Free your Camera: 3D Indoor Scene Understanding from Arbitrary Camera Motion

Axel Furlan[1]
furlan@disco.unimib.it
Stephen Miller[2]
sdmiller@stanford.edu
Domenico G. Sorrenti[1]
sorrenti@disco.unimib.it
Li Fei-Fei[2]
feifeili@stanford.edu
Silvio Savarese[2]
ssilvio@stanford.edu

[1] Computer Science Department
University of Milano - Bicocca
Milano, Italy

[2] Computer Science Department
Stanford University
Stanford, CA, USA

Abstract

Many works have been presented for indoor scene understanding, yet few of them combine structural reasoning with full motion estimation in a real-time oriented approach. In this work we address the problem of estimating the 3D structural layout of complex and cluttered indoor scenes from monocular video sequences, where the observer can freely move in the surrounding space. We propose an effective probabilistic formulation that allows us to generate, evaluate and optimize layout hypotheses by integrating new image evidence as the observer moves. Compared to state-of-the-art work, our approach makes significantly less limiting hypotheses about the scene and the observer (e.g., Manhattan world assumption, known camera motion). We introduce a new challenging dataset and present an extensive experimental evaluation, which demonstrates that our formulation reaches near-real-time computation time and outperforms state-of-the-art methods while operating in significantly less constrained conditions.

1 Introduction

The indoor scene reconstruction problem has sparked lively interest in the research community in the past few years. The ability to understand the structural geometry of living spaces allows, for example, autonomous robots to safely move in the surrounding environment, as well as the development of augmented reality applications. Many contributions have been proposed to either address the problem of recognizing semantically meaningful components, such as floor or walls in 2D images, or the problem of generating sparse 3D point-cloud maps of the observed scene while moving within it. By contrast, few works have addressed the problem of reconstructing semantically consistent indoor structures (also referred to as 3D scene layout, see Figure 1), and refining layout hypotheses by integrating new evidence acquired as the observer moves around. Among those that do, simplifications are made to

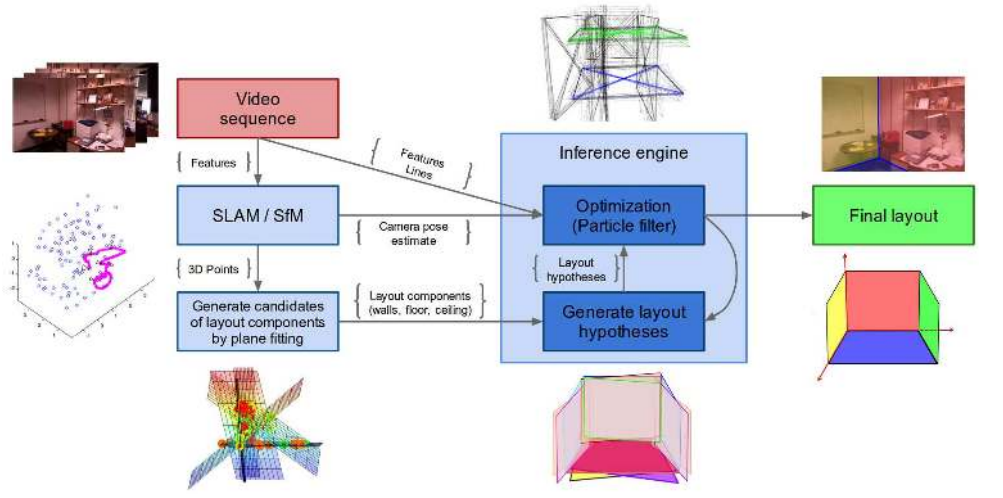


Figure 1: **3D scene layout estimation process.** The video sequence is first processed to obtain camera localization and sparse 3D point cloud reconstruction. Layout components (e.g. floor, ceiling, walls) are generated from the sparse 3D points and combined to generate layout hypotheses (see Section 2). Each layout hypothesis is evaluated and optimized by incorporating new image evidence. The final 3D scene layout is represented by the hypothesis that better describes the scene (see Section 3).

solve the problem, such as the Manhattan world assumption, partially or fully known observer motion, and occasionally human intervention for a correct initialization.

In this paper we pursue a semantically consistent reconstruction of the structural elements in indoor scenes, excluding any a priori knowledge of the observer's motion, human intervention, or hard Manhattan constraints. We start by observing in our experiments that state-of-the-art 3D reconstruction techniques, such as SLAM and SfM, are far from achieving reliable results that can be directly translated to a higher semantic reconstruction. This phenomenon is particularly noticeable when the observed scenes are highly cluttered and/or mostly featureless. We therefore propose a probabilistic framework, to be fed with such noisy elaborations, that allows us to efficiently integrate new evidence to generate, evaluate and refine 3D layout hypotheses as the observer moves through the scene. Different types of information extracted from images (points, lines, regions) and, potentially, from other kind of sensors, can be easily merged into our probabilistic formulation, allowing us to cope with both clutter and featureless surfaces.

We present extensive experimental results on challenging sequences, demonstrating improvement over state-of-the-art approaches while operating in significantly less constraining conditions and in near-real-time (in the order of $\sim$20fps). We propose a new challenging dataset to prove the full capacities of the proposed method and which is available for future comparisons [1].

1.1 Related work

Sparse 3D reconstruction. While the objective of this work is 3D indoor scene layout understanding, our approach hinges on the success of sparse reconstruction methods, such as structure from motion (SfM), simultaneous localization and mapping (SLAM), or parallel tracking and mapping (PTAM). SfM methods usually process sparse set of images from which partial geometry is recovered and use optimization methods (e.g., bundle adjustment)

to obtain reconstructions that are globally consistent [13, 32, 35]. They are often used to model large scale environments such as cities [4, 22, 28] and are capable of fusing data from different types of sensors like mono/stereo cameras, IMU and GPS. SLAM methods are primarily based on Bayesian filtering [5, 6, 20, 21, 29]. A key feature of all SLAM approaches is to update the filter status with new evidence extracted from each image in the sequence. An alternative strategy is taken by approaches such as PTAM [17] which, instead, rely on optimizing a smaller number of key-frames using bundle adjustment. [12] move toward a higher level representation by introducing lines and planes into a standard EKF SLAM system, which allows them to collapse the EKF state space. Most of these reconstruction methods, however, just recover the environment as a collection of sparse 3D points, lines or planes, without being able to identify the important semantic phenomena in the scene.

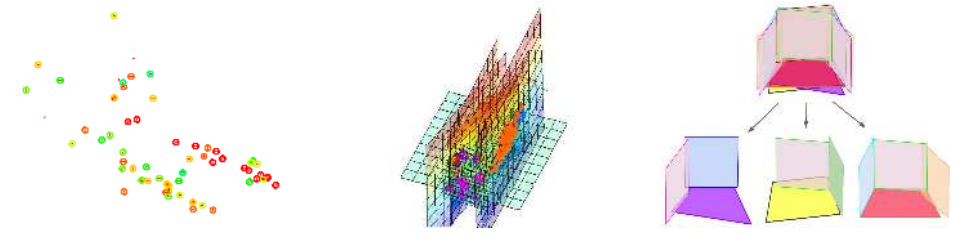
Single image. Many works have been recently proposed to solve the problem of indoor scene layout estimation from single images. Most of these methods leverage machine learning techniques for solving the inherent ambiguity of the 3D-to-2D mapping [2, 14, 15, 16, 18, 24, 25, 27]. The advantage of these techniques is that they reconstruct the scene as a collection of semantically meaningful components such as floors, walls or doors and identify those in 2D images. Reasoning about vanishing points or lines of the scene is often used to help segment regions of interest. These methods, however, mostly focus on obtaining a 2D or, often, a 2.5D layout estimation of the scene (i.e. identify walls and floors in the image) and fail to achieve an accurate metric 3D estimation of the geometrical properties of the environment as we seek to do. Moreover, none of these achieve real-time performances, except for [25].

Multiple images/sensors. A number of works have looked at inferring the scene layout from multiple images. Sinha *et al*. [27] proposed a multi-view stereo method to generate piecewise planar depth maps from sparse images. This approach assumes that strong line elements can be extracted from images. Flint *et al*. [7, 8, 9] present a model that leverages the Manhattan world assumption to estimate dominant vanishing points and relates them with learned photometric cues. Similarly, Furukawa *et al*. [10, 11] rely on Manhattan assumption, but model the reconstruction problem as an MRF. Xiao *et al*. [36] propose a method (Inverse CSG) based on dense 3D laser scans and imagery to obtain photo-realistic reconstructions of museums.

Real-time. All of the approaches to semantic layout estimation perform well below real-time speeds and either rely on slow robust SfM camera pose estimation and/or on scene-dependent learning. Tsai *et al*. [33, 34] focus on real-time indoor scene estimation tasks for mobile robot applications. They exploit video sequences to validate and support layout hypotheses and drop the Manhattan world assumption. However, they tackle a simplified problem, where the observer is a robotic agent with either known roto-translation between camera and floor plane [34] or known full pose of the robot at each time step [33]. Furthermore, they generate layout hypotheses by projecting floor-wall intersection boundaries (observed in the images) onto the known ground plane, which means that if those lines are not observed, the entire layout hypothesis will not be generated.

1.2 Contributions

In this paper we propose an efficient method for estimating 3D indoor layout from an arbitrary 6DoF moving monocular observer, whose motion is estimated usingstate-of-the-art



(a) (b) (c)

Figure 2: Left: a noisy sparse 3D reconstruction. Center: a large number of candidate layout components obtained with plane fitting. Right: The process of generating layout hypotheses as random combinations of candidate layout components.

techniques such as SLAM and key-framed SfM. Our method builds upon prior work in the following ways:

- Eliminating the hard Manhattan world assumption.
- Requiring no *a priori* knowledge of the observer motion with respect to the scene.
- Operating at near-real-time speeds (∼20fps).
- Introducing a new challenging dataset which features cluttered, non-Manhattan and non-box-shaped scenes.

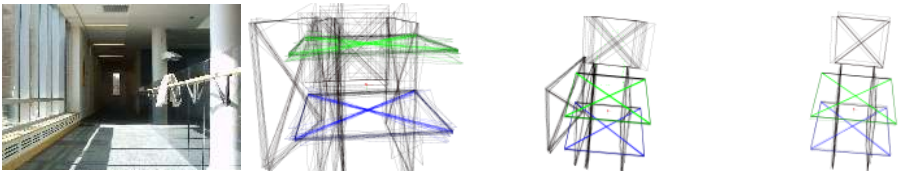
2 Method Overview

In this section we describe the framework within which layout hypotheses are generated, evaluated and updated or rejected. Figure 1 shows a pictorial representation and a schematic diagram of the whole process.

Sparse 3D reconstruction. As the observer moves in the surrounding environment and the image stream is acquired, we first pre-process sequences with a localization and sparse 3D reconstruction algorithm. Since our main contribution is not related to this problem, we designed our framework to be able to work with any algorithm that can provide a camera motion estimation and a cloud of 3D points. In our experiments we compare two such approaches: a real-time implementation of the Monocular V-SLAM approach proposed in [23] and the non-real-time VisualSfM [35]. Notice that these 3D reconstructions are in general noisy and sparse (Figure 2(a)).

Hypotheses initialization. The second step consists of generating a higher level representation of the 3D points estimated in the pre-processing phase. Several types of geometrical primitives are suitable for this purpose. In our case, we believe a piecewise planar representation is the most appropriate for indoor scene representation. We fit a large number of planes to the 3D points so as to generate a large number of (potentially inaccurate) candidates of layout components, i.e. walls, floor, ceiling (Figure 2(b)). In our experiments we implemented an iterative RanSaC plane fitting procedure, which we optimized for indoor scenes by allowing peripheral fitted points to be re-injected in the iteration process, since these points potentially lay on the intersection of two planes.

Layout estimation. In the last step, which constitutes the core of our proposed inference engine, we generate layout hypotheses as random combinations of candidate layout components (Figure 2(c)). Each layout hypothesis is evaluated at each time frame by measuring



(a) (b) (c) (d)

Figure 3: From left to right, an example of indoor scene and of the evolution in time of particles throughout the optimization process. Note how the unlikely layout hypotheses disappear while the most plausible ones survive and get refined.

its compatibility with observations (*e.g.* image points and lines) and geometrical constraints across frames. During this process, each layout is "perturbed" by locally adjusting, optimizing, merging or splitting layout components. There are different approaches to manage sets of hypotheses. In this paper we choose to integrate our probabilistic framework within a particle filter structure. This choice allows us to explicitly formulate the problem in a parallel-computing oriented fashion (particles are independent from each other), which can lead to high efficiency gains in computation time. The output of the optimization procedure is an estimation of the 3D scene layout, which is obtained by selecting the layout hypothesis with the best set of layout components (see Section 3).

Advantages. There are two important advantages stemming from the choices mentioned above: i) the local optimization step applied to each layout component helps recover from noisy initialization (transition from Figure 3(b) to Figure 3(c)) while keeping the computation amount affordable for real-time applications. This could not be achieved with standard "brute force hypothesize and test" approaches, *e.g.* [18, 19]; ii) the choice of embedding our probabilistic formulation within a particle filter also allows us to e exploit some critical properties of particle filters, like multi-modal posterior representation, re-sampling, particle clustering and, most importantly, recover from substantially wrong

initializations (transition from Figure 3(c) to Figure 3(d)).

3 Probabilistic Layout Estimation

As stated above, the input to our inference engine is an image sequence from which the camera motion and sparse 3D points are estimated. These 3D points are used to generate layout components (walls, floor, ceiling) by fitting a large number of planes through such points (see Section 2). The output of the inference engine is the layout hypothesis that better explains the scene in terms of compatibility with observations and geometrical constraints. At the heart of our approach is a particle filter-based optimization [30] that is capable of processing candidate layout components so as to obtain plausible scene layouts, where the layout hypotheses represent the particles of the filter. As with any optimization strategy, this requires four key components: a principled choice of the layout parameterization (3.1); an initialization strategy (3.2) to generate the initial layout hypotheses; a method for exploring the state space (3.3); a score function (3.4) to evaluate the quality of layout hypotheses. In the following section, we will discuss each of these aspects in details.

3.1 Layout Parametrization

While much prior work has leveraged the Manhattan world assumption, we believe this is a limiting hypothesis. To overcome this limitation, in this paper we adopt a representation similar to [33, 34] (sometimes referred to as Soft Manhattan), which makes the following

assumptions about the environment: i) ground plane and ceiling are parallel; ii) walls are only constrained to be orthogonal to the ground plane (and ceiling); iii) there can be any number of walls and each wall can be displaced at any angle with respect to other walls. Thus a room layout is fully parametrized by its gravity vector, the heights of its ground and ceiling, and a set of walls with only one degree of freedom for rotation and one for translation.

3.2 Initializing Layout Hypotheses

A layout hypothesis is generated as follows. We first determine the rough direction of gravity as given by an IMU or, if none is present, by assuming that the camera optical axis is roughly horizontal when the sequence begins. The height of the ground and ceiling are then approximated by the lowest and highest features tracked by SLAM/SfM along this direction. If they are not observed at initialization time, their heights will be underestimated at first and adjusted as more parts of the scene are observed. Subsequently, sampling from the initial set of planes fitted to the 3D points obtained from the SLAM/SfM reconstruction (Section 2), each layout hypothesis is assigned a random number of candidate walls (Figure 2(c)). While being assigned to the layout hypothesis, each wall is transformed by finding the minimal transformation needed to make it orthogonal to the ground.

3.3 Exploring the State Space

The above step gives a very rough estimate of the scene layout. In order to refine this estimate, we propose to "perturb" the layout components assigned to the layout hypothesis. This can be done within the hypothesis itself or by generating a set of new hypotheses. Given a layout hypothesis at time t-1, we may "perturb" it or generate a new hypothesis by:

- Rotating the ground plane about a random direction by some angle θ_g
- Translating the ground or ceiling by some distance, d_g or d_c
- Rotating wall i about the gravity vector by θ_w^i and optimizing its reprojection error
- Translating wall i by some distance d_w^i and optimizing its reprojection error
- Removing a wall which is currently hypothesized
- Adding a wall which had been removed
- Taking no action

where θ_g, d_g, d_c, θ_w^i, and d_w^i are normally distributed. For each particle in the filter, only a single action may be performed per time step, as determined by a weighted coin flip.

3.4 Scoring Hypotheses

At each timestep t, we wish to assign a probability to a particular hypothesis, taking into account new observations and geometrical constraints:

$$P_t = \prod_i P_f^i P_o^i(\theta_i) P_r^i(e_r^i) \prod_j P_m^{ij}(\phi_{ij}) P_s^{ij}(d_{ij}^{-1})^{p_{ij}} (P_w^{ij})^{a_{ij}} \tag{1}$$

where the binary terms p_{ij} and a_{ij} are initially zero. p_{ij} is set to 1 if walls i and j are near-parallel by some angular threshold and a_{ij} is set to 1 if they are adjacent. We have designed this probability to enforce a number of desirable properties.

Fitness: The initial pre-processing yields planes with varying goodness of fit. Thus, to each visible wall, we assign a corresponding fitness term P_f^i, represented by a zero-mean Gaussian over the residual least-square error after the plane fitting process.

Orthogonality to Ground: In Section 3.2 we described a method for generating orthogonal wall candidates from non-orthogonal planes. The more these planes are altered, the

less they are supported by data. We capture this with a zero-mean Gaussian $P_o^i(\theta)$, where θ denotes the amount of rotation required to orthogonalize the wall to the ground.

Low Reprojection Error: Many feature points tend to fall on the ground, ceiling, or walls of a scene. To make use of these visual cues, we track features using the Kanade-Lucas-Tomasi method [31]. To be robust to outliers, we discard those matches which i) could not have come from a plausible camera motion, or ii) don't share a homography with a minimum number of other points in the scene. At time t, we project the keypoints at $t-1$ onto all possible walls, and evaluate their hinge-loss 2D reprojection error. Each point is then assigned to the wall which minimizes its reprojection error, and the average e_r is computed. For each wall we assign a probability $P_r^i(e_r)$, which is normally distributed about 0 pixels.

Manhattan Layout: While we do not leverage the Manhattan assumption to generate layout hypotheses, we recognize that angles are far more likely to fall in 90° or 45° increments than, e.g., 87°. To capture this, for each pair of visible walls we include a term $P_m^{ij}(\phi)$, where ϕ is their relative angle modulo 90° (or 45°) and P_m^{ij} is a zero-mean Gaussian.

Simplicity: Adding more walls will always improve reprojection error. Actual layouts, however, are fairly simple: they are far more likely to contain one large wall than many small ones. To enforce this, for two near-parallel walls we assign a probability $P_s^{ij}(d^{-1})$ which captures how redundant wall i is given the presence of wall j, d meters away. This is normally distributed about $\frac{1}{d}=0$, or $d\to\infty$.

Wall-wall intersection: Small errors in the estimation of wall rotations can not be reliably captured by the reprojection error term. Yet, such small errors can lead to substantial errors in the displacement of the intersection between two walls. We exploit this intuition by assigning a probability P_w^{ij} that weights the image evidence supporting an intersection between walls i and j. To obtain this evidence, the 3D line segment resulting from the intersection is projected into the image and there compared against 2D line segments extracted with a Canny edge detector [3].

The final output of the optimization procedure is an estimation of the 3D scene layout, obtained by selecting the layout hypothesis which best describes the scene in terms of the score function in Eq. 1.

4 Results

In this section we show experimental results of our method when tested on the state-of-the-art dataset [34], as well as on a new challenging dataset that we introduce in this paper and that is available for future comparison [1]. To the best of our knowledge, the dataset [34] is the only state-of-the-art dataset that can be used for comparison for this type of problem. Since our method requires video sequences as inputs, some datasets cannot be used for evaluation because they feature single images [14], non-video (i.e. sparse) images [10, 11] or they are no longer available [7, 9].

4.1 Experimental setup

We first run two sparse 3D reconstruction techniques, RT-SLAM [23] and VisualSfM [35], and feed the generated 3D point clouds and camera pose estimations to our algorithm, which outputs the final 3D layout reconstruction. Final reconstruction results are compared to:

- **State-of-the-art approaches:** the video-based approach proposed in [34] and two well known single image methods, [16] and [14] (Section 1.1). For completeness, in Table 1 we report the results of [16] composed with a MRF over image frames. Please refer to [34] for a comprehensive description of this composition.

Method	Excl. ceil	Incl. ceil
[34]	**90.58**	82.17
[16]	82.62	83.30
[16]+MRF	81.44	82.13
[14]	84.70	84.33
Our + VSLAM	86.92	**87.01**

Table 1: Classification accuracy on the Michigan Indoor Corridor Video Dataset. Our results are compared to the results obtained with [34], [16] and [14].

Method	Clas. acc.	Avg. fps
Baseline	70.64	—
[16]	59.29	0.17
[14]	73.59	0.03
Our + VSLAM	**86.24**	**21.63**
Our + VSfM	75.94	16.90

Table 2: Classification accuracy on the proposed dataset. Our results (with SLAM and SfM) are compared to the results obtained with a naive baseline method, [16] and [14].

- **Baseline method:** in order to show the importance of the evaluation and optimization process (Section 3), we built a baseline method consisting in projecting all the possible combination of layout components (i.e. fitted planes, see Section 2) into the image and picking the combination that achieves the best classification accuracy.

In our experiments we evaluate the quality of the final reconstruction by means of the classification accuracy, which is a commonly adopted metric [14, 26, 34]. It is defined as the percentage of correctly labeled pixels when projecting the estimated 3D scene layout into the image. In order to evaluate if a pixel is correctly labeled, a groundtruth image is provided. Labels indicate if the pixel should belong to the ground floor, to the ceiling or to a wall numbered with an incremental counter. Please note that, for all the parameters described in Sections 3.3 and 3.4, the same configuration was used for all sequences.

4.2 Michigan Indoor Corridor Video Dataset

This dataset was proposed in [34] and consists of a set of image sequences collected in various indoor environments with a calibrated camera mounted on a mobile robot. The camera is set up to present zero tilt (pitch) and roll angles and known fixed height with respect to the ground floor. The authors in [34] state that their approach strictly relies on these specific setup constraints and on the ground-walls' boundaries detected in the images. This implies that, if the observer does not move parallel to the ground with known height and if those boundary lines are not observed, the approach will not be able to generate initial layout hypotheses. On the other hand, our approach does not require any of these assumptions.

The quantitative results of the tests on this dataset are presented in Table 1, while Figure 4 shows a visual overview of our performance. There are a few sequences for which neither SLAM nor VisualSfM are able to produce any 3D reconstruction due to the very small amount of motion of the observer (insufficient parallax). These sequences were not taken into account for the evaluation. Please note that the method in [34] cannot recover the ceiling part of the scene layout, therefore the authors did not include these pixels in the evaluation of the performances. Since our approach as well as [16] and [14] are able to estimate the ceiling component of the scene layout, and in order to present a more complete comparison, we add in Table 1, beside the original values, the results where ceilings are included in the evaluation. Please note that, when excluding the ceiling, the proposed method is second only to [34] (which was designed to work in specifically such constrained scenarios), while, when taking into account the whole scene, including ceiling, the proposed method outperforms all other approaches, while operating in significantly less constraining conditions.

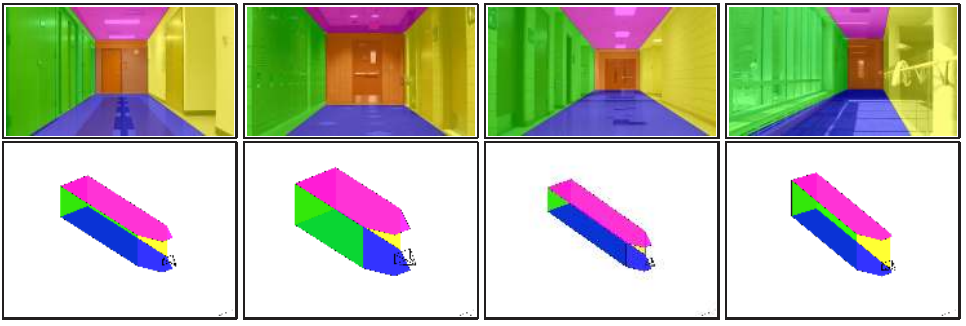


Figure 4: Some examples of our results on the dataset presented in [34]. The top row shows the reprojection of the best layout hypothesis into the image, while the bottom row shows the same layout hypothesis in the 3D space along with the camera pose.

4.3 Proposed dataset

The sequences in the dataset [34] feature substantially simple environments, as can be seen in the 3D reconstructions in Figure 4. With this paper we introduce a new dataset [1] to evaluate the full capabilities of our approach. As opposed to the previous dataset, we let the observer freely move (6DoF) around to observe the scene. We collected 10 sequences in a variety of environments, spanning offices, corridors and large rooms. Most of the sequences frame ground-walls boundaries for short periods or do not frame them at all; some present scenes that cannot be represented by a simple box layout model or relying on the Manhattan world assumption. All the sequences were collected with common smartphones, in the attempt to test the proposed method in real-life scenarios with low-cost sensors.

The classification accuracy results and the mean execution speed (in fps) of the tests on this dataset are presented in Table 2, while Figure 5 shows a visual overview of the dataset and of our performance. In Table 2, please note that: i) the proposed method significantly outperforms state-of-the-art methods in both classification accuracy and execution time; ii) when feeding the proposed approach with the SfM reconstructions, in order to keep the execution time reasonable, both SfM and the optimization procedure were run on a small subset of frames which, despite the ability of SfM to produce denser reconstruction with respect to SLAM, led to worst reconstruction results.

Please refer to the supplementary material [1] for a discussion of failure and success cases, additional images and the complete table of the experimental results.

5 Conclusions

In this paper we presented a real-time oriented approach for indoor scene understanding, addressing the problem of estimating the 3D structural layout of complex and cluttered indoor scenes from monocular video sequences, where the observer can freely move in the surrounding space. The proposed probabilistic framework allows us to generate, evaluate and optimize layout hypotheses by integrating new image evidence as the observer moves. The proposed effective inference engine allows us to make less limiting assumptions than other state-of-the-art methods (e.g., Manhattan world assumption, known and fixed camera height). In the extensive experimental evaluation we demonstrate that our formulation reaches near-real-time computation time and outperforms state-of-the-art methods in both classification accuracy and computation time, while operating in significantly less constraining conditions.

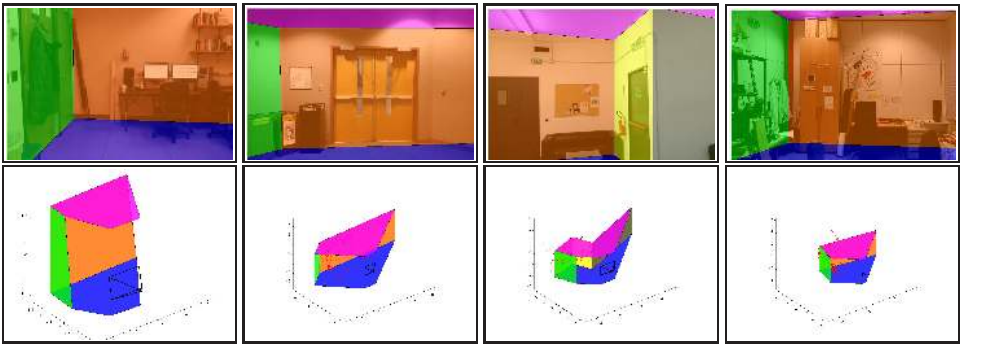


Figure 5: Some examples of our results on the proposed dataset. The top row shows the reprojection of the best layout hypothesis into the image, while the bottom row shows the same layout hypothesis in the 3D space along with the camera pose.

6 Acknowledgments

We acknowledge the support of a NSF CAREER award N.1054127 and a gift award from HTC.
This work has been partially supported by the Italian Ministry of University and Research (MIUR) through the PRIN 2009 grant "ROAMFREE".
Stephen Miller is supported by the Hertz Foundation Google Fellowship and the Stanford Graduate Fellowship.

References

[1] http://vision.stanford.edu/3Dlayout/
http://www.ira.disco.unimib.it/free_your_camera.

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Brian Curless, Steven M. Seitz, and Richard Szeliski. Reconstructing rome. *IEEE Computer*, 43, 2010.

[3] J. Canny. A computational approach to edge detection. *PAMI*, 8(6), 1986.

[4] Nico Cornelis, Bastian Leibe, Kurt Cornelis, and Luc Van Gool. 3d urban scene modeling integrating recognition and reconstruction. *IJCV*, 78, 2008.

[5] A.J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *ICCV*, 2003.

[6] Ethan Eade and Tom Drummond. Monocular slam as a graph of coalesced observations. In *ICCV*, 2007.

[7] Alex Flint, Christopher Mei, David Murray, and Ian Reid. A dynamic programming approach to reconstructing building interiors. In *ECCV*, 2010.

[8] Alex Flint, Christopher Mei, Ian Reid, and David Murray. Growing semantically meaningful models for visual slam. In *CVPR*, 2010.

[9] Alex Flint, David Murray, and Ian Reid. Manhattan scene understanding using monocular, stereo, and 3d features. In *ICCV*, 2011.

[10] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. Manhattan-world stereo. In *CVPR*, 2009.

[11] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. Reconstructing building interiors from images. In *ICCV*, 2009.

[12] Andrew P. Gee, Denis Chekhlov, Andrew Calway, and Walterio Mayol-Cuevas. Discovering higher level structure in visual slam. *IEEE Transactions on Robotics*, 24(5), 2008.

[13] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[14] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, 2009.

[15] Varsha Hedau, Derek Hoiem, and David A. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, 2010.

[16] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Recovering surface layout from an image. *IJCV*, 75(1), 2007.

[17] Georg Klein and David Murray. Improving the agility of keyframe-based slam. In *ECCV*, 2008.

[18] David C. Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. In *CVPR*, 2009.

[19] David Changsoo Lee, Abhinav Gupta, Martial Hebert, and Takeo Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. *NIPS*, 2010.

[20] Daniele Marzorati, Matteo Matteucci, Davide Migliore, and Domenico G. Sorrenti. On the use of inverse scaling in monocular slam. In *ICRA*, 2009.

[21] Pedro Piniés and J. D. Tardós. Large scale slam building conditionally independent local maps: Application to monocular vision. *TRO*, 24(5), 2008.

[22] M. Pollefeys, D. Nistér, J. M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3d reconstruction from video. *IJCV*, 78(2-3), 2008.

[23] Cyril Roussillon, Aurélien Gonzalez, Joan Solà, Jean-Marie Codol, Nicolas Man-sard, Simon Lacroix, and Michel Devy. Rt-slam: A generic and real-time visual slam implementation. *CoRR*, 2012.

[24] Scott Satkin, Jason Lin, and Martial Hebert. Data-driven scene understanding from 3D models. In *BMVC*, 2012.

[25] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient structured prediction for 3d indoor scene understanding. In *CVPR*, 2012.

[26] Alexander G. Schwing and Raquel Urtasun. Efficient exact inference for 3d indoor scene understanding. In *ECCV*, 2012.

[27] Sudipta N. Sinha, Drew Steedly, and Richard Szeliski. Piecewise planar stereo for image-based rendering. In *ICCV*, 2009.

[28] Keith N. Snavely. *Scene reconstruction and visualization from internet photo collections*. PhD thesis, Seattle, WA, USA, 2009.

[29] J. Solá. Consistency of the monocular ekf-slam algorithm for three different landmark parametrizations. In *ICRA*, 2010.

[30] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[31] C. Tomasi and T. Kanade. Detection and tracking of point features. *IJCV*, 1991.

[32] Bill Triggs, Philip Mclauchlan, Richard Hartley, and Andrew Fitzgibbon. Bundle adjustment - a modern synthesis. In *Vision Algorithms: Theory and Practice, LNCS*. Springer Verlag, 2000.

[33] Grace Tsai and Benjamin Kuipers. Dynamic visual understanding of the local environment for an indoor navigating robot. In *IROS*, 2012.

[34] Grace Tsai, Changhai Xu, Jingen Liu, and Benjamin Kuipers. Real-time indoor scene understanding using bayesian filtering with motion cues. In *ICCV*, 2011.

[35] Changchang Wu, Sameer Agarwal, Brian Curless, and Steven M. Seitz. Multicore bundle adjustment. In *CVPR*, 2011.

[36] Jianxiong Xiao and Yasutaka Furukawa. Reconstructing the world's museums. In *ECCV*, 2012.