Notes of “Human-centric Indoor Scene Synthesis Using Stochastic Grammar”

Jihang Li

October 12, 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.

[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.

- 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

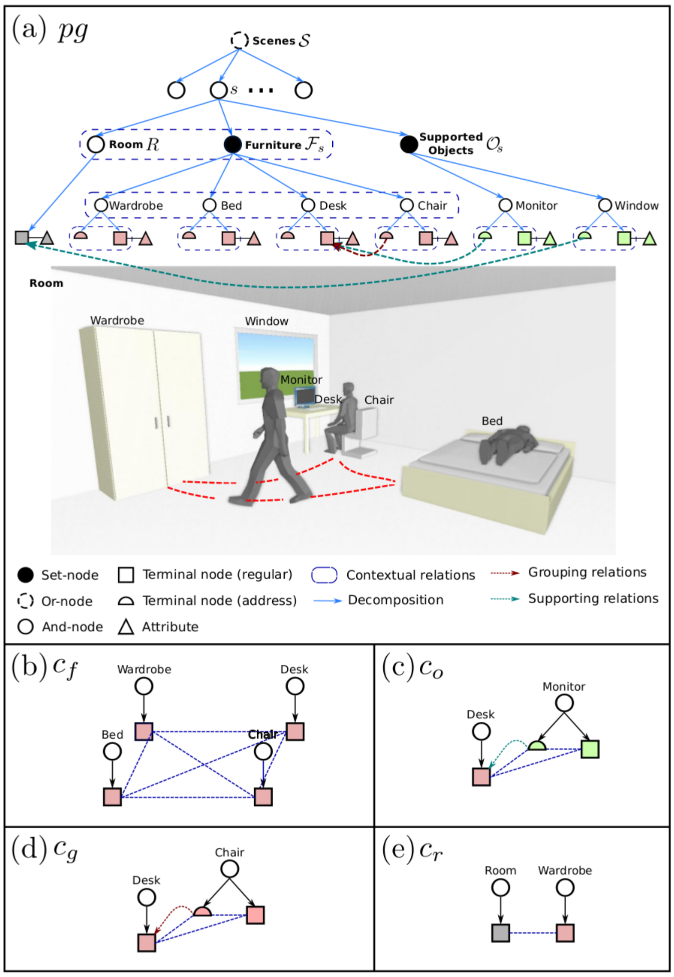

Figure 3: (a) A simplified example of a parse graph of a bedroom. The terminal nodes of the parse graph form an MRF in the terminal layer. Cliques are formed by the contextual relations projected to the terminal layer. Examples of the four types of cliques are shown in (b) — (e), representing four different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

3 Probabilistic Formulation of S-AOG

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

$$\begin{aligned} p(pg|\Theta) &= \frac{1}{Z}\{-\mathcal{E}(pg|\Theta)\} && (1) \\ &= \frac{1}{Z}\{-\mathcal{E}(pt|\Theta) - \mathcal{E}(E_{pt}|\Theta)\}, && (2) \end{aligned}$$

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:

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

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