Exploiting 2D Floorplan for Building-scale Panorama RGBD Alignment Supplementary Material

Erik Wijmans Yasutaka Furukawa
Washington University in St. Louis
{erikwijmans, furukawa}@wustl.edu

The supplementary material provides algorithmic and implementation details of the data preparation process (Sect. 3 in the main paper) and the scan-to-scan consistency potential (Sect. 4.2 in the main paper).

1. Data preparation

Clutter removal: We remove clutter in a floorplan image by discarding small connected components of black pixels. In particular, we find connected components of black pixels (i.e., intensity below 0.6), compute their average size (i.e., number of pixels), and discard a component if its size is less than the average.

Scale ruler: In each floorplan image, we measure the distance of a ruler in pixels and calculate the conversion ratio. See Fig. 1 for examples of scale rulers.

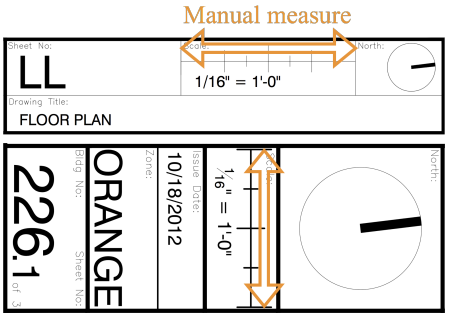

Figure 1: We manually measure the size of a scale ruler in pixels to calculate the pixel to a metric scale ratio.

Manhattan axis extraction: First, we subsample 3D points in each scan to a uniform density of 85mm, and estimate a surface normal at each point by using a search radius of 3cm. We use Point Cloud Library [1] for these steps. Dominant directions are extracted by RANSAC on surface normal estimates, followed by the linear least squares on the inlier-set. A tolerance of 1.15 degrees ($\approx \arcsin(0.02)$) determine the inliers. Inlier ratios are often small, and we extract the first axis, then the second and the third axes in two steps.

The scanner is placed near vertically and the extracted Manhattan axis closest to the default Z-axis is identified as the gravity direction. The floor (resp. ceiling) height is estimated by looking at the point density along the gravity direction and picking the highest peak below (resp. above) the scanner center.

Point evidence and free-space evidence: For each scan, both the point evidence and free-space evidence images are defined in the axis-aligned bounding of the 3D points, where the pixel size is set equal to that of the floorplan image. The 3D points are first sub-sampled to have a uniform density, then the point evidence is equal to the number of 3D points projected inside, while we normalize by linearly mapping the range of counts $[\mu + \sigma, \mu + 4\sigma]$ to the range $[0, 1.0]$. μ and σ denote the mean and the standard deviation of non-zero pixels in the point evidence image. Similarly, the free-space evidence is the number of times the ray (between a 3D point and the scanner center) passes through. We normalize the range of $[\mu - \sigma, \mu]$ to $[0, 1]$.

In scan-to-scan consistency potential, we also need to compute the point and the free-space evidence over the voxel-grid in 3D. The scale of the voxel gird is set such that there are 20^3 voxels per cubic meter (in other words, a voxel is a cube with length $5\ cm$). For both the 3D point evidence and 3D free space evidence we binarize the voxel grids by mapping everything greater than $\mu - \sigma$ to 1 and everything else to 0.

Door detection in a floorplan: In a floorplan image, we manually specify a bounding box containing a door symbol as a template then perform a standard sliding window template matching: 1) computing squared differences from the template, 2) smoothing the scores with a 2D Gaussian ($\sigma = 2$ pixels), and 3) extracting the peaks after the non-local max suppression. Note that we allow the template to be mirrored and/or rotated by a multiple of 90 degrees, where the squared difference is calculated as the minimum over all the augmented templates.

Door detection in a 3D scan: For every pixel in the depth image, we hypothesize that the pixel is at the bottom cor-

ner of a door and perform the following procedure to verify the hypothesis. Note that there are four possible door directions (positive and negative two horizontal Manhattan directions), and the process is repeated for four times. First, one side of the door border must be a wall while the other side should see through the door-way. Therefore, starting from the corresponding 3D point, we trace the depth image vertically-up and continue while the difference of the depth value at the right and the left of the pixel (with a margin of 5 pixels) is more than 0.5 meter. If this process stops below 2 meter, we reject this hypothesis as the door is too low. We also reject a hypothesis if the bottom part of the door is too high from the floor, in particular, when the height difference is more than 10% of the door height. We estimate the width of a door by tracing the top of the door horizontally, until the same depth difference test fails.

2. Scan-to-scan consistency potential

2.1. Photometric cue

The photometric cue calculates the Normalized Cross Correlation scores between feature points in the pair of panorama images. We first detect Harris corner features in each panorama image. Given the 2D placements and the vertical translation adjusted by the floor height estimates, we can reproject each feature point into another panorama image by using the depth information. However, 3D scans are often far apart, and most features have significant viewing angle differences. In practice, we use a feature if 1) the viewing angle difference is less than 20 degrees; 2) the difference in the estimated surface normals is less than 20 degrees; 3) it passes the visibility test during reprojection with a margin of 30 cm; and 4) the difference between the depth value and the average of its 8 neighbors is less than 30 cm (i.e., preferring a planar surface). We use feature points in both panorama images that pass the above test, and calculate the Normalized Cross Correlation (NCC) score over 11×11 pixel colors, while properly adjusting the scales of pixel sampling. One minus the average NCC score over all the features, followed by the normalization process in Sect. 2.3 is the photometric score.

2.2. Geometric cue

The geometric cue measures the consistency of 3D point evidence and 3D free-space evidence between a pair of 3D scans. We first binarize the 3D point and 3D free-space evidence with a threshold $\mu - \sigma$, computed based on each voxel image. The geometric cue is the amount of point agreement minus the amount of point disagreement. The agreement is measured by picking each 3D point, projecting to the other voxel grid, and checking if the point 3D evidence is 1 (i.e., contains points). The number of such 3D points is the amount of the agreement. Similarly, the amount of disagreement is the number of 3D points that project into voxels where the free-space evidence is 1. We divide the difference of the agreement and the disagreement by the total number of points in the two scans to map the score to the range of $[0, 1.0]$.

2.3. Normalization

In order to handle scans that cover a diverse range of rooms containing varying amounts of clutter, we perform the following normalization to compute the final scan-to-scan potential. Let E_p and E_g denote the photometric and the geometric score, respectively. We lower- and upper-bound each score by $(0.1, 0.6)$ for E_p and $(0.3, 0.7)$ for E_g. We then affinely map the score of $[\mu - \sigma, \mu + 2.5\sigma]$ to $[0, 1]$, where μ and σ are the mean and the standard deviation of the scores. With abuse of notation, let E_p and E_g be the normalized scores, the scan-to-scan consistency potential is defined as

$$\left(E_p + 0.5E_g\right) / max(0.25, min(\delta, 1.5)).$$

δ is the distance between the two scanner centers, and the denominator avoids placing scans at nearly the same location (a form of anti stacking bias). $1.5(m)$ is considered to be the minimum expected distance between adjacent scans.

References

[1] Point cloud library. http://pointclouds.org. 1