Scan and 3D Model Alignments in Indoor Scenes for Augmented Reality

Ruizhongtai (Charles) Qi
rqi@stanford.edu

1. Motivation

Recently, we have witnessed quite a lot of industrial interests and products in augmented reality. Head mounted devices (HMD) like Microsoft HoloLens are expected to have huge impacts in entertainment, design and medical treatment. One use scenario is as shown in 1, where a user can see an artificial scene of a Mincraft world seamlessly embedded in a normal living room.

While localization and reconstruction techniques mature, to achieve the effects shown in the Minecraft game we still need to fill the semantic gap. For example, computer needs to understand the 3D layout of the room such as floors, support planes and walls in order to place the virtual objects. Beyond that, if we can recognize what categories (sofa, table, bookshelf etc.) the real objects are, the application can benefit from it and support deeper immersion of virtual objects/characters into the scene. Ideally we can have virtual character navigating in the room and sit on the sofa beside us! One way to achieve such goal is to have a "semantic" reconstruction of the scene, i.e. we align scan data (depth images or point clouds) with 3D models, which can be either from our shape database or from models reconstructed from previous scans.

Figure 1. Snapshot from Microsoft HoloLens advertisement on a potential applicaiton of AR Minecraft game.

2. Related Works

3D indoor scene understanding. Scene understanding has been a hot topic in computer vision for more then ten years. Recently there are works that focus on semantic reconstruction from a single image [1]. The input to the system is a RGB or RGB-D image of an indoor scene, where the depth information is from commercial low-cost scanners like Kinect. The outputs are alignments of 3D models to the objects in the image. However, in their method, the 3D models only match the objects in the sense of category, not in style or shape (a square table will be matched to a round table). We think the methods proposed in [3] can mitigate this problem.

Scan and model alignment. 3D point cloud registration has been a classic problem in computer graphics and computer vision - given two scans of the same object, how do we match them? or given a scan and the geometry of the model how do we match the mesh and the scan? The ICP (Iterative Closest Points) algorithm and its extensions [5] are usually used. However, ICP requires that the two scans or the scans and models are of the same identity, which is unknown in the scene understanding case. Morever, ICP may rely on a good initialization to work. In general for AR application in 3D scenes, we need to first have object detection, model retrieval and pose estimation modules before we are able to do fine-grained alignments like ICP.

Augmented reality. Nowadays, most augmented reality applications either has no interaction with the geometry of the world (just displaying images in the air) or rely on a pre-configured environment (like a flat large talbe). This project is to combine recent advances in computer vision and explore how much we can achieve in terms of semantic understanding of indoor environment, which can be extremely helpful for AR.

3. Project Overview

In this project, we will try to understand a 3D scene (represented as multiple RGB-D images or fused point clouds from commercial RGB-D scanners) by aligning 3D models with objects in the scene. By achieving this goal, we can equip augmented reality application with the understanding of the room geometry and object semantics knowledge so that more advanced interactions become possible. For example, if the computer knows where the sofa is and where the table is, it's possible to render an avatar sitting on the sofa and pick up a teapot from the coffee table.

To achieve this goal, we would rather not develop computer vision algorithms from scratch. Instead we will use state-of-the-art object detection algorithms and 3D model retrieval algorithms and combine them in a good way so that we can semantically reconstruct the 3D scene. For object recognition, current best system is operating on 2D images but there also exists systems on RGB-D images or point clouds. To achieve object localizationa and recognition in 3D, one applicable approach is to do multipe detections in a series of RGB frames of the same room and stitch the prediction results together. However, how to effectively combine the 2D prediction results in 3D is not trivial and requires exploration. After localizing the object in 3D scene, we can retrieve 3D CAD models that are similar to the object and then estimate coarse pose of the object before a final CAD model and scan data alignment.

A sample input is like the fused point clouds shown in 2. Illustrative input-output of the system is shown in 3 (note that this figure is from [2] and is jsut temporarily put here). With 3D models aligned in the scene, we can not only support AR games with avatars or moving objects but also support design applicaitons like furniture product browsing and matching.

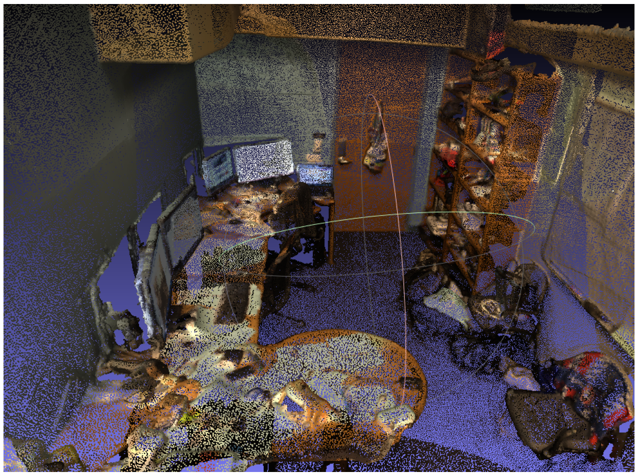

Figure 2. A 3D point scan of an office in Gates building at Stanford.

4. Milestones

Stage 1: Collecting and cleaning datasets. Feb 8 - Feb 15. We will collect testing/evaluation datasets for this project while object detection models are from author's previous projects or public available code repository. The evaluation dataset is ideally a series of RGB-D frames of an indoor living room or office.

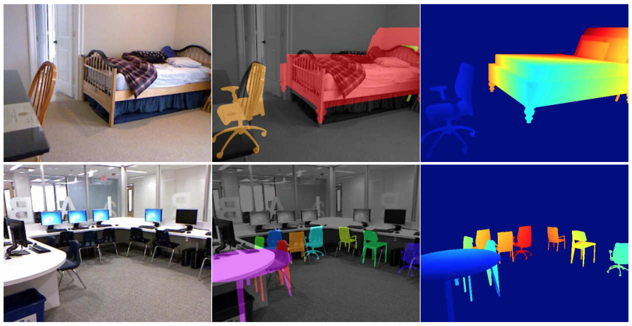

Figure 3. Illustration of sysetm input and output from [1]. **Input:** single RGB-D image of an indoor scene. **Output:** a 3D scene where each major object is replaced by a similar 3D model.

Stage 2: Verifying existing methods on the collected dataset. Feb 16 - Feb 23 Existing methods include object detection algorithm (faster-RCNN [4]), object pose estimation system [6] and 3D model retrieval system [3].

Stage 3: Align 3D models to scans. Feb 24 - Mar 8. Now we combine all the outputs from existing algorithms and apply local adjustment and optimization methods to align the 3D model to the scans and possibly show some simple AR applicaiton based on our alignments.

References

[1] S. Gupta, P. A. Arbeláez, R. B. Girshick, and J. Malik. Aligning 3D models to RGB-D images of cluttered scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 2

[3] Y. Li, H. Su, C. R. Qi, N. Fish, D. Cohen-Or, and L. J. Guibas. Joint embeddings of shapes and images via cnn image purification. *ACM Trans. Graph.*, 2015. 1, 2

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[5] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152. IEEE, 2001. 1

[6] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2