ENS/INRIA CVML Summer School
45 rue d'Ulm, Paris
July 26, 2013

Modeling and visual recognition of human actions

Ivan Laptev

ivan.laptev@inria.fr

WILLOW, INRIA/ENS/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…

Human Actions:
Why do we care?

Why video analysis?

Data:

BBC Motion Gallery

ina

TV-channels recorded since 60's

YouTube Broadcast Yourself

>34K hours of video uploads every day

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

GLASS

Why video analysis?

Applications:

First appearance of
N. Sarkozy on TV

Sociology research:
Influence of character
smoking in movies

Education: How do I
make a pizza?

Where is my cat?

Predicting crowd behavior
Counting people

Motion capture and animation

Why human actions?

How many person-pixels are in the video?

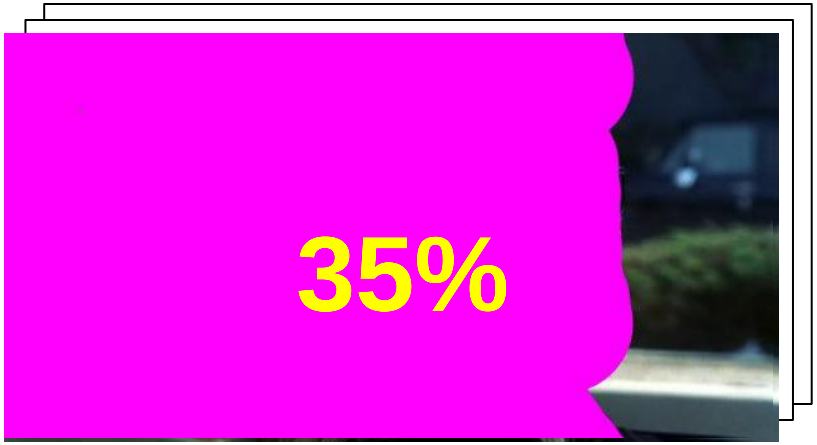
Movies

TV

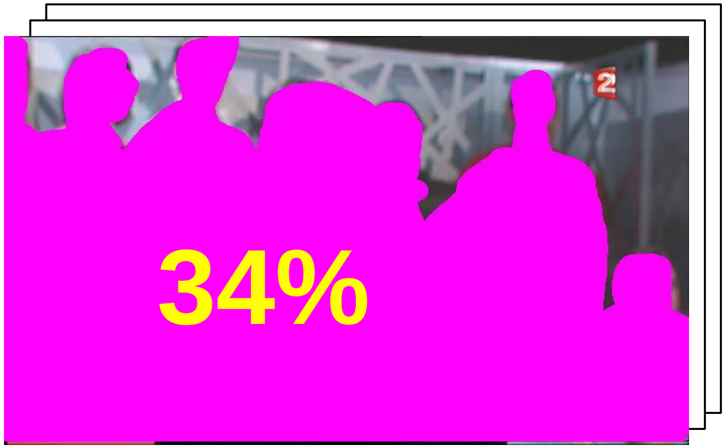
YouTube

Why human actions?

How many person-pixels are in the video?



Movies



TV

YouTube

How many person pixels in our daily life?

- Wearable camera data: Microsoft SenseCam dataset

How many person pixels in our daily life?

- Wearable camera data: Microsoft SenseCam dataset

Why do we prefer to watch other people?

- Why do we watch TV, Movies, … at all?
- Why do we read books?

"… books teach us new patterns of behavior…"

Olga Slavnikova
Russian journalist and writer

Why action recognition is difficult?

Challenges

- **Large variations in appearance:** occlusions, non-rigid motion, view-point changes, clothing…

Action *Hugging*:

…

- **Manual collection of training samples is prohibitive:** many action classes, rare occurrence

…

- **Action vocabulary is not well-defined**

Action *Open*:

…

How to recognize actions?

A HOUGHTON MIFFLIN PRODUCTION

A Teaching Resource

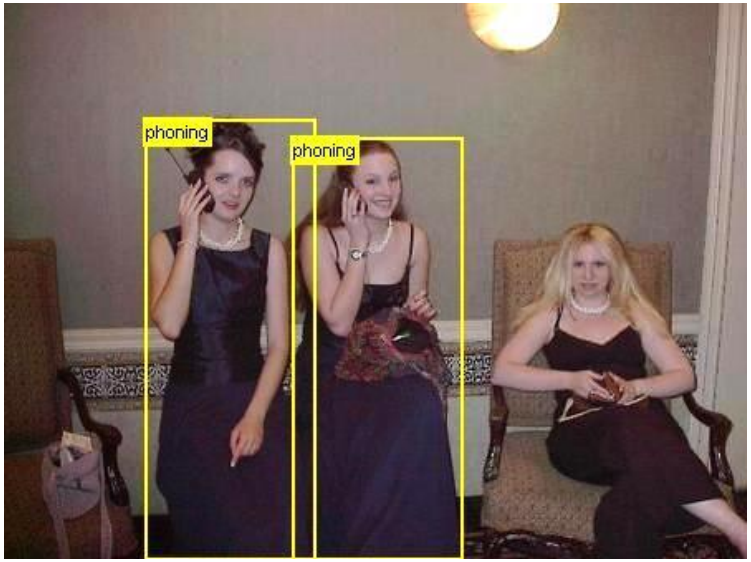
At the Frontiers of Psychological Inquiry

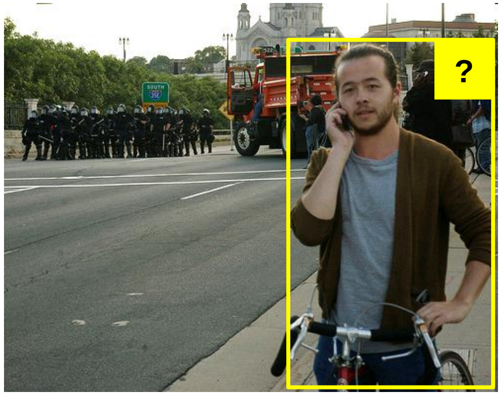
Activities characterized by a pose

Slide credit: A. Zisserman

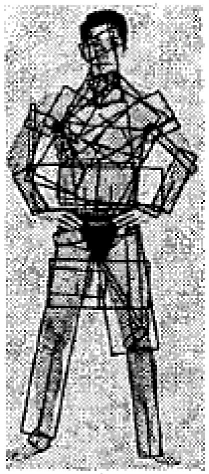
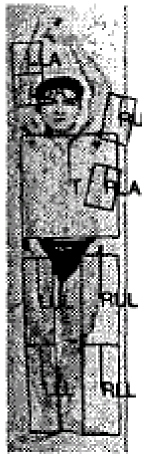
Activities characterized by a pose

Examples from VOC action recognition challenge



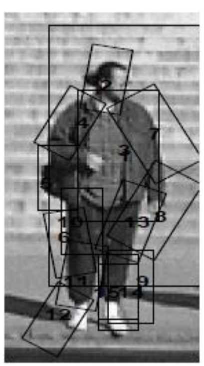


Human pose estimation (1990-2000)

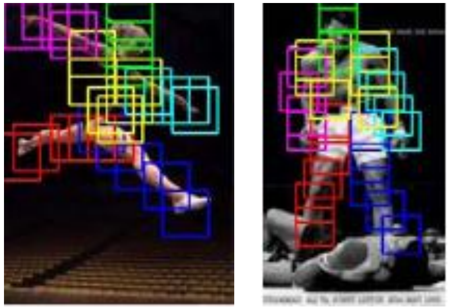
Finding People by Sampling
Ioffe & Forsyth, ICCV 1999

Pictorial Structure Models for Object Recognition
Felzenszwalb & Huttenlocher, 2000

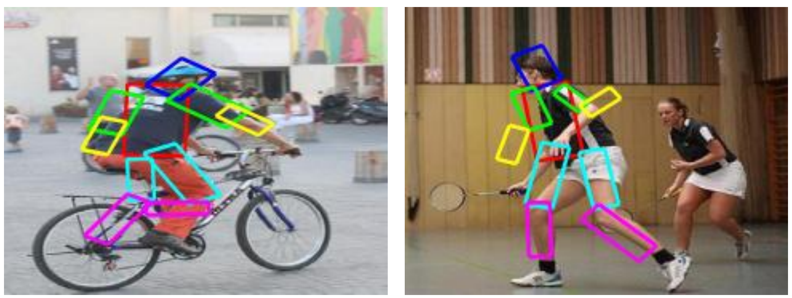
Learning to Parse Pictures of People
Ronfard, Schmid & Triggs, ECCV 2002

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.

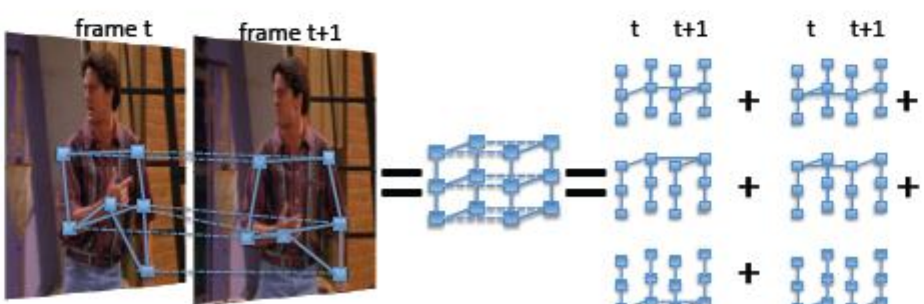


Y. Wang, D. Tran and Z. Liao. Learning Hierarchical Poselets for Human Parsing. In Proc. **CVPR 2011**.

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**.

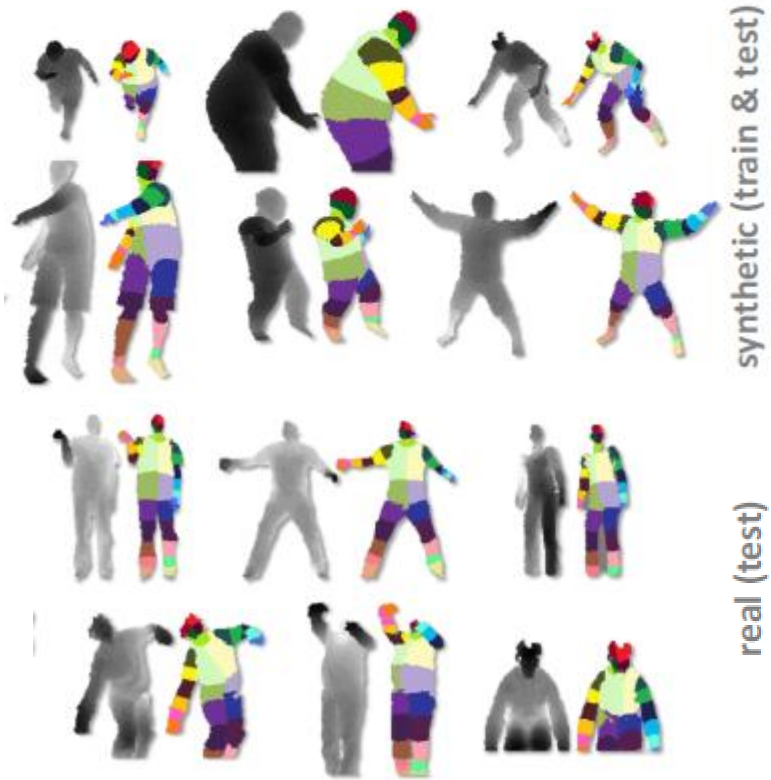
Learns from lots of noisy annotations

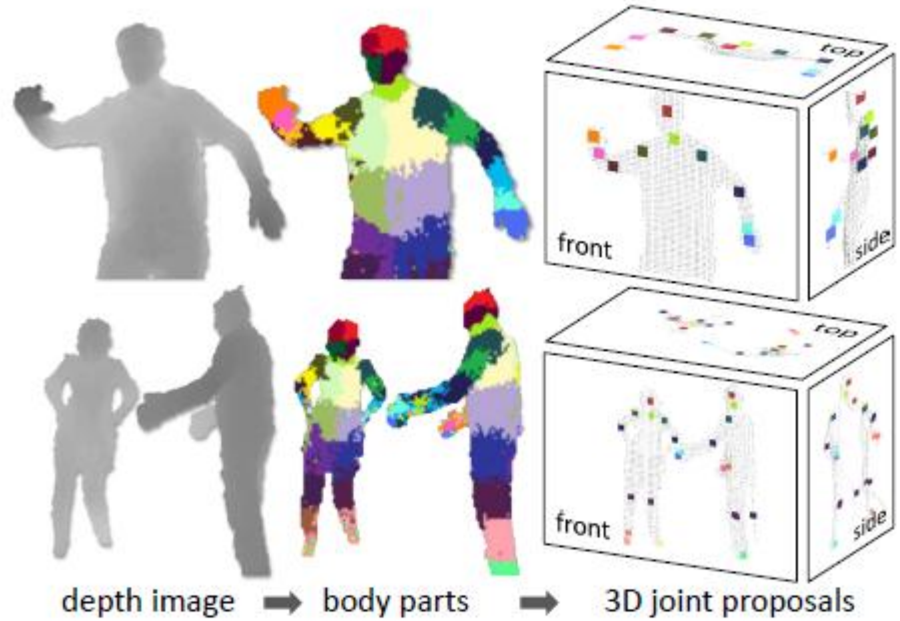


B. Sapp, D.Weiss and B. Taskar. Parsing Human Motion with Stretchable Models. In Proc. **CVPR 2011**.

Explores temporal continuity

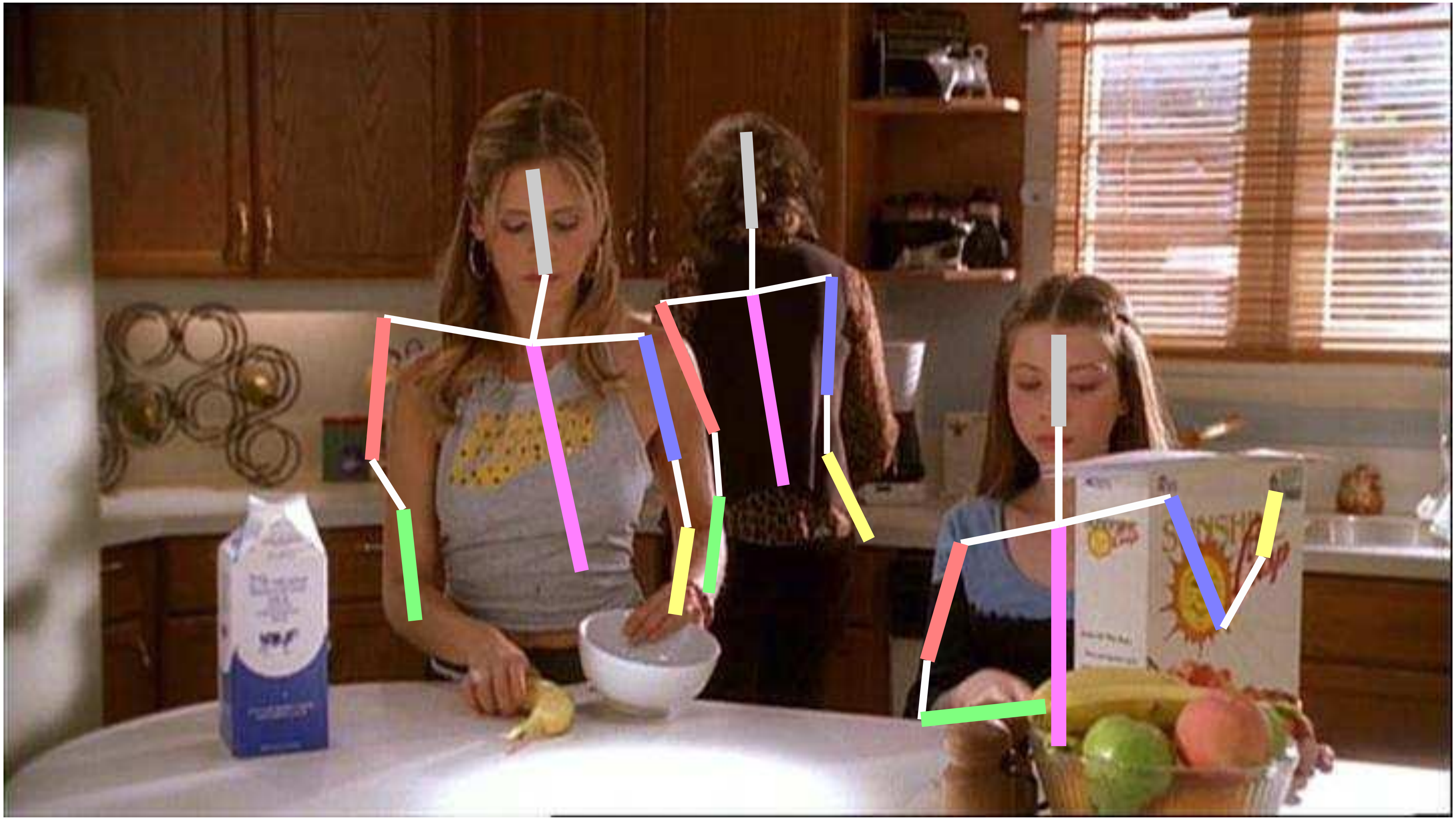
Human pose estimation





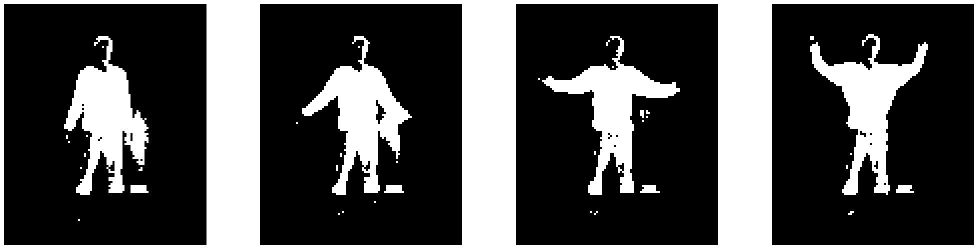
J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. (Best paper award at CVPR 2011)

Pose estimation is still a hard problem



Issues:
- occlusions
- clothing and pose variations

Appearance methods: 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

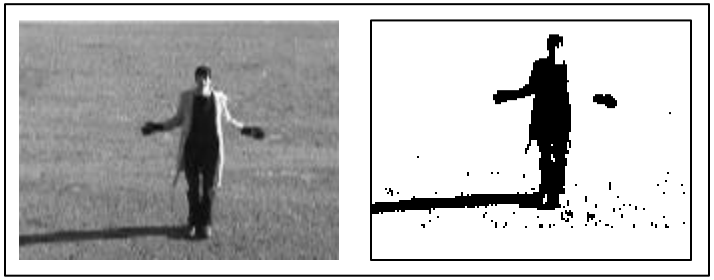
Appearance methods: Shape

Pros:

+ Simple and fast
+ Works in controlled settings

Cons:

- Prone to errors of background subtraction



Variations in light, shadows, clothing…

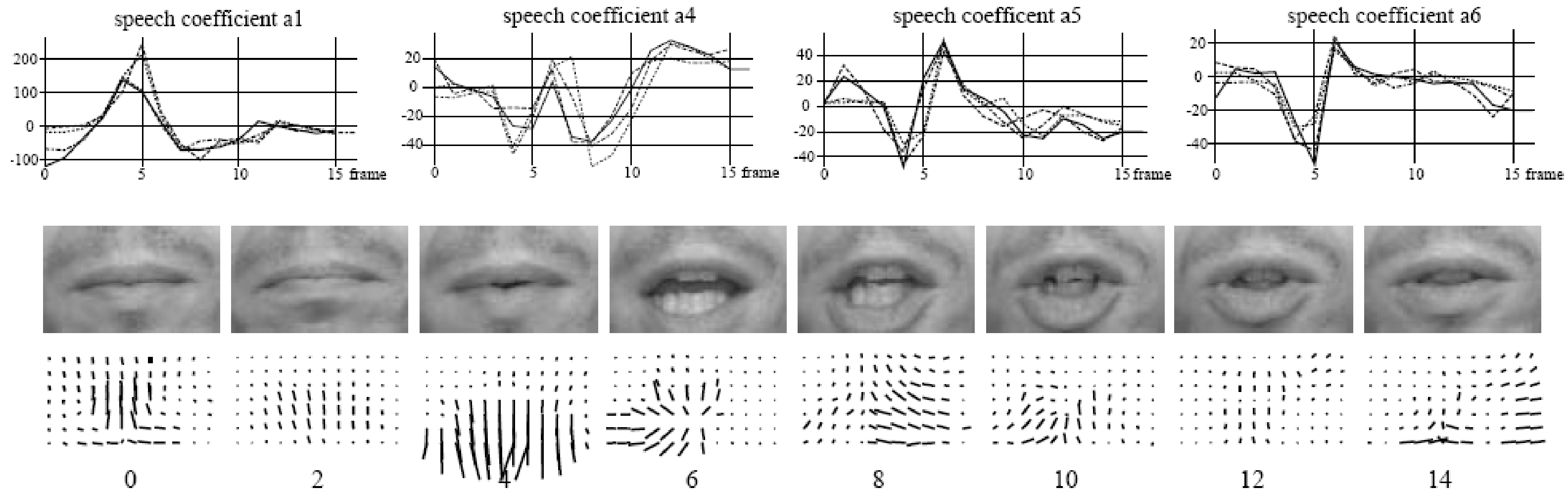
What is the background here?

- Does not capture *interior* Structure and motion

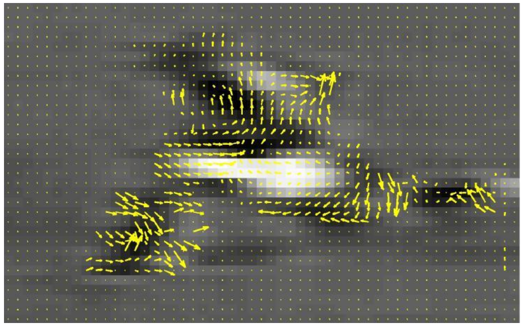
Silhouette tells little about actions

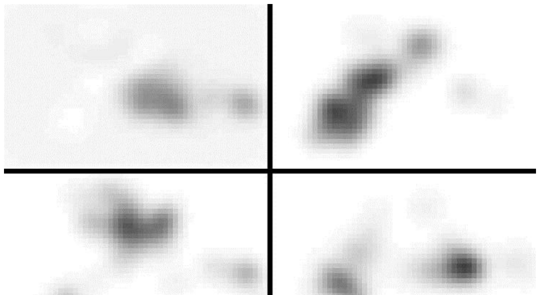
Appearance methods: Motion

Learning Parameterized Models of Image Motion
M.J. Black, Y. Yacoob, A.D. Jepson and D.J. Fleet, 1997



Recognizing action at a distance
A.A. Efros, A.C. Berg, G. Mori, and J. Malik., 2003.



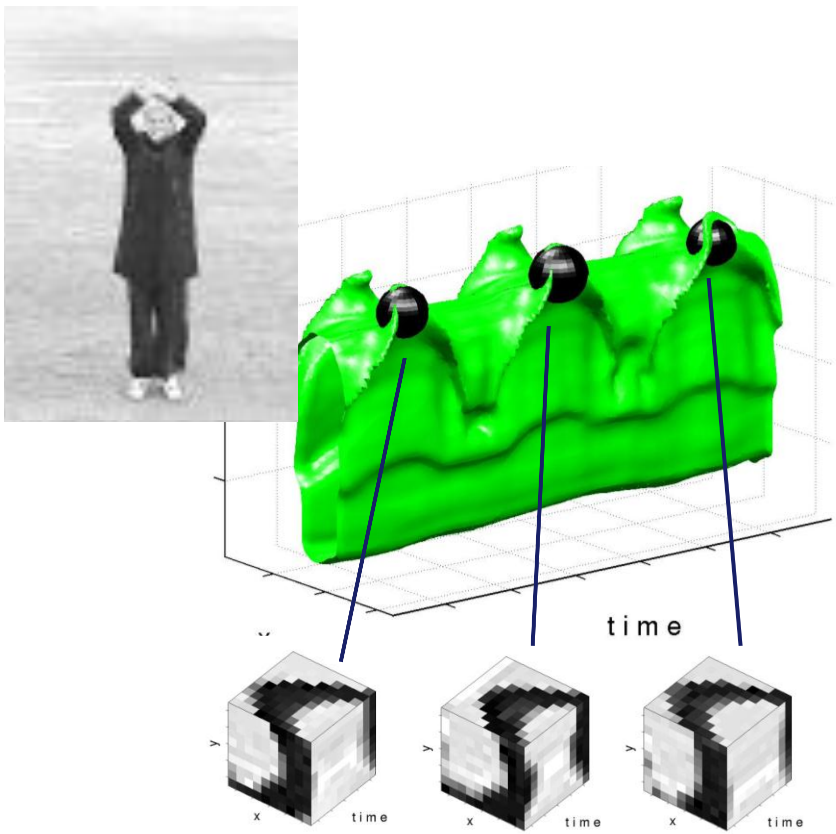


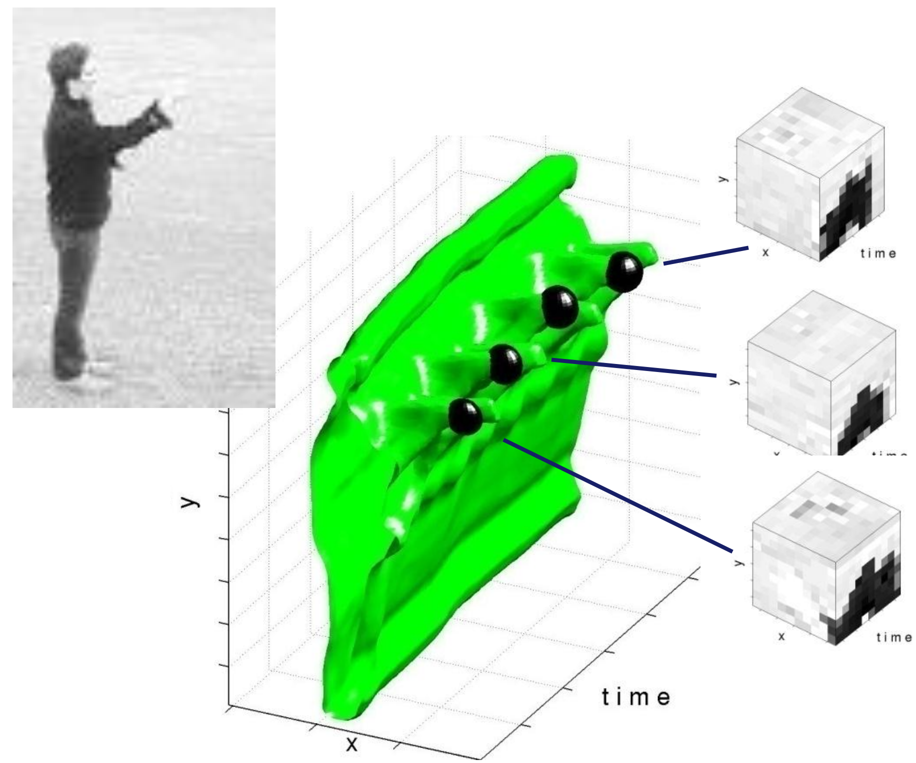
blurred $F_x^-, F_x^+, F_y^-, F_y^+$

Action recognition with local features

Local space-time features

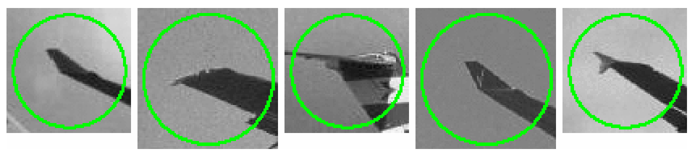
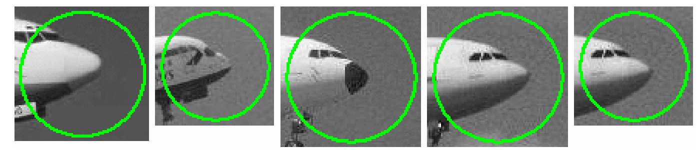
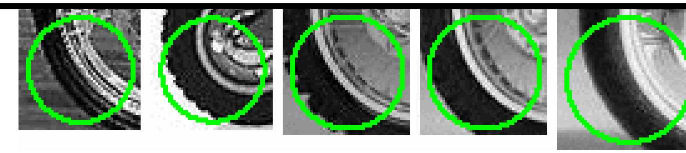
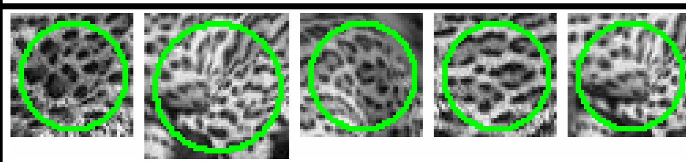
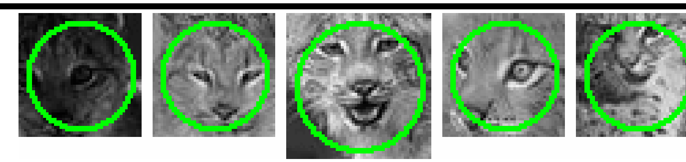
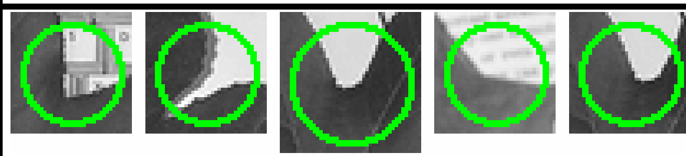
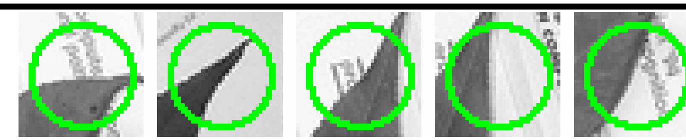
+ No segmentation needed
+ No object detection/tracking needed
− Loss of global structure





[Laptev 2005]

Local approach: Bag of Visual Words

Airplanes		
Motorbikes		
Faces		
Wild Cats		
Leaves		
People		
Bikes		

Space-Time Interest Points: Detection

What neighborhoods to consider?

Distinctive neighborhoods $\Rightarrow$ High image variation in space and time $\Rightarrow$ Look at the distribution of the gradient

Definitions:

$f: \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}$ Original image sequence

$g(x, y, t;\ \Sigma)$ Space-time Gaussian with covariance $\Sigma \in \mathrm{SPSD}(3)$

$L_\xi(\cdot;\ \Sigma) = f(\cdot) * g_\xi(\cdot;\ \Sigma)$ Gaussian derivative of f

$\nabla L = (L_x, L_y, L_t)^T$ Space-time gradient

$$\mu(\cdot;\ \Sigma) = \nabla L(\cdot;\ \Sigma)(\nabla L(\cdot;\ \Sigma))^T * g(\cdot;\ s\Sigma) = \begin{pmatrix} \mu_{xx} & \mu_{xy} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix}$$

Second-moment matrix

[Laptev 2005]

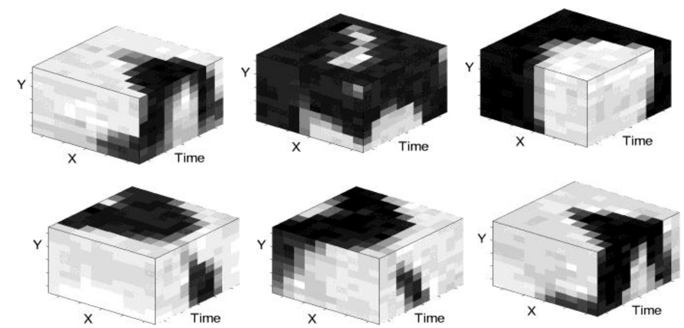
Local features: Proof of concept

- Finds similar events in pairs of video sequences

Bag-of-Features action recogntion

Extraction of Local features

space-time patches

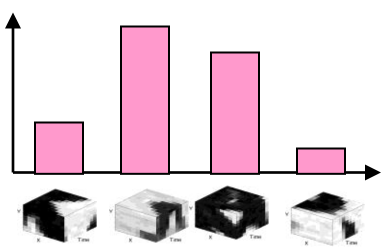


Feature description

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 results

KTH dataset

	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	.99	.01	.00	.00	.00	.00
Jogging	.04	.89	.07	.00	.00	.00
Running	.01	.19	.80	.00	.00	.00
Boxing	.00	.00	.00	.97	.00	.03
Waving	.00	.00	.00	.00	.91	.09
Clapping	.00	.00	.00	.05	.00	.95

Hollywood-2 dataset

	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

[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Action classification

Test episodes from movies “The Graduate”, “It’s a Wonderful Life”,
“Indiana Jones and the Last Crusade”

Evaluation of local feature detectors and descriptors

Four types of detectors:

- Harris3D [Laptev 2003]
- Cuboids [Dollar et al. 2005]
- Hessian [Willems et al. 2008]
- Regular dense sampling

Four types of descriptors:

- HoG/HoF [Laptev et al. 2008]
- Cuboids [Dollar et al. 2005]
- HoG3D [Kläser et al. 2008]
- Extended SURF [Willems’et al. 2008]

Three human actions datasets:

- KTH actions [Schuldt et al. 2004]
- UCF Sports [Rodriguez et al. 2008]
- Hollywood 2 [Marszałek et al. 2009]

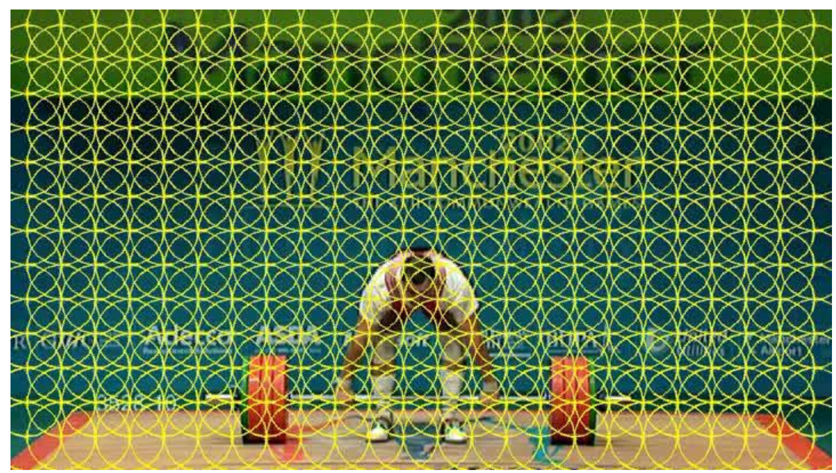
Space-time feature detectors

Harris3D

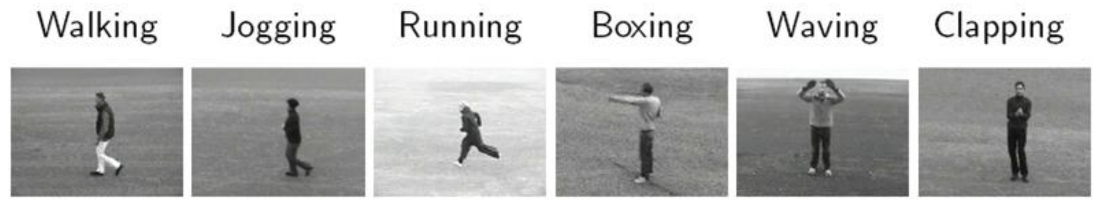
Hessian

Cuboids

Dense



Results on KTH Actions



6 action classes, 4 scenarios, staged

Detectors

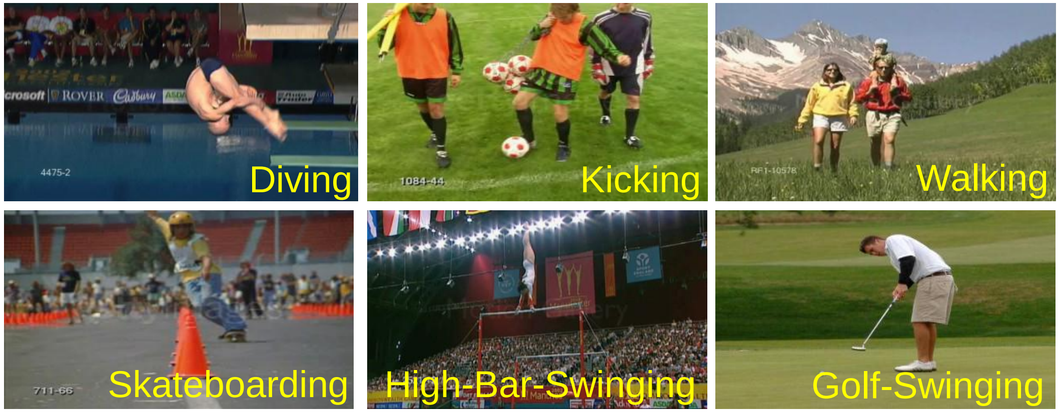
Descriptors	Harris3D	Cuboids	Hessian	Dense
HOG3D	89.0%	90.0%	84.6%	85.3%
HOG/HOF	91.8%	88.7%	88.7%	86.1%
HOG	80.9%	82.3%	77.7%	79.0%
HOF	**92.1%**	88.2%	88.6%	88.0%
Cuboids	-	89.1%	-	-
E-SURF	-	-	81.4%	-

(Average accuracy scores)

- Best results for **sparse** Harris3D + HOF
- Dense features perform relatively poor compared to sparse features

[Wang, Ullah, Kläser, Laptev, Schmid, 2009]

Results on UCF Sports



10 action classes, videos from TV broadcasts

Detectors

Descriptors	Harris3D	Cuboids	Hessian	Dense
HOG3D	79.7%	82.9%	79.0%	**85.6%**
HOG/HOF	78.1%	77.7%	79.3%	81.6%
HOG	71.4%	72.7%	66.0%	77.4%
HOF	75.4%	76.7%	75.3%	82.6%
Cuboids	-	76.6%	-	-
E-SURF	-	-	77.3%	-

(Average precision scores)

- Best results for **dense** + HOG3D

[Wang, Ullah, Kläser, Laptev, Schmid, 2009]

Results on Hollywood-2

12 action classes collected from 69 movies

Detectors

Descriptors

	Harris3D	Cuboids	Hessian	Dense
HOG3D	43.7%	45.7%	41.3%	45.3%
HOG/HOF	45.2%	46.2%	46.0%	**47.4%**
HOG	32.8%	39.4%	36.2%	39.4%
HOF	43.3%	42.9%	43.0%	45.5%
Cuboids	-	45.0%	-	-
E-SURF	-	-	38.2%	-

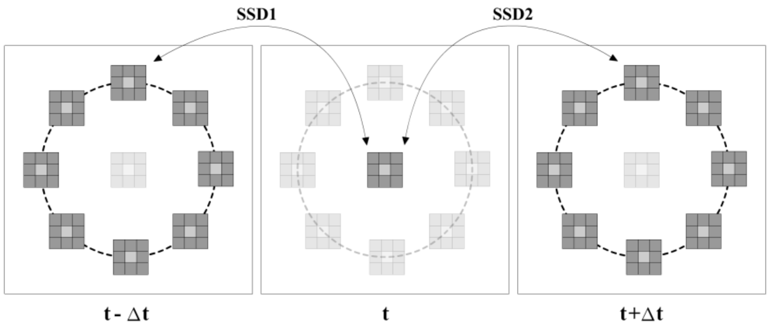
(Average precision scores)

- Best results for **dense** + HOG/HOF

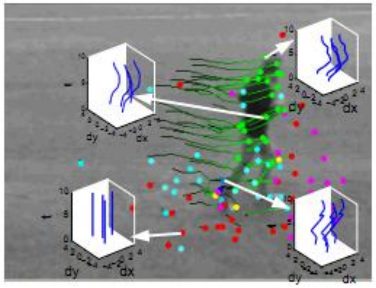
[Wang, Ullah, Kläser, Laptev, Schmid, 2009]

Other recent local representations

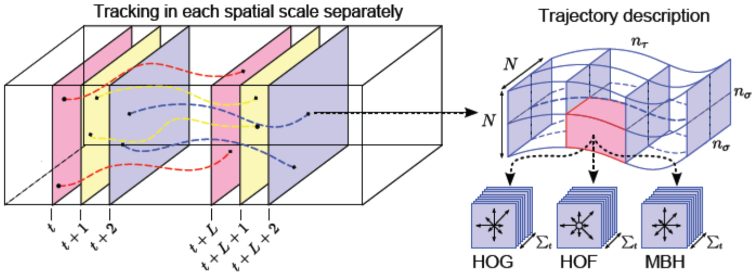
- Y. and L. Wolf, "Local Trinary Patterns for Human Action Recognition ", ICCV 2009



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



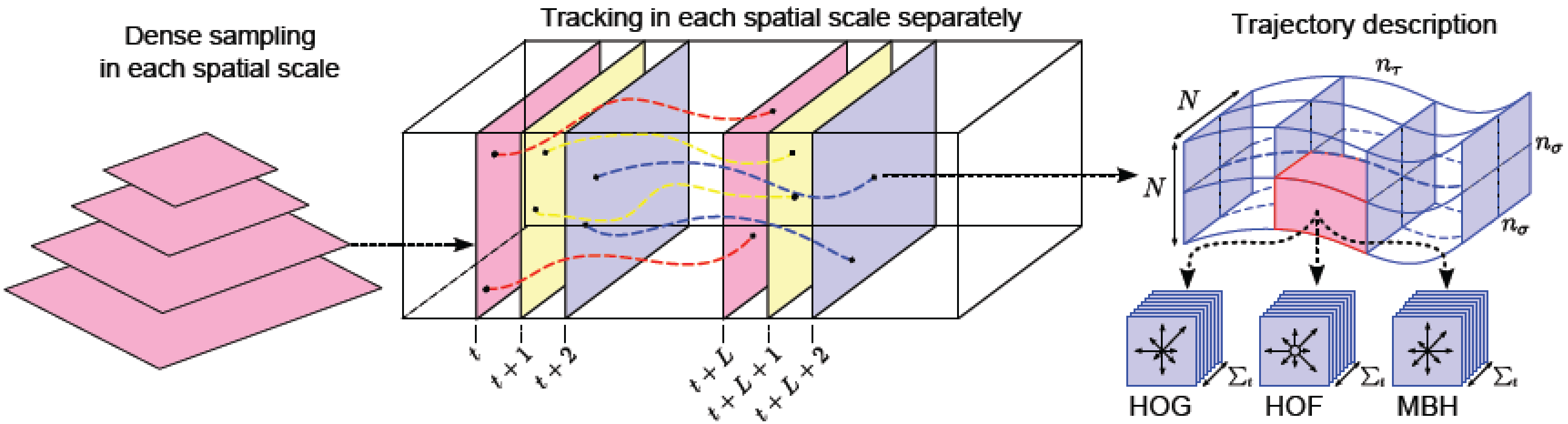
- H. Wang, A. Klaser, C. Schmid, C.-L. Liu, "Action Recognition by Dense Trajectories", CVPR 2011

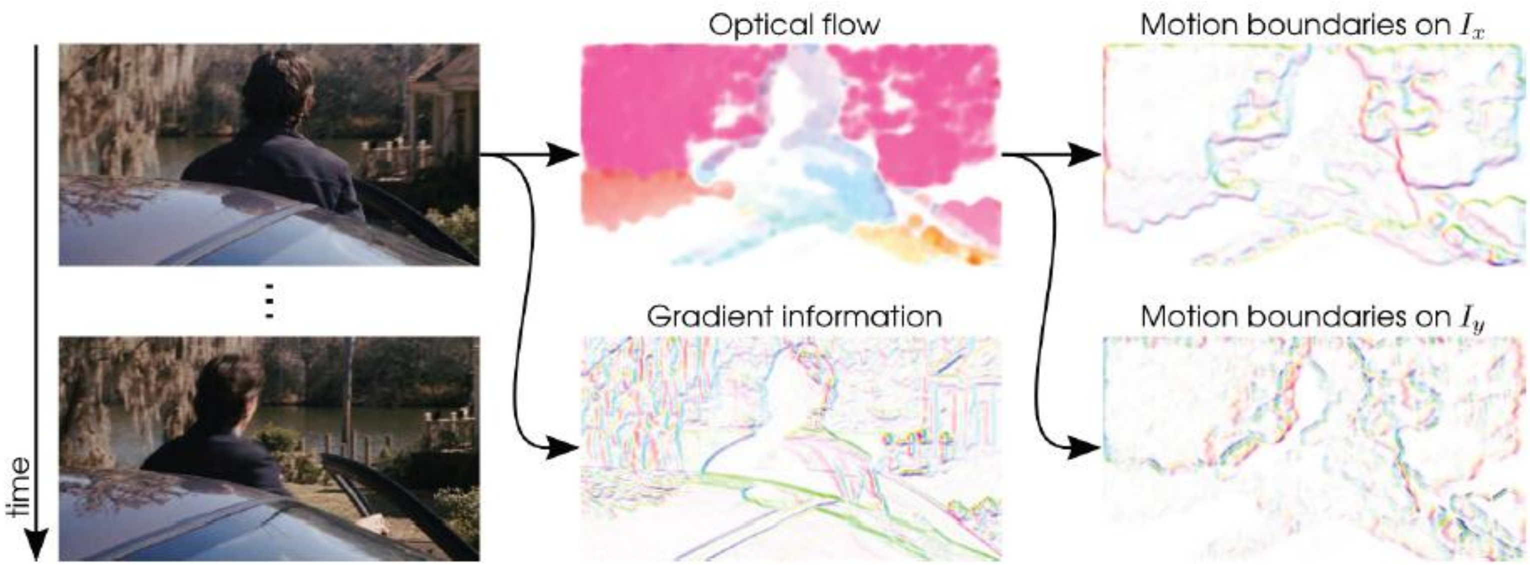


- Recognizing Human Actions by Attributes J. Liu, B. Kuipers, S. Savarese, CVPR 2011

Dense trajectory descriptors

[Wang et al. CVPR'11]





Dense trajectory descriptors

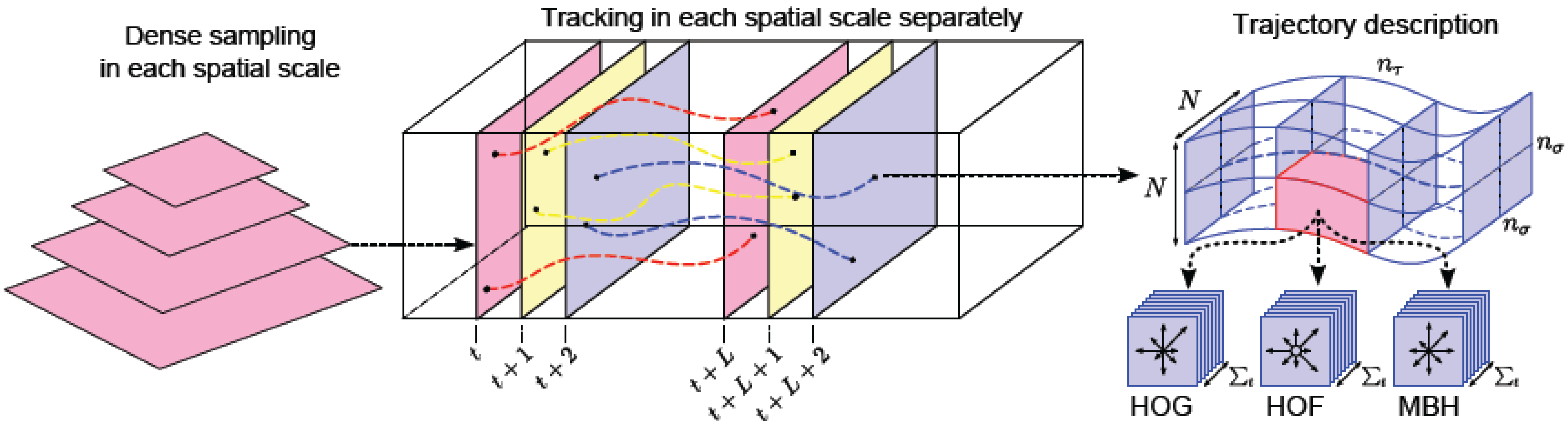
[Wang et al. CVPR'11]

	KTH		YouTube		Hollywood2		UCF sports	
	KLT	Dense trajectories	KLT	Dense trajectories	KLT	Dense trajectories	KLT	Dense trajectories
Trajectory	88.4%	90.2%	58.2%	67.2%	46.2%	47.7%	72.8%	75.2%
HOG	84.0%	86.5%	71.0%	74.5%	41.0%	41.5%	80.2%	**83.8%**
HOF	92.4%	93.2%	64.1%	72.8%	48.4%	50.8%	72.7%	77.6%
MBH	93.4%	**95.0%**	72.9%	**83.9%**	48.6%	**54.2%**	78.4%	**84.8%**
Combined	**93.4%**	**94.2%**	**79.9%**	**84.2%**	**54.6%**	**58.3%**	82.1%	**88.2%**

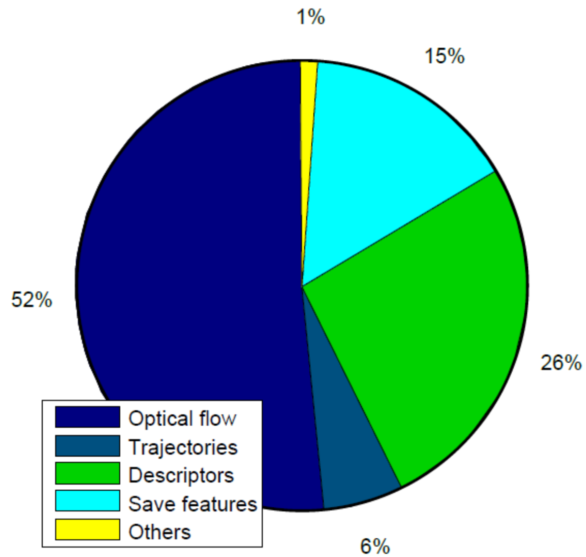
KTH		YouTube		Hollywood2		UCF sports	
Laptev *et al.* [14]	91.8%	Liu *et al.* [16]	71.2%	Wang *et al.* [32]	47.7%	Wang *et al.* [32]	85.6%
Yuan *et al.* [35]	93.3%	Ikizler-Cinbis *et al.* [9]	75.21%	Gilbert *et al.* [8]	50.9%	Kovashka *et al.* [12]	87.27%
Gilbert *et al.* [8]	94.5%			Ullah *et al.* [31]	53.2%	Kläser *et al.* [10]	86.7%
Kovashka *et al.* [12]	**94.53%**			Taylor *et al.* [29]	46.6%		
[Wang et al.]	94.2%	[Wang et al.]	**84.2%**	[Wang et al.]	**58.3%**	[Wang et al.]	**88.2%**

Dense trajectory descriptors

[Wang et al. CVPR'11]

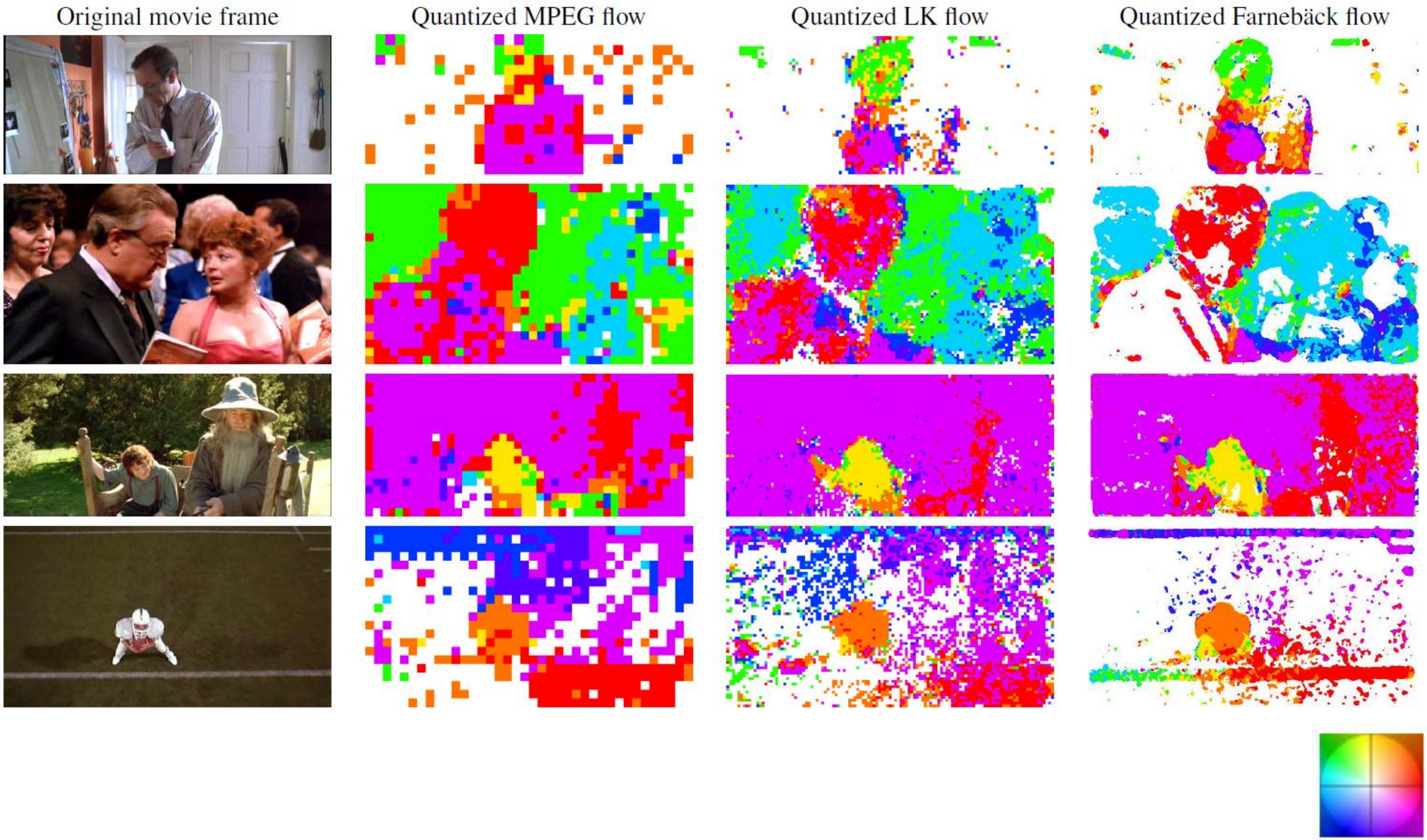


Computational cost:



Highly-efficient video descriptors

Optical flow from MPEG video compression



Highly-efficient video descriptors

Evaluation on Hollywood2

	Acc.	Feat. (fps)	Quant. (fps)	Total (fps)
CD FLANN(4-32)	55.8%		52.4	40.0
CD VLAD(4)	56.7%	168.4	167.5	84.0
CD FV(32)	58.2%		40.9	32.9
DT [Wang et al.'11]	59.9%	1.2	5.1	1.0

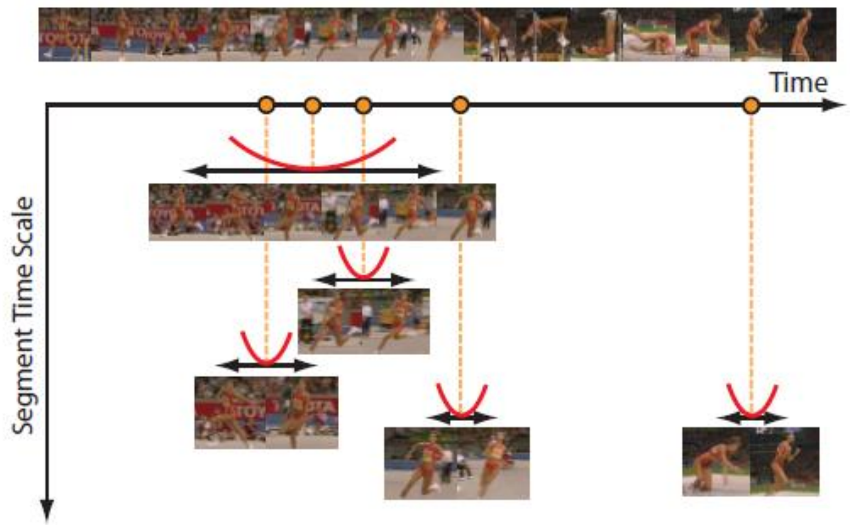
Evaluation on UCF50

	Acc.	Feat. (fps)	Quant. (fps)	Total (fps)
CD FLANN(4-32)	81.6%		52.4	48.1
CD VLAD(4)	80.6%	591.8	671.4	314.6
CD FV(32)	82.2%		171.3	132.8
DT [Wang et al.'11]	85.6%	2.8	5.1	1.8

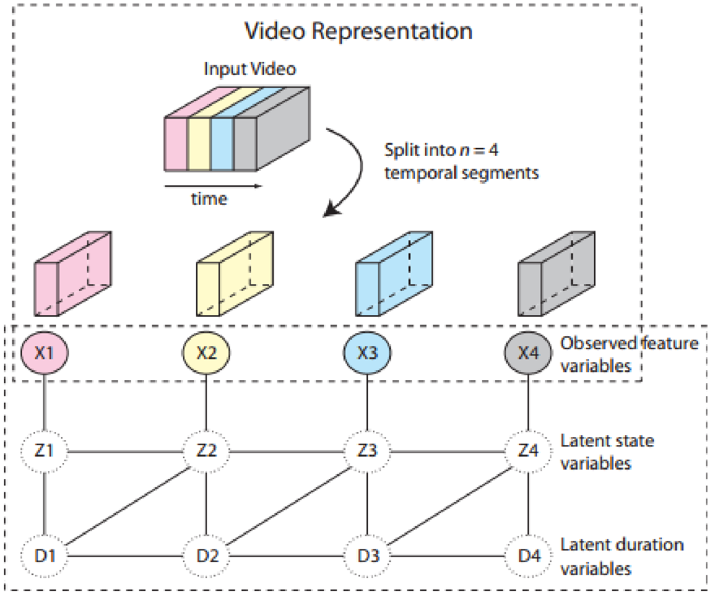
[Kantorov & Laptev, 2013]

Beyond BOF: Temporal structure

- Modeling Temporal Structure of Decomposable Motion Segments for Activity Classication, J.C. Niebles, C.-W. Chen and L. Fei-Fei, ECCV 2010

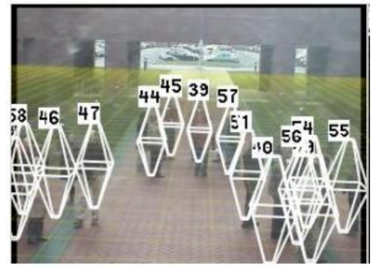


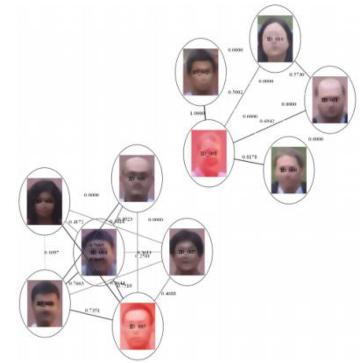
- Learning Latent Temporal Structure for Complex Event Detection. Kevin Tang, Li Fei-Fei and Daphne Koller, CVPR 2012



Beyond BOF: Social roles

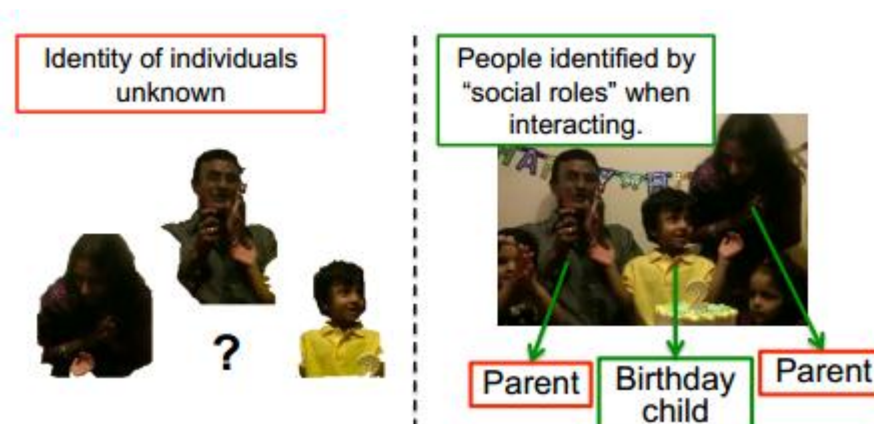
- T. Yu, S.-N. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In CVPR, 2009.





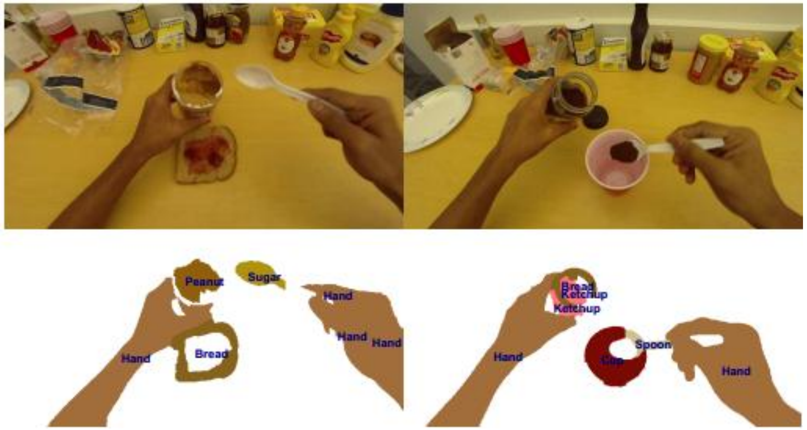
- L. Ding and A. Yilmaz. Learning relations among movie characters: A social network perspective. In ECCV, 2010

- V. Ramanathan, B. Yao, and L. Fei-Fei. Social Role Discovery in Human Events. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013.

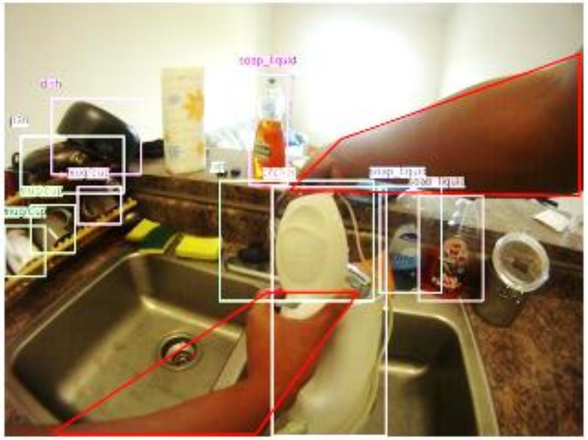


Beyond BOF: Egocentric activities

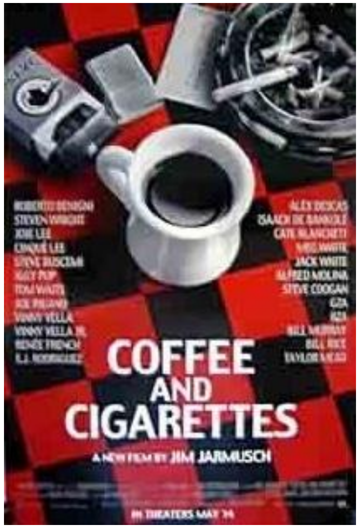
- A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In ICCV, 2011.



- H. Pirsiavash, D. Ramanan. Recognizing Activities of Daily Living in First-Person Camera Views, In CVPR, 2012.



Beyond BOF: Action localization



Manual annotation of drinking actions in movies:
"Coffee and Cigarettes"; "Sea of Love"

"*Drinking*": 159 annotated samples

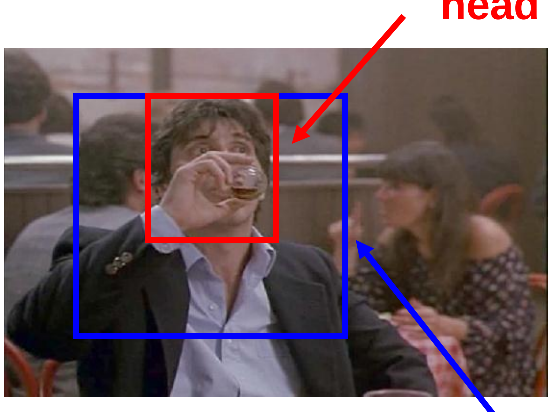
"*Smoking*": 149 annotated samples

Temporal annotation

First frame **Keyframe** **Last frame**

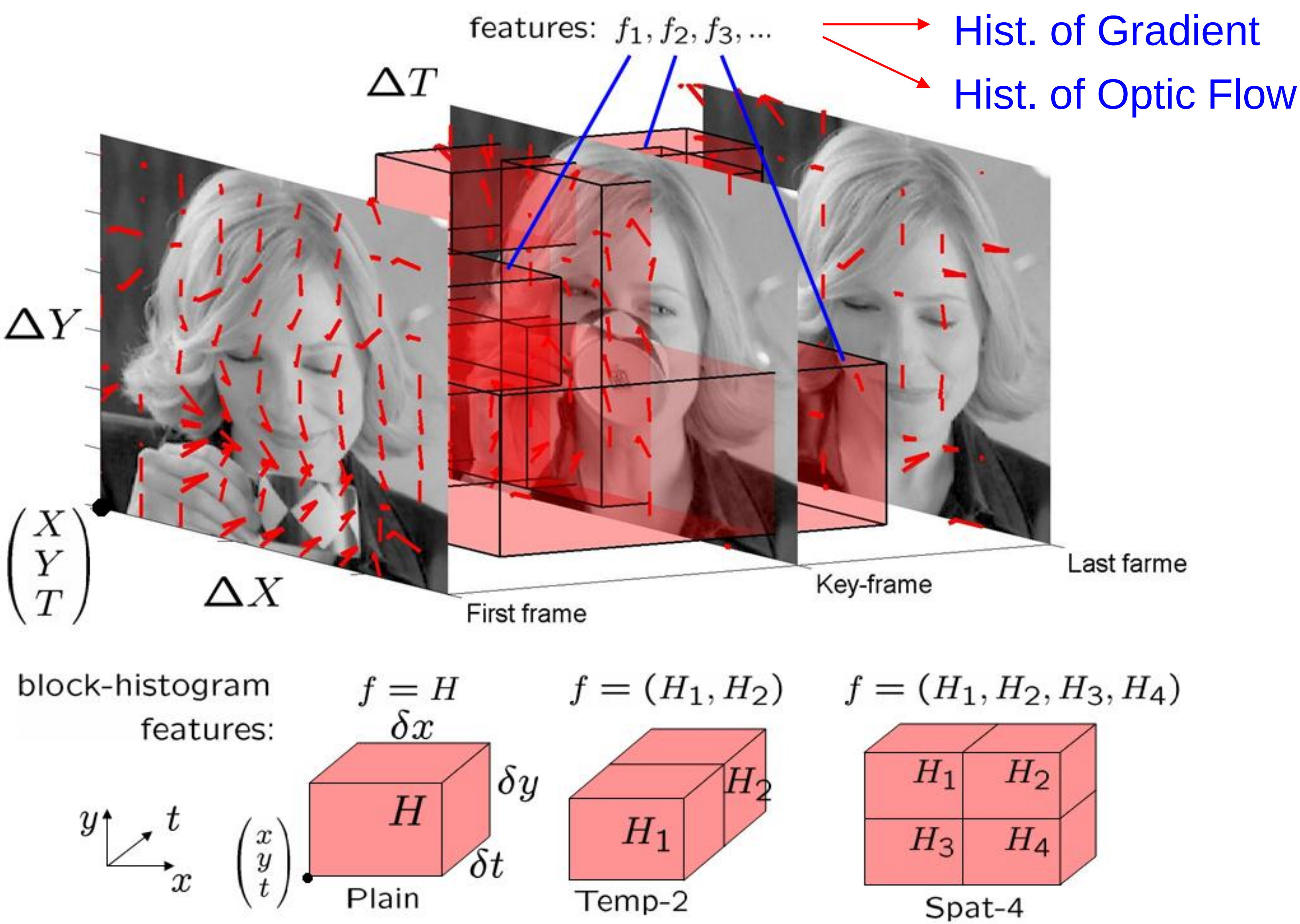
Spatial annotation

head rectangle

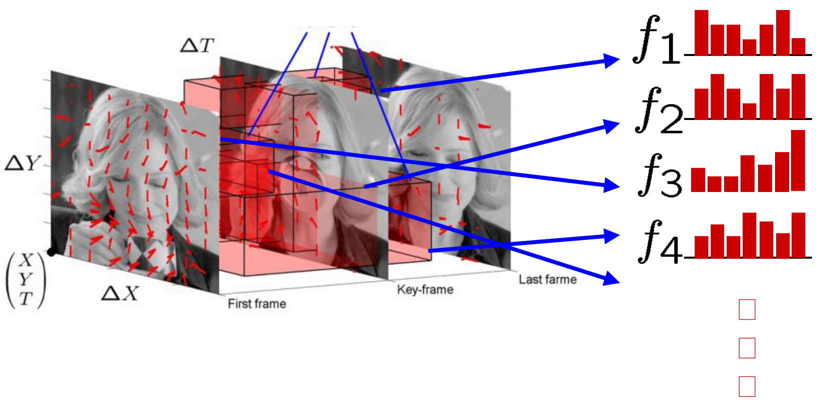


torso rectangle

Action representation



Action learning



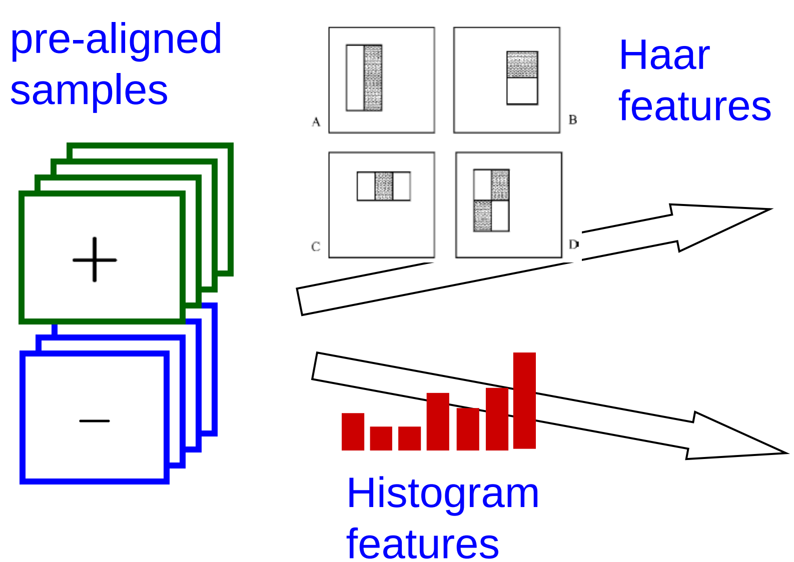
boosting

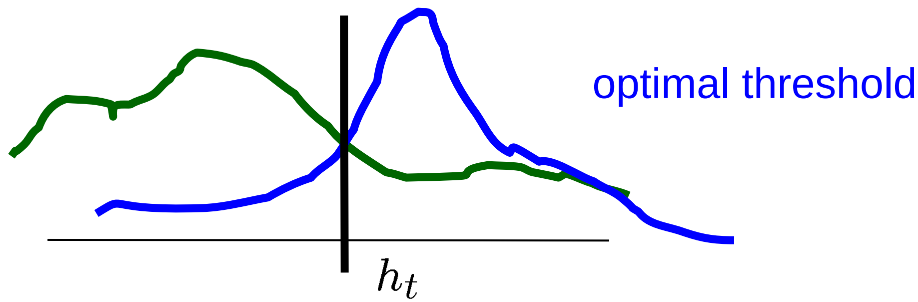
selected features

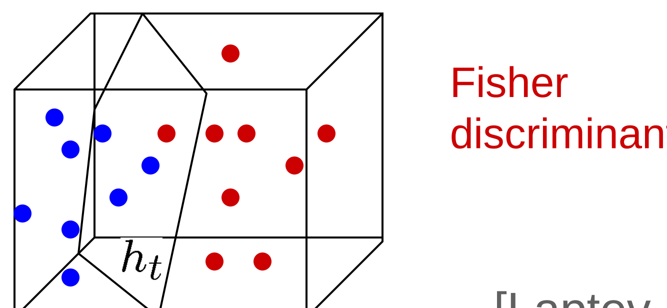
$$H(z) = \mathrm{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(f_t)\right)$$

weak classifier

AdaBoost:
- Efficient discriminative classifier [Freund&Schapire'97]
- Good performance for face detection [Viola&Jones'01]







[Laptev, Perez 2007]

Action Detection

Test episodes from the movie “Coffee and cigarettes”

[Laptev, Perez 2007]

20 most confident detections

Where to get training data?

➡ **Weakly-supervised learning**

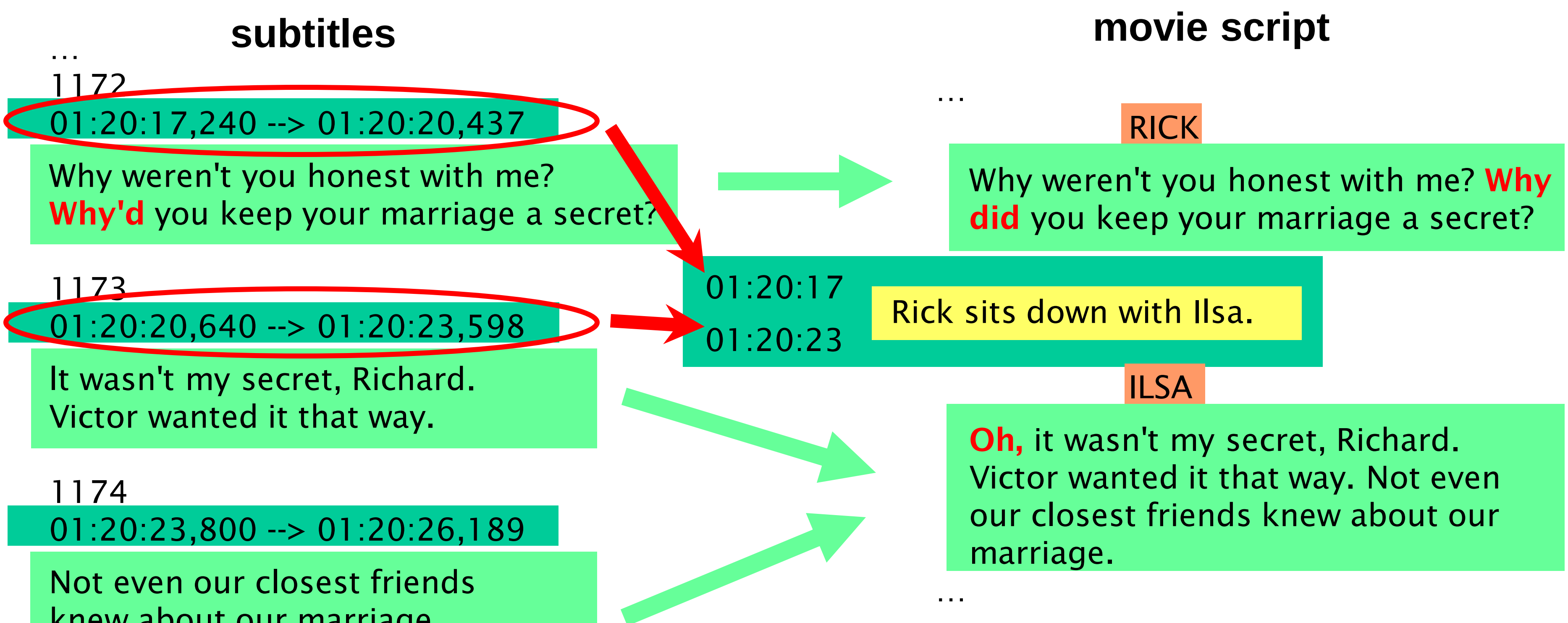
Actions in movies

- Realistic variation of human actions
- Many classes and many examples per class

- Typically only a few class-samples per movie
- Manual annotation is <u>very</u> time consuming

Script-based video annotation

- Scripts available for >500 movies (no time synchronization)
  www.dailyscript.com, www.movie-page.com, www.weeklyscript.com …
- Subtitles (with time info.) are available for the most of movies
- Can transfer time to scripts by text alignment



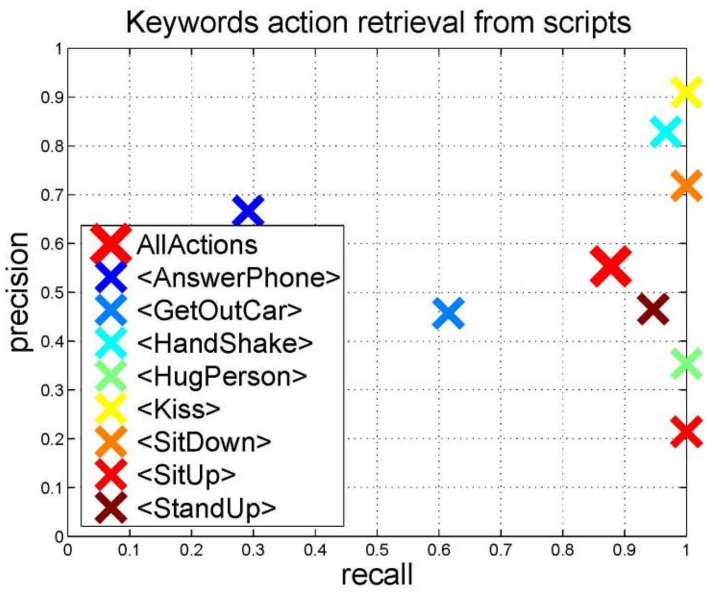
[Laptev, Marszałek, Schmid, Rozenfeld 2008]

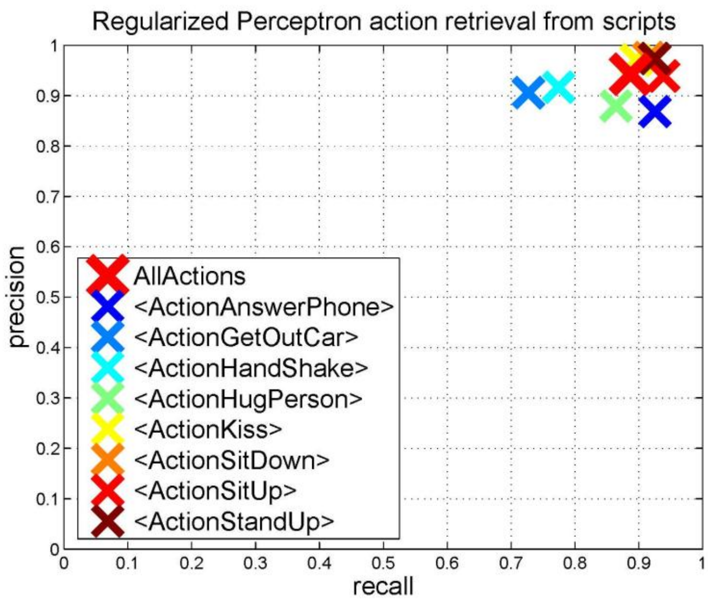
Text-based action retrieval

- Large variation of action expressions in text:

GetOutCar action:	*"… Will gets out of the Chevrolet. …"* *"… Erin exits her new truck…"*
Potential false positives:	*"…About to sit down, he freezes…"*

- => Supervised text classification approach





[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Hollywood-2 actions dataset

Actions			
	Training subset (clean)	Training subset (automatic)	Test subset (clean)
AnswerPhone	66	59	64
DriveCar	85	90	102
Eat	40	44	33
FightPerson	54	33	70
GetOutCar	51	40	57
HandShake	32	38	45
HugPerson	64	27	66
Kiss	114	125	103
Run	135	187	141
SitDown	104	87	108
SitUp	24	26	37
StandUp	132	133	146
All Samples	**823**	**810**	**884**

Training and test samples are obtained from 33 and 36 distinct movies respectively.

Hollywood-2 dataset is on-line:
http://www.irisa.fr/vista/actions/hollywood2

[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Action classification results

	Clean		Automatic		
	hoghof		hoghof		Chance
Channel	bof	flat	bof	flat	
mAP	47.9	50.3	31.9	36.0	9.2
AnswerPhone	15.7	20.9	18.2	19.1	7.2
DriveCar	86.6	84.6	78.2	80.1	11.5
Eat	59.5	67.0	13.0	22.3	3.7
FightPerson	71.1	69.8	52.9	57.6	7.9
GetOutCar	29.3	45.7	13.8	27.7	6.4
HandShake	21.2	27.8	12.8	18.9	5.1
HugPerson	35.8	43.2	15.2	20.4	7.5
Kiss	51.5	52.5	43.2	48.6	11.7
Run	69.1	67.8	54.2	49.1	16.0
SitDown	58.2	57.6	28.6	34.1	12.2
SitUp	17.5	17.2	11.8	10.8	4.2
StandUp	51.7	54.3	40.5	43.6	16.5

Average precision (AP) for Hollywood-2 dataset

Actions in the context of scenes

- Human actions are frequently correlated with particular scene classes

  Reasons: *physical properties* and *particular purposes* of scenes

Eating -- *kitchen*

Eating -- *cafe*

Running -- *road*

Running -- *street*

Mining scene captions

01:22:00
01:22:03

ILSA

I wish I didn't love you so much.

She snuggles closer to Rick.

CUT TO:

EXT. RICK'S CAFE - NIGHT

Laszlo and Carl make their way through the darkness toward a side entrance of Rick's. They run inside the entryway.

The headlights of a speeding police car sweep toward them.

They flatten themselves against a wall to avoid detection.

The lights move past them.

01:22:15
01:22:17

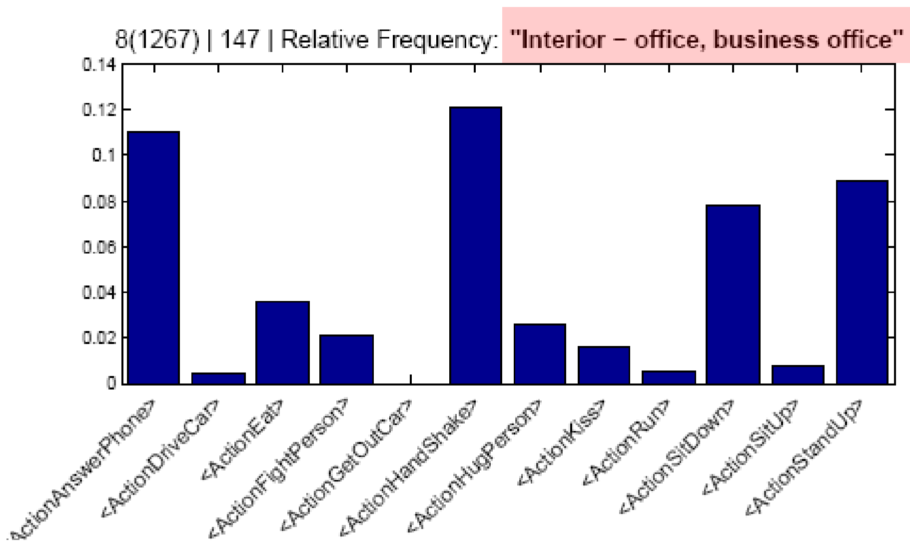
CARL

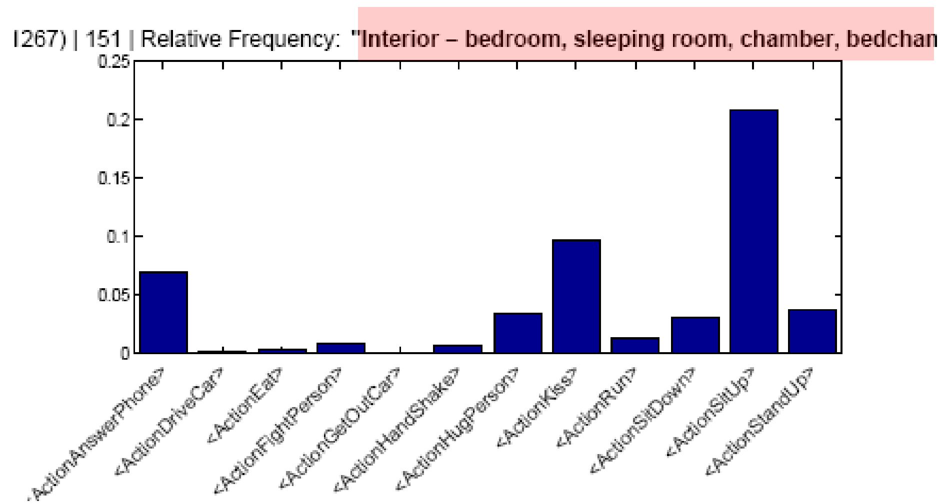
I think we lost them.

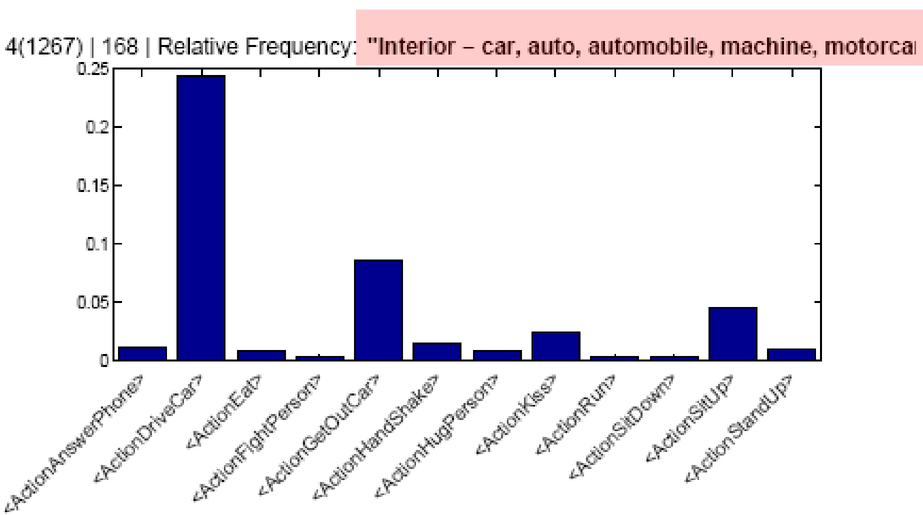
...

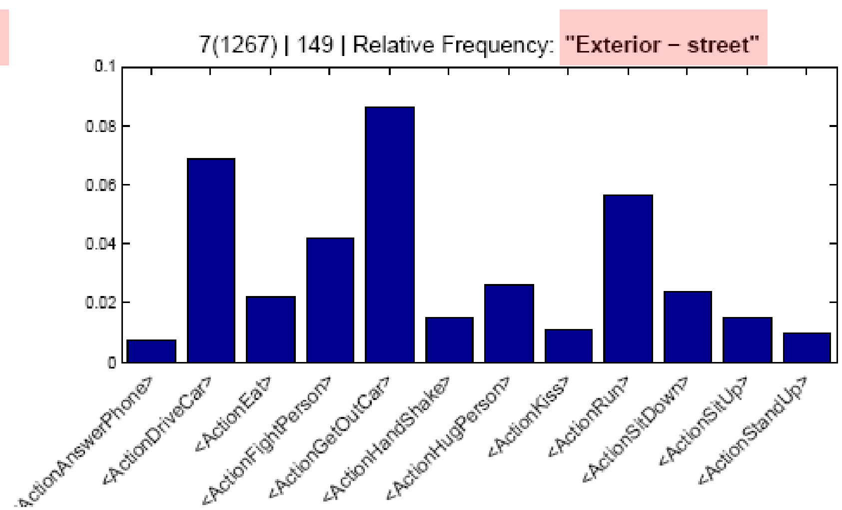
[Marszałek, Laptev, Schmid 2008]

Co-occurrence of actions and scenes in scripts





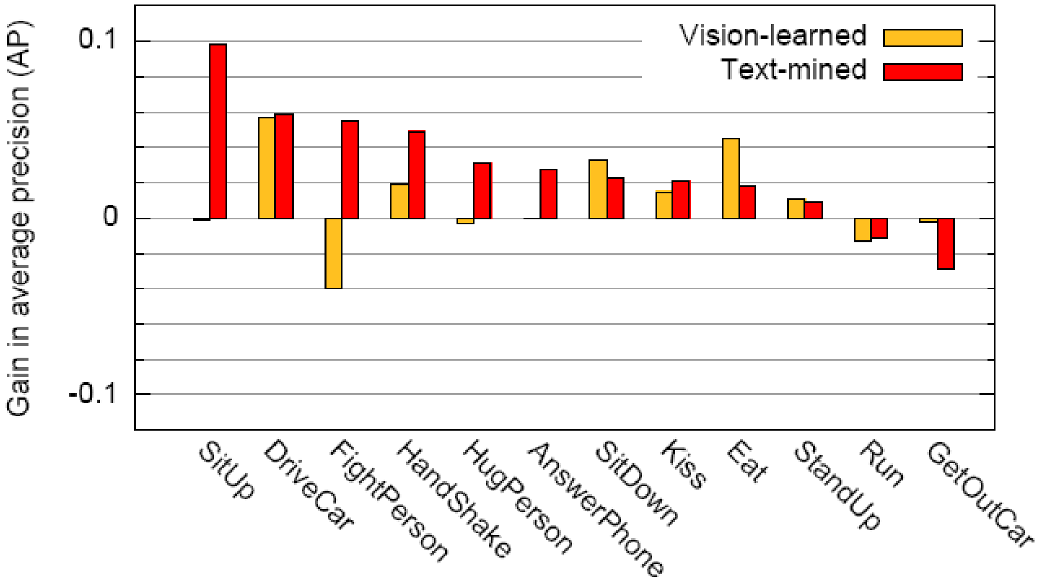




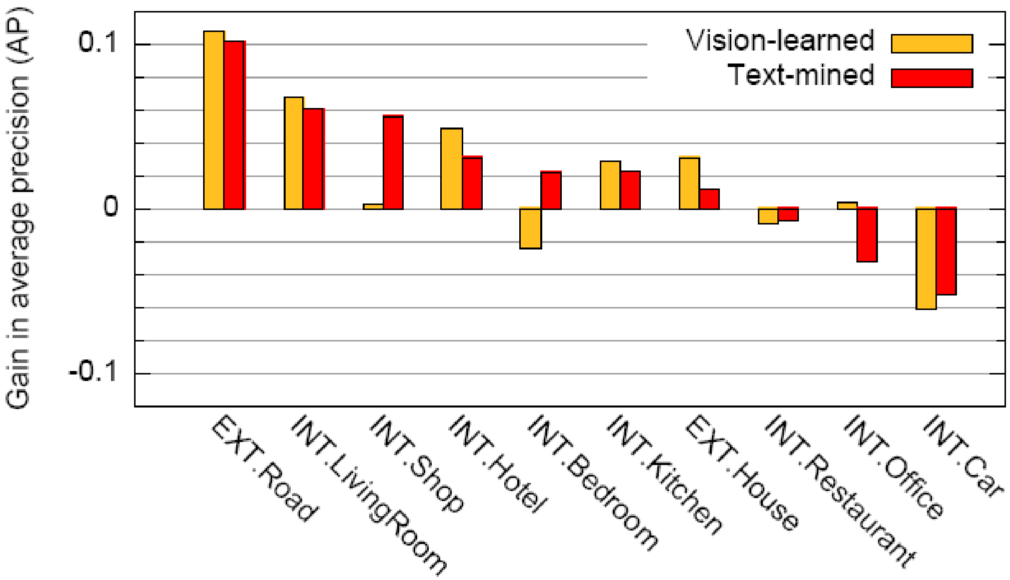
[Marszałek, Laptev, Schmid 2008]

Results: actions and scenes (jointly)

Actions
in the
context
of
Scenes



Scenes
in the
context
of
Actions



[Marszałek, Laptev, Schmid 2008]

Handling temporal uncertainty

Subtitles

00:24:22 –> 00:24:25
– Yes, Monsieur Laszlo. Right this way.

00:24:51 –> 00:24:53
Two Cointreaux, please.

Script

Speech

Monsieur Laszlo. Right this way.

Scene description

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Speech

Two cointreaux, please.

[Duchenne, Laptev, Sivic, Bach, Ponce, 2009]

Discriminative action clustering

Input:

- Action type, e.g.
  "Person opens door"
- Videos + aligned scripts

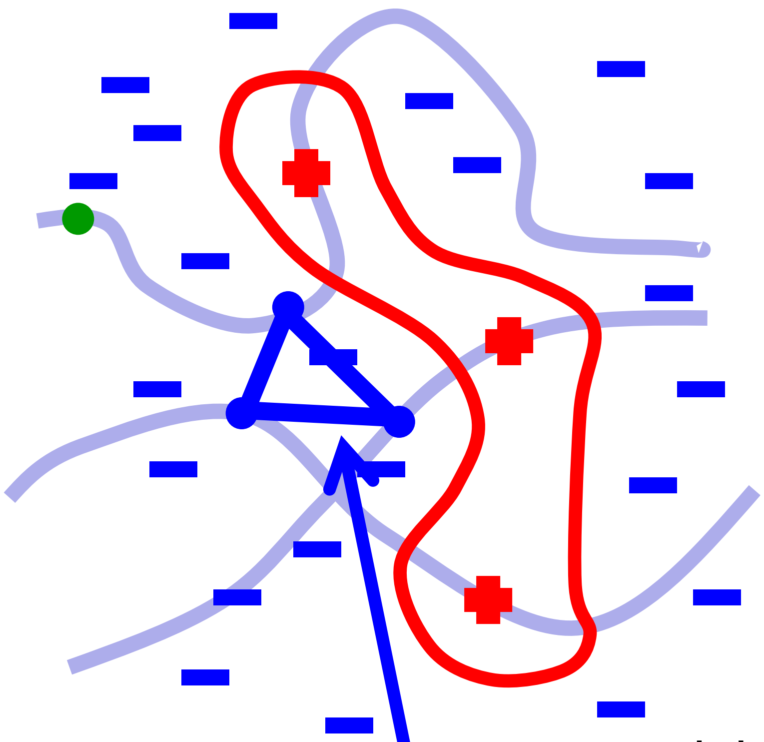
Automatic collection of video clips

... **Jane** jumps up and **opens** the **door** ...
... **Carolyn opens** the front **door** ...
... **Jane opens** her bedroom **door** ...

[Duchenne, Laptev, Sivic, Bach, Ponce, 2009]

Discriminative action clustering

Feature space



Nearest neighbor solution: wrong!

Video space

Negative samples

Random video samples: lots of them, very low chance to be positives

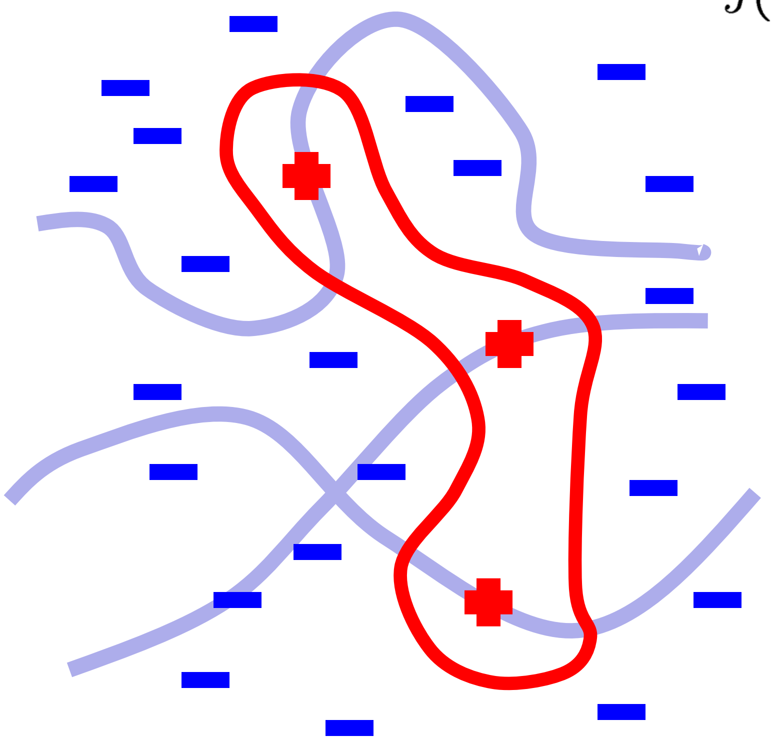
[Duchenne, Laptev, Sivic, Bach, Ponce, 2009]

Action clustering

Formulation

[Xu et al. NIPS'04]
[Bach & Harchaoui NIPS'07]

Feature space



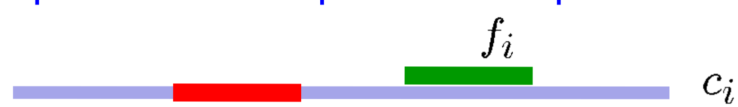
discriminative cost

$$J(f,w,b) = C_+ \underbrace{\sum_{i=1}^{M} \max\{0, 1 - w^\top \Phi(c_i[f_i]) - b\}}_{\text{Loss on positive samples}} +$$

$$+ C_- \underbrace{\sum_{i=1}^{P} \max\{0, 1 + w^\top \Phi(x_i^-) + b\}}_{\text{Loss on negative samples}} + \|w\|^2$$

x_i^- negative samples

$c_i[f_i]$ parameterized positive samples

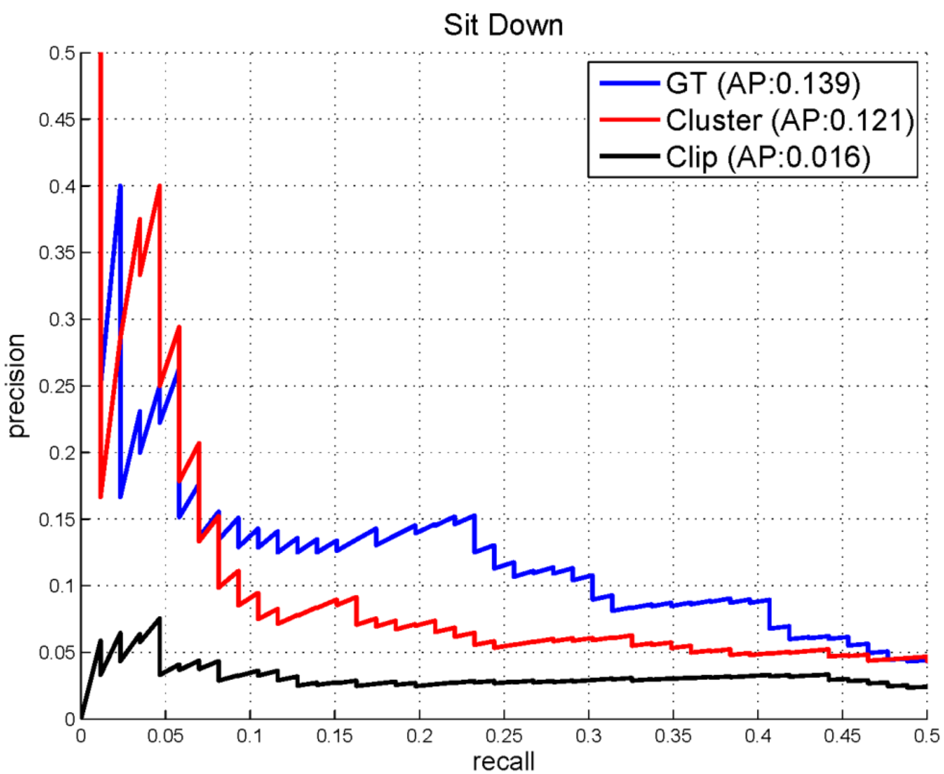


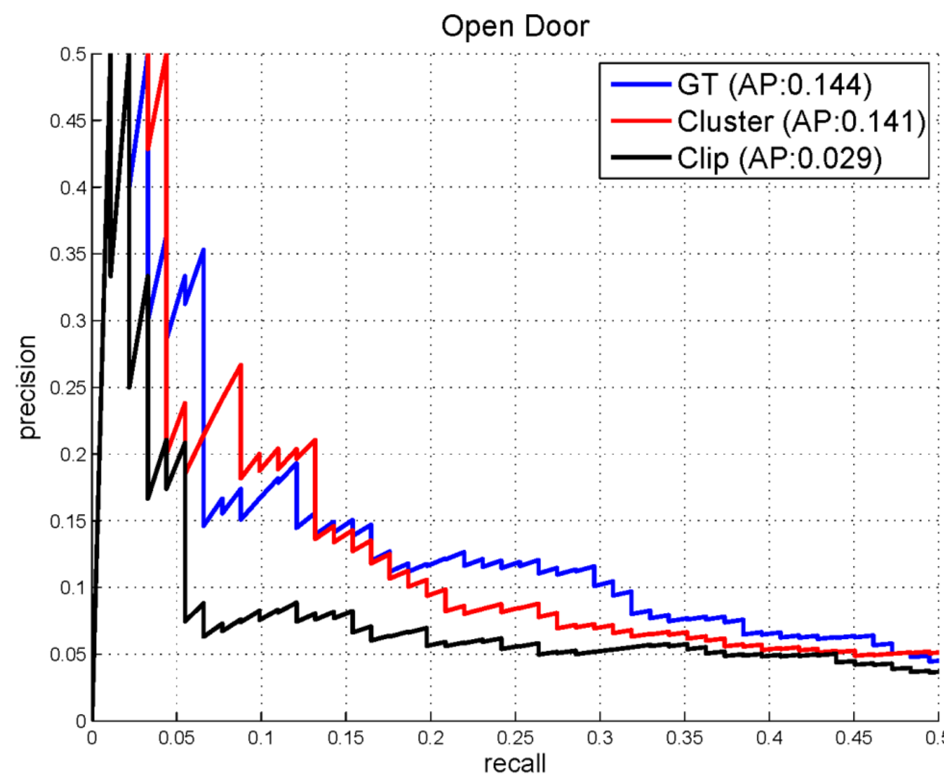
Optimization

SVM solution for w, b
Coordinate descent on f_i

Action detection: Sliding time window

"Sit Down" and "Open Door" actions in ~5 hours of movies

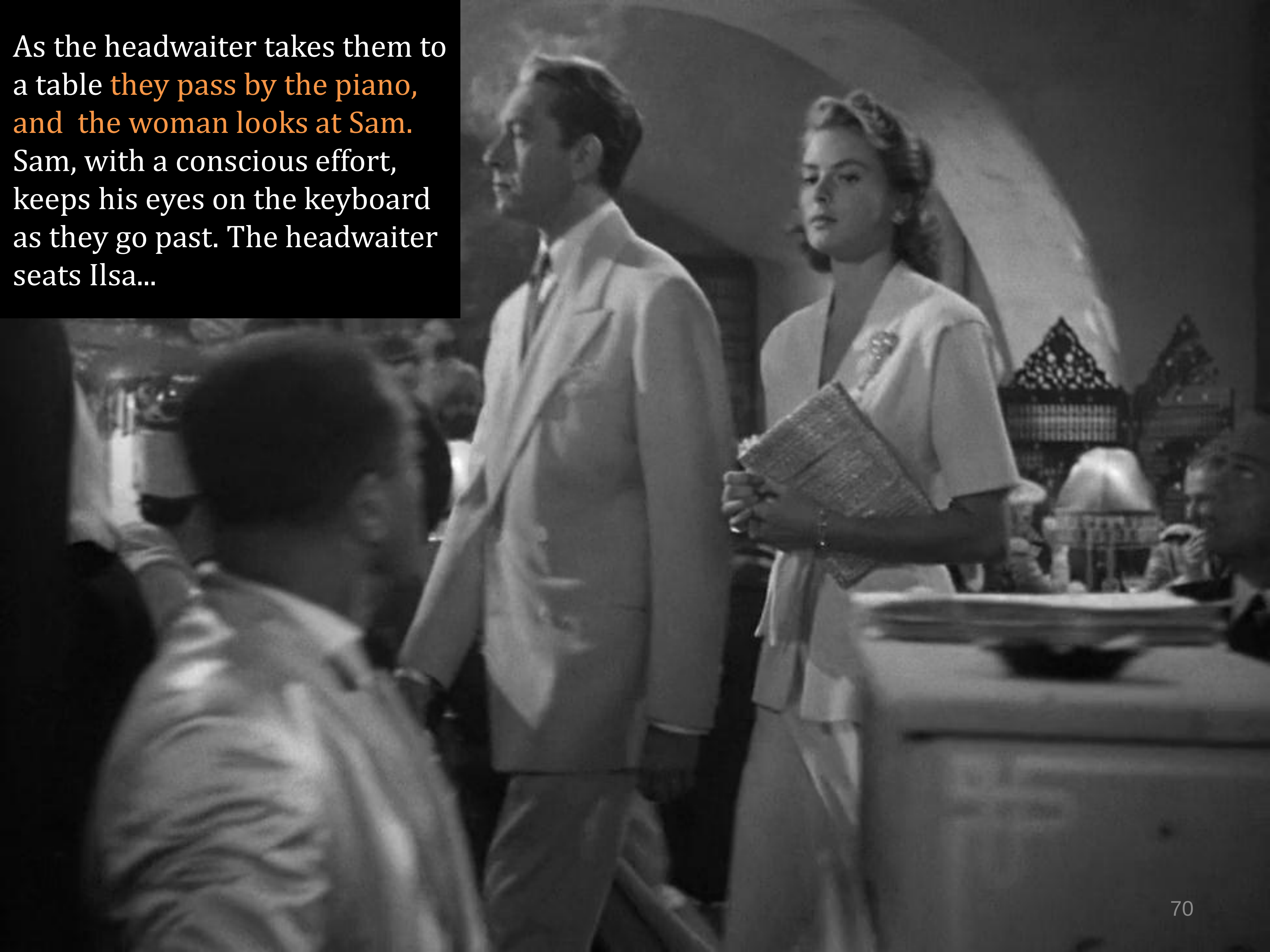




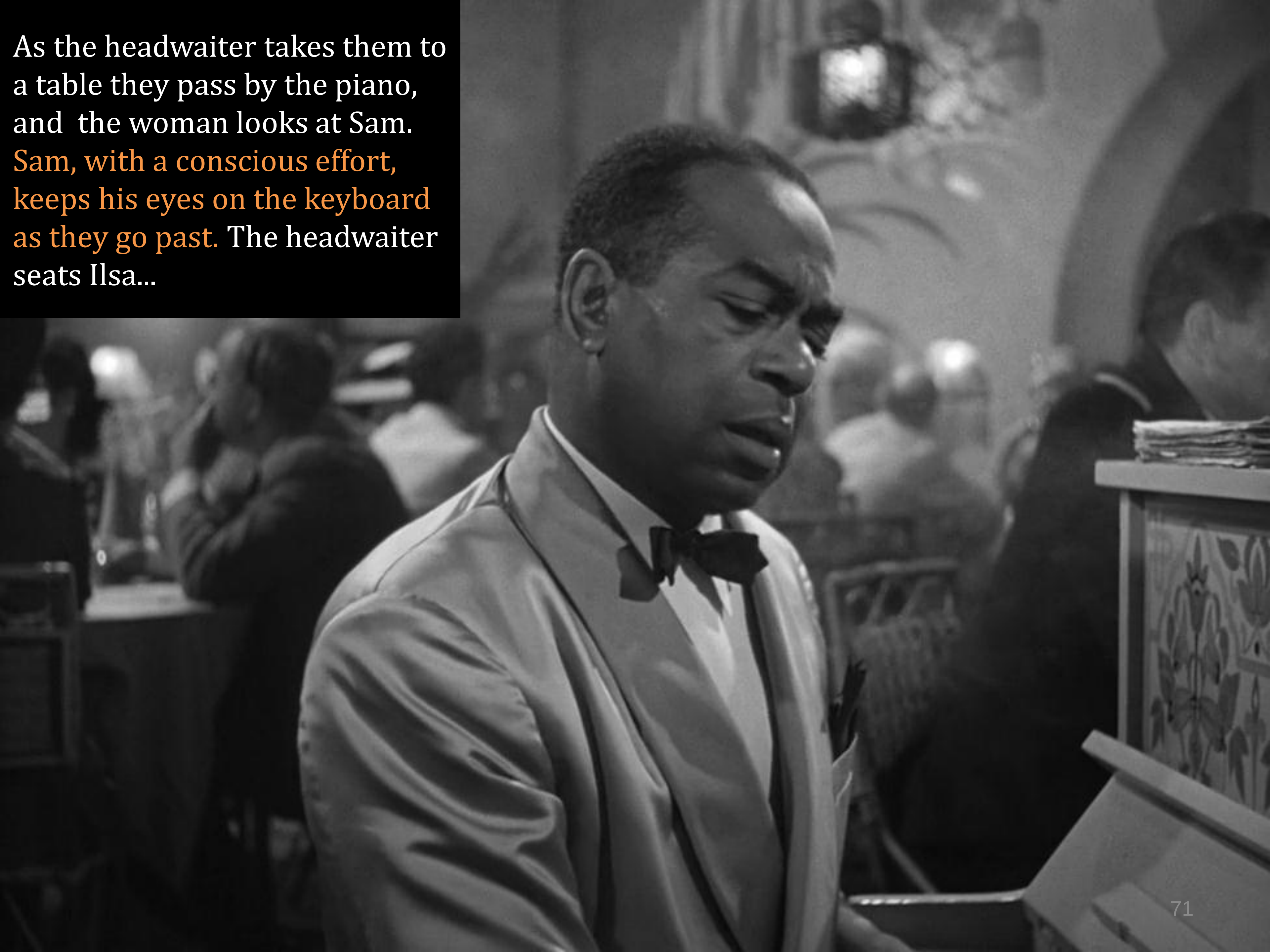
Temporal detection of “Sit Down” and “Open Door” actions in movies:
The Graduate, The Crying Game, Living in Oblivion [Duchenne et al. 09]

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

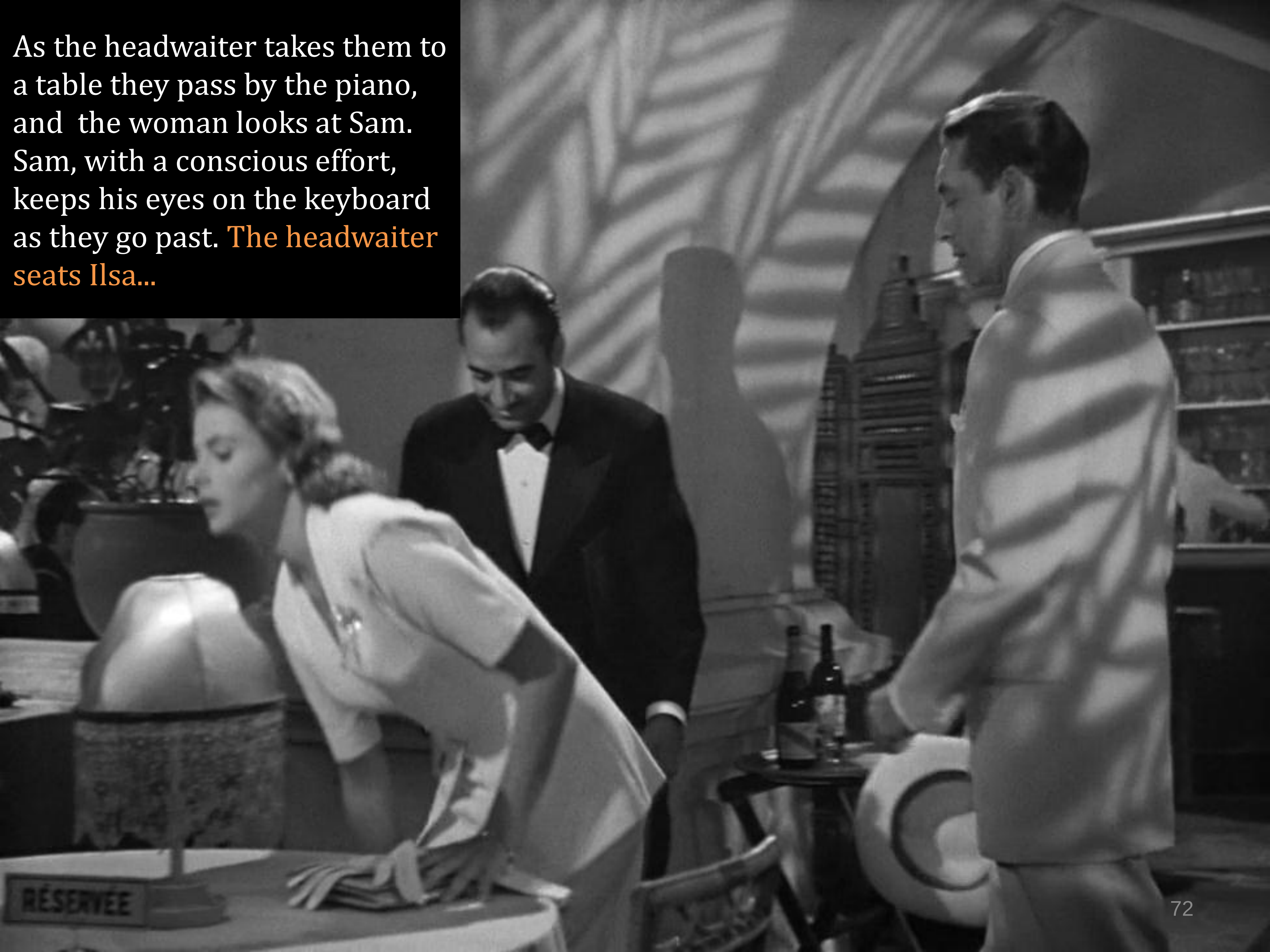
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



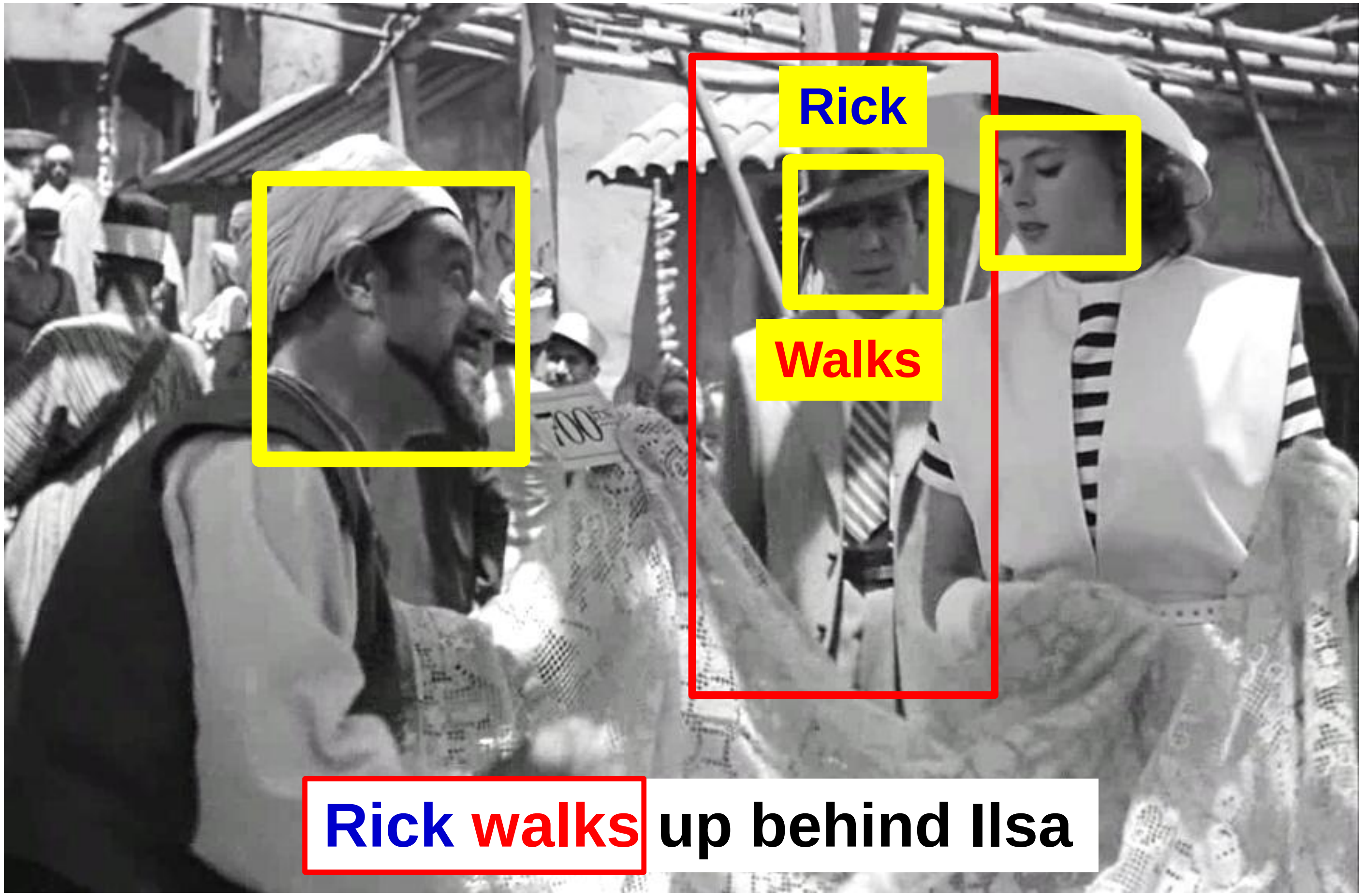
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



On-going: Joint Recognition of Actions and Actors

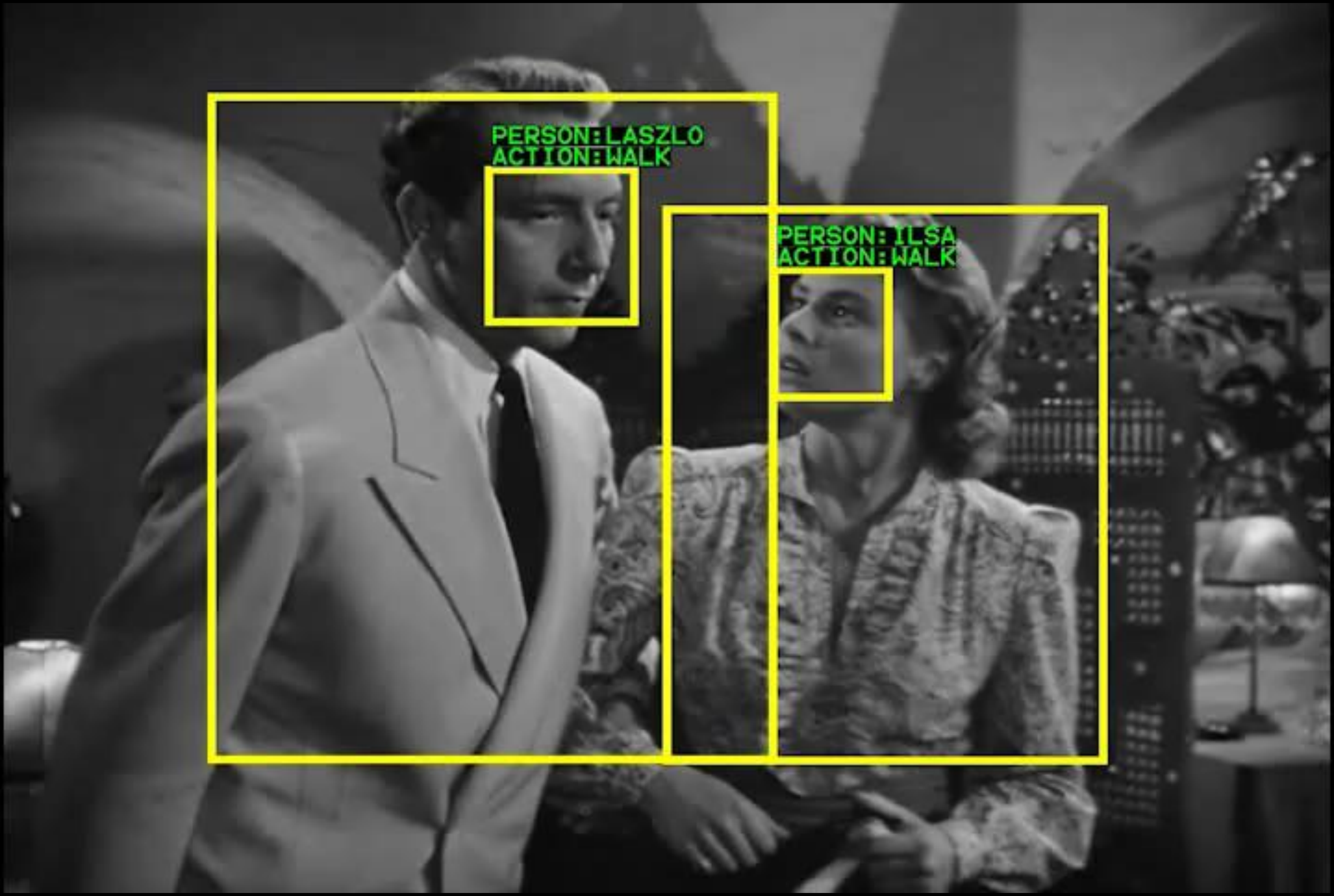
[Bojanowski, Bach, Laptev, Ponce, Sivic, Schmid, 2013, in submission]

On-going: Joint Recognition of Actions and Actors



[Bojanowski, Bach, Laptev, Ponce, Sivic, Schmid, 2013, in submission]

Recognition of Actions and Actors



[Bojanowski, Bach, Laptev, Ponce, Sivic, Schmid, 2013]

What we have seen so far

Actions understanding in realistic settings:

Action classification (and localization)

Is classification the final answer?

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?

Crowdsourcing action definitions

(Joint work with T.H. Vu, C. Olsson, A. Oliva and J. Sivic)

MTurk interface :

The movie to the left depicts one or more people doing something. Please watch the movie as many times as you would like and answer the questions below about the people in the movie. If at first you cannot see the movie, please try using a different browser (such as Chrome). **If you cannot successfully watch the movie, please do NOT accept this HIT.**

Please describe what each person in the video is doing. For example, "playing piano" or "walking to the table."

What is P1 doing?

What is P2 doing?

What is P3 doing?

Have you ever seen this clip before? ○ Yes, I think so ○ No, I don't think so

Submit

Crowdsourcing action definitions

Input video:

Five responses for each video and person:

situation 1:

Similar expressions

P1:

P1 is dancing with P2.
P1 dances with P2.
P1 is dancing with P2.
P1 is dancing with P2.
P1 is dancing with P2.

Crowdsourcing action definitions

Input video:

Action responses:

P1:

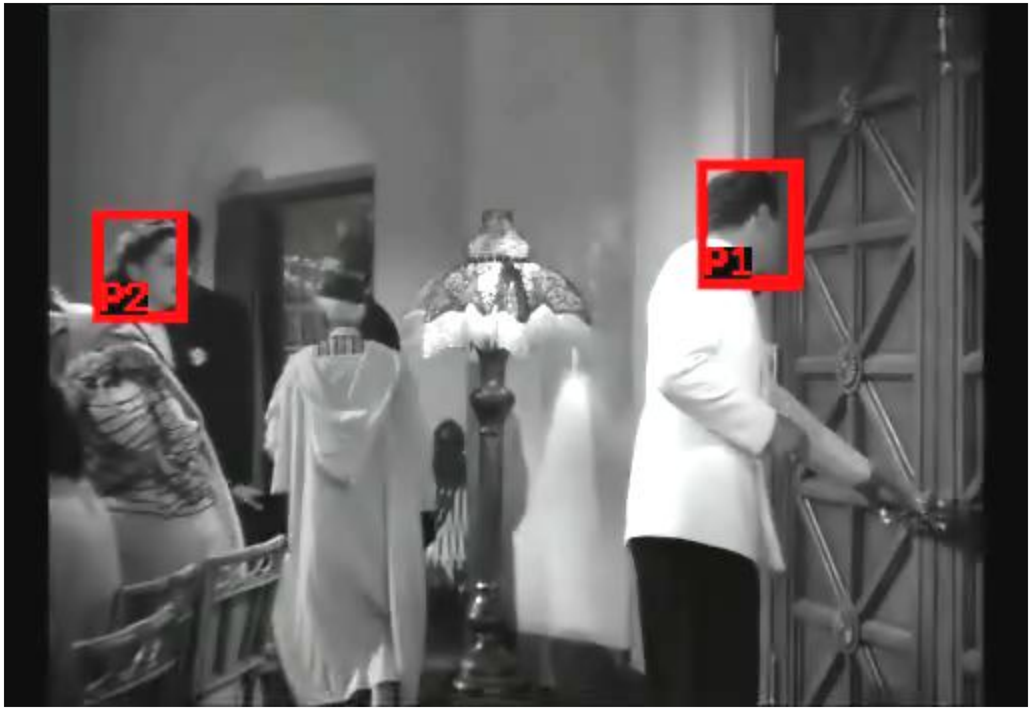
P1 greets P2 and shakes hands
P1 shakes P2's hand and greets him.
P1 is shaking P2's hand
P1 is shaking hands.
P1 shakes hands with P2.

situation 1:

Similar expressions

Crowdsourcing action definitions

Input video:



Action responses:

P2:

P2 is walking up to P1 and talking to him.
P2 approaches P1.
P2 runs towards P1 and speaks to him.
P2 is rushing to P1 before he leaves.
P2 stops P1 before he can leave to talk to him

situation 2:

Similar meaning
Different expressions

Crowdsourcing action definitions

Input video:

Action responses:

P1:
P1 is leaving the room
P1 gets up and leaves the table
P1 storms from the table.
P1 gets up and leaves to the back of the room.
P1 is walking away from an interaction with P2.

situation 2:

Similar meaning
Different expressions

Crowdsourcing action definitions

Input video:

Action responses:

P1:

P1 is carrying his money to the casino banker.
P1 is leading P3 and P4.
P1 walks in front of a group of people
P1 is leading P3 and P4 through the room.
P1 is walking up to the cage

situation 3:

Different expressions
Different meanings

Crowdsourcing action definitions

Input video:

Action responses:

P1:
P1 is walking through a crowd carrying cases
P1 is walking.
P1 is looking perplexed and walking away.
P1 scans the area.
P1 is looking for someone.

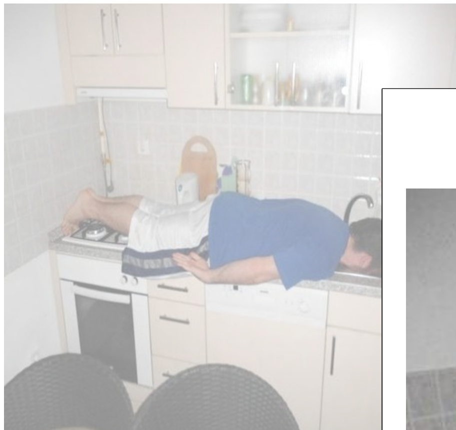
situation 3:

Different expressions
Different meanings

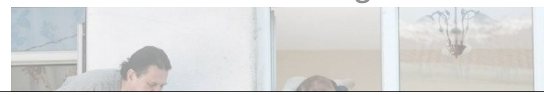
What current methods cannot do?

Limitations of Current Methods

What is unusual in this scene?



Is this scene dangerous?



What is intention of this person?

What is unusual in this scene?

Next challenge

Shift the focus of computer vision

Object, scene and action recognition

Recognition of objects' function and people's intentions

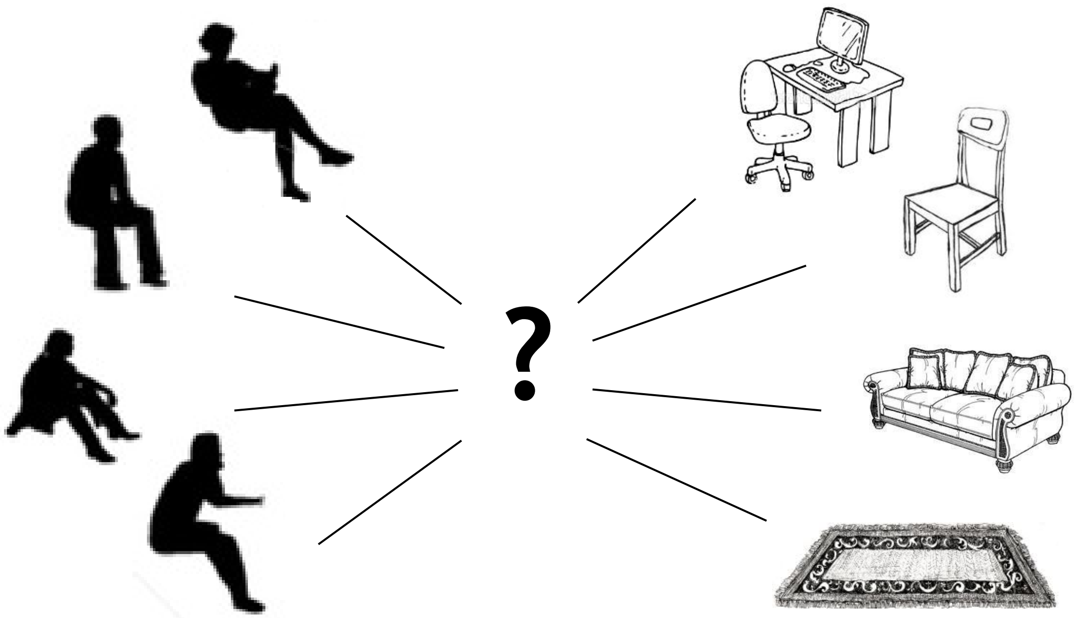
Is this a picture of a dog?
Is the person running in this video?

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.

[Delaitre, Fouhey, Laptev, Sivic, Gupta, Efros, 2012]

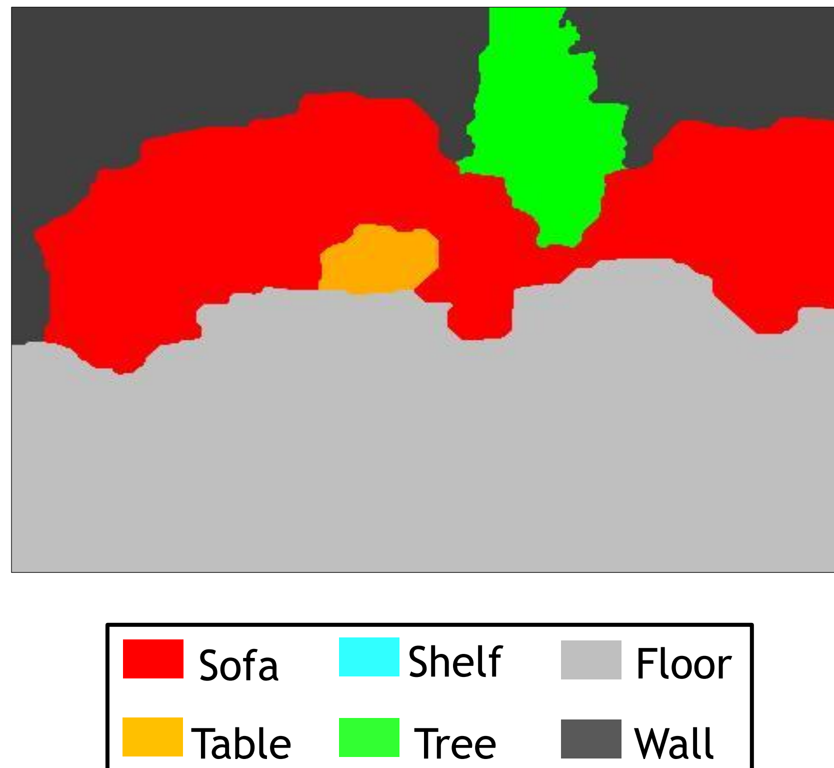
Goal

Recognize objects by the way people interact with them.

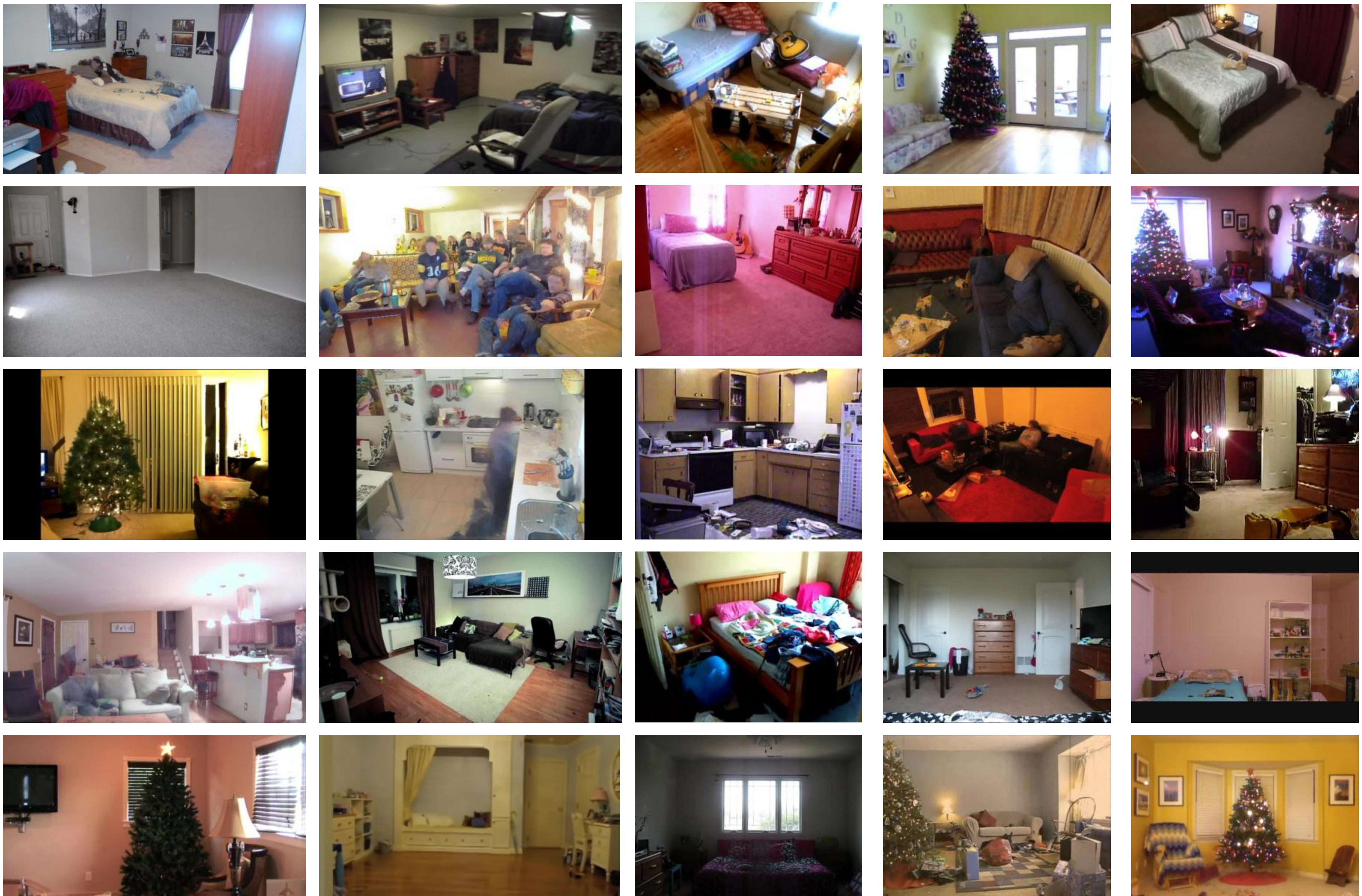
Time-lapse “Party & Cleaning” videos

Lots of person-object interactions,
many scenes on YouTube

Semantic object segmentation



New “Party & Cleaning” dataset



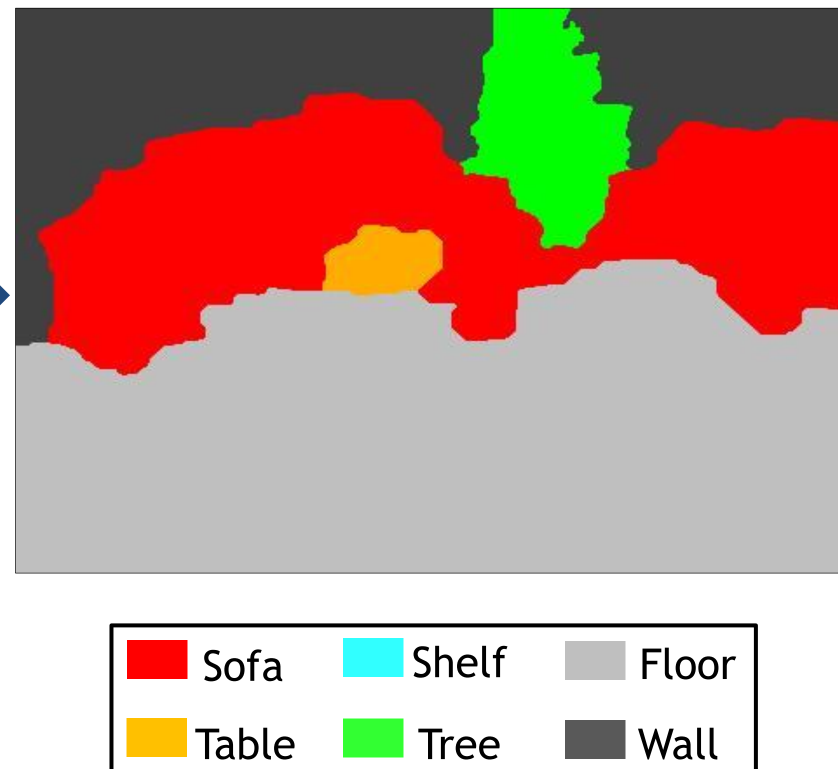
Goal

Recognize objects by the way people interact with them.

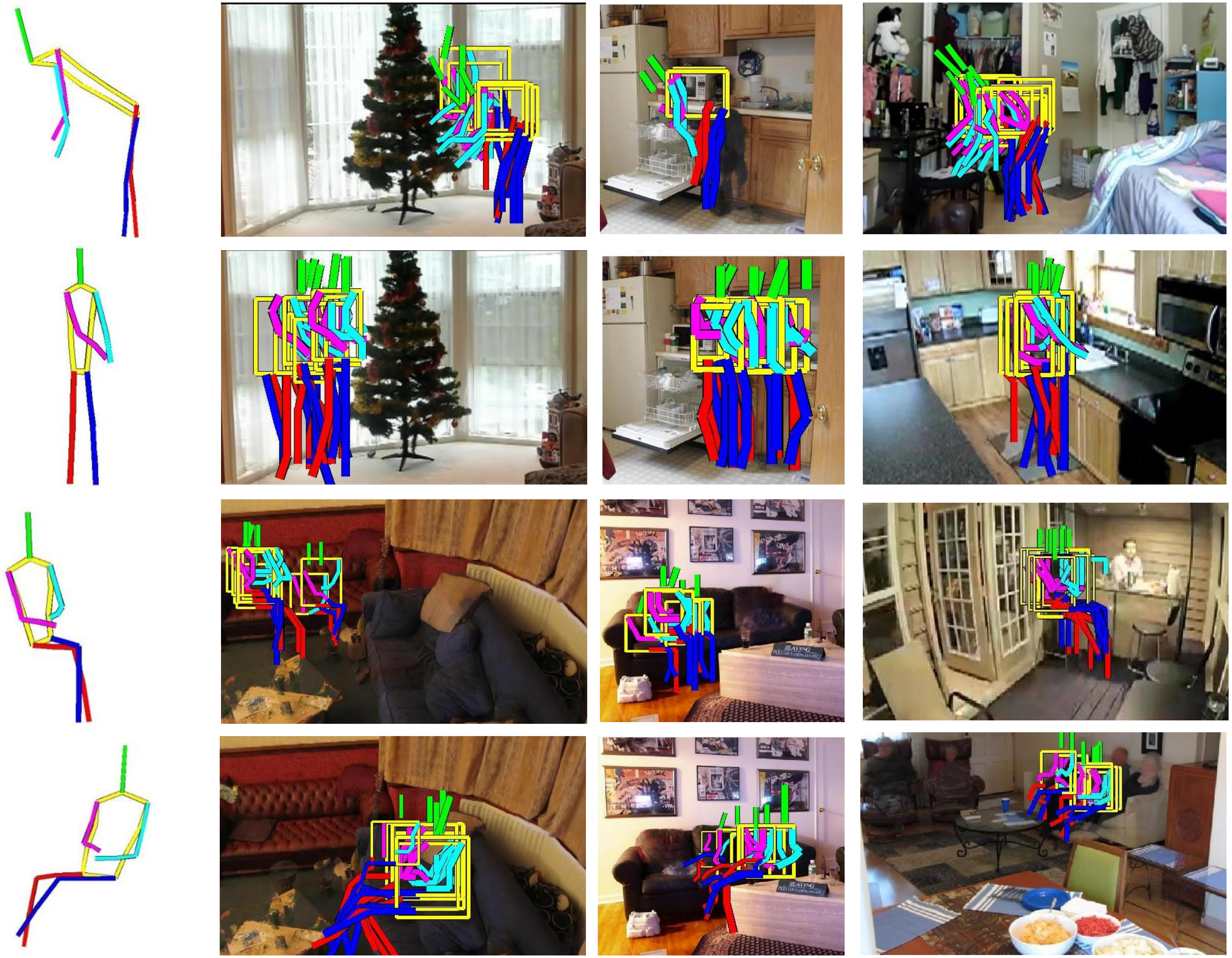
Time-lapse “Party & Cleaning” videos

Lots of person-object interactions, many scenes on YouTube

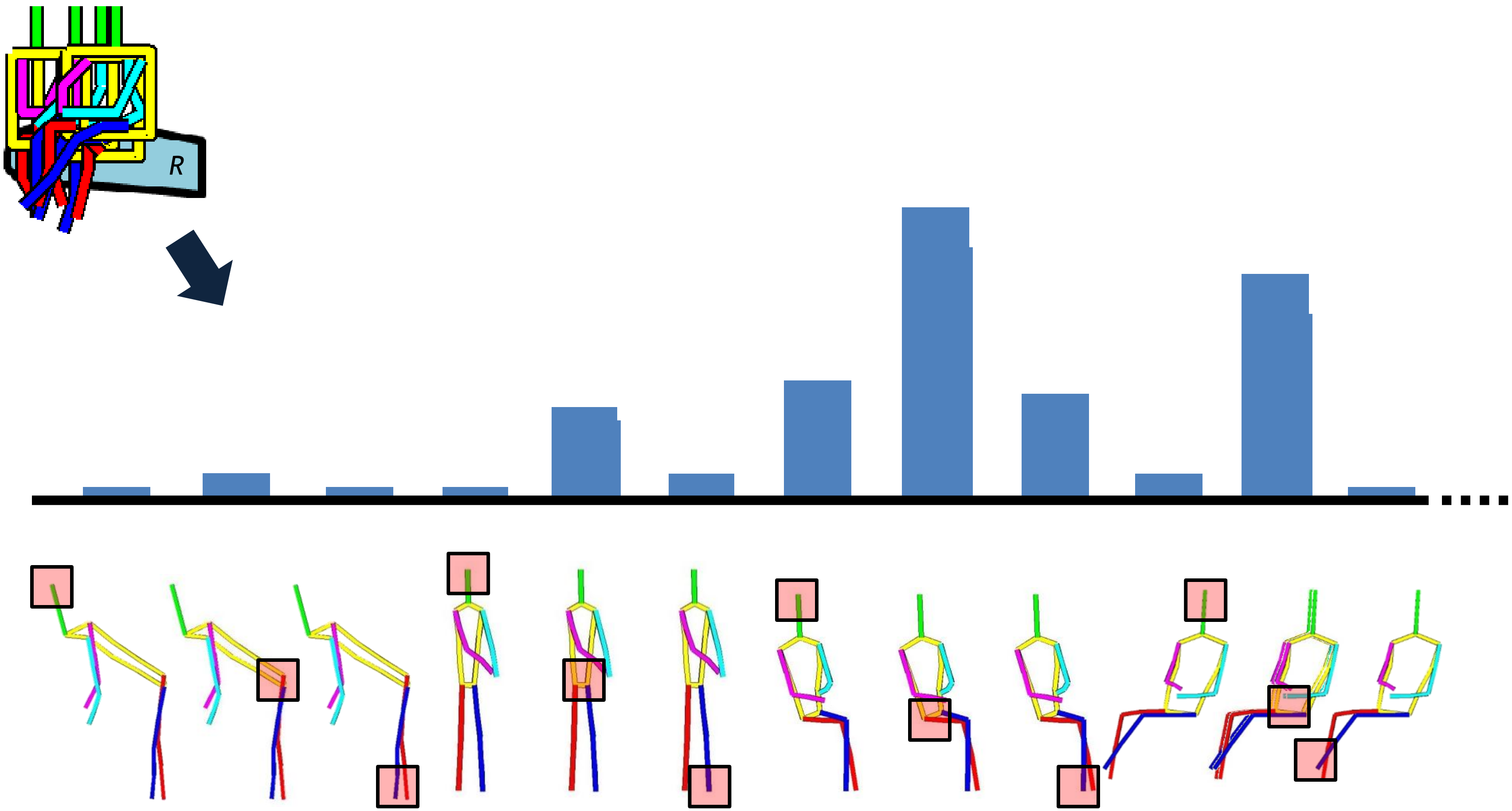
Semantic object segmentation



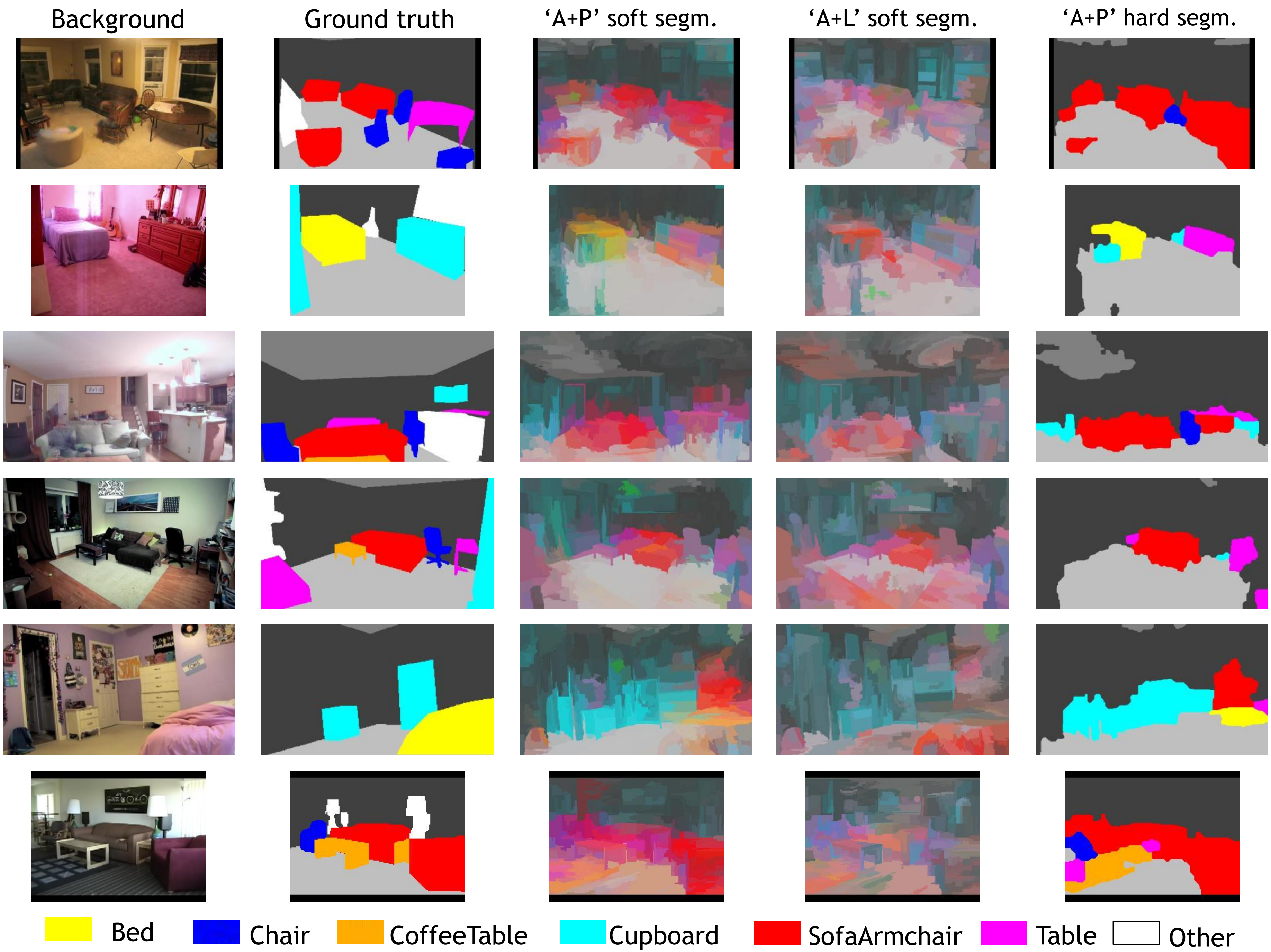
Pose vocabulary



Pose histogram



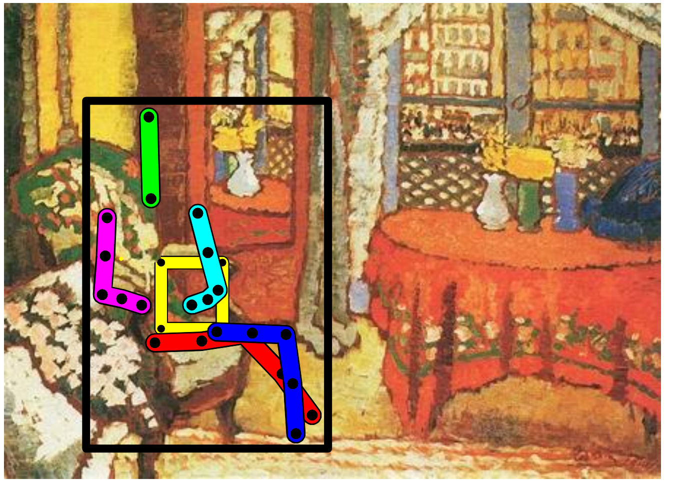
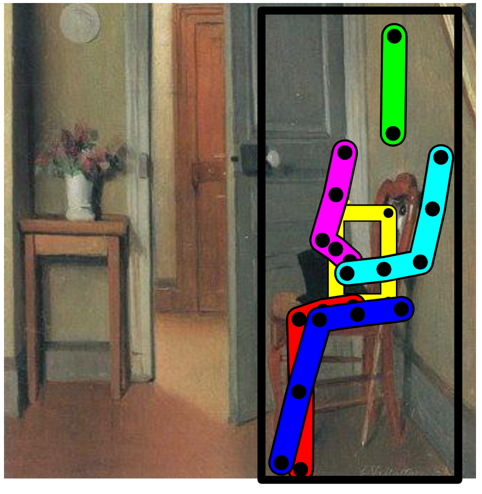
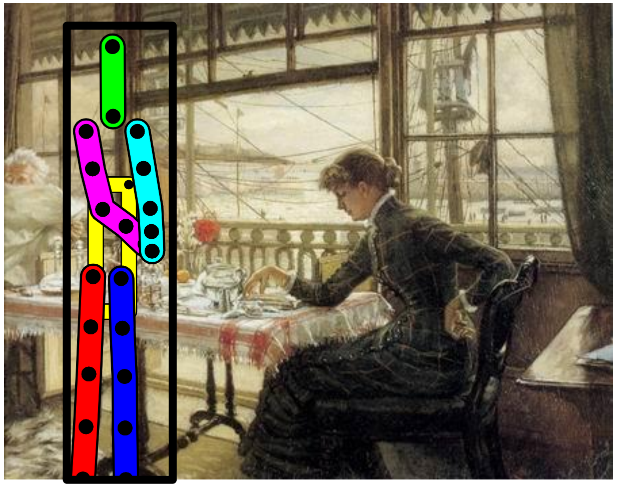
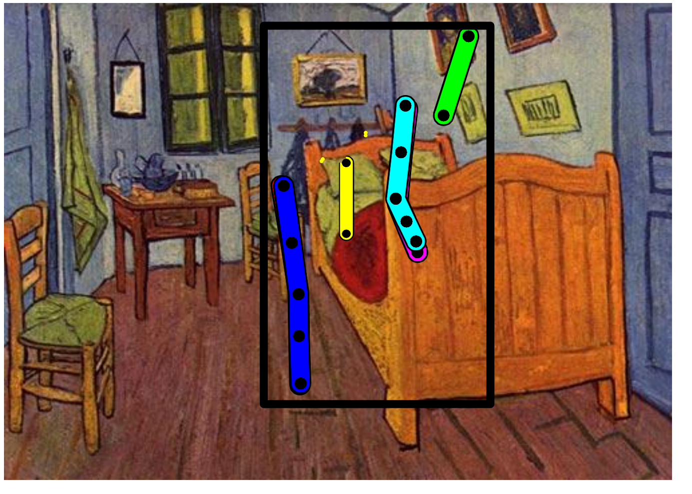
Some qualitative results


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

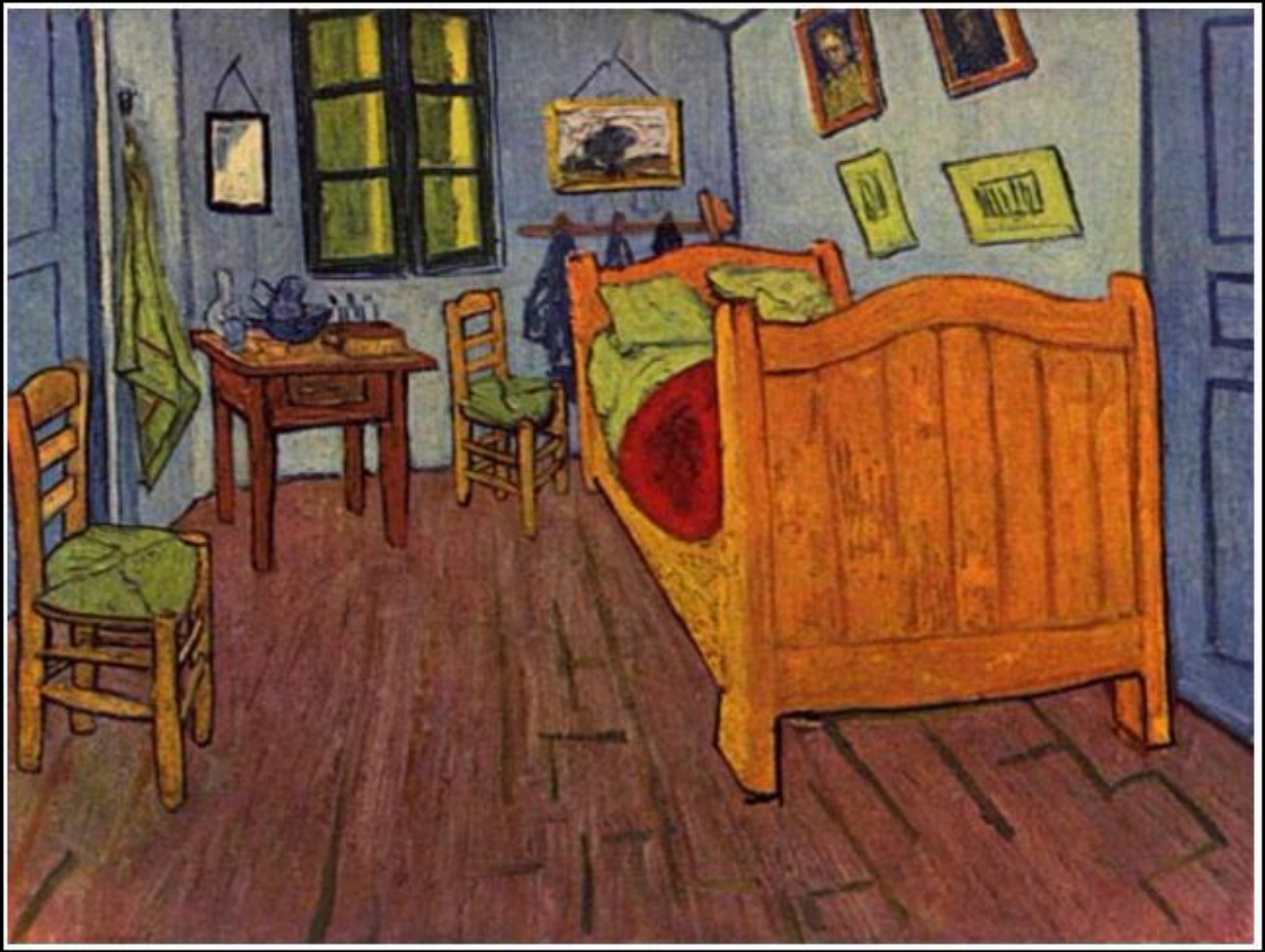
Using our model as pose prior

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

- Bag-of-Features methods give state-of-the-art results for action recognition in realistic data. Better models are needed
- Weakly-supervised methods crucial to address large-scale and large diversity of the video data.
- Video labeling by action classes is not the end of the story. New challenging problems are waiting.

Ad:
We are looking for Postdocs!

Willow, Paris