Infinite Latent Conditional Random Fields

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

Abstract

In this paper, we present Infinite Latent Conditional Random Fields (ILCRFs) that model the data through a mixture of CRFs generated from Dirichlet processes. Each CRF represents one possible explanation of the data. In addition to visible nodes and edges that exist in classic CRFs, it generatively models the distribution of different CRF structures over the latent nodes and corresponding edges, imposing no restriction on the number of both nodes and types of edges. We apply ILCRFs to several applications, such as robotic scene arrangement and scene labeling, where a scene is modeled through, not only objects, but also latent human poses and human-object relations. In extensive experiments, we show that our model outperforms the state-of-the-art results as well as helps a robot placing objects in a new scene. [1]

1. Introduction

In this work, we are interested in modeling hidden causes in the data, including latent variables as well as how they related to observations. Let us consider scene modeling as an example. A human environment is constructed under two types of relations: *object-object* and *human-object relations*. When only considering object-object relations, Conditional random fields (CRFs) are a natural choice, as each object can be modeled as a node in a Markov network and the edges in the graph can reflect the object-object relations. In fact, CRFs and their variants have thus been applied to many scene modeling tasks (e.g., [20, 1, 19, 21]). While objects are easy to observe, humans are often not. However, modeling them as latent factors in the scene can be beneficial, as it is human who made the scene as it is.

Modeling possible human poses and human-object interactions (or *object affordances*) is not trivial because of several reasons. First, there can be any number of possible humans in a scene—e.g., in an office scene such as Fig. 1, some sitting on the couch/chair, some standing by the shelf/table. Second, there can be various types of human-object interactions in a scene, such as watching TV in distance, eating from dishes, or working on a laptop, etc. Third, an object can be used by different human poses, such as a book on the table can be accessed by either a sitting pose on the couch or a standing pose nearby. Last, there can be multiple possible usage scenarios in a scene. Therefore, we need models that can incorporate latent factors, latent structures, as well as different alternative possibilities.

To admit those properties, we propose infinite latent conditional random fields (ILCRFs). Intuitively, it is a mixture of CRFs where each CRF can have two types of nodes: existing nodes (e.g., object nodes, which are given in the graph and we only have to infer the value) and latent nodes (e.g., human nodes, where an unknown number of humans may be hallucinated in the room). The relations between the nodes (object-object edges and human-object edges) could also be of different types. Unlike traditional CRFs, where the structure of the graph is given, the structure of our ILCRF is sampled from Dirichlet Processes (DPs). DPs are widely used as nonparametric Bayesian priors for mixture models, the resulting DP mixture models can determine the number of components from data, and therefore is also referred as infinite mixture model. ILCRFs are inspired by this, and we call it 'infinite' as it can sidestep the difficulty of finding the correct number of latent nodes as well as latent edge types. Our learning and inference methods are based on Gibbs sampling that samples latent nodes, existing nodes, and edges from their posterior distributions.

The idea of ILCRF can be applied to many tasks, such as scene labeling [7], robotic scene arrangement [11] and even document modeling [10]. We instantiate two specific ILCRFs for two applications: scene arrangement where the objective is to find proper placement (including 3D location and orientation) of given objects in a scene, and scene labeling where the objective is to identify objects in a scene. Despite the disparity of the tasks at the first look, we relate them through one common hidden cause—imaginary humans and object affordances. For both tasks, our ILCRF models each object placement as an existing node, hallucinated human poses as latent nodes and spatial relationships

[1] This work was previously published as [11, 7].

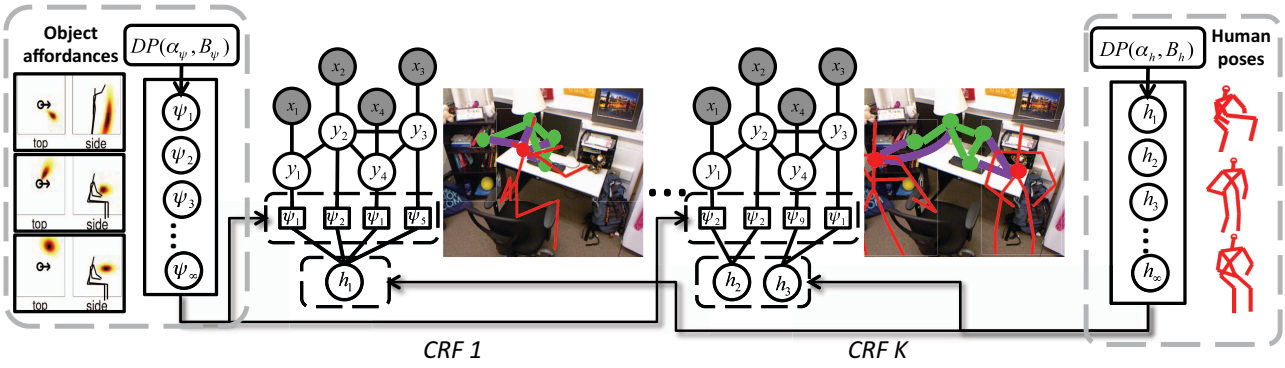

Figure 1: An example of instantiated ILCRF for scene understanding. A scene with objects and hallucinated humans in it can be explain by a latent CRF: Each y node represents an object and x represents its observation; h nodes are latent human configurations, and their links to y, parameterized by ψ, describe the interactions between humans and objects (referred as 'object affordances'). Our ILCRF reasons a scene through a mixture of multiple latent CRFs, generated by two DPs: one for human poses (right) and one for object affordances (left). See Sec. 2.2 for more details.

among objects or between objects and humans as edges. We demonstrate in the experiments that this unified model achieves the state-of-the-art results. We demonstrate in the experiments that this model achieves the state-of-the-art results, and it also helps a robot successfully identify the class of objects in a new room, and placed several objects correctly in it.

2. Infinite Latent Conditional Random Fields

In this paper, we propose a type of mixture CRFs—infinite latent conditional random fields (ILCRFs), which can capture the following properties:

1. *Unknown number of latent nodes.*
2. *Unknown number of* ***types*** *of potential functions.*
3. *Mixture CRFs.*
4. *Informative priors on the structure of CRFs.*

We achieve this by imposing Bayesian nonparametric priors—Dirichlet processes (DPs)—to the latent variables, potential functions and graph structures.

2.1. Background: Dirichlet Process Mixture Model

Dirichlet process [23] is a stochastic process to generate distributions that are used to model clustering effects in the data, especially when the number of clusters is *unknown*.

A DP mixture model, $DP(\alpha, B)$, defines the following generative process, with a concentration parameter α and a base distribution B:

1. Generate infinite number of mixture components, parameterized by $\Theta = \{\theta_1, \ldots, \theta_\infty\}$, and their mixture weights $\boldsymbol{\pi}$:

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

2. Assign the z_i^{th} component to each data point x_i and draw from it:

$$z_i \sim \boldsymbol{\pi}, \qquad x_i \sim F(\theta_{z_i}). \tag{2}$$

2.2. ILCRF

ILCRF uses DPs to admit an arbitrary number of latent variables and potential functions. In brief, it generates latent variables and potential functions from two DPs respectively, and each data point builds a link, associated with one potential function, to one latent variable. Different samples thus form different CRFs.

Definition 1 *A ILCRF$(\mathcal{X}, \mathcal{Y}, E_Y, \alpha_h, B_h, \alpha_\psi, B_\psi)$ is a mixture of CRFs, where the edges in $\mathcal{Y}$ are defined in graph E_Y and latent variables $\mathcal{H}$ as well as the edges between $\mathcal{H}$ and $\mathcal{Y}$ are generated through the following process:*

1. *Generate infinite number of latent nodes $\mathcal{H} = \{h_1, h_2, \ldots, h_\infty\}$ and a distribution $\boldsymbol{\pi}_h$ from a DP process $DP(\alpha_h, B_h)$ following Eq. (1); Assign one edge to each label y_i that links to h_{z_i}, where $z_i \sim \boldsymbol{\pi}_h$ following Eq. (2).*

2. *Generate infinite number of potential functions ('types' of edges) $\Psi = \{\psi_1, \ldots, \psi_\infty\}$ and a distribution $\boldsymbol{\pi}_\psi$ from a DP process $DP(\alpha_\psi, B_\psi)$ following Eq. (1); Assign one potential function ψ_{ω_i} to each edge (y_i, h_{z_i}), where $\omega_i \sim \boldsymbol{\pi}_\psi$ following Eq. (2).* ■

We will illustrate the process using Figure 1. Consider first sampled CRF ('CRF-1' in the figure) with four visible nodes y_i $(i = 1 \ldots 4)$. In the first step, all the y_i's are connected to h_1, because z_i's are sampled as 1 from $DP(\alpha_h, B_h)$. Meanwhile, we also draw the value of h_1 from $DP(\alpha_h, B_h)$. Thus, we get a CRF with one latent node. In the second step, the potential function of edge (y_1, h_1) is assigned to ψ_1, (y_2, h_1) to ψ_2, (y_3, h_1) to ψ_5 and (y_4, h_1) to ψ_1. This is because ω_i's are sampled as $(1, 2, 5, 1)$ from $DP(\alpha_\psi, B_\psi)$. Since, only (ψ_1, ψ_2, ψ_5) are active, we have three edge types in this CRF. We draw their parameters from $DP(\alpha_\psi, B_\psi)$. Repeating this procedure

Table 1: The Gibbs sampling of ILCRF in the two applications.

Application	Phase	Gibbs sampling (Sect. 2.3)				
		$\boldsymbol{z}$ (3)	$\mathbf{h}$ (4)	$\boldsymbol{\omega}$ (5)	$\boldsymbol{\psi}$ (6)	$\mathcal{Y}$ (7)
Scene arrangement	Training	✓	✓		✓	
	Testing	✓	✓			✓
Scene labeling	Training	✓	✓		✓	
	Testing	✓	✓	✓		✓

may generate different latent CRFs such as *'CRF-K'* which has two different latent nodes and three different edge types. In the end, their mixture forms the ILCRF.

2.3. Gibbs Sampling for Learning and Inference

Inspired by the Gibbs sampling algorithm for DP mixture models [17], here we present our sampling algorithm:

- Sample the graph structure, i.e., one edge for each y_i to one latent node:

$$z_i = z \propto \begin{cases} \frac{n^h_{-i,z}}{n+m-1+\alpha_h}\psi_{\omega_i}(y_i, h_z) & n^h_{-i,z} \geq 0, \\ \frac{\alpha_h/m}{n+m-1+\alpha_h}\psi_{\omega_i}(y_i, h_z) & \text{otherwise} \end{cases} \quad (3)$$

- Sample values for each latent node in the graph:

$$h_k = h \propto B_h(h) \times \prod_{i:z_i=k} \psi_{\omega_i}(y_i, h) \quad (4)$$

- Assign the type of potential functions to each edge:

$$\omega_i = \omega \propto \begin{cases} \frac{n^\psi_{-i,\omega}}{n+m-1+\alpha_\psi}\psi_{\omega}(y_i, h_{z_i}) & n^\psi_{-i,\omega} \geq 0, \\ \frac{\alpha_\psi/m}{n+m-1+\alpha_\psi}\psi_{\omega_i}(y_i, h_{z_i}) & \text{otherwise} \end{cases} \quad (5)$$

- Sample the parameters of each selected potential function:

$$\psi_k = \psi \propto B_\psi(\psi) \times \prod_{i:\omega_i=k} \psi_{\omega}(y_i, h_{z_i}) \quad (6)$$

- Sample labels:

$$y_i = y \propto \psi_{\omega_i}(y, h_{z_i}) \times \prod_{(y_i,y_j)\in E} \psi(y_i, y_j) \quad (7)$$

As for learning the E_Y, when labels are given in the training data, E_Y is independent with latent variables $\mathcal{H}$ (if the partition function is ignored), and therefore can be learned separately.

2.4. Scene Arrangement

We apply ILCRFs to the application of 3D scene arrangement, where the goal is to find appropriate locations and orientations for placing given objects. We define $y_i \in \mathcal{Y}$ as the placement (location and orientation) of an object and $x_i \in \mathcal{X}$ as its given object class. The edges between the visible nodes $\mathcal{Y}$ model the object-object spatial relationships.[2]

[2]It is defined as a multi-variate Gaussian distribution of the location and orientation difference between the two objects.

We model possible human poses as latent nodes $\mathcal{H}$. A human pose is specified by its pose, location and orientation. Following [8], we use six types covering different sitting and standing poses.

We model object affordances as the potential functions Ψ. We use the spatial relationship between a human pose and the object to represents its affordance. It is defined as a product of several terms: Euclidean distance, relative angle, orientation difference, and height (vertical) distance. (See [8] for details.)

Learning. During training, our goal is to learn the object affordances (i.e., a set of potential functions ψ in the ILCRF). As shown in Table 1, we perform sampling on the human-object edges, human poses and object affordances, given placements $\mathcal{Y}$ and edge types ω_i, according to Eq. (3), (4) and (6). (Here, since x_i as the object class label is given, we set $\omega_i = x_i$ in this application.) The object-object structure, E_Y is learned based on object co-occurence, as computed from the training data.

Inference. During testing on a new scene, our goal is to predict placements y_i, given objects $\mathcal{X}$. We perform inference by sampling the human-object edges, human poses and placements, using the learned object affordances (see Eq. (3), (4) and (7)). In order to predict the most likely placement for object i, we choose the placement area sampled most because that represents the highest probability.

2.5. Scene Labeling

In this task, the goal is to identify the class of each object (represented as a segment in a 3D point cloud) in the scene. We define $y_i \in \mathcal{Y}$ as the object class of the segment and $x_i \in \mathcal{X}$ as its location and appearance features. The object-object edges are the spatial relationships as described in [1].

During training, we learn object affordances by sampling human-object edges, human poses and object affordances according to Eq. (3), (4) and (6). E_Y is learned separately using max-margin learning [1].

During testing a new scene, we sample all other variables in ILCRF, except the object affordances already learned from the training data, using Eq. (3), (4) and (5).

2.6. Related Work

To our best knowledge, there is little work about arranging/placing objects in robotics (e.g., [4, 22, 6, 12, 9]), and none of these works consider reasonable arrangements for *human usage*. Recent works [8, 7] considered hallucinating humans for object placements and scene labeling, but did not model human-object and object-object relationships in a joint model.

There are other recent works applying object affordances in tasks of predicting human workspaces [5], 3D geometry when humans are observed [3], improving human robot interactions [18], detecting and anticipating human activity [14, 15, 16].

Table 2: Results of arranging partially-filled scenes and arranging empty scenes in synthetic dataset, evaluated by the location and height difference to the labeled arrangements.

Algorithms	partially-filled scenes		empty scenes	
	location (m)	height (m)	location (m)	height (m)
Chance	2.35±0.23	0.41±0.04	2.31±0.23	0.42±0.05
CRF	1.69±0.05	0.12±0.01	2.17±0.07	0.39±0.01
Human+obj [8]	1.44±0.18	**0.09** ±0.01	1.63±0.19	0.11±0.01
FLCRF	1.55±0.06	0.12±0.01	1.63±0.06	0.14±0.01
ILCRF	**1.33±0.19**	**0.09±0.01**	**1.52±0.06**	**0.10±0.01**

Table 3: Object and Attribute Labeling Results. The table shows average micro precision/recall, and average macro precision and recall for 52 scenes. Computed with 4-fold cross-validation.

Algorithms	**Object Labeling**			**Attribute Labeling**			
	micro	macro		micro		macro	
	P/R	prec	recall	prec	recall	prec	recall
chance	5.88	5.88	5.88	12.50	12.50	12.50	12.50
Affordances	31.38	16.33	15.99	50.93	34.06	42.02	28.02
Appearance [13]	67.24	53.31	50.48	81.81	60.85	73.30	52.36
Afford. + Appear.	69.36	56.16	53.65	83.04	63.95	78.85	56.00
Koppula et al. [13]	78.72	68.67	63.72	85.52	70.98	80.04	63.07
ILCRF	**78.86**	**71.14**	**65.07**	**85.91**	**73.51**	**82.76**	**69.22**

Modeling latent variables has been successfully applied to many vision problems, such as scene understanding [24], object recognition [21] and gesture recognition [25, 2]. However, the labels and hidden states are discrete and take only finite number of values.

3. Experiments

In our application, the scenes (including objects/furniture) are perceived as point-clouds, either generated from 3D models in synthetic datasets or obtained using Microsoft Kinect camera in real datasets. For more details, see [11] and [7].

3.1. Scene Arrangement

Dataset. We use the same datasets as in [8, 9]: a synthetic dataset consisting of 20 rooms and 47 objects from 19 categories (book, laptop, light, utensil, etc.). we conduct 5-fold cross validation on 20 rooms so that the test rooms are new to the algorithms.

Algorithms. We compare all the following methods:
1) CRF, a ILCRF with only object-object edges, without latent human nodes. *2) Human+obj.* Linear combine the human- and object-context. [8]. *3) FLCRF,* a ILCRF with fixed number of latent nodes, same for all the scenes. *4) ILCRF*, our full model.

Results. Table 2 presents the results on the synthetic dataset, where the predicted arrangements are evaluated by two metrics, same as in [8]: location difference and height difference (in meters) to the labeled arrangements. From Table 2 we can see the performance gain of our full ILCRF model against ILCRF with only object edges (CRF) or Obj. that uses heuristic object context. Even methods that use human pose in naive ways (ILCRF-NSH and FLCRF) achieve better results than methods don't consider latent humans.

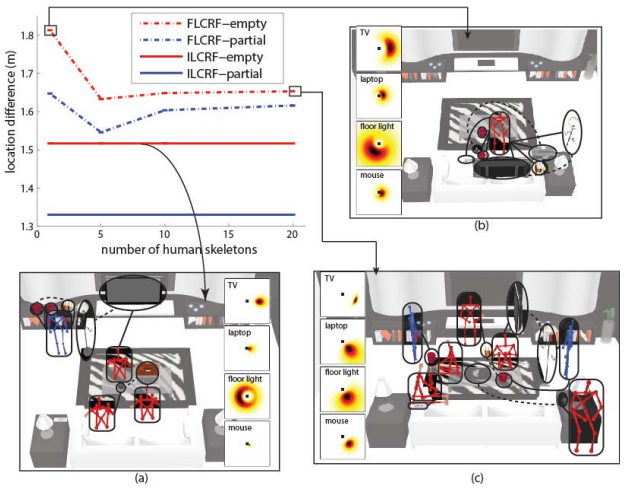

Figure 2: Results of FLCRF with different number of human poses versus ILCRF. We also show exemplar sampled CRFs and learned object affordances (in top-view heatmaps) by different methods.

The advantage of using DP mixture models in ILCRF is being able to determine the number of human poses from the data instead of guessing manually. In Fig. 2, we compare ILCRF against FLCRF with the number of human poses varying from 1 to 20. While having five poses in FLCRF gives the best result, it is still outperformed by ILCRF. This is because scenes of different sizes and functions prefer different number of human poses and incorrect number of humans may lead to meaningless affordances, either under-fit (Fig. 2-b) or over-fit (Fig. 2-c).

3.2. Scene Labeling

Dataset. We use the Cornell RGB-D indoor dataset [13, 1] for our experiments. It consists of 52 scenes labeled with 26 object classes and 10 attribute classes.

Algorithms.
1) Appearance. Only uses local image and shape features.
2) Affordances. Only uses human configuration features.
3) Afford.+Appear. ILCRF without object-object edges.
4) CRF, full model in [13]. *5) ILCRF*, our full model.

Results. Table 3 shows that our algorithm performs better than the state-of-the-art in both object as well as attribute labeling experiment.

One of our goals is to also learn reasonable object affordance for each class. Fig. 3 shows the affordances from the top-view and side-view respectively for typical object classes. From the side views, we can see that for objects such as wall and books, the distributions are more spread out compared to objects such as floor, bed and monitor. From the top view, some objects have strong angular preference such as laptop and keyboard, compared to objects such as floor and tableTop.

3.3. Robotic Experiment

Given a RGB-D scene, our robot first uses ILCRF to label the scene as well as hallucinating humans. Then, it uses ILCRF and those human poses to infer proper place-

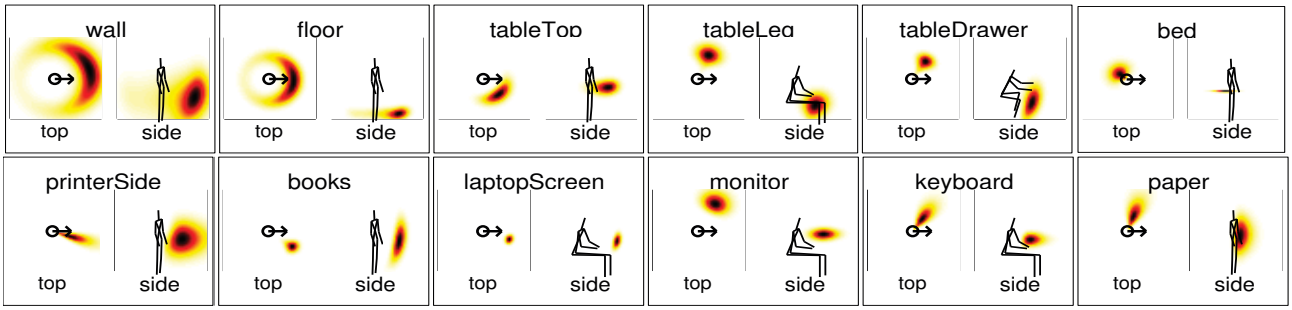

Figure 3: Examples of learned object-affordance topics. An affordance is represented by the probabilistic distribution of an object in a $5 \times 5 \times 3$ space given a human pose. We show both projected top views and side views for different object classes.

ments for new objects. To see our PR2 arranging the scene in action (along with code and data), please visit: http://pr.cs.cornell.edu/hallucinatinghumans

4. Conclusion

In this paper, we considered two challenging tasks of scene understanding, which require an algorithm that can handle: 1) unknown number of latent nodes (for potential human poses), 2) unknown number of edge types (for human-object interactions), and 3) a mixture of different CRFs (for the whole scene). We therefore presented a new algorithm, called Infinite Latent Conditional Random Fields (ILCRFs), together with learning and inference algorithms. Through extensive experiments, we showed that our ILCRF can understand scenes better by hallucinating reasonable human poses and learning their relations to objects. ILCRF also helps our robot to arrange a room in practice.

References

[1] A. Anand, H. Koppula, T. Joachims, and A. Saxena. Contextually guided semantic labeling and search for 3d point clouds. *IJRR*, 32(1):19–34, 2012. 1, 3, 4

[2] K. Bousmalis, S. Zafeiriou, L.-P. Morency, and M. Pantic. Infinite hidden conditional random fields forbreak human behavior analysis. In *TNNLS*, 2013. 4

[3] V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. Efros. Scene semantics from long-term observation of people. In *ECCV*, 2012. 3

[4] A. Edsinger and C. Kemp. Manipulation in human environments. In *Humanoid Robots*, 2006. 3

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

[6] D. Jain, L. Mosenlechner, and M. Beetz. Equipping robot control programs with first-order probabilistic reasoning capabilities. In *ICRA*, 2009. 3

[7] Y. Jiang, H. Koppula, and A. Saxena. Hallucinated humans as the hidden context for labeling 3d scenes. In *CVPR*, 2013. 1, 3, 4

[8] Y. Jiang, M. Lim, and A. Saxena. Learning object arrangements in 3d scenes using human context. In *ICML*, 2012. 3, 4

[9] Y. Jiang, M. Lim, C. Zheng, and A. Saxena. Learning to place new objects in a scene. *IJRR*, 2012. 3, 4

[10] Y. Jiang and A. Saxena. Discovering different types of topics: Factored topics models. 2013. 1

[11] Y. Jiang and A. Saxena. Infinite latent conditional random fields for modeling environments through humans. In *RSS*, 2013. 1, 4

[12] Y. Jiang, C. Zheng, M. Lim, and A. Saxena. Learning to place new objects. In *ICRA*, 2012. 3

[13] H. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *NIPS*, 2011. 4

[14] H. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 2013. 3

[15] H. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. In *RSS*, 2013. 3

[16] H. Koppula and A. Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *ICML*, 2013. 3

[17] R. Neal. Markov chain sampling methods for dirichlet process mixture models. *J comp graph statistics*, 9(2):249–265, 2000. 3

[18] A. Pandey and R. Alami. Taskability graph: Towards analyzing effort based agent-agent affordances. In *RO-MAN, IEEE*, 2012. 3

[19] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *NIPS*. Citeseer, 2004. 1

[20] A. Saxena, S. Chung, and A. Ng. Learning depth from single monocular images. In *NIPS 18*, 2005. 1

[21] P. Schnitzspan, S. Roth, and B. Schiele. Automatic discovery of meaningful object parts with latent crfs. In *CVPR*, 2010. 1, 4

[22] M. Schuster, J. Okerman, H. Nguyen, J. Rehg, and C. Kemp. Perceiving clutter and surfaces for object placement in indoor environments. In *Humanoid Robots*, 2010. 3

[23] Y. W. Teh. Dirichlet process. *Encyc. of Mach. Learn.*, 2010. 2

[24] H. Wang, S. Gould, and D. Koller. Discriminative learning with latent variables for cluttered indoor scene understanding. In *ECCV*, 2010. 4

[25] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *CVPR*, 2006. 4