PROCEEDINGS OF THE 2006 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION
VOLUME II, PAGES 2410–2417, JUNE 2006

Depth from Familiar Objects: A Hierarchical Model for 3D Scenes

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

Abstract

We develop an integrated, probabilistic model for the appearance and three-dimensional geometry of cluttered scenes. Object categories are modeled via distributions over the 3D location and appearance of visual features. Uncertainty in the number of object instances depicted in a particular image is then achieved via a transformed Dirichlet process. In contrast with image-based approaches to object recognition, we model scale variations as the perspective projection of objects in different 3D poses. To calibrate the underlying geometry, we incorporate binocular stereo images into the training process. A robust likelihood model accounts for outliers in matched stereo features, allowing effective learning of 3D object structure from partial 2D segmentations. Applied to a dataset of office scenes, our model detects objects at multiple scales via a coarse reconstruction of the corresponding 3D geometry.

1. Introduction

Detailed geometric models have played an important role in the design of methods for the detection of particular objects in cluttered scenes. However, most algorithms for generic object categorization use a simple 2D pixel representation. In discriminative approaches, scale invariance is often achieved by resizing an image in small steps, and detecting objects via a "sliding window". Alternatively, some part–based models include variables which account for global scaling of expected feature distances [3], or local affine warpings of feature templates [7]. Other methods discard geometry entirely following an initial stage of feature extraction [1, 13]. In all cases, scale invariance is based on transformations of the observed pixels or low–level features, and underlying 3D structure is ignored.

While a purely image based approach to object recognition is sometimes adequate, many applications require more explicit knowledge about the 3D world. For example, if robots are to navigate in complex environments and manipulate objects, they require more than a flat segmentation of the image pixels into object categories. Motivated by these challenges, we instead cast object recognition as a 3D problem, and develop methods which partition estimated 3D structure into object categories.

A few recent models ignore objects, learning direct mappings from images to 3D geometry [5, 12, 17]. However, knowledge of the objects present in a scene provides information about their expected 3D shape, regularizing the often ambiguous depth estimates produced by low–level features. In addition, geometry provides important cues for object recognition. To exploit these relationships, we use binocular stereo training images to train an approximately calibrated model of multiple objects' 3D geometry. Using this model, we achieve scale invariant object recognition via translations of 3D objects, rather than image transformations. Because we consider objects with predictable 3D structure, we also automatically recover a coarse reconstruction of the underlying scene depths.

Rather than learning classifiers for isolated objects, we propose a hierarchical, probabilistic model of multiple object scenes [14, 18]. Our approach extends an earlier 2D scene model based on the *transformed Dirichlet process* (TDP) [15]. Dirichlet processes are a flexible tool from nonparametric Bayesian statistics [2, 11, 16], which we use to allow uncertainty in the *number* of object instances depicted in each image. Generalizing [15], we automatically learn part–based descriptions of an *a priori* unknown set of visual categories. Previous TDP models also assembled 2D object models in a "jigsaw puzzle" fashion, and thus assumed images were normalized to a common scale. Extending the TDP to 3D scenes, we propose a robust stereo likelihood which captures ambiguities in low–level feature matching. We then develop Monte Carlo methods which learn 3D object models from partial stereo segmentations, and estimate 3D scene structure from monocular images.

We begin in Sec. 2 by introducing our feature representation, and formulate a robust stereo likelihood function. Sec. 3 then uses the TDP to develop a generative, part–based model for 3D feature locations and appearance. In Sec. 4, we describe a blocked Gibbs sampler which learns scene geometry from labeled stereo training images. We conclude in Sec. 5 with results on office scenes.