Notes of "Human-centric Indoor Scene Synthesis Using Stochastic Grammar"

Jihang Li

October 17, 2019

Contents

Abstract

Human contexts as contextual relations are encoded by **Markov Random Field (MRF)** on the terminal nodes.

Distributions are learned from an indoor scene dataset.

New layers are sampled using **Monte Carlo Markov Chain (MCMC)**.

Sampling is based on three criteria:

1. Visual realism compare to SOTA room arrangement method.
2. Accuracy of the affordance maps w.r.t. ground-truth.

3. The functionality and naturalness of synthesized rooms evaluated by human subjects.

1 Introduction

Traditional methods of image data collection and labeling limitations:

1. High-quality ground truths are hard to obtain, as depth and surface normal obtained from sensors are always noisy.
2. Impossible to label certain ground truth information.
3. Manual labeling of massive ground-truth is tedious and error-prone.

The proposed algorithm is useful for (but not limited to):

1. Learning and inference for computer vision tasks.
2. 3D content generation.
3. 3D reconstruction and robot mappings.
4. Benchmarking of both low-level and high-level task-planning problems in robotics.

Four major difficulties in synthesizing:

1. The number of pieces may vary in a functional group. (e.g. a dinning set.)
2. There is already a quadratic number of object-object relation even if only pair-wise relations are considered.
3. Most object-object relations are not obviously meaningful. (e.g. a pen and a monitor both on a desk.)
4. An excessive number of constrains are generated due to the previous difficulties.

Contributions:

1. Jointly model objects, affordances, and activity planning for indoor scene configurations.
2. Provide a general learning and sampling framework for indoor scene modeling.
3. Demonstrate the effectiveness of this structured joint sampling by extensive comparative experiments.

2 Representation of Indoor Scenes

An attributed S-AOG combines:

1. A **probabilistic context free grammar (PCFG)**.

2. Contextual relations defined on an MRF, i.e. the horizontal links among the nodes.

An S-AOG is defined as a 5-tuple: $\mathcal{G} = \langle S, V, R, P, E \rangle$, where:

- S: the root node of the scene grammar
- V: the vertex set
  - $V = V_{NT} \cup V_T$
  - $V_{NT} = V^{And} \cup V^{Or} \cup V^{Set}$[1]
  - $V_T = V_T^r \cup V_T^a$
    1. Regular terminal node: $v \in V_T^r$, represents a spatial entity in a scene with **internal attributes** of object sizes (w, l, h), and **external attributes** A_{ext} of object position (x, y, z) and orientation ($x - y$ plane) θ and sampled human positions.
    2. Address terminal node: $v \in V_T^a$, encodes interactions that only occur in a certain context but a absent in all others.
- R: the production rules
- P: the probability model
- E: the contextual relations, $E = E_f \cup E_o \cup E_g \cup E_r$
  - E_f: relations among furniture
  - E_o: relations between supported objects and their supporting objects
  - E_g: relations in a functional pair
  - E_r: relations between furniture and the room

[1] V^{Set}: A set of Or-nodes serving as child branches are grouped by an And-node, and each child branch may include different numbers of objects.

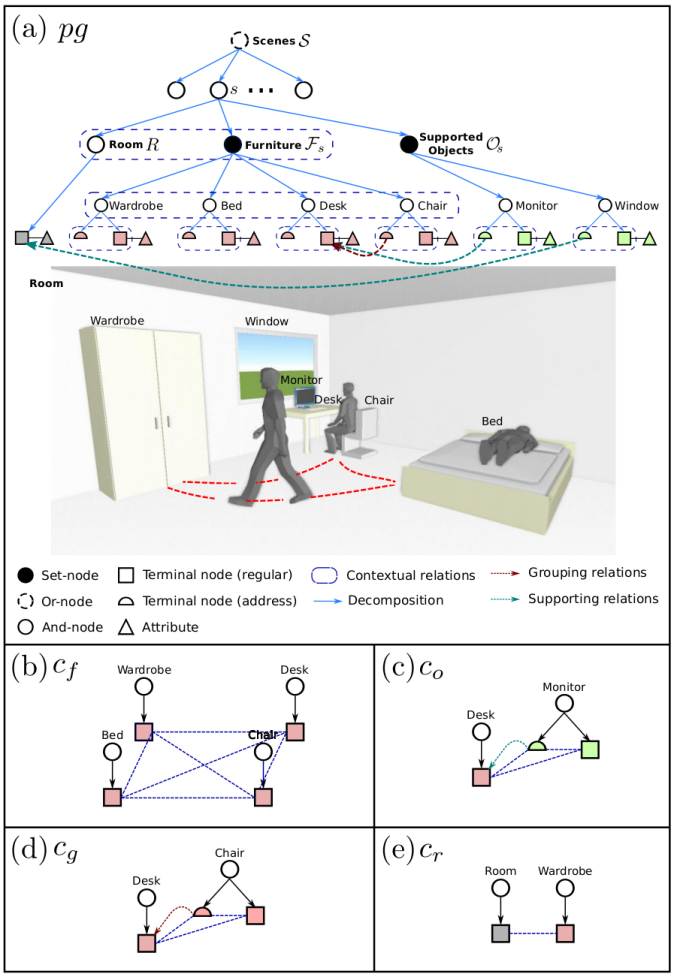


Figure 3: (a) A simplified example of a parse graph of a bedroom. The terminal nodes of the parse graph form an MRF in the terminal layer. Cliques are formed by the contextual relations projected to the terminal layer. Examples of the four types of cliques are shown in (b) — (e), representing four different types of contextual relations.

3 Probabilistic Formulation of S-AOG

The **prior probability** of pg generated by an S-AOG parameterized by Θ is formulated as a Gibbs distribution:

$$p(pg|\Theta) = \frac{1}{Z}\{-\mathcal{E}(pg|\Theta)\} \tag{1}$$

$$= \frac{1}{Z}\{-\mathcal{E}(pt|\Theta) - \mathcal{E}(E_{pt}|\Theta)\}, \tag{2}$$

where:

- $\mathcal{E}(pg|\Theta)$: the energy function of a pg
- $\mathcal{E}(pt|\Theta)$: the energy function of a pt
- $\mathcal{E}(E_{pt}|\Theta)$: the energy term of the contextual relations

$\mathcal{E}(pt|\Theta)$ can be decomposed into:

$$\mathcal{E}(pt|\Theta) = \underbrace{\sum_{v\in V^{Or}} \mathcal{E}_{\Theta}^{Or}(v) + \sum_{v\in V^{Set}} \mathcal{E}_{\Theta}^{Set}(v)}_{\text{non-terminal nodes}} + \underbrace{\sum_{v\in V_T^r} \mathcal{E}_{\Theta}^{A_{in}}(v)}_{\text{terminal nodes}}, \tag{3}$$

where the choice of the child node of an $v \in V^{Or}$ and the child branch of a $v \in V^{Set}$ follows **different multinomial distributions**. A_{in} of terminal nodes follows a non-parametric probability distribution learned by **kernel density estimation**.

$\mathcal{E}(E_{pt}|\Theta)$ combines the potentials of the 4 types of cliques, integrating human attributes and A_{ex} of V_T^r:

$$p(E_{pt}|\Theta) = \frac{1}{Z}\exp\{-\mathcal{E}(E_{pt}|\Theta)\} \tag{4}$$

$$= \prod_{c\in C_f} \phi_f(c) \prod_{c\in C_o} \phi_o(c) \prod_{c\in C_g} \phi_g(c) \prod_{c\in C_r} \phi_r(c). \tag{5}$$

Human Centric Potential Functions:

- $\phi_f(c)$[2]:

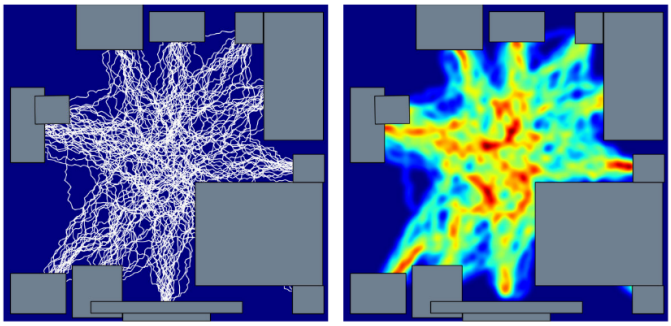
$$\phi_f(c) = \frac{1}{Z}\exp\{-\lambda_f \cdot \langle \sum_{f_i\neq f_j} l_{col}(f_i, f_j), l_{ent}(c)\rangle\}, \tag{6}$$

  where:

  - $c = \{f_i\} \in C_f$ includes all the V_T representing furniture.
  - λ_f is a weight vector.
  - $\langle\cdot,\cdot\rangle$ denotes a vector.

[2]The subscripts, f, o, g, and r, hold the similar meanings as those in the contextual relation notations.

- $l_{col}(f_i, f_j)$ (cost function) is the overlapping volume of the two pieces of furniture, serving as the **penalty of collision**.
- $l_{ent}(c) = -H(\Gamma) = \sum_i p(\gamma_i) \log p(\gamma_i)$ yields better utility of the room space by sampling human trajectories, where Γ is the set of planned trajectories in the room, and $H(\Gamma)$ is the **entropy**. The probability map is first obtained by planning a trajectory γ_i from the center of every piece of furniture to another one using **bi-directional rapidly-exploring random tree (RRT)**, which forms a heatmap. $H(\Gamma)$ is computed from the heatmap as shown in Figure 4.



(a) Planned trajectories (b) Probability map

Figure 4: Given a scene configuration, we use bi-directional RRT to plan from every piece of furniture to another, generating a human activity probability map.

- $\phi_o(c)$:

$$\phi_o(c) = \frac{1}{Z} \exp\{-\lambda_o \cdot \langle l_{num}(f, o), l_{add}(a) \rangle\}, \tag{7}$$

where:

- $c = \{f, a, o\} \in C_o$ includes a supported object $o \in V_T$, $a \in V_T^a$ connected to o, and a furniture $f \in V_T$ pointed by a.
- $l_{num}(f, o)$ defines the human usability cost — a favorable human position should enable agent to access or use both the furniture and the object. To compute, human position h_i^o are first sampled based on position, orientation, and the affordance map of o. Given an f:

$$l_{num}(f, o) = \max_i p(h_i^o | f). \tag{8}$$

- $l_{add}(a)$ is the **negative log probability** of a $v \in V_T^a$, treated as a certain $v \in V_T^r$, following a multinomial distribution.

- $\phi_g(c)$:

$$\phi_g(c) = \frac{1}{Z} \exp\{-\lambda_g \cdot \langle l_{num}(f_i, f_j), l_{add}(a) \rangle\}, \tag{9}$$

where $c = \{f_i, a, f_j\} \in C_g$ consists of terminal nodes of a core functional f_i, pointed by the a of an associated f_j.

Other Potential Functions:

- $\phi_r(c)$:

$$\phi_r(c) = \frac{1}{Z}\exp\{-\lambda_r \cdot \langle l_{dis}(f,r), l_{ori}(f,r)\rangle\}, \tag{10}$$

  where:

  - $c = \{f, r\} \in C_r$ includes a furniture $f \in V_T$ and a room r.
  - $l_{dis}(f,r) = -\log p(d|\Theta)$ is the distance cost function, in which $d \sim \ln\mathcal{N}(\mu, \sigma^2)$ is the distance between the f and the nearest wall.
  - $l_{ori}(f,r) = -\log p(\theta|\Theta)$, where $\theta \sim p(\mu,\kappa) = \frac{e^{\kappa\cos(x-\mu)}}{2\pi I_0(\kappa)}$ is the relative orientation between the model and the nearest wall modeled by a **von Mises distribution**.

4 Learning S-AOG

Use SUNCG dataset as training data and collect the statistics of:

- room types
- room sizes
- furniture occurrences
- furniture sizes
- relative distances
- orientations between furniture and walls
- furniture affordance
- grouping occurrences
- supporting relations

The parameters Θ of the probability model P can be learned in a **supervised** way by **maximum likelihood estimation (MLE)**.

Weights of Loss Functions: Recall

$$p(E_{pt}|\Theta) = \frac{1}{Z}\{-\mathcal{E}(E_{pt}|\Theta)\} \tag{11}$$

$$= \frac{1}{Z}\{-\langle\lambda, l(E_{pt})\rangle\}, \tag{12}$$

where:

- λ: weight vector.
- $l(E_{pt})$: the loss vector given by the 4 types of potential functions.

To learn the weight vector, the standard MLE maximizes the average log-likelihood:

$$\mathcal{L}(E_{pt}|\Theta) = -\frac{1}{N}\sum_{n=1}^{N}\langle\lambda, l(E_{pt_n})\rangle - \log Z. \tag{13}$$

which is usually maximized by following the gradient:

$$\frac{\partial\mathcal{L}(E_{pt}|\Theta)}{\partial\lambda} = -\frac{1}{N}\sum_{N=1}^{n} l(E_{pt_n}) - \frac{\partial \log Z}{\partial\lambda} \tag{14}$$

$$= -\frac{1}{N}\sum_{N=1}^{n} l(E_{pt_n}) - \frac{\partial \log \sum_{pt}\exp\{-\langle\lambda, l(E_{pt})\rangle\}}{\partial\lambda} \tag{15}$$

$$= -\frac{1}{N}\sum_{N=1}^{n} l(E_{pt_n}) + \sum_{pt}\frac{1}{Z}\exp\{-\langle\lambda, l(E_{pt})\rangle\} l(E_{pt}) \tag{16}$$

$$= -\frac{1}{N}\sum_{N=1}^{n} l(E_{pt_n}) + \frac{1}{\widetilde{N}}\sum_{\widetilde{n}=1}^{\widetilde{N}} l(E_{pt_{\widetilde{n}}}), \tag{17}$$

where $\{E_{pt_{\widetilde{n}}}\}_{\widetilde{n}=1,\ldots,\widetilde{N}}$ is the set of synthesized examples from the current model.

Original Texts

It is usually computationally infeasible to sample a Markov chain that burns into an *equilibrium distribution* at every iteration of gradient ascent. Hence, instead of waiting for the Markov chain to converge, we adopt the **contrastive divergence (CD)** learning that follows the gradient of difference of two divergences:

$$\mathrm{CD}_{\widetilde{N}} = \mathrm{KL}(p_0\|p_\infty) - \mathrm{KL}(p_{\widetilde{n}}\|p_\infty), \tag{18}$$

where $\mathrm{KL}(p_0\|p_\infty)$ is the **Kullback-Leiber** divergence between the data distribution p_0 and the model distribution p_∞, and $p_{\widetilde{n}}$ is the distribution obtained by a Markov chain started at the data distribution and run for a small number $\widetilde{n}$ of steps. In this paper, $\widetilde{n} = 1$.

The gradient of CD is given by:

$$\frac{\partial \mathrm{CD}_{\widetilde{N}}}{\partial\lambda} = \frac{1}{N}\sum_{n=1}^{N} l(E_{pt_n}) - \frac{1}{\widetilde{N}}\sum_{\widetilde{n}=1}^{\widetilde{N}} l(E_{pt_{\widetilde{n}}}) - \frac{\partial p_{\widetilde{n}}}{\partial\lambda}\frac{\partial \mathrm{KL}(p_{\widetilde{n}}\|p_\infty)}{\partial p_{\widetilde{n}}}, \tag{19}$$

where the third term can be ignored.

Finally, the weight vector is learned by gradient descent computed by generating a small number $\widetilde{N}$ of examples from the Markov chain:

$$\lambda_{t+1} = \lambda_t - \eta_t\frac{\partial \mathrm{CD}_{\widetilde{N}}}{\partial\lambda} \tag{20}$$

$$= \lambda_t + \eta_t\left(\frac{1}{\widetilde{N}}\sum_{\widetilde{n}=1}^{\widetilde{N}} l(E_{pt_{\widetilde{n}}}) - \frac{1}{N}\sum_{n=1}^{N} l(E_{pt_n})\right) \tag{21}$$

Branching Probabilities: The MLE of the branch probabilities ρ_i of V^{Or}, V^{Set}, and V^a_T is the frequency of each alternative choice: $\rho_i = \#(v \to u_i) / \sum_{n=1}^{n(v)} \#(v \to u_j)$.

Grouping Relations: The grouping relations are **hand-defined**. The probability of occurrence is learned as a multinomial distribution, and the supporting relations are **automatically** extracted from SUNCG.

Room Size and Object Sizes: Of which the distribution is learned as a non-parametric distribution. The size information is extracted from the 3D models inside SUNCG, and then a non-parametric distribution is fit using **kernel density estimation**. The distances and relative orientations of the furniture and objects to the nearest wall are computed and fitted into a **log normal** and a mixture of **von Mises distributions**, respectively.

Affordance: Affordance maps of all the furniture and supported objects are learned by computing the heatmap of possible human positions. These position include annotated humans, and we assume that the center of chairs, sofas, and beds are positions that humans often visit. By accumulating the relative positions, we get reasonable affordance maps as non-parametric distributions as shown in Figure 5.

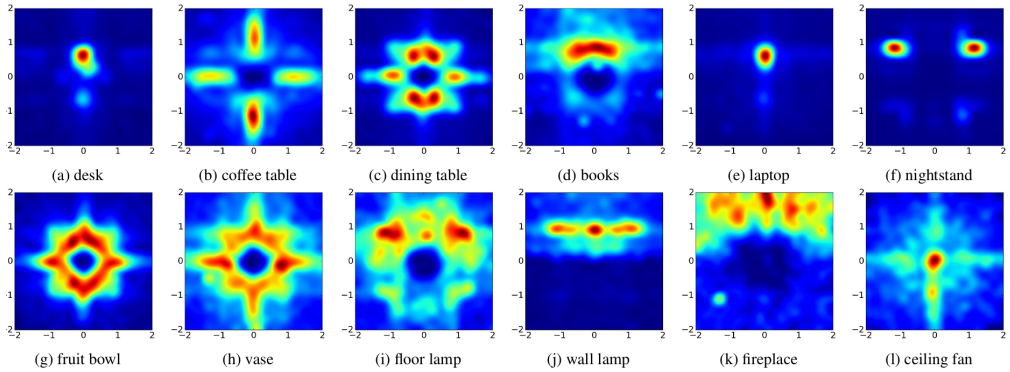


Figure 5: Examples of the learned affordance maps. Given the object positioned in the center facing upwards, i.e., coordinate of (0, 0) facing direction (0, 1), the maps show the distributions of human positions. The affordance maps accurately capture the subtle differences among desks, coffee tables, and dining tables. Some objects are orientation sensitive, e.g., books, laptops, and night stands, while some are orientation invariant, e.g., fruit bowls and vases.

5 Synthesizing Scene Configurations

Synthesizing scene configurations is accomplished by:

- Sampling a pg from the prior probability $p(pg|\Theta)$ defined by the S-AOG.
- The structure of a pt and A_{in} (sizes) can be sampled from the closed-form distributions or non-parametric distributions.
- For A_{ex} (positions and orientations), a **Markov Chain Monte Carlo (MCMC)** sampler is utilized to draw a typical state in the distribution:

1. Directly sample the structure of pt and A_{in}: (i) sample the child node for V^{Or}; (ii) determine the state of each child branch of V^{Set}; (iii) for each V_T^r, sample the sizes and human positions from learned distributions.
2. Use an MCMC scheme to sample the values of V^a and A_{ex} by making proposal moves. A sample will be chosen after the Markov chain converges.

Two types of Markov chain dynamics $q_i, i = 1, 2$ are designed and used at random with probabilities to make proposal moves:

- q_1: **translation** of objects. It chooses a $v \in V_T^r$, and samples a new position based on the current position $x \to x + \delta x$, where δx follows a **bivariate normal distribution**.
- q_2: **rotation** of objects. It chooses a $v \in V_T^r$, and samples a new orientation based on the current orientation $\theta \to \theta + \delta\theta$, where $\delta\theta$ follows a **normal distribution**.

Adopting the **Metropolis-Hastings** algorithm, the proposed new $pg\prime$ is accepted according to the acceptance probability:

$$\alpha(pg\prime|pg, \Theta) = \min(1, \frac{p(pg\prime|\Theta)p(pg|pg\prime)}{p(pg|\Theta)p(pg\prime|pg)}) \tag{22}$$

$$= \min(1, \exp(\mathcal{E}(pg|\Theta) - \mathcal{E}(pg\prime|\Theta))), \tag{23}$$

Original Texts

where the proposal probability rate is canceled since the proposal moves are symmetric in probability. A **simulated annealing** scheme is adopted to obtain samples with high probability as shown in Figure 6.

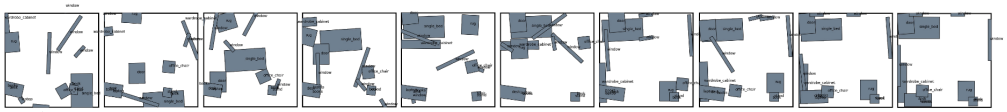


Figure 6: MCMC sampling process (from left to right) of scene configurations with simulated annealing

6 Experiments

(Omitted.)

7 Conclusion

Original Texts

We propose a novel general framework for human-centric indoor scene synthesis by sampling from a spatial And-Or graph. The experimental results demonstrate the effectiveness of our approach over a large variety of scenes based on different criteria. In the future, to synthesize physically plausible scenes, a physics engine should be integrated. We hope the synthesized data can contribute to the broad AI community