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

Axel Furlan[1]
furlan@disco.unimib.it

Stephen David Miller[2]
sdmiller@stanford.edu

Domenico G. Sorrenti[1]
sorrenti@disco.unimib.it

Li Fei-Fei[2]
feifeili@stanford.edu

Silvio Savarese[3]
silvio@umich.edu

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

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

[3] EECS Building
University of Michigan
Ann Arbor, MI, USA

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 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.

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, we first pre-process sequences with a localization and sparse 3D reconstruction algorithm. In our experiments we compare two such approaches: a real-time implementation of the Monocular V-SLAM approach proposed in [4] and the non-real-time VisualSfM [6]. These 3D reconstructions are in general noisy and sparse.

Candidate layout components. The second step consists of generating a higher level representation of the 3D points estimated in the preprocessing 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. 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 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 [5] (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.

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

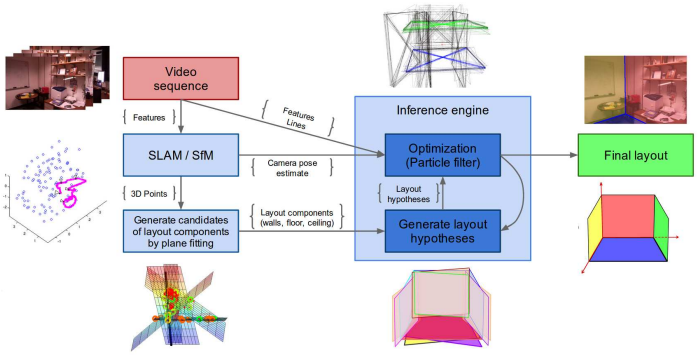

Figure 1: **3D scene layout estimation process.** The video sequence is first processed to obtain camera localization and sparse 3D point cloud reconstruction. Layout components (e.g. floor, ceiling, walls) are generated from the sparse 3D points and combined to generate layout hypotheses. Each layout 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.

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

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

We have designed this probability to enforce a number of desirable properties (*e.g.* low reprojection error, description simplicity, physical plausibility, etc.). Please refer to the main paper for a complete description of this scoring function and its components.

Experimental validation. We show experimental results of our method when tested on the state-of-the-art dataset [5], as well as on a new challenging dataset [1] that we introduce in this paper. Final reconstruction results are compared to three state-of-the-art approaches (the video-based approach [5] and the two well known single image methods [2, 3]) and a baseline method (to explicitate the importance of the evaluation and optimization process).

Conclusions. In this paper we present a real-time oriented approach for indoor scene understanding in cluttered environments. The proposed probabilistic framework allows to generate, evaluate and optimize layout hypotheses by integrating new image evidence as the observer moves. In the extensive experimental evaluation we demonstrate that our formulation reaches near-real-time computation time and, while operating in significantly less constraining conditions (*e.g.* soft Manhattan assumption, complex scene geometry, freely moving observer), outperforms state-of-the-art methods in both classification accuracy and computation time.

[1] URL `http://www.ira.disco.unimib.it/free_your_camera`.

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

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

[4] 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.

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

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