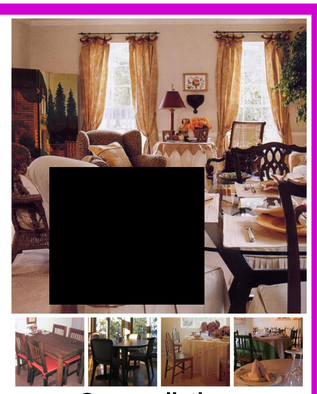
Incorporating Scene Context and Object Layout into Appearance Modeling

Hamid Izadinia*, Fereshteh Sadeghi*, Ali Farhadi
{izadinia,fsadeghi,ali}@cs.washington.edu

IEEE 2014 Conference on Computer Vision and Pattern Recognition
CVPR 2014

What is behind the black box?

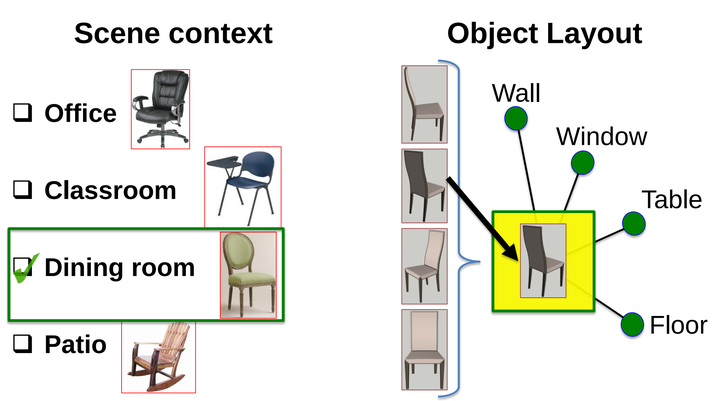
- Joint reasoning over scene and objects
- Scene type
  - Expected objects and their style
- Nearby objects
  - Scene Type
  - Scene Layout
  - Object category



Our predictions

Our Approach

- Reasoning over interlacing components of a scene
  - Scene category
  - Scene specific appearance of objects (style, pose)
  - Objects layout in the scene



Model

- Joint optimization on **topology** and **structure parameters**

$$\min_{\mathcal{G}_c,\mathcal{W}_c^a,\mathcal{W}_c^d,\mathcal{H},\xi} \quad \sum_{j=1}^{p_c}\|W_{c,j}^a\|_2^2+\sum_{j,k=1}^{p_c}\|W_{c,j,k}^d\|_2^2+\lambda_1\sum_{i=1}^{n}\xi_i+\lambda_2\|\mathcal{G}_c\|_\bullet$$

$$\mathcal{D}(x_i,\mathcal{H}_i,\mathcal{W}_c^a,\mathcal{W}_c^d,\mathcal{G}_c)\geq\max_{\mathcal{H}^*}\mathcal{D}(x_i,\mathcal{H}_i^*,\mathcal{W}_c^a,\mathcal{W}_c^d,\mathcal{G}_c)+\Delta(\mathcal{H}_i,\mathcal{H}_i^*)-\xi_i,\xi_i\geq 0\ \ \forall i$$

Structure parameters | Topology

Learning

- Decoupling optimization of **G** and other parameters
- Optimize **G** by weighted maximum spanning tree
  - Nodes: Frequency of objects
  - Edges: Spatial consistency of object pairs
- Optimize **W** using large margin structured learning

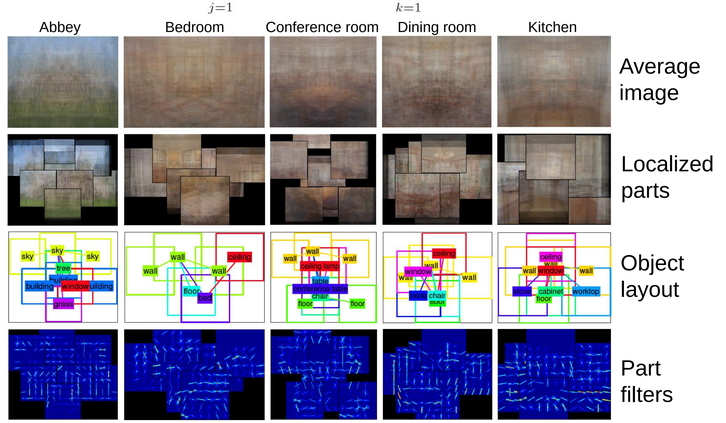
$$\min_{\mathcal{W},\mathcal{H},\xi} \quad \|\mathcal{W}\|_2^2+\lambda_1\sum_{i=1}^{n}\xi_i$$

GT layout | Predicted layout

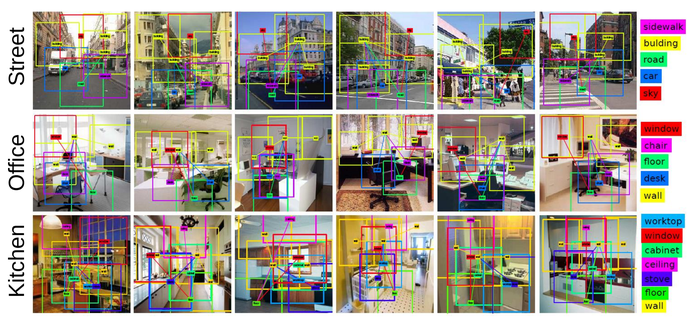
$$\mathcal{D}(x_i,\mathcal{H}_i,\mathcal{W})\geq\max_{\mathcal{H}^*}\mathcal{D}(x_i,\mathcal{H}_i^*,\mathcal{W})+\Delta(\mathcal{H}_i,\mathcal{H}_i^*)-\xi_i,\xi_i\geq 0\ ,\ \forall i$$

HOG | Quadratic distance transform

$$\mathcal{D}(x_i,\mathcal{H}_i,\mathcal{W}_c^a,\mathcal{W}_c^d,\mathcal{G}_c)=\sum_{j=1}^{p_c}(W_{c,j}^a\phi(x_i,\mathcal{H}_i,\mathcal{G}_c)+\sum_{k=1}^{p_c}W_{c,j,k}^d\psi(\mathcal{H}_{i,j},\mathcal{H}_{i,k},\mathcal{G}_c))$$



Samples of trained models

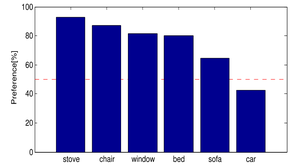


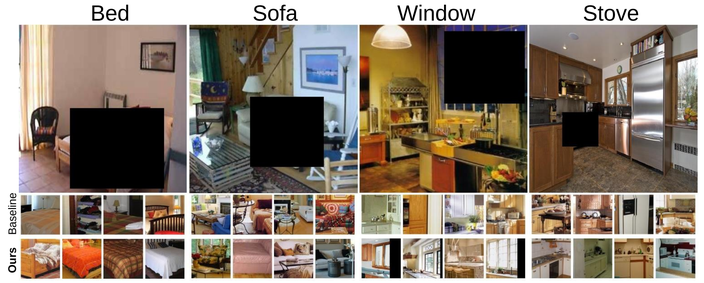
Samples of inferred scene layout

Experimental results

Black Box Test:

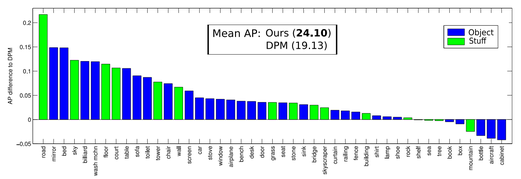
- Predicting object pose
- Forced choice human study
- Ours wins in 74.7% of cases

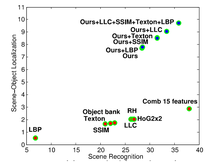




Object Detection:

- Simultaneous scene recognition and object detection
- Scene structures improve localizing objects
- Intersection over union of object & ground truth





Scene Recognition:

- Use scene structures to generate features
  - Best scoring structures per scene category
  - Normalized locations of objects
  - Relative locations of objects

Method	Accuracy	Method	Accuracy
LBP	6.84	Ours	28.45
SSIM	21.06	Comb 15 features [22]	38
Texton	22.04	Ours+LBP	28.59
Object bank [8]	22.93	Ours+Texton	31.57
LLC	26.23	Ours+SSIM	31.58
RH [4]	26.9	Ours+LLC	33.45
HoG2x2 [22]	27.2	Ours+LLC+SSIM+Texton+LBP	35.95

SUN database (390 classes)

Method	Accuracy	Method	Accuracy
LBP	18.12	GIST-color+SP+DPM [13]	43.1
HOG [13]	22.8	LPR [15]	44.84
ROI-GIST [14]	26.5	Ours	45.91
GIST-color [13]	29.7	Ours+LBP	47.64
DPM [13]	30.4	Ours+Texton	49.36
SSIM	33.45	Ours+LLC	49.38
Spatial Pyramid (SP)[7]	34.4	MLD Patches+GIST+SP+DPM [17]	49.4
Texton	35.98	Ours+SSIM	49.62
Object bank [8]	37.6	Ours+LLC+SSIM+Texton+LBP	52.41
LLC	37.53	BoP+IFV [6]	63.10
MLD Patches [17]	38.1	Midlevel elements+IFV [2]	66.87

MIT-indoor (67 classes)