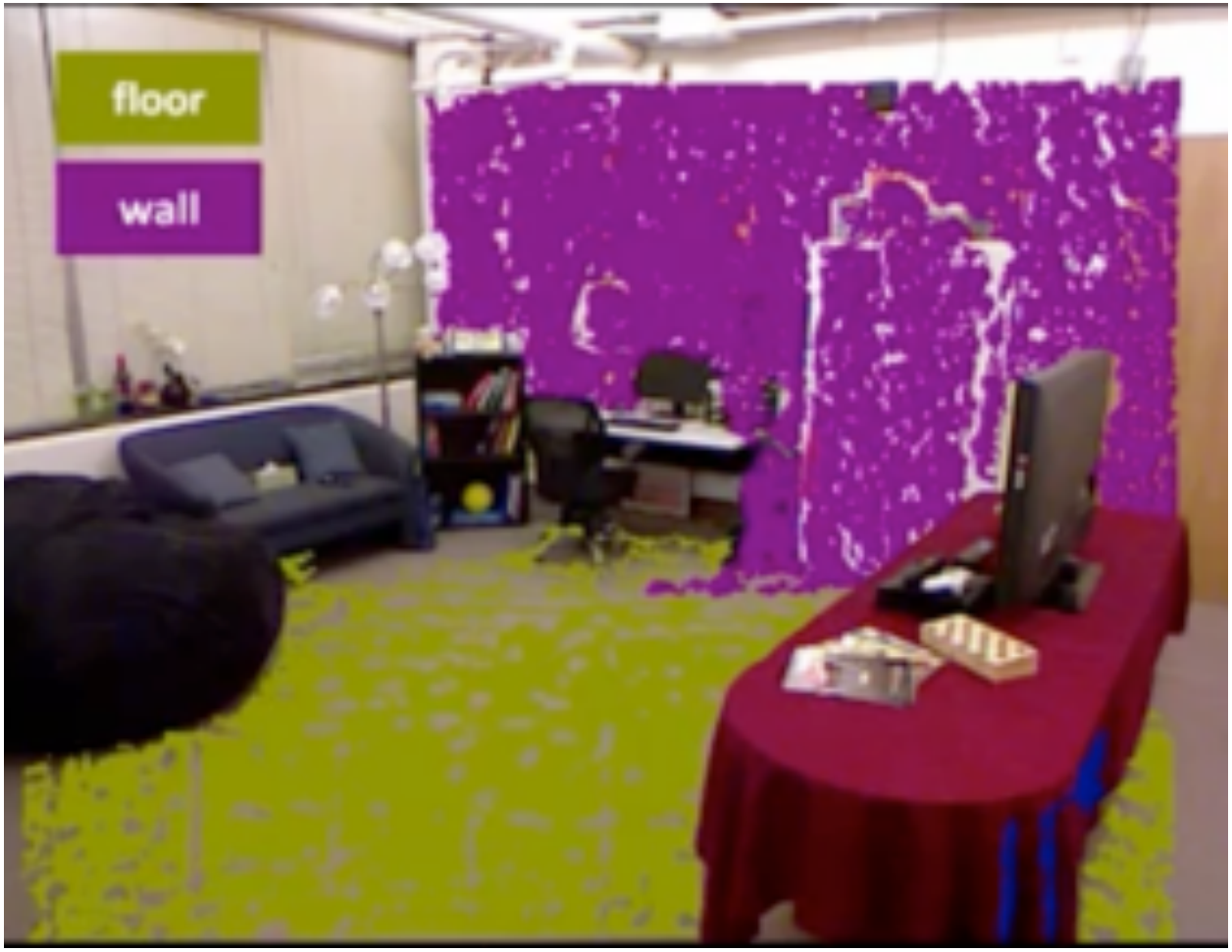
Hallucinated Humans as the Hidden Context for Labeling 3D Scenes

CVPR 2013

Yun Jiang, Hema Koppula and Ashutosh Saxena
Department of Computer Science, Cornell University.
{yunjiang,hema,asaxena}@cs.cornell.edu



Colin Lea – June 2013

Overview

We make the world we live in and shape our own environment.
Orison Swett Marden (1894).

Goal: learn hidden **human-object** relationships and use this as a cue for labeling scenes

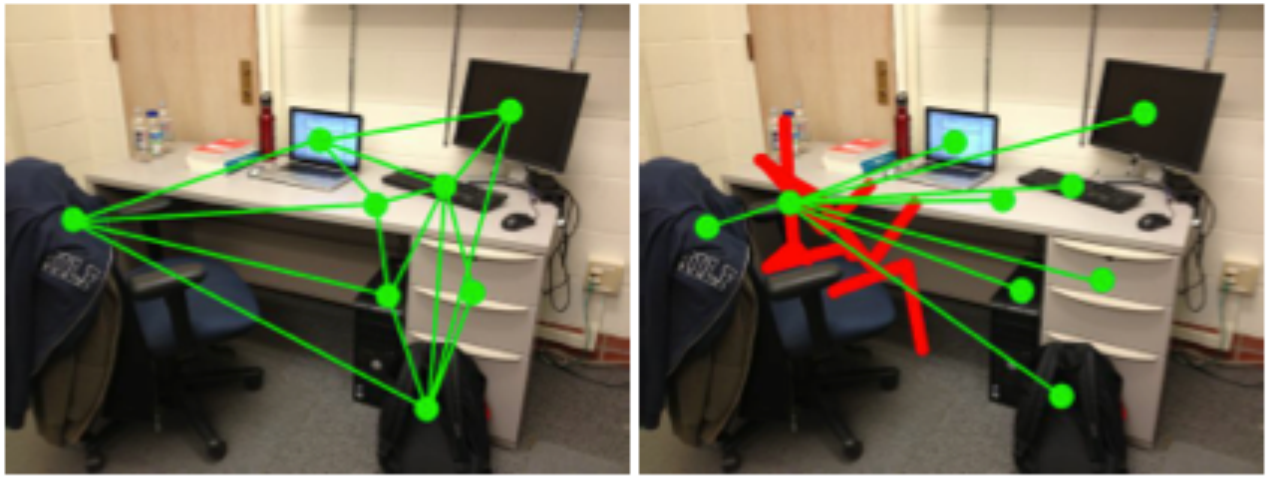


Figure 1: Left: Previous approaches model the relations between observable entities, such as the objects. Right: In our work, we consider the relations between the objects and hidden humans. Our key hypothesis is that even when the humans are never observed, the human context is helpful.

Data

Cornell RGB-D indoor dataset

24 offices, 28 homes, 550 RGB-D views.

Segments can have 1 object label, multiple attributes

Attributes: {*wall, floor, flat horizontal surfaces, furniture, fabric, heavy, seating-areas, small-objects, table-top-objects, electronics*} [10x]

Objects: {*wall, floor, tableTop, tableDrawer, tableLeg, chairBackRest, chairBase, chairBack, monitor, printerFront, printerSide, keyboard, cpuTop, cpuFront, cpuSide, laptop, book, paper, sofaBase, sofaArm, sofaBackRest, bed, bedSide, quilt, pillow, shelfRack*} [26x]

Hallucinating Humans for Robotic Scene Understanding

Yun Jiang and Ashutosh Saxena

Cornell University

Approach

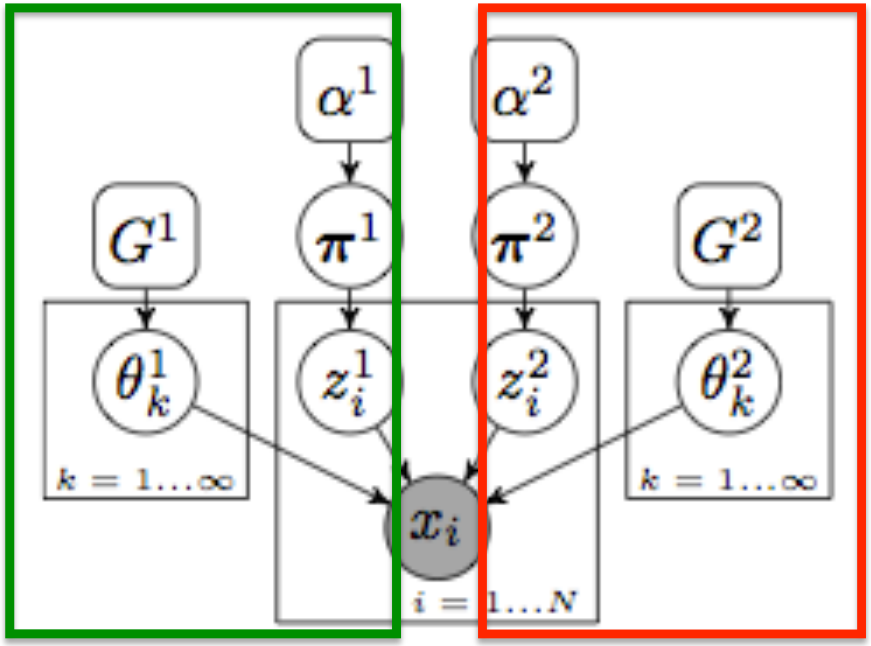
Procedure: Take 3D pointcloud and label objects using shape, appearance and hallucinated human context

New Model: Infinite Factored Topic Model (IFTM)

Topics for (a) human context (b) object affordance

Topics used as features for scene label classifier

Human Context



(b) 2D infinite factored topic model

Object Affordance

Representation

Human Configuration:

Pose library (6 configs)

{Pose, X, Y, Z, Theta}

[From CAD-60 dataset]

Figure 2: Six types of human poses extracted from Kinect

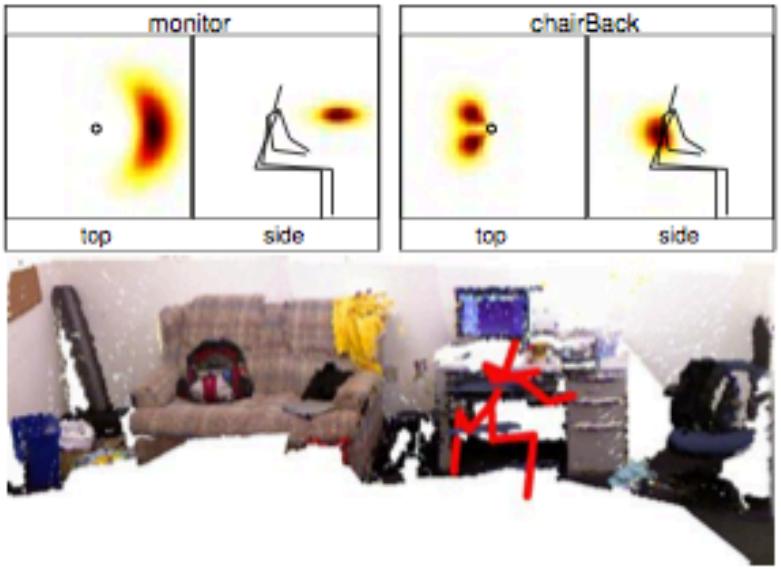
Object Affordance:

Distribution over {X,Y,Z,Theta}

relative to human pose

e.g. Small objects close to human

Books can be close to *or* far away



Mixture of topics!

Priors

Human Configuration:

Physics:

1) Kinematics: Collisions detection

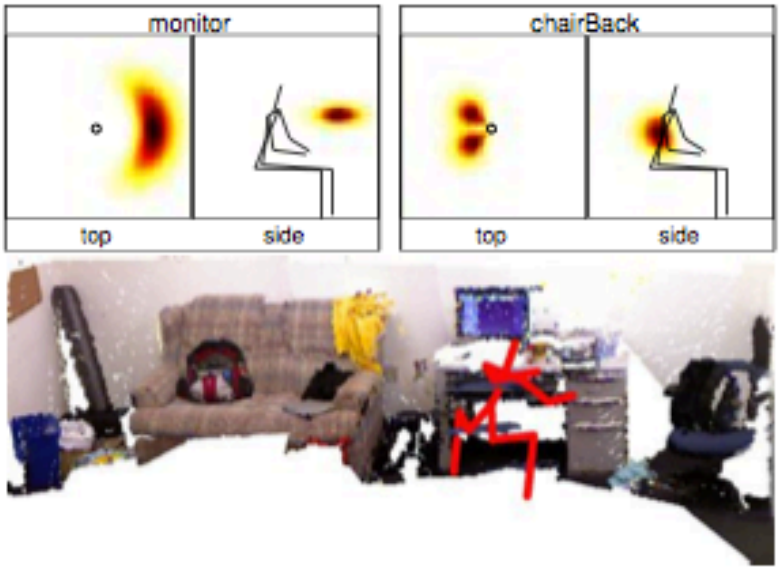
2) Dynamics: supported by ground?

Figure 2: Six types of human poses extracted from Kinect

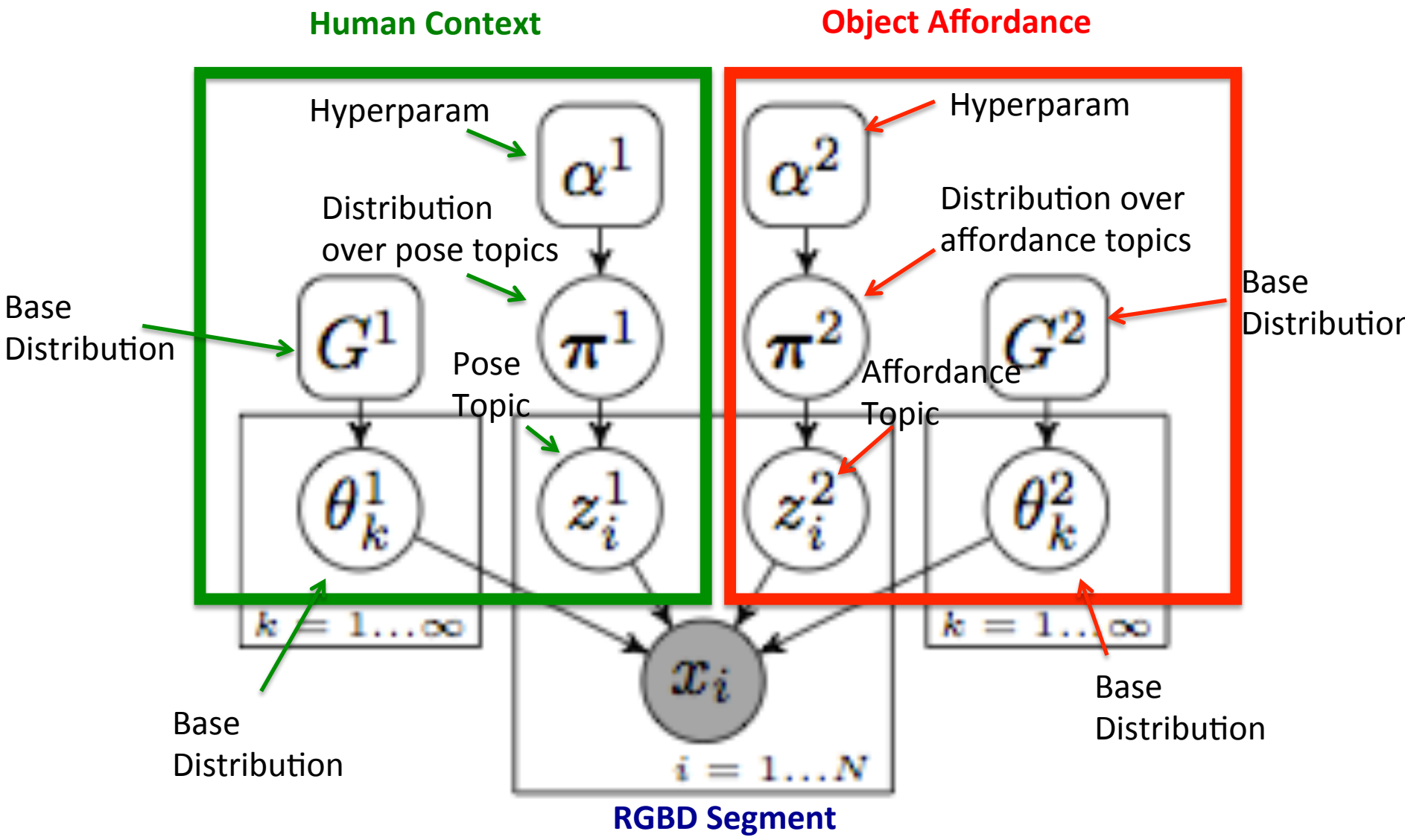
Object Affordance:

1) Proximity (should be close)

2) Symmetry (left/right)



Infinite Factored Topic Model

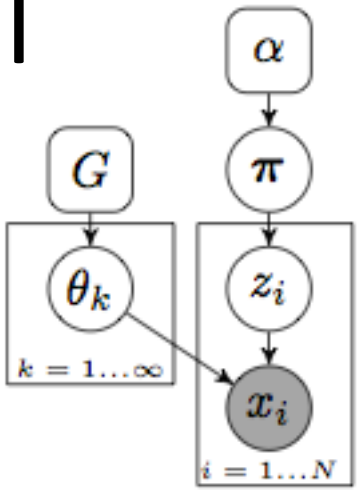


Background: DP Mixture Model

Process for drawing data from set of topics

$$\theta_k \sim G, \;\; b_k \sim Beta(1, \alpha), \;\; \pi_k = b_k \prod_{i=1}^{k-1}(1 - b_i).$$

$$x|z, \theta \sim F(\theta_z).$$



-Gibbs sampling to get marginal distrbution of z_i, $theta_k$

-Get pi from stick breaking

-Sample topic (z) from Chinese Restaurant Process

$$z|\boldsymbol{\pi} \sim \boldsymbol{\pi}; \qquad z_i = z|\boldsymbol{z}^{-i} = \begin{cases} \frac{n_z^{-i}}{N-1+\alpha} & \text{if } z \text{ is previously used} \\ \frac{\alpha}{N-1+\alpha} & \text{otherwise} \end{cases}$$

Benefits over GMM:

Prior over topics

Variable number of topics

Infinite Factored Topic Model

Each (L) topic determined independently

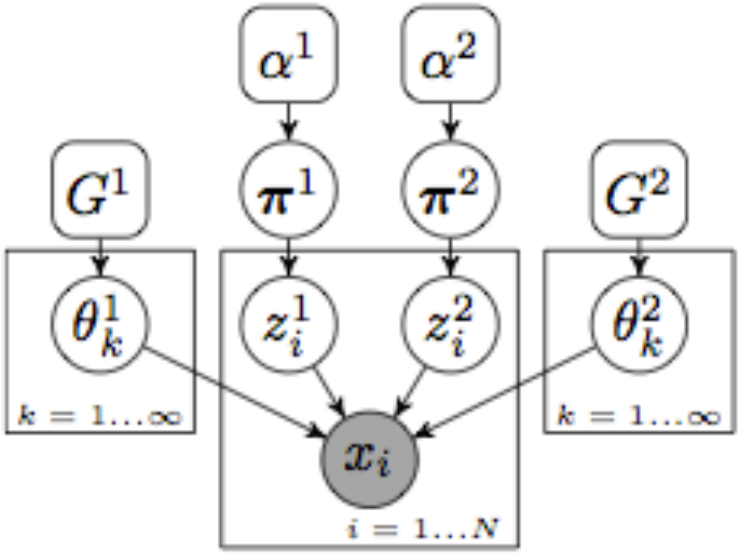
xi = 3D location of ith object in the scene

$F(xi;Theta^H,Theta^O) = F_{distance}F_{rel\_angle}F_{height}$ [log-normal,von Mises, normal]

$Theta^H$ = human pose

G^O = Normal distribution

G^H = uniform over valid poses



IFTM for L=2

Learning Topics

Gibbs sampling to sample Thetas

Step 1) Sample topic assignments

$$z_i^H = z \quad \propto \quad \begin{cases} \frac{n^H_{-i,z}}{N+m-1+\alpha^H} F(x_i, \theta^H_z, \theta^O_{z_i^O}) & n^H_{-i,z} \geq 0, \\ \frac{\alpha^H/m}{N+m-1+\alpha^H} F(x_i, \theta^H_z, \theta^O_{z_i^O}) & \text{otherwise} \end{cases}$$

$$z_i^O = z \quad \propto \quad \begin{cases} \frac{n^O_{-i,z}}{N+m-1+\alpha^O} F(x_i, \theta^H_{z_i^H}, \theta^O_z) & n^O_{-i,z} \geq 0, \\ \frac{\alpha^O/m}{N+m-1+\alpha^O} F(x_i, \theta^H_{z_i^H}, \theta^O_z) & \text{otherwise} \end{cases}$$

Step 2) MAP of Thetas [Posterior is too expensive!]

Calculate argmax of means, variances, concentrations

Scene Labeling

Use sampled affordances as features

Set affordance topics as top K sampled topics $Theta^{O}$

For new scene:

Repeatedly sample $z_i^{O}, z_i^{H}, theta^{H}$

Create histogram of sampled z_i^{O} as feature

Classify with “other” algorithm [No details!!]

Results & Discussion

Results

Node features for segment i.

Description	Count
Visual Appearance	**48**
N1. Histogram of HSV color values	14
N2. Average HSV color values	3
N3. Average of HOG features of the blocks in image spanned by the points of a segment	31
Local Shape and Geometry	**8**
N4. linearness (λ_{i0} - λ_{i1}), planarness (λ_{i1} - λ_{i2})	2
N5. Scatter: λ_{i0}	1
N6. Vertical component of the normal: $\hat{n}_{iz}$	1
N7. Vertical position of centroid: c_{iz}	1
N8. Vert. and Hor. extent of bounding box	2
N9. Dist. from the scene boundary (Fig. 2)	1

Algorithm	Image & Shape	Human Context	Obj-obj Context	Object Labeling					
				Office Scenes			Home Scenes		
				micro	macro		micro	macro	
				P/R	prec	recall	P/R	prec	recall
chance				5.88	5.88	5.88	5.88	5.88	5.88
max class				26.33	26.33	5.88	29.38	29.38	5.88
Affordances		✓		29.13	16.28	16.67	33.62	16.37	15.30
Appearance	✓			77.97	69.44	66.23	56.50	37.18	34.73
Afford. + Appear.	✓	✓		79.71	73.45	69.76	59.00	38.86	37.54
Koppula et al. [19]	✓		✓	84.06	80.52	72.64	**73.38**	56.81	54.80
Full Model	✓	✓	✓	**85.22**	**83.20**	**74.11**	72.50	**59.07**	**56.02**

Algorithm	Image & Shape	Human Context	Obj-obj Context	Attribute Labeling							
				Office Scenes				Home Scenes			
				micro		macro		micro		macro	
				prec	recall	prec	recall	prec	recall	prec	recall
chance				12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5
max class				22.89	22.89	22.89	12.5	31.4	31.4	31.4	12.5
Affordances		✓		47.93	32.04	42.85	29.83	53.92	36.07	41.19	26.21
Appearance	✓			85.82	66.48	86.58	62.52	77.80	55.21	60.01	42.20
Afford. + Appear.	✓	✓		87.05	68.88	**87.24**	65.42	79.02	59.02	70.45	46.57
Koppula et al. [19]	✓		✓	87.92	71.93	84.04	67.96	83.12	70.03	76.04	58.18
Full Model	✓	✓	✓	**88.40**	**76.73**	85.58	**74.16**	**83.42**	**70.28**	**79.93**	**64.27**

Results

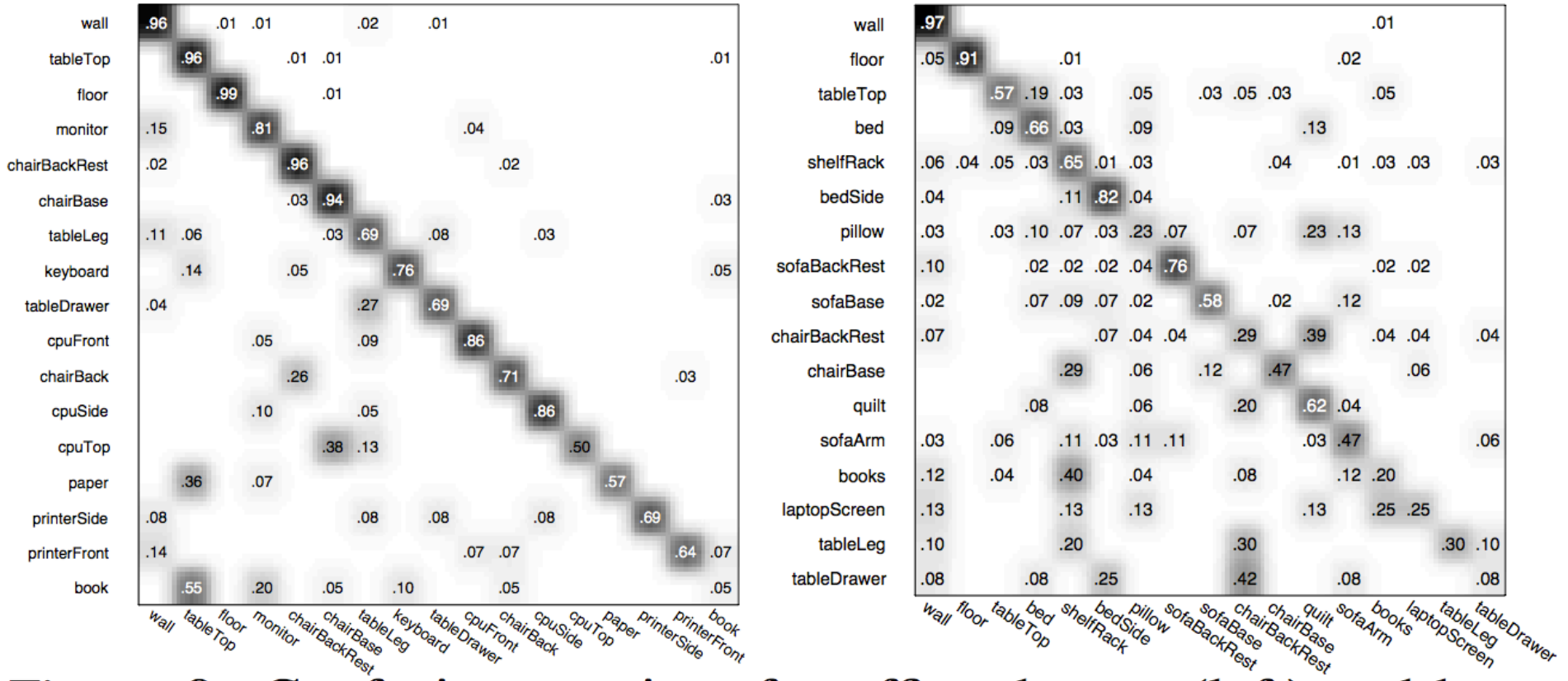


Figure 8: Confusion matrices for office dataset (left) and home dataset (right) using the full model.

Discussion

Are the sampled human poses meaningful?

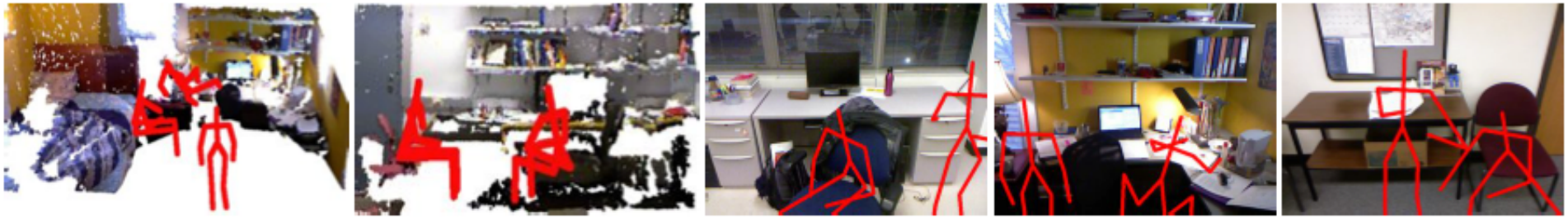


Figure 5: Top sampled human poses in different scenes. The first two are from stitched point-cloud from multiple RGB-D views, and the last three scenes are shown in RGB-D single views.

-sitting on the edge of the bed

-standing close to the desk (easy access to table+shelf)

-on chairs with correct orientation (office scene)

-successfully identifies the workspaces in the office

-naturally explain arrangement of monitors, keyboards and computers

Discussion

Are the discovered affordances meaningful?

+chairBase is often associated with a sitting pose

+computers can either be on the table or on the floor

-wall is more to the front than back

-monitor is biased to the side

Biases are attributed to lack of data and imperfect "valid" poses (errors in physics model)

Discussion

Are the discovered affordances meaningful?

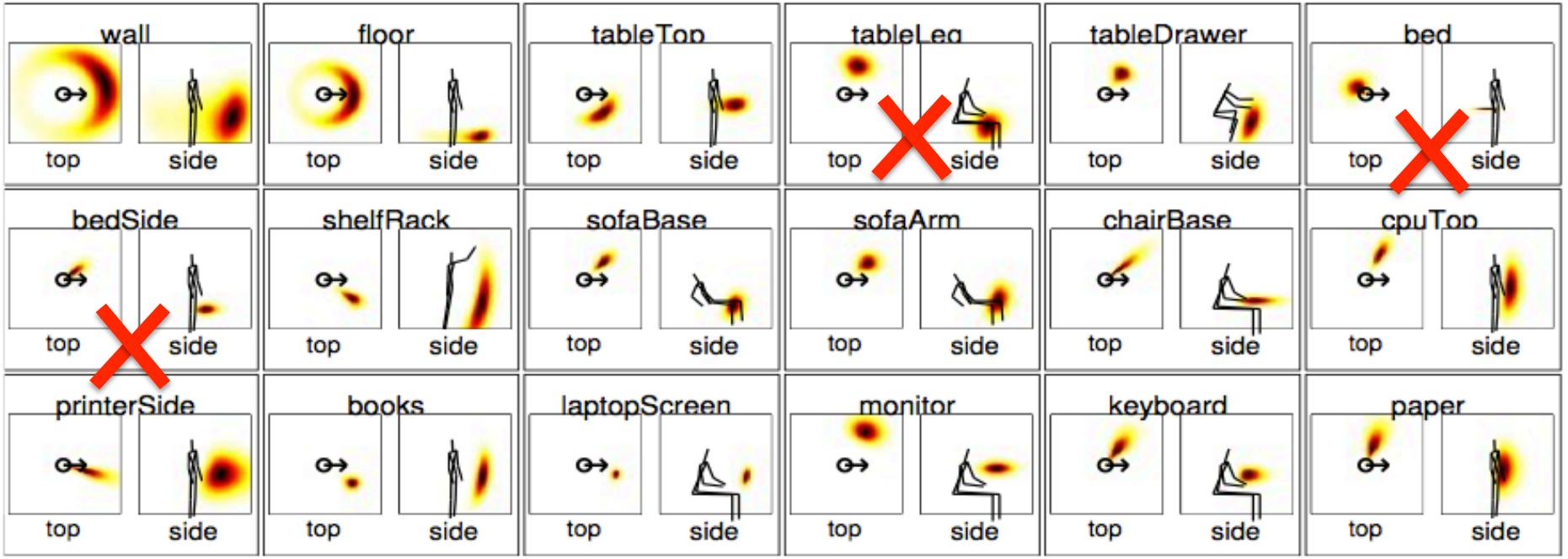


Figure 6: Examples of learned object-affordance topics. An affordance is represented by the probabilistic distribution of an object in a $5 \times 5 \times 3$ space given a human pose. We show both projected top views and side views for different object classes.

Discussion

Can we obtain object-object relations from object affordances?

Yes – Convolve human-object relations

e.g. keyboard-human x human-monitor

Can model N^2 obj-obj relations w/ only N human-obj relations!

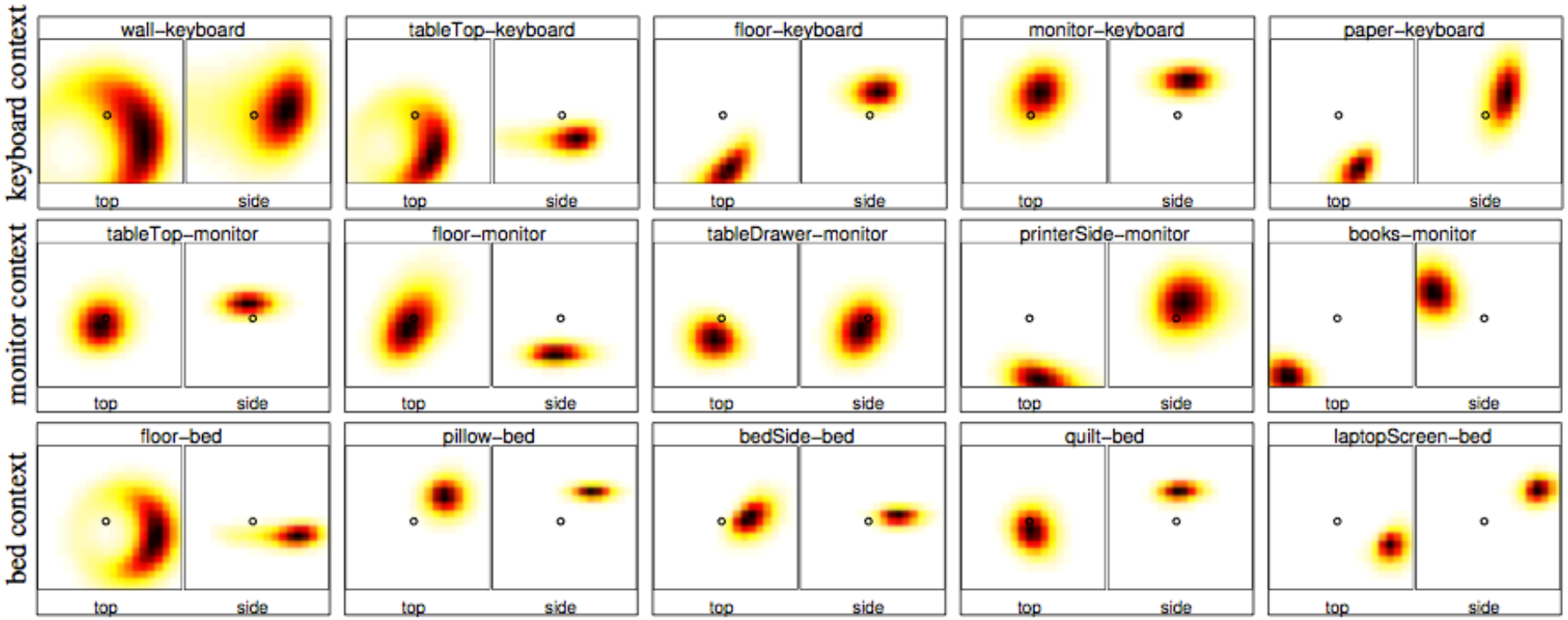


Figure 7: Object-object context obtained from our learned human context. Each pair of the top- and side-view of a heatmap with the title of 'obj1-obj2' shows the distribution of obj1 given obj2 at the center facing right. For example, in the first row the keyboard is in the center of the image and the heat-maps show the probability of finding other related objects such as table top, monitor, etc.

Discussion

Does human context helps in scene labeling? 

Yes… see results

However, using both human-object + obj-obj is better.

(Especially for small objects)

Takeaways

Good:

- Requires less training data because of pose sampling
- Naturally discovers hidden relationships and affordances

Bad:

- Appears to have a lot of bias [unsure if the model or the lack of data is the problem]