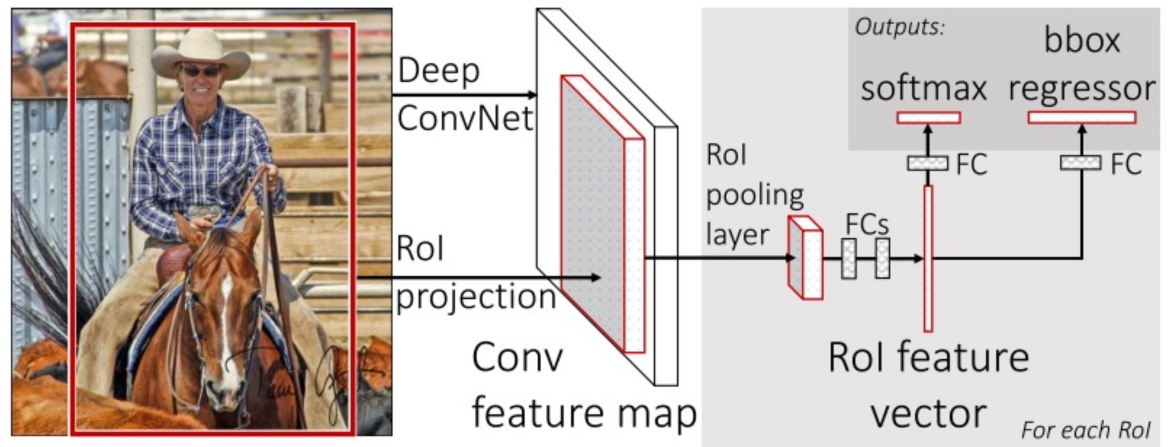
IVMSP-P6.3

Markov Random Field Based Pruning and Learning Based Rescoring for Object Detection

Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, Kiyoharu Aizawa

Dept. Information and Communication Engineering, The University of Tokyo

Background

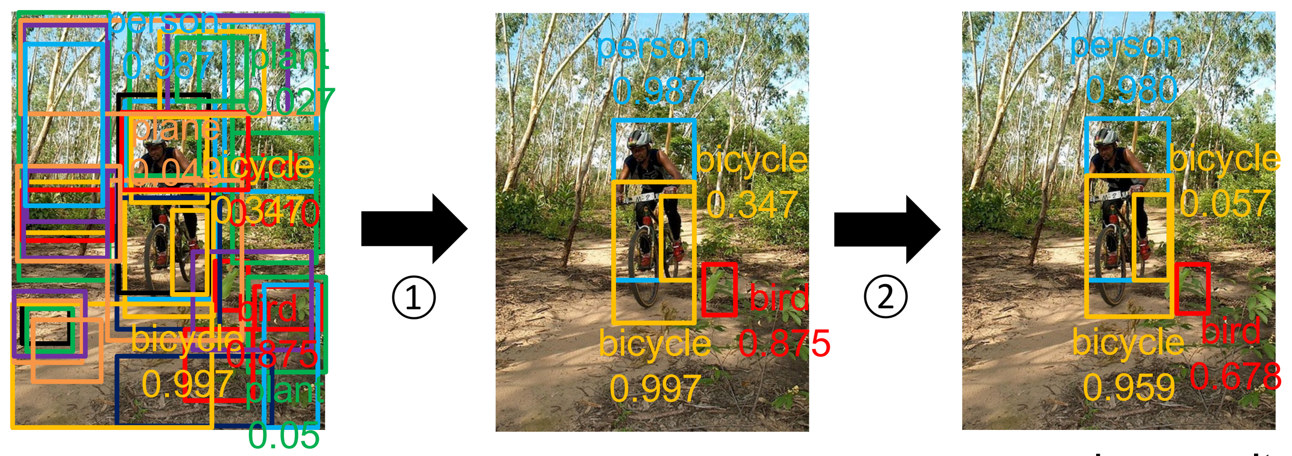


Fast R-CNN [Girshick, 2015]

"Context" in object detection
= information out of each candidate window
= co-occurrence of objects, spatial layout, scale, background, scene of image, ..

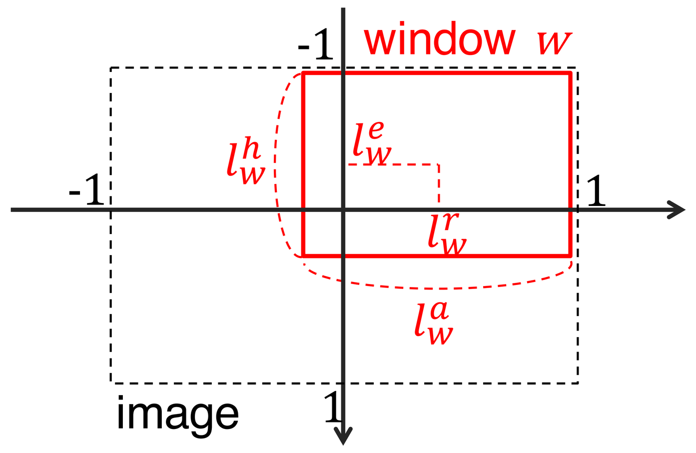
R-CNN based methods do not consider context.

Proposed Method



detector output — pruning result — rescoring result

Use coordinate in [Hoiem+, 2008] to model context.



b_w : 1 if detection is correct else 0

c_w : class of candidate window w

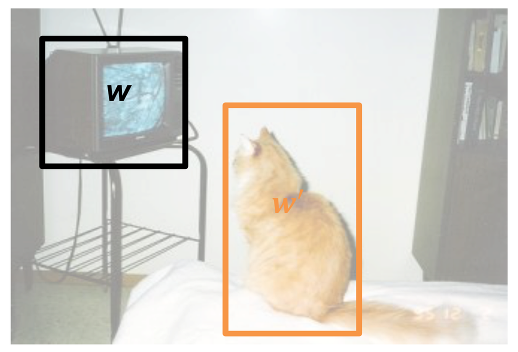
H_{c_w} : physical height of c_w

$L_y^w = \frac{l_w^e}{l_w^h} H_{c_w}$ (= vertical position)

$L_z^w = \frac{1}{l_w^h} H_{c_w}$ (= size)

Distribution of windows are fitted by Cauchy distribution.
Conditional probability of each relation is calcurated.

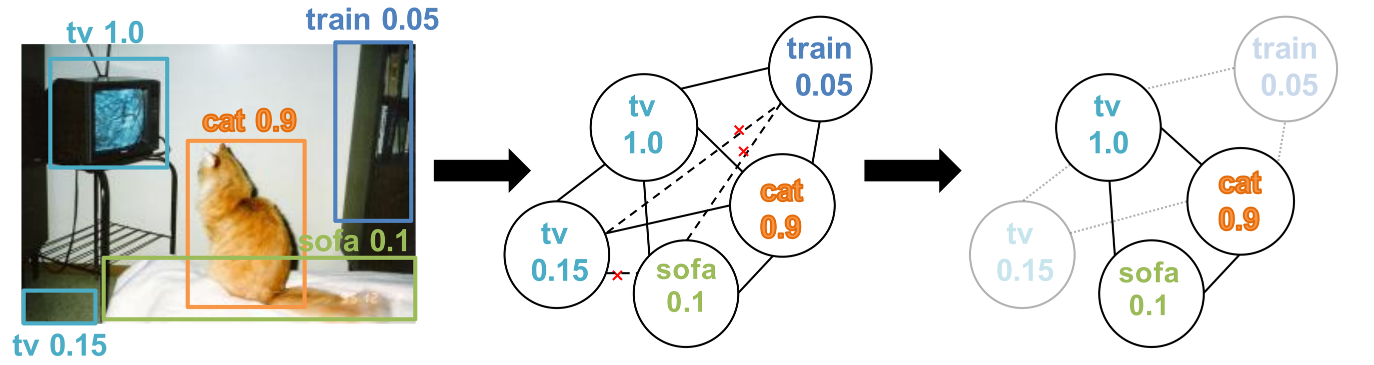
$$p_{pos}(w, w') = p(b_w = 1, b_{w'} = 1 \mid d(L_w^y, L_{w'}^y), c_w, c_{w'})$$
$$p_{scale}(w, w') = p(b_w = 1, b_{w'} = 1 \mid d(L_w^z, L_{w'}^z), c_w, c_{w'})$$



① Pruning by MRF

Decide whether each window should be pruned (y_w =1) or not (y_w =0).
Solved by QPBO [Kolmogorov+, 2007] energy minimization.
Better than just setting threshold.

$$\boldsymbol{y} = \{y_{w_1}, y_{w_2}, ..\}, E(\boldsymbol{y}) = \sum_i \phi(y_{w_i}) + \beta \sum_{i,j} \varphi(y_{w_i}, y_{w_j})$$
$$\to \boldsymbol{y}^* = argmin E(\boldsymbol{y})$$



candidate windows — constructed graph — energy minimized

② Rescoring by SVM

Predict whether each window is correct by SVM.
New score s_w is calcurated based on decision value d_w.

$$s_w = \frac{1}{1+e^{-3d_w}} \quad (0 \le s_w \le 1)$$

Type of feature	Dim.
Score of the window	1
Probability of vertical position ($p(b_w = 1 \mid L_y^w, c_w)$)	1
Vertical position likelihood between windows	1
Scale likelihood between windows	1
Scene probability by Places-CNN [zhou+, 2014] ($\boldsymbol{p}_{scene}$)	205

Experimental Results

Evaluation

Baseline detector: Fast R-CNN [Girchick+, 2015].
Dataset: VOC2007 and MSCOCO.

Dataset	VOC2007-test		MSCOCO-val	
Evaluation metric	mAP [%]	F1 [%]	mAP [%]	F1 [%]
Baseline (Fast R-CNN)	66.9	3.5	32.3	8.4
+ Tree Context [Choi+, 2012]	60.9 (-6.0)	3.5 (+0.0)	-	-
+ HOOD [Cao+, 2015]	57.9 (-9.0)	67.2 (+63.7)	21.0 (-10.3)	34.0 (+25.6)
+ threshold	66.5 (-0.4)	24.4 (+20.9)	32.1 (-0.2)	11.8 (+3.4)
+ pruning	66.5 (-0.4)	26.2 (+22.7)	32.2 (-0.1)	11.0 (+2.6)
+ pruning + rescoring	67.3 (+0.4)	26.2 (+22.7)	33.0 (+0.7)	11.0 (+2.6)

Examples

ground truth

Fast R-CNN results

proposed results

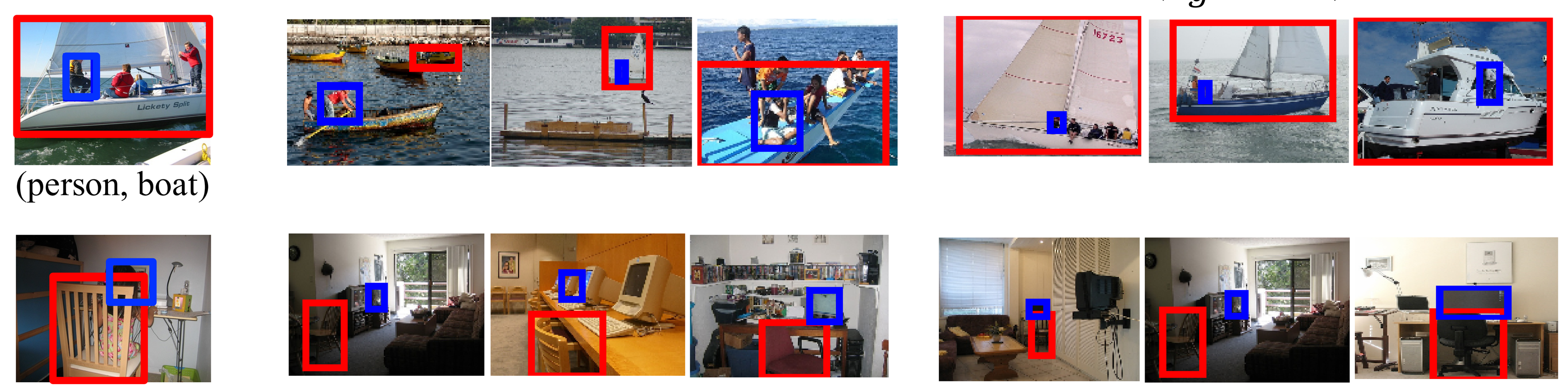
(※) windows whose score are > 0.05 are shown.

Application: Structured Retrieval

Given a query q containing image i_q and a window pair (w_q, w_q'),
find target t contaiting i_t and (w_t, w_t') that has smallest distance.
Compararison with HOOD [Cao+, 2015]

$$dist(q,t) = (1-\lambda_g)\left\| \begin{pmatrix} p_{pos}(w_q, w_q') \\ p_{scale}(w_q, w_q') \end{pmatrix} - \begin{pmatrix} p_{pos}(w_t, w_t') \\ p_{scale}(w_t, w_t') \end{pmatrix} \right\|_2 + \lambda_g \|\boldsymbol{p}_{scene}(i_q) - \boldsymbol{p}_{scene}(i_t)\|_2$$

Query — HOOD — Ours ($\lambda_g = 0.5$)



(person, boat)

(chair, person)

Conclusion

- Reducing and rescoring candidate windows by considering contextual model.
- Fast R-CNN detectors are improved by
  +0.4% on mAP and +22.7% on F1 in VOC2007-test.
  +0.7% on mAP and +2.6% on F1 in MSCOCO-val.
- Applications to structured retrieval are also presented.