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

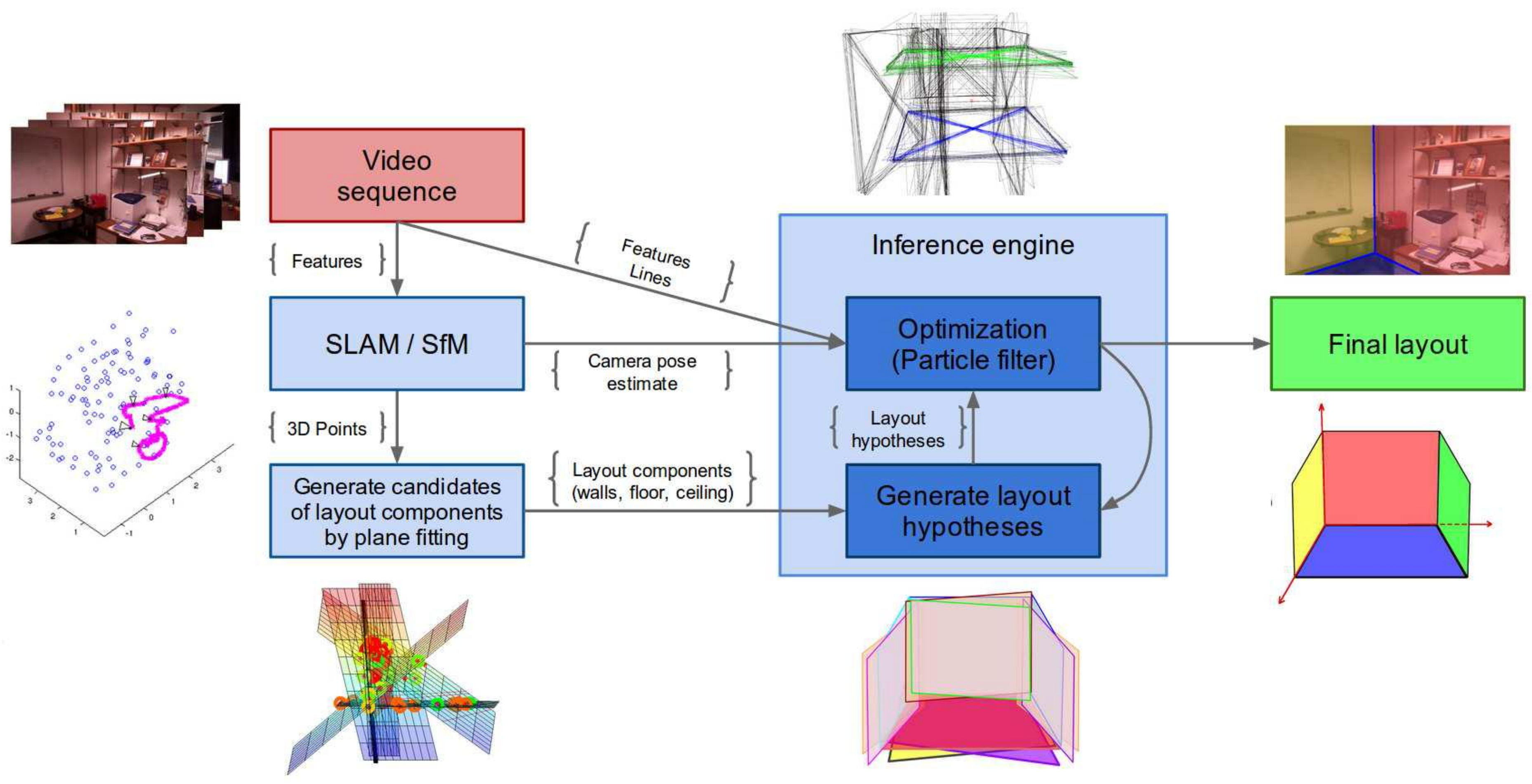
A. Furlan[1], S. Miller[2], D. G. Sorrenti[1], L. Fei-Fei[2], S. Savarese[2]

[1]Computer Science Department – University of Milano - Bicocca
[2]Computer Science Department – Stanford University

{furlan, sorrenti}@disco.unimib.it *{sdmiller, feifeili, ssilvio}@stanford.edu*

Overview

- Problem statement
  - 3D Indoor semantic layout estimation
  - Full 6DoF freely moving observer
  - No hard Manhattan assumptions
  - Near real-time performances
- Experiments
  - Tested on the Michigan Indoor Corridor dataset [1]
  - Introduction of a new challenging dataset

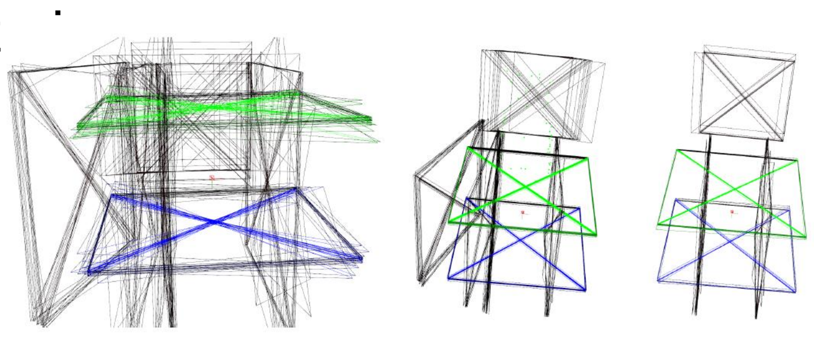


Proposed approach

- Sparse 3D reconstruction
  - Estimate camera pose and a sparse map with:
    - Fast Monocular V-SLAM – All frames in real-time
    - Slow VisualSfM – Few frames to preserve real-time
- Layout definition
  - Made of layout components (walls, ground, floor)
  - Walls are orthogonal to the ground plane
  - Arbitrary number of walls, not mutually orthogonal
- Layout estimation
  - Iterative RanSaC plane fitting
    - Large number of inaccurate layout components
  - Initialize layout hypotheses as random combinations of layout components
  - Local perturbation and optimization of hypotheses
  - Each hypothesis is a particle in a particle filter
- Scoring hypotheses

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

  - Terms in the score function enforce *fitness* (P_f), *orthogonality to ground* (P_o), *reprojection error* (P_r), *wall-to-wall orientation* (P_m), *simplicity* (P_s), *wall-to-wall intersection* (P_w).
- Advantages:
  - No hard Manhattan assumptions
  - No *a priori* knowledge of the observer motions w.r.t. the scene
  - Near-real-time performances (~20fps)
  - Particle filter implementation allows recovering from noisy and wrong initialization exploiting multimodal posterior, re-sampling and particle clust



Experiments

- Michigan Indoor Corridor dataset [1]
  - Indoor video sequences from a mobile robot
  - Object-free corridor scenes
- Proposed dataset
  - Indoor video sequences from hand-held smartphone
  - Various cluttered scenes
    - Offices, corridors, large rooms
    - Complex layouts (not box-room, not Manhattan)
- Results
  - Our method significantly outperforms [1], [2] and [3] in both classification accuracy and execution time
  - Table below:
    - Left – Results on the Michigan Indoor Corridor dataset [1] (excluding and including ceiling)
    - Right – Results on the proposed dataset (classification accuracy and computation time)

Method	Excl. ceil	Incl. ceil
[1]	**90.58**	82.17
[2]	82.62	83.30
[2]+MRF	81.44	82.13
[3]	84.70	84.33
Our + VSLAM	86.92	**87.01**

Method	Clas. acc.	Avg. fps
Baseline	70.64	—
[2]	59.29	0.17
[3]	73.59	0.03
Our + VSLAM	**86.24**	**21.63**
Our + VSfM	75.94	16.90

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

Conclusions

- Real-time oriented approach for indoor scene understanding
- Probabilistic framework to generate, evaluate and optimize layout hypotheses
- Extensive experimental evaluation, that demonstrates that our formulation outperforms state-of-the-art methods in both classification accuracy and computation time
- **Dataset available:** http://www.ira.disco.unimib.it/free_your_camera
  http://vision.stanford.edu/3Dlayout/

This work was partially funded by a NSF CAREER award N.1054127 and a gift award from HTC, the PRIN 2009 grant "ROAMFREE" and the Hertz Foundation Google Fellowship.