Estimating the Aspect Layout of Object Categories

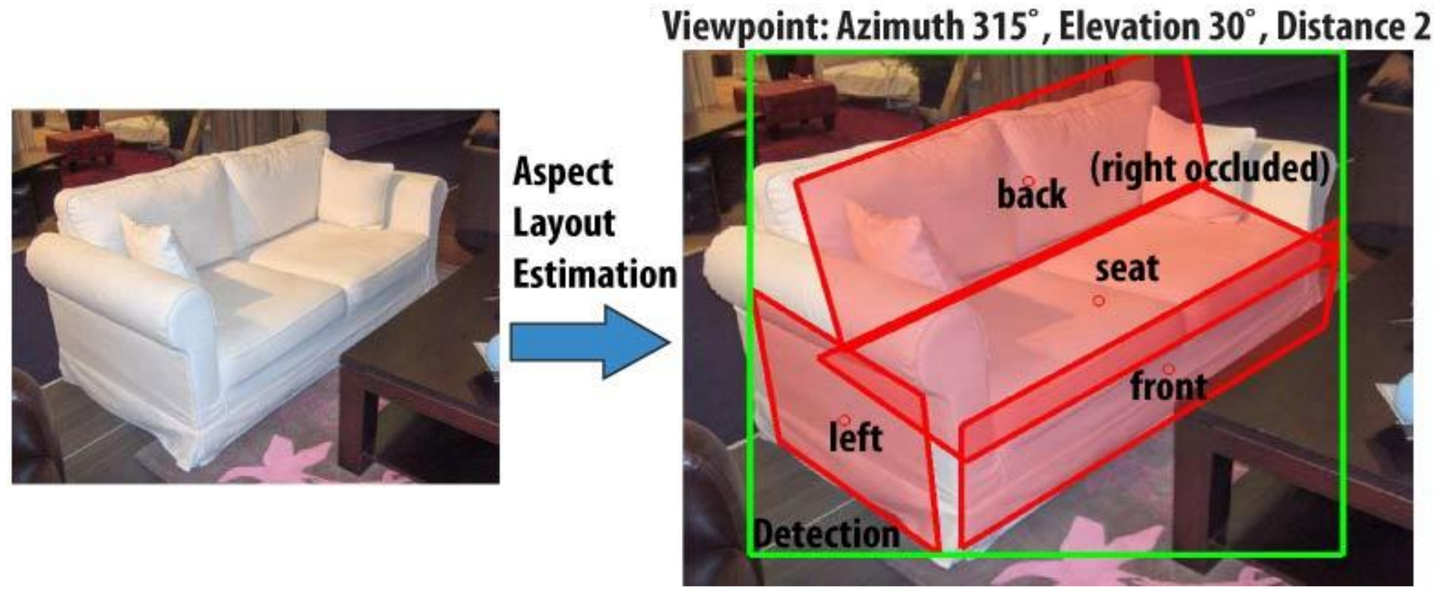
Yu Xiang and Silvio Savarese
Electrical and Computer Engineering, University of Michigan at Ann Arbor
{yuxiang, silvio}@eecs.umich.edu

VISION LAB

Introduction

Goal

Detect objects, identify objects' 3D poses, and estimate objects' 3D layout from a single image



Motivation

- Beyond 2D bounding boxes: provide richer 3D characterization of detected objects
- Relevant to applications such as robotics, autonomous navigation and object manipulation

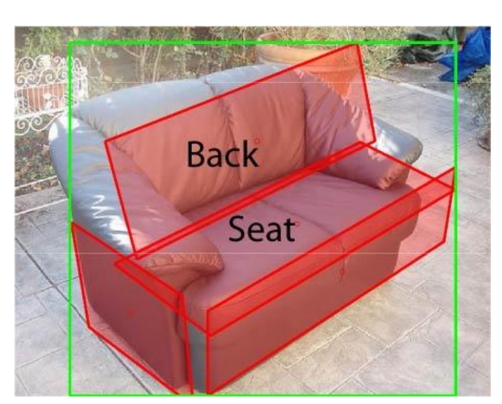
Contributions

- Joint object detection, pose estimation and aspect layout estimation
- Training by view-invariant part templates; inject rectification process into inference
- Obtain significant improvement in viewpoint accuracy over state-of-the-art on public datasets

Aspect Part

Definition

A portion of the object whose entire 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., occluded).



Related Concepts

Aspect graph; object affordance; functional part; geometrical attributes of objects; object-human interaction

Acknowledgements

We acknowledge the support of ARO grant W911NF-09-1-0310, NSF CPS grant #0931474 and a KLA-Tencor Fellowship.

Aspect Layout Model

Input: single 2D image I

Output: object label for a category $Y \in \{+1, -1\}$

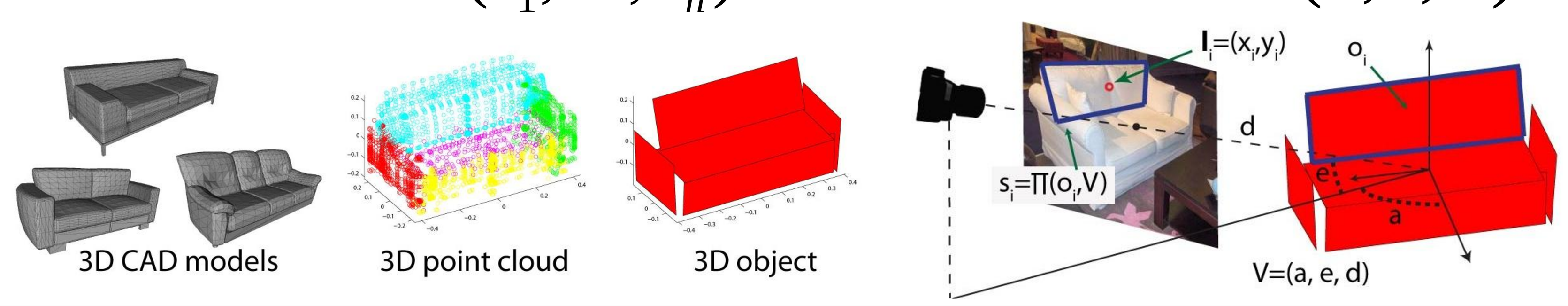
part configuration in 2D $C = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$, $\mathbf{c}_i = (x_i, y_i, s_i)$

part center coordinates x_i and y_i, part shape in 2D s_i

Posterior distribution:

$$P(Y, C \mid I) = P(Y, L, O, V \mid I),\ L = (\mathbf{l}_1, \ldots, \mathbf{l}_n),\ \mathbf{l}_i = (x_i, y_i)$$

3D object $O = (o_1, \ldots, o_n)$ Viewpoint $V = (a, e, d)$

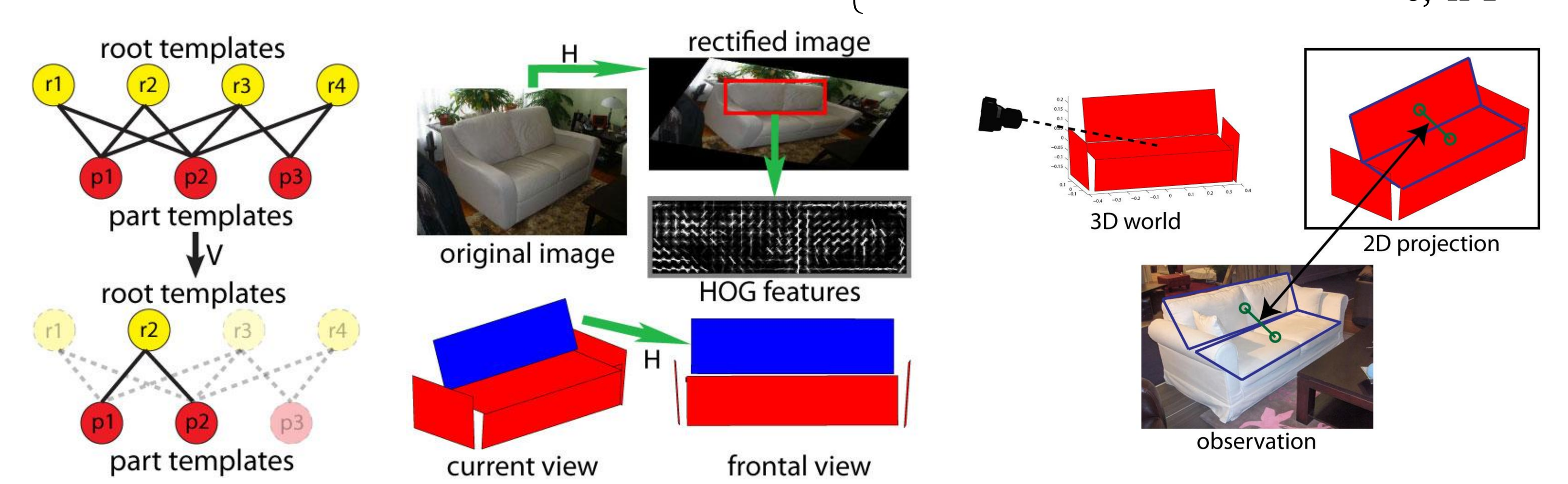


Modeling

Conditional Random Field $P(Y, L, O, V \mid I) \propto E(Y, L, O, V, I)$

Energy function

$$E(Y, L, O, V, I) = \begin{cases} \sum_i V_1(\mathbf{l}_i, O, V, I) + \sum_{(i,j)} V_2(\mathbf{l}_i, \mathbf{l}_j, O, V), & \text{if } Y = +1 \\ 0, & \text{if } Y = -1 \end{cases}$$



Unary potential

$$V_1(\mathbf{l}_i, O, V, I) = \begin{cases} \mathbf{w}_i^T \phi(\mathbf{l}_i, O, V, I), & \text{if unoccluded} \\ \alpha_i, & \text{if occluded} \end{cases}$$

Pairwise potential

$$V_2(\mathbf{l}_i, \mathbf{l}_j, O, V) = -w_x(x_i - x_j + d_{ij,O,V}\cos(\theta_{ij,O,V}))^2 - w_y(y_i - y_j + d_{ij,O,V}\sin(\theta_{ij,O,V}))^2$$

Maximal margin learning: structural SVM

Model inference: belief propagation for each O and V

Reference

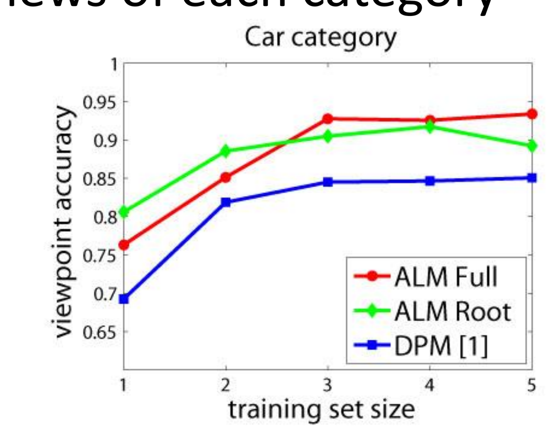
[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. TPAMI, 2010.
[2] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In ECCV, 2010.
[3] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In ICCV, 2007.
[4] N. Payet and S. Todorovic. From contours to 3d object detection and pose estimation. In ICCV, 2011.
[5] J. Liebelt and C. Schmid. Multi-view object class detection with a 3D geometric model. In CVPR, 2010.
[6] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and pose estimation. In ICCV, 2011.
[7] M. Stark, M. Goesele, and B. Schiele. Back to the future: Learning shape models from 3d cad data. In BMVC, 2010.
[8] H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a dense multiview representation for detection, viewpoint classification and synthesis of object categories. In ICCV, 2009.
[9] M. Arie-Nachimson and R. Basri. Constructing implicit 3d shape models for pose estimation. In ICCV, 2009.
[10] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In CVPR, 2009.

Experiments

1. 3DObject dataset [3]

Train on 5 instances, test on 5 instances for 8 views of each category

Method	ALM Full	ALM Root	DPM [1]	[2]	[3]
Viewpoint	**80.7**	77.7	67.9	74.2	57.2
Detection	81.8	81.3	**83.9**	n/a	n/a



Category	Bicycle			Car						
Method	ALM	[4]	[5]	ALM	[4]	[6]	[7]	[5]	[8]	[9]
Viewpoint	**91.4**	80.8	75.0	**93.4**	85.4	85.3	81	70	67	48.5
Detection	**93.0**	n/a	n/a	98.4	n/a	**99.2**	89.9	76.7	55.3	n/a

2. EPFL Car dataset [10]

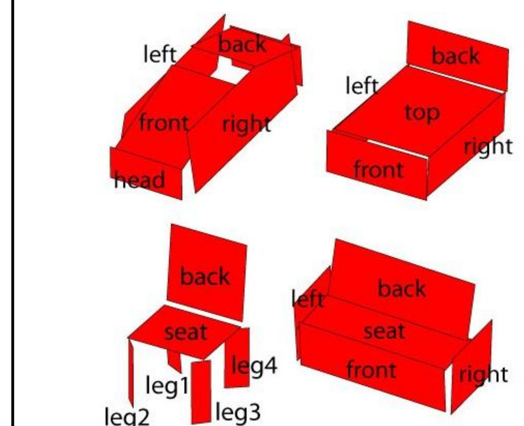
Train on 10 instance
Test on 10 instances for 16 views

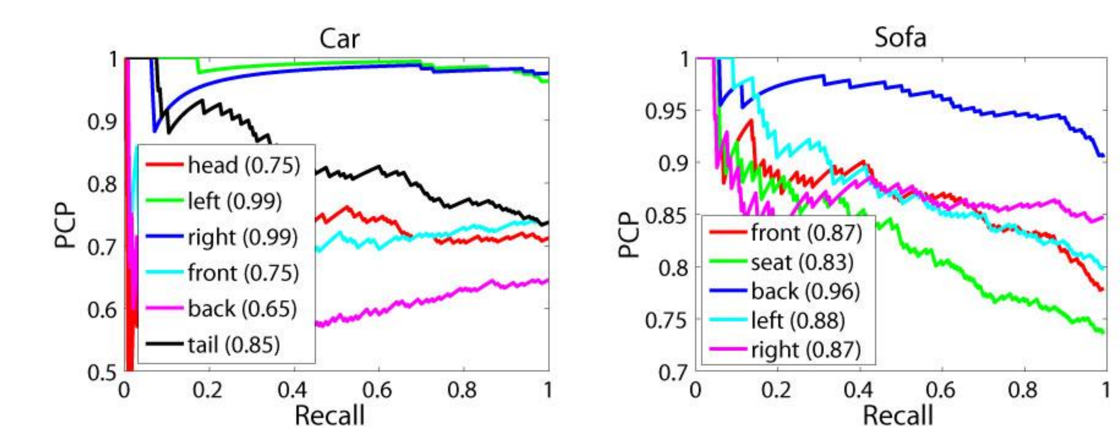
Method	ALM Full	ALM Root	DPM [1]	[10]
Viewpoint	**64.8**	58.1	56.6	41.6
Detection	96.4	97.5	**98.1**	85.4

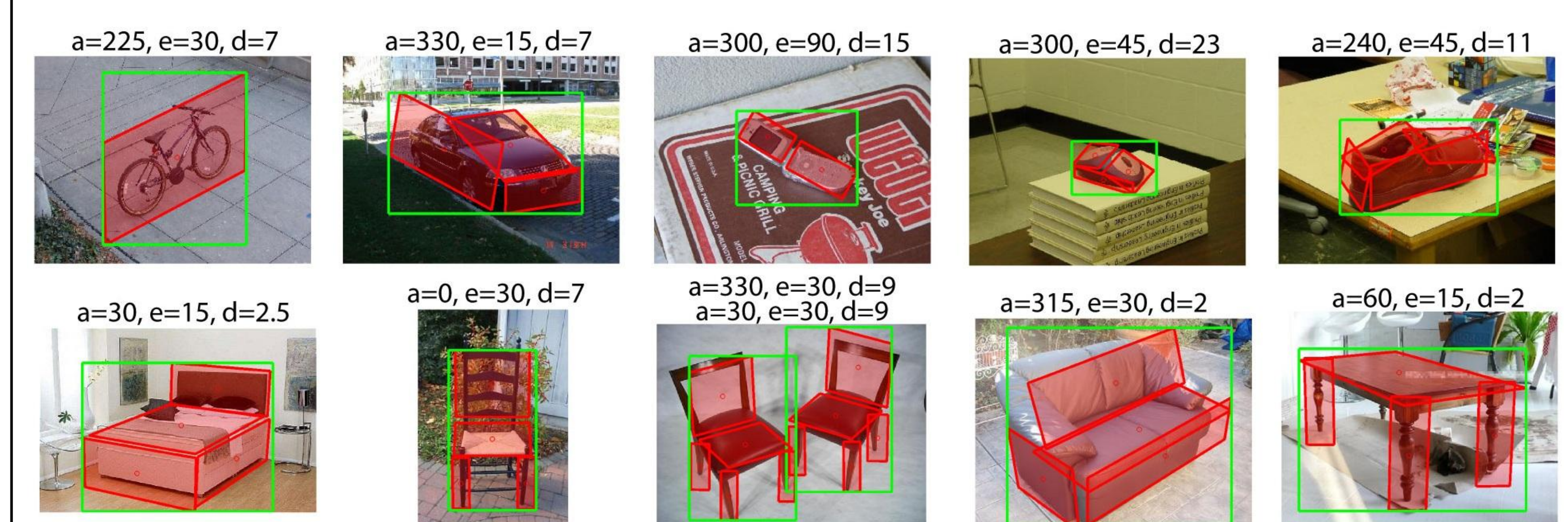
3. New ImageNet dataset

Train on half of the instances
Test on half for 7 views

Category	Bed	Chair	Sofa	Table	Mean
DPM [1]	56.2	41.2	44.0	**56.4**	49.5
ALM Root	37.5	23.4	39.6	35.4	34.0
ALM Full	**62.7**	**73.1**	**65.0**	52.6	**63.4**







Conclusion

- Presented a new model for joint object detection, pose estimation and aspect part localization
- Able to handle large number of viewpoints, localize parts with approximately correct shapes, and reason about self-occlusions
- Potentially useful for recognizing functional parts or estimating object affordances