C280, Computer Vision

Prof. Trevor Darrell
trevor@eecs.berkeley.edu

Lecture 16: Recognition in Context

Last Lecture

- Naïve-Bayes Nearest Neighbor (Irani)
- ISM (Liebe)
- Constellation Models (Fergus)
- Transformed LDA Models (Sudderth)
- 3-D view models (Saravese)

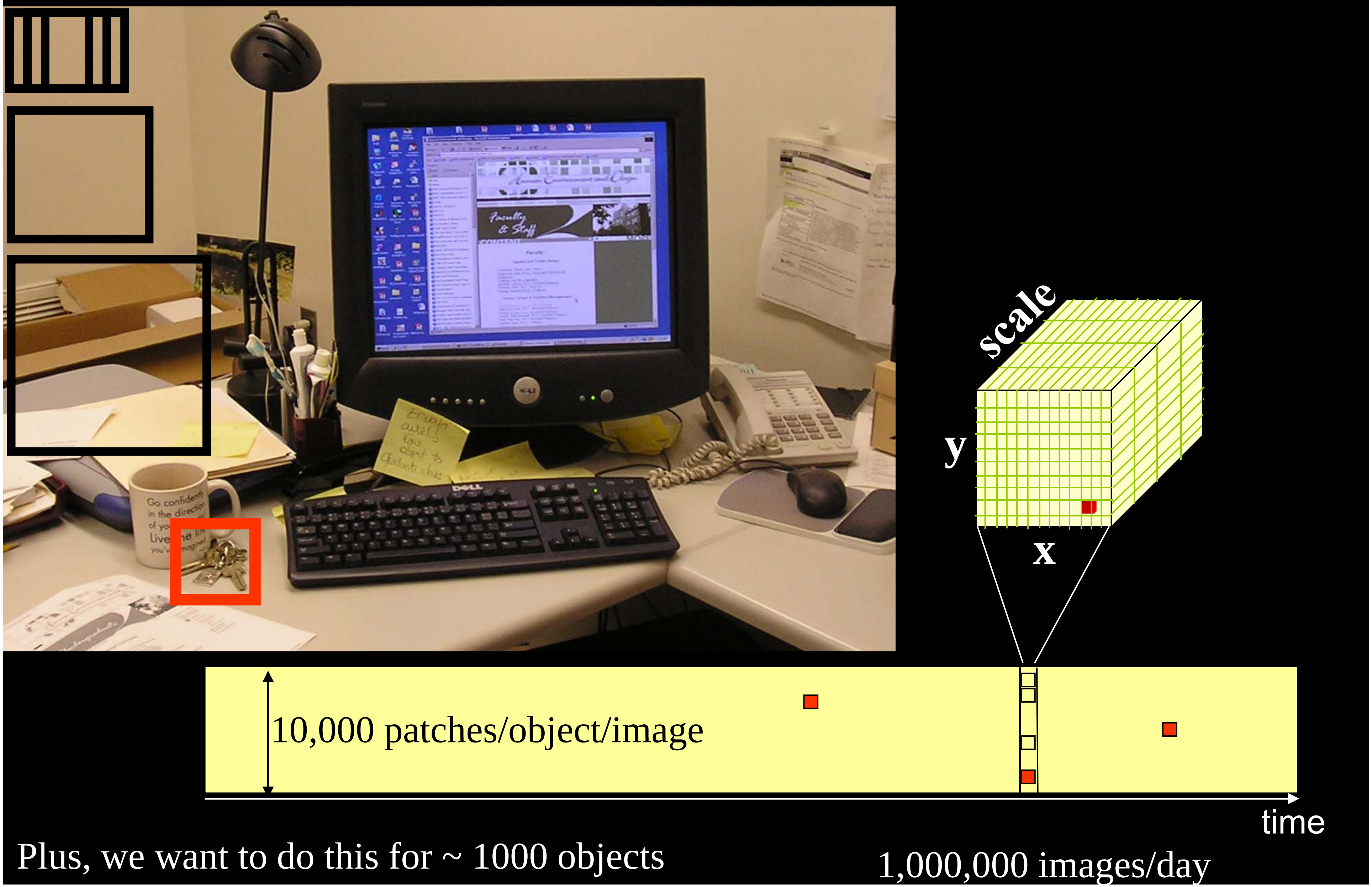
This week

- Two last topics in recognition:
  - Context
  - Articulation

Today: Three papers on computational models of context:

- A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in Advances in Neural Information Processing Systems 17 (NIPS), 2005.
- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006
- G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43.

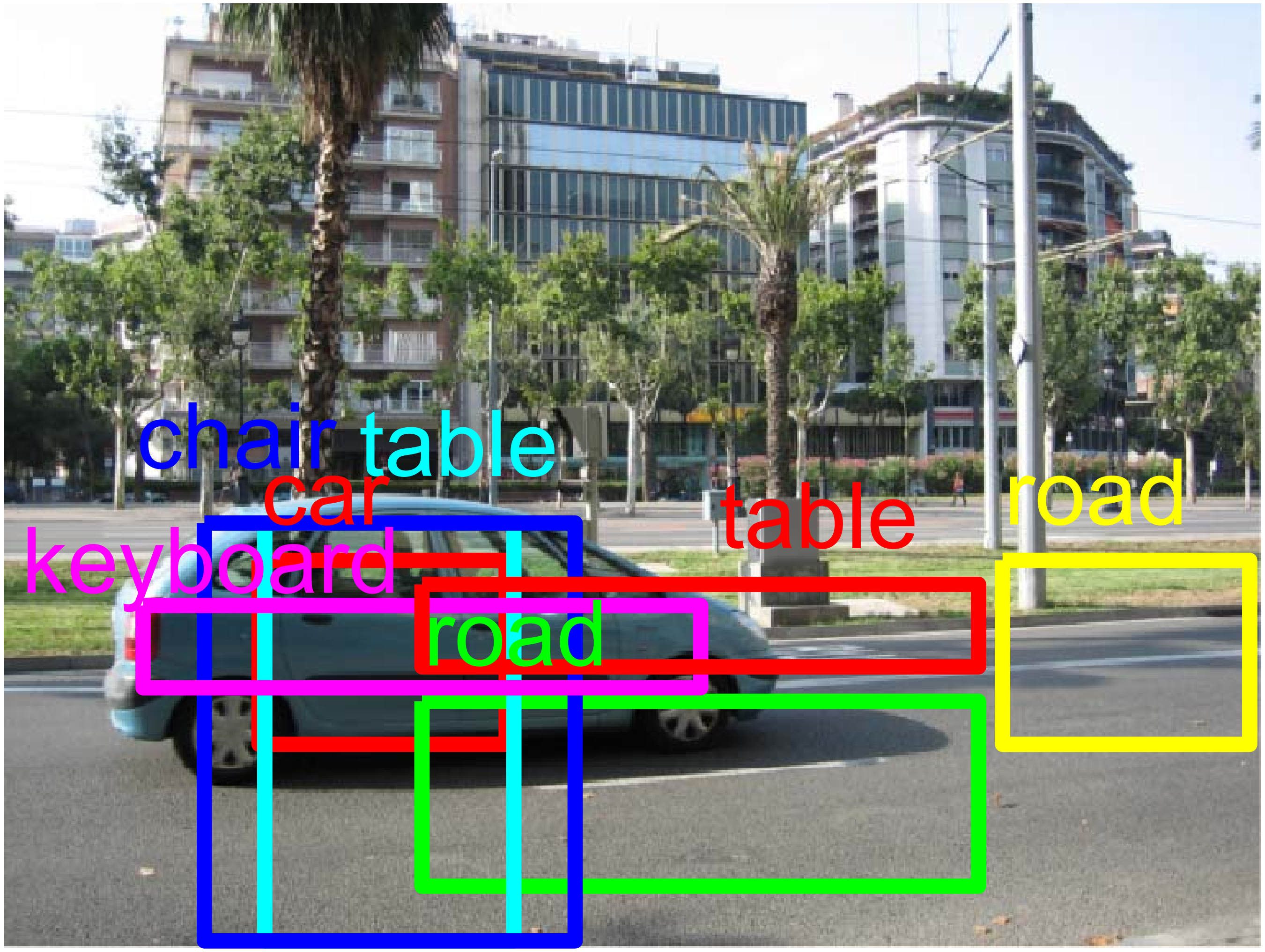
Why is detection hard?



Is local information enough?

Slide credit: A. Torralba

With hundreds of categories



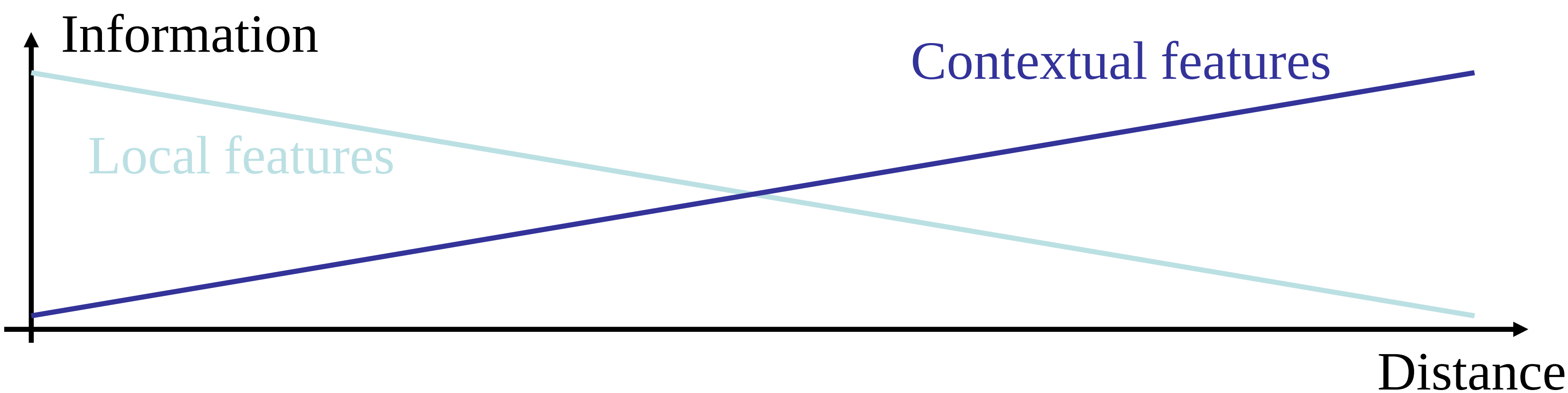
If we have 1000 categories (detectors), and each detector produces 1 fa every 10 images, we will have 100 false alarms per image… pretty much garbage…

Slide credit: A. Torralba

Is local information even enough?

Slide credit: A. Torralba

Is local information even enough?

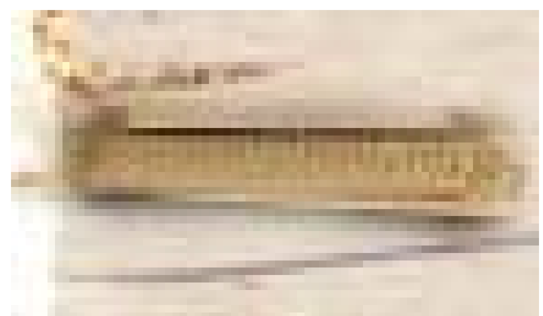


Slide credit: A. Torralba

The system does not care about the scene, but we do…

We know there is a keyboard present in this scene even if we cannot see it clearly.

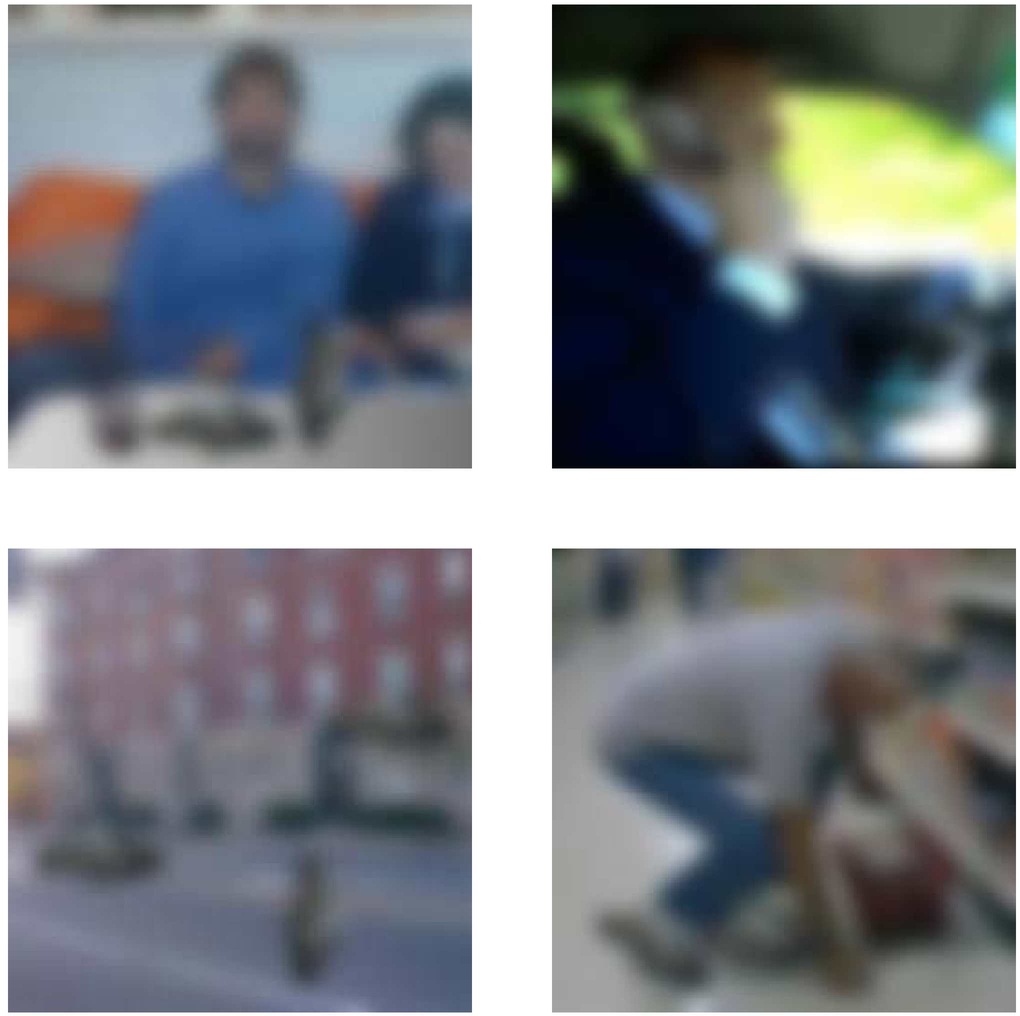
We know there is no keyboard present in this scene



… even if there is one indeed.

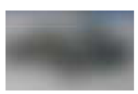
Slide credit: A. Torralba

The multiple personalities of a blob



Slide credit: A. Torralba

The multiple personalities of a blob



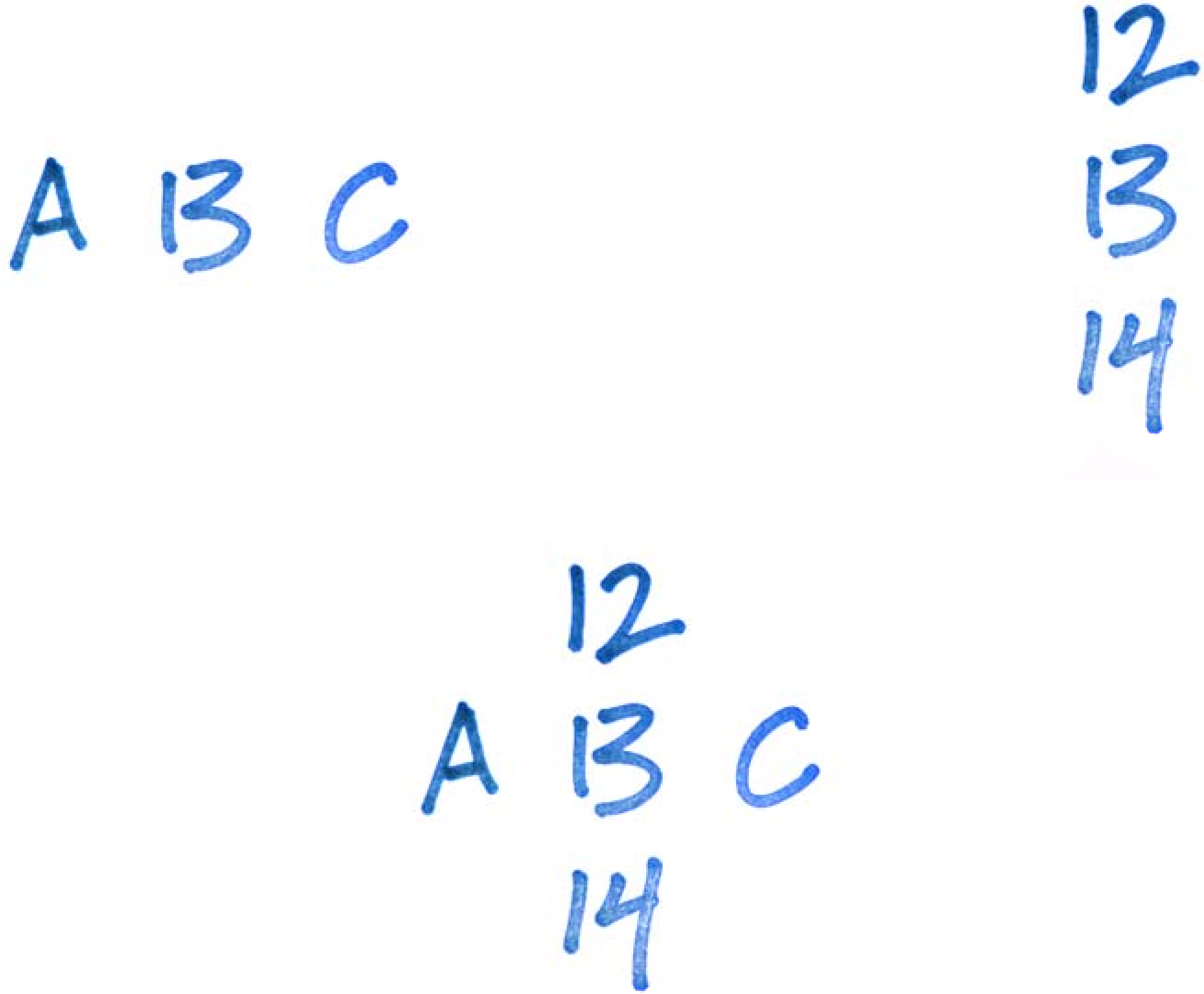
Slide credit: A. Torralba

A B C

Slide credit: A. Torralba

12
13
14

Slide credit: A. Torralba

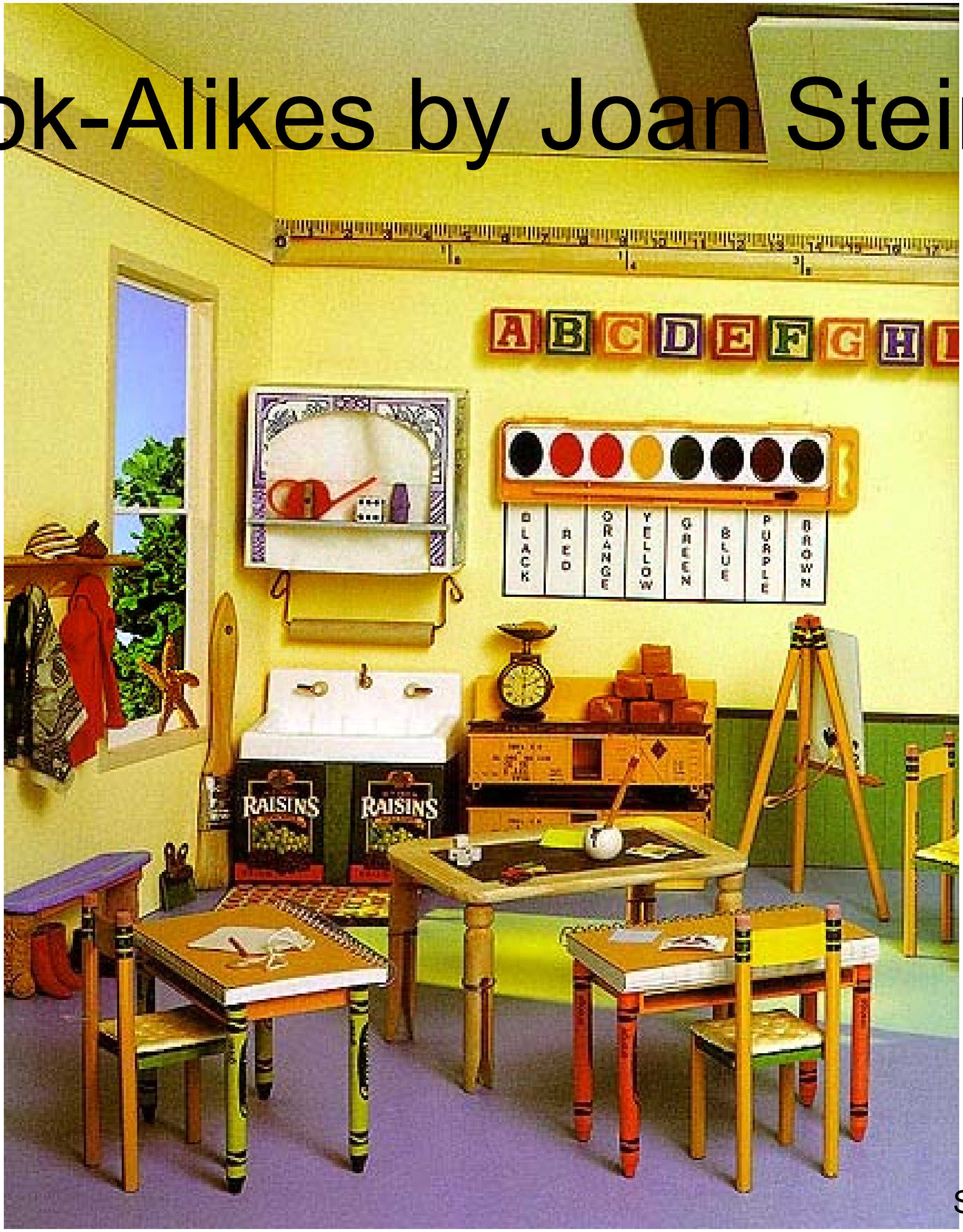


Slide credit: A. Torralba

Look-Alikes by Joan Steiner

Slide credit: A. Torralba

Look-Alikes by Joan Steiner



Slide credit: A. Torralba

Look-Alikes by Joan Steiner

Slide credit: A. Torralba

The context challenge

How far can you go without using an object detector?

Slide credit: A. Torralba

What are the hidden objects?

Slide credit: A. Torralba

What are the hidden objects?

Slide credit: A. Torralba

The importance of context

- Cognitive psychology
  - Palmer 1975
  - Biederman 1981
  - …

- Computer vision
  - Noton and Stark (1971)
  - Hanson and Riseman (1978)
  - Barrow & Tenenbaum (1978)
  - Ohta, kanade, Skai (1978)
  - Haralick (1983)
  - Strat and Fischler (1991)
  - Bobick and Pinhanez (1995)
  - Campbell et al (1997)

Class	Context elements	Operator
SKY	ALWAYS	ABOVE-HORIZON
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE	BLUE
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE	WHITE
SKY	SPARSE-RANGE-IS-AVAILABLE	SPARSE-RANGE-IS-UNDEFINED
SKY	CAMERA-IS-HORIZONTAL	NEAR-TOP
SKY	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(complete-sky)	ABOVE-SKYLINE
SKY	CLIQUE-CONTAINS(sky)	SIMILAR-INTENSITY
SKY	CLIQUE-CONTAINS(sky)	SIMILAR-TEXTURE
SKY	RGB-IS-AVAILABLE ∧ CLIQUE-CONTAINS(sky)	SIMILAR-COLOR
GROUND	CAMERA-IS-HORIZONTAL	HORIZONTALLY-STRIATED
GROUND	CAMERA-IS-HORIZONTAL	NEAR-BOTTOM
GROUND	SPARSE-RANGE-IS-AVAILABLE	SPARSE-RANGES-FORM-HORIZONT
GROUND	DENSE-RANGE-IS-AVAILABLE	DENSE-RANGES-FORM-HORIZONTA
GROUND	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(complete-ground)	BELOW-SKYLINE
GROUND	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(geometric-horizon) ∧ ¬ CLIQUE-CONTAINS(skyline)	BELOW-GEOMETRIC-HORIZON
GROUND	TIME-IS-DAY	DARK

Slide credit: A. Torralba

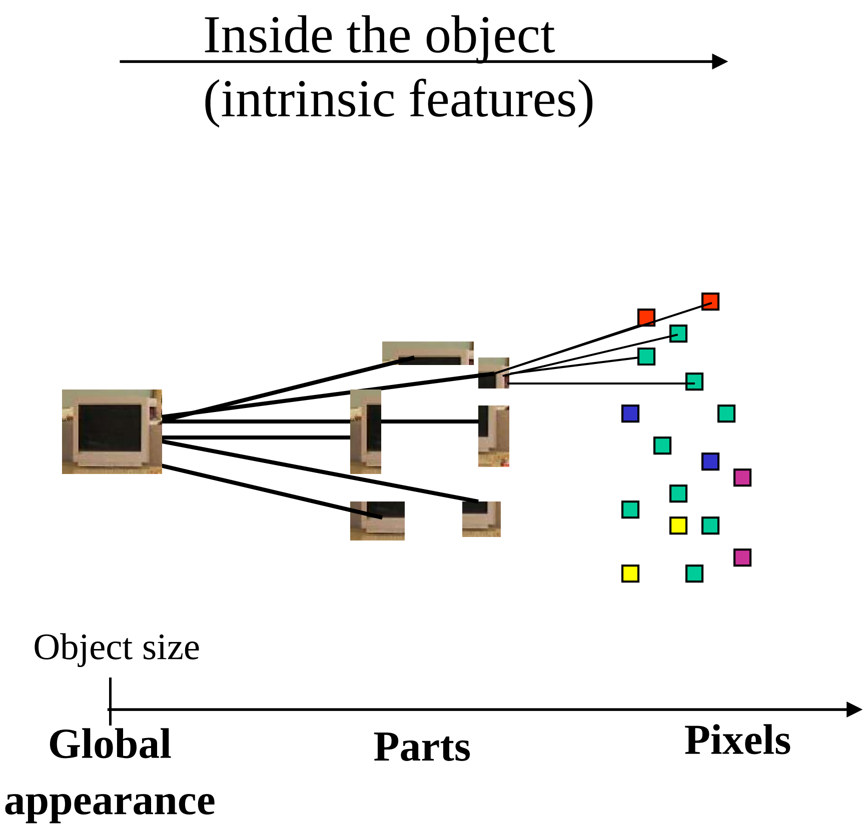
Multiclass object detection and context modeling

Antonio Torralba

In collaboration with

Kevin P. Murphy and William T. Freeman

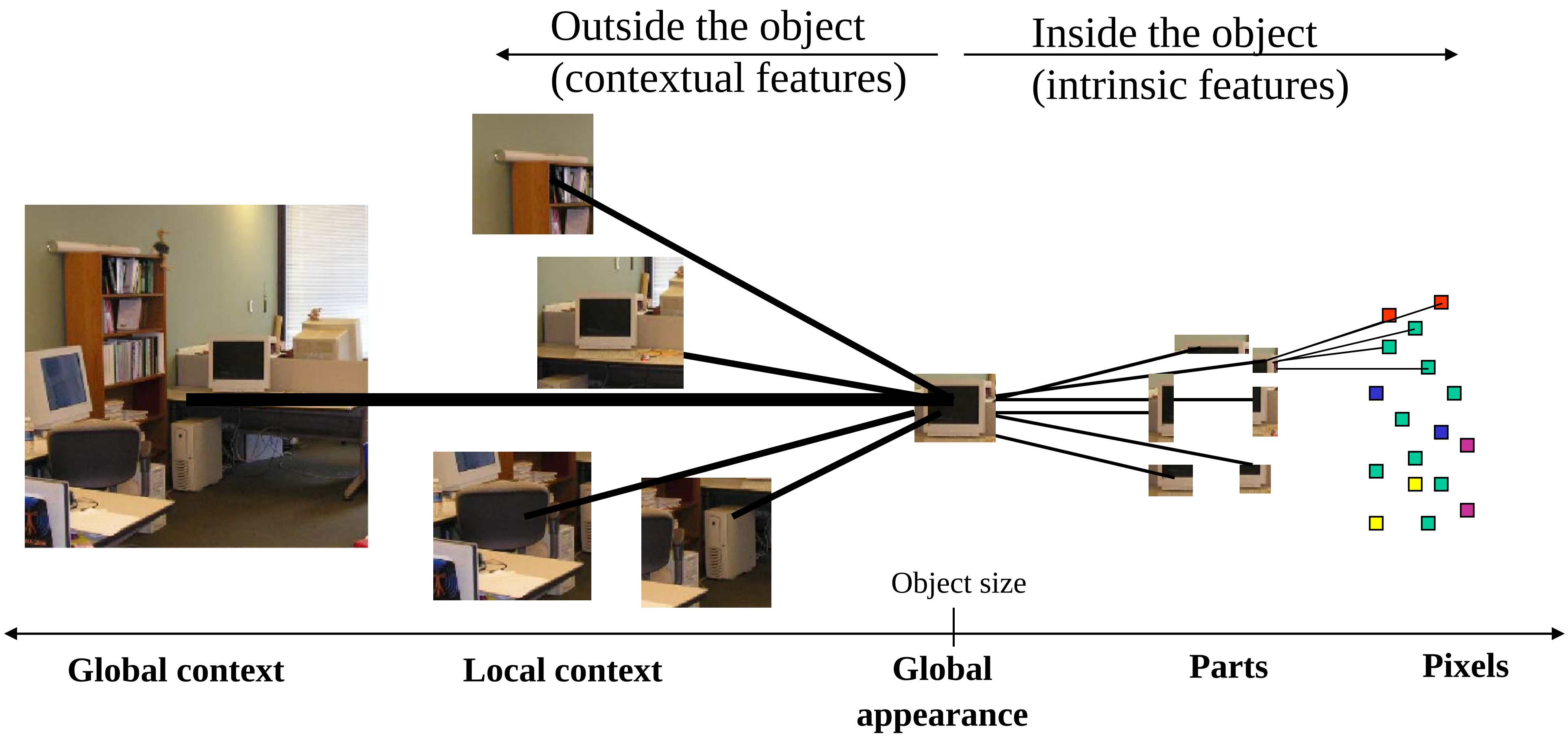
Object representations



Agarwal & Roth, (02), Moghaddam, Pentland (97), Turk, Pentland (91),Vidal-Naquet, Ullman, (03)

Heisele, et al, (01), Agarwal & Roth, (02), Kremp, Geman, Amit (02), Dorko, Schmid, (03)
Fergus, Perona, Zisserman (03), Fei Fei, Fergus, Perona, (03), Schneiderman, Kanade (00), Lowe (99)
Etc.

Object representations



Kruppa & Shiele, (03), Fink & Perona (03)

Carbonetto, Freitas, Barnard (03), Kumar, Hebert, (03)

He, Zemel, Carreira-Perpinan (04), Moore, Essa, Monson, Hayes (99)

Strat & Fischler (91), Murphy, Torralba & Freeman (03)

Agarwal & Roth, (02), Moghaddam, Pentland (97), Turk, Pentland (91),Vidal-Naquet, Ullman, (03)

Heisele, et al, (01), Agarwal & Roth, (02), Kremp, Geman, Amit (02), Dorko, Schmid, (03)
Fergus, Perona, Zisserman (03), Fei Fei, Fergus, Perona, (03), Schneiderman, Kanade (00), Lowe (99)
Etc.

Previous work on context

- Strat & Fischler (91)

Context defined using hand-written rules about relationships between objects

#	Class	Context elements	Operator
41	SKY	ALWAYS	ABOVE-HORIZON
42	SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	BRIGHT
43	SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	UNTEXTURED
44	SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE	BLUE
45	SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	BRIGHT
46	SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	UNTEXTURED
47	SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE	WHITE
48	SKY	SPARSE-RANGE-IS-AVAILABLE	SPARSE-RANGE-IS-UNDEFINED
49	SKY	CAMERA-IS-HORIZONTAL	NEAR-TOP
50	SKY	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(complete-sky)	ABOVE-SKYLINE
51	SKY	CLIQUE-CONTAINS(sky)	SIMILAR-INTENSITY
52	SKY	CLIQUE-CONTAINS(sky)	SIMILAR-TEXTURE
53	SKY	RGB-IS-AVAILABLE ∧ CLIQUE-CONTAINS(sky)	SIMILAR-COLOR
61	GROUND	CAMERA-IS-HORIZONTAL	HORIZONTALLY-STRIATED
62	GROUND	CAMERA-IS-HORIZONTAL	NEAR-BOTTOM
63	GROUND	SPARSE-RANGE-IS-AVAILABLE	SPARSE-RANGES-FORM-HORIZONTAL-SURFACE
64	GROUND	DENSE-RANGE-IS-AVAILABLE	DENSE-RANGES-FORM-HORIZONTAL-SURFACE
65	GROUND	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(complete-ground)	BELOW-SKYLINE
66	GROUND	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(geometric-horizon) ∧ ¬ CLIQUE-CONTAINS(skyline)	BELOW-GEOMETRIC-HORIZON
67	GROUND	TIME-IS-DAY	DARK
71	FOLIAGE	ALWAYS	HIGHLY-TEXTURED
72	FOLIAGE	ALWAYS	HIGH-VEGETATIVE-TRANSPARENCY
73	FOLIAGE	CAMERA-IS-HORIZONTAL	NEAR-TOP
74	FOLIAGE	RGB-IS-AVAILABLE	GREEN
76	RAISED-OBJECT	SPARSE-RANGE-IS-AVAILABLE	SPARSE-HEIGHT-ABOVE-GROUND
77	RAISED-OBJECT	DENSE-RANGE-IS-AVAILABLE	DENSE-HEIGHT-ABOVE-GROUND
78	RAISED-OBJECT	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(complete-sky)	ABOVE-SKYLINE

Table 5: Type II Context Sets: Candidate Evaluation

Previous work on context

- Fink & Perona (03)

Use output of boosting from other objects at previous iterations as input into boosting for this iteration

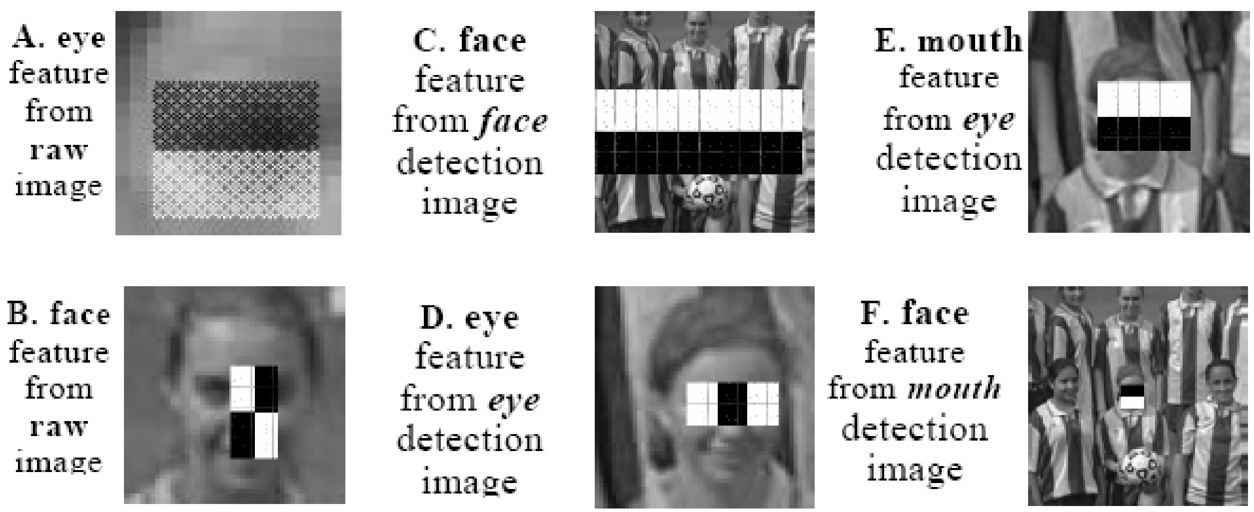
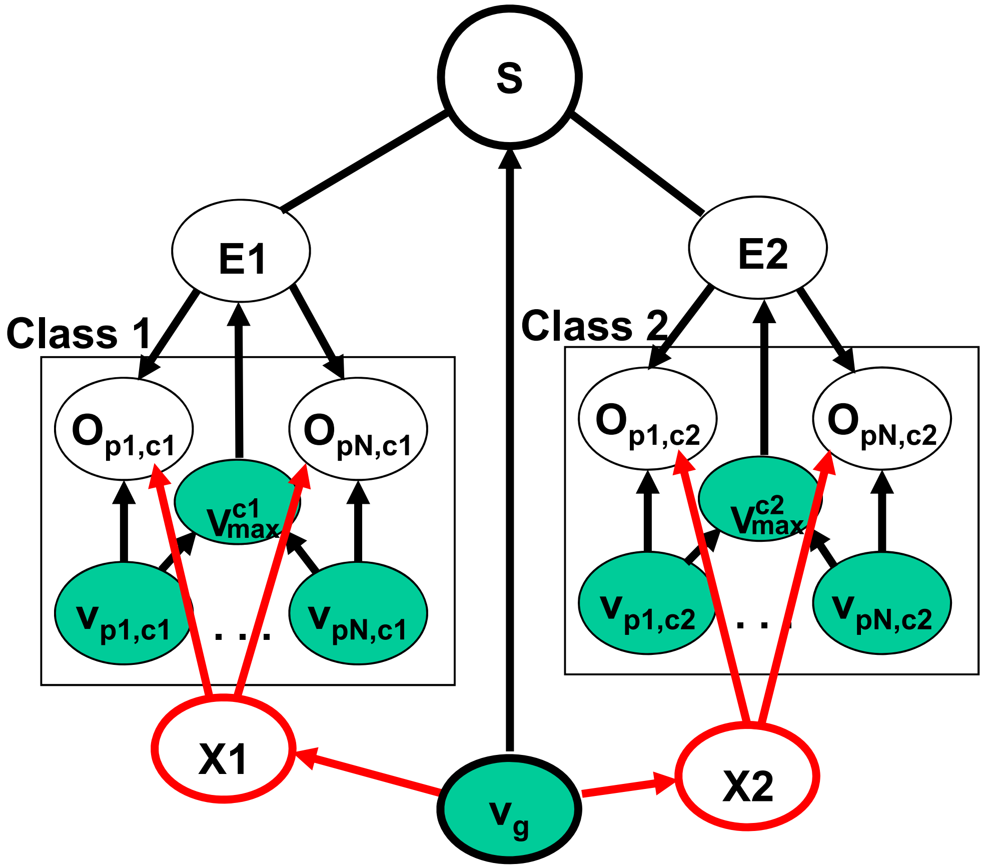


Figure 5: **A-E.** Emerging features of eyes, mouths and faces (presented on windows of raw images for legibility). The windows' scale is defined by the detected object size and by the map mode (local or contextual). **C.** faces are detected using face detection maps H^{Face}, exploiting the fact that faces tend to be horizontally aligned.

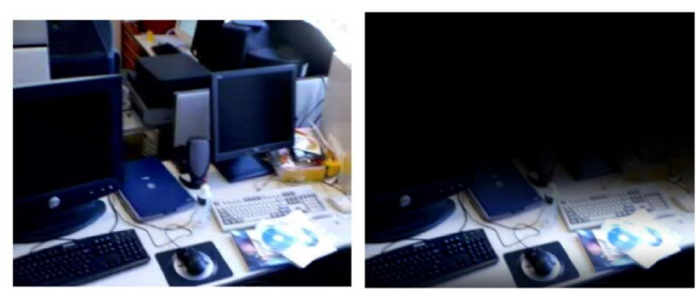
Previous work on context

- Murphy, Torralba & Freeman (03)

Use global context to predict objects but there is no modeling of spatial relationships between objects.

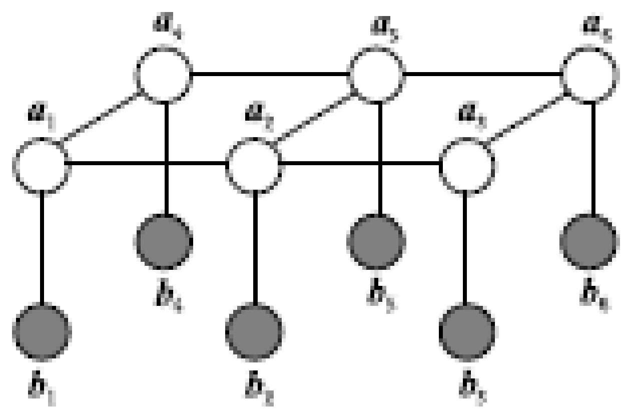


Keyboards

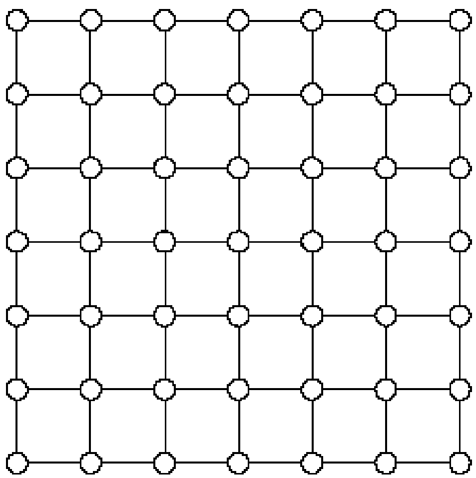


Previous work on context

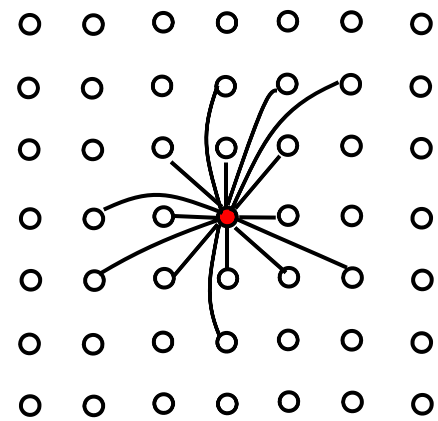
- Carbonetto, de Freitas & Barnard (04)
- Enforce spatial consistency between labels using MRF



Graphical models for image labeling



Nearest neighbor grid



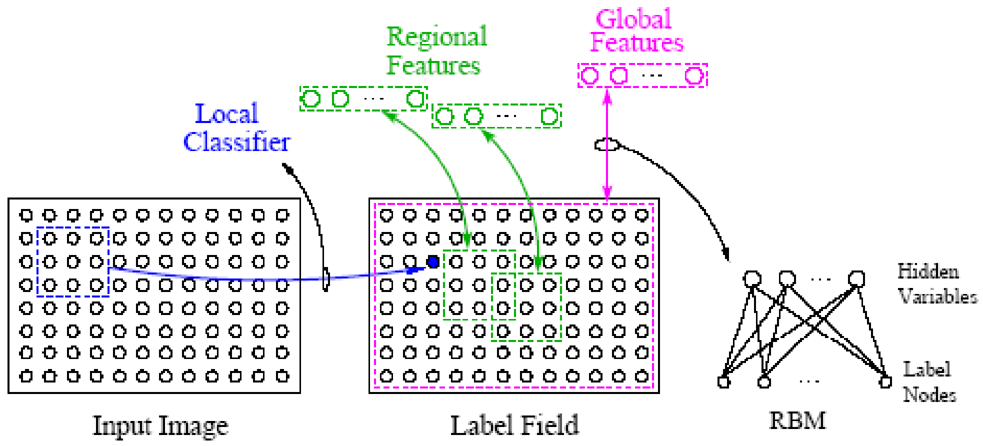
Densely connected graphs
with low informative connections

Want to model long-range correlations between labels

Previous work on context

- He, Zemel & Carreira-Perpinan (04)

Use latent variables to induce long distance correlations between labels in a Conditional Random Field (CRF)



Outline of this talk

- Use global image features (as well as local features) in boosting to help object detection
- Learn structure of dense CRF (with long range connections) using boosting, to exploit spatial correlations

Image database

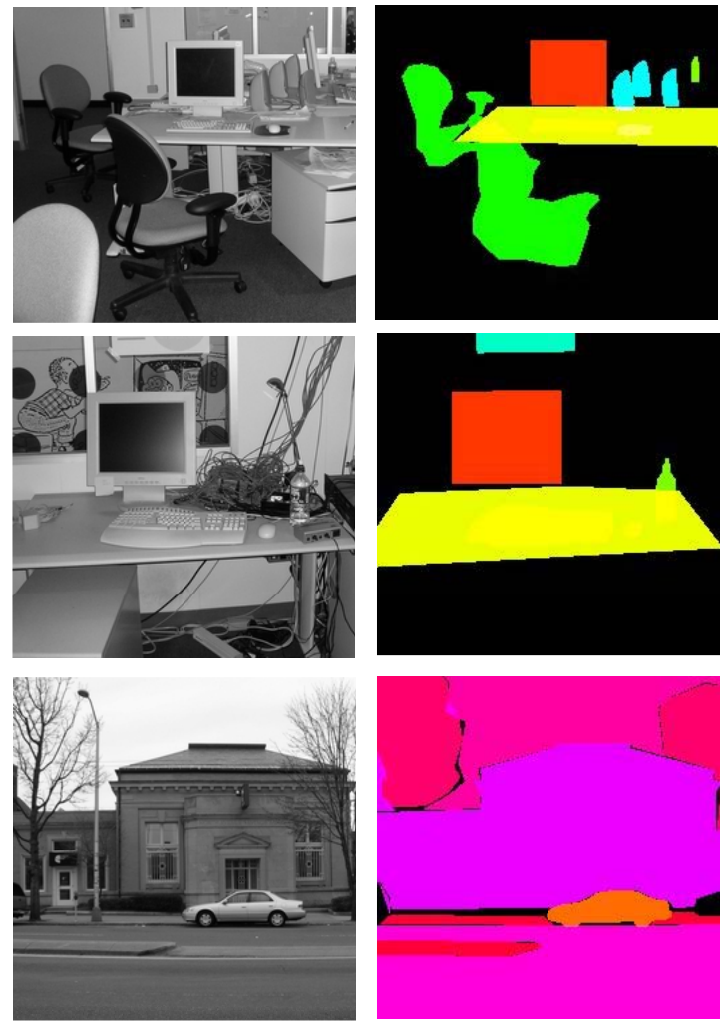
• ~2500 hand labeled images with segmentations

• ~30 objects and stuff

• Indoor and outdoor

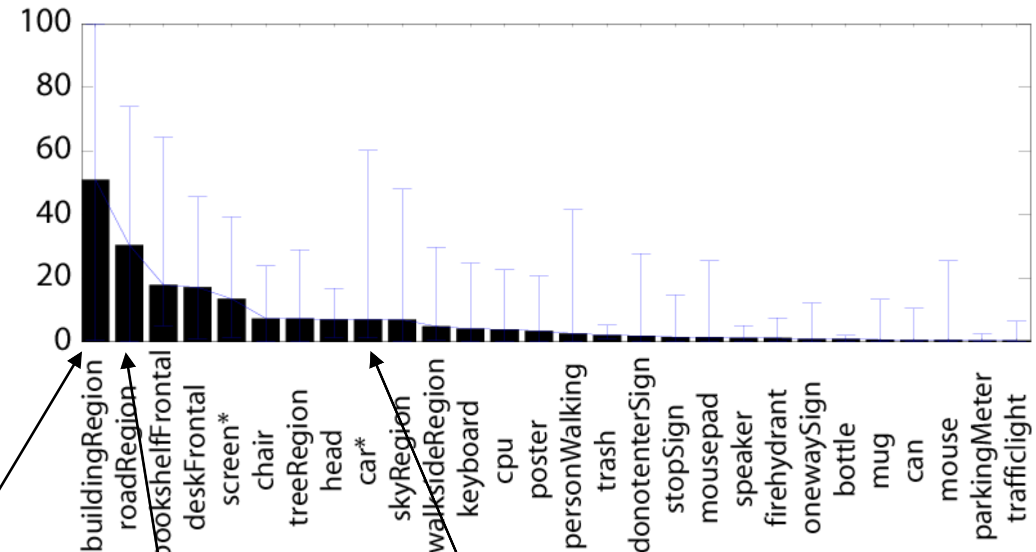
• Sets of images are separated by locations and camera (digital/webcam)

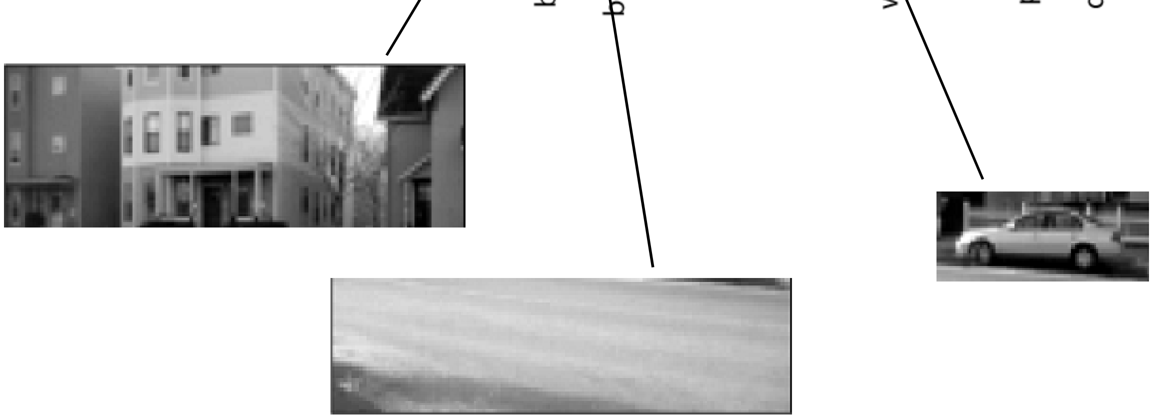
• No graduate students or low-income-student-class exploited for labeling.



Which objects are important?

Average percentage of pixels occupied by each object.

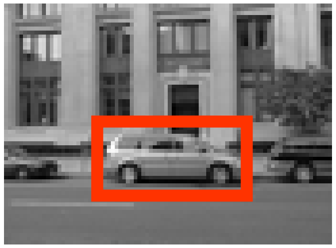




Object representation

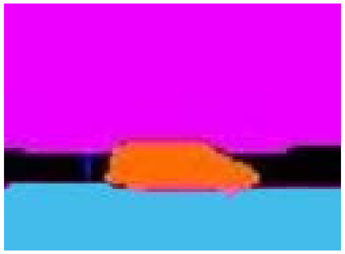
- **Discrete/**bounded/rigid

Screen, car, pedestrian, bottle, …



- **Extended**/unbounded/deformable

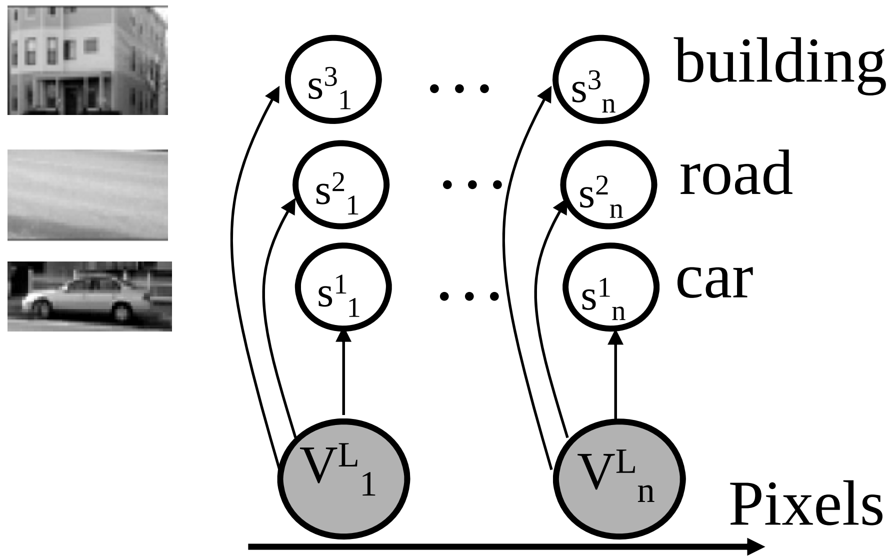
Building, sky, road, shelves, desk, …



We will use region labeling as a representation.

Learning local features

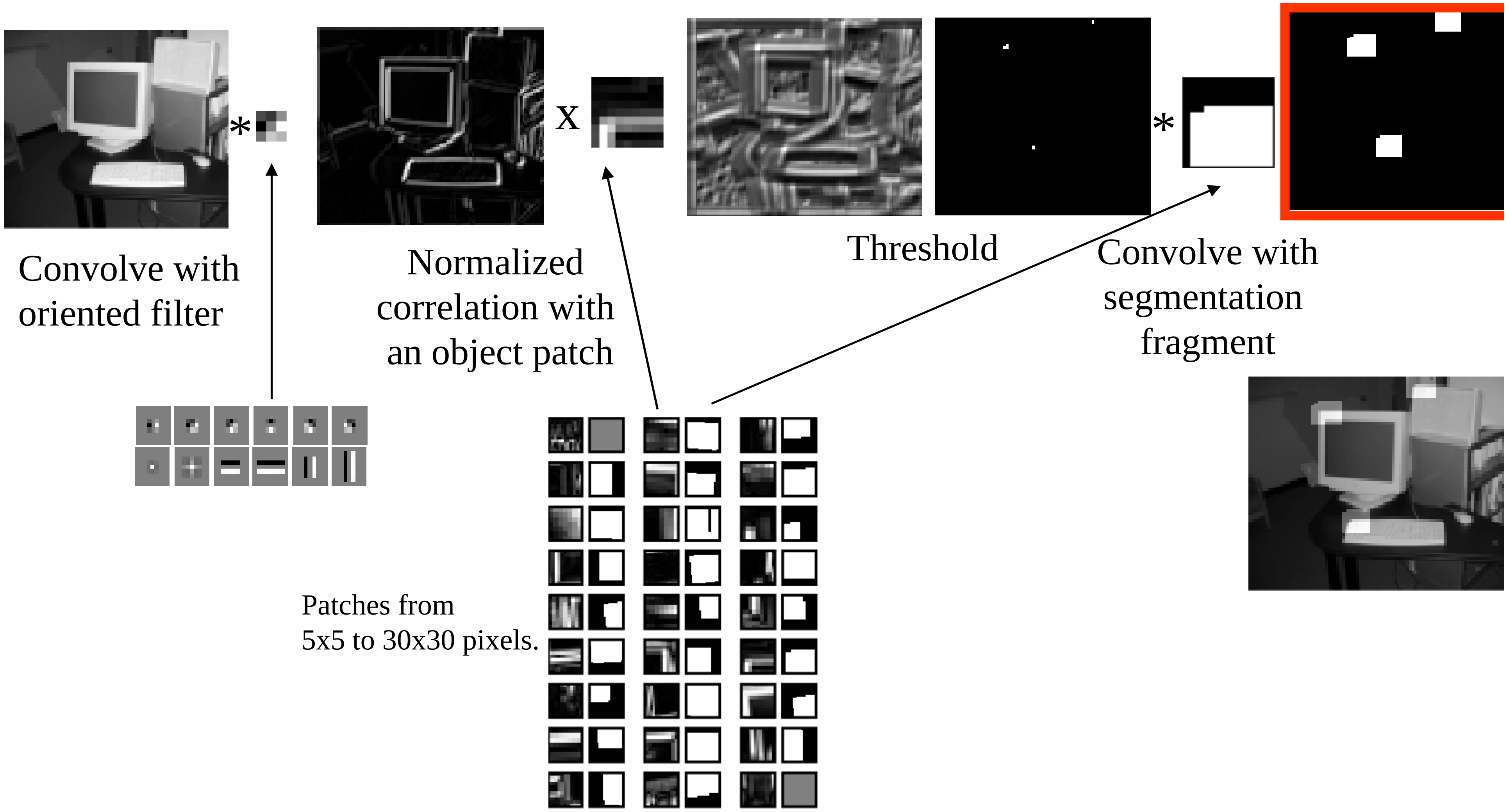
(intrinsic object features)



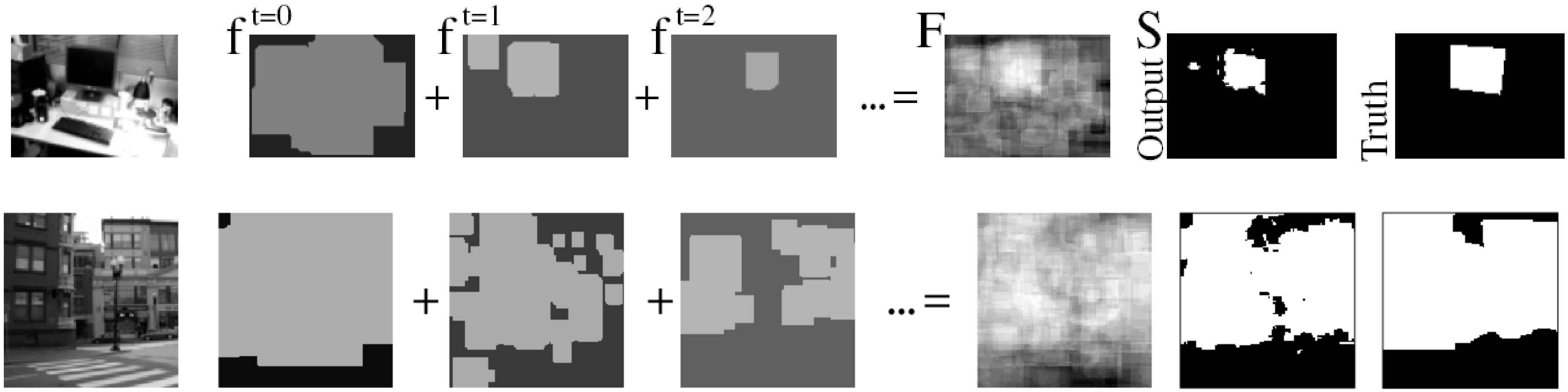
We maximize the probability of the true labels using Boosting.

Object local features

(Borenstein & Ullman, ECCV 02)

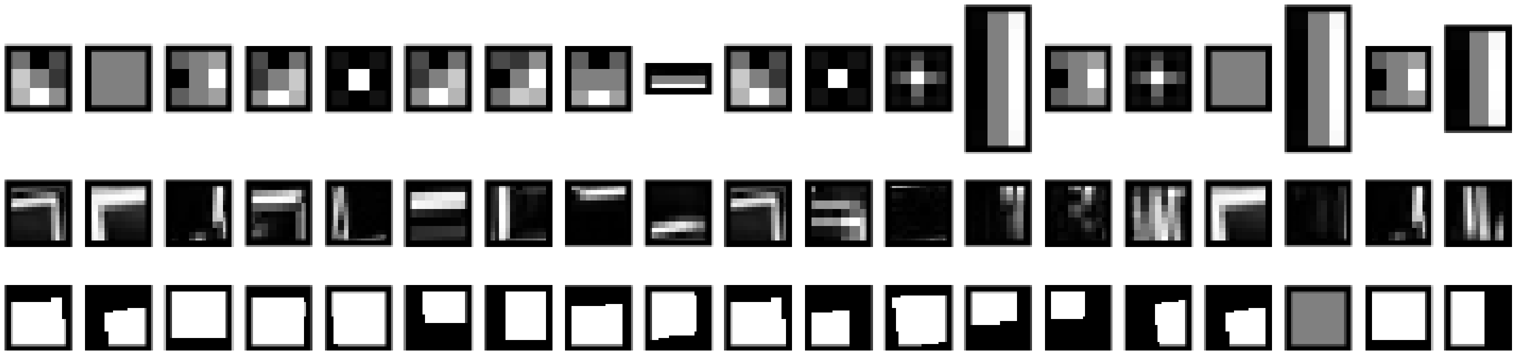


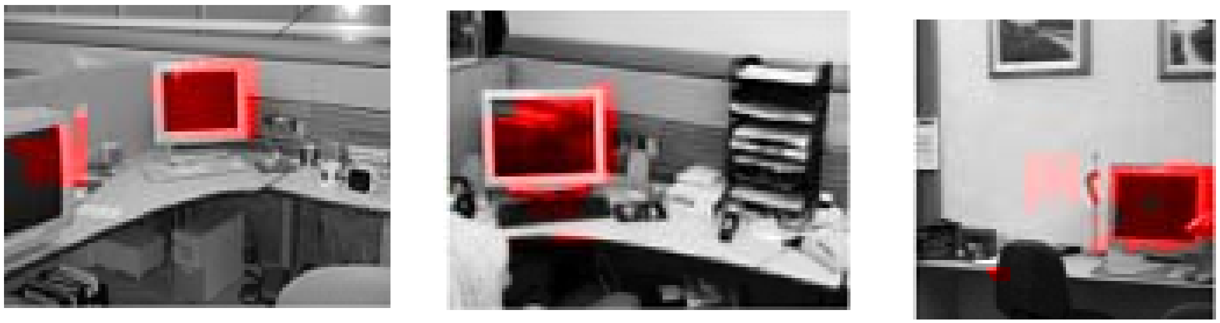
Results with local features



Results with local features

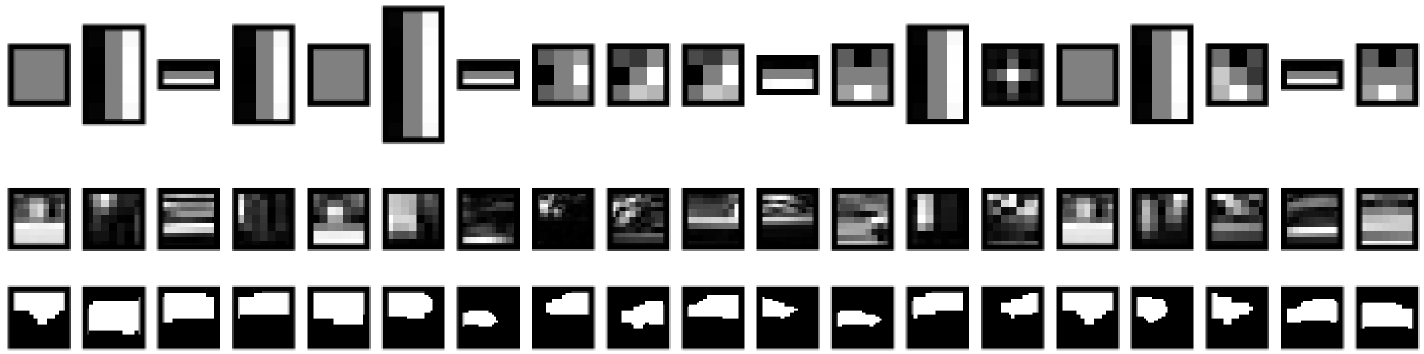
Screen

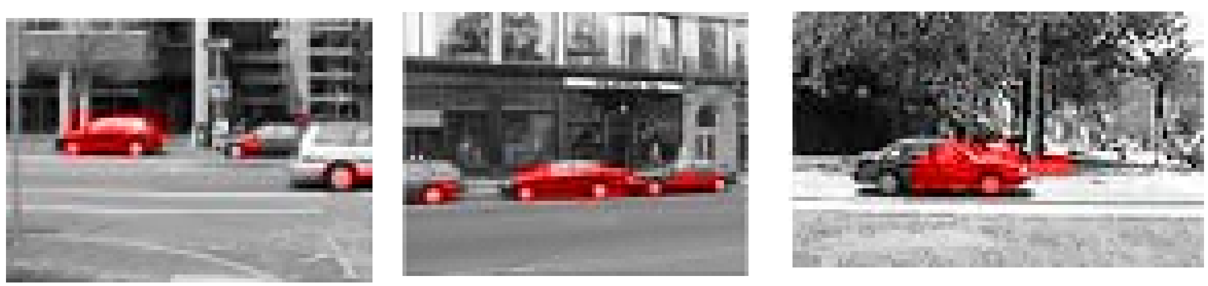




Results with local features

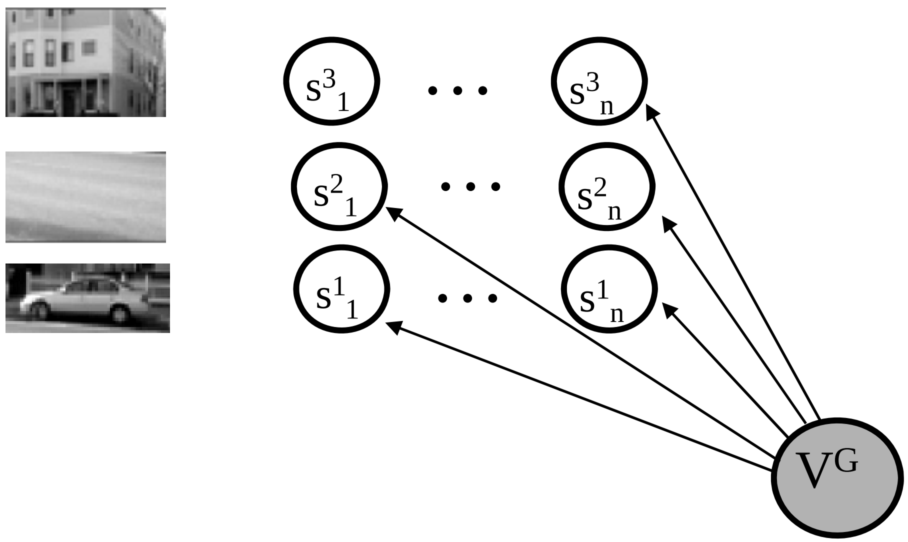
Car





Global context: location priming

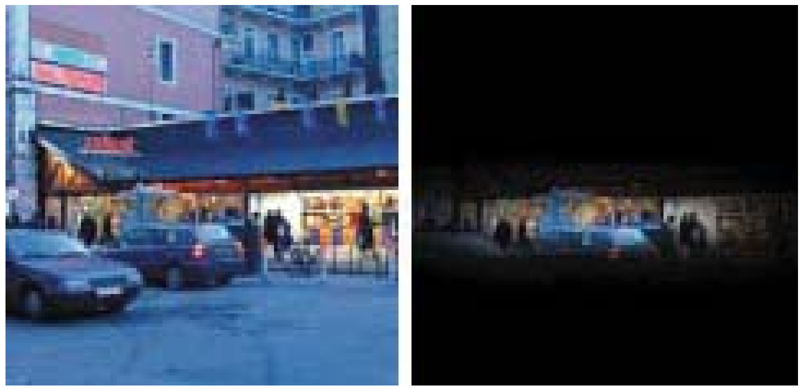
How far can we go without object detectors?



Context features that represent the scene instead of other objects.

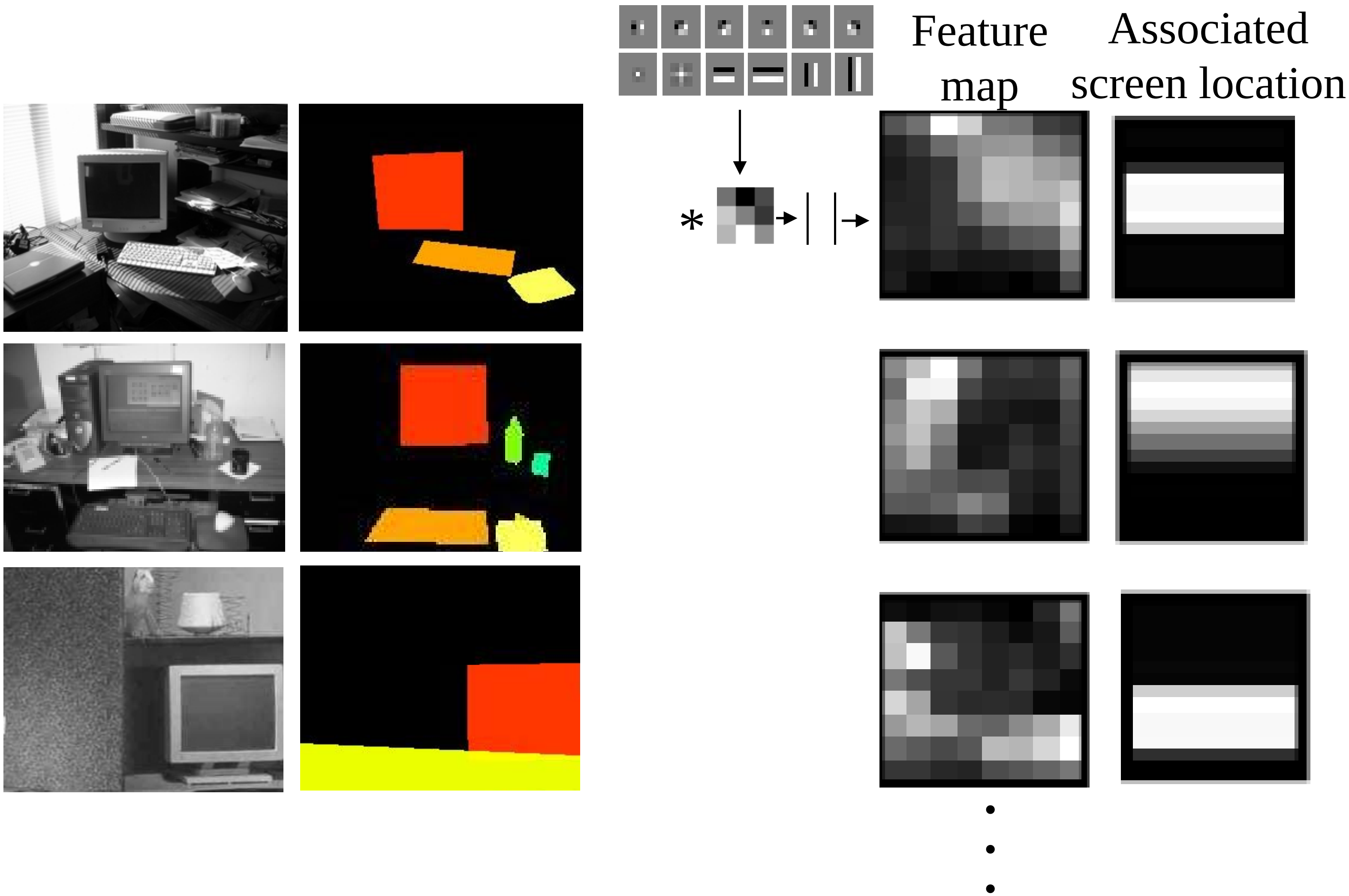
The global features can provide:

- Object presence
- Location priming
- Scale priming



Object global features

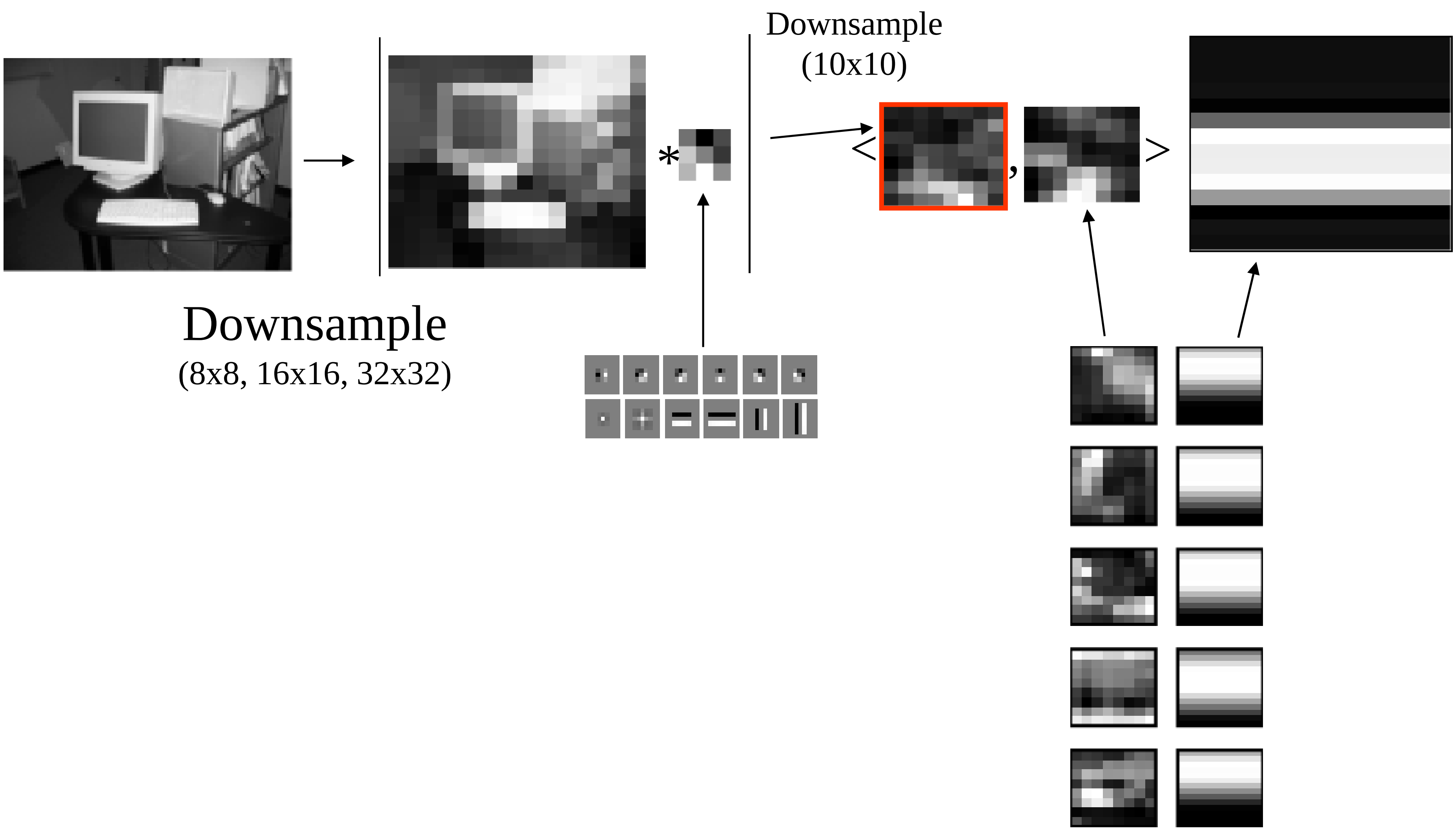
First we create a dictionary of scene features and object locations:



Only the vertical position of the object is well constrained by the global features

Object global features

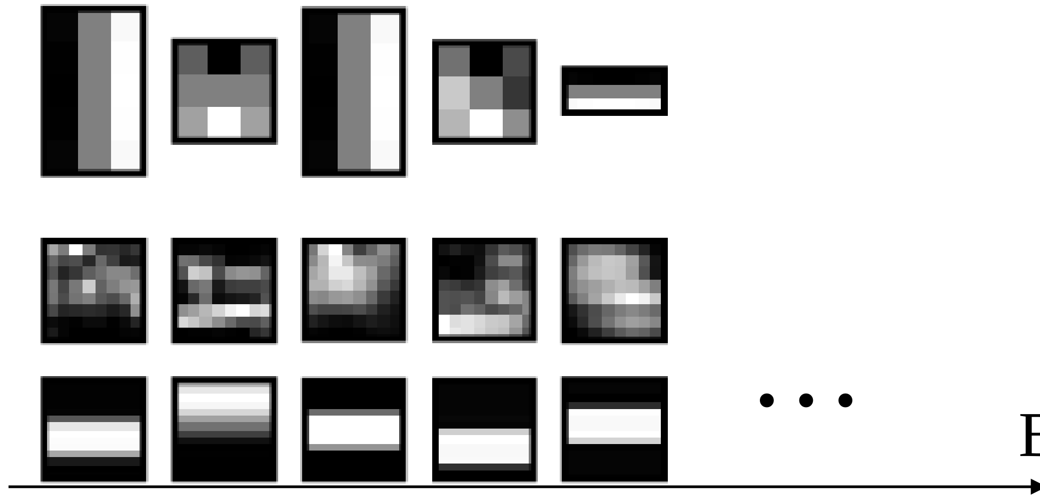
How to compute the global features



Car detection with global features

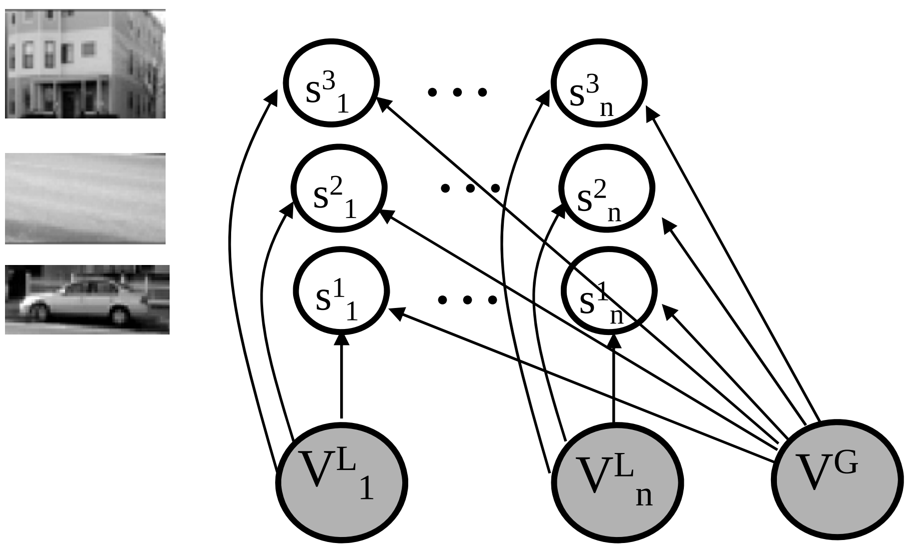
Features selected by boosting:

Car

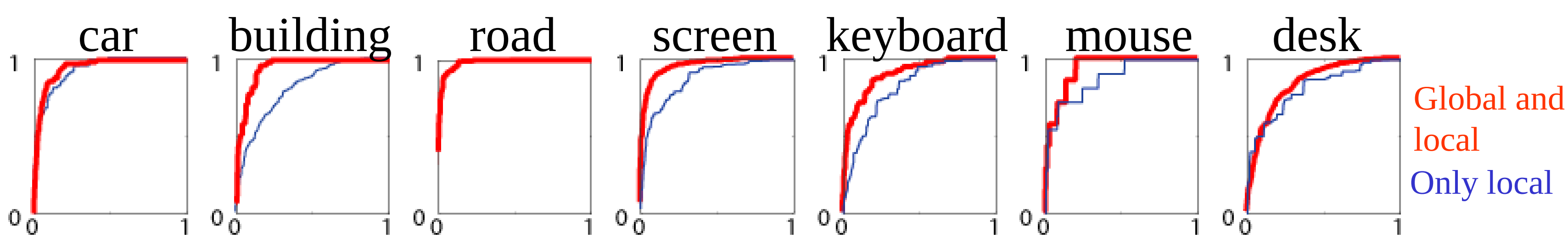


Boosting round

Combining global and local

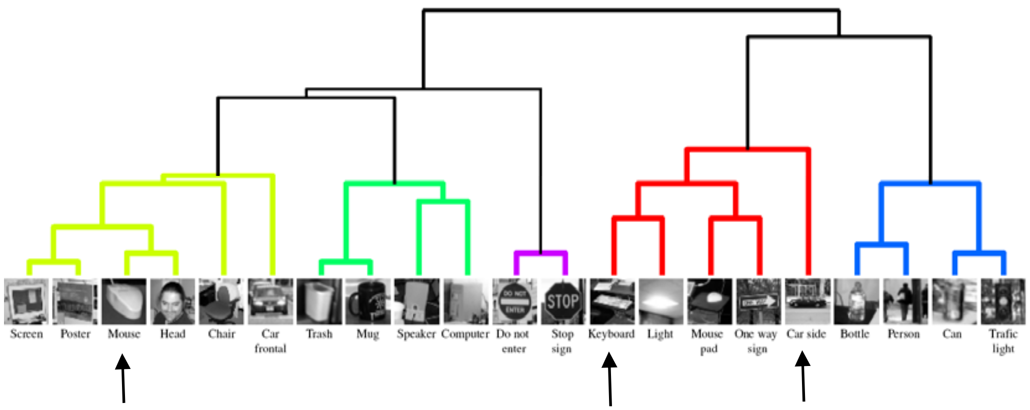


ROC for same total number of features (100 boosting rounds):

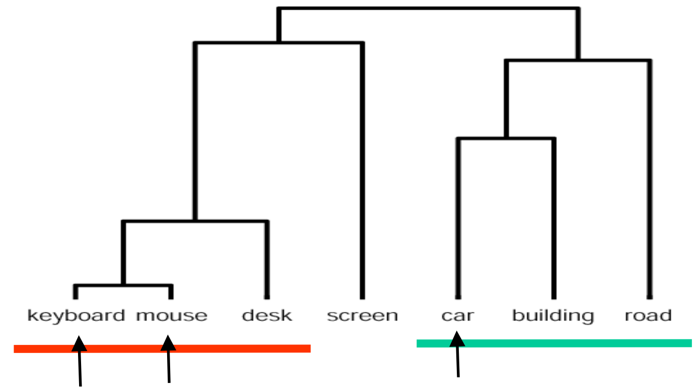


Clustering of objects with local and global feature sharing

Clustering with local features



Clustering with global and local features

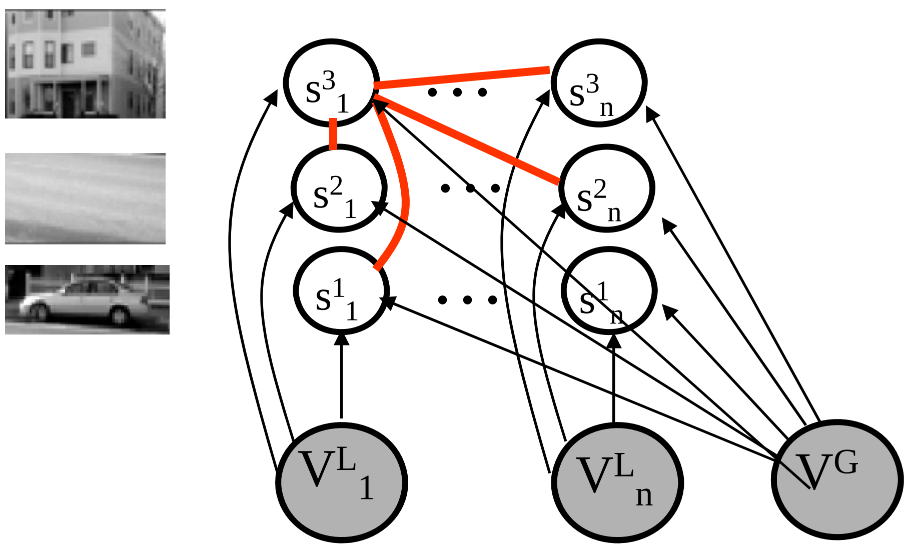


Objects are similar if they share local features and they appear in the same contexts.

Outline of this talk

- Use global image features (as well as local features) in boosting to help object detection
- Learn structure of dense CRF (with long range connections) using boosting, to exploit spatial correlations

Adding correlations between objects



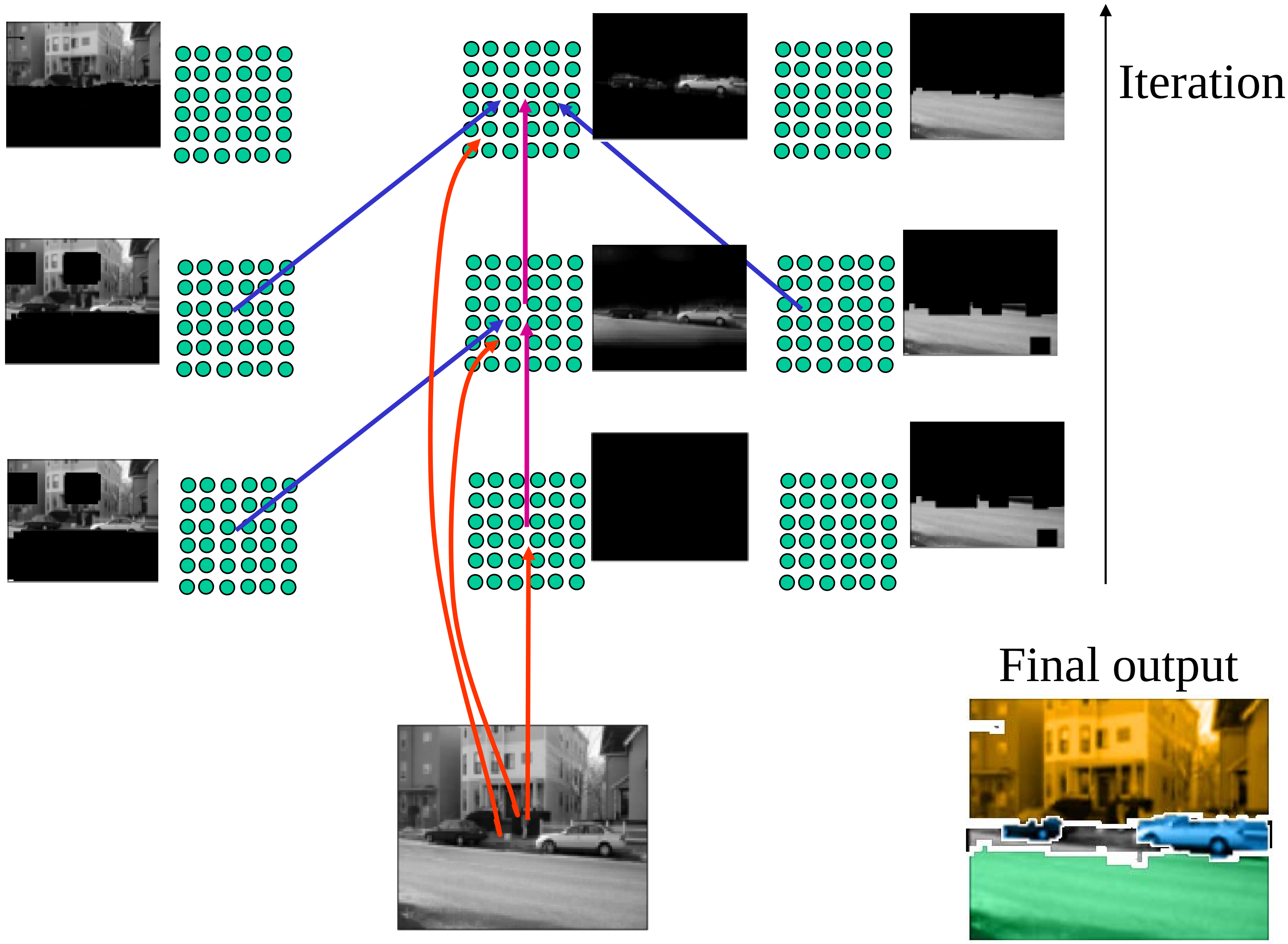
We need to learn

- The structure of the graph
- The pairwise potentials

Learning in CRFs

- Parameters
  - Lafferty, McCallum, Pereira (ICML 2001)
    - Find global optimum using gradient methods plus exact inference (forwards-backwards) in a chain
  - Kumar & Herbert, NIPS 2003
    - Use pseudo-likelihood in 2D CRF
  - Carbonetto, de Freitas & Barnard (04)
    - Use approximate inference (loopy BP) and pseudo-likelihood on 2D MRF
- Structure
  - He, Zemel & Carreira-Perpinan (CVPR 04)
    - Use contrastive divergence
  - Torralba, Murphy, Freeman (NIPS 04)
    - Use boosting

Sequentially learning the structure

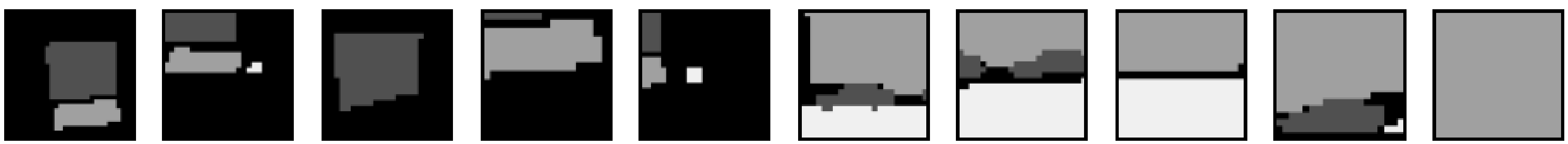


Sequentially learning the structure

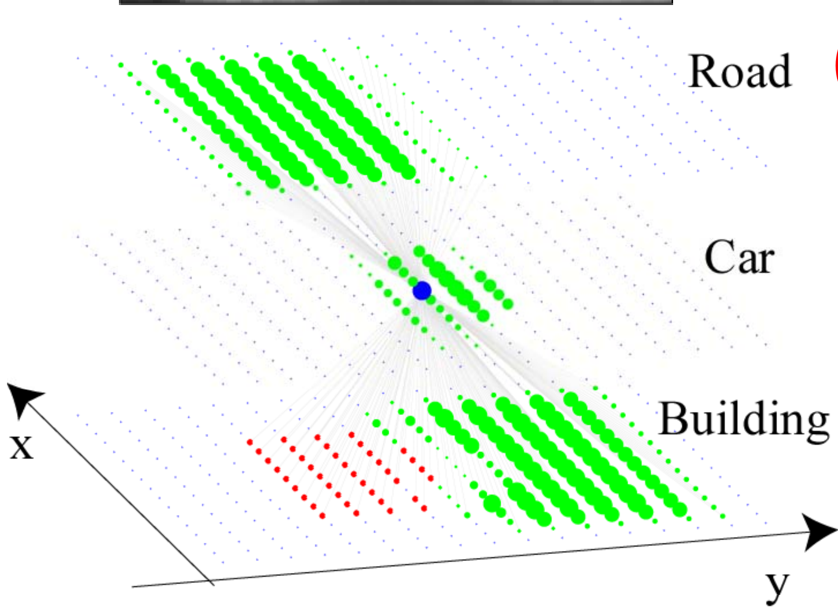
At each iteration of boosting

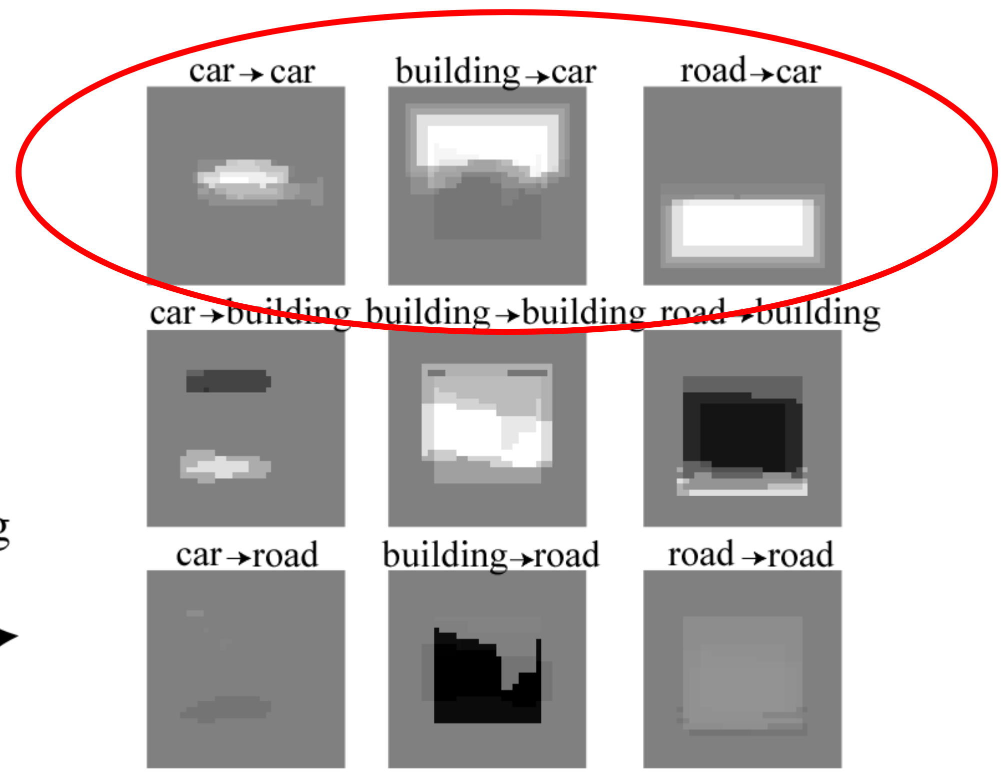
•We pick a weak learner applied to the image
(local or global features)

•We pick a weak learner applied to a subset of the label-beliefs at the previous iteration. These subsets are chosen from a dictionary of labeled graph fragments from the training set.



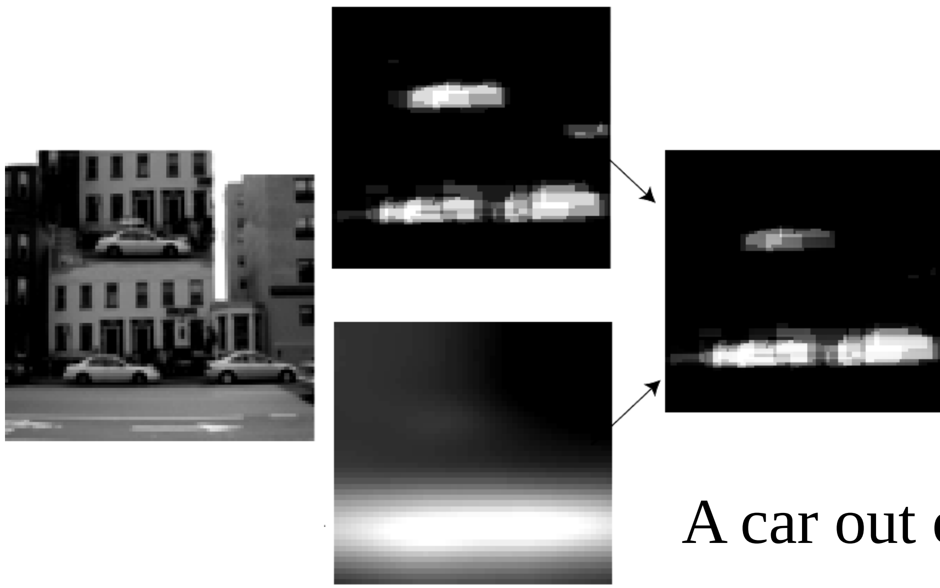
Car detection



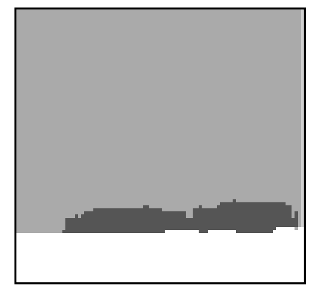


Car detection

From intrinsic features

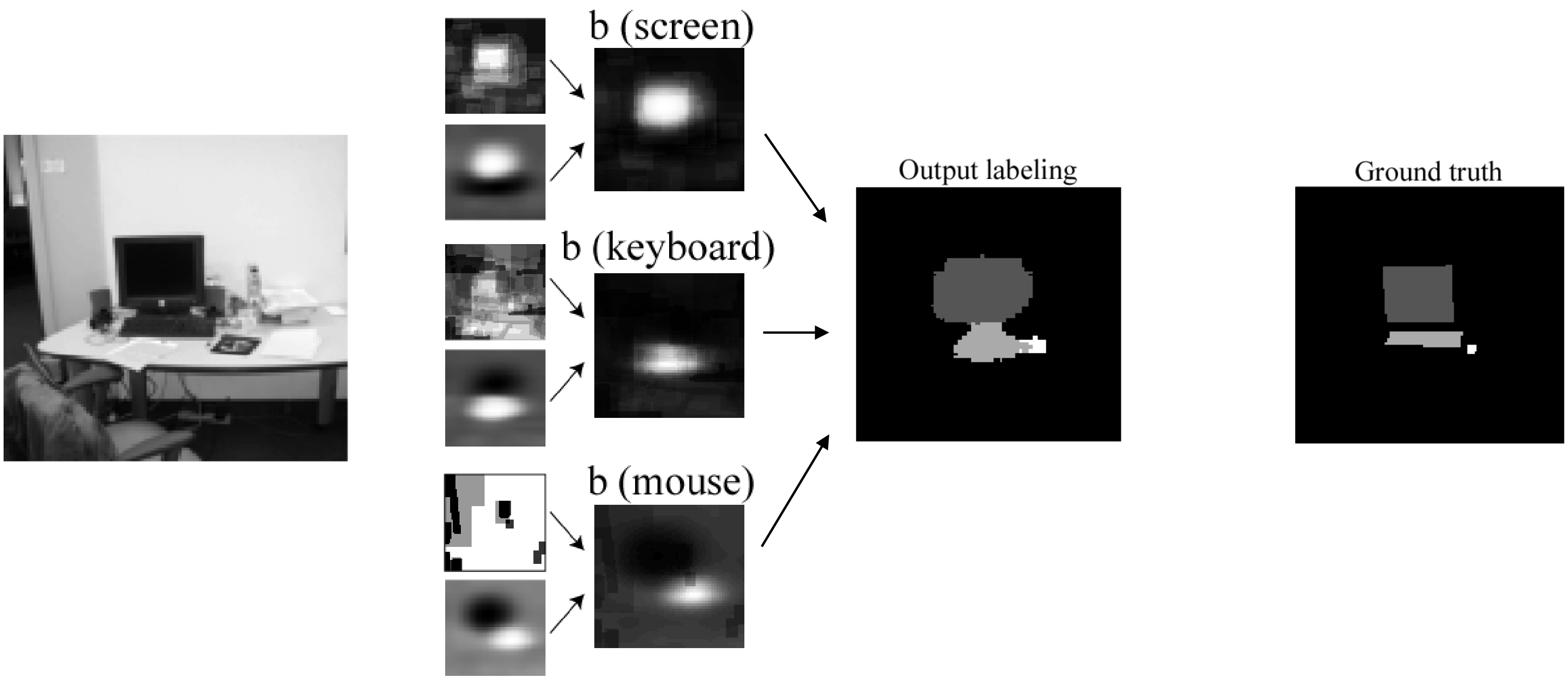


From contextual features



A car out of context is less of a car

Screen/keyboard/mouse



Cascade

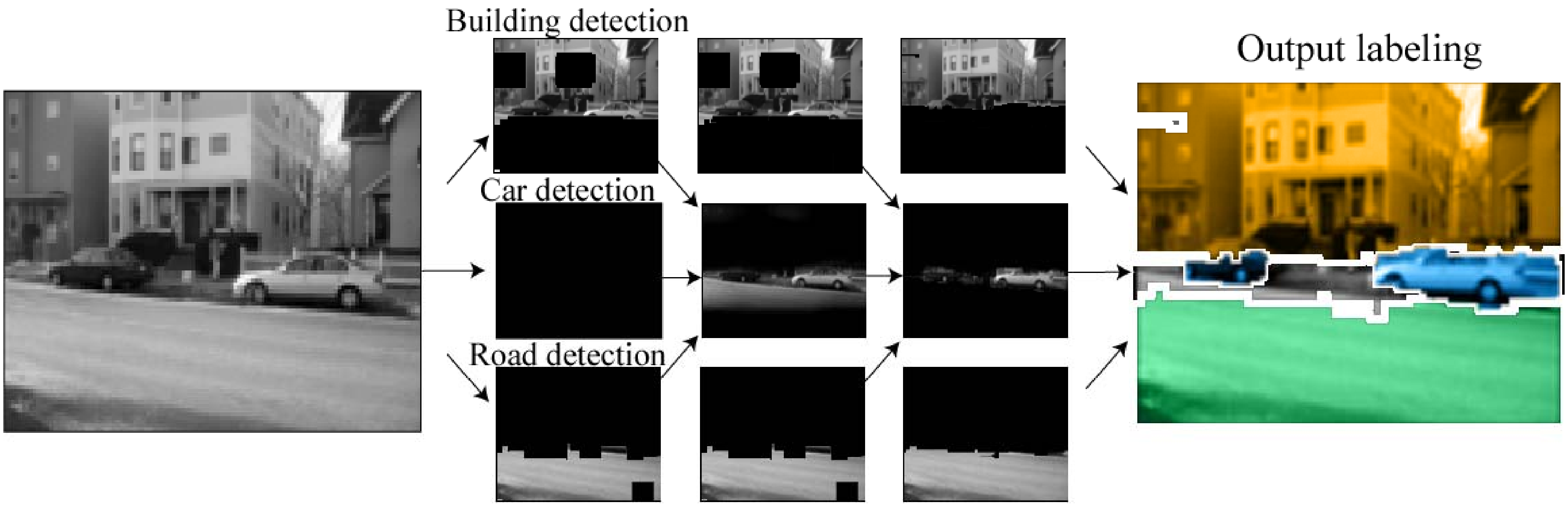
Viola & Jones (2001)

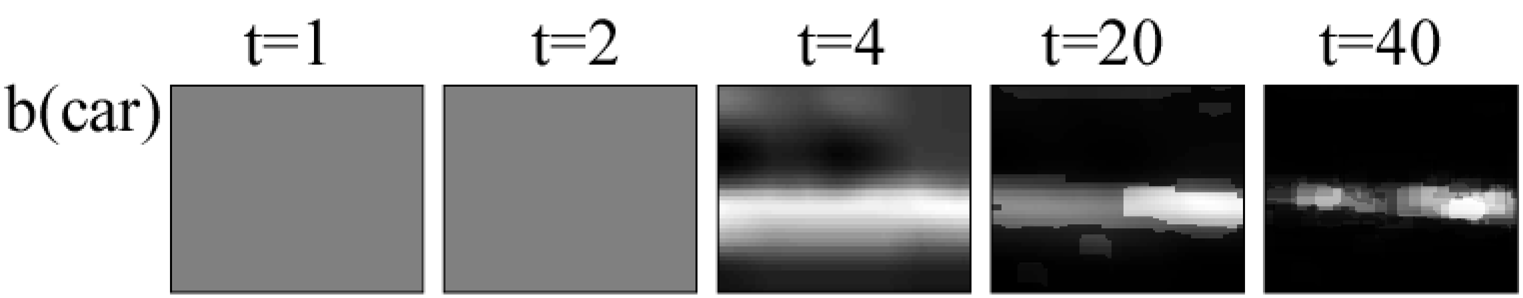
Set to zero the beliefs of nodes with low probability of containing the target.

Perform message passing only on undecided nodes

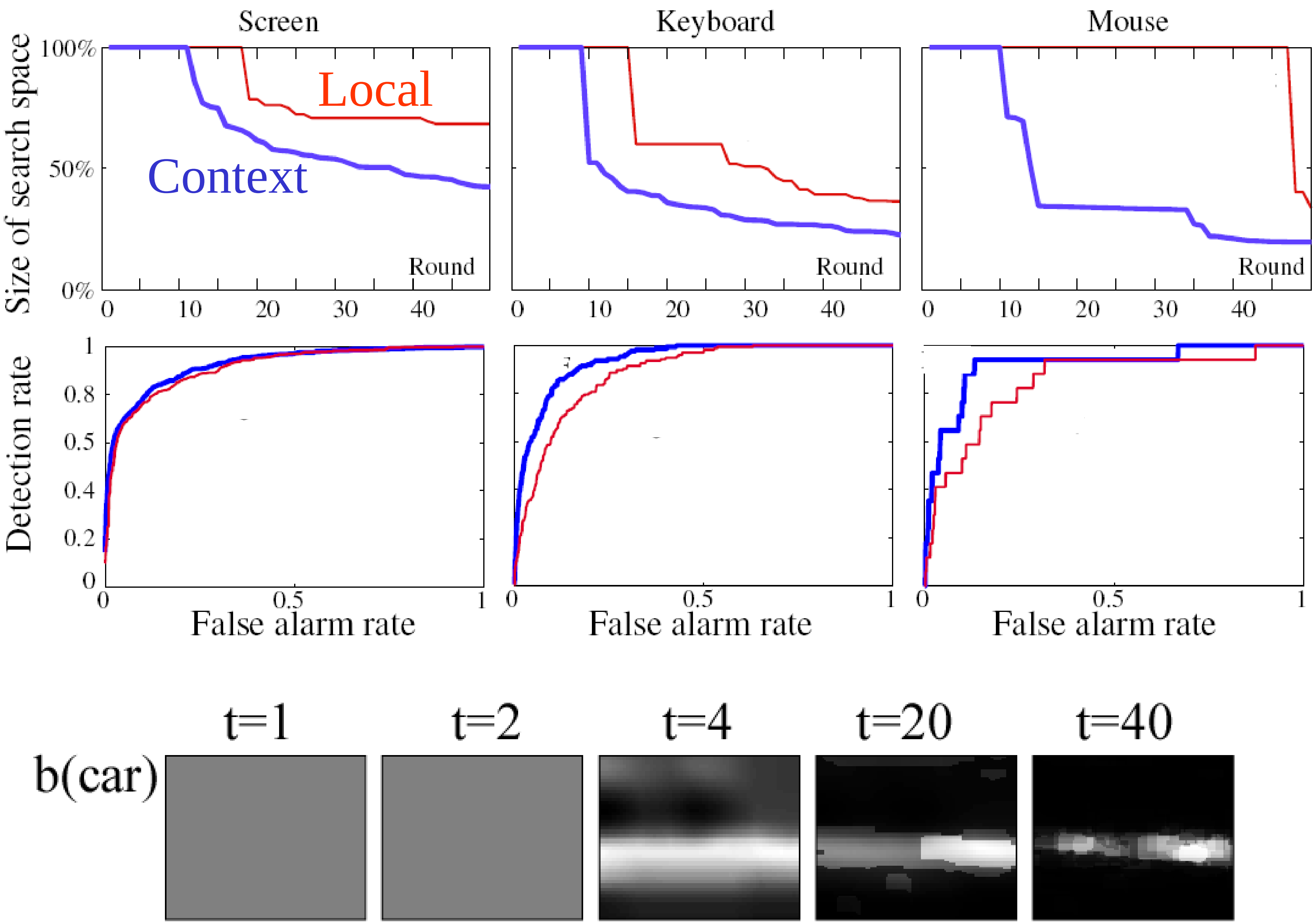
The detection of the screen reduces the search space for the mouse detector.

Cascade


Building detection
Car detection
Road detection
Output labeling


t=1
t=2
t=4
t=20
t=40
b(car)

Cascade


Screen
Keyboard
Mouse
Size of search space
100%
50%
0%
Local
Context
Round
0
10
20
30
40
Detection rate
1
0.8
0.5
0.4
0.2
0
0.5
False alarm rate
t=1
t=2
t=4
t=20
t=40
b(car)

Putting Objects in Perspective

Derek Hoiem

Alexei A. Efros

Martial Hebert

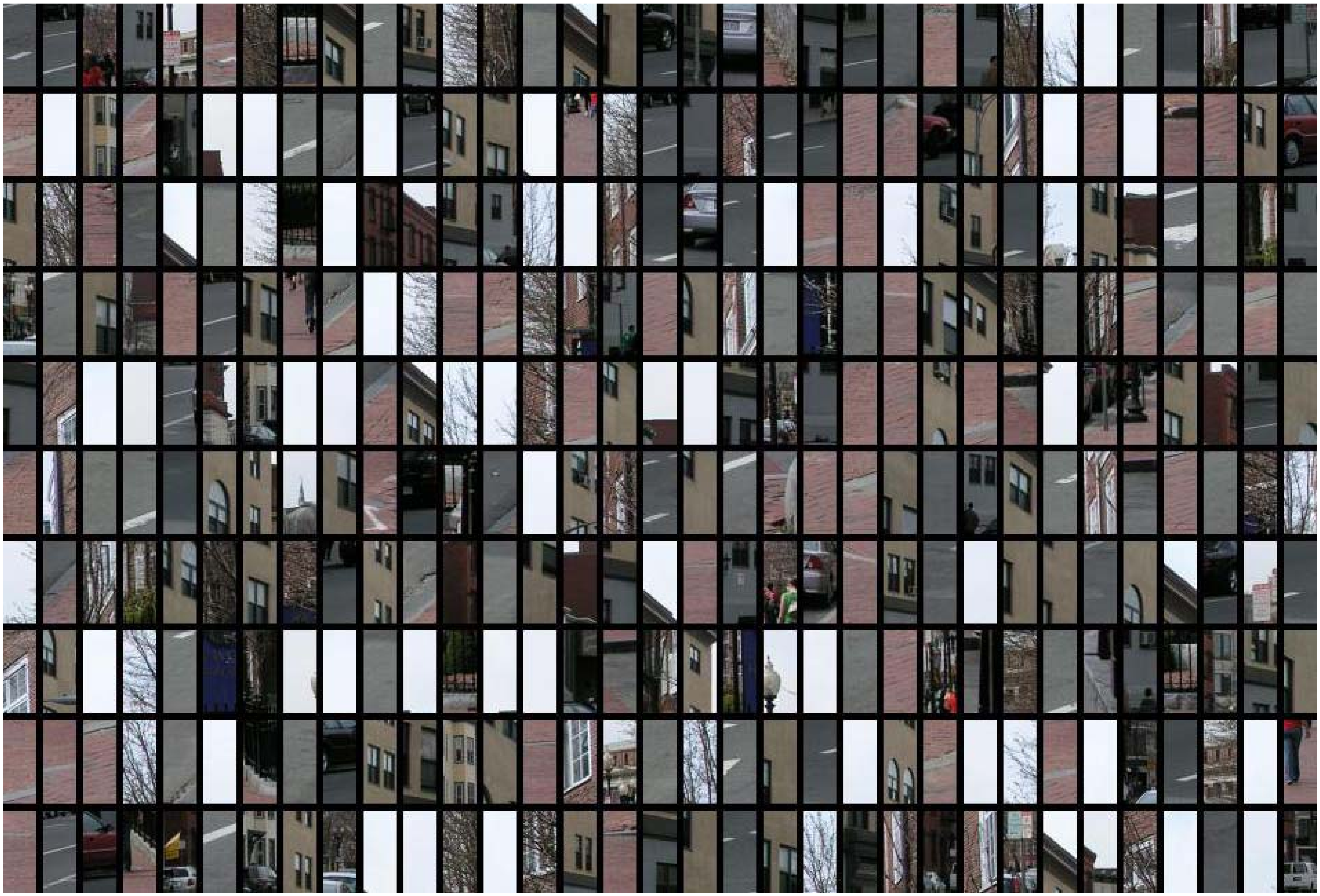
Carnegie Mellon University

Robotics Institute

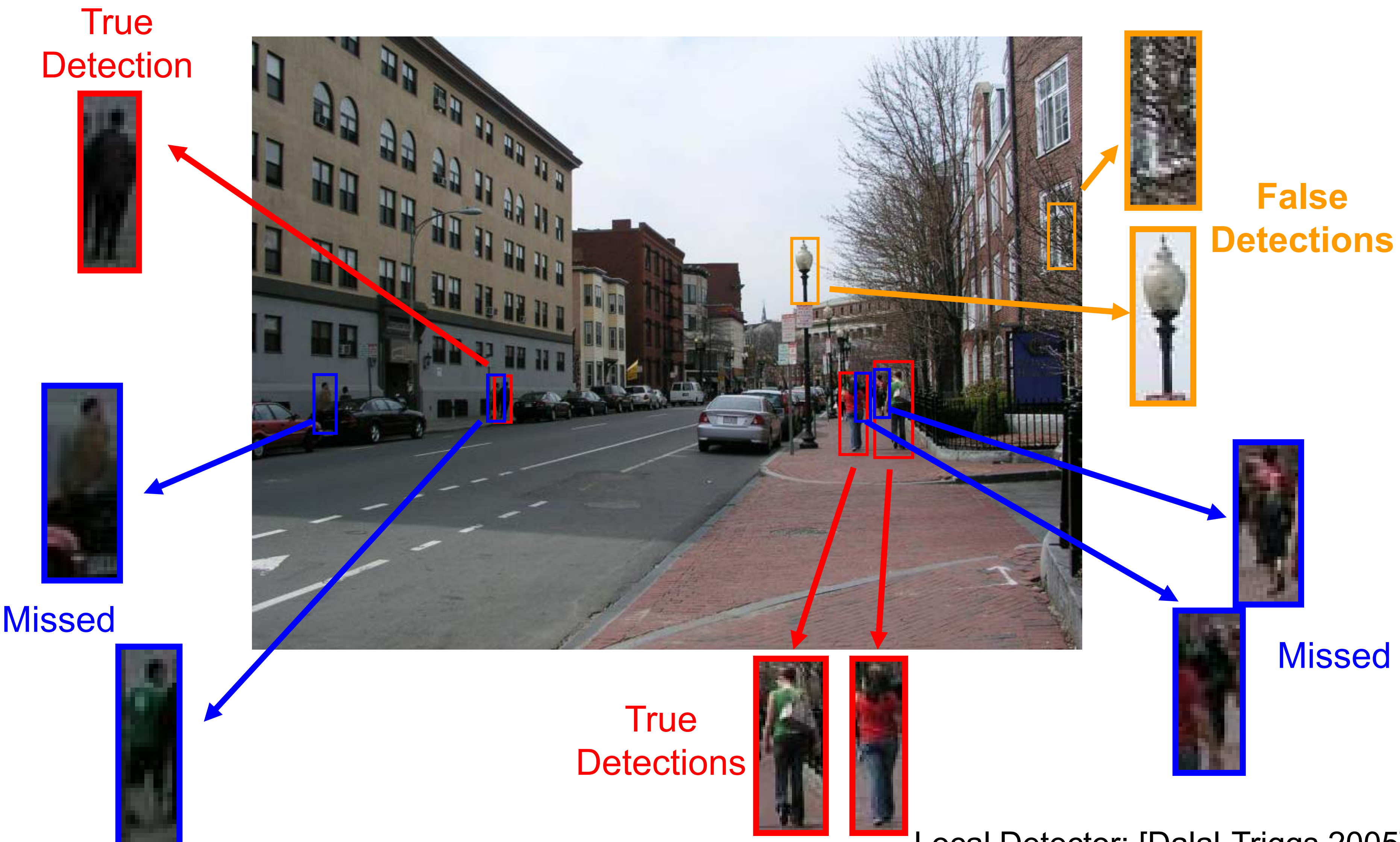
Understanding an Image

Today: Local and Independent

What the Detector Sees



Local Object Detection



Local Detector: [Dalal-Triggs 2005]

Work in Context

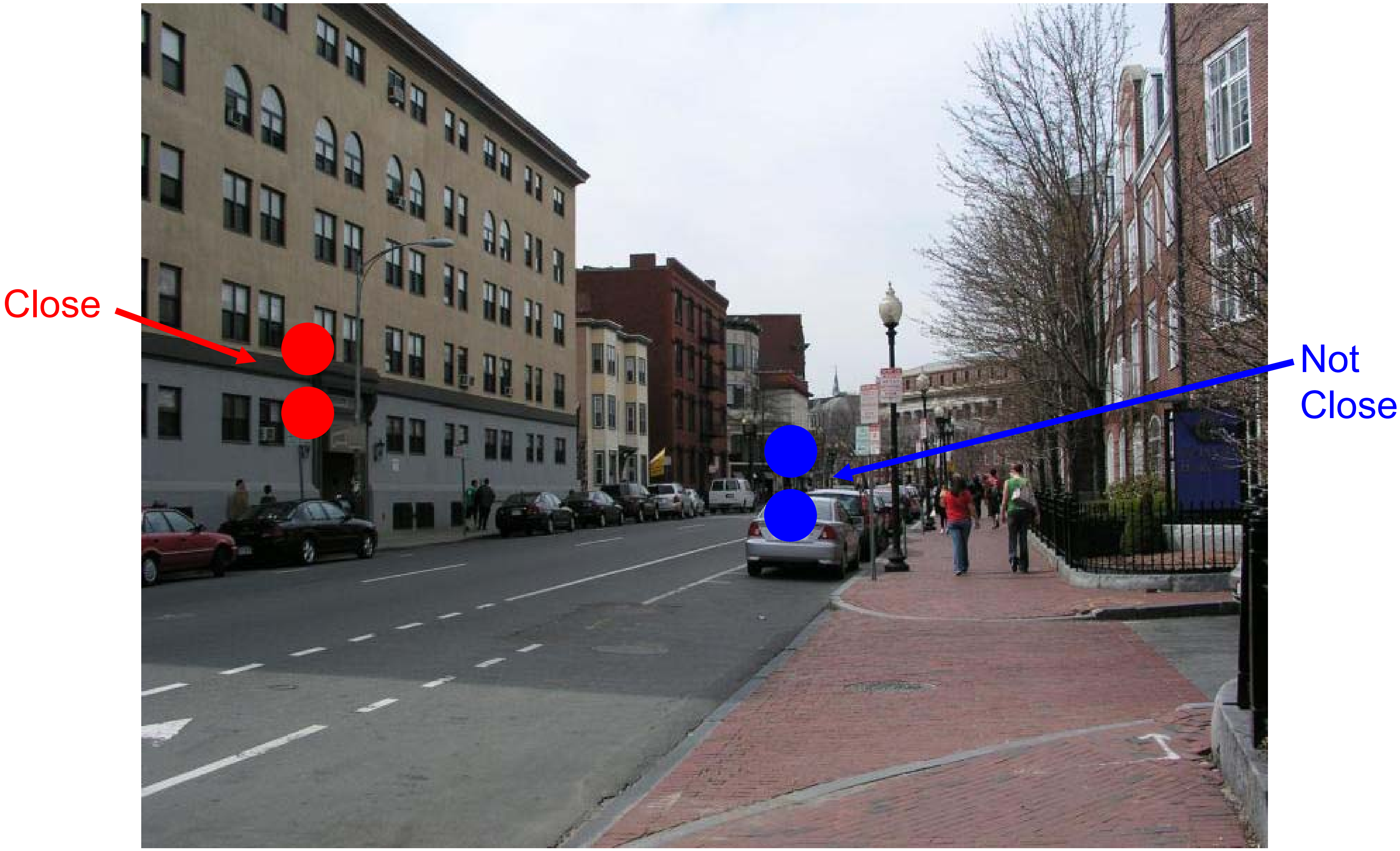
- Image understanding in the 70's

Guzman (*SEE*) 1968
Hansen & Riseman (*VISIONS*) 1978
Barrow & Tenenbaum 1978
Yakimovsky & Feldman 1973
Brooks (*ACRONYM*) 1979
Marr 1982
Ohta & Kanade 1973

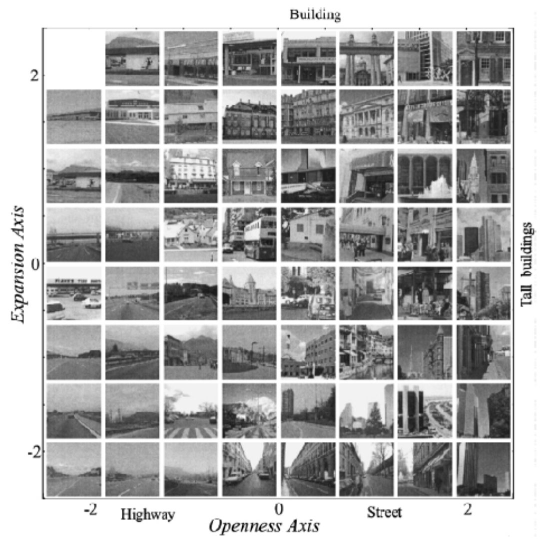
- Recent work in 2D context

Kumar & Hebert 2005
Torralba, Murphy, Freeman 2004
Fink & Perona 2003
He, Zemel, Cerreira-Perpiñán 2004
Carbonetto, Freitas, Banard 2004
Winn & Shotton 2006

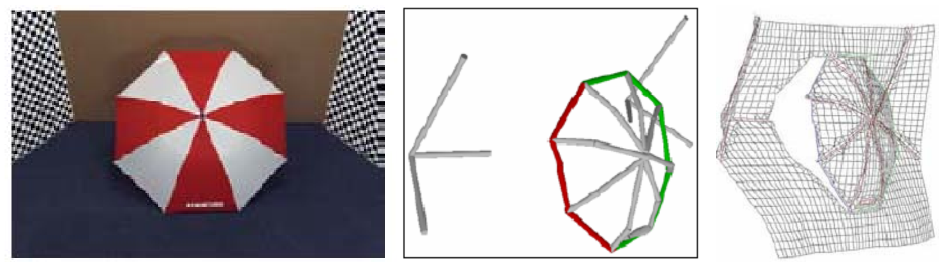
Real Relationships are 3D



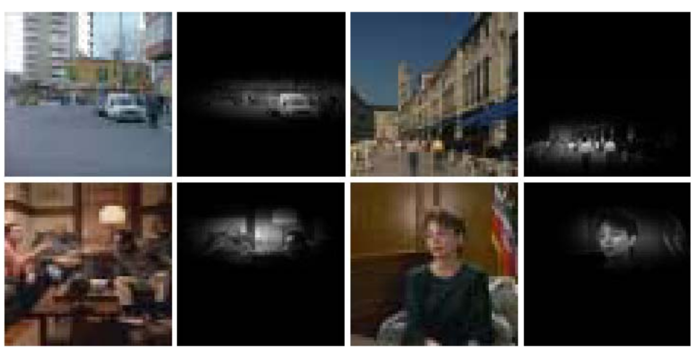
Recent Work in 3D



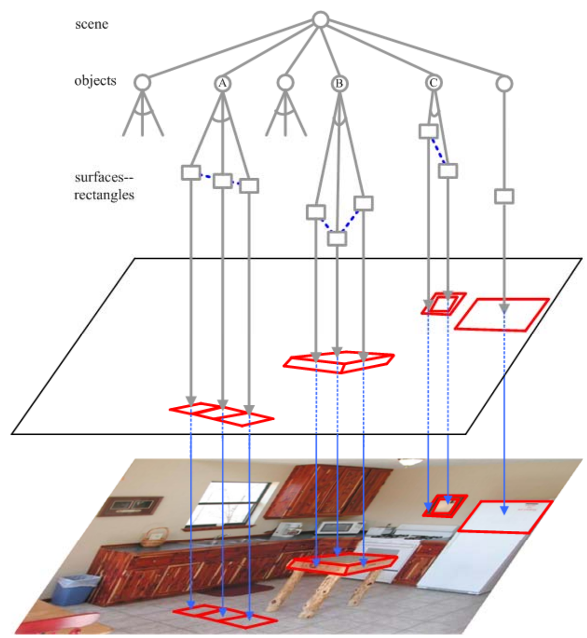
[Oliva & Torralba 2001]



[Han & Zu 2003]

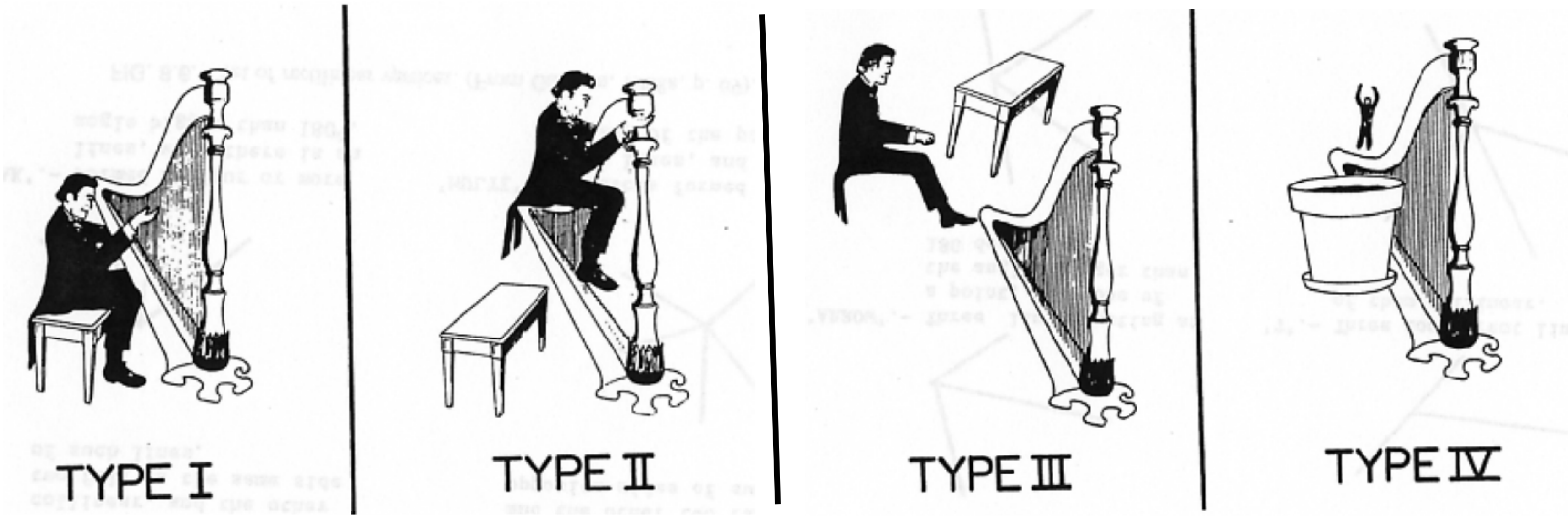


[Torralba, Murphy & Freeman 2003]



[Han & Zu 2005]

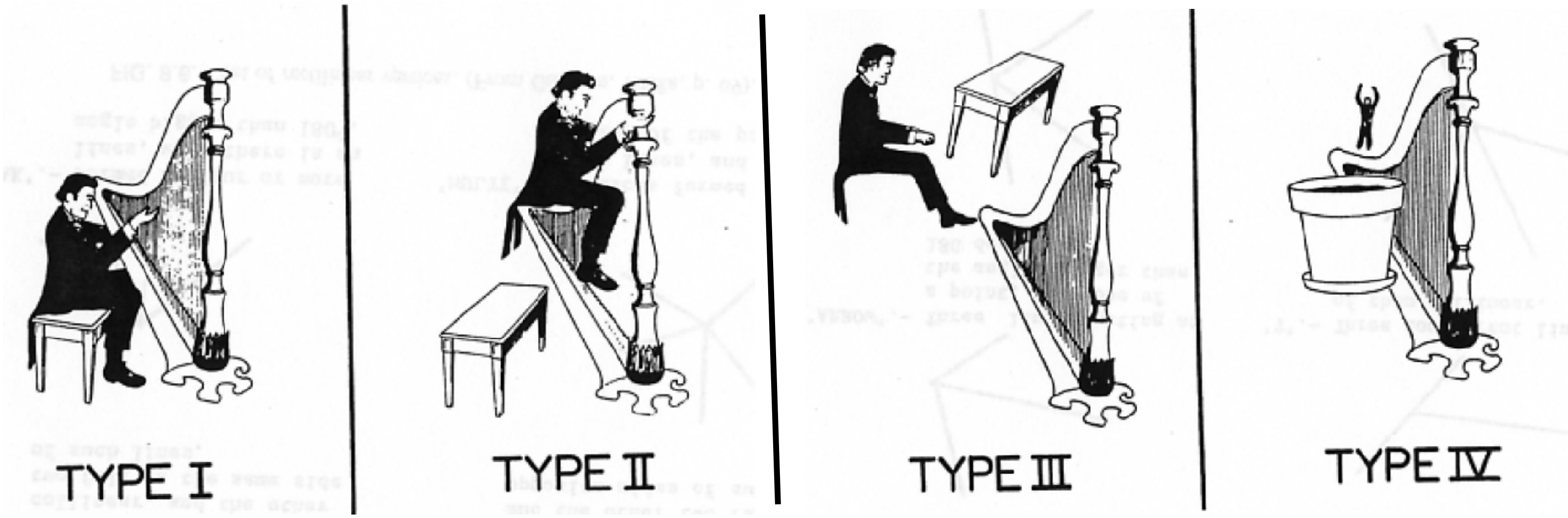
Objects and Scenes



Hock, Romanski, Galie, & Williams 1978

- Biederman's Relations among Objects in a Well-Formed Scene (1981):
  - Support
  - Size
  - Position
  - Interposition
  - Likelihood of Appearance

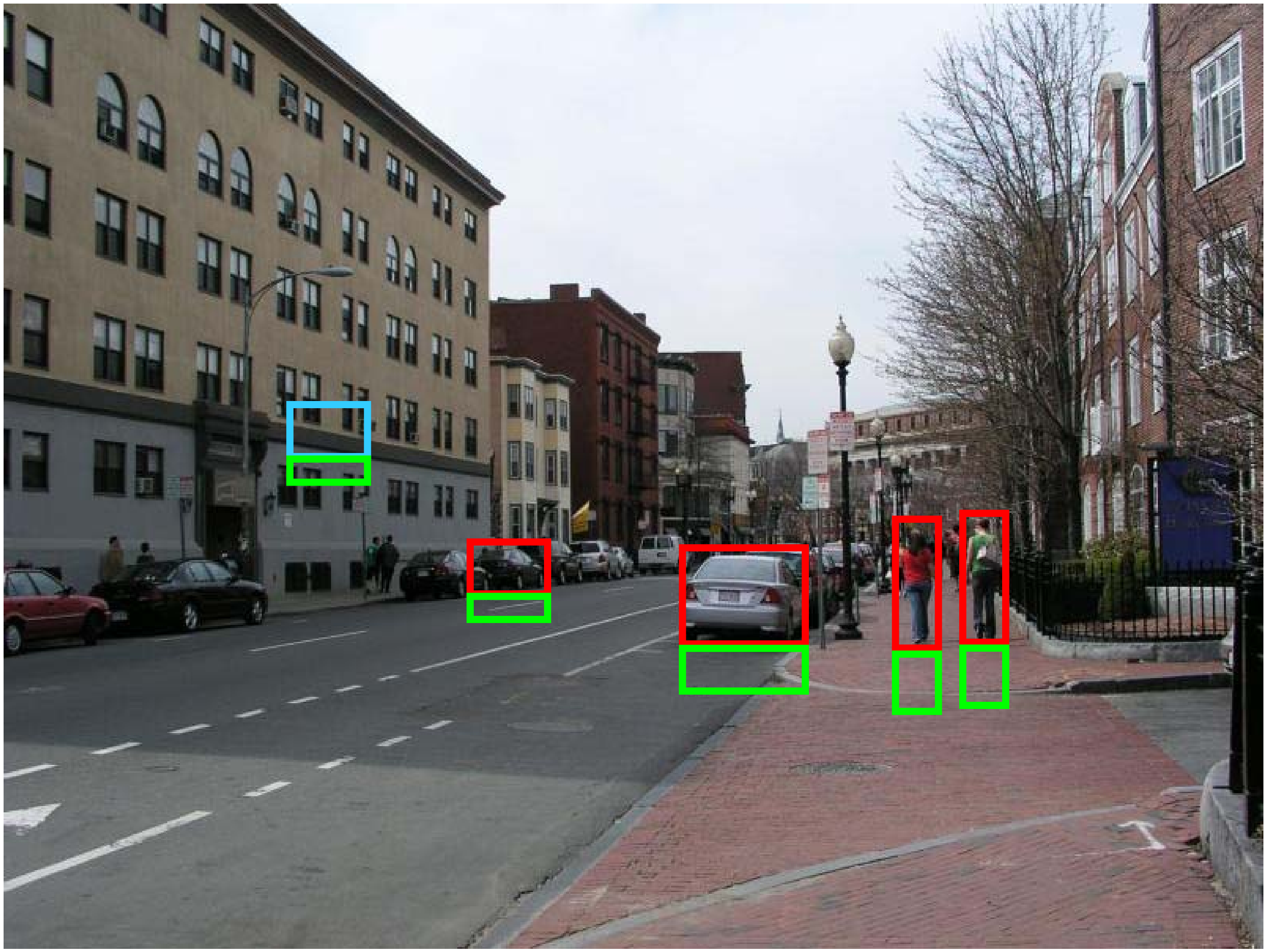
Contribution of this Paper



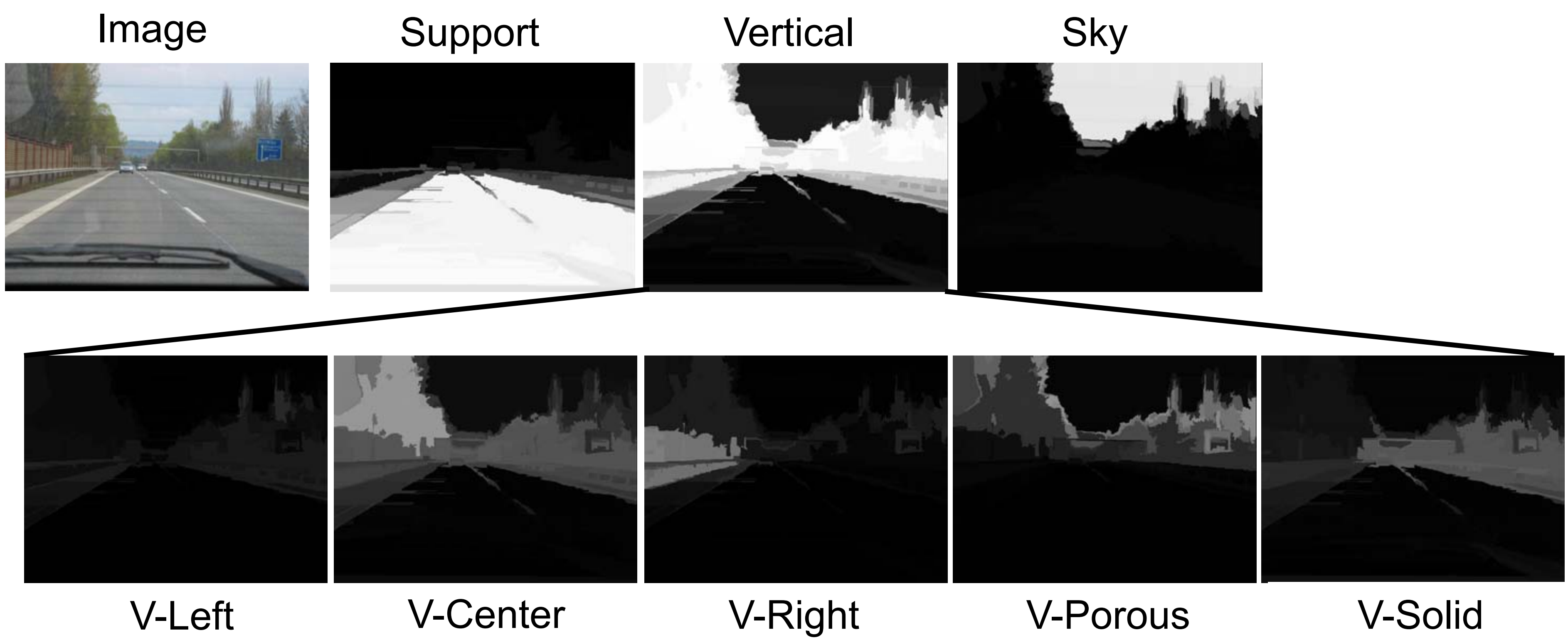
Hock, Romanski, Galie, & Williams 1978

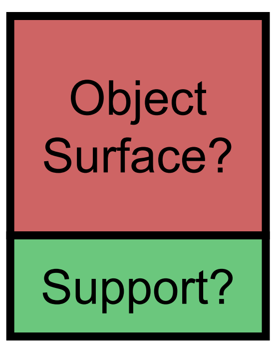
- Biederman's Relations among Objects in a Well-Formed Scene (1981):
  - Support
  - Size
  - Position
  - Interposition
  - Likelihood of Appearance

Object Support



Surface Estimation

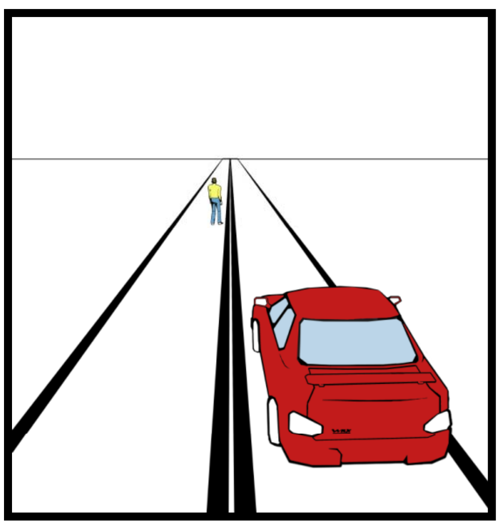




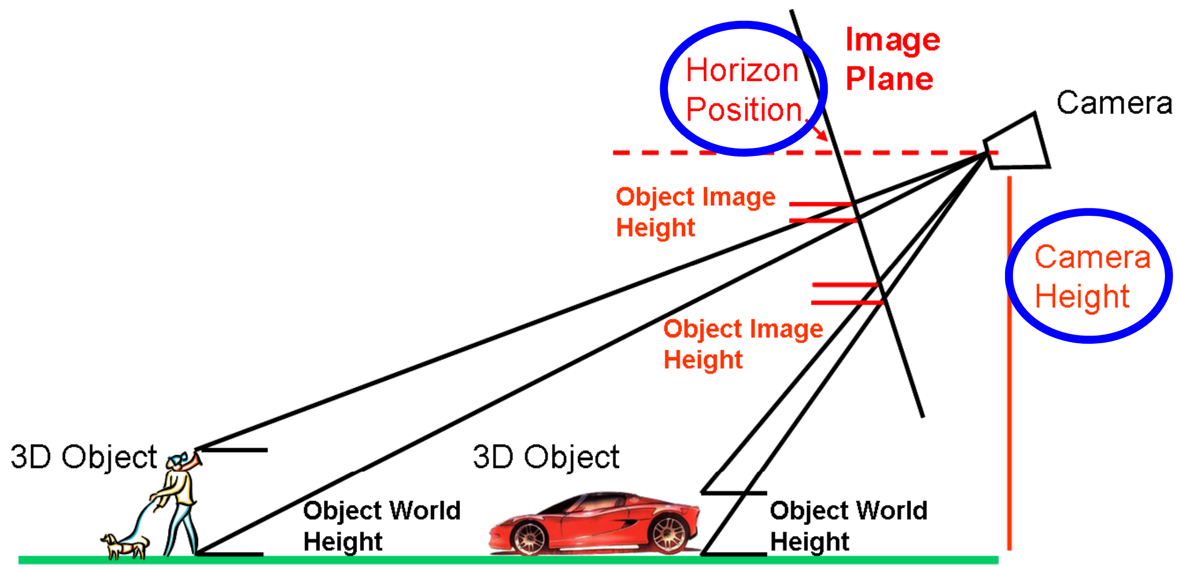
[Hoiem, Efros, Hebert ICCV 2005]

Software available online

Object Size in the Image



Image



World

Object Size ↔ Camera Viewpoint

Input Image

Loose Viewpoint Prior

Object Size ↔ Camera Viewpoint

Input Image

Loose Viewpoint Prior

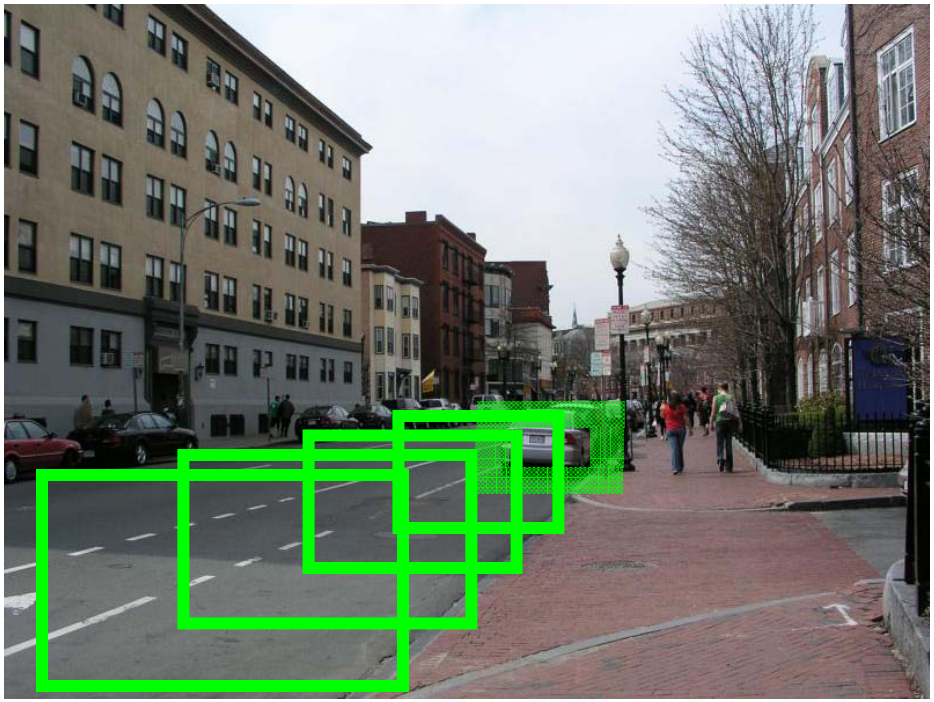
Object Size ↔ Camera Viewpoint

Object Position/Sizes

Viewpoint

Object Size ↔ Camera Viewpoint

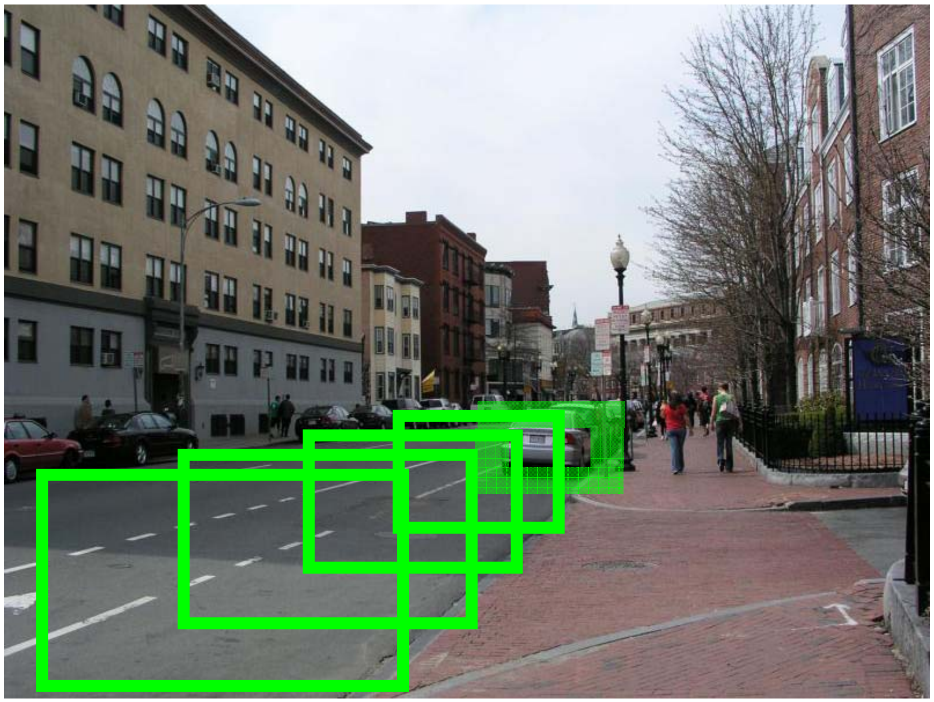
Object Position/Sizes



Viewpoint

Object Size ↔ Camera Viewpoint

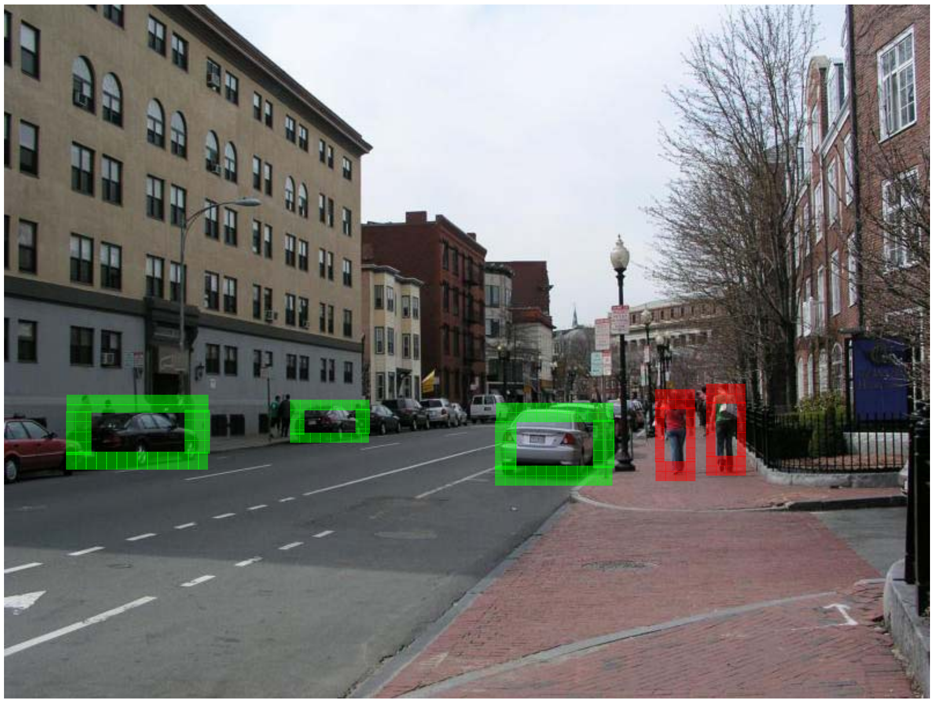
Object Position/Sizes



Viewpoint

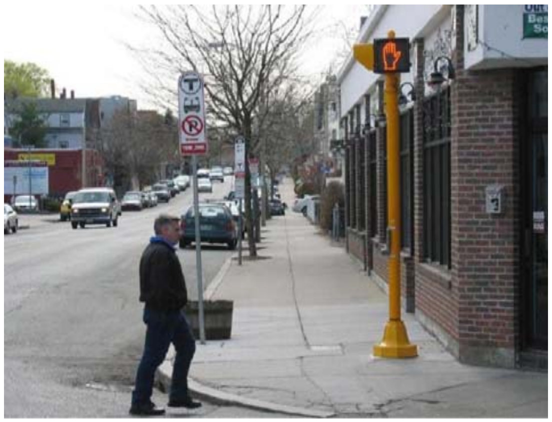
Object Size ↔ Camera Viewpoint

Object Position/Sizes

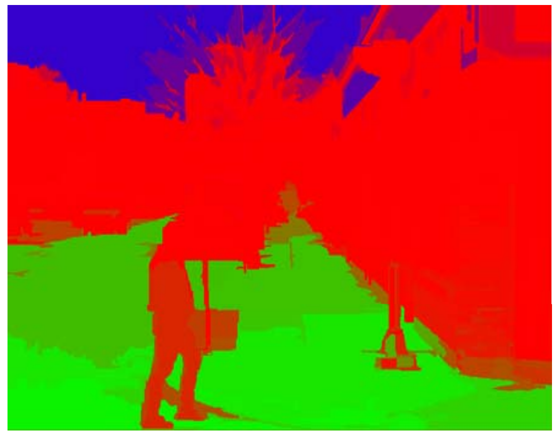


Viewpoint

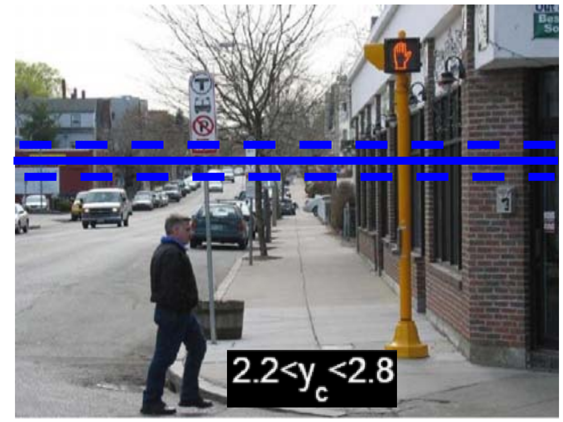
What does surface and viewpoint say about objects?



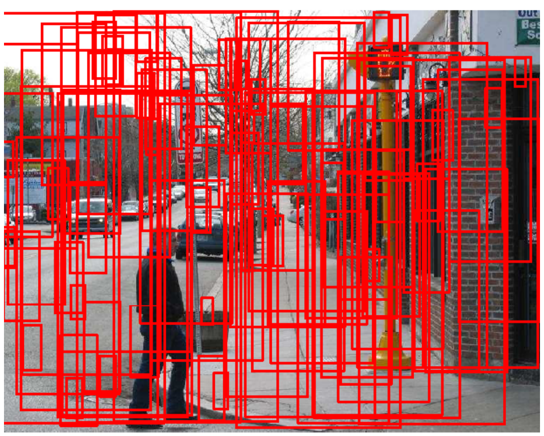
Image



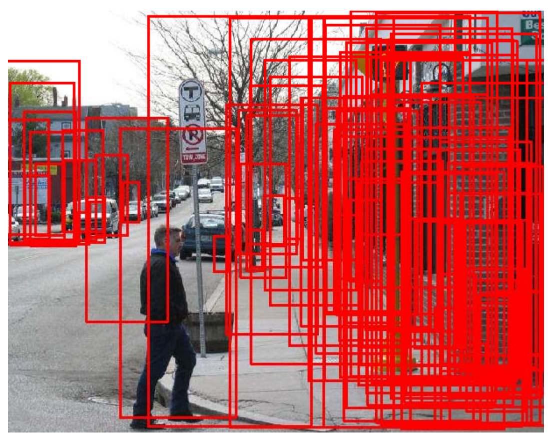
P(surfaces)



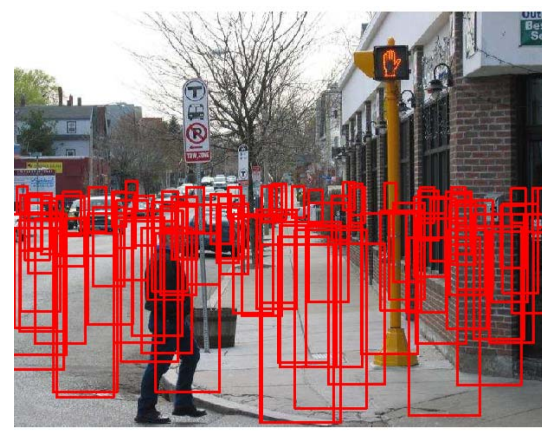
P(viewpoint)



P(object)

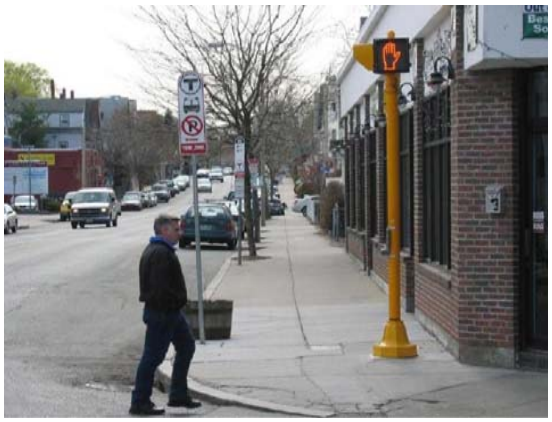


P(object | surfaces)



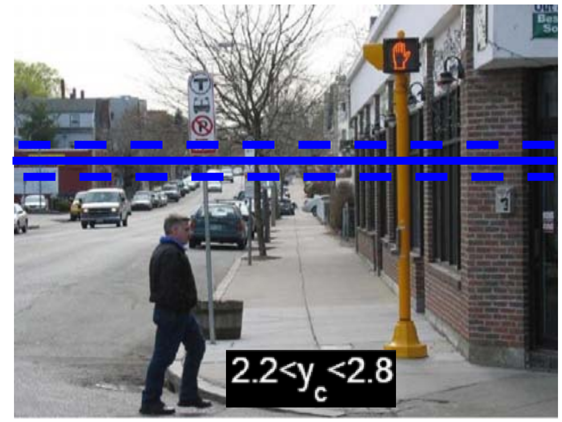
P(object | viewpoint)

What does surface and viewpoint say about objects?

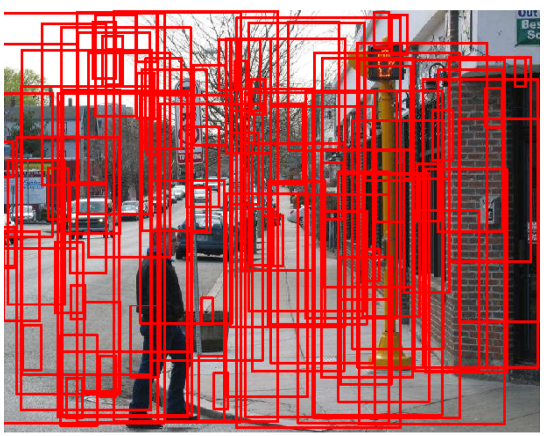


Image

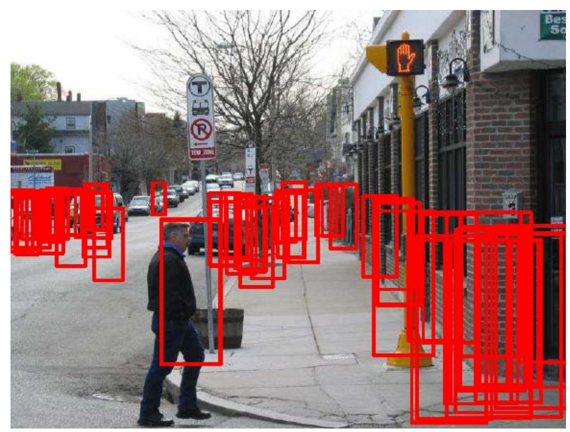
P(surfaces)



P(viewpoint)

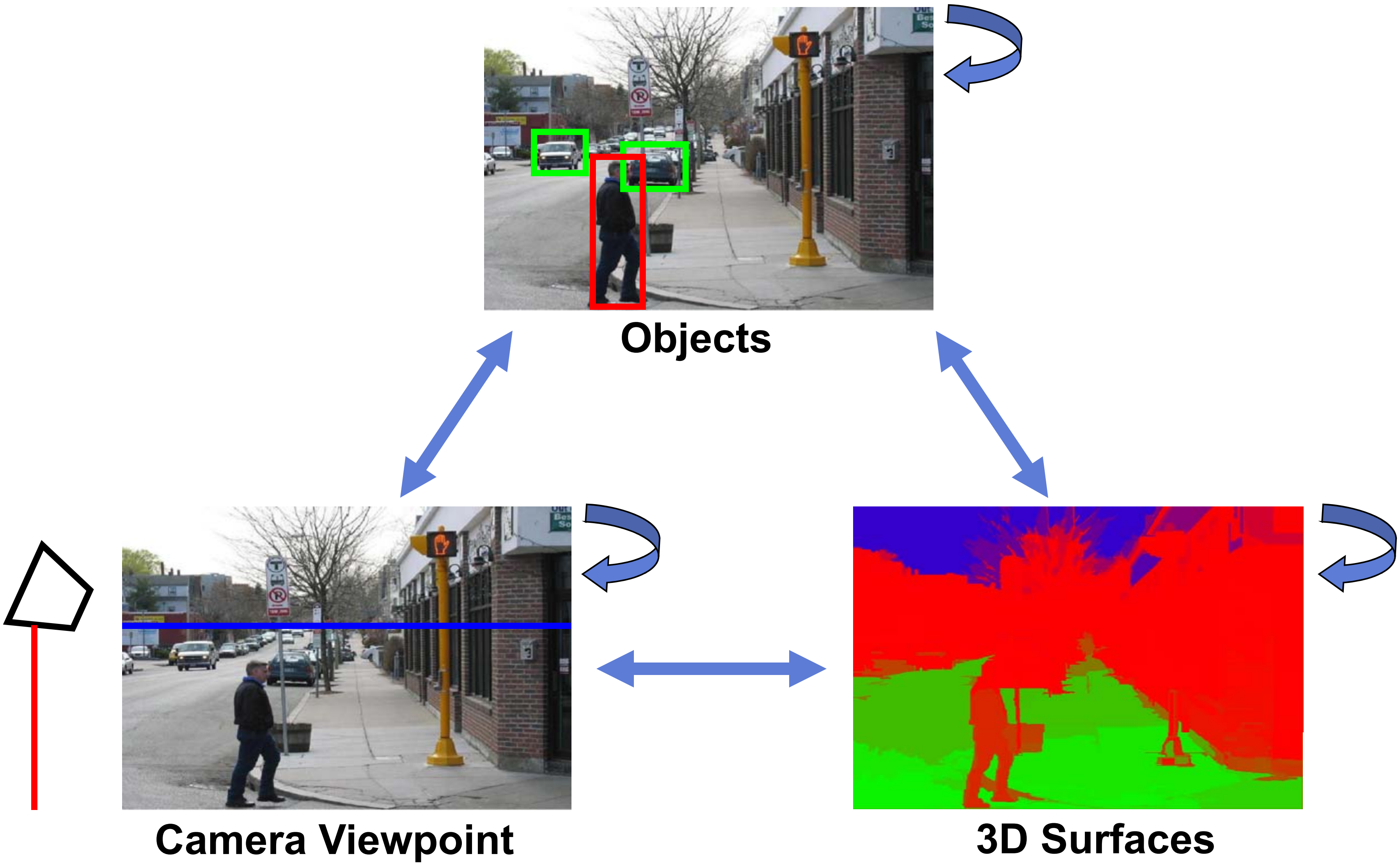


P(object)



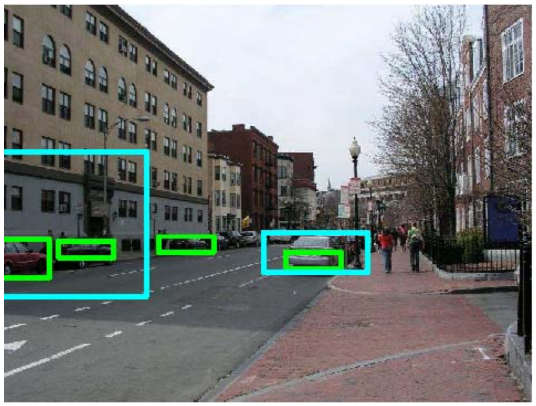
P(object | surfaces, viewpoint)

Scene Parts Are All Interconnected

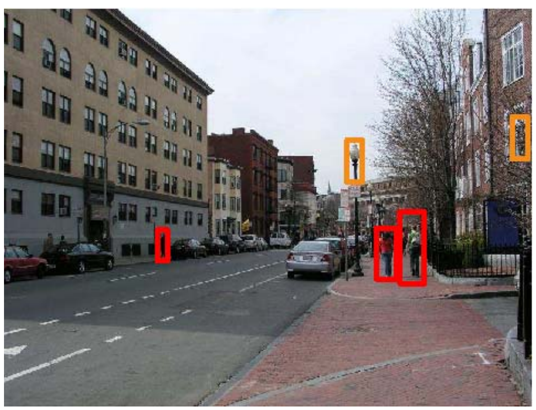


Input to Our Algorithm

Object Detection



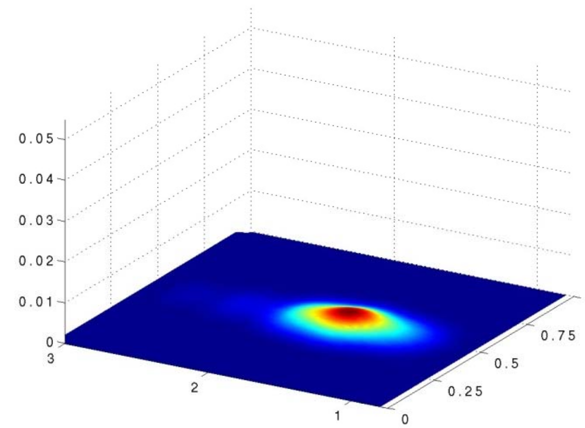
Local Car Detector



Local Ped Detector

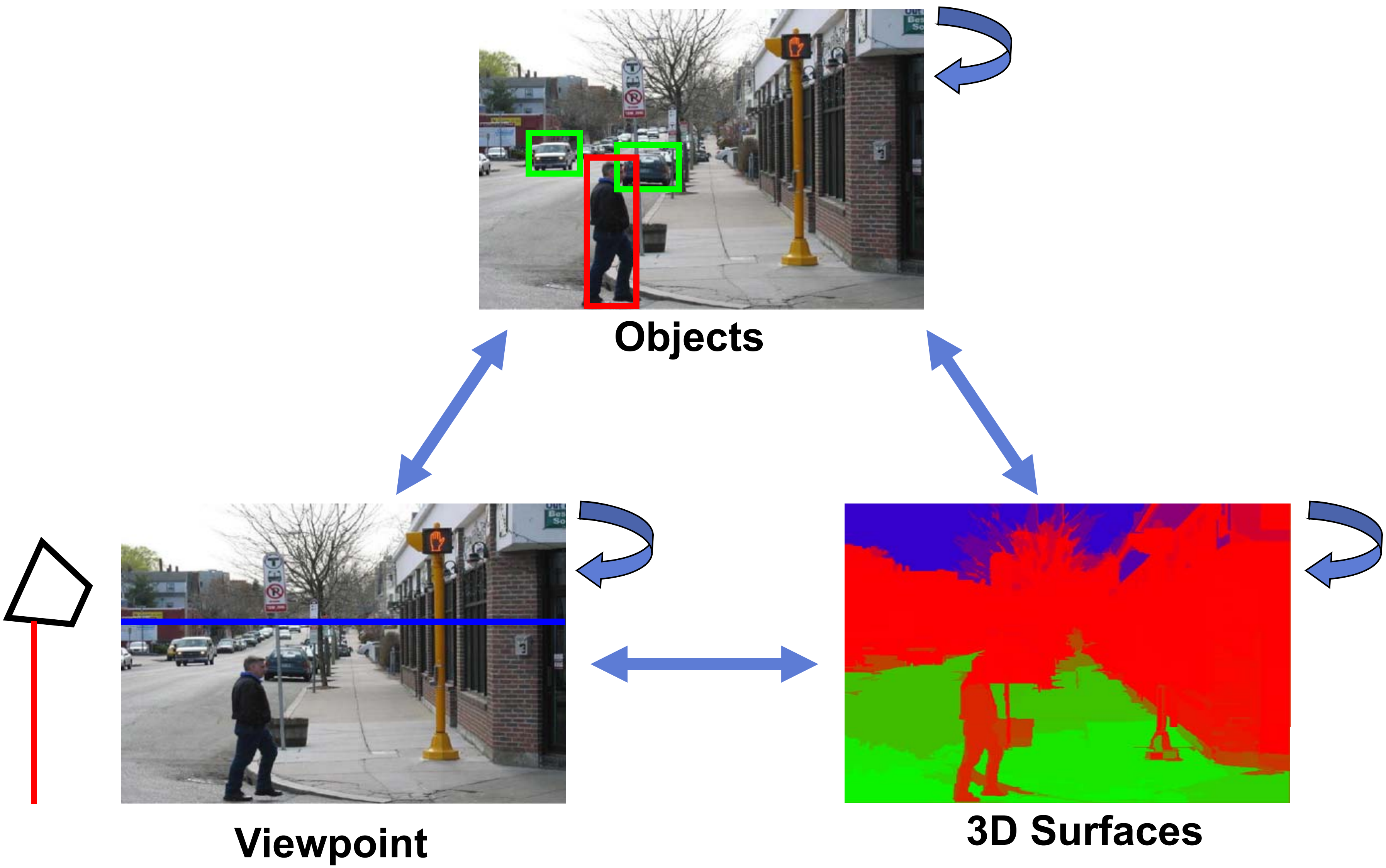
Surface Estimates

Viewpoint Prior

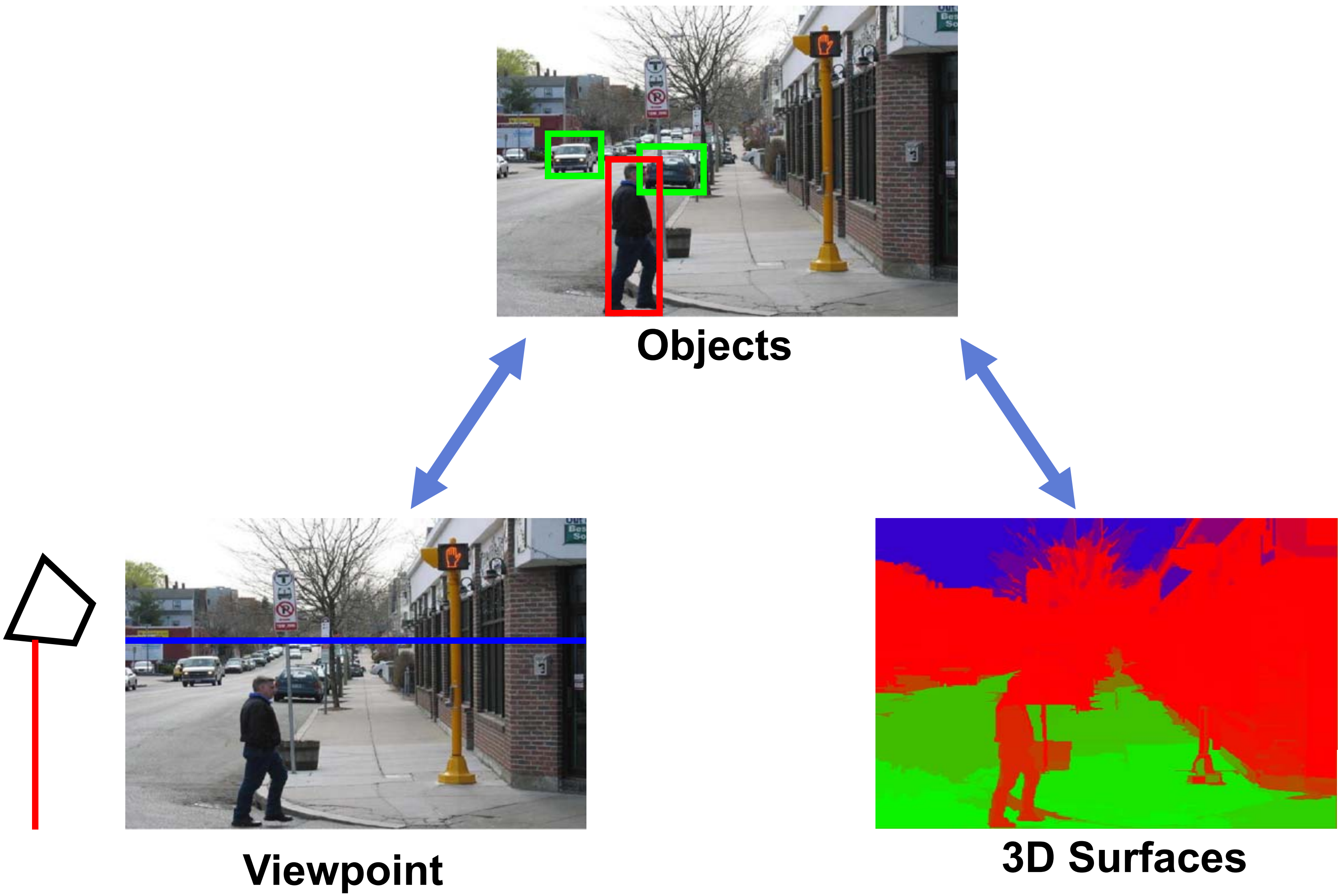


Local Detector: [Dalal-Triggs 2005]   Surfaces: [Hoiem-Efros-Hebert 2005]

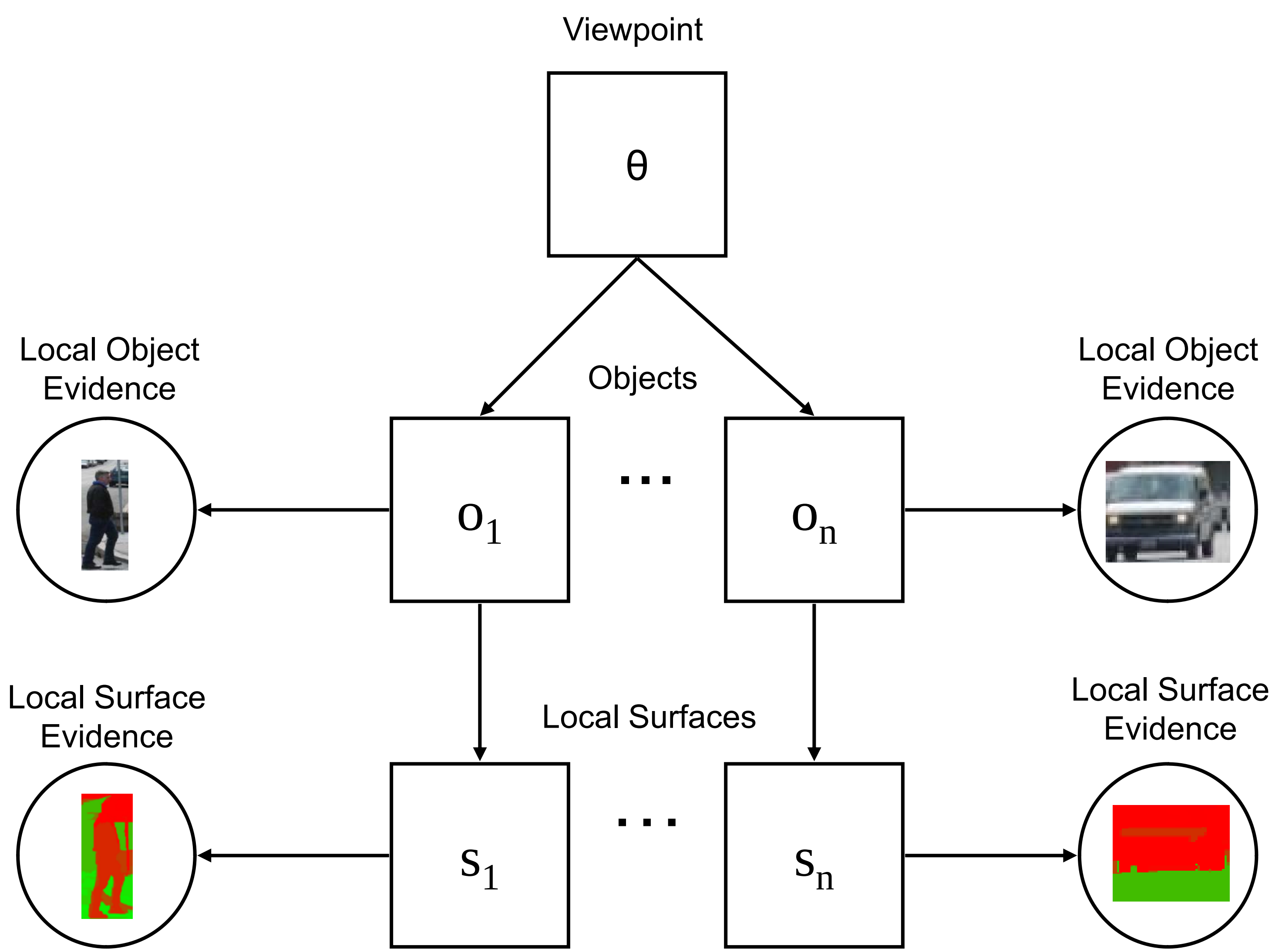
Scene Parts Are All Interconnected



Our Approximate Model

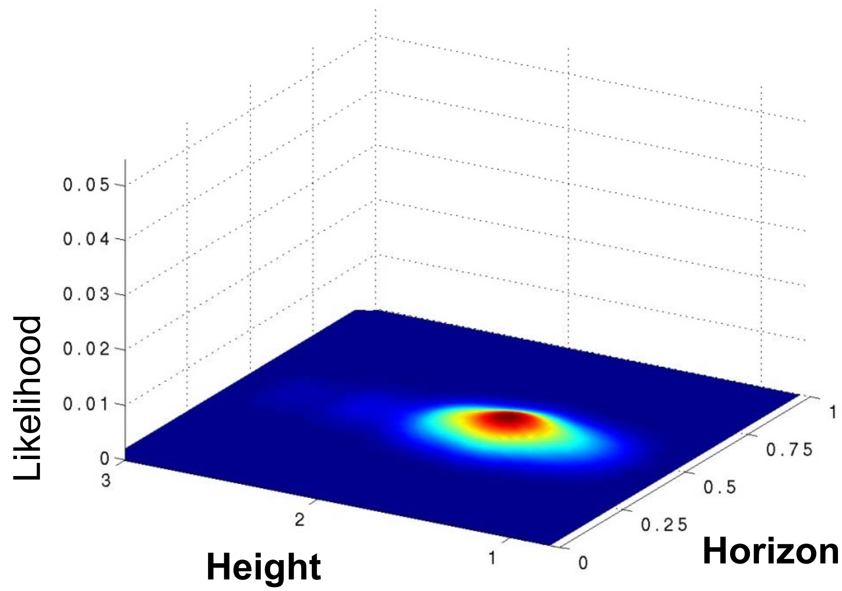


Inference over Tree Easy with BP

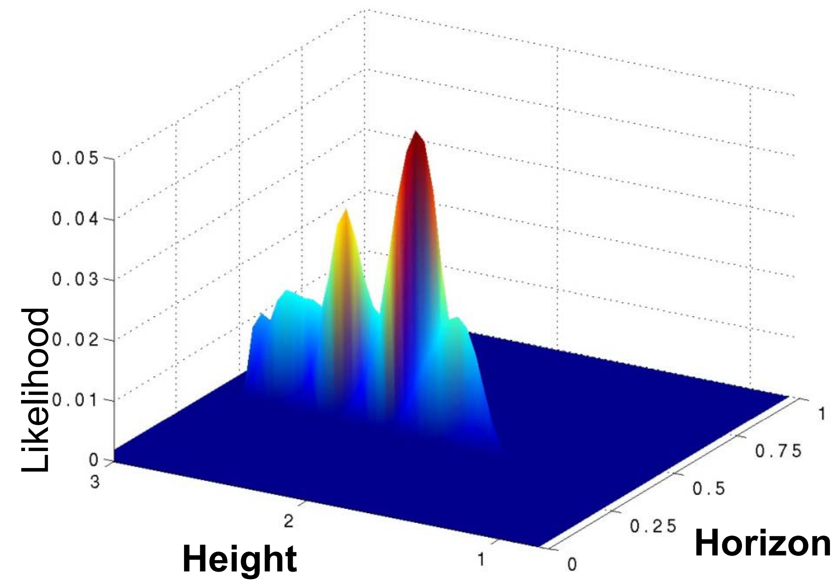


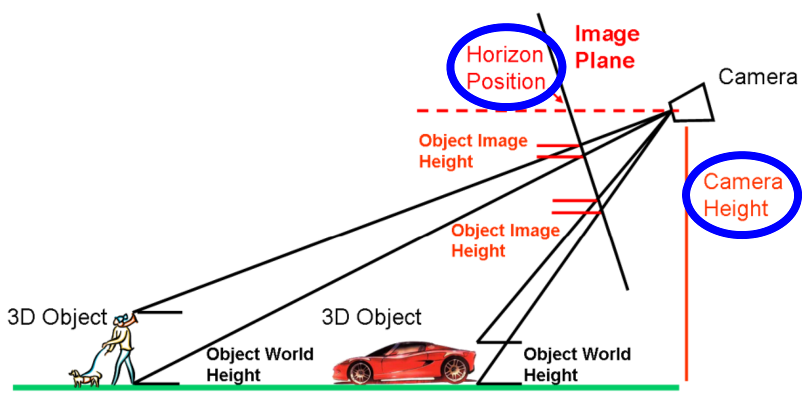
Viewpoint estimation

Viewpoint Prior



Viewpoint Final





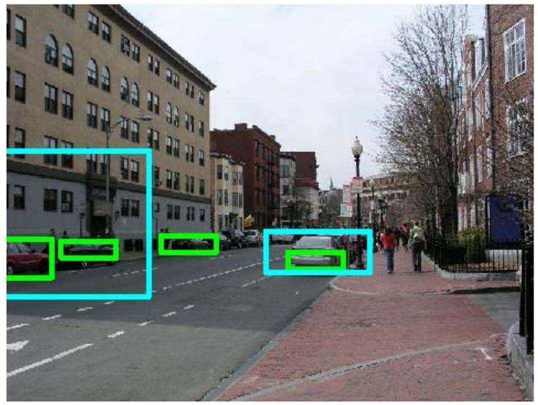
Object detection

Car: TP / FP

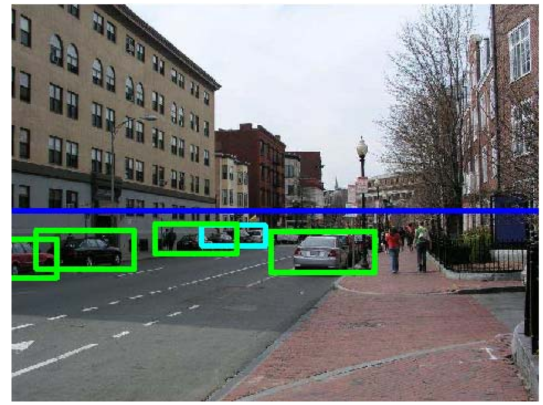
Ped: TP / FP

Initial (Local) | Final (Global)

Car Detection

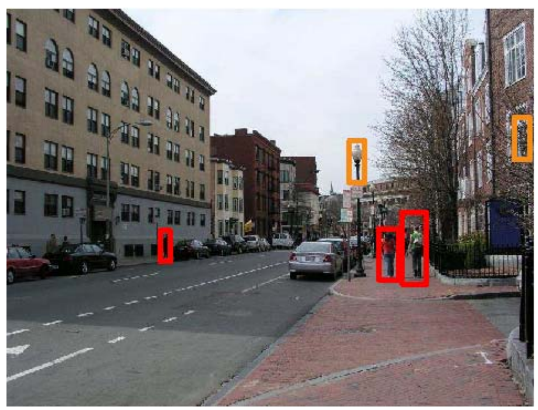


4 TP / 2 FP

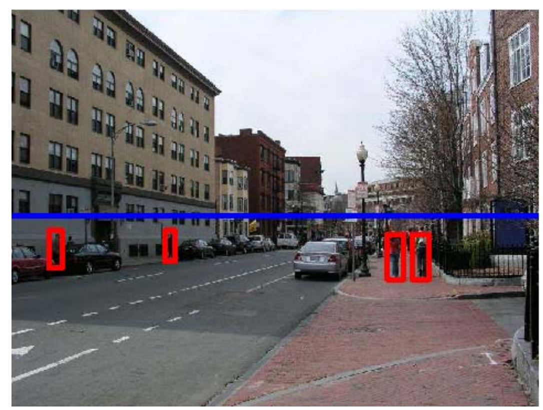


4 TP / 1 FP

Ped Detection



3 TP / 2 FP



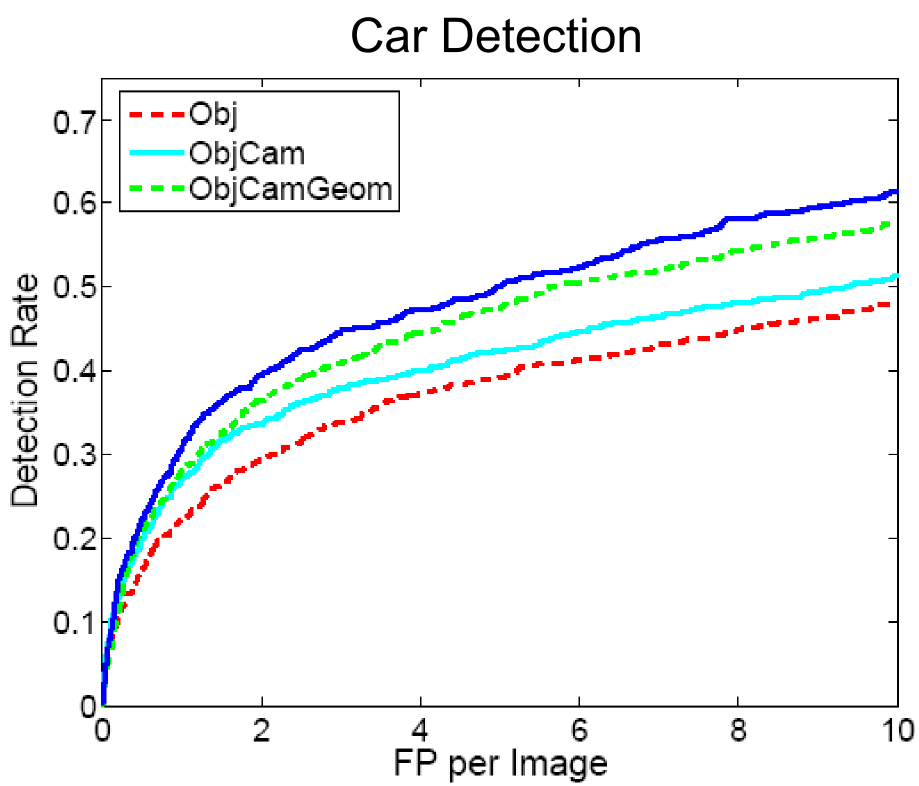
4 TP / 0 FP

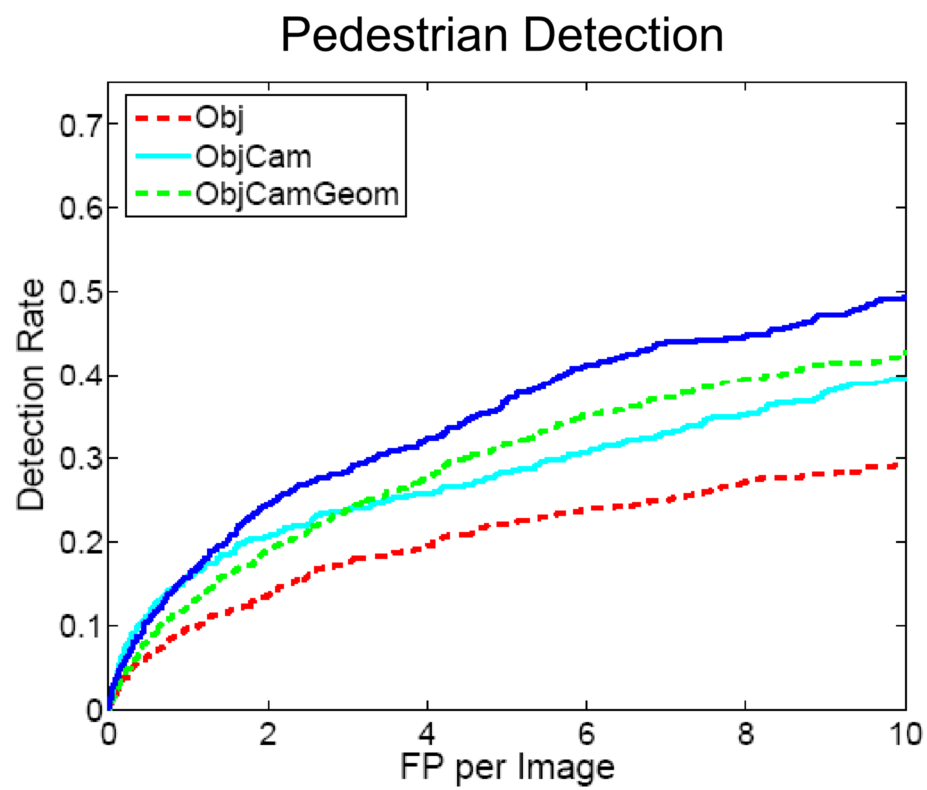
Local Detector: [Dalal-Triggs 2005]

Experiments on LabelMe Dataset

- Testing with LabelMe dataset: 422 images
  - 923 Cars at least 14 pixels tall
  - 720 Peds at least 36 pixels tall

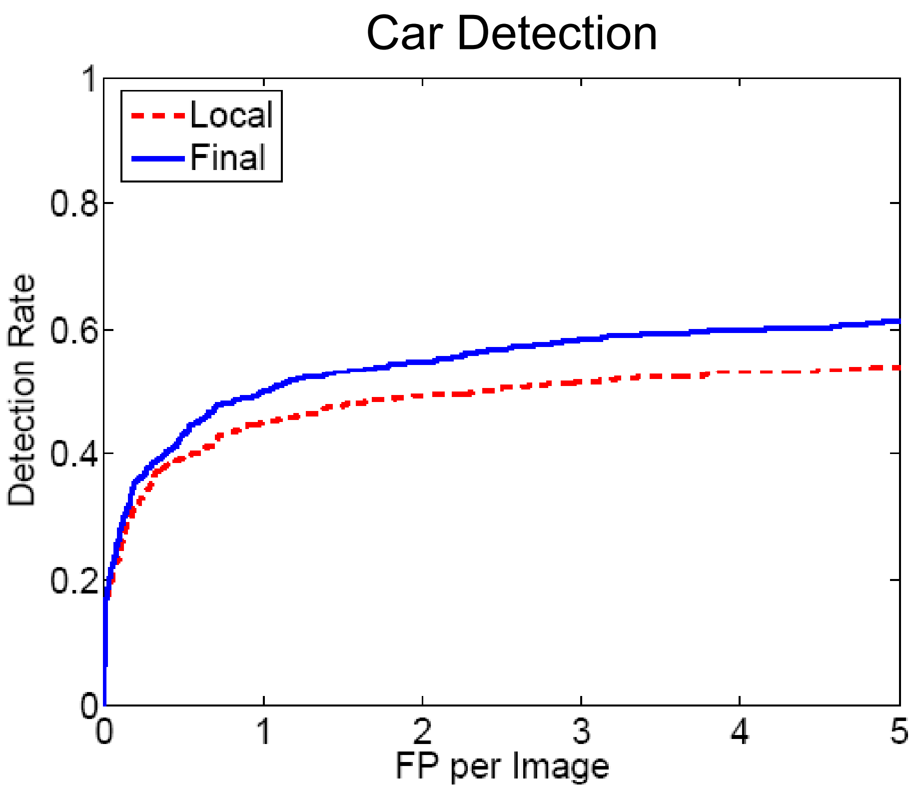
Each piece of evidence improves performance

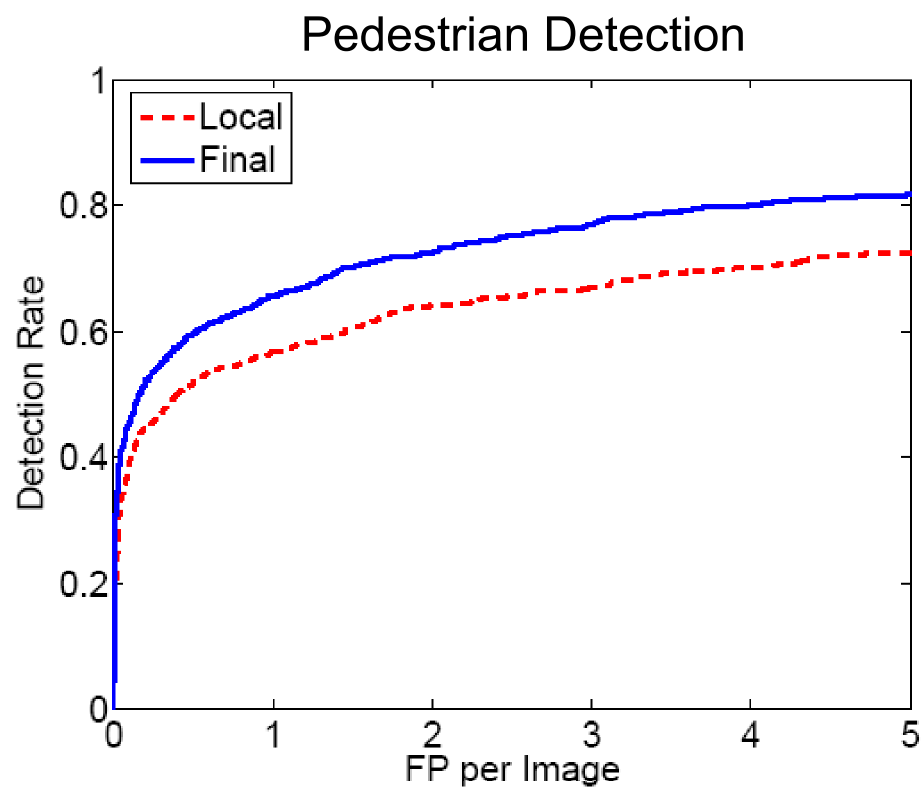




Local Detector from [Murphy-Torralba-Freeman 2003]

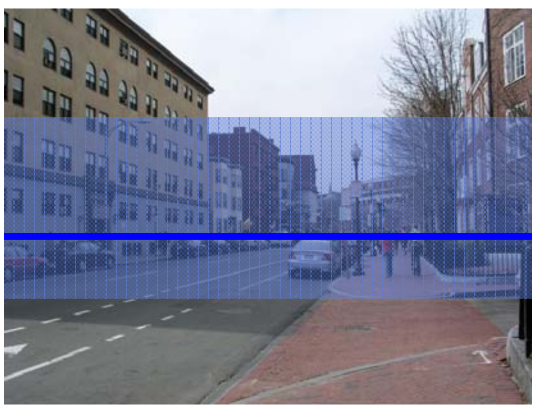
Can be used with any detector that outputs confidences





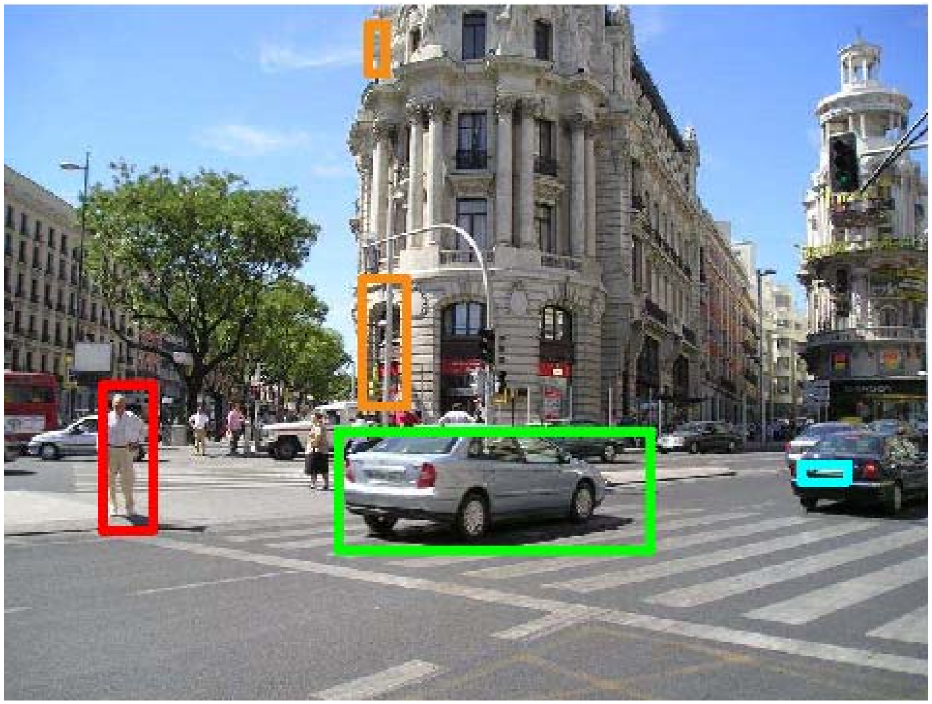
Local Detector: [Dalal-Triggs 2005] (SVM-based)

Accurate Horizon Estimation

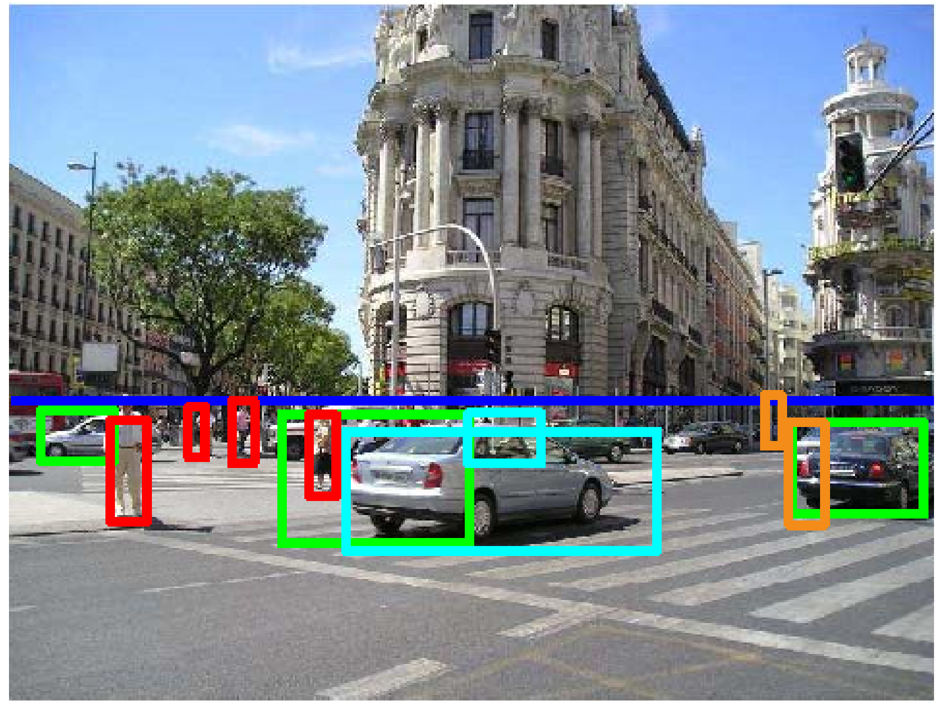
	Horizon Prior	[Murphy-Torralba-Freeman 2003]	[Dalal-Triggs 2005]
Median Error:	8.5%	4.5%	3.0%
90% Bound:			

Qualitative Results

Car: TP / FP Ped: TP / FP



Initial: 2 TP / 3 FP

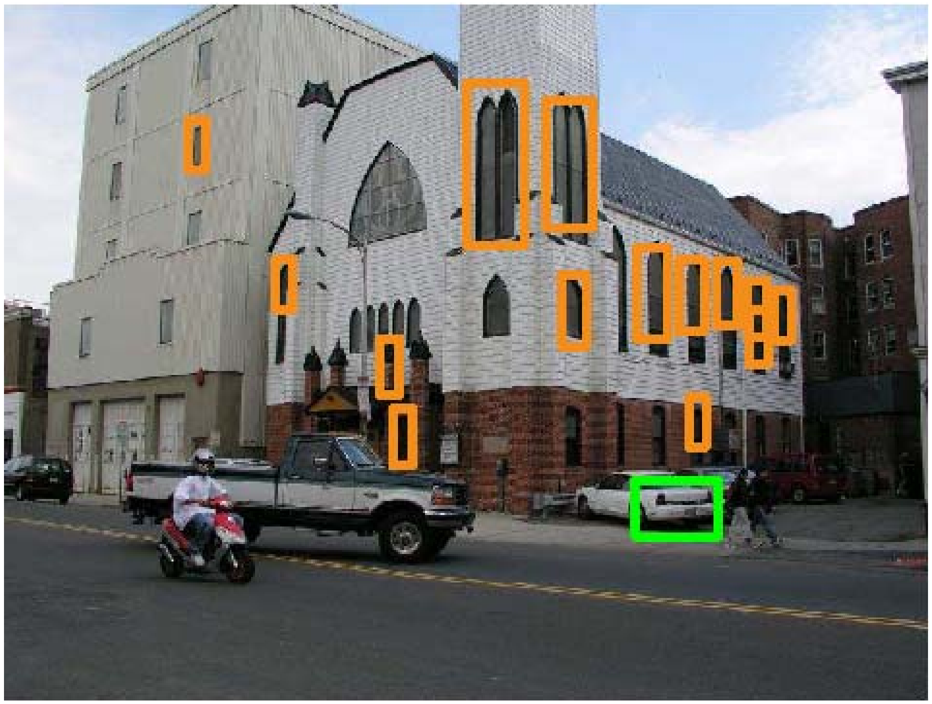


Final: 7 TP / 4 FP

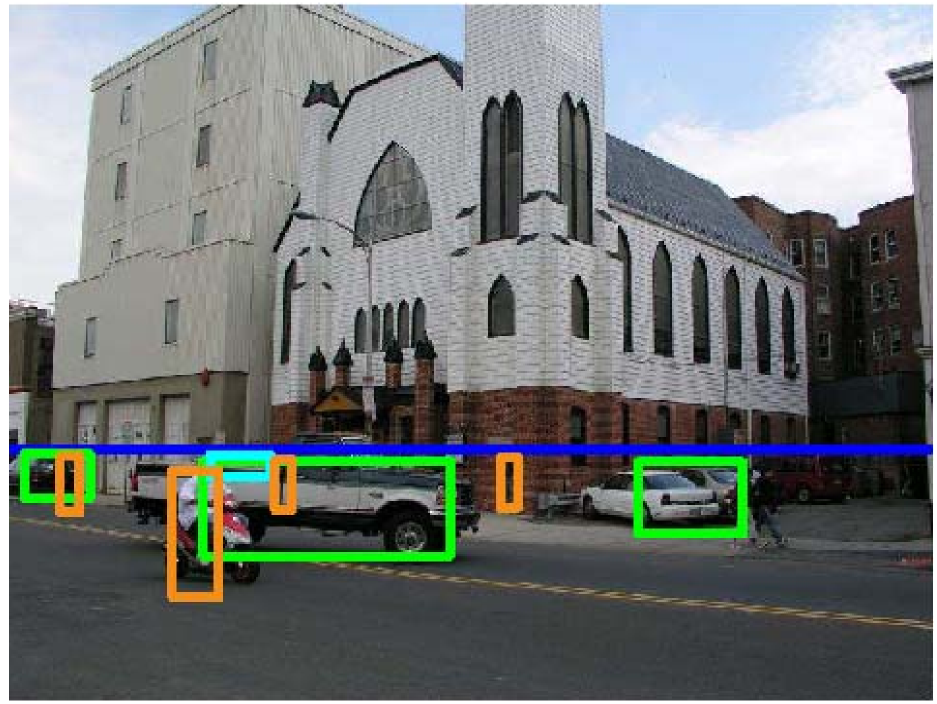
Local Detector from [Murphy-Torralba-Freeman 2003]

Qualitative Results

Car: TP / FP Ped: TP / FP



Initial: 1 TP / 14 FP

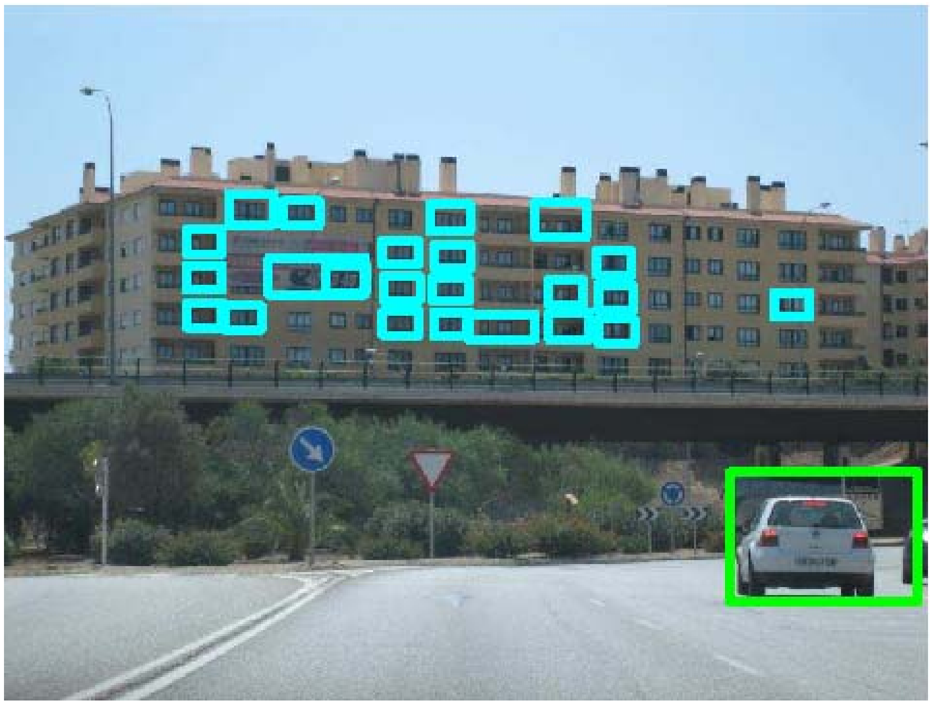


Final: 3 TP / 5 FP

Local Detector from [Murphy-Torralba-Freeman 2003]

Qualitative Results

Car: TP / FP Ped: TP / FP



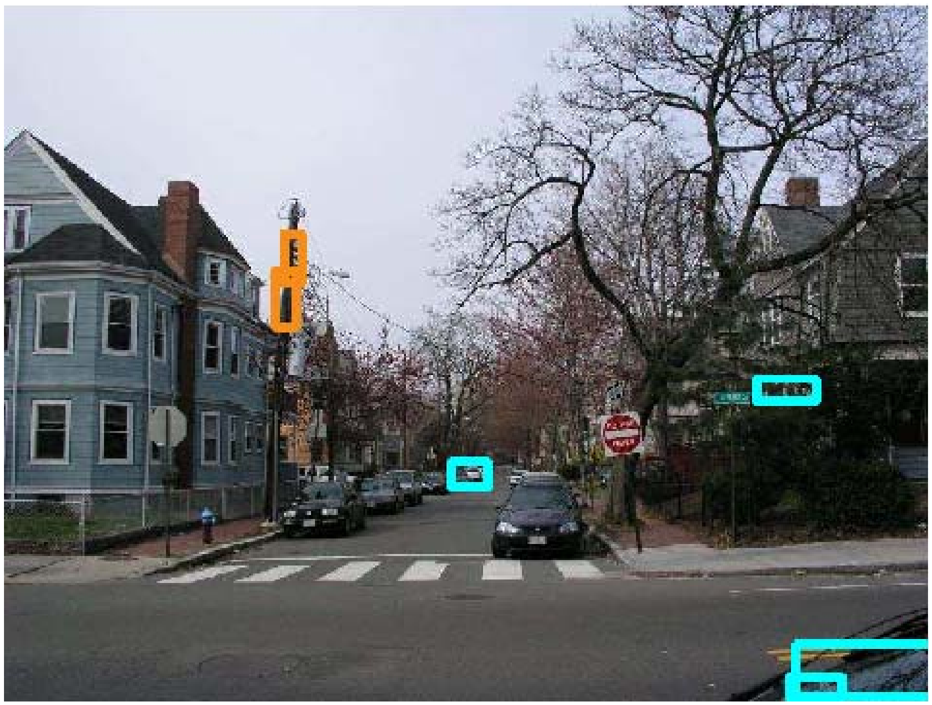
Initial: 1 TP / 23 FP

Final: 0 TP / 10 FP

Local Detector from [Murphy-Torralba-Freeman 2003]

Qualitative Results

Car: TP / FP Ped: TP / FP

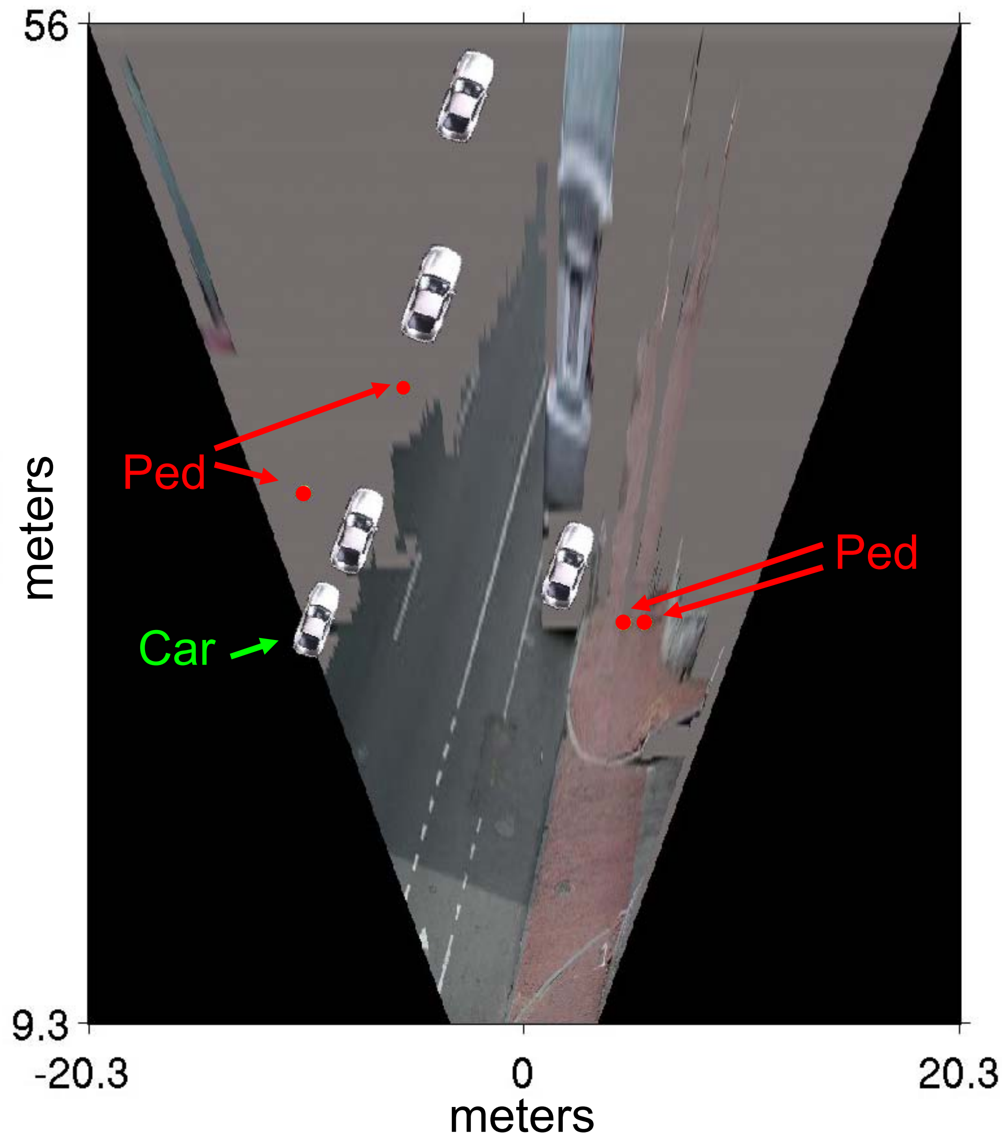


Initial: 0 TP / 6 FP

Final: 4 TP / 3 FP

Local Detector from [Murphy-Torralba-Freeman 2003]

Summary & Future Work



Reasoning in 3D:

- Object to object
- Scene label
- Object segmentation

Conclusion

- Image understanding is a 3D problem
  - Must be solved jointly
- This paper is a small step
  - Much remains to be done

Learning Spatial Context:
Using stuff to find things

Geremy Heitz
Daphne Koller

Stanford University

October 13, 2008
ECCV 2008

Things vs. Stuff

From: Forsyth et al. **Finding pictures of objects in large collections of images**. *Object Representation in Computer Vision*, 1996.

Thing (n): An object with a specific size and shape.

Stuff (n): Material defined by a homogeneous or repetitive pattern of fine-scale properties, but has no specific or distinctive spatial extent or shape.

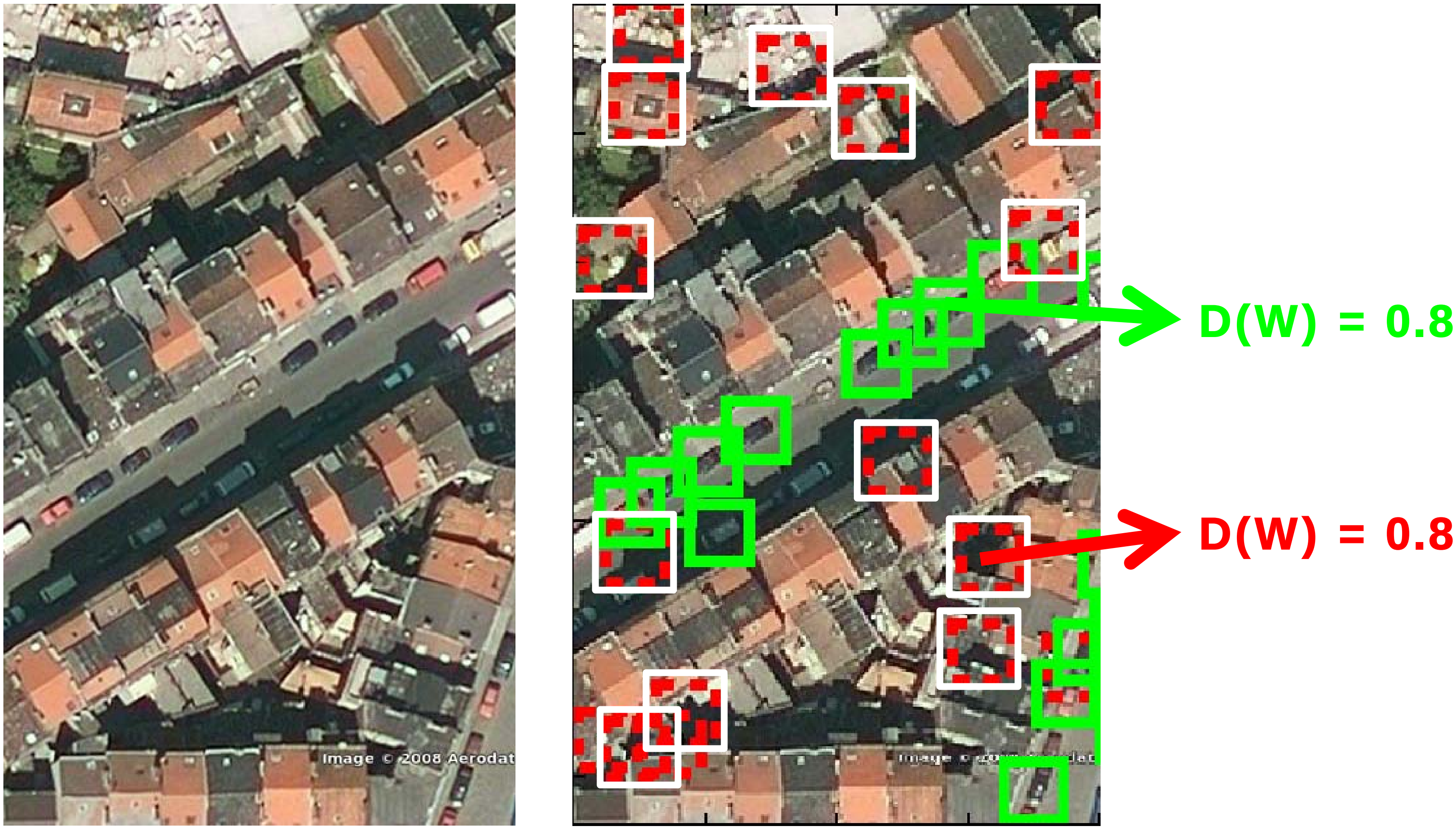
Finding Things

Context is key!

Outline

- What is Context?
- The Things and Stuff (TAS) model
- Results

Satellite Detection Example



Error Analysis

Typically…

True Positives are IN CONTEXT

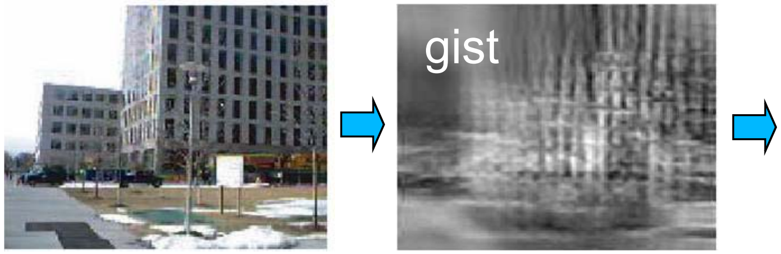
False Positives are OUT OF CONTEXT

We need to look outside the bounding box!

Types of Context

- Scene-Thing:

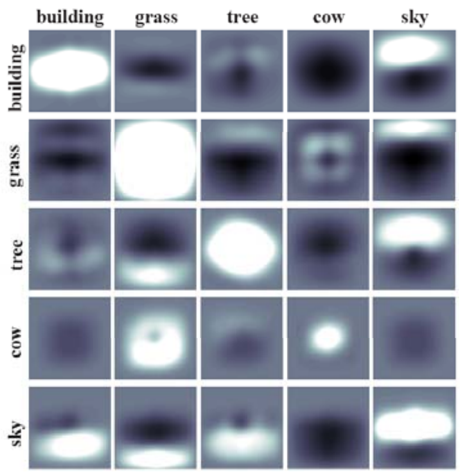
[Torralba et al., LNCS 2005]

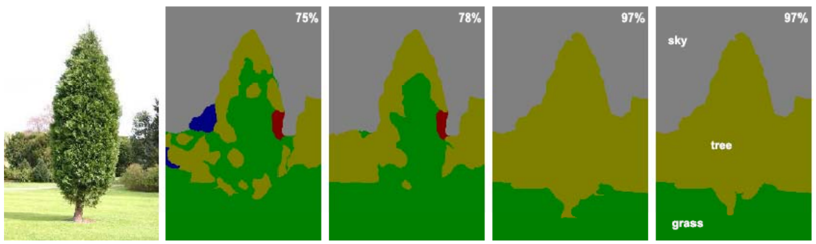


car “likely”

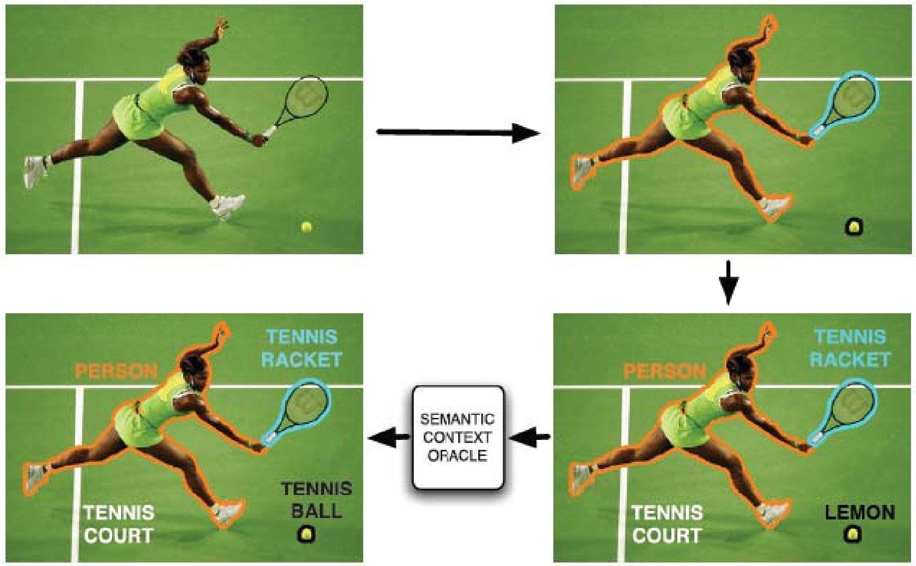
keyboard “unlikely”

- Stuff-Stuff: [Gould et al., IJCV 2008]





- Thing-Thing: [Rabinovich et al., ICCV 2007]



Types of Context

- Stuff-Thing:
  - Based on spatial relationships

- Intuition:

"Cars drive on roads"

"Cows graze on grass"

"Boats sail on water"

Outline

- What is Context?
- The Things and Stuff (TAS) model
- Results

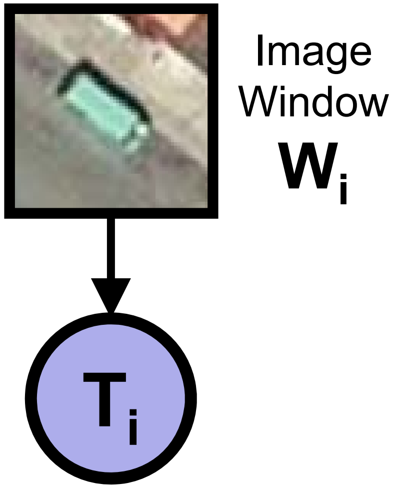
Things

- Detection “candidates”
  - Low detector threshold -> “over-detect”
  - Each candidate has a detector score

Things

- Candidate detections
  - Image Window $\mathbf{W_i}$ + Score
- Boolean R.V. $\mathbf{T_i}$
  - $T_i = 1$: Candidate is a positive detection

- Thing model



$$P(T_i|W) = \frac{1}{1+\exp(\alpha + \beta \cdot D(W))}$$

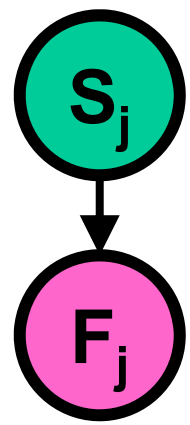
Stuff

- Coherent image regions
  - Coarse “superpixels”
  - Feature vector $\mathbf{F_j}$ in R^n
  - Cluster label $\mathbf{S_j}$ in {1…C}

- Stuff model
  - Naïve Bayes

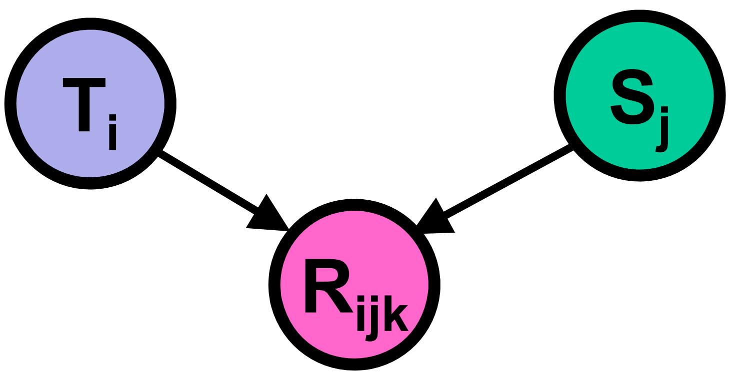
$$P(S_j, F_j) = P(S_j)P\left(F_j \middle| S_j\right)$$

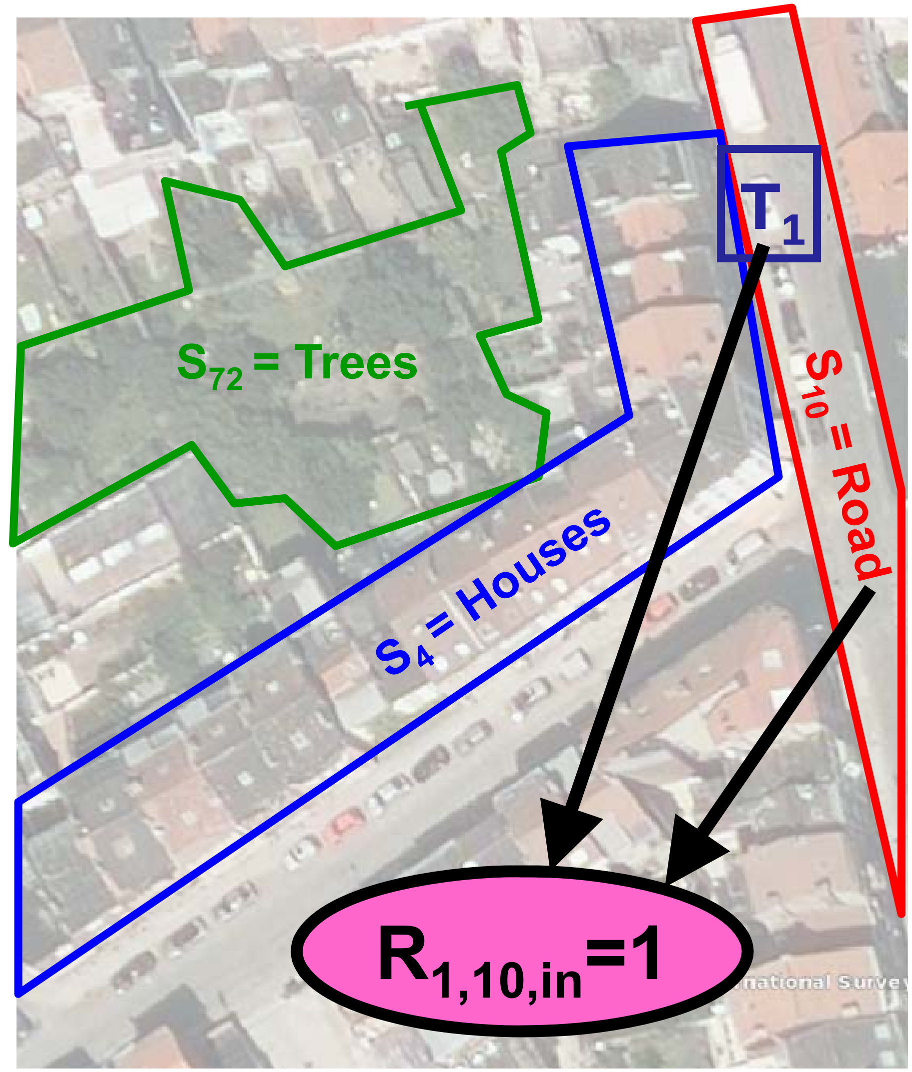
$$F_j \left| \left(S_j = s\right) \sim \mathrm{N}\left(\mu_s, \Sigma_s\right)\right.$$



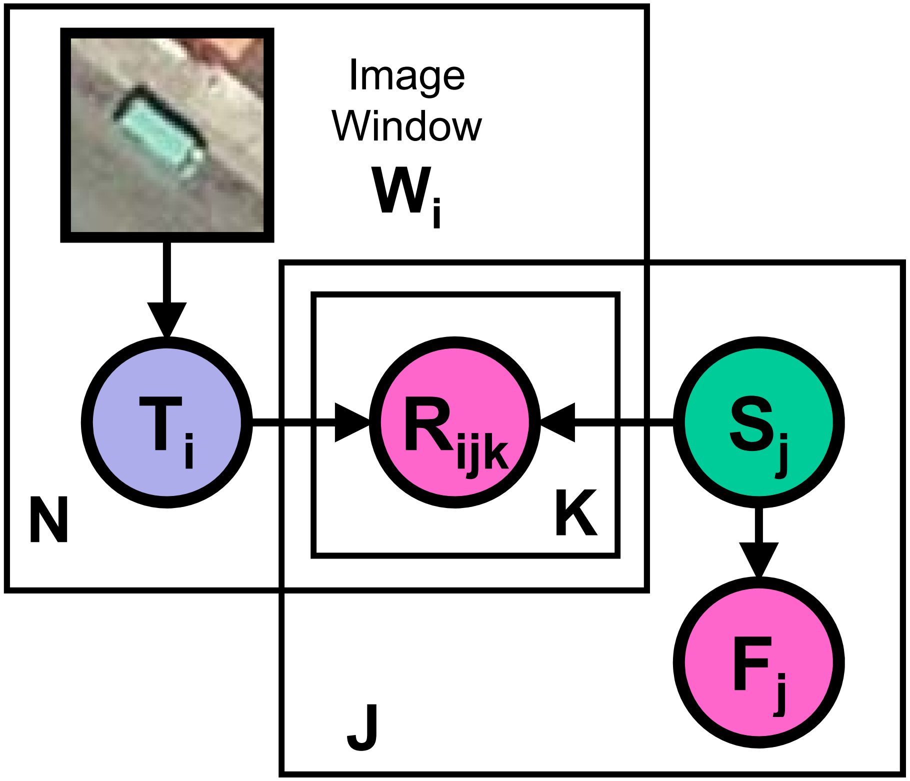
Relationships

- Descriptive Relations
  - "Near", "Above", "In front of", etc.
- Choose set $\mathbf{R} = \{r_1 \ldots r_K\}$
- $R_{ijk}=1$: Detection i and region j have relation k
- Relationship model





The TAS Model



W_i: Window

T_i: Object Presence

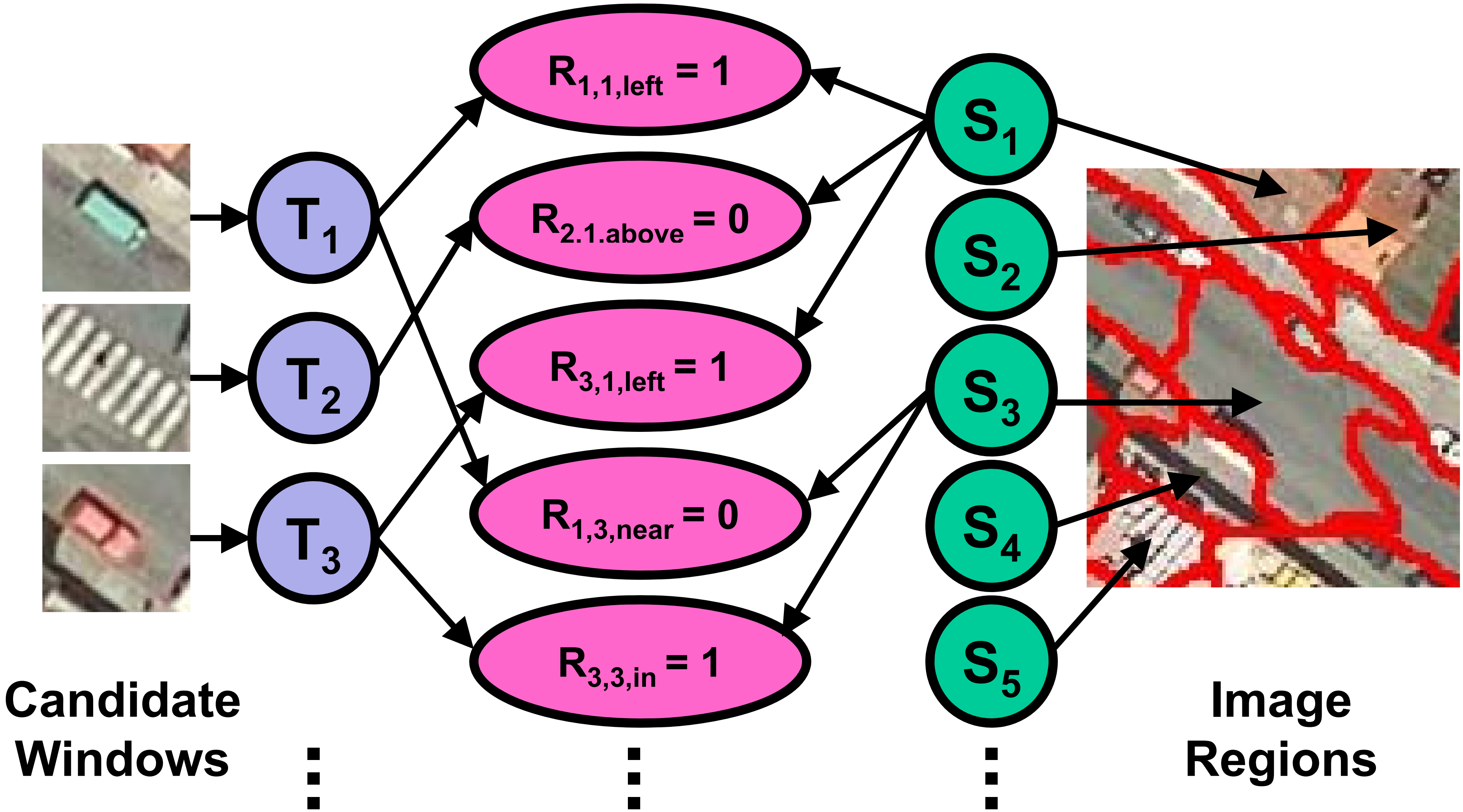
S_j: Region Label

F_j: Region Features

R_{ijk}: Relationship

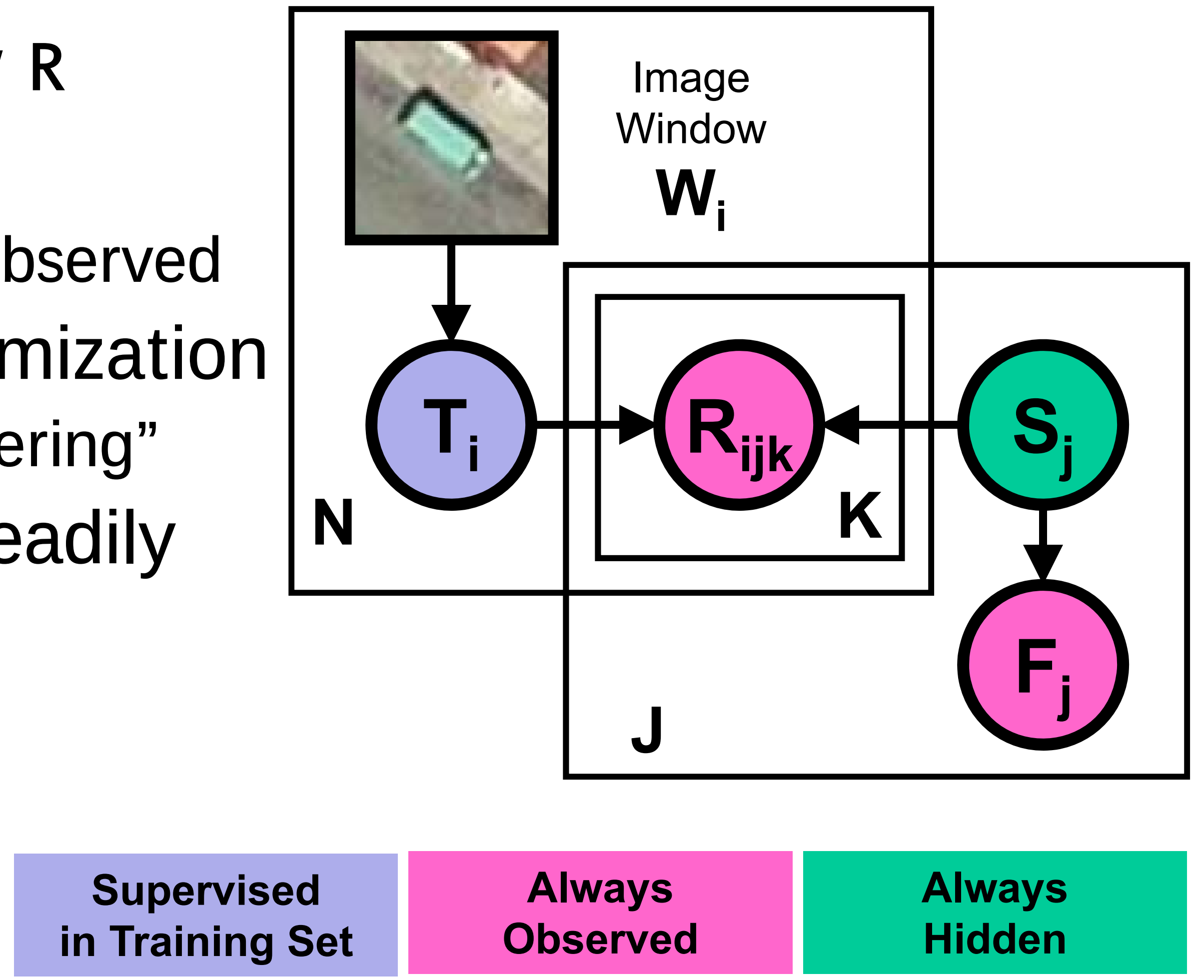
Supervised in Training Set	Always Observed	Always Hidden

Unrolled Model

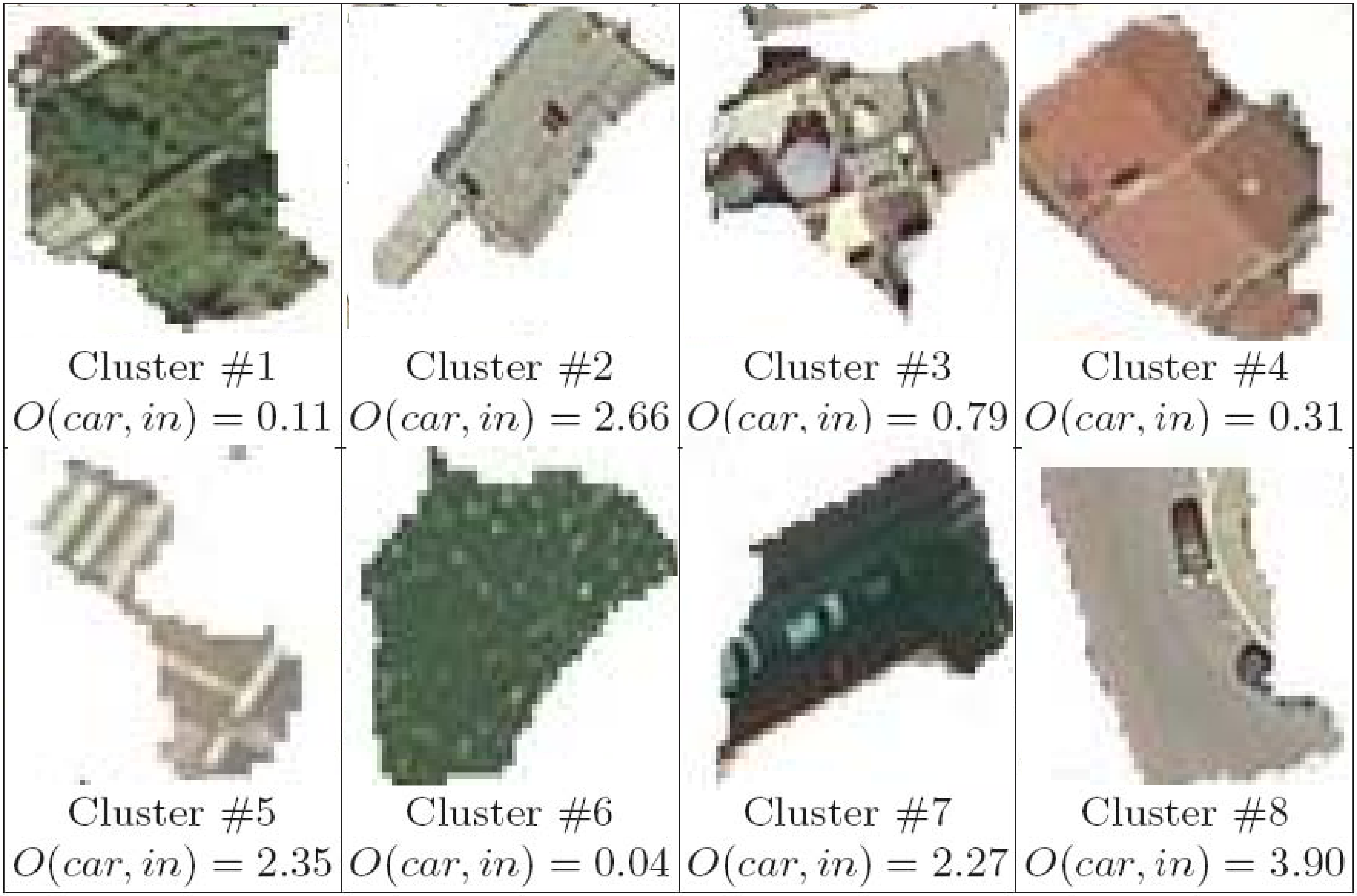


Learning the Parameters

- Assume we know R
- S_j is hidden
  - Everything else observed
- Expectation-Maximization
  - "Contextual clustering"
- Parameters are readily interpretable



Learned Satellite Clusters



Which Relationships to Use?

- Rijk = spatial relationship between candidate i and region j

Rij1 = candidate in region

Rij2 = candidate closer than 2 bounding boxes (BBs) to region

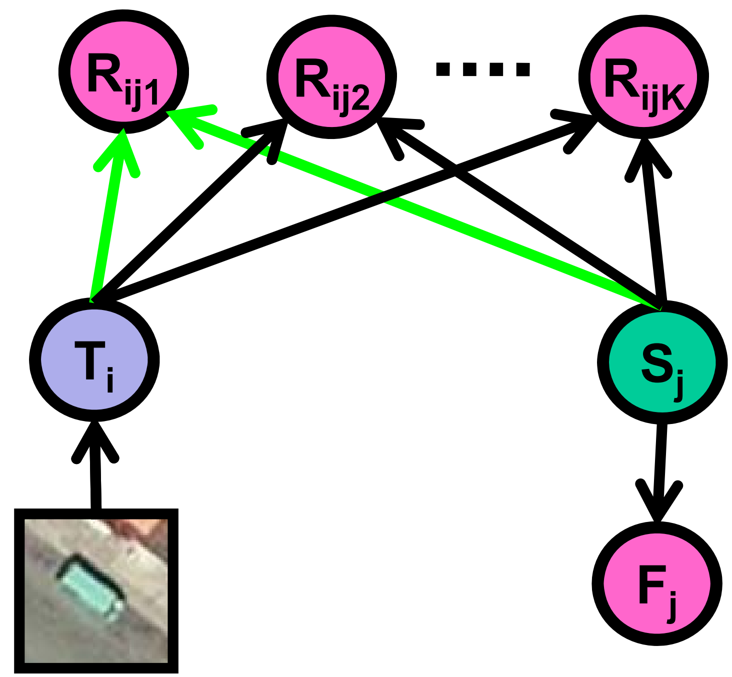
How do we avoid overfitting?

...

RijK = candidate near region boundary

Learning the Relationships

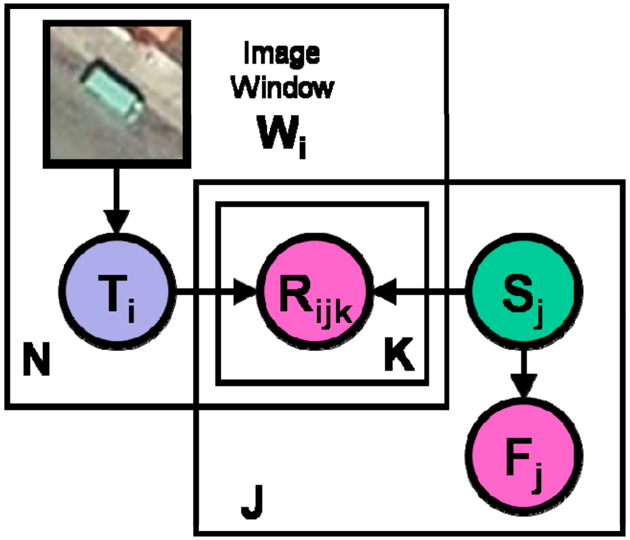
- Intuition
  - "Detached" R_{ijk} = inactive relationship
- Structural EM iterates:
  - Learn parameters
  - Decide which edge to toggle
- Evaluate with $\ell(T|F,W,R)$
  - Requires inference
  - Better results than using standard $E[\ell(T,S,F,W,R)]$



Inference

- Goal:

$$P(\boldsymbol{T} \mid \boldsymbol{F}, \boldsymbol{R}, \boldsymbol{W}) = \sum_{\boldsymbol{S}} P(\boldsymbol{T}, \boldsymbol{S} \mid \boldsymbol{F}, \boldsymbol{R}, \boldsymbol{W})$$



- Block Gibbs Sampling
  - Easy to sample T_i's given S_j's and vice versa

$$P(S_j \mid \boldsymbol{T}, \boldsymbol{F}, \boldsymbol{R}, \boldsymbol{W}) \propto P(S_j) P(F_j \mid S_j) \prod_i P(R_{ij} \mid T_i, S_j)$$

$$P(T_i \mid \boldsymbol{S}, \boldsymbol{F}, \boldsymbol{R}, \boldsymbol{W}) \propto P(T_i \mid W_i) \prod_j P(R_{ij} \mid T_i, S_j).$$

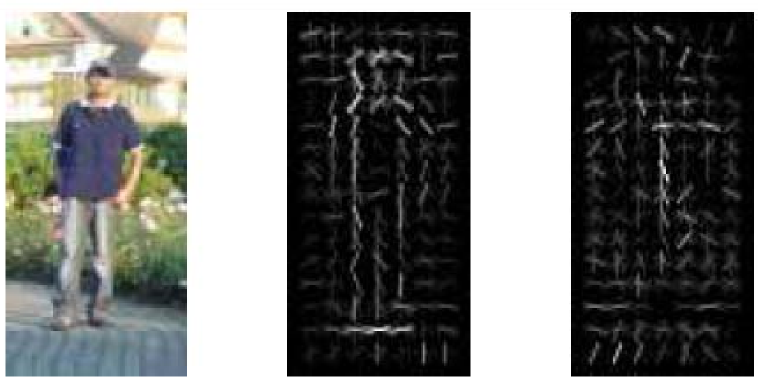
Outline

- What is Context?
- The Things and Stuff (TAS) model
- Results

Base Detector - HOG

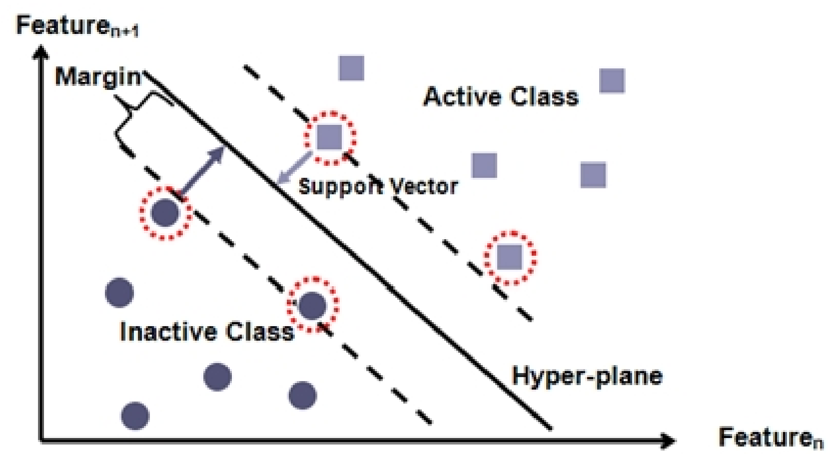
- HOG Detector: **[Dalal & Triggs, CVPR, 2006]**

Feature Vector X



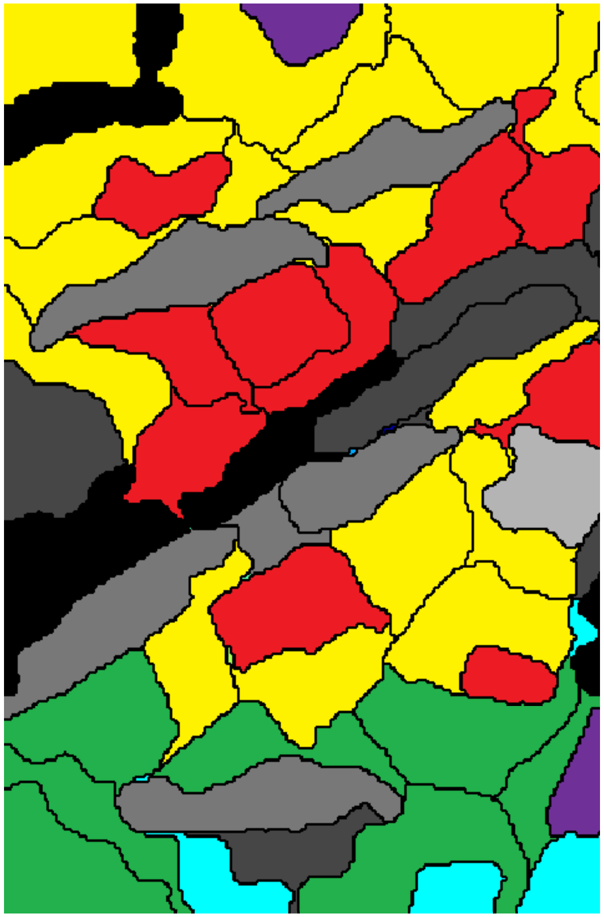
input image | weighted pos wts | weighted neg wts

SVM Classifier



Results - Satellite

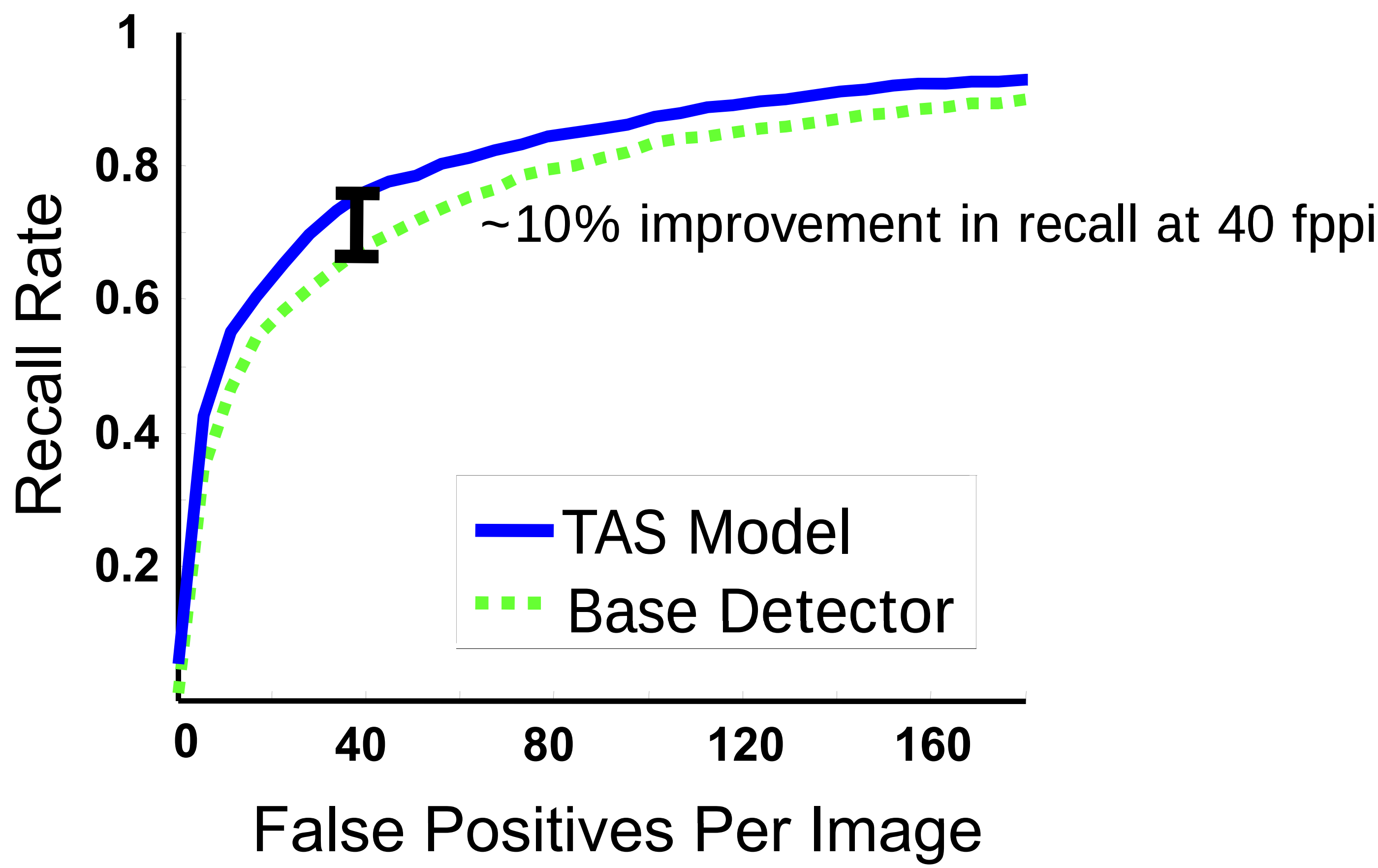
Prior:
Detector Only



Posterior:
Region Labels

Posterior:
Detections

Results - Satellite



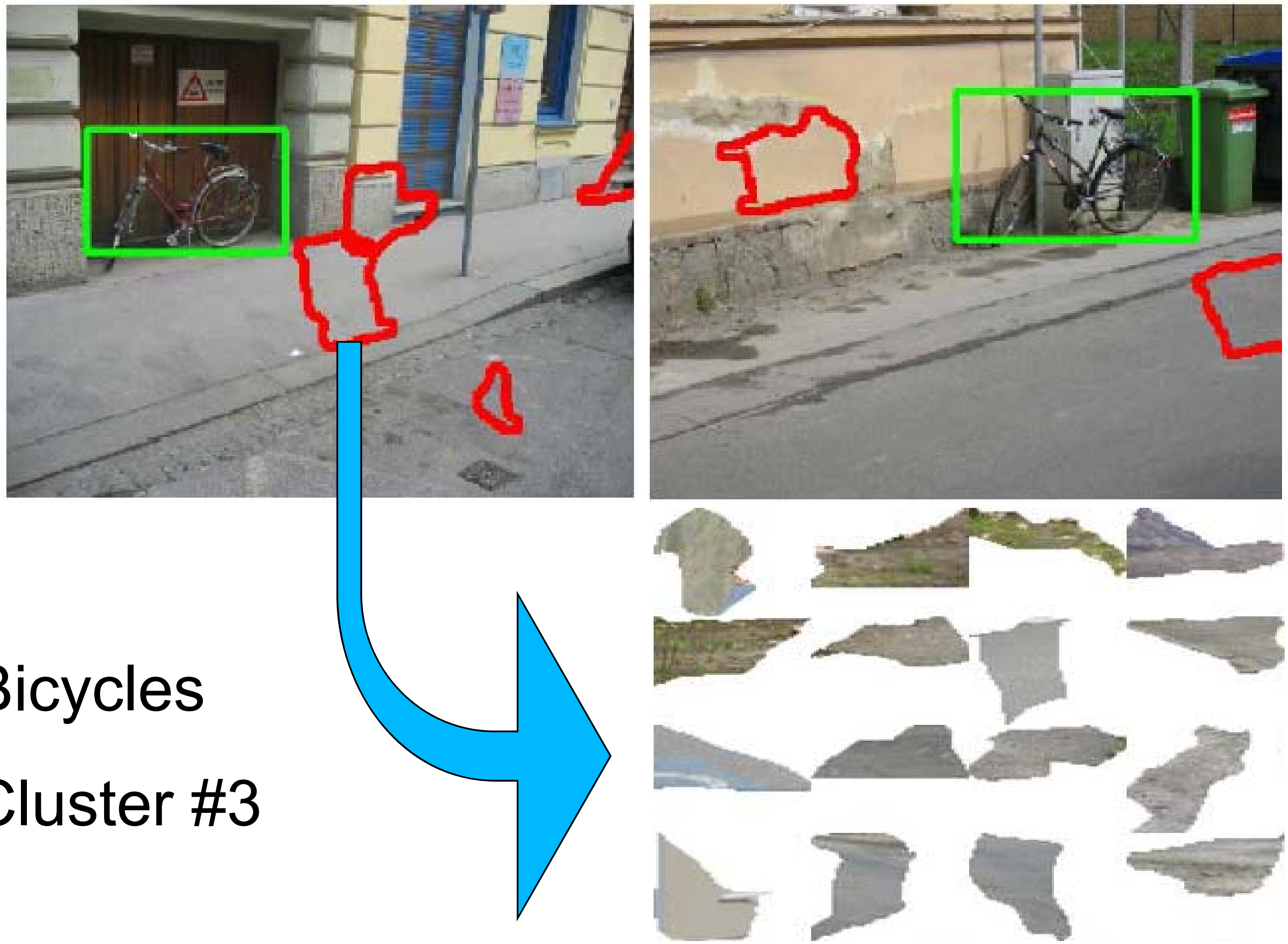
PASCAL VOC Challenge

- 2005 Challenge
  - 2232 images split into {train, val, test}
  - Cars, Bikes, People, and Motorbikes
- 2006 Challenge
  - 5304 images plit into {train, test}
  - 12 classes, we use Cows and Sheep

Base Detector Error Analysis

Cows

Discovered Context - Bicycles

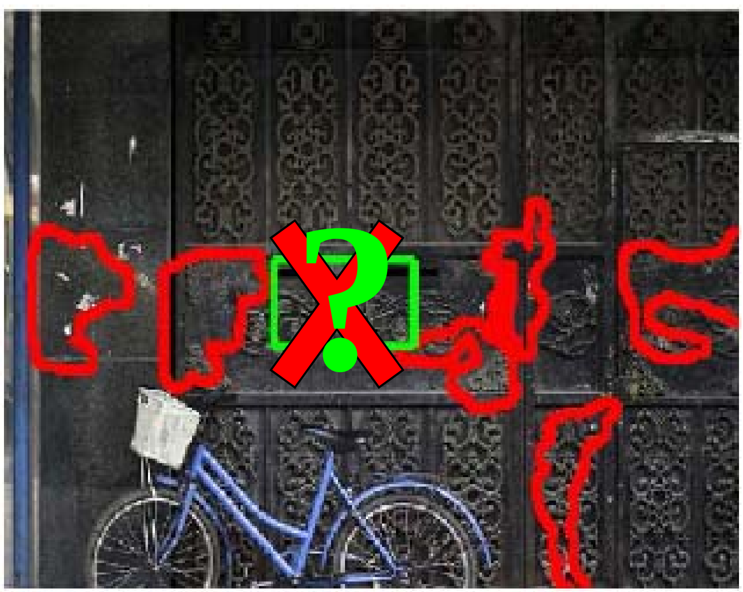


Bicycles

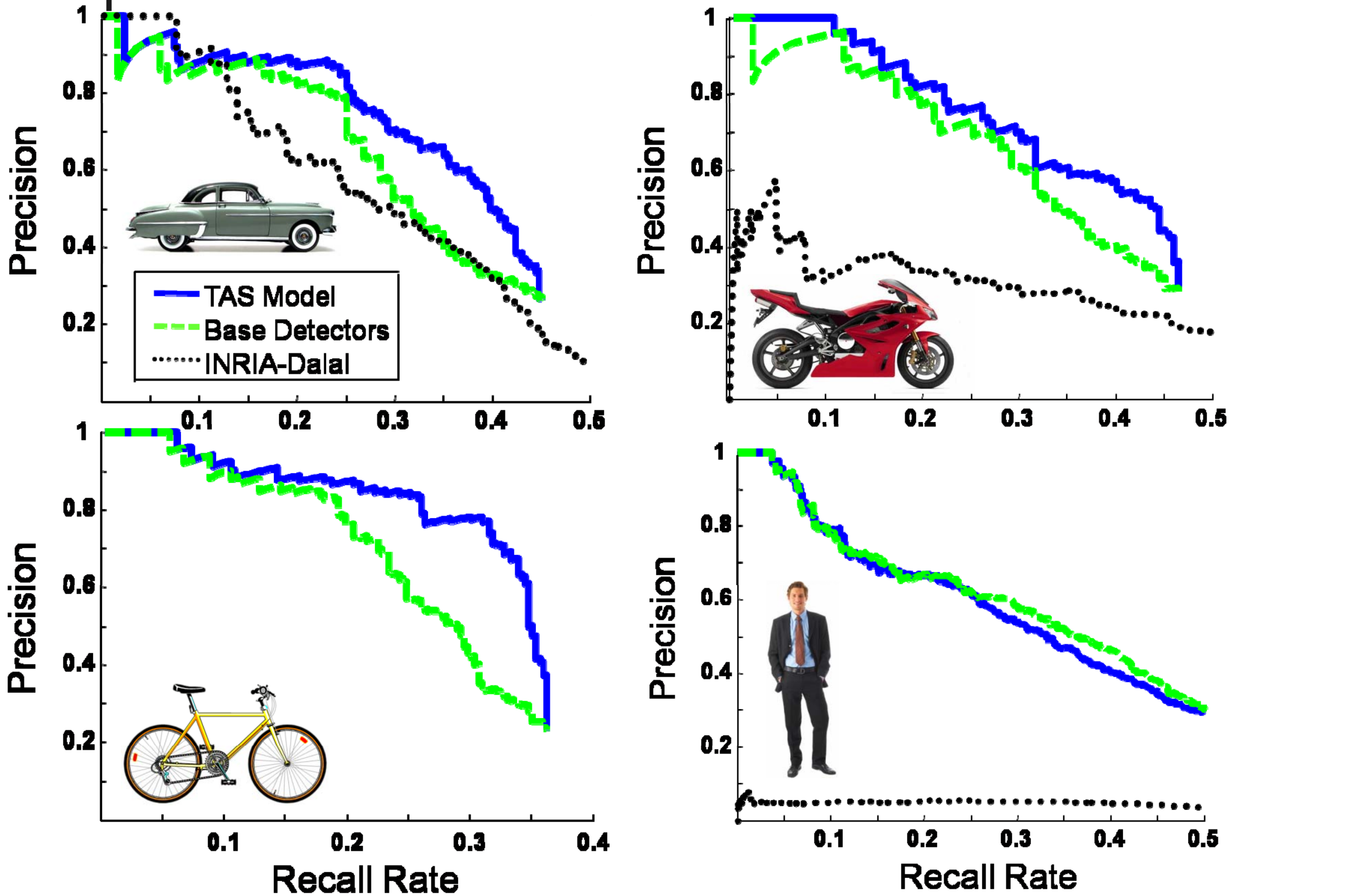
Cluster #3

TAS Results – Bicycles

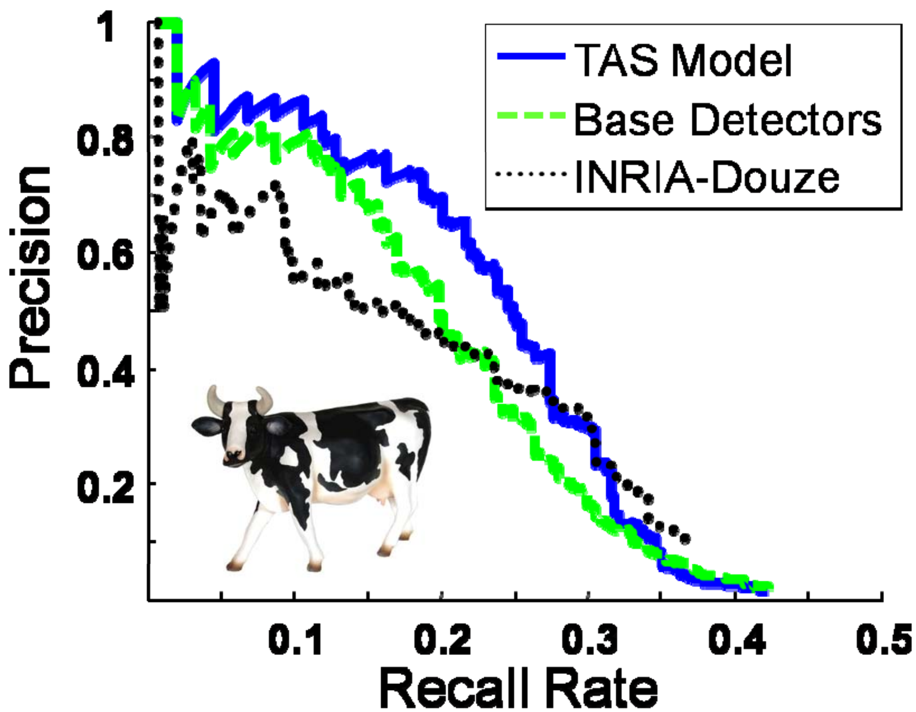
- Examples
  - Discover “true positives”
  - Remove “false positives”

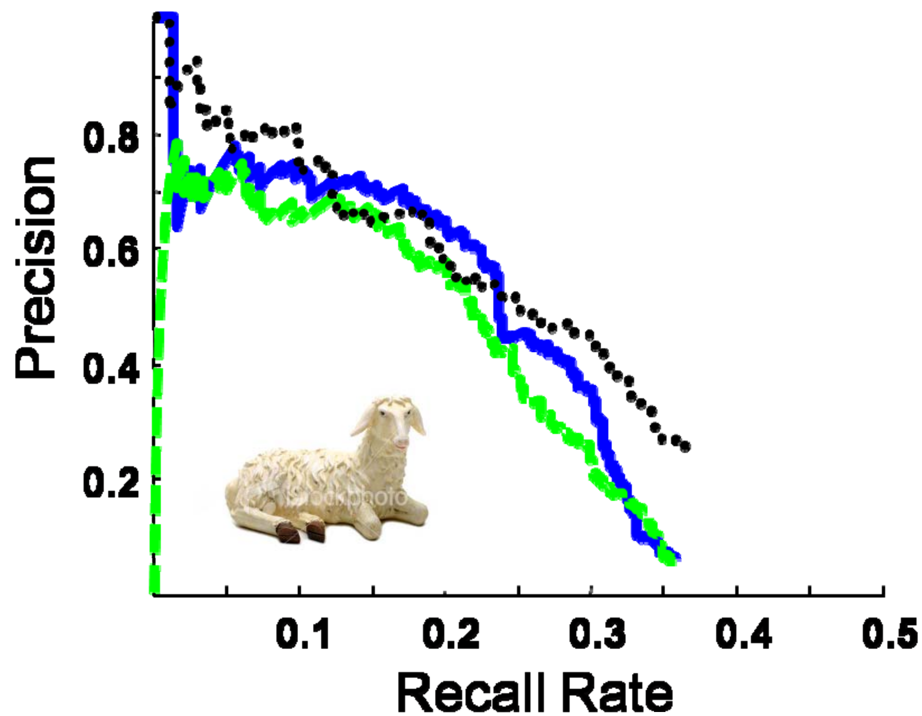


Results – VOC 2005



Results – VOC 2006





Conclusions

- Detectors can benefit from context
- The TAS model captures an important type of context
- We can improve *any* sliding window detector using TAS
- The TAS model can be interpreted and matches our intuitions
- We can learn which relationships to use

Today: Three papers on computational models of context:

- A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in Advances in Neural Information Processing Systems 17 (NIPS), 2005.
- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006
- G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43.

Who needs context anyway?

We can recognize objects even out of context

Banksy

Slide credit: A. Torralba