Object-Level Generative Models for 3D Scene Understanding

Ehsan Jahangiri, René Vidal, Laurent Younes, Donald Geman
Center for Imaging Science, Johns Hopkins University.

Abstract

A core challenge in computer vision is to develop generative models of the world that capture rich contextual relationships among scene entities. Such models are broadly applicable in scene understanding, computer graphics, and robotics, including serving as prior models in a Bayesian framework, constructing model-based visual Turing tests, robotic manipulation, etc. This paper proposes a new probabilistic, generative model of 3D scenes consisting of multiple objects lying on a plane. The proposed model is a probability distribution over random attributed graphs that can encode favored layouts while accounting for variations in the number and relative poses of objects. Each graph node corresponds to an object instance that is labeled with a category and a 3D pose in the world coordinate system, while relationships among nodes reflect the generative process. Finally, we illustrate how to learn the model parameters from annotated images of dinning room tables.

1 Introduction

In the past decade there have been considerable advances in building models for object categorization, mostly aimed at discriminative learning and at reasoning primarily in 2D [2, 3, 11]. Recently, several attempts have been made at building models that reason about 3D surfaces of scenes and the interaction of objects with the supporting surfaces [1, 6, 7, 8, 9, 10]. However, such models are not generative and to the best of our knowledge do not encode contextual relations among objects on supporting surfaces beyond their coplanarity. Many man-made scenes are composed of multiple parallel supporting surfaces upon which instances from different object categories are placed [1], often with considerable structure. In this paper, we focus on modeling the 3D arrangement of objects on one supporting plane. Our 3D world model is at the level of objects and allows for encoding expected properties and multi-object relationships among a distinguished family of objects, including the numbers and relative poses of scene objects.

Such models can serve scene understanding in several ways. For one, they can be coupled with data models to complete a conventional Bayesian framework and thereby "regularize" the output from image descriptors; at a semantic level, the descriptors might be discriminatively trained classifiers for detecting and localizing object instances and the prior model integrates these results with expected relationships among objects. Another application is to generate sequences of unpredictable queries for testing computer vision systems; such a visual Turing test was described in [4], where the likelihoods of answers were estimated empirically from labeled data, relying on heuristics to address data fragmentation. Having an appropriate model would allow one to identify longer streams and more accurate estimates of unpredictability by sampling from the scene model conditional on oracle answers (in effect perfect classifiers). However, the focus of this paper is primarily in the design and learning of such a prior model.

The underlying graph in the model is a forest of directed trees which captures a natural generative process in which objects are placed down on the surface in stages; the number of root nodes (e.g., a place setting instance) refers to the conditional placement of an object instance relative to the size and geometry of the scene; an edge from a parent (e.g., a plate instance) to a child (e.g., a utensil instance) refers to the conditional placement of the child relative to the parent; and so forth (see Figure 1). Moreover, edges are only allowed between certain types of object categories and these restrictions are imposed by a "Master Graph". Designing 3D models which encode favored relationships but still accommodate real-world variability is not straightforward. In particular, given purely object-annotated image data, learning is complicated because the graph structure is hidden. Another challenge is conditional sampling: whereas simulating from the full model is simple and feedforward, applications to Bayesian inference and constructing query streams require generating samples of attributed graphs under multiple constraints on object instances and relationships.

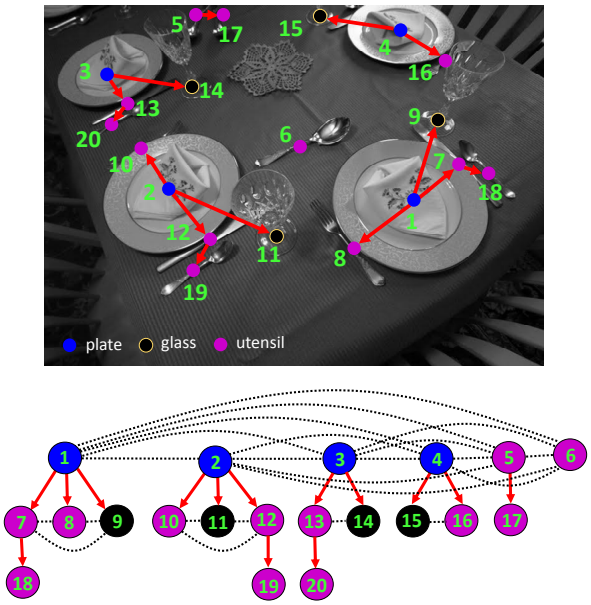

Figure 1: A table-setting scene and its corresponding category-labeled base graph where the categories (plate, bottle, glass, and utensil) are color-coded in the graph. Root nodes V_0 initialize the generative process; here there are six. The terminal nodes for this instance are $V_T = \{6,8,9,10,11,14,15,16,17,18,19,20\}$. According to the base graph $n_{(0,\text{plate})} = 4$, $n_{(0,\text{bottle})} = 0$, $n_{(0,\text{glass})} = 0$, and $n_{(0,\text{utensil})} = 2$. Removing the undirected edges (dashed lines) leads to a Bayesian tree.

2 Proposed Model

In the proposed model, a scene is described as a collection of object instances from different categories at different poses. Each object instance is associated with a vertex $v \in V$ of a *base* graph $g_0 \in \mathcal{G}_0$ which captures contextual relationships among object instances. An attributed graph is a triple $g = (g_0, c_V, \theta_V)$, where $c_V = \{c_v, v \in V\}$ and $\theta_V = \{\theta_v, v \in V\}$ denote the set of category labels and 3D poses of objects, respectively. The categories are restricted to a set $\mathcal{C}$ of size K. The root nodes are denoted by $V_0 \subset V$, the terminal nodes by $V_T \subset V$, the parent of node v by $pa(v)$ and the set of its children by $ch(v)$. The model is a probability distribution $p(g|T),\ g \in \mathcal{G}$ on the space of attributed graphs $\mathcal{G}$ conditioned on the environment's geometric properties T. The model is specified by four sets of distributions:

(1) $p^{(0)}(n_{(0,1)}, \cdots, n_{(0,K)}|T)$: the conditional joint distribution of the number of root nodes $\{n_{(0,k)}\}_{k=1}^K$ from each object category k given the geometric properties T of the supporting surface.

(2) $\{p^{(c)}(n_1, \cdots, n_K),\ c \in \mathcal{C}\}$: the joint distribution of the number of children $\{n_k\}_{k=1}^K$ from each object category k of a parent from category c.

(3) $p(\theta_{V_0}|c_{V_0}, T)$: the conditional joint distribution of the poses of the root nodes given their corresponding category labels.

(4) $\{p(\theta_{ch(v)}|c_{ch(v)}, c_v, \theta_v, T),\ v \in V \backslash V_T\}$: the conditional joint distribution of the poses of the children of v given their parent's pose and the corresponding category labels.

The category-labeled base graph distribution is:

$$p(g_0, c_V|T) = p^{(0)}(n_{(0,1)}, \cdots, n_{(0,K)}|T) \times \prod_{v \in V} p^{(c_v)}(n_{(v,1)}, \cdots, n_{(v,K)}), \quad (1)$$

where, $n_{(v,k)}$ denotes the number of children from the k-th category of vertex v. The full model is:

$$\begin{aligned} p(g|T) &= p(g_0, c_V|T) \times p(\theta_V|g_0, c_V, T) \\ &= p^{(0)}(n_{(0,1)}, \cdots, n_{(0,K)}|T)\ p(\theta_{V_0}|c_{V_0}, T) \times \\ &\quad \prod_{v \in V \backslash V_T} p^{(c_v)}(n_{(v,1)}, \cdots, n_{(v,K)}) \times p(\theta_{ch(v)}|c_{ch(v)}, c_v, \theta_v, T). \end{aligned} \quad (2)$$

Assuming the children are conditionally independent given the corresponding category labels, the pose of their parent, and the geometric properties of the supporting surface, the graph reduces to the standard *Bayesian* forest structure. Otherwise, we will have a hybrid graph with both directed and undirected edges. Figure 1 illustrates an example scene and its corresponding base graph.

Obviously, some category pairs have stronger expected contextual relationships than others, and it is reasonable to assume $p^{(c_i)}(n_1,\ldots,n_K)=0$ for configurations with $n_j>0$ for certain pairs c_i and c_j. To capture these preferred relationships, we define a directed *Master* graph, $G_M=(\mathcal{V}_M,\mathcal{E}_M)$, over the set of object categories $\mathcal{C}$; this constrains the branching probabilities $p^{(c)}(\cdot)$. Every vertex in G_M corresponds to one object category in $\mathcal{C}$ and every edge $(c_i\rightarrow c_j)\in\mathcal{E}_M$ indicates existence of a considerable contextual relation (not necessarily causal) between categories c_i and c_j. A vertex of category c_i will not generate a vertex of category c_j if there is no directed edge from c_i to c_j in the master graph G_M *i.e.,* if $(c_i\rightarrow c_j)\notin\mathcal{E}_M$. An undirected version of the master graph can be computed by thresholding the fully-connected context graph over the set of object categories whose edge weights are proportional to the local co-occurrence of the corresponding categories. We then give direction to the edges usually from landmark (larger) objects to more peripheral (smaller) objects. The master graph can be fully or partially user-determined.

3 Model Learning

Assume we have a data set of J annotated scenes from which we obtain object attributes, namely we can get $\mathcal{D}=\{c_V[j],\theta_V[j]\}_{j=1}^J$. However, $\mathcal{D}$ is not a sufficient statistic for learning the model parameters since the set of corresponding base graphs $\mathcal{M}=\{g_0[j]\}_{j=1}^J$ is missing. We normally do not directly observe the set of corresponding base graphs from annotated scenes. Therefore, we are facing a learning-from-incomplete-data problem where $\mathcal{D}$ is given and $\mathcal{M}$ is missing. The combination of missing and incomplete data constitutes the complete data composed of attributed graphs for each scene $\mathcal{D}^+=\langle\mathcal{D},\mathcal{M}\rangle=\{g[j]=(g_0[j],c_V[j],\theta_V[j])\}_{j=1}^J$. We propose a parameter learning method based on the stochastic *Expectation-Maximization* (EM) algorithm. According to the proposed learning strategy, we sample the conditional base graph distribution given the object attributes using Gibbs sampling to complete $\mathcal{D}$ and estimate the parameters by iteratively maximizing the complete data likelihood over the parameters.

Let Φ denote the set of all the parameters in the model including the parameters of the four sets of distributions summarized earlier, we iteratively estimate the parameters until convergence according to:

$$\Phi^{t+1}=\arg\max_{\Phi}\sum_{j=1}^{J}\sum_{g_0[j]}p(g_0[j]\mid c_V[j],\theta_V[j],\Phi^t)\times\log p(g[j]\mid\Phi),\quad(3)$$

which is obtained assuming that the base graphs for different images are independent given their corresponding object attributes. Note that we dropped T for notational convenience in (3). Also, note that the parameters of $p(g|T)$ should be learned from annotated scenes whose environment's geometric properties roughly match T. We are not describing sampling of data completion distribution in the interest of space but let $g_0^{(l)}[j]$ denote a base graph sample for the j-th annotated scene from $p(g_0[j]\mid c_V[j],\theta_V[j],\Phi^t)$ and N is a sufficiently large number of such samples, then using Monte-Carlo integration we have:

$$\Phi^{t+1}\approx\arg\max_{\Phi}\sum_{j=1}^{J}\sum_{l=1}^{N}\log p(g^{(l)}[j]=(g_0^{(l)}[j],c_V[j],\theta_V[j])\mid\Phi).\quad(4)$$

4 The JHU Table-Setting Dataset

We applied the proposed attributed graph model to table-setting scenes in the world with four object categories including plates, bottles, glasses, and utensils. We learned the model from a fully-annotated in-house data set of about 3000 table-setting scene images. Figure 2 shows a snapshot of this data set. The left hand side photo in Figure 3 shows a sample annotated image from the data set. Each annotation is represented by a polygon containing the object and its corresponding category label. We represent the 2D pose of an object in the image by an enclosing ellipse to its corresponding polygon vertices. We assume that objects are planar which is reasonable if the heights of objects are small relative to their distances from the camera. To estimate the pose of objects in the 3D world coordinate system for every image, namely θ_V, we back-projected the 2D fitting ellipses onto the table (world) coordinate system using the corresponding manually-estimated homography matrix. The homography matrix for every image in the dataset is manually estimated and scaled appropriately (using objects' typical size in real world) such that after back-projection all distances can be measured in meters. The right hand side photo in Figure 3 shows a top-view of the table-setting on the left after back-projection using the corresponding manually-estimated homography.

Figure 2: A snapshot of the JHU table-setting data set.

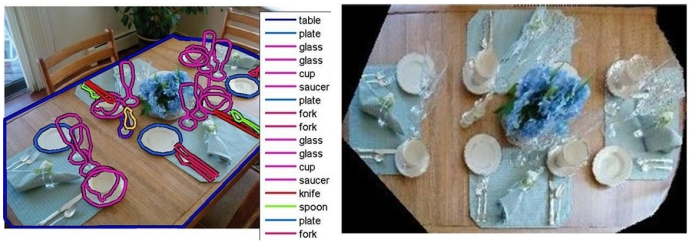

Figure 3: A sample from the JHU table-setting data set. On the left is an annotated image and on the right is its corresponding top-view after back-projection.

According to the pinhole camera model, the world model samples from $p(g|T)$ can be projected to the image coordinate system given camera's intrinsic and extrinsic parameters. Such a "projected model" could in principle be used for 2D scene understanding. However, in applications such as 3D scene understanding and robotics the 3D world model is directly used [5].

References

[1] S. Y. Bao, M. Sun, and S. Savarese. Toward coherent object detection and scene layout understanding. In *CVPR*, 2010.

[2] M. J. Choi, A. Torralba, and A. S. Willsky. A tree-based context model for object recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(2): 240–252, February 2012.

[3] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. *Int. J. Comput. Vision*, 2011.

[4] D. Geman, S. Geman, N. Hallonquist, and L. Younes. A visual turing test for computer vision systems. In *PNAS*, 2014.

[5] G. D. Hager and B. Wegbreit. Scene parsing using a prior world model. *Int. J. of Rob. Res.*, 30(12):1477–1507, 2011.

[6] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *Int. J. Comput. Vision*, 75(1):151–172, October 2007.

[7] D. C. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010.

[8] X. Liu, Y. Zhao, and SC. Zhu. Single-view 3d scene parsing by attributed grammar. In *CVPR*, 2014.

[9] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(5):824–840, May 2009.

[10] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[11] M. Sun, B. Kim, P. Kohli, and S. Savarese. Relating things and stuff via objectproperty interactions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1370–1383, 2014.