Estimating the Aspect Layout of Object Categories

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

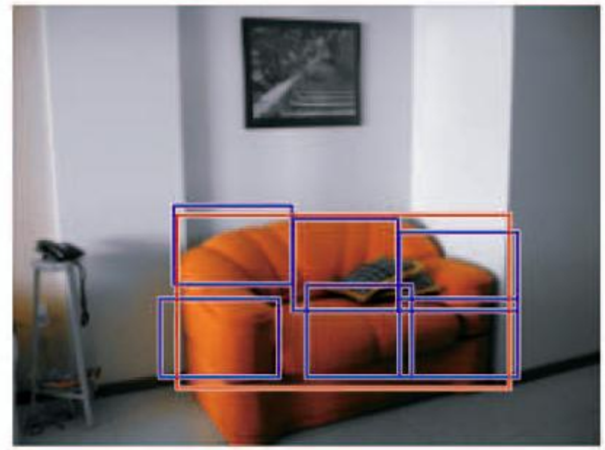
Traditional object recognition

- Uses 2D bounding boxes

Face

From Viola & Jones, 01

Rigid object



From Felzenszwalb et al., 10

Human

From Barinova et al., 12

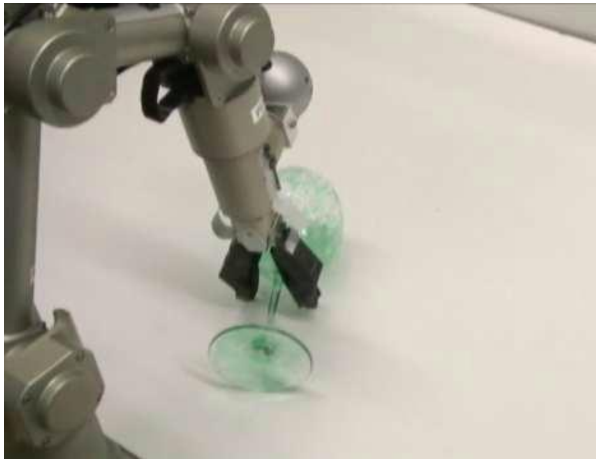
Body part

From Ramanan & Sminchisescu, 06

Beyond 2D bounding boxes

- Model the 3D properties of objects
  - 3D pose
  - 3D part location

- More suitable for robotics, autonomous navigation and manipulation

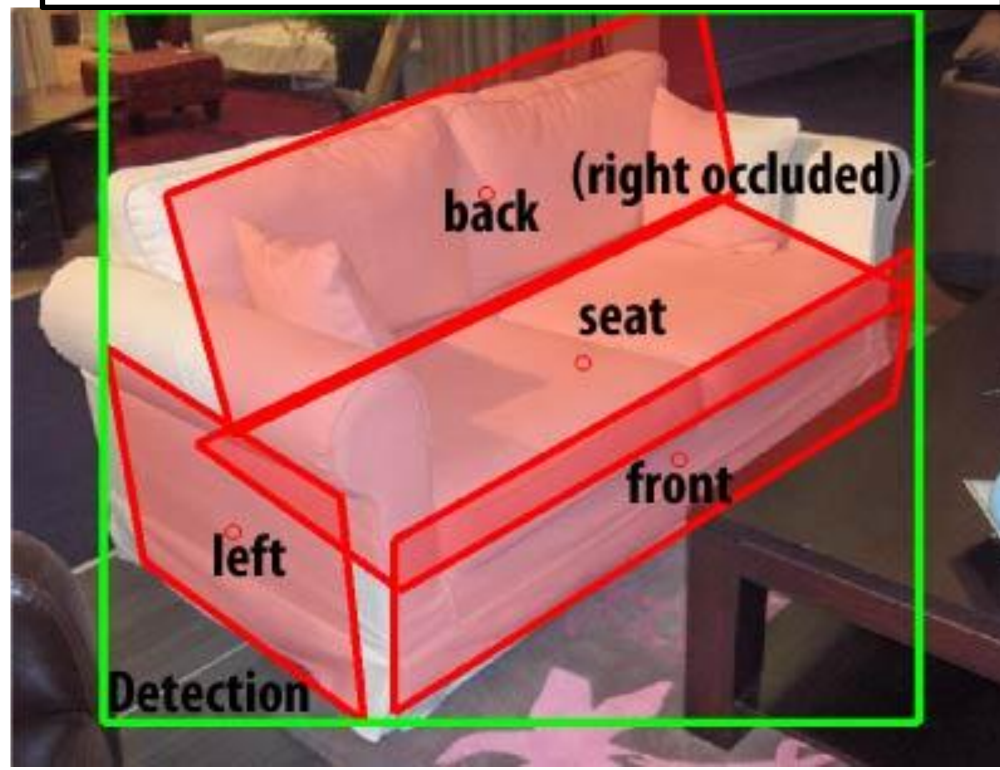


From Saxena et al., 08

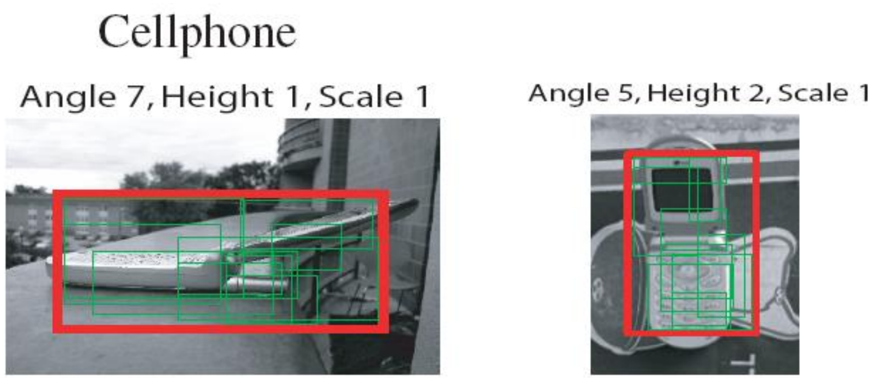
MICHIGAN VISION LAB

Our goals

Viewpoint: Azimuth 315˚, Elevation 30˚, Distance 2



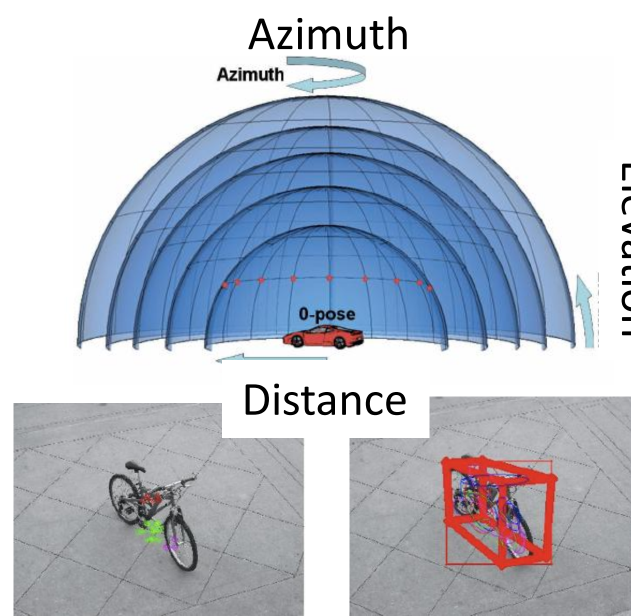
Related work: joint object detection and pose estimation



From Savarese & Fei-Fei ICCV'07

From Glasner et al. ICCV'11

- Savarese et al. 07, 08
- Ozuysal et al. 08
- Liebelt et al. 08, 10
- Xiao et al. 08
- Thomas et al. 08
- Sun et al. 09
- Su et al. 09
- Arie-Nachimson & Barsi 09
- Stark et al. 10
- Gu & Ren. 10
- Glasner et al. 11
- Payet & Todorovic 11
- Zia et al., 3DRR'11
- Pepik et al., CVPR'12
- Schels et al., CVPR'12
- Xiang and Savarese, CVPR'12



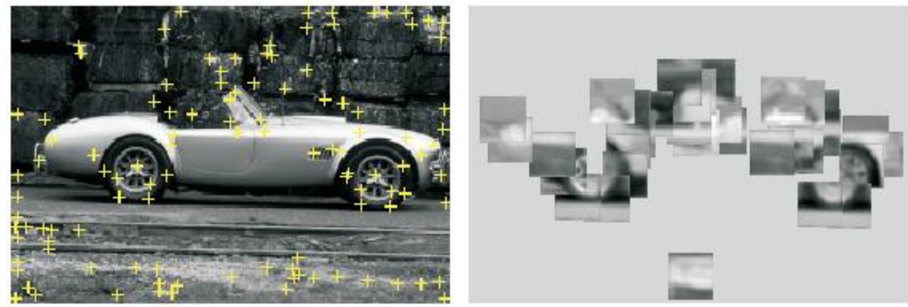
From Liebelt et al. 08, 10

Related work: 2D part-based model

Constellation Model

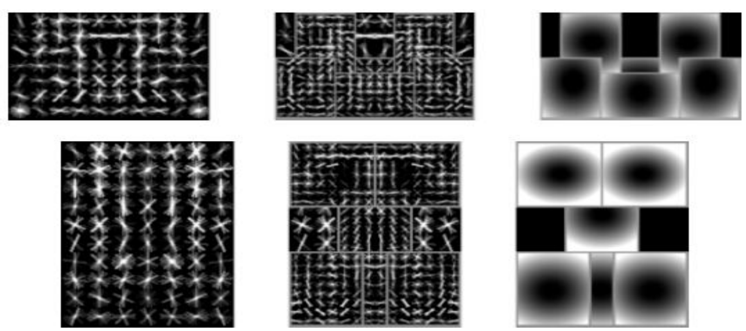
From Fergus et al. CVPR'03

Implicit Shape Model



From Leibe et al. ECCV'04 workshop

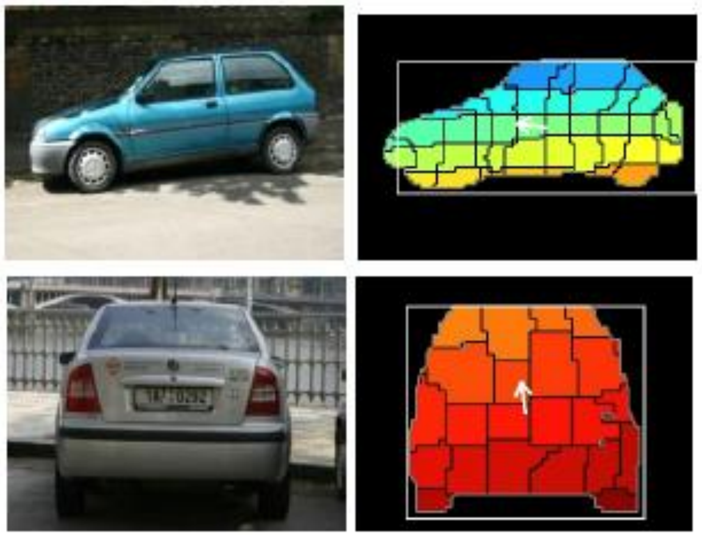
Deformable Part Model (DPM)



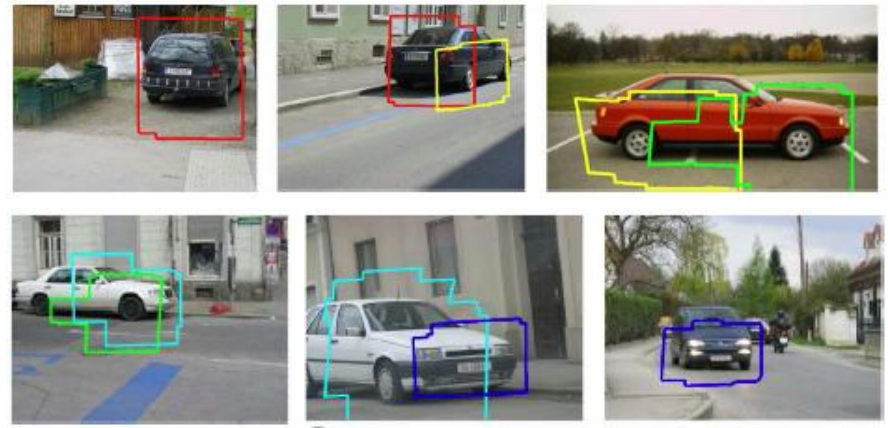
From Felzenszwalb et al. CVPR'08

MICHIGAN
VISION LAB

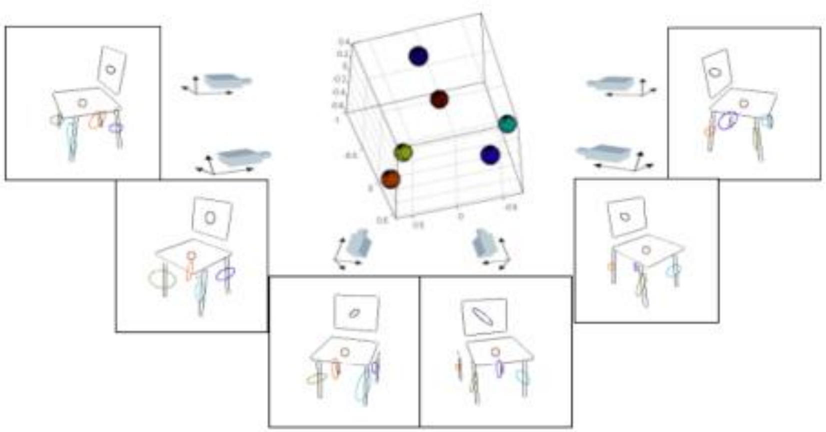
Related work: 3D part-based model



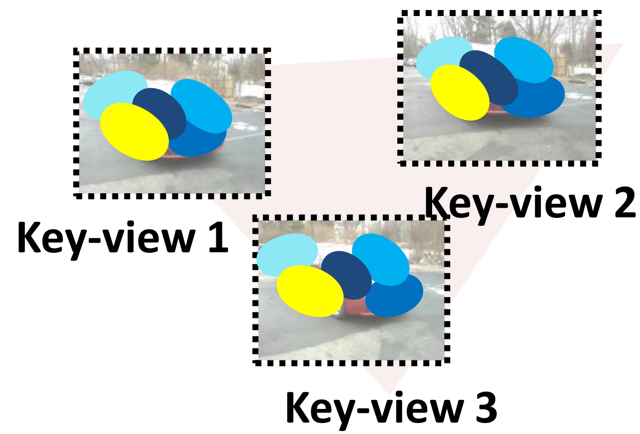
From Hoiem et. al., CVPR'07



From Kushal et. al., CVPR'07



From Chiu et. al., CVPR'07



From Sun et. al. ICCV'09

Our contributions

- Propose a 3D part based representation for object categories
- Introduce the concept of *aspect parts*
- Jointly solve object detection, pose estimation and aspect part localization
- Significantly improve pose estimation accuracy, evaluate rigid part localization

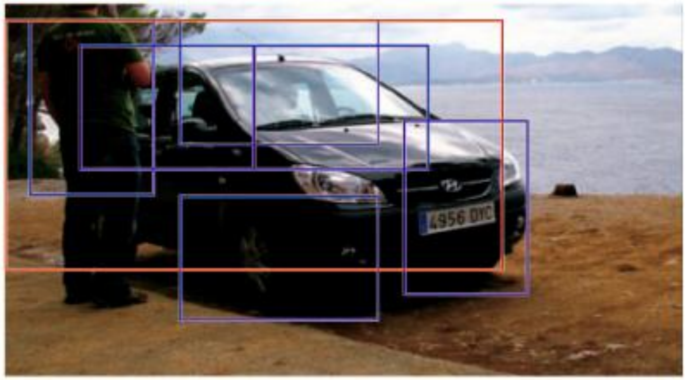
Yu Xiang and Silvio Savarese. Estimating the aspect layout of object categories. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

MICHIGAN VISION LAB

Aspect Part

- Parts are arbitrarily defined in previous work

From Fergus et al. CVPR'03

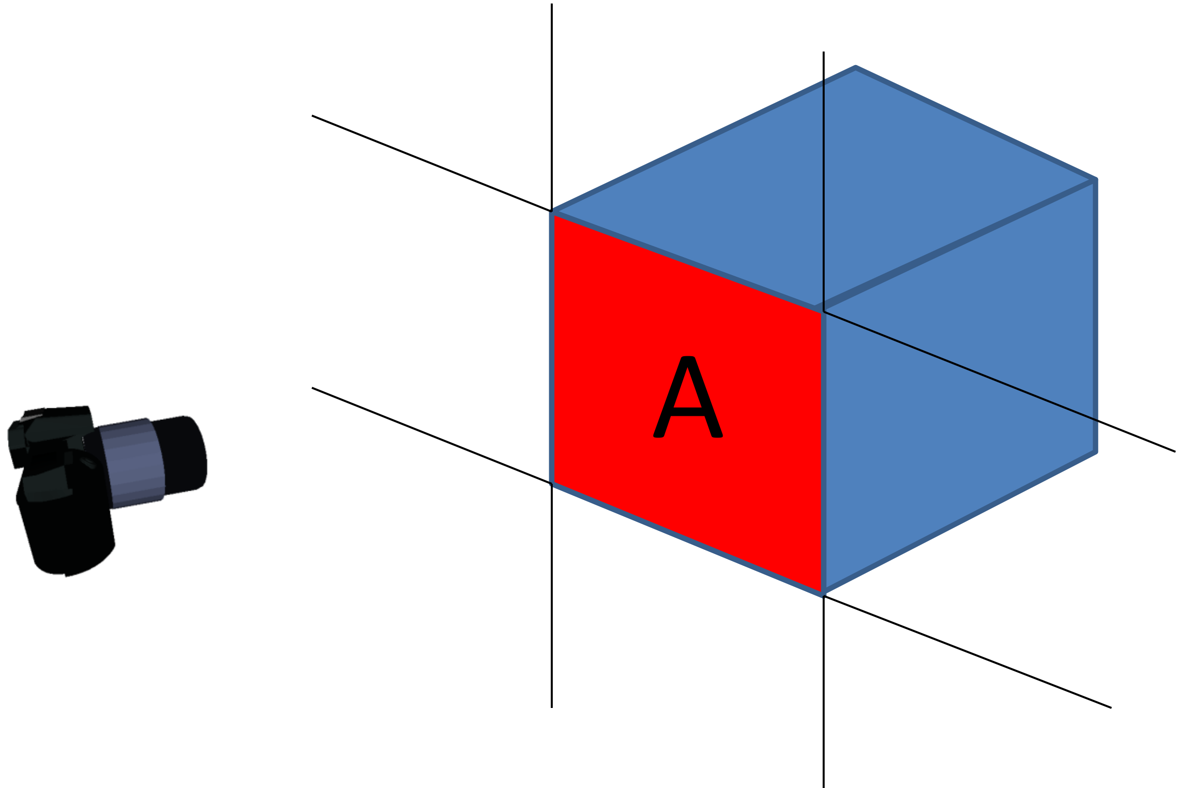


From Felzenszwalb et al., 2010.

- Introduce parts with geometrical and topological properties, called *aspect parts*

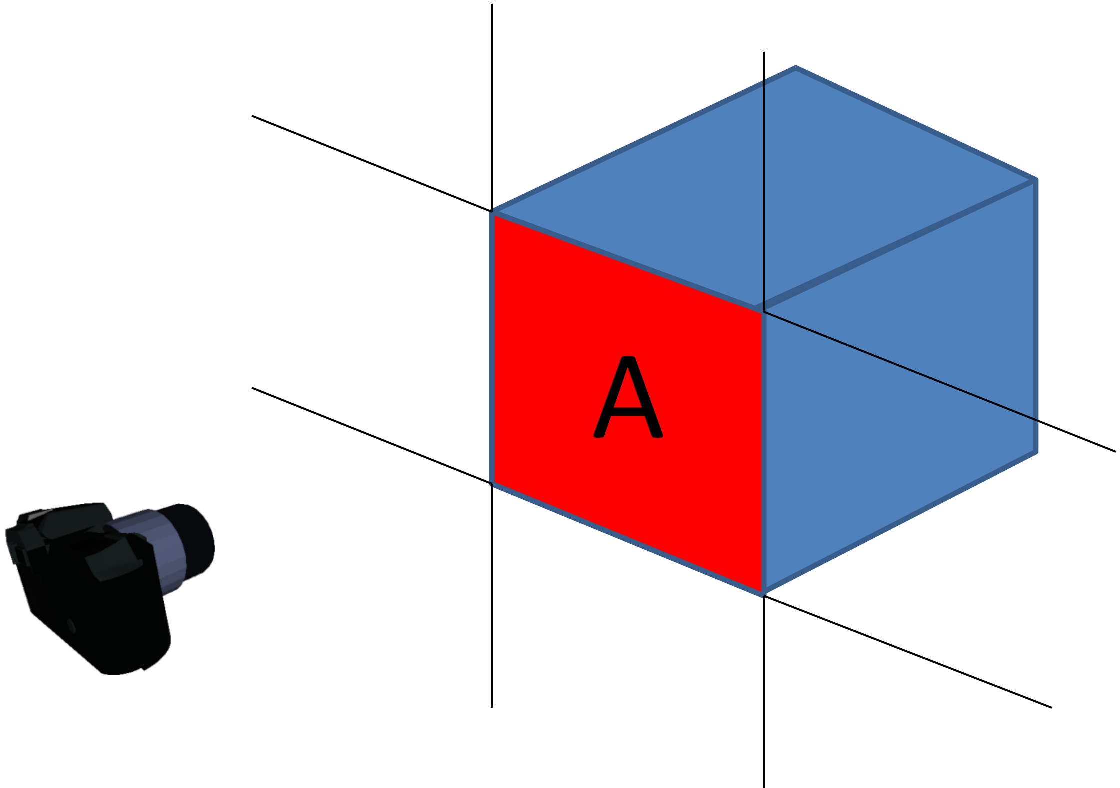
Aspect Part

- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)



Aspect Part

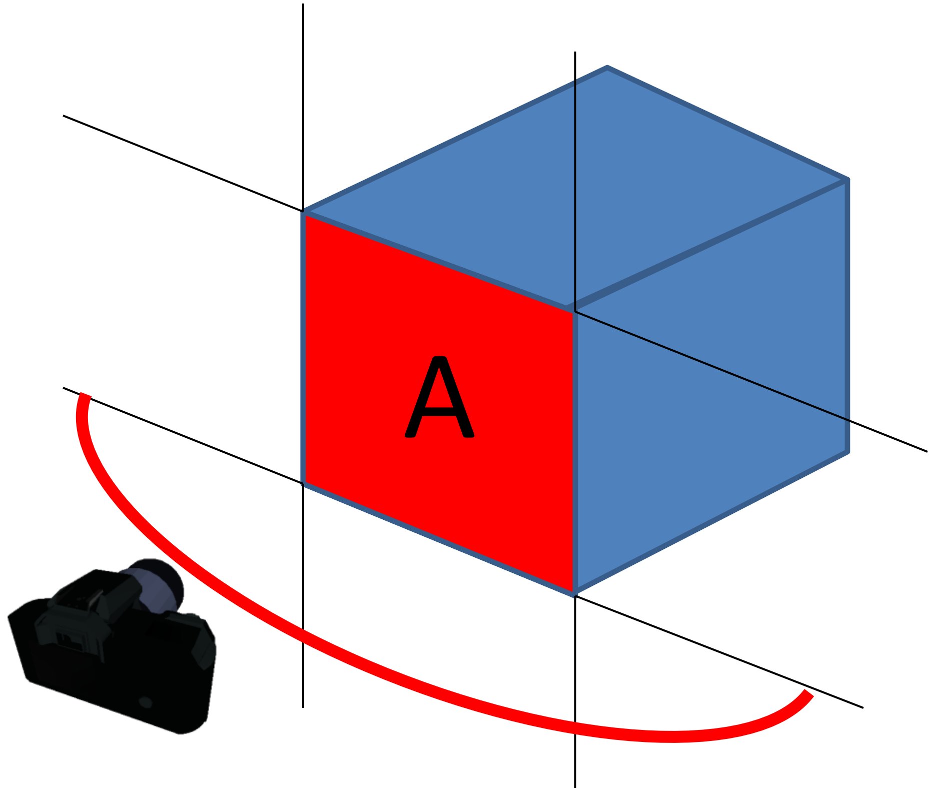
- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)



VISION LAB

Aspect Part

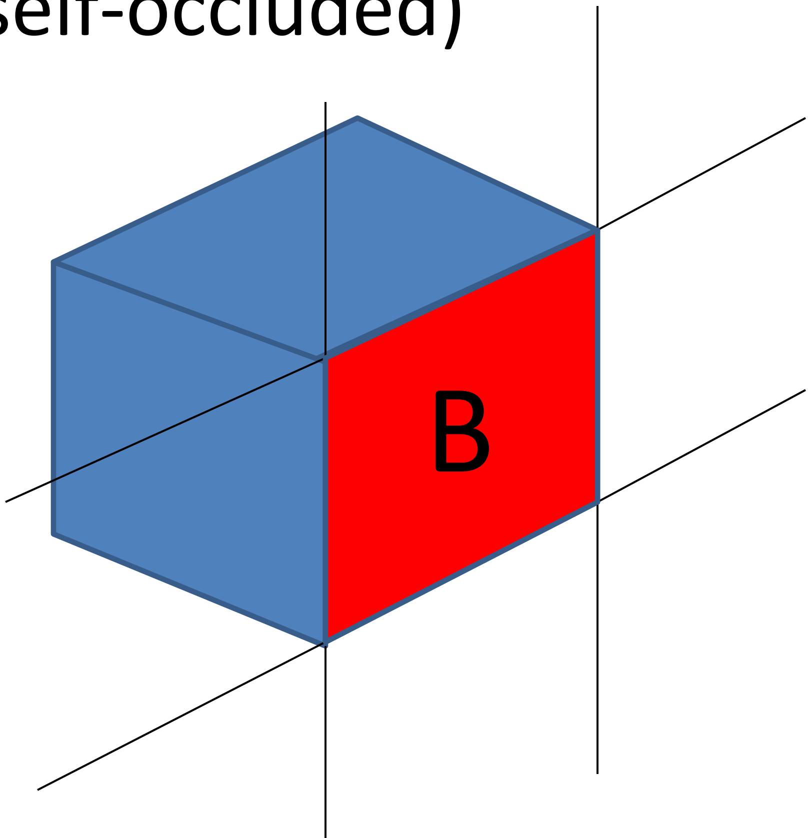
- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)



MICHIGAN
VISION LAB

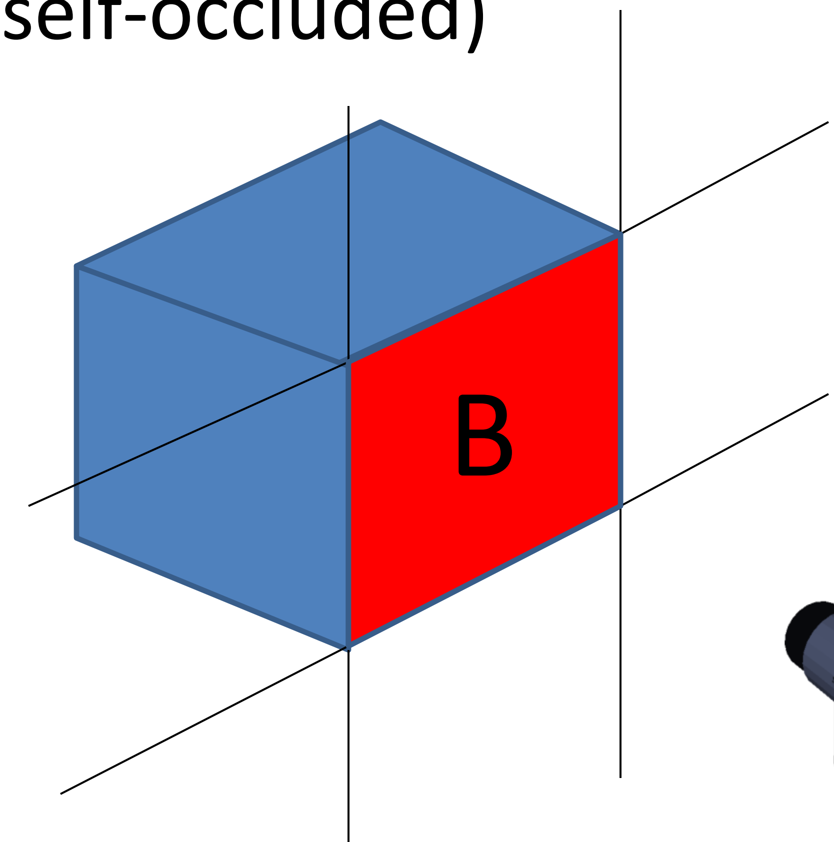
Aspect Part

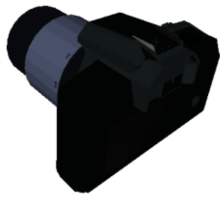
- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)



Aspect Part

- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)

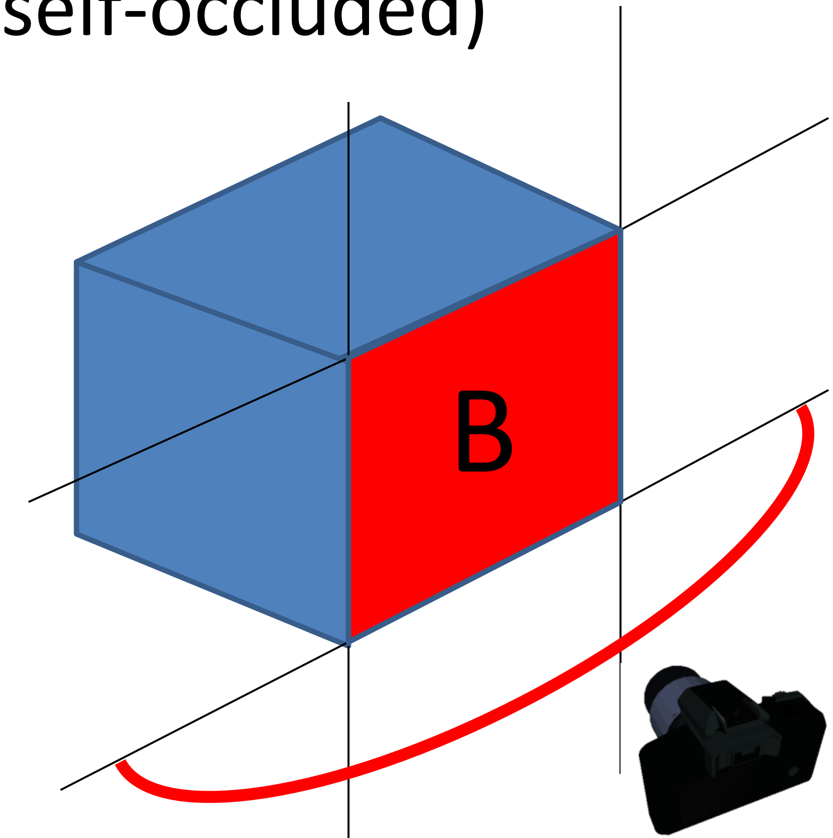




MICHIGAN VISION LAB

Aspect Part

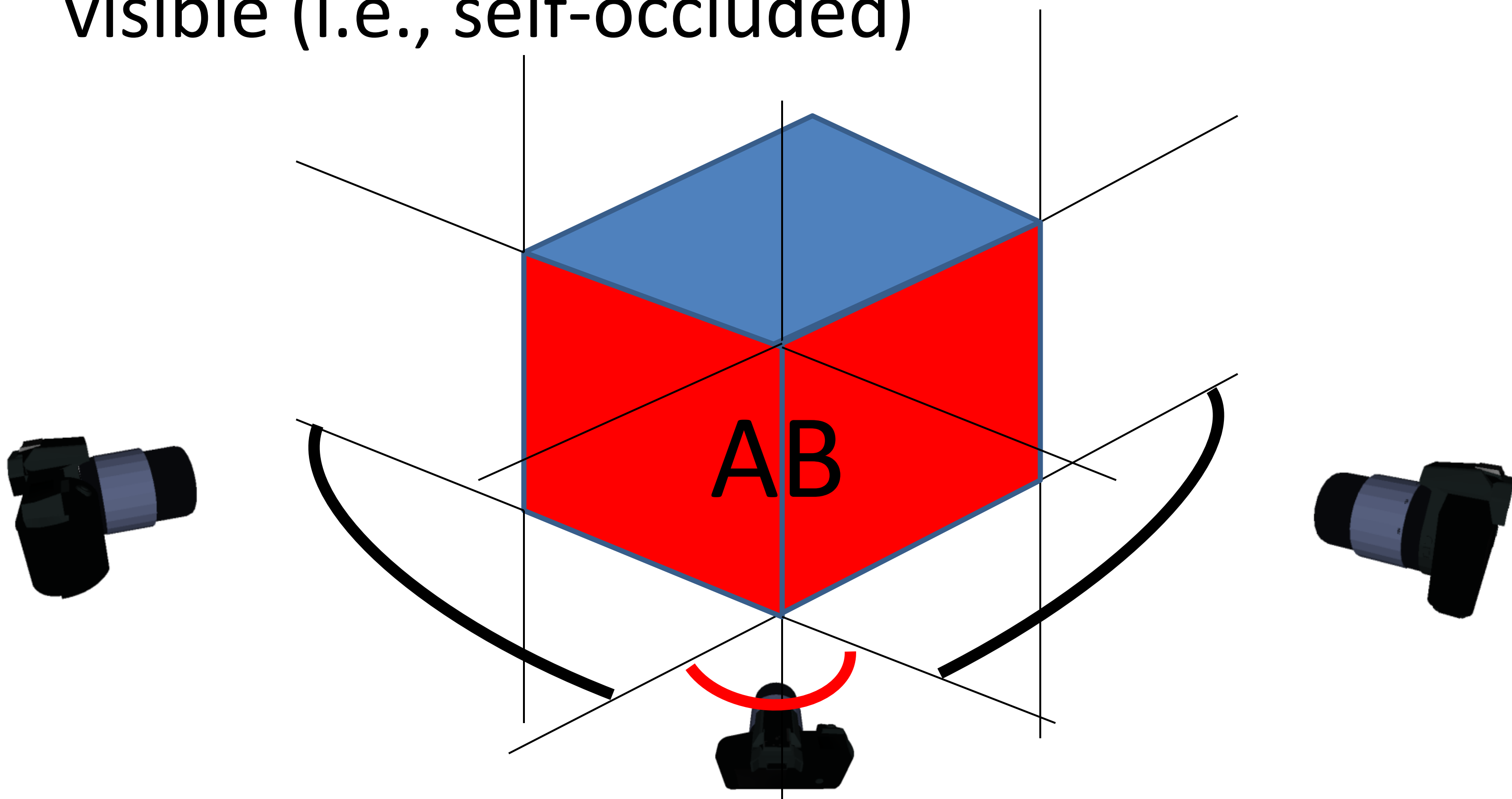
- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)



MICHIGAN VISION LAB

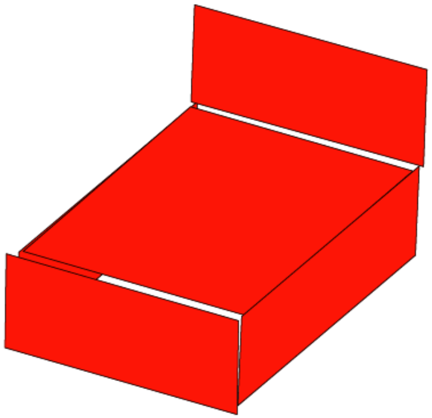
Aspect Part

- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)

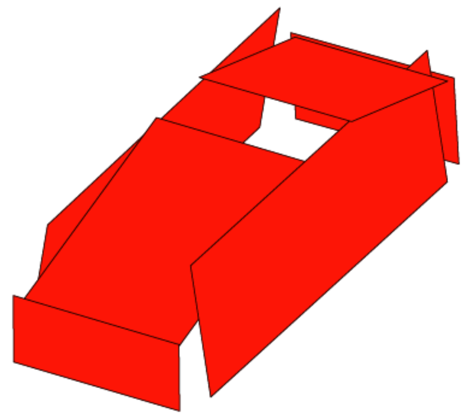


Aspect Part

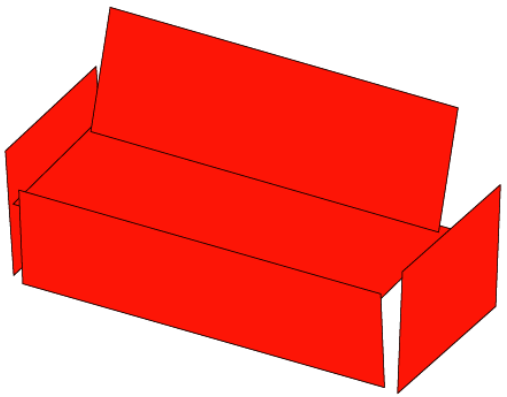
- Examples



Bed



Car



Sofa

Aspect Part

- Related to aspect graph [1]
- Related to discriminative aspect, Farhadi et al, 07

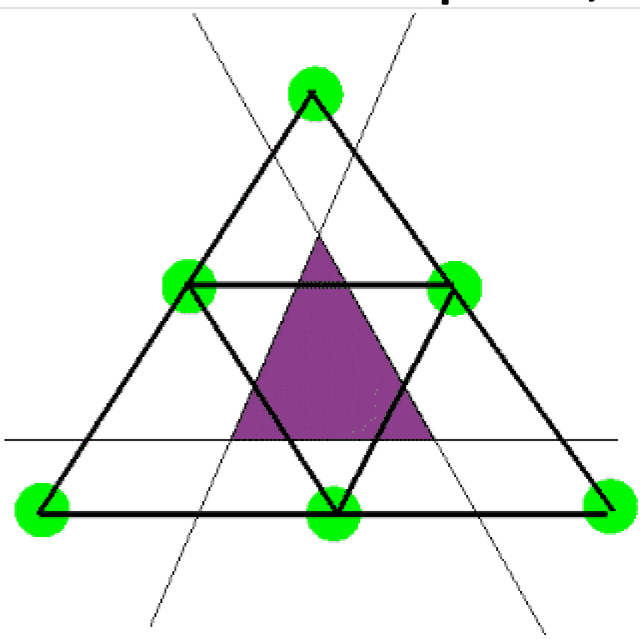


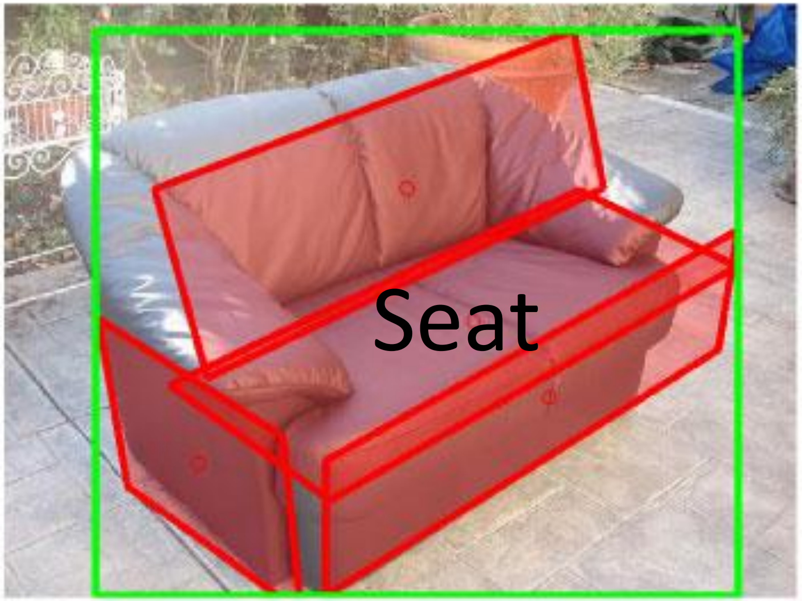
Figure from Barb Culter, MIT

[1] J. J. Koenderink and A. J. Doorn. The internal representation of solid shape with respect to vision. Biological Cybernetics, 1979.

MICHIGAN VISION LAB

Aspect Part

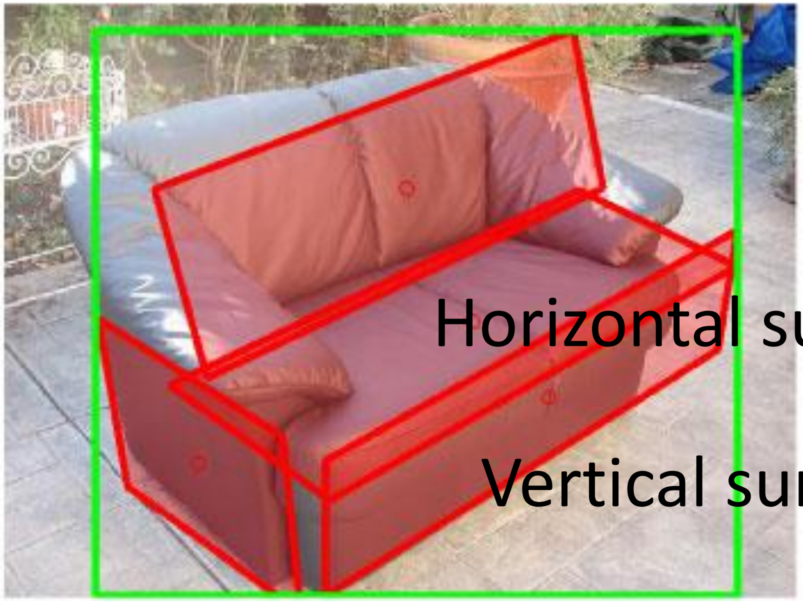
- Related to object affordance or functional part



VISION LAB

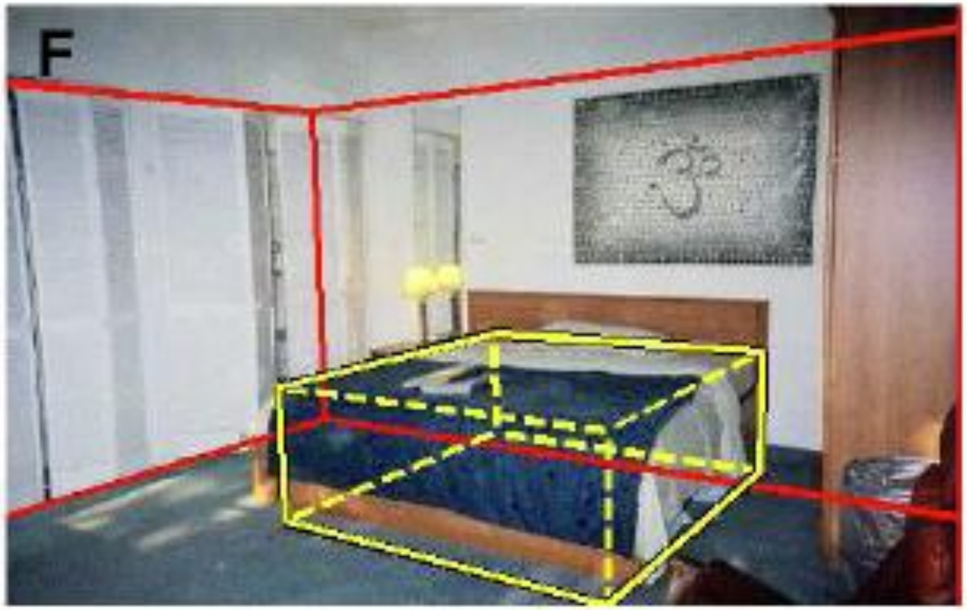
Aspect Part

- Related to geometrical attributes of object



Aspect Part

- Related to scene layout estimation

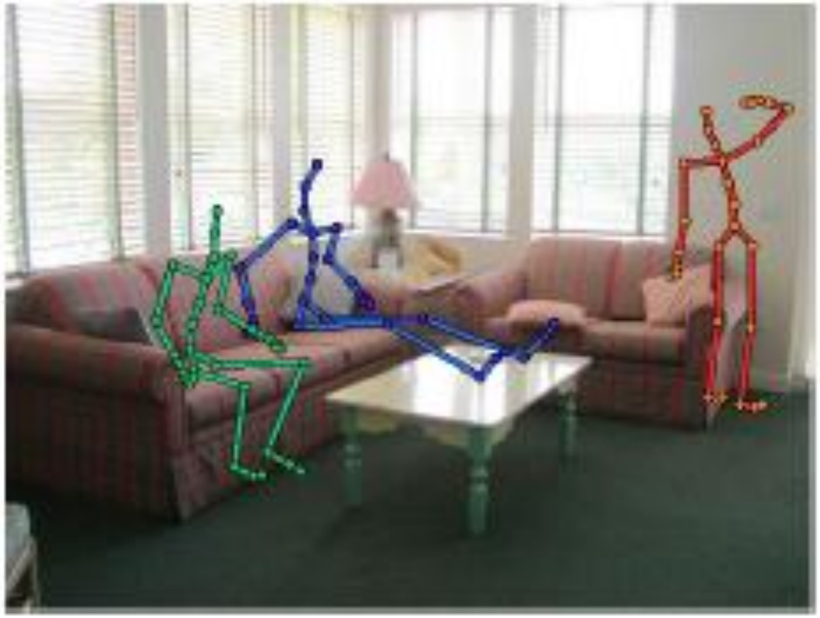


From Hedau, Hoiem & Forsyth, ECCV'10

MICHIGAN
VISION LAB

Aspect Part

- Enables the modeling of object-human interactions



From Gupta et al., CVPR'11

Outline

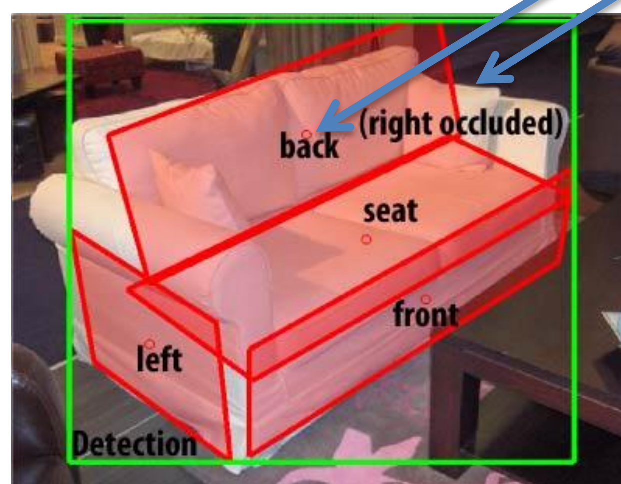
- Aspect layout model
- Maximal margin parameter estimation
- Model inference
- Experiments
- Conclusion

Input & output

- Input
  - 2D image I
- Output
  - Object label $Y \in \{+1, -1\}$
  - Part configuration in 2D $C = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$ $\mathbf{c}_i = (x_i, y_i, s_i)$

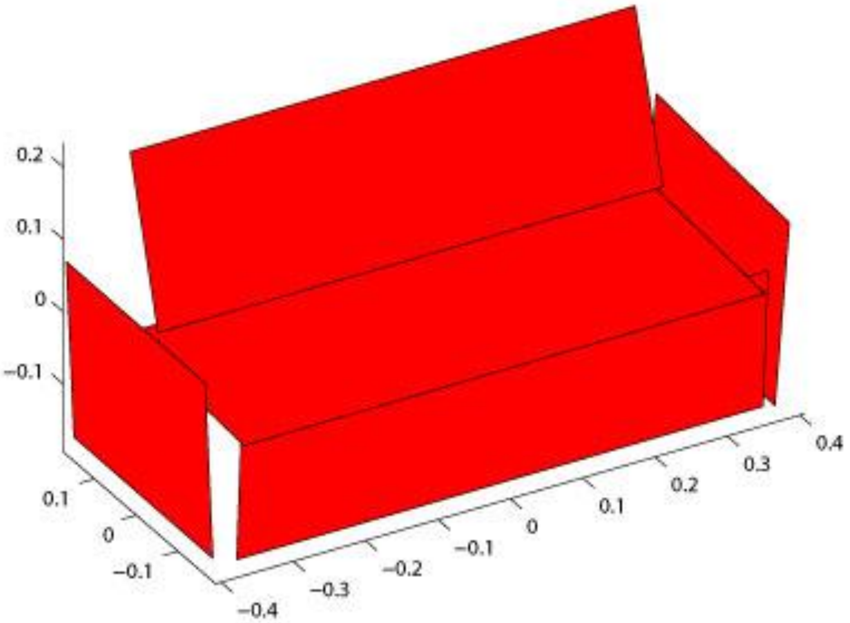
2D part center coordinates

2D part shape



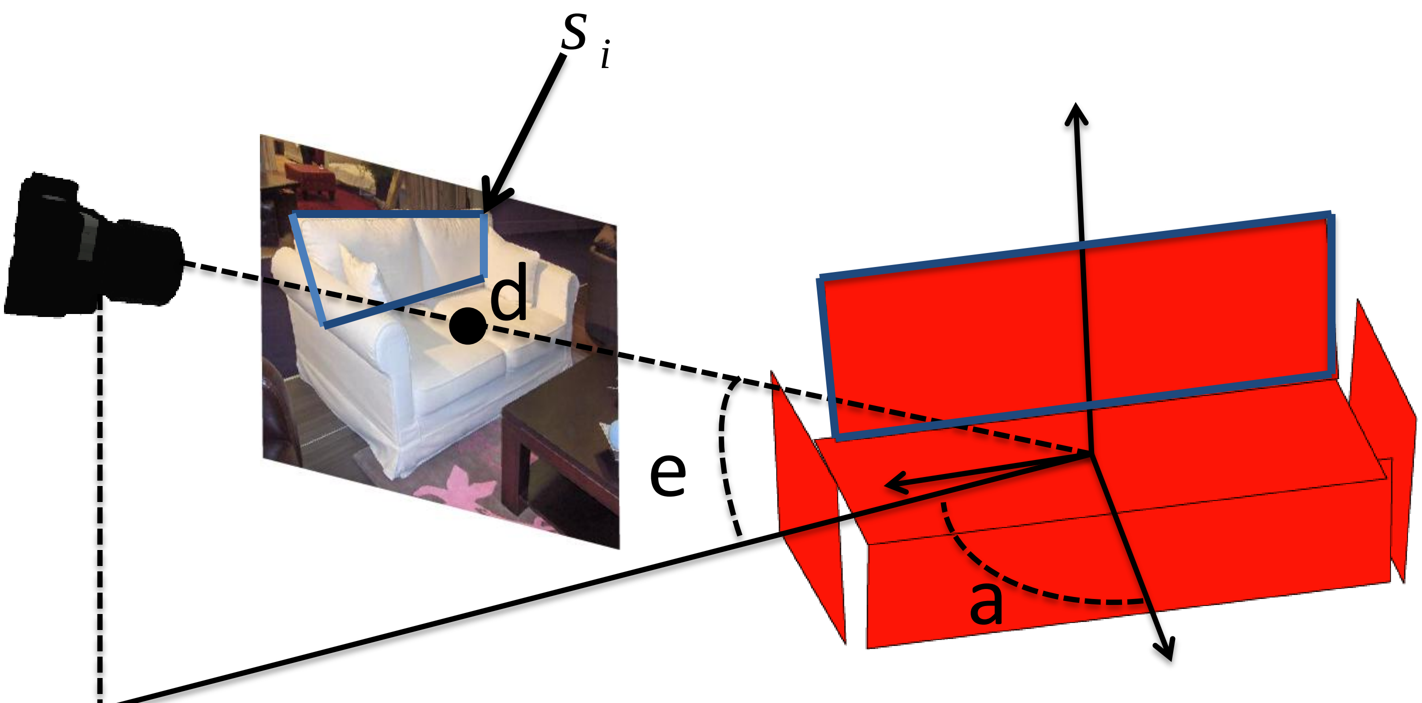
Aspect Layout Model

- 3D Object $O = (o_1, \ldots, o_n)$



Aspect Layout Model

- Viewpoint representation V=(a,e,d)
- 2D part shape from 3D



Azimuth, elevation and distance

Aspect Layout Model

- Model the posterior distribution

$$P(Y, C \mid I) \quad C = (\mathbf{c}_1, \ldots, \mathbf{c}_n), \mathbf{c}_i = (x_i, y_i, s_i)$$

$$= P(Y, L, O, V \mid I)$$

object label

2D part center coordinates

3D object

viewpoint

$$L = (\mathbf{l}_1, \ldots, \mathbf{l}_n),\ \mathbf{l}_i = (x_i, y_i)$$

Aspect Layout Model

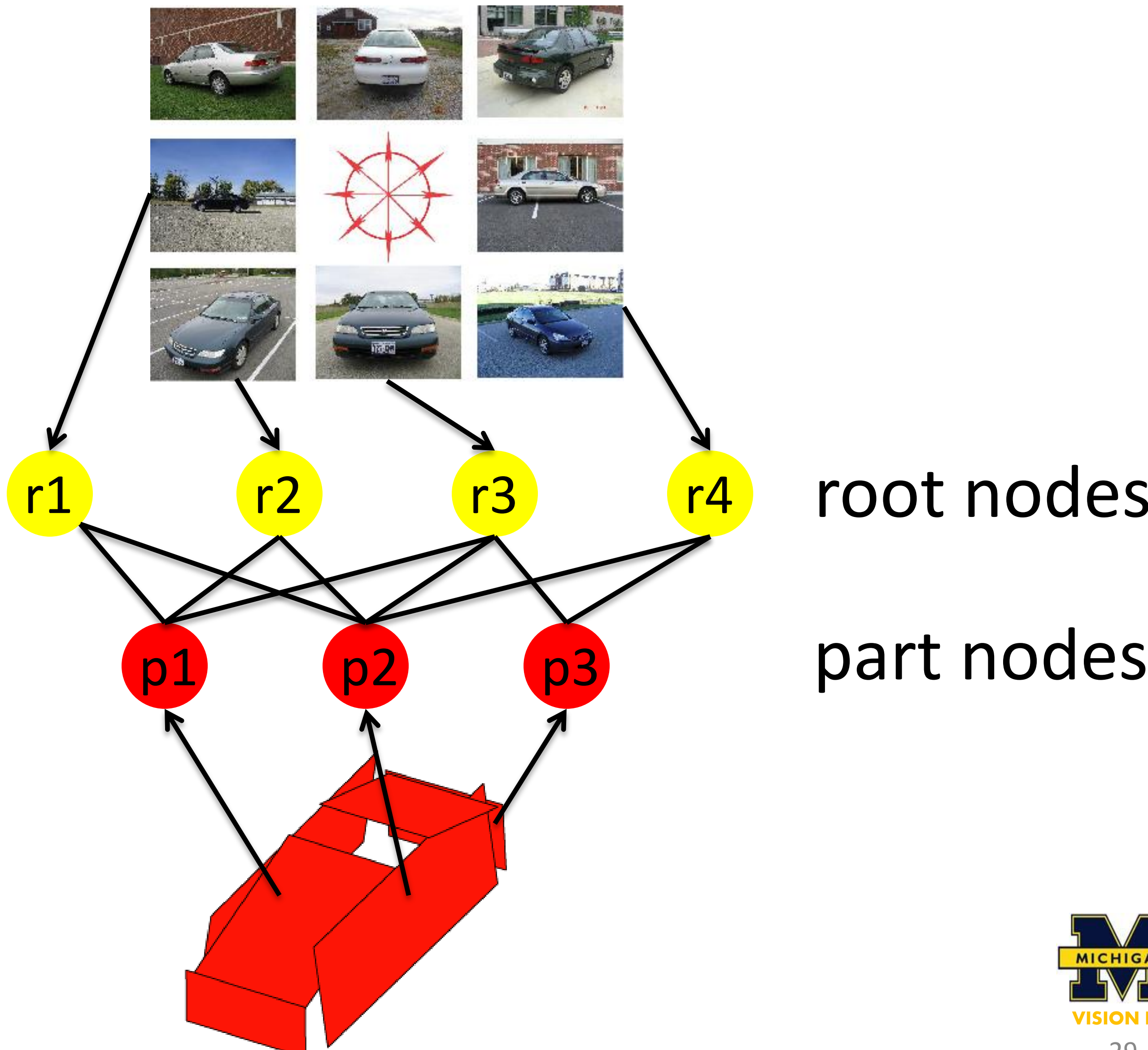
- Conditional Random Field (CRF) [1]

$$P(Y, L, O, V \mid I) \propto \exp(E(Y, L, O, V, I))$$

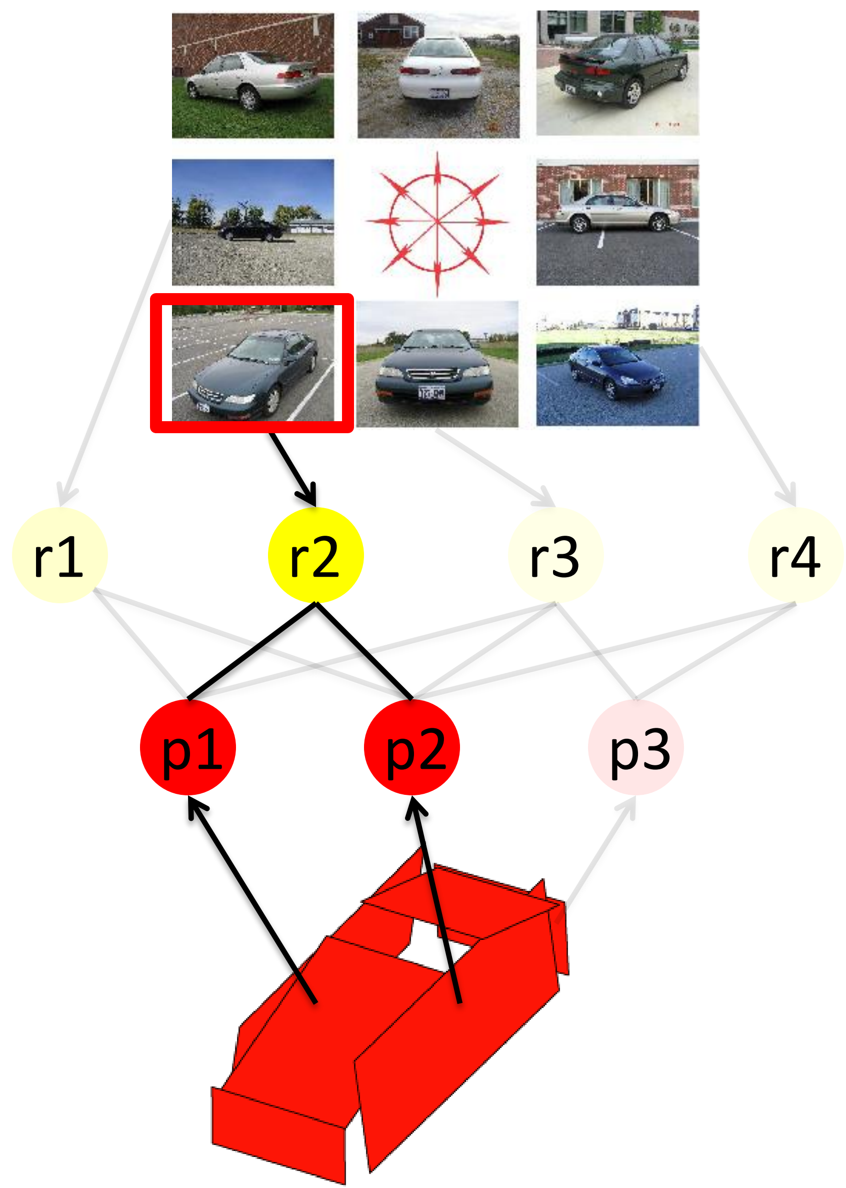
- Graph structure of the CRF

[1] J. Lafferty, A. McCallum and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML, 2001.

Aspect Layout Model



Aspect Layout Model

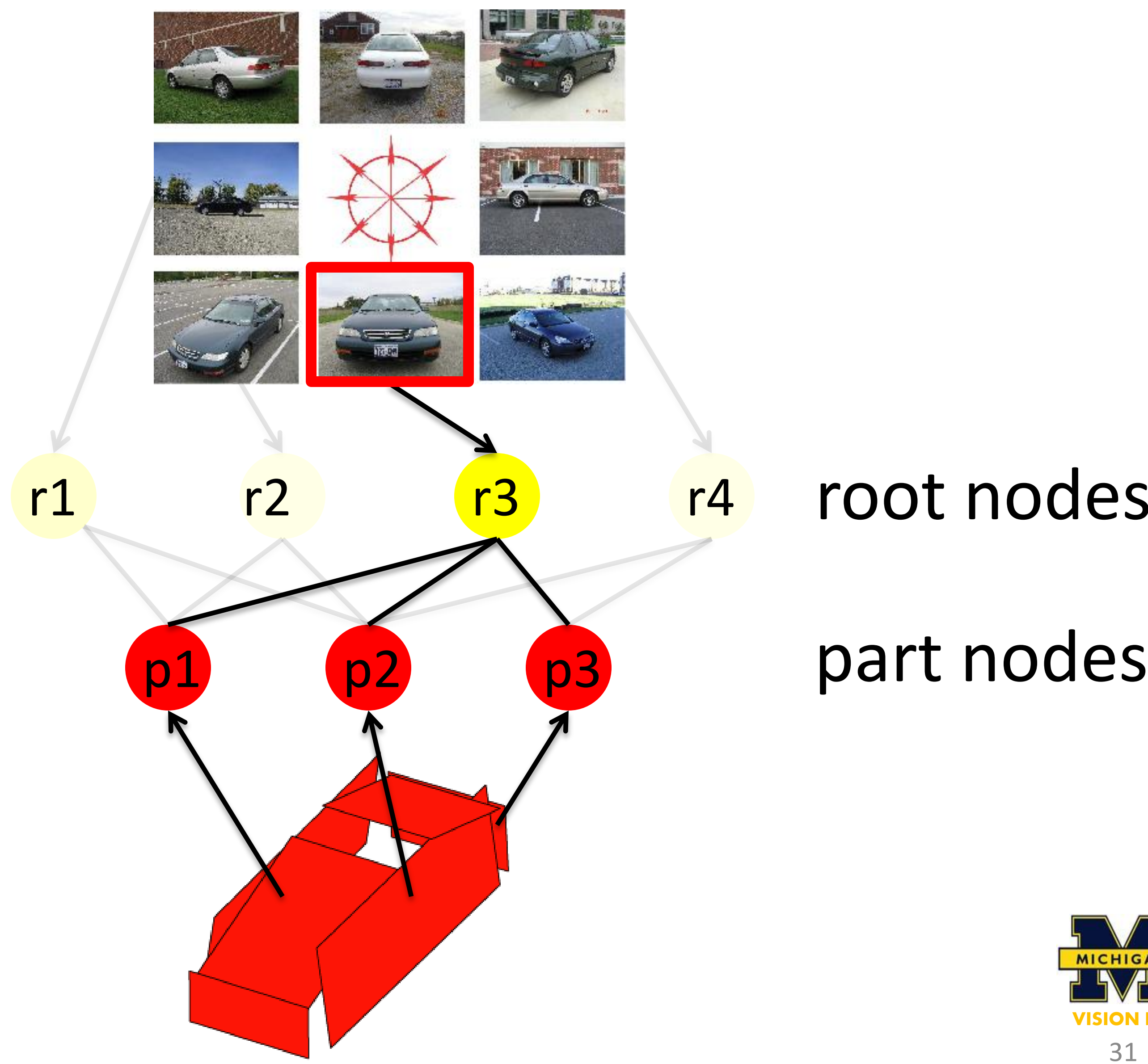


root nodes

part nodes

VISION LAB

Aspect Layout Model



Aspect Layout Model

- 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

pairwise potential

MICHIGAN VISION LAB

Aspect Layout Model

- Viewpoint invariant unary potential
  - Models part appearances

part template

feature vector

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

self-occlusion weight

MICHIGAN
VISION LAB

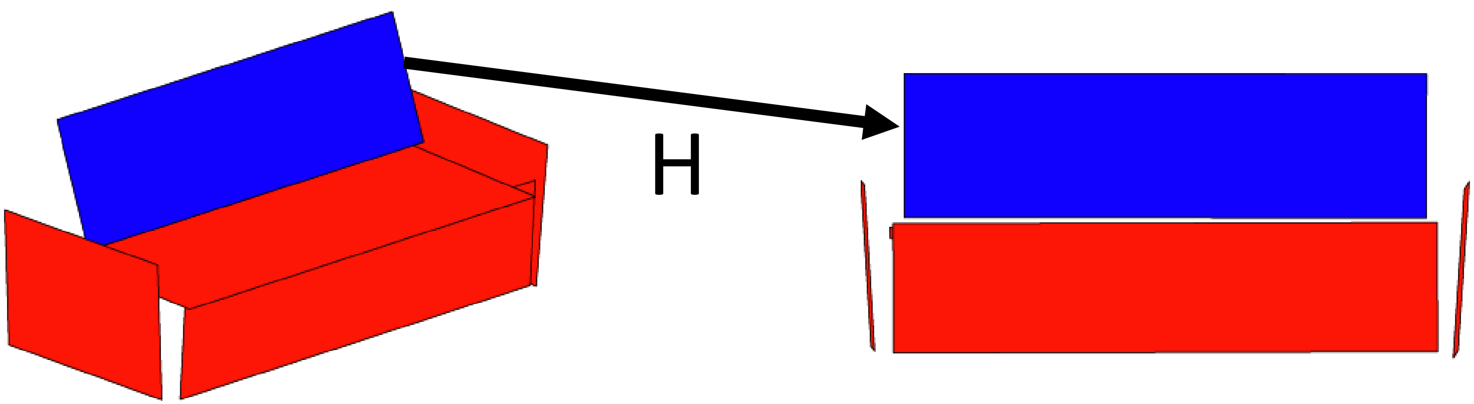
Aspect Layout Model

- Rectified HOG features

original image

rectified image

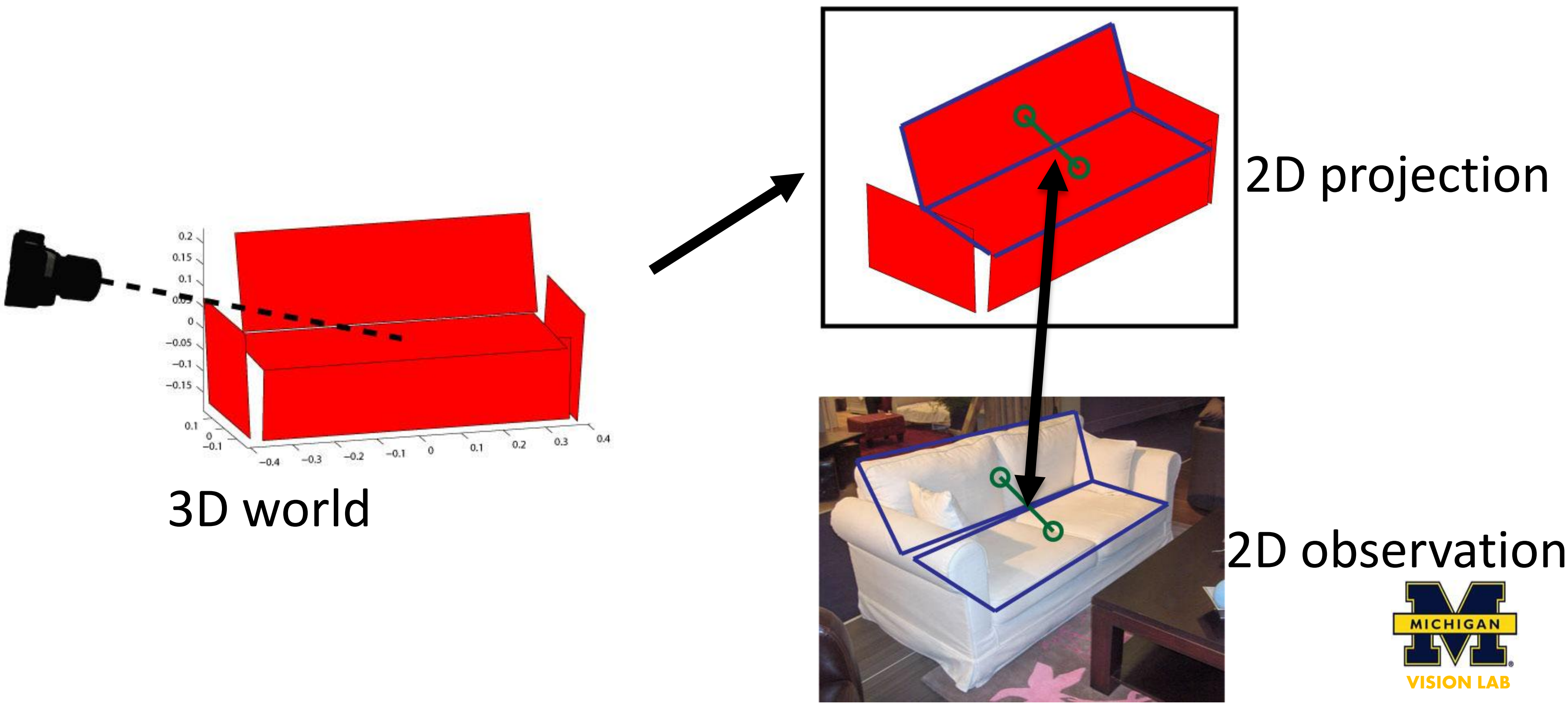
HOG features



ALM only needs one template for each part across all the viewpoints.

Aspect Layout Model

- Pairwise potential
  - Constrains 2D relative locations of parts

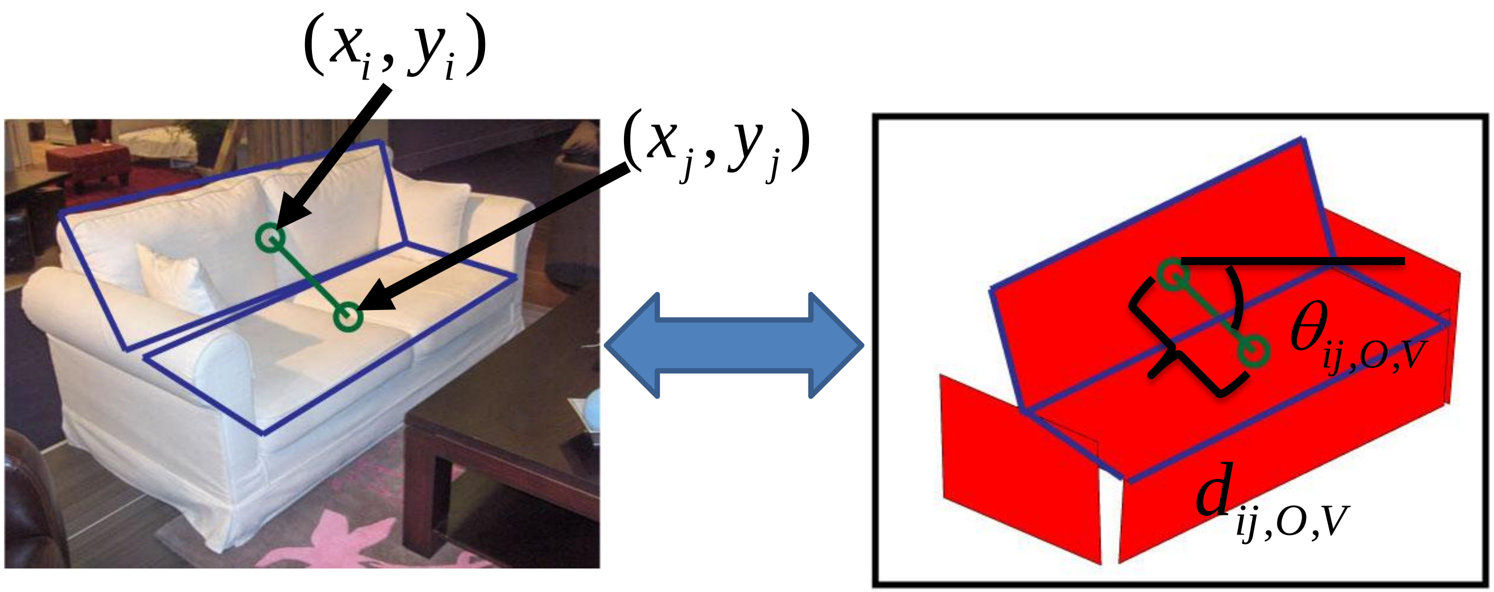


Aspect Layout Model

- Pairwise potential

$$V_2(\mathbf{l}_i, \mathbf{l}_j, O, V)$$

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



Aspect Layout Model

- Energy function

$$E(Y, L, O, V, I \mid \theta) = \theta^T \Psi(Y, L, O, V, I)$$

  – Parameters

$$\theta = (\mathbf{w}_{i,\forall i}, \alpha_{i,\forall i}, w_x, w_y)$$

  – Linear energy function

MICHIGAN VISION LAB

Aspect Layout Model

- Maximal margin parameter estimation
  - Energy based learning [1]: find an energy function which outputs the maximal energy value for the correct label configuration of an object
  - Training set

    $$T = \{(I^t, Y^t, L^t, O^t, V^t), t = 1, \ldots N\}$$
  - Structural SVM optimization [2]

[1] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato and F. J. Huang. A tutorial on energy-based learning. In Predicting Structured Data, MIT Press, 2006.

[2] I. Tsochantaridis, T. Hofmann, T. Joachims and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In ICML, 2004.

MICHIGAN VISION LAB

Aspect Layout Model

- Model inference

$$(Y^*, L^*, O^*, V^*) = \arg\max_{Y,L,O,V} E(Y, L, O, V, I \mid \theta)$$

  - Run Belief Propagation (BP) [1] for each combination of O and V to obtain $E(Y=+1, L^*, O^*, V^*)$
  - Recall the graph structure
  - $Y^* = +1$ if $E(Y=+1, L^*, O^*, V^*) > \gamma$ (detection threshold)

[1] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In Exploring artificial intelligence in the new millennium, 2003.

MICHIGAN VISION LAB

Experiments

- Datasets
  - 3DObject dataset [1]: 10 categories, 10 instances each category
  - VOC 2006 Car dataset [2]: 921 car images
  - EPFL Car dataset [3]: 2299 images, 20 instances
  - Our new ImageNet dataset [4]: Bed (400), Chair (770), sofa (800), table (670)

[1] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In ICCV, 2007.
[2] M. Everingham, A. Zisserman, I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 Results.
[3] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In CVPR, 2009.
[4] http://www.image-net.org.

MICHIGAN VISION LAB

Experiments

- Datasets
  - 3DObject dataset [1]: 10 categories, 10 instances each category
  - VOC 2006 Car dataset [2]: 921 car images
  - EPFL Car dataset [3]: 2299 images, 20 instances
  - Our new ImageNet dataset [4]: Bed (400), Chair (770), sofa (800), table (670)

[1] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In ICCV, 2007.

[2] M. Everingham, A. Zisserman, I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 Results.

[3] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In CVPR, 2009.

[4] http://www.image-net.org.

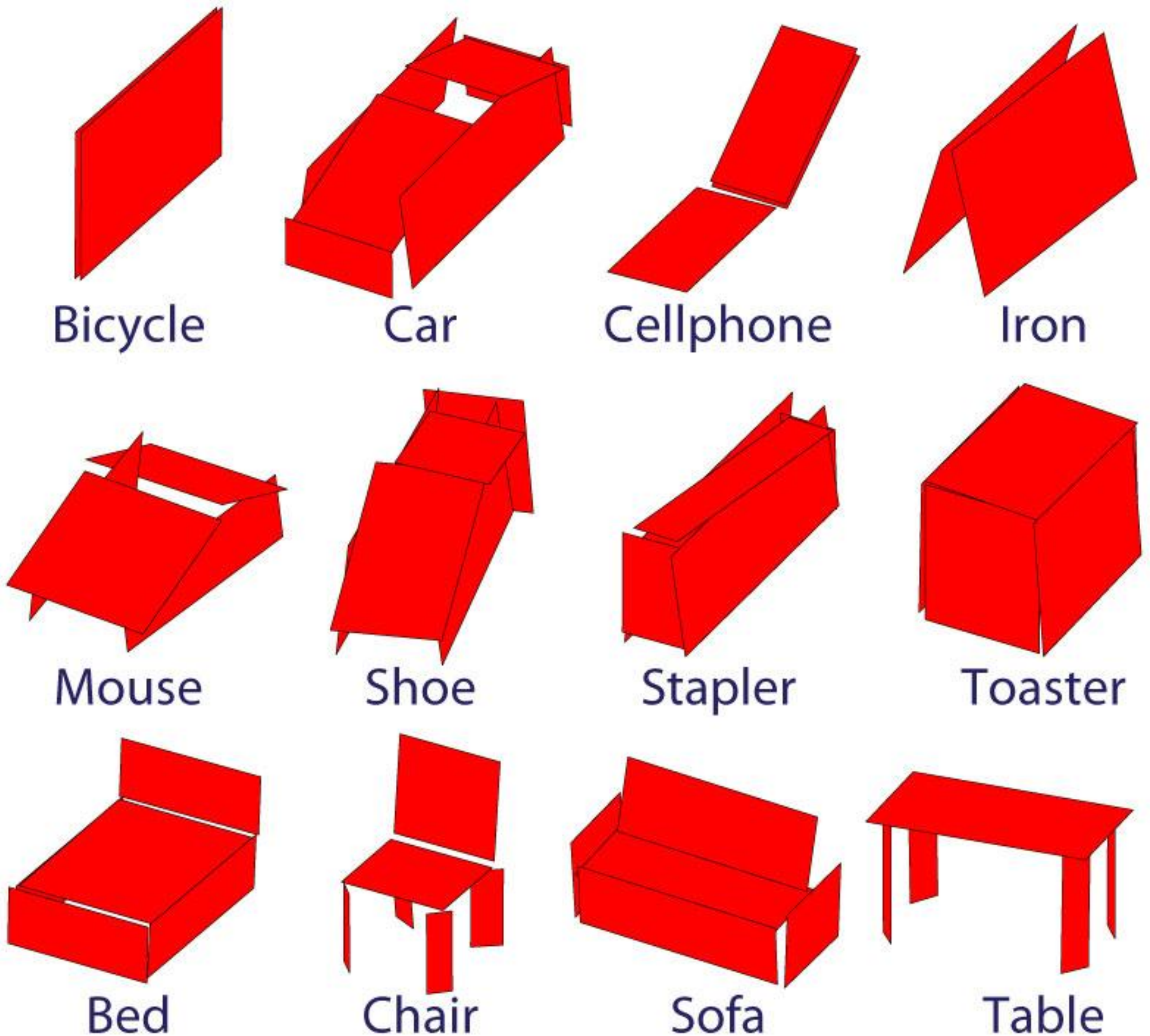
Experiments

- Evaluation measures
  - Detection: Average Precision (AP)
  - Viewpoint: average viewpoint accuracy (the average of the elements on the main diagonal of the viewpoint confusion matrix)
  - Part localization: Percentage of Correct Parts (PCP)-recall curve

MICHIGAN VISION LAB

Experiments

- 3D models



Experiments

- Average results for eight categories on the 3DObject dataset (8 views)

Method	ALM	[1]	[2]
Viewpoint	**80.7**	74.2	57.2
Detection	81.8	n/a	n/a






[1] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In ECCV, 2010.
[2] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In ICCV, 2007.

MICHIGAN VISION LAB

Experiments

- Results on the Bicycle Category in the 3DObject dataset

Method	ALM	[1]	[2]
Viewpoint	**91.4**	80.8	75.0
Detection	**93.0**	n/a	69.8



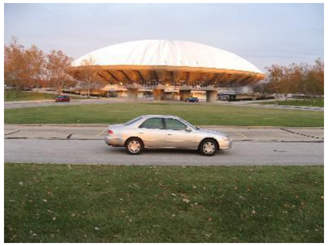


[1] N. Payet and S. Todorovic. From contours to 3d object detection and pose estimation. In ICCV, 2011.
[2] J. Liebelt and C. Schmid. Multi-view object class detection with a 3D geometric model. In CVPR, 2010.

Experiments

- Results on the Car Category in the 3DObject dataset

Method	ALM	[1]	[2]	[3]	[4]	[5]	[6]
Viewpoint	**93.4**	85.4	85.3	81	70	67	48.5
Detection	98.4	n/a	**99.2**	89.9	76.7	55.3	n/a


[1] N. Payet and S. Todorovic. From contours to 3d object detection and pose estimation. In ICCV, 2011.
[2] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and pose estimation. In ICCV, 2011.
[3] M. Stark, M. Goesele, and B. Schiele. Back to the future: Learning shape models from 3d cad data. In BMVC, 2010.
[4] J. Liebelt and C. Schmid. Multi-view object class detection with a 3D geometric model. In CVPR, 2010.
[5] 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.
[6] M. Arie-Nachimson and R. Basri. Constructing implicit 3d shape models for pose estimation. In ICCV, 2009.

Experiments

- Detailed average viewpoint accuracy on the 3DObject dataset

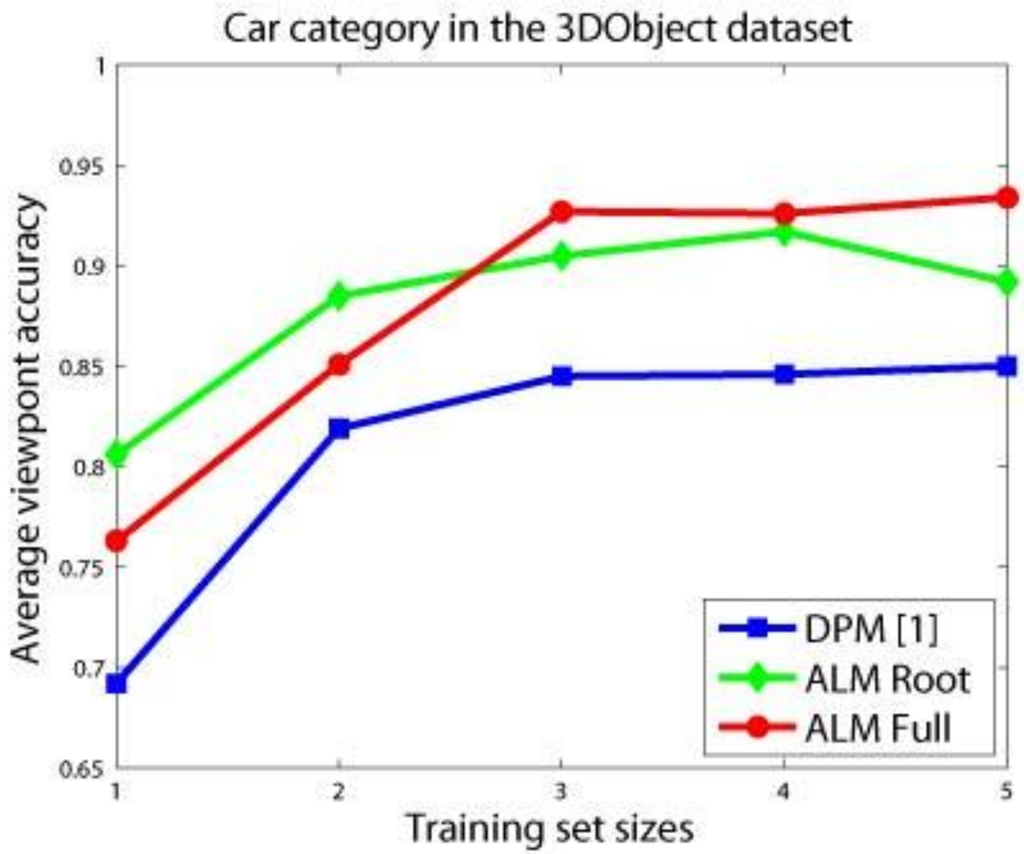
Category	Bicycle	Car	Cellphone	Iron	Mouse	Shoe	Stapler	Toaster
DPM [1]	88.4	85.0	62.1	82.7	40.0	71.7	58.5	55.0
ALM Root	**92.5**	89.2	83.4	**86.0**	58.7	82.7	69.2	59.6
ALM Full	91.4	**93.4**	**85.0**	84.6	**66.5**	**87.0**	**72.8**	**65.2**

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. TPAMI, 2010.

MICHIGAN VISION LAB

Experiments

- Effect of training set sizes for viewpoint

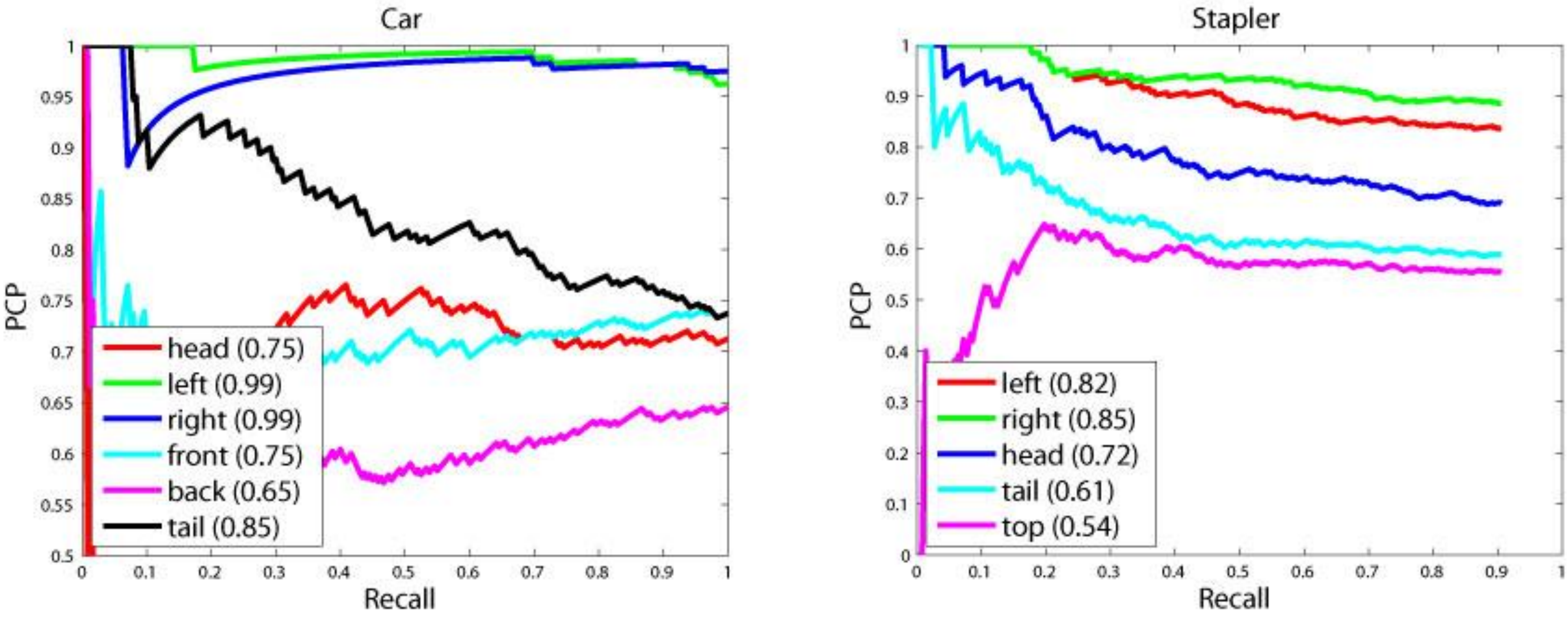


[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. TPAMI, 2010.

MICHIGAN VISION LAB

Experiments

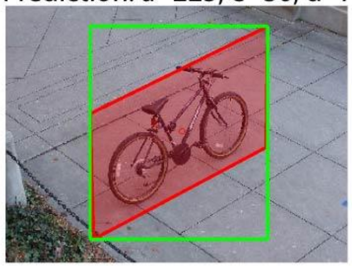
- Part localization on the 3DObject dataset



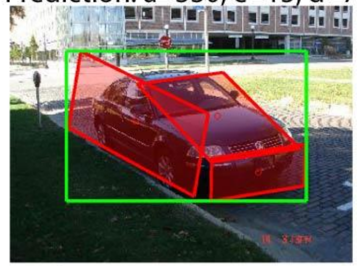
Experiments

Prediction: a=45, e=15, d=5

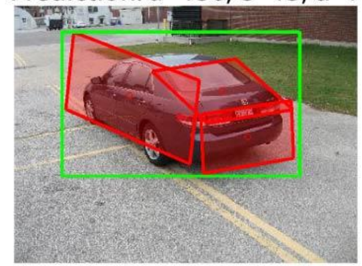
Prediction: a=225, e=30, d=7



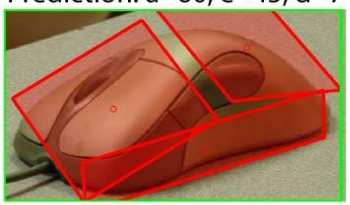
Prediction: a=330, e=15, d=7



Prediction: a=150, e=15, d=7



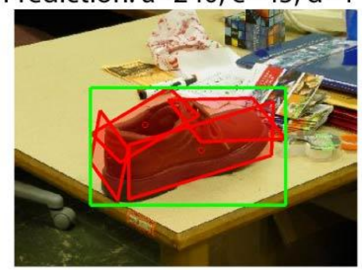
Prediction: a=60, e=45, d=7



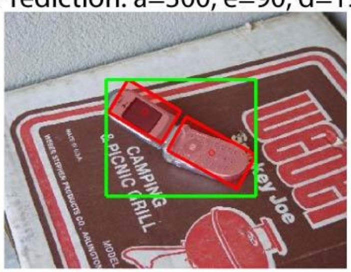
Prediction: a=300, e=45, d=23

Prediction: a=45, e=90, d=5

Prediction: a=240, e=45, d=11

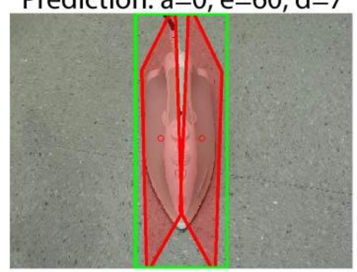


Prediction: a=300, e=90, d=15

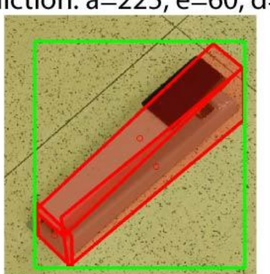


Prediction: a=135, e=0, d=11

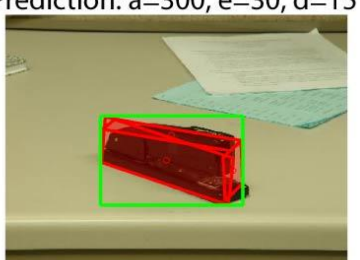
Prediction: a=0, e=60, d=7



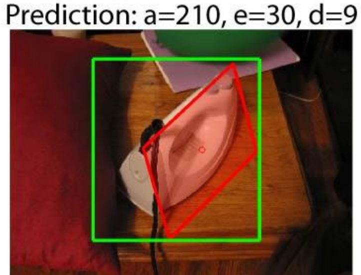
Prediction: a=225, e=60, d=7



Prediction: a=300, e=30, d=15



Prediction: a=210, e=30, d=9

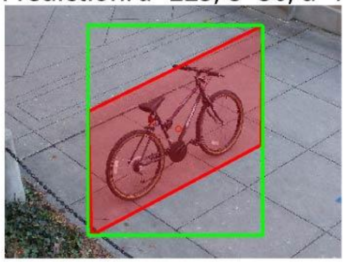


VISION LAB

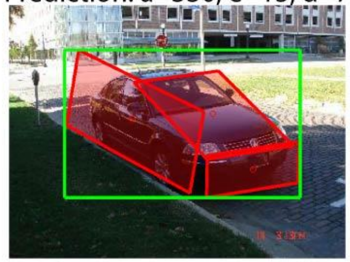
Experiments

Prediction: a=45, e=15, d=5

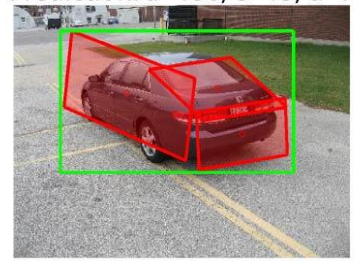
Prediction: a=225, e=30, d=7



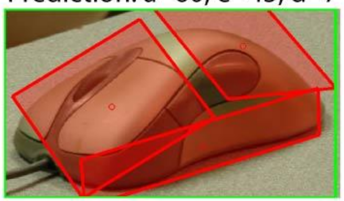
Prediction: a=330, e=15, d=7



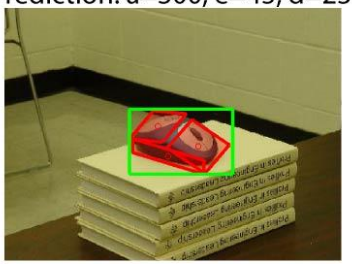
Prediction: a=150, e=15, d=7



Prediction: a=60, e=45, d=7

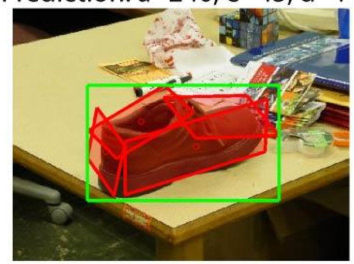


Prediction: a=300, e=45, d=23

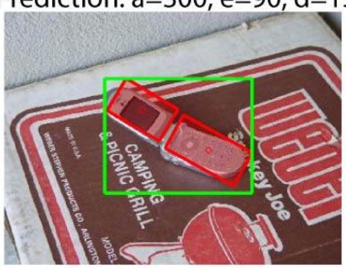


Prediction: a=45, e=90, d=5

Prediction: a=240, e=45, d=11

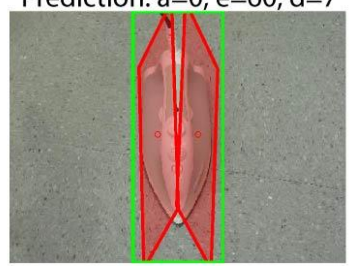


Prediction: a=300, e=90, d=15



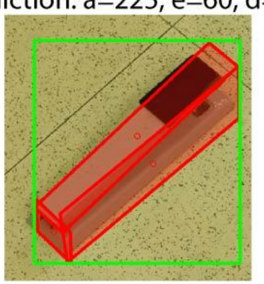
Prediction: a=135, e=0, d=11

Prediction: a=0, e=60, d=7

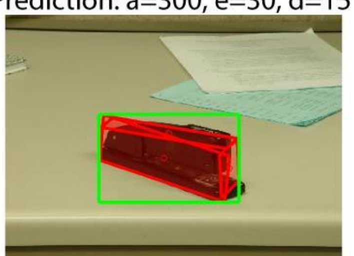


Prediction: a=330, e=15, d=7

Prediction: a=225, e=60, d=7

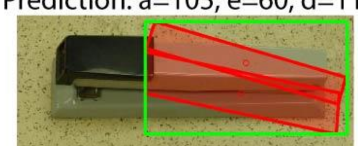


Prediction: a=300, e=30, d=15



Prediction: a=210, e=30, d=9

Prediction: a=105, e=60, d=11



MICHIGAN
VISION LAB

Experiments

- Average results on the ImageNet dataset

Method	ALM Full	ALM Root	DPM [1]
3 views	**86.5**	79.0	84.6
7 views	**63.4**	34.0	49.5

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. TPAMI, 2010.

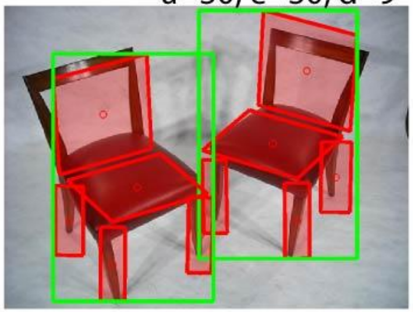
Experiments

Prediction: a=30, e=15, d=2.5

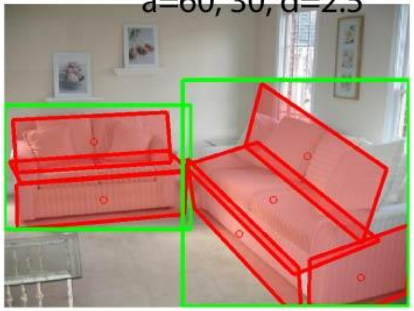
Prediction: a=0, e=15, d=1.5

Prediction: a=0, e=30, d=7

Prediction: a=330, e=30, d=9
a=30, e=30, d=9

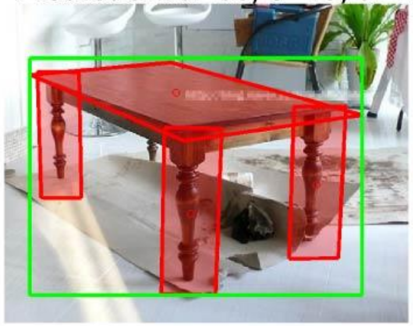


Prediction: a=345, e=15, d=3.5
a=60, 30, d=2.5



Prediction: a=315, e=30, d=2

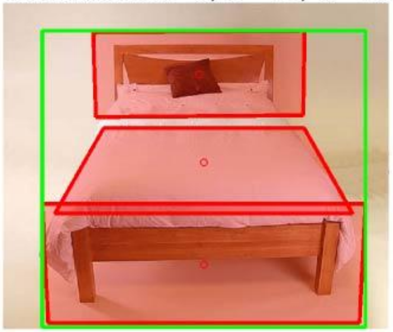
Prediction: a=60, e=15, d=2



Experiments

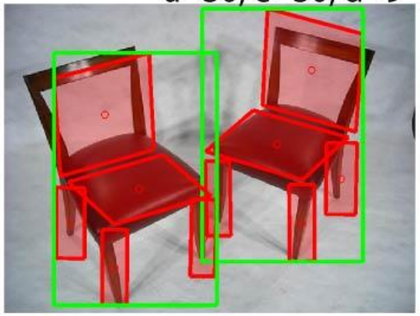
Prediction: a=30, e=15, d=2.5

Prediction: a=0, e=15, d=1.5

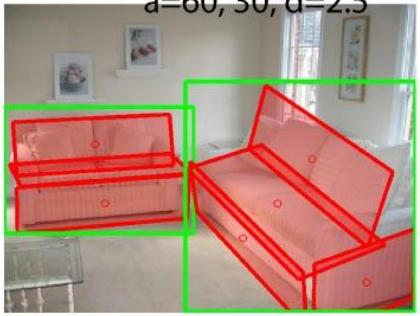


Prediction: a=0, e=30, d=7

Prediction: a=330, e=30, d=9
a=30, e=30, d=9

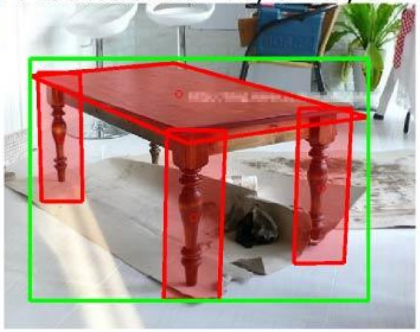


Prediction: a=345, e=15, d=3.5
a=60, 30, d=2.5

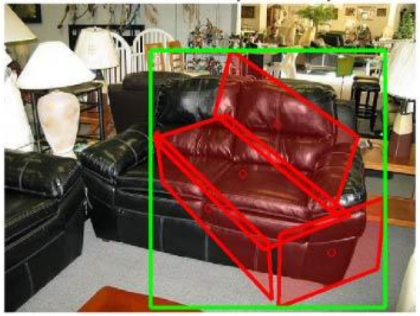


Prediction: a=315, e=30, d=2

Prediction: a=60, e=15, d=2



Prediction: a=60, e=30, d=2.5



Conclusion

- A new Aspect Layout Model (ALM) for object detection, pose estimation and aspect part localization.
- ALM is capable of handling large number of views, locating aspect parts and reasoning self-occlusion.
- ALM can be useful for estimating functional parts or object affordances.
- Our code and datasets are available online.

MICHIGAN VISION LAB

Acknowledgments

NSF

SCIENCE TO SHAPE THE FUTURE OF THE ARMY
ARMY RESEARCH OFFICE
1951

KLA Tencor
Accelerating Yield

Thank you!

MICHIGAN
VISION LAB