SUN3D: A Database of Big Spaces Reconstructed using SfM and Object Labels

Jianxiong Xiao
Princeton University

Andrew Owens
MIT

Antonio Torralba
MIT

Abstract

Existing scene understanding datasets contain only a limited set of views of a place, and they lack representations of complete 3D spaces. In this paper, we introduce SUN3D, a large-scale RGB-D video database with camera pose and object labels, capturing the full 3D extent of many places. The tasks that go into constructing such a dataset are difficult in isolation – hand-labeling videos is painstaking, and structure from motion (SfM) is unreliable for large spaces. But if we combine them together, we make the dataset construction task much easier. First, we introduce an intuitive labeling tool that uses a partial reconstruction to propagate labels from one frame to another. Then we use the object labels to fix errors in the reconstruction. For this, we introduce a generalization of bundle adjustment that incorporates object-to-object correspondences. This algorithm works by constraining points for the same object from different frames to lie inside a fixed-size bounding box, parameterized by its rotation and translation. The SUN3D database, the source code for the generalized bundle adjustment, and the web-based 3D annotation tool are all available at http://sun3d.cs.princeton.edu.

The popularity of the Microsoft Kinect and other depth-capturing devices has led to a renewed interest in 3D for recognition. Researchers have extended traditional object and scene recognition datasets to incorporate 3D. For example, UW RGB-D object dataset is an evolution of popular 2D object datasets such as Caltech 101 to 3D objects captured by an RGB-D camera. The NYU Depth dataset and others go beyond objects by capturing RGB-D videos of scenes and labeling the objects within. However, these 3D datasets inherit many of the limitations of traditional 2D datasets: they contain a sample of views from the world, but the physical relationship *between* these views and the structure of the space containing them is mostly missing.

What we desire is a dataset that is *place-centric* rather than view-based, containing full 3D models of spaces (*e.g.* entire apartments) instead of a limited set of views (Fig. 1). Such a database would allow us to ask questions like: "what does this object look like from behind?" or "what can we expect the space to look like beyond the available field of view?" Such a database would be useful for learning complete 3D context models to be used for scene parsing (*e.g.* learning that there is always a bed in a bedroom); for obtaining an integrated understanding of a space instead of individual disconnected snapshots; and for reasoning about intuitive physics, functionality and human activity.

2D view-based SUN database 3D place-centric SUN3D database

Figure 1. **View-based vs. place-centric.** This example shows the difference between a view-based scene representation and a place-centric scene representation. SUN database contains a view of a living room. SUN3D database contains an RGB-D video for the whole apartment and 3D models with camera poses.

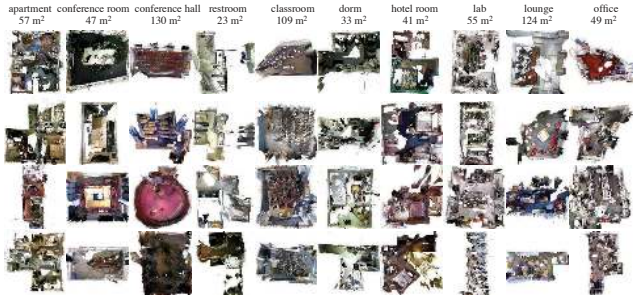

Figure 2. **SUN3D database.** Each column contains examples from a place category. The numbers are the median coverage areas.

With this goal in mind, we introduce *SUN3D*, a place-centric database (see Figure 2). The items in our database are full 3D models with semantics: RGB-D images, camera poses, object segmentations, and point clouds registered into a global coordinate frame.

This database requires camera poses, but estimating them reliably for large space from an RGB-D video is a difficult problem. And despite recent progress in RGB-

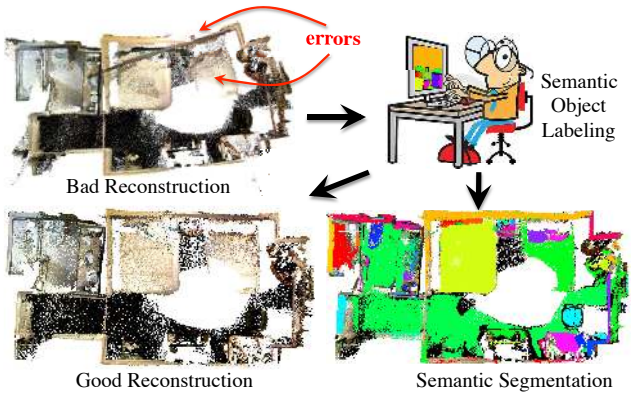

Figure 3. **Main idea.** Semantic object labeling as a way to correct pose estimation errors.

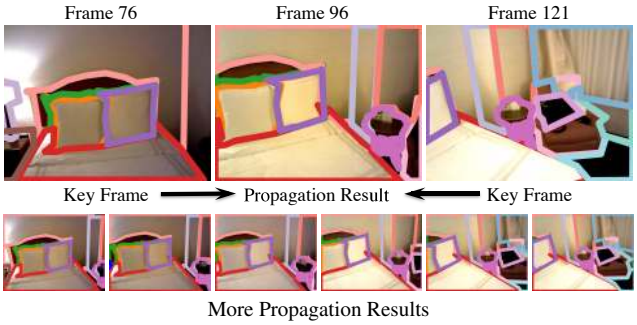

Figure 4. **3D label propagation.** Annotation of each frame is automatically populated from nearby key frames.

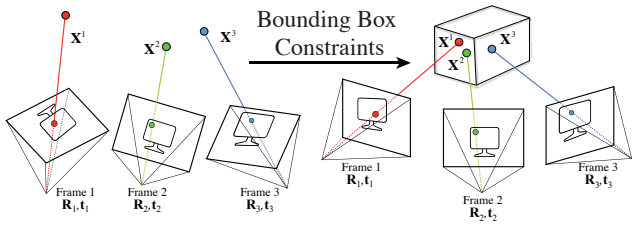

Figure 5. **Generalized bundle adjustment.** The object-to-object correspondence constraint essentially pulls a set of points belonging to the same object so that they fit into one 3D bounding box for that object. Together with constraints from other objects and key-points, the camera pose can be estimated more reliably.

D structure-from-motion (SfM), existing automatic reconstruction methods are not reliable enough for our purposes. Additionally, we desire a semantic segmentation, but labeling every frame in a full video is a painstaking task – for this reason, existing RGB-D video databases (e.g. NYU Depth) have semantic annotations only for a sparse keyframes.

To address this, we design our 3D reconstruction and object labeling tasks so that they mutually support one another (see Figure 3). Our approach is based on the idea that if the 3D reconstruction were perfect, then object labeling would be easy – one would merely need to label an object in one frame, and the reconstruction could be used to propagate these annotations to the rest of the images. On the other hand, if objects were annotated in every frame, then reconstruction would improve dramatically since consistencies between frames could be used as constraints in optimization. By combining the two tasks, we (a) produce better 3D reconstructions, and (b) provide an object annotation tool that makes it easy to label long RGB-D videos.

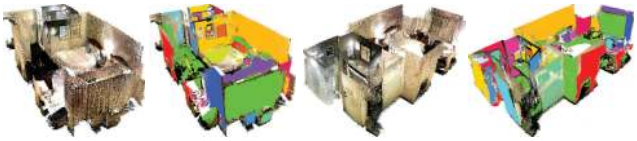

Figure 6. **Annotation and reconstruction correction result.** The point cloud is colored based on semantic object categories.

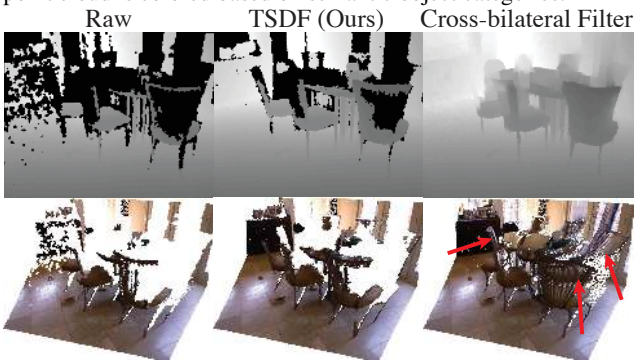

Figure 7. **Comparision of depth improvement algorithms.** The cross-bilateral filtering introduces large artifacts, *e.g.* it smooths over occlusion boundaries, as the 3D point cloud shows.

To produce better reconstructions, we incorporate object labels into our structure-from-motion algorithm and solve jointly for object locations and camera poses (see Figure 5). The resulting algorithm is based on standard bundle adjustment, and the addition of object labels helps to avoid errors due to drift and loop-closing failures, establishing "long-range" connections between frames that may be very far apart in a video but that nevertheless contain the same object instances.

Additionally, we use the 3D reconstruction to help with object annotation, creating a tool that speeds up the process of labeling a long video. A user labels an object in one frame, and the partially completed reconstruction is used to propagate object labels to other frames (see Figure 4).

Figure 6 shows an example of the corrected reconstruction and semantic segmentation of the 3D point cloud. With the camera pose, we also propose a way to improve the raw depth map from multiple frames. The raw depth maps are usually noisy, with many holes. To fill in the holes, cross-bilateral filtering is typically used to produce a visually pleasing depth map, but it introduces many artifacts (Figure 7). Instead, we improve the depth map using a Truncated Signed Distance Function (TSDF) to voxelize the space, accumulating the depth map from nearby frames (*e.g.* 40 closest frames) using the camera poses obtained above. Finally, we use ray casting to get a reliable depth map.