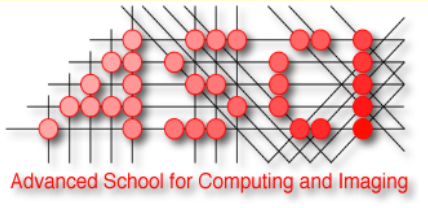
Computer Vision by Learning

Cees Snoek
Laurens van der Maaten
Arnold W.M. Smeulders
with Shih-fu Chang, Columbia University




UNIVERSITY OF AMSTERDAM

TU Delft

Administration

Tuesday to Friday

Lectures	0930-1215	D1.116 (1.115)
Lunch	1215-1330	on your own
Lab	1330-1700	D1.111

Please be on time

Note

Monday

Lecture **Shih-Fu Chang**	0930-1215	G2.10
Lunch	1215-1400	on your own
Lab	1400-1700	G2.02

Lab

Lab 1	Measuring invariance
Lab 2	Pedestrian detection
Lab 3	Learning object and sce
Lab 4	Fine-grained categor
Lab 5	Your own research pro

Demonstrate you have learned.
Do not make it your life's work.

Each team of 2 persons hands in a 10-page report using CVPR style sheet, 2 pages per lab.

Note

Deadline: **Monday April 21, 2014.**

Email to: cgmsnoek@uva.nl

The spatial extent of an object

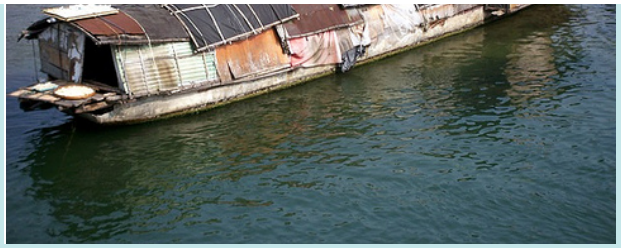
What is the context of an object? Where is the evidence for an object to be member of its class? What is the visual extent of an object? Does it stop at the visual projection of its physical boundary?

What is in the middle?

No segmentation …
Not even the pixel values of the object …

1. Segmentation out of context requires much experience.
2. Segmentation in context is easy.
3. Recognition precedes segmentation.

What makes a boat a boat?







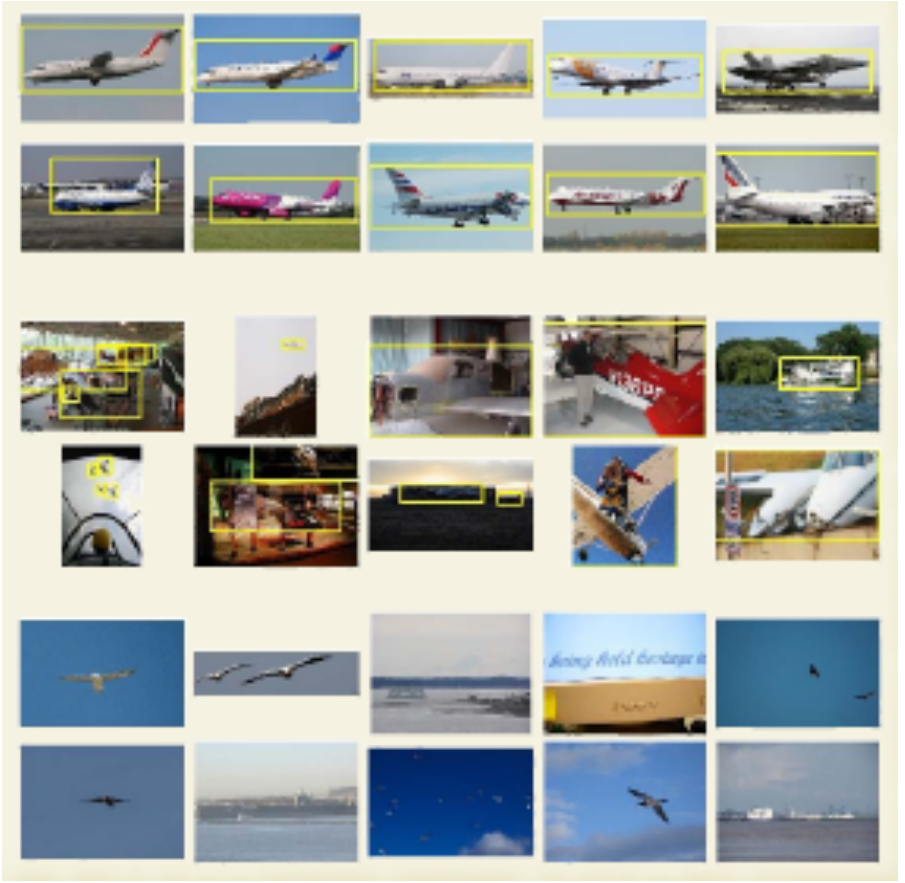


Context dominated objects

Highest ranked class

Lowest ranked class

Highest ranked non class



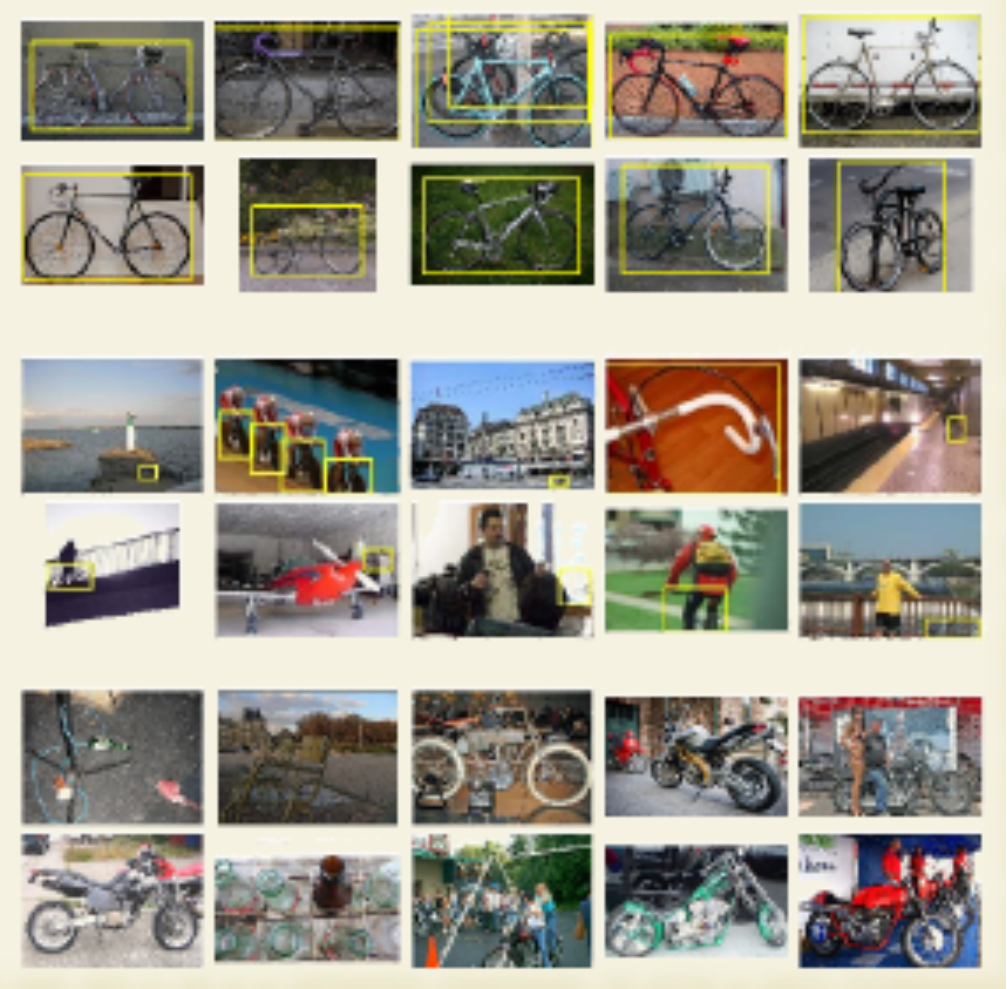
Slide: Mark Everingham

Object dominating the context

Highest ranked class

Lowest ranked class

Highest ranked non class



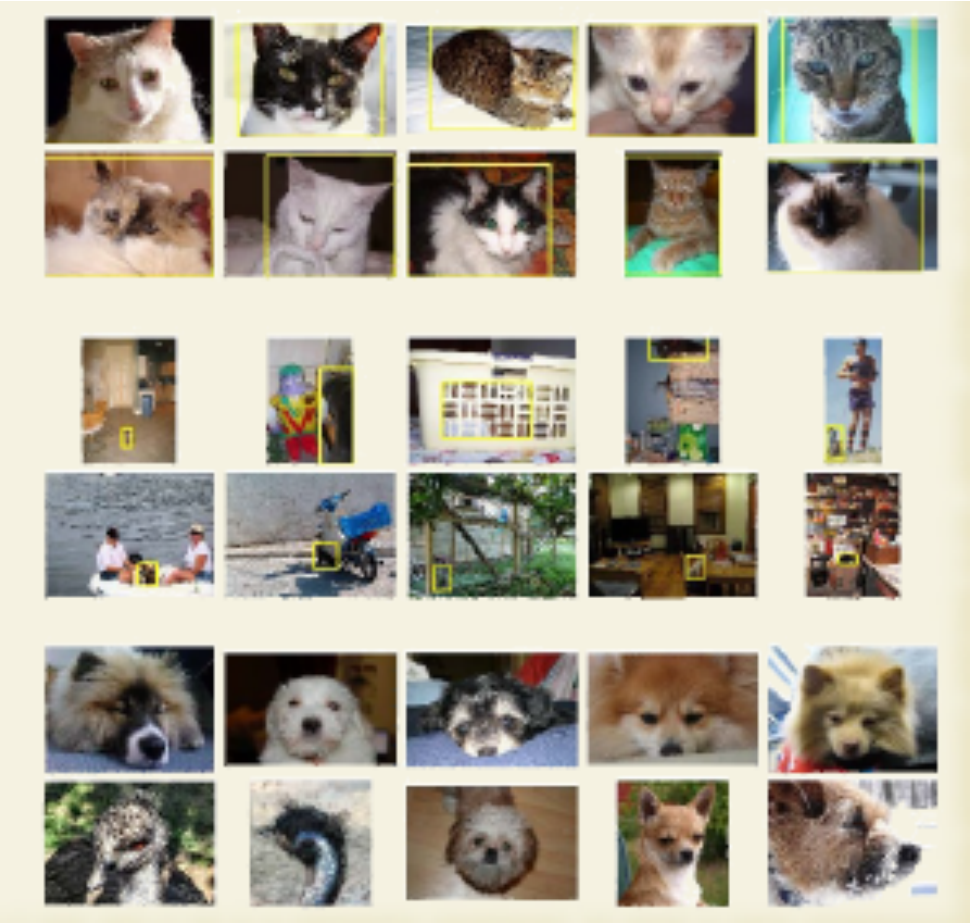
Slide: Mark Everingham

Object salient detail dominant

Highest ranked class

Lowest ranked class

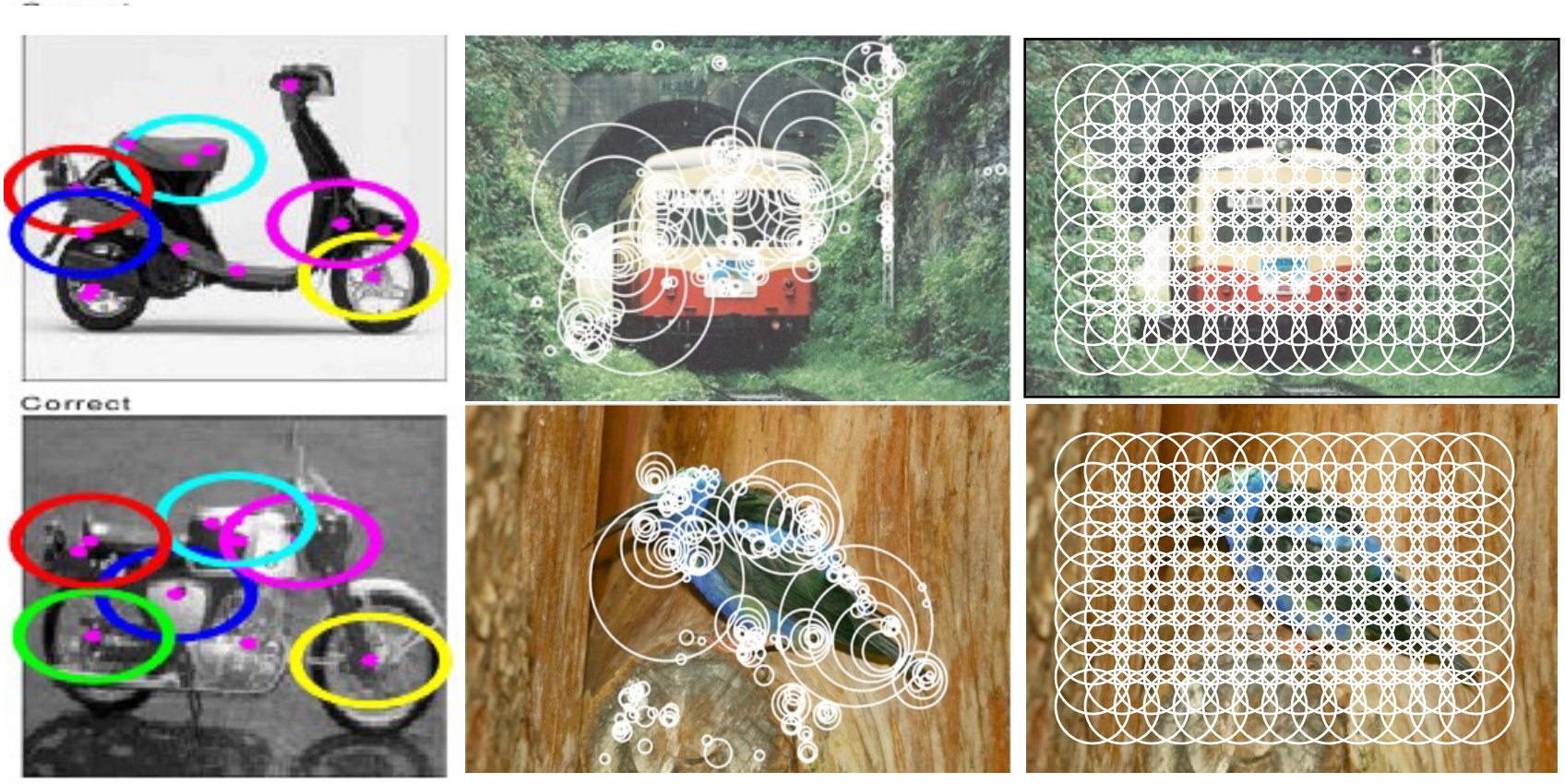
Highest ranked non class



Slide: Mark Everingham

Progress in 2003

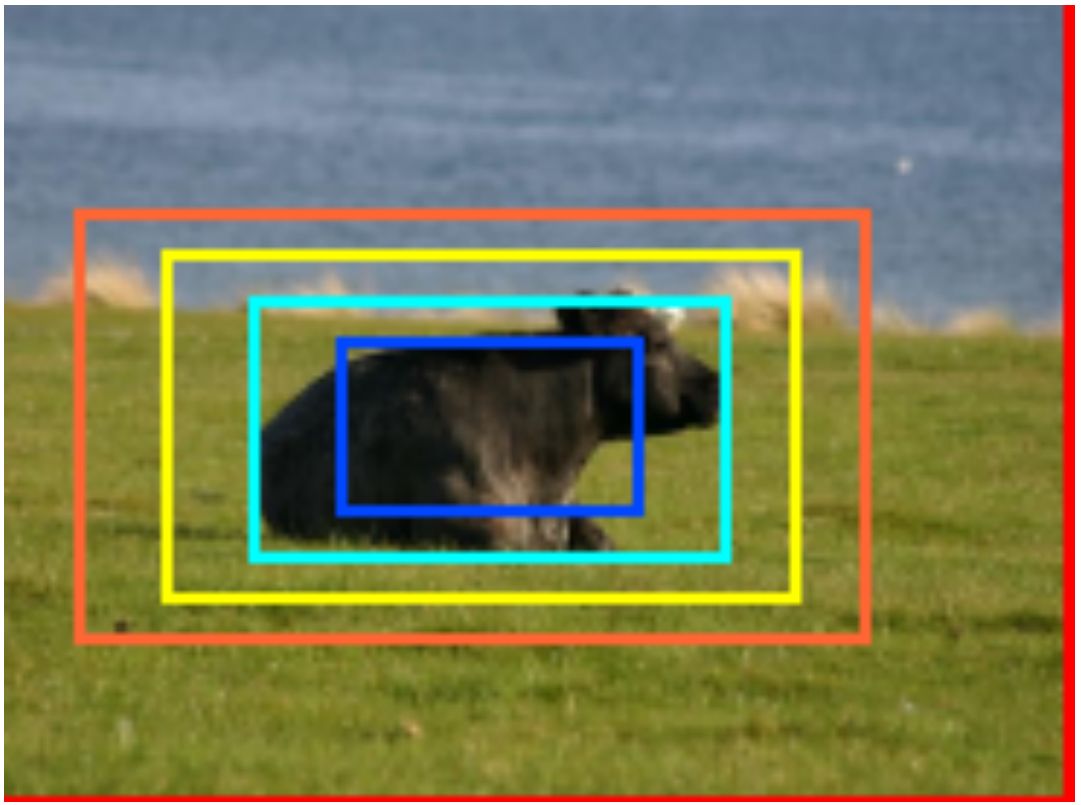
From the start by Fergus 2003 ICCV to the advances in 2008 we used more pixels but less and less locality of the object.



The spatial extent?

This trend cannot go on. Some object are non-context. When scene is cluttered, object's info drowns in the noise.

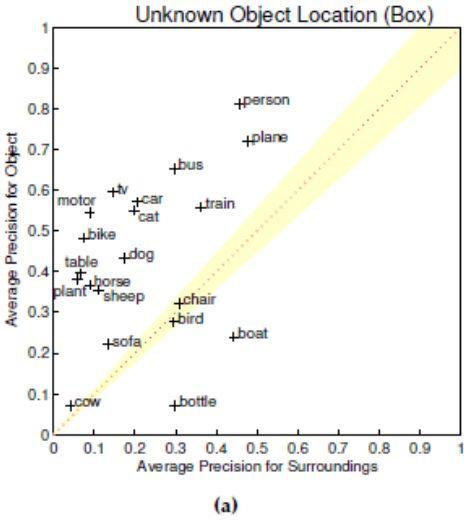
Back to the object.

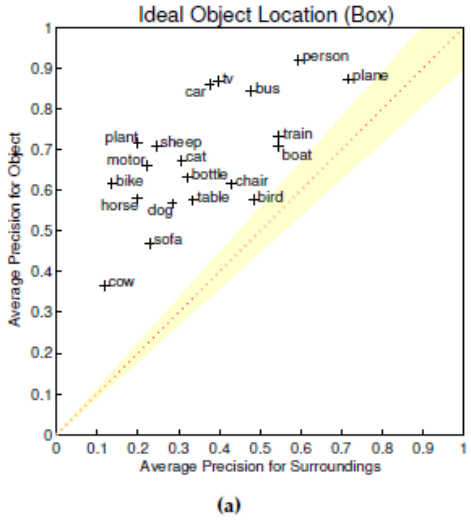


Uijlings IJCV 2012

The spatial extent?

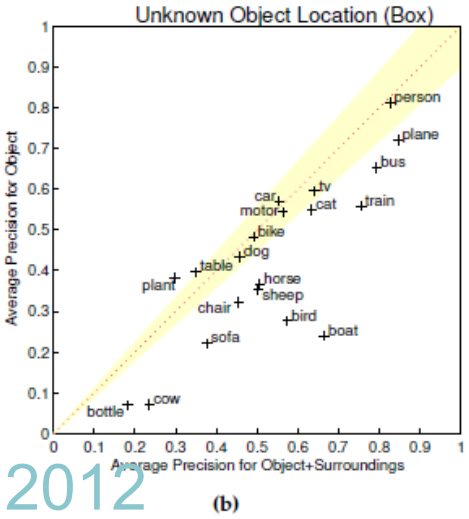
For bottle and boat, context outperforms object.

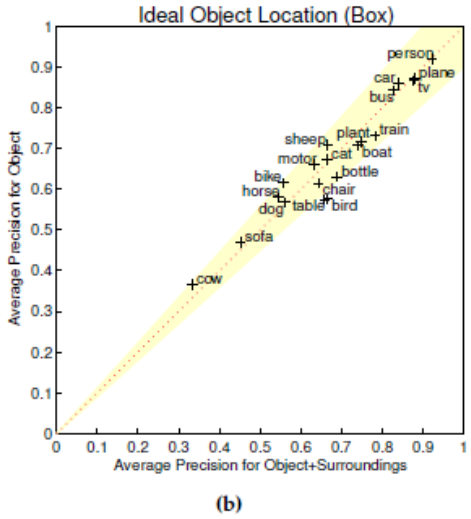




The ideal localization of objects always improves.

For bike, car, dog, table, context is irrelevant.





When ideal localization is known, context is not important.

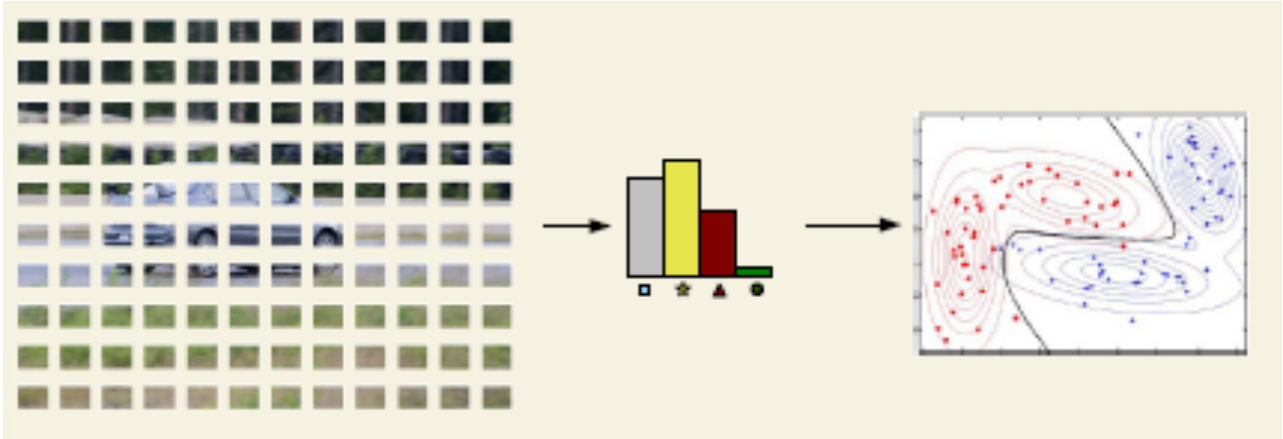
Uijlings IJCV 2012

Where is evidence?

The classification for intersection metric:

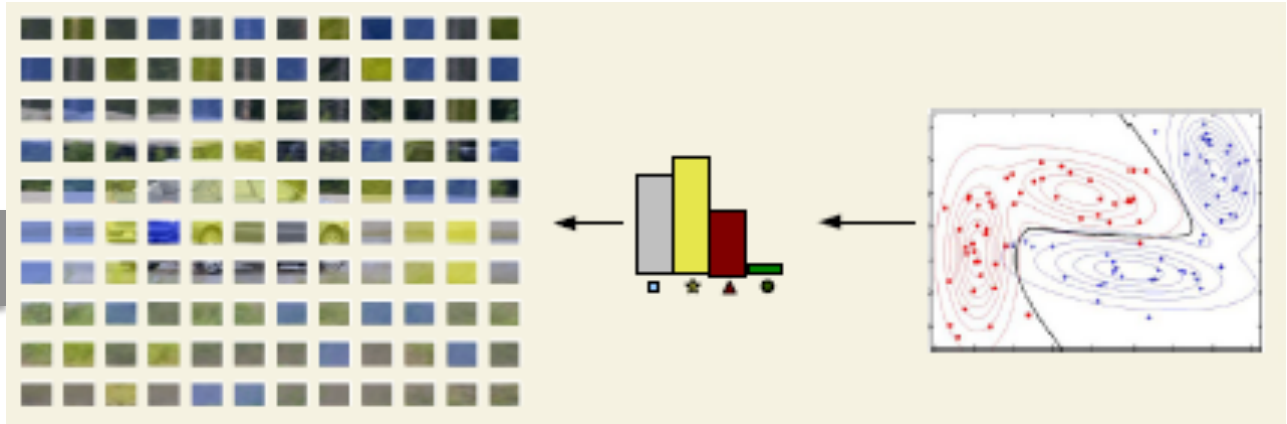
$$b + \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_j t_j \min(x_i, z_{ij})$$

Per word x and per supp. vector z take intersection and sum with alfa weight and label t.



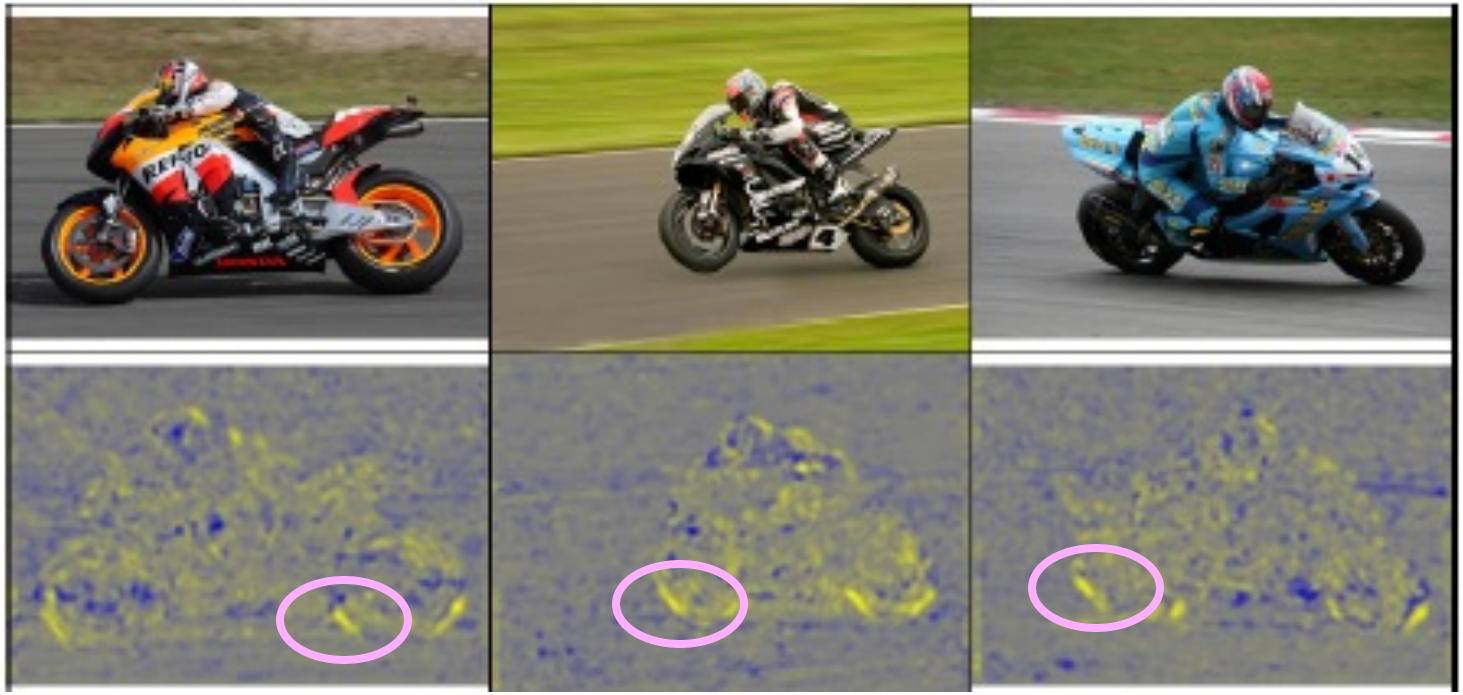
Inner sum is weight per word. Distribute over word instances

positive negative



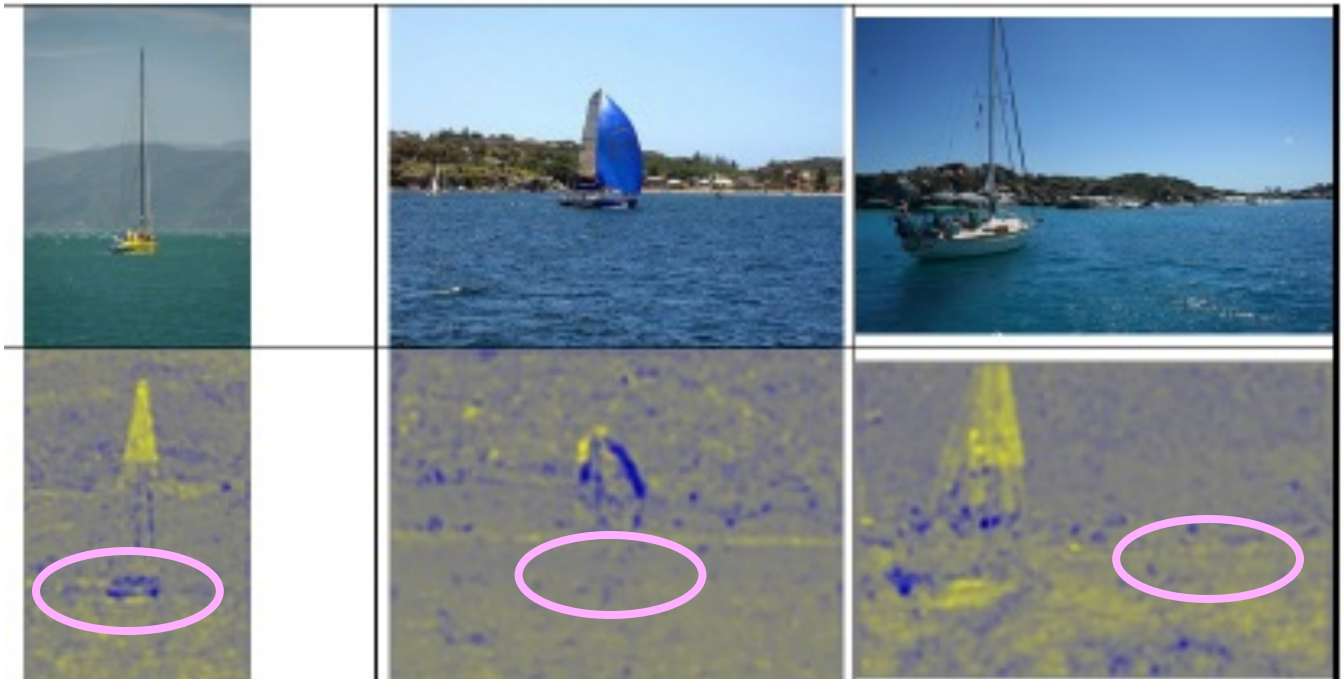
Uijlings ICCV 2011

Where is evidence for an object?



Uijlings ICCV 2011

Where is evidence for an object?



Uijlings ICCV 2011

Objects in context

Context plays an important role for some.

- Many-form objects in simple-form context.
- Cows, boats, bottle.

Hard objects rest on an integral view.

- Carry on things are context-free.
- Camera, bike, persons.

Details of the object may be decisive.

- The more classes the more details are important.
- Cats versus dogs, man versus woman.

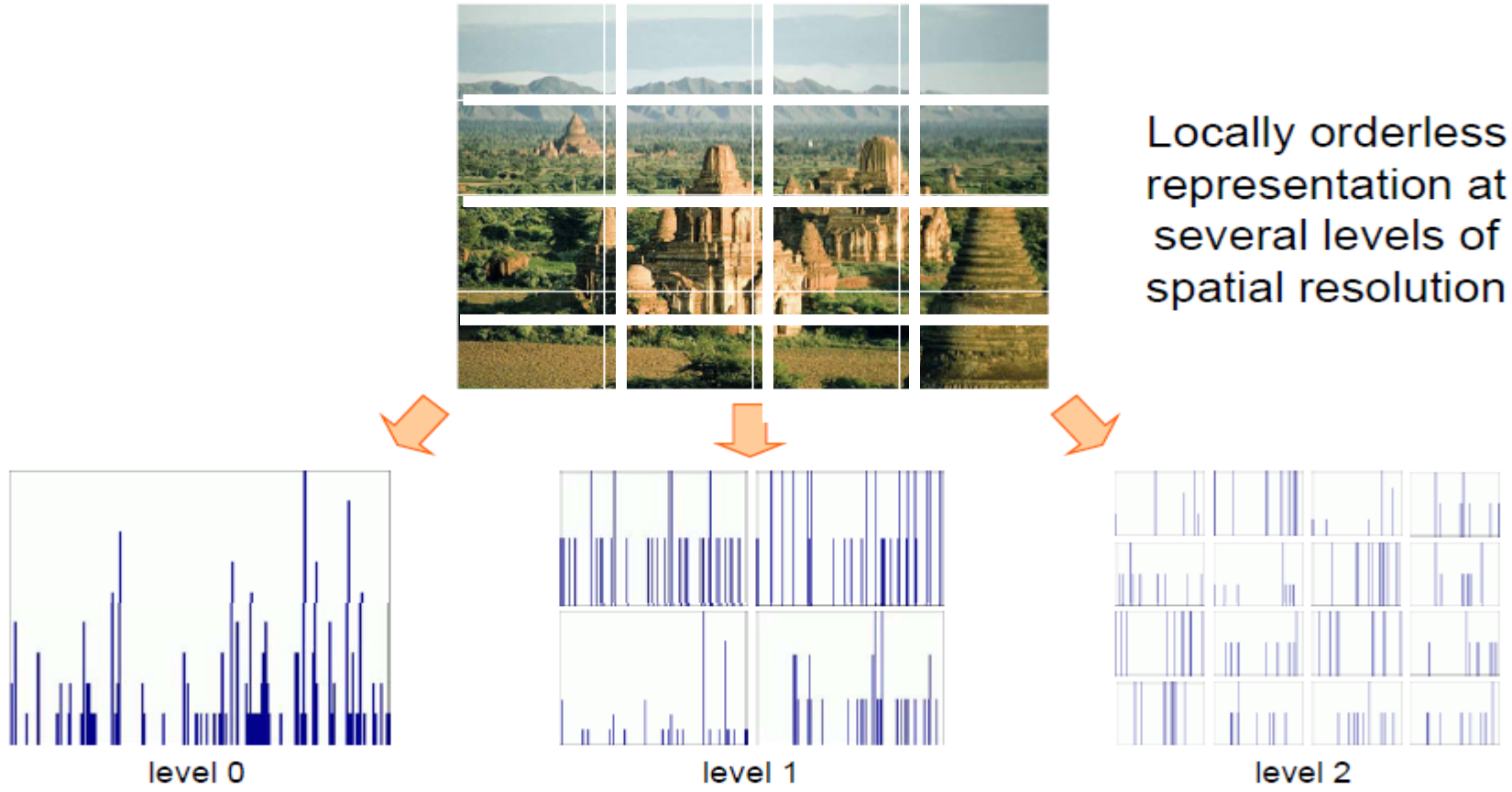
Localization

We need to reintroduce location.

Best way to do so is bottom-up. Selective search is describes the object roughly and hierarchically, exactly what is needed. A variety of features to group helps. With selective search, object class recognition goes up. Several alternatives, notably Objectness, Randomized PRIM and BING, improve the speed.

The photographer's role

Use the composition of an image to find object-related loci.



Slide credit: Cordelia Schmid

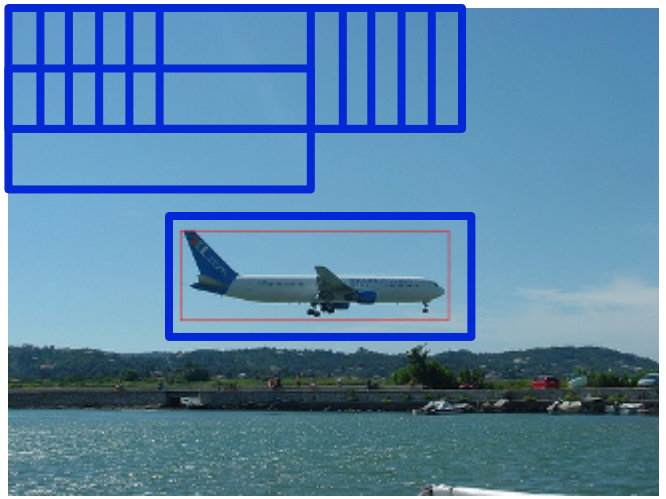
Exhaustive search for objects

Look everywhere for the object window

Imposes computational constraints on

- Very many locations and windows (coarse grid/fixed aspect ratio)
- Evaluation cost per location (weak features/classifiers)

Impressive results but takes long.



Viola IJCV 2004 Dalal CVPR 2005
Felzenszwalb TPAMI 2010 Vedaldi ICCV 2009

The need for hierarchy

An image is intrinsically semantically hierarchical.

Windows at one level of grouping will not find all objects.

The need for multiple scales

Objects may appear at different scales.

There is no fixed scale to find one object (type).

Uijlings IJCV 2013

The need for diversity

Objects are made up of image patches for many reasons.

- similar color
- similar texture
- similar shape
- enclosed shape
- same shading
- same color of the light

The need for high recall

For segmentation, find fewer but good windows

- accurate delineation
- low number of windows

For recognition, the emphasis is on rough localization

- Once discarded, an object will never be found again
- high recall (& reasonable compute time)
- less accuracy (as the context should be included)

Carreira CVPR 2010 Endres ECCV 2010 Uijlings CVPR 2009

Selective search: grouping

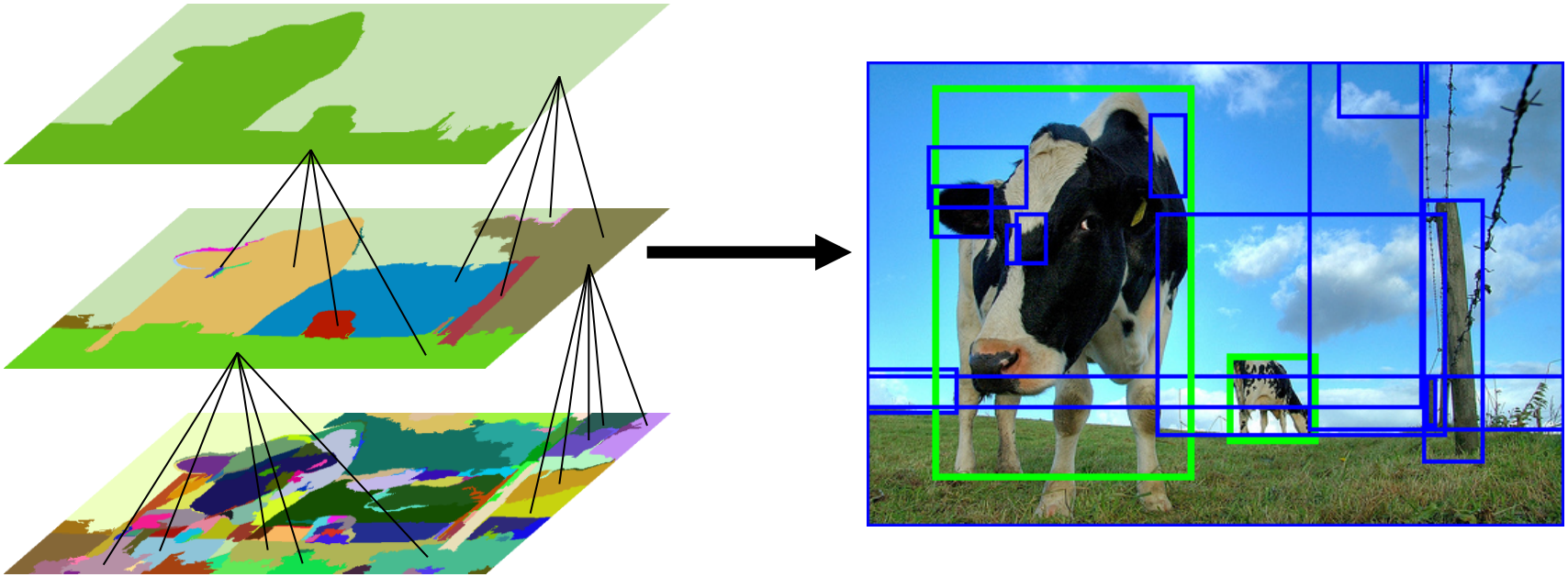
Initial over-segmentation

Ground truth

Felzenszwalb 2004

Selective search

Windows formed by hierarchical grouping.

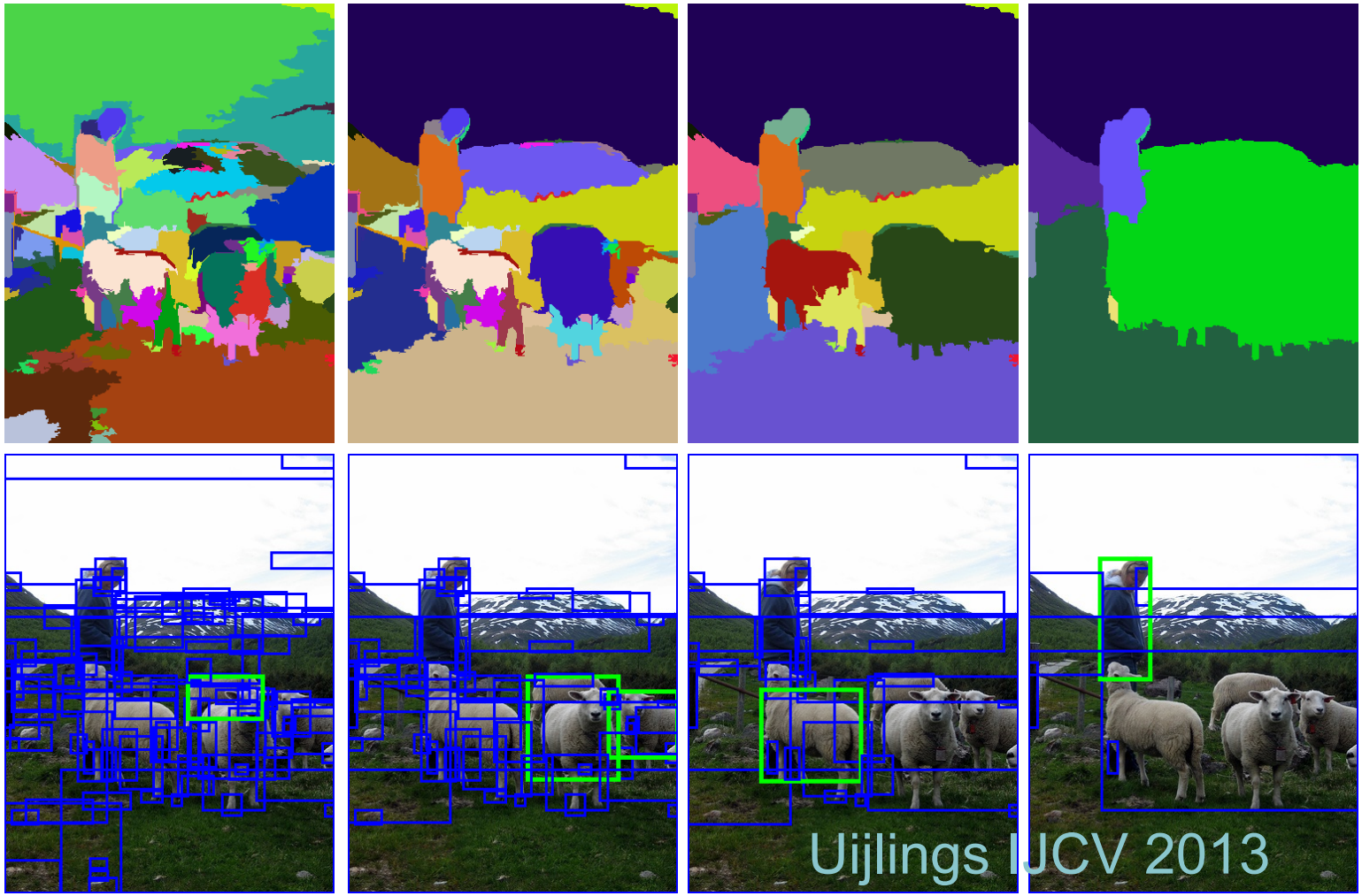


Group adjacent windows on color/texture/shape cues.

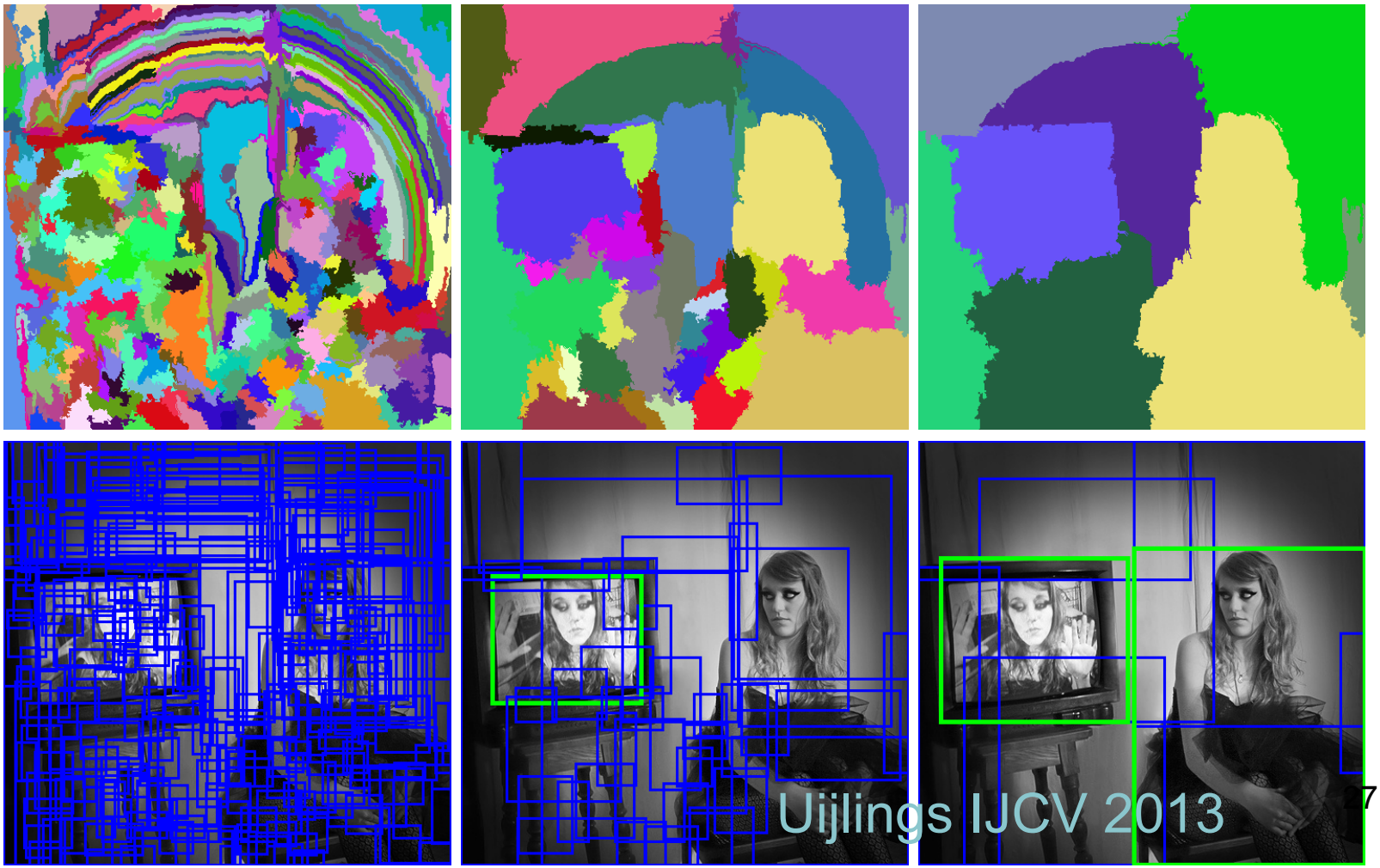
Gather all levels.

Van de Sande ICCV 2011

Selective search: grouping



Selective search: grouping



Uijlings IJCV 2013

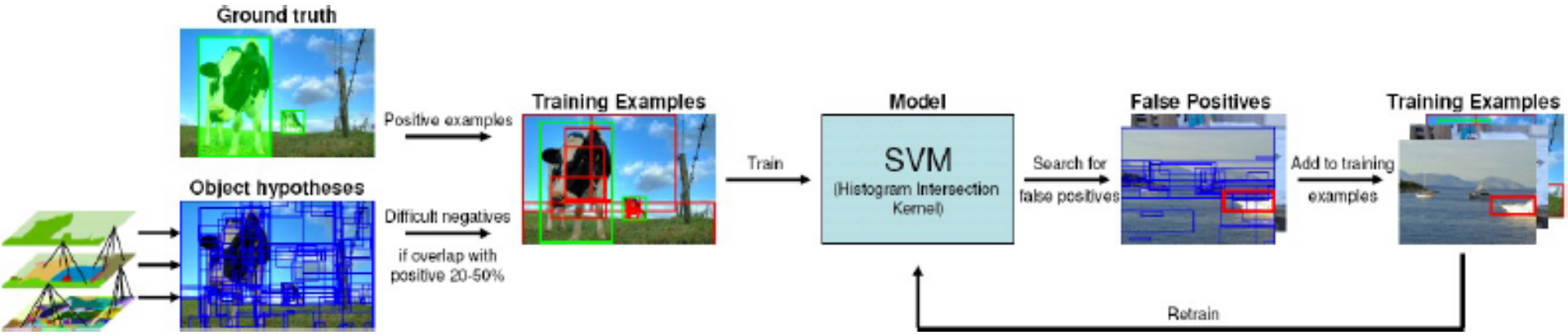
Selective search: classification

Positive window are the ones with data-driven overlap >50%.

(Hard) negatives are the ones with 20-50% overlap.

Add iteratively to training set to optimize location finding.

Use color-BoW on window to classify object.



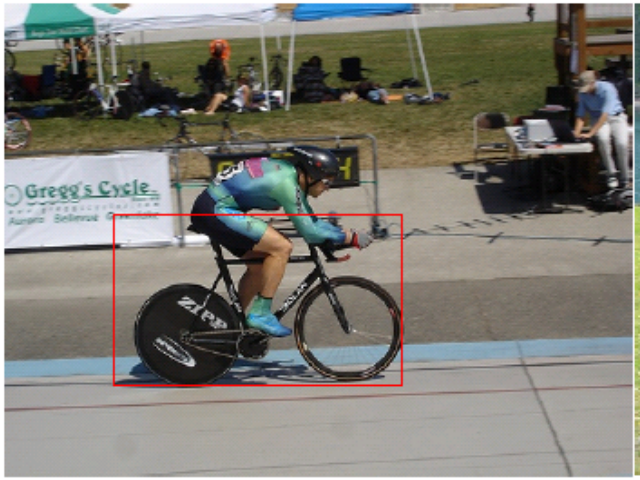
Uijlings IJCV 2013

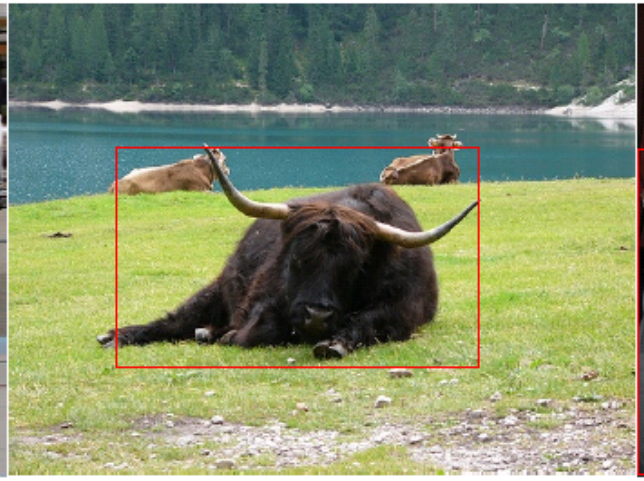
Mean Average Best Overlap ~88%

Mean over all 20 classes

Avarage within the class.

MABO of 88% looks like this:

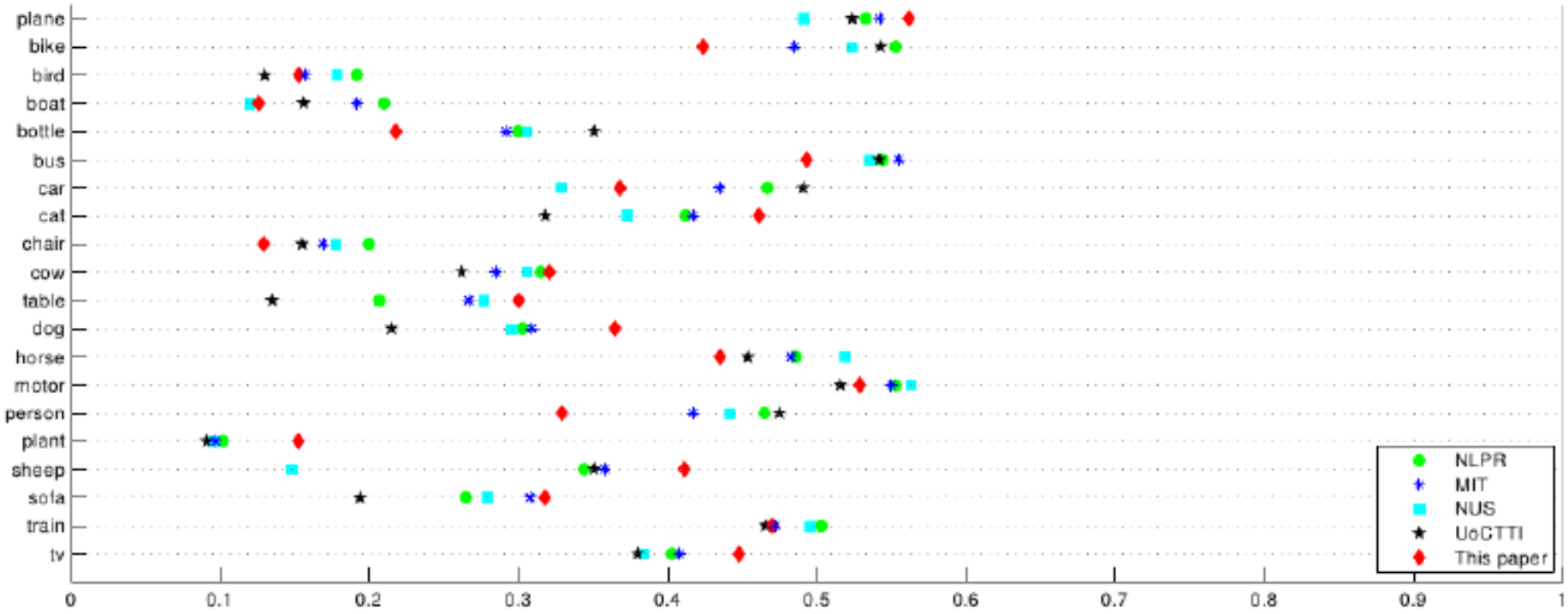





Van de Sande ICCV 2011

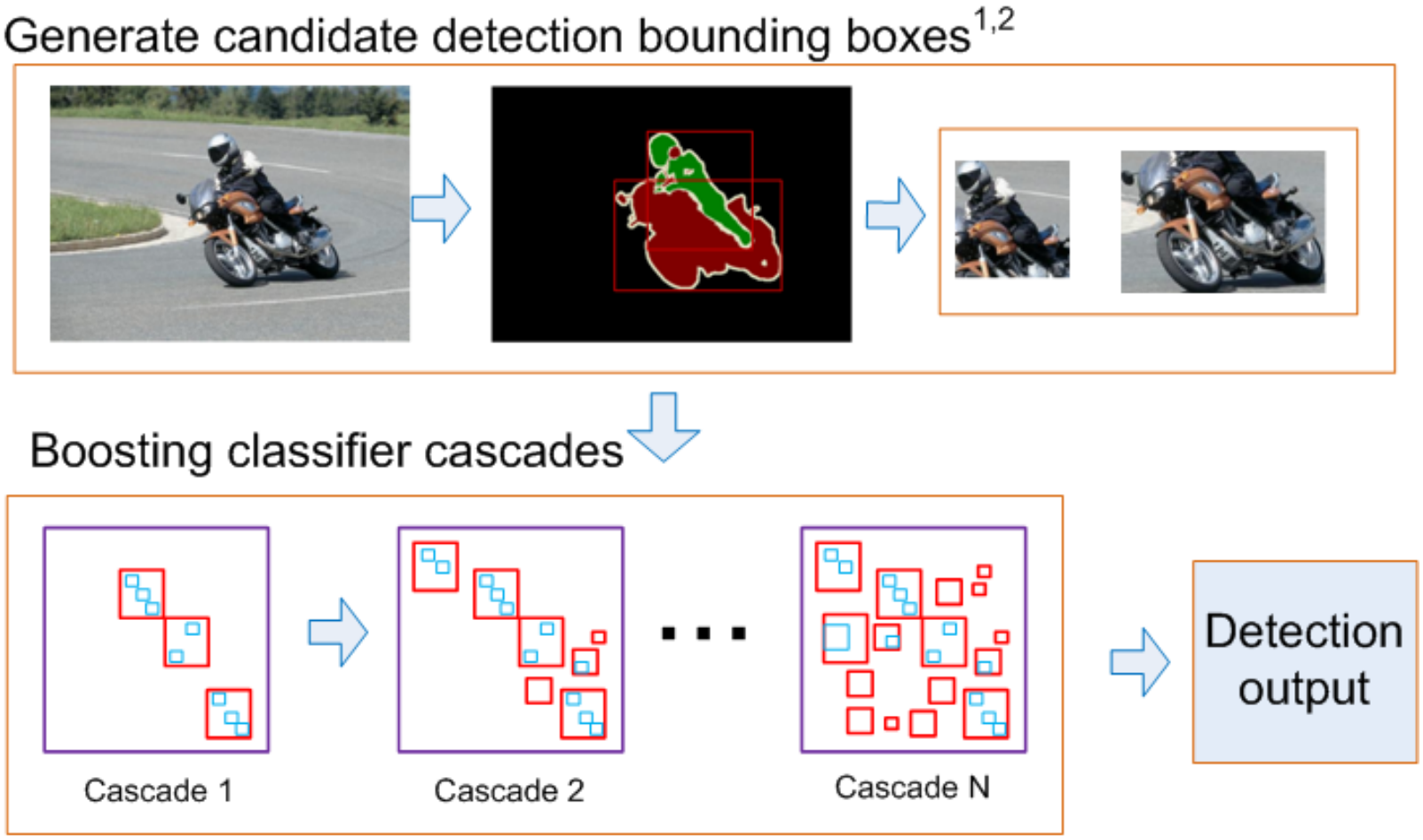
Results

Pascal VOC 2010, best in 9 out of 20.



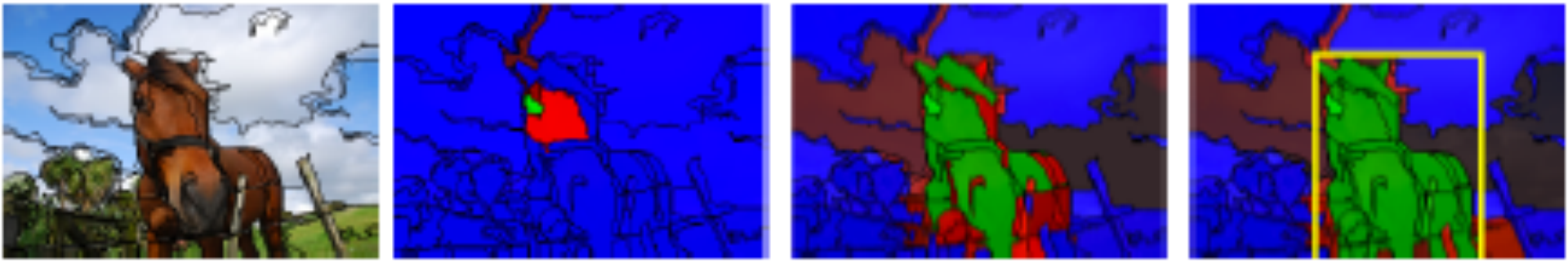
Van de Sande ICCV 2011

Alternative 1: Region-lets

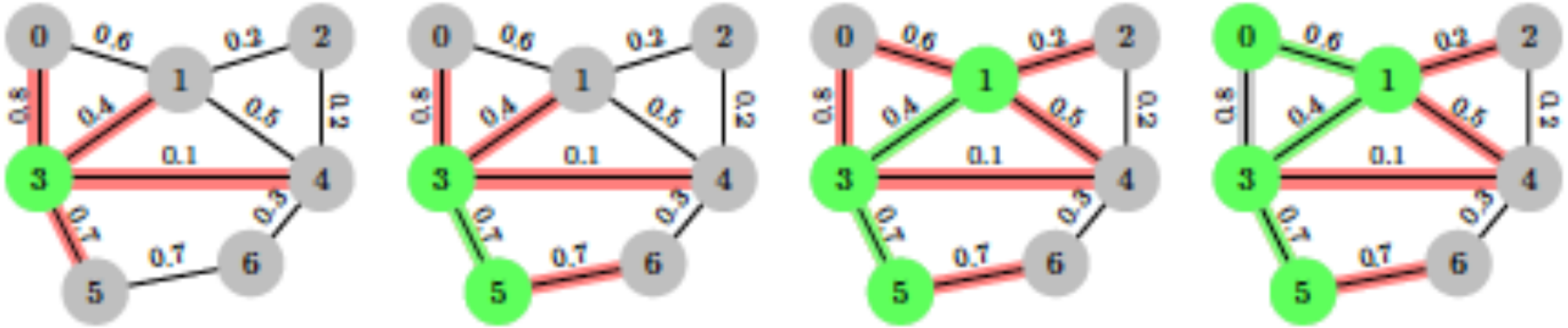


Xiaoyu Wang ICCV 2013

Alternative 2: Random-PRIM

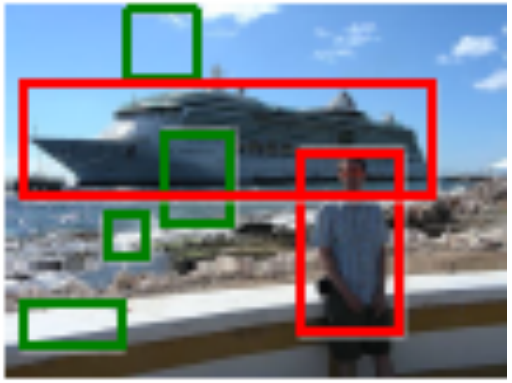


Superpixel segmentation + start at random superpixel (green) + expand with a *randomized* most similar neighbor or return a box.



Maanen ICCV 13

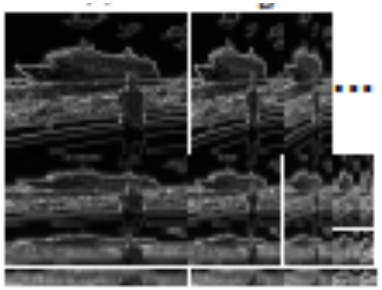
Alternative 3: BING

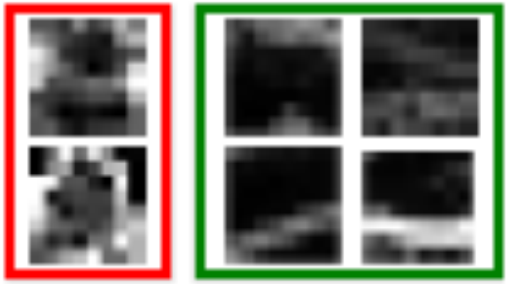


Original image

Red = true 1 & 2.

Green = false.





Gradient maps at various scales. Their normed gradients look similar after rescaling. NG holds a 64D normed gradient feature. Binarize and learn from NG, x, s by binSVM. Once learned also suits unseen types.

Two concepts

Two concepts tell a story, the story of the image.
Localization is needed to make it work.

Bi-concept by windows

The story an image tells is about pairs of things.

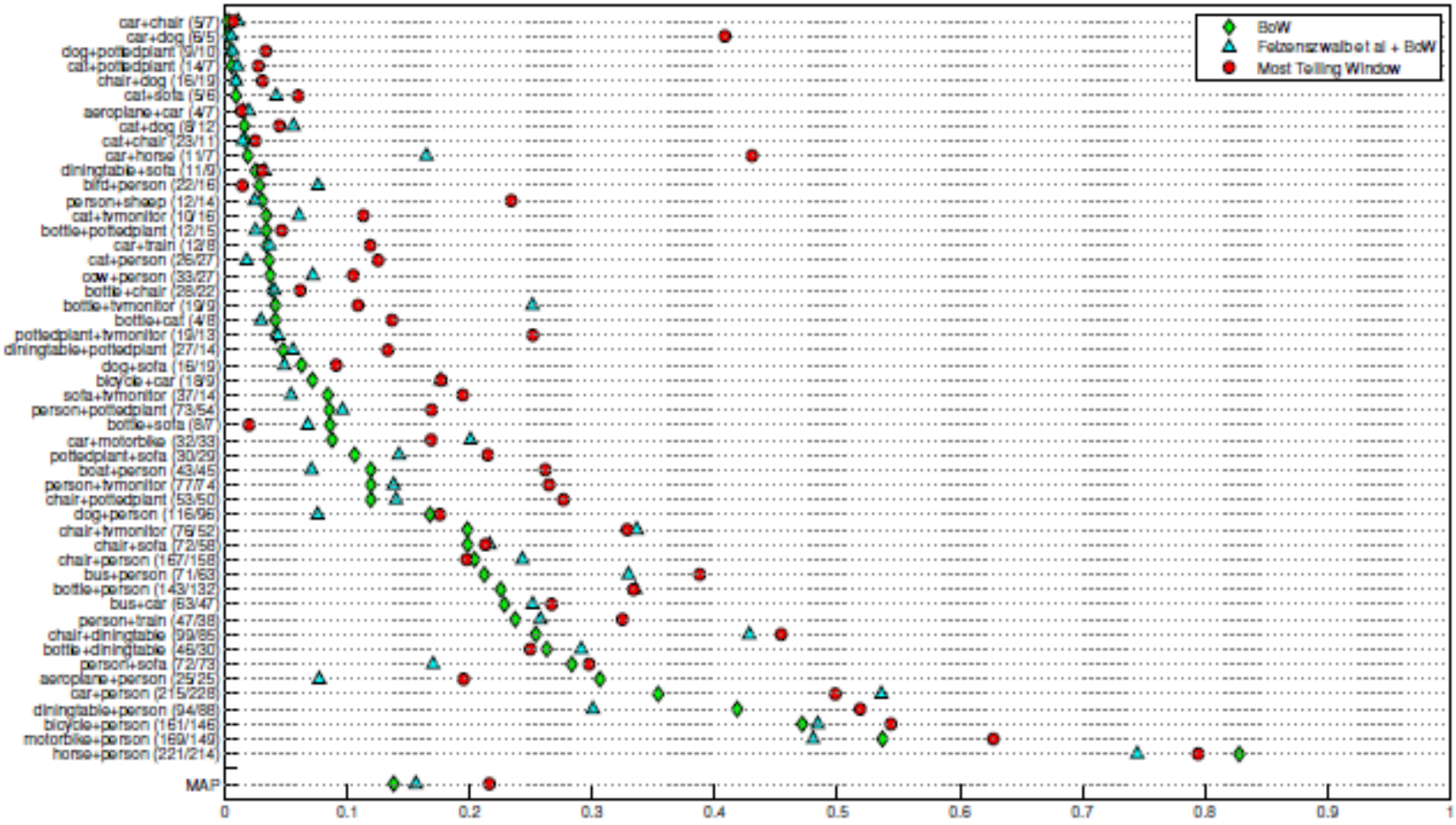
(a) (b) (c)

(d) (e) (f)

For pairs of things, one needs the most telling window.

Uijlings ICCV demo 2012

Bi-concepts by windows



Uijlings ICCV demo 2012

Bi-Concept by harvesting

Find images showing “a horse next to a car”.

Search in Google for “horse car”.

Horses in “horse car” do not look like normal “horses”.

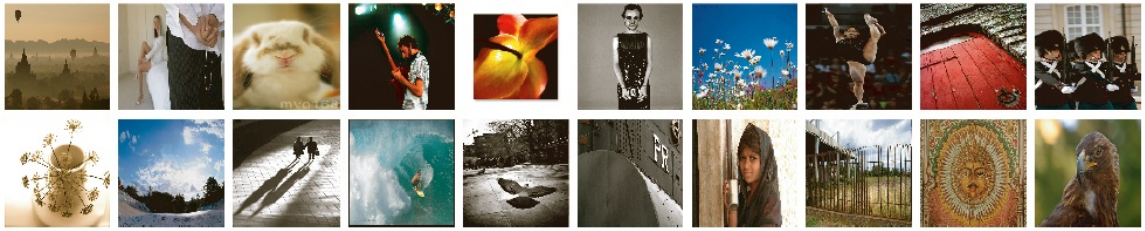
Xirong Li 2012 IEEE trans MM

Bi-Concept by harvesting

Combing single concepts does not work

p(car|x)

p(horse|x)



p(car|x)*p(horse|x)

Xirong Li 2012 IEEE trans MM

Bi-Concept by harvesting

Social data size: use single class images as hard negatives.

car +horse

car +showroom

cat+flower

cat +snow

Xirong Li 2012 IEEE trans MM

Two concepts

Bi-concepts are not two times the concept.

a. Location via selective search.

b. Harvesting with single class as negatives.

beach + girl + horse

Xirong Li 2012 IEEE trans MM

Some references

[1] K.E.A. van de Sande, Th. Gevers, C.G.M. Snoek. **Evaluating Color Descriptors for Object and Scene Recognition.** *PAMI*, 32(9):1582-1596, 2010.

[2] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders, J.M. Geusebroek. **Visual Word Ambiguity.** *PAMI*, 32(7): 1271-1283, 2010.

[3] E. Gavves, C.G.M. Snoek, A.W.M. Smeulders. **Convex Reduction of High-Dimensional Kernels for Visual Classification.** *CVPR*, 2012.

[4] X.Li, C.G.M. Snoek, M. Worring. **Unsupervised Multi-Feature Tag Relevance Learning for Social Image Retrieval**. *Proc ACM ICIVR*, Xi'an, 2010, best paper.

[5] C.G.M. Snoek, A.W.M. Smeulders. **Visual-Concept Search Solved?** *IEEE Comp*, 43:76-78, 2010.

[6] C.G.M. Snoek, et al. **The MediaMill TRECVID 2011 Semantic Video Search Engine**. In *Proc 8th TRECVID Workshop*, USA, 2011.

[7] J.R R. Uijlings, A.W.M. Smeulders, R.J.H. Scha. **Real-Time Visual Concept Classification.** *IEEE Trans. Multimedia*, 12(7):665-681, 2010, best paper.

[8] J.R.R. Uijlings, A.W.M. Smeulders, R.J.H. Scha. **The Visual Extent of an Object - Suppose We Know the Object Locations.** *IJCV*, 96 (1):46-63, 2012

[9] K.E.A. van de Sande, J.R.R. Uijlings, T.Gevers, A.W.M. Smeulders. **Segmentation As Selective Search for Object Recognition.** *ICCV,* 2011.

[10] A.W.M.Smeulders, M.Worring, S.Santini, A.Gupta, R.Jain. **Content Based Image Retrieval at the End of the Early Years**. *PAMI, 22*(12):1349-1380, 2000.