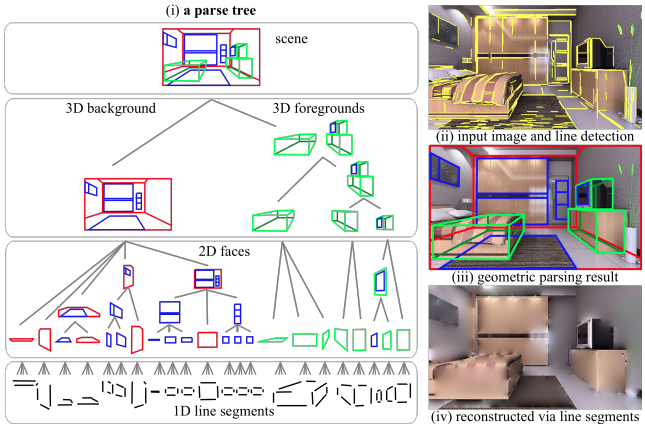
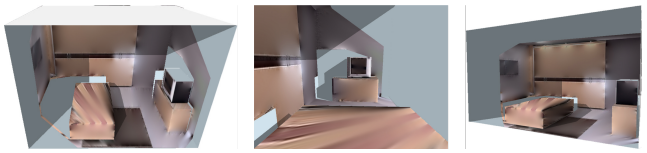
Image Parsing via Stochastic Scene Grammar

Yibiao Zhao and Song-Chun Zhu
University of California, Los Angeles

Introduction

This paper proposes a parsing algorithm for indoor scene understanding which includes four aspects: computing 3D scene layout, detecting 3D objects (e.g. furniture), detecting 2D faces (windows, doors etc.), and segmenting the background. The algorithm parse an image into a hierarchical structure, namely a parse tree. With the parse tree, we reconstruct the original image by the appearance of line segments, and we further recover the 3D scene by the geometry of 3D background and foreground objects.





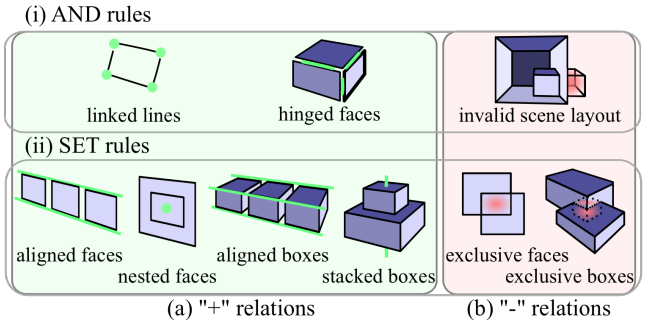
3D synthesis of novel views based on the parse tree

Results

Stochastic Scene Grammar

The grammar represents compositional structures of visual entities, which includes three types of production rules and two types of contextual relations:

- **Production rules:** (i) AND rules represent the decomposition of an entity into sub-parts; (ii) SET rules represent an ensemble of visual entities; (iii) OR rules represent the switching among sub-types of an entity.
- **Contextual relations:** (a) Cooperative "+" relations represent positive links between binding entities, such as hinged faces of a object or aligned boxes; (b) Competitive "-" relations represents negative links between competing entities, such as mutually exclusive boxes.



Bayesian Formulation

We define a posterior distribution for a solution (a parse tree) *pt* conditioned on an image I. This distribution is specified in terms of the statistics defined over the derivation of production rules.

$$P(pt|I) \propto P(pt)P(I|pt) = P(S) \prod_{v \in V^N} P(Ch_v|v) \prod_{v \in V^T} P(I|v) \quad (1)$$

The probability is defined on the Gibbs distribution: $P(pt|I) = \frac{1}{Z}\exp\{-E(pt|I)\}$ and the energy term is decomposed as three potentials:

$$E(pt|I) = \sum_{v \in V^{OR}} E^{OR}(A_T(Ch_v)) + \sum_{v \in V^{AND}} E^{AND}(A_G(Ch_v)) + \sum_{\Lambda_v \in \Lambda_I, v \in V^T} E^T(I(\Lambda_v)) \quad (2)$$

Inference by Hierarchical Cluster Sampling

We design an efficient MCMC inference algorithm, namely Hierarchical cluster sampling, to search in the large solution space of scene configurations. The algorithm has two stages:

- **Clustering**: It forms all possible higher-level structures (clusters) from lower-level entities by production rules and contextual relations.

$$P_+(Cl|I) = \prod_{v \in Cl^{OR}} P^{OR}(A_T(v)) \prod_{u,v \in Cl^{AND}} P_+^{AND}(A_G(u), A_G(v)) \prod_{v \in Cl^T} P^T(I(\Lambda_v)). \quad (3)$$

- **Sampling**: It jumps between alternative structures (clusters) in each layer of the hierarchy to find the most probable configuration (represented by a parse tree).

$$Q(pt^*|pt, I) = P_+(Cl^*|I) \prod_{u \in Cl^{AND}, v \in pt^{AND}} P_-^{AND}(A_G(u)|A_G(v)). \quad (4)$$

Experiment and Conclusion

Segmentation precision compared with Hoiem et al. 2007 [1], Hedau et al. 2009 [2], Wang et al. 2010 [3] and Lee et al. 2010 [4] in the UIUC dataset [2].

Segmentation precision	[1]	[2]	[3]	[4]	Our method
Without rules	73.5%	78.8%	79.9%	81.4%	80.5%
With 3D "-" constraints	-	-	-	83.8%	84.4%
With AND, OR rules	-	-	-	-	85.1%
With AND, OR, SET rules	-	-	-	-	85.5%

Compared with other algorithms, our contributions are

- A Stochastic Scene Grammar (SSG) to represent the hierarchical structure of visual entities;
- A Hierarchical Cluster Sampling algorithm to perform fast inference in the SSG model;
- Richer structures obtained by exploring richer contextual relations.

Website: http://www.stat.ucla.edu/~ybzhao/research/sceneparsing