PROCEEDINGS OF THE 2005 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION
VOLUME II, PAGES 1331–1338, OCT. 2005

Learning Hierarchical Models of Scenes, Objects, and Parts

Erik B. Sudderth, Antonio Torralba, William T. Freeman, and Alan S. Willsky
Electrical Engineering & Computer Science, Massachusetts Institute of Technology
esuddert@mit.edu, torralba@csail.mit.edu, billf@mit.edu, willsky@mit.edu

Abstract

We describe a hierarchical probabilistic model for the detection and recognition of objects in cluttered, natural scenes. The model is based on a set of parts which describe the expected appearance and position, in an object centered coordinate frame, of features detected by a low-level interest operator. Each object category then has its own distribution over these parts, which are shared between objects. We learn the parameters of this model via a Gibbs sampler which uses the graphical model's structure to analytically average over many parameters. Applied to a database of images of isolated objects, the sharing of parts among objects improves detection accuracy when few training examples are available. We also extend this hierarchical framework to scenes containing multiple objects.

1. Introduction

In this paper, we develop methods for the visual detection and recognition of object categories. We argue that multi–object recognition systems should be based on models which consider the relationships between different object categories during the training process. This approach provides several benefits. At the lowest level, significant computational savings can be achieved if different categories share a common set of features. More importantly, jointly trained recognition systems can use similarities between object categories to their advantage by learning features which lead to better generalization [4, 18]. This inter–category regularization is particularly important in the common case where few training examples are available.

In complex, natural scenes, object recognition systems can be further improved by using contextual knowledge about the objects likely to be found in a given scene, and common spatial relationships between those objects [7, 19, 20]. In this paper, we propose a hierarchical generative model for objects, the parts composing them, and the scenes surrounding them. The model, which is summarized in Figs. 1 and 5, shares information between object categories in three distinct ways. First, parts define distributions over a common low–level feature vocabularly, leading to computational savings when analyzing new images. In addition, and more unusually, objects are defined using a common set of parts. This structure leads to the discovery of parts with interesting semantic interpretations, and can improve performance when few training examples are available. Finally, object appearance information is shared between the many scenes in which that object is found.

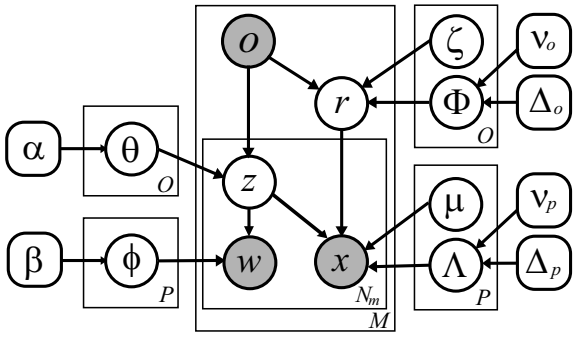

Figure 1. Graphical model describing how latent parts $\mathbf{z}$ generate the appearance $\mathbf{w}$ and position $\mathbf{x}$, relative to an image–specific reference location $\mathbf{r}$, of the features detected in an image of object $\mathbf{o}$. Boxes denote replication of the corresponding random variables: there are M images, with N_m observed features in image m.

We begin in Sec. 2 by describing our generative model for objects and parts, including a discussion of related work in the machine vision and text analysis literature. Sec. 3 then describes parameter estimation methods which combine Gibbs sampling with efficient variational approximations. In Sec. 4, we provide simulations demonstrating the potential benefits of feature sharing. We conclude in Sec. 5 with preliminary extensions of the object hierarchy to scenes containing multiple objects.

2. A Generative Model for Object Features

Our generative model for objects is summarized in the graphical model (a directed Bayesian network) of Fig. 1. The nodes of this graph represent random variables, where shaded nodes are observed during training, and rounded boxes are fixed hyperparameters. Edges encode the conditional densities underlying the generative process [12].

2.1. From Images to Features

Following [17], we represent each of our M grayscale training images by a set of SIFT descriptors [13] computed on affine covariant regions. We use K-means clustering to