Video Concept Detection by Deep Nets with FLAIR

Cees Snoek, Koen van de Sande, Daniel Fontijne

Qualcomm Technologies
Netherlands B.V.

University of Amsterdam
The Netherlands

Presented by **Thomas Mensink**, UvA

Summary of our efforts

Last year

Deep CNN for video concept detection and localization

This year

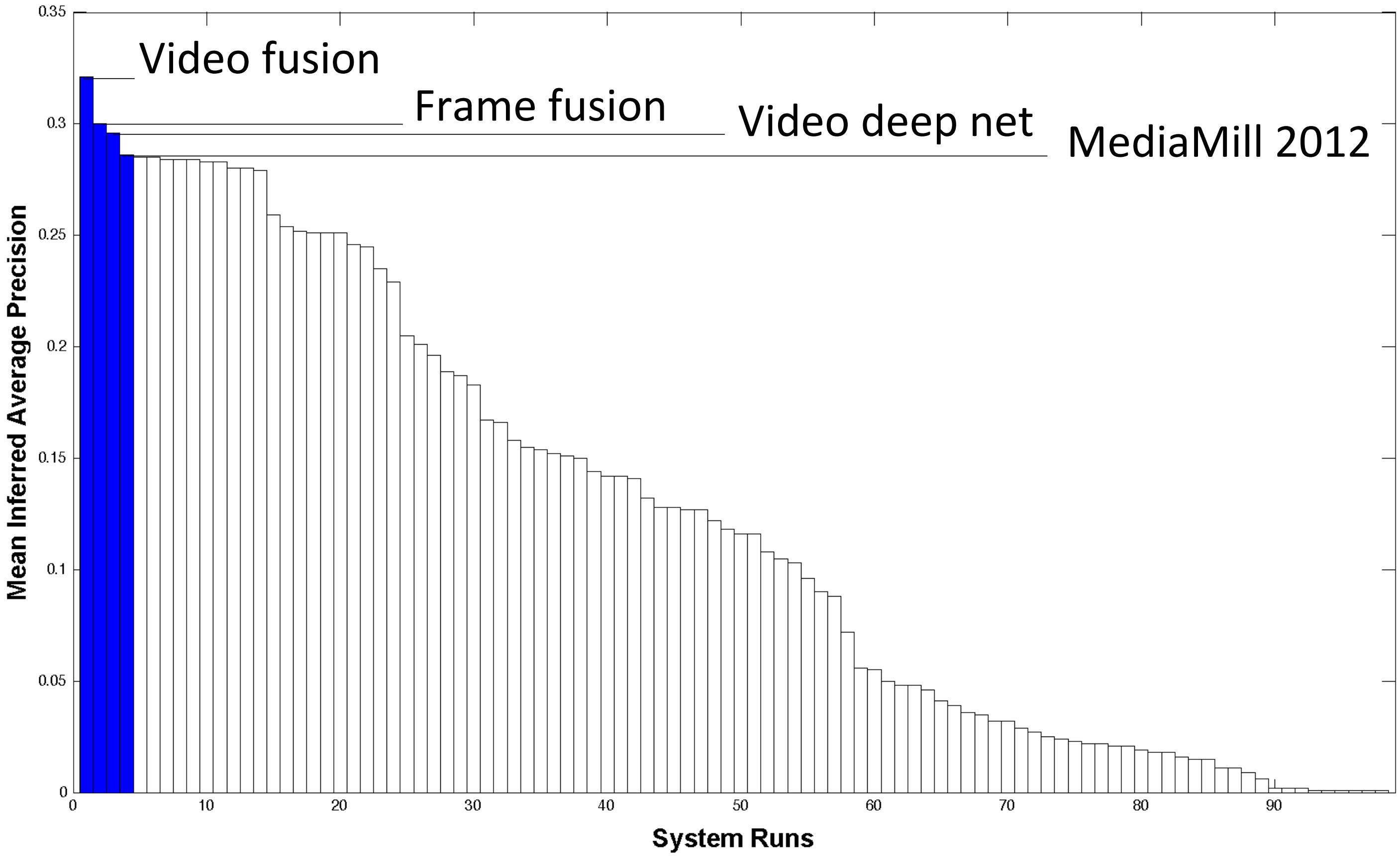
Tangential improvements for concept detection

Our main innovation is in concept localization

TASK I

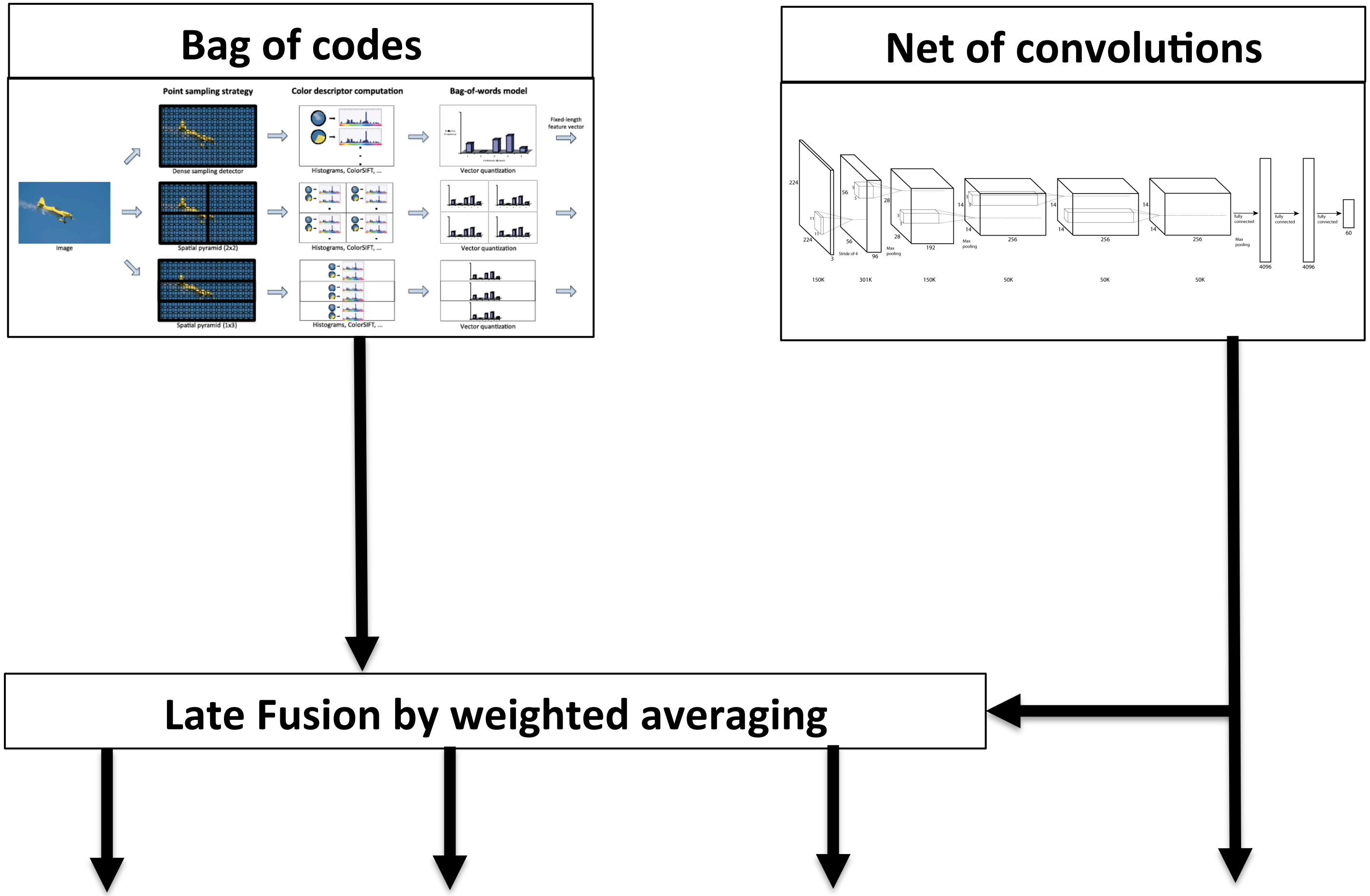
DETECTING CONCEPTS

Conclusion from TRECVID 2013



Bag of words and deep net profit from each other

MediaMill TRECVID 2014 runs



Bag of codes

Net of convolutions

Late Fusion by weighted averaging

Run: Fusion 1
Fuse-all

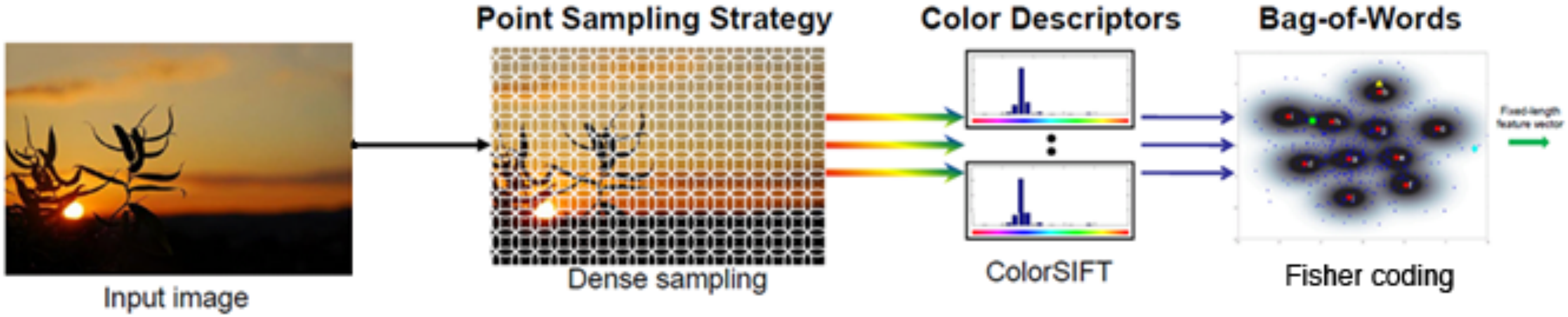
Run: Fusion 2
8x CNN + BoW

Run: Fusion 3
Best CNN + BoW

Baseline Run:
8x CNN

MediaMill: Color difference coding

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



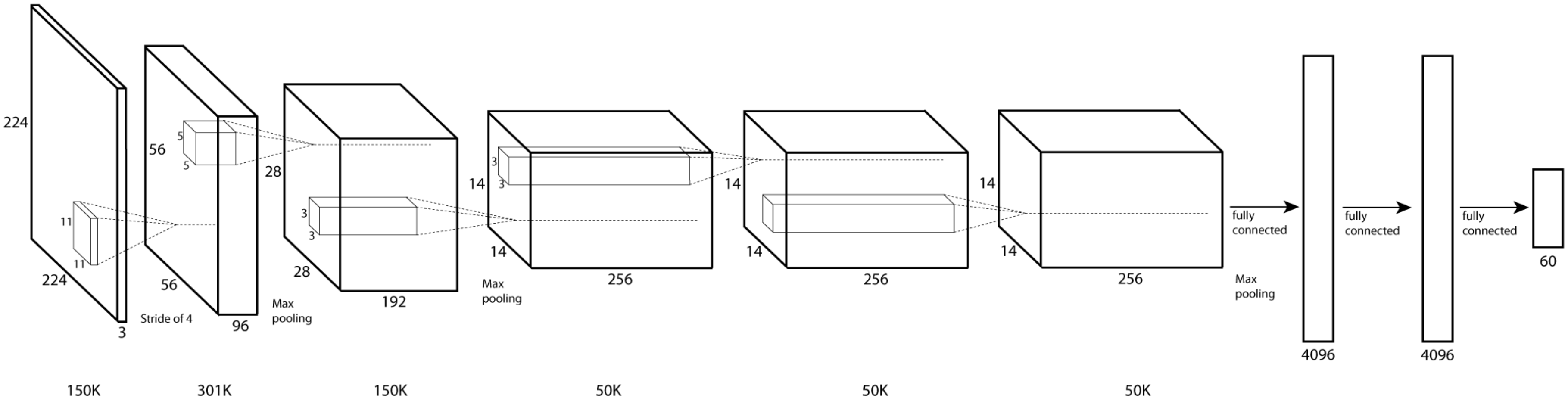
Color Descriptor software available for download at http://colordescriptors.com

MediaMill: Video deep learning

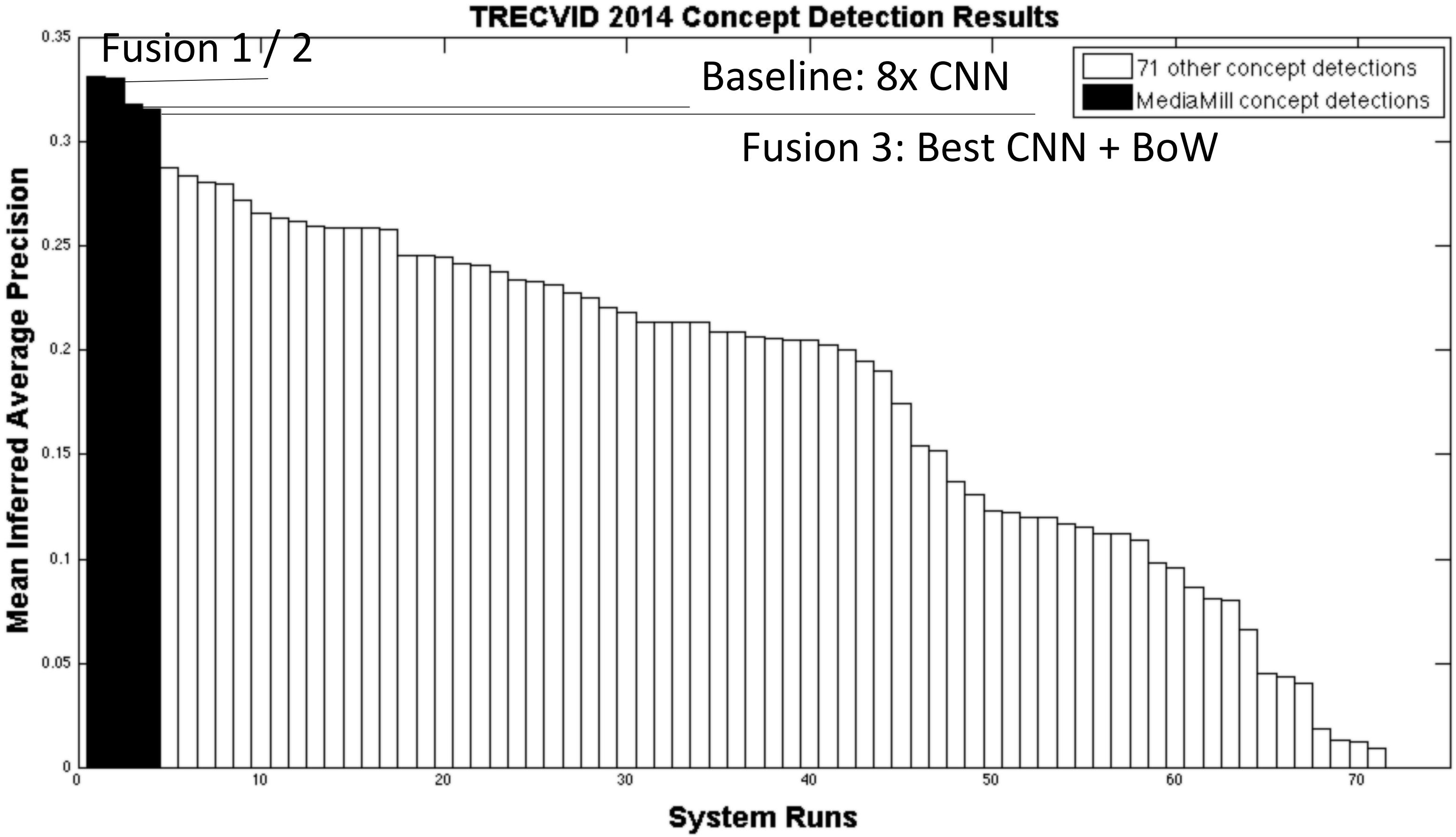
Convolutional neural network with 8 layers with weights

Trained using error back propagation

- ImageNet for pre-training

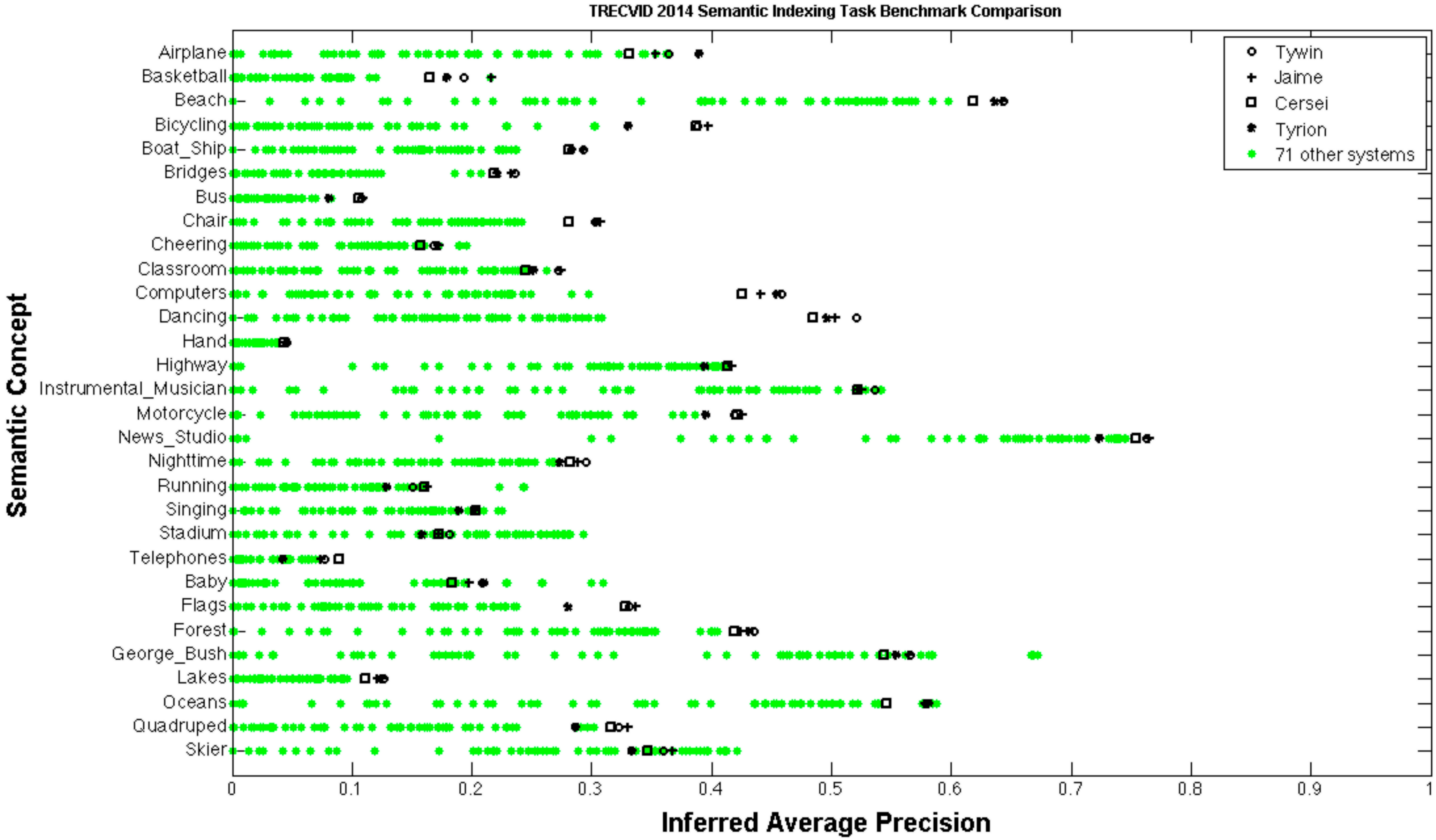


Results



Bag of words and deep net profit from each other, better results with more nets

Results per concept

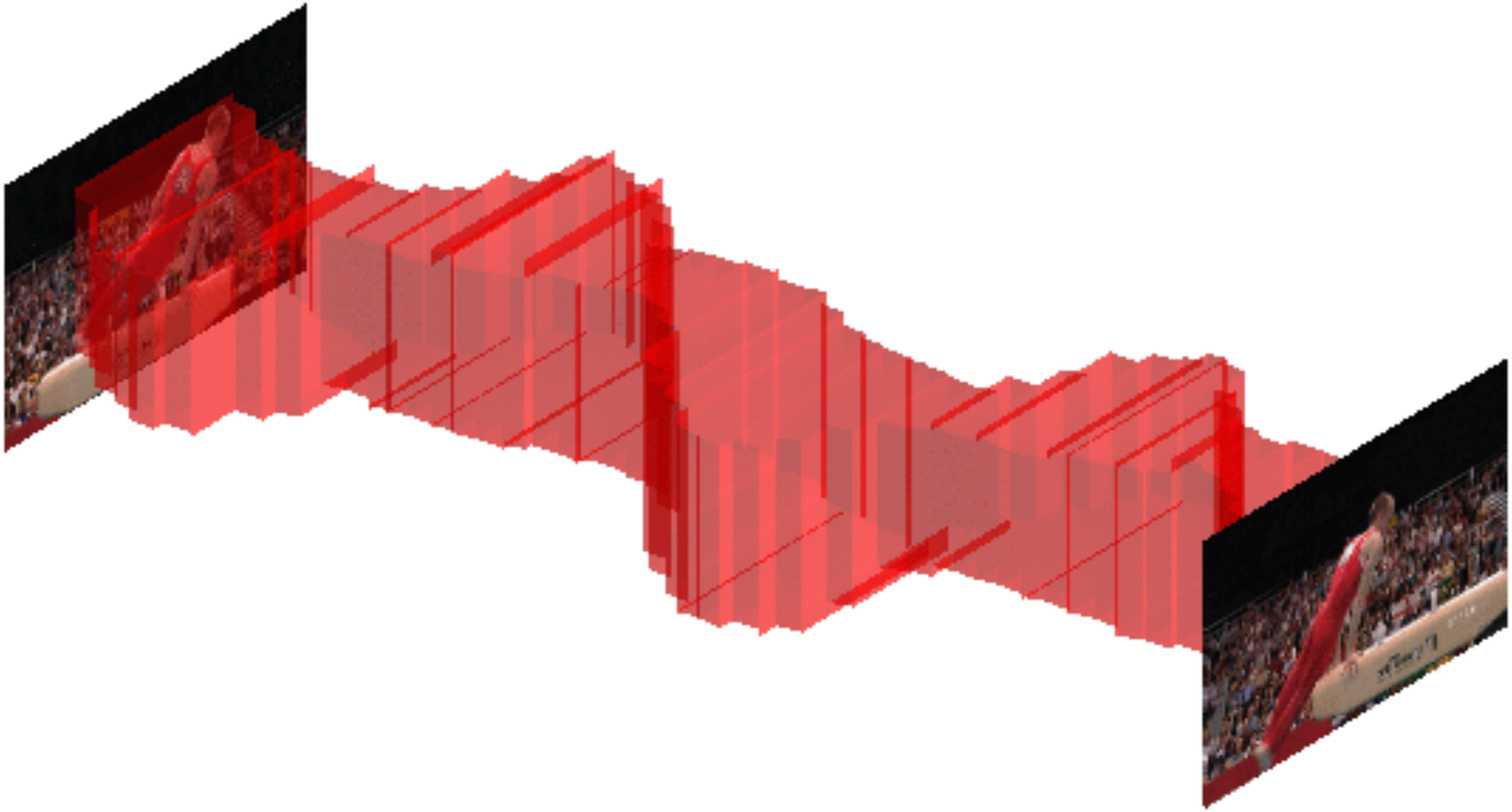


TASK II

LOCALIZING CONCEPTS

Fisher and VLAD with FLAIR, Koen van de Sande, Cees Snoek, and Arnold Smeulders
CVPR 2014

Goal: meaningful localization



Finding **where**, **when**, **what** is happening

Challenges: huge search space, non-rigid deformation

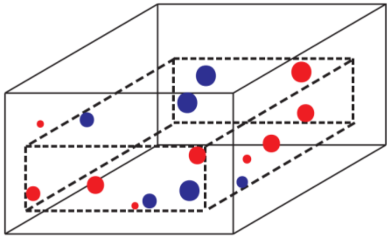
Related work

Sliding Window

Image

[Rowley, 1996]

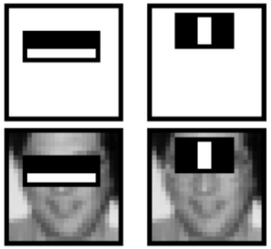
Video



[Rodriguez, 2008]

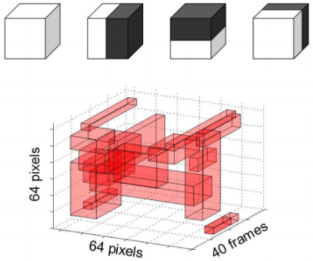
Boosting Cascade

Image



[Viola & Jones, 2001]

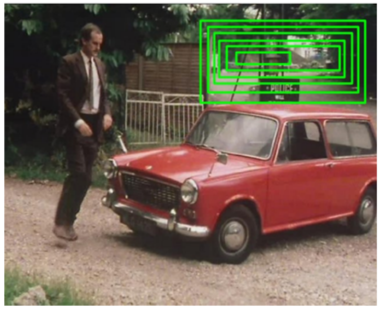
Video



[Ke, 2005]

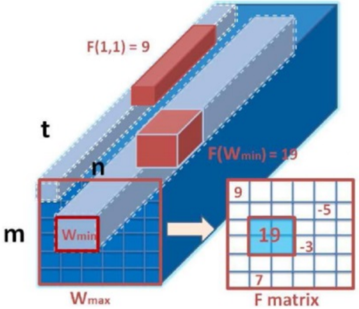
Branch and Bound

Image



[Lampert, 2009]

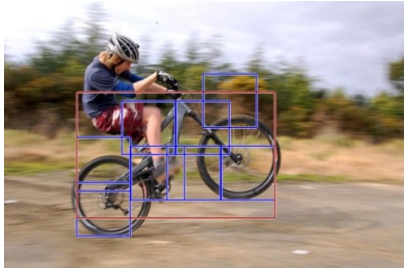
Video



[Yuan, 2011]

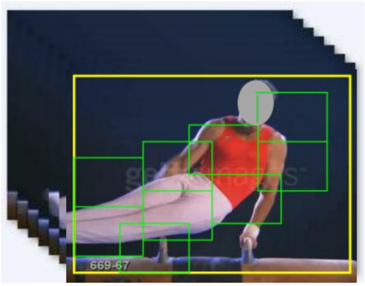
Deformable Parts

Image



[Felzenswalb, 2008]

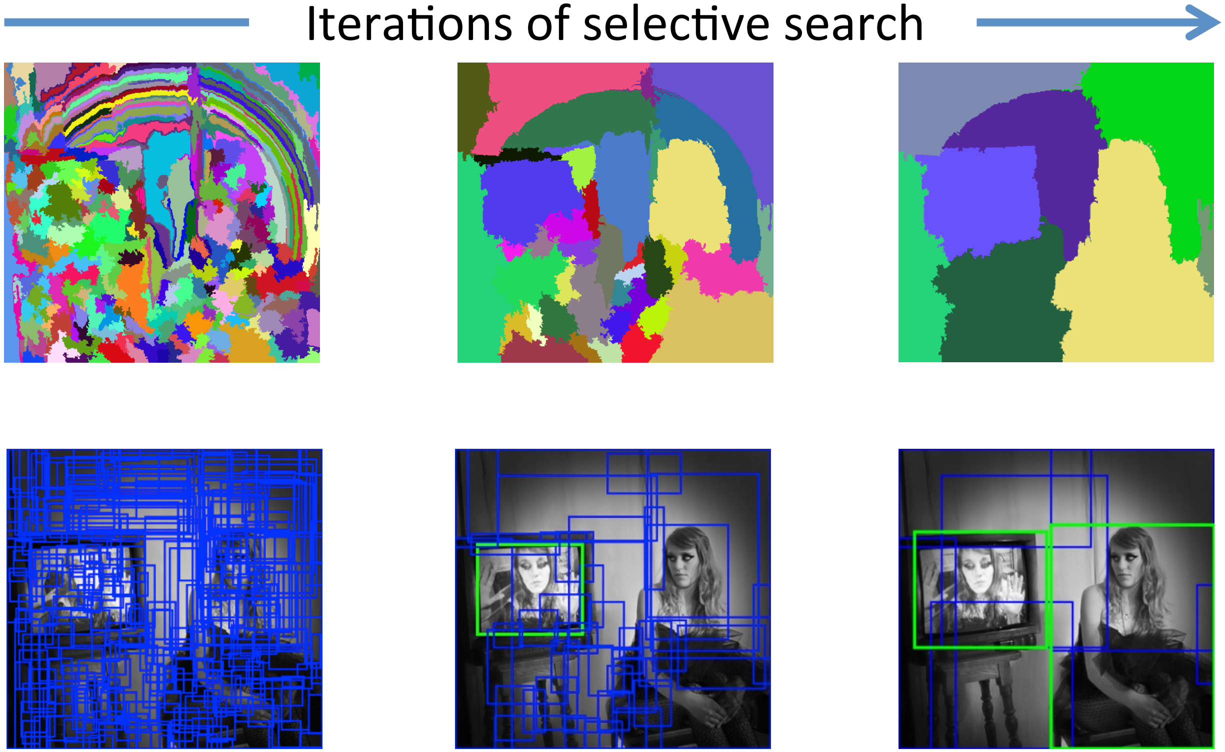
Video



[Tian, 2013]

Inspiration: Selective Search

[Uijlings, 2013]



Iterations of selective search

Hierarchical grouping of super-pixels

Object proposals

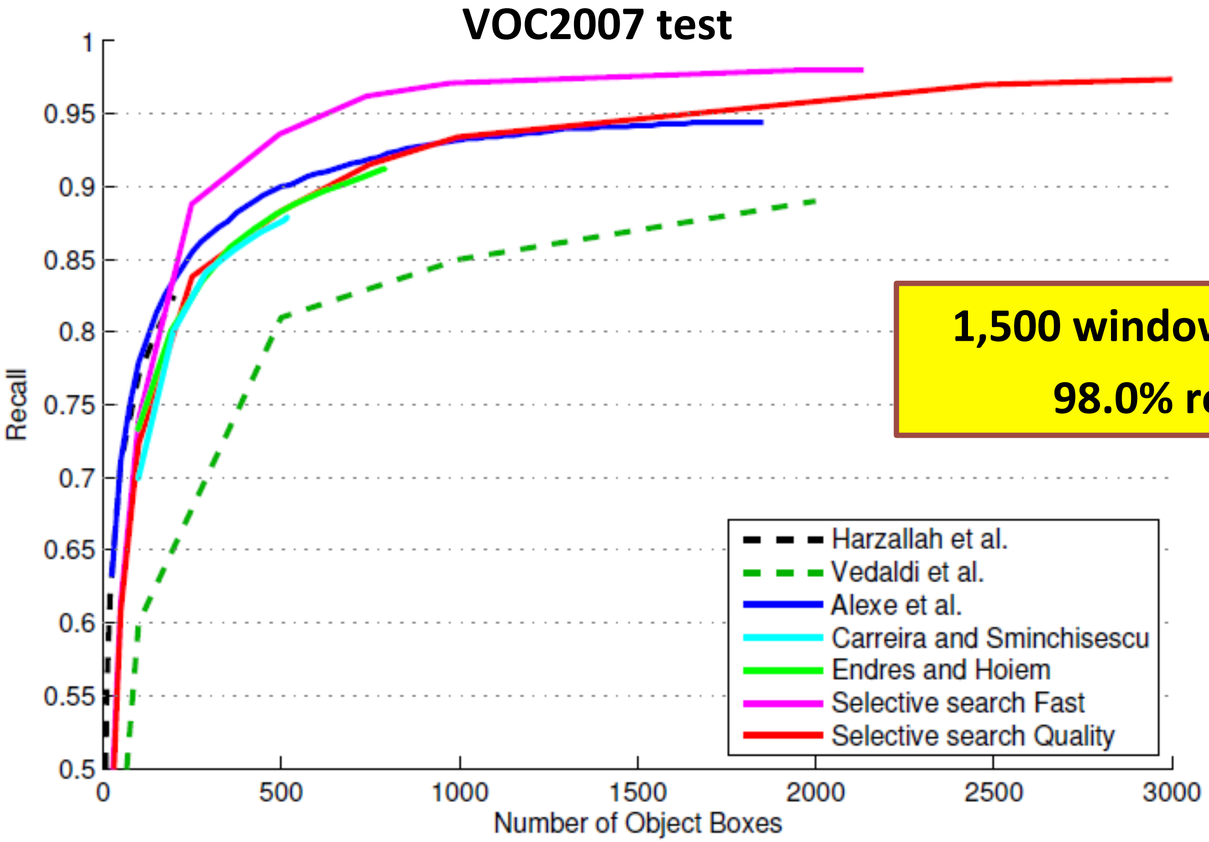
High recall with modestly sized object hypotheses set

Feasible to train an expensive classifier

Selective Search

Multiple complementary invariant color spaces

Location hypotheses are class-independent



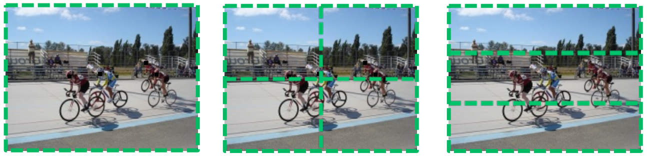
1,500 windows/image

98.0% recall

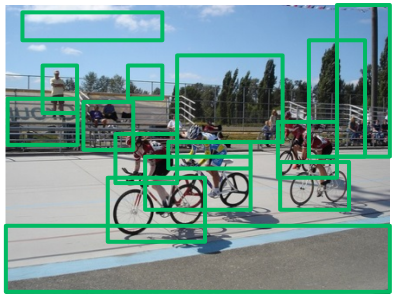
Software available for download at http://koen.me/research/selectivesearch/

Local object classification

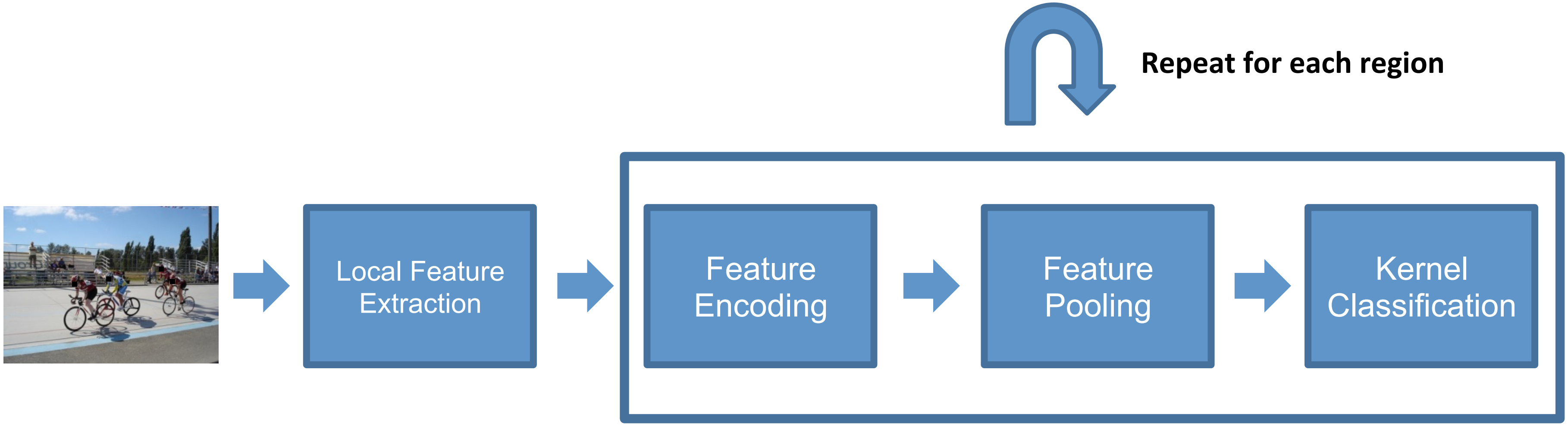
Requires **repetitive** computations on **overlapping** regions



Spatial Pyramids [*Lazebnik, CVPR06*]
(#regions: 10-100)



Object Detection [*Sande, ICCV11*]
(#regions: 1,000-10,000)



Repeat for each region

Local Feature Extraction

Feature Encoding

Feature Pooling

Kernel Classification

Features

Use SIFT and ColorSIFT descriptors

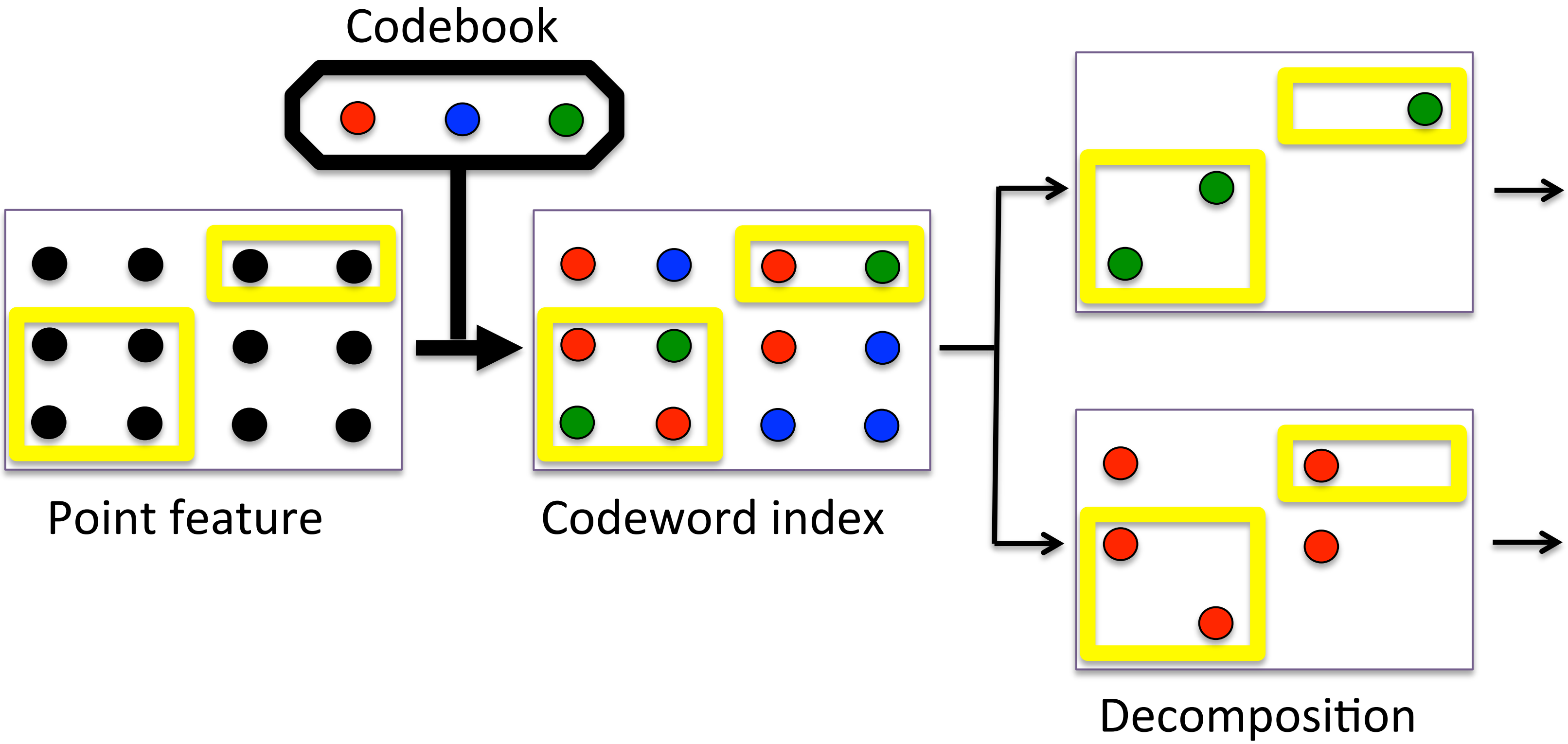
Bag-of-words, VLAD, Fisher vector encoding

Encoding 2000 boxes per image is expensive

Bag-of-words:	10s
VLAD:	30s
Fisher:	120s

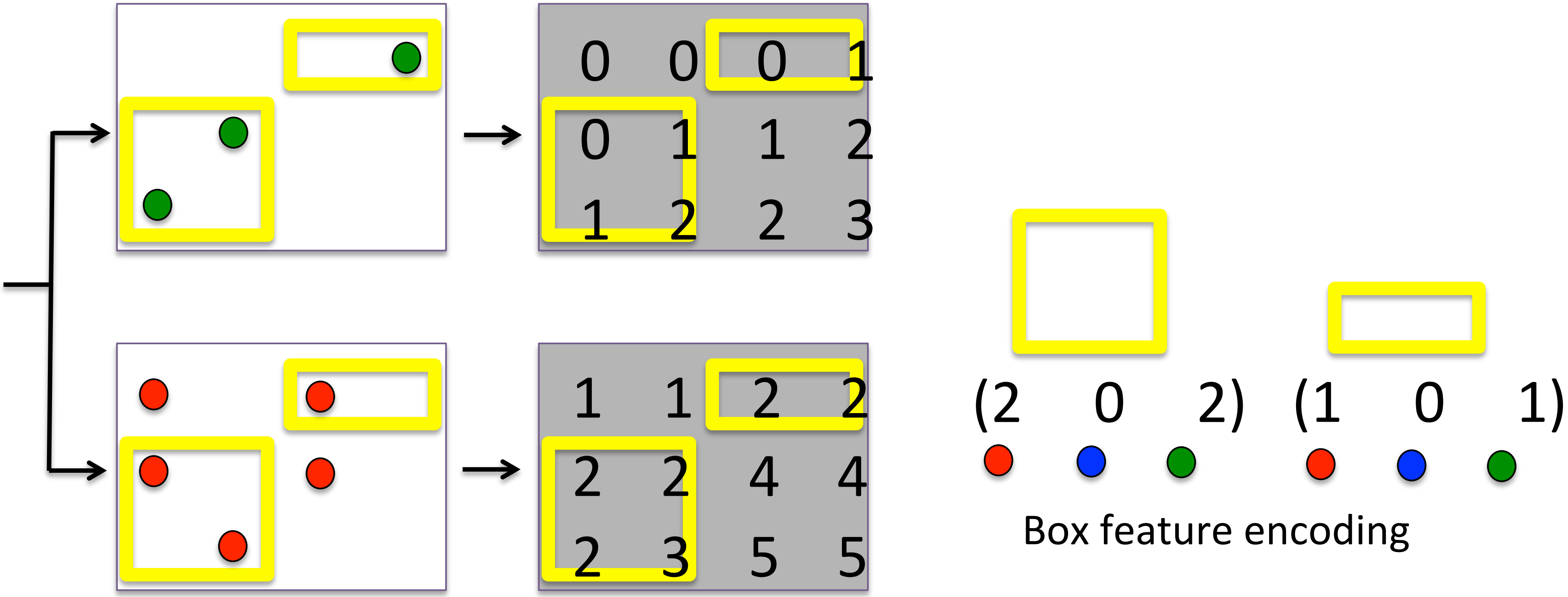
Key idea

Decompose assignment over codebook elements



Area-independent decomposition

Fast box evaluation with integral images



0	0	0	1
0	1	1	2
1	2	2	3

1	1	2	2
2	2	4	4
2	3	5	5

(2 0 2) (1 0 1)

Decomposition

Integral image

Box feature encoding

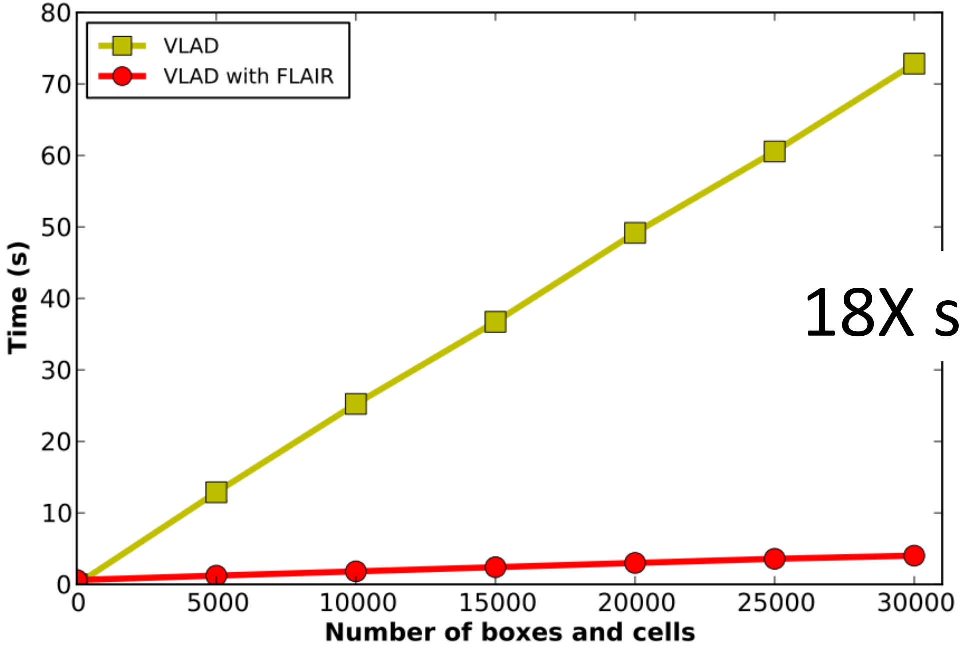
Porikil, CVPR 2005

VLAD with FLAIR

Decomposition as multi-dimensional integral image

Sparsity drops memory from 14GB to 1GB/image

Supports power norm, L2 norm and spatial pyramid



18X speedup

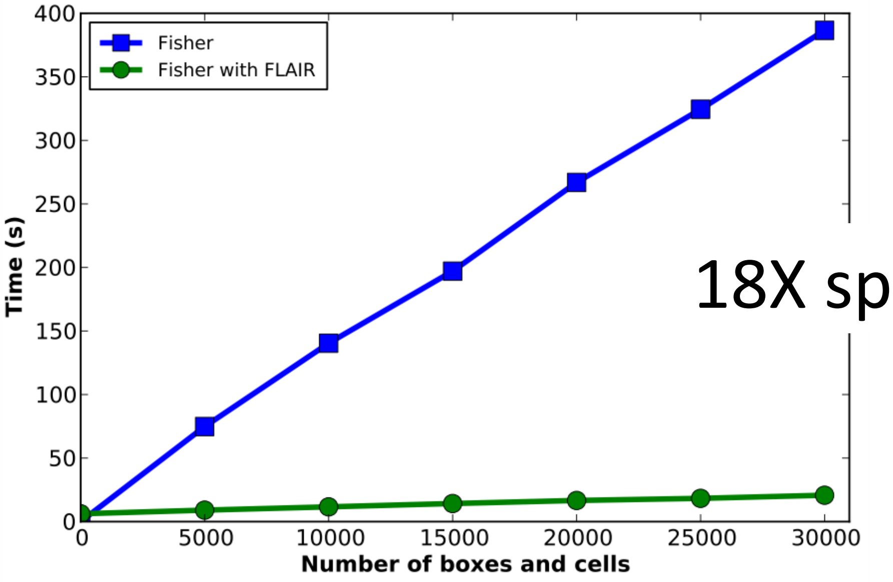
Fisher with FLAIR

Decomposition as four multi-dimensional integral images [See paper]

Supports power norm, L2 norm, spatial pyramids

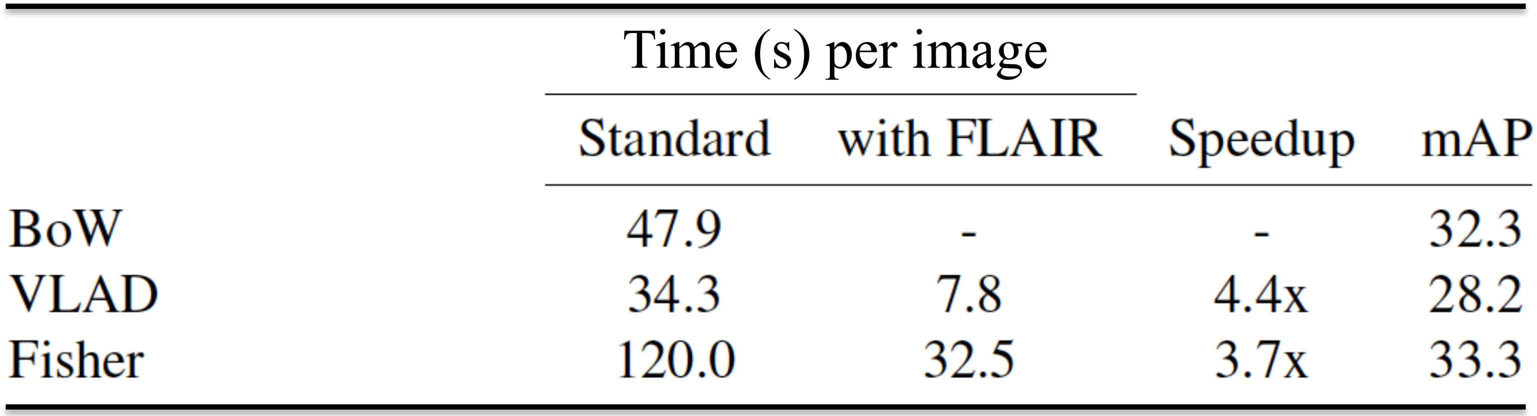
No need for approximations

Scalable to modern datasets



18X speedup

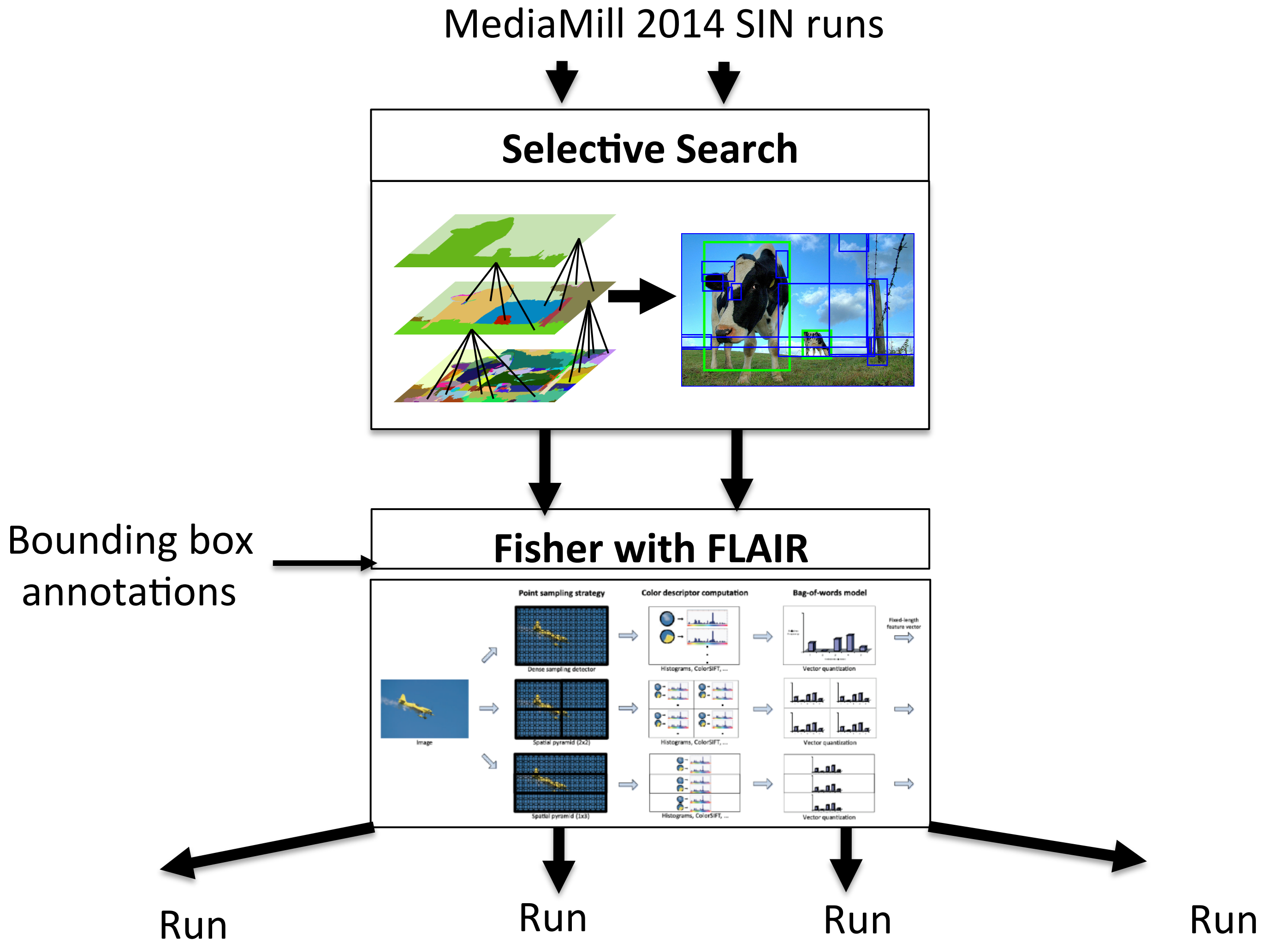
Overall detection speedup and accuracy



	Time (s) per image			
	Standard	with FLAIR	Speedup	mAP
BoW	47.9	-	-	32.3
VLAD	34.3	7.8	4.4x	28.2
Fisher	120.0	32.5	3.7x	33.3

Fisher with FLAIR is better and faster than BoW

MediaMill TRECVID 2014 runs



MediaMill 2014 SIN runs

Selective Search

Bounding box annotations

Fisher with FLAIR

Run

Run

Run

Run

Implementation details

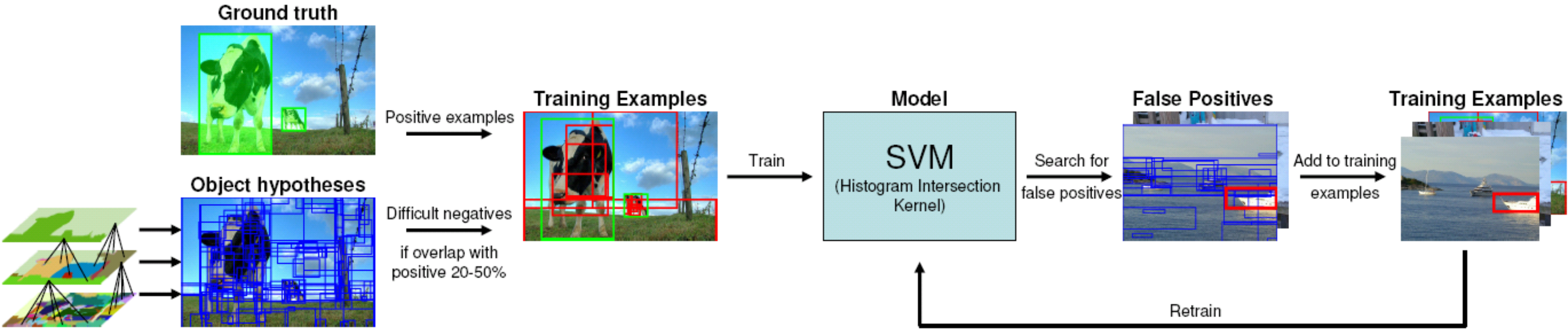
PCA-reduced ColorSIFT descriptors to 80D

Fisher with FLAIR encoding

Spatial pyramid

Linear SVM

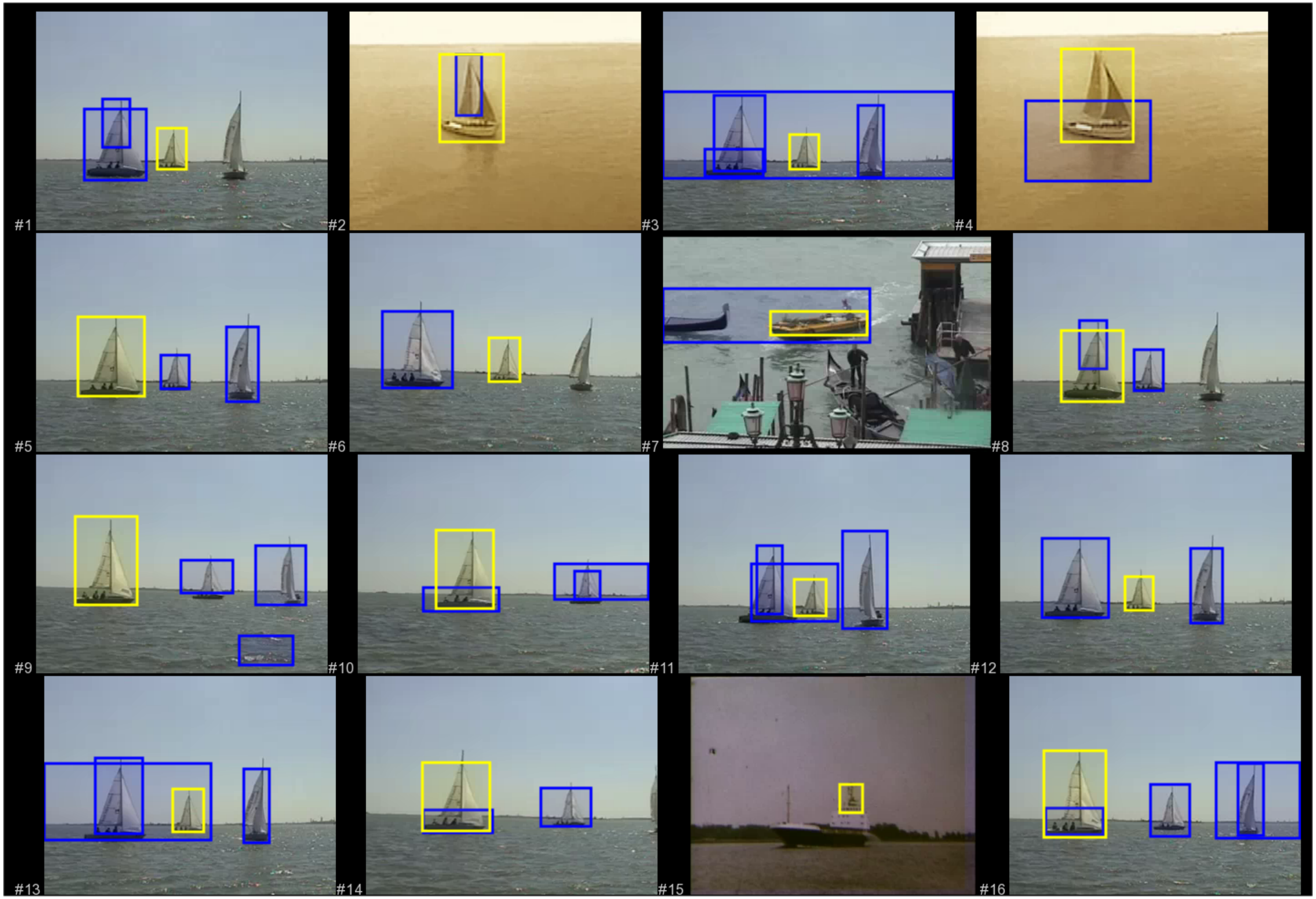
Hard negative mining



Boat

🟨 Best box

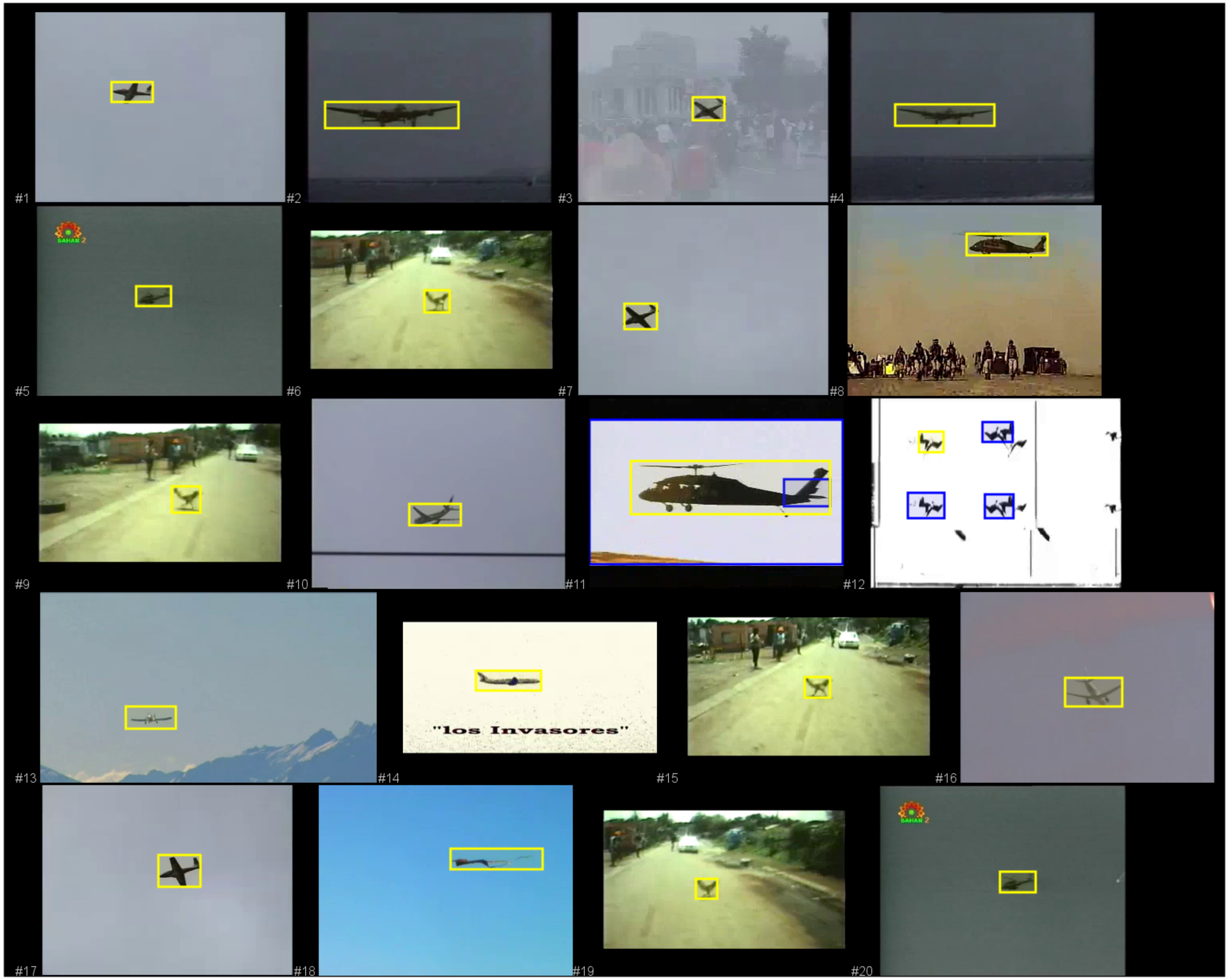
🟦 Other boxes



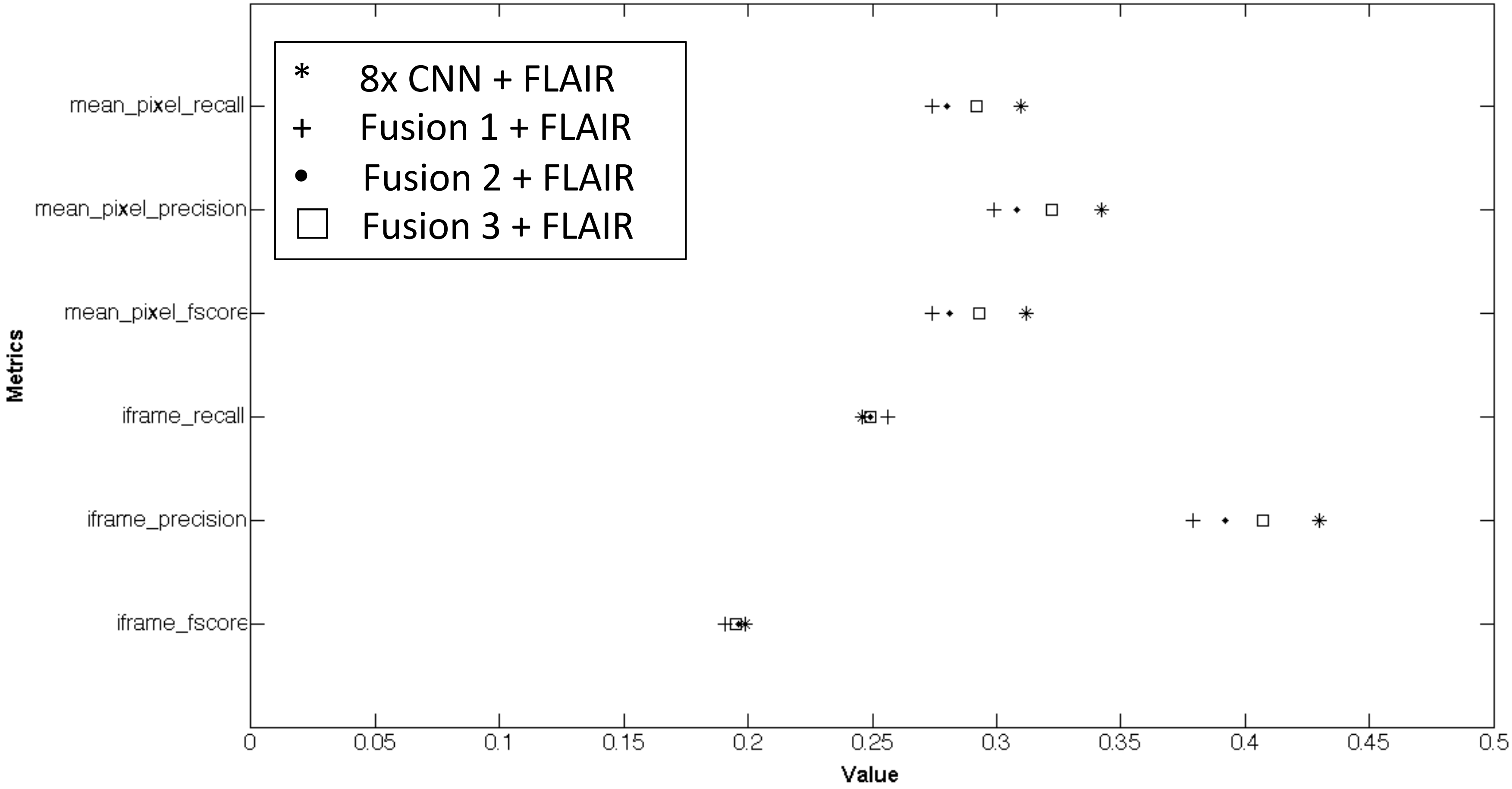
Airplane

🟨 Best box

🟦 Other boxes



Results



FLAIR after deep nets is best

Conclusions

Bag of words and deep net profit from each other

Encoding Fisher with FLAIR is 18x faster

- Area independent
- Supports spatial pyramids, power norm, L2 norm
- No approximation

Allows for large-scale localization in video