Semantic Structure From Motion with Points, Regions, and Objects

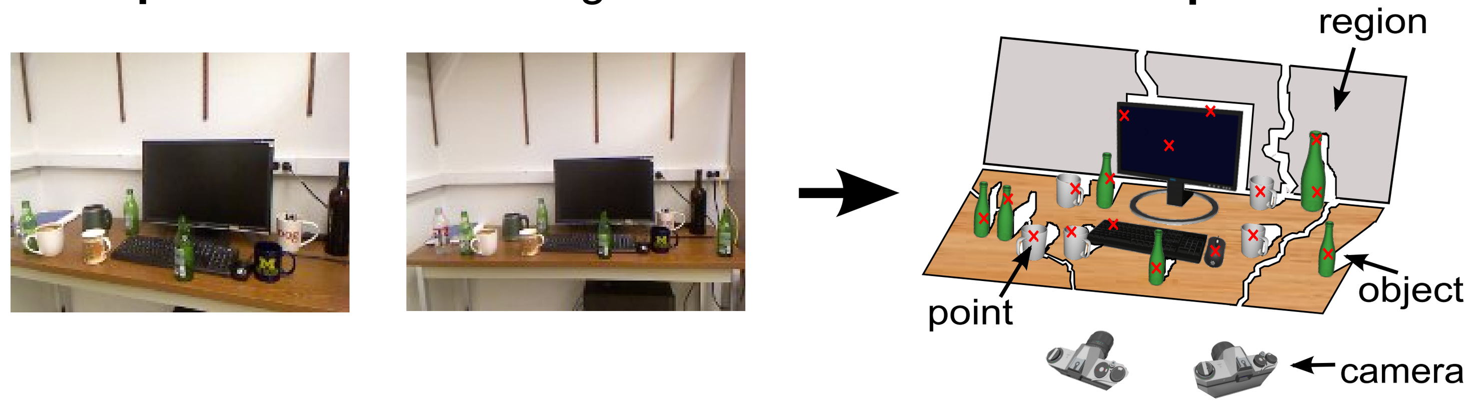
Sid Yingze Bao, Mohit Bagra, Yu-Wei Chao, Silvio Savarese
Electrical and Computer Engineering, University of Michigan at Ann Arbor

Project webpage: http://www.eecs.umich.edu/vision/projects/ssfm/index.html (Code will be available soon!)

Introduction

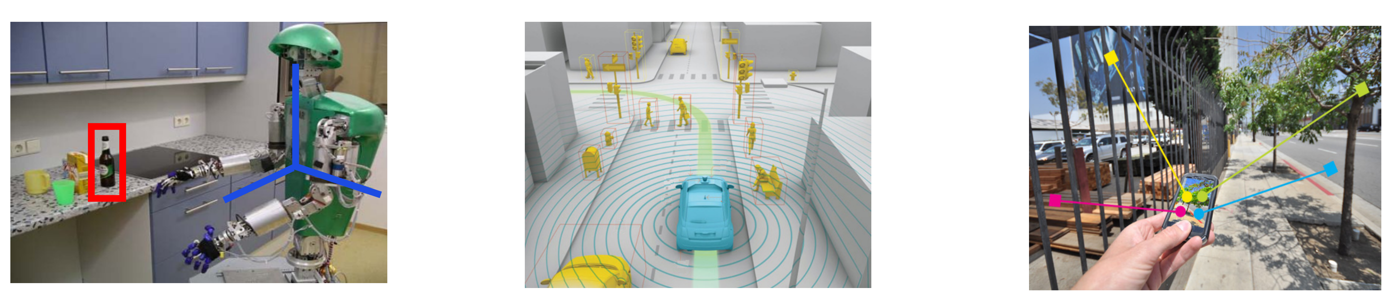
Semantic Structure from Motion (SSFM) is a new framework for jointly estimating semantic and geometrical information from multiple images:
- Detect object; segment and classify regions **(semantic)**
- Recover 3D geometry of objects, regions, and points **(structure)**
- Recover cameras location and pose **(motion)**

Input: two or more images **Output**



Motivation

Ability to jointly recover semantic and geometry information is critical in many applications.
- Most 3D reconstruction methods do not provide semantic.
- Most recognition methods do not localize objects in 3D physical space.

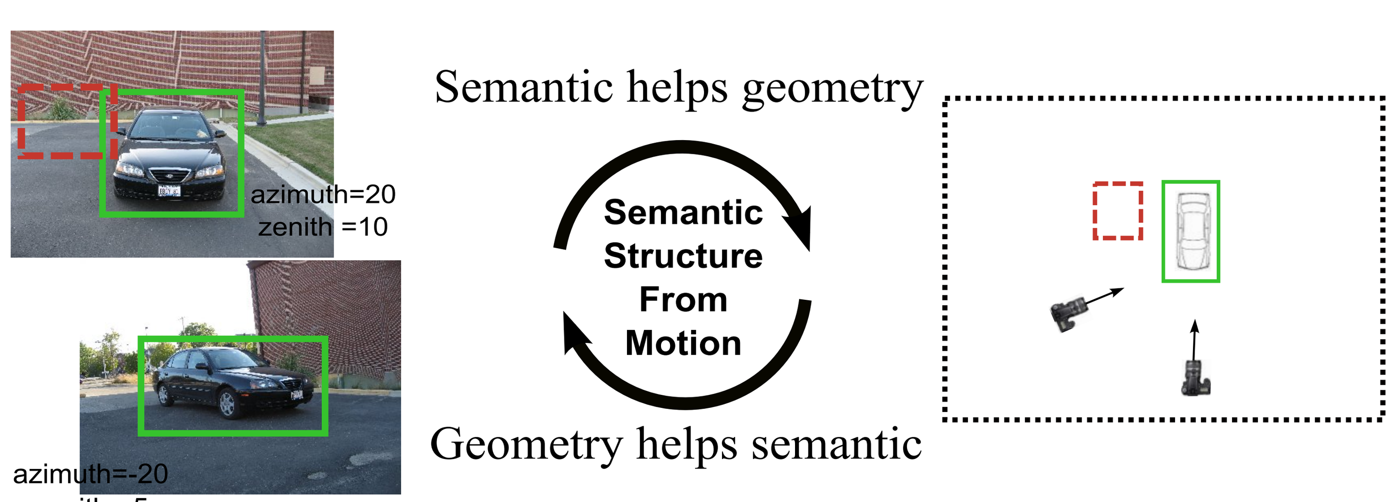


Robot manipulation | Autonomous driving | Augmented Reality

Main intuitions

Semantics and 3D geometry are mutually beneficial.
- Objects and regions help localize the observer.
- Geometric context helps object detection and region classification.
- Semantic reasoning guides the process of matching points and regions.

Interactions among objects, regions and points help regularize solution.



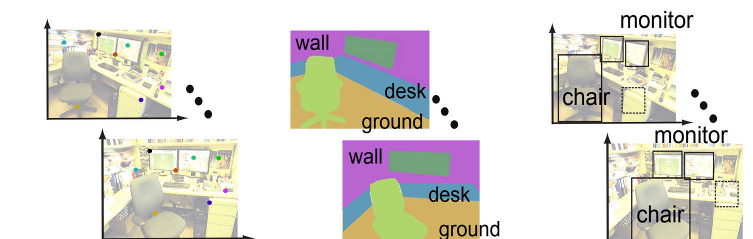
Reference

[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 of Pattern Analysis, 2009.
[3] Gaurav Pandey, James McBride, and Ryan Eustice. Ford campus vision and lidar data set. IJRR 2011
[4] Y. Bao, and S. Savarese, Semantic structure from motion, CVPR 2011
[5] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. ECCV 2010
[6] Y. Bao, M. Bagra, S. Savarese, Semantic structure from motion with object and point interactions, IEEE Workshop on Challenges and Opportunities in Robot Perception (in conjunction with ICCV-11). **Best Student Paper Award**

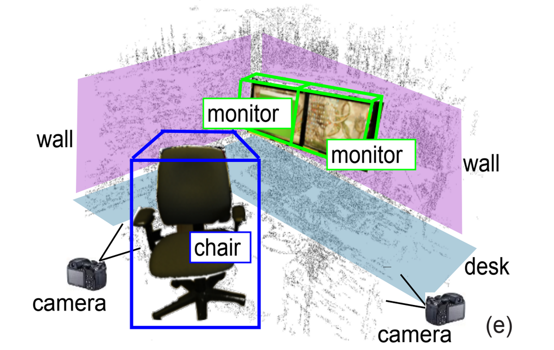
Notations

Inputs
- Two or more images **I**
- known internal parameters

Measurements (noisy)
- **q**: point features (e.g. DOG+SIFT)
- **u**: point matches (e.g. threshold test)
- **b**: 2D regions (e.g. superpixel)
- **o**: 2D objects (e.g. detected by [2])



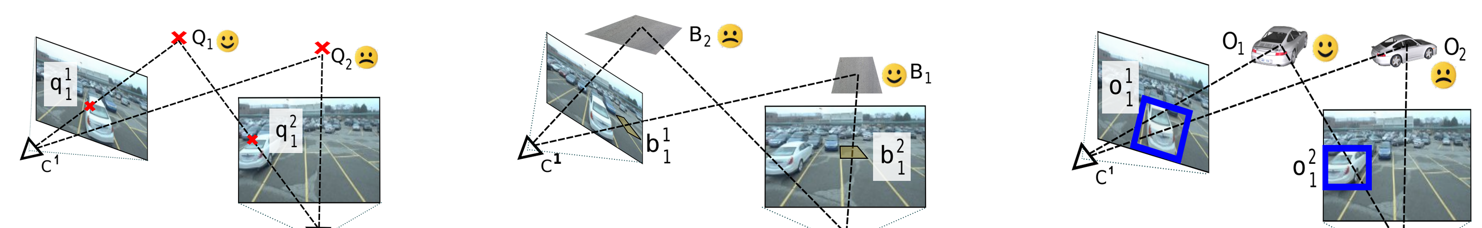
Model Parameters (unknowns)
- **C**: cameras (locations and poses)
- **Q**: 3D points (locations)
- **B**: 3D regions (locations, orientations, classes)
- **O**: 3D objects (locations, poses, categories)



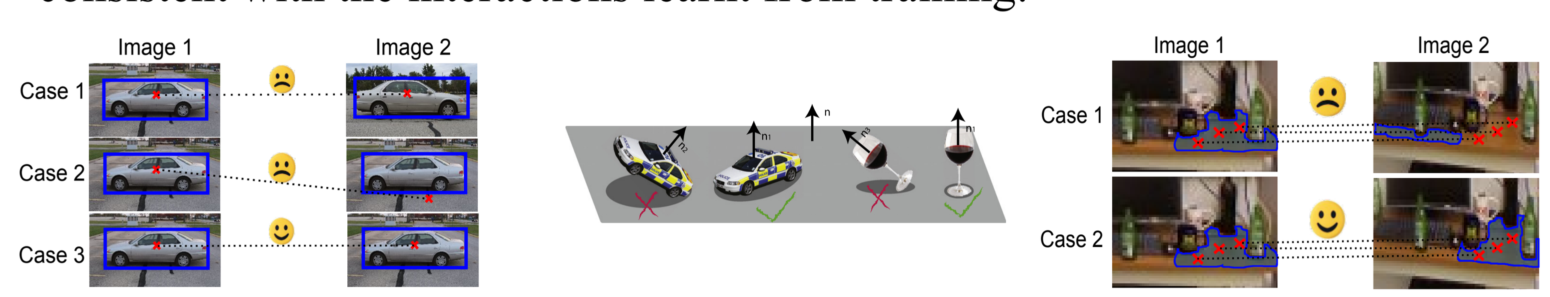
Model

Relationships among points, regions, objects, and cameras follow:

Intuition 1: The image projection of estimated objects, regions, and points are consistent with measurements (location, scale, and pose).



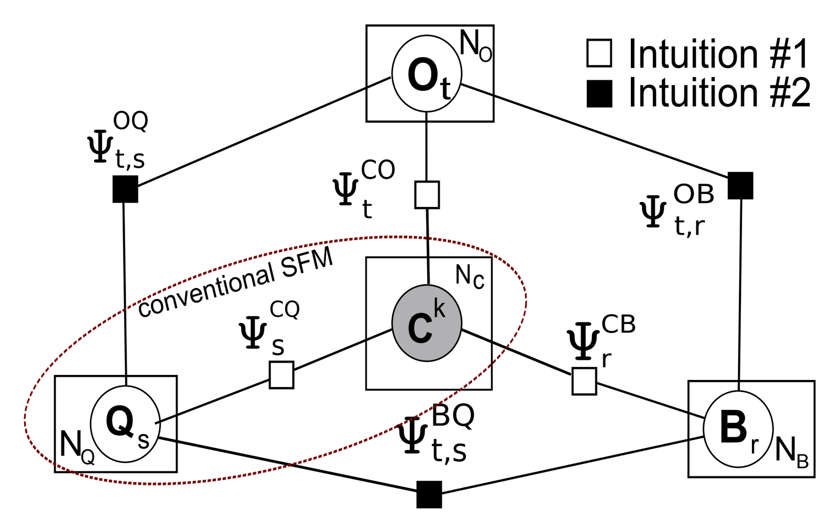
Intuition 2: The interactions among points, regions, and objects should be consistent with the interactions learnt from training.



Energy Formulation

Joint energy of objects, regions, points, cameras given images.

$$\Psi(\mathbf{O},\mathbf{B},\mathbf{Q},\mathbf{C};\mathbf{I}) = \prod_t \Psi_t^{CO} \prod_s \Psi_s^{CQ} \prod_r \Psi_r^{CB} \prod_{t,s} \Psi_{t,s}^{OQ} \prod_{t,r} \Psi_{t,r}^{OB} \prod_{r,s} \Psi_{r,s}^{BQ}$$



Ackowledgement

We acknowledge the support of NSF CAREER #1054127 and the Gigascale Systems Research Center.

Inference

Solving SSFM Problem:

$\{\mathbf{O},\mathbf{B},\mathbf{Q},\mathbf{C}\} = \mathrm{argmax}\ \Psi(\mathbf{O},\mathbf{B},\mathbf{Q},\mathbf{C};\mathbf{I})$

Sampling (Simulated Annealing)
- High dimensionality of unknowns

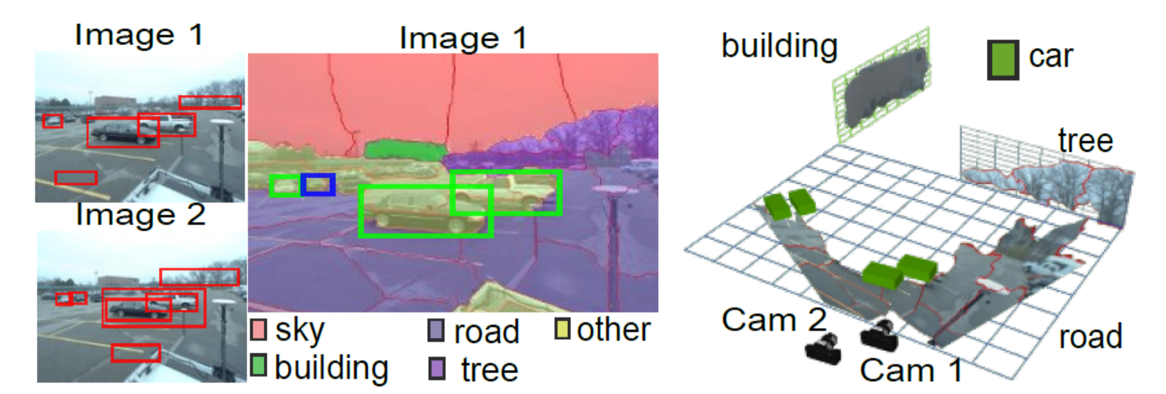
Propose initial guesses of cameras:
- Cameras estimated by point matches (SFM)
- Cameras estimated by matched object detections
- Cameras estimated by matched regions

Sampling Algorithm

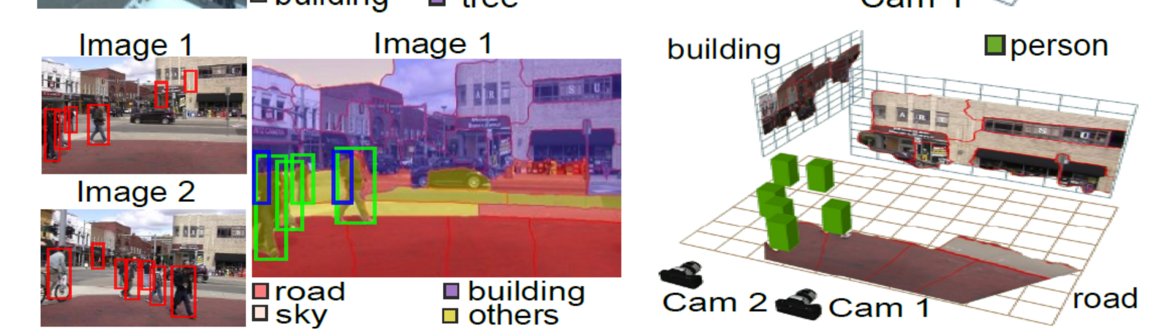
Propose initial guesses of cameras
FOR $\mathbf{C} \in$ initial guesses of cameras
  $\mathbf{C}_1=\mathbf{C}$
  FOR n = 1 : M (M is user-specified)
    $\mathbf{C}_{n+1} = \mathbf{C}_n + \mathbf{C}'$ ($\mathbf{C}'$ is 0-mean Gaussian r.v. whose variance decreases as n increases)
    $\mathbf{O}'_n = \mathrm{argmax} \prod \Psi_t^{CO}$
    $\mathbf{Q}'_n = \mathrm{argmax} \prod \Psi_s^{CQ}$ *(details in paper)*
    $\mathbf{B}'_n = \mathrm{argmax} \prod \Psi_r^{CB}$
    $\{\mathbf{O}_n, \mathbf{Q}_n, \mathbf{B}_n\} = \mathrm{argmax} \prod \Psi_{t,s}^{OQ} \prod \Psi_{t,r}^{OB} \prod \Psi_{r,s}^{BQ}$
    $\alpha = \Psi(\mathbf{O}_n; \mathbf{Q}_n; \mathbf{B}_n; \mathbf{C}_n; \mathbf{I}) / \Psi(\mathbf{O}_{n-1}; \mathbf{Q}_{n-1}; \mathbf{B}_{n-1}; \mathbf{C}_{n-1}; \mathbf{I})$
    IF $\alpha <$ uniform(0, 1)
      $\{\mathbf{O}_n, \mathbf{Q}_n, \mathbf{B}_n, \mathbf{C}_n\} = \{\mathbf{O}_{n-1}, \mathbf{Q}_{n-1}, \mathbf{B}_{n-1}, \mathbf{C}_{n-1}\}$
    END
  END
END
Identify the sample maximizing $\Psi(\mathbf{O}; \mathbf{Q}; \mathbf{B}; \mathbf{C}; \mathbf{I})$

Results

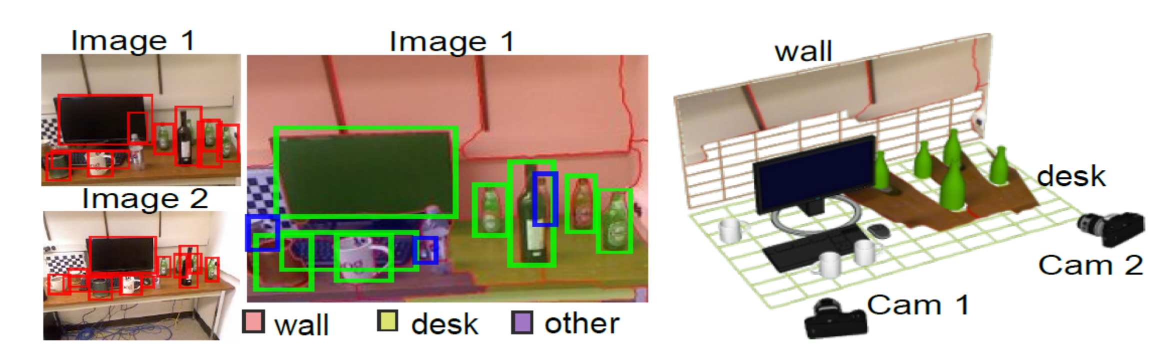
Datasets



Car Dataset [3]
- available online
- Lidar Points
- ~500 testing images
- 10 scenarios



Person Dataset
- Stereo cameras
- 400 image pairs
- 10 scenarios



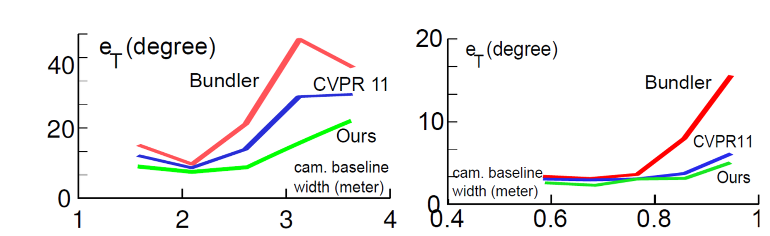
Kinect Office
- available online
- Kinect 3D range data
- Mouse, Monitor, and Keyboard
- 500 images
- 10 scenarios

Camera Pose Esimation

Camera Translation and Rotation Error

e_T / e_R	Car	Person	Office
[1]	26.5°/< 1°	27.1°/21.1°	8.5°/9.5°
[4]	19.9°/< 1°	17.6°/3.1°	4.7°/3.7°
[4] + Regions	18.0°/< 1°	15.7°/3.3°	4.9°/4.1°
This paper	**12.1°/< 1°**	**11.4°/3.0°**	**4.2°/3.5°**

Camera Translation Error v.s. Camera Baseline



(a) Car dataset (b) Kinect dataset

Object Detection Average Precision

Object Detection in 2D

	[2]	[4]	This Paper
Car	54.5%	61.3%	**62.8%**
Person	70.1%	75.1%	**76.8%**
Office	42.9%	45.0%	**45.7%**

Object Detection in 3D

	by single image.	Without interactions	Our full model.
Car	21.4%	32.7%	**43.1%**
Office	15.5%	20.2%	**21.6%**

Region Classification and 3D Geometry Estimation

Region Classification Accuracy

%	Car	Person	Office
[5]	88.9	82.9	50.8
Ours	**90.2**	**84.4**	**51.2**

Region 3D Localization Raletive Error

with / without interaction	$median(e_d)$	$var(e_d)$
Car	0.281 / **0.175**	0.54 / 0.44
Office	0.033 / **-0.011**	0.182 / 0.189