Exploiting 2D Floorplan for Building-scale Panorama RGBD Alignment

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

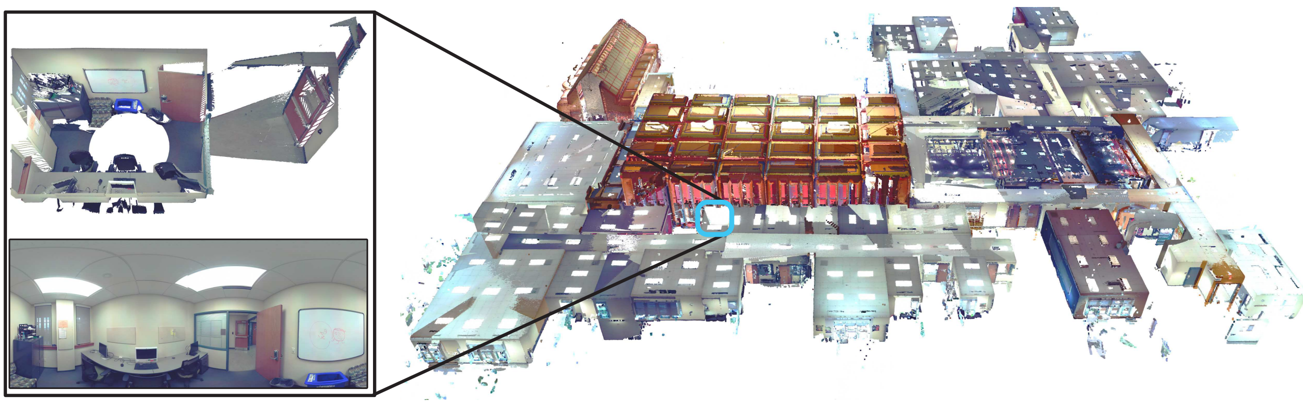

Figure 1: The paper tackles building-scale panorama RGBD image alignment. Our approach utilizes a floorplan image to significantly reduce the number of necessary scans and hence human operating costs.

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 are 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 [10] 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, 5] 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 [10]), 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

Figure 2: We have used a professional grade laser range finder to acquire building-scale panorama RGBD scans over five floors in two buildings. An orange circle indicates a rough scan location. Only one scan has been acquired in each room in most cases, making the intersection of adjacent scans minimal and the use of floorplan-image essential for our problem.

semi-automatic 3D alignment tool such as Autodesk ReCap 360 [1] to minimize the number of necessary scans.

This paper focuses on high-end 3D indoor scanning (See Fig. 1). A key observation is that building targets for high-end 3D scanning often have 2D floorplans. Our approach can significantly reduce the number of necessary RGBD scans with the aid of a 2D floorplan image. The key technical contribution lies in a novel Markov Random Field formulation as a scan placement problem as opposed to the conventional scan-to-scan alignment. Besides the standard visual and geometric feature matching between scans, we incorporate multi-modal geometric or semantic correspondence cues associating scans and a floorplan, as well as a novel "coverage potential" that avoids an inherent *stacking bias*. We have experimented with five challenging large indoor spaces and demonstrated near perfect alignment results, significantly outperforming existing approaches (See Fig. 2). While our work has focused on existing indoor structures, the technology can also be transformative to other domains, for example, Civil Engineering applications at construction sites for progress monitoring or safety inspection [15], where precise building blueprints exist.

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 [19, 8, 24] after the success of Kinect Fusion [14]. 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 [7] for the SLAM literature.

Structure from Motion (SfM) addresses the problem of automatic image alignment [12]. State-of-the-art SfM system can handle even millions of unorganized Internet pho-

Name	Area (ft^2)	Scans	Rooms	3D-points	Floorplan res.
Center1	$111,627$	50	13	2200×10^6	4808×7692
Center2	$111,627$	80	29	3520×10^6	6845×7477
Hall1	$8,896$	7	3	308×10^6	4400×3600
Hall2	$19,461$	75	42	3300×10^6	9336×4168
Hall3	$20,580$	65	40	2860×10^6	8256×4552

Table 1: Statistics of our building-scale panorama RGBD datasets. A single scan contains 44 million colored points.

tographs [4, 11]. The wider field-of-view (e.g., panorama images) further makes the alignment robust [21, 20, 16] as more features are visible. When depth data is available, geometry provides additional cues for alignment, where Iterative Closest Point (ICP) [6] has been one of the most successful methods. However, even state-of-the-art SfM or ICP systems face real challenges for indoor scenes that are full of textureless walls with limited visibility through narrow doorways. Existing approaches either 1) take extremely dense samples [2] or 2) rely on manual alignment [18, 13].

The idea of utilizing building information for scan alignment has been demonstrated in Civil Engineering applications [15]. However, the system requires a full 3D model for construction design and planning, as well as manual image-to-model correspondences to start the process. This paper utilizes 2D floorplan to align undersampled building-scale panorama RGBD scans with minimal user interactions (i.e., a few mouse strokes on a floorplan image).

3. Building-scale Panorama RGBD Datasets

We have used a professional grade laser range finder, Faro 3D [10], to acquire panorama RGBD scans over five floors in two buildings (Fig. 2 and Table 1). A floorplan is given as a rasterized image. This section summarizes standard preprocessing steps, necessary to prepare our datasets for our algorithm. We here briefly describe these steps and refer the details to the supplementary material (See Fig. 3).

• We remove clutter in a floorplan image by discarding small connected components of black pixels.

• Our floorplan image contains a scale ruler, which lets us calculate a metric scale per pixel with a few mouse strokes. The process may be imprecise, and our algorithm will be designed to tolerate errors. In the absence of a ruler, we can align one scan with a floorplan by hand to obtain a scale.

• We extract a Manhattan frame from each scan and a floorplan image, respectively. In each scan, we identify the vertical direction and the floor height based on the point density.

• We turn each scan into *point* or *free-space* evidence images in a top-down view. A point evidence is a score in the range of $[0, 1]$, while a free-space evidence is a binary mask.

• We compute a building mask over a floorplan image, which can quickly prune unlikely scan placements. A floorplan 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.

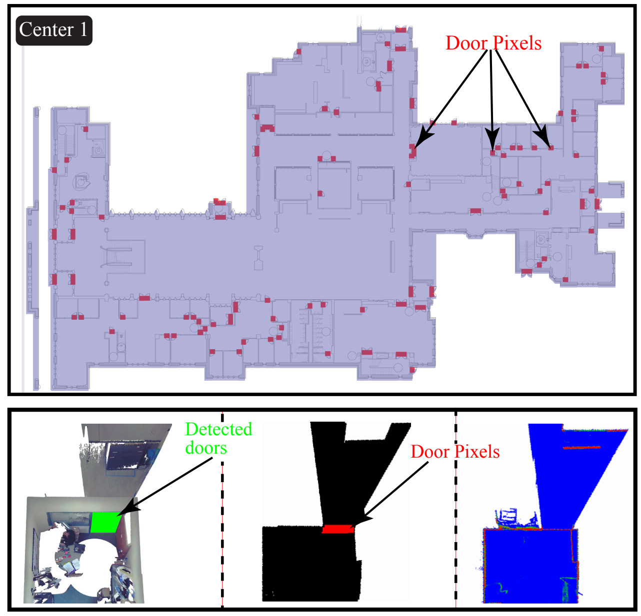

Figure 3: [Top] A floorplan image for Center 1 after the clutter removal. The blue overlay shows the building mask, and red pixels show the detected door pixels. [Bottom] The left shows the close-up of one scan with the result of 3D door detections. The middle shows the free-space image mask with detected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

• We detect doors both in a floorplan image and 3D scans. For a floorplan, we manually specify a bounding box containing a single door symbol, and use a standard template matching method to find the remaining symbols. 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 [22, 23]. 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 [22, 23]. 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

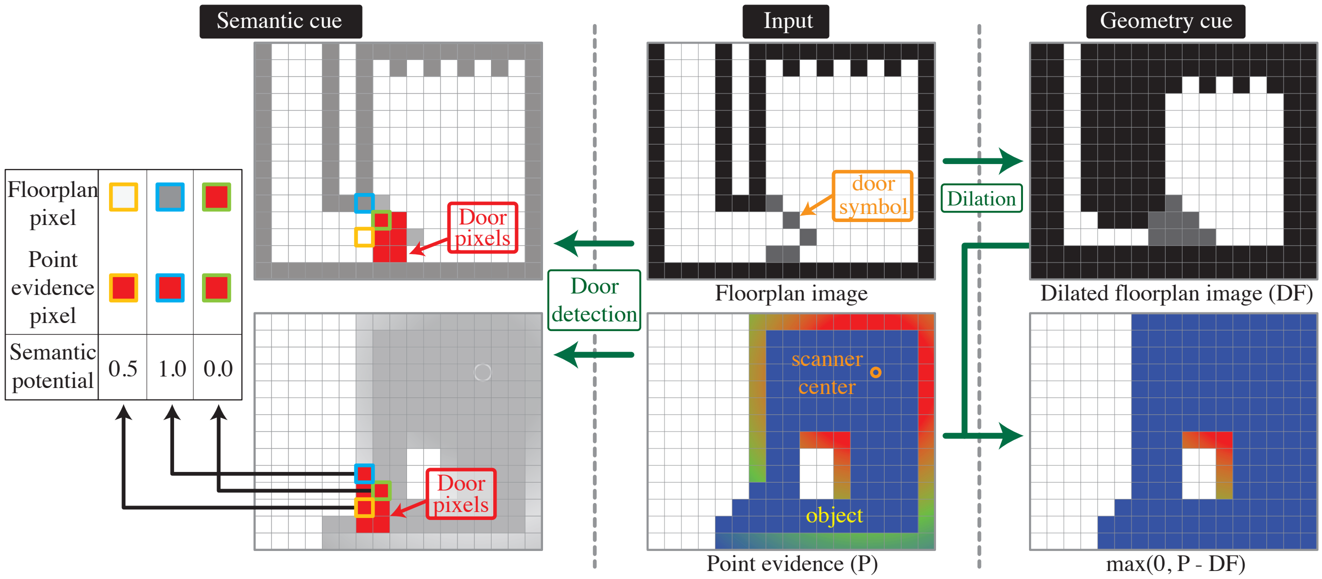

Figure 4: The scan-to-floor potential measures the consistency of the floorplan and a 3D scan in two ways. Left: The semantic cue checks the consistency of door detections over the door-detected pixels in the point evidence. Right: The geometric cue measures how much of the point evidence (P) is NOT explained by the dilated floorplan image.

the floorplan image. We seek to find the placements $\mathcal{S}$ that minimizes the following three potentials:

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

The first term (E_S) is a unary potential, measuring the consistency between the scan placement and the floorplan image. The second term ($E_{S \times S}$) is a binary potential, measuring the consistency between pairs of scan placements. The third term (E_F) counts the number of floorplan pixels covered by the scan placements, and is summed over floorplan image pixels $\mathcal{F}$. E_F^k is a higher order term, but will be approximated by a sum of pairwise terms as explained below. The first (E_S) and the third (E_F^k) terms are the main contributions, while the second term ($E_{S \times S}$) resembles existing approaches.

4.1. Scan-to-floorplan consistency potential

Scan-to-floorplan consistency (E_S) needs to overcome vastly different modalities between a floorplan image and real sensor data. Our measure is the sum of the semantic and geometric penalties (See Fig. 4).

Semantic cue: The semantic cues exploit door detections. The door detection are used to align pairs of scans in a recent work by Yan et al. [23]. We use door detections to align a 3D scan against a floorplan image. Door detection results are represented as a set of pixels in a floorplan image or per-scan evidence images. The semantic penalty is defined for every door-pixel in the evidence image. Let f_p denote the corresponding floorplan pixel under the scan placement. The penalty is 0 if f_p is also a door-pixel. Since not all the doors are marked in a floorplan, f_p may be an unmarked door-pixel. In that case, f_p must be in a door-way (i.e., white pixel). Therefore, if f_p is not a door pixel and has intensity more than 0.4, we set the penalty to 0.5, otherwise 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 geometric consistencies between the two scans given their placements. The photometric consistency uses normalized cross-correlation of local image patches. The geometric consistency measures the discrepancy between the point and free-space evidence information stored over the voxel grids. The consistency measures are based on standard techniques, and we refer the details to the supplementary material.

4.3. Floorplan coverage potential

The third potential seeks to increase the number of floorplan pixels covered by the scan. This acts as a counter-force against the scan-to-scan consistency potential, which has a strong bias to stack multiple scans at the same location, because 1) this potential was added for every pair of scans; and 2) the potential goes down only when scans overlap.

The floorplan coverage potential can be implemented by the sum of the sub-potentials over the floorplan pixels, each of which returns 1 if the pixel is not covered by any scans, otherwise 0. We define that "a scan covers a floorplan pixel" if the pixel is inside the free-space mask. This sub-potential depends on any scan, one of whose placement candidates covers the pixel, and usually becomes higher-order. In practice, most rooms are scanned only once or twice, and the approximation by pairwise terms has worked well. More precisely, for every floorplan pixel, we identify a set of scans, one of whose placement candidates covers the pixel. For every pair of such scans, we form a pairwise potential that becomes 0.0 if exactly one of the scans covers the pixel (ideal case), 0.5 if both cover the pixel, and 1.0 if none covers.

5. Inference

Naive inference will be infeasible to solve our MRF problem. The label space is massive (i.e., 4 rotations x 50 million translations per variable). The energy is not submodular. The key insight is that while indoor scenes are full of repetitions, there are not too many places or rooms that have exactly the same surrounding geometry and door placements. Therefore, simply identifying significant (negative) peaks in the unary potential can restrict a set of feasible placements for each scan. In practice, exhaustive evaluation of all the unary costs are still infeasible (200 million possible placements per scan), and we employ a standard hierarchical search scheme to identify a small number (5 in our experiments) of placement candidates per scan.

The hierarchical search scheme works as follows. First, we build an image pyramid of 5 levels for each floorplan image, an evidence image, or a door-detection image. Second, we exhaustively evaluate all the unary costs at the top level, and keep all the local minima less than a threshold. Then, by level by level, we iterate evaluating the unary costs at the children pixels under the current local minima, and applying non-local min suppression with a thresholding. This hierarchical search runs for each of the four orientations. The 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 intensity [1] 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 [17] 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, merged point cloud and colored floorplan-masks are shown. When multiple masks cover the same pixel, the color of the closest scan is chosen.

We have manually inspected every result to check the placement correctness, where erroneous ones are highlighted in the figure. Table 2 shows our placement error

[1] We assume that black pixels (walls) have a greater intensity than the white pixels (floor) in the *point* and *free-space* evidence, and the input floorplan

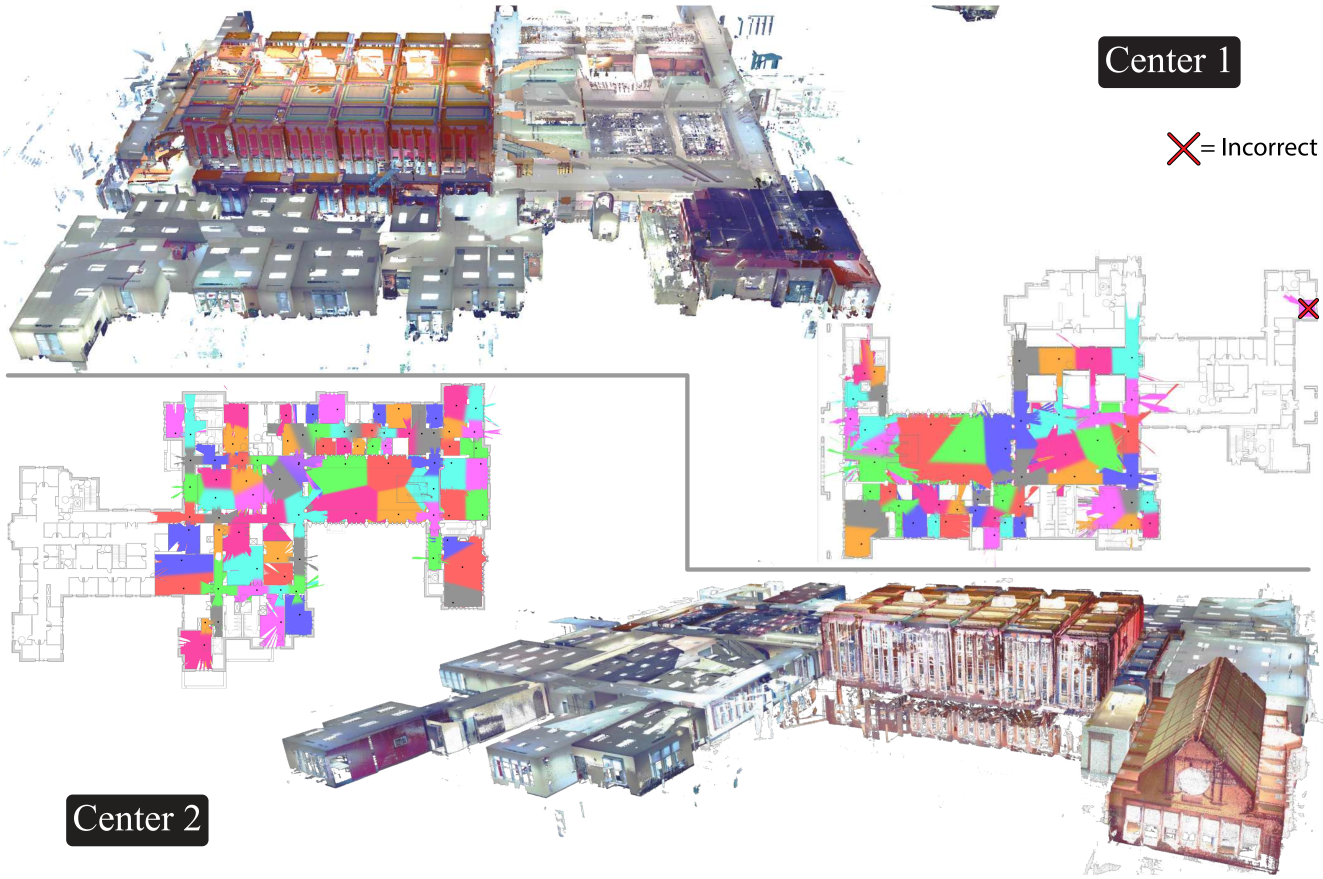

Figure 5: Placement results for Center1 and Center2. A merged 3D point-cloud, and 2D colored free-space masks are shown.

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.

rates (i.e., the ratio of incorrectly placed scans). Our algorithm 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 in the floorplan image. We will discuss failure cases later.

Our MRF formulation consists of the three potentials. We have run our algorithm with a few different combinations to assess their contributions. Table 2 shows relatively low error rates for the unary-only results (SF) and demonstrates the power of utilizing floorplans. The table also shows that the standard pairwise potential (SS), the main cue for existing approaches, has consistently improved the 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.

Furthermore, we have experimented the feasibility of conventional scan-to-scan alignment techniques, in particular, Autodesk ReCap 360 [1] and K-4PCS [22], which do not use floorplan data. Both methods have failed to generate any type of meaningful result, again confirming the importance of utilizing a floorplan image for our problems. Note that for fair comparison, we have evaluated the fully automated mode in Autodesk ReCap 360. We have also utilized its interactive mode in aligning the scans, but the process has been extremely painful and time-consuming (6 or 7 hours of intensive mouse clicking for large dataset). Furthermore, the final alignments have suffered from major errors due to the forced automatic refinement, which cannot be avoided.

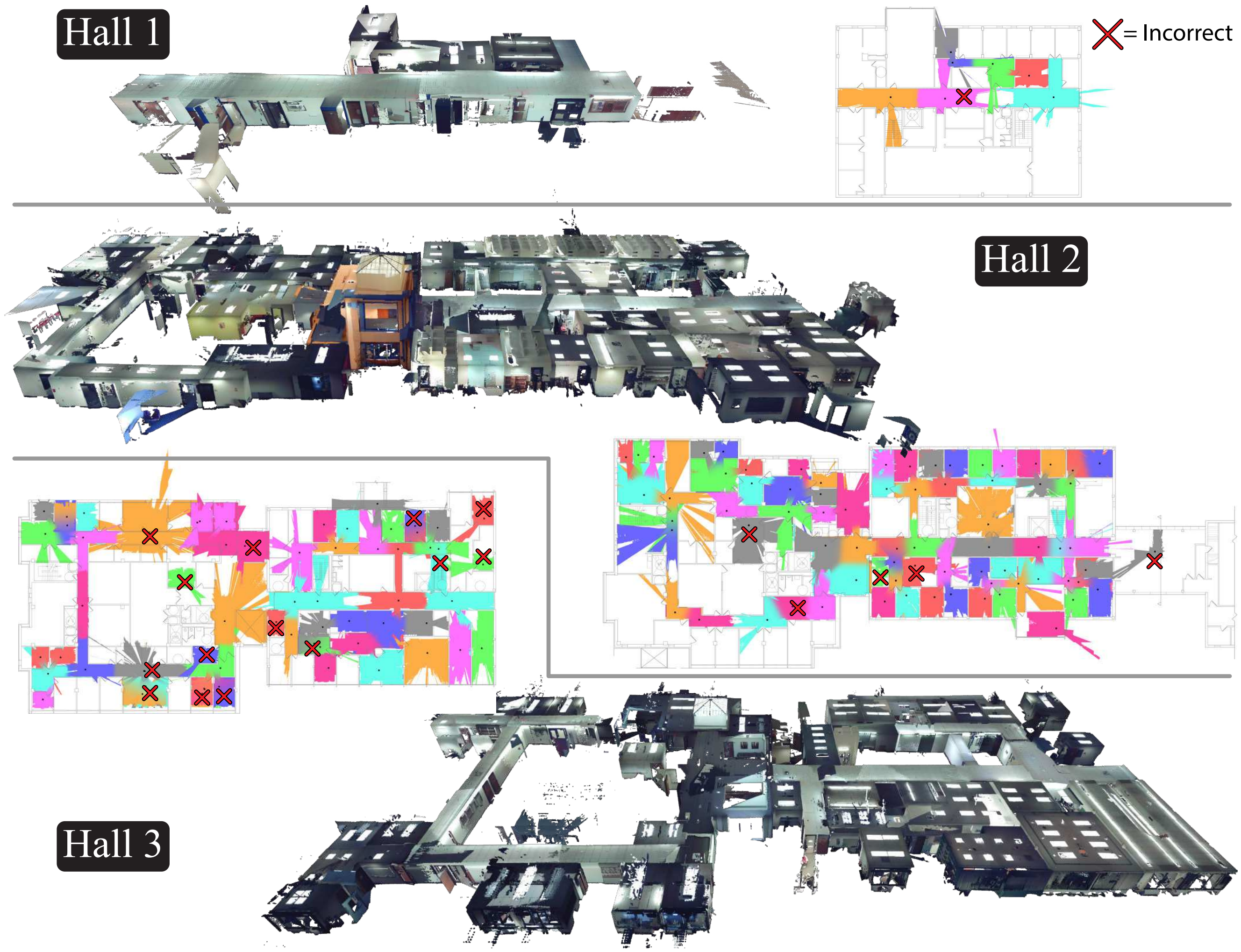

Figure 6: Placement results for Hall1, Hall2, and Hall3.

Figure 7: Results when replacing our unary potential with a distance transform on Center 2. Major placement errors occur in the middle of a big hall. The right cutaway point rendering shows the magnitude of the error.

To further evaluate the contribution of our unary potential and the effectiveness of utilizing a floorplan image, we 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

Name	#Scans	Naive SSD Top 1	Naive SSD Top 5	SSD Top 1	SSD Top 5	SSD Final	Distance transform Top 1	Distance transform Top 5	Distance transform Final	Ours (unary) Top 1	Ours (unary) Top 5	Ours (unary) Final
Center 1	50	96%	94%	24%	12%	12%	14%	4%	12%	6%	2%	2%
Center 2	80	94%	91%	18%	10%	10%	20%	12%	13%	1%	0%	0%
Hall 1	7	100%	100%	43%	0%	15%	29%	29%	29%	15%	0%	15%
Hall 2	75	100%	94%	48%	26%	37%	12%	8%	16%	22%	10%	7%
Hall 3	65	97%	96%	71%	56%	61%	45%	22%	37%	45%	32%	34%

Table 3: We have compared our results against several image distance metrics that replace the scan-to-floorplan (unary) potential. Columns "Top 1" and "Top 5" indicate if the ground-truth is in the top 1 or 5 placements based solely on the unary potential. Column "Final" reports the error rate after the MRF optimization with the replaced unary potential. Naive SSD simply takes the sum of squared differences between the floorplan image and the point-evidence image. SSD computes the same measure but only inside the free-space evidence mask. Distance transform makes the floorplan image into binary with a threshold 0.4, constructs its distance-transform image [9], then takes the sum of element-wise product with the point-evidence image inside the free-space evidence mask.

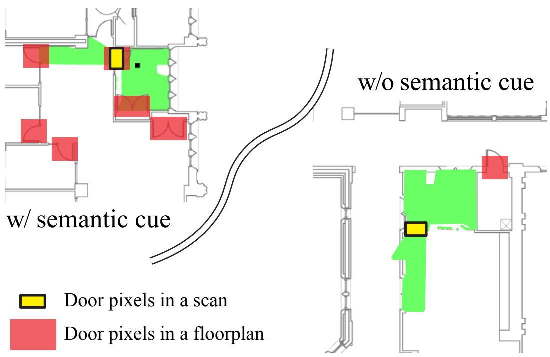

Figure 8: The semantic cue (i.e., door detection) resolves ambiguities. The figure shows the best placement based on the unary potential with or without the semantic cue.

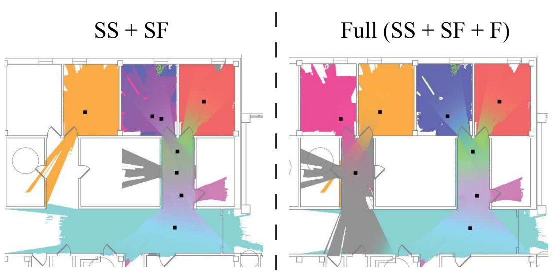

Figure 9: Final scan placements with or without the floorplan coverage potential, which mitigates the stacking bias visible on the left.

transform, which tends to concentrate scans in large rooms. Our analysis is that a large room tends to have non-architectural lines or symbols. The distance transform image contains significantly smaller values in this scenario than it would in absence, and allows incorrect placements to have a lower energy.

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 near identical. Lastly, our method has made major errors in Hall 3, simply because the floorplan has not reflected recent structural renovations. 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, inparticular, 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. Our source code and building-scale datasets are publicly available to further enhance indoor mapping research [3].

Acknowledgments This research is partially supported by National Science Foundation under grant IIS 1540012 and IIS 1618685.

References

[1] Autodesk recap 360. https://recap360.autodesk.com/. 2, 6

[2] Matterport. https://matterport.com. 1, 3

[3] Rgbd alignment project page. https://erikwijmans.github.io/WashU-Research. 8

[4] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 3

[5] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 1

[6] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Robotics-DL tentative*, pages 586–606. International Society for Optics and Photonics, 1992. 3

[7] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard. Simultaneous localization and mapping: Present, future, and the robust-perception age. *arXiv preprint arXiv:1606.05830*, 2016. 2

[8] S. Choi, Q.-Y. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5556–5565. IEEE, 2015. 2

[9] P. Felzenszwalb and D. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell University, 2004. 8

[10] L. S. F. Focus3D. http://www.faro.com. 1, 3

[11] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, et al. Building rome on a cloudless day. In *European Conference on Computer Vision*, pages 368–381. Springer, 2010. 3

[12] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[13] S. Ikehata, H. Yang, and Y. Furukawa. Structured indoor modeling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1323–1331, 2015. 3

[14] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011. 2

[15] K. Karsch, M. Golparvar-Fard, and D. Forsyth. Constructaide: analyzing and visualizing construction sites through photographs and building models. *ACM Transactions on Graphics (TOG)*, 33(6):176, 2014. 2, 3

[16] B. Klingner, D. Martin, and J. Roseborough. Street view motion-from-structure-from-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 953–960, 2013. 3

[17] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1568–1583, 2006. 5

[18] C. Mura, O. Mattausch, A. J. Villanueva, E. Gobbetti, and R. Pajarola. Automatic room detection and reconstruction in cluttered indoor environments with complex room layouts. *Computers & Graphics*, 44:20–32, 2014. 3

[19] 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

[20] 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

[21] 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

[22] 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, 6

[23] 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

[24] 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