theory is intuitive and elegant, there are, however, many flaws. Tarr [150, 149, 151], the chief critic of this theory, suggested three inherent problems with it. The first problem is that RBC lacks the generality to characterize a wide range of recognition conditions because the technique is conditioned for immediate viewpoint invariance. More specifically, general account of objects cannot be easily discovered from limited views from the object.

The second problem is that an extensive body of viewpoint-dependent results cannot be dismissed as processing 'by-product' or 'experimental artifacts.' Biederman suggests that all studies demonstrating viewpoint dependence fail to satisfy one or more of their conditions for immediate viewpoint invariance. Tarr claims that they have reviewed some of the many recognition experiments that provide converging evidence for multiple-views, instead of viewpoint-dependent.

The third problem is that geon structural descriptions cannot coherently account for category recognition, the domain they are intended to explain. For example, there are instances where geons will represent different entry-level objects as members of the same category (e.g., a cow and a horse). Additionally, Tarr suggests that Biederman's own results indicate that objects that are named as members of the same entry-level category are treated as separate representations by the recognition system.

Tarr argued that while this theory has been very influential, it is still unclear that any experimental approaches based on this model are robust enough for real world images. He added that the experiments carried out by Biederman and Gerhardstein [10] provided little meaningful evidence support.

A.5 Summary of Geometric Matching Methods

The geometric matching model was the first meaningful attempt by the computer vision community in tackling the problem of object recognition. The fundamental principle is that an object is represented as a collection of parts, which implies that recognition will be viewpoint invariant. Mundy, in [105], proposed four reasons why geometric representation played such an important part in the development

of recognition theory and resulting algorithms and systems, are invariance to viewpoint; invariance to illumination; well developed theory; and man-made objects.

- **Invariance to viewpoint.** Geometric object descriptions allow the projected shape of an object to be accurately predicted understanding perspective projection.

- **Invariance to illumination.** Recognizing geometric descriptions from images can be achieved using edge detection and geometric boundary segmentation.

- **Well developed theory.** Geometry has been under active investigation by mathematicians for a long time. The geometric framework has achieved a high degree of maturity and effective algorithms exist for analyzing and manipulating geometric structures.

- **Man-made objects.** A large fraction of manufactured objects are naturally described by primitive geometric elements, such as planes and spheres.

One of the major problems of the geometric approach is that an object can be seen from different points of view, resulting in different images which need to be recognized as portraying the same object [166]. This implies that the extraction of edges from natural images can be difficult when there is extensive illumination difference, background clutter and occlusion.

Appendix B

Category Level Recognition

This thesis is interested in the problem of learning and recognition of object categories. Unlike object instance recognition, the focus of category level recognition is not only matching concrete shapes to make sense of shape concepts. Indeed, traditional strategies like template matching, geometric models and texture region matching are no longer capable of handling such tasks. Not because of inflexibility of the models, but rather because the template database is no longer well-defined at the level of abstraction on which the system operates – because each instance of the category is no longer identical, hence the matching scheme must have a way of accounting for the variability across instances in the features extracted. Object categories are more general, require more complex representations, and are more difficult to learn; which is why most work today is focused on modeling and learning object categories.

In the last two decades, the research community has mainly focused on some challenging problems such as complex scenes, and large number of classes. This section reviews some of the most notable approaches in turn, starting from hand-

Figure B.1: Sample digits from the MNIST dataset.

written digits [82], pedestrian [69] and faces [161, 136].

B.1 Categorizing Handwritten Digits

The recognition of handwritten digits is a challenging problem, not only because there are different ways in which a digit can be written, but also as a result of strict requirements of specific problems, as shown in Figure B.1. The primary performance is measured by recognition accuracy and speed, and most researchers have adopted the classical pattern recognition approach in which image pre-processing is followed by feature extraction and classification. This section will not attempt to review in depth the work that has gone into this area in the past three decades. However, it will summarize research directions and methodologies in this field.

Work in this area can be roughly summarized in two dimensions: statistic/structural and local/global approaches. For the global statistical approach, Cash *et al.*, in [17] extract central and raw mathematical moments and use them as features, while Shridhar *et al.* [139] use features derived from the topological (e.g. crossings, endpoints, holes, etc) character profiles in the image, which are

dependent on the global property of the data.

For the local structural approach, Lam *et al.* [77] extracted local geometric information consisting of lines and convex polygons and used these as input to a structural classifier. Other notable attempts include automatically learning appropriate local features using feed forward neural networks [25]. Hinton *et al.* [55] argue that it is also possible to discriminate by fitting a separate probability density model to each class and then picking the class of the model that assigns the highest density to a test image. However, one disadvantage of this relative density approach is that it generally requires more computational time during recognition.

In recent years, a more intuitive approach [154] begin to emerge, in the form of deformable templates. In this approach, an image deformation is used to match testing images against a library of training images. Research in this approach has concentrated on taking the outlines of images, representing them with a number of a combinations of curve segments, and deformation of the image is achieved by altering the curve parameters [67].

Belonging to the same dimension of research, Lam and Suen [77] proposed a two-stage scheme. In their work, samples are first classified by their structure using a tree classifier. Samples which can not be confidently assigned to a class through this process will be passed to a slower, relaxation matching algorithm that uses deformable templates.

B.2 Pedestrian Recognition

The ability to detect people in images is key for several important applications, ranging from intelligent vehicle braking systems, to advanced user interface, to

Figure B.2: Some sample images from the INRIA dataset. The subjects are upright with a wide range of pose, clothing and background

robotics and surveillance, just to name a few. This review is concerned with automatic pedestrian detection in natural images. Advances in computer vision in the recent decades have shifted the research focus in this field from manually crafted models to the more intuitive learning approaches. Nevertheless, this research area has proven to be much more difficult than initially thought, owing to the variability in appearances of the subjects. Broadly speaking, pedestrian recognition is particularly challenging for a number of reasons:

- Pose – This is one of the primary reasons why the traditional template matching methods will not work in this task. There exist countless numbers of different poses and styles that are unique to each individuals (Figure B.2 shows some examples of a typical human detection dataset).

- Background clutter – Pedestrian detection systems are most likely to be used in cities. This implies that the algorithm must consider different background conditions.

- Clothing – Pedestrians possess different fashion styles and the temperature also plays an important role in what is worn.

- Change of context – The system must be able to distinguish the difference between an actual human and a human on a poster.

In general, a typical pedestrian detection system includes the following stages:

- Determining the candidate regions of interest with a systematic scan of the input, for the purpose of possible target filtering;
- Single frame classification for the initial recognition;
- Multiple frame classification. A two-stage process in detecting pedestrians from multiple frames; and
- Determining range measurement. The last stage is concerned with identifying the entire body of the subject for the purpose of gait recognition.

Research in this area can be summarized into two approaches: sliding window based and part-detection based. In the sliding window approach, a classifier is applied sequentially to all possible sub-windows in an image. For example, in [117] a Support Vector Machine (SVM) classifier is learned from Haar wavelets (other features were also used, such as covariance) as feature descriptors. Later, this work was further extended by Mohan *et al.* [103] with the use of multiple classifiers for improved accuracy, where category decisions from each sub-window are combined to give the final decision. Recently, a successful human detector was proposed by Dalal *et al.* [27], which trains a SVM classifier using densely sampled histograms of oriented gradients (HOG) from image patches. This work was later also extended to handle near real time detection, using the same HOG features [181].

The second group of approaches are focused more on detecting human parts or common shapes before constructing the final model based on the geometrical

layout of the detected parts or shapes [38, 100]. Others [83] have proposed using the co-occurrence of object parts, where the existence of pedestrians is determined by the maximum likelihood of the detected co-occurring parts. Furthermore, [102] proposed to divide the human body into seven parts and the representation of these body parts is made using SIFT [90] keypoints.

B.3 Face Recognition

Within the last two decades, there have been a significant number of algorithms proposed for the problem of face recognition. Progress has advanced to the point that face recognition systems are being demonstrated in real-world applications [120]. Interestingly, the key challenge of this problem is not the differences between people's images, instead, variations in pose and illumination actually provide a bigger challenge [180].

The ultimate goal of face recognition is to extract characteristics (features like shapes, colour and textures) of a face from images, despite other random variations included in the images. Blanz *et al.* [13] argued that these variations are different to background clutter and image noise, rather, these conditions are the results of camera and scene geometry, illumination direction and intensity. There are two main approaches in dealing with these image variations; one approach is to treat these conditions as another type of image feature and model their functional role explicitly. The second approach does not formally distinguish these conditions from other image features and treats them equally. Figure B.3 illustrates some examples of random variations in the Harvard Face Dataset.

Most of the work in computer face recognition is focused on the detection of

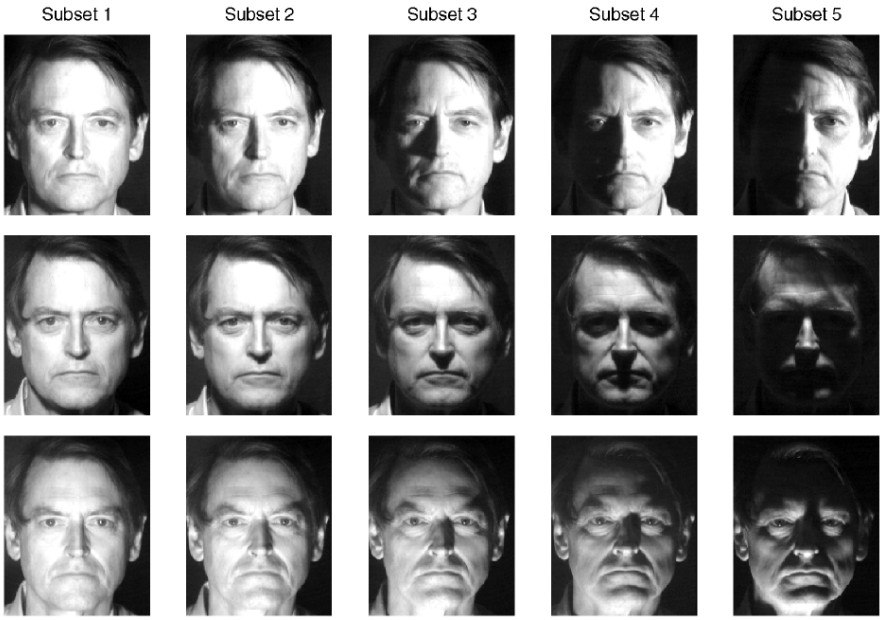

Figure B.3: Some examples of variations in the Harvard Face Dataset [171].

individual facial features in defining a face model by size, position and the geometrical relationship between these features. Similar to some of the early work in object instance recognition, research in this field can be divided into two dimensions: geometrical based and appearance based models.

In the geometrical based approaches [169, 72, 73], researchers have been relying on using properties such as eyes, nose, chin and the distance and angle between these features to perform recognition. This type of approach benefits from the traditional advantages of the geometric based object instance recognition model, such as economical representation of image features and relative insensitivity to variations in illumination and viewpoint changes. However, this type of approach also suffers from similar problems such as sensitivities to the feature extraction and measurement process. Fischler *et al.* [40] were the first to propose this type of

'parts and structure' model, where the model consists of various of 'parts' and is arranged in some geometrical 'structure'. Despite the fact that in their research, attempted recognition performance was poor, this idea nevertheless sparked numerous subsequent extensions based on this theory.

The other group of approaches [51, 140], the appearance based models, use low-dimensional representations of images of objects to perform recognition. This type of approach differs from geometrical based techniques in that their low-dimensional representation is a more suitable representation of the original image. However, research in this area suffers from criticism that in their original form, the technique cannot extrapolate or generalize to novel viewing conditions – a problem that is also hindering object instance recognition.

B.4 Summary of Category Level Recognition

Reflecting the historic development of the field, the literature on category level recognition has been reviewed in this section. Early attempts at solving this problem have concentrated on a small set of objects, for example, handwritten digits, pedestrian detection, and face recognition. In comparison to object instance recognition, this is a much harder problem. Each instance of an object class is no longer identical, hence recognition algorithms not only have to account for variations in viewpoint and illumination, but also in variability across instance shapes, colours, contour and region textures.

The following section will review some of the recent works in this area. These are appearance based models, shape based models, the BOW model, and the spatial pyramid extension of the BOW model.