First Croatian Computer Vision Workshop (CCVW 2012)
September 20-21, 2012, Zagreb, Croatia

Human action recognition: Recent progress, open questions and future challenges

Ivan Laptev

ivan.laptev@inria.fr

WILLOW, ENS/INRIA/CNRS, Paris

Computer vision grand challenge: Dynamic scene understanding

Objects:
cars, glasses, people, etc…

Actions:
drinking, running, door exit, car enter, etc…

constraints

Scene categories:
indoors, outdoors, street scene, etc…

Geometry:
Street, wall, field, stair, etc…

Why analyzing people and human actions?

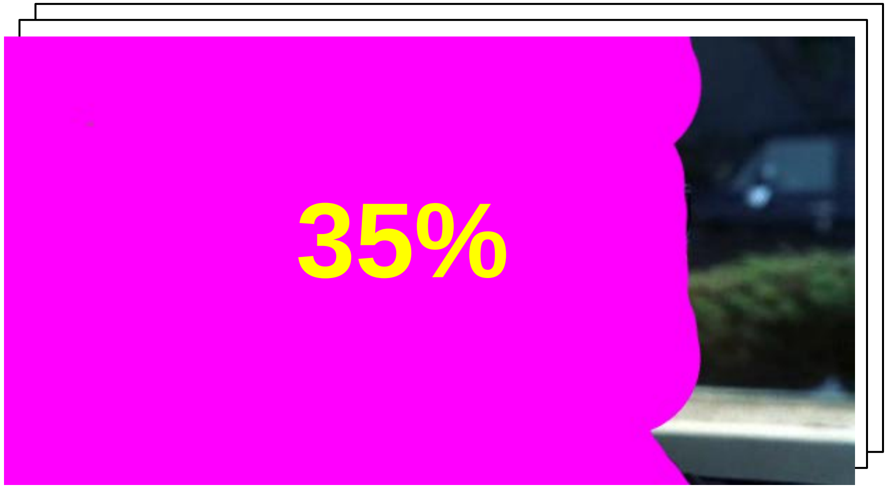
How many person pixels are in video?

Movies

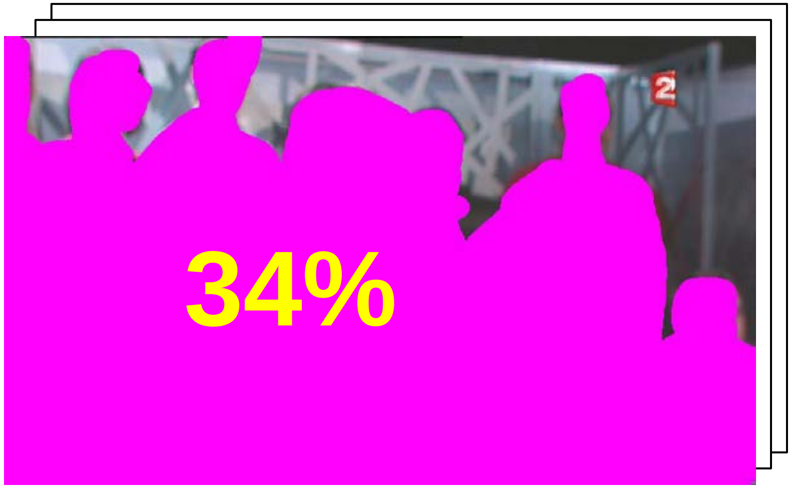
TV

YouTube

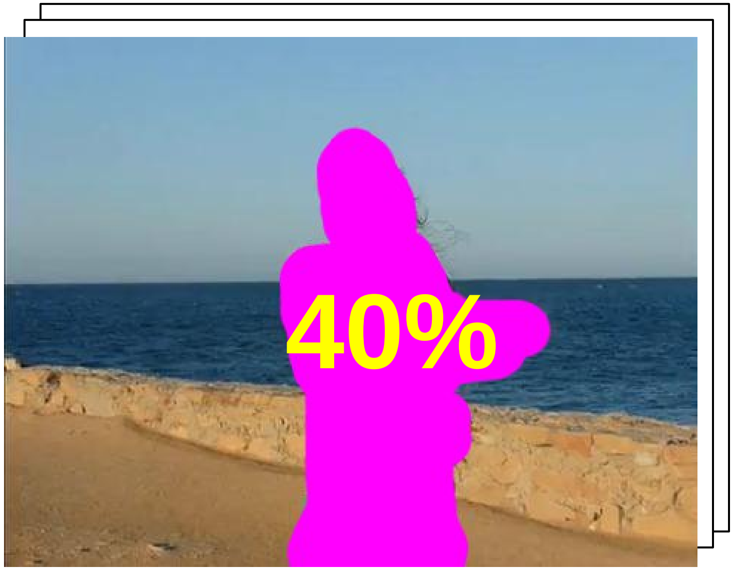
How many person pixels are in video?



Movies



TV



YouTube

Applications

- Analyzing video archives

First appearance of
N. Sarkozy on TV

Sociology research:
Influence of character
smoking in movies

Education: How do I
make a pizza?

- Surveillence

Where is my cat?

Predicting crowd behavior
Counting people

- Graphics

Motion capture and animation

Technology: Access to lots of data

- Huge amount of video is available and growing

BBC Motion Gallery

TV-channels recorded since 60's

>34K hours of video uploads every day

~30M surveillance cameras in US => ~700K video hours/day

Why action recognition is hard?

- Need to process very large amounts of video data
- Need to deal with large appearance variations, many classes

Drinking

Smoking

This talk:

Review of work on action recognition

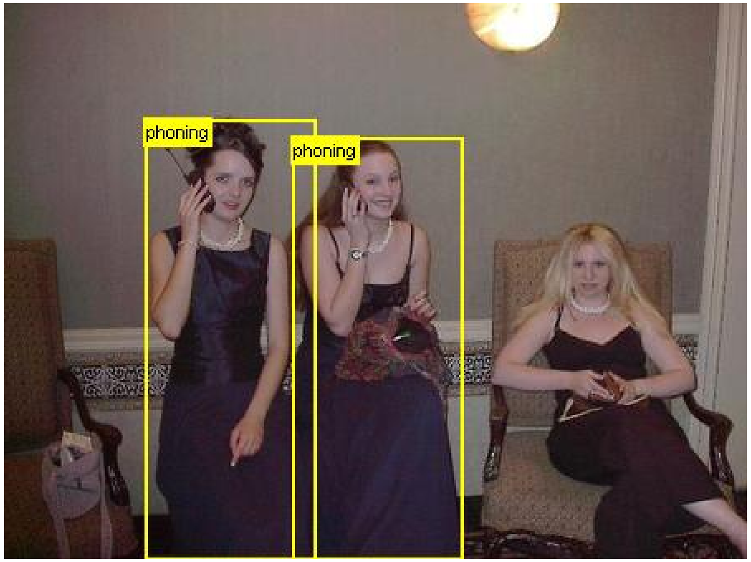
Discussion: Do we ask the right questions?

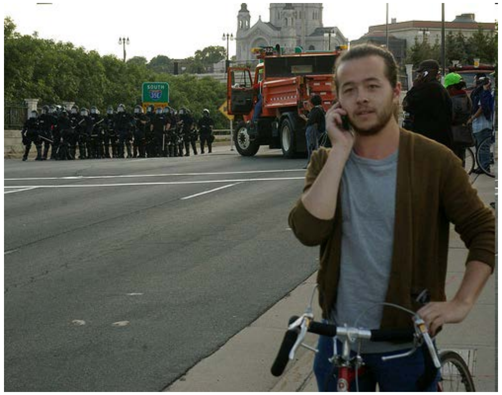
Our more recent work

Activities characterized by a pose

Slide credit: A. Zisserman

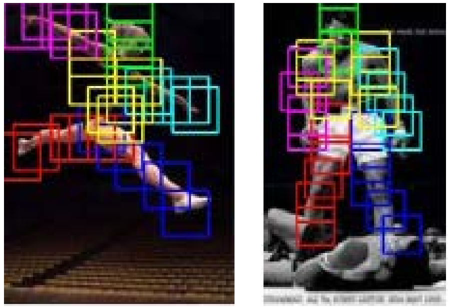
Activities characterized by a pose





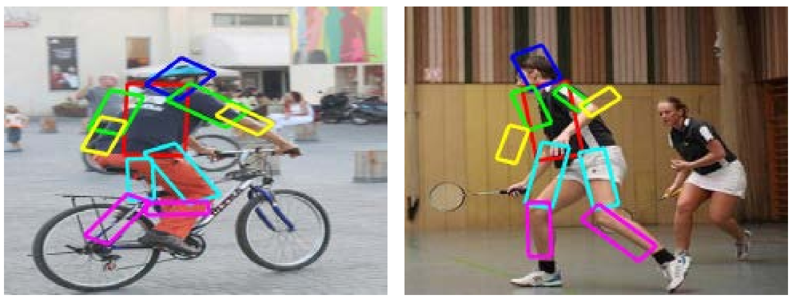
Slide credit: A. Zisserman

Human pose estimation



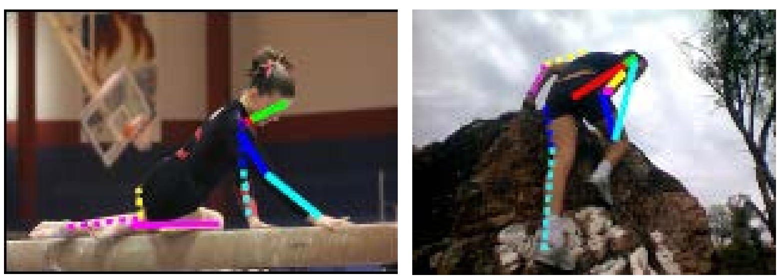
Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In Proc. **CVPR 2011**

Extension of LSVM model of Felzenszwalb et al.



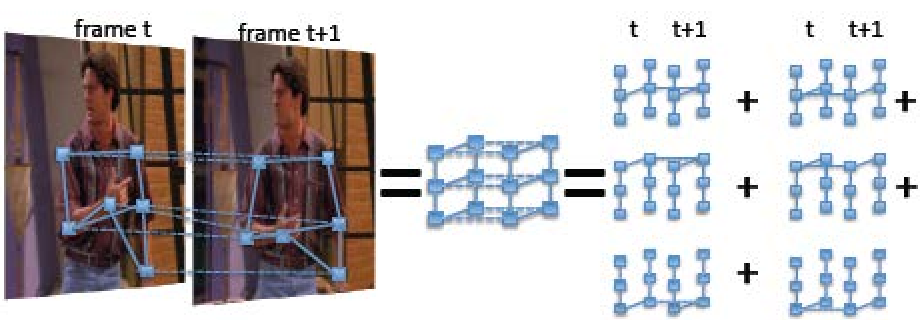
Y. Wang, D. Tran and Z. Liao. Learning Hierarchical Poselets for Human Parsing. In Proc. **CVPR 2011**.

Builds on Poslets idea of Bourdev et al.



S. Johnson and M. Everingham. Learning Effective Human Pose Estimation from Inaccurate Annotation. In Proc. **CVPR 2011**.

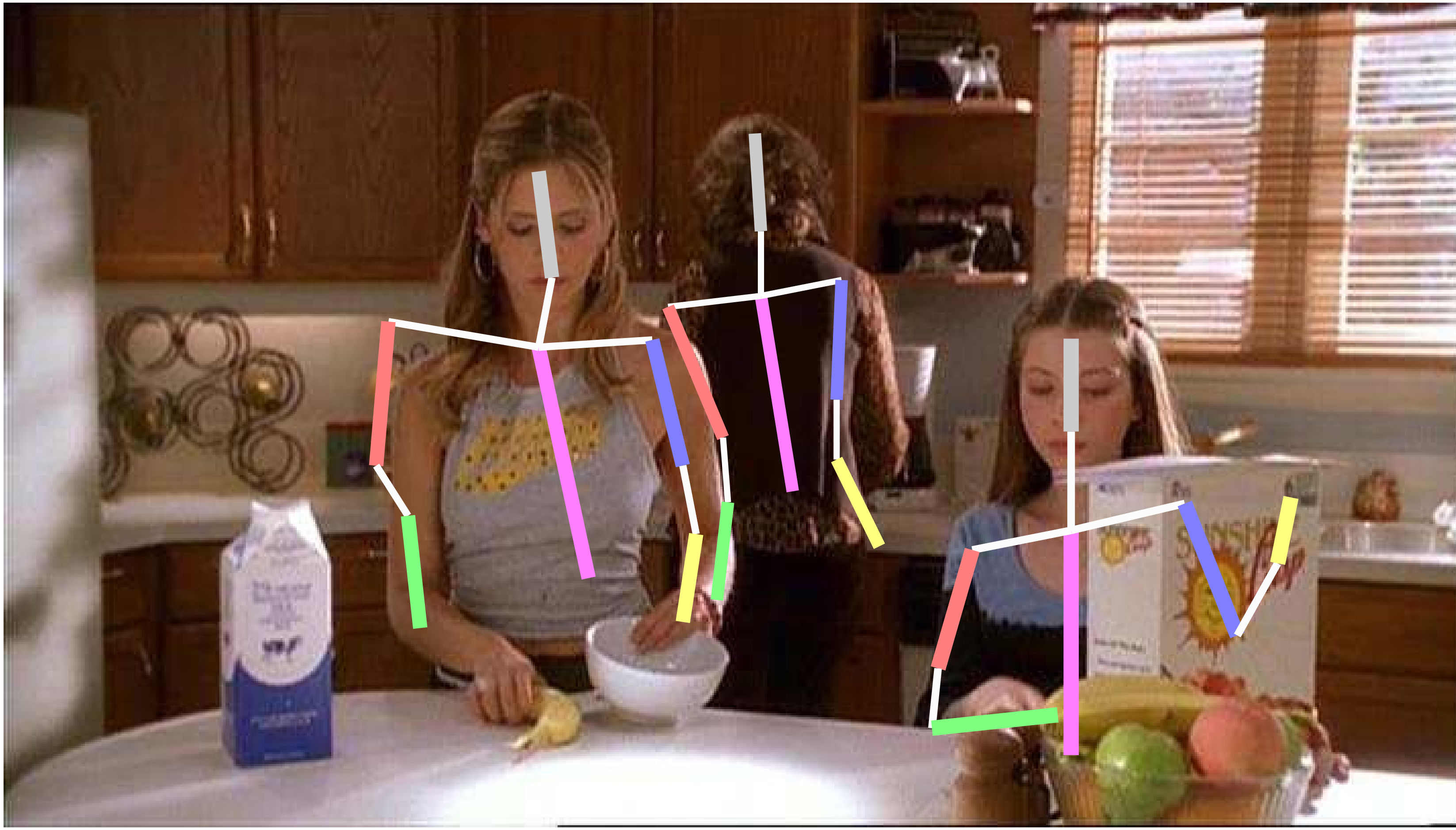
Learns from lots of noisy annotations



B. Sapp, D.Weiss and B. Taskar. Parsing Human Motion with Stretchable Models. In Proc. **CVPR 2011**.

Explores temporal continuity

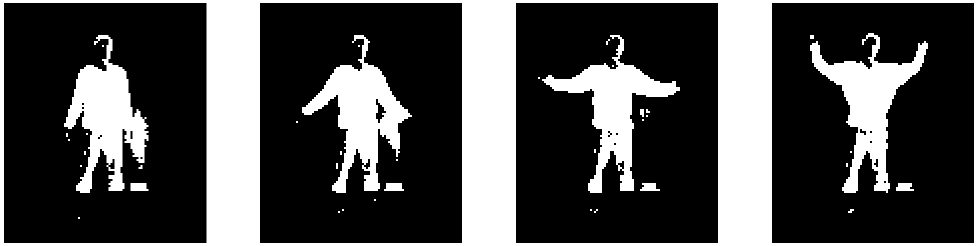
Pose estimation is still a hard problem



Issues:

- occlusions
- clothing and pose variations

Appearance-based methods: global shape



[A.F. Bobick and J.W. Davis, PAMI 2001]

Idea: summarize motion in video in a *Motion History Image (MHI)*:

L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as spacetime shapes. 2007

Appearance-based methods: shape tracking

[Baumberg and Hogg, ECCV 1994]

Goal:
Interpret complex dynamic scenes

Common methods:

- Segmentation using background model **-> hard**

- Tracking using appearance model **->hard**

Common problems:

- Complex & changing BG

- Changing appearance

⇒ **Global assumptions** about the scene are **unreliable**

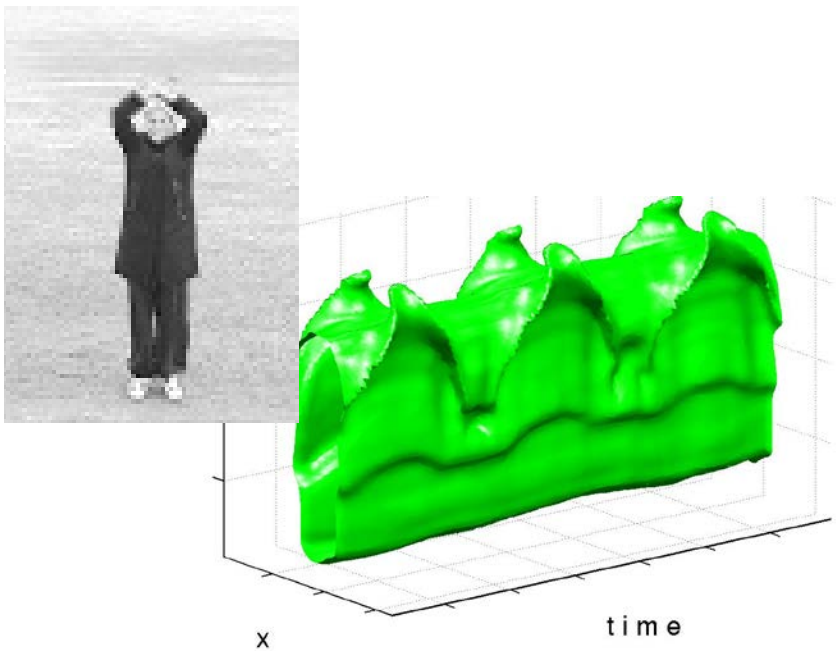
Space-time

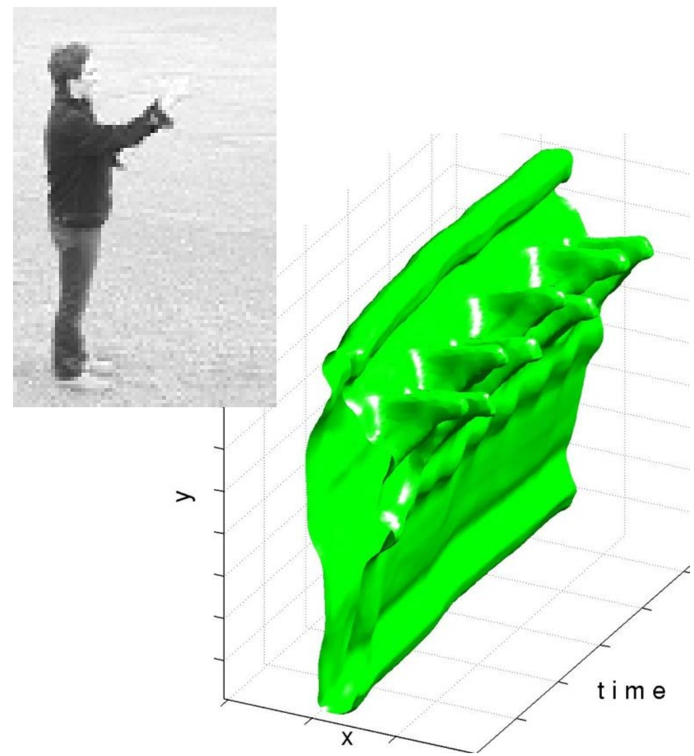
No global assumptions $\Rightarrow$
Consider local spatio-temporal neighborhoods

hand waving

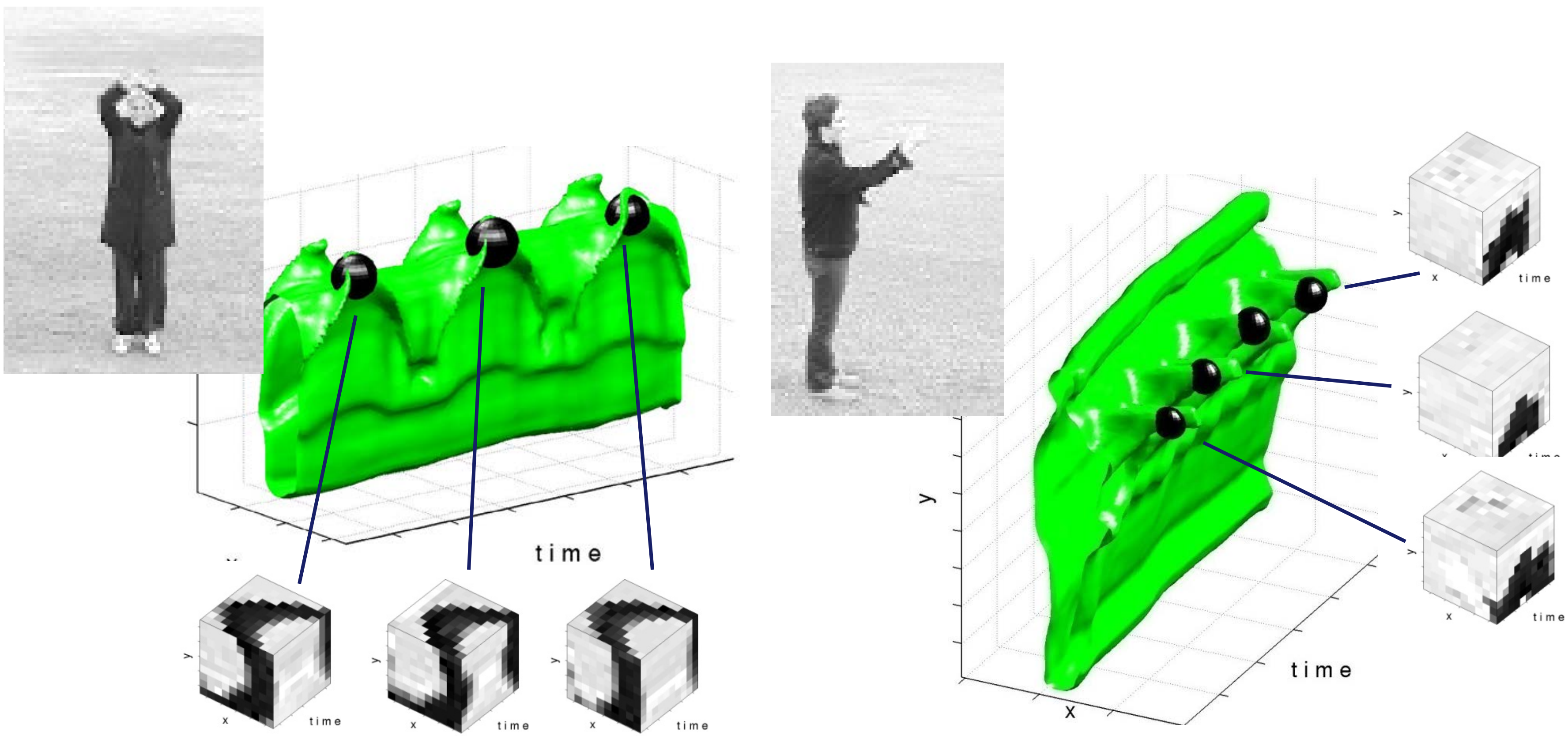
boxing

Actions == Space-time objects?

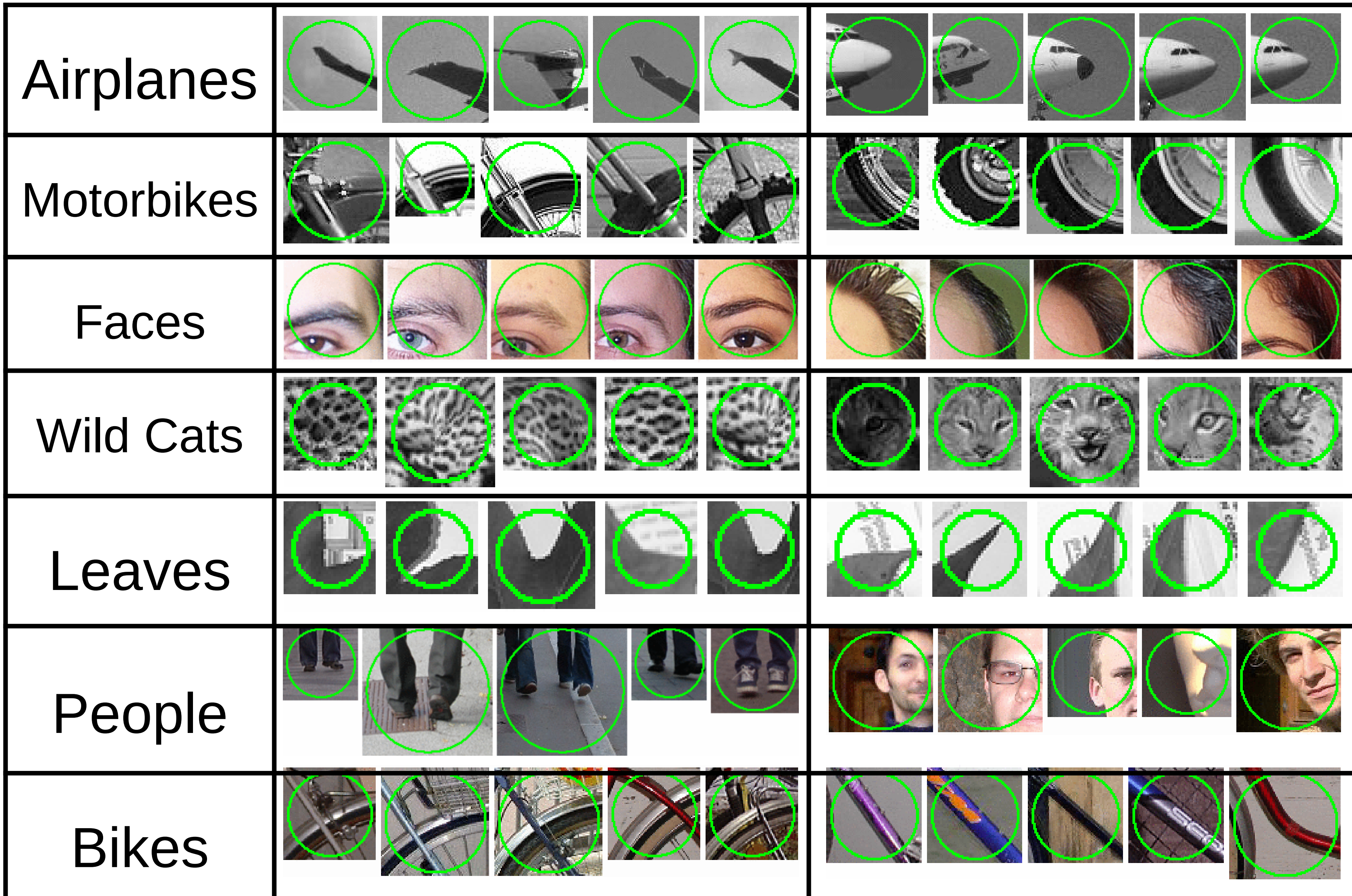




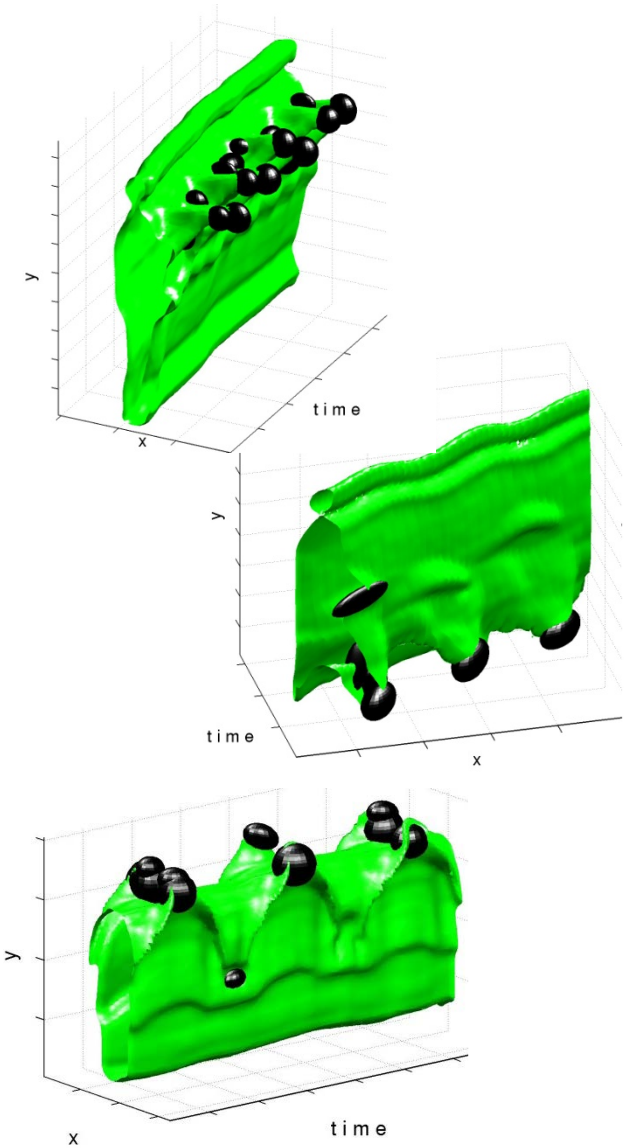
Space-time local features



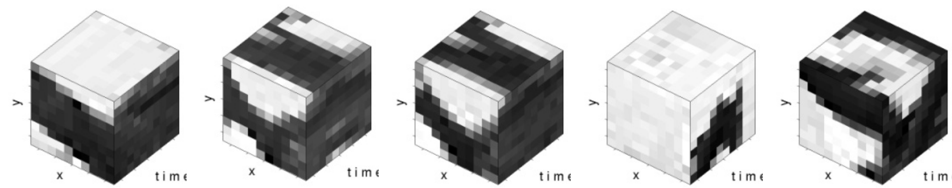
Local approach: Bag of Visual Words



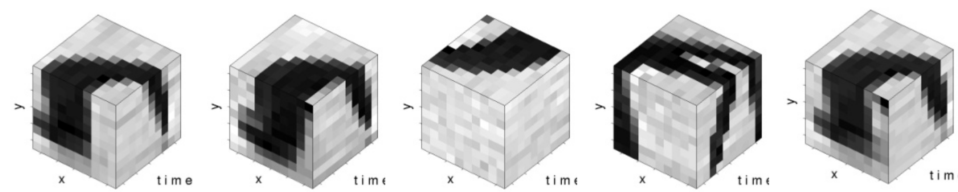
Local features for human actions



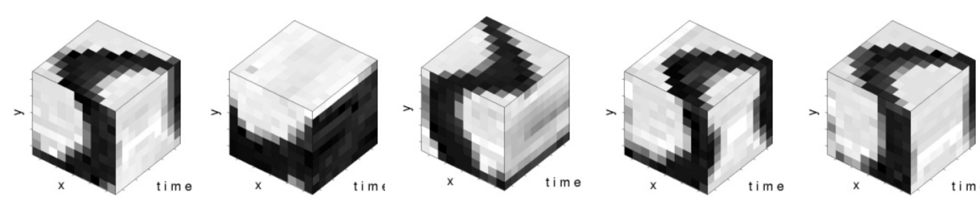
boxing



walking



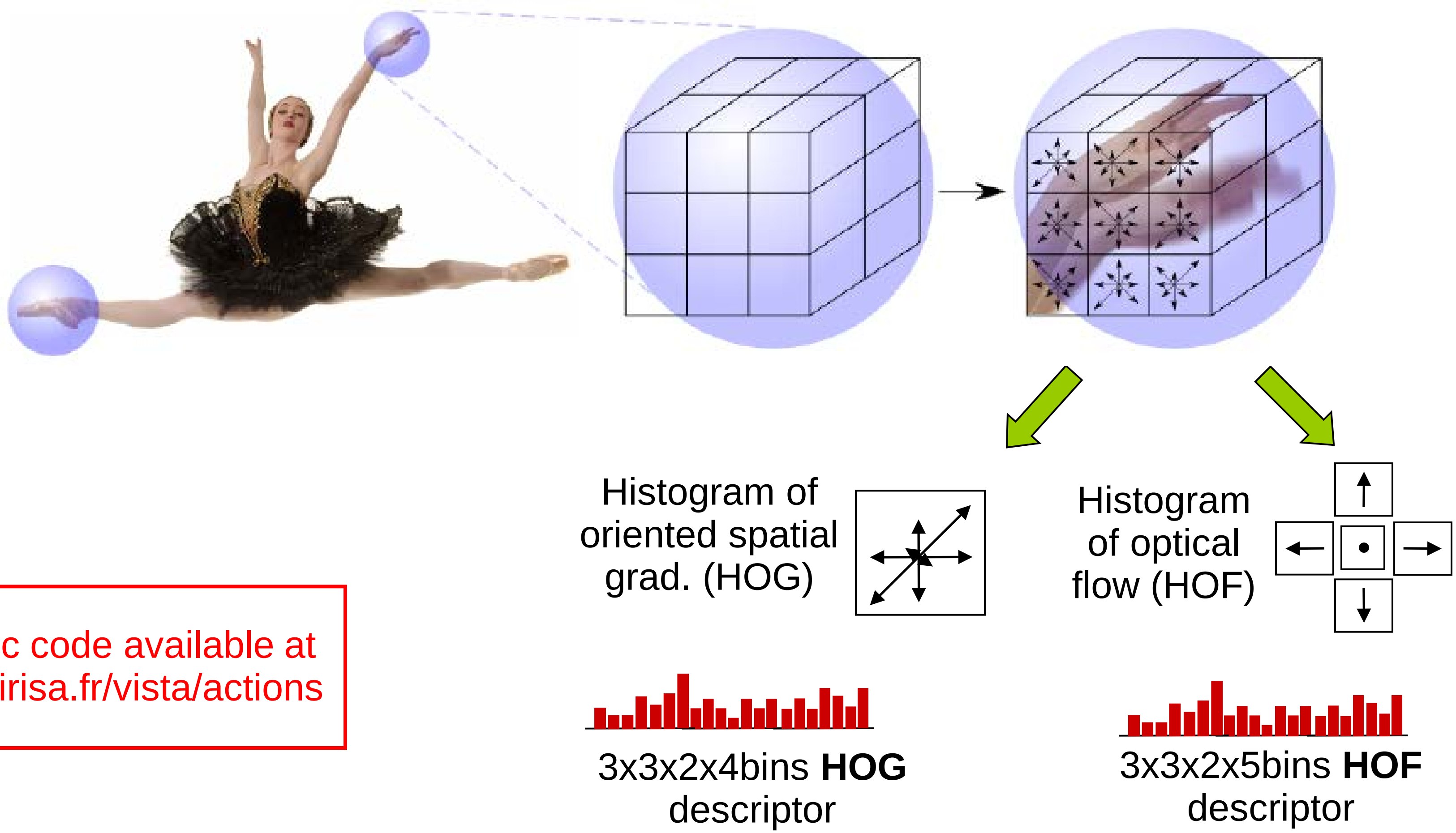
hand waving



[Laptev 2005]

Local space-time descriptor: HOG/HOF

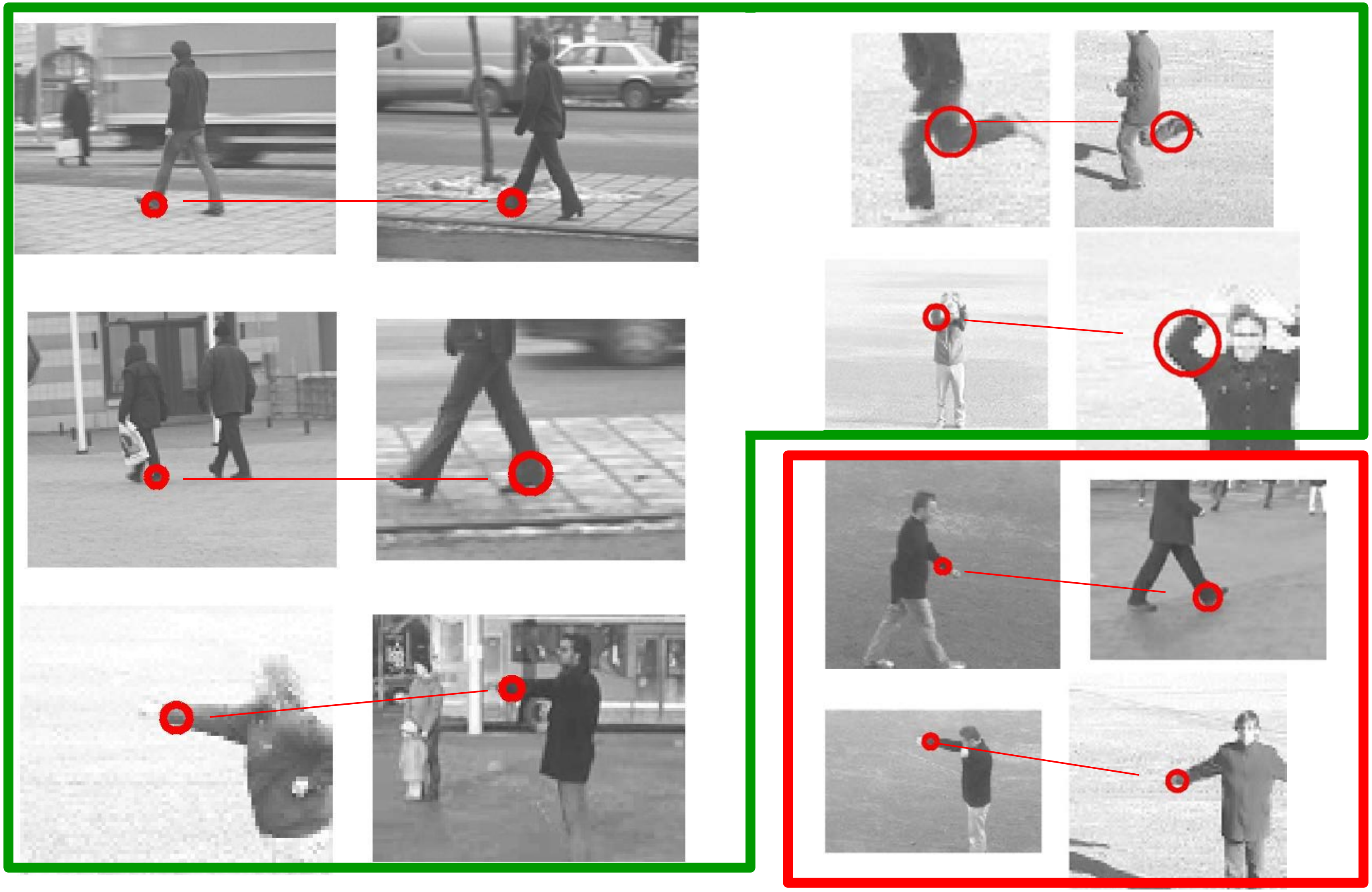
Multi-scale space-time patches



Public code available at www.irisa.fr/vista/actions

Local feature methods: Why working?

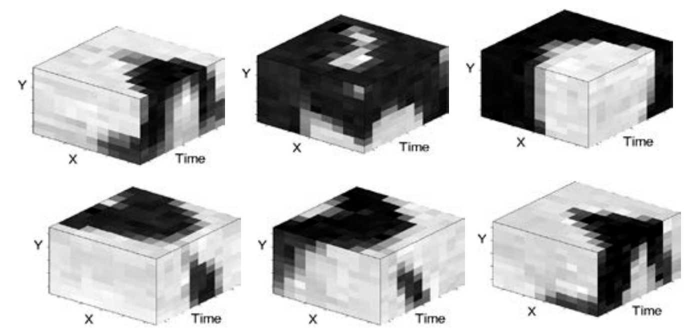
- Finds similar events in pairs of video sequences



Bag-of-Features action recogntion

Extraction of
Local features

space-time patches

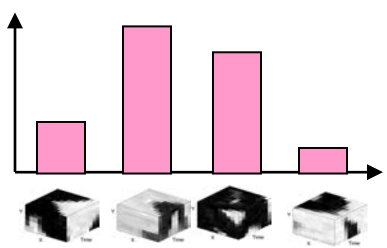


Feature description

K-means clustering (k=4000)

Feature quantization

Occurrence histogram of visual words



Non-linear SVM with χ^2 kernel

[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Action classification in movies

Test episodes from movies “The Graduate”, “It’s a Wonderful Life”,
“Indiana Jones and the Last Crusade” [Laptev et al. CVPR2008]

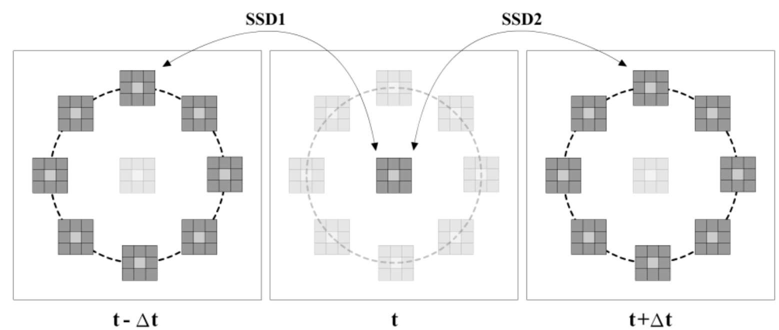
Action classification results

	hoghof		Chance
Channel	bof	flat	
mAP	47.9	50.3	9.2
AnswerPhone	15.7	20.9	7.2
DriveCar	86.6	84.6	11.5
Eat	59.5	67.0	3.7
FightPerson	71.1	69.8	7.9
GetOutCar	29.3	45.7	6.4
HandShake	21.2	27.8	5.1
HugPerson	35.8	43.2	7.5
Kiss	51.5	52.5	11.7
Run	69.1	67.8	16.0
SitDown	58.2	57.6	12.2
SitUp	17.5	17.2	4.2
StandUp	51.7	54.3	16.5

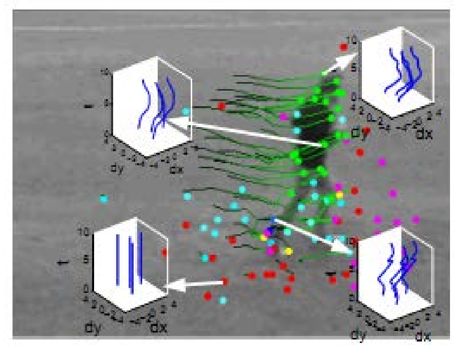
Average precision (AP) for Hollywood-2 dataset

More recent local representations

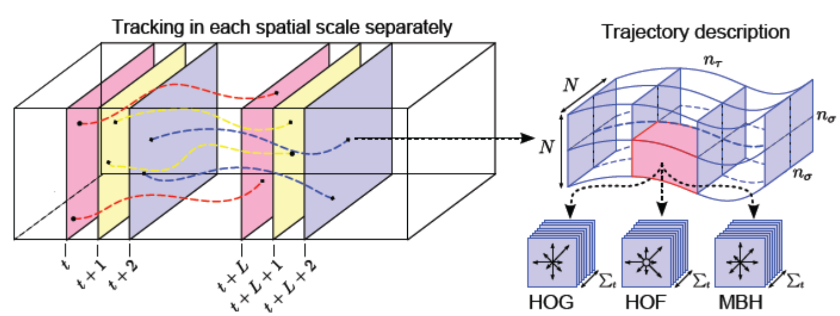
- Y. and L. Wolf, "Local Trinary Patterns for Human Action Recognition ",
  ICCV 2009
  + ECCV 2012 extension



- P. Matikainen, R. Sukthankar and M. Hebert "Trajectons: Action Recognition Through the Motion Analysis of Tracked Features"
  ICCV VOEC Workshop 2009,



- H. Wang, A. Klaser, C. Schmid, C.-L. Liu, "Action Recognition by Dense Trajectories",
  CVPR 2011



Dense trajectory descriptors

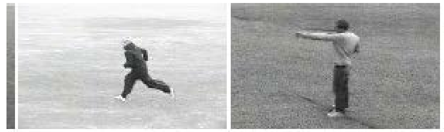
[Wang et al. CVPR'11]

KTH		YouTube		Hollywood2		UCF sports	
Laptev *et al.* [5]	91.8%	Liu *et al.* [45]	71.2%	Wang *et al.* [17]	47.7%	Wang *et al.* [17]	85.6%
Kovashka *et al.* [53]	94.53%	Ikizler-Cinbis *et al.*[35]	75.21%	Taylor *et al.* [58]	46.6%	Kläser *et al.* [59]	86.7%
Yuan *et al.* [60]	93.7%	Brendel *et al.* [51]	77.8%	Ullah *et al.* [43]	53.2%	Kovashka *et al.* [53]	87.27%
Le *et al.* [52]	93.9%	Le *et al.* [52]	75.8%	Gilbert *et al.* [61]	50.9%	Le *et al.* [52]	86.5%
Gilbert *et al.* [61]	94.5%	Bhattacharya *et al.* [62]	76.5%	Le *et al.* [52]	53.3%		
MBH	95.0%	MBH	80.6%	MBH	55.1%	MBH	84.2%
Combined	94.2%	Combined	84.1%	Combined	58.2%	Combined	88.0%
MBH+STP	**95.3%**	MBH+STP	83.0%	MBH+STP	57.6%	MBH+STP	84.0%
Combined+STP	94.4%	Combined+STP	**85.4%**	Combined+STP	**59.9%**	Combined+STP	**89.1%**

IXMAS		UIUC		Olympic Sports		UCF50	
Tran *et al.*[50]	80.22%	Tran *et al.* [50]	98.7%	Brendel *et al.* [56]	77.3%		
Junejo *et al.* [63]	79.6%			Niebles *et al.* [49]	72.1%		
Wu *et al.*[54]	88.2%						
MBH	91.8%	MBH	97.1%	MBH	71.6%	MBH	82.2%
Combined	**93.5%**	Combined	**98.4%**	Combined	74.1%	Combined	84.5%
MBH+STP	91.9%	MBH+STP	**98.1%**	MBH+STP	74.9%	MBH+STP	83.6%
Combined+STP	**93.6%**	Combined+STP	**98.3%**	Combined+STP	**77.2%**	Combined+STP	**85.6%**

Action recognition datasets

- KTH Actions, 6 classes, 2391 video samples [Schuldt et al. 2004]



Running Boxing

- Weizman, 10 classes, 92 video samples, [Blank et al. 2005]

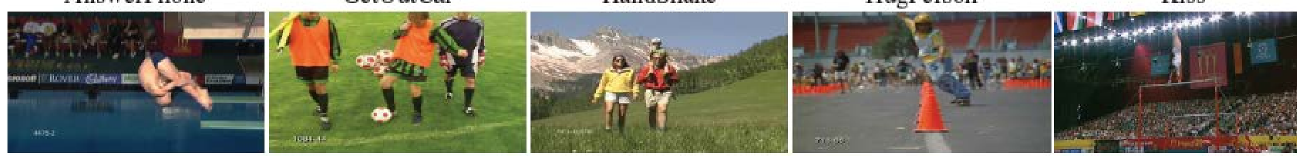
- UCF YouTube, 11 classes, 1168 samples, [Liu et al. 2009]

Biking Shooting Spiking Swinging Walking dog

- Hollywood-2, 12 classes, 1707 samples, [Marszałek et al. 2009]

AnswerPhone GetOutCar HandShake HugPerson Kiss

- UCF Sports, 10 classes, 150 samples, [Rodriguez et al. 2008]



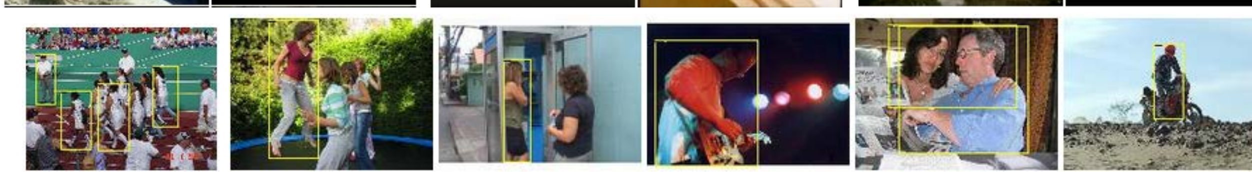
Diving Kicking Walking Skateboarding High-Bar-Swinging

- Olympic Sports, 16 classes, 783 samples, [Niebles et al. 2010]

springboard snatch clean-jerk vault bowling tennis-serve

- HMDB, 51 classes, ~7000 samples, [Kuehne et al. 2011]

- PASCAL VOC 2011 Action Classification Challenge, 10 classes, 3375 image samples



Where to go next?

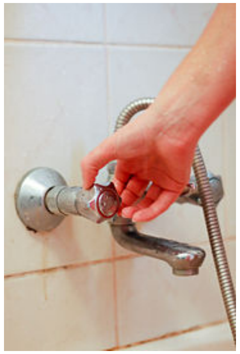
Is action classification the right problem?

- Is action vocabulary well-defined?

Examples of “Open” action:





- What granularity of action vocabulary shall we consider?

Source: http://www.youtube.com/watch?v=eYdUZdan5i8

Do we want to learn *person-throws-cat-into-trash-bin* classifier?

Limitations of Current Methods

What is unusual in this scene? *Is this scene dangerous?* *What is intention of this person?*

Next challenge

Shift the focus of computer vision

Object, scene and action recognition

Recognition of objects' function and people's intentions

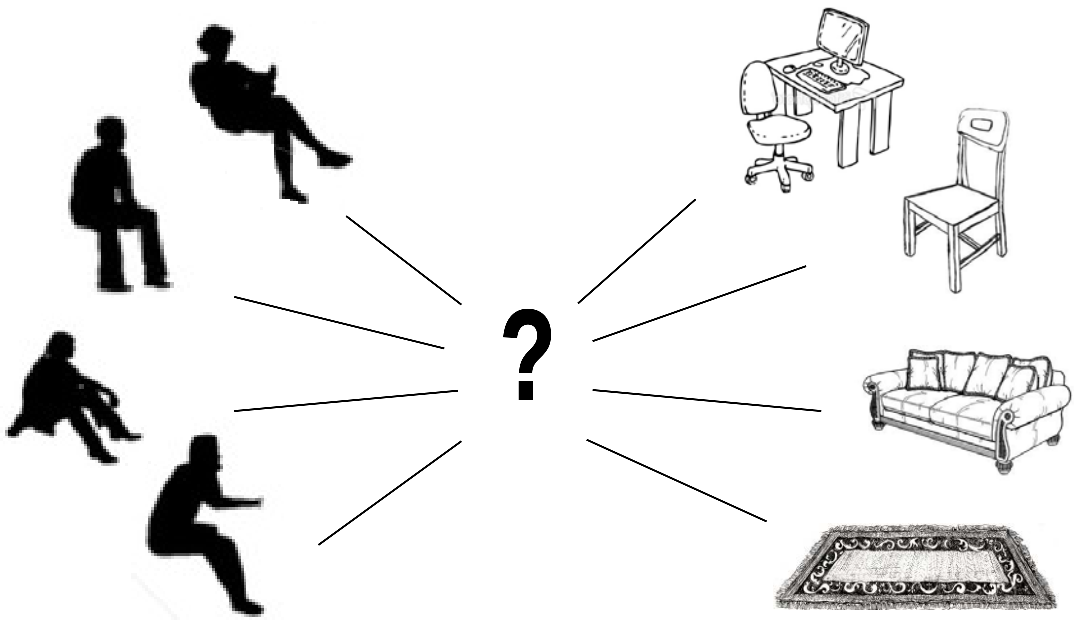
Is this a picture of a dog?
Is the person running in this video?

What people do with objects?
How they do it?
For what purpose?

Enable new applications

Motivation

- Exploit the link between human pose, action and object function.



- Use human actors as active sensors to reason about the surrounding scene.

Scene semantics from long-term observation of people

ECCV 2012

V. Delaitre, D. F. Fouhey, I. Laptev,
J. Sivic, A. Gupta, A. Efros

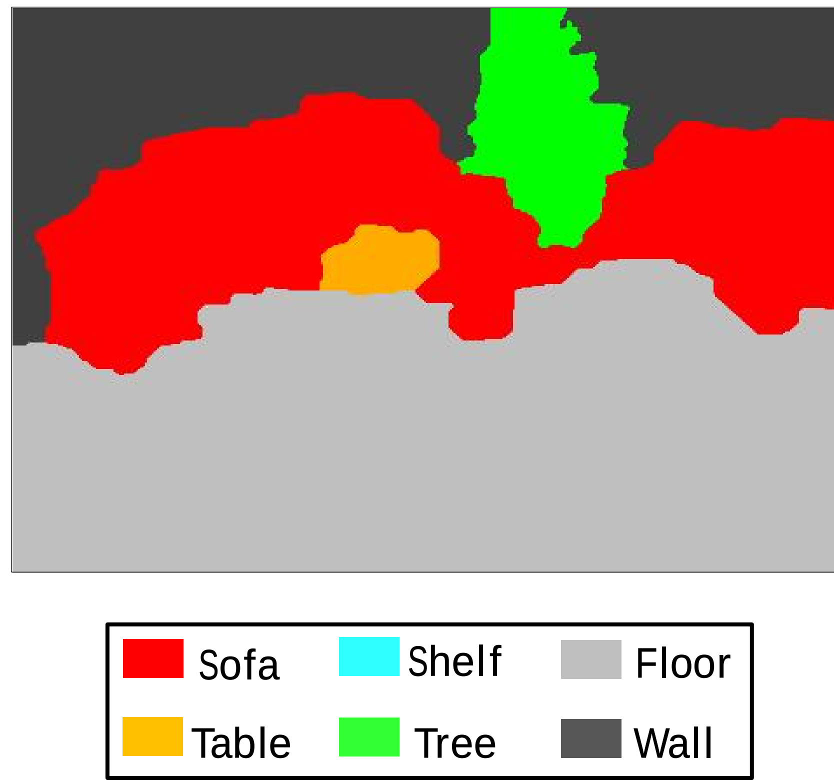
Goal

Recognize objects by the way people interact with them.

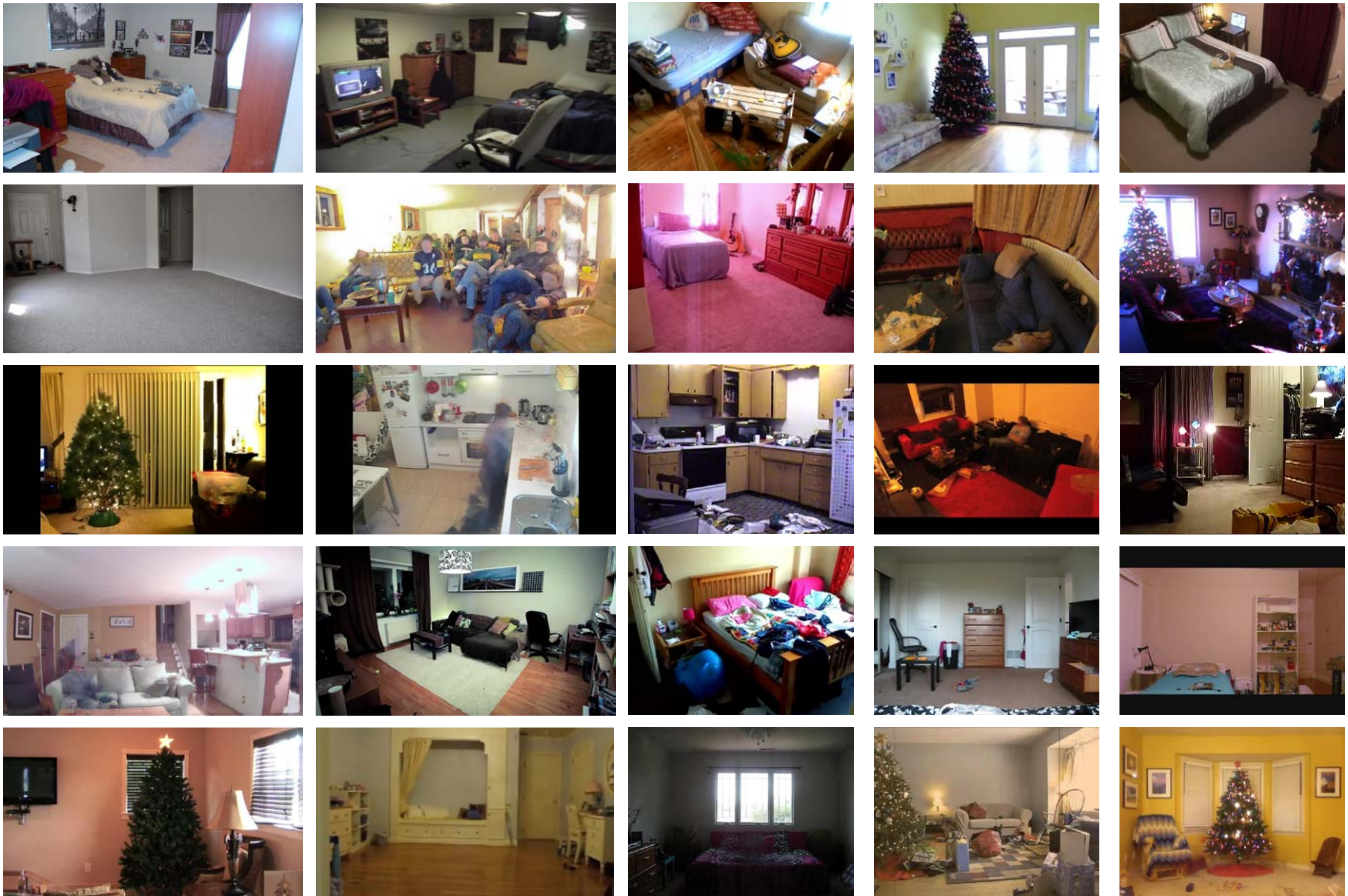
Time-lapse “Party & Cleaning” videos

Lots of person-object interactions, many scenes on YouTube

Semantic object segmentation



New “Party & Cleaning” dataset



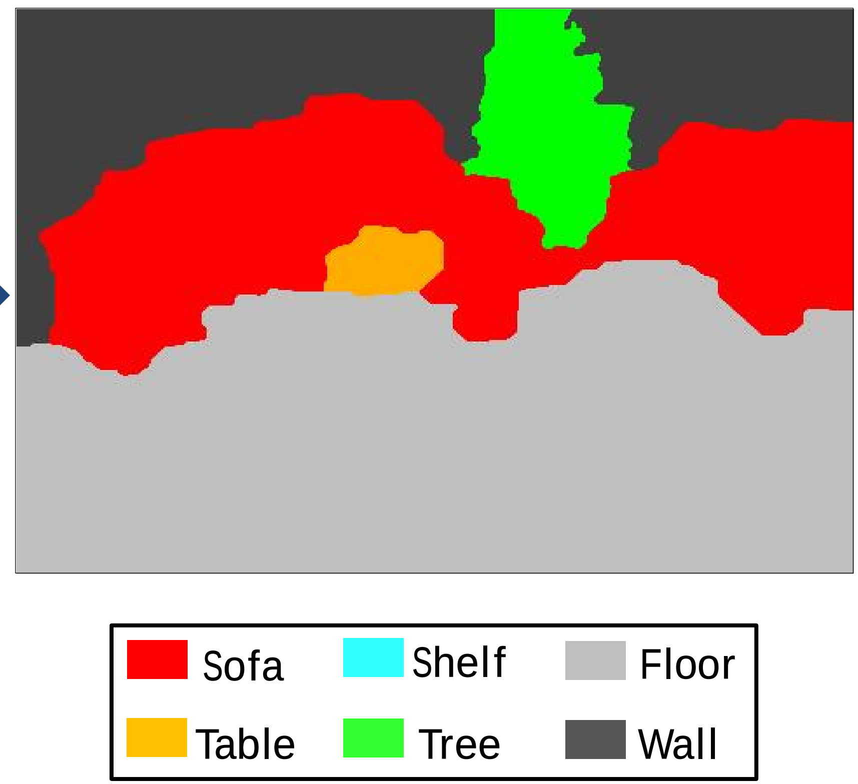
Goal

Recognize objects by the way people interact with them.

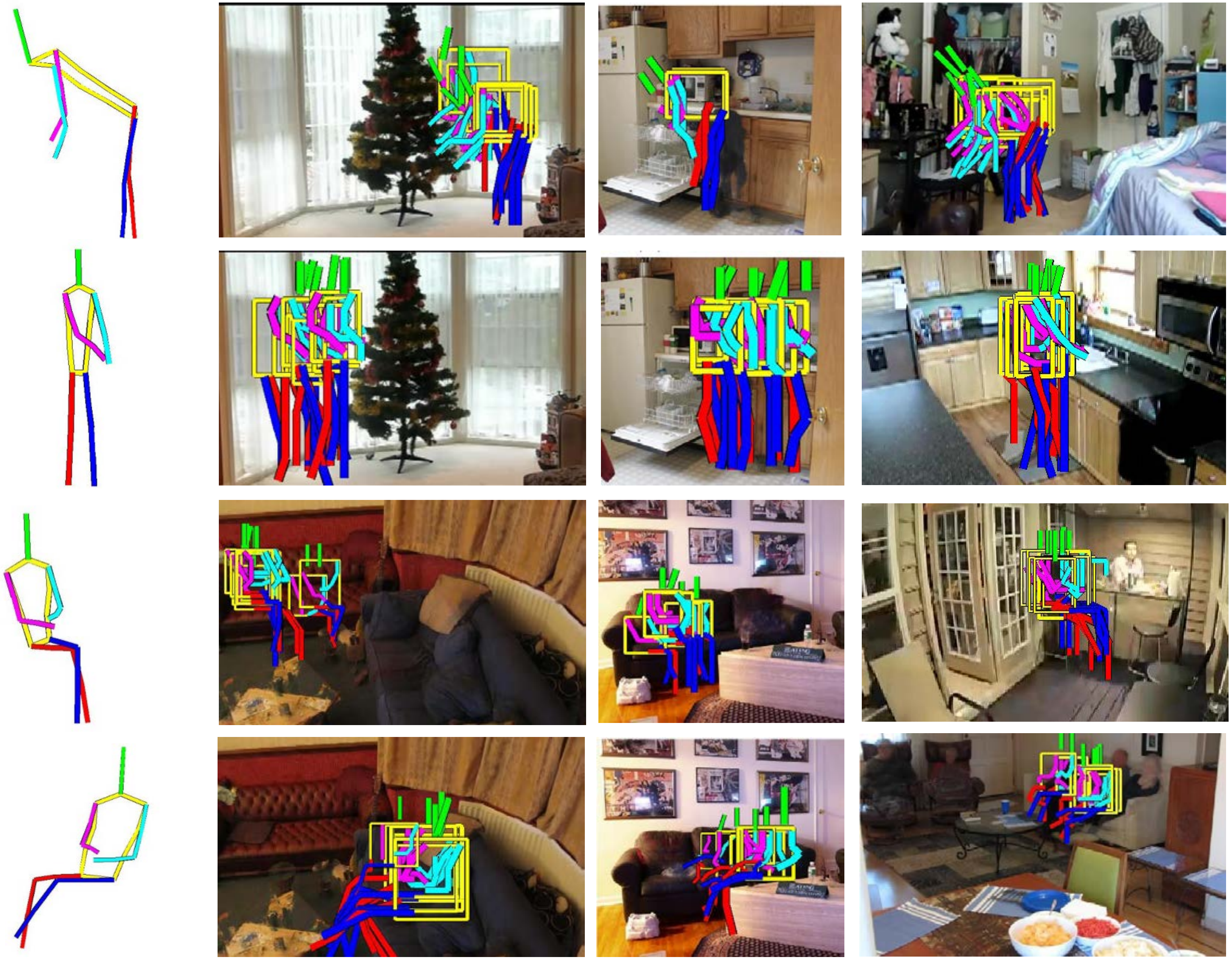
Time-lapse "Party & Cleaning" videos

Lots of person-object interactions, many scenes on YouTube

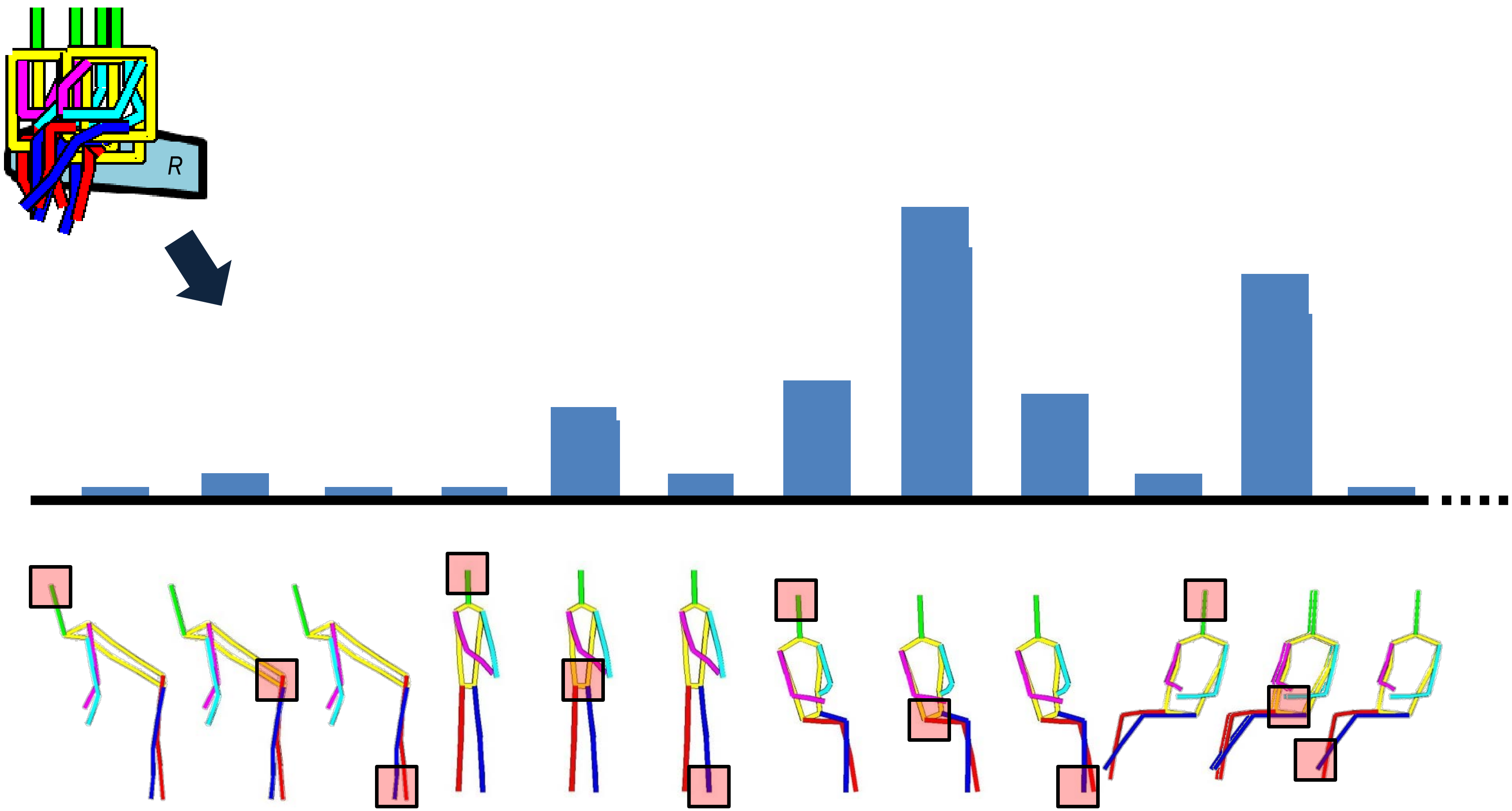
Semantic object segmentation



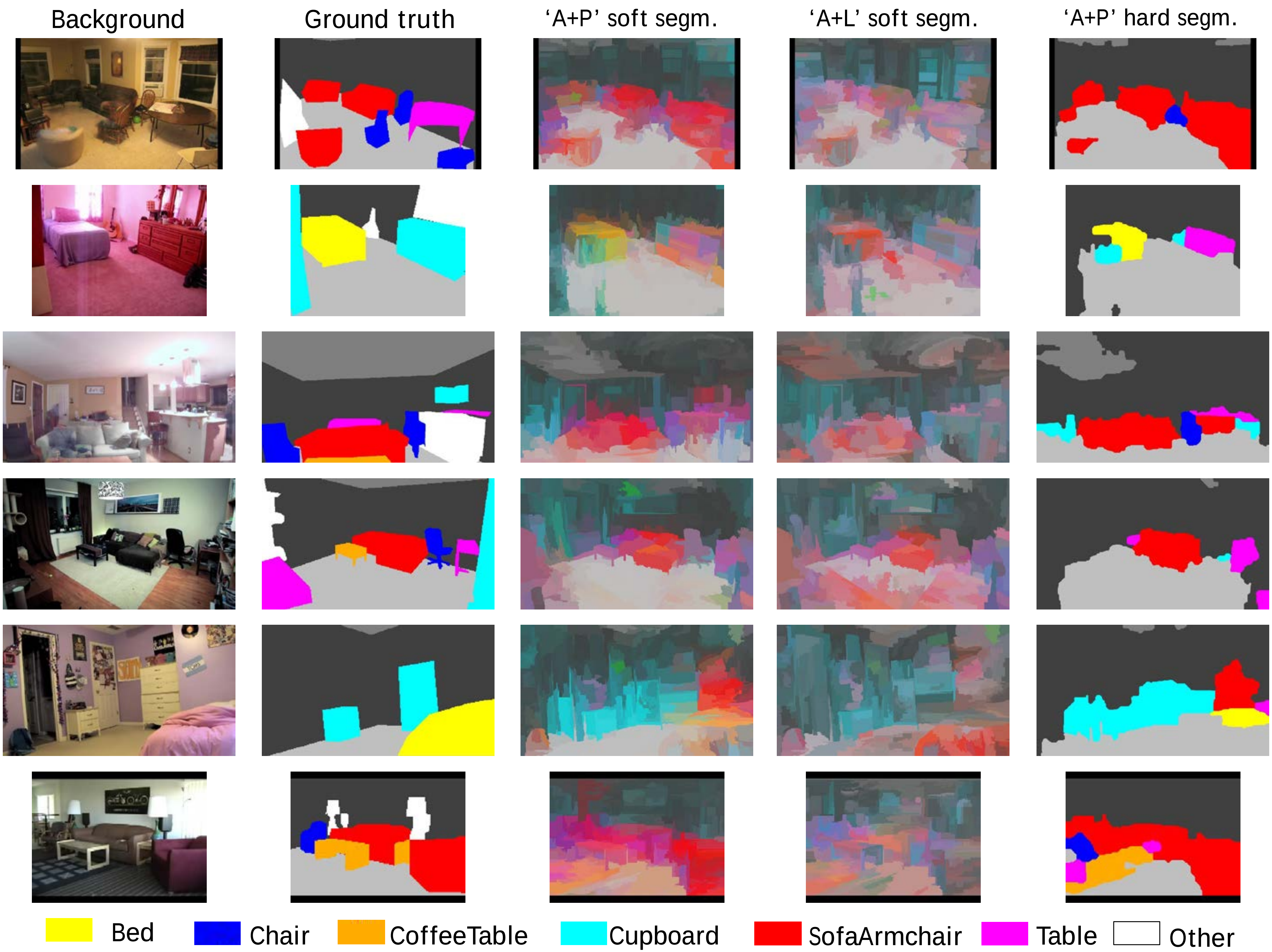
Pose vocabulary



Pose histogram



Some qualitative results


Background
Ground truth
'A+P' soft segm.
'A+L' soft segm.
'A+P' hard segm.
Bed
Chair
CoffeeTable
Cupboard
SofaArmchair
Table
Other

Quantitative results

	DPM	Hedau	(A+L)	(P)	(A+P)	(A+L+P)
Wall	—	75±3.9	76±1.6	76±1.7	**82**±1.2	81±1.3
Ceiling	—	47±20	53±8.0	52±7.4	**69**±6.7	**69**±6.6
Floor	—	59±3.1	64±5.5	65±3.6	**76**±3.2	**76**±2.9
Bed	**31**±20	12±7.2	14±5.0	21±5.8	27±13	26±13
Sofa/Armchair	26±9.4	26±10	34±3.3	32±6.5	**44**±5.4	43±5.8
Coffee Table	11±5.4	11±5.2	11±4.4	12±4.3	**17**±10	**17**±9.6
Chair	9.5±3.9	6.3±2.8	8.3±2.7	5.8±1.4	11±5.4	**12**±5.9
Table	15±6.4	18±3.8	17±3.9	16±7.1	**22**±6.2	**22**±6.4
Wardrobe/Cupboard	27±10	27±8.2	28±6.4	22±1.1	**36**±7.4	**36**±7.2
Christmas tree	50±3.3	55±12	72±1.8	20±6.0	76±6.2	**77**±5.5
Other Object	12±6.4	11±1.2	7.9±1.9	13±4.2	**16**±8.3	**16**±8.2
Average	23±1.8	31±2.0	35±2.4	30±1.7	**43**±4.4	**43**±4.3

A: Appearance (SIFT) histograms;

L: Location;

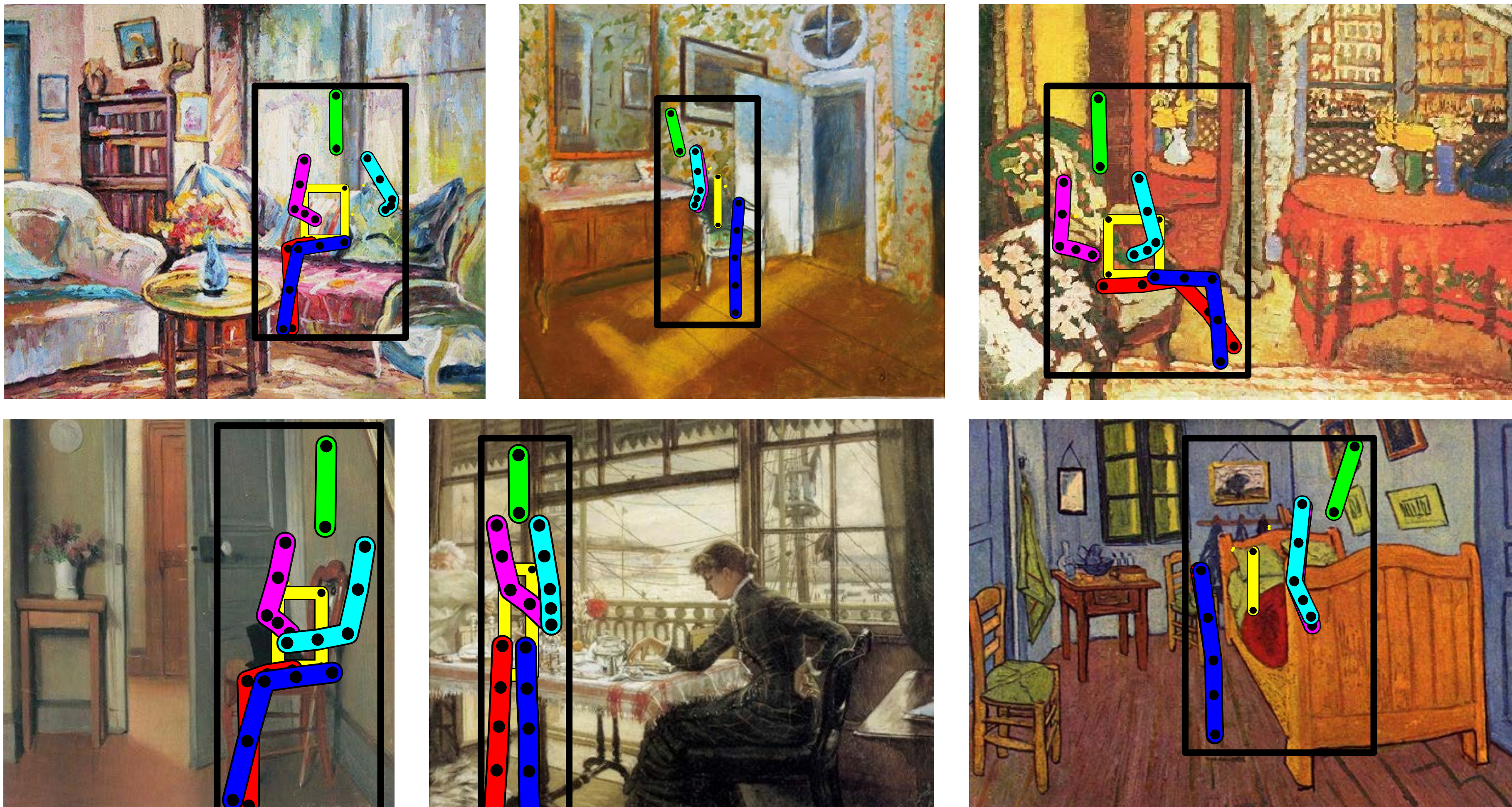
P: Pose histograms

Hedau: Hedau et al., Recovering the spatial layout of cluttered rooms. In: ICCV. (2009)

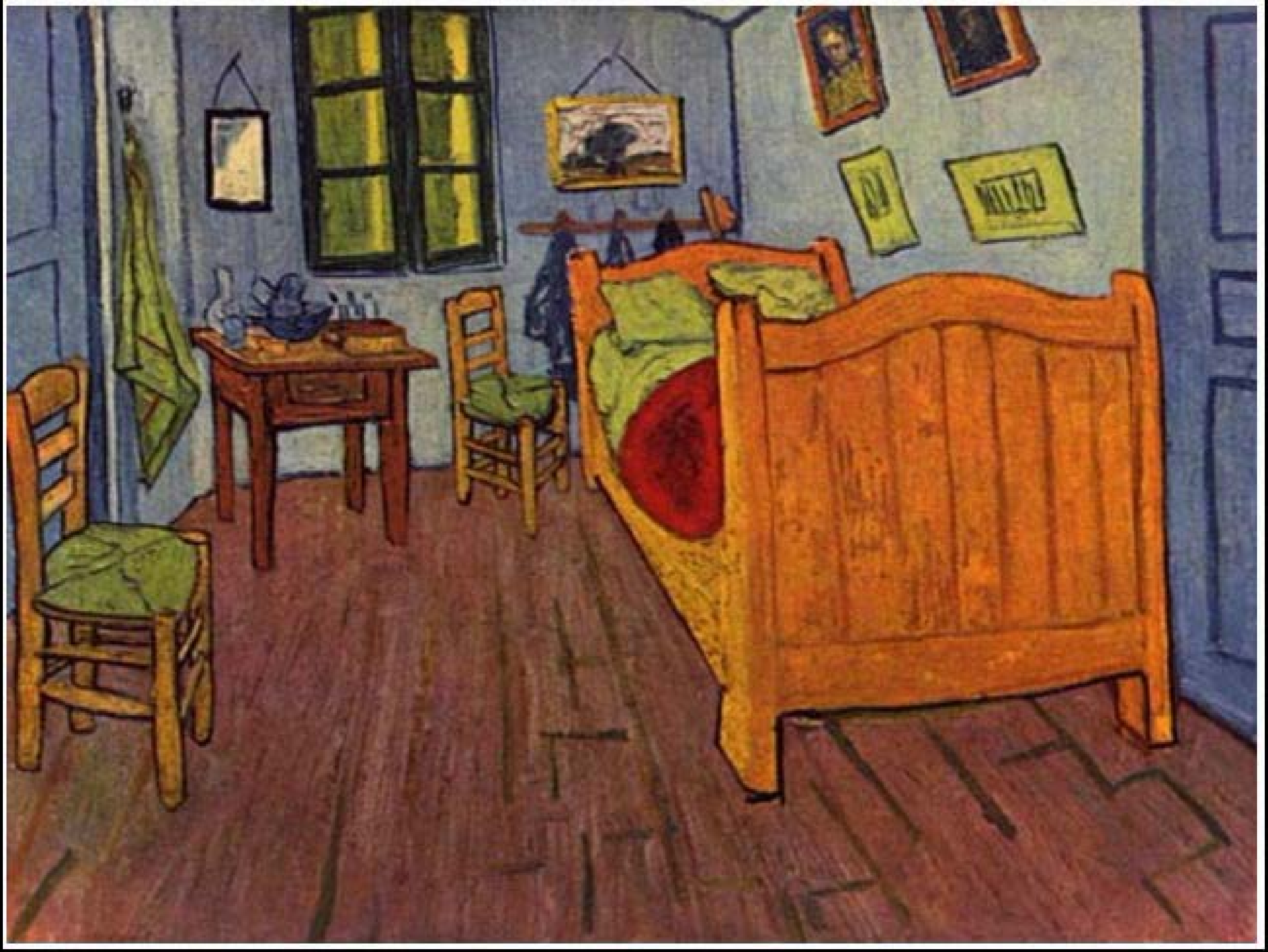
DPM: Felzenszwalb et al., Object detection with discriminatively trained part based models. PAMI (2010)

Using our model as pose prior

Given a bounding box and the ground truth segmentation, we fit the pose clusters in the box and score them by summing the joint's weight of the underlying objects.



Input image



Conclusions

- BOF methods give state-of-the-art results for action recognition in realistic data. Better models are needed

- Action classification (and temporal action localization) are often ill-defined problems

- Targeting more realistic problems with functional models of objects and scenes can be the next challenge.

Willow, Paris