Semantic Structure From Motion

Sid Yingze Bao and Silvio Savarese

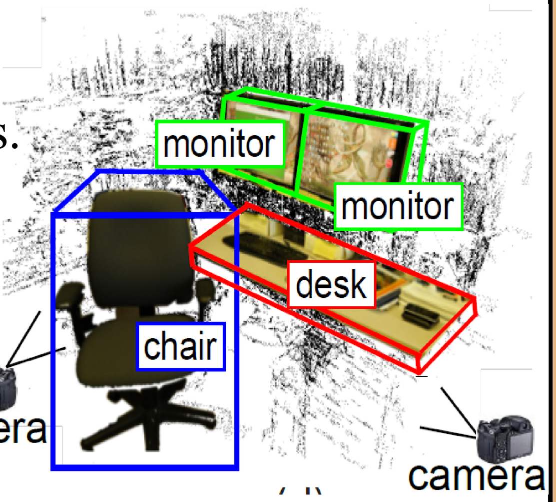
Electrical and Computer Engineering, University of Michigan at Ann Arbor

Source code and data: http://www.eecs.umich.edu/vision/projects/ssfm/index.html

Introduction

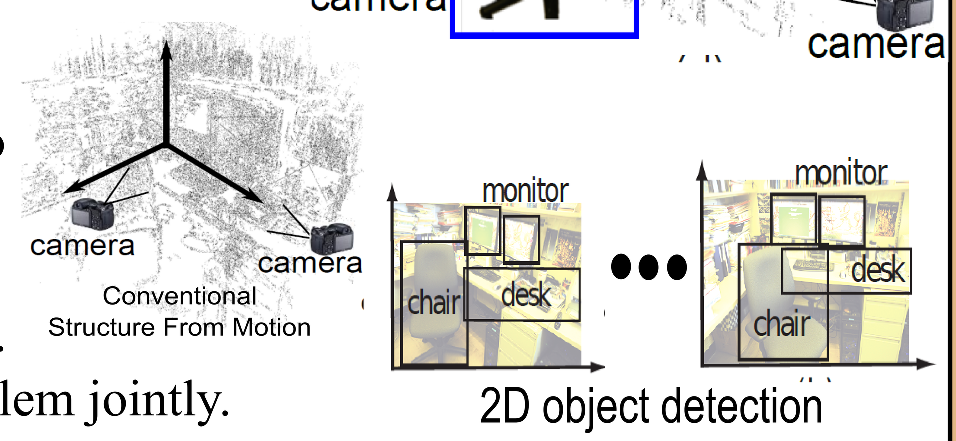
Goal:

Estimate 3D location and pose of objects, 3D location of points, and camera parameters from 2 or more images.



Motivation:

- Most 3D reconstruction methods do not povide semantic information.
- Most recognition methds do not provide geometry and camera pose.
- We propose to solve these two problem jointly.



Advantages:

- Improve camera pose estimation, compared to feature-point-based SFM.
- Improve object detections given multiple images, compared to independently detecting objects from each single images.
- Establish object correspondences across views.

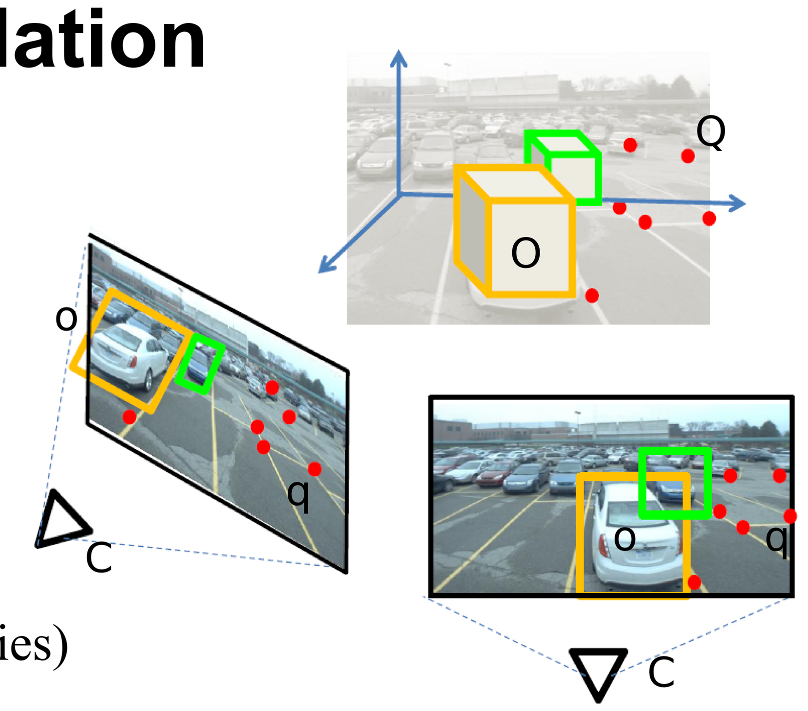
SSFM Problem Formulation

Measurements

- **q**: point features (e.g. DOG+SIFT)
- **u**: point matches (e.g. threshold test)
- **o**: 2D objects (e.g. [2])

Model Parameters (unknowns)

- **C**: camera (K is known)
- **Q**: 3D points (locations)
- **O**: 3D objects (locations, poses, categories)



Intuition:

In addition to point features, measurements of objects across views provide additional geometrical constraints that allow to relate cameras and scene parameters.

Reference

[1] N. Snavely, S. M. Seitz, and R. S. Szeliski. Modeling the world from internet photo collections. IJCV. 2008.

[2] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009.

[3] Gaurav Pandey, James McBride,,and Ryan Eustice,.Ford campus vision and lidar data set. International Journal of Robotics Research. 2011

Model Overview

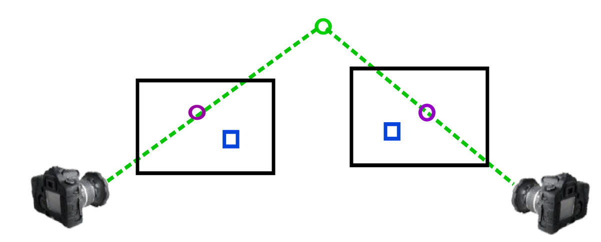
{O,Q,C}=arg max P(q,u,o | C,O,Q)

=arg max P(q,u | C,Q)P(o | C,O)

Assumption:
Given camera hypothesis, objects and points are independent

Point Likelihood P(q,u | C,Q)

$$P(\mathbf{q},\mathbf{u}|\mathbf{Q},\mathbf{C}) \propto \prod_i^{N_Q}\prod_k^{N_k} \exp(-(q_i^k - q_{u_i^k}^k)^2/\sigma_q)$$

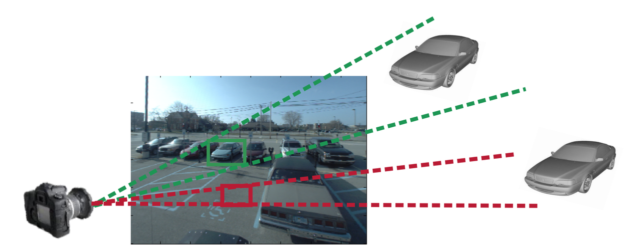


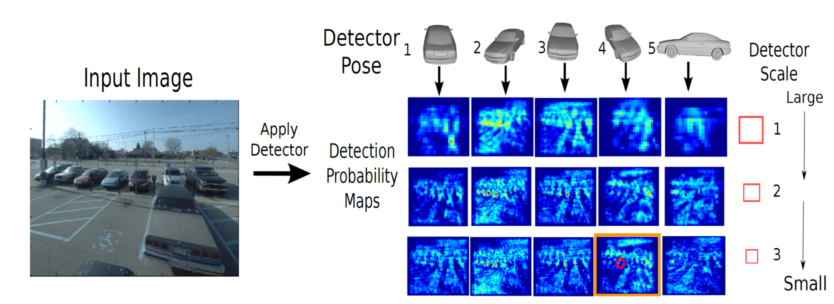
Object Likelihood P(o|C,O)

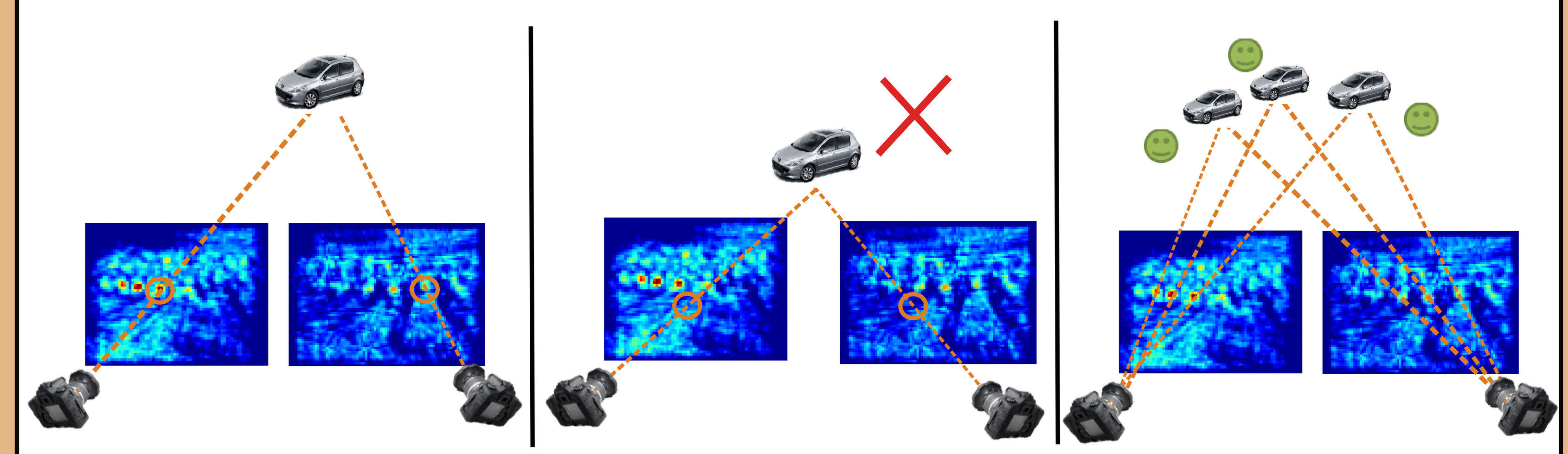
- Estimate 3D object likelihood by 2D projection appearance:

$$P(\mathbf{o}|\mathbf{O},\mathbf{C}) \propto \prod_t^{N_t} P(\mathbf{o}|O_t,\mathbf{C})$$

$$\propto \prod_t^{N_t}(1 - \prod_k^{N_k}(1 - P(o|O_t,C^k)))$$







Joint Likelihood Maximization

Main challenge:

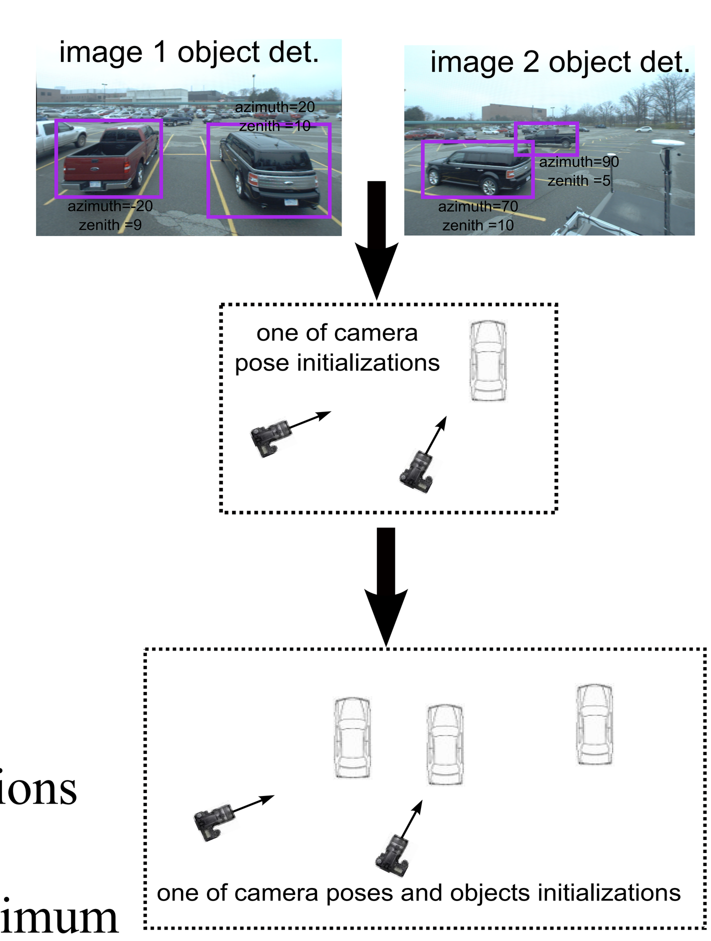
High dimensionality of unknowns =>
Sample P(q,u,o|C,O,Q) with MCMC

Parameter Initialization

- Use object detection scale and pose to initialize cameras relative poses
- Theorem: camera parameters can be estimated given:
  i) 3 objects with scale; ii) 2 objects with pose; iii) 1 object with scale and pose.

Monte Carlo Markov Chain

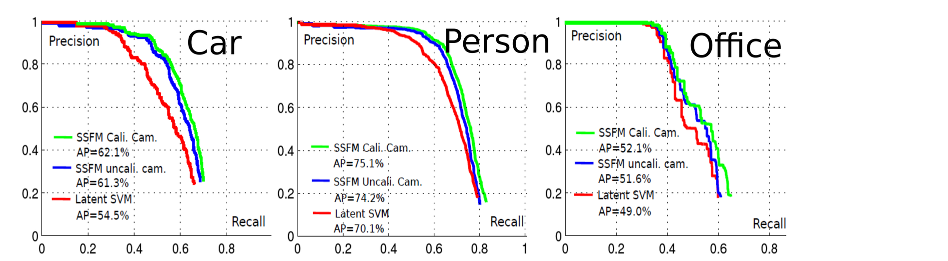
- Sampling starts from different initializations
- Proposal distribution P(q,u,o|C,O,Q)
- Combine all samples to identify the maximum



Results

Comparison Baselines

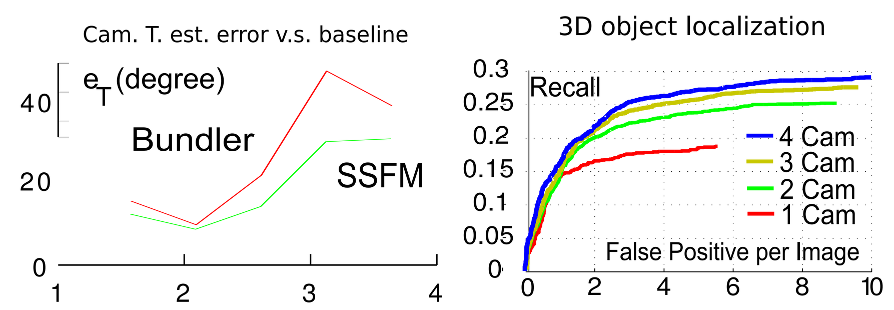
- Camera Pose Est.: Bundler [1]
- Object Detection: LSVM [2]



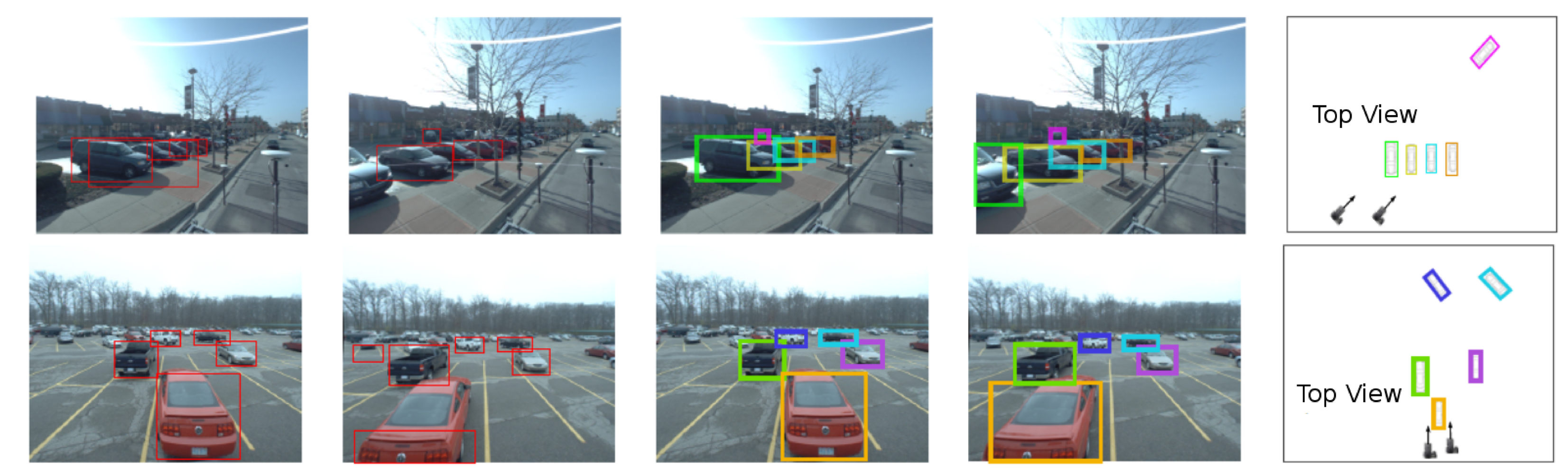
1. Car Dataset [3] (available online)

- Images and Dense Lidar Points
- ~500 testing images in 10 scenarios

Dataset	$\bar{e}_T$ Bundler/SSFM	$\bar{e}_R$ Bundler/SSFM
Ford Campus Car	26.5/**19.9**°	**0.47**°/0.78°
Street Pedestrian	27.1°/**17.6**°	21.1°/**3.1**°
Office Desktop	8.5°/**4.7**°	9.6°/**4.2**°

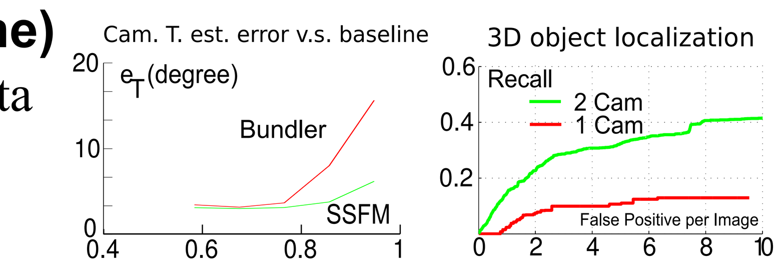


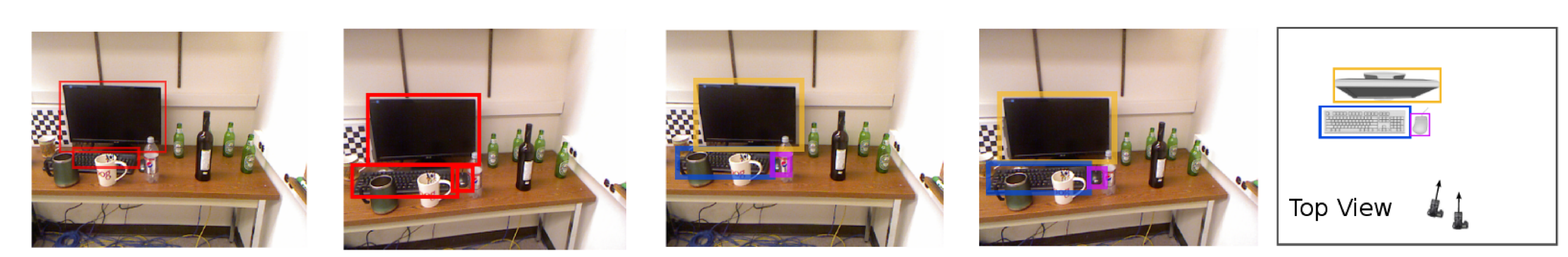
Camera #	2	3	4
Det. AP (Cali. Cam.)	62.1%	63.6%	64.2%
Det. AP (Uncali. Cam.)	61.3%	61.7%	62.6%
$\bar{e}_T$	19.9°	16.2°	13.9°



2. Kinect Office Dataset (available online)

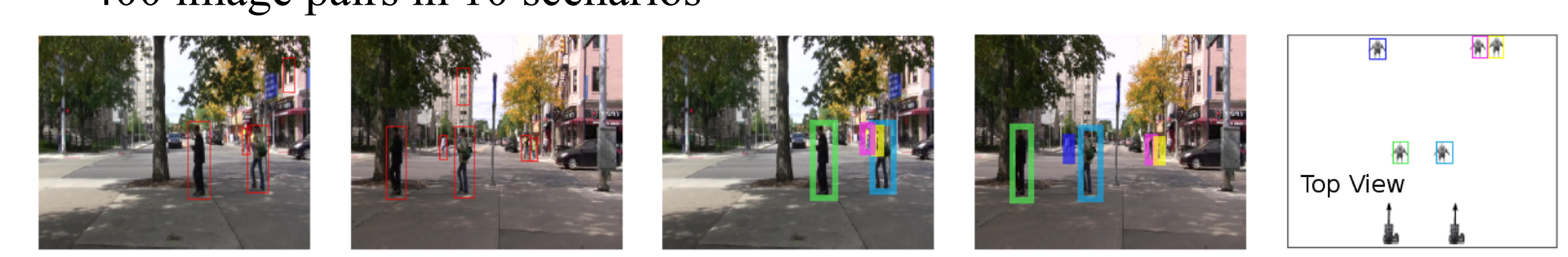
- Images and calibrated Kinect 3D range data
- Mouse, Monitor, and Keyboard
- 500 images in 10 scenarios





3. Person Dataset

- A pair of stereo cameras
- 400 image pairs in 10 scenarios



Ackowledgement

We acknowledge the support of NSF CAREER #1054127 and the Gigascale Systems Research Center. We thank Mohit Bagra for collecting the Kinect dataset and Min Sun for helpful feedback.