Computer Vision by Learning

Cees Snoek, UvA
Arnold W.M. Smeulders, UvA
Efstratios Gavves, UvA
Laurens van der Maaten Facebook

Advanced School for Computing and Imaging

University of Amsterdam

Tomorrow

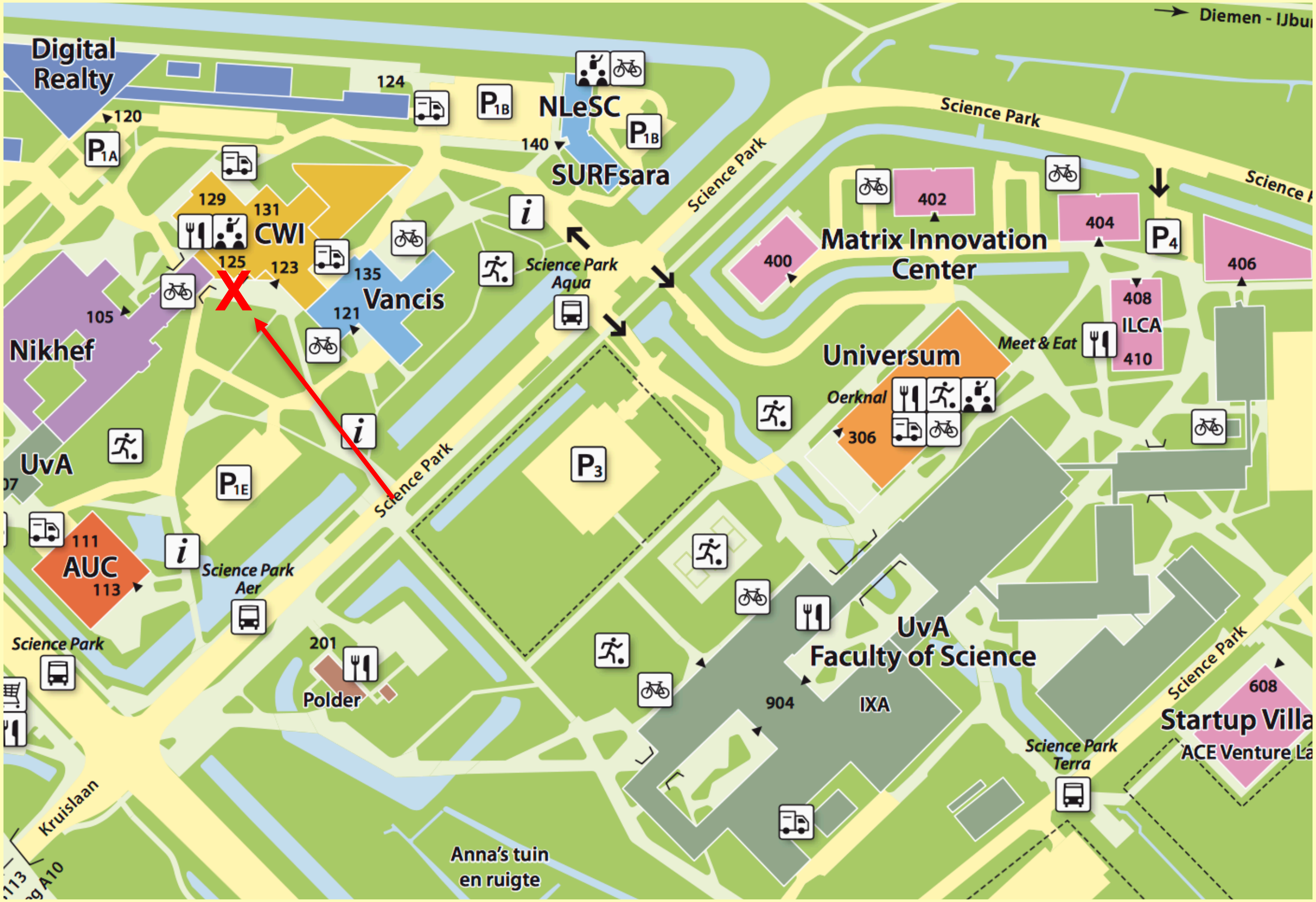
Invited tutorial by **Laurens van der Maaten**

- Understanding and Improving Convolutional Networks
- From Visual Recognition to Visual Reasoning

Note change of location

- **CWI, Z009 Eulerzaal**

CWI, Z009 Eulerzaal

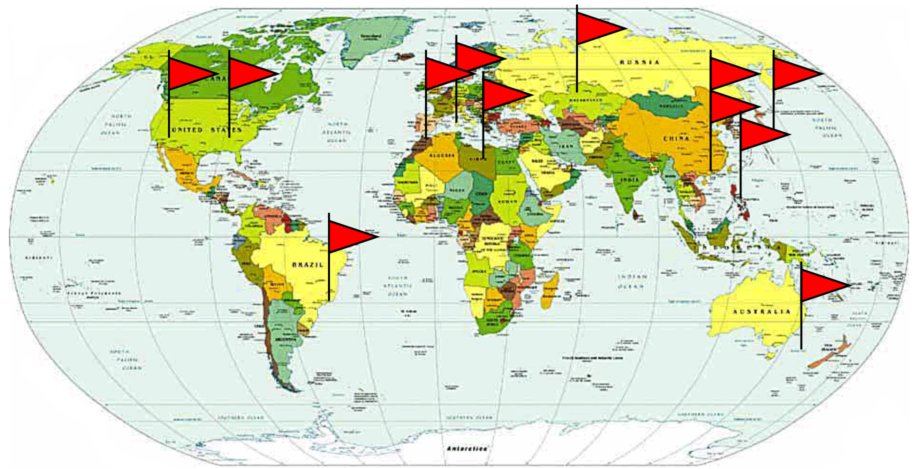


Overview

1. Image benchmarks, PASCAL, ImageNet, MSCOCO
2. Video benchmarks, TRECVID, ActivityNet
3. Labels from humans, experts, volunteers, crowdsourcing
4. Labels from similarity, nearest neighbor, simple features
5. Weakly-supervised computer vision

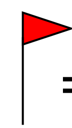
6. Event recognition by learning

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

1. 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

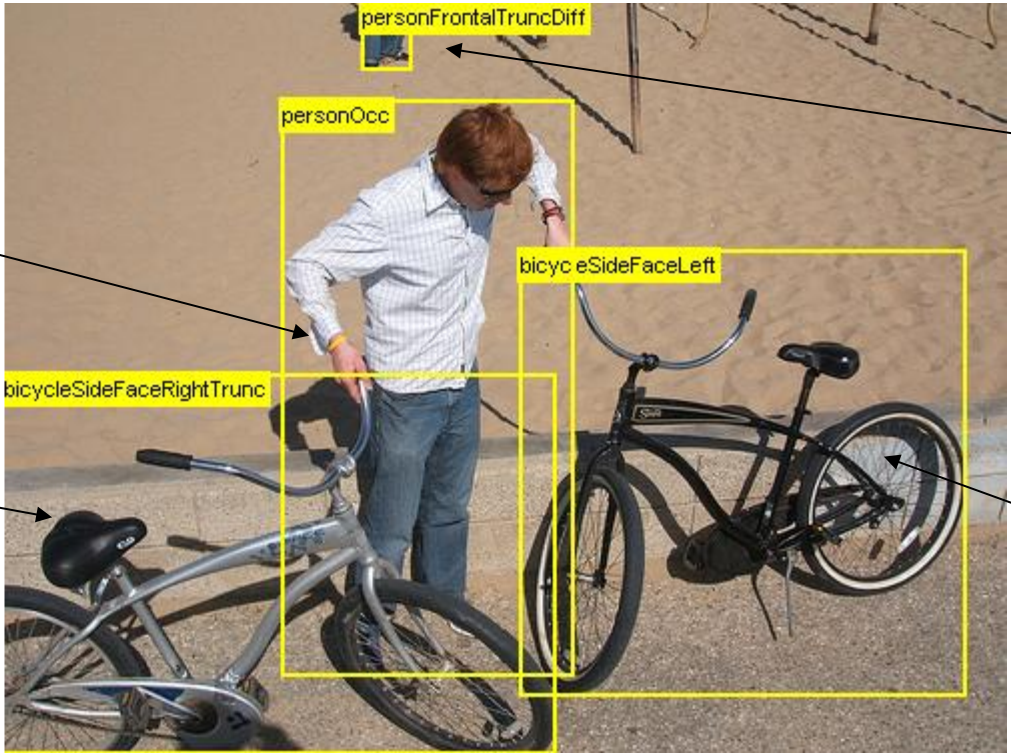
Pascal Dataset Collection

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

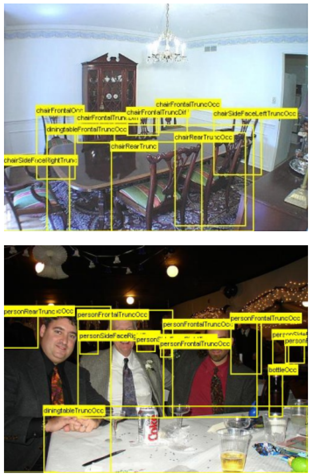


Difficult
Not scored in evaluation

Pose
Facing left

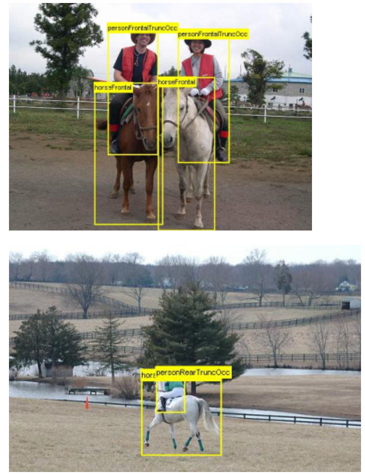
Examples

Dining Table

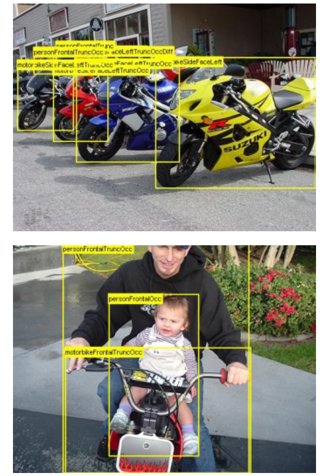


Dog

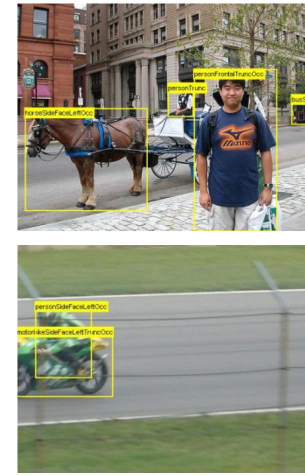
Horse



Motorbike

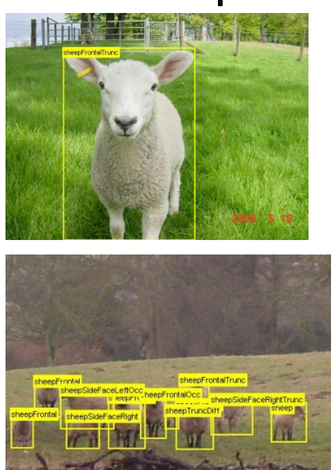


Person

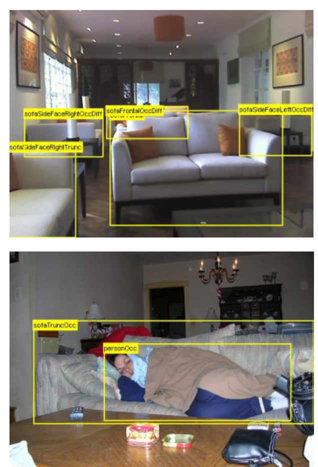


Potted Plant

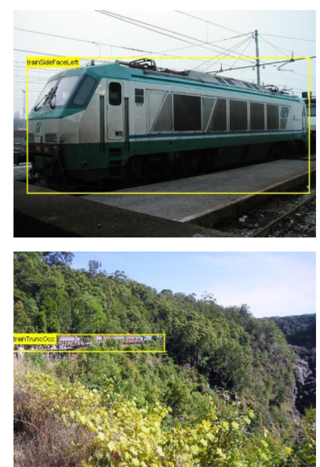
Sheep



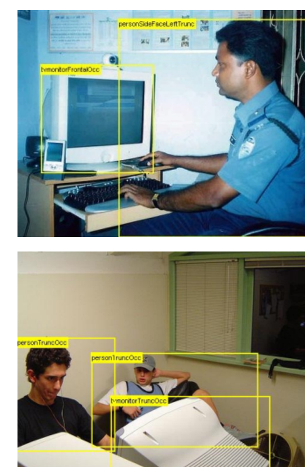
Sofa



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

ImageNet Challenge

Yearly competition

- Automatically label 1.4M images with 1K objects
- Measure top-5 classification error

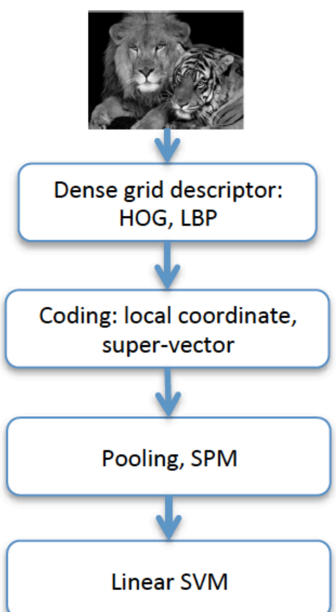
Output	Output
Scale	Scale
T-shirt	T-shirt
Steel drum ✔	Giant panda ✗
Drumstick	Drumstick
Mud turtle	Mud turtle

Slide credit: Andrej Karpathy

Some highlights

Year 2010

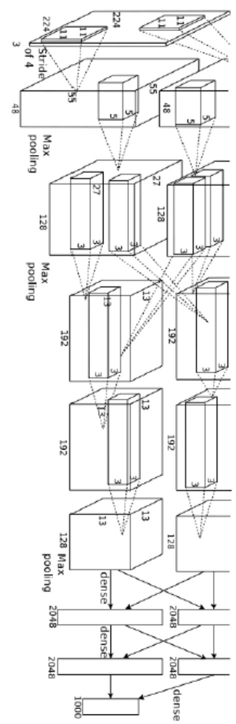
NEC-UIUC



Lin *et al.* CVPR11

Year 2012

SuperVision



Krizhevsky *et al.* NIPS12

Year 2014

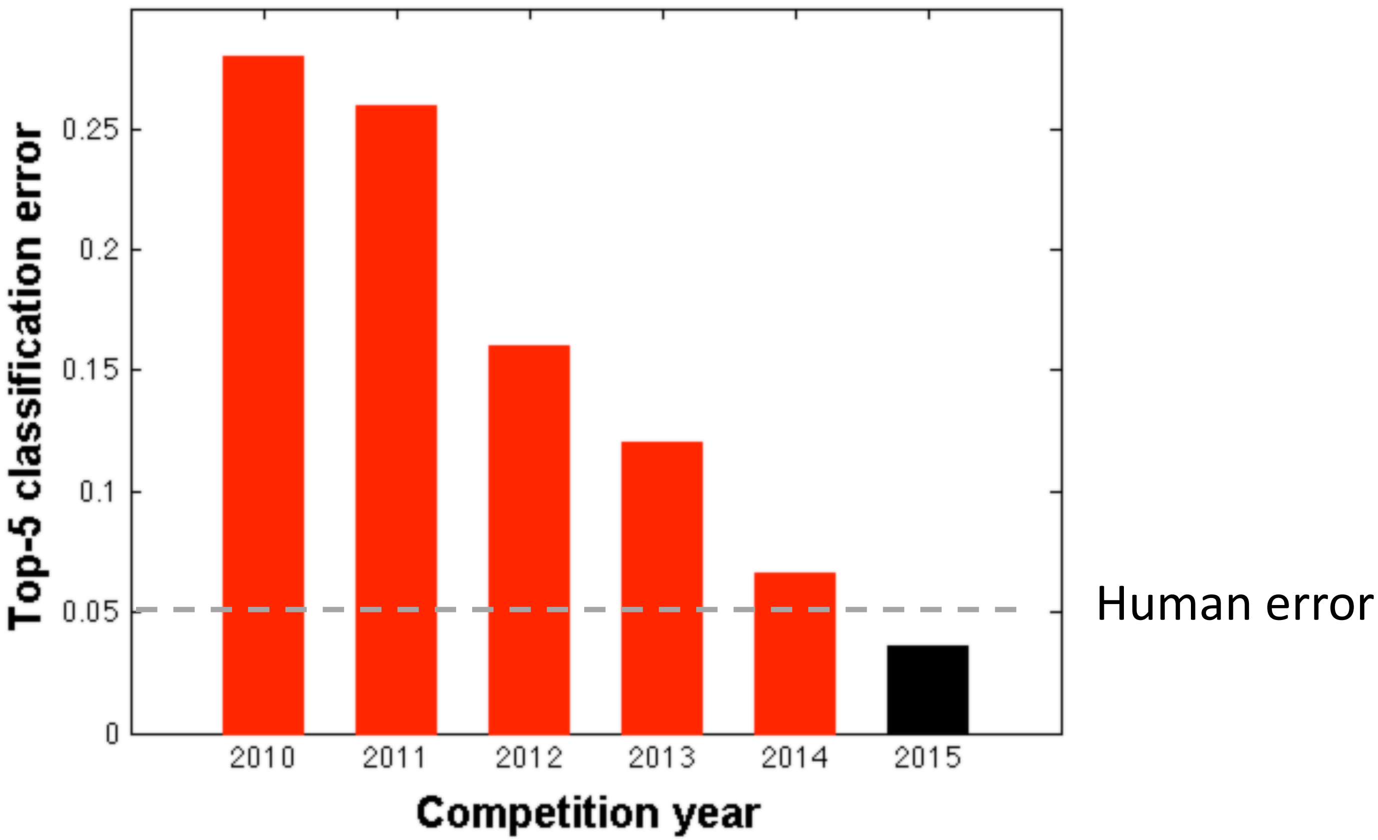
GoogLeNet

Szegedy *et al.* CVPR15

VGG

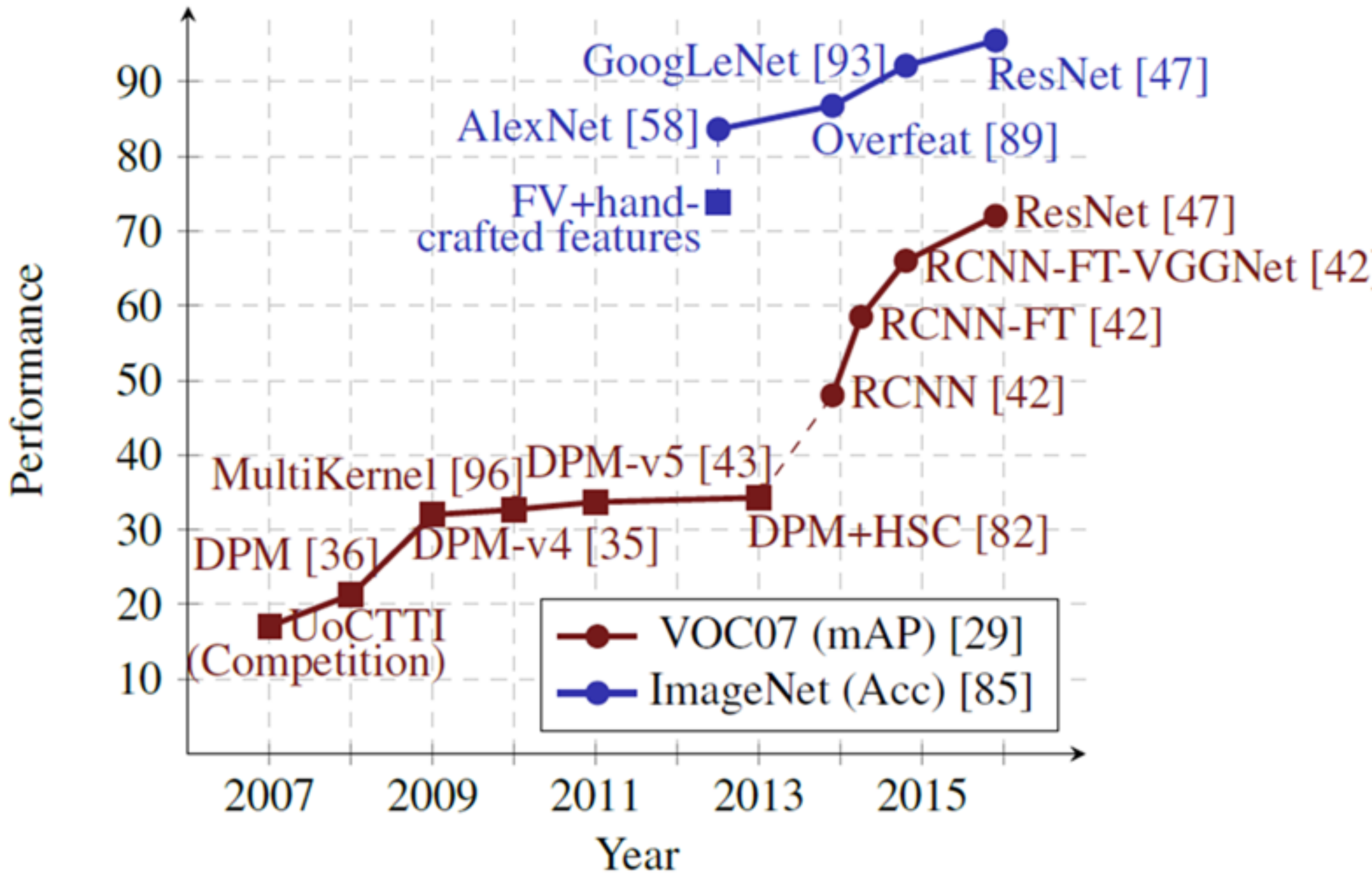
Simonyan *et al.* ICLR15

Progress in ImageNet



Machine makes less mistakes than human

Progress: Classification & Detection



ImageNet object detection

Modeled after PASCAL VOC

Algorithm outputs a list of bounding box detections with confidences

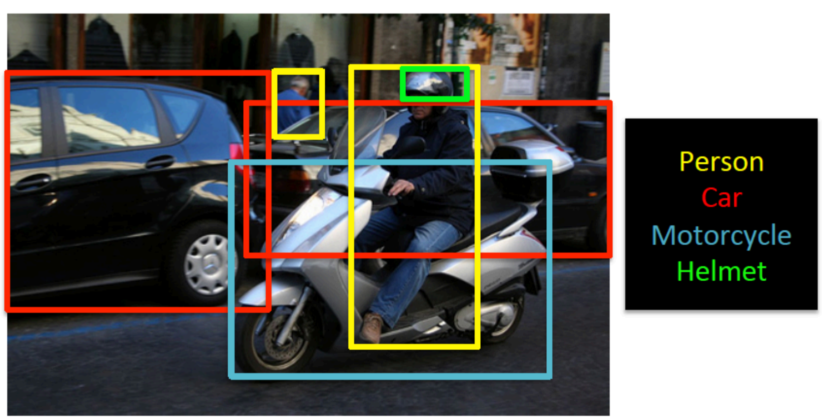
A detection is considered correct if intersection over union (IoU) overlap with ground truth > threshold (0.5)

Evaluated by average precision per object class

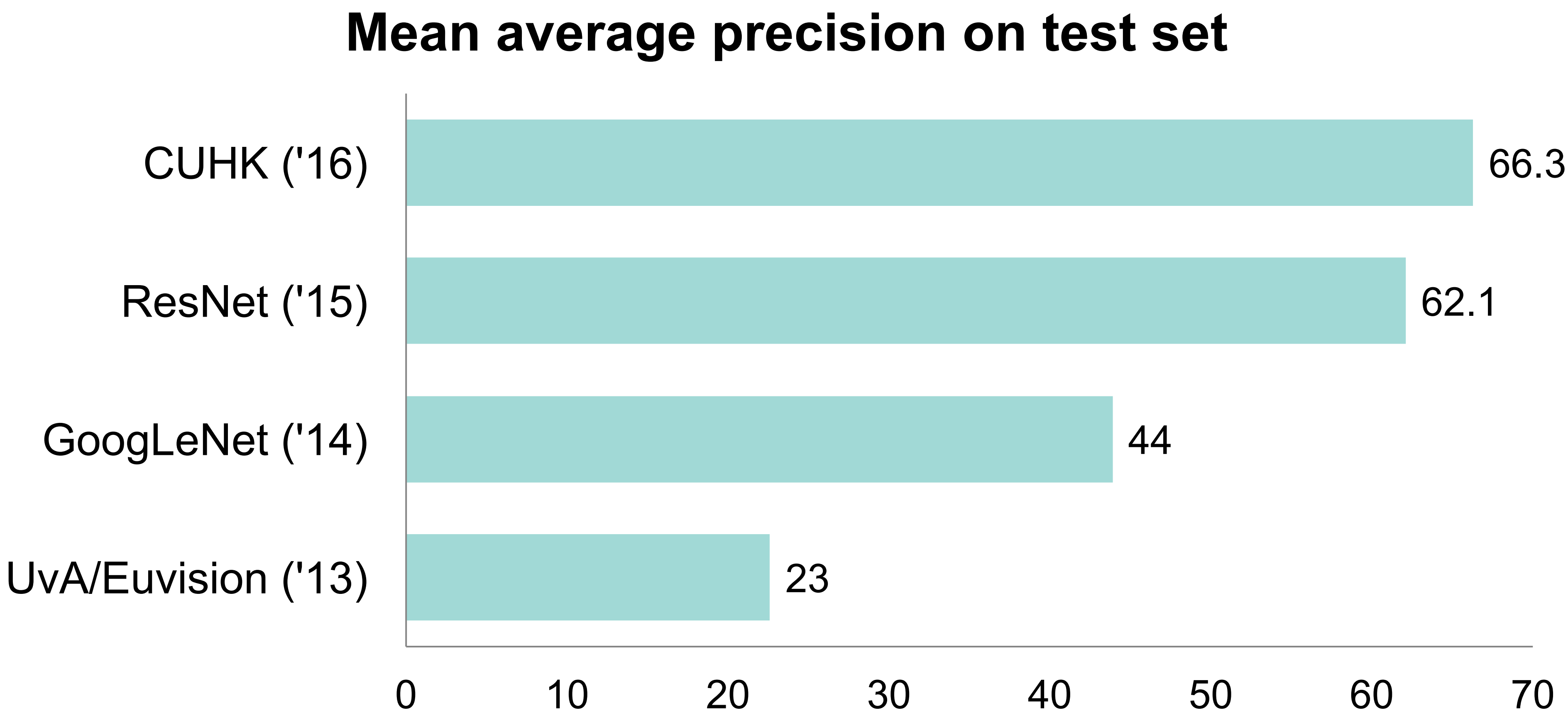
Winner is the team that wins the most object categories

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	---		---



Progress



Mean average precision on test set
CUHK ('16)
ResNet ('15)
GoogLeNet ('14)
UvA/Euvision ('13)
66.3
62.1
44
23
0
10
20
30
40
50
60
70

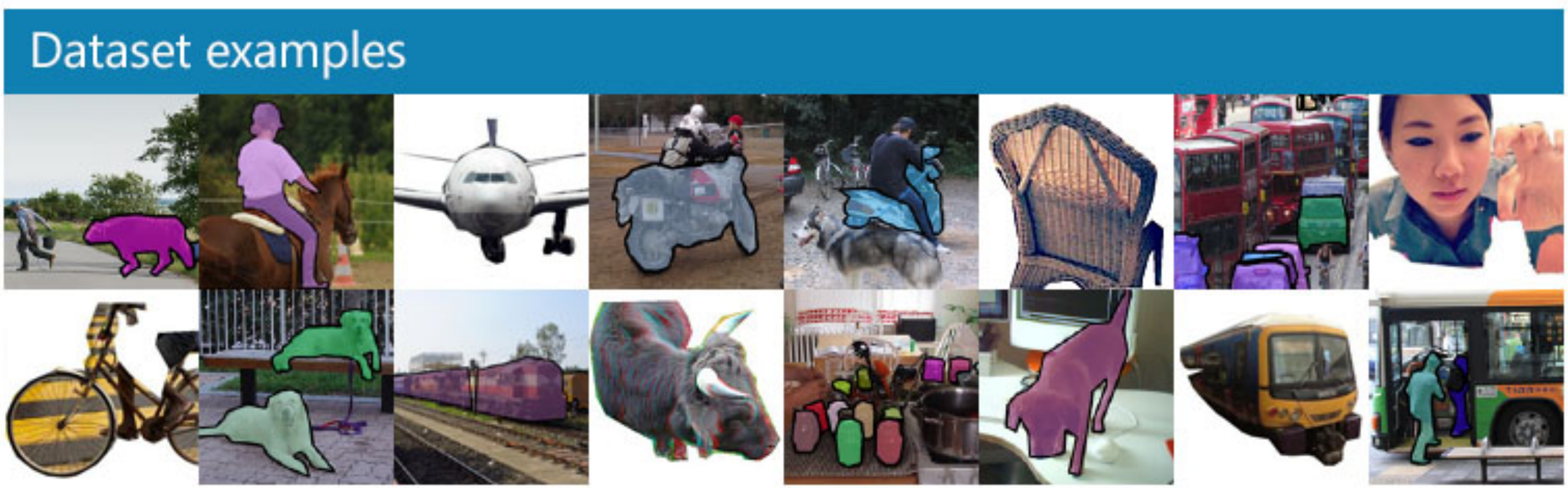
MSCOCO

80 object categories

200k images

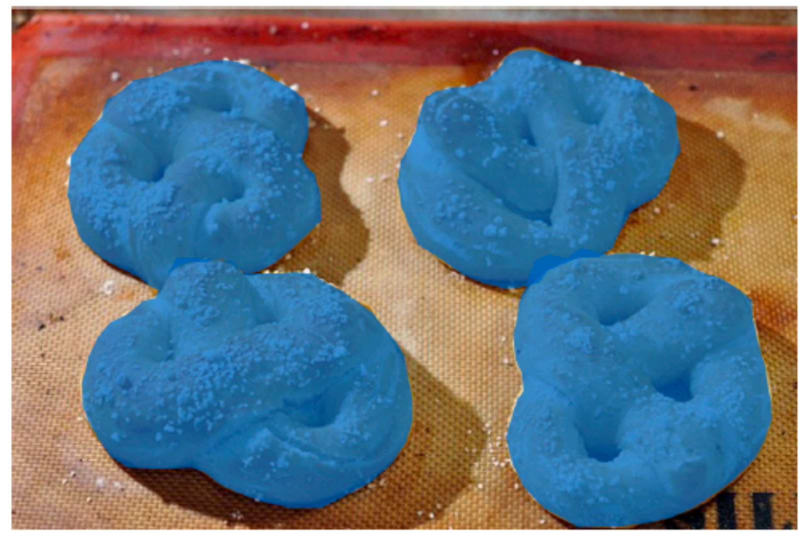
1.2M instances (350k people)

106k people with keypoints



MSCOCO.org

Instance segmentations



Every instance segmented in MSCOCO

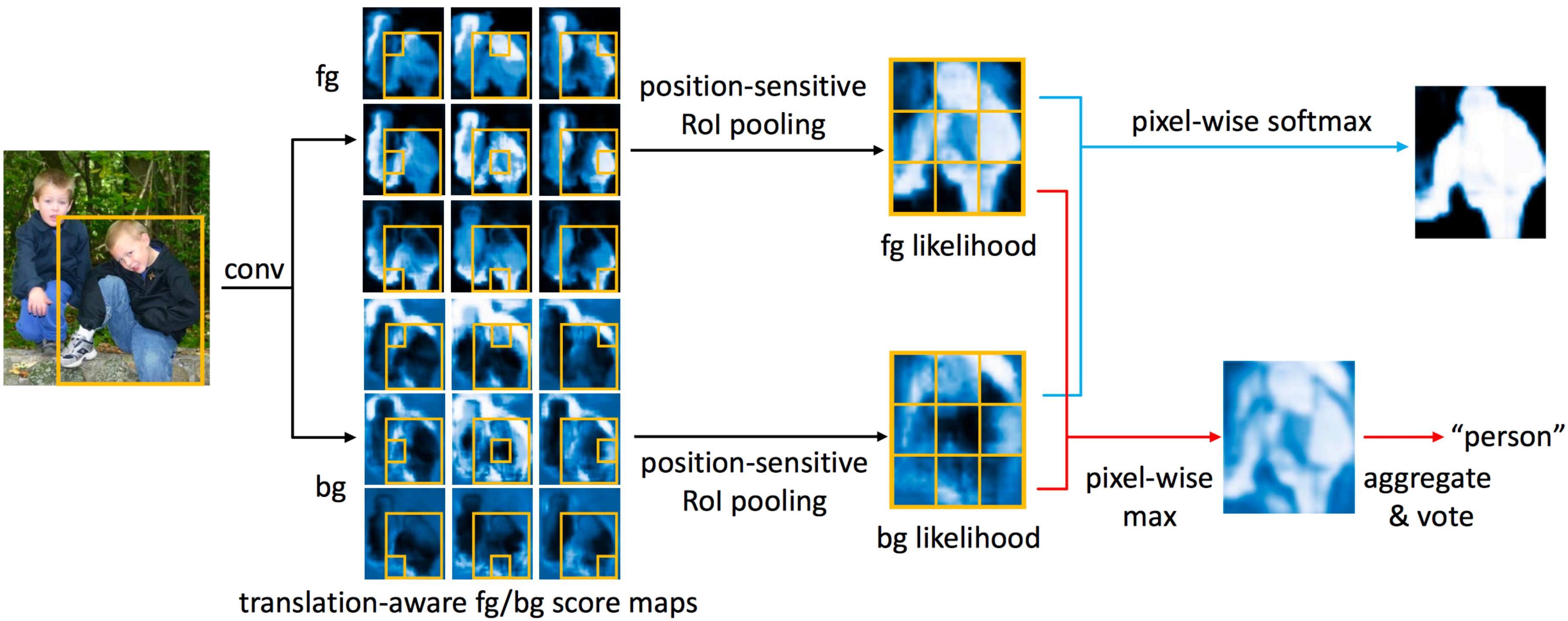
Picture from Kovashka *et al.* FnTCGV 2016

Challenges in 2016

Segmentation winner

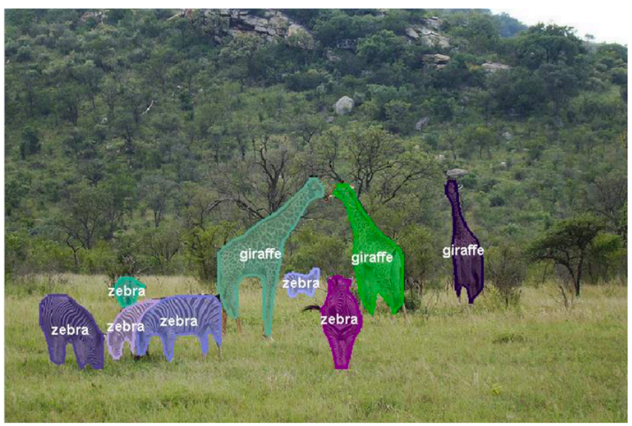
Fully convolutional end-to-end for instance segmentation

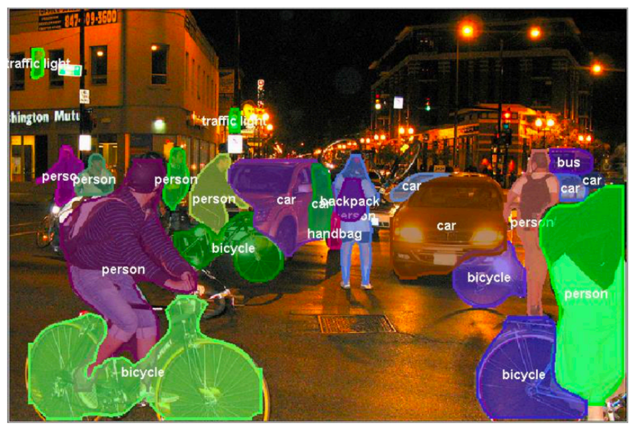
Based on ResNet-101

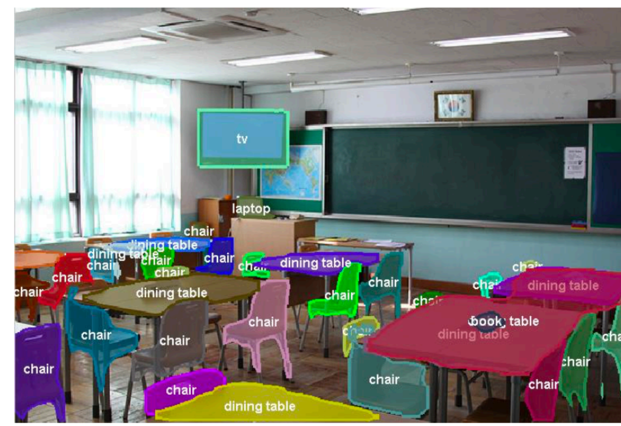


Dai *et al.* MSRA

Some results



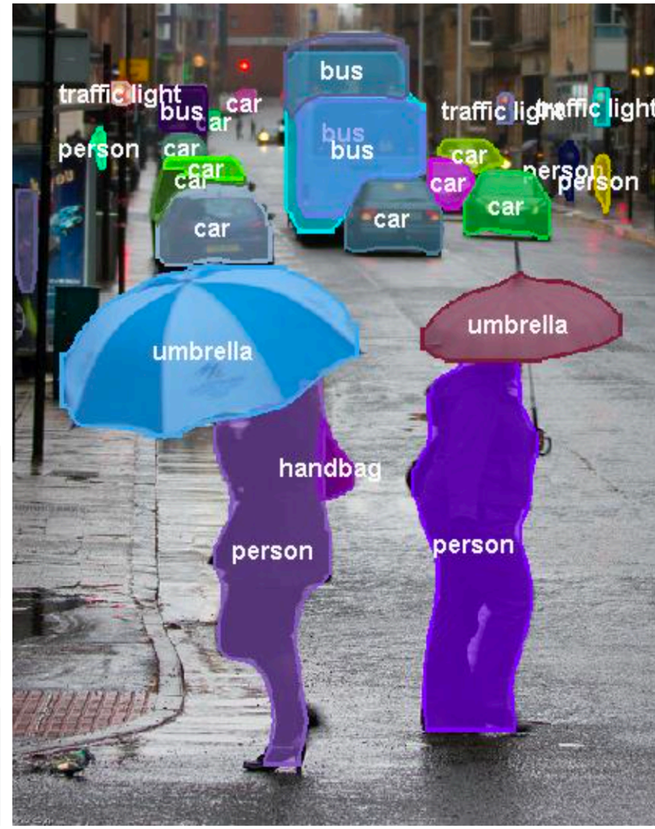


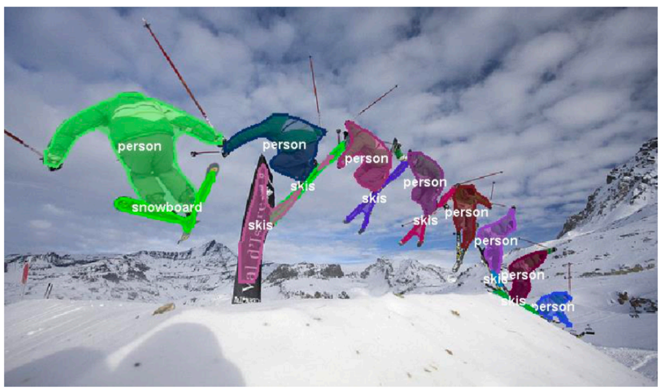


Some more


potted plant
tv
person
handbag
backpack
suitcase
laptop


boat
person


bus
traffic light
car
person
umbrella
handbag


person
snowboard
skis


potted plant
vase
car

2. 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/

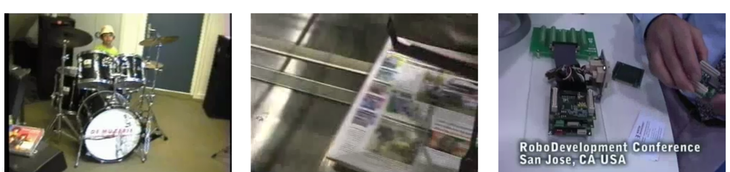
Video data sets

US TV news (`03/`04)

International TV news (`05/`06)

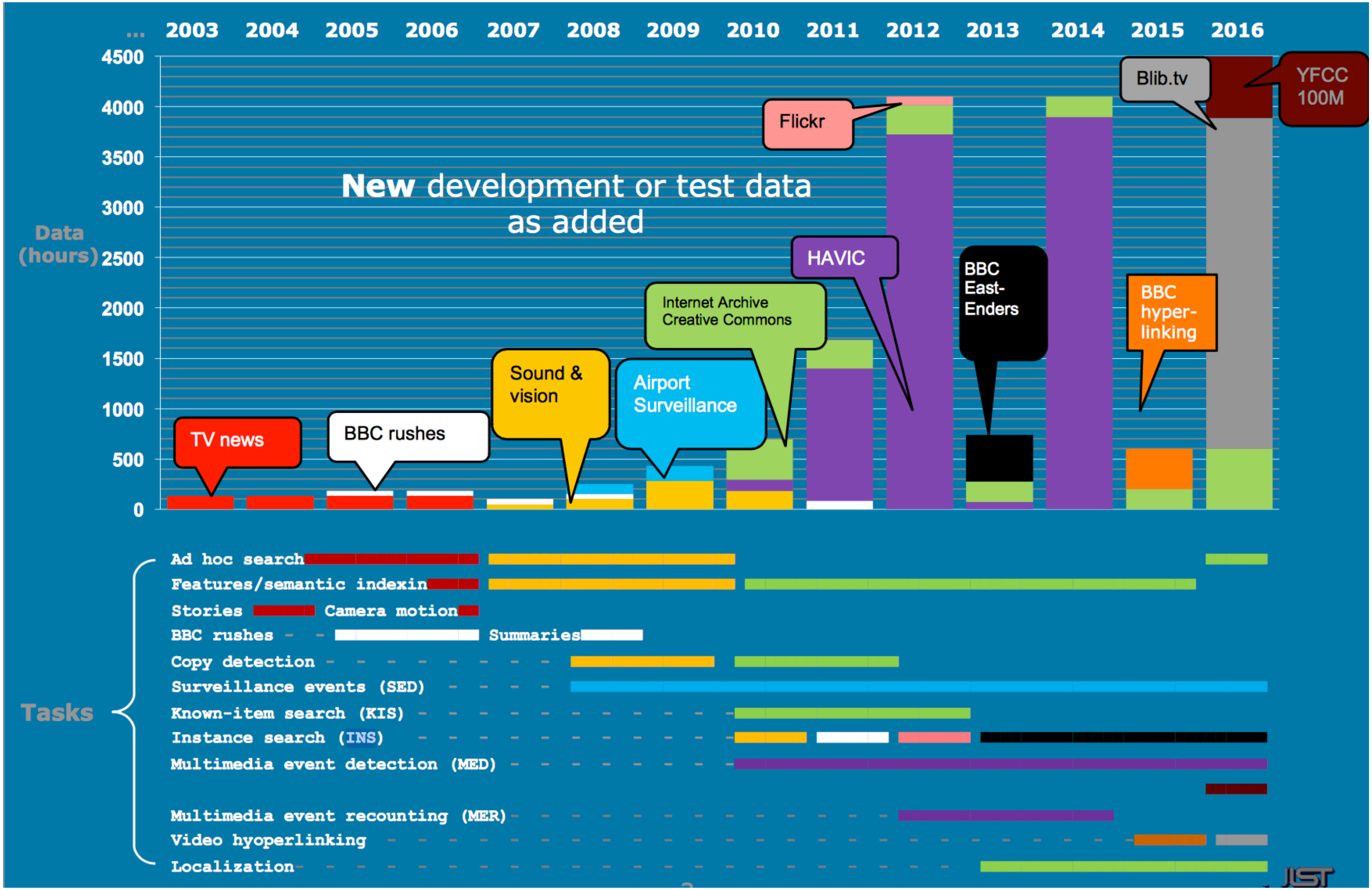
Dutch TV infotainment (`07/`08/`09)

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

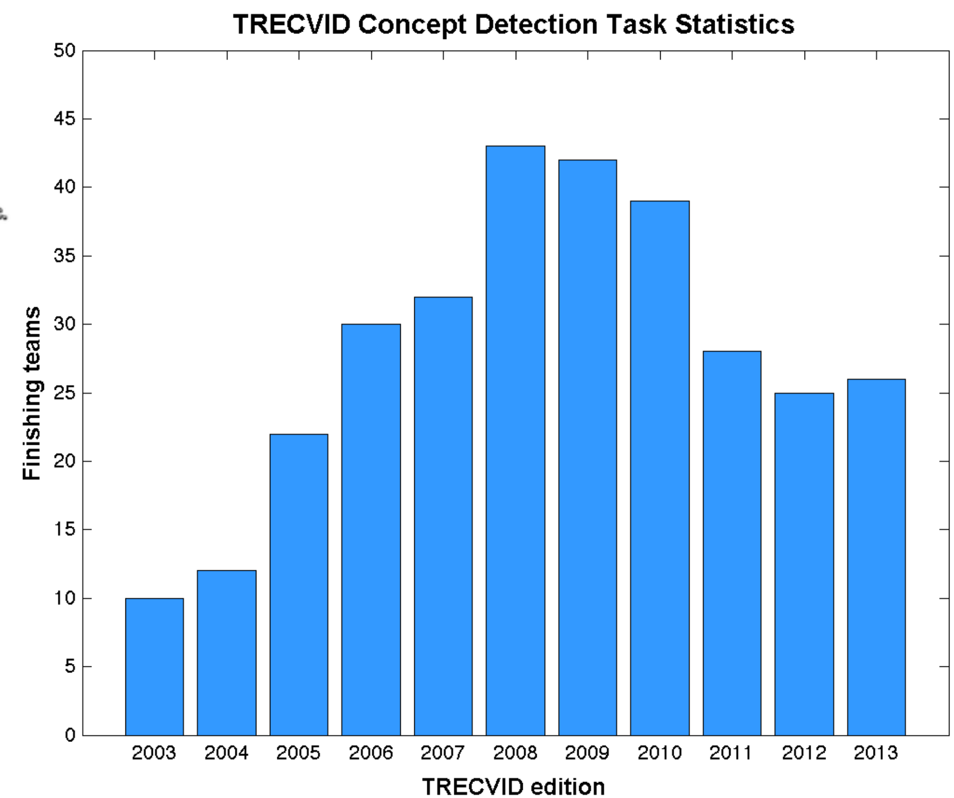
Aircraft

Beach

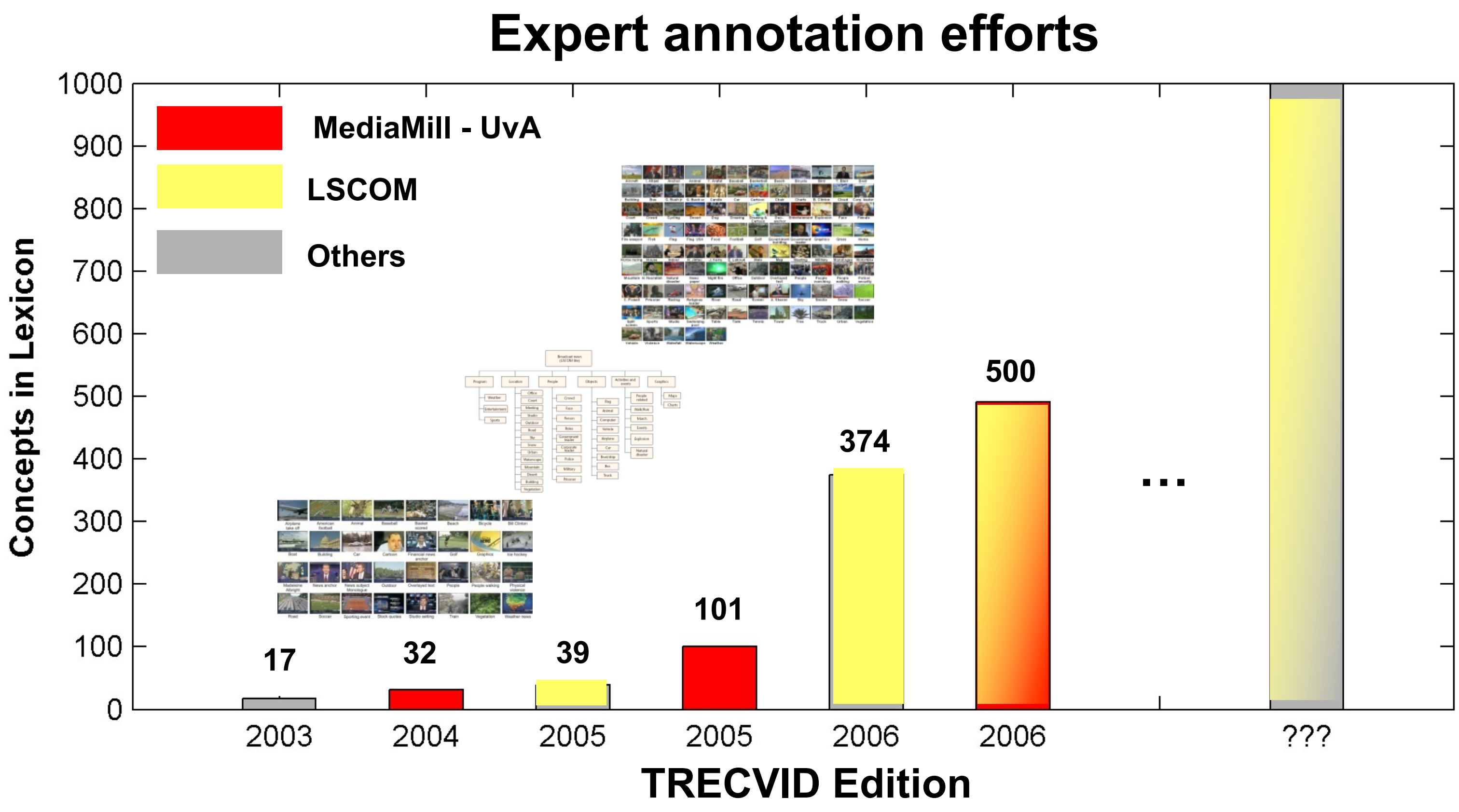
Mountain

Note the variety in visual appearance

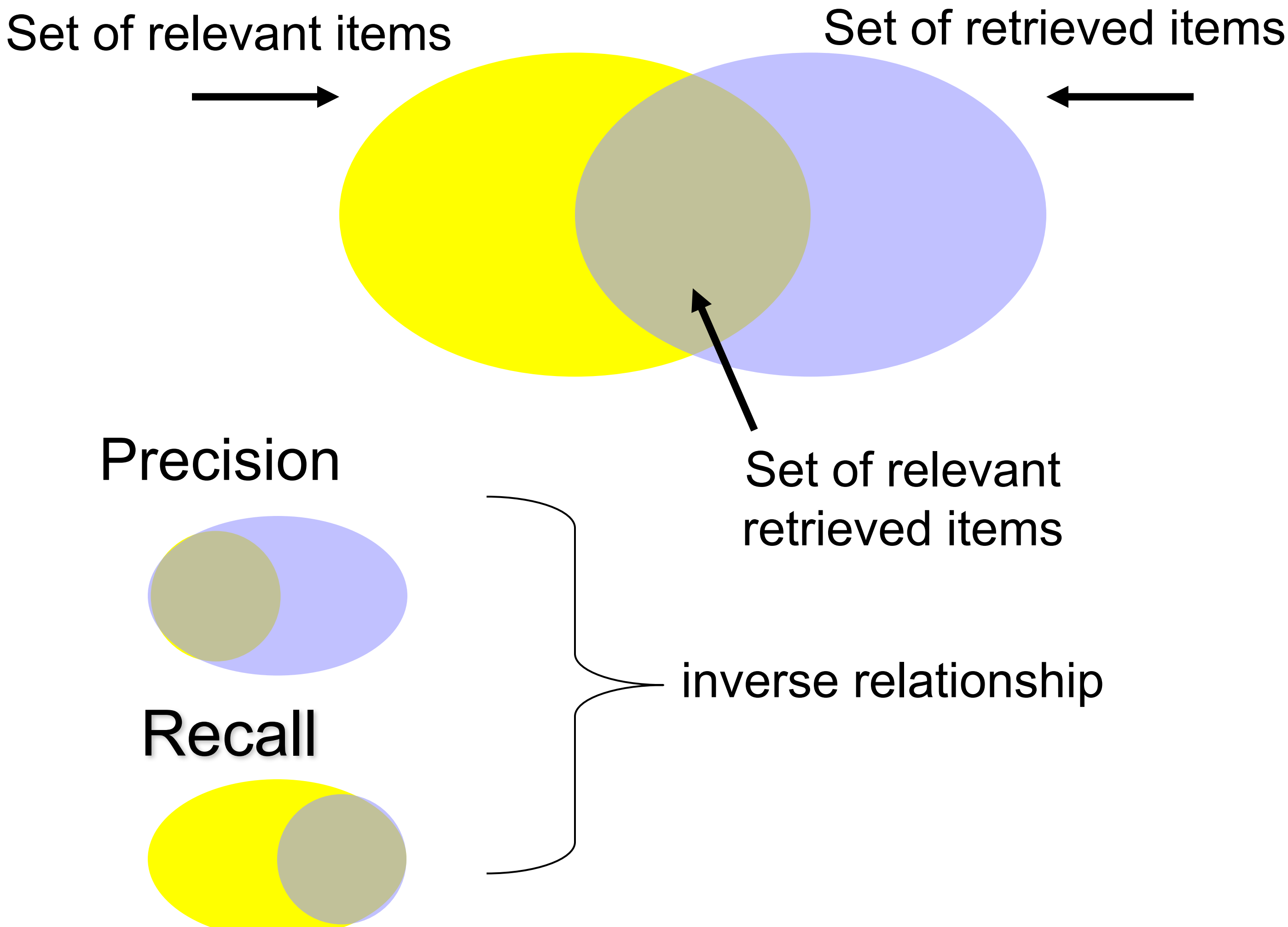
De facto evaluation standard



Annotation efforts



Measuring performance



Evaluation measure

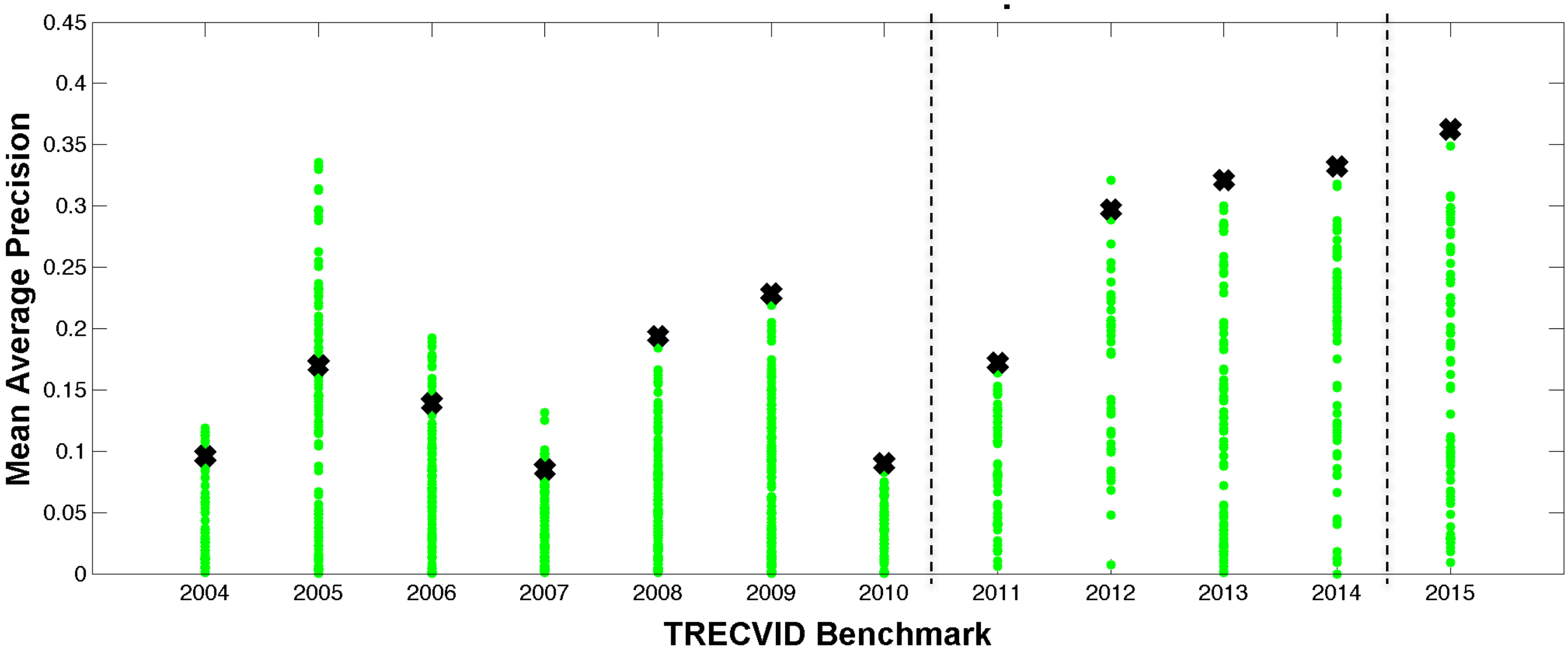
Average Precision

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

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

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

Progress in video concept search



• = 1000+ others

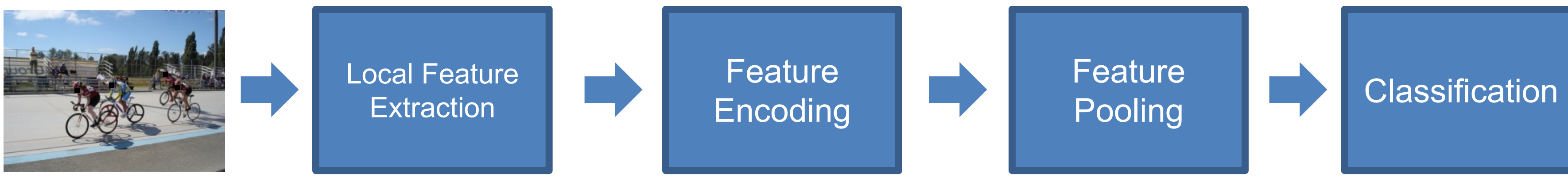
* = UvA / Euvision / Qualcomm

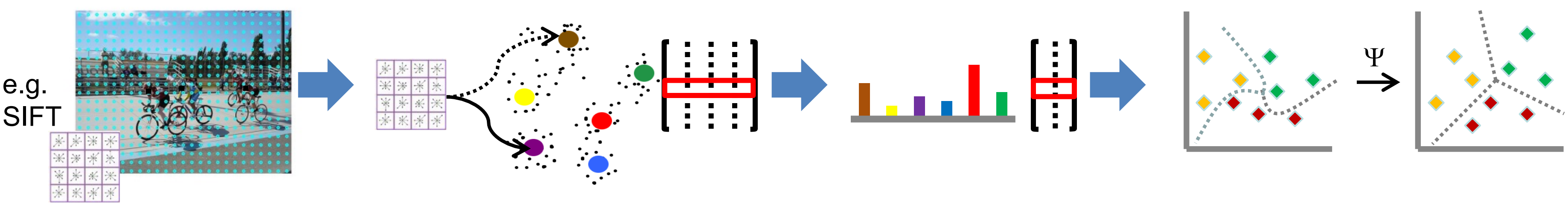
Snoek et al. TRECVID 2004-2015

2010: Bag-of-words

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

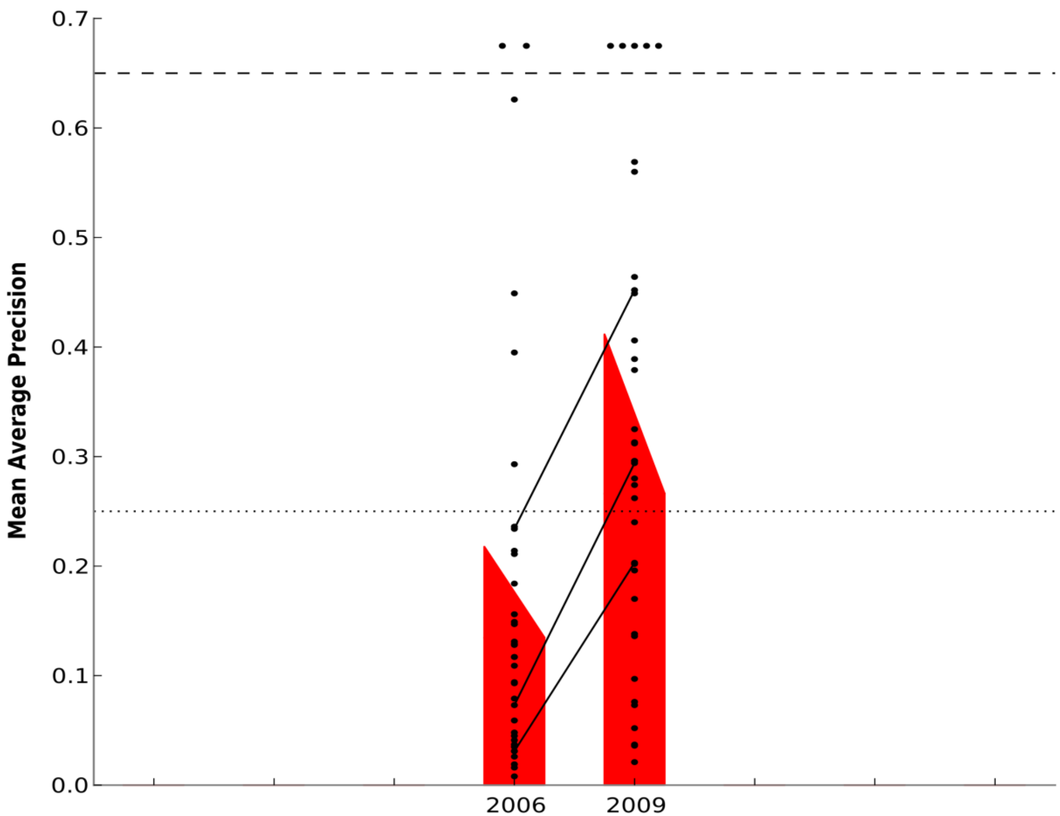
Benchmarking is compute intensive

Distributed ASCI super computer: *priceless*

Advanced School for Computing and Imaging

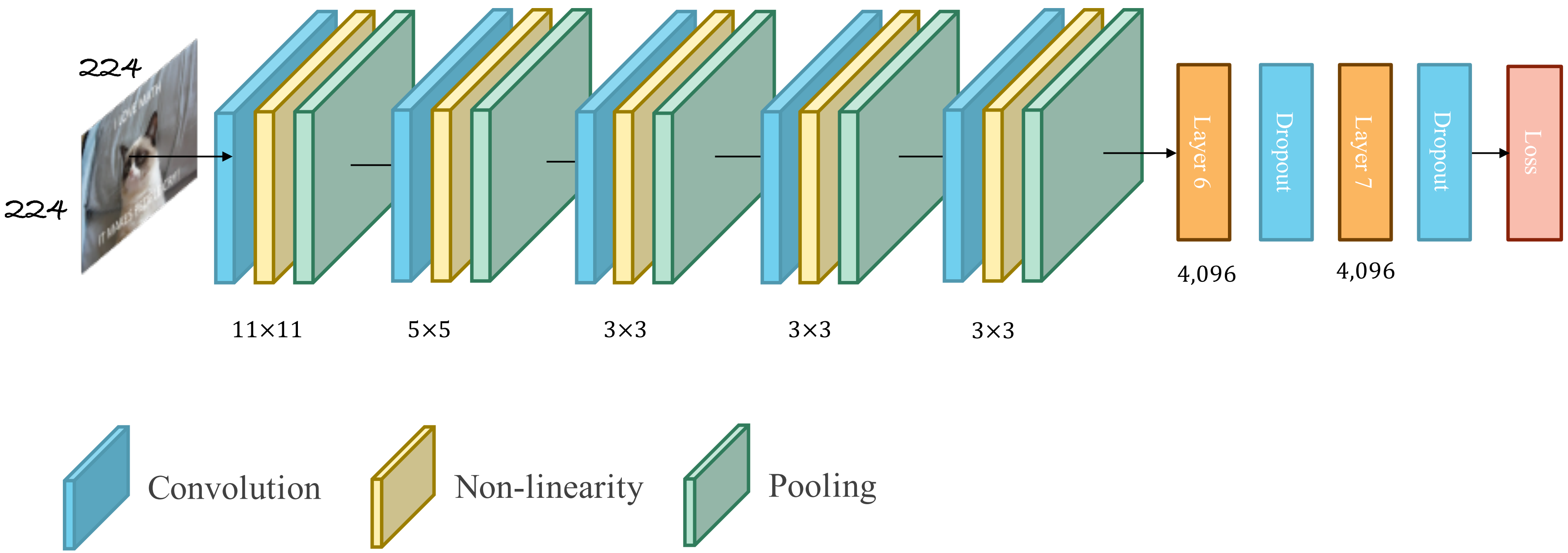
Performance doubled in 3 years

- 36 concept detectors

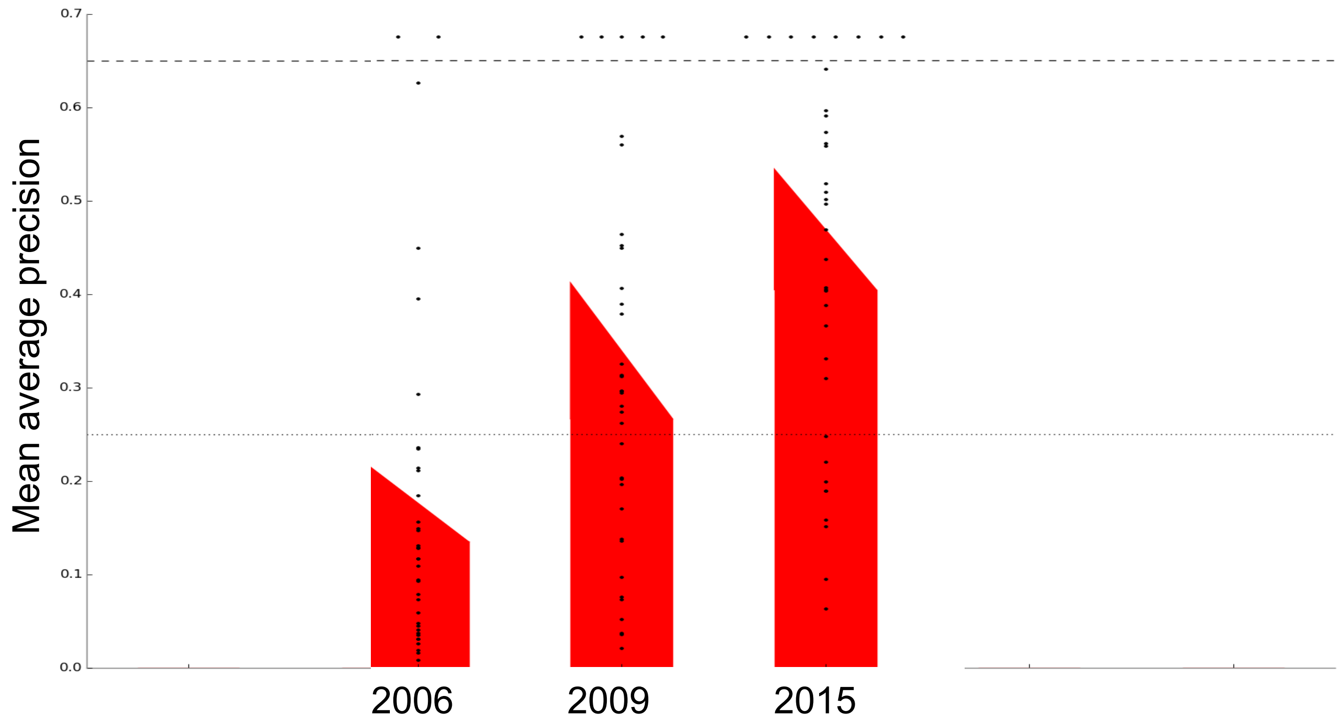


Snoek & Smeulders, IEEE Computer 2010

2013: AlexNet-variant



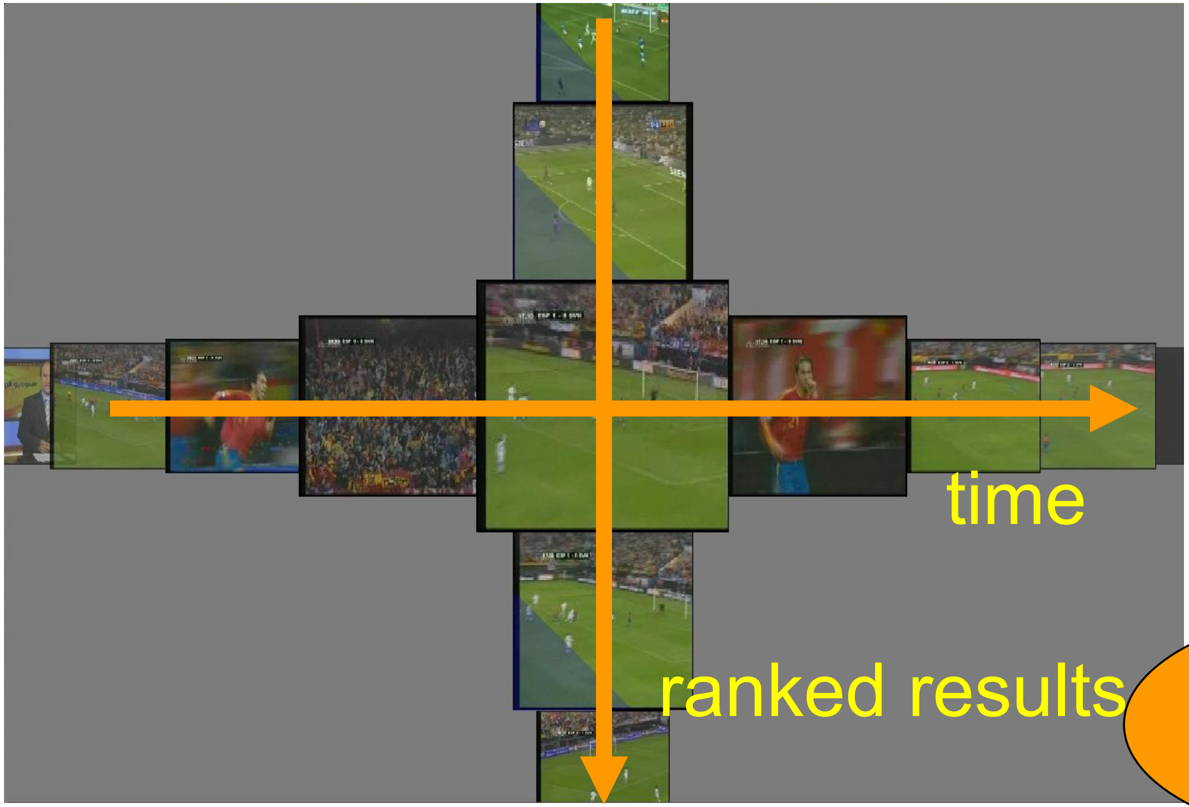
Latest jump due to deep learning



Snoek, TMM 2007

MediaMill video search engine

CrossBrowser combines query results and time

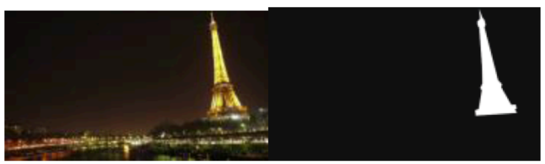


2010 Version

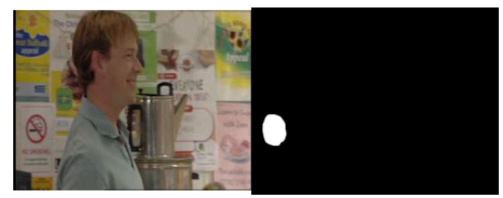
Other challenge: Instance search

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

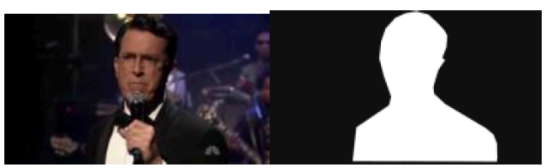
instance “Eiffel tower”



instance “a circular ‘no smoking’ logo”



instance “Stephen Colbert”



instance “an Audi logo”

instance “this man”

Other 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

ACTIVITYNET

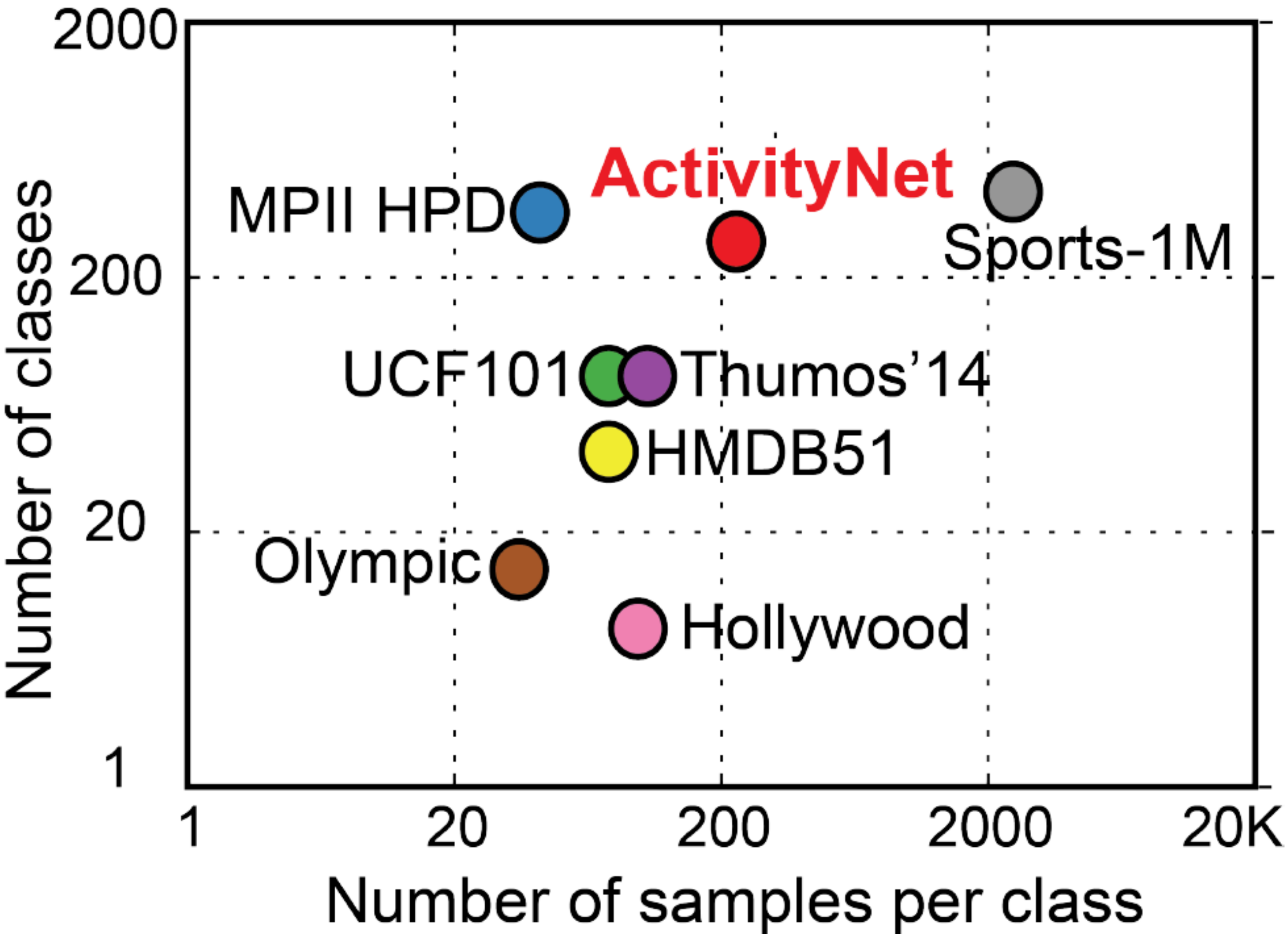
Large Scale Activity Recognition Challenge

ACTIVITYNET

Goal

Recognize all activities in daily life

ActivityNet



Challenges

- Task I: Untrimmed Video Classification

- Task II: Activity Detection

http://activity-net.org/challenges/2016/

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.

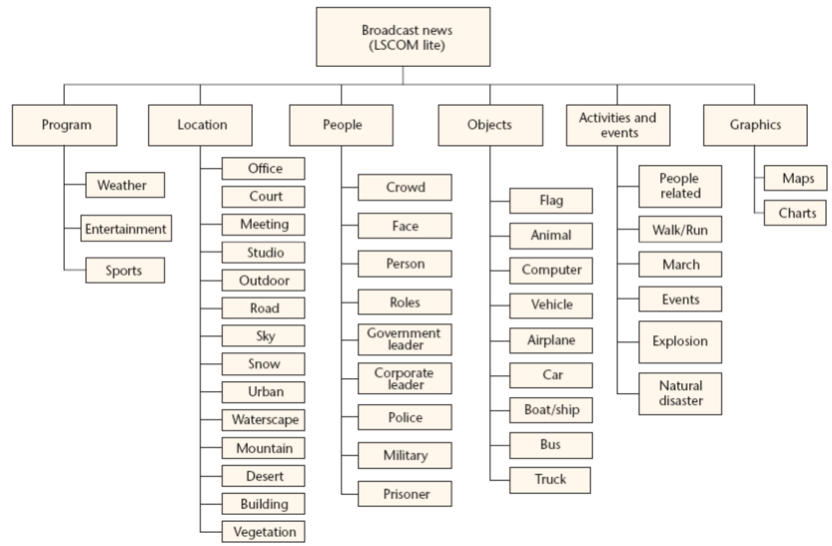
Labeling by library experts

LSCOM (Large Scale Concept Ontology for Multimedia)

Provides manual annotations for 449 concepts

– In international broadcast TV news

Connection to Cyc ontology



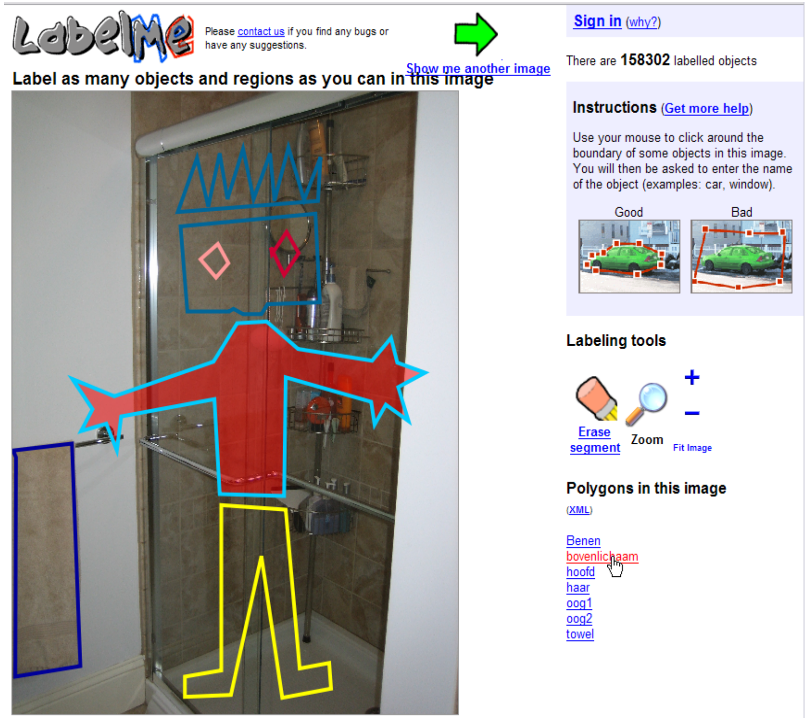
http://www.lscom.org/

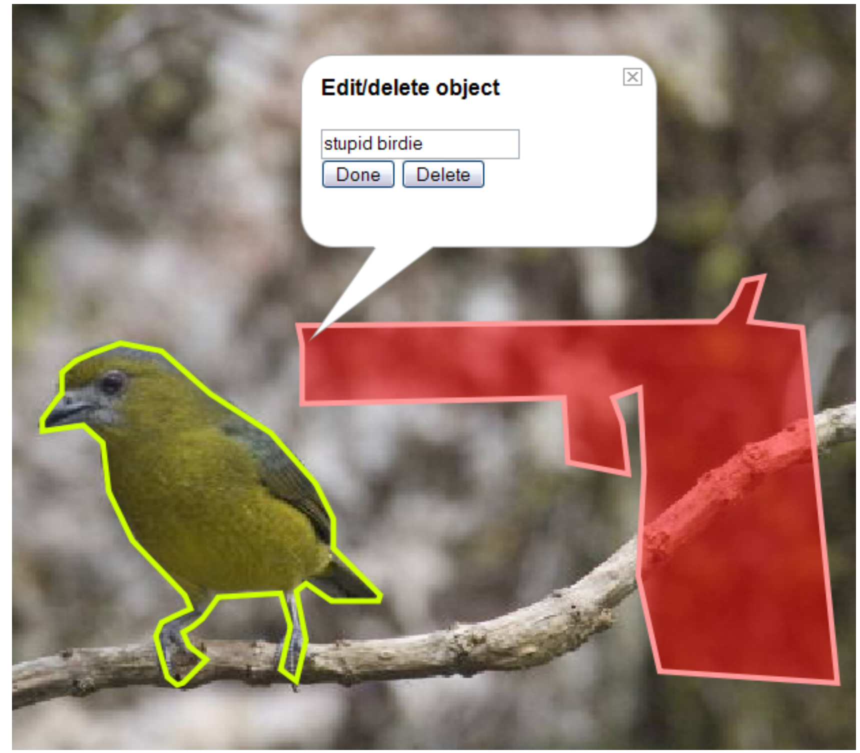
Labeling by volunteers

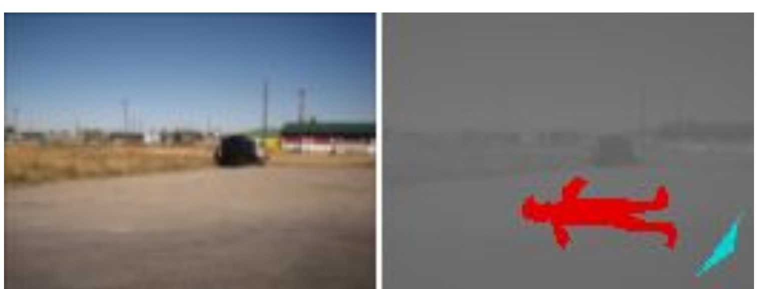
Russell IJCV 2008

Polygon quality

Online hooligans

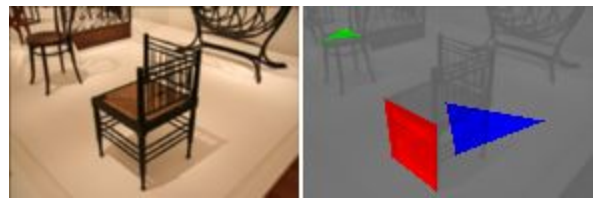






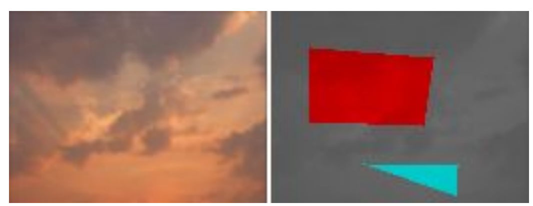
Testing

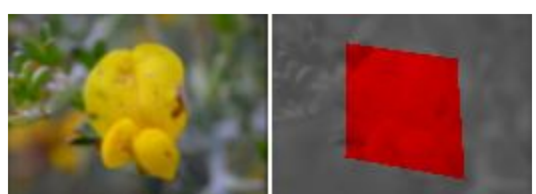
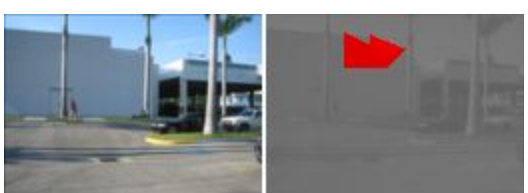








Most common labels:

test

adksdsa

woiieiie

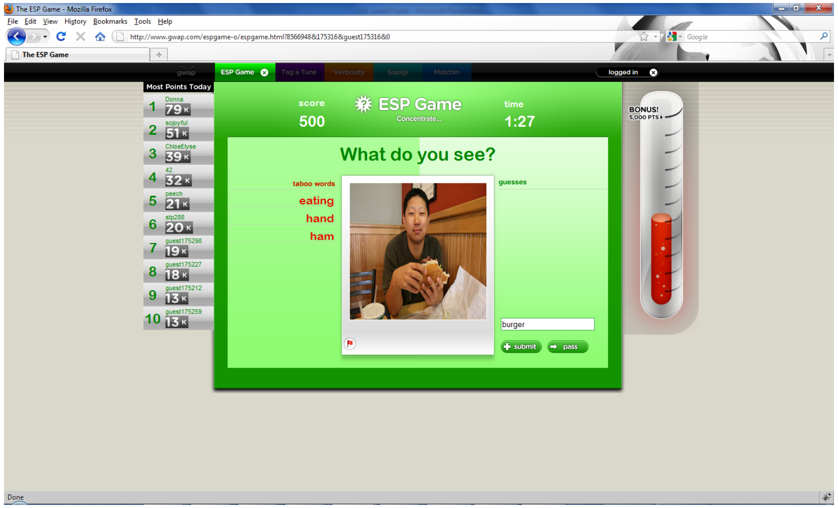
Quiz: downside of volunteers?

Lack of incentive

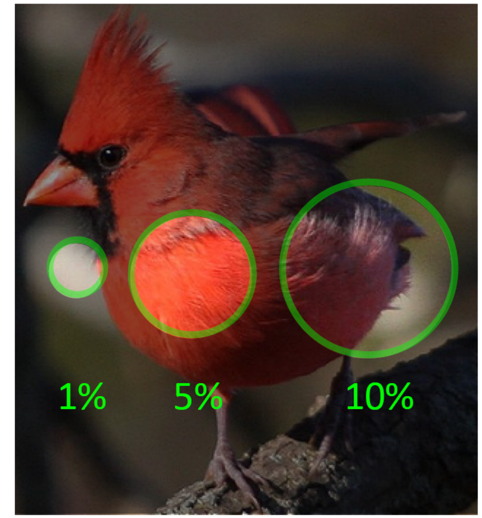
Limited quality control

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

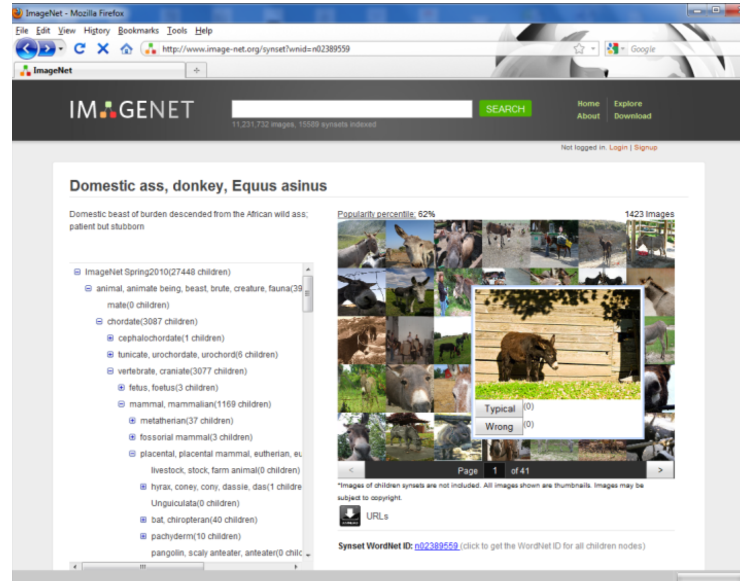
- Adding semantic structure
- Linking labels to regions

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

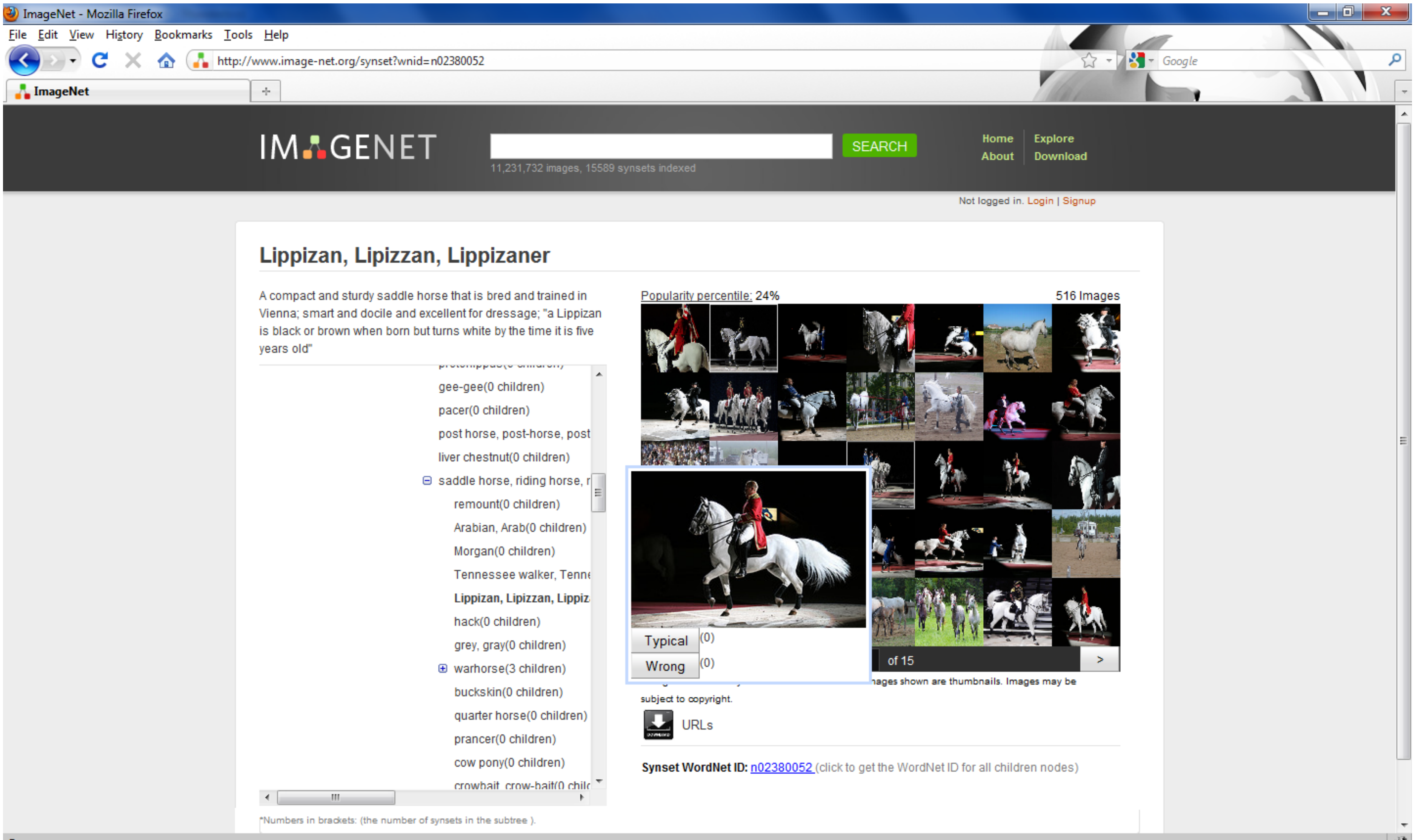


Deng et al, CVPR 2009

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

YAHOO!
picsearch
flickr
Live Search
Google Image Search

amazon mechanical turk
beta
Artificial Artificial Intelligence

IMAGENET is built by crowdsourcing

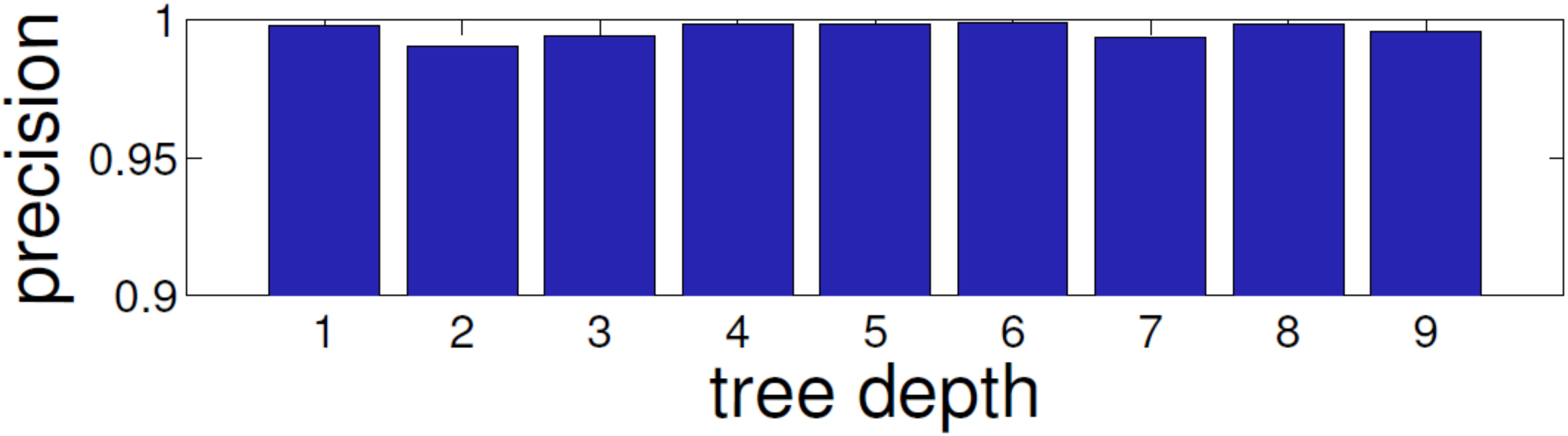
July 2008: 0 images

Dec 2008: 3 million images, 6K+ synsets

April 2010: 11 million images, 15K+ synsets

Yesterday: 14 million images, 21K synsets indexed

Accuracy



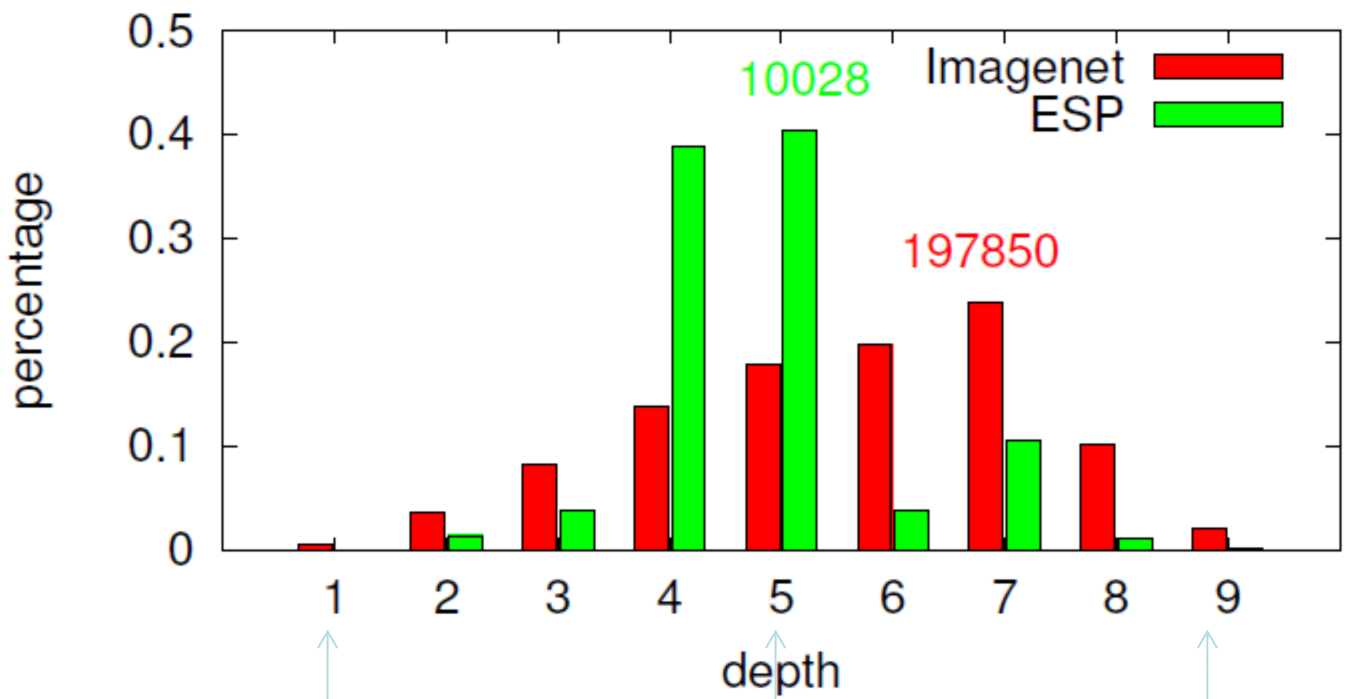
e.g. mammal

e.g. dog

e.g. German Shepherd

Deng CVPR, 2009

Diversity



e.g. mammal

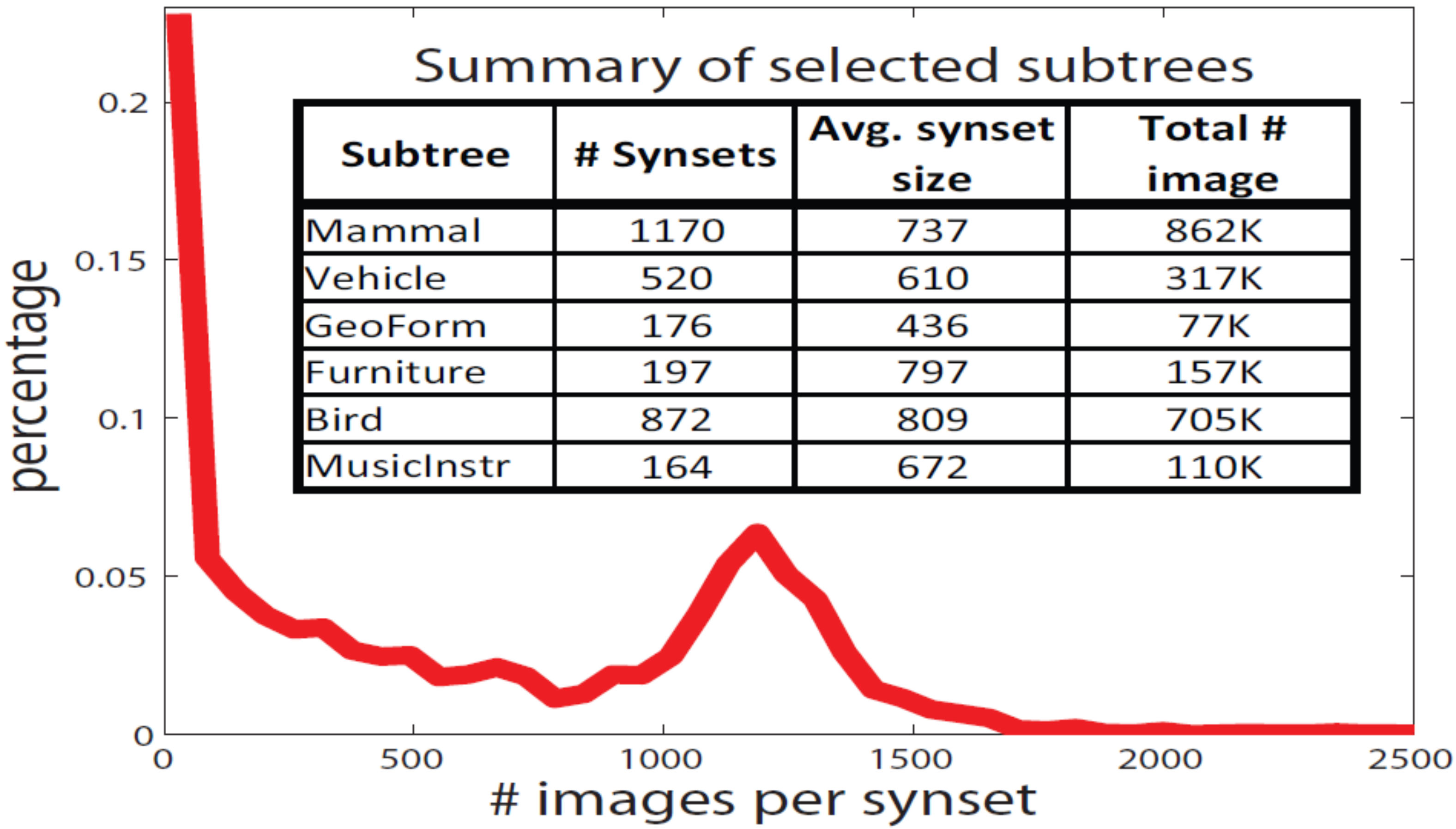
e.g. dog

e.g. German Shepherd

ESP: Ahn et al. 2006

Deng CVPR, 2009

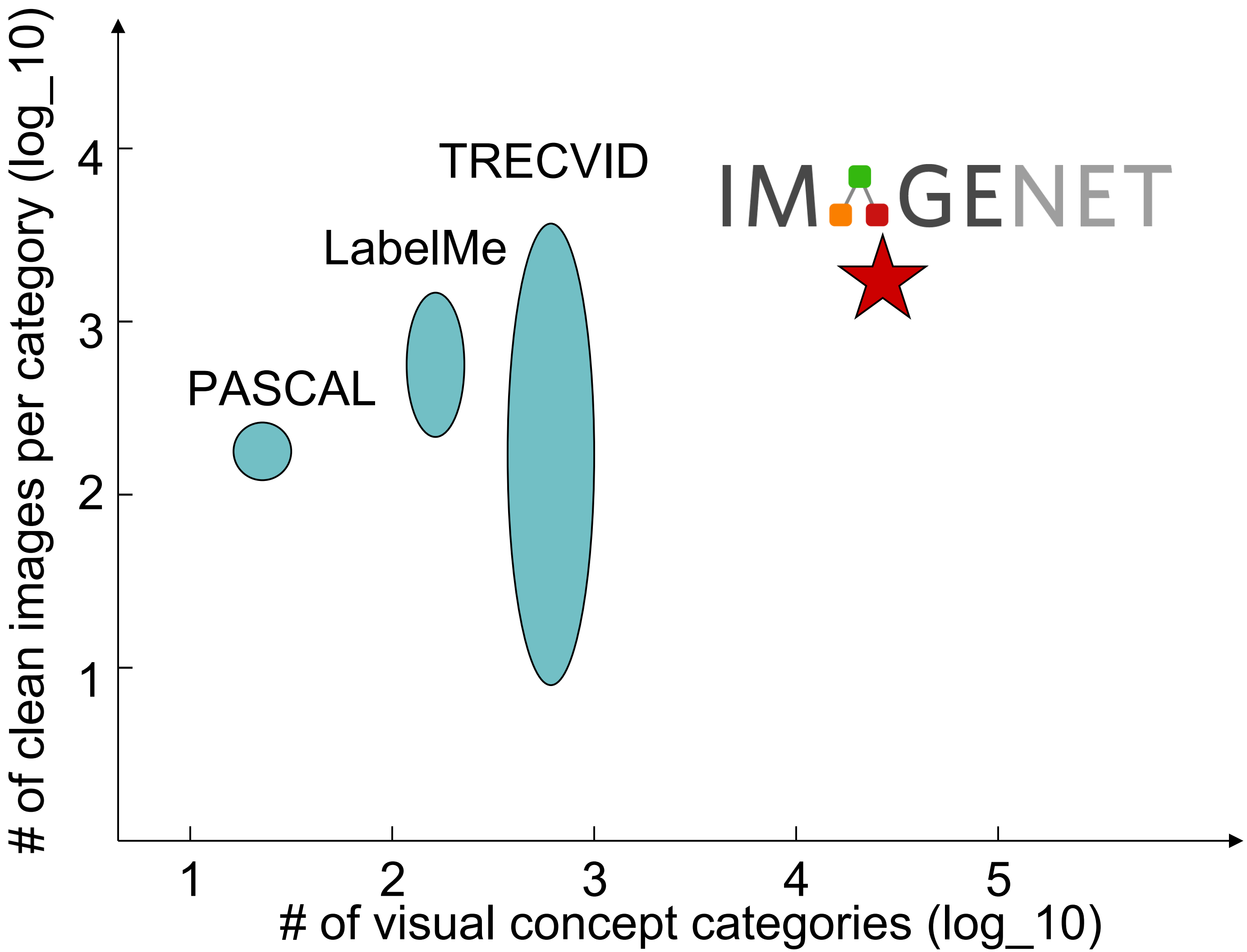
Scale



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

YAHOO!

picsearch™

flickr

Live Search

Google™ Image Search

amazon mechanical turk
beta
Artificial Artificial Intelligence

Constructing ImageNet

Free

Step 2:
Clean up the candidate
Images by humans

YAHOO!
picsearch™
flickr
Live Search
Google™ Image Search

amazon mechanical turk
beta
Artificial Artificial Intelligence

Constructing ImageNet

Free → $$$

YAHOO!

picsearch™

flickr

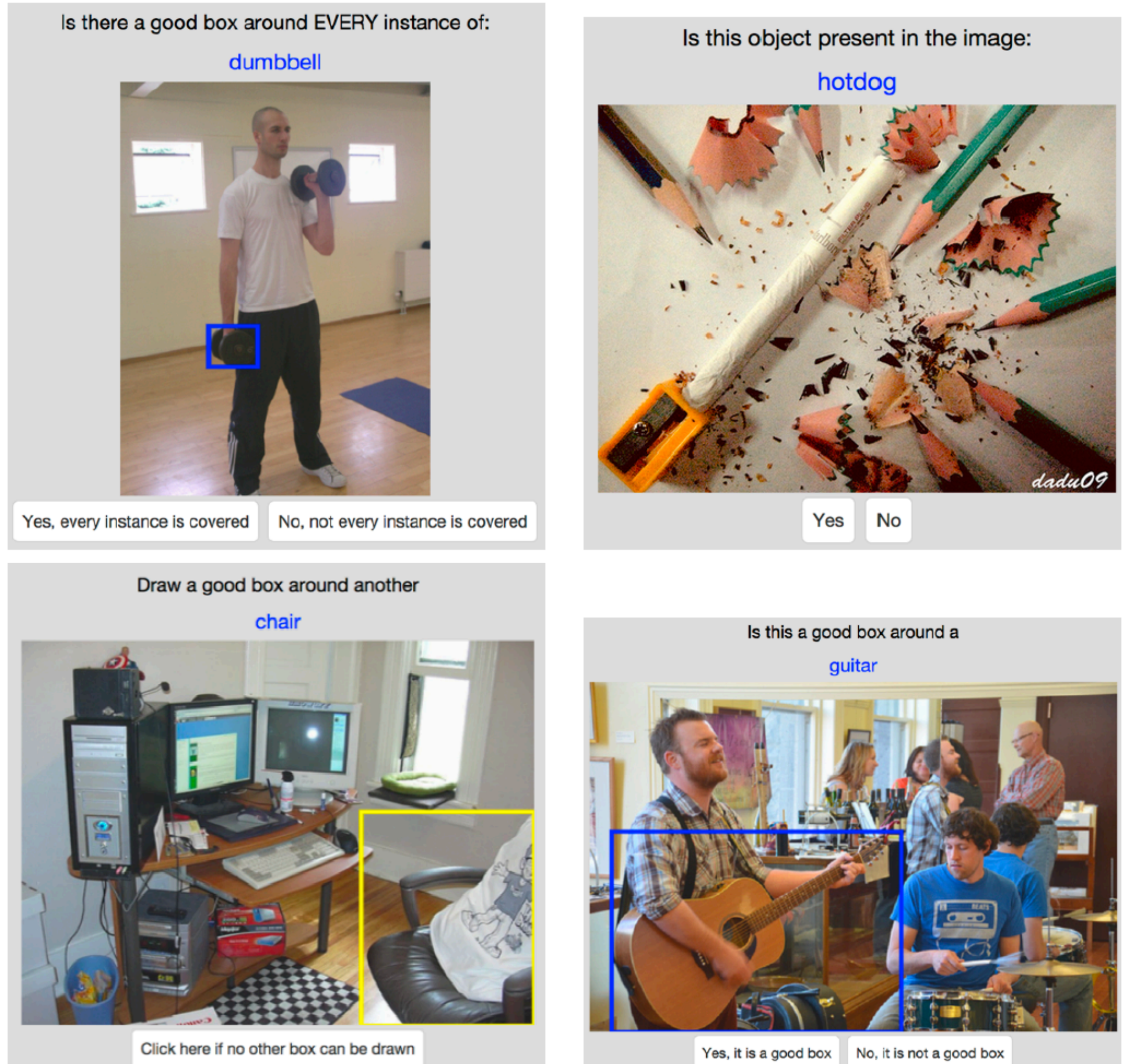
Live Search

Google™ Image Search

amazon mechanical turk
Artificial Artificial Intelligence
beta

User interfaces

For image labeling



Kovashka *et al.* FnTCGV 2016

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
2 MILLION Searches
Spotify
14 New Songs Added
DOMAINS
70 New Registered
15 THOUSAND Tracks downloaded from iTunes
WordPress
347 New Blog Posts
WEBSITES
571 New Created
Walmart
Transactions 17 THOUSAND
tumblr.
New Tumblr Photos 20 THOUSAND
EMAIL
emails sent 204 MILLION
Instagram
photos every second 3,600
Pinterest
Active Users 11 THOUSAND
LinkedIn
Professional searches 11 THOUSAND
amazon.com
Amazon Sales $83,000
snapchat
104 THOUSAND Photos Shared
twitter
278 THOUSAND Tweets
flickr
20 MILLION Photo Views
skype
1.4 MILLION Minutes Connecting with Each other
facebook
41 THOUSAND posts every second
1.8 MILLION likes
350GB of data
YouTube
72 HOURS of Video uploaded
55
50
45
40
35
30
25
20
15
10
5

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’

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
...

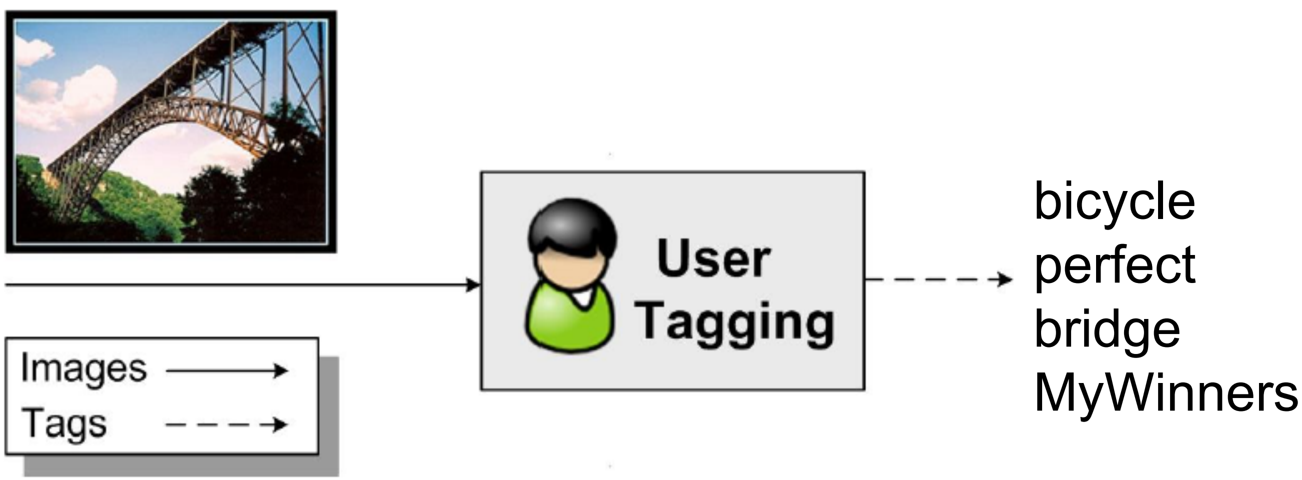
view details

tour
tampere
church
upload
...

view details

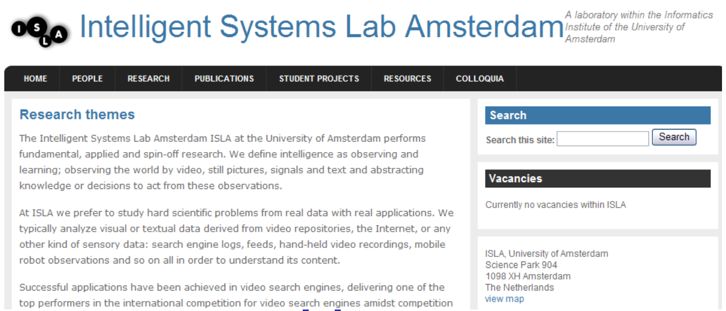
Quiz

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



Computer vision is essential

Free text



User tags

bridge
bicycle
perfect
MyWinners

?

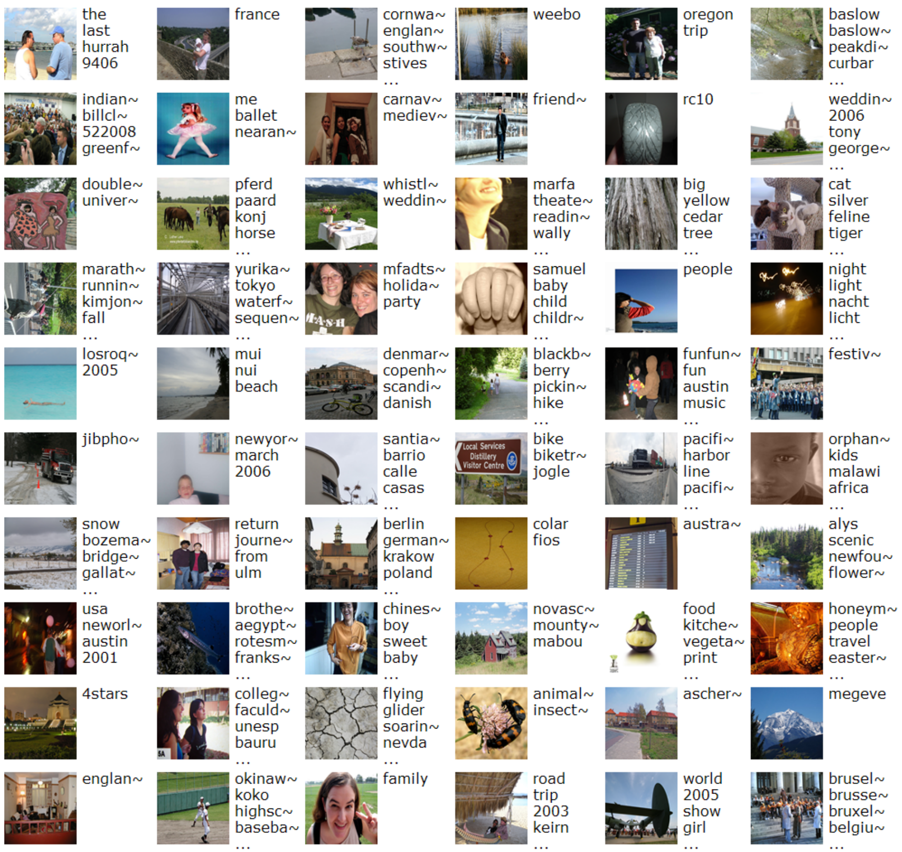
bridge
bicycle
perfect
MyWinners

Challenges

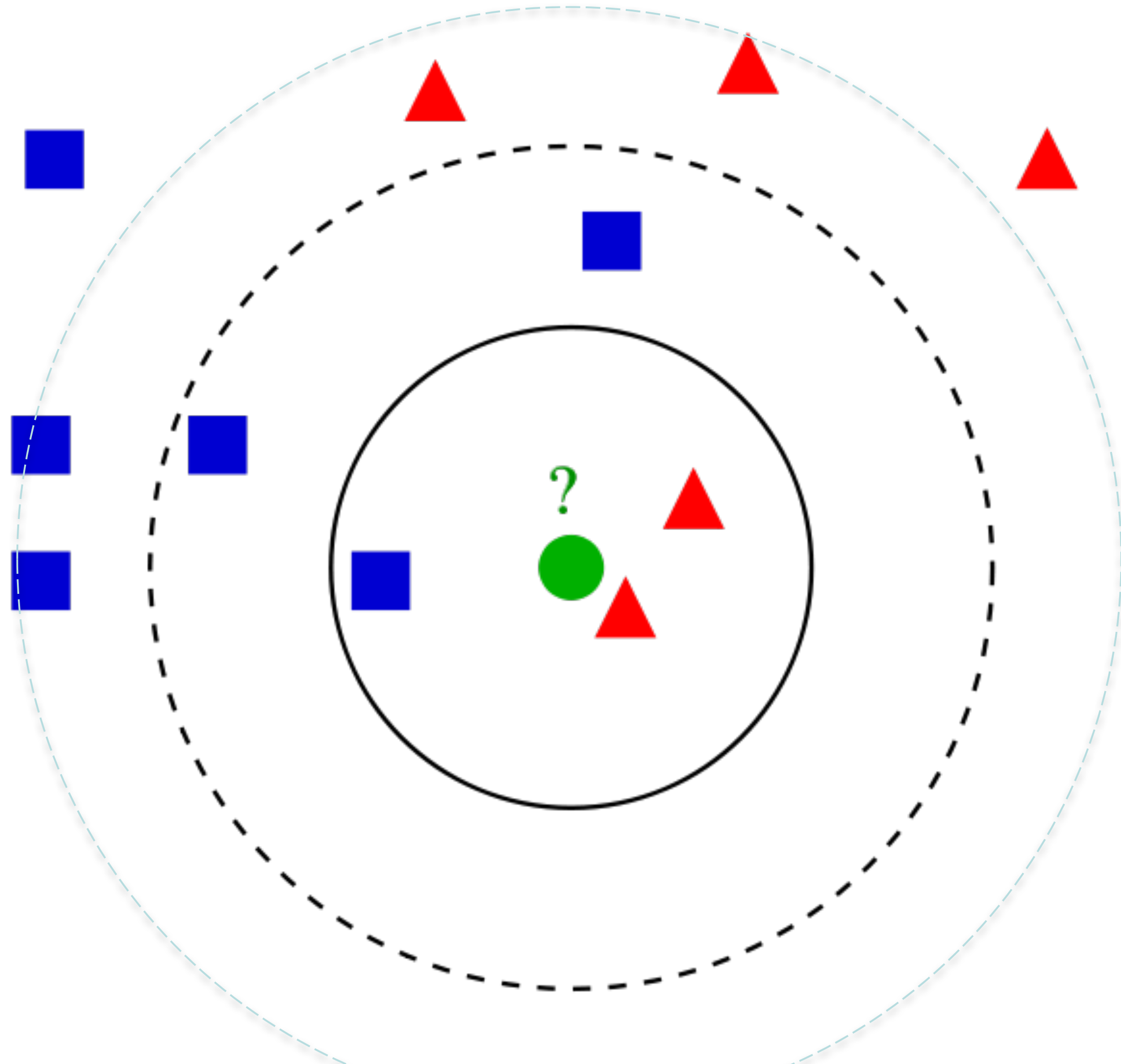
Many tags & many images

A prospective algorithm

- scalable
- unsupervised

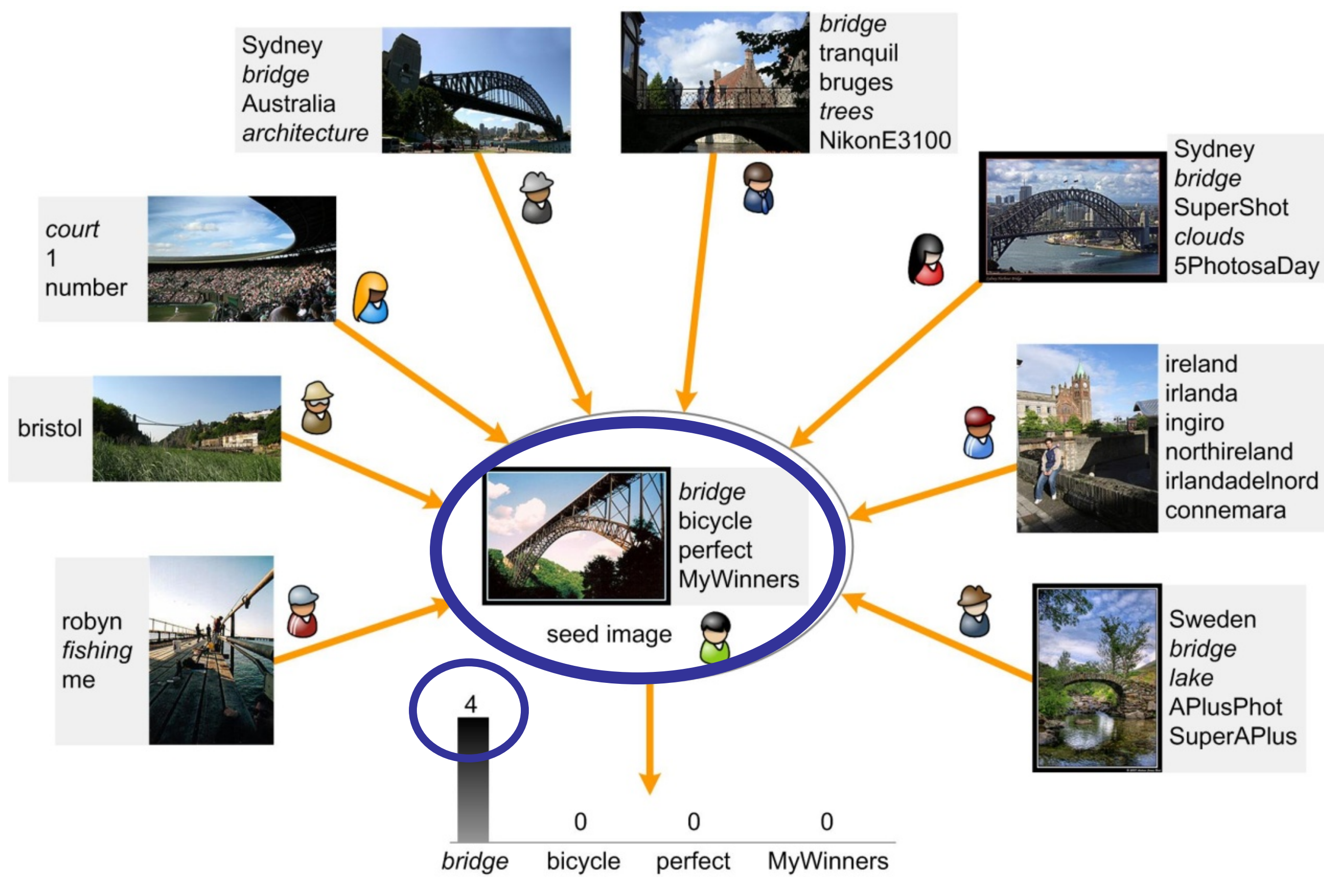


Nearest neighbor



Intuition for tagged images

Similar images with similar tags are reliable



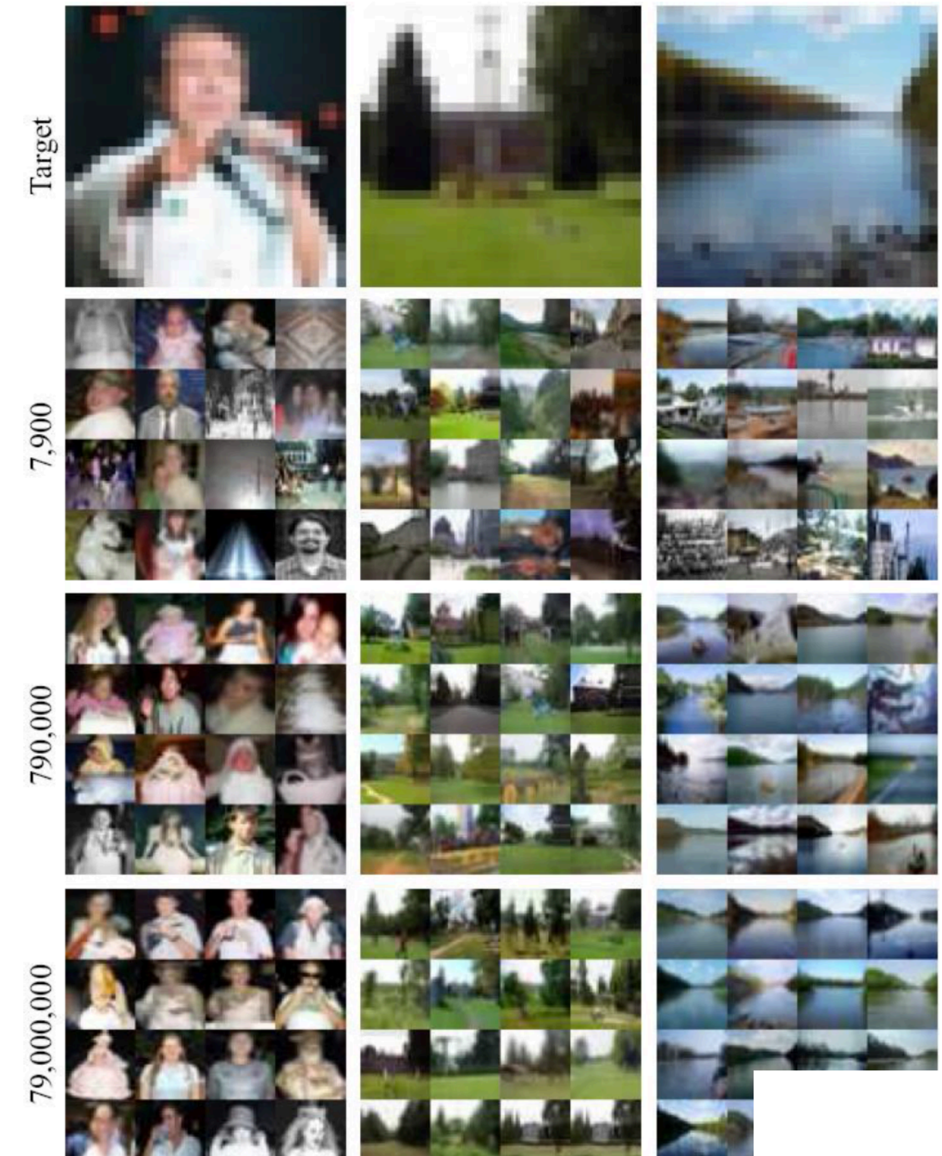
Xirong Li, TMM 2009, best paper

Nearest neighbor for tag relevance

Objective tags are identified

Based on 3.5 Million images downloaded from Flickr

Even more efficient with tiny images



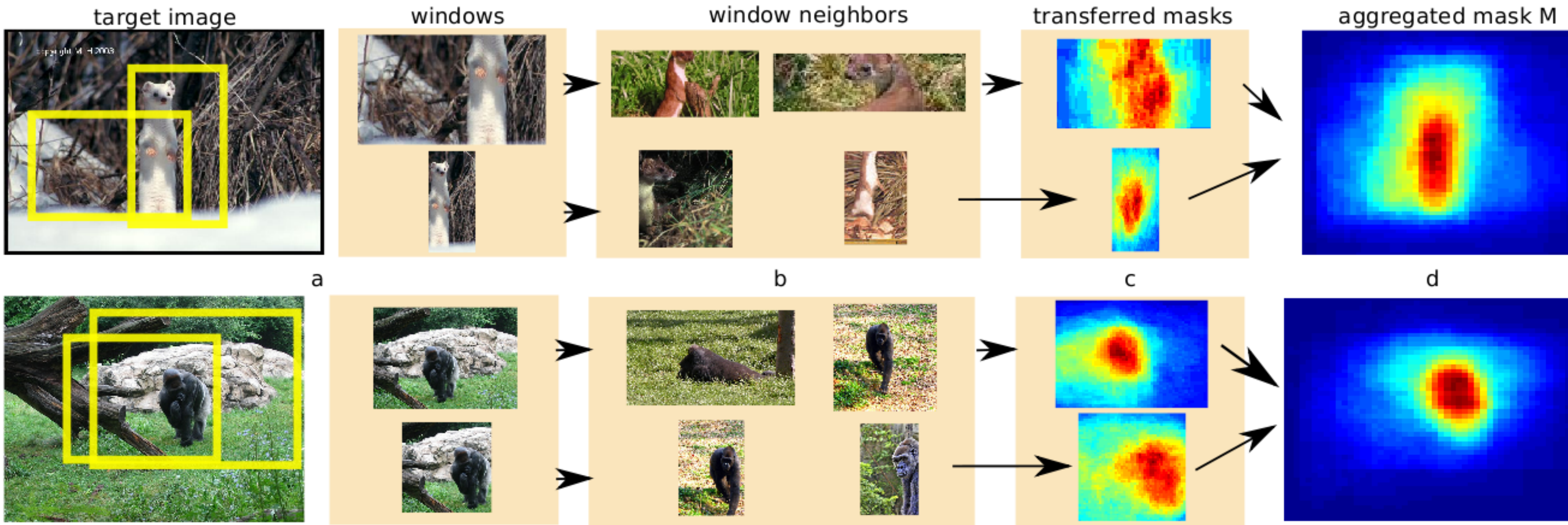
32x32 resolution

80M images

Nearest neighbor

Torralba, PAMI 2008

Nearest neighbor for segments



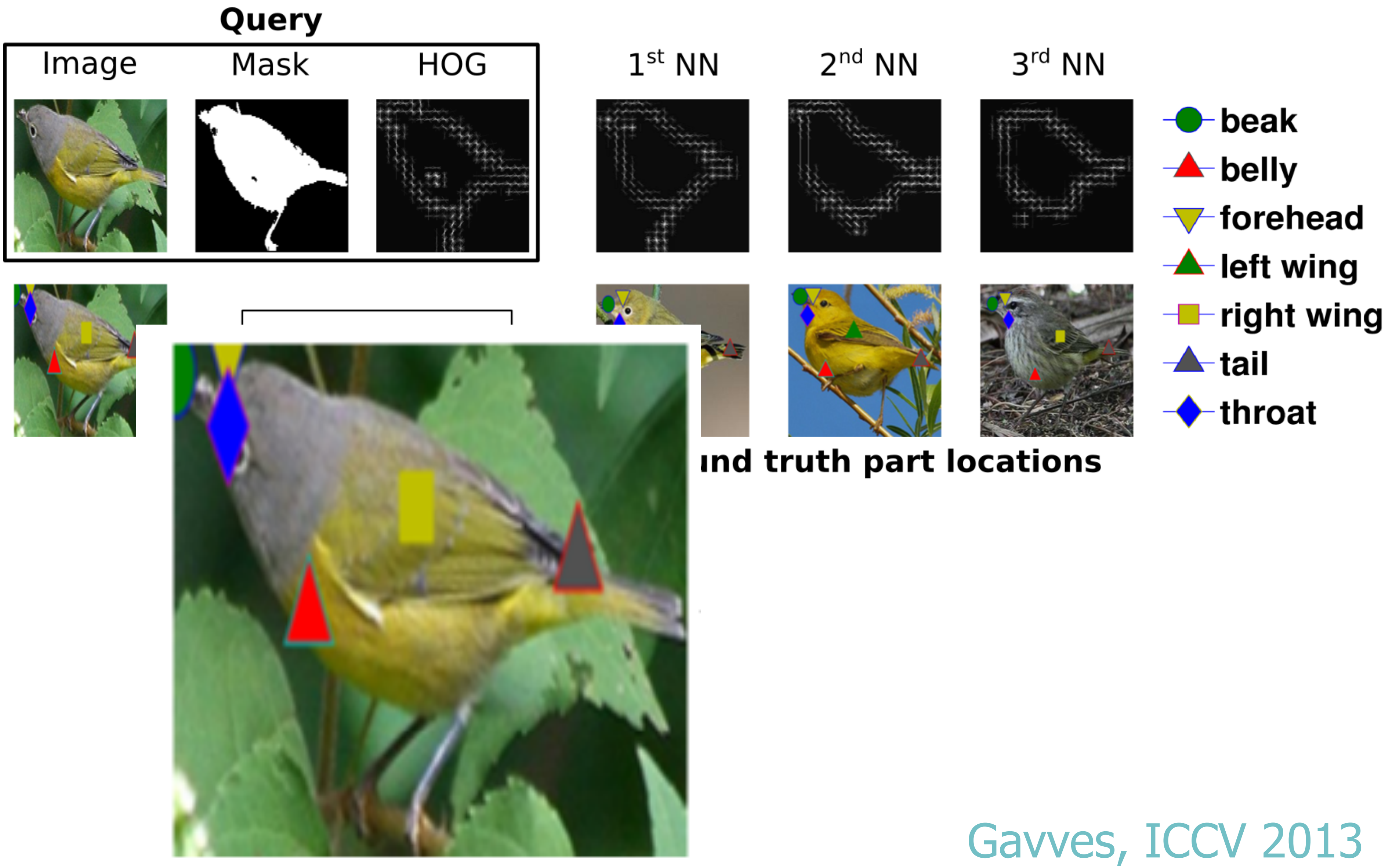
Annotates many classes with accurate segmentations

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. Weakly-supervised vision

In this Chapter we consider computer vision by weakly-supervised learning. In such scenarios some limited supervision is available at train time, typically an object or action class label. The goal is then to enrich this label, for example by predicting bounding boxes, segments or spatio-temporal tubes.

Weakly-supervised object detection

Typically casted as Multiple Instance Learning problem

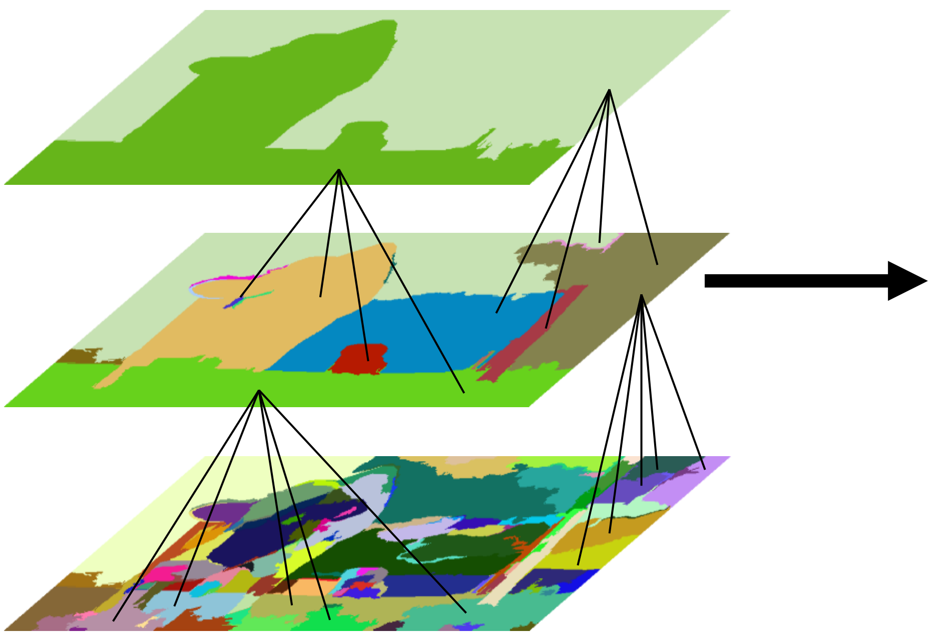
Each image is considered as a “bag” of examples given by object proposals.

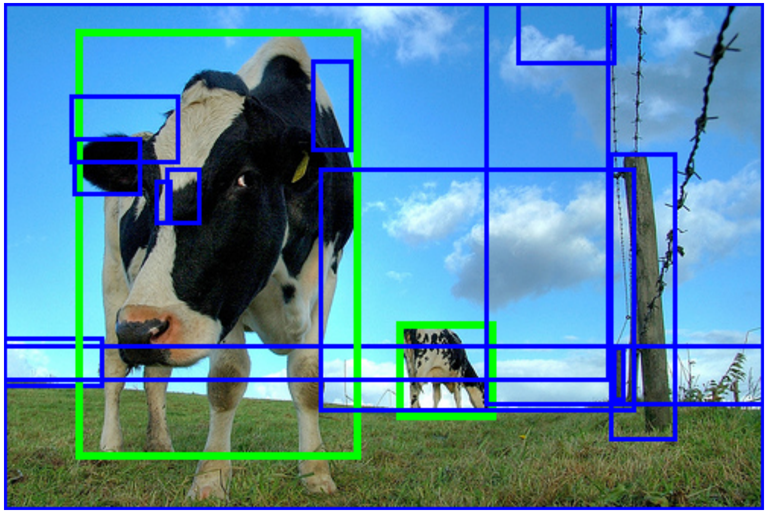
Positive images are assumed to contain at least one positive object proposal

The object detector is obtained by alternating detector training, and using the detector to select the single most likely object instance in each positive image.

Object proposals

Hypotheses from hierarchical grouping of super-pixels





Uijlings, IJCV 2013

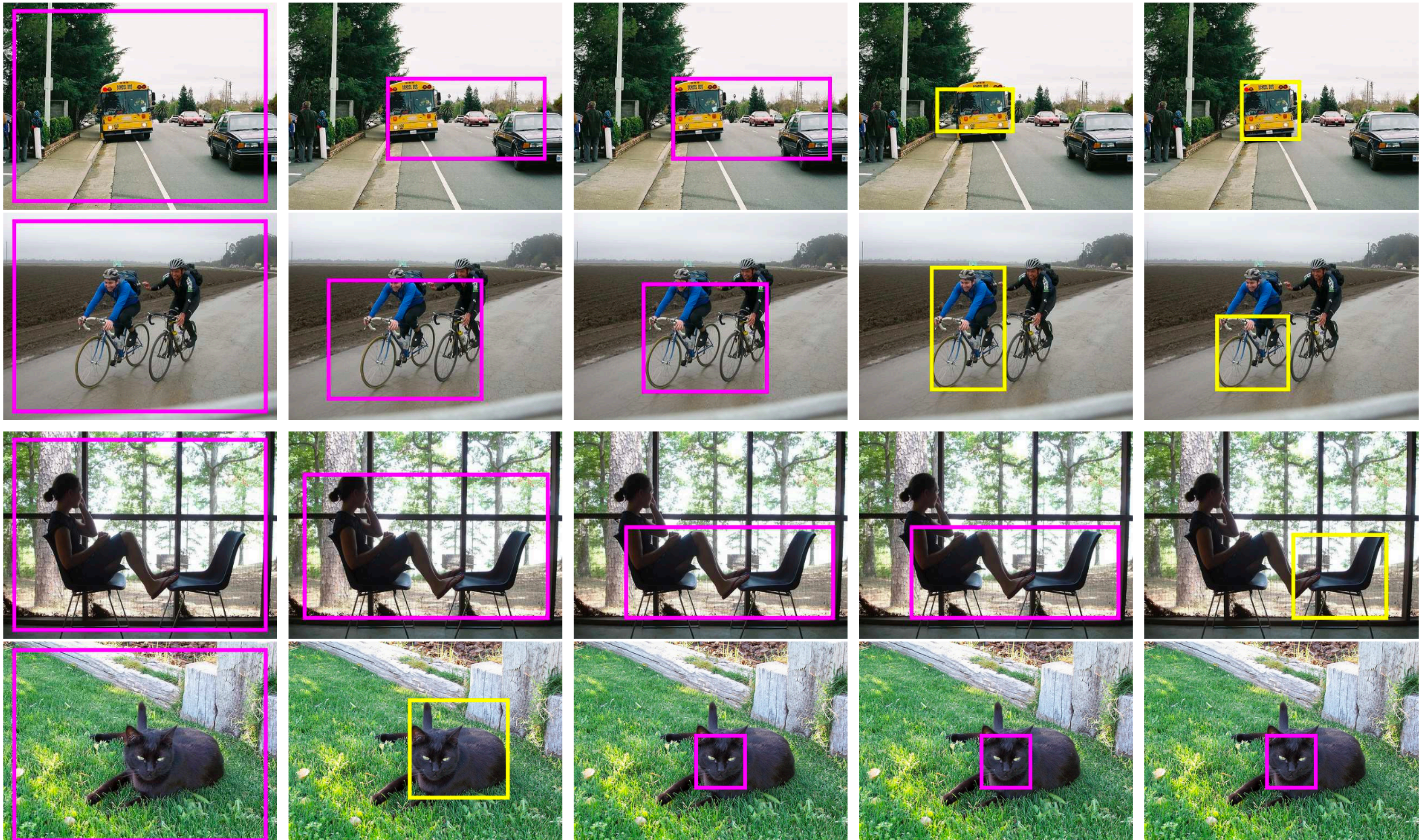
Multi-fold multiple instance learning

Algorithm 1 — Multi-fold weakly supervised training

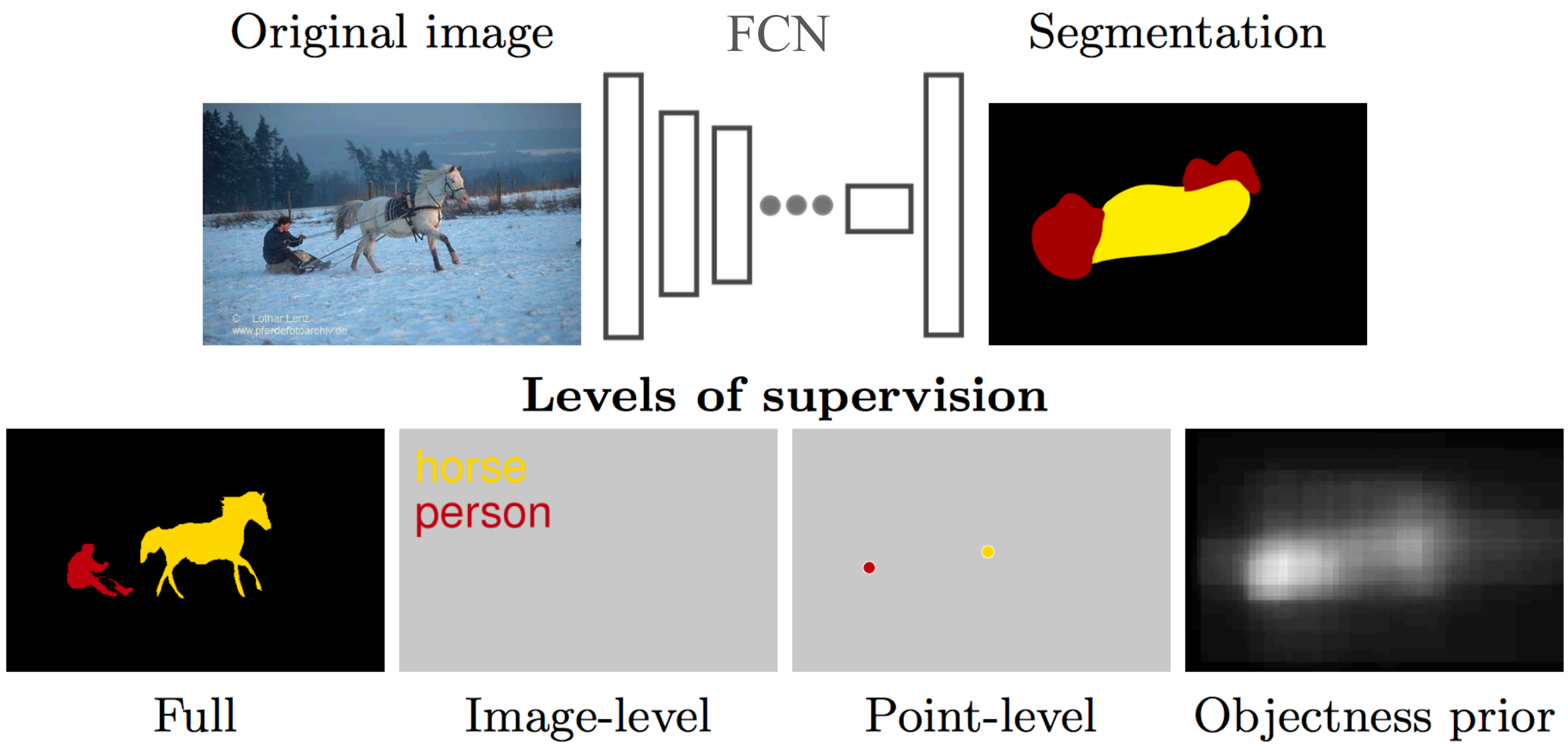
1. Initialization: positive and negative windows are set to entire images up to a 4% border.
2. For iteration $t = 1$ to T
   (a) Divide positive images randomly into K folds.
   (b) For $k = 1$ to K
      i. Train using positives in all folds but k.
      ii. Re-localize positives in fold k using this detector.
   (c) Train detector using positive windows from all folds.
   (d) Perform hard-negative mining using this detector.
3. Return final detector and object windows in train data.

Cinbis *et al.* CVPR 2014 / PAMI 2017

Re-localization process



Weakly-supervised segmentation



Adapt loss function depending on supervision scheme

Bearman *et al.* ECCV 2016

Crowdsourcing point annotations

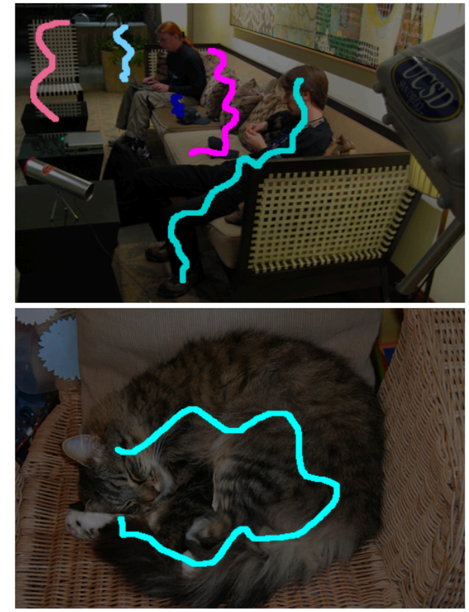


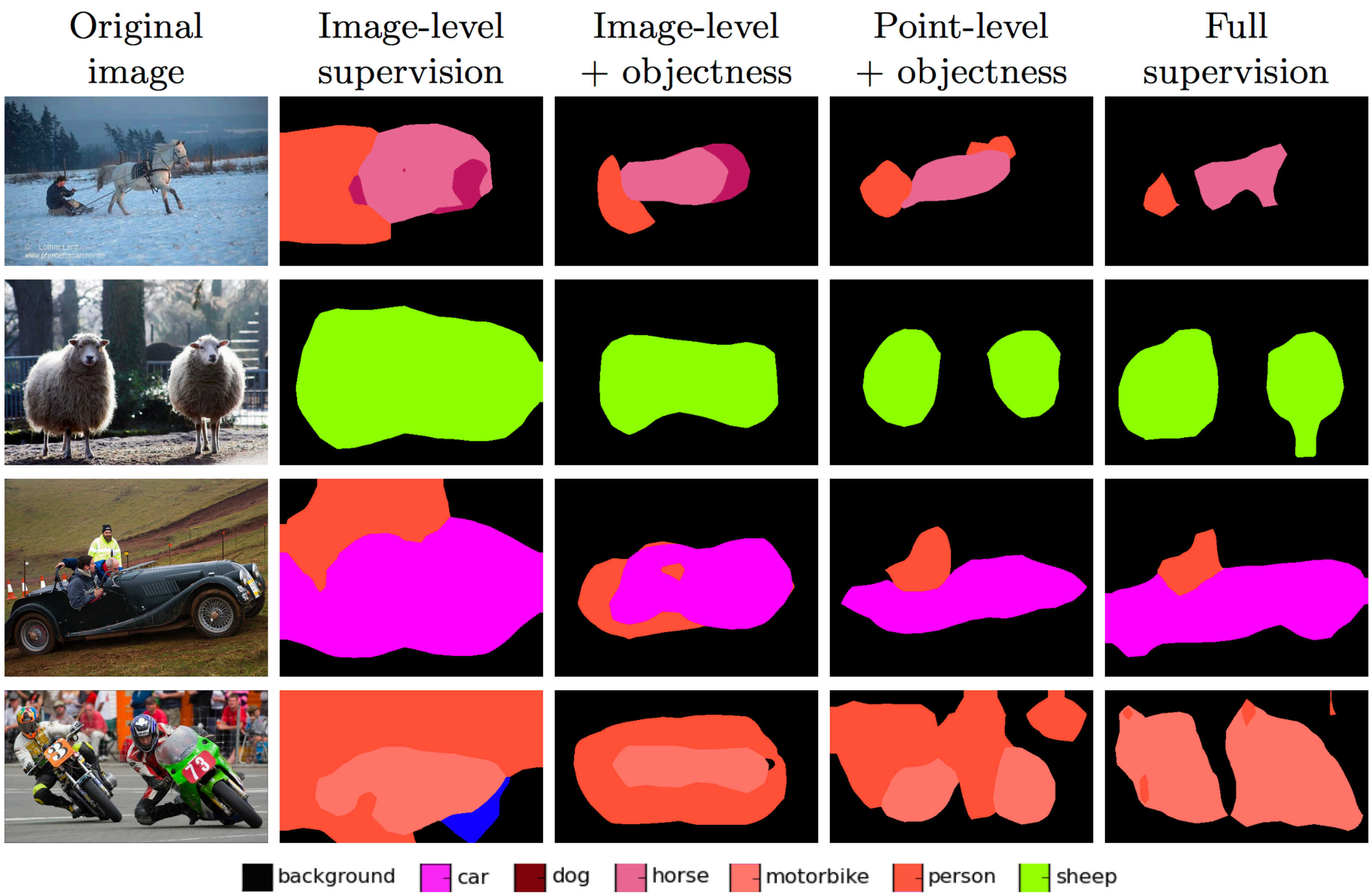
Image-level labels: 20.0 sec/image

Points: **22.1** sec/image

Squiggles: 34.9 sec/image

Full supervision: **239.7** sec/image

Some results

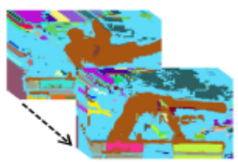


Bearman *et al.* ECCV 2016

Recap: Action proposals

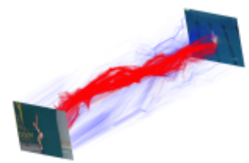
Supervoxels

Jain et al. *CVPR'14*
Oneata et al. *ECCV'14*



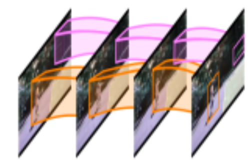
Trajectories

van Gemert et al. *BMVC'15*
Puskas et al. *ICCV'15*

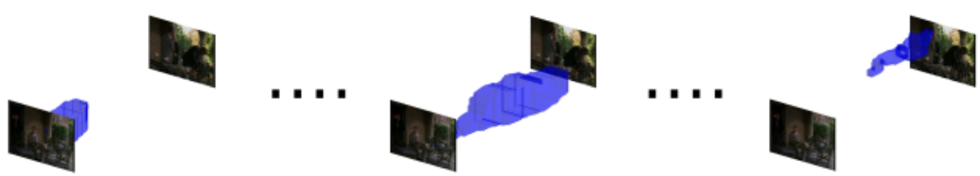


Detect & Track

Yu et al. *CVPR'15*
Weinzaepfel et al. *ICCV'15*



Action proposals



Action localization with proposals

At train time

Annotate spatiotemporal tubes with class labels

Extract video representation from tubes

Train favorite classifier

At test time

Extract action proposals

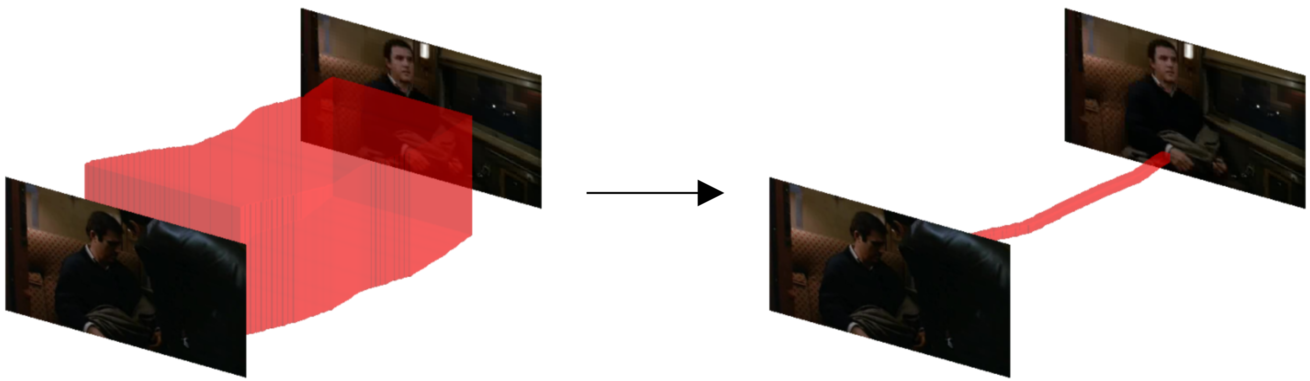
Extract video representation from each proposal

Classify all proposals, select proposal with maximum response

Hypothesis

Training on bounding boxes not required.

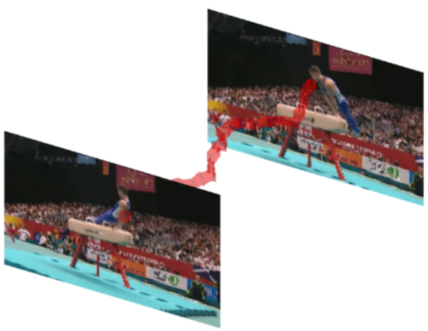
Training on proposals with fast point annotations is as effective.



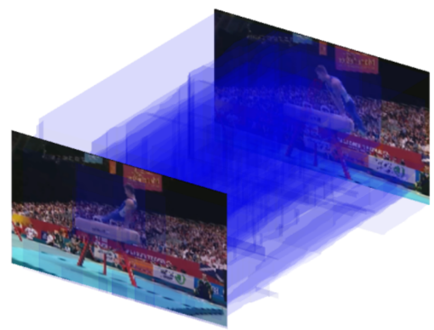
Annotation time for video:
5 min. 11 sec.

Annotation time for video:
25 sec.

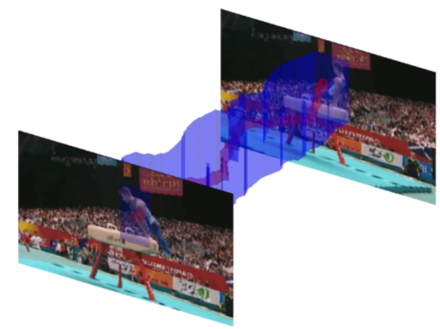
Idea



Human point supervision



Compute proposal affinity



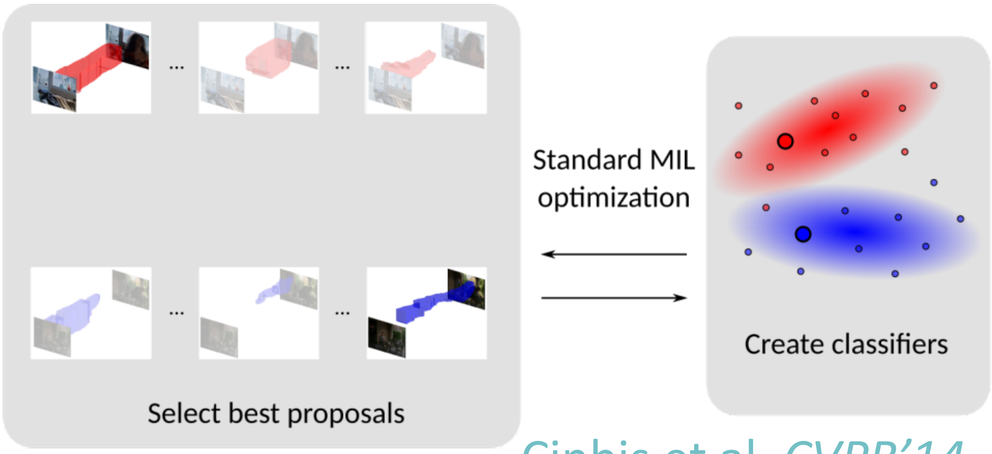
Mine best proposal

Mettes *et al.* ECCV 2016

Mining the best proposal

Train action classifiers using best proposals only.

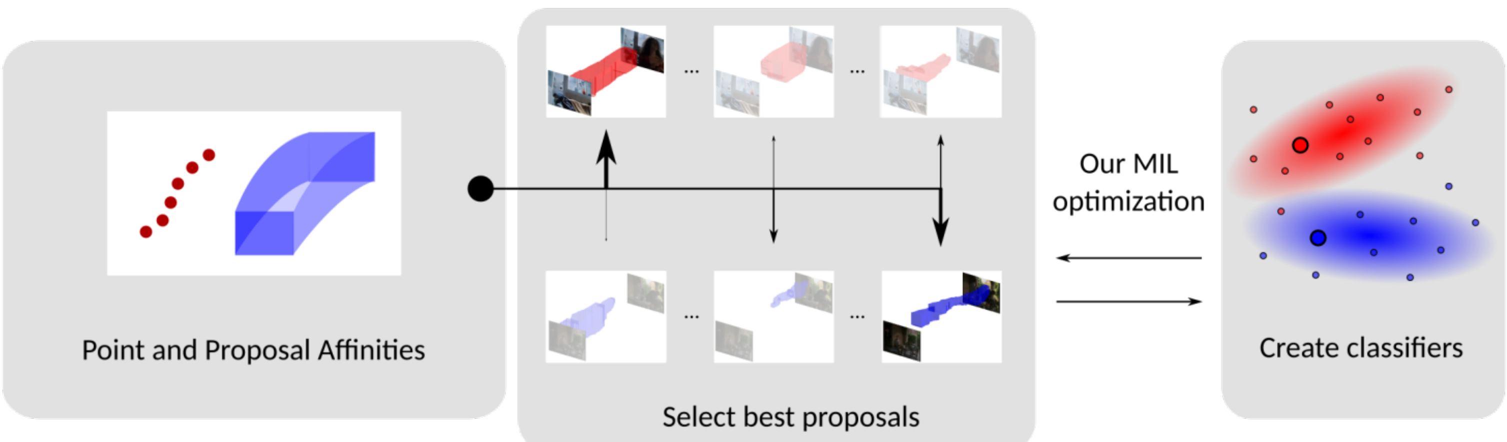
Casted as a Multiple Instance Learning problem.



Cinbis et al. *CVPR'14*

Mining the best proposal

Train action classifiers using best proposals only.
Casted as a Multiple Instance Learning problem.

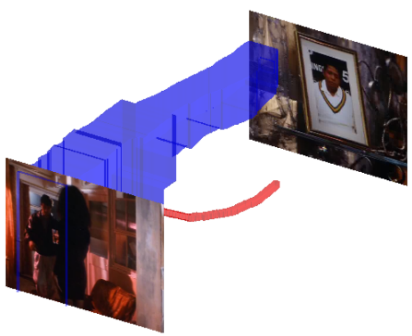


Use affinity with point annotations to guide the mining.

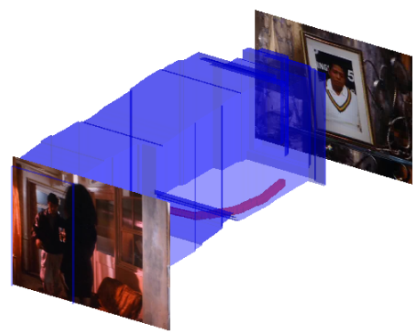
Mettes *et al.* ECCV 2016

Proposal affinity

Novel overlap measure between point annotations and proposals.



No overlap

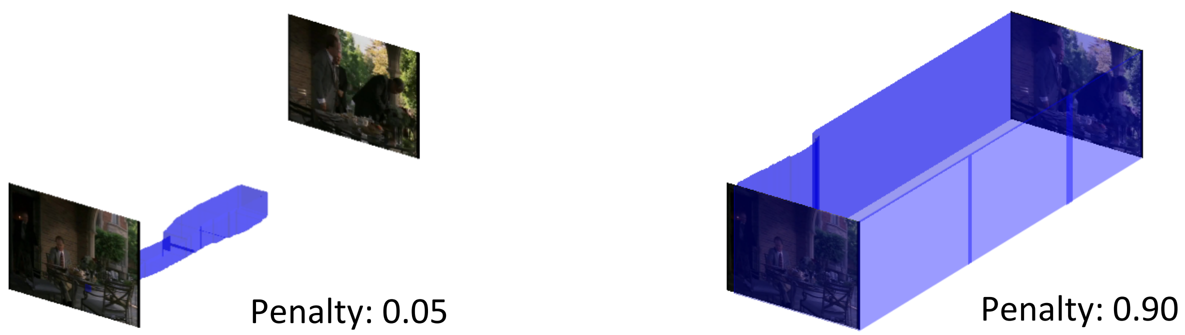


Small overlap

High overlap

Mind the center bias

Subtract the size of the proposal from the match.
To alleviate center bias of large proposals.



Experiments

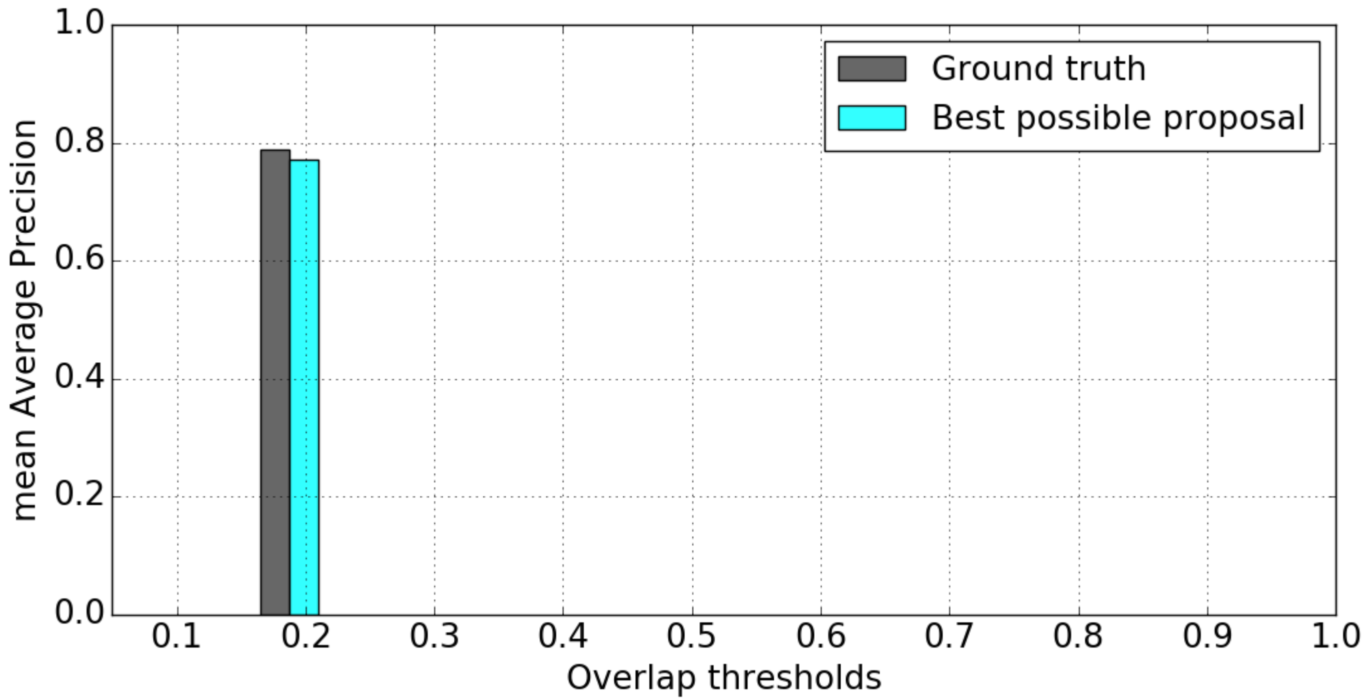
UCF Sports

UCF 101 (in paper)

Unsupervised proposals from clustered trajectory features.
Evaluated with Fisher Vectors and SVMs.

van Gemert *et al.* BMVC'15

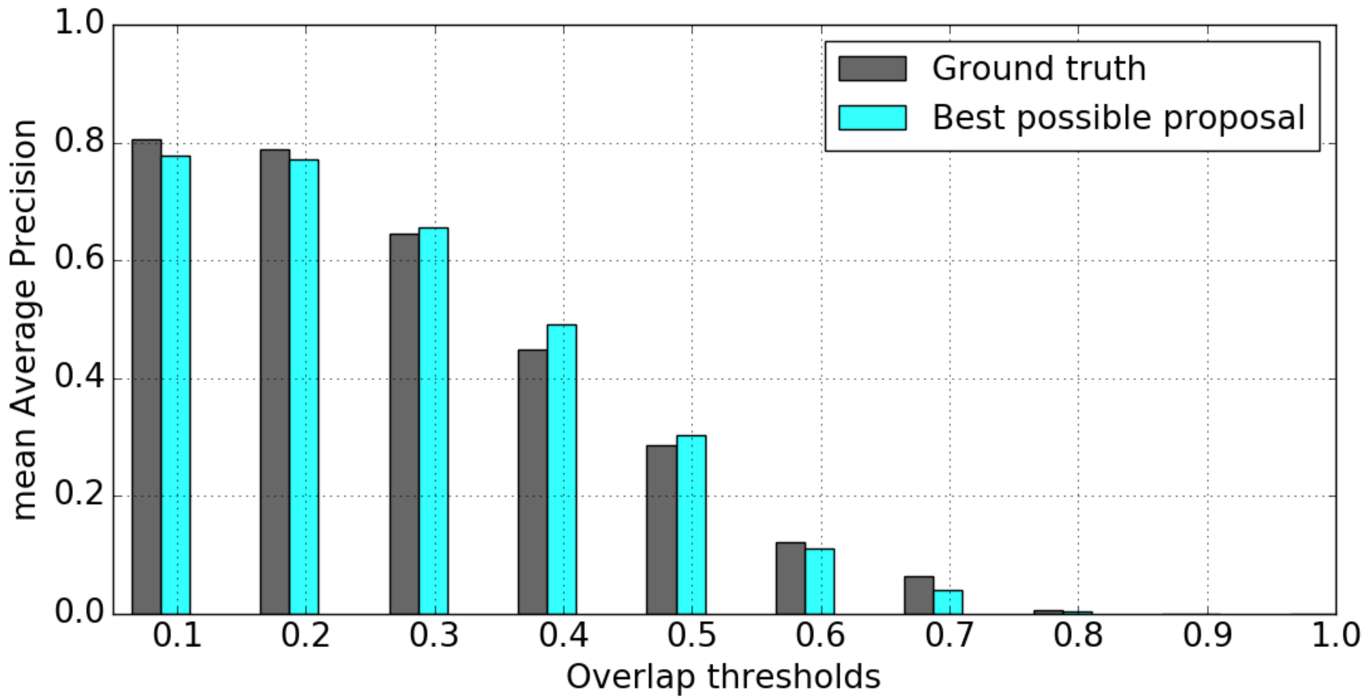
Training without ground truth boxes



UCF Sports

Best possible proposal performs as well as ground truth boxes.

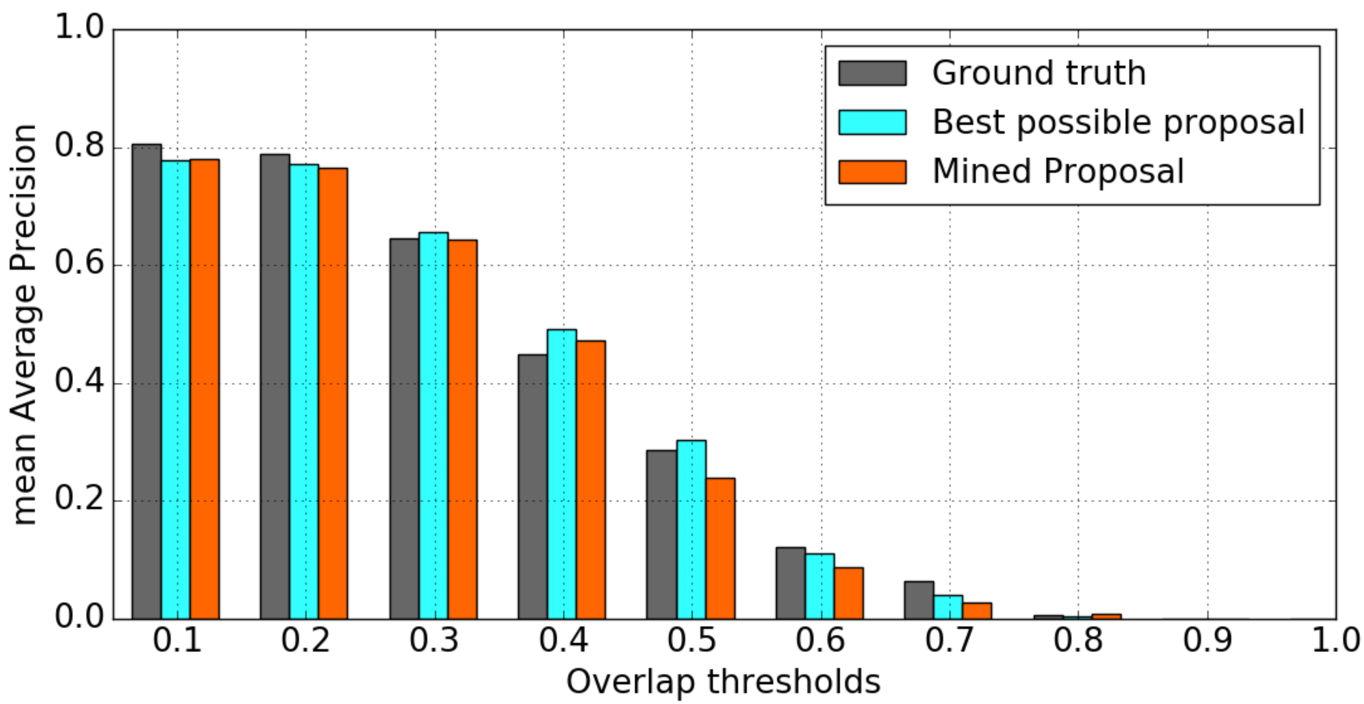
Training without ground truth boxes



UCF Sports

Best possible proposal performs as well as ground truth boxes.

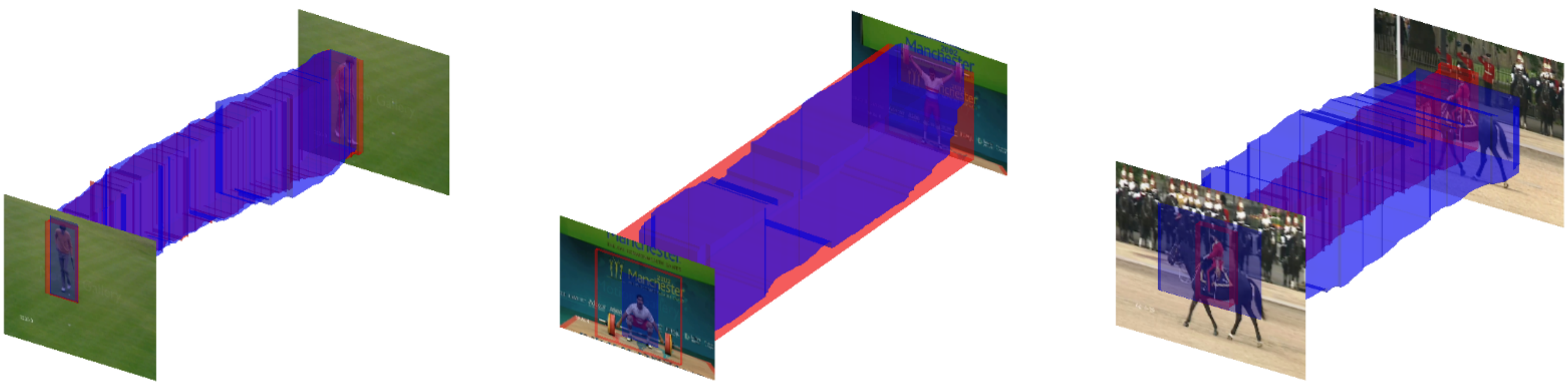
Training without ground truth boxes



UCF Sports

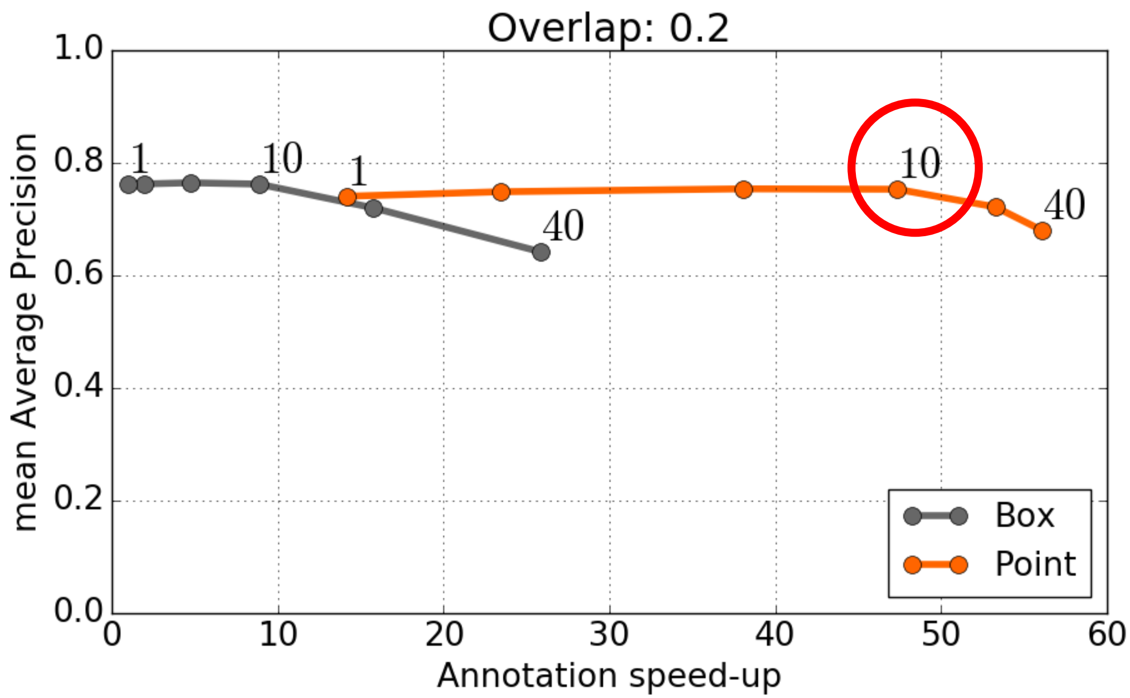
Mean AP maintained using our mined proposals.

Qualitative results



- Ground truth boxes
- Mined proposal

Lowering annotation frame-rate



UCF Sports

Up to 50 times speed-up at similar performance.

Hollywood2Tubes

Dataset to demonstrate how easy action annotation becomes.
Contains actions and instances new to action localization.

a. 

Multi-label videos.

b. 

Contextual actions.

c. 

Group interactions.

Download:

tinyurl.com/hollywood2tubes

Videos from Hollywood2 by Marszalek et al. *CVPR'09*

Take home message

Weakly-supervised computer vision is aided by reasonable proposals for objects, segments and/or actions.

Proposals are further refined with point annotations. Especially useful for precise annotations, like segments and actions.

Facilitates dataset construction and/or enrichment.

Overview

1. Image benchmarks, PASCAL, ImageNet, MSCOCO
2. Video benchmarks, TRECVID, ActivityNet
3. Labels from humans, experts, volunteers, crowdsourcing
4. Labels from similarity, nearest neighbor, simple features
5. Weakly-supervised computer vision

6. Event recognition by learning