A Tree-Based Context Model for Object Recognition

Maha ELBAYAD

Probabilistic graphical models 2015/2016

Motivation

- Exploit contextual information + local features to detect and localise multiple object categories coexisting in an image.
- Rule out incoherent combinations or locations of objects and guide detectors to interpret the analysed scene..
- One probabilistic framework: global image features, dependencies between object categories, and outputs of local detectors. to improve object recognition performance.

The Context Model

Given $M = |Images|$, $N = |objects|$

(I) The Co-Occurrences prior captures dependencies between object categories.

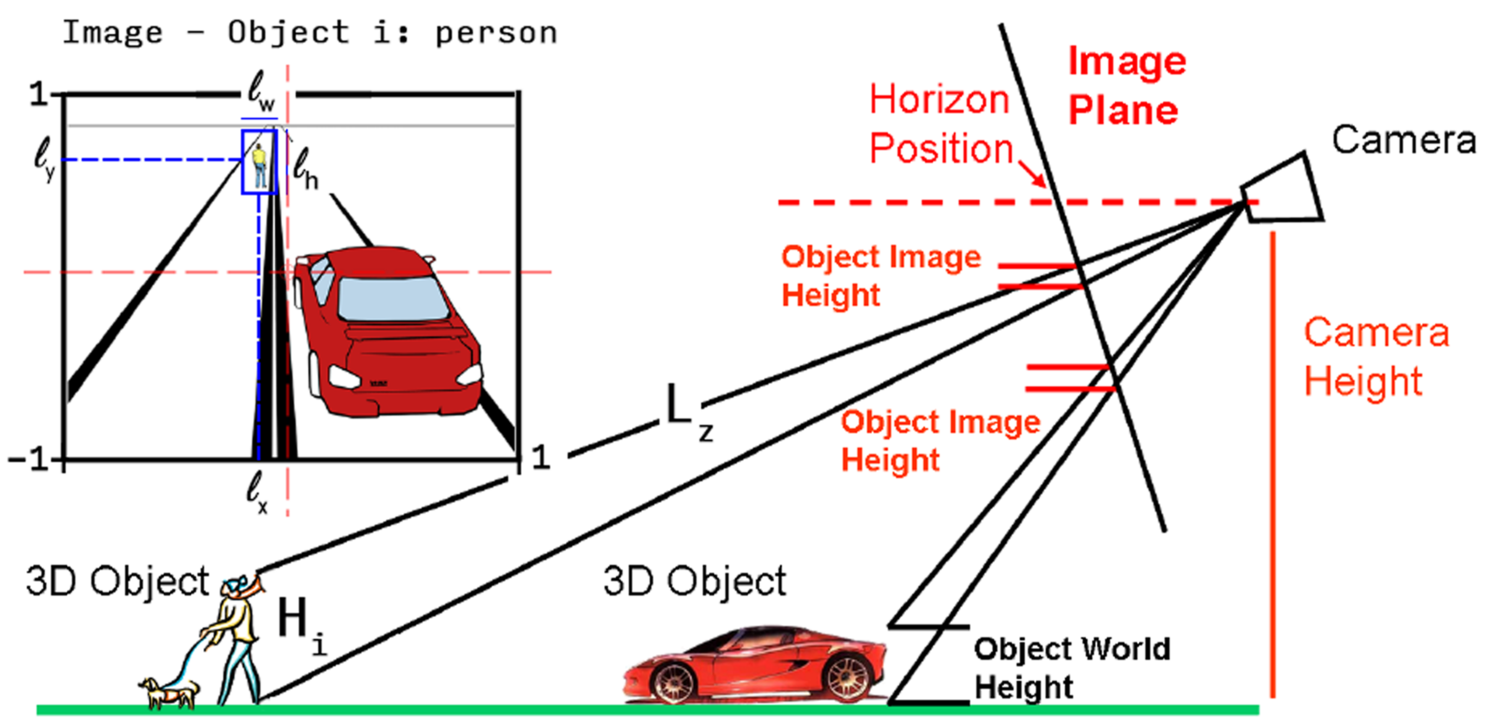
Given N nodes of binary variables $b_i = \mathbb{I}(object_i \in Image)$

Learn the dependency structure via **Chow-Liu's algorithm**: MST on the complete graph with weights $w_{i,j} = I(b_i, b_j)$ (mutual information)

$$\mathbb{P}(b) = \mathbb{P}(b_{root}) \prod_i \mathbb{P}(b_i | b_{\pi_i}) \qquad \text{(Co-Occurrences prior)}$$

(II) The Spatial prior each object occurring in an image is encoded with 2 coordinates:

$$L_i = (L_y, \log L_z) = \underset{o_i \in Image}{\mathbf{Median}} \left[(l_y, \log(.)) \frac{H_i}{l_h} \right]$$



Adapted from Hoiem, Efros, Hebert [2004]

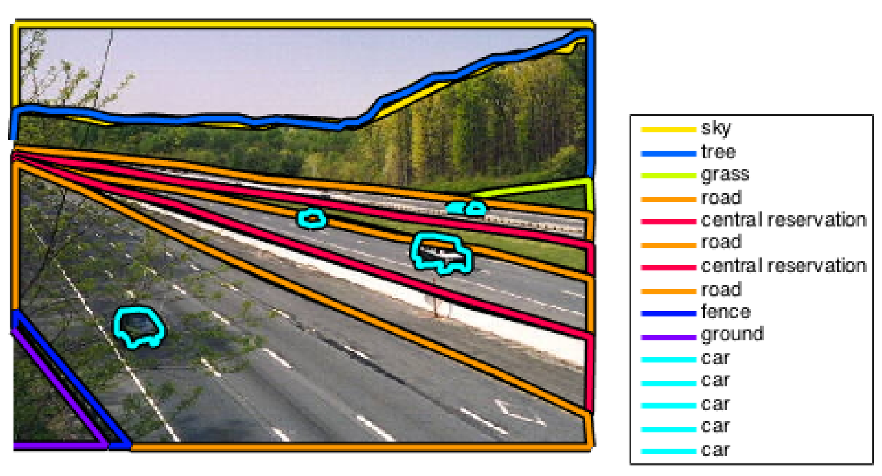
- Horizontal locations dropped since they tend to have weak contextual information!
- We assume $(L_y^{(i)})_i$, $(\log L_z^{(i)})_i$ are jointly Gaussians and that $L|b$ inherits the binary prior tree structure.

$$\mathbb{P}(L|b) = \mathbb{P}(L_{root}|b_{root}) \prod_i \mathbb{P}(L_i | L_{\pi_i}, b_i, b_{\pi_i}) \qquad \text{(Spatial prior)}$$

The Measurement Model

(IV) Baseline detectors

We apply **baseline single-object detectors** to obtain a set of candidate windows (as in L_i) for each object category.



- Candidates : $(W_{i,k})_k$
- scores : $(s_{i,k})_k$
- verdicts : $(c_{i,k})_k = \mathbb{I}(\text{correct})$

We assume:

$$W_{i,k} | c_{i,k} = 1 \sim \mathcal{N}(L_i) \quad W_{i,k} | c_{i,k} = 0 \sim \mathcal{U} \perp\!\!\!\perp L_i$$

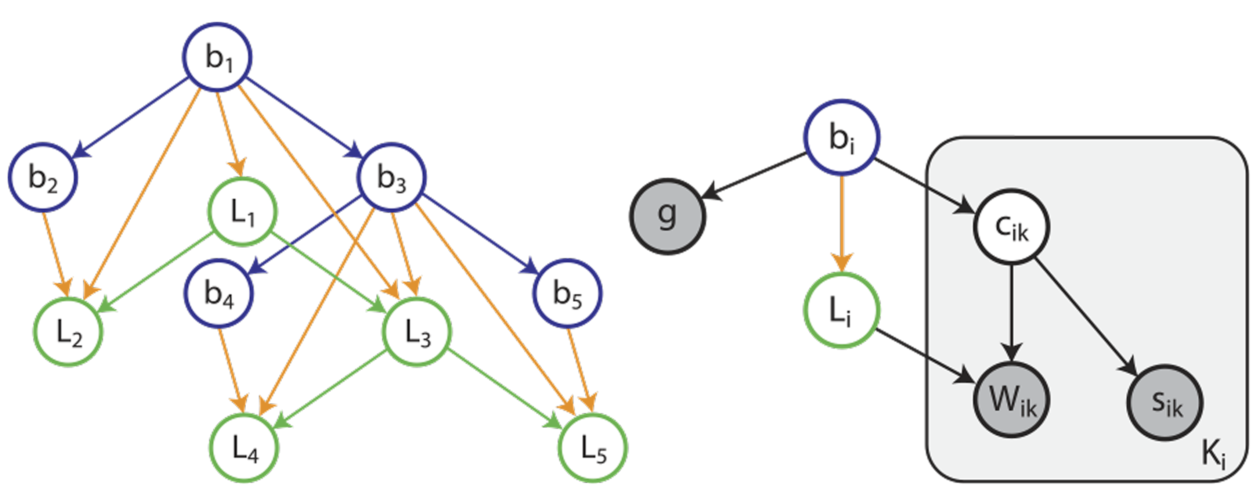
(III) GIST

We also integrate **global features** encoded in Gist for each image:

(i) Convolve with 32 **Gabor filters** at 4 scales, 8 orientations.

(ii) Average pooling in a 4x4 grid

(iii) Concatenate output: 16x32 descriptor $\equiv$ g.



Learning

(II) The Spatial prior

Infere $\mathbb{P}(L_i | L_{\pi_i}, b_i, b_{\pi_i})$ in 3 scenarios:

$\mathbf{b_i = 1, b_{\pi_i} = 1}$ $L_i | L_{\pi_i} \sim \mathcal{N}$

$\mathbf{b_i = 1, b_{\pi_i} = 0}$ $L_i \perp\!\!\!\perp L_{\pi_i}$

$\mathbf{b_i = 0}$ $L_i \perp\!\!\!\perp L_j \, \forall j$, set $L_i = \mathbb{E}(L_i)$

(III) GIST

For each category fit $\mathbb{P}(b_i|g)$ with a logistic regression.

(IV) Baseline detectors

For the local detectors outputs, we fit $\mathbb{P}(c_{i,k}|s_{i,k})$ with a logistic regression.

And estimate $\mathbb{P}(c_{i,k}|b_i)$ by counting the correct detections in the training set.

Alternating inference on trees - Sum-Product

Inputs: GIST g, candidate windows $W = W_{i,k}$ and their scores $s = s_{i,k}$

Infer: Presence $b = b_i$, detections' verdicts $c = c_{i,k}$ and the locations $L = L_i$ as:

$$\hat{b}, \hat{c}, \hat{L} = \arg\max_{b,c,L} \mathbb{P}(b, c, L | g, s, W)$$

Noting that $L|b, c$ is a Gaussian tree and $b, c|L$ is a Binary tree

Approach:

Initialisation:

$$\hat{b}, \hat{c} = \arg\max_{b,c} \mathbb{P}(b, c | g, s) \qquad \text{(Ignoring W)}$$

Iterate:

$$\hat{L} = \arg\max_{L} \mathbb{P}(L | \hat{b}, \hat{c}, W) \qquad \text{(Gaussian tree : SUM-PRODUCT)}$$

$$\hat{b}, \hat{c} = \arg\max_{b,c} \mathbb{P}(b, c | g, s) \mathbb{P}(\hat{L}, W | b, c) \qquad \text{(Binary tree: SUM-PRODUCT)}$$

Final outputs:

Compute marginal probability $\mathbb{P}(b_i = 1 | g, s, \hat{L}, W)$

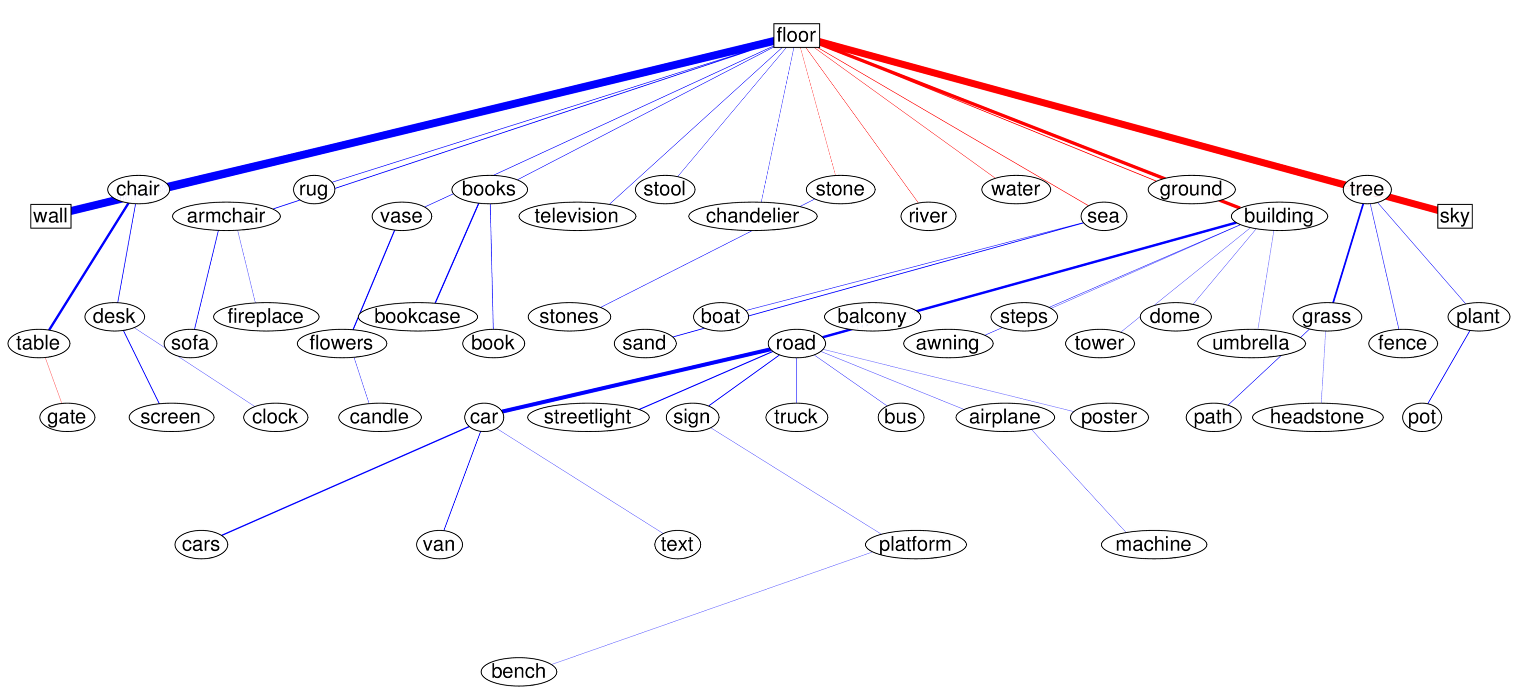
And the marginal $\mathbb{P}(c_{i,k} = 1 | g, s, \hat{L}, W)$

To deal with re-occurring objects of class i, we set all the messages from node b_i to $(c_{i,k})_{1 \le k \le K_i}$ as 1, except a single occurence.

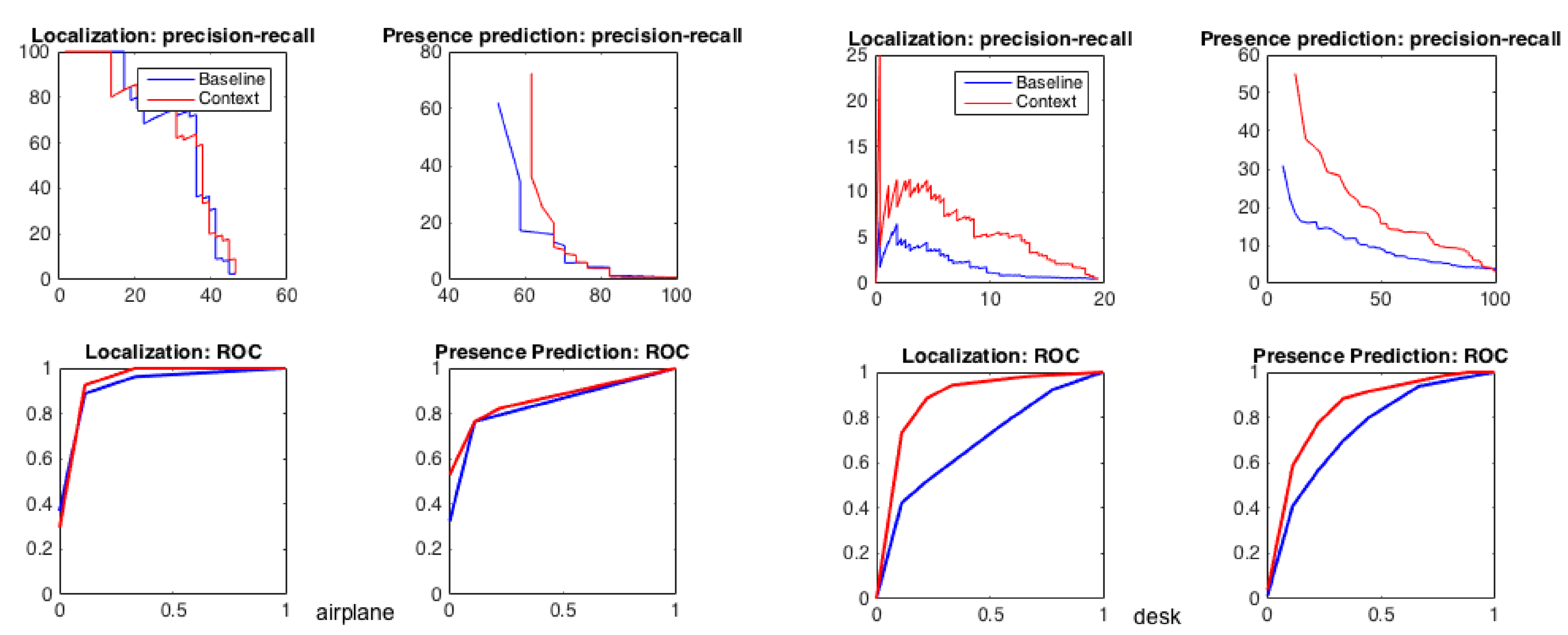
Experiments - SUN 09

Dataset: SUN 09, Training: 4367 images, Test: 4317, N=111 categories.

Object dependency structure learned from SUN 09 - subtree `Floor`:



Localisation and presence prediction performance:



Average recognition performance for the top N most confident detections