Visual Recognition: Computational Models and Human Psychophysics

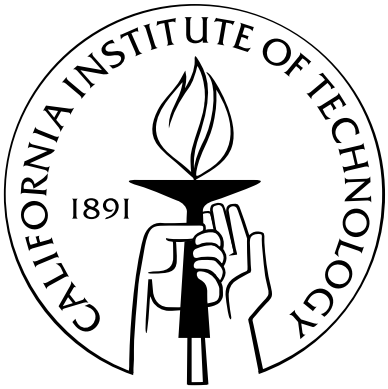
Thesis by

Fei-Fei Li

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2005

(Defended)

for my parents

Acknowledgements

I would like to thank my two advisors, Professor Pietro Perona and Professor Christof Koch, for all their help and advice during my studies at Caltech.

Over the past five years it has been my great honor to collaborate with (in alphabetical order) Rob Fergus, Asha Iyer, Silvio Savarese, Rufin VanRullen and Dirk Walther.

Many people have contributed in various ways to make this ph.D study an exciting and memorable journey. I owe sincere gratitude to Yaser Abu-Mostafa, Irvine Biederman, Christopher Bishop, Jochen Braun, David McKay, Max Welling, Andrew Zisserman, members of the Caltech Vision Lab, members of the Caltech Klab and the undergraduate students with whom I worked. I also thank all my friends, near and far, for their support and care.

Abstract

Object and scene recognition is one of the most essential functionalities of human vision. It is also of fundamental importance for machines to be able to learn and recognize meaningful objects and scenes. In this thesis, we explore the following four aspects of object and scene recognition.

It is well known that humans can be "blind" even to major aspects of natural scenes when we attend elsewhere. The only tasks that do not need attention appear to be carried out in the early stages of the visual system. Contrary to this common belief, we show that subjects can rapidly detect animals or vehicles in briefly presented novel natural scenes while simultaneously performing another attentionally demanding task. By comparison, they are unable to discriminate large T's from L's, or bisected two-color disks from their mirror images under the same conditions. We explore this phenonmenon further by removing color from the natural scenes, or increasing the number of images peripherally. We find evidence that suggests that familiarity and meaningfulness might be among the factors that determine attentional requirements for both natural and synthetic stimuli.

So what exactly do we see when we glance at a natural scene? And does what we see change as the glance becomes longer? We asked naive subjects to report what they saw in nearly a hundred briefly presented photographs. After each presentation subjects reported what they had just seen as completely as possible. Afterward, another group of sophisticated individuals who were not aware of the goals of the experiment were instructed to score each of the descriptions produced by the subjects in the first stage. Individual scores were assigned to more than a hundred different attributes. Given the evaluation of the responses, we show that within a single glance, much object and scene level information is perceived by human subjects. But the richness of our perception seems asymmetrical. Subjects tend to have a bias to natural scenes being perceived as outdoor rather than indoor.

In computer vision, it is commonly known that learning visual models of object categories notoriously requires thousands of training examples. We show that it is possible to learn much information about a category from just one image, or a handful of images. The key insight is that, rather than learning from scratch, one can take advantage of knowledge coming from previously learnt categories, no matter how different these categories might be. We explore a Bayesian implementation of this idea. Object categories are represented by probabilistic models. Prior knowledge is represented as a probability density function on the parameters of these models. The posterior model for an object category is obtained by updating the prior in the light of one or more observations. We test a simple implementation of our algorithm on a database of 101 diverse object categories. We compare category models learnt by a simple implementation of our Bayesian approach to models learnt from *maximum likelihood* (ML) and *maximum a posteriori* (MAP) methods. We find that in a database of more than 100 categories the Bayesian approach produces informative models when the number of training examples is too small for other methods to operate successfully.

We also propose a novel approach to learn and recognize natural scene categories. Unlike previous work,

it does not require experts to annotate the training set. We represent the image of a scene by a collection of local regions, denoted as codewords obtained by unsupervised learning. Each region is represented as part of a "theme." In previous work, such themes were learnt from hand-annotations of experts, while our method learns the theme distributions as well as the codewords distribution over the themes without supervision. We report satisfactory categorization performances on a large set of 13 categories of complex scenes.

Contents

Acknowledgements **iv**

Abstract **v**

Table of Contents **vii**

List of Figures **xi**

I Introduction 1

II Natural Scene Categorization in the Near Absence of Attention 3

1 Introduction **4**
1.1 Background . . . 4
1.2 Contribution . . . 5

2 General Method **6**
2.1 Subjects . . . 6
2.2 Apparatus . . . 6
2.2.1 Database . . . 6
2.2.2 Equipment . . . 6
2.3 Procedure . . . 7
2.3.1 Experimental Paradigm . . . 7
2.3.2 Central Letter Discrimination Task . . . 7
2.3.3 Peripheral Task . . . 8
2.3.4 Training Procedure . . . 8
2.3.5 Data Analysis . . . 8

3 Experiments and Results **10**
3.1 The Main Experiment: Natural Scene Categorization without Attention . . . 10
3.1.1 Method . . . 10
3.1.2 Results . . . 11
3.1.3 Discussion . . . 15
3.2 Control Experiment 1: Effects of Training . . . 16
3.2.1 Method . . . 17
3.2.2 Results . . . 18
3.2.3 Discussion . . . 21
3.3 Control Experiment 2: Effect of Color . . . 21
3.3.1 Method . . . 21
3.3.2 Results . . . 22
3.3.3 Discussion . . . 23

3.4 Control Experiment 3: Evidence for Parallel Processing . . . 23
3.4.1 Method . . . 24
3.4.2 Results . . . 25
3.4.3 Discussion . . . 26
3.5 Control Experiment 4: Multiple Copies of the Synthetic Stimuli . . . 26
3.5.1 Method . . . 28
3.5.2 Results . . . 28
3.5.3 Discussion . . . 29
3.6 Control Experiment 5: Evidence for Well-learned Categories of Objects Entailing Less Attentional Load During Recognition . . . 30
3.6.1 Method . . . 31
3.6.2 Results . . . 31
3.6.3 Discussion . . . 33

4 Summary **35**
4.1 Natural Scene Categorization Requires Little Attention . . . 35
4.2 Natural Scene Categorization Is an “Easy” Task to Learn and to Perform . . . 36
4.3 Parallel Processing . . . 37
4.4 Meaningful Categories . . . 37
4.5 Conclusion . . . 38

III Gist of Natural Scenes: Perception in a Glance **39**

5 Introduction **40**
5.1 Background . . . 40
5.2 Contributions . . . 41

6 General Method **43**
6.1 Dataset . . . 43
6.2 Experimental Stage I: Free Recall . . . 46
6.2.1 Subjects . . . 46
6.2.2 Apparatus . . . 46
6.2.3 Procedure . . . 46
6.3 Experimental Stage II: Description Evaluation . . . 47
6.3.1 Subjects . . . 47
6.3.2 Apparatus . . . 48
6.3.3 Procedure . . . 49

7 Experiments and Results **52**
7.1 Experiment I: The ‘Content’ of a Single Fixation . . . 52
7.1.1 Method . . . 52
7.1.2 Result and Discussion . . . 53
7.2 Experiment II: Outdoor and Indoor Categorization . . . 56
7.2.1 Method . . . 56
7.2.2 Results and Discussion . . . 58
7.3 Experiment III: Sensory-level Recognition vs. Object/Scene-level Recognition . . . 61
7.3.1 Method . . . 62
7.3.2 Results and Discussion . . . 63
7.4 Experiment IV: Hierarchies of Objects and Scenes . . . 65
7.4.1 Method . . . 65
7.4.2 Results and Discussion . . . 66
7.5 Experiment V: Object and Scene Perception: Are They Correlated? . . . 68
7.5.1 Method . . . 69

7.5.2 Results and Discussion . . . 69

8 Summary **73**
8.1 The Gist of Gist . . . 73
8.2 Shapes, Objects and Scenes . . . 74
8.3 Two Puzzling Asymmetries? . . . 75

IV Computational Models I: Object Recognition **77**

9 Introduction **78**
9.1 Introduction and Motivation . . . 78
9.2 Literature Review . . . 79
9.3 Contribution . . . 80

10 A Bayesian Model **81**
10.1 Overall Bayesian Framework . . . 81
10.2 The Object Category Model . . . 82
10.2.0.1 Appearance . . . 83
10.2.0.2 Shape . . . 84
10.2.1 Discussion of model . . . 85
10.2.2 Form of the Parameter Posterior . . . 86
10.2.3 Maximum Likelihood (ML) and Maximum A Posteriori (MAP) . . . 86
10.2.3.1 Other Inference Methods . . . 87
10.2.4 Conjugate Densities . . . 88
10.3 Recognition Using a Conjugate Density Parameter Posterior . . . 88
10.3.1 Parameter Distribution . . . 88
10.3.2 Closed-form Calculation of R . . . 89
10.4 Learning Using a Conjugate Density Parameter Posterior . . . 89

11 Experiments and Results **91**
11.1 Implementation . . . 91
11.1.1 Feature detection and representation . . . 91
11.1.2 Learning . . . 91
11.1.2.1 Choice of Prior . . . 91
11.1.2.2 Details of Bayesian One-Shot algorithm . . . 92
11.2 Experimental Results . . . 93
11.2.1 Datasets . . . 93
11.2.2 Experimental Setup . . . 94
11.2.3 Walkthrough for the Motorbike Category . . . 94
11.2.4 Caltech 4 Dataset . . . 98
11.2.5 101 Object Categories . . . 99
11.2.5.1 Overall Results: ML vs. MAP vs. Bayesian . . . 99
11.2.5.2 Good models and Bad models . . . 104
11.2.5.3 A Further Investigation on Prior Models and Feature Detectors . . . 108
11.2.5.4 Bayesian One-Shot Algorithm: Shape-Only vs. App-Only vs. Shape-App models . . . 109
11.2.5.5 Bayesian One-Shot Algorithm: Discrimination Amongst 101 Categories . . . 109
11.2.5.6 Discussions . . . 111

12 Summary **114**

V Computational Model II: Natural Scene Classification 116

13 Introduction 117
13.1 Background . . . 117
13.2 Contributions . . . 118

14 Hierarchical Bayesian Model and Learning 119
14.1 Model Structure . . . 119
14.1.1 The Theme Models . . . 120
14.1.2 Bayesian Decision . . . 123
14.1.3 Learning: Variational Inference . . . 124
14.1.4 A Brief Comparison . . . 125
14.2 Features and Codebook . . . 125
14.2.1 Local Region Detection and Representation . . . 125
14.2.2 Codebook Formation . . . 126

15 Experiments & Results 127
15.1 Dataset and Experimental Setup . . . 127
15.2 Results . . . 129

16 Summary 135

Bibliography 137

A 149
A.1 Definition of Various Densities and Functions . . . 149
A.1.1 Dirichlet Distribution . . . 149
A.1.2 Normal-Wishart Distribution . . . 149
A.1.3 Gamma Distribution . . . 150
A.1.4 Multivariate Student's T Distribution . . . 150
A.1.5 Kullback-Leibler Distance . . . 150
A.1.6 Digamma Function . . . 150
A.2 Learning using a conjugate density parameter posterior . . . 150
A.2.1 Variational methods . . . 151
A.2.2 Variational Bayesian EM . . . 151
A.2.3 The E-step of Bayesian One-Shot . . . 152
A.2.4 The M-step in Bayesian One-Shot . . . 152
A.3 MAP learning . . . 153
A.3.1 Expectation Maximization (EM) for MAP . . . 154

List of Figures

2.1 Dual-Task experimental setup for a single trial. . . . 7

3.1 Experimental Protocol. . . . 12
3.2 Main Results. . . . 13
3.3 Normalized performances of several experiments. . . . 14
3.4 Cross-Training experiment I and results. . . . 18
3.5 Cross-Train experiment II and results. . . . 19
3.6 Natural scene categorization without color. . . . 22
3.7 Experimental setup for single-image versus double-image experiment. . . . 24
3.8 Results of the single-image versus double-image experiment. . . . 27
3.9 Multiple stimuli experiment. . . . 29
3.10 Rotated versus fixed rotation versus upright letter experiments. . . . 32

6.1 46 images of outdoor scenes in our dataset of 90 grayscale images. . . . 44
6.2 44 images of indoor scenes in our dataset of 90 grayscale images. . . . 45
6.3 A single trial in Stage I. . . . 47
6.4 Attribute Tree. . . . 48
6.5 Experiment Stage II: Evaluating the free recall responses. . . . 48
6.6 A sample score plot for the building attribute. . . . 50

7.1 Subject description samples. . . . 53
7.2 Results of single fixation experiment. . . . 55
7.3 Power spectral analysis of indoor and outdoor scenes. . . . 57
7.4 Template analysis for indoor and outdoor scenes. . . . 58
7.5 Categorization results of indoor and outdoor scenes. . . . 59
7.6 Categorization results of manmade outdoor and natural outdoor scenes. . . . 59
7.7 Summary plot of average categorization performances of all 7 SOAs. . . . 60
7.8 Sensory information and object perception in outdoor and indoor scenes. . . . 60
7.9 Samples of subjects' free recall responses to images at different SOAs. . . . 62
7.10 Perceptual performances of different attributes across all 7 presentation times. . . . 63
7.11 Perceptual performances of different object attributes across all 7 presentation times. . . . 65
7.12 Perceptual performances of different scene attributes across all 7 presentation times. . . . 65
7.13 Object recognition performance versus scene recognition performance at various different SOAs. 70
7.14 Overall correlation coefficients for scene versus objects and breakdowns. . . . 70

10.1 Schematic comparison of ML and MAP methods. . . . 87

11.1 Output of the feature detector on samples images from four categories. . . . 92
11.2 The 101 object categories and the background clutter category. . . . 95
11.3 Training images for the motorbike category. . . . 96
11.4 A visualization of the prior parameter density. . . . 97
11.5 The learning process for motorbike category. . . . 98
11.6 The mean model for motorbike category. . . . 99

11.7 Summary of the face model. . . . 100
11.8 Summary of the motorbike model. . . . 101
11.9 Summary of the spotted cat model. . . . 102
11.10 Summary of the airplane model. . . . 103
11.11 Performance on 101 categories using three different learning methods. . . . 105
11.12 Results for the grand-piano category. . . . 106
11.13 Results for the “cougar face” category. . . . 107
11.14 Two categories with poor performance. . . . 107
11.15 Effect of prior models on object categorization. . . . 108
11.16 Quality of feature detection compared with object detection performances. . . . 108
11.17 Full model versus shape only and appearance only models. . . . 109
11.18 Discrimination experiment. . . . 112
11.19 A toy example to illustrate the effect of prior knowledge. . . . 113

14.1 Flow chart of the algorithm. . . . 120
14.2 Different models of scene categorization. . . . 121
14.3 Codebook of textons. . . . 122

15.1 13 categories of natural scenes. . . . 128
15.2 Internal structure of the models learnt for each category. . . . 130
15.3 Examples of testing images for each category. . . . 131
15.4 Confusion and rank statistics tables of Theme Model 1. . . . 132
15.5 Example of themes for the forest category. . . . 132
15.6 Dendrogram of the relationships of the 13 category models. . . . 133
15.7 Codeword number vs. category number and performance comparison of the 3 algorithms. . . . 133
15.8 Performance vs. training examples, theme numbers and codewords. . . . 133

Part I

Introduction

To understand how humans see and to build machines to see are two important goals in science and engineering. The purpose of this thesis is two-fold. On the human vision side, we explore properties of natural scene recognition through psychophysics experiments. On the computer vision side, we propose two algorithms that learn and recognize objects and natural scenes.

We will introduce the significance of our questions as well as our contribution separately in each of the following part:

In Part II, we present a series psychophysics studies that shows rapid natural scene categorization requires very little attention. We further explore various aspects of this astonishing ability.

In Part III, we take the question of natural scene recognition further by studying what human subjects perceive in a glance of a real-world image.

In Part IV, a computer vision model is proposed for learning object categories with very few training examples. We use Bayesian learning technique to incorporate useful prior information to achieve this goal.

Finally in Part V, we show that natural scene categorization could be done without much human supervision in a hierarchical Bayesian model.

Part II

Natural Scene Categorization in the Near Absence of Attention

Chapter 1

Introduction

1.1 Background

Psychologists have long known that certain visual search tasks require minimal or no attention. A hallmark of preattentive vision is that it is achieved in a seemingly parallel fashion: a preattentive task may be carried out simultaneously with other visual tasks [17]; target detection does not become significantly more difficult when the number of distractors is increased [15, 140]. However, none of the known preattentive tasks approaches the sophistication of everyday vision where complex scenes must be scrutinized in order to assess high level properties such as the presence of danger and the structure of a social interaction. Virtually all of the visual tasks that may be performed preattentively have been explained, either in detail or in principle, by quasi-linear models that replicate mechanisms found in the early stages of our visual system [7, 83]. While much can be accomplished by these simple mechanisms, it is quite clear that they are inadequate for explaining "high level" perception such as recognition and categorization, i.e., visual processes that rely on neural activities in the inferior temporal cortex and beyond [43, 68, 78]. This would suggest that there is no sophisticated property of the scene that we can see without paying attention. In agreement with this view, change blindness and inattentional blindness studies demonstrate that without visual attention, significant changes in a large part of the visual field can easily escape our awareness [82, 96, 115, 128].

On the other hand, some complex visual tasks can be rapidly accomplished by our visual system. RSVP (rapid serial visual presentation) experiments have demonstrated that natural objects belonging to a specified category may be classified remarkably fast [109, 130]. Thorpe and colleagues have found that complex natural scenes can be categorized in as little as 150ms [22, 29, 135, 136, 146]. This astonishing speed relative to the time constant of information processing and transmission in networks of neurons raises the question of whether attention plays a critical role in this type of rapid visual processing.

In this study, we would like to explore the relationship between rapid natural-scene categorization and

visual attention. More specifically, we ask whether a complex scene recognition task require much attention and to what extent attention plays a role in various types of visual tasks.

1.2 Contribution

To the surprise of current attention theories [140], we find that there is little or no attentional cost in rapid visual categorization of complex, natural images [77]: detecting the presence of an animal, or a vehicle, in a natural photograph, a genuinely challenging task for today's computer vision algorithms, can be carried out by human observers in the near-absence of attention. Subjects could perform this task equally well alone or simultaneously with another attention-demanding task (i.e., deciding whether 5 randomly rotated Ts and Ls presented at fixation were all identical or whether one of them differed from the others). Under the same dual-task conditions, subjects could not perform apparently simpler discrimination tasks involving synthetic stimuli (discriminating between a single peripherally rotated T or L, or discriminating between red-green and green-red bisected disks).

While the main result implies that natural scenes probably hold a special status for our visual systems, it is unclear exactly what about these stimuli is responsible for this distinctiveness: is it the mere fact that a picture is natural rather than synthetic, or are there some associated (or confounded) factors that could be responsible for determining attentional requirements? We follow these questions in the second set of experiments that investigate a variety of such potential factors and establish how they can affect the attentional requirements of recognition tasks using natural and artificial stimuli. We find that natural scene categorization without attention requires little stimulus-specific training. It is robust to lack of color information or increasing the set size of the stimuli presented. In contrary, multiple redundant copies of synthetic stimuli do not improve the performances of recognition without attention. Some simple tasks, such as single letter discrimination, require much attentional assistance unless the letters are presented in a familiar, upright position. We hypothesize that attention is particularly important for tasks that do not have neuronal representations in the visual pathway. Natural scene categorization, a well-learned and familiarized task for most human observers, does not require much attention.

Chapter 2

General Method

2.1 Subjects

Fifteen highly motivated California Institute of Technology undergraduates and graduate students (from 20 to 26 years old) served as subjects in all or part of the following experiments. Each subject enrolled for at least 15 daily sessions of 1 hour and received payment. Subjects reported normal color vision and visual acuity (sometimes with corrective lenses or glasses), but underwent no tests in this respect. All subjects were right-handed. All subjects were naive about the purpose of the experiments.

2.2 Apparatus

2.2.1 Database

The pictures were complex color scenes taken from a large commercially available CD-ROM library allowing access to several thousand stimuli. The animal category images included pictures of mammals, birds, fish, insects and reptiles. The vehicle category images included pictures of cars, trucks, trains, airplanes, ships and hot-air balloons. There was also a very wide range of distractor images, which included natural landscapes, city scenes, photos of food, fruits, plants, houses and artificial objects.

2.2.2 Equipment

Subjects were seated in a dark room especially designed for psychophysics experiments. The seat was approximately 100cm from a computer screen, which was connected to a Macintosh (OS9) computer. The refresh rate of the monitor was 75Hz. All experimental software was programmed using the Psychophysics Tool box [14, 101] and Matlab.

2.3 Procedure

2.3.1 Experimental Paradigm

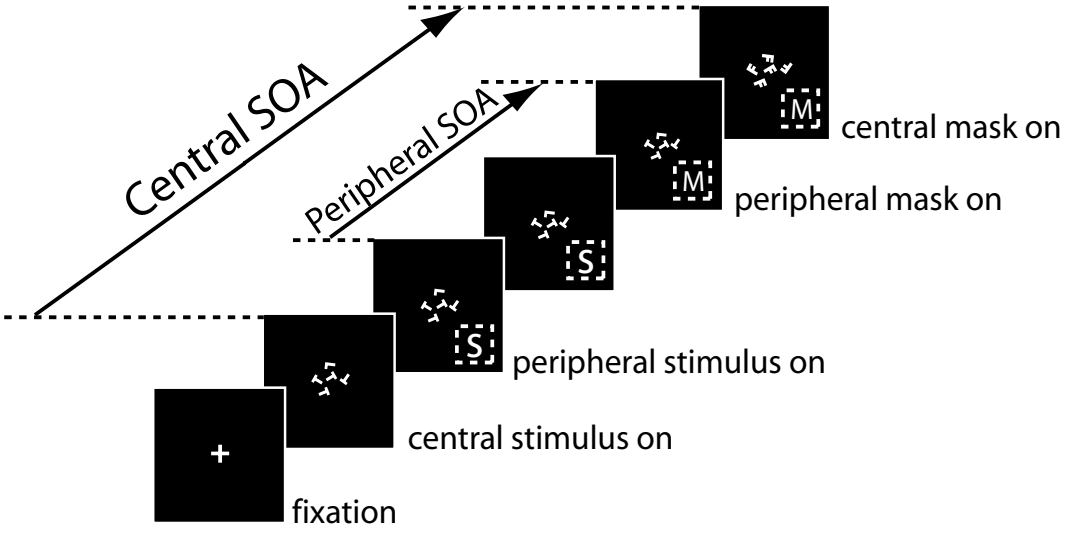


Figure 2.1: Dual-Task experimental setup for a single trial. A fixation cross of 1-degree visual angle is flashed for 300ms at the onset of each trial, started by the subject. After that, the central letter discrimination task stimuli are presented for a Central SOA (Stimulus Onset Asynchrony) amount of time. The central stimuli are then masked by an appropriate perceptual mask. Central SOA is determined individually for each subject so that the performances of the central letter discrimination task center around 80% correctness. 52ms after the onset of the central stimuli, a peripheral stimulus is presented randomly at a peripheral location centering at 6-degree eccentricity. For different experiments reported in this paper, peripheral stimuli vary. Each peripheral stimulus is then masked by its corresponding perceptual mask, after Peripheral SOA amount of time. It is important to note that to ensure that attention is properly withdrawn by the central task under the dual task condition, the peripheral mask always onsets earlier than or at the same time as the central stimulus mask.

We use a Dual-Task paradigm in all of our experiments [15, 129]. Each experiment consists of three different conditions: the primary task– an attentionally demanding central task (identical in all experiments), a secondary peripheral task (in which the role of attention is investigated) and a dual task condition in which both the central and peripheral tasks are performed concurrently. In each experiment, all trials are organized in the same way irrespective of the experimental condition (i.e., single-task condition or dual-task condition). Only the number of required responses varies between conditions.

2.3.2 Central Letter Discrimination Task

In all experiments, each trial starts with a fixation cross 300 ± 100ms before the onset of the central stimulus. At 0ms, the central stimulus (a combination of five letters) is presented. The five letters (Ts and Ls, either all identical, or one differing from the other four), appear at nine possible locations within 1.2° eccentricity.

Each letter is randomly rotated. After the central SOA (Stimulus Onset Asynchrony, the time between the appearance of the central stimulus and the onset of the central mask), each stimulus letter is masked by the letter 'F' rotated according to the random orientation of the stimulus letter. For a given subject, the central SOA is the same for both single-task and dual-task conditions. All trial types are presented with equal probability. Subjects are instructed to respond by pressing 'S' on the keyboard if the five letters are the same, or 'D' if one of the letters differs from the other four. Fig. 1 illustrates schematically the setup of a sample trial of the dual task paradigm. In earlier studies it was found that the central task performance is quite a sensitive measure to indicate the allocation of attentional resources [15]. Subjects' performances on this central task decreased drastically if the SOA was slightly decreased [77].

2.3.3 Peripheral Task

In each peripheral task, the stimulus is always presented 53ms after the central stimulus onset. Subjects respond to these tasks in a speeded fashion. They are instructed to continuously hold down the mouse button and release it as fast as possible (within 1000ms) when they have detected the target. For a given trial, the location of the peripheral stimulus is randomly determined, keeping a distance of 6-degree eccentricity (Fig. 2.1).

2.3.4 Training Procedure

Each novel subject to the dual task paradigm underwent a training process. It usually took more than 10 hours for a new subject to coordinate his/her motor responses well enough to answer both a speeded peripheral task and the central task. The central SOA, starting at 500ms, was decreased after each block where the performance of this task exceeded 85% correct. The training procedure was terminated after the subject's performance had stabilized and the central SOA was below 250ms. This value is chosen to limit the possibility of switching attention or eye movement during stimulus presentation. Central task and peripheral task always received the same amount of training.

2.3.5 Data Analysis

For each subject in a given experiment, we obtain two baseline performances: central letter discrimination with attention (single-task condition) and peripheral recognition task with attention (single-task condition). Each of these two performances consists of performances of 9-15 blocks (depending on the experiment) of 96-trial experiments unless otherwise specified. Similarly, we also obtain the corresponding performances for the central letter discrimination task with attention (primary task) and peripheral recognition task without

attention under the dual-task condition. Each of these two performances consists of performances of 9-15 blocks (depending on the experiment) of 96-trial experiments unless otherwise specified. T-tests are computed for each experiment to compare single- and dual-task performances. An alpha value of 0.05 is used for all statistical tests.

To visualize results, we summarize each of the experiment using a "normalized performance" figure. The "normalized performance" for each task is obtained in the following way. The averages of the two baseline performances are linearly scaled to 100% such that chance level performance remains at 50%. Then the same scaling factor for each subject is used to obtain normalized performance levels of the two tasks under the dual-task condition. In other words,

$$\text{Normalized performance} \quad = \quad 0.5[(P_d - 0.5)/(P_s - 0.5)] + 0.5 \tag{2.1}$$

where P_d and P_s refer to performance in the dual-task and single-task conditions, respectively. It is important to point out that since the 100% baseline is the average performance of a given task under the single-task condition (with attention), it is possible that the normalized performance, same task's performance under the dual-task condition (without attention) might sometimes be larger than 100%. This simply means that the actual performance has a higher average under the dual-task condition than under the single-task condition. Statistical tests will determine whether this difference is significant or not.

Chapter 3

Experiments and Results

In this chapter, we present first the main experiment of our study. The key question of our investigation is whether natural scene categorization requires visual attention. Following the main experiments, we present a series of control experiments targeted to investigate various aspects of our findings in the main experiments.

3.1 The Main Experiment: Natural Scene Categorization without Attention

We studied the role of attention in natural scene categorization using a dual task paradigm, in which a natural scene categorization task, where target scenes were defined by the presence of one or more animals, was performed concurrently with another visual task that required visual attention [17, 74, 129] (Fig. 3.1). The idea is to compare subjects' performances of the categorization task under two conditions: the "single task" condition where attention is available, and the "dual task" condition where attention is drawn away by the other task. If the rapid natural scene categorization task demands attention, we should observe a significant decrease in performance under the dual task condition. If the rapid natural scene categorization does not entail much attentional cost, performances should be comparable.

3.1.1 Method

As described in Chapter 2, we use a dual task paradigm. The central task is attentionally demanding. It involves discriminating displays composed of five randomly rotated Ts and Ls at the center of the visual field. Subjects needed to respond by pressing one key when all five letters were the same and another key when one of the letters differed from the other four. This task engages attention at the center of the display, preventing attention from focusing on the natural scene in the periphery [17, 74] (see also Fig. 3.3d-e). When our subjects performed this task alone, their performances averaged around 77% (varied between 68% and

82%, Fig. 3.2). This value can be used as a reference for the dual task condition: if a subject has continuously engaged full attention to the central task, we expect the performance to be maintained at the same level; any significant distraction or withdrawing of attention would decrease performance.

The peripheral task is our locus of interest–natural scene categorization. It is a modification of the one used by Thorpe and colleagues [135]. A picture was flashed for only 27ms at a random location in the periphery of the visual field, followed by a perceptual mask (Fig. 3.1). Subjects had to decide whether the image contained an animal (or animals) or not, as fast and accurately as possible [135]. When subjects performed this task alone, their performance averaged around 76% (ranging from 75% to 79%; Fig. 3.2).

3.1.2 Results

Under the dual task condition, subjects were instructed to focus attention at the center of the display, and to try to perform both tasks as accurately as possible. Since we were interested in the reaction times of the natural scene categorization task, we asked subjects to respond as fast as possible to the peripheral task before answering the central one. For each subject, the central task performance under the dual task condition showed no difference ($p>0.05$) from its counterpart under the single task condition (Fig. 3.2). This is a clear indication that attention was locked at the center under the dual task condition. Furthermore, for each individual subject the average peripheral categorization performance under the dual task condition was not significantly ($p > 0.05$, t-test) different from the corresponding performance under the single task condition (Fig. 3.2), suggesting that natural scene categorization can still be performed when attention is drawn away (see also Fig. 3.3a-c).

One might argue that subjects could first attend to the peripheral stimulus before switching attention to the central one. In that case, however, the time available to process the central stimulus would be much shorter by at least 80ms than the actual central SOA (the peripheral stimulus is turned off 80ms after the onset of the central stimulus). This strategy would result in a strong decrease in performance of the central task. Indeed, in a separate control experiment, we asked all six subjects to perform the central letter task with an SOA shortened by only 66ms. Their average performance dropped from 77% to 66% (individual t-test for each subject, $p = 0.01$). This confirms that our results do not reflect a systematic switch of attention between the two tasks.

Because of its high motor coordination demands, the dual task required extensive training. During this period, our subjects were repetitively trained with the same set of 288 images. It could be argued that such training could serve to optimize feature detection mechanisms for specific stimuli, reducing the attentional demands for this task [16, 65]. However, the above results were obtained with a set of 1056 novel images that

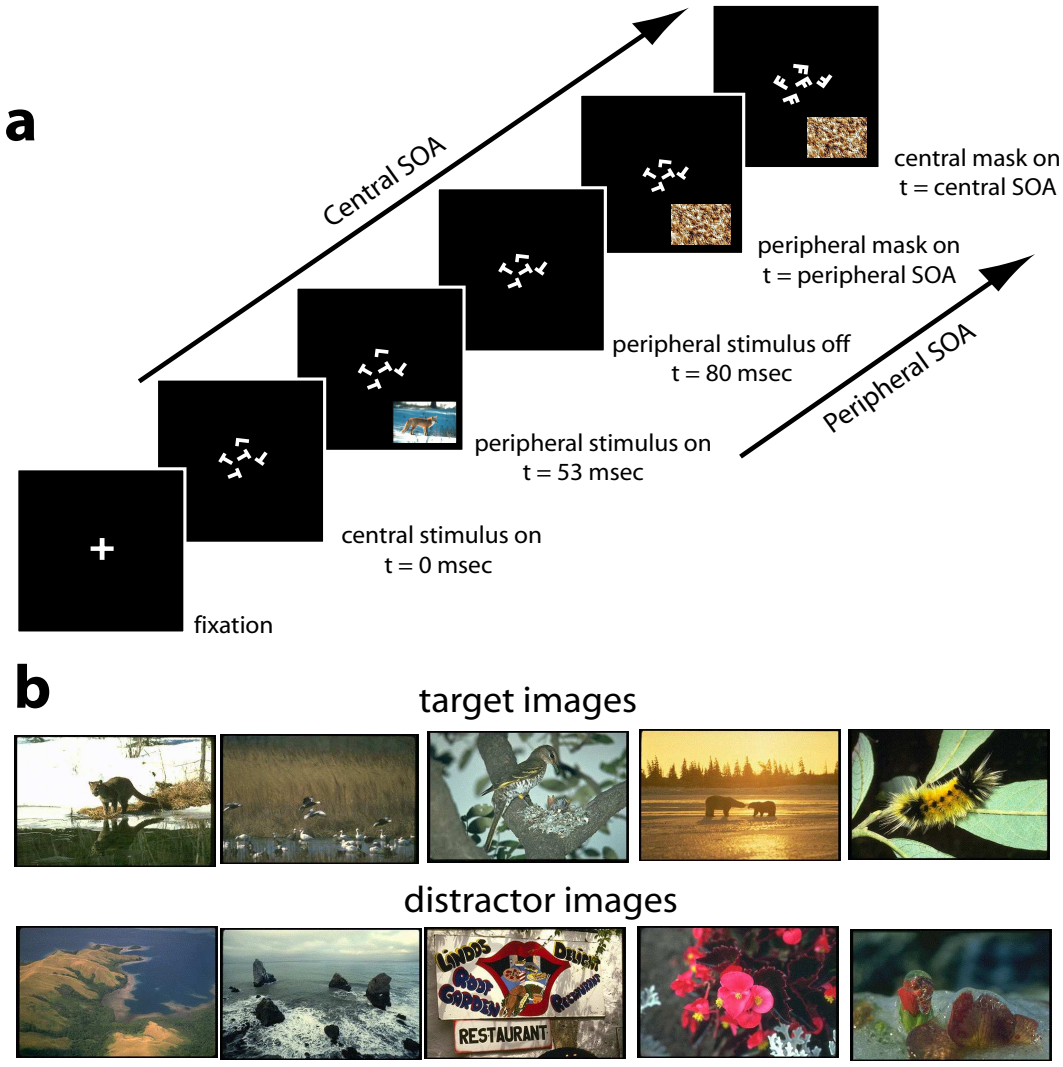


Figure 3.1: Experimental Protocol. (a) Schematic illustration of one trial. After a fixation cross presented at the center of the visual field, an attentionally demanding letter discrimination task is presented centrally. The central stimulus (combination of five Ts and Ls) is then replaced by a perceptual mask (five Fs) after a time interval commonly called the stimulus onset asynchrony (SOA, ranging from 133ms-240ms for different subjects). Subjects are instructed to respond whether all five letters are the same or one of them is different. In the peripheral natural scene categorization task, an image is presented peripherally for 27ms at a random location and 53ms after the onset of the central stimulus. The peripheral stimulus is followed (after peripheral SOA) by a perceptual mask. The peripheral SOA varies individually for each subject, ranging from 53ms to 80ms. The peripheral mask always appears before the central stimulus is replaced by its own mask. Subjects make a speeded response to the presence of animals. Under the dual task condition, subjects are required to perform both tasks concurrently. (b) Sample Images of the Stimulus Database. The pictures are complex color scenes taken from a large commercially available CD-ROM library allowing access to several thousand stimuli. The animal category images include pictures of mammals, birds, fish, insects, and reptiles. In a separate experiment (Fig. 3.3b-c), an additional target category is used-vehicles. The vehicle category images include pictures of cars, trucks, trains, airplanes, ships and hot-air balloons. There is also a very wide range of distractor images, which include natural landscapes, city scenes, pictures of food, fruits, plants, houses, and artificial objects.

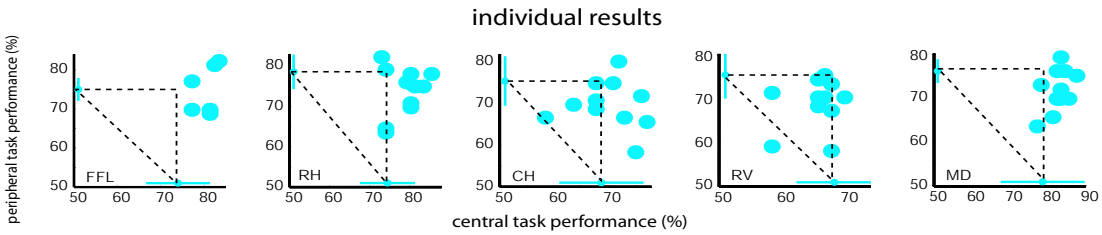


Figure 3.2: Main Results. Individual subject's results for dual vs. single task performance (5 subjects). The horizontal axis represents performance of the central task (attentionally demanding letter discrimination task). The vertical axis represents performance of the peripheral task (natural scene categorization). Each solid • represents the performance of one block (96 trials/block) under the dual task condition. All images used for testing were novel for the subjects. Each open ○ represents the average performance under the single task condition. For each subject, performances of the letter discrimination task do not differ significantly (t-test, $p>0.05$) under the single and dual task conditions, suggesting that attention was fully allocated to the center in the dual task condition. Furthermore, the performances of the natural scene categorization task do not differ significantly (t-test, $p>0.05$) either under the single and dual task conditions, suggesting that the task may be performed while attention is engaged elsewhere.

were never presented during training. Furthermore, we show later (Fig. 3.3d and 3.3e) that the same amount of training in other dual tasks did not reduce attentional demands. This makes it unlikely that our results are a direct consequence of the training process. In addition to our experiments, a study done by Rousselet et al. reaches a compatible conclusion with untrained subjects [120].

Reaction times measured under the single task condition are compatible with results observed by Thorpe and colleagues, suggesting that our natural scene categorization task is performed in an ultra-rapid mode [135]. Note that this task involves a speeded response under both single and dual task conditions. Under the dual task condition, while categorization performance is unaffected, we observe an average delay of 117ms in response times compared to the single task condition (single task: 491ms; dual task: 608ms). This delay is likely to arise due to central rather than perceptual attentional competition [100]. Indeed, when subjects are required to perform two tasks simultaneously, interference is known to occur at several different stages: task preparation [50], response selection [99, 154] and response production [57, 92]. These limitations, often referred to as the "psychological refractory period" [134], could easily account for the observed delay [100]. Moreover, a number of studies have shown that the presence of attention decreases perceptual latencies [58] and reaction times to a significant extent [69, 107, 112]. This could also explain the observed delay.

Are the above results due to the high biological and evolutionary relevance of the target category "animal?" In other words, could we obtain a similar result using a man-made object category, e.g., vehicles [146]? We tested one group of five subjects with both categorization tasks. In the vehicle task, target images included cars, trains, airplanes, ships, etc. Half of the distractors were animal scenes, while the other half contained neither animals nor vehicles (Fig. 3.3c). The animal task was essentially the same as in the main experiment

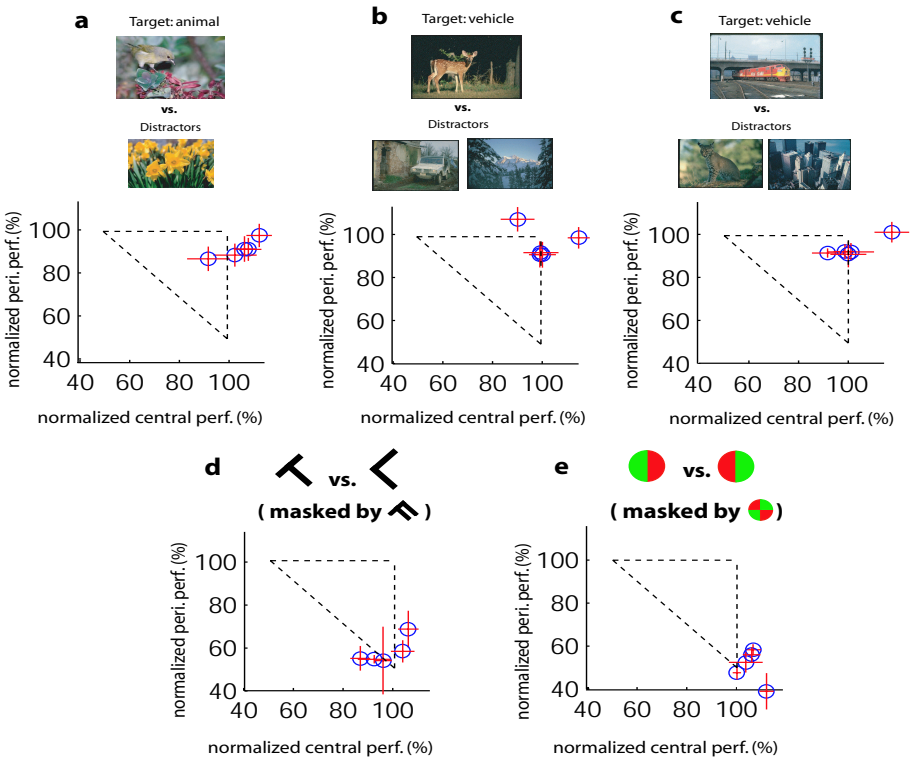


Figure 3.3: (a) Summary: Categorization of Masked Animal Images. This panel corresponds to the normalized average performance of the main experiment (Fig. 3.2). Each open circle ($\circ$) is the average value of one subject's dual task performance, normalized according to his/her own single task performance: a linear scaling transforms the average single task performance into 100%, leaving chance at 50%. Performance under the dual task condition that is higher than the corresponding performance under the single task condition would result in a normalized performance higher than 100%. Error bars reflect the standard error of the means. (b and c) Categorization of Natural and Artificial Objects. The same five subjects performed the following two categorization tasks in alternating blocks. (b) Categorization of masked animal images among vehicles and other distractors. Distractors for this task include fifty percent vehicle scenes and fifty percent non-animal/non-vehicle scenes, randomly drawn from the same database described in Fig. 3.1b. Task performance for each of the 5 subjects is comparable under dual task and single task conditions (t-test, $p > 0.05$). This panel presents a summary of normalized average performance of each subject as detailed in panel (a). (c) Categorization of masked vehicle images. Subjects are instructed to perform the natural scene categorization task using vehicles as targets (including cars, trucks, trains, airplanes, ships and hot-air balloons). Distractors for this task include fifty percent animal scenes and fifty percent non-animal/non-vehicle scenes. The panel illustrates normalized dual task performances of the 5 subjects. For each subject, task performance is comparable under dual task and single task conditions ($p > 0.05$). These experiments provide evidence that artificial as well as natural target categories can be detected in the near absence of attention. (d - e) Control Experiments. (d) Peripheral letter discrimination task. 5 subjects are instructed to discriminate between the letters T and L presented in the periphery. The letter, randomly rotated, is masked by the letter F after the peripheral SOA (ranging from 53ms to 160ms). For each subject, this peripheral letter discrimination task cannot be performed above chance in the absence of attention (paired t-test, $p > 0.05$). This panel presents a summary of normalized average performance of each subject. (e) Peripheral color pattern discrimination task. 5 subjects are instructed to discriminate a red/green color disc from a green/red color disc. The stimulus is masked after the peripheral SOA (ranging from 66ms to 106ms). For each subject this peripheral color pattern discrimination task cannot be performed above chance in the absence of attention (paired t-test, $p > 0.05$). The results from these control experiments demonstrate that our central discrimination task effectively withdraws attention away from the peripheral task (1). This panel presents a summary of normalized average performance of each subject.

(Fig. 3.2 and Fig. 3.3a), with the exception that 50% of the distractor images contained vehicles (Fig. 3.3b). The two tasks were presented in alternation and all stimuli were masked. Our results show that for each individual subject there is no significant decrease in categorization performance under the dual task condition compared to the single task condition in both cases (Fig. 3.3b - c, t-test, $p > 0.05$). This result suggests that categorization of natural scenes in the near absence of attention might well be a general phenomenon not limited to evolutionarily relevant object categories. Another possible confound is that the subjects may not be performing an animal (or vehicle) detection task, but rather may be detecting the presence of a "foreground object." Foreground objects may be more frequent in images containing animals or vehicles than in images containing scenery only. However, the fact that animal photographs were used as distractors for the vehicle task and vice versa makes this possibility implausible since "foreground objects" were contained both in the target and distractor images.

The interpretation of our findings relies on the assumption that attention is allocated to the center of the visual field under the dual task condition. This assumption is supported by the fact that there is no decrease in the central performance under dual task compared to single task conditions. This implies that when the peripheral task does demand attention, performance should suffer. To examine this question, we conducted two control experiments in which the peripheral tasks involved either discriminating a briefly presented letter followed by a mask (T or L followed by F; Fig. 3.3d) or discriminating a briefly presented and masked color disk (red/green or green/red; Fig. 3.3e). These tasks have been shown by Braun and colleagues to require attention [74]. In both of these control experiments, the central task was the same as in our previous experiments (five Ts and Ls discrimination). We observed a sharp drop in performance of both peripheral tasks ($p < 0.0001$ in Fig. 3.3d; $p < 0.0001$ in Fig. 3.3e). While subjects can perform at 74% and 78% in peripheral single letter and color tasks, respectively, they cannot do any better than chance (individual paired t-test for each subject, $p > 0.05$; average over all subjects is 51% for letter task; 51% for color task) during the dual task scenarios. These results demonstrate that attention is effectively allocated to the central task and provide further evidence that extensive training does not necessarily result in an improvement of performances. Subjects performing these dual tasks received the same amount of training as those performing the natural categorization tasks.

3.1.3 Discussion

Our findings show that rapid visual categorization of novel natural scenes requires very little or no focal attention. Perception outside the focus of attention has mostly been reported for simple salient stimuli [17, 140]. In our task, however, human subjects are actively searching for a complex category of objects whose

appearance is highly variable. It thus appears that a sophisticated high level of representation (e.g. semantic) can be accessed outside the focus of attention. It has already been argued that the "gist" of a visual scene could be available preattentively [8, 157]. In this context, the contents of the "gist" could in fact be extended to include information about the presence of a complex target category whose appearance is not known in advance.

3.2 Control Experiment 1: Effects of Training

We had demonstrated the human visual system's amazing efficiency in natural scene categorization with little or no attention [77]. In these experiments, an average training period of 10-15 hours on dual-task was necessary for each subject. It is likely that this training helps sharpen the executive control necessary for performing different tasks, particularly when they are carried out simultaneously [100, 126]. However, training sometimes also decreases the attentional demand on perceptual processing [65]. So, could it be that this superb efficiency in natural scene processing is mainly due to the training process that each subject had received in these experiments? We had argued that this is unlikely since the same amount of training was applied both to the natural scene categorization and the seemingly simpler synthetic stimuli tasks (rotated single T versus L, bisected disk versus its mirror image). Our data showed, however, a large discrepancy between the attentional requirements of these two types of tasks. It is difficult to explain this by the same training process.

Here we further investigated the effects of training (or lack of it), particularly its influence on the natural scene stimuli. If training indeed helped in performing natural scene categorization with little attention, this might be achieved through learning specific visual features critical for performing this task. It is important to note that all the data collected in the testing phase was from a set of novel images that the subjects were never trained on. Therefore simple image-based learning or memorization cannot account for the observed results. The testing stimuli, however, were drawn from the same set of images as the training images. Could subjects have, therefore, learned to categorize animal versus non-animal (or vehicle versus non-vehicle) images because the same image types were presented repetitively? In a computational framework, it is conceivable that a specific set of "animal filters" (or "vehicle filters") were sharply tuned and enhanced during this training period. But if this were the case, training on a specific task would only help to tune the specific "filters" for that particular categorization task. If we tested on a different natural scene categorization task, we should be able to observe a difference in performance.

3.2.1 Method

We tested this hypothesis in two separate experiments. In the first experiment, we divided a group of 6 subjects into 2 groups. Both groups received the same amount of training in all tasks. Specifically, both groups were trained on the central letter discrimination task (see Chapter 2.3.2) and a natural scene categorization task under both single- and dual-task conditions. Recall that the "single-task" condition refers to the situation where attention is available to perform the current peripheral task, while the "dual-task" condition refers to the situation where attention is drawn to the center, leaving the peripheral task in the near absence of attention. There are 96 trials in any given block of task under all conditions. In Group I, the 3 subjects were trained on the "animal vs. non-animal" categorization task. They were instructed to categorize natural scenes with or without animals in a go/no-go fashion by releasing a mouse button. The task was speeded so that any lack of response after 1000ms was automatically registered as a "no target" answer. The test data were then acquired by having them perform vehicle vs. non-vehicle categorization without any additional training. Similarly to the animal categorization task, subjects responded by releasing the mouse button when a target (vehicle(s)) was detected. In Group II, the 3 subjects were trained on vehicle categorization and tested on animal categorization (Fig. 3.4). Previous experiments [77] have already established that a trained natural scene categorization task requires little attention. We are, therefore, interested in seeing whether such performance can be transferred from one type of categorization (e.g., animal) to another (e.g., vehicle). Namely, will the performance of vehicle categorization without attention be comparable to the performance of animal categorization without attention for Group I subjects, and vice versa for Group II?

One might argue that in the above manipulation, even though the tested natural scene category was not trained, it was nevertheless learned during training because natural scene photographs shared many commonalities [95]. When one is trained on one type of natural scene categorization, say "animal scenes," it is possible that similar image statistics help to tune the "filters" on other types of natural scene categories (e.g., "vehicle scenes"). If this were the case, however, such performance should not hold for a recognition task that does not share similar stimulus statistics. For the second experiment in this section, we tested this hypothesis with another four subjects who were previously trained on the dual-task paradigm in a task that did not involve natural scene photographs. Specifically, these subjects performed a face gender discrimination task in a dual-task paradigm [114]. The gender discrimination task utilizes very different stimuli that bore little commonality with the natural scene images [141]. After the subjects completed their training on this task, we tested them on both animal and vehicle categorization tasks (Fig. 3.5).

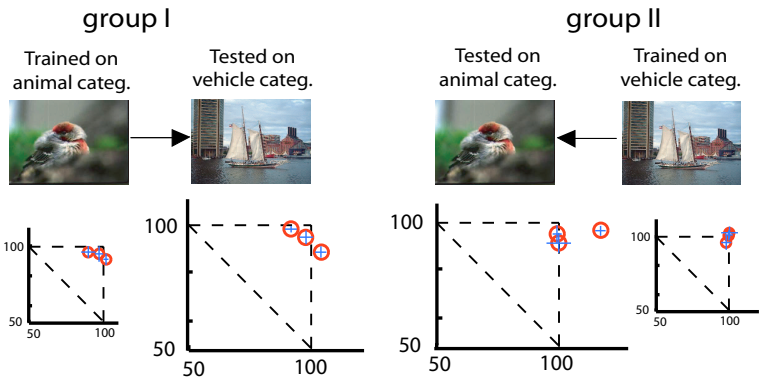


Figure 3.4: Cross-Training experiment. In this experiment 6 subjects are divided into 2 groups of 3 each. In Group I, illustrated on the left, the subjects are trained with central letter discrimination and animal categorization task, as well as the dual task condition using these two tasks. After training, they are immediately tested on the vehicle categorization task in both the single task condition (with attention) and the dual task condition (without attention). Results are shown in normalized performance plots. In Group II, illustrated on the right, the subjects are trained with central letter discrimination and vehicle categorization task, as well as the dual task condition using these two tasks. After training, they are immediately tested on the animal categorization task in both the single task condition (with attention) and the dual task condition (without attention). Results are shown in normalized performance plots. Our results illustrate that subjects need not be trained for the specific natural scene categorization task in order to perform it without attention, suggesting category-specific training is not necessary to carry out this high-level task without attention.

3.2.2 Results

The bottom panels of Fig. 3.4 illustrate the normalized performance results from the cross-training experiment between two groups of subjects. In Group I, three subjects were trained on animal categorization and then tested on vehicle categorization. Their central performances during the training phase show that under the dual task condition, they had successfully maintained their attention at the central task (single central task, average over 9 blocks: $75.9 \pm 3.5\%$, $85.4 \pm 3.9\%$ and $83.2 \pm 2.5\%$ for each subject, respectively; dual central task, average over 9 blocks: $70.3 \pm 5.5\%$, $86.11 \pm 2.7\%$ and $80.8 \pm 4.1\%$, respectively; t-test results: $t(16) < 1.75$, $p > 0.05$ for each subject). During the testing condition, only one subject has a slight drop in central task performance under the dual task condition (single central task, average over 9 blocks: $75.9 \pm 3.6\%$, $85.4 \pm 3.9\%$ and $83.2 \pm 2.5\%$ for each subject, respectively; dual central task, average over 9 blocks: $71.2 \pm 2.9\%$, $87.6 \pm 3.7\%$ and $81.3 \pm 4.1\%$; t-test results: $t(16) = 2.40$, $p = 0.01$ for the first subject; $t(16) < 1.75$, $p > 0.05$ for the rest). During the training phase, the subjects performed the animal categorization task without any interference when comparing the performances under the dual task condition with the single task condition (single peripheral task, average over 9 blocks: $79.5 \pm 0.6\%$, $84.3 \pm 0.6\%$ and $77.1 \pm 1.8\%$; dual peripheral task, average over 9 blocks: $77.3 \pm 5.2\%$, $78.1 \pm 5.6\%$ and $74.3 \pm 6.7\%$; t-test results: $t(16) < 1.75$, $p > 0.05$ for each subject). A similar performance pattern is observed for these three

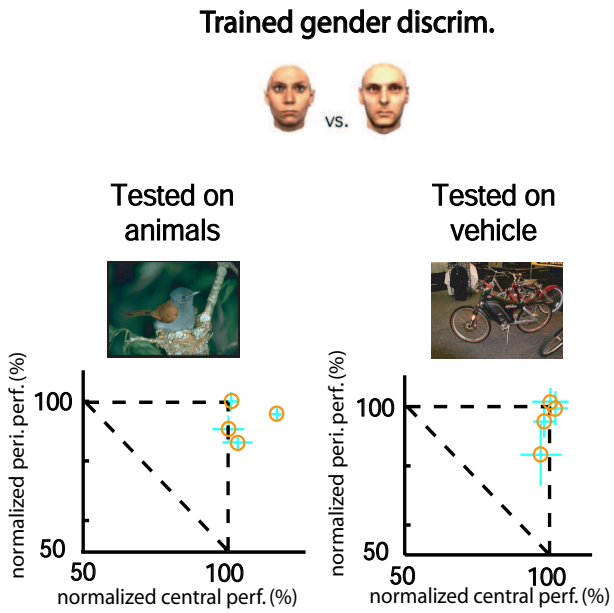


Figure 3.5: Gender-Train Dual-Task experiment. In this experiment, 4 subjects are trained on the central letter discrimination task, gender discrimination as well as the dual task condition using these two tasks. Two examples of the stimuli of the gender discrimination task are shown here [141]. Subjects are instructed to respond whether a briefly presented, then masked hair-less face peripherally is of a female or a male face [114]. After the training process is completed, subjects are tested on two natural scene categorizations without attention: animal and vehicle. Normalized performances of the single and dual task conditions are presented for each natural scene categorization. Our results show little training on natural scenes is needed to perform natural scene categorization.

subjects during the testing phase, in which they were put directly on the vehicle categorization task without any prior training (single peripheral task, average over 9 blocks: $83.2 \pm 4.5\%$, $87.3 \pm 4.4\%$ and $80.2 \pm 2.6\%$; dual peripheral task, average over 9 blocks: $82.2 \pm 2.1\%$, $79.3 \pm 3.9\%$ and $77.3 \pm 3.7\%$; t-test results: $t(16) = 3.22$, $p = 0.003$ for the second subject; $t(16) < 1.75$, $p > 0.05$ for the rest). Note that there is a slight drop in the vehicle categorization task under dual task condition for the second subject. This small decrease, although significant, should be viewed in the light of the results of [74] and [77]: when attention is taken away, performances for a simple rotated T versus L task (or Red-Green disk versus Green-Red disk) dropped much more dramatically, often to chance level (50%). Compared to Group I, Group II subjects went through reversed training and testing categorization tasks. During the training stage, three subjects were trained on vehicle categorization only. All of them have successfully allocated attention at the center under both the single task and dual task condition (single central task, average over 9 blocks: $68.6 \pm 4.6\%$, $78.5 \pm 6.1\%$, and $88.9 \pm 3.2\%$; dual central task, average over 9 blocks: $67.9 \pm 2.8\%$, $78.6 \pm 7.2\%$, and $88.8 \pm 3.5\%$; t-test results: $t(16) < 1.75$, $p > 0.05$ for each subject). Their vehicle categorization task results also show that they were able to perform this task without attention (single peripheral task, average over 9

blocks: $72.2 \pm 7.9\%$, $77.4 \pm 1.2\%$ and $81.6 \pm 1.6\%$; dual peripheral task, average over 9 blocks: $70.0 \pm 4.8\%$, $78.3 \pm 5.0\%$ and $81.4 \pm 4.7\%$; t-test results: $t(16) < 1.75$, $p > 0.05$ for each subject). During the testing stage, where subjects were tested directly on animal categorization without any prior training (training was done using vehicle categorization), they maintained good performances on central task under both conditions just as they did during the training sessions (single central task, average over 9 blocks: $68.6 \pm 4.6\%$, $78.5 \pm 6.1\%$ and $88.9 \pm 3.2\%$; dual central task, average over 9 blocks: $68.6 \pm 4.2\%$, $88.3 \pm 2.7\%$ and $88.4 \pm 3.4\%$; t-test results: $t(16) = 3.79$, $p = 0.001$ for the second subject; $t(16) < 1.75$, $p > 0.05$ for the rest). Similarly, all three subjects performed the animal categorization task under the dual task condition as well as under the single task condition (single peripheral task, average over 9 blocks: $74.8 \pm 5.8\%$, $84.2 \pm 1.8\%$ and $81.1 \pm 4.3\%$; dual peripheral task, average over 9 blocks: $70.4 \pm 4.5\%$, $81.5 \pm 3.0\%$ and $77.8 \pm 5.1\%$; t-test results: $t(16) < 1.75$, $p > 0.05$ for each subject).

Fig. 3 illustrates the results from the second experiment in this section. Four new subjects were trained on the dual task paradigm with the same central letter discrimination task but a peripheral face gender discrimination task [114]. Testing of whether animal categorization and vehicle categorization require attention followed after the training phase. In all but one case, all of the four subjects showed natural scene categorization performances (both animal and vehicle) in the near absence of attention statistically indistinguishable from the same tasks performed with attention available (single peripheral animal categorization task performance, averaged over 9 blocks for each: $83.0 \pm 3.2\%$, $81.9 \pm 0.6\%$, $83.3 \pm 3.4\%$ and $76.9 \pm 4.3\%$; dual peripheral animal categorization task performance, averaged over 9 blocks: $79.9 \pm 4.3\%$, $77.4 \pm 3.7\%$, $74.3 \pm 3.3\%$ and $77.1 \pm 3.4\%$, t-test results: $t(16) = 4.46$, $p = 0.0002$ for the third subject; $t(16) < 1.75$, $p > 0.05$ for the rest. Single peripheral vehicle categorization task performance, averaged over 9 blocks for each: $81.8 \pm 3.9\%$, $83.0 \pm 4.5\%$, $85.1 \pm 2.6\%$ and $78.1 \pm 7.3\%$; dual peripheral vehicle categorization task performance, averaged over 9 blocks: $79.6 \pm 2.5\%$, $80.0 \pm 2.7\%$, $73.6 \pm 14.8\%$ and $77.8 \pm 6.4\%$, t-test results: $t(16) < 1.75$, $p > 0.05$ for each subject). Note that one subject showed a small decrease in animal categorization performance when attention was drawn away (single task performance, average over 9 blocks: $83.3 \pm 3.4\%$; dual task performance, average over 9 blocks: $74.3 \pm 3.3\%$; t-test results: $t(16) = 4.46$, $p = 0.0002$). Here again, this slight decrease is much smaller than those observed on known "attentionally-demanding" tasks [74, 77]. Note that the same subject's performance on vehicle categorization was not significantly different in the single and dual task conditions.

Both experiments in this section demonstrate that little training is needed for natural scene categorization without attention. Subjects are able to perform natural scene categorization tasks in the near absence of attention without previous training on the specific task or type of stimuli.

3.2.3 Discussion

The first experiment in this section has shown that when trained on animal categorization (or vehicle) in the near absence of attention, subjects can perform vehicle categorization (or animal) with no further training. This effect is even more dramatically demonstrated by the second experiment, in which a totally different type of stimulus is used during training. But as soon as subjects have learned to perform this recognition task under the dual task paradigm, they can apply this ability to a natural scene categorization task. On the contrary, our previous experiments have shown that single letter discrimination or color disk discrimination cannot be performed without attention given the same amount of training as the natural scene categorization task [77]. Hochstein and colleagues argue that "higher-level" tasks can be more easily transferred than "lower-level" tasks [59]. Could it be that less attentional resource is needed because natural scene categorization is carried out in "higher" areas of the visual system than the other synthetic tasks? We will revisit this point in both Control Exp 3 and Control Exp 5.

3.3 Control Experiment 2: Effect of Color

We set out to explore different factors that might contribute to the fast recognition of natural scene categories with little or no attention. A simple question to ask is whether some low-level features might have been useful cues to the categorization task. For example, it has been shown that color histograms are very informative for natural scene recognition in both human and computer vision [93, 97, 111]. In today's computer vision field, some image retrieval algorithms have utilized color information to categorize different images [123]. Delorme and colleagues have shown that color is not a critical component in fast categorization, under conditions where attention was not explicitly controlled [22]. In the present experiment, we question the role of color information in natural scene categorization with little or no attention, by changing the stimuli to grayscale.

3.3.1 Method

We use the same dual task paradigm for this experiment as in the previous experiment. The central task is an attentionally demanding letter discrimination. The peripheral task is natural scene categorization, using novel grayscale images (examples of the grayscale images are shown in Fig. 4). Five subjects participated in this experiment. They were instructed to respond as fast as possible when they detected the presence of an animal in the image shown at a random position peripherally. Subjects performed 15 blocks of dual task and 12 blocks of single task. Each block consisted of 96 trials.

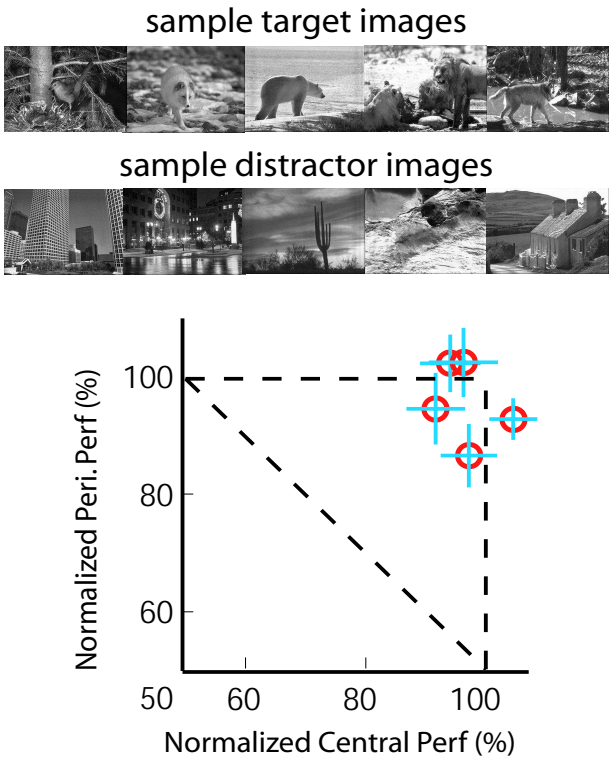


Figure 3.6: Natural scene categorization without color. In this experiment, the peripheral task is animal categorization with grayscale natural scene images. All experimental conditions remain identical to the one introduced in [77]. The only difference is that all peripheral task stimuli as well as the masks are presented in grayscale. The top two rows show some samples of the target stimuli and distractor stimuli, respectively. The bottom panel indicates subjects' normalized performances of this task, showing that there is little cost in grayscale natural scene categorization when attention is withdrawn elsewhere.

3.3.2 Results

It is first interesting to observe that individual SOAs for the natural scene categorization task are not much different from the ones observed in [77] where color information was included in the images (average natural scene categorization SOA of 5 subjects in [77]: 61msec; average SOA of 5 subjects in current experiment: 85msec; t-test result: $t(8) < 1.86, p > 0.05$). Fig. 3.6 illustrates the normalized performances of the subjects' dual task performances. Note that each subject has achieved a central task performance at his/her baseline level in dual task (single central task, average over 12 blocks for each subject: $76.9 \pm 6.1\%$, $76.0 \pm 3.2\%$, $74.2 \pm 4.9\%$, $75.0 \pm 4.4\%$ and $74.3 \pm 6.6\%$; dual central task, average over 15 blocks for each subject: $75.4 \pm 5.7\%$, $78.4 \pm 4.7\%$, $71.3 \pm 5.5\%$, $70.8 \pm 5.5\%$ and $72.5 \pm 6.3\%$; t-test results: $t(25) < 1.71$, $p > 0.05$ for each subject). This result assures that much of attentional resource is rightfully allocated for the demanding letter discrimination task under the dual-task condition. Four of the five subjects' peripheral natural scene categorization task performances remain comparable to their respective baseline performances

(single peripheral task, average over 12 blocks for each subject: $79.3 \pm 2.7\%$, $73.2 \pm 7.1\%$, $76.9 \pm 6.5\%$ and $74.2 \pm 7.4\%$; dual peripheral task, average over 15 blocks for each subject: $75.2 \pm 4.7\%$, $74.4 \pm 5.1\%$, $71.7 \pm 5.9\%$ and $72.9 \pm 5.8\%$; t-test results: $t(25) < 1.71$, $p > 0.05$ for each subject). Only one subject's natural scene categorization task performance decreases slightly while attention is allocated elsewhere (single peripheral task, average over 12 blocks: $71.9 \pm 5.6\%$; dual peripheral task, average over 15 blocks: $66.1 \pm 5.3\%$; t-test results: $t(25) = 3.82$, $p = 0.0004$). Overall, grayscale natural scene images can be categorized rather efficiently in the near absence of attention.

3.3.3 Discussion

Our results indicate that when much of their attention is locked elsewhere, subjects can perform rapid natural scene categorization task without using color information. This result suggests that color information is not critical in performing such a task in the near absence of attention. Dunai and colleagues have also found that color cues are only more informative at a longer time scale to an attentionally demanding detection task [26]. In addition, Delorme and colleagues have already shown that color information is not critical in the initial recognition of the same set of natural scenes that we are using [22]. These findings, together with ours, suggest that natural scene categorization might be carried out by a rapid and efficient process that does not require much of the slower color information. Torralba and colleagues suggested that some natural scene images can be categorized based on second order statistics derived from power spectral analysis [138]. It would be fruitful to test their hypothesis on object recognition such as animal or vehicle categorizations. To conclude, removal of color information failed to make natural scene categorization an "attentionally demanding" task. Thus, if this type of natural scene processing only relies on low-level features, color cannot be counted as one such feature.

3.4 Control Experiment 3: Evidence for Parallel Processing

We have established so far the amazing robustness of the human visual system when categorizing natural scenes with little attention. In an attempt to search for a "breaking point" of this ability, we investigate the effects of natural scene categorization when the number of peripheral images is increased to two. In other words, instead of searching for a possible target in just one image, the subjects have to now search for a possible single occurrence of the target in two images. We ask whether by effectively halving the "signal to noise ratio" would the efficiency of this task decrease? Our rationale is that by identifying the condition in which such natural scene processing is no longer doable, we can start comparing and contrasting different

conditions in order to understand the underlying neuronal mechanisms.

3.4.1 Method

The attentionally demanding central letter discrimination task remained the same as in the previous experiments. Five Subjects performed a letter discrimination task at the central SOAs individually adjusted for each of them. The peripheral task was a colored animal scene categorization. Each block consisted of 48 trials. In half of the trials, there were two peripheral natural scene images ("double-image" condition), with two equally likely configurations-either one of the two images contained a scene with animal(s); or neither image contained an animal. In the other half of the trials, there was only one image, just like in the previous experiments on scene categorization ("single-image" condition). Subjects are told to respond by lifting the mouse button when they detect the presence of an animal (or animals) in both conditions. These two types of trials were intermixed randomly throughout the experiment. Fig. 3.7 shows the schematic setup of this experiment. For the double-image condition, the separation between the two images varied randomly between 4 and 12 degrees (each maintaining an eccentricity of 6 degrees, just as in the previous experiments). Subjects were informed before the experiments of the two different possible conditions. No one reported any confusion or difficulty with the instructions. There was a total of 15 blocks for the dual task condition and 15 blocks for each single task condition.

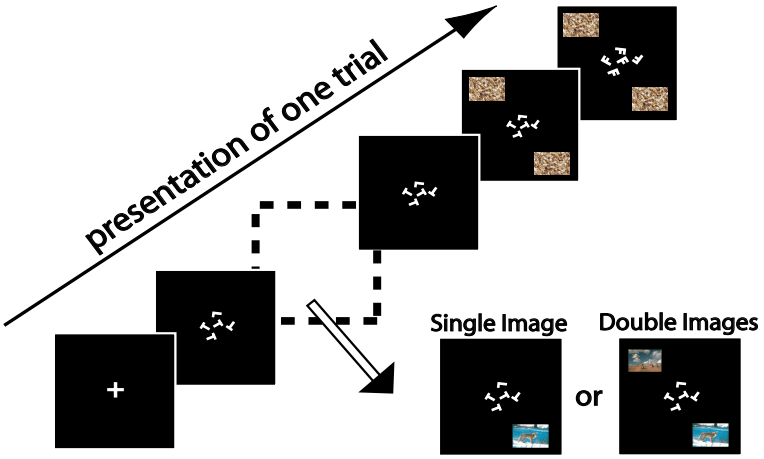


Figure 3.7: Experimental setup for single-image versus double-image experiment. We illustrate here the setup of a single trial for this experiment. The basic procedure is the same as in Fig. 2.1. For a given trial, there are two possible peripheral stimulus presentation setups. For 50% of the trials, there are two unrelated natural scene images presented randomly in the periphery. Both images are of 6-degree eccentricity with respect to the center of the screen. Their mutual separation varies randomly. Two similar perceptual masks follow the images after the peripheral SOA amount of time. For the other 50% of the trials, there is only one natural scene image presented randomly in the periphery, which is exactly the same case as in [77]. At the end of the presentation, a random perceptual mask follows the image stimulus.

3.4.2 Results

All 5 subjects were able to categorize novel natural scenes without attention. The top two panels of Fig. 3.8 illustrate the normalized dual task performances for the single-image condition and double-image condition, respectively. First we observe that all subjects maintained their central letter discrimination task performances under single task and dual task conditions (single central task performance, average over 15 blocks for each subject: $75.4 \pm 3.3\%$, $71.5 \pm 1.3\%$, $67.7 \pm 5.0\%$, $77.6 \pm 4.9\%$ and $77.1 \pm 3.8\%$; dual central task performance, average over 15 blocks for each subject: $73.8 \pm 5.4\%$, $71.5 \pm 5.5\%$, $67.3 \pm 4.3\%$, $75.0 \pm 6.0\%$ and $77.3 \pm 5.9\%$; t-test results: $t(28) < 1.70$, $p > 0.05$ for each subject). These results indicate that attention was successfully locked at the central task for all these subjects. Now we are interested in comparing subjects' natural scene categorization performances with or without attention under single-image and double-image conditions. Fig. 3.8 shows the performance pattern of double-image categorization (single task performance, average over 15 blocks for each subject: $64.6 \pm 3.1\%$, $71.7 \pm 5.8\%$, $74.7 \pm 7.6\%$, $75.4 \pm 6.1\%$ and $73.6 \pm 8.4\%$; dual task performance, average over 15 blocks for each subject: $64.0 \pm 6.4\%$, $68.1 \pm 6.5\%$, $70.2 \pm 5.8\%$, $72.3 \pm 6.3\%$ and $70.7 \pm 6.0\%$; t-test results: $t(28) < 1.70$, $p > 0.05$ for each subject) as well as the single-image case (single task performance, average over 15 blocks for each subject: $77.1 \pm 5.4\%$, $80.1 \pm 7.1\%$, $87.5 \pm 6.9\%$, $81.9 \pm 4.1\%$ and $79.2 \pm 11.0\%$; dual task performance, average over 15 blocks for each subject: $76.3 \pm 4.9\%$, $74.0 \pm 4.3\%$, $79.0 \pm 8.3\%$, $83.8 \pm 5.9\%$ and $70.4 \pm 9.0\%$; t-test results: $t(28) = 2.48$, $p = 0.01$ for the second subject, $t(28) = 1.98$, $p = 0.03$ for the third subject, $t(28) < 1.70$, $p > 0.05$ for the rest). All subjects' results show that when there are two images to process, the categorization performances with attention (single-task condition) are statistically no different from the performances without attention (dual-task condition). There is, however, a small but significant drop for one subject when categorizing the single-image without attention compared to with attention. Similarly to the previous arguments, we think this is a rather small effect in the light of the comparative results obtained from synthetic stimuli [74,77]. The average baseline performances (i.e., single task condition when attention is available) show an overall decrease in the double-image categorization condition (single task condition for single-image case: $81.2 \pm 3.9\%$; single task condition for double-image case: $72.0 \pm 4.4\%$; t-test result: $t(8) = 3.04$, $p = 0.008$). This small set size effect, which appears to contradict previous results by [120], might simply be attributed to our keeping SOAs constant between the single- and double-image conditions. In addition, stimulus location was totally unpredictable in our study, whereas it was fixed in the experiments of [120]. But the main result is that, when attention is taken away, subjects were able to perform double-image categorization just as well as they did when attention was available. This result suggests that by halving the "signal-to-noise ratio" of the stimuli, natural scene categorization can still be efficiently carried out in the near absence of attention.

Since we have randomly varied the visual angle distance between the two images under the double-image condition, we can ask whether the subjects performances differ for different visual angle separations. The bottom two panels of Fig. 3.8 show double-image condition performances sorted by the visual angle separation. The left panel corresponds to the condition where attention is available (single task condition), whereas the right panel corresponds to no-attention condition (dual task condition). The two leftmost bars in each panel, placed at the 0 degree angles, indicate the average single-image performances of the subjects with or without attention, respectively. We investigated the effects of visual angle separation between the two stimulus images with a 2-way ANOVA (attentional condition vs. inter-stimulus separation). In accordance with our previous results, there was no main effect of attention on performances (single task condition vs. dual task condition: $F(1, 96) = 1.2112$, $p > 0.05$). There was no main effect of the visual angle separations between stimuli either ($F(11, 96) = 0.5817$, $p > 0.05$). Additionally, there was no significant interaction between these two factors ($F(11, 96) = 1.1654$, $p > 0.05$). Note that due to the size of the image itself, the minimal separation distance between two images is 4 degrees.

3.4.3 Discussion

Our finding suggests that natural scene categorization task not only demands little attentional resource, but is also highly parallel. When two images are presented simultaneously, subjects are able to process both of the images in search of a target object in a comparable fashion as when there is only one image. Physiological data from ERP recording also supports this finding. Rousselet et al. found that subjects are as fast for animal categorization with two images as with one image [120]. Together our results suggest that high-level information can be accessed by the visual system in a parallel fashion with little attentional assistance. It suggests that some categorical information might be able to reach higher areas of the visual hierarchy rather efficiently, without much serial focal-attention selection.

3.5 Control Experiment 4: Multiple Copies of the Synthetic Stimuli

We have so far probed in a number of ways to what extent natural scene categorization can be carried out by the human visual system without attention. Our results tell us that such categorization is highly efficient and robust to the lack of attentional resource. By contrast, seemingly much simpler tasks involving synthetic stimuli do not enjoy this freedom of attention [18, 77]. In the following two experiments, we turn to the question of these synthetic stimuli: what type of manipulation would decrease attentional requirements for these stimuli? In other words, through which dimension of manipulation can we make the synthetic stimuli

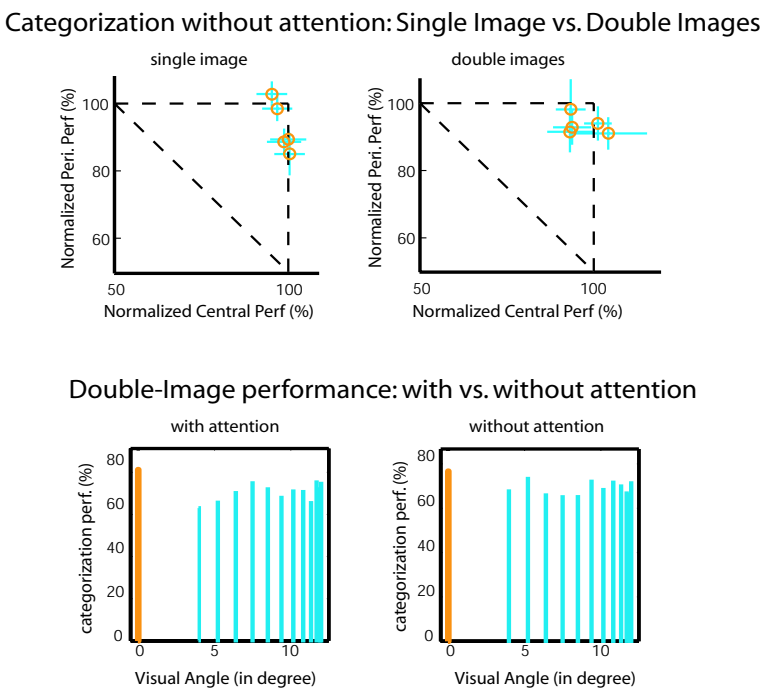


Figure 3.8: Results of the single-image versus double-image experiment. The top two panels illustrate normalized performances of the single-image dual task and the double-image dual task, respectively. Five subjects participated in this experiment. The results show that there is little difference between the single-image case and the double-image case, suggesting that natural scene categorization without attention is a highly parallel process. The bottom two panels break down the performances of the double-image case by the angle of separation between the two images. The two leftmost bars in each panel, placed at the 0 degree angles, indicate the average single-image performances of the subjects with or without attention, respectively. The left panel shows the result when attention is available; whereas the right panel shows when attention is withdrawn. For each attentional condition, there is no apparent pattern of performance difference as a function of visual angle separation.

task "easier" without attention?

One possible hypothesis of this contrast between natural scene images and synthetic stimuli is that an object target (e.g., an elephant) in a natural image might carry multiple "diagnostic features" for its detection or recognition. The exact nature of these "diagnostic features" is unknown. But it is conceivable that many of the body parts of an animal in the animal categorization task, for example, are potential cues for the detection of the object. The synthetic stimuli, on the other hand, do not enjoy the luxury of multiple potential "diagnostic features" [77]. In that study, there were two types of synthetic stimuli, a rotated T versus a rotated L, and a Red-Green bisected disk versus a Green-Red bisected disk. In the case of T versus L, the only obvious "diagnostic feature" is the T-junction versus the L-junction in the letters, respectively. A failure to detect such junction would result into an ambiguous decision of the stimulus. Similarly, for the bisected disks, since the stimuli position is random from trial to trial, it is not possible to determine whether it is a Red-Green disk

or a Green-Red disk by detecting the color on half of the disk. The only "diagnostic feature" is the junction between the two colored semicircles. Hence we predict that if the advantage of scene categorization without attention lies in the higher probability of detecting one or more of the possible "diagnostic features," then increasing the number of stimuli in the synthetic stimulus task could result in an increase of performance.

3.5.1 Method

We test this hypothesis using the bisected color disks. Four subjects participated in this experiment. The basic setup remained the same. The attentionally demanding central task was letter discrimination. In the periphery, subjects were instructed to respond when the bisected disk(s) was (were) arranged in a red-green fashion, as opposed to an equally likely green-red pattern. In half of the trials, there were four identical disks. Subjects were assured of the fact that all stimuli were redundant. In the other half of the trials, there was only one such bisected disk (which is the same condition as in [77]). These two types of trials were intermixed randomly within a block of trials. All disks were the same size. Fig. 3.9 illustrates the arrangement of the stimuli. In the single-disk condition, the disk was centered at 6-degree eccentricity. In the four-disk condition, the center of the 4-disk array was located at 6-degree eccentricity. Each block consisted of 96 trials. Subjects performed 18 blocks of experiments.

3.5.2 Results

The hypothesis described above predicts that an increased number of peripheral stimuli might result in an increase of dual-task performance for recognition of the color disks. The intuition is that there are more potential "diagnostic features" to be sampled by the visual system when the number of redundant stimuli is greater. Contrary to our prediction, we observe no improvement in dual-task performances for the trials where there are four disks rather than one. For both the single-disk and the four-disk conditions, subjects' dual task performances of the peripheral color disk recognition are not significantly better than chance (one-disk performance under dual-task condition, averaged over 18 blocks for each subject: $50.4 \pm 7.63\%$, $49.3 \pm 6.6\%$, $48.9 \pm 8.0\%$, and $52.4 \pm 6.2\%$; four-disk performance under dual-task condition, averaged over 18 blocks for each subject: $47.9 \pm 8.1\%$, $49.7 \pm 6.1\%$, $50.5 \pm 3.4\%$ and $46.9 \pm 7.4\%$; t-test results: $t(17) < 1.74$, $p > 0.05$ for each subject and each task). Note that when the peripheral color disk recognition task is carried out with attention available, subjects' performances center around 85% at their individual SOAs (one-disk performance under single task condition, average over 18 blocks for each subject: $88.8 \pm 8.5\%$, $81.3 \pm 5.0\%$, $89.6 \pm 3.3\%$ and $77.0 \pm 7.5\%$; four-disk performance, averaged over 18 blocks: $91.7 \pm 6.3\%$, $87.1 \pm 5.1\%$, $90.5 \pm 5.4\%$ and $74.5 \pm 7.5\%$). It seems that the subjects cannot take advantage of the increased number of

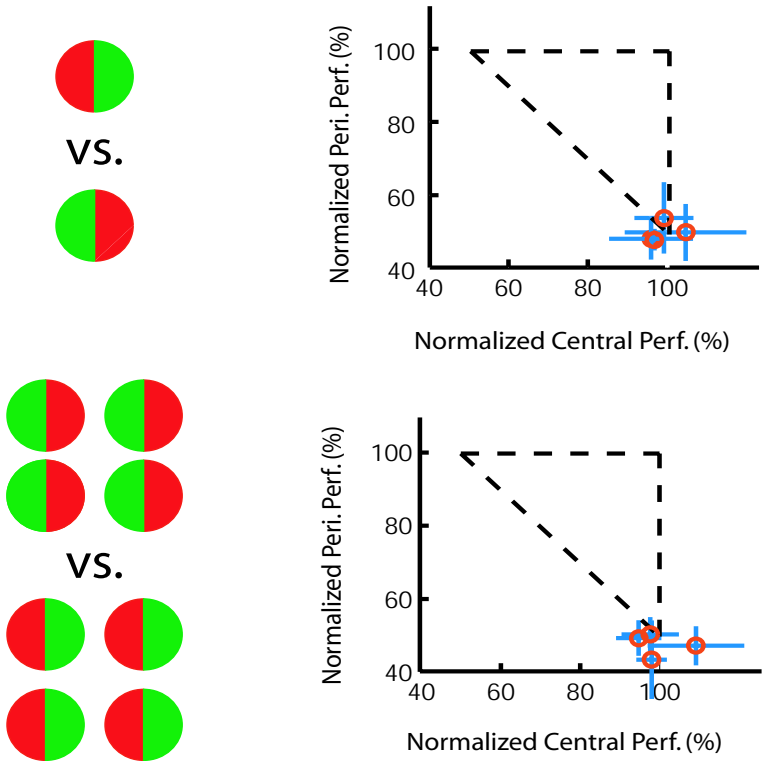


Figure 3.9: Multiple stimuli experiment. The basic setup of this experiment is also Dual-Task paradigm. A color disk discrimination task is used peripherally. Two types of peripheral stimuli are mixed randomly during the experiment. The top row shows the first peripheral stimulus and the normalized dual task performances. Peripheral recognition task is a red-green color disk versus its mirror image, green-red color disk. Subjects' performances of this task without attention is at chance level compared to their baseline performances, centered around 80% before normalization. The bottom row shows the second type of peripheral stimulus and the corresponding normalized dual task performances. In this case, the recognition task remains the same as the top row, with the exception that there are four copies of the same stimuli arranged in the indicated pattern. We show here that subjects' performances of this task without attention is also at chance level, no better than the case with one copy of the stimulus.

possible "diagnostic features," even when attention is fully available.

3.5.3 Discussion

We test the hypothesis that independent but "diagnostic features" might contribute greatly in recognition without attention. The assumption was natural scene categorization might be potentially "easier" than the synthetic stimuli recognition due to the multitude of "diagnostic features." In other words, different body parts of an animal (or vehicle) might increase the chance of detection while the synthetic stimuli tend to have a very localized, nearly singular point of "diagnostic feature" (e.g., bisecting line of the double color disks). We therefore increased the probable number of features by replicating the number of stimuli from 1 to 4. It is important to point out that this manipulation is not comparable to the one that we did in Exp 3. In

Exp 3, we left the amount of "target signal" (or probability of the presence of an "animal" scene) constant while doubling the amount of "distracting noise" (or probability of the presence of a "non-animal" scene). Here the absolute number of potential "diagnostic features" is increased through having multiple, redundant copies of stimuli. Our results show clearly that such increase of potential diagnostic features did not help at all in recognition of synthetic stimuli without attention. This observation implies that it is unlikely that the bottleneck of such synthetic stimuli recognition without attention is the number of available "diagnostic features." Natural scenes might have an overall advantage over the synthetic stimuli used here due to the intrinsic image statistics or different processing mechanisms.

An alternative explanation also deserves further investigation. It is true that we have replicated the target four times in each trial. But if one imagines that features related to the targets and distractors for the stimuli live in a high dimensional "feature space." Then it is possible that "diagnostic features" in the synthetic stimuli case might lie too closely to the "distractor features" of the synthetic stimuli in the "feature space". On the other hand, in the rich natural scene stimuli case, the "diagnostic features" of the targets might be much more easily isolated from the distractors than the synthetic stimuli case. If this hypothesis were true, simply repeating the number of targets in the synthetic stimuli task would not increase the discriminability of the target from the distractor, just as what we have observed here.

3.6 Control Experiment 5: Evidence for Well-learned Categories of Objects Entailing Less Attentional Load During Recognition

So far our attempt to "increase" the difficulty of natural scene recognition without attention by reducing the amount of signal or decreasing the amount of training has not broken down the system dramatically. Similarly, adding copies of stimuli to the synthetic recognition task does not "ease" the task difficulty either. Hence, we want to test whether task "predictability" can be an influential factor in the recognition task without attention.

Our observations, however, also point to the direction that it could be the different levels of processing that result in such different performances between natural scenes and synthetic stimuli. It has been long known that object categories are encoded in higher level visual areas such as the inferior temporal lobe (IT) [78, 133]. The most prominent object category is face for the human visual system [27, 118]. Haxby et al. have shown differentiable fMRI patterns in IT and related areas when responding to different types of stimuli of a wide range of visual categories [54]. So could it be that existing neuronal representations of natural scene categories are responsible for such efficient and fast recognition of natural images with little attention? If this is the case, can we find meaningful categories of objects in synthetic stimuli to test this

hypothesis?

We test here in this experiment two independent hypotheses by using the letter discrimination task as the peripheral task. The first hypothesis is that stimulus predictability might affect the attentional requirement in recognition. It is conceivable that less attention is required when subjects know beforehand the exact shapes of the stimuli to be discriminated. The second hypothesis is that well learned object categories can be recognized with significantly less attentional load. Evidence from the visual search paradigm using familiar and unfamiliar letter-like patterns indicates that visual search speed is strongly facilitated by more familiar objects [127, 151]. Peripheral letter discrimination task in dual task paradigm has previously been set up in such a way that the single peripheral letter stimulus is randomly rotated on each trial [16, 77]. Though these letters can be considered as well-learned categories of objects, letter recognition is better trained for upright letters for obvious reasons (try reading this page upside down). Hence we might observe some performance difference under the dual task condition between upright letter discrimination and the original, rotated letter discrimination.

3.6.1 Method

We use the Dual-Task paradigm to test these hypotheses. As usual, we use the central letter discrimination task as the attention-demanding central task. Three conditions are tested for the peripheral letter discrimination task: randomly rotated letter, fixed rotation and upright positions. For all conditions, the letter discrimination task is between T and L (L is the target in a go/no-go setup where subjects have to release a mouse button when the target is detected). The letter and its mask are located at a random position at 6-degree eccentricity. For a given block of 96 trials, one of the three different tasks is run, and subjects are informed beforehand about the block task. There are 10 blocks tested for each of the three conditions. Fig. 3.10 depicts the three different conditions. Four subjects participated in this experiment. A short training period of 3-4 hours proceeded the actual testing. During this period only the randomly rotated letter discrimination was trained concurrently with the central discrimination task. All three conditions were presented for an equal amount of time in the subsequent real data collection. For each subject, the peripheral letter task SOA was determined based solely on their single task performance on the randomly rotated letter discrimination task.

3.6.2 Results

When attention is available, the single task performances for all three different letter discrimination tasks are highly comparable (randomly rotated, average over 4 subjects and 10 blocks: $77.4 \pm 6.6\%$; fixed rotation, average over 4 subjects and 10 blocks: $78.8 \pm 7.0\%$; upright, average over 4 subjects and 10 blocks: $84.9 \pm$

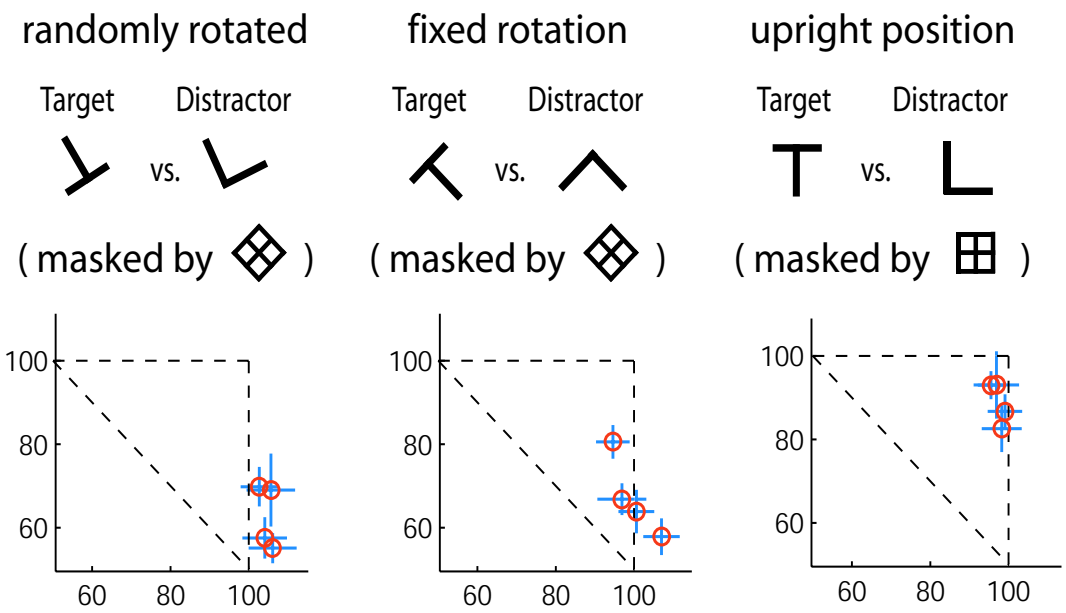


Figure 3.10: Rotated versus fixed rotation versus upright letter experiments. The basic setup of this experiment is also Dual-Task paradigm. A letter discrimination task is used peripherally. Three different peripheral conditions and their corresponding performances are shown in three columns. In the first column, the peripheral task is a randomly rotated T versus randomly rotated L, masked by a perceptual mask. Subjects perform barely at or above chance when attention is not available. This result confirms [16, 77]. The second column shows the peripheral task of a fixed rotation of T versus L, indicated in the figure. Subjects' performances of this task without attention are slightly better than the randomly rotated letter condition. In the last column, we show results of subjects' performances for the peripheral task in which the letter T and L are in their upright position, a configuration that is familiar and well-learned for all of our subjects. The normalized performances indicate a clear advantage of this condition, showing a near-baseline performance when attention is withdrawn. This result suggests that tasks that are meaningful and well-learned can be carried out by the visual system in a much more efficient manner.

4.8%; pair-wise t-test shows no statistically significant difference between each pair of the three different tasks ($t(9) < 1.83$, $p > 0.05$ for all cases)). This suggests when there is abundant attentional resource, carrying out the letter discrimination task for all rotation conditions is similar. It is important to point out that the fixed rotation and upright conditions have similar performances as the randomly rotated condition. This indicates that the SOA that was determined for each subject for the peripheral task was effective for all conditions. We observed, however, different performance results under the dual-task paradigm. For the randomly rotated letter task, subjects' performances were congruent with what is reported in previous studies [16, 77]. Two subjects' performances were not significantly different from chance (randomly rotated letter task performance under single-task condition, average over 10 blocks for each of the 2 subjects: $74.2 \pm 3.3\%$ and $85.9 \pm 4.3\%$; randomly rotated letter task performance under dual-task condition, average over 10 blocks for each of the 2 subjects: $53.7 \pm 4.8\%$ and $53.7 \pm 5.2\%$; t-test results comparing the dual-task condition with chance: $t(9) < 1.83$, $p > 0.05$ for each subject). The other two subjects performed slightly better

than chance, but significantly lower than their baseline performances obtained when attention was available (randomly rotated letter task performance under single-task condition, average over 10 blocks for each of the 2 subjects: $70.6 \pm 4.4\%$ and $78.9 \pm 4.1\%$; randomly rotated letter task performance under dual task condition, average over 10 blocks for each of the 2 subjects: $57.8 \pm 7.2\%$ and $61.5 \pm 5.5\%$; t-test results comparing dual-task condition performances with single-task condition performances: $t(18) > 3.61, p < 0.001$ for each subject). For the fixed but uncommon rotation (non-upright), subjects' performances varied more than the random rotation condition. Three subjects' dual-task performances were comparable to chance level or just slightly higher (fix-rotation letter task performance under dual task condition, average over 10 blocks for each of the 3 subjects: $56.0 \pm 4.5\%$, $56.0 \pm 6.7\%$ and $59.1 \pm 4.2\%$, t-test results: $t(9) < 1.83$, $p > 0.05$ for the first two subject; $t(9) > 4.30$, $p < 0.01$ for the third subject). One subject performed at 81% of his baseline level performance (single task condition: $78.1 \pm 8.1\%$; dual task condition: $67.2 \pm 4.6\%$). We should then compare this result to the last condition: upright letter discrimination. Here all subjects performed above 80% of their baseline performance level (single task condition performance, average over 10 block for each subject: $85.2 \pm 4.4\%$, $87.2 \pm 5.7\%$, $78.1 \pm 8.1\%$ and $89.1 \pm 4.3\%$; dual task condition performance, average over 10 blocks for each subject: $72.9 \pm 7.9\%$, $77.3 \pm 6.1\%$, $74.2 \pm 9.1\%$ and $83.6 \pm 5.3\%$). One subject's performance was not significantly different from her performance when attention was available (single-task condition: $78.1 \pm 8.1\%$; dual-task condition: $74.2 \pm 9.1\%$; t-test result: $t(18) < 1.73$, $p > 0.05$). There is thus a clear pattern that the least amount of attention resource is needed when letters are in their upright positions.

3.6.3 Discussion

Two independent hypotheses were tested in this set of experiments. We found that subjects could not recognize rotated letters without attention even when the rotation is fixed to one angle throughout the entire testing period of several hours. This result indicates that stimulus predictability alone cannot reduce much of the attentional load required for such a task.

On the other hand, while keeping everything else exactly the same, recognizing upright letters shows a clear advantage over randomly rotated letters when attentional resource is scarce. One might argue that a fixed upright rotation is easier for a mental template matching algorithm because the stimuli are much more predictable whereas such a method is less useful for a random rotation. This is why we observe a significantly improved performance for the upright letter condition compared to the random letter condition. But this argument does not support our observation for the fixed but uncommon rotation condition. If the template-matching theory works, it should work for any fixed rotation, not just upright. In fact a modern

computer vision algorithm can easily implement such a template-matching method to carry out this task. Our results suggest that there is a clear bias for people to recognize upright letters better than other rotated versions, even highly predictable ones. This is not surprising if we consider how much more learning one receives on upright letter recognition over his or her lifetime. Similar evidence has also been found in the visual search tasks. Wang et al. have reported that familiar letters take less time to search than unfamiliar ones [151]. This observation supports our hypothesis that familiar and meaningful categories of objects can be recognized with much less attentional load.

Chapter 4

Summary

4.1 Natural Scene Categorization Requires Little Attention

Natural scene categorization is one of the most evolutionarily relevant tasks of the human visual systems. The superb efficiency of this task has stimulated much research across the fields of neural psychophysics, physiology and modeling [8,77,109,135,138]. Contrary to the daily experience of the effortlessness of natural scene recognition, it is one of the hardest tasks that the modern state-of-the-art computer vision algorithms have yet to accomplish. This difficulty is mostly due to the vast variability across similar categories of the natural scenes. Unlike low-level tasks such as orientation discrimination or texture recognition, in which much of the tasks can be accomplished through filtering of the primary visual cortex [83], understanding the categorical information across different examples of natural scenes is usually thought to be a high-level visual task. In an effort to understand the processing of natural scene categorization, we have chosen an approach to study the efficiency of this process in the near absence of attention.

Visual attention is considered to be one of the first and foremost means of controlling the flow of information between different levels of visual processing. Numerous studies have probed the function of attention, demonstrating much attentional control over stimuli with complex and conjugate features [140, 157]. Needless to say, the function of attention is tightly associated with the computational function of recognition in the human visual system. We hope that by manipulating the attentional condition of various natural scene categorization tasks, as well as comparing it with other recognition tasks, much light can be shed in the understanding of the fundamental mechanisms that enable us such a rapid and fast ability of scene categorization.

Our findings show that rapid visual categorization of novel natural scenes requires very little or no focal attention. Perception outside the focus of attention has mostly been reported for simple salient stimuli [17, 140]. In our task, however, human subjects are actively searching for a complex category of objects whose appearance is highly variable. It thus appears that a sophisticated high level of representation (e.g., semantic)

can be accessed outside the focus of attention. It has already been argued that the "gist" of a visual scene could be available preattentively [8, 157]. In this context, the contents of the "gist" could in fact be extended to include information about the presence of a complex target category whose appearance is not known in advance.

These results suggest that if attention gates visual information processing at early stages of the visual system, such as V1 and V2 [3, 55, 80, 140], it cannot do so in an "all-or-nothing" fashion. At least some information from unattended parts of the visual field can reach higher level areas of the infero-temporal cortex and medial temporal lobe, where selective neuronal responses to various categories of objects [2, 21, 27, 71] have been found.

4.2 Natural Scene Categorization Is an "Easy" Task to Learn and to Perform

Perceptual learning is closely linked to the mechanisms of recognition and attention. Our results show that contrary to common belief, certain seemingly much simpler stimuli are harder to learn to discriminate than the complex natural scenes, when attention is not available. In fact, little stimulus specific training is necessary for subjects to perform the natural scene categorization task. Given the current models of visual recognition, this result is highly counterintuitive. Today's start-of-the-art computer algorithms take much more training to recognize natural scenes than simple geometric configurations [148, 153]. Hochstein and colleagues have coined the term "easy" for task conditions where considerable learning transfer occurs [59]. Under this definition, natural scene categorization is a much "easier" task given the results of our first set of experiments. They hypothesized that "easier" tasks involve higher cortical level processing than lower ones. We will revisit this point in our discussion of "meaningful categories," in which our further findings with differently rotated letters also give a hint to this possible architecture of the visual processing.

Another piece of indirect evidence of the "easiness" for natural scene categorization is its performance pattern without color information. In Control Experiment 2 we found that there is virtually no cost in removing color information from the natural scenes. Again, state-of-the-art computer algorithms for image retrieval often utilize color features as one of the most informative cues for categorization. Our results show a clear discrepancy between such algorithms and the actual properties of the human visual system. It is suggested that rapid natural scene categorization might take advantage of coarser, achromatic information from the magnocellular pathway earlier than the finer, chromatic parvocellular pathway [22]. Several studies on face recognition [27] also suggest that there is a response of IT neurons for the early phasic component of the

stimuli rather than more fine-tuned information. This suggests that categorical recognition might be achieved in higher level visual areas using early waves of neuronal information, where more detailed features such as colors and fine edges have not yet been computed or incorporated.

4.3 Parallel Processing

The robustness of natural scene categorization is further confirmed through the experiment of double-image recognition (Control Experiment 3). Our human visual system is surprisingly parallel in processing the complex stimuli of natural scenes [120]. In contrast, Exp 4 shows that such ability does not apply to stimuli that are defined by their low-level differences, such as the configuration of the red semicircle and a green semicircle. Attention has long been considered to be deployed preferentially to tasks that require much scrutiny and processing. This experiment confirms further that the seemingly much simpler stimuli requires more attention to be categorized.

It has been long suggested that the more a recognition task requires feature conjunctions and binding, the more attention will be needed for this task [139, 140]. Therefore only "elementary features" are processed in a parallel fashion (i.e., under visual search, where serial focal attentional scan is not required). Our results suggest the possibility that natural scene categories might belong to the set of "elementary features" while the color disks or rotated letters do not. But this type of feature is unlikely to be encoded in lower level visual pathways where receptive fields are small and neurons tend to respond to primitive features such as orientations and brightness.

4.4 Meaningful Categories

In an attempt to understand the efficiency and robustness of natural scene categorization, so far we have gathered much indirect evidence that visual tasks involving higher cortical levels are recognized easier, learned faster and deploy less attentional resource. In the last set of experiments, we find a stronger evidence suggesting that meaningfulness and familiarity might participate in determining attentional load and more efficient recognition and learning of the natural scene categorization task. Everything else being equal, an upright letter is discriminated much better than one rotated to a fixed, but uncommon orientation. There is little low-level difference between these two sets of stimuli, but they do differ in terms of familiarity and meaningfulness. A similar result was recently discovered by Reddy et al. [114]. In their study, they contrast gender discrimination of hairless upright faces versus inverted faces in the near-absence of attention. They find that little attention is required to perform the task with upright faces (which are both familiar and meaningful)

while the attentional cost is rather high with inverted faces.

This observation is also consistent with the recent development of change blindness studies. Change blindness has shown that attention is critical for our visual awareness [115, 128]. Changes of large patches of the visual world can escape our awareness without attending to them. But the amount of attention needed to discern such changes seems to depend also on the meaningfulness of the stimuli. Semantically relevant information is less likely to be neglected in change blindness than less relevant information [61].

In short, we hypothesize that natural scene categorization requires little or no attentional cost because of the familiarity and "meaningfulness" of the stimuli and task. Attention acts as a gauge for information processing. When the task or stimuli are unfamiliar, hence are not directly associated with previous neuronal representations, attention helps to select and process features for the recognition task. When there are pre-existing neuronal representations for a given task or stimulus, for example natural scene categorization, little attention is needed.

4.5 Conclusion

We have shown that natural scene categorization does not require attention and is independent of training with the specific type of stimuli. It is robust to lack of color information or increasing the set size of the stimuli presented. In contrary, multiple redundant copies of synthetic stimuli do not improve the performances of recognition without attention. Some simple tasks, such as single letter discrimination, require much attentional assistance unless the letters are presented in a familiar, upright position. We hypothesize that attention is particularly important for tasks that do not have neuronal representations in the visual pathway. Natural scene categorization, a well-learned and familiarized task for most human observers, does not require much attention.

Part III

Gist of Natural Scenes: Perception in a Glance

Chapter 5

Introduction

5.1 Background

When we visit a family member or a close friend, we often enjoy browsing through their family albums. It probably never occurs to us that while flipping through the pages of the albums, our visual system is working with superb efficiency and accuracy in rapidly grasping the meaning of every photograph. We never experience any 'perceptual glitch' if the previous page was set in downtown Manhattan and the next one switches to a conference room full of people and posters.

It has long been known that humans can understand a real-world scene quickly and accurately. Film makers first demonstrated this ability through a technique called 'flash cut.' In a commercial motion picture called 'The Pawnbroker' [81], S. Lumet inserted an unusually brief scene representing a distant memory. Lumet found that a presentation lasting a third of a second, though unexpected and unrelated to the flow of the main narrative, was sufficient for the audience to capture the meaning of the interposed scene [11].

Pioneering studies extended these anecdotal findings, bolstering the claim that humans could rapidly apprehend a real-world scene. Potter et al. utilized rapid serial visual presentations (RSVP) of images and revealed that subjects could perceive scene content in less than 200ms [108, 110]. Furthermore, she demonstrated that while the semantic understanding of a scene is quickly extracted, it requires a few hundred milliseconds to be consolidated into memory [108]. Later studies however documented limits to our perception of a scene. Rensink et al. showed that changes to retinotopically large portions of the scene will sometimes go unobserved. It is likely that this occurs if the regions are not linked to the scene's overall 'meaning' [115].

Other hallmark investigations attempted to elucidate the information involved in this 'overall meaning'; their conclusions regarding scene perception paralleled concepts in auditory studies of sentence and word comprehension. Biederman et al. found that recognition of objects is impaired when those objects are embedded in a randomly jumbled rather than a coherent scene [8]. They further identified several physical (support,

interposition) and semantic (probability, position, size) constraints, which objects must satisfy when within a scene, similar to the syntactic and grammatical rules of language [9]. They investigated how object recognition was modulated by violating these constraints. Their results suggested that the *schema* of a scene-or the overall internal representation of a scene that includes objects and object relations-is perceived within a single fixation [9], regardless of expectation and familiarity [11]. Boyce and colleagues also demonstrated that objects are more difficult to identify when located against an inconsistent background given a briefly flashed scene (150ms), further suggesting that both recognition of objects and global contextual understanding are quickly and deftly accomplished [13].

5.2 Contributions

While it has become clear then that some comprehension of scene is rapidly attained, the conceptual 'content' of this scene gist is still somewhat nebulous. What is it exactly that we perceive and understand when we glance at the world?

In this study, we focus on particular facets of this question:

- There has been no commonly accepted definition of the content of 'gist.' Mandler and Parker have suggested that three types of information are remembered from a picture: i) an inventory of objects; ii) descriptive information of the physical appearance and other details of the objects; iii) spatial relations between the objects [84]. In addition to this object information, propositional relationships between objects, spatial layout of the scene, and a general impression of the low-level features that fill the scene (e.g., texture, etc.) are speculatively incorporated into the scene gist [156]. Finally, Biederman has proposed that global semantic meaning or context also contributes to the initial surmisal of a scene [9]. Positing the 'contents' of a glance as an operational definition of *scene gist*, we would like to ascertain the visual and semantic information comprising scene gist.

- Rosch suggested that one distinguishes between 'basic-level,' 'super-ordinate level' and 'sub-ordinate level' object categories [117]. Similarly, Tversky and Hemenway proposed the same taxonomy for scene categories [142]. These authors motivate their theory with arguments of maximizing the visual and linguistic information conveyed during naming. Does human perception of natural complex scenes reveal a similar hierarchy of objects and scenes? What patterns arise during free recalling of cluttered scenes of various objects? Would responses reveal similar hierarchies or different ones?

- One parameter to vary in examining scene perception is the length of presentation times. We are curious to see whether there is a natural ordering in the range of percepts that becomes available under

increasing temporal constraints.

- In all previous studies of scene perception, the experimenters have a set of predetermined hypotheses to test. Their experiments are hence constructed to illuminate certain parameters relevant to their claims and questions. Our design however might broaden the scope of scene perception research. Through unbiased responses, we hope to uncover new aspects of scene perception previously not considered.

Keeping the above issues in mind, we propose to examine unbiased real-world scene perception as a function of display time. We have designed an experiment in which subjects view one of nearly a hundred natural scenes for a brief interval of time without any priming, pre- or post-stimulus cuing, as to its content. We ask them to type freely what they have seen in as much detail as possible. We vary the presentation time of the image between 27ms to half of a second.

Chapter 6

General Method

Our subjects were asked to freely recall what they perceived in briefly displayed images of real-world scenes. We explored the evolution of our subjects' reports as a function of the length of presentation times. Our data was collected in Stage I and analyzed in Stage II.

In Stage I, subjects viewed briefly a picture of a scene on a computer monitor and were then asked to type what they had seen, using a free recall method to collect responses. Chapter 6.2 explains the details of this stage of the experiment.

In Stage II, we asked an independent group of subjects to evaluate and classify the free recall responses collected in Stage I. Chapter 6.3 is a detailed account of this evaluation process.

6.1 Dataset

In most previous studies of scene perception or object recognition, line drawings were used as stimuli [9, 60]. Recently, several studies have used a large commercial database of photographs for studying the perception of scenes and categories [77, 135, 138]. This dataset, unfortunately, is a collection of professionally photographed scenes, mostly shot with the goal of capturing a single type of objects or specific themes of scenes. We are, however, interested in studying images of everyday scenes, as commonly seen by most people in a naturalistic setting. Therefore, we assembled a collection of images trying to minimize this sampling bias.

Fig. 6.1 and Fig. 6.2 show our dataset of 44 indoor images and 46 outdoor images collected from the internet in the following way. We asked a group of 10 naive subjects to randomly call out 5 names of scenes that first came to their minds. Some of the names overlapped. After pruning, we had at hand about 20 to 30 different words/word phrases that corresponded to different environments. We then typed each of these words/word phrases in the Google image search engine. On the first one or two pages of the search results, we randomly selected 3-6 images that are sensibly related to the keyword. The Google image search engine

Figure 6.1: 46 images of outdoor scenes in our dataset of 90 grayscale images.

tended to return images that are found on people's personal websites, most often taken with a snapshot camera. While everyone has a bias when taking a picture, we believed the large number of images from

Figure 6.2: 44 images of indoor scenes in our dataset of 90 grayscale images.

different unknown sources would help to average out these biases.

A number of authors have suggested that color information is not critical for the rapid categorization

of scenes [29, 33, 36]. While color could be diagnostic in a later stage of recognition [93], we are mostly concerned with the initial evolution of scene perception; thus we decided to use only gray scale versions of our images for our experiments.

6.2 Experimental Stage I: Free Recall

6.2.1 Subjects

Twenty two highly motivated California Institute of Technology students (from 18 to 35 years old) proficient in English served as subjects in Experiment Stage I. One author (A.I.) was among the subjects. All subjects (including A.I.) were naive about the purpose of the experiments until all data were collected. Subjects reported normal color vision and visual acuity (sometimes with corrective lenses or glasses), but underwent no tests in this respect.

6.2.2 Apparatus

Subjects were seated in a dark room especially designed for psychophysics experiments. The seat was approximately 100cm from a computer screen, which was connected to a Macintosh (OS9) computer. The refresh rate of the monitor was 75Hz. All experimental software was programmed using the Psychophysics Toolbox [14, 101] and Matlab.

6.2.3 Procedure

Fig. 6.3 illustrates a single trial of Stage I. An image from our dataset was presented for one of 7 different possible SOAs: 27ms, 40ms, 53ms, 67ms, 80ms, 107ms, and 500ms. For each trial, the particular SOA was randomly selected with equal probability from these choices. The image was then masked by one of eight natural image perceptual masks, constructed by superposing white noise band-passed at different spatial frequencies [145]. The subject was then shown a screen with the words:

Please describe in detail what you see in the picture. Two sample responses are: 1. City scene. I see a big building on the right, and some people walking by shops. There are also trees. Most of the trees are on the left of the picture, against some background buildings. 2. Possibly outdoor. I really cannot tell much. Probably some animals, maybe mammals...

Subjects were given an unlimited amount of time to write down their responses.

Each subject was shown all 90 images in the database, broken into 5 20-trial sessions. The images were presented in random order. At the beginning of each session, 4 images outside of the database were used to

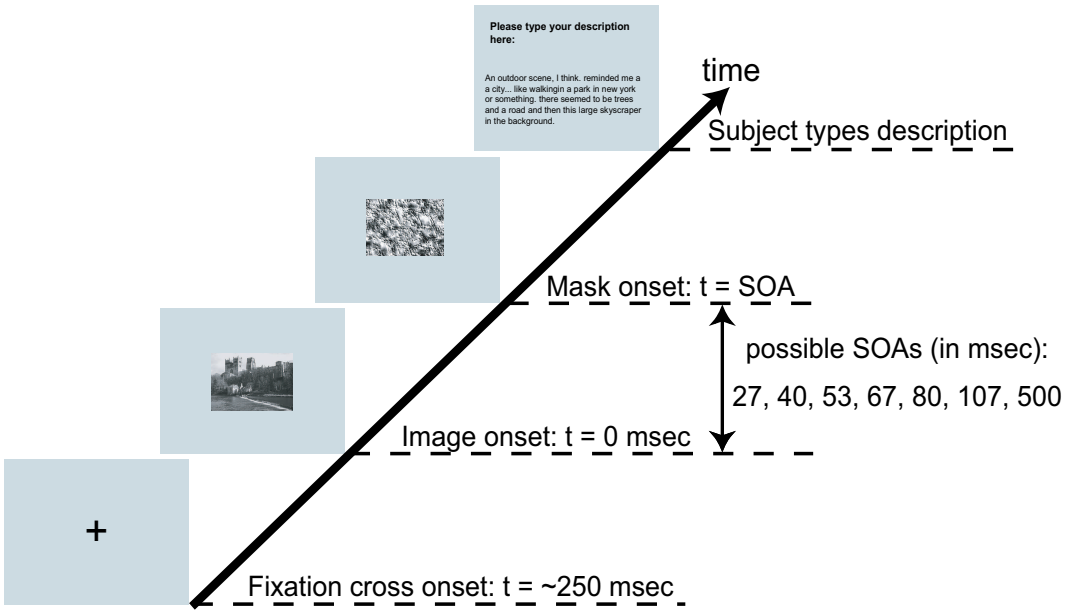


Figure 6.3: A single trial in Stage I: A fixation cross appeared for about 500ms. An image from our dataset was then presented at the center, subtending $6^\circ \times 8^\circ$ in visual angle. After a variable SOA, the image was masked by one of 8 natural image perceptual masks. The mask was displayed for $\sim$ 250ms. The time between the onset of the image and the onset of the mask is called Stimulus Onset Asynchrony (SOA). The mask was presented for 500ms. Afterward, the subject was prompted to a screen in which he/she was asked to type in the what he/she had seen of the image. Subjects were given an unlimited amount of time to write down their responses. When they were ready to continue, they could initiate the next trial by pressing the space bar.

familiarize the subject with the responses and SOAs. Free recall responses for these 20 (4×5) images were excluded from all data analysis. Order of image presentation, as well as the choice of SOA for each image, were randomized and counter-balanced among all subjects. Each subject thus contributed one description for each image at one of the SOAs. Overall, our 22 subjects provided 1980 descriptions, i.e., we obtained between 3 and 4 descriptions for each image and each SOA.

6.3 Experimental Stage II: Description Evaluation

6.3.1 Subjects

Five paid volunteer undergraduate students from different schools in the Los Angeles area (from 18 to 35 years old) served as scorers in Experiment Stage II. As scorers needed to analyze and interpret unstructured written responses, they were required to be native English speakers. All scorers were naive about the purpose of the experiments until all response evaluation was finished. Subjects reported normal visual acuity (sometimes with corrective lenses or glasses), but underwent no tests in this respect.

6.3.2 Apparatus

The scorers' task was to evaluate and classify the image descriptions obtained in the previous stage. For this purpose they used Response Analysis software, which we designed and implemented for this purpose (Fig. 6.5). Subjects were seated in a lighted office room. The seat was approximately 100cm from a computer screen, which was connected to a Macintosh (OS9) computer. The refresh rate of the monitor was 75Hz. All Response Analysis user interface software was programmed using MATLAB and the GUI toolbox.

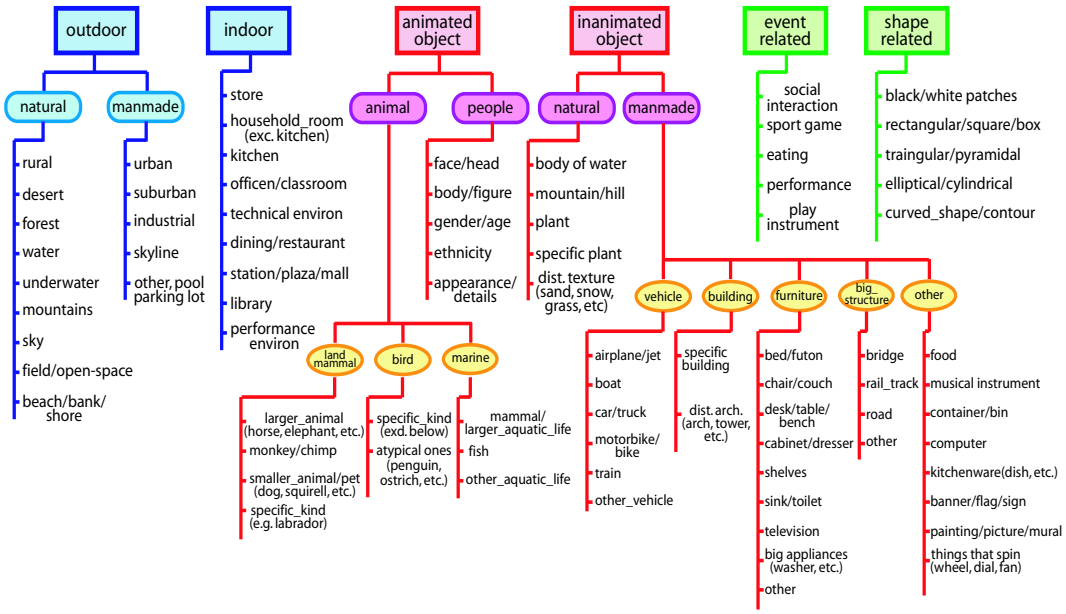


Figure 6.4: Attribute Tree. The list of attributes was constructed by examining the entire set of free recall responses/descriptions to extract a comprehensive inventory of terms referred to in these descriptions.

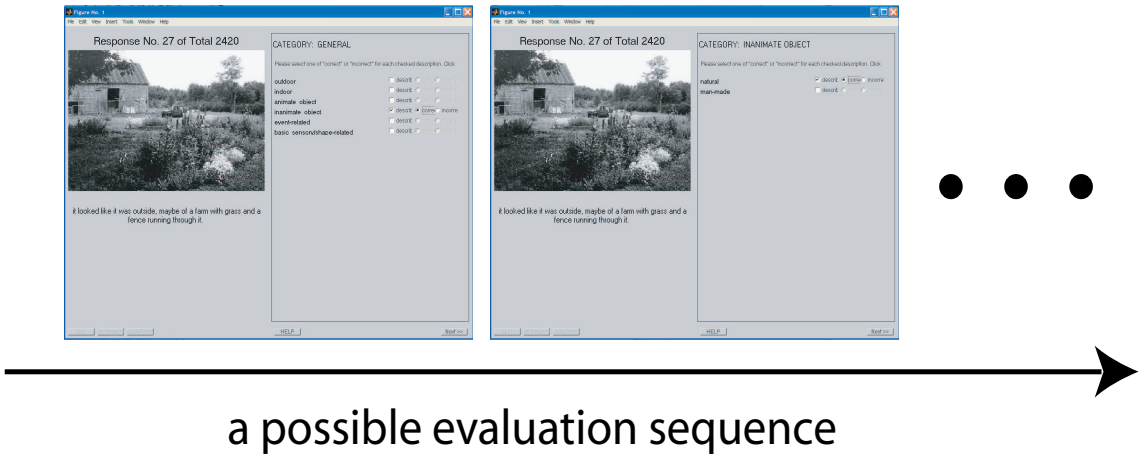


Figure 6.5: Experiment Stage II: Evaluating the free recall responses.

6.3.3 Procedure

Our aim was to evaluate free recall responses in a consistent and uniform manner for all subjects. To do this, the content of all responses was assessed with respect to a standardized list of attributes.

The list of attributes was constructed by the experimenters, who examined the entire set of free recall responses/descriptions to extract a comprehensive inventory of terms referred to in these descriptions. Most attributes described fell into one of six categories: inanimate objects; animate objects; outdoor scenes; indoor scenes; visual/perceptual features (i.e., shapes, lines, etc.); or event-related (this category comprised a more cognitive understanding of the picture, in which human behavior related to the scene was inferred, i.e., social interaction, sports/games, performances, concert, etc.). See Fig. 6.4 for the entire list of attributes.

The attribute list consisted of 105 terms. We organized these attributes into a hierarchical tree structure, where the highest level represented the most general level of description (e.g., inanimate object); the intermediate stages exhibited a greater degree of specificity (e.g., manmade inanimate object, building); and the lowest level corresponded to the most detailed level of description (e.g., Capitol building). This taxonomy schema stems from conventional notions of object and scene categorization, as originally developed by Rosch (1978) and Tversky and Hemenway (1983), predicated on the superordinate level; the entry, or basic, level; and the subordinate level. The findings of these authors formed the basis of our hierarchical classification for the animate object, inanimate object, indoor, and outdoor branches of the tree. The last two branches–sensory-related and event-related–have received less investigation, and thus were classified parsimoniously with only two levels, more general (e.g., sensory-related) and more detailed (e.g., lines, shapes, etc).

Each of the 5 scorers read every response (22 subjects who each responded to the same 90 images = 1980 responses) and assayed them for mention or description of each attribute as well as correctness. The scorer was guided through this task with the *Response Analysis* interface tool (Fig. 6.5). For each response, the scorer proceeded as follows: the first screen contained the text of one of the responses, the image described in the response, and a box with labels for the most general attributes: *indoor, outdoor, animate object, inanimate object, event-related, shape-related*. Next to each attribute, a button allowed the scorer to indicate whether the attribute had been described in the written response. If an attribute was checked as 'described,' the scorer was additionally required to indicate whether the description of the attribute was either an 'accurate' or 'inaccurate' depiction of the corresponding image. This completed the first screen. For any attribute checked, a successive screen was displayed, which comprised again the text of the response and the image, but now the next level of more detailed attributes; for example, if *inanimate object* had been checked in the first screen, a following screen would have contained the labels: *manmade* and *natural* (Fig. 6.4), for which the user would again be prompted to indicate whether these attributes were described in the response, and if

so, whether accurately described or not. If the user had then checked *natural*, a following screen would have contained: the text of the response, the image, and the next level of attributes: *body of water, plant, specific plant, mountain/hill, distinctive texture.* The entire branch was thus traversed.

If, on the first screen, the scorer had also checked *indoor*, then following screens would have also displayed: the text of the response, the image, and the next level of attributes: *store, household room, kitchen, office/classroom, technical environment, dining/restaurant, station/plaza, library, performance venue*. In this manner, the relevant portions of the tree were traversed, one branch at a time. This process was repeated for each response.

As explicated earlier, 3-4 responses were provided for a given image at a given SOA. For a given attribute, each scorer judged whether each of these 3-4 responses accurately described the attribute in the respective image. The percentage of responses rated as accurate measured the 'degree' to which the attribute was perceived in this image. This initial score thus reflected a particular image, SOA, and scorer. The scores were then normalized: the seven scores for a given image (one for each SOA) were divided by the highest score achieved for that image (across all SOAs). All evaluation scores were therefore between 0 and 1. Due to this 'within-image' normalization, inherent differences in 'difficulty' of perceiving or understanding scenes between different images were eliminated.

In all analyses, the scores were then averaged over all 5 scorers. In some analyses, the scores were additionally averaged over images, so that the averaged evaluation score represented the degree to which the attribute was perceived at a given SOA across the entire image set.

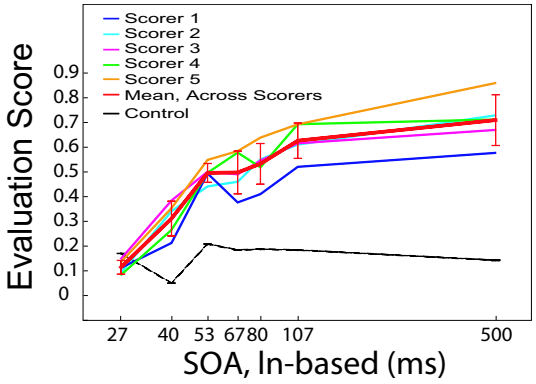


Figure 6.6: A sample score plot for the building attribute.

To better illustrate the parameters just discussed, we will focus on the evaluation of one attribute, *'building,'* depicted in Fig. 6.6. On the x-axis are the seven SOAs for which images were displayed. The y-axis reflects normalized accuracy evaluation score.

For the SOA of 80ms, for example, each scorer sees roughly 3 responses for each image. For each

response, the scorer determines whether the attribute ‘building’ was accurately reported with respect to the corresponding image (the other 104 attributes were also checked, but we will not follow those for the purposes of this example). Suppose the scorer indicates that building was described accurately in only one response. The initial evaluation score for the attribute ‘building’ for this image at SOA 80ms is therefore $1/3$, or 0.33. Suppose also that the maximum accuracy score achieved in describing this image occurred at SOA 500ms, where $2/3$ of the responses accurately reported a building. This maximum score of 0.67 would be used to normalize all scores, so that the evaluation score at SOA 80ms is now 0.5 and the score at 500ms is 1.0. This normalization allows each image to be its own baseline; therefore differences in the quality of the image (i.e., simple vs. cluttered) will not affect scores. Finally, all normalized ‘building’ scores at SOA 80ms–one for each image–are averaged to obtain the final evaluation score at this SOA for this particular scorer.

This process of normalization per image and then averaging over all images is done for each SOA. Again, the resulting values are per scorer. Thus, in Fig. 6.6, the yellow, blue, green, cyan, and magenta lines each represent the normalized evaluation scores (averaged over images) for one scorer.

These curves are then averaged over all the scorers. The resulting means are plotted in the red line in Fig. 6.6, with error bars representing standard error of the mean (s.e.m.).

In addition, there is a black line resting at the bottom of the plot. It consists of scores given by our scorers when the responses/descriptions are randomly matched to the images. This corresponds to our controls in the response evaluation process. As this evaluation process is subjective, scorer bias in judging accuracy of responses could be a potential confound, i.e., a scorer might be inclined to generally interpret vague or nebulous responses as ‘probably correct,’ giving ‘the benefit of the doubt’ even for inaccurate descriptions. To probe for this bias, each scorer was presented with 220 responses that were paired with an incorrect image (e.g., not the image the subject was viewing when making the response). The scorer had to indicate whether the response accurately described the image with which it was presented, the same task as for the real response-image pairings. Since these are incorrect pairings, responses associated with longer SOAs will not contain a more accurate description of any attribute (here, building) of the image with which it is presented to the scorer. Therefore, assuming no scorer bias, the line should remain flat, as observed in Fig. 6.6. If scorers do exhibit a propensity to liberally give credit to subjects for their responses, we would not expect to see a low, flat line from the controls. Instead, we would anticipate other patterns, such as increasing scores with SOA (since responses become more verbose and contain more content that could mistakenly be viewed as accurate); or scores that rise up to intermediate SOAs (more verbose but still somewhat ambiguous responses) but then decrease at the longest SOAs (where responses become specific enough that scorers have little opportunity to give subjects the benefit of the doubt). The control curves from all scorers were averaged.

Chapter 7

Experiments and Results

7.1 Experiment I: The 'Content' of a Single Fixation

How much of a scene can be initially perceived within the first glance?

Biederman's findings implied that some kind of global context of the scene is registered in the early stages of scene and object recognition [9].

Friedman and colleagues proposed that early scene recognition involves the identification of at least one 'obligatory' object [6, 44]. In this 'priming model,' the obligatory object serves as a contextual pivotal point for the recognition of other parts of the scene [56]. There is also evidence that objects could be independently recognized without facilitation by global scene context [56]. Despite this discrepancy between all these models, one thing is clear: object recognition is an important aspect of early scene perception. Humans appear to be able to recognize at least some objects in a naturally cluttered scene in a single glance.

So what is the content of the first glance of a scene? Does it include a list of objects, and/or relations of objects, and/or background textures, and/or layout of space [156]?

In this first experiment, we try to extract as much information as possible from subjects' reports of scenes in a single fixation.

7.1.1 Method

We compare the subjects' descriptions of scenes in two SOAs: 107ms and 500ms. While the average fixation length during scene viewing can be as high as 339ms [113], numerous previous studies have used presentation times between 100ms to 200ms to investigate the effect of single fixation [9, 13, 108]. Here we follow the tradition and use 107ms as an estimate of the length of the first fixation of a scene. 500ms is chosen as a baseline presentation time for viewing a scene. It is commonly accepted that this amount of time is sufficient for perceiving a natural scene and most of its contents. Fig. 7.1 shows two different example scenes and sam-

ple descriptions at the two SOAs. In the first row, the scene is grasped with relative ease. Subjects are nearly as good at perceiving the details of the scene at SOA 107ms as compared to the baseline viewing condition. In the second row, the scene is much more cluttered and complex. We see that the extra presentation time for SOA 500ms helps greatly in perceiving the details of the scene.

Figure 7.1: Subject description samples. In the first row, the scene is relatively easy. Subjects are nearly as good at perceiving the details of the scene at SOA 107ms as compared to SOA 500ms. In the second row, the scene is more cluttered and complex. When the paper is accepted for publication, we will publish all descriptions collected for the entire dataset.

Several attributes were examined, from five branches of the analysis tree and at various levels of abstraction, from super-ordinate to subordinate. The evaluation scores for each of these attributes were averaged over all images and all scorers. The scores for SOA 107ms and for SOA 500ms were compared; a pair of bars representing the scores at these two SOAs are plotted for each attribute of interest.

7.1.2 Result and Discussion

Since we are interested in eliciting a fuller description of the semantic 'content' of a brief look at a scene, five categories of attributes are considered: animate objects (including humans and animals); inanimate objects; outdoor scenes; indoor scenes; and human activities/events. Fig. 7.2 summarizes all results.

In Fig. 7.2(a) and (b), we show these comparisons for *objects*. As our focus is primarily on scene recognition, we will consider object recognition only briefly. In the super-ordinate category of *animate objects*(Fig. 7.2(a), most levels are equivalently perceived within a single fixation as compared to the baseline viewing condition. The super-ordinate level of *animate object*, more detailed levels such as *people*, and attributes *large mammal* (subordinate to *animal*) and *ethnicity*, *appearance and details* and *body/figure* (subordinate to *people*) are reported with similar accuracy and are insignificantly different (One-Way ANOVA: $0.06 < t(8) < 4.07$, $p > 0.05$). Three attributes differ weakly in a One-Way ANOVA: *animal* ($t(8) = 7.70$,

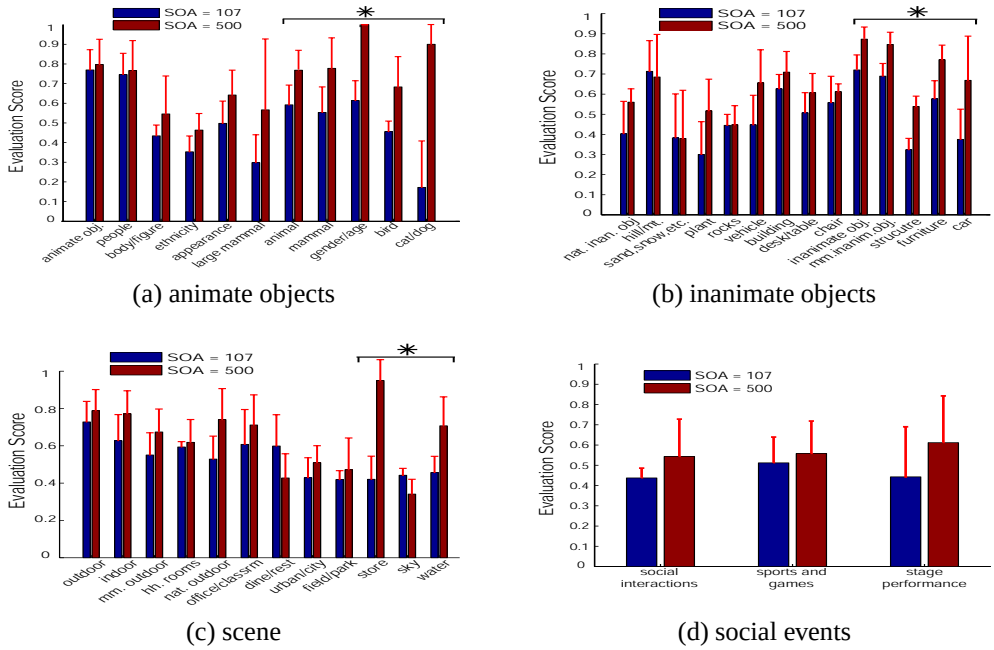
$p = 0.024$); *mammal* ($t(8) = 6.16$,$p = 0.04$); and *gender/age* ($t(8) = 9.73$, $p = 0.01$), and two others strongly differ (One-Way ANOVA: bird $t(8) = 73.32$, $p < 0.001$; dogs/cats $t(8) = 33.98$, $p < 0.001$). While several detailed attributes of people, such as ethnicity, appearance, and body figures, are perceived with adroitness, recognition of non-human animals does not appear to enjoy the same ease. Entry level animals, such as dogs, cats, and birds, are more reliably discriminated with longer presentation times, with dogs and cats being particularly poorly recognized at 107ms. These propensities speak to a large body of literature claiming an innate visual preference for faces and humans [30, 31, 116].

Fig. 7.2(b) displays the results for the inanimate objects contained in the image dataset. Several attributes pertaining to inanimate object categories are perceived within a single fixation, namely the super-ordinate category *inanimate natural objects*, plus more basic level objects such as *rocks*, *plants*, *mountain/hills*, *grass*, *sand* and *snow* (One-Way ANOVA: $4.24\text{e-}4 < t(8) < 4.02$, $p > 0.05$). In the realm of manmade objects, the findings are less clear. Super-ordinate levels, such as *manmade inanimate object*, *furniture*, and *structures* (roads, bridges, railroad tracks), and the basic level attribute *car* are more accurately reported at 500ms than at 107ms (One-Way ANOVA: $14.20 < t(8) < 31.95$, $p < 0.01$; except car, weakly significant: $t(8) = 6.10$, $p = 0.04$). Other super-ordinate and entry-level objects, including *vehicle*, *building*, *chair*, and *desk or table* exhibit equal accuracy at both SOAs (One-Way ANOVA: $0.80 < t(8) < 4.50$, $p > 0.05$). The lack of an unequivocal advantage for recognition of basic-level categories versus super-ordinate categories connotes a discrepancy from Rosch's study on object categories [117]. We observe that one of the main differences between our setup and Rosch's is the clutter and fullness of our scenes. In her study, objects are presented in isolation, segmented from background. In our setup, objects are viewed under more natural conditions, with clutter and occlusion.

Fig. 7.2(c) displays comparisons for the scene environments portrayed in our dataset. At SOA 107ms, subjects easily name the super-ordinate level categories, *outdoor*, *indoor*, *natural outdoor* and *manmade outdoor*. In addition, scenes such as *office/classroom*, *field/park*, *urban streets*, *household rooms (dining rooms, bedrooms, living rooms)*, and *restaurant* scenes are recognized within a single fixation (One-Way ANOVA: $0.20 < t(8) < 5.23$, $p > 0.05$). Only shop/store and water scenes require longer presentations (One-Way ANOVA: $9.93 < t(8) < 50.40$, $p < 0.02$; except sky, weakly significant: $t(8) = 6.73$, $p = 0.03$).Compared to objects then, scene context is more uniformly described by our subjects in a single fixation. Our results suggest that semantic understanding of scene environments can be grasped rapidly and accurately after a brief glance, with a hierarchical structure consistent with Tversky and Hemenway [142].

We have seen that both objects as well as global scene environments can be processed given a single fixation. These attributes, however, are explicitly denoted by properties of a still image, where the physical

features defining an object or the quintessential components of an environment can be readily rendered. Can a more cognitive appraisal of the transpiring scenario be inferred with the same ease? In Fig. 7.2(d), we look at attributes related to human activities and social events. Given our dataset, only five types of activities are included: sport/game, social interaction, eating/dining, stage performance, and instrument playing. Of the 5 activities, sport/game, social interactions and possibly stage performance can be reported after a single glance (One-Way ANOVA: $0.25 < t(8) < 1.54$, $p > 0.05$). Only one image each involved humans either eating or playing instruments; thus these event-related attributes were not statistically meaningful and were excluded from our analysis. These findings suggest that subjects cannot only extract objects as well as their embedded environment, but in addition can infer the interaction of the objects in order to consolidate an abstract, semantic meaning of their visual world.



(a) animate objects

(b) inanimate objects

(c) scene

(d) social events

Figure 7.2: Fixation results for animate objects (a), inanimate objects (b), scenes (c) as well as social events and human activities (d).

In summary, within this brief period of time, humans seem to be able to recognize objects at super-ordinate category level as well as a variety of basic category level. Furthermore, a single fixation seems sufficient for recognition of most common scenes and activities, many of them coinciding with the basic level scene categories suggested by Tversky and Hemenway [142].

7.2 Experiment II: Outdoor and Indoor Categorization

In recent years, several computer vision studies have suggested efficient algorithms for categorizing scenes, exploiting both global and local image information [35, 94, 132, 144, 150]. While these methods shed light on how coarse classification of scenes can be achieved in a feed forward fashion after supervised learning, little is known in the human vision literature about the actual cues and mechanisms allowing categorization of different scene classes. In their work on scene taxonomy, Tversky and Hemenway examined in particular people's understanding of the disparate components of indoor and outdoor scenes [142]. Their methods however treated indoor and outdoor environments symmetrically, presuming no obvious preference or bias.

After reviewing the free recall responses in this experiment, we observed a proclivity towards recognition and naming of one of these two kinds of environments. We thus sought to further explore this unexpected affinity.

7.2.1 Method

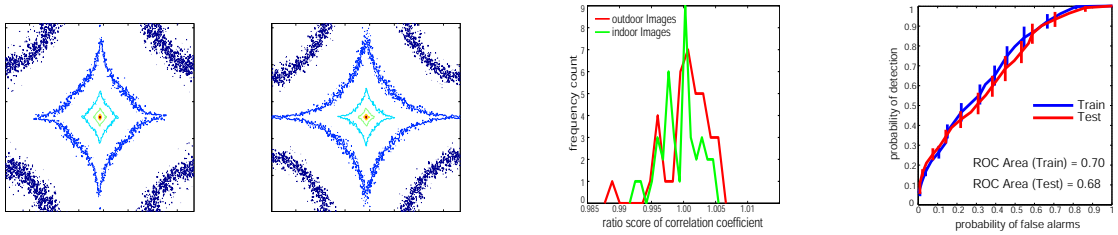
To probe for a possible bias, we examined how the outdoor and indoor images in our dataset were classified by our subjects, and how this classification changed as a function of presentation time (SOA). For each SOA, a scatter plot was generated, each dot representing an image–red dots correspond to ground-truth outdoor images, green dots to ground-truth indoor images. 'Ground-truth' is determined in the following way: for each image, we take all responses of all subjects at SOA 500ms. If a majority of the subjects accurately described the image as 'outdoor,' then the ground-truth label for the image is 'outdoor.' The same is true for the 'indoor' images. For each image, we are able to ascertain the percentage of subjects that labelled the image as 'indoor' or as 'outdoor' at a particular SOA time. Fig. 7.5 shows how the images are perceived at different times. We shall discuss this more in the Results section.

Before we proceeded, we wished to know whether a bias in subject performance could be accounted for by simple, low-level global cues. Indeed many studies have explored the usage of global cues for categorizing natural scenes, and computer vision algorithms have demonstrated relative success in utilizing such cues to accurately achieve a variety of classifications [94, 132, 144]. Following the same line of reasoning, we carried out two control analyses of the global statistics of the scenes in our dataset.

In the first control experiment, we assessed whether indoor and outdoor scenes in our database could be separated by simple frequency information [94]. Both the indoor and outdoor images were randomly divided into two halves—a 'training set' and a 'test set.' Two power spectrum templates were then created: 1) an outdoor template, which averaged the power spectra of all outdoor images in the outdoor training set, and 2) an indoor template, which averaged the power spectra of all indoor images in the indoor training set.

Fig. 7.3(a) and (b) show two example outdoor and indoor templates for randomly drawn training sets. For the images in the test sets, a two-dimensional correlation was performed between the power spectrum of each image and the outdoor template, and between the power spectrum of each image and the indoor template. We then obtained a ratio of correlation coefficients (outdoor correlation coefficient: indoor correlation coefficient) for each image in the test sets. This correlation analysis was repeated, with training and test sets reversed, i.e., the images previously in the training sets formed the new test sets, and the images formerly used in the test sets were used to generate the templates. Ratios of correlation coefficients were obtained for images of the new test sets. In this way, correlations were performed on every image in the dataset, with templates formed from a disjoint set of images. This procedure was reiterated 10 times, with a random segregation of images into either the training sets or test sets each time.

Fig. 7.3(c) shows the distribution of this ratio score for all of the outdoor and indoor images. We use this ratio score of the images to perform indoor versus outdoor classification. Fig. 7.3(d) is a Receiver Operating Characteristic (ROC) curve of the result. A weak classification result of 68.0% is achieved for separating indoor images from outdoor ones based on the average power spectra (chance classification by an ROC analysis is considered to be 50%). Compared to the average performance of human observers at SOA 500ms (90.5% in Fig. 7.5), this result indicates that little information could be used to classify indoor and outdoor scenes based on low-level power spectral information.

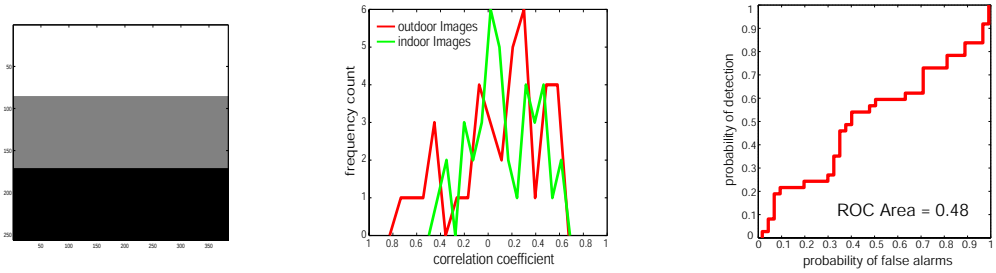


(a) outdoor template (b) indoor template (c) distribution of correlations (d) ROC curve

Figure 7.3: Power spectral analysis. **(a)** A sample outdoor template, which averaged the power spectra of all outdoor images in the dataset (excluding the image itself if it is outdoor). **(b)** A sample indoor template, which averaged the power spectra of all indoor images in the dataset (excluding the image itself if it is indoor). **(c)** Distribution of the ratio score for outdoor and indoor images. The ratio score of correlation coefficients is obtained from the outdoor correlation coefficient and indoor correlation coefficient for each image. **(d)** shows two Receiver Operating Characteristic (ROC) curves (training and testing) of the classification results based on the correlation ratios. A weak classification result of 68.0% is achieved for separating indoor images from outdoor ones based on the average power spectra in the testing case.

Our second control addressed the argument that outdoor scenes tend to have a lighter top partly due to the contrast of the sky, while there is no such cue in an indoor image. We therefore used a simple 'sky' template to explore this possibility (Fig. 7.4(a)). Three horizontal layers constituted this template, the top consisting of high-intensity pixels, the middle median-intensity pixels, and the the bottom of low-intensity

pixels. A two-dimensional correlation was performed between each image in the dataset and the template. The correlation coefficient for each image was used for classification. Fig. 7.4(b) shows the distributions of the correlation coefficients of all the indoor and outdoor images, while Fig. 7.4(c) shows the classification results in ROC curve. Only a 47.5% performance is achieved by using the template method. This is no better than chance, compared to a high human observer performance at SOA 500ms (90.5% in Fig. 7.5).



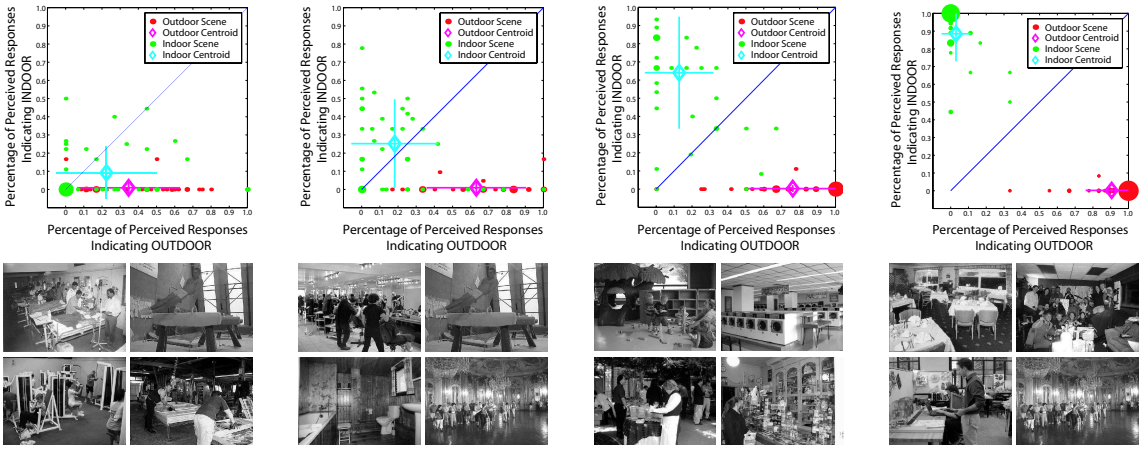
(a) light dark template (b) distribution of correlation scores (c) ROC curve

Figure 7.4: Light top dark bottom correlation analysis. **(a)** Two horizontal layers constituted this template, the top consisting of high-intensity pixels, and the the bottom of low-intensity pixels. **(b)** A two-dimensional correlation was performed between each image in the dataset and the template. The correlation coefficient for each image was used for classification. **(c)** shows the classification results in ROC curve. Only a 47.5% performance is achieved by using the template method.

7.2.2 Results and Discussion

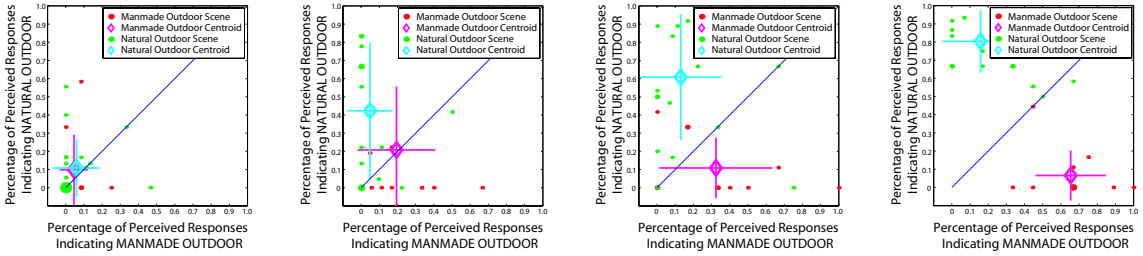
The recall performances for indoor versus outdoor scenes are shown in Fig. 7.5. We sampled the responses as a function of stimulus presentation times: 40ms, 67ms, 120ms and 500ms. Ideally, all outdoor images (green dots) would cluster at the $(0, 1)$ corner of each of the panel in Fig. 7.5, while all indoor images (red dots) would cluster at the $(1, 0)$ corner of the panel. At short SOAs, however, fewer subjects mention the indoor/outdoor category, while at 500ms, virtually all do. At the baseline SOA of 500ms (Fig. 7.5(d)), most of the red dots are indeed located on the x-axis, as subjects correctly identified the outdoor images as outdoor. Similarly, most of the green dots are located on the y-axis. In Fig. 7.5(a)-(d), we observe a very clear trend of an early bias for outdoor images. At SOA 40ms, if subjects chose to make the indoor/outdoor dichotomous distinction in their responses, they tended to identify asymmetrically outdoor images as outdoor, despite the fact that there is a similar number of indoor and outdoor images in the dataset. This preference for outdoor labelling continues even at SOA 107ms (Fig. 7.5(c)). In Fig. 7.5(a)-(d), we also present the four indoor images that were most frequently misclassified as 'outdoor' at the corresponding SOA. Several of them are consistent over a range of SOAs. By considering these images, it is possible that predominantly vertical structures give rise to the 'outdoor' percept more easily when there is less than 107ms for viewing the image.

In Fig. 7.7(c), we summarize the change of indoor and outdoor classification over presentation time in one plot. Each diamond represents the average performance score at one SOA.



(a) SOA 40ms (b) SOA 67ms (c) SOA 107ms (d) SOA 500ms

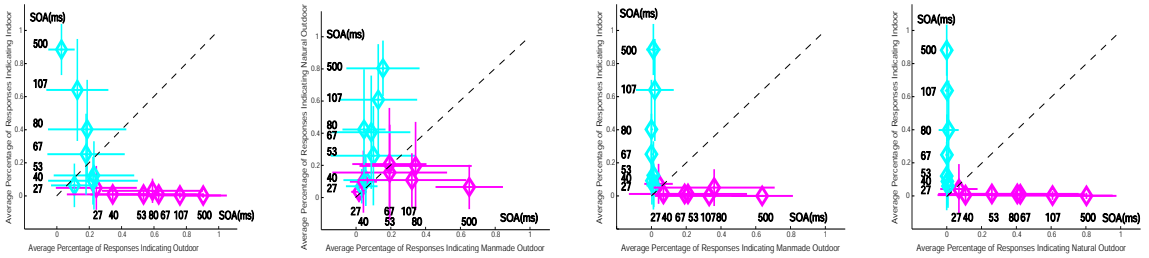
Figure 7.5: Categorization results of indoor and outdoor scenes. Each column illustrates the result in a specified presentation SOA. The top panel of each column is a scatter plot of the categorization results. Each dot represents an image in the database, red for ground-truth outdoor and green for ground-truth indoor. The x-axis indicates the percentage of subjects labelling an image as an outdoor image, and the y-axis indicates the percentage of subjects labelling an image as an indoor image. A diamond shape with error bars indicates the average performance. The bottom panel shows the four indoor images that were often confused as outdoor scenes given this SOA.



(a) SOA 40ms (b) SOA 67ms (c) SOA 107ms (d) SOA 500ms

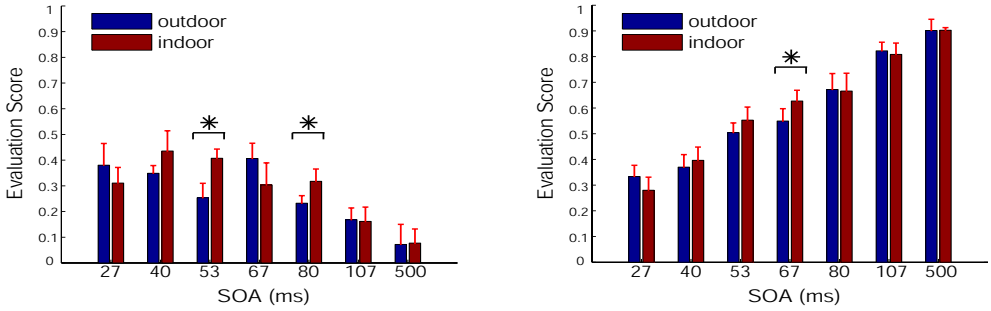
Figure 7.6: Categorization results of manmade outdoor and natural outdoor scenes. Each dot represents an image in the database. The ground-truth labelling is represented by a color: red for manmade outdoor and green for natural outdoor scenes. The x-axis indicates the percentage of subjects labelling an image as a manmade outdoor image, while the y-axis indicates the percentage of subjects labelling an image as a natural indoor image. A diamond shape with error bars is also plotted for each class of images (manmade outdoor and natural outdoor) to indicate the average percentage.

While we observe this strong bias in favor of outdoor over indoor classification of natural scenes for short display times, we do not see a large difference between manmade outdoor over natural outdoor images (Fig. 7.6). Subjects labelled both natural and manmade outdoor scenes with similar accuracy. Given shorter SOAs (less than 107ms), manmade outdoor scenes are at times confused with natural outdoor scenes, hence



(a) (b) (c) (d)

Figure 7.7: Summary plot of average categorization performances of all 7 SOAs. **(a)** Indoor versus outdoor scenes; **(b)** Manmade outdoor versus natural outdoor scenes; **(c)** Indoor versus manmade outdoor scenes; **(d)** Indoor versus natural outdoor scenes.



(a) sensory information: shape recognition (b) object recognition

Figure 7.8: Sensory information and object perception in outdoor and indoor scenes. **(a)** Sensory information perception performance comparison between indoor and outdoor scenes across all SOAs. **(b)** Overall object recognition performance comparison between indoor and outdoor scenes across all SOAs.

a lower average performance. But overall the trend is not nearly as pronounced as the bias between indoor and outdoor scenes (Fig. 7.7(b)).

Fig. 7.7(c) and (d) summarize average classification results for indoor vs. manmade outdoor images, and indoor vs. natural outdoor images, respectively. Unlike Fig. 7.7(a), there exists no indication of a bias in either of these conditions. This suggests that while indoor scenes tend to be confused as outdoor scenes, there is little confusion with manmade or natural outdoor scenes.

Where does this bias arise? Given the limited amount of information available when stimuli are presented very briefly (less than or about a single fixation), did outdoor pictures have an advantage over indoor pictures because subjects could perceive low-level, sensory related information more clearly? Or was it due to greater ease in identifying objects in the outdoor scenes versus the indoor scenes, as the 'priming model' would suspect [6, 44]? Fig. 7.8 illustrates the evaluation results in both indoor and outdoor scenes for sensory level information (panel (a)) as well as object level information (panel (b)), from the shortest presentation time (27ms) to the maximum (500ms). For sensory information perception, we see that the evaluation scores for

both indoor and outdoor images do not differ at most presentation times except SOAs 53ms and 67ms (One-Way ANOVA: $0.15 < \mathrm{t}(8) < 5.26$, $p > 0.05$; SOAs 53 and 67ms $11.46 < \mathrm{t}(8) < 26.60$, $p < 0.05$). Overall, there is no evident trend to suggest that outdoor scenes permit better sensory information recognition as compared to indoor scenes. Similarly, little trend is detected with respect to object level perception. For both indoor and outdoor images, the evaluation scores for the subjects' descriptions are not statistically different (One-Way ANOVA: $0.01 < \mathrm{t}(8) < 11.46$, $p > 0.05$, except SOA 67ms $\mathrm{t}(8) = 26.60$, $p < 0.001$). These results indicate that while there is an obvious preference for perceiving outdoor images for short presentation times, this bias does not seem to stem from a preference for perceiving the sensory information or object contents of the different environments.

7.3 Experiment III: Sensory-level Recognition vs. Object/Scene-level Recognition

Humans possess a superb ability in categorizing complex natural scenes. Thorpe and colleagues have demonstrated that the presence of an animal (or vehicle) in a photograph can be rapidly detected by subjects, and a neurophysiological correlate of this detection is observed in the prefrontal cortex area in as little as 150ms [135]. This ability is robust to multiple stimuli as well as withdrawal of attention [77, 120]. Further studies also suggest that a low-level, object-independent mechanism precedes the detection or recognition of semantically meaningful stimuli [64, 146]. A key question following these findings is that of the natural evolution of scene perception. In other words, what is the time course of such recognition? Although the exact timing differs in these studies, an overall consensus stipulates that recognition starts with the perception of low-level features and is followed by categorical recognition.

Similarly, traditional models of object recognition posit that a low-level visual processing precedes the high-level object recognition, in which segmentation takes place before recognition [25, 90, 122]. Other evidence suggests that semantically meaningful object recognition might in turn influence low-level, object-independent segmentation [103–105]. Recently, Grill-Spector and Kanwisher have found that humans are as accurate at categorizing objects as at detecting their presence [52]. Moreover, analysis of response time suggests not only that a similar amount of information is needed for these two processes, but also the same amount of neuronal processing time [52].

The conclusions above are drawn from experiments that rely on multiple forced choices paradigm, in which subjects are given a short list of possible answers before viewing the image. We are interested in examining the same question in a free recall scenario, one that is closer to the natural experience of scene

perception. Intuition tells us that different levels of recognition might occur upon processing different levels of information. While coarser or lower frequency information is sufficient to detect the existence of a dog, it is not necessarily adequate to know the dog is a husky or a German shepherd. We would like to, therefore, scrutinize subjects' descriptions of natural scenes at different presentation times in order to investigate the evolution of different levels of recognition. In particular, we would like to contrast whether higher-level conceptual information (e.g., object identification, object categorization, scene categorization, etc.) is perceived simultaneously with low-level or 'sensory' information (e.g., shape recognition/parsing).

7.3.1 Method

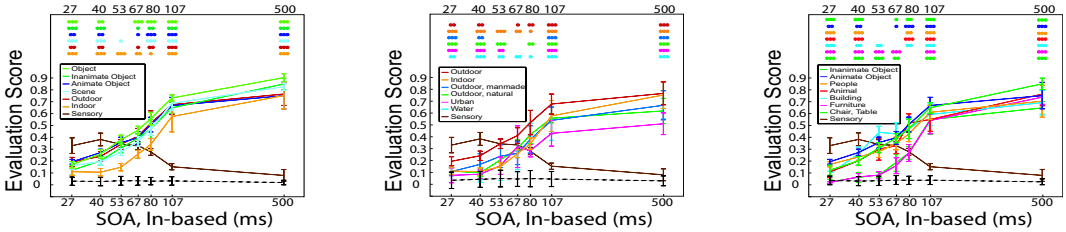
SOA 27ms	There was a range of dark sploches in the middle of the picture, running from most of the way on the left side, to all the way on the right side. This was surrounded primarily by a white or light gray color. (Subject: KM)	Couldn't see much; it was mostly dark w/ some square things, maybe furniture. (Subject: AM)	looked like something black in the center with four straight lines coming out of it against a white background. (Subject: AM)
SOA 40ms	I saw a very brght object, shaped in a pyramidal shape. There was somethin black in the front, but I couldn't tell what it was. (Subject: JB)	This looked like an indoor shot. Saw what looked like a large framed object (a painting?) on a white background (i.e., the wall). (Subject: RW)	The first thing I could recognize was a dark sploch in the middle. It may have been rectangular-shaped, with a curved top…but, that's just a guess. (Subject: KM)
SOA 67ms	possibly outdoors. maybe a few ducks, or geese. Water in the background. (Subject: JL)	I saw the interior of a room in a house. There was a picture to the right, that was black, and possibly a table in the center. It seemed like a formal dining room. (Subject: JB)	a person, I think, sitting down or crouching. Facing the left side of the picture We see their profile mostly. They were at a table or were some object was in front of them (to their left side in the picture). (Subject: EC)
SOA 500ms	It was definately on a coast byt hte ocean with a large [r]ock in the forground and atleast three bird sitting on the rock. (Subject: CC)	Some fancy 1800s living room with ornate single seaters and some portraits on the wall. (Subject: WC)	This looks like a father or somebod helping a little boy. The man had something in his hands, like a LCD screen or laptop. they looked like they were standing in a cubicle. (Subject: WC)

Figure 7.9: Samples of subjects' free recall responses to images at different SOAs.

In Chapter 6, we gave a detailed account of how subjects viewed and recorded their responses to each of the natural scene images in our database. Fig. 7.9 shows three of the images and some of their free recall responses at four different SOAs. Notice that when the presentation time is short (e.g., SOA $=$ 27ms or 40ms), the terminology used in the free recall responses tends to be shape and low-level sensory feature related, such as 'dark,' 'light,' 'rectangular,' etc. As the display time increases, subjects seem more confident at identifying the identity of the objects as well as the category of scenes. More conceptual and semantic terms, such as 'people,' 'room,' 'chair,' appear with increasing frequency.

We quantify the above observation by comparing the evaluation scores of the shape/sensory-related at-

tribute, as a function of presentation time, with scores of other, more semantically meaningful, attributes. In Chapter 6.3, we explained how both such evaluation scores as well as baseline performances are obtained. Note that our images are highly cluttered and objects tend to occlude each other. A correct label is given to a description as long as the shape information given is correct, not necessarily complete.



(a) Overall object and scene (b) Different scenes (c) Different objects

Figure 7.10: Perceptual performances of different attributes across all 7 presentation times. The perceptual performance is based on evaluation scores detailed in the Method section. The sensory related perception is plotted as a benchmark in all three panels. Perceptual performances for **(a)** overall scene and object attributes; **(b)** scene level attributes; and **(c)** object level attributes.

7.3.2 Results and Discussion

Fig. 7.10 summarizes our results. The y-axis of each panel is the evaluation score computed for each selected attribute(s). For comparison, we plot the sensory information response in all three panels of Fig. 7.10. The general trend in sensory information accuracy indicates that its score decreases, relative to other attributes, as the presentation time increases. This pattern is intuitive and predictable, as subjects cease to report shape or sensory related information when they are able instead to ascribe higher-level descriptions to the image, such as object labels, scene context, and semantic relationships among the objects.

In contrast, evaluation scores for attributes such as object names and scene types rise as the SOA lengthens. The accuracy and frequency with which these attributes are reported increases as more information becomes available. All panels of Fig. 7.10 also include a black line (at the bottom) corresponding to the random control responses. Since the scorers are evaluating descriptions that are randomly matched to images and various SOAs, it is expected that the average evaluation score is low and similar across all SOAs. This is indeed what we observe.

In Fig. 7.10(a), we compare the responses of low-level visual/sensory information to the high-level information related to object, animate object, inanimate object, scene, indoor scene, and outdoor scene superordinate categorizations. At SOA 27ms and 40ms, subjects report sensory level information more frequently and accurately than object- and scene-related information (One-Way ANOVA: $16.28 < \mathrm{t}(8) < 97.29$, $p > 0.05$). The object, inanimate object and animate object attribute information dominates over sensory

information at SOA 67ms (One-Way ANOVA: $2.21 < \mathrm{t}(8) < 36.86$, $p < 0.05$). Similarly, the outdoor scene attribute becomes indistinguishable to that for sensory level information at SOA 53ms (One-Way ANOVA: $\mathrm{t}(8) = 0.003$, $p = 0.96$), while the indoor scene curves overtake the sensory curve at 80ms (One-Way ANOVA: $\mathrm{t}(8) = 36.86$, $p = 0.03$). Once again, we find an obvious advantage for accurate report of outdoor scenes over indoor scenes, confirming our results in Experiment II.

In Fig. 7.10(b), the relation between sensory information and scene information is dissected at finer levels. Interestingly, if we analyze outdoor scene information at a finer level, for example manmade outdoor and natural outdoor, in both cases attribute report is inferior to that of sensory level information until SOA 80ms. On the other hand, the trajectory of manmade outdoor scene perception statistically coincides with the perception of even more subordinate categorizations of outdoor scenes, such as urban scenes (One-Way ANOVA: $0.37 < \mathrm{t}(8) < 5.22$, $p > 0.05$). Once again, the randomized control results remain a stagnant flat line at the bottom of the plot. This provides a glimpse into the order in which various kinds of semantic information becomes available for conscious report, as a function of presentation time.

In an analogous assessment, Fig. 7.10(c) displays evaluation scores as a function of SOA for object information. Somewhere between 45ms and 67ms presentation time, various levels of object perception become more pronounced than sensory level information (at SOA 67ms, animate and inanimate object are both significantly more reported than sensory information. One-Way ANOVA: $5.34 < \mathrm{t}(8) < 7.30$, $p < 0.05$). This switch in the predominant information reported transpires with shorter SOAs as compared to the reports of scene-related attributes discussed in the previous paragraph.

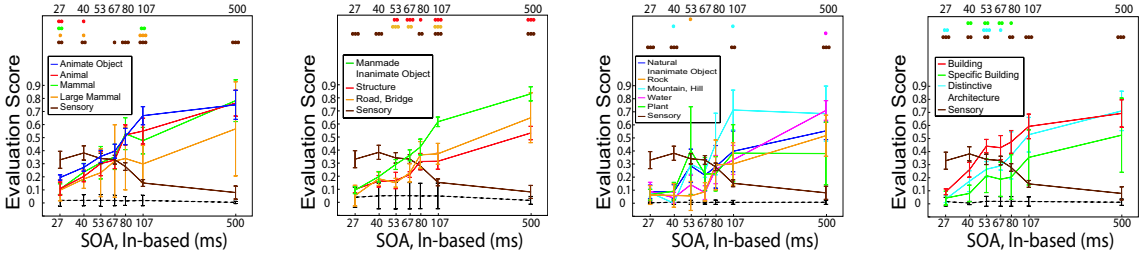
While our results cannot attest directly for the time course of information processing while viewing an image, our evidence suggests that on average, less information is needed to access some level of non-semantic, shape-related information in a scene compared to semantically meaningful, object- or scene-related information. This result is different from what Grill-Spector and Kanwisher reported in [52]. One major difference in our experimental design is that their subjects are forced to make a multiple choice while our subjects are instructed to write down whatever they recall. In addition, in their database, scenes that contain objects have very different statistics compared to the scenes that do not contain objects, namely randomized pixels. Studies have suggested that some reliable structural information of a scene may be quickly extracted based on coarse spatial scale information [93]. Consistent with these findings, our data seem to also show that coarse spatial information about shape segmentation can be perceived with less presentation of the image.

7.4 Experiment IV: Hierarchies of Objects and Scenes

It has been shown that some level of categorization of objects is most natural for identifying the object as well as for discriminating it from others. Rosch developed this category hierarchy for object recognition and identification; Tversky and Hemenway suggested a similar taxonomy for natural environments [142]. We were therefore interested in seeing if any correlation existed between our subjects' reports of scene and object recognition and those findings in [117, 142].

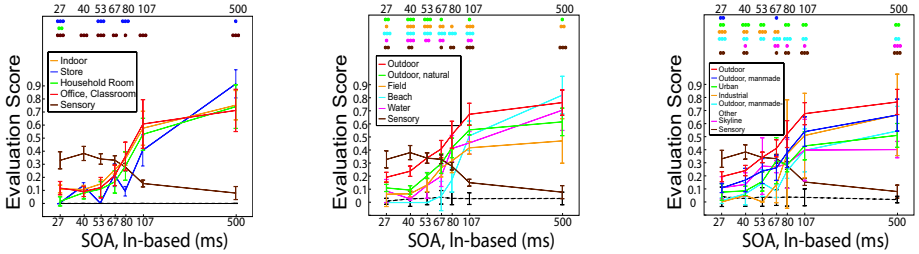
7.4.1 Method

We studied how different levels of object and scene categorization evolved as a function of presentation time (SOAs). We follow the same method as described in Chapters 7.3.1 and 6.3. Evaluation scores were averaged over images to provide an estimate of perception for each attribute at each SOA. Baselines were also constructed for these attributes and were averaged and displayed on each graph.



(a) Animate objects (b) Manmade inanimate (c) Natural inanimate (d) Building

Figure 7.11: Perceptual performances of different object attributes across all 7 presentation times. The perceptual performance is based on evaluation scores detailed in Chapter 6. The shape segmentation related perception is plotted as a benchmark in all three panels. **(a)** Animate object related attributes; **(b)** Manmade inanimate object related attributes; **(c)** Natural inanimate object related attributes; **(d)** Building and sub-ordinate building categories.



(a) Indoor scenes (b) Outdoor natural scenes (c) Outdoor manmade scenes

Figure 7.12: Perceptual performances of different scene attributes across all 7 presentation times. The perceptual performance is based on evaluation scores detailed in the Chapter 6. The shape segmentation related perception is plotted as a benchmark in all three panels. **(a)** Indoor scenes; **(b)** Outdoor natural scenes; **(c)** outdoor manmade scenes.

7.4.2 Results and Discussion

First we explored the relationship between levels of the *animate object* hierarchy. Subjects seem able to perceive coarser level animate objects more accurately than finer levels, particularly at shorter presentation times (Fig. 7.11(a)). We show in this plot three levels of animate objects: the super-ordinate level *animate objects*, *animal*, and *mammal*. At SOA 27ms, there exists a clear advantage for more accurate and frequent report of animate objects versus the other three categories (One-Way ANOVA between the following attribute and *animate objects*. *animal*: $t(8) = 12.38$, $p = 0.01$; *mammal*: $t(8) = 19.27$, $p = 0.002$; *large mammal*: $t(8) = 6.61$, $p = 0.03$). This advantage decreases by SOA 40ms, though it still retains statistical significance with respect to *animal* and *large mammal* (One-Way ANOVA: *animal* $t(8) = 9.99$, $p = 0.01$; *mammal* $t(8) = 1.25$, $p = 0.30$; *large mammal* $t(8) = 6.55$, $p = 0.03$). In short, given a very limited amount of information, subjects tend to form a vague percept of an animate object, but little beyond that.

A comparable advantage is found for manmade inanimate objects. Fig. 7.11(b) shows that while the evolution of structure and road/bridge are very similar, subjects tend to accurately report an overall impression of a manmade inanimate object rather than provide a more detailed level categorization. At short SOAs (27ms and 40ms), recognition of all levels of this hierarchy is poor. With longer presentation times (from SOA 53ms on), recognition improves, preferentially for the most super-ordinate level of 'manmade inanimate object' (significantly greater than structure and road/bridge for SOAs 53ms–500ms. One-Way ANOVA: $13.13 < t(7) < 40.5578$, $p < 0.05$. except at 80ms, vs. road/bridge (One-Way ANOVA: $t(7) = 1.24$, $p = 0.30$) and at 500ms, vs. road/bridge (One-Way ANOVA: $t(7) = 4.35$, $p = 0.08$). The trend is replicated in the hierarchy of structure recognition (Fig. 7.11(d)). In this plot, we observe that there is very clear gradation in terms of perception accuracy among buildings, distinctive architectural styles (e.g., Gothic building, triangular roof, etc.) and specific buildings (e.g., Capitol hill, Golden Gate, etc.). As with Fig. 7.11(b), accuracy is poor for all levels at SOA 27ms. With increasing presentation time, the more general attribute of 'building' is better discerned than finer level discriminations. From 40ms to 80ms, 'building' evaluation scores are significantly greater than those for the finest level of descriptive resolution 'specific building' (One-Way ANOVA: $9.82 < \mathrm{t}(8) < 23.02$, $p < 0.05$). For the earlier part of the same interval (53ms and 67ms), building perception is also superior to the intermediate level attribute of 'distinctive architectural features' (One-Way ANOVA: $10.12 < \mathrm{t}(8) < 25.73$, $p < 0.05$). Less of an overall trend is seen in natural inanimate objects, largely due to the high noise level of the plot (Fig. 7.11(c)). It seems that different levels of categorization occur more or less at similar times.

Our results on object hierarchies and the change of perceptual accuracy over increasing SOAs are not necessarily in conflict with the findings of Rosch [117]. In her study, the goal is to determine the level of

categorical representation that is most 'informative' and useful to identify and distinguish an object. An unspoken assumption is that this categorization is achieved given the full amount of perceptual information. In our setup, however, subjects do not have unlimited access to the images. They have to make a decision given the perceptual limitation due to the particular exposure length of the image. We find that under this setting, coarser level object categorization is in general more accurate than finer level ones. As information becomes more and more available (i.e., longer SOA), this difference becomes smaller. Whenever there is enough information, subjects would attempt to make a finer level categorization.

We adopted a similar strategy in examining the evolution of scene-related perceptions, as represented in Fig. 7.12. Fig. 7.12(a) shows, as a function of presentation times, the accuracy scores of 'indoor scenes' and three different 'basic-level' indoor environments: 'household rooms' (e.g., livingroom, bedroom, etc), 'office/classroom' and 'dining/restaurant' [142]. Unlike the hierarchical perception of objects, different levels of indoor scenes do not exhibit clear discrepancies in recognition frequency and accuracy at any SOA. Curves, representing the accuracy and frequency of indoor; household rooms, including living rooms, bedrooms, and dining rooms; offices; and classrooms, seem to overlap and are statistically equivalent (One-Way ANOVA: $0.06 < \text{t} < 5.04$, $p > 0.05$). The accuracy scores for store show a minor but significant deviation from the indoor curve at a few SOAs (One-Way ANOVA at 27 and 53 ms, for example $68.45\text{t}(8) < 111.70$, $p < 0.05$). However, only 3 images in our dataset correspond to store environments. This small sample may not be representative. Overall, it seems that once subjects decided that an image was an indoor scene, they had also determined what type of scene it was.

Fig. 7.12(b) shows the evaluation results for different levels of natural outdoor scenes (natural outdoor scene, field, beach and water). The coarsest level of the hierarchy, 'outdoor scene,' has a clear advantage over all other levels from the shortest SOA (27ms) till about 500ms (One-Way ANOVA: $5.96 < \text{t}(8) < 183.45$, $p < 0.05$) except at 80ms: outdoor natural $\text{t}(8) = 3.13$, $p = 0.11$ water $\text{t}(8) = 2.71$, $p = 0.14$). Outdoor scenes can be then further identified as a 'natural outdoor' scene. Analogous to the indoor scenario, once subjects have classified an image as a natural outdoor scene, they are capable of further identifying its basic-level category. There is no statistical difference among the evaluation scores for natural outdoor and many of its subordinate categories, such as field, mountains, and water. The one notable exception is the entry-level scene 'beach,' which is significantly lower at all SOAs until 107ms (One-Way ANOVA $< \text{t}(8) <$, $p > 0.05$).

A commensurate hierarchical trend is observed in manmade outdoor scenes (Fig. 7.12(c)). Again, here the most abstract level attribute (outdoor scene) is more accurately perceived than the basic-level scenes from SOA 27ms on (One-Way ANOVA for most attributes at all SOAs $4.16 < \text{t}(8) < 128.94$, $p < 0.05$. except 'industrial' at SOA 80 and 107 $0.64 < \text{t}(8) < 1.25$, $p > 0.05$; and 'skyline' at SOA 27 and 53ms

$0.47 < \text{t}(8) < 4.16, p > 0.05$). But the perceptual accuracy scores of manmade outdoor scene, urban centers, skylines, industrial environments and other manmade outdoor environments are essentially indistinguishable. A few instances of significant but small differences were noted between manmade outdoor and industrial, and between manmade outdoor and 'other manmade' scenes (for example, One-Way ANOVA for other manmade $\text{t}(8) = 27.41$, $p < 0.001$). These categories comprised images of construction sites, parking lots, and swimming pools; such scenes have not been mapped out in terms of their taxonomy and could conceivably be specific subordinate rather than basic level categories. This may in part account for these findings.

Tversky and Hemenway have suggested a taxonomy of scenes similar to that of objects [142]. Their study follows a similar line of arguments as Rosch [117]. Our results show, however, that scene perception differs from object perception. While object recognition reveals some hierarchical structure, only the overall categorization of 'outdoor' environment seems to need less information than recognition of other scene types. In general, super-ordinate level scene categories (e.g., indoor, manmade outdoor, natural outdoor) seems to require the same amount of information in recognition as the basic-level scenes (e.g., field, beach, skyline, urban centers, etc.).

7.5 Experiment V: Object and Scene Perception: Are They Correlated?

Intuitively, much of the meaning of a scene is defined by the objects that comprise the scene. Biederman has shown that recognition of objects is impaired when embedded in jumbled scenes rather than coherent scenes [8]. On the other hand, recent computational work has suggested that global features such as the spatial frequencies of the images are often sufficient for categorizing different environments without explicit recognition of the objects [138]. So are the objects in the scene perceived first? Or is the scene context grasped independently, and perhaps prior to recognizing the objects? How are the two perceptions related? Such questions have been open for debate for more than two decades [46, 51, 60].

Supported by studies of scene consistency and object detection, the *perceptual schema model* proposes that expectations derived from knowledge about the composition of a scene type interact with the perceptual analysis of objects in the scene [9, 13, 85, 98]. This view suggests that scene context information can be processed and accessed early enough to influence recognition of objects contained in scene, even inhibiting recognition of inconsistent ones [11].

The *priming model*, on the other hand, proposes that the locus of the contextual effect is at the stage when a structural description of an object is matched against long-term memory representations [6,44]. Regardless of

the mechanism, both the priming model and the perceptual schema model claim that scene context facilitates consistent objects more so than inconsistent ones. These theories predict that we should observe a correlation of object identification performance with scene context catetgorization performance.

In contrast, a third theory called the *functional isolation model* proposes that object identification is isolated from expectations derived from scene knowledge [60]. It predicts that experiments examining the perceptual analysis of objects should find no systematic relation between object and scene recognition performance [60].

In this experiment, we do not attempt to resolve the debate between these models directly. Instead, we look at the correlation between subjects' perceptions of different levels of object categorization with scenes as the presentation time changes. If scene and object perception follow from unrelated and disparate mechanisms as the functional isolation model suggests, little correlation between the two should be observed regardless of the presentation time. Conversely, if they share computational resources or facilitate each other in some way, we expect a correlation between the perception of objects and scenes. Furthermore, if there is a correlation between object and scene, we would like to know how this correlation is affected by the amount of available information—in other words, how different levels of object categorization relate to overall scene perception.

7.5.1 Method

For each SOA, we were interested in the correlation between perception of the overall scene context and that of the various levels of object categories (e.g., animate objects, animals, large mammals, etc.). If for example we wanted to look at the correlation between overall scene perception and overall object perception at SOA 40ms, we found the evaluation scores for these two attributes for every image at this SOA. We then performed a straightforward correlation between the two sets of scores for all images. The same process can be repeated for any pair of attributes.

7.5.2 Results and Discussion

We show the relationship between object level information and scene level information in Fig. 7.13. Each of the 8 panels in Fig. 7.13 is a scatter plot of the evaluation scores for these two attributes. Let us take Fig. 7.13(a) as an example, where we show the object and scene recognition at SOA 40ms. Each dot on the scatter plot represents one image. If more than one image falls on the same coordinate, the size of the dot increases linearly with the number of images. Fig. 7.13(a)-(d) uses the *scene* attribute as a bench mark. The red dots represent the images with the top 20% of evaluation scores for *scene*, at the baseline condition (SOA

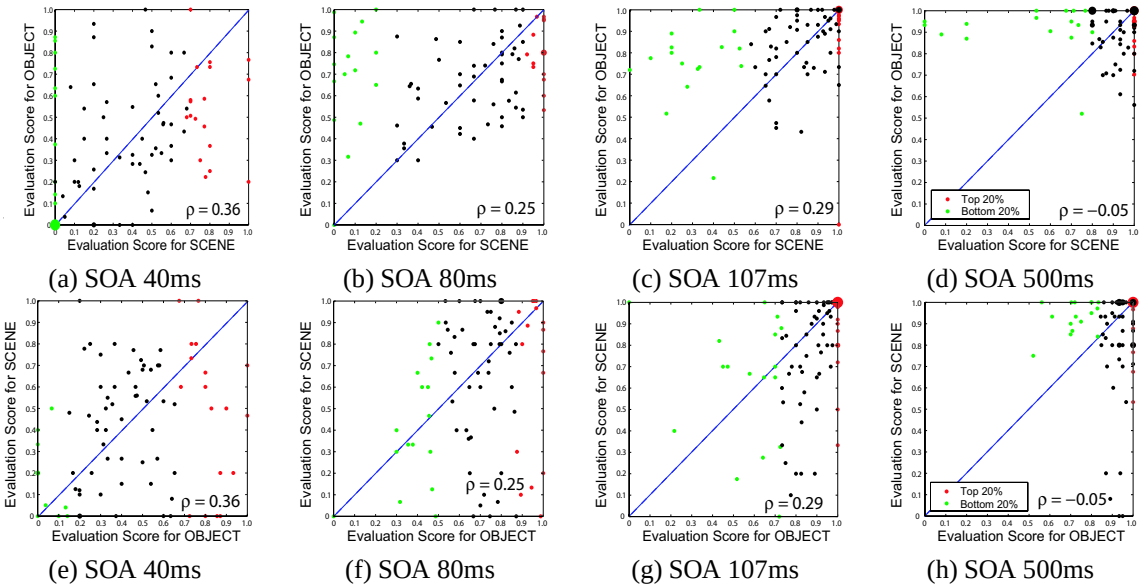


Figure 7.13: Object recognition performance versus scene recognition performance at various different SOAs. Performance is based on evaluation scores. See Results and Discussion sections for detailed explanations.

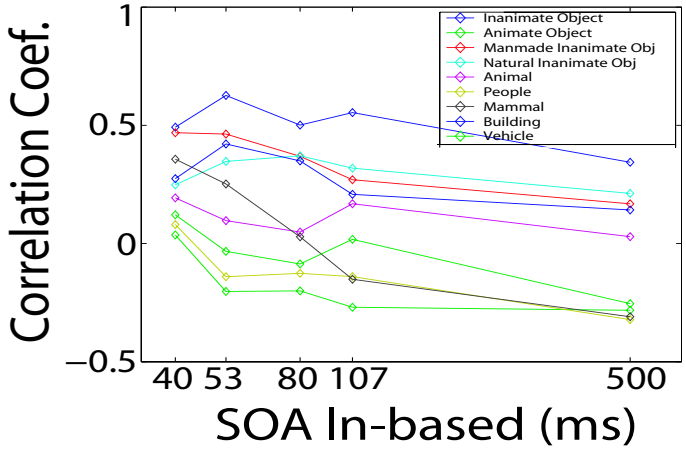


Figure 7.14: Overall correlation coefficients for scene versus objects and breakdowns.

500ms). The green dots are the images with the lowest 20% of evaluation scores for *scene* at the baseline condition. The black dots represent the remaining images. For Fig. 7.13(a)-(d), we are interested in these images' evaluation scores for the *scene* attribute (x-axis) as well as *object* attribute (y-axis) at SOA 40ms, 67ms, 107ms and 500ms. On each scatter plot, we also show the correlation coefficient computed across all images. From 40ms to 107ms, there is a weak correlation between the scene attribute and the object attribute ($\rho(40ms) = 0.38$, $\rho(80ms) = 0.26$, $\rho(107ms) = 0.29$), suggesting that subjects will perceive objects a little more accurately when they perceive scenes more accurately. At SOA 500ms, this correlation becomes

nearly 0. But both scene and object scores cluster near the upper right corner of the plot, indicating very high accuracy of perception for both of these attributes. Similar to Fig. 7.13(a)-(d), Fig. 7.13(e)-(h) show the relationship between scene and object recognition using the *object* attribute as a bench mark. In this case, the red dots are images that have the top 20% of evaluation scores for *object* under the baseline condition, and the green dots are those images with the lowest 20% of evaluation scores. Since correlation does not reflect causality, we should obtain the same correlation score whether the object or the scene attribute is used as a bench mark. Our data in Fig. 7.13(e)-(h) show the same correlation scores as each of their counterpart plots in Fig. 7.13(a)-(d).

We can further explore the different relationships between scene perception and various level object attributes at different presentation times (Fig. 7.14). The x-axis is the log scale of SOA times, ranging from 40ms to 500ms. Most of the object attributes receive very low evaluation scores at 27ms, hence the omission. The y-axis is the correlation score between a given attribute (e.g., inanimate object) and overall scene perception.

Compared to objects, the inanimate object attribute possesses a much stronger correlation with scene perception (average correlation score between 40ms to 107ms is 0.55 for inanimate object, and 0.30 for overall object, $p <$ 10e-3). This relatively stronger correlation between scene and inanimate object perception continues as we break it down to manmade inanimate objects and natural inanimate objects. They each have an average correlation score of 0.39 ($p \leq 0.01$) and 0.32 ($p \leq 0.04$), respectively (for SOA 40ms to 107ms). In Fig. 7.14, we also show two manmade objects, *vehicle* and *building*. Interestingly, while *building* is very similar to *manmade inanimate object* in terms of correlation between its recognition accuracy with scene perception (average correlation score of 0.31 for SOA 40ms to 107ms, $p \leq 0.02$, except SOA 107ms, $p = 0.09$), *vehicle* attribute seems to have a near 0 correlation with the scene (average correlation score of 0.01 for SOA 40ms to 107ms, $0.40 \leq p \leq 0.92$).

Curiously, the predominantly strong correlation between inanimate object perception and scene perception does not hold for those attributes involving animate objects. At the coarsest level, *animate object* recognition has an average correlation score of -0.15 with scene perception (for SOA 40ms to 107ms, $0.02 \leq p \leq 0.77$). At various levels of animate object recognition, the correlations with scene perception oscillate between no correlation (e.g., people, an average correlation of -0.08 for SOA 40ms to 107ms, $0.25 \leq p \leq 0.51$) and a very weak correlation (e.g. animal and mammal, both with average correlation of 0.12 for SOA 40ms to 107ms, $0.12 \leq p \leq 0.92$).

Our observations do not suggest causality. We merely indicate the correlation or lack of correlation between scene perception and various levels of object perception. Overall we see a weak but significant

correlation between scene and object perception (Fig. 7.13) at and up to presentation times of 107ms. This correlation might suggest several possibilities: i) object and scene perceptions might share at least some resources in processing; and/or ii) object (or scene) perception facilitate processing of scene (or object) perception. Both the schema model and the priming model would support the present observation.

What is curious is that this correlation is not evenly shared by inanimate and animate objects (plus vehicles). Fig. 7.14 demonstrates clearly that there is a qualitatively different correlative relation between inanimate objects and scenes versus animate objects and scenes. The general trend is that recognition of inanimate objects is dramatically more correlated with the perception of scene context than is perception of animate objects. Given this observation, one possiblity is that familiarity may account for the diminished facilitation between animate objects and overall scene perception. If we are innately more familiar with animals, especially human figures, the perception of these objects may depend less on the facilitation from other factors. Interestingly, vehicle is among the least correlated object categories with scenes. Given our modern lifestyle, subjects are in general very familiar with various kinds of vehicles in the pictures in our database. Another highly speculative hypothesis would be that there is less mutual facilitation between the recognition of mobile objects (such as animals, people and vehicles) and scenes. If prior knowledge of these objects informs us that they are likely to move from scene to scene, there might be less expectation for recognizing them in any particular scene. Admittedly, much still needs to be done to fully understand this unexpected asymmetry between inanimate and animate objects.

Chapter 8

Summary

We have shown a novel method to study scene perception. We collected free recall responses from subjects who were instructed to view 90 different real-world scenes under different presentation times. An independent group of subjects then evaluated the free recall responses. In this chapter, we summarize several interesting findings from this novel approach.

8.1 The Gist of Gist

The term 'gist' has long been used to refer generally to the overall crux or meaning of something. In the world of human vision, the term 'gist' has frequently been applied to scene understanding, yet the central question remains as to what actually constitutes this scene gist. We would like to suggest that the term 'gist' is used to denote the perceived contents of a scene given a certain amount of viewing time. A sensible and intuitive proposal would be a single glance or fixation. Many studies have shown much can be seen within a single glance of a scene [8, 13, 52, 77, 135, 145]. These experiments, however, are all conducted with some form of forced multiple choices. In Experiment I, we have collected a list of scene attributes perceived by subjects within a single glance of real-world scenes. This list includes most common scene types, super-ordinate categories of objects and a variety of basic categories of objects, as well as social activities and human interactions. We suggest that these are all part of the scene 'gist.' It is also important to point out that in our list of scene attributes, we do not include any sensory level information, such as shapes and illumination contrast. Experiment III shows that such information can clearly be accessed within a single glance. In fact, other semantically more meaningful attributes quickly predominate over sensory description in subjects' reports. Since our assumption limits the definition of gist within semantically meaningful attributes, we do not include the sensory and shape information.

Information contained in the 'gist' of a real-world scene seems to enjoy a tremendous privilege in visual

processing. Temporally, this privilege is reflected through the ultra-rapid speed with which the brain categorizes natural scenes [135]. Spatially, this complex scene categorization is not affected when spatial attention is deployed elsewhere [77]. Our results in Experiment I further suggest that a rich collection of perceptual attributes is represented and rises to conscious memory within a single fixation. Beyond a list of objects and scene environment [156], more cognitive appraisals of the event—such as social interaction and sports events—can be recognized effortlessly. It would be highly interesting for future studies to investigate the neural correlates that are responsible for such superb ability of real-world scene perception.

8.2 Shapes, Objects and Scenes

A key question in perception is the neuronal time course a given perceptual task follows, in other words, through what stages is a stimulus processed in order to manifest as semantically meaningful concepts.

The ventral visual pathway, linking the primary visual cortex through Inferior Temporal cortex to the prefrontal cortex, is generally known as the 'what' visual pathway, as it is responsible for object recognition through integrating features [28, 70, 89, 143]. Given the hierarchical structure of the visual system, many have proposed a model in which elementary features of objects are first processed and then bound together for object recognition [140, 157]. An ongoing debate in this picture is whether shape segmentation is a necessary intermediate step between low-level feature processing and high-level object recognition [25, 90, 122]. Recently, Grill-Spector and Kanwisher have found that categorization of super-ordinate to basic level objects (e.g., vehicle, musical instrument, bird, car, dog, etc.) is as accurate and fast as the mere detection of the object [52]. Their conclusion is based on an experiment in which subjects are asked to either choose one of the possible object categories or respond simply if an object is detected. Comparing their non-object distractors, it is obvious that the low-level image statistics of the distractors (mostly pixel noise) are drastically different from the images containing objects (all containing a central blob). Given this expectation, subjects are likely to heighten their search for a centrally located blob when detecting objects. In our experiments, subjects viewed freely a naturally cluttered real-world scene. Because our scenes are highly variable, they cannot expect a centrally located blob when looking at an image. In Experiment III, we found that shape related information has a slight advantage over semantically meaningful information of a scene. Our dataset shows less information seems needed for lower-level shape recognition compared to higher-level semantically meaningful recognition. This temporal constraint implicates a lower, feature level processing in facilitation of the initial stages of complex scene recognition.

Another major question regards object recognition in cluttered scenes. Several psychological models have been proposed to suggest different mechanisms of scene and object perception [6, 8, 9, 44, 60, 84, 98].

information can serve as facilitating media for more accurate object recognition [8, 9]. Similar to this view, Friedman et al. proposes in the priming model that a pivotal object, serving as the locus of the context, becomes a seed in the long term memory of scenes [6, 44, 98]. Opposite to these two views, Henderson and colleagues argue that object identification is independent from scene knowledge [60, 84]. This model predicts that recognition of objects in a scene and the scene context itself should have little effect on each other. In Experiment V, we show very weak evidence that object and scene recognition might be correlated when information is scarce. But this correlation is not uniformly distributed among different level of object categories. Results in Experiment V tells us that there is a stronger correlation between various levels of inanimate objects and scenes compared to animate objects and scenes. We will come back to this point in more detail in the next section.

In general, the question of the processing stages of cluttered scenes is still largely unsolved. Our experiments add evidence that there might exist a mutual facilitation between overall scene recognition and object recognition. In addition, low-level shape processing seems to require less information and possibly time compared to more high-level, semantically meaningful categorizations of objects and scenes. Traditionally, scene comprehension tends to be viewed in a serial fashion–in the order of sensory information, object features, objects, and the overall scene. Many new studies have now suggested that contrary to this view, high-level perception of natural scenes might be a highly efficient and parallel process [52, 77, 120, 135]. It would be interesting to examine an alternative hypothesis in which most of the recognition stages occur in parallel and constantly feed back information to each other to enhance the overall recognition of various components of the scene. In this possible scenario, early sensory information extraction stages still precede most of the semantic recognition stages. But as soon as there is any information for any possible level(s) of recognition, our brains take advantage of this.

8.3 Two Puzzling Asymmetries?

In Experiment II, we observe a strong preference for outdoor scenes over indoor scenes when visual information is scarce. Subjects seem to assume by default that an ambiguous image is more likely to be outdoor than indoor. This effect diminishes as the presentation time lengthens. At 500ms, outdoor and indoor scene categorization becomes nearly perfect. Our results further show that the bias only appears at the most super-ordinate level. When indoor scenes are compared with manmade or natural outdoor scenes, the bias disappears. Furthermore, neither segmentation nor object recognition seem influenced by this bias between these two categories of scenes. So what is it that causes this bias? Recent computational models have shown that using global and local cues such as edge and color information, it is possible to separate most outdoor

and indoor scenes [35, 132, 138, 144]. This strongly suggests that whatever feature(s) enables this discrimination is(are) either missing or inaccessible when information is scarce. More studies should be performed to pinpoint exactly what it is. This might be a very useful entry point to investigate the features needed for rapid scene categorization.

Another curious asymmetry we observe in Experiment V is the stronger correlation between inanimate object recognition and overall scene context versus that between animate object recognition and overall scene context. One possible explanation of this phenomenon is the effect of familiarity. It has been long known that there might be special neuronal resources designated for human parts such as faces and bodies [24, 30, 31, 67, 116]. We have also found recently that familiarity might modulate the level of attentional requirement in object recognition tasks [36]. If there is indeed an innate preference for animate objects such as animals and humans, there might also exist efficient computational mechanisms for the visual system to process this information rapidly and accurately. Compared to other object categorization, it might therefore be less dependent on possible mutual facilitation mechanisms with scene gist perception. As this is largely speculation, more experiments need to be done to address these hypotheses and account for this asymmetry.

Part IV

Computational Models I: Object Recognition

Chapter 9

Introduction

9.1 Introduction and Motivation

Recognition is one of the most useful functions of our visual system. We recognize materials (marble, orange peel), surface properties (rough, cold), objects (my car, a willow) and scenes (a thicket of trees, my kitchen) at a glance and without touching them. We recognize both individuals (my mother, my office), as well as categories (a 1960's hairdo, a frog). By the time we are six years old we recognize more than 10^4 categories of objects [10] and keep learning more throughout our life. As we learn, we organize both objects and categories into useful and informative taxonomies and relate them to language. Replicating these abilities in the machines that surround us would profoundly affect the practical aspects of our lives, mostly for the better. Certainly, this is the most exciting and difficult puzzle that faces computational vision scientists and engineers in this decade.

A rich palette of diverse ideas has been proposed during the past few years, especially on the problem of recognizing objects and object categories (see our brief review of the literature below). There is broad consensus of the fact that models need to capture the great diversity of forms and appearances of the objects that surround us. This means models containing hundreds, sometimes thousands, of parameters. It is common knowledge in statistics that estimating a given number of parameters requires a many-fold larger number of training examples–as a consequence, learning one object category requires a batch process involving thousands or tens of thousands of training examples [39, 125, 148, 153].

Unfortunately, it is often difficult and expensive to acquire large sets of training examples. Compounding this problem, most algorithms for learning categories require that each training exemplar be aligned (typically by hand) with a prototype. This becomes particularly problematic when fiducial points are not readily identifiable (can we find a natural alignment for images of octopus, of cappuccino machines, of bonsai trees?). This is a large, practical obstacle on the way to learning thousands of object categories. It would be far better

if we managed to find ways to train new categories with few examples.

Additionally, learning should be incremental, rather than batch. Imagine a machine placed in a new environment. It would be useful for that machine to learn a new object class, maybe only tentatively, as soon as it encounters a few exemplars, rather than waiting, perhaps in vain, for hundreds of examples to show up. Online learning of object categories has not yet been approached in the literature.

Is there any hope? We believe so. A young child learns many categories per day [10]. It seems unlikely that this would require a large set of training images for each category as well as much supervision. Informal observation also tells us that for an adult, learning a new category is both fast and easy, sometimes requiring very few training examples: given 2 or 3 images of an animal you have never seen before, later on you can usually recognize with some reliability other exemplars of the same species.

We hypothesize that, once a few categories have been learnt the hard way, some information may be abstracted from that process to make learning further categories more efficient. In other words, we should be able to make use of the knowledge that has been gained so far rather than starting from scratch each time we learn a new category. We pursue here this hypothesis in a Bayesian setting: we extract "general knowledge" from previously learnt categories and represent it in the form of a prior probability density function in the space of model parameters. Given a training set, no matter how small, we update this knowledge and produce a posterior density, which is then used for detection/recognition. Our experiments show that this is a productive approach and that indeed some useful information about categories may be obtained from a few, even one, training example.

9.2 Literature Review

In order to place our work in context we make a few observations and mention the relevant literature on object recognition.

Researchers in this area face three main challenges. Representation: how should we model objects and categories? Learning: how may we acquire such models? Detection/recognition: given a new image, how do we detect the presence of a known object/category amongst clutter, and despite occlusion, viewpoint and lighting changes? The great richness and diversity of methods and ideas in the literature indicates that these issues are far from being settled. However, there is broad consensus on a few significant points. First of all, the shape and appearance of the objects that surround us are complex and diverse, therefore models should be rich (lots of parameters, heterogeneous descriptors). Second, the appearance of objects within a given category may be highly variable, therefore models should be flexible (allow for some slop in the parameters). Third, in order to handle intra-class variability and occlusion, models should be composed of features, or

parts, which are not required to be detected in all instances; the mutual position of these parts constitutes further model information. Fourth, it is difficult, if not impossible, to model class variability using principled a-priori techniques; it is best to learn the models from training examples. Fifth, computational efficiency must be kept in mind.

Work on recognition may be divided into two groups: recognition of individual objects [42, 53, 79, 119] and recognition of categories [4, 19, 39, 72, 75, 121, 124, 125, 131, 148, 153]. Individual objects are easier to handle, therefore more progress has been made on efficient recognition [79], lighting-invariant [79, 86] and viewpoint-invariant [63, 119] representations and recognition. Classes are more general, require more complex representations, and are more difficult to learn; most work has therefore focused on modeling and learning. Viewpoint and lighting have not been treated explicitly (an exception is [152]), but rather treated as an additional source of in-class variability. With the exception of work on handwritten digits [72], researchers have only dealt with detection (a given category is present/absent) rather than recognition (recognizing one out of many possible categories).

We are interested in the problem of learning and recognition of categories (as opposed to individual objects). While the literature proposed learning methods that require batch processing of thousands of training examples, the present work focuses on the previously unexplored problem of efficient learning: how could we estimate models of categories from very few, one in the limit, training examples. Most researchers have focused on special-interest categories: human faces [125, 148], pedestrians [149], hand-written digits [72] and automobiles [39, 125]. Instead, we wish to develop techniques that apply equally well to any category that a human would readily recognize. With this objective in mind, we carried out our experiments on a large number of categories.

Another aspect that we wish to emphasize is the ability to learn with minimal supervision. We prefer to develop methods that do not rely on hand-alignment of the training examples, for the reasons mentioned in the introduction. For this reason, we use statistical models and probabilistic detection techniques developed by [19, 39, 75, 153], which will be reviewed in Chapter 10.2.

9.3 Contribution

We show in this study that by utilizing prior information of the object world, our algorithm is able to learn a completely new object category given very few training examples. This result is compared favorably to today's state-of-the-art computer vision algorithms in object recognition. We introduce an advanced machine learning method, variational Bayesian method. Our algorithm is tested on a large object category database of 101 object categories.

Chapter 10

A Bayesian Model

10.1 Overall Bayesian Framework

Let's say that we are looking for a flamingo bird in a query image that is presented to us. To decide whether there is a flamingo bird or not, we compare the probability of a flamingo being present in the image with the probability of only background clutter being present in the image. The decision is simple: if the probability for a flamingo present is higher, we decide this image contains an instance of a flamingo. If it is the other way around, we decide there is no flamingo. To compute the probability of a flamingo being present in an image, we need a model of a flamingo, which we learn from a set of training images containing examples of flamingos. Then we could compare this probability with the background model, and in turn make our final decision.

We can now translate the above events into a probabilistic framework. Let $\mathcal{I}$ be the query image, which may contain an example of the foreground class $\mathcal{O}_{fg}$, say flamingo. The alternative is that it contains background clutter belonging to a generic background class $\mathcal{O}_{bg}$. $\mathcal{I}_t$ is the set of training images that we have seen while learning the flamingo class. Now the decision of whether this query image $\mathcal{I}$ has a flamingo or not can be written in the following way:

$$R = \frac{p(\mathcal{O}_{fg}|\mathcal{I},\mathcal{I}_t)}{p(\mathcal{O}_{bg}|\mathcal{I},\mathcal{I}_t)} \tag{10.1}$$

$$= \frac{p(\mathcal{I}|\mathcal{I}_t,\mathcal{O}_{fg})\,p(\mathcal{O}_{fg})}{p(\mathcal{I}|\mathcal{I}_t,\mathcal{O}_{bg})\,p(\mathcal{O}_{bg})} \tag{10.2}$$

If R, the ratio of the class posteriors, is greater than some threshold, T, then we decide the image contains an instance of a flamingo. If it is less than T then the image does not contain a flamingo. Note that in Eq. 10.2, we use Bayes Rule to expand Eq. 10.1, giving us a ratio of likelihoods and a ratio of priors on the object classes. We can now further expand Eq. 10.2 by introducing a parametric model for the foreground

and background class, whose parameters are $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{bg}$, respectively:

$$R \propto \frac{\int p(\mathcal{I}|\boldsymbol{\theta}, \mathcal{O}_{fg}) p(\boldsymbol{\theta}|\mathcal{I}_t, \mathcal{O}_{fg})\, d\boldsymbol{\theta}}{\int p(\mathcal{I}|\boldsymbol{\theta}_{bg}, \mathcal{O}_{bg}) p(\boldsymbol{\theta}_{bg}|\mathcal{I}_t, \mathcal{O}_{bg})\, d\boldsymbol{\theta}_{bg}} \tag{10.3}$$

$$= \frac{\int p(\mathcal{I}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{I}_t, \mathcal{O}_{fg})\, d\boldsymbol{\theta}}{\int p(\mathcal{I}|\boldsymbol{\theta}_{bg}) p(\boldsymbol{\theta}_{bg}|\mathcal{I}_t, \mathcal{O}_{bg})\, d\boldsymbol{\theta}_{bg}} \tag{10.4}$$

The ratio of priors, $\frac{p(\mathcal{O}_{fg})}{p(\mathcal{O}_{bg})}$, is a constant, thus it is omitted in Eq. 10.3 since it may be incorporated in the decision threshold. In addition, we have simplified $p(\mathcal{I}|\boldsymbol{\theta}, \mathcal{O}_{fg})$ into $p(\mathcal{I}|\boldsymbol{\theta})$ for the sake of simplicity. Similarly, $p(\mathcal{I}_t|\boldsymbol{\theta}_{bg}, \mathcal{O}_{bg})$ is abbreviated to $p(\mathcal{I}_t|\boldsymbol{\theta}_{bg})$. The learning procedure involves estimating $p(\boldsymbol{\theta}|\mathcal{I}_t, \mathcal{O}_{fg})$, the distribution of model parameters given the training images. Once this is known, we can evaluate R by integrating out over $\boldsymbol{\theta}$. We now look at the particular object model used.

10.2 The Object Category Model

Our chosen representation is a Constellation model [20, 39, 153]. Given a query image, $\mathcal{I}$, we find a set of N interesting regions in the image. From these N regions, we obtain two variables: $\mathcal{X}$–the locations of the regions and $\mathcal{A}$–the appearances of the regions. Section 14.2 gives details of how $\mathcal{X}$ and $\mathcal{A}$ are obtained. It is $\mathcal{X}$ and $\mathcal{A}$ that we now model, $\mathcal{I}$ no longer being used directly. Similarly, in the case of the training images $\mathcal{I}_t$, we obtain $\mathcal{X}_t$ and $\mathcal{A}_t$. Thus Eq. 10.3 becomes:

$$R \propto \frac{\int p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta}, \mathcal{O}_{fg}) p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg})\, d\boldsymbol{\theta}}{\int p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta}_{bg}, \mathcal{O}_{bg}) p(\boldsymbol{\theta}_{bg}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{bg})\, d\boldsymbol{\theta}_{bg}} \tag{10.5}$$

$$= \frac{\int p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg})\, d\boldsymbol{\theta}}{\int p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta}_{bg}) p(\boldsymbol{\theta}_{bg}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{bg})\, d\boldsymbol{\theta}_{bg}} \tag{10.6}$$

We now examine likelihoods $p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta})$ and $p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta}_{bg})$, where in the general case, we have a mixture of constellation models, with Ω components:

$$p(\mathcal{X}, \mathcal{A}|\,\boldsymbol{\theta}) = \sum_{w=1}^{\Omega} \sum_{\mathbf{h} \in H} p(\mathcal{X}, \mathcal{A}, \mathbf{h}, w|\,\boldsymbol{\theta}) = \sum_{w=1}^{\Omega} p(w|\boldsymbol{\pi}) \sum_{\mathbf{h} \in H} \underbrace{p(\mathcal{A}|\mathbf{h}, \boldsymbol{\theta}_w^{\mathcal{A}})}_{Appearance} \underbrace{p(\mathcal{X}|\mathbf{h}, \boldsymbol{\theta}_w^{\mathcal{X}})}_{Shape} p(\mathbf{h}|\boldsymbol{\theta}_w) \tag{10.7}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\theta}^{\mathcal{A}}, \boldsymbol{\theta}^{\mathcal{X}}\}$ and $p(\mathbf{h}|\boldsymbol{\theta}_w)$ is a constant. Note that the shape, $\mathcal{X}$, and appearance, $\mathcal{A}$, are assumed to be independent. Typically, a constellation model would have P ($3 \sim 7$) diagnostic features, or parts. But there are N (up to 100) interest points, or candidate features in the image. We therefore introduce an indexing variable $\mathbf{h}$, which we call a *hypothesis*. $\mathbf{h}$ is a vector of length P, where each entry is between 1 and N, which allocates a particular feature to a model part. Any unallocated features are assumed to belong to

the background of the image. The set of all hypotheses H consists of all valid allocations of features to the parts; consequently $|H^n|$, the total number of hypotheses in image n is $O(N^P)$. For simplicity, we assume the background model is fixed and has a single parameter value, $\boldsymbol{\theta}_{bg}$, thus the integral in the denominator of Eq. 10.6 collapses to $p(\mathcal{X}, \mathcal{A}, |\boldsymbol{\theta}_{bg})$. If we believe no object to be present (the $\mathcal{O}_{bg}$ case), then only one hypothesis exists, $\mathbf{h}_0$, the null hypothesis, where all detections are assigned to be background. Hence the denominator becomes:

$$p(\mathcal{X}, \mathcal{A}, |\boldsymbol{\theta}_{bg}) = p(\mathcal{X}, \mathcal{A}, \mathbf{h}_0, |\boldsymbol{\theta}_{bg}) = p(\mathcal{A}|\mathbf{h}_0, \boldsymbol{\theta}^{\mathcal{A}}_{bg}) p(\mathcal{X}|\mathbf{h}_0, \boldsymbol{\theta}^{\mathcal{X}}_{bg}) p(\mathbf{h}_0|\boldsymbol{\theta}_{bg}) \tag{10.8}$$

Since this expression is constant for given $\mathcal{X}$ and $\mathcal{A}$, we can use it to cancel terms in the numerator of Eq. 10.6.

The model encompasses the important properties of an object: shape and appearance, both in a probabilistic way. This allows the model to represent both geometrically constrained objects (where the shape density would have a small covariance, e.g., a face) and objects with distinctive appearance but lacking geometric form (the appearance densities would be tight, but the shape density would now be looser, e.g., an animal principally defined by its texture such as a zebra). Note, that in the model the following assumptions are made: shape is independent of appearance; for shape the joint covariance of the parts' position is modeled, whilst for appearance each part is modeled independently. In the experiments reported here we use a slightly simplified version of the model presented in [39] by removing the terms involving occlusion and statistics of the feature finder, since these are relatively unimportant when we only have a few images to train from.

10.2.0.1 Appearance

Each feature's appearance is represented as a point in some appearance space, defined in Chapter 11.1.1. For a given mixture component, each part p has a Gaussian density within this space, with mean and precision parameters $\boldsymbol{\theta}^{\mathcal{A}}_{p,w} = \{\boldsymbol{\mu}^{\mathcal{A}}_{p,w}, \boldsymbol{\Gamma}^{\mathcal{A}}_{p,\omega}\}$ that are independent of other parts' densities. The background model has the same form, with fixed parameters $\boldsymbol{\theta}^{\mathcal{A}}_{bg} = \{\boldsymbol{\mu}^{\mathcal{A}}_{bg}, \boldsymbol{\Gamma}^{\mathcal{A}}_{bg}\}$. Note that $\boldsymbol{\Gamma}^{\mathcal{A}}_{p,w}$ and $\boldsymbol{\Gamma}^{\mathcal{A}}_{bg}$ are diagonal matrices. Each feature selected by the hypothesis is evaluated under the appropriate part density with features not selected being evaluated under the background model:

$$p(\mathcal{A}|\mathbf{h}, \boldsymbol{\theta}^{\mathcal{A}}_{w}) = \prod_{p=1}^{P} \mathcal{G}(\mathcal{A}(\boldsymbol{h}_p)|\boldsymbol{\mu}^{\mathcal{A}}_{p,w}, \boldsymbol{\Gamma}^{\mathcal{A}}_{p,w}) \prod_{j=1,\, j \setminus \boldsymbol{h}}^{N} \mathcal{G}(\mathcal{A}(j)|\boldsymbol{\mu}^{\mathcal{A}}_{bg}, \boldsymbol{\Gamma}^{\mathcal{A}}_{bg}) \tag{10.9}$$

where $\mathcal{G}$ is the Gaussian distribution and j represents features not assigned to a part in hypothesis $\boldsymbol{h}$. If no object is present then all features are modeled by the background:

$$p(\mathcal{A}|\mathbf{h}_0, \boldsymbol{\theta}_{bg}^{\mathcal{A}}) = \prod_{j=1}^{N} \mathcal{G}(\mathcal{A}(j)|\boldsymbol{\mu}_{bg}^{\mathcal{A}}, \boldsymbol{\Gamma}_{bg}^{\mathcal{A}}) \tag{10.10}$$

Note that $p(\mathcal{A}|\mathbf{h}_0, \boldsymbol{\theta}_{bg})$ is a constant for a given image, therefore it can be brought inside the integral and summation over all hypotheses in Eq. 10.6 and 10.7. This cancels with all other background hypotheses except the true foreground hypothesis $\boldsymbol{h}$ in Eq. 10.9:

$$\frac{p(\mathcal{A}|\mathbf{h}, \boldsymbol{\theta}_{w}^{\mathcal{A}})}{p(\mathcal{A}|\mathbf{h}_0, \boldsymbol{\theta}_{bg}^{\mathcal{A}})} = \prod_{p=1}^{P} \frac{\mathcal{G}(\mathcal{A}(\boldsymbol{h}_p)|\boldsymbol{\mu}_{p,w}^{\mathcal{A}}, \boldsymbol{\Gamma}_{p,w}^{\mathcal{A}})}{\mathcal{G}(\mathcal{A}(\boldsymbol{h}_p)|\boldsymbol{\mu}_{bg}^{\mathcal{A}}, \boldsymbol{\Gamma}_{bg}^{\mathcal{A}})} \tag{10.11}$$

10.2.0.2 Shape

The shape of each constellation model component is represented by a joint Gaussian density of the locations of features within a hypothesis, after they have been transformed into a scale and translation-invariant space. Translation invariance is achieved by using the leftmost part as a landmark and modeling all parts relative to it. Scale invariance is obtained by taking the scale of the landmark feature and using it to normalize the relative locations of the other parts. We assume uniform densities α^{-1} for the position of the object, where α is the image area. The relative location of the parts is modeled by a $2(P-1)$ dimensional Gaussian, with a uniform background model for unallocated features:

$$p(\mathcal{X}|\mathbf{h}, \boldsymbol{\theta}_{w}^{\mathcal{X}}) = \alpha^{-1}\, \mathcal{G}(\mathcal{X}(\mathbf{h})|\boldsymbol{\mu}_{w}^{\mathcal{X}}, \boldsymbol{\Gamma}_{w}^{\mathcal{X}})\ \alpha^{-(N-P)} \tag{10.12}$$

where $\boldsymbol{\theta}_{w}^{\mathcal{X}} = \{\alpha, \boldsymbol{\mu}_{w}^{\mathcal{X}}, \boldsymbol{\Gamma}_{w}^{\mathcal{X}}\}$. For the null hypothesis, $p(\mathcal{X}|\mathbf{h}_0, \boldsymbol{\theta}_{bg}^{\mathcal{X}}) = \alpha^{-N}$, which is also a constant, so we cancel with all other background hypotheses except the true foreground hypothesis $\boldsymbol{h}$ in Eq. 10.12:

$$\frac{p(\mathcal{X}|\mathbf{h}, \boldsymbol{\theta}_{w}^{\mathcal{X}})}{p(\mathcal{X}|\mathbf{h}_0, \boldsymbol{\theta}_{bg}^{\mathcal{X}})} = \alpha^{P-1}\, \mathcal{G}(\mathcal{X}(\mathbf{h})|\boldsymbol{\mu}_{w}^{\mathcal{X}}, \boldsymbol{\Gamma}_{w}^{\mathcal{X}}) \tag{10.13}$$

Additionally, to reduce the number of hypotheses that must be considered in each frame, we impose an ordering constraint on each hypothesis' shape, such that the x-coordinate of each part much be monotonically increasing. This reduces the number of hypotheses that must be considered by $P!$ and provides a useful constraint in the learning process.

10.2.1 Discussion of model

We make some comments concerning the model:

1. $\mathcal{X}(\mathbf{h}) \in \mathbb{R}^{2P-2}$ and $\mathcal{A}(\mathbf{h}) \in \mathbb{R}^{kP}$ thus for $k = 10$, $P = 4$, the shape term has $6+21 = 27$ (mean + full covariance matrix) parameters. The appearance term has $40 + 40 = 80$ (mean + diagonal covariance matrix) parameters, thus the model has $27 + 80 = 107$ parameters in total.

2. The total number of hyperparameters for $k = 10$, $P = 4$ is 109, since $\boldsymbol{m}$ and $\boldsymbol{B}$ are the same dimensionality as $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}$ but additionally β and a (both real numbers) exist for both shape and appearance terms: $107 + 2 = 109$.

3. The constellation model is a generative model of the output of an interest region detector, not the image pixels. Hence the performance of the model is dependent on the performance of the detectors themselves. See Chapter **??** for an investigation into this dependency.

4. In our representation, there is nothing to prevent patches from overlapping that could lead to overcounting of the evidence for the model. However, given relatively low number of features per image, this should not be a major problem.

5. The shape model presented above uses a joint density over all parts, thus the data association problem has complexity $O(N^P)$. While this is the most thorough approach to modelling the location of parts, it presents a major computational bottleneck. Imposing conditional independence by the use of a tree-structured model would reduce the complexity to $O(N^2P)$ in learning and $O(NP)$ in recognition [37, 41]. However, in doing so, other issues arise such as how the optimal graph structure should be chosen. Since these issues are in themselves complex and are outside the focus of this paper, for the sake of simplicity, we stick with the complete representation, despite its drawbacks.

6. Our model and representation of shape is suited to compact objects that do not have large amounts of articulation (e.g., human bodies). For such categories, different graph structures and coordinate frames (i.e., the angles between parts) may be more appropriate.

7. Our feature representation is currently confined to textured image patches. Alternative representations, such as curve contours, which model the outline of the object could also be used with little modification to the underlying model [38, 40]. This would allow the model to handle categories where the outline of the object is more important than its interior (e.g., bottles).

8. Currently the background model is very simple: a uniform shape distribution and a single Gaussian distribution for appearance. Their crude nature is a consequence of the requirement, for efficiency,

that the denominator in Eq. 10.5 must be able to cancel with the numerator, making evaluation of the likelihood ratio simple. The parametric assumptions of the background model were tested by examining the distribution of thousands of detections from an assorted collection of images. Our observation was that these assumptions were reasonably accurate.

9. The framework describes object detection (i.e., object present or absent), however it can easily be extended to localization by using the best hypothesis in each image (e.g., by taking a bounding box around it). Multiple instances per image can also be found by a greedy approach: finding the best hypothesis; summing over all hypotheses around its neighborhood to give a value of R for a sub-window of the image; and removing all features within the sub-window and repeating until no sub-windows with R greater than a given threshold can be found.

10. Our model is formulated as a mixture of Gaussians (Eq.10.7). In practice, we use a single mixture component in this paper for all of the experiments. Weber et al. have demonstrated that by increasing the number of mixture components, the model is capable of representing different aspects of the object due to pose variations [152].

10.2.2 Form of the Parameter Posterior

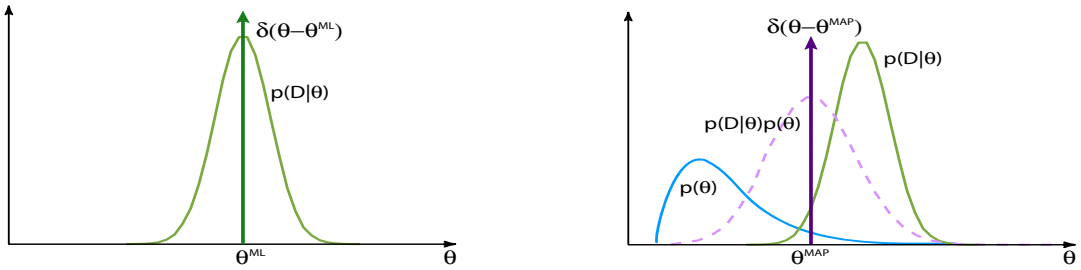
In computing R, we must evaluate the integral $\int p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O})\, d\boldsymbol{\theta}$. In Chapter 10.2 the form of $p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta})$ was considered. We now look at the posterior of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O})$. Before we consider how this density might be estimated, its form must be decided upon. Since the integral above is typically impossible to solve analytically, we look at various forms of $p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O})$ that approximate the true density whilst making the integral tractable.

10.2.3 Maximum Likelihood (ML) and Maximum A Posteriori (MAP)

If we assume that the model distribution $p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O})$ is highly peaked, we could approximate it with a δ function at $\boldsymbol{\theta}^*$: $\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. This allows the integral in Eq. 10.6 to collapse to $p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta}^*)$, whose functional form is given by Eq. 10.7.

There are two ways of obtaining $\boldsymbol{\theta}^*$, illustrated in Fig. 10.1. The simplest one is Maximum Likelihood (ML) estimation [39, 153]. Here $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{\mathrm{ML}}$ is computed by picking the $\boldsymbol{\theta}$ that gives rise to the highest likelihood value of the training data:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{\mathrm{ML}} = \underset{\theta}{argmax}\, p(\mathcal{X}_t, \mathcal{A}_t|\boldsymbol{\theta}) \tag{10.14}$$



(a) Maximum Likelihood (ML) (b) Maximum a posteriori (MAP)

Figure 10.1: Schematic comparison of ML and MAP methods.

If we had some prior knowledge about $\boldsymbol{\theta}$, we could also use this information to help estimate $\boldsymbol{\theta}^*$. The idea is to weigh the likelihood of training examples at $\boldsymbol{\theta}$ by the prior probability of $\boldsymbol{\theta}$ at that point. This is called the Maximum *a posteriori* (MAP) estimation.

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{\text{MAP}} = \underset{\theta}{argmax}\, p(\mathcal{X}_t, \mathcal{A}_t|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{10.15}$$

The form of $p(\boldsymbol{\theta})$ needs to be chosen carefully to ensure that the estimation procedure is efficient. In Appendix A.3, we revisit this equation and give a more detailed account of $p(\boldsymbol{\theta})$ and methods for estimating $\boldsymbol{\theta}^{\text{MAP}}$.

Both ML and MAP assume a very well peaked $p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O})$ so that $\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ is a suitable estimate of the entire distribution. But when there is a very limited number of training examples, the distribution may not be well peaked, in which case both ML and MAP are likely to yield poor models.

10.2.3.1 Other Inference Methods

Sampling methods. At the other extreme, we can use numerical methods such as Gibbs Sampling [47] or Markov-Chain Monte-Carlo (MCMC) [48] to give an accurate estimate of the integral in Eq. 10.6, but these can be computationally very expensive. In the constellation model, the dimensionality of $\boldsymbol{\theta}$ is large (~ 100) for a reasonable number of parts, making MCMC methods impractical for our problem. Additionally, the use of sampling-based methods is something of an art: issues such as what sampling regime to use have no simple answer. Hence they are less attractive as compared with methods giving a distinct solution.

Recursive Approximations A variety of variational approximations exist that are recursive in nature [62]. In such schemes, the data points are processed sequentially with the (approximate) marginal posterior $p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O})$ being updated after each new data point. The major drawback to using them is that the final solution is dependent on the ordering of the data points. In our problem the data has no obvious ordering, hence such methods would complicate the learning procedure so we choose not to adopt them [33].

10.2.4 Conjugate Densities

The final approach is to assume that $p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg})$ has a specific parametric form, such that the integral in Eq. 10.5 has a closed-form solution. Recalling the numerator of Eq. 10.5:

$$\int p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg})\, d\boldsymbol{\theta} \tag{10.16}$$

Our goal is to find a parametric form of $p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg})$ such that the learning of $p(\boldsymbol{\theta})$ is feasible and the evaluation of Eq.10.16 is tractable. This could be achieved by taking advantage of a class of prior distributions that are conjugate to their posterior distributions. In other words, a conjugate prior for a given probabilistic model is one for which the resulting posterior has the same functional form as the prior. In the case of $p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg})$, we use a Normal-Wishart distribution as its conjugate prior. Given that $p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta})$ was chosen to be a product of Gaussians (in Chapter 10.2), the entire integral of Eq.10.16 becomes a multivariate Student's T distribution. Efficient learning schemes exist for estimating the hyper-parameters of the Normal-Wishart distribution [5], having the same computational complexity as standard ML methods. These are introduced in Chapter 10.4.

10.3 Recognition Using a Conjugate Density Parameter Posterior

Having specified a functional form for the parameter posterior, we now give the actual equations for use in recognition.

10.3.1 Parameter Distribution

Recall the mixture of constellation models from Eq. 10.3:

$$p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta}) = \sum_{\omega=1}^{\Omega} p(\omega|\boldsymbol{\pi}) \sum_{h=1}^{|H|} p(\mathcal{X}_h|\boldsymbol{\mu}_\omega^{\mathcal{X}}, \boldsymbol{\Gamma}_\omega^{\mathcal{X}}) p(\mathcal{A}_h|\boldsymbol{\mu}_\omega^{\mathcal{A}}, \boldsymbol{\Gamma}_\omega^{\mathcal{A}}) \tag{10.17}$$

Each component ω has a mixing coefficient π_ω; a mean of shape and appearance $\boldsymbol{\mu}_\omega^{\mathcal{X}}, \boldsymbol{\mu}_\omega^{\mathcal{A}}$; and a precision matrix of shape and appearance $\boldsymbol{\Gamma}_\omega^{\mathcal{X}}, \boldsymbol{\Gamma}_\omega^{\mathcal{A}}$. The $\mathcal{X}$ and $\mathcal{A}$ superscripts denote shape and appearance terms, respectively. Collecting all mixture components and their corresponding parameters together, we obtain an overall parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}^{\mathcal{X}}, \boldsymbol{\mu}^{\mathcal{A}}, \boldsymbol{\Gamma}^{\mathcal{X}}, \boldsymbol{\Gamma}^{\mathcal{A}}\}$. Assuming we have now learnt the model distribution $p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t)$ from a set of training data $\mathcal{X}_t$ and $\mathcal{A}_t$, we define the model distribution in the following way:

$$p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t) = p(\boldsymbol{\pi}) \prod_{\omega} p(\boldsymbol{\mu}_\omega^{\mathcal{X}}|\boldsymbol{\Gamma}_\omega^{\mathcal{X}}) p(\boldsymbol{\Gamma}_\omega^{\mathcal{X}}) p(\boldsymbol{\mu}_\omega^{\mathcal{A}}|\boldsymbol{\Gamma}_\omega^{\mathcal{A}}) p(\boldsymbol{\Gamma}_\omega^{\mathcal{A}}) \tag{10.18}$$

where the mixing component is a symmetric Dirichlet: $p(\boldsymbol{\pi}) = \mathcal{D}ir(\lambda_\omega \boldsymbol{I}_\Omega)$, the distribution over the shape precisions is a Wishart $p(\boldsymbol{\Gamma}_\omega^{\mathcal{X}}) = \mathcal{W}(\boldsymbol{\Gamma}_\omega^{\mathcal{X}} | a_\omega^{\mathcal{X}}, \boldsymbol{B}_\omega^{\mathcal{X}})$ and the distribution over the shape mean conditioned on the precision matrix is Normal: $p(\boldsymbol{\mu}_\omega^{\mathcal{X}} | \boldsymbol{\Gamma}_\omega^{\mathcal{X}}) = \mathcal{G}(\boldsymbol{\mu}_\omega^{\mathcal{X}} | \boldsymbol{m}_\omega^{\mathcal{X}}, \beta_\omega^{\mathcal{X}} \boldsymbol{\Gamma}_\omega^{\mathcal{X}})$. Together the shape distribution $p(\boldsymbol{\mu}_\omega^{\mathcal{X}}, \boldsymbol{\Gamma}_\omega^{\mathcal{X}})$ is a Normal-Wishart density [5, 102]. Note $\{\lambda_\omega, a_\omega, \boldsymbol{B}_\omega, \boldsymbol{m}_\omega, \beta_\omega\}$ are hyper-parameters for defining their corresponding distributions of model parameters. Identical expressions apply to the appearance component in Eq. 10.18.

10.3.2 Closed-form Calculation of R

Recall that:

$$R = \frac{p(\mathcal{X}, \mathcal{A} | \mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg})}{p(\mathcal{X}, \mathcal{A} | \mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{bg})} = \frac{\int p(\mathcal{X}, \mathcal{A} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg}) \, d\boldsymbol{\theta}}{\int p(\mathcal{X}, \mathcal{A} | \boldsymbol{\theta}_{bg}) p(\boldsymbol{\theta}_{bg} | \mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{bg}) \, d\boldsymbol{\theta}_{bg}} \tag{10.19}$$

Due to the use of conjugate densities, the integral in the numerator becomes a multi-modal multivariate Student's T distribution (denoted by $\mathcal{S}$):

$$p(\mathcal{X}, \mathcal{A} | \mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg}) = \sum_{\omega=1}^{\Omega} \sum_{h=1}^{|H|} \tilde{\pi}_\omega \, \mathcal{S}(\mathcal{X}_h |\, g_\omega^{\mathcal{X}}, \mathbf{m}_\omega^{\mathcal{X}}, \boldsymbol{\Lambda}_\omega^{\mathcal{X}}) \, \mathcal{S}(\mathcal{A}_h |\, g_\omega^{\mathcal{A}}, \mathbf{m}_\omega^{\mathcal{A}}, \boldsymbol{\Lambda}_\omega^{\mathcal{A}}) \tag{10.20}$$

$$\text{where} \quad g_\omega = a_\omega + 1 - d \quad \text{and} \quad \boldsymbol{\Lambda}_\omega = \frac{\beta_\omega + 1}{\beta_\omega g_\omega} \boldsymbol{B}_\omega \quad \text{and} \quad \tilde{\pi}_\omega = \frac{\lambda_\omega}{\sum_{\omega'} \lambda_{\omega'}}$$

Note d is the dimensionality of the parameter vector $\boldsymbol{\theta}$. The denominator of Eq. 10.19 is a constant, since we only consider a single value of $\boldsymbol{\theta}_{bg}$: $\boldsymbol{\theta}_{bg}^{ML}$ i.e. $p(\boldsymbol{\theta}_{bg} | \mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{bg}) = \delta(\boldsymbol{\theta}_{bg} - \boldsymbol{\theta}_{bg}^{ML})$.

10.4 Learning Using a Conjugate Density Parameter Posterior

The process of learning an object category is unsupervised [39, 153]. The algorithm is presented with a number of training images labeled as "foreground images." It assumes there is an instance of the object category to be learnt in each image. But no other information, e.g., location, size, shape, appearance, etc., is provided. The algorithm first detects interesting features in these training images, and then estimates the parameters of the densities from these regions. Since the model is linear and Gaussian with conjugate priors it should have a closed-form solution. However, the discrete indexing variable $\mathbf{h}$, representing the assignment of features to parts prevents such a solution. Instead an iterative variational method that resembles the Expectation Maximization (EM) algorithm [23] is used to estimate the variational posterior. Afterwards recognition is performed on a query image by repeating the process of detecting regions and then evaluating the regions, using the model parameters estimated in the learning process.

The goal of learning is to obtain a posterior distribution $p(\boldsymbol{\theta} | \mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg})$ of the model parameters given

a set of training data $\{\mathcal{X}_t, \mathcal{A}_t\}$ as well as some prior information. We formulate this learning problem using Variational Bayesian Expectation Maximization (VBEM), applied to a multi-dimensional Gaussian mixture model as introduced by Attias [5]. Detailed derivations of VBEM are given in Appendix A.2. In addition, we also give a detailed derivation of the MAP parameter estimation in Appendix A.3.

Chapter 11

Experiments and Results

11.1 Implementation

11.1.1 Feature detection and representation

We use the same features as in [39]. They are found using the detector of Kadir and Brady [66]. This method finds regions that are salient over both location and scale. Gray-scale images are used as the input. The most salient regions are clustered over location and scale to give a reasonable number of features per image, each with an associated scale. The coordinates of the center of each feature give us $\mathcal{X}$. Fig. 11.1 illustrates this on images from four datasets. Once the regions are identified, they are cropped from the image and rescaled to the size of a small (11×11) pixel patch. Each patch exists in a 121 dimensional space. We then reduce this dimensionality by using PCA. A fixed PCA basis, pre-calculated from the background datasets, is used for this task, which gives us the first 10 principal components from each patch. The principal components from all patches and images form $\mathcal{A}$.

11.1.2 Learning

Practical aspects of the Bayesian One-Shot learning procedure are now discussed, including: the choice of the prior density, $p(\boldsymbol{\theta})$ and details of the Bayesian One-Shot implementation.

11.1.2.1 Choice of Prior

One critical issue is the choice of priors for the Dirichlet and Norm-Wishart distributions. In this paper, learning is performed using a single mixture component, i.e., $\Omega = 1$. So λ is set to 1, since π_ω will always be 1. Ideally, the values for the shape and appearance priors should reflect object models in the real world. In other words, if we have already learnt a sufficient number of classes of objects (e.g., hundreds or thousands),

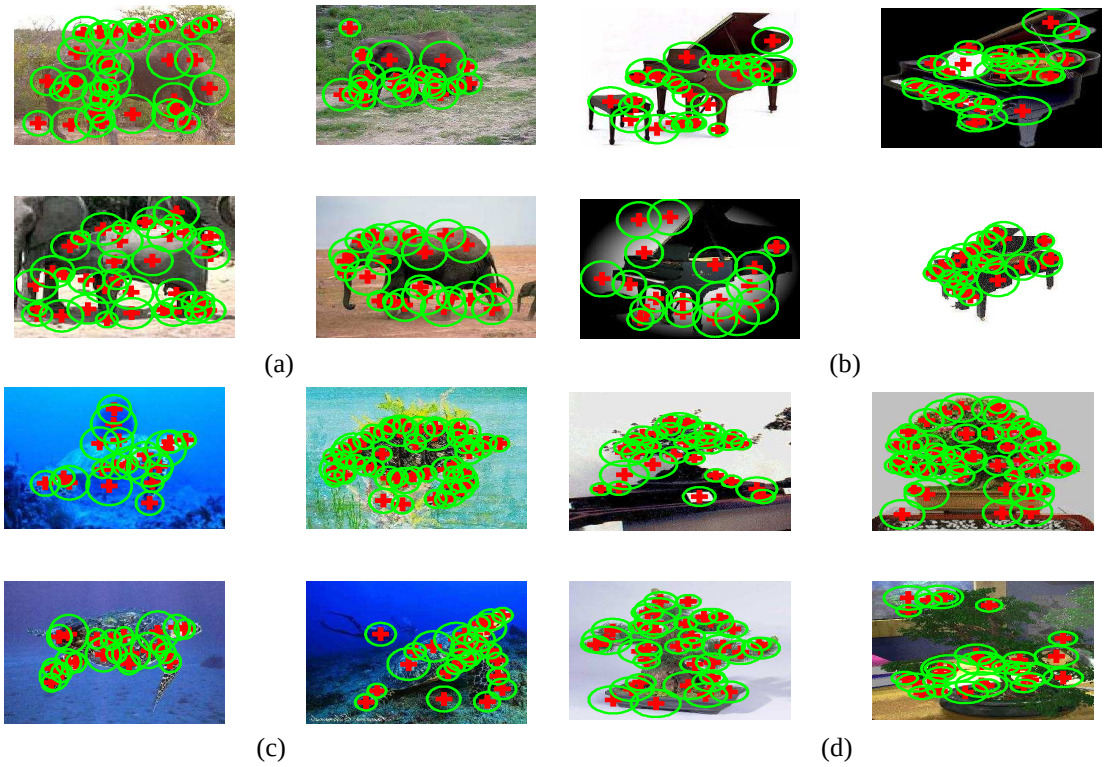


Figure 11.1: Output of the feature detector on samples images from four categories. (a) Elephant, (b) Grand piano, (c) Hawksbill, (d) Bonsai tree.

we would have a pretty good idea of the average shape (appearance) mean and variances given a new object category. In reality we do not have the luxury of such a number of object classes. We use four classes of object models learnt in a ML manner from [39] to form our priors. They are: spotted cats, motorbikes, faces and airplanes. Since we wish to learn the same four datasets with our algorithm, we use a "leave one out" strategy. For example, when learning motorbikes we obtain priors by averaging the learnt model parameters from the other three categories (i.e., spotted cats, faces and airplanes), hence avoiding the incorporation of an existing motorbike model. The hyper-parameters of the prior are then estimated from the parameters of the existing category models. An example of this process is given in Chapter 11.2.3.

11.1.2.2 Details of Bayesian One-Shot algorithm

- Initial conditions are chosen in the following way. Shape and appearance means are set to the means of the training data itself. Covariances are chosen randomly within a sensible range. Namely, they are initialized to be roughly in the order of the average dimensions of the training images.

- Learning is halted when the largest parameter change per iteration (across all parameters) falls below a

certain threshold (10^{-4}) or the maximum number of iterations is exceeded (typically 500). In general, convergence occurs within less than 100 iterations.

- Since the model is a generative one, the background images are not used in learning except for one instance: the appearance model has a distribution in appearance space modeling background features. Estimating this from foreground data proved inaccurate so the parameters are estimated from a set of background images and not updated within the Bayesian One-Shot iteration.

- Learning a class takes roughly less than a minute on a 2.8 GHz machine when the number of training images is less than 10 and the model is composed of 4 parts. The algorithm is implemented in Matlab. It is also worth mentioning that the current algorithm does not utilize any efficient search method, unlike [39]. It has been shown that increasing the number of parts in a constellation model results in greater recognition power provided enough training examples are given [39]. Were efficient search techniques used, 6-7 parts could be learnt, since the Bayesian One-Shot update equations require the same amount of computation as the traditional ML ones. However, all our experiments currently use 4 part models for both the current algorithm and ML.

11.2 Experimental Results

11.2.1 Datasets

In the first set of experiments, the same four object categories as in [32,39] were used[1], namely: human faces, motorbikes, airplanes and spotted cats. These datasets contain a fair amount of background clutter and scale variation, although each category is presented from a consistent viewpoint.

In addition, two naive subjects collected another dataset of 101 object categories for the second set of experiments. The names of 101 categories were generated by flipping through the pages of the Webster Collegiate Dictionary [1] and picking a subset of categories that were associated with a drawing. Using a script, all images returned by Google Image Search engine for each category name were downloaded. The two subjects then sorted through the images for each category, getting rid of irrelevant images (e.g., a zebra-patterned shirt for the "zebra" category). Fig. 11.2 shows examples from 101 foreground object categories as well as the background clutter category (obtained by typing "things" into Google).

Minimal preprocessing was performed on the categories. Categories such as motorbike, airplane, cannon, etc. where two mirror image views were present, were manually flipped, so all instances faced in the same direction. Additionally, categories with a predominantly vertical structure were rotated to an arbitrary angle.

[1] Available from www.vision.caltech.edu

This is due to the convention that the leftmost part of each hypothesis is used as a reference point to translate the rest of the parts (see Section 10.2.0.2). With vertically orientated structures, the horizontal ordering of the features will be somewhat arbitrary, thereby artificially giving a large vertical variability.

11.2.2 Experimental Setup

Each experiment is carried out in the following way. Each dataset is randomly split into two disjoint sets of equal size. N training images are drawn randomly from the first. A fixed set of 50 are selected from the second, which form the test set. We then learn models using both Variational Bayesian and ML approaches and evaluate their performance on the test set. For evaluation purposes, we also use 50 images from a background dataset of assorted junk images from the Internet. For each category, we vary N from 1 to 6, repeating the experiments 10 times for each value (using a different set of N training images each time) to obtain a more robust estimate of performance. When $N = 1$, ML fails to converge, so we only show results for the Bayesian One-Shot algorithm in this case.

When evaluating the models, the task is a binary decision–object present or absent. All performance values are quoted as equal error rates from the receiver-operating characteristic (ROC) (i.e., p(True positive) = 1 - p(False alarm)). The ROC curve is obtained by testing the model on 50 foreground test images and 50 background images . For example, a value of 85% means that 85% of the foreground images are correctly classified but 15% of the background images are incorrectly classified (i.e., false alarms). In all the experiments, the following parameters are used: number of parts in model $= 4$; number of PCA dimensions for each part appearance $= 10$; and average number of detections of interest point for each image $= 20$. It is also important to point out that except for the different priors, as described in section 11.1.2.1, all parameters remain the same for learning all different categories. In other words, exactly the same piece of software was used in all experiments.

11.2.3 Walkthrough for the Motorbike Category

We now go through the experimental procedure step-by-step for the motorbike category. 6 training images are selected, as shown in Fig. 11.3. The Kadir interest operator is applied to them, giving $\mathcal{X}_t$. Each of these regions is then transformed into the fixed PCA basis, to give $\mathcal{A}_t$. Next we consider the prior we will use in learning. This has been constructed from models trained using ML from the three other datasets: spotted cats, faces and airplanes. 10 ML models were trained for each category, giving a total of 30 models, each being a point in θ-space. The parameters of the prior, $\{\boldsymbol{m}_0, \beta_0, a_0, \boldsymbol{B}_0\}$ for both the shape and appearance components of the model, are then directly computed from these points in the following manner:

Figure 11.2: The 101 object categories and the background clutter category. Each category contains between 45 and 400 images. Two randomly chosen samples are shown for each category. The categories were selected prior to the experiments and the images were collected by operators not associated with the experiment. The last row shows examples from the background dataset. This dataset is obtained by collecting images through the Google image search engine (www.google.com). The keyword "things" is used to obtain hundreds of random images. Note only gray-scale information is used in our system. Complete datasets can be found at http://vision.caltech.edu/feifeili/101_ObjectCategories.

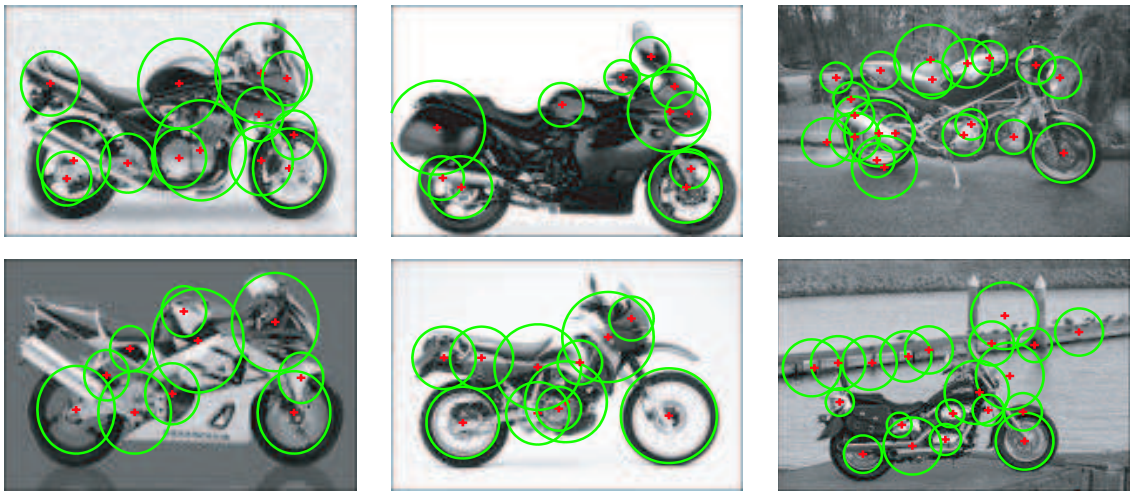
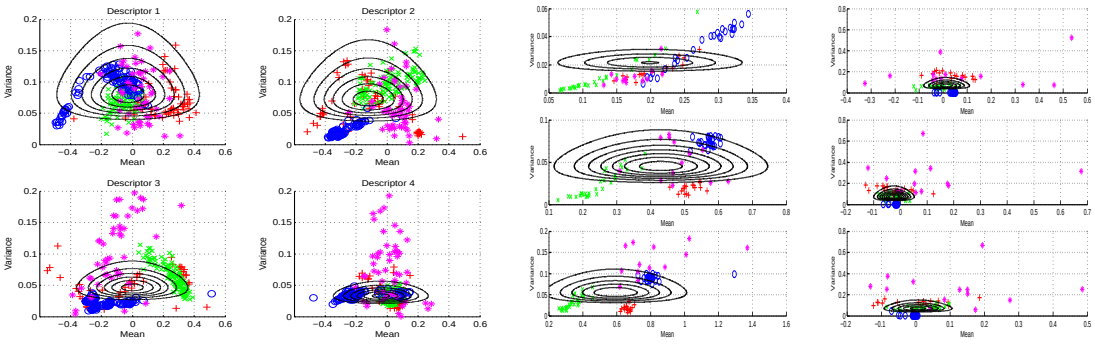


Figure 11.3: The training images for the motorbike category, with the output of the feature detector overlaid.

- $\boldsymbol{m}_0$ is estimated by computing the mean of $\boldsymbol{\mu}^{ML}$ over the $M = 30$ ML models: $\boldsymbol{m}_0 = \frac{1}{M}\sum_m \boldsymbol{\mu}_m^{ML}$.
- a_0 is fixed to be number of degress of freedom in the precision matrix $\boldsymbol{\Gamma}^{ML}$, which differs between the shape and appearance terms. For shape, $a_0^{\mathcal{X}} = 2(P-1)(P-2)$, while $a_0^{\mathcal{A}} = kP$.
- $\boldsymbol{B}_0$ is estimated by letting $a_0\boldsymbol{B}_0^{-1}$, the mean of the precision be $\frac{1}{M}\sum_m \boldsymbol{\Gamma}^{ML}$, and using the previously calculated value of a_0 to give $\boldsymbol{B}_0$.
- β_0 is estimated as the ratio between the precision of the mean and the mean of the precision: $\beta_0 = \frac{\|1/M\sum_m(\boldsymbol{\mu}_m^{ML}-\boldsymbol{m}_0)^2\|}{\|a_0\boldsymbol{B}_0^{-1}\|}$.

Fig. 11.4 illustrates shows both the ML models (as points colored by category) and the prior density fitted to them. Since the parameter space is high dimensional it is difficult to visualize but by considering each appearance descriptor separately, the mean and variance of the part from each model can be plotted in 2-D. Note the all parts use the same prior density for appearance. For shape, the mean and variance of location of each part relative to the landmark part is shown. To understand how the prior assists in learning, models were trained on background data alone and their parameters also plotted in Fig. 11.4 (as magenta $*$'s). Note that the prior density was estimated only from the ML category models, not these background models. However, they serve to illustrate the point that models that do not correspond to visual consistency occupy a different part of parameter space to models trained on images with a consistent visual appearance. The prior captures this knowledge so in the learning process it biases $p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg})$ to areas of θ-space corresponding to models of visual consistency.



(a) (b) (c)

Figure 11.4: A visualization of the prior parameter density, estimated from ML models of spotted cats (green $\times$'s), face (red $+$'s) and airplanes (blue $\circ$'s). Models trained on background data are shown as magenta $*$'s but are not used in estimating the prior density. In all figures the mean is plotted on the x-axis and the variance on the y-axis. **(a)** Appearance parameter space for the first 4 descriptors. **(b)** X component of the shape term for each of the non-landmark model parts. **(c)** Y component of shape.

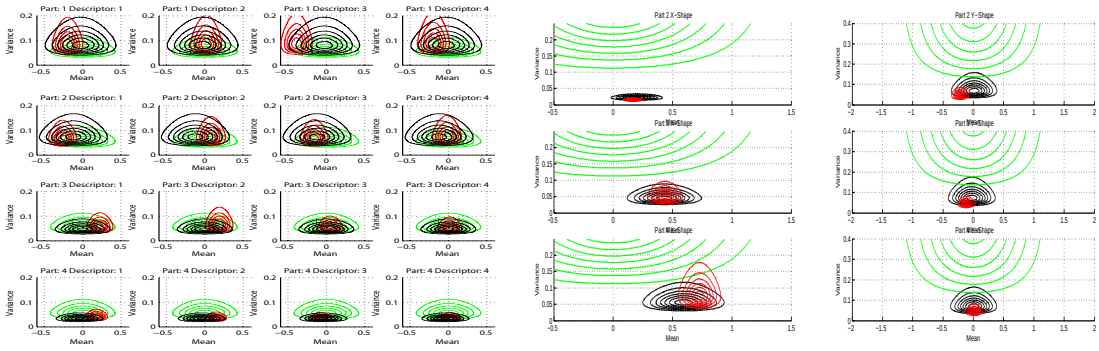
Now that the prior and training data $\mathcal{X}_t$ and $\mathcal{A}_t$ have been obtained, we commence the learning process described in Chapter 10.4. We only use one mixture component, so $\Omega = 1$. The initial values of the hyperparameters $\{\lambda_\omega, a_\omega, \boldsymbol{B}_\omega, \boldsymbol{m}_\omega, \beta_\omega\}$ are initialized as in Table 11.1. Note that since we only have one component, we do not need to worry about setting λ.

Hyperparameter	Shape	Appearance
$\boldsymbol{M}$	$\frac{1}{I}\sum_i \frac{1}{\lvert H\rvert}\sum_h \mathcal{X}(h)$	$\frac{1}{I}\sum_i \frac{1}{\lvert H\rvert}\sum_h \mathcal{A}(h)$
β	β_0	β_0
a	$2(P-1)(P-2)$	kP
$\boldsymbol{B}$	$0.1\boldsymbol{I}_{2(P-1)}$	$0.1\boldsymbol{I}_{kP}$

Table 11.1: Initial values of the hyperparameters of the parameter posterior for shape and appearance terms.

The initial posterior densities are illustrated in green in Fig. 11.5. Then we run Bayesian One-Shot until convergence is reached. Fig. 11.5 shows the final parameter densities after learning in red. They can be seen to be much tighter then the initial density, often lying close the prior density, which is likely to exert a large influence with so few training images. The model corresponding to the mean of the parameter density is shown in Fig. 11.6.

In the recognition phase, the learnt model is applied to 50 images containing motorbikes and 50 images of scenes not containing motorbikes. For each image in both sets, the likelihood ratio R is computed (using Eq. 10.19 and Eq. 10.21), giving an ROC-curve measuring the detection performance of the model. Fig. 11.6 shows the ROC curve for the model, along with sample images when the threshold, T, is set so as to give equal numbers of false alarms and missed detections.



(a) (b) (c)

Figure 11.5: The learning process. **(a)** Appearance parameter space, showing the mean and variance distributions for each of the models' 4 parts for the first 4 descriptors. The parameter densities are colored as follows: Black for the prior; green for the initial posterior density and red for the density after 30 iterations of Bayesian One-Shot, when convergence is reached. **(b)** X component of the shape term for each of the model parts. **(c)** Y component of shape. Note that in both **(b)** and **(c)**, only the variance terms along the diagonal are visualized - not the covariance terms.

In the recognition phase, the learnt model is applied to 50 images containing motorbikes and 50 images of scenes not containing motorbikes. For each image in both sets, the likelihood ratio R is computed (using Eq. **??**), giving an ROC-curve measuring the detection performance of the model. Fig. 11.6 shows the ROC curve for the model, along with sample images when the threshold, T, is set so as to give equal numbers of false alarms and missed detections.

11.2.4 Caltech 4 Dataset

We first tested our algorithm on the four object categories used by Weber et al. [153] and Fergus et al. [39]. They are faces, motorbikes, airplanes and spotted cats. Our experiments demonstrate the benefit of using prior information as well as using a full Bayesian computation in learning new object categories (Figs. 11.7-11.10). Note that in Figs. 11.7-11.10, given 0 training images, the detection rate for each category is at chance level 50%. This tells us that given only the prior model, it is not sufficient to capture characteristic information of the particular categories we are interested in. Only by incorporating this prior knowledge into the training data, is the algorithm capable of learning a sensible model with only 1 training example. For instance, in Fig. 11.7(c), we see that the 4-part model has captured the essence of a face (e.g., eyes and nose). In this case it achieves an average detection rate of 82%, given only 1 training example.

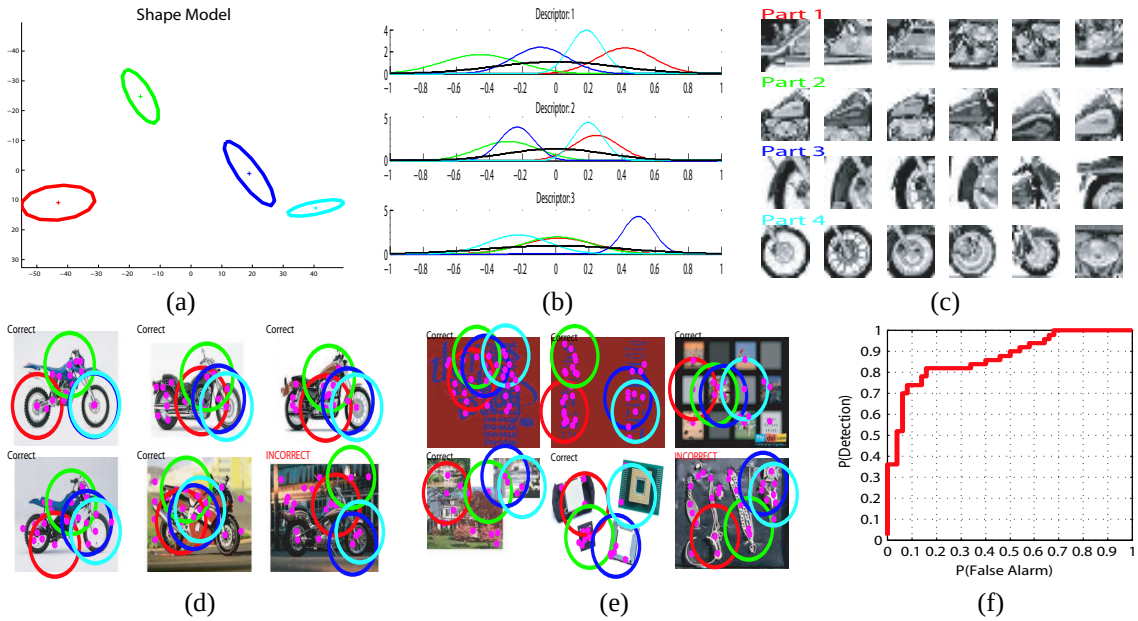


Figure 11.6: The mean model. **(a)** Shows the shape component of the model. The four $+$'s and ellipses indicate the mean and variance in position of each part. The inter-part covariance terms are not shown. **(b)** Shows the mean appearance distributions for the first 3 PCA dimensions. Each color indicates one of the four parts. The background density is shown in black. **(c)** Shows the detected feature patches in the training image closest to the mean of the appearance densities for each of the four parts. **(d)** Some examples of foreground test images for the model, with a mix of correct and incorrect classifications. The pink dots are features found on each image and the colored circles indicate the best hypothesis in the image. The size of the circles indicates the score of the hypothesis (the bigger the better). **(e)** The model running on some background query images. **(f)** The ROC curve for the model on the test set. The equal error rate is around 18%.

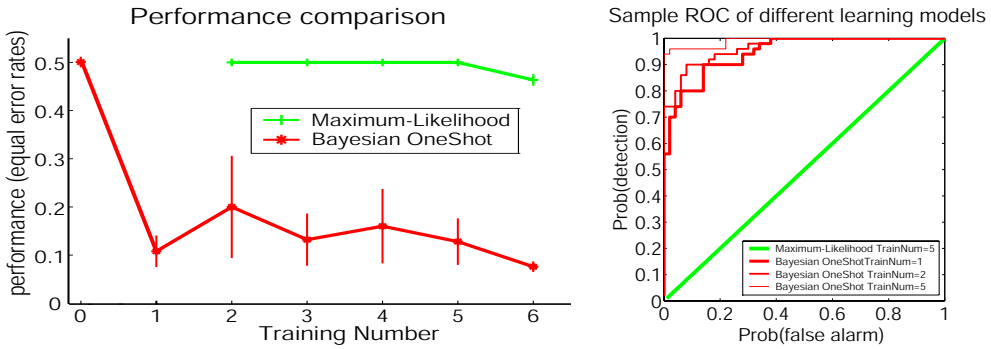
11.2.5 101 Object Categories

We have tested our algorithm on a large dataset of 101 object categories (Fig. 11.2). We summarize different aspects of our experiments in the following sections.

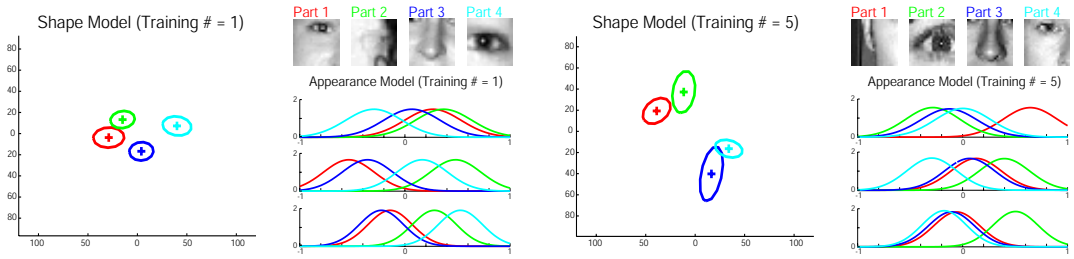
11.2.5.1 Overall Results: ML vs. MAP vs. Bayesian

Using the Bayesian formulation, we are able to incorporate prior knowledge of the object world into the learning scheme. In addition, we are also capable of averaging over the uncertainties of models by integrating over the model distributions. Do both of these two factors contribute in the efficient learning of our algorithm? Or is it only the prior that truly matters?

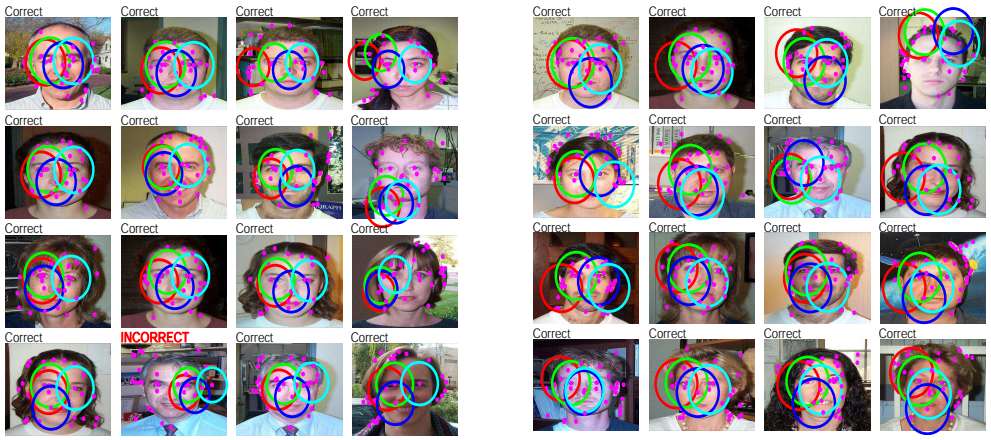
We are able to answer this question by comparing the detection result of the Bayesian One-Shot algorithm not only to the ML method, but also to the MAP algorithm (as derived in Appendix A.3). Both the Bayesian One-Shot and the MAP algorithms are given exactly the same prior distributions (estimated from faces, airplanes and spotted cats models) for learning for each of the 101 categories. While Fig. 11.11 illustrates



(a) (b)

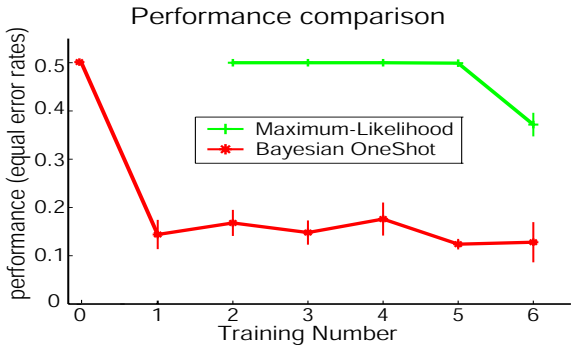


(c) Model from 1 training example. (d) Model from 5 training example.

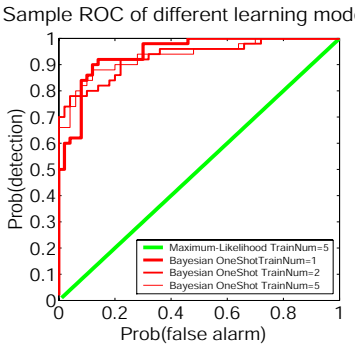


(e) Testing examples from model in (c). (f) Testing examples from model in (d).

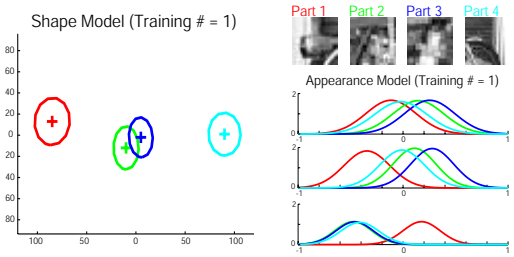
Figure 11.7: Summary of face model. **(a)** Test performances of the algorithm given $0-6$ number of training image(s) (red line). 0 number of training images is when only the prior model is used. Note that the prior alone is not sufficient for categorization. Each data point is obtained by 10 repeated runs with different randomly drawn training and testing images. Error bars show one standard deviation from the mean performance. This result is compared with the maximum-likelihood (ML) method (green). Note ML cannot learn the degenerate case of a single training image. **(b)** Sample ROC curves for the Bayesian One-Shot algorithm (red) compared with the ML algorithm (green line). The curves shown here use typical models drawn from the repeated runs summarized in **(a)**. **(c)-(f)** show typical models learnt with 1 and 5 training images. **(c)** Shape model; appearance samples and appearance densities (of the first 3 descriptors) for a model trained on 1 image. **(e)** Sample foreground test images for the model shown in **(c)**. **(d)** and **(f)** correspond to a model trained on 5 images.



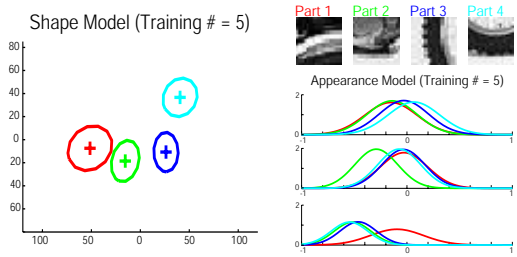
(a)



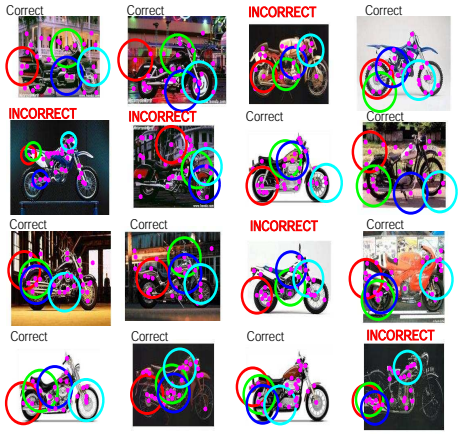
(b)



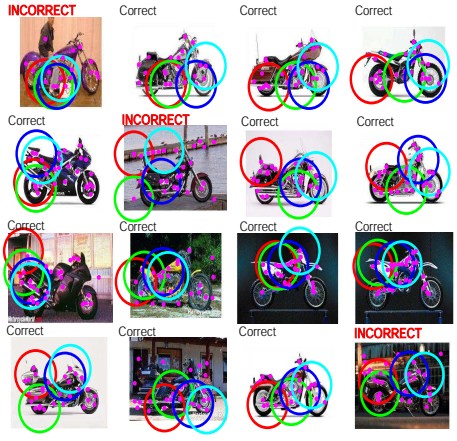
(c) Model from 1 training example.



(d) Model from 5 training example.

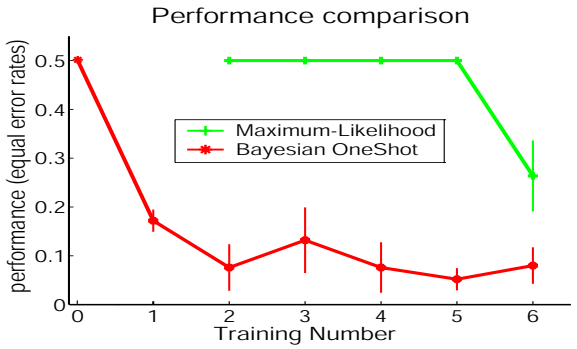


(e) Testing examples from model in (c).

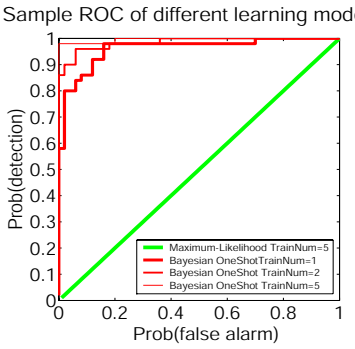


(f) Testing examples from model in (d).

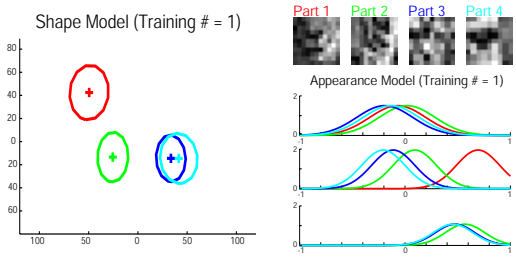
Figure 11.8: Summary of the motorbike model.



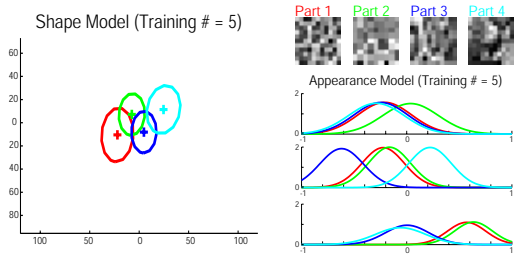
(a)



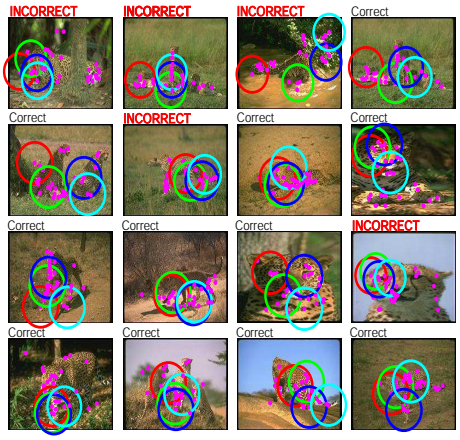
(b)



(c) Model from 1 training example.



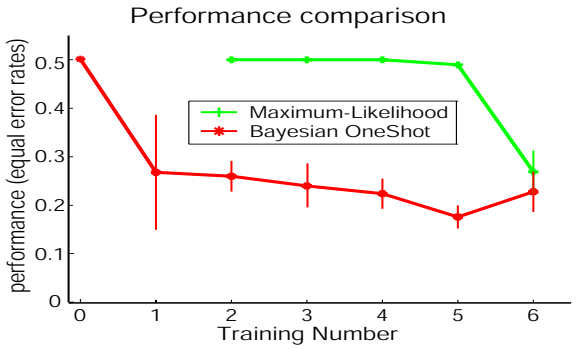
(d) Model from 5 training example.



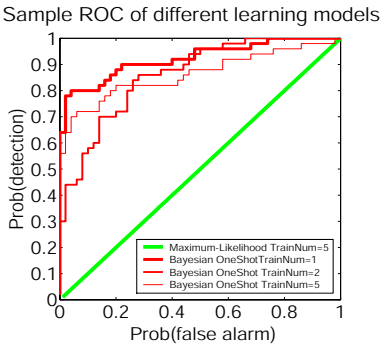
(e) Testing examples from model in (c).

(f) Testing examples from model in (d).

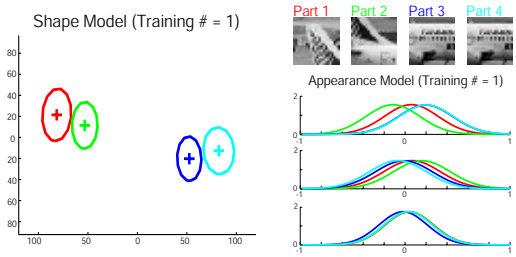
Figure 11.9: Summary of spotted cat model.



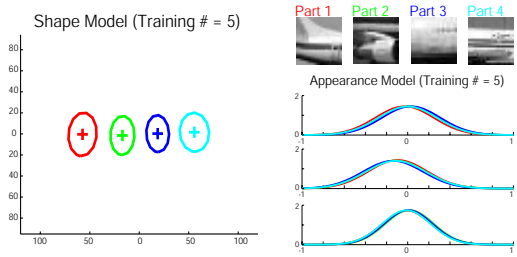
(a)



(b)

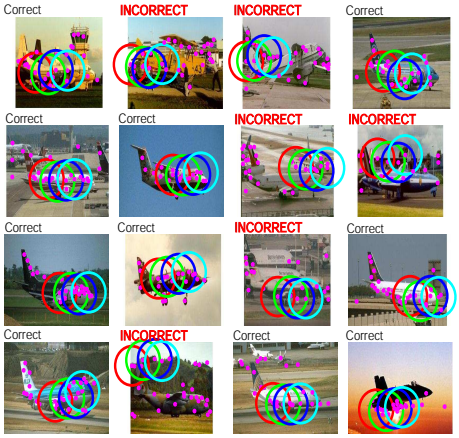


(c) Model from 1 training example.



(d) Model from 5 training example.

(e) Testing examples from model in (c).



(f) Testing examples from model in (d).

Figure 11.10: Summary of airplane model.

that prior knowledge helps in learning new object categories, the introduction of priors alone cannot account for all the advantages of our Bayesian formulation. The Bayesian algorithm consistently performed better than both the ML and MAP methods given few training examples. While MAP learning takes advantage of the prior density, it is fundamentally the same as maximum likelihood in that a single parameter set is estimated for the object category. Given few training examples, such an assumption is likely to overfit the few data points. The Bayesian algorithm reduces the overfit by averaging over model uncertainties.

11.2.5.2 Good models and Bad models

Figs. 11.12 and 11.13 show in detail the results from the grand-piano and cougar-face categories, both of which have achieved reasonable performances given few training examples (equal error rates of 84% and 85% respectively for 15 training examples). In the left-most columns, four examples of feature detection results are presented. The center of each detection circle indicates the location of the feature detected while the size of the circle indicates its scale. The second column shows the resulting shape model for the Bayesian One-Shot method for $\{1, 3, 6, 15\}$ training images. As the number of training examples increases, we observe that the shape model is more defined and structured with a reduction in variance. This is expected since the algorithm should be more and more confident of what is to be learned. The third column shows examples of the part appearance that are closest to the mean distribution of the appearance. Notice that distinctive features such as keyboards for the piano and eyes or whiskers for the cougar-face are successfully learned by the algorithm. Two learning methods' performances are compared in the top panel of the last column. The Bayesian methods clearly show a big advantage over the ML method when training number is small.

It is also useful to look at the other end of the performance spectrum–those categories that have low recognition performance. We give some informal observations into the cause of the poor performance. Feature detection is a crucial step for both learning and recognition. On both the crocodile and mayfly figures in Fig. 11.14, notice that some testing images marked "INCORRECT" have few detection points on the target object itself. When feature detection fails either in learning or recognition, it affects the performance results greatly. Furthermore, Fig. 11.12(a) shows that a variety of viewpoints is present in each category. In this set of experiments we have only used one mixture component, hence only a single viewpoint can be accommodated. Our model is also a simplified version of Burl, Weber and Fergus' constellation model [20, 39, 153] as it ignores the possibility of occluded parts.

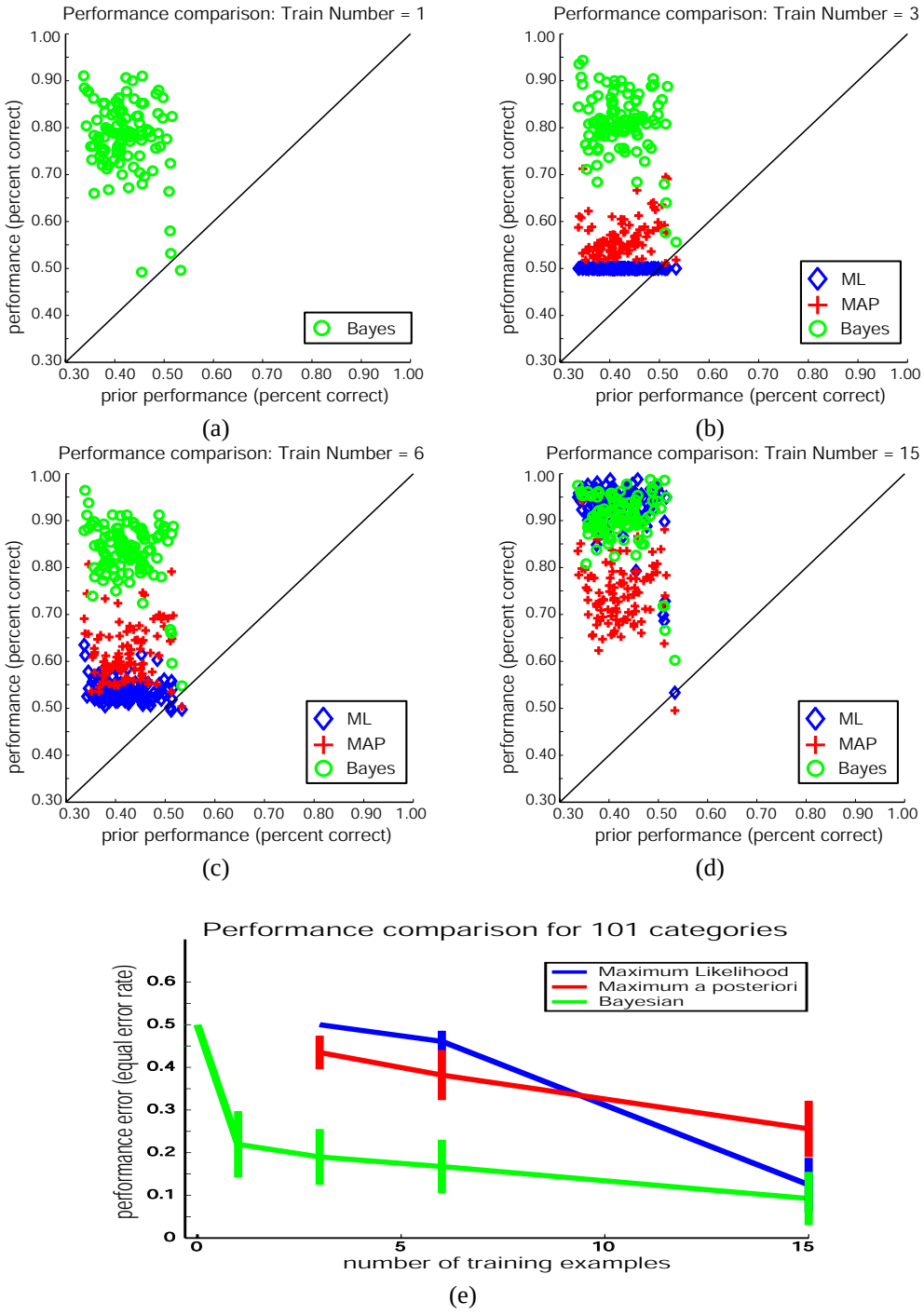


Figure 11.11: Performance on 101 categories using three different learning methods: Maximum Likelihood (ML), maximum a posteriori (MAP), and the Bayesian One-Shot algorithm. **(a) - (d)** show the performance given training number(s) 1, 3, 6, and 15 and compare them with performance of the prior alone. "Percent correct" is measured as $1 -$ Eq. Error Rate. **(e)** summarizes the four panels above, showing the mean performance (Eq. Error Rate). The error bars indicate one standard deviation.

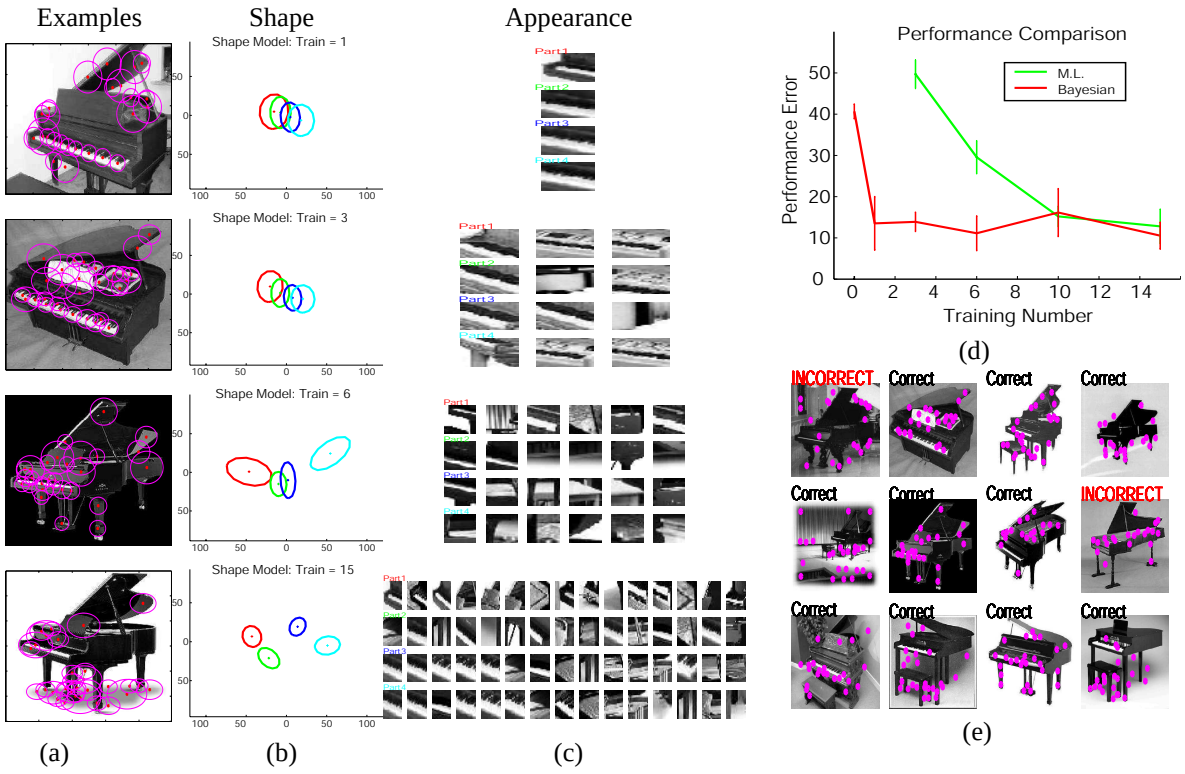


Figure 11.12: Results for the "grand-piano" category. Column 1 shows examples of feature detection. Column 2 shows the shape models learned from $\{1, 3, 6, 15\}$ training images. Column 3 shows the appearance patches for the model learned from $\{1, 3, 6, 15\}$ training images. The top panel of Column 4 shows the comparative results between ML and Bayesian methods (the error bars show the variation over the 10 runs). The bottom panel of Column 4 shows the recognition result for the Bayesian One-Shot algorithm for one training image. Pink dots indicate the center of detected interest points.

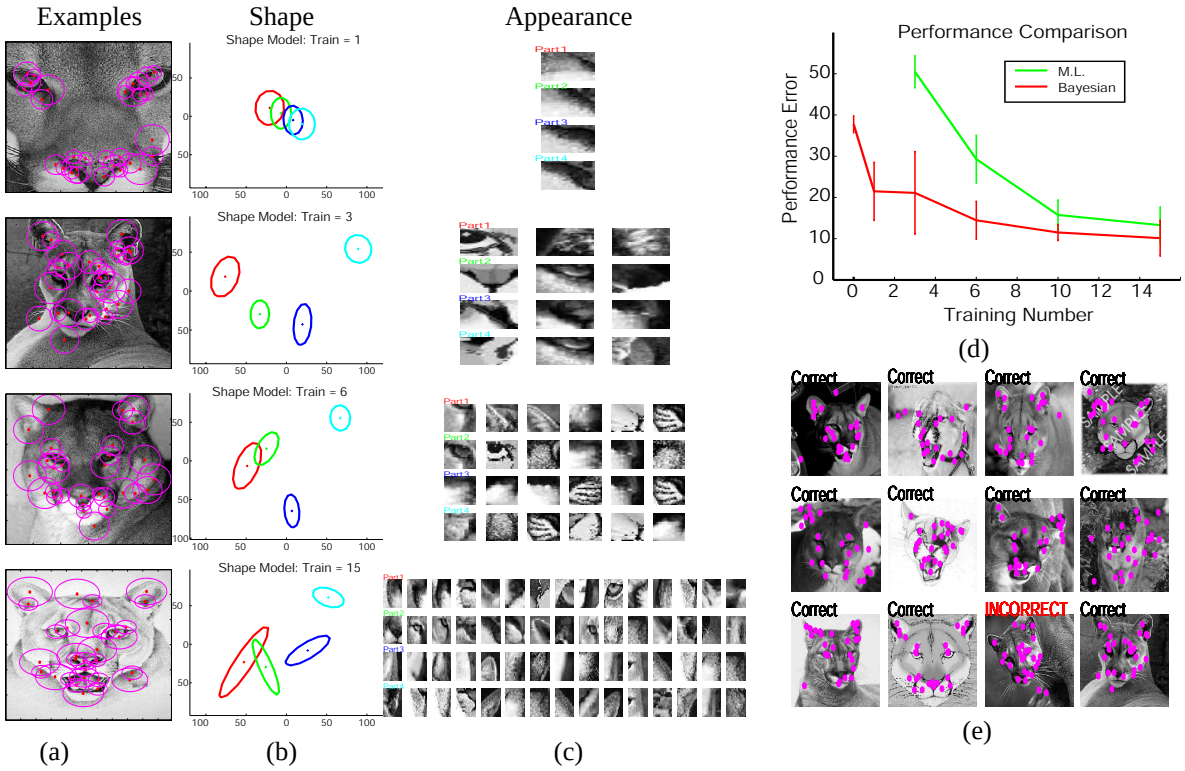


Figure 11.13: Results for the "cougar face" category.

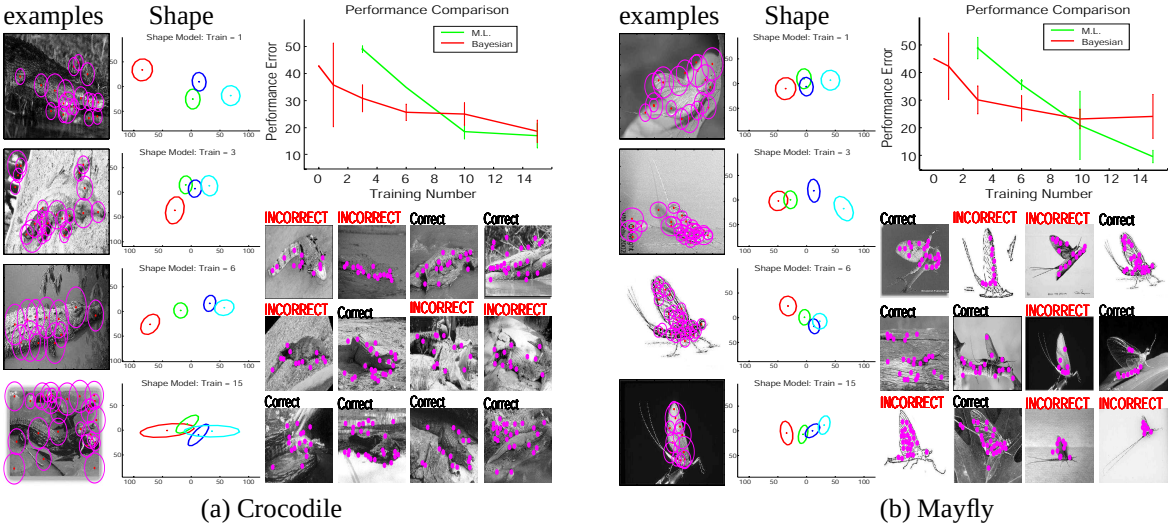


Figure 11.14: Two categories with poor performance. **(a)** Crocodile (equal error rate = 35% for 1 training example). **(b)** Mayfly (equal error rate = 42% for 1 training example).

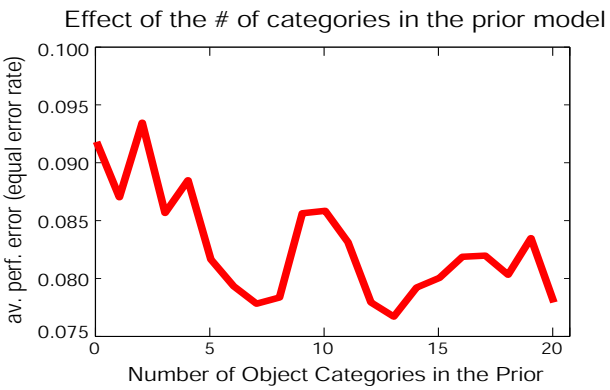
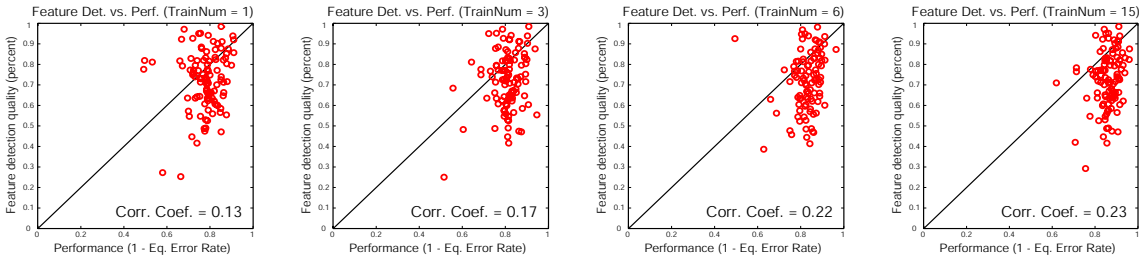


Figure 11.15: Effect of number of object categories in the prior model on the performance of testing categories. There are 20 randomly drawn object categories for training the prior model. There are 30 other randomly drawn object categories in the testing category set. The x-axis indicates the number of object categories in the prior model. The y-axis indicates the average performance error of the 30 test categories given the prior model.



(a) TrainNum = 1 (b) TrainNum = 3 (c) TrainNum = 6 (d) TrainNum = 15

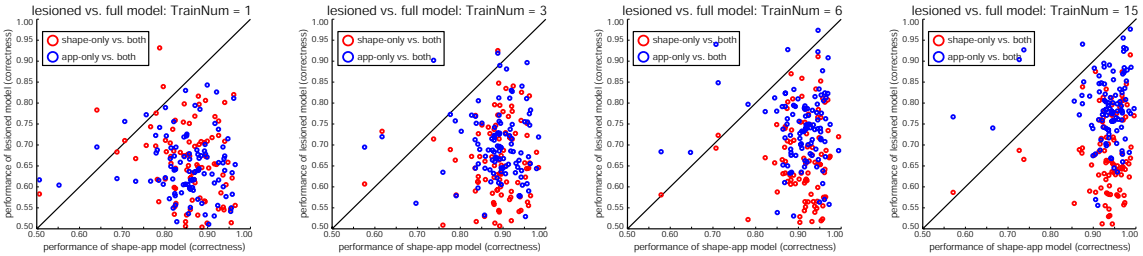
Figure 11.16: Quality of feature detection compared with object detection performances of the 101 categories given $\{1, 3, 6, 15\}$ training images (**(a) - (d)**). The x-axis of each plot is the detection performance of the model. The y-axis is the quality of feature detection, defined by the percentage of detection points landed within the outline of the object over the total number of detections. For each category, we average the percentage over all images within this category.

11.2.5.3 A Further Investigation on Prior Models and Feature Detectors

One useful question to ask is whether learning is improved by constructing the prior model from more categories. To investigate this, we randomly select 20 object categories that will incrementally contribute to the prior model. We learn a model for each of the 20 categories, forming a set of models C. We also randomly select 30 object categories from the rest of the dataset, calling this set S. We train a model for each category in S using a prior constructed from N models drawn from C. We vary N from 0 to 20. For $N = 0$, the prior model is a broad, non-informative distribution over the shape and appearance space. For $N > 0$, we pick a model from C, and update the prior as a weighted average between the old prior model and the new category model, the weighting being $N - 1$ and 1 respectively. Fig.11.15 shows the relationship between the number of categories contributing to the prior model and the performances averaged over all categories in S. We see

a trend of decreasing error when the number of categories in the prior model is between 1 and 8, although this trend becomes less clear beyond 8.

We also explored the effect of feature detections on the overall object detection performances. Two human subjects annotated the whole dataset, giving ground truth information of the location and the contours of the objects within each image. Given this information, we are able to compute the proportion of features detected within the object boundary as a fraction of the total number in the image. In Fig.11.16 we show the relationship between the quality of the feature detections and the performances for each training number. In general, a very weak positive correlation is observed between feature detection quality and performance. This correlation seems to increase slightly as the training number increases.



(a) TrainNum = 1 (b) TrainNum = 3 (c) TrainNum = 6 (d) TrainNum = 15

Figure 11.17: Shape only models and appearance only models compared with models using shape and appearance for each of the 101 categories given $\{1, 3, 6, 15\}$ training images (**(a) - (d)**). The x-axis of each plot is the detection performance of models using both shape and appearance. The y-axis is the detection performance of shape only models and appearance only models for each category.

11.2.5.4 Bayesian One-Shot Algorithm: Shape-Only vs. App-Only vs. Shape-App models

In Chapter 10.2 we detailed the formulation of object class models. Each model of an object category carries two sources of information: shape and appearance. We show in Fig. 11.17 that the contributions of shape and appearance components of the model vary when the object category to be learnt differs. While some categories depend more on the shape component (e.g., faces, electrical guitars, side view of cars, etc.), others rely more on the appearance (leopards, octopus, ketch, etc.). For most categories, learning is slightly more effective when the appearance component is included, as opposed to the shape part.

11.2.5.5 Bayesian One-Shot Algorithm: Discrimination Amongst 101 Categories

So far we have tested our algorithm in a *detection* scenario: for a particular object category we are only deciding if it is present or not. We now test the algorithm in a *discrimination* scenario: one where we have multiple categories (i.e., more than two) and must correctly classify the query images from each. In our

experiment, we first learn a model for each of the 101 object categories. Query images are then drawn from the test set of each category in turn and evaluated by all 101 models. For a given image, the assignment of category it belongs to is in the "winner-take-all" fashion. In other words, the category model that achieved the highest likelihood score is assigned to the image. For each category of images, we repeat the experiment 50 times with different randomly chosen training and test images. This gives a vector of 101 entries, each being the average of the "winner-take-all" assignment over the 50 repetitions. We do this for each of the 101 categories, thereby obtaining the confusion table in Fig. 15.4.

By averaging the correct discrimination rates, i.e., the entries along the diagonal of Fig. 15.4, we obtain the average correct discrimination rates for 3, 6 and 15 training examples of, respectively, 10.4, 13.9, and 17.7%. These rates would be approximately 1% if the classifiers were making random decisions. Recall that the corresponding correct detection rates are 73.6, 76.2 and 80.1%. Is there a way to predict discrimination rates from detection rates? We propose a simple approximation that produces good predictions and allows us to evaluate our current results in the context of our long-term goal of classifying thousands of categories.

Simulation relating detection and discrimination performance. What is the difference between detection and recognition? When detecting objects, e.g., bonsai trees, a single detector is first obtained for that category (e.g., by training on an appropriate collection of images). That detector is then used to compute the likelihood ratio of whether a given image contains a bonsai tree or not. If this number exceeds a threshold, then a bonsai tree is believed to be present. When discriminating (or recognizing) objects, the same image is inspected by a collection of detectors (e.g., bonsai, ceiling fan, automobile, etc.), each one of which is used to calculate a likelihood ratio. The highest likelihood ratio is taken as the indicator of the most likely category to be present. Whether we detect or discriminate, each detector will behave identically and will produce two densities of likelihood ratios: one conditioned on the preferred object category being present and one conditioned on the preferred category being absent from the image. The difference between detection and discrimination in our experiments is purely the number of competing hypotheses. There are two hypotheses in the case of detection, 101 hypotheses in the case of discrimination. Notice that one may in effect regard all incorrect classification hypotheses (100 of them) as one by taking the hypothesis that is associated with the highest likelihood ratio. This is the only incorrect hypothesis that has a chance of 'winning.' In our simple model we hypothesize that all detectors are independent (this hypothesis is clearly wrong, it is only justified as a coarse approximation). Furthermore, in our model all detectors have identical Gaussian densities describing both the response to the favorite category and to images not containing the favorite category. Such densities may be adjusted (by modifying the mean and variance) so that a given detection performance is obtained. From such densities one may also calculate (numerically, in our simulations) the density corresponding to the

best incorrect hypothesis, again by taking the highest likelihood result of a number of competing incorrect hypotheses. The probability of the discrimination error is then easily computed by Monte-Carlo simulations as a function both of the number of competing models, and as a function of the average detection performance. Such data is shown in Fig. 15.4(c). Notice that the red curve (100 categories) fits reasonably close to our experimental findings. We therefore predict that, in order to obtain 90% correct discrimination rate on 20 categories, we need detection errors smaller than 1%. For 100 and 1000 categories, respectively, in order to obtain a 90% correct discrimination rate we need, respectively, fewer than 3 errors every 1000 and 2 errors every 10^4 images. These are sobering requirements! It is clear that, as we improve the quality of our detector beyond, say, 95% correct, a more sensitive measure of performance will be given by discrimination rates, and thus recognition experiments should be preferred to detection experiments when comparing different approaches.

11.2.5.6 Discussions

Our results highlight a number of issues that we continue to investigate. The most important one is the choice of priors. We have used a very general prior constructed from three categories and would like to explore further the effects of different priors. Notice that in Fig. 11.11 the Maximum Likelihood method, on average, gives a similar level of performance to the Bayesian One-Shot algorithm for 15 training images. This is surprising, given the large number of parameters in each model, and therefore a few hundred training examples are in principle required by a Maximum Likelihood method–one might have expected the ML method to converge with the Bayesian One-Shot method at only around 100 training examples. The most likely reason for this result is that the prior that we employ is very simple. Bayesian methods live and die by the quality of the prior that is used. Our prior density is derived from only three object categories. Given the variability of our training set, it is realistic that a prior based on many more categories would yield a better performance. We have tested this hypothesis using a simple, synthetic example in Fig. 11.19. Our goal is to learn a simple triangular shape model (Fig. 11.19(a)). We test the effect of priors on the Bayesian One-Shot algorithm by giving the system three different priors: a triangular shape prior (similar to the synthetic model in Fig. 11.19(a) used to generate the data), a trapezium shape prior and a square shape prior. The Bayesian One-Shot algorithm with three different priors is compared to the maximum likelihood method. We observe that it takes more than 100 training examples for the ML method to "catch up" with the Bayesian One-Shot learning method given the triangular shape prior. On the contrary, it takes much smaller number of training examples for the ML method to converge with the other two Bayesian One-Shot learning methods with non-effective priors.

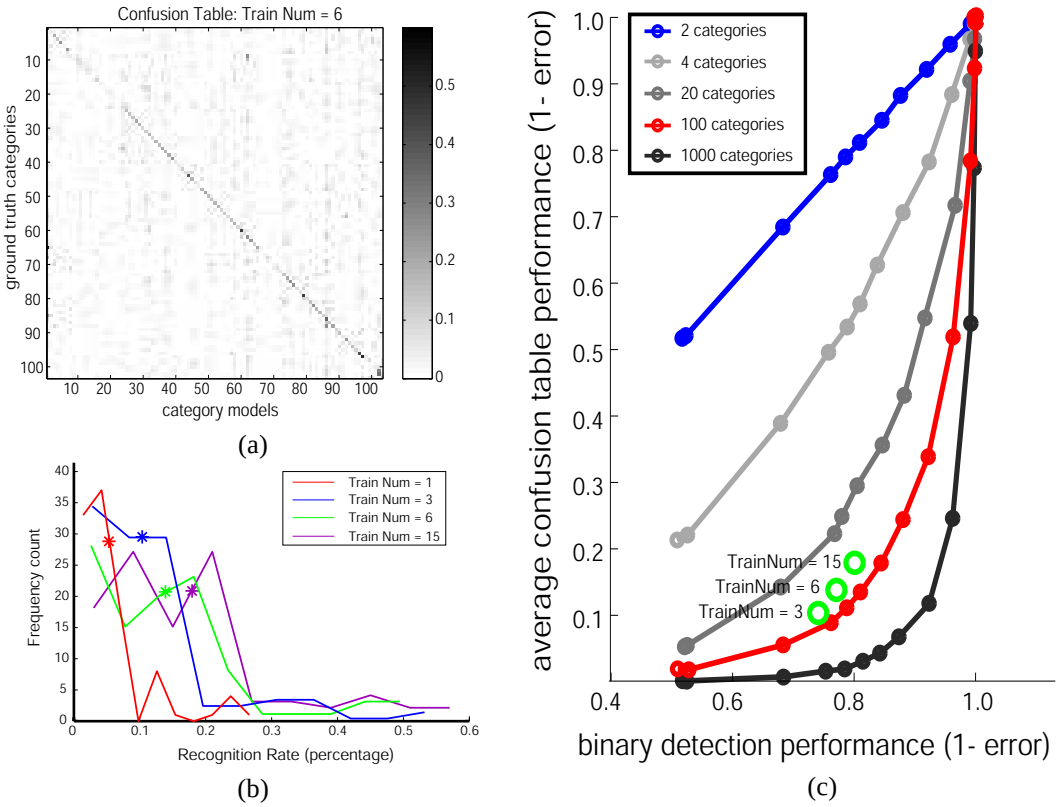
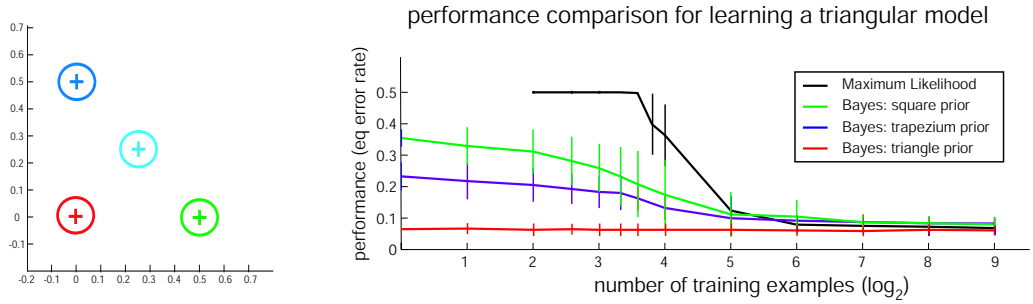


Figure 11.18: **(a)** A confusion table for 6 training examples. The x-axis enumerates the category models, one for each category, giving 101 in total. The y-axis is the ground truth category for the query image. The intensity of an entry in the table corresponds to the probability of a given query image being classified as a given category. Since the categories are consistently ordered on both axes, the ideal case would consist of a completely black diagonal line, showing perfect discrimination power of all category models over all categories of objects. **(b)** Histogram summary of diagonal entries of confusion tables for $\{1, 3, 6, 15\}$ training examples. The x-axis represents the recognition percentage of the discrimination task. The y-axis is a frequency count of the number of categories. The $*$ indicates the average confusion table performance given each training number. **(c)** Relationship between the binary *detection* performance and *discrimination* performance for differing number of categories using a one-dimensional Gaussian simulation.

Another important issue is the robustness of feature detection. We saw in Fig. 11.14 that the performance of models is highly dependent on obtaining a good set of stable and distinctive features from each object instance. We find that for some of the categories we experimented with, the Kadir-Brady feature detector [66] fails to detect consistently useful features, hence performance is impaired. Thus we are currently working on improving the quality and consistency of the feature detection stage.



(a) (b)

Figure 11.19: **(a)** The synthetic triangle model used in **(b)**. Note the triangle is characterized by a 4-part model. **(b)** Effect of different priors for learning a triangle model. Note that the point of convergence between the ML method and the Bayesian One-Shot method depends on the choice of prior distribution. When a prior is very effective (e.g., a triangular prior for learning a triangular model), it takes more than 100 training examples to converge. But when the prior is not very effective (e.g., square or trapezium priors for learning a triangular model), it takes less than 30 training examples for the two methods to converge.

Chapter 12

Summary

We have demonstrated that, contrary to intuition, useful aspects of a new object category may be learnt from a single training example (or just a few). As Table 16.1 shows, this is beyond the capability of existing algorithms.

The key insight we have exploited is that categories we have already learnt give us information that helps in learning new categories with fewer training examples. To pursue this idea we developed a Bayesian learning framework based on representing object categories with probabilistic models. Prior information from previously learnt categories is represented with a suitable prior probability density function on the parameters of their models. These 'prior' models are updated with the few training examples available to produce 'posteriors' that, in turn, may be used for both detection and recognition.

Our experiments, conducted on images from four categories, are encouraging in that they show that very few (1 to 5) training examples produce models that are already able to achieve a detection performance of around 10-20%. Our detection experiments conducted on 101 categories show that the method is applicable to a great variety of appearances. Furthermore: that the categories from which the 'prior' knowledge is learnt do not need to be visually similar to the categories that one wishes to learn.

While our experiments are very encouraging, they are by no means satisfactory from a practical standpoint. As our recognition experiments show, the margin for improvement of the recognition rates is enormous. Unless detection (object present/absent) error rates drop to almost zero, recognition rates of one out of 100-1000 categories will be disappointing (Figure 15.4(c)). Much can be done towards the goal of obtaining better error rates, as our current implementation is, at the moment, just a toy. In order to contain the complexity of our experiments, we have simplified the probabilistic models that are used for representing objects. For example, a probabilistic model for occlusion ([20, 39, 153]) was not implemented, and we only used four parts in our models, definitely not enough to represent the full complexity of object appearance. Furthermore, we only used three known categories to derive a prior. This is clearly a very small set that ought to be

Authors	*Categories*	*# Categories*	*# Training Images*	*Framework*	*Hand Alignment*	*Segmented*
Fei-Fei *et al.*	Assorted	101	1-5	Gen.	N	N
Fergus *et al.* [39]	Assorted	6	> 100	Gen.	N	N
Weber *et al.* [153]	Cars, Faces	2	> 100	Gen. + Disc.	N	N
Viola & Jones [148]	Faces	1	$\sim 10{,}000$	Disc.	Y	Y
Schneiderman & Kanade [125]	Cars	1	2,000	Disc.	Y	N
Rowley *et al.* [121]	Cars	1	500	Disc.	Y	N
Amit *et al.* [4]	Faces, Characters	3	300	Gen.	Y	Y
LeCun *et al.* [72]	Digits	10	60,000	Disc.	N	Y
LeCun *et al.* [73]	Assorted	5	$\sim$300,000	Disc.	Y	N

Table 12.1: A comparison among a variety of object recognition approaches. The framework column specified in the approach is generative (Gen.) or discriminative (Disc.) or both. The hand alignment and segmented columns indicate if the training data needs to be hand-aligned or hand-segmented for a given approach.

substantially broadened in a real-world situation.

However, at this point it is probably more important to make progress at the conceptual level, and much still needs to be done. First of all: the issue of priors. How much does prior knowledge improve as the number of known categories increases? Is it easier to learn new categories which are similar to some of the 'prior' categories? Second, the issue of representations: How should one best represent prior knowledge? Is there any other productive point of view, besides the Bayesian one that we have adopted here, which allows one to incorporate prior knowledge? Third, as we have pointed out in the introduction, it would be highly valuable to learn incrementally, where each training example will update the probability density function defined on the parameters of each object category; we presented a few ideas towards this in [33, 91].

One last note of optimism: we feel that the problem of recognizing automatically hundreds, perhaps thousands, of object categories does not belong to a hopelessly far future. We hope that the positive outcome of our experiments on the large majority of 101 very diverse and challenging categories, despite the simplicity of our implementation and the rudimentary prior we employ, will encourage other vision researchers to test their algorithms on larger and more diverse datasets.

Part V

Computational Model II: Natural Scene Classification

Chapter 13

Introduction

13.1 Background

The ability to analyze and classify accurately and rapidly the scene in which we find ourselves is highly useful in everyday life. Thorpe and colleagues found that humans are able to categorize complex natural scenes containing animals or vehicles very quickly [135]. Li and colleagues later showed that little or no attention is needed for such rapid natural scene categorization [77]. Both of these studies posed a serious challenge to the conventional view that to understand the context of a complex scene, one needs first to recognize the objects and then in turn recognize the category of the scene [140].

Can we recognize the context of a scene without having first recognized the objects that are present? A number of recent studies have presented approaches to classify indoor versus outdoor, city versus landscape and sunset versus mountain versus forest using global cues (e.g., power spectrum, color histogram information) [49, 132, 144]. Oliva and Torralba further incorporated the idea of using global frequency with local spatial constraints [94]. The key idea in their study is to use intermediate representations before classifying scenes: scenes are first labelled with respect to local and global properties by human observers. Similarly to Oliva and Torralba's work, Vogel and Schiele also used an intermediate representation obtained from human observers in learning the semantic context of a scene [150].

A main requirement of such approaches is the manual annotation of "intermediate" properties. In Oliva and Torralba's work, human subjects are instructed to rank each of the hundreds of training scenes into 6 different properties (e.g., ruggedness, expansiveness, roughness, etc.). In Vogel and Schiele's work, human subjects are asked to classify $59,582$ local patches from the training images into one of 9 different "semantic concepts" (e.g., water, foliage, sky, etc.). Both cases involve tens of hours of manual label. These works clearly point to the usefulness of these intermediate representations and motivate us to think of methods for learning these representations directly from the data: both because hand-annotating images is tedious and

expensive, and because expert-defined labels are somewhat arbitrary and possibly sub-optimal.

Much can also be learnt from studies for classifying different textures and materials [76, 106, 147]. Traditional texture models first identify a large dictionary of useful textons (or codewords). Then for each category of texture, a model is learnt to capture the signature distribution of these textons. We could loosely think of a texture as one particular intermediate representation of a complex scene. Again, such methods yield a model for this representation through manually segmented training examples. Another limitation of the traditional texture model is the hard assignment of one distribution for a class. This is fine if the underlying images are genuinely created by a single mixture of textons. But this is hardly the case in complex scenes. For example, it is not critical at all that trees must occupy 30% of a suburb scene and houses 60%. In fact, one would like to recognize a suburb scene whether there are many trees or just a few.

13.2 Contributions

The key insights of previous work, therefore, appear to be that using intermediate representations improves performance, and that these intermediate representations might be thought of as textures, in turn composed of mixtures of textons, or codewords. Our goal is to take advantage of these insights, but avoid using manually labelled or segmented images to train the system, if at all possible. To this end, we adapt to the problems of image analysis in recent work by Blei and colleagues [12], which was designed to represent and learn document models. In this framework, local regions are first clustered into different intermediate themes, and then into categories. Probability distributions of the local regions as well as the intermediate themes are both learnt in an automatic way, bypassing any human annotation. No supervision is needed apart from a single category label to the training image. We summarize our contribution as follows.

- Our algorithm provides a principled approach to learning relevant intermediate representations of scenes automatically and without supervision.
- Our algorithm is a principled probabilistic framework for learning models of textures via textons (or codewords) [76, 106, 147]. These approaches, which use histogram models of textons, are a special case of our algorithm. Given the flexibility and hierarchy of our model, such approaches can be easily generalized and extended using our framework.
- Our model is able to group categories of images into a sensible hierarchy, similar to what humans would do.

Chapter 14

Hierarchical Bayesian Model and Learning

Fig. 14.1 is a summary of our algorithm in both learning and recognition. We model an image as a collection of local patches. Each patch is represented by a codeword out of a large vocabulary of codewords (Fig. 14.3). The goal of learning is then to achieve a model that best represents the distribution of these codewords in each category of scenes. In recognition, therefore, we first identify all the codewords in the unknown image. Then we find the category model that fits best the distribution of the codewords of the particular image.

Our algorithm is based on the *Latent Dirichlet Allocation* (LDA) model proposed by Blei et al. [12]. We differ from their model by explicitly introducing a category variable for classification. Furthermore, we propose two variants of the hierarchical model (Fig. 14.2(a) and (b)).

14.1 Model Structure

It is easier to understand the model (Fig. 14.2(a)) by going through the generative process for creating a scene in a specific category. To put the process in plain English, we begin by first choosing a category label, say a mountain scene. Given the mountain class, we draw a probability vector that will determine what intermediate theme(s) to select while generating each patch of the scene. Now for creating each patch in the image, we first determine a particular theme out of the mixture of possible themes. For example, if a "rock" theme is selected, this will in turn privilege some codewords that occur more frequently in rocks (e.g., slanted lines). Now that the theme favoring more horizontal edges is chosen, one can draw a codeword, which is likely to be a horizontal line segment. We repeat the process of drawing both the theme and codeword many times, eventually forming an entire bag of patches that would construct a scene of mountains. Fig. 14.2(a) is a graphical illustration of the generative model. We will call this model the Theme Model 1. Fig. 14.2(b) is a slight variation of the model in Fig. 14.2(a). We call it the Theme Model 2. Unless otherwise specified, the

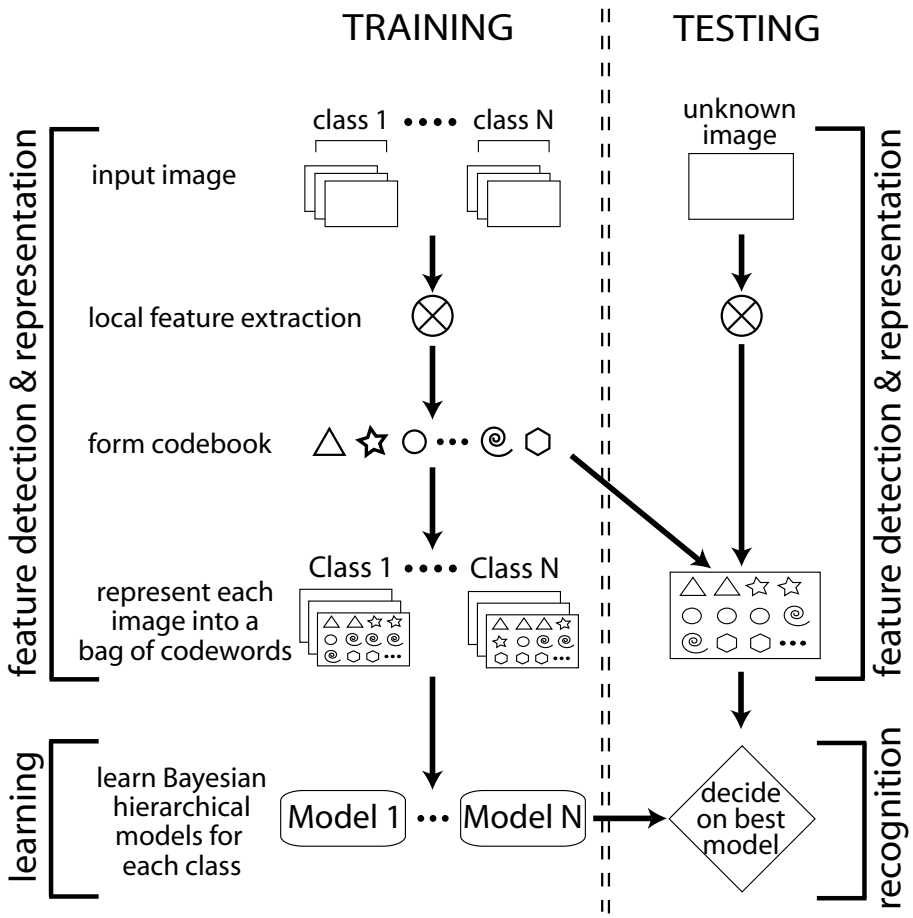


Figure 14.1: Flow chart of the algorithm.

rest of the chapter will focus on Theme Model 1. Now we are ready to show the mathematical details of the formulation of this model and how we learn its parameters.

14.1.1 The Theme Models

We begin with some notations and definitions for the Theme Model 1 in Fig. 14.2(a). We will contrast explicitly the use of terminology with both [12] and the texture studies [76, 147].

- A *patch* x is the basic unit of an image, defined to be a patch membership from a dictionary of codewords indexed by $\{1, \ldots, T\}$. The t^{th} codeword in the dictionary is represented by a T-vector x such that $x^t = 1$ and $x^v = 0$ for $v \neq t$. In Fig. 14.2(a), x is shaded by common convention to indicate that it is an observed variable. All other nodes in the graph are unobserved, hence no shading. The equivalent of an image in [12] is a "document." And a codeword (or patch) in our model is a "word" in theirs. In texture and material literature, a codeword is also referred as to a "texton" [76, 147].

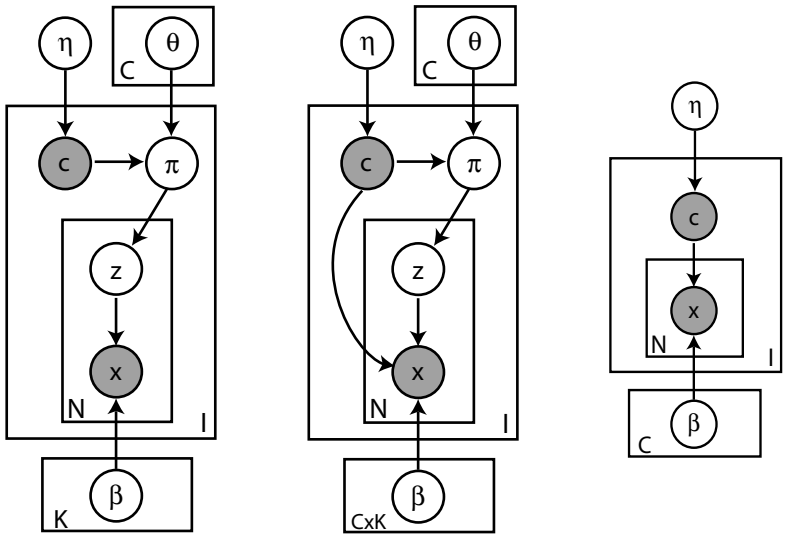


Figure 14.2: **(a)** Theme Model 1 for scene categorization that shares both the intermediate level themes as well as feature level codewords. **(b)** Theme Model 2 for scene categorization that shares only the feature level codewords; **(c)** Traditional texton model [76, 147].

- An *image* is a sequence of N patches denoted by $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$, where x_n is the n^{th} patch of the image.
- A *category* is a collection of I images denoted by $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_I\}$. In [12], this is equivalent to a "corpus."

We can now write down the process that generates an image i formally from the model.

1. Choose a category label $c \sim p(c|\boldsymbol{\eta})$ for each image, where $c = \{1, \ldots, C\}$. C is the total number of categories. $\boldsymbol{\eta}$ is a C-dimensional vector of a multinomial distribution.
2. Now for this particular image in category c, we want to draw a parameter that determines the distribution of the intermediate themes (e.g., how "foliage," "water," "sky" etc. are distributed for this scene). This is done by choosing $\boldsymbol{\pi} \sim p(\boldsymbol{\pi}|c, \boldsymbol{\theta})$ for each image. $\boldsymbol{\pi}$ is the parameter of a multinomial distribution for choosing the themes. $\boldsymbol{\theta}$ is a matrix of size $C \times K$, where $\boldsymbol{\theta}_{c\cdot}$ is the K-dimensional Dirichlet parameter conditioned on the category c. K is the total number of themes.
3. For each N patches x_n in the image:
   - Choose a theme $z_n \sim \text{Mult}(\boldsymbol{\pi})$. z_n is a K-dim unit vector. $z_n^k = 1$ indicates that the k^{th} theme is selected (e.g., "rock" theme).

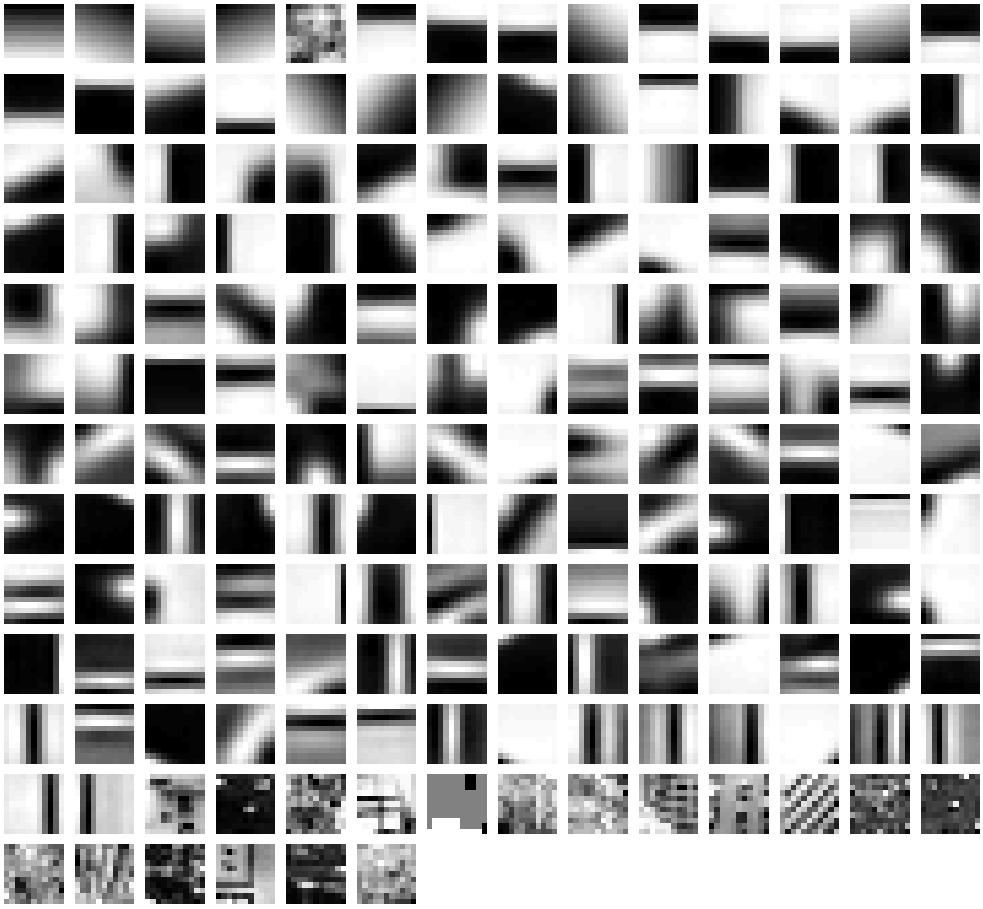


Figure 14.3: A codebook obtained from 650 training examples from all 13 categories (50 images from each category). Image patches are detected by a sliding grid and random sampling of scales. The codewords are sorted in descending order according to the number of patches that belong to the codeword. Interestingly most of the codewords appear to represent simple orientations and illumination patterns, similar to the ones that we would find in the early human visual system.

.

- Choose a patch $x_n \sim p(x_n|z_n, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a matrix of size $K \times T$. K is again the number of themes, and T is the total number of codewords in the codebook. Therefore we have $\boldsymbol{\beta}_{kt} = p(x_n^t = 1|z_n^k = 1)$.

A $K-$dimensional Dirichlet random variable $\boldsymbol{\pi}$ has the property such that $\pi_i \geq 0, \sum_{i=1}^{K} \pi_i = 1$. It is a conjugate distribution of a multinomial distribution. Since the themes $\boldsymbol{z}$ are best described as a discrete variable over the multinomial distribution, Dirichlet distribution becomes the natural choice to describe distribution of $\boldsymbol{\pi}$ [45]. It has the following probability density:

$$Dir(\boldsymbol{\pi}|\boldsymbol{\theta}_{c\cdot}) = \frac{\Gamma\left(\sum_{i=1}^{K} \theta_{ci}\right)}{\prod_{i=1}^{K} \Gamma(\theta_{ci})} \pi_i^{(\theta_{ci}-1)} \dots \pi_K^{(\theta_{cK}-1)} \tag{14.1}$$

Given the parameters $\boldsymbol{\theta}$, $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$, we can now write the full generative equation of the model. It is the joint probability of a theme mixture $\boldsymbol{\pi}$, a set of N themes $\boldsymbol{z}$, a set of N patches $\boldsymbol{x}$ and the category c is

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, c|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\beta}) &= p(c|\boldsymbol{\eta})p(\boldsymbol{\pi}|c, \boldsymbol{\theta}) \cdot \\ &\quad \prod_{n=1}^{N} p(z_n|\boldsymbol{\pi})p(x_n|z_n, \boldsymbol{\beta}) \qquad (14.2) \\ p(c|\boldsymbol{\eta}) &= \text{Mult}(c|\boldsymbol{\eta}) \qquad (14.3) \\ p(\boldsymbol{\pi}|c, \boldsymbol{\theta}) &= \prod_{j=1}^{C} Dir(\boldsymbol{\pi}|\boldsymbol{\theta}_{j\cdot})^{\delta(c,j)} \qquad (14.4) \\ p(z_n|\boldsymbol{\pi}) &= \text{Mult}(z_n|\boldsymbol{\pi}) \qquad (14.5) \\ p(x_n|z_n, \boldsymbol{\beta}) &= \prod_{k=1}^{K} p(x_n|\boldsymbol{\beta}_{k\cdot})^{\delta(z_n^k,1)} \qquad (14.6) \end{aligned}$$

As Fig. 14.2(a) shows, Theme Model 1 is a hierarchical representation of the scene category model. The Dirichlet parameter $\boldsymbol{\theta}$ for each category is a category-level parameters, sampled once in the process of generating a category of scenes. The multinomial variables $\boldsymbol{\pi}$ are scene-level variables, sampled once per image. Finally, the discrete theme variable $\boldsymbol{z}$ and patch $\boldsymbol{x}$ are patch-level variables, sampled every time a patch is generated.

If we wish to model the intermediate themes for each category without sharing them amongst all categories, we would introduce a link between the class node c to each patch x_n, such that $x_n \sim p(x_n|z_n, \boldsymbol{\beta}, c)$, where there are C different copies of $\boldsymbol{\beta}$, each of the size $K \times T$. Then we have $\beta_{kt}^c = p(x_n^t|z_n^k = 1)$. The generative equations above (Eq. 14.2-14.6) are hence changed according to this dependency on c.

14.1.2 Bayesian Decision

Before we show how we could proceed to learn the model parameters, let us first look at how decisions are made given an unknown scene. An unknown image is first represented by a collection of patches, or codewords. We keep the notation $\boldsymbol{x}$ for an image of N patches. Given $\boldsymbol{x}$, we would like to compute the probability of each scene class:

$$p(c|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta}) \propto p(\boldsymbol{x}|c, \boldsymbol{\theta}, \boldsymbol{\beta})p(c|\boldsymbol{\eta}) \propto p(\boldsymbol{x}|c, \boldsymbol{\theta}, \boldsymbol{\beta}) \qquad (14.7)$$

where $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are parameters learnt from a training set. For convenience, the distribution of $p(c|\boldsymbol{\eta})$ is always assumed to be a fixed uniform distribution in which $p(c) = 1/C$. Therefore we will omit to estimate $\boldsymbol{\eta}$ from now on. Then the decision of the category is made by comparing the likelihood of $\boldsymbol{x}$ given each category: $c = \arg\max_c p(\boldsymbol{x}|c, \boldsymbol{\theta}, \boldsymbol{\beta})$. The term $p(\boldsymbol{x}|c, \boldsymbol{\theta}, \boldsymbol{\beta})$ is in general obtained by integrating over the

hidden variables $\boldsymbol{\pi}$ and $\boldsymbol{z}$ in Eq. 14.2:

$$p(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\beta},c) = \int p(\boldsymbol{\pi}|\boldsymbol{\theta},c)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\boldsymbol{\pi})p(x_n|z_n,\beta)\right) d\boldsymbol{\pi} \tag{14.8}$$

Unfortunately Eq. 14.8 is not tractable due to the coupling between $\boldsymbol{\pi}$ and $\boldsymbol{\beta}$ [12]. However, a wide range of approximate inference algorithms can be considered, including Laplace approximation, variational approximation and MCMC method [12]. In the following section, we briefly outline the variational method based on Variational Message Passing (VMP) [155], a convenient framework to carry out variational inferences.

14.1.3 Learning: Variational Inference

In learning, our goal is to maximize the log likelihood term $\log p(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\beta},c)$ by estimating the optimal $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Using Jensen's inequality, we can bound this log likelihood in the following way:

$$\begin{aligned}\log p(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\beta}) &\geq \int\sum_{\boldsymbol{z}} q(\boldsymbol{\pi},\boldsymbol{z})\log p(\boldsymbol{\pi},\boldsymbol{z},\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\beta})d\boldsymbol{\theta} - \\ &\quad \int\sum_{\boldsymbol{z}} q(\boldsymbol{\pi},\boldsymbol{z})\log q(\boldsymbol{\pi},\boldsymbol{z}) \\ &= E_q\left[\log p(\boldsymbol{\pi},\boldsymbol{z},\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\beta})\right] - E_q\left[\log q(\boldsymbol{\pi},\boldsymbol{z})\right]\end{aligned}$$

where the probability density function $q(\boldsymbol{\pi},\boldsymbol{z}|\gamma,\phi)$ could be any arbitrary variational distribution. By letting $L(\gamma,\phi;\boldsymbol{\theta},\boldsymbol{\beta})$ denote the RHS of the above equation, we have:

$$\begin{aligned}\log p(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\beta}) &= L(\gamma,\phi;\boldsymbol{\theta},\boldsymbol{\beta}) + \\ &\quad KL(q(\boldsymbol{\pi},\boldsymbol{z}|\gamma,\phi)\parallel p(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\beta}))\end{aligned} \tag{14.9}$$

where the second term on the RHS of the above equation stands for the Kullback-Leibler distance of two probability densities. By maximizing the lower bound $L(\gamma,\phi;\boldsymbol{\theta},\boldsymbol{\beta})$ with respect to γ and ϕ is the same as minimizing the KL distance between the variational posterior probability and the true posterior probability.

Given Eq. 14.9, we first estimate the variational parameters γ and ϕ. Substituting the variational lower bound as a surrogate for the (intractable) marginal likelihood, we can then in turn estimate the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. The iterative algorithm alternates between the following two steps till convergence.

1. (E-step) For each class of images, optimize values for the variational parameters γ and ϕ. The update

rules are:

$$\gamma_i = \boldsymbol{\theta}_i + \sum_{n=1}^{N} \phi_{ni} \tag{14.10}$$

$$\phi_{ni} \propto \beta_{i\nu} \exp\left[\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j)\right] \tag{14.11}$$

where i is the image index, n the patch index and $\Psi(\cdot)$ a digamma function.

2. (M-step) Maximize the resulting lower bound on the log likelihood with respect to the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. We can do this by finding the maximum likelihood estimates with expected sufficient statistics computed in the E-step [12, 88].

14.1.4 A Brief Comparison

We can compare this hierarchical model with a traditional texton model for texture recognition, for instance [76, 147]. Fig. 14.2(c) is a graphical representation of a traditional texton model. We see here that for a given class of textures or materials, only a single multinomial parameter $\boldsymbol{\beta}$ is associated with the class. In other words, to generate an image, all patches are drawn from a single "theme." This might be fine when the training data are "pure" textures segmented manually. Since there are no "contaminations" of other "themes", the single mixture learnt from the codewords might suffice. As shown by [76], this framework may be further extended by training different models for the same category of textures under different lighting and view point conditions. This again requires manual separations of data and labelling of the segmented textures. In Chapter **??**, we will show empirically that by explicitly modelling the intermediate themes in these complex scenes, our model achieves better recognition performances than the traditional "texton" model in Fig. 14.2(c).

14.2 Features and Codebook

In the formulation of the theme model, we represent each image as a collection of detected patches, each assigned a membership to a large dictionary of codewords. We show now how these patches are obtained and memberships assigned.

14.2.1 Local Region Detection and Representation

While most previous studies on natural scene categorization have focused on using global features such as frequency distribution, edge orientations and color histogram [49, 132, 144], recently it has been shown local

regions are very powerful cues [150]. Compared to the global features, local regions are more robust to occlusions and spatial variations. We have tested four different ways of extracting local regions from images.

1. *Evenly Sampled Grid.* An evenly sampled grid spaced at 10×10 pixels for a given image. The size of the patch is randomly sampled between scale 10 to 30 pixels.

2. *Random Sampling.* 500 randomly sampled patches for a given image. The size of the patch is also randomly sampled between scale 10 to 30 pixels.

3. *Kadir and Brady Saliency Detector.* Roughly $100 \sim 200$ regions that are salient over both location and scale are extracted using the saliency detector [66]. Scales of each interest point is between 10 to 30 pixels.

4. *Lowe's DoG Detector.* Roughly $100 \sim 500$ regions that are stable and rotationally invariant over different scales are extracted using the DoG detector [79]. Scales of each interest point vary between 20 to 120 pixels.

We have used two different representations for describing a patch: normalized 11×11 pixel gray values or a $128-$dim SIFT vector [79].

14.2.2 Codebook Formation

Given the collection of detected patches from the training images of all categories, we learn the codebook by performing k-means algorithm [76]. Clusters with too small a number of members are further pruned out. Codeswords are then defined as the centers of the learnt clusters. Fig. 14.3 shows the 174 codewords learnt from the gray value pixel intensities.

Chapter 15

Experiments & Results

15.1 Dataset and Experimental Setup

Our dataset contains 13 categories of natural scenes (Fig. 15.1): highway ([94], 260 images), inside of cities ([94], 308 images), tall buildings ([94], 356 images), streets ([94], 292 images), suburb residence (241 images), forest ([94], 328 images), coast ([94], 360 images), mountain ([94], 374 images), open country ([94], 410 images), bedroom (174 images), kitchen (151 images), livingroom (289 images) and office (216 images). The average size of each image is approximately 250×300 pixels. The 8 categories that are provided by Oliva and Torralba were collected from a mixture of COREL images as well as personal photographs [94]. The rest of the 5 categories are obtained by us from both the Google image search engine as well as personal photographs. It is also worth noting that 4 (coast, forest, open country and mountain) of the categories are similar to the 4 of the 6 categories reported in [150]. But unlike them, we only use grayscale images for both learning and recognition. We believe that this is the most complete scene category dataset used in literature thus far.

Each category of scenes was split randomly into two separate sets of images, N (100) for training and the rest for testing. A codebook of codewords was learnt from patches drawn from a random half of the entire training set. A model for each category of scenes was obtained from the training images. When asked to categorize one test image, the decision is made by assigning a category label to the image from the category model that gives the highest likelihood probability. A confusion table is used to illustrate the performance of the models. On the confusion table, the x-axis represents the models for each category of scenes. The y-axis represents the ground truth categories of scenes. The orders of scene categories are the same in both axes. Hence in the ideal case one should expect a completely white diagonal line to show perfect discrimination power of all category models over all categories of scenes. Unless otherwise specified, all performances in Chapter **??** are quoted as the average value of the diagonal entries of the confusion table. For a 13-category

Figure 15.1: Our dataset consists of 13 categories, the largest natural scene category dataset to date.

Descriptor	Grid	Random	Saliency [66]	DoG [79]
11×11 Pixel	64.0%	47.5%	45.5%	N/A
128-dim Sift	65.2%	60.7%	53.1%	52.5%

Table 15.1: Performance comparison given different feature detectors and representations. The performance is quoted from the mean of the confusion table similar to that of Fig. 15.4. SIFT representation seems to be in general more robust than the pixel grayvalue representation. The sliding grid, which yields the most number of patches, outperforms all other detectors.

recognition task, random chance would be 7.7%. Excluding the preprocessing time of feature detection and codebook formation, it takes a few minutes (less than 10) to obtain 13 categories of models (100 training images for each category) on a 2.6 Ghz machine.

15.2 Results

Fig. 15.4 is an overview of the performance of the Theme Model 1 trained with 100 images from each of the 13 categories. Fig. 14.3 shows the corresponding codebook learnt. 650 testing images (50 from each class) are used. There are a total number of 40 themes. Our model achieved an average performance of 64.0% (random chance is 7.7%). A closer look at the confusion table (Fig. 15.4(a)) reveals that the highest block of errors occurs among the four indoor categories: bedroom, livingroom, kitchen and office. Another way to evaluate the performance is to use the rank statistics of the categorization results (Fig. 15.4(b)). Using both the best and second best choices, the mean categorization result goes up nearly 20% to 82.3%.

Both Fig. 15.2 and Fig. 15.5 demonstrate some of the internal structure of the models learnt for each category. Take the "highway" category as an example in Fig. 15.2. The left panel shows the average distribution of the 40 intermediate themes for generating highway images. In the right panel, we show the average distribution of all codewords for generating highway images, after a large number of samplings (1000). Clearly, this distribution of codewords (174, Fig. 14.3) is very much influenced by the distribution of themes. We show in the right panel 10 of the top 20 codewords that are most likely to occur in highway images. Note that horizontal lines dominate the top choices. This is to be contrasted, for instance, to the likely codewords for the tall building category. We see that most of the top-choice codewords are vertical edges in the case of tall buildings. The 4 indoor categories all tend to have sharp horizontal and vertical edges. This is quite revealing of the scene statistics for these manmade, indoor structures. Further, by looking at the distribution of both the themes and the codewords of these four indoor categories, it is not surprising that they are easily confused among each other. Fig. 15.3 then shows some testing image examples.

Taking the distributions of themes from each category, we can further establish some relationship among

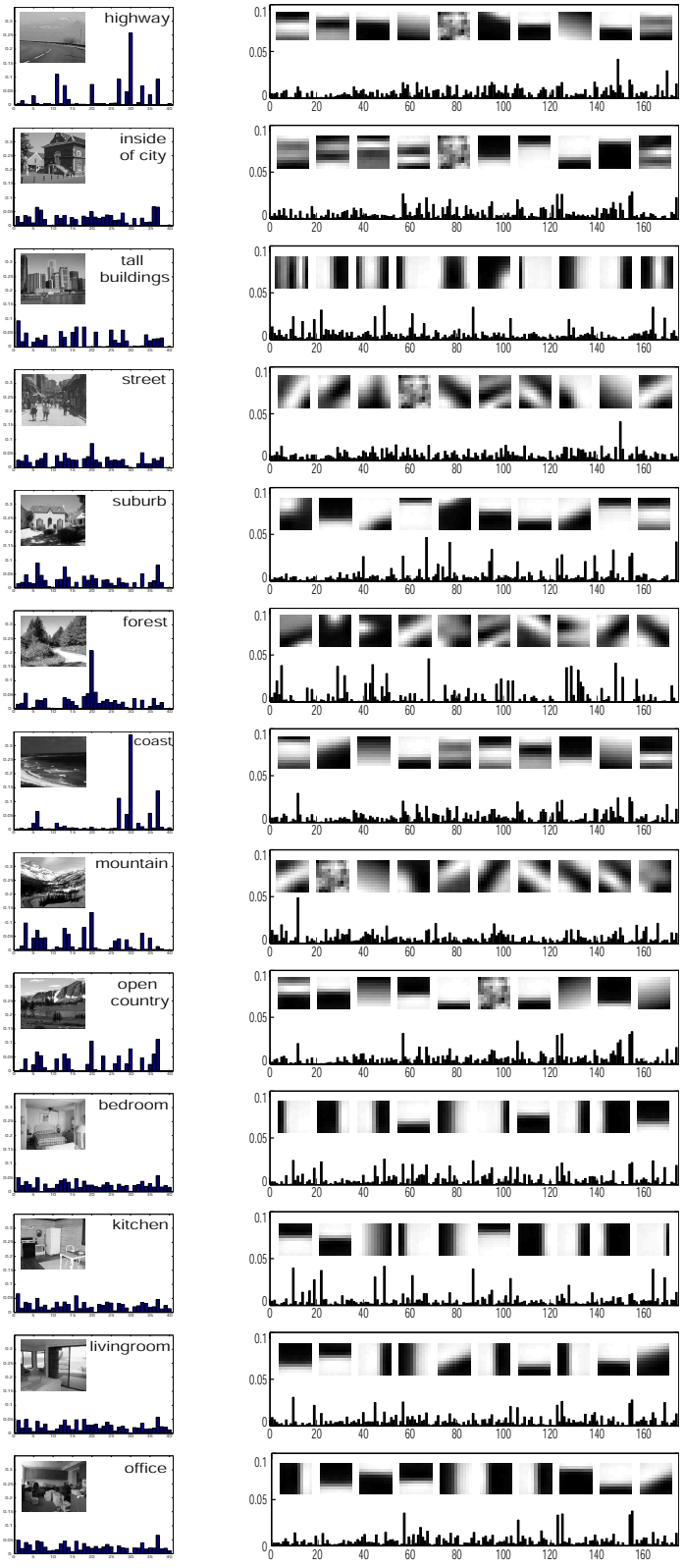


Figure 15.2: Internal structure of the models learnt for each category. Each row represents one category. The left panel shows the distribution of the 40 intermediate themes. The right panel shows the distribution of codewords as well as the appearance of 10 codewords selected from the top 20 most likely codewords for this category model.

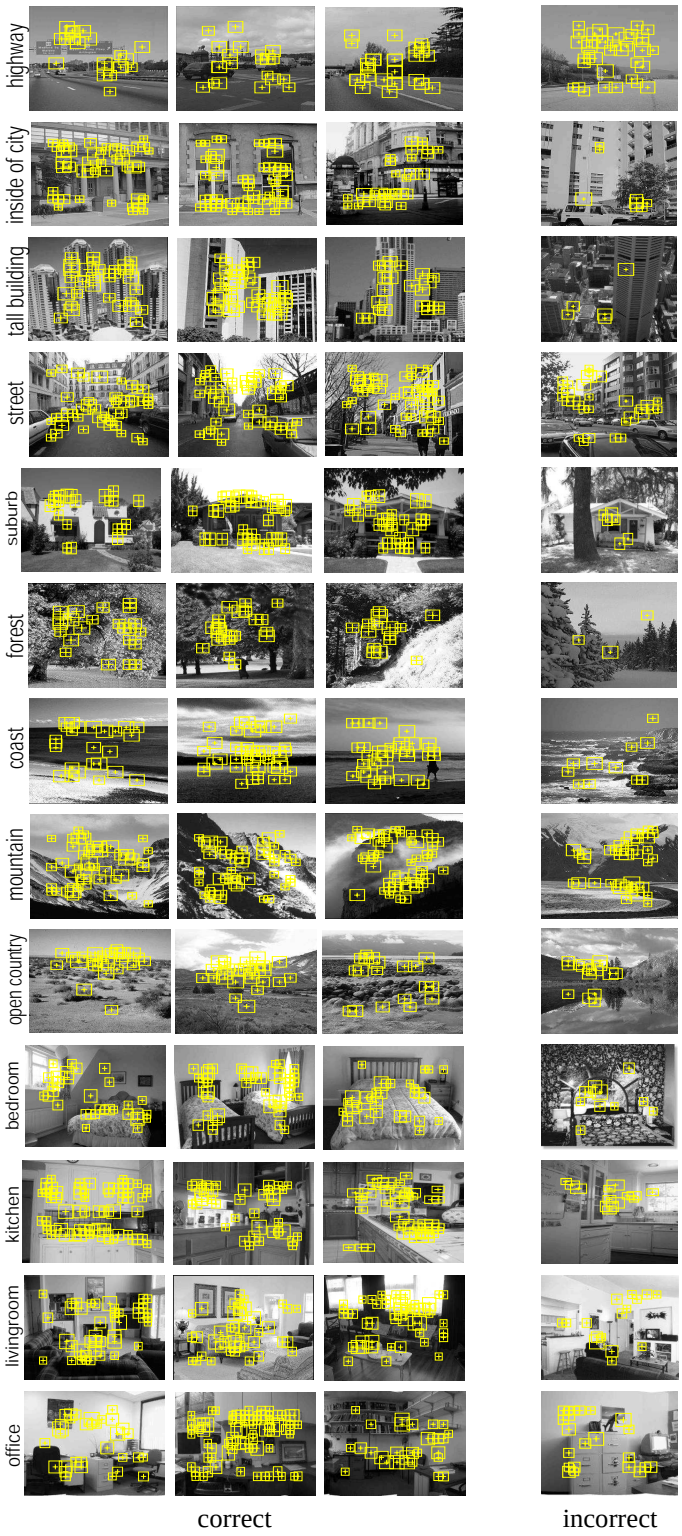


Figure 15.3: Examples of testing images for each category. Each row is for one category. The first 3 columns on the left show 3 examples of correctly recognized images; the last column on the right shows an example of an incorrectly recognized image. Superimposed on each image, we show samples of patches that belong to the most significant set of codewords given the category model. Note that for the incorrectly categorized images, the number of significant codewords of the model tends to occur less likely. (This figure is best viewed in color.)

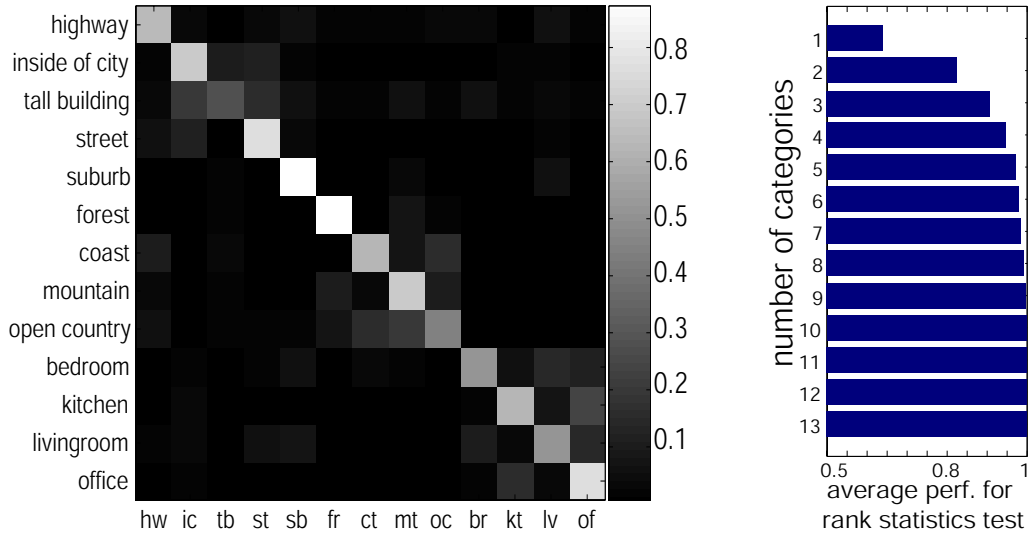


Figure 15.4: **Left Panel.** Confusion table of Theme Model 1 using 100 training examples and 50 test examples from each category, the grid detector and patch based representation. **Right Panel.** Rank statistics of the confusion table, which shows the probability of a test scene correctly belonging to one of the top N highest probability categories. N ranges from 1 to 13.

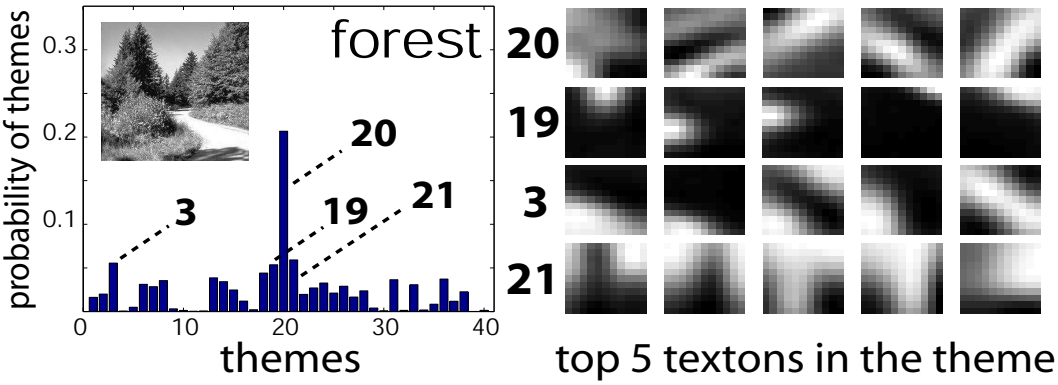


Figure 15.5: Example of themes for the forest category. **Left Panel** The distribution of all 40 themes. **Right Panel** The 5 most likely codewords for each of the 4 dominant themes in the category. Notice that codewords within a theme are visibly consistent. The "foliage" (#20, 3) and "tree branch" (#19) themes appear to emerge automatically from the data.

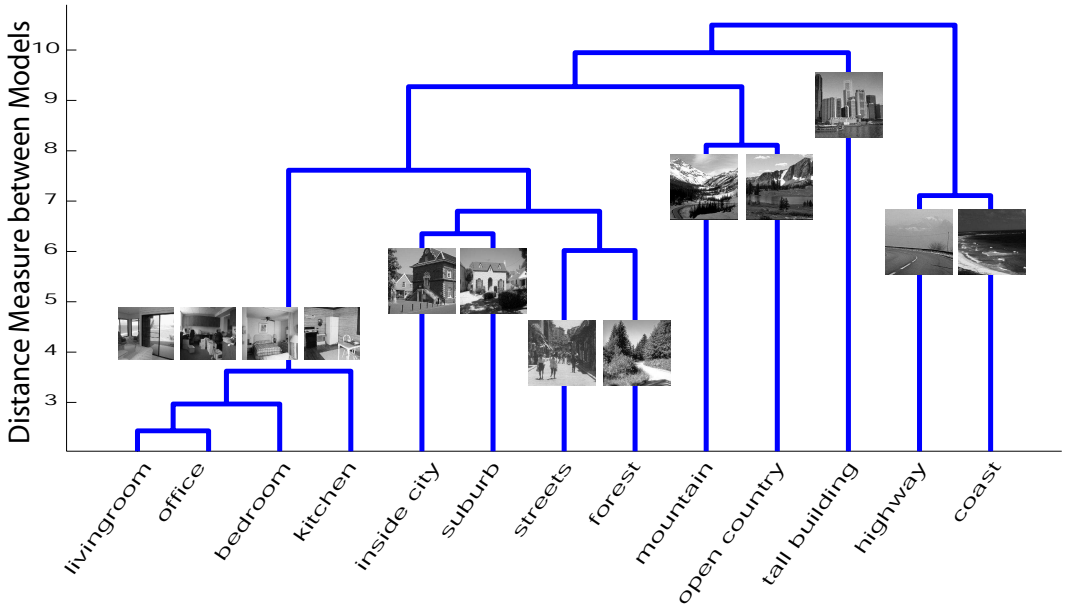


Figure 15.6: Dendrogram of the relationship of the 13 category models based on theme distribution. y-axis is the pseudo-euclidean distance measure between models.

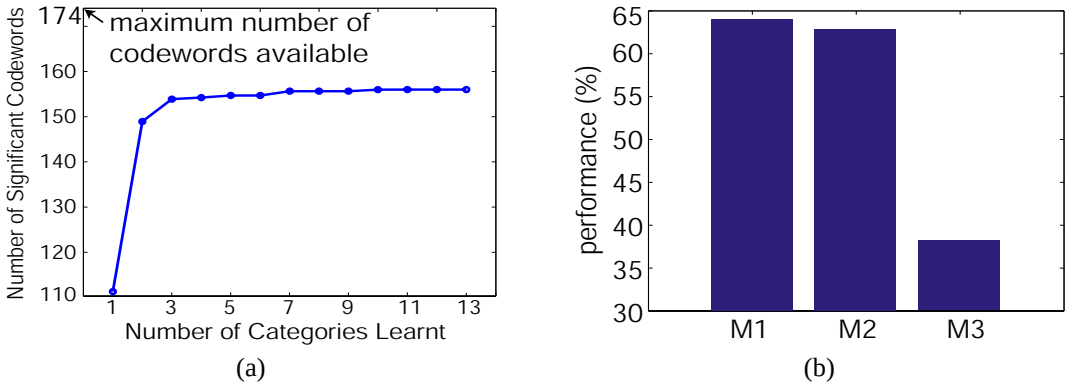


Figure 15.7: **(a)** Number of significant codewords as a function of the number of categories learnt. "Significance" is defined as 90% of the probabilistic weight. **(b)** Performance comparison among Theme Model 1 (M1), Theme Model 2 (M2) and the traditional texton model (M3, e.g. [147]).

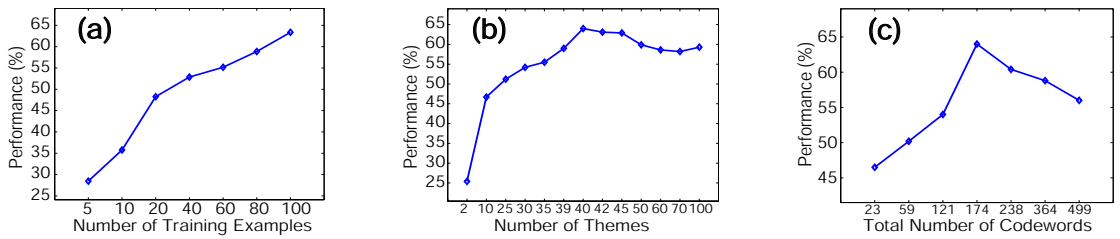


Figure 15.8: **(a)** Number of training examples vs. performance. **(b)** Number of themes vs. performance. **(c)** Number of codewords vs. performance. All performances are quotes from the mean of the confusion table.

the categories by looking at the model distances among them (see the dendrogram in Fig. 15.6). When the distribution of the themes are close, the categories would also be close to each other on the dendrogram. For example, the closest categories are the 4 indoor environments. Fig. 15.7(a) shows that by sharing the resources of codewords and intermediate themes, the number of significant codewords for learning more and more new models tends to level off quickly [137].

Fig. 15.8 illustrates 3 different aspects of the algorithm: performances versus the number of training examples (a), of themes (b) and of codewords in the codebook (c). Table 15.1 shows how different feature detection and representation influences the performance.

Chapter 16

Summary

We have proposed a Bayesian hierarchical model to learn and recognize natural scene categories. The model is an adaptation to vision of ideas proposed recently by Blei and collaborators [12] in the context of document analysis . While previous schemes [94,150] require a detailed manual annotation of the images in the training database, our model can learn characteristic intermediate "themes" of scenes with no supervision, nor human intervention, and achieves comparable performance to Vogel and Schiele (see Table 16.1 for details.).

Our model is based on a principled probabilistic framework for learning automatically the distribution of codewords and the intermediate-level themes, which might be thought to be akin to texture descriptions. Intermediate-level descriptions are shown to be useful [94]; Fig. 15.7(b) shows that indeed this model outperforms the traditional "texton models" where only a fixed codeword mixing pattern is estimated for each category of scenes [147]. One way to think about our model is as a generalization of the the "texton models" [76, 147] for textures, which need samples of "pure" texture to be trained. By contrast, our model may be trained on complete scenes and infer the intermediate "themes" from the data. In the future, it is important to further explore this relationship between the "themes" to meaningful textures such as the semantic concepts suggested by [94,150]. In addition, we provide a framework to share both the basic level codewords as well as

	# of categ.	training # per categ.	training requirements	perf. (%)
Theme Model 1	13	100	unsupervised	76
[150]	6	~ 100	human annotation of 9 semantic concepts for $60,000$ patches	77
[94]	8	$250 \sim 300$	human annotation of 6 properties for thousands of scenes	89

Table 16.1: Comparison of our algorithm with other methods. The average confusion table performances are for the 4 comparable categories (forest, mountain, open country and coast) in all methods. We use roughly $1/3$ of the number of training examples and no human supervision than [94]. Fig. 15.8(a) indicates that given more training examples, our model has the potential of achieve higher performances.

intermediate level themes amongst different scene categories. Similarly to what Torralba and colleagues [137] found, the number of features to be learnt increases sub-linearly as the number of new categories increases.

We tested our algorithm on a diverse set of scene types, introducing a number of new categories (13 here, as opposed to 4+4 in [94] and 6 in [150]). The lackluster performances for the indoor scenes suggest that our model is not complete. At a minimum, we need a richer set of features: by using different cues as well as a hierarchy of codewords, we might be able to form much more powerful models for these difficult categories.

Bibliography

[1] *Merriam-Webster's Collegiate Dictionary.* Merriam-Webster, Inc., Springfield, Massachusetts, USA, 10th edition, 1994.

[2] G.K. Aguirre, E. Zarahn, and D. Esposito. The variability of human, BOLD hemodynamic responses. *Neuron*, 21:373–383, 1998.

[3] T. Allison, A. Puce, D.D. Spencer, and G. McCarthy. Electrophysiological studies of human face perception. *Cereb. Cortex*, 9:415–430, 1999.

[4] Y. Amit and D. Geman. A computational model for visual selection. *Neural Computation*, 11(7):1691–1715, 1999.

[5] H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *15th conference on Uncertainty in Artificial Intelligence*, pages 21–30, 1999.

[6] M. Bar and S. Ullman. Spatial context in recognition. *Perception*, 25:343–352, 1996.

[7] J.R. Bergen and B. Julesz. Parallel versus serial processing in rapid pattern discrimination. *Nature*, 303:696–698, 1983.

[8] I. Biederman. Perceiving real-world scenes. *Science*, 177:77–80, 1972.

[9] I. Biederman. On the semantics of a glance at a scene. In M. Kubovy and J.R. Pomerantz, editors, *Perceptual Organization*, pages 213–254. Erlbaum, Hillsdale, N.J., 1981.

[10] I. Biederman. Recognition-by-Components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[11] I. Biederman, R.C. Teitelbaum, and R.J. Mezzanotte. Scene perception: a failure to find a benefit from prior expectancy or familiarty. *J. of Exp. Psychol.*, 9(3):411–429, 1983.

[12] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[13] S.J. Boyce, A. Pollatsek, and K. Rayner. Effect of background information on object identification. *J. Exp. Psychol. Hum. Perc. and Perf.*, 15(3):556–566, 1989.

[14] D.H. Brainard. The Psychophysics Toolbox. *Spatial Vision*, 10:433–436, 1997.

[15] J. Braun. Visual search among items of different salience: Removal of visual attention mimics a lesion in extrastriate area v4. *J. Neuroscience*, 14:554–567, 1994.

[16] J. Braun. Vision and attention: the role of training. *Nature*, 393:424–425, 1998.

[17] J. Braun and B. Julesz. Withdrawing attention at little or no cost: Detection and discrimination tasks. *Perception & Psychophysics*, 60(1):1–23, 1998.

[18] J. Braun, J. Wen, and C. Koch. Visual attention is undifferentiated: Concurrent discrimination of shape, color and motion. *Investigative Ophthalmology & Visual Science*, 38:5455–5455, 1997.

[19] M. Burl and P. Perona. Recognition of planar object classes. In *Proc. Computer Vision and Pattern Recognition*, pages 223–230, 1996.

[20] M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *Proc. European Conference on Computer Vision*, pages 628–641, 1996.

[21] L.L. Chao, A. Martin, and J.V. Haxby. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nat. Neurosci.*, 2:913–919, 1999.

[22] A. Delorme, G. Richard, and M. Fabre-Thorpe. Rapid categorization of natural scenes is color blind: A study in monkeys and humans. *Vision Research*, 40(16):2187–2200, 2000.

[23] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 29:1–38, 1976.

[24] P. Downing, Y. Jiang, M. Shuman, and N. Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293:2470–2473, 2001.

[25] J. Driver and G.C. Baylis. Edge-assignment and figure-ground segmentation in short-term visual matching. *Cognitive Psychology*, 31(3):248–306, 1996.

[26] J. Dunai, U. Castiello, and Y. Rossetti. Attentional processing of colour and location cues. *Exp. Brain Res.*, 138(4):520–526, 2001.

[27] R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

[28] D.C. Van Essen. Functional organization of primate visual cortex. In E.G. Jones and A. Peters, editors, *Cerebral Cortex*, volume 3, pages 259–329. Plenum Press, 1985.

[29] M. Fabre-Thorpe, A. Delorme, C. Marlot, and S. Thorpe. A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *J. Cognitive Neurosci.*, 13(2):171–180, 2001.

[30] M.J. Farah. Dissociable systems for visual recognition: A cognitive neuropsychology approach. In S.M. Kosslyn and D.N. Osherson, editors, *Visual cognition: An invitation to cognitive science*, pages 101–119. MIT Press, Cambridge, MA, 1995.

[31] M.J. Farah, K.D. Wilson, M. Drain, and J.N. Tanaka. What is "special" about face perception? *Psychol. Rev.*, 105:482–498, 1998.

[32] L. Fei-Fei, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-Shot learning of object categories. In *Proceedings of the 9th International Conference on Computer Vision, Nice, France*, pages 1134–1141, October 2003.

[33] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *Workshop on Generative-Model Based Vision*, 2004.

[34] L. Fei-Fei, R. Fergus, and P. Perona. One-Shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, in press.

[35] L. Fei-Fei and P. Perona. A Bayesian hierarchy model for learning natural scene categories. *CVPR*, 2005.

[36] L. Fei-Fei, R. VanRullen, C. Koch, and P. Perona. Why does natural scene recognition require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition*, in press.

[37] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 1:55–79, 2005.

[38] P. Felzenszwalb and D. Huttenlocher. Representation and detection of deformable shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, To appear.

[39] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. Computer Vision and Pattern Recognition*, pages 264–271, 2003.

[40] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for Google images. In *Proc. 8th European Conf. on Computer Vision*, 2004.

[41] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *Proc. Computer Vision and Pattern Recognition*, 2005.

[42] D. Forsyth and A. Zisserman. Shape from shading in the light of mutual illumination. *Image and Vision Computing*, pages 42–29, 1990.

[43] D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291:312–316, 2001.

[44] A. Friedman. Frame pictures: the role of knowledge in automatized encoding and memory for gist. *J. of Exp. Psychol. General*, 108(3):316–355, 1979.

[45] A. Gelman, J.B. Carlin, Stern H.S., and Rubin D.B. *Bayesian Data Analysis*. Chapman Hall/CRC, 1995.

[46] F. Germeys and G. d'Ydewalle. Revisiting scene primes for object locations. *Quart. J. Exp. Psychol.*, 54A(3):683–693, 2001.

[47] R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman Hall, 1992.

[48] R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41:337–348, 1992.

[49] M. Gorkani and R. Picard. Texture orientation for sorting photos at a glance. In *Int. Conf. on Pattern Recognition*, 1994.

[50] R. Gottsdanker. The ubiquitous role of preparation. In G.E. Stelmach and J. Requin, editors, *Tutorials in motor behavior*, pages 355–371. Elsevier, North-Holland, Amsterdam, 1980.

[51] P. De Graef, D. Christiaens, and G. d'Ydewalle. Perceptual effects of scene context on object identification. *Psychol. Res.*, 52:317–329, 1990.

[52] K. Grill-Spector and N. Kanwisher. Visual recognition: as soon as you know it is there, you know what it is. *Psychol. Sci.*, in press.

[53] W. Grimson and D. Huttenlocher. On the sensitivity of the Hough transform for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-12(3):255–274, 1990.

[54] J. Haxby, M.I. Gobbini, M.L. Furey, A. Ishai, J.L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2425–2430, 2001.

[55] H.J. Heinze, G.R. Mangun, W. Burchert, H. Hinrichs, M. Scholz, T.F. Munte, A. Gos, M. Scherg, S. Johannes, H. Hundeshagen, M.S. Gazzaniga, and S.A. Hillyard. Combined spatial and temporal imaging of brain activity during visual selective attention in humans. *Nature*, 372:543–546, 1994.

[56] J.M. Henderson and A. Hollingworth. High-level scene perception. *Annu. Rev. Psychol.*, 50(243-271), 1999.

[57] H. Heuer. Intermanual interactions during simultaneous execution and programming of finger movements. *J. Motor Behav.*, 17:335–354, 1985.

[58] O. Hikosaka, S. Miyauchi, and S. Shimojo. Focal visual attention produces illusory temporal order and motion sensation. *Vision Res.*, 33:1219–1240, 1992.

[59] S. Hochstein and M. Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36:791–804, 2002.

[60] A. Hollingworth and J.M. Henderson. Object identification is isolated from scene semantic constraint: evidence from object type and token discrimination. *Acta Psychol (Amst).*, 102(2-3):319–343, 1999.

[61] A. Hollingworth and J.M. Henderson. Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition: Special issue on change detection and visual memory*, 7:213–235, 2000.

[62] K. Humphreys and M. Titterington. Some examples of recursive variational approximations for Bayesian inference. In M. Opper and D. Saad, editors, *Advanced mean field methods*. MIT Press, 2001.

[63] D. Huttenlocher, G. Klanderman, and W. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

[64] J.S. Johnson and B.A. Olshausen. Timecourse of neural signatures of object recognition. *J. of Vis.*, 3:499–512, 2003.

[65] J.S. Joseph, M.M. Chun, and K. Nakayama. Reply to Braun. *Nature*, 387:805–808, 1998.

[66] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[67] N. Kanwisher. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.*, 17:4302–4311, 1997.

[68] C. Keysers, D.K. Xiao, P. Foldiak, and D.I. Perrett. The speed of sight. *J. Cognitive Neurosci.*, 13(1):90–101, 2001.

[69] A. Kingstone. Combining expectancies. *Q. J. Exp. Psychol.*, 44:69–104, 1992.

[70] S.M. Kosslyn, R.A. Flynn, J.B. Amsterdam, and G. Wang. Components of high-level vision: A cognitive neuroscience analysis and accounts of neurological syndromes. *Cognition*, 34:203–277, 1990.

[71] G. Kreiman, C. Koch, and I. Fried. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat. Neurosci.*, 3:946–953, 2000.

[72] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[73] Y. LeCun, F. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proc. CVPR*, 2004.

[74] D.K. Lee, C. Koch, and J. Braun. Attentional capacity is undifferentiated: Concurrent discrimination of form, color, and motion. *Perc. & Psych.*, 61(7):1241–1255, 1999.

[75] T. Leung, M. Burl, and P. Perona. Finding faces in cluttered scenes using labeled random graph matching. In *Proc. International Conference on Computer Vision*, pages 637–644, 1995.

[76] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, June 2001.

[77] F.F. Li, R. VanRullen, C. Koch, and P. Perona. Rapid natural scene categorization in the near absence of attention. *Proc. Nat. Acad. Sci.*, 99(14):9596–9601, 2002.

[78] N.K. Logothetis and D.L. Sheinberg. The role of temporal cortical areas in perceptual organization. *Proc. Nat. Acad. Sci.*, 94:3408–3413, 1997.

[79] D. Lowe. Object recognition from local scale-invariant features. In *Proc. International Conference on Computer Vision*, pages 1150–1157, 1999.

[80] S.J. Luck, L. Chelazzi, S.A. Hillyard, and R. Desimone. Neural mechanisms of spatial attention in areas V1, V2 and V4 of macaque visual cortex. *J. Neurophysiol.*, 77:24–42, 1997.

[81] S. Lumet. The Pawnbroker. Film, 1965.

[82] A. Mack and I. Rock. *Inattentional Blindness.* MIT Press, Cambridge, MA, 1998.

[83] J. Malik and P. Perona. Preattentive texture discrimination with early vision mechanisms. *J. Opt. Soc. Am. A.*, 7(5):923–932, 1990.

[84] J.M. Mandler and R.E. Parker. Memory for descriptive and spatial information in complex pictures. *J. of Exp. Psychol.*, 2(1):38–48, 1976.

[85] R.L. Metzger and J.R. Antes. The nature of processing early in picture perception. *Psychol. Res.*, 45:267–274, 1983.

[86] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. European Conference on Computer Vision*, volume 1, pages 128–142, 2002.

[87] T. Minka. Using lower bounds to approximate integrals. `http://www.stat.cmu.edu/~minka/papers/rem.html`, 2001.

[88] T. Minka and J. Lafferty. Expectation-Propagation for the generative aspect model. In *Proc. of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 2002.

[89] M. Mishkin, L.G. Ungerleider, and K.A. Macko. Object vision and spatial vision: Two cortical pathways. *Trends Neurosci.*, 6:414–417, 1983.

[90] K. Nakayama, Z.J. He, and S. Shimojo. Visual surface representation: a critical link between lower-level and higher-level vision. In S.M. Kosslyn and D.N. Osherson, editors, *An invitation to cognitive science: visual cognition.* MIT, Cambridge, Massechussetts, 1995.

[91] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variants. In M.I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer academic press, Norwell, 1998.

[92] A. Netick and S.T. Klapp. Hesitations in manual tracking: a single-channel limit in response programming. *J. Exp. Psychol. Human*, 20:766–782, 1994.

[93] A. Oliva and P.G. Schyns. Colored diagnostic blobs mediate scene recognition. *Cognitive Psychology*, 41:176–210, 2000.

[94] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. Journal of Computer Vision.*, 42, 2001.

[95] B.A. Olshausen and D.J. Field. Natural image statistics and efficient coding. *Network-Comp. Neural*, 7(2):333–339, 1996.

[96] J.K. O'Regan, R.A. Rensink, and J.J. Clark. Change blindness as a result of mud splashes. *Nature*, 398:34, 1999.

[97] A.L. Ostergaard and J.B. Davidoff. Some effects of color on naming and recognition of objects. *J. of Experimental Psychology: Learning, Memory and Cognition*, 11:579–587, 1985.

[98] S.E. Palmer. Visual perception and world knowledge: notes on a model of sensory-cognitive interaction. In D.A. Norman and D.E. Rumelhart, editors, *Explorations in Cognition*, pages 279–307. LNR Res. Group, San Francisco, 1975.

[99] H. Pashler. Processing stages in overlapping tasks: Evidence for a central bottleneck. *J. Exp. Psychol. Hum. Perc. and Perf.*, 10:358–377, 1984.

[100] H. Pashler. *The Psychology of Attention.* MIT Press, 1998.

[101] D.G. Pelli. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10:437–442, 1997.

[102] W. Penny. Variational bayes for d-dimensional Gaussian mixture models. Tech. rep., University College London, 2001.

[103] M.A. Peterson and B.S. Gibson. Shape recognition contributions to figure-ground organization in three-dimensional display. *Cognitive Psychology*, 25:383–429, 1993.

[104] M.A. Peterson and B.S. Gibson. Must shape recognition follow figure-ground organization? an assumption in peril. *Psychological Science*, 5:253–259, 1994.

[105] M.A. Peterson and J.H. Kim. On what is bound in figures and grounds. *Visual Cognition.*, 8:329–348, 2001.

[106] J. Portilla and E.P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *IJCV*, 40(1):49–70, October 2000.

[107] M.I. Posner, C.R.R. Snyder, and B.J. Davidson. Attention and the detection of signals. *J. Exp. Psychol. General*, 109:160–174, 1980.

[108] M.C. Potter. Short-term conceptual memory for pictures. *J. of Exp. Psychol. Hum. Learning and Mem.*, 2:509–522, 1976.

[109] M.C. Potter and E.I.J. Levy. Recognition memory for a rapid sequence of pictures. *Exp Psychol.*, 81(1):10–15, 1969.

[110] M.C. Potter, A. Staub, J. Rado, and D.H. O'Connor. Recognition memory for briefly presented pictures: the time course of rapid forgetting. *J. Exp. Psychol: Hum. Perc. and Perf.*, 28(5):1163–1175, 2002.

[111] C.J. Price and G.W. Humphreys. The effects of surface detail on object categorization and naming. *The Quarterly Journal of Experimental Psychology*, 41(A):797–828, 1989.

[112] A.M. Proverbio and G.R. Mangun. Electrophysiological and behavioural "costs" and "benefits" during sustained visual-spatial attention. *Int. J. Neurosci.*, 79:221–233, 1994.

[113] K. Rayner. Visual selection in reading, picture perception, and visual search: A tutorial review. In H. Bouma and D. Bouwhuis, editors, *Attention and performance X*. Erlbaum, Hillsdale, N.J., 1984.

[114] L. Reddy, P. Wilken, and C. Koch. Face-Gender discrimination is possible in the near absence of attention. *Journal of Vision*, 4(2):106–117, 2004.

[115] R.A. Rensink, J.K. O'Regan, and J.J. Clark. To see or not to see: the need for attention to perceive changes in scenes. *Psychol. Sci.*, 8(5):368–373, 1997.

[116] T. Ro, C. Russell, and N. Lavie. Changing faces: A detection advantage in the flicker paradigm. *Psychol. Sci.*, 12:94–99, 2001.

[117] E. Rosch. Principles of categorization. In E. Rosch and B.B. Lloyd, editors, *Cognition and categorization*. Erlbaum, Hillsdale, N.J., 1978.

[118] B. Rossion and I. Gauthier. How does the brain process upright and inverted faces? *Behavioral and Cognitive Neuroscience Reviews*, 1(1):63–75, 2002.

[119] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3D object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *Proc. Computer Vision and Pattern Recognition*, pages 272–280, 2003.

[120] G. Rousselet, M. Fabre-Thorpe, and S. Thorpe. Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, 5:629–630, 2002.

[121] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(1):23–38, 1998.

[122] E. Rubin. Figure and ground. In D.C.W. Beardslee, editor, *Readings in perception.* Van Nostrand, New York, 1958.

[123] Y. Rubner, C. Tomasi, and L.J. Guibas. Adaptive color-image embeddings for database navigation. *Proc. of Asian Conf. on Comp. Vision*, pages 104–111, 1998.

[124] E. Sali and S. Ullman. Combining class-specific fragments for object classification. In *Proc. British Machine Vision Conference*, volume 1, pages 203–213, 1999.

[125] H. Schneiderman and T. Kanade. A statistical approach to 3D object detection applied to faces and cars. In *Proc. Computer Vision and Pattern Recognition*, pages 746–751, 2000.

[126] K. Shapiro, editor. *The limits of attention.* Oxford University Press, Oxford, 2001.

[127] R.M. Shiffrin and W. Schneider. Controlled and automatic human information processing II: Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84:127–188, 1977.

[128] D.J. Simons and D.T. Levin. Change blindness. *Trends Cog. Sci.*, 1:261–267, 1997.

[129] G. Sperling and B. Dosher. Strategy and optimization in human information processing. In K.R. Boff, L. Kaufman, and J.P. Thomas, editors, *Handbook of perception and human performance.*, pages 1–65. Wiley, New York, 1986.

[130] S. Subramaniam, I. Biederman, and S. Madigan. Accurate identification but no priming and chance recognition memory for pictures in RSVP sequences. *Vis. Cogn.*, 7(4):511–535, 2000.

[131] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(1):39–51, 1998.

[132] M. Szummer and R. Picard. Indoor-outdoor image classification. In *Int. Workshop on Content-based Access of Image and Vedeo Databases*, Bombay, India, 1998.

[133] K. Tanaka. Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139, 1996.

[134] C.W. Telford. The refractory phase of voluntary and associative responses. *J. Exp. Psychol.*, 14:1–36, 1931.

[135] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

[136] S. Thorpe, K.R. Gegenfurtner, M. Fabre-Thorpe, and H.H. Bulthoff. Detection of animals in natural images using far peripheral vision. *Eur. J. Neurosci.*, 14:869–876, 2001.

[137] A. Torralba, K. Murphy, and W.T. Freeman. Sharing features: efficient boosting procedures for multi-class object detection. In *Proc. of the 2004 IEEE CVPR.*, 2004.

[138] A. Torralba and A. Oliva. Statistics of natural images categories. *Network: Computation in Neural Systems*, 14:391–412, 2003.

[139] A. Treisman. The perception of features and objects. In A. Baddeley and L. Weiskrantz, editors, *Attention: selection, awareness and control. A tribute to Donald Broadbent.*, pages 5–35. Clarendon University Press, Oxford, 1993.

[140] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cogn. Psychol.*, 12:97–136, 1980.

[141] N.F. Troje and H.H. Bulthoff. Face recognition under varying poses: the role of texture and shape. *Vision Res.*, 36:1761–1771, 1996.

[142] B. Tversky and K. Hemenway. Categories of environmental scenes. *Cog. Psychol.*, 15:121–149, 1983.

[143] L.G. Ungerleider and M. Mishkin. Two cortical visual systems. In D.J. Ingle, M.A. Goodale, and R.J.W. Mansfield, editors, *Analysis of Visual Behavior.*, pages 549–586. MIT press, Cambridge, Mass., 1982.

[144] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image classification for content-based indexing. *IEEE Trans. on Image Processing*, 10, 2001.

[145] R. VanRullen and C. Koch. Competition and selection during visual processing of natural scenes and objects. *J. of Vis.*, 3:75–85, 2003.

[146] R. VanRullen and S.J. Thorpe. The time course of visual processing: from early perception to decision-making. *J. Cognitive Neurosci.*, 13(4):454–461, 2001.

[147] M. Varma and A. Zisserman. Texture classification: are filter banks necessary? In *CVPR03*, pages II: 691–698, 2003.

[148] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001.

[149] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. International Conference on Computer Vision*, pages 734–741, 2003.

[150] J. Vogel and B. Schiele. A semantic typicality measure for natural scene categorization. In *DAGM'04 Annual Pattern Recognition Symposium*, Tuebingen, Germany, 2004.

[151] Q. Wang, P. Cavanagh, and M. Green. Familiarity and pop-out in visual search. *Perception & Psychophysics*, 56:495–500, 1994.

[152] M. Weber, W. Einhaeuser, M. Welling, and P. Perona. Viewpoint-invariant learning and detection of human heads. In *Proc. 4th Int. Conf. Autom. Face and Gesture Rec.*, pages 20–27, 2000.

[153] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. European Conference on Computer Vision*, volume 2, pages 101–108, 2000.

[154] A.T. Welford. The psychological refractory period and the timing of high-speed performance–a review and a theory. *Brit. J. Psychol.*, 43:2–19, 1952.

[155] J. Winn. *Variational Message Passing and its applications.* PhD thesis, University of Cambridge, 2003.

[156] J.M. Wolfe. Visual memory: what do you know about what you saw? *Curr. Bio.*, 8:R303–R304, 1998.

[157] J.M. Wolfe. Visual search. In H. Pashler, editor, *Attention.*, pages 13–74. Psychology Press Ltd., 1998.

Appendix A

A.1 Definition of Various Densities and Functions

A.1.1 Dirichlet Distribution

A Dirichlet, with variables $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_\Omega\}$ and parameters $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_\Omega\}$:

$$\mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\lambda}) = \frac{\Gamma(\sum_{\omega=1}^{\Omega} \lambda_\omega)}{\prod_{\omega=1}^{\Omega} \Gamma(\lambda_\omega)} \prod_{\omega=1}^{\Omega} \pi_\omega^{\lambda_\omega - 1} \tag{A.1}$$

where $\pi_1, \ldots, \pi_\Omega \geq 0$ and $\lambda_1, \ldots, \lambda_\Omega > 0$ and $\sum_{\omega=1}^{\Omega} \pi_\omega = 1$. The Dirichlet distribution is the conjugate to the multinomial density.

A *symmetric* Dirichlet density has all the λ's the same value, e.g., λ_0, and the density reduces to:

$$\mathcal{D}(\boldsymbol{\pi}|\lambda_0) = \frac{\Gamma(\Omega\lambda_0)}{\Gamma(\lambda_0)^\Omega} \prod_{\omega=1}^{\Omega} \pi_\omega^{\lambda_0 - 1} \tag{A.2}$$

A.1.2 Normal-Wishart Distribution

A Normal-Wishart density is the conjugate density for the parameters of a normal distribution. We first condition on the Σ:

$$p(\boldsymbol{\mu}, \Sigma) = p(\boldsymbol{\mu}|\Sigma)p(\Sigma) \tag{A.3}$$

then we model the $\boldsymbol{\mu}$ term with a normal density and the Σ term with a Wishart:

$$p(\boldsymbol{\mu}|\Sigma) = \mathcal{G}(\boldsymbol{\mu}|\mathbf{m}, \frac{1}{\beta}\Sigma) \tag{A.4}$$

where $\mathbf{m}, \beta$ are hyperparameters.

Technically, since Σ is a covariance matrix rather than a precision matrix, we actually use an inverse-Wishart density:

$$p(\Sigma) = \mathcal{W}^{-1}(\Sigma|a, B) = \frac{1}{\Gamma_d(a/2)|\Sigma|^{(a+d+1)/2}} \left|\frac{B}{2}\right|^{a/2} \exp\left(-\frac{1}{2}Tr(\Sigma^{-1}B)\right) \tag{A.5}$$

where d is the dimensionality of Σ, a, B are hyperparameters, with a being the number of degrees of freedom in Σ and:

$$\Gamma_d(n/2) = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma((n+1-i)/2) \tag{A.6}$$

A Wishart density (as supposed to an inverse-Wishart) has the following form:

$$\mathcal{W}(\Gamma|a,B) = \frac{|\Gamma|^{(a+d+1)/2}\left|\frac{B}{2}\right|^{a/2}\exp\left(-\frac{1}{2}Tr(\Gamma B)\right)}{\Gamma_d(a/2)} \tag{A.7}$$

A.1.3 Gamma Distribution

A Gamma density with variable x and parameters b, c is:

$$\Gamma(x|b,c) = \frac{1}{\Gamma(c)}\frac{x^{c-1}}{b^c}\exp\left(\frac{-x}{b}\right) \tag{A.8}$$

where $x \geq 0$ and $\Gamma(c)$ is the Gamma *function*. The mean of the density is bc and the variance b^2c.

A.1.4 Multivariate Student's T Distribution

A d dimensional Student's T distribution having the vector $\mathbf{m}$ as the variable, k degrees of freedom and parameters $\mathbf{b}, C$ has the form:

$$\mathcal{S}(\mathbf{m}|k,\mathbf{b},C) = \frac{1}{\Delta(\frac{1}{2}\mathbf{1},\frac{1}{2}(k-d+1))\,|kC|^{1/2}\,(1+(\mathbf{m}-\mathbf{b})^T(kC)^{-1}(\mathbf{m}-\mathbf{b}))^{(k+1)/2)}} \tag{A.9}$$

where $\mathbf{1}$ is a vector of ones, of length d; C must be a positive definite matrix and $k > d-1$. $\Delta()$ is a Dirichlet function:

$$\Delta(\mathbf{g},h) = \frac{\Gamma(g_1)\ldots\Gamma(g_d)\Gamma(h)}{\Gamma(\sum_i g_i + h)} \tag{A.10}$$

A.1.5 Kullback-Leibler Distance

Given two p.d.f's $p()$ and $q()$ that exist over the same space, x, the Kullback-Leibler distance between the two distributions is:

$$\mathcal{KL}(p||q) = \sum_x p(x)\,\log\frac{p(x)}{q(x)} \tag{A.11}$$

Note that $\mathcal{KL}(p||q) \geq 0$ always holds.

A.1.6 Digamma Function

The Digamma function is defined as:

$$\mathbf{\Psi}(z) \equiv \frac{d\log\Gamma(z)}{dz} = \frac{\Gamma'(z)}{\Gamma(z)} \tag{A.12}$$

A.2 Learning using a conjugate density parameter posterior

Recall the mixture of constellation models from Eq. 17 in [34]:

$$p(\mathcal{X},\mathcal{A}|\boldsymbol{\theta}) = \sum_{\omega=1}^{\Omega} p(\omega|\boldsymbol{\pi}) \sum_{h=1}^{|H|} p(\mathcal{X}_h|\boldsymbol{\mu}_\omega^{\mathcal{X}},\boldsymbol{\Gamma}_\omega^{\mathcal{X}})p(\mathcal{A}_h|\boldsymbol{\mu}_\omega^{\mathcal{A}},\boldsymbol{\Gamma}_\omega^{\mathcal{A}}) \tag{A.13}$$

Each component ω has a mixing coefficient π_ω; a mean of shape and appearance $\boldsymbol{\mu}_\omega^{\mathcal{X}}, \boldsymbol{\mu}_\omega^{\mathcal{A}}$; and a precision matrix of shape and appearance $\boldsymbol{\Gamma}_\omega^{\mathcal{X}}, \boldsymbol{\Gamma}_\omega^{\mathcal{A}}$. The $\mathcal{X}$ and $\mathcal{A}$ superscripts denote shape and appearance terms, respectively. Collecting all mixture components and their corresponding parameters together, we obtain an

overall parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}^{\mathcal{X}}, \boldsymbol{\mu}^{\mathcal{A}}, \boldsymbol{\Gamma}^{\mathcal{X}}, \boldsymbol{\Gamma}^{\mathcal{A}}\}$. Assuming we have now learnt the model distribution $p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t)$ from a set of training data $\mathcal{X}_t$ and $\mathcal{A}_t$, we define the model distribution in the following way:

$$p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t) = p(\boldsymbol{\pi}) \prod_{\omega} p(\boldsymbol{\mu}_{\omega}^{\mathcal{X}}|\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) p(\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) p(\boldsymbol{\mu}_{\omega}^{\mathcal{A}}|\boldsymbol{\Gamma}_{\omega}^{\mathcal{A}}) p(\boldsymbol{\Gamma}_{\omega}^{\mathcal{A}}) \tag{A.14}$$

where the mixing component is a symmetric Dirichlet: $p(\boldsymbol{\pi}) = \mathcal{D}ir(\lambda_{\omega}\boldsymbol{I}_{\Omega})$, the distribution over the shape precisions is a Wishart $p(\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) = \mathcal{W}(\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}|a_{\omega}^{\mathcal{X}}, \boldsymbol{B}_{\omega}^{\mathcal{X}})$ and the distribution over the shape mean conditioned on the precision matrix is Normal: $p(\boldsymbol{\mu}_{\omega}^{\mathcal{X}}|\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) = \mathcal{G}(\boldsymbol{\mu}_{\omega}^{\mathcal{X}}|\boldsymbol{m}_{\omega}^{\mathcal{X}}, \beta_{\omega}^{\mathcal{X}}\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}})$. Together the shape distribution $p(\boldsymbol{\mu}_{\omega}^{\mathcal{X}}, \boldsymbol{\Gamma}_{\omega}^{\mathcal{X}})$ is a Normal-Wishart density [5, 102]. Note $\{\lambda_{\omega}, a_{\omega}, \boldsymbol{B}_{\omega}, \boldsymbol{m}_{\omega}, \beta_{\omega}\}$ are hyper-parameters for defining their corresponding distributions of model parameters. Identical expressions apply to the appearance component in Eq. A.14.

A.2.1 Variational methods

We first review briefly the variational method for model learning. We have some integral we wish to evaluate: $F = \int_{\theta} f(\theta)\, d\theta$. We write $f(\theta)$ as a function of its parameters and some hidden variables, S: $f(\theta) = \int_S g(\theta, S)\, dS$. Applying Jensen's inequality to give us a lower bound on the integral, we get:

$$\begin{aligned} F &= \int_{\theta,S} g(\theta, S)\, dS\, d\theta \\ &\geq \exp\left(\int_{\theta,S} q(\theta, S) \log \frac{g(\theta, S)}{q(\theta, S)}\, dS\, d\theta \right) \quad \text{provided} \int_{\theta,S} q(\theta, S)\, dS\, d\theta = 1 \end{aligned} \tag{A.15}$$

Variational Bayes makes the assumption that $q(\theta, S)$ is a probability density function that can be factored into $q_{\theta}(\theta)q_S(S)$. We then iteratively optimize q_{θ} and q_S using expectation maximization (EM) to maximize the value of the lower bound to the integral (see [87, 102]). If we consider $g(\theta, S)$ the "true" p.d.f., by using the above method, we are effectively decreasing the Kullback-Leibler distance between $g(\theta, S)$ and $q(\theta, S)$, hence obtaining a $q(\theta, S)$ that approximates the true p.d.f.

A.2.2 Variational Bayesian EM

Recall that we have a mixture model with Ω components. Collecting all mixture components and their corresponding parameters together, we have an overall parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}^{\mathcal{X}}, \boldsymbol{\mu}^{\mathcal{A}}, \boldsymbol{\Gamma}^{\mathcal{X}}, \boldsymbol{\Gamma}^{\mathcal{A}}\}$. For n training images, we have $\{\mathcal{X}_t^n, \mathcal{A}_t^n\}$ with $n = 1...N$. In the constellation model, each image n has $|H^n|$ hypotheses, each one of which picks out P features from $\{\mathcal{X}^n, \mathcal{A}^n\}$ to give $\{\mathcal{X}_h^n, \mathcal{A}_h^n\}$. We have two latent variables, the hypothesis $\boldsymbol{h}$ and the mixture component $\boldsymbol{\omega}$. We assume that the prior on any hypothesis always remains uniform, namely $1/|H^n|$, so it is omitted from the update equations since it is constant. We can now express the likelihood of an image n as:

$$p(\mathcal{X}^n, \mathcal{A}^n|\boldsymbol{\theta}) = \sum_{\omega=1}^{\Omega} \sum_{h=1}^{|H^n|} p(\omega_n = \omega|\boldsymbol{\pi}) p(\mathcal{X}_h^n|\boldsymbol{\mu}_{\omega}^{\mathcal{X}}, \boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) p(\mathcal{A}_h^n|\boldsymbol{\mu}_{\omega}^{\mathcal{A}}, \boldsymbol{\Gamma}_{\omega}^{\mathcal{A}}) \tag{A.16}$$

where $p(\boldsymbol{\omega} = \omega|\boldsymbol{\pi}) = \pi_{\omega}$. Both the terms involving $\mathcal{X}, \mathcal{A}$ above have a Normal form. The prior on the model parameters has the same form as the model distribution in Eq. A.14:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_{\omega} p(\boldsymbol{\mu}_{\omega}^{\mathcal{X}}|\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) p(\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) p(\boldsymbol{\mu}_{\omega}^{\mathcal{A}}|\boldsymbol{\Gamma}_{\omega}^{\mathcal{A}}) p(\boldsymbol{\Gamma}_{\omega}^{\mathcal{A}}) \tag{A.17}$$

where the mixing prior is $p(\boldsymbol{\pi}) = \mathcal{D}ir(\lambda_0\boldsymbol{I}_{\Omega})$, and the shape prior is a Normal-Wishart distribution $p(\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}})p(\boldsymbol{\mu}_{\omega}^{\mathcal{X}}|\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) = \mathcal{G}(\boldsymbol{\mu}_{\omega}^{\mathcal{X}}|\boldsymbol{m}_0^{\mathcal{X}}, \beta_0^{\mathcal{X}}\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}})\mathcal{W}(\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}|a_0^{\mathcal{X}}, \boldsymbol{B}_0^{\mathcal{X}})$. Identical expressions apply to the appearance component of Eq. A.17.

A.2.3 The E-step of Bayesian One-Shot

The central idea of Bayesian One-Shot is to approximate the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{h}|\mathcal{X}, \mathcal{A})$ by an optimal approximation $q(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{h})$ that is factorisable $q(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{h}) = q(\boldsymbol{\theta})q(\boldsymbol{\omega}, \boldsymbol{h})$, where $\boldsymbol{\omega}$ and $\boldsymbol{h}$ are hidden variables while $\boldsymbol{\theta}$ is the actual model parameter. In the E-step of Bayesian One-Shot, $q(\boldsymbol{\omega}, \boldsymbol{h})$ is updated according to:

$$q(\boldsymbol{\omega}, \boldsymbol{h}) \propto \exp\left[I(\boldsymbol{\omega}, \boldsymbol{h})\right] \qquad \text{where} \quad I(\boldsymbol{\omega}, \boldsymbol{h}) =< \log p(\mathcal{X}_{\sqcup}, \mathcal{A}_{\sqcup}, \boldsymbol{\omega}, h|\boldsymbol{\theta}) >_{\boldsymbol{\theta}} \tag{A.18}$$

and the expectation is taken w.r.t. $q(\boldsymbol{\theta})$ [102]. $I(\boldsymbol{\omega}, \boldsymbol{h})$ can be further written as:

$$I(\boldsymbol{\omega}, \boldsymbol{h}) = \langle \log p(\mathcal{X}|\boldsymbol{\omega}, \boldsymbol{h}, \boldsymbol{\theta}) p(\mathcal{A}|\boldsymbol{\omega}, \mathcal{X}, \boldsymbol{h}, \boldsymbol{\theta}) p(\omega|\boldsymbol{\theta}) p(\boldsymbol{h}) \rangle_{\boldsymbol{\theta}} \tag{A.19}$$

If we define $I(\boldsymbol{\omega}, \boldsymbol{h})$ for each image n, each mixture component ω and each hypothesis h as $\tilde{\gamma}^n_{\omega,h}$, the indicator posterior, we then have the update rule:

$$\tilde{\gamma}^n_{\omega,h} = \tilde{\pi}_\omega \tilde{\gamma}_\omega(\mathcal{X}^n_h) \cdot \tilde{\gamma}_\omega(\mathcal{A}^n_h) \tag{A.20}$$

where:

$$\log(\tilde{\pi}_\omega) = \Psi(\lambda_\omega) - \Psi(\sum_{\omega'} \lambda_{\omega'}) \tag{A.21}$$

$$\tilde{\gamma}_\omega(\mathcal{X}^n_h) = \exp\left[-\frac{1}{2}(\mathcal{X}^n_h - \boldsymbol{m}^{\mathcal{X}}_\omega)^T \bar{\boldsymbol{\Gamma}}^{\mathcal{X}}_\omega (\mathcal{X}^n_h - \boldsymbol{m}^{\mathcal{X}}_\omega)\right] \cdot (\tilde{\Gamma}^{\mathcal{X}}_\omega)^{1/2} \exp\left[\frac{-d^{\mathcal{X}}}{2\beta^{\mathcal{X}}_\omega}\right] \tag{A.22}$$

$$\log \tilde{\Gamma}^{\mathcal{X}}_\omega = \sum_{i=1}^{d^{\mathcal{X}}} \Psi((a^{\mathcal{X}}_\omega + 1 - i)/2) - \log|\boldsymbol{B}^{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\omega}}| + d^{\mathcal{X}} \log 2 \tag{A.23}$$

$$\bar{\boldsymbol{\Gamma}}^{\mathcal{X}}_\omega = a^{\mathcal{X}}_\omega (\boldsymbol{B}^{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\omega}})^{-1} \tag{A.24}$$

where $\Psi()$ is the Digamma function and $d^{\mathcal{X}}$ is the dimensionality of $\mathcal{X}^n_h$. Superscript $\mathcal{X}$ indicates that the parameters are related to the shape component of the model. The RHS of the above equations consists of hyper-parameters for the parameter posteriors (i.e., λ, $\boldsymbol{m}$, $\boldsymbol{B}$, β and a). $\tilde{\gamma}_\omega(\mathcal{A}^n_h)$ is computed exactly the same way as $\tilde{\gamma}_\omega(\mathcal{X}^n_h)$, using the corresponding parameters of the appearance component. We then normalize to give:

$$\gamma^n_{\omega,h} = \frac{\tilde{\gamma}^n_{\omega,h}}{\sum_{\omega',h'} \tilde{\gamma}^n_{\omega',h'}} \tag{A.25}$$

which is the probability that component ω is responsible for hypothesis $\boldsymbol{h}$ of the n^{th} training image.

A.2.4 The M-step in Bayesian One-Shot

In the M-step, $q(\boldsymbol{\theta})$ is updated according to:

$$q(\boldsymbol{\theta}) \propto \exp\left[I(\boldsymbol{\theta})\right] p(\boldsymbol{\theta}) \qquad \text{where} \quad I(\boldsymbol{\theta}) =< \log p(\mathcal{X}_{\sqcup}, \mathcal{A}_{\sqcup}, \boldsymbol{\omega}, h|\theta) >_{\boldsymbol{\omega},\boldsymbol{h}} \tag{A.26}$$

Again, the above equation can be written as:

$$I(\boldsymbol{\theta}) = \langle \log p(\mathcal{X}_t|\boldsymbol{\omega}, \boldsymbol{h}, \boldsymbol{\theta}) p(\mathcal{A}_t|\boldsymbol{\omega}, \mathcal{X}_t, \boldsymbol{h}, \boldsymbol{\theta}) p(\boldsymbol{\omega}|\theta) p(\boldsymbol{h}) \rangle_{\boldsymbol{\omega},\boldsymbol{h}} \tag{A.27}$$

and the expectation is taken w.r.t. $q(\boldsymbol{\omega}, \boldsymbol{h})$.

We show here the update rules for the shape components. The equations are exactly the same for the

appearance components. We define the following variables:

$$\bar{\pi}_\omega = \frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{|H^n|}\gamma_{\omega,h}^n \tag{A.28}$$

$$\bar{N}_\omega = N\bar{\pi}_\omega \tag{A.29}$$

$$\bar{\boldsymbol{\mu}}_\omega^{\mathcal{X}} = \frac{1}{\bar{N}_\omega}\sum_{n=1}^{N}\sum_{h=1}^{|H^n|}\gamma_{\omega,h}^n\mathcal{X}_h^n \tag{A.30}$$

$$\bar{\boldsymbol{\Sigma}}_\omega^{\mathcal{X}} = \frac{1}{\bar{N}_\omega}\sum_{n=1}^{N}\sum_{h=1}^{|H^n|}\gamma_{\omega,h}^n(\mathcal{X}_h^n-\bar{\boldsymbol{\mu}}_\omega^{\mathcal{X}})(\mathcal{X}_h^n-\bar{\boldsymbol{\mu}}_\omega^{\mathcal{X}})^T \tag{A.31}$$

Now we are ready to re-estimate the model distribution $p(\boldsymbol{\theta}|\mathcal{X}_t,\mathcal{A}_t,\mathcal{O})$ through updating the hyper-parameters ($\{\lambda_\omega, a_\omega, \boldsymbol{B}_\omega, \boldsymbol{m}_\omega, \beta_\omega\}$) that govern the shape of the distribution. For the mixing coefficients we have a Dirichlet distribution $q(\boldsymbol{\pi}) = \mathcal{D}ir(\boldsymbol{\lambda})$ where the hyper-parameters are updated by:

$$\lambda_\omega = \bar{N}_\omega + \lambda_0 \tag{A.32}$$

For the means, we have $q(\boldsymbol{\mu}_\omega^{\mathcal{X}}|\boldsymbol{\Gamma}_\omega^{\mathcal{X}}) = \mathcal{G}(\boldsymbol{m}_\omega^{\mathcal{X}}, \beta_\omega^{\mathcal{X}}\boldsymbol{\Gamma}_\omega^{\mathcal{X}})$ where:

$$\boldsymbol{m}_\omega^{\mathcal{X}} = \frac{\bar{N}_\omega\bar{\boldsymbol{\mu}}_\omega^{\mathcal{X}} + \beta_0^{\mathcal{X}}\boldsymbol{m}_0^{\mathcal{X}}}{\bar{N}_\omega + \beta_0^{\mathcal{X}}} \tag{A.33}$$

$$\beta_\omega^{\mathcal{X}} = \bar{N}_\omega + \beta_0^{\mathcal{X}} \tag{A.34}$$

For the noise precision matrix we have a Wishart density $q(\boldsymbol{\Gamma}_\omega^{\mathcal{X}}) = \mathcal{W}(a_\omega^{\mathcal{X}}, \boldsymbol{B}_\omega^{\mathcal{X}})$ where:

$$\boldsymbol{B}_\omega^{\mathcal{X}} = \frac{\bar{N}_\omega\beta_0^{\mathcal{X}}(\bar{\boldsymbol{\mu}}_\omega^{\mathcal{X}}-\boldsymbol{m}_0^{\mathcal{X}})(\bar{\boldsymbol{\mu}}_\omega^{\mathcal{X}}-\boldsymbol{m}_0^{\mathcal{X}})^T}{\bar{N}_0+\beta_0^{\mathcal{X}}} + \bar{N}_\omega\bar{\boldsymbol{\Sigma}}_\omega^{\mathcal{X}} + \boldsymbol{B}_0^{\mathcal{X}} \tag{A.35}$$

$$a_\omega^{\mathcal{X}} = \bar{N}_\omega + a_0^{\mathcal{X}} \tag{A.36}$$

A.3 MAP learning

Having laid out the variational Bayesian framework for learning and recognition, we now give the learning equations for the MAP scenario. In Section 6 of [34] we will compare the performance of ML [39], MAP and the Variational Bayesian approaches.

In the MAP scenario, the integral in Eq.6 in [34] can be simplified as:

$$\int p(\mathcal{X},\mathcal{A}|\boldsymbol{\theta},\mathcal{O})p(\boldsymbol{\theta}|\mathcal{X}_t,\mathcal{A}_t,\mathcal{O})\,d\boldsymbol{\theta} \approx p(\mathcal{X},\mathcal{A}|\boldsymbol{\theta}^{\text{MAP}},\mathcal{O})$$
$$\text{where}\quad \boldsymbol{\theta}^{\text{MAP}} = \underset{\theta}{argmax}\; p(\mathcal{X}_t,\mathcal{A}_t|\boldsymbol{\theta},\mathcal{O})p(\boldsymbol{\theta}) \tag{A.37}$$

The prior distribution of $p(\boldsymbol{\theta})$ in MAP has the same form as Eq. A.17 in Variational Bayes:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})\prod_\omega p(\boldsymbol{\mu}_\omega^{\mathcal{X}}|\boldsymbol{\Gamma}_\omega^{\mathcal{X}})p(\boldsymbol{\Gamma}_\omega^{\mathcal{X}})p(\boldsymbol{\mu}_\omega^{\mathcal{A}}|\boldsymbol{\Gamma}_\omega^{\mathcal{A}})p(\boldsymbol{\Gamma}_\omega^{\mathcal{A}}) \tag{A.38}$$

where the mixing prior is $p(\boldsymbol{\pi}) = \mathcal{D}ir(\lambda_{\omega,0}\boldsymbol{I}_\Omega)$, and the shape prior is a Normal-Wishart distribution $p(\boldsymbol{\mu}_\omega^{\mathcal{X}}|\boldsymbol{\Gamma}_\omega^{\mathcal{X}})p(\boldsymbol{\Gamma}_\omega^{\mathcal{X}}) = \mathcal{G}(\boldsymbol{\mu}_\omega^{\mathcal{X}}|\boldsymbol{m}_{\omega,0}^{\mathcal{X}}, \beta_{\omega,0}^{\mathcal{X}}\boldsymbol{\Gamma}_\omega^{\mathcal{X}})\mathcal{W}(\boldsymbol{\Gamma}_\omega^{\mathcal{X}}|a_{\omega,0}^{\mathcal{X}}, \boldsymbol{B}_{\omega,0}^{\mathcal{X}})$. Identical expressions apply to the appearance component of Eq. A.38.

A.3.1 Expectation Maximization (EM) for MAP

The central idea of EM for MAP is to obtain the optimal $\boldsymbol{\theta}$ through iteratively optimizing the $\boldsymbol{q}(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1})$ function, where $\boldsymbol{\theta}$ contains the set of parameters $\{\boldsymbol{\pi}, \boldsymbol{\mu}^{\mathcal{X}}, \boldsymbol{\mu}^{\mathcal{A}}, \boldsymbol{\Gamma}^{\mathcal{X}}, \boldsymbol{\Gamma}^{\mathcal{A}}\}$:

$$\boldsymbol{q}(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1}) = \sum_{\omega}^{\Omega} \log p(\omega, \mathcal{X}, \mathcal{A}, \boldsymbol{\theta}^t)\ p(\omega|\mathcal{X}, \mathcal{A}, \boldsymbol{\theta}^{t-1}) \tag{A.39}$$

In the E-step, we calculate the indicator function for each hypothesis $\boldsymbol{h}$ and each mixture component ω:

$$\begin{aligned} &\tilde{\gamma}^n_{\omega,h} = \pi_\omega\ \tilde{\gamma}_\omega(\mathcal{X}^n_h)\ \tilde{\gamma}_\omega(\mathcal{A}^n_h) \\ &\text{where} \qquad \tilde{\gamma}_\omega(\mathcal{X}^n_h) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_\omega|^{-1}} \exp\left[-\frac{1}{2}(\mathcal{X}^n_h - \boldsymbol{m}^{\mathcal{X}}_{\omega,0})^T \boldsymbol{\Sigma}^{-1}_\omega (\mathcal{X}^n_h - \boldsymbol{m}^{\mathcal{X}}_{\omega,0})\right] \end{aligned} \tag{A.40}$$

The same form of equation applies to $\tilde{\gamma}_\omega(\mathcal{A}^n_h)$. Finally we obtain the normalized indicator function:

$$\gamma^n_{\omega,h} = \frac{\tilde{\gamma}^n_{\omega,h}}{\sum_{\omega',h'} \tilde{\gamma}^n_{\omega',h'}} \tag{A.41}$$

In the M-step, we maximize $\boldsymbol{q}(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1})$ over each of the parameters $\{\boldsymbol{\pi}, \boldsymbol{\mu}^{\mathcal{X}}, \boldsymbol{\mu}^{\mathcal{A}}, \boldsymbol{\Sigma}^{\mathcal{X}}, \boldsymbol{\Sigma}^{\mathcal{A}}\}$. For convenience, we only show here the update rules for the mixture component as well as the shape related parameters. Appearance related parameters have the same update form as their shape counterparts.

$$\pi_\omega = \frac{\sum_{n=1}^{N}\sum_{h=1}^{|H^n|}\gamma^n_{\omega,h} + \lambda_{\omega,0} - 1}{N + \sum_{\omega=1}^{\Omega}\lambda_{\omega,0} - \Omega} \tag{A.42}$$

$$\boldsymbol{\mu}^{\mathcal{X}}_\omega = \frac{\sum_{n=1}^{N}\sum_{h=1}^{|H^n|}\gamma^n_{\omega,h}\mathcal{X}^n_h + \beta^{\mathcal{X}}_{\omega,0}\boldsymbol{m}^{\mathcal{X}}_{\omega,0}}{\sum_{n=1}^{N}\sum_{h=1}^{|H^n|}\gamma^n_{\omega,h} + \beta^{\mathcal{X}}_{\omega,0}} \tag{A.43}$$

$$\boldsymbol{\Sigma}^{\mathcal{X}}_\omega = \frac{\sum_{n=1}^{N}\sum_{h=1}^{|H^n|}\gamma^n_{\omega,h}(\mathcal{X}^n_h - \boldsymbol{\mu}^{\mathcal{X}}_\omega)(\mathcal{X}^n_h - \boldsymbol{\mu}^{\mathcal{X}}_\omega)^T + \beta^{\mathcal{X}}_{\omega,0}(\boldsymbol{\mu}^{\mathcal{X}}_\omega - \boldsymbol{m}^{\mathcal{X}}_{\omega,0})(\boldsymbol{\mu}^{\mathcal{X}}_\omega - \boldsymbol{m}^{\mathcal{X}}_{\omega,0})^T + \boldsymbol{B}^{\mathcal{X}}_{\omega,0}}{\sum_{n=1}^{N}\sum_{h=1}^{|H^n|}\gamma^n_{\omega,h} + \beta^{\mathcal{X}}_{\omega,0} + a^{\mathcal{X}}_{\omega,0} + d^{\mathcal{X}} + 1} \tag{A.44}$$