Automatic 3D Reconstruction of Structured Indoor Environments

Tutorial Notes

Giovanni Pintore
CRS4, Italy

Claudio Mura
University of Zurich, Switzerland

Fabio Ganovelli
ISTI-CNR, Italy

Lizeth Fuentes-Perez
University of Zurich, Switzerland

Renato Pajarola
University of Zurich, Switzerland

Enrico Gobbetti
CRS4, Italy

ABSTRACT

Creating high-level structured 3D models of real-world indoor scenes from captured data is a fundamental task which has important applications in many fields. Given the complexity and variability of interior environments and the need to cope with noisy and partial captured data, many open research problems remain, despite the substantial progress made in the past decade. In this tutorial, we provide an up-to-date integrative view of the field, bridging complementary views coming from computer graphics and computer vision. After providing a characterization of input sources, we define the structure of output models and the priors exploited to bridge the gap between imperfect sources and desired output. We then identify and discuss the main components of a structured reconstruction pipeline, and review how they are combined in scalable solutions working at the building level. We finally point out relevant research issues and analyze research trends.

CCS CONCEPTS

• **Computing methodologies** → **Computer graphics**; *Shape modeling*; **Computer vision**; *Computer vision problems*; *Shape inference*; *Reconstruction*; • **Applied computing** → *Computer-aided design.*

KEYWORDS

indoor reconstruction, indoor scanning, structured reconstruction

ACM Reference Format:
Giovanni Pintore, Claudio Mura, Fabio Ganovelli, Lizeth Fuentes-Perez, Renato Pajarola, and Enrico Gobbetti. 2020. Automatic 3D Reconstruction of Structured Indoor Environments: Tutorial Notes. In *Proceedings of SIGGRAPH 2020 Courses (SIGGRAPH2020 Courses).* ACM, New York, NY, USA, 10 pages.

1 FORMAT AND PRE-REQUISITES

Format. Long (3 hours).
Necessary background. The tutorial is at the intermediate level. Basic computer-vision and graphics background is a pre-requisite.
Intended audience. The target audience includes researchers in geometric modeling, as well as practitioners in the relevant application fields. Researchers will find a structured overview of the field, which organizes the various problems and existing solutions, classifies the existing literature, and indicates challenging open problems. Domain experts will, in turn, find a presentation of the areas where automated methods are already mature enough to be ported into practice, as well as an analysis of the kind of indoor environments that still pose major challenges.
Previous presentations. This tutorial builds on an extensive state-of-the-art survey that has been presented at Eurographics 2020 [Pintore et al. 2019b]. The Eurographics presentation version was a condensed STAR aimed at experts, and focused on the presentation of the literature survey. This course significantly extends it with tutorial-style presentations to accommodate a much more varied audience and to make the content more self-contained.

SIGGRAPH2020 Courses, July 2020, Washington, DC, USA
2020.

2 COURSE DESCRIPTION

The automated reconstruction of 3D models from acquired data, be it images or 3D point clouds, has been one of the central topics in computer graphics and computer vision for decades. This field is now thriving, as a result of complementing scientific, technological and market trends. In particular, in recent years, the widespread availability and proliferation of high-fidelity visual/3D sensors (e.g., smartphones, commodity and professional stereo cameras and depth sensors, panoramic cameras, low-cost and high-throughput scanners) has been matched with increasingly cost-effective options for large data processing (e.g., cloud and GPU-accelerated computation), as well as with novel means of visual exploration, from mobile phones to immersive personal displays.

In this context, one of the rapidly emerging sub-fields is concerned with the automatic reconstruction of indoor environments. That is, a 3D representation of an interior scene must be inferred from a collection of measurements that sample its shape and/or appearance, exploiting and/or combining sensing technologies ranging from passive methods, such as single- and multi-view image capturing, to active methods, such as infrared or time-of-flight cameras, optical laser-based range scanners, structured-light scanners, and LiDAR scanners [Berger et al. 2017]. Based on the raw data acquired by these devices, many *general* surface reconstruction methods focus on producing accurate and dense 3D models that faithfully replicate even the smallest geometry and appearance details. In this sense, their main goal is to provide the most accurate representation possible of all the surfaces that compose the input scene, disregarding its structure and semantics or possibly only exploiting them to maximize the fidelity of the output surface model. A number of more *specialized* indoor reconstruction solutions focus, instead, on abstracting simplified high-level structured models that optimize certain application-dependent characteristics [Ikehata et al. 2015].

The focus on high-level structured models is motivated by several reasons. First of all, their availability is necessary in many fields. For example, applications such as the generation or revision of building information models (BIM) require, at least, the determination of the bare architectural structure [Mura et al. 2014b; Turner et al. 2015]. On the other hand, information on the interior clutter, in terms of 3D footprint of major indoor objects, is necessary in many other use cases, such as guidance, energy management, security, evacuation planning, location awareness or routing [Ikehata et al. 2015]. Even when the goal is solely for visualization, structured simplified models need to be extracted as a fundamental component of a renderable model. This is because narrow spaces, windows, non-cooperative materials, and abundant clutter make the transition from the acquisition of indoor scenes to their modeling and rendering a very difficult problem. Thus, applying standard dense surface reconstruction approaches, which optimize for completeness, resolution and accuracy, leads to unsatisfactory results.

Automatic 3D reconstruction and modeling of indoor scenes, has thus attracted a lot of research in recent years, making it an emerging well-defined topic. In particular, the focus has been on developing specialized techniques for very common and very structured multi-room environments, such as residential, office, or public buildings, which have a substantial impact on architecture, civil engineering, digital mapping, urban geography, real estate, and more [Ikehata et al. 2015]. In this context, the fundamental tasks are the discovery of structural elements, such as rooms, walls, doors, and indoor objects, and their combination in a consistent structured 3D shape and visual representation. The research community working on these problems appears, however, fragmented, and many different vertical solutions have been proposed for the various motivating applications. In this course, we provide an up-to-date integrative view of the field, bridging complementary views coming from computer graphics and computer vision.

3 COURSE RATIONALE

Reconstruction of visual and geometric models from images or point clouds is a very broad topic in computer graphics and computer vision. This course focuses on the specific problems and solutions relating to the reconstruction of *structured 3D indoor models*, that is rapidly emerging as a very important and challenging problem, with specific solutions and very important applications. Thus, we complement existing courses and surveys focusing on reconstructing detailed surfaces from dense high-quality data or on assigning semantic to existing geometry, by covering the extraction of an *approximate structured geometry* connected to a *visual representation* from sparse and incomplete measurements.

The tutorial content is based on a recent survey of the state-of-the-art that we have published in Computer Graphics Forum [Pintore et al. 2019b], and presented at the 2020 Eurographics conference. We refer the audience to that STAR for an in-depth presentation of the concept and a detailed reasoned bibliography.

A general coverage of methods for 3D surface reconstruction and primitive identification is available in recent surveys [Berger et al. 2017; Kaiser et al. 2019], and we will build on them for the definition of general problems and solutions. In the same spirit, we do not specifically cover interactive or online approaches; those interested in online reconstruction can find more detail on the topic in the survey by Zollhöfer et al. [Zollhöfer et al. 2018]. We also will refer the audience to an established state-of-the-art report on urban reconstruction [Musialski et al. 2013] for an overview of the companion problem of reconstructing (from the outside) 3D geometric models of urban areas, individual buildings, façades, and further architectural details.

The techniques surveyed in this course also have an overlap with the domains of Scan-to-BIM or Inverse-CAD, where the goal is the automatic reconstruction of full (volumetric) information models from measurement data. However, the overlap is only partial, since we do not cover the assignment of full semantic information and/or the satisfaction of engineering construction rules, and Scan-to-BIM generally does not cover the generation of visual representations, which is necessary for rendering. Moreover, most Scan-to-BIM solutions are currently targeting (dense) point cloud data, while we cover solutions starting from a variety of input sources. It should be noted that, obviously, relations do exist, and many of the solutions surveyed here can serve as good building blocks to tackle the full Scan-to-BIM problem. We will refer the audience to established surveys in the Scan-to-BIM area for a review of related techniques based on point-cloud data [Pătrăucean et al. 2015; Tang et al. 2010; Volk et al. 2014], general computer vision [Fathi et al. 2015], and RGB-D data [Chen et al. 2015a].

In addition, commodity mobile platforms are emerging as a very common solutions both for capture and for exploration of mobile environments. On this specific topics, we refer the audience to two recent tutorials on the subject, which also contain sections devoted to indoor environments [Agus et al. 2017a,b].

4 DETAILED OUTLINE

The course will be organizes in two sessions of 1.5 hours. After providing a general overview of the subject (Session 1.1), we will discuss shape and color sources generated by indoor mapping devices and describe several open datasets available for research purposes (Session 1.2). We will then provide an abstract characterization of the typical structured indoor models, and of the main problems that need to be solved to create such models from imperfect input data, identifying the specialized priors exploited to address significantly challenging imperfections in visual and geometric input (Session 1.3). The various solutions proposed in the literature, and their combination into global reconstruction pipelines will be then analyzed by providing a general overview, pointing out the various solutions proposed in the literature, and discussing their pros and cons. Session 1.4 will be dedicated to room segmentation, while Session 1.5 will cover boundary surface reconstruction from dense 3D data. After a break, we will continue with a presentation of boundary surface reconstruction from images and/or sparse 3D data (Session 2.1), object detection and reconstruction (Session 2.2), final model assembly (Session 2.3), and visual representation generation (Session 2.4). We will finally point out relevant research issues and analyze research trends (Session 2.5).

SESSION 1.1: Opening and introduction

In the introductory session, we will define the topic of structured indoor reconstruction and point out to the many applications of it. We will then provide an outline of the rest of the presentation.

SESSION 1.2: Data capture and representation

Indoor reconstruction starts from measured data obtained by surveying the indoor environment. Many options exist for performing capture, ranging from very low-cost commodity solutions to professional devices and systems. In this session, we first provide a characterization of the various input sources and then provide a link to the main public domain datasets available for research purposes.

Input data sources. Indoor mapping is required for a wide variety of applications, and an enormous range of 3D acquisition devices have been proposed over the last decades. From LiDAR to portable mobile mappers, these sensors gather shape and/or color information in an effective, often domain-specific, way [Lehtola et al. 2017; Xiong et al. 2013]. In addition, many general-purpose commodity solutions, e.g., based on smartphones and cameras, have also been exploited for that purpose [Pintore et al. 2014; Sankar and Seitz 2012]. However, a survey of acquisition methods is out of the scope of this survey. We rather provide a classification in terms of the characteristics of the acquired information that have an impact on the processing pipeline. Our classification will differentiate *Purely visual input sources*, *Purely geometric input sources*, and *Multimodal colorimetric and geometric input sources*.

Open research data. A notable number of freely available datasets containing indoor scenes have been released in recent years for the purposes of benchmarking and/or training learning-based solutions. However, most of them are more focused on scene understanding [University of Zurich 2016] than reconstruction, and often only cover portions of rooms [Cornell University 2012; New York University 2012; Princeton University 2015; Stanford University 2016b; Technical University of Munich 2015; Washington University 2014]. Many of them have been acquired with RGB-D scanners, due to the flexibility and low-cost of this solution (see an established survey [Firman 2016] for a detailed list of them). We will summarize the major open datasets that have been used in general 3D indoor reconstruction research, detailing their characteristics and possible usage. These will include *SUN360 Database* [Massachussets Institute of Technology 2012; Pintore et al. 2018a,b; Xiao et al. 2012; Yang and Zhang 2016; Zhang et al. 2014], *SUN3D Database* [Chang et al. 2017; Choi et al. 2015; Dai et al. 2017c; Princeton University 2013; Xiao et al. 2013], *UZH 3D Dataset* [Mattausch et al. 2014; Mura et al. 2014b, 2016; University of Zurich 2014], *SUNCG Dataset* [Armeni et al. 2017; Chang et al. 2017; Liu et al. 2018b; Princeton University 2016; Song et al. 2017], *BundleFusion Dataset* [Dai et al. 2017c; Fu et al. 2017; Huang et al. 2017; Stanford University 2016a], *ScanNet Data* [Chang et al. 2017; Dai et al. 2017a,b], *Matterport3D Dataset* [Chang et al. 2017; Matterport 2017], *2D-3D-S Dataset* [Armeni et al. 2017; Stanford University 2017], *FloorNet Dataset* [Chen et al. 2019; Liu et al. 2018b,c], *CRS4/ViC Research Datasets* [CRS4 Visual Computing 2018; Pintore et al. 2019a, 2018a,b], *Replica Dataset* [Straub et al. 2019], and *Structured3D Dataset* [Sun et al. 2019; Zheng et al. 2019a].

SESSION 1.3: Targeted structured 3D model

The goal of structured 3D indoor reconstruction is to transform an input source containing a sampling of a real-world interior environment into a compact structured model containing both geometric and visual abstractions. Each distinct input source tends to produce only partial coverage and imperfect sampling, making reconstruction difficult and ambiguous. For this reason, research has concentrated on defining priors in order to combat imperfections and focus reconstruction on very specific expected indoor structures, shapes, and visual representations. In this session, we first characterize the artifacts typical of indoor model measurement, before defining the structure and priors commonly used in structured 3D indoor reconstruction research, and the sub-problems connected to its generation.

Artifacts. In this session, we will introduce the characterization provided by Berger et al. [Berger et al. 2017] for point clouds, which characterized sampled sources according to the properties that have the most impact on reconstruction algorithms, identifying them into *sampling density*, *noise*, *outliers*, *misalignment*, and *missing data*. We will then show how this characterization extends to visual and mixed data. We will then discuss how the artifacts associated with each one of these characteristics have some specific forms for indoor environments.

Reconstruction priors. We will show how, without prior assumptions, the reconstruction problem for indoor environments is ill-posed, since an infinite number of solutions may exist that fit under-sampled or partially missing data. We will discuss how structured indoor reconstruction has focused its efforts on formally or implicitly restricting the target output model, in order to cover a large variety of interesting use-cases while making reconstruction tractable, introducing in particular the separation between permanent structures and movable objects, and the organization of permanent structures into a graph of rooms connected by passages. We will then survey very specific geometric priors for structural recovery that have been introduced in the indoor reconstruction literature, including *floor-wall* [Delage et al. 2006], *cuboid* [Hedau et al. 2009], *Manhattan world* [Coughlan and Yuille 1999], *Atlanta world* (a.k.a. *Augmented Manhattan World*) [Schindler and Dellaert 2004], *Indoor World Model* [Lee et al. 2009], *Vertical Walls* [Pintore et al. 2018a], and *Piece-wise planarity* [Furukawa et al. 2009].

Main problems. Starting from the above definitions, we identify a core set of basic problems that need to be solved to construct the model from observed data, which are then discussed in the following sessions: *room segmentation*, *bounding surfaces reconstruction*, *indoor object detection and reconstruction*, *integrated model computation*, and *visual representation generation*.

SESSION 1.4:
Room segmentation

While a number of early methods focused on reconstructing the bounding surface of the environment as a single entity, without considering the problem of recognizing individual sub-spaces within it, structuring the 3D model of an indoor environment according to its subdivision into different rooms has gradually become a fundamental step in all modern indoor modeling pipelines, regardless of the type of input they consider (e.g. visual vs. 3D data) or of their main intended goal (e.g. virtual exploration vs. as-built BIM) [Ikehata et al. 2015]. In this session we will discuss approaches that segment the *input* before the application of the reconstruction pipeline, as well as approaches that structure the *output* 3D model according to its subdivision into different rooms.

SESSION 1.5:
Bounding surfaces reconstruction - part 1

While room segmentation deals with the problem of decomposing an indoor space into disjoint spaces (e.g., hallways, rooms), the goal of bounding surface reconstruction is to further parse those spaces into the structural elements that bound their geometry (e.g. floor, ceiling, walls, etc.). This task is one of the major challenges in indoor reconstruction, since building interiors are typically cluttered with furniture and other objects. Not only are these elements not relevant to the structural shape of a building, and should therefore considered as outliers for this task, but they also generate viewpoint occlusions resulting in large amounts of missed sampling of the permanent structures. Larger amounts of missed 3D samplings are also present in visual input sources. Thus, generic surface reconstruction approaches are doomed to fail. In this session, we will discuss an array of specific state-of-the-art approaches, focusing primarily on the extraction of walls, ceilings, and floors. Given the complexity of the topic, the session is subdivided in two parts. In this first session, we will introduce the topic and discuss methods for reconstruction *with dense geometric measures*, acquired either by stereo or by direct measurement of depth.

SESSION 2.1:
Bounding surfaces reconstruction - part 2

The second part of the bounding surface reconstruction session will be devoted to techniques that perform reconstruction *without geometric measures as input sources* and *with sparse geometric measures*. As we will see, these technique exploit mostly visual input data (single- and multi-view).

SESSION 2.2:
Object detection and reconstruction

Modeling objects that occur in indoor scenes is a recurrent problem in computer graphics and computer vision research. In this context, the term *object* refers to a part of the environment that is movable (typically, furniture) and thus does not belong to the architectural structure. In this session, we will survey those aspects of indoor object modeling that are integrated in the reconstruction of the entire indoor scene. In particular, we will present approaches where object detection is exploited for clutter removal, methods where 3D indoor objects are approximately reconstructed, and specialized techniques targeting the detection and modeling of flat objects attached to walls and ceilings.

SESSION 2.3:
Integrated model computation

The structured reconstruction of a complex environment requires not only the analysis of isolated structures, permanent or not, but also to ensure their integration into a coherent structured model. In this session, we will first discuss how the boundary models of the different rooms are made geometrically and structurally consistent, ensuring for instance that the separating wall boundaries between adjacent rooms are correctly modeled based on the specific output representation of choice. Secondly, we will show methods that find connections among rooms, so that adjacent rooms are connected by doors or large passages that directly reflect the intended functionality of the environment and that can therefore be integrated in its structured representation in the form of graph edges. Moreover, the structure of a multi-room environment goes beyond the plain geometric description of its rooms and is strongly related to the way such rooms are connected. For this reason, we will also present approaches for the extraction of a graph that encodes the room interconnections in multi-room and multi-floor environments.

SESSION 2.4:
Visual representation generation

The geometric and topological description coming out of the previous steps may not be enough for the applications that should ultimately visualize the reconstructed model. It is therefore necessary to enrich the structured representation with information geared towards visual representation. In this session, we will discuss how generating visual representations translates into two different problems: the improvement of appearance of reconstructed models with additional geometric and visual data, and the generation of structures to support exploration and navigation. We will then discuss techniques to improve the appearance of reconstructed models by refining the color or by refining the geometry. We will finally show how providing support for visualizing/exploring the dataset has especially been tackled in the context of applications that link the structured reconstruction to the original data, and will present current approaches.

SESSION 2.5:
Wrap-up and discussion

In this concluding session, we will summarize the main result coming out of the literature survey and provide examples of applications in which the techniques are exploiting, focusing especially on emerging software-as-a-service approaches. We will then provide a view on open problems and current and future works. We will particularly mention work that exploits less constraining priors, performing data fusion to combine visual and depth cues into multi-modal feature descriptors to help reconstruction, improving reconstruction from visual input from commodity cameras and smartphones, as well as exploiting data-driven priors to learn hidden relations from the available data.

5 TUTORIAL NOTES CONTENTS

At the end of this tutorial, we include a full bibliography, as well as commented slides for all the tutorial sessions.

6 SCHEDULE

Duration	Lecturer	Topic	Sub-topics
10'	Gobbetti	Opening and introduction	Topic definition; Main applications; Course outline
10'	Gobbetti	Data capture and representation	Input data sources; Capture setups; Open research data
15'	Gobbetti	Targeted structured 3D model	Artifacts; Reconstruction priors; Main problems
25'	Mura	Room segmentation	Segmentation of input; Segmentation of output
25'	Pajarola	Bounding surfaces reconstruction - part 1	With dense geometric measures
		BREAK	
25'	Pintore	Bounding surfaces reconstruction - part 2	Without geometric measures as input sources; With sparse geometric measures
20'	Pintore	Indoor object detection and reconstruction	Object detection for clutter removal; 3D indoor objects detection and reconstruction; Flat indoor objects detection and reconstruction
15'	Ganovelli	Integrated model computation	Ensuring consistency of multi-room models; Finding and modeling connections; Multi-room and multi-floor graphs
15'	Ganovelli	Visual representation generation	Geometry refinement; Texture refinement; Visual exploration
15'	Gobbetti	Wrap-up and discussion	Summary of techniques and assessment of capabilities; Open problems; Q&A

7 AUTHORS AND LECTURERS

- **Giovanni Pintore** is a senior research engineer at the Visual Computing (ViC) group at the Center for Advanced Studies, Research, and Development in Sardinia (CRS4). He holds a Laurea (M. Sc.) degree (2002) in Electronics Engineering from the University of Cagliari. His research interests include methods for 3D reconstruction of structured indoor scenes from images, multi-resolution representations of large and complex 3D models, as well as visual computing applications of mobile graphics. He has published a number of works in the field of both interactive and automatic reconstruction of indoor structures and has given several courses in international conferences, such as Eurographics, SIGGRAPH Asia, and 3DV, focusing on mobile capture and metric reconstruction of architectural scenes. He has contributed as key developer and manager in international industrial and research projects in the areas of security, space exploration and smart cities. He served as program chair, editor and reviewer in international conferences and journals.
- **Claudio Mura** is a postdoctoral researcher and lecturer at the Visualization and MultiMedia Lab of the University of Zurich, from which he obtained a Ph.D. in Informatics in 2017 while working as an Early-Stage Researcher in the EU FP7 MSCA-ITN project DIVA. Before that, he received a M.Sc. degree in Computer Science from the University of Cagliari, Italy. His research, for which he has obtained direct funding from several public and private institutions, has been awarded with the Best Student Paper Award at the 2016 Pacific Graphics Conference and the 2nd Best Paper Award at the 2018 Computer Graphics International Conference. He has also collaborated with industry partners in R&D and technology transfer projects. His current research interests include 3D modeling and semantic understanding of interiors, point-based shape analysis and point cloud processing.
- **Lizeth Fuentes** is a doctoral candidate at the Visualization and MultiMedia Lab of the University of Zurich, working as an Early-Stage Researcher in the H2020 MSCA-ITN project EVOCATION. She obtained a B.Sc. degree in Computer Science from the National University of Saint Augustine, Peru, and a M.Sc. degree (2017) in Computer Science from the Federal Fluminense University, Rio de Janeiro, Brazil. Her research interests are geometry processing, computer vision, shape analysis and machine learning.
- **Fabio Ganovelli** is a research scientist at the Istituto di Scienza e Tecnologie dell'Informazione (ISTI) of the National Research Council (CNR) in Pisa, Italy. He received his PhD from University of Pisa in 2001. Since then, he published in the fields of deformable objects, geometry processing, out-of-core rendering and manipulation of massive models, photorealistic rendering, image-to-geometry registration, indoor reconstruction, and education. He is a core developer of the Visualization and Computer Graphics Library and served as reviewer and/or chair for all the main journals and conferences in Computer Graphics.
- **Renato Pajarola** is a full Professor in the Department of Informatics at the University of Zürich (UZH). He received a Dipl. Inf-Ing ETH as well as a Dr. sc. techn. degree in computer science from the Swiss Federal Institute of Technology (ETH) Zurich in 1994 and 1998 respectively. Subsequently he was a post-doctoral researcher and lecturer in the Graphics, Visualization and Usability Center at Georgia Tech. In 1999 he joined the University of California Irvine as an Assistant Professor where he established the Computer Graphics Lab. Since 2005 he has been leading the Visualization and MultiMedia Lab at UZH. He is a Senior Member of ACM and IEEE as well as a Fellow of the Eurographics Association. Dr. Pajarola's research interests include interactive large-scale data visualization, real-time 3D graphics, 3D scanning and reconstruction, geometry processing, as well as remote and parallel rendering. He has published a wide range of internationally peer-reviewed research articles in top journals and conferences. Prof. Pajarola regularly serves on program committees, such as for example the IEEE Visualization Conference, Eurographics, EuroVis Conference, IEEE Pacific Visualization or ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games. He organized and co-chaired the Eurographics Conference in 2015, chaired the 2010 EG Symposium on Parallel Graphics and Visualization and was papers co-chair in 2011, and also of the 2007 and 2008 IEEE/EG Symposium on Point-Based Computer Graphics. His recent co-authored papers received a SPIE Best Paper Award in 2013, a Best Student Paper at the Pacific Graphics Conference and an Honorable Mention Award at the ACM SIGGRAPH Symposium on Visualization both in 2016, as well as a (2nd) Best Paper Award at the Computer Graphics International Conference in 2018.

- **Enrico Gobbetti** is the director of Visual Computing (ViC) and Data-Intensive Computing (DiC) at the Center for Advanced Studies, Research, and Development in Sardinia (CRS4), Italy. He holds an Engineering degree (1989) and a Ph.D. degree (1993) in Computer Science from the Swiss Federal Institute of Technology in Lausanne (EPFL). Prior to joining CRS4, he held research and/or teaching positions at EPFL, University of Maryland, and NASA. His main research interests span many areas of visual and distributed computing, with emphasis on scalable technology for acquisition, storage, processing, distribution, and interactive exploration of complex objects and environments. Systems based on these technologies have been used in as diverse real-world applications as internet geoviewing, scientific data analysis, surgical training, and cultural heritage study and dissemination. Enrico has (co-)authored over 200 papers, eight of which received best paper awards. He regularly serves the scientific community through participation in editorial boards, conference committes, and working groups, as well as through the organization and chairing of conferences. He is a Fellow of Eurographics.

ACKNOWLEDGMENTS

This work has received funding from Sardinian Regional Authorities under projects VIGECLAB, AMAC, and TDM (POR FESR 2014-2020 Action 1.2.2). We also acknowledge the contribution of the European Union's H2020 research and innovation programme under grant agreements 813170 (EVOCATION) and 820434 (ENCORE).

REFERENCES

3DVista. 1999. 3DVista:Professional Virtual Tour software. https://www.3dvista.com.

Antonio Adan and Daniel Huber. 2011. 3D reconstruction of interior wall surfaces under occlusion and clutter. In *Proc. 3DIMPVT*. 275–281.

A. Agarwala, A. Colburn, A. Hertzmann, B. Curless, and M. F. Cohen. 2013. Image-Based Remodeling. *IEEE TVCG* 19, 01 (2013), 56–66.

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

Marco Agus, Enrico Gobbetti, Fabio Marton, Giovanni Pintore, and Pere-Pau Vázquez. 2017b. Mobile Graphics. In *Proc. EUROGRAPHICS Tutorials*, Adrien Bousseau and Diego Gutierrez (Eds.).

Mohamed Aly and Jean-Yves Bouguet. 2012. Street view goes indoors: Automatic pose estimation from uncalibrated unordered spherical panoramas. In *Proc. WACV*. 1–8.

Rareş Ambruş, Sebastian Claici, and Axel Wendt. 2017. Automatic Room Segmentation From Unstructured 3-D Data of Indoor Environments. *IEEE Robotics and Automation Letters* 2, 2 (2017), 749–756.

Abhishek Anand, Hema Swetha Koppula, Thorsten Joachims, and Ashutosh Saxena. 2013. Contextually guided semantic labeling and search for three-dimensional point clouds. *The International Journal of Robotics Research* 32, 1 (2013), 19–34.

Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R. Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 2019. 3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera. In *Proc. ICCV*.

I. Armeni, A. Sax, A. R. Zamir, and S. Savarese. 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints* (Feb. 2017). arXiv:1702.01105

Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 2016. 3D Semantic Parsing of Large-scale Indoor Spaces. In *Proc. CVPR*. 1534–1543.

S. Y. Bao, A. Furlan, L. Fei-Fei, and S. Savarese. 2014. Understanding the 3D layout of a cluttered room from multiple images. In *Proc. IEEE WACV*. 690–697.

Matthew Berger, Andrea Tagliasacchi, Lee M. Seversky, Pierre Alliez, Gaël Guennebaud, Joshua A. Levine, Andrei Sharf, and Claudio T. Silva. 2017. A Survey of Surface Reconstruction from Point Clouds. *Computer Graphics Forum* 36, 1 (2017), 301–329.

Dmytro Bobkov, Martin Kiechle, Sebastian Hilsenbeck, and Eckehard Steinbach. 2017. Room Segmentation in 3D Point Clouds using Anisotropic Potential Fields. In *Proc. ICME*. 727–732.

András Bódis-Szomorú, Hayko Riemenschneider, and Luc Van Gool. 2014. Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In *Proc. CVPR*. 469–476.

Alexandre Boulch, Martin de La Gorce, and Renaud Marlet. 2014. Piecewise-Planar 3D Reconstruction with Edge and Corner Regularization. *Computer Graphics Forum* 33, 5 (2014), 55–64.

Alexandre Boulch, Simon Houllier, Renaud Marlet, and Olivier Tournaire. 2013. Semantizing Complex 3D Scenes using Constrained Attribute Grammars. *Computer Graphics Forum* 32, 5 (2013), 33–42.

Yuri Boykov, Olga Veksler, and Ramin Zabih. 2001. Fast Approximate Energy Minimization via Graph Cuts. *IEEE TPAMI* 23, 11 (November 2001), 1222–1239.

Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. 2008. Segmentation and Recognition Using Structure from Motion Point Clouds. In *Proc, ECCV*, David Forsyth, Philip Torr, and Andrew Zisserman (Eds.). 44–57.

Angela Budroni and Jan Böhm. 2010. Automated 3D Reconstruction of Interiors from Point Clouds. *International Journal of Architectural Computing* 8, 1 (2010), 55–73.

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

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *Proc. 3DV*. 667–676.

Anne-Laure Chauve, Patrick Labatut, and Jean-Philippe Pons. 2010. Robust Piecewise-Planar 3D Reconstruction and Completion from Large-scale Unstructured Point Data. In *Proc. CVPR*. 1261–1268.

Jiacheng Chen, Chen Liu, Jiaye Wu, and Yasutaka Furukawa. 2019. Floor-SP: Inverse CAD for Floorplans by Sequential Room-wise Shortest Path. *Proc. ICCV* (2019).

Kang Chen, Yu-Kun Lai, and Shi-Min Hu. 2015a. 3D indoor scene modeling from RGB-D data: a survey. *Computational Visual Media* 1, 4 (2015), 267–278.

Kang Chen, Yu-Kun Lai, Yu-Xin Wu, Ralph Martin, and Shi-Min Hu. 2014. Automatic Semantic Modeling of Indoor Scenes from Low-quality RGB-D Data Using Contextual Information. *ACM TOG* 33, 6 (Nov. 2014), 208:1–208:12.

Kang Chen, Kun Xu, Yizhou Yu, Tian-Yi Wang, and Shi-Min Hu. 2015b. Magic Decorator: Automatic Material Suggestion for Indoor Digital Scenes. *ACM TOG* 34, 6 (Oct. 2015), 232:1–232:11.

Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. 2015. Robust Reconstruction of Indoor Scenes. In *Proc. CVPR*. 5828–5839.

Marc Christie, Patrick Olivier, and Jean-Marie Normand. 2008. Camera control in computer graphics. *Computer Graphics Forum* 27, 8 (2008), 2197–2218.

Ibrahim Cinaroglu and Yalin Bastanlar. 2016. A direct approach for object detection with catadioptric omnidirectional cameras. *Signal, Image and Video Processing* 10, 2 (2016), 413–420.

Daniel Cohen-Or, Yiorgos L Chrysanthou, Claudio T. Silva, and Frédo Durand. 2003. A survey of visibility for walkthrough applications. *IEEE TVCG* 9, 3 (2003), 412–431.

Cornell University. 2012. Cornell RGBD dataset. http://pr.cs.cornell.edu/sceneunderstanding/data/data.php. [Accessed: 2019-09-25].

James M Coughlan and Alan L Yuille. 1999. Manhattan world: Compass direction from a single image by bayesian inference. In *Proc. ICCV*, Vol. 2. 941–947.

CRS4 Visual Computing. 2018. CRS4 ViC Research Datasets. http://vic.crs4.it/download/datasets/. [Accessed: 2019-09-25].

Y. Cui, Q. Li, B. Yang, W. Xiao, C. Chen, and Z. Dong. 2019. Automatic 3-D Reconstruction of Indoor Environment With Mobile Laser Scanning Point Clouds. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2019), 1–14.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. 2017a. ScanNet Data. http://www.scan-net.org/. [Accessed: 2019-09-25].

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. 2017b. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proc. CVPR*.

Angela Dai, Matthias Niessner, Michael Zollhofer, Shahram Izadi, and Christian Theobalt. 2017c. BundleFusion: Real-Time Globally Consistent 3D Reconstruction Using On-the-Fly Surface Reintegration. *ACM TOG* 36, 4 (2017), 24:1–24:18.

Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Niessner. 2018. ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. 2012. Image melding: Combining inconsistent images using patch-based synthesis. *ACM TOG* 31, 4 (2012), 82–1.

L. Del Pero, J. Bowdish, D. Fried, B. Kermgard, E. Hartley, and K. Barnard. 2012. Bayesian geometric modeling of indoor scenes. In *Proc. CVPR*. 2719–2726.

L. Del Pero, J. Bowdish, B. Kermgard, E. Hartley, and K. Barnard. 2013. Understanding Bayesian Rooms Using Composite 3D Object Models. In *Proc. CVPR*. 153–160.

E. Delage, Honglak Lee, and A. Y. Ng. 2006. A Dynamic Bayesian Network Model for Autonomous 3D Reconstruction from a Single Indoor Image. In *Proc. CVPR*, Vol. 2. 2418–2428.

M. Di Benedetto, F. Ganovelli, M. Balsa Rodriguez, A. Jaspe Villanueva, R. Scopigno, and E. Gobbetti. 2014. ExploreMaps: Efficient Construction and Ubiquitous Exploration of Panoramic View Graphs of Complex 3D Environments. *Computer Graphics Forum* 33, 2 (2014), 459–468.

Youli Ding, Xianwei Zheng, Yan Zhou, Hanjiang Xiong, et al. 2019. Low-Cost and Efficient Indoor 3D Reconstruction through Annotated Hierarchical Structure-from-Motion. *Remote Sensing* 11, 1 (2019), 58.

Herbert Edelsbrunner, Joseph O'Rourke, and Raimund Seidel. 1986. Constructing Arrangements of Lines and Hyperplanes with Applications. *SIAM J. Comput.* 15, 2 (May 1986), 341–363.

ETH Zurich. 2017. ETH3D Dataset. https://www.eth3d.net/datasets. [Accessed: 2019-09-25].

Habib Fathi, Fei Dai, and Manolis Lourakis. 2015. Automated as-built 3D reconstruction of civil infrastructure using computer vision: Achievements, opportunities, and challenges. *Advanced Engineering Informatics* 29, 2 (2015), 149–161.

Michael Firman. 2016. RGBD Datasets: Past, Present and Future. In *Proc. CVPR Workshop on Large Scale 3D Data: Acquisition, Modelling and Analysis*.

Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J Brostow. 2016. Structured prediction of unobserved voxels from a single depth image. In *Proc. CVPR*. 5431–5440.

Matthew Fisher, Manolis Savva, Yangyan Li, Pat Hanrahan, and Matthias Niessner. 2015. Activity-centric Scene Synthesis for Functional 3D Scene Modeling. *ACM TOG* 34, 6 (2015), 170:1–179:13.

Alex Flint, Christopher Mei, David Murray, and Ian Reid. 2010. A Dynamic Programming Approach to Reconstructing Building Interiors. In *Proc ECCV*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.). 394–407.

A. Flint, D. Murray, and I. Reid. 2011. Manhattan scene understanding using monocular, stereo, and 3D features. In *Proc. ICCV*. 2228–2235.

Stephen Friedman, Hanna Pasula, and Dieter Fox. 2007. Voronoi Random Fields: Extracting Topological Structure of Indoor Environments via Place Labeling. In *IJCAI*, Vol. 7. 2109–2114.

Qiang Fu, Xiaowu Chen, Xiaotian Wang, Sijia Wen, Bin Zhou, and Hongbo Fu. 2017. Adaptive Synthesis of Indoor Scenes via Activity-associated Object Relation Graphs. *ACM Trans. Graph.* 36, 6 (Nov. 2017), 201:1–201:13.

Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. 2009. Manhattan-world stereo. In *Proc. CVPR*. 1422–1429.

Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. 2009. Reconstructing building interiors from images. In *Proc. ICCV*. 80–87.

David Gallup, Jan-Michael Frahm, and Marc Pollefeys. 2010. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Proc. CVPR*. 1418–1425.

Christopher Geyer and Kostas Daniilidis. 2000. A Unifying Theory for Central Panoramic Systems and Practical Implications. In *Proc. ECCV*. 445–461.

Enrico Gobbetti. 2019. Creation and Exploration of Reality-based Models. *Computers Graphics Forum* 38, 2 (2019), xvii.

Enrico Gobbetti, Dave Kasik, and Sung-eui Yoon. 2008. Technical strategies for massive model visualization. In *Proc. ACM Symp. on Solid and physical modeling*. 405–415.

Mani Golparvar Fard, Feniosky Pea-Mora, Carlos A. Arboleda, and Sanghyun Lee. 2009. Visualization of construction progress monitoring with 4D simulation model overlaid on time-lapsed photographs. *Journal of Computing in Civil Engineering* 23, 6 (2009), 391–404.

Ruiqi Guo and Derek Hoiem. 2013. Support Surface Prediction in Indoor Scenes. In *Proc. ICCV*. 2144–2151.

A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. 2011. From 3D scene geometry to human workspace. In *Proc. CVPR*. 1961–1968.

A. Handa, T. Whelan, J.B. McDonald, and A.J. Davison. 2014. A Benchmark for RGB-D Visual Odometry, 3D Reconstruction and SLAM. In *Proc. ICRA*.

David Harel and Yehuda Koren. 2001. On Clustering Using Random Walks. In *Proc. FST TCS*. 18–41.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proc. ICCV*. 2961–2969.

V. Hedau, D. Hoiem, and D. Forsyth. 2009. Recovering the spatial layout of cluttered rooms. In *Proc. ICCV*. 1849–1856.

Varsha Hedau, Derek Hoiem, and David Forsyth. 2010. Thinking Inside the Box: Using Appearance Models and Context Based on Room Geometry. In *Proc. ECCV*. 224–237.

V. Hedau, D. Hoiem, and D. Forsyth. 2012. Recovering free space of indoor scenes from a single image. In *Proc. CVPR*. 2807–2814.

Derek Hoiem, Alexei A. Efros, and Martial Hebert. 2007. Recovering Surface Layout from an Image. *International Journal of Computer Vision* 75, 1 (01 Oct 2007), 151–172.

Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. 2016. SceneNN: A Scene Meshes Dataset with aNNotations. In *Proc. 3DV*.

Jingwei Huang, Angela Dai, Leonidas Guibas, and Matthias Niessner. 2017. 3Dlite: Towards Commodity 3D Scanning for Content Creation. *ACM TOG* 36, 6 (2017), 203:1–203:14.

ICL. 2017. ICL-NUIM RGB-D Dataset. https://www.doc.ic.ac.uk/~ahanda/VaFRIC/iclnuim.html. [accessed: 2019-09-24].

Satoshi Ikehata, Hang Yang, and Yasutaka Furukawa. 2015. Structured Indoor Modeling. In *Proc. ICCV*. 1323–1331.

A. Iraqui, Y. Dupuis, R. Boutteau, J. Y. Ertaud, and X. Savatier. 2010. Fusion of Omnidirectional and PTZ Cameras for Face Detection and Tracking. In *Proc. Int. Conf. on Emerging Security Technologies*. 18–23.

Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. 2011. KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *Proc. UIST*. 559–568.

Philipp Jenke, Benjamin Huhle, and Wolfgang Strasser. 2009. Statistical Reconstruction of Indoor Scenes. In *Proc. WSCG*. 17–24.

Hou Ji, Angela Dai, and Matthias Niessner. 2019. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *Proc. CVPR*.

Z. Jia, A. Gallagher, A. Saxena, and T. Chen. 2013. 3D-Based Reasoning with Blocks, Support, and Stability. In *Proc. CVPR*. 1–8.

Adrien Kaiser, Jose Alonso Ybanez Zepeda, and Tamy Boubekeur. 2019. A Survey of Simple Geometric Primitives Detection Methods for Captured 3D Data. *Computer Graphics Forum* 38, 1 (2019), 167–196.

Sungil Kang, Annah Roh, Bodam Nam, and Hyunki Hong. 2011. People detection method using graphics processing units for a mobile robot with an omnidirectional camera. *Optical Engineering* 50 (2011), 50:1–50:9.

Z. Sadeghipour Kermani, Z. Liao, P. Tan, and H. Zhang. 2016. Learning 3D Scene Synthesis from Annotated RGB-D Images. In *Proc. SGP*. 197–206.

Young Min Kim, Niloy J Mitra, Dong-Ming Yan, and Leonidas Guibas. 2012. Acquiring 3D indoor environments with variability and repetition. *ACM TOG* 31, 6 (2012), 138:1–138:10.

Kujiale.com. 2019. Structured3D Data. https://structured3d-dataset.org/. [Accessed: 2019-09-25].

Avanish Kushal, Ben Self, Yasutaka Furukawa, David Gallup, Carlos Hernandez, Brian Curless, and Steven M Seitz. 2012. Photo tours. In *Proc. 3DIMPVT*. 57–64.

K. Lai, L. Bo, and D. Fox. 2014. Unsupervised feature learning for 3D scene labeling. In *Proc. ICRA*. 3050–3057.

David C. Lee, Abhinav Gupta, Martial Hebert, and Takeo Kanade. 2010. Estimating Spatial Layout of Rooms Using Volumetric Reasoning About Objects and Surfaces. In *Proc. NIPS*. 1288–1296.

David C Lee, Martial Hebert, and Takeo Kanade. 2009. Geometric reasoning for single image structure recovery. In *Proc. CVPR*. 2136–2143.

K. Lee, S. Ryu, S. Yeon, H. Cho, C. Jun, J. Kang, H. Choi, J. Hyeon, I. Baek, W. Jung, H. Kim, and N. L. Doh. 2016. Accurate Continuous Sweeping Framework in Indoor Spaces With Backpack Sensor System for Applications to 3-D Mapping. *IEEE Robotics and Automation Letters* 1, 1 (2016), 316–323.

Ville Lehtola, Harri Kaartinen, Andreas Nüchter, Risto Kaijaluoto, Antero Kukko, Paula Litkey, Eija Honkavaara, Tomi Rosnell, Matti Vaaja, Juho-Pekka Virtanen, et al.

2017. Comparison of the selected state-of-the-art 3D indoor scanning and point cloud generation methods. *Remote sensing* 9, 8 (2017), 796.

Yangyan Li, Angela Dai, Leonidas Guibas, and Matthias Niessner. 2015. Database-Assisted Object Retrieval for Real-Time 3D Reconstruction. *Computer Graphics Forum* 34, 2 (May 2015), 435–446.

Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. 2019. Planercnn: 3D Plane Detection and Reconstruction from a Single Image. In *Proc. CVPR*. 4450–4459.

C. Liu, P. Kohli, and Y. Furukawa. 2016. Layered Scene Decomposition via the Occlusion-CRF. In *Proc. CVPR*. 165–173.

Chen Liu, Jiaye Wu, and Yasutaka Furukawa. 2018a. Data. https://github.com/art-programmer/. [Accessed: 2019-09-25].

Chen Liu, Jiaye Wu, and Yasutaka Furukawa. 2018b. FloorNet: A Unified Framework for Floorplan Reconstruction from 3D Scans. In *Proc. ECCV*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). 203–219.

Chen Liu, Jiaye Wu, and Yasutaka Furukawa. 2018c. FloorNet Data. https://github.com/art-programmer/FloorNet. [Accessed: 2018-10-24].

Chen Liu, Jiajun Wu, Pushmeet Kohli, and Yasutaka Furukawa. 2017. Raster-to-vector: Revisiting Floorplan Transformation. In *Proc. ICCV*. 2195–2203.

Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. 2018d. Planenet: Piece-wise Planar Reconstruction from a Single RGB Image. In *Proc. CVPR*. 2579–2588.

Andelo Martinovic and Luc Van Gool. 2013. Bayesian grammar learning for inverse procedural modeling. In *Proc. CVPR*. 201–208.

Eleonora Maset, Federica Arrigoni, and Andrea Fusiello. 2017. Practical and efficient multi-view matching. In *Proc. ICCV*. 4568–4576.

Massachussets Institute of Technology. 2012. SUN360 Database. http://people.csail.mit.edu/jxiao/SUN360/. [Accessed: 2019-09-25].

Oliver Mattausch, Daniele Panozzo, Claudio Mura, Olga Sorkine-Hornung, and Renato Pajarola. 2014. Object Detection and Classification from Large-Scale Cluttered Indoor Scans. *Computer Graphics Forum* 33, 2 (2014), 11–21.

Matterport. 2017. Matterport3D. https://github.com/niessner/Matterport. [Accessed: 2019-09-25].

Kevin Matzen, Michael F. Cohen, Bryce Evans, Johannes Kopf, and Richard Szeliski. 2017. Low-cost 360 Stereo Photography and Video Capture. *ACM TOG* 36, 4 (2017), 148:1–148:12.

Paul Merrell, Eric Schkufza, Zeyang Li, Maneesh Agrawala, and Vladlen Koltun. 2011. Interactive Furniture Layout Using Interior Design Guidelines. *ACM TOG* 30, 4 (July 2011), 87:1–87:10.

Aron Monszpart, Nicolas Mellado, Gabriel J. Brostow, and Niloy J. Mitra. 2015. RAPter: Rebuilding Man-Made Scenes with Regular Arrangements of Planes. *ACM TOG* 34, 4 (2015), 103:1–103:12.

Claudio Mura, Alberto Jaspe Villanueva, Oliver Mattausch, Enrico Gobbetti, and Renato Pajarola. 2014a. Reconstructing Complex Indoor Environments with Arbitrary Walls Orientations. In *Eurographics Posters*.

Claudio Mura, Oliver Mattausch, Alberto Jaspe Villanueva, Enrico Gobbetti, and Renato Pajarola. 2014b. Automatic room detection and reconstruction in cluttered indoor environments with complex room layouts. *Computers & Graphics* 44 (2014), 20–32.

Claudio Mura, Oliver Mattausch, and Renato Pajarola. 2016. Piecewise-planar Reconstruction of Multi-room Interiors with Arbitrary Wall Arrangements. *Computer Graphics Forum* 35, 7 (2016), 179–188.

Claudio Mura and Renato Pajarola. 2017. Exploiting the Room Structure of Buildings for Scalable Architectural Modeling of Interiors. In *ACM SIGGRAPH Posters*. 4:1–4:2.

Srivathsan Murali, Pablo Speciale, Martin R. Oswald, and Marc Pollefeys. 2017. Indoor Scan2BIM: Building Information Models of House Interiors. In *Proc. IROS*. 6126–6133.

Przemyslaw Musialski, Peter Wonka, Daniel G Aliaga, Michael Wimmer, Luc Van Gool, and Werner Purgathofer. 2013. A survey of urban reconstruction. *Computer graphics forum* 32, 6 (2013), 146–177.

Liangliang Nan, Ke Xie, and Andrei Sharf. 2012. A Search-classify Approach for Cluttered Indoor Scene Understanding. *ACM TOG* 31, 6 (2012), 137:1–137:10.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *Proc. ECCV*.

NavVis. 2012. TUMViewer. https://www.navvis.lmt.ei.tum.de/view/. [Accessed: 2019-09-25].

New York University. 2012. NYU-Depth V2. https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html. [Accessed: 2019-09-25].

Sebastian Ochmann, Richard Vock, , and Reinhard Klein. 2019. Automatic Reconstruction of Fully Volumetric 3D Building Models from Oriented Point Clouds. *ISPRS Journal of Photogrammetry and Remote Sensing* 151 (2019), 251 – 262.

Sebastian Ochmann, Richard Vock, Raoul Wessel, and Reinhard Klein. 2016. Automatic Reconstruction of Parametric Building Models from Indoor Point Clouds. *Computers & Graphics* 54 (February 2016), 94–103.

Sebastian Ochmann, Richard Vock, Raoul Wessel, Martin Tamke, and Reinhard Klein. 2014. Automatic generation of structural building descriptions from 3D point cloud scans. In *Proc. GRAPP*. 1–8.

Sven Oesau, Florent Lafarge, and Pierre Alliez. 2014. Indoor Scene Reconstruction using Feature Sensitive Primitive Extraction and Graph-cut. *ISPRS Journal of Photogrammetry and Remote Sensing* 90 (2014), 68–82.

Sven Oesau, Florent Lafarge, and Pierre Alliez. 2016. Planar Shape Detection and Regularization in Tandem. *Computer Graphics Forum* 35, 1 (2016), 203–215.

Viorica Pătrăucean, Iro Armeni, Mohammad Nahangi, Jamie Yeung, Ioannis Brilakis, and Carl Haas. 2015. State of Research in Automatic As-Built Modelling. *Advanced Engineering Informatics* 29, 2 (2015), 162–171.

Mark Pauly, Niloy J. Mitra, Joachim Giesen, Markus Gross, and Leonidas J. Guibas. 2005. Example-based 3D Scan Completion. In *Proc. SGP*. 23:1–23:10.

Giovanni Pintore, Marco Agus, and Enrico Gobbetti. 2014. Interactive mapping of indoor building structures through mobile devices. In *Proc. 3DV*, Vol. 2. 103–110.

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*.

Giovanni Pintore, Fabio Ganovelli, Alberto Jaspe Villanueva, and Enrico Gobbetti. 2019a. Automatic modeling of cluttered floorplans from panoramic images. *Computer Graphics Forum* 38, 7 (2019), 347–358.

Giovanni Pintore, Fabio Ganovelli, Ruggero Pintus, Roberto Scopigno, and Enrico Gobbetti. 2018a. 3D floor plan recovery from overlapping spherical images. *Computational Visual Media* 4, 4 (2018), 367–383.

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, Claudio Mura, Fabio Ganovelli, Lizeth Fuentes-Perez, Renato Pajarola, and Enrico Gobbetti. 2019b. State-of-the-art in Automatic 3D Reconstruction of Structured Indoor Environments. *Computer Graphics Forum* 39, 2 (2019), 667–699.

Giovanni Pintore, Ruggero Pintus, Fabio Ganovelli, Roberto Scopigno, and Enrico Gobbetti. 2018b. Recovering 3D existing-conditions of indoor structures from spherical images. *Computers & Graphics* 77 (2018), 16–29.

Ruggero Pintus, Enrico Gobbetti, Marco Callieri, and Matteo Dellepiane. 2017. *Techniques for seamless color registration and mapping on dense 3D models*. Springer, 355–376.

Princeton University. 2013. SUN3D Database. https://sun3d.cs.princeton.edu/. [Accessed: 2019-09-25].

Princeton University. 2015. SUNRGBD Database. http://3dvision.princeton.edu/projects/2015/SUNrgbd/. [Accessed: 2019-09-25].

Princeton University. 2016. SceneCG Dataset. https://sscnet.cs.princeton.edu/. [Accessed: 2019-09-25].

Andrzej Pronobis, Barbara Caputo, Patric Jensfelt, and Henrik I Christensen. 2010. A realistic benchmark for visual indoor place recognition. *Robotics and autonomous systems* 58, 1 (2010), 81–96.

K. Pulli, H. Abi-Rached, T. Duchamp, L. G. Shapiro, and W. Stuetzle. 1998. Acquisition and visualization of colored 3D objects. In *Proc. Pattern Recognition*, Vol. 1. 11–15.

Reconstruct Inc. 2016. Reconstruct: A Visual Command Center. https://www.reconstructinc.com/.

Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *Proc. CVPR*. 7263–7271.

Kensaku Saitoh, Takashi Machida, Kiyoshi Kiyokawa, and Haruo Takemura. 2006. A 2D-3D integrated interface for mobile robot control using omnidirectional images and 3D geometric models. In *Proc. ACM/IEEE Int. Symp. on Mixed and Augmented Reality*. 173–176.

Victor Sanchez and Avideh Zakhor. 2012. Planar 3D Modeling of Building Interiors from Point Cloud Data. In *Proc. ICIP*. 1777–1780.

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

Scott Satkin, Maheen Rashid, Jason Lin, and Martial Hebert. 2015. 3DNN: 3D Nearest Neighbor. Data-Driven Geometric Scene Understanding Using 3D Models. *International Journal of Computer Vision* 111 (2015), 69–97.

Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Niessner. 2016. PiGraphs: Learning Interaction Snapshots from Observations. *ACM TOG* 35, 4 (2016).

G. Schindler and F. Dellaert. 2004. Atlanta world: an expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Proc. CVPR*, Vol. 1. I–I.

Ruwen Schnabel, Roland Wahl, and Reinhard Klein. 2007. Efficient RANSAC for Point-Cloud Shape Detection. *Computer Graphics Forum* 26, 2 (2007), 214–226.

D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stueckler, and D. Cremers. 2018. The TUM VI Benchmark for Evaluating Visual-Inertial Odometry. In *Proc. IROS*.

A. G. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun. 2013. Box in the Box: Joint 3D Layout and Object Reasoning from Single Images. In *Proc. ICCV*. 353–360.

T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. 2017. A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos. In *Proc. CVPR*. 2538–2547.

Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms.

In *Proc. CVPR*, Vol. 1. 519–528.
Tianjia Shao, Weiwei Xu, Kun Zhou, Jingdong Wang, Dongping Li, and Baining Guo. 2012. An interactive approach to semantic modeling of indoor scenes with an RGBD camera. *ACM TOG* 31, 6 (2012), 136:1–136:10.
Chao-Hui Shen, Hongbo Fu, Kang Chen, and Shi-Min Hu. 2012. Structure recovery by part assembly. *ACM TOG* 31, 6 (2012), 180:1–180:10.
H. Shin, Y. Chon, and H. Cha. 2012. Unsupervised Construction of an Indoor Floor Plan Using a Smartphone. *IEEE TPAMI* 42, 6 (2012), 889–898.
S. N. Sinha, D. Steedly, and R. Szeliski. 2009. Piecewise planar stereo for image-based rendering. In *Proc. ICCV*. 1881–1888.
Sudipta N. Sinha, Drew Steedly, Richard Szeliski, Maneesh Agrawala, and Marc Pollefeys. 2008. Interactive 3D Architectural Modeling from Unordered Photo Collections. *ACM TOG* 27, 5 (Dec. 2008), 159:1–159:10.
Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2008. Modeling the World from Internet Photo Collections. *International Journal of Computer Vision* 80, 2 (2008), 189–210.
S. Song, S. P. Lichtenberg, and J. Xiao. 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 567–576.
Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. 2017. Semantic Scene Completion from a Single Depth Image. *Proc. CVPR* (2017).
Stanford University. 2016a. Bundle Fusion Dataset. https://graphics.stanford.edu/projects/bundlefusion. [Accessed: 2019-09-25].
Stanford University. 2016b. PiGraphs Dataset. https://graphics.stanford.edu/projects/pigraphs/. [Accessed: 2019-09-25].
Stanford University. 2017. BuildingParser Dataset. http://buildingparser.stanford.edu/dataset.html. [Accessed: 2019-09-25].
Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. arXiv:cs.CV/1906.05797
M. Stroila, A. Yalcin, J. Mays, and N. Alwar. 2012. Route Visualization in Indoor Panoramic Imagery with Open Area Maps. In *Proc. ICME Workshops*. 499–504.
StructionSite. 2016. VideoWalk. https://www.structionsite.com/products/videowalk/.
Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. 2015. Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views. In *Proc. ICCV*. 2686–2694.
Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. 2019. HorizonNet: Learning Room Layout With 1D Representation and Pano Stretch Data Augmentation. In *Proc. CVPR*.
Pingbo Tang, Daniel Huber, Burcu Akinci, Robert Lipman, and Alan Lytle. 2010. Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques. *Automation in Construction* 19, 7 (2010), 829–843.
Technical University of Munich. 2015. TUM LSI Dataset. https://hazirbas.com/datasets/tum-lsi/. [Accessed: 2019-09-25].
G. Tsai, Changhai Xu, Jingen Liu, and B. Kuipers. 2011. Real-time indoor scene understanding using Bayesian filtering with motion cues. In *Proc. ICCV*. 121–128.
Shubham Tulsiani, Abhishek Kar, Joao Carreira, and Jitendra Malik. 2016. Learning category-specific deformable 3D models for object reconstruction. *IEEE TPAMI* 39, 4 (2016), 719–731.
E. Turner, P. Cheng, and A. Zakhor. 2015. Fast, Automated, Scalable Generation of Textured 3D Models of Indoor Environments. *IEEE Journal of Selected Topics in Signal Processing* 9, 3 (2015), 409–421.
Eric Turner and Avideh Zakhor. 2012. Watertight as-Built Architectural Floor Plans Generated from Laser Range Data. In *Proc. 3DIMPVT*. 316–323.
Eric Turner and Avideh Zakhor. 2013. Watertight Planar Surface Meshing of Indoor Point-Clouds with Voxel Carving. In *Proc. 3DV*. 41–48.
Eric Turner and Avideh Zakhor. 2014. Floor Plan Generation and Room Labeling of Indoor Environments from Laser Range Data. In *Proc. Int. Conf. on Computer Graphics Theory and Applications*. 22–33.
University of Zurich. 2014. UZH 3D Dataset. https://www.ifi.uzh.ch/en/vmml/research/datasets.html. [Accessed: 2019-09-25].
University of Zurich. 2016. SceneNN Dataset. https://www.ifi.uzh.ch/en/vmml/research/datasets.html. [Accessed: 2019-09-25].
Andrea Vedaldi and Andrew Zisserman. 2018. Object instance recognition. http://www.robots.ox.ac.uk/~vgg/practicals/instance-recognition/index.html. [Accessed: 2018-10-24].
Rebekka Volk, Julian Stengel, and Frank Schultmann. 2014. Building Information Modeling (BIM) for existing buildings — Literature review and future needs. *Automation in Construction* 38 (2014), 109 – 127.
F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers. 2017. Image-based localization using LSTMs for structured feature correlation. In *Proc. ICCV*.
Huayan Wang, Stephen Gould, and Daphne Koller. 2010. Discriminative Learning with Latent Variables for Cluttered Indoor Scene Understanding. In *Proc. ECCV*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.). 435–449.
M. L. Wang and H. Y. Lin. 2009. Object recognition from omnidirectional visual sensing for mobile robot applications. In *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics*. 1941–1946.
Washington University. 2014. Washington RGBD dataset. https://rgbd-dataset.cs.washington.edu/dataset.html. [Accessed: 2019-09-25].
R. T. Whitaker, J. Gregor, and P. F. Chen. 1999. Indoor scene reconstruction from sets of noisy range images. In *Proc. 3-D Digital Imaging and Modeling*. 348–357.
Erik Wijmans and Yasutaka Furukawa. 2017. WUSTL Indoor RGBD Dataset. https://cvpr17.wijmans.xyz/data/. [Accessed: 2019-09-25].
T. Wu, J. Liu, M. Li, R. Chen, and J. Hyyppä. 2018. Automated large scale indoor reconstruction using vehicle survey data. In *Proc. UPINLBS*. 1–5.
Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. 2014. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *Proc. WACV*. 75–82.
J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba. 2012. Recognizing scene viewpoint using panoramic place representation. In *Proc. CVPR*. 2695–2702.
Jianxiong Xiao and Yasutaka Furukawa. 2014. Reconstructing the World's Museums. *International Journal of Computer Vision* 110, 3 (Dec 2014), 243–258.
J. Xiao, A. Owens, and A. Torralba. 2013. SUN3D: A Database of Big Spaces Reconstructed Using SfM and Object Labels. In *2013 IEEE International Conference on Computer Vision*. 1625–1632.
Xuehan Xiong, Antonio Adan, Burcu Akinci, and Daniel Huber. 2013. Automatic creation of semantically rich 3D building models from laser scanner data. *Automation in Construction* 31 (2013), 325–337.
Xuehan Xiong and Daniel Huber. 2010. Using Context to Create Semantic 3D Models of Indoor Environments. In *Proc. BMVC*. BMVA Press, 45.1–45.11.
J. Xu, B. Stenger, T. Kerola, and T. Tung. 2017. Pano2CAD: Room Layout from a Single Panorama Image. In *Proc. WACV*. 354–362.
Kun Xu, Kang Chen, Hongbo Fu, Wei-Lun Sun, and Shi-Min Hu. 2013. Sketch2Scene: Sketch-based Co-retrieval and Co-placement of 3D Models. *ACM TOG* 32, 4 (July 2013), 123:1–123:15.
Kai Xu, Hui Huang, Yifei Shi, Hao Li, Pinxin Long, Jianong Caichen, Wei Sun, and Baoquan Chen. 2015. Autoscanning for Coupled Scene Reconstruction and Proactive Object Analysis. *ACM TOG* 34, 6 (2015), 177:1–177:14.
Bo Yang, Stefano Rosa, Andrew Markham, Niki Trigoni, and Hongkai Wen. 2018b. Dense 3D object reconstruction from a single depth view. *IEEE TPAMI* (2018).
Fan Yang, Gang Zhou, Fei Su, Xinkai Zuo, Lei Tang, Yifan Liang, Haihong Zhu, and Lin Li. 2019c. Automatic Indoor Reconstruction from Point Clouds in Multi-room Environments with Curved Walls. *Sensors* 19, 17 (Sep 2019), 3798.
Fengting Yang and Zihan Zhou. 2018. Recovering 3D Planes from a Single Image via Convolutional Neural Networks. In *Proc. ECCV*. 85–100.
H. Yang and H. Zhang. 2016. Efficient 3D Room Shape Recovery from a Single Panorama. In *Proc. CVPR*. 5422–5430.
Jingyu Yang, Ji Xu, Kun Li, Yu-Kun Lai, Huanjing Yue, Jianzhi Lu, Hao Wu, and Yebin Liu. 2019b. Learning to Reconstruct and Understand Indoor Scenes from Sparse Views. *CoRR* (2019). http://arxiv.org/abs/1906.07892
Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. 2019a. DuLa-Net: A Dual-Projection Network for Estimating Room Layouts from a Single RGB Panorama. In *Proc. CVPR*.
Yang Yang, Shi Jin, Ruiyang Liu, , and Jingyi Yu. 2018a. Automatic 3D Indoor Scene Modeling From Single Panorama. In *Proc. CVPR*. 3926–3934.
Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. 2019. Recurrent MVSNet for High-Resolution Multi-View Stereo Depth Inference. In *Proc. CVPR*.
Edward Zhang, Michael F. Cohen, and Brian Curless. 2016. Emptying, Refurnishing, and Relighting Indoor Spaces. *ACM TOG* 35, 6 (2016), 174:1–174:14.
Jian Zhang, Chen Kan, Alexander G Schwing, and Raquel Urtasun. 2013. Estimating the 3D layout of indoor scenes and its clutter from depth sensors. In *Proc. ICCV*. 1273–1280.
Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. 2015a. Minimum Barrier Salient Object Detection at 80 FPS. In *Proc. ICCV*. 1404–1412.
Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. 2014. PanoContext: A Whole-Room 3D Context Model for Panoramic Scene Understanding. In *Proc. ECCV*. 668–686.
Yizhong Zhang, Weiwei Xu, Yiying Tong, and Kun Zhou. 2015b. Online structure analysis for real-time indoor scene reconstruction. *ACM TOG* 34, 5 (2015), 159:1–159:13.
Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. 2019a. Structured3D: A Large Photo-realistic Dataset for Structured 3D Modeling. arXiv:cs.CV/1908.00222
Liang Zheng, Yi Yang, and Qi Tian. 2017. SIFT meets CNN: A decade survey of instance retrieval. *IEEE TPAMI* 40, 5 (2017), 1224–1244.
Lintao Zheng, Chenyang Zhu, Jiazhao Zhang, Hang Zhao, Hui Huang, Matthias Niessner, and Kai Xu. 2019b. Active Scene Understanding via Online Semantic Reconstruction. *Computer Graphics Forum* 38, 7 (2019), 103–114.

J. Zhu, Y. Guo, and H. Ma. 2018. A Data-Driven Approach for Furniture and Indoor Scene Colorization. *IEEE TVCG* 24, 9 (2018), 2473–2486.
S. Zingg, D. Scaramuzza, S. Weiss, and R. Siegwart. 2010. MAV navigation through indoor corridors using optical flow. In *Proc. IEEE IROS*. 3361–3368.
Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Niessner, Reinhard Klein, and Andreas Kolb. 2018. State of the Art on 3D Reconstruction with RGB-D Cameras. *Computer Graphics Forum* 37, 2 (2018), 625–652.
Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. 2018. LayoutNet: Reconstructing the 3D Room Layout from a Single RGB Image. In *Proc. CVPR*. 2051–2059.
Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. 2019. 3D Manhattan Room Layout Reconstruction from a Single 360 Image. arXiv:cs.CV/1910.04099

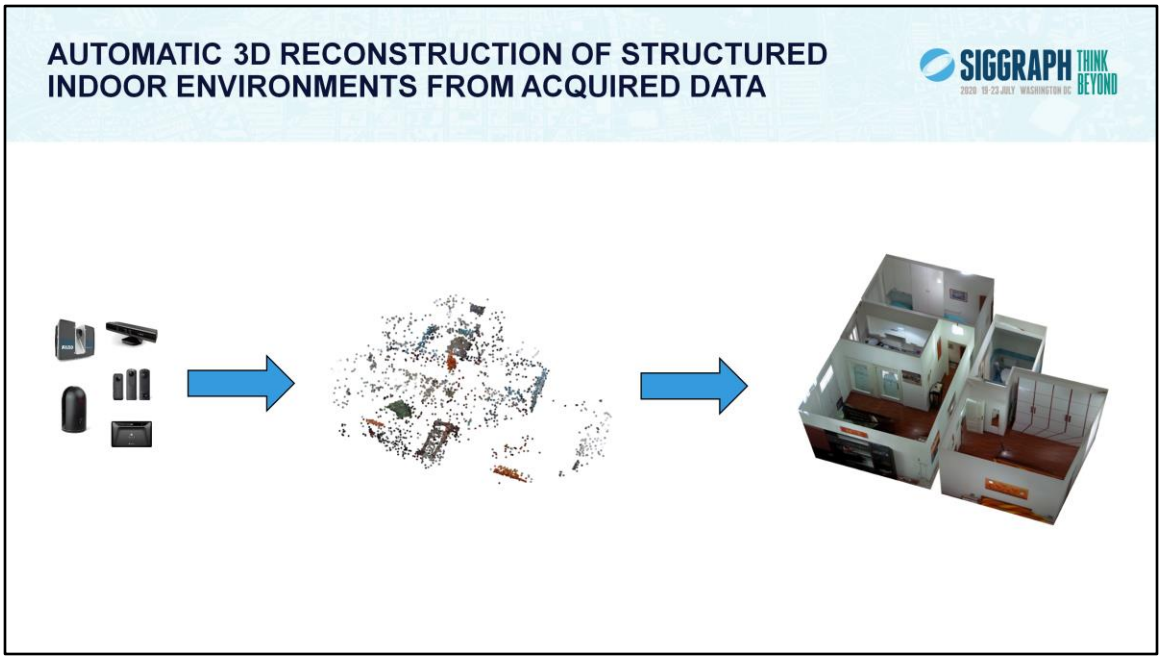
Tutorial slides

Good day everybody, here is Enrico Gobbetti welcoming you to the tutorial on automatic 3D reconstruction of structured indoor environments.

This is a very timely topic, since myself and my co-authors had to prepare and record this work from our respective homes, while enjoying various flavors of lockdown…

Let's start with some bureaucracy: this slide is here just because, in this tutorial, we will sometimes be using some images from previously published works under the "Fair Use" exception…

Now, let's move to the interesting stuff!

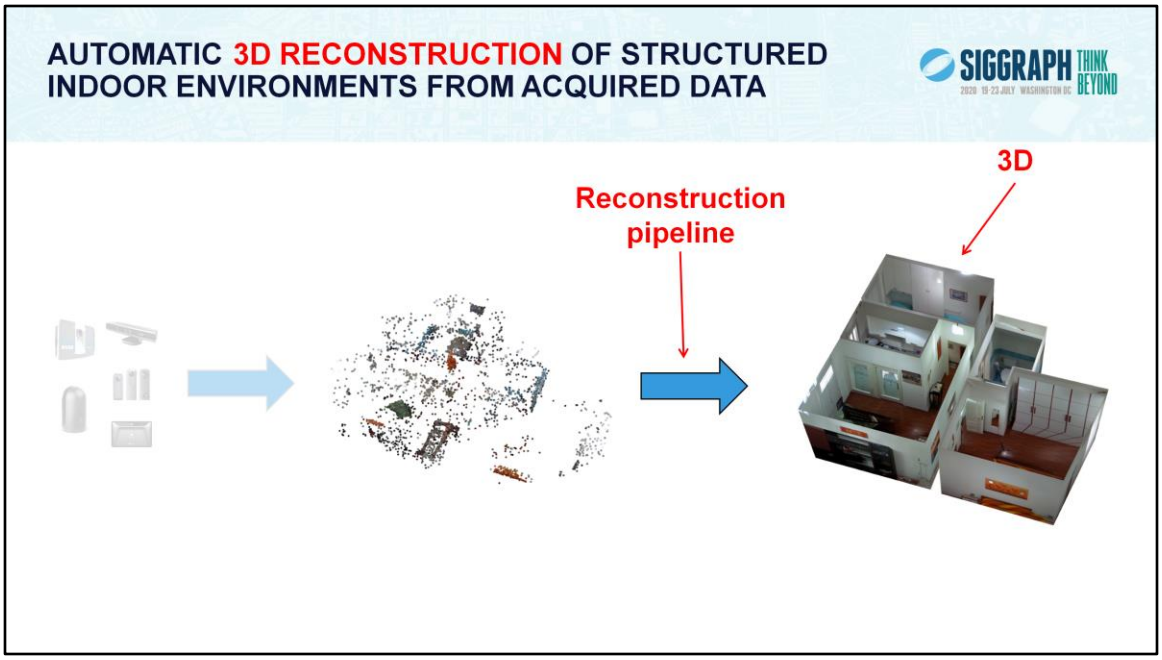


Reconstruction of indoor environments from acquired samples is one of the most rapidly developing sub-fields of 3D reconstruction.

As for all 3D reconstruction processes, devices sample a collection of possibly noisy and sparse measures of the environment, from which a clean application-dependent 3D model must be inferred.

As we will see, indoor environments themselves, as well as target applications, have very peculiar features that make the problem very challenging.

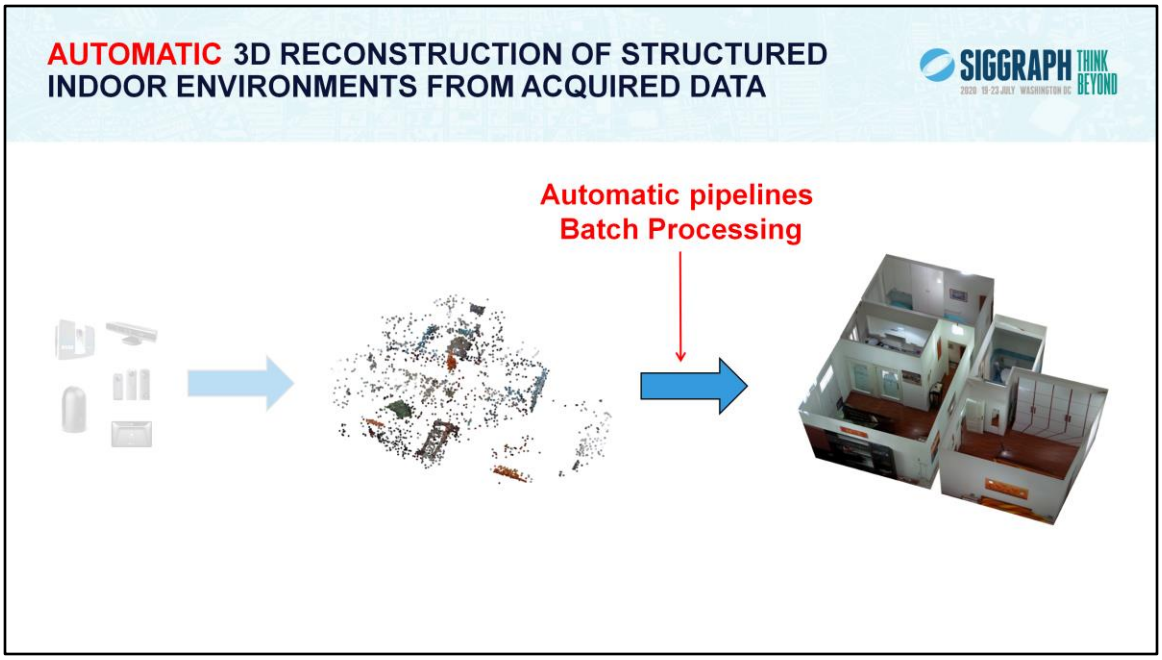
In this tutorial, we strive to provide an up-to-date integrative view of the vast amount of computer graphics and computer vision research in this area.



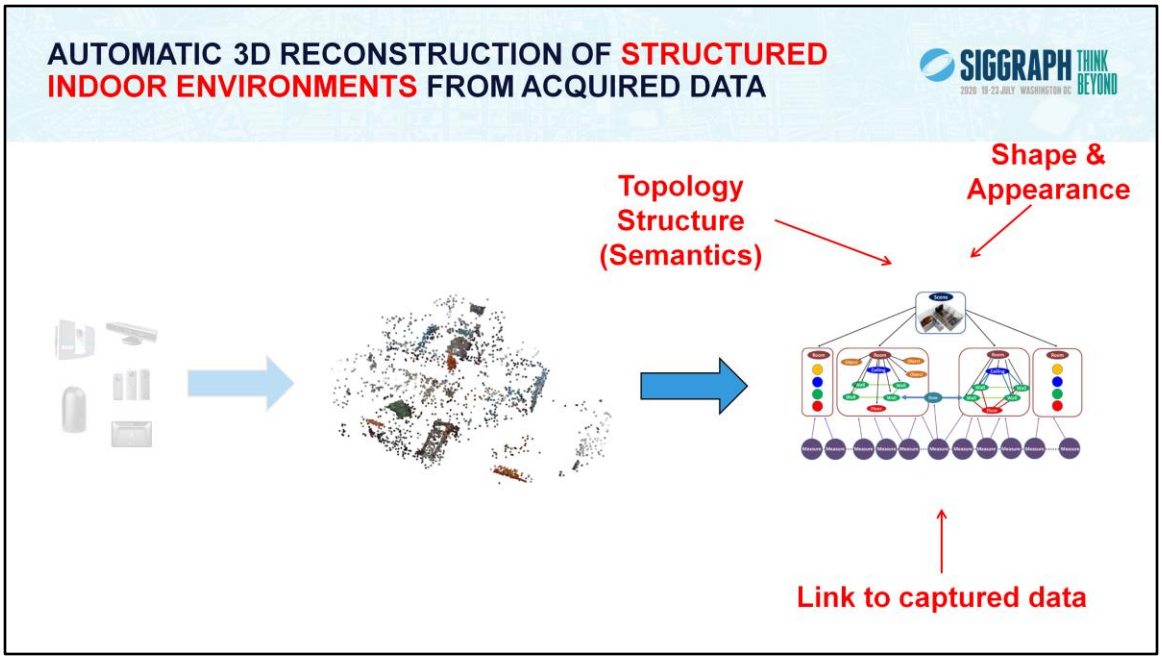
Let's start by defining more precisely our focus.

First of all, we do not specifically discuss the acquisition process, but concentrate on the reconstruction pipeline from the acquired geometric or visual samples.

Second, we mostly target full 3D reconstruction, as opposed to solutions that only strive to generate 2D floor plans.



In terms of pipelines, moreover, we do not target interactive acquisition or modeling, but, rather, processing pipelines that auromatically reconstruct 3D models from samples.



Most of the surveyed solutions, as we will see, work in a batch rather than online fashion, since they need to solve complex global optimization problems to generate their output.

As a matter of fact, as opposed to plain surface reconstruction, the goal is not to replicate dense surface and appearance details, but to discover architectural elements and indoor objects, as well as to organize them in a consistent visual and geometric structure that captures their relations.

THE STAR

- 1. Introduction
- 2. Related surveys
- 3. Background on data capture and representation
- 4. Targeted structured 3D model
- 5. Room segmentation
- 6. Bounding surfaces reconstruction
- 7. Indoor object detection and reconstruction
- 8. Integrated model computation
- 9. Visual representation generation
- 10. Conclusion

EUROGRAPHICS 2020
R. Mantiuk and V. Sundstedt
(Guest Editors)

Volume 39 (2020), Number 2
STAR – State of The Art Report

State-of-the-art in Automatic 3D Reconstruction of Structured Indoor Environments

Giovanni Pintore[1*], Claudio Mura[2*], Fabio Ganovelli[3], Lizeth Fuentes-Perez[2], Renato Pajarola[2], and Enrico Gobbetti[1]

[1] Visual Computing, CRS4, Italy
[2] Visualization and MultiMedia Lab, University of Zurich, Switzerland
[3] Visual Computing Group, ISTI CNR, Italy

Abstract

Creating high-level structured 3D models of real-world indoor scenes from captured data is a fundamental task which has important applications in many fields. Given the complexity and variability of interior environments and the need to cope with noisy and partial captured data, many open research problems remain, despite the substantial progress made in the past decade. In this survey, we provide an up-to-date integrative view of the field, bridging complementary views coming from computer graphics and computer vision. After providing a characterization of input sources, we define the structure of output models and the priors exploited to bridge the gap between imperfect sources and desired output. We then identify and discuss the main components of a structured reconstruction pipeline, and review how they are combined in scalable solutions working at the building level. We finally point out relevant research issues and analyze research trends.

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics;*** *Shape modeling;* ***Computer vision;*** *Computer vision problems; Shape inference; Reconstruction;* • ***Applied computing*** → *Computer-aided design;*

1 Introduction

The automated reconstruction of 3D models from acquired data, be it images or 3D point clouds, has been one of the central topics in computer graphics and computer vision for decades. This field is now thriving, as a result of complementing scientific, technological and market trends. In particular, in recent years, the widespread availability and proliferation of high-fidelity visual/3D sensors (e.g., smartphones, commodity and professional stereo cameras and depth sensors, panoramic cameras, low-cost and high-throughput scanners) has been matched with increasingly cost-effective options for large data processing (e.g., cloud and GPU-accelerated computation), as well as with novel means of visual exploration, from mobile phones to immersive personal displays.

In this context, one of the rapidly emerging sub-fields is concerned with the automatic reconstruction of indoor environments. That is, a 3D representation of an interior scene must be inferred from a collection of measurements that sample its shape and/or appearance, exploiting and/or combining sensing technologies ranging from passive methods, such as single- and multi-view image capturing, to active methods, such as infrared or time-of-flight cameras, optical laser-based range scanners, structured-light scanners, and LiDAR scanners [BTS*17].

Based on the raw data acquired by these devices, many *general* surface reconstruction methods focus on producing accurate and dense 3D models that faithfully replicate even the smallest geometry and appearance details. In this sense, their main goal is to provide the most accurate representation possible of all the surfaces that compose the input scene, disregarding its structure and semantics or possibly only exploiting them to maximize the fidelity of the output surface model. A number of more *specialized* indoor reconstruction solutions focus, instead, on abstracting simplified high-level structured models that optimize certain application-dependent characteristics [IYF15].

The focus on high-level structured models is motivated by several reasons. First of all, their availability is necessary in many fields. For example, applications such as the generation or revision of building information models (BIM) require, at least, the determination of the bare architectural structure [MMJV*14, TCZ15]. On the other hand, information on the interior clutter, in terms of 3D footprint of major indoor objects, is necessary in many other use cases, such as guidance, energy management, security, evacuation planning, location awareness or routing [IYF15]. Even when the goal is solely for visualization, structured simplified models need to be extracted as a fundamental component of a renderable model. This is because

* Joint first authors

We have organized the massive amount of research done on this topic in a large survey, that we have recently published in Computer Graphics Forum, and presented as a state-of-the-art report at Eurographics 2020.

In this presentation, we are going to summarize our main findings in a tutorial form, providing not only a coverage of related work but a full introduction to the subject.

We encourage you, anyway, to refer to the article text for much more details and a very detailed bibliography.

This survey is a collaboration of six authors from three institutions distributed across Europe and funded by several collaborative projects, also mentioned in this slide.

Like myself, my colleagues also prepared this presentation while enjoying various flavors of lockdown in their respective indoor environment.

SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC

GIOVANNI PINTORE

SENIOR RESEARCHER

CRS4, Italy

Giovanni is a senior research engineer at the CRS4 research center in Italy. He has published a number of works in the field of indoor reconstruction and given courses on the same topic at SIGGRAPH Asia, Eurographics and 3DV. His recent research interests include methods for 3D reconstruction of structured indoor scenes from images, multi-resolution representations of large and complex 3D models and visual computing applications of mobile graphics.

10

All the authors collaborated on the preparation of this work, and will be presenting various parts of this work. Let's start with a very short introduction.

Giovanni Pintore is a senior research engineer at the CRS4 research center in Italy, in the group under my direction, and is a well-known expert in mobile graphics and 3D reconstruction from images. In this tutorial, he will be presenting the sessions on bounding surfaces reconstruction from images and on indoor object detection and reconstruction.

SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC

CLAUDIO MURA

POST-DOCTORAL RESEARCHER

University of Zurich, Switzerland

Claudio is a postdoctoral researcher at the University of Zurich, Switzerland. He holds a Ph.D. in Informatics from the same university and a M.Sc. as well as a B.Sc. degree in Computer Science from the University of Cagliari, Italy. His research work, for which he has obtained direct funding from both public and private institutions, focuses on 3D modeling and understanding of interiors, point-based shape analysis and point cloud processing.

11

Claudio Mura is a Post-Doctoral researcher at the University of Zurich, in Switzerland, whose research focus is on 3D modeling and understanding of interiors, point-based shape analysis and point cloud processing.

In this tutorial, he will be presenting the session on Room Segmentation.

SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC

FABIO GANOVELLI

SENIOR RESEARCHER

ISTI-CNR, Italy

Fabio is a senior research scientist at the National Research Council of Italy (CNR). He holds a PhD in Computer Science from the University of Pisa (2001).

He worked in several areas of Computer Graphics ranging from modeling of soft objects, geometry processing, real time rendering of massive datasets and acquisition of shape and appearance.

Fabio Ganovelli is a Senior Researcher at the National Research Council in Italy, who has contributions in several areas of shape and appearance capture, processing, and display.

In this tutorial, he will be presenting the sessions on Integrated model computation and Visual representation generation.

SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC

LIZETH FUENTES-PEREZ

EARLY-STAGE RESEARCHER

University of Zurich, Switzerland

Lizeth is a doctoral candidate at the Visualization and MultiMedia Lab of the University of Zurich, working as an Early-Stage Researcher in the H2020 MSCA-ITN project EVOCATION. She obtained a B.Sc. degree in Computer Science from the National University of Saint Augustine, Peru, and a M.Sc. degree (2017) in Computer Science from the Federal Fluminense University, Rio de Janeiro, Brazil.

Her research interests are geometry processing, computer vision, shape analysis and machine learning.

13

Lizeth is a doctoral candidate at the Visualization and MultiMedia Lab of the University of Zurich, and an Early-Stage Researcher in the European project EVOCATION that involves all the groups represented here.

She will not be directly presenting a session, but contributed to the preparation of the course material as well as of the STAR on which it is based.

Renato Pajarola is full professor at the University of Zurich Switzerland. His research interests include interactive large-scale data visualization, real-time 3D graphics, 3D scanning and reconstruction, geometry processing, as well as remote and parallel rendering.

In this tutorial, he will be presenting the sessions on Bounding Surfaces Reconstruction from Point Clouds.

SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC

ENRICO GOBBETTI

DIRECTOR OF VISUAL COMPUTING

CRS4, Italy

Enrico Is a research director at the CRS4 research center in Italy. He holds an Engineering degree (1989) and a Ph.D. degree (1993) in Computer Science from the Swiss Federal Institute of Technology in Lausanne (EPFL). His research spans many areas of visual computing and is widely published in major journals and conferences. The primary focus is the development of technology for acquisition, storage, processing, distribution, and interactive exploration of complex objects and environments.

15

Finally, myself, Enrico Gobbetti. I am a research director at the CRS4 research center in Italy, where my group focus on the study, development, and application of technology for acquisition, storage, processing, distribution, and interactive exploration of complex objects and environments.

In this tutorial, in addition to the opening and closing sessions, I will provide background on data capture, model representations, artifacts, priors, and pipeline structures.

THIS TUTORIAL

- 1. Introduction
- 2. Related surveys
- 3. Background on data capture and representation
- 4. Targeted structured 3D model
- 5. Room segmentation
- 6. Bounding surfaces reconstruction
- 7. Indoor object detection and reconstruction
- 8. Integrated model computation
- 9. Visual representation generation
- 10. Conclusion

EUROGRAPHICS 2020
R. Mantiuk and V. Sundstedt
(Guest Editors)

Volume 39 (2020), Number 2
STAR – State of The Art Report

State-of-the-art in Automatic 3D Reconstruction of Structured Indoor Environments

Giovanni Pintore[1*], Claudio Mura[2*], Fabio Ganovelli[3], Lizeth Fuentes-Perez[2], Renato Pajarola[2], and Enrico Gobbetti[1]

[1] Visual Computing, CRS4, Italy
[2] Visualization and MultiMedia Lab, University of Zurich, Switzerland
[3] Visual Computing Group, ISTI CNR, Italy

Abstract

Creating high-level structured 3D models of real-world indoor scenes from captured data is a fundamental task which has important applications in many fields. Given the complexity and variability of interior environments and the need to cope with noisy and partial captured data, many open research problems remain, despite the substantial progress made in the past decade. In this survey, we provide an up-to-date integrative view of the field, bridging complementary views coming from computer graphics and computer vision. After providing a characterization of input sources, we define the structure of output models and the priors exploited to bridge the gap between imperfect sources and desired output. We then identify and discuss the main components of a structured reconstruction pipeline, and review how they are combined in scalable solutions working at the building level. We finally point out relevant research issues and analyze research trends.

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics;*** *Shape modeling;* ***Computer vision;*** *Computer vision problems; Shape inference; Reconstruction;* • ***Applied computing*** → *Computer-aided design;*

1 Introduction

The automated reconstruction of 3D models from acquired data, be it images or 3D point clouds, has been one of the central topics in computer graphics and computer vision for decades. This field is now thriving, as a result of complementing scientific, technological and market trends. In particular, in recent years, the widespread availability and proliferation of high-fidelity visual/3D sensors (e.g., smartphones, commodity and professional stereo cameras and depth sensors, panoramic cameras, low-cost and high-throughput scanners) has been matched with increasingly cost-effective options for large data processing (e.g., cloud and GPU-accelerated computation), as well as with novel means of visual exploration, from mobile phones to immersive personal displays.

In this context, one of the rapidly emerging sub-fields is concerned with the automatic reconstruction of indoor environments. That is, a 3D representation of an interior scene must be inferred from a collection of measurements that sample its shape and/or appearance, exploiting and/or combining sensing technologies ranging from passive methods, such as single- and multi-view image capturing, to active methods, such as infrared or time-of-flight cameras, optical laser-based range scanners, structured-light scanners, and LiDAR scanners [BTS*17].

Based on the raw data acquired by these devices, many *general* surface reconstruction methods focus on producing accurate and dense 3D models that faithfully replicate even the smallest geometry and appearance details. In this sense, their main goal is to provide the most accurate representation possible of all the surfaces that compose the input scene, disregarding its structure and semantics or possibly only exploiting them to maximize the fidelity of the output surface model. A number of more *specialized* indoor reconstruction solutions focus, instead, on abstracting simplified high-level structured models that optimize certain application-dependent characteristics [IYF15].

The focus on high-level structured models is motivated by several reasons. First of all, their availability is necessary in many fields. For example, applications such as the generation or revision of building information models (BIM) require, at least, the determination of the bare architectural structure [MMJV*14, TCZ15]. On the other hand, information on the interior clutter, in terms of 3D footprint of major indoor objects, is necessary in many other use cases, such as guidance, energy management, security, evacuation planning, location awareness or routing [IYF15]. Even when the goal is solely for visualization, structured simplified models need to be extracted as a fundamental component of a renderable model. This is because

* Joint first authors

So, let’s start with the technical presentations….

We will loosely follow the organization of our mentioned state-of-the-art report….

THIS TUTORIAL

- 1. Introduction
- 2. Related surveys
- 3. Background on data capture and representation
- 4. Targeted structured 3D model
- 5. Room segmentation
- 6. Bounding surfaces reconstruction
- 7. Indoor object detection and reconstruction
- 8. Integrated model computation
- 9. Visual representation generation
- 10. Conclusion

EUROGRAPHICS 2020
R. Mantiuk and V. Sundstedt
(Guest Editors)

Volume 39 (2020), Number 2
STAR – State of The Art Report

State-of-the-art in Automatic 3D Reconstruction of Structured Indoor Environments

Giovanni Pintore[1*], Claudio Mura[2*], Fabio Ganovelli[3], Lizeth Fuentes-Perez[2], Renato Pajarola[2], and Enrico Gobbetti[1]

[1] Visual Computing, CRS4, Italy
[2] Visualization and MultiMedia Lab, University of Zurich, Switzerland
[3] Visual Computing Group, ISTI CNR, Italy

Abstract

Creating high-level structured 3D models of real-world indoor scenes from captured data is a fundamental task which has important applications in many fields. Given the complexity and variability of interior environments and the need to cope with noisy and partial captured data, many open research problems remain, despite the substantial progress made in the past decade. In this survey, we provide an up-to-date integrative view of the field, bridging complementary views coming from computer graphics and computer vision. After providing a characterization of input sources, we define the structure of output models and the priors exploited to bridge the gap between imperfect sources and desired output. We then identify and discuss the main components of a structured reconstruction pipeline, and review how they are combined in scalable solutions working at the building level. We finally point out relevant research issues and analyze research trends.

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics;*** *Shape modeling;* ***Computer vision;*** *Computer vision problems; Shape inference; Reconstruction;* • ***Applied computing*** → *Computer-aided design;*

1 Introduction

The automated reconstruction of 3D models from acquired data, be it images or 3D point clouds, has been one of the central topics in computer graphics and computer vision for decades. This field is now thriving, as a result of complementing scientific, technological and market trends. In particular, in recent years, the widespread availability and proliferation of high-fidelity visual/3D sensors (e.g., smartphones, commodity and professional stereo cameras and depth sensors, panoramic cameras, low-cost and high-throughput scanners) has been matched with increasingly cost-effective options for large data processing (e.g., cloud and GPU-accelerated computation), as well as with novel means of visual exploration, from mobile phones to immersive personal displays.

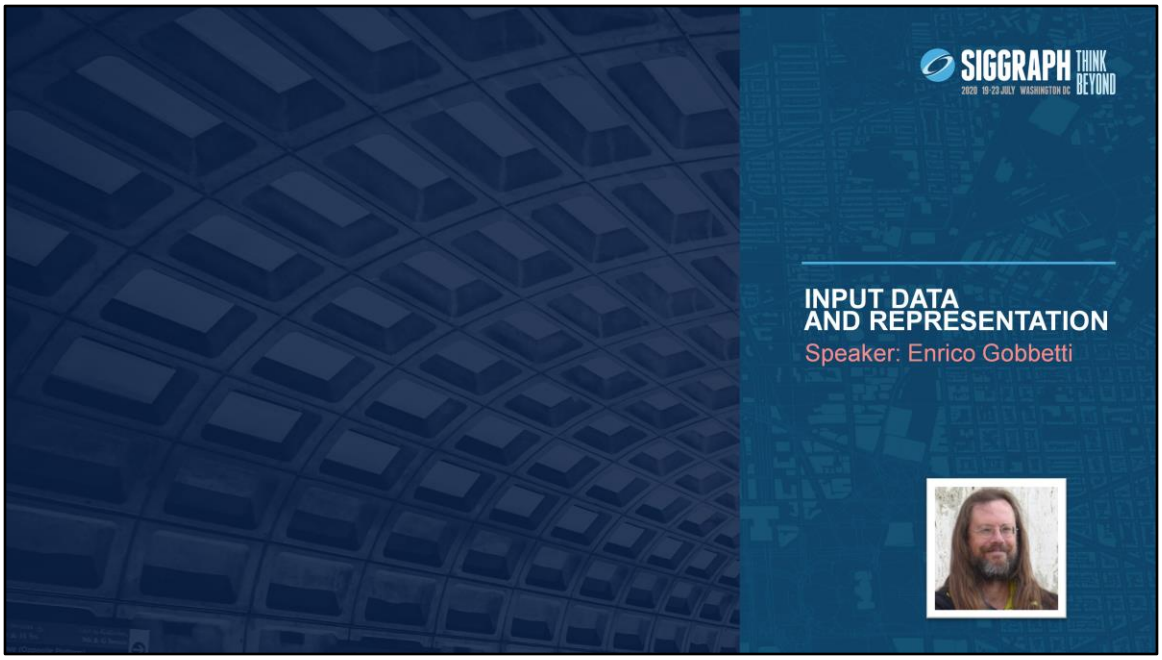
In this context, one of the rapidly emerging sub-fields is concerned with the automatic reconstruction of indoor environments. That is, a 3D representation of an interior scene must be inferred from a collection of measurements that sample its shape and/or appearance, exploiting and/or combining sensing technologies ranging from passive methods, such as single- and multi-view image capturing, to active methods, such as infrared or time-of-flight cameras, optical laser-based range scanners, structured-light scanners, and LiDAR scanners [BTS*17].

Based on the raw data acquired by these devices, many *general* surface reconstruction methods focus on producing accurate and dense 3D models that faithfully replicate even the smallest geometry and appearance details. In this sense, their main goal is to provide the most accurate representation possible of all the surfaces that compose the input scene, disregarding its structure and semantics or possibly only exploiting them to maximize the fidelity of the output surface model. A number of more *specialized* indoor reconstruction solutions focus, instead, on abstracting simplified high-level structured models that optimize certain application-dependent characteristics [IYF15].

The focus on high-level structured models is motivated by several reasons. First of all, their availability is necessary in many fields. For example, applications such as the generation or revision of building information models (BIM) require, at least, the determination of the bare architectural structure [MMJV*14, TCZ15]. On the other hand, information on the interior clutter, in terms of 3D footprint of major indoor objects, is necessary in many other use cases, such as guidance, energy management, security, evacuation planning, location awareness or routing [IYF15]. Even when the goal is solely for visualization, structured simplified models need to be extracted as a fundamental component of a renderable model. This is because

* Joint first authors

Starting by providing background on data capture and representation…


SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC
INPUT DATA
AND REPRESENTATION
Speaker: Enrico Gobbetti

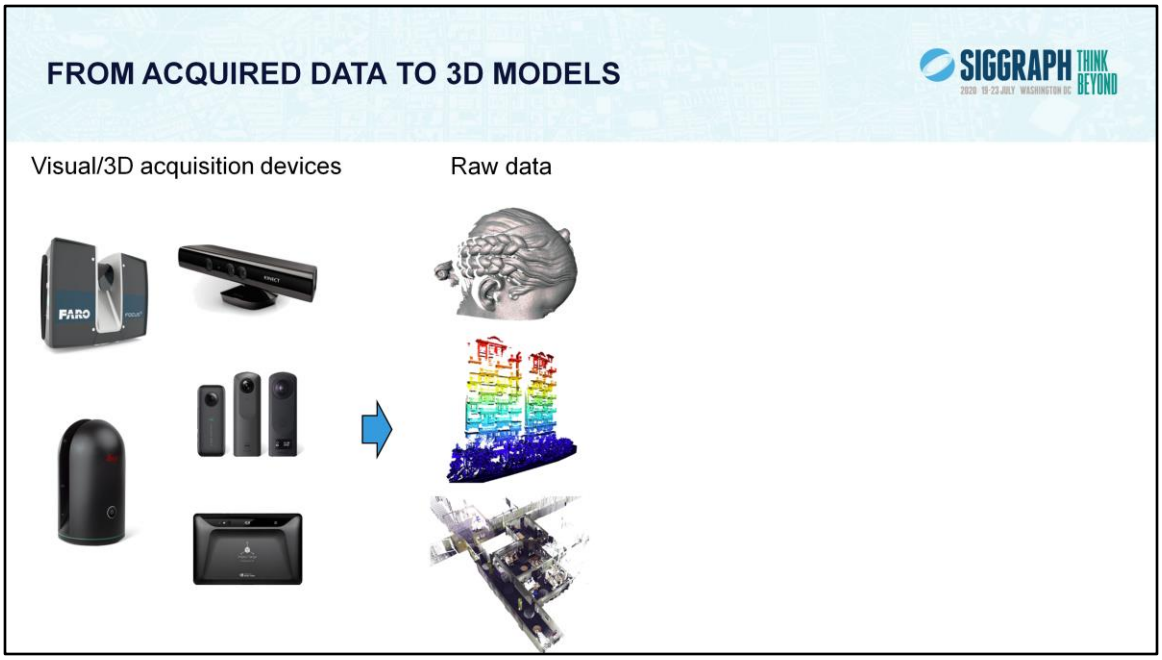
FROM ACQUIRED DATA TO 3D MODELS

SIGGRAPH THINK BEYOND 2020 19-23 JULY WASHINGTON DC

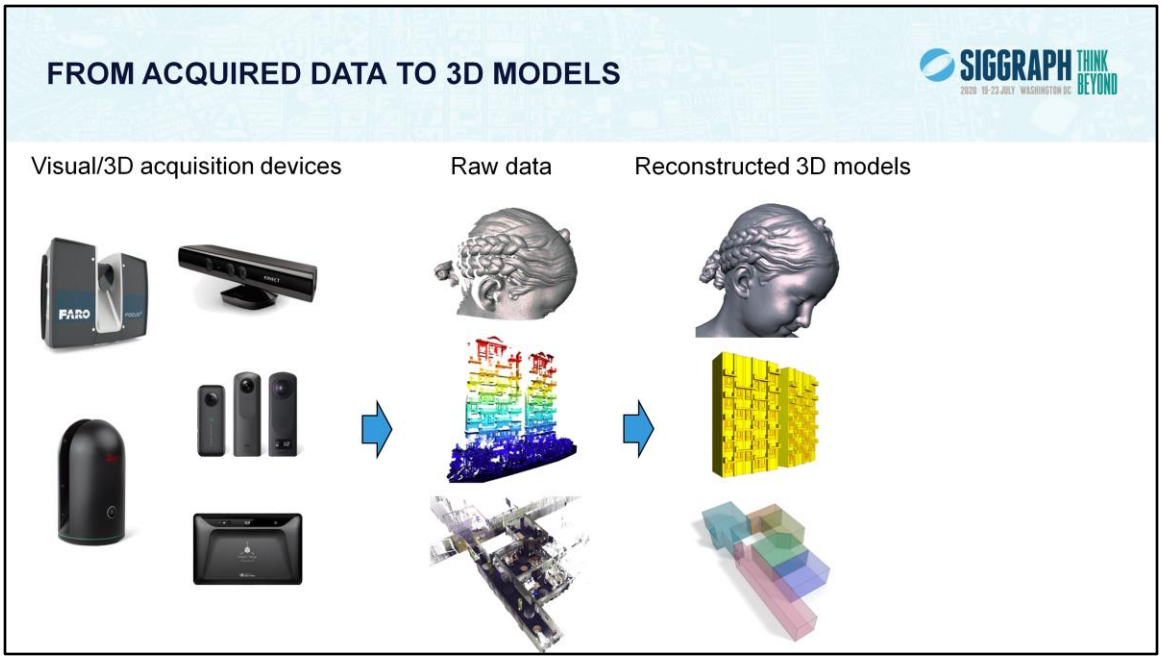
Visual/3D acquisition devices

We have all seen in recent years an unprecedented proliferation of devices to capture the visual appearance and the 3D shape of objects and environment.
These devices vary greatly in terms of cost as well as quality of the capture data, ranging from high-quality yet expensive 3D laser scanners to low-cost panoramic cameras.

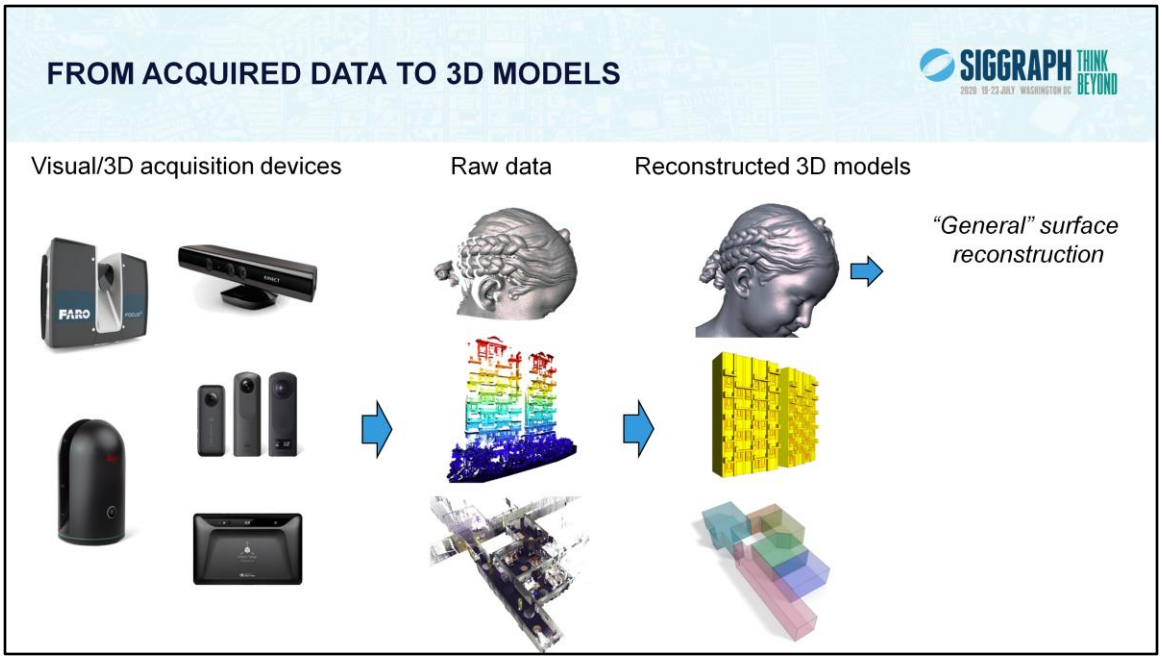
FROM ACQUIRED DATA TO 3D MODELS



Despite their differences, they all generally allow to generate large amounts of digital measurements of real-world entities with unprecedented ease. Be it images of 3D point clouds, these raw data are typically noisy, unorganized, highly redundant and often massive in size.

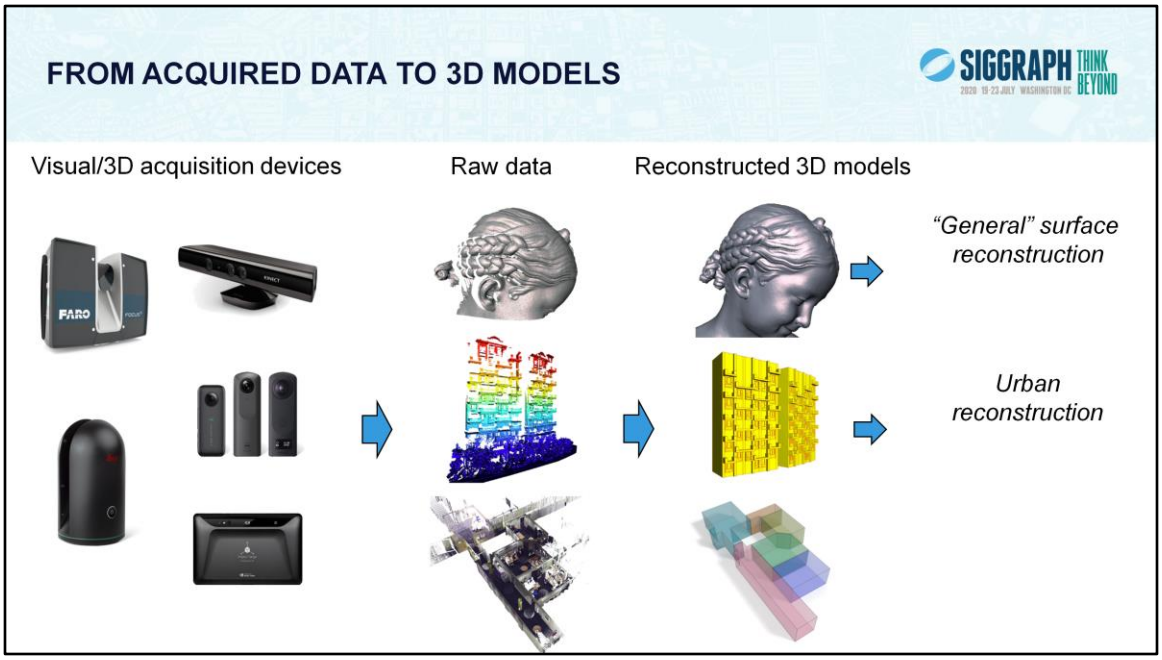


Converting them into more compact and manageable 3D models in an automatic manner has been a central topic in computer graphics and computer vision, and is today more than ever of paramount relevance.



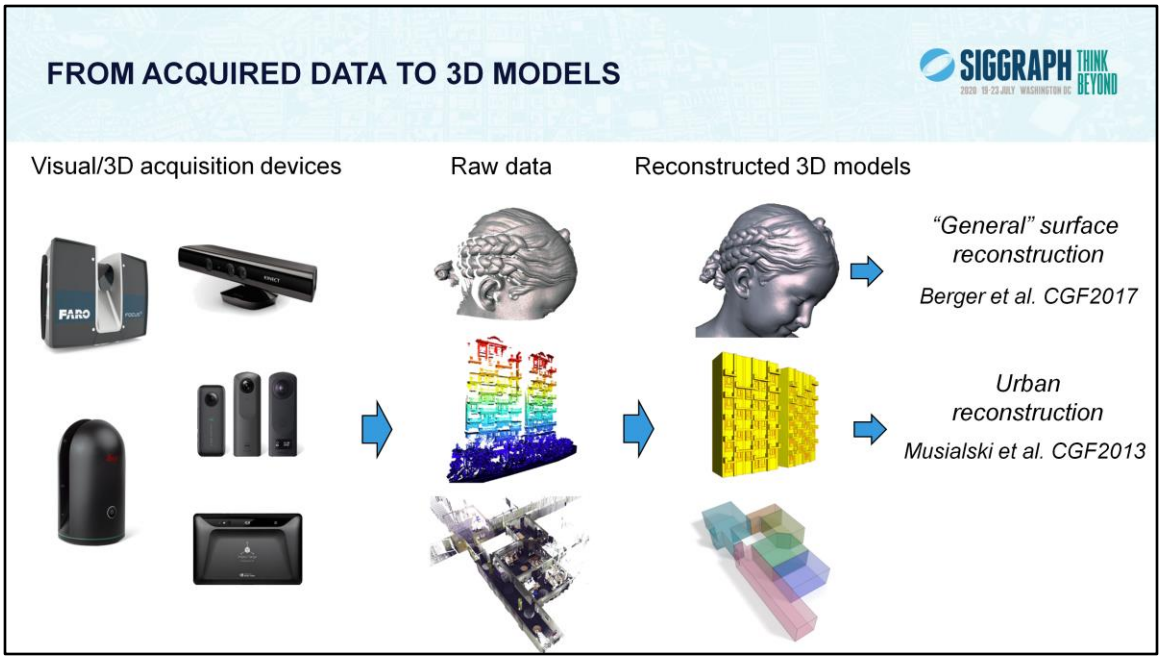
In this context, a number of methods have been proposed over the years that target *general* surface reconstruction: their goal is to produce 3D models that replicate with the highest accuracy possible the geometric and appearance details of the entities represented – but they do so disregarding their semantics and structures.

There are indeed more specialized approaches that focus on specific domains and include specific priors in the reconstruction process, which also allows the contextual inclusion of semantic information in the reconstructed models.



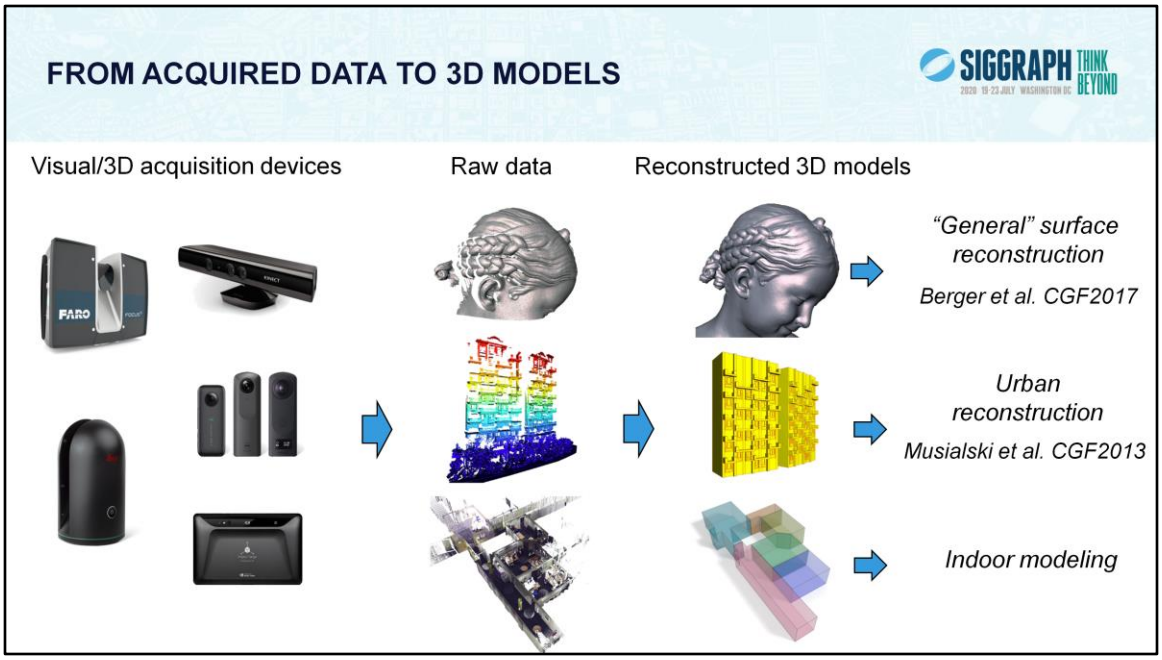
A prime example of this is given by the topic of urban reconstruction.

In fact, both general surface reconstruction and the more specific urban reconstruction are nowadays well-established topics …

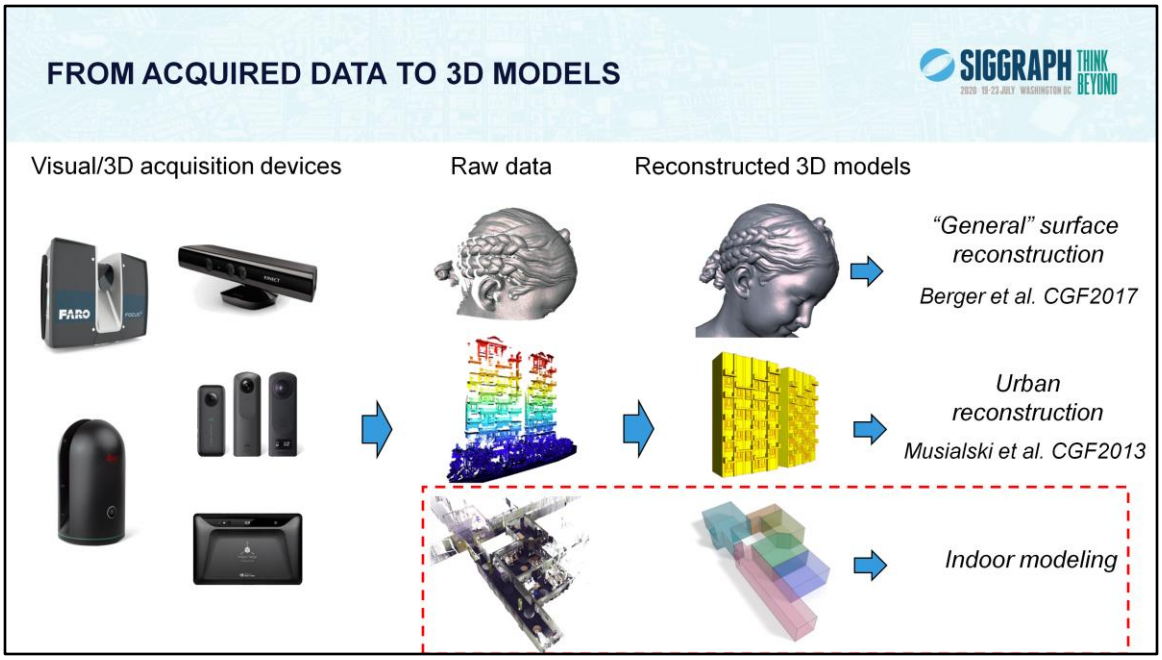


… which have been recently surveyed in a complete and exhaustive manner. We refer you to these extensive surveys by Berget et al. and Musialski et al. for a coverage of those topics.

For other domains, the development of specialized reconstruction approaches has only recently been tackled in a systematic way – and among these domains …



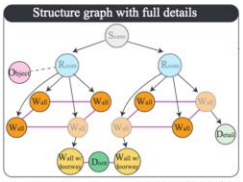
… that of interior scenes …



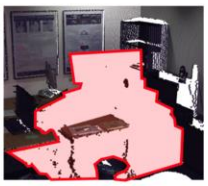
… is one of the most prominent ones, with countless application scenarios in different sectors of human activity.

WHY SPECIALIZED SOLUTIONS FOR INTERIORS?

- Strong need for *structured indoor models*
  - High-level representation of main elements and their relations
  - Optimized to meet requirements of specific fields of application
    - Building Information Models (AEC domain): bare architectural structure
    - Emergency management, location awareness, routing: also interior clutter
  - Standard surface reconstruction does not guarantee this
- Deal with specific challenges of input data
  - Technological limitations of acquisition devices
  - Artifacts caused by properties of real-world interiors
    - Clutter, unreachable areas
    - Transparent/reflective + textureless surfaces



Ikehata et al. ICCV2015



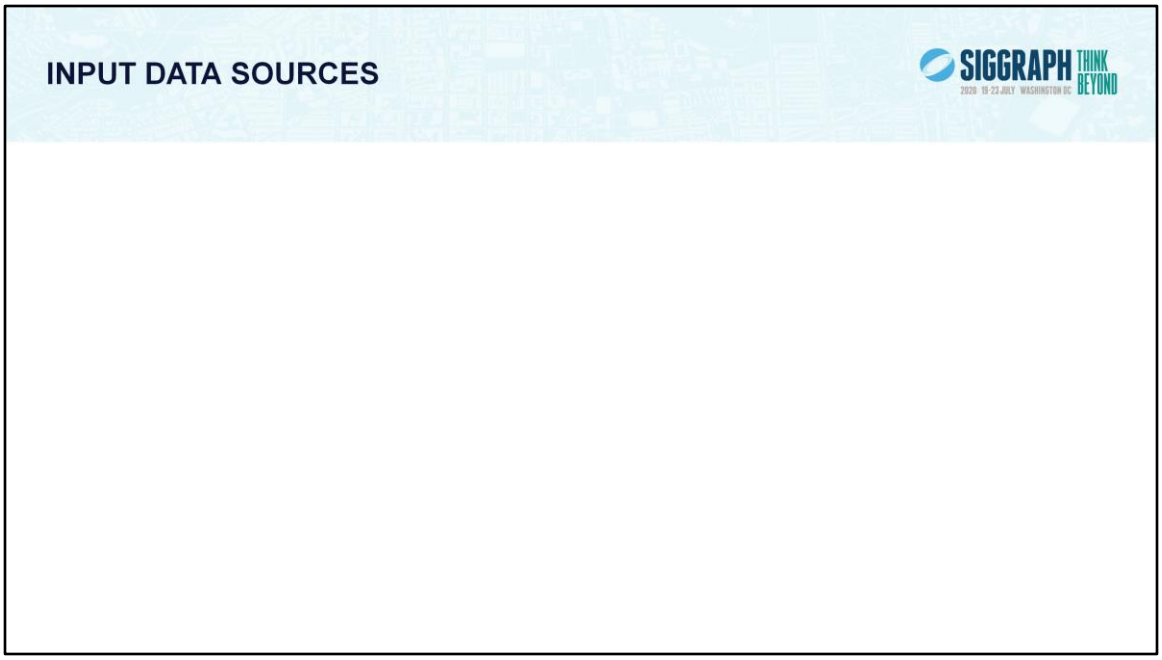
An obvious question at this point is: why do we need specialized techniques for interior scenes?

It turns out that in many specific fields of application there is a strong need for structured indoor models, high-level representations that abstract the low-level measurements into the main scene elements and their relations. Such representations are optimized to meet specific needs: for instance, in the Architecture, Engineering and Construction domain it is often required to generate and update Building Information Models that are largely focused on the architectural structure of the environment. On the other hand, applications like emergency management and indoor navigation additionally require a description of the cluttering elements inside these scenes. All these needs are not guaranteed by standard surface reconstruction.

On top of this, specialized solutions are also needed to deal with the specific problems that are typical of raw input representations of interiors. These are partly linked to the technological limitations of the devices used for indoor capture, but more importantly by the properties of the scene themselves: interior environments are often cluttered and have many unreachable areas; moreover, they contain

transparent as well as textureless surfaces, which give rise to a number of specific artifacts, as we will discuss shortly.

INPUT DATA SOURCES



In fact, the specific type of input data is a factor that influences many aspects of the indoor modeling process. We can distinguish three main types of data sources that can be used to survey an indoor environment:

Purely visual data…

INPUT DATA SOURCES

SIGGRAPH THINK BEYOND 2020 19-23 JULY WASHINGTON DC

1. Purely visual


Single still image


Full-view panorama


Registered images/ panoramas

… – in the form of plain RGB images – are probably the most ubiquitous type of input data, since they can be generated with low-cost, easy-to-operate cameras. Besides individual still image, which have a limited field of view, 360 full-view panoramic images are popular purely visual representations, since they can provide the full context of a single room. To properly describe multi-room environments, however, it is necessary to use collections of images or panoramas, each taken from a different view point and registered together. Obviously, these representations do not explicitly include any 3D information on the scene

Conversely, …

INPUT DATA SOURCES

SIGGRAPH THINK BEYOND 2020 19-23 JULY WASHINGTON DC

1. Purely visual

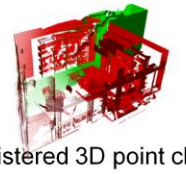
Single still image

Full-view panorama

Registered images/ panoramas

2. Purely geometric

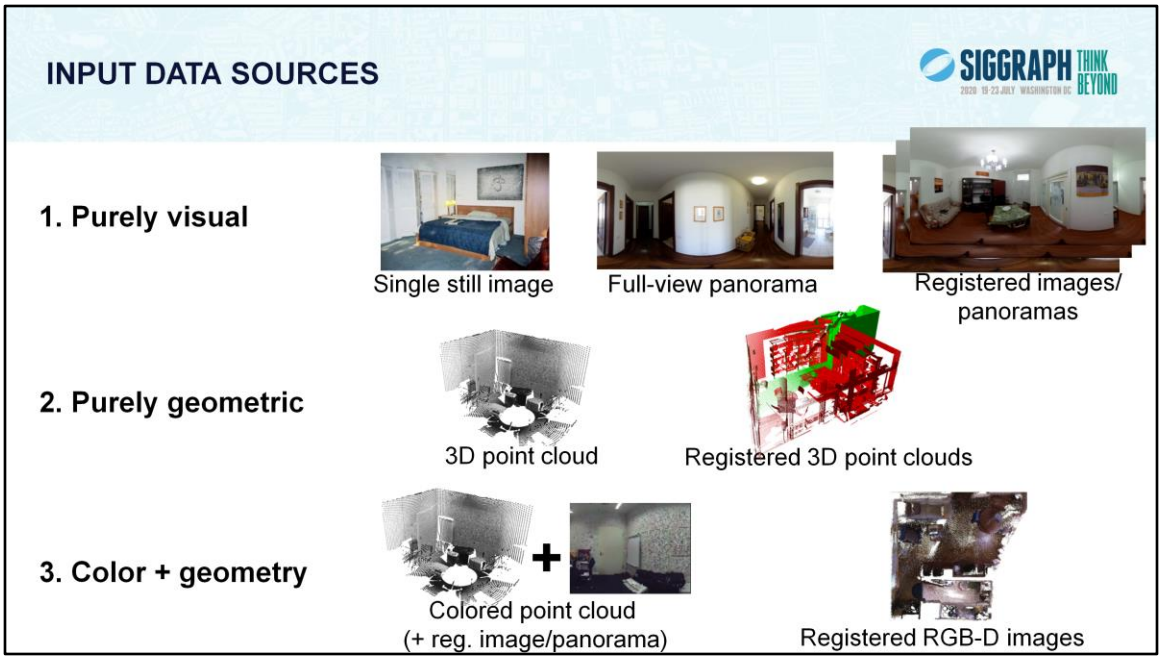
3D point cloud



Registered 3D point clouds

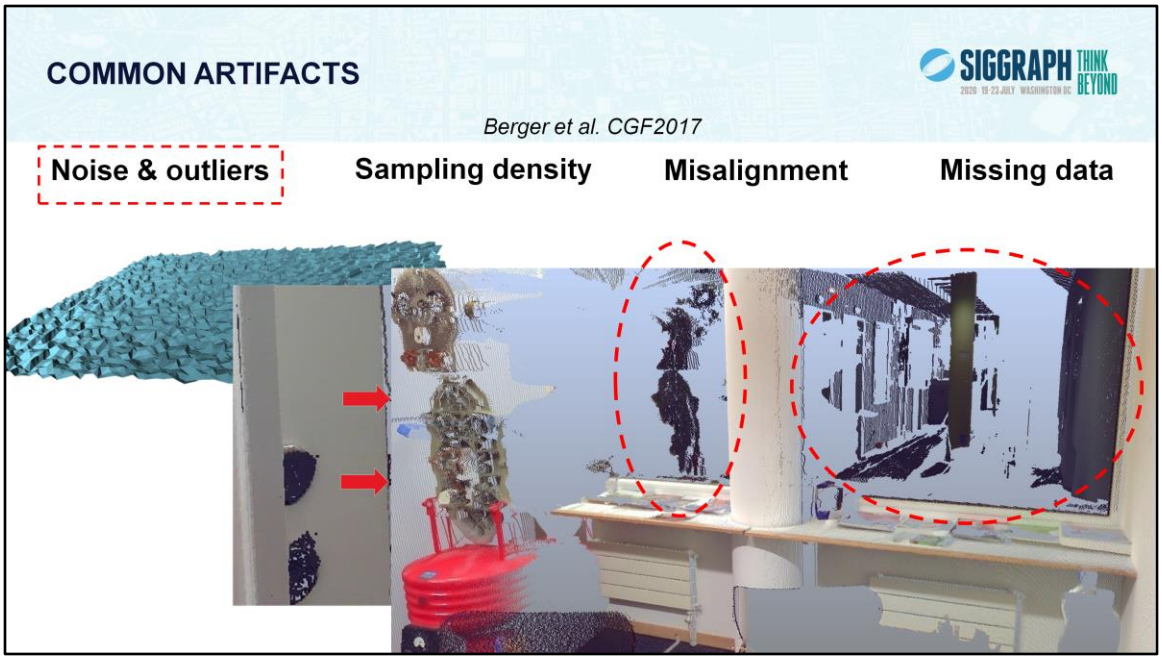
… pure geometric information is provided by representations like 3D point clouds, which consist of dense collections of individual 3D points. Normally, these data have very low measurement noise and are acquired using expensive and bulky terrestrial laser range scanners, though recently there has been a shift towards solutions that are faster, more mobile – and significantly cheaper. Just like for images and panoramas, multiple point clouds taken from different positions AND registered into a single reference frame are needed to properly describe large and complex environments

It is increasingly common to see input representations that combine …



… colorimetric and geometric information. Colored 3D point clouds can be obtained directly from multi-modal devices that combine depth sensors and color cameras, or by registering 3D point clouds and color images acquired separately. Often, multi-modal devices output registered color images together with the generated point clouds. With the recent widespread diffusion of cheap RGB-D cameras, registered collections of color and range images are becoming increasingly popular, and are reaching quality levels comparable to those offered by more expensive laser scanners.

It is often useful to categorize indoor modeling approaches based on the type of input data, as this factor plays a role in the techniques of choice for the reconstruction process.



…also due to the fact that the specific data source used in the acquisition is directly linked to the artifacts that appear in the input data.

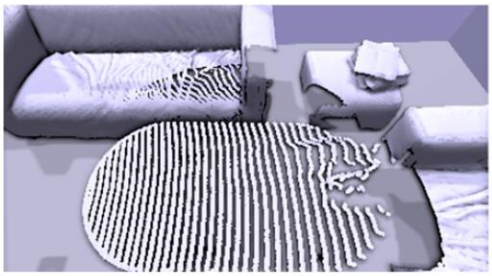
Generally speaking, the artifacts that one can expect to find are the same as those indicated by Berger and colleagues in their survey on “general” surface reconstruction.
However, in the case of indoor environments, these take specific forms and are more or less evident depending on the specific acquisition device used.

All acquisition technologies generally exhibit some degree of measurement noise and have problems with transparent and highly reflective surfaces; this can cause sparse scattered outliers off the actual sensed surfaces or larger and more structured ghosting artifacts - essentially, measurements corresponding to non-existent geometry caused by mirror-like reflections of the light rays that hit these surfaces.

COMMON ARTIFACTS

Berger et al. CGF2017

Noise & outliers | **Sampling density** | **Misalignment** | **Missing data**



Mattausch et al. CGF2014

An insufficient or irregular sampling density is also a recurring issue. In laser-scanned point clouds, this is mainly due to the fact that the rays emitted have uniform angular spacing and can hit scanned geometry in a non-uniform way. However, non-uniform sampling is also an issue for 3D data generated from visual sources, for instance in the presence of texture-less surfaces

COMMON ARTIFACTS

Berger et al. CGF2017

Noise & outliers | **Sampling density** | **Misalignment** | **Missing data**

Choi et al. CVPR2015

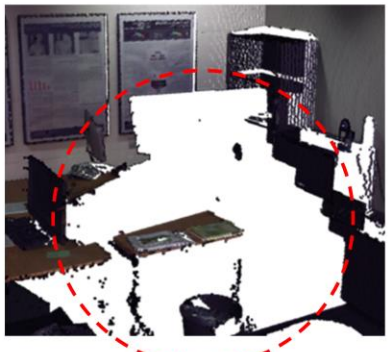
Misalignments can also occur, either between registered 3D point clouds or between individual frames of an RGB-D stream, for instance due to loop closure failures caused by drift

This is a relatively standard problem, only exacerbated by the possible scarcity of feature points detected in texture-less indoor environments, …

COMMON ARTIFACTS

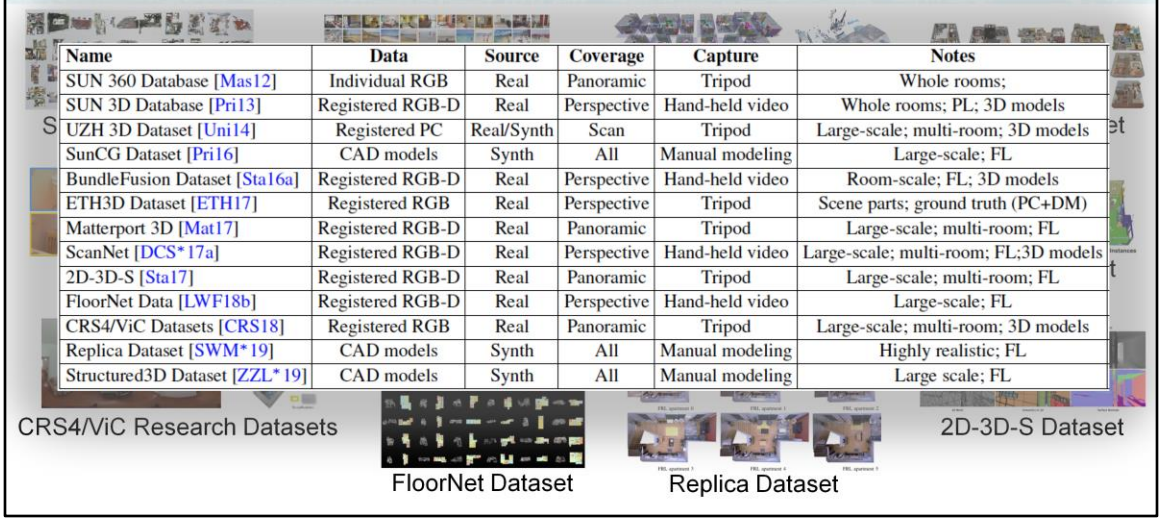
Berger et al. CGF2017

Noise & outliers **Sampling density** **Misalignment** **Missing data**



On the other hand, high amounts of missing data are a distinctive trait of indoor scenes, which are typically highly cluttered. This makes it difficult to capture unaccessible parts of the scene and originates many shadowed areas, caused by objects occluding the line of sight of the sensing device to the structures of interest.

OPEN RESEARCH DATASETS



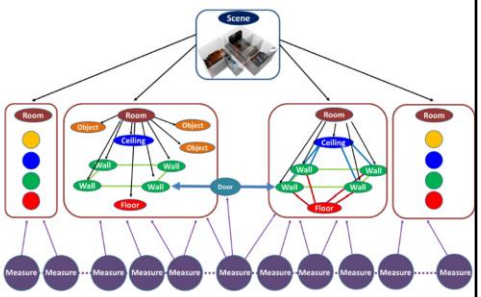
Name	Data	Source	Coverage	Capture	Notes
SUN 360 Database [Mas12]	Individual RGB	Real	Panoramic	Tripod	Whole rooms;
SUN 3D Database [Pri13]	Registered RGB-D	Real	Perspective	Hand-held video	Whole rooms; PL; 3D models
UZH 3D Dataset [Uni14]	Registered PC	Real/Synth	Scan	Tripod	Large-scale; multi-room; 3D models
SunCG Dataset [Pri16]	CAD models	Synth	All	Manual modeling	Large-scale; FL
BundleFusion Dataset [Sta16a]	Registered RGB-D	Real	Perspective	Hand-held video	Room-scale; FL; 3D models
ETH3D Dataset [ETH17]	Registered RGB	Real	Perspective	Tripod	Scene parts; ground truth (PC+DM)
Matterport 3D [Mat17]	Registered RGB-D	Real	Panoramic	Tripod	Large-scale; multi-room; FL
ScanNet [DCS*17a]	Registered RGB-D	Real	Perspective	Hand-held video	Large-scale; multi-room; FL;3D models
2D-3D-S [Sta17]	Registered RGB-D	Real	Panoramic	Tripod	Large-scale; multi-room; FL
FloorNet Data [LWF18b]	Registered RGB-D	Real	Perspective	Hand-held video	Large-scale; FL
CRS4/ViC Datasets [CRS18]	Registered RGB	Real	Panoramic	Tripod	Large-scale; multi-room; 3D models
Replica Dataset [SWM*19]	CAD models	Synth	All	Manual modeling	Highly realistic; FL
Structured3D Dataset [ZZL*19]	CAD models	Synth	All	Manual modeling	Large scale; FL

One can find plenty of examples of these artifacts in the many research datasets of indoor scenes that have been publicly released in recent years.

We don’t have the time for an in-depth description here, so for additional details we refer you to the tutorial notes in the proceedings, as well as to our survey articles.

TARGETED STRUCTURED 3D MODEL

- Abstraction of input data into main elements and their relations
  - Can be defined as scene graph [Ikehata 15, Armeni 19]
- Architectural graph-based data structure
  - Nodes: elements (geometry + appearance)
    - rooms, walls, floors, ceilings, objects
  - Edges: geometric relations (e.g. adjacency)



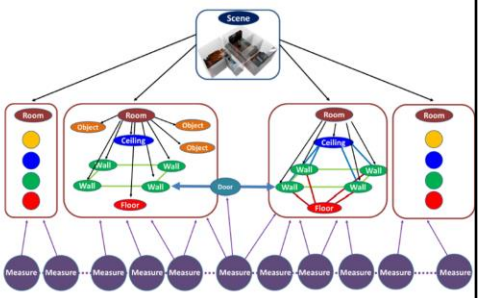
Given the acquired input measurements of an indoor environment, the ultimate goal of automatic structured indoor modeling approaches is to convert them into a structured model.

We mentioned already that this is a high-level model that abstracts the main elements of the environment and their relations. More specifically, also based on the work by Ikehata and colleagues and the more recent paper by Armeni et al., we can describe this model as an architectural graph-based data structure: the nodes correspond to elements of the scene (rooms, walls, floors, ceilings, but also movable objects contained in the scene) and are associated to both geometry and visual appearance; the edges correspond to geometric relationships between the nodes – typically, adjacency.

So based on this, a scene can be described as a graph of rooms bounded by walls, floor and ceiling , connected by portals (doors and passages), and possibly containing furniture or other movable items.

TARGETED STRUCTURED 3D MODEL

- Abstraction of input data into main elements and their relations
  - Can be defined as scene graph [Ikehata 15, Armeni 19]
- Architectural graph-based data structure
  - Nodes: elements (geometry + appearance)
    - rooms, walls, floors, ceilings, objects
  - Edges: geometric relations (e.g. adjacency)
- Topology + geometry + appearance
- Two main goals:
  - Semantic analysis, domain-specific applications
  - Implicitly constrain the output generated, make the modeling tractable



So this graph has a topological part (the connection graph), a geometric part (the shape of the various components) and a visual part (the appearance model of the different nodes).

This representation serves two fundamental purposes:

- First, it simplifies semantic analysis of the scene and enables a number of domain-specific applications, for instance indoor navigation
- Second, it makes the indoor modeling process tractable by implicitly constraining the possible output – and this is particularly important, since indoor scenes normally exhibit extreme variability in both architectural shapes and object arrangements

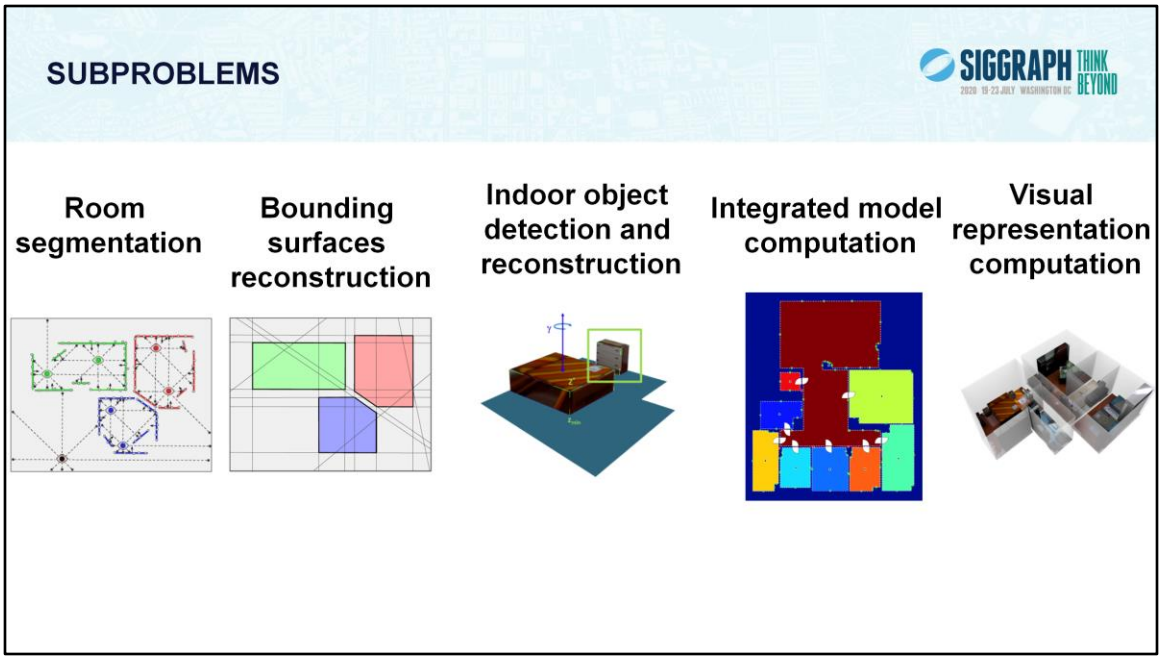
In fact, besides the implicit constraints set by the graph itself, there are a number of explicit assumptions that are commonly made on the architecture of the environment – which are presented here in increasing order of complexity

The most restrictive one is the FLOOR-wall prior , which assumes that the environment is composed of a single flat floor and straight vertical wall, essentially ignoring the ceiling.

The cuboid prior assumes that the room has a cube-like shape and is therefore bounded by six rectangles placed at right angles

According to the indoor world model, the environment must have a horizontal floor and ceiling and vertical walls which all meet at right angles; this is a slightly more restrictive version of the more widely used Manhattan World prior, which allows floors and ceilings to be at different elevations. The Atlanta World prior – or Augmented Manhattan World – lifts the restriction that walls must meet at right angles.

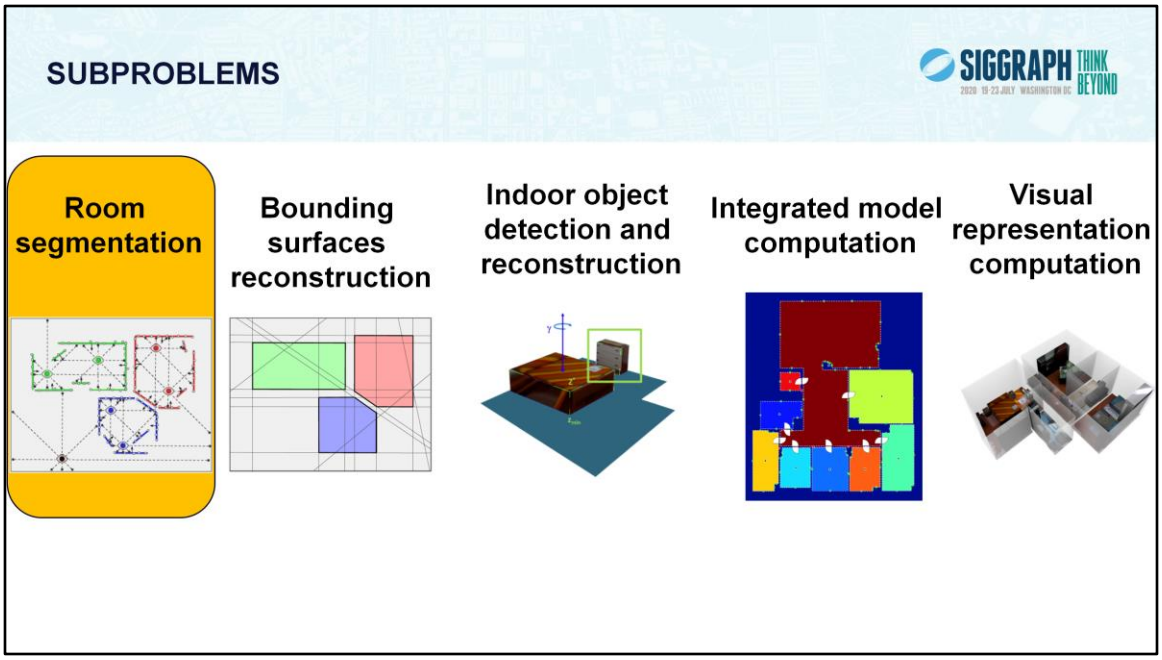
Allowing floors and ceilings to be sloping – while keeping the verticality of walls – results in the Vertical Walls prior; finally, the Piecewise Planarity assumption also includes sloping walls, thus allowing rooms to be general polyhedra. This imposes the least restrictions, but at the same time requires full 3D reasoning on the scene.



These priors influence the different aspects of the structured modeling process, allowing to manage its complexity at different levels.
To better capture this complexity, it is useful to identify a number of basic sub-problems that should be solved to obtain the final model from the measured data.

We denote as **room segmentation** the problem of separating the measured data based on the room to which they belong
The task of **bounding surfaces reconstruction** aims to recover the geometry bounding the individual room shapes
Another conceptual sub-problem is the **detection and reconstruction of the indoor objects**, with the goal to remove elements that represent clutter and possibly reconstruct their footprint or shape
After the individual rooms and the contained objects have been reconstructed, they are fused into a single consistent model – a step that we denote as **integrated model computation**
The last conceptual problem is focused on **adding visual attributes** to the generated model, making it suitable for interactive visualization and navigation.

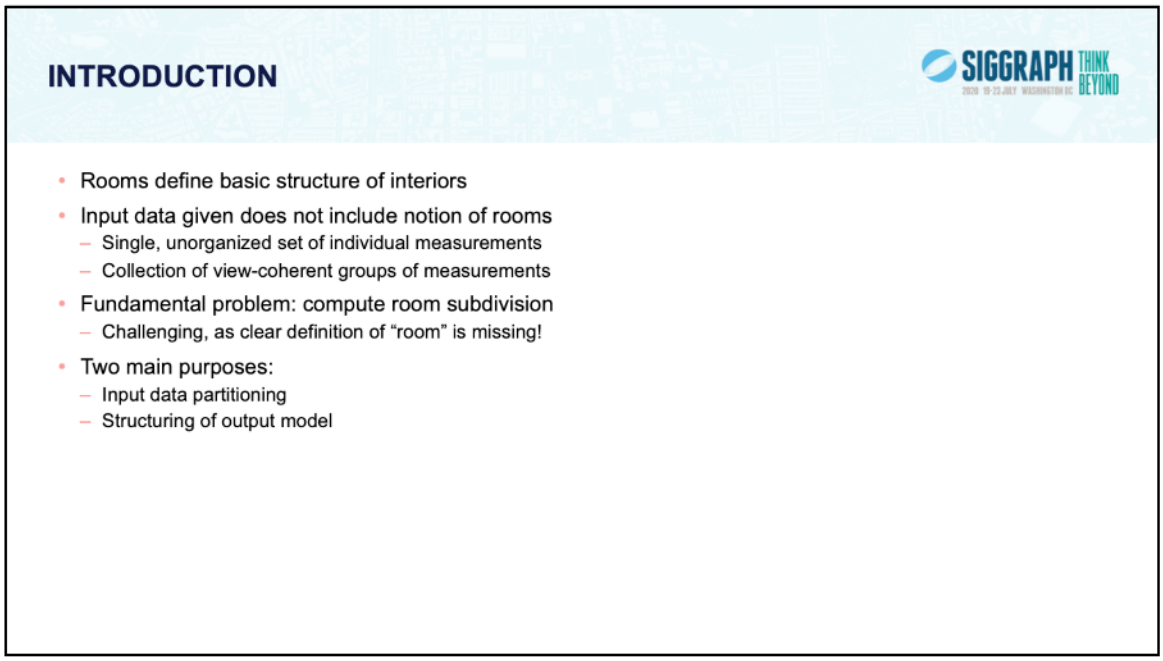
So let's now delve into the details of these five problems…



…starting from room segmentation.

We now move from Sardinia to Switzerland, where my colleague Claudio will discuss this topic…


SIGGRAPH THINK BEYOND
ROOM SEGMENTATION
Speaker: Claudio Mura



Extracting the individual rooms that compose an indoor environment has emerged in the recent years as a very important part of indoor modeling . This is because rooms are the fundamental sub-spaces in which interiors are structured.

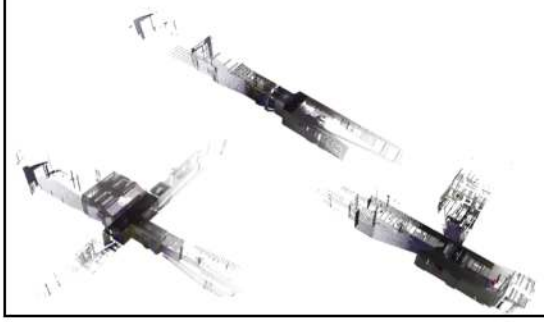
Normally, the raw input measurements do not include any notion of rooms: the input consists in a single list of individual samples (e.g., in the case of a single point cloud or a single image) or possibly in multiple lists of samples, with each list corresponding to a different viewpoint.

Computing the subdivision into different rooms is as important as it is challenging, also due to the lack of a clear definition of what a room is.

In the state-of-the-art, room detection is mainly used for two purposes: first, to partition the input data into separate chunks; second, to actually organize the output model in a way that reflects the structure of the environment

ROOM SEGMENTATION AS INPUT DATA PARTITIONING

- Use room structure to cluster input data into coherent groups
  - *Prior* to actual reconstruction steps
  - More efficient processing, higher accuracy through data filtering
- Typical assumption: input as *collection* of scans/images



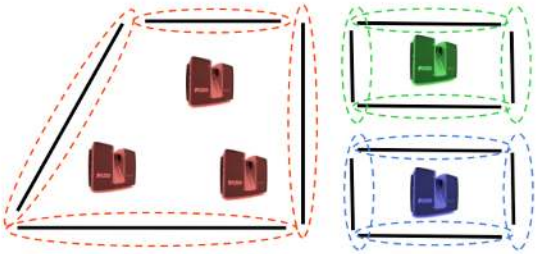
Taking into account the room subdivision **prior** to the actual reconstruction allows to partition the input data into separate, coherent chunks.

This allows for more efficient processing, as the reconstruction can be performed on each chunk separately and possibly in parallel, but also for a more effective filtering of outliers, which is done early on in the pipeline and does not affect subsequent steps.

A common practical assumption is that the input is given as a collection of individual scans or images: the partitioning then amounts to clustering these scans or images according to the room in which they where taken.

INPUT PARTITIONING: 3D DATA

- Survey done with 1 scan per room [Ochmann 14]
  - Trivial clustering
- 1+ scans per room: *visibility clustering* [Mura 16, Ochmann 19]
  - Inspired by viewpoint generation for scene exploration [DiBenedetto 14]
  - Clustering viewpoints based on visible surface overlap

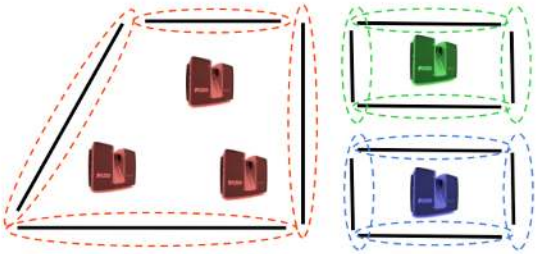


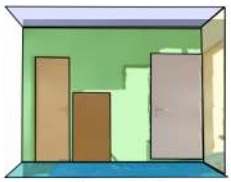
When the input consists in 3D measurements (for instance, 3D point clouds) the simplest case is to assume that the surveying was done by taking exactly one scan for each room of the environment. In this case, the clustering is already implicitly provided as part of the input data.

If one or more scans are taken for each room, some clustering needs to be performed. A well-established technique that has emerged recently is to cluster the scans based on the visibility information from their origin, that is, the viewpoint from which the scans were taken. This is based on a technique originally developed for visual scene exploration and consists in computing, for each viewpoint, which parts of the scene are visible, and then grouping the viewpoints based on the visible surface overlap, that is, based on the amount of visible scene parts they have in common.

INPUT PARTITIONING: 3D DATA

- Survey done with 1 scan per room [Ochmann 14]
  - Trivial clustering
- 1+ scans per room: *visibility clustering* [Mura 16, Ochmann 19]
  - Inspired by viewpoint generation for scene exploration [DiBenedetto 14]
  - Clustering viewpoints based on visible surface overlap

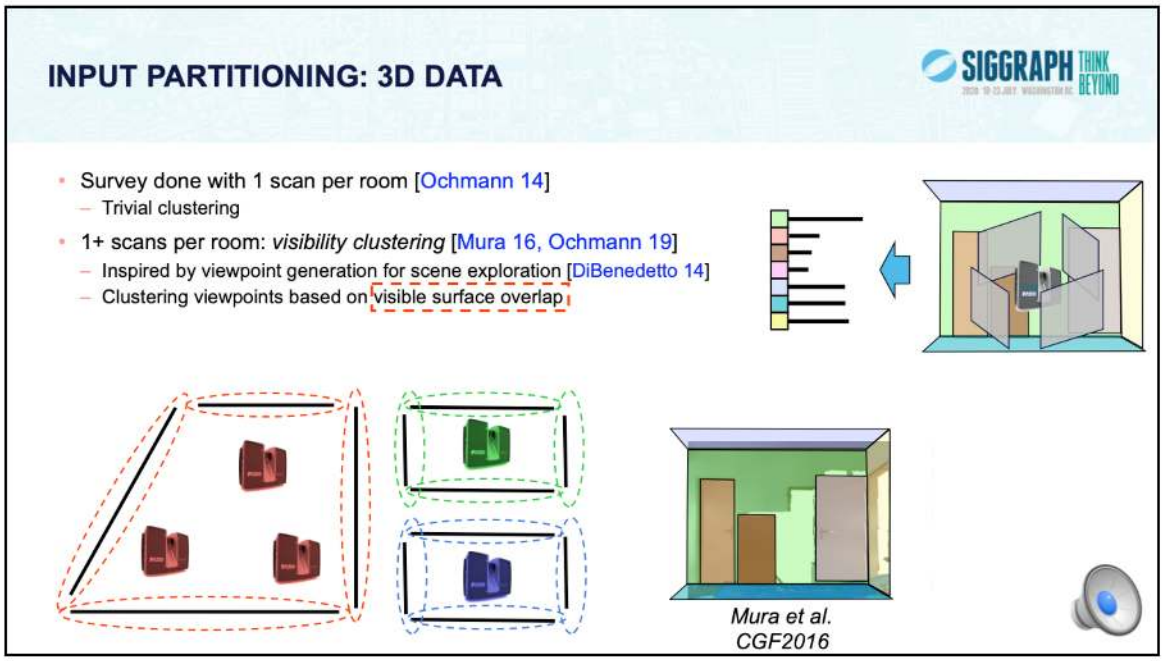




Mura et al. CGF2016

These scene parts can be defined in a number of ways: they could be 2D line segments corresponding to vertical walls projected onto the ground plane, but they can also be defined directly in 3D space using 3D geometric proxies, typically computed from patches of co-planar 3D points extracted in a pre-processing step.

In the work by Mura et al. of 2016, the scene parts are represented as fitting rectangles, oriented based on their normal and on the global up vector of the scene. in the more recent work from 2019, Ochmann and colleagues additionally build an occupancy grid on top of these planar shapes; in the later steps of their pipeline, the occupancy information is used to evaluate to what extend these shapes are densely covered by input scan points.

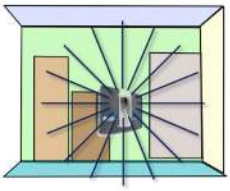


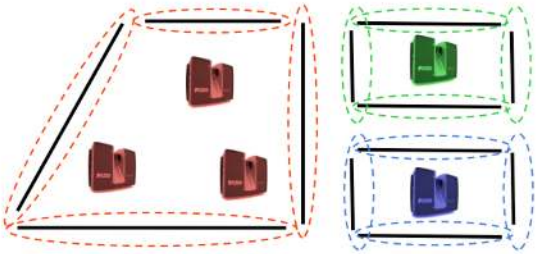
Given these planar 3D proxies, the visible surface overlap can be defined and computed by exploiting the standard graphics pipeline. One can render the scene - defined as the set of planar proxies - from each viewpoint, partitioning the entire field of view into a number of separate views and then counting how many pixels are visible in all views for each planar proxy, essentially building a histogram with a bin for each rectangle.

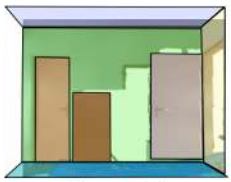
Given two histograms corresponding to two different viewpoints, their visible surface overlap can be defined as the similarity of the two histograms, using for instance cosine similarity or a more ad-hoc definition.

INPUT PARTITIONING: 3D DATA

- Survey done with 1 scan per room [Ochmann 14]
  - Trivial clustering
- 1+ scans per room: *visibility clustering* [Mura 16, Ochmann 19]
  - Inspired by viewpoint generation for scene exploration [DiBenedetto 14]
  - Clustering viewpoints based on visible surface overlap







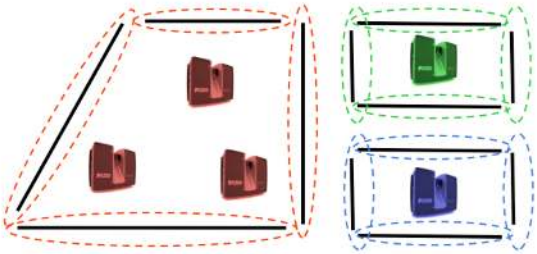
Mura et al. CGF2016

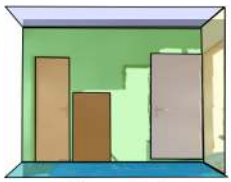
Note that it is also possible to obtain the same visibility information by ray casting, using for instance toolkits like Intel Embree or NVIDIA Optix for efficient implementation.

INPUT PARTITIONING: 3D DATA

- Survey done with 1 scan per room [Ochmann 14]
  - Trivial clustering
- 1+ scans per room: *visibility clustering* [Mura 16, Ochmann 19]
  - Inspired by viewpoint generation for scene exploration [DiBenedetto 14]
  - Clustering viewpoints based on visible surface overlap

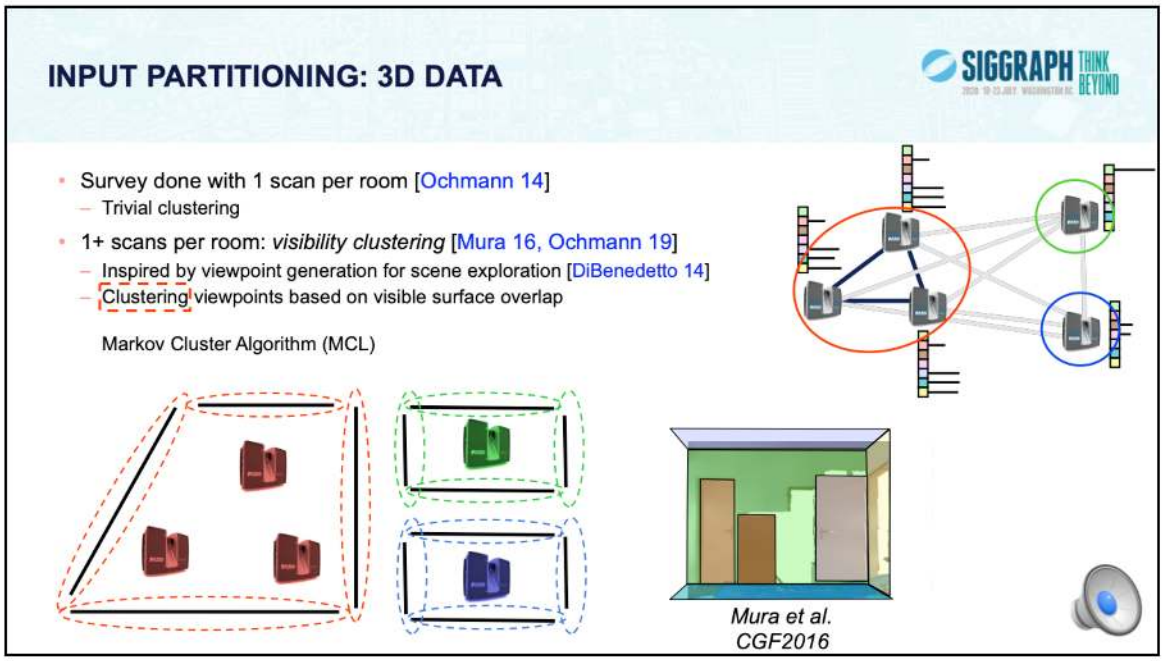
Markov Cluster Algorithm (MCL)





Mura et al. CGF2016

Besides the definition of the visible surface overlap, the other major ingredient of visibility clustering is the clustering algorithm itself. In principle, any clustering algorithm capable of determining the number of clusters automatically from the data can be used.
Nevertheless, the algorithm chosen by Di Benedetto et al. in their work from 2014 on scene exploration − namely, the Markov Cluster Algorithm (MCL) − has proven to work quite robustly in this context.



In MCL, the data items to be clustered (the viewpoints, in our case) are represented as nodes of a fully-connected graph. The algorithm simulates random walks within this graph, based on weights defined on the edges; specifically, each edge is associated to the probability of transitioning from one endpoint node to the other. In this setting, these probabilities correspond to the visible surface overlap between the viewpoints at the endpoints of the edge. MCL computes random walks of increasing lengths; the probability of longer paths is higher within clusters rather than between different clusters. The algorithm progressively updates the probability on the edges according to this principle, eventually leading to inter-cluster connections being removed and to links between nodes in the same cluster being preserved.
This yields clusters of viewpoints that reflect the room structure of the environment.

The visibility clustering process described so far is specifically tailored for the case of 3D data as input.

INPUT PARTITIONING: VISUAL DATA

- Conceptually similar, yet more robustness needed
- Grouping based on 3D features [Furukawa 09, Pintore 18]
  - Might be too sparse
  - Images taken in adjacent rooms near open door too similar
- Grouping based on global + local similarity between images [Zhang 17]
  - Many false positives in settings with standardized furniture (e.g. office spaces)

However, a conceptually similar approach can be used when the input is represented by pure visual data, that is, by a collection of color images or panoramas. In this case, instead of considering patches of 3D points, one can extract 3D features from the input images – for instance using Structure from Motion – and group the images accordingly. However, this approach can fail if the set of features obtained is too sparse. Even if this is not the case, there is another more conceptual problem: two images taken in different, adjacent rooms at locations close to an open door connecting these rooms are likely to contain many common features, thus being considered similar and assigned to the same room cluster.

As an alternative, one can perform the grouping using similarity metrics that take into account both local and global properties of the images. The problem with this is that many indoor environments (like offices spaces) contain standardized furniture and have similar architectural traits, which results in many false positives in the image matching process.

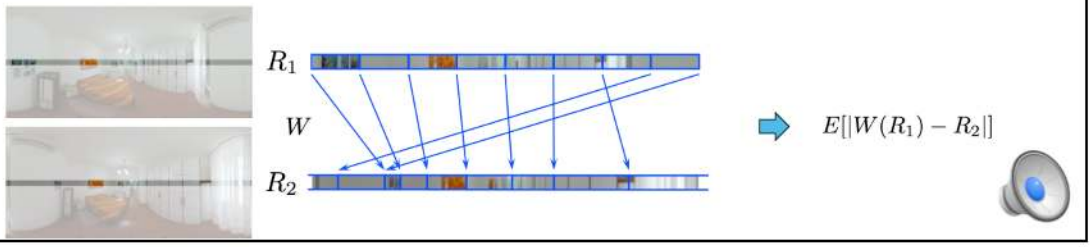
INPUT PARTITIONING: VISUAL DATA

- Conceptually similar, yet more robustness needed
- Grouping based on 3D features [Furukawa 09, Pintore 18]
  - Might be too sparse
  - Images taken in adjacent rooms near open door too similar
- Grouping based on global + local similarity between images [Zhang 17]
  - Many false positives in settings with standardized furniture (e.g. office spaces)
- Measure how well panoramas are warped into one another [Pintore 19]
  - Using only horizontal central slices for robustness

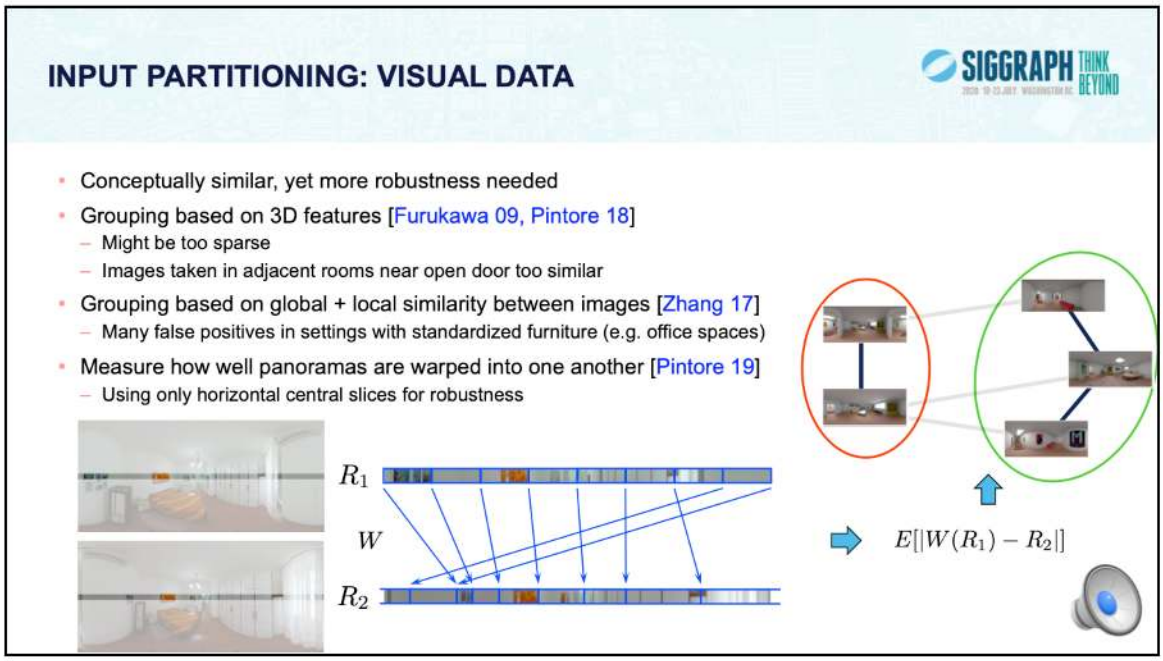
To solve these issues, an alternative solution specifically targeting panoramic images has been proposed very recently by Pintore and colleagues. The idea is that if two panoramic images are taken inside the same room, it should be possible to warp an unoccluded portion of an image onto a matching portion of the other. Assuming that all panoramic images are taken from the same height, one can consider the horizontal central slice of the input panoramas. It is very reasonable to assume that this slice shows a largely unoccluded profile of the room, as many of the occluding objects are either lying on the floor and are lower than this height or span a much larger vertical extent and are all intersected by this slice.

INPUT PARTITIONING: VISUAL DATA

- Conceptually similar, yet more robustness needed
- Grouping based on 3D features [Furukawa 09, Pintore 18]
  - Might be too sparse
  - Images taken in adjacent rooms near open door too similar
- Grouping based on global + local similarity between images [Zhang 17]
  - Many false positives in settings with standardized furniture (e.g. office spaces)
- Measure how well panoramas are warped into one another [Pintore 19]
  - Using only horizontal central slices for robustness



Let's call these slices for two given images R1 and R2. We can divide R1 into a certain number of blocks and use a simple optimization algorithm to compute the best warping function W that warps the sequence of blocks in R1 into the best matching sequence of blocks in R2. Here, "best" means minimizing, for each pair of matching boxes, the average distance between the colors of the pixels that correspond under the warping. This distance is defined as the error of warping R1 onto R2. Pintore et al. use a hierarchical formulation for this warping process, with the number of blocks as well as the pixel resolution of the slices increase in a coarse-to-fine manner.



This definition of error between images is used to derive weights for the arcs of a graph similar to the one already described for visibility clustering for 3D data. The nodes of this graph are in this case panoramic images and the weights of the arcs are considered as probabilities of transitioning from the panorama on one endpoint to the panorama on the other endpoint. A clustering based on random walks on this weighted graph yields the room-based grouping of the panoramas.

As an interesting detail, note that this graph is not fully-connected: the nodes of two panoramas are connected with an edge only if they share a sufficient number of 3D features obtained from Multi-View Stereo. So while these features may not be dense enough to compute a reliable similarity measure between two panoramas, they do allow to prune connections between panoramas that are highly likely to not belong to the same room.

ROOM SEGMENTATION AS STRUCTURING PROCESS

- Focus on including semantics and structural information in output model
- Room segmentation integrated in reconstruction process
- Moves from in/out segmentation [Oesau 14, Turner 12]
  - Label regions of a space subdivision as inside/outside bounding walls

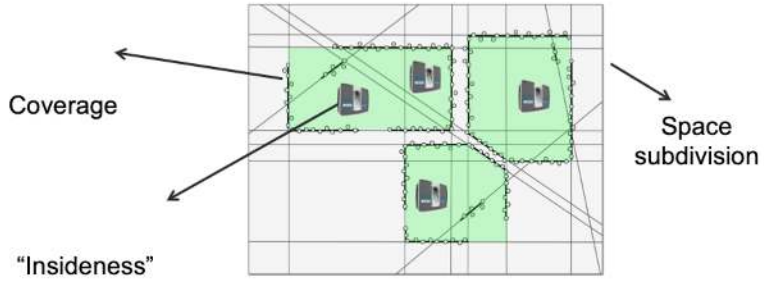
So far we have considered how the room subdivision can be exploited for input data partitioning.

More commonly, the focus is on how to include the semantic and structural information represented by rooms in the output, for the purpose of producing a structured model. This is normally tightly coupled with the reconstruction process, which in turn moves from a basic scheme designed to simply reconstruct the indoor space as a single entity – essentially simply separating the space that is inside the bounding walls from the outer space.

A detailed description of this technique will be provided in the next part of the course. We briefly introduce here how this works, considering the case of a 2D domain corresponding to a top-down view of the environment for simplicity.

ROOM SEGMENTATION AS STRUCTURING PROCESS

- Focus on including semantics and structural information in output model
- Room segmentation integrated in reconstruction process
- Moves from in/out segmentation [Oesau 14, Turner 12]
  - Label regions of a space subdivision as inside/outside bounding walls

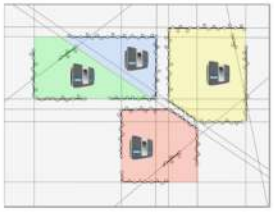


Given the input samples (e.g., scanned 3D points projected onto the horizontal 2D domain) a number of 2D line segments are fitted to them. These are further clustered into representative lines, which can be used to subdivide the 2D spatial domain into a number of polygonal regions of space, which define a space subdivision. This way of creating a space subdivision is very common in the state-of-the-art, but note that it is not the only way: it could be extended to the third dimension and the regions could also be shaped differently.
Regardless of the specific traits of the space subdivision, the regions that correspond to the interior space can be extracted using a variety of techniques – often, min-cut optimization – using mainly two types of information.
The first is the point coverage at the interface between two adjacent regions, which indicates how likely it is that the two regions are separated by a wall.
The second is a measure of "insideness", which is computed on individual regions (not on pairs of adjacent ones) and denotes how likely it is that the region corresponds to interior space. This measure is often defined as the fraction of visibility rays shot from the center of the region that hit input measurements.

To extend this technique to the multi-room case, one typically includes the scan positions in this process.

OVERSEGMENTATION + MERGING

- Initially assume 1 scan per room, use 1 label per scan
  - Multi-label clustering of regions of space subdivision [Ochmann 16]
  - Solved using graph-cuts [Boykov 14]

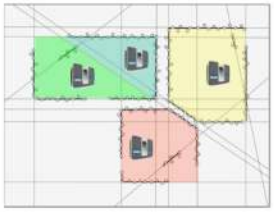


Perhaps the simplest extension of this model is to initially assume that each room is covered by one and only one scan. The set of scans defines a set of labels, with one label for each room. The binary inside/outside reconstruction is then extended to a multi-label classification of the regions of the space subdivision, normally computed by approximate energy minimization based on graph-cuts.

Obviously, if the assumption of having exactly one scan per room does not hold, this process results in over-segmentation,

OVERSEGMENTATION + MERGING

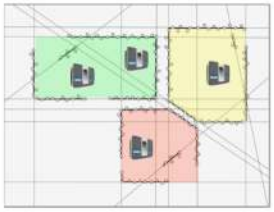
- Initially assume 1 scan per room, use 1 label per scan
  - Multi-label clustering of regions of space subdivision [Ochmann 16]
  - Solved using graph-cuts [Boykov 14]
- Merge over-segmented clusters corresponding to same room
  - Thresholding based on coverage at border between adjacent clusters [Mura14]
  - Classify candidate boundary as true or virtual using SVM [Ochmann 16]
  - Detect *peak-gap-peak* pattern between clusters [Armeni 16]

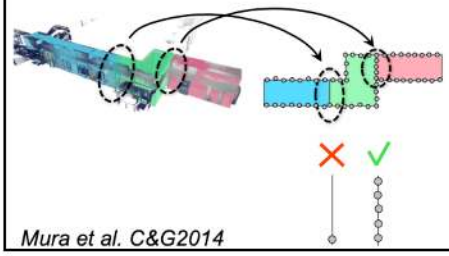


…which should be fixed by merging the over-segmented clusters belonging to the same room. Several approaches have been used for this, typically based on well-founded heuristics.

OVERSEGMENTATION + MERGING

- Initially assume 1 scan per room, use 1 label per scan
  - Multi-label clustering of regions of space subdivision [Ochmann 16]
  - Solved using graph-cuts [Boykov 14]
- Merge over-segmented clusters corresponding to same room
  - Thresholding based on coverage at border between adjacent clusters [Mura14]
  - Classify candidate boundary as true or virtual using SVM [Ochmann 16]
  - Detect *peak-gap-peak* pattern between clusters [Armeni 16]





Mura et al. C&G2014

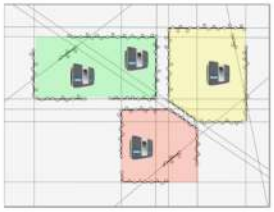
Basic intuition tells us that the border between two rooms should correspond to a solid wall; hence, the interface between two adjacent rooms should be highly covered by scanned measurements. Working in the 2D domain of the ground plane, we can consider the 2D line segments that define the interface between each pair of adjacent rooms. We can bin the extent of each segment and compute the fraction of bins that are occupied by points. If this is close to 1, we can conclude that the rooms are actually separated by a solid wall and that they should not be merged.

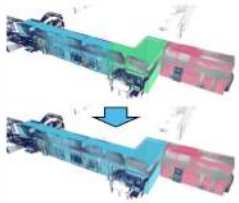
On the other hand, if the fraction of the extent occupied by scanned points is close to 0, we can conclude that the two rooms should be merged.

In practice, even in the presence of viewpoint occlusions and missing data the covered extent tends to be very low in the case of fake separations and very high otherwise, so that a plain thresholding-based rule is sufficient to unambiguously distinguish between the two cases.

OVERSEGMENTATION + MERGING

- Initially assume 1 scan per room, use 1 label per scan
  - Multi-label clustering of regions of space subdivision [Ochmann 16]
  - Solved using graph-cuts [Boykov 14]
- Merge over-segmented clusters corresponding to same room
  - Thresholding based on coverage at border between adjacent clusters [Mura14]
  - Classify candidate boundary as true or virtual using SVM [Ochmann 16]
  - Detect *peak-gap-peak* pattern between clusters [Armeni 16]





Mura et al. C&G2014

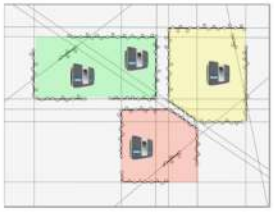
Ochmann et al. C&G2016

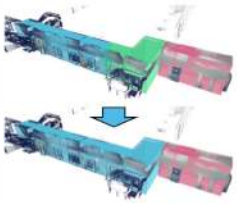
More elaborate techniques can be used, especially if one also aims to reconstruct a possible door located at the actual interface between two adjacent rooms.

To do this, in their work from 2016 Ochmann and colleagues analyze the candidate borders between adjacent rooms using machine learning. In particular, for each border, they consider the scan positions on the two opposite sides of the boundary and cast rays from each of them towards this boundary.
Then, they cluster the intersection points between these rays and the vertical plane at the candidate boundary. Each cluster of points is associated to a six-dimensional feature vector (which includes the width and height of its bounding box and the approximate coverage); the feature vector is fed to an SVM classifier, which determines whether the candidate boundary actually corresponds to a separation – possibly with a door or a window – or to an artifact of the reconstruction process.

OVERSEGMENTATION + MERGING

- Initially assume 1 scan per room, use 1 label per scan
  - Multi-label clustering of regions of space subdivision [Ochmann 16]
  - Solved using graph-cuts [Boykov 14]
- Merge over-segmented clusters corresponding to same room
  - Thresholding based on coverage at border between adjacent clusters [Mura14]
  - Classify candidate boundary as true or virtual using SVM [Ochmann 16]
  - Detect *peak-gap-peak* pattern between clusters [Armeni 16]





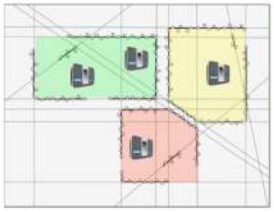
Mura et al. C&G2014 *Ochmann et al. C&G2016* *Armeni et al. CVPR2016*

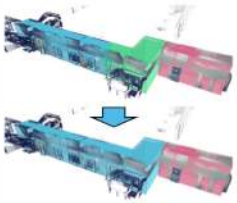
In case of high levels of clutter, a more robust approach to determining whether two rooms are actually separate is to look for a specific peak-to-peak pattern. This corresponds to the presence of a wall surface, some void space, and another wall surface.

Armeni et al have implemented this approach in the case of Manhattan World environments, in which walls are vertical and all meet at right angles. Let us assume that the z axis corresponds to the vertical directions and the x and y axes span the horizontal domain of the floorplan. Armeni and colleagues consider the x and y directions separately and, for each of them, analyze the density histogram of the points. A peak in this histogram at a location on the horizontal axis indicates the presence of a wall sheet, as many points at different height levels must insist on that location. In the presence of empty space the histogram value is null and in the case of a single horizontal surface it has a low value.

OVERSEGMENTATION + MERGING

- Initially assume 1 scan per room, use 1 label per scan
  - Multi-label clustering of regions of space subdivision [Ochmann 16]
  - Solved using graph-cuts [Boykov 14]
- Merge over-segmented clusters corresponding to same room
  - Thresholding based on coverage at border between adjacent clusters [Mura14]
  - Classify candidate boundary as true or virtual using SVM [Ochmann 16]
  - Detect *peak-gap-peak* pattern between clusters [Armeni 16]



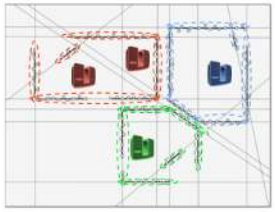


Mura et al. C&G2014 *Ochmann et al. C&G2016* *Armeni et al. CVPR2016*

Based on this, the presence of a separating wall can be detected by convolving the x and y histograms with a pre-defined filter that corresponds to the profile of a wall, followed by empty space, followed by another wall. The highest peak in the convolved histogram is the best candidate location for a boundary wall, so based on the value of this peak one can determine whether an actual separation between two rooms occurs. Note that, in practice, a bank of filters is used instead of a single one.

VISIBILITY-BASED CLUSTERING

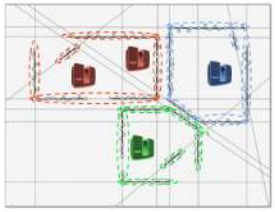
- Avoid oversegmentation by clustering scan viewpoints
  - Actual number of rooms available when grouping regions of space subdivision
  - Visible surface overlap drives the clustering
  - Same idea of input partitioning, applied during reconstruction

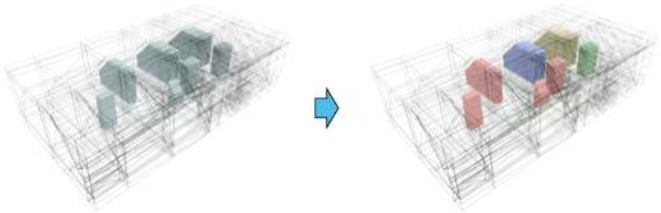


Clearly, the merging step can be avoided if over-segmentation is avoided, which can be done by clustering the scan positions before the actual labeling: this ensures that the correct number of room labels is available when the multi-label optimization is performed.
A visibility clustering similar to the one shown for input data partitioning can be applied in this case.

VISIBILITY-BASED CLUSTERING

- Avoid oversegmentation by clustering scan viewpoints
  - Actual number of rooms available when grouping regions of space subdivision
  - Visible surface overlap drives the clustering
  - Same idea of input partitioning, applied during reconstruction
- Cluster regions of space subdivision containing a viewpoint [Mura 16]



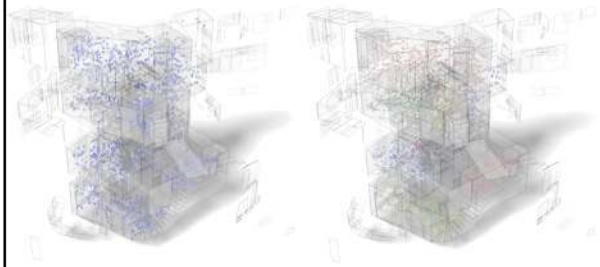


Mura et al. CGF2016

In practice, the only minor difference is that this procedure is normally applied to the regions of the space subdivision that contain a viewpoint, rather than to the viewpoints themselves. For instance, if the space subdivision is a three-dimensional structure, the visibility clustering is applied to the polyhedral regions of space in which the viewpoints lie.

ROOM SEGMENTATION, WITHOUT SCAN POSITIONS

- Extension: visibility clustering on synthetic scan position
  - Dense sampling of space around input data [Mura 17]
  - View probes on planar regions detected from input [Ochmann 19]
  - In-between input partitioning and room reconstruction



Mura & Pajarola Siggraph Posters 2017

Ochmann et al. IJPRS2019

This approach can be adapted to the case in which the scan positions are not available, which typically happens when the input is given as a single, unorganized list of 3D measurements rather than as a collection of scans with viewpoint information.

A direct extension is to generate synthetic viewpoints, for instance by performing a dense sampling of the space around the input data, and then applying the same procedure to these synthetic viewpoints. Doing so in a trivial way incurs a high computational cost.

An alternative has been proposed by Ochmann and colleagues, who actually generate view probes on planar patches fitted to the input data in a pre-processing step. More precisely, these patches correspond to the cells of occupancy grids fitted to vertical planar segments of 3D points, detected using the RANSAC algorithm. In this case, the room segmentation is done by clustering patches of wall points.

Regardless of the technical details, it is important to notice that performing visibility clustering using synthetic viewpoints goes beyond the plain inclusion of room structure in the reconstruction process and actually moves in the direction of an input partitioning scheme.

ROOM SEGMENTATION ON 2D TOP-DOWN VIEWS

- Synthetic viewpoints on medial axis of interior space [Ambrus17]

More specialized and well-established techniques have been proposed for the case in which the room segmentation process is applied to a top-down view of the environment – specifically, to a map of the interior space defined in this domain.

Ambrus et al. have proposed an approach that relies on the generation of synthetic viewpoints, a technique that we have just examined for the more general case of input 3D data.
In their pipeline, the rooms are obtained as clusters of pixels of an image-based representation of the 2D map of the interior space. The pixels corresponding to each room are obtained by applying a multi-label optimization: each pixel is assigned one of N labels corresponding to the N rooms of the environment, in a way that minimizes a specially-crafted energy function.

To get the number N of room labels, Ambrus and colleagues compute the medial axis of the interior space and sample a number of points on this axis. The intuition behind this is that points on the medial axis are maximally distant from the bounding walls and therefore have a high visibility on most of the surrounding space. New viewpoints are sampled from the medial axis in a greedy, iterative process, until most of the locations of the map are within a minimum distance from a viewpoint.

ROOM SEGMENTATION ON 2D TOP-DOWN VIEWS

- Rooms as clusters of pixels that "see" same boundary pixels [Ikehata 15]

Other approaches that work on the 2D top-down view of the environment do not require the use of viewpoints, be them synthetic viewpoints or actual ones. An example is the work by Ikehata and colleagues. As in the work by Ambrus et al., rooms are obtained as clusters of pixels of the image-based top-down representation of the interior space. After a simple refinement that gets rid of outliers, a number of pixels are sampled from the interior space, and for each of them a binary visibility vector is computed, encoding whether the pixel is visible from each boundary pixel.

The subsampled pixels are clustered using the k-medoids algorithm based on the distance between their binary visibility vectors. An initial set of 200 clusters is extracted; these results are further refined by running the k-medoids algorithm again and then merging clusters whose centroids are closer than a given threshold. The segmentation results are then propagated to all pixels of the interior map by nearest neighbor.

ROOM SEGMENTATION ON 2D TOP-DOWN VIEWS

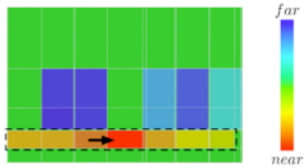
- Rooms as clusters of pixels that "see" same boundary pixels [Ikehata 15]

- Recently, room masks computed as output of R-CNNs [Chen 19]

In their 2019 ICCV contribution, Chen and colleagues have improved on this technique by detecting coarse room segments using instance semantic segmentation. Specifically, they represent the top-down view of the environment as a 4-channel image, where the 4 channels at a pixel correspond to the density of scanned points and to the 3 coordinates of the normal vector at that pixel. This image is processed by a specific type of recurrent convolutional neural network (called Mask R-CNN) that extracts a segmentation masks for the rooms of the environment.
Note that these masks are quite crude approximations of the true room shapes and are converted to a more regular and consistent vector-based representation in a subsequent optimization step.

ROOM SEGMENTATION ON 2D TOP-DOWN VIEWS

- Clustering of space subdivision regions using diffusion distances [Mura 14]



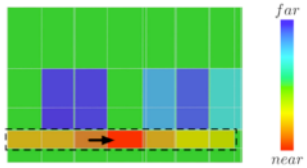
Mura et al. C&G2014

Although based on different techniques, all the specialized approaches described so far use a raster, image-based representation of the top-down view of the environment. There are however examples of pipelines that model this 2D domain differently.

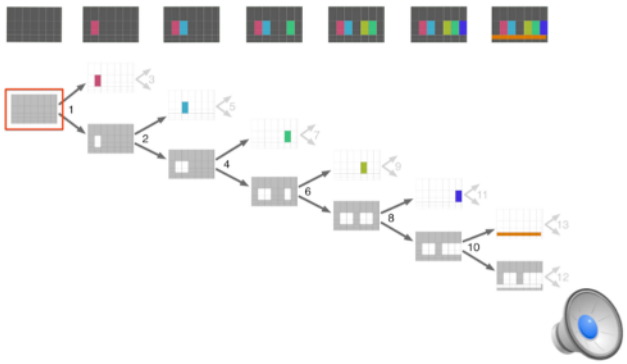
In their work from 2014, Mura and colleagues detect rooms by appropriately grouping regions of a 2D space subdivision, using diffusion maps to establish a similarity measure between these regions. This approach simulates the diffusion of heat from sources placed in the regions: high-coverage interfaces between regions (corresponding to walls) block heat diffusion, resulting in a uniform heat level within the same room. Hence, any two regions corresponding to the same room have a low diffusion distance, whereas the distance between regions that lie in different rooms is high.

ROOM SEGMENTATION ON 2D TOP-DOWN VIEWS

- Clustering of space subdivision regions using diffusion distances [Mura 14]



Mura et al. C&G2014



These diffusion distance are used to drive an iterative clustering process: at each step, a binary k-medoid algorithm is applied to detach one room cluster from the rest of the regions of the space subdivision. The termination criterion for this iterative clustering is formulated based on the knowledge of the scan positions from which the environment was acquired. The assumption is that, at acquisition time, each room was scanned by at least one position. Then, each time a room cluster is extracted in the clustering process, the scan positions that fall inside that cluster are marked as assigned. The iterative process can be terminated when all scan positions are marked as assigned, since at that point all rooms will have been extracted.

DISCUSSION

- Dual role of room segmentation
  - Provide criterion to pre-partition input data => more efficient computation
  - Add basic structuring to output model
- Room-based input partitioning still largely unexplored
  - Especially for purely visual inputs
- Room segmentation more well-established
  - Especially for 3D input data, working on 2D top-down views
  - Largely based on visibility reasoning
  - Often embedded in reconstruction of room models

Pintore et al. CGF2019

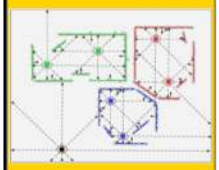
Mura et al. CGF2016

To summarize the discussion on room segmentation, it is important to highlight that this step has a dual role within the structured indoor modeling process: on the one hand, it allows to partition the input data into separate chunks, allowing for independent processing of each room and thus for increased computational efficiency; on the other hand, and even more importantly, it provides fundamental information about the structure of the environment that is included directly in the output model.

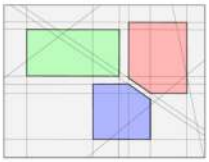
Exploiting the room structure for input partitioning is still largely uncharted territory, especially in the case of purely visual input, while applying room segmentation to structure the output model is a more well-established technique. A number of approaches have been developed, especially when the input is represented by 3D data and the reconstruction domain is the 2D top-down view of the environment. Such techniques are typically based on visibility reasoning, but some alternative ideas have been explored as well.

SUBPROBLEMS

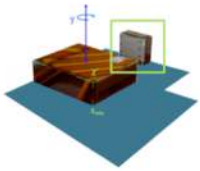
Room segmentation



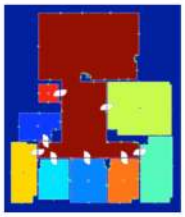
Bounding surfaces reconstruction



Indoor object detection and reconstruction



Integrated model computation

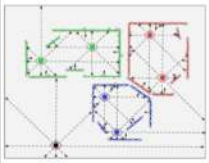


Visual representation computation

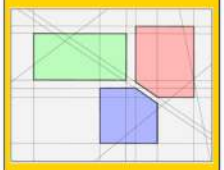
SUBPROBLEMS

SIGGRAPH THINK BEYOND

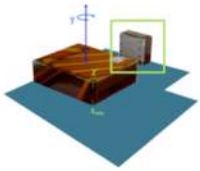
Room segmentation



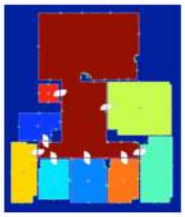
Bounding surfaces reconstruction



Indoor object detection and reconstruction



Integrated model computation



Visual representation computation


SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC
BOUNDING SURFACES RECONSTRUCTION / 1
Speaker: Renato Pajarola
University of Zürich

INTRODUCTION

- Goal: recover structural elements from defined indoor spaces
  - Input: rooms partitioning
    - At least one space/room
  - Output: room layout
    - floor, ceiling, walls, etc.

The goal of this task is to recover the structure of elements from a partitioned indoor space.

The typical input is a room segmentation that is obtained in a previous step of the reconstruction pipeline or generated as an integral part of this step,
and the expected output is a room layout, floor plan with the appropriate wall elements.

INTRODUCTION

- Goal: recover structural elements from defined indoor spaces
  - Input: rooms partitioning
    - At least one space/room
  - Output: room layout
    - floor, ceiling, walls, etc.

- Major challenge for indoor reconstruction
  - Main problem: clutter
    - Not relevant for structure identification
    - Causes occlusion and missing sampling
  - Generic surface reconstruction often fails

This step is one of the major challenges of indoor reconstruction.
In particular, one of the main problems is the clutter in the data as the scanned indoor environment usually includes various objects and artifacts.
Clearly, this clutter is not relevant for the architectural structure identification and also the geometry to be reconstructed.
Occlusions and missing sampling further complicate this step.

Thus generic methods for surface reconstruction tend to fail in this particular context.

APPROACHES BASED ON INPUT DATA TYPE

- Reconstruction without geometric measures as input source
  - No 3D information is explicitly present (e.g., single RGB image)
  - Geometric information from image features through strong priors
  - Top-down (fitting) or bottom-up (clues assembling)
- Reconstruction from sparse geometric measures as input sources
  - Data fusion techniques to integrate known 3D data and 2D image geometric reasoning
  - Less restrictive priors
- Reconstruction from dense geometric measures as input sources
  - High-density sampling is required to recover high level geometric primitives
  - Boundary reconstruction from patch-based representation

The proposed approaches can usually be summarized according to their input.

First, we have methods that reconstruct the layout without any geometric measures available, like it is the case for single image methods for example.
In this scenario the geometric information is obtained from image features typically assuming very strong priors.

Then we have methods that reconstruct from sparse geometric information like from data fusion of 2D image and 3D data, or from structure from motion from images.
In this case we already see less restrictive priors.

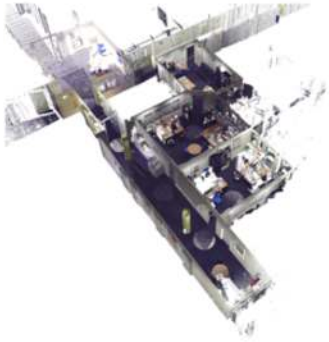
APPROACHES BASED ON INPUT DATA TYPE

- Reconstruction without geometric measures as input source
  - No 3D information is explicitly present (e.g., single RGB image)
  - Geometric information from image features through strong priors
  - Top-down (fitting) or bottom-up (clues assembling)
- Reconstruction from sparse geometric measures as input sources
  - Data fusion techniques to integrate known 3D data and 2D image geometric reasoning
  - Less restrictive priors
- **Reconstruction from dense geometric measures as input sources**
  - **High-density sampling is required to recover high level geometric primitives**
  - **Boundary reconstruction from patch-based representation**

Then we have reconstruction methods from dense geometric measures, in this case a dense sampling is required to recover all high level geometric primitives and to join them into the final model.

SEGMENTATION INTO CO-PLANAR PATCHES

- Dense input for indoor
  - 3D point cloud from laser scanning or RGB-D
  - Problems
    - Size, redundancy and lack of structure



Mura et al. C&G2014

In this context we can have as dense input a sampled 3D point cloud from laser scanning or from RGB-D sensors, which can be very precise and reach billions of geometric measures.

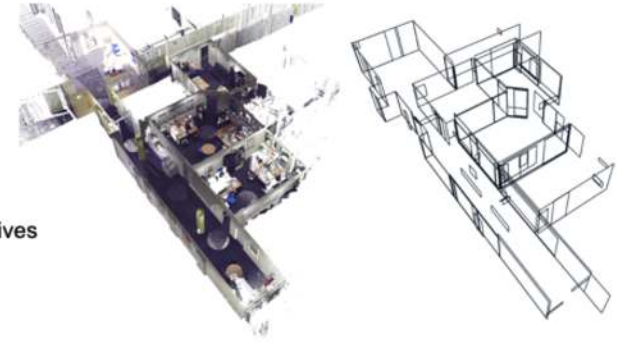
However, these point clouds are not well-suited for reasoning on the structure of the environment and only convey very fundamental information about local geometric features.

To reduce the complexity of subsequent steps in the reconstruction pipeline, in particular the construction of a 2D or 3D space partitioning data structure, it is important to identify the higher-level geometric elements that are the most likely priors for the room boundaries.

Thus we are looking for strong evidence of planar regions which eventually define the final wall locations.

SEGMENTATION INTO CO-PLANAR PATCHES

- Dense input for indoor
  - 3D point cloud from laser scanning or RGB-D
  - Problems
    - Size, redundancy and lack of structure
- Common step
  - Converting the point cloud into higher level geometric primitives
    - RANSAC aggregation (e.g., Jenke 09)
    - Region growing (e.g., Mura 14)
- Different solutions for layout extraction
  - Geometric primitives->boundary structures



Mura et al. C&G2014

A common step for most dense techniques is to convert the input point clouds into simple geometric primitives like planes, indicating potential wall candidates.

Here we have standard approaches such as RANSAC based or region growing based aggregation in order to obtain planar elements.
These planar patches should be free of noise, artifacts and structurally sound to robustly separate the permanent architecture from clutter in a meaningful way.

Hence the point cloud data is commonly converted into a higher-level graph structure that encodes the planar components of the scene, along with their adjacency relations.

Methods usually differentiate themselves starting from this point on how they recover the final layout.

SEGMENTATION INTO CO-PLANAR PATCHES

- Based on observation that man-made structures are dominated by planar parts
- The well-known RANSAC algorithm offers a robust method to efficiently detect planes, spheres, cylinders, cones and tori in point cloud data
  - Many indoor modeling pipelines use it in a pre-process step to identify planar patches
    - Generating plane hypotheses in a randomized way and selecting the best candidate matching the points
    - Directly in 3D (e.g., Ochmann 16) or in a simplified 2D view (e.g., Oesau 14)

Many indoor modeling pipelines effectively use RANSAC in a pre-processing step, either directly in 3D space or in a simplified 2D view of the environment, although this can result in missing regions and in non-deterministic results due to its randomized nature.

SEGMENTATION INTO CO-PLANAR PATCHES

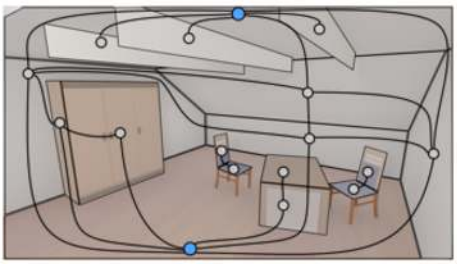
- Based on observation that man-made structures are dominated by planar parts
- The well-known RANSAC algorithm offers a robust method to efficiently detect planes, spheres, cylinders, cones and tori in point cloud data
  - Many indoor modeling pipelines use it in a pre-process step to identify planar patches
    - Generating plane hypotheses in a randomized way and selecting the best candidate matching the points
    - Directly in 3D (e.g., Ochmann 16) or in a simplified 2D view (e.g., Oesau 14)
- Region growing can reduce problems due to missing regions and non-deterministic results
  - Expanding planar patches from seed points based on normal deviation and plane offset (e.g., Mura 14)
  - Less robust but highly applicable for high-quality laser-scanned data

Mura et al. C&G2014

To overcome these issues, some approaches rather opt for a region-growing formulation, in which planar patches are expanded from a set of seed points based on normal deviation and plane offset [MMJV* 14, MMP16]. This less robust yet more systematic way of detecting planar patches is common when the input comes from high-quality laser-scanned data [CLP10, BdLGM14].

STRUCTURALLY SOUND ADJACENCY

- Detected planar primitives enhanced with adjacency relation
  - Graph based on their spatial proximity and configuration
  - Extract cuboids from six adjacent wall planes (e.g., Jenke 09, Murali 17)



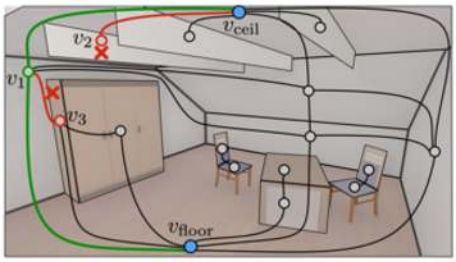
Mura et al. CGF2016

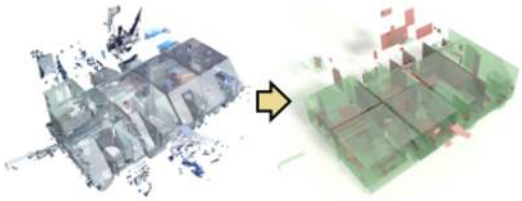
Regardless of the specific approach used, the detected primitives are often arranged in an adjacency graph based on their spatial proximity and mutual configuration.

This can for example be exploited to detect cuboids formed from six adjacent wall planes detected as a special configuration in the adjacency graph of primitives, which are then used to reconstruct the bounding surfaces.

STRUCTURALLY SOUND ADJACENCY

- Detected planar primitives enhanced with adjacency relation
  - Graph based on their spatial proximity and configuration
  - Extract cuboids from six adjacent wall planes (e.g., Jenke 09, Murali 17)
- Encode structurally plausible relations between patches
  - Permanent structures arranged in stable configuration conforming to a set of structurally valid patterns (e.g., Mura 16)
    - Filter out unnecessary and non-plausible parts
    - Reduces size and complexity of subsequent reconstruction steps
    - Improves on robustness to clutter and outliers





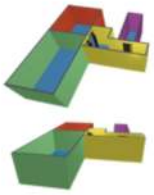
Mura et al. CGF2016

Using the intuition that permanent structures carry the physical load from top to bottom in a consistent and stable configuration, a number of structurally valid adjacency relation patterns
have been proposed by Mura et al. which define valid transitions from one planar patch to another.

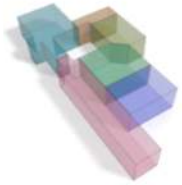
The non-plausible configurations lead to the removal of planar primitives with the effect of reducing the computational complexity of subsequent stages in the reconstruction pipeline.
However, in complex environments this type of structural analysis can also increase the robustness to clutter and outliers.

FROM GEOMETRIC PRIMITIVES TO ROOMS

- Approaches differ considerably in how to use the input to extract the boundary surface of rooms
  - Given the decomposition of the input into simple geometric primitives such as planar patches
- Depending on the assumption of the targeted room shapes and quality of input
  - Recap typical structural priors:



Manhattan world (all right angles)



Atlanta world (horizontal floors/ceilings)

Vertical walls (sloping floors and ceilings)

Piecewise planar

12

Given a decomposition of the input into simpler geometric primitives, existing approaches differ considerably in the way such primitives are used to extract the boundary surface of the rooms.

In general, the complexity of the technical solutions adopted depends on the assumptions made on the quality of the input data and on the shape of the rooms.

VOLUMETRIC BOUNDARY RECONSTRUCTION

- Clean and complete input under the Manhattan world prior allows efficient reconstruction using the union of one of more cuboids
  - Cuboids formed by intersection of six wall planes detected as special configuration from adjacency graph of primitives (e.g., Jenke 09, Murali 17)

Some early approaches exploit the Manhattan world prior and voxel based representations, which allow only right angles between adjacent wall planes. Thus this requires the identification and alignment of the Cartesian coordinate grid to the dominant directions in the data.

In this context, for relatively clean and complete inputs, rooms can be reconstructed as the union of one or more cuboids, each obtained by intersecting a group of six adjacent wall planes detected as a special configuration in the adjacency graph of primitives.

VOLUMETRIC BOUNDARY RECONSTRUCTION

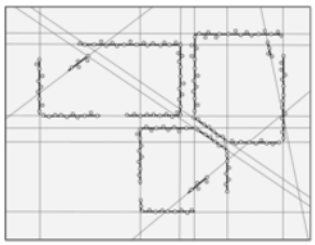
- Clean and complete input under the Manhattan world prior allows efficient reconstruction using the union of one of more cuboids
  - Cuboids formed by intersection of six wall planes detected as special configuration from adjacency graph of primitives (e.g., Jenke 09, Murali 17)
  - Restricting cuboids to equal sides allows a reconstruction based on a voxelization of the input space
  - E.g. voxel carving along the lines of sight as implied by the scanner (e.g., Turner 13)

Restricting the cuboids to have equal sides of a pre-determined length results in a voxel-based reconstruction. In this case, it is possible to devise more specialized approaches that work directly on a voxelization of the input scene, for instance by extracting the internal volume of a room by carving out voxels that intersect the lines of sight from scanned points to the positions of the acquisition device (if available in the input model)
3d rasterization

However, these methods typically result in a blocky reconstruction of all permanent structures that are not nicely aligned with the axes of the Cartesian grid.

ARRANGEMENT OF HYPERPLANES

- Avoid hard constraints of aligning walls with the Cartesian grid for more flexible room layouts
- 2.5D structure and Atlanta World prior
  - With arbitrary angles between adjacent vertical walls and horizontal floors and ceilings



dominant planes from input geometry
form a line arrangement and 2D cell complex

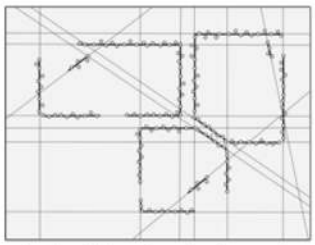
This representation is very constraint, so many methods relax the restrictions to a 2.5 dimensional structure which allows for a more flexible extraction of the indoor architecture.

This is achieved by allowing the piecewise planar room boundaries to align in more general configurations than just the major axis directions.

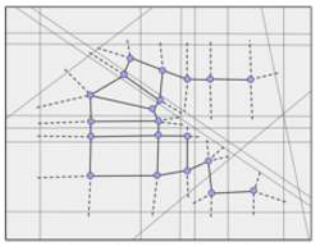
The Atlanta world assumption with arbitrary angles between vertical wall planes but horizontal floors and ceilings corresponds to this scenario.

ARRANGEMENT OF HYPERPLANES

- Avoid hard constraints of aligning walls with the Cartesian grid for more flexible room layouts
- 2.5D structure and Atlanta World prior
  - With arbitrary angles between adjacent vertical walls and horizontal floors and ceilings
- Adjacency relations maintained in a 2D cell complex
  - Convex regions partitioning by hyperplanes
  - Correspondence between structures and planes



dominant planes from input geometry
form a line arrangement and 2D cell complex



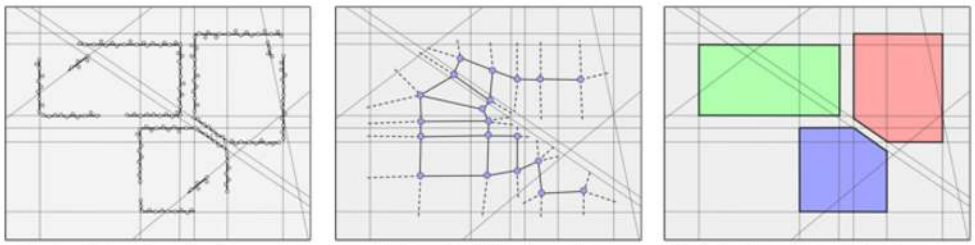
dual graph of 2D cell complex
captures spatial adjacency relations

The vertical walls allow the pipeline to work in a planar projection and to subdivide the floor space surrounding the input model into convex regions formed by an arrangement of hyperplanes in 2D.

These convex regions form a 2D cell complex with corresponding adjacency relations, and establish a correspondence between structure and planes.

ARRANGEMENT OF HYPERPLANES

- Avoid hard constraints of aligning walls with the Cartesian grid for more flexible room layouts
- 2.5D structure and Atlanta World prior
  - With arbitrary angles between adjacent vertical walls and horizontal floors and ceilings
- Adjacency relations maintained in a 2D cell complex
  - Convex regions partitioning by hyperplanes
  - Correspondence between structures and planes
- 3D model by labelling cell complex and extrusion from 2D floor plan



dominant planes from input geometry form a line arrangement and 2D cell complex

dual graph of 2D cell complex captures spatial adjacency relations

2D footprint and boundaries of rooms given by union of labelled cells

17

The shape of rooms or the entire indoor environment is eventually defined by the union of cells, classified and labelled in the dual graph to represent the interior space.

Eventually, the 3D model can be formed by extruding the 2D floor plan boundaries along the vertical axis to the height of the rooms.

2.5D RECONSTRUCTION

- Lines can be detected by 2D projection of the point cloud and line fitting
  - Multi-scale line fitting (e.g., Oesau 14)
- Lines induced from 3D primitives, planar patches
  - Considering minimum surface area (e.g., Ochmann 16)
  - Analyzing occlusion shadows to reduce false-positives (e.g., Mura 14)
- Formation of 2D line hypothesis for vertical wall candidates must be robust to clutter and occlusions
  - Line normals orthogonal to vertical axis direction
  - Clustering line orientations (e.g., Oesau 14, Mura 14)

Mura et al. C&G2014

Based on this concept, we have several different approaches which vary in how the hyperplanes are selected from the input geometry matching the architectural surfaces.

Using 2D line fitting techniques, the input point samples can directly be projected onto the floor. Alternatively, the point cloud data can be filtered and processed in 3D to detect vertical planar patches which then induce the line arrangement and 2D cell complex partitioning in the floor plane.

Basic robustness against noise is commonly achieved by filtering and clustering the normal and line orientations.

2.5D RECONSTRUCTION

- Lines can be detected by 2D projection of the point cloud and line fitting
  - Multi-scale line fitting (e.g., Oesau 14)
- Lines induced from 3D primitives, planar patches
  - Considering minimum surface area (e.g., Ochmann 16)
  - Analyzing occlusion shadows to reduce false-positives (e.g., Mura 14)
- Formation of 2D line hypothesis for vertical wall candidates must be robust to clutter and occlusions
  - Line normals orthogonal to vertical axis direction
  - Clustering line orientations (e.g., Oesau 14, Mura 14)
- Generic 2D shapes from 3D primitives
  - Cell complex as a 2D triangulation (e.g., Turner 14)
  - Curved-line fitter on the horizontal plane (e.g., Yang 19)

Mura et al. C&G2014

Extensions to more generic shapes of the boundary walls as well as non-planar wall structures have been proposed, for example through the us of a Delaunay triangulation of the cell complex or curve fitting techniques.

RASTERIZED FLOOR PLAN

- Voxelization of indoor space with per-voxel surface or free-space evidence indicators
  - Room segmentation in 2D by clustering projected voxel indicators from RGB-D panoramas (e.g., Ikehata 15)

Further methods use some form of a rasterization of the floor plan, also resulting in a 2.5D room boundary reconstruction.

An example exploiting RGB-D input panoramas by Ihehata et al. is basically a free-or-occupied space analysis. In this case solved through the voxelization of space and projection per-voxel free space indicators into the ground plane, where a room segmentation through clustering is performed.

RASTERIZED FLOOR PLAN

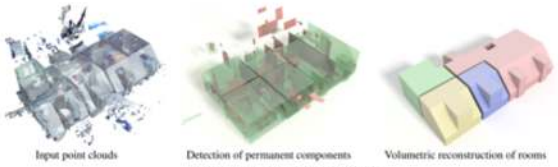
- Voxelization of indoor space with per-voxel surface or free-space evidence indicators
  - Room segmentation in 2D by clustering projected voxel indicators from RGB-D panoramas (e.g., Ikehata 15)
- Data-driven neural network based approaches
  - Conversion of RGB-D scans into rasterized floor plan with {I,L,T,X}-junction type geometry predictions (e.g., Liu 18)
    - Raster-to-vector aggregation into Manhattan world boundary lines
  - Optimization problem on planar graph from 2D point density/normal map to form multiple polygonal loops bounding the rooms (e.g., Chen 19)
    - Considering data discrepancy, consistency and model complexity terms

Also data-driven approaches have been proposed based on trained neural networks. Also starting from RGB-D data, Liu et al. perform a rasterization of the floor plan into pixel-wise predictions of floorplan geometry. Based on the junction-indicators of wall segments, a raster-to-vector transformation eventually produces the boundary outlines.

Stargin from a combined point-density and normal map, Chen et al. reconstruct a floorplan graph through formulating the task as an optimization problem. The objective function penalizes discrepancies between the input measurements, inconsistencies in the polygon loops as well as model complexity.

FULL 3D RECONSTRUCTION

- Complete 3D cell complex from planar 3D primitives
  - Generic convex polyhedral models (e.g., Mura 16)
    - Structurally coherent set of 3D dominant planes from planar patches
      - plane hypothesis limitation for reduced computational cost
    - Multi-class label optimization to assign cells to rooms
    - Supporting sloped wall boundaries and multi-story environments



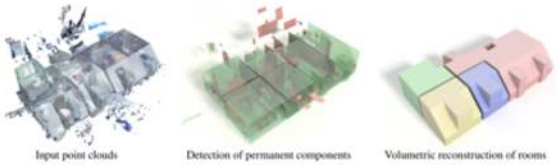
Mura et al. CGF2016

More recent approaches lift the 2.5D restriction and can deal with a complete 3D cell complex formed from the detected planar primitives representing the possible wall locations.

For example, Mura et al. extract a set of 3D planes from planar patches and then form a multi-label optimization problem to assign cells to individual rooms.
A key contributions is to use an analysis of structural patterns to limit the number of plausible plane hypothesis to reduce the complexity of the hyperplane arrangement and 3D cell complex.
As a result, sloped walls and multi-story configuration of environments can be handled.

FULL 3D RECONSTRUCTION

- Complete 3D cell complex from planar 3D primitives
  - Generic convex polyhedral models (e.g., Mura 16)
    - Structurally coherent set of 3D dominant planes from planar patches
      - plane hypothesis limitation for reduced computational cost
    - Multi-class label optimization to assign cells to rooms
    - Supporting sloped wall boundaries and multi-story environments
  - Layered 3D complex by vertically stacking multiple 2D complexes (e.g., Ochmann 19)
    - Allows reconstruction of rooms having multiple floors and ceilings
    - Only horizontal and vertical orientations

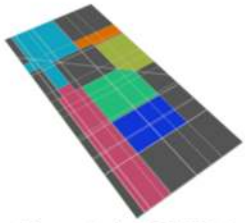


Mura et al. CGF2016

Similarly supporting multi-story is the approach by Ochmann et al. who build a layered 3D cell complex by stacking multiple 2D complexes.
This allows the reconstruction of indoor environments with multiple floors and ceilings but still only allows horizontal or vertical wall orientations.

LABELING OF CELL COMPLEX

- Identify the interior space and possibly individual rooms by inside/outside or multi-room labelling
- Clustering of cells by a distance metric
  - Measure of the likelihood that cells belong to the same room → low distance
  - k-medoids clustering from over segmented point cloud (e.g., Mura 14, Ikehata 15)



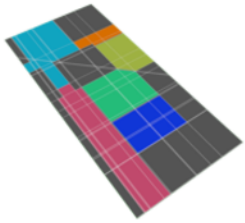
Mura et al. C&G2014

Once the cell complex has been constructed from the 2D or 3D hyperplane arrangement, a fundamental question is how to select the groups of cells that correspond to the inside or outside and in fact to each room.

A natural option is to apply a clustering algorithm to the set of cells, based on a metric that ensures that cells belonging to a same room have low distance.

LABELING OF CELL COMPLEX

- Identify the interior space and possibly individual rooms by inside/outside or multi-room labelling
- Clustering of cells by a distance metric
  - Measure of the likelihood that cells belong to the same room → low distance
  - k-medoids clustering from over segmented point cloud (e.g., Mura 14, Ikehata 15)
- Energy minimization to better support controllable regularization
  - Data term quantifying the error for assigning a room-label to a cell
  - Smoothness term penalizing incoherent labelling
  - Binary in/out labelling using data area coverage or visibility measures (e.g., Budroni 10, Oesau 14)
    - Efficient exact solution using combinatorial algorithms



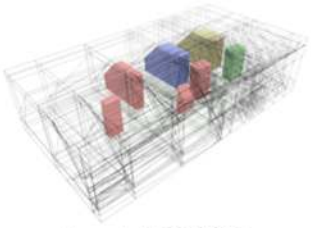
Mura et al. C&G2014

Since a controllable regularization is difficult to integrate in clustering techniques, a majority of methods cast the cell-to-room assignment as an energy minimization problem.
The data term is typically based on an initial guess of the most likely label, while the smoothness terms is used to avoid jumps in the labeling of neighboring cells, hence avoiding jagged or implausible boundaries.

Note that binary inside-outside labelling can be solved by an energy function that can be optimized efficiently and exactly using combinatorial algorithms.

MULTI-LABELING OF CELL COMPLEX

- Assigning different labels can be done efficiently using Markov random field and graph-cut multi-class labelling optimization algorithms
  - Data term based on visibility of view positions or some pre-assigned geometry
  - Smoothness term used to penalize structurally inconsistent room shapes and ill-defined boundaries between room (e.g., Mura 16, Ochmann 19)



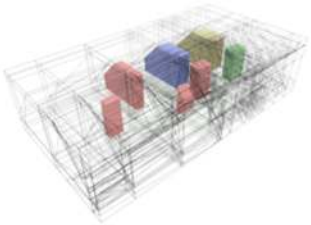
Mura et al. CGF2016

The assignment of cells to multiple rooms can also be expressed as a multi-label energy minimization problem.

The smoothness or regularization term has been extended to favor structurally consistent room configurations and ensure solid walls separating the rooms.

MULTI-LABELING OF CELL COMPLEX

- Assigning different labels can be done efficiently using Markov random field and graph-cut multi-class labelling optimization algorithms
  - Data term based on visibility of view positions or some pre-assigned geometry
  - Smoothness term used to penalize structurally inconsistent room shapes and ill-defined boundaries between room (e.g., Mura 16, Ochmann 19)
- Commonly used combinatorial multi-label optimization techniques only yield approximations of the globally optimal solution
  - No noticeable decrease of output model quality in practice (e.g., Mura 16, Ochmann 16, Ambrus 17)

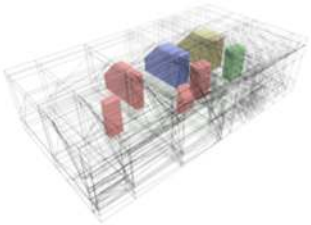


Mura et al. CGF2016

While multi-label optimizations using combinatorial techniques only yield an approximation of the globally optimal solution in this case, the experimental results do not show a limited quality in practical scenarios.

MULTI-LABELING OF CELL COMPLEX

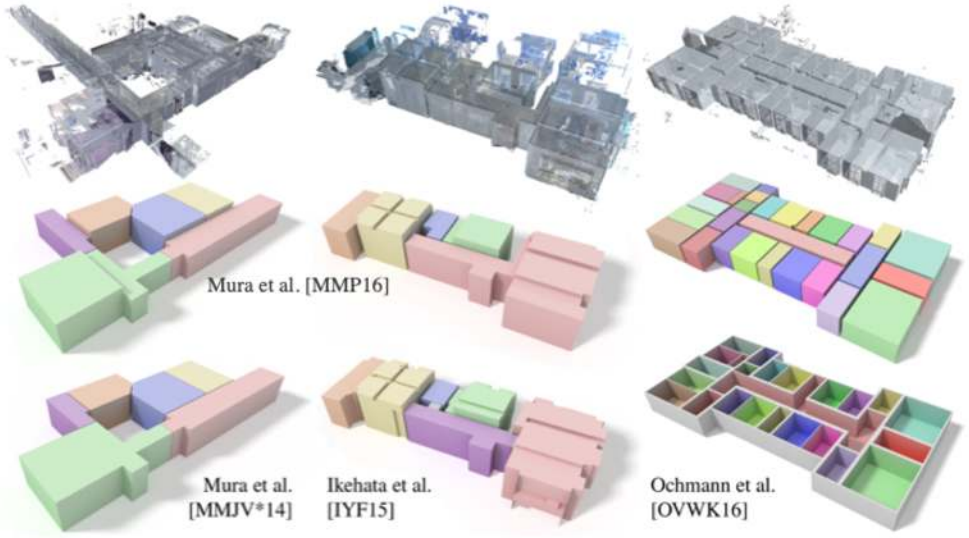
- Assigning different labels can be done efficiently using Markov random field and graph-cut multi-class labelling optimization algorithms
  - Data term based on visibility of view positions or some pre-assigned geometry
  - Smoothness term used to penalize structurally inconsistent room shapes and ill-defined boundaries between room (e.g., Mura 16, Ochmann 19)
- Commonly used combinatorial multi-label optimization techniques only yield approximations of the globally optimal solution
  - No noticeable decrease of output model quality in practice (e.g., Mura 16, Ochmann 16, Ambrus 17)
- Recently integer linear programming (ILP) solution proposed
  - Can support for set of structural rules as hard constraints (e.g., Ochmann 19)



Mura et al. CGF2016

Nevertheless, a notable exception has recently been proposed by Ochmann et al. who propose an integer linear programming solution which is computationally efficient and can incorporate hard constraints, for example enforcing solid walls between separate rooms.

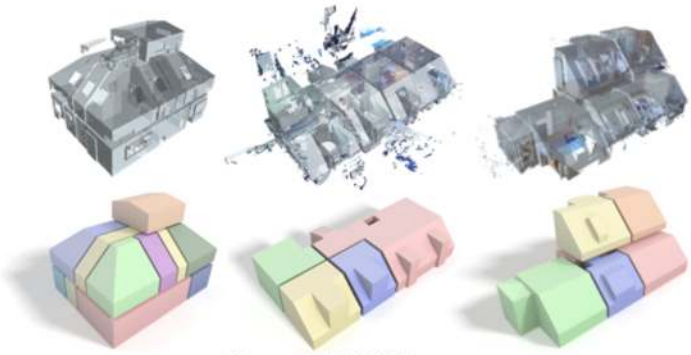
RESULTS FOR 2.5D ENVIRONMENTS



These examples for boundary reconstruction from dense point cloud data demonstrate that state-of-the-art approaches work very well in real-world cases where the 2.5D or Manhattan world assumptions are actually applicable.

Complete room layouts can be recovered, as well as varying ceiling heights using some of the methods.

RESULTS FOR 3D ENVIRONMENTS



Mura et al. [MMP16]

In cases where the 2.5D and Manhattan world assumptions are clearly violated, fully 3D approaches can reconstruct arbitrary wall and ceiling orientations which cannot be recovered using the more-constraint methods.

In particular, these approaches show the ability to reconstruct very fine differences in orientations of the permanent structures, of walls as well as the roof lines.

DISCUSSION (1/2)

- Modeling room boundaries is a major challenge in indoor reconstruction
  - High level of noise, clutter, and missing data->heavy use of priors
- Almost all solutions works totally or partially in 2D or 2.5D
  - Extracting floorplan and extending it to 3D
  - Sometimes special handling for ceiling
- Common assumptions
  - Large planar surfaces…
  - Vertical walls…
    - … ok for the majority of buildings
  - Horizontal floor and ceiling
    - … much more limiting!

So to summarize, modeling the room boundaries usually requires the use of priors.

Many solutions work in the 2D and 2.5D domain and using common assumptions like that one expects large planar surfaces, vertical walls and often assuming horizontal floors and ceilings.

DISCUSSION (2/2)

- Open problems
  - Fully 3D approaches
    - Handle complex overhanging structures, floors and ceilings
  - Free-form 3D interiors
    - Also remove large planar surface assumptions
- RGB-D cameras could help fusion between CV and CG approaches
- Recent data-driven methods are promising to increase convergence
  - E.g., 3D plane and depth estimation from images
    - Total3DUnderstanding – CVPR2020

In contrast, fully 3D approaches still pose major challenges and only few have been proposed. Another major problem is to detect and reconstruct free-form 3D interiors.

In this context, RGB-D cameras could help in data fusion and also recent data-driven methods are promising to increase the convergence between computer graphics and computer vision techniques.


SIGGRAPH
THINK
BEYOND
2020 19-23 JULY WASHINGTON DC
BREAK
© 2020 SIGGRAPH. ALL RIGHTS RESERVED.


SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC
BOUNDING SURFACES RECONSTRUCTION / 2
Speaker: Giovanni Pintore
CRS4

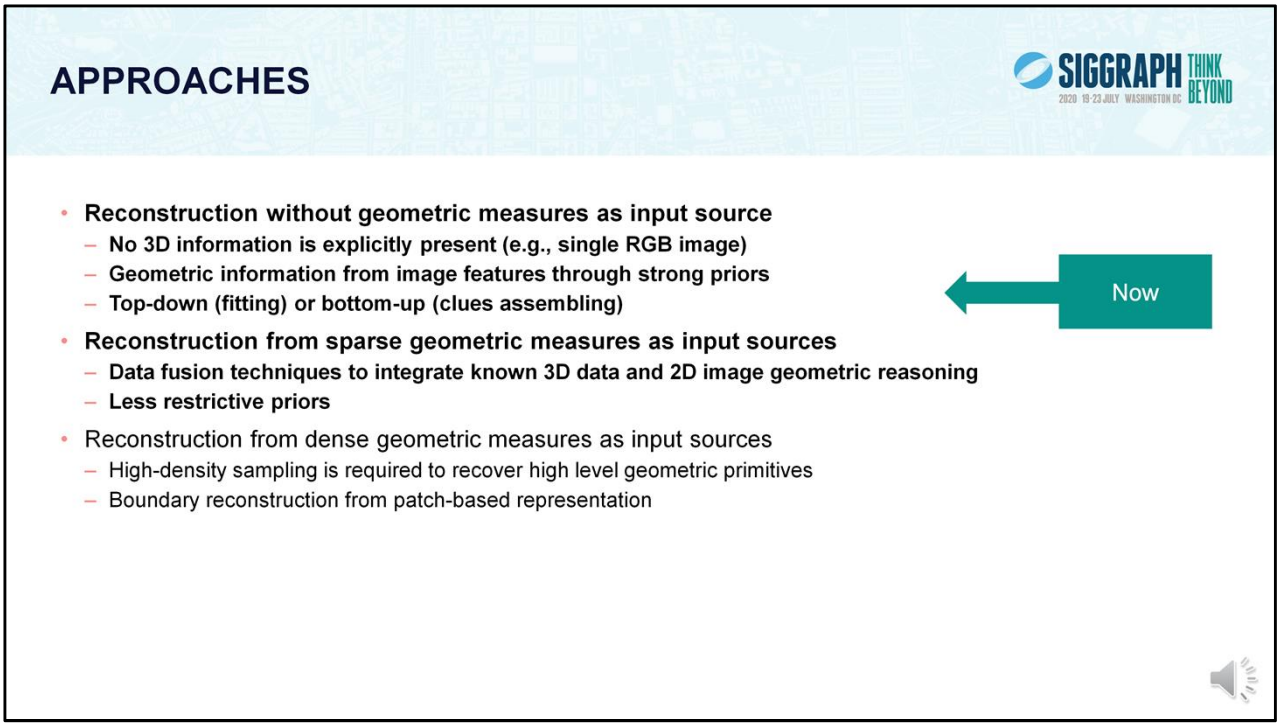
APPROACHES

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **Reconstruction without geometric measures as input source**
  - No 3D information is explicitly present (e.g., single RGB image)
  - Geometric information from image features through strong priors
  - Top-down (fitting) or bottom-up (clues assembling)
- **Reconstruction from sparse geometric measures as input sources**
  - Data fusion techniques to integrate known 3D data and 2D image geometric reasoning
  - Less restrictive priors
- **Reconstruction from dense geometric measures as input sources**
  - High-density sampling is required to recover high level geometric primitives
  - Boundary reconstruction from patch-based representation

As we have seen before the break, the approaches can be usually summarized according to their input.
We have already discussed the reconstruction from dense geometric input

APPROACHES



In this specific part of this talk, we focus instead on the first two options.
In the first case, we have methods that reconstruct the layout without geometric measures available, that is the case of single image methods.
Here the geometric information is obtained from image features assuming strong priors.
Then we have methods that recover the structure from sparse geometric information, like in the case of data fusion or structure from motion.
In this case we have less restrictive priors.
These, in many ways, are very important in real-world applications, since obtaining a real dense and uniform coverage, in terms of geometric samples, is not trivial.
Moreover the modern trend in terms of acquisition is to use rgb-d devices that combine high-definition visual data with less dense depth data.

RECONSTRUCTION WITHOUT GEOMETRIC MEASURES: LAYOUT FROM A SINGLE RGB IMAGE

- **Early approaches (perspective)**
  - Heavy constrains
  - **Delage 06**: floor-wall boundary for each image column
    - Floor-Wall (FW) model
    - Partial model
  - **Hedau 09**: geometric context (GC) for indoor scene
    - Cuboid (CB) prior
    - Room box and surface labels jointly estimated (floor, ceiling, wall, objects)
  - **Lee 09**: orientation maps (OM) from MW vanishing lines
    - Indoor World Model (IWM) geometric reasoning
    - Manhattan world planes bounding the room

Starting from the first option, we have early methods from a single perspective image and based on heavy constrains.
For example the approach of Delage, recover the floor-wall boundary form each image column, and then returns an indoor model composed by a single floor and the main walls.
Later we have the introduction of geometric context concept, where a cuboid-like model of the room is recovered from surface labels assigned to the image,
and then we have orientation maps method based on Manhattan World vanishing lines, returning Manhattan planes bounding the room.

RECONSTRUCTION WITHOUT GEOMETRIC MEASURES: LAYOUT FROM A SINGLE RGB IMAGE

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **Geometric context (GC) and Orientation Map (OM)**
  - Basis of indoor geometric reasoning
- **Geometric reasoning on the IWM**
  - **Flint 10**: dynamic programming
  - Horizontal floor and ceiling related by a homography
- **Geometric reasoning on panorama**
  - Panoramic image is converted into virtual perspective images (e.g, cubemaps)
  - GC and OM applied
  - Result re-projected on the original panorama

Geometric context and orientation maps are the basis even of modern indoor geometric reasoning.
For example, Flint proposes an efficent dynamic programming approach based on GC and OM, assuming that horizontal floor and ceiling
Are related by an homography.
Such kind of geometric reasoning can be extended also to panoramic images:
In this case input sheremaps are convered into virtual perspsective, and then results are projected back to sphere.

RECONSTRUCTION WITHOUT GEOMETRIC MEASURES: LAYOUT FROM A SINGLE RGB IMAGE

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **Limitations of perspective images**
  - Restricted FOV -> limited geometric context
  - Small and simple scenes
- **Emerge of full (360°x 180°) panoramic format**
  - Scene captured with one or at least few shots
  - Adopted also for RGB-D capture
- **Whole-room 3D context from panorama**
  - **Zhang 14**: combines GC and OM to fit 3D bounding box of the room and of all major objects inside

Indeed the main limitation on single image is given by the field of view, that results in a limited geometric context,
So in the recent years research focused in expoliting full panoramic format images. With such a format a scene can be captured with just one or at least few shots, moreover this format has been also adopted for modern geometry capture devices, Such as RGB-D cameras.
One of first and prominent work in this field is the work of Zhang, where GC and OM are combined to recover room bouding box joined with the majors obtects inside.

RECONSTRUCTION WITHOUT GEOMETRIC MEASURES: LAYOUT FROM A SINGLE PANORAMA

SIGGRAPH THINK BEYOND 2020 19-23 JULY WASHINGTON DC

- **Geometric reasoning approaches**
  - Yang 16: 3D shape from oriented super-pixel facets
    - Manhattan World locally applied to oriented facets (embedding GC+OM)
  - Pintore 16: top-down 2D domain by E2P transform
    - Atlanta World assumption: horizontal ceiling and floor; vertical walls but free walls layout
- **Data-driven approaches**
  - MW rectification exploiting GC+OM (pre-processing)
  - 2D image segmentation (encoder-decoder scheme)
  - MW regularization (post-processing)
  - LayoutNet 18: corners position as sparse features on the image
  - DulaNet 19: room shape from E2P and panorama
  - HorizonNet 19: corners position from 1D encoding of panorama

When dealing with panoramic images, first we have methods based on geometric reasoning, like the works of Yang and Pintore.
In the first case the room shape is obtained from a set of oriented super-pixels facets under canonical Manhattan World constrain,
in the second case the problem is solved in a top-down 2D domain using a specific spatial transform, adopting the less restrictive Atlanta World model.
More recently arise many data-driven approches, and i some ways this is a very active research field, leading to impressive results in terms of accuracy and speed.
All these methods have a common pipeline: first they perform a MW rectification of the image, based on the well know GC and OM,
then they segment the rectified image through an encoder-decoder scheme, and finally the recover the layout after a regularization of the result.
Some prominent methods are: Layoutnet, where the corner positions are obtained as sparse features, Dulanet, which employes the same E2P transform of Pintore 2016,
And Horizonnet, where the corners are obtained from an 1 dimensional encoding of the panorama.

RECONSTRUCTION FROM SPARSE GEOMETRIC MEASURES

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- Single pose limits
  - Important structures must be visible from a single point-of-view
  - Multiple registered images
  - Indoor problems
    - Untextured surfaces -> sparse 3D features
    - Fatal occlusions
- Main approaches
  - Sparse features->MW densification->volumetric fusion
    - E.g., Furukawa 09
    - Unstructured 3D mesh
  - Combining single and multi-view analysis
    - E.g., Flint 11

Indeed working with a single pose has limits. For example all the important structures of the room must be visible from a single point of view.
To this end, more recent research is focused on exploiting multiple poses, possibly registered between them.
Dealing with multiview in the indoor environment is very challenging, since we have poor untextured surfaces, fatal occlusion and complex visibility reasoning,
So is not possible to have a dense nad regular sampling just from images in the indoor.
So, common approaches usually follow two approaches to recover a 3D model: the first one try to densify the sparse multiview features using MW stereo and then using a volumetric fusion approch to find the model. However such pipelines usually return unstructured meshes.
A second option instead, is a data fusion approach, which try to combine the sparse 3D features with single image analysis.
So, since we are interested in structured models, we only discuss this second option

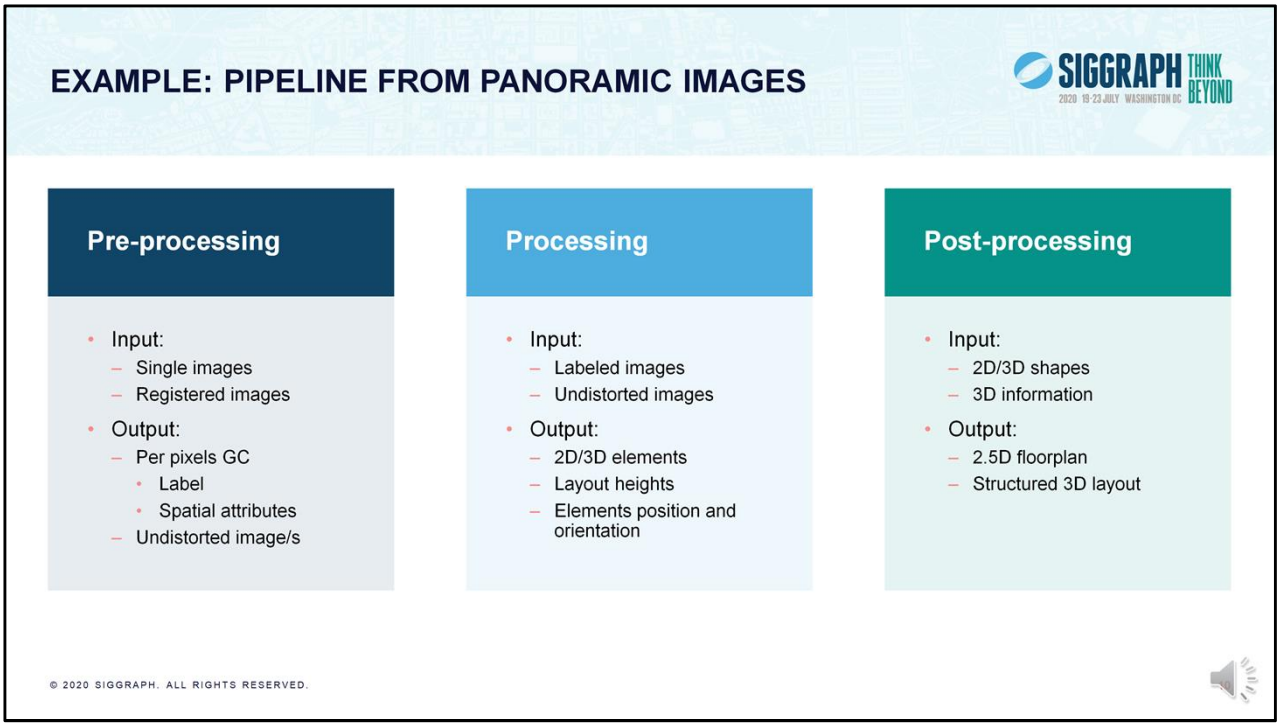
RECONSTRUCTION FROM SPARSE GEOMETRIC MEASURES

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- Combining single and multi-view analysis
  - Cabral 14: geometric reasoning on single panorama to integrate semi-dense point-cloud
    - MW piece-wise planarity
    - Externally calculated point cloud from MW-MVS needed
    - 3D anchor points from single image geometric reasoning
  - Pintore 18: 3D facets representation from registered panoramas
    - Assuming VW (vertical walls): less restrictive than MW
    - Sparse 3D features recovered by panoramas registration
    - E2P transform locally applied to each super-pixel
      - 2D super-pixel -> 3D facet
      - 3D Facets from different images are joined to identify layout

Pintore et al. CGF 2019

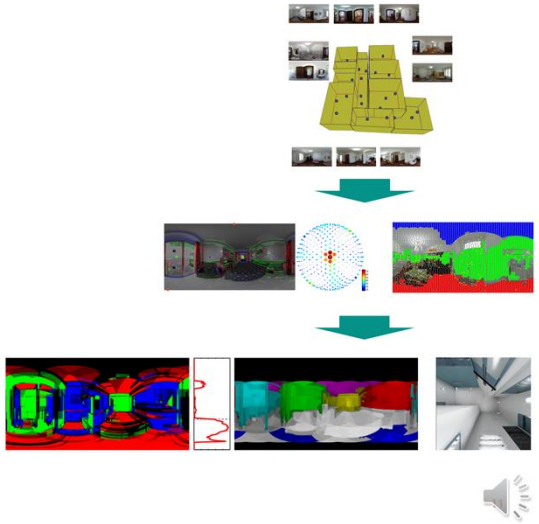
Here we have some prominent example: in the work of Cabral the geometric reasoning on single panoramas is exploited to integrate a densified point cloud.
They adopt MW piecewise planarity as assumption, exploiting image data to complete the holes in the externally calculated point cloud.
In the other example, insted, a 3D facets representation is extraceted directly form the panoramic images, which are registered in common floorplan space.
they assume only vertical walls, so the can reconstruct also sloped ceiling and walls with not right angles.
In this work 3D information is recovered directly form image registration and a spatial transform is locally applied to each super pixel to transform it in a 3D facet,
So 3D facets from different images are joined toghether.



Once we have seen an overview of methods, we introduce an example of application, and in particular a typical pipeline to recover rooms layout from a collection of panoramic images.
A common structure for such a pipeline includes a pre-processing step, before the proper model processing and then a post-processing final step.
As an input we can have a single panoramic image covering the scene, or, for larger and complex spaces, a set of registered images.
As output of the pipeline instead we expect to have a 2.5D floorplan, or in general a structured 3D layout composed by one or more rooms.

PIPELINE EXAMPLE: PRE-PROCESSING

- **Input**
  - One single equirectangular image
  - Multiple registered images
    - MV registration information
      - Common reference frame / 3D features
- **Scheme**
  - First step: low-level features extraction
  - Second step: high-level features computation
- **Output**
  - Per pixels geometric context
    - Labeled super pixels, Orientation maps
    - 2D/3D attributes
  - Undistorted images
    - Warped panorama
    - E2P/A2P transform

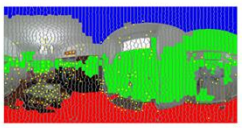


In the pre-processing step we have as input one single equirectangular image covering the environment, or multiple registered images.
In this second case we also have available multi-view information, like a reference frame for each camera and a bunch of 3D features.
The typical pre-processing step is similar to dense 3D data processing, but here starting from 2D input, so that we extract low-level features from images and we then we aggregate them in to high-level features and into spatial elements.
As output we obtain a per pixel geometric context, that is spatial attributes for each point of the image.
Additionally many methods require a trasformation of the image, in order to adapt equirectangular projection to specific priors.

PIPELINE EXAMPLE: PRE-PROCESSING

- **Low-level features extraction**
  - **Edge maps**
    - Vanishing lines aligned with MW **Zhang 2014, Yang 2016**
    - Room edges projected to floorplan **Pintore 2016**
      - **Provided geometric context and/or orientation maps (GC+OM)**
  - **Superpixels**
    - **Cabral 2014, Yang 2016, Pintore 2016, Pintore 2018, Yang 2018**
  - **Multi-view features**
    - Too sparse for dense reconstruction
    - Sparse 3D information associated **Cabral 2014, Pintore 2018**
  - **Spatial transforming**
    - Different domain different features
    - Cubemaps from equirectangular **Zhang 2014, Yang 2016**
    - E2P transform **Pintore 2016, Pintore 2018, Yang 2019**

Pintore 2016: edges transform

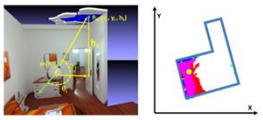


Pintore 2018: superpixels and MV features

The pre-processing step starts extracting low level features from the image.
Common features are edge maps, to identify vanishing Manhattan lines or to identify room edges and corners projected to the floorplan.
Edges are usually exploited to provide geometric context and orientation maps, as seen in previous methods.
Then we have super-pixels, based on the assumption that there is a relationship between color distribution and spatial properties, and then we have multi-view features associated to sparse parts of the images.
It is also common when dealing with panoramic images to perform spatial transforming on the image itself.
This is common for for two main reasons: first, because fundamental methods like geometric context and orientation map estimation are basically targeted to perspective images,
So many methods convert equirectangula images to a set of virtual perspective images, calculate the features, and project back the result to equirectangular space.
Other approch instead applied priors like atlanta or vertical walls to work directly on a floorplan space where room structure is highlighted.

PIPELINE EXAMPLE: PRE-PROCESSING

- **High-level features computation**
  - Low-level features aggregation by priors
  - **Labeling propagation**
    - Manhattan World **Cabral 2014**
    - Atlanta World **Pintore 2016, Pintore 2018**
  - **Vanishing lines – super pixels merging**
    - Manhattan World **Yang 2016**
  - **Multi-view features – super pixels merging**
    - Manhattan World **Cabral 2014**
    - Vertical walls **Pintore 2018**
  - **Planes and shapes**
    - Manhattan World **Cabral 2014, Yang 2019**
    - Vertical walls **Pintore 2018**



Pintore 2018: MV / SP merging

In the second step of pre-processing low-level features are aggregated by using typical indoor priors.
For example several approaches propagate ceiling, wall and floor labeling to each pixel of the image, assuming that the upper part of the image is surely ceiling, the bottom floor and the middle part wall.
Then geometric properties or measures are associated to image patches, for example Manhattan World segments, or multi-view 3D features.

PIPELINE EXAMPLE: PRE-PROCESSING

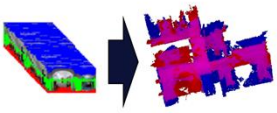
- **Output**
  - Per pixel geometric context
    - **2D/3D attributes**
    - Labeled super pixels **Cabral 2014**
      - Ceiling, wall, floor
    - Orientated maps **Zhang 2014**
    - Oriented facets **Yang 2016, Pintore 2018, Yang 2018**
  - Undistorted images
    - Warped panorama **Zou 2018, Yang 2019, Sun 2019**
    - E2P/A2P transform **Yang 2019, Pintore 2020**

Pintore 2020: A2P transform

As a results we obtain a per pixel geometric context providing 3D properties for each poin tof the image, and associated oriented facets, usually aligned with the reference indoor model.
Optionally several methods adopt such spatial information to transform the original equirectangular image.
As, for example, for data driven methods which expect images aligned with the three main MW axes.

PIPELINE EXAMPLE: PROCESSING

- **Input**
  - Labeled images
  - Undistorted images
- **Approaches**
  - Fitting, optimization, dynamic programming,etc.
    - Often multiple images
  - Data-driven
    - Performing, single image
- **Output**
  - 2D/3D elements
  - Layout height
  - Elements position and orientation

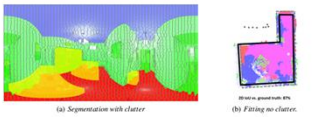


Pintore 2018: facets merging

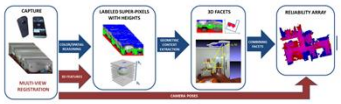
Then, starting from the pre-processed images, different approaches can be exploited to recover the main elements of the structure, such as room shape, the layout height or the rooms position. These generally alternative approaches can be classified in the conventional optimization methods or in the more recent data-driven approaches. The first branch is more often used in the case of multiple images combined, while the second, although more performing, is usually used for single images.

PIPELINE EXAMPLE: PROCESSING

- **Optimization and fitting**
  - Simplified segmentation into **wall**, **ceiling**, **floor** super-pixels/facets
  - constraint graph of facets
  - **Single view**
    - 3D information from **Manhattan World** GC+OM **Yang 2016**
    - Shape is defined in a 2D top-down domain under **Atlanta World** **Pintore 2016**
  - **Multi-view**
    - 3D information from MV features **Cabral 2014, Pintore 2018**
      - **Manhattan World** prior
        - **wall** super-pixels used as 2D anchor points **Cabral 2014**
      - **Vertical walls** prior
        - 2D footprint of **ceiling**/**floor** facets merging **Pintore 2018**
  - **Output**
    - 2D footprint, 2D corners, heights



Pintore 2019: 2D footprint of ceiling&floor SP



Pintore 2018: MV merging

So, the first approach is usually based on optimization methods. We start from a simplified segmentation into wall, ceiling a floor facets, then a constrained graph of facets is exploited to recover each room.
Single view approaches recover the 3D information applying Manhattan World or less restrictive Atlanta World, while multi view approaches recover 3D from sparse multi-view features.
For these multi image methods the 2D fooprint of the room can be recovered just projecting and merging ceiling and floor facets.
As a result we obtain the 2D floorplan of one or more rooms, coupled with the ceiling heights.

PIPELINE EXAMPLE: PROCESSING

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

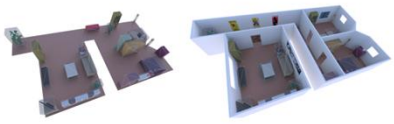
- **Data-driven approaches**
  - **Expected input**: single image aligned to MW directions
  - **Ground truth/network output**
    - **LayoutNet 2018, HorizonNet 2019**
      - Corners position in image space
      - Wall-ceiling and wall-floor boundaries in image space
    - **DulaNet 2019**
      - Wall-ceiling and wall-floor boundaries in image space (height estimation)
      - 2D shape on the floorplan (E2P transform)
  - **Result**
    - 2D footprint, 2D corners, heights
    - Same as optimization methods
      - 2D footprint of **ceiling**/**floor**

A more recent option to recover room layout is instead using modern data-driven methods.
In this case the input must be at least an undistorted image, so that the structure depicted in image is aligned to the canonical manhattan axes.
This is basically for two reason: the output of the network, that is also the ground truth for training, is a representation of the room layout as corners and boundaries in the equirectangular image. In this way we expect that at least corners of the same edge lie on the same vertical line.
Second, the output is noisy, so heavy priors need to be applied on the shape at post-processing time, based on Manhattan alignment.
However, these methods out-perform standard optimization approaches at least for the single images, returning 2D footprint, corners and layout heights.

PIPELINE EXAMPLE: POST-PROCESSING

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **Input**
  - 2D/3D elements
    - Footprint shapes, corners, oriented planes
  - Layout information
    - Priors, heights, rooms position
- **Methods**
  - 2D regularization
  - 3D extrusion and displacement
- **Output**
  - 2.5D floorplan
  - 3D layout



Pintore 2018: 3D extrusion

Finally, at post-processing time, all elements are joined to obtain the final layout.
In this step we usually perform a 2 dimensional regularization on the room footprint, 3D extrusion and , eventually, rooms displacement.

PIPELINE EXAMPLE: POST-PROCESSING

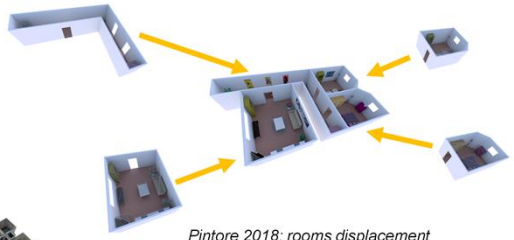
SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **2D regularization**
  - Room shape on a 2D floorplan
  - Manhattan World prior
    - Image warped so that walls are aligned with horizontal/vertical lines
  - Walls are regressed and clustered into horizontal and vertical lines **DulaNet 2019**
  - Walls are fitted from corners position and ceiling/floor boudaries **HorizonNet 2019**

The regularization process is common when dealing with noisy or uncomplete shapes.
A typical approach is to consider the room shape as a 2D projection on the floorplan.
In the case of data/driven output, for example, is common to exploit manhattan world priors, by warping the equirectangular image so that walls are aligned with horizontal and vertical directions.
Here we have some approaches, in the first case, the wall segments are regressed and clustered into horizontal and vertical lines,
In the second one instead the wall lines are fitted starting from corners position with a voting scheme based on ceiling and floor boudaries.

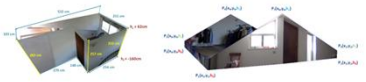
PIPELINE EXAMPLE: POST-PROCESSING

- **3D extrusion and room displacement**
  - Layout height
    - MW/ Atlanta World model : single height
    - Vertical walls model: multiple heights
  - Room reference system: camera center
  - Camera position from MV registration
- **Output examples**
  - 2.5D floorplan
  - 3D layout



Pintore 2018: rooms displacement

Pintore 2018: VW multiple images



Pintore 2018: multiple heights

Finally, in order to define the 3d layout, is common to perform an extrusion using the layout height, which is a single value for manhattan and atlanta models, or multiple values for the vertical walls model.
Moreover, for multiroom and complex scene, we need to join several rooms in the same model. This is usually accomplished using camera reference frames from the multi-view registration,
Since each room reference system is the camera center, that is the sphere center.

DISCUSSION (1/2)

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- Modeling room boundaries is a major challenge in indoor reconstruction
  - High level of noise, clutter, and missing data->heavy use of priors
- Almost all solutions works totally or partially in 2D or 2.5D
  - Extracting floorplan and extending it to 3D
  - Sometimes special handling for ceiling
- Common assumptions
  - Large planar surfaces…
  - Vertical walls…
    - … ok for the majority of buildings
  - Horizontal floor and ceiling
    - … much more limiting!

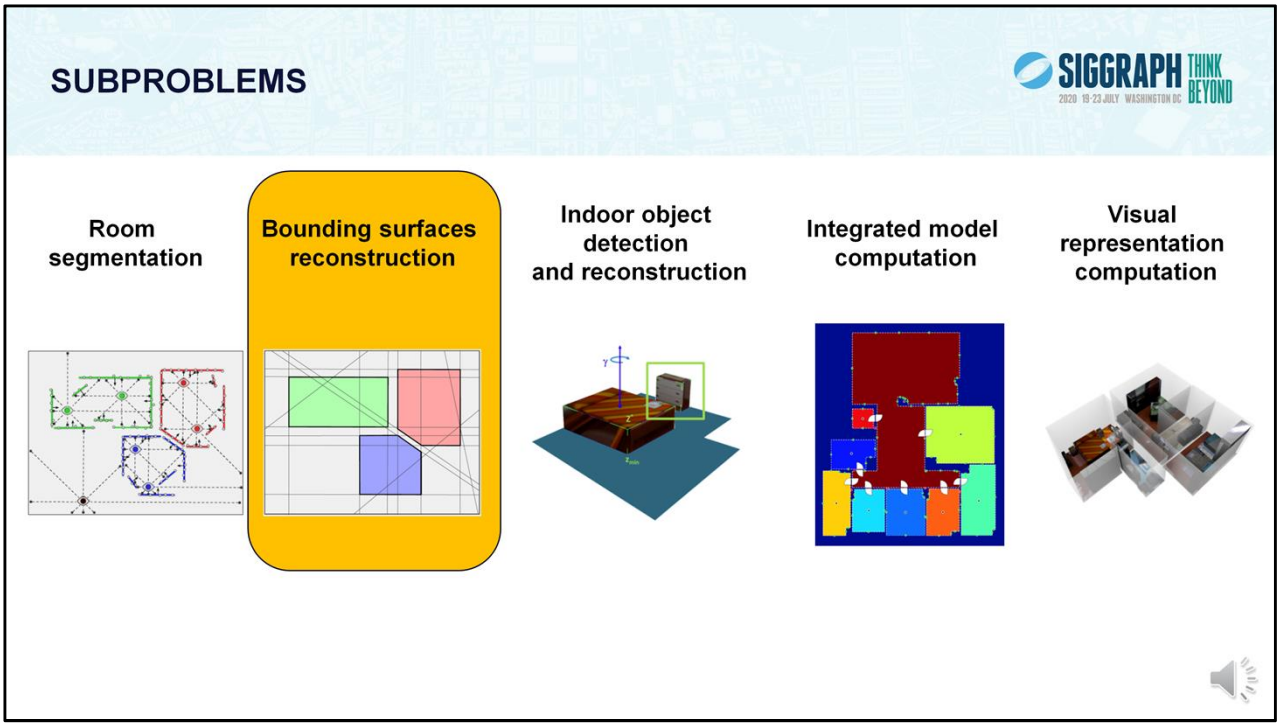
So to summarize:
Modeling the room boudaries usually requires an heavy use of priors.
Moreover, almos all solutions work in 2D and 2.5D domain, using common assumptions like:
Large planar surfaces, vertical walls and often assuming horizontal floor and ceiling.

DISCUSSION (2/2)

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- Open problems
  - Fully 3D approaches
    - Handle complex overhanging structures, floors and ceilings
  - Free-form 3D interiors
    - Also remove large planar surface assumptions
- RGB-D cameras could help fusion between CV and CG approches
- Recent data-driven methods are promising to increase convergence
  - E.g., 3D plane and depth estimation from images
    - Total3DUnderstanding – CVPR2020

Fully 3D approaches are still open problems, in order to obtain, for example, free-form shapes.
In this context RGB_D cameras can help data fusion, as well as recent data-driven methods.

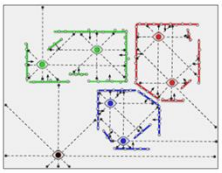


So, once we have recovered the room boundaries we need to deal with the modeling of the elements inside, as for example furnitures and functional objects, in order to complete the model of the scene

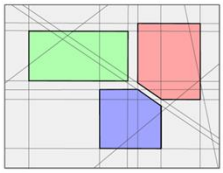
SUBPROBLEMS

SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC

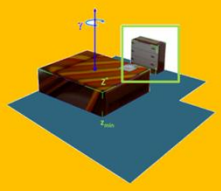
Room segmentation



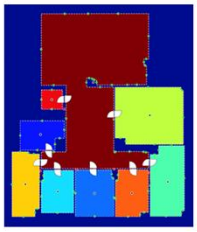
Bounding surfaces reconstruction



Indoor object detection and reconstruction

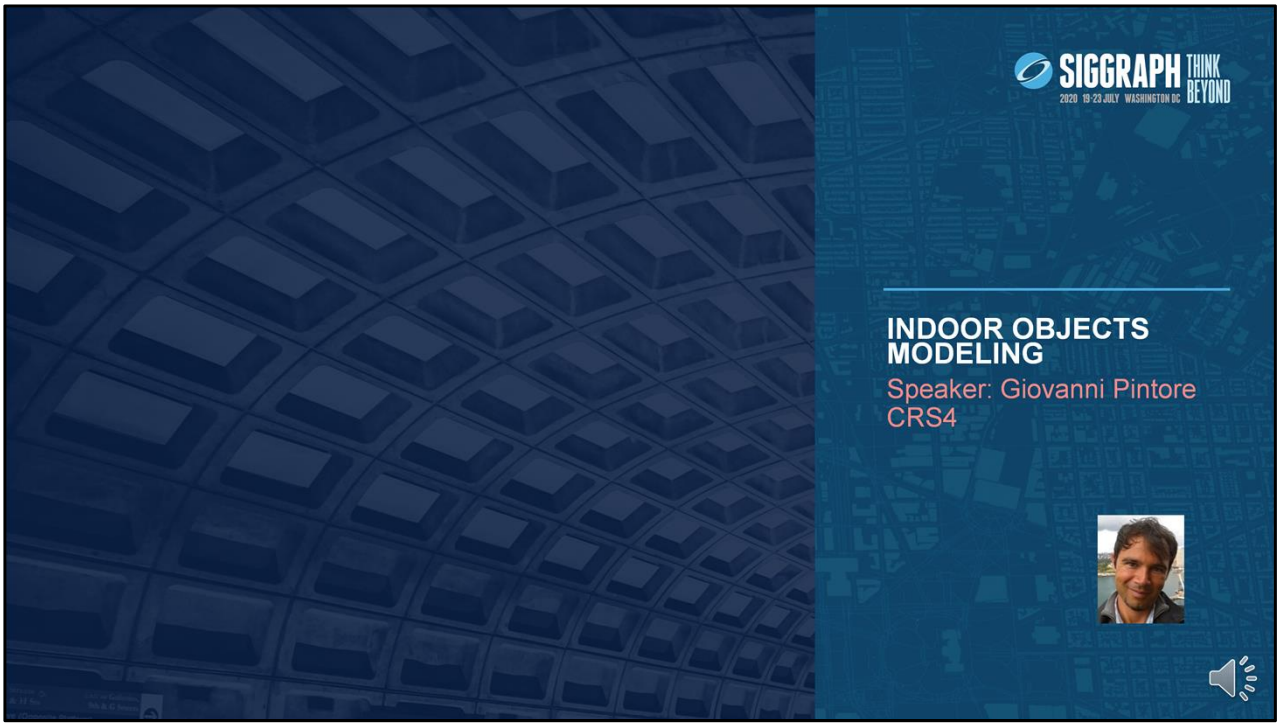


Integrated model computation



Visual representation computation

We are going to see this in the next session.


SIGGRAPH
2020 19-23 JULY WASHINGTON DC
THINK
BEYOND
INDOOR OBJECTS MODELING
Speaker: Giovanni Pintore
CRS4

INTRODUCTION

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **Part of the scene not belonging to the architectural structure**
  - E.g., furniture, lamps, outlets…
- **Object detection and reconstruction is a large topic in itself**
  - Many works in CG and CV!
- **Our focus on object detection and reconstruction integrated with structured indoor modeling**

In our context, we define an indoor object as a part of the scene not belonging to the architectural structure.
Typical examples are furniture, lamps or some everyday items.
Indeed object detection is a large topic itself,
in our course we focus only on object detection and reconstruction embedded in a structured indoor model

GOALS OF OBJECT RECONSTRUCTION IN INDOOR MODELING

SIGGRAPH THINK BEYOND 2020 19-23 JULY WASHINGTON DC

- **Object detection as clutter removal**
  - Focus on permanent structures vs. movable objects
  - Clutter as noise
  - Mostly methods using dense 3D data without imagery (but not only)
- **Reconstruction of 3D objects**
  - Visual data involved
  - Objects are part of the model for many real-world applications
  - Non-zero, finite volume contained inside the room boundary
    - E.g., furniture
- **Reconstruction of flat objects**
  - Many functional objects are flat
    - Outlets, lights, airvents, etc.

Here the main goals for indoor modeling are basically three:
First, object detection as clutter removal, that is the case of methods focused on the permanent structures instead of the movable objects modeling.
this is the typical goal of methods starting from dense 3D data without images.
Then we have the methods that actually reconstruct the objects.
In this case we have objects that are intendended as functional part of model, such as 3d objects with a non-zero volume, or even flat objects, such as outlets or lights.

OBJECT DETECTION AS CLUTTER REMOVAL

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **Focus on the *as-built***
  - Clutter must be removed as early as possible to make reconstruction more efficient
    - less data, less noise
- **Solutions mainly depend on input data**
- **Techniques designed to maximize recall of the permanent components vs. precision**

In the first case, methods are focused on the as-bulit modeling, so clutter needs to be removed as early as possible.
In this context the solutions usually depend on the input data, and the techniques are designed to maximize the recall of the permanent components.

OBJECT DETECTION AS CLUTTER REMOVAL

SIGGRAPH 2020 THINK BEYOND 19-23 JULY WASHINGTON DC

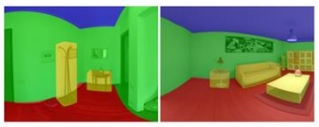
- **Most common approaches from 3D input**
  - First fit 3D primitives (e.g., large planes), then remove outlier points
    - **Ochmann 16, Murali 17 and Ochmann 19**
  - First project 3D data to floor plane, then remove all points not belonging to 2D primitives (e.g., lines)
    - **Oesau 14, Turner 12** and **Turner 14**
  - First partition 3D data into oriented rectangles, then check if rectangles connect floor and ceiling
    - **Mura 14** check vertical extent
    - **Mura 16** proximity graph

From 3D input there are different ways to filter the clutter.
Some techniques start directly fitting 3D primitives and then remove the outliers points,
Other methods first project the 3D data on a floorplan to remove all points not belonging to primitives.
Or other methods first partion 3D data into oriented rectangles and then check if the rectangles are connecting floor and ceiling.

OBJECT DETECTION AS CLUTTER REMOVAL

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **Image pixels classification used with RGB input**
  - Goal: boundary reconstruction enhancement
  - Exploit visual information to label scene parts
  - Common approach: indentify foreground and background elements
    - Exploiting modern machine learning approaches
      - Semantic and/or salient segmentation
  - Foreground objects are accounted for the reconstruction of the boudaries
    - 2D models removed from image **Yang 18**
    - 3D models used as anchor points **Pintore 19**



Pintore et al. CGF 2019

There are also some cases of clutter removal for methods starting from visual input.
In this case the visual information is exploited to improve the boundary reconstruction.
A typical approach is identify foreground and background elements of the scene, and then remove foreground objects from boundary computation.

RECONSTRUCTION OF 3D OBJECTS: SINGLE POSE

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **Two strong assumptions**
  - The object planes are parallel to the walls
  - The object base touches the floor
- **Common approaches**
  - Fitting with a small set of candidates
    - **Lee 10**, **Hedau 10**
  - Generative models
    - **Del Pero 12** and **Del Pero 13**
    - Many constrains about camera object position and size
  - Branch-and-bound strategy
    - **Schwing 13**

In terms of reconstruction of 3D objects, a first branch of methods start from a single pose input.
These methods leverage on two strong assumptions:
The object planes are parallel to the walls and the object lie on the floor.
Some common approaches are based on object model fitting, for example fitting a small set of candidates, or using generative models,
Or a larger set of candidates using a branc and bound strategy.

RECONSTRUCTION OF 3D OBJECTS: SINGLE POSE

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **Data-driven solutions**
  - **Satkin 15**: top-down image matching to align 3D models from a database exploiting GC+OT
  - **Su 15**: CNN for pose estimation using rendered models
- **Objects and layout from a single panorama**
  - **Zhang 14**: bottom-up object hypotheses from image edges
  - **Xu 17**: top-down detection from a library of 3D models

More recently data-driven approaches are becoming more popular, exploiting in some cases top-down matching between models in a database and the image,
Or estimating object poses using rendered and rastered models.
Again, methods migrate to panoramic images, using both bottom-up and top-down strategies.

RECONSTRUCTION OF 3D OBJECTS: MULTIPLE POSES

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **Single and multi-view reasoning on RGB images**
  - **Bao 14**: robust scene understanding from pin-hole images
    - Small scenes
    - Dense image coverage required
  - **Pintore 19**: 3D boxes of major objects from panoramic images
    - Plane-sweeping approach to join multiple image segmentation
    - Large floorplan
- **Volumetric segmentation from RGB-D images**
  - **Hou 19** and **Zheng 19** : deep neural network for real-time voxel-based semantic labeling
- **Data-driven solutions from RGB-D images and database**
  - **Nan 12**: non-rigid pose estimation by deforming and matching

Pintore et al. CGF 2019

Compared to external boundary extraction, recovering objects requires more 3D clues and in general more view-points.
To this end several methods exploits joined single and multi-view reasoning on images,
Starting from pin-hole images to reconstruct a full layout of a small indoor scene,
Or from panoramic images, exploiting a plane-sweeping approach to join multiple image segmentations and reconstruct cluttered floorplans.
On the other hand several solutions start from RGB-D images, exploiting volumetric segmentation and a database of objects.

RECONSTRUCTION OF 3D OBJECTS: MULTIPLE POSES

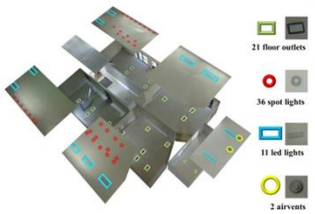
- **Data-driven solutions from RGB-D input and database**
  - 3D object matching
    - **Shao 12**: from virtual scanning
    - **Kim 12**: from depth segmentation
  - RGB data to complete parts that are missing in the 3D scan
    - **Shen 12**
  - Recurrent objects
    - **Mattausch 14**: find instances without database
  - **Yang 18**: GAN to infer fine-grained 3D indoor objects
  - **3D-SIS 19**: real-time voxel-based semantic labeling
- **Data-driven solutions from RGB input and RGB-D database**
  - **Nie 20**: object label, 3D pose and mesh from a database
    - Contextual prediction of layout, camera pose, 3D object bounding boxes and meshes

Many data-driven solutions to match the objects adopt, for example, virtual scanning, depth segmetation,
Other explit rgb data to complete missing parts instead work without a specific database, for example finding recurrent object instances.
More recently some methods exploit supervised training on generative adversarial networks, or voxel-level semantic labeling.
A recent example works directly on rgb images, recovering a contextual prediction of room layout, camera pose, object bounding boxes and their meshes.

RECONSTRUCTION OF FLAT OBJECTS

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **No 3D evidence: 3D approaches are not effective**
  - Image-based methods
- **Multiple pin-hole images**
  - **Vedaldi 18**: SIFT geometric consistency evaluation
- **Single panoramic images**
  - **Pintore 18**: distortion-aware object recognition exploiting underlying 3D structure
    - Multi-room mapping
  - **Chou 20**: SoA semantic segmentation (e.g., FCRN, BlitNet, etc.) trained with equirectagular images
    - Single room mapping

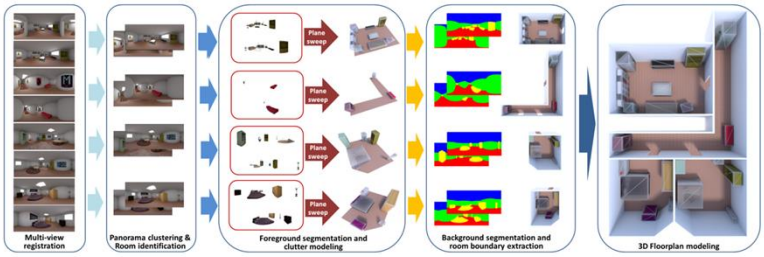


Pintore et al. C&G 2018

A specific mention is deserved flat objects recognition.
Infact, many important functional objects have no 3D evidence, so previous mentioned approaches are ineffective.
In this case the approches for detection are based on images, meanwhile 3D clues are needed to map the objects into a structured model.
In this context some approches evaluate geometric consistency of multi-view features,
Or in the context of panoramic images, some methods adopt a distortion-aware recognition exploiting the underlying 3D structure.

PIPELINE EXAMPLE: OBJECT MODELING FROM PANORAMIC IMAGES

- **Method to model a large indoor scene from a small set of panoramic images**
- **Output: structured 3D model**
  - Rooms subdivision
  - Defined walls, ceilings, floors
  - Major objects inside rooms (i.e., clutter)



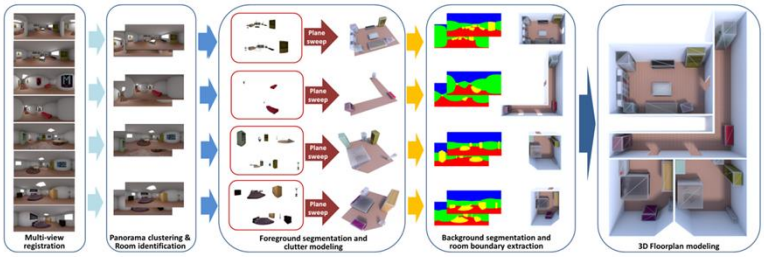
Pintore et al. Automatic modeling of cluttered multi-room floorplans from panoramic images. CGF 2019

Now we'll see an example of a full pipeline for structured object modeling. The presented example is a part of larger system to model a large and complex indoor scene, just form a small set of panoramic images.
Such a system provided the multi-room model with the major objects inside the rooms, modeled with their 3D pose and size.

PIPELINE EXAMPLE: APPROACH 1/2

SIGGRAPH THINK BEYOND 2020 19-23 JULY WASHINGTON DC

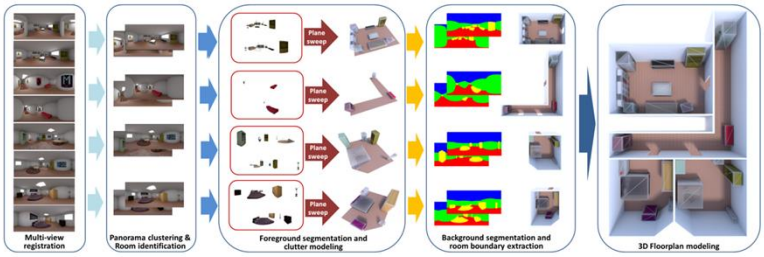
- **Images clustered by visibility**
  - Input clustering per room
- **For each cluster**
  - Plane-sweeping to combine **data-driven SV** analysis and **MV reasoning**
    - **3D clutter model**



Since we have a multi-room environment, first we start clustering the images in to rooms.
Then, for each cluster, we adopt a plane-sweeping approach to combine data-driven object detection and multi-view geometric reasoning.

PIPELINE EXAMPLE: APPROACH 2/2

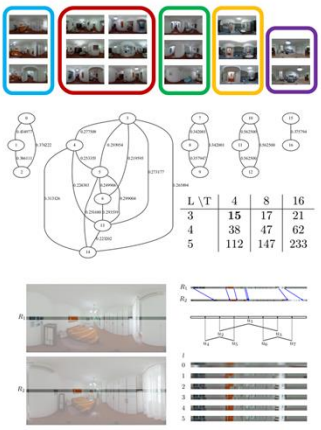
- **Priors**
  - Structure: **vertical walls**
    - Less restrictive than Manhattan World
  - Clutter model: **cuboids**
  - Model suitable for most real-world applications
    - Content creation for: security, guidance, path-planning, energy management, etc.



Such a system adopts some of the priors we have seen in previsious parts of this talk.
In general the whole structure is a vertical walls model, and we represent the objects in the environment with their oriented bounding volume.
Such a model is suitable for most real-world applications, like content creation for security, path planning or guidance.

PIPELINE EXAMPLE: IMAGES CLUSTERING

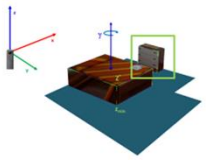
- **Graph connecting images**
  - Arc linking images with common MV features
  - Arc weight: depends by a similarity function
- **Similarity function**
  - 1D image warping
    - Textured strip from the image horizon
    - Warping cost
- **Random walks to find cuts**



As a first task we need to group the images in to different rooms. We have already seen this step in the room partitioning section before.
Just to summarize, here we build a graph connectiong images, where each arc links images which have common multi view features, and we assign to each arc weight depending by a specific similarity function.
Such a function is the 1D warping cost between two images, calculated on a textured strip from the images horizon.

PIPELINE EXAMPLE: CLUTTER MODELING BY PLANE-SWEEPING

- **Virtual camera setting**
  - Top-down view along Z
- **Parametrization and cost function**
- **Z parametrization**
  - *Sphere to plane transform (E2P/A2P)* **Pintore 16**
    - Recover piece-wise floor/ceiling footprint
    - Commonly applied to infer room shape
      - Background segmentation **Pintore 18**
      - Whole image **Yang2019**
    - **Assume only one horizontal plane**
      - Not applicable to objects



(a) *Object model*

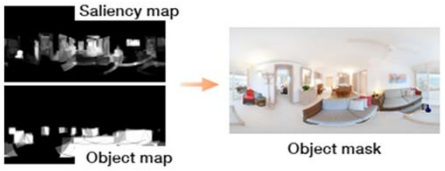
Pintore 2018

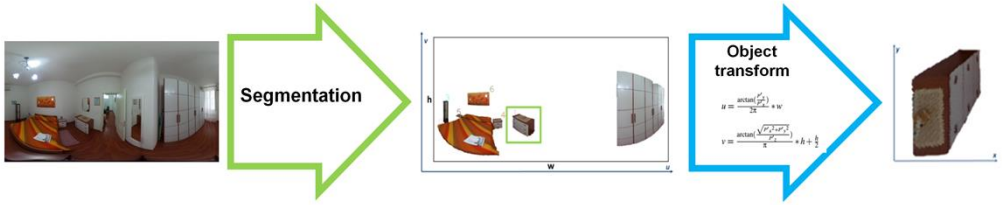
Now, for each room, we model the clutter by introducing virtual plane sweeping approach.
To do this, firstly we set a virtual camera on the top-down view along Z axis, that is along the height of the layout.
Then we need to define a parametrization and a cost function.
We adopt a Z parametrization, like the proposed sphere to plane trasform seen in the previous section.
Such trasnform is commonly adopted to recover piece-wise floor and ceiling footprint, a so to infer the 2D shape of the room.
As it is this transform is suitable only for the external boundary, since it expects to project the spherical image on a single horizontal plane.
In other words is not applicable directly to objects since each one has different height.

PIPELINE EXAMPLE: PLANE-SWEEPING PARAMETRIZATION

SIGGRAPH THINK BEYOND 2020 19-23 JULY WASHINGTON DC

- **Local spatial transform**
  - Applied to each segmented object
  - Image foreground segmentation needed
    - Object detection **YOLO9000 CVPR2017**
    - Saliency **Zhang ICCV2015**
    - **Segmentation mask for each object in the scene**
  - **Returns a 2D object footprint hypothesis for each Z**



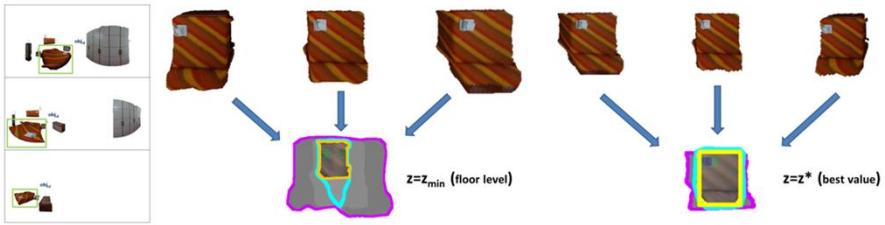


So, for this reason, we need to introduce a local spatial transform, applied time by time to each single object visible in the image.
To do this we need to segment the image in to background, or permanent structure, and foreground objects.
We combine in this case state-of-the-art methods to for image foreground segmentation, such as data-driven object detection and saliecy detection.
In the indoor environment object detection algorithms work well when the entire object is visible.
However, they fail in the cases where the object is occluded or the object is unusual.
So saliency detection helps in these situations because partially visible or unusual objects typically show up as being salient.
Starting from the segmentation mask of each object the local transform returns the 2D object footprint hypotesis, parametrized for each Z.

PIPELINE EXAMPLE: PLANE-SWEEPING COST FUNCTION

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- For each object combines multiple transforms
  - $E(z)=E_s(z)+E_c(z)+E_e(z)$
    - E_s : transformed shape consistency (1-IoU)
    - E_c : color consistency (color STD – H component HSV)
    - E_e : significant edges consistency (2D distance)

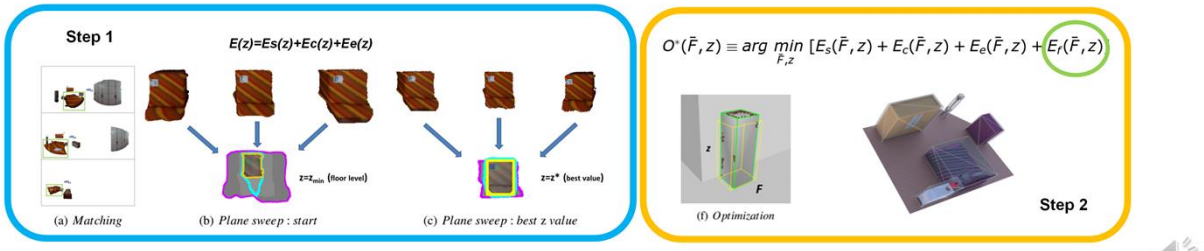


(a) *Matching* (b) *Plane sweep : start* (c) *Plane sweep : best z value*

Now, for each object, we combine multi-view information through the plane-sweeping.
So, varying Z, we minimize a cost function based on three components:
The first component takes in account of the shape consistency, in terms of intersection over union between each object masks from different images.
As we can see we are working in the transformed space.
The second component take in account the color consistency between object projections
While the third component evaluates the significant edges consistency, that is the case of objects higher than the camera level, we will see later an example.

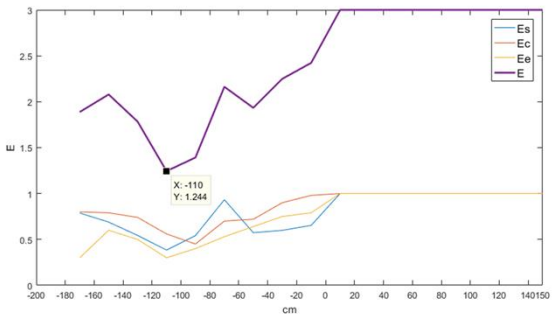
PIPELINE EXAMPLE: PLANE-SWEEPING OPTIMIZATION

- **Step 1**
  - Varying Z we obtain a first cuboid approximation O*(z)
  - 2D footprint (2D position and size, rotation around Z axis) + z* (height)
- **Step 2**
  - Optimization of the 6 cuboid parameters bound to O*(z)
    - Inserting $\mathbf{E_f}$ component: fitting of closest 3D points (MV features)

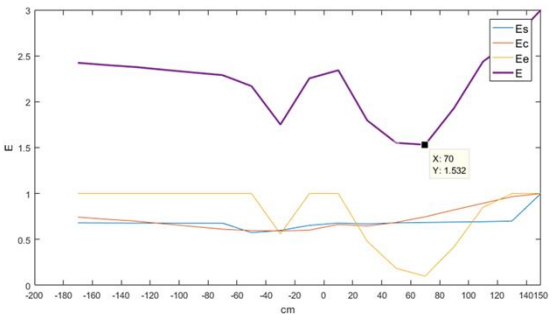


So, varying Z as parameter we obtain a first cuboid approximation, which is given by a 2D footprint, that is the bounding rectangle of best intersection over union, with its associated height.
Then, in a second step, we optimize all the 6 parameters defining the cuboid by levmar minimization , also inserting a further component, which takes in account fitting of the model with the closest sparse 3D points.

PIPELINE EXAMPLE: COMPONENTS MINIMIZATION



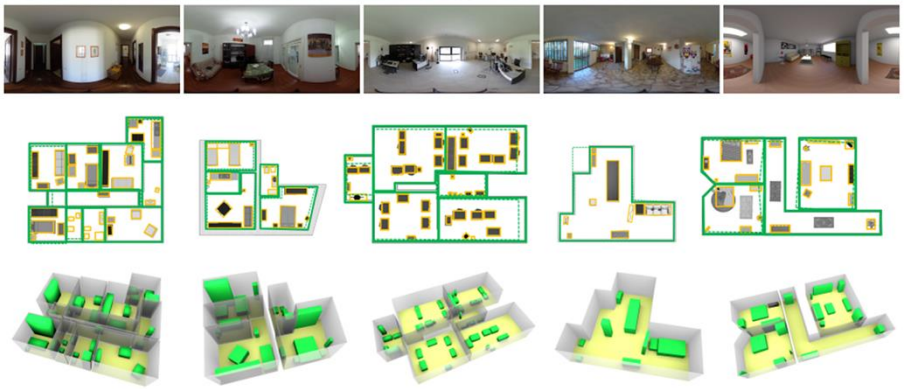
Below camera (bed)



Above camera (cabinet)

Here we have an example of trend of the different components, shape, color, edge at the first optimization step, when Z is varying.
In the first case, the object, that is a bad, is below the camera, so we found that the most discriminative components are the shape and the color.
In the second case on the right instead, the cabinet is higher than the camera position, in this case the most discriminative component is the edge, relative to cabinet visible edges.

PIPELINE EXAMPLE: RECONSTRUCTION PERFORMANCE

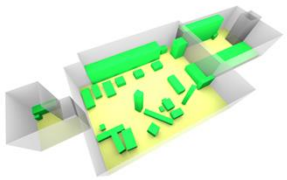


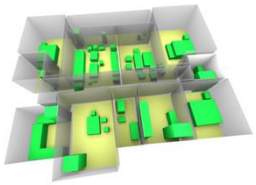
Scene			Clutter error				Imgs per room		Imgs assignement		Room 3D IoU		
Name	Objects	m^2	2D Pos.	Orient.	Area	Height[cm]	Our	[PGP*18]	Our	[PGP*18]	Our	[PGP*18]	[YZ16]
Real-data R1	35/36	96	2 ± 5 *cm*	0.2 ± 1.8 *deg*	2 ± 26 %	2 ± 15 *cm*	2.5	2.6	97 %	76 %	89 %	83 %	75 %
Real-data R2	19/20	78	3 ± 21 *cm*	0.7 ± 2.3 *deg*	2 ± 18 %	1 ± 8 *cm*	3	4	100 %	99 %	90 %	82 %	74 %
Real-data R4	43/44	196	2 ± 8 *cm*	0.4 ± 2.1 *deg*	3 ± 1 %	3 ± 2 *cm*	3	6	91 %	72 %	88 %	74 %	70 %
Real-data R5	10/10	55	4 ± 11 *cm*	0.5 ± 1.0 *deg*	2 ± 8 %	2 ± 3 *cm*	2.5	5	100 %	70 %	91 %	84 %	61 %
Synthetic data S1	20/21	188	4 ± 16 *cm*	0.1 ± 1.0 *deg*	3 ± 32 %	1 ± 5 *cm*	2.5	4	100 %	86 %	90 %	72 %	49 %

		Coverage	Bedroom1	Living	Bedroom2	Restroom	Kitchen
Number of poses	Min	3	5	3	2	2	
	Med	5	9	5	3	3	
	Max	7	13	7	4	4	
3D IoU no clutter info	Min	85%	87%	80%	88%	87%	
	Med	87%	90%	84%	90%	88%	
	Max	89%	91%	86%	92%	89%	
3D IoU with clutter info	Min	90%	91%	89%	91%	89%	
	Med	90%	92%	89%	92%	90%	
	Max	91%	92%	90%	93%	90%	

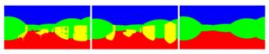
Here we have some numerical results of the presented system on large mulit-room scenes, in term of position and size error.
We also see as considering the object poses improves the whole layout recovery, also in terms of rooms boundaries reconstruction.
In particular, using clutter information reduces the number of images needed to estimate the floorplan structure.

PIPELINE EXAMPLE: RECONSTRUCTION PERFORMANCE





	Our	[BFFFS14]	[CF14]
Completeness	100%	91%	100%
Accuracy	91%	80%	68%



(a) *Ground truth* (b) *Our* (c) [CF14]

(d) *D2 : Coat hanger* (e) *D5 : Round table* (f) *D5 : Plant*

	Our		[XSKT17]	
	Pos. Err.	Orient. Err.	Pos. Err.	Orient. Err.
Bed	2 ± 4 *cm*	0.0 ± 1.5 *deg*	25 ± 17 *cm*	1.0 ± 1.4 *deg*
Chair	1 ± 2*cm*	0.5 ± 1.5 *deg*	52 ± 66 *cm*	10.7 ± 15 *deg*
Plant	2 ± 6*cm*	–	9 ± 12 *cm*	–
Overall	3 ± 21*cm*	1.0 ± 3.0 *deg*	28 ± 32 *cm*	4.3 ± 5.7 *deg*

(a) *Wrong matching* (b) *Fatal occlusion*

Failure cases

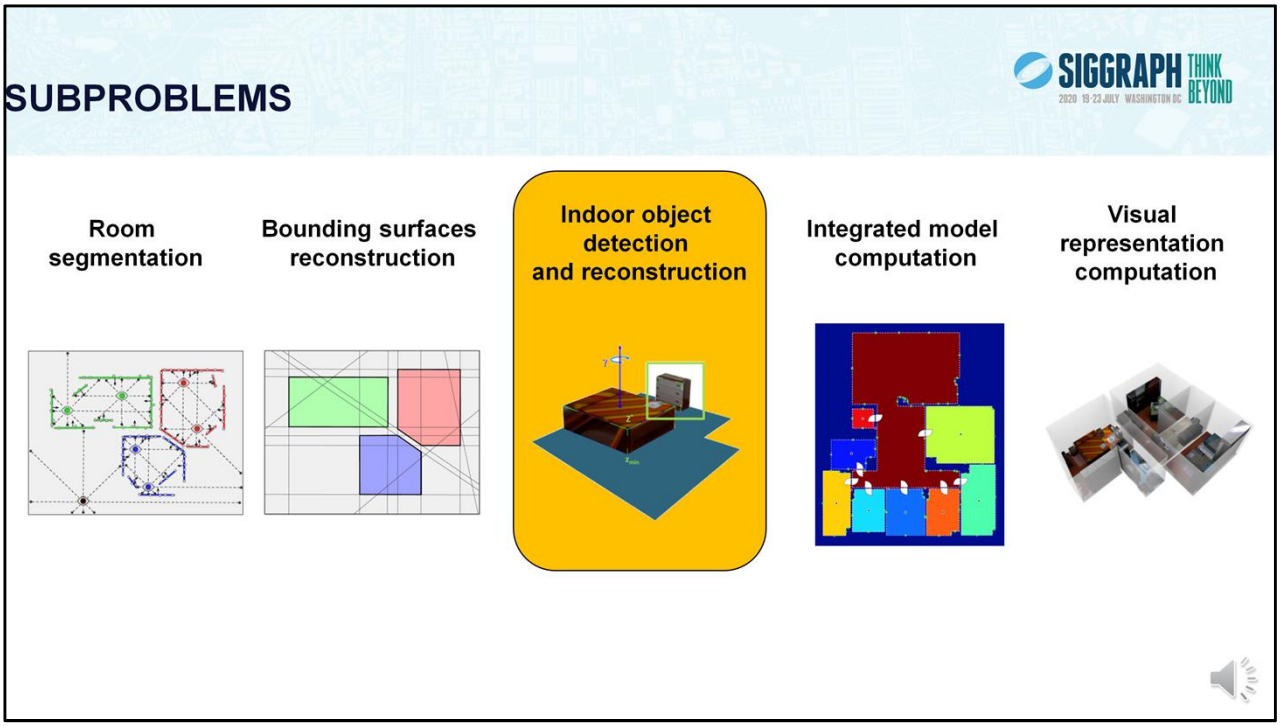
Here we have performances in terms of object recovery and some failure cases. Tipically such a system fails if there is a wrong object matching between different images, or in case of a fatal occlusion in the object mask.

DISCUSSION

SIGGRAPH 2020 19-23 JULY WASHINGTON DC THINK BEYOND

- **Approaches from pure 3D data only focus on the permanent structure**
  - Clutter is filtered as noise
- **Approaches from purely visual or mixed data provide a coarse approximation of the objects**
  - 3D poses and size of the objects
  - 3D shape from standard indoor objects database
  - Recent prominent work: ***Total 3D Understanding CVPR 2020***
    - From a single RGB perspective image provides contextual room layout, 3D object poses and refined meshes
    - Limited to small scene
- **Recovering the real 3D shape of multiple objects in a complex environment is an open problem**
  - Using panoramic images extends the ability to model larger and more complex scenes
  - Combining visual and depth information can bridge the gap between scene understanding and geometric reasoning

To summarize, we have seen that the approaches from pure 3D data only focus on the permanent structure and the clutter is filtered a noise.
On the other hand approaches from pure visual or mixed 3D data have a broader focus, but provide just a coarse approximation of the object., or at least a matching with an object library.
Recent works from visual data provide contextual room layout, 3d object poses and refined meshes exploiting data driven methods, however they are still limited to very small scenes.
So recovering the real, detailed shape of multiple objects, especially in a large and complex environment is still an open research problem.
Also in this case using data fusion approaches can bridge the gap between scene understanding and geometric reasoning.

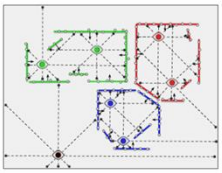


We have seen the methods for the indoor object modeling embedded in a whole 3D room layout.
Indeed a complex and large environment it is defined by many rooms and spaces joined togheter.

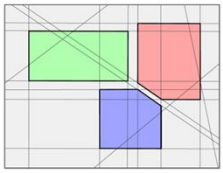
SUBPROBLEMS

SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC

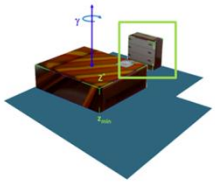
Room segmentation



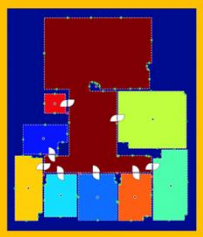
Bounding surfaces reconstruction



Indoor object detection and reconstruction



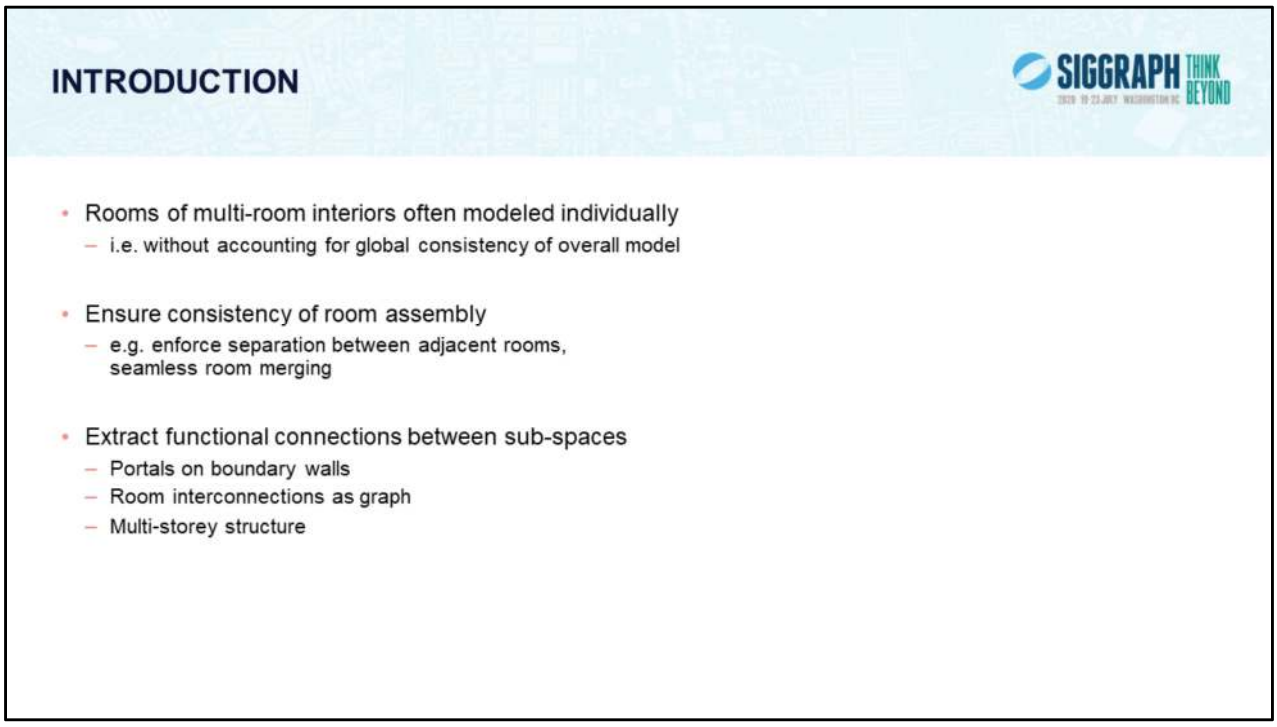
Integrated model computation



Visual representation computation

To this end we need to integrate recovered rooms in a consistent way, as we are going to see in the next section.


SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC
INTEGRATED MODEL COMPUTATION
Speaker: Fabio Ganovelli

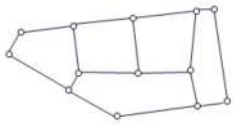


Many modern pipelines model individual rooms as individual entities, reconstructing a single 3D model independently for each of them. This is useful under many aspects – first of all, it increases computational efficiency – but introduces the need for an explicit step to integrate these partial results into a single coherent structured model.
In this context, it is first of all important to ensure that the individual room models are assembled correctly – which means, ensuring a consistent treatment of the separation between the two sheets of a wall shared by different rooms, or avoiding geometric artifacts when merging two room models into a single one.
On top of that, once all the room models are given it is important to extract basic information about their interconnections, which requires the systematic detection of doors and passages on the reconstructed boundary walls and, possibly, handling the case of interiors that span multiple storys.

ROOM CONSISTENCY: WALL BOUNDARIES

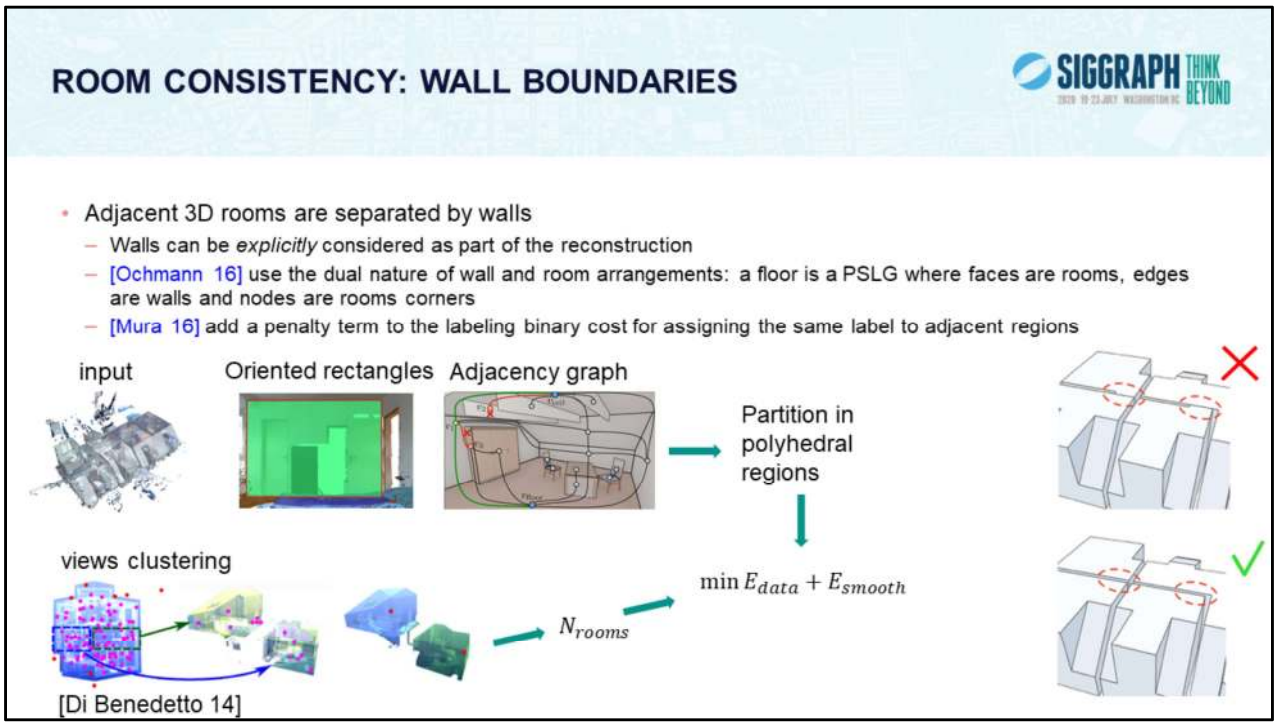
- Adjacent 3D rooms are separated by walls
  - Walls can be *explicitly* considered as part of the reconstruction
  - [Ochmann16] use the dual nature of wall and room arrangements: a floor is a PSLG where faces are rooms, edges are walls and nodes are rooms corners

PLSG representation

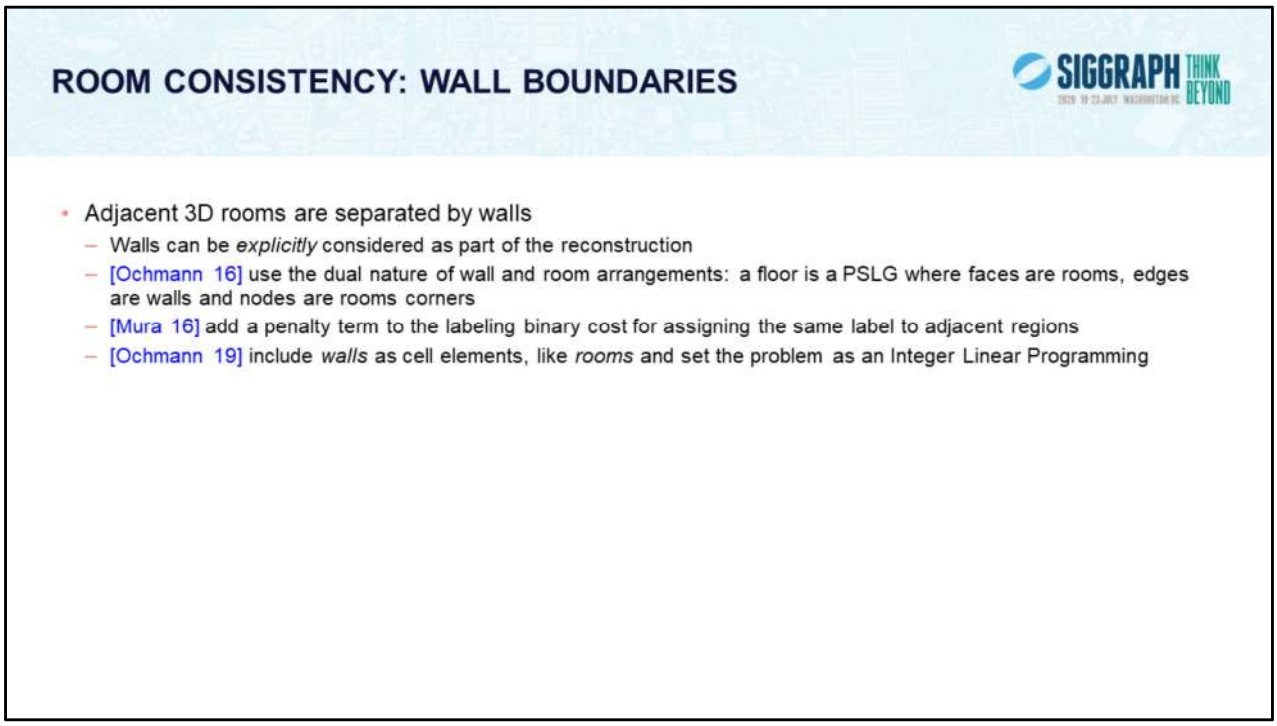


Enforcing coherent boundaries between adjacent rooms involves different techniques depending on how exactly boundaries are represented.

Walls can be considered as part of the reconstruction just like rooms are. After all the 3D samplings return surfaces which are boundaries both of the walls and the rooms alike. Ochmann and colleagues [Ochmann 16] harnessed this duality and consider the floor as a planar straight lines graph where the edges are the walls, the faces are the rooms and the corners are the points where walls cross each others.
In their approach, the input data are analyzed to find wall as pairs of parallels vertical surfaces. The middle line of each wall is considered and extended so that the set of extended walls induces a partition of the dataset in cells. Also, the scan position, which are assumed to be known, provide a coarse subdivision of the space in rooms. Than a labeling problem is set up where the goal is to assign a label to each of the cells from the partition. In this labeling problem, the smoothness term penalizes the assignment of the same label to adjacent regions when they are separated by a wall supported by enough points of the dataset in the boundary between the cells.



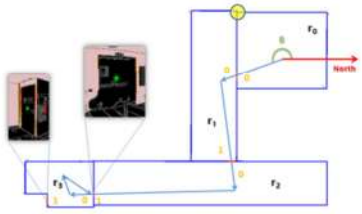
A labelling approach is also used in the work by Mura and colleagues. They use planar priors and fit oriented rectangles with the point cloud. In order to find the structural rectangles, that is, the floor, the ceilings and the walls, they encode the relations between rectangles in an adjacency graph. By visiting the adjacency graph, they are able to identify non-structural rectangles as those which are not in a path from the floor to the ceiling such as the surface of a table in the image. Thanks to the adjacency graph, the can then extend only structural planes and have a robust partition of space in cells over which to define the labelling problem. The number of rooms is determined with a clustering approach [Di Benedetto 14] applied to the scan positions (at least one scan is assumed for room).



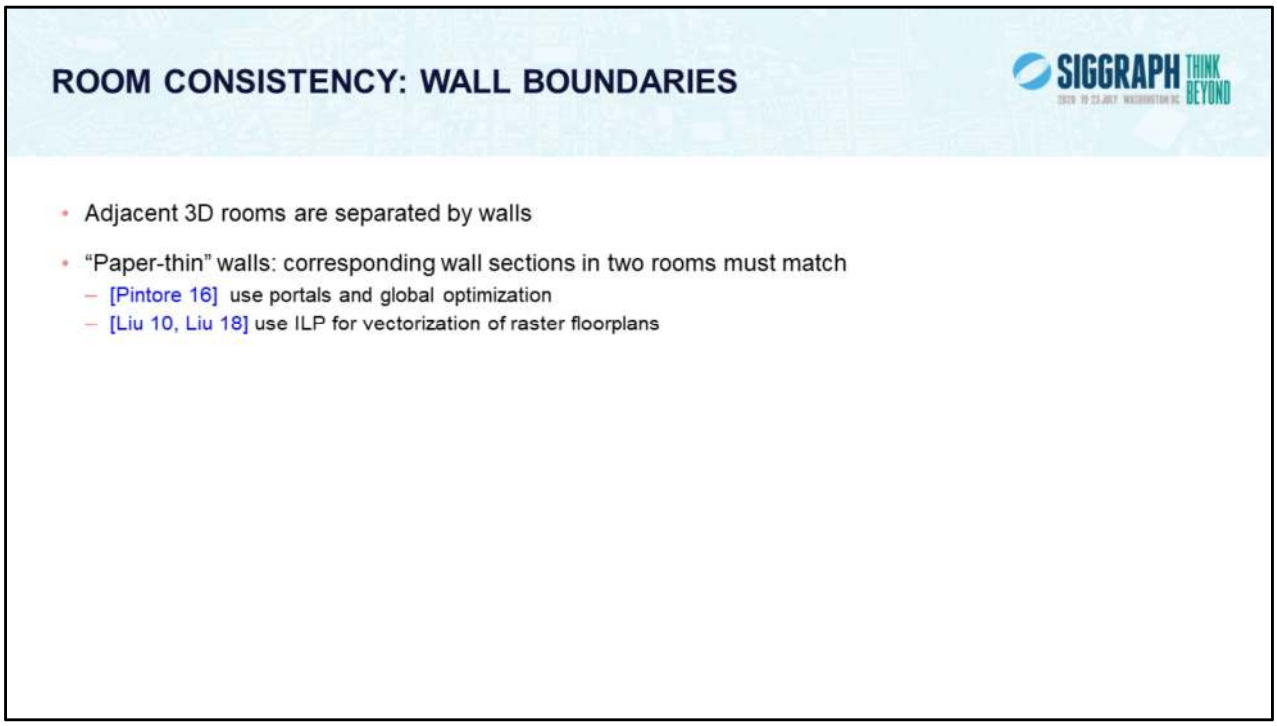
In a more recent effort, Ochmann and colleagues [Ochmann 19] push forward the idea of walls/rooms duality by considering walls as cells themselves. In other words, the volumetric partition is found by using all planes, that is, both side of each wall, so that the resulting cells can be either portion of rooms or portion of walls. They use ILP where the solution is expressed as a series of binary variables, one per couple (cell,label) and a set of constraints ensure that any found solution is sound, for example that the boundary of a room cell is a wall cell and so on .

ROOM CONSISTENCY: WALL BOUNDARIES

- Adjacent 3D rooms are separated by walls
- "Paper-thin" walls: corresponding wall sections in two rooms must match
  - [Pintore 16] use portals and global optimization
    - Portals positions are *estimated* by tracking device during acquisition and *refined* with image segmentation
    - Portals provide room graph and their position provide constraints on the rooms alignment



There are contexts where the thickness of the walls is unimportant and a paper-thin assumption can be adopted. In these cases the problem consists only in finding a consistent arrangement of neighbor rooms.
For example Pintore and colleagues propose a system where the position of the acquisition device, which in this case is a panoramic camera, is tracked during acquisition. Then, after the geometric reconstruction of each room, the tracking is used to estimate the position of doors, further refined with image segmentation techniques. Since that each door is shared by two adjacent rooms, their position provide information both for the definition of a room graph and for the mutual position of the rooms.



In the work by Liu and colleagues [Liu 18], the definition of rooms borders is essentially a post processing step of the system. Their pipeline consists in the combined use of NNs which produces a rasterized data with features points, such as room corners and door corners) which is used as input to a final stage which perform a vectorization of the data in order to produce the final footprint. The vectorization problem is tackled with integer linear programming, by defining a set of junction types and a set of constraints on how these junction can be connected to for a valid footprint.

ROOM CONSISTENCY: WALL BOUNDARIES

- Adjacent 3D rooms are separated by walls
- "Paper-thin" walls: corresponding wall sections in two rooms must match
  - [Pintore 16] use portals and global optimization
  - [Liu 10, Liu 18] use ILP for vectorization of raster floorplans
  - [Chen 19] shortest path with penalty term + post-processing
    - Express each room as a loop of pixels
    - Define a global cost function that accounts for input data and wall consistency

Coherency with the data | Global consistency | Regularization term

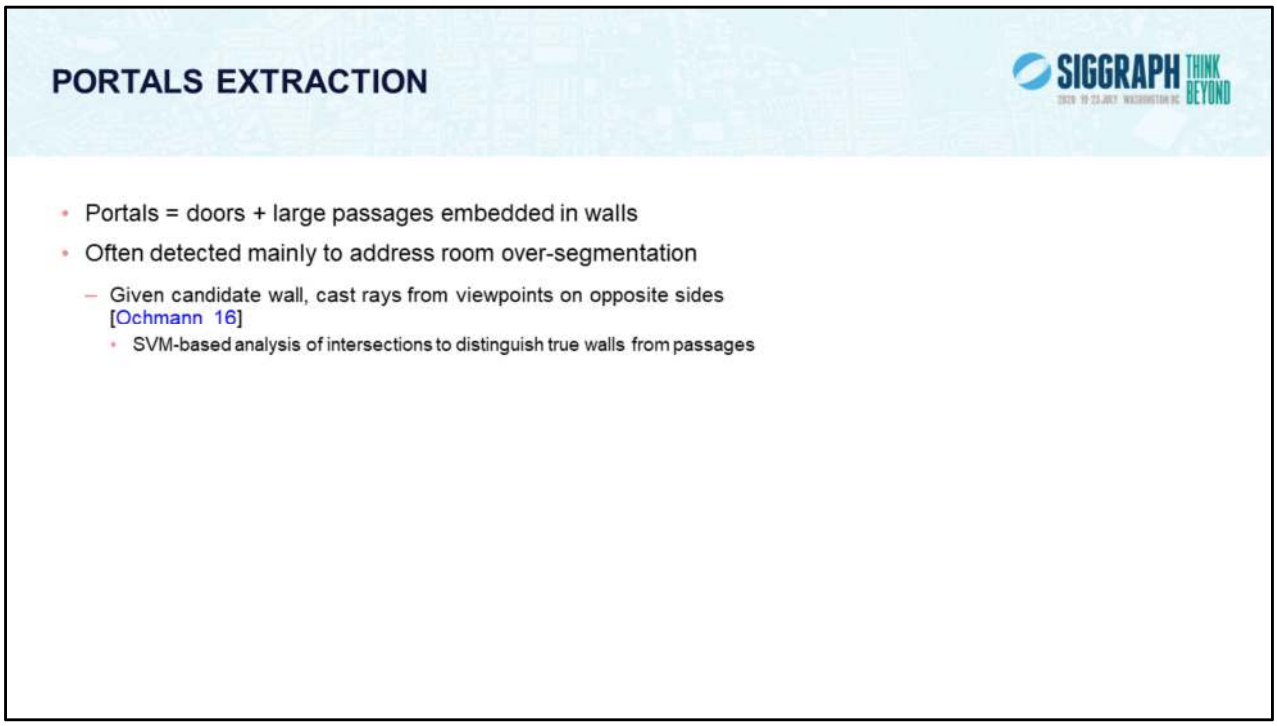
$$\sum_{L_i \in \mathcal{L}} E_{data}(L_i) \quad + \quad E_{consis}(\mathcal{L}) \quad + \quad \sum_{L_i \in \mathcal{L}} E_{model}(L_i)$$

A similar solution for the same type of input data is proposed by Chen and colleagues [Chen 19]. Their idea is to reduce the problem of vectorization to solving a series of shortest path problems where each path is a loop that identifies a single room.
They define a cost function over the entire sets of paths which considers the global consistency by penalizing sides of neighbor loops that are closest than a threshold and are not coincident. They express the cost function as a summation of values over adjacent pixels so that the shortest path around a room is the minimum of the cost function for that room.

ROOM CONSISTENCY: WALL BOUNDARIES

- Adjacent 3D rooms are separated by walls
- "Paper-thin" walls: corresponding wall sections in two rooms must match
  - [Pintore 16] use portals and global optimization
  - [Liu 10, Liu 18] use ILP for vectorization of raster floorplans
  - [Chen 19] shortest path with penalty term + post-processing
    - Express each room as a loop of pixels
    - Define a global cost function that accounts for input data and wall consistency
    - Solve for each loop by reduction to a shortest path (of pixels around each room) problem

At a global scale, they iterate the shortest path algorithm sequentially for all rooms in the dataset until convergence, thus performing a gradient descent optimization.



Turning to aspects more closely related to the navigable structure of the environment, portals – that is, doors and large passages in walls – are an element of fundamental importance, as they explicitly mark transitions between different rooms.

For this reason, their presence has often been used to address room oversegmentation, that is, the case of whole rooms being wrongly split into several distinct rooms. For example, Ochmann and colleagues detect non-existent walls between candidate rooms by (NEXT) considering viewpoints on the two sides of the wall (NEXT)(NEXT) and casting rays from them through the wall. The intersections on the walls can correspond either to scanned points or to empty space; a support vector machine classifier is fed with the features of these intersections and classifies the wall as real or virtual

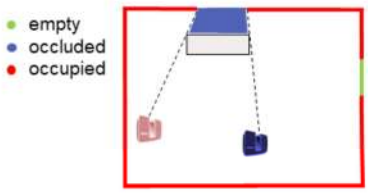
PORTALS EXTRACTION

- Portals = doors + large passages embedded in walls
- Often detected mainly to address room over-segmentation
  - Given candidate wall, cast rays from viewpoints on opposite sides [Ochmann 16]
    - SVM-based analysis of intersections to distinguish true walls from passages
  - Analyze empty regions in 3D vertical planes of candidate walls
    - [Ambrus 17] projects the point on the wall planes and evaluate the shape of openings

The same rationale is behind the approach of Ambrus and colleagues, who essentially resort to image processing to analyze the empty regions on the vertical planes of candidate walls. If a sufficiently large and high empty area is detected, the wall is considered as a real one.

PORTALS EXTRACTION

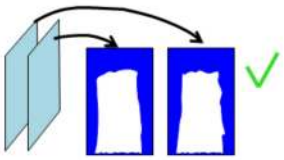
- Portals = doors + large passages embedded in walls
- Often detected mainly to address room over-segmentation
  - Given candidate wall, cast rays from viewpoints on opposite sides [Ochmann 16]
    - SVM-based analysis of intersections to distinguish true walls from passages
  - Analyze empty regions in 3D vertical planes of candidate walls
    - [Ambrus 17] projects the point on the wall planes and evaluate the shape of openings
    - [Adan 11] distinguish between empty and occluded voxels by volumetric ray casting from all scan positions



It's worth mentioning that holes in the sampling of the wall do not necessarily imply the presence of doors or windows, in that the presence of occluders, such a closet, may lead to false positives. To this regard, the solution proposed by Adan and colleagues [Adan 11] uses a voxelization of the domain and label as occluded those voxels which are not visible by any of the scan positions.

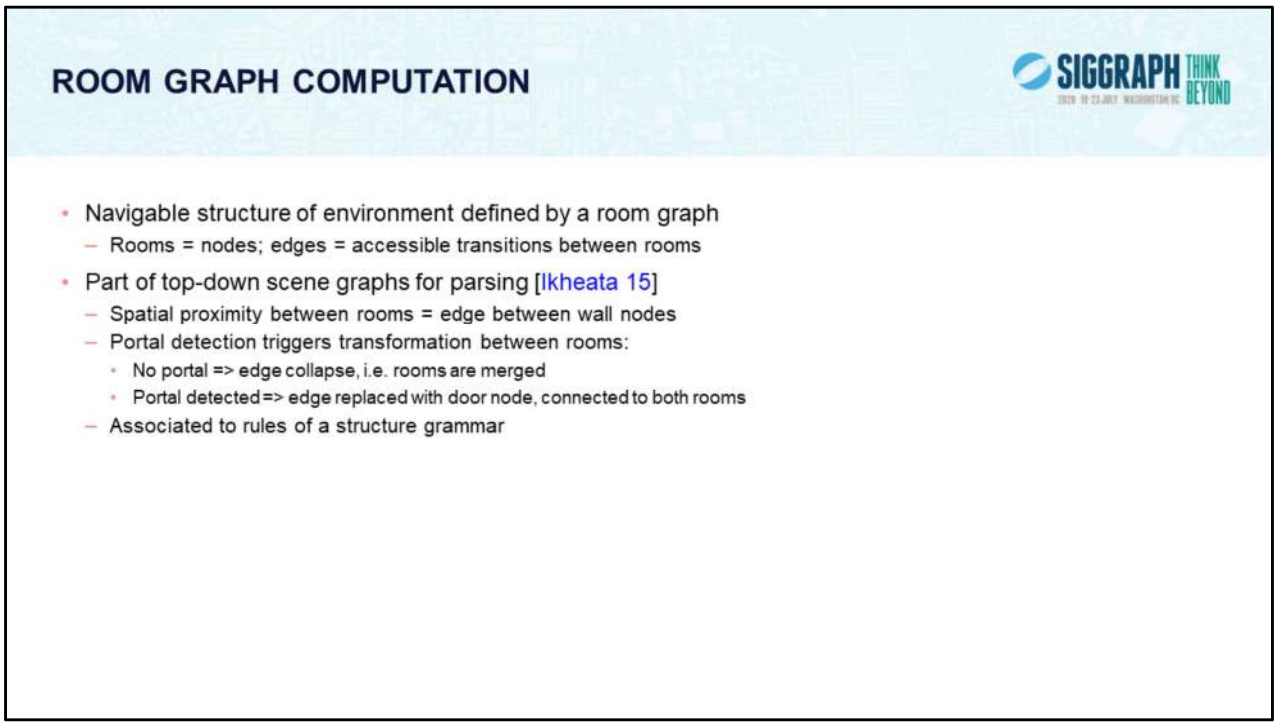
PORTALS EXTRACTION

- Portals = doors + large passages embedded in walls
- Often detected mainly to address room over-segmentation
  - Given candidate wall, cast rays from viewpoints on opposite sides [Ochmann 16]
    - SVM-based analysis of intersections to distinguish true walls from passages
  - Analyze empty regions in 3D vertical planes of candidate walls
    - [Ambrus 17] projects the point on the wall planes and evaluate the shape of openings
    - [Adan 11] distinguish between empty and occluded voxels by volumetric ray casting from all scan positions
- More principled detection denotes room transitions
  - Detect matching empty regions on pair of parallel, vertical walls [Ikehata 15]



However, portals detection can be applied in a more principled manner to detect transitions between rooms. In the work by Ikehata et al [Ikeata 15], portals are also extracted by analyzing the empty space on vertical wall planes;
In this case, however, pairs of adjacent, parallel planes are considered: if sufficiently large empty regions are detected on both planes, and if their shapes match, then a portal is inserted between the two rooms

and this information is explicitly used towards the definition of a ….



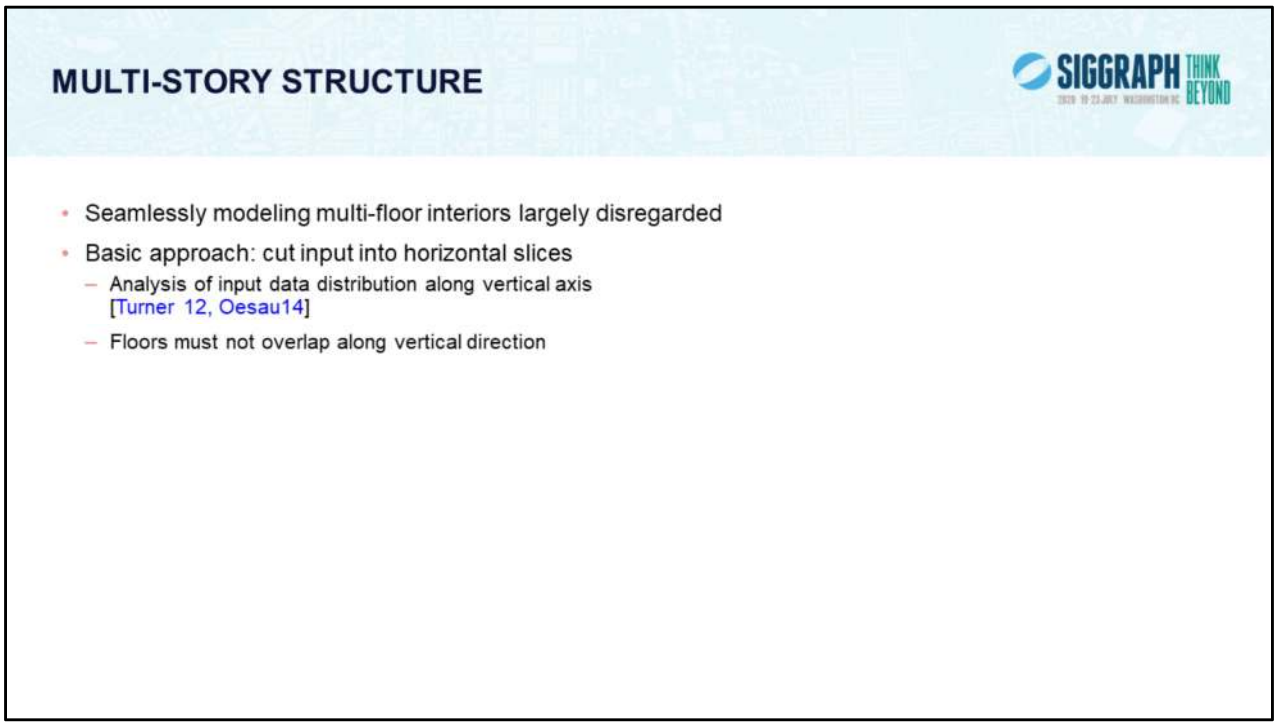
Room graph, a fundamental piece of information to define the navigable structure of the environment. This graph represents the rooms as nodes, and physically accessible transitions between rooms as edges.

This graph can be seen as a subset of the scene graphs recently proposed for top-down representations of indoor scenes, which focus on semantic parsing. In the work of Ikheata and colleagues, this graph also encodes a structure grammar, with grammar rules defining transformations between nodes.
In particular, spatially close rooms have wall nodes that are connected by an edge; the outcome of a portal detection operation corresponding conceptually to that edge defines a precondition for one of two rules:

- A room-merge rule, which merges the room nodes into one
- A door-addition rule, which adds an explicit door node connected to each wall node of the rooms

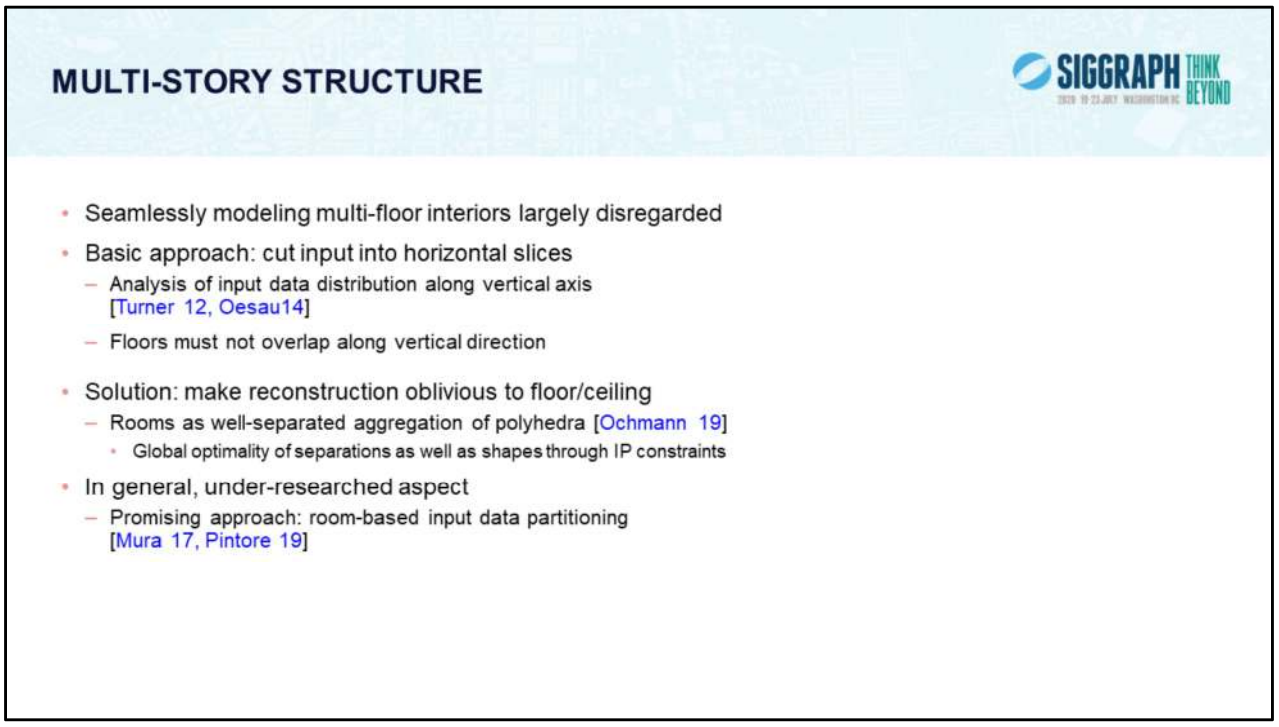
Such an explicit and holistic modeling of room interconnections is however the exception rather than the rule: most approaches focus on modeling adjacencies rather than room interconnections, and aim to do this in the most accurate way possible.
For instance, in the recent work by Chen an colleagues, rooms are modeled as 2D boundary loops in a top-down view of the environment; adjacent rooms share portions of their

boundaries, and specific penalty terms are added to the optimization formulation to favor the perfect overlap of the shared segments



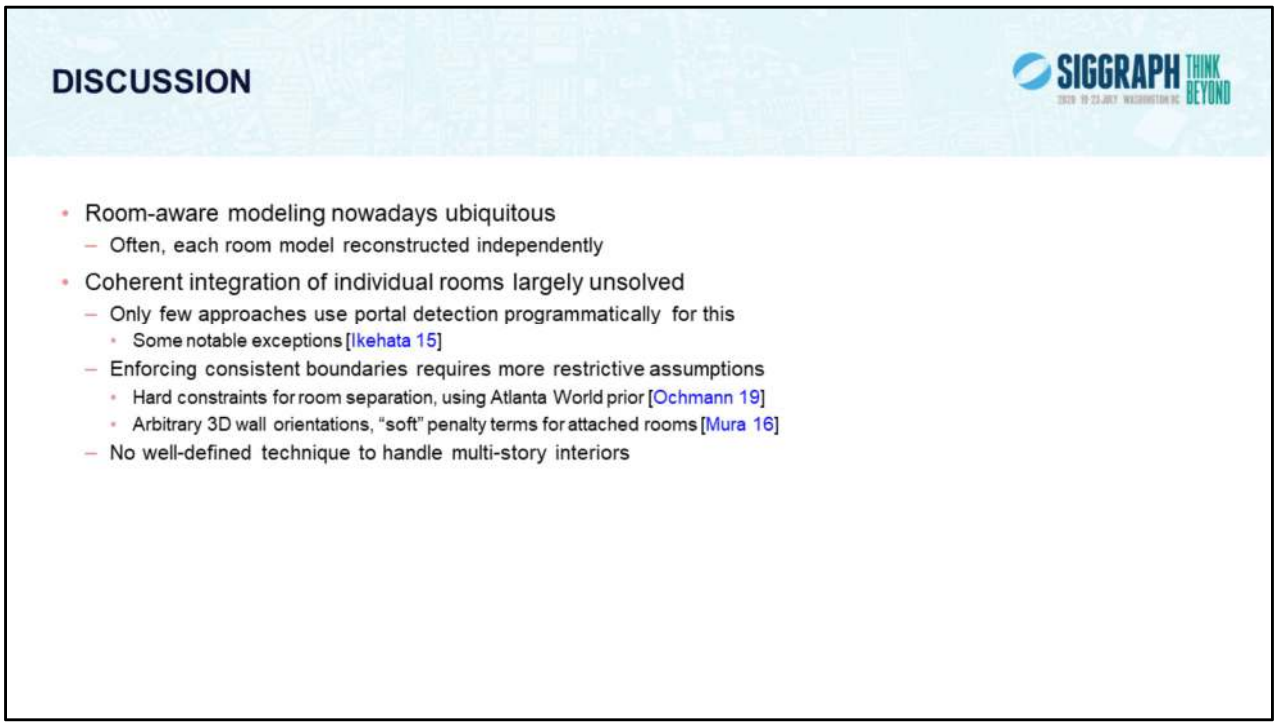
For large indoor spaces that span multiple storys it is important to ensure that the transitions between storys do not affect the correct assembly of the individual room models. Interestingly, this aspect is widely disregarded in state-of-the-art pipelines.

The most basic approach to handle multi-story interiors is to slice the input data along the vertical direction, obtaining horizontal slices, each corresponding to a story level. This can be done in a relatively robust manner by analyzing the distribution of input samples along the vertical axis, for instance using a 1D mode finding algorithm to detect peaks in this distribution, as done by Oesau and colleagues. Two subsequent peaks separated by a relatively small gap denote the transition from one floor level to another.
Obviously, this sets the very restrictive assumption that the floors must not overlap along the vertical direction



One way of overcoming this limitation is to simply make the reconstruction process oblivious of the notions of floor and ceiling. This result is obtained as a nice side-effect in the recent work by Ochmann and colleagues, who simply compute rooms through the aggregation of polyhedral regions of space, ensuring that these aggregates of polyhedra are well-separated using a global formulation based on integer programming. Enforcing this separation does not require the definition of global floor and ceiling heights, so each room is allowed to have its floor leveled at an arbitrary height, independent of other rooms.

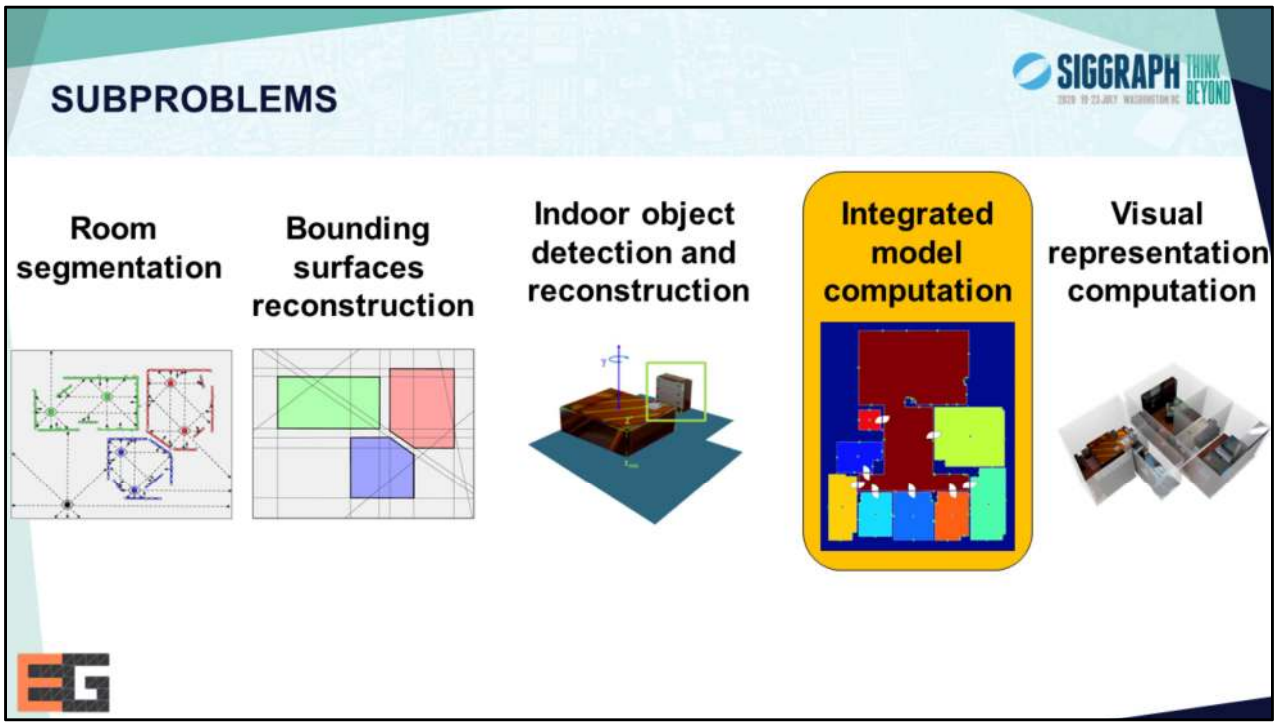
In general, though, a flexible and robust handling of multi-story interiors is an under-researched problem, and a promising solution comes from modeling rooms as individual entities by input data partitioning strategies.



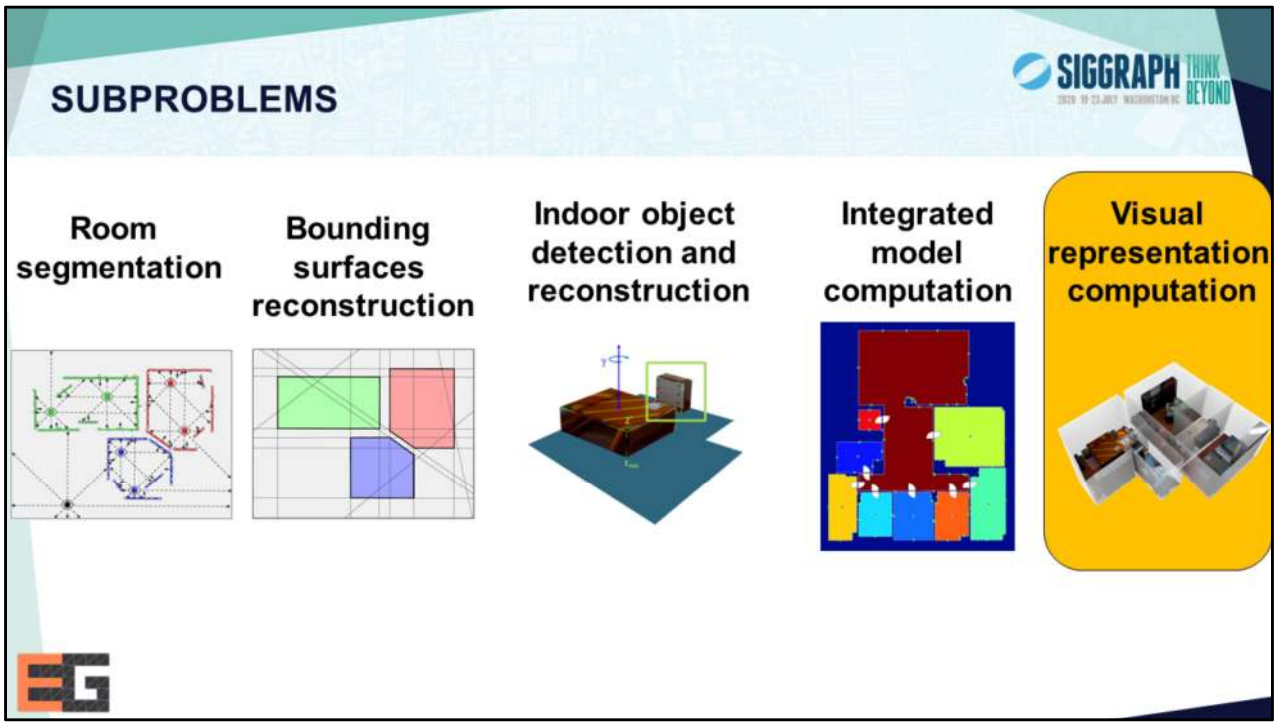
In fact, this is just one of the open problems when it comes to producing globally consistent structured models of interiors.

Making indoor modeling room-aware is nowadays a need-to-have feature rather than simply a nice-to-have, but how to integrate the individual room models correctly while also recovering the navigable structure of the environment is not a solved problem.
Only some approaches – mainly, the work by Ikheata et al – programmatically use portals to determine the interconnections of rooms.

Also when it comes to the geometric consistency of room boundaries, obtaining this result often requires using restrictive assumptions on the architectural shapes of the environment. For instance, the work by Mura et al. recovers walls with arbitrary orientations in 3D space, but only ***favors*** an explicit separation between adjacent rooms; the integer programming formulation by Ochmann and colleagues rigidly enforces room separation, but uses the Atlanta World prior, requiring vertical walls and horizontal floors and ceilings.



This concludes the review of room segmentation techniques



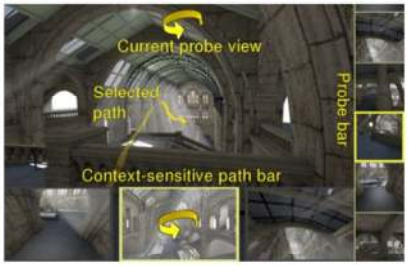
This concludes the review of room segmentation techniques


SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC
VISUAL REPRESENTATION GENERATION
Speaker: Fabio Ganovelli

VISUAL REPRESENTATION GENERATION

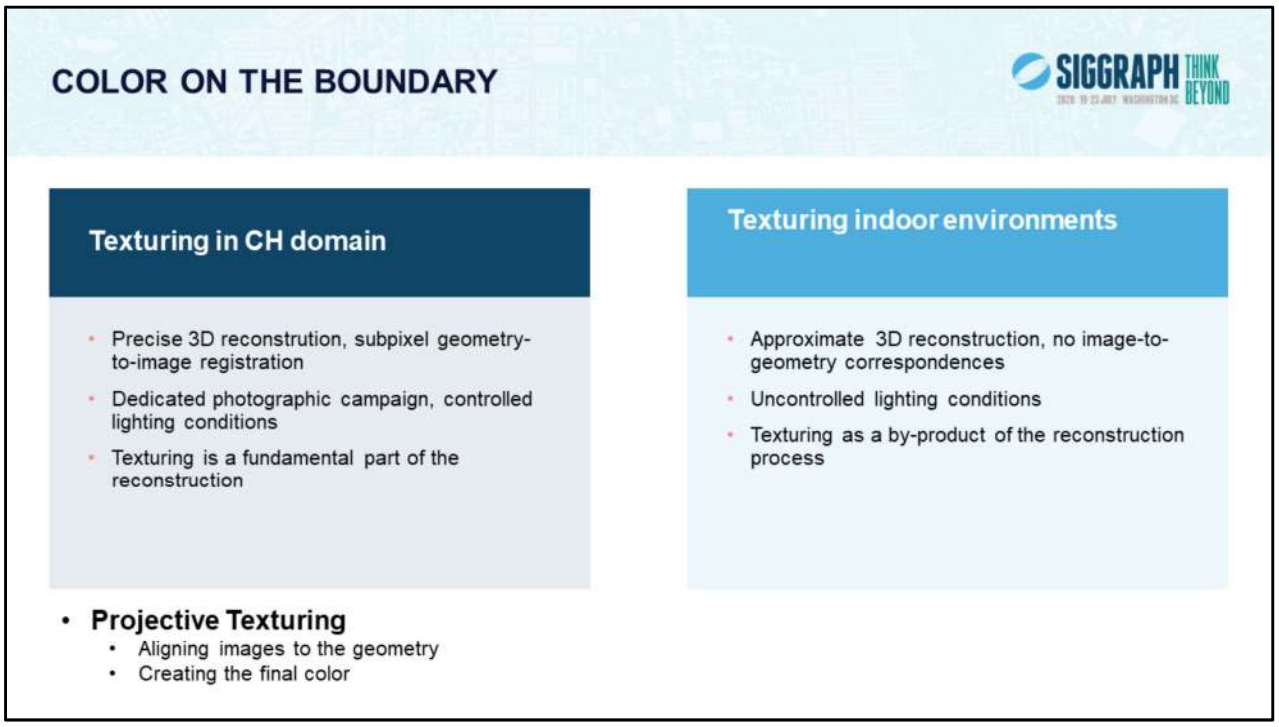
SIGGRAPH THINK BEYOND

- Several methods target recovering blueprint or 2.5D models. Producing a compelling visual experience requires:
  - Adding color detail to the reconstructed geometry
  - Geometric refinement: not just walls, indoor elements (i.e. furniture)
  - Algorithms to support visual inspection



We have seen how many methods are targeted to reconstruct geometry and topology of the environment. However, there are many applications where also a realistic and immmersive visual representatation can be useful.
We may find examples in the real estate market, in the virtual museums domain or in the applications about building renovation, to mention a few.
In these and other cases we may want to be able to browse the reconstructed model as if we were inside it, and this needs at least 3 things:
Having color detail: no matter how much we are used look at shaded geometry of acquired artifacts such as a statue or a church, the great majority of indoor spaces are simple and repetitive and geometry alone would not be enough to make sense of them.
Having indoor elements is also important. What is generically defined as «clutter» in the context of boundary reconstruction, when it comes to visualization it is a precious visual aid.
Finally, we need algorithms for visualizing the models. As much as rendering itself is usually not an issue, we need proper metaphores and interaction modes, which also depend on the specific type of data used (that is images, panoramic images, textured geometry) and the device we want to use.
In the next few minutes, we will look at how these problems are solved at the

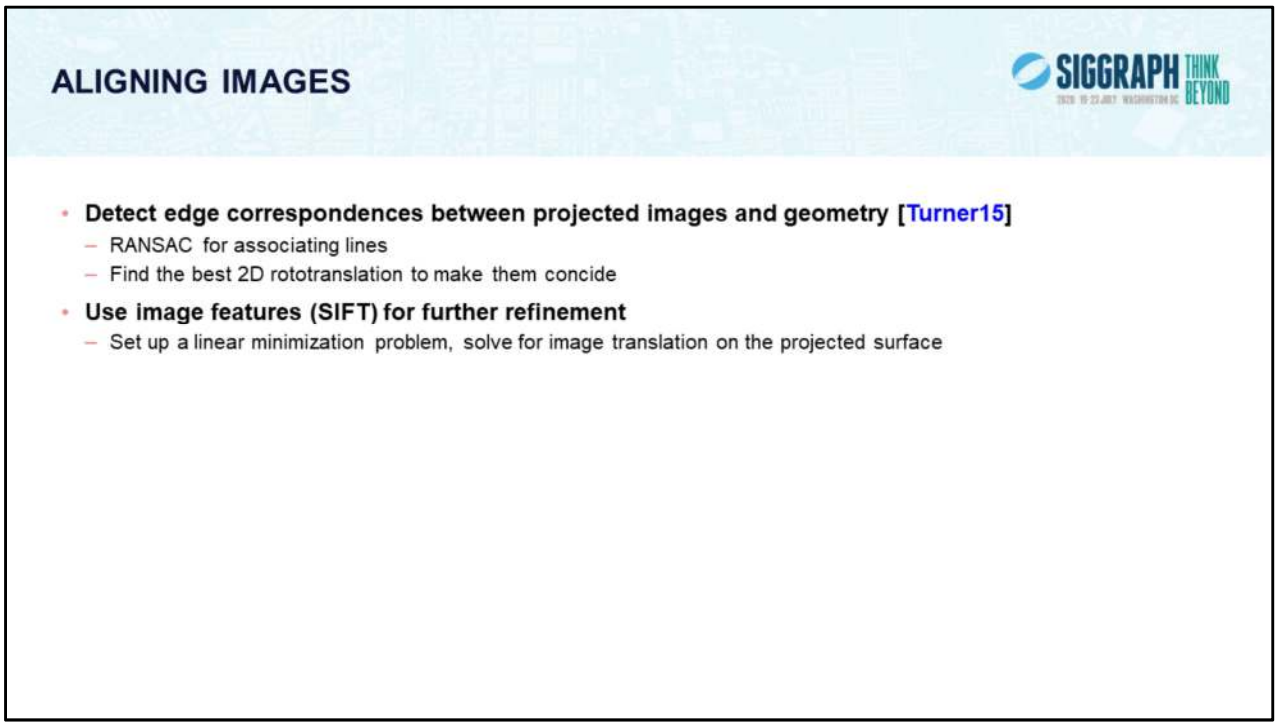
current state of the art.



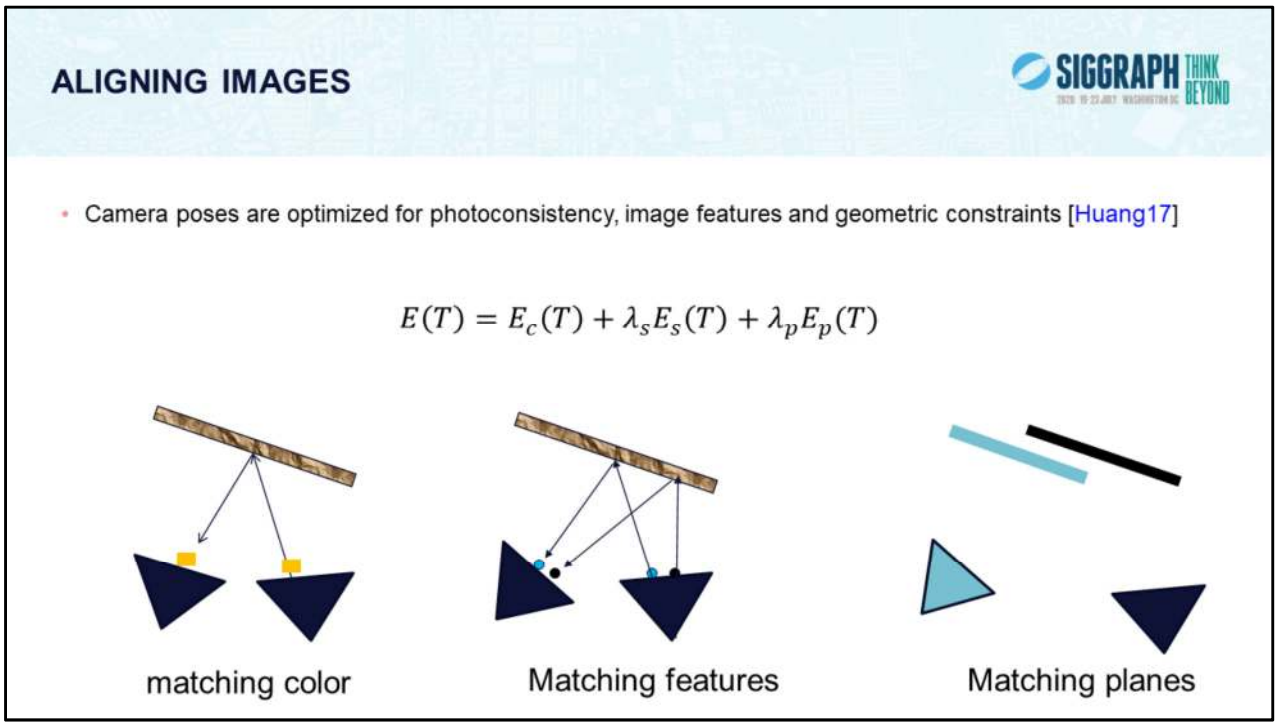
First of all, let us point out a few differences between how we are used to think to texturing in the field of 3D scanning when targeted to accurate reconstruction of specific objects and texturing indoor environements. In the first case we aim at precise and accurate 3D reconstruction, which gives us a reliable geometry which will match with a photograph of the same object down to pixel precision. On the other hand, when reconstructing indoor environments we may easily end up with approximations of the order of centimeters (especially with image-based methods) that would often serve as a little more that a proxy geometry.

Second, in 3D scanning we usually perform an acquisition campaign specifically targeted to texturing, or even to BRDF, in a controlled lighting environments or, anyway, color data are acquired in the most favorable time and point of view. In contrast, lighting in indoor acquisition depends on windows and light sources present in the rooms.
In summary, in 3D scanning texturing is a fundamental part of the reconstruction process while in reconstruction of indoor environmenst it is almost a by-product and for this reasons it needs ad-hoc tecniques.
In this context, we may reduce the problem o texturing to two sub-problems: finding a correct alignment between images and reconstructed geometry; assign the color to

each texel by appropriately considering the contribution from the aligned images.

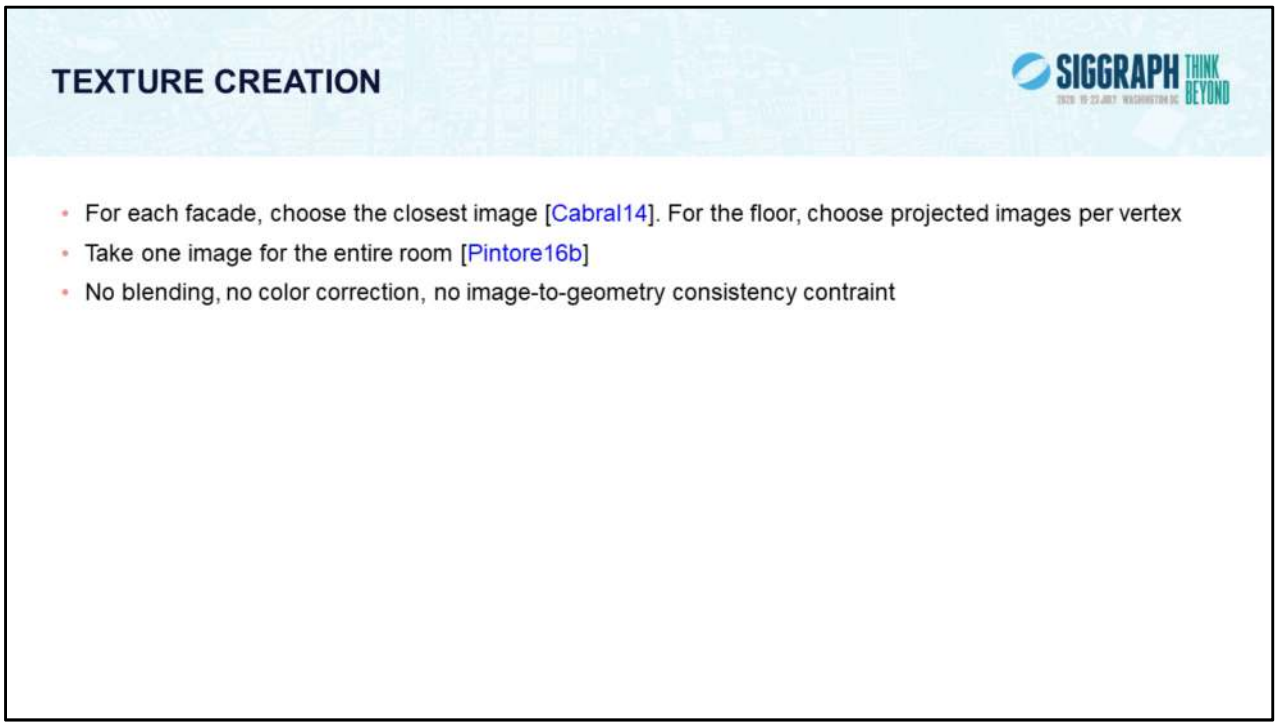


Concerning the image-to-geometry correspondences, Turner use Hough transform to find edges in the images and they try to find the correspondences with the edges of the triangulation, specifically those separating walls from the floor. Then they rototranslate the cameras so to minimize the difference between corresponding edges in the projection (shown in blue and red, respectively).
The consistency of texture on the projecting surface is addressed with a second minimization procedure on the SIFT features of overlapping images, where only 2D translation are considered.



Huang and colleagues [Hunah 17] proposed a system called «3D Lite» to reconstruct 3D environements by using consumer RGB-D sensors. They use bundle fusion to reconstruct the geometry and than use plane priors to abstract the whole 3D asa set of planar primitives. At this point, the depth associated with each image is no more the original depth image but the rendered depth of the finalized geometry.

The alignement of images is obtained by minimizing a 3 term energy function over the camera poses. A term accounts for the photoconsistency of all points, a term for the matching of point features and a term for correspondence of planes.



The most straighforward way to produce a textured geometry is a simple projection of the images onto the geometry. This is done by several algorithms, when producing a faitfhul visualization is not their main goal.
For example, in their seminal paper on image-based reconstruction from panoramic images, Cabral and Furukawa [Cabral 14] output the wall boundary as a set of planar quads and, for each quad, choose the closest camera and project the relative image. The floor instead is provided as a triangulated surface and the images are chosen per point. A similar approach is used by Pintore et al. but only one image is used per room, which is assumed to be star-shaped.
This simple approach falls short in several regards: First, all the 3D objects in the room will be splat onto the boundary and on the floor; second, even the smallest error in geometry estimation will be made apparent by projective distortion, as it can be seen on the part of walls projected onto the floor; Third, different choices of the camera in nearby surfaces are immediately noticeable both for the difference in projection and the different exposure and lighting.

TEXTURE CREATION

- The goal is avoiding ghosting, seams, blurring etc...
  - Exposure correction [Zangh16,Huang17]: auto white-balance returns image with different exposures
  - Graph-cut optimization: solve the labelling problem to assign texel to images

$$E(L) = \sum_{f\_i} D(l_i) + \lambda \sum_{(f_i, f_j) \in N} V(l_i, l_j)$$

[Sinha08,Xiao14]
$D(l_i)$: combination of angle and distance from the median color + user's strokes on the images
$V(l_i, l_j)$: sum of color and gradient difference of the projections

[Huang17]
$D(l_i)$: difference with the sharpest score on images
$V(l_i, l_j)$: sum of color difference
Further smoothing step

$$E(F) = \sum_{p} ||F(p) - \bar{C}(p)||^2 + \lambda ||\triangle F(p) - \triangle C(p)||^2$$

The goal of combining the contributions of aligned images to produce a final texture is avoiding visible artifacts due to remaining misalignments, differences in exposures, blurred images and so on.
Current approaches typically use exposure correction by finding a mapping function to map the color between images.
Then a labelling problem is solved to assign image sources to point on the geometry, minimizing a two term energy function.
The first term accounts for the error on choosing a source for a specific point, the second term is used for spatial coherency, that is, to avoid too many changes of assignment for neighbor elements which would tend to create visible borders.
It is typical for the first term to incorporate the view angle, however more information can contribute to the term. In the work Sinha and colleagues uses the distance from the median color of all projecting images is used to avoid outliers. Also, they allow user input as strokes in the source images to favor or prevent specific regions to be used.
The second term typically contains a measure of the difference between the projection of neighbors in both associated images plus a measure of the respective gradients. In the variation proposed by Huang, the first terms is instead designed to favor the texture that is locally sharper w.r.t. the projection of the point (which implicitly includes projection angle). Also, they introduce a final smoothing step by reintroducing a contribution from the simple average of all images. This is done by including the difference between the laplacian of their projections in the energy function.

BEYOND SUBJECTIVE COLOR: EXPOSURE CALIBRATION

- [Zhang16] uses a series of LDR images to compute to compute *exposure* and *radiance*, assuming pure diffuse materials

$$\min_{t_j, b_i} \sum_{i,j} \left(t_i b_j - X_{ij}\right)^2$$

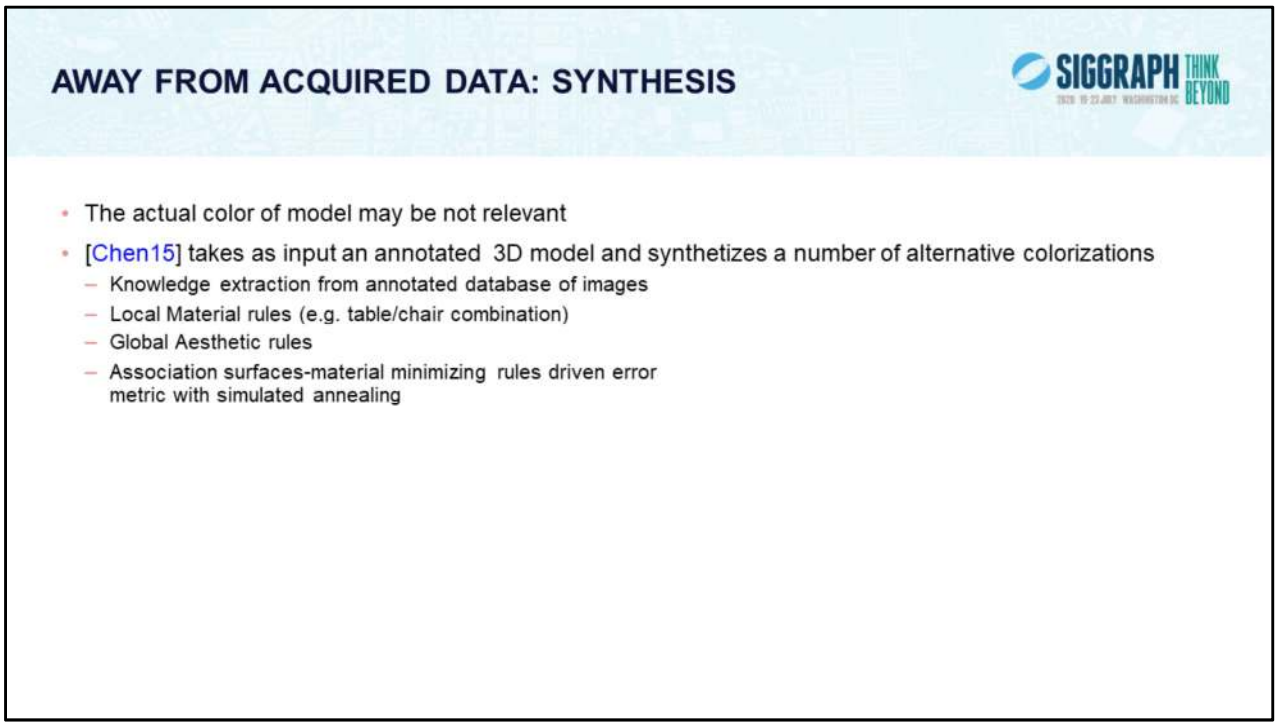
t_j *exposure image* j
b_i *radiance vertex* i
X_{ij} *pixel value of vert* i *proj on* j

As aforementioned, the image acquisition is typically done in a uncontrolled lighting environment and images will have to be corrected prior their combination.
When using RGB-D cameras, many images will be available and, thanks to the automatic white balancing of the camera, they will be in a possibly wide range of exposures.
Zhang et al. Took advantage of it by assuming a purely diffuse environment and setting a minimazion problem over radiance and exposure values for each vertex of the mesh.
I short, they created and HDR textures from a set of LDR images, which of course allowed them to relight the scene at different exposure.

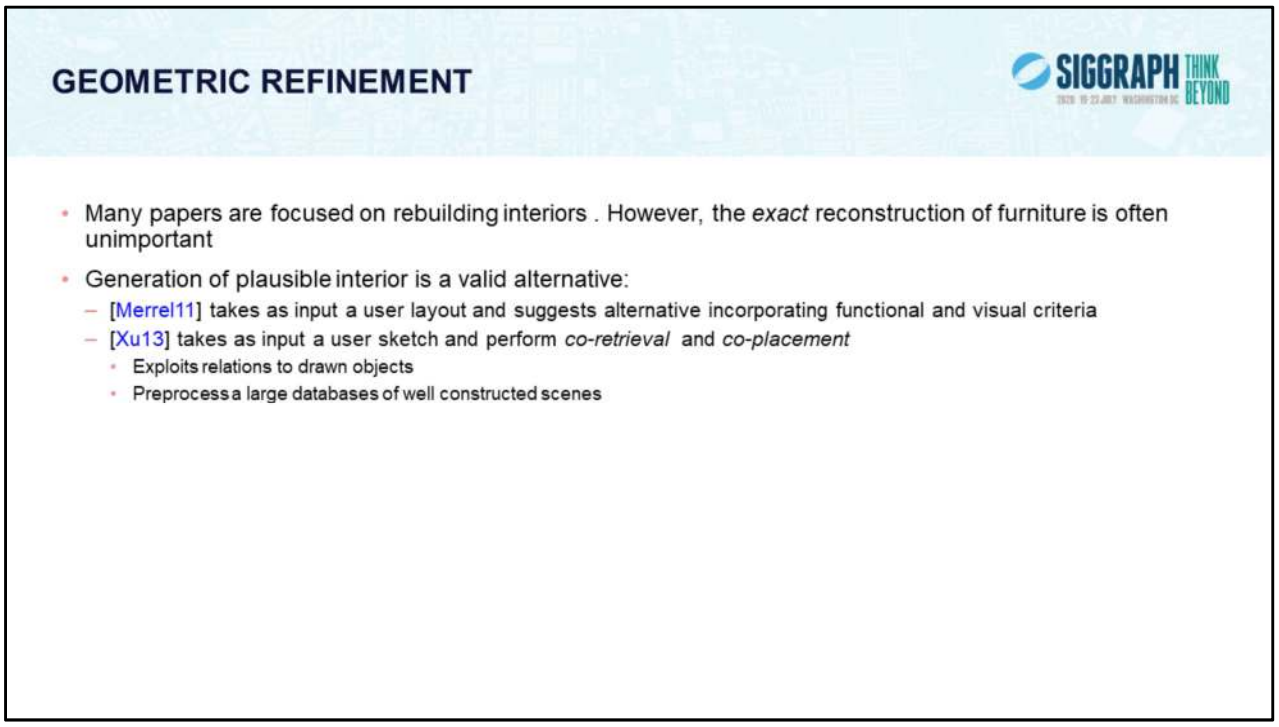
BEYOND TEXTURED GEOMETRY: VDT

- View Dependent Texturing in the context of remodelling [Colburn13]
  - Editing of the scene from a set of predefined points of views and transitions
  - Dynamically project HQ images on the visible geometry, optimized for the point of view and LQ images during transitions
- View Dependent Texture Atlas
  - For each view, pack the contribution of all images to the geometry in the **view frustum**

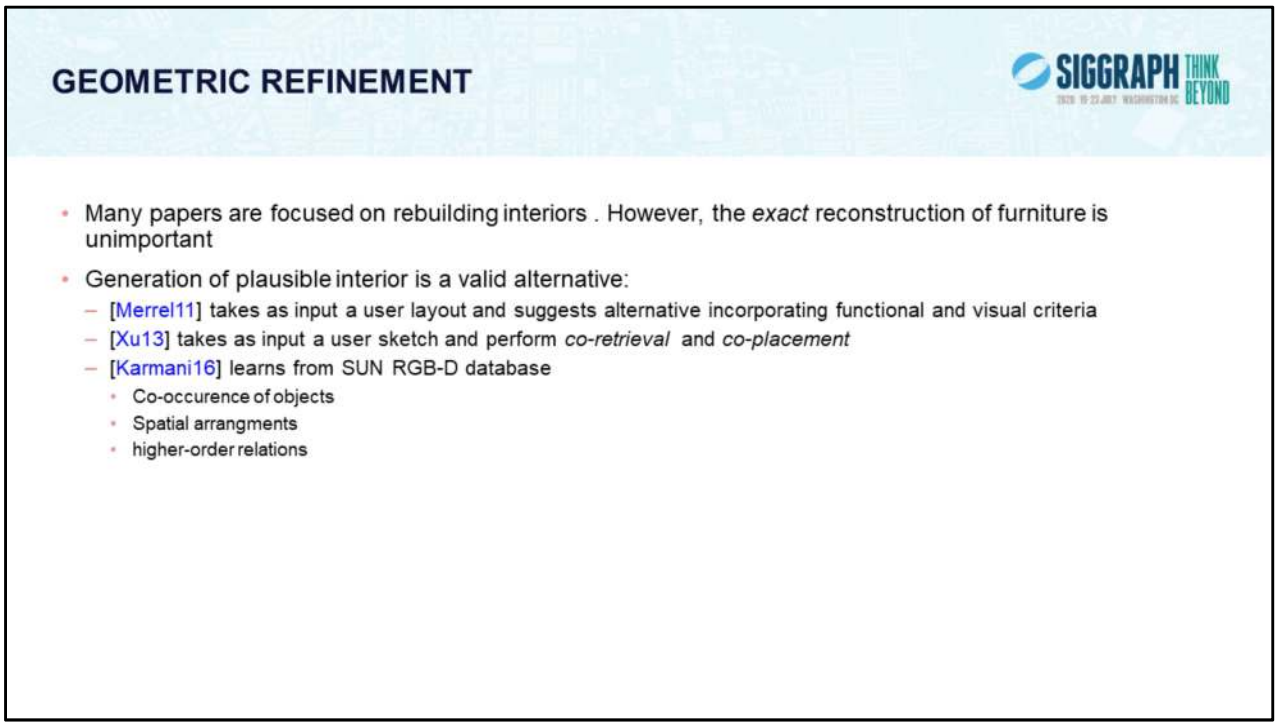
So far we considered only solutions that target the creation of a view independent textured geometry, but there are applications where this is not necessary.
For example Colburn and colleagues [Colburn 15] implemented a system for indoor remodelling that uses view dependent texturing. The advantage of VDT is that from the point of view of the camera the rendering is very realistic, since it's the actual photograph of the scene. On the other hand, if you want to change the geometry of the scene the image would not correspond to the new geometry. Their idea is to store, for each editing view, the set of images projecting onto the geometry within the view frustum so, in this example, the images of the room unveiled by the new door.



There are also many cases where the actual color appearance of the environment is not important and all we need is just a plausible colorization.
In these cases a viable alternative is to synthetize the color. For example Chen and colleagues [Chen 15] propose a method that takes as input an annotated 3D scene and porpose a number of alternative colorizations.
They use annotated databases of images (Opensurface) and infer local materials rules, that is, rules that bound the color of the table and then chair, or the sofa and the arm chair and on, and added global ahestetic rules, then set up a minimization problem over the material type.



As for color appearance, there as many applications where the actual layout of the furniture itself is unimportant and a plausible one can be syntetized. User input can be taken for creating room layout. For example Merrel and colleagues proposed a system where the user edits the 3D scene and the elements of the scene are re-arranged following visual and functional criteria. Xu et al instead take user input in the form of a sketch and infer placement rules from a database of over 700 well costructed scenes taken from google warehouse



Kermani et al instead use the SUN RGB-D database, which contains 10000 depth images from real world scenes for learning co-occurrences, spatial arrangments and higher order relations.

GEOMETRIC REFINEMENT

- Many papers are focused on rebuilding interiors . However, the *exact* reconstruction of furniture is unimportant
- Generation of plausible interior is a valid alternative:
  - [Merrel11] takes as input a user layout and suggests alternative incorporating functional and visual criteria
  - [Xu13] takes as input a user sketch and perform *co-retrieval* and *co-placement*
  - [Karmani16] learns from SUN RGB-D database
  - [Fisher15,Fu17] create activity-centered layouts
    - Activity-maps are 2D maps that spatially describe actions done in the environment
    - Derive AM from partial scans by a trained classifier

Finally, there are approaches such as done proposed by Fisher et al and by Fu et al that create layouts based on the functionality of spaces. Here the idea is to infer activity maps from partial scans using a pretrained classifier, and then to synthetize the scene accordingly to the activity map.

VISUALIZING INDOOR ENVIRONMENTS

- Rendering itself mainly requires a good visibility culling strategy [Cohen-Or03]
- Solutions vary on:
  - Methaphore for the interface
    - How to move, where to look
  - Type of data visualized
    - Images
    - Textured geometry
    - Videos
    - Maps

When it comes to visualizing indoor environments, the rendering itself is not more difficult than in the general case. If anything, the room structure allows to take full advantage of visibility culling techniques.
What is more interesting is the way the environment can be browsed, the methaphores used to control the browsing activity and the type of data that are used.
We all have experience with first person shooter games and keyboard-mouse combination to control point and direction of view, respectively.
However, very often freely browsing every hidden corner of every room is not what the application needs and the keyboard and mouse interation are not suitable for touch screen devices that simply don't have them.

BROWSING INDOOR ENVIRONMENTS

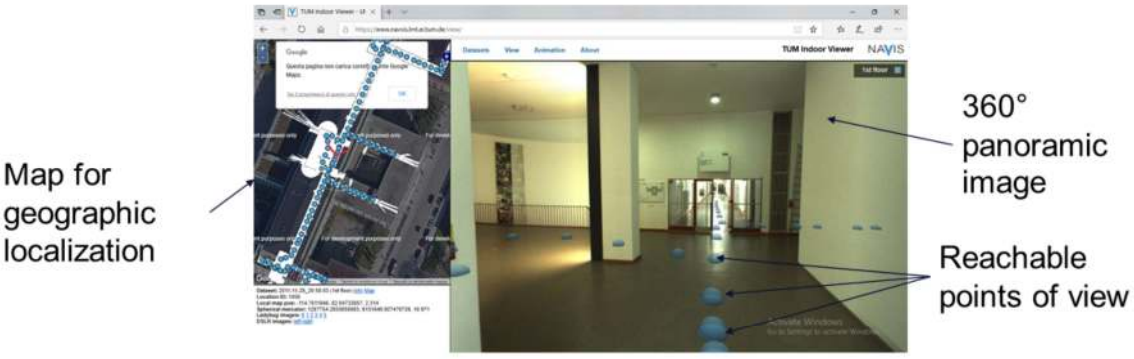
- [Sankar12] uses cylindrical panoramic images for each viewpoint
- Slides and tap for browsing
- Videos recorded during the acquisition for transition

Sankar, in its seminal paper on indoor acquisition with a smarphone, shows the environment as a collection of panoramic images, one per room. In their system, a clickable arrow is overlayed to move to the next room.

The transition is implemented by showing the video that was acquired during the acquisition process, while moving from a room the the next.

BROWSING INDOOR ENVIRONMENTS

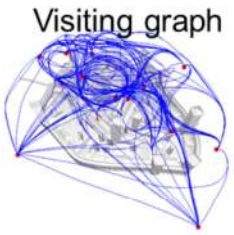
- TUMViewer [Nav12] uses a set panoramic images
  - Overlayed positions of other panoramas as spheres
  - Transitions with zoom-in and fade

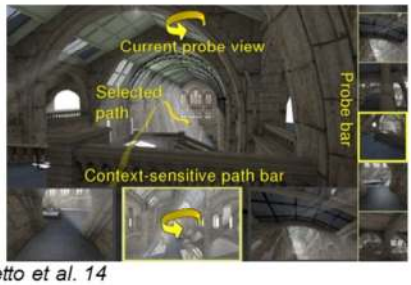


In the TumViewer (developed at the University of Munich and then a commercial product) each image is overlayed an icon (a small sphere) to indicate the location of the other images. Upon clicking a sphere, the view is moved to the corresponding panorama. During the transition, the current panoramic images is zoomed in to convey the sense of movement towards the next image and then faded into the new one.

BROWSING INDOOR ENVIRONMENTS

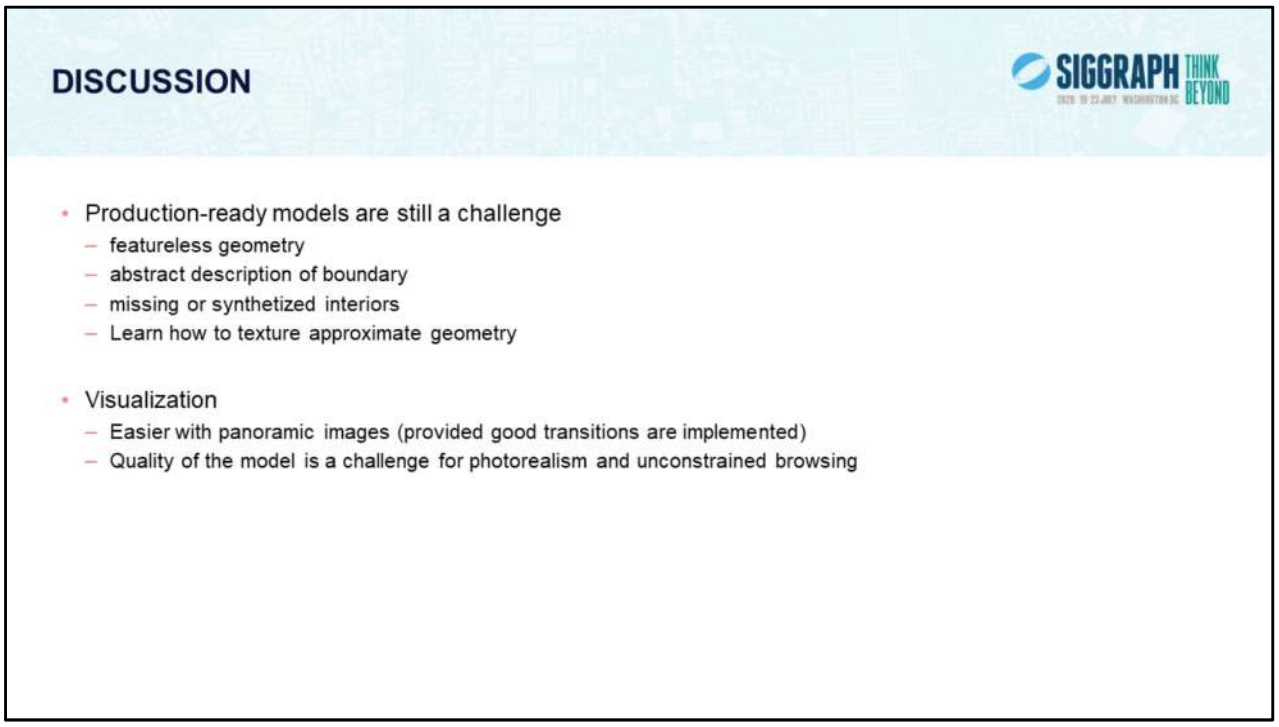
- [Di Benedetto14] also constrains the viewing position to a number of viewpoints and transitions
  - Visiting graph computed from the model
  - Panoramic images for view points
  - Arrows to show reachable viewpoints
  - Panoramic videos for transitions (or textured geometry [Pintore16b,Mat17])



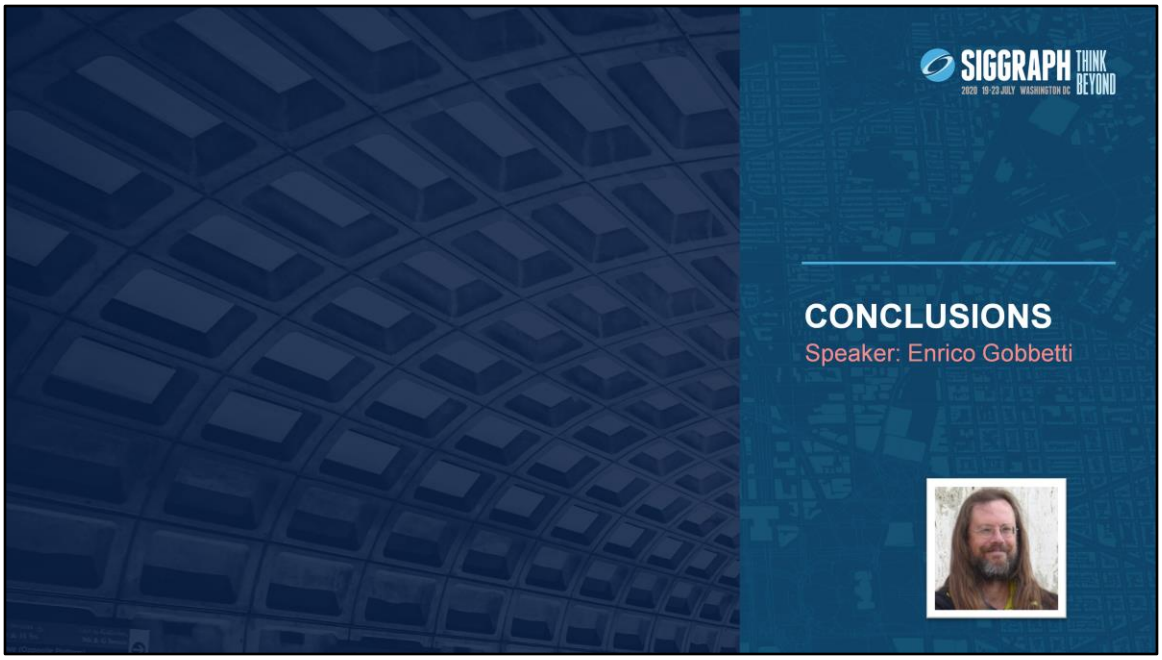


Di Benedetto et al. 14

Di Benedetto et al also use a set of panoramas. In their approach, a panoramic video is precomputed for each transition in a visiting graph, which is computed in a preprocessing phase. Since the video is panoramic, the viewing direction can be changed while moving from a point of view to the next.
Pintore uses the same technique by projecting the panoramic image onto the geometry, that is also what the Matterport showcase viewer does.



Translating a reconstructed indoor environment into a production-ready 3D model is still challenging.
The featureless nature of interior walls and, more importantly, the goal of providing an abstracted description of the boundary as a set of planes generate an approximation that clashes with the need to align the images to project.
Furthermore, the removal or poor sampling, or abstraction of interiors also breaks the consistency between images and reconstructed geometry.
In summary, the images show the real scene while the final geometry shows onnly an approximation of it. In this sense,
an interesting under researched avenue concerns developing new algoritmhs that allow to project, on, more in general, to adapt real images to approximated geometry.
On the visualization side, we can say that panoramic images offer the great advantage of having perfect rendering, even if from a finite number of locations and, provided that the transitions are well handled, are currently a preferred choice. Again, the quality of the final models is still a challenge for photorealistic visualization and unconstrained browsing.


SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC
CONCLUSIONS
Speaker: Enrico Gobbetti

THE STAR – PLEASE REFER TO IT!

- 1. Introduction
- 2. Related surveys
- 3. Background on data capture and representation
- 4. Targeted structured 3D model
- 5. Room segmentation
- 6. Bounding surfaces reconstruction
- 7. Indoor object detection and reconstruction
- 8. Integrated model computation
- 9. Visual representation generation
- 10. Conclusion

EUROGRAPHICS 2020
R. Mantiuk and V. Sundstedt
(Guest Editors)

Volume 39 (2020), Number 2
STAR – State of The Art Report

State-of-the-art in Automatic 3D Reconstruction of Structured Indoor Environments

Giovanni Pintore[1*], Claudio Mura[2*], Fabio Ganovelli[3], Lizeth Fuentes-Perez[2], Renato Pajarola[2], and Enrico Gobbetti[1]

[1] Visual Computing, CRS4, Italy
[2] Visualization and MultiMedia Lab, University of Zurich, Switzerland
[3] Visual Computing Group, ISTI CNR, Italy

Abstract

Creating high-level structured 3D models of real-world indoor scenes from captured data is a fundamental task which has important applications in many fields. Given the complexity and variability of interior environments and the need to cope with noisy and partial captured data, many open research problems remain, despite the substantial progress made in the past decade. In this survey, we provide an up-to-date integrative view of the field, bridging complementary views coming from computer graphics and computer vision. After providing a characterization of input sources, we define the structure of output models and the priors exploited to bridge the gap between imperfect sources and desired output. We then identify and discuss the main components of a structured reconstruction pipeline, and review how they are combined in scalable solutions working at the building level. We finally point out relevant research issues and analyze research trends.

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics;*** *Shape modeling;* ***Computer vision;*** *Computer vision problems; Shape inference; Reconstruction;* • ***Applied computing*** → *Computer-aided design;*

1 Introduction

The automated reconstruction of 3D models from acquired data, be it images or 3D point clouds, has been one of the central topics in computer graphics and computer vision for decades. This field is now thriving, as a result of complementing scientific, technological and market trends. In particular, in recent years, the widespread availability and proliferation of high-fidelity visual/3D sensors (e.g., smartphones, commodity and professional stereo cameras and depth sensors, panoramic cameras, low-cost and high-throughput scanners) has been matched with increasingly cost-effective options for large data processing (e.g., cloud and GPU-accelerated computation), as well as with novel means of visual exploration, from mobile phones to immersive personal displays.

In this context, one of the rapidly emerging sub-fields is concerned with the automatic reconstruction of indoor environments. That is, a 3D representation of an interior scene must be inferred from a collection of measurements that sample its shape and/or appearance, exploiting and/or combining sensing technologies ranging from passive methods, such as single- and multi-view image capturing, to active methods, such as infrared or time-of-flight cameras, optical laser-based range scanners, structured-light scanners, and LiDAR scanners [BTS*17].

Based on the raw data acquired by these devices, many *general* surface reconstruction methods focus on producing accurate and dense 3D models that faithfully replicate even the smallest geometry and appearance details. In this sense, their main goal is to provide the most accurate representation possible of all the surfaces that compose the input scene, disregarding its structure and semantics or possibly only exploiting them to maximize the fidelity of the output surface model. A number of more *specialized* indoor reconstruction solutions focus, instead, on abstracting simplified high-level structured models that optimize certain application-dependent characteristics [IYF15].

The focus on high-level structured models is motivated by several reasons. First of all, their availability is necessary in many fields. For example, applications such as the generation or revision of building information models (BIM) require, at least, the determination of the bare architectural structure [MMdV*14, TCZ15]. On the other hand, information on the interior clutter, in terms of 3D footprint of major indoor objects, is necessary in many other use cases, such as guidance, energy management, security, evacuation planning, location awareness or routing [IYF15]. Even when the goal is solely for visualization, structured simplified models need to be extracted as a fundamental component of a renderable model. This is because

* Joint first authors

Hello, here is Enrico Gobbetti again, continuing to enjoy lockdown in the same indoor environment…

All good things come to an end (hopefully also bad ones), and it's time to wrap-up.

Our presentation has quickly covered, in tutorial form, the state-of-the-art in structured indoor reconstruction.

Again, we refer you to our recently published survey article for much more information and a detailed bibliography.

WRAP-UP

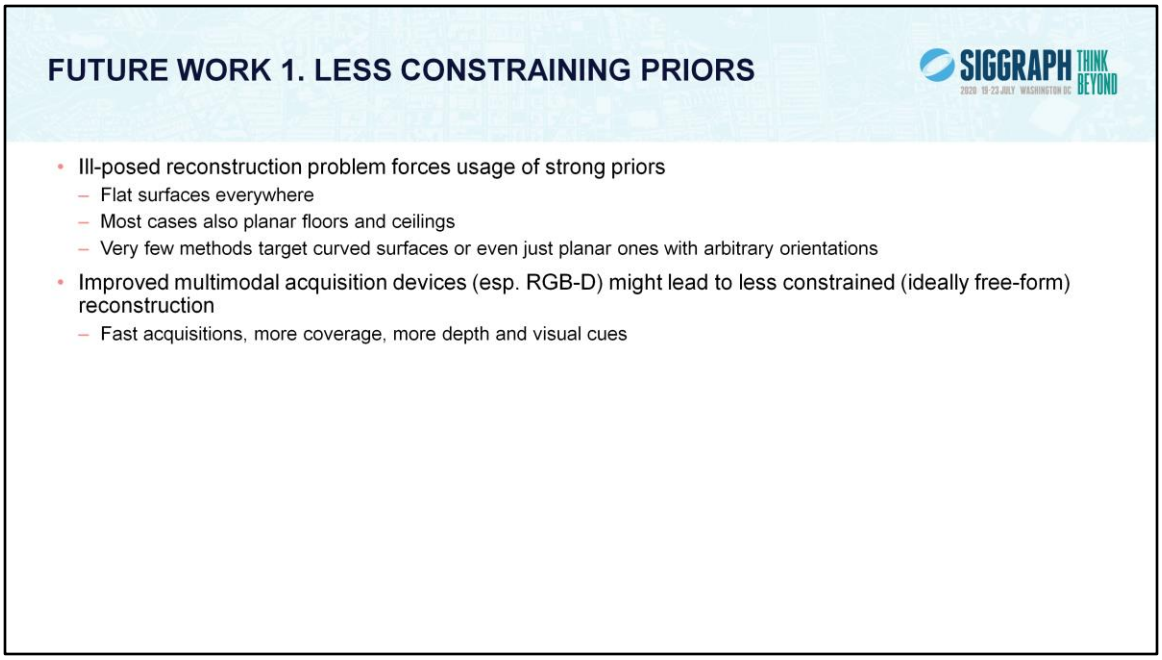
- Strong need for structured indoor models
  - High-level representation of main elements and their relations
  - Optimized to meet requirements of specific fields of application
    - Building Information Models (AEC domain): bare architectural structure
    - Emergency management, location awareness, routing: also interior clutter
- Many specialized solutions successfully deal with specific challenges of input data on many common building shapes
  - Mostly tuned for residential/office buildings
- What about future work?

To summarize, our survey has highlighted that structured indoor reconstruction is a well-defined area of research in itself, that requires specialized solutions to produce the desired structured models from captured samples.

The overall problem is very challenging, but we have shown that research has witnessed substantial progress in the past decade, growing from methods handling small-scale single-room simple environments, to techniques that handle substantial artifacts and produce high-level structured models of large-scale complex multi-room buildings.

However, several areas remain for future work.

I'll conclude my talk, and the overall tutorial, with just a quick overview of the main identified ones.



The first important point is that, in order to offer robust solutions to the ill-posed reconstruction problems, very strong priors are typically imposed.

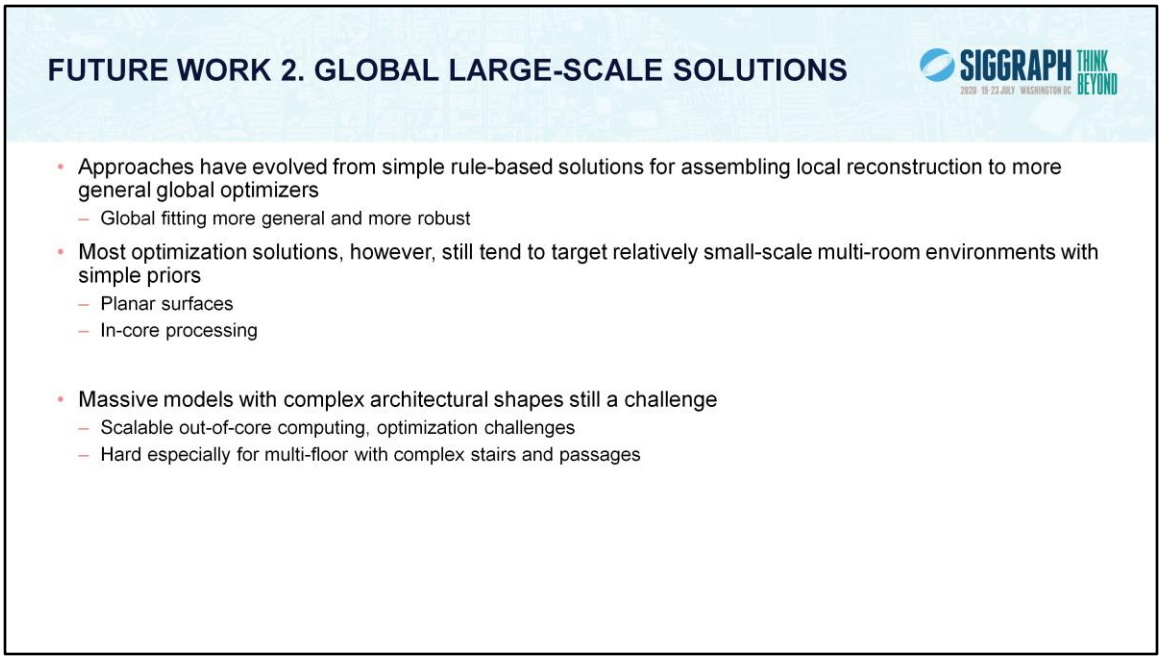
For instance, very few methods target complex ceilings or curved surfaces.

Relaxing such constraints is a main avenue of future work.

Besides better solvers, an emerging aspect is that improved multi-modal acquisition devices (for instance RGB-D cameras) might be very helpful in this area.

This is because by providing denser visual and depth coverage they reduce reconstruction ambiguities.

Usage of RGB-D data, as opposed to purely visual or purely geometric input is thus becoming a constant for most solvers tackling complex models.



An second important take-home message is that research approaches are increasingly evolving from simple rule-based methods to assemble local reconstruction to more general optimizers that strive to globally minimize some fitting function.

However, most optimization solutions still tend to target relatively small-scale multi-room environments with simple priors, using, for instance, planar surfaces and fully in-core processing.

The global reconstruction of massive models with complex architectural shapes is still a challenge.

This requires progress both on global solution methods and on their efficient parallel multi-scale implementation on large environments.

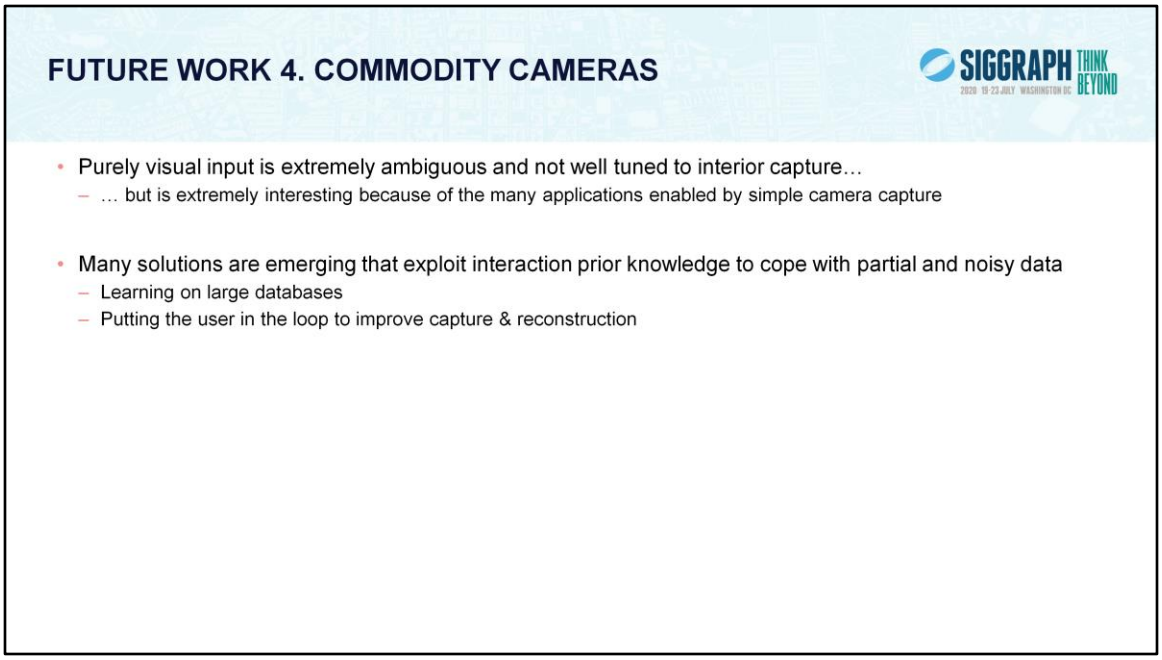
FUTURE WORK 3. DATA FUSION

- Recent years have seen the consolidation of multi-modal capture systems
  - Especially RGB-D cameras
- Very few solutions, however, jointly exploit both color and 3D data
  - Most methods using both channels handle them in separate stages
- Performing data fusion to combine visual and depth cues into multi-modal feature descriptors on which to base further analysis is an important avenue for future work
  - Joint analysis promises to better cope with heavily cluttered and partial acquisitions

As already mentioned, moreover, the recent years have seen the consolidation of multi-modal capture systems, which provide lots of information to work with.

Very few solutions, however, jointly exploit color and 3D data, which are typically handled in different stages of the pipelines.

Performing data fusion to combine visual and depth cues into multi-modal feature descriptors is therefore a very important avenue of future work, since joint analysis promises to better cope with heavily cluttered and partial acquisitions



At the opposite, purely visual input is extremely ambiguous and makes reconstruction very hard.

Regular and panoramic cameras are therefore not the best capture device to use for inferring the shape of an indoor environment.

On the other hand, indoor reconstruction from camera capture is extremely interesting for a number of applications and is possibly the only way to provide tools usable for the masses.

Therefore, many many solutions are emerging in this area, and given the strong ambiguities they need to strongly exploit prior knowledge or additional capture information.

Two recent trends in this area are to exploit knowledge learned from large databases, as well as putting the user in the loop to improve capture & reconstruction.

In particular, it is not uncommon to exploit data from additional sensors such as inertial units, for instance by tracking paths.

FUTURE WORK 5. DATA-DRIVEN APPROACHES

- Research is increasingly moving from exploiting hard-coded priors to learning those priors from data
  - General trend in computer vision and computer graphics
  - Potentially well adapted to man-made indoors, which exhibit strong repetition
- Combination of geometric reasoning with purely data-driven solutions very promising for large-scale reconstruction
  - Need for annotated model databases for training and evaluation purposes

Finally, the trend of moving from hard-coded priors to priors learned from data is a general one, and not limited to purely visual input.

This is not surprising, since learning-based solutions are a becoming a general approach for many problems in computer vision and computer graphics.

In our context, they are also especially well adapted, since man-made indoors exhibit strong regularities that can be learned from examples and are transferable across model kinds.

For the same reasons, a combination of geometric reasoning with purely data-driven solutions is particularly promising because of the mix of geometrically regular architectural elements and freely dispersed indoor clutter.

In this context, the rapid emergence of annotated real and synthetic models databases is very important for both training and evaluation purposes.

Our survey includes a detailed list of the currently available ones.

And with this final remark, and yet another referral to our article, we have come to an end.

That's it for today!

Due to the exceptional period we are living in, this tutorial could not be given live, and it's not possible to have the standard Q&A Session.

However, the virtual conference format should also provide some way to interact, which is however still not clear at the moment of preparation of these notes and of recording of our talk.

In addition all authors may be contacted by e-mail, and some of them will even sometimes answer…

Thanks for your attention!


SIGGRAPH
THINK
BEYOND
2020 19-23 JULY WASHINGTON DC
THE END
© 2020 SIGGRAPH. ALL RIGHTS RESERVED.