PROBABILISTIC CONTEXTUAL MODELS FOR OBJECT CLASS RECOGNITION IN UNCONTRIVED IMAGES

MOZAHERUL HOQUE ABUL HASANAT

UNIVERSITI SAINS MALAYSIA

2011

PROBABILISTIC CONTEXTUAL MODELS FOR OBJECT CLASS RECOGNITION IN UNCONTRIVED IMAGES

by

MOZAHERUL HOQUE ABUL HASANAT

Thesis submitted in fulfillment of the requirements
for the degree of
Doctor of Philosophy

May 2011

ACKNOWLEDGEMENTS

All praises and thanks to Allah, the most wise and knowledgeable for giving me the energy and talent to undertake this research. I pray to Him for His guidance throughout my life.

Gratitudes beyond words to my father, mother and brother, who has always ushered me with the best of their love, support and assistance in every step of my life. I deeply acknowledge their enormous sacrifice behind my success.

I would like to express my sincere appreciation to my thesis supervisor, Dr. Dhanesh Ramachandram for his guidance, encouragement, friendship, and support over this course of my study. I am also grateful to my co-supervisor Dr. Mandava Rajeswari for all the assistance, invaluable advice and motherly kindness throughout my research.

I would also like to thank the academic and technical support staff of the School of Computer Sciences for providing a conducive environment during the course of my research. My sincere gratitude to Alfian and Tan Ewe Hoe for their generous help in translating the Abstract of this thesis to Bahasa Malaysia. I would also wish to acknowledge the financial support under the prestigious USM Fellowship Scheme by the Institute of Postgraduate Studies, Universiti Sains Malaysia, for the full length of this research.

All my friends and members of the Computer Vision Research Group: Mahmoud Jawarneh, Nor Idayu, Anusha Achutan, Khairul Azrin, Nor Hafizah, Siti Zubaidah, Tan Alwin, Ahmad Adel, Osama Alia, Alfian Abdulhalin, Ehsan Abbasbejad – you have been a great source of support and lots of fun to work with. I should not forget to thank Dr. Michael Hartley of the University of Nottingham for exposing me to the world of artificial intelligence research and encouraging me to join a PhD programme.

Last but not the least, my endless thanks to my beloved wife Mazkiah, lovely sons Muaz and Marwan for their sacrifice, constant love, encouragement and patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BBN Bayesian Belief Network

BIC Bayesian Information Criteria

ConVeS Contextual Verification System

CRF Conditional Random Field

FIAMB Fast Incremental Association Markov Blanket algorithm

HC Hill Climbing algorithm

LAR Local Appearance-Based Recognition

MLE Maximum Likelihood Estimator

MMHC Max-Min Hill Climbing algorithm

MRF Markov Random Field

pdf Probability Distribution Function

ROC Receiver Operating Characteristic

SCM Semantic Context Model

SpCM Spatial Context Model

SVM Support Vector Machine

LIST OF SYMBOLS

θ	Angular projection in radians
$+\!\!+$	Concatenation operator for two vectors
μ_p	Centre of mass of object p
I	Uncontrived image
Γ	Set of regions in image I
Λ	Set of object classes to be recognised
$\mathcal{N}$	Set of neighbouring regions of any γ_q
Ψ	Set of spatial relations that a ${\lambda_n}^{\nu_m}$ can have with respect to ${\lambda_n}^{\gamma_q}$ where ${\lambda_n}^{\gamma_q}$ is said to be the *reference object* and ${\lambda_n}^{\nu_m}$ is the *neighbouring object*
$\mathcal{H}_{LAR}^{\gamma_q}$	Initial hypotheses given by LAR for region γ_q
$\mathcal{H}_{SCM}^{\gamma_q}$	Initial hypotheses given by SCM for region γ_q
$\mathcal{H}_{SpCM}^{\gamma_q}$	Initial hypotheses given by SpCM for region γ_q

MODEL KONTEKSTUAL BERASASKAN KEBARANGKALIAN UNTUK PENGECAMAN KELAS OBJEK DALAM IMEJ-IMEJ YANG TIDAK DIBUAT-BUAT

ABSTRAK

Konteks merupakan suatu elemen penting dalam mendapatkan penjelasan yang bererti untuk sesuatu imej bagi kedua-dua sistem visual biologi dan buatan. Tesis ini mencadangkan permodelan hubungan konteks di antara objek dunia nyata di dalam imej yang tidak dibuat-buat bagi meningkatkan prestasi pengecaman kelas objek. Dua model kebarangkalian dicadangkan iaitu Semantic Context Model (SCM) dan Spatial Context Model (SpCM) - untuk memodelkan hubungan kontekstual semantik dan ruangan peringkat tinggi. SCM mempelajari struktur graf terarah dari suatu set data imej yang tidak dibuat-buat untuk memodelkan hubungan kebersandaran di antara objek dunia nyata. Nod graf mewakili objek bagi kawasan liputan permasalahan manakala sisi terarah mewakili hubungan kebersandaran di antara dua objek. Merujuk kepada SpCM pula, ia mempelajari taburan kebarangkalian bagi hubungan ruangan berpasangan bagi kesemua objek di dalam ruang lingkup permasalahan. Kedua-dua SCM dan SpCM mampu mempelajari hubungan kontekstual secara berasingan dan juga bebas dari mana-mana proses pembelajaran lain di dalam sesuatu sistem pengecaman objek. Ini membolehkan kedua-dua model ini diintegrasikan secara bermodul dengan sistem pengecaman kelas objek yang sedia ada. Dalam hal ini, tesis ini juga mencadangkan suatu rangka kerja iaitu ConVeS yang mampu mengintegrasikan model-model yang dicadangkan dengan sistem pengecaman kelas objek. Peningkatan prestasi yang dicapai melalui penggunaan model-model yang dicadangkan telah dibandingkan dengan dua sistem pengecaman berasaskan penampilan tempatan. Metrik-metrik

prestasi yang telah digunakan untuk minilai prestasi pengecaman adalah: matriks kekeliruan, ketepatan, kepersisan, ingatan, ukuran-F, lengkok ROC dan kawasan di bawah lengkok ROC. Keputusan eksperimen membuktikan bahawa konteks semantik dan ruangan peringkat tinggi menyumbang secara positif kepada prestasi sistem-sistem pengecaman berasaskan penampilan tempatan. Keputusan ini juga menunjukkan bahawa maklumat kontekstual amat berguna apabila pengecaman berasaskan penampilan tempatan adalah lemah.

PROBABILISTIC CONTEXTUAL MODELS FOR OBJECT CLASS RECOGNITION IN UNCONTRIVED IMAGES

ABSTRACT

Context is a vital element in deriving meaningful explanation of an image for both biological, as well as, artificial vision systems. This thesis proposes to model contextual relation among real-world objects in uncontrived images in order to improve object class recognition performance. Two probabilistic models are proposed – Semantic Context Model (SCM), and Spatial Context Model (SpCM) to model high-level semantic and spatial contextual relations respectively. SCM learns a directed graph structure from a given dataset of uncontrived images to model the dependency relation among real-world objects. The nodes of the graph represent the objects of the problem domain and the directed edges represent the dependency relation between a pair of objects. With respect to SpCM, it learns probability distributions of pair-wise spatial relations for all the objects in the problem domain. Both SCM and SpCM can learn contextual relation independently of each other and of any other learning process within an object class recognition system. This allows for modular integration of these models with an existing object class recognition system. In this regard, this thesis also proposes a framework dubbed as ConVes that integrates the proposed models with object class recognition systems. The performance improvements achieved due to the usage of the proposed models were compared against two local appearance-based recognition systems. Performance metrics used to evaluate the recognition performance are: confusion matrix, accuracy, precision, recall, F-measure, ROC curve, and area under the ROC curve. Experimental results proved that high-level se-

mantic and spatial context positively contribute to the performance of local appearance-based recognition systems. The results also shows that contextual information is more useful when the local appearance-based recognition algorithms do not perform well.

CHAPTER 1

INTRODUCTION

One of the fundamental concerns of computer vision is object class recognition. It refers to discriminating a class of objects from every other object or pattern in the world not belonging to the target object class (Zhang et al., 2005). Object class recognition in uncontrived images is of particular importance in computer vision as it is relevant to applications such as autonomous navigation, robot vision, and satellite image analysis. Human beings accomplish the enormously complex task of visual recognition almost effortlessly in part because of their ability to integrate contextual cues from the surroundings and interpret the information based on their accumulated knowledge. While computers at present are not nearly as capable, nonetheless this human ability defines the overarching goal to which this research contributes: to enable computers understand uncontrived images better with the help of context.

1.1 Uncontrived Images

Uncontrived images mean images of natural scenes that are encountered commonly in our surroundings. This means that the objects in an uncontrived image appear in their natural settings, without any intervention or alteration by the photographer. Furthermore, the photographer does not compose or contrive out the scene in such images. Uncontrived images do not necessarily imply photographs of nature only. They may include both indoor and outdoor themes, and will always contain objects from our real world such as car, building, sky, table, chair etc. Figure 1.1 shows several examples of uncontrived images from outdoor and indoor settings. Some prominent objects seen in the images are car, ground, tree, road, building, sign, books, moni-

Figure 1.1: A collection of uncontrived images. *Photo courtesy of LabelMe (Russell et al., 2008a)*

tor, mouse and keyboard. Many researchers used the "natural images" to refer to uncontrived images. In this thesis, the term "uncontrived image" is used to stress on the unaltered nature of scenes captured in such images, which preserves the natural contextual relation among the real-world objects within the image.

Uncontrived images deserve special attention in computer vision as many application systems such as surveillance, autonomous navigation, or robot vision encounter such images as their input and are required to *understand* them to make decisions. Uncontrived images exhibit strong regularities in their structural properties. This is extremely useful in modelling, synthesis, and recognition tasks (Gousseau and Morel, 2002; Lee et al., 2003). A good primer on the statistical properties of uncontrived images can be found in Zhu (2003). Due to the unique underlying statistical properties, uncontrived images have been a subject of interest in the field of neuroscience (Karklin and Lewicki, 2003; Wainwright et al., 2002), cognitive science (Kay et al., 2008; Yuille and Kersten, 2006), and applied mathematics (Gousseau and Morel, 2002; Srivastava et al., 2003) in addition to the field of computer vision, such as the works in (Carlsson et al., 2008; Heidemann, 2006; Jain and Seung, 2009; Kanan and Cottrell, 2010; Kavukcuoglu et al., 2010; Li, Su, Xing and Fei-Fei, 2010; Maire et al., 2008; Pajares et al., 2009; Weiss and Freeman, 2007; Zoran and Weiss, 2009).

Figure 1.2: An example of typically co-existing objects in an uncontrived image. *Photo courtesy of LabelMe (Russell et al., 2008a)*

The aspects of uncontrived images relevant to this research come from the composition and spatial arrangement of objects in the natural world. In a natural setting, whether indoor or outdoor open country, objects do not co-exist without any relation. For instance, a tree will not exist without a ground or, a desktop computer monitor will most likely co-exist with a keyboard and a mouse (Figure 1.2). When an image of an uncontrived scene is taken, these contextual relationships are retained unaltered within the image. So far, very few attempts were made to model these relationships among objects of the real world and put into use. The aim of this research is to model these contextual relationships exhibited in uncontrived images and make it available for application.

1.2 Object Class Recognition in Uncontrived Images

Object class recognition includes diverse problems from recognition of single object in an image, to recognizing multiple object instances of varying shapes, sizes and poses; or from high resolution images with clearly identifiable objects to low resolution images with faint object boundaries. Typically, researchers in computer vision narrow down their focus of attention to specific problem classes with an objective to tackle the broader object class recognition prob-

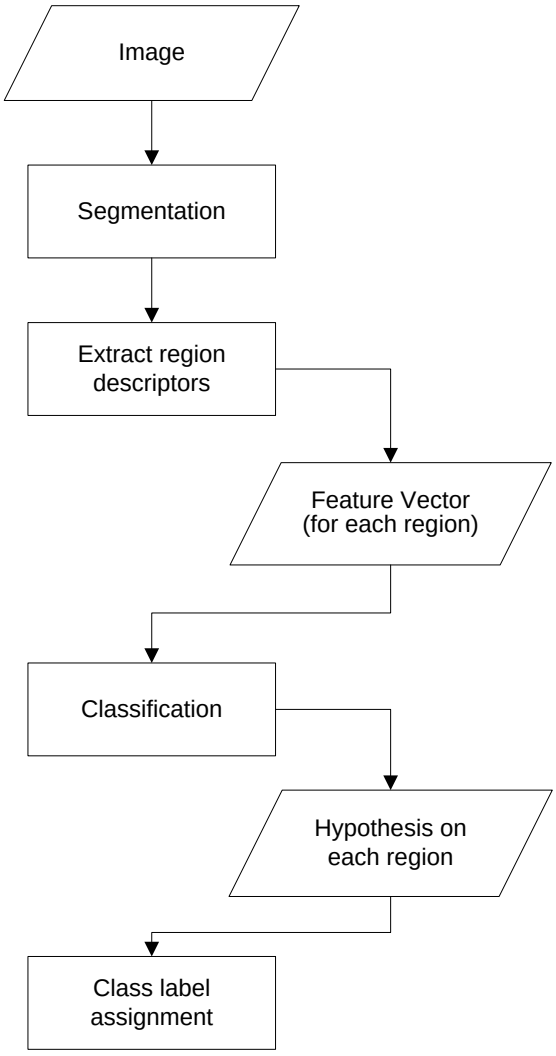

Figure 1.3: A general framework of an object class recognition system.

lem in a piecemeal fashion. Uncontrived image recognition is a type of object class recognition problem which is particularly difficult due to the unconstrained nature of objects that appear in such images. Presence of high intra-class variation, clutter, occlusion, and pose changes are the commonly cited difficulties for object class recognition in uncontrived images. The difficulty increases greatly when the objects within the image are too small or the resolution of the image is too low; in other words, when there is not enough information available to reliably identify an object. Use of contextual information becomes essential in such cases.

Fischler (1978) defined a general scheme for object class recognition (Batlle et al., 2000), modified versions of which have been implemented by many modern object class recognition

systems (Abul Hasanat et al., 2008; Kanan and Cottrell, 2010; Laptev, 2009; Lee and Grauman, 2010; Pantofaru et al., 2008; Parikh and Zitnick, 2010; Perko and Leonardis, 2010; Shotton et al., 2009; Vedaldi et al., 2010; Winn et al., 2011). The general object class recognition system starts with segmenting an image into homogeneous distinct regions or dividing an image into certain number of blocks of regions that are believed to represent real-world objects. The regions are then described with a set of descriptors local to the regions. For each region, the region descriptors are then passed to an inference engine to generate hypotheses on candidate class labels. Finally, each region is assigned the best candidate class label by the labelling process. Figure 1.3 presents an illustration of a general object class recognition system.

1.3 Context

Human beings can recognize an object in a variety of poses, illumination, and even in partially occluded conditions. A good reason behind this ability is the use of context in visual perception process. Context helps to disambiguate the visual perception when two objects appears to be visually similar. As such, context is regarded to be a vital element in both biological as well as synthetic vision systems. In the real world, objects often co-appear with other objects and in particular environments, providing contextual associations to be learned by intelligent vision systems like that of a human being (Bar and Ullman, 1996; Kleinschmidt et al., 1998) or perhaps a machine vision system.

The concept of context exists in fields other than vision as well. For instance, context is important in natural language processing to understand the meaning of a word within a sentence. In human-computer interaction, researchers are more and more interested in building applications that are aware of the user's context and can suggest solutions accordingly.

In computer vision, there are several types of contextual relations that can be used to im-

prove object class recognition performance. The very basic type of contextual relation is based on the notion that spatially neighbouring pixels are most likely to belong to the same class. This notion is generally appropriate at pixel level. Contextual relations also exists at higher levels, such as relationships among various objects or named regions within the image. The tendency of objects to co-appear with certain other objects (known as semantic context), is an example of this type of relation. Objects also tend to appear at certain spatial configuration with respect to other objects, which is known as spatial context. This research proposes methodologies to model contextual relations and develop a framework that combines different types of contexts with typical object class recognition systems.

1.4 Motivation

Real world objects do not appear in isolation in nature. They co-exist, subject to complex natural rules and nature's elaborate arrangement strategy. A car is likely to appear in any place on the ground but not in the sky. This is a fact primarily due to laws of gravitation. Similarly, the appearance of sky, mountain, and green vegetation in an open country landscape is also due to some complex natural organizational rules. But whatever the underlying reasons for the co-existence of objects in certain configurations are, human beings are capable of learning a set of complex relationships to interpret his or her surrounding visible world. These relationships provide important cues as to *what* object is most likely to co-exist with other objects for a given scene. For instance, on a city road, the human vision system would expect cars to exist. As illustrated in Figure 1.4 for a human being it is not very difficult to guess the missing object on the road in Figure 1.4a; Figure 1.4b shows the actual photograph with the missing object. This characteristic of human vision provides the key motivation to model the relationships among objects in uncontrived images in a theoretically sound manner.

Apart from relationships among objects that dictate the likelihood of co-appearance, the

(a) (b)

Figure 1.4: Photographs illustrating human perception of the likelihood of an object based on its neighbouring objects. *Photo courtesy of LabelMe (Russell et al., 2008a)*

arrangement or spatial organization of objects in a given natural environment dictate human vision *where* to expect what object. For instance, when we humans look at the sky, we do not expect a car to be there since that is a highly unlikely place to find a car. Previous research do acknowledge the importance of modelling such spatial relationships, but are mostly confined to pixel-pixel spatial relations. Taking the inspiration from human visual perception, this research is keen to model spatial relations (known as spatial context) at object level rather than pixel level relations.

In object class recognition research, context is typically modelled as region descriptors or global descriptors, in addition to local appearance-based descriptors such as texture, colour, or edge features of a given image. The combined feature set leads to a better classification compared to using local appearance features alone. While, this is encouraging, this approach has some serious drawbacks too. Since the contextual feature extraction is coupled with local appearance-based feature extraction process in such a recognition system, it is not possible to readily reuse the extracted contextual information in a different recognition system. Moreover, maintaining and updating the contextual information is difficult without disrupting the other components of the system. These predicaments motivated the proposed modular approach for integrating independent context models with typical object class recognition systems.

1.5 Problem Statement

Context is proven to be crucial for object class recognition tasks in uncontrived images where the content of the image is unconstrained. Existing context-based systems use various types of contextual information to facilitate the visual appearance based recognition process. Low-level contextual information are used by many researchers to augment appearance based local features of image regions (Kruppa and Schiele, 2003; Millet et al., 2005; Russell et al., 2008b). Such systems do not require the identity of the neighbouring objects to recognize a region of interest and rely on correlation among nearby image regions only. This leaves such systems indiscriminant to two neighbouring image regions of two different objects having similar visual attributes. Systems utilizing high-level contextual information do not suffer from this problem. High-level contextual information such as semantic and spatial context helps to disambiguate appearance based recognition hypothesis.

Prior works utilizing high-level context modelled the contextual relations in different ways. Rule-based context models are used in early vision systems (Fischler and Elschlager, 1973; Strat, 1993). Although, rule-based modelling techniques are simple and intuitive, they lack the power to accommodate uncertainty in real-world uncontrived images. Probabilistic graphical methods (directed and undirected) are proposed by many recent researchers to address the uncertainty issue. Rabinovich et al. (2007), Shotton et al. (2008), and Galleguillos et al. (2010) proposed using undirected graphs such as, Markov Random Fields or Conditional Random Fields to represent high-level semantic contextual relations. Besbes et al. (2009) and Lee and Grauman (2010) proposed undirected graph based models to represent high-level spatial contextual relations. But due to the high computational burden, they applied a threshold to limit the maximum number of neighbouring objects that can be considered. Furthermore, due to the undirected nature, such graphical models are not able to represent the dependency relations among the objects in uncontrived images. Alternatively, directed graphical methods such as

Bayesian network-based models are proposed by some researchers to overcome these issues. Im and Cho (2006) used a manually defined Bayesian network to represent contextual dependency relations among objects, their location, and low-level image features. Similarly, Sinha and Jain (2008) used a manually defined Bayesian network to model the context derived from optical meta-data associated with an image. In an attempt to model the contextual relation between an object and its neighbours Zhang and Izquierdo (2006) proposed a star-graph structured Bayesian network for every object. But, this model disregards the complex inter-object relationships for more than two objects that can be observed in real-world images. Moreover, manually defining a Bayesian network in order to represent contextual relations among object in uncontrived images is not feasible due to the complexity of the problem domain. This thesis is concerned with solving these problems by proposing a data-driven Bayesian network model that explicitly encodes the contextual relations among arbitrary number of neighbouring objects.

The success of any modelling exercise much depends on the way the model is being used. This is perhaps true for the context models developed in this research as well. The lack of a generic framework that is able to integrate probabilistic context models with existing appearance based recognition algorithms prompted this research to propose an integration framework which is extendible and allows integration of context models with existing appearance based recognition systems.

1.6 Objectives

The key objectives of this thesis are:

1. Propose a method to model semantic contextual relations among objects, independent of object class recognition process.

2. Propose a method to model spatial contextual relations among objects, independent of object class recognition process.

3. Develop a framework which integrates the proposed contextual models with typical object class recognition systems.

4. Improve object class recognition performance of existing local appearance based recognition systems through disambiguation, using the proposed context models.

1.7 Scope

The scope of the research presented in this research is outlined as follows:

1. The focus of the research is on modelling contextual relations in uncontrived imagery rather than object class recognition. Consequently, efforts are made to evaluate and demonstrate the contribution of well-modelled contextual relations in improving object class recognition performance instead of to achieve the best possible recognition performance.

2. The methodologies developed in this research is applicable to any set of objects in uncontrived images. Since an almost infinite number of objects appear in uncontrived imagery, in this thesis experimentations with the proposed methods are performed over eight object classes.

3. This research is concerned with modelling contextual relations among objects in any given dataset of uncontrived images. In this regard, it assumes that the given dataset sufficiently represents the actual contextual relations among the real-world objects.

1.8 Overview of Methodology

This section provides an overview of the proposed methodologies for modelling semantic and spatial contextual relations and the proposed framework for integration of these models with typical object class recognition systems. A schematic illustration of the methodology is presented in Figure 1.5. The key processes involved are as follows:

- Input Dataset: Serves as a common input for model learning and testing phases
- Semantic context model
  - Generate semantic context dataset
  - Learn the Semantic Context Model
- Spatial context model
  - Determine spatial configuration of neighbouring objects
  - Generate spatial context dataset
  - Learn the Spatial Context Model
- Model integration and testing
  - Local appearance-based recognition of object classes in test images
  - Query the Semantic Context Model
  - Query the Spatial Context Model
  - Adjust the initial hypotheses

Each of these processes are described in details in the following sections.

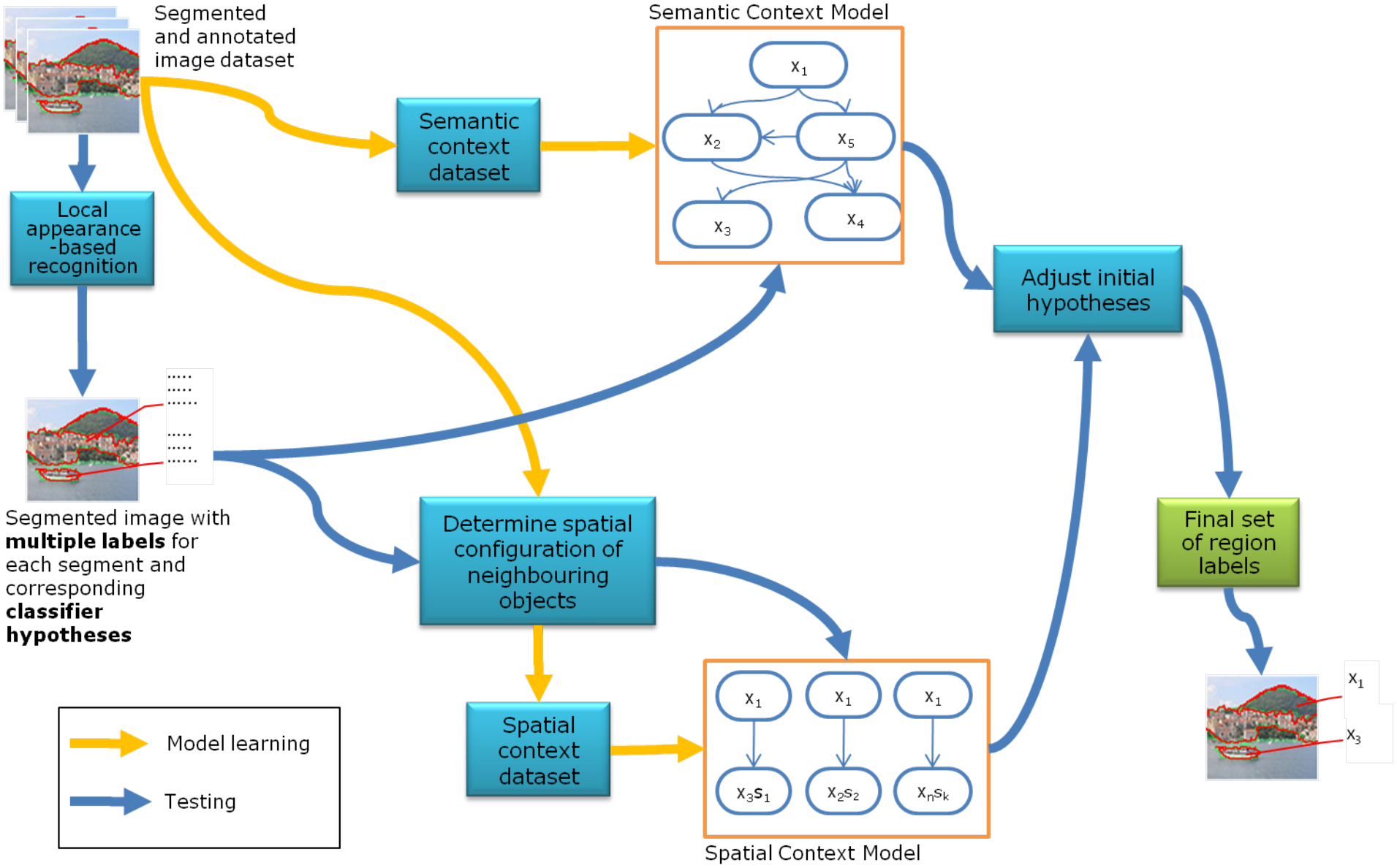

Figure 1.5: Schematic illustration of the methodology. *Photos in figure courtesy of FreeFoto.com*

1.8.1 Input Dataset

Dataset is the base for both semantic and contextual modelling methods. Since this research is interested in contextual relations among real-world objects, a suitable dataset should contain a large collection of uncontrived images. For learning contextual relations and validating the proposed models, it is also imperative to have a well segmented and properly labelled dataset. At present, the largest freely available dataset of uncontrived imagery is the LabelMe dataset from MIT Computer Science and Artificial Intelligence Laboratory (Torralba et al., 2010). With over $187,000$ images and $659,000$ labelled objects (Russell et al., 2008a), this is the most appropriate dataset for this research. Hence, in this research, non-overlapping subsets from LabelMe dataset are used for learning and testing tasks. The annotations provided with each image in the dataset are used for the learning process of the proposed context models and as the ground truth for performance evaluations. An example annotation file is is provided in Appendix A.

1.8.2 Generating Semantic Context Dataset

The semantic context dataset is created as table where the columns represent the objects and the rows represent each image. Each record in the table is a vector of binary values that encode the contextual relations within an image in terms of the appearance of each object: where the value "0" stands for absence and the value "1" stands for presence of an object. A snapshot of the semantic context dataset is provided in Appendix B.

1.8.3 Learning the Semantic Context Model (SCM)

Semantic context is the likelihood of objects to co-appear with certain other objects in uncontrived images. The aim of SCM is to model the semantic contextual relation among a set of objects in uncontrived images. A Bayesian network-based graphical model is proposed to en-

code the semantic relations among the objects appearing in uncontrived images. The nodes of the proposed graph structure represent the objects and the directed edges represent the relations. A hill climbing learning strategy is used to learn the structure of the network from the given *semantic context dataset*. Conditional probability parameters for each node is learned based on the same dataset using the maximum likelihood estimation (MLE) algorithm. The key steps involved in learning the SCM are as follows:

1. *Input:* Semantic context dataset
2. Generate random directed acyclic graph (DAG) that encodes the joint probability distribution over the variables in context dataset
3. Compute Bayesian Information Criterion (BIC) score for the DAG
4. Maximize the BIC score using hill climbing strategy
5. Determine the final DAG
6. Compute the conditional probability distribution from the semantic context dataset for each node of the DAG using MLE
7. *Output: Semantic Context Model*

1.8.4 Determining Spatial Configuration of Neighbouring Objects

This step is a perquisite step for generating the spatial context dataset in order to learn the spatial context model. This step is also required to query the spatial context model for a given test image. The process proceeds by finding the centre of mass for every segmented region within an image. Since spatial relations depends on the position of the observer, it assumes the observer is in parallel to the global horizon or the Cartesian x-axis. The angular projection between the centre of mass of the reference object and of neighbouring object is measured with

respect to the Cartesian x-axis in anti-clockwise direction. The angular projection value is then discretised into suitable number of spatial relations.

1.8.5 Generating Spacial Context Dataset

A dataset of spatial relations identified for every possible pair of objects in each image is created to be used for learning the proposed spatial context model. Each record of the spatial context dataset encodes the reference object, the neighbouring object and the corresponding spacial relation. A snapshot of the spatial context dataset is provided in Appendix C.

1.8.6 Learning the Spatial Context Model (SpCM)

Spatial context is the likelihood of finding an object in a certain spatial configuration with respect to other objects in an uncontrived image. The aim of SpCM is to model the spatial contextual relations between an object and a set of neighbouring objects in an uncontrived image. In this thesis a probabilistic model is proposed based on Bayesian formalism that can represent contextual relations between a given object and any number of neighbouring objects in an uncontrived image.

For each ordered pair of objects, and each spatial configuration of the dependent object with respect to the independent object, a conditional probability distribution is computed based on the *spatial context dataset* using the MLE algorithm. The key steps involved in learning the SCM are as follows:

1. *Input:* Spatial context dataset

2. Create DAGs for all possible pairs of objects in the dataset

3. Compute the conditional probability distribution from the spatial context dataset for each node of the DAG using MLE

4. *Output: Spatial Context Model*

1.8.7 Local Appearance-Based Recognition

Given a test image the local appearance-based recognition process take local descriptors (low-level visual features such as: texture features and shape descriptors) of each image region and employ a classifier algorithm for generating hypotheses on their candidate semantic labels with associated detection scores or probability values. These hypotheses are then refined by querying the proposed context models.

1.8.8 Querying SCM

The process of querying SCM proceeds by ranking the candidate semantic labels for each region of a test image provided by the local appearance-based recognition process based on their associated probability values. A set of considerable candidate labels for each object is then selected from the top of the ranked list. The SCM is then used to determine the most probable set of objects that may exist in the image based on the initial detection by the classifier. For this purpose, the joint probabilities of all the subsets of the set of considerable candidates with cardinality of at least 2 are calculated from the SCM and the subset with the highest probability is selected. The posterior probability of each candidate label in the selected subset is then computed using SCM by providing the remaining objects of the subset as evidence.

1.8.9 Querying SpCM

Given a new test image, the process of querying SpCM requires identifying the spatial relations between every possible region pairs in the image as mentioned in 1.8.4. The process then

proceeds by ranking the candidate labels for each region provided by the local appearance-based recognition process in descending order based on their associated probability values. The set of highest ranking candidate object labels for each region is then selected from the ranked lists. The SpCM is then used to provide the posterior probabilities of each candidate object label given its spatial relation with each neighbouring object.

1.8.10 Adjusting the Initial Hypotheses

Adjustment of the initial hypothesis process accepts input from either or both of the context models and uses a meta-classification scheme to produce revised hypothesis on each candidate label. A support vector machine (SVM) based meta-classification scheme is proposed in order to generate adjusted hypotheses by combining the hypotheses given by the local appearance-based recognition process and the context models. The final hypothesis on the correct class labels for each region is made using the standard winner-takes-all rule.

Figure 1.5 illustrates the relations among the components of the proposed methodology. In the figure, yellow coloured edges relate to the model learning processes and blue coloured edges relate to model testing processes.

1.9 Thesis Contributions

The contributions of this research are:

- Developed SCM – a probabilistic model to represent semantic contextual relations in uncontrived images.
  - SCM has a directed graphical structure allowing it to capture the dependence re-

lations among real-world objects, as opposed, to fuzzy logic based models and undirected graphical models.

  - The inference process in SCM has a time linear complexity in comparison with undirected graphical models where the complexity is NP-hard in general.

- Developed SpCM – a probabilistic model to represent spatial contextual relations in uncontrived images.

  - A unique feature of SpCM is that, it allows modelling of contextual relations with all or any arbitrary number of objects in the given image as opposed to other spatial context models which are limited to local neighbourhood objects only

- Both SCM and SpCM are self contained, can be learned and applied independent of any application systems, allowing them to be integrated with applications such as object class recognition systems in a modular fashion.

- Both the proposed models represent contextual relations in a principled manner based on sound Bayesian probability theory (Bayes, 1763; Pearl, 1988; Shen, 2007).

- Developed ConVeS – a framework to integrate contextual models with object class recognition systems in a modular fashion.

1.10 Organization of Thesis

The remainder of this thesis is organized as follows:

Chapter 2 *Context* is the key concept that this thesis is developed upon. Hence, it is imperative to explicate *context* in the context of computer vision. Chapter 2 provides a detailed discussion on what *context* is, different types of contextual information, and sources of

contextual information. This chapter also provides a functional definition of context which acts as the basis for understanding the term as used in this thesis.

Chapter 3 A detailed discussion on the existing scholarly works utilizing context for object class recognition is presented in this chapter. The chapter starts with describing the past research on object class recognition in uncontrived imagery and challenges involved. The answer for *what*, *how*, and *why* of context with regard to past research works is then provided. *What* – refers to the types of context used, *how* – refers to the different ways context was represented in object class recognition systems, and *why* – refers to the various purposes context was used in object class recognition systems. While presenting the previous works, this chapter also makes relevant criticisms and provides discussion on how this thesis is addressing the issues raised with regard to the existing works.

Chapter 4 In this chapter the semantic context model (SCM) is proposed. A detailed and systematic study on how to construct (*learn*) such a model is provided.

Chapter 5 The spatial context model (SpCM) is proposed in this chapter. It includes discussion on the probabilistic formulation of SpCM, data acquisition and model construction steps.

Chapter 6 The ConVeS framework to integrate SCM and SpCM with typical object class recognition system is presented in this chapter. This chapter reports the results of a set of experiments that were carried out to evaluate the performance of the SCM and SpCM models when used with local appearance-based recognition systems.

Chapter 7 This chapter summarize the thesis, presents the findings and concludes with a note on the limitations of the proposed models and possible future extensions.

CHAPTER 2

CONTEXT IN COMPUTER VISION

2.1 Definition of *Context* in Computer Vision

Although there is a commonly accepted lexical meaning of *context*, but the meaning widely varies based on the context it is being used. Some researchers are in the opinion that *context* should not be defined in the first place and be regarded as a primitive as Hirst explained in Hirst (2000) *"... context is what context does."* and avoided giving any strict definition of context.

This section starts with general lexical definitions of context in order to understand the notion of "context" and then attempts to narrow down the meaning of context and focuses on the computer vision domain by introducing a domain specific definition of context.

In lexical terms, Cambridge English dictionary (Cambridge University Press, 2011) defines context as:

> "the situation within which something exists or happens, and that can help explain it".

Oxford English dictionary (Oxford University Press, 2011) defines context as:

> "the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood".