Sung Ju Hwang

Context

Outline

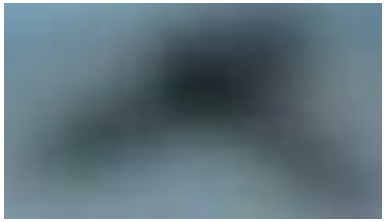
- The role of context in object recognition
- The types of context
- The use of the global context in object localization
  - Global context on object presence estimation
  - Global context on location estimation
  - Global context on scale estimation
- The use of co-occurrence and spatial relation in object categorization
  - Co-occurrence context on object categorization
  - Spatial relation context on object categorization

Covered Paper

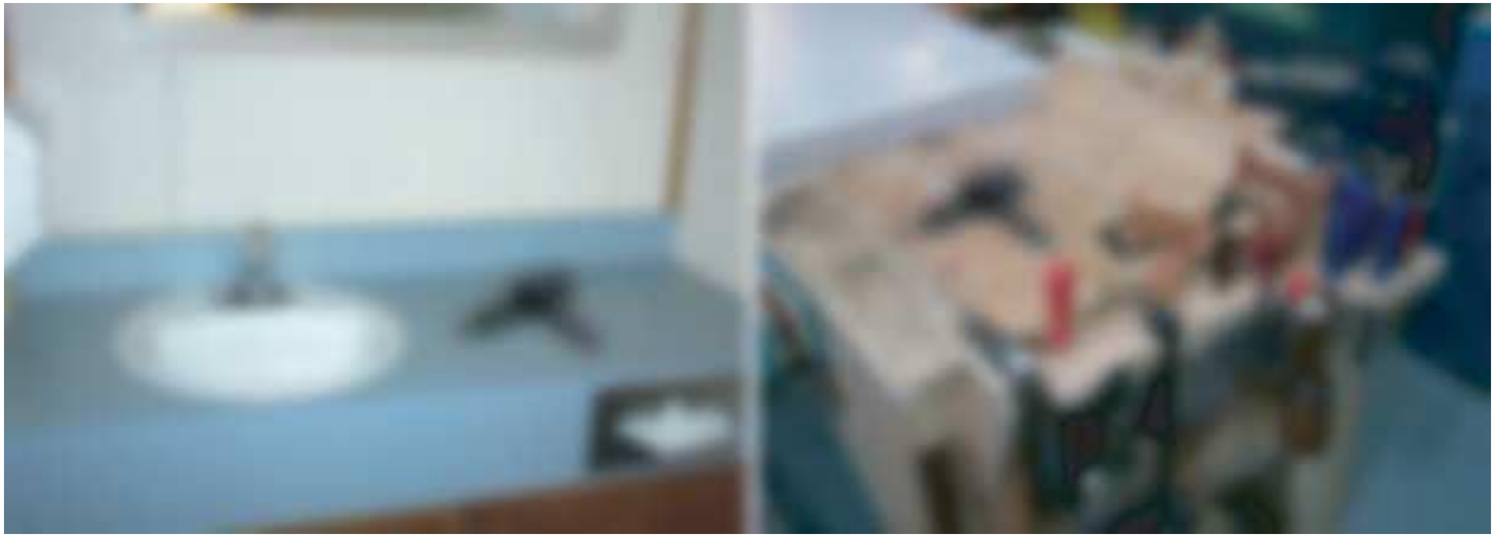
- Contextual Priming for Object Detection, Antonio Torralba, IJCV, 2003
    - Related papers
        - Using the forest to see the trees: a graphical model relating features, objects and scenes
          P. Murphy, A. Torralba and W. T. Freeman, NIPS, 2003.
        - Object detection and localization using local and global features
          K. Murphy, A. Torralba, D. Eaton, W. T. Freeman
        - Contextual Guidance of Attention in Natural scenes: The role of Global features on object search, Psychological Review. 2006.

- Object Cetegorization using Co-Occurrence, Location and Appearance, Carolina Galleguillos, Andrew Rabinovich, Serge Belongie, CVPR, 2008
    - Related papers
        - S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-Class Segmentation with Relative Location Prior, IJCV 2008.

What is this object?

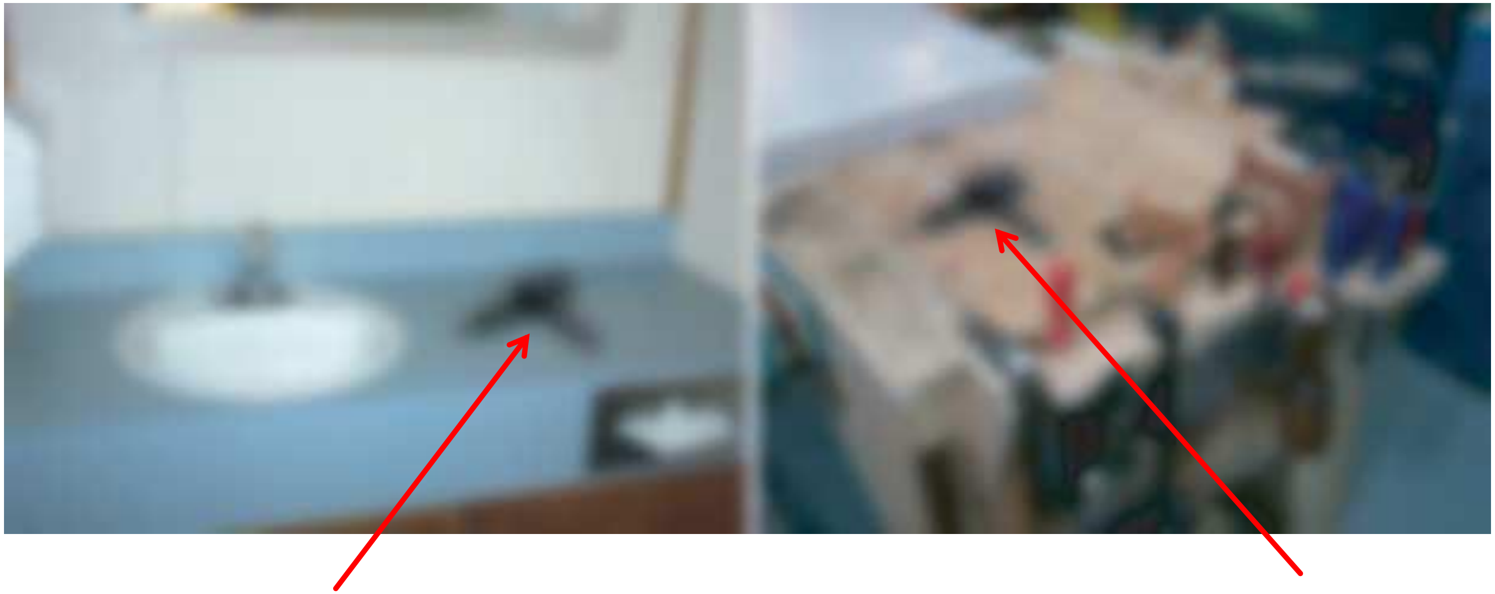


Object can be better understood with Context



Can you guess what the object is now?

Object can be better understood with Context



A hairdryer

A drill

How did you infer the object identity in both scenes?

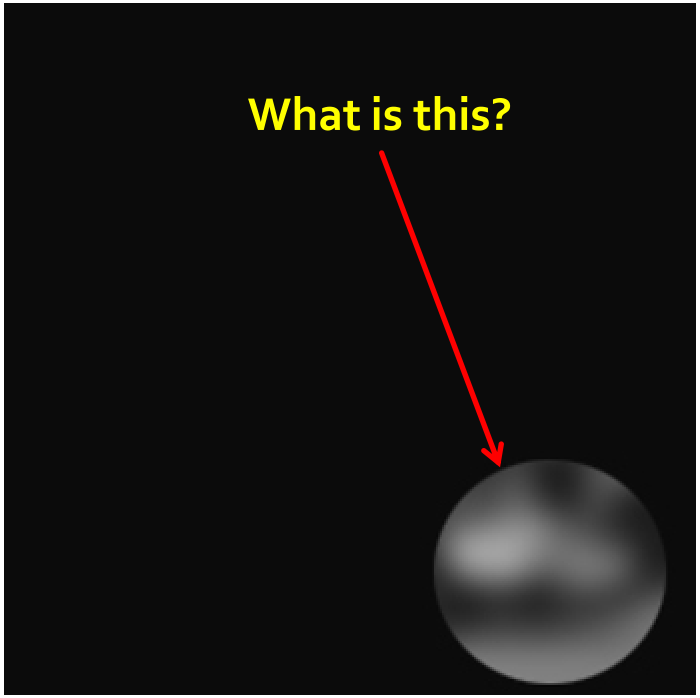
1) Scene information : Bathroom vs Worktable

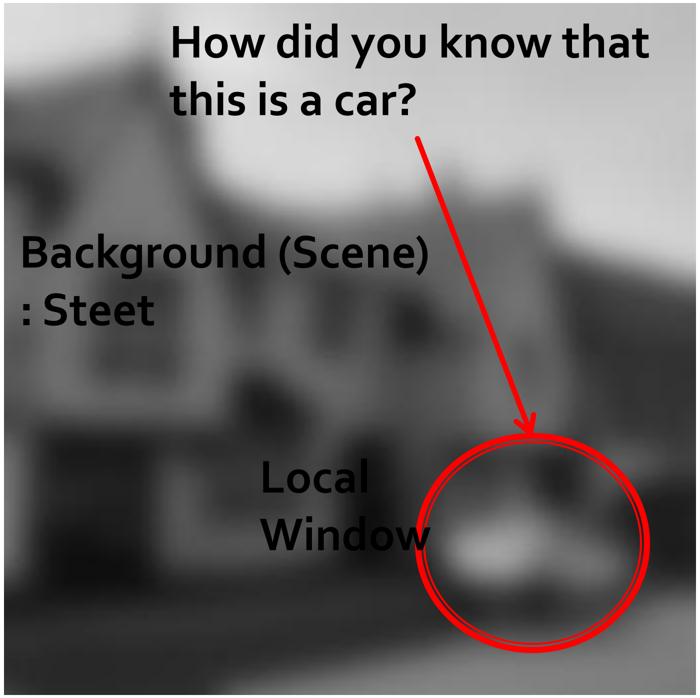
2) Other object : Sink vs Hammer

Types of context

- What is context? (in computer vision perspective?)
  - Any information that is not directly provided by the appearance of an object
    - Can be obtained by nearby image data
    - Can be obtained from the scene that contains the object
    - Can be obtained from Image tags, or annotations
    - Can be obtained from presence and location of other objects

Global Context

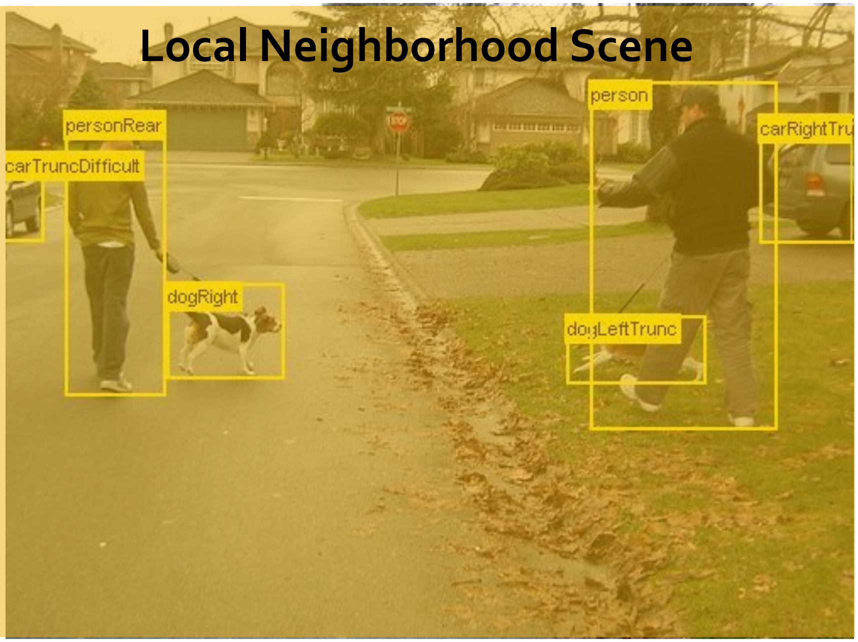



How did you know that this is a car?
Background (Scene) : Steet
Local Window

- Global Context
  - In local window search, the background is often treated as just clutters that distracts the object detection or recognition
  - However, they could give us more information than just few pixels in the local window can give us, especially when the image is degraded

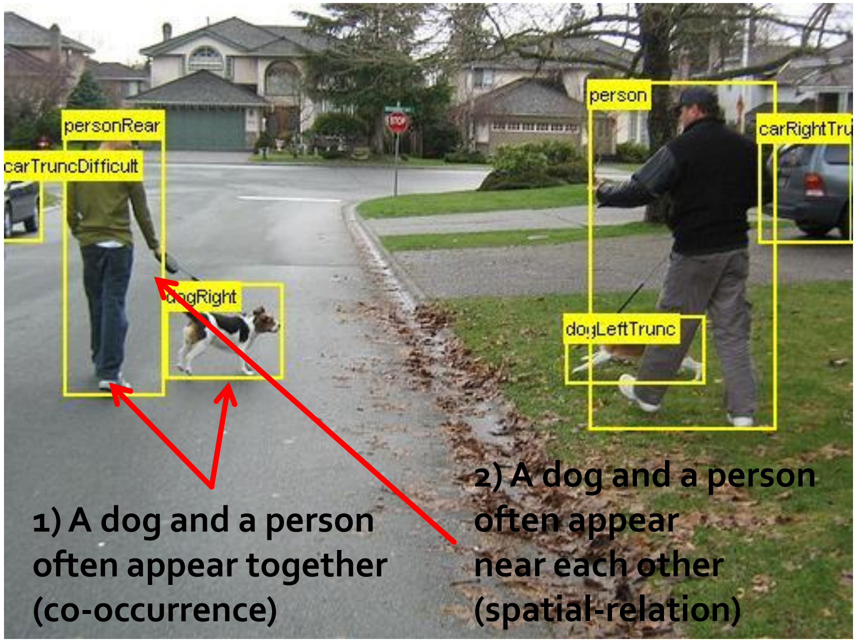
Types of Context

Context can be the scene that contains the object


Local Neighborhood Scene

Contextual Priming for Object Detection [Torralba 2003]

Context can be the relationship with other objects


1) A dog and a person often appear together (co-occurrence)

2) A dog and a person often appear near each other (spatial-relation)

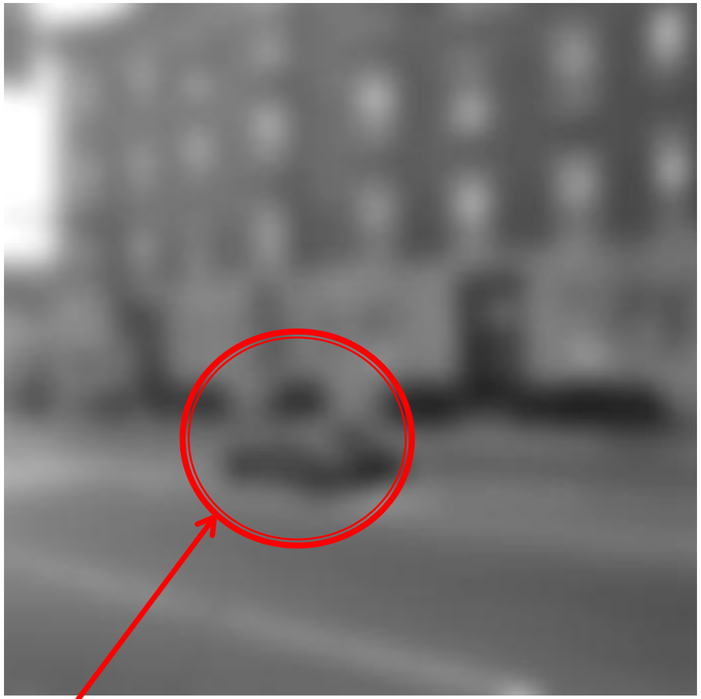
Object Categorization using Co-Occurrence, Location and Apperance [Galleguillos 2008]

Antonio Torralba [Torralba 2003]

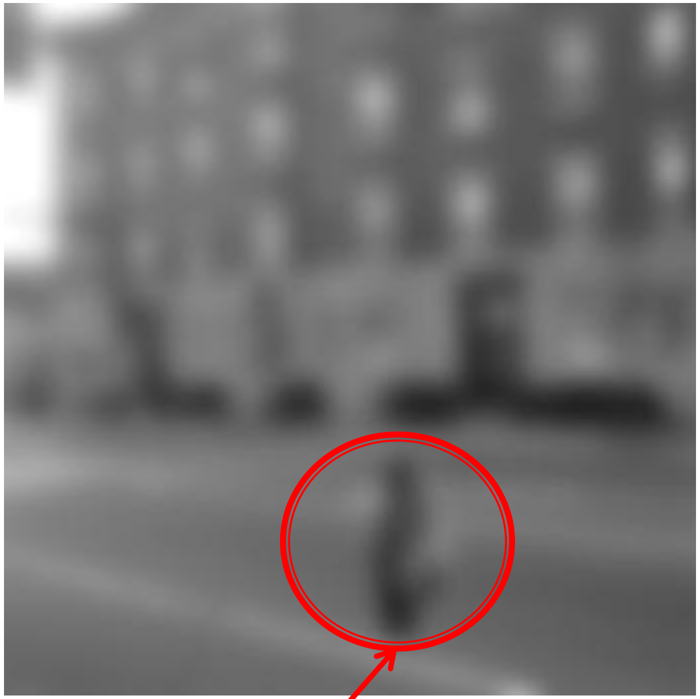
Contextual Priming for object detection [IJCV 2003]

What is this object?

What are these objects?

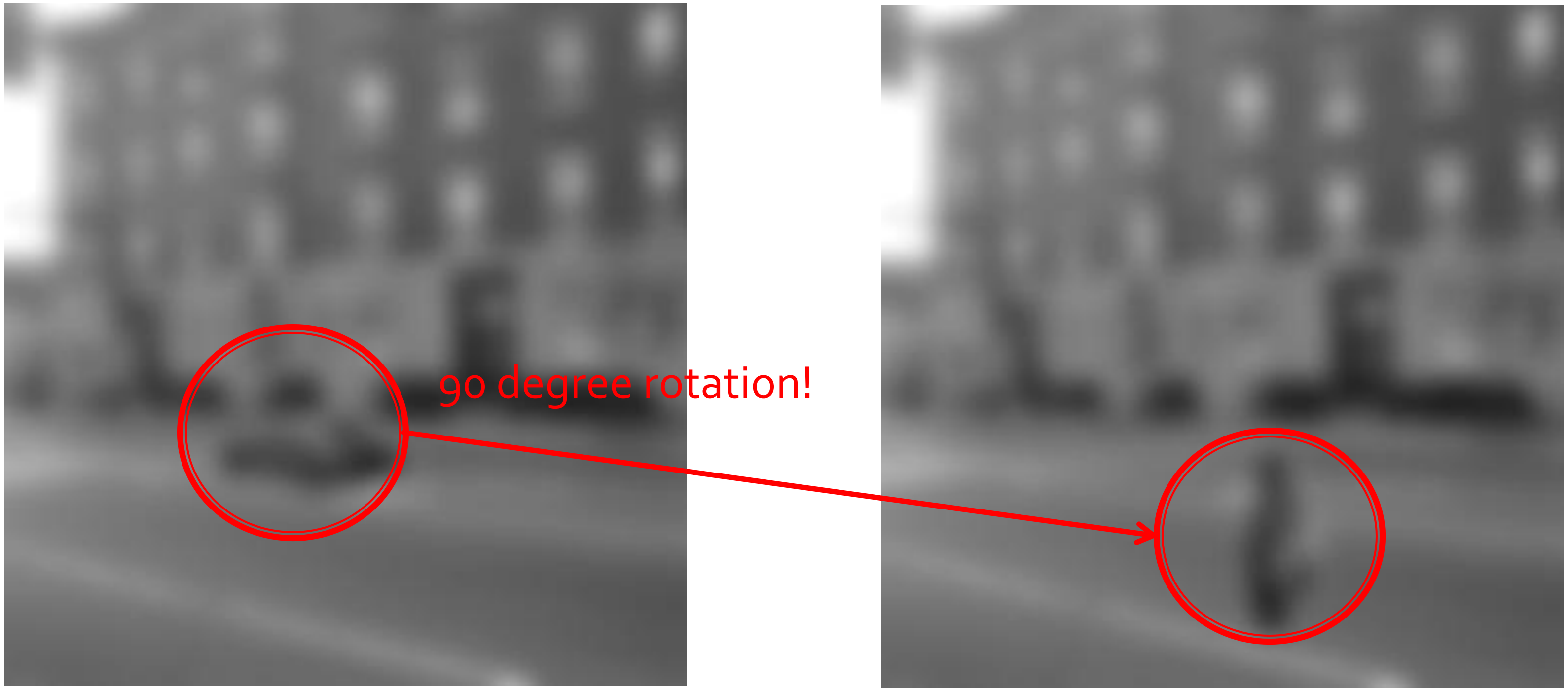


A car



A pedestrian?

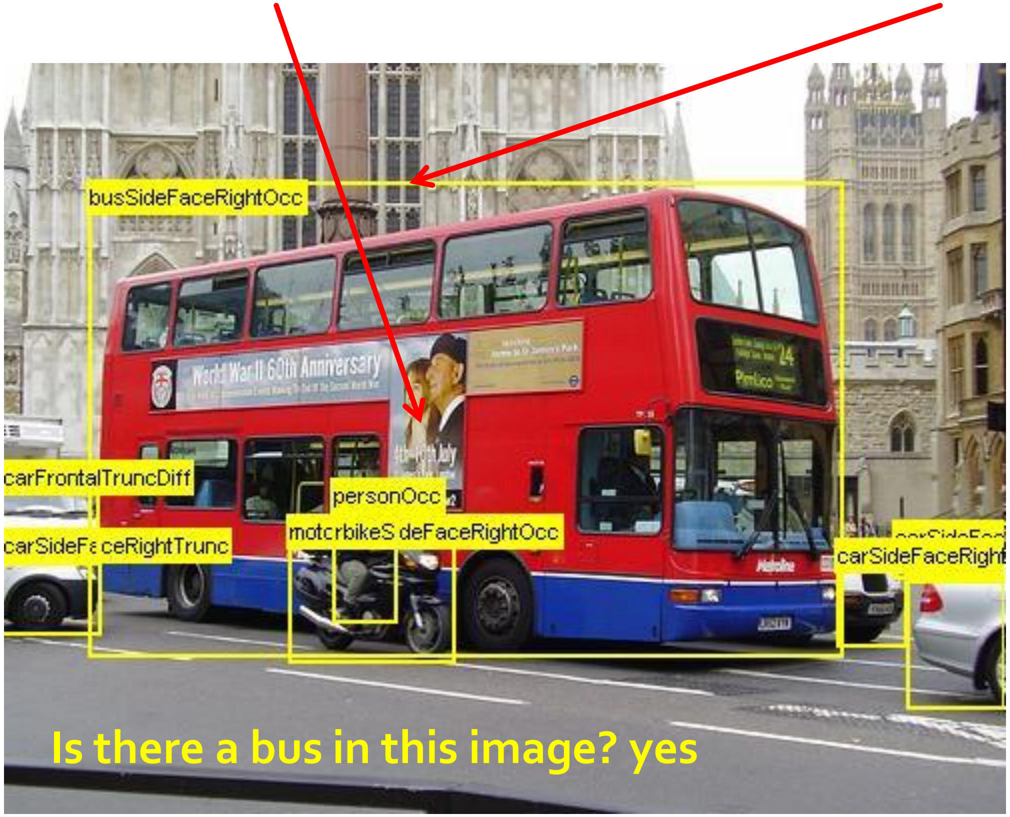
Location and the scale in the scene is an important cue in object understanding



Actually, the object in the second image is just a 90 degree rotation of the car object in the first image.

Why did we think that it was a pedestrian?

Object detection



Location of the object center

Scale of the bounding box

Is there a bus in this image? yes

- What do we mean by "detecting an object" ?
  - Determining if the object is present in the image or not
  - Finding the location of the object
  - Finding the scale of the object

$O = \{o, x, \sigma, . . .\}$

o = object category, x = location, σ = scale

Statistical object detection

- The goal of an object detection algorithm is to obtain O from image features v
  - O = {o, x, σ, . . .}
    - o is the label of the object category (car, person)
    - x is the location of the image in image coordinates
    - σ is the size of the object
  - v = image measurement
    - pixel intensity, color distributions, shape features
- We want to compute the PDF $P(O \mid \mathbf{v}) = \frac{P(\mathbf{v} \mid O)}{P(\mathbf{v})} P(O)$
  - The probability distribution of O given the image measurment
  - The dimensionality of the image measurement could be enormously high!

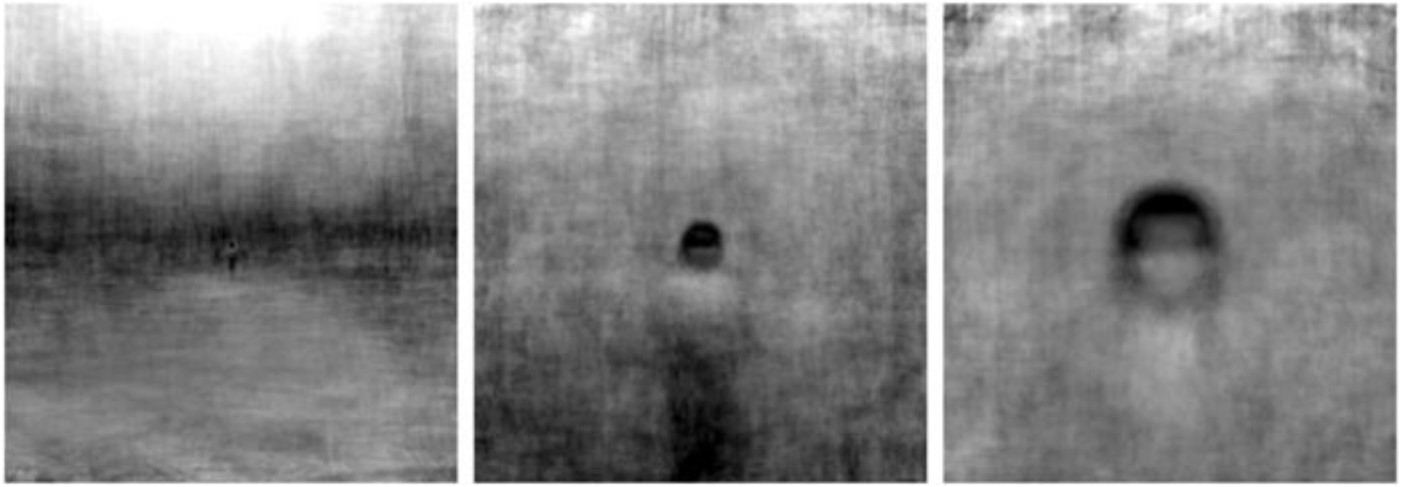
Classical object detection

- Dimensionality issue
  - Typically only the local image measurements are used
  - O is modeled on the given set of local (the image part belongs to the object) image measurment V_L
  - Context is treated as distractors

$$P(O \mid \mathbf{v}) \simeq P(O \mid \mathbf{v}_L) = \frac{P(\mathbf{v}_L \mid O)}{P(\mathbf{v}_L)} P(O)$$

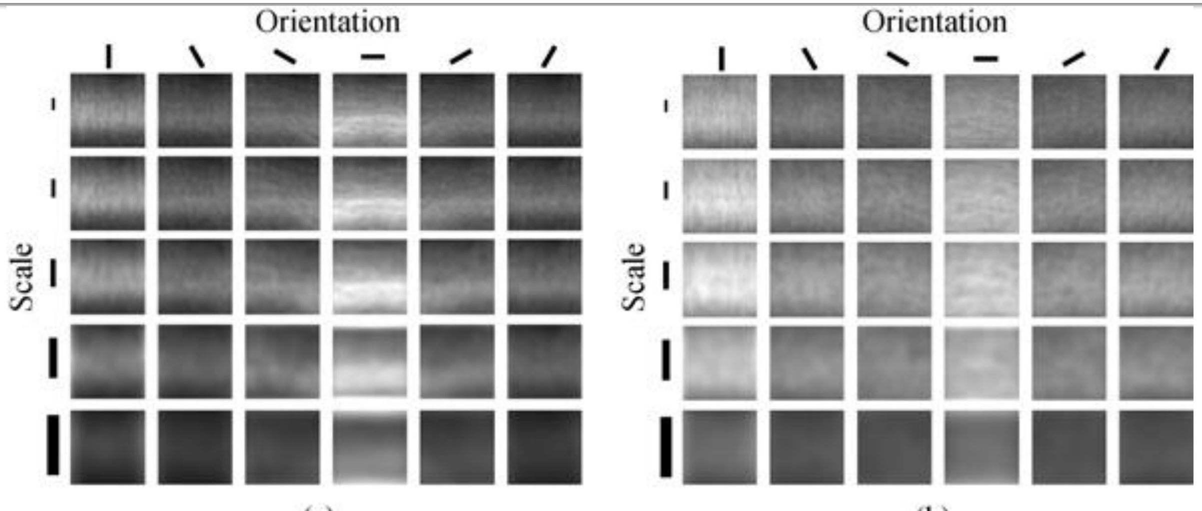
- What is the problem?
  - The object-based approach fails when the image is degraded
  - Exhaustive exploration of a large search space
    - Sliding window search

Object detection in context



- The average of pictures containing head in three different scales. What can we learn from these images?
  - There is also a regular pattern on the background
    - The background does not average to a mean gray
    - Background patterns are distinct for different scales
  - If we can learn these patterns, then we can predict the characteristics of an object

Object detection in context

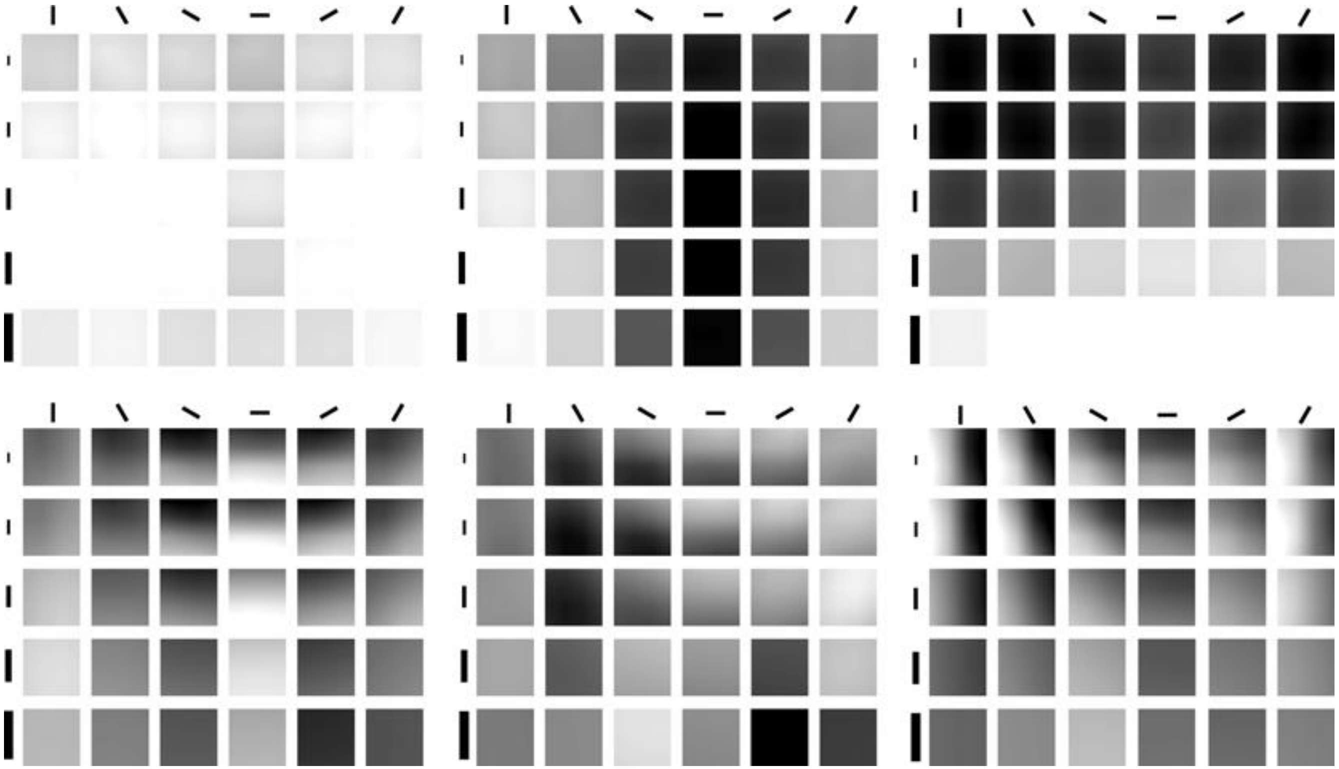


(a) (b)

$$v(\mathbf{x}, k) = \left| \sum_{\mathbf{x}'} i(\mathbf{x}') g_k(\mathbf{x} - \mathbf{x}') \right|$$

- Oriented band-path filters are used in the early stage of the visual pathway
  - This feature was proved useful in scene recognition task
  - Set of filters organized in 4 frequency bands and 6 orientations are used
- Resulting global context feature can tell us the presence of an object
  - (a) is a global context feature of the scenes that have cars but no peron
  - (b) s a global context feature of the scenes that have people but no car

Object detection in context



- Since the feature is very high-dimensional, principled component analysis (PCA) is done to reduce the dimensionality
  - The feature is decomposed using the basis function
  - The above are the top 1-3, 6-9 components

$$v(\mathbf{x}, k) \simeq \sum_{n=1}^{D} a_n \psi_n(\mathbf{x}, k)$$

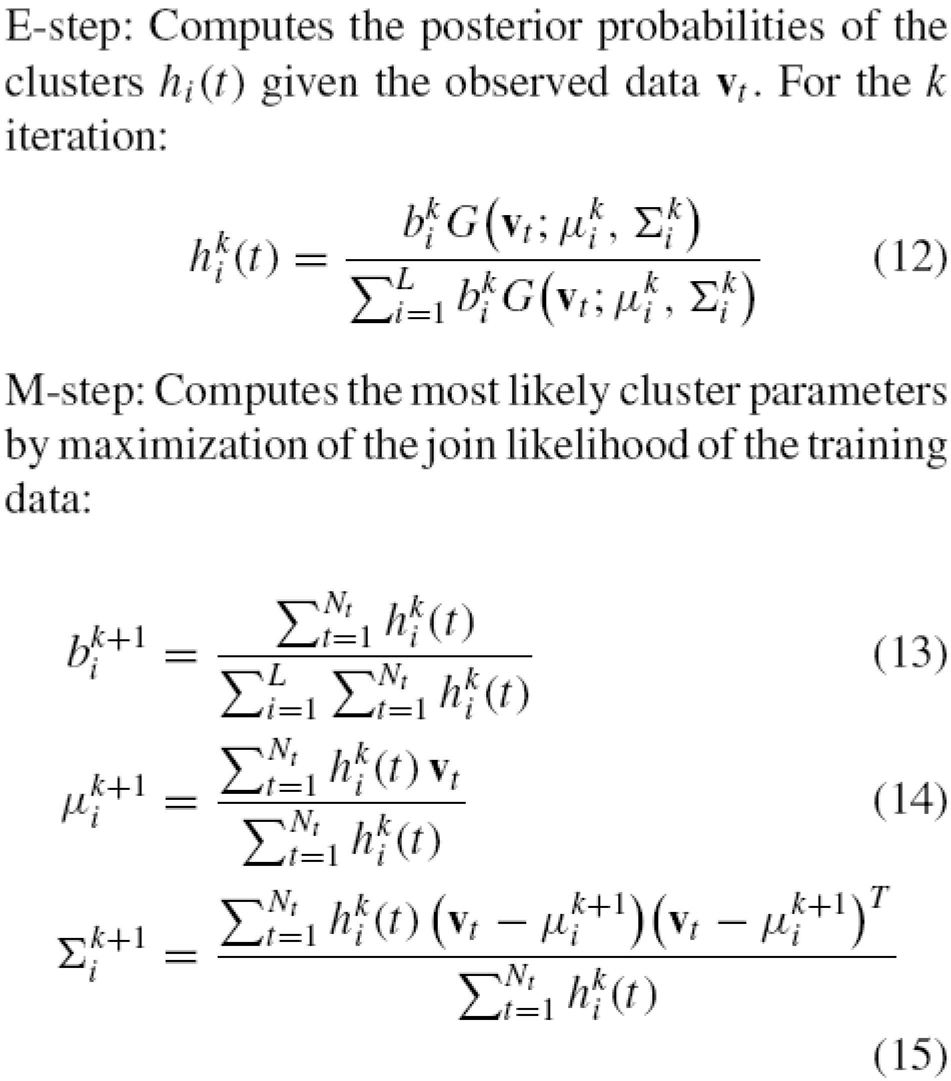
Object Priming

$$P(\mathbf{v}_C \mid o) = \sum_{i=1}^{M} b_i G(\mathbf{v}_C; \mu_i, \Sigma_i)$$

- The probability distribution of Vc given the object class o is modeled as mixture of Gaussians
  - $G(V_c;\mu_i,\Sigma_i)$ is a multivariate Gaussian function of Vc
  - Bi is the mixing coefficient (weights for each Gaussian cluster)
  - M is the number of Gaussian cluster

- How is it learned?
  - Using the EM algorithm!

Learning the parameter of the GMM using the EM algorithm



E-step: Computes the posterior probabilities of the clusters $h_i(t)$ given the observed data $\mathbf{v}_t$. For the k iteration:

$$h_i^k(t) = \frac{b_i^k G\left(\mathbf{v}_t; \mu_i^k, \Sigma_i^k\right)}{\sum_{i=1}^{L} b_i^k G\left(\mathbf{v}_t; \mu_i^k, \Sigma_i^k\right)} \qquad (12)$$

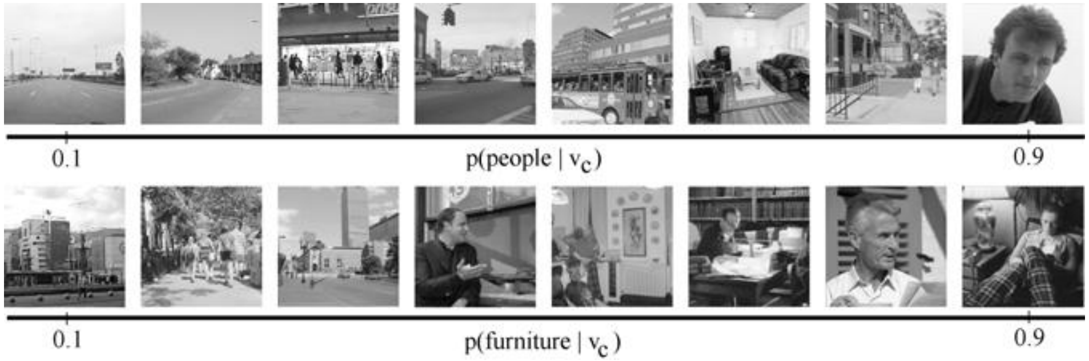
M-step: Computes the most likely cluster parameters by maximization of the join likelihood of the training data:

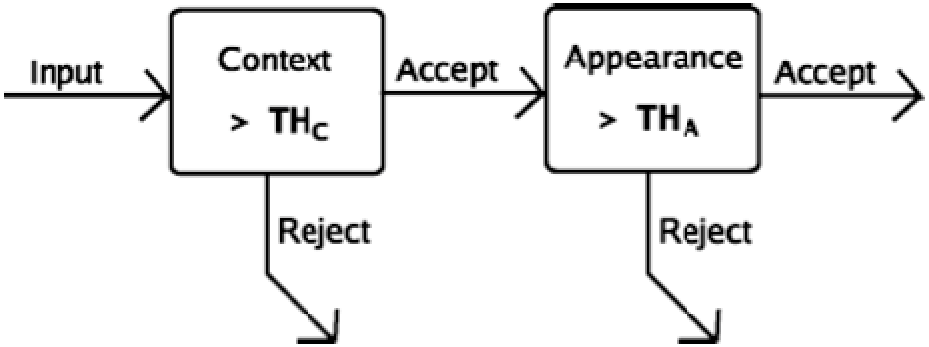
$$b_i^{k+1} = \frac{\sum_{t=1}^{N_t} h_i^k(t)}{\sum_{i=1}^{L} \sum_{t=1}^{N_t} h_i^k(t)} \qquad (13)$$

$$\mu_i^{k+1} = \frac{\sum_{t=1}^{N_t} h_i^k(t)\, \mathbf{v}_t}{\sum_{t=1}^{N_t} h_i^k(t)} \qquad (14)$$

$$\Sigma_i^{k+1} = \frac{\sum_{t=1}^{N_t} h_i^k(t) \left(\mathbf{v}_t - \mu_i^{k+1}\right)\left(\mathbf{v}_t - \mu_i^{k+1}\right)^T}{\sum_{t=1}^{N_t} h_i^k(t)} \qquad (15)$$

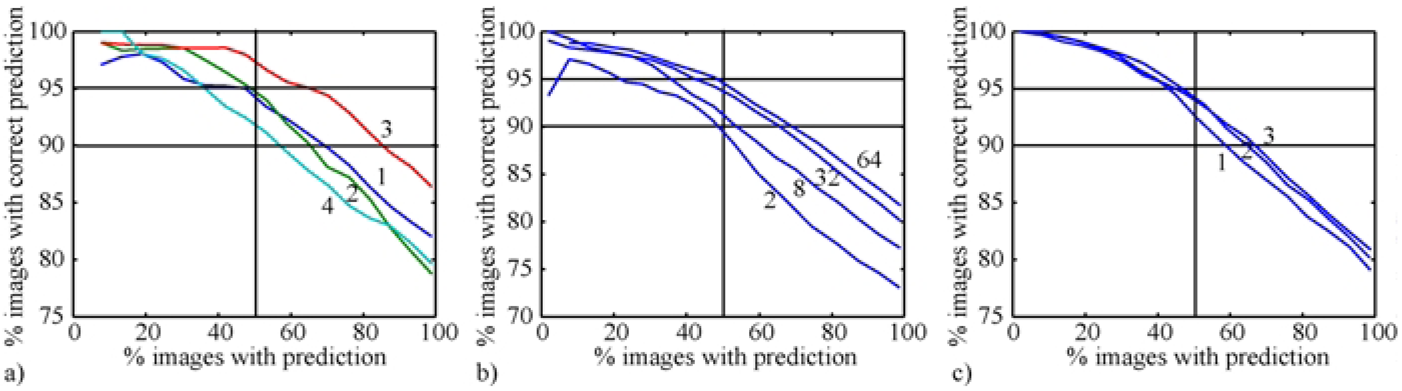
Global context can be helpful in object presence prediction





- Object presence could be predicted without actually detecting an object
  - This is relevant to what humans do
  - Cascade structure for faster object detection

Performance of object presence prediction using the learned prior



Performances as a function of decision threshold th and

(a) target object (o1 = people, o2 = furniture, o3 = vehicles and o4 = trees),

(b) number of contextual features and

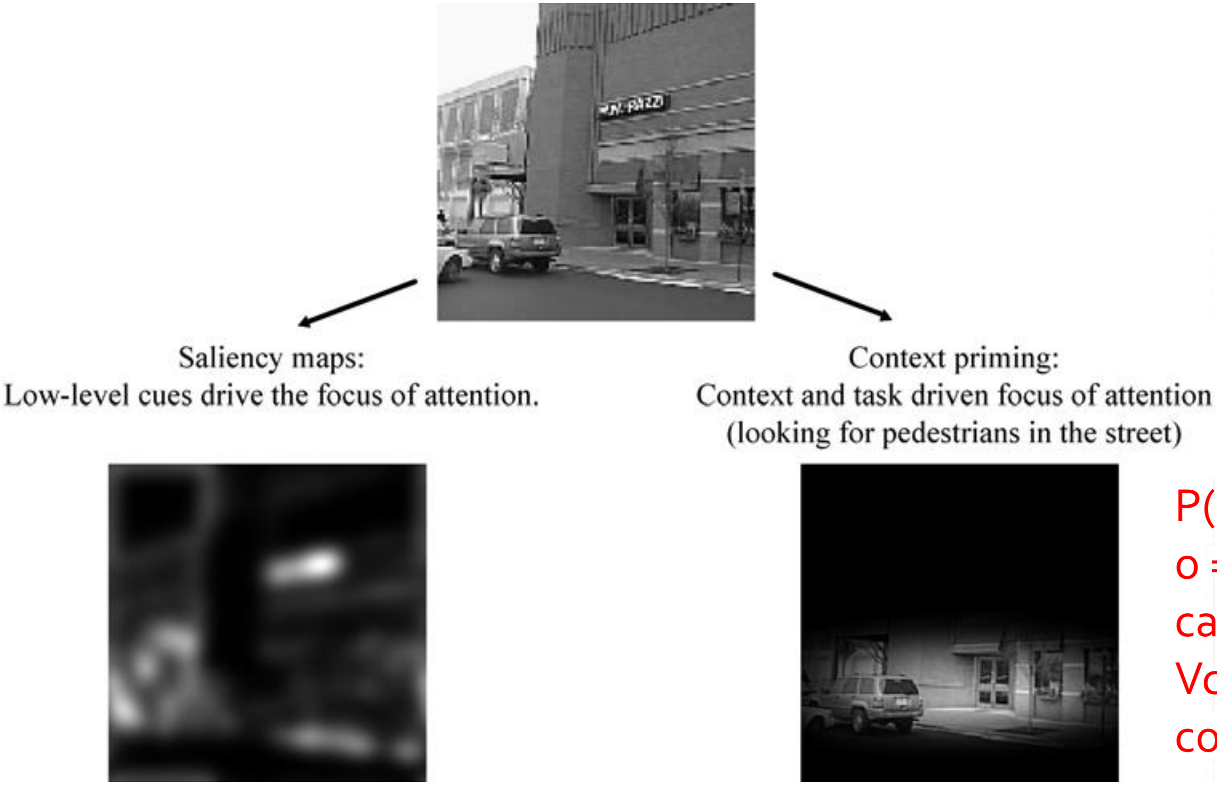
(c) number of gaussians for modeling the PDF *P(o |* ***vC).***

- Objects that have more tight scene-object pair constraints can benefit more from the global context
  - Ex) Cars are almost always found in the street scenes, while people can be found in any scenes

- Using more contextual feature can increase the accuracy
- Using a large number of Gaussians does not really help predicting the object presence
  - Using too many Gaussians will result in the model overfitting the training data, which in turn will result in the lost of generalization ability

Contextual focus of attention

What do we do when we search for a car on the street?



Saliency maps:
Low-level cues drive the focus of attention.

Context priming:
Context and task driven focus of attention
(looking for pedestrians in the street)

P(X | o, Vc)
o = object category (car)
Vc = global context (street)

some specific part of the scene are more likely for certain object to appear.
For example, the pedestrians almost always appear on the street,
not on the middle of the building

Location Prediction

$$P(\mathbf{x}, \mathbf{v}_C \mid o) = \sum_{i=1}^{M} b_i \, G(\mathbf{x}; \mathbf{x}_i, \mathbf{X}_i) \, G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)$$

- We wish to know the probability density function for the given object class and the Vc
  - P(x | o, vc)
  - How can we get this conditional probability?
    - From learning the joint probability P(x,vc|o)

- How is it learned?
  - Using the EM algorithm!

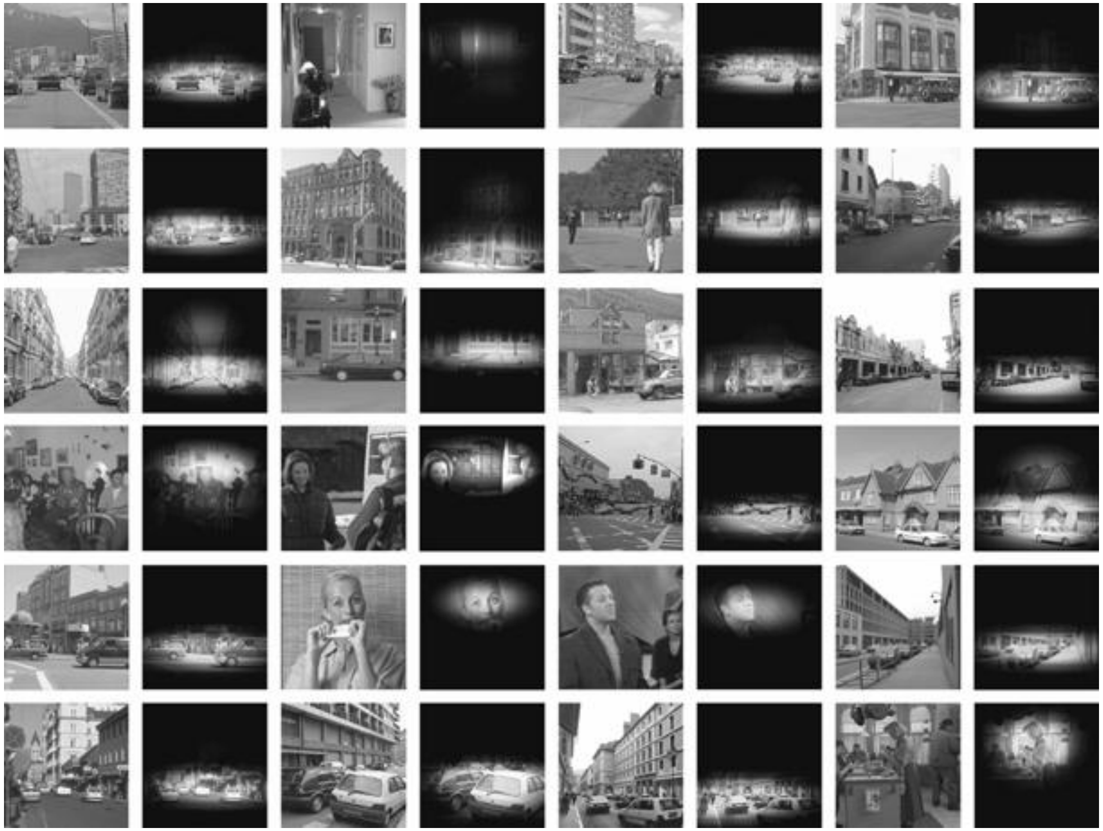
Location Prediction

$$P(\mathbf{x}\,|\,o, \mathbf{v}_C) = \frac{\sum_{i=1}^{M} b_i\, G(\mathbf{x}; \mathbf{x}_i, \mathbf{X}_i) G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i\, G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}$$

$$P(\mathbf{x}, \mathbf{v}_C\,|\,o) = \sum_{i=1}^{M} b_i \underbrace{G(\mathbf{x}; \mathbf{x}_i, \mathbf{X}_i)}_{\text{Distribution of object locations}} \underbrace{G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}_{\text{Distribution of contextual features for each cluster}}$$

- We wish to know the conditional PDF of the location, given the object class and the contextual feature Vc
  - How can we get this conditional probability?
    - From learning the joint probability $P(x,v_c|o)$
  - $P(x,v_c|o)$ is defined as the sum of Gaussian clusters
- How is $P(x,v_c|o)$ learned?
  - Using the EM algorithm!

Location prediction



Estimated center of region

$$(\bar{x}, \bar{y}) = \int \mathbf{x} P(\mathbf{x} \,|\, o, \mathbf{v}_C)\, d\mathbf{x}$$

$$= \frac{\sum_{i=1}^{M} b_i \mathbf{x}_i G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}$$

Estimated width of the region

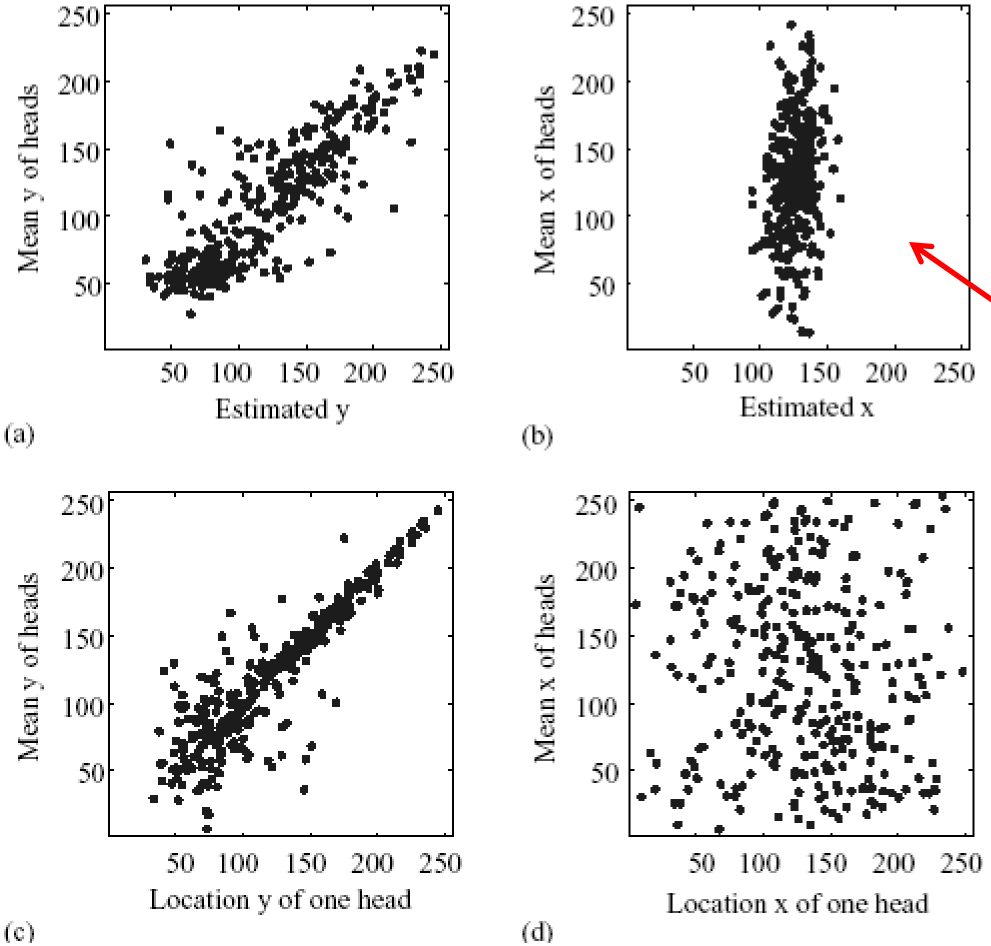
$$\sigma_r^2 = \int r^2 P(\mathbf{x} \,|\, o, \mathbf{v}_C)\, d\mathbf{x}$$

$$r^2 = (x - \bar{x})^2 + (y - \bar{y})^2 \text{ and } \mathbf{x} = (x, y)$$

This visualization is done by performing pixelwise multiplication of the prior map with the grayscale image

Estimated x y vs the mean x and y

Y positions can be predicted well since the global context can tell where the ground level is

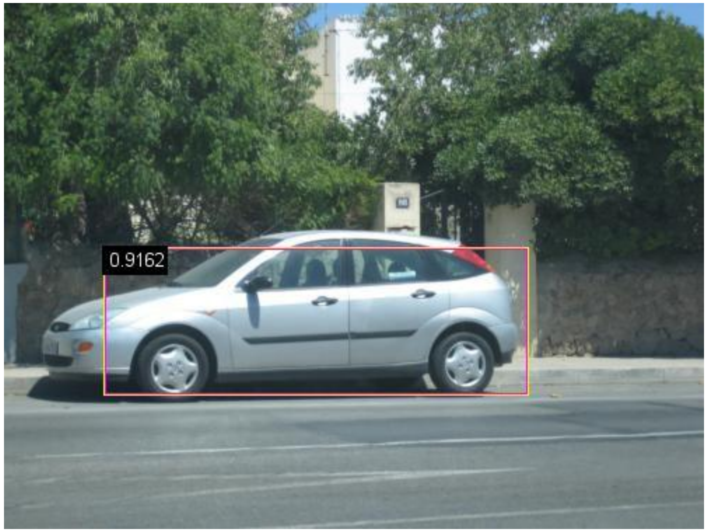


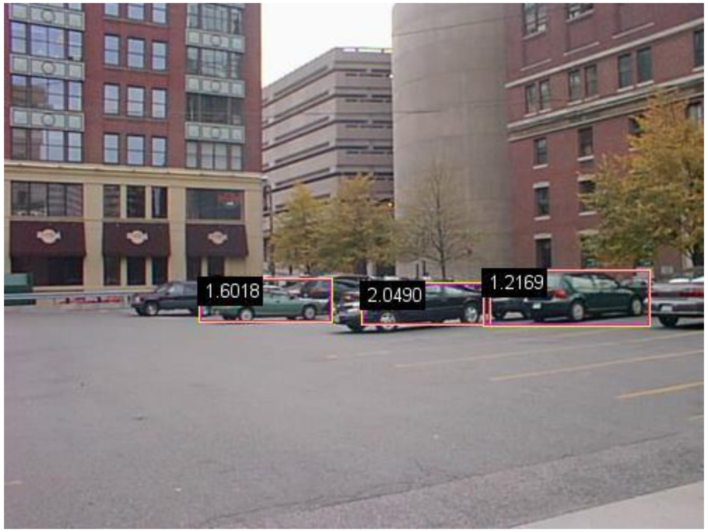
X positions are hard to estimate!

Estimation is clustered at the center, but real object x position appears everywhere

While it was possible to estimate the vertical position of the object correctly, the horizontal position was not estimated well.

Scale selection





- Scale of an object is related to the scene composition
  - Close up scenes have a larger chance of having larger object
  - Complex scenes that contains more of the backgrounds will have a larger chance of having small object

Scale selection

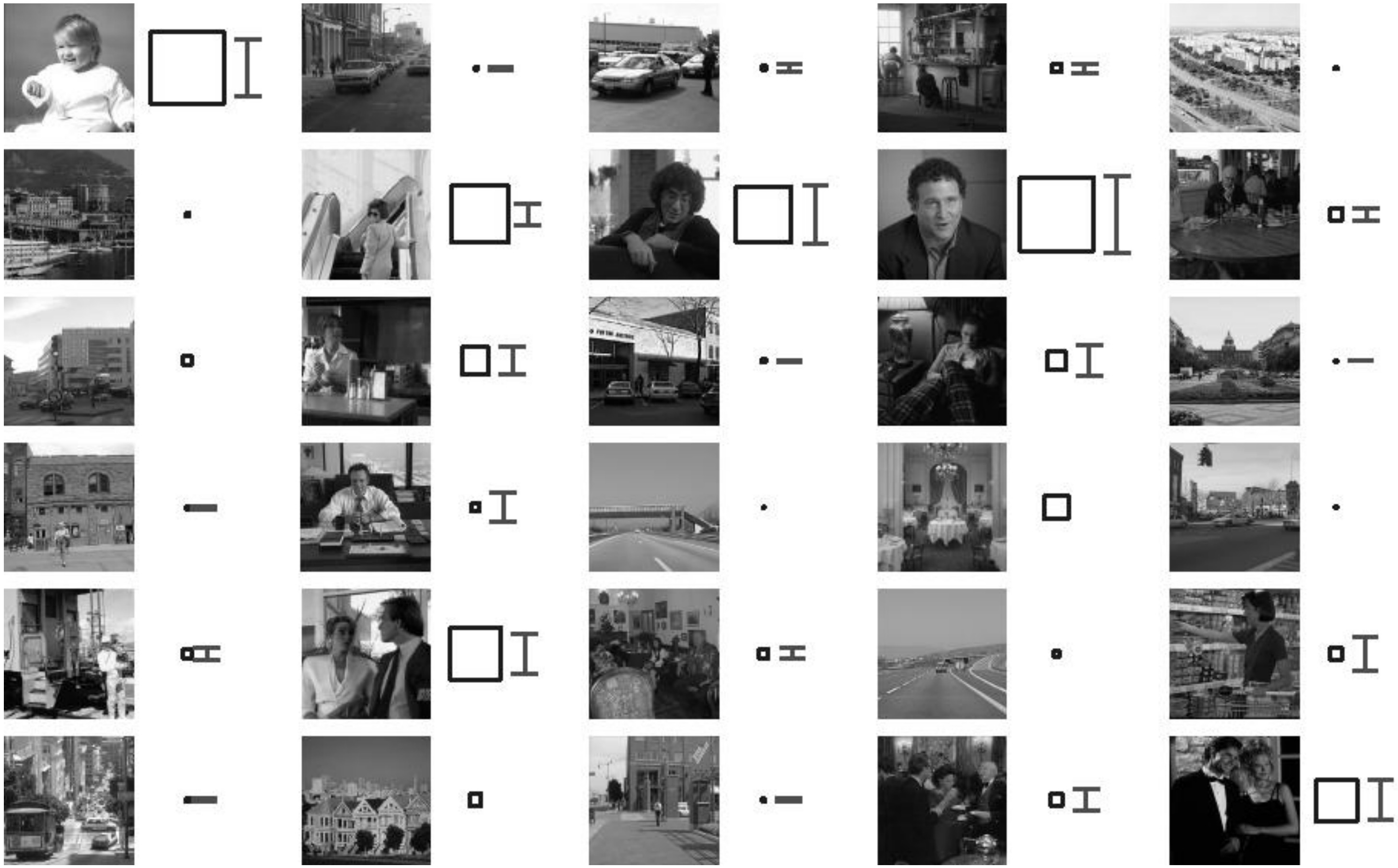
$$P(\sigma \mid \mathbf{x}, o, \mathbf{v}_C) \simeq P(\sigma \mid o, \mathbf{v}_C)$$

$$P(\sigma \mid o, \mathbf{v}_C) = \frac{\sum_{i=1}^{M} b_i \overbrace{G(\sigma; \sigma_i, \mathbf{S}_i)}^{\text{Distribution of object scales}} \overbrace{G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}^{\text{Distribution of contextual features for each cluster}}}{\sum_{i=1}^{M} b_i G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}$$

- The probability distribution of scale σ given the object class o and the contextual feature Vc
  - Assumption : location and scale of an object are independent of one other

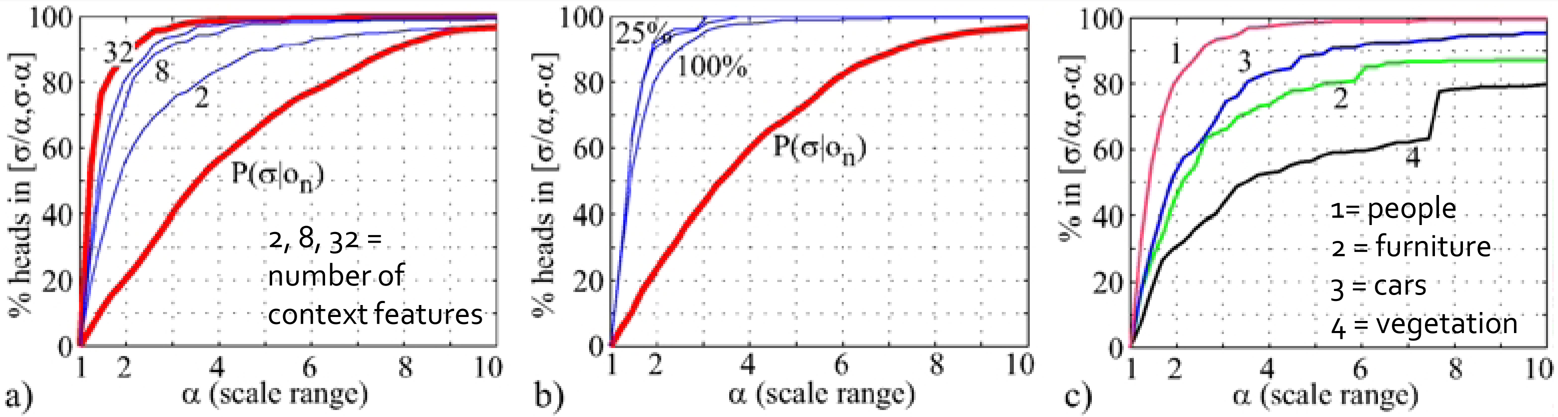
- How is it learned?
  - Using the EM algorithm

Predicted scales using the scale priming



Predicted scale of the objects using the learned prior $p(\sigma|o, v_c)$,
Where o = object category and v_c = context feature.

Predicted object scales



- Scale prediction performance increased with increasing number of contextual features
- Objects with more regular scales can benefit more from the scale prediction using the context
  - Humans have almost similar size while vegetations can vary greatly in size

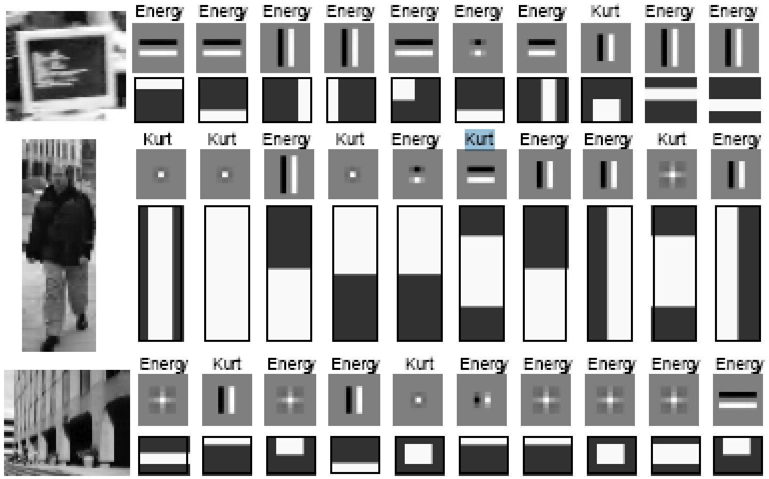
Contextual Priming for object detection

- So how can we use this gained knowledge?
  - We could use it to improve the detection speed.
    - Instead of searching for the whole image, we could only search the area of high priority, at the expense of the some accuracy
  - We could use it to improve the detection accuracy
    - We could put more weights on the primed area
    - Or we could incorporate the conditional probability P(X|O) in the score to improve the accuracy.

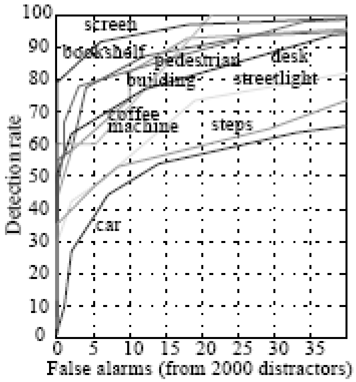
Contextual Priming for object detection

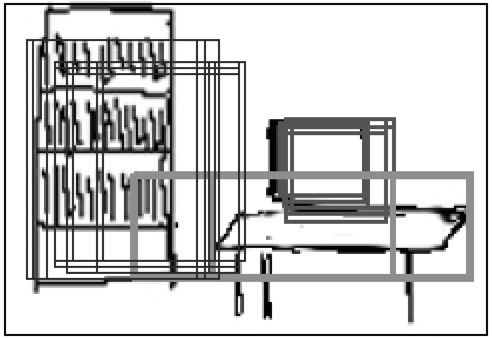
- Related papers
  - Using the forest to see the trees: a graphical model relating features, objects and scenes
    P. Murphy, A. Torralba and W. T. Freeman
    Adv. in Neural Information Processing Systems 16 (NIPS), Vancouver, BC, MIT Press, 2003.
  - Object detection and localization using local and global features
    K. Murphy, A. Torralba, D. Eaton, W. T. Freeman, 2006
  - Contextual Guidance of Attention in Natural scenes: The role of Global features on object search
    A. Torralba, A. Oliva, M. Castelhano and J. M. Henderson
    Psychological Review. 2006.

Using the forest to see the trees [Murphy 2003]

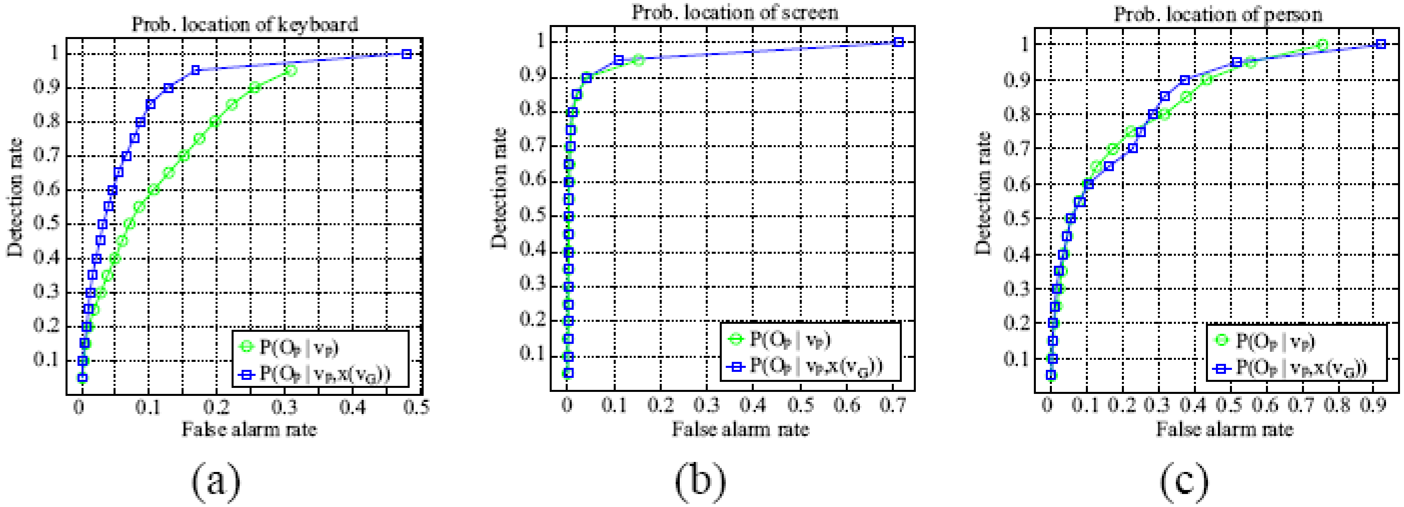

Energy Energy Energy Energy Energy Energy Energy Kurt Energy Energy
Kurt Kurt Energy Kurt Energy Kurt Energy Energy Kurt Energy
Energy Kurt Energy Energy Kurt Energy Energy Energy Energy Energy

- The local detector is a binary classifier trained using Gentleboost
  - Similar to the Viola-Jones detetor
  - Gentleboost gives much higher performance than Adaboost, and requires fewer iterations
  - Box filters are used to train the cascade of weak classifiers



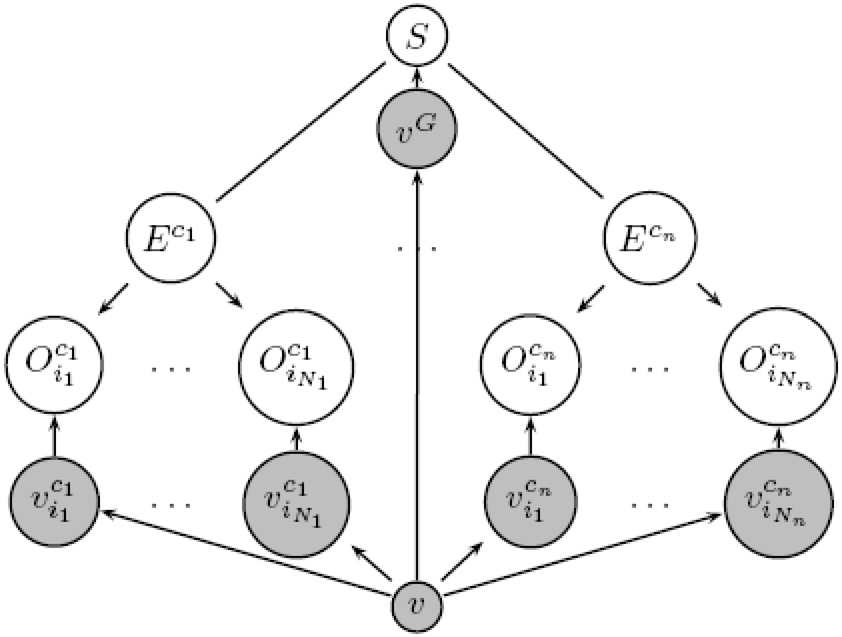


Using the forest to see the trees [Murphy 2003]



- How can we improve the speed and accuracy of a detector?
  - The detector can be run only on the location and scale with high probability, but this can Risk missing objects
  - Instead, the object is run everywhere, but detections that are far from the predicted location/scale are penalized
- Objects that do not have a uniform appearance model benefit more from the use of global context
  - Keyboards

Using the forest to see the trees [Murphy 2003]



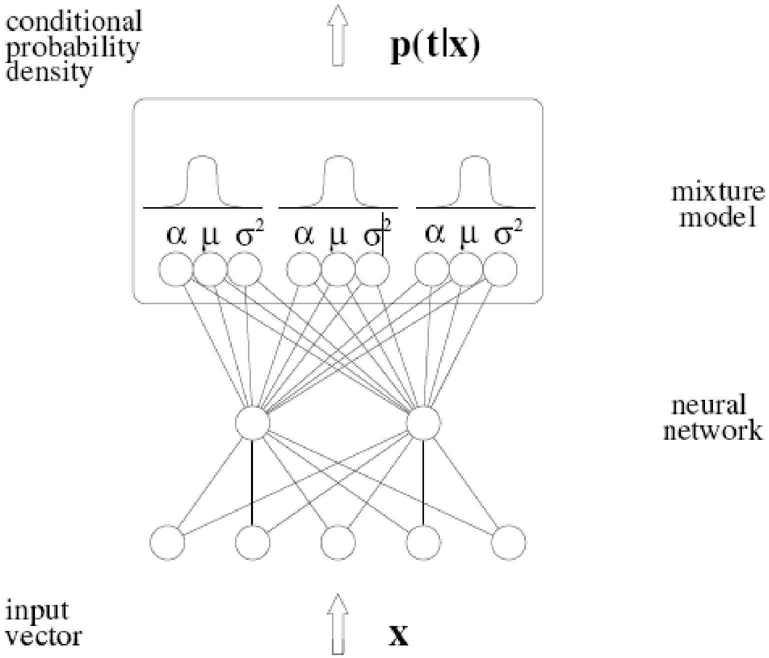
- By the holistic structured model, the object detection and scene classification is done simultaneously
  - Result of the scene recognition helps object recognition
  - Result of the object recognition helps scene recognition
  - "The whole influencing the part, and the part influencing the whole"

Object detection and localization using local and global features [Murphy 2006]

- Basically, the idea is the same, but there is some change in the methodology
  - Used steerable pyramids to describe global context
    - [Tor 2003] uses Gabor filters
  - P(X|G) is modeled directly (discriminative approach), using the Mixture Density Network (MDN)
    - For location priming, the conditional probability P(X = x, y, s | G) (x = x coordinate, y = y coordinate, G = gist feature) should be learned.
    - [Tor 2003] uses weighted regression to present p(X, G)
  - The primed priors are incorporated to improve the detection accuracy

Object detection and localization using local and global features [Murphy 2006]

Outputs the conditional probability distribution of (presence, location, scale) given the feature vectors



Feature vectors (GIST) are fed as an input to the neural network

- Mixture Density Network
  - Combination of a mixture model and a neural network
  - Parameter estimation is done by training the neural network.
  - Can model the conditional probability of P(t|x) directly

Object detection and localization using local and global features [Murphy 2006]

- Local Detectors are trained using boosting and box filters
- How can we combine the local and global estimates to improve the detection accuracy?
  - Estimates based on local and global features are combined using the "product of experts" model

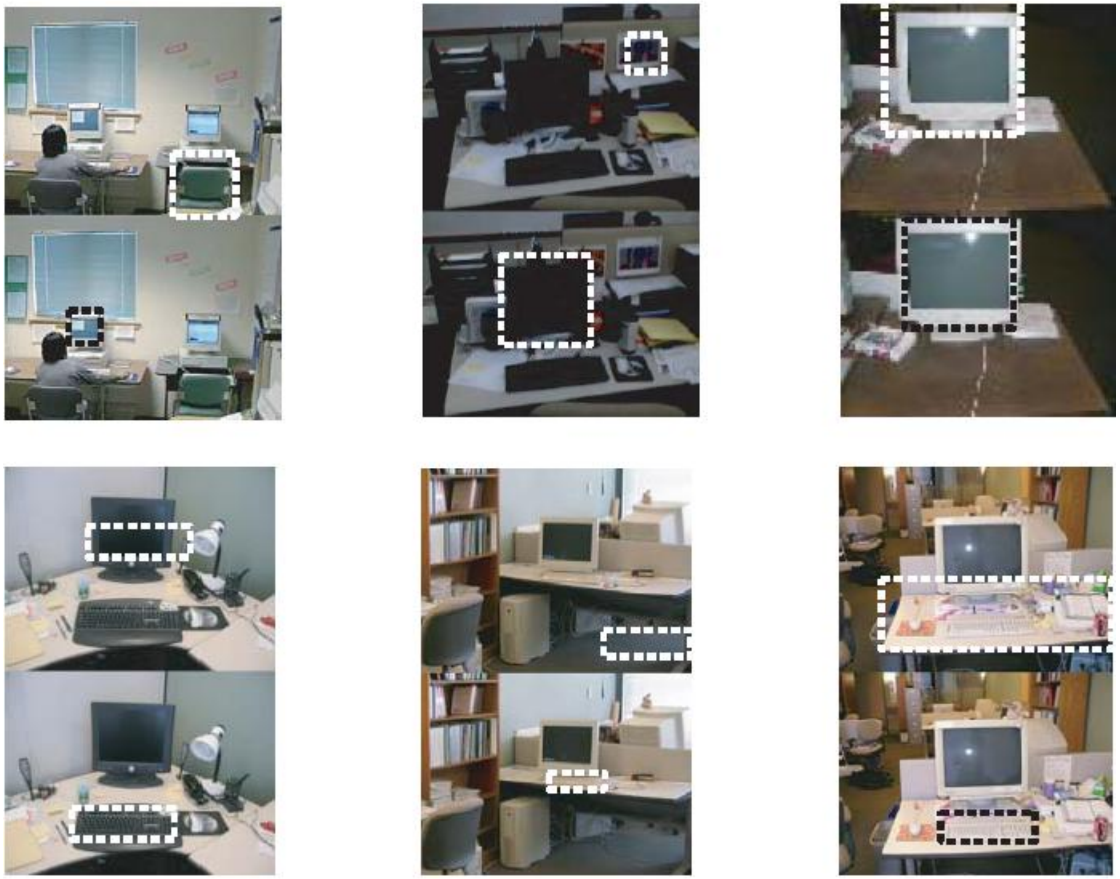
$$P(X = i|L, G) = \frac{1}{Z} P(X = i|L)^{\gamma} P(X = i|G)$$

We use the product of both the local estimation and the global estimation

The accuracy is improved as the global estimate can correct many false positives.

This exponent is used to balance the relative confidence of two detectors
It is learned offline and fixed

Object detection and localization using local and global features [Murphy 2006]



Location and scale is fixed by the global estimate to improve the detection accuracy

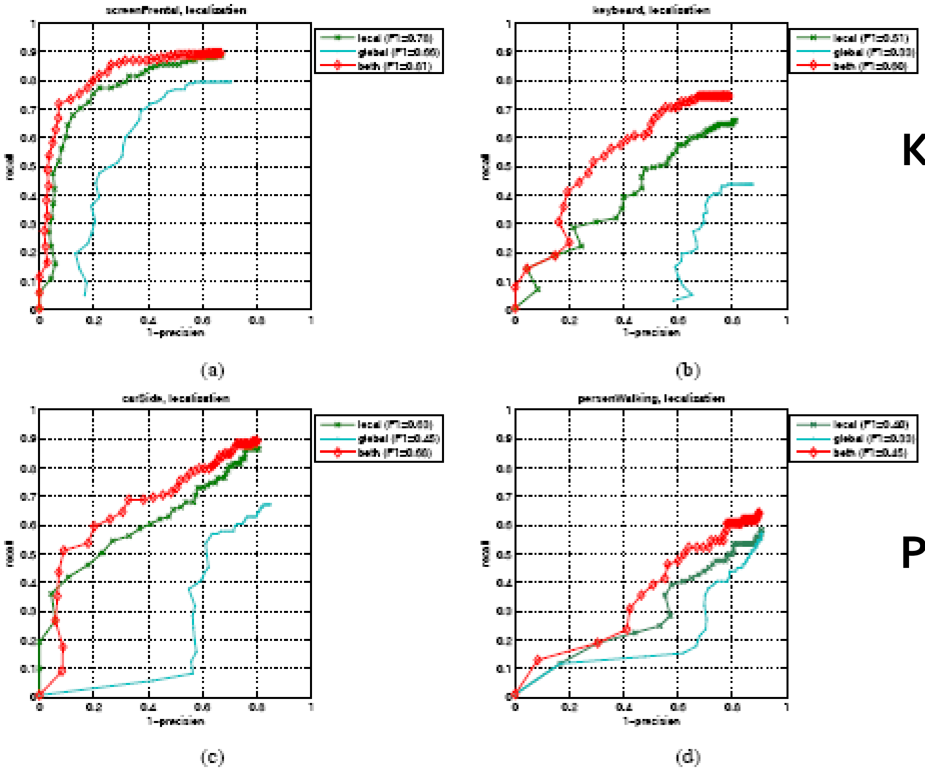
Object detection and localization using local and global features [Murphy 2006]

Screen Frontal

Keyboard

Car Side

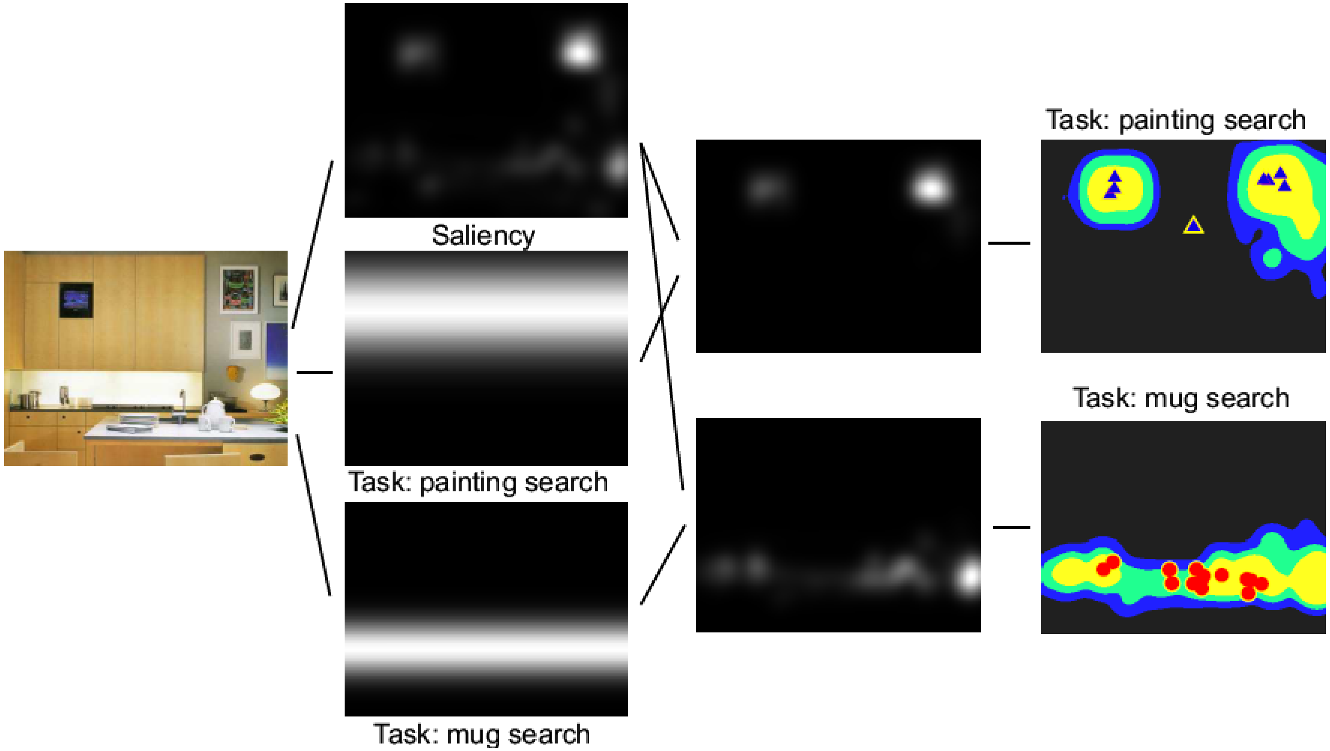
Person Walking



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

Fig. 12. Localization performance for (a) screens, (b) keyboards, (c) cars, (d) pedestrians. Each curve shows precision-recall curves for $P(X|G)$ (bottom blue line with dots), $P(X|L)$ (middle green line with crosses), and $P(X|L,G)$ (top red line with diamonds). We see significant performance improvement by using both global and local features.

Contextual Guidance of Eye Movements and attention in Real-World Scenes [Torralba 2006]

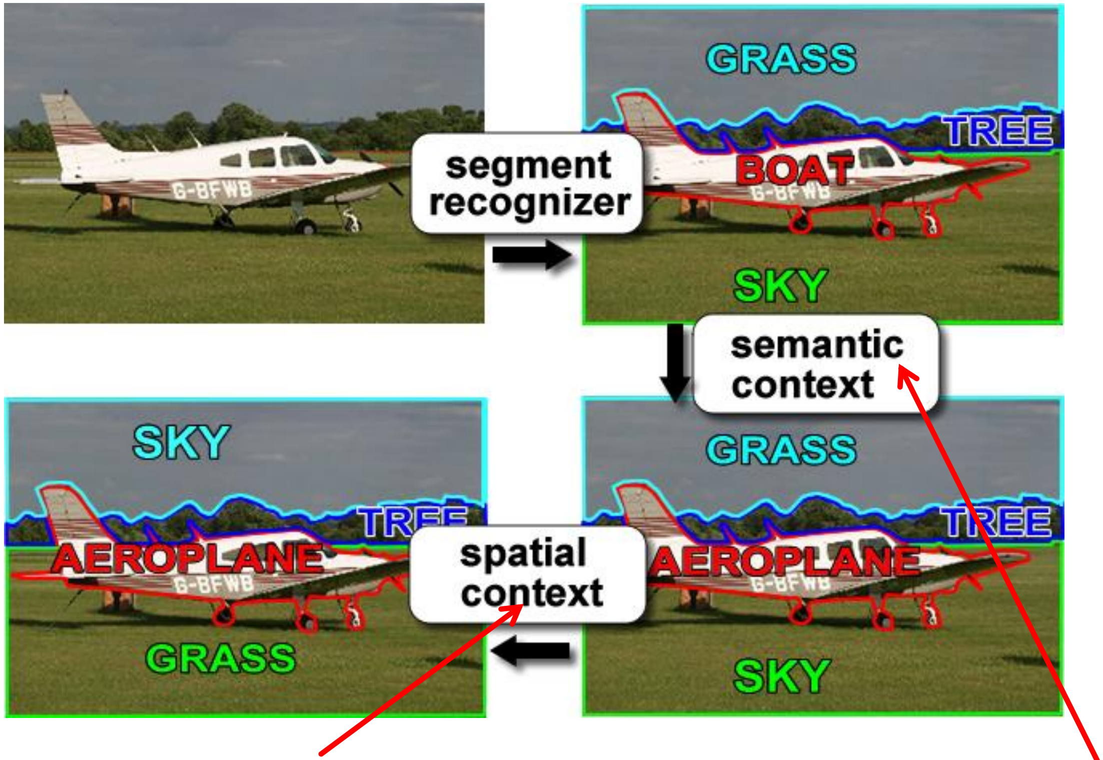


- Combined the low-level saliency map with the high-level task-specific prior to find the most likely positions for the object
  - The resulting prior agrees with the human eye movement analysis on the detection task

Carolina Galleguillos, Andrew Rabinovich, Serge Belongie [Galleguillos 2008]

Object Categorization using Co-Occurrence, Location and Apperance [CVPR 2008]

Object Categorization using Co-Occurrence, Location and Apperance

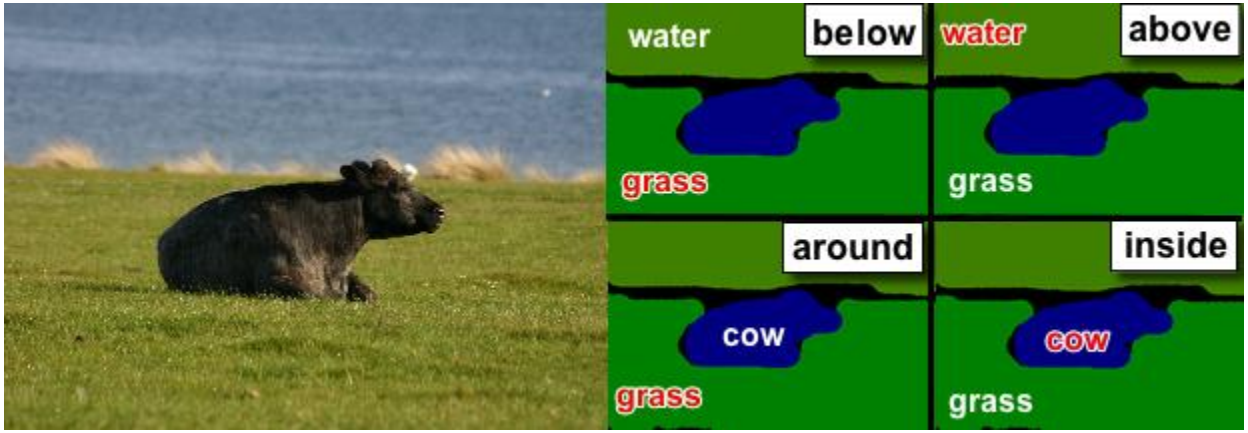


Again, spatial context is used to force spatial restrictions among objects in the scene

Semantic context (Co-Occurrence) is used to correct the result of object categorization

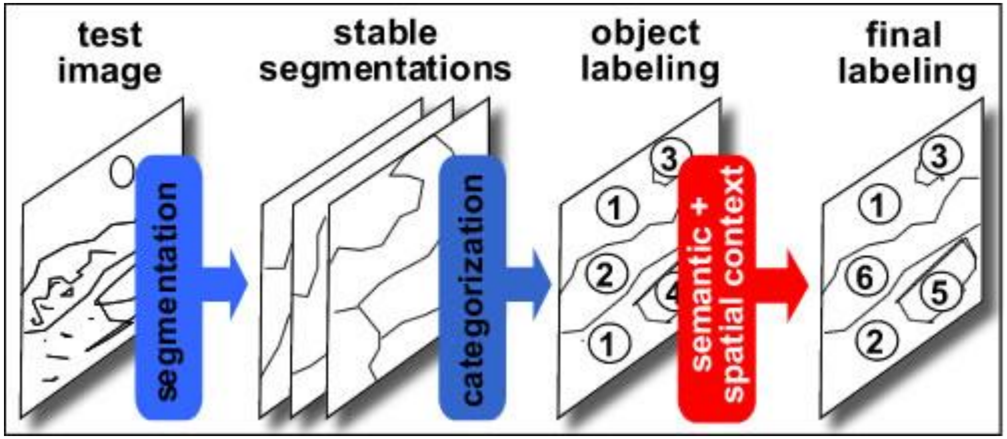
- Abstract
  - Goal
    - Use of context in object categorization
  - The type of context used
    - Co-Occurrence
    - Relative Location

Four basic spatial relationships



- Four spatial relationships are used
  - Below : Object A appears below Object B
    - Grass is below the water
  - Above : Object A appears above Object B
    - The water is above the grass
  - Around : Object A appears around Object B
    - The grass is around the cow
  - Inside : Object A appears inside Object B
    - The cow is inside the grass

The process of object categorization using semantic and spatial context



- The categorization process
  - 1) The test image is segmented
  - 2) The segmented parts are then categorized using only the local statistics
  - 3) Co-occurrence and spatial context are used to fix the labels

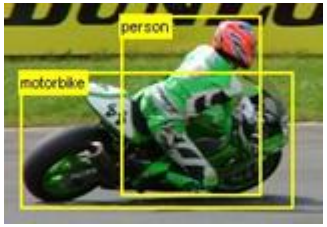
Simple pairwise feature

Difference between y component of the centroids of the objects ci, and cj

Overlap percentage of the object cj with respect to the object ci

$$F_{ij} = (\mu_{ij}, O_{ij}, O_{ji})^{\top} \quad \forall i, j \in \mathcal{C},\ i \neq j, \tag{1}$$

$$O_{ij} = \frac{\beta_i \,/\, \beta_j}{\beta_i} \text{ and } \mu_{ij} = \mu_{yi} - \mu_{yj} \tag{2}$$

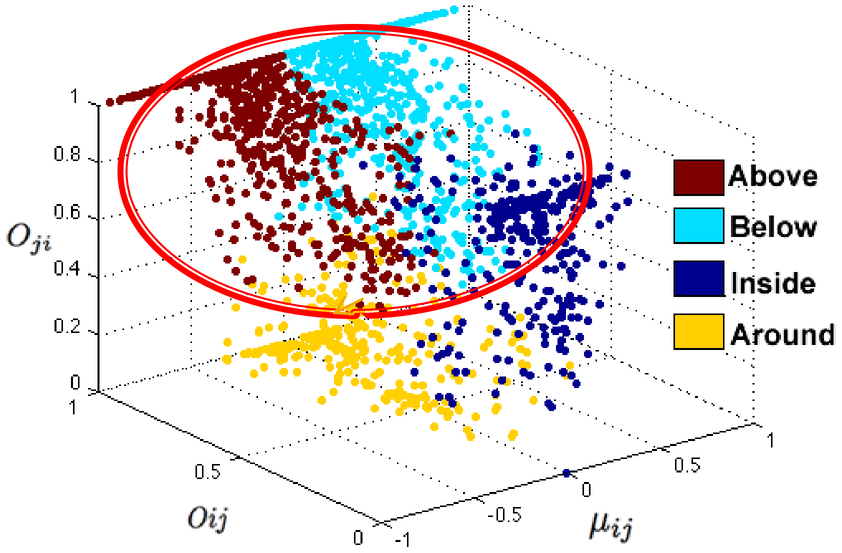


If uij < 0, then the object I is above the object j
If uij > 0, then the object I is below the object j

When Oij is very small while Oji is very large then object i is inside the object j
For the opposite,
the object i is around the object j

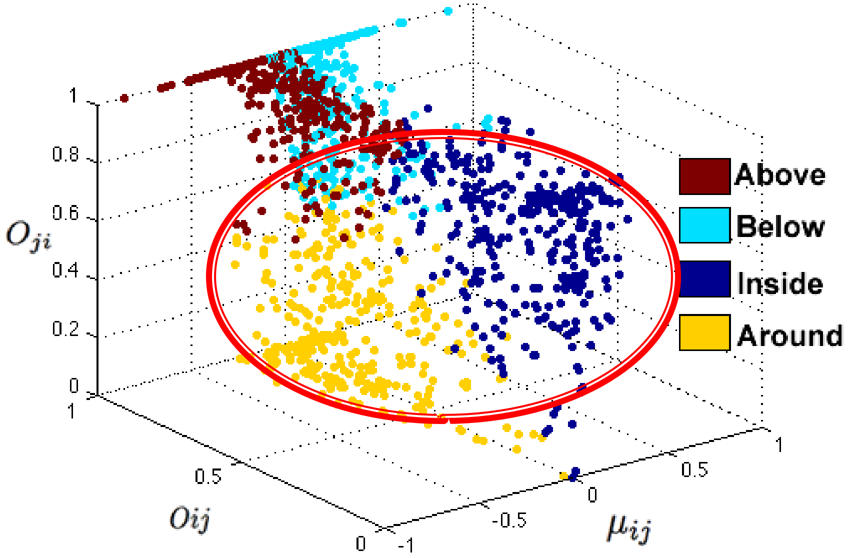
Distribution of the dataset

MSRC DATASET



More vertical relationship

PASCAL 2007 DATASET



More overlapping relationship

(Object labels are specified by Boudning boxes)

Context Model

Number of class labels and the segments are the same = labeling problem

Class labels

Segments

$$p(c_1 \ldots c_k | S_1 \ldots S_k) = \frac{B(c_1 \ldots c_k) \prod_{i=1}^{k} p(c_i | S_i)}{Z(\phi_0, \ldots \phi_r, S_1 \ldots S_k)},$$

r = spatial relation (inside, below, around, above)

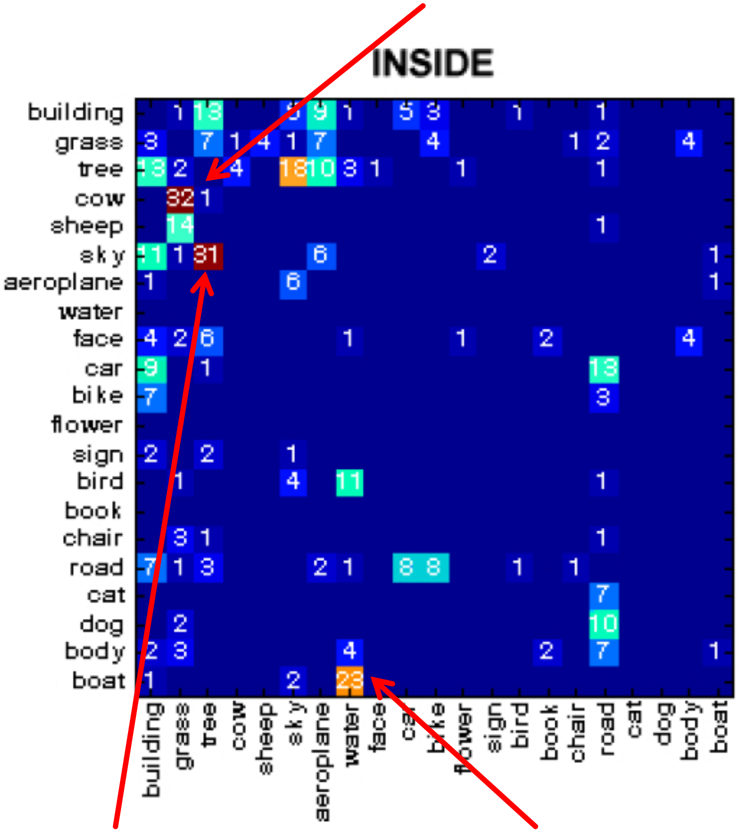
$$\text{with } B(c_1 \ldots c_k) = \exp\left(\sum_{i,j=1}^{k} \sum_{r=0}^{q} \alpha_r \phi_r(c_i, c_j) \right),$$

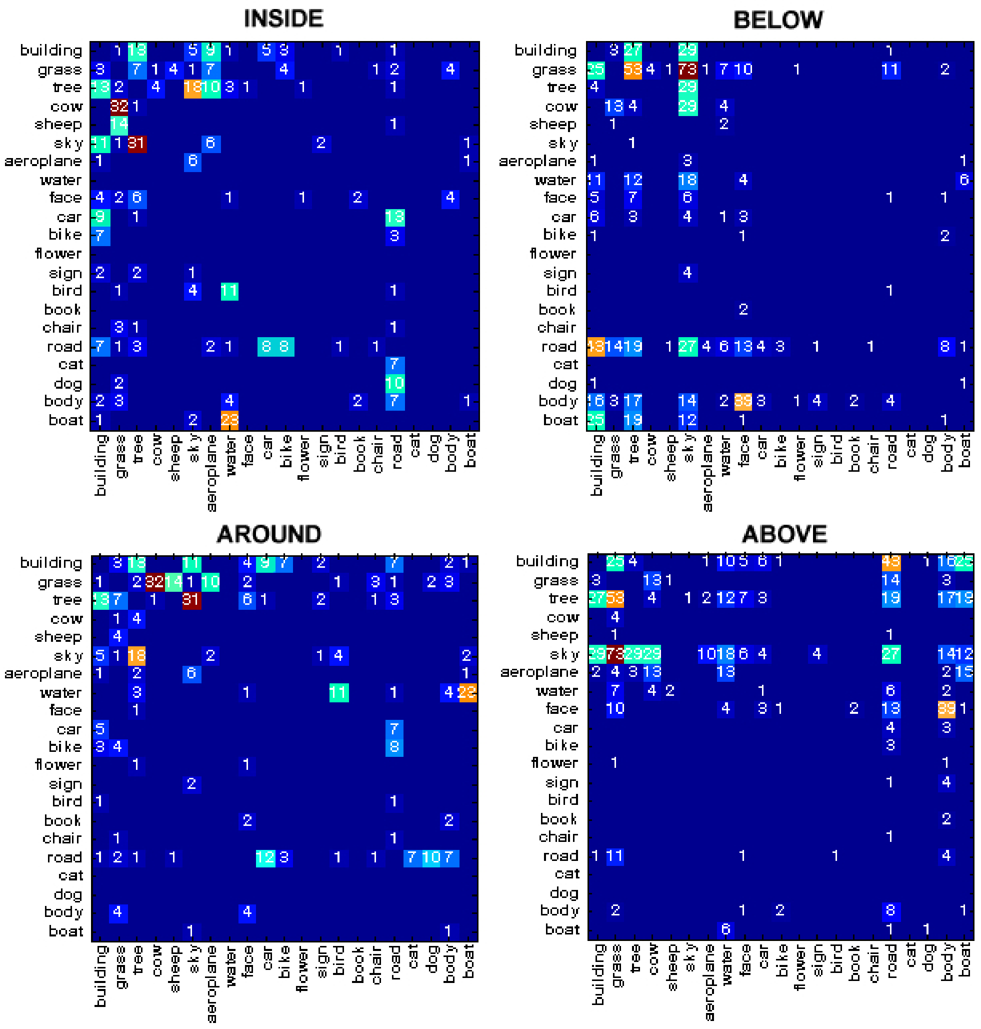
Parameter estimated from the training data

- Given a set of segmentation for each image, we wish to know the class label for each segmentation
  - CRF (Conditional Random Field) is used to learn this conditional probability distribution
  - Phi is the co-occurrence matrix for each spatial relation (r = 1, 2, 3, 4)

Frequency Matrices (co-occurrence matrices) for the MSRC database

The number of times the object in the row (cow) is INSIDE the object in the column (grass) : 32 times

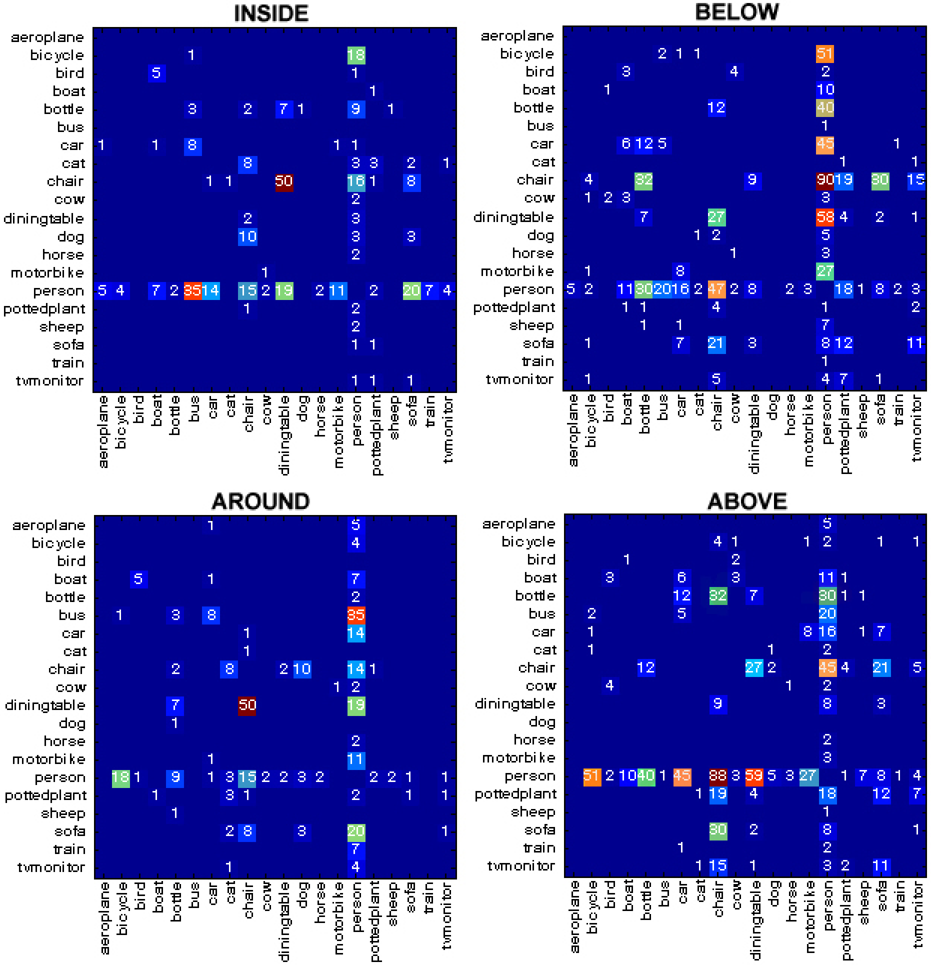




The sky appeared inside the tree for 31 times

The boat appeared inside the water for 23 times

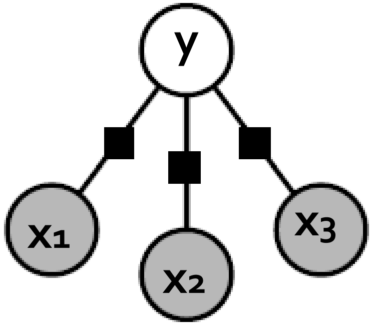
Frequency Matrices (co-occurrence matrices) for the PASCAL database



- Notable spatial relations
  - A person appears inside a bus
  - A chair appears below a person
  - A diningtable appears around a chair
  - A person appears above a bicycle
- So how can we use these co-occurrence statistics gained?
  - By using it in learning a CRF

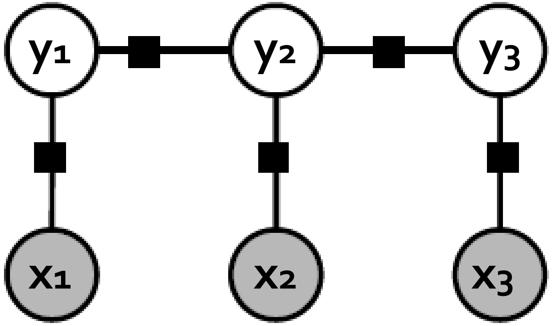
Conditional Random Field

Class label



Input vector

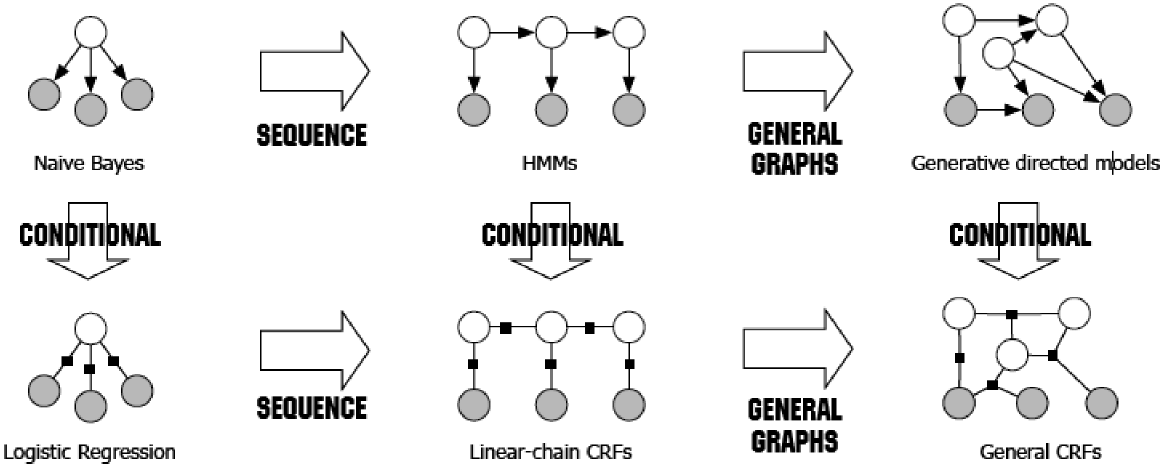
Label sequence
(ex: POS tags for each word)



Input sequence (ex: sentence)

- Classification
  - We wish to know the single class label y for the input vector
    - Conditional probability $P(y|\mathbf{x})$ should be modeled
- Sequence Labeling
  - We wish to know the labels for the input sequence x
    - $P(\mathbf{y}|\mathbf{x})$
  - There are dependencies among the labels

Conditional Random Field



- CRF is a discriminative alias for the generative directed models
  - Focused on modeling the conditional PDF p(y|x)
    - Without modeling p(x) that has complex dependencies
  - Uses iterative learning method similar to gradient descent learning
  - Linear-chain CRFS are used extensively for the labeling tasks in NLP such as in part-of-speech tagging, and named entity recognition.
  - The CRF model used in this paper is a more general CRF with fully connected dependencies between all segment labels

Conditional Random Field

Linear chain conditional random field

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left\{ \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}$$

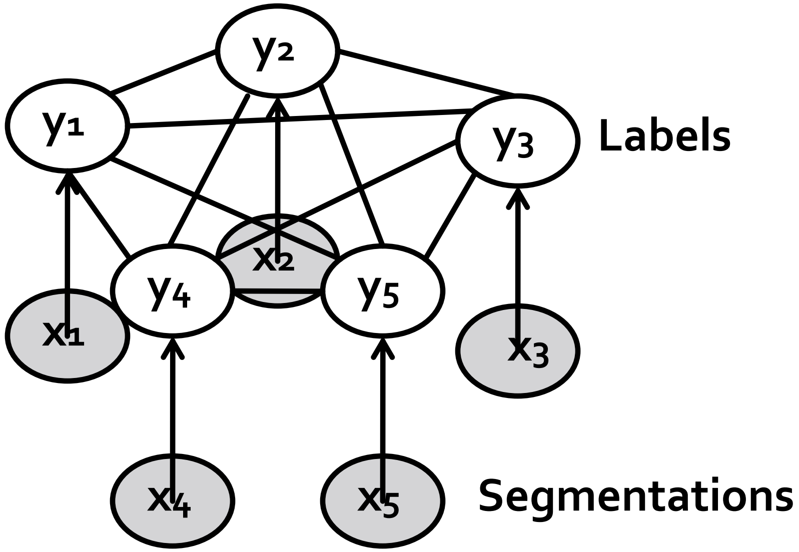
Ensures the p(y|x) to be positive

Instance Specific Normalization function Ensures the sum of p(y|x) to be 1

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp \left\{ \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}$$

- The state y_{t-1} is determined by the previous state y_t and the input x_t,for the linear chain CRF
- For more general CRFs, we can have different feature functions

Conditional Random Field used in the object categorization task



- The CRF model used in this paper has a fully-connected dependencies between all segment labels
  - Every labels have co-occurrence frequency and spatial relation with other labels

Integration of the contextual model into the CRF

$$\phi_0(c_i, c_j) = \phi'(c_i, c_j) + \sum_{k=1}^{|\mathcal{C}|} \phi'(c_i, c_k)$$

Sum of all four matrices
For each pair of classes

Presence of the label i

Parameter estimated from the training data

$$p(l_1 \dots l_{|\mathcal{C}|}) = \frac{1}{Z(\phi)} \exp\left(\sum_{i,j \in \mathcal{C}} \sum_{r=0}^{q} l_i l_j \cdot \alpha_r \cdot \phi_r(c_i, c_j) \right)$$

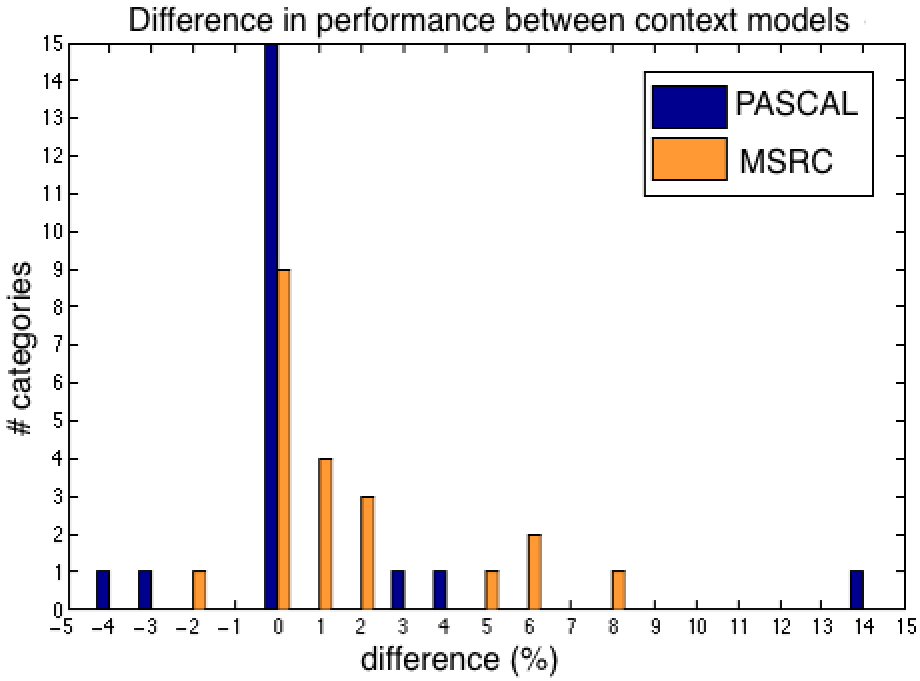
Label function (if the object i exist then 1, and 0 otherwise)

Partition function

Frequency matrix for the given label pair and the spatial relation

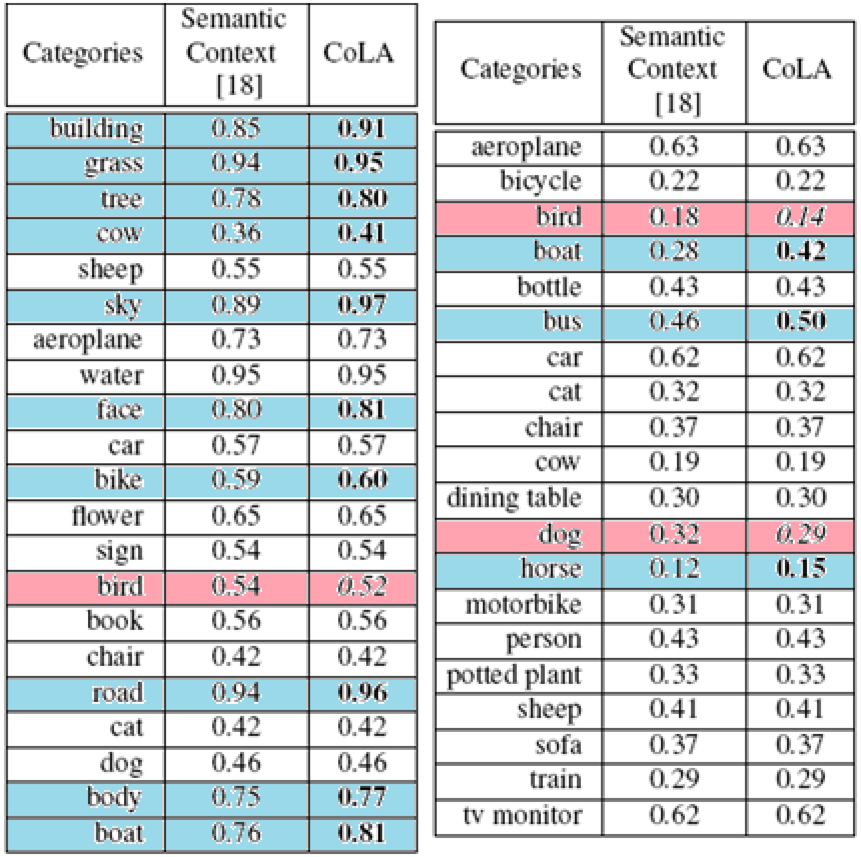
- We wish to find a ϕ that maximize the log-likelihood of the observed label co-occurrence
  - Evaluating the partition function is intractable
  - The partition function is approximated using the Monte Carlo integration

Performance improvement



- Accuracy of only few categories were worsened with the introduction of contexts.
  - Most are unchanged or improved

Performance improvement



Categories	Semantic Context [18]	CoLA
building	0.85	**0.91**
grass	0.94	**0.95**
tree	0.78	**0.80**
cow	0.36	**0.41**
sheep	0.55	0.55
sky	0.89	**0.97**
aeroplane	0.73	0.73
water	0.95	0.95
face	0.80	**0.81**
car	0.57	0.57
bike	0.59	**0.60**
flower	0.65	0.65
sign	0.54	0.54
bird	0.54	*0.52*
book	0.56	0.56
chair	0.42	0.42
road	0.94	**0.96**
cat	0.42	0.42
dog	0.46	0.46
body	0.75	**0.77**
boat	0.76	**0.81**

Categories	Semantic Context [18]	CoLA
aeroplane	0.63	0.63
bicycle	0.22	0.22
bird	0.18	*0.14*
boat	0.28	**0.42**
bottle	0.43	0.43
bus	0.46	**0.50**
car	0.62	0.62
cat	0.32	0.32
chair	0.37	0.37
cow	0.19	0.19
dining table	0.30	0.30
dog	0.32	*0.29*
horse	0.12	**0.15**
motorbike	0.31	0.31
person	0.43	0.43
potted plant	0.33	0.33
sheep	0.41	0.41
sofa	0.37	0.37
train	0.29	0.29
tv monitor	0.62	0.62

Table 1. Comparison of recognition accuracy between the models for MSRC and PASCAL categories. Results in **bold** indicate an increase in performance by our model. A decrease in performance is shown in *italics*.

- For almost all categories, the use of spatial relation improved, or unchanged the scene categorization accuracy of using the co-occurrence alone
  - Average Categorization Accuracy
    - 68.38% MSRC
    - 36.7 % for PASCAL
- The categories that gained high accuracy are ones that have some strong spatial relations with other categories
  - A horse appears under a person
  - A sky appears above almost all categories
  - A bird does not have strong relationship with its environment

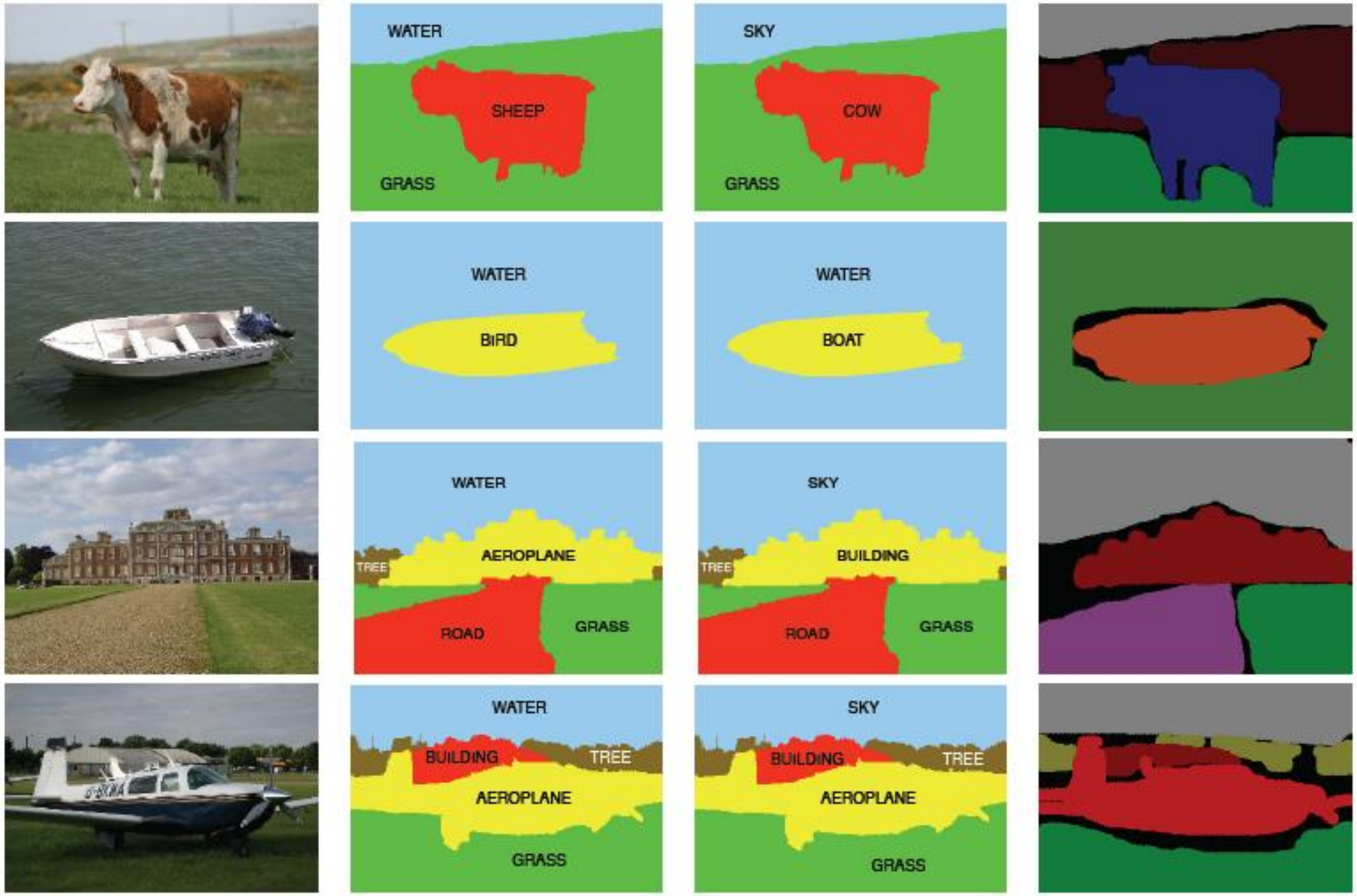
Positive results on the MSRC dataset

input image

co-occurence

co-occurrence
+ spatial relation

segmentation
Result



Water -> sky
(The sky should be
above everything)c

Bird->boat
(The water should be
around a boat, not
around a bird)

Aeroplane ->building
Water->sky
(A building is more
likely to be above the
grass and roads)

Water->sky
(The sky should be
above everything)

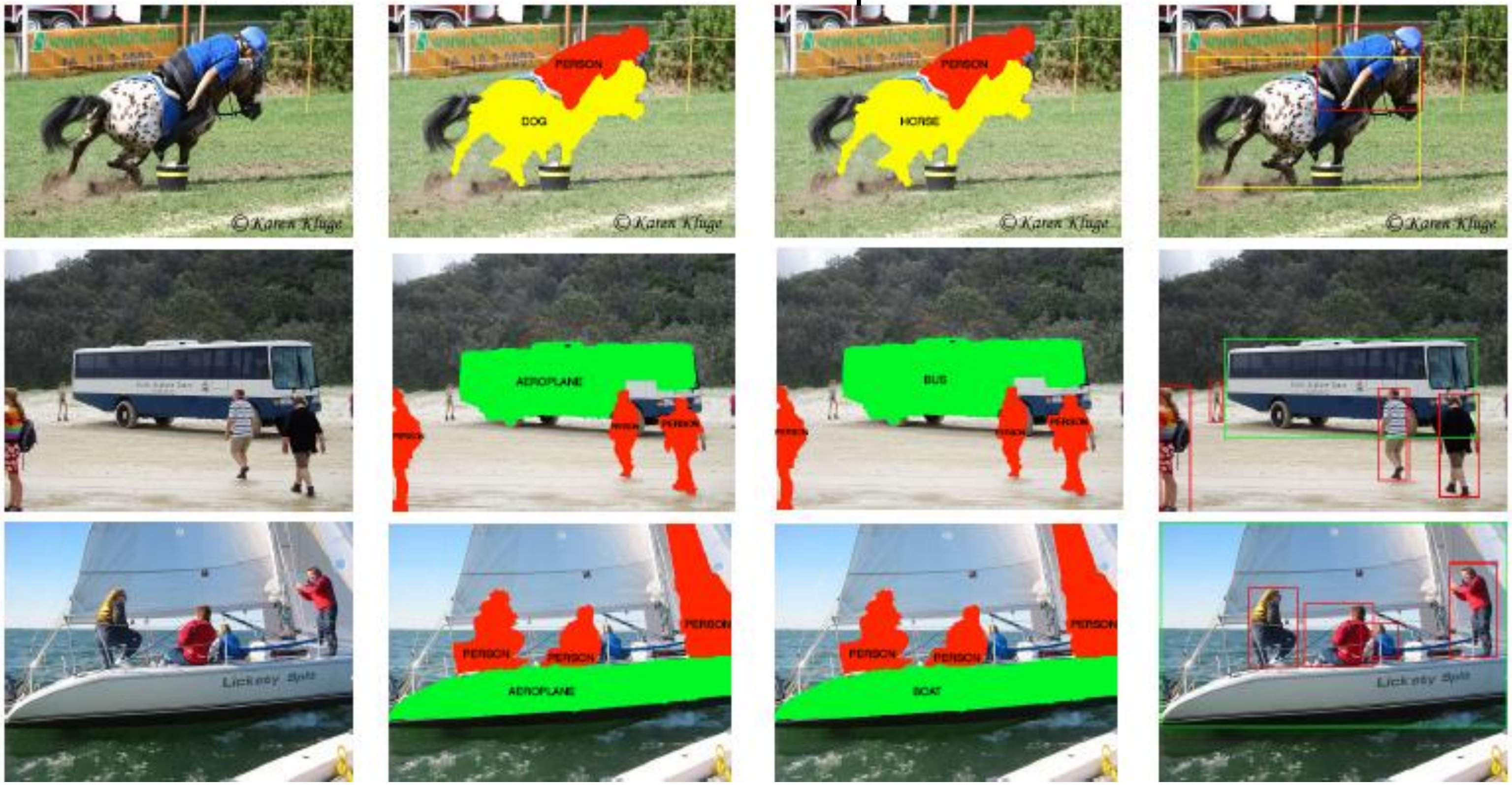
Positive results on the PASCAL 2007 dataset

input image

co-occurence

co-occurrence
+ spatial relation

Bounding boxes

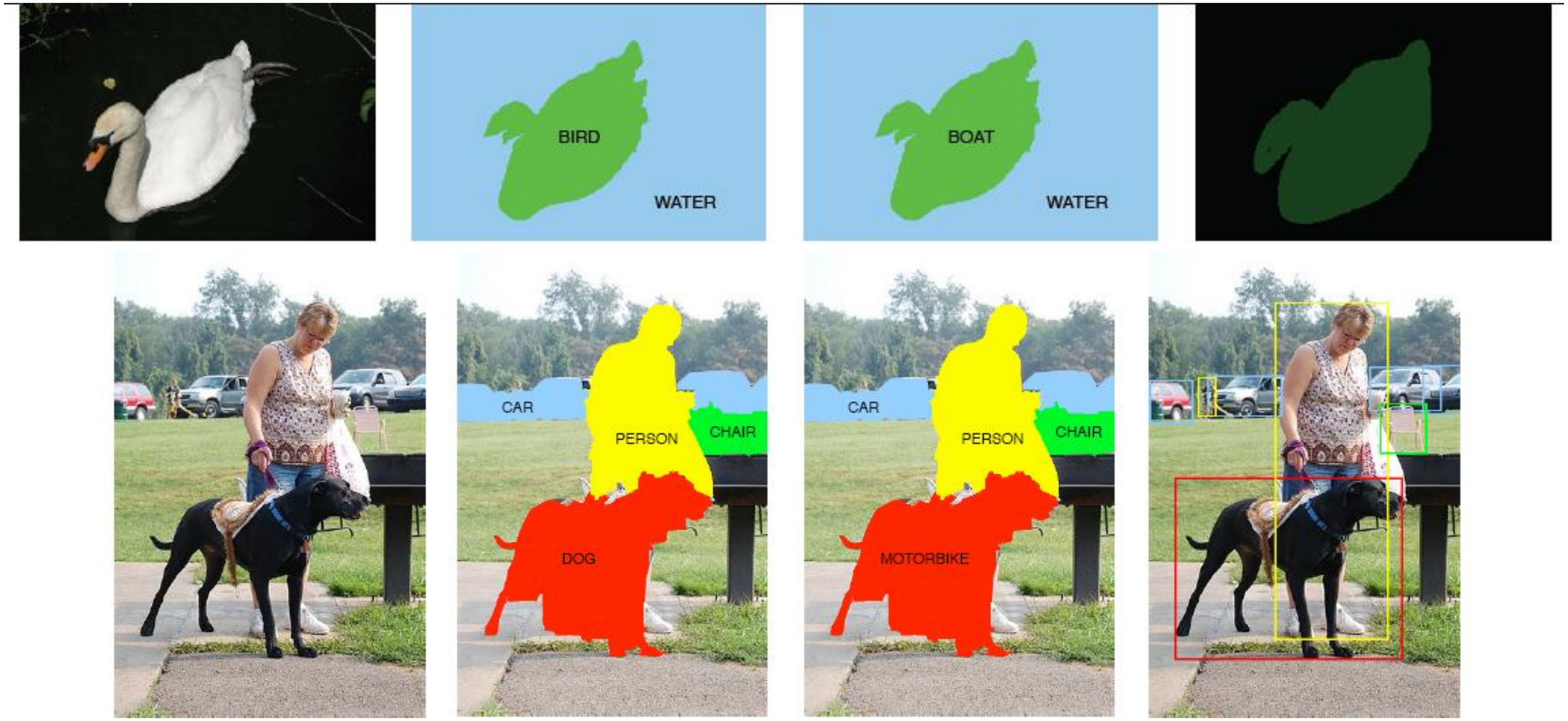


Dog->horse
(above, below)

Aeroplane->bus
(above, below)

Aeroplane->boat
(inside, around)

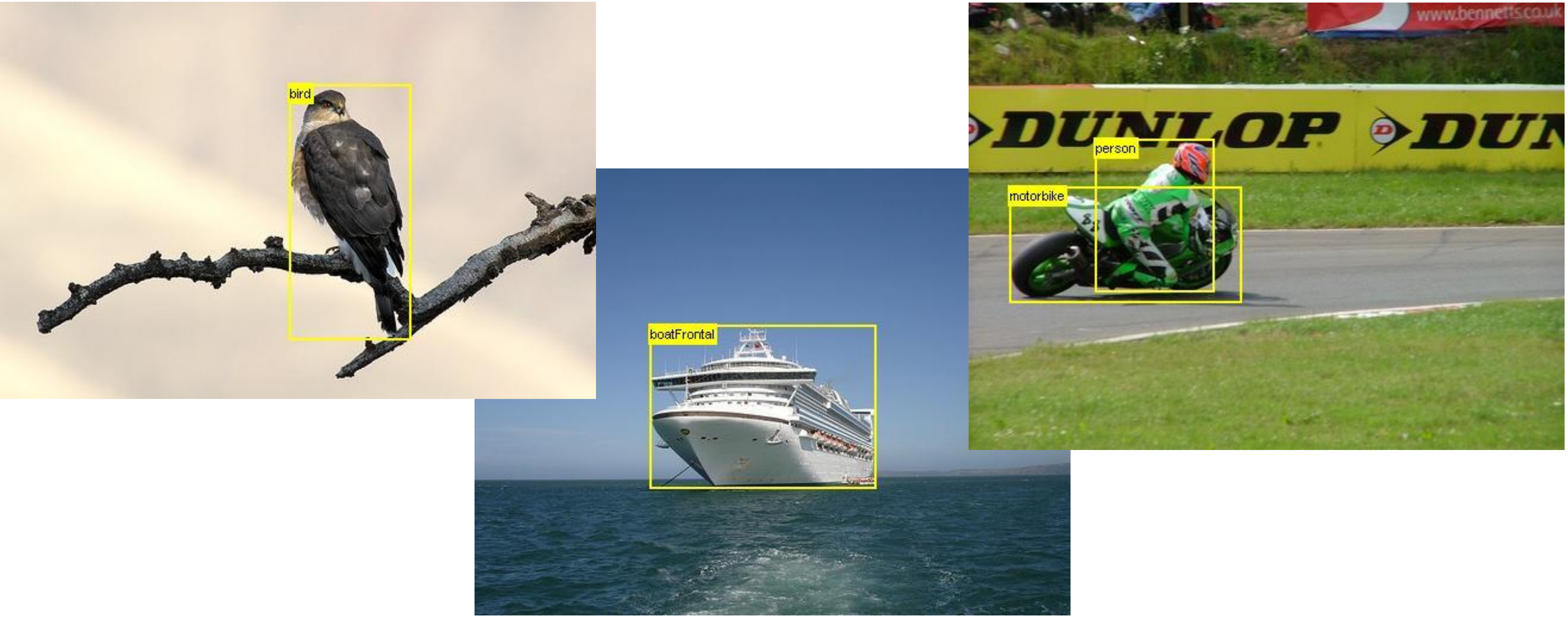
Using the spatial context can sometimes mislead to wrong categorization



Why did these fail?

1) A boat is more likely to appear inside water than a bird which usually appear in the sky.
   -> But it is a duck…
2) A person is more likely to appear on top of the motorbike, than on top of a dog.
   -> The person is not actually above the dog, while it appears so.

Using the spatial context can sometimes mislead to wrong categorization



Why did these fail?

1) The boast is more likely to appear inside water than the bird which usually appear in the sky.
   -> But it is a duck…
2) person is more likely to appear on top of the motorbike, than on top of the dog.
   -> The person is not actually above the dog, while it appears so.

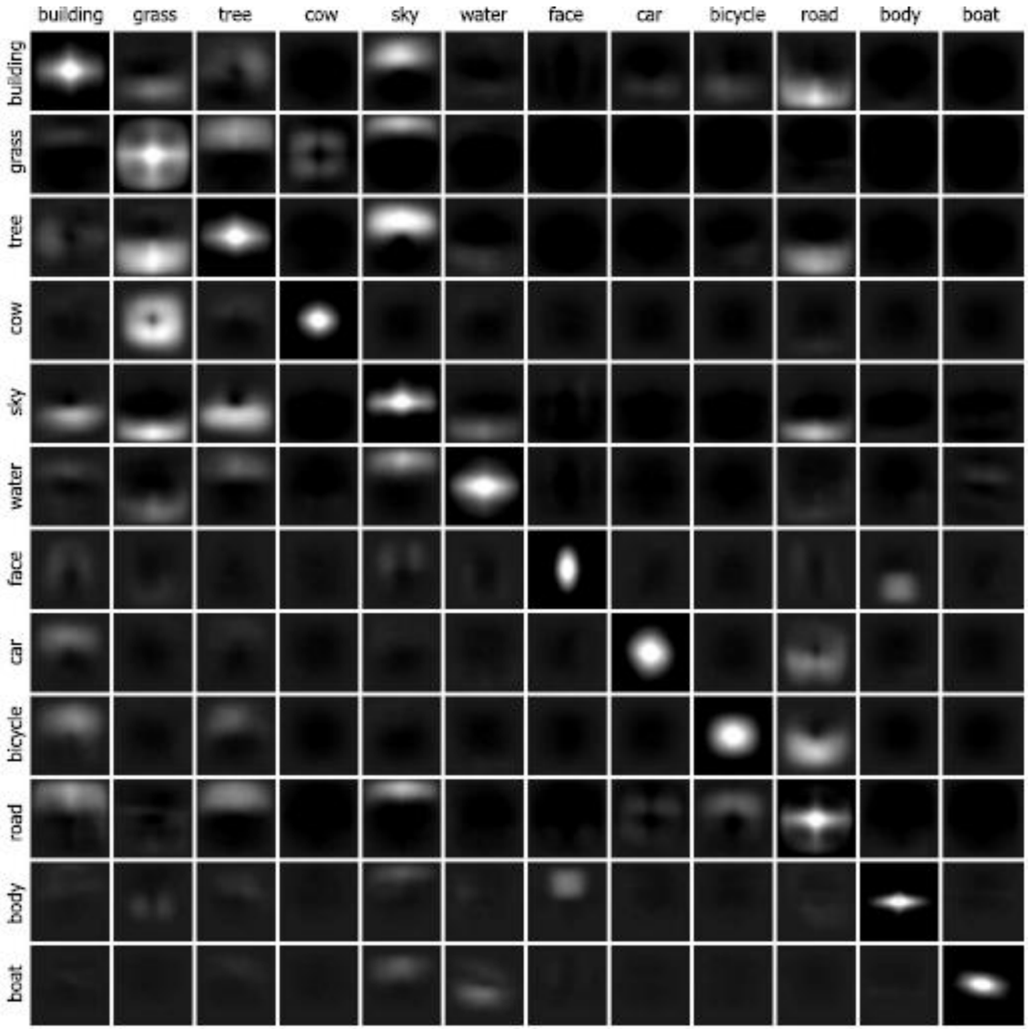
Summary

- Goal
  - Use of context in object categorization
- The type of context used
  - Co-Occurrence
    - Grass and Cow often appear together
  - Relative Location
    - Sky is above the grass.
- Related paper
  - S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-Class Segmentation with Relative Location Prior, IJCV 2008.

Multi-Class Segmentation with Relative Location Prior [Gould 2008]

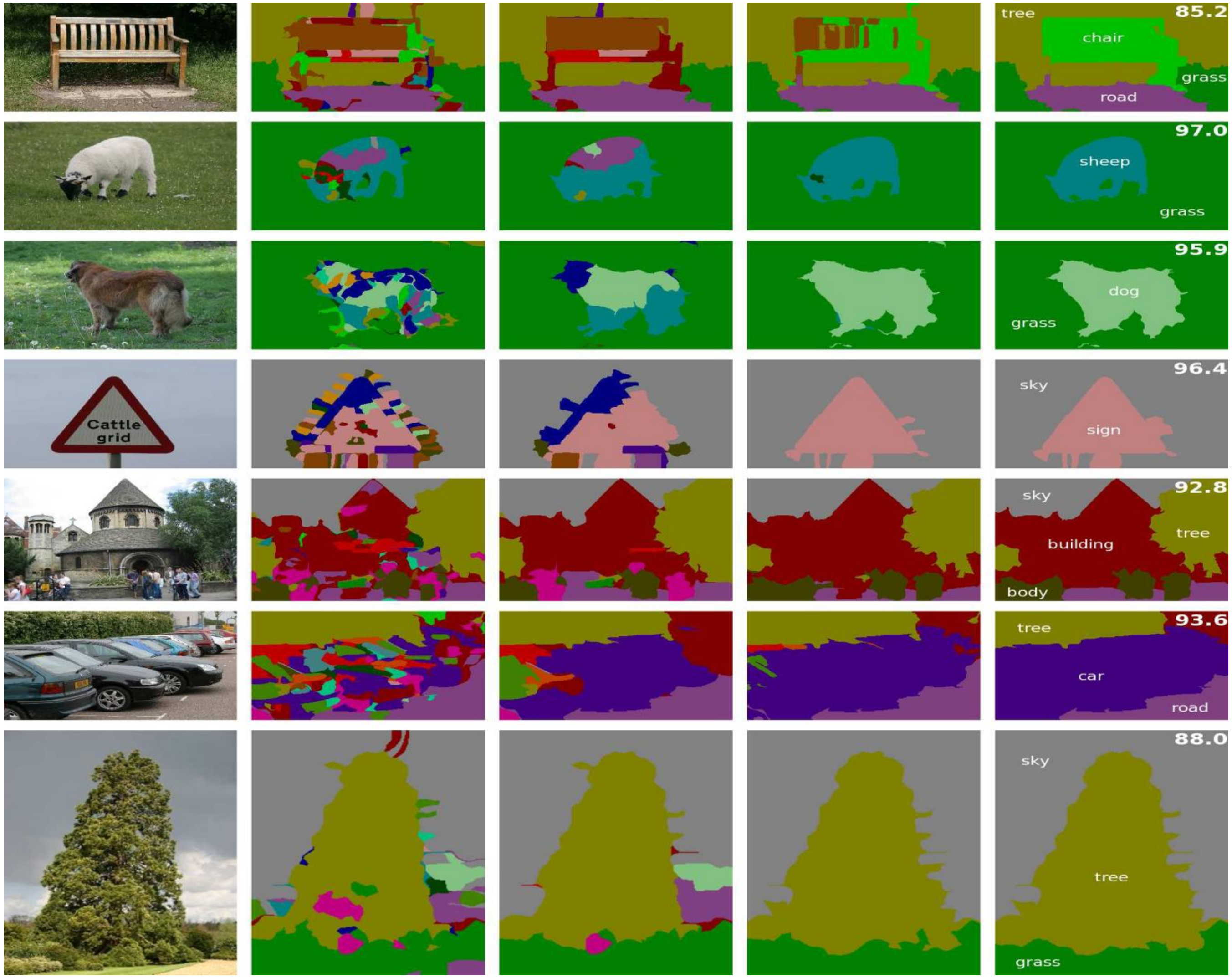
- Segmentation categorization accuracy is improved with the use of relative location prior
  - Same idea of using the spatial relation between object categories, but with a different approach
  - Relative location map is created between all categories
  - Obtained Average of 76.5% accuracy on the 21-class MSRC dataset (better than 68.38% in the [Galleguillos 2008])

Multi-Class Segmentation with Relative Location Prior [Gould 2008] – Relative prior maps



Instead of using only 4 types of spatial relation, The spatial relation between all object categories are modeled as prior maps

Multi-Class Segmentation with Relative Location Prior [Gould 2008] – The result on the MSRC dataset



Average Accuracy :
76.5% > 68.38%
in [Galleguillos 2008]

Conclusion

- There are different types of context
  - Global context
  - Inter-object relations

- Context can improve the object detection and recognition
  - We can learn important object properties from learning the scene it appears
  - The learned priors can be used in improving the speed and accuracy of object detection and localization
  - The Semantic and spatial relation between the objects can help to recognize the object better

Discussion

- Are there any other types of context that we can use to improve the object detection/recognition result?

Thank you