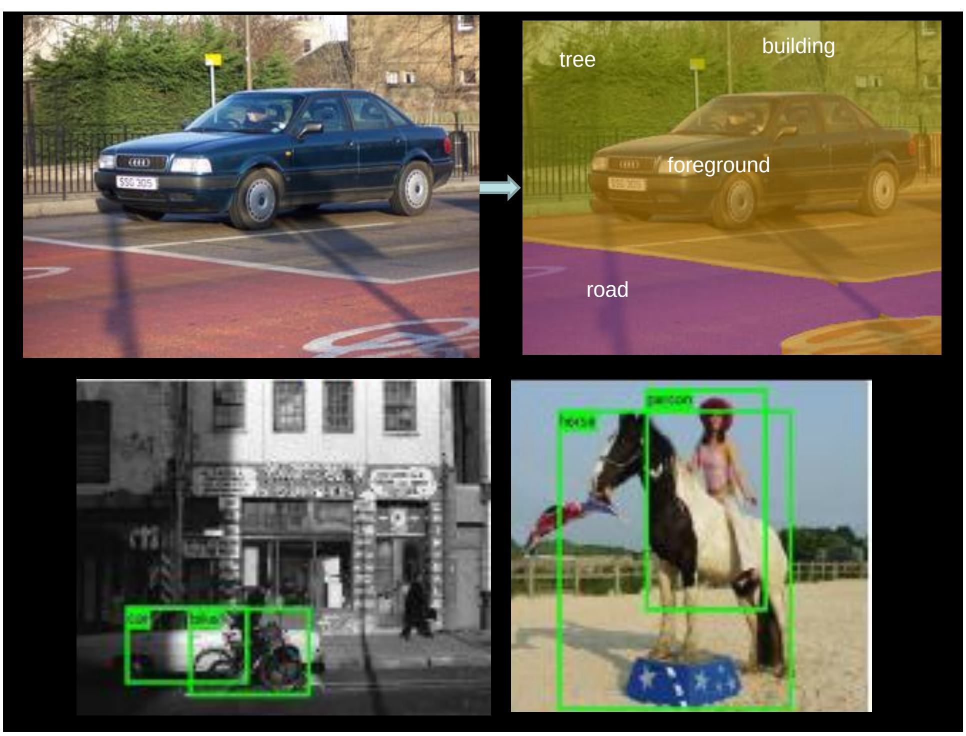
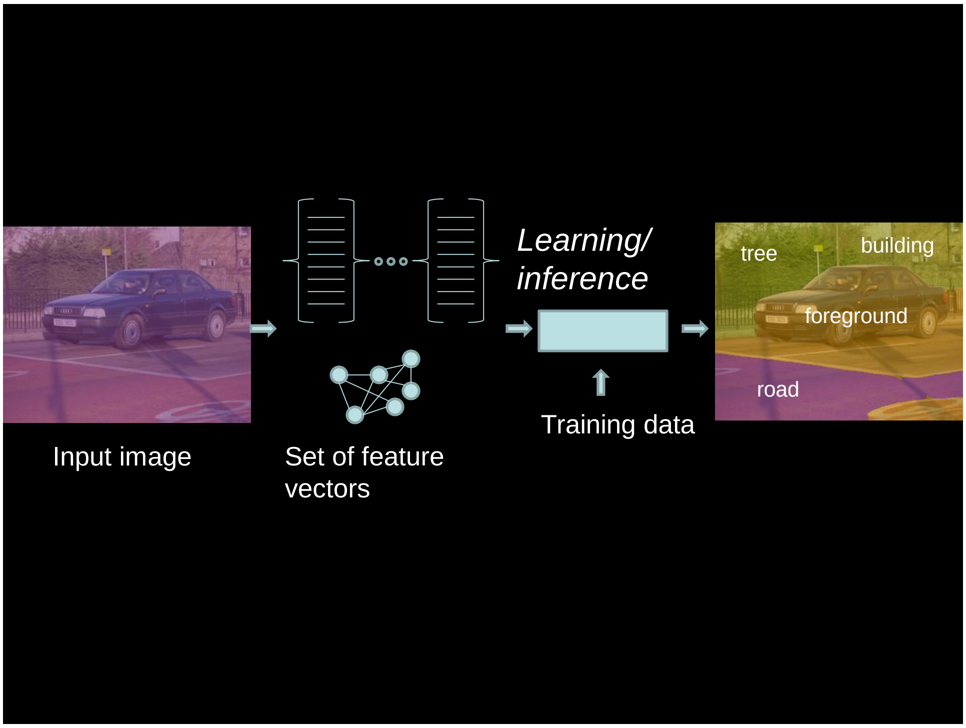
Using 3D cues

Martial Hebert

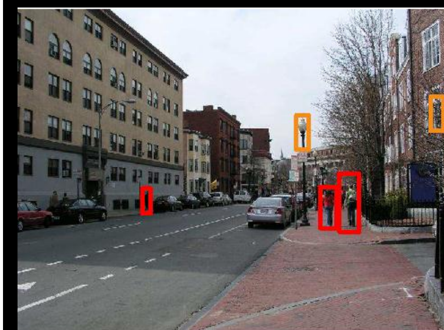
Carnegie Mellon University

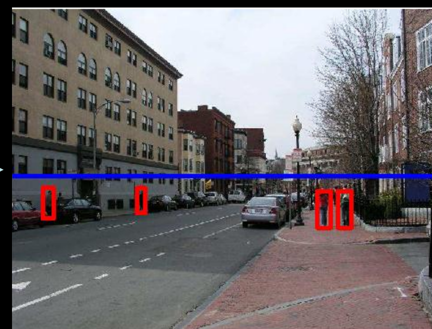



Learning/
inference
Training data
tree
building
foreground
road
Input image
Set of feature
vectors


Close
Not
Close
[Example from D. Hoiem]

Geometric context

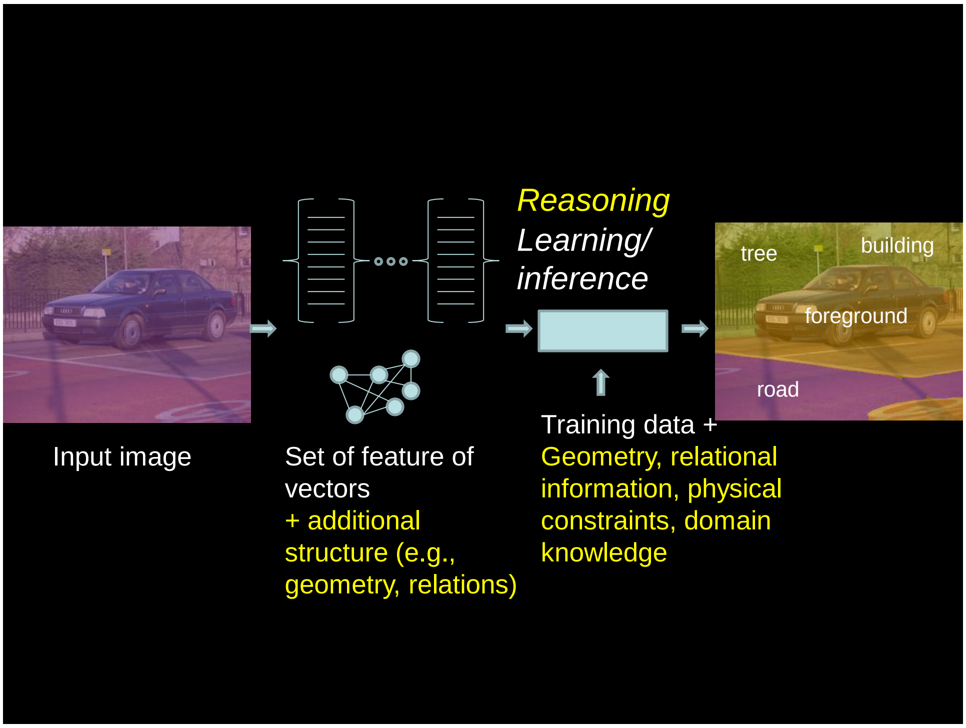




D. Hoiem, A. A. Efros, and M. Hebert. *Putting Objects in Perspective*. International Journal of Computer Vision, Vol. 80, No. 1, October, 2008.

Geometric context

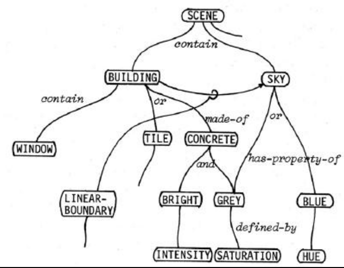
S.Y. Bao, M. Sun, S.Savarese. *Toward Coherent Object Detection And Scene Layout Understand*ing. CVPR 2010.

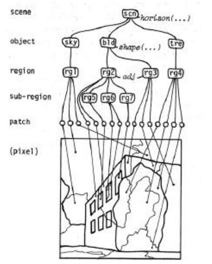


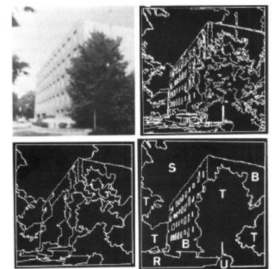
Questions

- How to estimate geometric properties from an image?
- How to incorporate geometric constraints?

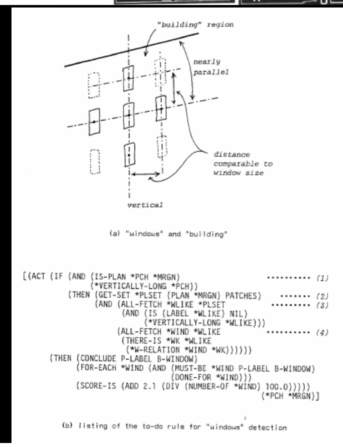
- How to combine reasoning tools with statistical classification/regression tools?







[Ohta & Kanade 1978]



- Guzman (*SEE*), 1968
- Yakimovsky & Feldman, 1973
- Hansen & Riseman (*VISIONS*), 1978
- Barrow & Tenenbaum 1978
- Brooks (*ACRONYM*), 1979
- Ohta & Kanade, 1978

Chronology

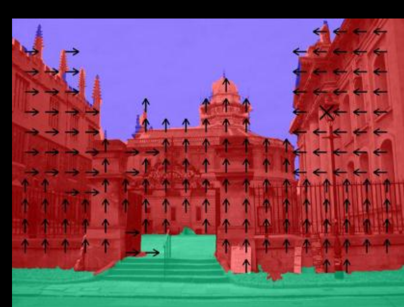
Bottom up classifiers — More explicit constraint+reasoning

Qualitative — Explicit/Quantitative

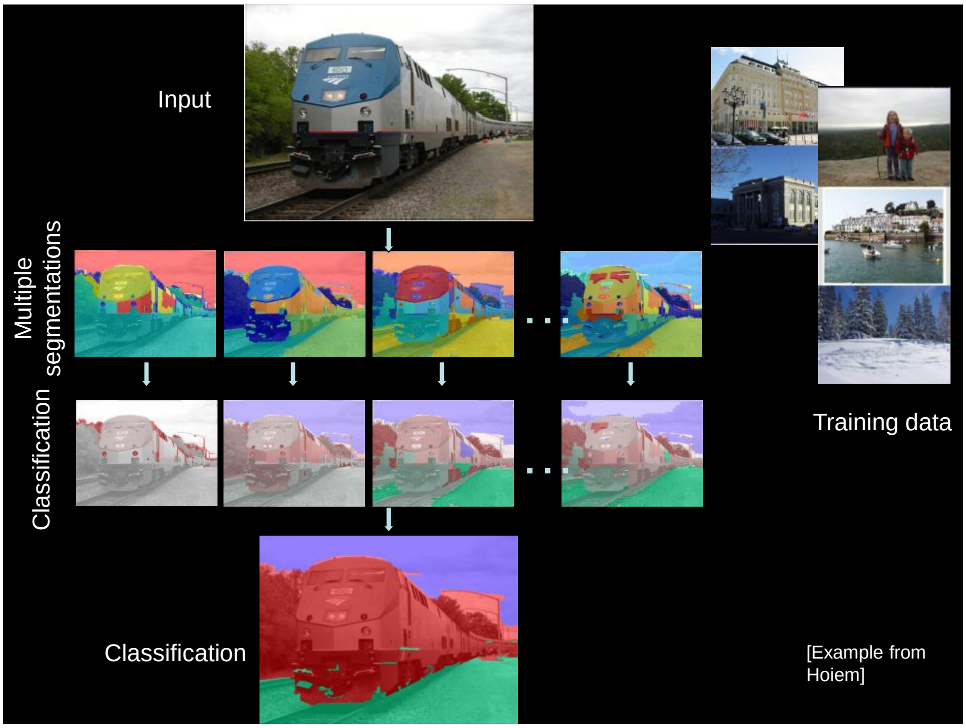
Region labels

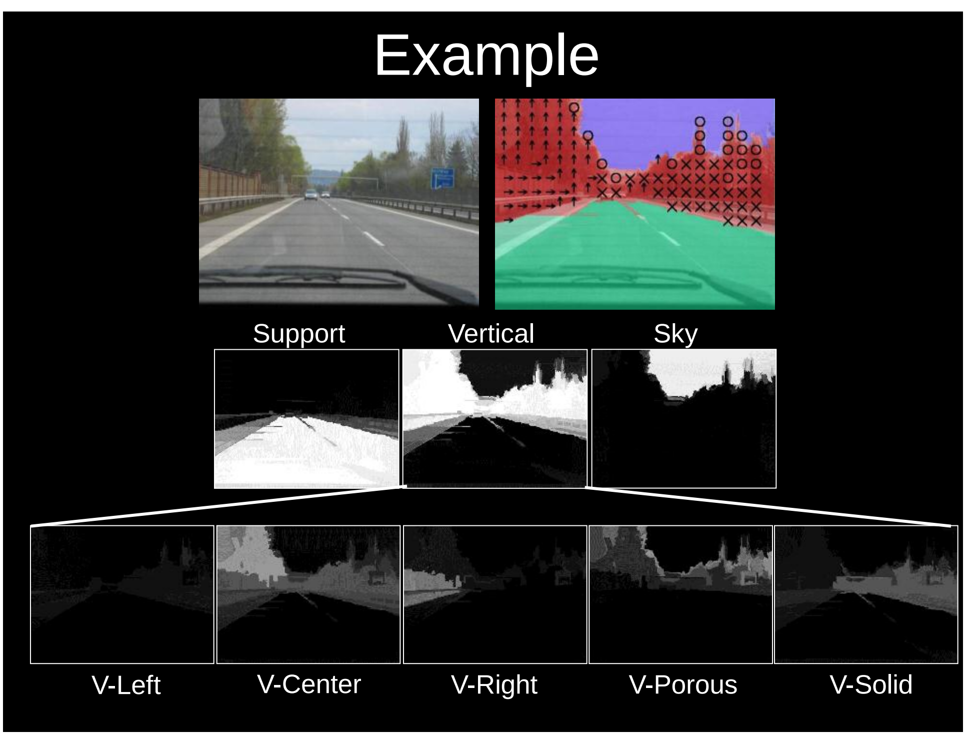
Qualitative

First attempt: Estimate surface labels

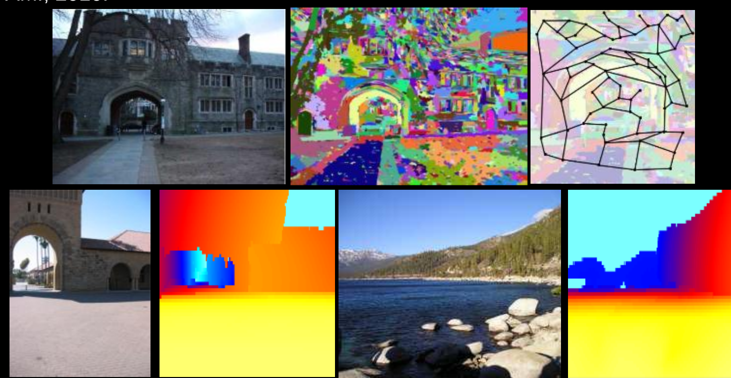


[D. Hoiem, A. A. Efros, and M. Hebert. *Recovering surface layout from an image*. IJCV, 75(1):151–172, 2007]


Input
Multiple segmentations
Classification
Classification
Training data
[Example from Hoiem]

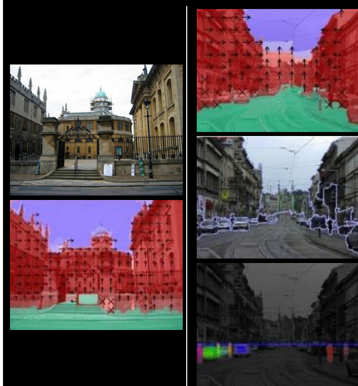

Example
Support
Vertical
Sky
V-Left
V-Center
V-Right
V-Porous
V-Solid

- *Learning from image features to depth + MRF:* A. Saxena, S. H. Chung, and A. Y. Ng. *3-D depth reconstruction from a single still image*. IJCV, 76, 2007.
- Make3D: Learning 3D Scene Structure from a Single Still Image: A. Saxena, M. Sun and A. Y. Ng. TPAMI, 2010.



MRF on superpixels → Depth map
- Unary potentials:
  - Depth prediction from local features+ co-occurrence of superpixels over multiple segmentations
- Binary potentials:
  - Colinearity along edges
  - Connectivity along neighboring superpixels
  - Co-planarity along neighboring superpixels

- Is a more precise representation possible?
- For example:
  - We would like to include reasoning about interposition (relations between object relative to a viewpoint induced by occlusion boundaries)
  - We would like to include constraints about object semantics (when known)

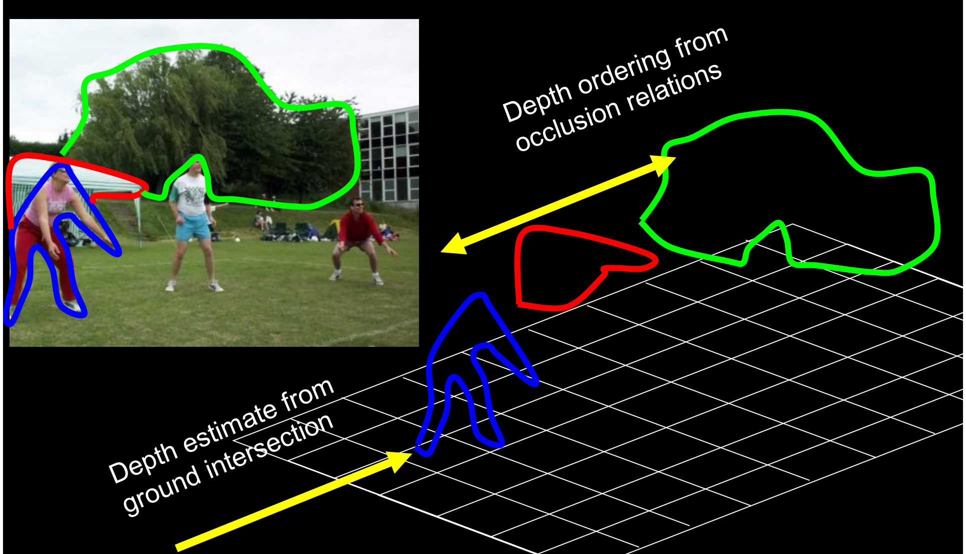


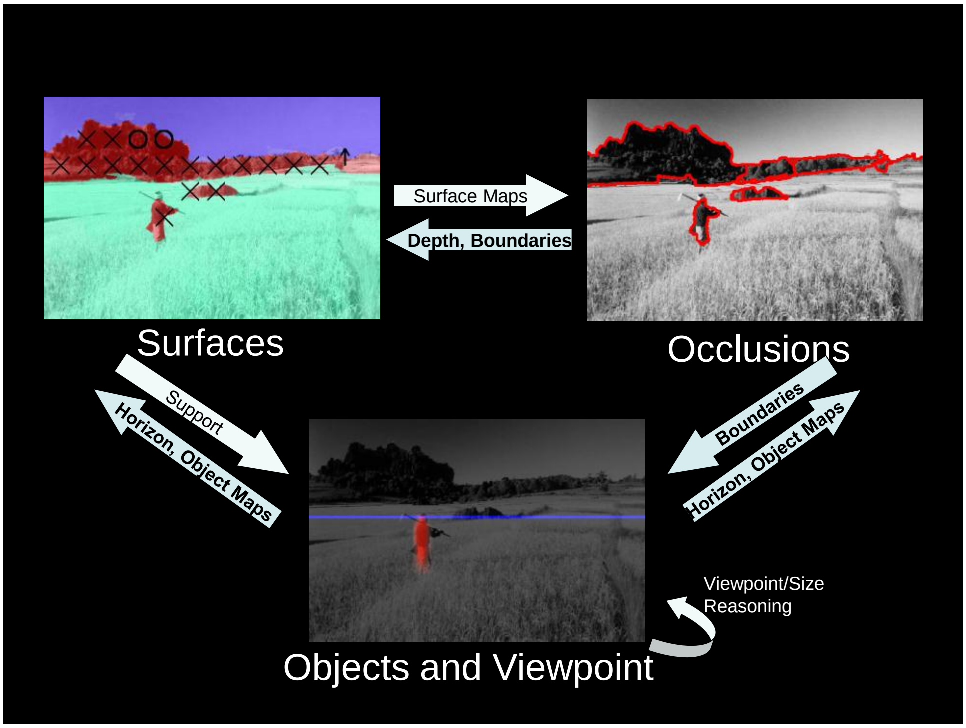
Region labels

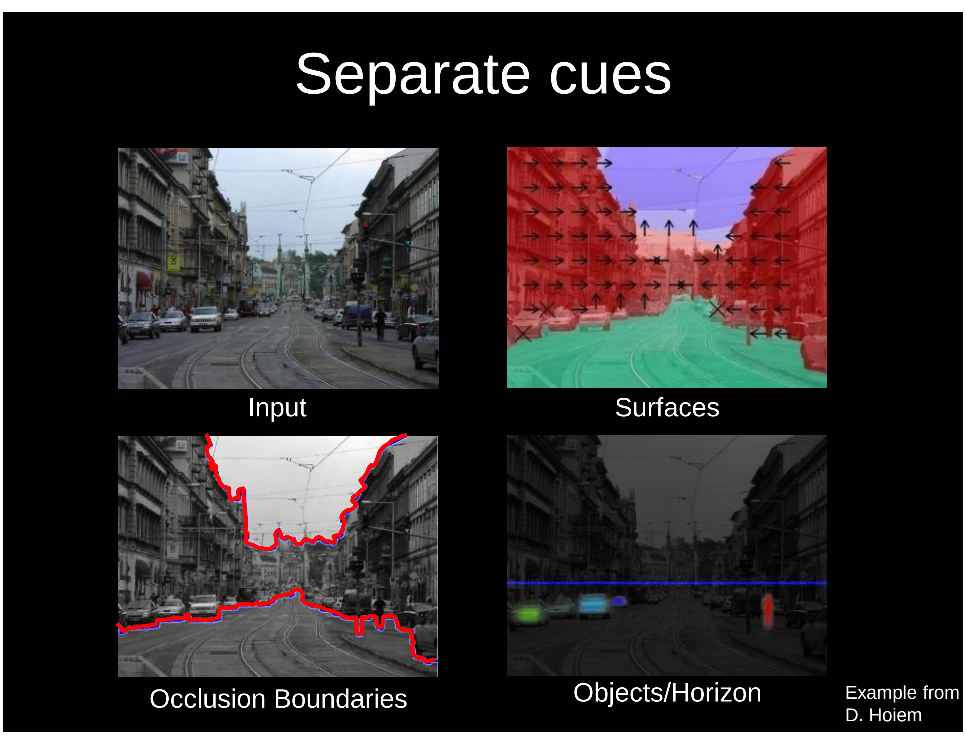
+ Boundaries and objects

Qualitative

Using occlusion cues: Depth ordering and depth estimation

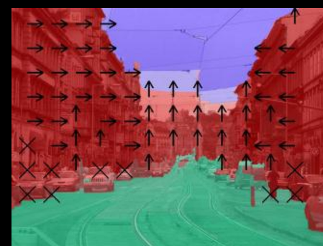



Surface Maps
Depth, Boundaries
Surfaces
Occlusions
Support
Horizon, Object Maps
Boundaries
Horizon, Object Maps
Viewpoint/Size
Reasoning
Objects and Viewpoint


Separate cues
Input
Surfaces
Occlusion Boundaries
Objects/Horizon
Example from
D. Hoiem

Combined reasoning

Input

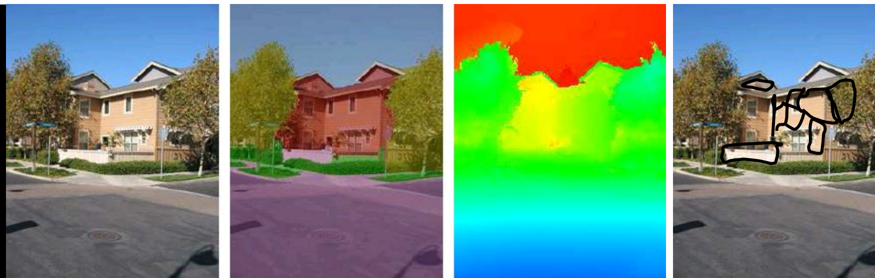


Surfaces

Occlusion Boundaries

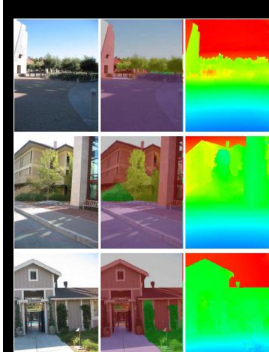
Objects and Horizon

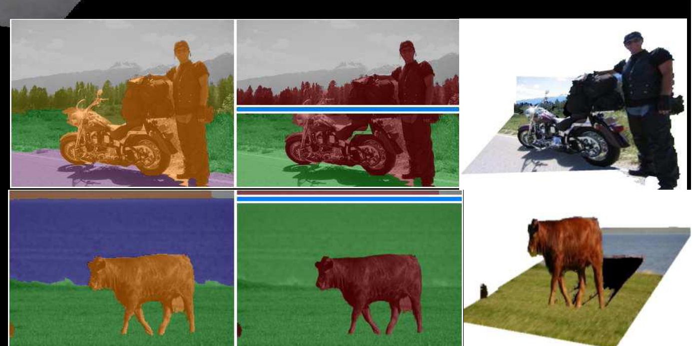
Example from D. Hoiem



- Semantic labels provide strong constraints on local surface orientation
- Semantic labels also provide an estimate of relative depth ordering and occlusion relations

- Outline:
  1. Estimate labels from image features
  2. Estimate point-wise depth (with depth constraints from labels)
  3. Estimate local orientations (with orientation constraints from labels)

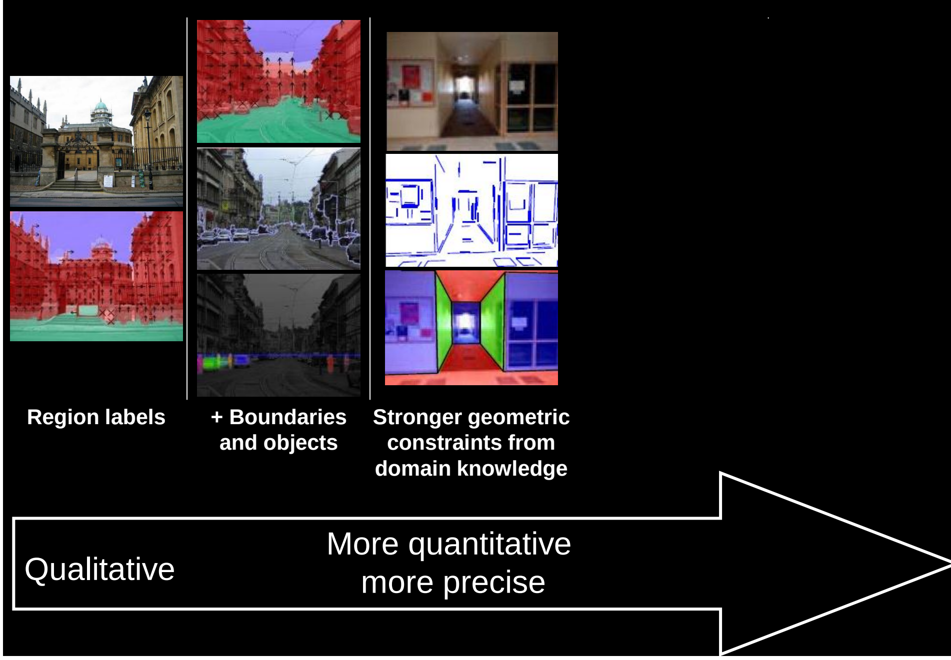




B. Liu, S. Gould, D. Koller. Single image depth estimation from predicted semantic labels. CVPR 2010

Comments

- Plus:
  - Scene geometry (surface geometry and object relations) estimated from image data
  - Scene geometry used explicitly in scene understanding
- Minus:
  - Still mostly bottom-up classification approach
  - No use of domain constraints or constraints governing the physical world



Example

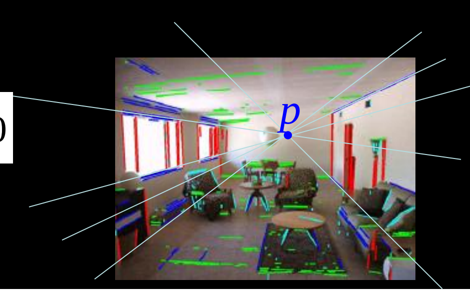
- Using constraints induced by man-made environments in interpreting images
- Examples: Manhattan world, limited vocabulary of object configurations, etc.

Constraint: Manhattan world assumption

- Three dominant directions corresponding to three "orthogonal" vanishing points

$$n_i = K^{-1} v_i$$

$$n_j \cdot n_i = v_j^{\ T} K^{-T} K^{-1} v_i = 0$$

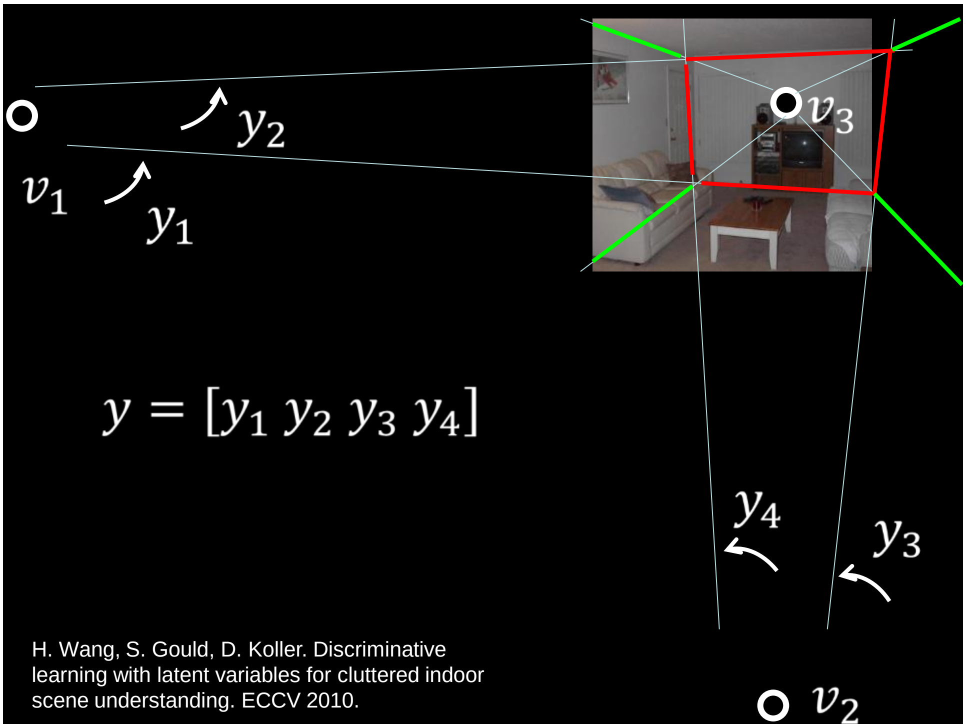


We need to design 4 things

- Parameterization: y
- Features: x
- Scoring/hypothesis evaluation:

$$y_o = \text{argmax}_y\, f(x,y,w)$$

- A way to sample, or generate hypotheses y



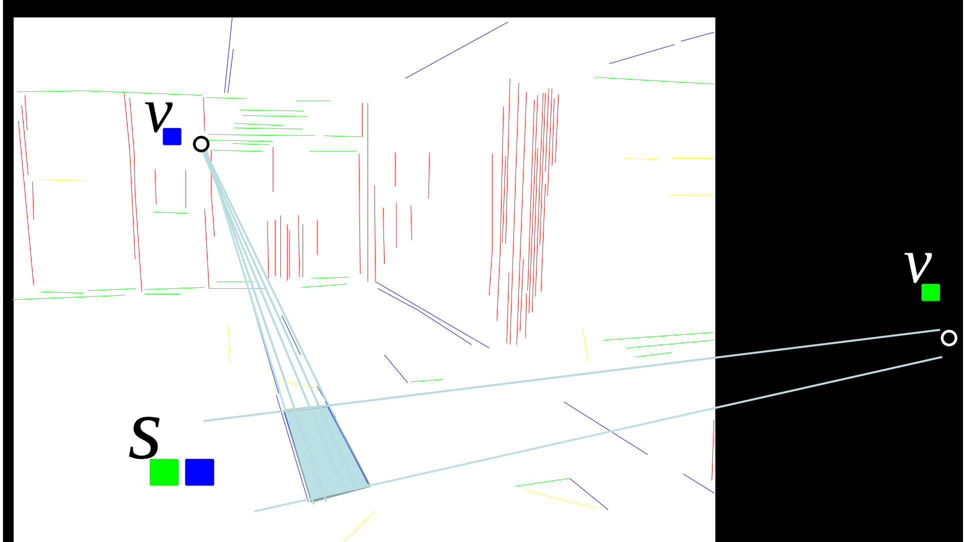
$$y = [y_1\ y_2\ y_3\ y_4]$$

H. Wang, S. Gould, D. Koller. Discriminative learning with latent variables for cluttered indoor scene understanding. ECCV 2010.

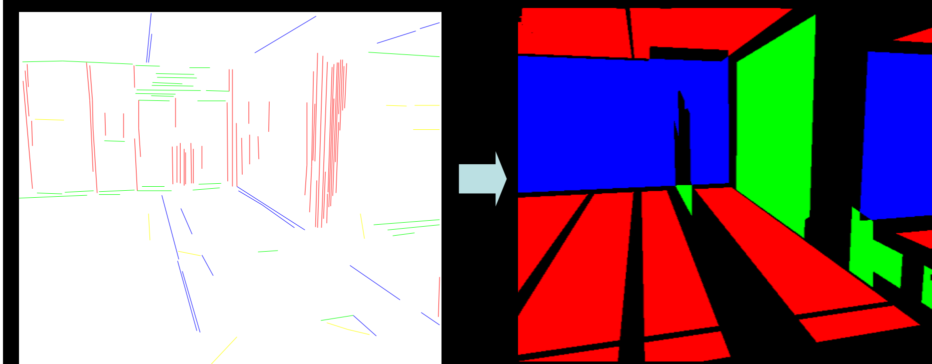
Features

- Features: x
  - Surface layout (see earlier)
  - Lines, regions,…
  - Orientation maps
  - Junctions

Orientation maps: Sweep algorithm



D. Lee, T. Kanade, M. Hebert. Geometric Reasoning for Single Image Structure Recovery. CVPR09.



p is of orientation if

it is in one S

It is in one S

It is in none of S S

Scoring the hypotheses

- Structured prediction

$$y_o = \text{argmax}_y \; w^T \varphi(x,y)$$

V. Hedau, D. Hoiem, D.Forsyth, “Recovering the Spatial Layout of Cluttered Rooms,” International Conference on Computer Vision (ICCV), 2009.

A.G. Shwing, T. Hazan, M. Pollefeys, R. Urtasun, “Efficient Structured Prediction for 3D Indoor Scene Understanding,” Computer Vision and Pattern Recognition (CVPR), 2012.

Vanishing points in input image

x

Score = output of predictor

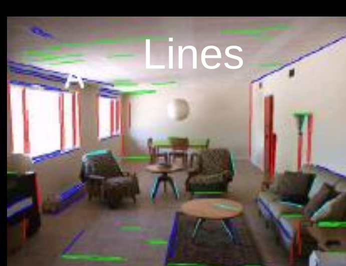
Many initial layout hypotheses

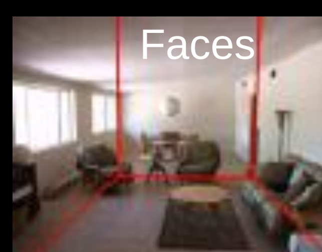
y

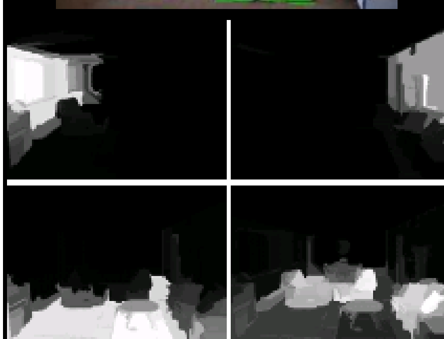
$\rightarrow f(x,y,w)$

[Example from Hedau et al.]

Data for evaluating y







Surface labels	Floor	Left	Middle	Right	Ceiling	Objects
Floor	74/68	0/0	0/1	0/1	0/0	24/30
Left	1/0	75/43	14/44	0/0	1/1	9/12
Middle	1/0	5/2	76/82	4/6	2/1	13/9
Right	1/1	0/0	14/48	73/42	3/2	10/7
Ceiling	0/0	4/3	28/47	2/5	66/45	0/0
Objects	16/12	1/1	5/10	1/2	0/0	76/76

Label confidences from classifier

Definition of mapping function

$$f(x,y,w) = w^{\mathrm{T}}\varphi(x,y)$$

Learned weight vector

Feature vector measuring agreement between lines, faces and labels

$\varphi(x,y)$= Relative sum of lengths of line segments in each face agreeing with labels from appearance-based classifiers

[Example from Hedau et al.]

Learning

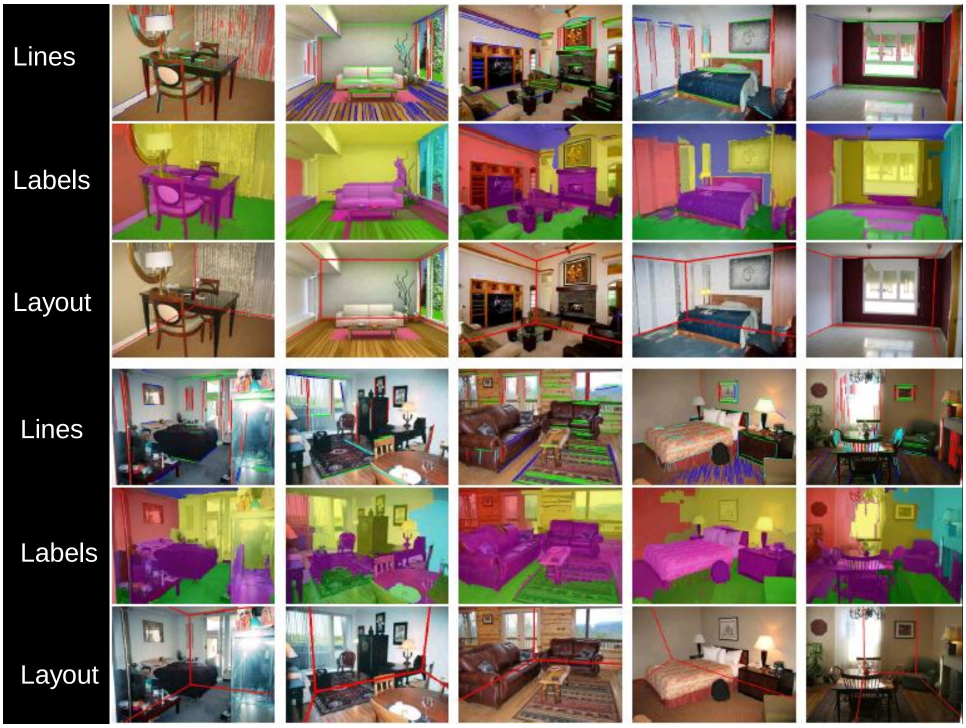
$$\min_{w,\xi} \tfrac{1}{2}||w||^2 + C\textstyle\sum_i \xi_i$$
$$\text{s.t.} \quad \xi_i \geq 0 \ \forall i, \quad \text{and}$$
$$w^T\psi(x_i,y_i) - w^T\psi(x_i,y) \geq \Delta(y_i,y) - \xi_i,$$

Loss function:
- Distance between centroids of true and estimated faces
- Overlap between true and estimated faces
- Number of missing faces

True: y_i Estimated: y

$\Delta(y_i,y)$

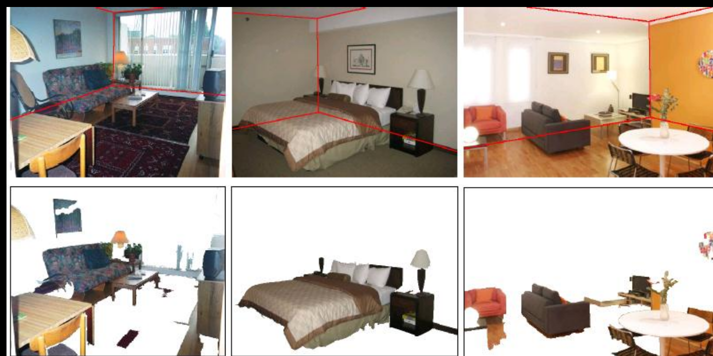
[Example from Hedau et al.]



More detailed interpretation: Clutter vs. free space

Example from H. Wang

H. Wang, S. Gould, D. Koller. *Discriminative Learning with Latent Variables for Cluttered Indoor Scene Understanding*. In Communications of the ACM, 2013.



$$f(x,y,h;w) = w^{T}\varphi(x,y,h)$$

Clutter mask

228-dim feature vector

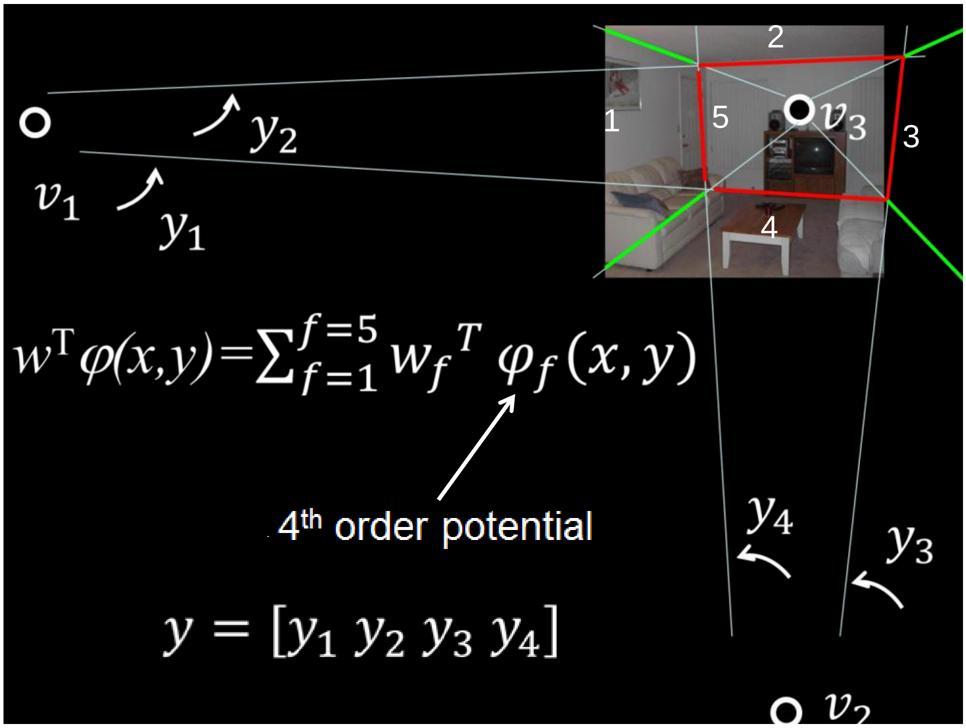
- Measures how the fraction of each face not in clutter agrees with y
- Measures the fraction of the faces not in clutter

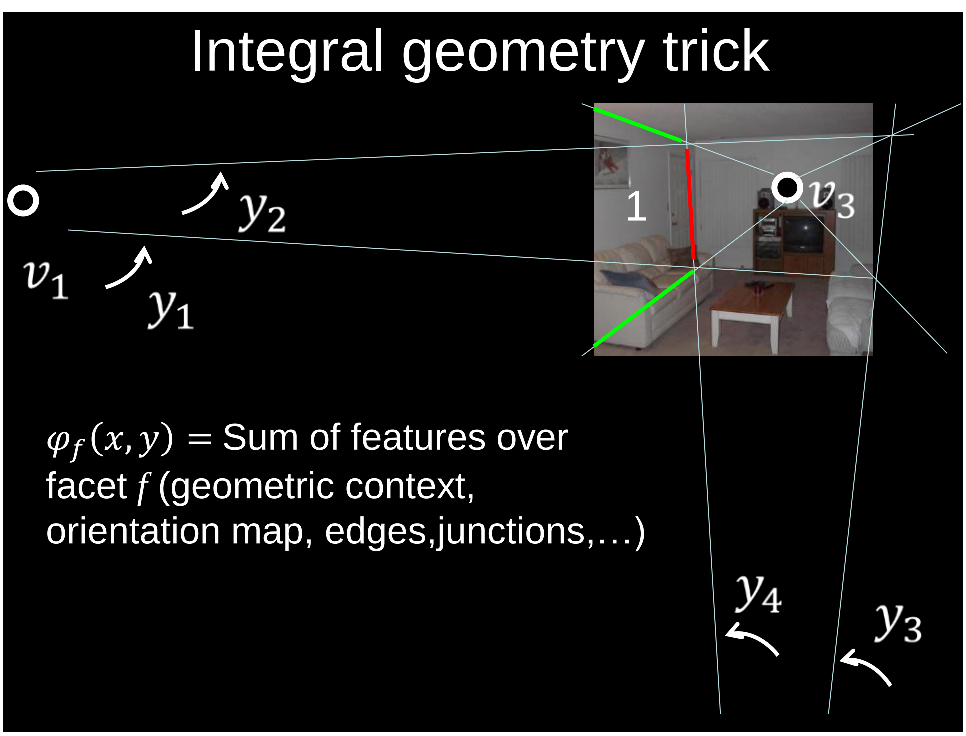
We need to design 4 things

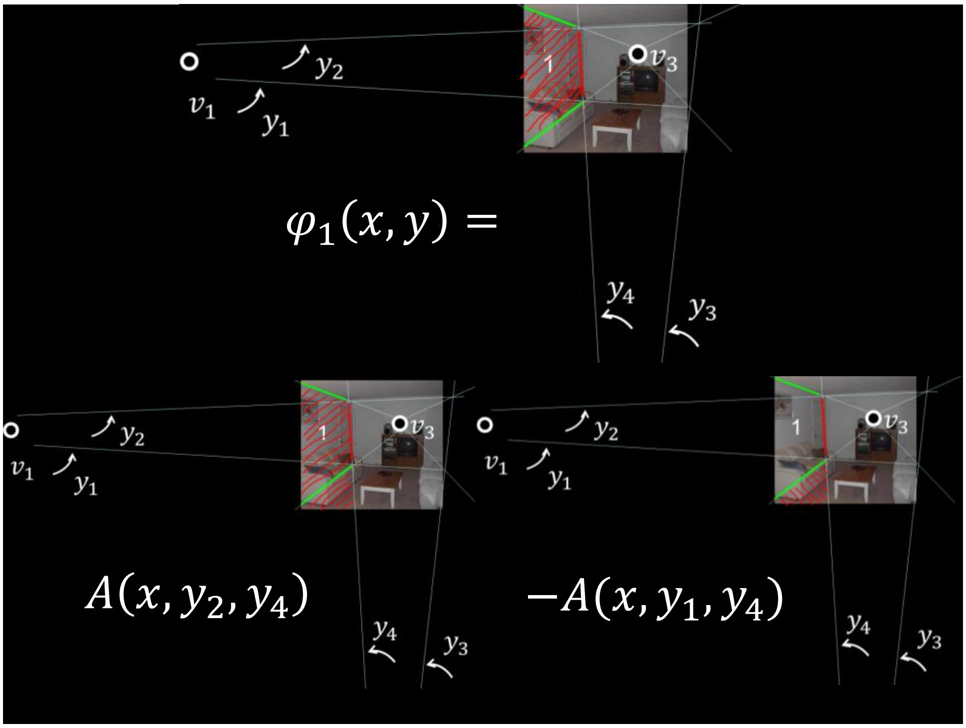
- Parameterization: y
- Features: x
- Scoring/hypothesis evaluation:

$$y_o = \text{argmax}_y\, f(x,y,w)$$

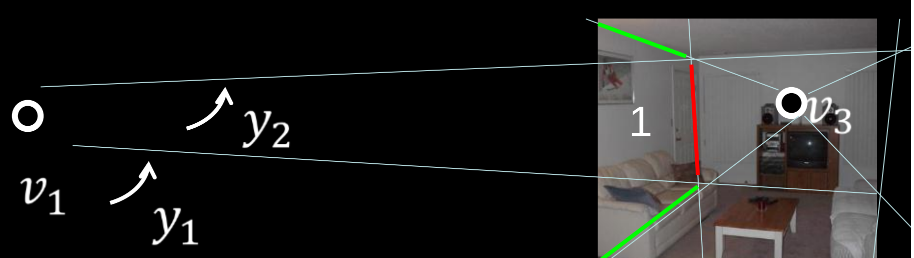
- **A way to sample, or generate hypotheses y**


2
1
5
v_3
3
4
y_2
v_1
y_1
$w^T\varphi(x,y) = \sum_{f=1}^{f=5} w_f{}^T \varphi_f(x,y)$
4th order potential
y_4
y_3
$y = [y_1\ y_2\ y_3\ y_4]$
v_2


Integral geometry trick
y_2
1
v_3
v_1
y_1
$\varphi_f(x,y)$ = Sum of features over facet f (geometric context, orientation map, edges,junctions,…)
y_4
y_3



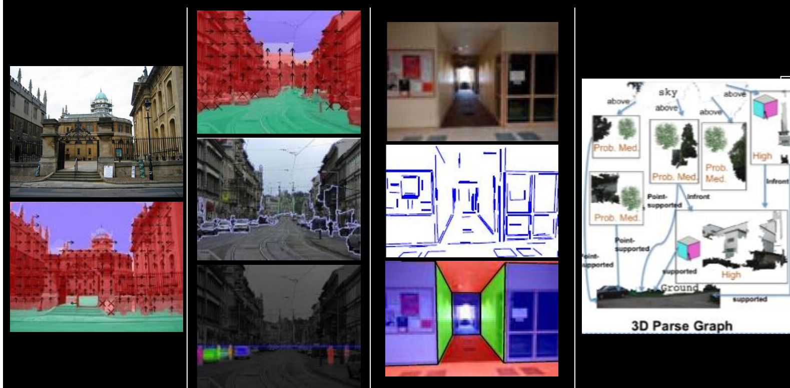
Integral geometry trick



$$\varphi_1(x, y) = A(x, y_2, y_4) - A(x, y_1, y_4)$$

2nd order potential

A.G. Schwing, T. Hazan, M. Pollefeys, R. Urtasun. Efficient Structured prediction for 3D Indoor Scene Understanding. CVPR 2012.
A. Schwing, T. Hazan, M. Pollefeys and R. Urtasun. Efficient Structured Prediction with Latent Variables for General Graphical Models. ICML 2012.



Region labels | + Boundaries and objects | Stronger geometric constraints from domain knowledge | + more constraints

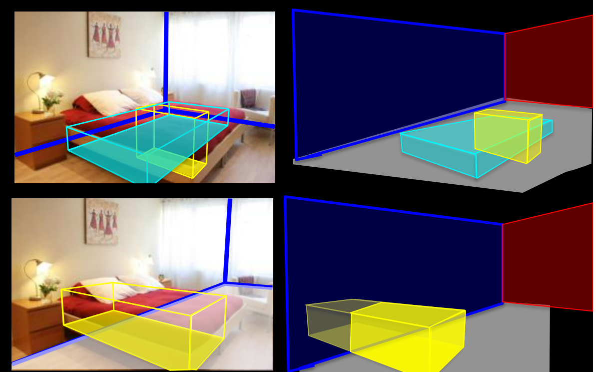
Qualitative — More quantitative more precise — Explicit

Integrating more constraints

- *Constraints*
  - Volumetric constraints
- *Techniques*
  - Structured prediction

Constraints: Solid objects must satisfy volumetric/physical constraints

- Finite volume
- Spatial exclusion
- Containment



D. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating Spatial Layout of Rooms using Volumetric Reasoning about Objects and Surfaces. Advances in Neural Information Processing Systems (NIPS), Vol. 24, 2011.

$$f(x, y)$$

Image

$$y = [r_1 \ldots r_n \; o_1 \ldots o_m]$$

Labeling: Indicator vector of scene configurations + object hypothesis

Penalty term for incompatible configurations

Compatibility of image data with geometric configuration

$$f\left(x,y\right)=w^{T}\psi\left(x,y\right)+w_{\varphi}{}^{T}\varphi\left(y\right)$$

Features from image (surface labels, vanishing points, etc.)

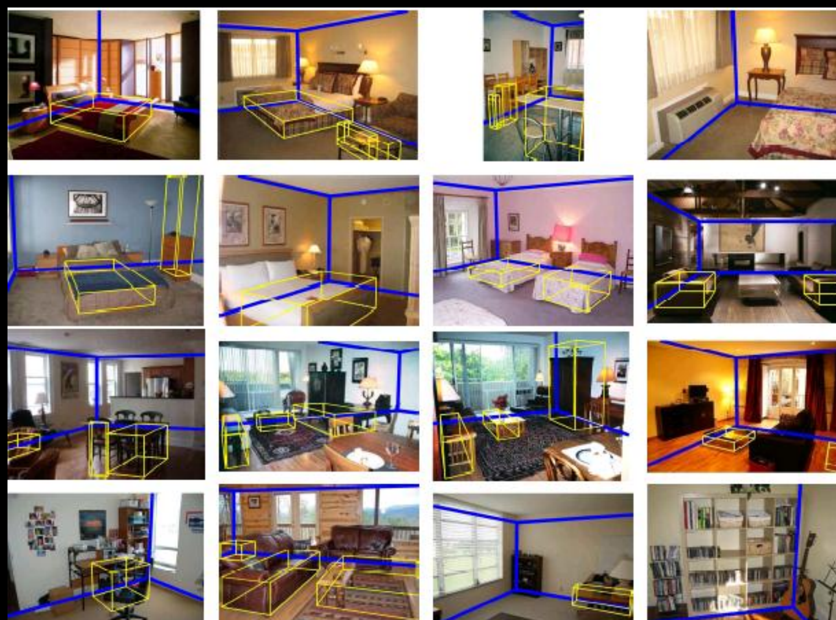
Features of the scene configuration to evaluate constraint violations

$$f\left(x,y\right)=w^{T}\psi\left(x,y\right)+w_{\varphi}{}^{T}\varphi\left(y\right)$$

- Inference:

$$y^{*}=\arg\max_{y} f(x,y)$$

- Training
  - Use structured SVM to estimate w



UIUC 10%

Integrating more constraints

- *Constraints*
  - Volumetric constraints
- *Techniques*
  - Structured prediction
  - Sampling

Representation

- θ = set of all parameters describing scene *and* camera parameters

$$\theta = (r,c)$$

Scene model

Camera model

$$r = (r_b, n, o_1, .., o_n)$$

$$c = (\psi, \varphi, f)$$

Scene layout:
One block +
orientation

$$r = (x,y,z,w,h,l,\gamma)$$

Object model: block

$$o = (x,y,z,w,h,l)$$

L. Del Pero, J. Guan, E. Brau, J. Schlecht, K. Barnard. Sampling Bedrooms. CVPR 2011.

Score function

- *E* = edge points in input image
- θ = hypothesis

Unmatched
Model edges

Unmatched
Input edges

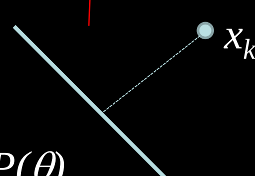
$$p(E\,|\,\theta) = e_{bg}^{N_{bg}} e_{miss}^{N_{miss}} \prod_k e(x_k)$$

How well edge point
matches model line

- Sampling:

Sample hypothesis by
MCMC sampling of

$$P(\theta|E) \sim P(E|\theta)P(\theta)$$



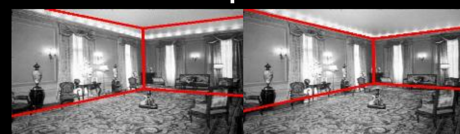
Sampling: Diffusion moves

- Diffusion moves change some part of θ
- Multiple types of moves used in random order

Sample room boundary

Change $r = (x,y,z,w,h,l,\gamma)$

Sample camera

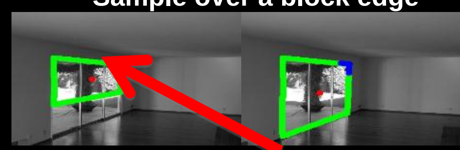


Change $c = (\psi,\varphi,f)$

Sample object parameters

Change $o = (x,y,z,w,h,l)$

Sample over a block edge

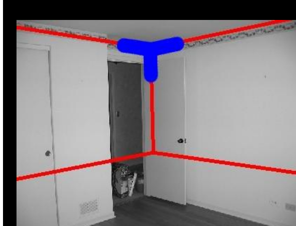


[Example from DelPero et al.]

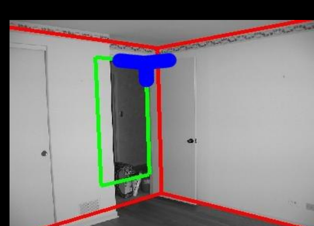
Sampling: Jump moves

- Continuous parameters only so far
- Number of objects is fixed in θ
- We need to sample over the possible number of objects

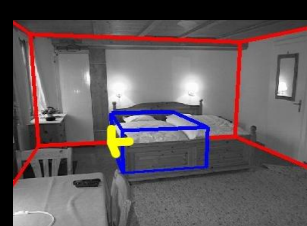
- Jump proposal generated based on corner features
- A corner feature can generate a new block or a new layout



Propose layout

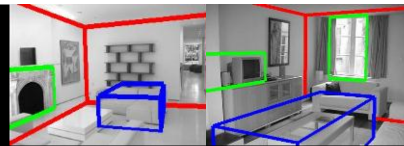


Propose frame



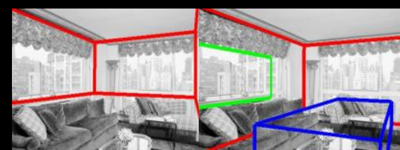
Propose block

[Example from DelPero et al.]



Full fit

Focal length estimation



Blocks explain occlusions

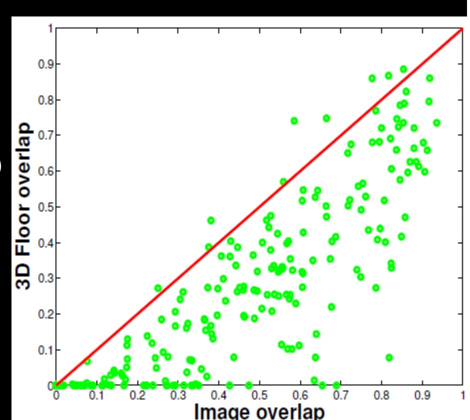
- Shows how to search through hypotheses using MCMC sampling
- No training
- Samples through continuous (sizes, etc.) and discrete (# objects) parameters
- Recovers camera parameters as well

- http://civs.ucla.edu/old/MCMC/MCMC_tutorial.htm

[Example from DelPero et al.]

3D geometry refinement

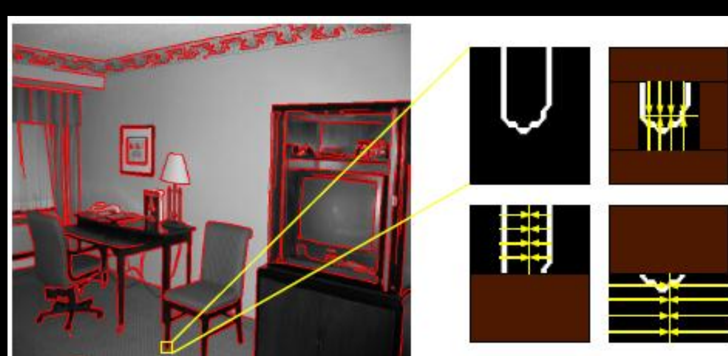
- Two (related) problems:
  – Discrepancy between 2D and 3D error evaluation
  – Large errors in object placement



V. Hedau, D. Hoiem, D. Forsyth. *Thinking Inside the Box: Using Appearance Models and Context Based on Room Geometry*. ECCV 2010.
V. Hedau, D. Hoiem, D. Forsyth. *Recovering Free Space of Indoor Scenes from a Single Image*. CVPR 2012.

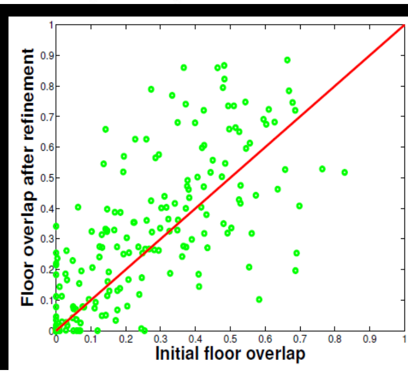
- Possible approach:
  - Error evaluation directly in 3D: Free space estimation
  - Pose refinement by contact estimation
    - For each object
      - Search through micro-hypothesis by varying location of vertices
      - Score using SVM to classify contact/non-contact

[Example from Hoiem et al.]



- Possible approach:
  - Error evaluation directly in 3D: Free space estimation
  - Pose refinement by contact estimation
    - For each object
      - Search through micro-hypothesis by varying location of vertices
      - Score using SVM to classify contact/non-contact

[Example from Hoiem et al.]



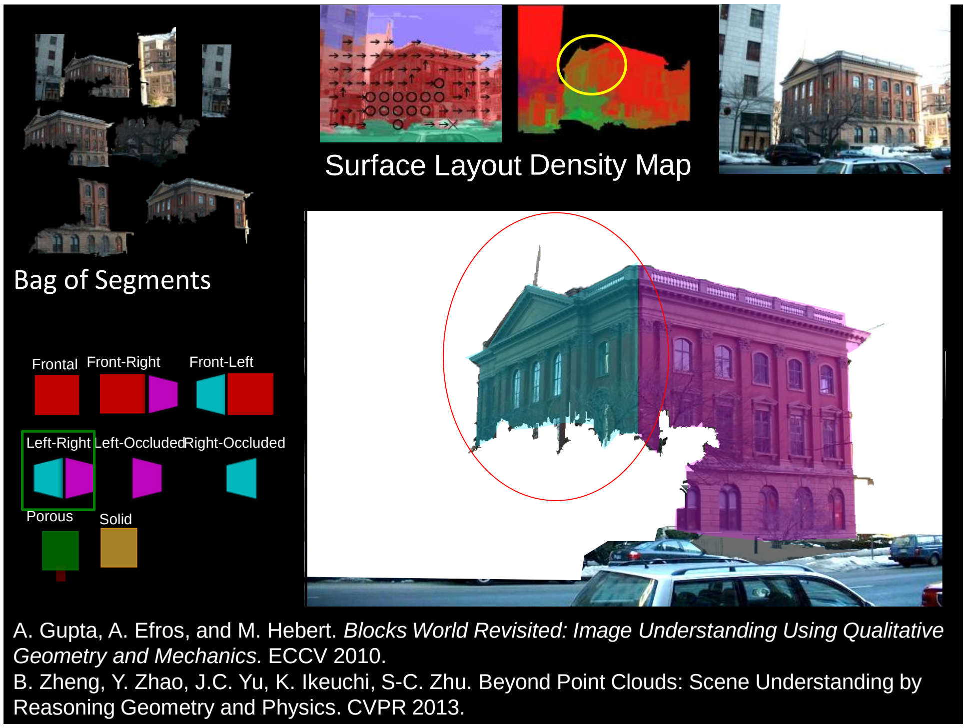
V. Hedau, D. Hoiem, D. Forsyth. Recovering Free Space of Indoor Scenes from a Single Image. Computer Vision and Pattern Recognition (CVPR). 2012.

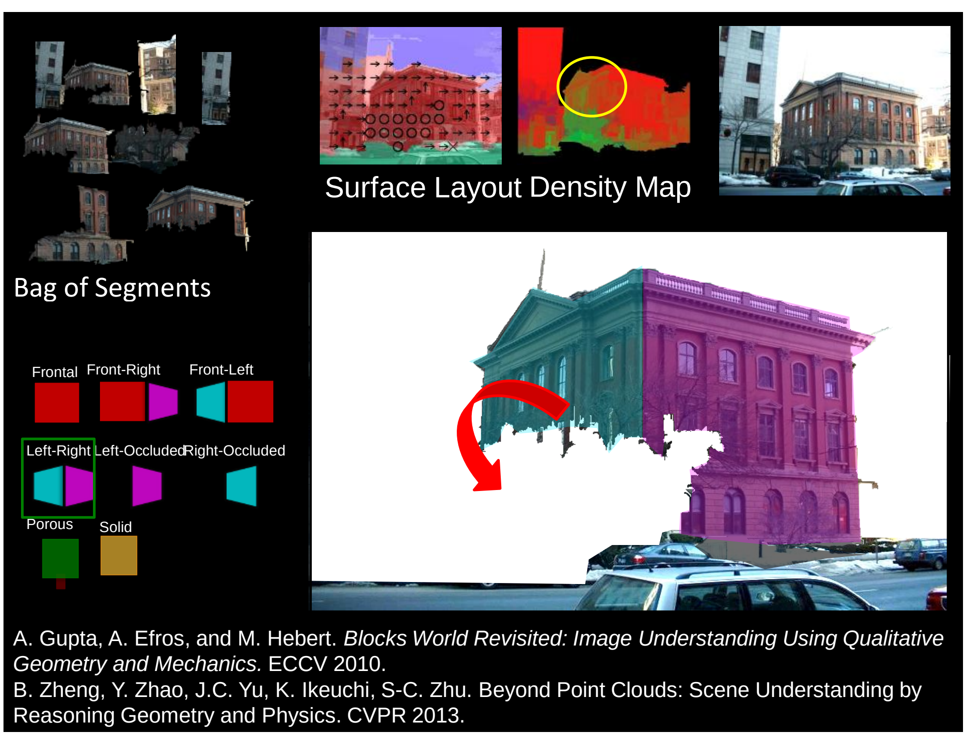
[Example from Hoiem et al.]

Integrating more constraints

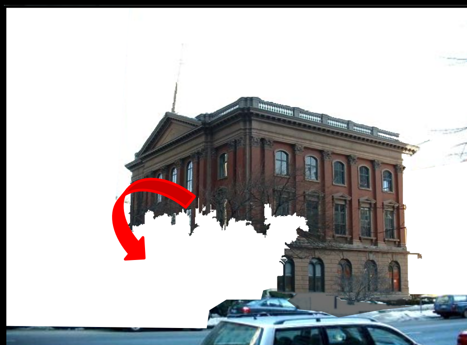
- *Constraints*
  – Volumetric constraints
  – Physical constraints

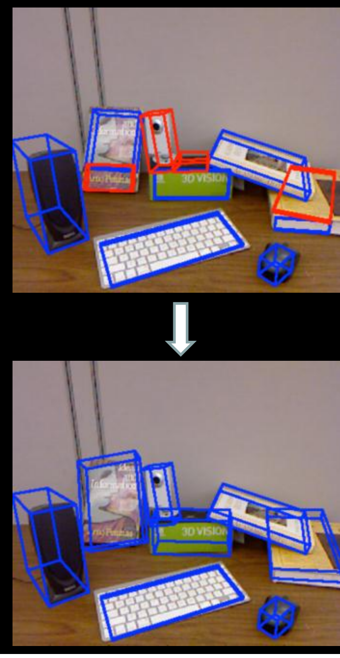
- *Techniques*
  – Structured prediction
  – Sampling
  – Search through hypothesis space


Surface Layout Density Map
Bag of Segments
Frontal
Front-Right
Front-Left
Left-Right
Left-Occluded
Right-Occluded
Porous
Solid
A. Gupta, A. Efros, and M. Hebert. *Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics.* ECCV 2010.
B. Zheng, Y. Zhao, J.C. Yu, K. Ikeuchi, S-C. Zhu. Beyond Point Clouds: Scene Understanding by Reasoning Geometry and Physics. CVPR 2013.


Surface Layout Density Map
Bag of Segments
Frontal
Front-Right
Front-Left
Left-Right
Left-Occluded
Right-Occluded
Porous
Solid
A. Gupta, A. Efros, and M. Hebert. *Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics.* ECCV 2010.
B. Zheng, Y. Zhao, J.C. Yu, K. Ikeuchi, S-C. Zhu. Beyond Point Clouds: Scene Understanding by Reasoning Geometry and Physics. CVPR 2013.

Physical constraints





A. Gupta, A. Efros, and M. Hebert. *Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics.* ECCV 2010.
B. Zheng, Y. Zhao, J.C. Yu, K. Ikeuchi, S-C. Zhu. Beyond Point Clouds: Scene Understanding by Reasoning Geometry and Physics. CVPR 2013.
Z. Jia, A. Gallagher, A. Saxena, T.Chen. 3D-Based Reasoning with Support and Stability. CVPR 2013.

Integrating more constraints

- *Constraints*
  - – Volumetric constraints
  - – Physical constraints
  - – Relative placement
- *Techniques*
  - – Structured prediction
  - – Sampling
  - – Search through hypothesis space

Using relative placement statistics

[Example from Del Pero]

Layout: statistics on relative size

$$r_1 = \frac{\max(w,l)}{\min(w,l)}$$

$$r_2 = \frac{\max(w,l)}{h}$$

Objects: statistics on relative size and contact for each object type i

$$r_{i1} = \frac{h_i}{\max(w_i,l_i)} \quad r_{i2} = \frac{\max(w_i,l_i)}{\min(w_i,l_i)}$$

$$r_{i3} = \frac{h}{h_i} \quad d_i = 1 \text{ if surface contact}$$

Gaussian distribution estimated from prior data
Sampling technique as before but incorporating the priors

Using relative placement statistics

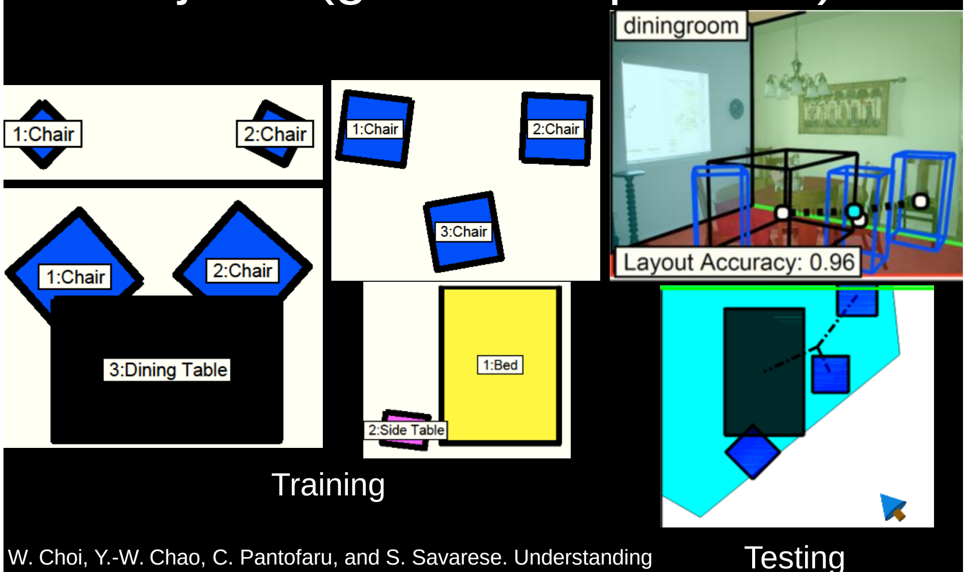
- Is it worth it?

only edge likelihood (no blocks)	26.0 %
+ camera and room prior (no blocks)	24.7 %
+ orientation likelihood (no blocks)	21.3 %
+ random blocks	19.7 %
+ objects	16.3 %

- Yes but still simplified representation of the prior distribution (independent distributions of a few parameters)

L. Del Pero, J. Bowdish, D. Fried, B. Kermgard, E. Hartley and K. Barnard. Bayesian Geometric Modeling of Indoor Scenes. CVPR 2012.
L. Del Pero, J. Bowdish, B. Kermgard, E. Hartley, K. Barnard. Understanding Bayesian Roooms Using Composite 3D Object Models. CVPR 2013.

Generalization to groups of objects (geometric phrases)

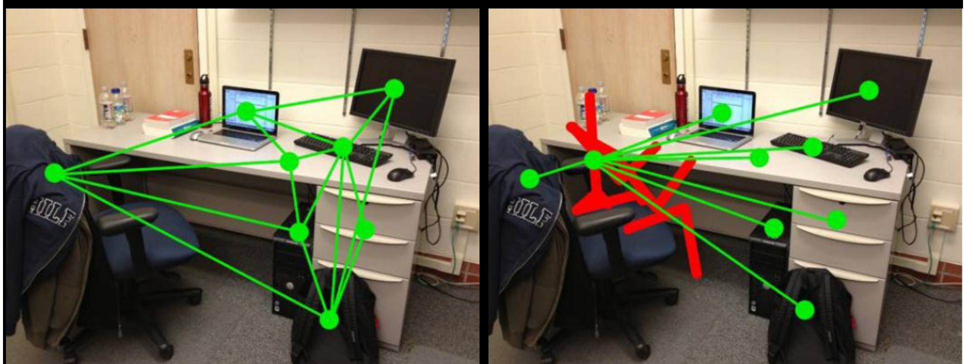


W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese. Understanding Indoor Scenes using 3D Geometric Phrases. CVPR 2013.

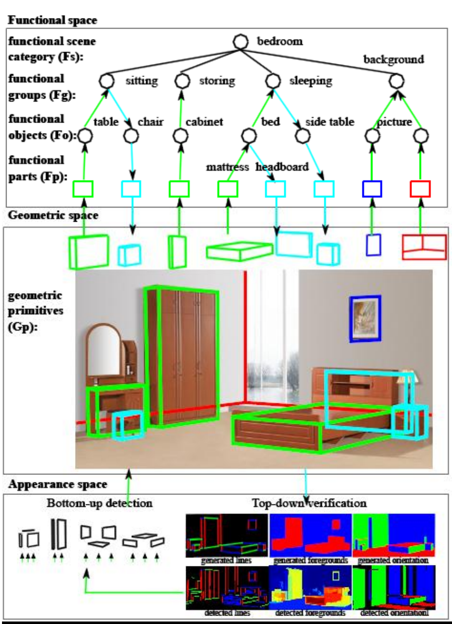
Integrating more constraints

- *Constraints*
  - Volumetric constraints
  - Physical constraints
  - Relative placement
  - Functional constraints
- *Techniques*
  - Structured prediction
  - Sampling
  - Search through hypothesis space

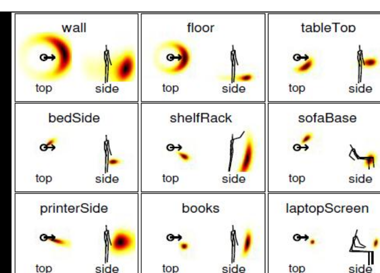
Even more constraints: Functional



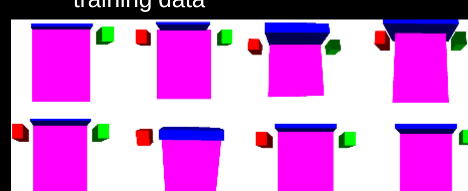
J.J. Gibson. The Theory of Affordances. Lawrence Erlbaum, 1977.



1. Structural model function-geometry-appearance



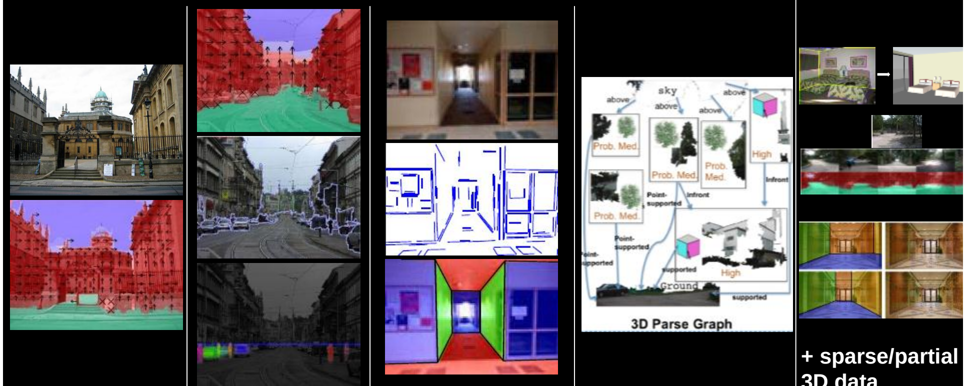
2. Estimate distributions from training data



3. Sample using model

Y.Z. Zhao, S-C. Zhu. Scene Parsing by Integrating Function, Geometry, and Appearance Models. CVPR 2013.

Y. Jiang, H. Koppula, A. Saxena. Hallucinated Humans as the Hidden Context for Labeling 3D Scenes. CVPR 2012.



Region labels

+ Boundaries and objects

Stronger geometric constraints from domain knowledge

+ more constraints

+ sparse/partial 3D data
+ other constraints
+ (large) prior data

Qualitative

More quantitative
more precise

Summary

- Estimating qualitative geometry from input image
- Combining geometric cues with interpretation
- Incorporating more and more constraints
  - Volumetric
  - Physical
  - Relative placement
  - Functional

- Classifiers/regressors + multiple segmentations
- Sampling techniques
- Search through discrete hypothesis space
- Structured prediction
- *Grammars*
- *Using (large) prior data*

Summary

- Estimating qualitative geometry from input image
- Combining geometric cues with interpretation
- Incorporating more and more constraints
  - Volumetric
  - Physical
  - Relativ
  - Functi

How to represent (3D, imprecise) spatial information?
How to generate hypotheses?
How to score hypotheses?
How to search through hypotheses?

- Classifier
- Sampling techniques
- Search through discrete hypothesis space
- Structured prediction
- Grammars
- Using (large) prior data