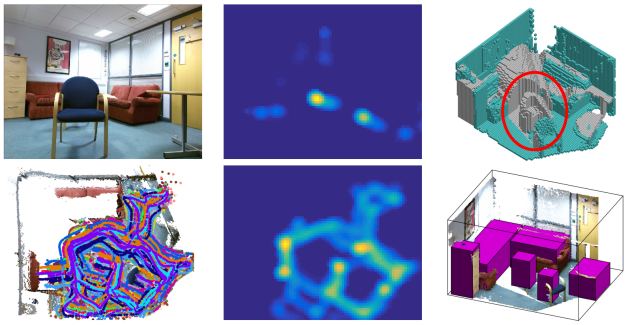

(a) (b) (c)

What: 3D scene reconstruction from a single-view RGB-D video and human motions.

Why: This is the first method that use human motions to reconstruct fine scene geometry including occluded areas.

Reconstructing 3D scene geometry from a single RGB-D image is challenging because of occlusions. If a scene contains humans moving in it, however, we can estimate occluded regions by assuming that area where people standing is empty. Though there are some previous scene reconstruction methods using human motions, the proposed method generates more accurate geometry by combining depth and human motion to a volumetric occupancy map. The occupancy map is also used to estimate occluded areas. We also demonstrate object (large furniture) detection of fully occluded objects based on the occupancy map.

How: Initialize occupancy map from depth data while assuming all occluded areas are occupied, and carve voxels where humans moved.

From the depth capture, we firstly initialize a volumetric occupancy map, which is a 3D matrix that incorporates a whole scene. Each voxel has a binary value that represents the position is occupied by scene objects or not. Because the depth of occluded areas is unknown, we assume that the occluded areas are all filled with voxels, and carve the voxels if humans walk through the position. Thus, the geometry of occluded areas is estimated from human motions.

3D body poses and positions were tracked with an off-the-shelf method, and classified into standing or sitting. The area where humans stood or sat was projected onto the floor, and obtained as heat maps (figure 1b). According to whether people stood or sat, voxels in the area with corresponding height were carved.

The object (large furniture) detection was performed by finding cuboids in the occupancy map. For that, we compute surface normals in the occupancy map, cluster surfaces with the same direction, and fit planes to surfaces. Only planes that align world axes are considered. Planes that intersected to form a cuboid were detected as an object.

The detected cuboids are further selected by minimizing a cost function (Eq. 2). The cost function comprises three terms. 1) The unary potential (U) gives a penalty when an object is too big or an object is overlapping with the area where people move, and encourages the contact between scene and sitting areas. 2) Overlap term (O) penalizes overlapping of two objects. 3) Coverage term encourages to select as many as cuboids to avoid not detecting any objects.

The three terms above are weighted manually. The minimization problem was solved with integer programming by branch and bound.

Results: Outperforms past methods in occupancy map estimation and object detection. The effectiveness of human motion is validated through ablation studies.

Evaluated on NYU dataset, which contains RGB-D images and true geometries of rooms. We created ground-truth occupancy maps by gridding the true geometries. Each grid size is 0.05m. Because the NYU dataset does not contain humans moving in it, we manually annotated maps where humans can walk or sit. Only the occluded area was used for evaluation. For comparison, we use two learning-based methods that are trained to estimate occluded areas from training examples. Tab. 1 shows the precision, recall, IoU between the estimated and ground-truth occupancy maps. The proposed method outperforms the past methods. The IoU drops when voxel carving with human motions was not applied (Base-NC-completion).

We also created an in-house dataset to test object detection performance. We captured humans exploring in a room for 1-2 minutes with a single RGB-D camera. We manually annotated bounding boxes of large furniture objects. An object is assumed to be detected when the IOU between predicted and true cuboids is over 0.3. Tab. 2 shows the results. The proposed method shows better precision and recall compared to the results from only depth images or non-carved occupancy maps.

Figure 3 shows qualitative results, and we can see that the use of human motions is effective especially for estimating occluded regions.

Thoughts: This method tends to output false-positive results. The scene dynamics and semantics are not considered. Extend this method to consider patterns of human behavior may be interesting.

Because the proposed method initializes occluded areas as filled with voxels and carve the voxels after that, false-positive voxels tend to appear. This method is not applicable when the scene changes over time. This method only considers standing and sitting, but we can learn more from human motions (*e.g.* if someone falls, there will be a small thing below the person).