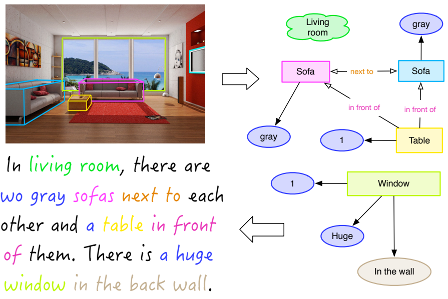
Generating Multi-sentence Natural Language Descriptions of Indoor Scenes

Dahua Lin[1]
dhlin@ie.cuhk.edu.hk
Sanja Fidler[2]
fidler@cs.toronto.edu
Chen Kong[3]
chenk@cs.cmu.edu
Raquel Urtasun[2]
urtasun@cs.toronto.edu

[1] Department of Information Engineering,
The Chinese University of Hong Kong.
[2] Department of Computer Science,
University of Toronto.
[3] Robotics Institute,
Carnegie Mellon University.



In living room, there are two gray sofas next to each other and a table in front of them. There is a huge window in the back wall.

Figure 1: Our method visually parses an RGB-D image to get a *scene graph* that represents objects, their attributes and relations between objects. Based on the scene graph we then generate a multi-sentence textual description via a learned grammar. The description generation takes into account co-reference and saliency of how people describe scenes.

This paper proposes a novel framework for generating lingual descriptions of indoor scenes. This is an important problem, as an effective solution to this problem can enable many exciting real-world applications, such as human robot interaction, image/video synopsis, and automatic caption generation. Whereas substantial efforts have been made to tackle this problem, previous approaches focusing primarily on generating a single sentence for each image, which is not sufficient for describing complex scenes. We attempt to go beyond this, by generating coherent descriptions with multiple sentences.

Particularly, we are interested in generating multi-sentence descriptions of cluttered indoor scenes. Complex, multi-sentence output requires us to deal with challenging problems such as consistent co-referrals to visual entities across sentences. Furthermore, the sequence of sentences needs to be as natural as possible, mimicking how humans describe the scene. This is especially important for example in the context of social robotics to enable realistic communications.

Towards this goal, we develop a framework with three major components: (1) a *holistic visual parser* based on [3] that couples the inference of objects, attributes, and relations to produce a semantic representation of a 3D scene (Fig. 1); (2) a *generative grammar* automatically learned from training text; and (3) a *text generation algorithm* that takes into account subtle *dependencies across sentences*, such as logical order, diversity, saliency of objects, and co-reference resolution.

From RGB-D Images to Semantics. Given an RGB-D image, we extract semantics, such as objects of interest, their attributes, and their physical relations, via visual parsing, and thereon construct a *scene graph*. The detailed procedure is as follows. First, a set of *"objectness" regions* are generated following [1], which are encouraged to respect intensity as well as occlusion boundaries in 3D. These regions are projected to 3D via depth and then cuboids are fit tightly around them, under the constraint that they are parallel to the ground floor.

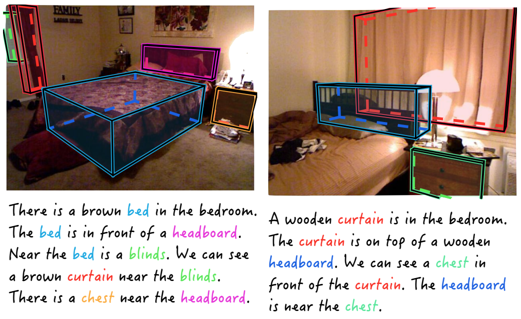
A *holistic CRF model* is then constructed to jointly reason about the cuboid classes as well as the scene class (*e.g.* kitchen, bathroom). The model exploits various geometric and semantic relations, including *scene appearance*, *cuboid appearance*, *object geometry*, *co-occurrence relations*, and *spatial relations*. These features and relations are incorporated into the CRF formulation as *potentials*. The CRF weights to combine the potentials are learned with a primal dual learning framework [2], and inference of class labels is done with an approximated algorithm [4].

Based on the extracted visual information, we construct a *scene graph*, with *nodes* representing objects and their attributes, and *edges* representing relations between nodes. There are three kinds of edges: *attribute edges* that link objects to their attributes, *position edges* that represent the positions of objects relative to the scene, (*e.g. corner-of-room*), and *pairwise edges* that characterize the relative positions between objects (*e.g. on-top-of* and *next-to*).

Generating Lingual Descriptions. Given a *scene graph*, we generate a descriptive paragraph in two steps. First, we transform the scene graph into a sequence of *semantic trees*. Then, we produce sentences, one from each semantic tree, following a *generative grammar*. A *semantic tree* contains a set of *terminal nodes* corresponding to individual entities or their attributes and *relational nodes* that express relations among them. Such a tree can capture information like *what entities are being described* and *what are the relationships between them*. To generate sentences in a coherent manner, we further devise a method that transforms a *scene graph* into a sequence of *semantic trees*, with multiple kinds of dependencies among sentences taken into account, including *logical order*, *diversity*, *saliency*, *co-reference*, and *richness vs. conciseness*.

For each *semantic tree*, we produce a sentence via a *generative grammar*, *i.e.* a map from each semantic relation to a set of templates (derivation rules). The *grammar* for generating sentences are often specified manually in previous work. This way, however, is time consuming, unreliable, and tends to oversimplify the language. Here, we explore a new approach, that is, to learn the grammar from data. The basic idea is to construct a semantic tree from each sentence through linguistic parsing, and then derive the templates by matching nodes of the semantic tree to parts of the sentence.

Experimental Evaluation. We tested the proposed framework on the NYU-v2 dataset [5] augmented with an additional set of textual descriptions, one for each image. Below are two representative examples. Please refer to the paper for more details about the experiments.



There is a brown bed in the bedroom. The bed is in front of a headboard. Near the bed is a blinds. We can see a brown curtain near the blinds. There is a chest near the headboard.

A wooden curtain is in the bedroom. The curtain is on top of a wooden headboard. We can see a chest in front of the curtain. The headboard is near the chest.

Figure 2: Examples of descriptions generated using our framework.

[1] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *TPAMI*, 2012.

[2] T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, 2010.

[3] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, 2013.

[4] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In *CVPR*, 2011.

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