Segmentation as Selective Search for Object Recognition

Koen E. A. van de Sande [1], Jasper R. R. Uijlings* [2], Theo Gevers[1], Arnold W. M. Smeulders[1]*

**Both authors contributed equally, [1]University of Amsterdam (The Netherlands), [2]University of Trento (Italy)*

Universiteit van Amsterdam

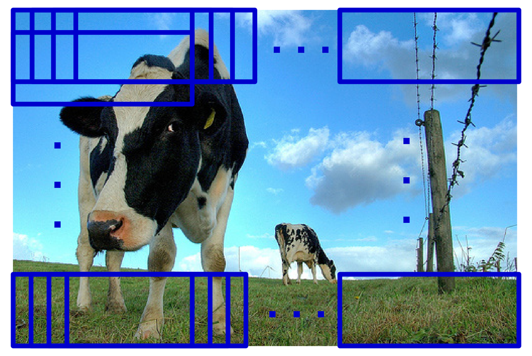
University of Trento

42

Object recognition

Object recognition seeks to answer 2 questions:
- What is it?
- **Where is it?**

Exhaustive Search



Exhaustive search:
- Current state-of-the-art
- # windows to evaluate: 100,000 – 1,000,000

➔ Simple-to-compute features

➔ Weak classifiers

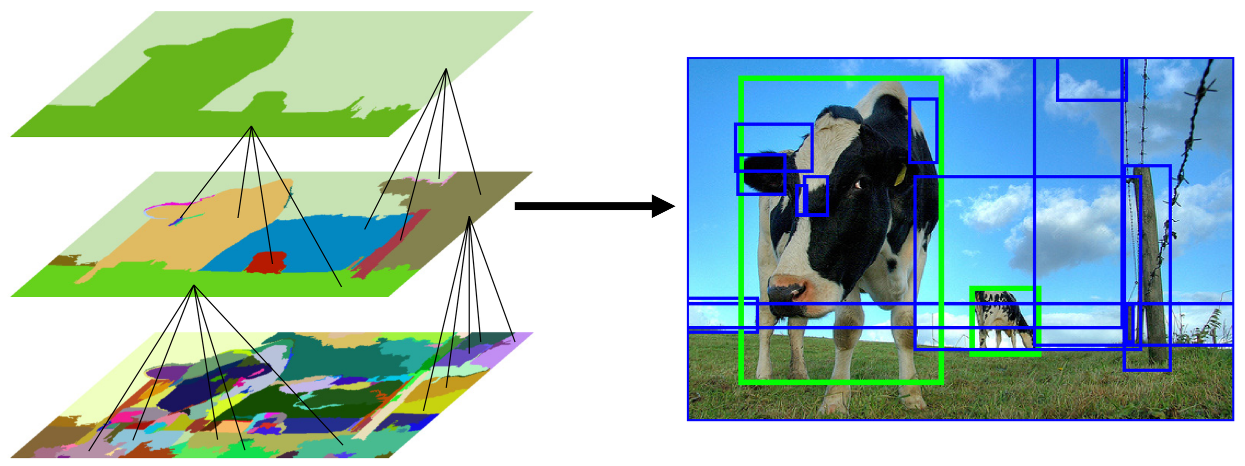
Selective Search

Adopt segmentation as selective search strategy

Different goal from segmentation:
prefer to generate many approximate locations over few and precise object delineations, because:
1. objects whose locations are not generated can never be recognised
2. appearance and immediate nearby context are effective for object recognition.

Design considerations:
- High recall
  - ➔ Details on the right
- Coarse locations are sufficient
  - ➔ Use bounding boxes
- Fast to compute
  - ➔ Less than 10s/image

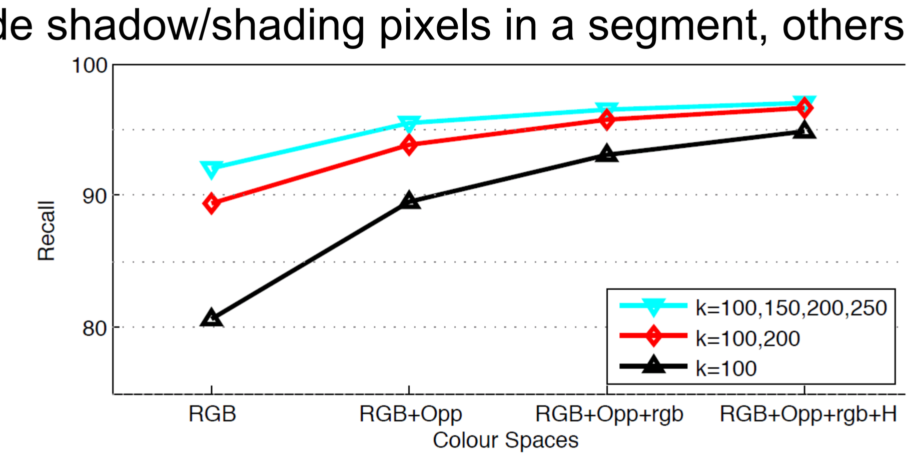
Segmentation as Selective Search



Selective search based on hierarchical grouping
- Initial segments from oversegmentation [Felzenszwalb2004]
- Group adjacent regions on region-level similarity:
  - Texture (gradient orientations)
  - Region size
- Consider *all* scales of the hierarchy

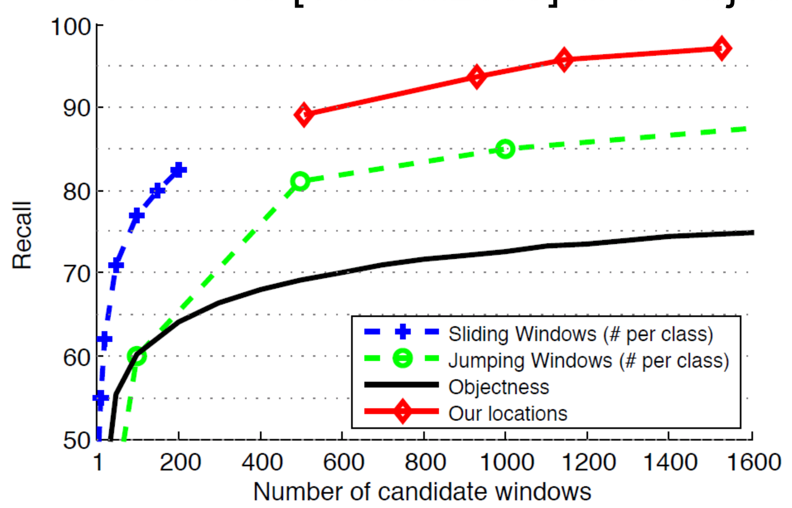
Multiple Complementary Color Spaces

- It is important to diversify the set of segmentations used: we combine multiple initial segmentations and different color spaces
- Color spaces with complementary invariance properties: some include shadow/shading pixels in a segment, others do not



Recall of Selective Search

- Our object location windows are class-independent
- Achieves higher recall than Sliding Windows [Harzallah2009], Jumping Windows [Vedaldi2009] and 'Objectness' [Alexe2010]



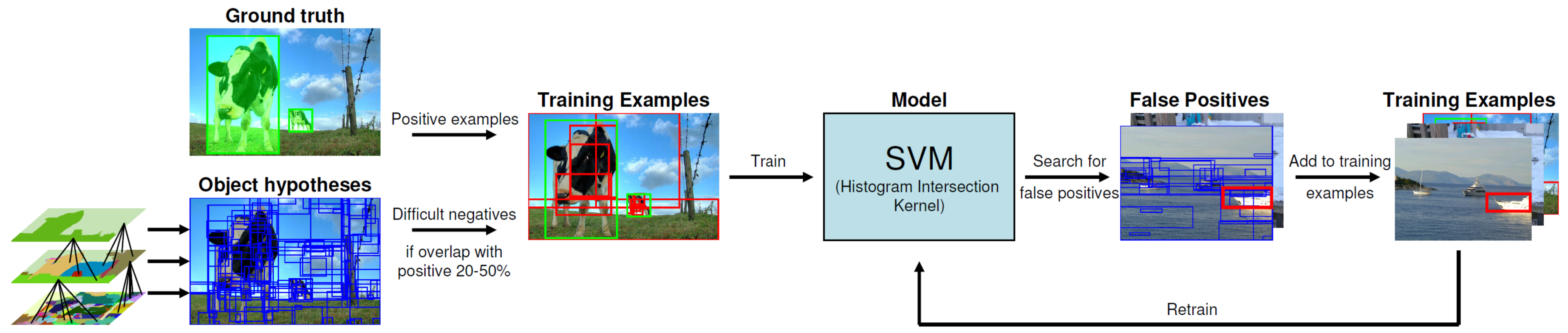
1,536 windows/image
96.7% recall

Experiment 2: Maximum Recall of Selective Search for Recognition

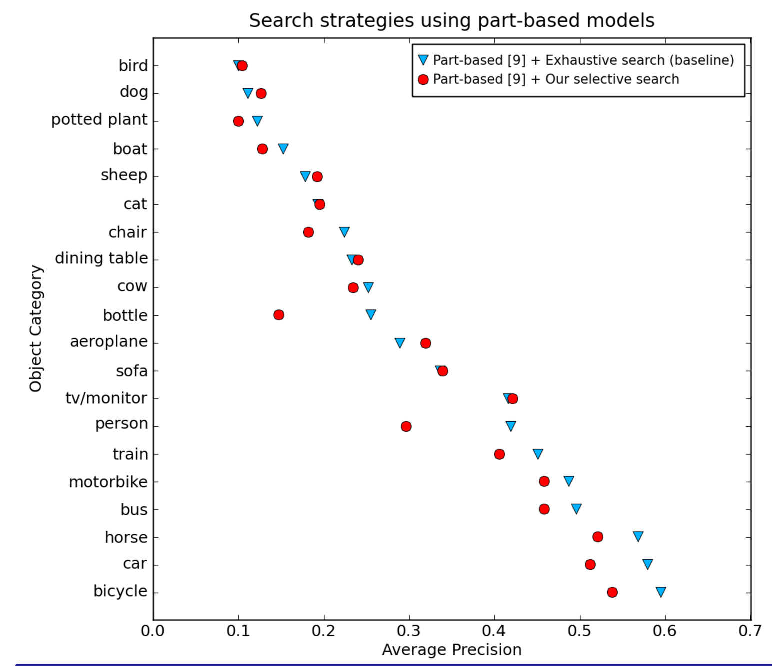
	Max. recall (%)	# windows
Sliding Windows [13]	83.0	200 per class
Jumping Windows [27]	94.0	10,000 per class
'Objectness' [1]	82.4	10,000
Our hypotheses	96.7	1,536

Object Recognition Accuracy

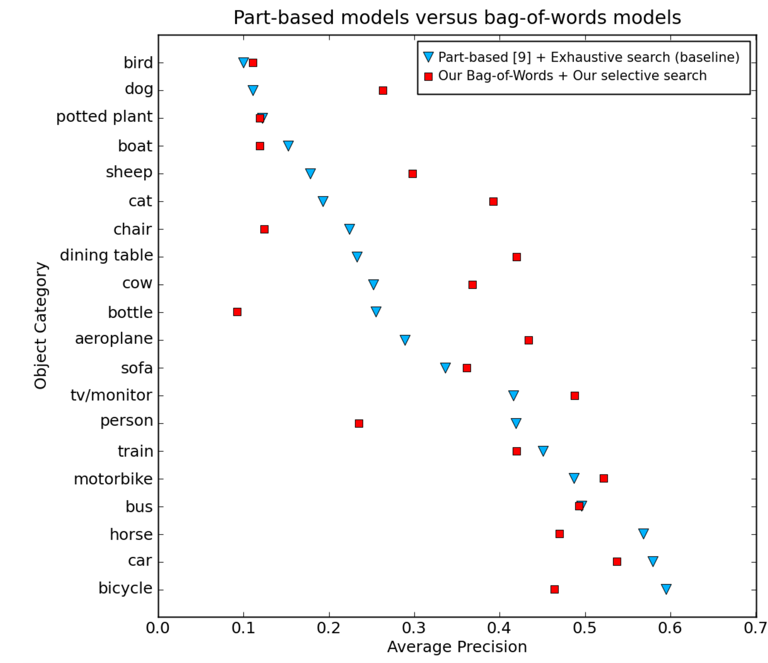
Object Recognition System: Training Pipeline



Selective search enables the use of more powerful features and classifiers:
- Dense SIFT, OpponentSIFT and RGB-SIFT sampled at every pixel (using software from www.colordescriptors.com)
- Codebook size 4,096; spatial pyramid with depth 4
- SVM classifier with Histogram Intersection Kernel and Fast Approximation [Maji2009]
- Initial negatives overlap 20-50% with positive examples
- Retrain with false positives (found in the train set) as extra negatives



Constrain [Felzenszwalb2010] from exhaustive to selective search:
20x fewer boxes -3% MAP



Bag-of-words features instead of HOG:
- Improvements for 10 out of 20 objects
- Oracle combination: **+5% MAP**

Benchmarks

- **#1 localisation** in IMAGENET Large Scale Visual Recognition Challenge 2011
- PASCAL VOC2010 test set (through independent evaluation server): Improves the state-of-the-art by up to 8.5% AP (absolute) for 8 out of 20 objects

System	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv
NLPR	.533	**.553**	**.192**	**.210**	.300	.544	.467	.412	**.200**	**.315**	.207	.303	.486	.553	.465	.102	.344	.265	**.503**	.403
MIT UCLA [29]	.542	.485	.157	.192	.292	**.555**	.435	.417	.169	.285	.267	.309	.483	.550	.417	.097	.358	.308	.472	.408
NUS	.491	.524	.178	.120	.306	.535	.328	.373	.177	.306	.277	.295	**.519**	**.563**	.442	.096	.148	.279	.495	.384
UoCTTI [9]	.524	.543	.130	.156	**.351**	.542	**.491**	.318	.155	.262	.135	.215	.454	.516	**.475**	.091	.351	.194	.466	.380
This paper	**.582**	.419	**.192**	.140	.143	.448	.367	**.488**	.129	.281	**.287**	**.394**	.441	.525	.258	**.141**	**.388**	**.342**	.431	**.426**

Conclusion

- Adopted segmentation as selective search strategy: prefer to generate many approximate locations over few and precise object delineations, as (1) objects whose locations are not generated can never be recognised and (2) appearance and immediate nearby context are effective for object recognition.
- Highest recall to date for Pascal VOC 2007 test set: only 1,536 class-independent locations/image capture 96.7% of all objects.
- Highly effective for object recognition: improve the state-of-the-art for 8 out of 20 classes for up to 8.5% AP

http://koen.me/research/selectivesearch