CS294-43: Recognition in Context

Prof. Trevor Darrell
Spring 2009

April 14th, 2009

Last Lecture – **Kernel Combination, Segmentation, and Structured Output**

- M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007,
- Q. Yuan, A. Thangali, V. Ablavsky, and S. Sclaroff, "Multiplicative kernels: Object detection, segmentation and pose estimation," in Computer Vision and Pattern Recognition, 2008. CVPR 2008
- M. B. Blaschko and C. H. Lampert, "Learning to localize objects with structured output regression," in ECCV 2008.

- C. Pantofaru, C. Schmid, and M. Hebert, "Object recognition by integrating multiple image segmentations," CVPR 2008,
- Chunhui Gu, Joseph J. Lim, Pablo Arbelaez, Jitendra Malik, Recognition using Regions, CVPR 2009, to appear

Today – **Image 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
- L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in Computer Vision, 2007.
- G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43.
- S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. R. Bradski, P. Baumstarck, S. Chung, A. Y. Ng: Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video. IJCAI 2007
- Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008

Today – **Image 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. **[Patrick Sundberg]**

- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 **[Robert Carroll]**

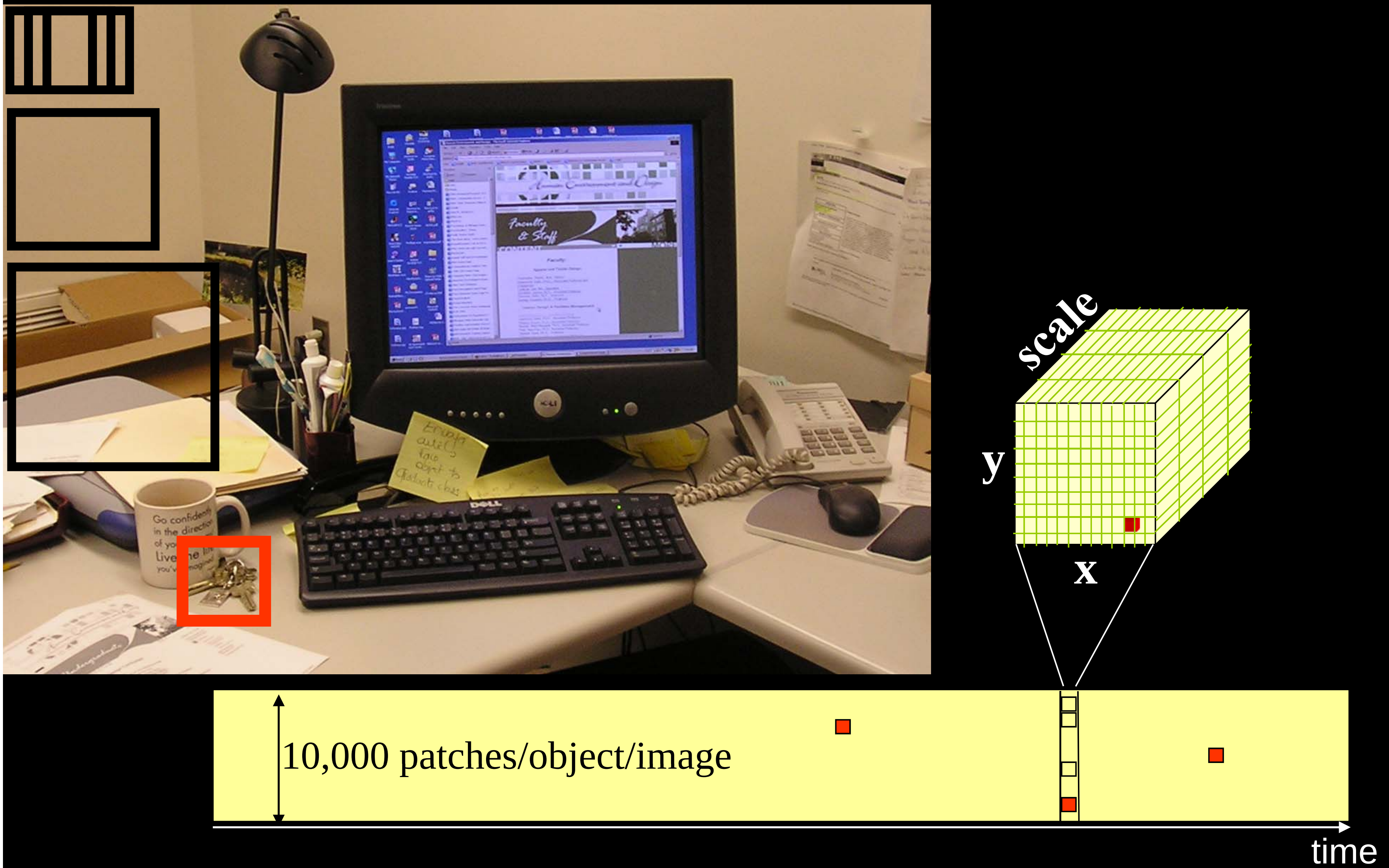
- L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in Computer Vision, 2007.

- G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. **[Brain Kazian]**

- S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. R. Bradski, P. Baumstarck, S. Chung, A. Y. Ng: Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video. IJCAI 2007

- Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008

Why is detection hard?



Plus, we want to do this for ~ 1000 objects

1,000,000 images/day

Standard approach to scene analysis

1) Object representation based on intrinsic features:

2) Detection strategy:

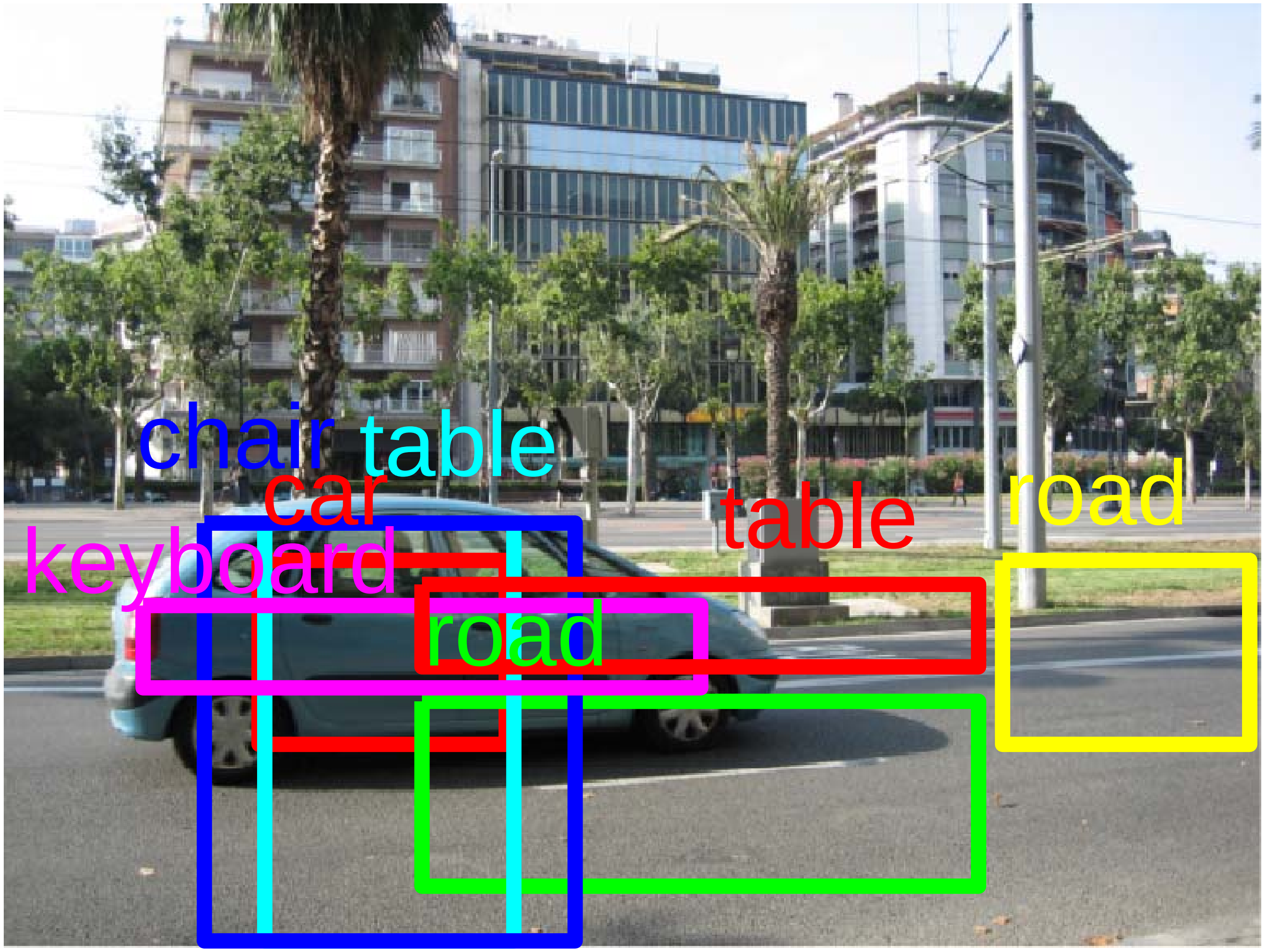
3) The scene representation

Slide credit: A. Torralba

Is local information enough?

Slide credit: A. Torralba

With hundreds of categories



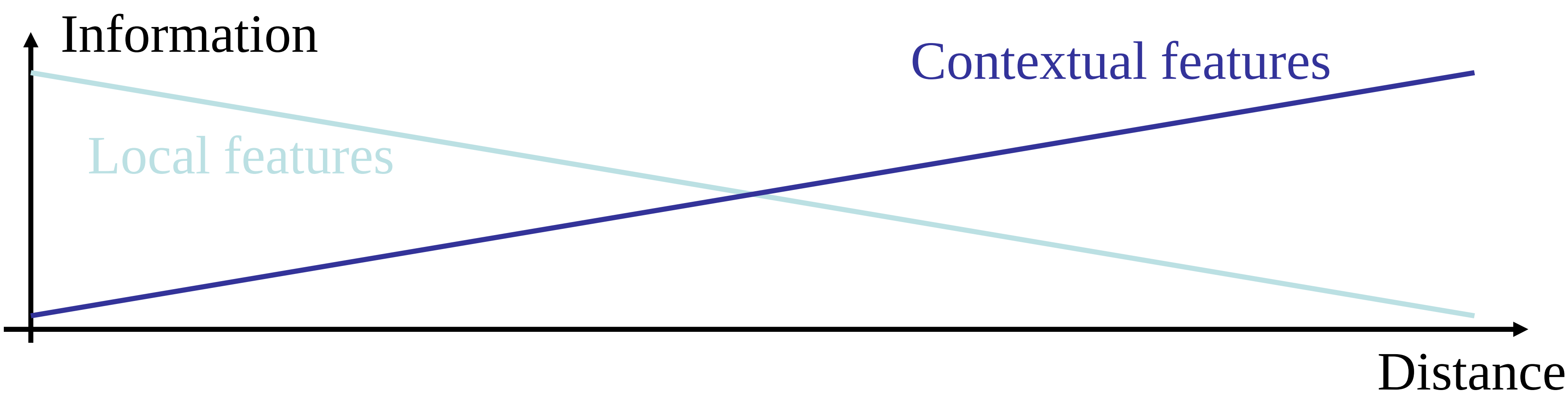
If we have 1000 categories (detectors), and each detector produces 1 fa every 10 images, we will have 100 false alarms per image… pretty much garbage…

Slide credit: A. Torralba

Is local information even enough?

Slide credit: A. Torralba

Is local information even enough?

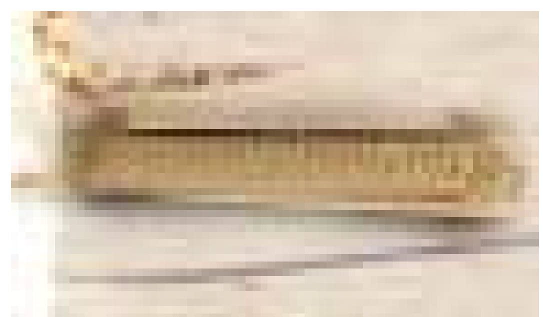


Slide credit: A. Torralba

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



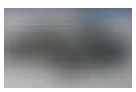
… even if there is one indeed.

Slide credit: A. Torralba

The multiple personalities of a blob

Slide credit: A. Torralba

The multiple personalities of a blob

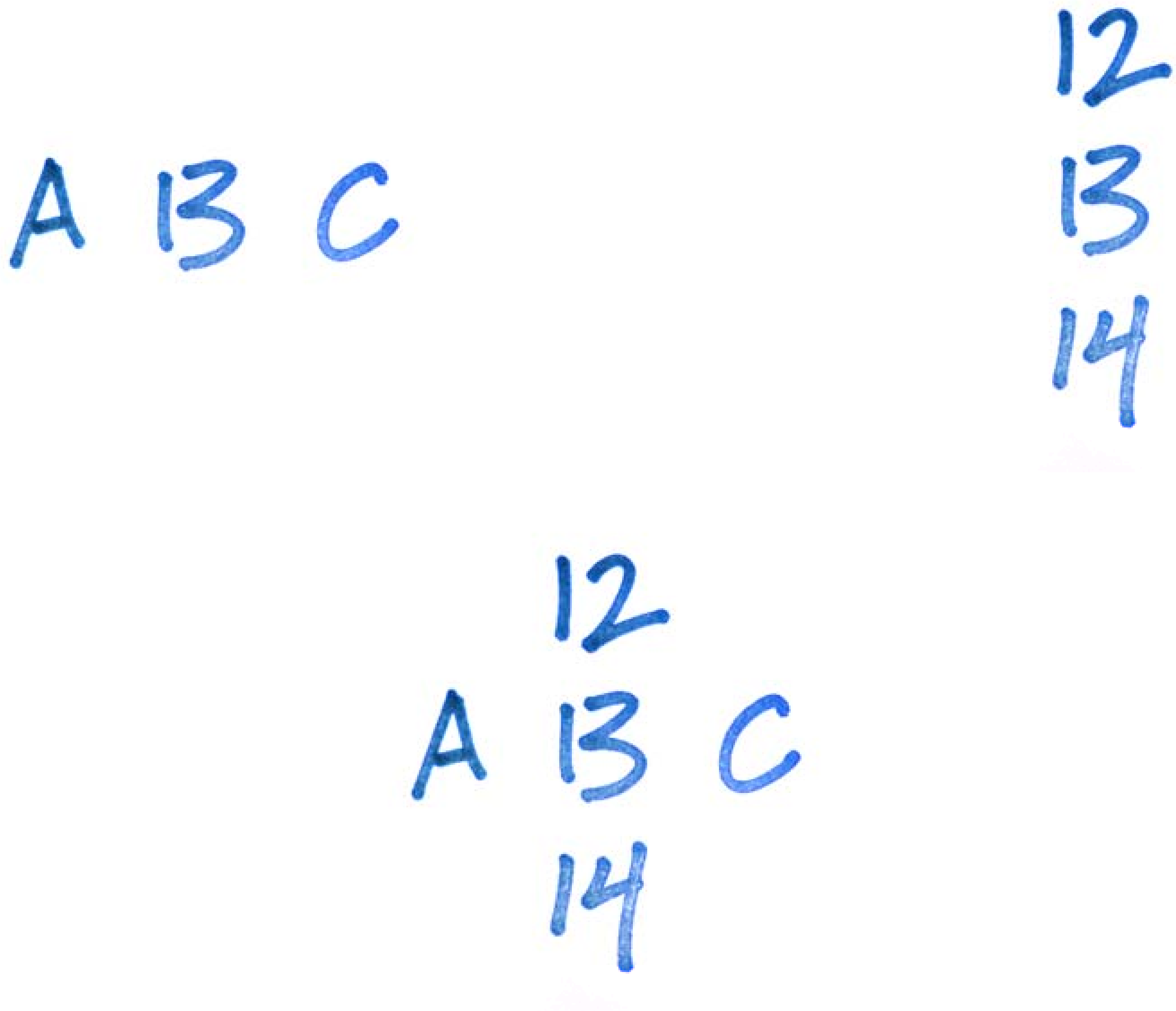


Slide credit: A. Torralba

Slide credit: A. Torralba

12
13
14

Slide credit: A. Torralba

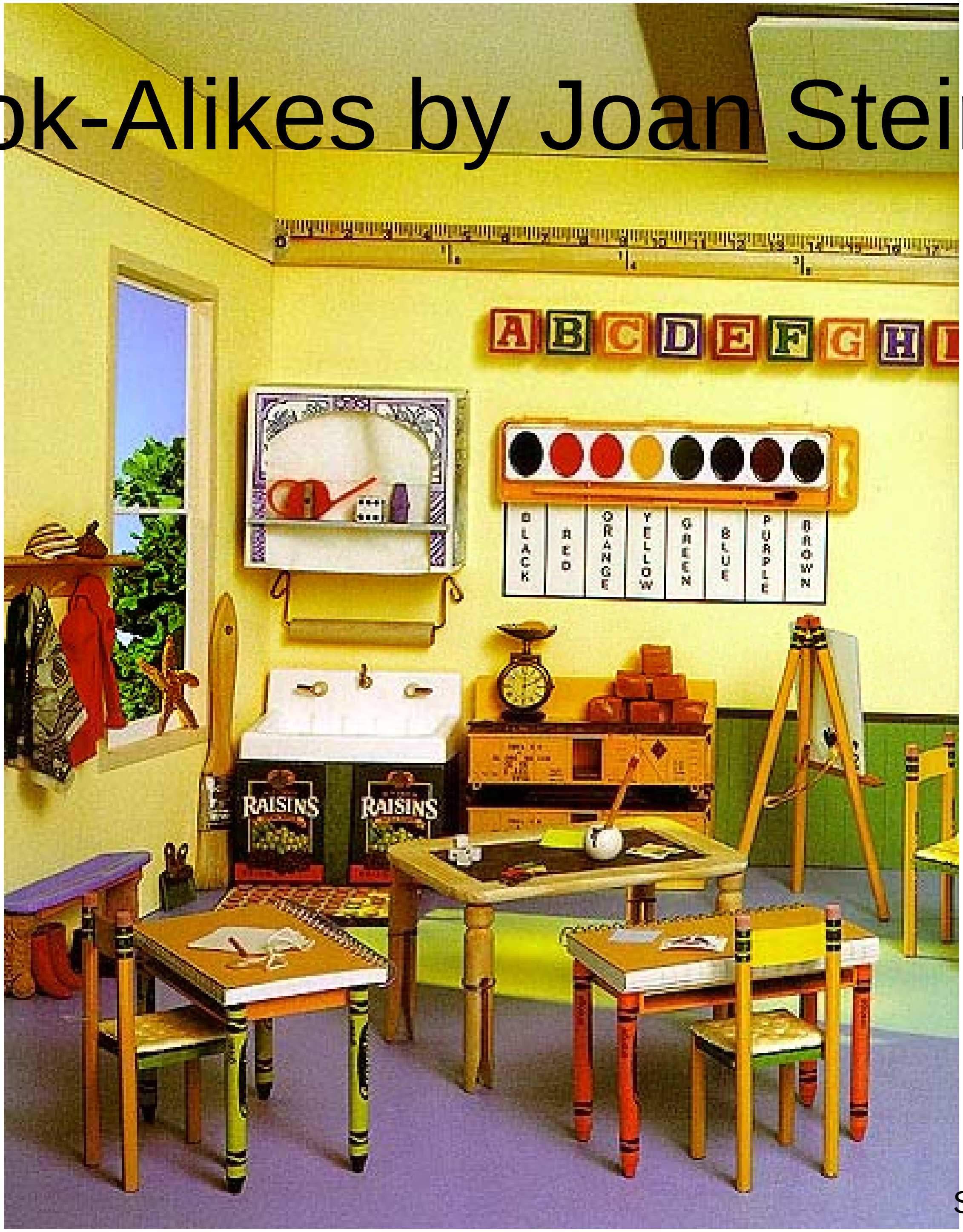


Slide credit: A. Torralba

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 $\wedge$ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-CLEAR $\wedge$ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-CLEAR $\wedge$ TIME-IS-DAY $\wedge$ RGB-IS-AVAILABLE	BLUE
SKY	SKY-IS-OVERCAST $\wedge$ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-OVERCAST $\wedge$ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-OVERCAST $\wedge$ TIME-IS-DAY $\wedge$ RGB-IS-AVAILABLE	WHITE
SKY	SPARSE-RANGE-IS-AVAILABLE	SPARSE-RANGE-IS-UNDEFINED
SKY	CAMERA-IS-HORIZONTAL	NEAR-TOP
SKY	CAMERA-IS-HORIZONTAL $\wedge$ CLIQUE-CONTAINS(complete-sky)	ABOVE-SKYLINE
SKY	CLIQUE-CONTAINS(sky)	SIMILAR-INTENSITY
SKY	CLIQUE-CONTAINS(sky)	SIMILAR-TEXTURE
SKY	RGB-IS-AVAILABLE $\wedge$ 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-HORIZONTA
GROUND	DENSE-RANGE-IS-AVAILABLE	DENSE-RANGES-FORM-HORIZONTA
GROUND	CAMERA-IS-HORIZONTAL $\wedge$ CLIQUE-CONTAINS(complete-ground)	BELOW-SKYLINE
GROUND	CAMERA-IS-HORIZONTAL $\wedge$ CLIQUE-CONTAINS(geometric-horizon) $\wedge$ $\neg$ CLIQUE-CONTAINS(skyline)	BELOW-GEOMETRIC-HORIZON
GROUND	TIME-IS-DAY	DARK

Slide credit: A. Torralba

Today – **Image 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. **[Patrick Sundberg]**
- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 **[Robert Carroll]**
- L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in Computer Vision, 2007.
- G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. **[Brain Kazian]**
- S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. R. Bradski, P. Baumstarck, S. Chung, A. Y. Ng: Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video. IJCAI 2007
- Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008

Multiclass object detection and context modeling

Antonio Torralba

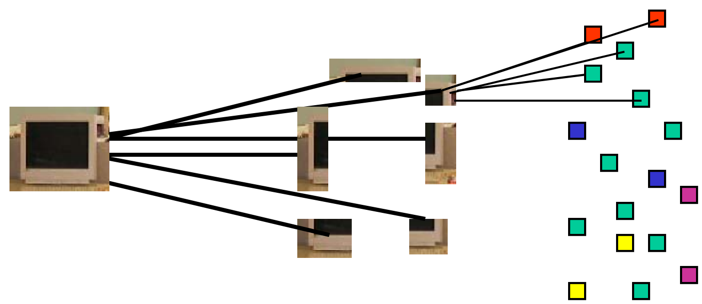
In collaboration with

Kevin P. Murphy and William T. Freeman

Object representations

Inside the object

(intrinsic features)



Object size

Global appearance

Parts

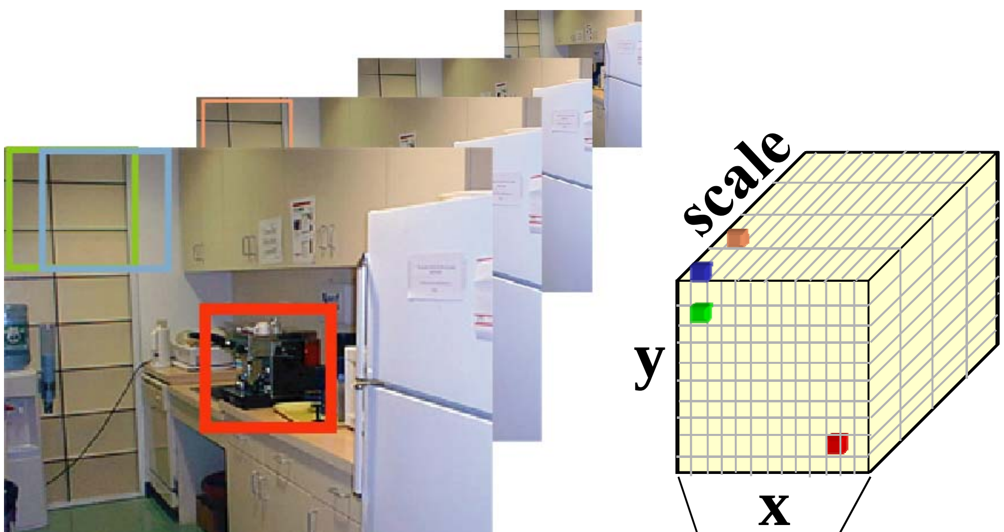
Pixels

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.

1) Search space is HUGE

“Like finding needles in a haystack”

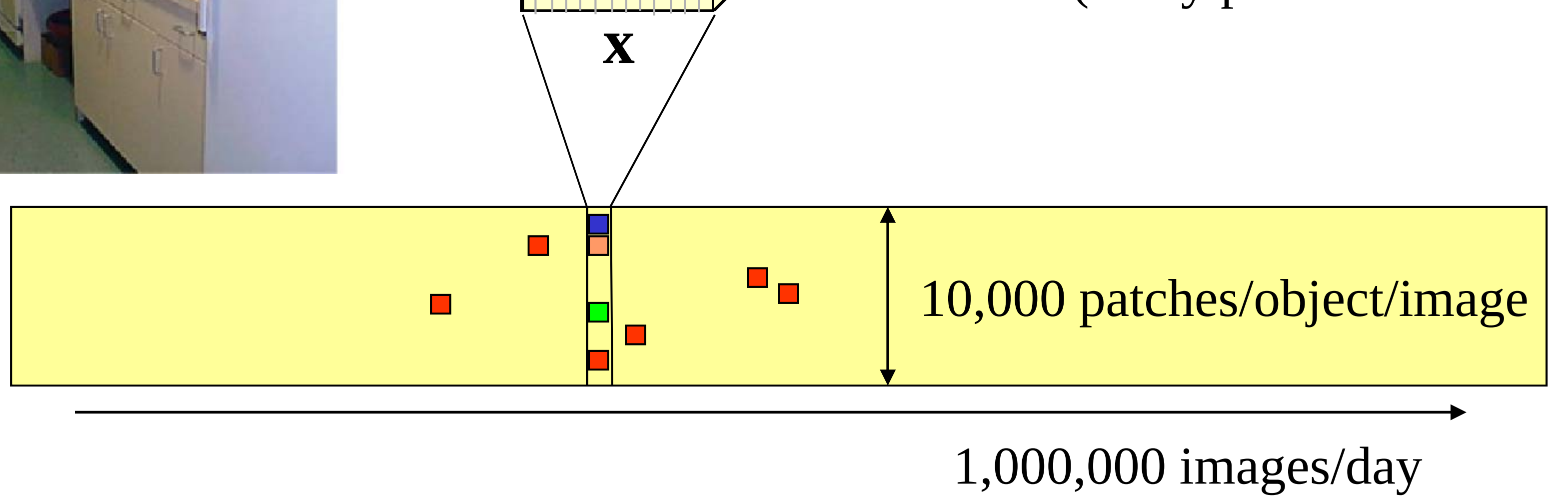


For each object:

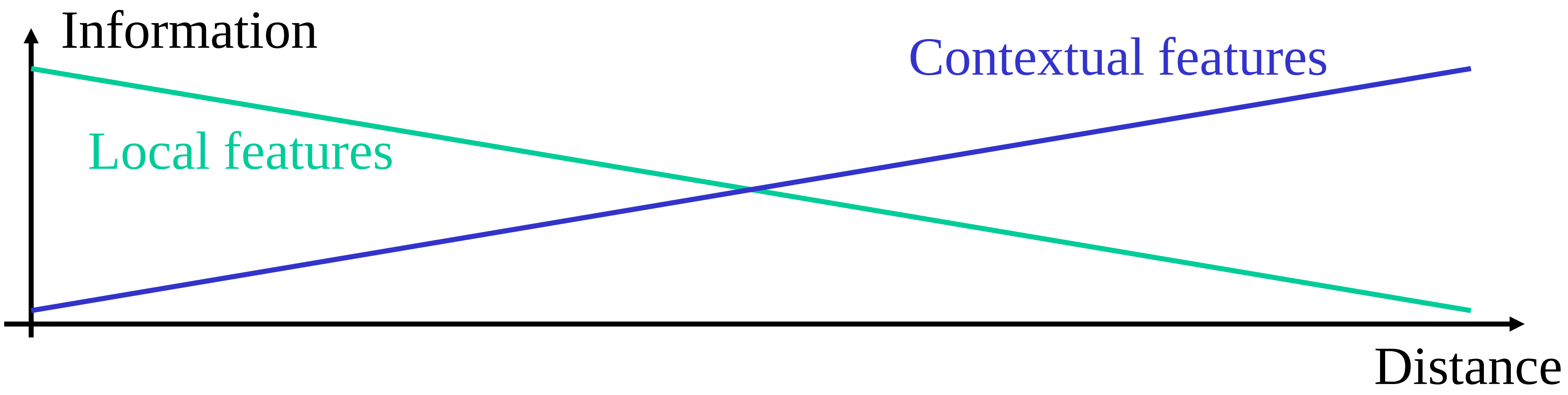
- Need to search over locations and scales

- Error prone (classifier must have very low false positive rate)

- Slow (many patches to examine)



2) Local features are not even sufficient

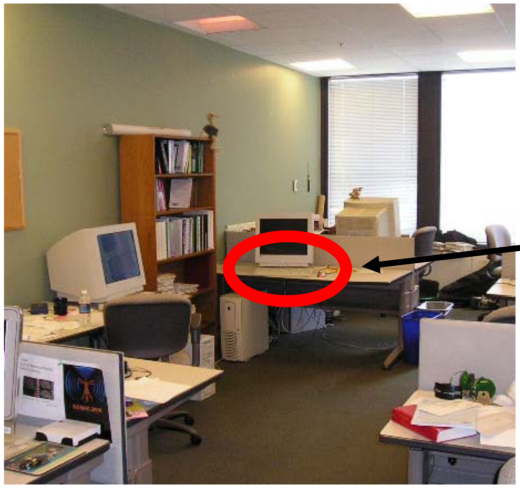


Symptoms of local features only

Some false alarms occur in image regions in which is impossible for the target to be present given the context.

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

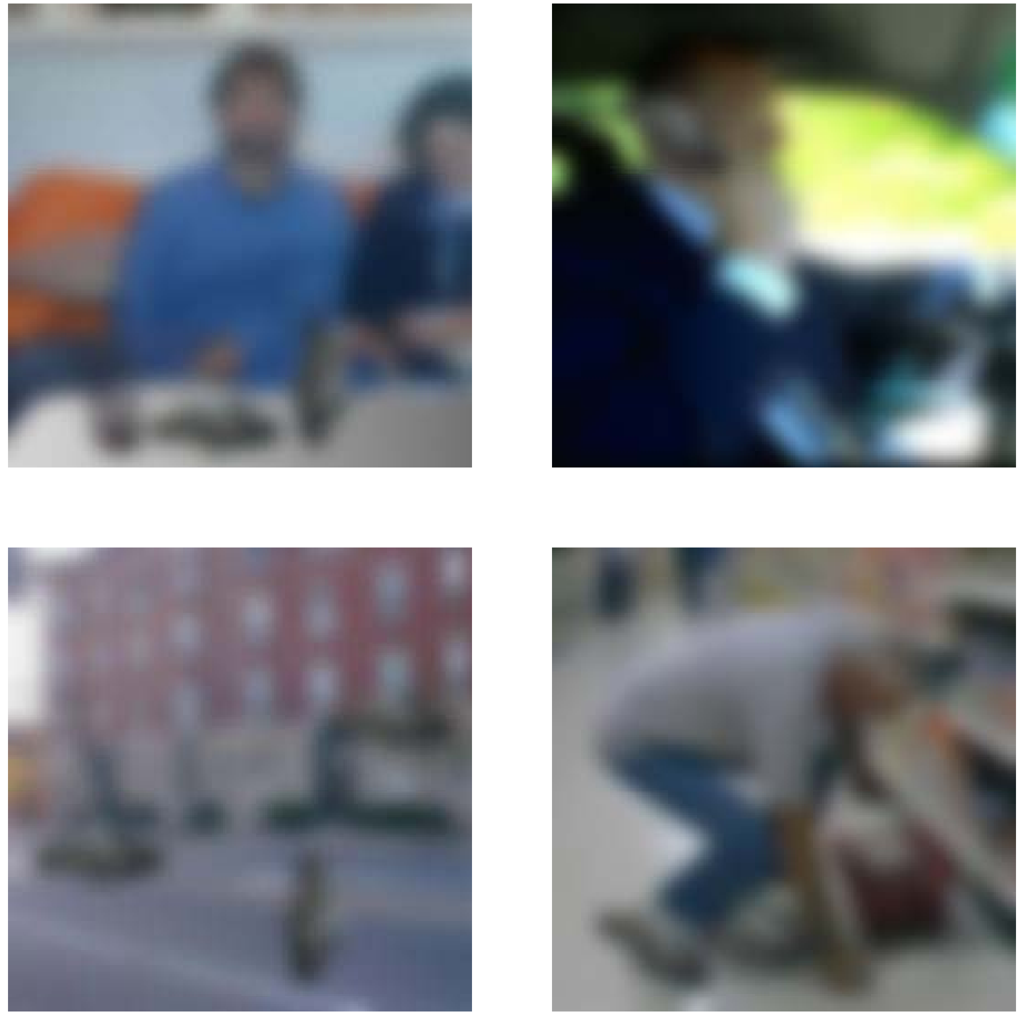
We know there is a keyboard present in this scene even if we cannot see it clearly.



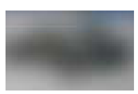
We know there is no keyboard present in this scene

… even if there is one indeed.

The multiple personalities of a blob



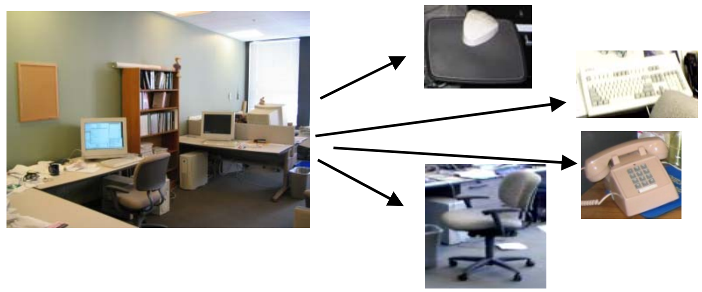
The multiple personalities of a blob



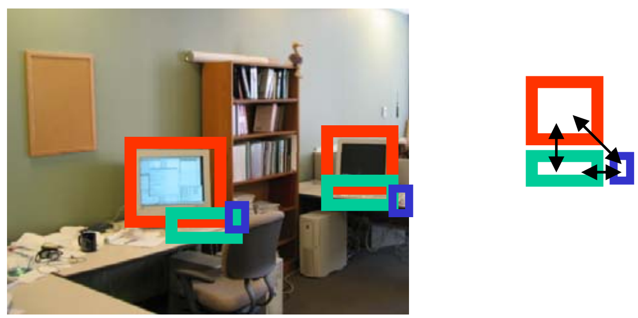
Human vision: Biederman, Bar & Ullman, Palmer, …

What is context

- Scenes



- Other objects



- Properties of objects and scenes (pose, style, etc.)

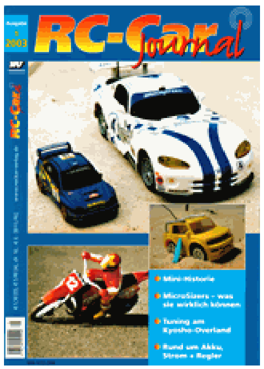
Conditional random fields

Conditional random fields

Conditional random fields

Why is context important?

- Changes the interpretation of an object (or its function)

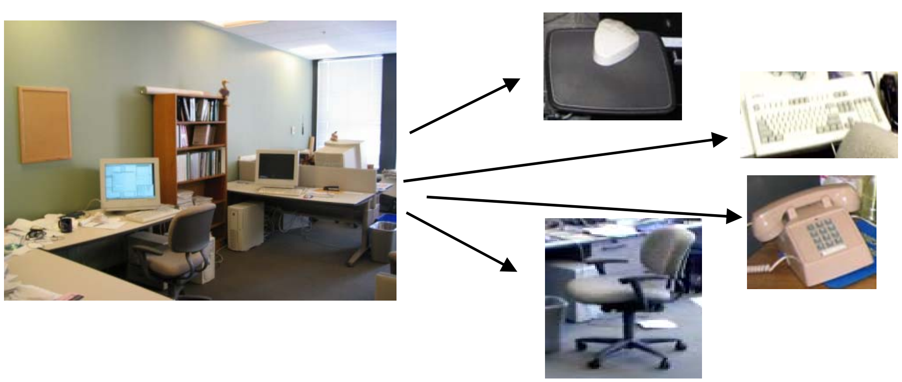


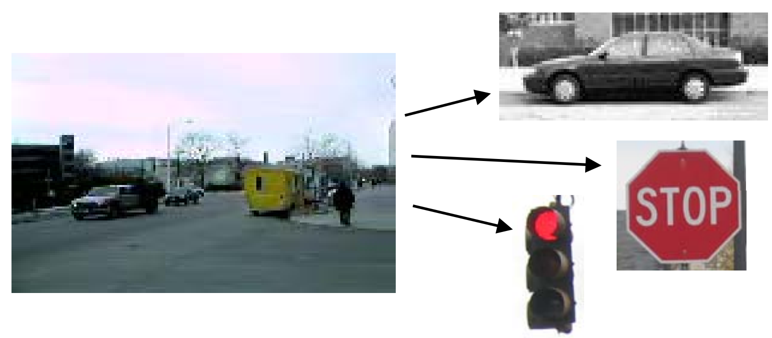
- Context defines what an unexpected event is

Why is context important?

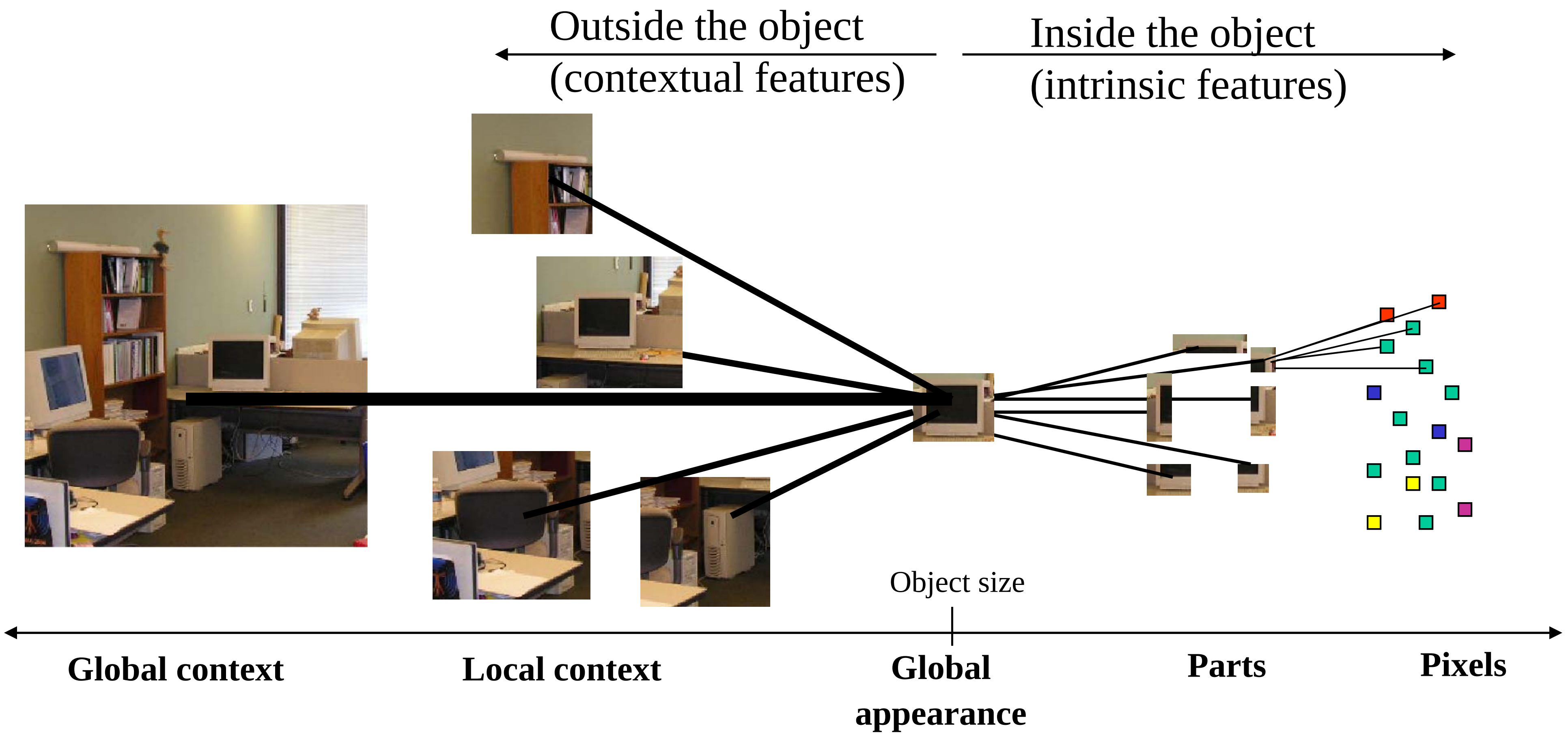
• Reduces the search space

• Context features can be shared among many objects across locations and scales: more efficient than local features.





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 $\wedge$ TIME-IS-DAY	BRIGHT
43	SKY	SKY-IS-CLEAR $\wedge$ TIME-IS-DAY	UNTEXTURED
44	SKY	SKY-IS-CLEAR $\wedge$ TIME-IS-DAY $\wedge$ RGB-IS-AVAILABLE	BLUE
45	SKY	SKY-IS-OVERCAST $\wedge$ TIME-IS-DAY	BRIGHT
46	SKY	SKY-IS-OVERCAST $\wedge$ TIME-IS-DAY	UNTEXTURED
47	SKY	SKY-IS-OVERCAST $\wedge$ TIME-IS-DAY $\wedge$ 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 $\wedge$ 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 $\wedge$ 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 $\wedge$ CLIQUE-CONTAINS(complete-ground)	BELOW-SKYLINE
66	GROUND	CAMERA-IS-HORIZONTAL $\wedge$ CLIQUE-CONTAINS(geometric-horizon) $\wedge$ $\neg$ 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 $\wedge$ 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

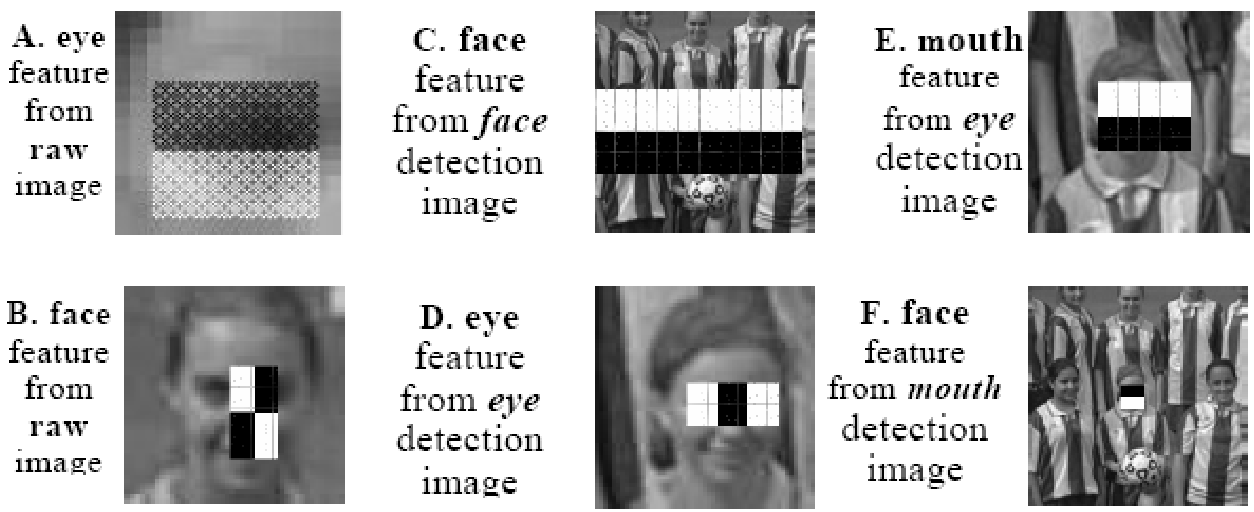
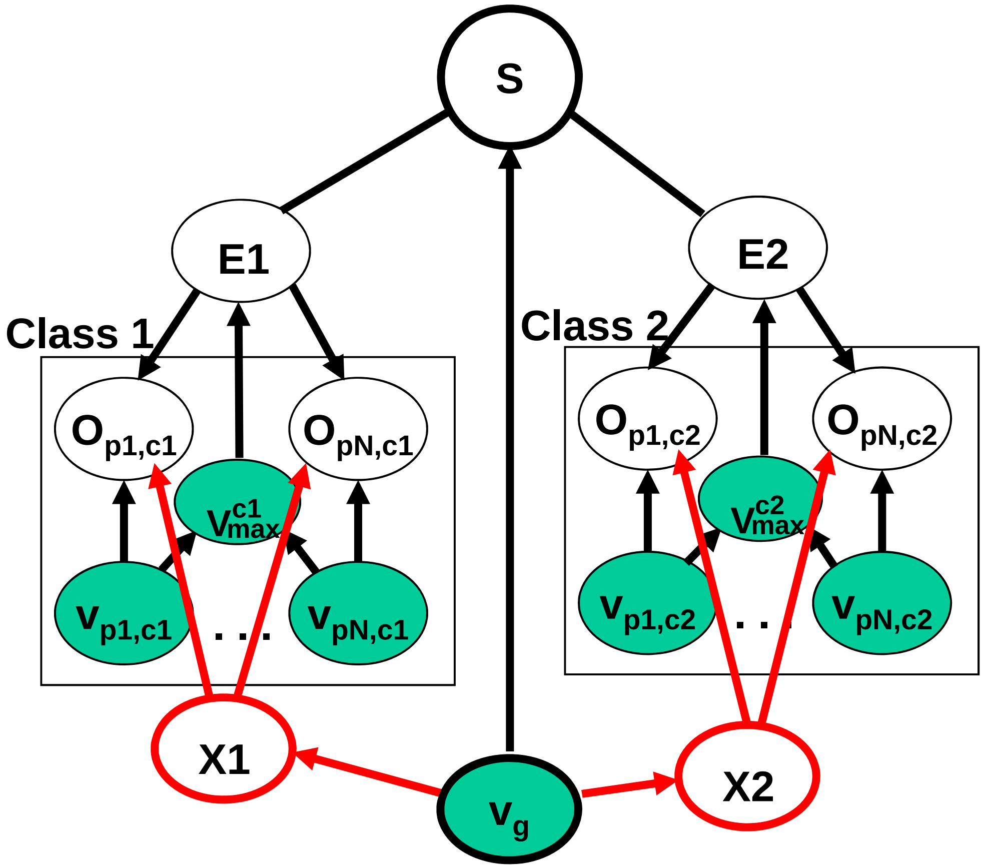


Figure 5: **A-E.** Emerging features of eyes, mouths and faces (presented on windows of raw images for legibility). The windows' scale is defined by the detected object size and by the map mode (local or contextual). **C.** faces are detected using face detection maps H^{Face}, exploiting the fact that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

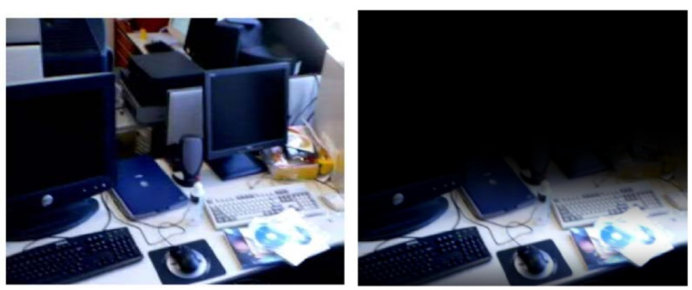
Previous work on context

- Murphy, Torralba & Freeman (03)

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

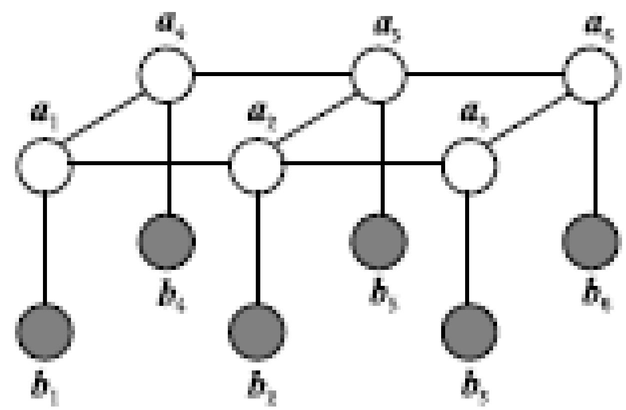


Keyboards

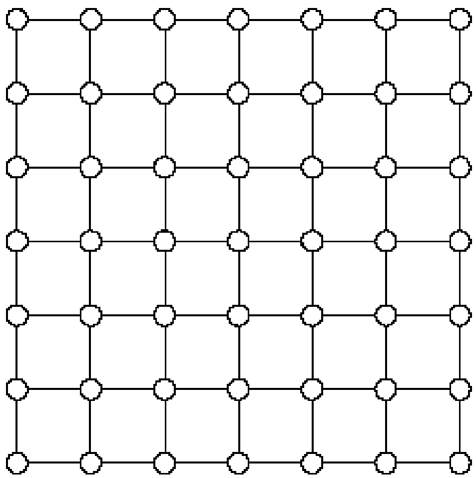


Previous work on context

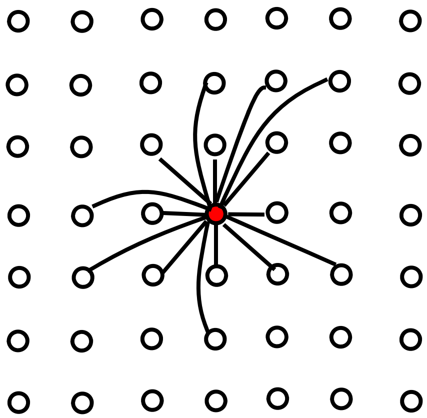
- Carbonetto, de Freitas & Barnard (04)
- Enforce spatial consistency between labels using MRF



Graphical models for image labeling



Nearest neighbor grid



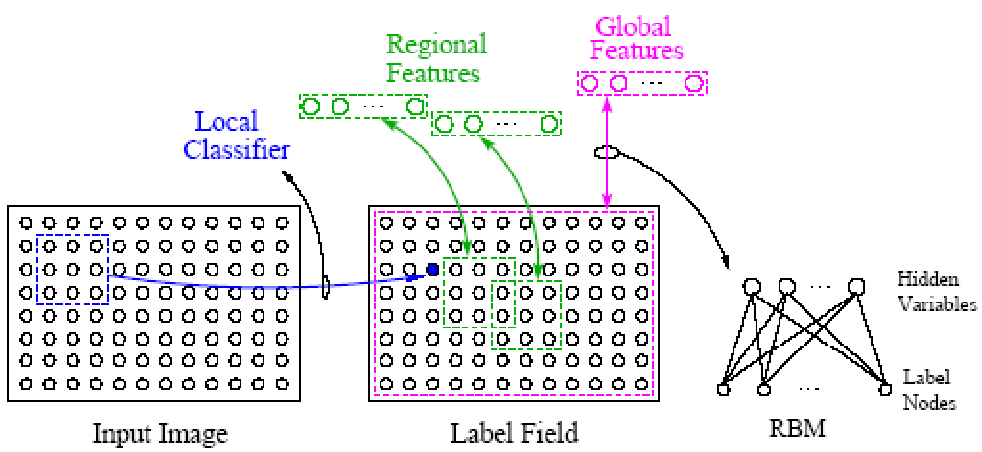
Densely connected graphs
with low informative connections

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)

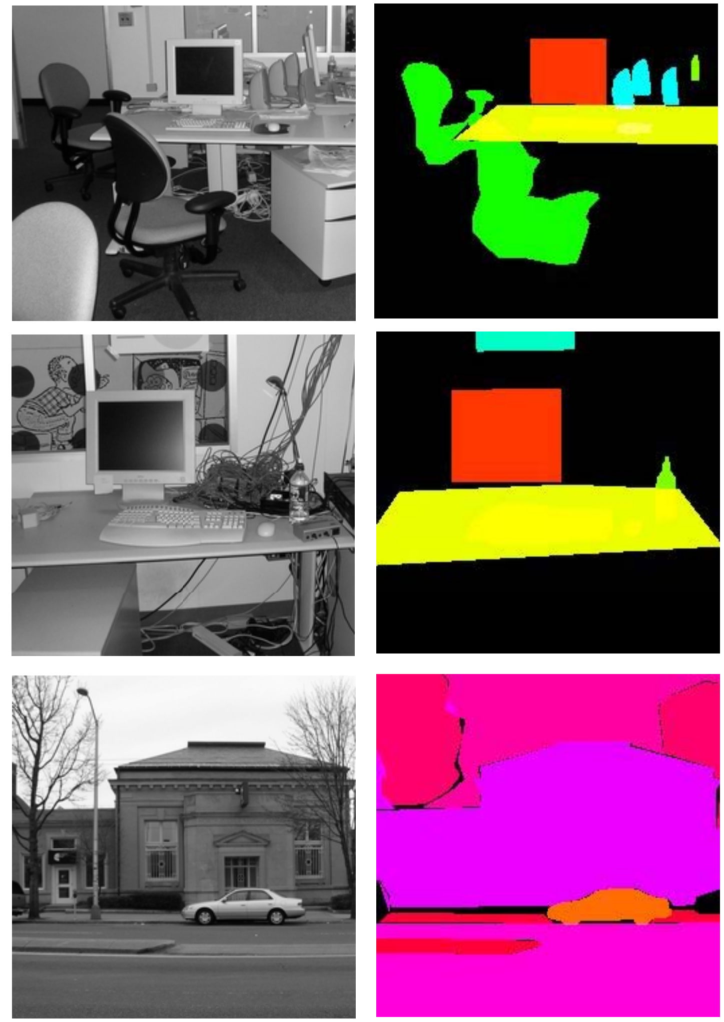


Outline of this talk

- 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

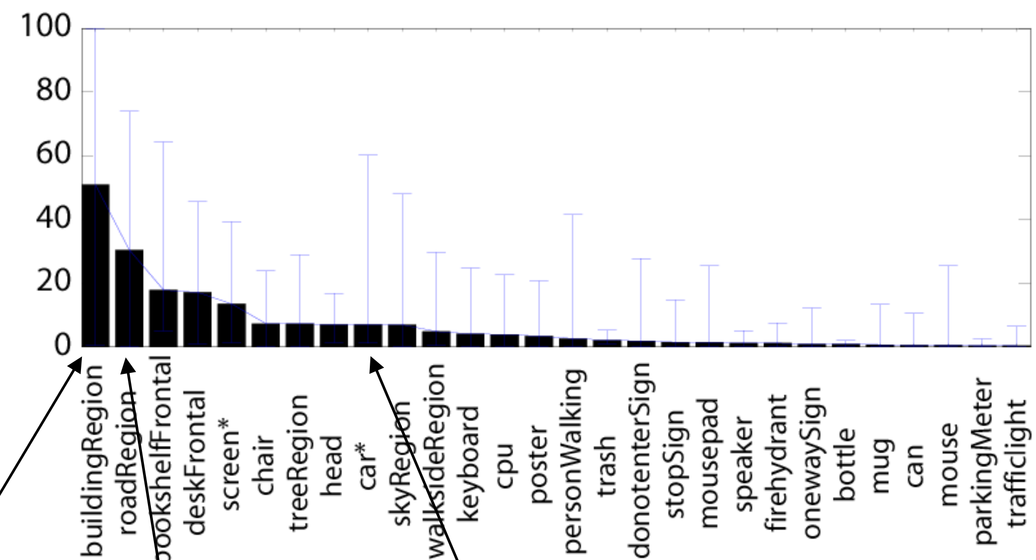
Image database

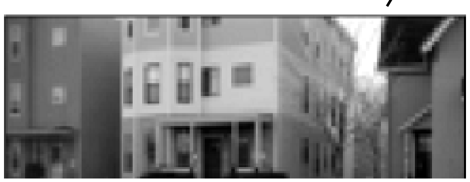
- ~2500 hand labeled images with segmentations
- ~30 objects and stuff
- Indoor and outdoor
- Sets of images are separated by locations and camera (digital/webcam)
- No graduate students or low-income-student-class exploited for labeling.

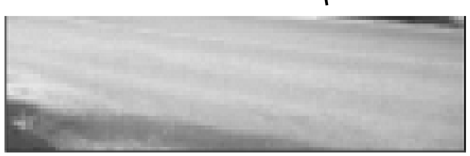


Which objects are important?

Average percentage of pixels occupied by each object.



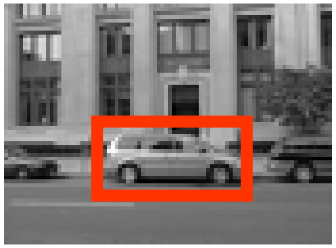




Object representation

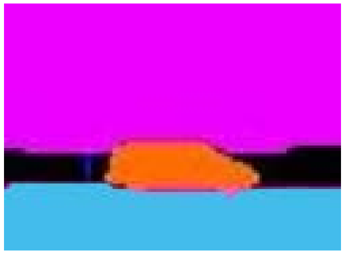
- **Discrete/**bounded/rigid

Screen, car, pedestrian, bottle, …



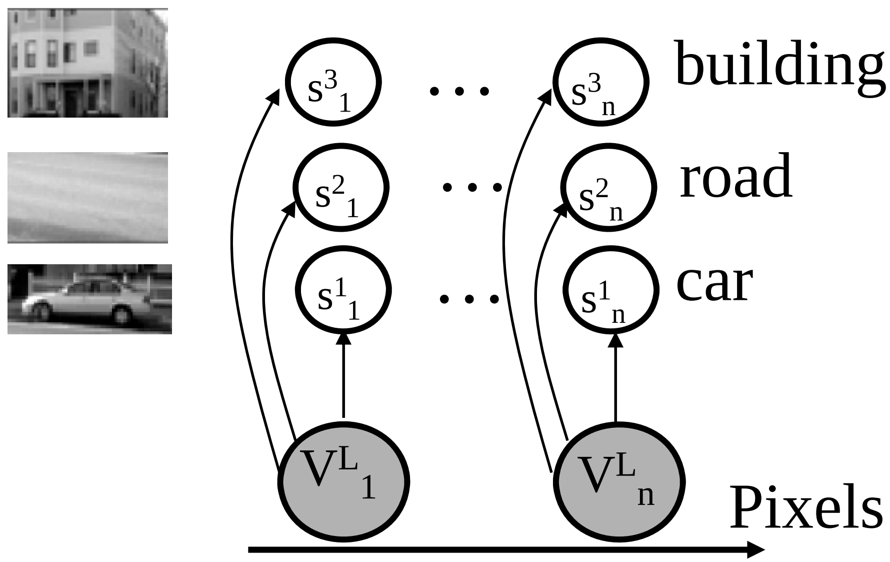
- **Extended**/unbounded/deformable

Building, sky, road, shelves, desk, …



We will use region labeling as a representation.

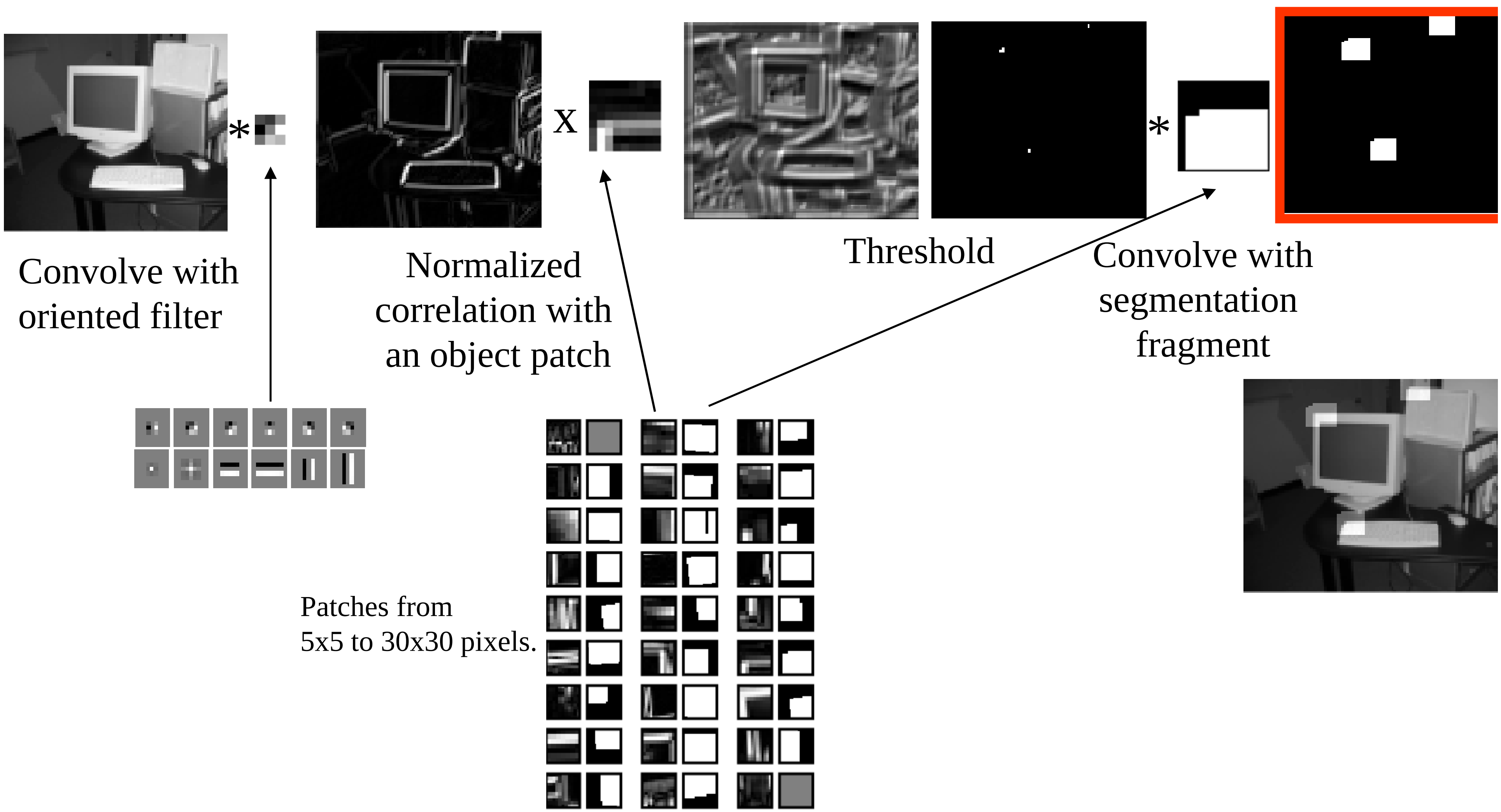
Learning local features
(intrinsic object features)



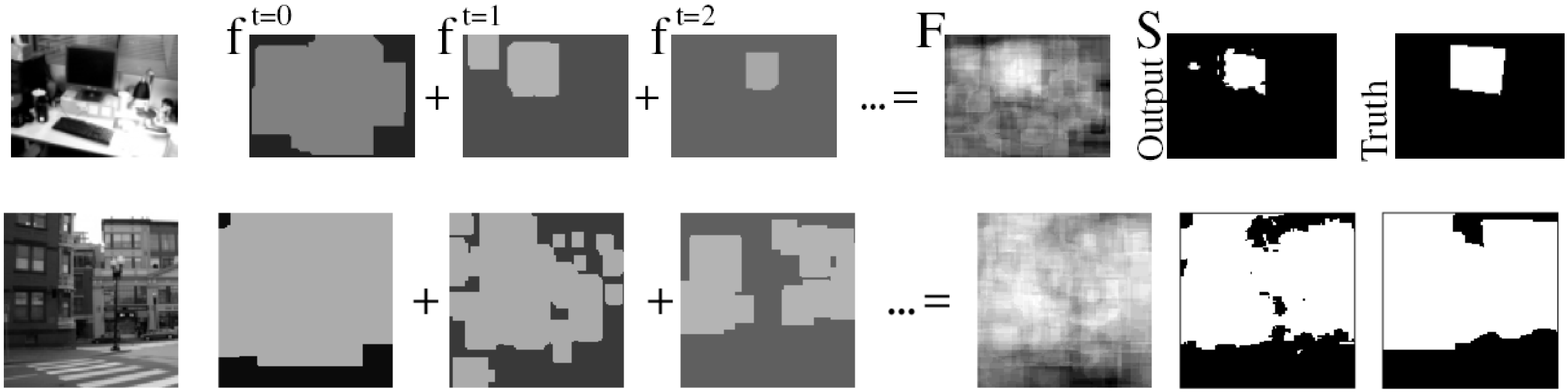
We maximize the probability of the true labels using Boosting.

Object local features

(Borenstein & Ullman, ECCV 02)

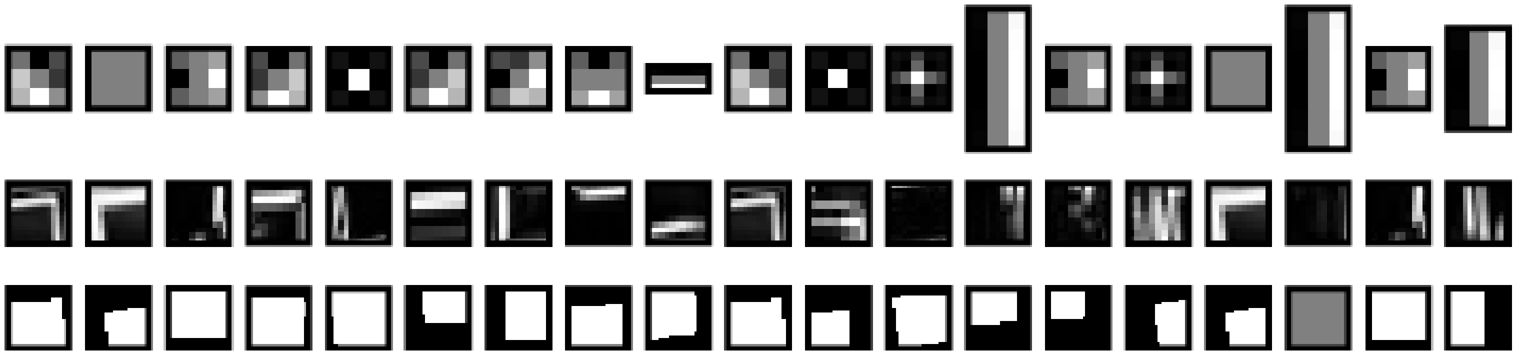


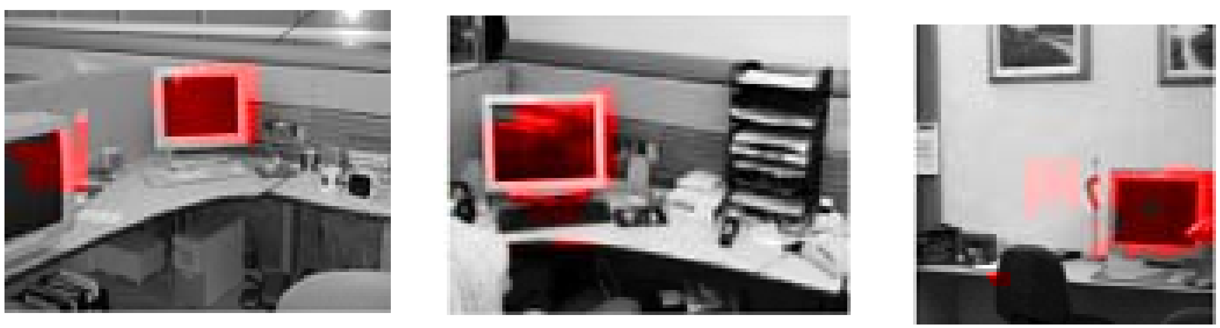
Results with local features



Results with local features

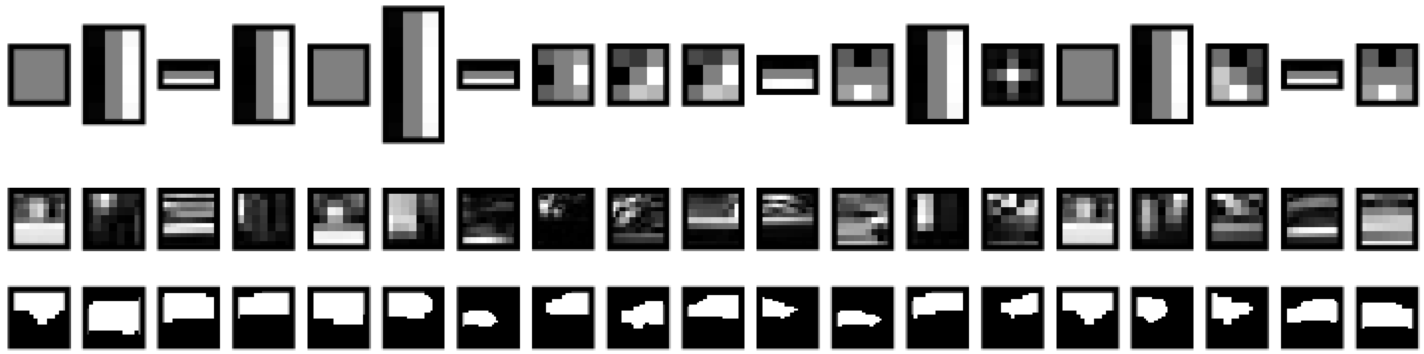
Screen

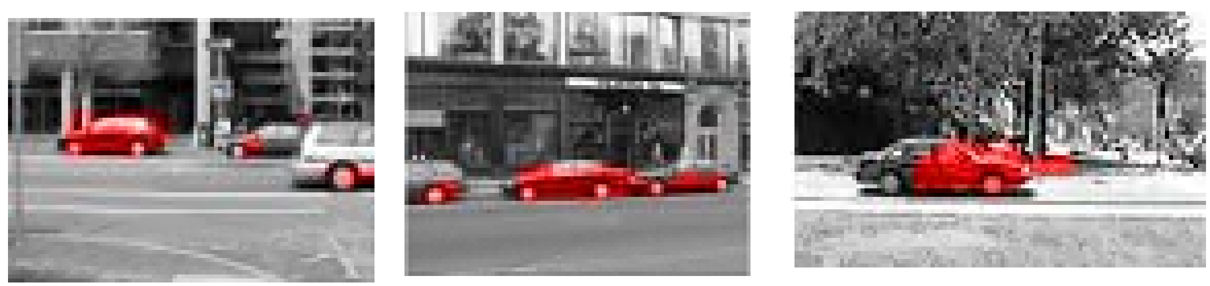




Results with local features

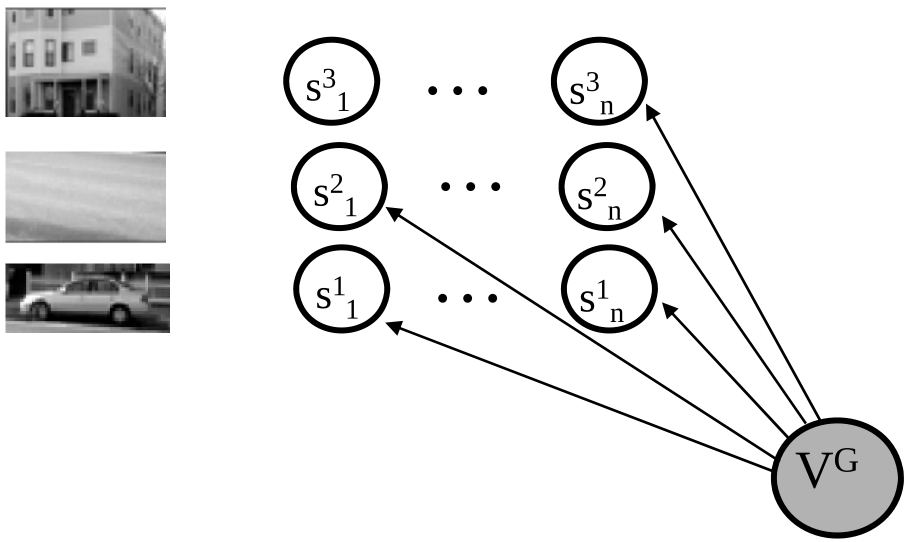
Car





Global context: location priming

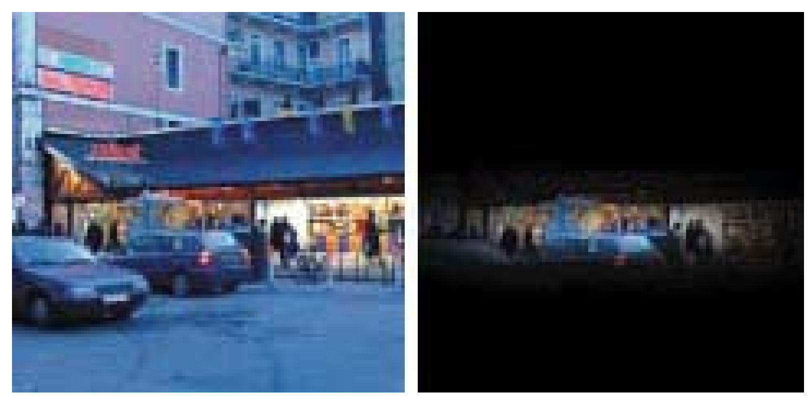
How far can we go without object detectors?



Context features that represent the scene instead of other objects.

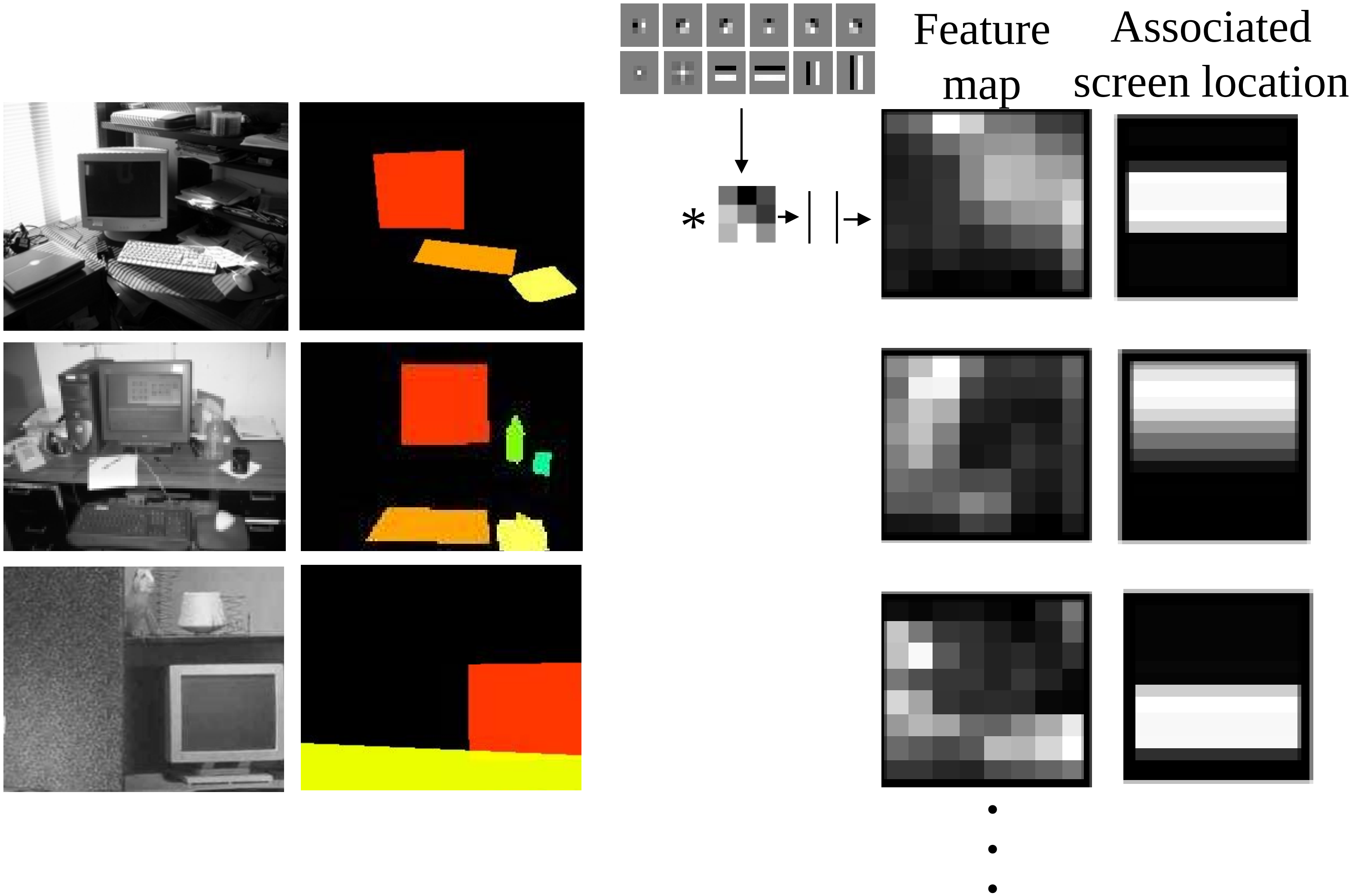
The global features can provide:

- Object presence
- Location priming
- Scale priming



Object global features

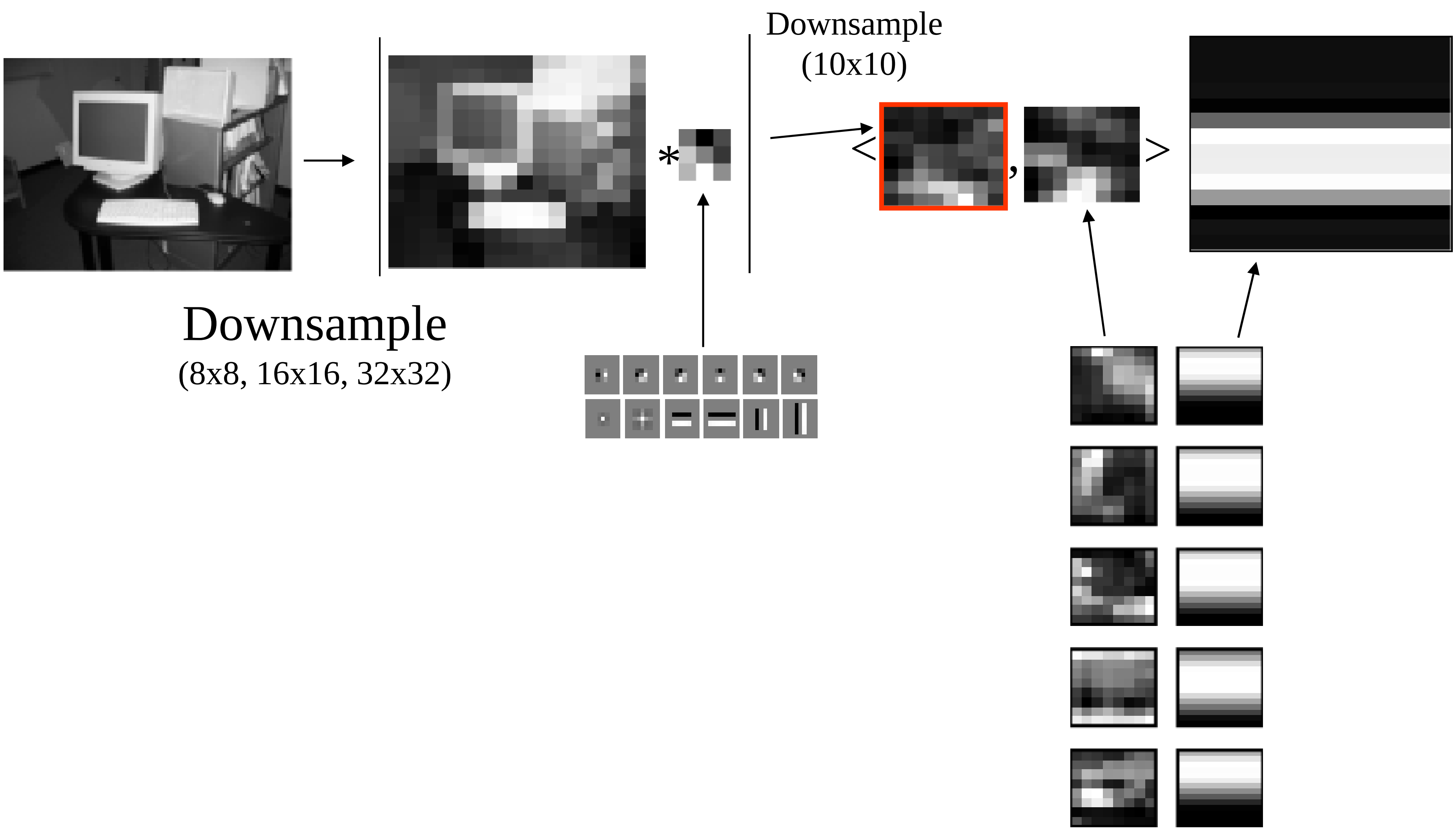
First we create a dictionary of scene features and object locations:



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

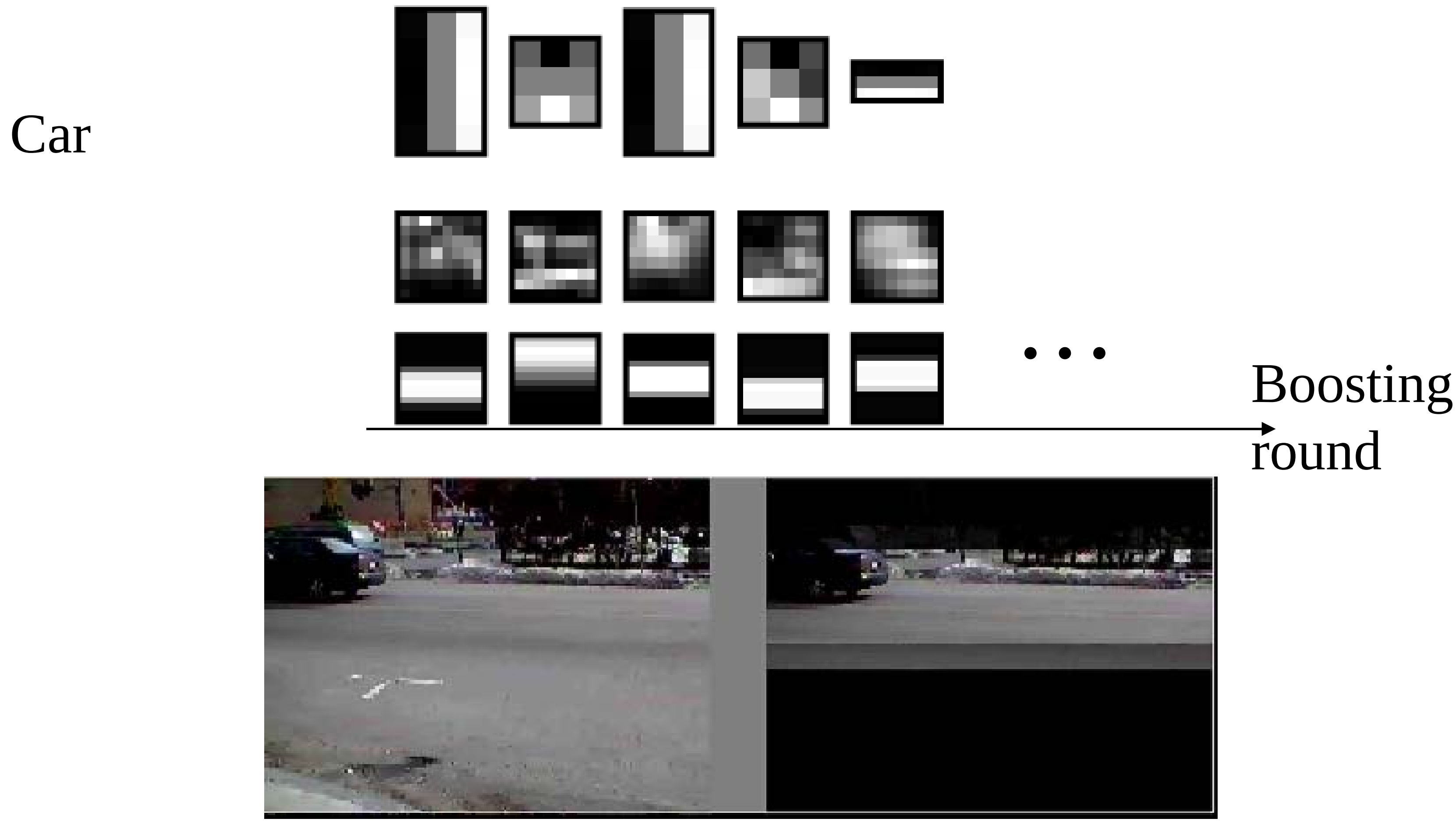
Object global features

How to compute the global features

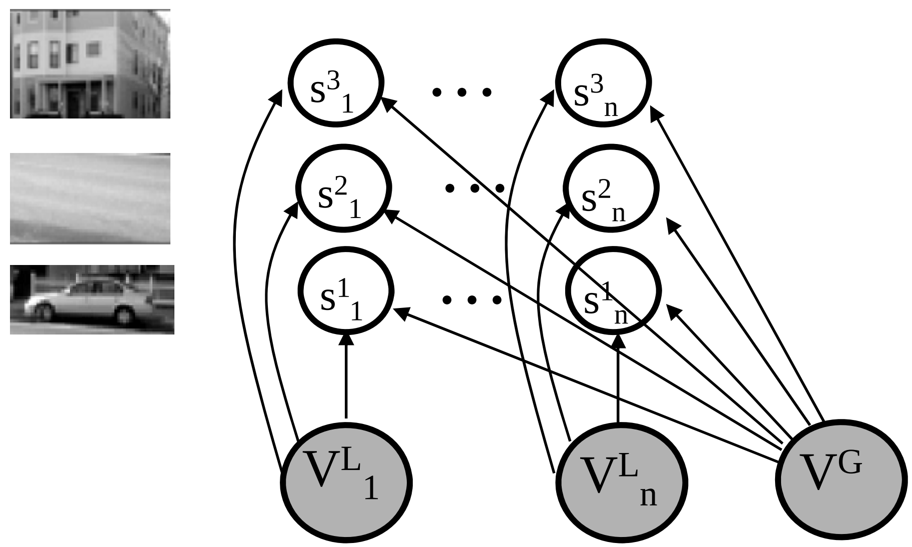


Car detection with global features

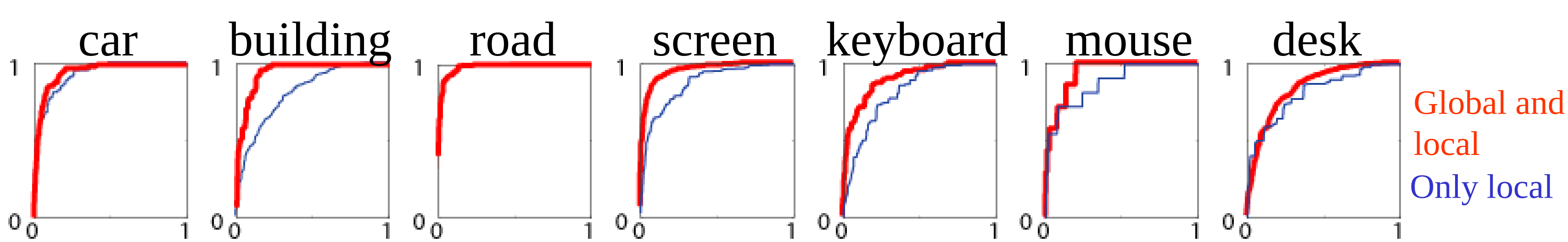
Features selected by boosting:



Combining global and local

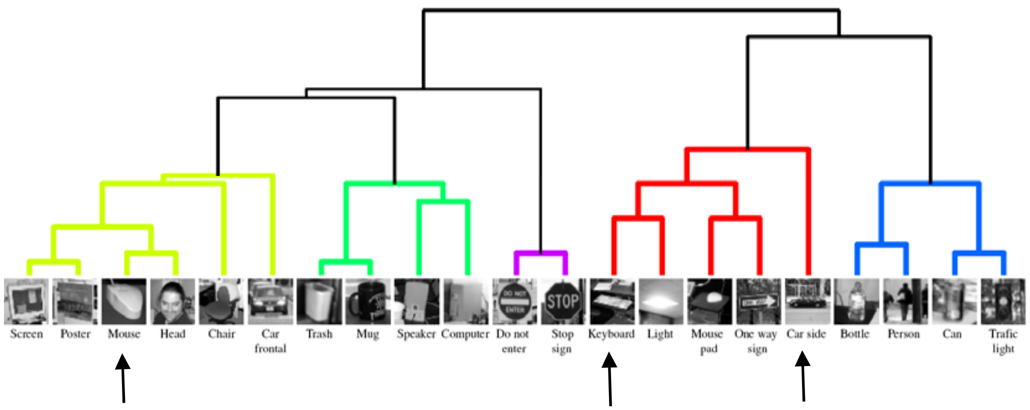


ROC for same total number of features (100 boosting rounds):

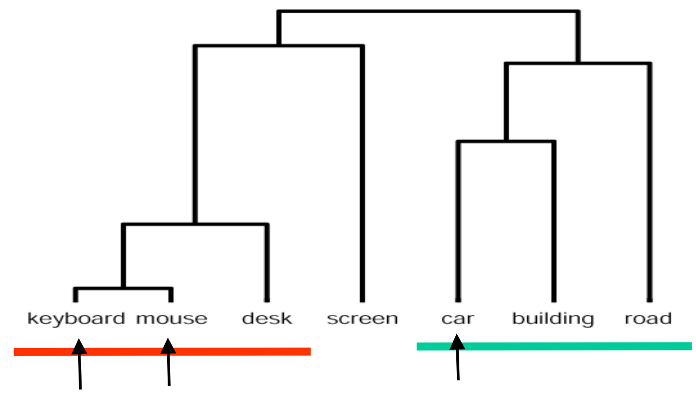


Clustering of objects with local and global feature sharing

Clustering with local features



Clustering with global and local features

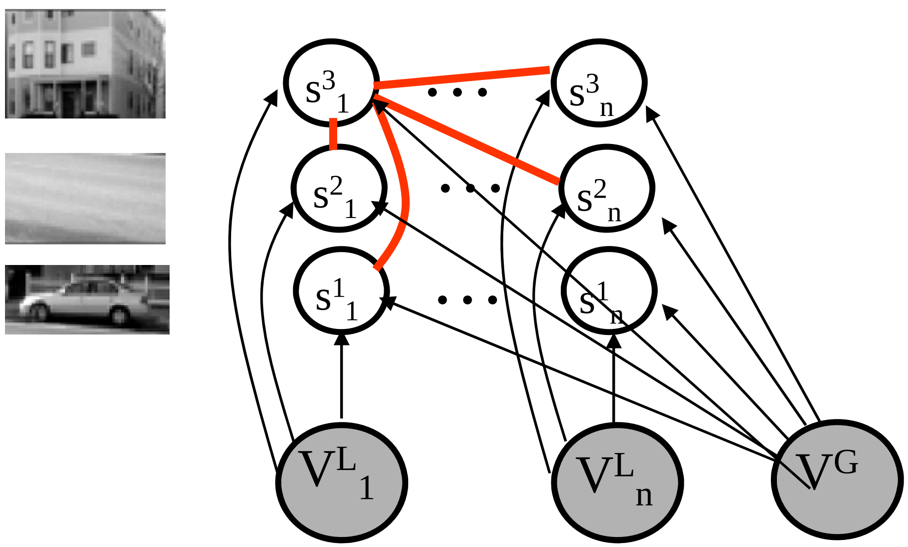


Objects are similar if they share local features and they appear in the same contexts.

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



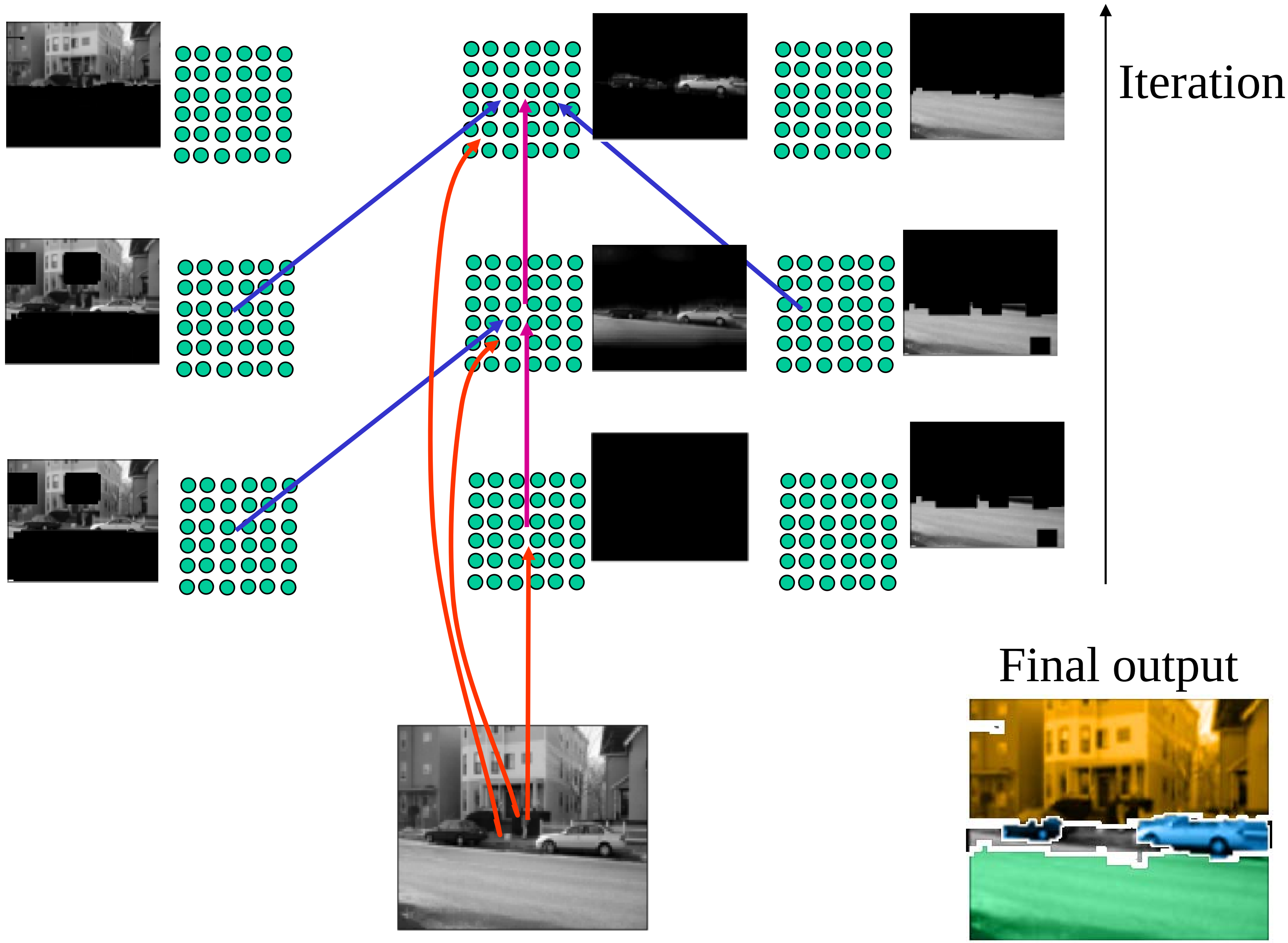
We need to learn

- 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

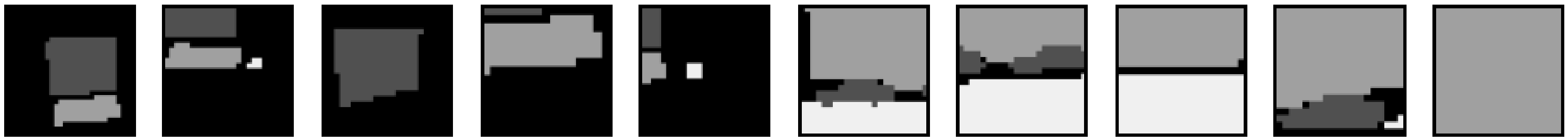


Sequentially learning the structure

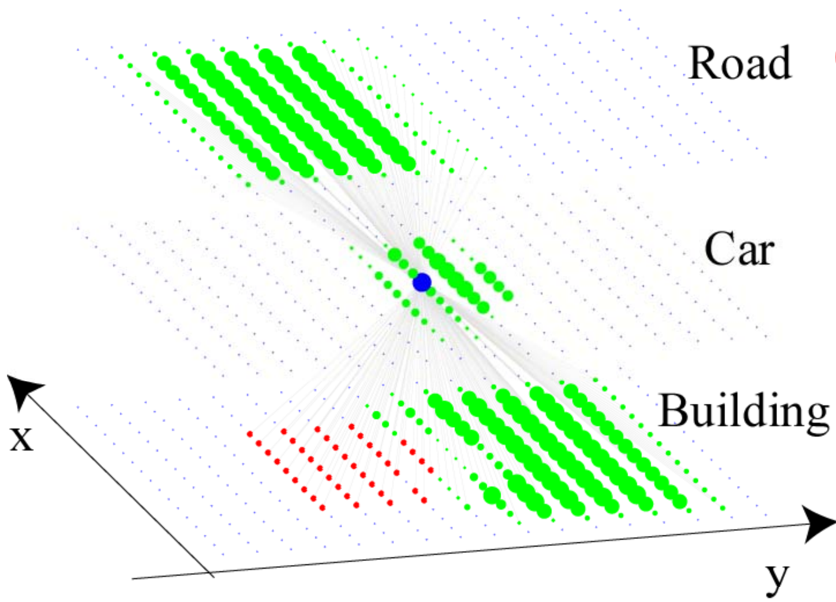
At each iteration of boosting

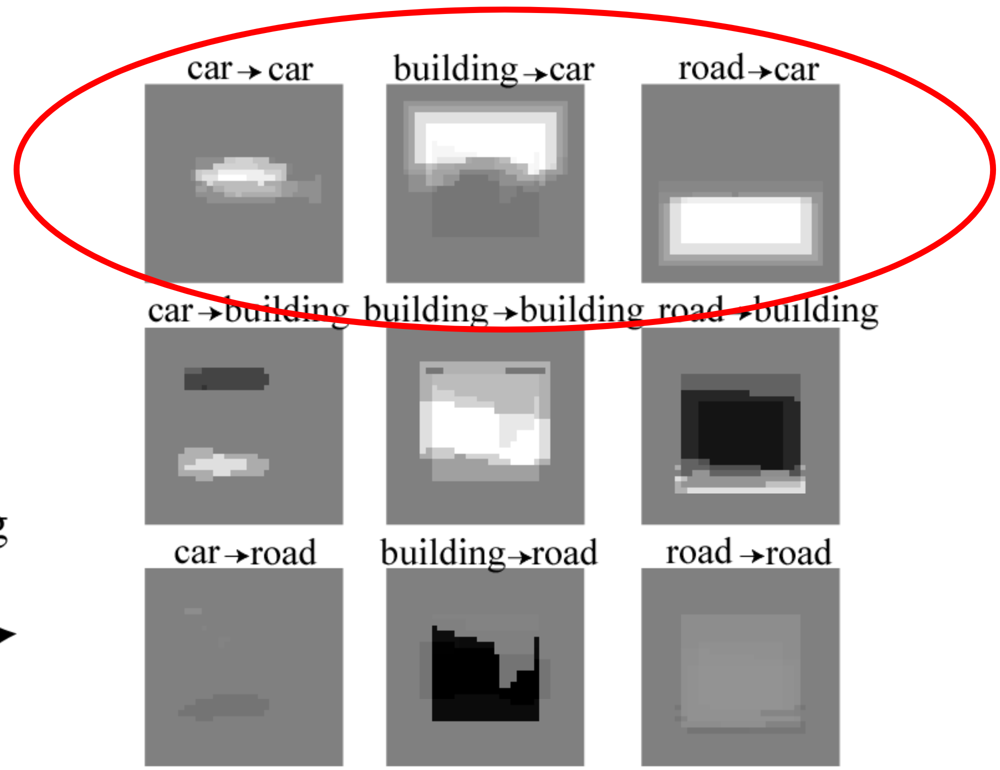
•We pick a weak learner applied to the image
(local or global features)

•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.



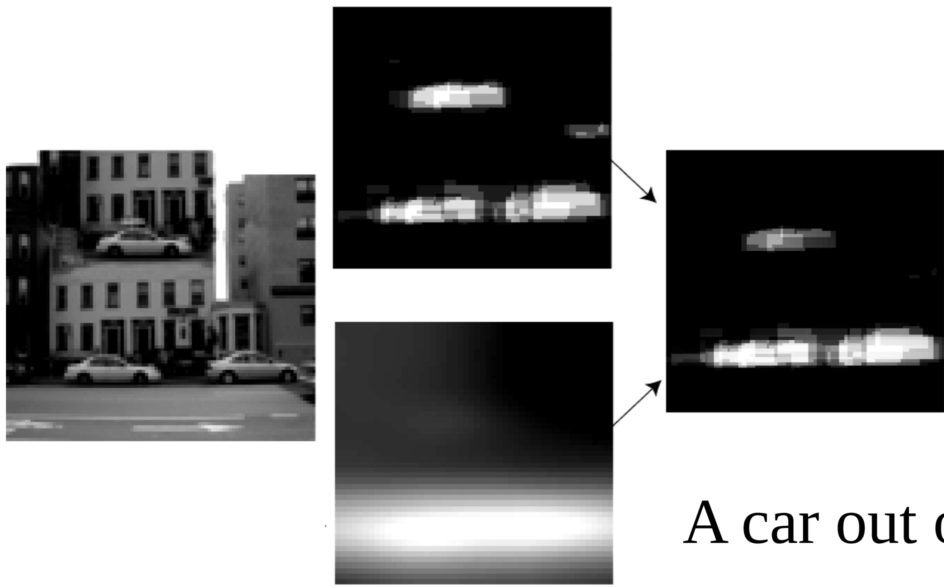
Car detection



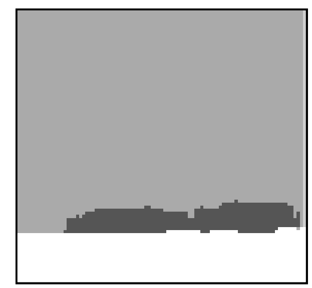


Car detection

From intrinsic features

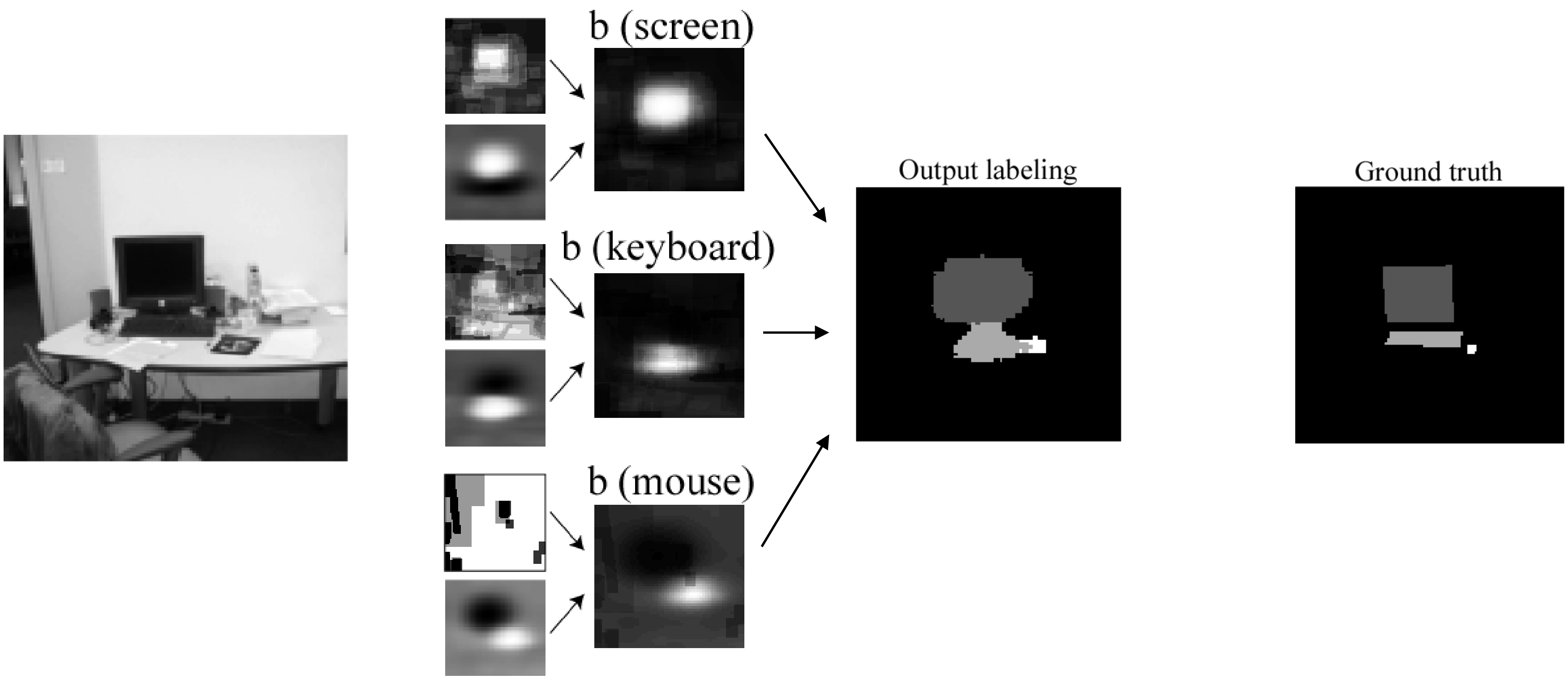


From contextual features



A car out of context is less of a car

Screen/keyboard/mouse



Cascade

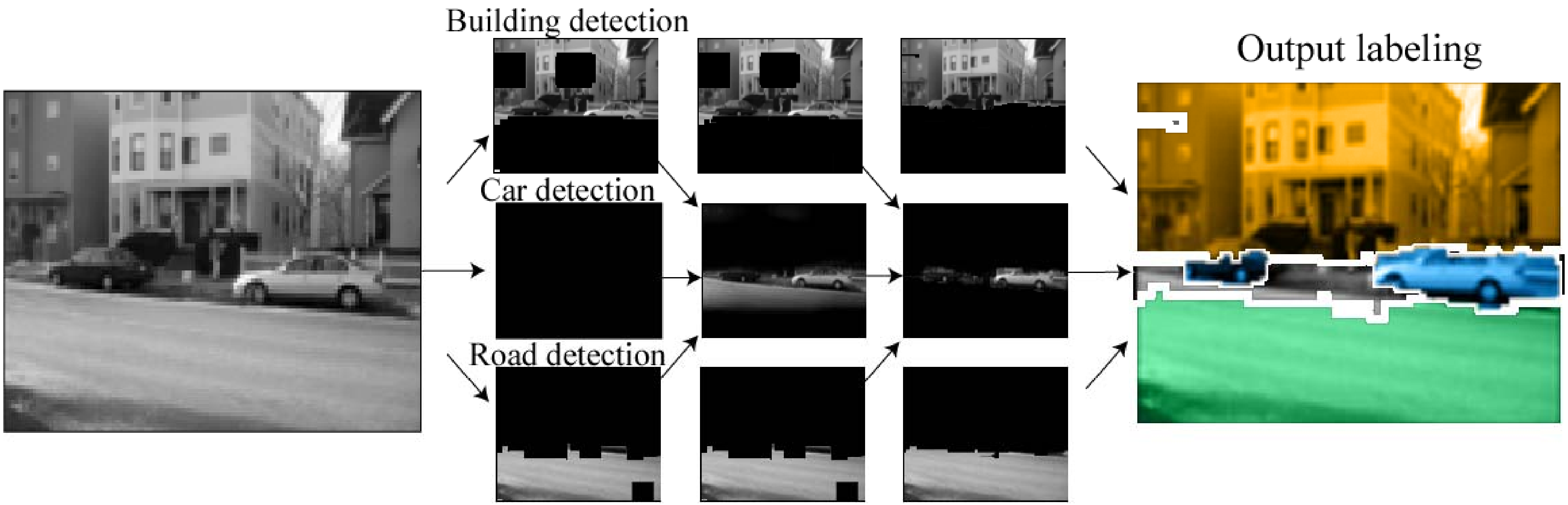
Viola & Jones (2001)

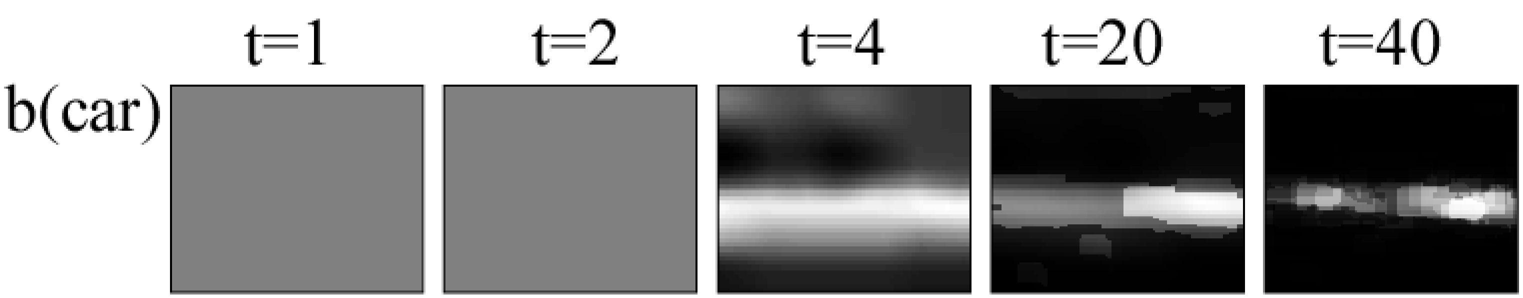
Set to zero the beliefs of nodes with low probability of containing the target.

Perform message passing only on undecided nodes

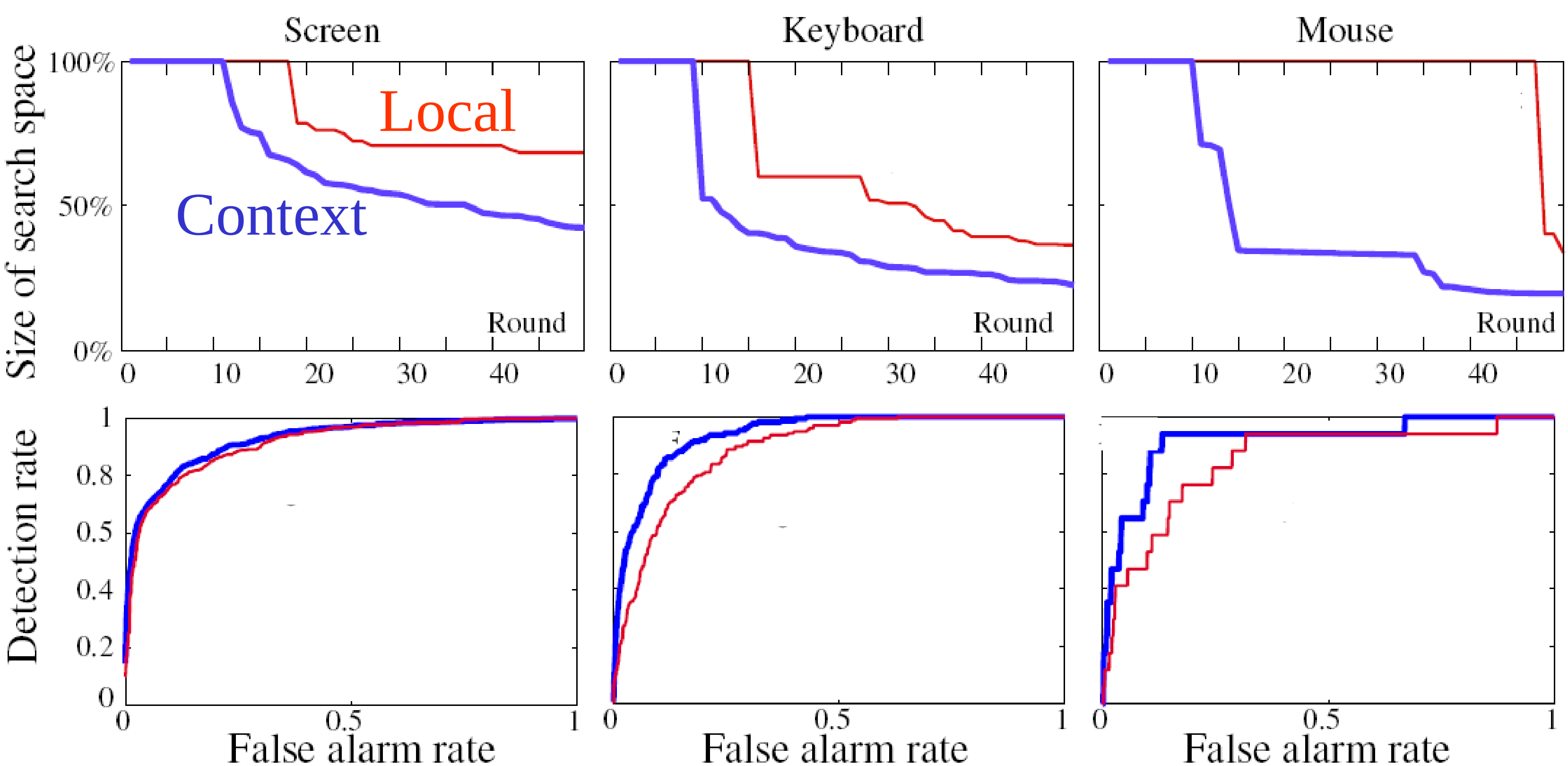
The detection of the screen reduces the search space for the mouse detector.

Cascade

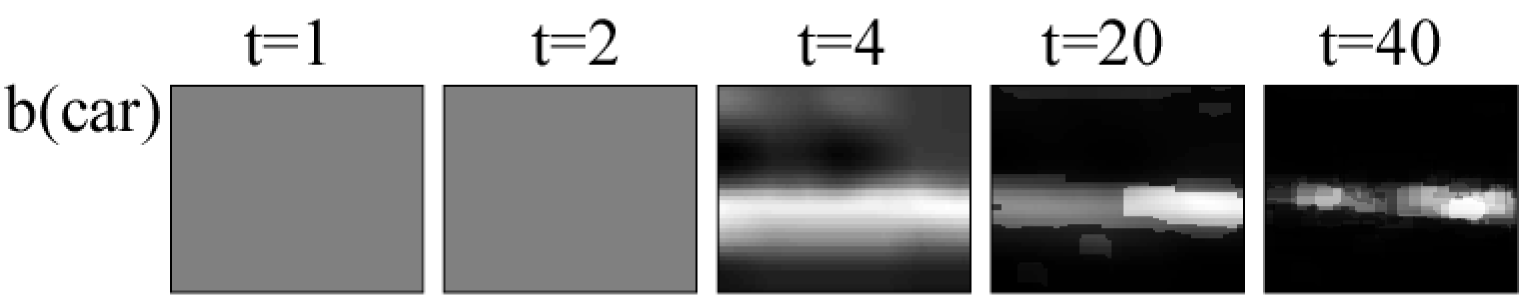

Building detection
Car detection
Road detection
Output labeling


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

Cascade



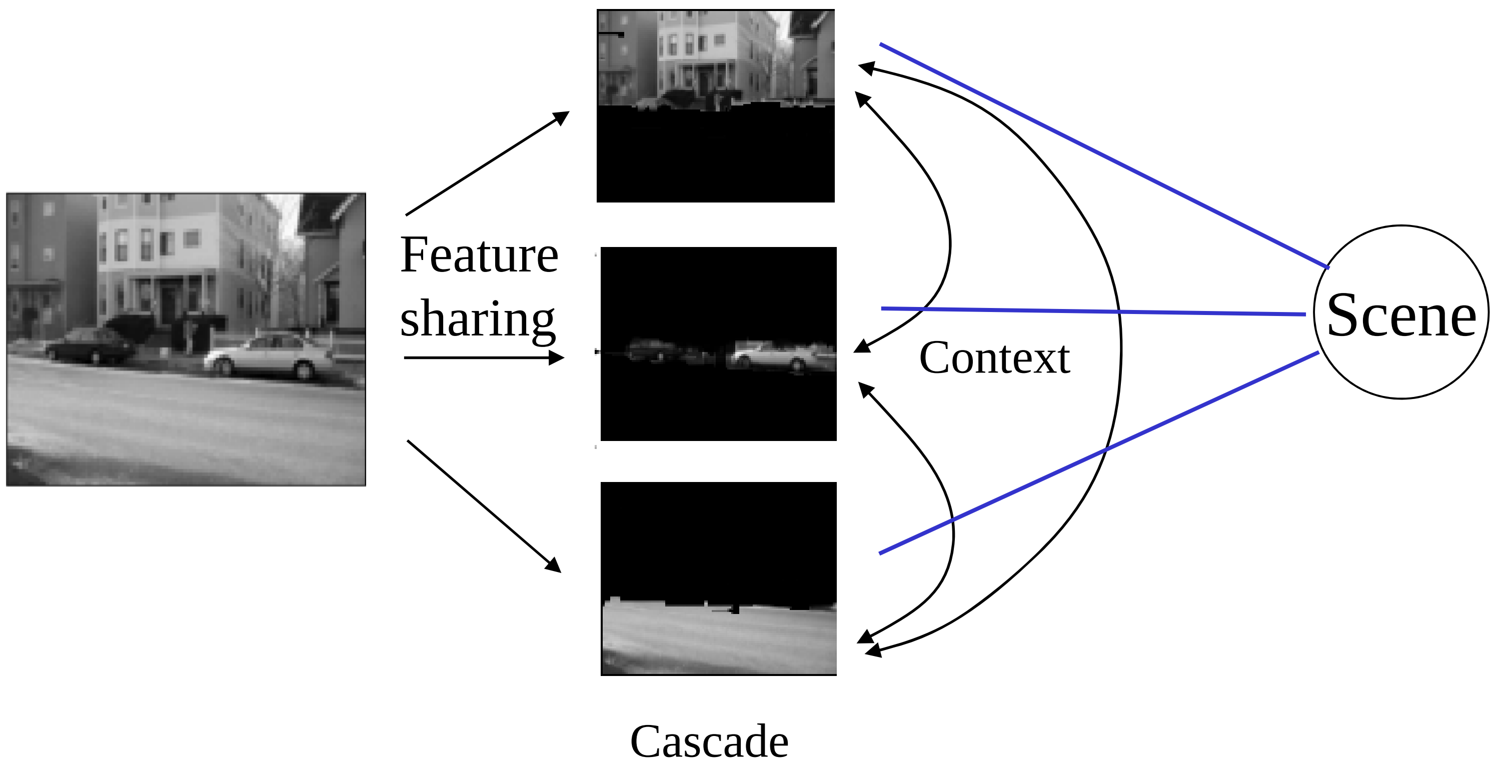
Screen
Keyboard
Mouse
Size of search space
100%
50%
0%
Local
Context
Round
0
10
20
30
40
Detection rate
1
0.8
0.5
0.4
0.2
0
0.5
False alarm rate



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

Future work

- Learn relationships between more objects (things get interesting beyond the 10 objects bar)
- Integrate segmentation and multiscale detection
- Add scenes/places



Today – **Image 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**. [Patrick Sundberg]**

- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 **[Robert Carroll]**

- L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in Computer Vision, 2007.

- G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. **[Brain Kazian]**

- S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. R. Bradski, P. Baumstarck, S. Chung, A. Y. Ng: Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video. IJCAI 2007

- Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008

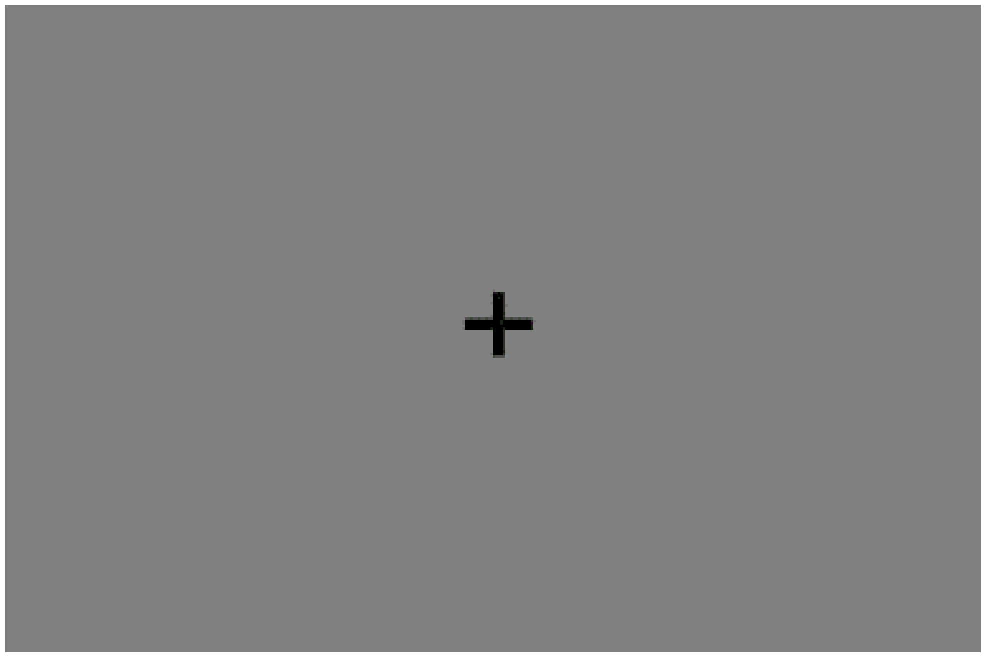
Today – **Image Context**

- 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. **[Patrick Sundberg]**
- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 **[Robert Carroll]**
- L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in Computer Vision, 2007.
- G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. **[Brain Kazian]**
- S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. R. Bradski, P. Baumstarck, S. Chung, A. Y. Ng: Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video. IJCAI 2007
- Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008

A picture tells us many stories

Fei-Fei et al. JoV 2007

A picture tells us many stories



Fei-Fei et al. JoV 2007

PT = 500ms

Some kind of game or fight. Two groups of two men? The foregound pair looked like one was getting a fist in the face. Outdoors seemed like because i have an impression of grass and maybe lines on the grass? That would be why I think perhaps a game, rough game though, more like rugby than football because they pairs weren't in pads and helmets, though I did get the impression of similar clothing. maybe some trees? in the background. (Subject: SM)

PT = 27ms

This was a picture with some dark sploches in it. Yeah. . .that's about it. (Subject: KM)

PT = 40ms

I think I saw two people on a field. (Subject: RW)

PT = 67ms

Outdoor scene. There were some kind of animals, maybe dogs or horses, in the middle of the picture. It looked like they were running in the middle of a grassy field. (Subject: IV)

PT = 107ms

two people, whose profile was toward me. looked like they were on a field of some sort and engaged in some sort of sport (their attire suggested soccer, but it looked like there was too much contact for that). (Subject: AI)

Fei-Fei et al. JoV 2007

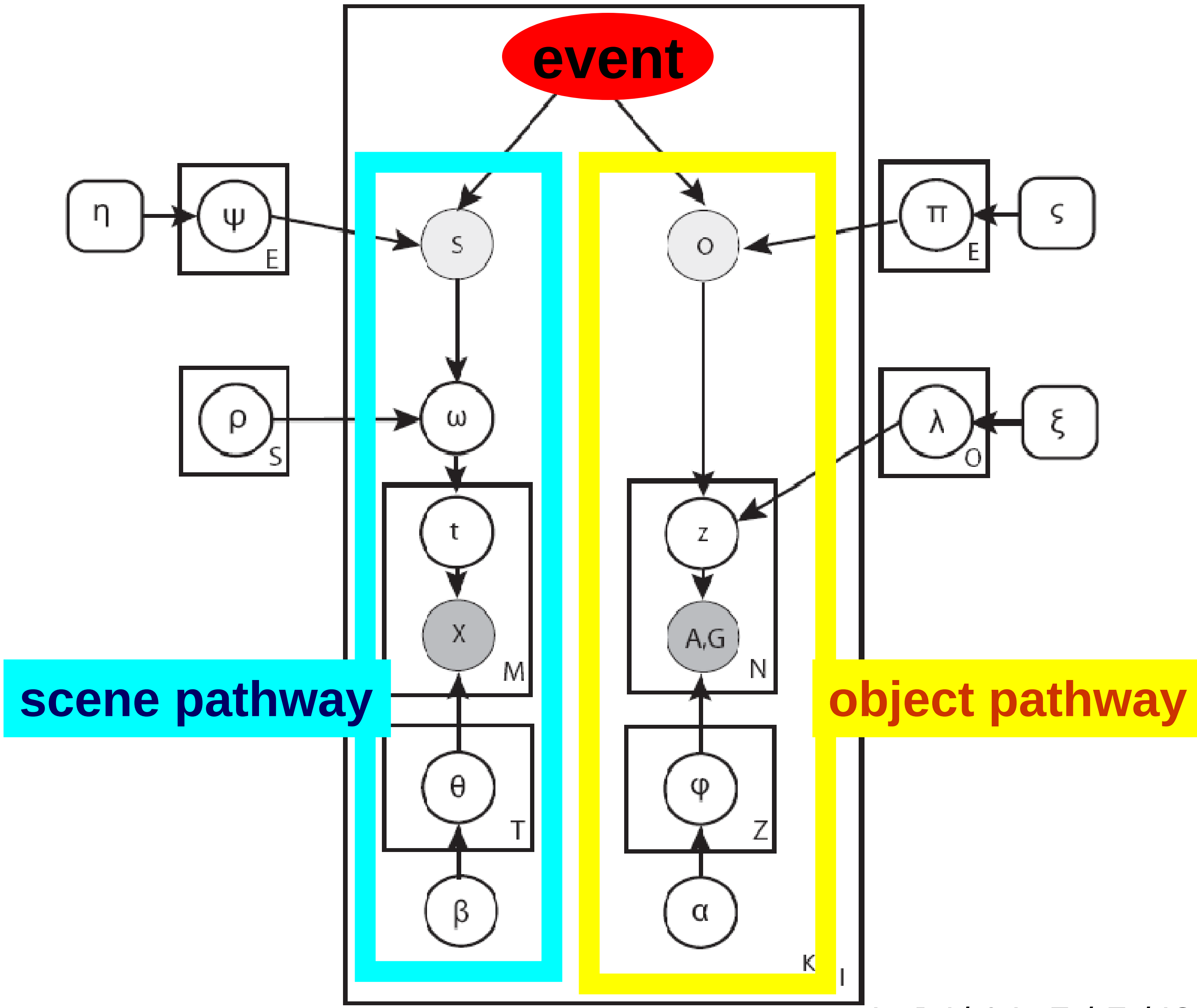
What, where and who? Classifying events by scene and object recognition

L-J Li & L. Fei-Fei, ICCV 2007

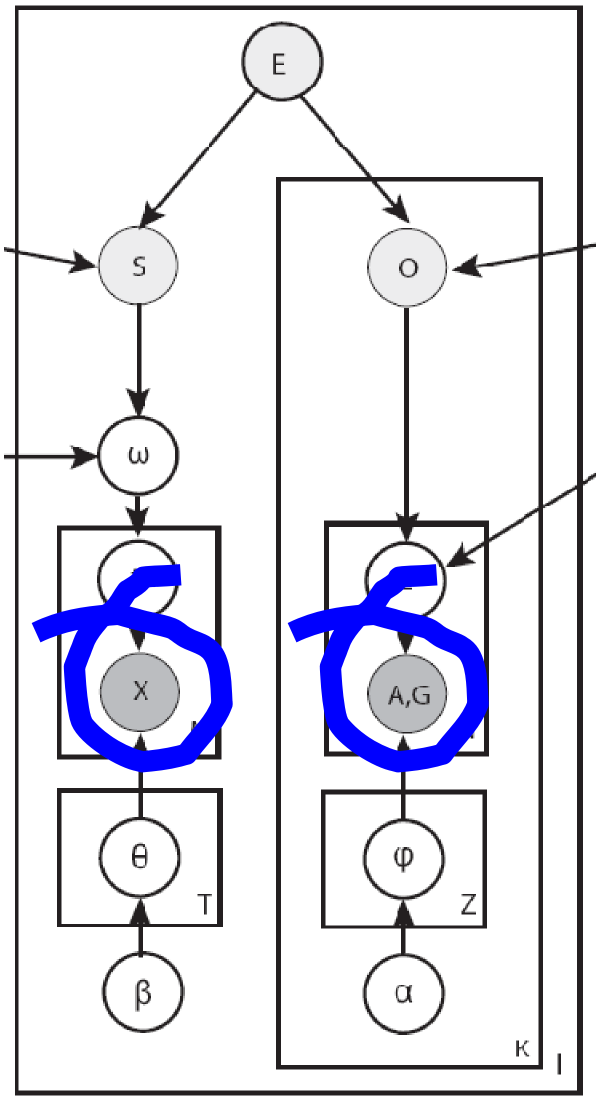
Related vision works toward holistic interpretation of images

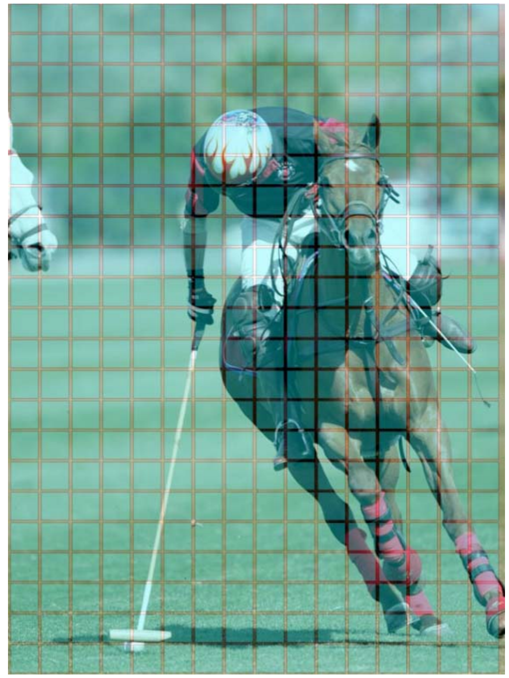
- Barnard, Duygulu, de Freitas, Forsyth, Blei & Jordan (‘Matching Pictures with Words’, 2003)
- Tu, Chen, Yuille & Zhu (‘Image Parsing’, 2003)
- Murphy, Torralba & Freeman (‘Seeing Forest and Trees’, 2003)
- Jin & Geman (‘Hierarchical Image Model’, 2006)
- Hoiem, Efros & Herbert (‘Objects in Perspective’, 2006)
- …


rowing
bocce
badminton
snow boarding
polo
croquet
sailing
rock climbing

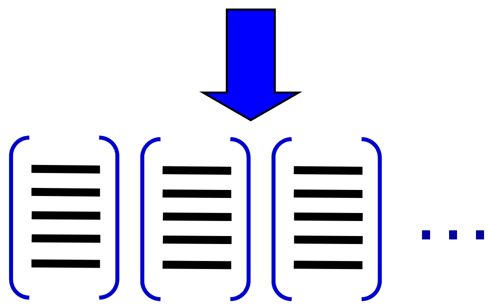


L.-J. Li & L. Fei-Fei ICCV 2007





X, A:
Appearance features



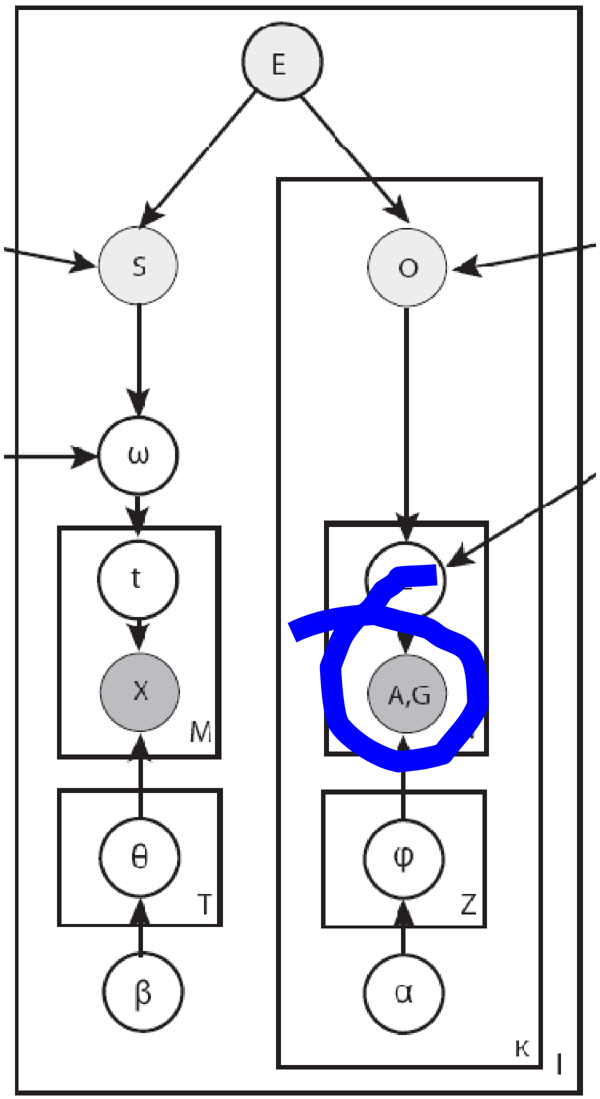
Compute SIFT descriptor

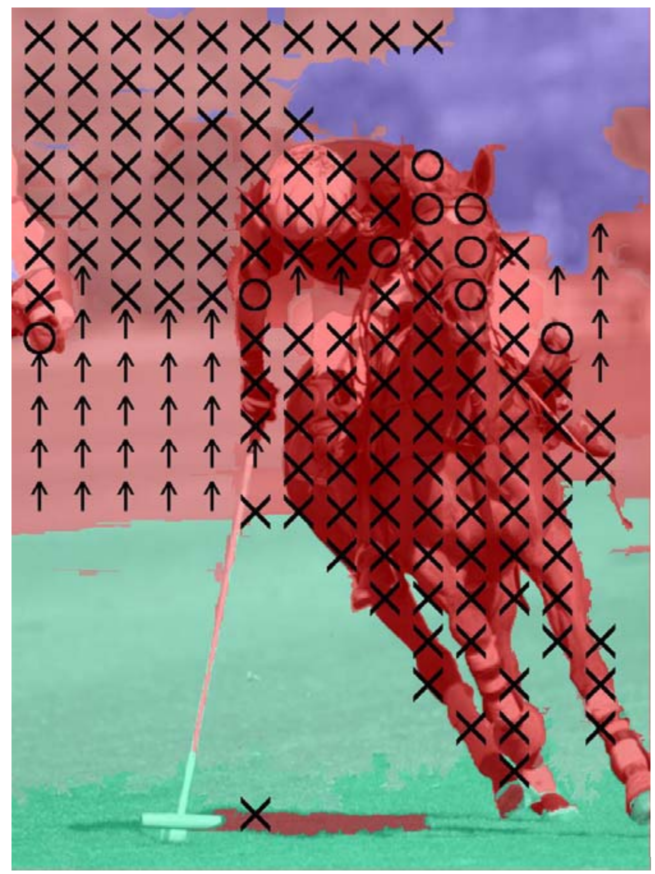
[Lowe'99]

Codewords representation

L.-J. Li & L. Fei-Fei ICCV 2007

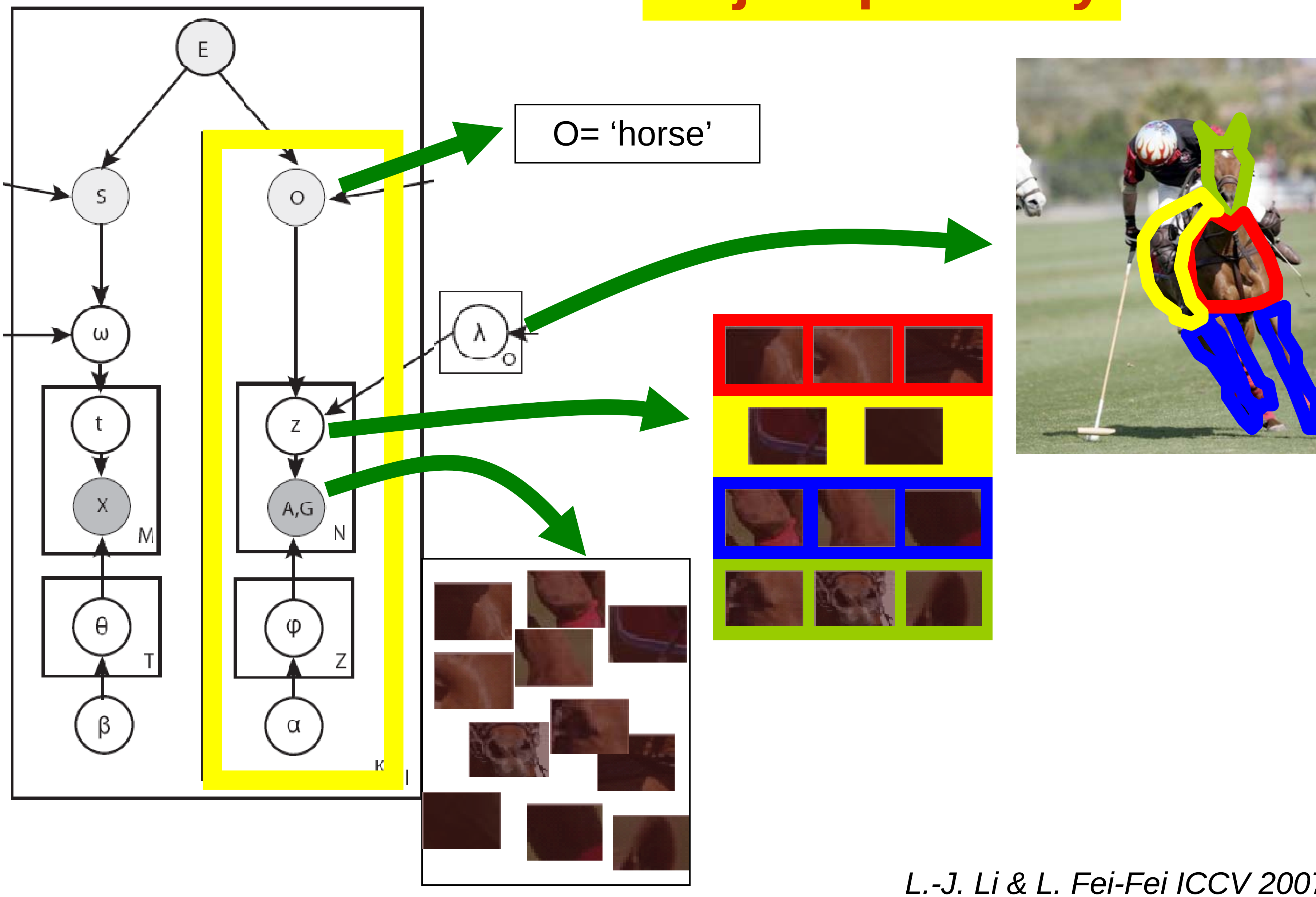
G:
Geometry features





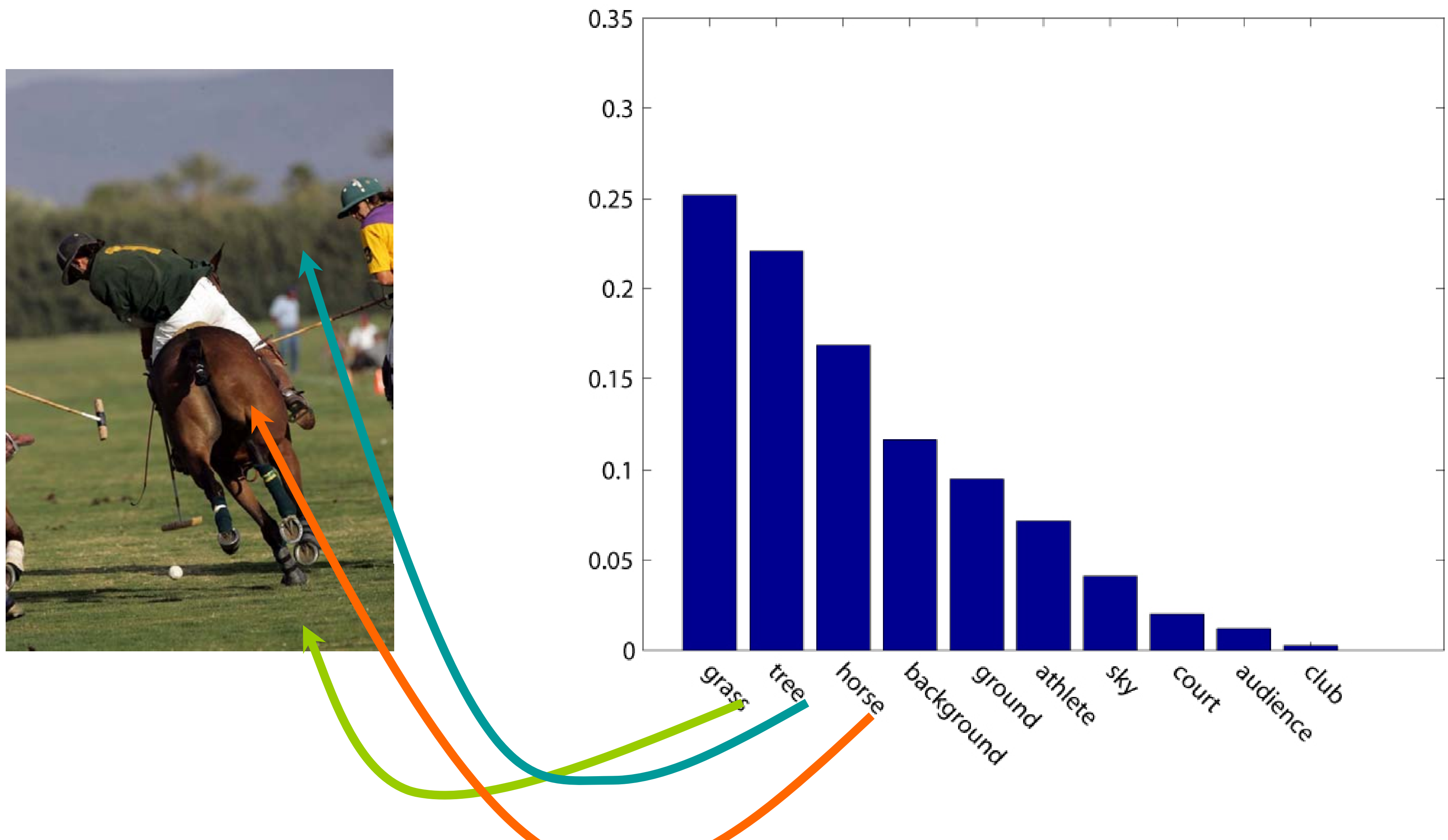
Hoiem et al. Siggraph 2006

object pathway



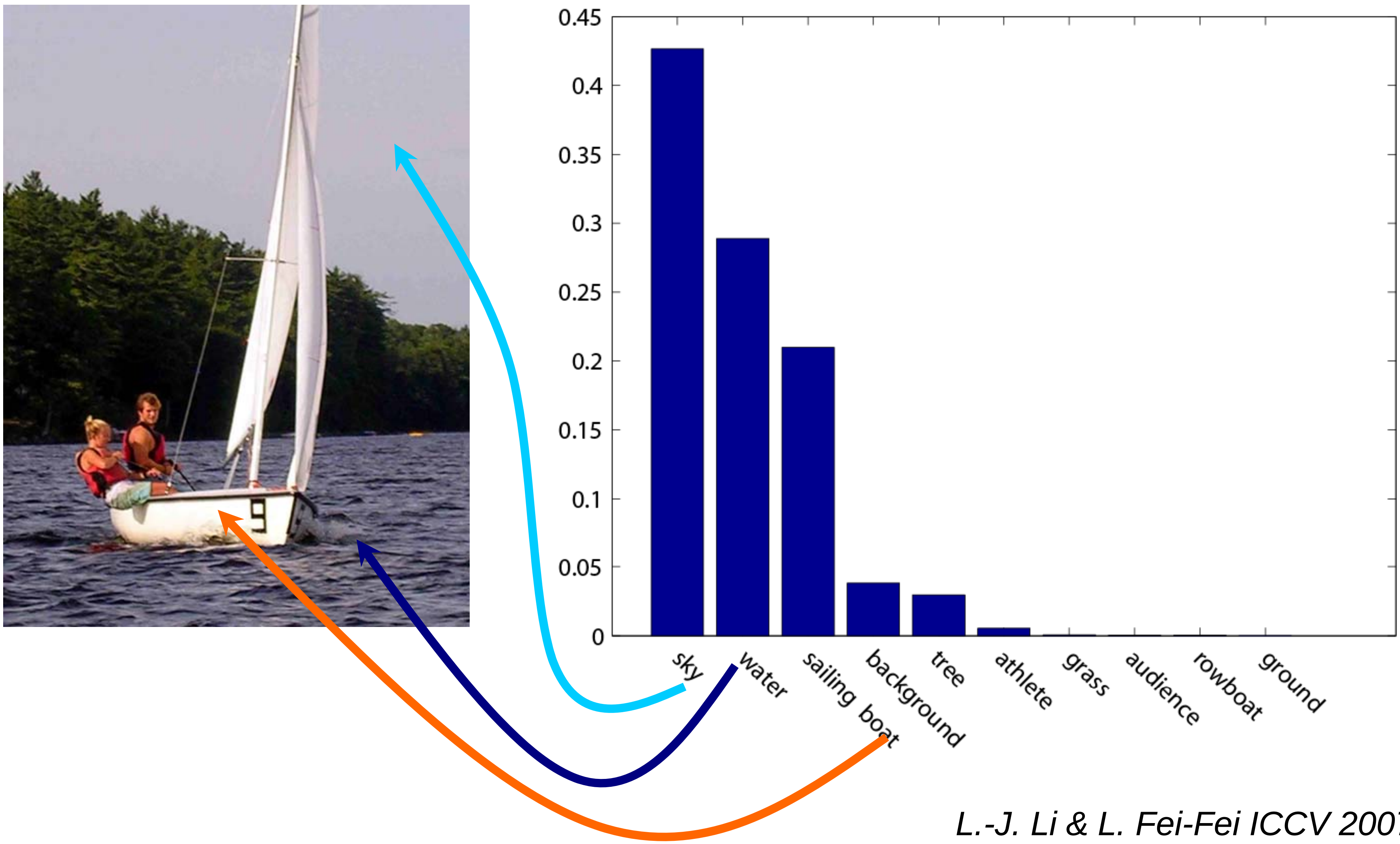
L.-J. Li & L. Fei-Fei ICCV 2007

Example: Polo



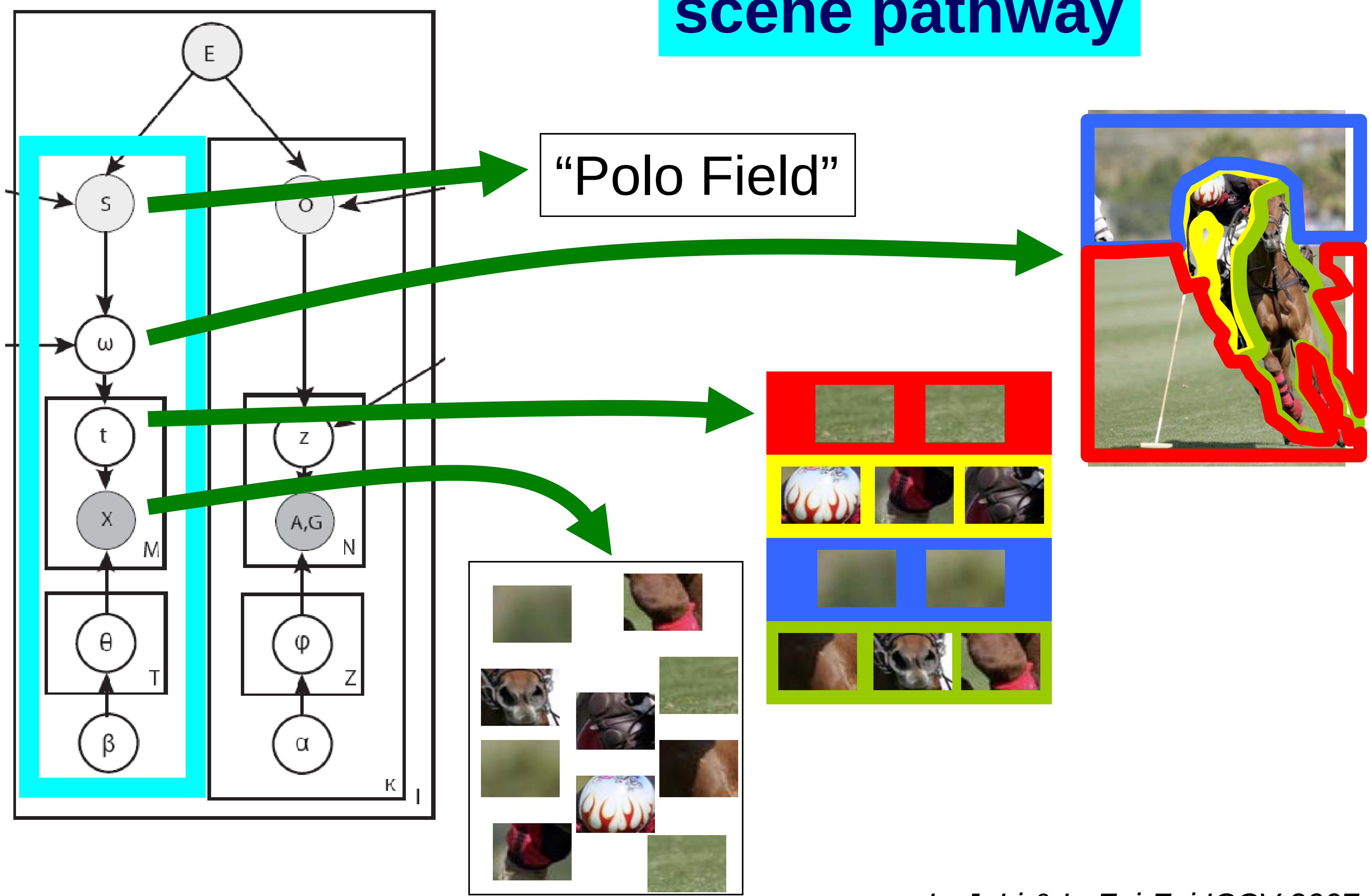
L.-J. Li & L. Fei-Fei ICCV 2007

Example: Sailing



L.-J. Li & L. Fei-Fei ICCV 2007

scene pathway



L.-J. Li & L. Fei-Fei ICCV 2007

Recognition in an Unknown Image

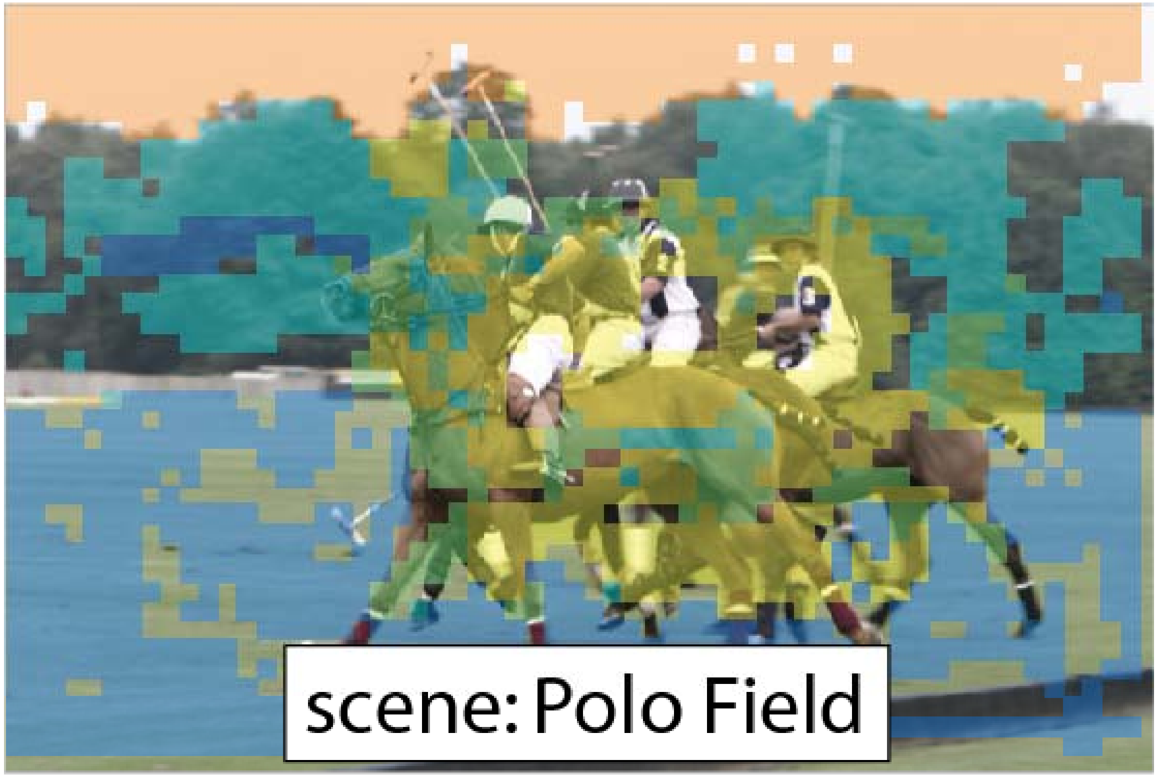
L.-J. Li & L. Fei-Fei ICCV 2007

Labeling objects in an unknown image -- objects (who)

$$p(A_n, G_n | O_k) = \sum_z P(A_n, G_n | z) P(z | O_k)$$

L.-J. Li & L. Fei-Fei ICCV 2007

Labeling objects in an unknown image -- scene (where)



$$p(I|S,\rho,\theta) = \int p(\omega|\rho,S)(\prod_{m=1}^{M}\sum_{t_m} p(t_m|\omega)\cdot p(X_m|t_m,\theta))d\omega$$

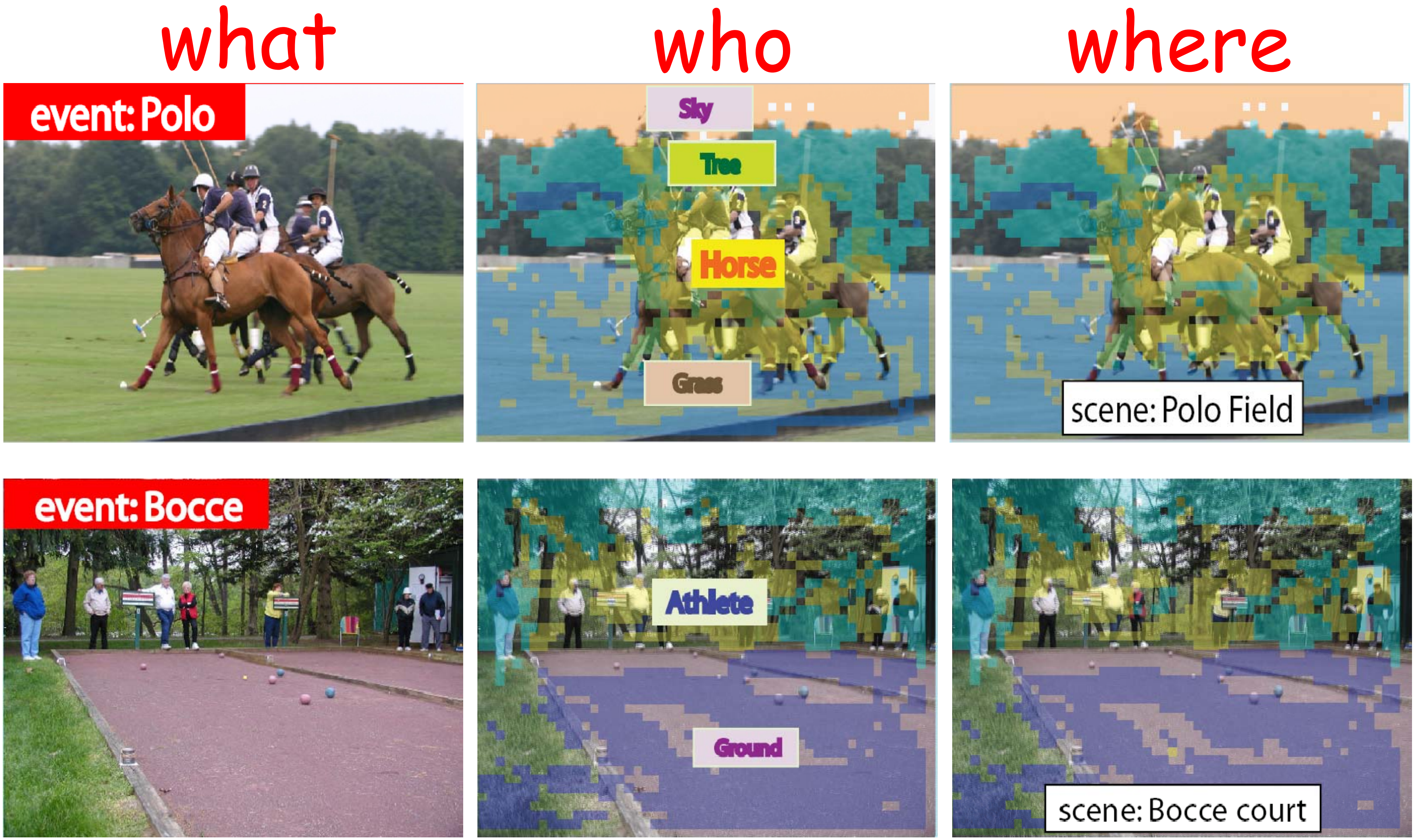
L.-J. Li & L. Fei-Fei ICCV 2007

Labeling objects in an unknown image -- events (what)

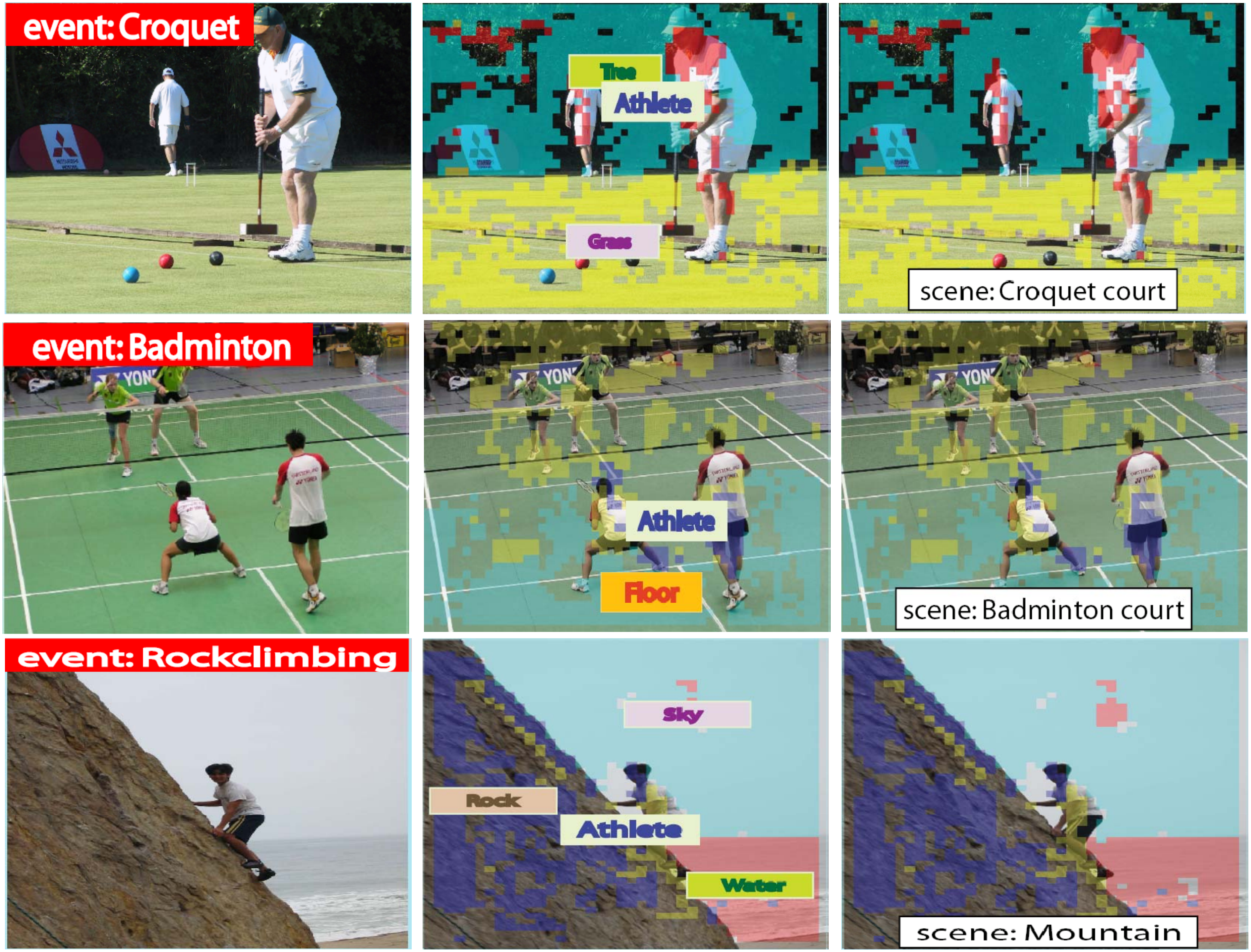
$$p(I|E) \propto P(I|S)P(S|E)\prod_{n=1}^{N}\sum_{O} P(A_n, G_n|O)P(O|E)$$

L.-J. Li & L. Fei-Fei ICCV 2007

The 3W stories

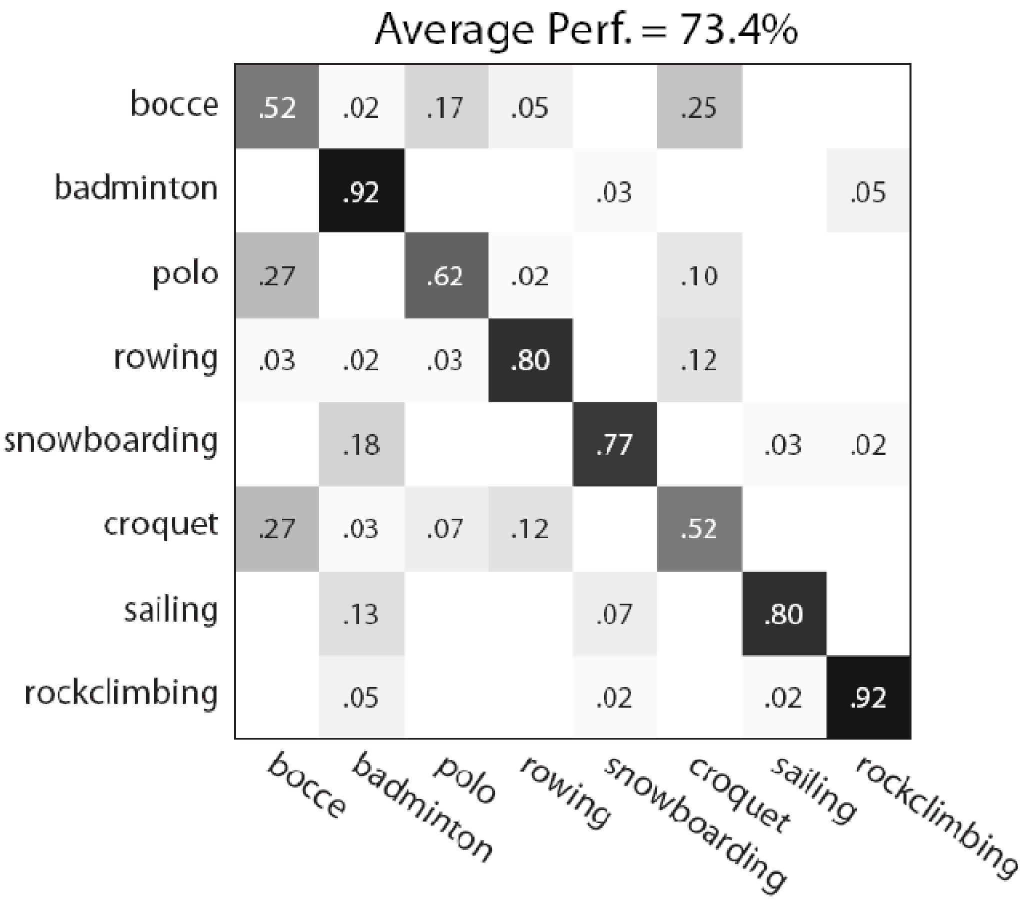


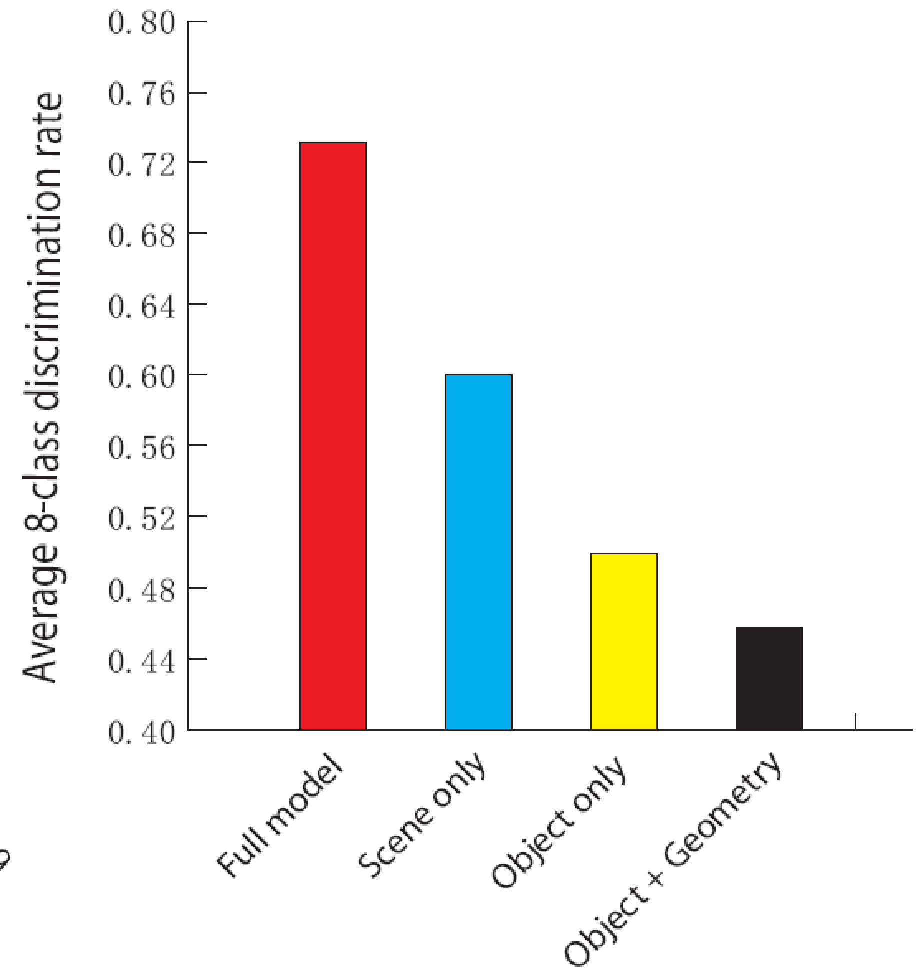
L.-J. Li & L. Fei-Fei ICCV 2007


event: Croquet
Tree
Athlete
Grass
scene: Croquet court
event: Badminton
Athlete
Floor
scene: Badminton court
event: Rockclimbing
Sky
Rock
Athlete
Water
scene: Mountain


event: Rowing
Tree
Athlete
Rowing boat
Water
scene: Lake
event: Sailing
Sky
Sailing boat
Water
scene: Lake
event: Snowboarding
Sky
Athlete
Snowfield
scene: Snow mountain

Quantitative result





L.-J. Li & L. Fei-Fei ICCV 2007

Today – **Image 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**. [Patrick Sundberg]**
- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 **[Robert Carroll]**
- L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in Computer Vision, 2007.
- G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. **[Brain Kazian]**
- S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. R. Bradski, P. Baumstarck, S. Chung, A. Y. Ng: Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video. IJCAI 2007
- Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008

Today – **Image 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**. [Patrick Sundberg]**
- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 **[Robert Carroll]**
- L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in Computer Vision, 2007.
- G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. **[Brain Kazian]**
- S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. R. Bradski, P. Baumstarck, S. Chung, A. Y. Ng: Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video. IJCAI 2007
- Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008

Peripheral-Foveal Vision for Real-time Object Recognition

Stephen Gould, Benjamin Sapp, Morgan Quigley, Andrew Y. Ng

http://ai.stanford.edu/~sgould/vision/nips06poster.pdf

1

Overview

• Human object recognition in a 3d environment is far superior to that of any robotic vision system.

• One reason (out of many) for this is that humans use a **fovea** to fixate on, or near an object, thus obtaining a very high resolution image of the object and rendering it easy to recognize.

• We present a novel method for identifying and tracking objects using a two camera system.

• **Our method** is motivated by biological vision systems:

  • uses a learned **"attentive" interest map** on a low resolution view to direct a high resolution **"fovea."**

  • objects that are recognized in the **fovea** can then be tracked using **peripheral vision**.

http://ai.stanford.edu/~sgould/vision/nips06poster.pdf

2

STAIR robot

http://ai.stanford.edu/~sgould/vision/nips06poster.pdf

Visual attention model

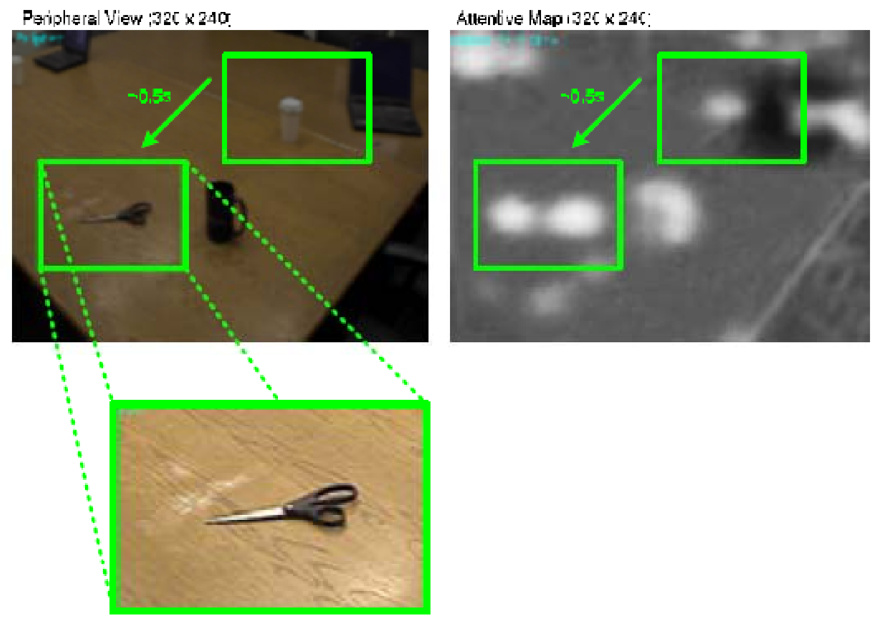
• Our system uses two separate cameras:

  • fixed **wide-angle camera** for **peripheral vision**

  • controllable **pan-tilt-zoom** (PTZ) camera for **foveal vision**.

• The PTZ camera can focus on any region of the scene to obtain a high-resolution image for object recognition.

• Previously identified objects are tracked using peripheral vision.



http://ai.stanford.edu/~sgould/vision/nips06poster.pdf

- The attention system periodically decides between the following actions:

  - **Confirmation** of a tracked object by fixating the fovea over the predicted location of the object;

  - **Search** for unidentified objects by moving the fovea to some new part of the scene.

- Estimating the reduction in entropy, *H*, by taking each action (**Confirmation** or **Search**), we take the action which maximizes the expected reduction in entropy.

http://ai.stanford.edu/~sgould/vision/nips06poster.pdf

4

Interest modeling

• Our **interest model** allows us to choose which foveal region to examine next by rapidly identifying pixels which have a high probability of containing objects that we can classify.

• We define a pixel to be interesting if it is **part of an unknown, yet classifiable object**.

  • a consequence of this definition is that our model automatically encodes the biological phenomena of **saliency** and **inhibition of return** [Itti and Koch, 2001].

• *Interestingness* of every pixel in the peripheral view is modeled using a dynamic Bayesian network (DBN) whose parameters are learned from training videos.

http://ai.stanford.edu/~sgould/vision/nips06poster.pdf

5

Object recognition and tracking

• A **Kalman filter** tracks the location and velocity of identified objects in the 2d image plane.

• Use subset of biologically inspired **C1 features** [Serre *et. al*, 2004] and learn a boosted decision tree classifier for each object.

6

Experimental results

- Our method compared to three naive approaches:
  (i) fixing the foveal gaze to the center of view,
  (ii) linearly scanning over the scene from top-left to bottom-right, and,
  (iii) randomly moving the fovea around the scene.

Fovea control	**Recall**	**Precision**	**F_1-Score**
Fixed at center	9.49%	97.4%	17.3%
Linear scanning	13.6%	100.0%	24.0%
Random scanning	27.7%	84.1%	41.6%
Our method	62.2%	83.9%	**71.5%**

- Videos demonstrating our results are available at **http://ai.stanford.edu/~sgould/vision/**

http://ai.stanford.edu/~sgould/vision/nips06poster.pdf

Today – **Image 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**. [Patrick Sundberg]**
- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 **[Robert Carroll]**
- L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in Computer Vision, 2007.
- G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. **[Brain Kazian]**
- S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. R. Bradski, P. Baumstarck, S. Chung, A. Y. Ng: Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video. IJCAI 2007
- Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008

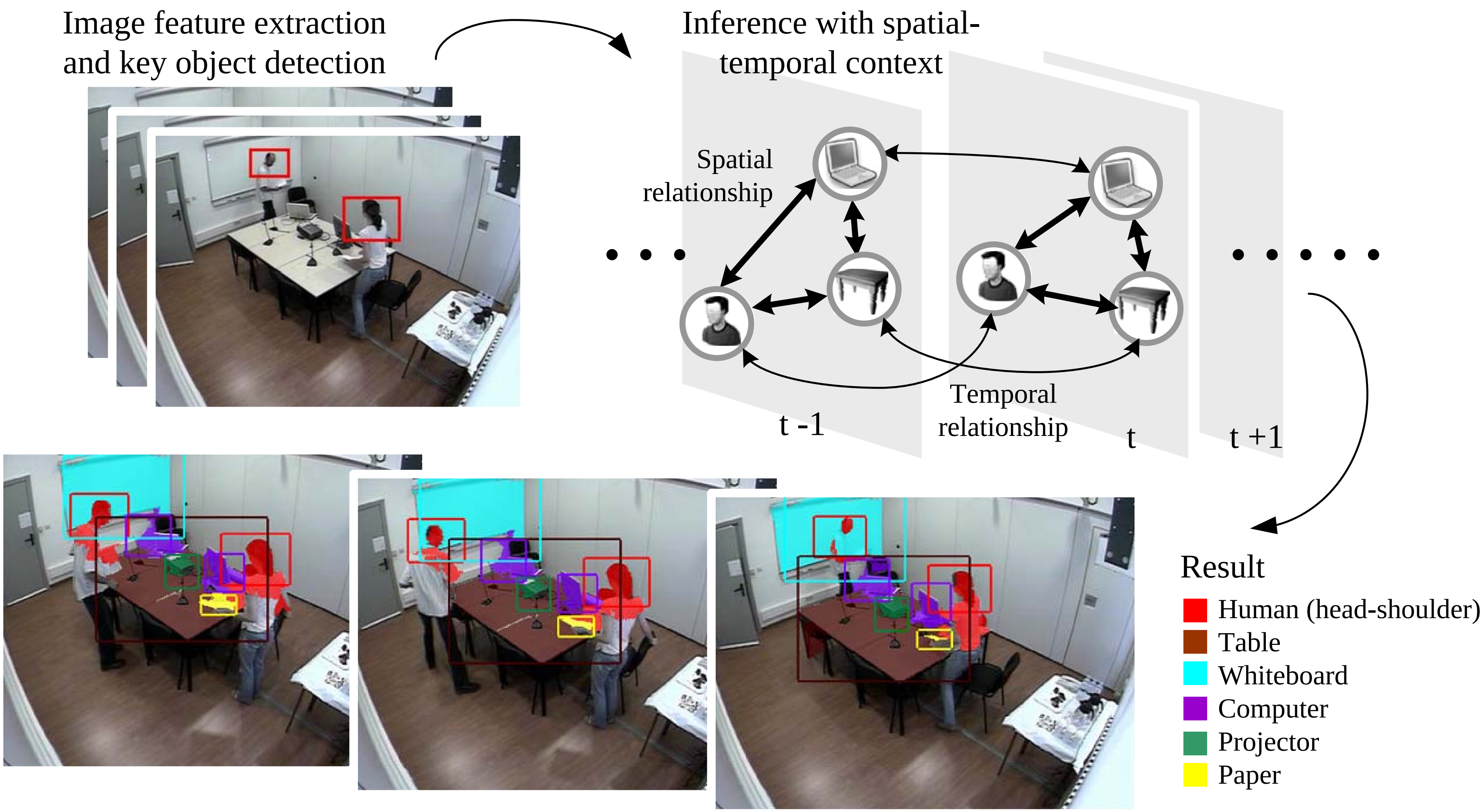
Key Object Driven Multi-Category Object Recognition, Localization and Tracking Using Spatio-Temporal Context

Yuan Li and Ram Nevatia
Institute for Robotic and Intelligent Systems
University of Southern California

The Idea

- **Objective**: recognizing and tracking multiple categories of objects that are involved in interaction with humans in video.
- **Motivation**: the significance of recognizing objects involved in interaction with humans lies not only in the static image domain but also in video understanding and analysis.
- **Difficulty**: varying appearance of objects, lack of image detail, multi-object co-occurrence, motion make it difficult for purely appearance-based approach.
- **Approach**: incorporate spatial and temporal contextual information to aid object recognition and localization.

Approach Overview



Approach Overview

- **Spatial relationships** between different object categories are utilized so that co-inference enhances accuracy;
- **Temporal context** is utilized to accumulate object evidence and to track objects continuously;
- **Robust key object detection**: boosting + edgelet features; use key objects to reduce inference space for other objects.
- **Observation model for other objects**: interest point detection + SIFT + image region features.
- **Modeled by a dynamic MRF**:
  - Node: object state and observation in one frame;
  - Intra-frame edge: spatial relationship;
  - Inter-frame edge: temporal relationship;
  - Inference done by nonparametric belief propagation.

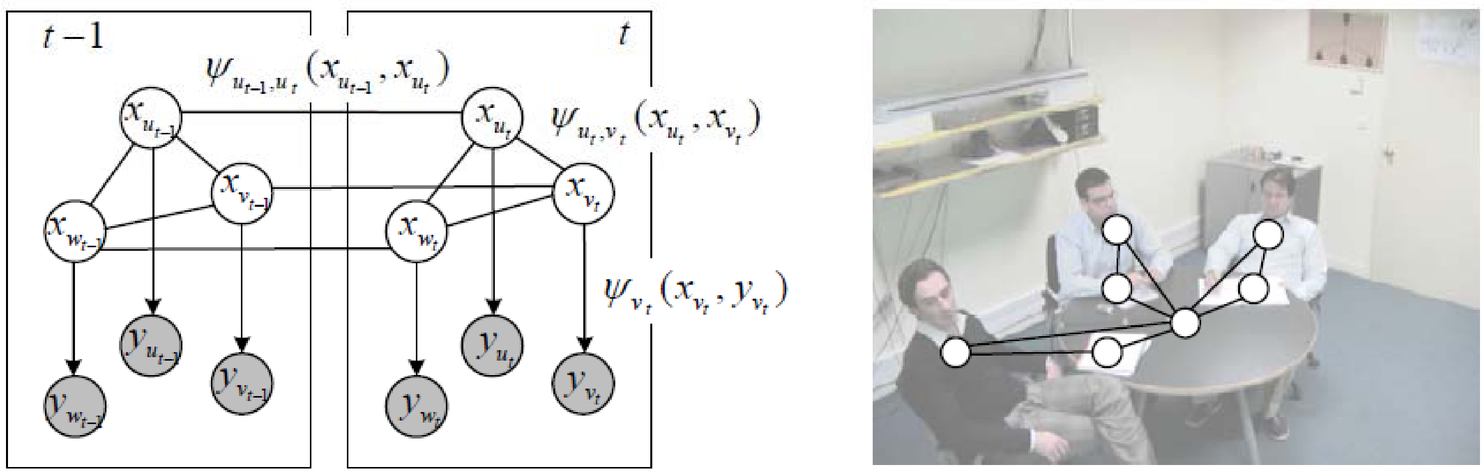


Fig. 2. the MRF defined in our problem (left) and an ideal graph structure for one input frame (right). Section 5 explains how to build such a graph.

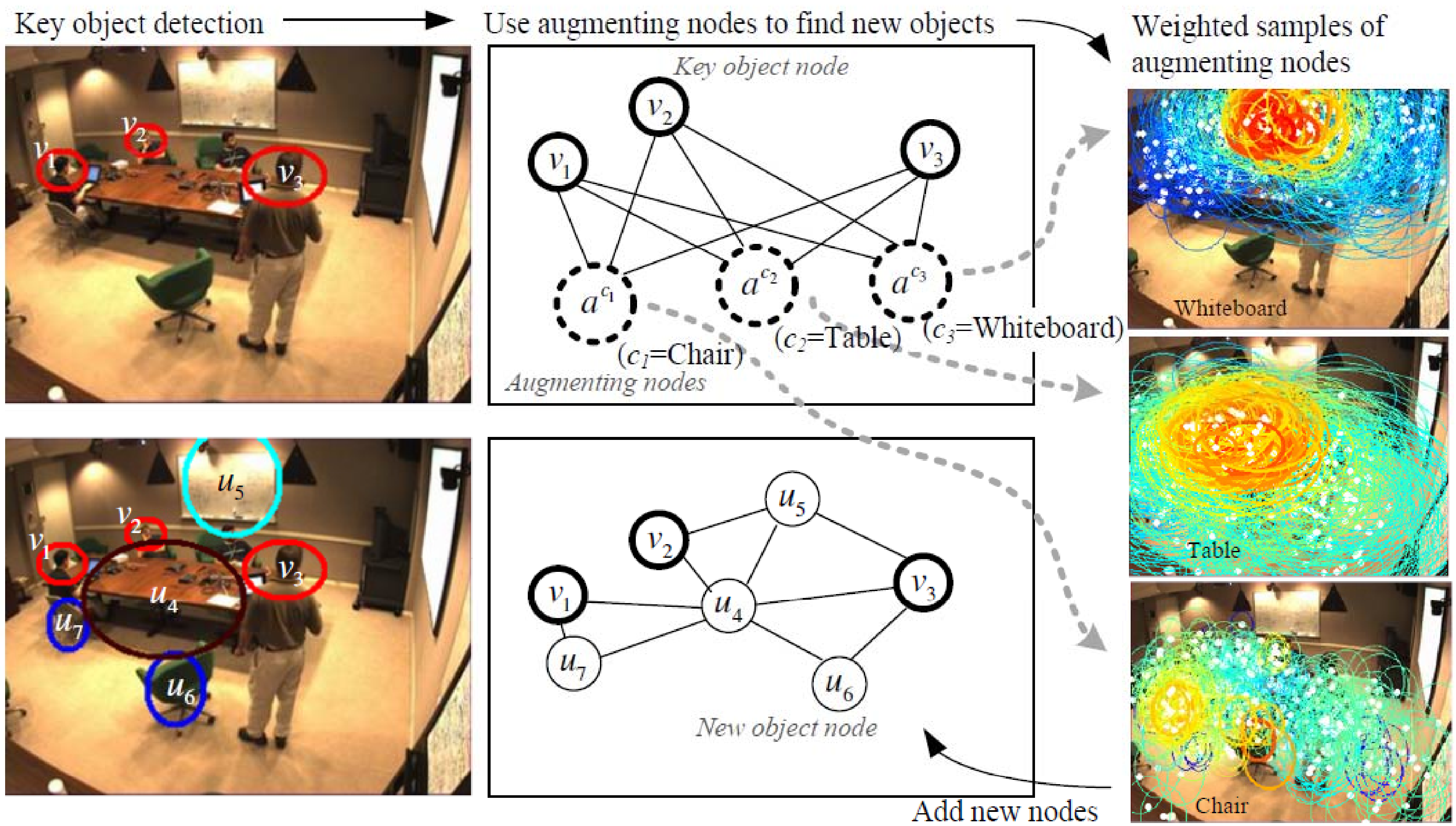
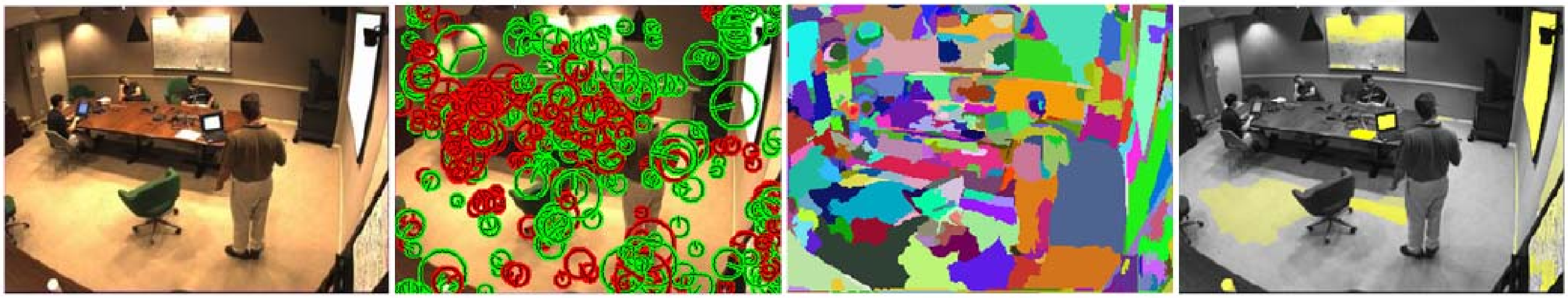


Fig. 4. Use of augmenting nodes to update graph structure. Augmenting nodes for each category are shown as one (dotted circle). For weighted samples, red indicates the highest possible weight, while blue indicates the lowest.

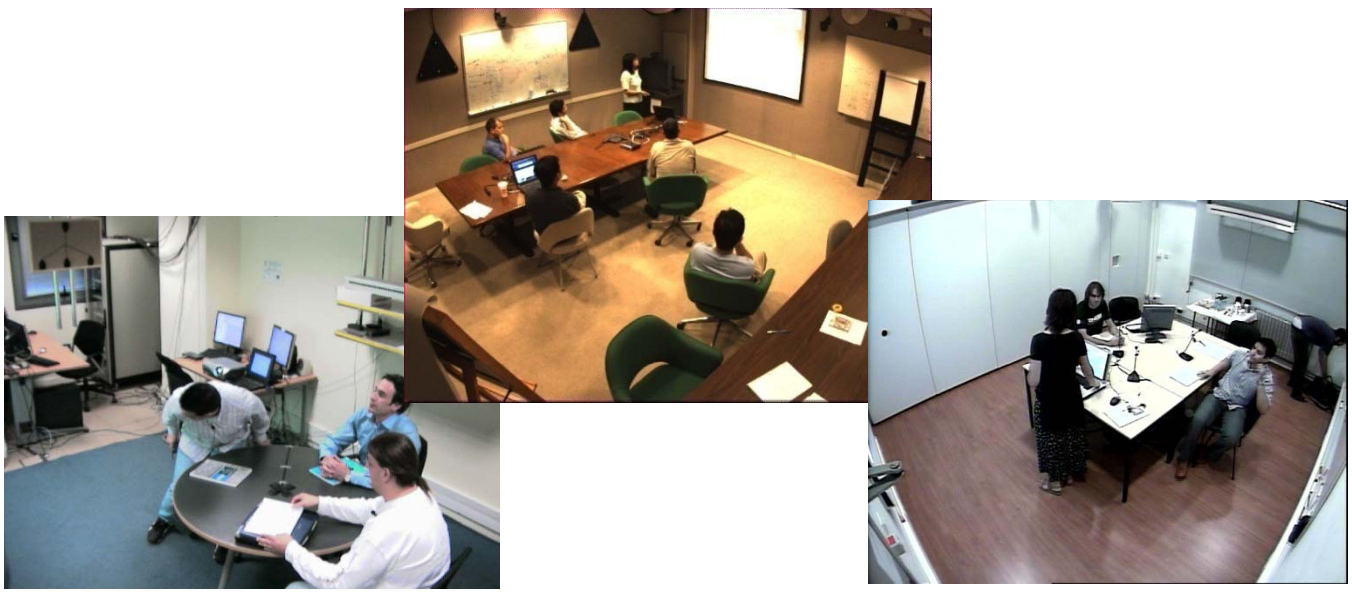


(a) (b) (c) (d)

Fig. 3. An example of finding paper based on appearance. (a) Input image; (b) SIFT features (green: feature with positive weight in the classifier, red: feature with negative weight); (c) Segmentation; (d) observation likelihood $p(paper|r_i)$ for each region r_i (yellow: high likelihood).

Experiment setting

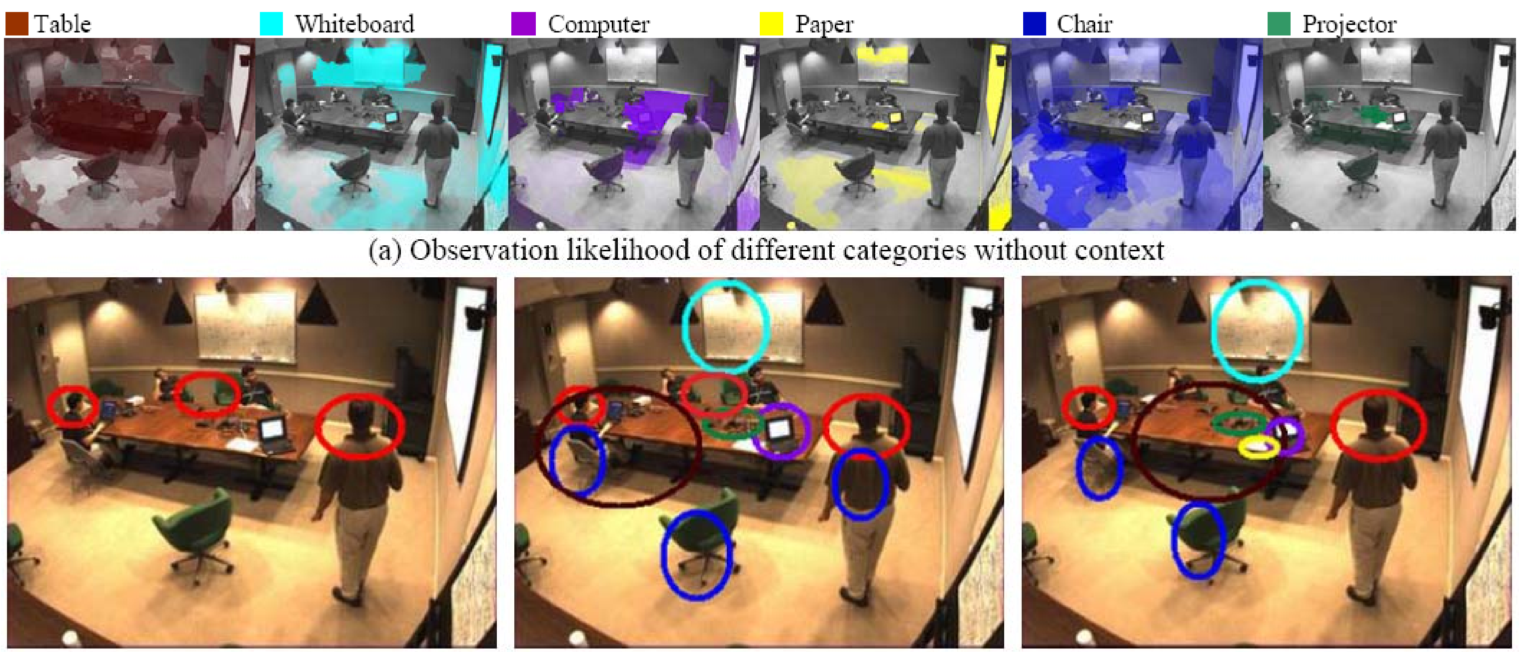
- CHIL meeting video corpus
  – Test on 16 videos from 3 sites (IBM, AIT and UPC), 3 camera views for each site, 400 frames / seq.



……

Experiment setting

- Compare three methods with different levels of context:
  - **No context**, *i.e.* object observation model is directly applied to each frame;
  - **Spatial context only**, *i.e.* a MRF without the temporal edges is applied in a frame-by-frame manner;
  - **Spatio-temporal context**, *i.e.* the full model with both spatial and temporal edges applied to the sequence.



(a) Observation likelihood of different categories without context

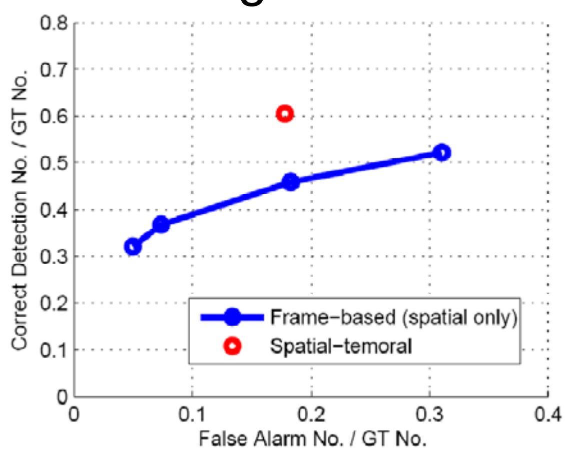
(b) Key object detection

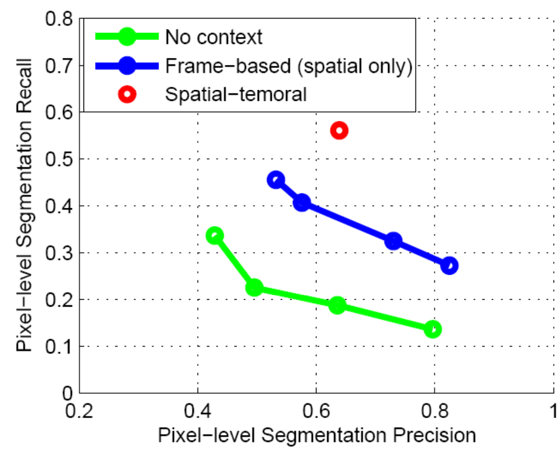
(c) Frame-based inference (spatial-only)

(d) Inference using the complete model (spatio-temporal)

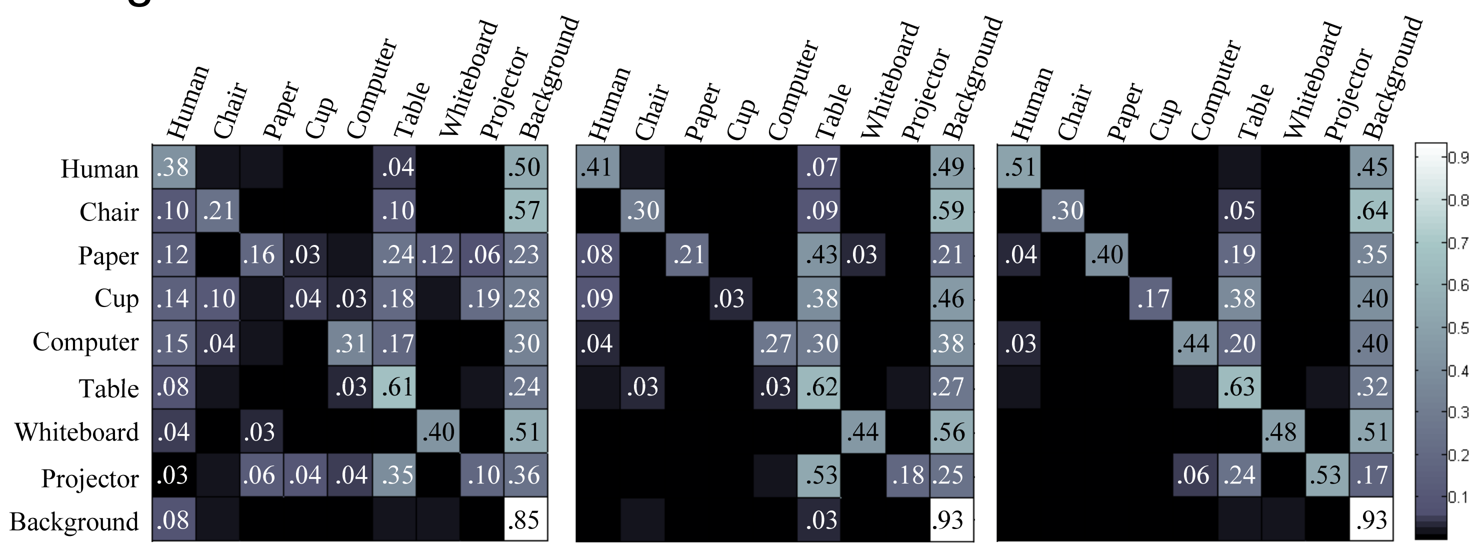
Results

- Detection and segmentation



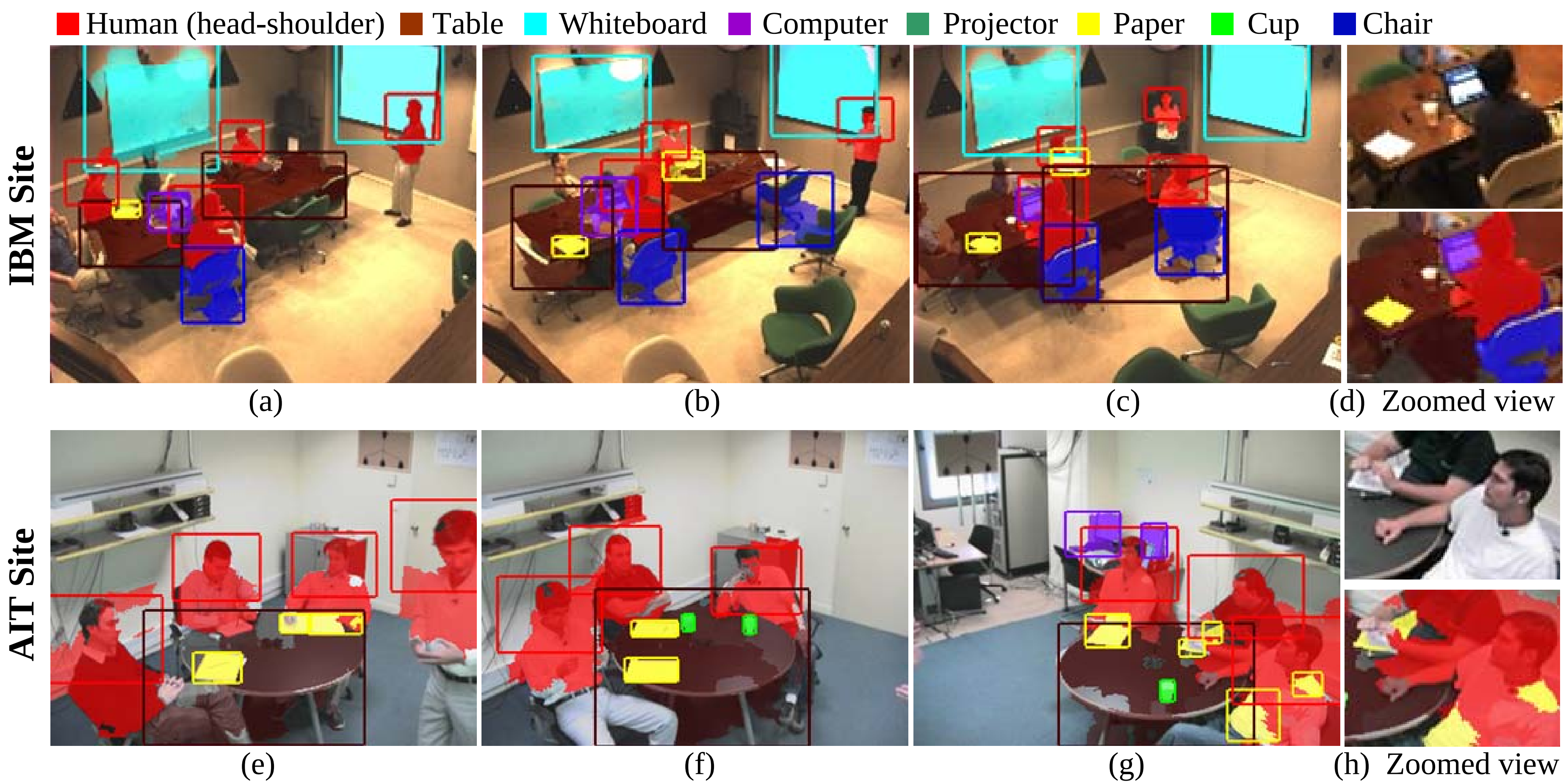


- Recognition confusion matrix



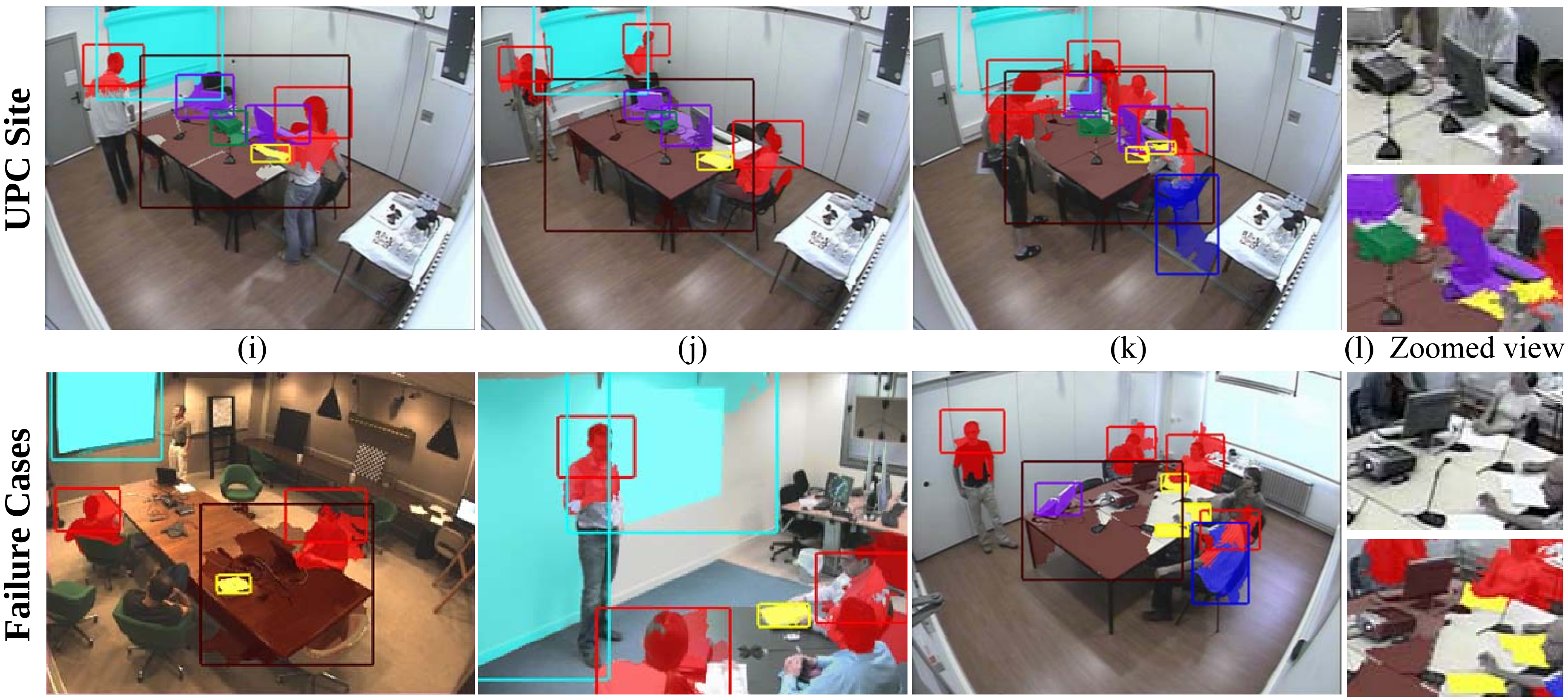
(a) Without context (b) Spatial-only (c) Spatio-temporal

Results



Human (head-shoulder)
Table
Whiteboard
Computer
Projector
Paper
Cup
Chair
IBM Site
AIT Site
(a)
(b)
(c)
(d) Zoomed view
(e)
(f)
(g)
(h) Zoomed view

Results



(i) (j) (k) (l) Zoomed view

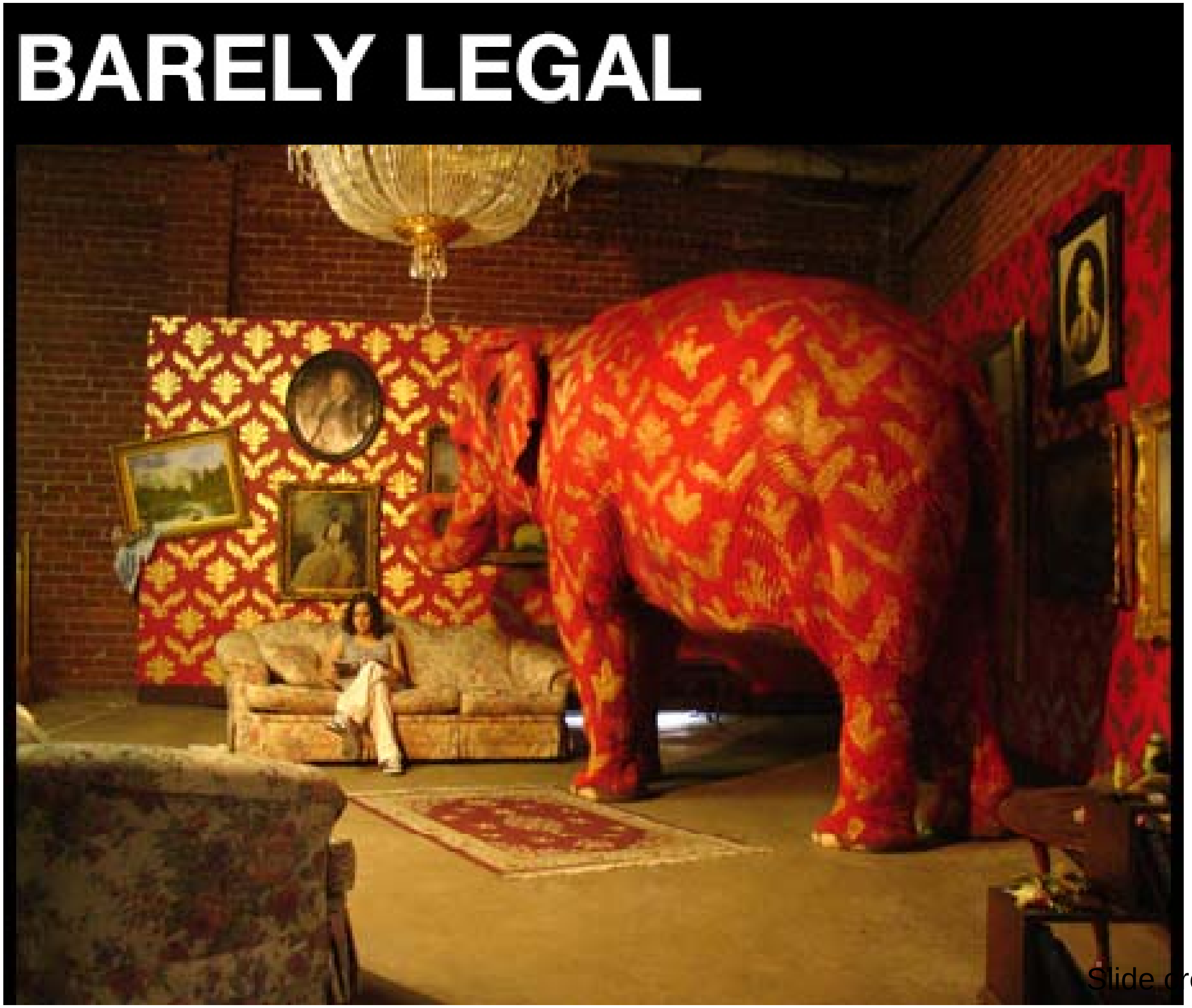
(m) IBM (n) AIT (o) UPC (p) Zoomed view

Today – **Image 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
- L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in Computer Vision, 2007.
- G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43.
- S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. R. Bradski, P. Baumstarck, S. Chung, A. Y. Ng: Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video. IJCAI 2007
- Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008

Who needs context anyway?

We can recognize objects even out of context



Banksy

Slide credit: A. Torralba

Next Class – **Shared Structures (Features, Parts)**

- A. Quattoni, M. Collins, and T. Darrell, "Transfer learning for image classification with sparse prototype representations," in IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008., pp. 1-8.
- A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, no. 5, pp. 854-869, 2007.
- S. Fidler and A. Leonardis, "Towards scalable representations of object categories: Learning a hierarchy of parts," in Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, 2007, pp. 1-8.
- T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Object recognition with cortex-like mechanisms. PAMI, 29(3):411–426, 2007.