Computer Vision by Learning

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

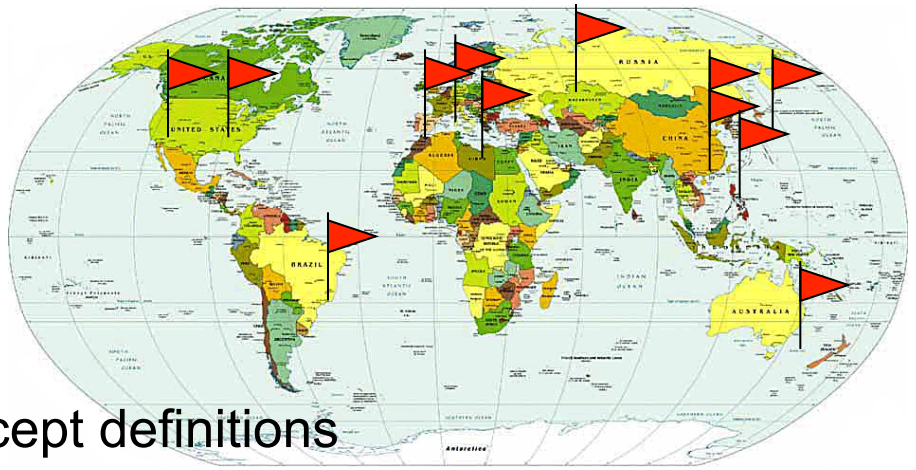
Advanced School for Computing and Imaging
University of Amsterdam
TU Delft

Administration

Today

Lectures	0930-1215	D1.115
Lunch	1215-1330	on your own
Lab	1330-1600	D1.111
Borrel	1600-zzz	D1.111

Evaluation of computer vision



Situation in 2000

- Various video concept definitions
- *Specific* and *small* data sets
- Hard to compare methodologies

= Researchers

For object tracking still the case in 2013

Overview

1. Video benchmarks, TRECVID, average precision, progress
2. Image benchmarks, PASCAL, ImageNet, lessons learned
3. Labels from humans, experts, volunteers, crowdsourcing
4. Labels from similarity, nearest neighbor, simple features
5. Negative labels, negative bootstrapping, model compression

6. Learning using attributes

1. Video benchmarks

Crucial drivers for progress in large-scale computer vision are international search engine benchmarks. The National Institute of Standards and Technology's TRECVID (TREC Video Retrieval) benchmark has played a significant role. The main goal of TRECVID is to promote progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation. TRECVID is a laboratory-style evaluation that attempts to model real world situations or significant component tasks involved in such situations.

International competition

NIST TRECVID Benchmark

Promote progress in video retrieval research

Open data, tasks, evaluation *and* innovation

http://trecvid.nist.gov/

Video data sets

US TV news (`03/`04)

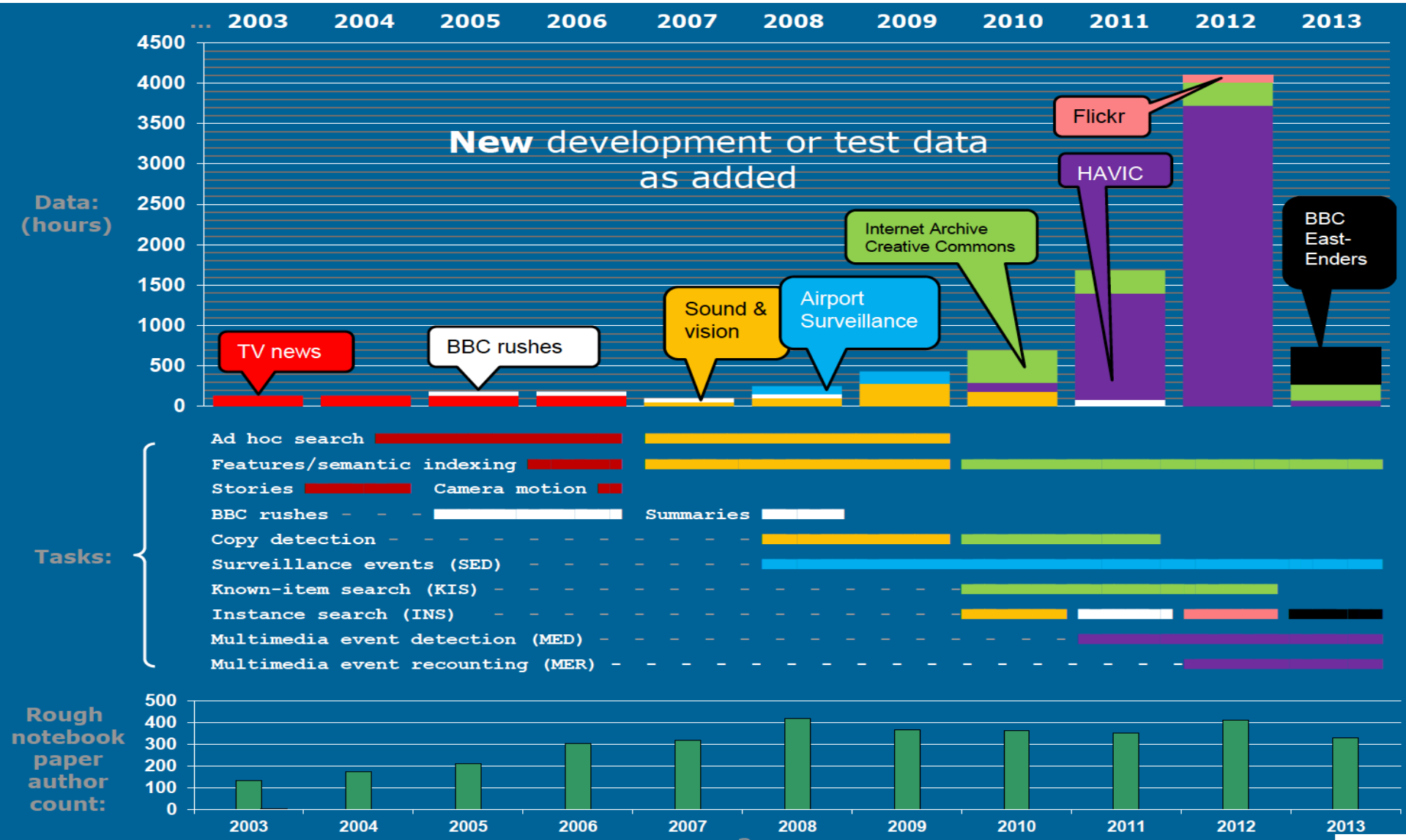
International TV news (`05/`06)

Dutch TV infotainment (`07/`08/`09)

Web video (since 2010)

Slide Credit: Paul Over, NIST

NIST TRECVID evolution



Task: concept detection

Goal

- Build benchmark collection for visual concept detection methods

Secondary goals

- encourage generic (scalable) methods for detector development
- semantic annotation is important for search/browsing

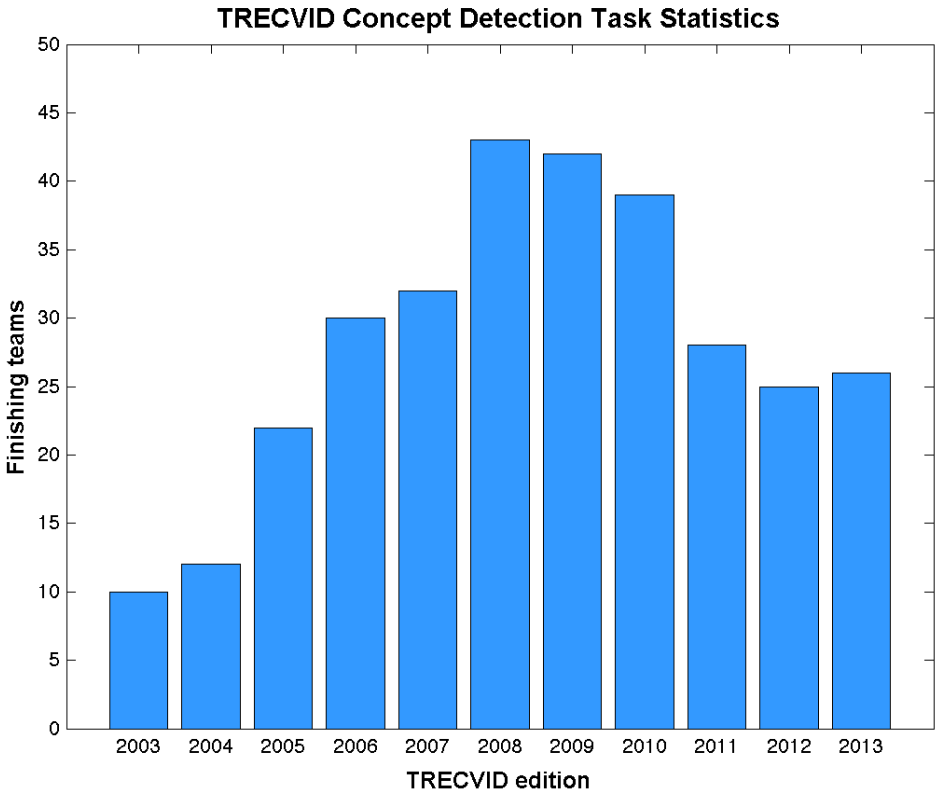
Aircraft

Beach

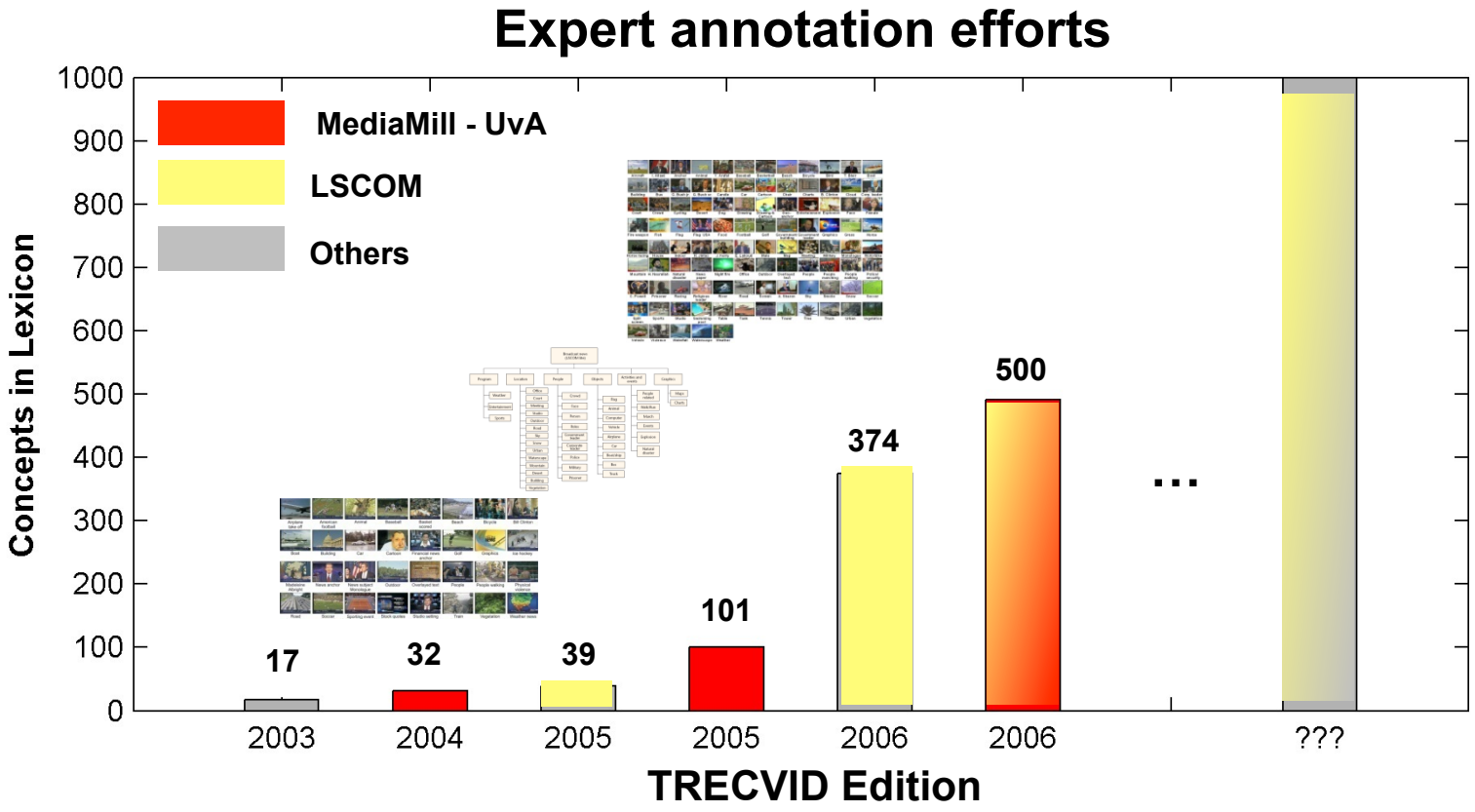
Mountain

Note the variety in visual appearance

De facto evaluation standard



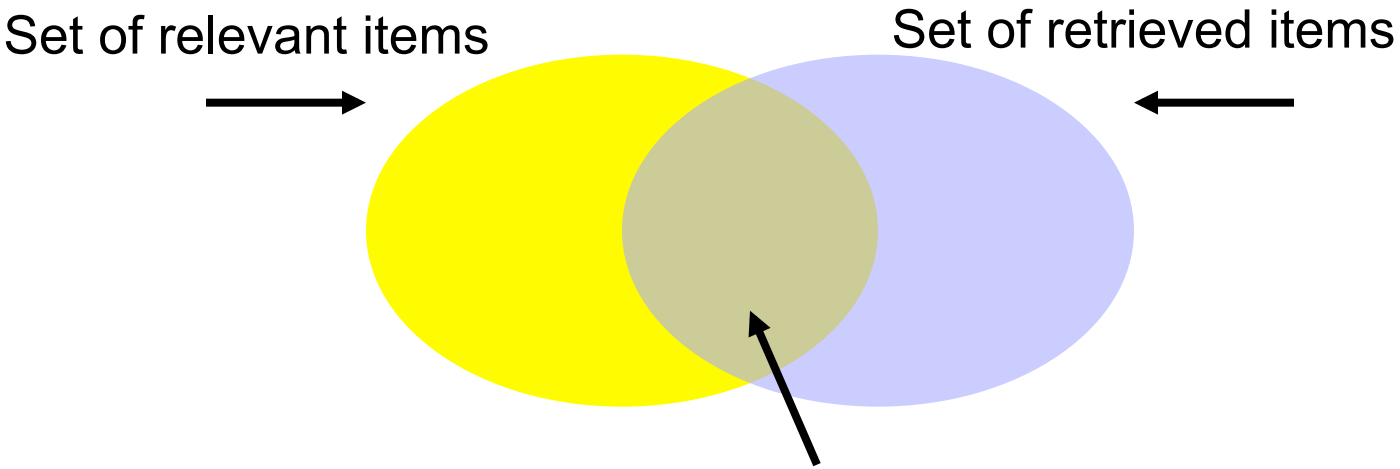
Annotation efforts



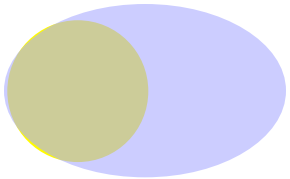
Measuring performance

Results

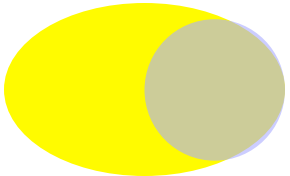
1. 
2. 
3. 
4. 
5. 



Precision



Recall



inverse relationship

Evaluation measure

Average Precision

- Combines precision and recall
- Averages precision after relevant shot
- Top of ranked list most important

$$AP = \frac{\sum_{r=1}^{N}(P(r) \times \mathrm{rel}(r))}{\text{number of relevant documents}}$$

$$AP = \frac{1/1 + 2/3 + 3/4 + \ldots}{\text{number of relevant documents}}$$

Results

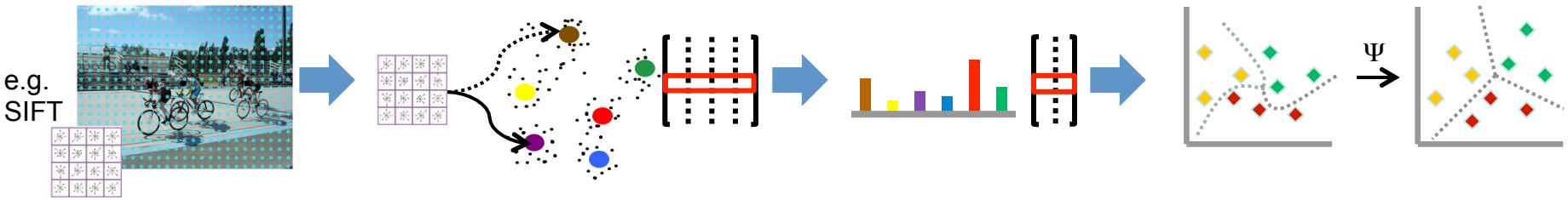
1. 
2. 
3. 
4. 
5. 

2003: no clue!

2010: Bag-of-words

Snoek et al, TRECVID 2008-2010
Van de Sande et al, PAMI 2010
Van Gemert et al, PAMI 2010

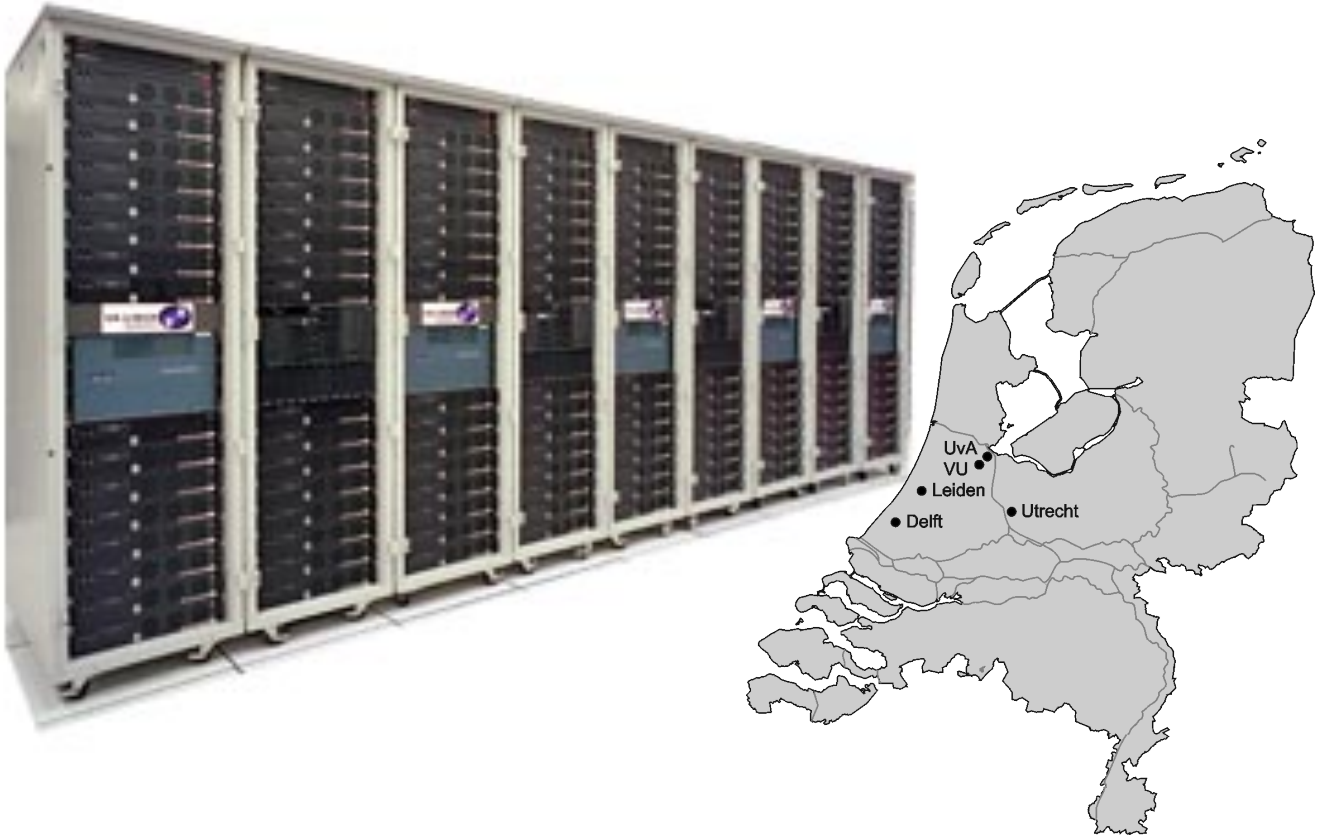
Color SIFT, soft assignment and kernel approximations.



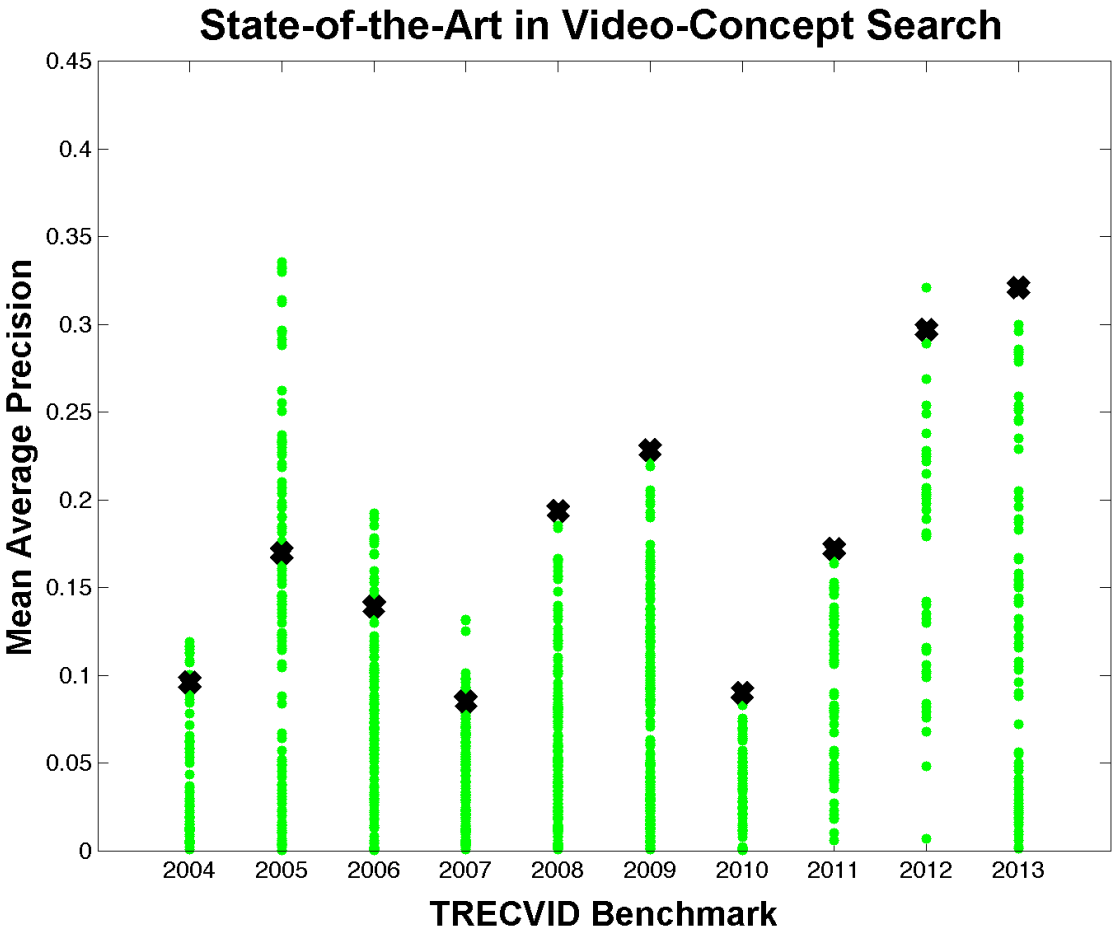
Software available for download at http://colordescriptors.com

Benchmarking is compute intensive

Distributed ASCI super computer: *priceless*



Are we making progress?

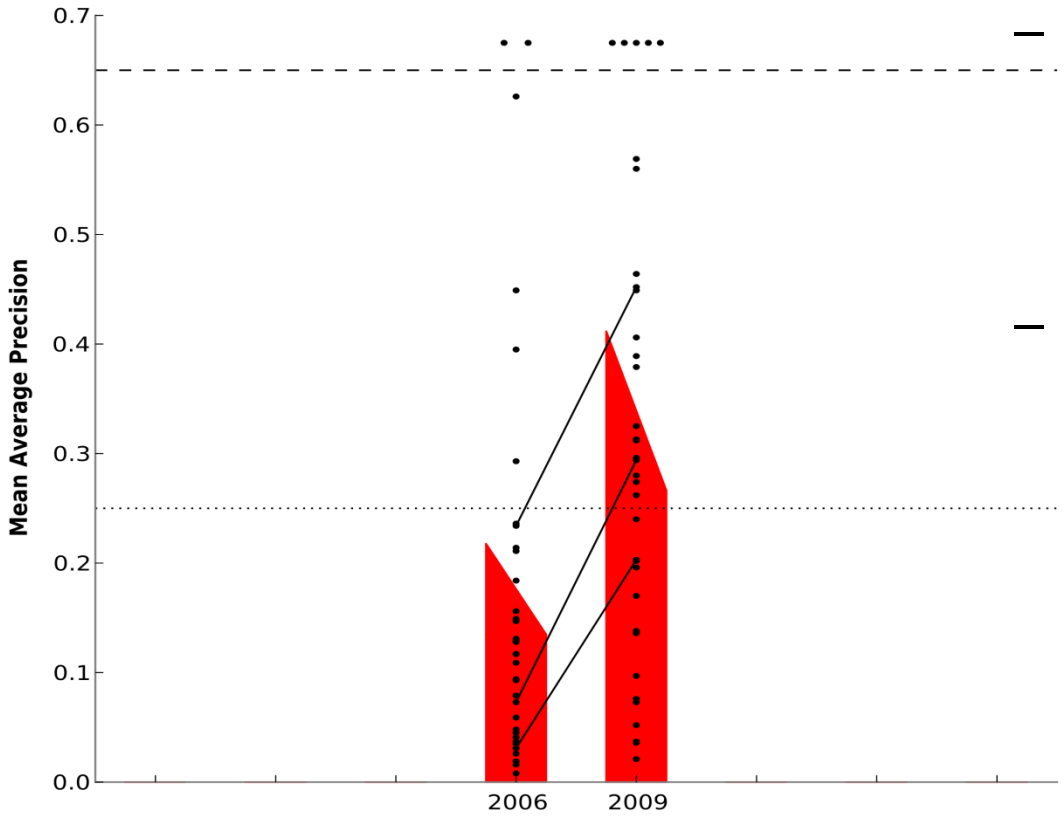


x **MediaMill team**

- 1000+ others

Performance doubled in 3 years

- 36 concept detectors

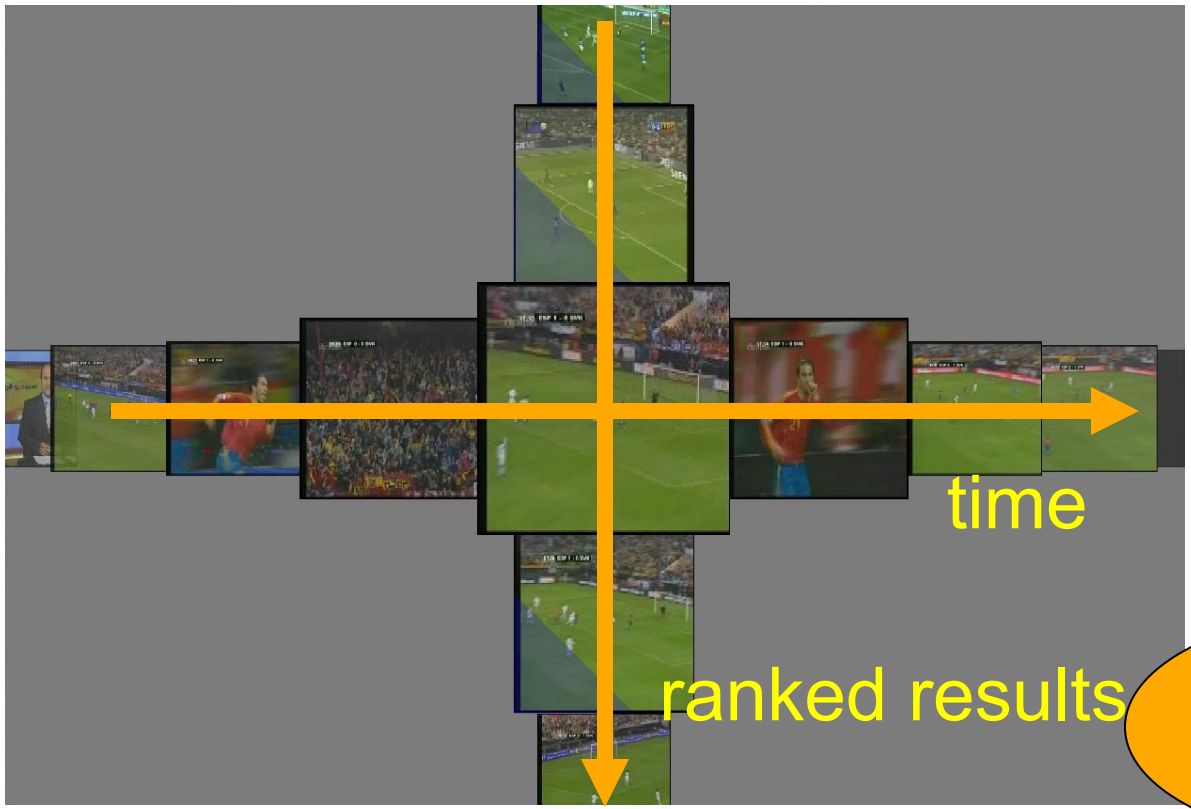


- Even when using training data of different origin
- Vocabulary still limited

Snoek & Smeulders, IEEE Computer 2010

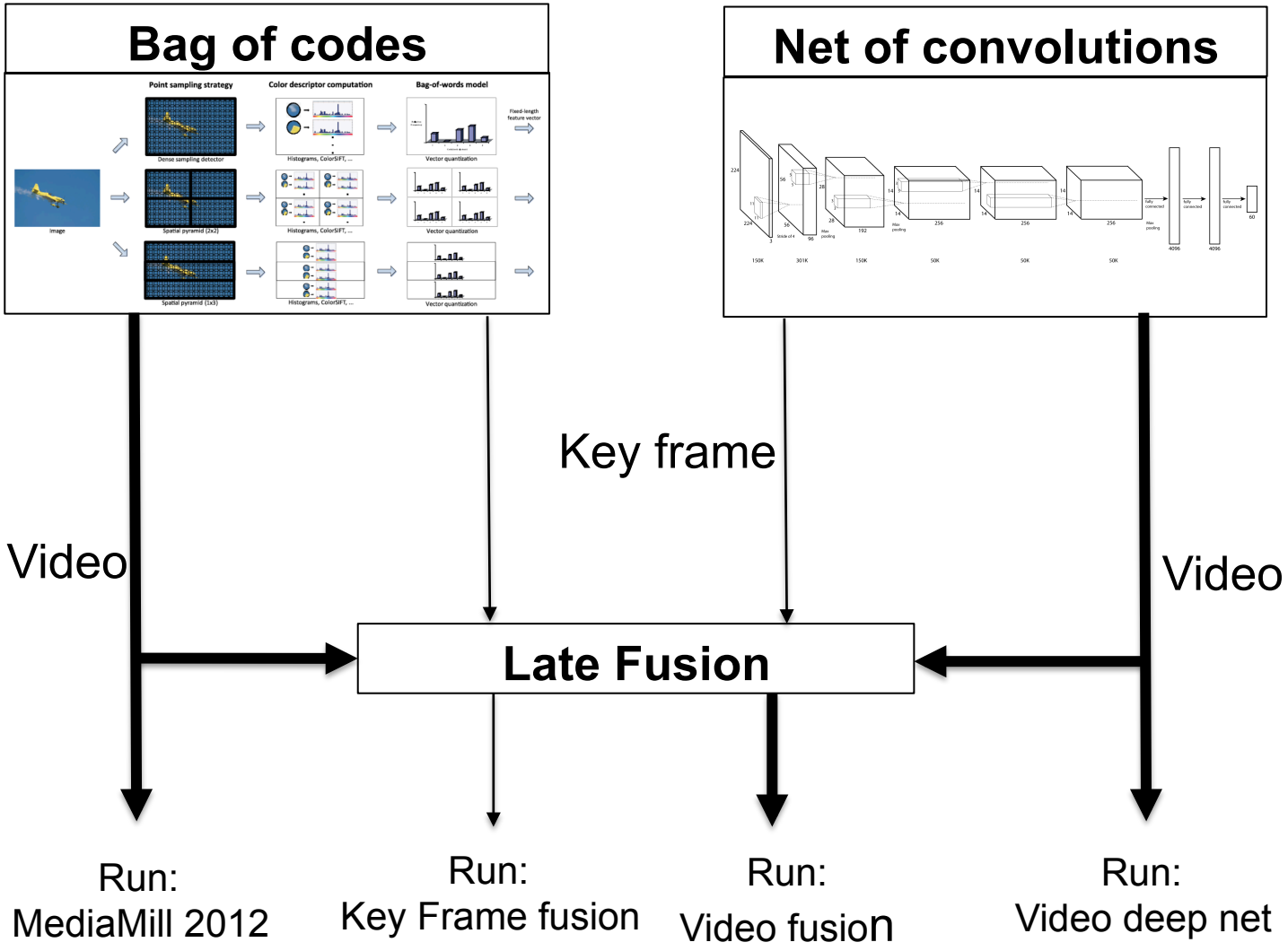
MediaMill video search engine

CrossBrowser combines query results and time



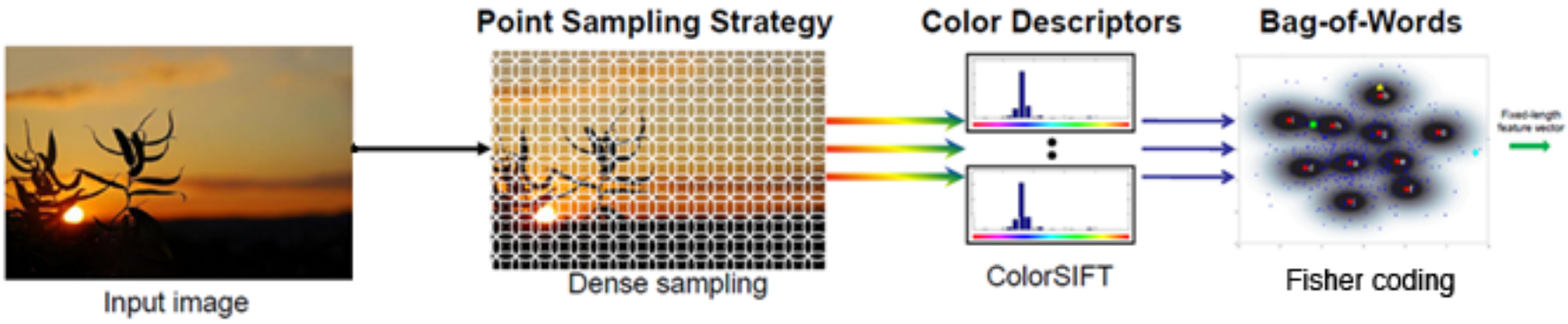
2010 Version

MediaMill TRECVID 2013



MediaMill: Color Fisher coding

Densely sampled points
SIFT, RGB-SIFT and T-SIFT descriptors
PCA reduction to 80D
Fisher vector coding with codebook size 256
Spatial pyramid 1x1+1x3
Linear classifier



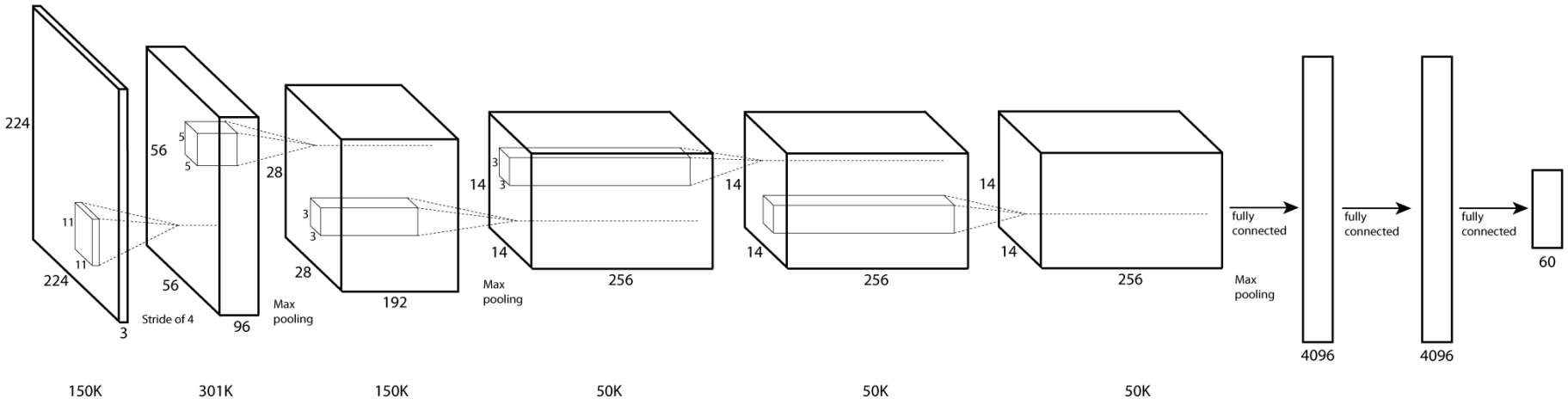
Color Descriptor software available for download at http://colordescriptors.com

MediaMill: Video deep learning

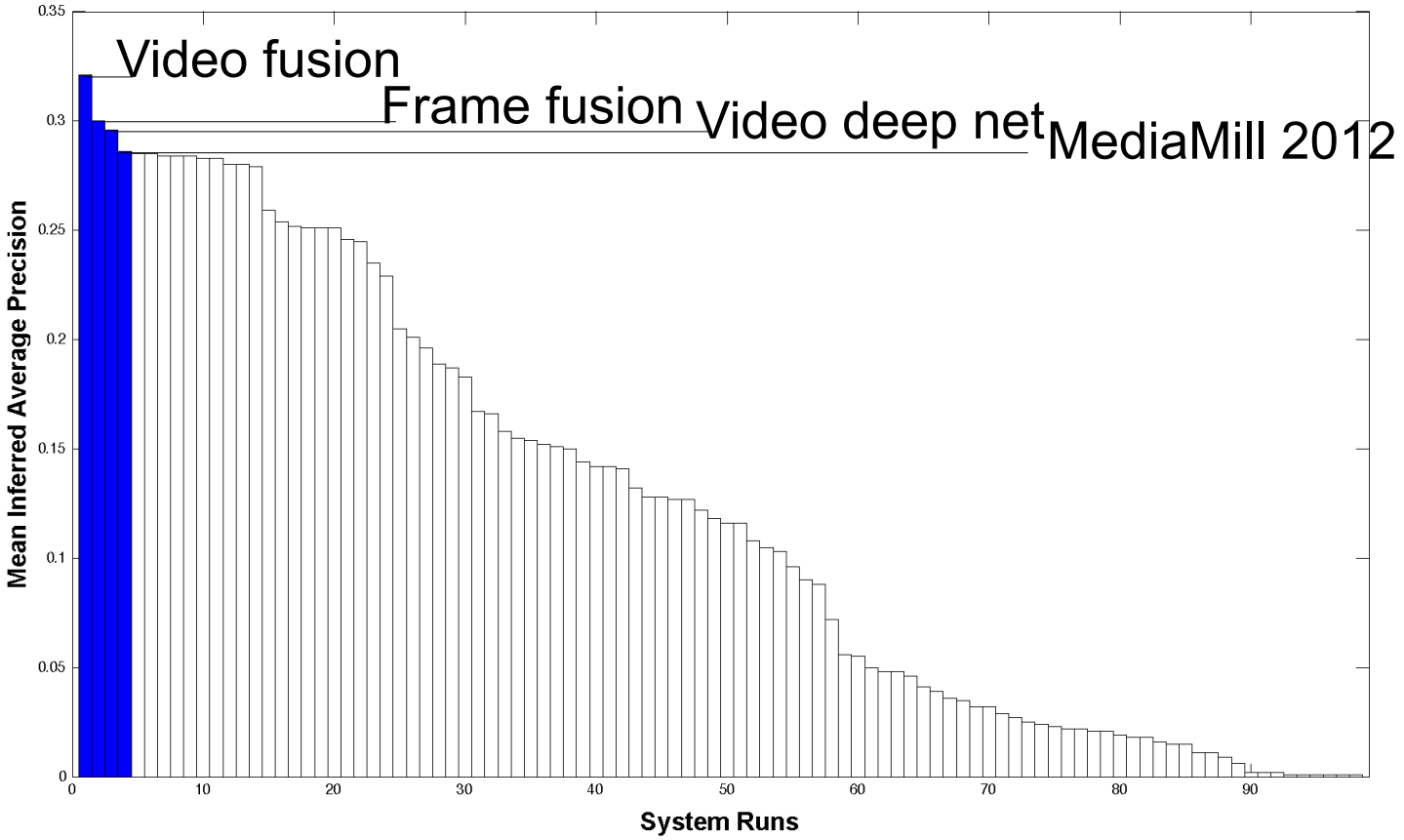
Convolutional neural network with 8 layers with weights

Trained using error back propagation

- ImageNet for pre-training

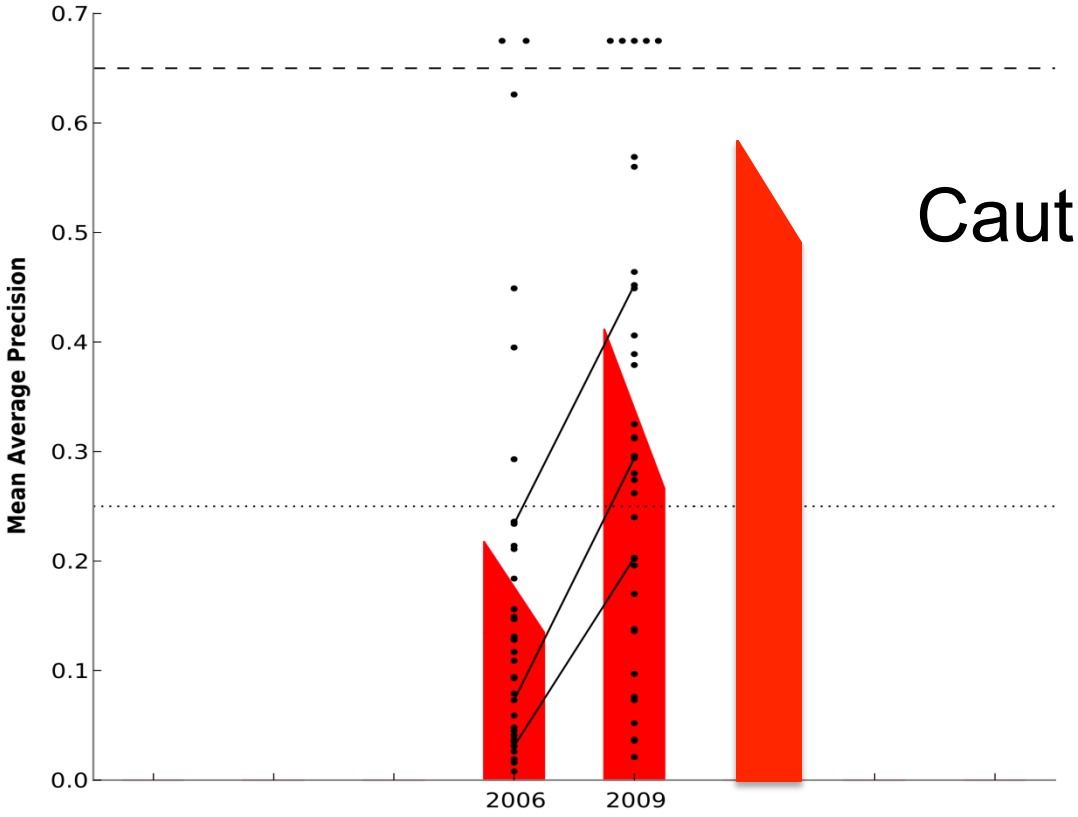


Results



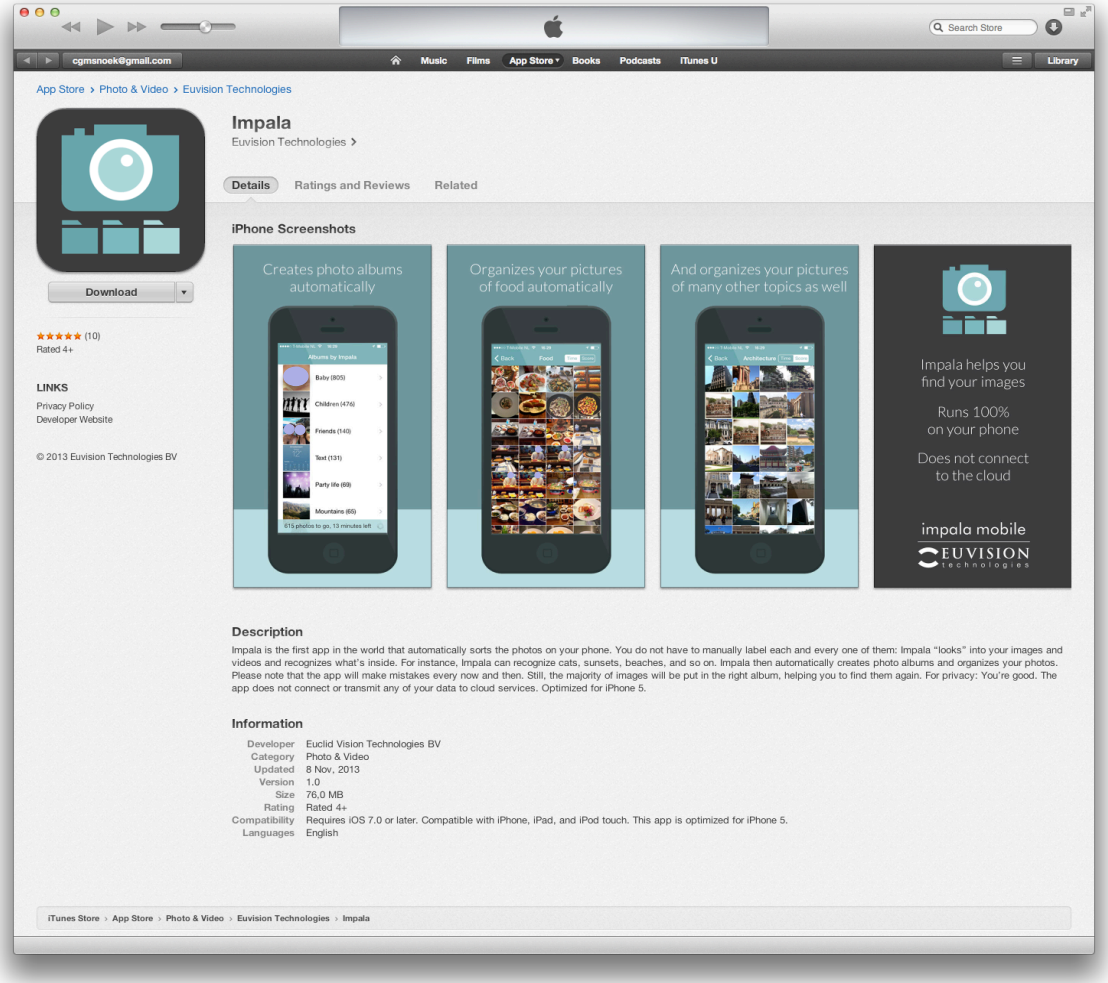
Bag of codes and deep net profit from each other

Performance doubled again?



Caution: To be tested…

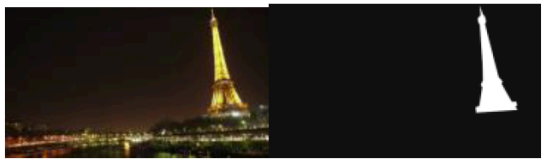
Impala iPhone App



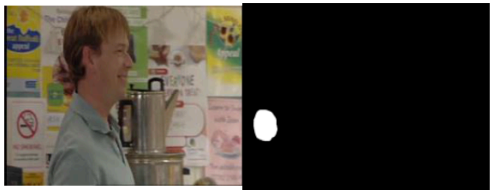
Future challenge: Instance search

Given a single query example, including a segmentation mask, find similar occurrences of the named instance in a collection of video.

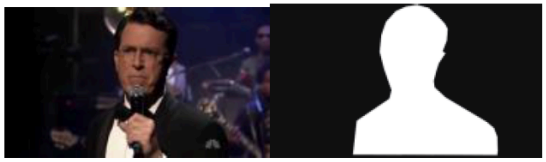
instance “Eiffel tower”



instance “a circular ‘no smoking’ logo”



instance “Stephen Colbert”



instance “an Audi logo”

instance “this man”

Future challenge: event recognition

Given 100, 10 or 0 training example videos, recognize and recount videos in a huge test collection containing the event of interest.

Working on a metal project

Cleaning an appliance

Object, scene and action detectors are believe to be part of the solution.

2. Image benchmarks

The PASCAL Visual Object Classes (VOC) challenge is a benchmark in visual object category recognition and detection, which provides challenging images and high quality annotation, together with a standard evaluation methodology. Measured the state-of-the-art on a yearly basis from 2005 to 2012. It has been succeeded by the ImageNet challenge which evaluates algorithms for object detection and image classification at large scale.

Slide credit: Mark Everingham

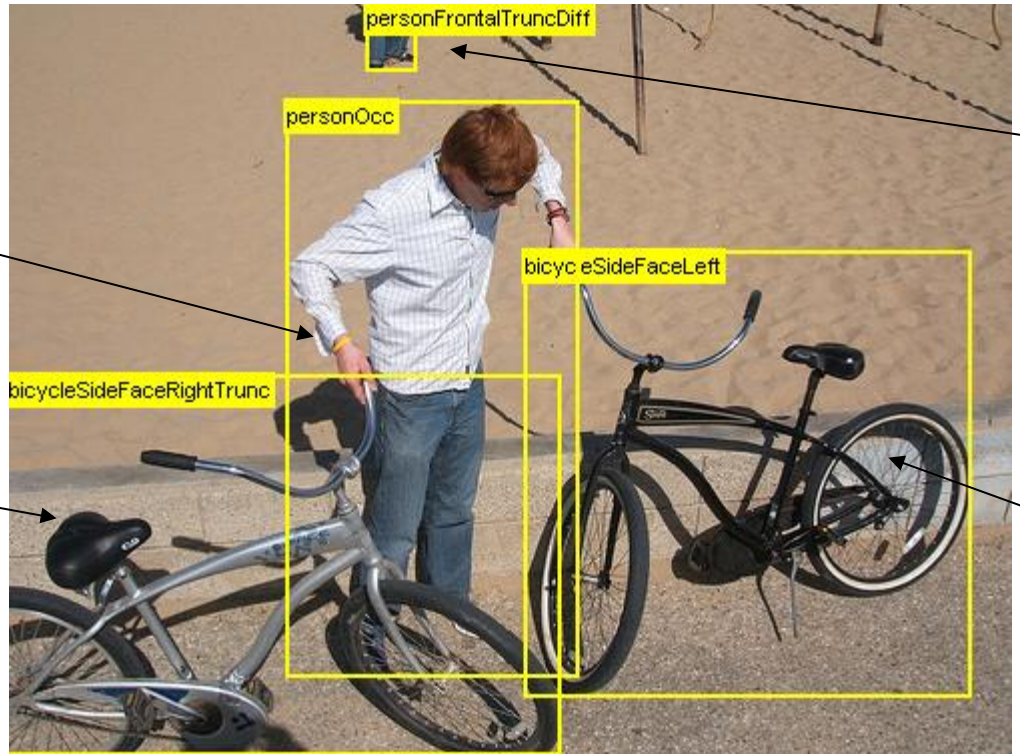
Dataset Collection

500K Images downloaded from **flickr** and random subset selected for annotation

Complete annotation of all objects from 20 categories

Occluded
Object is significantly occluded within BB

Truncated
Object extends beyond BB

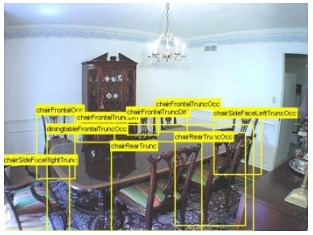


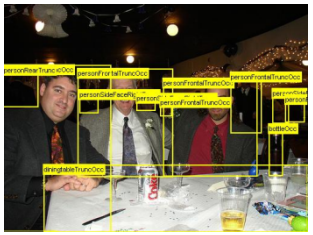
Difficult
Not scored in evaluation

Pose
Facing left

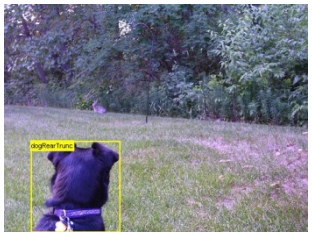
Examples

Dining Table

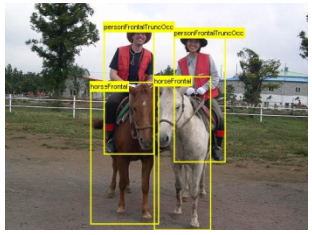




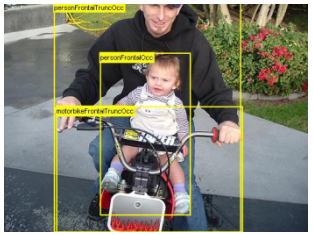
Dog



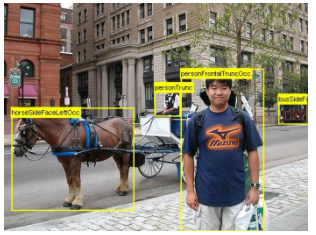
Horse

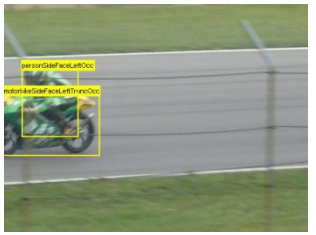


Motorbike

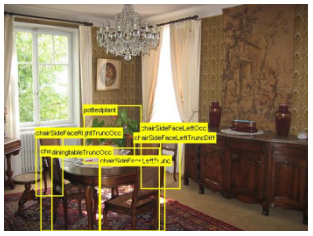


Person



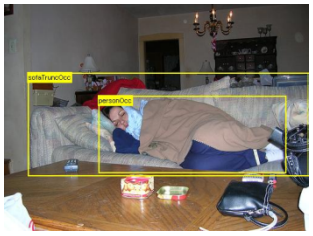


Potted Plant



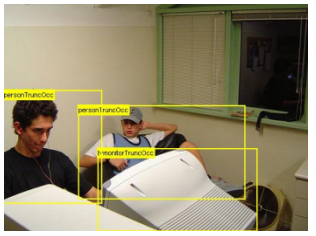
Sheep

Sofa



Train

TV/Monitor



2010 Dataset Statistics

	Training		Testing	
Images	10,103	(7,054)	9,637	(6,650)
Objects	23,374	(17,218)	22,992	(16,829)

VOC2009 counts shown in brackets

Minimum ~500 training objects per category

~1700 cars, 1500 dogs, 7000 people

~Equal distribution across training and test sets

PASCAL VOC Challenges

Object classification

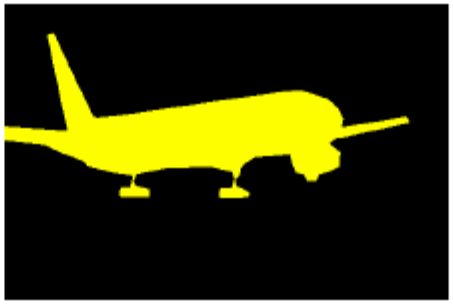
- Does the image contain an airplane?

Object deteciton

- Where is the airplane, (if any)?

Object segmentation

- Which pixels are part of an airplane, (if any)?



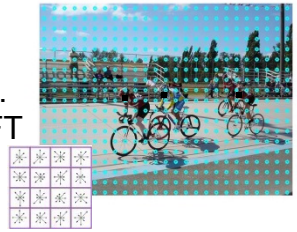
Slide credit: Andrew Zisserman

Lessons learned (day 1)

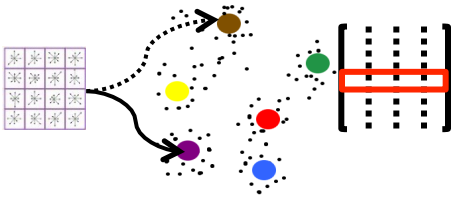
Reliable object classification with bag-of-words and SVMs

Local Feature Extraction → Feature Encoding → Feature Pooling → Classification

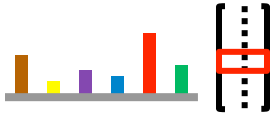
e.g. SIFT



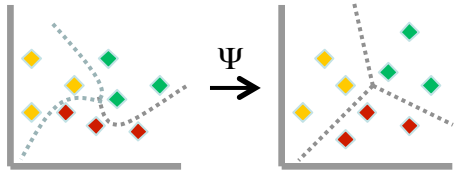
dense sampling



BoW
Sparse coding
Fisher
VLAD



avg/sum pooling
max pooling

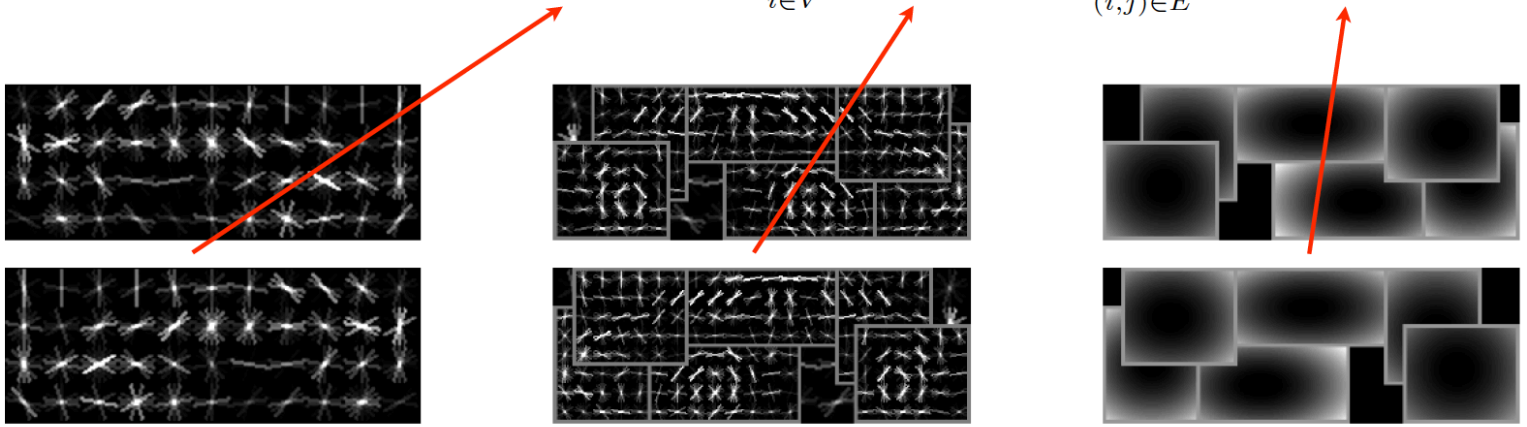


kernel approximation [*Maji, ICCV09*]
explicit feature maps [*Vedaldi, PAMI11*]

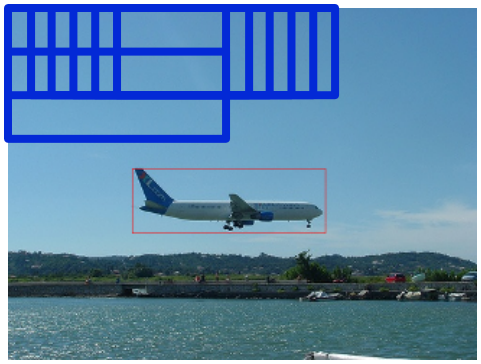
Lessons learned (day 2)

Model parts and local deformations with latent SVM

$$s(\mathbf{I}; x_0, y_0, \ldots, x_{|V|}, y_{|V|}) = \mathbf{w}_0^{\mathrm{T}} \phi(\mathbf{I}; x_0, y_0) + \sum_{i \in V} \mathbf{w}_i^{\mathrm{T}} \phi(\mathbf{I}; x_i, y_i) + \sum_{(i,j) \in E} d_{ij} \phi_d(x_i - x_j, y_i - y_j)$$



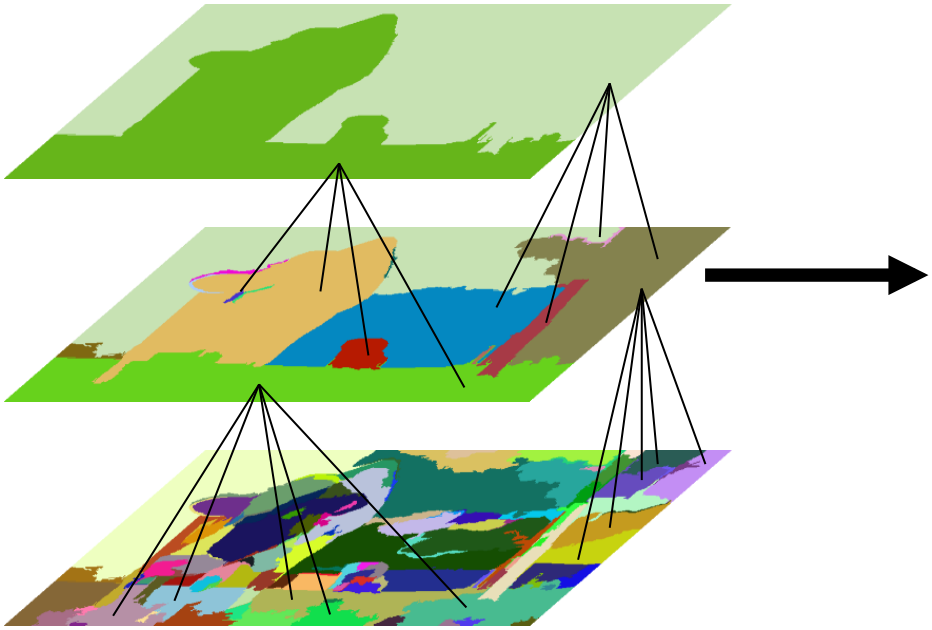
Global object model | Object part models | Deformation model

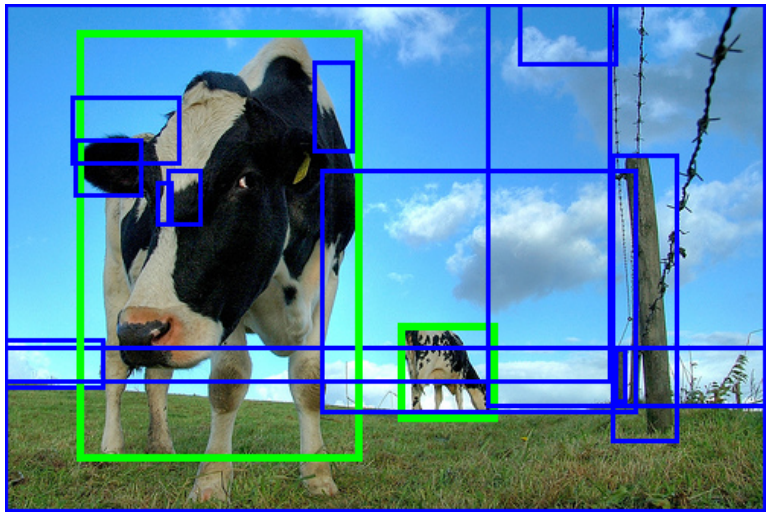


Felzenswalb, PAMI 2010

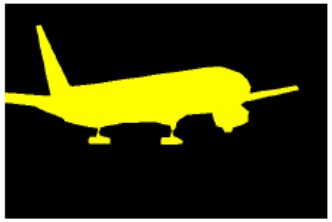
Lessons learned (day 3)

Hypotheses from hierarchical grouping, strong encodings



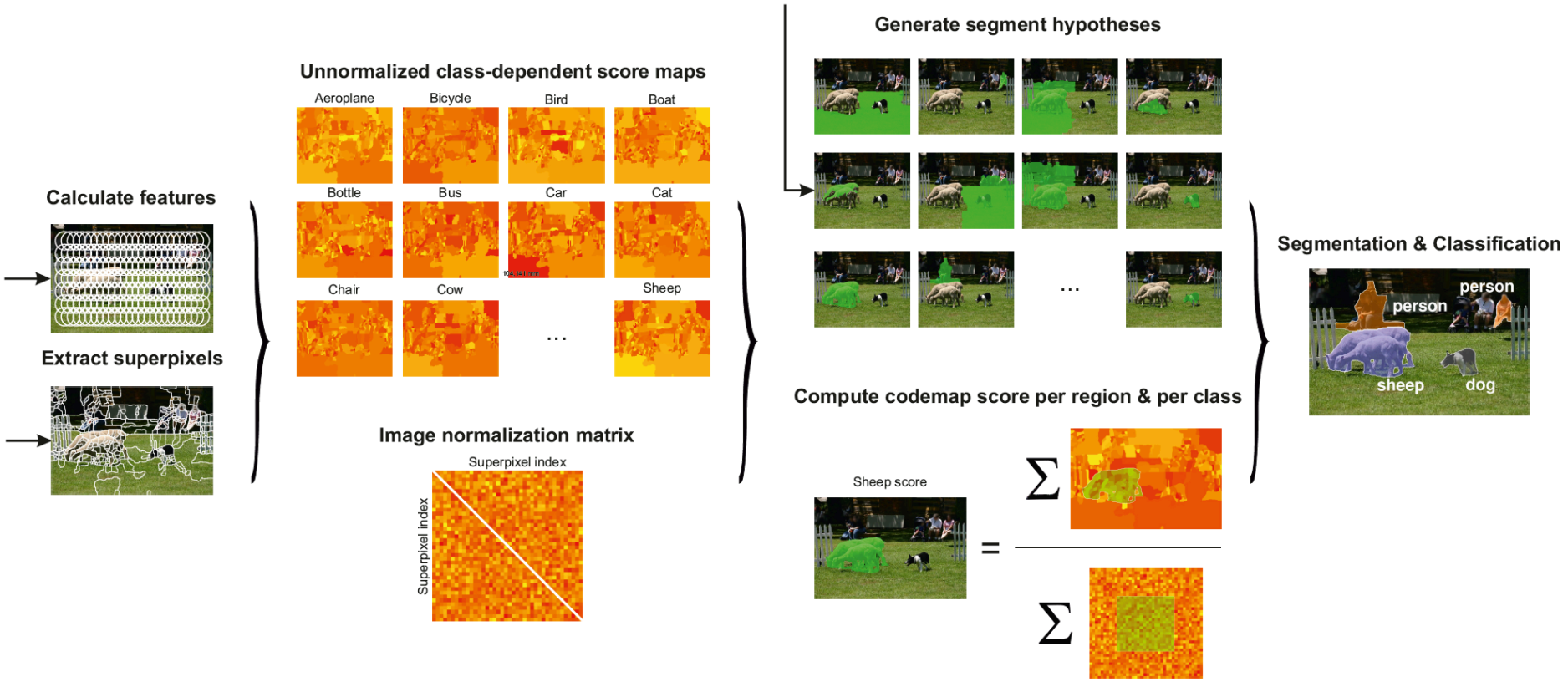


Uijlings, IJCV 2013



Lessons learned (day 1)

Codemaps for localized L2-normalized encoding



Gavves, PAMI submitted

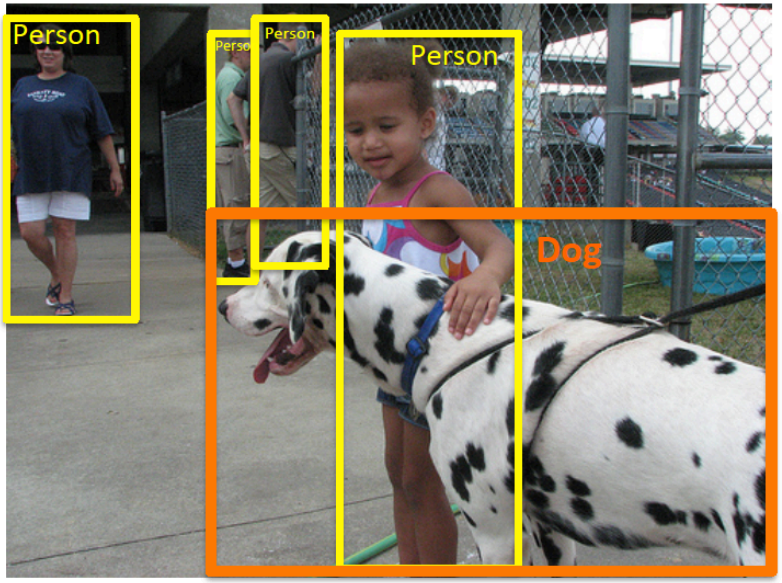
ImageNet large-scale challenge

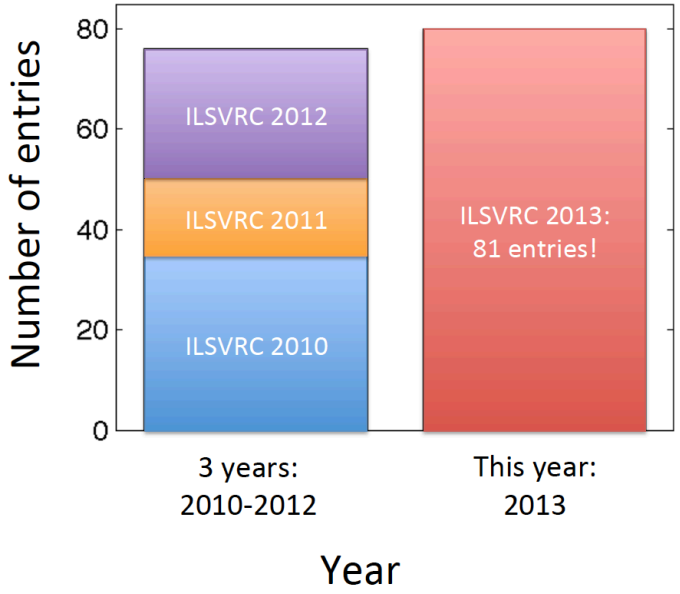
~~20 object classes~~ ~~22,591 images~~

NEW

200 object classes **456,191 images** **DET**

1000 object classes **1,431,167 images** **CLS-LOC**



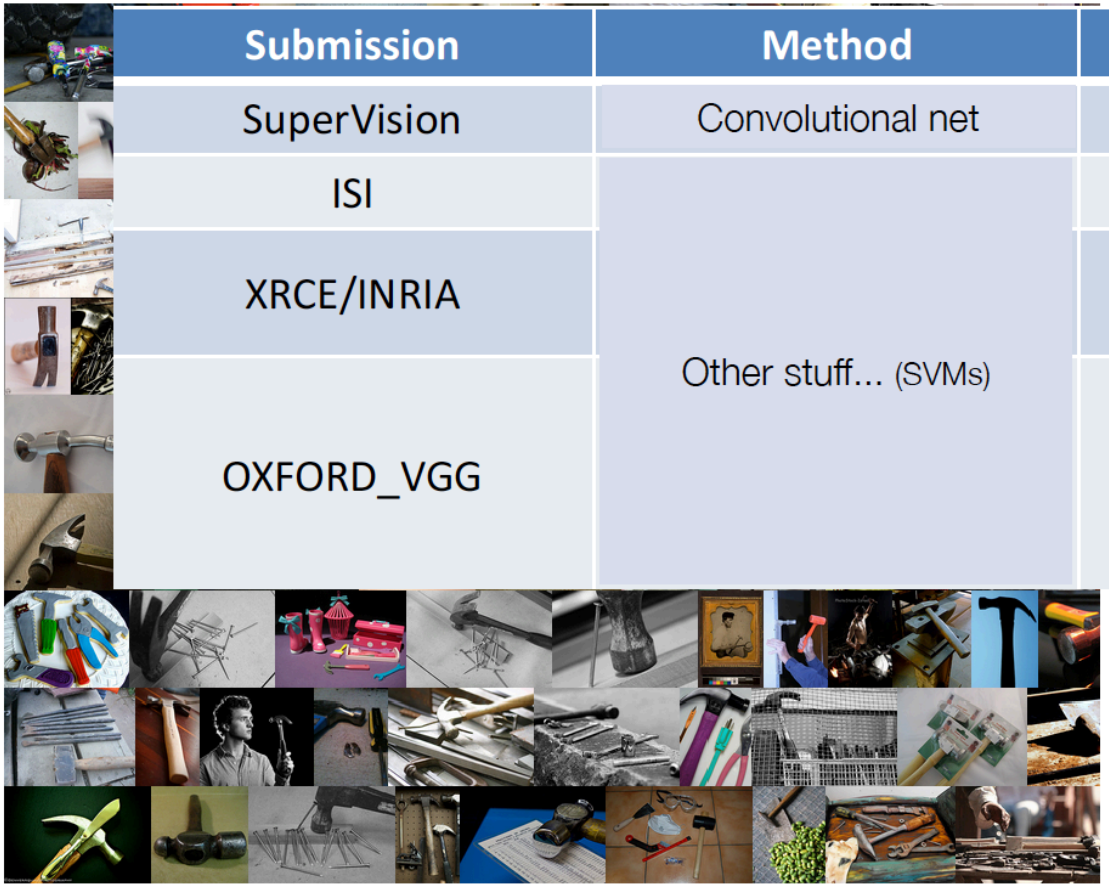


http://image-net.org/challenges/LSVRC/{2010,2011,2012,2

Slide credit: Olga Russakovsky

ImageNet 2012 classification

ImageNet Challenge 2012



Submission	Method	Error rate
SuperVision	Convolutional net	0.16422
ISI	Other stuff... (SVMs)	0.26172
XRCE/INRIA		0.27058
OXFORD_VGG		0.27302

9.8%

1.1%

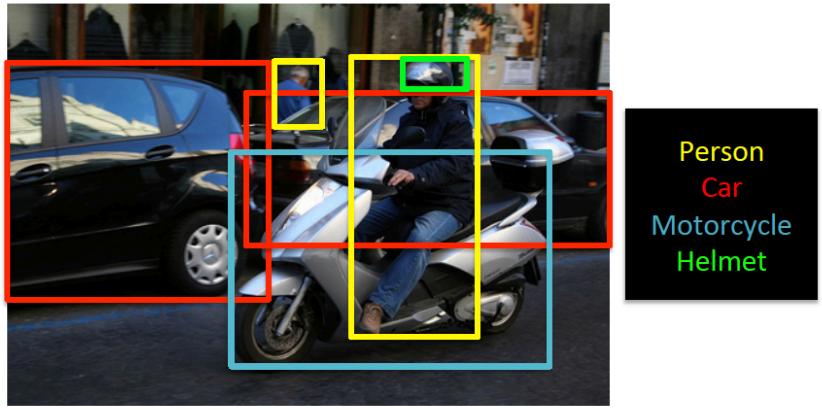
ImageNet 2013 classification

Top10 is using deep nets.

Team Name	Error
Clarifai (with outside data)	0.112
Clarifai	0.117
NUS	0.130
ZF	0.135
Andrew Howard	0.136
OverFeat – NYU	0.142
UvA-Euvision	0.143
Adobe	0.152
VGG	0.152
Cognitive Vision	0.161
Decaf	0.192
IBM Multimedia Team	0.207
… Deep Punx, Minerva-MSRA, MIL, Orange, BUPT-Orange, Trimps-Soushen1, Quantum Leap	

ImageNet detection challenge

Statistics		PASCAL VOC 2012		ILSVRC 2013
Object classes		20	10x	200
Training	Images	5.7K		395K
	Objects	13.6K	25x	345K
Validation	Images	5.8K		20.1K
	Objects	13.8K	4x	55.5K
Testing	Images	11.0K		40.1K
	Objects	---		---



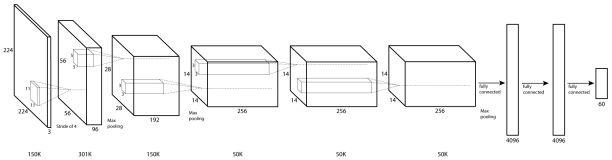
ImageNet 2013 detection results

Team Name	mAP	# categories won
UvA-Euvision	0.226	130
NEC-MU (with outside data)	0.209	---
NEC-MU	0.196	25+35 (2 entries)
OverFeat-NYU (with outside data)	0.194	---
Toronto A	0.115	6+1 (2 entries)
SYSU_Vision	0.105	3
GPU_UCLA	0.098	0
Delta	0.061	0
UIUC-IFP	0.010	0

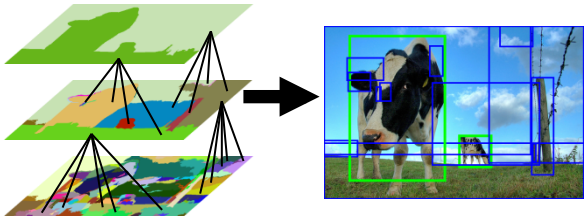
[Sande, Fontijne et al. ImageNet 2013]

Winning approach

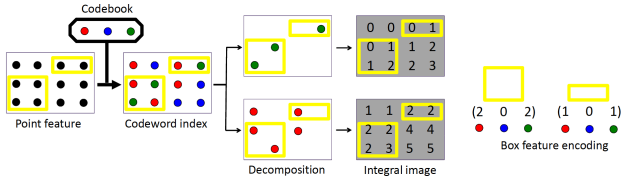
Classification priors



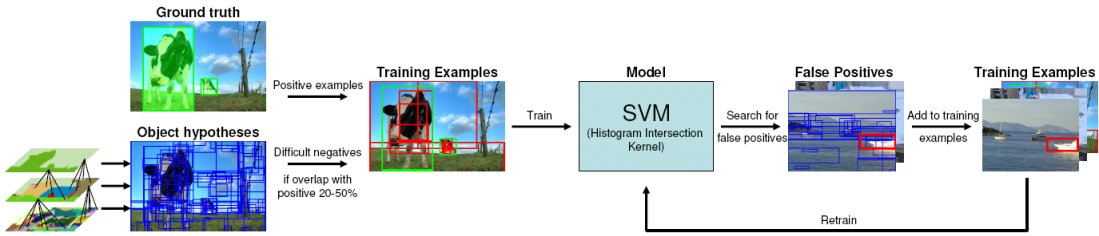
Selective search



Fisher vector with FLAIR



Retraining



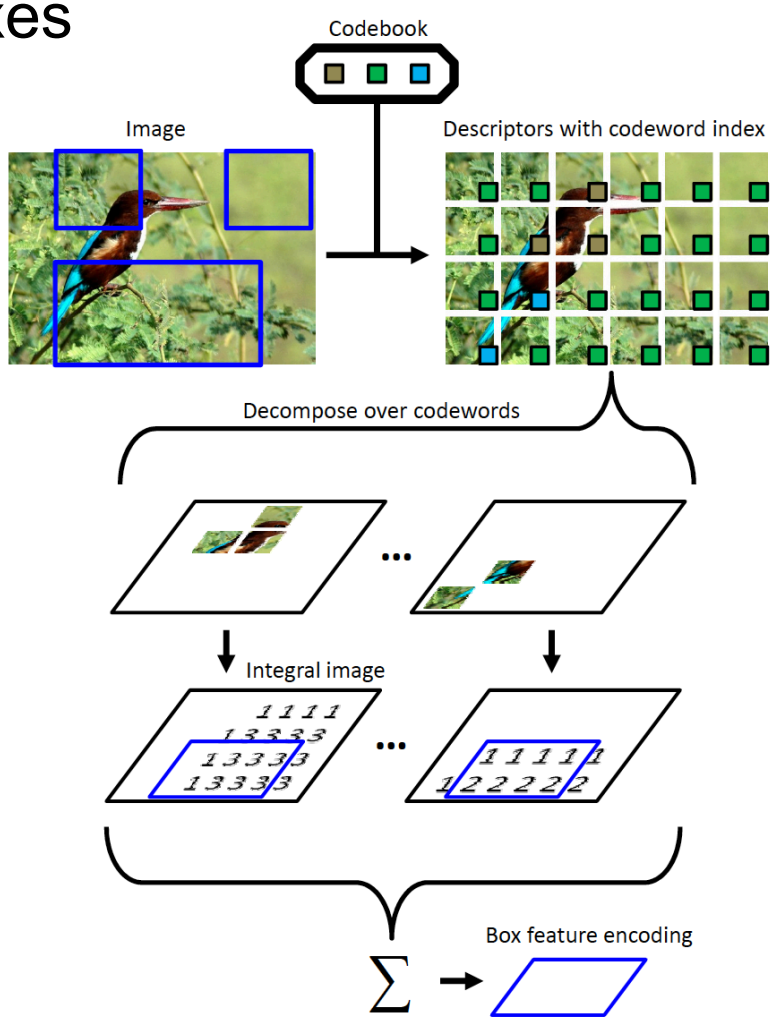
Fisher vector with FLAIR

FLAIR is a data structure for which it is as efficient to evaluate one box as it is many boxes

Decomposes Fisher vector per codeword into integral image

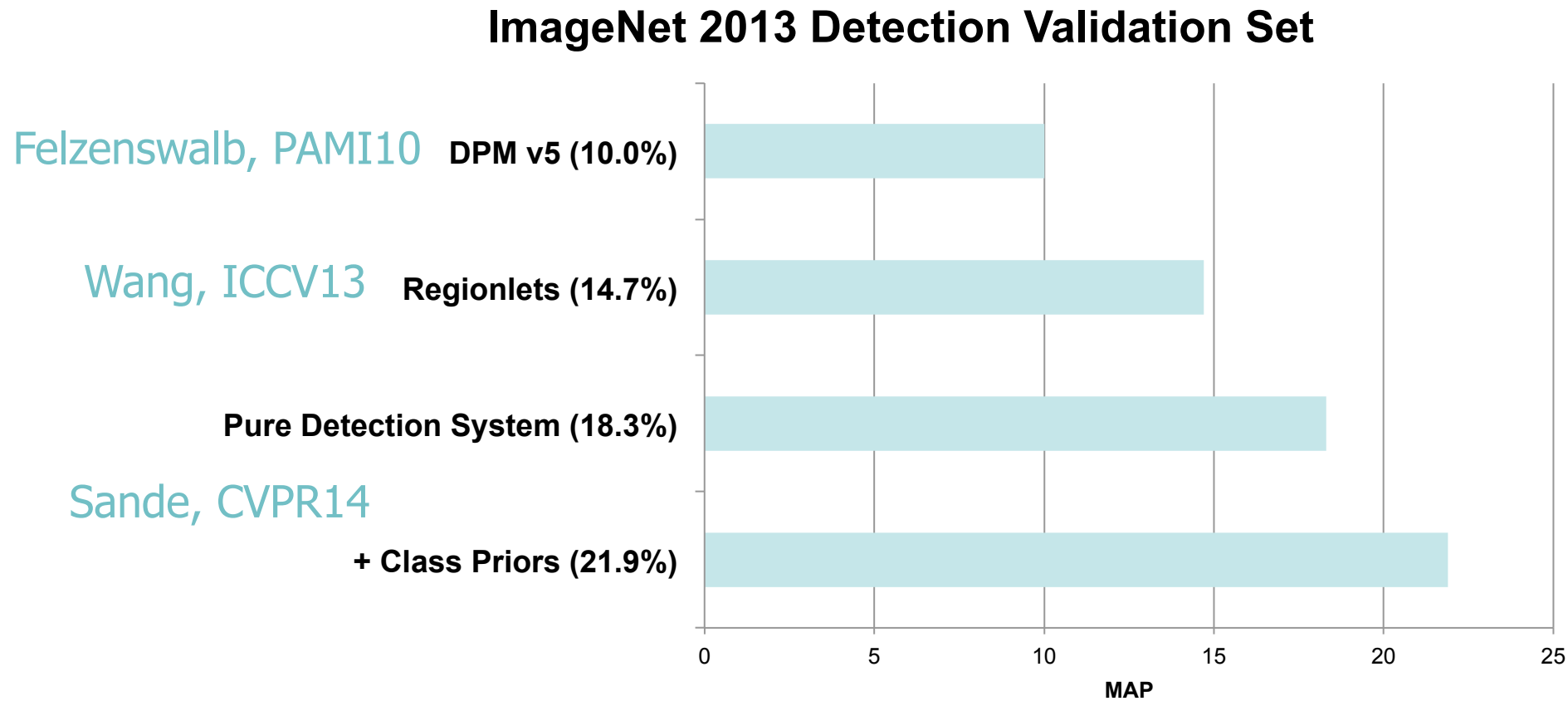
Maintains L2 and power-norm

18X speedup, same accuracy

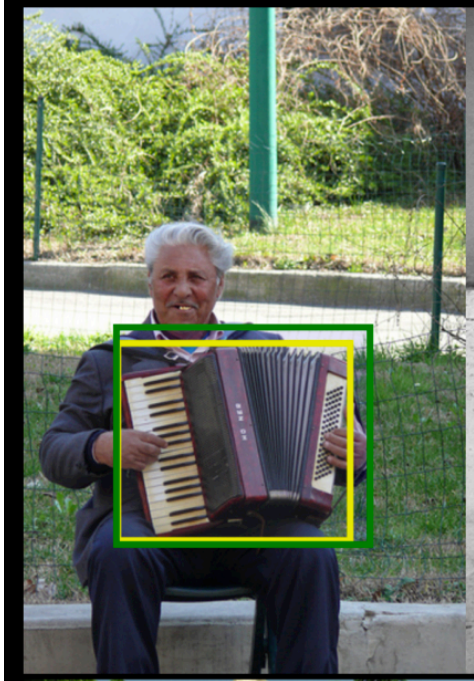
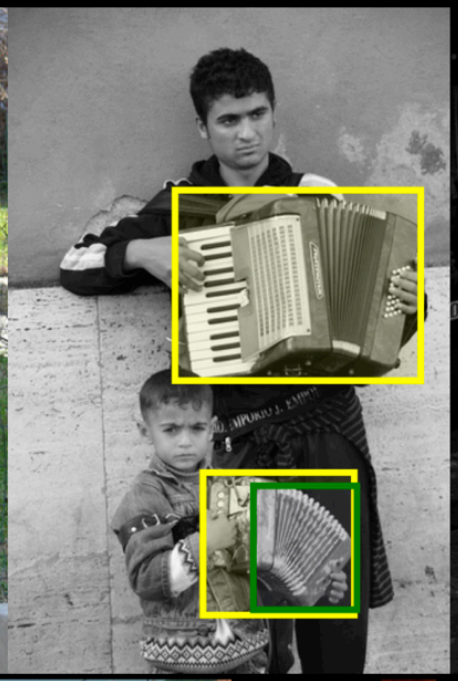


Sande CVPR 2014

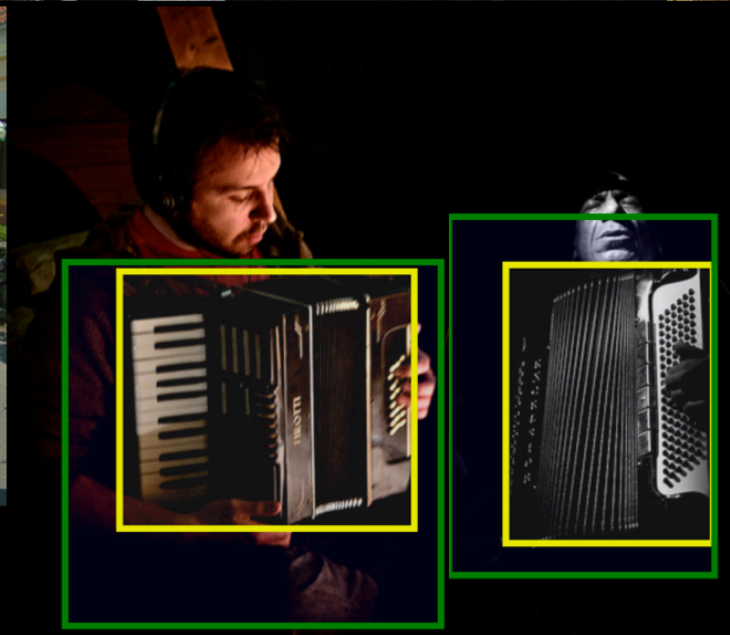
Detection results

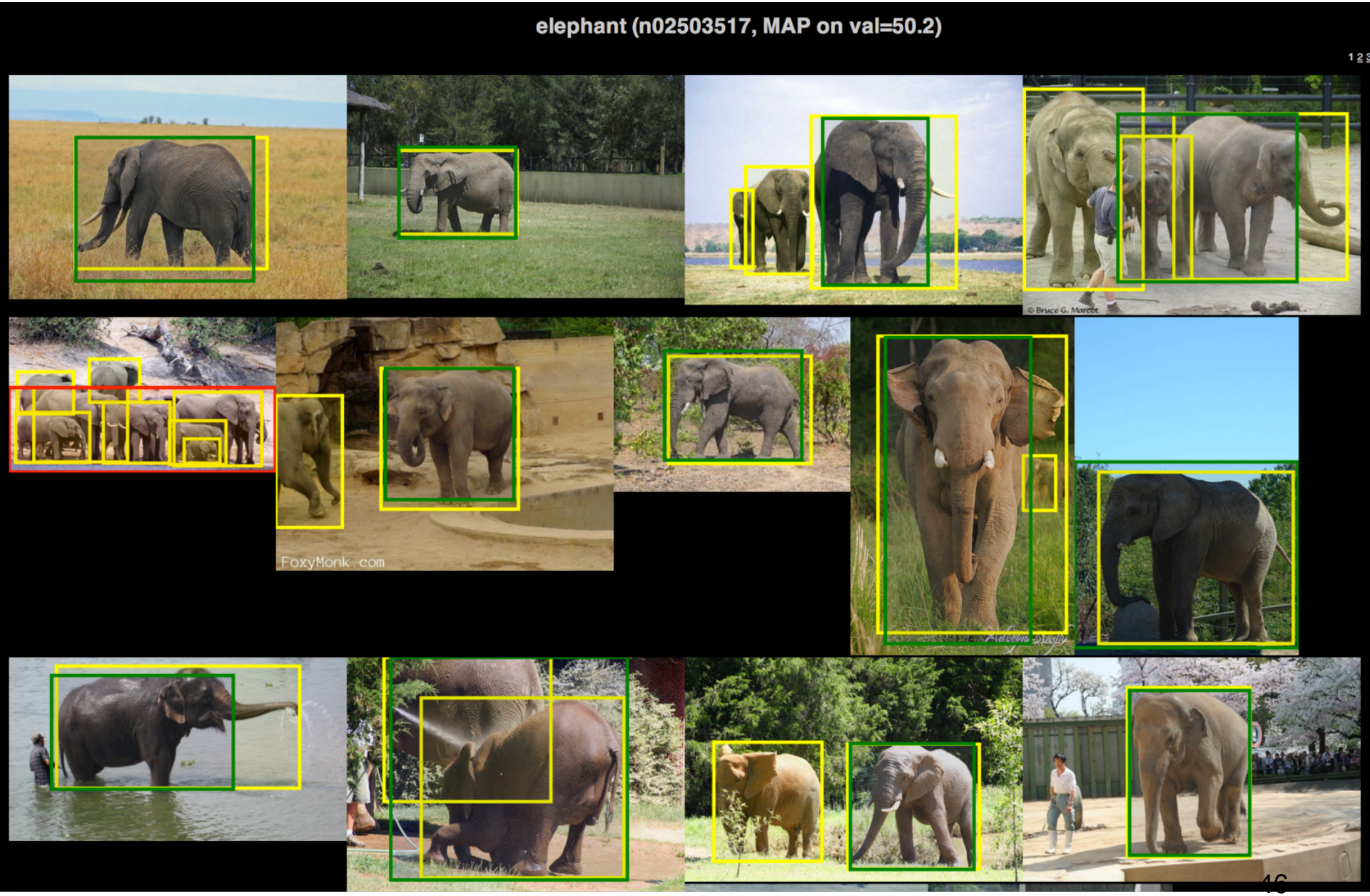


accordion (n02672831, MAP on val=40.9)

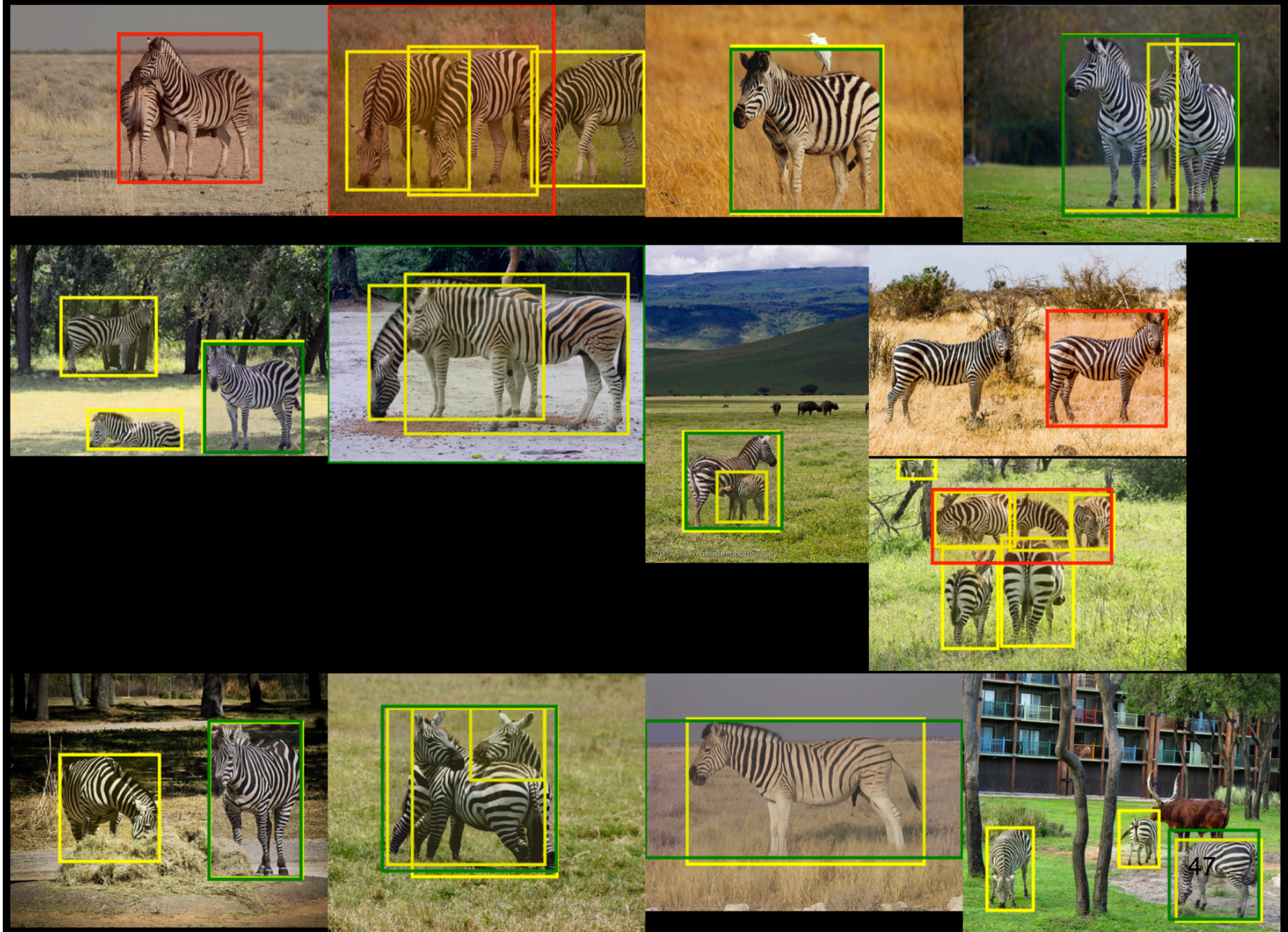







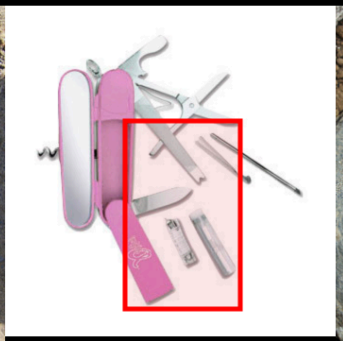

elephant (n02503517, MAP on val=50.2)
© Bruce G. Marcot
FoxyMonk.com

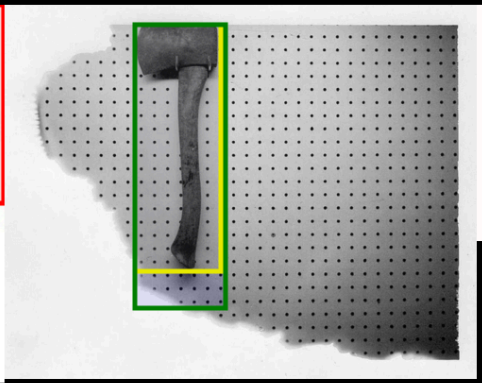
zebra (n02391049, MAP on val=44.0)



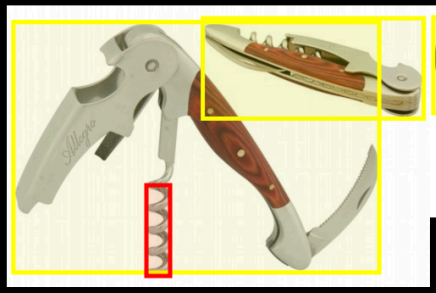
axe (n02764044, MAP on val=1.1)

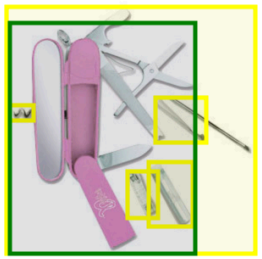





corkscrew (n03109150, MAP on val=5.6)



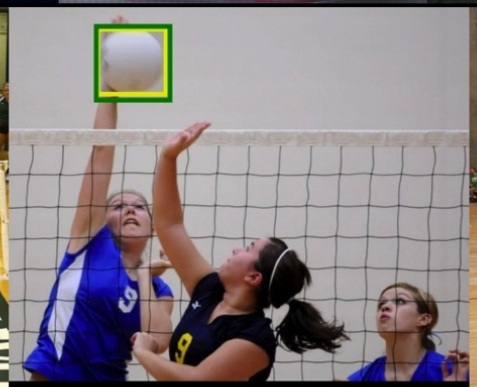

volleyball (n04540053, MAP on val=28.3)









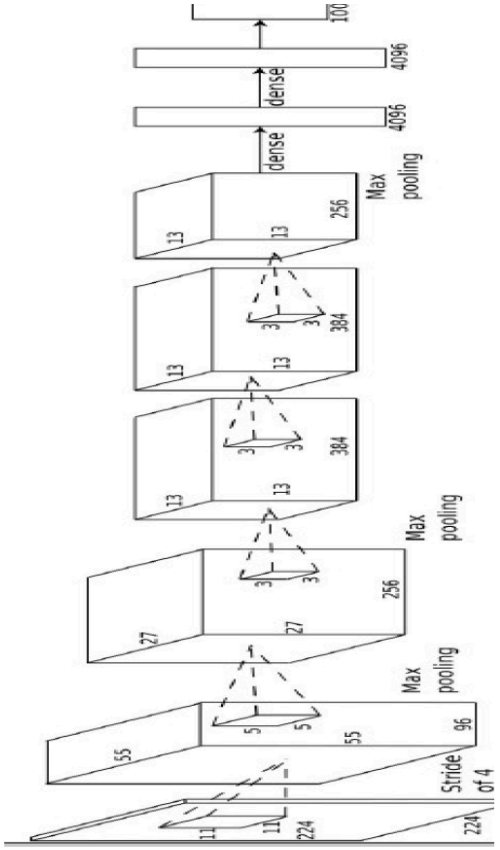






Quiz: how many parameters?

How many parameters to learn in a state-of-the-art seven layer deep convolutional neural network?



Quiz: how many concepts?

How many object, scene, and action detectors do we need for effective visual retrieval?

Counting dictionary words

Slide credit: Li Fei-Fei

Biederman, Psychological Review 1987

3. Labels from humans

The most precious resource in computer vision by learning is data.

The most traditional source for obtaining labeled examples is to rely on human experts. The Internet has launched the trend to let volunteers label visual content, either for fun, for winning a game or for a small compensation. ImageNet is a labeled image database organized according to the WordNet hierarchy in which each node of the hierarchy is depicted by hundreds of images.

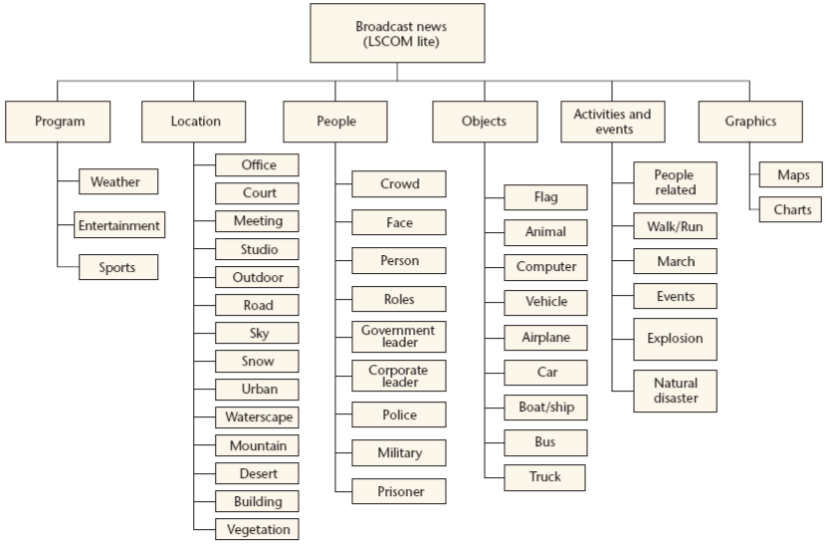
Labeling by library experts

LSCOM (Large Scale Concept Ontology for Multimedia)

Provides manual annotations for 449 concepts

- In international broadcast TV news

Connection to Cyc ontology



http://www.lscom.org/

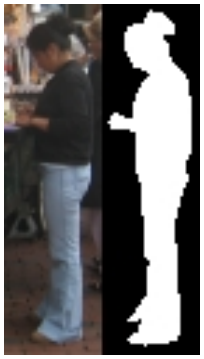
Labeling by volunteers

Russell IJCV 2008

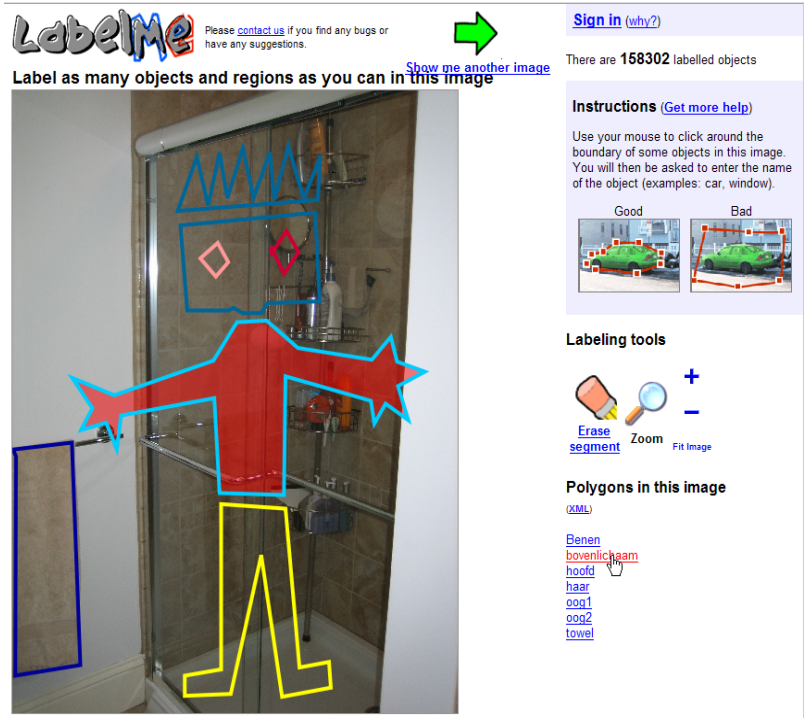
Polygon quality

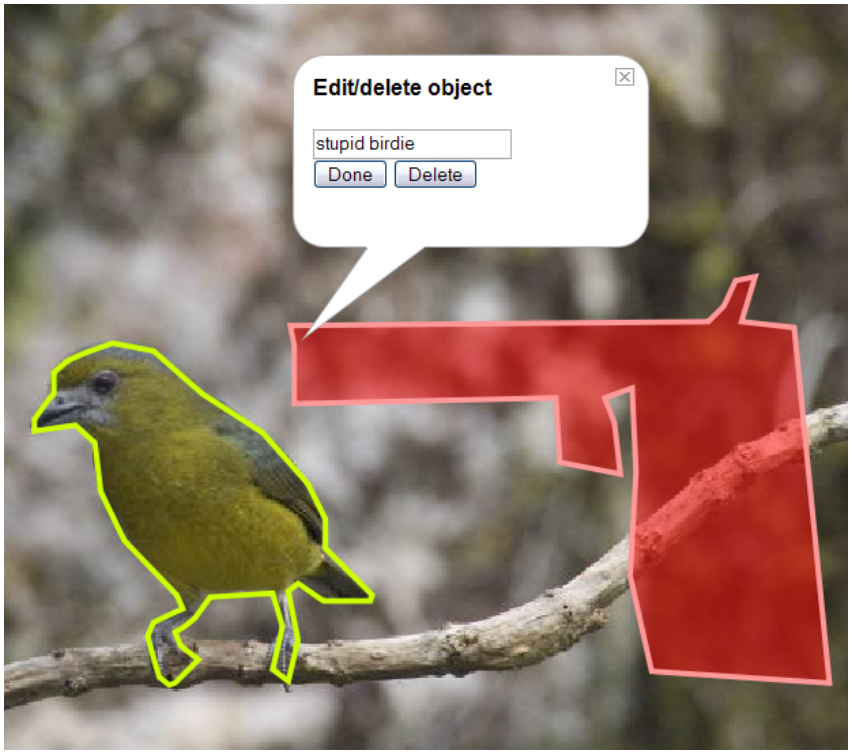






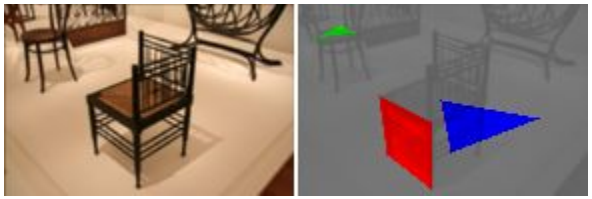

Online hooligans





Testing










Most common labels:

test

adksdsa

woiieiie

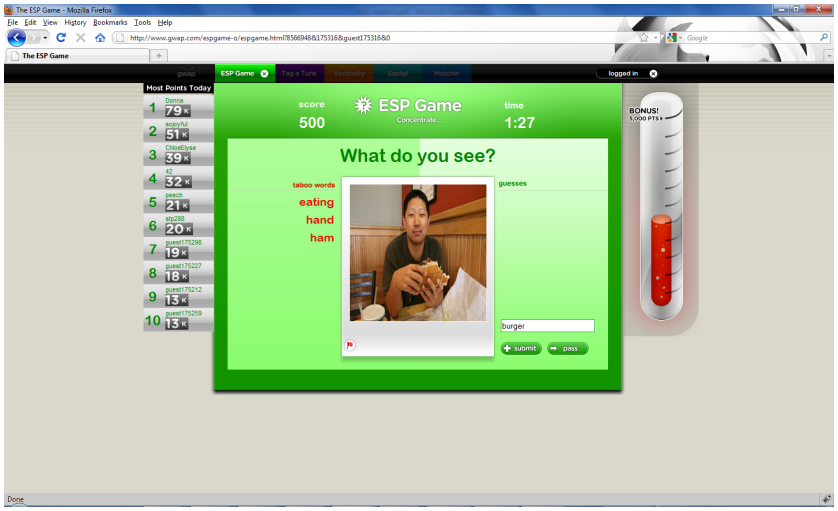
Downside of volunteers

Lack of incentive

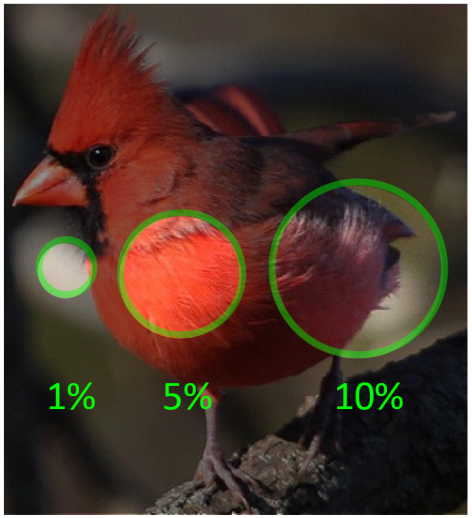
Limited quality control

Limited number of labels

Labels from games

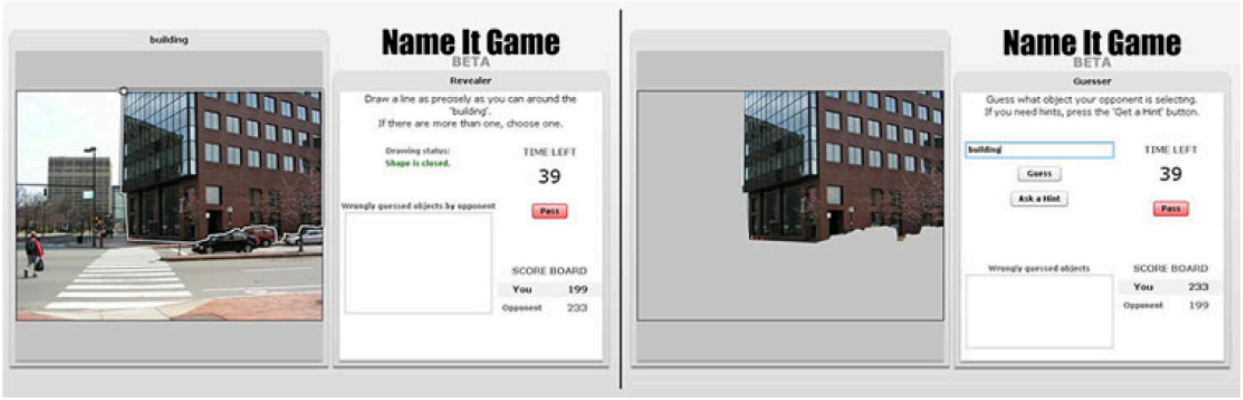


von Ahn, ESP Game



Bubble sizes as proportions of image

Deng CVPR 2013



Steggink, MM Sys, 2011

Labels from games

Games are a fun way to motivate volunteers

- Words are often too abstract
- Requires some sort of label validation

More descriptive labels by

- Adding semantic structure
- Linking labels to regions

Any game suffers from lack of popularity

Labels from micro-payments

ImageNet (11M images)

- 4000 categories
- > 100 examples

SUN (130K images)

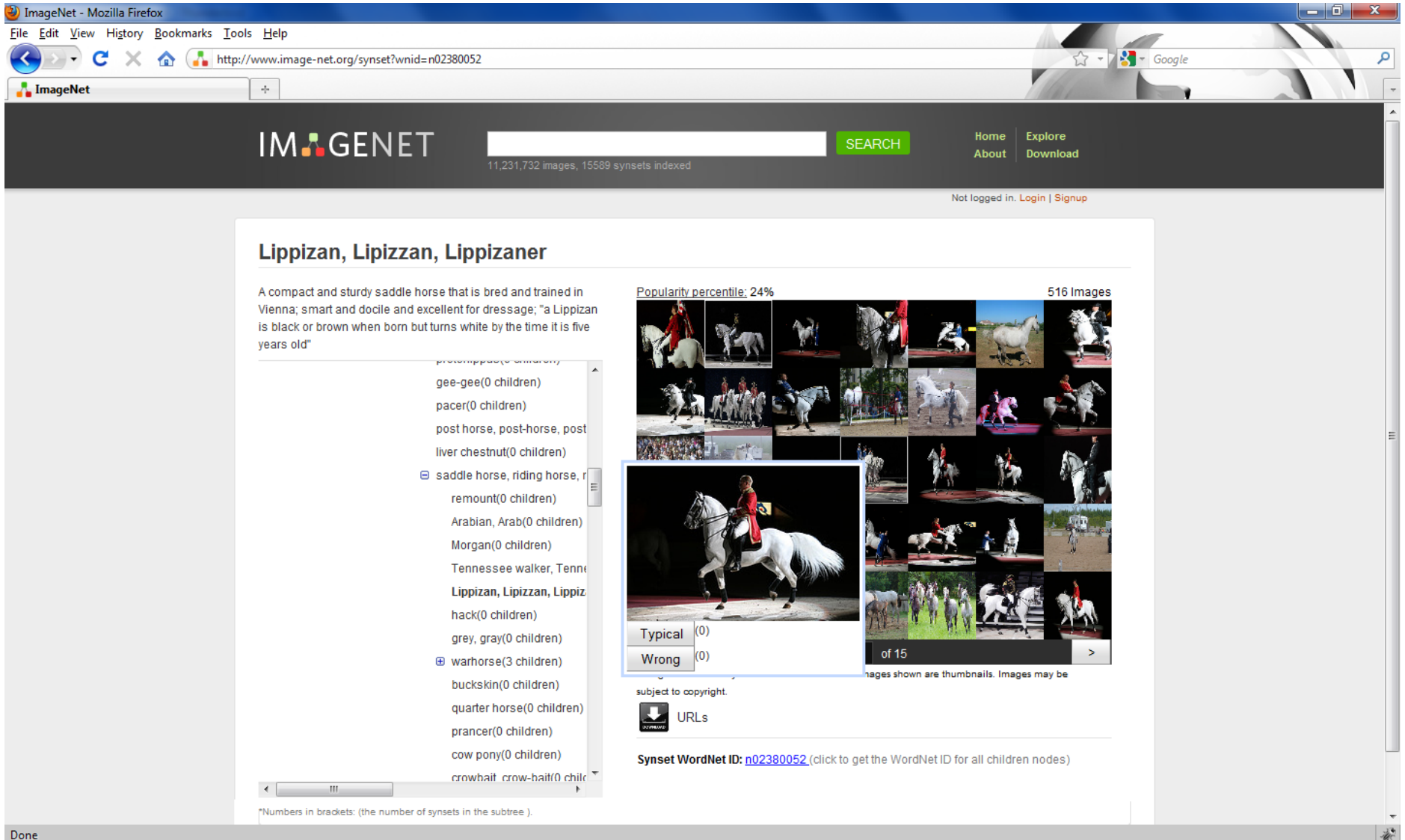
- 397 scene categories
- > 100 examples

Deng et al, CVPR 2009

Xiao et al, CVPR 2010

http://www.image-net.org

IMAGENET demo



Constructing ImageNet

Step 1:
Collect candidate images via the Internet

Step 2:
Clean up the candidate Images by humans

amazon mechanical turk
beta
Artificial Artificial Intelligence

IMAGENET is built by crowdsourcing

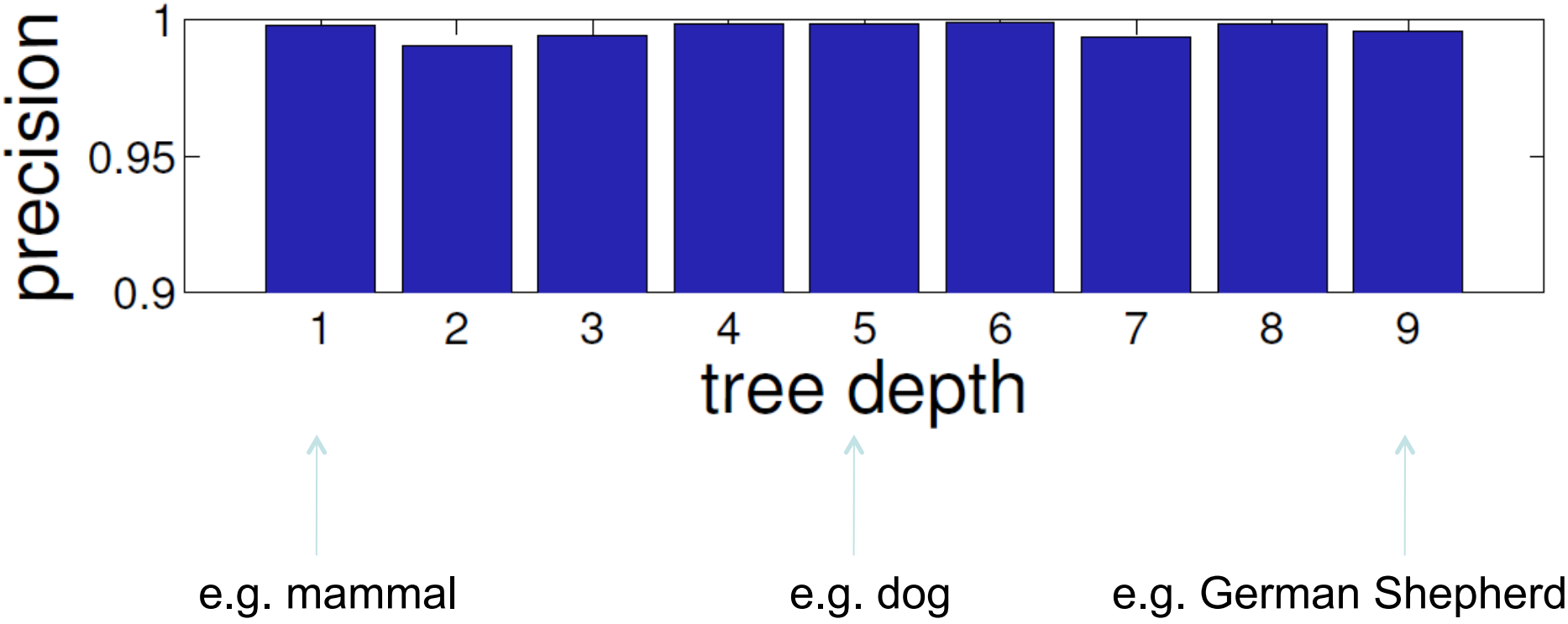
July 2008: 0 images

Dec 2008: 3 million images, 6K+ synsets

April 2010: 11 million images, 15K+ synsets

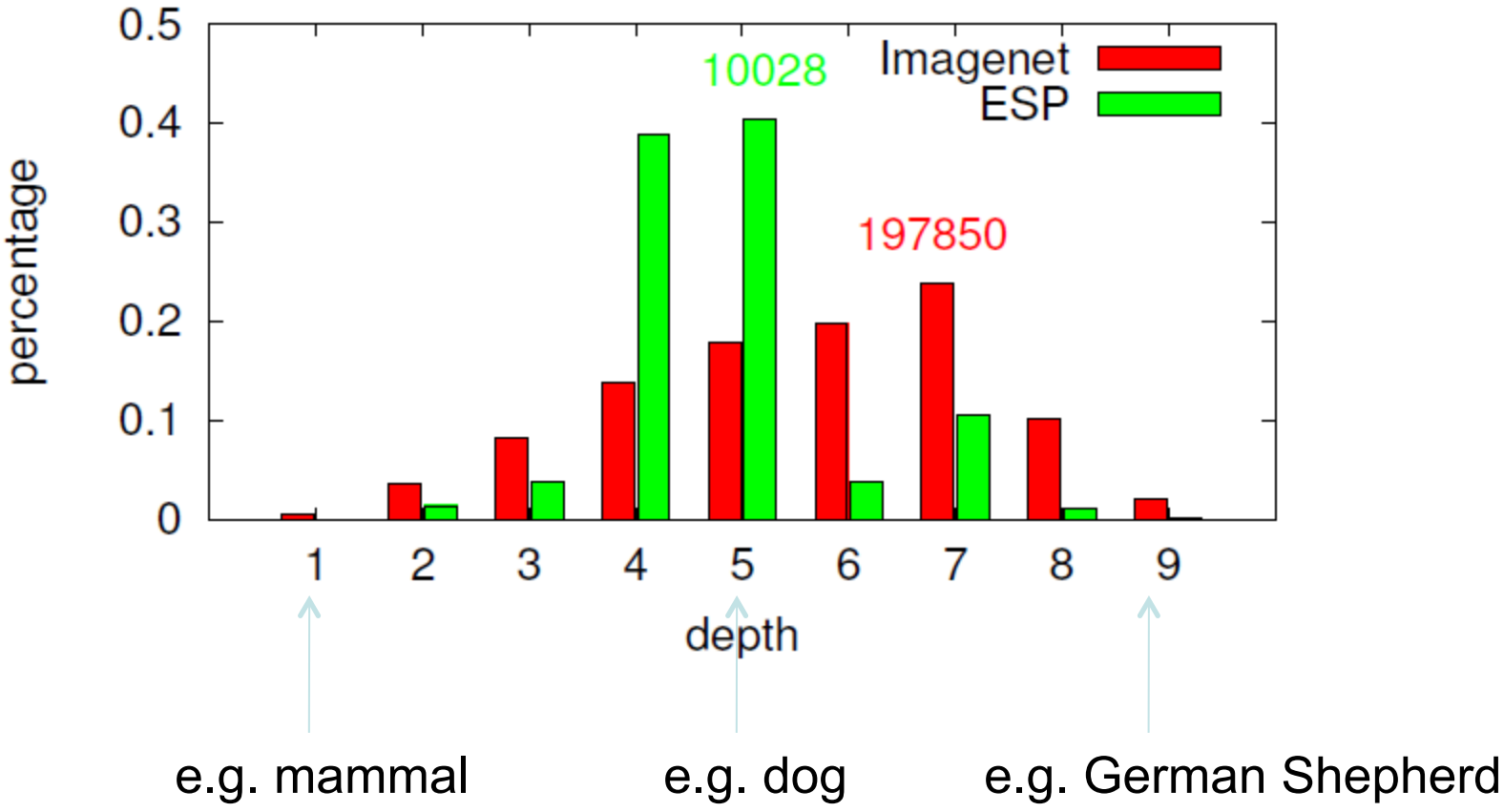
Yesterday: 14 million images, 21K synsets indexed

Accuracy



Deng CVPR, 2009

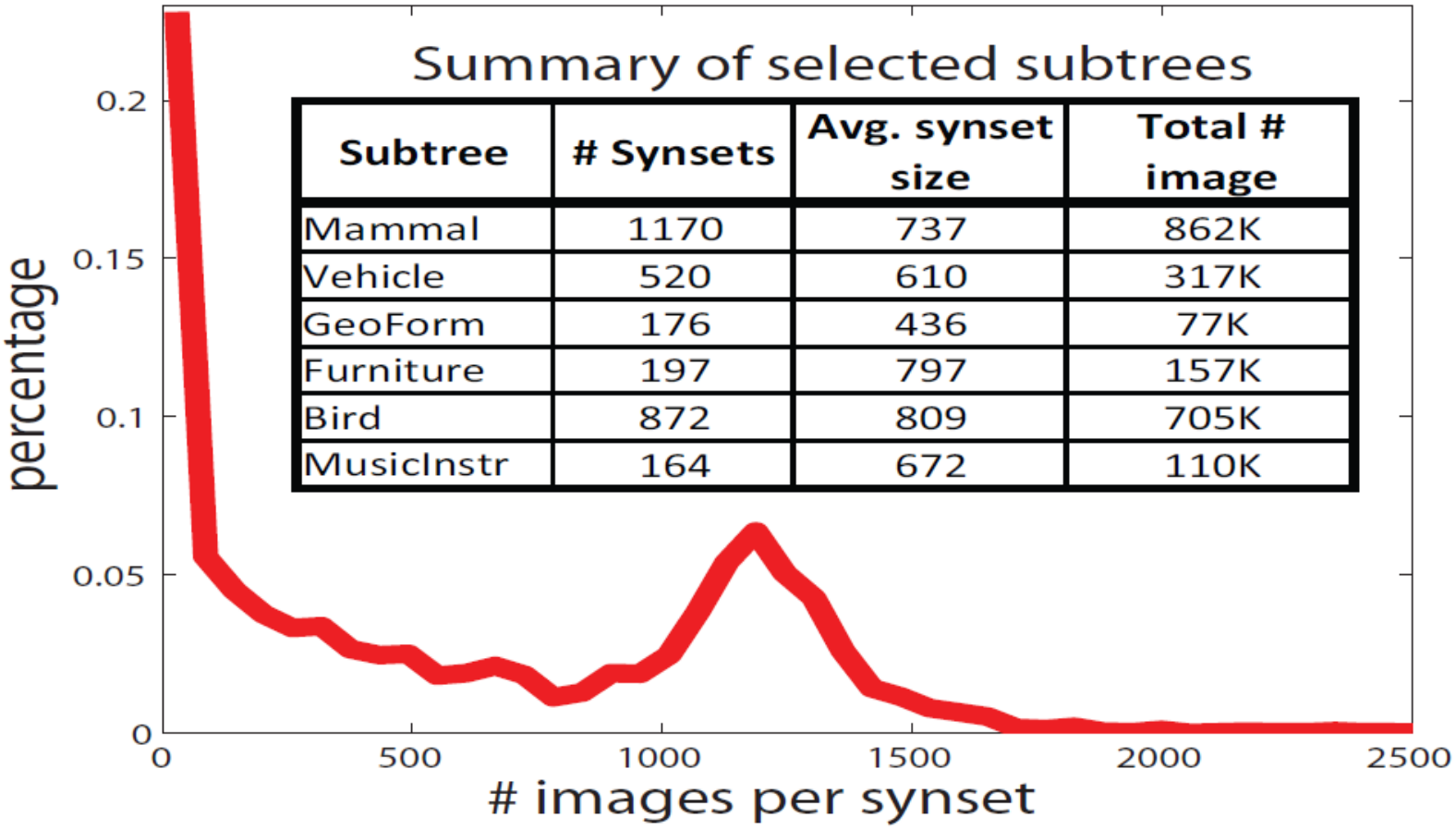
Diversity



ESP: Ahn et al. 2006

Deng CVPR, 2009

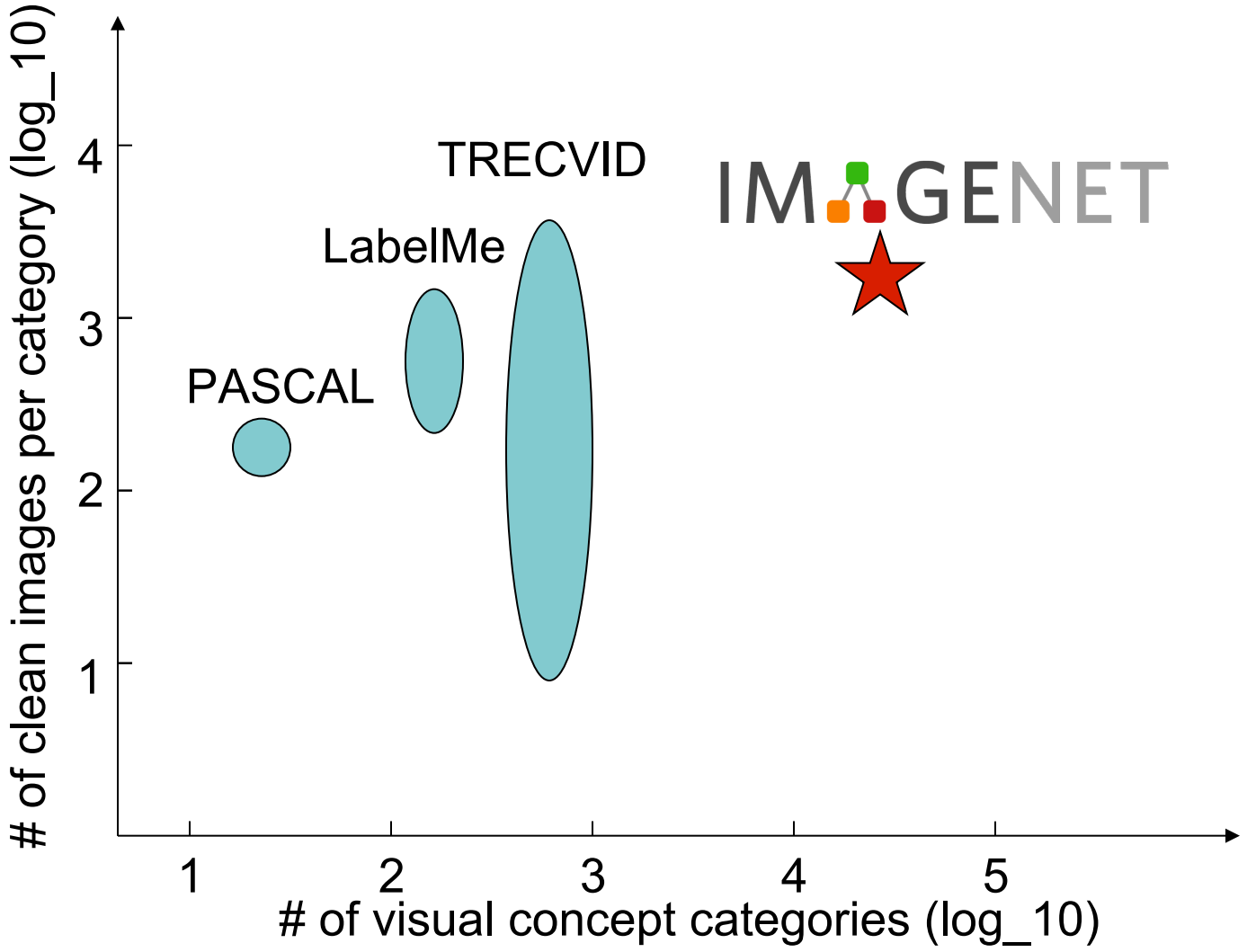
Scale



Subtree	# Synsets	Avg. synset size	Total # image
Mammal	1170	737	862K
Vehicle	520	610	317K
GeoForm	176	436	77K
Furniture	197	797	157K
Bird	872	809	705K
MusicInstr	164	672	110K

Deng CVPR, 2009

Datasets comparison



Constructing ImageNet

Step 1:
Collect candidate images via the Internet

Step 2:
Clean up the candidate Images by humans

amazon mechanical turk
beta
Artificial Artificial Intelligence

Constructing ImageNet

Free

Step 2:
Clean up the candidate Images by humans

amazonmechanical turk
beta
Artificial Artificial Intelligence

Constructing ImageNet

Free

$$$

YAHOO!

picsearch™

flickr

Live Search

Google Image Search

amazon mechanical turk
beta
Artificial Artificial Intelligence

4. Labels from similarities

The most precious resource in computer vision by learning is data.

Huge amounts of weakly labeled images and videos are available online. How reliable are these tags? Can we use them for learning classifiers, segment images, or localize distinctive parts? It turns out that ‘good old’ nearest neighbor with simple visual features provides a free, scalable and effective means to collect valuable data.

Many slides by Xirong Li


Google
Spotify
YouTube
DOMAINS
facebook
41
THOUSAND
posts every
second
72
HOURS
of Video
uploaded
2
MILLION
Searches
14
New Songs
Added
70
New
Registered
15
THOUSAND
Tracks
downloaded
from iTunes
skype
1.8
MILLION
likes
1.4
MILLION
Minutes Connecting
with Each other
350GB
of data
WordPress
347
New Blog
Posts
flickr
20
MILLION
Photo
Views
571
New Created
WEBSITES
278
THOUSAND
Tweets
twitter
Transactions
17
THOUSAND
Walmart
104
THOUSAND
Photos
Shared
New Tumblr
Photos
20
THOUSAND
snapchat
Amazon Sales
$83,000
Professional
searches
11
THOUSAND
Active
Users
11
THOUSAND
emails sent
204
MILLION
tumblr.
photos every
second
3,600
amazon.com
Linked in
Pinterest
Instagram
EMAIL

Fundamental problem

Social tags for image and video were never meant to meet professional standards, consequently they are

- subjective
- ambiguous,
- overly personalized, and
- limited.

Tagged images are notoriously difficult to find.

Searching for ‘tiger’

beach
car
lotus
caterham
...

view details

washington
animals
2005
z o
.

view details

december
2005
cat
tiger
...

view details

cute
cat
yilan
tiger

view details

cute
cat
yilan
iger

view details

cute
cat
yilan
iger

view details

Searching for ‘classroom’

kindergarten classroom layout of ...

view details

stone age stone age ...

view details

365days me of me ...

view details

cfc minnesotawoods minnesota forest ...

view details

12 favorite classroom incubator ...

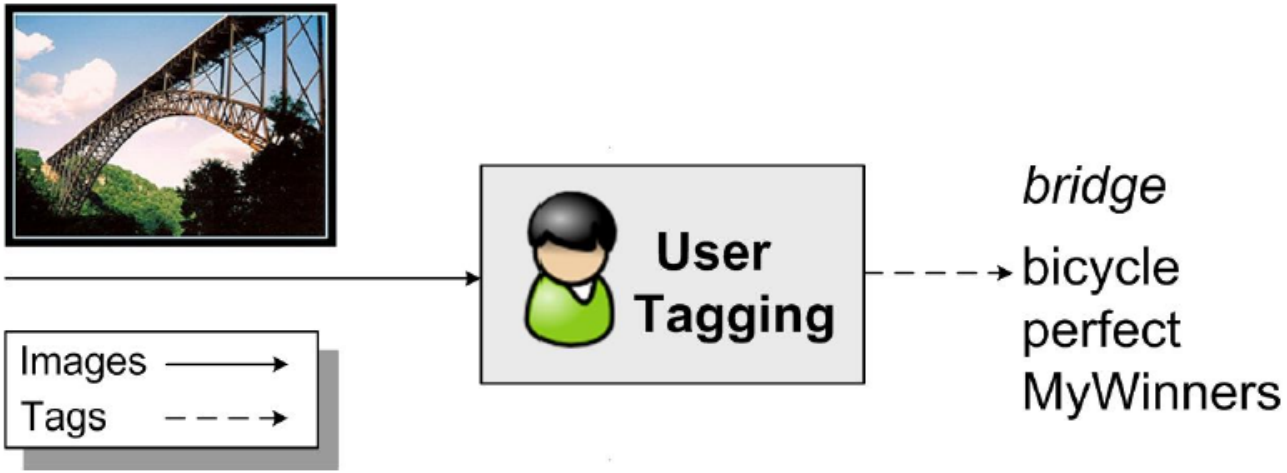
view details

tour tampere church unload ...

view details

Quiz

What image tags in this example are suited as training label?



Computer vision is essential

Free text

User tags

bridge
bicycle
perfect
MyWinners

?

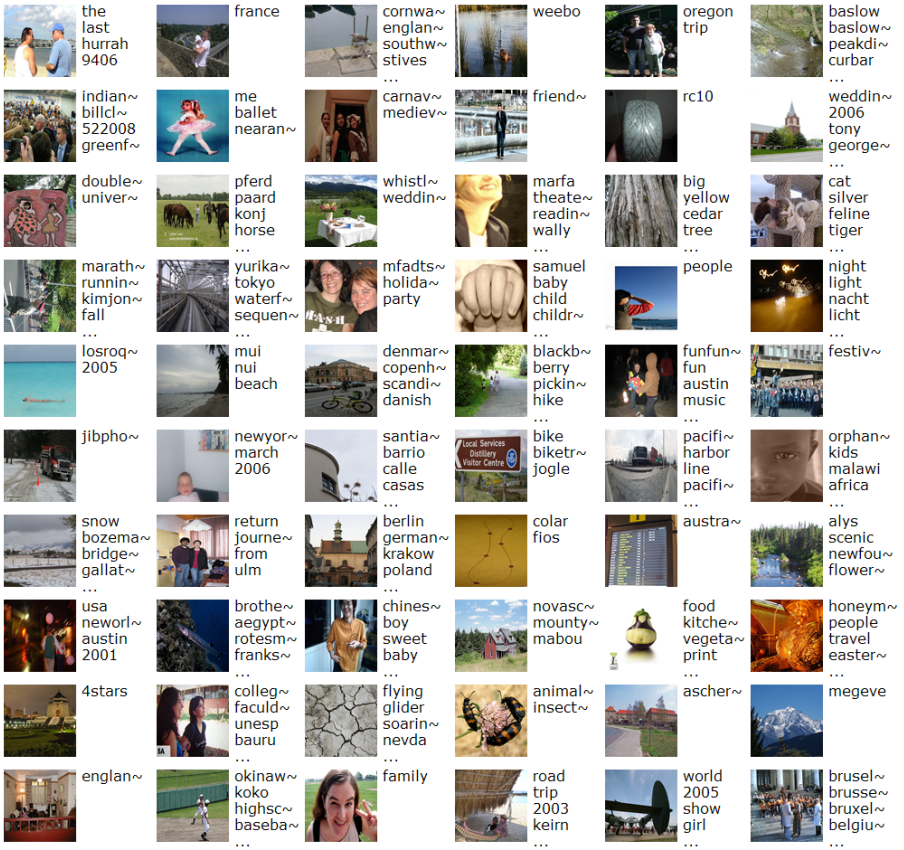
bridge
bicycle
perfect
MyWinners

Challenges

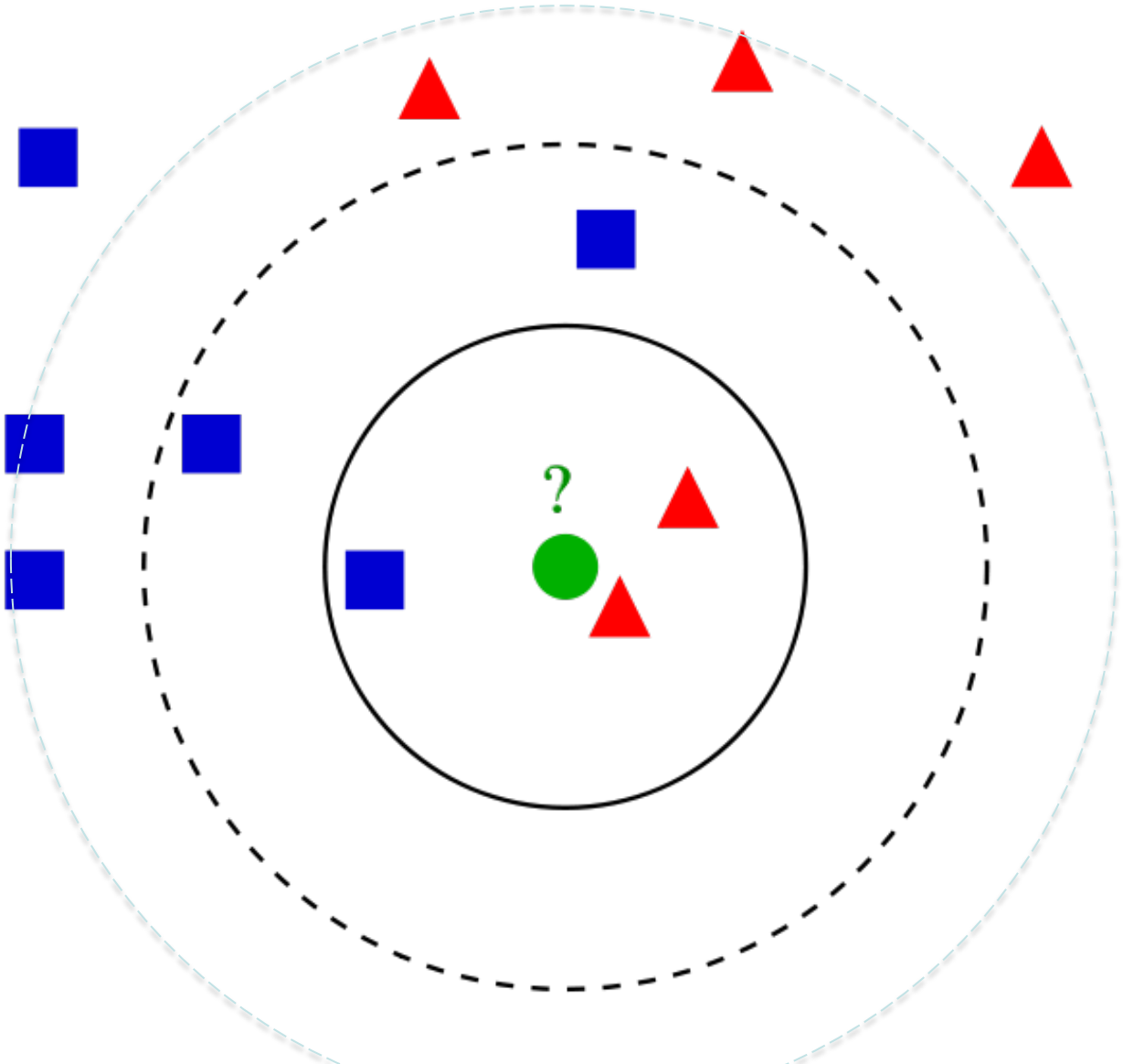
Many tags & many images

A prospective algorithm

- scalable
- unsupervised

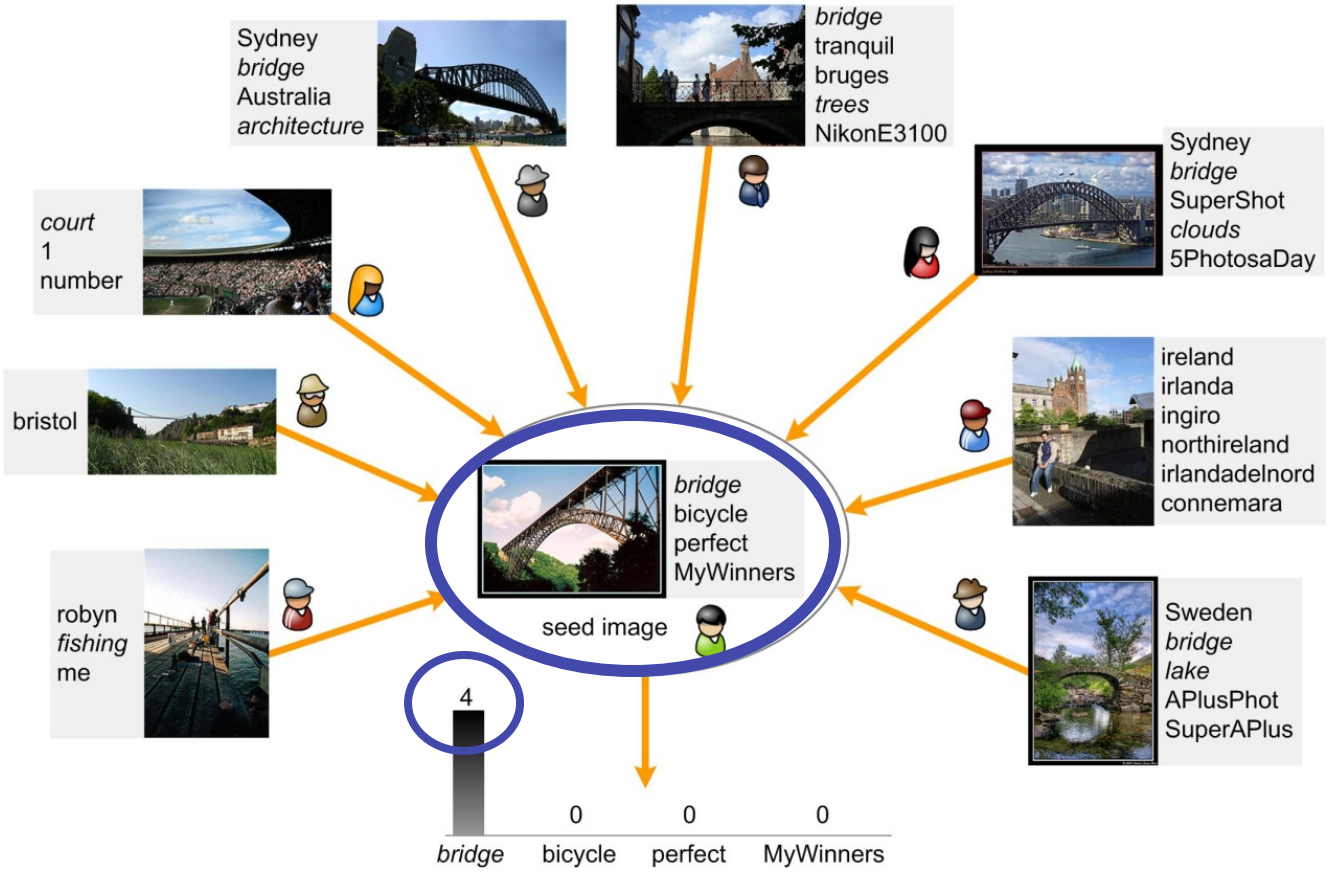


Nearest neighbor



Intuition for tagged images

Similar images with similar tags are reliable



Xirong Li, TMM 2009, best paper

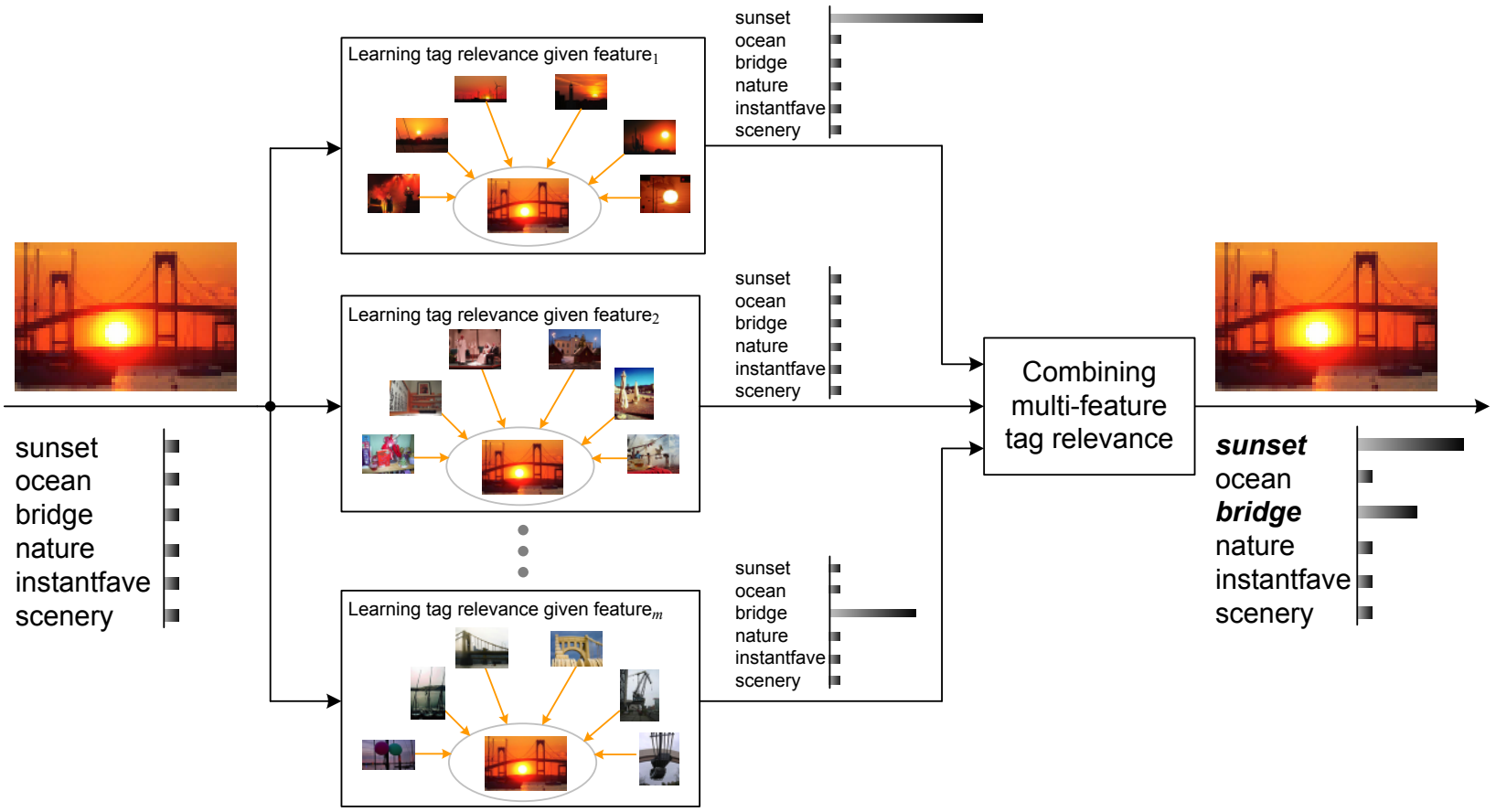
Nearest neighbor for tag relevance

Objective tags are identified

Based on 3.5 Million images downloaded from Flickr

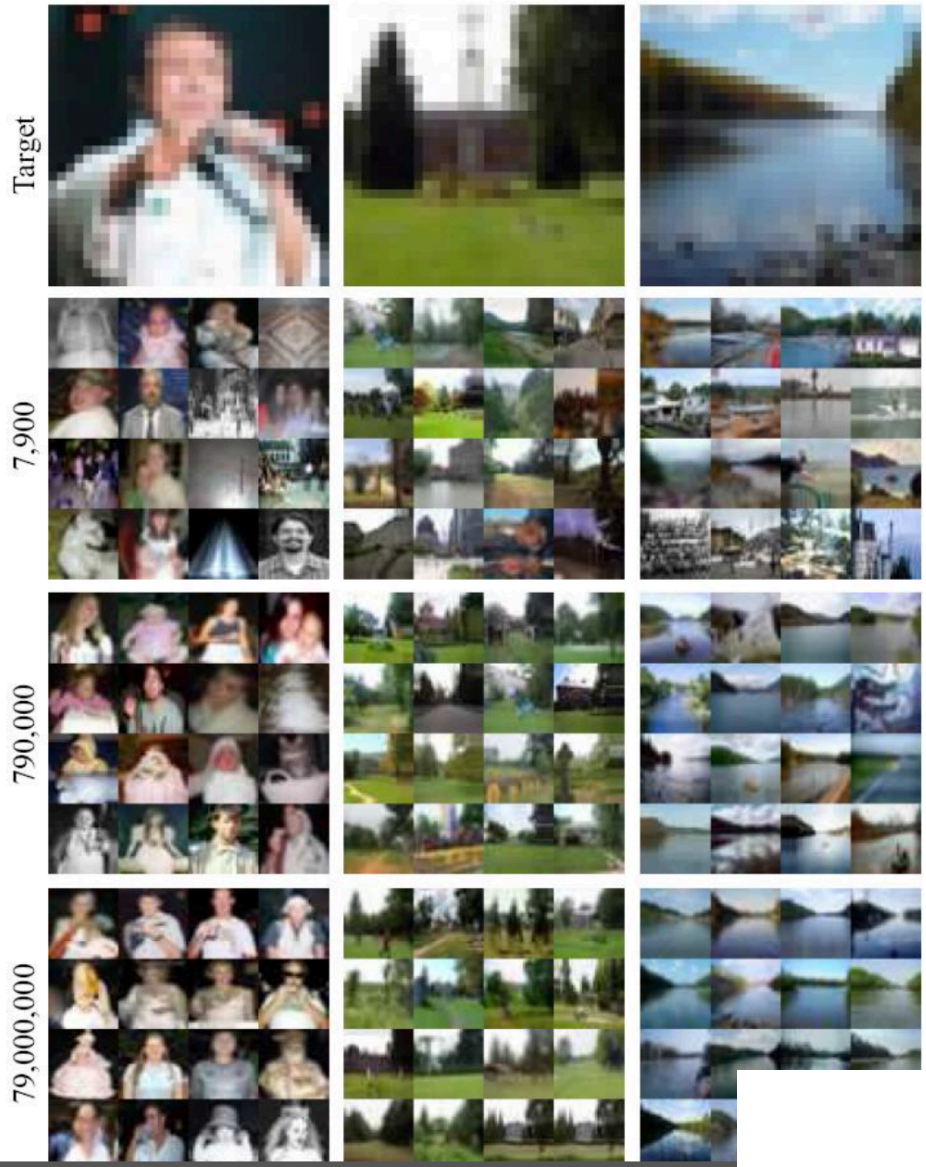
Same principle, diverse features

Fully unsupervised, adds 10% in performance



Xirong Li, CIVR 2010, best paper

Even more efficient with tiny images



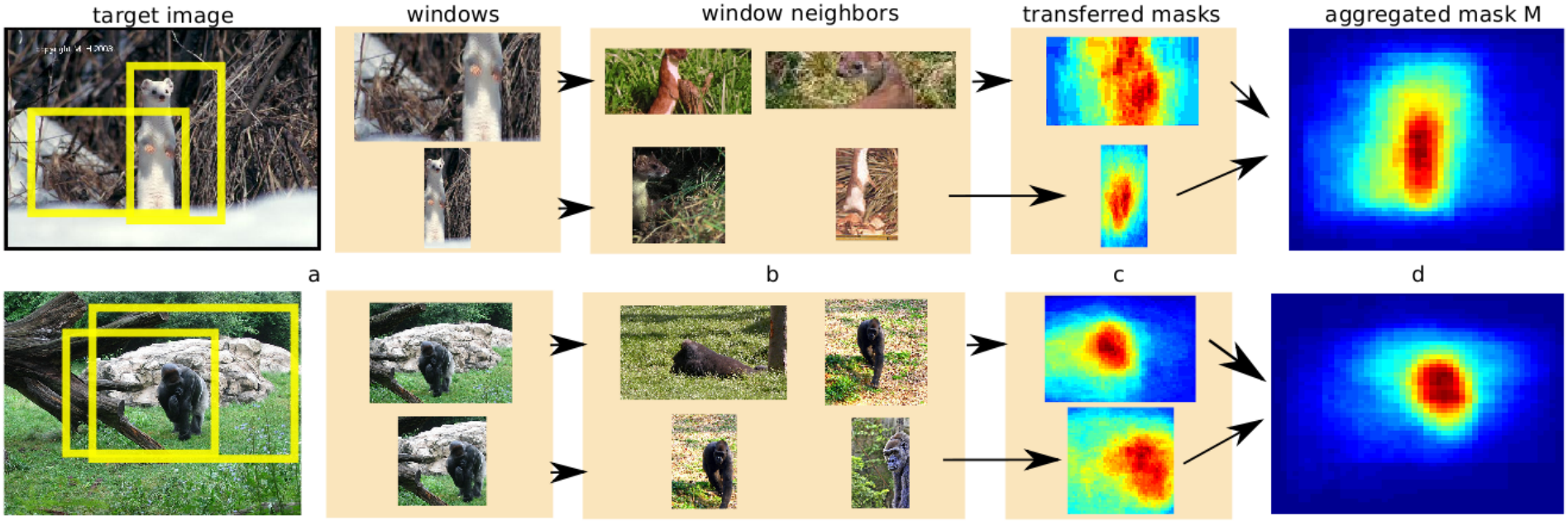
32x32 resolution

80M images

Nearest neighbor

Torralba, PAMI 2008

Nearest neighbor for segments



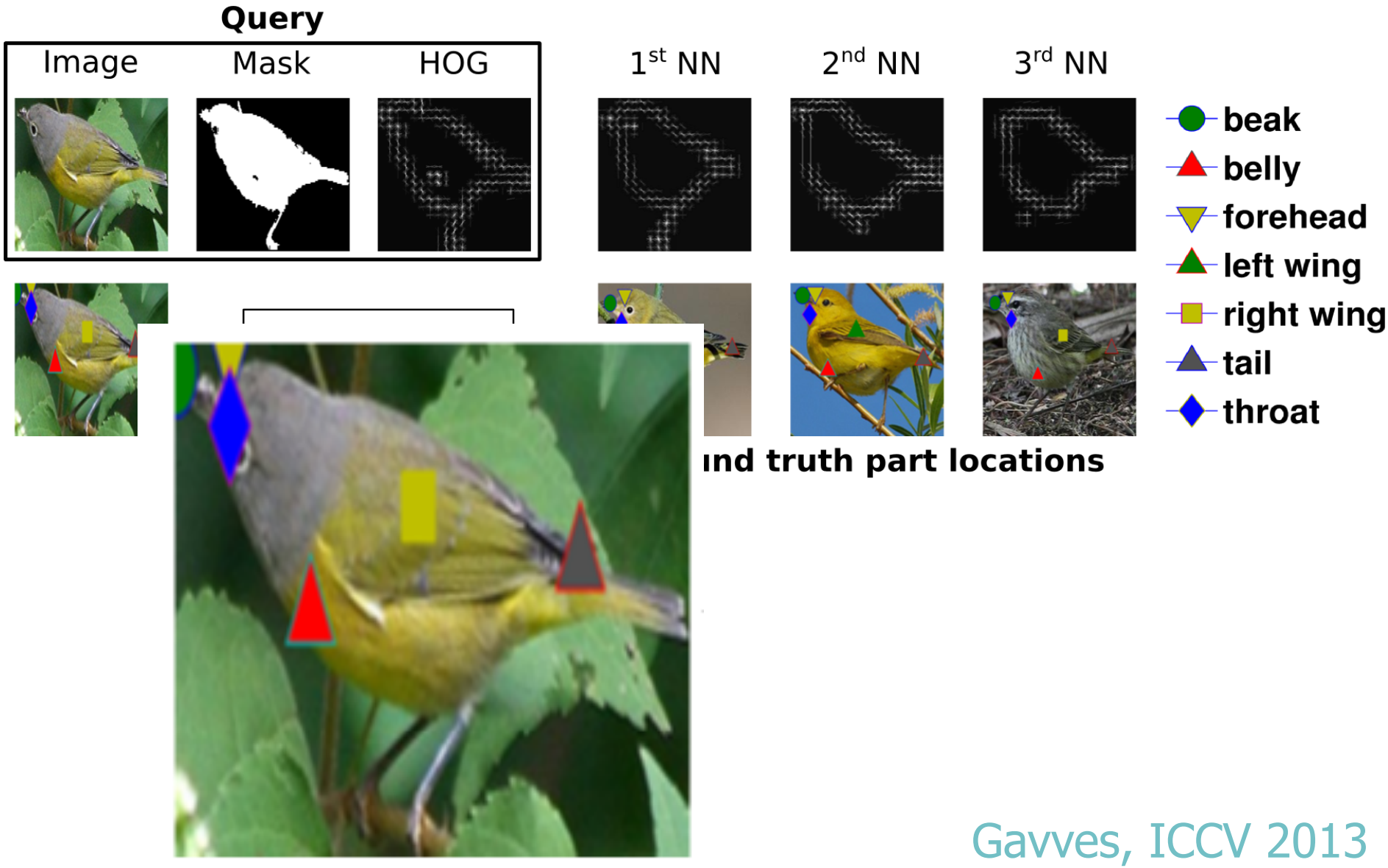
Annotates many classes with accurate segmentations

Scales efficiently

Segmentations available

Kuettel, ECCV 2012, best paper

Nearest neighbor for parts



Gavves, ICCV 2013

Nearest neighbor localized actions?

Write paper.

Take home message

Nearest neighbor with simple visual features provides a free, scalable and effective means to collect valuable data for many computer vision by learning problems.

5. Negative labels

Computer vision by learning tends to misclassify negative examples which are visually similar to positive ones, inclusion of such misclassified and thus relevant negatives should be stressed during learning. User-tagged images are abundant online, but which images are the relevant negatives remains unclear. We consider Negative Bootstrap, which iteratively finds relevant negatives. Per iteration, it learns from a small proportion of many user-tagged images, yielding an ensemble of meta classifiers. For efficient classification, it uses Model Compression such that the classification time is independent of the ensemble size.

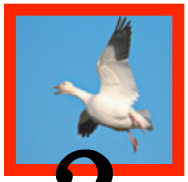
Many slides by Xirong Li

Which images are relevant negatives?

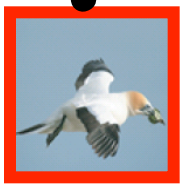
Random negatives are not necessarily relevant

Negatives

Decision boundary

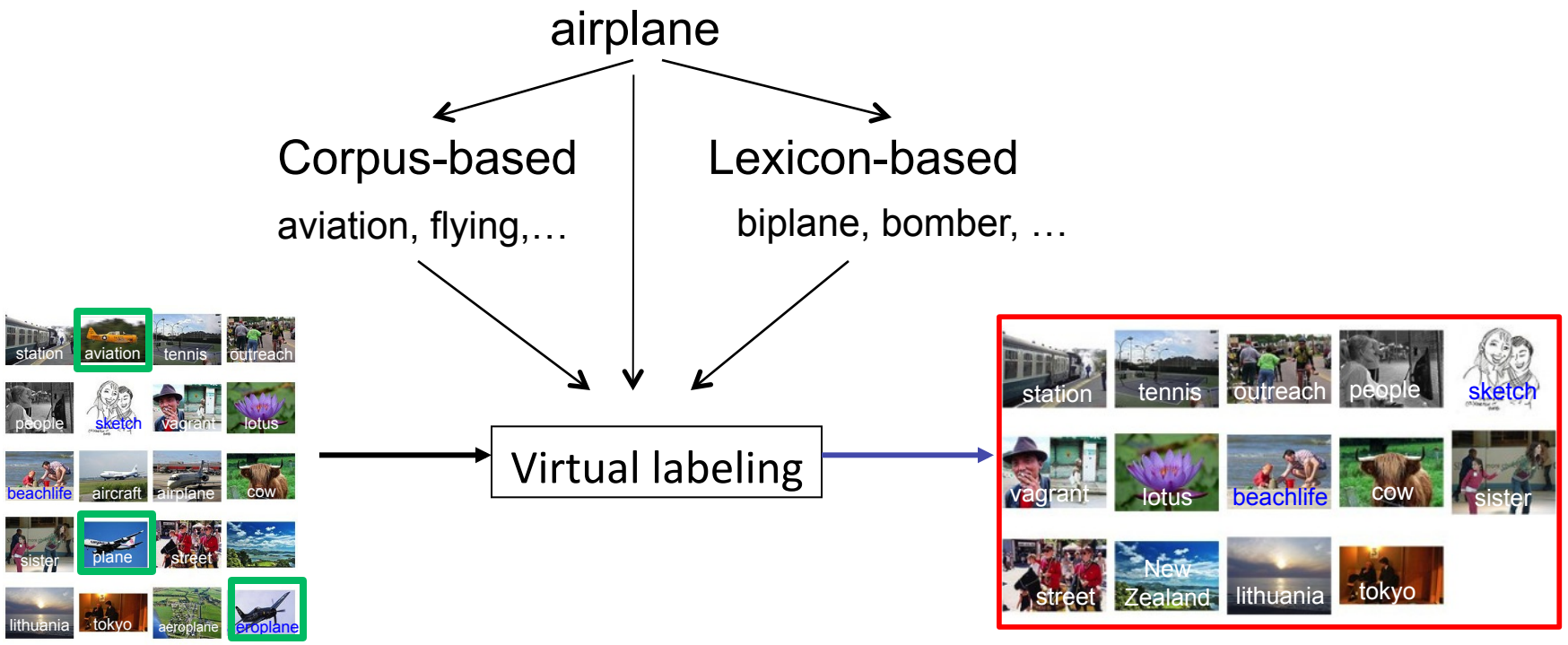


?



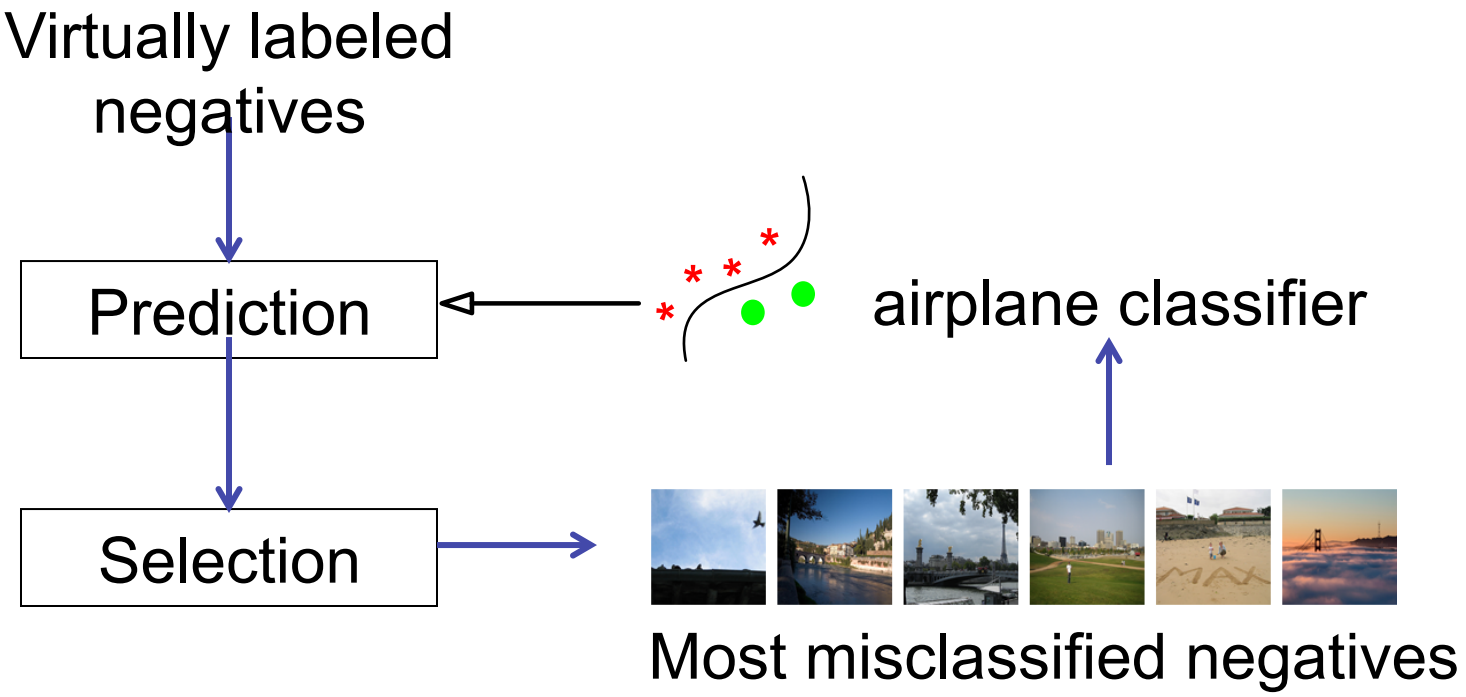
Positives

Negatives for free by virtual labeling

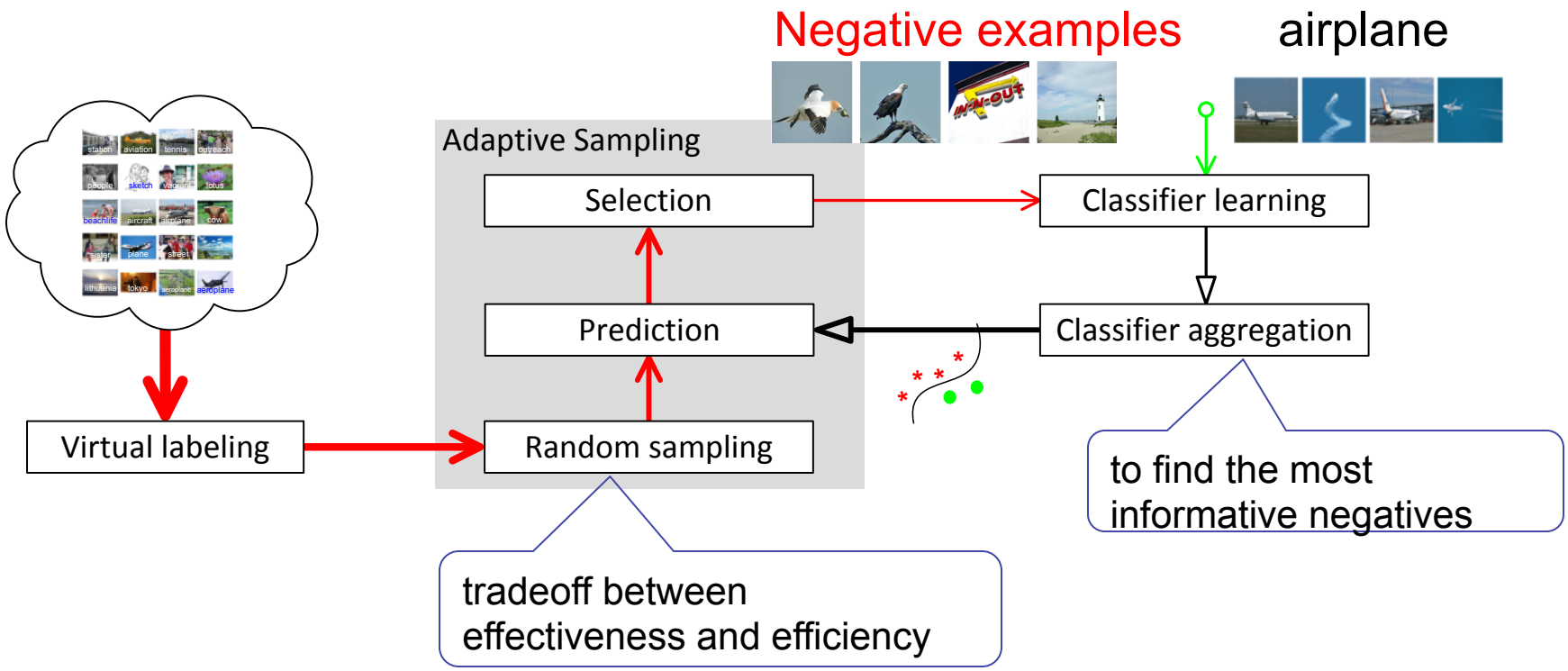


Identifying most relevant negatives

Select most misclassified negatives as the relevant negatives
Then iterate

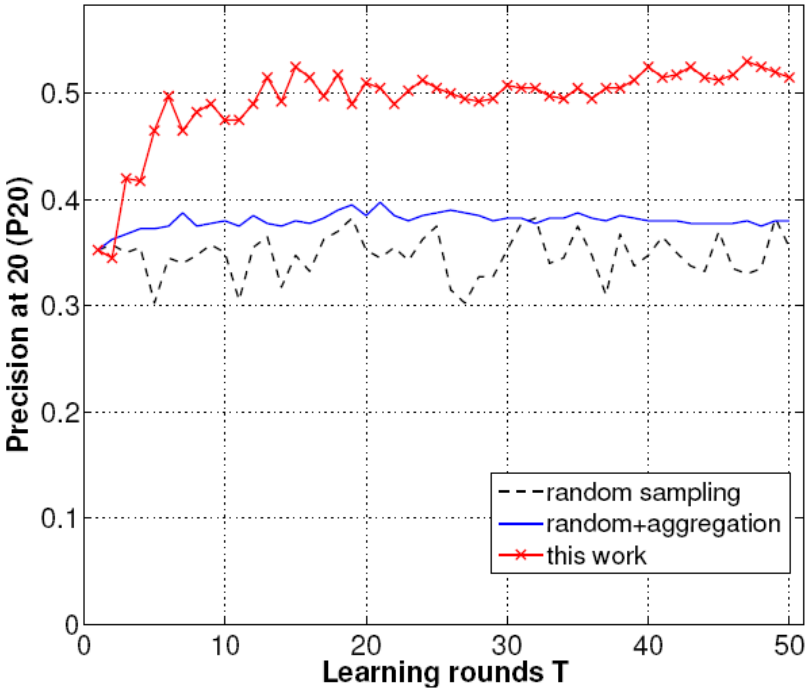


Negative Bootstrap

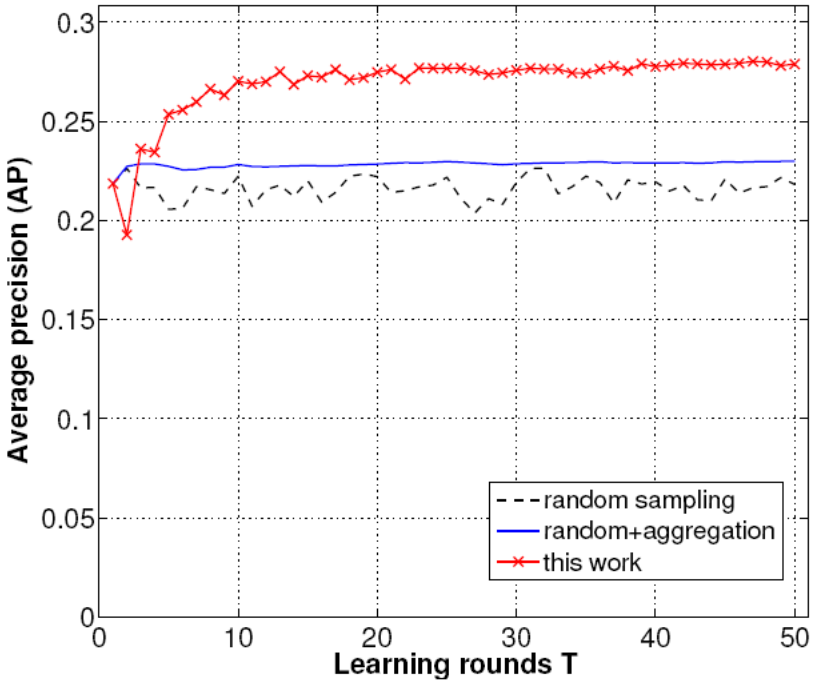


Negative Bootstrap vs State of the Art

Negatives are more useful when very few positives are available. Random negatives are not always informative

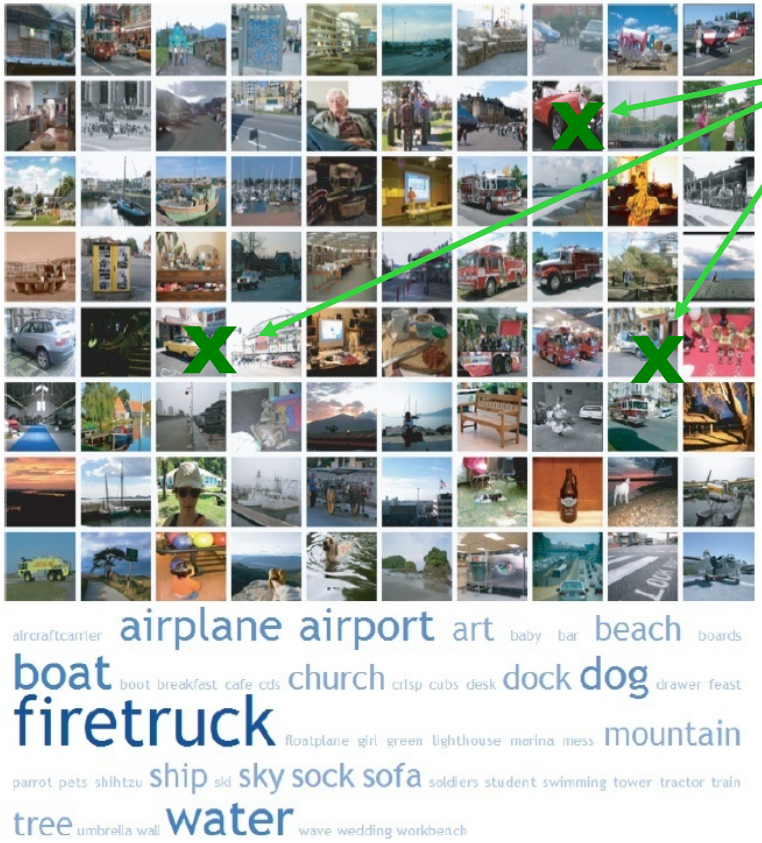


(a) Overall comparison



(a) Test set VOC08-val

Relevant negatives of ‘car’



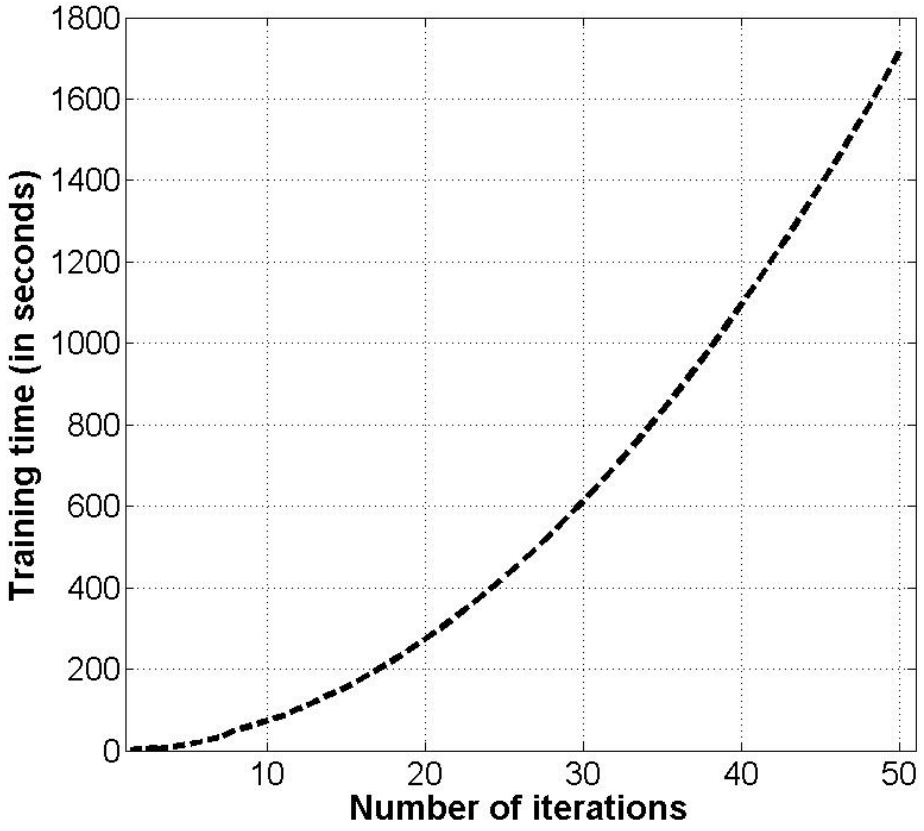
As genuine positives are in the minority, their impact is minimal

Because the tag **bus** is related to ‘**car**’, examples of ‘bus’ are excluded.

As an alternative, examples of `**firetruck**’ are identified as informative negatives.

The Efficiency Problem

Classification time is proportional to the number of classifiers



Histogram Intersection Kernel SVM

SVM decision function

Support vectors Kernel

$$g(x) = b + \sum_{j=1}^{n} \alpha_j \cdot y_j \cdot \mathcal{K}(x, x_j)$$

Weight per support vector

Features

$$\mathcal{K}(x, x') = \sum_{i=1}^{d} \min(x(i), x'(i))$$

Constraints

$$\sum_{j=1}^{n} \alpha_j \cdot y_j = 0,$$

$$0 \leq \alpha_j \leq C, j = 1, \ldots, n$$

Fast Intersection Kernel SVM

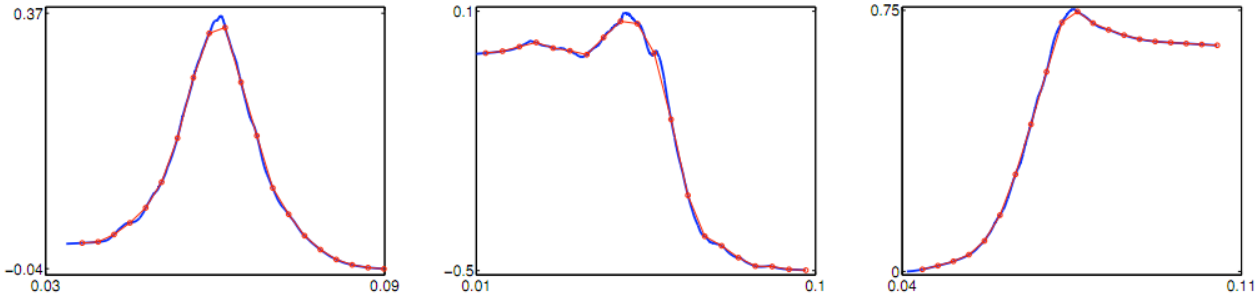
Histogram Intersection Kernel is additive

Support vectors # Features

$$
\begin{aligned}
g(x) &= \sum_{l=1}^{m} \alpha_l y_l \left(\sum_{i=1}^{n} \min\left(x(i), x_l(i)\right) \right) + b \\
&= \sum_{i=1}^{n} \left(\sum_{l=1}^{m} \alpha_l y_l \min\left(x(i), x_l(i)\right) \right) + b \\
&= \sum_{i=1}^{n} h_i(x(i)) + b
\end{aligned}
$$

Support vectors per dimension

$h_i(x(i))$ can be approximated by linear interpolation



Maji PAMI 2013

An ensemble of classifiers

Linearly combined classifiers

$$G_T(x) = \sum_{t=1}^{T} \lambda_t \cdot g_t(x, w)$$

Weight per classifier

Linearly combined Histogram Intersection Kernel SVMs

$$G_T(x) = \sum_{t=1}^{T} \lambda_t \cdot b_t + \underbrace{\sum_{i=1}^{d} \sum_{t=1}^{T} \sum_{j=1}^{n_t} \lambda_t \cdot \alpha_{t,j} \cdot y_{t,j} \cdot \min(x(i), x_{t,j}(i))}_{\text{decision value per dimension}}$$

Model Compression

Extending FIK-SVM to classifier ensembles

Decision function per dimension

$$H_i(z) = \sum_{t=1}^{T} \sum_{j=1}^{n_t} \lambda_t \cdot \alpha_{t,j} \cdot y_{t,j} \cdot \min(z, x_{t,j}(i))$$

Sort the *i*-th dimension of support vectors in all meta classifiers

$$\{x_{1,1}(i), \ldots, x_{1,n_1}, \ldots, x_{T,1}, \ldots, x_{T,n_T}(i)\} \longrightarrow \bar{x}_j(i)$$

$$H_i(z) = \sum_{j=1}^{M} \bar{\lambda}_j \cdot \bar{\alpha}_j \cdot \bar{y}_j \cdot \min(z, \bar{x}_j(i))$$

For any *z* within $[\bar{x}_1(i), \bar{x}_M(i)]$

$$H_i(z) = \beta_i \cdot H_i(\bar{x}_{r_i}(i)) + (1 - \beta_i) \cdot H_i(\bar{x}_{r_i+1}(i))$$

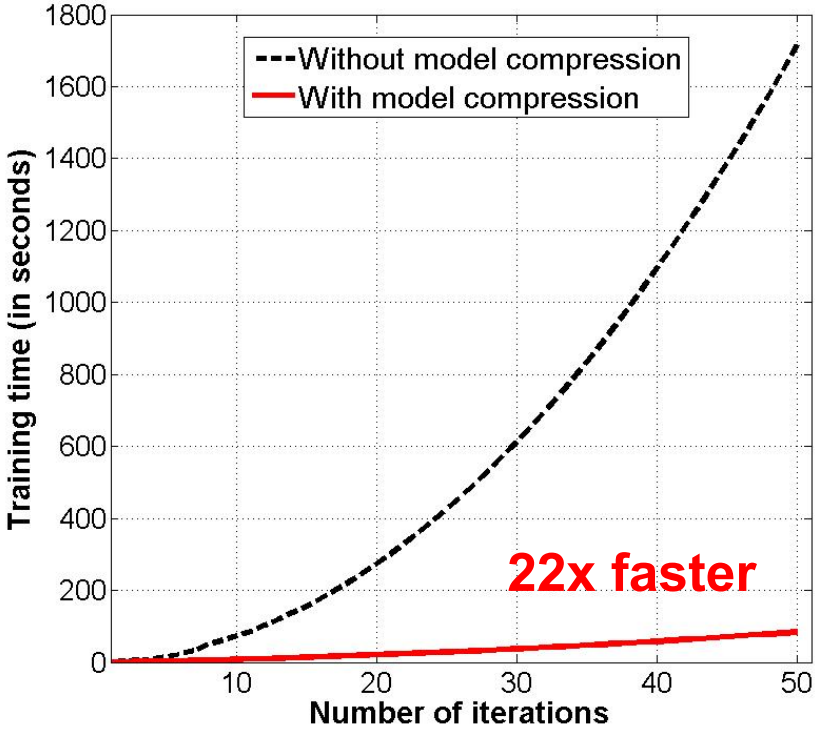
Li TMM 2013

The Influence of Model Compression

Effectiveness

Test data	Metric	Model Compression	
		No	Yes
VOC08val	Training time (seconds)	1,736	77
	Test time (seconds)	190	0.6
	mAP	0.306	0.304
	P@20	0.503	0.512
NUSfuture	Test time (seconds)	5,664	18
	mAP	0.171	0.171
	P@20	0.436	0.443

Efficiency



Take home message

Computer vision by learning without the need of labeling any negative examples

Negative Bootstrap is much better than random sampling

Negative Bootstraps & model compression find relevant negatives: effective *and* efficient.

Overview

1. Video benchmarks, TRECVID, average precision, progress
2. Image benchmarks, PASCAL, ImageNet, lessons learned
3. Labels from humans, experts, volunteers, crowdsourcing
4. Labels from similarity, nearest neighbor, simple features
5. Negative labels, negative bootstrapping, model compression

6. Learning using attributes