1. Introduction

Significant advances have been made towards mapping the exteriors of urban environments through the large-scale acquisition efforts of Google, Microsoft, Nokia, etc. Capturing 3D indoor environments, however, remains challenging. Unlike mapping of exterior spaces, which focuses on flat surfaces of building facades, interior mapping would focus on interior objects, which are geometrically complex, can be located in cluttered setting and undergo significant variations: doors and windows are opened and closed, chairs are moved around, cubicles are rearranged, etc.

The growing popularity of fast, easy-to-use, affordable range cameras (e.g., the Microsoft Kinect) presents new acquisition possibilities. High frame-rate and increased portability of these cameras, however, come at the cost of resolution and data quality: the scans are at best of modest resolution, often noisy, invariably contain outliers, and suffer from missing parts due to occlusion (see Figure 2).

The work described in this abstract focused on reconstructing and understanding objects in interiors of public buildings (e.g., schools, hospitals, hotels, restaurants, airports, train stations) or office buildings from Kinect scans of such interiors. We exploited three observations to make the problem of indoor 3D acquisition tractable: (i) most such building interiors are composed of basic elements such as walls, doors, windows, furniture (e.g., chairs, tables, lamps, computers, cabinets), which come from a small number of prototypes and repeat many times. (ii) such building components usually consist of rigid parts of simple geometry, i.e., they have surfaces that are well approximated by planar, cylindrical, conical, and spherical proxies. Further, although variability and articulation are dominant (e.g., a chair is moved or rotated, a lamp arm is bent or adjusted), such variability is limited and low-dimensional (e.g., translational motion, hinge joint, and telescopic joint). (iii) mutual relationships among the basic objects satisfy strong priors (e.g., a chair stands on the floor, a monitor rests on the table).

We present a simple yet practical system to acquire models of indoor objects such as furniture, together with their variability modes, and discover object repetitions and exploit them to speed up large-scale indoor acquisition towards high-level scene understanding.

More detailed explanation and results are available in the original submission of the abstract [2].

2. Algorithm

Our framework works in two main phases: a learning phase and a recognition phase (see Figure 1).

First, in a learning phase we start from a few scans of individual objects to construct primitive-based 3D models while explicitly recovering respective joint attributes and modes of variation. Second, in a fast recognition phase, we start from a single-view scan to segment and classify it into plausible objects, recognize them, and extract the pose parameters for the low complexity models generated in the learning phase. Intuitively, we use priors for primitive types and their connections, thus greatly reducing the number of unknowns to enable model fitting even from very sparse and low resolution datasets, while hierarchically solving for part association. We also demonstrate that simple inter- and intra-object relations simplify segmentation and classification tasks necessary for high-level scene understanding (see [3] and references therein).
3. Results and Discussion

The real test of our system is on scanned data since it is difficult to synthetically recreate all the artifacts encountered during scanning. We tested our framework on a range of real-world examples each consisting of multiple objects arranged over large spaces (e.g., office area, and conference rooms). For both the learning and the recognition phases, we acquired the scenes using a Microsoft Kinect scanner with an open source scanning library [1].

Our recognition phase is lightweight and fast taking on an average 200ms to compare a point cluster to a model on a 2.4Hz CPU with 6GB RAM. We summarize the results in Figure 2 for (cluttered) office setups, auditoriums, and seminar rooms. We overlay the unresolved points on the recognized parts for comparison. Our algorithm has access to only the geometry, but no color or texture attributes. The complexity of our problem setting can be appreciated by looking at the input scan, which is hard even to parse visually. We detect the chairs, monitors, whiteboards, and trash bins across different rooms, and the rows of auditorium chairs in different configurations.

The segmentation of real-world scenes are challenging with naturally cluttered set-ups. In office #2, we also miss a couple of the chairs that are mostly occluded and beyond what our framework can handle. The challenge is better demonstrated in the seminar rooms because of closely spaced chairs or chairs leaning against the wall. The quality of data also deteriorates because of thin metal legs with specular highlights. Still we correctly recognized most of the chairs along with correct configurations by first detecting the larger parts and refining the segmentation based on the learned geometry (in 3-4 iterations).

While in our tests the recognition results were mostly satisfactory, we observed two main failure cases. First, we fail to detect objects when large amounts of data are missing. We missed some of the monitors because the material property of the screens were probably not favorable to Kinect capture. The missed monitors in office #2 have big rectangular holes within the screen in the scans. Second, we cannot overcome the limitations of our initial segmentation. While in certain cases the algorithm can recover segmentations with the help of other visible parts, this becomes difficult as we allow objects to deform and hence have variable extent.

Generally, we reliably recognized scans with 1000-3000 points per scan since in the learning phase we extracted the important degrees of variation, thus providing a compact, yet powerful, model (and deformation) abstraction.

References

