Pacific Graphics (2018)
H. Fu, A. Ghosh, and J. Kopf (Editors)

Short Paper

Recovering 3D indoor floor plans by exploiting low-cost spherical photography

Giovanni Pintore [1], Fabio Ganovelli [2], Ruggero Pintus [1], Roberto Scopigno [2], and Enrico Gobbetti [1]

[1] Visual Computing, CRS4, Italy
[2] Visual Computing Group, ISTI CNR, Italy

Abstract

We present a novel approach to automatically recover, from a small set of partially overlapping panoramic images, an indoor structure representation in terms of a 3D floor plan registered with a set of 3D environment maps. Our improvements over previous approaches include a new method for geometric context extraction based on a 3D facets representation, which combines color distribution analysis of individual images with sparse multi-view clues, as well as an efficient method to combine the facets from different point-of-view in the same world space, considering the reliability of the facets contribution. The resulting capture and reconstruction pipeline automatically generates 3D multi-room environments where most of the other previous approaches fail, such as in presence of hidden corners, large clutter and sloped ceilings, even without involving additional dense 3D data or tools. We demonstrate the effectiveness and performance of our approach on different real-world indoor scenes.

CCS Concepts

•***Computing methodologies*** → *Computer graphics;* ***Computational photography; Shape inference; Reconstruction;***

1. Introduction

With the emergence of consumer-level 360° cameras, large and complex environments can now be captured with very few single-shot panoramic images, whose overlap can provide global registration information from just a few features. Such geometrically sparse, but visually rich, coverage is a very interesting and simple alternative to dense shape capture, as done with scanners or dense multi-view, especially in applications where location awareness and structure reconstruction is more important than fine geometric acquisition, such as acquiring and recovering *as-built* models of an indoor environment. Creating models of indoor environments just from visual data is, however, not an easy task, due, for instance, to poor texture detail, large occlusions, and complex floor-plan topology, which often lead to solutions that need elaborate acquisition and stitching processes, requiring complex reasoning to reconstruct invisible parts, often including manual intervention, especially in multi-room environments. In recent years, see Sec. 2, research has focused on extending conventional image-based approaches for indoor reconstruction by exploiting panoramic imagery. However, these solutions still have many limitations. Solutions based on dense capture typically require long processing times and features to extract a dense point cloud. Faster solutions typically focus on one panoramic image per room, but are capable to infer 3D clues only under very limiting constrains (e.g., *Manhattan World*). Furthermore, all these methods are limited by the strict condition that all the corners of the room must be visible from a single point of view, which make them ineffective in many common indoor environments (e.g., L-shapes, multi-room scenes, corridors). In order to address these issues, we propose a novel and lightweight approach, which efficiently combines the analysis of individual images with multi-view clues (see Sec. 3). Our main contribution to the state-of-the-art in indoor reconstruction are a novel geometric context extraction approach based on the combination of color/spatial reasoning with sparse multi-view 3D features, dubbed *3D facets* (Sec. 3.2), an efficient method to combine 3D facets from different images and evaluating their reliability (Sec. 3.3), as well as a novel and practical image-based pipeline to automatically retrieve a multi-room indoor 3D layout from a small set of panoramic images.

2. Related Work

Purely image-based techniques for indoor capture have to cope with texture-less walls and occlusion. This has led to the emergence of methods that aid reconstruction by imposing domain-specific constraints. For example, several authors (e.g., [FCSS09, FMR11, TXLK11]) exploit the heavily constraining *Manhattan World* [CY99] assumption to reconstruct the 3D structure of moderately cluttered interiors. Bao et al. [BFFFS14], similarly to our work apply instead both multi-view geometry and single-view analysis, but focus on estimating a single room layout and the fore-

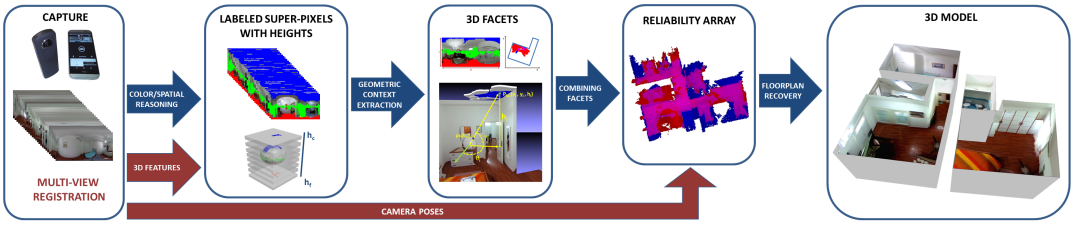


Figure 1: **Method pipeline.** Starting from few overlapping spherical images, we recover sparse 3D features and registered cameras. Such global information is used to enrich a per-image labeling into wall, ceiling and floor-superpixels, to robustly create 3D facets and segmentation reliability maps. A fusion approach finally generates a global 3D model.

ground objects rather than multi-room structures. In general, however, methods based on pin-hole image capture require a large number of shots. The recent emergence of consumer spherical cameras promises to improve visual capture of indoor environment, since each image covers the complete environment around the viewer, simplifying geometric reasoning, and very few images are required for a large coverage, simplifying the capture process and the features tracking. In recent years, several research efforts have focused on approaches for indoor reconstruction from panoramic images coming from commodity 360° cameras or from stitching regular photographs. Cabral et al. [CF14] adopted stitched equirectangular images to improve indoor reconstruction provided by a dense multi-view pipeline [FCSS09]. As clutter and homogeneous zones in indoor scenes tend to leave large reconstruction holes for image-based methods, their method exploits the labeling of the panoramas to complete the multi-view reconstruction obtained from pin-hole images. However, the approach assumed a considerable number of images and a dense point cloud, thus requiring considerable efforts in terms of user capture and processing time. Moreover, dense point clouds are not easily obtained in indoor environments lacking features. With the goal of minimizing user's burden, simplify geometric reasoning, and obtain solutions working on indoors with many featureless areas, recent state-of-the-art approaches [YZ16, PGG*16] focus on using only one panoramic image per room. Yang et al. [YZ16] propose an efficient method to recover the 3D shape of a single room based on a constraint graph encoding the spatial configurations of Manhattan World line segments and super-pixels of a single panoramic image. Although effective in many indoor layouts, this approach is limited only to single room environment where all the corners are visible from the same point-of-view. Similarly to Yang et al. [YZ16], Pintore et al. [PGG*16] integrate the super-pixel labeling through the analysis of the image's edgemap, extending the result for the single room to multi-room environments with the aid of motion sensors embedded in a mobile device. Although in a less restrictive way than Manhattan World, their approach works only by imposing fixed horizontal floor and ceiling plans, and with environments where all the structural features of the room are visible from an unique position. In this work, we improve over previous solutions by presenting an approach that, starting from a small set of panoramic images, recovers the 3D floor plan of a multi-room environment, by exploiting at the same time multi-view 3D data and single-view image analysis. Such approach is more robust to errors and provide a consistent reconstruction even when previous methods fail.

3. Method

Our pipeline, summarized in Fig. 1, starts from a set of partially overlapping equirectangular images aligned to the gravity vector. For each spherical image, we perform an image-based classification based on *super-pixels*, labeling only those super-pixels that can be unambiguously assigned to floor, walls, and ceilings (Sec. 3.1). In parallel, we recover global camera and features alignment through a multi-view registration of the images. Local and global information are jointly exploited to recover 3D world space points from image-space super-pixels, and to generate world-space 3D facets distribution (Sec. 3.2), which are then fused to recover the scene floor-plan and the relative 3D rooms shapes (Sec. 3.3).

3.1. Color/spatial reasoning: image labeling

We perform a super-pixel segmentation and geometric context labeling, which assigns regions of the image to *ceiling*, *floor*, *wall* zones, leaving undecided areas labeled as *unknown*. We start by labeling as *ceiling* the top most super-pixels row, *floor* the bottom ones, and *wall* the ones lying on the image horizon (i.e., middle of the equirectangular image). All others are set to *unknown*. We iteratively propagate known structure labels by enforcing the super-pixel label order, defining a distance function D which appropriately combines color similarity and spatial proximity [ASS*12]. Since our final goal is to integrate many partial, but reliable, image contributions (Sec. 3.2), we perform labeling propagation following a *conservative* strategy (Fig. 2 left), that is we do not assign a geometric context (i.e. ceiling, floor, wall) to super-pixel in the image if the distance to known neighbors is too large (i.e., D is above a threshold).

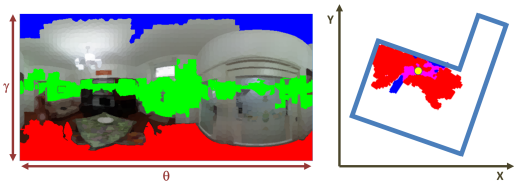


Figure 2: **Image segmentation and resulting facets.** Left: our super-pixels conservative segmentation. Right: the same ceiling and floor super-pixels represented as *facets* (see Sec. 3.2).

3.2. Geometric context extraction: 3D facets

In order to infer a reliable geometric context for each point-of-view, we introduce the concept of *3D facets*, which associate a height

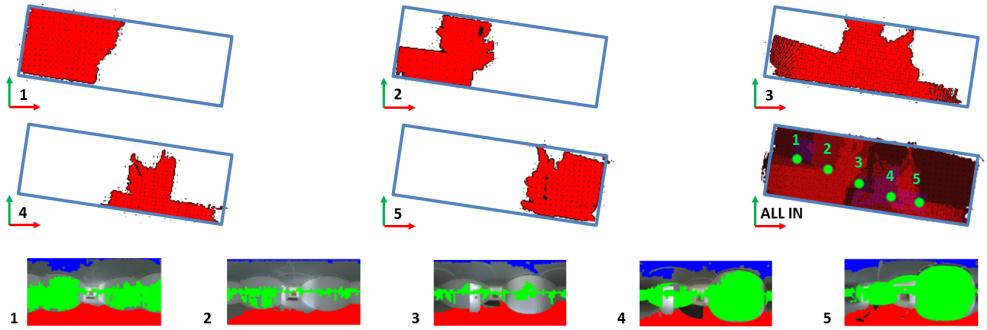


Figure 3: **Facets combination.** Example of facets joining from five labeled images (only floor facets are showed to simplify the illustration).

value to each labeled superpixel. Super-pixel points are thus associated to the following transform:

$$P_{loc}(\theta,\gamma,h_k) = \begin{cases} x_l = h_k/\tan\gamma * \cos\theta \\ y_l = h_k/\tan\gamma * \sin\theta \\ z_l = h_k \end{cases} \quad (1)$$

where $P_{loc} \in F_k$ is the 3D position of the pixel $(\theta,\gamma) \in SP_k$. Specifically, a *floor* or *ceiling* facet F_k is a horizontal patch corresponding to a specific super-pixel SP_k, parametrized on its height h_k. Such representation has several advantages in order to identify the underlying structure. The footprint of *floor* and *ceiling* facets, infact, highlights the shape of the room (i.e., Fig. 2, right and Fig. 3). *Wall* super-pixels on the other hand, since are distributed on vertical planes, can be represented as *anchor points* in the transformed space, to enforce 3D reconstruction (see Sec. 3.3).

To estimate the height of labeled super-pixels, we exploit the 3D position of matched features coming out of the SfM pipeline, which also extracts each spherical camera pose. Initially only super-pixels on which 3D features fall have an height, thus we propagate heights to *all* the labeled super-pixels, according to their neighbors connectivity, assuming that there is at least one height coming from SfM in a connected labeled region. This ensures that height values will be assigned to all super-pixels in the floor and ceiling regions, which are two single connected regions by construction.

3.3. Combining 3D facets from different images

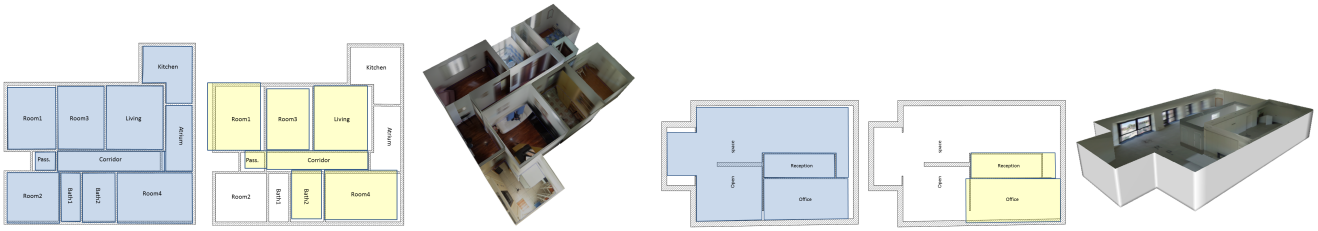
Combining the ceiling and floor facets from many view should highlight the wall rooms shape (see Fig. 3), that is the 2D contour of the merged facets. However, simply joining all the facets assumes that their generative super-pixels have been perfectly segmented and classified. Actually, mostly due to indoor imagery quality and spherical distortion, important errors could affect the final result, such as a noisy segmentation or an inaccurate height assignment due to texture-less regions. Compared to competing solutions [YZ16, PGG*16, CF14], we improved the robustness with respect to noisy segmentation and texture-less regions through multi-view fusion. Since for each part of the scene we have more than one labeled view, we exploit this redundancy to assign a *reliability score* to each 3D point projected, and possibly to discard unreliable results. We project all the 3D points from the *ceiling* and *floor* facets on a XY grid (spacing of 4 cm in all the presented experiments), arranging the projected points in an *accumulation array*. Each cell contains the occurrences of a each labeled point, that is how many images cover the same spatial position. Furthermore each cell can be at the same time covered by ceiling and floor facets. Joining in the same cell both ceiling and floor contributions makes room shape reconstruction more robust against many clutter problems (e.g., furniture covering the floor but not the ceiling). Unreliable cells are those seen only from a single point-of-view. A side effect of just removing values that are seen only from few points of views, is obviously the eventuality to filter out also correct details, for example small peripheral parts of the structure that are barely seen and labeled only from a single image. To compensate for this effect or to eventually complete parts that are not been labeled as ceiling or floor at all, we exploit the data labeled as *wall* to optimize the room shape. We represent the contour of the room as a 2D polygon of k corners and relative $S_k(\bar{s_0},\cdots,\bar{s_k})$ segments. The contour is initialized just from the merged ceiling/floor facets. We then exploit such *wall* contribution as 2D *anchor points*, with the goal of minimizing the distance between the floor/ceiling segmentation and the wall segmentation. We evaluate then distances of the S_k segments to their closest anchor points k $(W_0,\cdots,W_k)$. Given the elements count $W_{i_{count}}$ of each subset points, with $i \in [0,\cdots,k]$, we formalize the optimization problem as (Eq. 2):

$$R_{2k} \equiv \underset{\bar{R}}{argmin} \sum_{i=0}^{k} \sum_{j=0}^{W_{i_{count}}} dist(W_i(j),\bar{s_i})^2 \quad (2)$$

which, once expressed in matrix form, can be solved as non-linear least squares problem with Levenberg-Marquardt iterations. Once all the corners of the floor-plan have been recovered, we can reconstruct the 3D model by exploiting the 3D information (i.e., the height) contained in their closest facets.

4. Results

To demonstrate our approach we developed a reconstruction pipeline that, starting from a collection of spherical images and their multi-view alignment, automatically produces a structured 3D floor plan in terms of interconnected rooms bounded by walls. Most of the benefits of our method are in its use in multiple and structured environments and, in general, where single view approaches are ineffective or less reliable. In terms of multi-room structure extraction our method is comparable with the method of Pintore et al. [PGG*16], which is the most close to ours, although limited by many more assumptions, among which having a single image per room. We show in Fig. 4 the comparison against a real and



(a) *D1 : our* (b) *D1 : [PGG*16]* (c) *D1 : 3D reconstruction* (d) *D4 : our* (e) *D4 : [PGG*16]* (f) *D4 : 3D reconstruction*

Figure 4: **Recovered footprint and 3D models vs. ground truth floor plan.** Comparisons against real, metrically scaled, ground truth (grey footprint), of our method (first column) and the multi-room approach of Pintore et al. [PGG*16] (second column). We show in the third column our final textured 3D floor plan. Ceilings and septal wall have been removed from 3D reconstruction to make the illustration more clear.

metrically scaled ground truth (background layer), and the resulting textured models returned by our system (e.g., texturing only for illustrative reasons). In the first case (*D1* dataset), we show the reconstruction of a typical apartment layout. As each room is a fairly regular structure, the main challenges are the splitting of spaces (eleven rooms) and the clutter. Our method (Fig. 4(a)) returns almost perfect spatial mapping and shape for each room, with an overall area error (calculated on real footprint *including* walls thickness), with respect to ground truth, of about *5%*. In the second column (Fig. 4(b)), we show the same environment reconstructed with the approach of Pintore et al. [PGG*16], where, mainly because of clutter, the reconstruction of some rooms failed. Furthermore, due to the rooms joined through doors matching, considerable global mapping errors are present. The second case (*D4* dataset) is a larger and complex structure, where an office layout has been created into a former factory, having *Non-Manahttan World* corners and very thick walls. Such layout is arranged in 3 functional spaces (reception, office, open space) along 290 square meters. In particular, the open space is distributed around the central reception and a septal wall, describing an U-shape, impossible to be captured with only one view. Also in this case, our method returns a reliable reconstruction 4(d) and a very low error, 8%, especially when considering the large size and the peculiar topology. On the other hand, the compared approach, as it expect rooms where all corners are visible from a single point-of-view, definitely fails the reconstruction of the main room 4(e). As we expected, our approach returns a

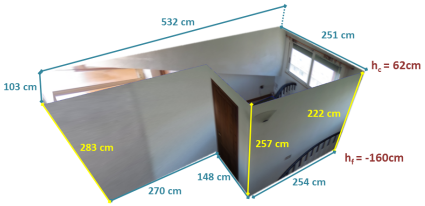


Figure 5: **sloped ceiling case.** *D3-Attic*.

reliable reconstruction also when the compared approaches fail to find the room structure, such as in presence of hidden corners (i.e., D1:Atrium, D1:Room2), large clutter (i.e., D1:Kitchen), Sloped ceilings (i.e., D3:Attic, Fig. 5) or complex environment containing more than one of these issues at the same time (i.e., D4:Open space).

5. Conclusions

We presented a novel and practical approach for recovering 3D indoor structures using 360° images. Differently from most previous works, we exploit a reasoning approach based on few registered image features coming from multiple images to recover a consistent geometric context and resolve occlusions from single viewpoint, without the need to impose strictly Manhattan World constraints. As a result, only few overlapping images are required to generate a 3D floor plan, even when other previous approaches fail, such as in presence of hidden corners, large clutter and more complex multi-room structures. We envision, for the future, to extend the mixing of single-view and multi-view labeling to extract other structural information from the data, such as the clutter in the rooms, in order to create a complete furnished 3D model.

Acknowledgments. This work was partially supported by projects VIGEC and 3DCLOUDPRO. The authors also acknowledge the contribution of Sardinian Regional Authorities.

References

[ASS*12] ACHANTA R., SHAJI A., SMITH K., LUCCHI A., FUA P., SUSSTRUNK S.: SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI 34*, 11 (2012), 2274–2282. 2

[BFFFS14] BAO S. Y., FURLAN A., FEI-FEI L., SAVARESE S.: Understanding the 3D layout of a cluttered room from multiple images. In *Proc. IEEE WACV* (2014), pp. 690–697. 1

[CF14] CABRAL R., FURUKAWA Y.: Piecewise planar and compact floorplan reconstruction from images. In *Proc. CVPR* (2014), pp. 628–635. 2, 3

[CY99] COUGHLAN J. M., YUILLE A. L.: Manhattan world: Compass direction from a single image by bayesian inference. In *Proc. IEEE ICCV* (1999), vol. 2, pp. 941–947. 1

[FCSS09] FURUKAWA Y., CURLESS B., SEITZ S. M., SZELISKI R.: Reconstructing building interiors from images. In *Proc. IEEE ICCV* (2009), pp. 80–87. 1, 2

[FMR11] FLINT A., MURRAY D., REID I.: Manhattan scene understanding using monocular, stereo, and 3D features. In *Proc. IEEE ICCV* (2011), pp. 2228–2235. 1

[PGG*16] PINTORE G., GARRO V., GANOVELLI F., AGUS M., GOBBETTI E.: Omnidirectional image capture on mobile devices for fast automatic generation of 2.5D indoor maps. In *Proc. IEEE WACV* (2016), pp. 1–9. 2, 3, 4

[TXLK11] TSAI G., XU C., LIU J., KUIPERS B.: Real-time indoor scene understanding using bayesian filtering with motion cues. In *Proc. IEEE ICCV* (2011), pp. 121–128. 1

[YZ16] YANG H., ZHANG H.: Efficient 3D room shape recovery from a single panorama. In *Proc. IEEE CVPR* (2016), pp. 5422–5430. 2, 3