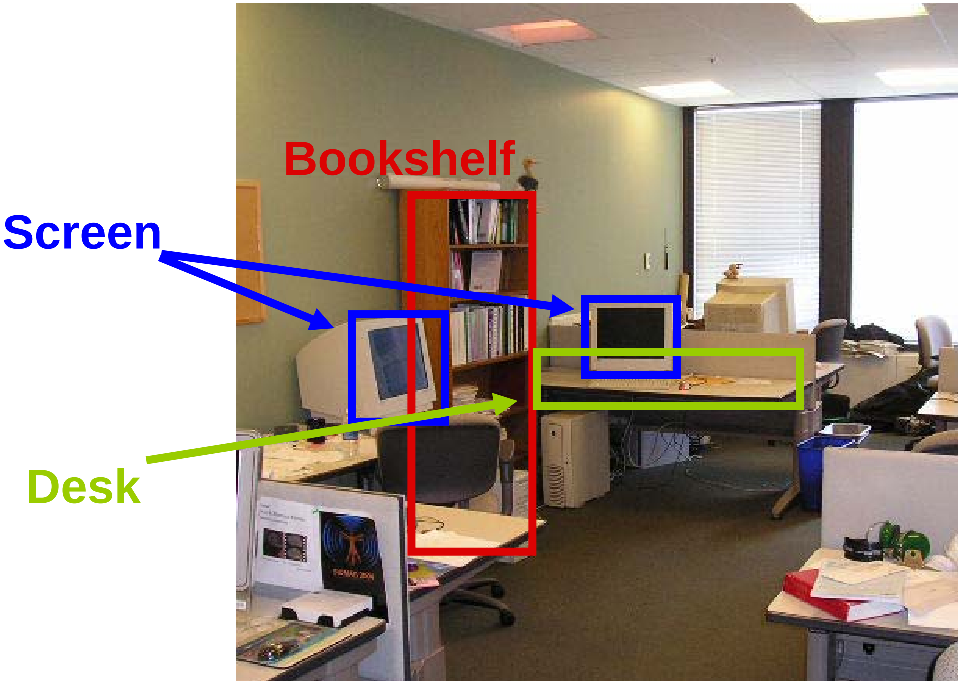
Context in vision

Antonio Torralba

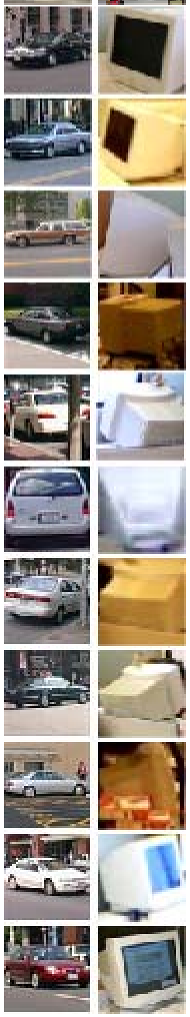
The goal

Office scene



Why object detection is a hard problem

Object classes →



viewpoints ↓

Styles, lighting conditions, etc, etc, etc…

Need to detect Nclasses * Nviews * Nstyles, in clutter.
Lots of variability within classes, and across viewpoints.

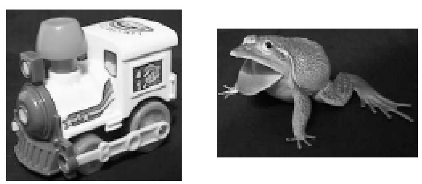
Where is the field of computer vision?

There are efficient solutions for

- Detecting few single object categories:

- Detecting particular objects:

Lowe, 1999



- Recognizing objects in isolation

From Leibe & Schiele, 2003

But the problem of multi-class and multi-view object detection in a scene with clutter is still largely unsolved.

The ingredients

- Object representations
- Scene representations

- Classifiers
- Graphical models

- Object features
- Scene features

OBJECTS

Object representations

Models

- Constellations of parts
- Holistic representations
  - Shape-appearance models
- Shapes, silohuetes
- 3D models

Object representations

Features

- Pixel intesities
- Patches
- SIFT
- Basic geometric forms (Geons, quadrics)

Learning representations

- Generative models
- Discriminative models

Shape-appearance models

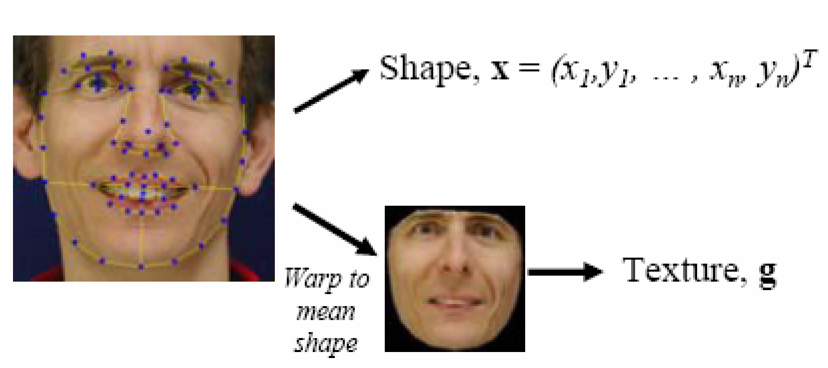
- Idea

- Features
  - Pixel intensities

- Representation
  - Subspace model of shape and appearance variations
  - Generative model

AAM = T. F. Cootes, C.J. Taylor, G. J. Edwards
Morphable models = Blanz, T. Vetter

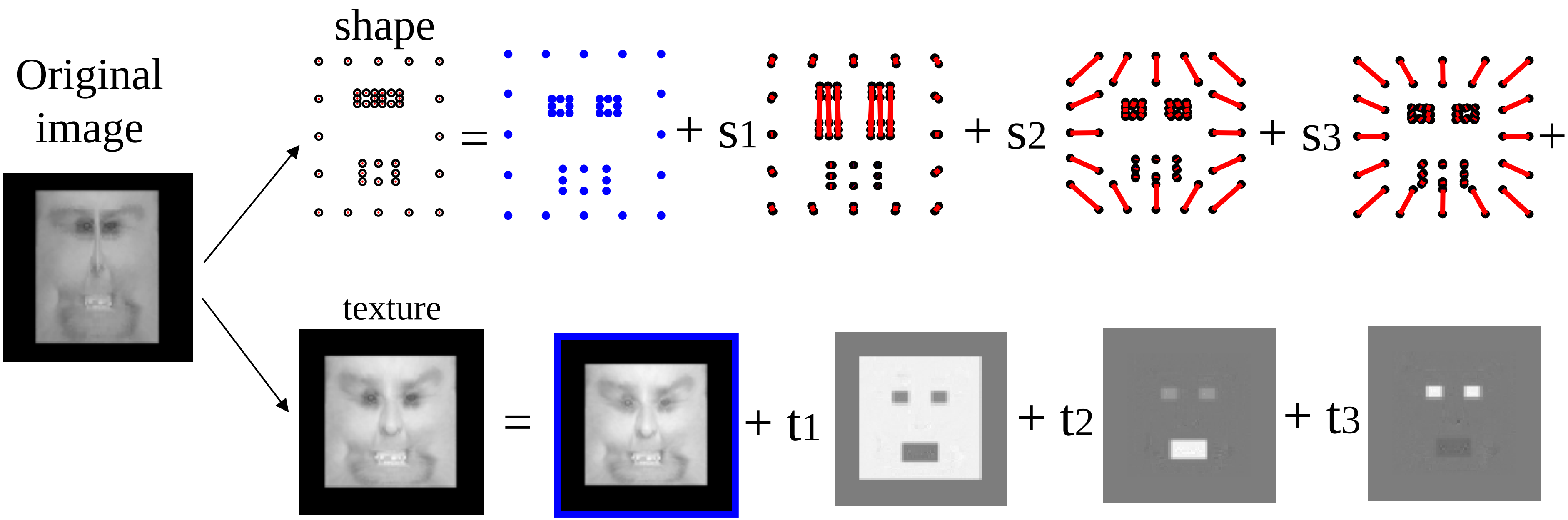
Shape-appearance models



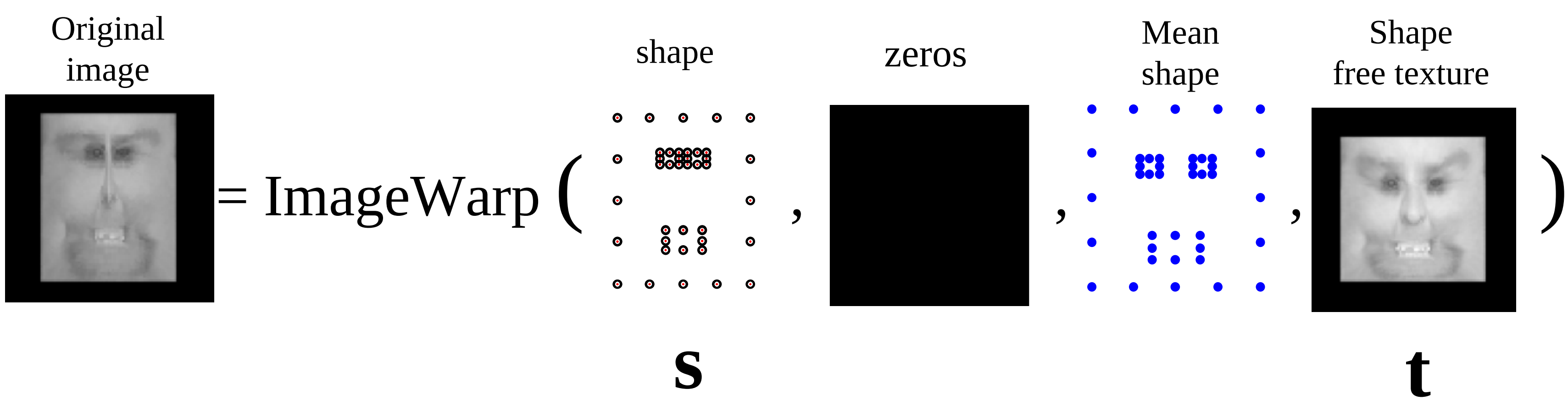
- Statistical analysis
  - shape model: $\mathbf{x} = \mathbf{x}_{mean} + \mathbf{P}_s\mathbf{b}_s$
  - texture model: $\mathbf{g} = \mathbf{g}_{mean} + \mathbf{P}_g\mathbf{b}_g$
- Parameters $\mathbf{b}_i$ control modes of variation

AAM = T. F. Cootes, C.J. Taylor, G. J. Edwards
Morphable models = Blanz, T. Vetter

Shape-appearance models



AAM uses an additional PCA, to reduce redundancy between texture and shape.



Constelation models

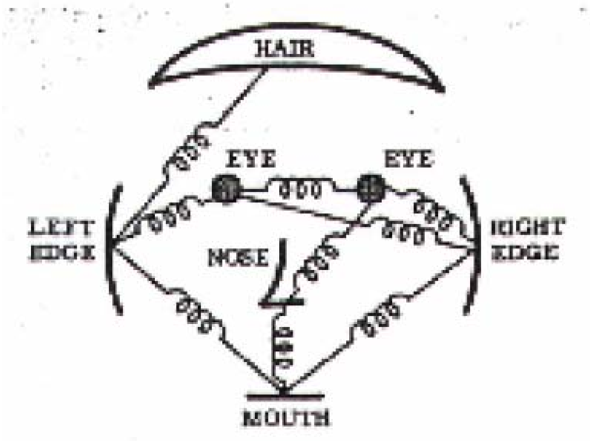
- Idea

- Features
  - Intensities, patches, SIFT features.

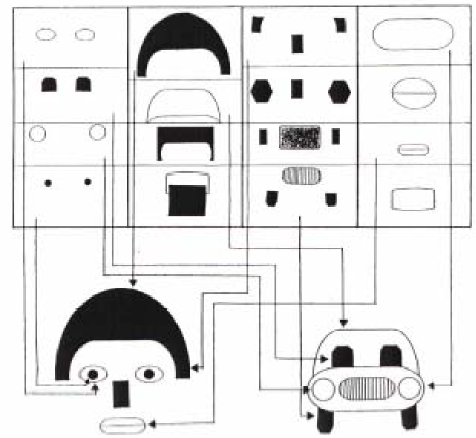
- Representation
  - Parts base representation.

AAM = T. F. Cootes, C.J. Taylor, G. J. Edwards
Morphable models = Blanz, T. Vetter

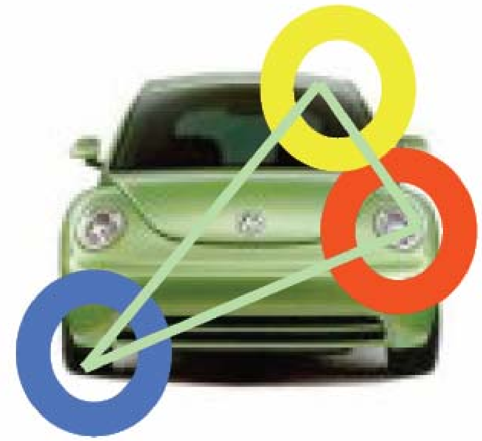
Constelations of parts



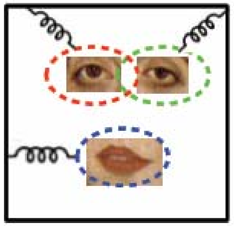
Fischler & Elschlager, 1973



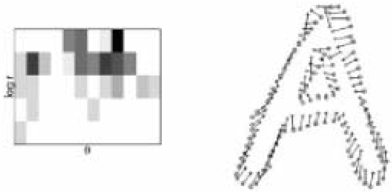
Perrett & Oram, 1993



Perona et al. '95



Schmid '99,
Lowe '99, Moreels '04



Belongie et al. '02

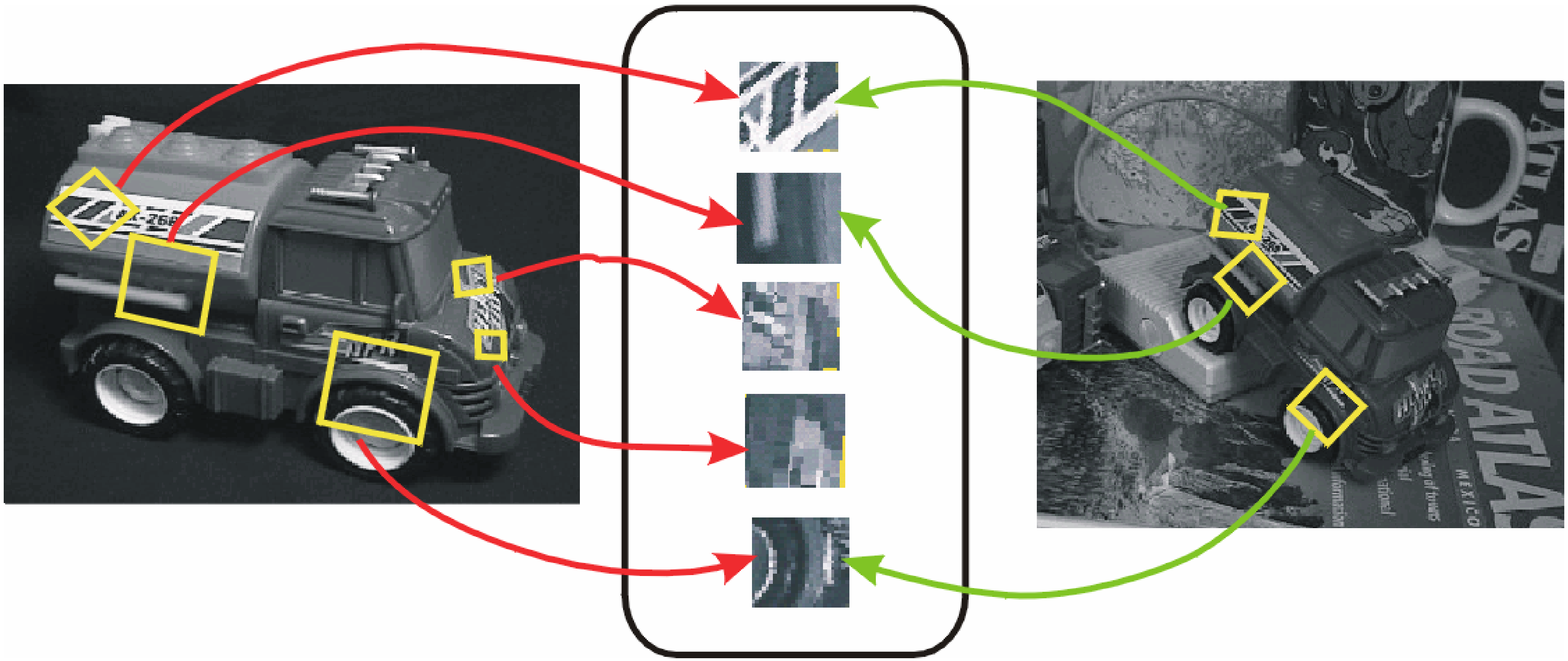
(Interest points)
Local appearance
Shape / deformation
(Clutter)
Correspondence

Slide from Perona 2005

SIFT features

Invariant Local Features

- Image content is transformed into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters

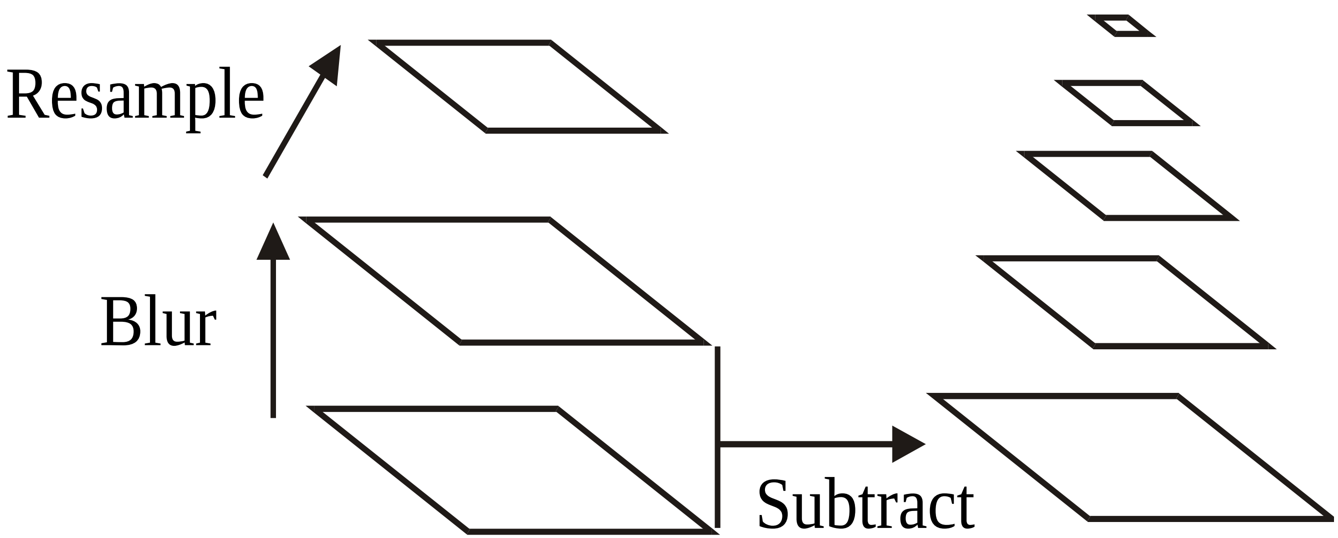


SIFT Features

Slide from David Lowe

Build Scale-Space Pyramid

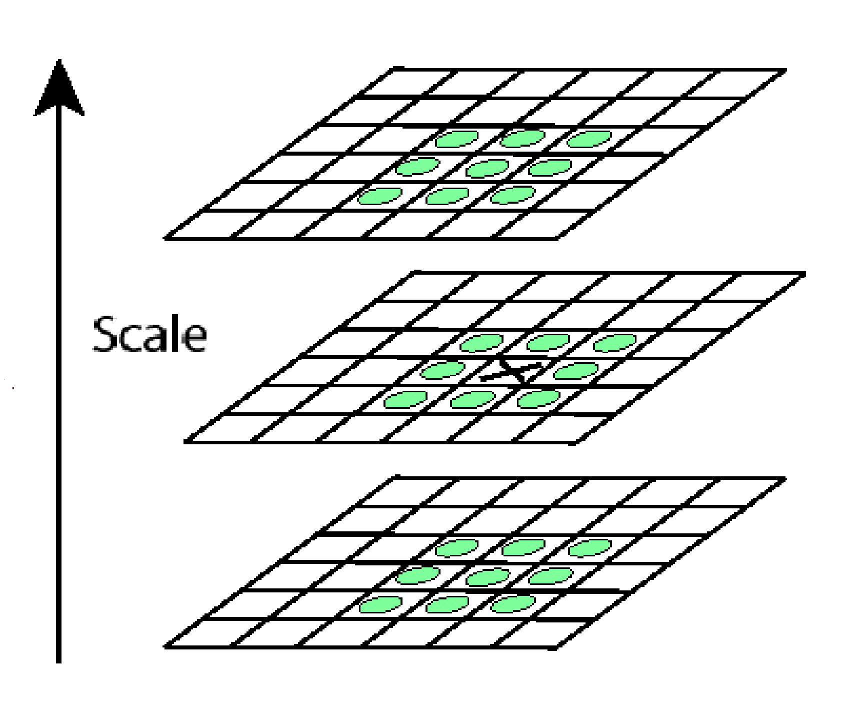
- All scales must be examined to identify scale-invariant features
- An efficient function is to compute the Difference of Gaussian (DOG) pyramid (Burt & Adelson, 1983)



Slide from David Lowe

Key point localization

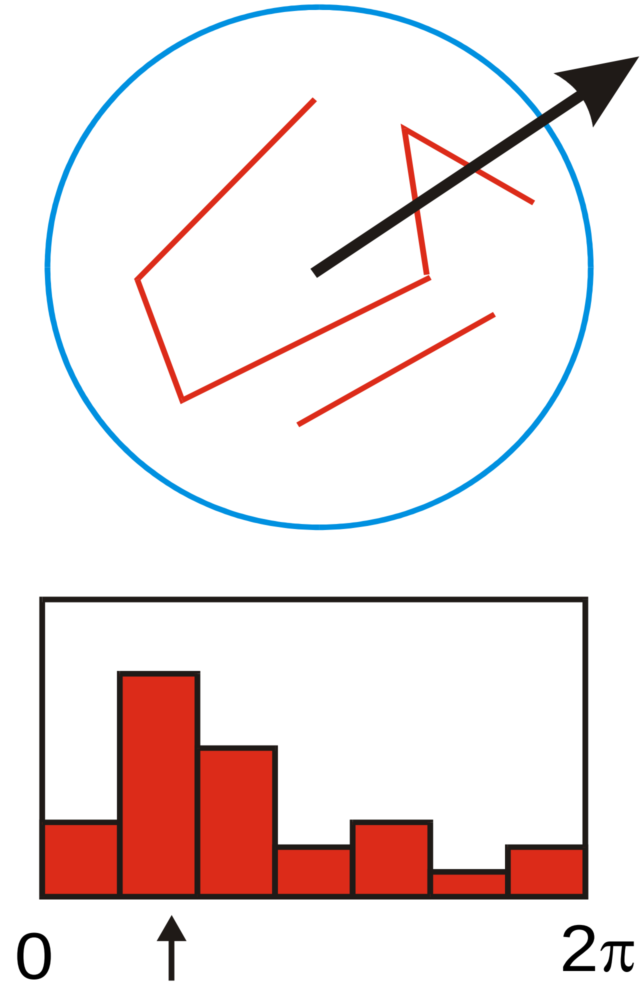
- Detect maxima and minima of difference-of-Gaussian in scale space



Slide from David Lowe

Select dominant orientation

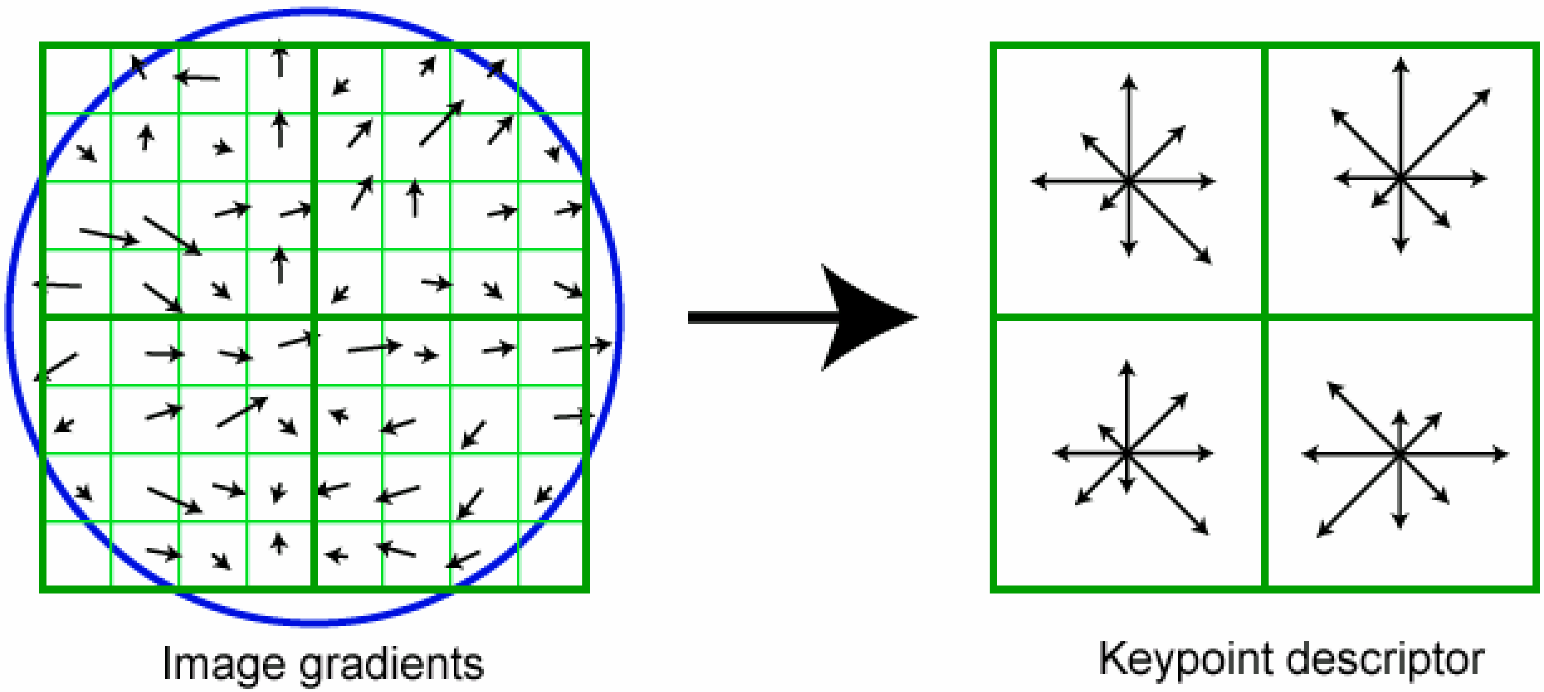
- Create histogram of local gradient directions computed at selected scale
- Assign canonical orientation at peak of smoothed histogram



Slide from David Lowe

SIFT vector formation

- Thresholded image gradients are sampled over 16x16 array of locations in scale space
- Create array of orientation histograms
- 8 orientations x 4x4 histogram array = 128



Slide from David Lowe

Invariant Local Features

- Detecting particular objects:

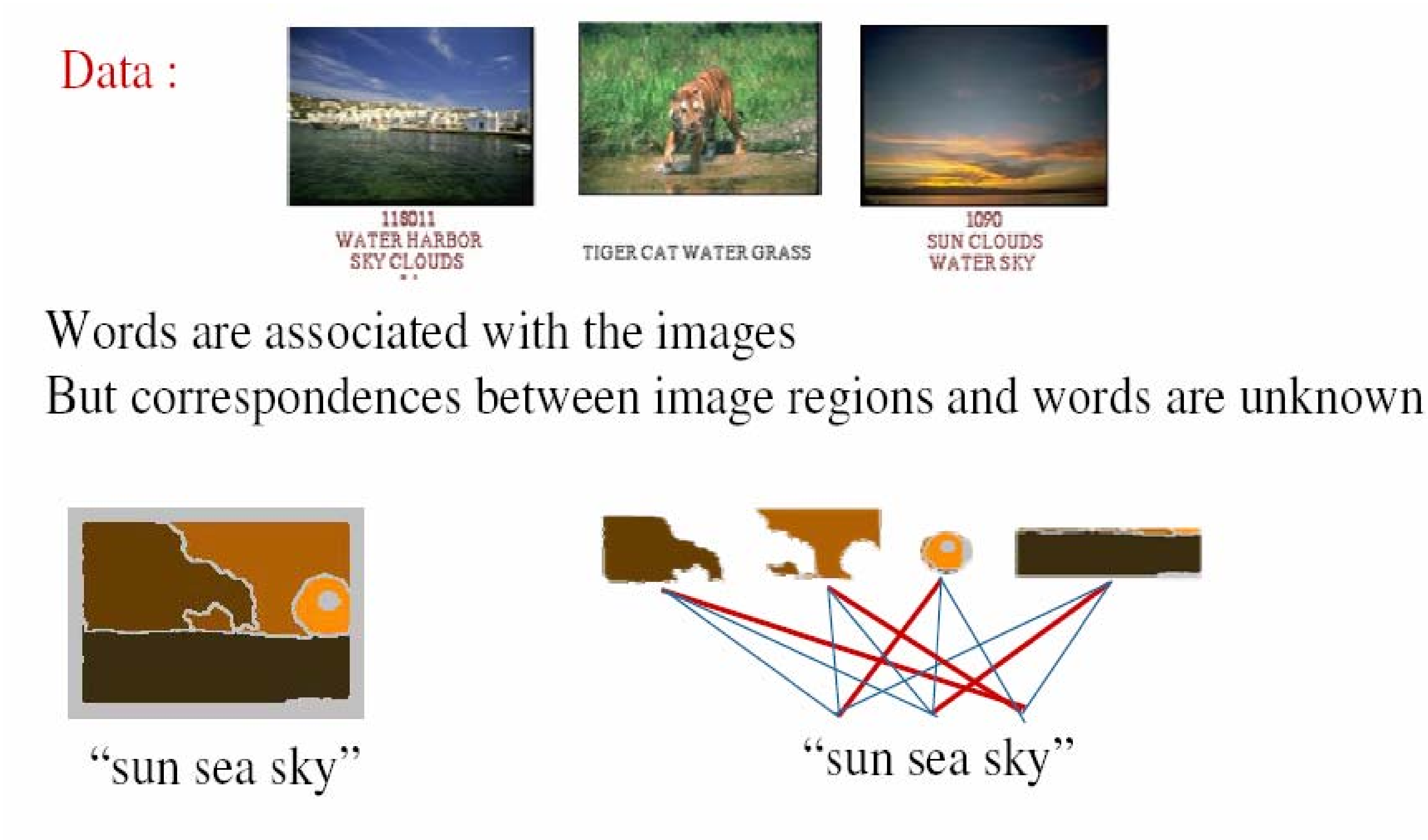
.

Lowe, 1999

Segmentation driven

- Idea
  - Avoid scaning and reduce number of candidates
- Features
  - Blobs and image regions
- Representation
  - An image is an arrangement of regions

Segmentation-recognition



Slide from Duygulu, 04

P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth. ECCV 02

Discriminative approach

- Idea

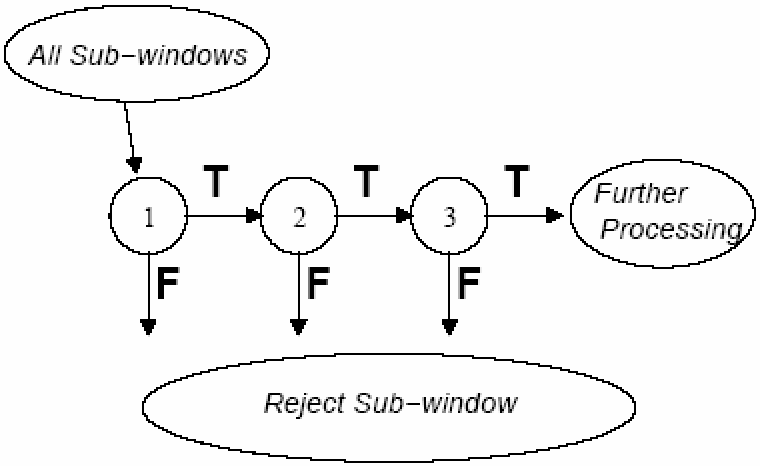
- Features
  - Pixel intensities, wavelets, patches

- Representation
  - Any of the representations before

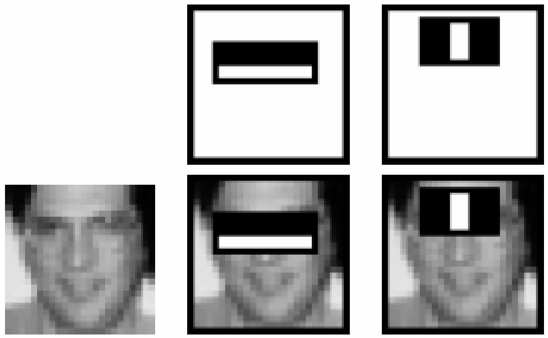
Cascade of classifiers

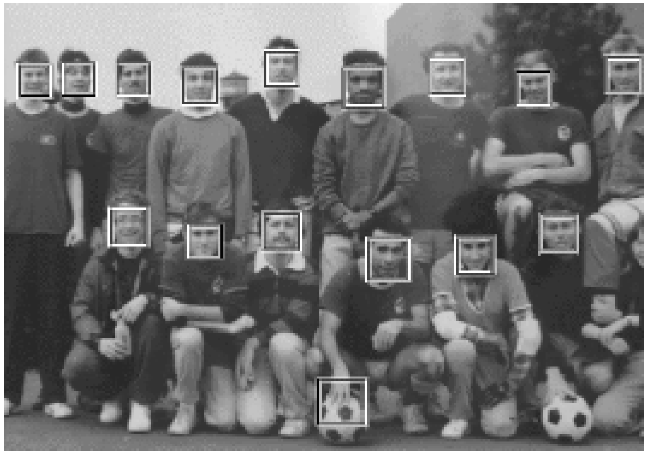
- Graded Learning for Object Detection - Fleuret, Geman (1999)
- Robust Real-time Object Detection - Viola, Jones (2001)

Cascade: classifiers of increasing complexity. Low miss rate.



Features: stumps, inspired from haar wavelets





Short introduction to Boosting

Why use boosting?

- Creates very accurate, very fast classifiers.
- Training is fast and easy to implement.
- Can handle high-dimensional data (stumps perform feature selection).
- Robust to overfitting (implicitly maximizes margin).

Boosted decision trees

- "Best off-the-shelf classifier in the world"
  – Leo Breiman, 1998
- 1 node tree = "stump"

$$f(x; \theta = (a, b, d, \phi)) = a[x_d > \phi] + d$$

- Can be used for feature selection.
- Pick best dimension *d* and threshold ϕ by exhaustive search.
- Pick best slope *a* and offset *b* using weighted least squares.

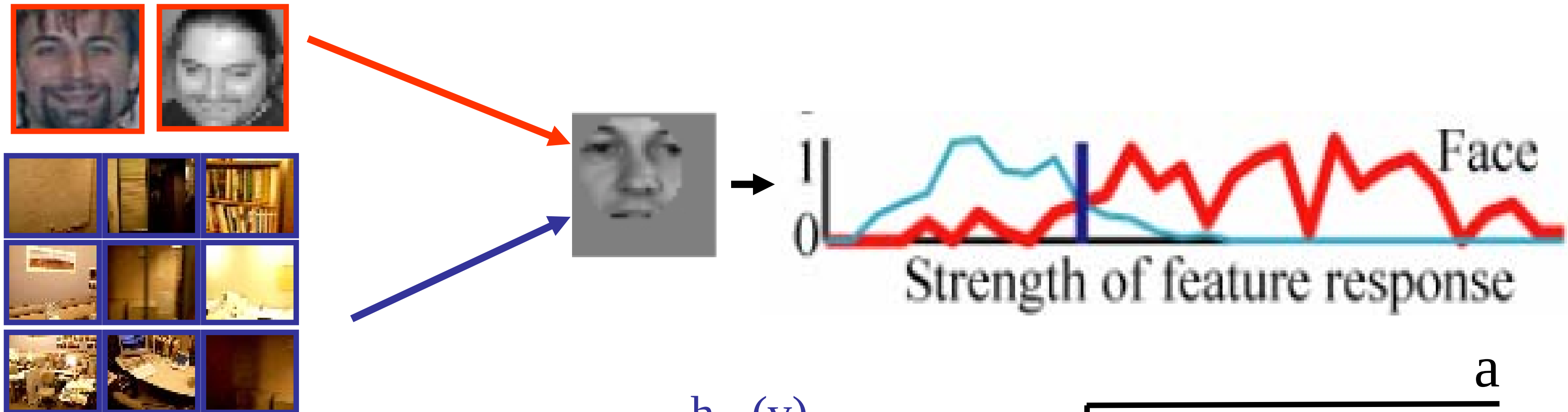
Additive models for classification

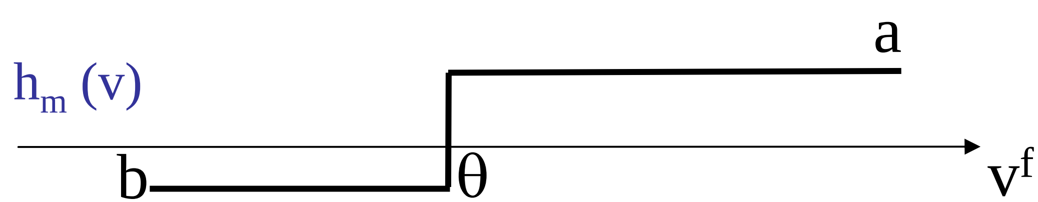
$$\underbrace{H(v,c)}_{\text{+1/-1 classification}} = \sum_{m=1}^{M} h_m(\underbrace{v}_{\text{feature responses}}, \underbrace{c}_{\text{classes}})$$

$h_m(v,c)$ is a weak classifier (performs better than chance)

$H(v,c)$ is the strong classifier obtained as a sum of weak classifiers

Example of weak classifier (stumps)





A decision stump is a threshold on a single feature

Each decision stump has 4 parameters: {f, θ, a, b}

f = template index (selected among a dictionary of 2000 templates)

θ = Threshold,

a,b = average class value (-1, +1) at each side of the threshold

Flavors of boosting

• Different boosting algorithms use different loss functions or minimization procedures (Freund & Shapire, 1995; Friedman, Hastie, Tibshhirani, 1998).

• We base our approach on Gentle boosting: learns faster than others (Friedman, Hastie, Tibshhirani, 1998; Lienahart, Kuranov, & Pisarevsky, 2003).

Multi-class Boosting

We use the exponential multi-class cost function

classes

$$J = \sum_{c=1}^{C} E\left[e^{-z^c H(v,c)}\right]$$

cost function

membership in class c, +1/-1

classifier output for class c

Freund & Shapire, 1995; Friedman, Hastie, Tibshhirani, 1998

Weak learners are shared

At each boosting round, we add a perturbation or “weak learner” which is shared across some classes:

$$H(v_i, c) := H(v_i, c) + h_m(v_i, c)$$

We add the weak classifier that provides the best reduction of the exponential cost

$$J = \sum_{c=1}^{C} E\left[e^{-z^c H(v,c)}\right] = \sum_{c=1}^{C} E\left[e^{-z^c \left(H(v_i, c) + h_m(v_i, c)\right)}\right]$$

Freund & Shapire, 1995; Friedman, Hastie, Tibshhirani, 1998

Use Newton's method to select weak learners

Treat h_m as a perturbation, and expand loss J to second order in h_m

$$\arg\min_{h_m} J(H+h_m) \simeq \arg\min_{h_m} \sum_{c=1}^{C} E\left[e^{-z^c H(v,c)}(z^c - h_m)^2\right]$$

cost function

classifier with perturbation

reweighting

squared error

Freund & Shapire, 1995; Friedman, Hastie, Tibshhirani, 1998

Multi-class Boosting

Replacing the expectation with an empirical expectation over the training data, and defining weights $w_i^c = e^{-z_i^c H(v_i, c)}$ for example i and class c, this reduces to minimizing the weighted squared error:

$$J_{wse} = \sum_{c=1}^{C} \sum_{i=1}^{N} w_i^c (z_i^c - h_m(v_i, c))^2.$$

Weight squared error over training data

weight

squared error

Freund & Shapire, 1995; Friedman, Hastie, Tibshhirani, 1998

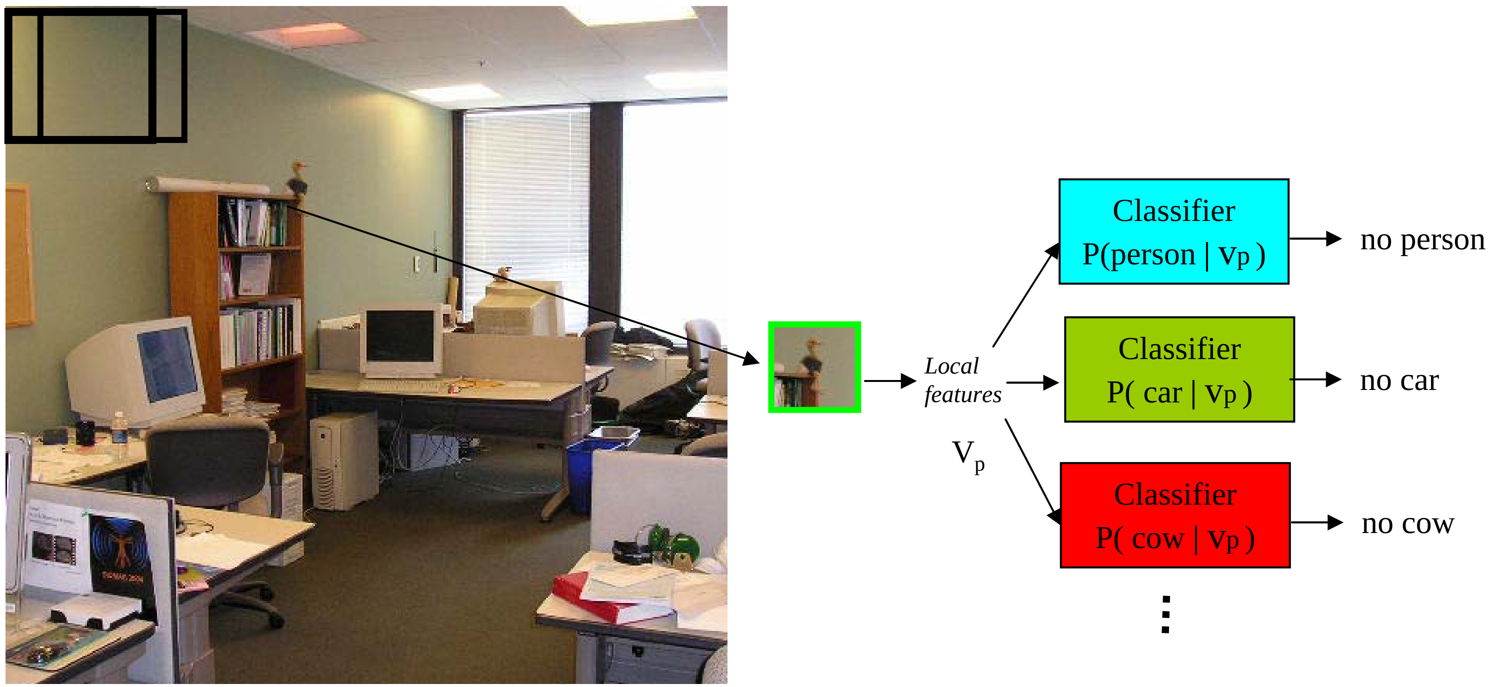
Demo
Boosting for object detection

Summary

1) Object representation based on **local** features:

Summary

2) Search strategy:

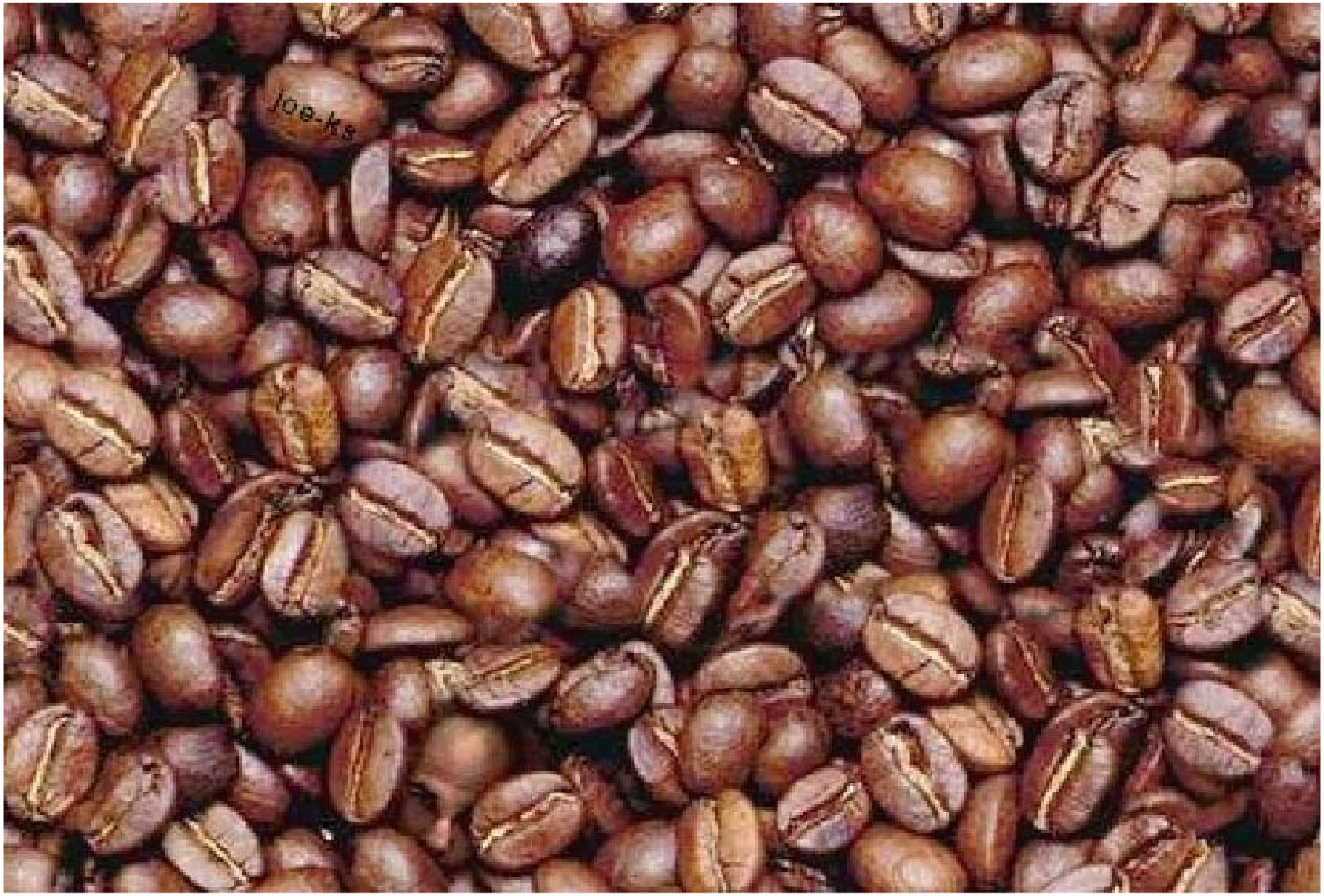


Agarwal & Roth, (02), Moghaddam, Pentland (97), Turk, Pentland (91),Vidal-Naquet, Ullman, (03)

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.

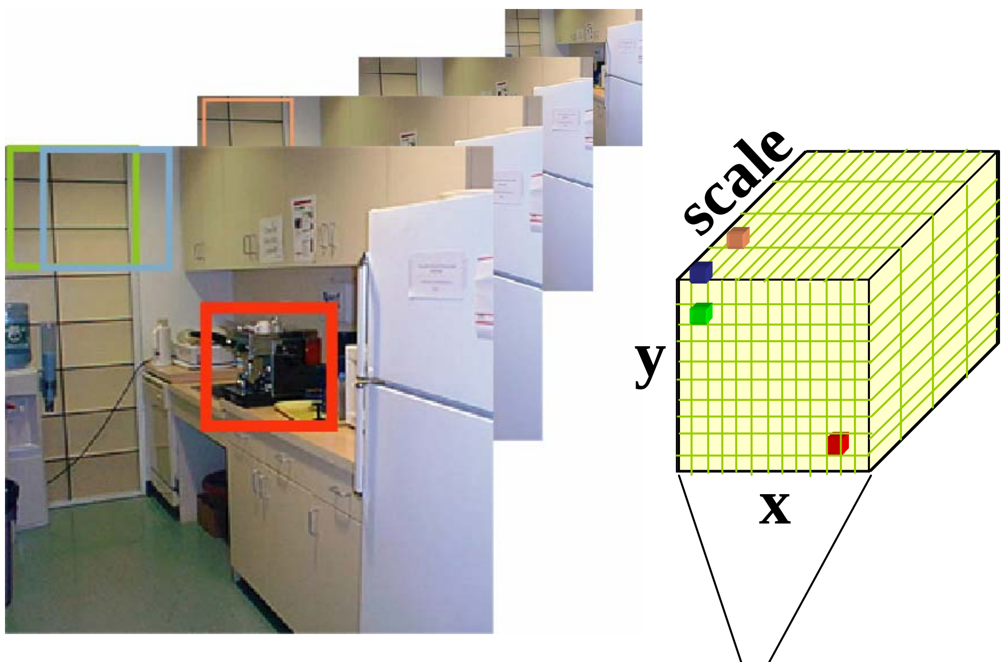
SCENES

Try to find the face in this image



The search space is huge

"Like finding needles in a haystack"

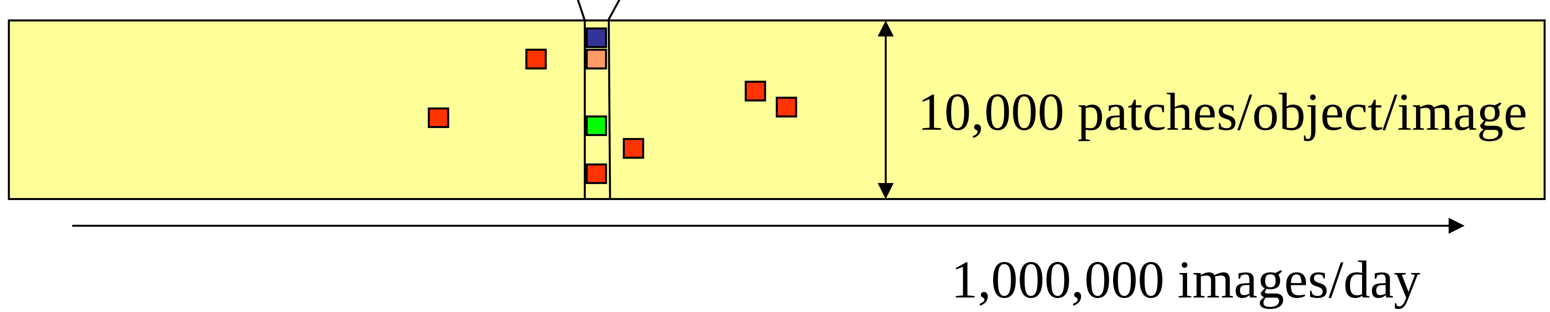


For each object:

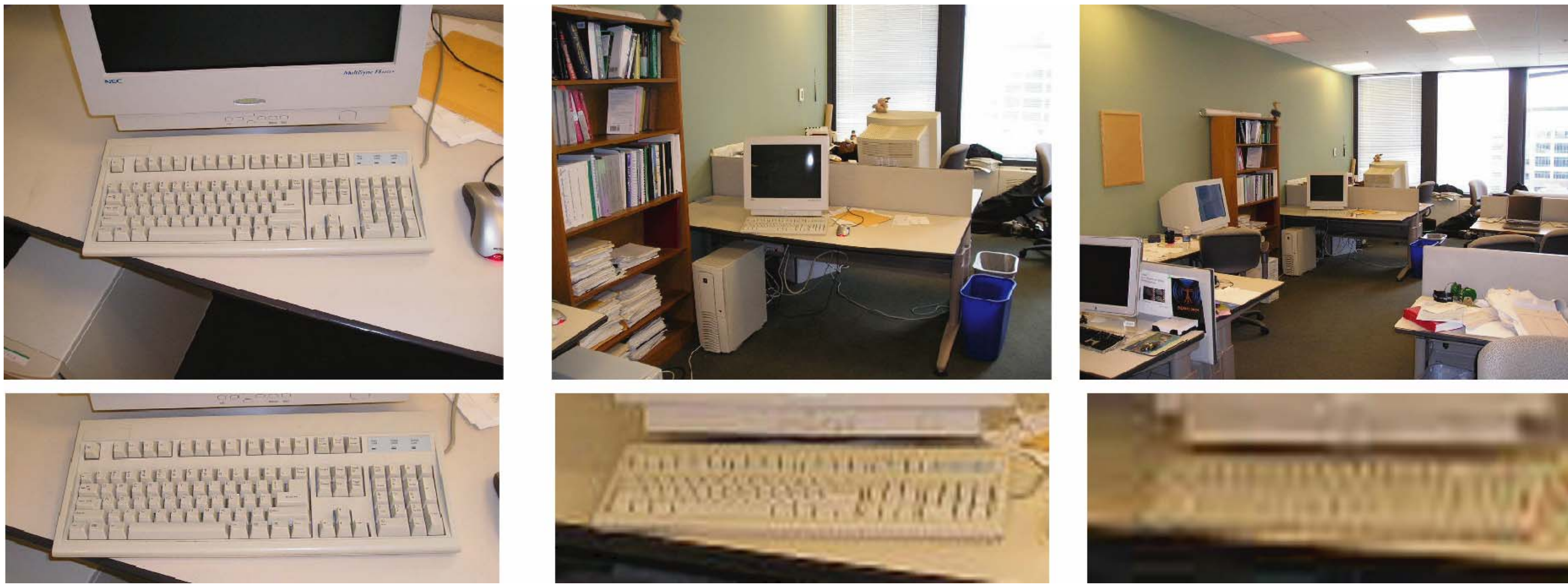
- Need to search over locations and scales

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

- Slow (many patches to examine)



Local features are not even sufficient

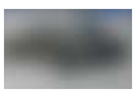


The multiple personalities of a blob

The multiple personalities of a blob

The multiple personalities of a blob

The multiple personalities of a blob

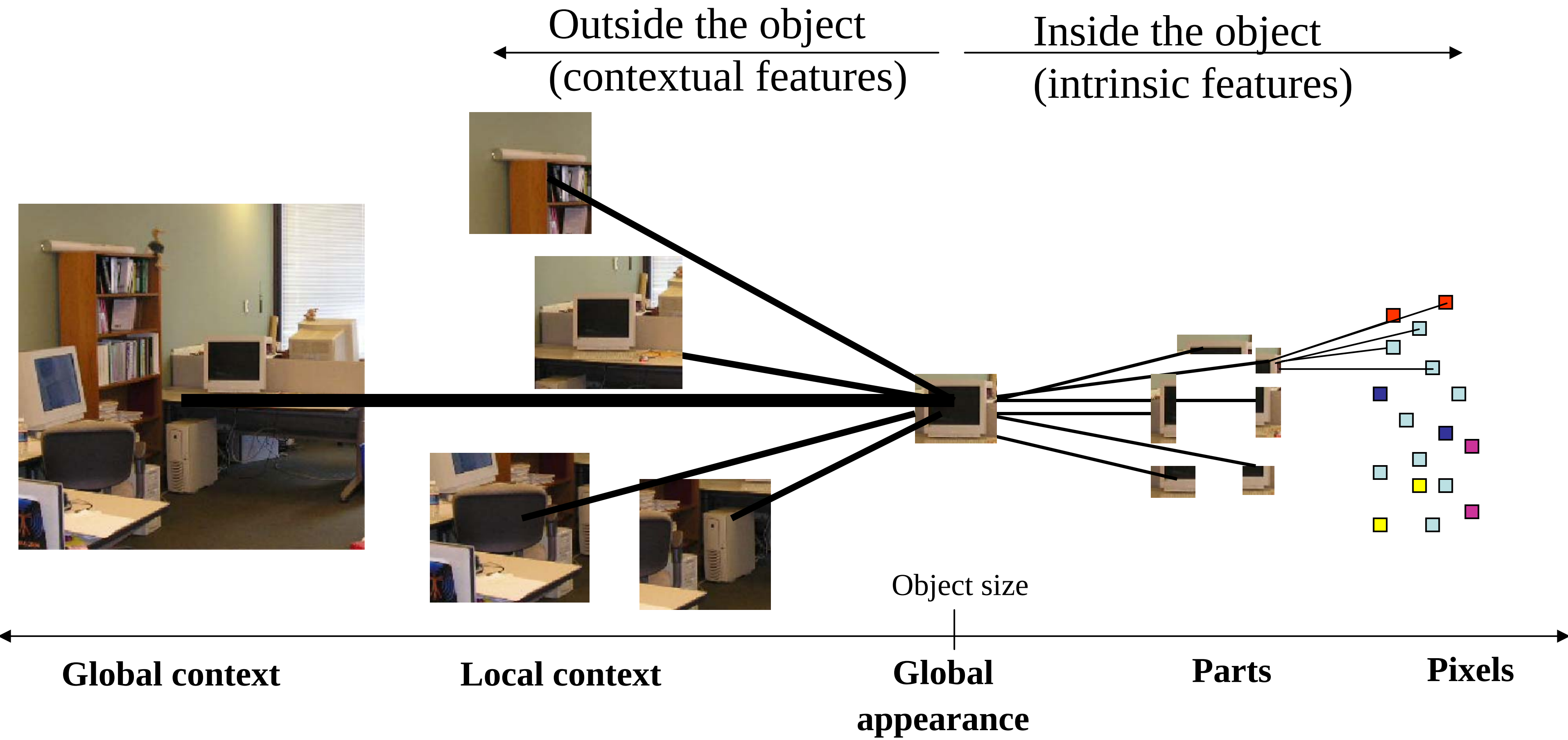


Not everything fits inside a rectangle

- e.g., detecting irregularly-shaped “stuff”
  - Grass, trees, roads, building facades
- e.g., detecting non-rigid/ articulated/ “wiry” things
  - - people, chairs, desk lamps

Source: MIT-CSAIL database of Objects and Scenes

Looking outside the box



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.

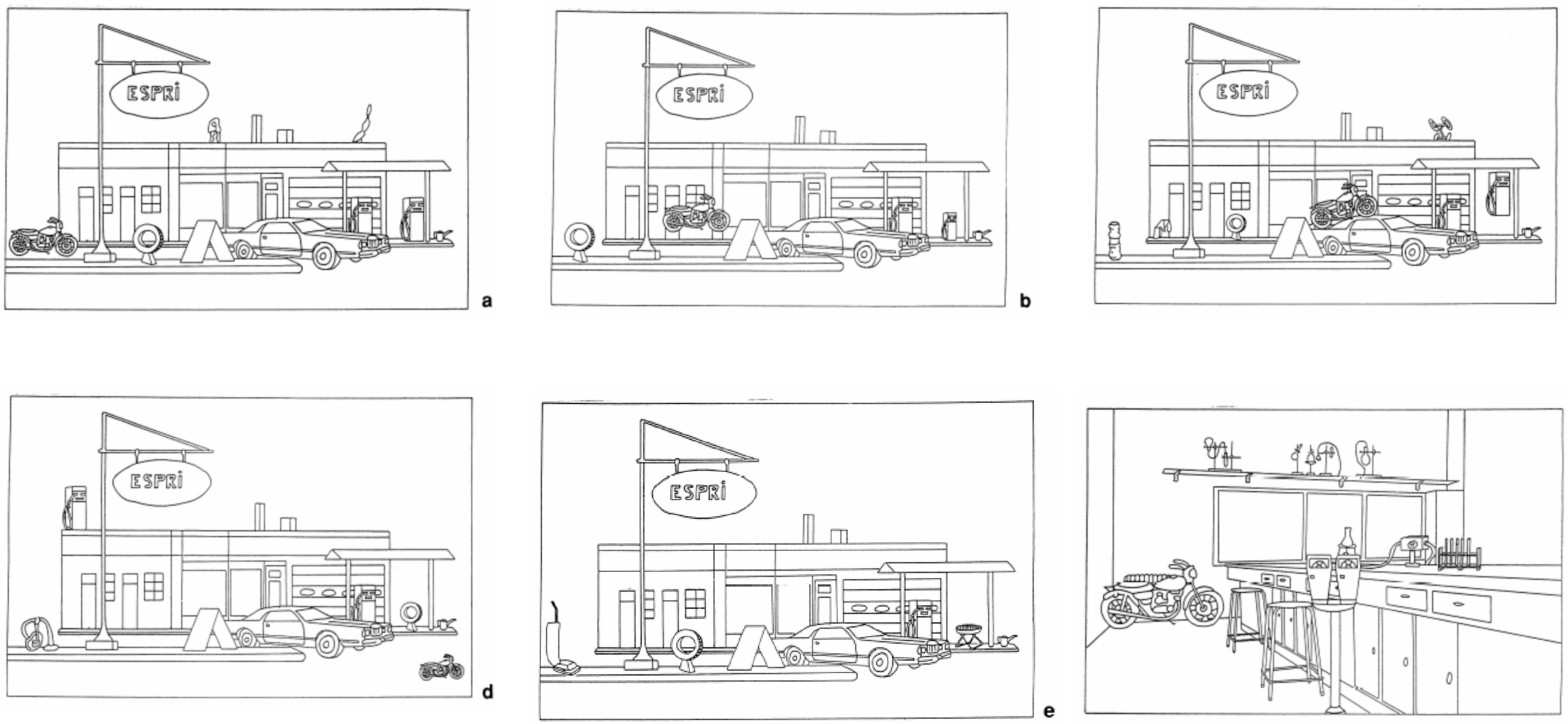
What is visual scene *context*?

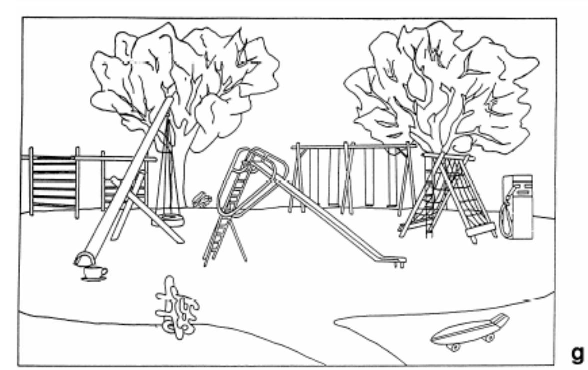
- A specific scene category (a coffeemaker is usually in a kitchen)
- The structure of the scene background (a chair is on the ground, not the ceiling)
- A combination of objects of shapes (TV+sofa+rug+bookshelf = living-room)
- Spatial relationships between shapes

Scene Context and Object Consistencies

- Biederman et al (82) proposed that five classes of relations exist between an object and its scene background:
- (1) **Interposition** (object interrupts their background)
- (2) **Support** (objects tend to rest on surfaces)
- (3) **Probability** (objects tend to be found in some scenes but not others)
- (4) **Position** (given an object is probable in a scene, it often is found in position but not others)
- (5) **Familiar size** (objects have a limited set of size relations with other objects)

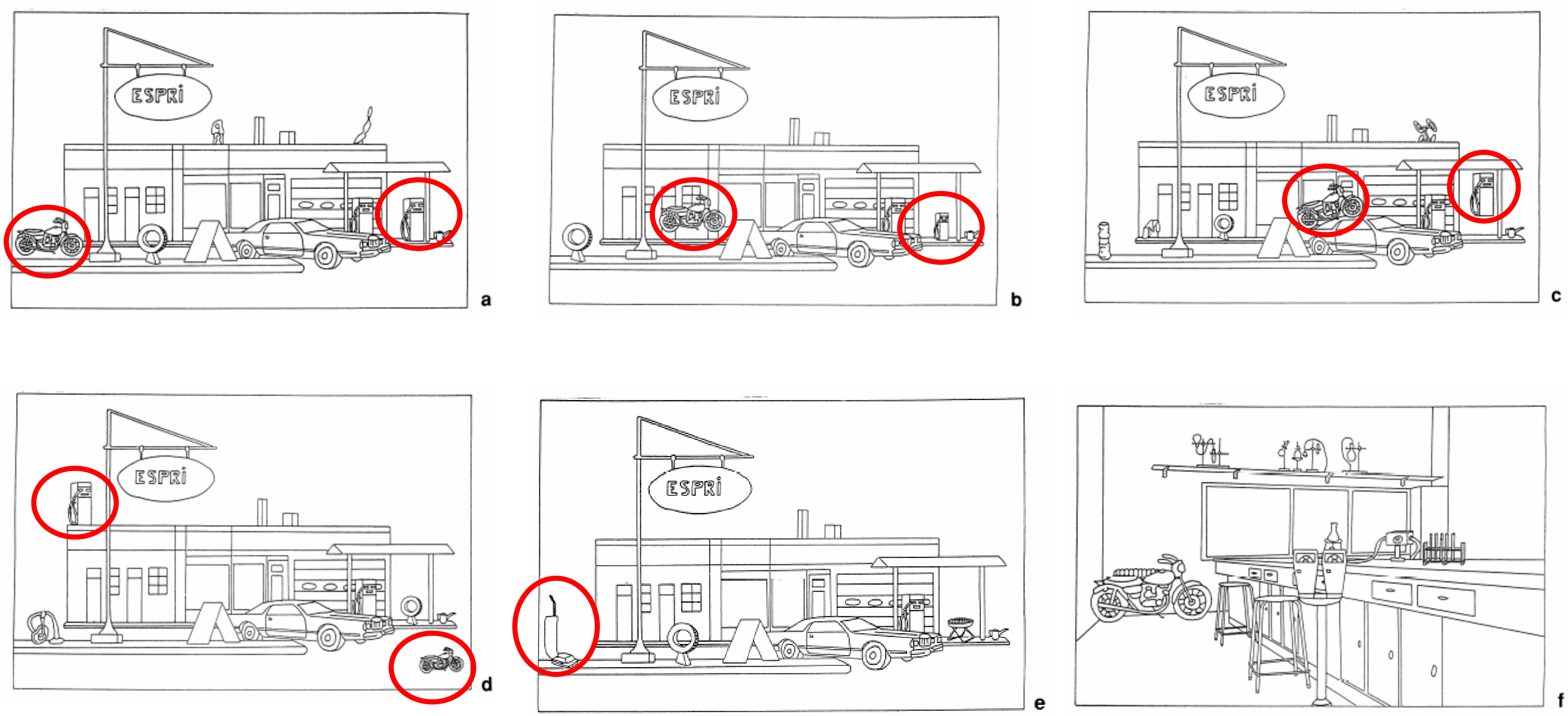
Object Consistencies



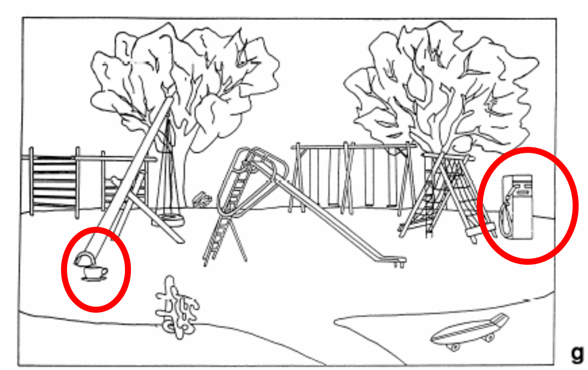


Biederman et al (1982), DeGraef(1990).

Object Consistencies



Examples of inconsistencies



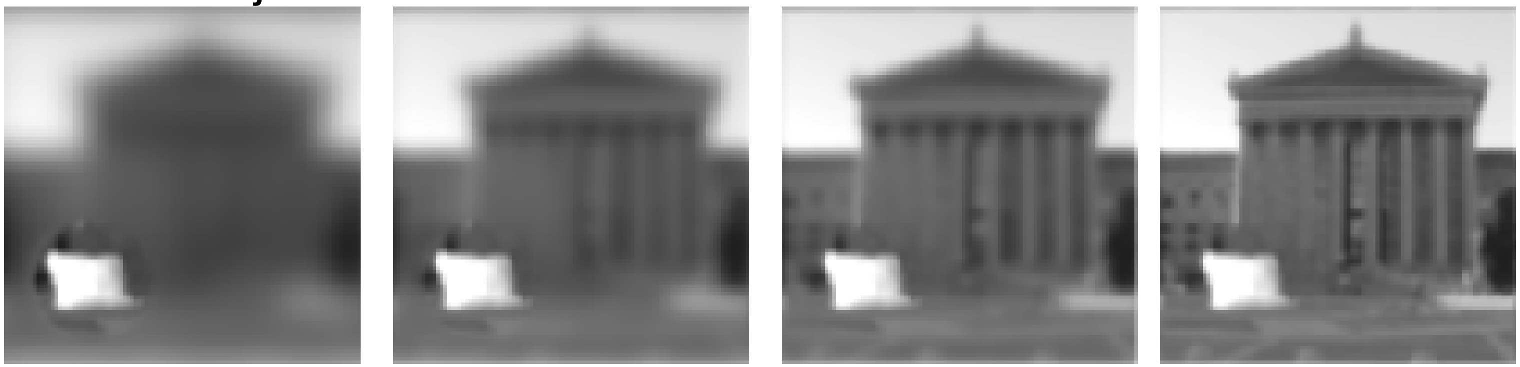
Biederman et al (1982), DeGraef(1990).

Rapid scene processing

- Conceptual information about a picture is available with a glimpse of > 100 ms (M. Potter)
- Scene processing can be quickly done without much object information (Schyns & Oliva, 1994)

Object priming

Inconsistent object

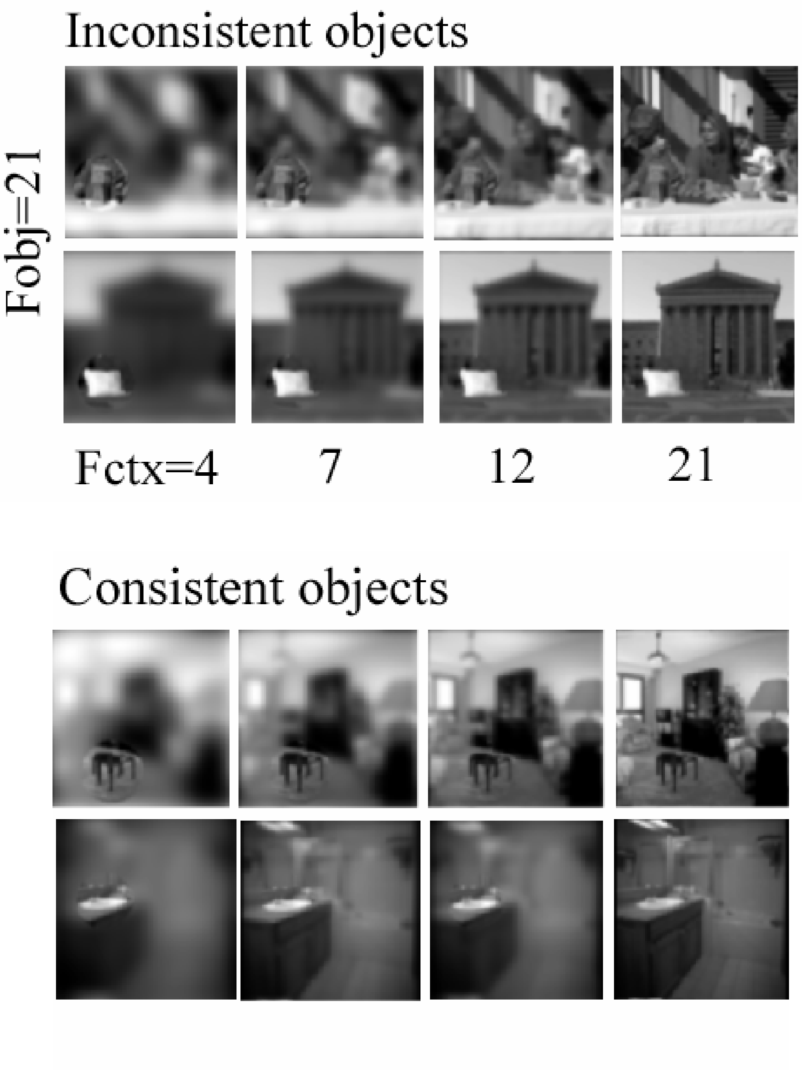


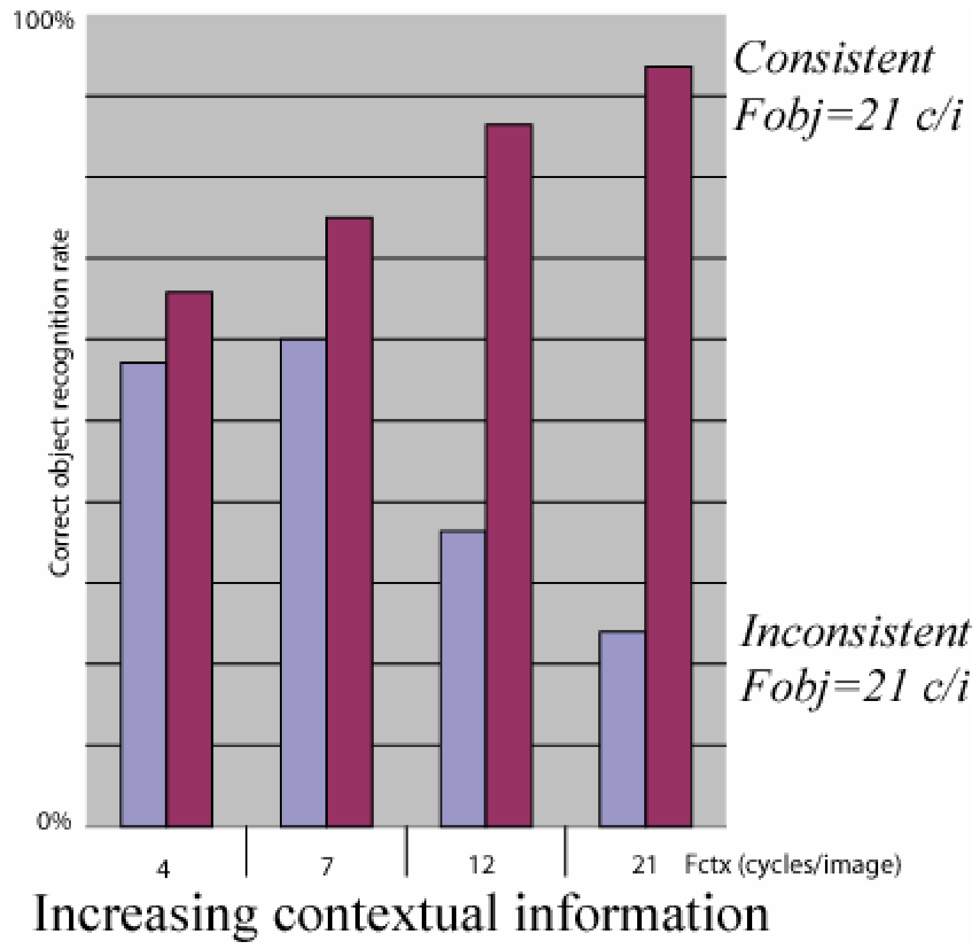
Consistent object

Increasing contextual information

Torralba, Sinha, Oliva, VSS 2001

Object priming

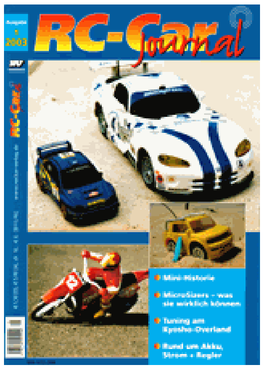




Torralba, Sinha, Oliva, VSS 2001

Why is context important?

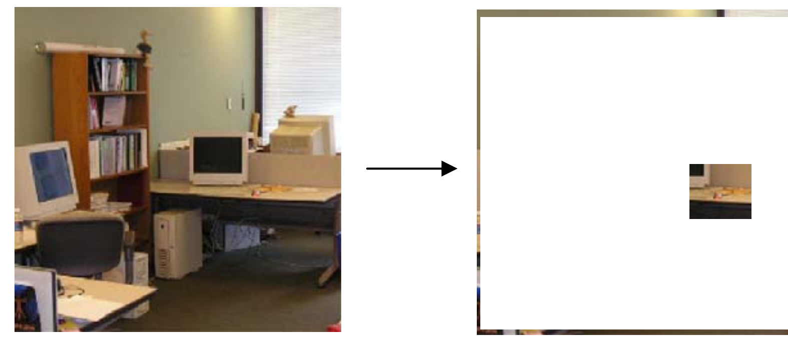
- Changes the interpretation of an object (or its function)



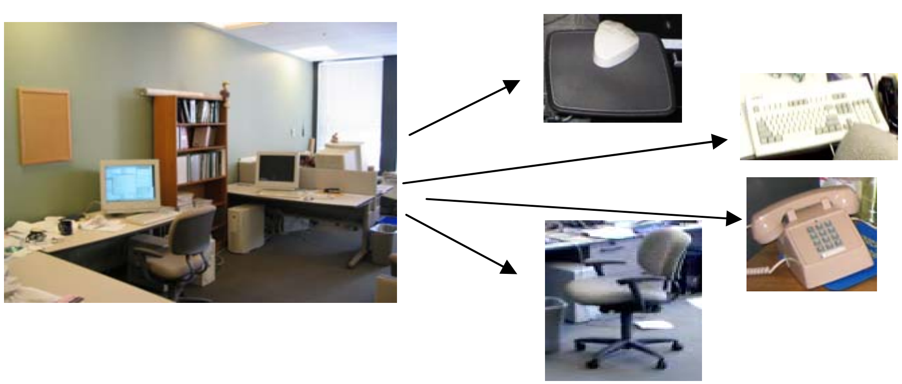
- Context defines what an unexpected event is

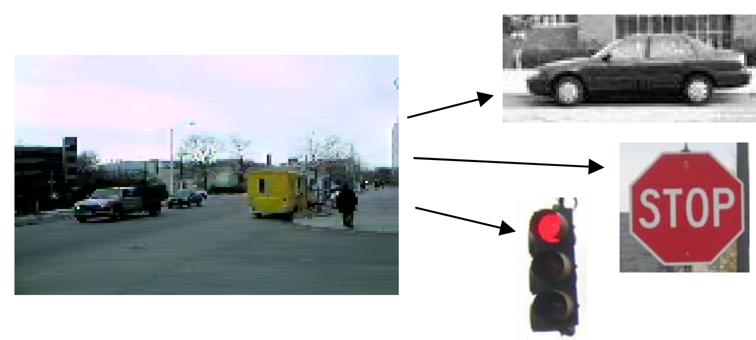
Why is context important?

- Reduces the search space



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





Context models

The problem: how to represent context?

V_C might have a very high dimensionality. There are as many ways of breaking down the dimensionality of V_C as there are possible definitions of contextual representations.

How far can we go without object detectors?

Previous work on context

- Strat & Fischler (91)

Context defined using hand-written rules about relationships between objects

- Torralba & Sinha (01), Torralba (03)

Global context to predict objects.

- Fink & Perona (03)

Use boosting incorporating the output of multiple detectors to generate contextual weak-classifiers.

- Murphy, Torralba & Freeman (03)

Use graphical models to represent the relation between global context and objects.

- Carbonetto, Freitas & Barnard (04)

They extend the work on “words and images” by adding spatial consistency between labels.

- He, Zemel & Carreira-Perpinan (04)

Use dense connectivity for incorporating spatial context using Multiscale conditional random fields.

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

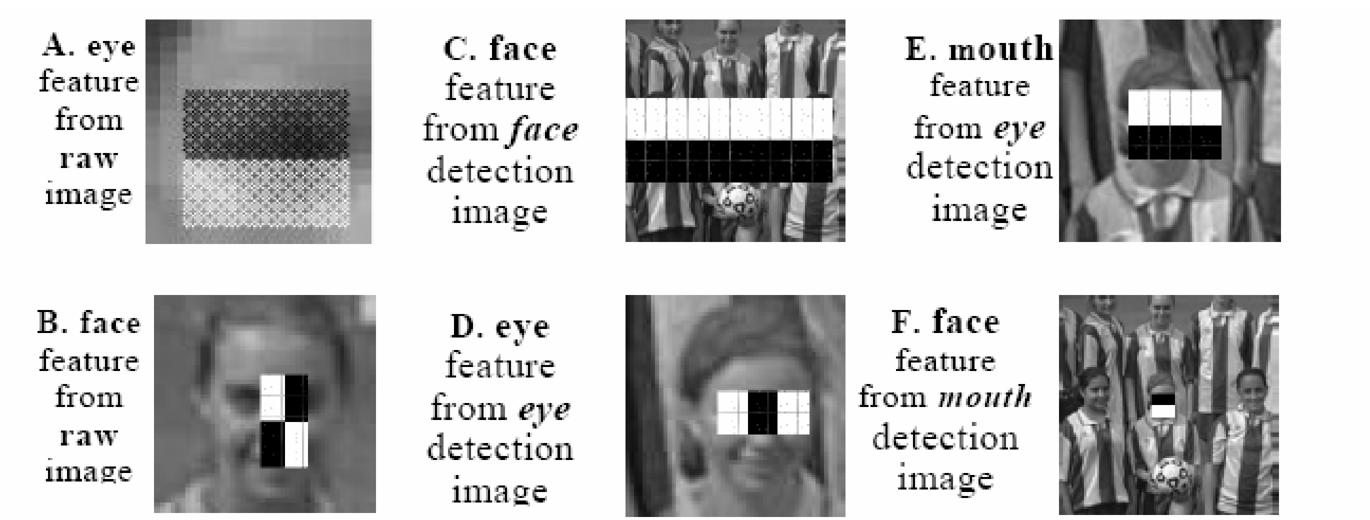
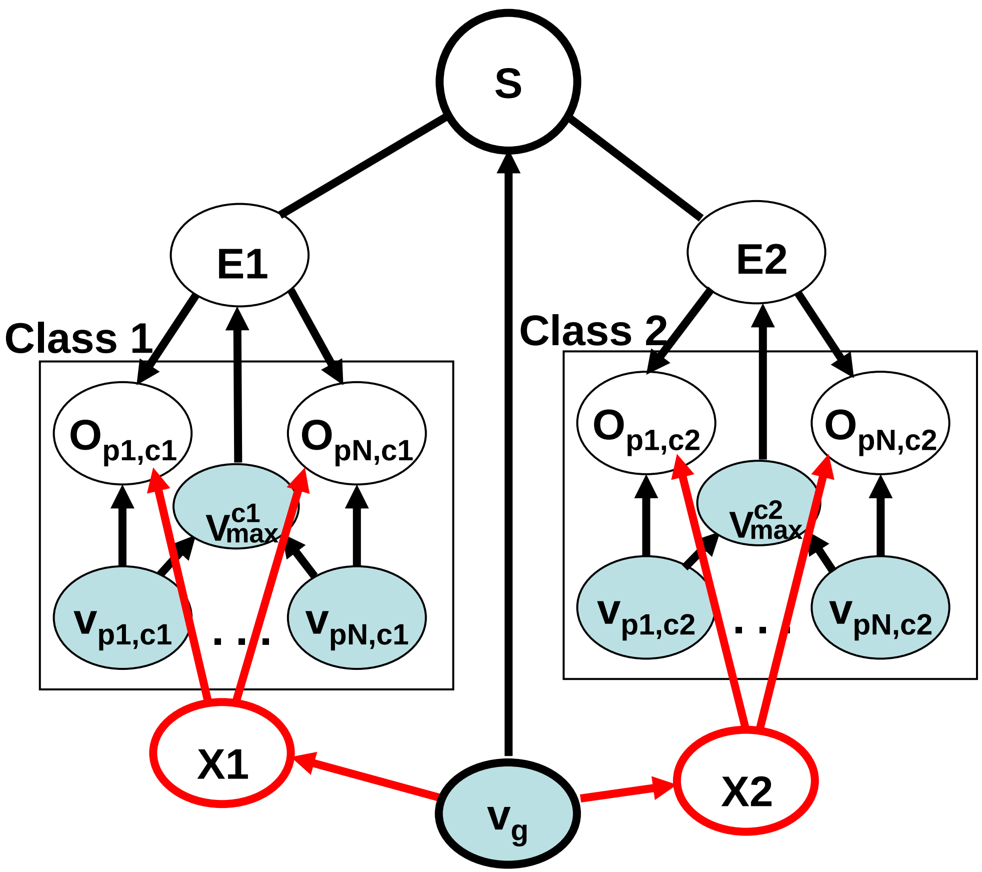


Figure 5: **A-E.** Emerging features of eyes, mouths and faces (presented on windows of raw images for legibility). The windows' scale is defined by the detected object size and by the map mode (local or contextual). **C.** faces are detected using face detection maps H^{Face}, exploiting the fact that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

Previous work on context

- Murphy, Torralba & Freeman (03)

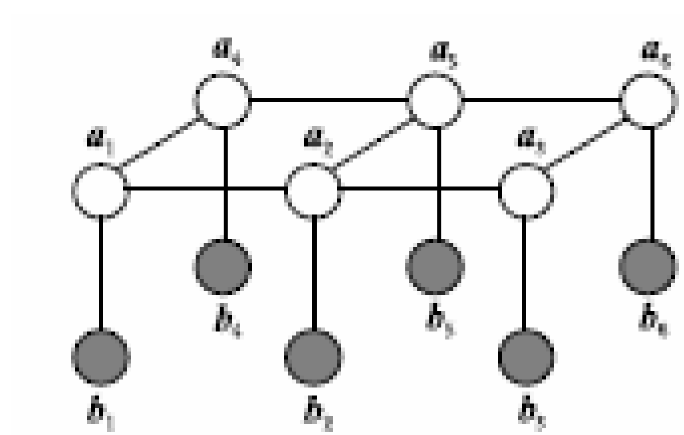
Use global context to predict objects but there is no modeling of spatial relationships between objects.



Keyboards

Previous work on context

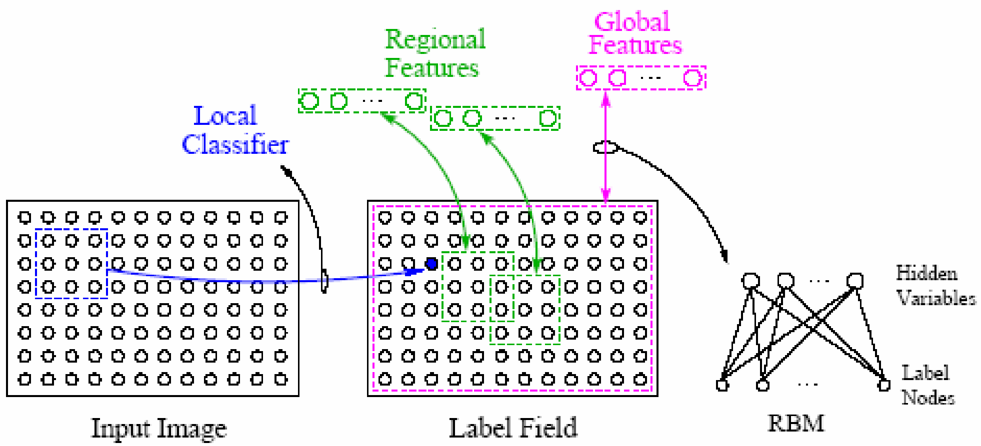
- Carbonetto, de Freitas & Barnard (04)
- Enforce spatial consistency between labels using MRF

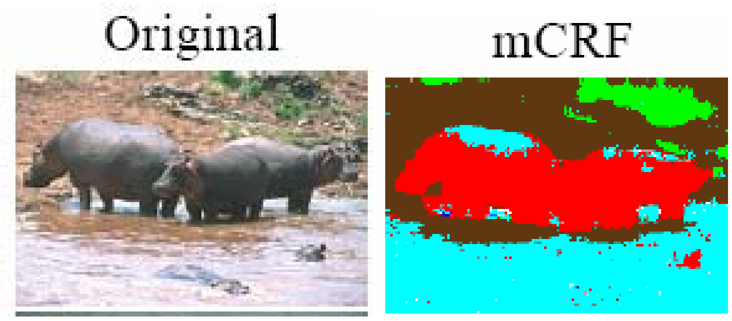


Previous work on context

- He, Zemel & Carreira-Perpinan (04)

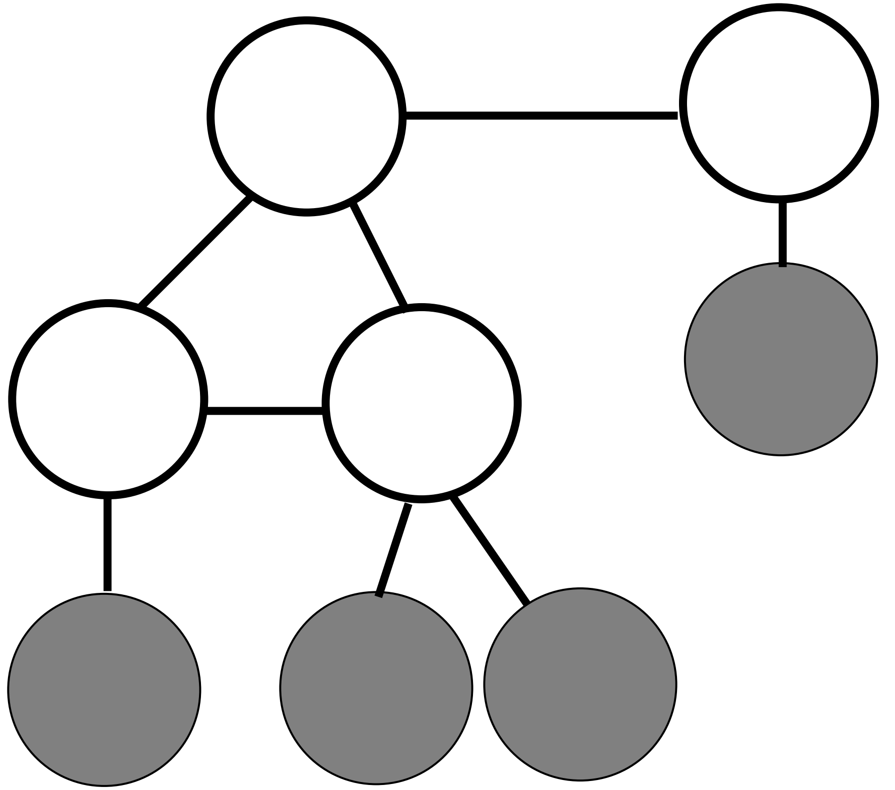
Use latent variables to induce long distance correlations between labels in a Conditional Random Field (CRF)



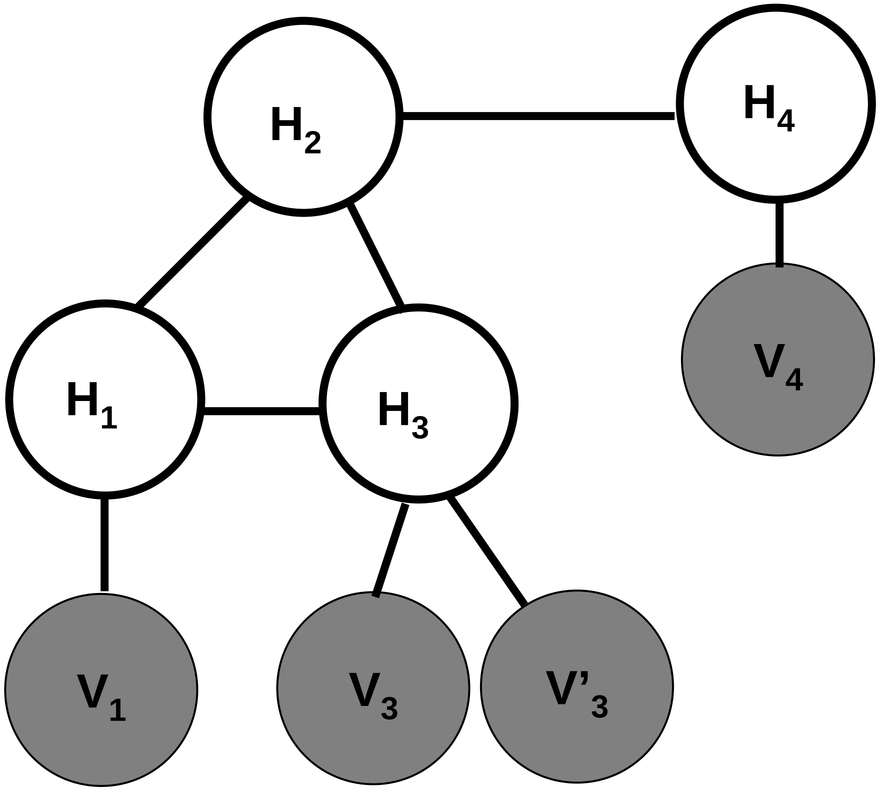


How do we exploit relationships between parts/ wholes to overcome local ambiguity?

Use probabilistic graphical models!



What is a graphical model?



- Nodes = random variables
  - Shaded = observed
  - Clear = hidden
- Arcs = (soft) constraints
- Bayes nets are a special case
- Goal of inference: state estimation

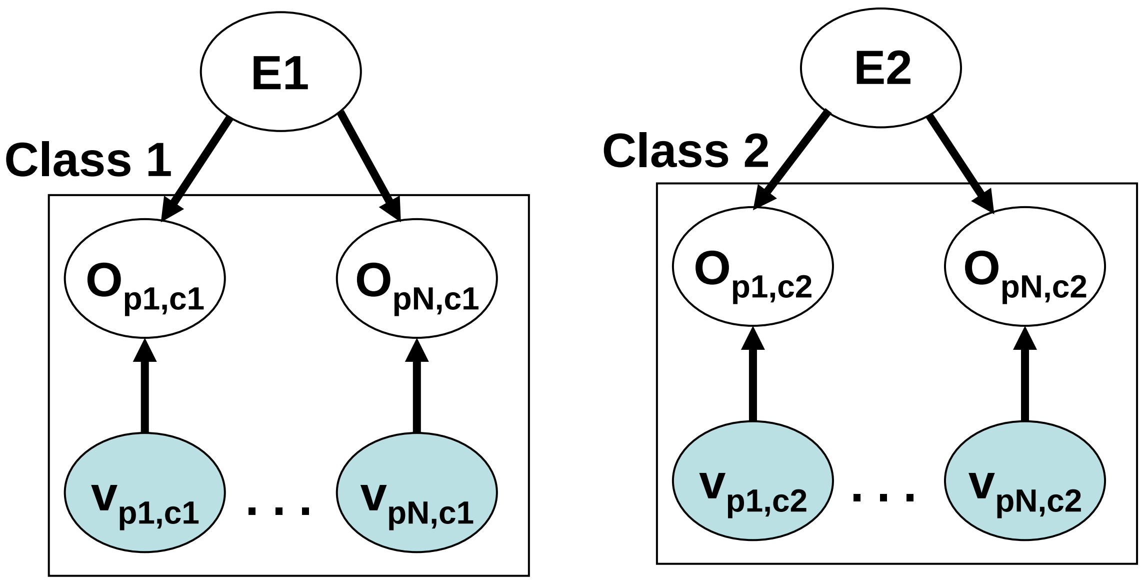
$$P_\theta(H_i | v_{1:4})$$

- Goal of learning: parameter estimation

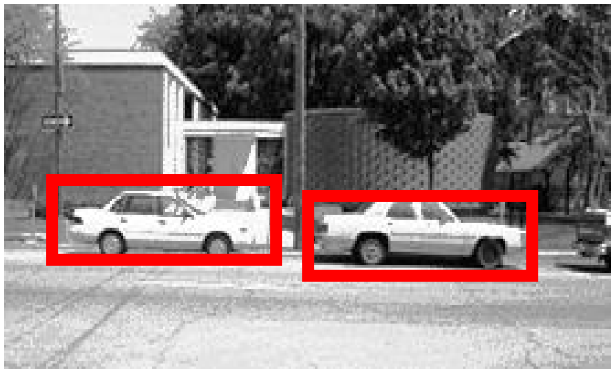
$$\arg\max_\theta P_\theta(h_{1:4} | v_{1:4})$$

Including scene-context for object detection

E_c = Exists object c anywhere in image?



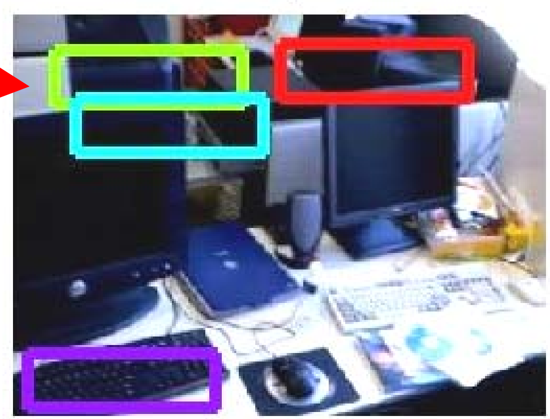
$O_{p,c}$ = Object c in patch p?

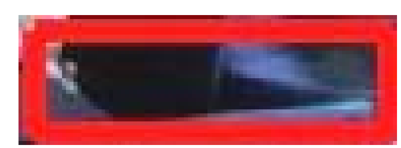


$V_{p,c}$ = Features for class c in patch p

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.

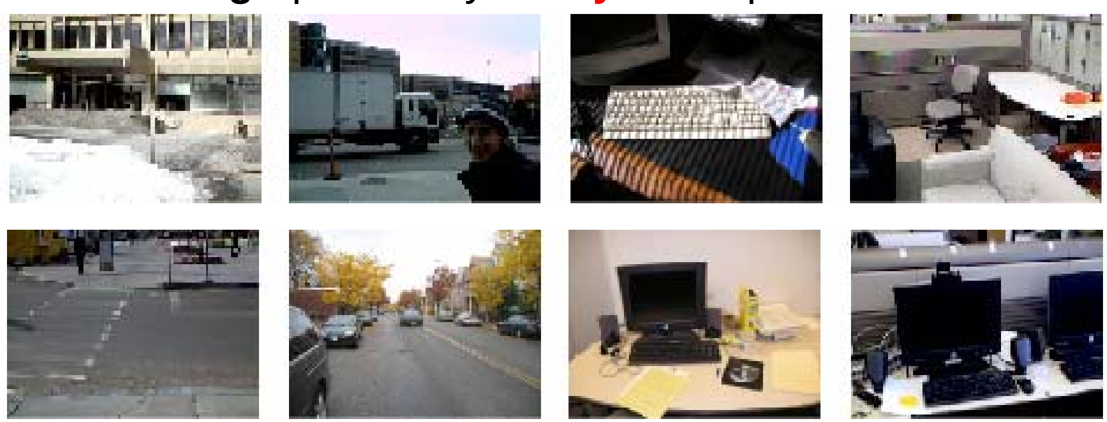




Symptoms of local features only

Low probability of **keyboard** presence

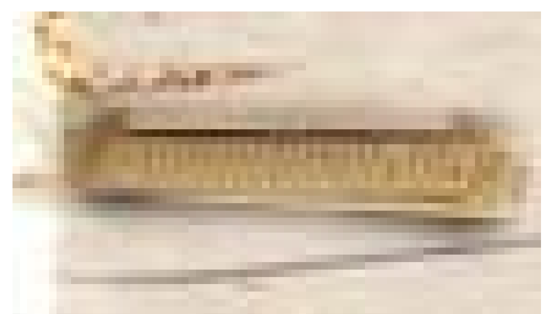
High probability of **keyboard** presence



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

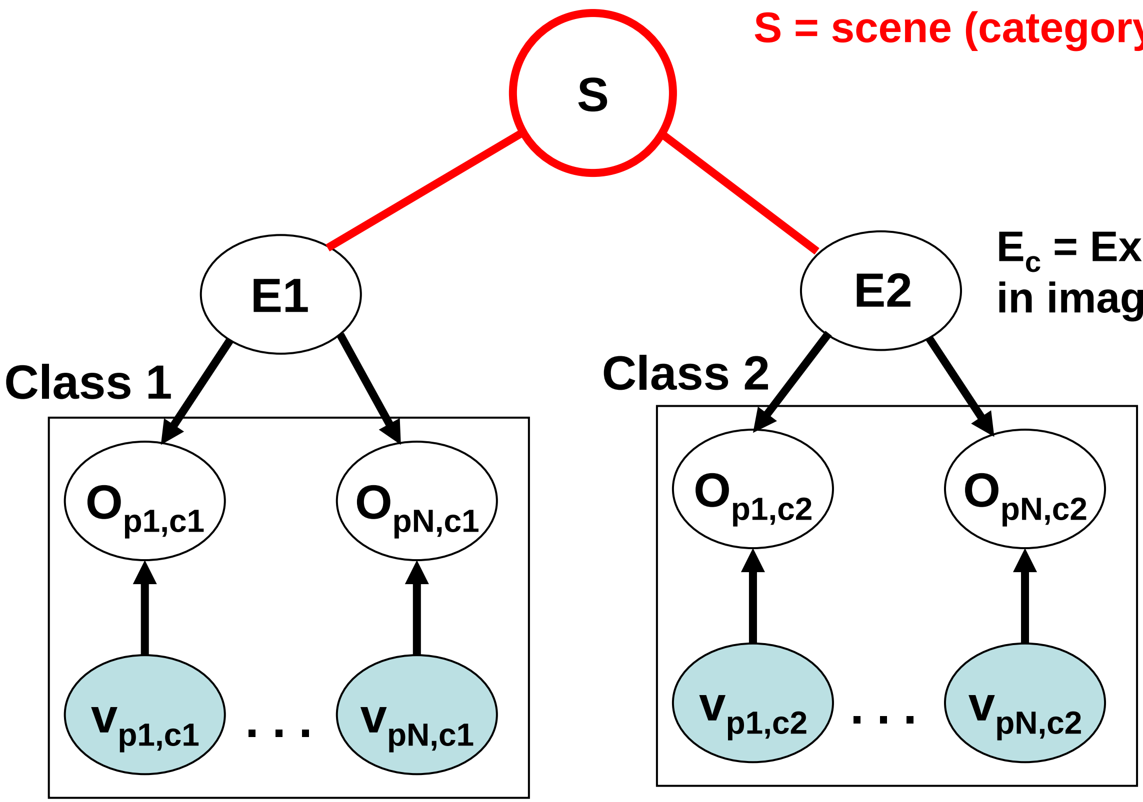
We know there is a keyboard present in this scene even if we cannot see it clearly.

We know there is no keyboard present in this scene



… even if there is one indeed.

Including scene-context for object detection



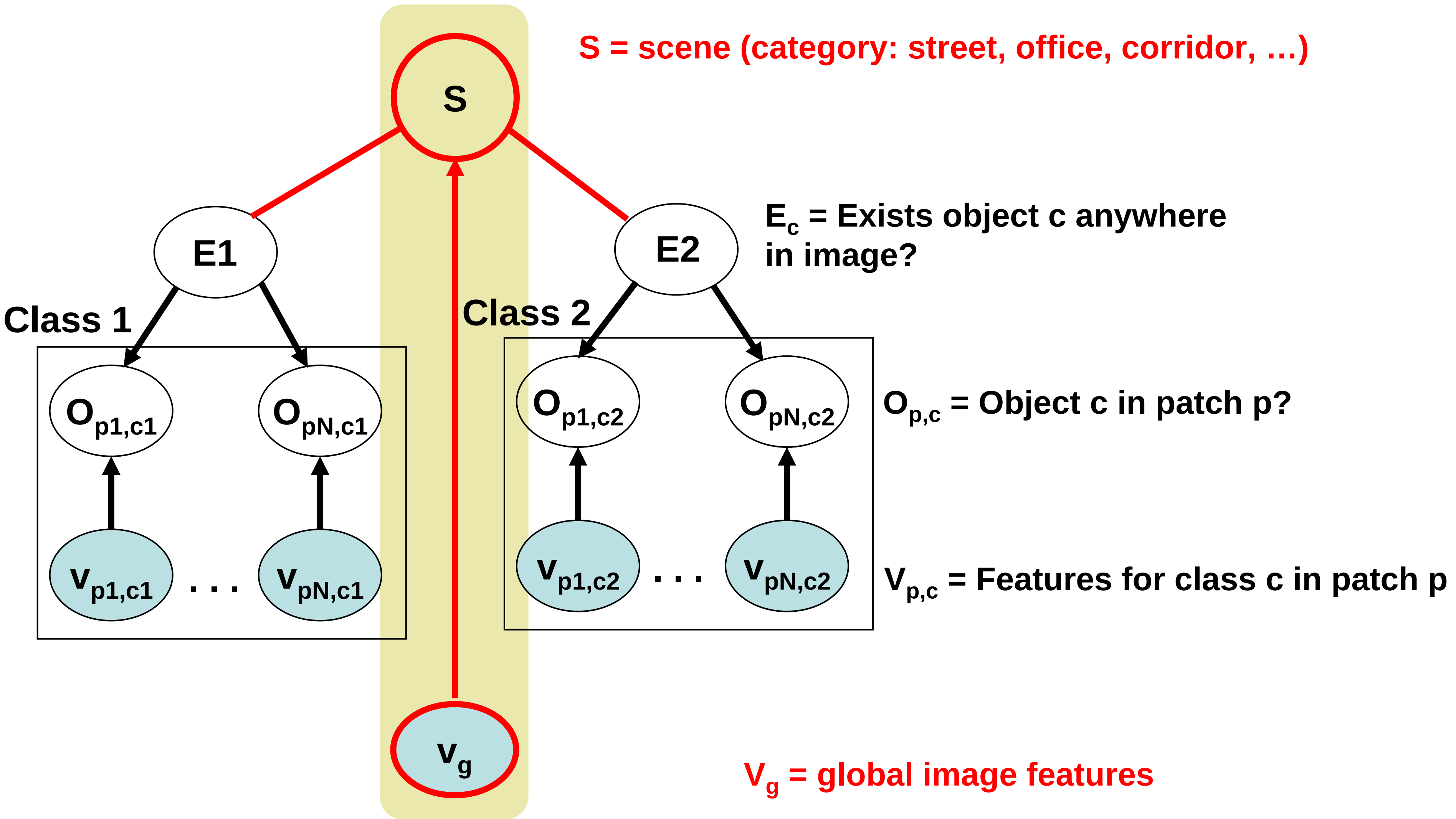
S = scene (category: street, office, corridor, …)

E_c = Exists object c anywhere in image?

$O_{p,c}$ = Object c in patch p?

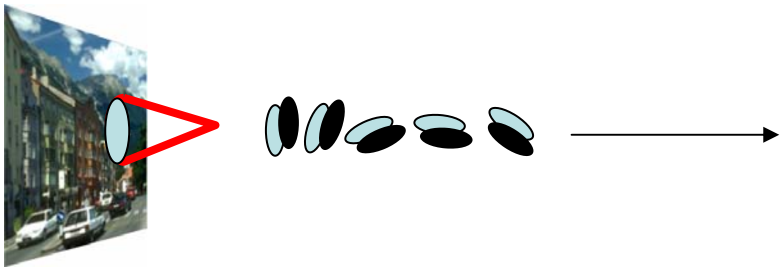
$V_{p,c}$ = Features for class c in patch p

Including scene-context for object detection



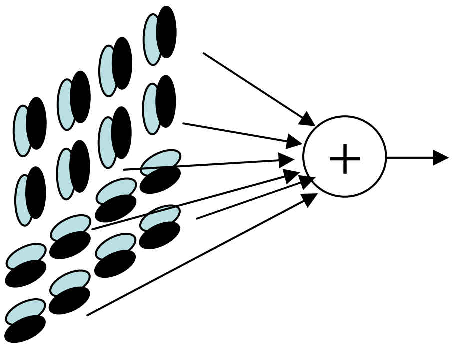
Local and Global features

A set of local features describes image properties at one particular location in the image:



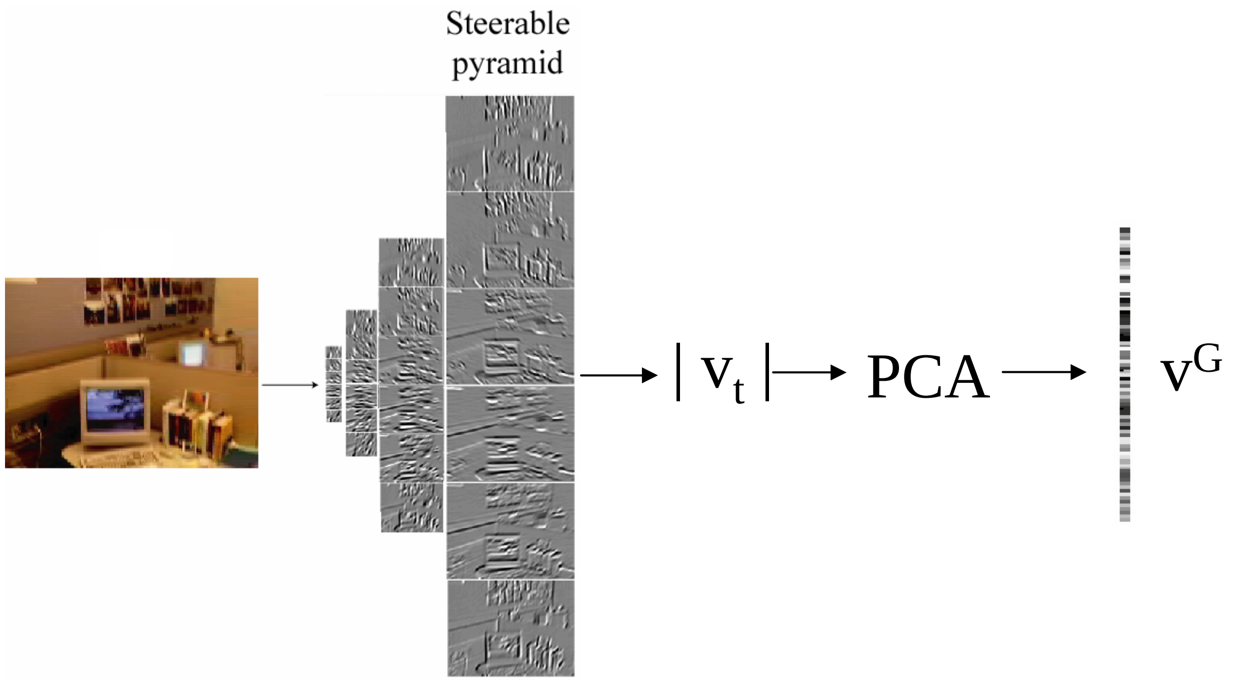
Jet of local orientations and scales

A set of global features provides information about the global image structure without encoding specific objects



This feature likes images with vertical structures at the top part and horizontal texture at the bottom part (this is a typical composition of an empty street)

Computing the global scene features



- Pipe image through steerable filter bank (here we use 6 orientations, 4 scales)
- Compute magnitude of filter outputs
- Downsample to 4 x 4 each scale/orientation
- PCA to 80 dimensions

Oliva, Torralba. IJCV 2001

Global features

64 global features

The representation preserves:
Low resolution structure
Phase is only preserved for very low spatial frequencies (2 cycles/image)

Goal

- To build a system that knows where it is
- That recognizes the main objects in the scene
- That can work on new environments
- Robust to user

Our mobile rig, version 1

Kevin Murphy

Torralba, Murphy, Freeman, Rubin, ICCV 2003; Murphy, Torralba, Freeman, NIPS 2003

Our mobile rig, version 2

Torralba, Murphy, Freeman, Rubin, ICCV 2003; Murphy, Torralba, Freeman, NIPS 2003

Training for scene recognition

Scene categorization:

office

street

corridor

3 categories

Place identification:

Office 610

Office 615

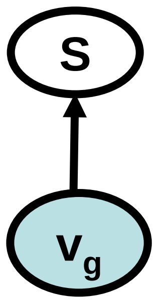
'Draper' Street

…

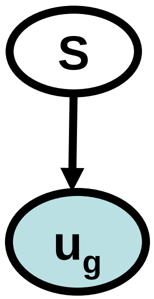
62 places

Scene classifier

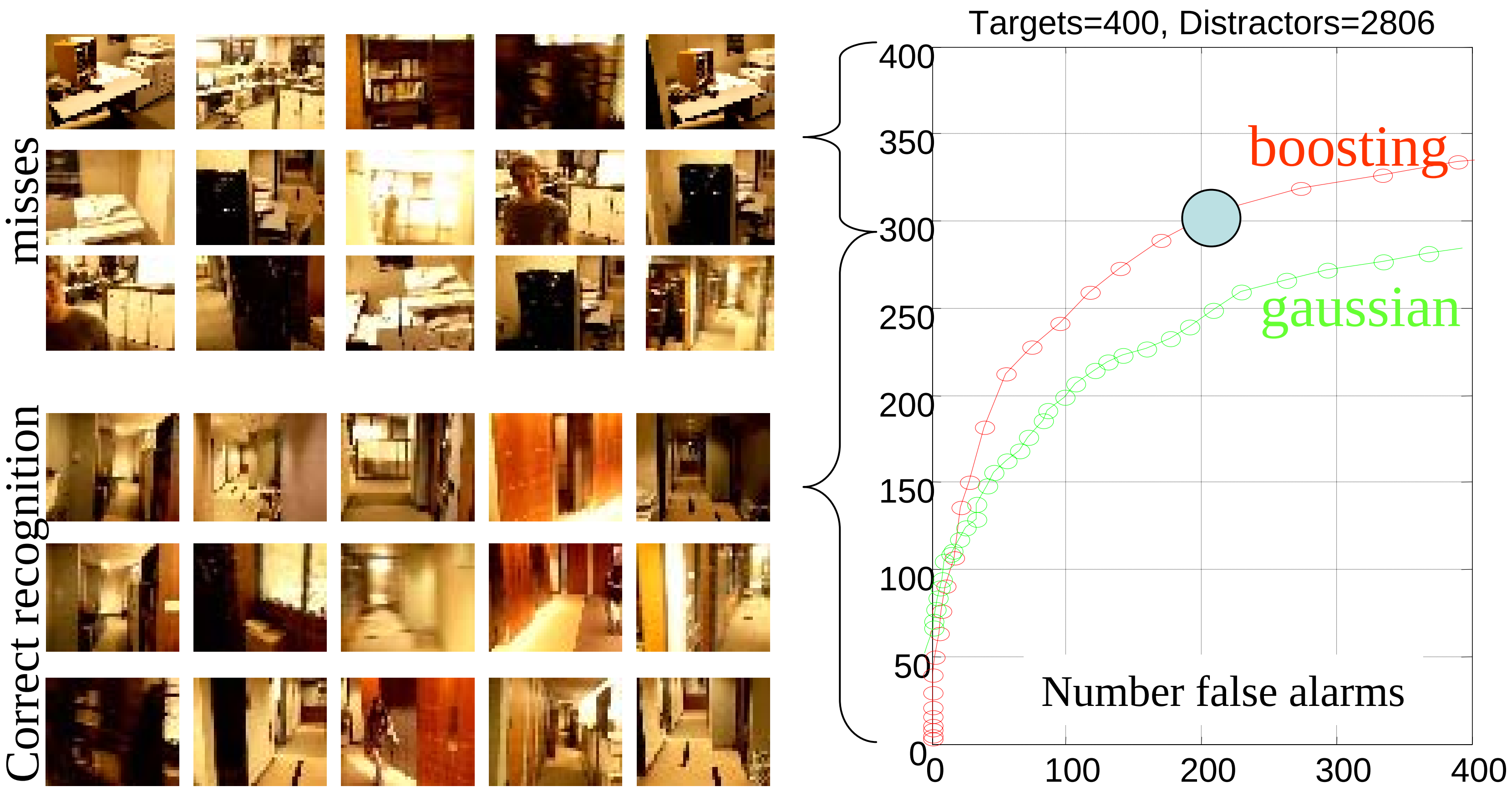
Discriminative
(boosting)



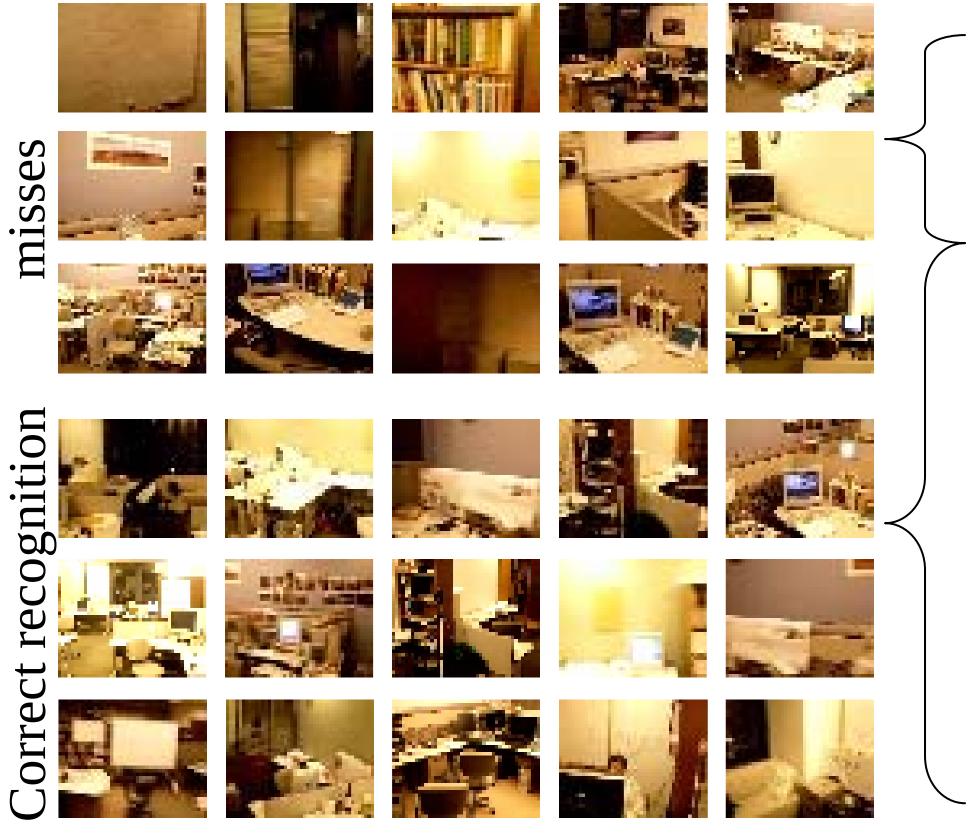
Generative
(mixture of Gaussians)

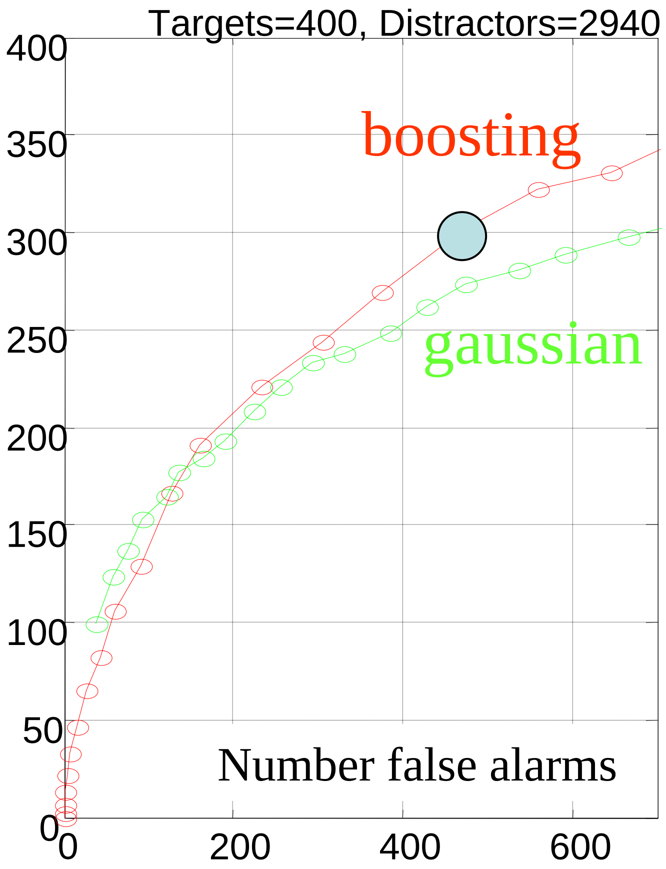


Corridor recognition



Office recognition





Temporal context helps

= ?

Temporal context helps

Place and object recognition

$p(\, o_t, \, q_t \mid v_{1:t} \,)$

Image sequence

Objects

Location

Place and object recognition

$$p(\, o_t, q_t \mid v_{1:t} \,) \;=\; p(\, o_t, q_t \mid v_{1:t}\, , v^{G}_{1:t}) \quad \alpha$$

$$p(\, o_t \mid q_t \,, v_{1:t} \,)\; P(\, q_t \mid v^{G}_{1:t} \,)$$

Location

Context features

Hidden Markov Model

$$p(\, o_t, q_t \mid v_{1:t}\,) \;\; \alpha$$

$$p(\, o_t \mid q_t\,,\, v_{1:t}\,)\; P(\, q_t \mid v^G_{1:t}\,)$$

Location

Context features

We use a HMM to estimate the location recursively:

$$P(\, q_t \mid v^G_{1:t}\,) \;\; \alpha \;\; p(\, v^G_t \mid q_t\,) \sum_{q'} P(q_t \mid q'_{t-1})\; P(\, q'_{t-1} \mid v^G_{1:t-1}\,)$$

Probability for each location

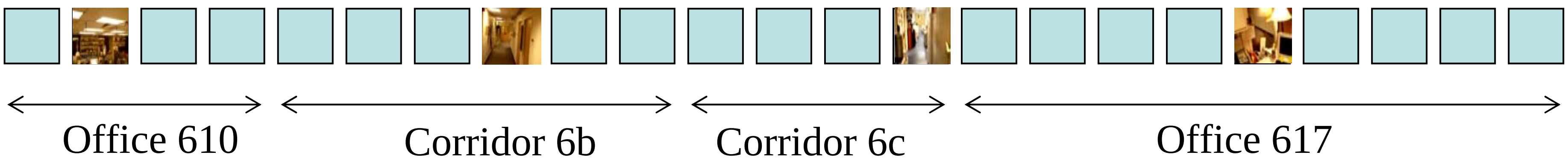
Observation likelihood

Transition matrix (encodes topology)

Previous estimation

Hidden Markov Model

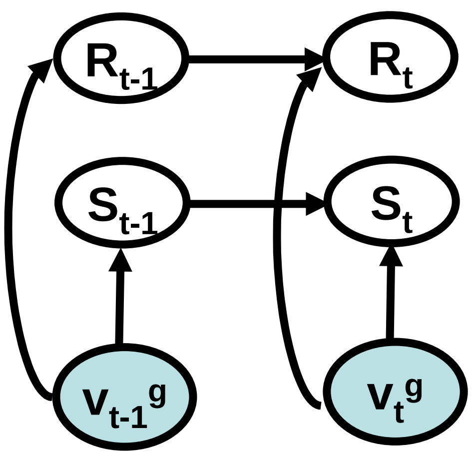
We use 17 annotated sequences for training



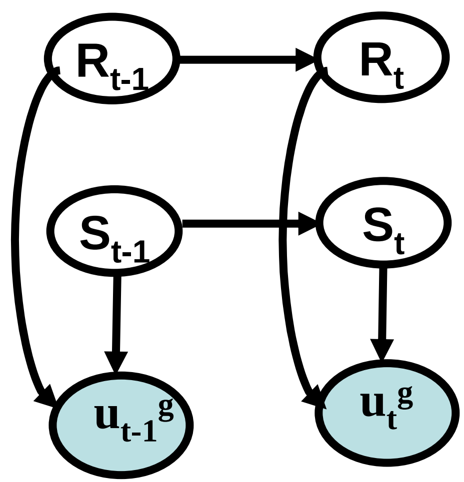
- Hidden states = location (63 values)
- Observations = v^G_t (80 dimensions)
- Transition matrix encodes topology of environment
- Observation model is a mixture of Gaussians centered on prototypes (100 views per place)

Temporal classifier

Discriminative (1D CRF)



Generative (HMM)

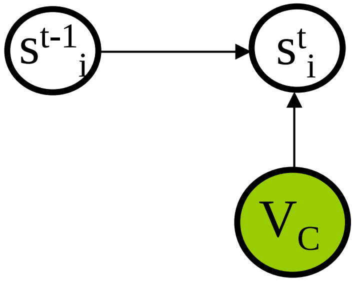


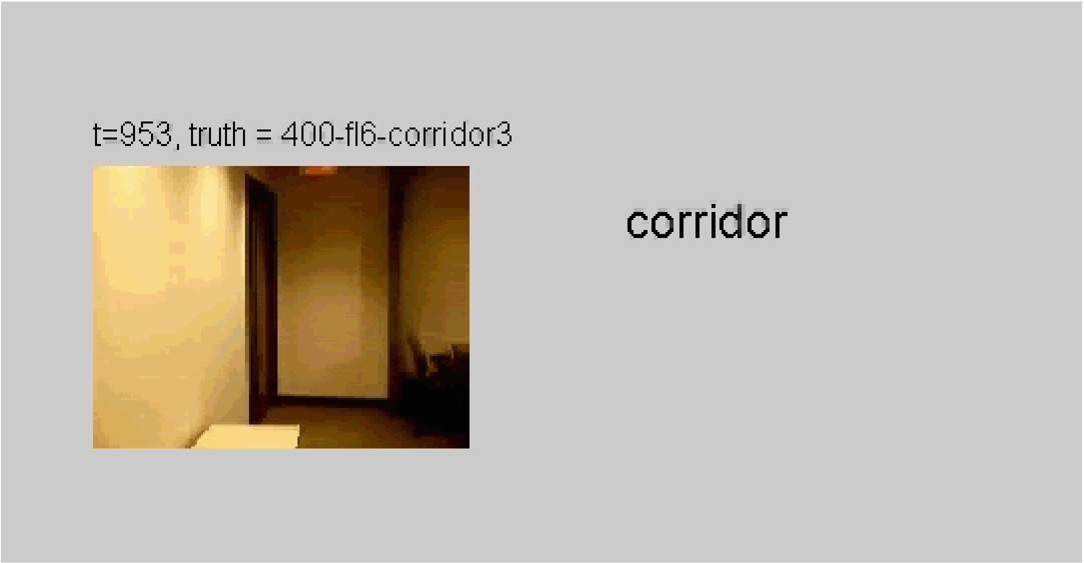
Room-name

Scene-type

Torralba, Murphy, Freeman, Rubin, ICCV 03

Place recognition demo

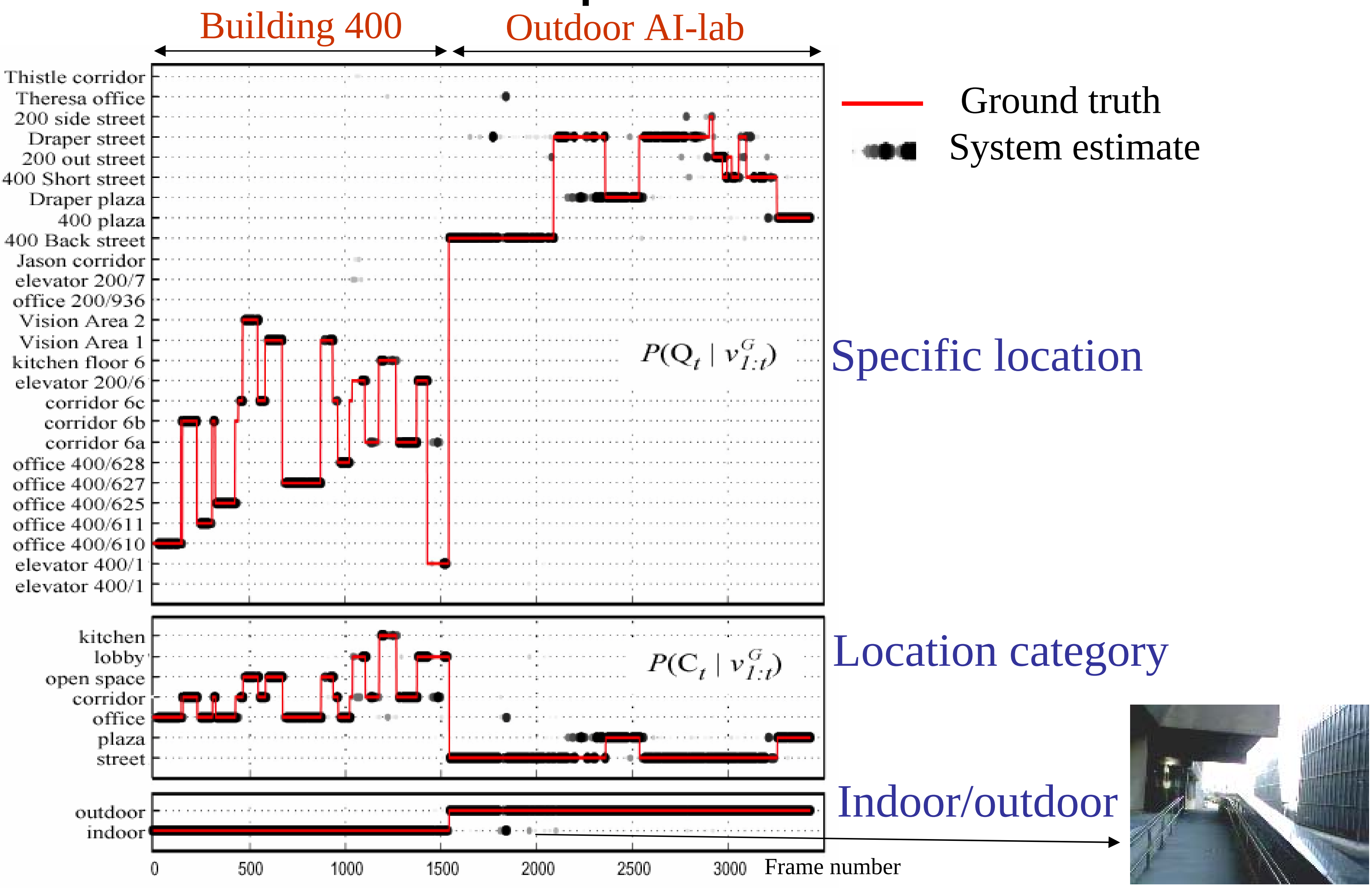




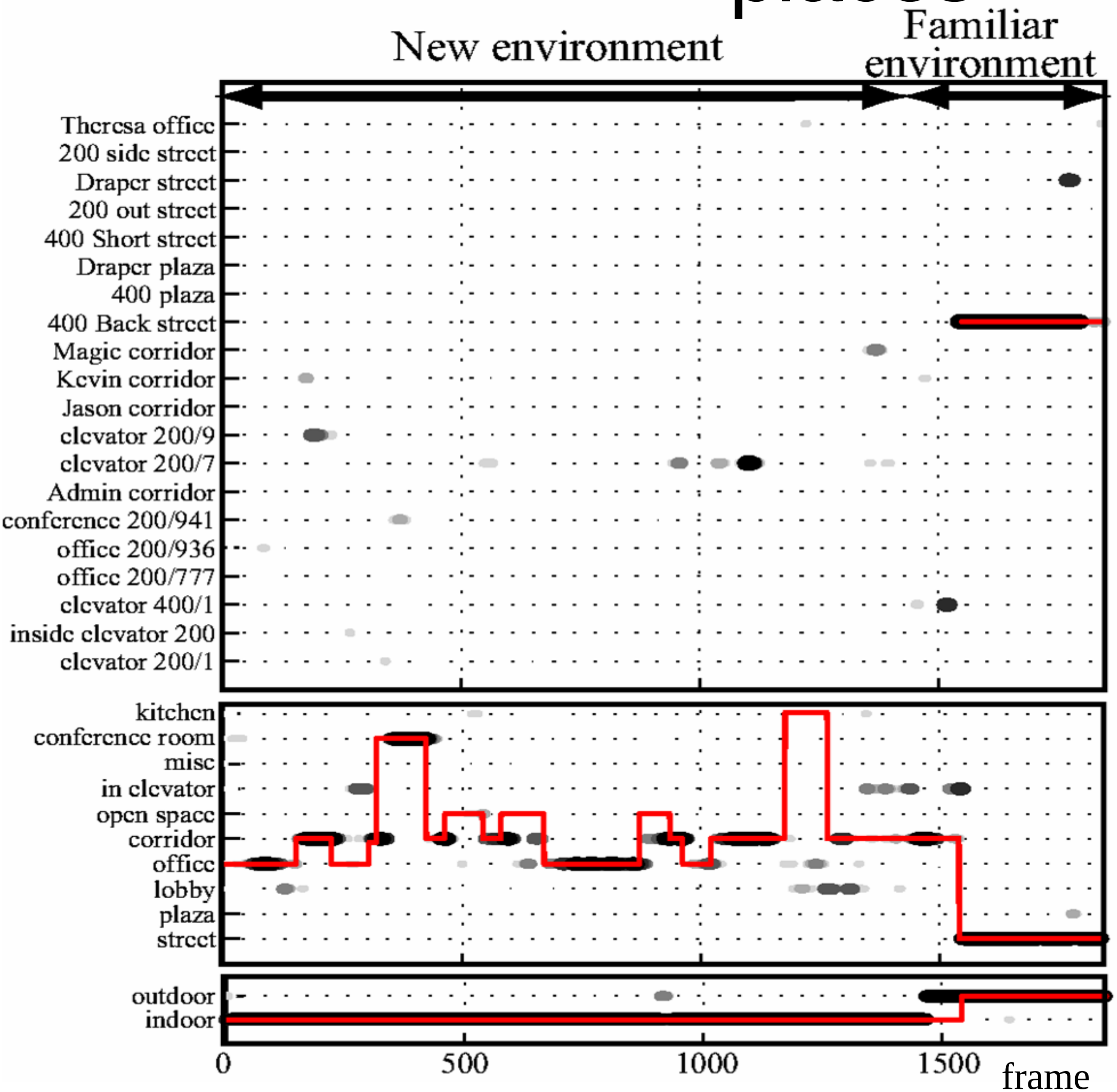
Input image (120x160)

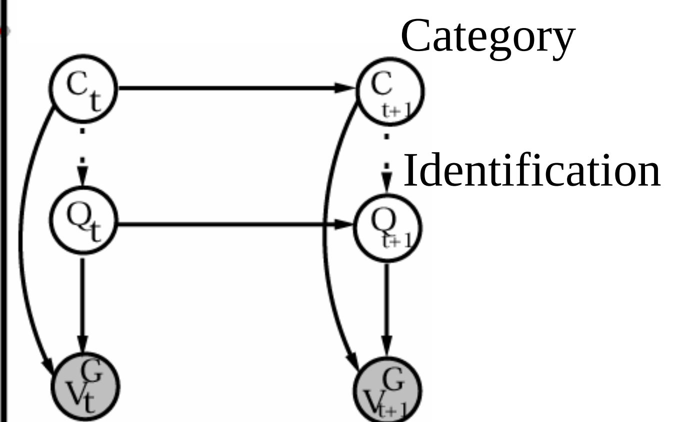
Shows the category and the identity of The place when the system is confident. Runs at 4 fps on Matlab.

Identification and categorization of known places

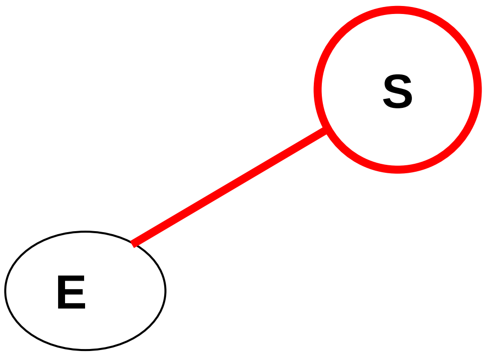


Identification and categorization of new places



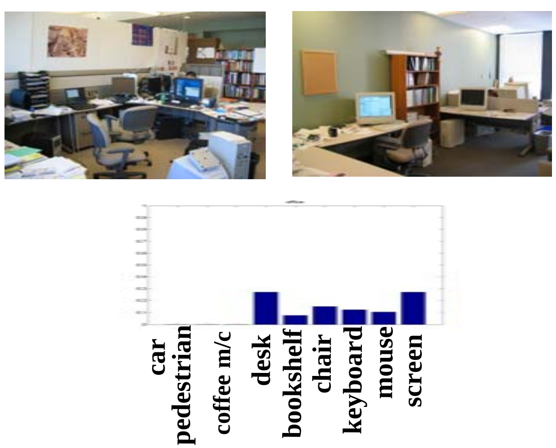


Predicting the presence of an object

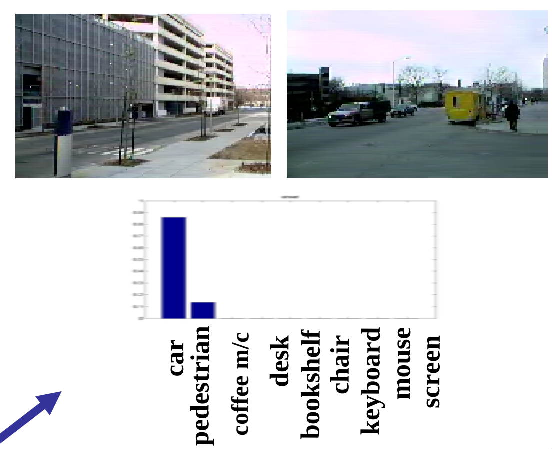


Φ(E, S) can be estimated by counting co-occurrences in labeled images

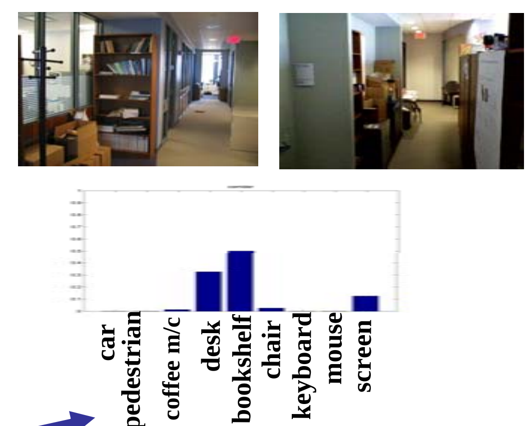
Office



Street

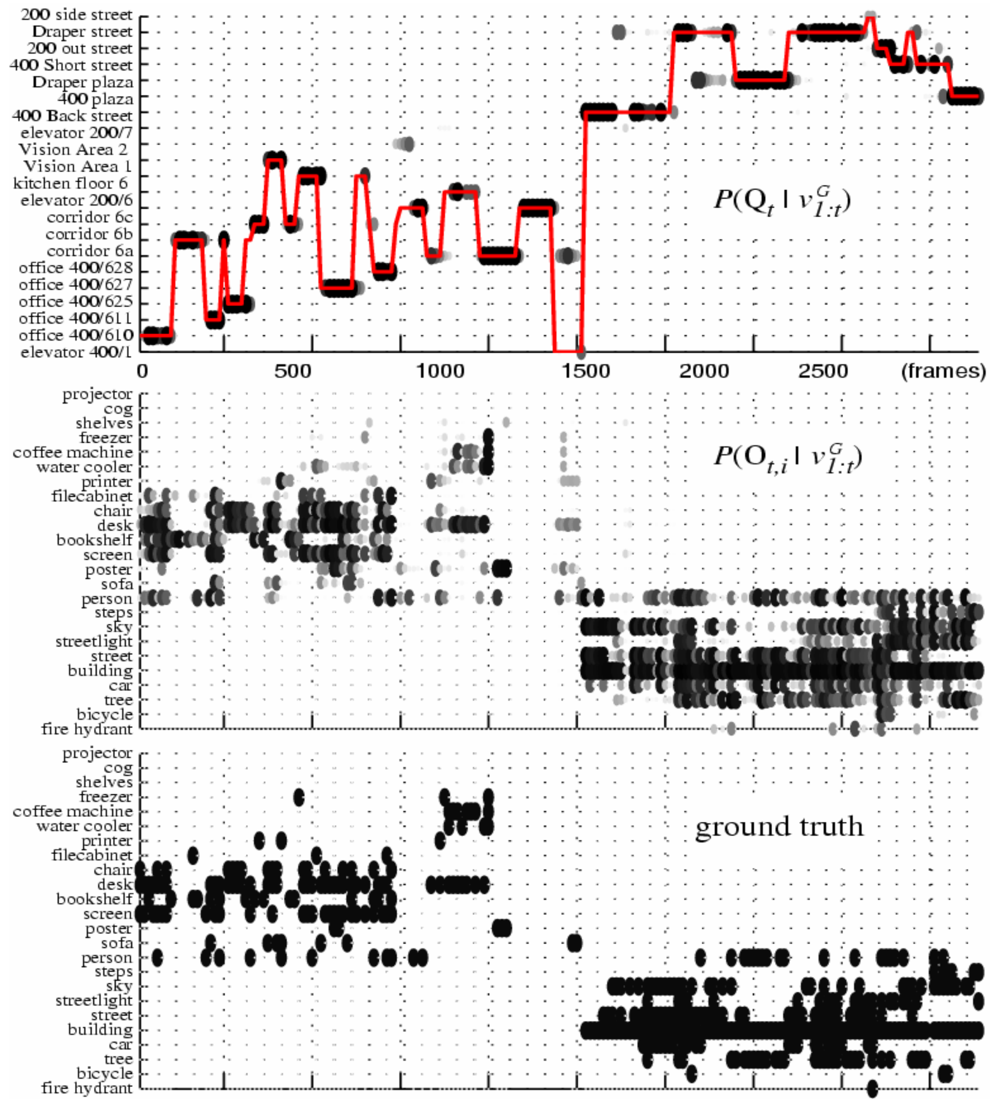


Corridor



"Cars are likely in streets, but not in offices or corridors"

Predicting the presence of an object



indoor outdoor

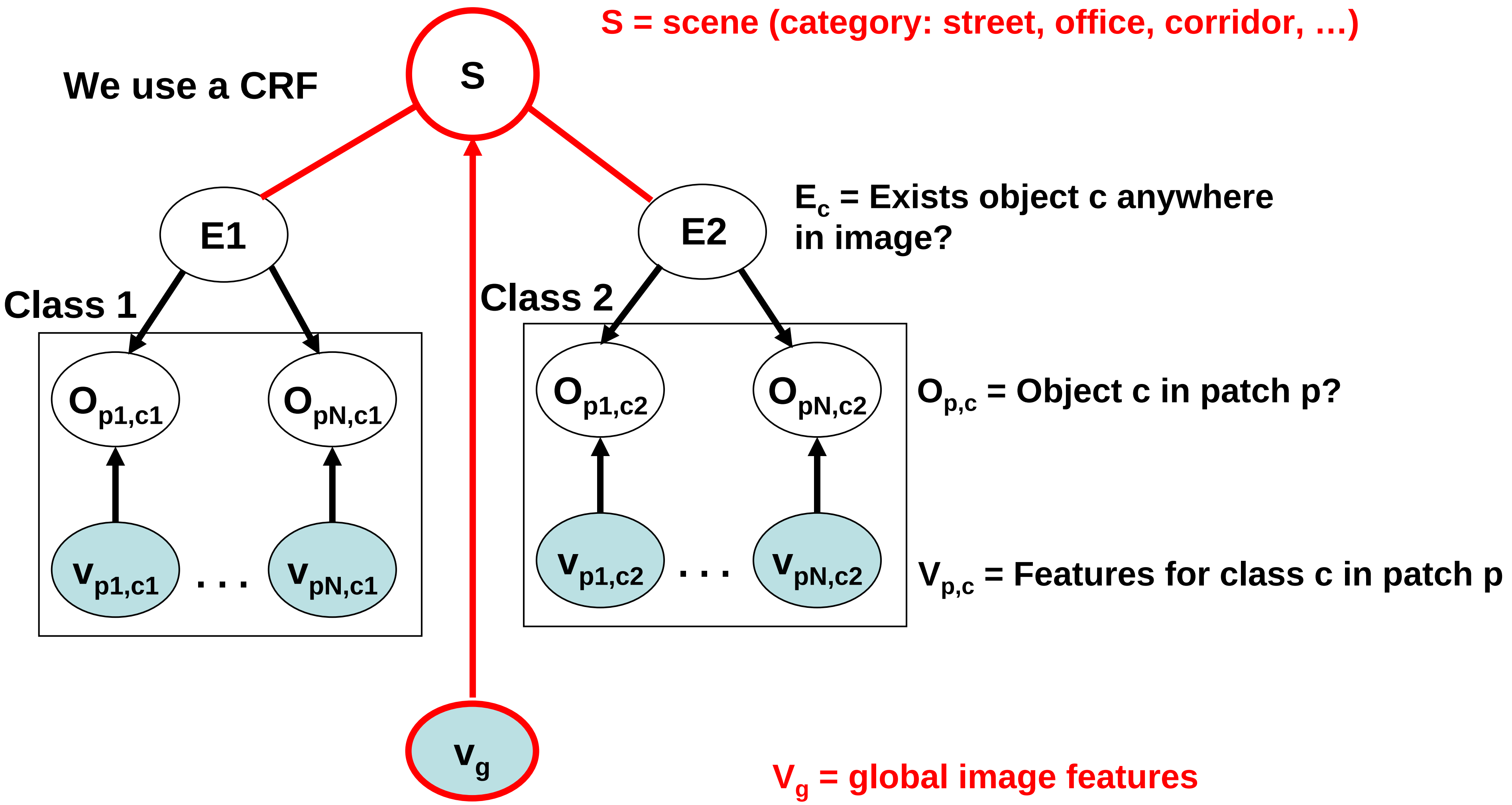
Place recognition

Object predictions

Ground truth for object presence

At each place it is not necessary to consider all possible objects for detection.

Combining scene Top-down predictions with detectors bottom-up signal



Application of object detection for image retrieval

Results using the keyboard detector alone

Low probability

High probability

Results using both the keyboard detector and the global scene features

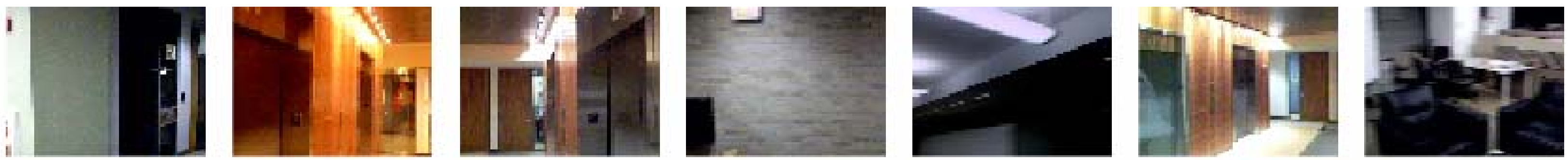
Low probability

High probability

Application of object detection for image retrieval

Results using the car detector alone

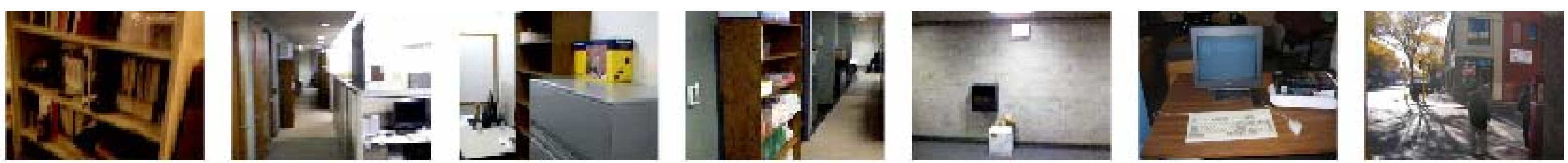
Low probability



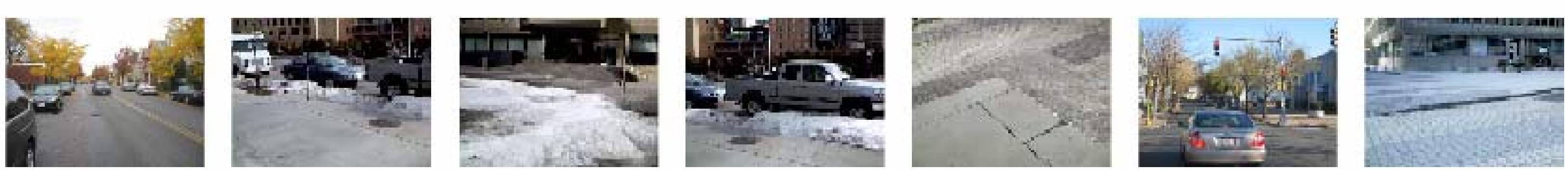
High probability

Results using both the car detector and the global scene features

Low probability



High probability

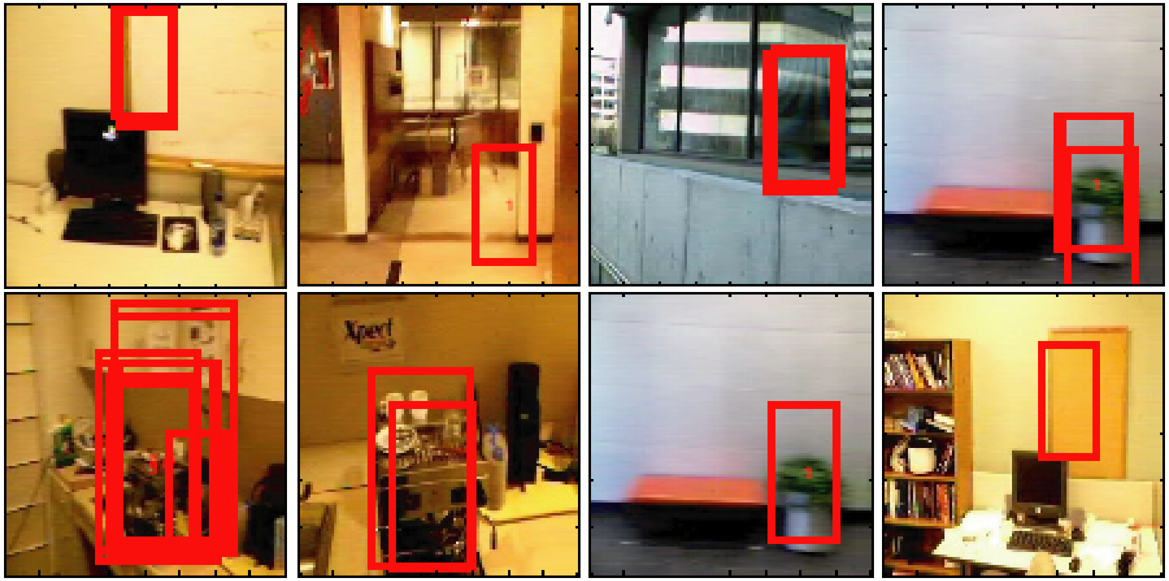


Application of object detection for image retrieval

Detecting the coffee machine:

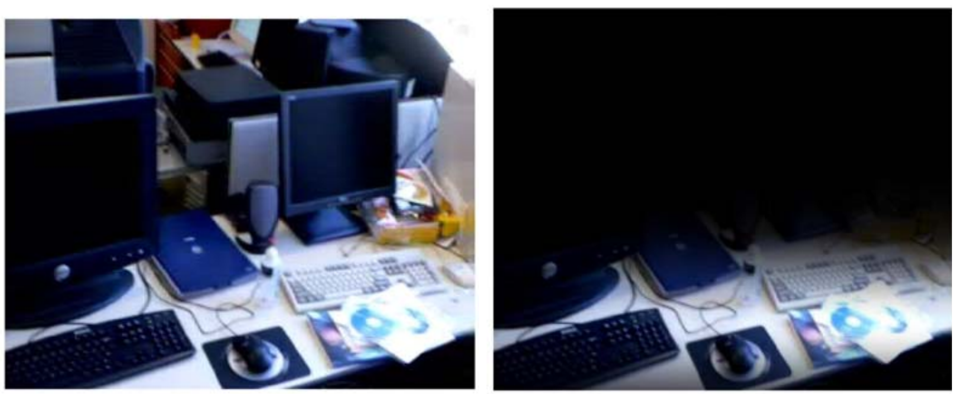
Without context

With context

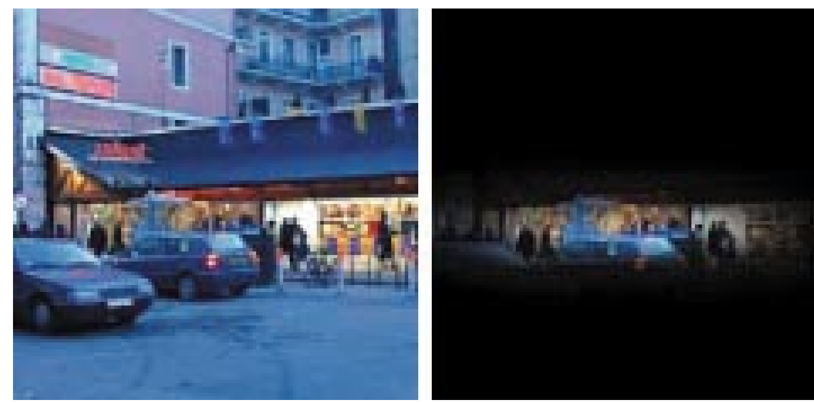


Global features can predict expected locations/scales of objects *before* running detectors

Keyboards

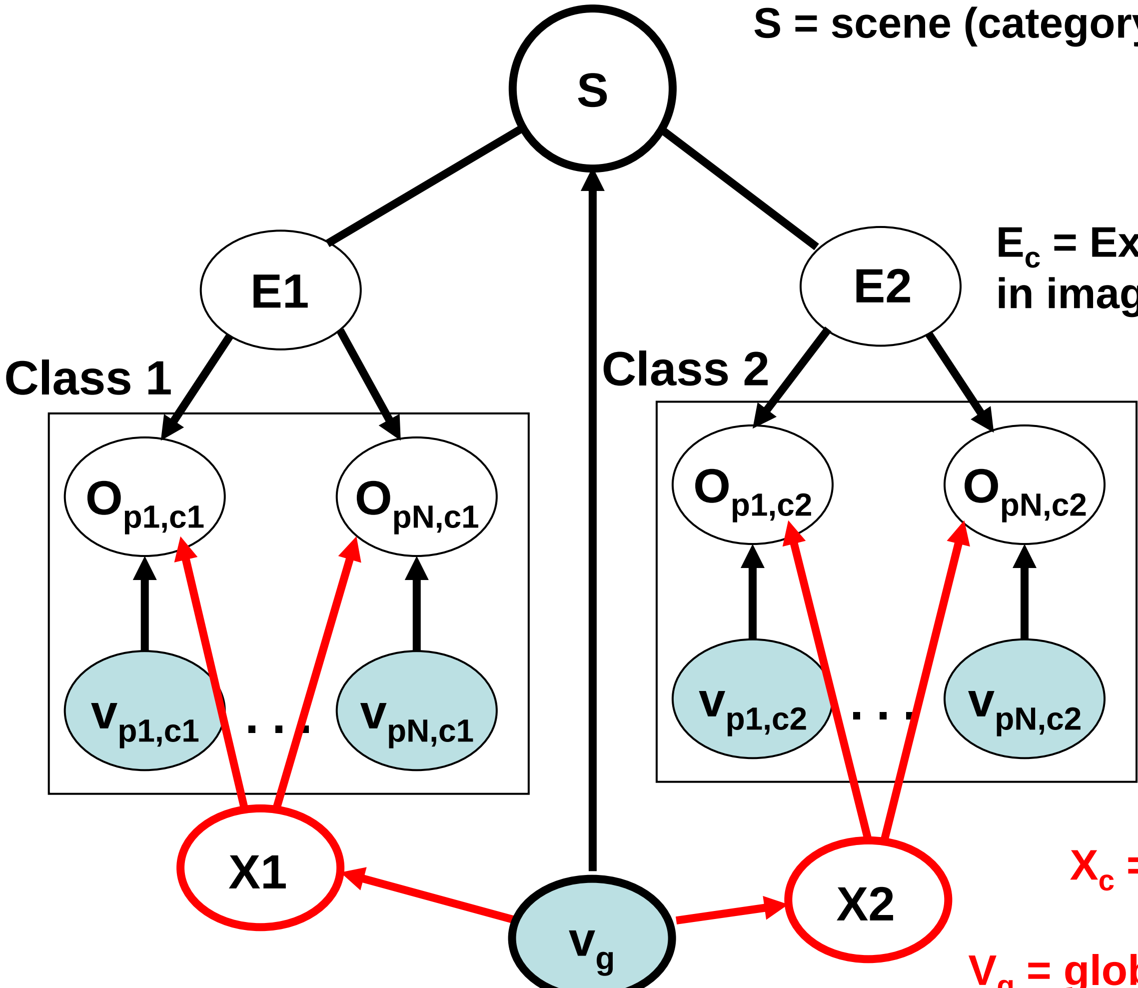


Pedestrians



There is a relationship between the aspect of the objects in a scene, and the aspect of the scene itself. For instance, the point of view of cars is correlated with the orientation of the street. But also, the location of the ground in the scene is correlated with the location of the objects in the scene.

Global scene features predicts location



S = scene (category: street, office, corridor, …)

E_c = Exists object c anywhere in image?

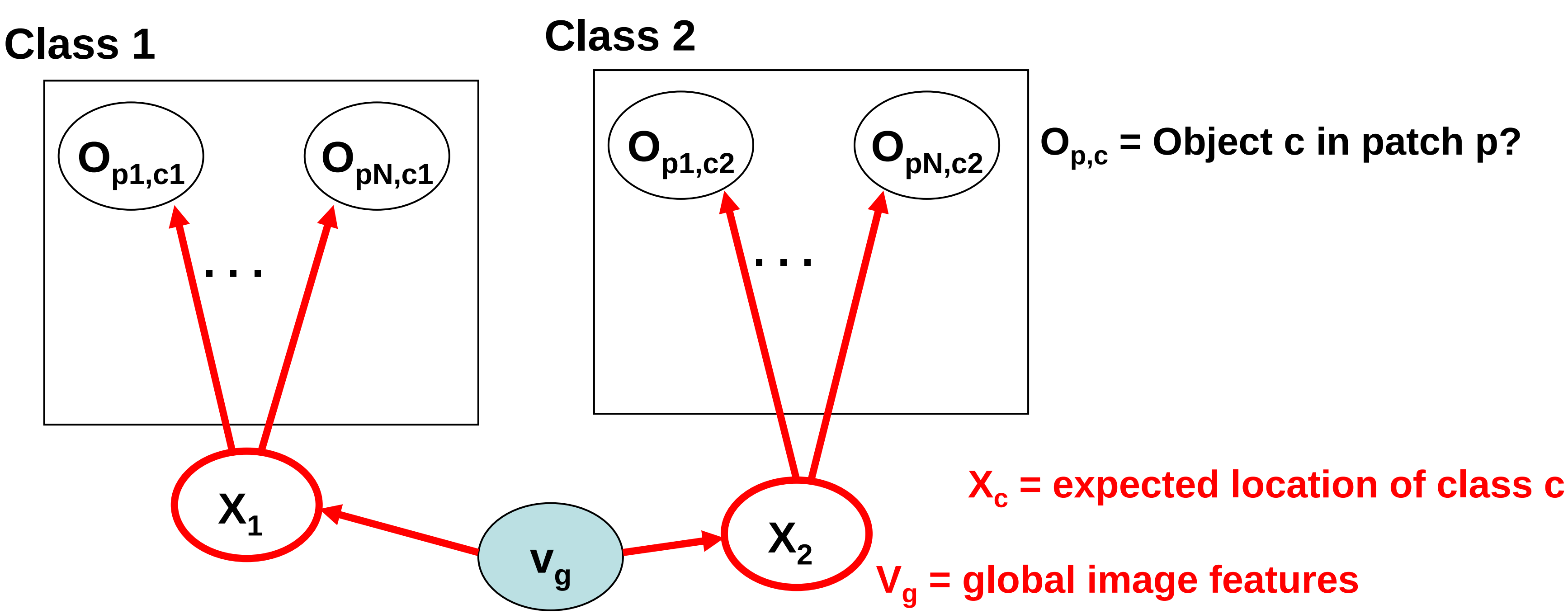
$O_{p,c}$ = Object c in patch p?

$V_{p,c}$ = Features for class c in patch p

X_c = expected location of class c

V_g = global image features

Global scene features predicts location



Global scene features predicts location

Training set (cars)

$\rightarrow \{V_g^1, X^1\}$

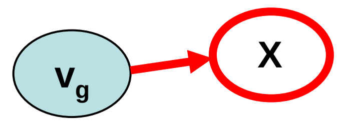
$\rightarrow \{V_g^2, X^2\}$

$\rightarrow \{V_g^3, X^3\}$

$\rightarrow \{V_g^4, X^4\}$

⋮

1) We learn the mapping between image global features and object location as a regression problem:



$$X = \sum h_m(Vg)$$

Minimize $E[(x_{true} - x)^2]$

We use boosting for regression.
h_m are regression stumps.

(We do the regression for the horizontal and vertical Components, and for scale)

Global scene features predicts location

Training set (cars)

→ $\{V_g^1, X^1\}$

→ $\{V_g^2, X^2\}$

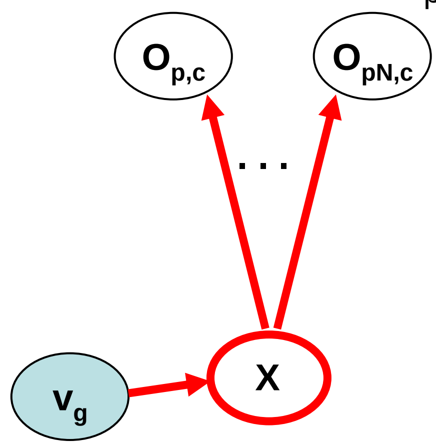
→ $\{V_g^3, X^3\}$

→ $\{V_g^4, X^4\}$

⋮

2) We fit a logistic function to compute the probability of object presence in a patch p given the expected location x:

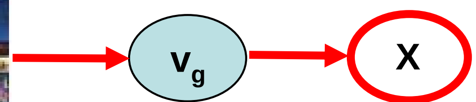
$$P(O_{p,c} \mid x) = \sigma\,(w^T\,[1 \;\; ||x_p - x||^2\,])$$



$$x = \sum h_m(Vg)$$

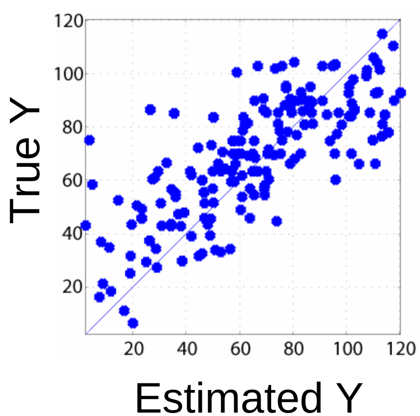
Global scene features predicts location

Given a new scene, we can predict the most expected location of an object based on the global features of the image

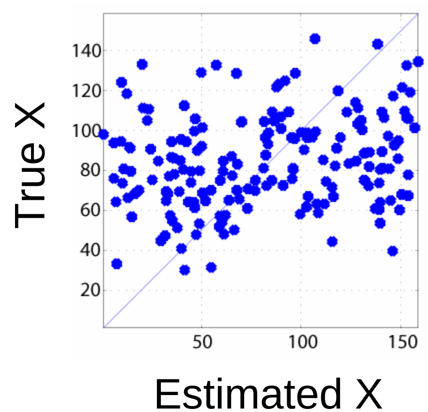


$$X = \sum h_m(Vg)$$

Results for predicting the vertical location of cars

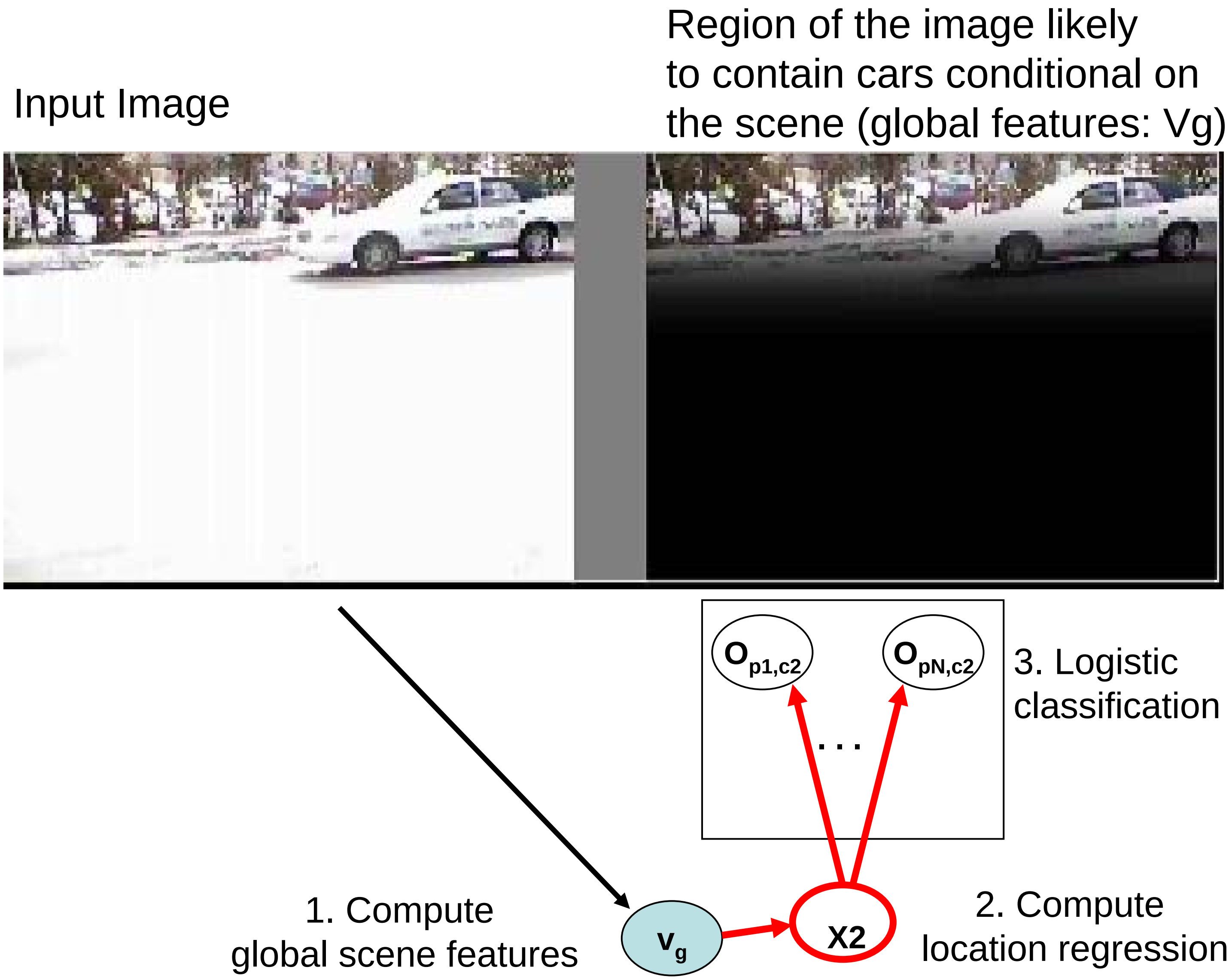


Results for predicting the horizontal location of cars

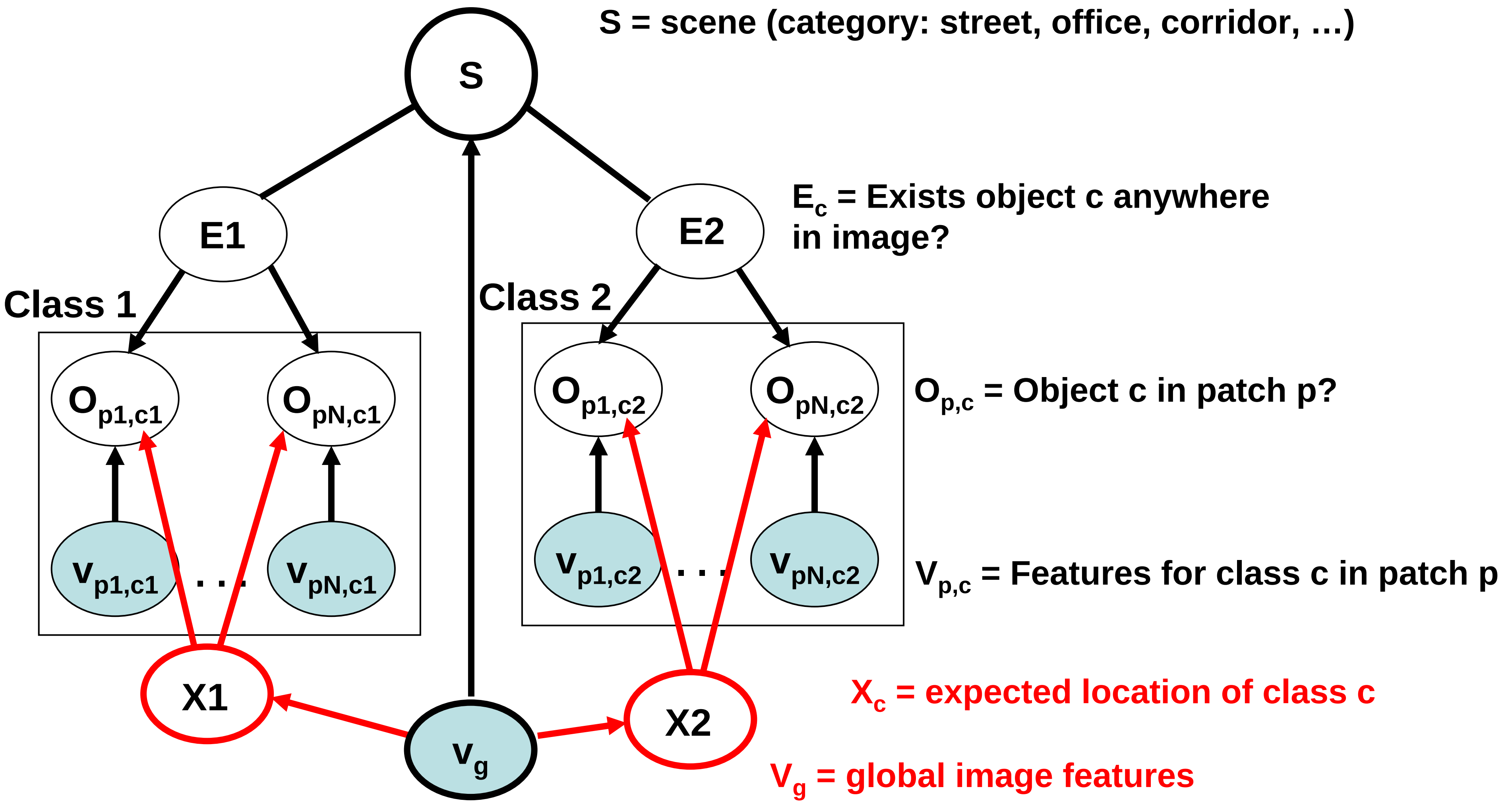


Scenes are arranged on horizontal layers.
We can predict the vertical component (ground level) but the horizontal component is poorly constrained by the global scene.

Global scene features predicts location



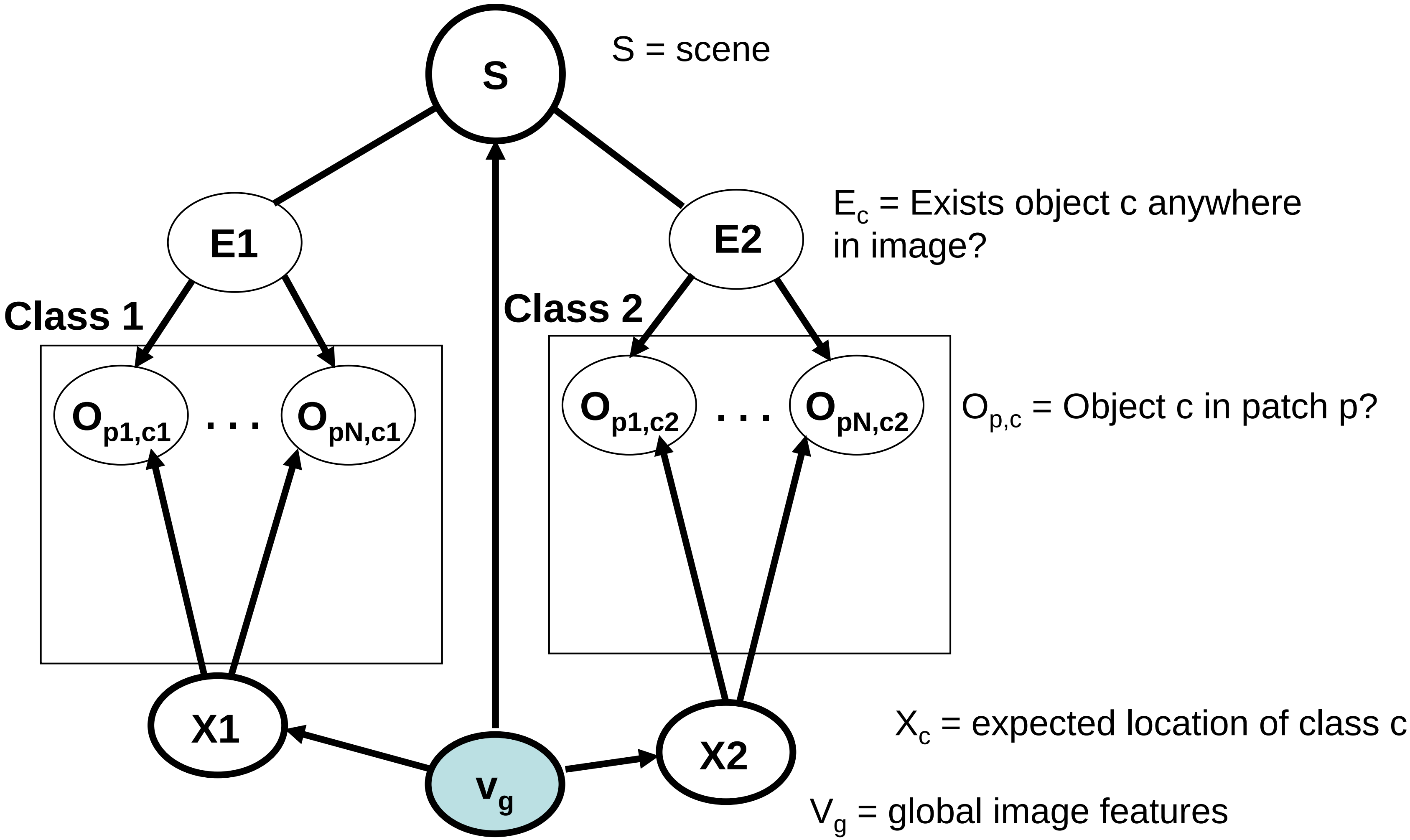
Full system



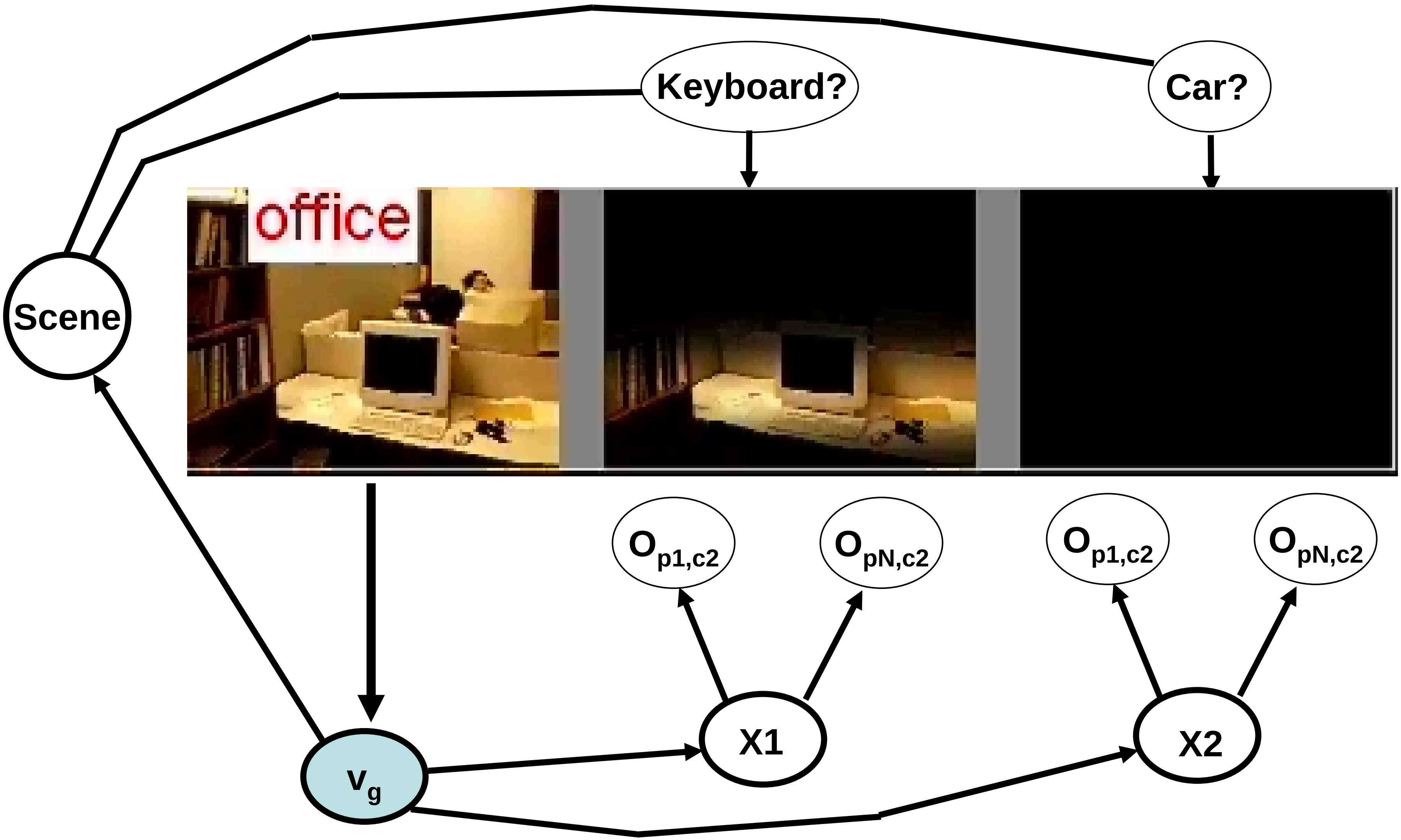
S
E1
E2
Class 1
Class 2
$O_{p1,c1}$
$O_{pN,c1}$
$O_{p1,c2}$
$O_{pN,c2}$
$V_{p1,c1}$
$V_{pN,c1}$
$V_{p1,c2}$
$V_{pN,c2}$
. . .
. . .
X1
X2
V_g
S = scene (category: street, office, corridor, …)
E_c = Exists object c anywhere in image?
$O_{p,c}$ = Object c in patch p?
$V_{p,c}$ = Features for class c in patch p
X_c = expected location of class c
V_g = global image features

The strength of context

Lets see how far can we get in object detection and localization without using detectors at all.



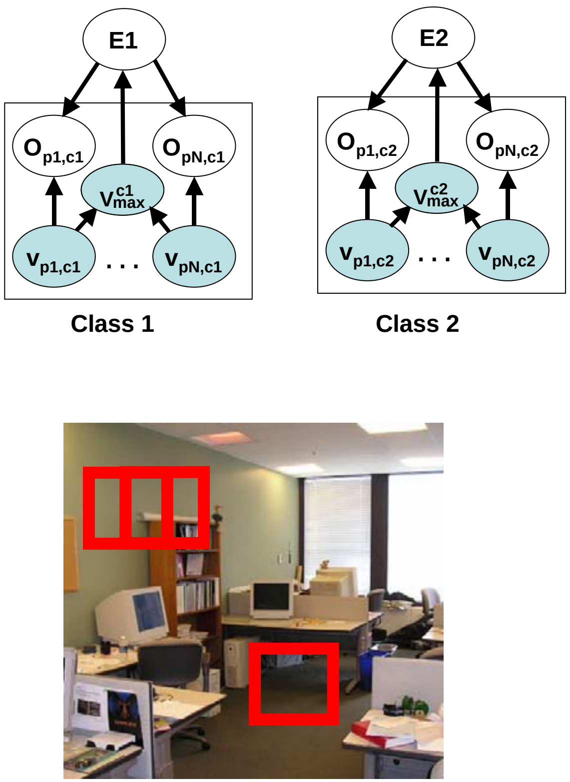
The strength of context



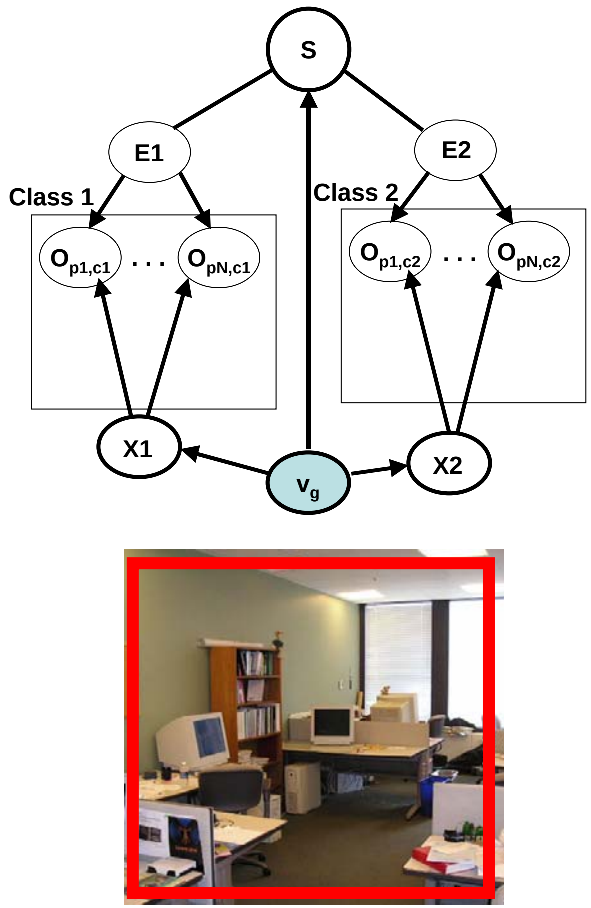
No temporal integration. Every frame is processed independently from the previous one.

The two sources of information and the final system

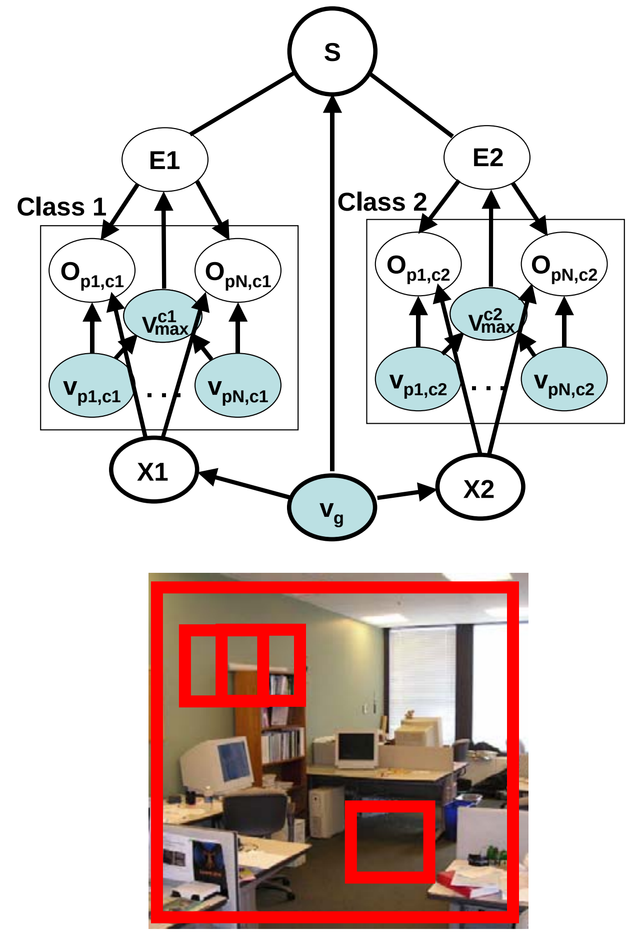
Local scene analysis



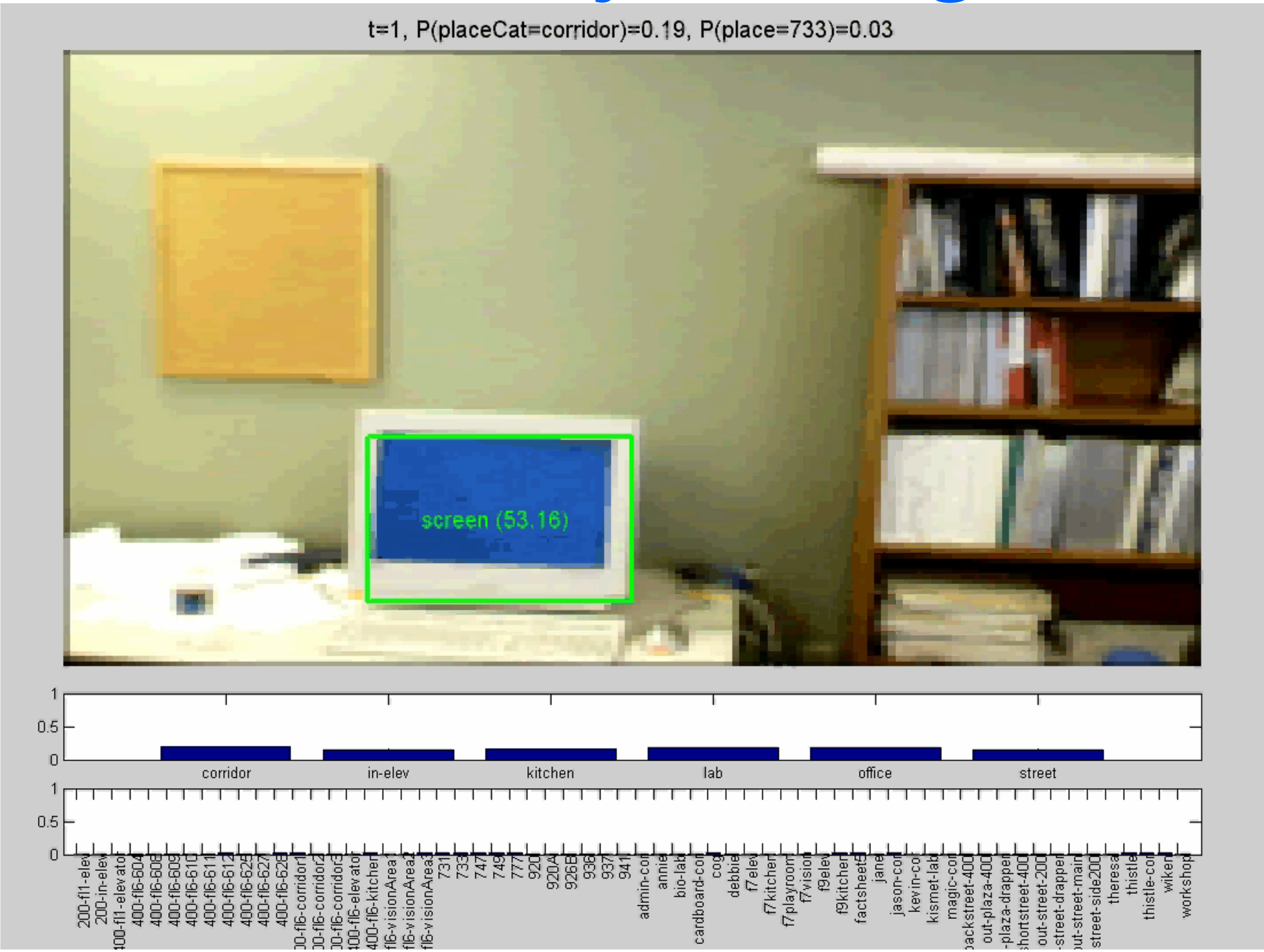
Global scene analysis



Integration of global and local features



Context-based vision system for place and object recognition



Object detection

Scene categorization

Place recognition

Learning joint object models

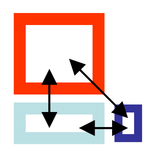
Multiclass object detection

- We want to recognize many object classes with efficient algorithms:

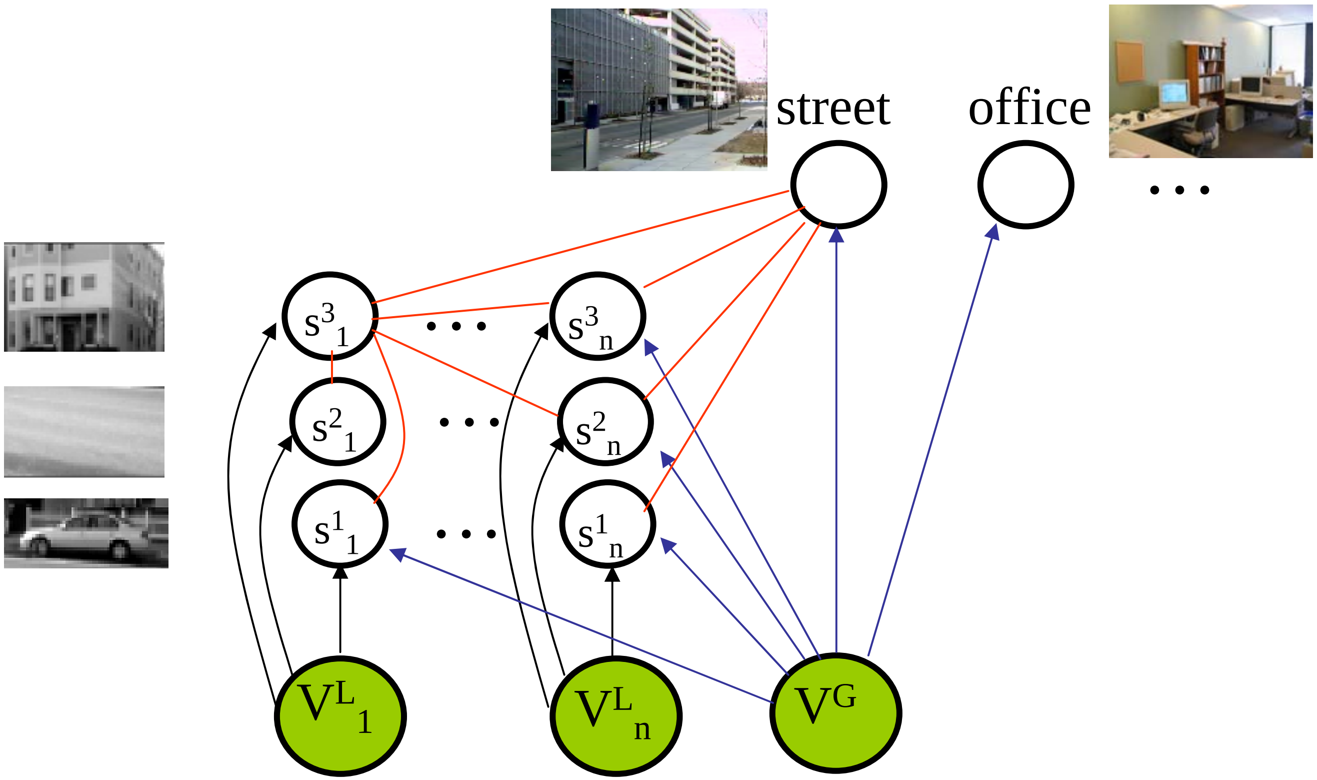
(Torralba, Murphy, Freeman, CVPR 04)

- We want to use contextual relationships between objects

(Torralba, Murphy, Freeman, NIPS 04)



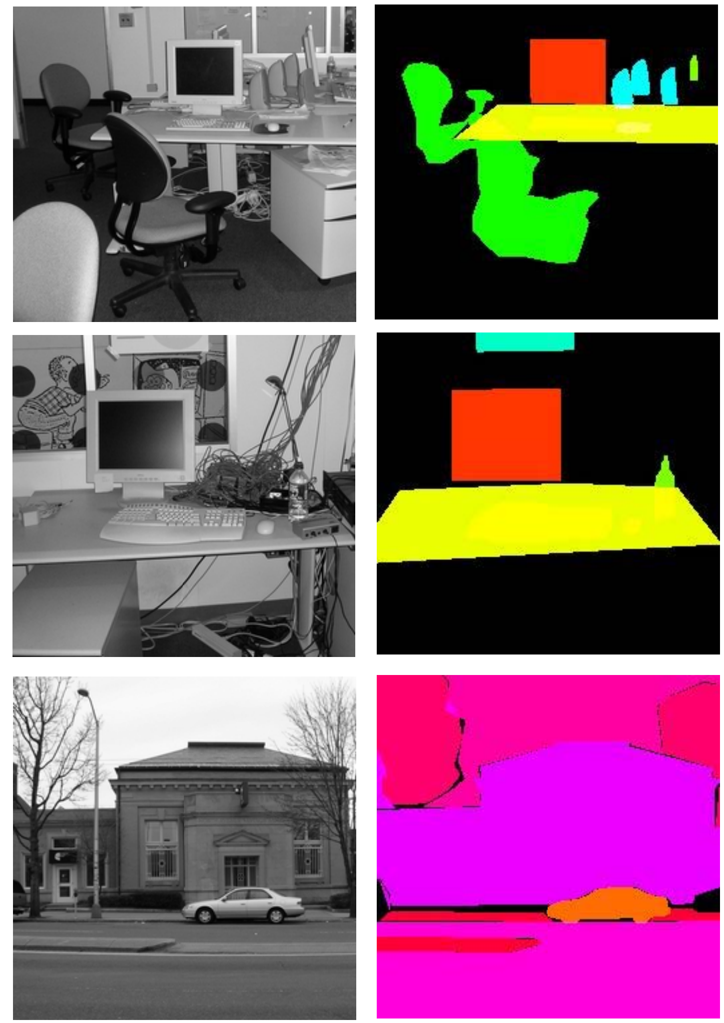
A more complete model of context



Torralba, Murphy, Freeman, NIPS 04

Image database

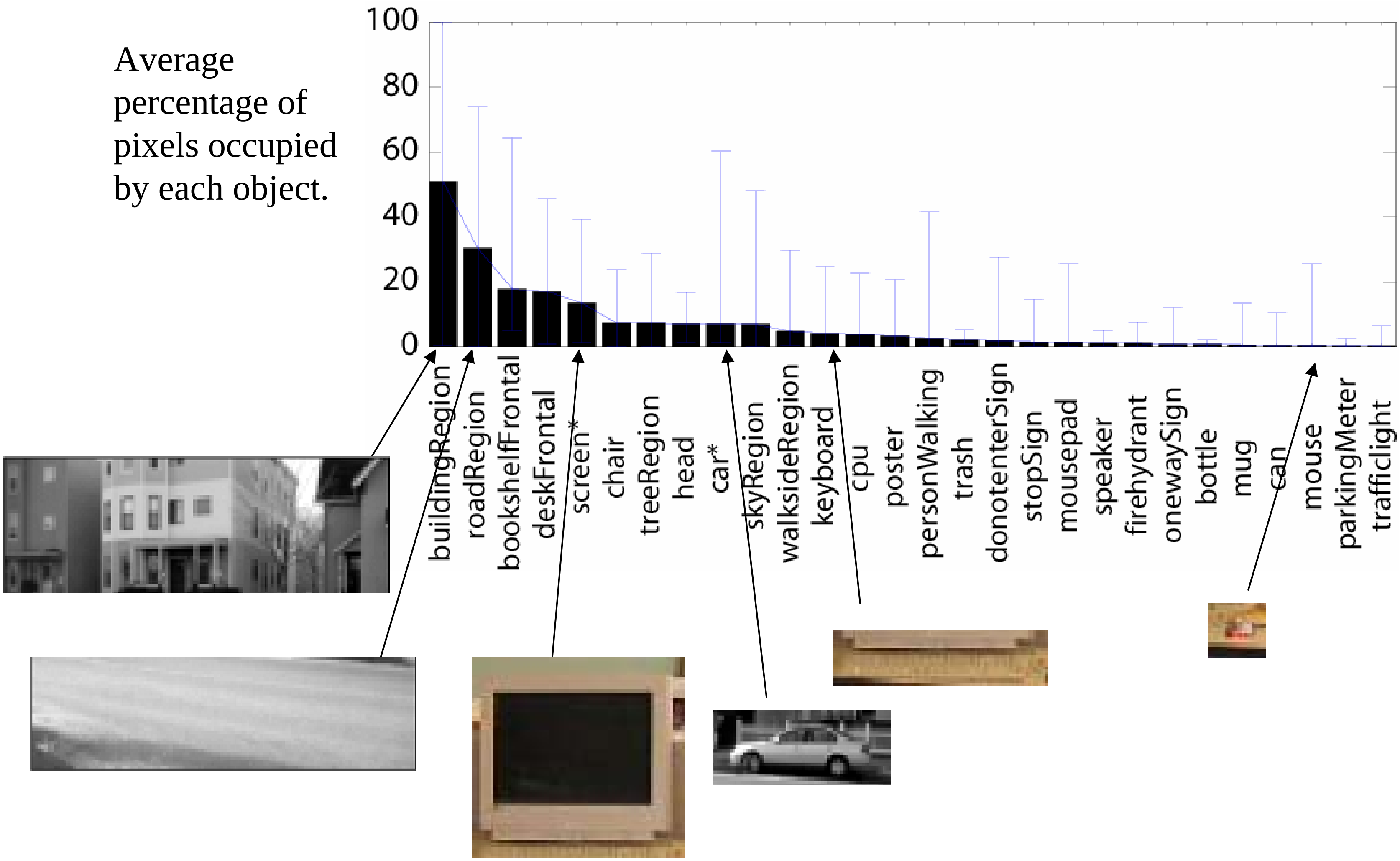
- ~2500 hand labeled images with segmentations
- ~30 objects and stuff
- Indoor and outdoor
- Sets of images are separated by locations and camera (digital/webcam)



Detecting difficult objects

There is a whole range of difficulties for the task of object detection:

Average percentage of pixels occupied by each object.



Detecting difficult objects

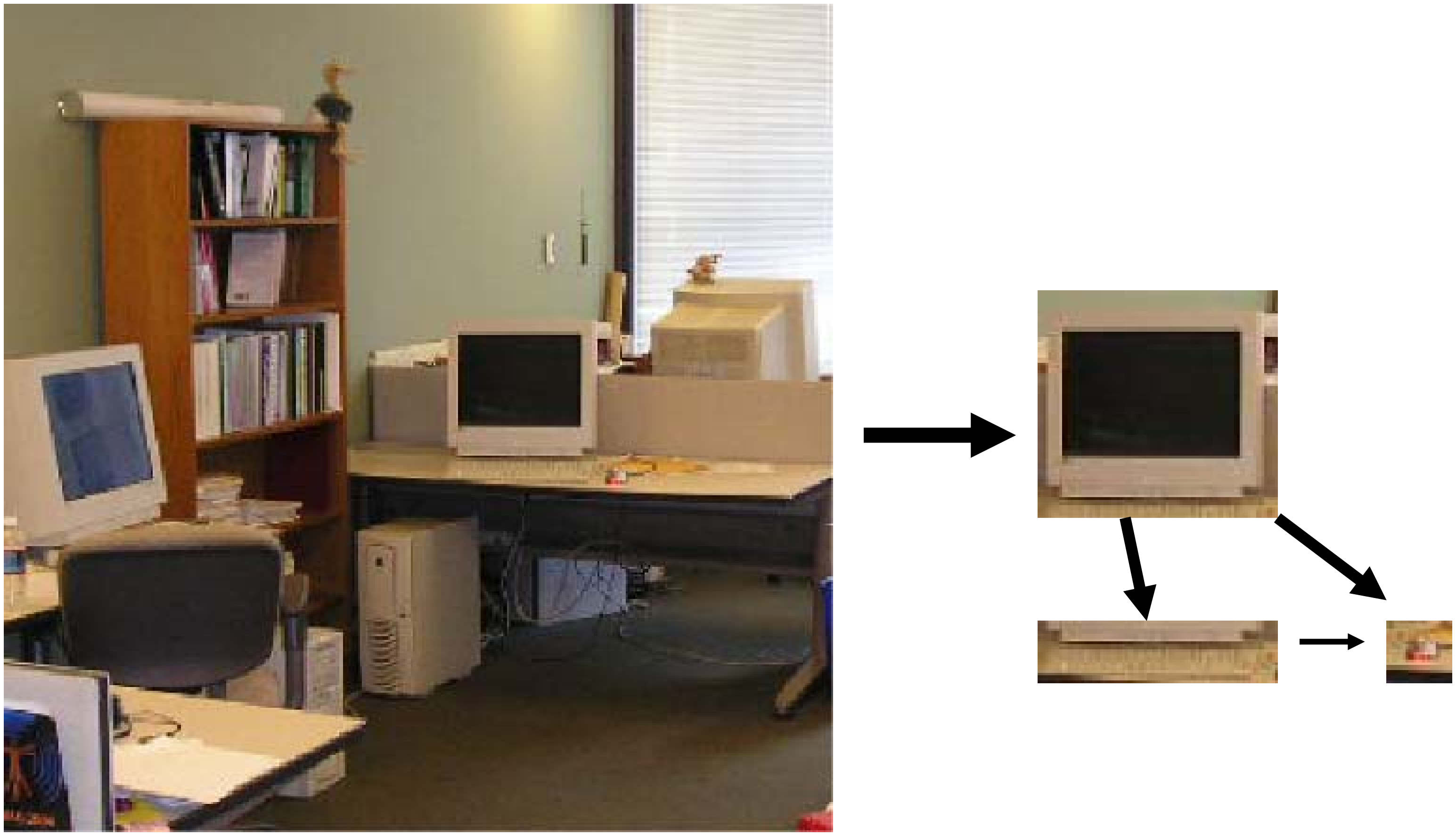
Start recognizing the scene

→ Office → Maybe there is a mouse

Detecting difficult objects

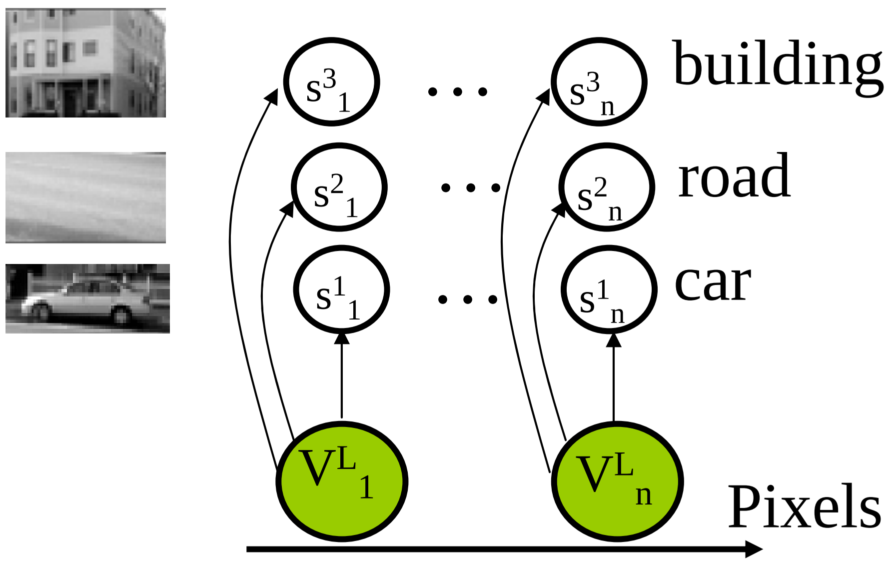
Detect first simple objects (reliable detectors) that provide strong contextual constraints to the target (screen -> keyboard -> mouse)

Segmenting difficult objects



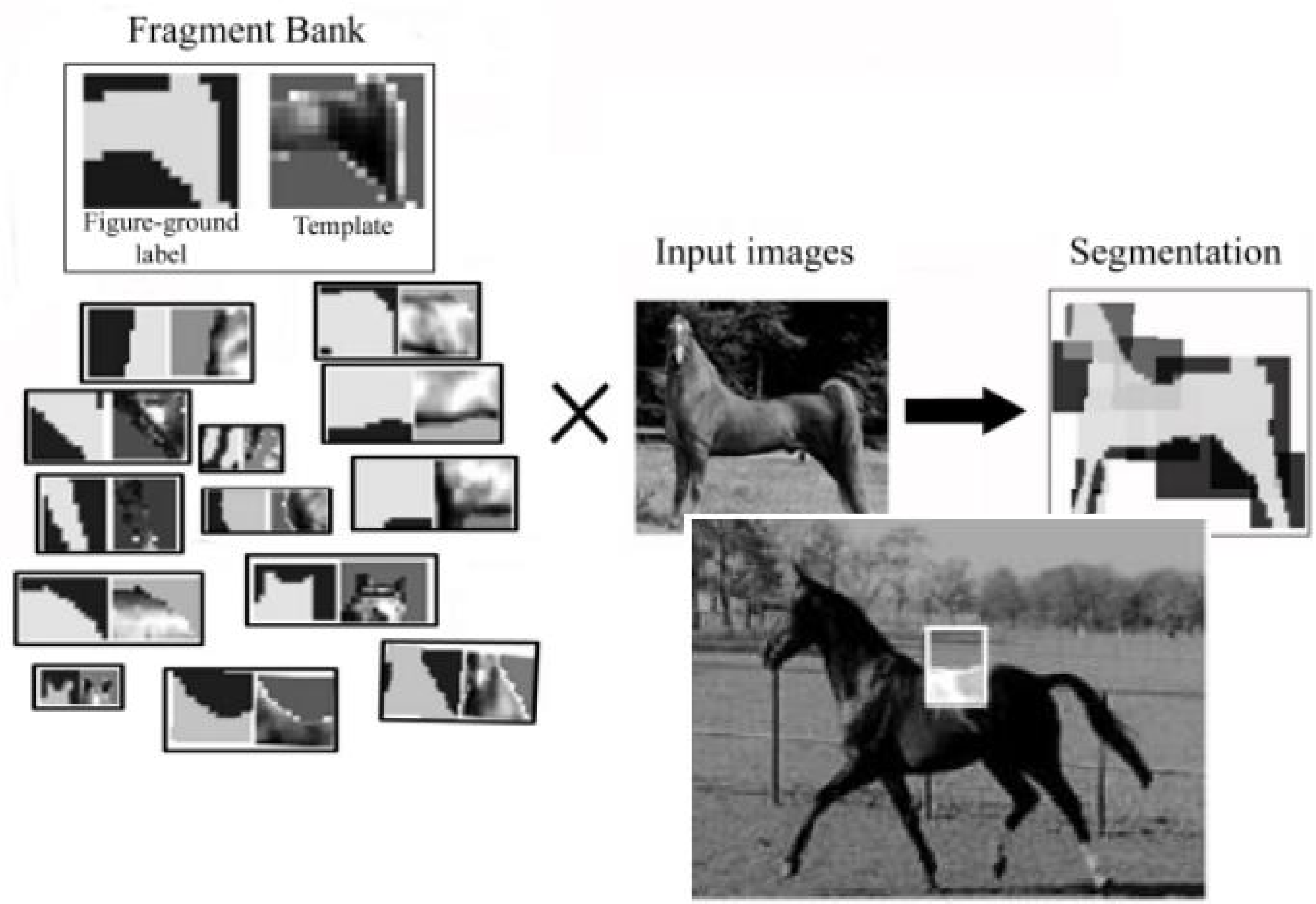
Detect first simple objects (reliable detectors) that provide strong contextual constraints to the target (screen -> keyboard -> mouse)

Learning local features
(First we need some intrinsic object features)



We maximize the probability of the true labels using **Boosting**.

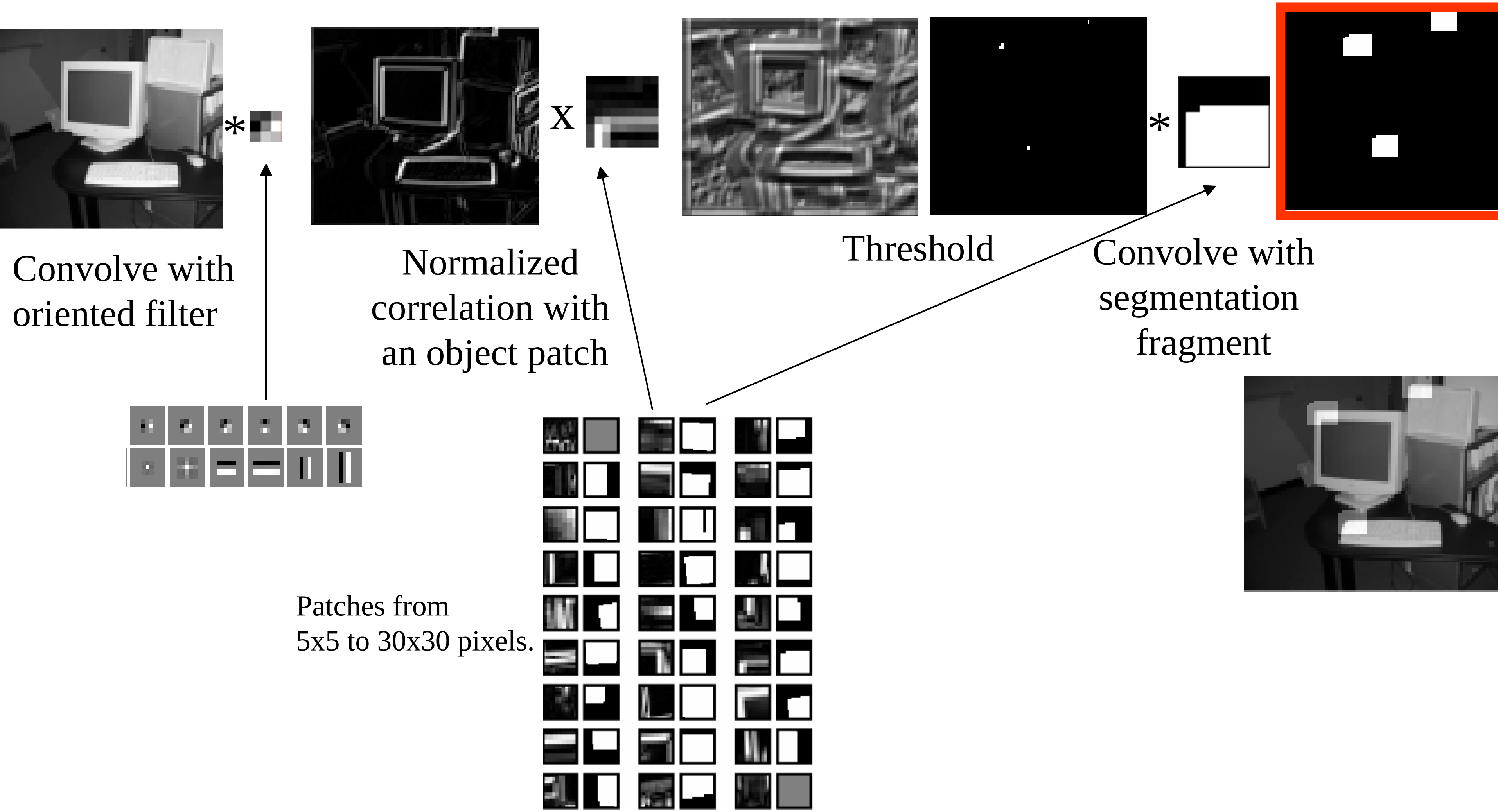
Fragments for class-specific segmentation



Source: Borenstein & Ullman, ECCV'02

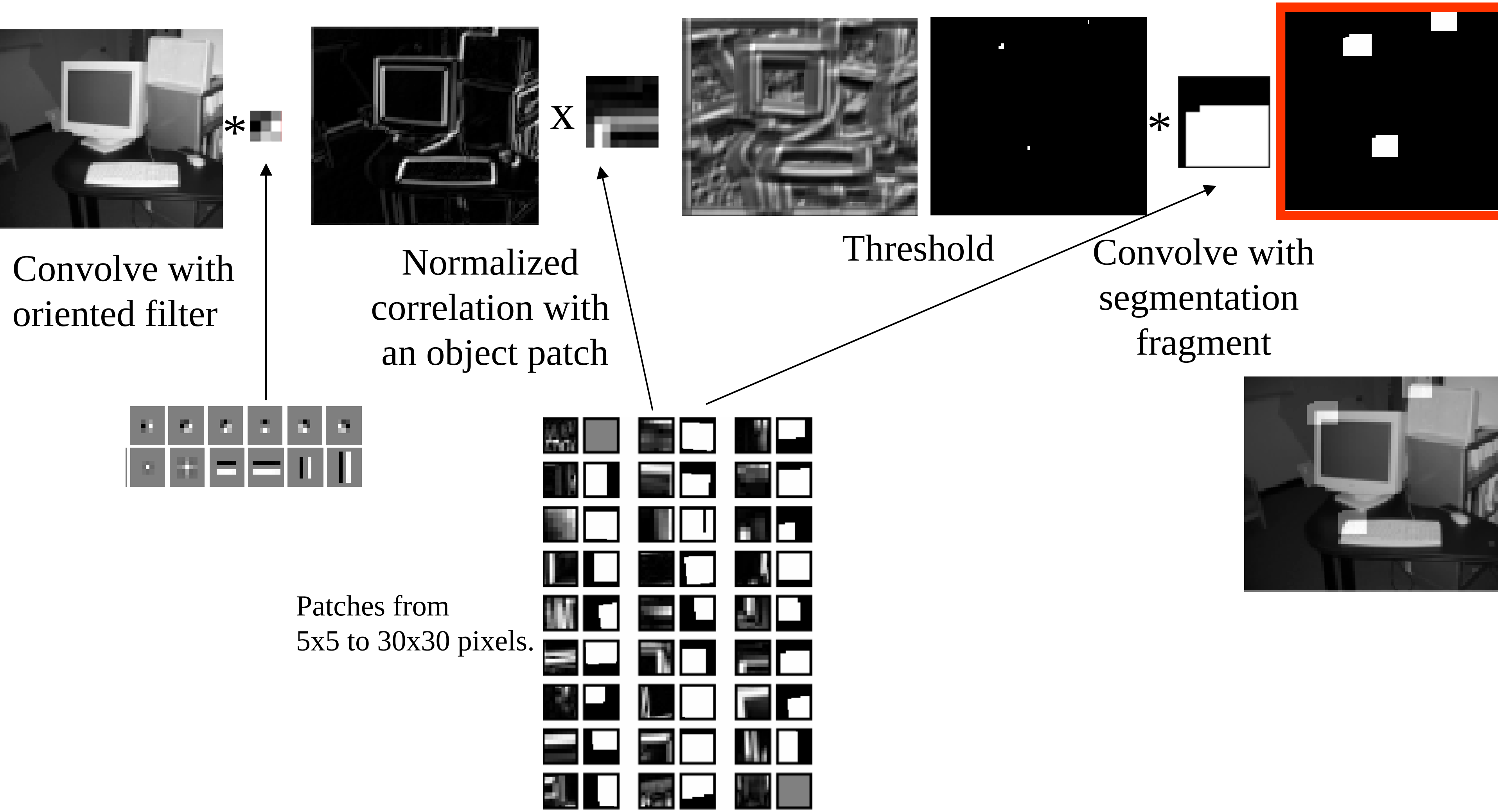
Object local features

(Borenstein & Ullman, ECCV 02)



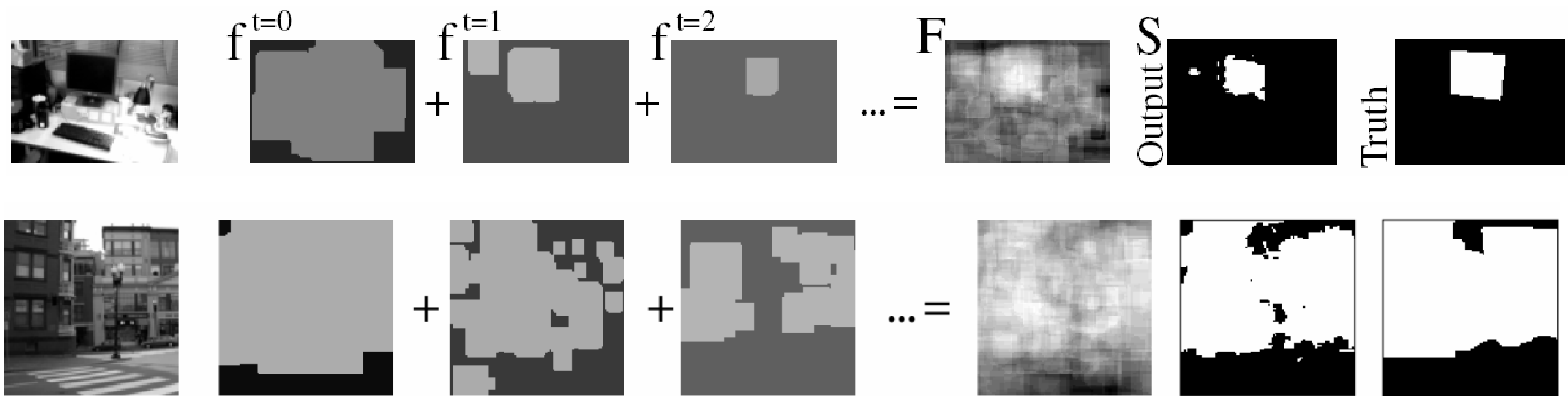
Object local features

(Borenstein & Ullman, ECCV 02)



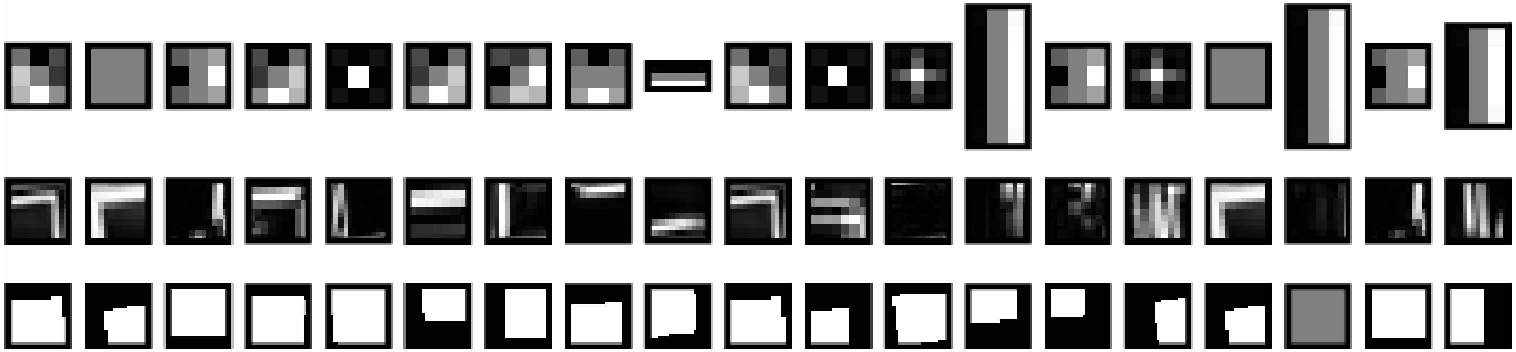
Results with local features

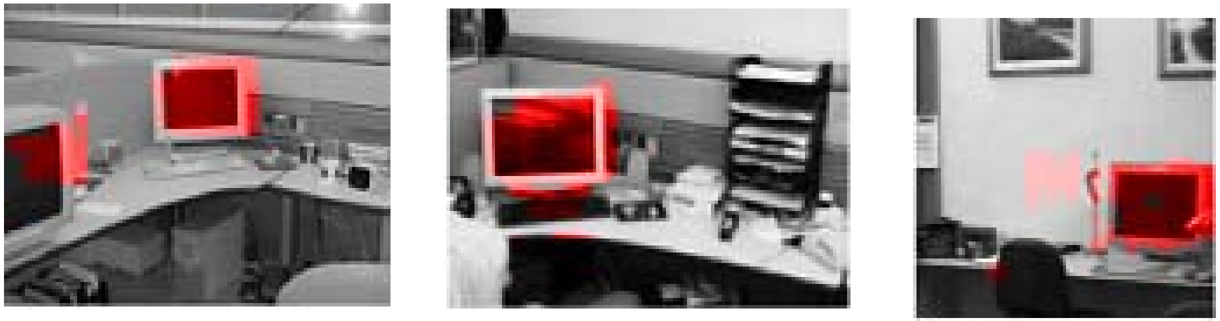
We use Boosting to build a classifier:



Results with local features

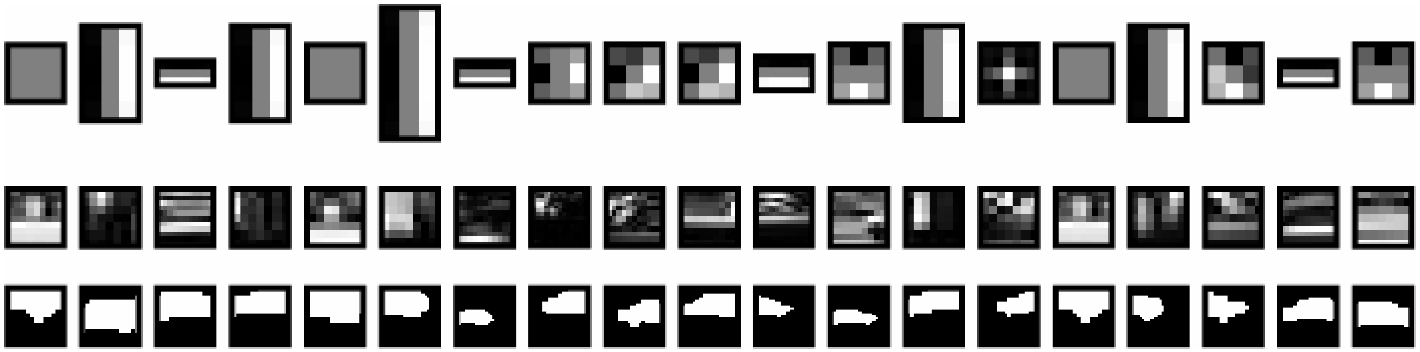
Screen

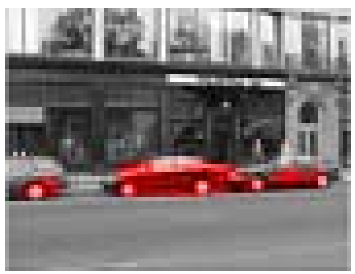




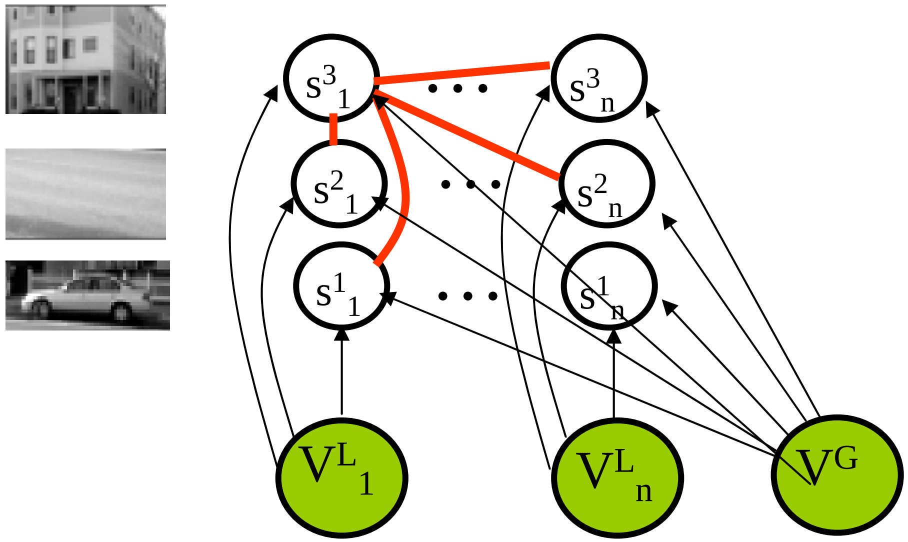
Results with local features

Car





Adding correlations between objects



We need to learn

- The structure of the graph
- The pairwise potentials

Previous work on joint object modeling

- Strat & Fischler (91)

Context defined using hand-written rules about relationships between objects

- Torralba & Sinha (01)

Global context to predict objects.

- Fink & Perona (03)

Use boosting incorporating the output of multiple detectors to generate contextual weak-classifiers.

- Murphy, Torralba & Freeman (03)

Use graphical models to represent the relation between global context and objects.

- Carbonetto, Freitas & Barnard (04)

They extend the work on “words and images” by adding spatial consistency between labels.

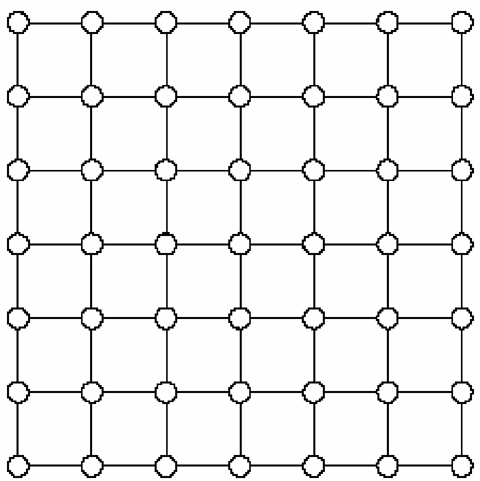
- He, Zemel & Carreira-Perpinan (04)

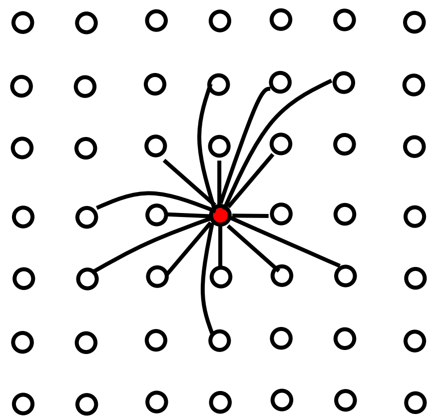
Use dense connectivity for incorporating spatial context using Multiscale conditional random fields.

Learning in conditional random fields

- 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

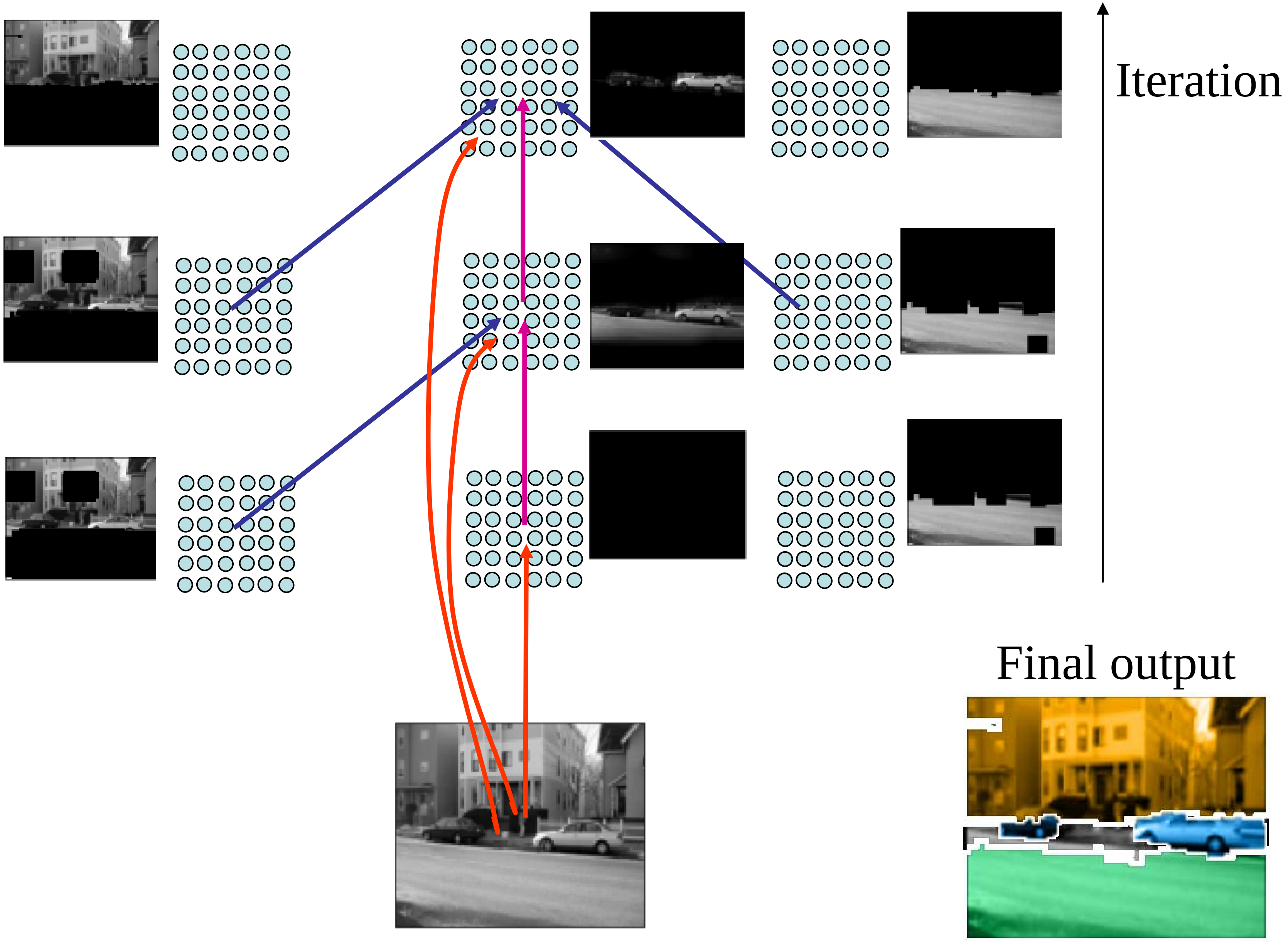
Graphical models for vision





Densely connected graphs
with low informative connections

Sequentially learning the structure

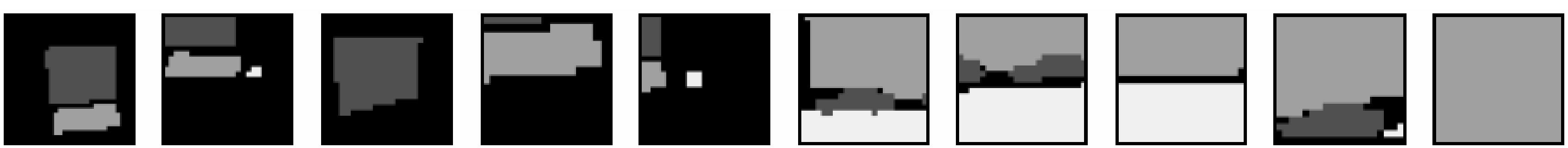


Sequentially learning the structure

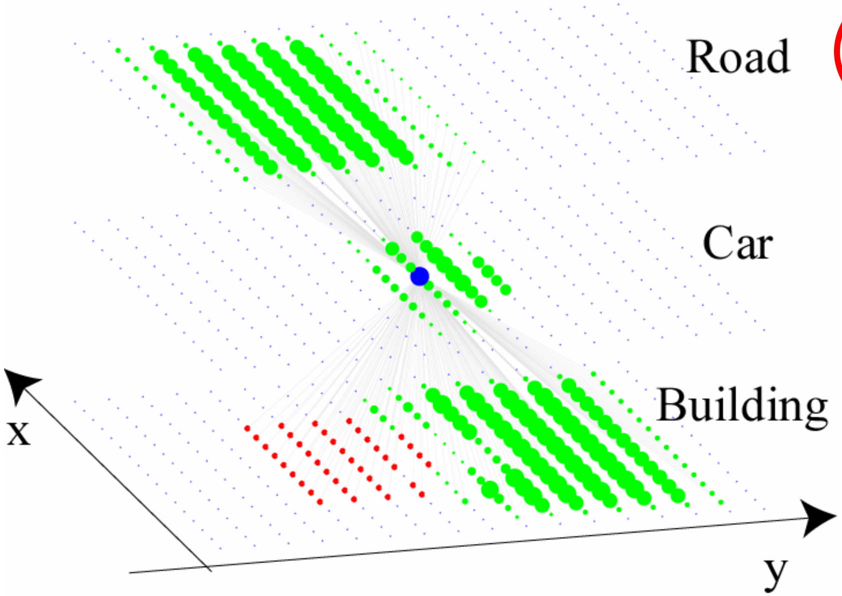
At each iteration of boosting

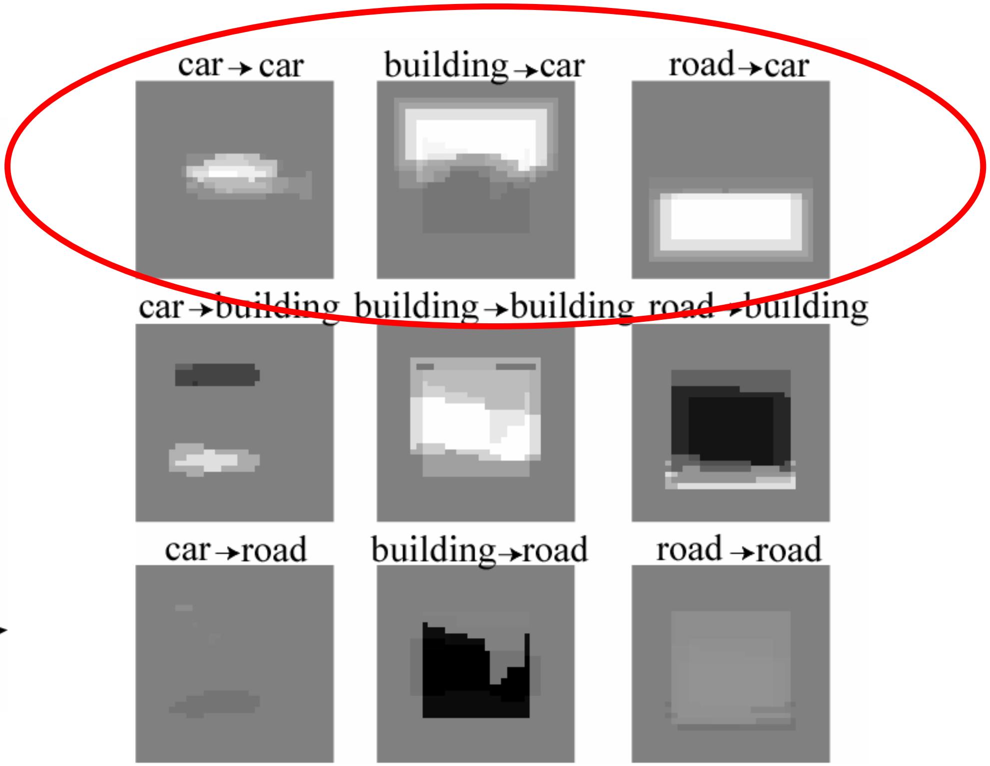
•We pick a weak learner applied to the image
(local or global features)

•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.

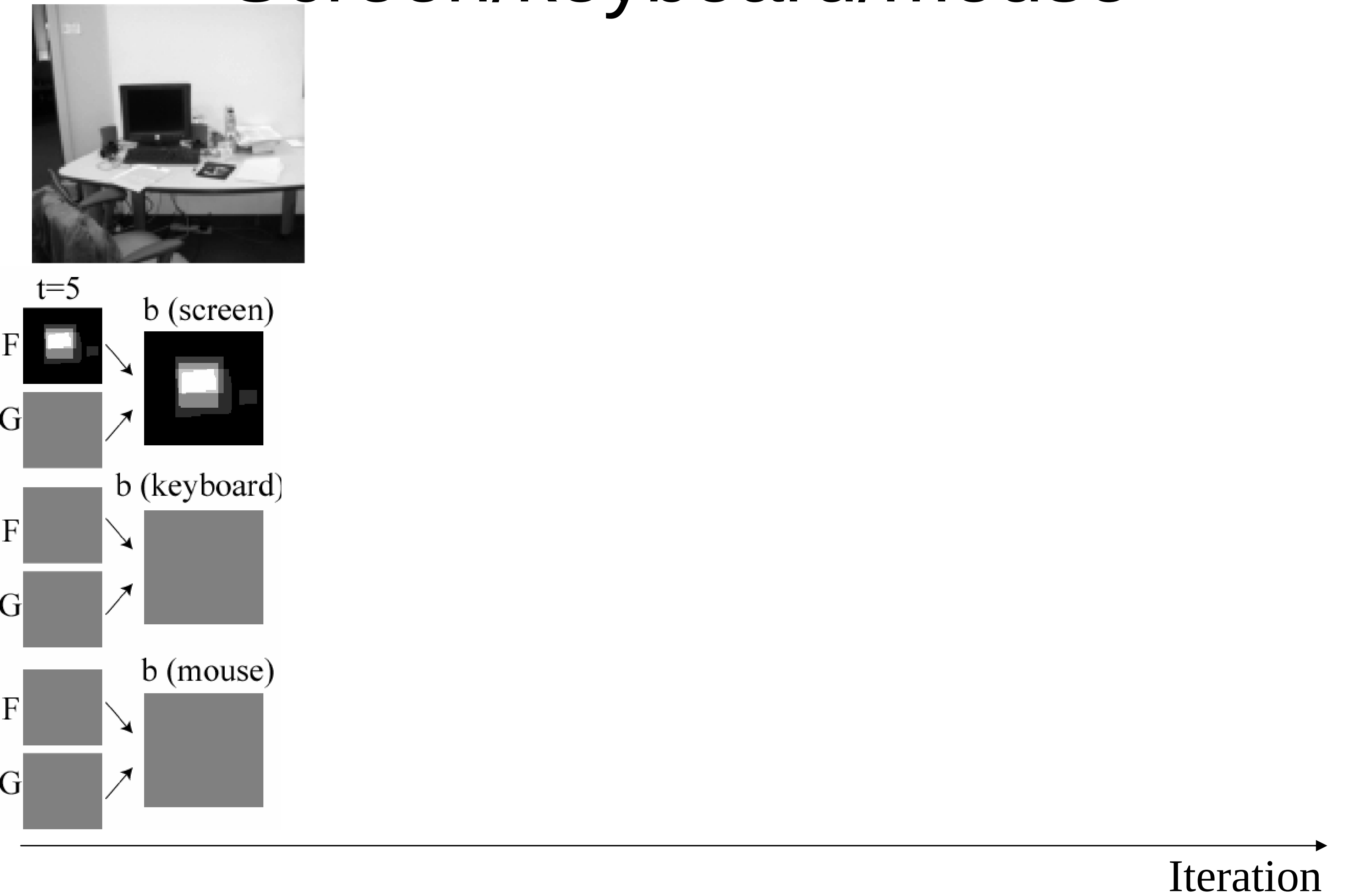


Car detection

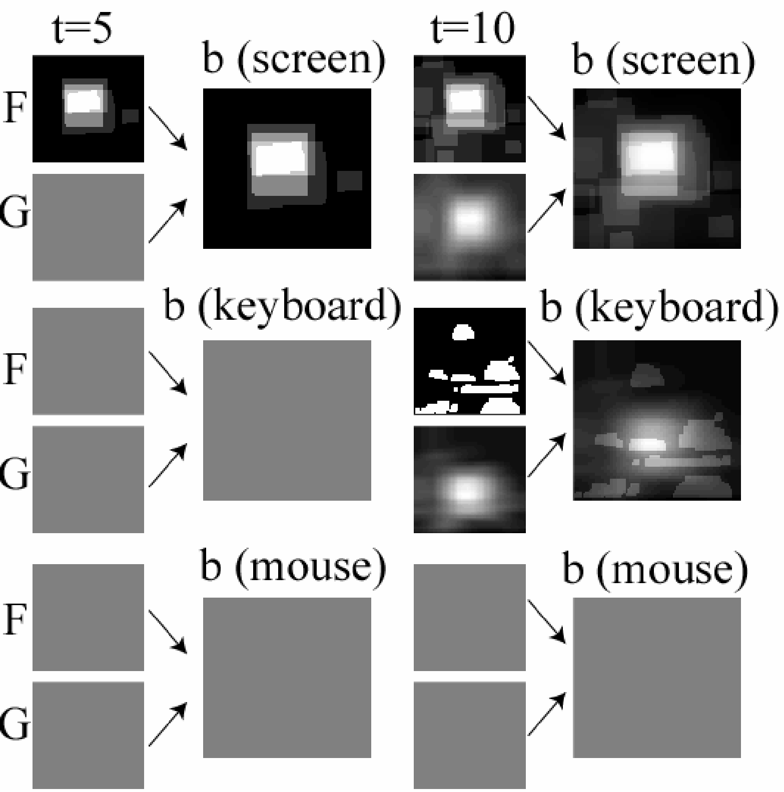




Screen/keyboard/mouse

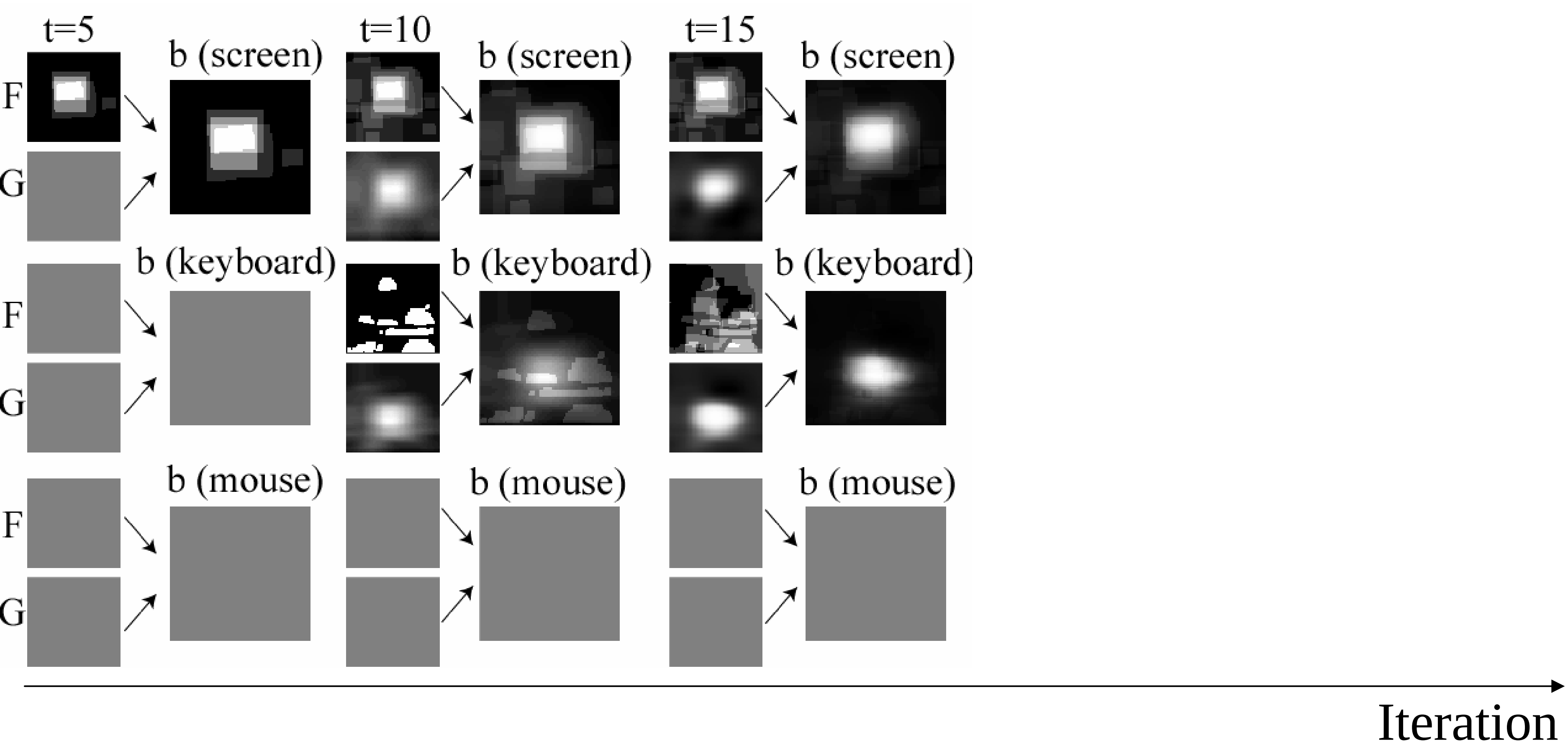


Screen/keyboard/mouse

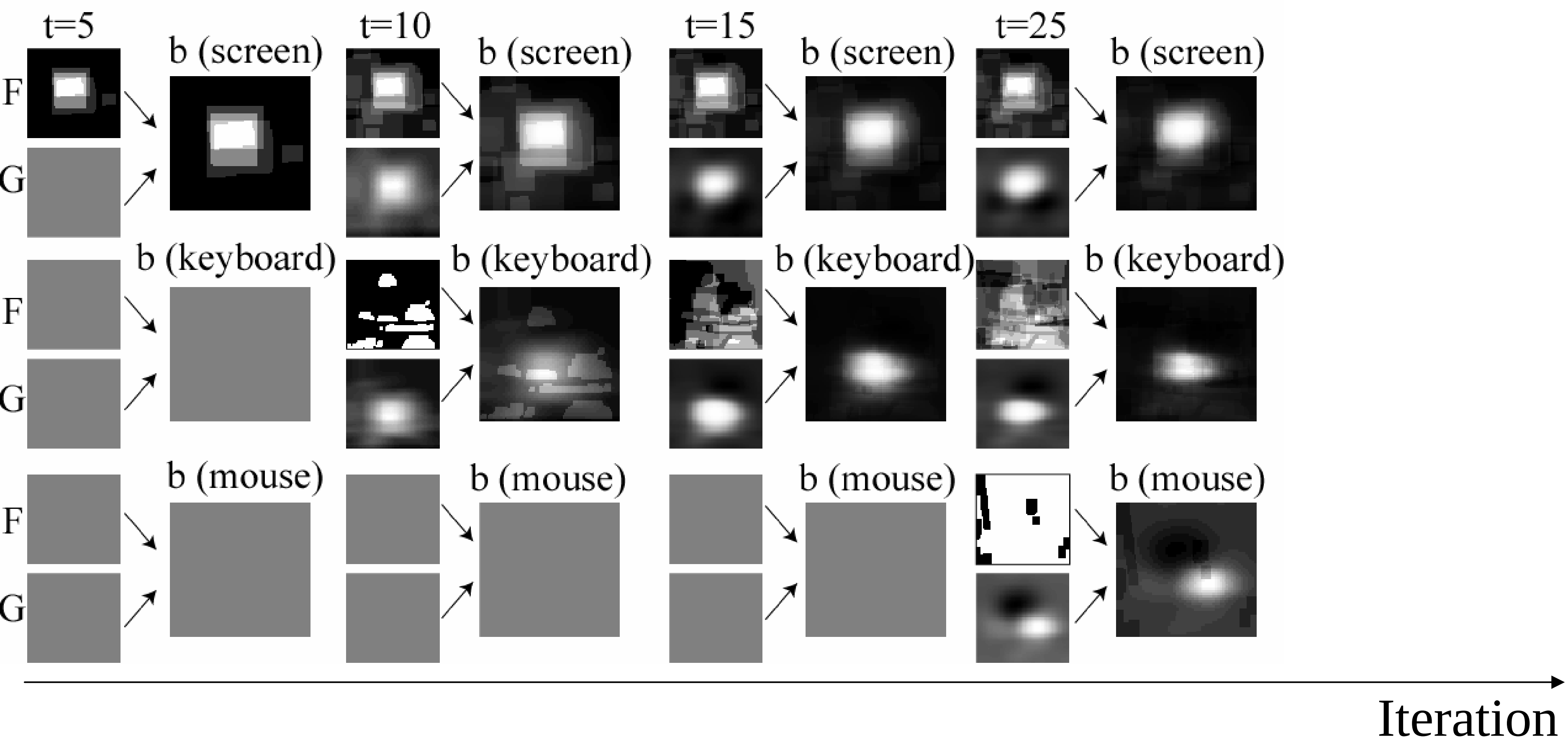


Iteration

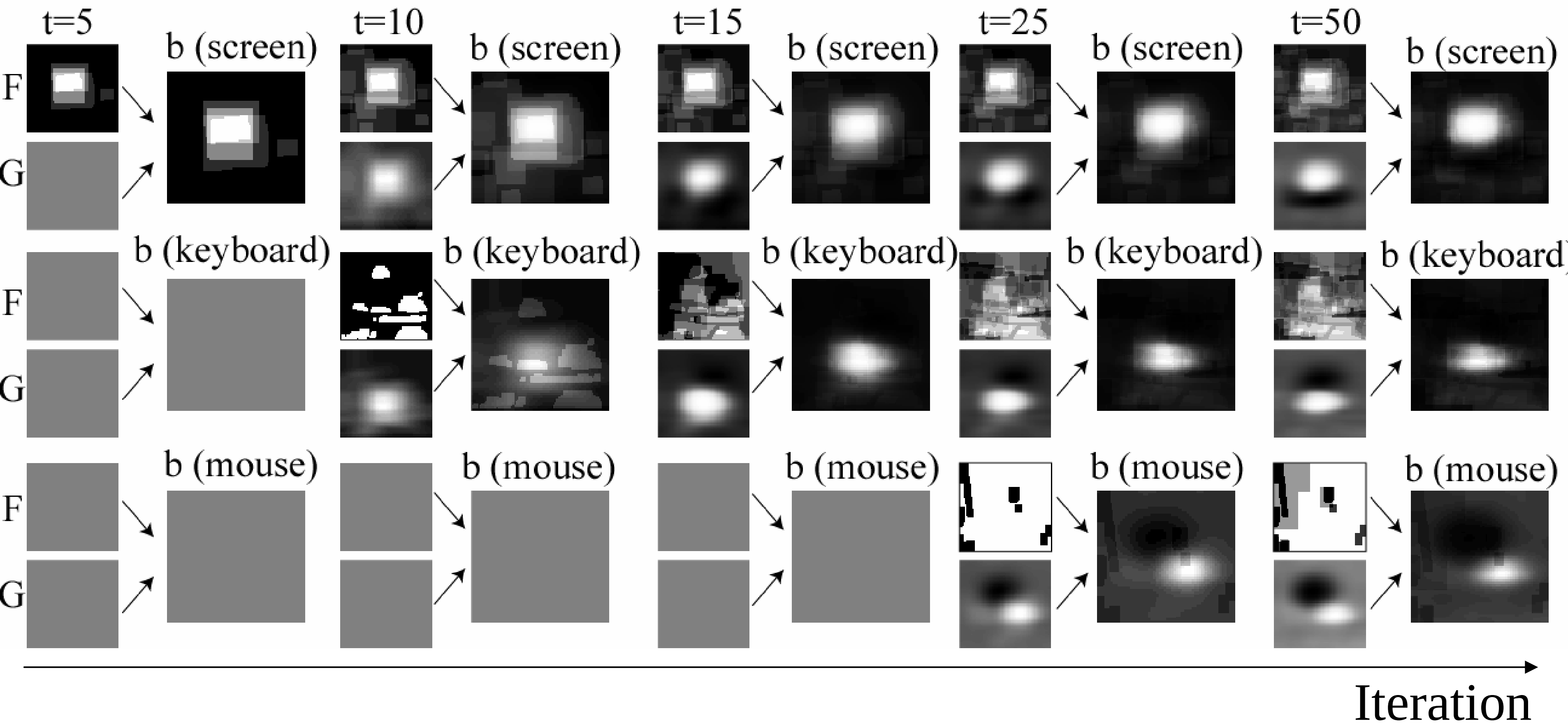
Screen/keyboard/mouse



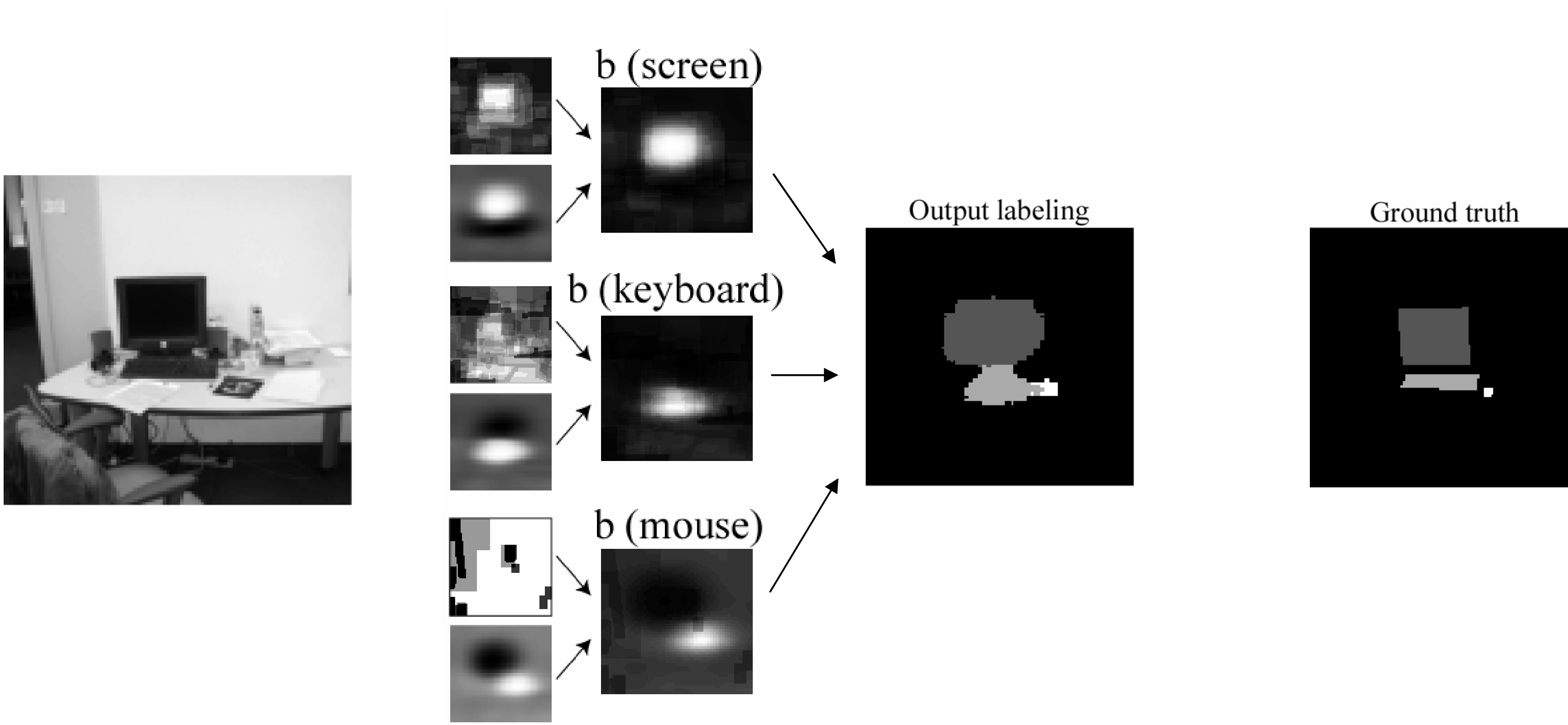
Screen/keyboard/mouse



Screen/keyboard/mouse



Screen/keyboard/mouse

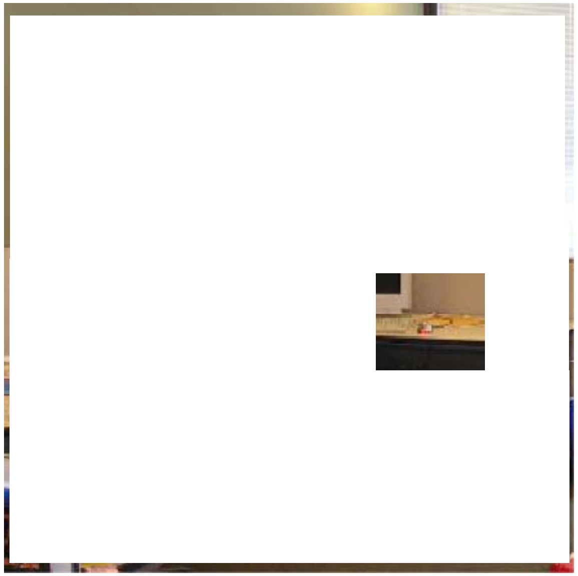


Cascade

Geman et al, 98; Viola & Jones, 01

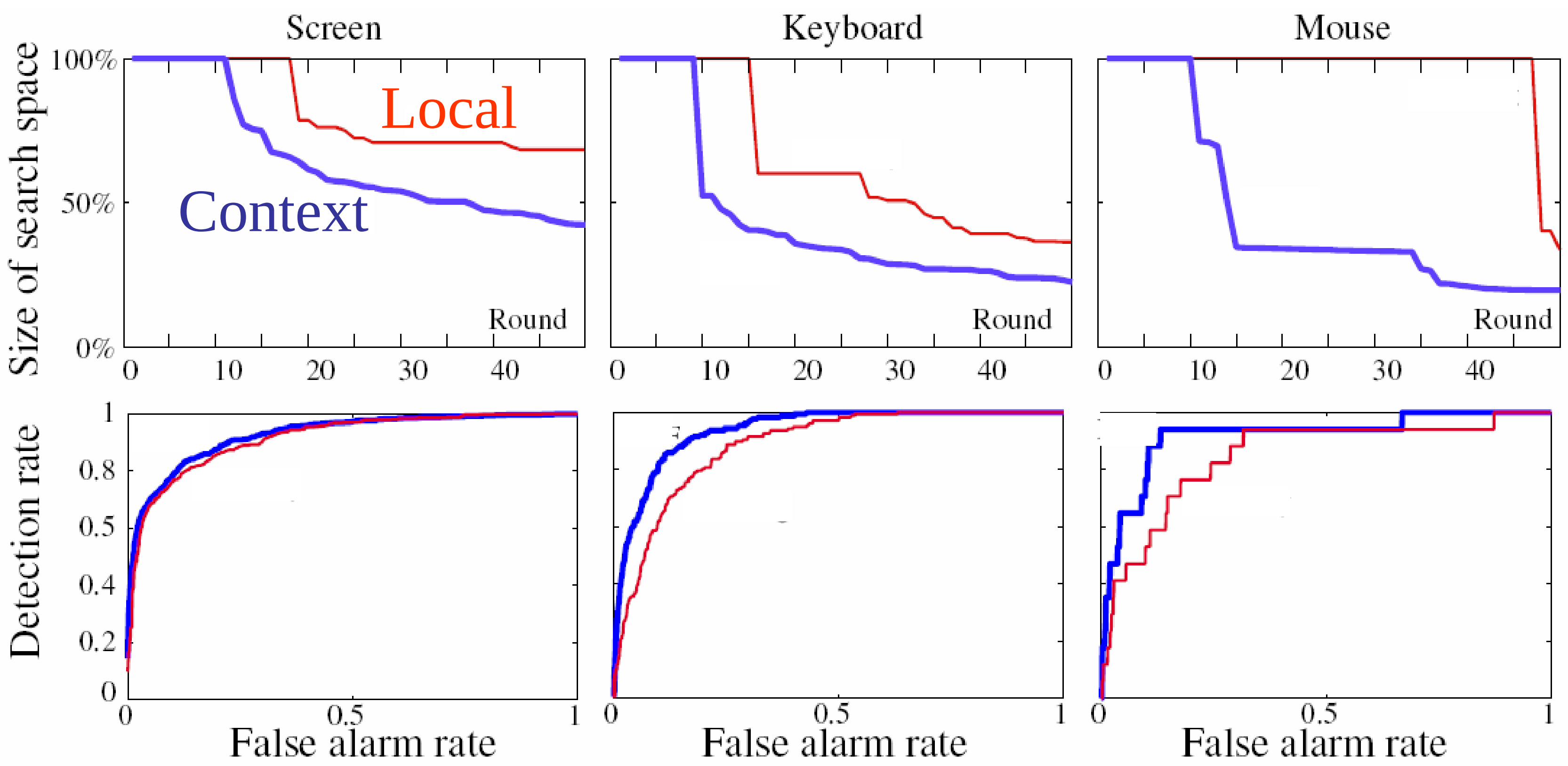
Set to zero the beliefs of nodes with low probability of containing the target.

Perform message passing only on undecided nodes



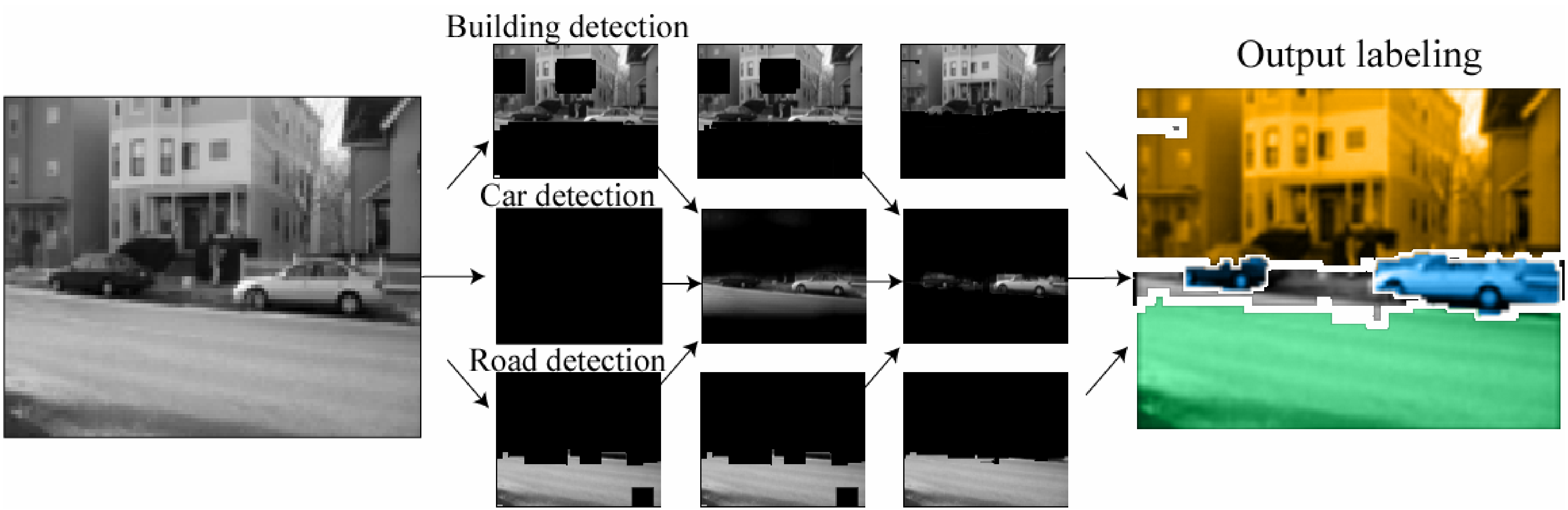
The detection of the screen reduces the search space for the mouse detector.

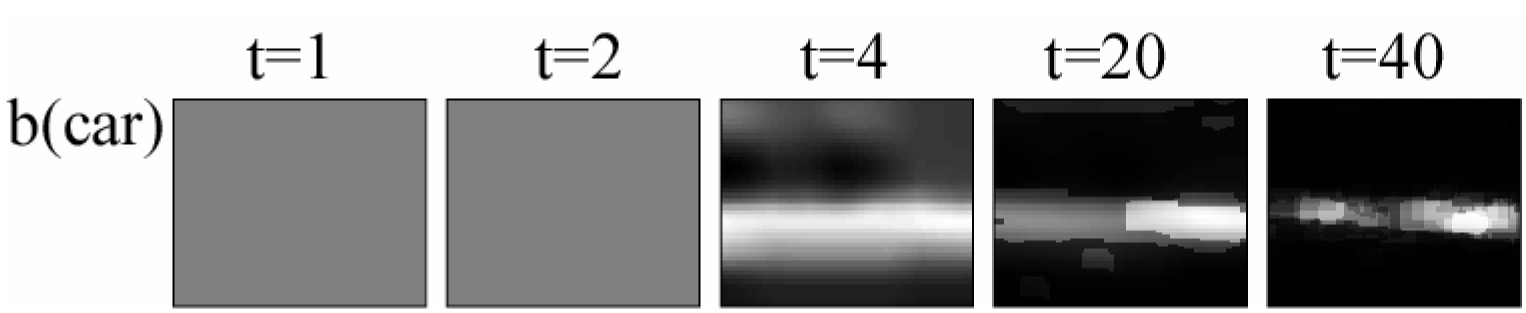
Cascade



Geman et al, 99;Viola and Jones 01

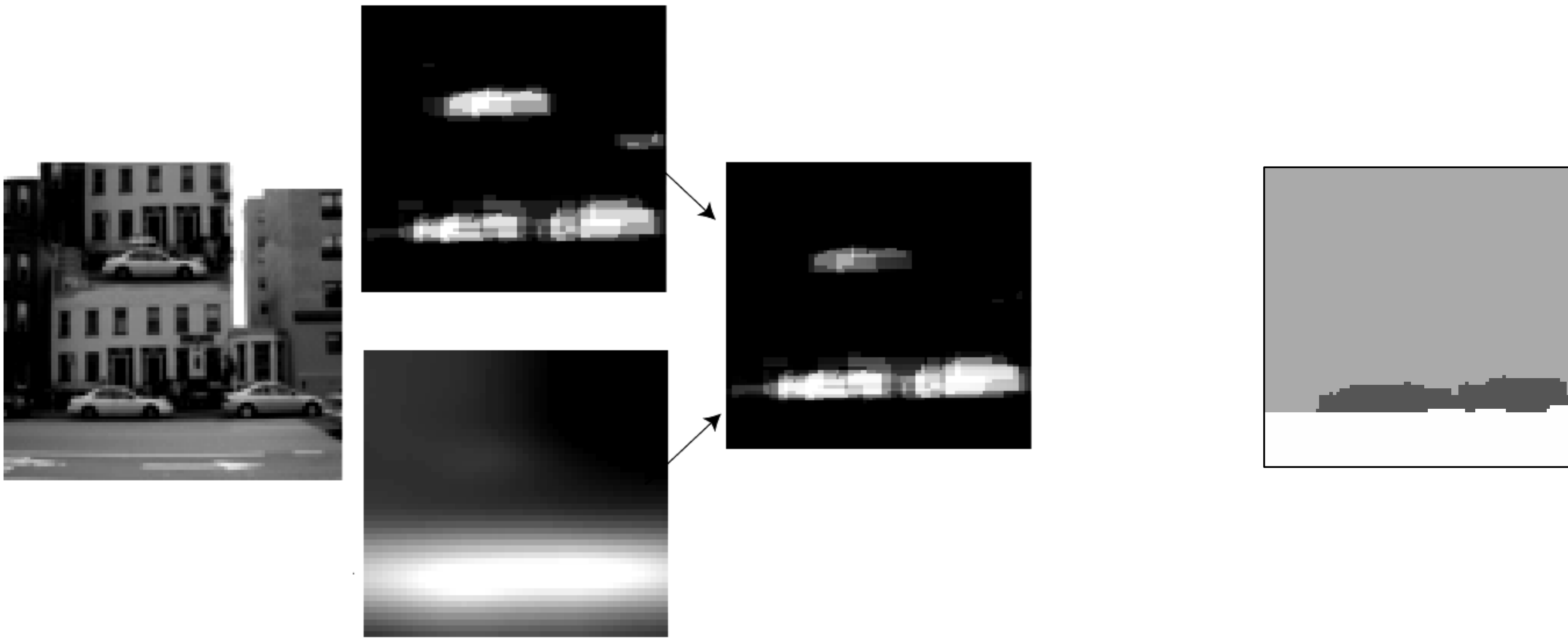
Cascade


Building detection
Car detection
Road detection
Output labeling


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

Car detection

From intrinsic features



From contextual features

A car out of context is less of a car

Future work

- Learn relationships between more objects (things get interesting beyond the 10 objects bar)