Mobile metric capture and reconstruction in indoor environments

Giovanni Pintore
CRS4
Cagliari, Italy
giovanni.pintore@crs4.it

Fabio Ganovelli
ISTI-CNR
Pisa, Italy
fabio.ganovelli@isti.cnr.it

Roberto Scopigno
ISTI-CNR
Pisa, Italy
roberto.scopigno@isti.cnr.it

Enrico Gobbetti
CRS4
Cagliari, Italy
enrico.gobbetti@crs4.it

ABSTRACT

Mobile devices have become progressively more attractive for solving environment sensing problems. Thanks to their multi-modal acquisition capabilities and their growing processing power, they can perform increasingly sophisticated computer vision and data fusion tasks. In this context, we summarize our recent advances in the acquisition and reconstruction of indoor structures, describing the evolution of the methods from current single-view approaches to novel mobile multi-view methodologies. Starting from an overview on the features and capabilities of current hardware (ranging from commodity smartphones to recent 360° cameras), we present in details specific real-world cases which exploit modern devices to acquire structural, visual and metric information.

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**; **Reconstruction**; *Image processing*;

KEYWORDS

Indoor reconstruction; omnidirectional images; mobile mapping

ACM Reference format:
Giovanni Pintore, Fabio Ganovelli, Roberto Scopigno, and Enrico Gobbetti. 2017. Mobile metric capture and reconstruction in indoor environments. In *Proceedings of SA '17 Symposium on Mobile Graphics & Interactive Applications , Bangkok, Thailand, November 27-30, 2017,* 5 pages.
https://doi.org/10.1145/3132787.3139202

1 INTRODUCTION

The last decade has been marked by the increased proliferation and availability of a large variety of mobile devices. *Mobility* is now a common feature shared by different consumer devices, such as smartphones, tablets and compact panoramic cameras, as well as by specialized embedded devices, such as the hardware for autonomous driving and assistive technology, or the sensing devices mounted on drones and robots. By integrating mobility with multi-core CPU-GPU capabilities, high-res and flexible cameras with interactive screens and absolute and relative motion sensors, these devices open up new visual computing possibilities. In particular, the evolution of sensors and processing power extends the range of possible uses from 2D applications (e.g., image enhancing, stitching or matching, object detection, texture classification, activity recognition) to novel 3D applications, such as camera localization, pose estimation, augmented reality, and 3D scene reconstruction [Agus et al. 2017a,b]. Capture and reconstruction of indoor environments [Pintore et al. 2016a] is an important and challenging field affected by these enhancements.

Figure 1: Modern mobile panoramic cameras driven by smartphones (top left) can easily capture sequences of spherical images (bottom), which can be exploited to automatically reconstruct structured 3D models (top right).

Nowadays, solutions to acquire the shape of indoor environments range from assisted manual sketching of floor plans to automated methods that process high-density scans (e.g., [Mura et al. 2014]). Such dense point clouds can be acquired with laser scanners or depth cameras, and, to a lesser extent, with photogrammetric techniques, which, however, often fail to reconstruct surfaces with poor texture detail. Generating and processing dense data is often lengthy, and the most advanced solutions require expensive equipment and specialized personnel. The use of these techniques is thus often restricted to specific application domains, such as cultural heritage, construction, architecture, or engineering. Current

SA '17 Symposium on Mobile Graphics & Interactive Applications , November 27-30, 2017, Bangkok, Thailand

ACM ISBN 978-1-4503-5410-3/17/11.
https://doi.org/10.1145/3132787.3139202

mobile devices, combined with computer vision techniques that directly extract sparse structural information, offer instead a very attractive platform to overcome these problems, especially in many real-world cases where mapping the structure and obtain a visual depiction of the scene is more important than capturing fine-scale 3D details (see Fig. 1). Furthermore, data fusion approaches which exploit integrated mobile instruments can solve the intrinsic scale problem of pure image-based methods, allowing to obtain models in real-world dimensions.

Following this trend, we present here our recent advances on indoor capture and metric reconstruction methods, from simple single pose approaches to novel improved multi-view techniques.

2 SINGLE-POSE-PER-ROOM CAPTURE APPROACH

The tight integration of a camera with an interactive screen supports the implementation of real-time interactive capturing applications on mobile devices. In the indoor environment, interaction is used, for instance, to enable non-technical people to create geometric models [Sensopia 2014] or to add visual information to support interactive virtual tours [Sankar and Seitz 2012]. Such user-assisted approaches have proven be effective for floor plan reconstruction, but have the drawback of requiring extensive and repetitive user inputs. Moreover, they are prone to errors caused by imprecise device handing or manual editing. To overcome these limitations, in the last years research focused on devising data fusion and computer vision techniques that integrate data from in-built sensors and images [AliAkbarpour et al. 2015; Martinelli 2012; Pintore and Gobbetti 2014], with the specific goal of automating the creation of structured models.

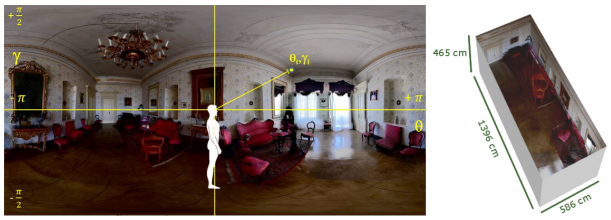

Figure 2: A full sphere capture and its relative indoor structure reconstruction [Pintore et al. 2016c].

It has been recently demonstrated [Pintore et al. 2016c; Pintore and Gobbetti 2014] that capturing the whole field-of-view (360° longitude and 180° latitude) of an ideal observer typically provides enough information to reconstruct an indoor scene without additional user interaction. Such an omnidirectional view can be captured with regular smartphones with off-the-shelf mobile guided applications (e.g., *Google PhotoSphere*, *Autostitch*), which guide stitching by exploiting the information coming from inertial and angular sensors (basically the accelerometer and the gyroscope). Moreover, the emergence of new 360° cameras is significantly reducing spherical images capturing efforts. Many consumer-level 360° cameras have just recently become available, and can readily be used as standalone mobile tools or driven by smartphones or tablets[1, 2] (see Figure 1) . Our solution [Pintore et al. 2016c]

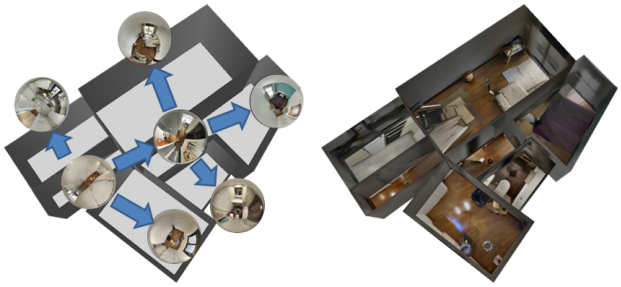

Figure 3: Exploiting a single spherical image for each room we obtain a structured and visually realistic model [Pintore et al. 2016b].

addresses the indoor capture and reconstruction problem directly from spherical images that map each view direction on a sphere centered around the observer (Fig. 2). We process a single panoramic image per room, resulting in a 2.5D floor plan describing an indoor environment in terms of rooms bounded by walls and connected by doors (Fig. 3). The basic idea of the method is to consider a room model in which vertical walls, aligned with the world gravity vector, are also aligned with the image gravity vector. We first extract edges from the images, and assume that a subset of those edges will correspond to the separation between vertical walls and floor/ceiling. We then segment the image by region growing and color similarity in order to filter out regions likely far from top/bottom edges of walls. The wall height is then found by a voting scheme, which determines the most likely wall height by maximizing pairs of matching wall-floor / wall-height edge pixels. Once the height is known, and a set of edge pixels are found, a full 2.5D room model is recovered from the edge map. The resulting model is then textured with the panoramic image. Such a structured model offers a simple scalable way to interactively explore visually realistic indoor environment on mobile setups [Pintore et al. 2016b]. Moreover, a full multiple-room environments can be obtained by mobile tracking of user's direction when moving between adjacent rooms in order to create a connected room graph. Doors position in the image can be identified by computer vision techniques, and matched to create graph edges and rooms displacements. A global optimization then produces a full combined model.

We implemented a minimal Android application (compatible with version 4.4 and higher) to capture a multi-room indoor scene using this method. This application keeps track of the user's movements between rooms and acquires the sphere-map of each environment. As demonstrated in Pintore et al. [2016b], this approach is capable to perform complex reconstructions with an area error for each single room of about 4% (with respect to ground truth), and an overall area error of about 13% (on a 70 rooms environment). The increased overall error is basically due to the fact that those single-pose methods do not consider walls thickness, a limitation lifted in our multi-view approach (see Sec. 4).

[1] http://www.samsung.com/global/galaxy/gear-360/
[2] https://theta360.com/en/about/theta/s.html

3 RESOLVING SCALE BY SENSOR FUSION

The above approach is capable to return a full 3D indoor environment made of rooms connected by doors, but, using images alone, is not able to resolve its scale. Even though scale information can be provided using external means (e.g., providing the height of the camera or the dimension of a reference object), a fully automated solution is made possible by the presence of multiple synchronized sensors on a mobile device.

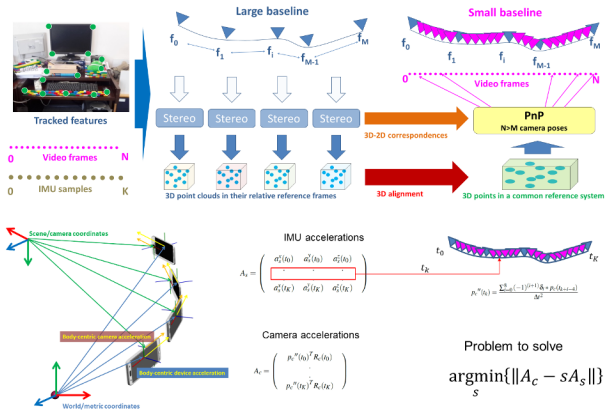

Figure 4: Scheme of the algorithm for mobile fast metric reconstruction [Garro et al. 2016]. Top pipeline: structure-from-motion synced with IMU samples and fully implemented on mobile device. Bottom: Illustration of the adopted reference systems and optimization strategy to recover the scale factor.

With the increasing processing power, it is now possible to have mobile pipelines integrating structure-from-motion and data fusion. For instance, Tanskanen et al. [Tanskanen et al. 2013], adopt a filtering approach based on *Kalman* filter to estimate a rough metric scale of the scene, exploiting *Verlet* integration to estimate the device physical trajectory from the inertial sensors and to compare it with the camera motion determined from the images. Using integration to go from accelerometer data to position introduces, however, important errors. We have thus recently presented a fast metric reconstruction method on mobile devices using a combination of image and inertial acceleration data [Garro et al. 2016], which, instead of integrating acceleration coming from inertial sensors, derives acceleration from the more precise position data derived from multi-view geometry (Fig. 4).

In order to estimate the scale (ratio meters/scene units), we thus compare the acceleration values from the inertial sensors with the ones inferred from images. The approach, fully working on a mobile phone, starts by detecting Shi-Tomasi features, and, when the baseline is large enough, it estimates the Essential Matrix determining the transformation from the current camera pose to the previous one. 3D points are then computed for each feature point. Feature correspondences are then used to align all the frame point clouds to the same reference system, therefore globally registering all the key camera poses in the traced path. The camera pose at each frame is then computed by PnP, in order to obtain a high-frequency sampled camera path. The scale factor is then recovered by matching the accelerations computed from the camera path with the acceleration from the IMU. In order to cope with outliers, this matching uses a RANSAC-based robust fitting method. Moreover, the fitting not only recovers the scale, but also the relative orientation of the IMU with respect to the camera.

Numerical results on scale estimation obtained with this approach have proven to provide an average scale error of about 3% on typical office scenes captured with common Android devices with a strong noise of inertial measurements (see detailed results in Garro et al. [2016]). Such results are similar to those obtained by post-processing desktop pipelines [Ham et al. 2014] based on long video sequences (about 1 minute). The results presented in Garro et al. [2016]) can be readily extended to modern 360° cameras, which are internally equipped with an IMU. In order to apply this method to the indoor capture technique of Sec. 2, common image features on the recovered structures (walls, ceiling, floor) must be matched to features present in the sequence used for detecting scale.

4 MULTIPLE-POSE-PER-ROOM CAPTURE APPROACH

The single-pose approaches introduced in Sec. 2 have a number of limitations. First, clutter within the room causes occlusion that can lead to hidden corners. Second, editing time or custom solutions based on user interaction are typically use to join multi-room scene (e.g., to detect doors). Third, scale detection is a post-hoc solution, which requires manual feature matching on recovered structures.

The emergence of new 360° cameras is significantly reducing omnidirectional capturing efforts, making it possible to consider overcoming the above limitations by acquiring more than one panoramic image per room and using multi-view reconstruction approaches.

We present here a brief overview of the approach we are developing to return a fully automatic metric reconstruction of the indoor environment also in presence of hidden corners or multi-room structures (Fig. 5).

The approach assumes that several (at least two) spherical images are captured per room. For each spherical image, we start by performing, in parallel, an image-based labeling and geometric context extraction based on *super-pixels*, similar to what is done in single-view approaches (e.g., [Cabral and Furukawa 2014; Yang and Zhang 2016]). Unlike previous methods, we label only those super-pixels that can be unambiguously assigned to floor, walls, and ceilings. Then, we recover the camera and feature alignment through a multi-view registration of the spherical images (many solutions are possible, ranging from the real-time approach described in Sec. 3 to off-line methods like for example the one presented in Kopf et al. [Kopf 2016]). Registration exploits the correspondences among individual sphere maps and the labeled super-pixel data, thus recovering real 3D relationships between poses and single super-pixels. We finally combine all the information in a single 3D structure through a specialized *spatial transform* based on an omnidirectional projection [Pintore et al. 2016c], unveiling the 3D floorplan shape. Since modern SPC (spherical panoramic cameras) are equipped with an IMU and support real-time image/video streaming (e.g. through custom connection or *http*, it is just a matter of implementation to apply the methods described in Sec. 3 to obtain

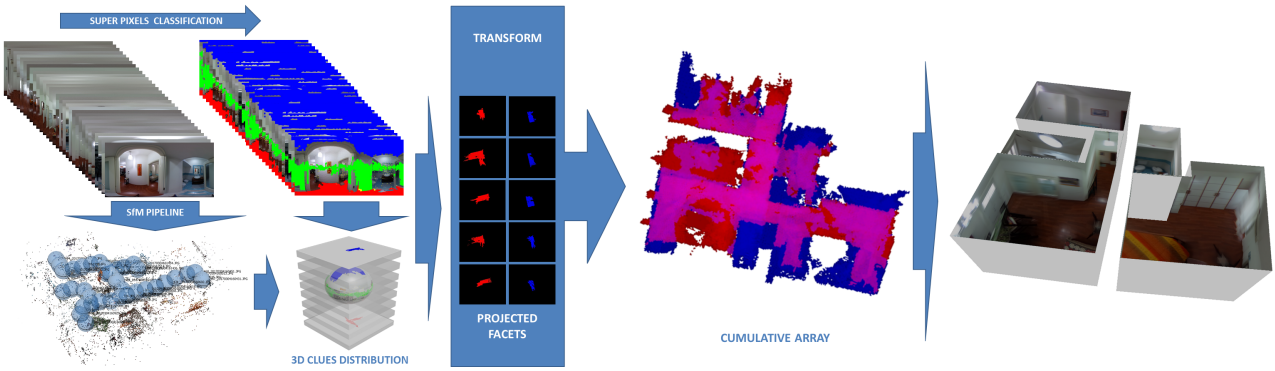

Figure 5: Indoor capture pipeline concept. From a sequence of panoramic images captures with a compact panoramic camera we perform an fully automatic image-based classification and a multi-view registration. By the merging of these information we automatically recover a 3D floorplan scaled in metric dimensions.

models scaled in real-world metric dimensions. We evaluated the

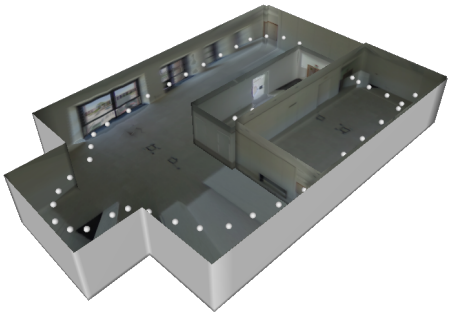

Figure 6: Office complex of 300 square meters reconstructed from 40 images (grey spheres in the illustration show user poses), acquired by a Ricoh Theta S camera driven by an HTC One M8 smartphone. This scene includes an open space and two big rooms embedded into the main space, resulting in a complex topology with several hidden points-of-view.

method on a variety of indoor environments, specifically on scenes where previous approaches fail, such as rooms with hidden corners, non-Manhattan World angles, and sloped ceilings. This multi-view approach outperforms previous single-pose approaches [Pintore et al. 2016c; Yang and Zhang 2016] both in terms of accuracy (overall area error 4% vs. 13%), both in terms of usability, covering a wider range of indoor scenes and returning a fully automatic and more accurate alignment between different rooms.

5 CONCLUSIONS AND FUTURE WORK

We have presented approaches capable to recover the metric structure of an indoor environment using commodity mobile devices. Our current work is focused on developing a fully mobile solution for indoor reconstruction based on the methods presented in this work. We are also extending the system to provide features and user interfaces aimed at exploiting the recovered models. 3D navigation is the more straightforward use, but the availability of visual data coupled to a coarse 3D model makes it possible to exploit the reconstruction for a number of important applications. A significant example is object retrieval within captured scenes. In the system we are currently developing, a query object is selected from one image, for example an electrical outlet, and through feature-based matching, its occurrences can be found across all images. As result the complete map of visible outlets can be automatically integrated with the 3D floor plan, providing useful additional information. We are also exploring how to improve data visualization and immersive experiences exploiting panoramic 360° videos and visors.

ACKNOWLEDGMENTS

This work was partially supported by project VIGEC. The authors also acknowledge the contribution of Sardinian Regional Authorities.

REFERENCES

Marco Agus, Enrico Gobbetti, Fabio Marton, Giovanni Pintore, and Pere-Pau Vázquez. 2017a. Mobile Graphics. In *Proc. EUROGRAPHICS Tutorials.*

Marco Agus, Enrico Gobbetti, Fabio Marton, Giovanni Pintore, and Pere-Pau Vázquez. 2017b. Mobile Graphics. In *Proc. SIGGRAPH Asia Tutorials.*

H. AliAkbarpour, K. Palaniappan, and G. Seetharaman. 2015. Fast structure from motion for sequential and wide area motion imagery. In *Proc. IEEE ICCVW Video Summarization for Large-scale Analytics Workshop.* https://doi.org/iccv15/

R. Cabral and Y. Furukawa. 2014. Piecewise Planar and Compact Floorplan Reconstruction from Images. In *Proc. CVPR.* 628–635.

Valeria Garro, Giovanni Pintore, Fabio Ganovelli, Enrico Gobbetti, and Roberto Scopigno. 2016. Fast Metric Acquisition with Mobile Devices. In *Proc. VMV.* 29–36.

Christopher Ham, Simon Lucey, and Surya Singh. 2014. *Proc. ECCV.* Chapter Hand Waving Away Scale.

Johannes Kopf. 2016. 360&Deg; Video Stabilization. *ACM Trans. Graph.* 35, 6, Article 195 (Nov. 2016), 9 pages. https://doi.org/10.1145/2980179.2982405

A. Martinelli. 2012. Vision and IMU Data Fusion: Closed-Form Solutions for Attitude, Speed, Absolute Scale, and Bias Determination. *IEEE Transactions on Robotics* 28, 1 (2012), 44–60. https://doi.org/10.1109/TRO.2011.2160468

Claudio Mura, Oliver Mattausch, Alberto Jaspe Villanueva, Enrico Gobbetti, and Renato Pajarola. 2014. Automatic Room Detection and Reconstruction in Cluttered Indoor Environments with Complex Room Layouts. *Computers & Graphics* 44 (2014), 20–32.

Giovanni Pintore, Fabio Ganovelli, Enrico Gobbetti, and Roberto Scopigno. 2016a. Mobile Mapping and Visualization of Indoor Structures to Simplify Scene Understanding and Location Awareness. In *Proc. ECCV Workshops.* 130–145.

Giovanni Pintore, Fabio Ganovelli, Enrico Gobbetti, and Roberto Scopigno. 2016b. Mobile reconstruction and exploration of indoor structures exploiting omnidirectional images. In *Proc. SIGGRAPH Asia Symposium on Mobile Graphics and Interactive Applications.* 1:1–1:4.

Giovanni Pintore, Valeria Garro, Fabio Ganovelli, Marco Agus, and Enrico Gobbetti. 2016c. Omnidirectional image capture on mobile devices for fast automatic generation of 2.5D indoor maps. In *Proc. IEEE WACV*. 1–9.

Giovanni Pintore and Enrico Gobbetti. 2014. Effective Mobile Mapping of Multi-room Indoor Structures. *The Visual Computer* 30, 6–8 (2014), 707–716.

Aditya Sankar and Steven Seitz. 2012. Capturing Indoor Scenes with Smartphones. In *Proc. ACM UIST*. 403–412.

Sensopia. 2014. MagicPlan. (2014). www.sensopia.com.

P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys. 2013. Live Metric 3D Reconstruction on Mobile Phones. In *Proc. ICCV*. 65–72. https://doi.org/10.1109/ICCV.2013.15

H. Yang and H. Zhang. 2016. Efficient 3D Room Shape Recovery from a Single Panorama. In *Proc. IEEE CVPR*. 5422–5430.