This CVPR2013 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Hallucinated Humans as the Hidden Context for Labeling 3D Scenes

Yun Jiang, Hema Koppula and Ashutosh Saxena
Department of Computer Science, Cornell University.
{yunjiang,hema,asaxena}@cs.cornell.edu

Abstract

For scene understanding, one popular approach has been to model the object-object relationships. In this paper, we hypothesize that such relationships are only an artifact of certain hidden factors, such as humans. For example, the objects, monitor and keyboard, are strongly spatially correlated only because a human types on the keyboard while watching the monitor. Our goal is to learn this hidden human context (i.e., the human-object relationships), and also use it as a cue for labeling the scenes. We present Infinite Factored Topic Model (IFTM), where we consider a scene as being generated from two types of topics: human configurations and human-object relationships. This enables our algorithm to hallucinate the possible configurations of the humans in the scene parsimoniously. Given only a dataset of scenes containing objects but not humans, we show that our algorithm can recover the human object relationships. We then test our algorithm on the task of attribute and object labeling in 3D scenes and show consistent improvements over the state-of-the-art.

1. Introduction

> *We make the world we live in and shape our own environment.* Orison Swett Marden (1894).

For reasoning about cluttered human environments, for example in the task of 3D scene labeling, it is critical we reason *through* humans. Human context provides a natural explanation of why the environment is built in particular ways. Specifically, consider the scene in Fig. 1, with a chair, table, monitor and keyboard. This particular configuration that is commonly found in offices, can be naturally explained by a sitting human pose in the chair and working with the computer. Moreover, from the point of view of modeling and learning, this explanation is parsimonious and efficient as compared to modeling the object-object relationships [19] such as chair-keyboard, table-monitor, monitor-keyboard, etc.[1]

[1]For n objects, we only need to model how they are used by humans, i.e., $O(n)$ relations, as compared with modeling $O(n^2)$ if we were to model object to object context naively.

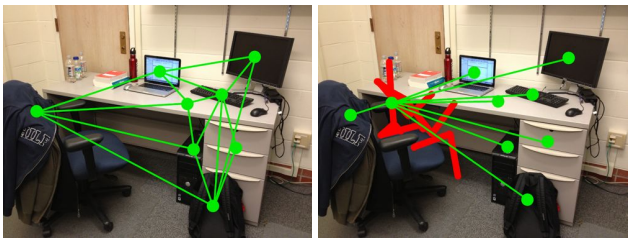


Figure 1: Left: Previous approaches model the relations between observable entities, such as the objects. Right: In our work, we consider the relations between the objects and hidden humans. Our key hypothesis is that even when the humans are never observed, the human context is helpful.

In fact, several recent works have shown promise in using human and object affordances to model the scenes. Jiang, Lim and Saxena [14, 17] used hallucinated humans for learning the object arrangements in a house in order to enable robots to place objects in human-preferred locations. However, they assumed that the objects have been detected. Our goal in this work is different, where we start with 3D point-clouds obtained from RGB-D sensors and label them using their shape, appearance and hallucinated human context. Gupta et al. [10] proposed predicting stable and feasible human poses given an approximate 3D geometry from an image. While inspired by these prior works, the key idea in our work is to hallucinate humans in order to learn a generic form of object affordance, and to use them in the task of labeling 3D scenes. While a large corpus of scenes with objects is available, humans and their interactions with objects are observed only a few times for some objects. Therefore, using hallucinated humans gives us the advantage of considering human context while not limited to data that contains real human interactions.

However, if the humans are not observed in the scene and we do not know the object affordances either (i.e., how humans use objects), then learning both of them is an ill-posed problem. For example, one trivial, but useless, solution would be having one human configuration for every object in the scene. The key idea in our work is to prefer parsimony in our model as follows. First, while the space of potential unobserved human configurations are large, only

few are likely, and so are the object affordances. For example, if standing on furniture (such as tables and chairs) is very unlikely in the prior, we are less likely to learn affordances such as humans stepping on books. Second, we encourage fewer humans per scene resulting in different objects sharing same human configurations. This allows us to explain, but not over-fit, a scene with as few human configurations as necessary.

In order to model the scene through hallucinated human configurations and object affordances, we propose a new topic model, which we call Infinite Factored Topic Model (IFTM). Each object in the scene is generated by two types of topics *jointly*: human-configuration topics and object-affordance topics. We use a sampling algorithm to estimate the human pose distribution in scenes and to optimize the affordance functions that best explain the given scenes. The learned topics are later used as features for building a scene labeling classifier.

We test our approach on the tasks of labeling objects and attributes in 3D scenes, and show that the human-object context is informative in that it increases performance over classifier based on object appearance and shapes. More interestingly, we show that object-object context and human-object context are complementary in nature and their combination improves the state-of-the-art.

2. Related Work

Context and 3D Scene Understanding. There is a significant body of work that captures the relations between different parts of the object [5] and between different objects [19]. In the past, 3D layout or depths have been used for improving object detection (e.g., [26, 27, 11, 21, 12, 22]), where an approximate 3D geometry is inferred from 2D images. Recent works [31, 19, 2, 28] address the problem of labeling 3D point clouds. Reasoning in 3D allows an algorithm to capture stronger context, such as shape, stability and orientation of the objects [15, 13, 16]. However, none of these works consider human context for scene understanding.

Human activities. In most previous works, object detection and activity recognition have been addressed as separate tasks. Only recently, some works [9, 32, 1, 25, 20] have shown that modeling the interaction between human poses and objects in 2D images and videos result in a better performance on the tasks of object detection and activity recognition. In contemporary work, Fouhey et al. [6] and Delaitre et al. [4] observe humans in videos for estimating 3D geometry and estimating affordances respectively. However, these works are unable to characterize the relationship between objects in 3D unless a human explicitly interacted with each of the objects and are also limited by the quality of the human poses inferred from 2D data. Our method can extract the hidden human context even from static scenes *without* humans, based on the object configurations found in the human environments.

Object Affordances. The concept of affordances, proposed by Gibson [7], has recently become the focus of many works in cognitive vision (e.g., [24]) and robotics [14, 17]. Grabner et al. [8] apply this idea to detect the functionality of the object (specifically, chairs), and then combine this information with visual object appearance to perform object classification. However, they require explicit training data specifying the human pose associated with an affordance and demonstrated their method on a single object category and affordance. In comparison, Jiang et al. [14] consider many affordances in the form of human-object relation topics which are obtained in a completely unsupervised manner. While they employ the learned affordances to infer reasonable object arrangements in human environments, in this work, we combine these affordances, as functional cues, with other visual and geometric cues to improve the performance of scene labeling.

3. Representation of Human Configurations and Object Affordances

We first define the representation of the human configurations and object affordances in the following:

The Space of Human Configurations. A human configuration is comprised of its pose (relative position for every joint), location and orientation. Each pose could be at any X-Y-Z location and in different orientations $\in [0, 2\pi)$ inside the scene. For poses, we considered human poses from real human activity data (Cornell Activity Dataset-60, [29]), and clustered them using k-means algorithm giving us six types (three sitting poses and three standing poses) of skeletons showing in Fig. 2.

Figure 2: Six types of human poses extracted from Kinect 3D data. From left: sitting upright, sitting reclined, sitting forward, reaching, standing and leaning forward.

The Space of Object Affordances. A human can use the objects at different distances and orientations from the human body. For example, small hand-held devices are typically held close to the human and in front of the person. Other objects are used typically at a distance, such as a TV and decoration pieces. For the goal of scene labeling, we are interested in probability distribution of the 3D location of the objects around humans. For visualization, we show these probability distribution as heat-maps from top-view and side-view (e.g., see Fig. 6 and Section 5.3).

4. Human Context: a Double-Edged Sword

The human context is very important for understanding our environment. In fact, even when no human is present in an indoor scene, the potential human-object interactions give such a strong cue for scene understanding that we want to model it as latent variables in our algorithms.

However, the human context cannot be easily harnessed because the space of possible human configurations and object affordances is rather large. Furthermore, the humans are not always observable and such latent nature leads to an ill-posed problem while using it. For example, one potential explanation of the scene could be humans floating in the air and prefer stepping on every object as the affordance! The key to modeling the large space of latent human context lies in building *parsimonious* models and providing *priors* to avoid physically-impossible models.

4.1. Model Parsimony

While there are infinite number of human configurations in a scene and countless ways to interact with objects, only a few human poses and certain common ways of using objects are needed to explain most parts of a scene. These could be shared across objects and be instantiated to numerous forms in reality. We will do so by representing them as 'topics,' according to which objects in a scene are generated. This is analogous to the document topics [30, 18], except that in our case topics will be continuous distributions and factored. Similar to document topics, our human-context topics can be *shared* across objects and scenes. As a result, the model's complexity, i.e., the number of parameters, is significantly reduced. We describe the two types of topics below:

Human Configuration Topics. In a scene, there are certain human configurations that are used more commonly than others. For instance, in an office a sitting pose on the chair and a few poses standing by the desk, shelf and whiteboard are more common. Most of the objects in an office are arranged for these human configurations.

Object Affordance Topics. An object affordance, despite its large variety, can often be represented as a mixture of several commonly shared object-affordance topics. For example, both using a keyboard and reading a book require a human pose to be close to objects. However, when books are not in use, they can be stored away from humans. Therefore, the affordance of a book would be a mixture of a 'close-to' and a 'spread-out' topic.

4.2. Physics-Based Priors

In order to obtain meaningful human-configuration and object-affordance topics, we impose prior that follows physics and conventions to those topics.

Human Configuration Prior. Our hallucinated human configurations need to follow basic physics. Encoding physics-based notions about objects has been shown to be successful in 3D geometric interpretation [28, 15]. We consider the following two properties as priors for the generated human configurations [10]: *1) Kinematics.* We perform collision check so that the human pose is kinematically feasible. *2) Dynamics.* We check if the human skeleton is supported by the nearby environments to ensure its stability.

Object Affordance Prior. In general, it is more likely for an object to be close to humans while being used. Furthermore, most objects' affordance should be symmetric in their relative orientation to the humans' left or right. We encode this information in the design of the function quantifying affordances and as Bayesian priors in the estimation of the function's parameters, see Section 5.3.

5. Infinite Factored Topic Model (IFTM)

In this work, we model the human configurations and object affordances as two types of 'factored' topics. In our previous work [18], we presented finite factored topic model that discovers different *types* of topics from text data. Each type of topic is modeled by an independent topic model and a data point is jointly determined by a set of topics, one from each type. By factorizing the original parameter (topic) space into smaller sub-spaces, it uses a small number of topics from different sub-spaces to effectively express a larger number of topics in the original space.

In this work, we extend our idea to Infinite Factored Topic Models (IFTM), which can not only handle multiple types of topics but also unknown number of topics in each type. Furthermore, unlike text data, our topics are *continuous* distributions in this work that we model using Dirichlet process mixture model (DPMM) [30]. In the following, we first briefly review DPMM, and then describe our IFTM and show how to address the challenges induced by the coupling of the topics from different types.

5.1. Background: Dirichlet Process Mixture Model

A DPMM describes a generative process of drawing data point x from a set of topics, each of which is parameterized by θ_k. Specifically, it first draws infinite number of topics from a base distribution G, and the topic proportion $\boldsymbol{\pi}$:

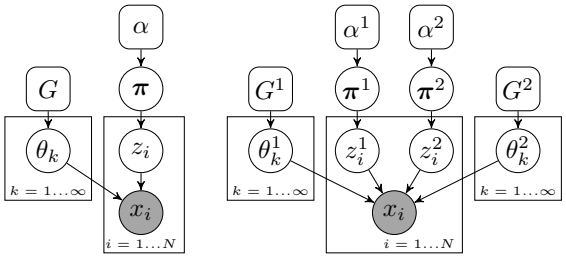
$$\theta_k \sim G, \;\; b_k \sim Beta(1, \alpha), \;\; \pi_k = b_k \textstyle\prod_{i=1}^{k-1}(1 - b_i).$$

Then each data point x is drawn from one of the topics. The topic assignment z is sampled from the topic proportion $\boldsymbol{\pi}$.

$$z|\boldsymbol{\pi} \sim \boldsymbol{\pi}; \qquad\qquad x|z, \theta \sim F(\theta_z).$$

Figure 3(a) shows the graphical representation of DPMM. In practice, z_i is sampled according to the Chinese restaurant process:

$$z_i = z|\boldsymbol{z}^{-i} = \begin{cases} \frac{n_z^{-i}}{N-1+\alpha} & \text{if } z \text{ is previously used} \\ \frac{\alpha}{N-1+\alpha} & \text{otherwise} \end{cases} \tag{1}$$



(a) DPMM (b) 2D infinite factored topic model

Figure 3: DPMM and our 2D infinite factored topic model.

where superscript $-i$ denotes everything except the i^{th} instance, n_z^{-i} equals the number of data points assigned to the component z excluding x_i, and N is the total number of data-points. Using this conditional distribution, one can apply Gibbs sampling method to approximate the marginal distribution of z_i and θ_k.

DPMM is different from traditional mixture models because of that it incorporates base (prior) distribution of topics and it allows the number of topics change according to data. These two properties are desired for modeling our human context because we need to encode prior as described in Section 4.2, and because the number of topics (i.e., the number of affordances and human poses) is unknown and can vary from scene to scene.

5.2. IFTM

In our IFTM, data is generated *jointly* by several independent topics. Particularly, an L-dimensional IFTM has L different mixture models, each having K^ℓ components parameterized by θ_k^ℓ. Now, generating a data point x involves choosing a topic $z^\ell \in \{1, \dots, K^\ell\}$ for *each* of the L dimensions. Given $\Theta = (\theta_k^\ell)_{k=1\dots K^\ell}^{\ell=1\dots L}$ and $\boldsymbol{z} = (z^1, \dots, z^L)$, we then draw x from the distribution parameterized by the selected L topics together:

$$z^\ell | \boldsymbol{\pi}^\ell \sim \boldsymbol{\pi}^\ell,\ \ell = 1, \dots, L; \qquad x | \boldsymbol{z}, \Theta \sim F(\theta_{z^1}^1, \dots, \theta_{z^L}^L).$$

Note that the domain of the density function F is now a Cartesian product of the domains of all L types of topics. In our scene labeling application, we have two types of topics, i.e., $L = 2$. In this case, the difference between DPMM to our ITFM is shown in Fig. 3.

Since the topic assignment $\boldsymbol{z}$ is now a L-dimensional vector. Directly applying Eq. (1) to compute the conditional distribution is computationally inappropriate: First, the complexity increases to $O(\prod_{\ell=1}^L K^\ell)$. Second, due to the sparsity in $n_{\boldsymbol{z}}$, $\boldsymbol{z}$ would tend to uniformly distributed. However, since our model assumes the two types of topic spaces are independent, it is easy to show that the distribution of $\boldsymbol{z}$ can be factorized into the distribution of z^ℓ, each of which follows Eq. (1).

5.3. IFTM for Human Context

We apply IFTM to learn the two types of human-context topics. In the following, we use superscript H and O to distinguish symbols for human-configuration topic and object-affordance topic. Namely, we use (G^H, G^O) to denote their prior distribution, (θ^H, θ^O) for the topic's parameters, and (z_i^H, z_i^O) for the two topic assignments for object x_i.

Specifically, x_i is the 3D location of i^{th} object in the scene. Its distribution F should reflect the likelihood of the object being at this location given the human configuration and affordance. We therefore define F as (see [14]):

$$F(x_i; \theta^H, \theta^O) = F_{\text{dist}} F_{\text{rel}} F_{\text{height}}, \tag{2}$$

where the three terms depict three types of spatial relationships between the object x_i and the human pose θ^H: Euclidean distance, relative angle and height (vertical) distance. We use log-normal, von Mises and normal distributions to characterize the probability of these measurements, and the parameters of these distributions are given by the object affordance topics, i.e., θ^O. The prior of these parameters, G^O, are set to Normal distributions with large variance. G^H is a uniform distribution over *valid* human poses in the scene (see Section 4.2).

5.4. Learning Human-Context Topics

Given a scene, the location of an object x_i is observed and our goal is to estimate likely human configurations and affordances in the scene. We use Gibbs sampling with auxiliary parameters [23] to sample θ^H and θ^O from their posterior distribution. The process consists of two steps:

Step 1: Sampling topic assignments. The general idea is, the distribution of z_i^H or z_i^O is proportional to two factors: 1) the likelihood, i.e., $F(x_i; \theta^H, \theta^O)$ with fixed topics; 2) the percentage of other objects also choosing the same topic. Moreover, to incorporate the growth of the number of topics, we add m auxiliary topics for a subject to choose from [23]. These auxiliary topics are drawn from the base distribution G^H or G^O, and the probability of choosing one of these topics is equal to α^H/m or α^O/m. So the topic assignments are sampled by:

$$z_i^H = z \quad \propto \begin{cases} \frac{n^H_{-i,z}}{N+m-1+\alpha^H} F(x_i, \theta_z^H, \theta^O_{z_i^O}) & n^H_{-i,z} \geq 0, \\ \frac{\alpha^H/m}{N+m-1+\alpha^H} F(x_i, \theta_z^H, \theta^O_{z_i^O}) & \text{otherwise} \end{cases}$$

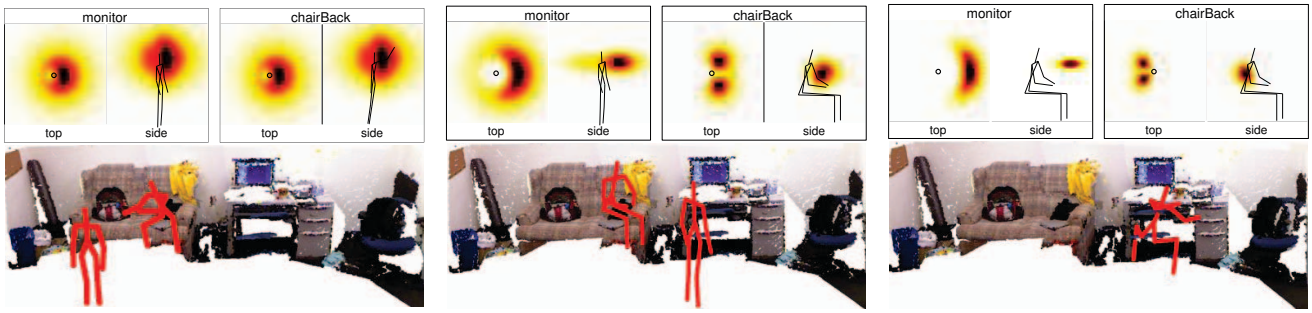
$$z_i^O = z \quad \propto \begin{cases} \frac{n^O_{-i,z}}{N+m-1+\alpha^O} F(x_i, \theta^H_{z_i^H}, \theta_z^O) & n^O_{-i,z} \geq 0, \\ \frac{\alpha^O/m}{N+m-1+\alpha^O} F(x_i, \theta^H_{z_i^H}, \theta_z^O) & \text{otherwise} \end{cases}$$

where $n^O_{-i,z}$ (or $n^H_{-i,z}$) is the number of other subjects x_j ($j \neq i$) with $z_j^O = z$ (or $z_j^H = z$), and N is the total number of objects in the scene.

Step 2: Sampling topics. Given topic assignments, we can compute the posterior distribution of topics and sample topics from it:

$$\theta_k^H = \theta^H \quad \propto \quad G^H(\theta^H) \times \textstyle\prod_{i:z_i^H=k} F(x_i, \theta^H, \theta^O_{z_i^O})$$

$$\theta_j^O = \theta^O \quad \propto \quad G^O(\theta^O) \times \textstyle\prod_{i:z_i^O=j} F(x_i, \theta^H_{z_i^H}, \theta^O)$$



(a) Iteration #1 (b) Iteration #10 (c) Iteration #100

Figure 4: An illustration of learning the object-affordance topics (top row, shown as heatmaps) and human-configuration topics (bottom) using IFTM. It shows two affordance topics, labeled with the most common object label for understanding in this figure. For each, it also shows the most probable human pose. In Iteration#1, the affordance is only based on the prior G^O and hence is same for all objects. In later iterations, our learning algorithm (Section 5.4) converges to reasonable topics.

In practice, the posterior sampling may be difficult when not using conjugate priors. To handle this, we use the maximum a posteriori (MAP) estimate instead of sampling. For example, an affordance topic $\theta^O_j = (\mu_d, \sigma_d, \mu_r, \kappa_r, \mu_h, \sigma_h)$ is updated as follows:

The mean and variance (μ_d, σ_d) *in* F_{dist}. Given the distance between each object x_i and its associated human pose $\theta^H_{z^H_i}$, denoted by d_i, μ_d and σ_d are given by,

$$\mu_d, \sigma_d = \arg\max_{\mu,\sigma} G^O_{\text{dist}}(\mu, \sigma) \prod_{i:z^O_i=j} F_{\text{dist}}(d_i; \mu, \sigma)$$

The mean and concentration (μ_r, κ_r) *in* F_{rel}. Let r_i be the angular difference between the object and the orientation of the human pose, then

$$\mu_r, \kappa_r = \arg\max_{\mu,\kappa} G^O_{\text{rel}}(\mu, \kappa) \prod_{i:z^O_i=j} F_{\text{rel}}(r_i; \mu, \kappa)$$

The mean and variance (μ_h, σ_h) *in* F_{height}. Let h_i be the vertical difference, then

$$\mu_h, \sigma_h = \arg\max_{\mu,\sigma} G^O_{\text{height}}(\mu, \sigma) \prod_{i:z^O_i=j} F_{\text{height}}(h_i; \mu, \sigma)$$

We illustrate the learning process in Fig. 4. It shows how θ^O and θ^H are sampled and refined progressively.

6. Affordance Features for Scene Labeling

Once having learned the human-context topics, how can we use them for scene labeling? In the task of 3D scene labeling, the goal is to identify the class of each object in the scene. As the object affordance is often strongly coupled with the object classes, we use the affordance derived from the learned topics as features that feed into other learning algorithms, similar to the ideas using in supervised topic models [3]. Although IFTM itself is an unsupervised method, in order to obtain more category-oriented topics, we initialize z^O_i to its object category to encourage topics to be shared by objects from the same class exclusively. Note that when computing the affordance features (for both training and test data), no object labels are used.

In detail, we compute the affordance features as follows. We set the affordance topics as the top K sampled topics θ^O_k, ranked by the posterior distribution. Given a test scene, repeatedly sample z^H_i and z^O_i and θ^H_k same as in the learning phase. Then we use the histogram of sampled z^O_i as the affordance features for object i.

7. Experiments and Results

7.1. Experimental Setting

Data. We used the Cornell RGB-D indoor dataset [19, 2] for our experiments. This data consists full-scene RGB-D point clouds of 24 offices and 28 homes obtained from 550 RGB-D views. The point-clouds are over-segmented based on smoothness, and the goal is to label these segments with object labels and attribute labels. Each segment can have multiple attribute labels but has only one object label. The attribute labels are: {*wall, floor, flat-horizontal-surfaces, furniture, fabric, heavy, seating-areas, small-objects, table-top-objects, electronics*} and the object labels are: {*wall, floor, tableTop, tableDrawer, tableLeg, chairBackRest, chairBase, chairBack, monitor, printerFront, printerSide, keyboard, cpuTop, cpuFront, cpuSide, book, paper, sofaBase, sofaArm, sofaBackRest, bed, bedSide, quilt, pillow, shelfRack, laptop*}.

Baselines. We perform 4-fold cross-validation where we train the model on data from three folds and tested on the fourth fold of the unseen data. Table 1 presents the results for object labeling and attribute labeling. In order to study the effects of different algorithms, we compare with the following algorithms:

(a) Appearance. We run versions with both local image and shape features [19].

(b) Human Context (Affordances). This is our affordance and human configurations information being used in prediction, without using object-object context.

(c) Object-Object context. In this case, we use the learning algorithm presented in [19] that uses Markov Random Field with log-linear node and pairwise edge potentials.

Table 1: **Object and Attribute Labeling Results.** The table shows average micro precision/recall, and average macro precision and recall for home and office scenes. Computed with 4-fold cross-validation.

Algorithm	Image & Shape	Human Context	Obj-obj Context	Object Labeling						Attribute Labeling							
				Office Scenes			Home Scenes			Office Scenes				Home Scenes			
				micro	macro		micro	macro		micro		macro		micro		macro	
				P/R	prec	recall	P/R	prec	recall	prec	recall	prec	recall	prec	recall	prec	recall
chance				5.88	5.88	5.88	5.88	5.88	5.88	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5
max class				26.33	26.33	5.88	29.38	29.38	5.88	22.89	22.89	22.89	12.5	31.4	31.4	31.4	12.5
Affordances		✓		29.13	16.28	16.67	33.62	16.37	15.30	47.93	32.04	42.85	29.83	53.92	36.07	41.19	26.21
Appearance	✓			77.97	69.44	66.23	56.50	37.18	34.73	85.82	66.48	86.58	62.52	77.80	55.21	60.01	42.20
Afford. + Appear.	✓	✓		79.71	73.45	69.76	59.00	38.86	37.54	87.05	68.88	**87.24**	65.42	79.02	59.02	70.45	46.57
Koppula et al. [19]	✓		✓	84.06	80.52	72.64	**73.38**	56.81	54.80	87.92	71.93	84.04	67.96	83.12	70.03	76.04	58.18
Full Model	✓	✓	✓	**85.22**	**83.20**	**74.11**	72.50	**59.07**	**56.02**	**88.40**	**76.73**	85.58	**74.16**	**83.42**	**70.28**	**79.93**	**64.27**

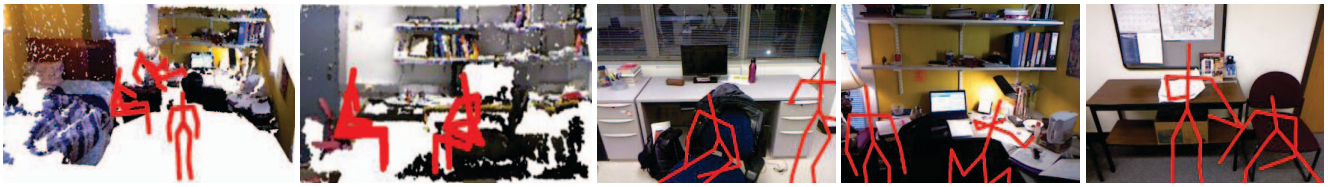


Figure 5: Top sampled human poses in different scenes. The first two are from stitched point-cloud from multiple RGB-D views, and the last three scenes are shown in RGB-D single views.

(d) Full model. Here we combine the human context (from affordances and human configurations) with object-object context. In detail, we append the node features of each segment with the affordance topic proportions derived from the learned object-affordance topics and learn the semantic labeling model as described in [19].

We report precision and recall using both micro and macro aggregation. Since we predict only one label for each segment in case of predicting object labels, our micro precision and recall is the same as the percentage of correctly classified segments. The macro precision and recall are the average of precision and recall of all classes respectively.

7.2. Results and Discussion

Table 1 shows that our algorithm performs better than the state-of-the-art in both object as well as attribute labeling experiment. Our approach is able to predict the correct labels for majority of the classes as can be seen from the strong diagonal in the confusion matrices. We discuss our results in the light of the following questions.

Are the sampled human poses meaningful? Being able to hallucinate sensible human poses is critical for learning object affordances. To verify that our algorithm can sample meaningful human poses, we plot a few top sampled poses in the scenes, shown in Fig. 5. In the first home scene, some sampled human poses are sitting on the edge of the bed while others standing close to the desk (so that they have easy access to objects on the table or the shelf-rack). In the next office scene (Fig. 5-b), there is one L-shaped desk and two chairs on each side. It can be seen that our sampled human poses are not only on these chairs but also with correct orientation. Also, as can be seen in Fig. 4-c, our algorithm successfully identifies the workspaces in the office scene. Note that these poses naturally explain why the monitors, keyboards and CPUs are arranged in this particular way. It is these correctly sampled human poses that give us possibility to learn correct object affordances.

Are the discovered affordances meaningful? During training, we are given scenes with the objects and their labels, but not humans. Our goal is to learn object affordance for each class. Fig. 6 shows the affordances from the top-view and side-view respectively for typical object classes. Here the X-Y dimensions of the box are 5m×5m, and the height axis's range is 3m. The person is in the center of the box. From the side views, we can see that for objects such as wall and cpuTop, the distributions are more spread out compared to objects such as floor, chairBase and keyboard. This is because that that chairBase is often associated with a sitting pose at similar heights, while CPUs can either be on the table or on the floor. While this demonstrates that our method can learn meaningful affordances, we also observe certain biases in our affordances. For example, the wall is more to the front as compared to the back, and monitor is biased to the side. We attribute to the limited data and imperfect generation of valid human skeletons. Note that while the affordance topics are unimodal, the affordance for each objects is a mixture of these topics and thus could be multimodal and more expressive.

Can we obtain object-object relations from object affordances? Since objects are related to humans, it turns out that we can infer object-object spatial relations (and object co-occurences) from the human-object relations. For example, if we convolve keyboard-human and human-monitor relations, we obtain the spatial relations between keyboard and monitor. More formally, we compute the conditional distribution of one object x_i given another x_j as,

$$P(x_i|x_j) = \int P(x_i|\theta^H)P(\theta^H|x_j)d\theta^H$$

$$\propto \int F(x_i;\theta^H,\theta^O_{z_i^O})F(x_j;\theta^H,\theta^O_{z_i^O})G_0(\theta^H)d\theta^H$$

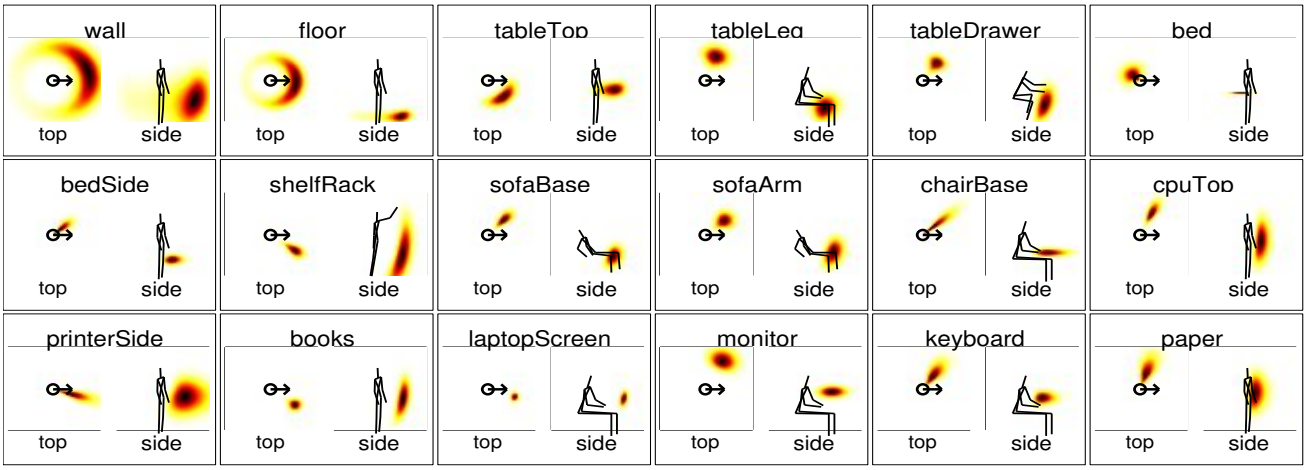


Figure 6: Examples of learned object-affordance topics. An affordance is represented by the probabilistic distribution of an object in a $5 \times 5 \times 3$ space given a human pose. We show both projected top views and side views for different object classes.

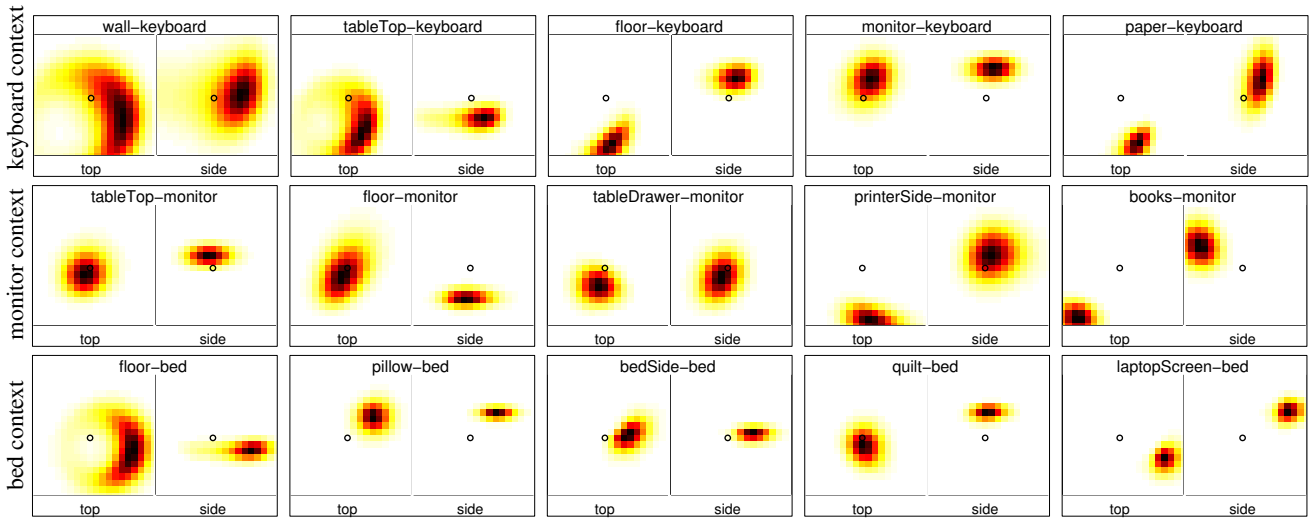


Figure 7: Object-object context obtained from our learned human context. Each pair of the top- and side-view of a heatmap with the title of 'obj1-obj2' shows the distribution of obj1 given obj2 at the center facing right. For example, in the first row the keyboard is in the center of the image and the heat-maps show the probability of finding other related objects such as table top, monitor, etc.

This illustrates that for n objects, we can model $O(n^2)$ object-object relations with only $O(n)$ human-object parameters. Some examples are shown in Fig. 7. We can find that many object-object relationships are recovered reasonably from our learned affordances. For example, given a keyboard, a monitor is likely to be found in front of and above it while tableTop at the same height as it (sometimes above it as the keyboard is often in a keyboard-tray in offices). In home scenes, given a bed, we can find a pillow on the head of the bed, quilt right above the bed and bedSide slightly below it. This supports our hypothesis that object-object relations are only an artifact of the hidden context of human-object relations.

Does human context helps in scene labeling? Table. 1 shows that the affordance topic proportions (human context) as extra features boosts the labeling performance. First, when combining human context with the image- and shape-features, we see a consistent improvement in labeling performance in all evaluation metrics, regardless of the object-object context. Second, when we add object-object context, the performance is further boosted in the case of office scenes and improves marco precision for home scenes. This indicates that there is some orthogonality in the human-object context and object-object context. In fact, adding object-object context to human-object context was particularly helpful for small objects such as keyboards and books that are not always used by humans together, but still have a spatial correlation between them.

We also show the confusion matrices in Fig. 8. We found that while our algorithm can distinguish most of the objects, it sometimes confuses objects with similar affordance. For example, it confuses pillow with quilt and confuses book

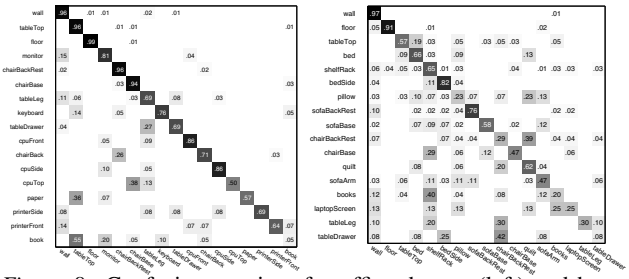


Figure 8: Confusion matrices for office dataset (left) and home dataset (right) using the full model.

and paper with tableTop. Similarly, it confuses cpuTop with chairBase because the CPU-top (placed on the ground) could also afford sitting human poses!

Finally, we applied our method in the task of a robot detecting and arranging objects in a room. For the robot video (along with the data and code), please visit:
http://pr.cs.cornell.edu/hallucinatinghumans/

8. Conclusions

We presented infinite factored topic models (IFTM) that enabled us to model the generation of a scene containing objects through hallucinated (hidden) human configurations and object affordances, both modeled as topics. Given only a set of scenes containing objects, we showed that we can discover meaningful human-object relations (affordances). We then showed that such modeling improved the performance of object and attribute labeling tasks over the state-of-the-art.

Acknowledgements: This research was funded by Microsoft Faculty Fellowship and Sloan Fellowship to Saxena.

References

[1] E. Aksoy, A. Abramov, F. Worgotter, and B. Dellen. Categorizing object-action relations from semantic scene graphs. In *ICRA*, 2010.

[2] A. Anand, H. S. Koppula, T. Joachims, and A. Saxena. Contextually guided semantic labeling and search for 3D point clouds. *IJRR*, 32(1):19–34, 2013.

[3] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *NIPS*, 2007.

[4] V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. Efros. Scene semantics from long-term observation of people. In *ECCV*, 2012.

[5] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[6] D. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single view geometry. In *ECCV*, 2012.

[7] J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

[8] H. Grabner, J. Gall, and L. J. V. Gool. What makes a chair a chair? In *CVPR*, 2011.

[9] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *PAMI*, 31(10):1775–1789, 2009.

[10] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011.

[11] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, 2010.

[12] G. Heitz, S. Gould, A. Saxena, and D. Koller. Cascaded classification models: Combining models for holistic scene understanding. In *NIPS*, 2008.

[13] Z. Jia, A. Gallagher, A. Saxena, and T. Chen. 3d-based reasoning with blocks, support, and stability. In *CVPR*, 2013.

[14] Y. Jiang, M. Lim, and A. Saxena. Learning object arrangements in 3d scenes using human context. In *ICML*, 2012.

[15] Y. Jiang, M. Lim, C. Zheng, and A. Saxena. Learning to place new objects in a scene. *IJRR*, 31(9):1021–1043, 2012.

[16] Y. Jiang, S. Moseson, and A. Saxena. Efficient grasping from rgbd images: Learning using a new rectangle representation. In *ICRA*. IEEE, 2011.

[17] Y. Jiang and A. Saxena. Hallucinating humans for learning robotic placement of objects. In *ISER*, 2012.

[18] Y. Jiang and A. Saxena. Discovering different types of topics: Factored topic models. In *IJCAI*, 2013.

[19] H. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *NIPS*, 2011.

[20] H. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 2013.

[21] 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.

[22] C. Li, A. Kowdle, A. Saxena, and T. Chen. Towards holistic scene understanding: Feedback enabled cascaded classification models. In *NIPS*, 2010.

[23] R. Neal. Markov chain sampling methods for dirichlet process mixture models. *J comp graph stats*, 9(2), 2000.

[24] D. Norman. *The Design of Everyday Things*. 1988.

[25] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

[26] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *NIPS 18*, 2005.

[27] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE PAMI*, 31(5):824–840, 2009.

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

[29] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from RGBD images. In *ICRA*, 2012.

[30] Y. W. Teh. Dirichlet process. *Encyclopedia of Machine Learning*, pages 280–287, 2010.

[31] X. Xiong and D. Huber. Using context to create semantic 3d models of indoor environments. In *BMVC*, 2010.

[32] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.