Cascaded Sparse Spatial Bins For Efficient And Effective Generic Object Detection

David Novotný[1,2], Jiří Matas[2]
[1]Visual Geometry Group, University of Oxford
[2]Center for Machine Perception, Czech Technical University, Prague

Introduction

- **A novel object proposal method.**
- Efficiency is achieved by the use of spatial bin pooling in a novel combination with sparsity-inducing group normalized SVM.
- Boundary Edge Vector (BEV), a new HoG-like "objectness" descriptor, proposed.
- State-of-the-art results on VOC07 and ILSVRC13 achieved.

Overview

- Window scoring method.
- Each bounding box described by "objectness" features pooled from **learned** spatial bins.
  - **Train time: ℓ_1/ℓ_2 normalized SVM automatically selects the set of relevant spatial bins.**
- The pooled features are scored by a **two-stage SVM cascade**.

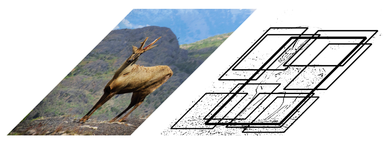
Objectness features:

CNN-SPP	CNN features obtained by spatial pyramid pooling [1].
EdgeBoxes score (EB)	The score by which EdgeBoxes rank proposals [2].
Boundary Edge Vector (BEV)	A novel HoG-like edge statistic.

Two-stage Cascade

Initialization:

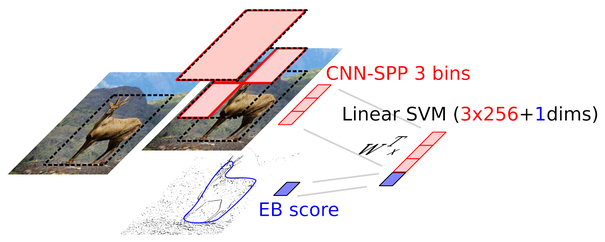
Obtain a large pool of 100K EdgeBoxes proposals.



Stage 1: Reduces the pool of boxes from 100K to 10K

For each box:

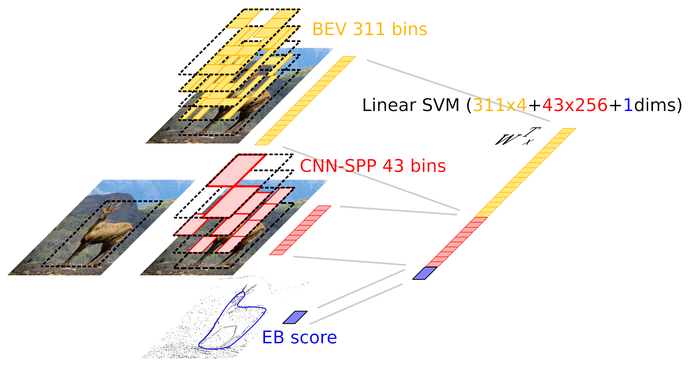
1. Pool CNN-SPP features from 3 **selected spatial bins**.
2. Append EdgeBoxes score.
3. Score with SVM.
4. Based on SVM score keep top 10K boxes.



Stage 2: Reduces the pool of boxes from 10K to the final requested size

For each box kept after stage 1:

1. Pool BEV features from k=311 **selected spatial bins**.
2. Pool CNN-SPP features from l=43 **selected spatial bins**.
3. Append EdgeBoxes score.
4. Score with SVM.
5. Apply non-maximum suppression to obtain requested number of boxes.



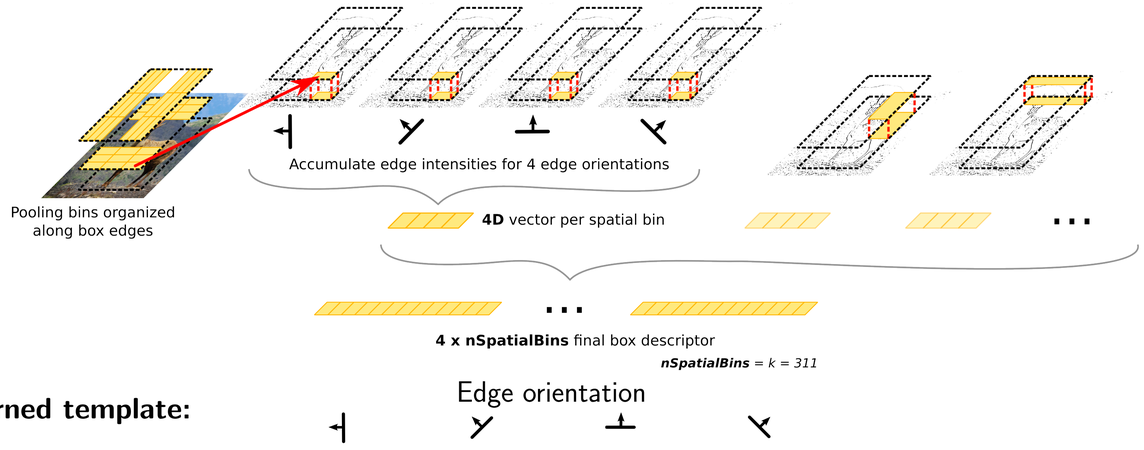
Parameters k and l were validated such that they give best compromise between execution speed and performance.

Spatial Bin Selection by ℓ_1/ℓ_2 Normalized SVM

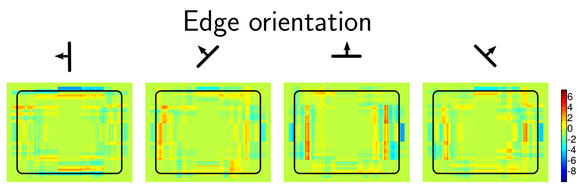
- The set of spatial bins for pooling CNN-SPP and BEV features is learned automatically.
- Group sparsity inducing ℓ_1/ℓ_2 SVM selects groups of dimensions that correspond to relevant spatial bins.
- **Significantly speeds-up "objectness" feature extraction** with negligible performance decrease.
- Group sizes: CNN-SPP ... 256 (# of conv5 filters), BEV ... 4 (# of orientation bins).

Boundary Edge Vector (BEV)

- A novel HoG-like edge statistic.
- Reuses the EdgeBoxes structured edge detector output = almost no additional cost.
- Quantizes edges based on their orientation and pools their intensities from selected spatial bins:



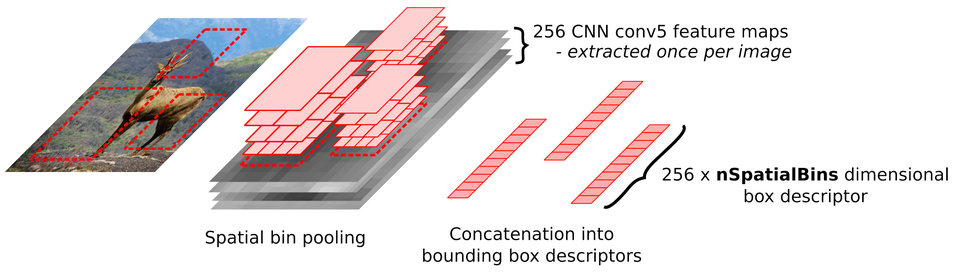
Learned template:



Note: BEV weights are learned to be complementary to the EdgeBoxes score and CNN-SPP descriptors.

CNN-SPP Features [1]

Box descriptors formed by spatial pyramid pooling of CNN conv5 features:

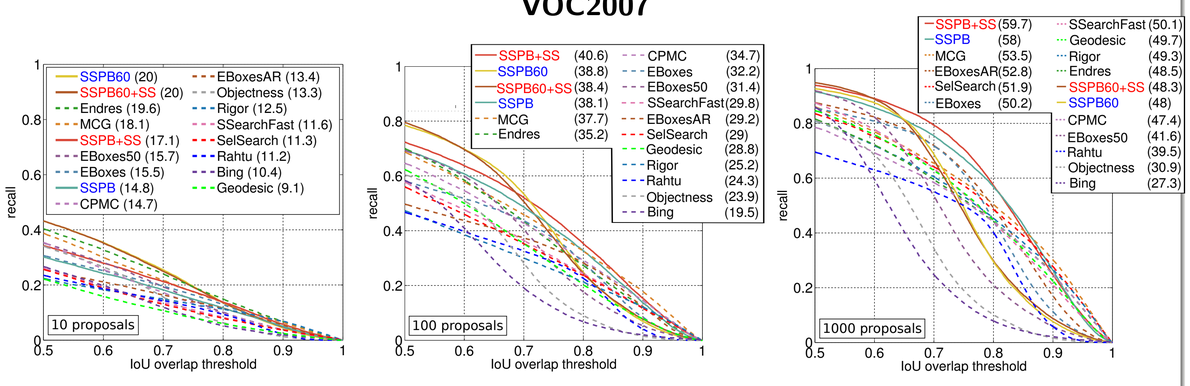


Overlap-Recall Curves

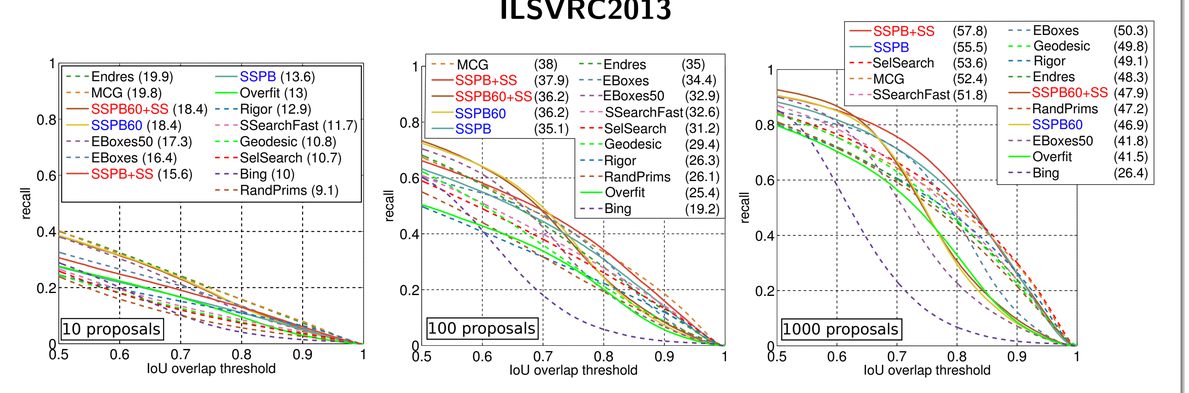
Proposed methods (solid lines in plots):
SSPB (**S**parse **SP**atial **B**ins) Basic method.
SSPB60 SSPB with non-max suppression (NMS) threshold optimized for a small number of candidates.
SSPB+SS Basic method improved with Selective Search [3] (slower than SSPB).
SSPB60+SS SSPB+SS with NMS threshold optimized for a small number of candidates.

All SSPB variants are trained solely on the trainval set of VOC2007.

VOC2007



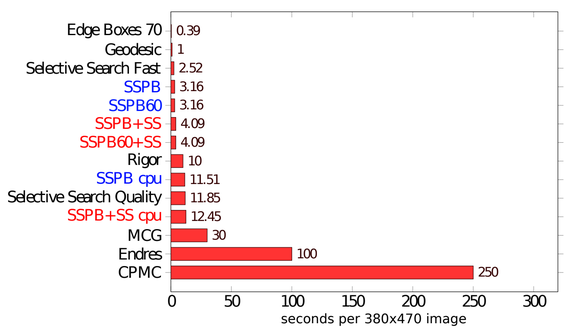
ILSVRC2013



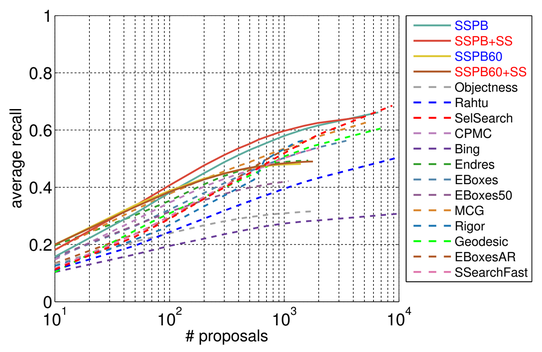
State-of-the-art results on VOC2007.
Competitive results on ILSVRC2013 confirm generalization capability.

Run-time and Performance

Run-time:



Average recall on VOC07:

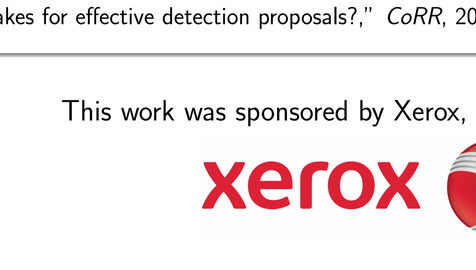


Highest performance of the methods with similar run-time.
Conv5 features calculated by SSPB can be reused by the final class-specific detector - lower detection times.

Object Detection with RCNN

RCNN [4] mAP when used in combination with different proposal methods.

	# proposals					
method	10	50	100	500	1000	10000
SelectiveSearchFast [3]	23.7	37.2	42.8	52.5	54.2	54.8
EdgeBoxes [2]	32.3	43.0	46.1	52.1	53.3	53.1
SSPB	**36.0**	46.7	50.0	53.1	56.4	**56.3**
SSPB+SS	35.7	**47.8**	**50.2**	**56.1**	**56.6**	**56.3**



SSPB and SSPB+SS achieve higher mAP than standard proposal methods.

Conclusions

- CNN features efficiently used for object proposals.
- VOC07: State-of-the-art recall performance on VOC07.
- ILSVRC13: Recall better than methods with similar execution time.
- GPU speed comparable to Selective Search in "fast mode", CPU as fast as "quality mode".
- Better than EdgeBoxes [2] or Selective Search [3] in combination with the RCNN class specific object detector.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," 2014.

[2] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014.

[3] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *IJCV*, 2013.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[5] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?," *CoRR*, 2015.

This work was sponsored by Xerox, S.A.S.

The code will be available shortly on http://cmp.felk.cvut.cz/software/SSPB