Exploiting 2D Floorplan for Building-scale Panorama RGBD Alignment

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

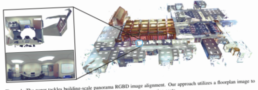


Figure 1: The paper tackles building-scale panorama RGBD image alignment. Our approach utilizes a floorplan image to significantly reduce the number of necessary scans and hence human operating costs.

Abstract

This paper presents a novel algorithm that utilizes a 2D floorplan to align panorama RGBD scans. While effective panorama RGBD alignment techniques exist, such a system requires extremely dense RGBD image sampling. Our approach can significantly reduce the number of necessary scans with the aid of a floorplan image. We formulate a novel Markov Random Field inference problem as a scan placement over the floorplan, as opposed to the conventional scan-to-scan alignment. The technical contributions lie in multi-modal image correspondence cues (between scans and schematic floorplan) as well as a novel coverage potential avoiding an inherent stacking bias. The proposed approach has been evaluated on five challenging large indoor spaces. To the best of our knowledge, we present the first effective system that utilizes a 2D floorplan image for building-scale 3D pointcloud alignment. The source code and the data will be shared with the community to further enhance indoor mapping research.

1. Introduction

3D scanning hardware has made remarkable progress in recent years, where successful products exist in industry for commercial applications. In particular, Panorama RGBD scanners have found real-world applications as the system produces both 3D geometry and immersive panorama images. For instance, Faro 3D [9] is a professional grade panorama RGBD scanner, which can reach more than 100 meters and produce 100 million points per scan within a millimeter accuracy. The device is perfect for 3D measurement, documentation, or surveillance in indoor mapping, civil engineering or GIS applications. Matterport [2, 4] is an emerging low-end solution that can reach only 5 meters, but is much quicker (i.e., 1 to 2 minutes per scan), and has demonstrated compelling results for Real Estate markets.

Given the success on the 3D scanning hardware, automated panorama RGBD alignment has become a crucial technology. The Matterport system provides a robust solution but requires extremely dense scanning (e.g., one scan every 2 to 3 meters). Dense scanning becomes infeasible for high-end scanners (e.g., Faro 3D [9]), whose single scan could take thirty minutes or an hour depending on the resolution. However, these scanners are the only option for large buildings such as department stores, airport terminals, or hotel lobbies, simply due to the required operating ranges (e.g., 20 to 30 meters). Furthermore, the precision of these high-end scanners is necessary for quantitative recovery of metric information for scientific and engineering data analysis. In practice, with high-end 3D scanning devices, people use calibration objects such as big bright balls and/or utilize

der to acquire building-scale panorama RGBD scans over five :an location. Only one scan has been acquired in each room in d and the use of floorplan-image essential for our problem.

2. Related work

Two approaches exist for indoor 3D scanning: "RGBD streaming" or "Panorama RGBD scanning". RGBD streaming continuously moves a depth camera and scans a scene. This has been the major choice among Computer Vision researchers [18, 7, 23] after the success of Kinect Fusion [13]. The input is a RGBD video stream, where Simultaneous Localization and Mapping (SLAM) is the core technology. Panorama RGBD scanning has been rather successful in industry, because 1) data acquisition is easy (i.e., picking a 2D position as opposed to 6 DoF navigation in RGBD streaming); 2) alignment is easier thanks to the panoramic field of views; and 3) the system produces panorama images, essential for many visualization applications. Structure from Motion (SfM) is the core technology in this approach. This paper provides an automated solution for Panorama RGBD alignment, and the remainder of the section focuses on the description of the SfM techniques, where we refer the reader to a survey article [6] for the SLAM literature.

Structure from Motion (SfM) addresses the problem of automatic image alignment [11]. State-of-the-art SfM system can handle even millions of unorganized Internet pho-

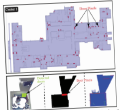


Figure 3: [Top] A floorplan image for Center 1 after the clutter removal. The blue overlay shows the building mask, and red pixels show the detected door pixels. [Bottom] The left shows the close-up of one scan with the result of 3D door detections. The middle shows the free-space image mask with detected door pixels. The right shows the point-evidence image with a heat-map color scheme

plan pixel becomes inside the mask if the pixel is between the left and the right most pixels in its row and between the top and the bottom most pixels in its column.

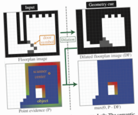
- We detect doors both in a floorplan image and 3D scans. For a floorplan, we manually specify a bounding box containing a door symbol, and use a standard template matching. For 3D scans, we use a heuristic to identify door-specific 3D patterns directly in 3D points.

4. MRF formulation

The multi-modal nature of the problem makes our formulation fundamentally different from existing ones [21, 22]. The first critical difference lies in the definition of the variables. In existing approaches, a variable encodes a *3D relative placement* between a pair of scans [21, 22]. In our formulation, a variable encodes a *2D absolute placement* of a single scan over a floorplan image.

Let $\mathcal{S} = \{s_1, s_2, \cdots\}$ be our variables, where s_i encodes the 2D placement of a single scan. s_i consists of two components: 1) rotation, which takes one of the four angle values by exploiting the Manhattan frames (0, 90, 180, or 270 degrees); and 2) translation, which is a pixel coordinate in the floorplan image. We seek to find the placements $\mathcal{S}$ that minimizes the following three potentials:

$$\sum_{s_i \in \mathcal{S}} E_S(s_i) + \sum_{(s_i, s_j) \in \mathcal{S} \times \mathcal{S}} E_{S \times S}(s_i, s_j) + \sum_{f_k \in \mathcal{F}} E_P^k(\mathcal{S}).$$



y of the floorplan and a 3D scan in two ways. Left: The semantic detected pixels in the point evidence. Right: The geometric cue ined by the dilated floorplan image.

1.0. The average penalty over all the door-pixels in the evidence image is the semantic penalty.

Geometric cue: Measuring the consistency between the floorplan image and the point evidence image is a real challenge: 1) A floorplan image contains extra symbols that are not in evidence images; 2) An evidence image contains objects/clutter that are not in a floorplan image; 3) The style of a floorplan (e.g., line thickness) may vary; and 4) Both are essentially line-drawings, making the comparison sensitive to small errors. In practice, we have found that the following consistency potential provides a robust metric.

We first apply a standard morphological dilation operation (as a gray-scale image) to a floorplan image, using the OpenCV default implementation with a 5×5 kernel. We then measure how much of the point evidence image is NOT explained by the floorplan, by 1) subtracting the dilated floorplan image (DF) from the point evidence image (P); then 2) clamping the negative intensities to 0. The sum of intensities in this residual image $(max(0, P - DF))$ divided by the sum of intensities in the original evidence image (P) calculates the amount of the discrepancy. We swap the role of the floorplan and a point evidence image, compute the other discrepancy measure, and take the average.

4.2. Scan-to-scan consistency potential

Different from standard MRF formulation, we do not know which pairs of variables (i.e., scans) should have interactions, because our variables encode the placements of the scans. Therefore, we set up a potential for every pair of scans. The potential measures the photometric and geomet-

Data	SF	SF+SS	All (SF+SS+F)
Center 1	12%	0%	2%
Center 2	3%	2%	0%
Hall 1	29%	15%	15%
Hall 2	12%	15%	7%
Hall 3	44%	47%	34%

Table 2: To assess the contributions of each potential, we have run our algorithm with different combinations of the potentials. SF, SS and F denotes the scan-to-floorplan, scan-to-scan, and floorplan coverage potentials, respectively. For example, the left column shows the error rate when only the unary (SF) potential is used.

best five placements at the bottom level of these searches are reported as the candidates.

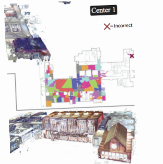
While this search strategy is relatively straightforward, a few algorithmic details are worth noting. First, we use the maximum (minimum for a floorplan) intensity over 2×2 pixels instead of the averaging in the image pyramid creation, as images are near binary. Second, the non-local min suppression looks at a much larger area than the direct neighbors, as the function tends to be peaky. Let W_B and H_B be the width and height of the tight bounding box containing the floorplan mask. We look at a square region whose size is $(W_B + H_B)/80$. Third, the threshold at the non-local min suppression is the mean minus the standard deviation of the evaluated scores at the same pyramid level. Fourth, we speeded up the unary potential evaluation by skipping scan placements when more than 30% of the corresponding free-space mask goes outside the building-mask. Lastly, 7×7 children pixels (every other pixel is chosen at the perimeter for speed) instead of 2×2 are searched under each local minimum for more robustness.

With five placement candidates per scan, we resort to the tree-reweighted message passing algorithm [16] to optimize our non-submodular energy. Each variable is initialized as the placement with the best unary potential. The optimization usually converges after 50 iterations.

6. Experimental results and discussions

We have used C++ for implementation and Intel Core I7 CPU with 16GB RAM PC. Three computational expensive steps are preprocessing, unary-potential evaluation, and TRW optimization, where the running time is roughly proportional to the number of the input scans. For large datasets with 70 to 80 scans, these steps roughly take 5 hours, 2.5 hours, and 30 minutes, respectively. The preprocessing is the bottleneck due to I/O and processing of the massive scan files, which can be parallelized if necessary.

Figures 5 and 6 show our main results. For each dataset, we first show the merged point cloud. The visualization of the scan placements are given by colored floorplan-masks.

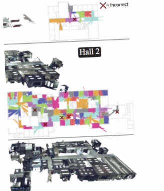


ged 3D point-cloud, and 2D colored free-space masks are shown.

	Distance transform			Ours (unary)		
Final	Top 1	Top 5	Final	Top 1	Top 5	Final
12%	14%	4%	12%	6%	2%	2%
10%	20%	12%	13%	1%	0%	0%
15%	29%	29%	29%	15%	0%	15%
37%	12%	8%	16%	22%	10%	7%
61%	45%	22%	37%	45%	32%	34%

age distance metrics that replace the scan-to-floorplan (unary) ıd-truth is in the top 1 or 5 placements based solely on the unary RF optimization with the replaced unary potential. Naive SSD orplan image and the point-evidence image. SSD computes the Distance transform makes the floorplan image into binary with a n takes the sum of element-wise product with the point-evidence

rithm has successfully aligned most of the scans. We have not scanned the right wing of the building in Center 1 and Center 2 (See Fig. 2), which makes a large space for scans to be misplaced. Nonetheless, our method has only one misplacement in that area (Center 1). Note that Hall 3 is an exception in which we make many errors due to the glitch



s for Hall1, Hall2, and Hall3.

istance transform on Center 2. Major placement errors occur in ows the magnitude of the error.

alignment accuracy for easier datasets (Center 1, Center 2, and Hall1), but not for the harder two cases. The floorplan coverage potential is crucial for challenging datasets (Hall2 and Hall3), which are full of repetitions and ambiguity.

Due to the large label space, our approach has required the unary potential to limit the feasible label-set. Nonetheless, we have taken the limited label-space and run our algorithm without the unary term, whose results are much

have experimented with three alternative image matching metrics to replace the unary term (See Table 3). The same hierarchical search scheme (Sect. 5) has been used. The naive SSD without the mask fails badly as expected. SSD and Distance transform utilizing our masks has achieved reasonable accuracy, which is remarkable, considering the fact that the pairwise scan alignment without floorplan (i.e., current state-of-the-art) has completely failed in all the examples. Columns "Top 1" and "Top 5" indicate if the ground-truth placement is in the top 1 or 5 candidates based solely on the unary potential. It is worth noting that expanding the candidate list did not help in reducing the error rate for Top 5, because failure cases are usually extreme.

Figure 7 illustrates typical failure modes of Distance transform, which tends to concentrate scans in large rooms. Our analysis is that a large room tends to have non-architectural lines or symbols to fill-in an empty canvas, which makes the distance transform image contain small values, and allow lower-energy placements.

Figure 8 illustrates the effects of the semantic cue in the unary potential (i.e., the door detection). Indoor scenes are full of symmetries and repetitions, which makes the comparison of pure geometry (i.e., geometric cue) susceptible to local minima. The figure demonstrates a representative case, where the door detections break such an ambiguity.

When the placement is ambiguous even with the geometric and the semantic cues, we rely on the MRF optimization with the full three potentials. Figure 9 compares the final scan placements with or without the floorplan coverage potential. The floorplan coverage potential seeks to avoid "stacking" and evenly distribute the placements.

Our method is not perfect and has exposed several failure modes. First, our approach tends to make mistakes for small storage-style rooms, where a small room with a lot of clutter makes the geometric cue very noisy. Second, there are genuinely ambiguous cases where the scene geometry, appearance, and door locations are exactly the same. Lastly, our method has made major errors in Hall 3, simply because the floorplan has not reflected renovations in the past. Unfortunately, it was difficult to identify erroneous scans based on the potentials. At the presence of problematic scans, the MRF optimization seems to shuffle around scan placements including correct ones to achieve a low energy state. Nonetheless, the total potential, in particular, the magnitude of the total potential divided by the number of scans is a good indicator of success. The quantity for Hall 3 is a few times larger than the others and indicates that "something is wrong". Our main future work is to develop a robust algorithm to detect potentially erroneous scan placements, which will allow a quick user feedback to correct mistakes. We will share our source code and high-end building-scale datasets to further enhance indoor mapping research.

[18] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (TOG)*, 32(6):169, 2013. 2

[19] A. Pagani and D. Stricker. Structure from motion using full spherical panoramic cameras. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 375–382. IEEE, 2011. 3

[20] O. Shakernia, R. Vidal, and S. Sastry. Structure from small baseline motion with central panoramic cameras. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 7, pages 83–83. IEEE, 2003. 3

[21] P. Theiler, J. Wegner, and K. Schindler. Fast registration of laser scans with 4-point congruent sets-what works and what doesn't. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(3):149, 2014. 3, 8

[22] F. Yan, L. Nan, and P. Wonka. Block assembly for global registration of building scans. *ACM Transactions on Graphics (TOG)*, 35(6):237, 2016. 3, 4

[23] Q.-Y. Zhou, S. Miller, and V. Koltun. Elastic fragments for dense scene reconstruction. In *2013 IEEE International Conference on Computer Vision*, pages 473–480. IEEE, 2013. 2