Human Activity and Vision Summer School
October 1-5, Sophia-Antipolis, France

Human Action Recognition

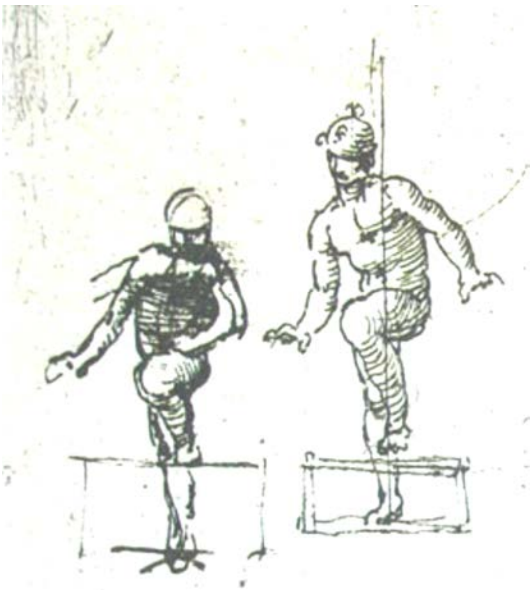
Ivan Laptev

ivan.laptev@inria.fr

INRIA, WILLOW, ENS/INRIA/CNRS UMR 8548
Laboratoire d'Informatique, Ecole Normale Supérieure, Paris

Includes slides from: Alyosha Efros and Andrew Zisserman

Lecture overview



Motivation

Historic review
Applications and challenges

Human Pose Estimation

Pictorial structures
Recent advances

Appearance-based methods

Motion history images
Active shape models & Motion priors

Motion-based methods

Generic and parametric Optical Flow
Motion templates

Space-time methods

Space-time features
Training with weak supervision

Why analyzing people and human actions?

History: Artistic Representation

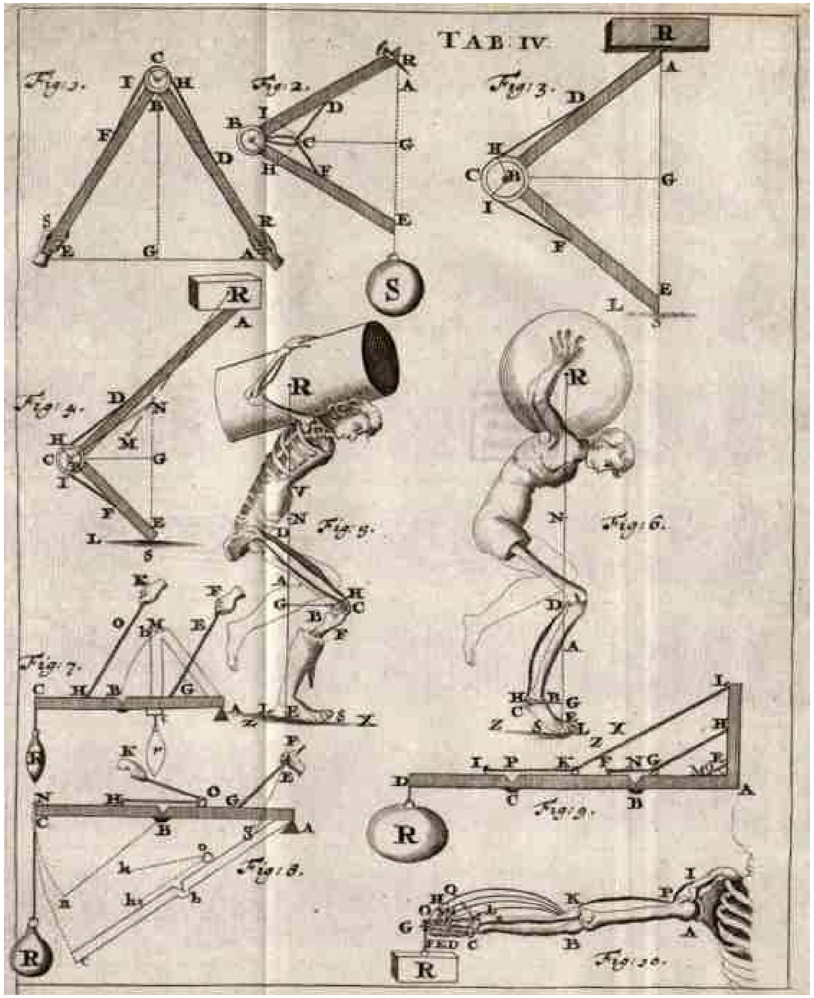
Early studies were motivated by human representations in Arts

Da Vinci: “it is indispensable for a painter, to become totally familiar with the anatomy of nerves, bones, muscles, and sinews, such that he understands for their various motions and stresses, which sinews or which muscle causes a particular motion”

“I ask for the weight [pressure] of this man for every segment of motion when climbing those stairs, and for the weight he places on *b* and on *c*. Note the vertical line below the center of mass of this man.”

Leonardo da Vinci (1452–1519): A man going upstairs, or up a ladder.

History: Biomechanics

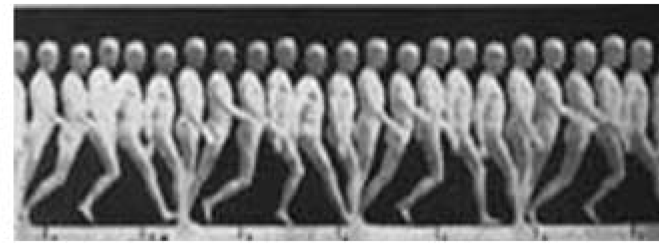


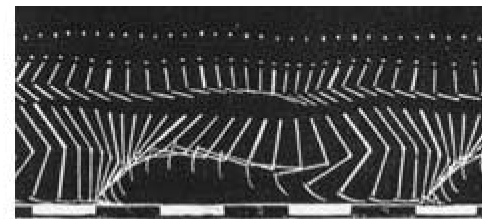
Giovanni Alfonso Borelli (1608–1679)

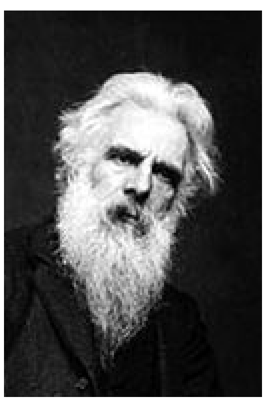
- The emergence of *biomechanics*
- Borelli applied to biology the analytical and geometrical methods, developed by Galileo Galilei
- He was the first to understand that bones serve as levers and muscles function according to mathematical principles
- His physiological studies included muscle analysis and a mathematical discussion of movements, such as running or jumping

History: Motion perception

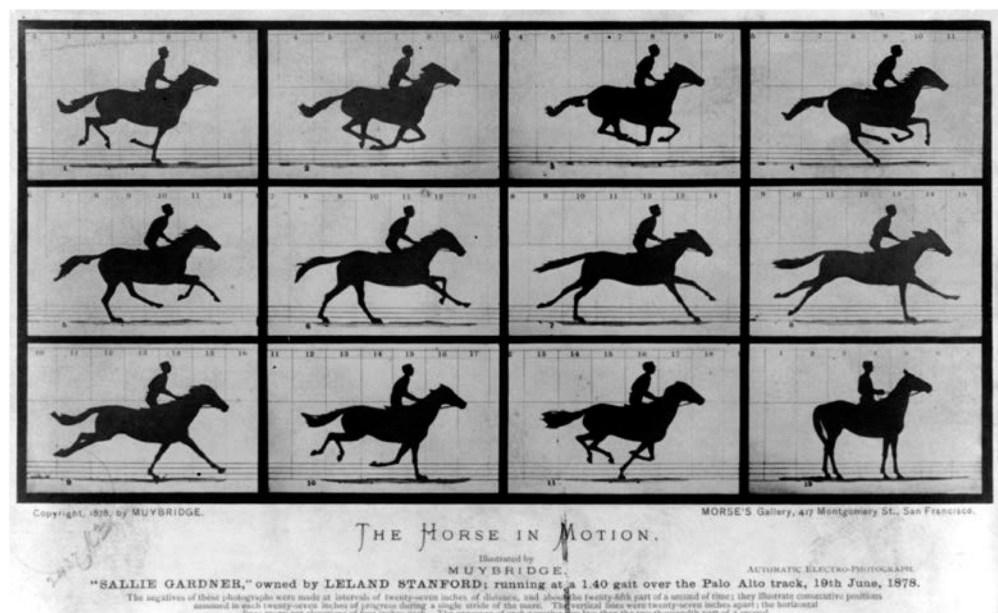
Etienne-Jules Marey: (1830–1904) made Chronophotographic experiments influential for the emerging field of *cinematography*





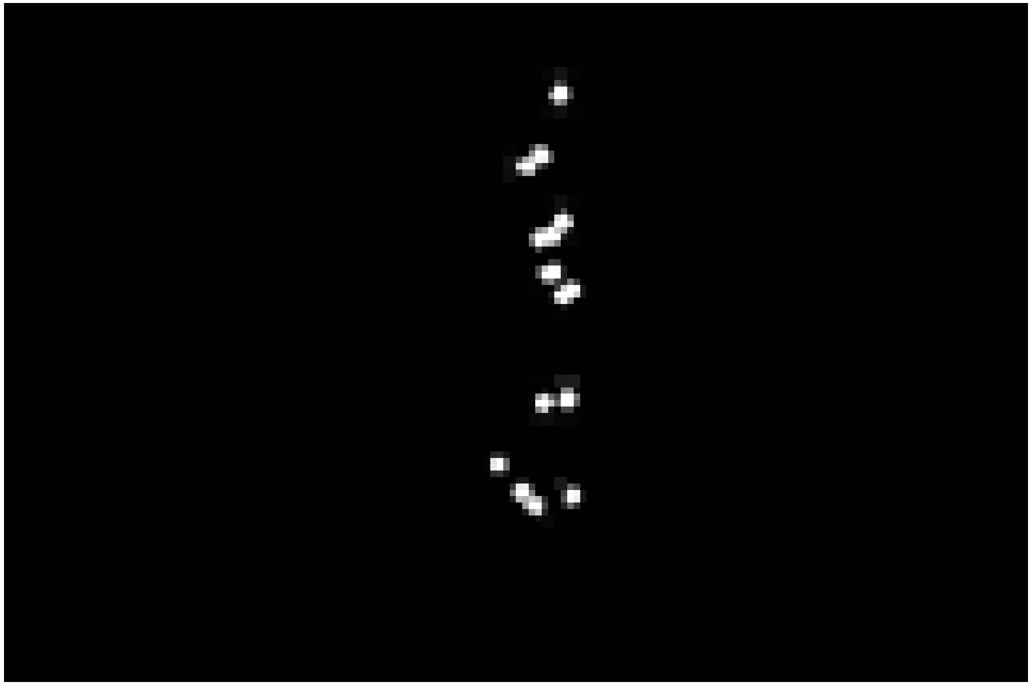


Eadweard Muybridge (1830–1904) invented a machine for displaying the recorded series of images. He pioneered motion pictures and applied his technique to movement studies



History: Motion perception

- Gunnar Johansson [1971] pioneered studies on the use of image sequences for a programmed human motion analysis
- “Moving Light Displays” (LED) enable identification of familiar people and the gender and inspired many works in computer vision.



Gunnar Johansson, **Perception and Psychophysics,** 1973

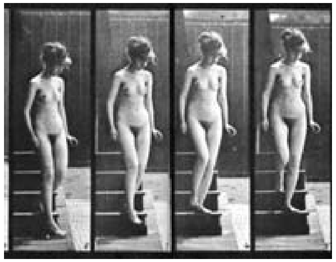
A HOUGHTON MIFFLIN PRODUCTION

A Teaching Resource
At the Frontiers of Psychological Inquiry

Human actions: Historic overview

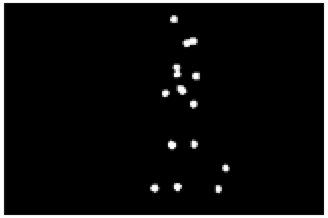
15th century
studies of
anatomy

17th century
emergence of
biomechanics



19th century
emergence of
cinematography

1971
studies of human
motion perception

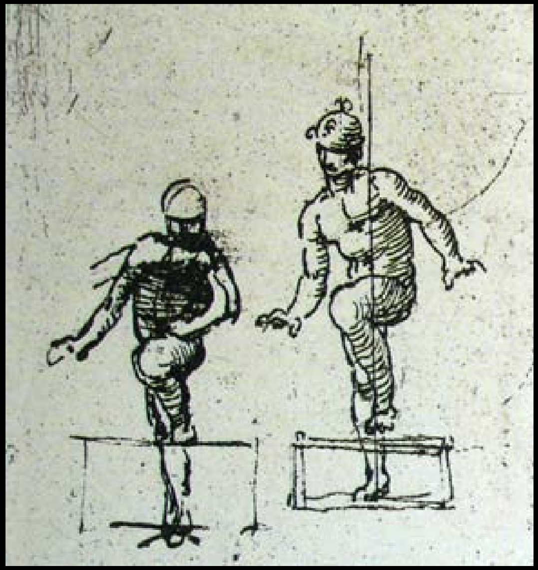


Modern computer vision

Modern applications: Motion capture and animation

Avatar (2009)

Modern applications: Motion capture and animation



Leonardo da Vinci (1452–1519)

Avatar (2009)

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?

- Surveillance

Where is my cat?

Predicting crowd behavior
Counting people

- Graphics

Motion capture and animation

How much data do we have?

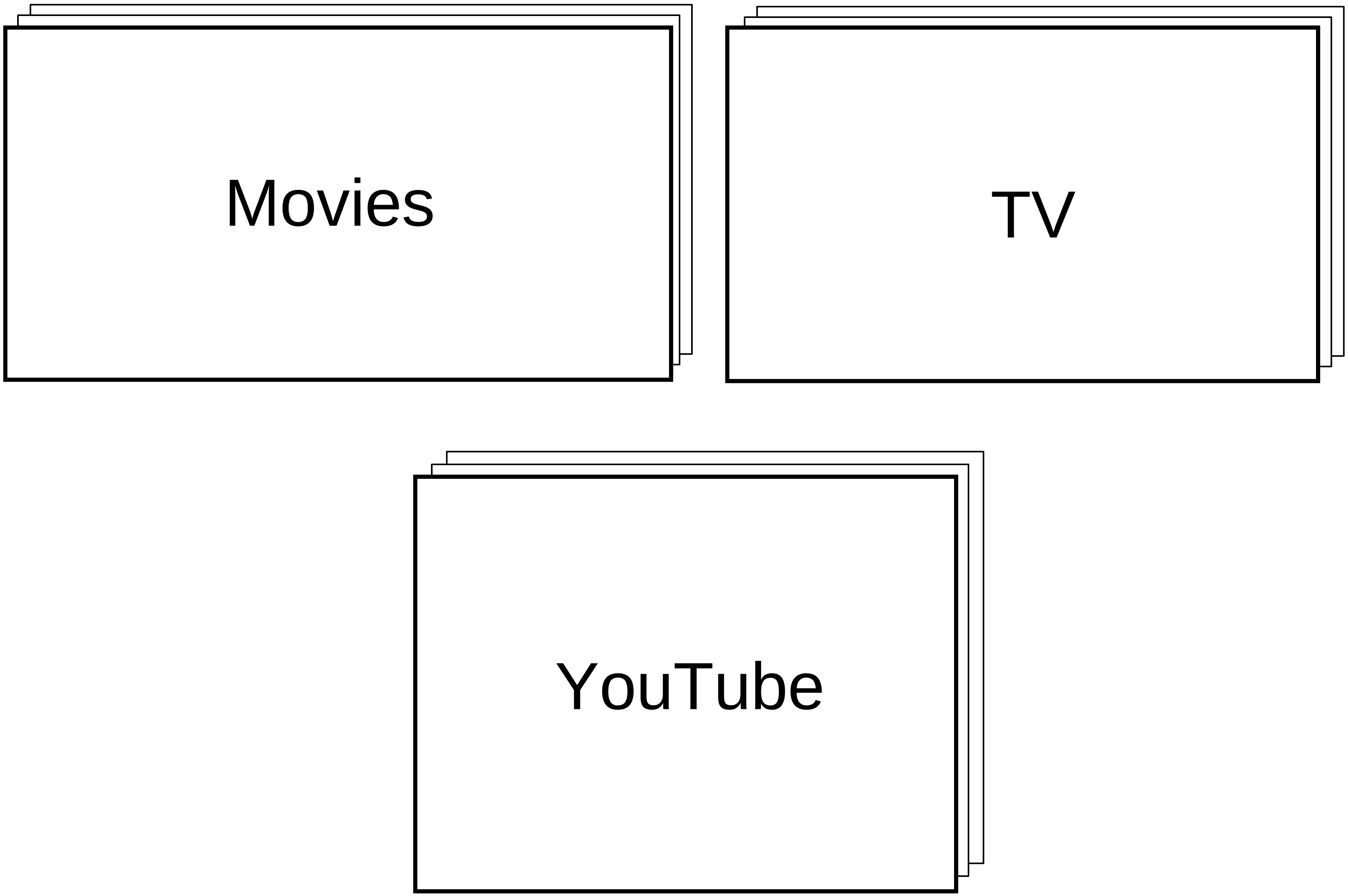
- Huge amount of video is available and growing

TV-channels recorded since 60's

>34K hours of video upload every day

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

How many person-pixels are there?



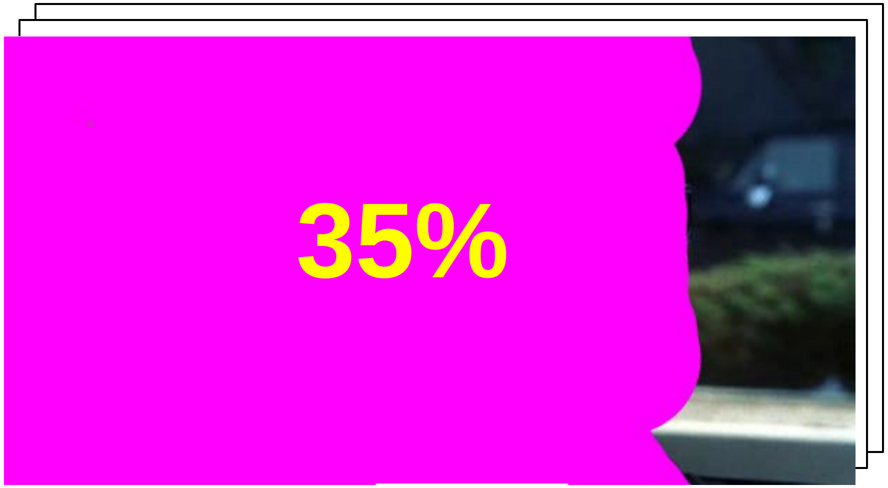
How many person-pixels are there?

Movies

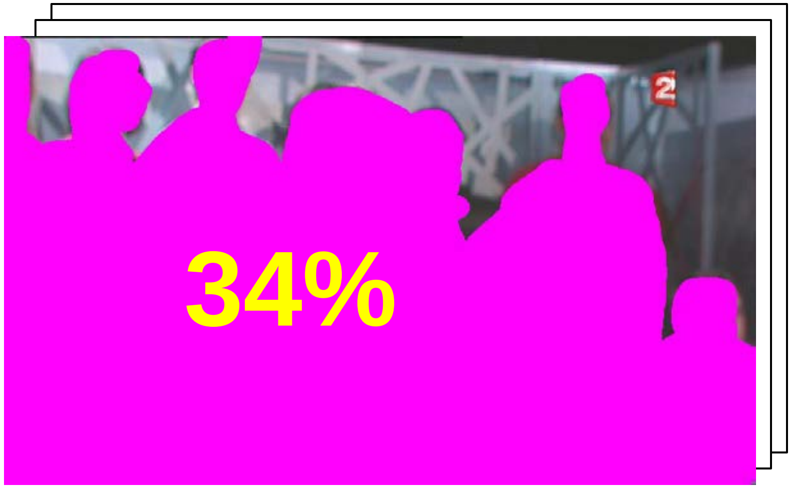
TV

YouTube

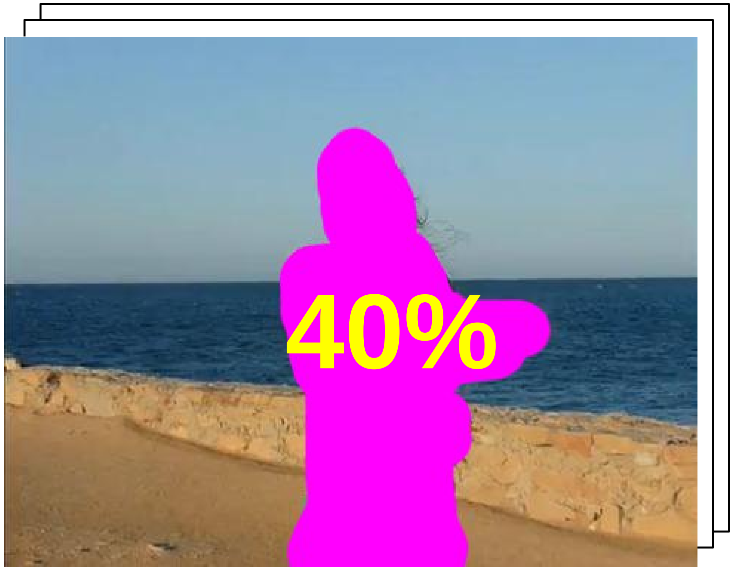
How many person-pixels are there?



Movies



TV



YouTube

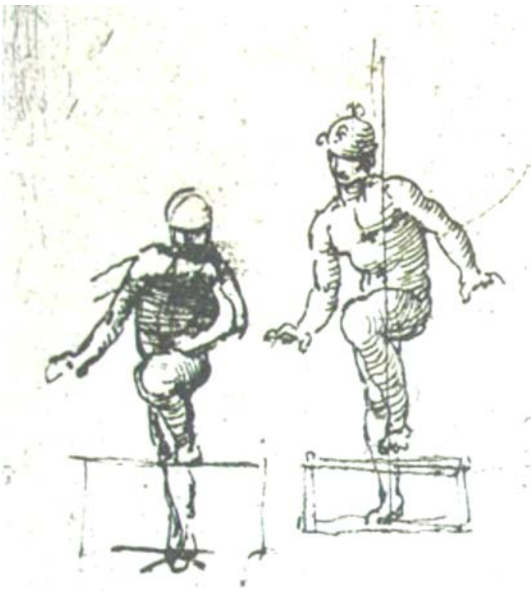
Why is action recognition hard?

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

Drinking

Smoking

Lecture overview



Motivation

Historic review
Applications and challenges

Human Pose Estimation

Pictorial structures
Recent advances

Appearance-based methods

Motion history images
Active shape models & Motion priors

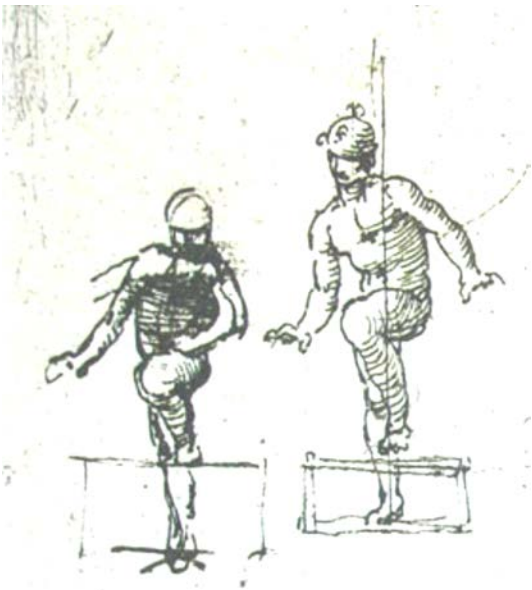
Motion-based methods

Generic and parametric Optical Flow
Motion templates

Space-time methods

Space-time features
Training with weak supervision

Lecture overview



Motivation

Historic review
Applications and challenges

Human Pose Estimation

Pictorial structures
Recent advances

Appearance-based methods

Motion history images
Active shape models & Motion priors

Motion-based methods

Generic and parametric Optical Flow
Motion templates

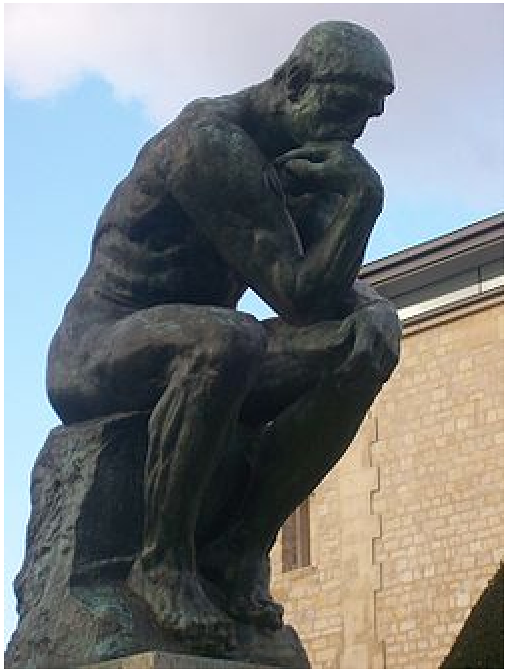
Space-time methods

Space-time features
Training with weak supervision

Activities characterized by a pose

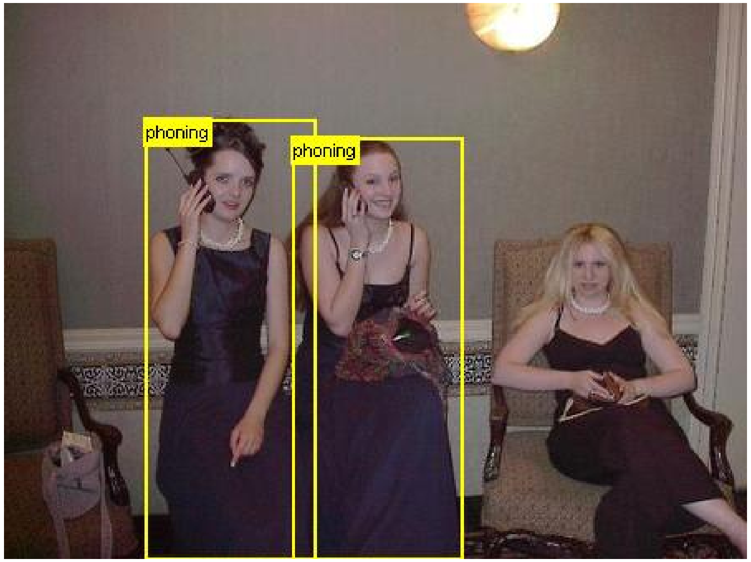
Slide credit: A. Zisserman

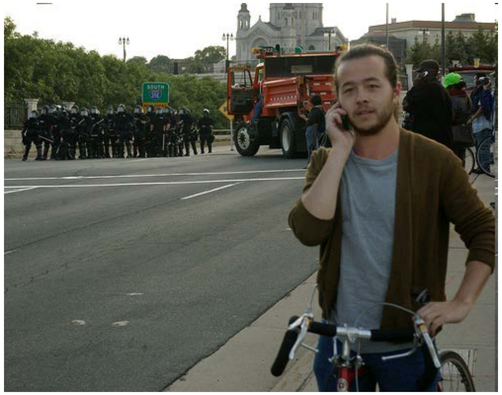
Activities characterized by a pose



Slide credit: A. Zisserman

Activities characterized by a pose



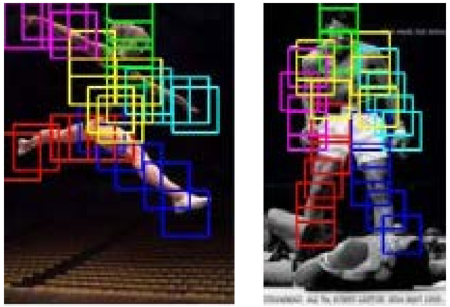


Challenges: articulations and deformations

Challenges: of (almost) unconstrained images

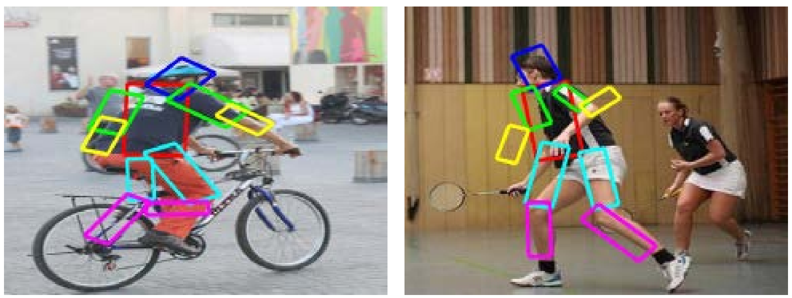
varying illumination and low contrast; moving camera and background; multiple people; scale changes; extensive clutter; any clothing

Pose estimation is an active research area



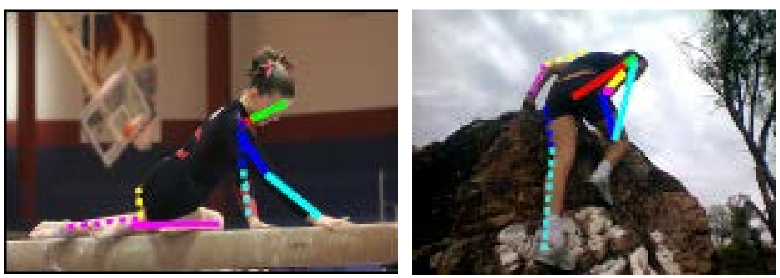
Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In Proc. **CVPR 2011**

Extension of LSVM model of Felzenszwalb et al.



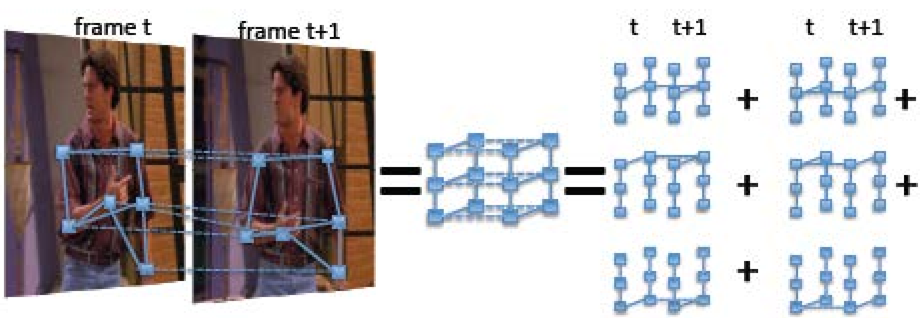
Y. Wang, D. Tran and Z. Liao. Learning Hierarchical Poselets for Human Parsing. In Proc. **CVPR 2011**.

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**.

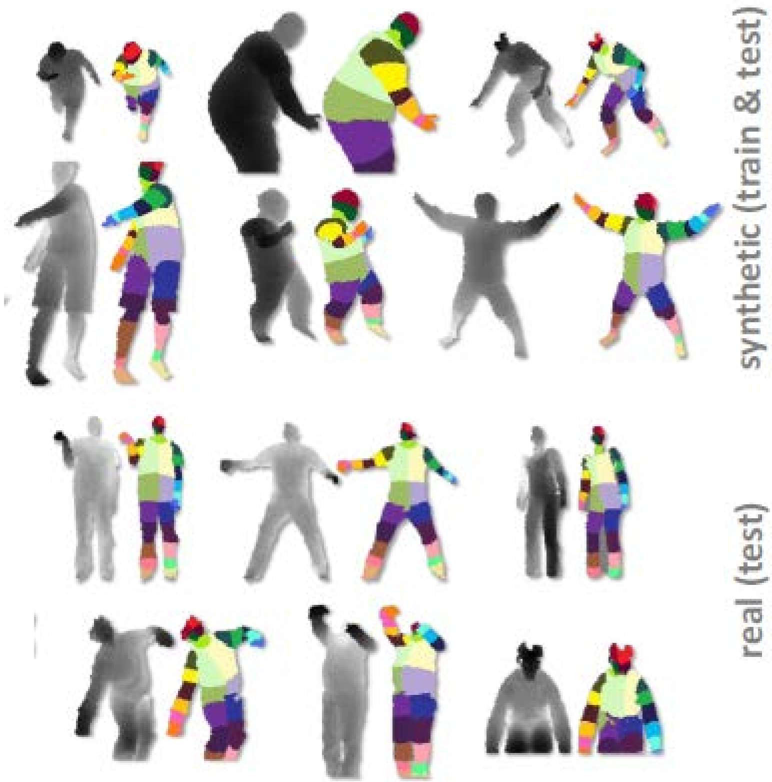
Learns from lots of noisy annotations

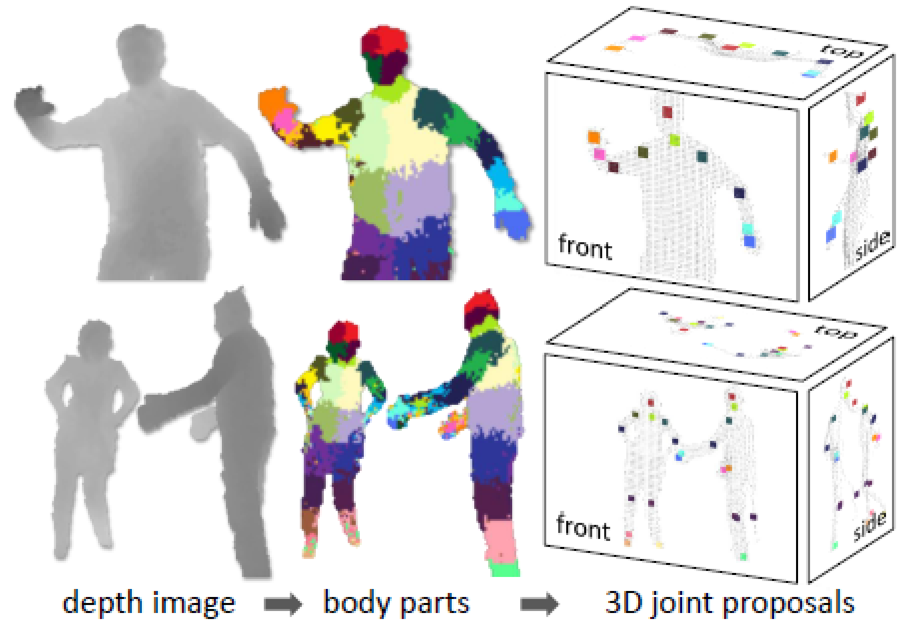


B. Sapp, D.Weiss and B. Taskar. Parsing Human Motion with Stretchable Models. In Proc. **CVPR 2011**.

Explores temporal continuity

Pose estimation is an active research area

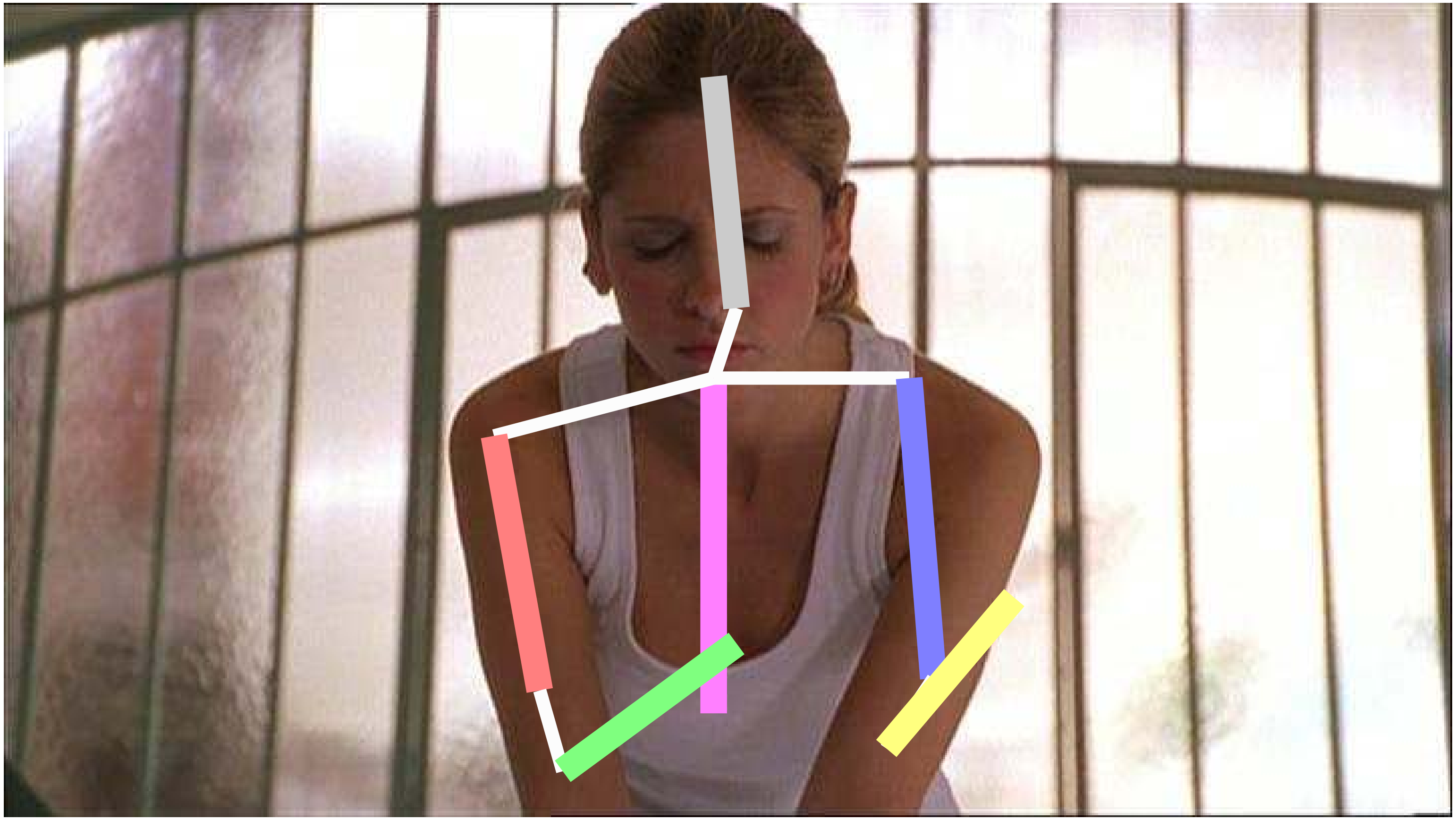




J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. **Best paper award at CVPR 2011**

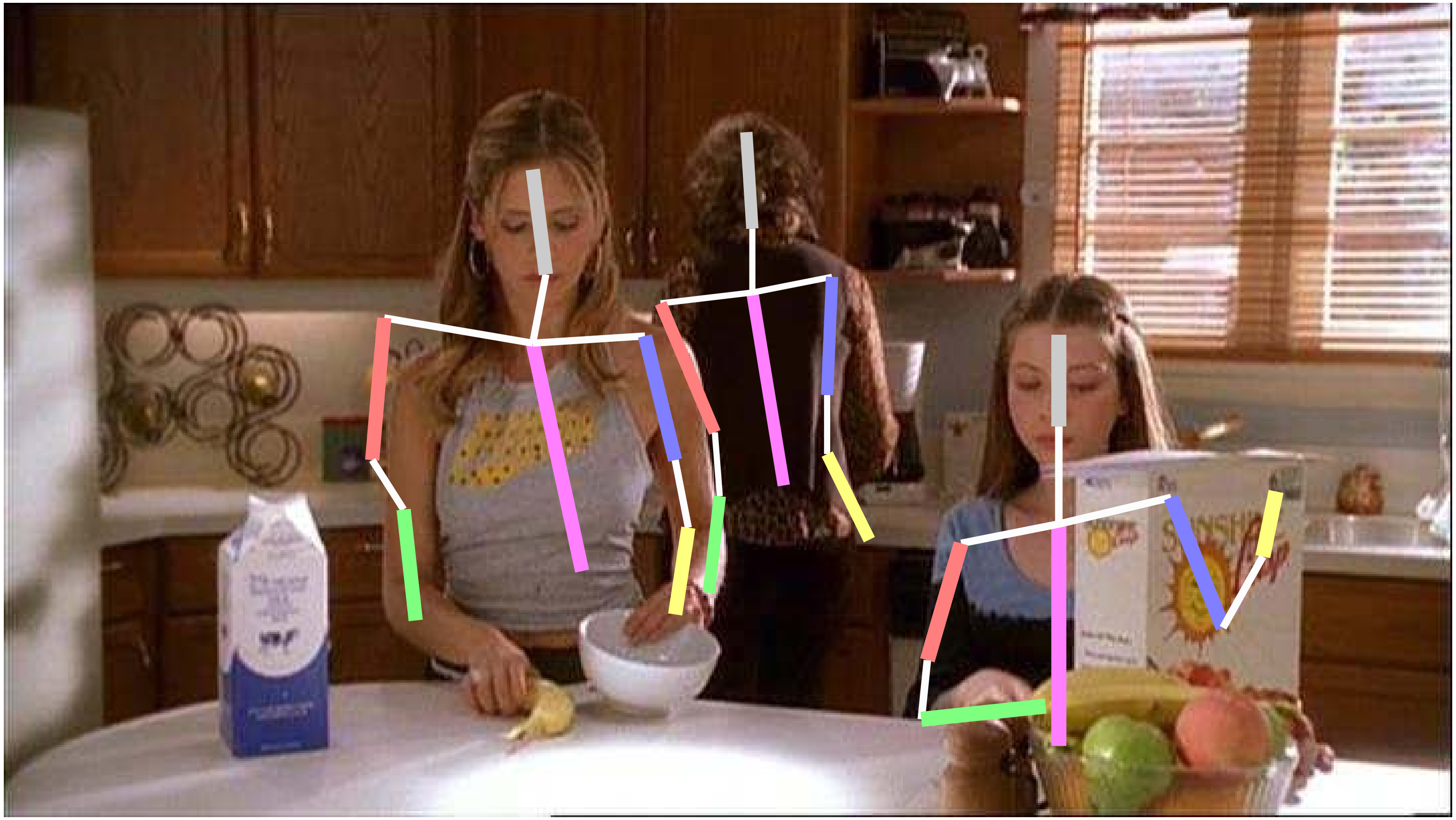
Exploits lots of synthesized depth images for training

What is missed?



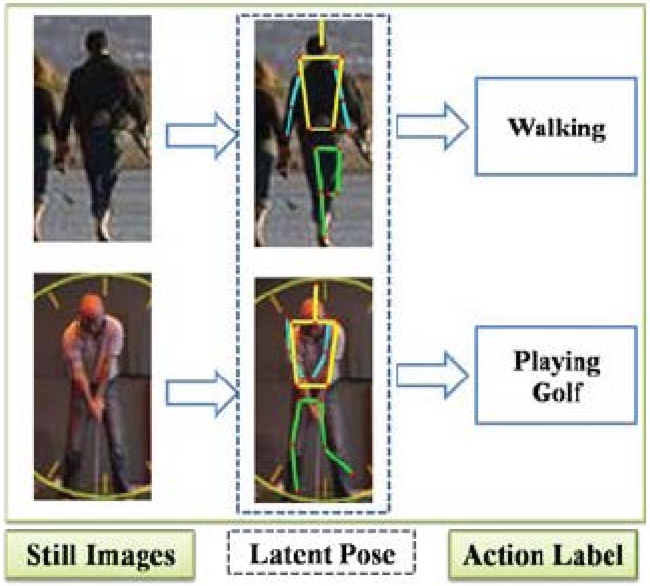
truncation is not modelled

What is missed?



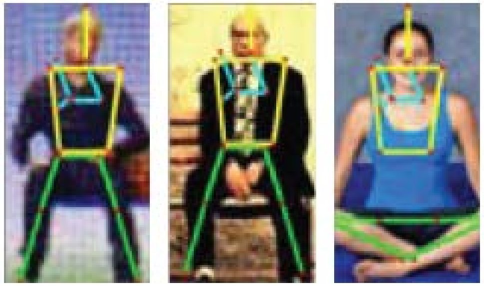
occlusion is not modelled

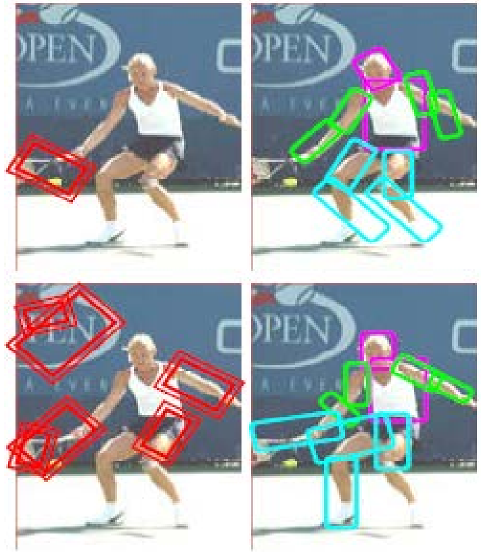
Modelling person-object-pose interactions



W. Yang, Y. Wang and Greg Mori. Recognizing Human Actions from Still Images with Latent Poses. In Proc. CVPR 2010.

Some limbs may not be important for recognizing a particular action (e.g. sitting)





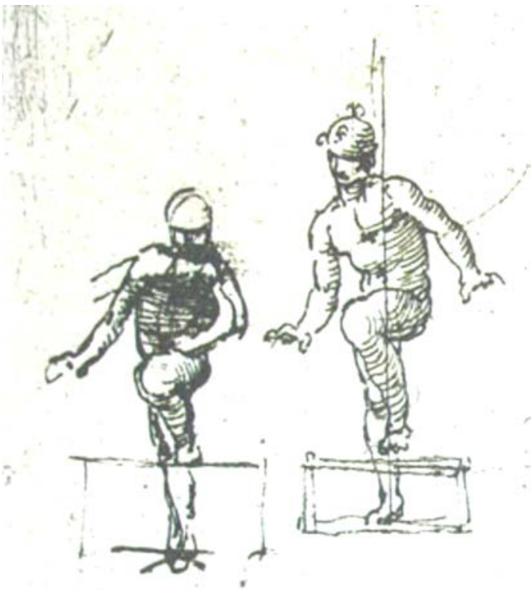
B. Yao and L. Fei-Fei. Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities. In Proc. CVPR 2010.

Pose estimation helps object detection and vice versa

Conclusion: Human poses

- Exciting progress in pose estimation in realistic still images and video.
- Industry-strength pose estimation from depth sensors
- Pose estimation from RGB is still very challenging
- Human Poses ≠ Human Actions!

Lecture overview



Motivation

Historic review
Applications and challenges

Human Pose Estimation

Pictorial structures
Recent advances

Appearance-based methods

Motion history images
Active shape models & Motion priors

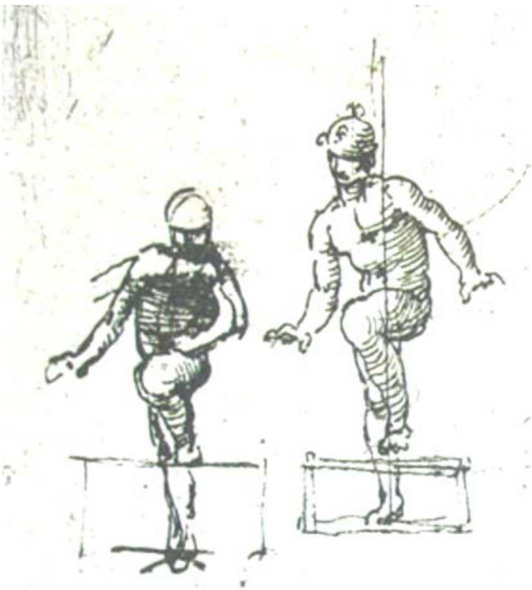
Motion-based methods

Generic and parametric Optical Flow
Motion templates

Space-time methods

Space-time features
Training with weak supervision

Lecture overview



Motivation

Historic review
Applications and challenges

Human Pose Estimation

Pictorial structures
Recent advances

Appearance-based methods

Motion history images
Active shape models & Motion priors

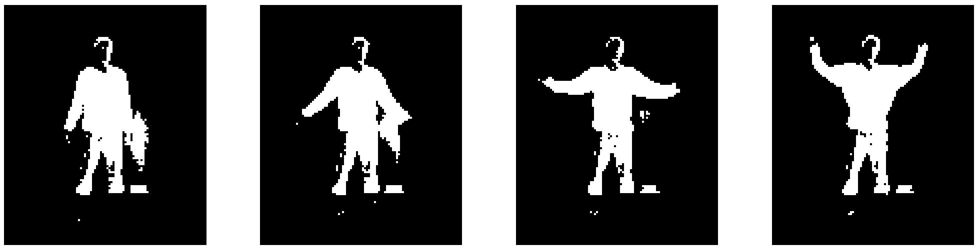
Motion-based methods

Generic and parametric Optical Flow
Motion templates

Space-time methods

Space-time features
Training with weak supervision

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

Person Tracking

[A. Baumberg and D. Hogg, ECCV'94]

Appearance methods: Shape

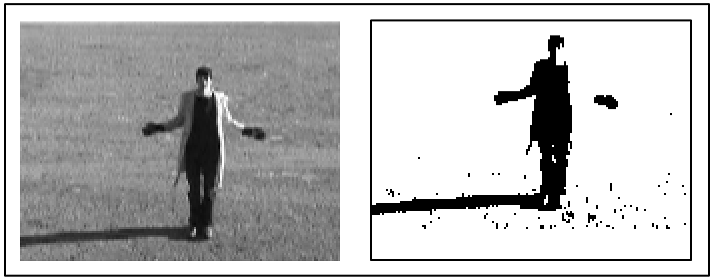
Pros:

+ Simple and fast

+ Works in controlled settings

Cons:

- Prone to errors of background subtraction



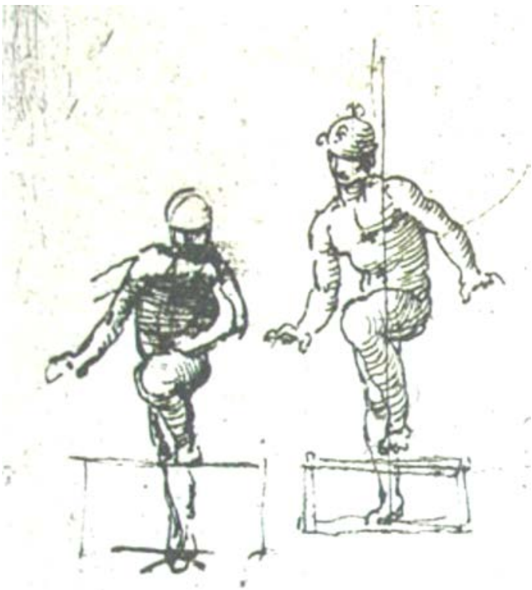
Variations in light, shadows, clothing…

What is the background here?

- Does not capture *interior* Structure and motion

Silhouette tells little about actions

Lecture overview



Motivation

Historic review
Applications and challenges

Human Pose Estimation

Pictorial structures
Recent advances

Appearance-based methods

Motion history images
Active shape models & Motion priors

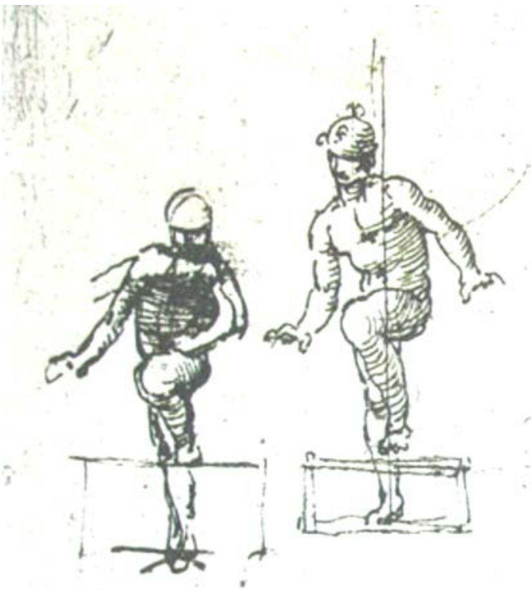
Motion-based methods

Generic and parametric Optical Flow
Motion templates

Space-time methods

Space-time features
Training with weak supervision

Lecture overview



Motivation

Historic review
Applications and challenges

Human Pose Estimation

Pictorial structures
Recent advances

Appearance-based methods

Motion history images
Active shape models & Motion priors

Motion-based methods

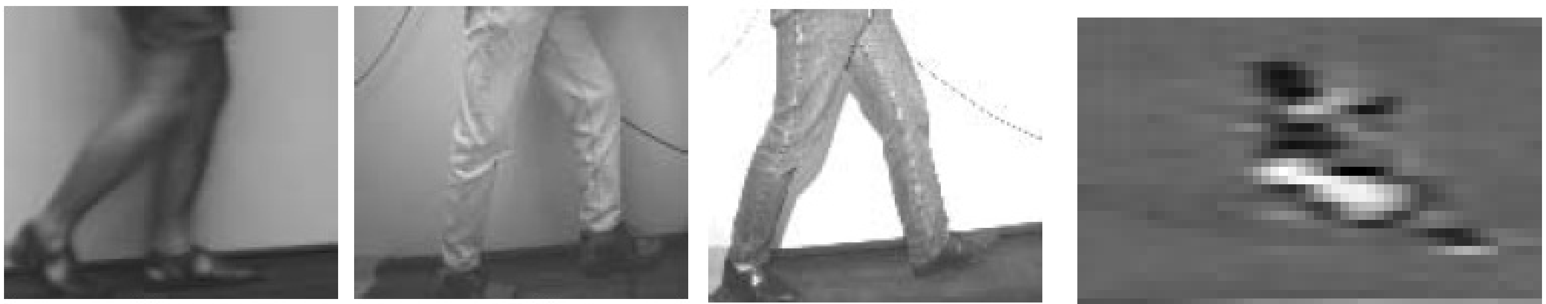
Generic and parametric Optical Flow
Motion templates

Space-time methods

Space-time features
Training with weak supervision

Shape and Appearance vs. Motion

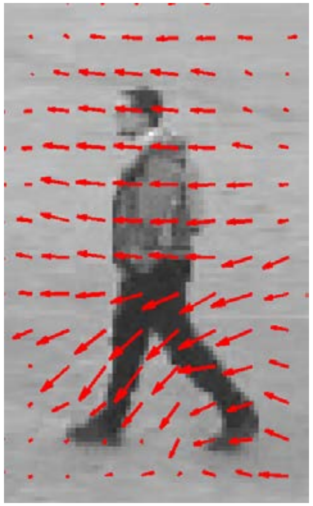
- Shape and appearance in images depends on many factors: clothing, illumination contrast, image resolution, etc…



[Efros et al. 2003]

- Estimated motion field is invariant to shape (in theory) and can be used directly to describe human actions

Motion estimation: Optical Flow



- Classic problem of computer vision [Gibson 1955]
- Goal: estimate motion field

How? We only have access to image pixels

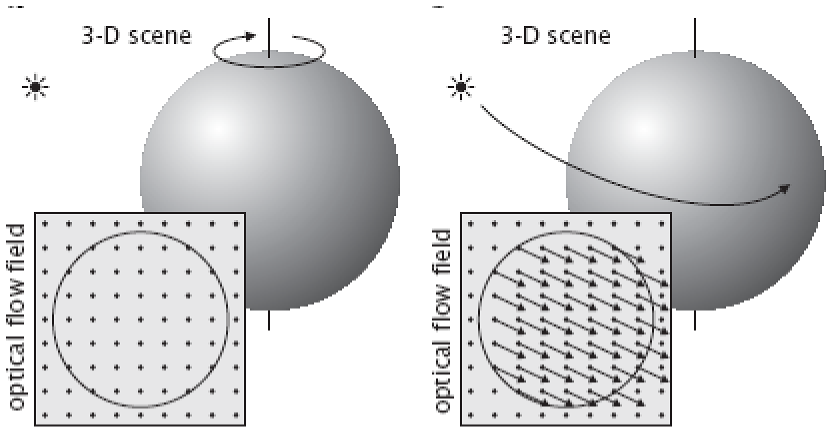
⇨ Estimate pixel-wise correspondence between frames = Optical Flow

- Brightness Change assumption: corresponding pixels preserve their intensity (color)

❖ Useful assumption in many cases

❖ Breaks at occlusions and illumination changes

❖ Physical and visual motion may be different



Parameterized Optical Flow

- Another extension of the constant motion model is to compute PCA basis flow fields from training examples

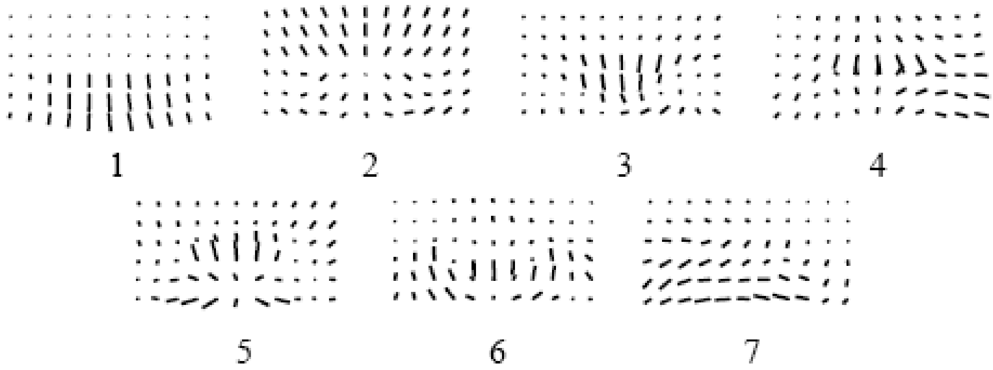
  1. Compute standard Optical Flow for many examples
  2. Put velocity components into one vector

  $$\mathbf{w} = (v_x^1, v_y^1, v_x^2, v_y^2, ..., v_x^n, v_y^n)^\top$$

  3. Do PCA on $\mathbf{w}$ and obtain most informative PCA flow basis vectors

Training samples

PCA flow bases

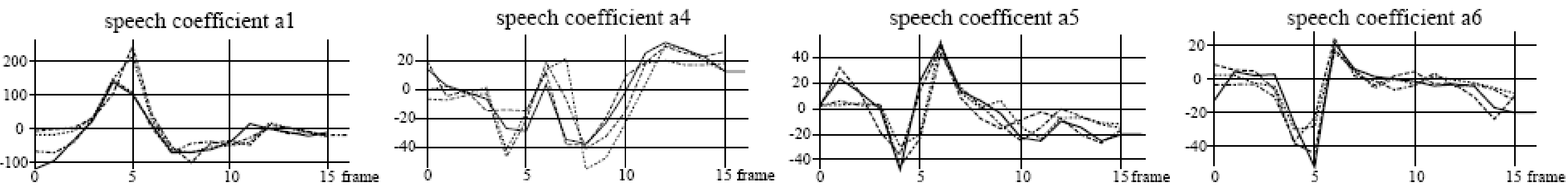


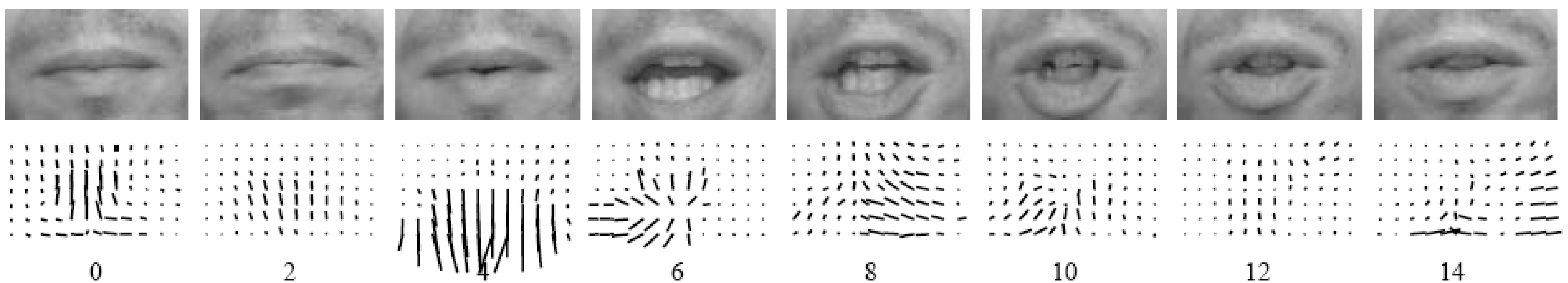
Learning Parameterized Models of Image Motion

[M.J. Black, Y. Yacoob, A.D. Jepson and D.J. Fleet, CVPR 1997]

Parameterized Optical Flow

- Estimated coefficients of PCA flow bases can be used as action descriptors





Frame numbers

➡ Optical flow seems to be an interesting descriptor for motion/action recognition

Spatial Motion Descriptor

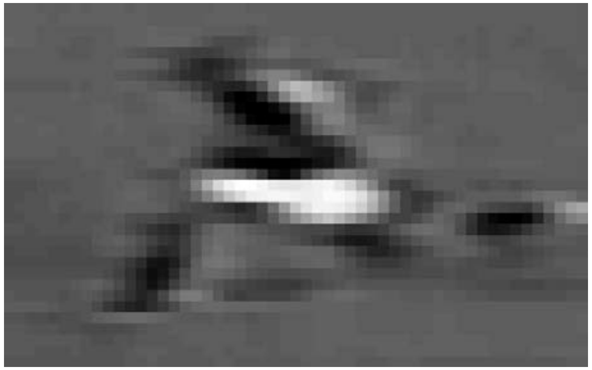
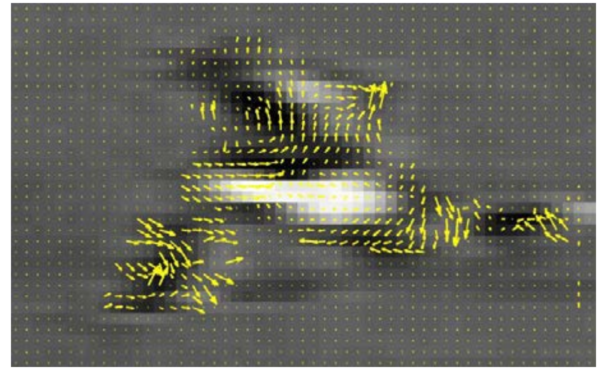
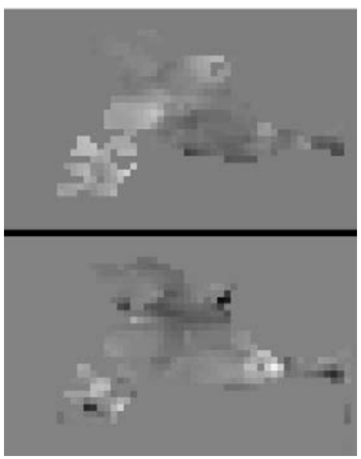


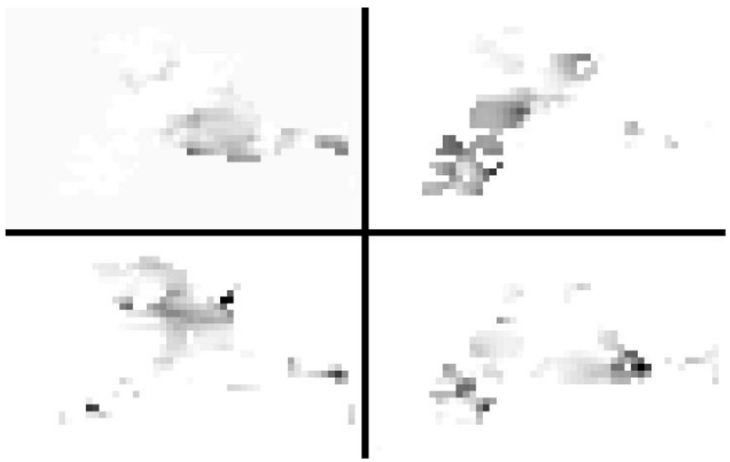
Image frame



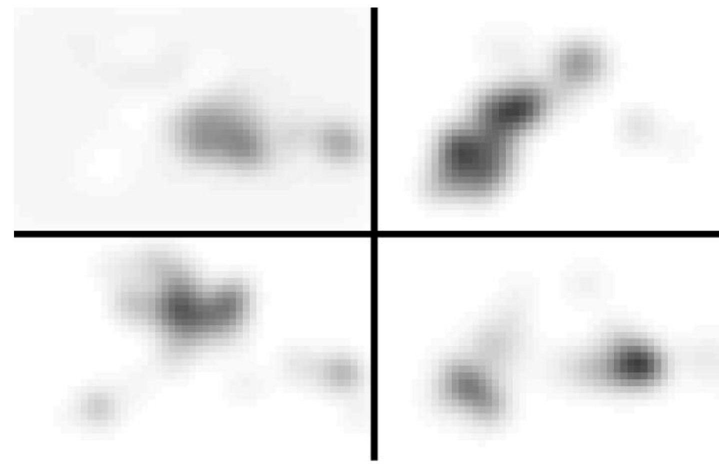
Optical flow $F_{x,y}$



F_x, F_y



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



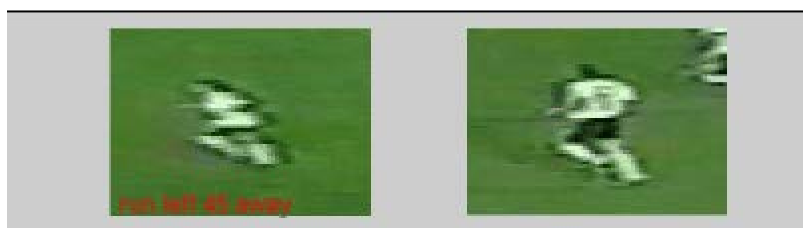
blurred $F_x^-, F_x^+, F_y^-, F_y^+$

A. A. Efros, A.C. Berg, G. Mori and J. Malik. Recognizing Action at a Distance. In Proc. ICCV 2003

Football Actions: matching

Input Sequence

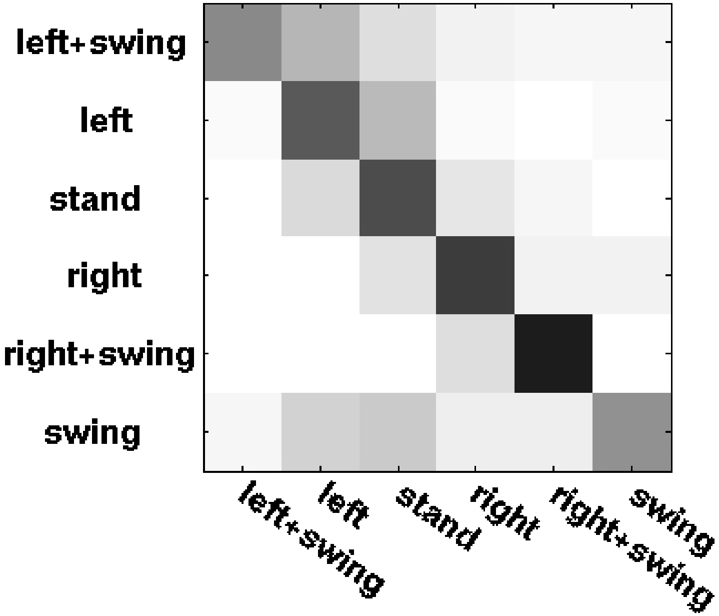
Matched Frames



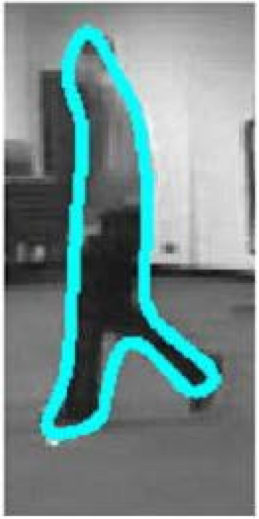
input matched

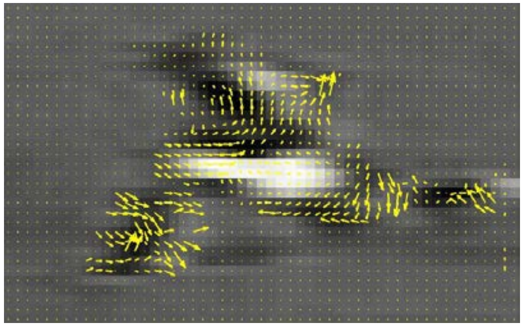
Classifying Tennis Actions

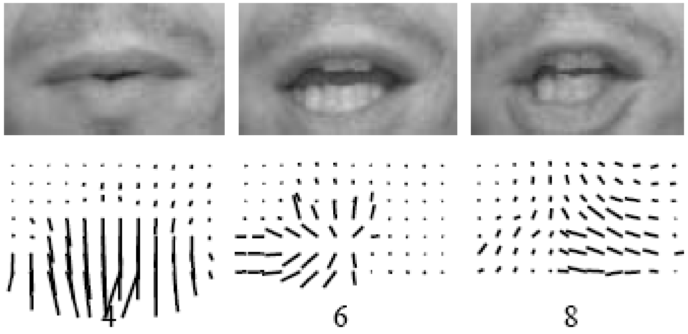
6 actions; 4600 frames; 7-frame motion descriptor
Woman player used as training, man as testing.



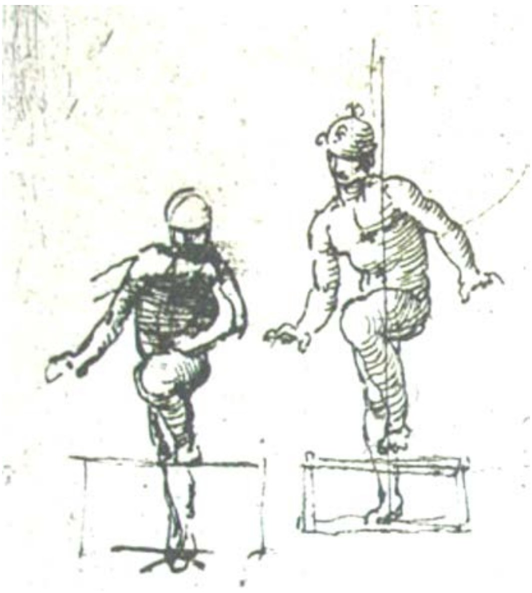
Summary so far…







Lecture overview



Motivation

Historic review
Modern applications

Human Pose Estimation

Pictorial structures
Recent advances

Appearance-based methods

Motion history images
Active shape models & Motion priors

Motion-based methods

Generic and parametric Optical Flow
Motion templates

Space-time methods

Space-time features
Training with weak supervision

Goal:
Interpret complex dynamic scenes

Common methods:

- Segmentation **?**
- Tracking **?**

Common problems:

- Complex & changing BG
- Changing appearance

⇒ No global assumptions about the scene

Space-time

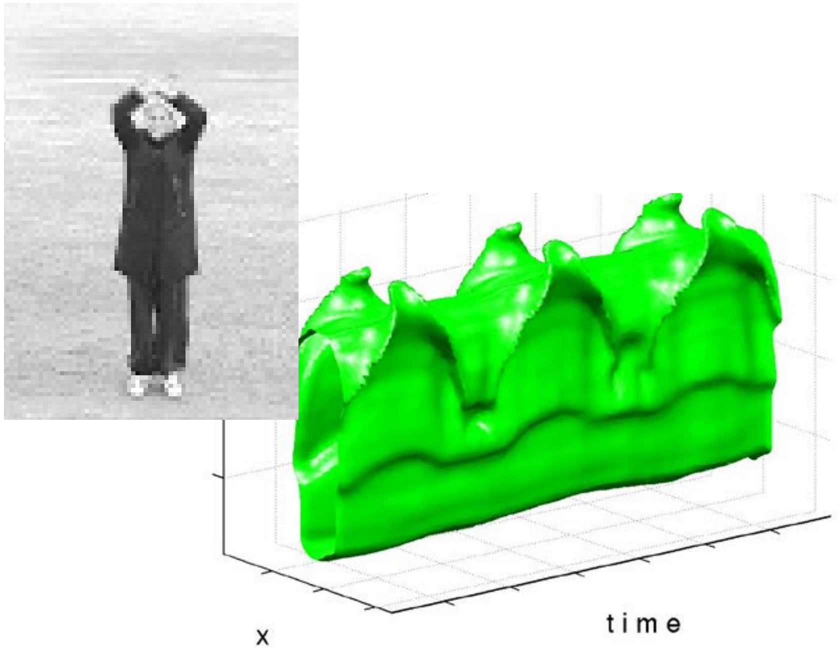
No global assumptions $\Rightarrow$

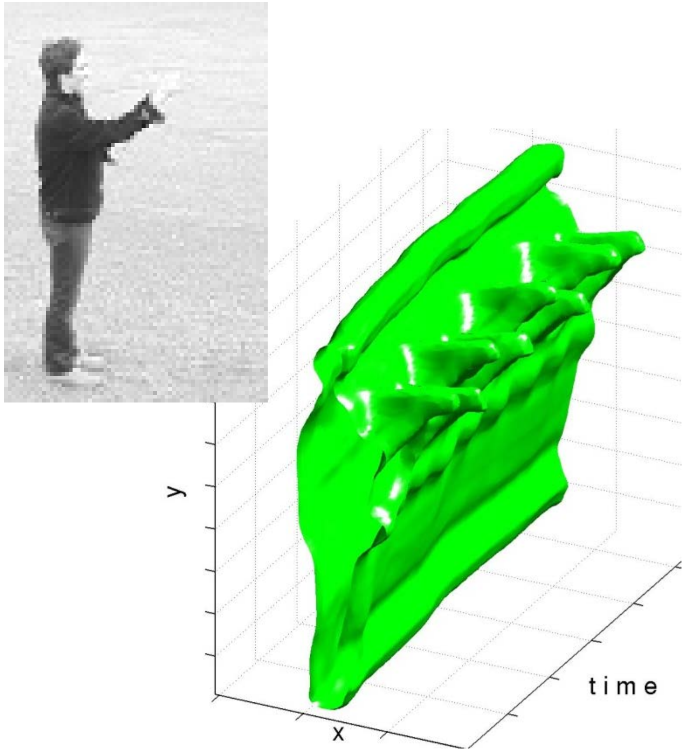
Consider local spatio-temporal neighborhoods

hand waving

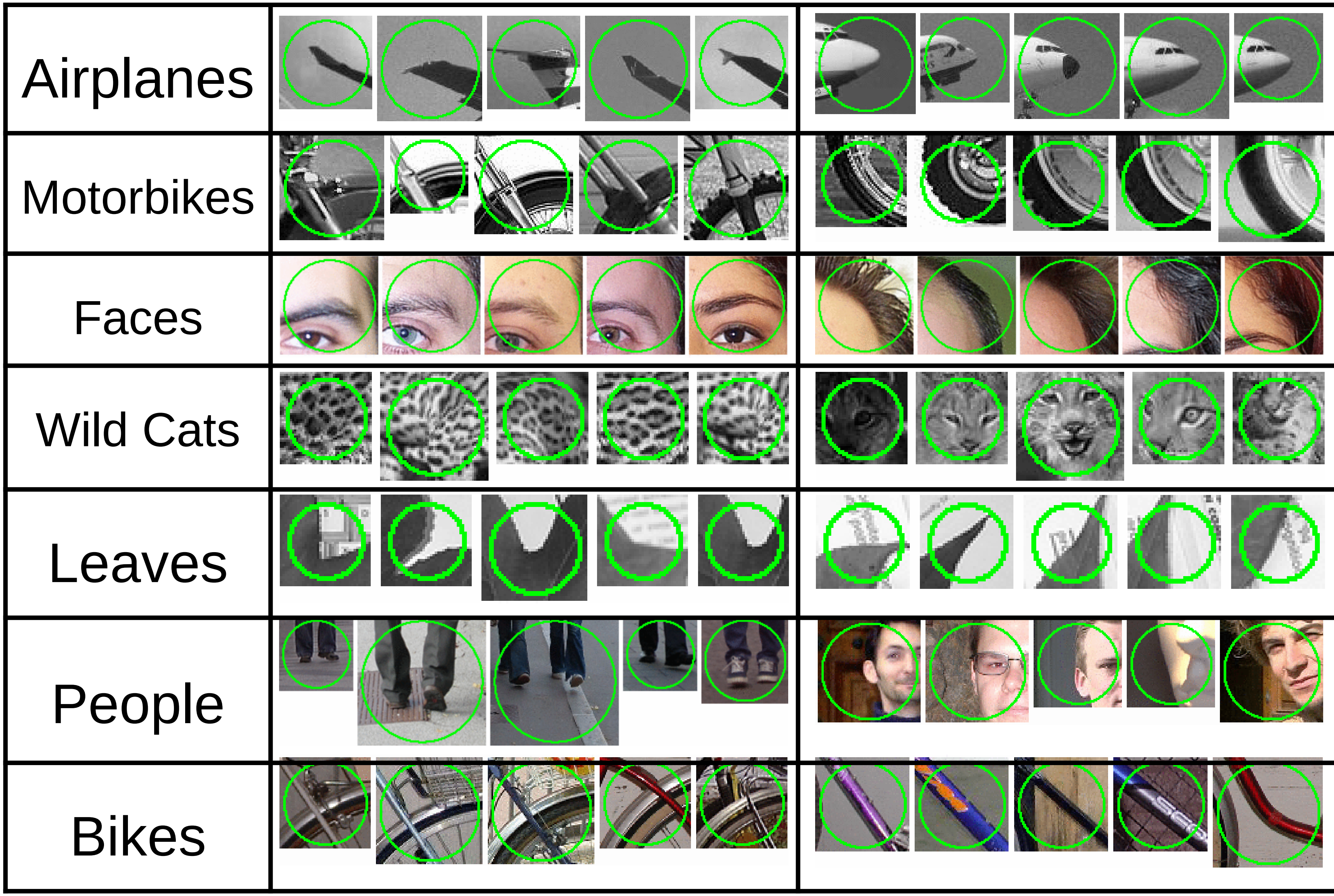
boxing

Actions == Space-time objects?

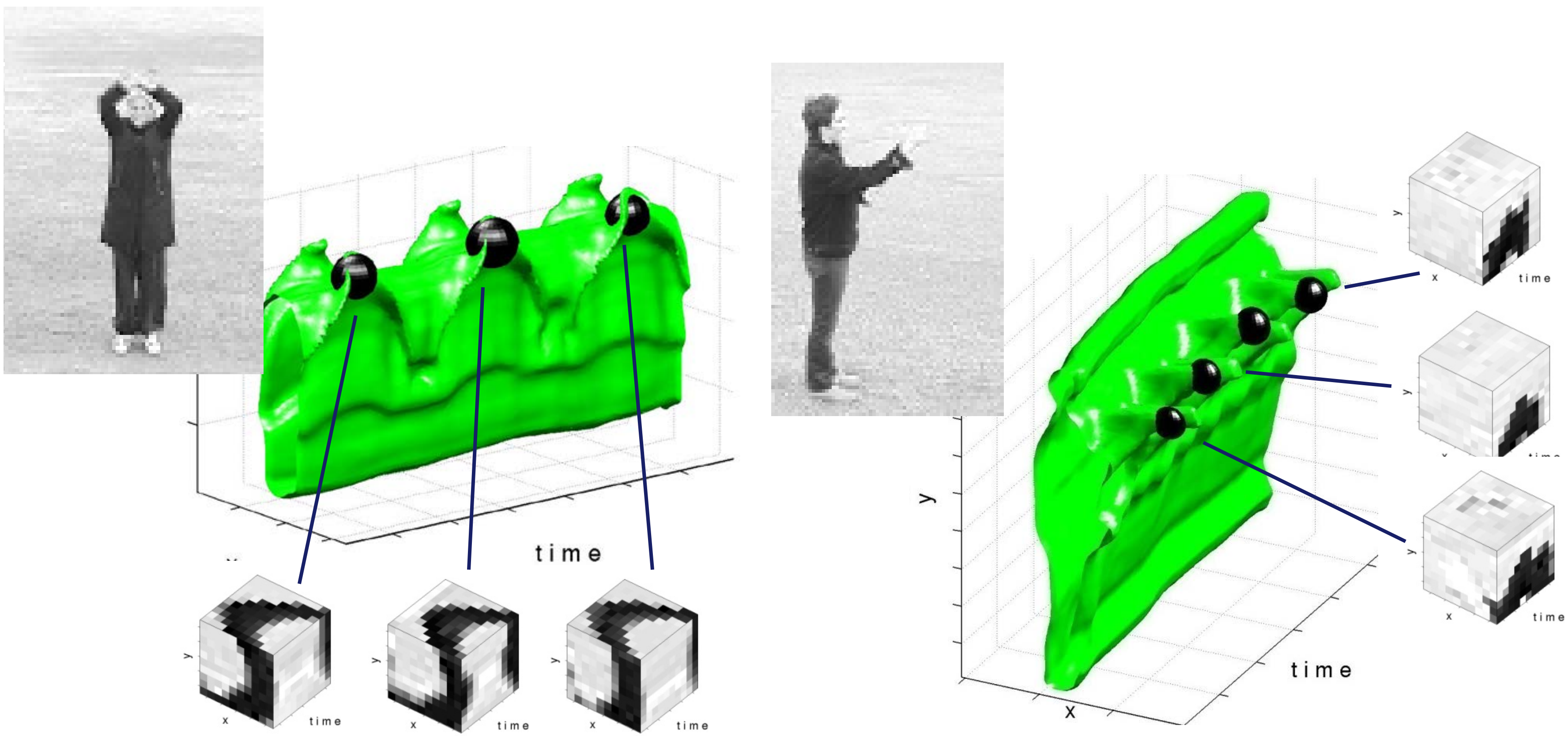




Local approach: Bag of Visual Words



Space-time local features



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

Space-Time Interest Points: Detection

Properties of $\mu(\cdot;\ \Sigma)$

$\mu(\cdot;\ \Sigma)$ defines second order approximation for the local distribution of ∇L within neighborhood Σ

$\text{rank}(\mu) = 1 \quad \Rightarrow$ 1D space-time variation of f e.g. moving bar

$\text{rank}(\mu) = 2 \quad \Rightarrow$ 2D space-time variation of f e.g. moving ball

$\text{rank}(\mu) = 3 \quad \Rightarrow$ 3D space-time variation of f e.g. jumping ball

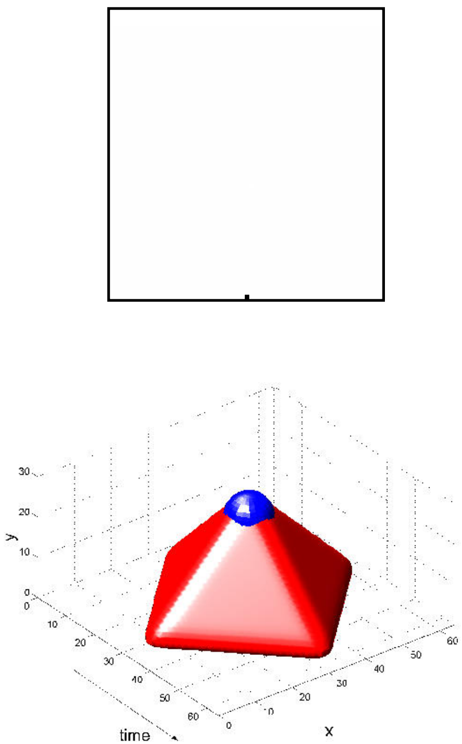
Large eigenvalues of μ can be detected by the local maxima of H over (x,y,t):

$$\begin{aligned} H(p;\ \Sigma) &= \text{det}(\mu(p;\ \Sigma)) + k\text{trace}^3(\mu(p;\ \Sigma)) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3 \end{aligned}$$

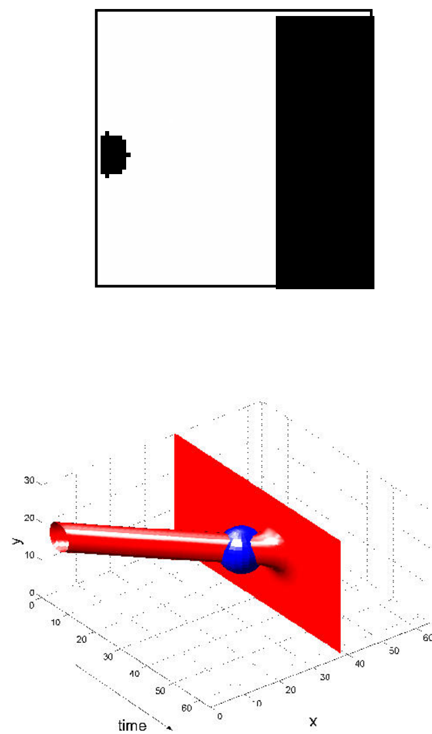
(similar to Harris operator [Harris and Stephens, 1988])

Space-Time interest points

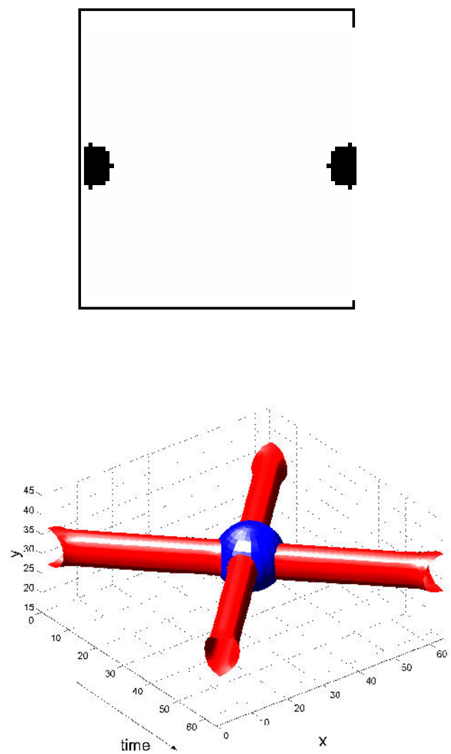
Velocity changes



appearance/ disappearance

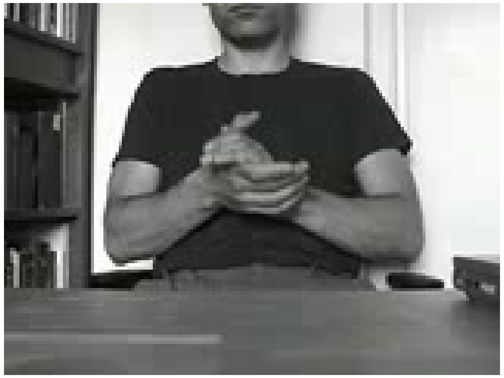


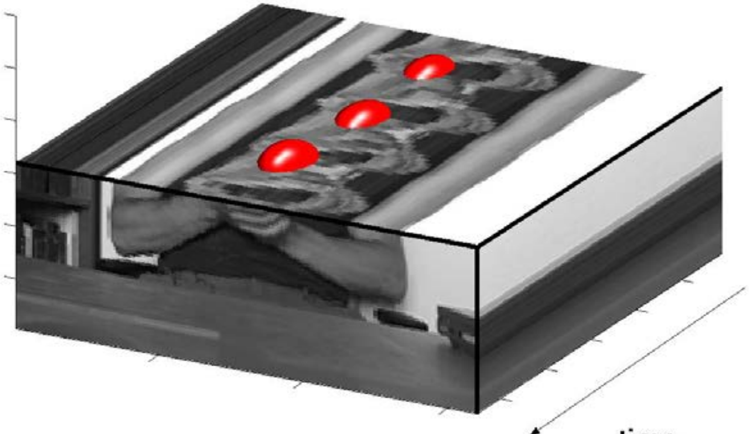
split/merge

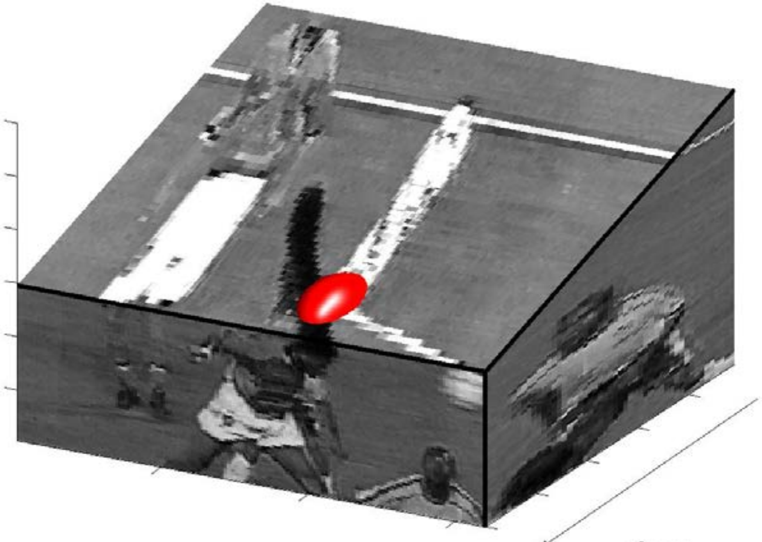


Space-Time Interest Points: Examples

Motion event detection

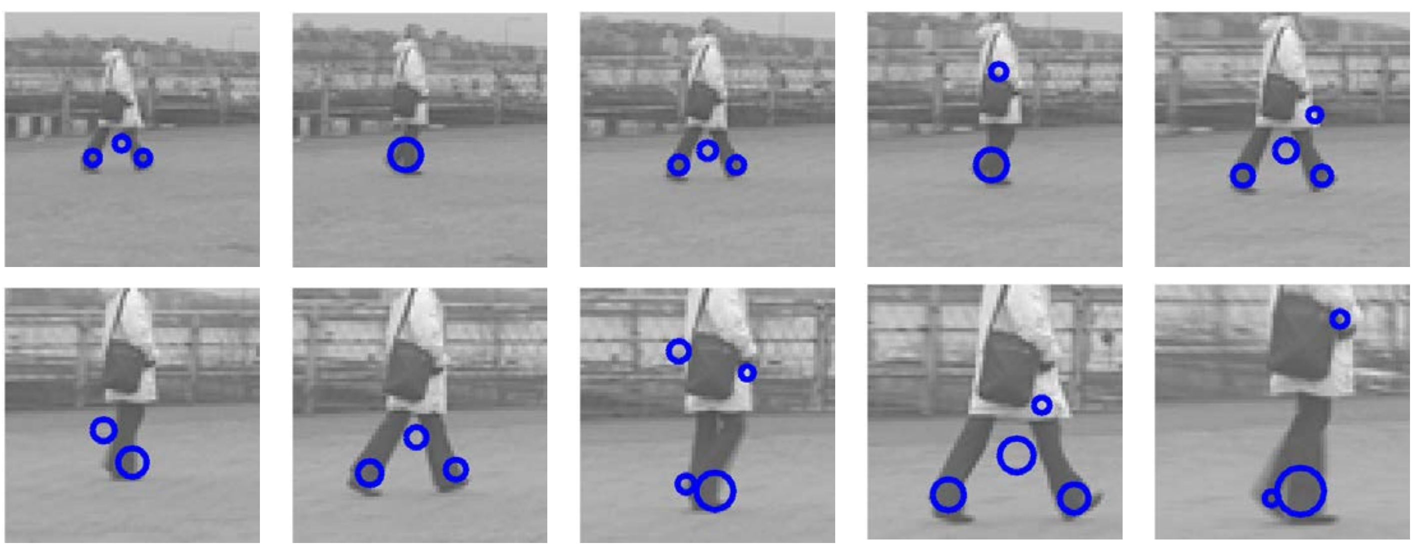




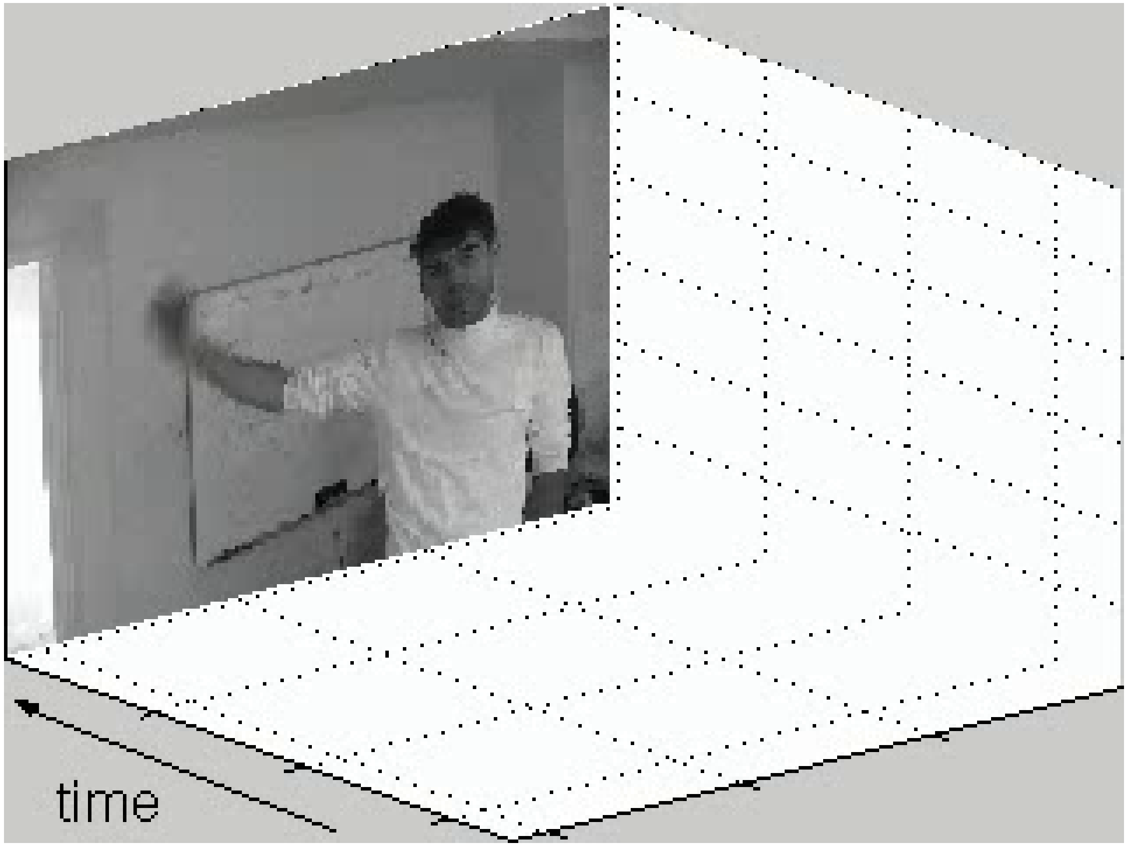


Spatio-temporal scale selection

Stability to size changes,
e.g. camera zoom

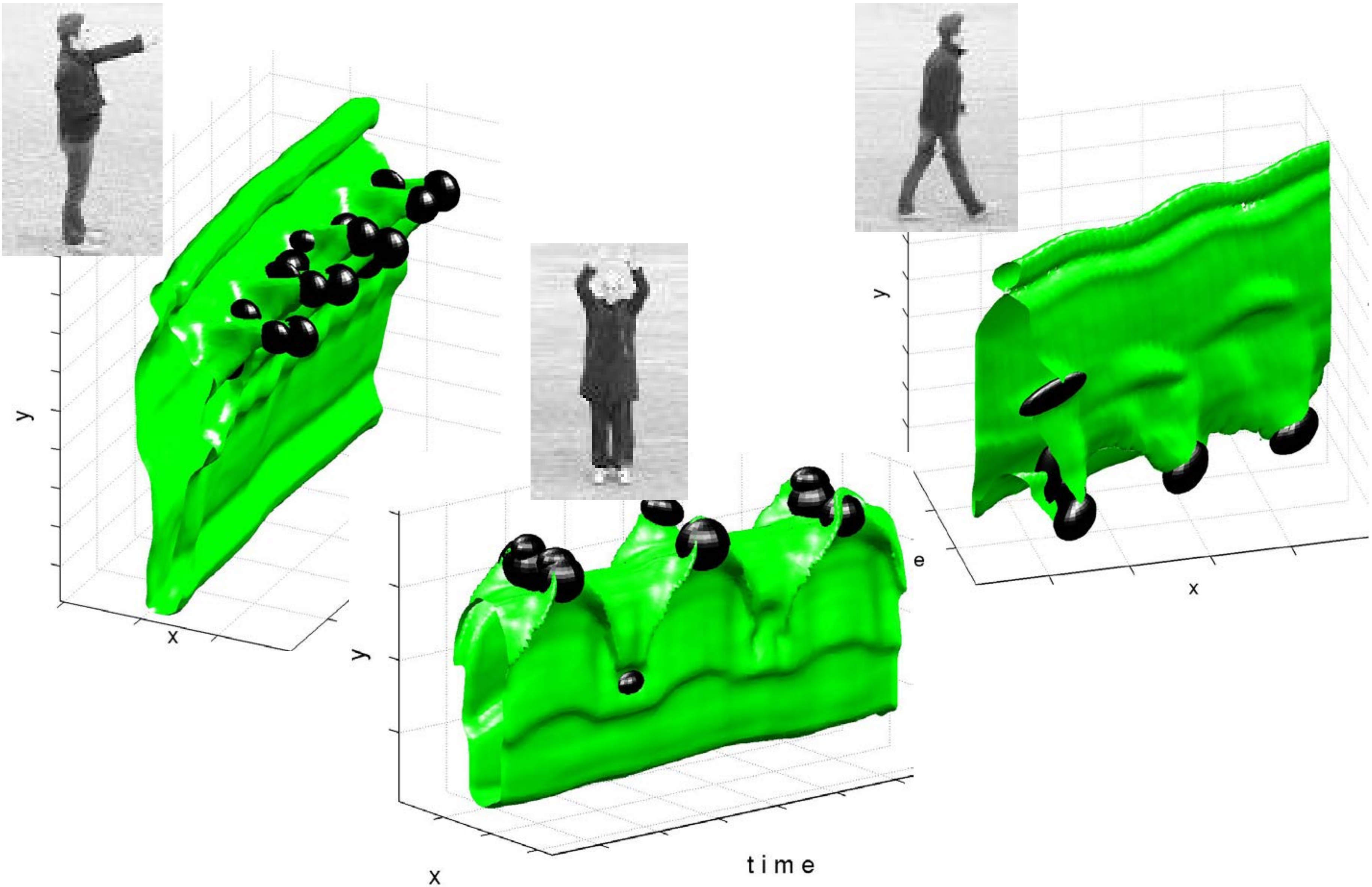


Spatio-temporal scale selection

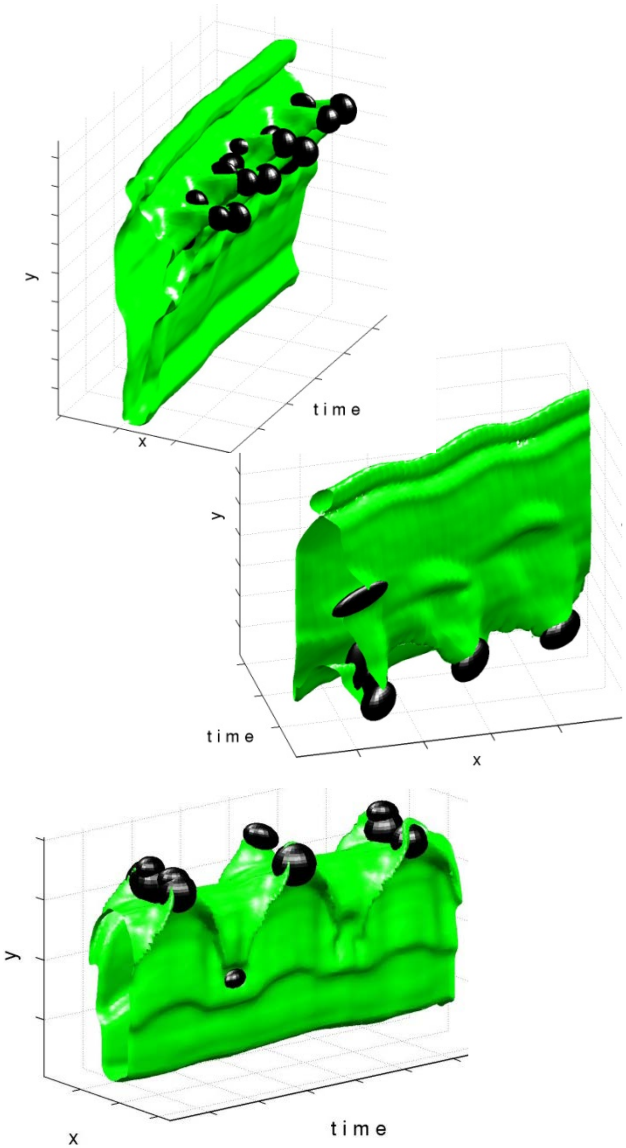


Selection of temporal scales captures the frequency of events

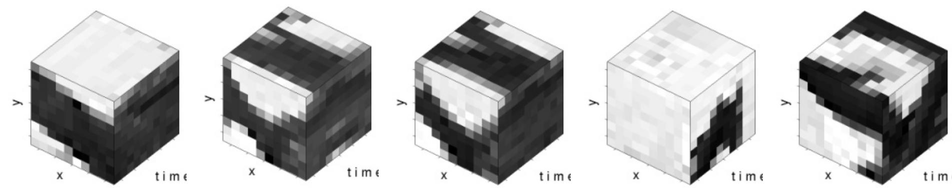
Local features for human actions



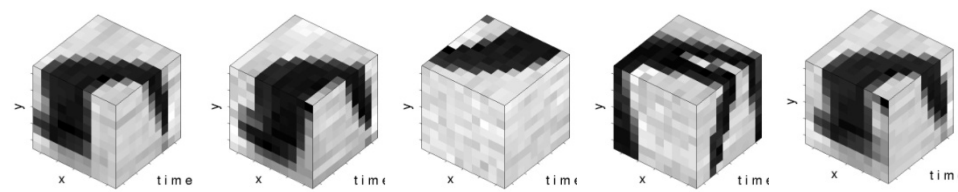
Local features for human actions



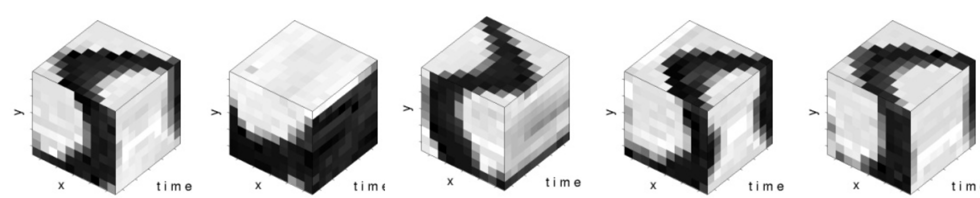
boxing



walking

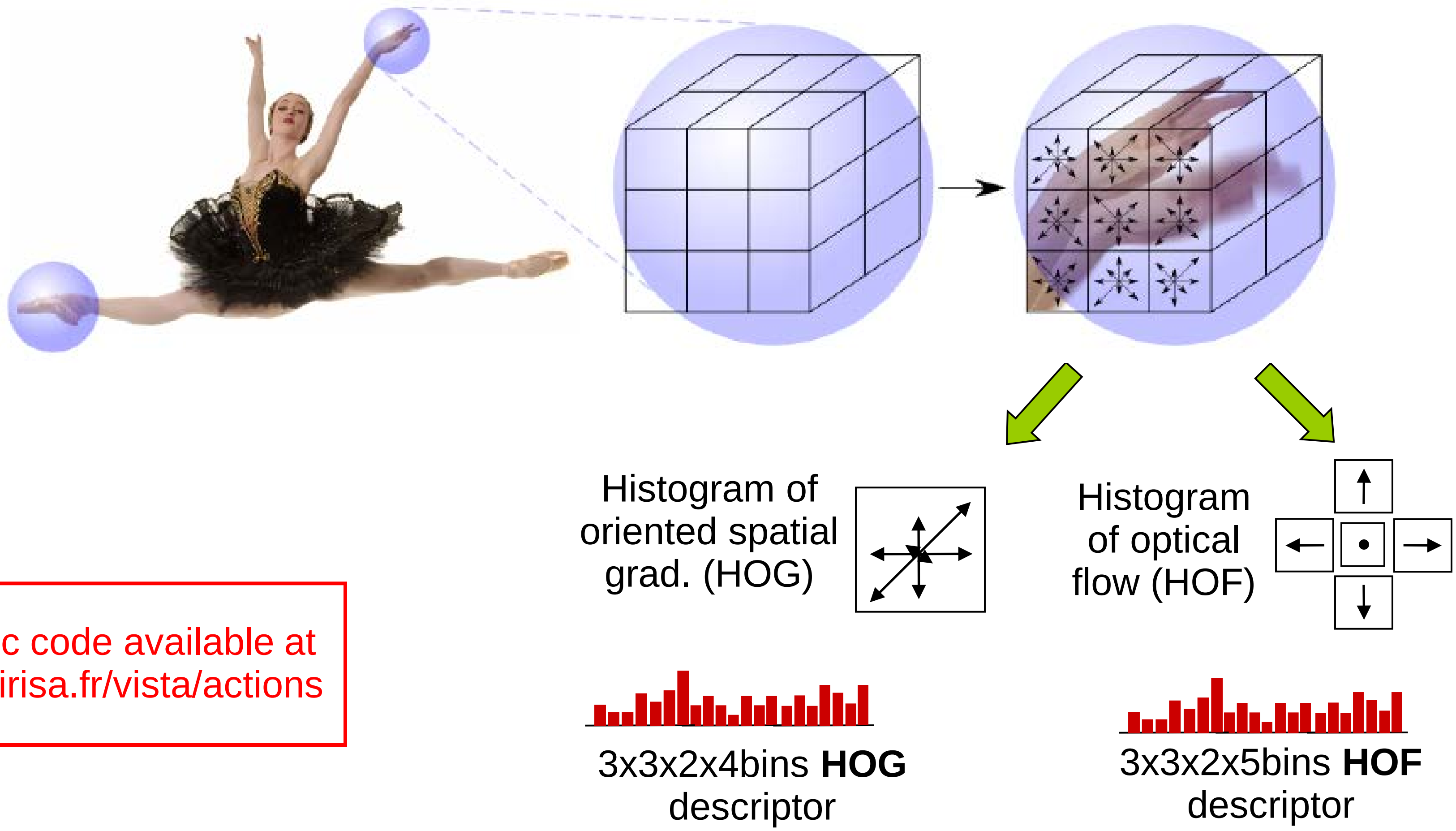


hand waving



Local space-time descriptor: HOG/HOF

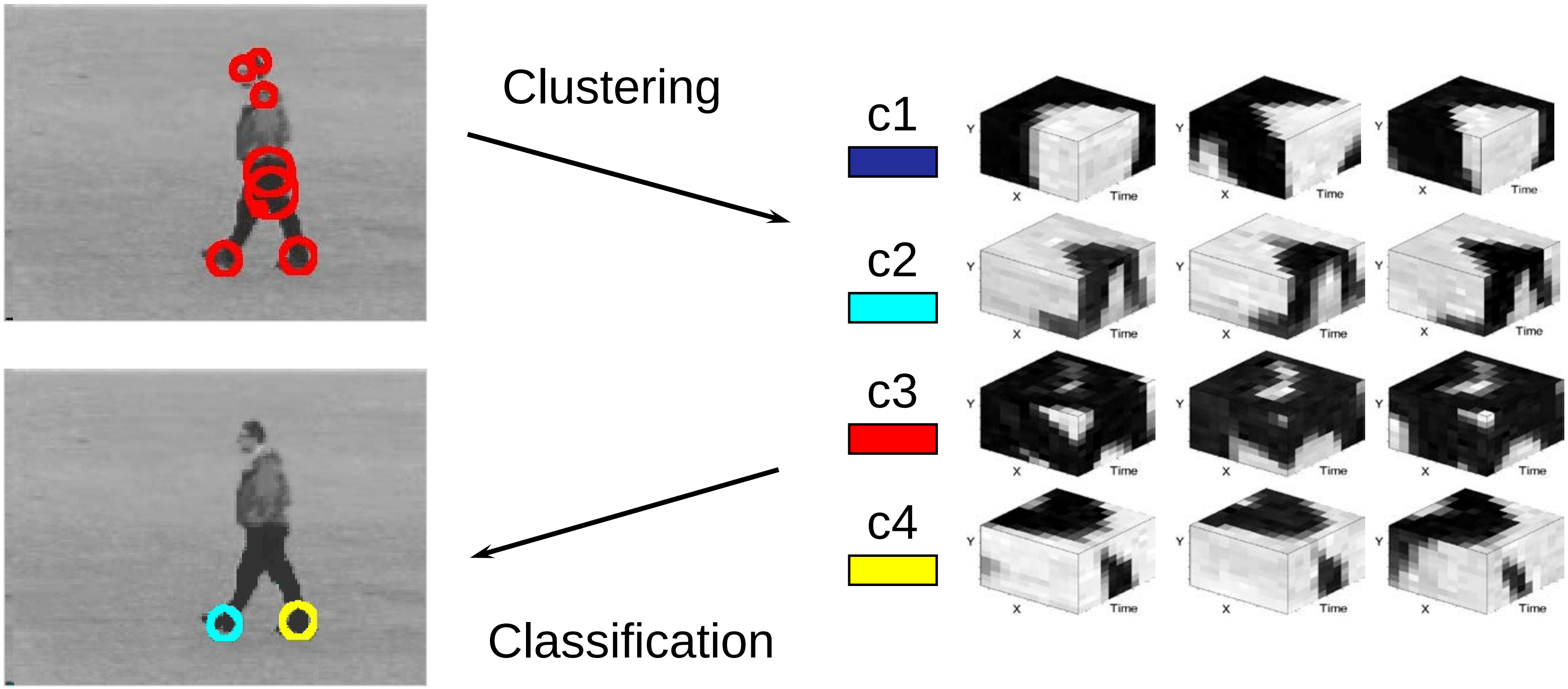
Multi-scale space-time patches



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

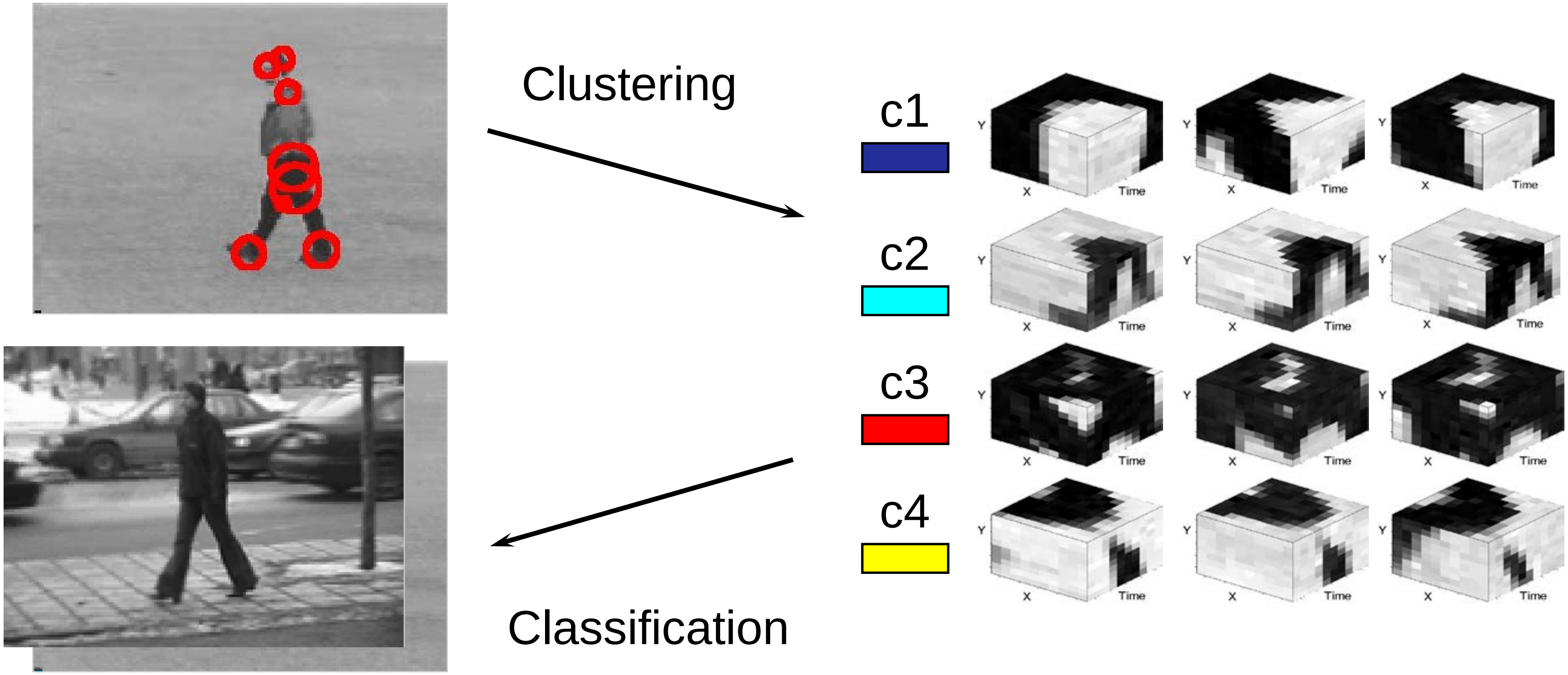
Visual Vocabulary: K-means clustering

- Group similar points in the space of image descriptors using K-means clustering
- Select significant clusters



Visual Vocabulary: K-means clustering

- Group similar points in the space of image descriptors using K-means clustering
- Select significant clusters



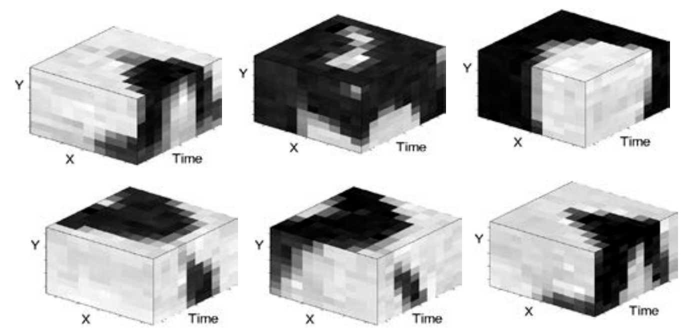
Local feature methods: Matching

- Finds similar events in pairs of video sequences

Bag-of-Features action recogntion

Extraction of Local features

space-time patches

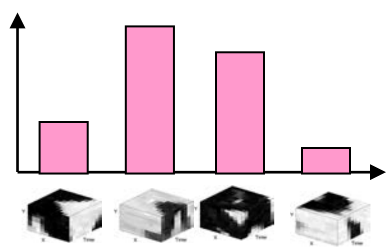


Feature description

K-means clustering (k=4000)

Feature quantization

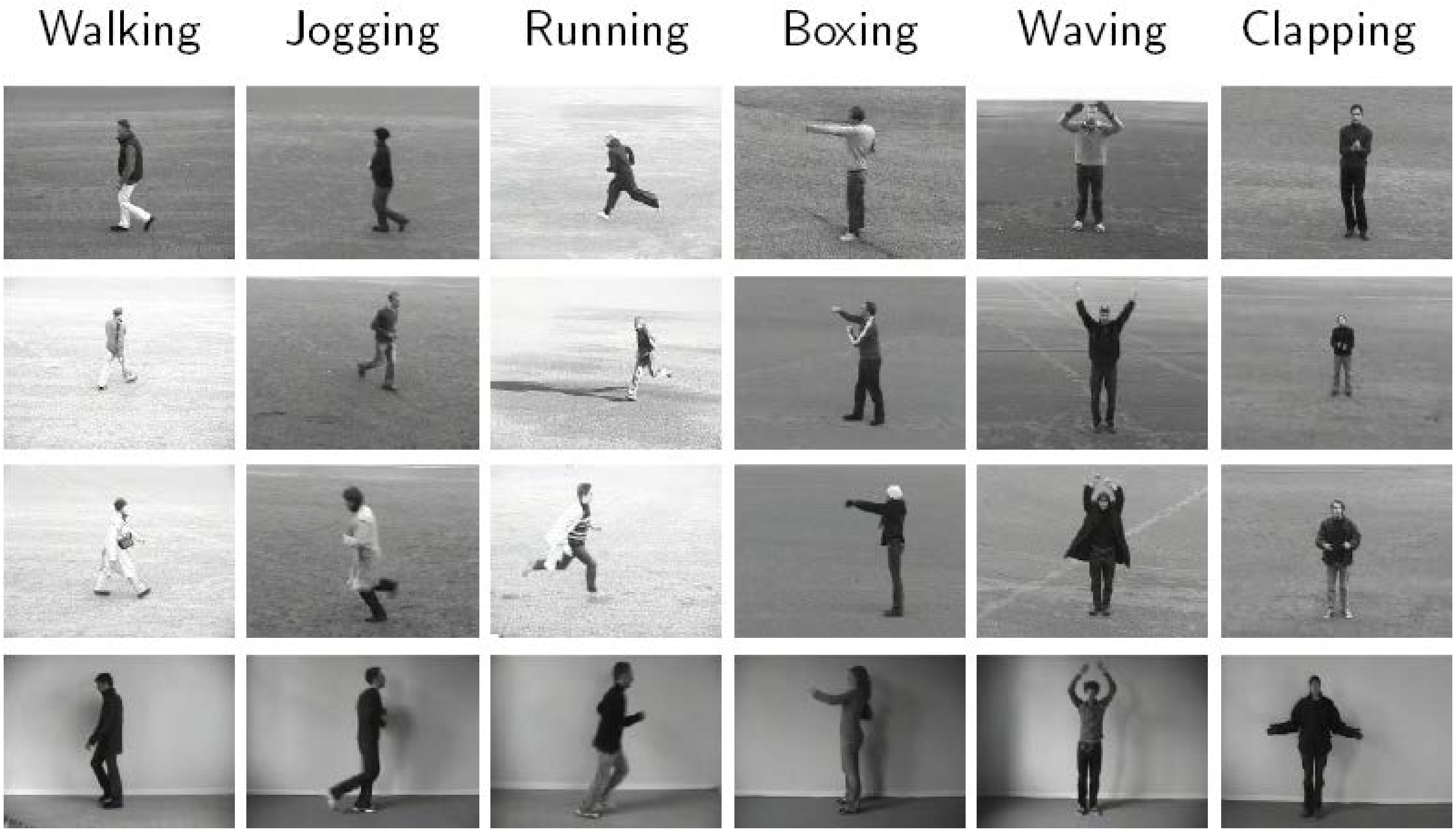
Occurrence histogram of visual words



Non-linear SVM with χ^2 kernel

[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Action recognition in KTH dataset



Sample frames from the KTH actions sequences, all six classes (columns) and scenarios (rows) are presented

Classification results on 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

Confusion matrix for KTH actions

Hollywood dataset

[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Action classification (CVPR08)

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

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

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:

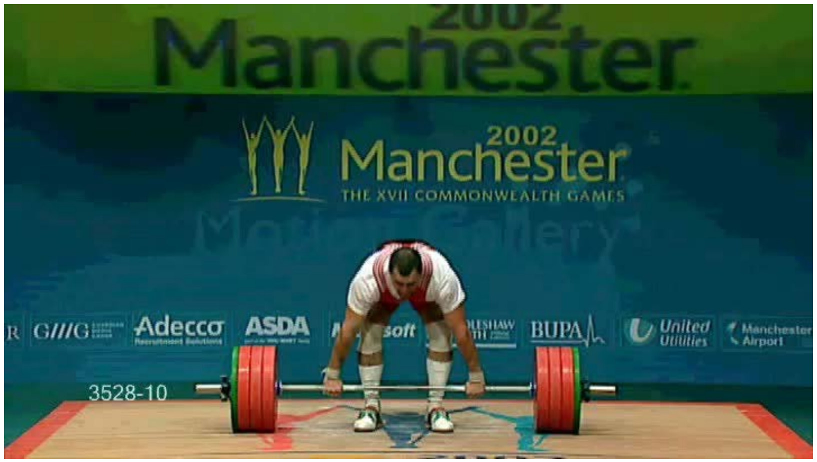
- HoG/HoF [Laptev et al. 2008]
- Cuboids [Dollar et al. 2005]
- HoG3D [Kläser et al. 2008]
- Extended SURF [Willems'et al. 2008]

Three human actions datasets:

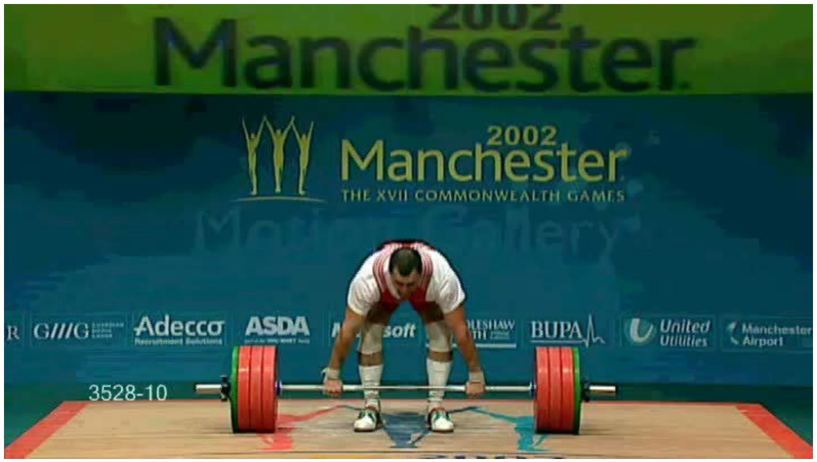
- KTH actions [Schuldt et al. 2004]
- UCF Sports [Rodriguez et al. 2008]
- Hollywood 2 [Marszałek et al. 2009]

Space-time feature detectors

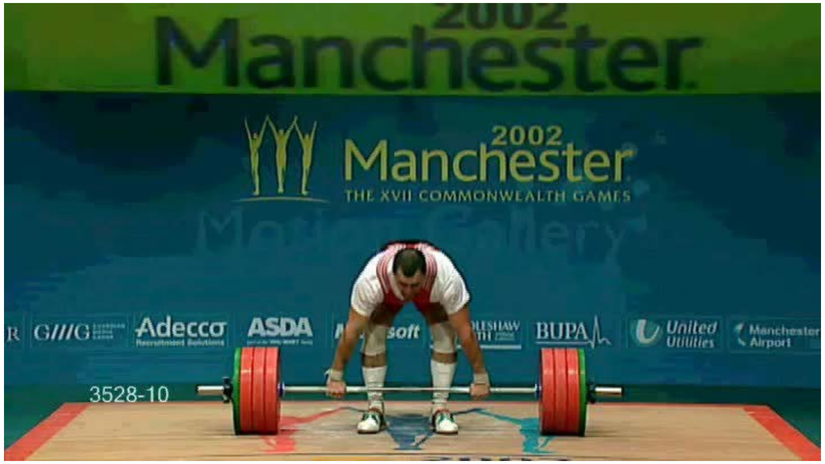
Harris3D



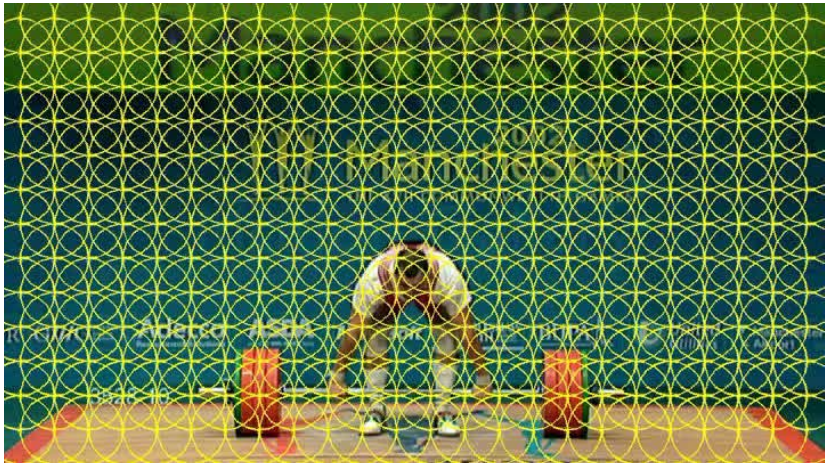
Hessian



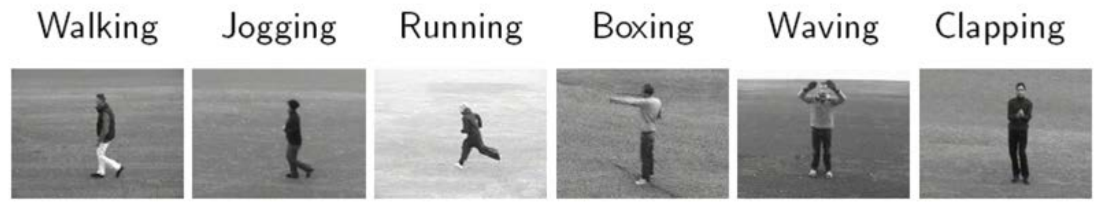
Cuboids



Dense



Results on KTH Actions



6 action classes, 4 scenarios, staged

Detectors

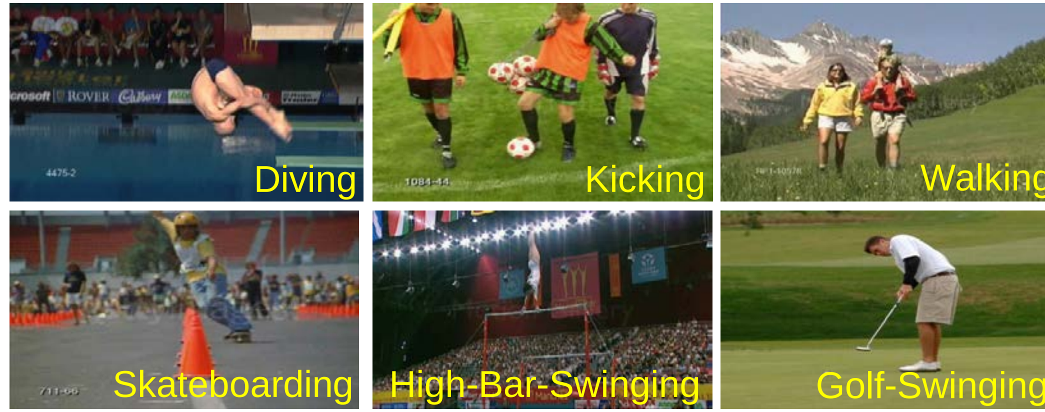
Descriptors	Harris3D	Cuboids	Hessian	Dense
HOG3D	89.0%	90.0%	84.6%	85.3%
HOG/HOF	91.8%	88.7%	88.7%	86.1%
HOG	80.9%	82.3%	77.7%	79.0%
HOF	**92.1%**	88.2%	88.6%	88.0%
Cuboids	-	89.1%	-	-
E-SURF	-	-	81.4%	-

(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%	-

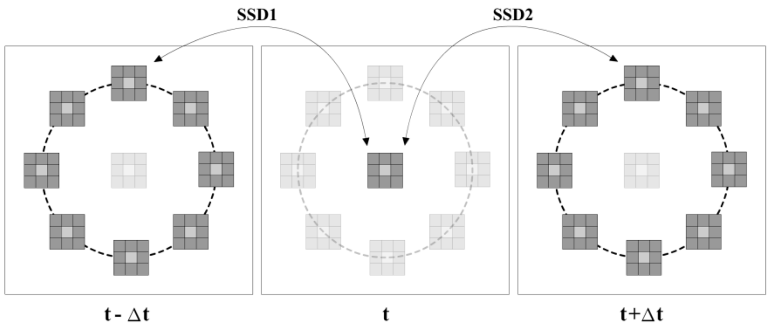
(Average precision scores)

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

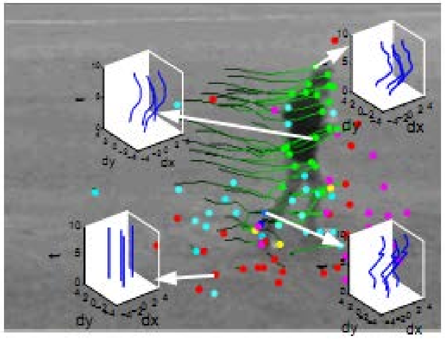
[Wang, Ullah, Kläser, Laptev, Schmid, 2009]

More recent local methods I

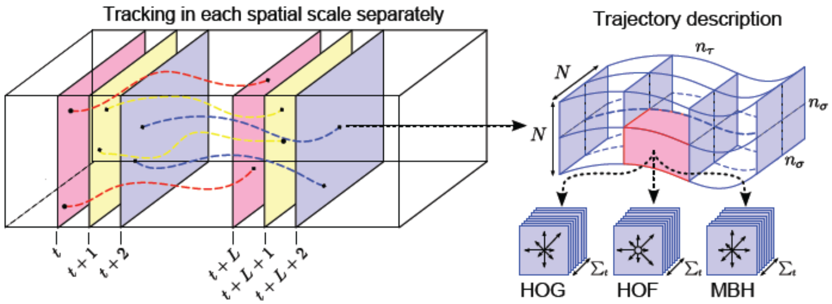
- Y. and L. Wolf, "Local Trinary Patterns for Human Action Recognition ",
  ICCV 2009
  + ECCV 2012 extension



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

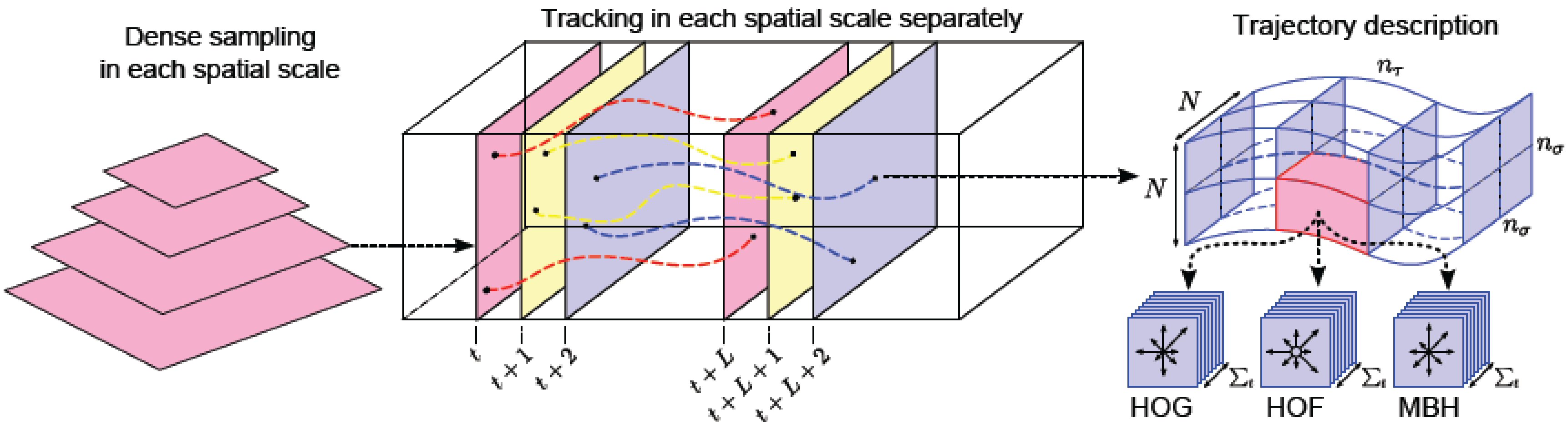


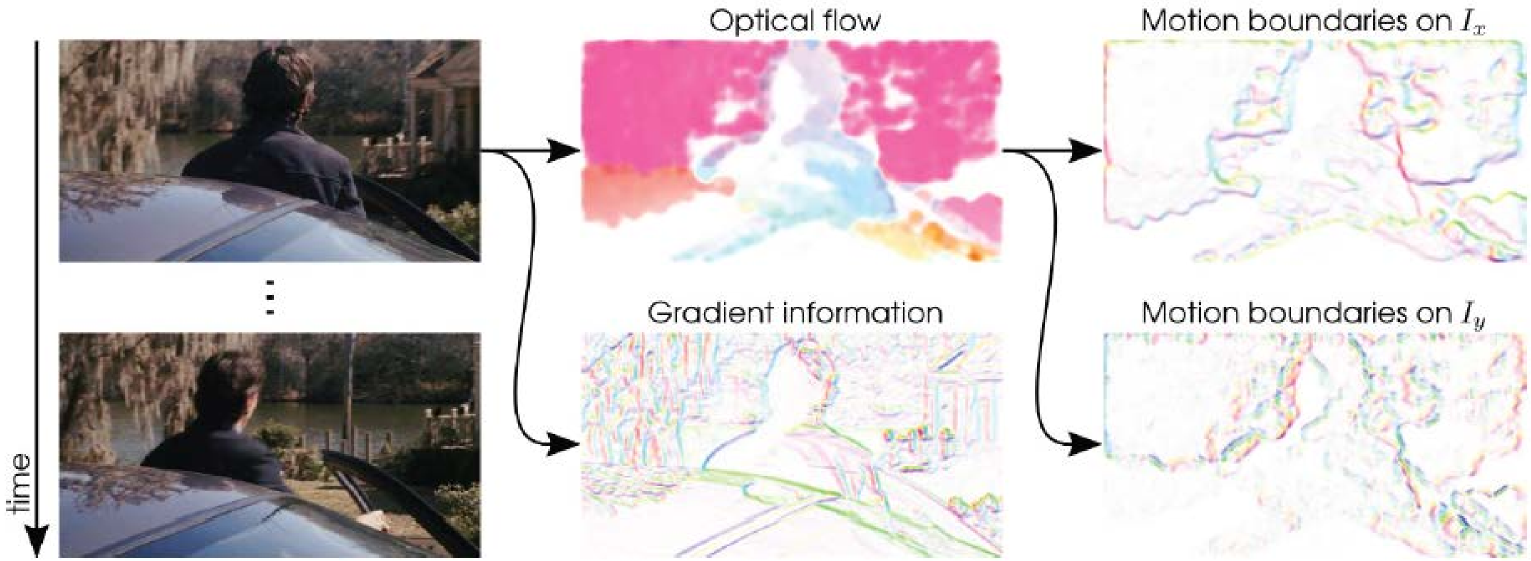
- H. Wang, A. Klaser, C. Schmid, C.-L. Liu, "Action Recognition by Dense Trajectories", CVPR 2011



Dense trajectory descriptors

[Wang et al. CVPR'11]





Dense trajectory descriptors

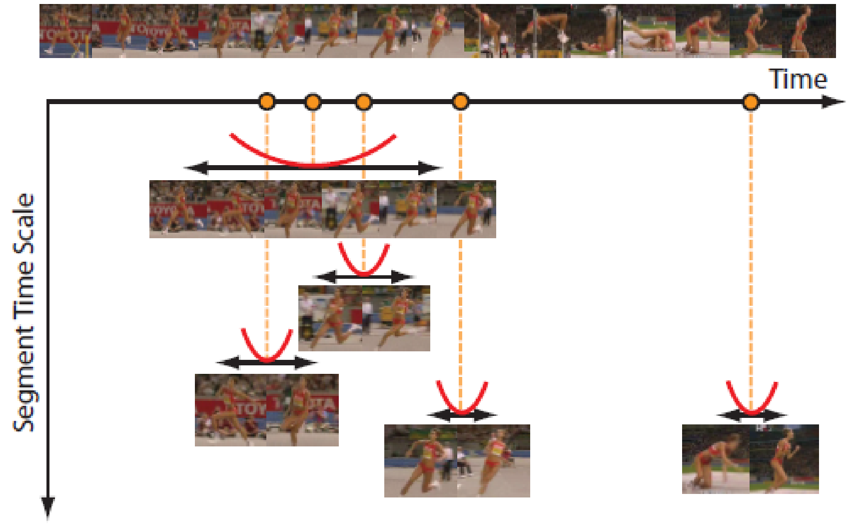
[Wang et al. CVPR'11]

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%**

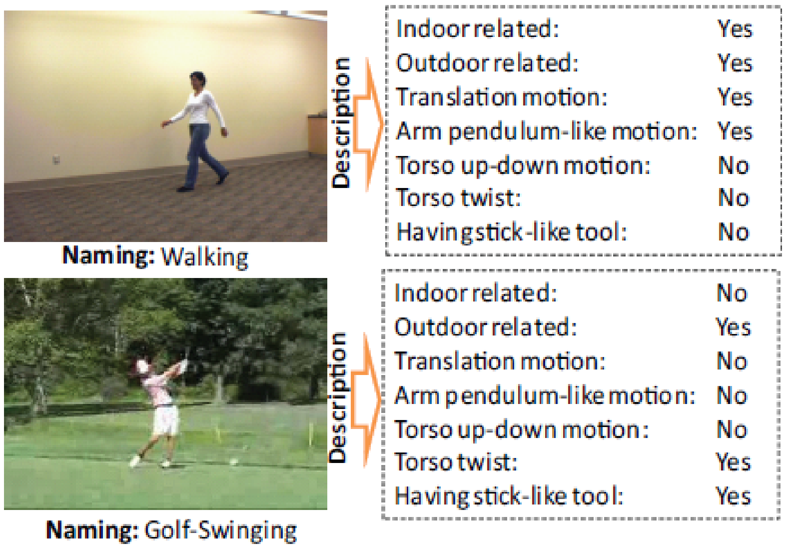
IXMAS		UIUC		Olympic Sports		UCF50	
Tran *et al.*[50]	80.22%	Tran *et al.* [50]	98.7%	Brendel *et al.* [56]	77.3%		
Junejo *et al.* [63]	79.6%			Niebles *et al.* [49]	72.1%		
Wu *et al.*[54]	88.2%						
MBH	91.8%	MBH	97.1%	MBH	71.6%	MBH	82.2%
Combined	**93.5%**	Combined	**98.4%**	Combined	74.1%	Combined	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%**

More recent local methods II

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



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



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

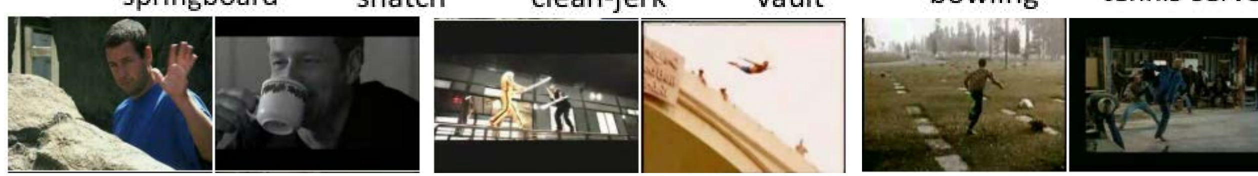
- UCF Sports, 10 classes, 150 samples, [Rodriguez et al. 2008]

Diving Kicking Walking Skateboarding High-Bar-Swinging

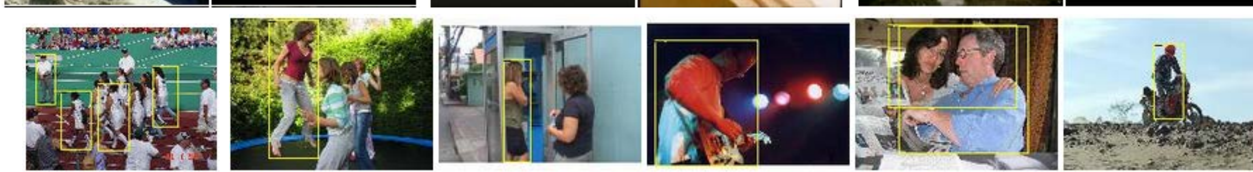
- Olympic Sports, 16 classes, 783 samples, [Niebles et al. 2010]

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



How to collect training data?

Learning Actions from Movies

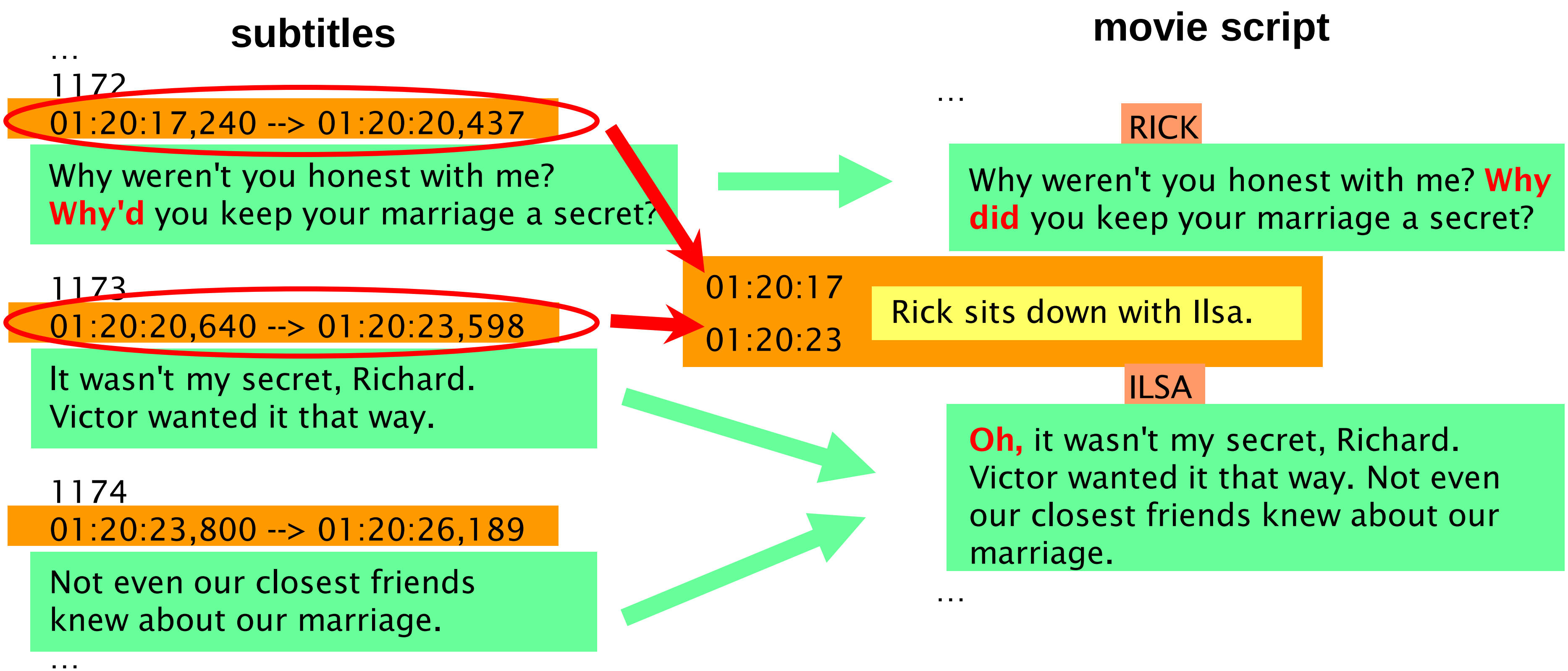
- Realistic variation of human actions
- Many classes and many examples per class

Problems:

- Typically only a few class-samples per movie
- Manual annotation is very time consuming

Automatic video annotation with scripts

- 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



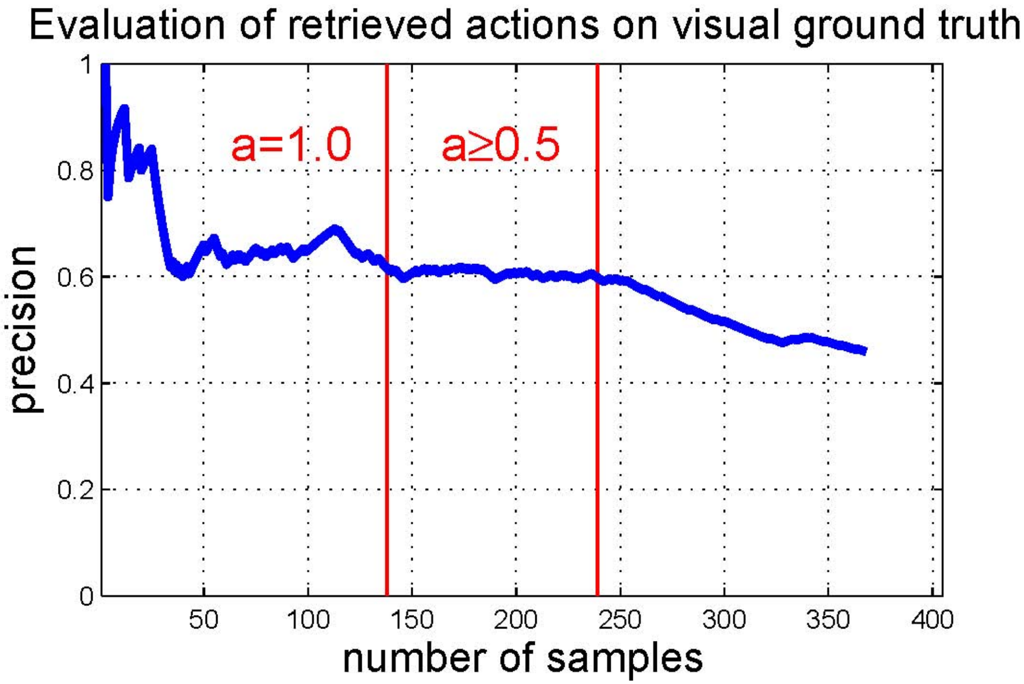
Script-based action annotation

- **On the good side:**
  - Realistic variation of actions: subjects, views, etc…
  - Many examples per class, many classes
  - No extra overhead for new classes
  - Actions, objects, scenes and their combinations
  - Character names may be used to resolve “who is doing what?”

- **Problems:**
  - No spatial localization
  - Temporal localization may be poor
  - Missing actions: e.g. scripts do not always follow the movie
  - Annotation is incomplete, not suitable as ground truth for testing action detection
  - Large within-class variability of action classes *in text*

Script alignment: Evaluation

- Annotate action samples *in text*
- Do automatic script-to-video alignment
- Check the correspondence of actions in scripts and movies



a: quality of subtitle-script matching

Example of a "visual false positive"

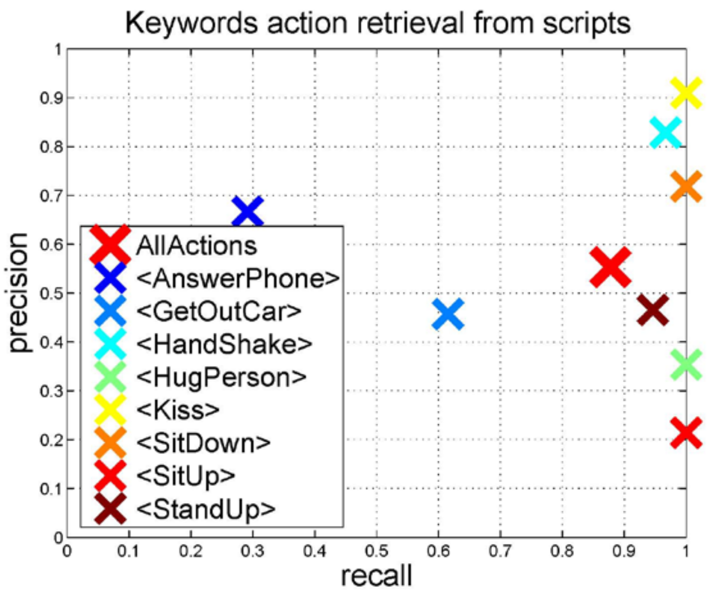
A black car pulls up, two army officers get out.

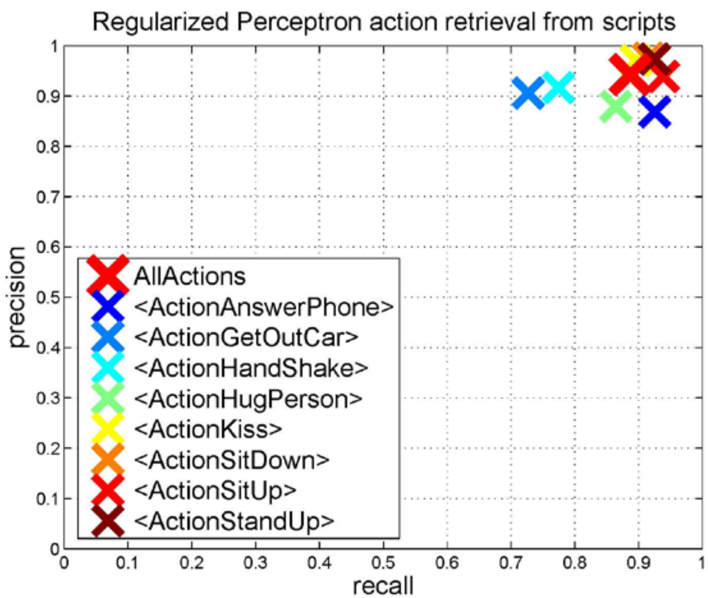
Text-based action retrieval

- 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





Automatically annotated action samples

[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]

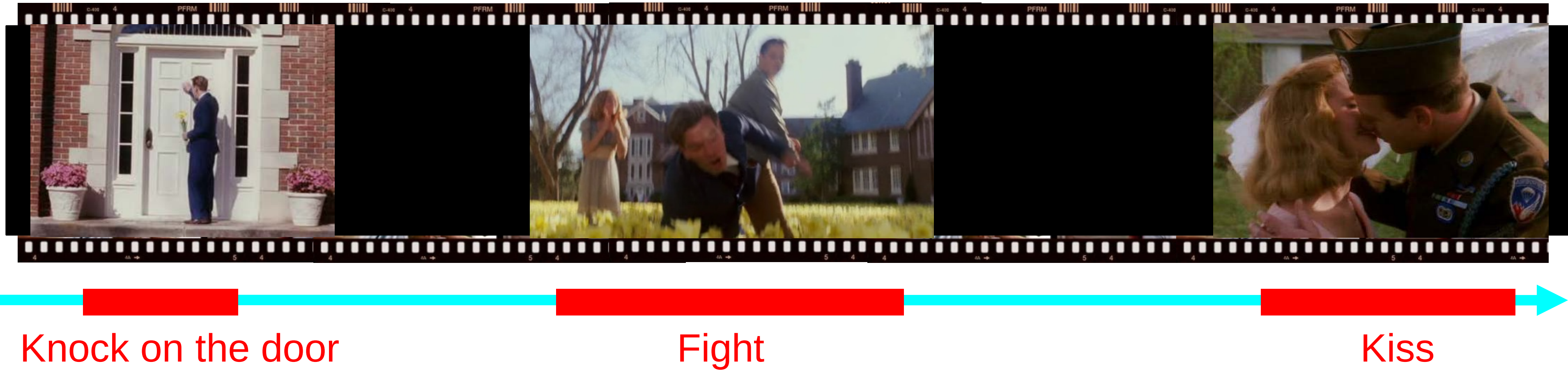
Action classification results

	Clean hoghof		Automatic 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

Weakly-Supervised
Temporal Action Annotation

- Answer questions: *WHAT actions and WHEN they happened* ?

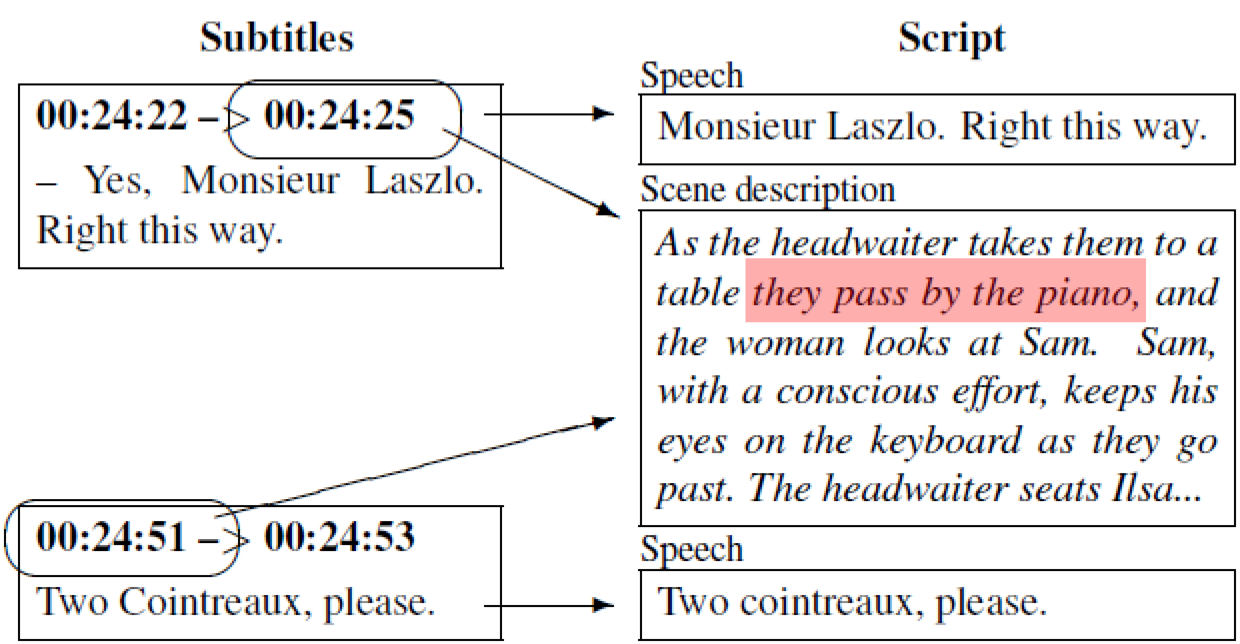


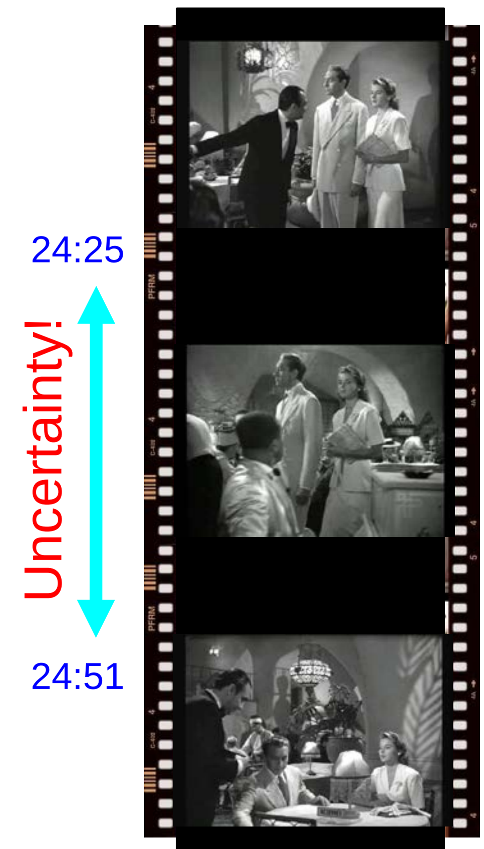
- Train visual action detectors and annotate actions with the minimal manual supervision

WHEN: Video Data and Annotation

- Want to target realistic video data
- Want to avoid manual video annotation for training

➡ Use movies + scripts for automatic annotation of training samples





Overview

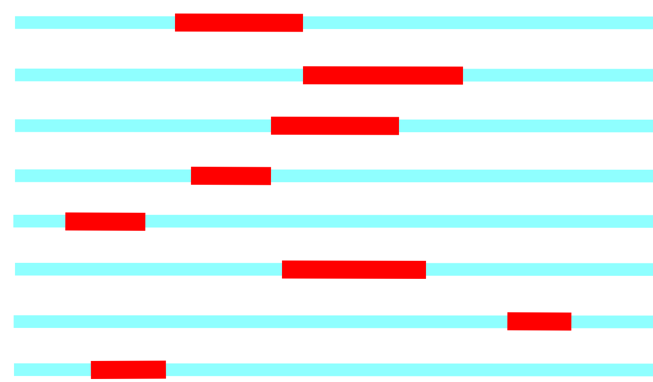
Input:

- Action type, e.g.
  Person Opens Door
- Videos + aligned scripts

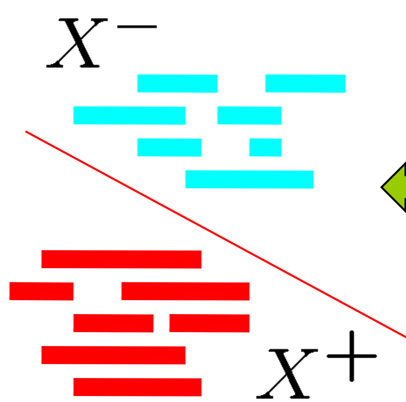
Automatic collection of training clips

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

Clustering of positive segments



Training classifier



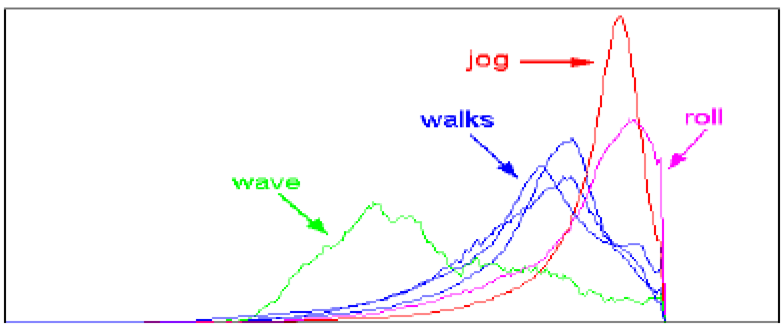
Output:

Sliding-window-style temporal action localization

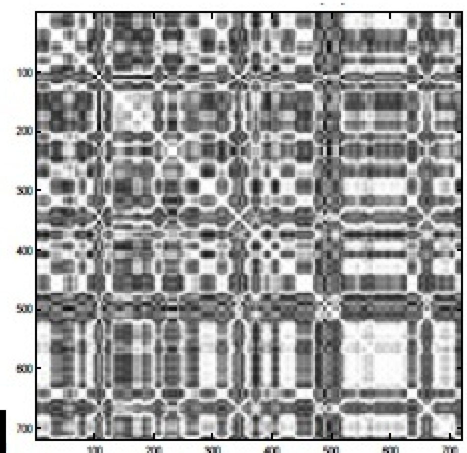
Action clustering

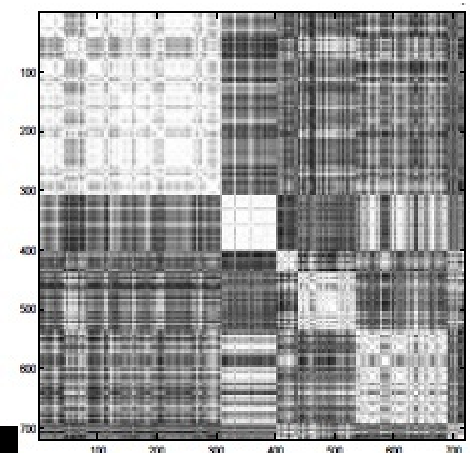
[Lihi Zelnik-Manor and Michal Irani CVPR 2001]

Descriptor space

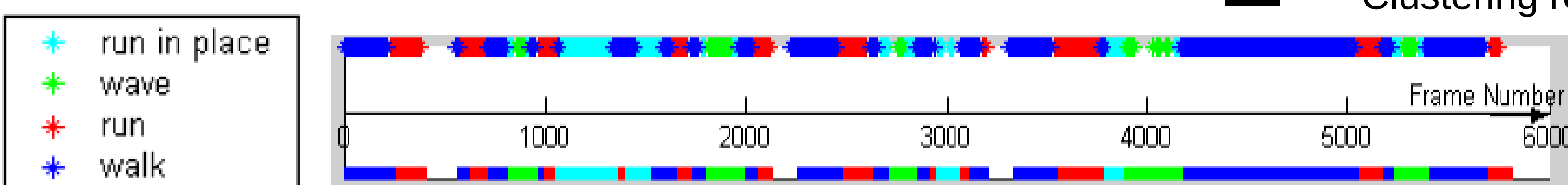


Spectral clustering





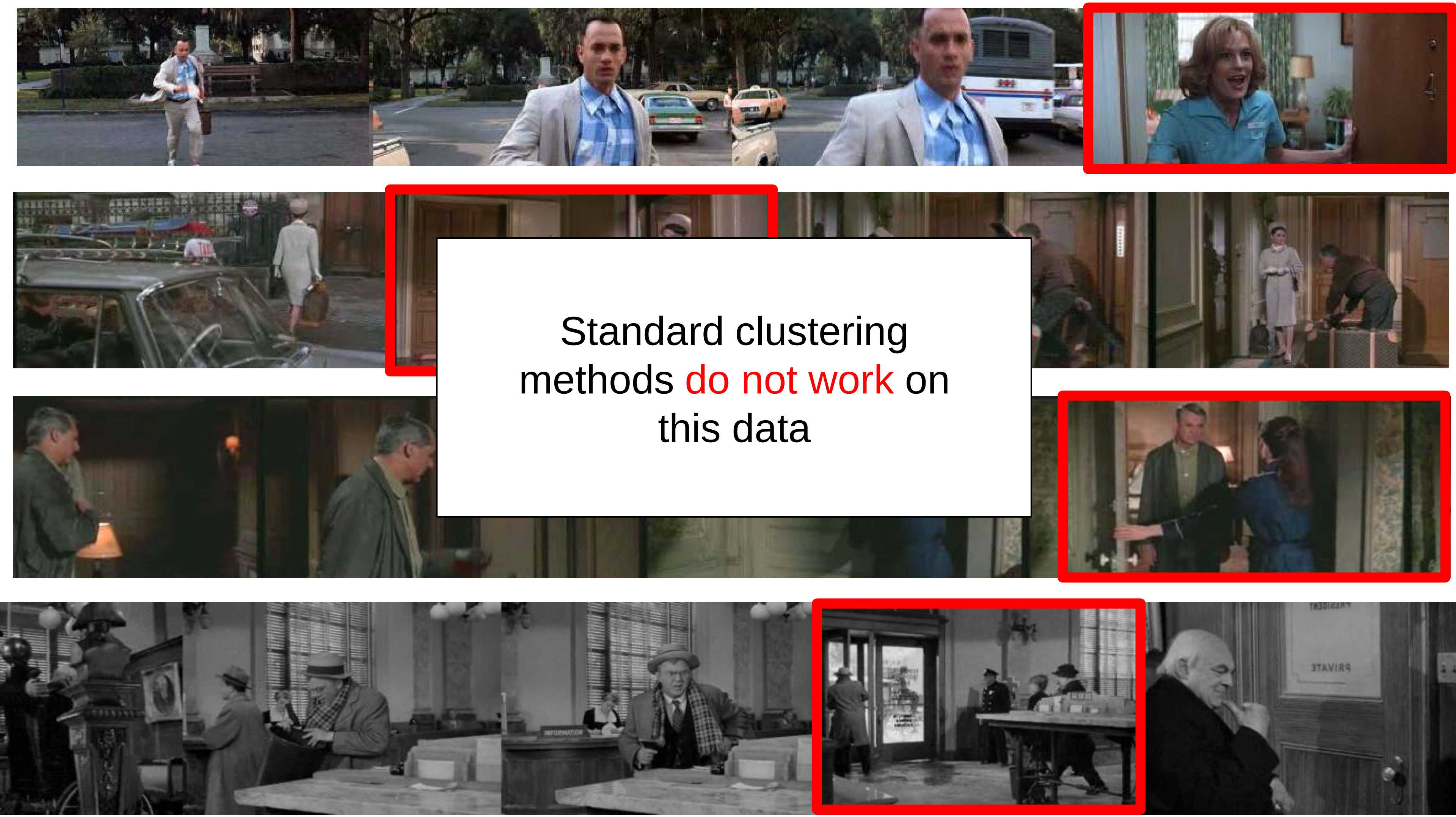
Clustering results



Ground truth

Action clustering

Complex data:

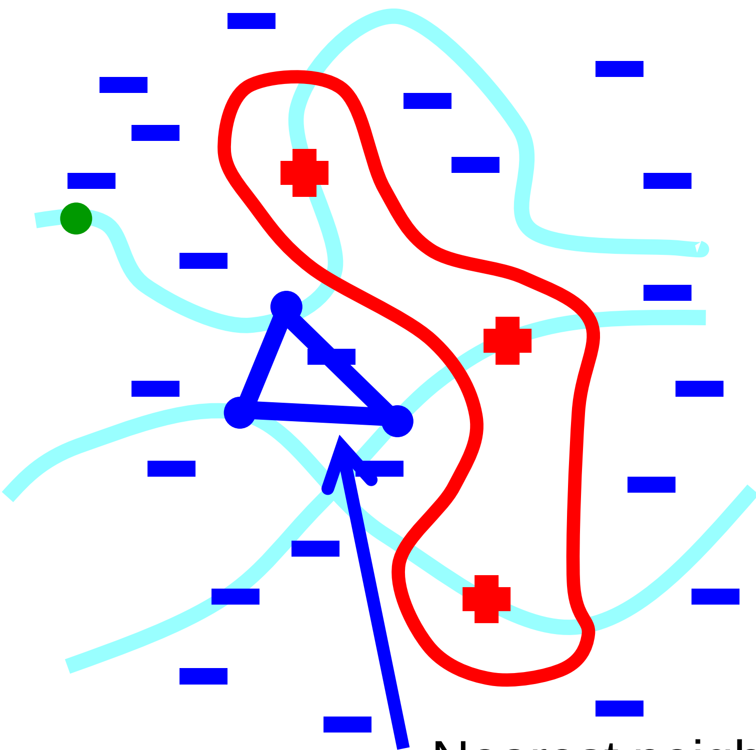


Standard clustering methods do not work on this data

Action clustering

Our view at the problem

Feature space



Nearest neighbor solution: wrong!

Video space

Negative samples!

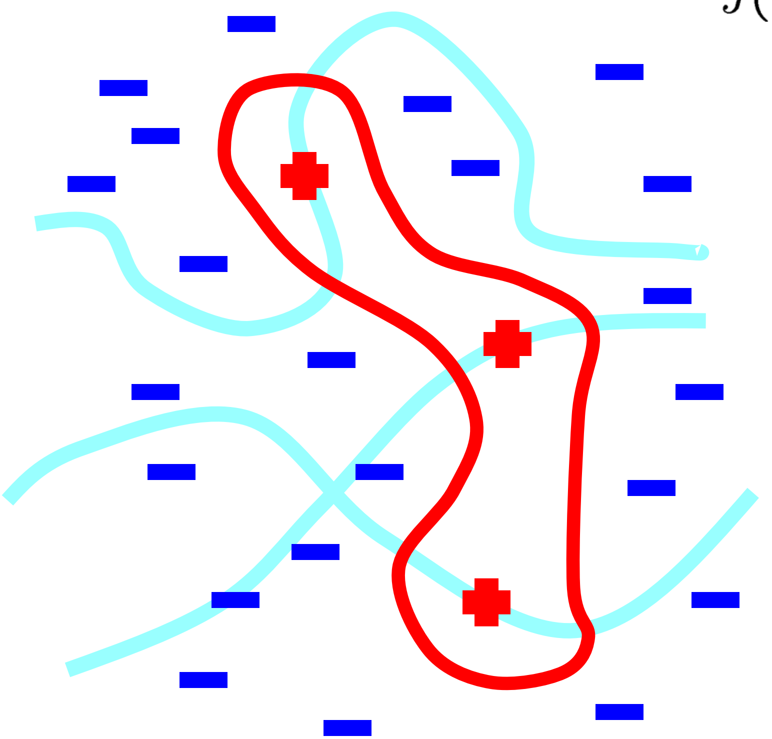
Random video samples: lots of them, very low chance to be positives

Action clustering

Formulation

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

Feature space



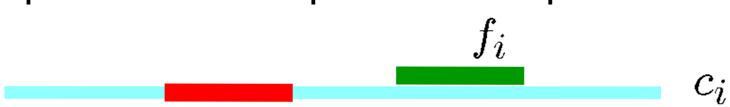
discriminative cost

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

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

x_i^- negative samples

$c_i[f_i]$ parameterized positive samples



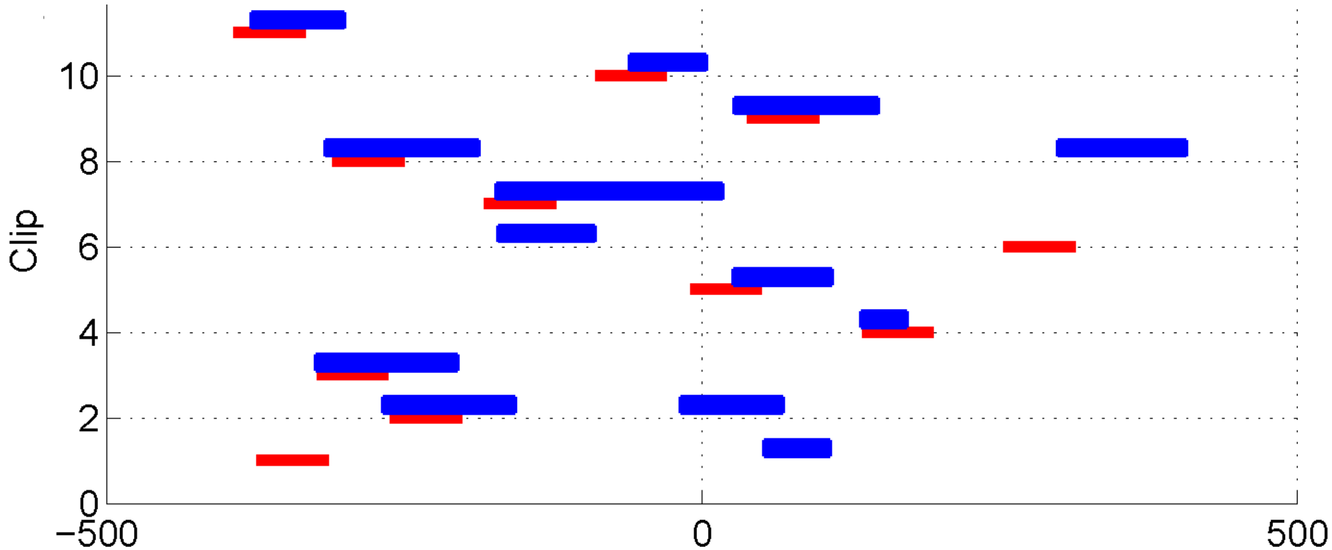
Optimization

SVM solution for w, b

Coordinate descent on f_i

Clustering results

Drinking actions in Coffee and Cigarettes

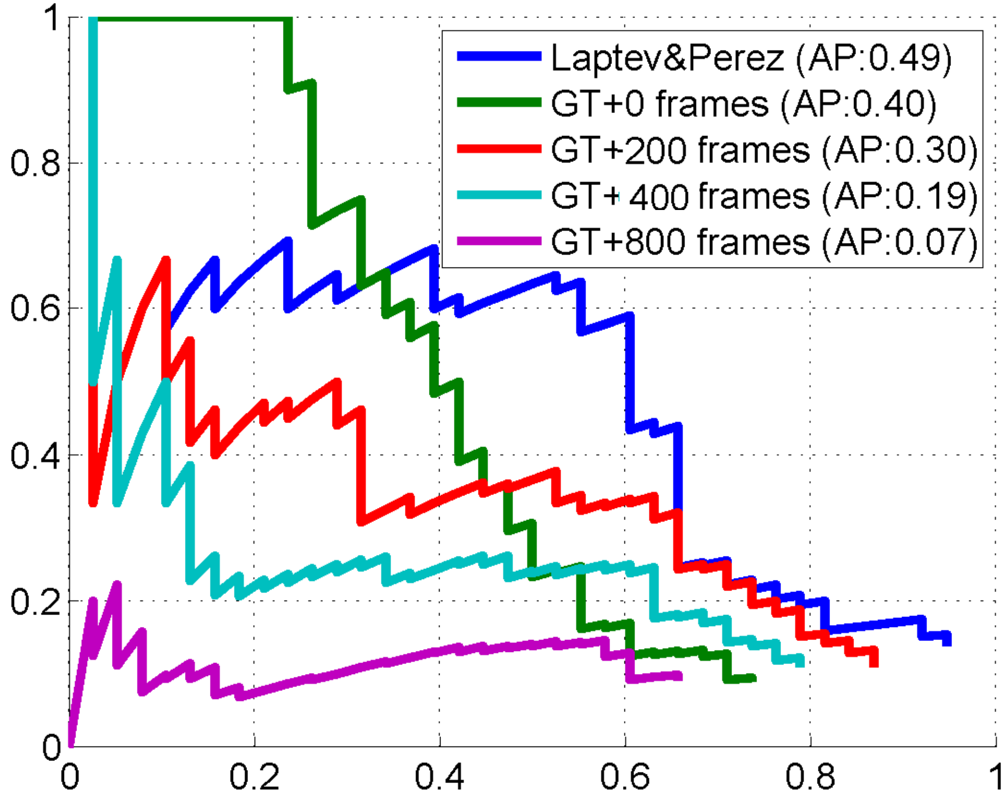


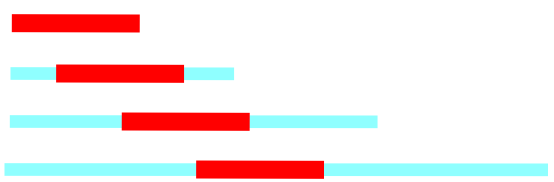
Detection results

Drinking actions in Coffee and Cigarettes

- Training Bag-of-Features classifier
- Temporal sliding window classification
- Non-maximum suppression

Detection trained on *simulated clusters*





Test set:

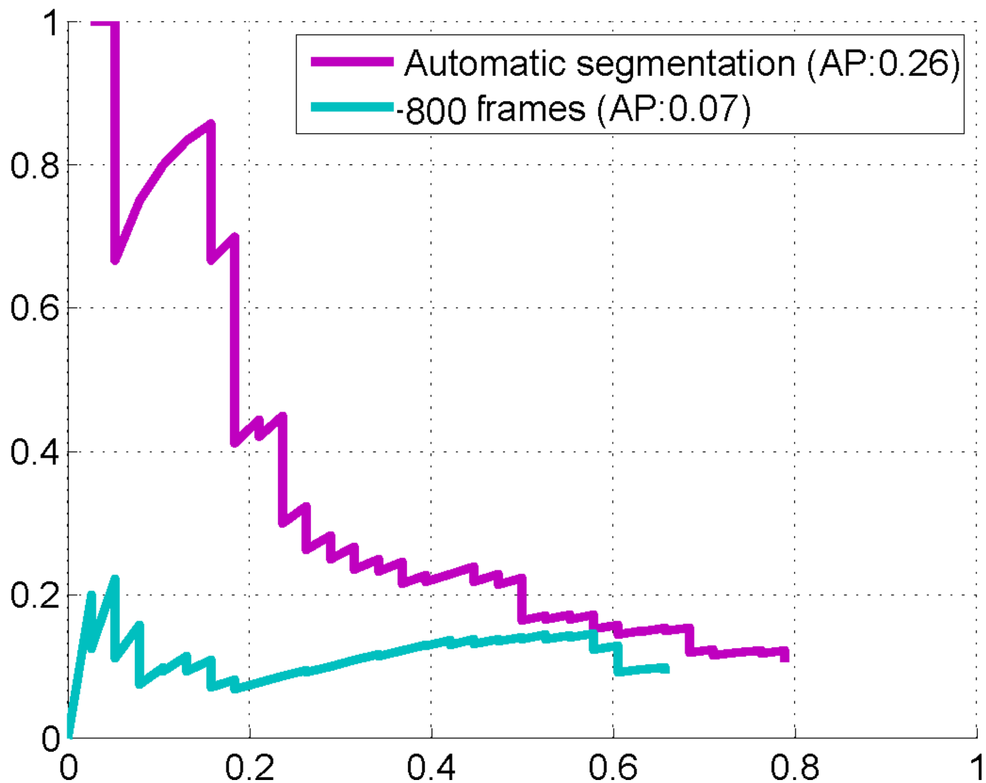
- 25min from "Coffee and Cigarettes" with GT 38 drinking actions

Detection results

Drinking actions in Coffee and Cigarettes

- Training Bag-of-Features classifier
- Temporal sliding window classification
- Non-maximum suppression

Detection trained on *automatic clusters*

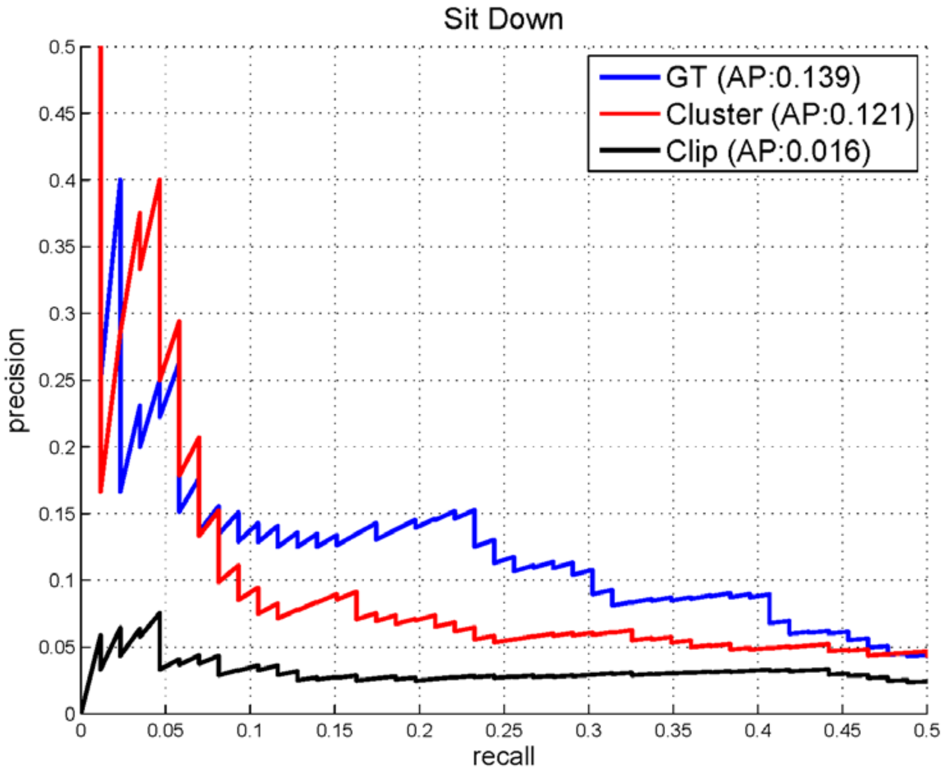


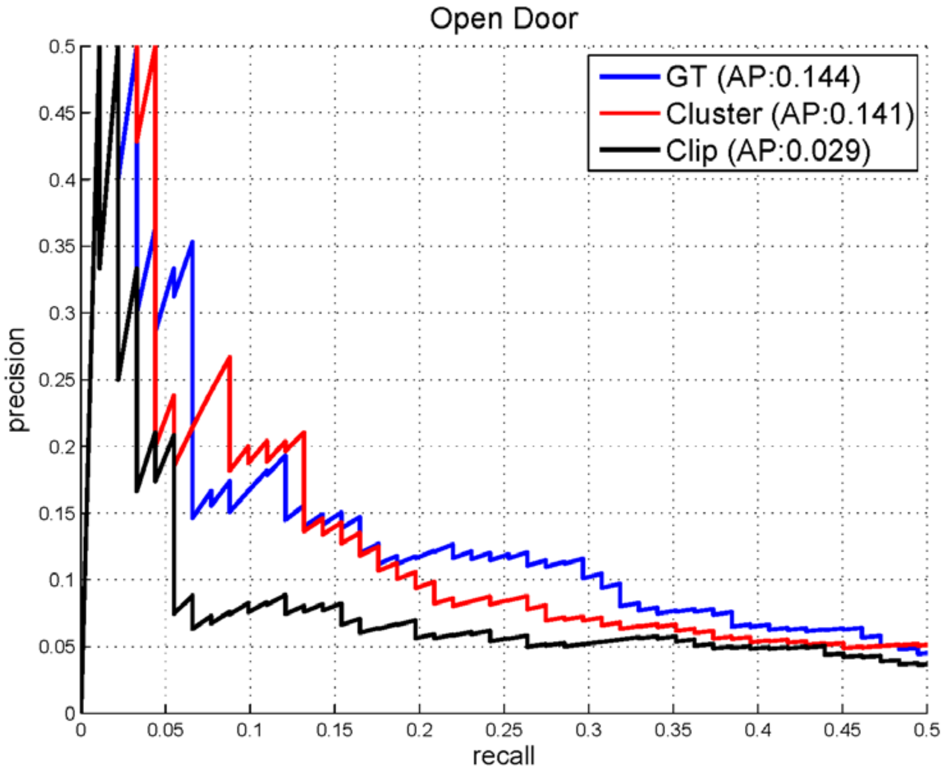
Test set:

- 25min from “Coffee and Cigarettes” with GT 38 drinking actions

Detection results

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





Automatic Annotation of Human Actions in Video

ICCV 2009 DEMO

O.Duchenne, I.Laptev, J.Sivic, F.Bach and J.Ponce

Temporal detection of actions OpenDoor and SitDown in episodes of
The Graduate, The Crying Game, Living in Oblivion

Temporal detection of “Sit Down” and “Open Door” actions in movies:
The Graduate, The Crying Game, Living in Oblivion

Mining scene captions

ILSA

I wish I didn't love you so much.

01:22:00
01:22:03

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.

CARL

I think we lost them.

01:22:15
01:22:17

…

Actions in Context

- Human actions are frequently correlated with particular scene classes

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

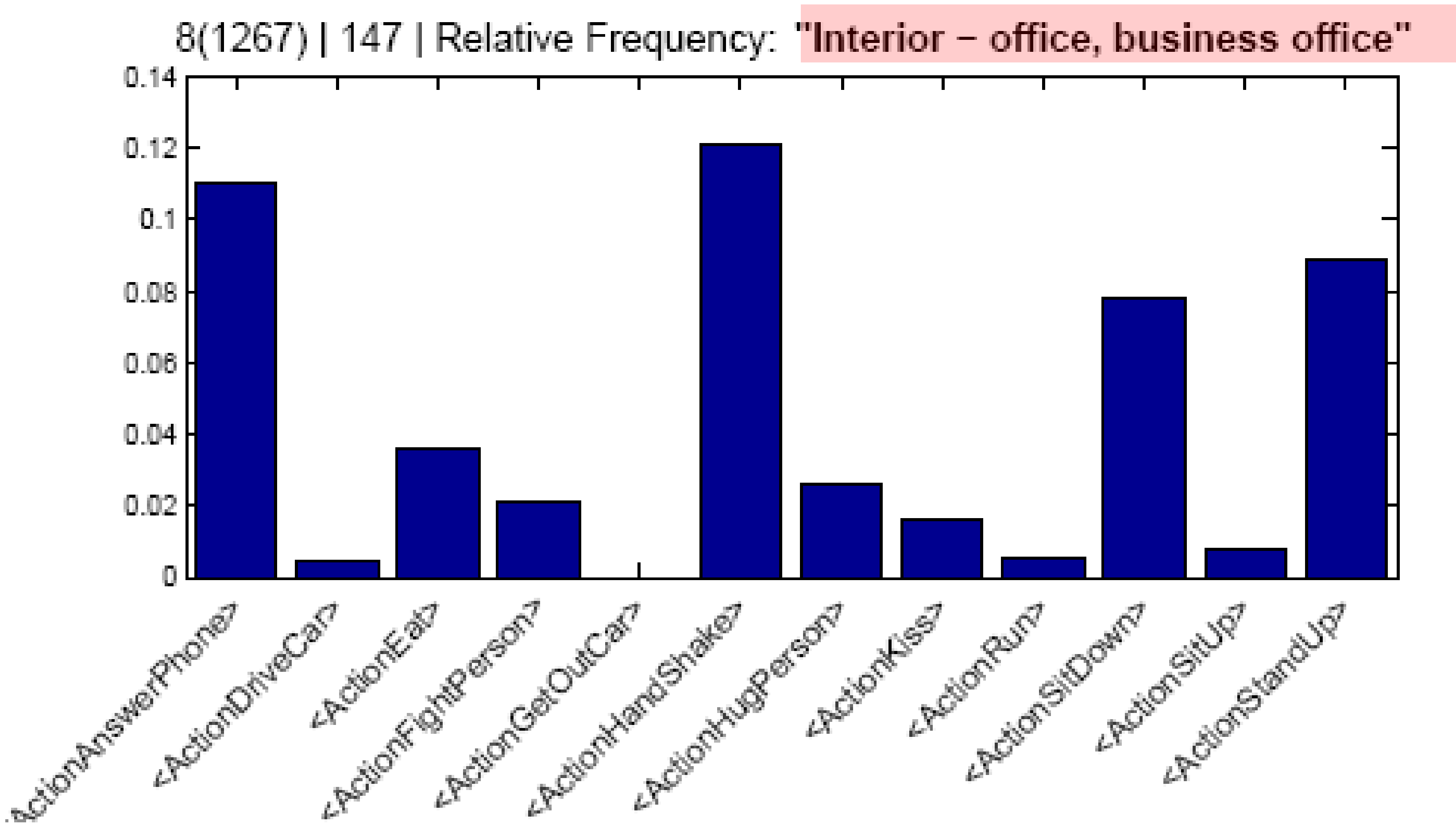
Eating -- *kitchen*

Eating -- *cafe*

Running -- *road*

Running -- *street*

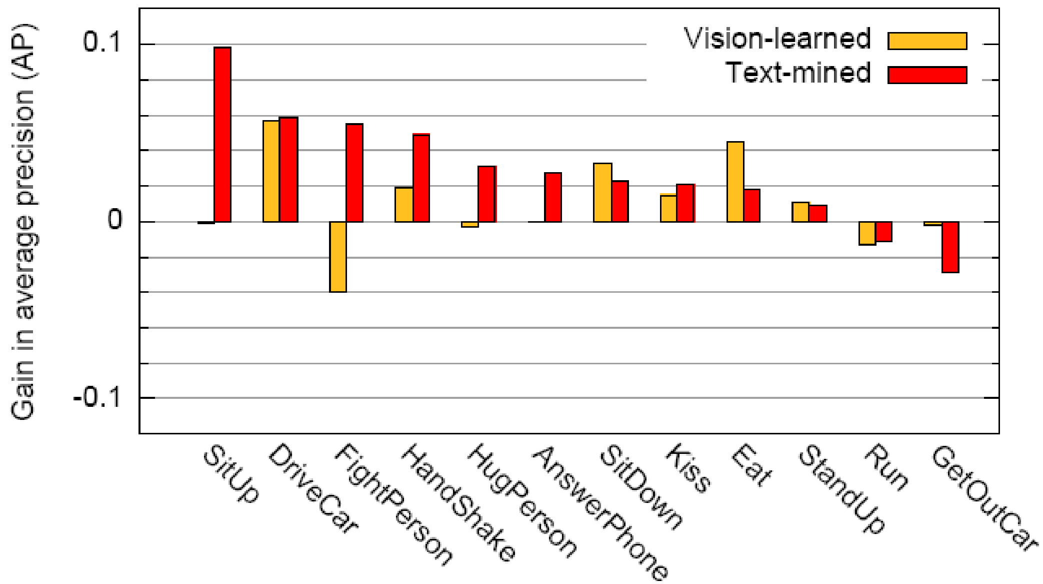
Co-occurrence of actions and scenes in scripts



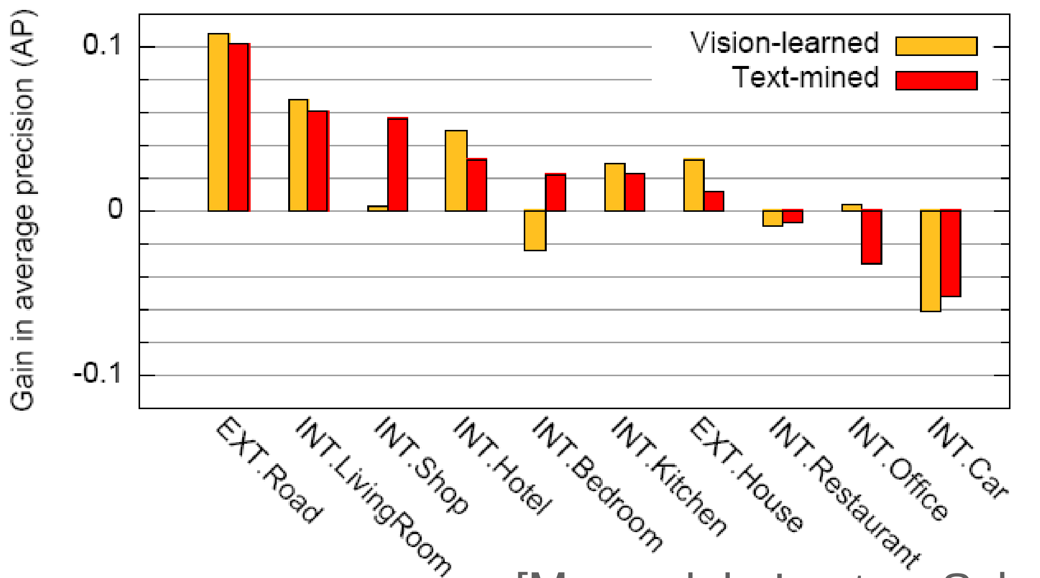
[Marszałek, Laptev, Schmid, 2009]

Results: Joint action and scene recognition

Actions in the context of Scenes



Scenes in the context of Actions



[Marszałek, Laptev, Schmid, 2009]

Where to go next?

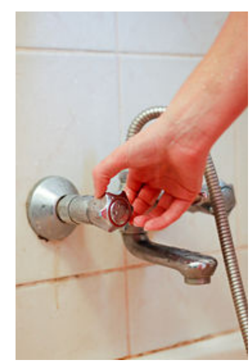
Is action classification the right problem?

- Is action vocabulary well-defined?

Examples of “Open” action:





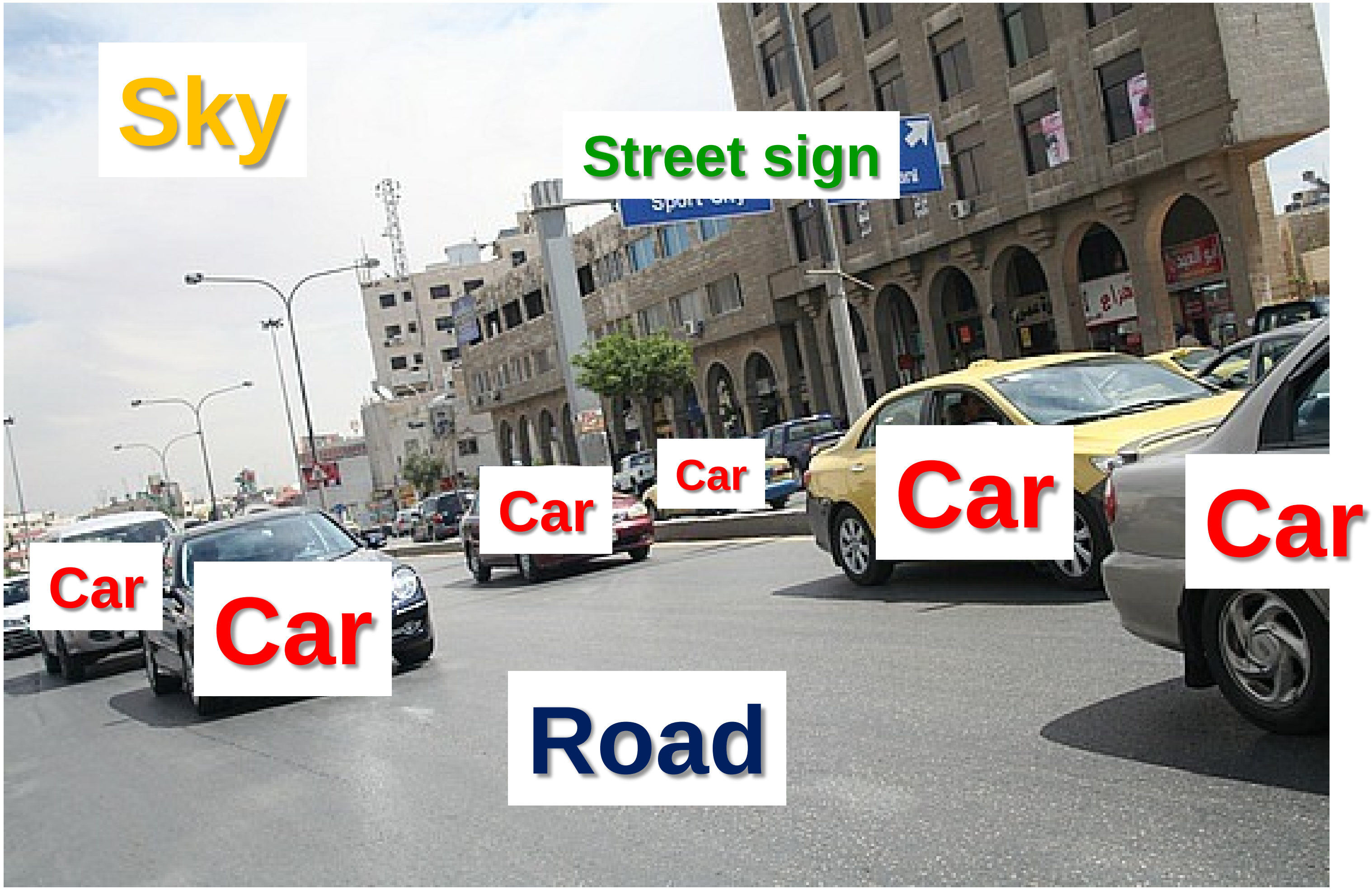



- What granularity of action vocabulary shall we consider?

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

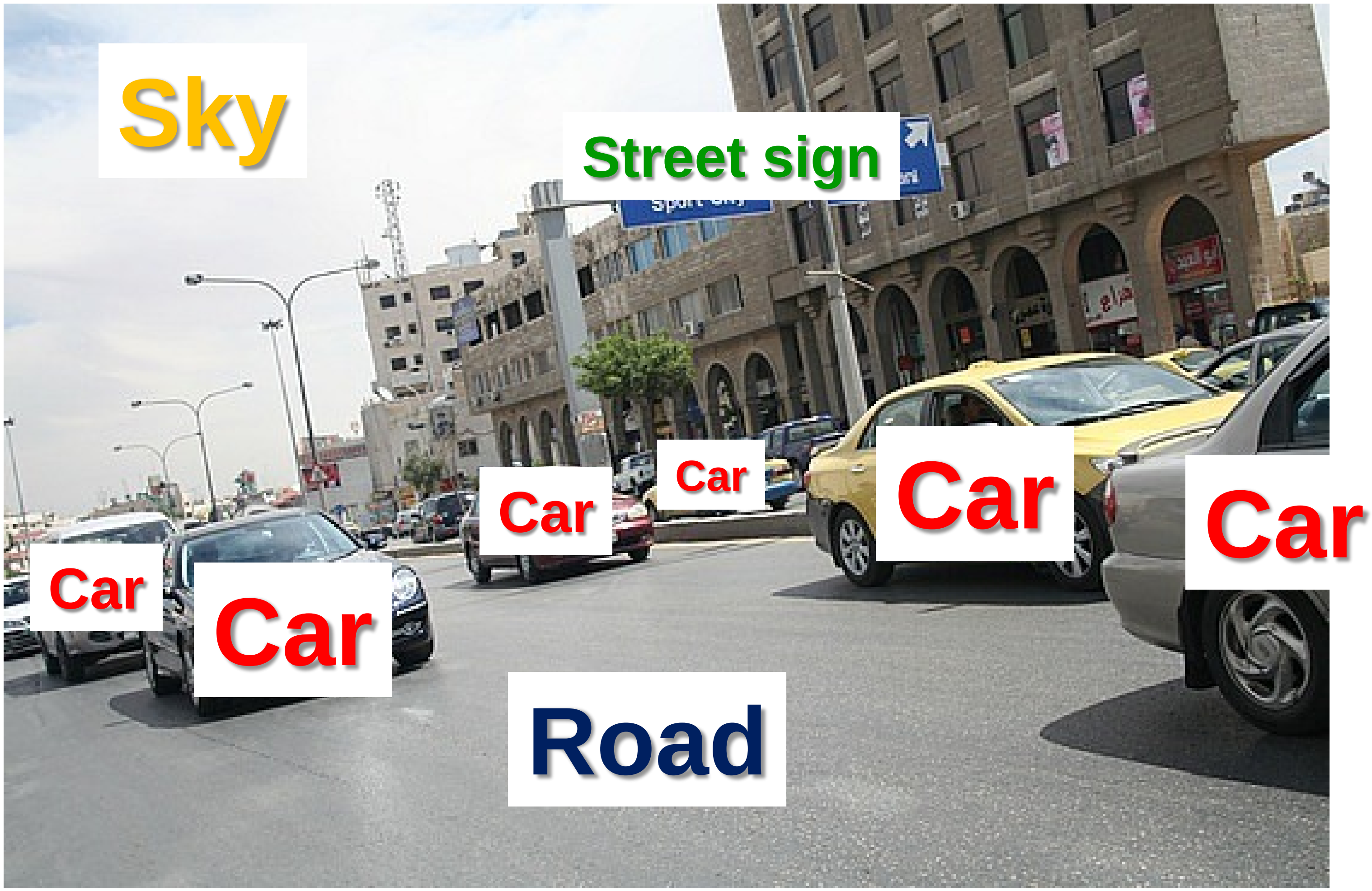
Do we want to learn *person-throws-cat-into-trash-bin* classifier?

How action recognition is related to other visual recognition tasks?

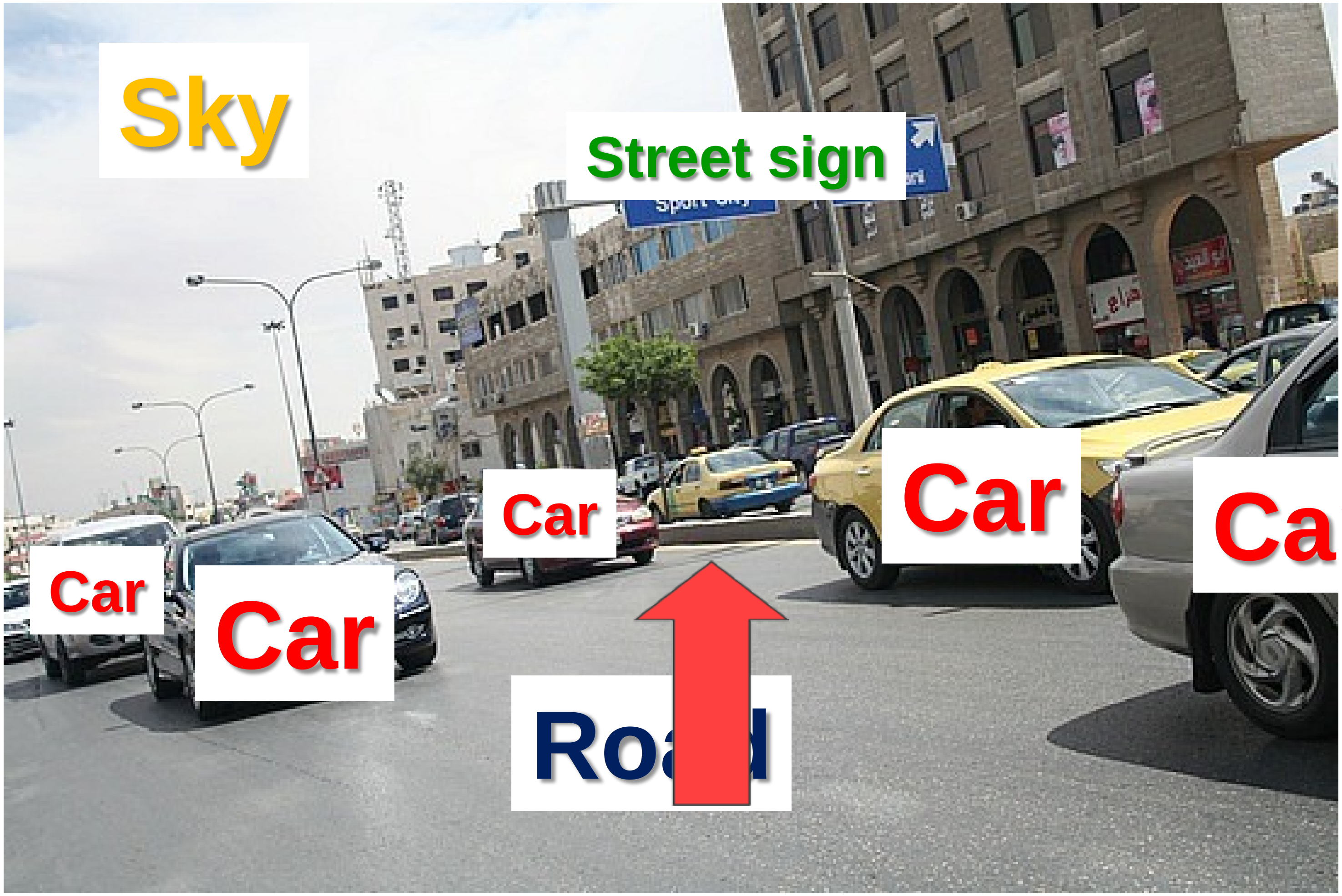


We can recognize cars and roads, What’s next?

What is missing in current methods?

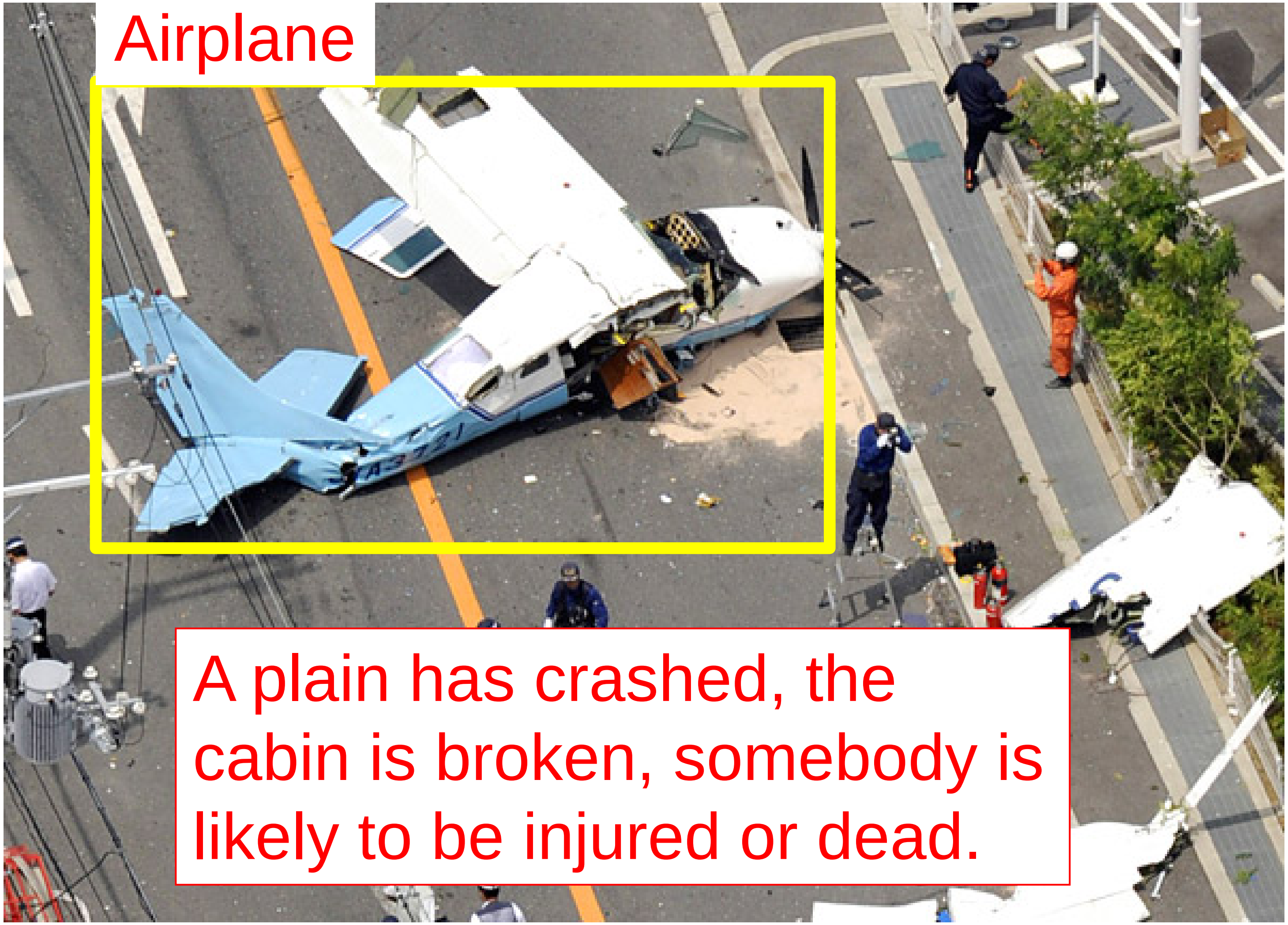


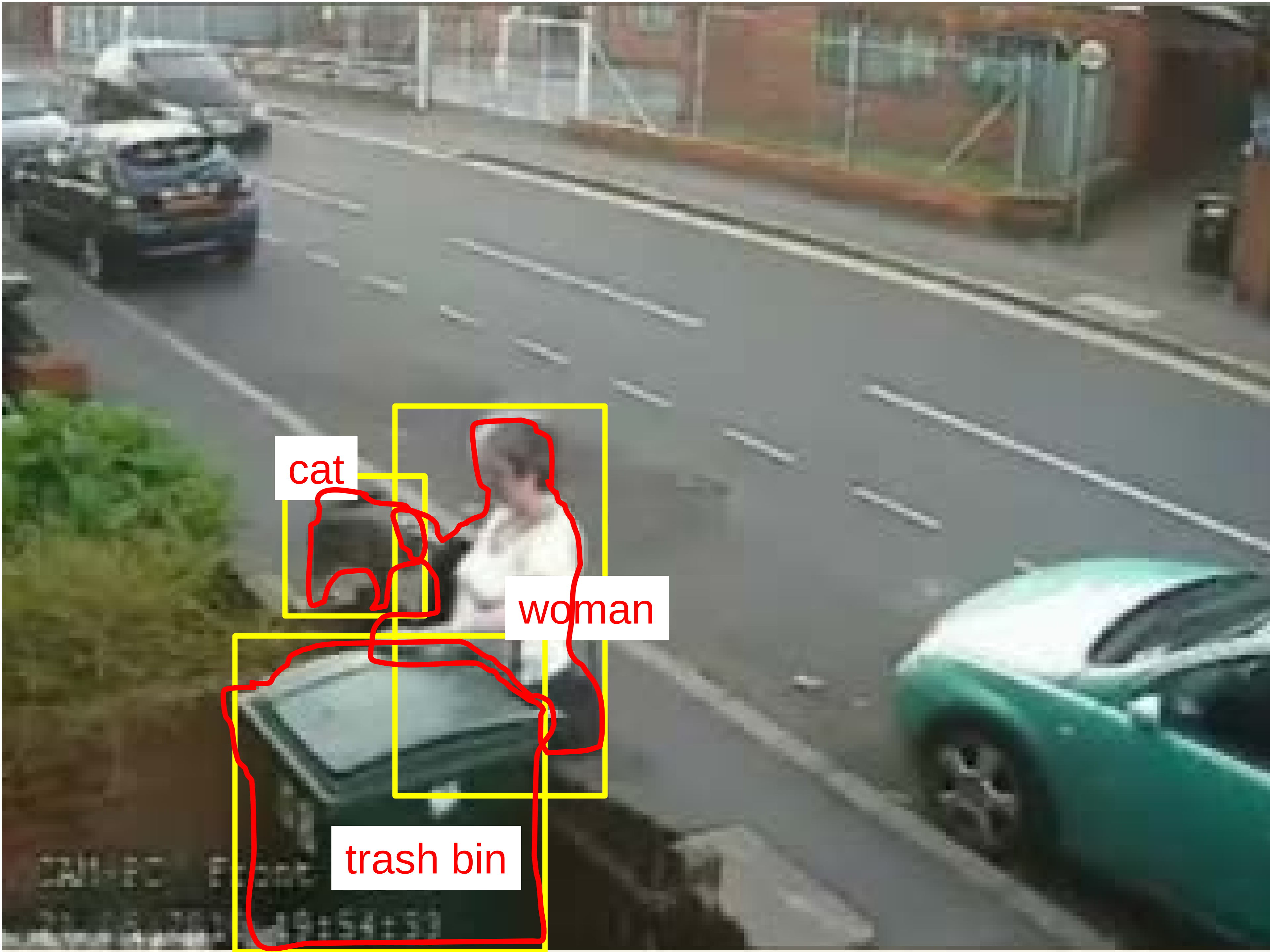
What is missing in current methods?



Object detection/classification won't help us to safely cross the street

What is missing in current methods?




cat
woman
trash bin

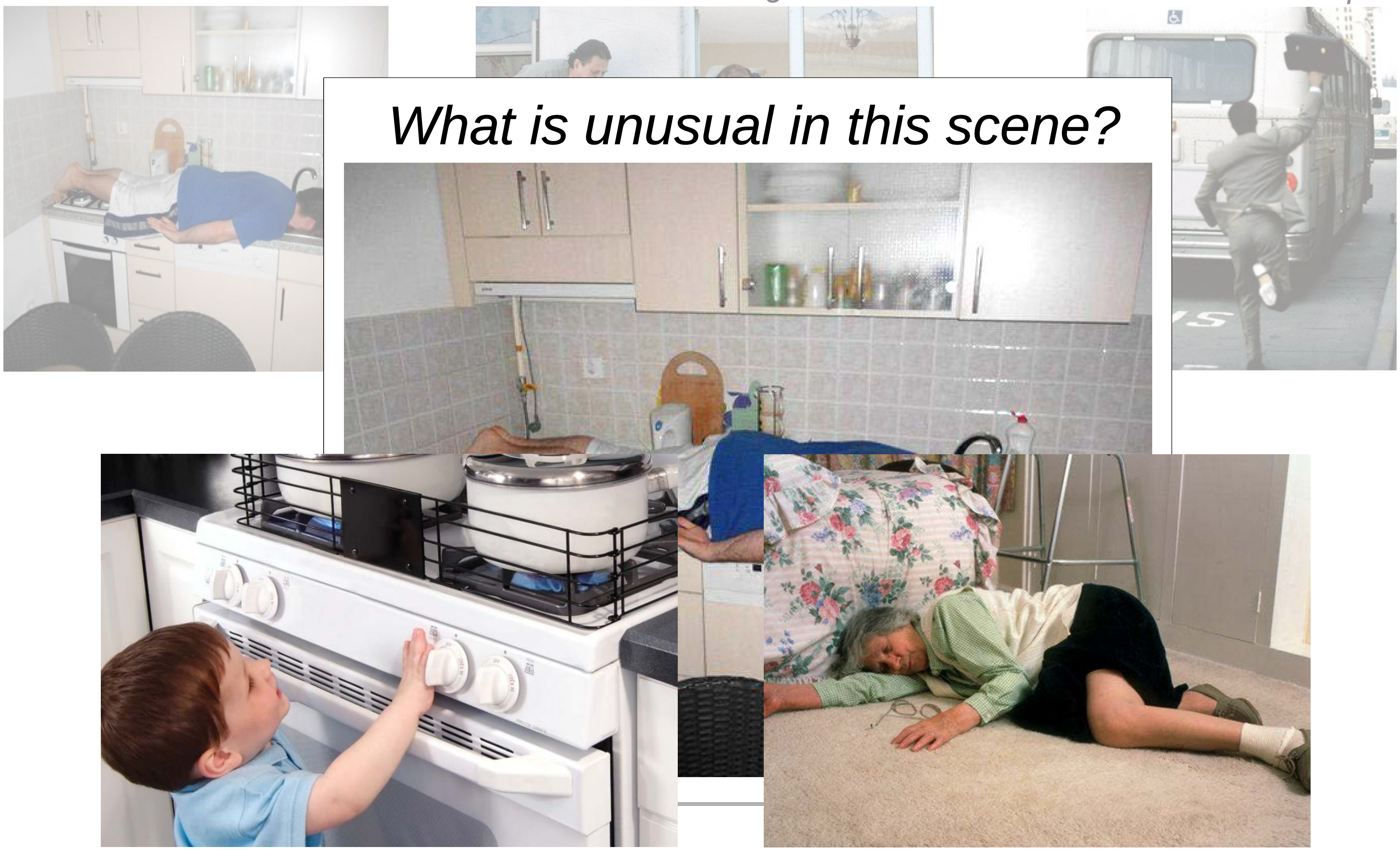
What current methods cannot do at all?

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

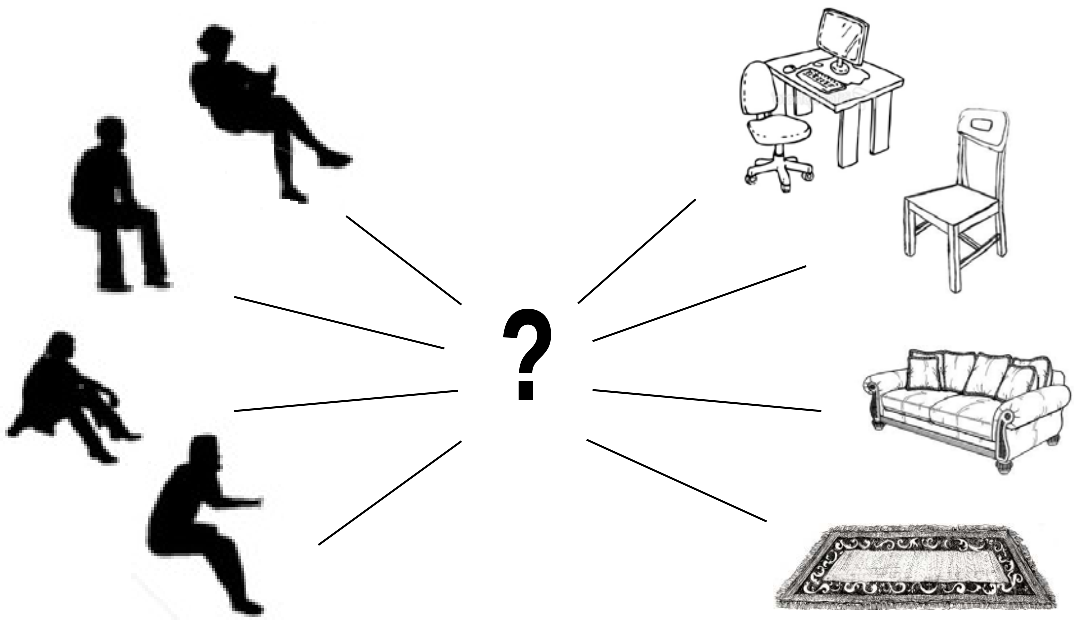
Is this a picture of a dog?
Is the person running in this video?

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

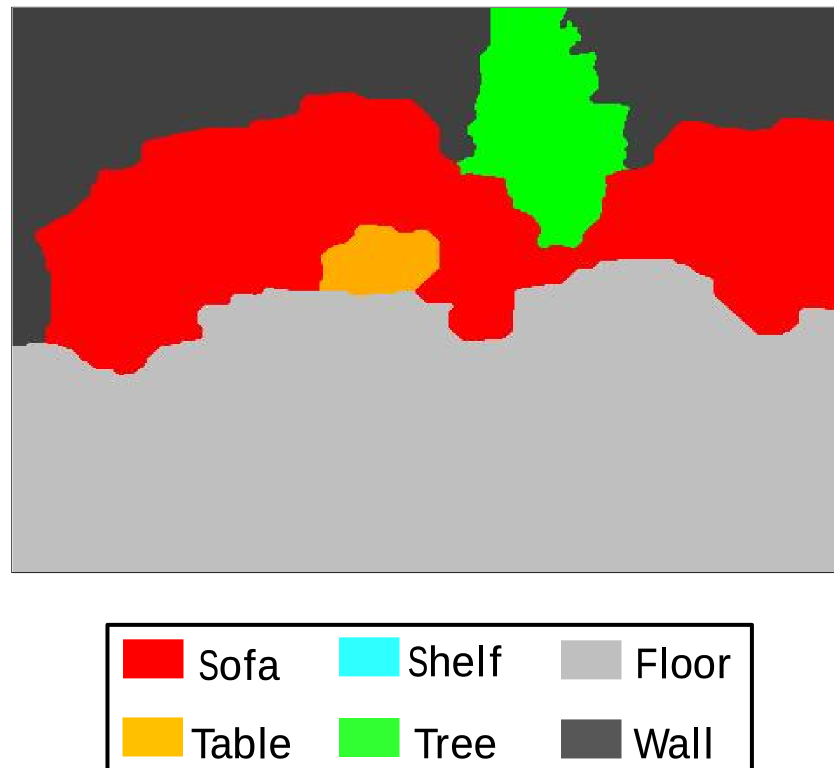
Goal

Recognize objects by the way people interact with them.

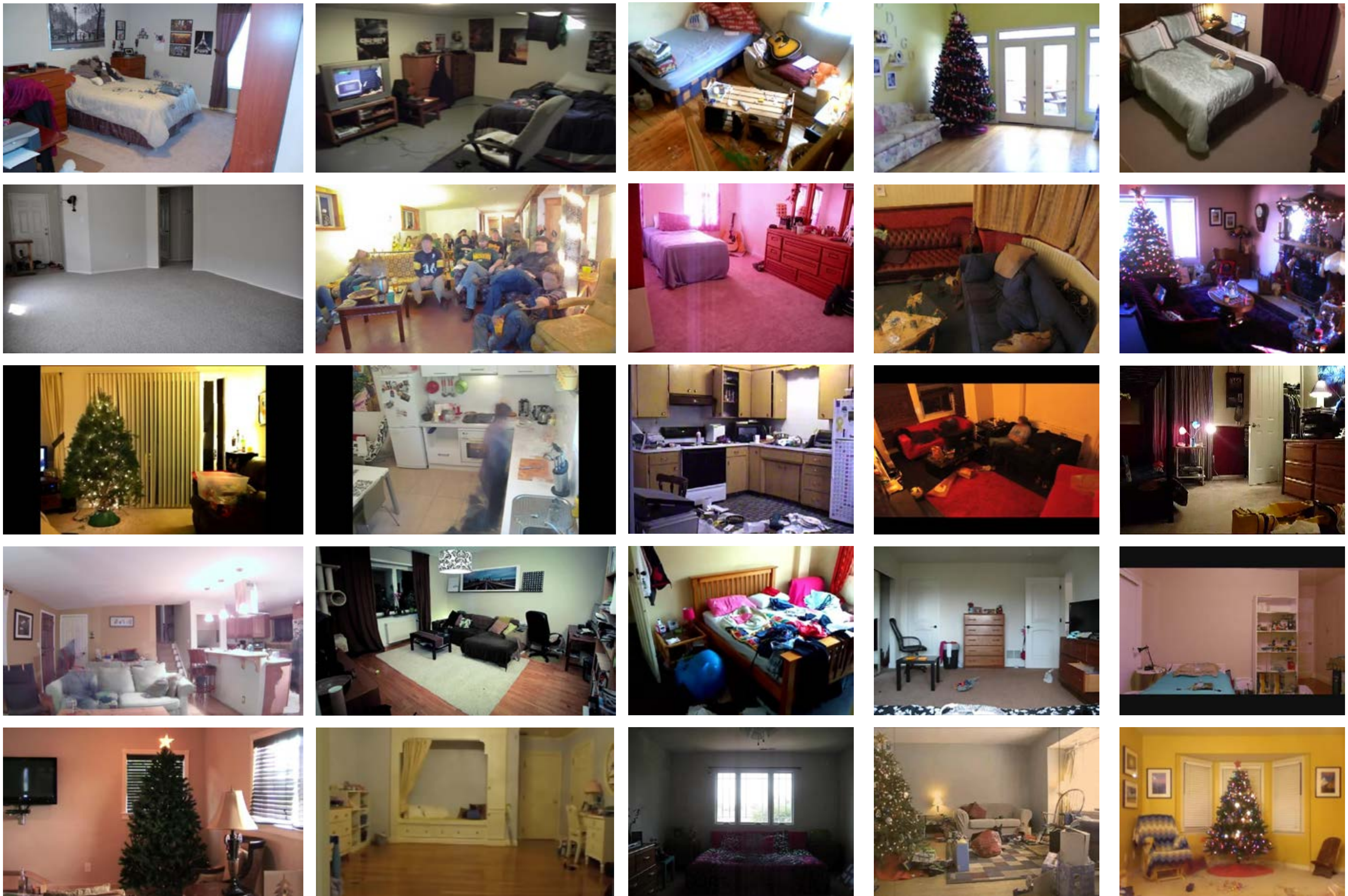
Time-lapse “Party & Cleaning” videos

Lots of person-object interactions,
many scenes on YouTube

Semantic object segmentation



New “Party & Cleaning” dataset



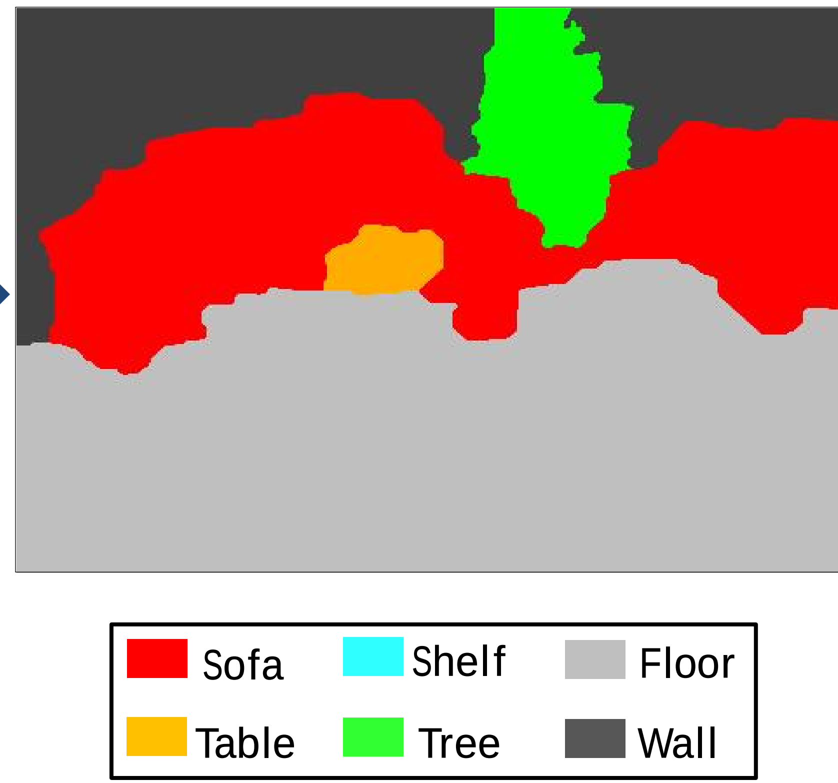
Goal

Recognize objects by the way people interact with them.

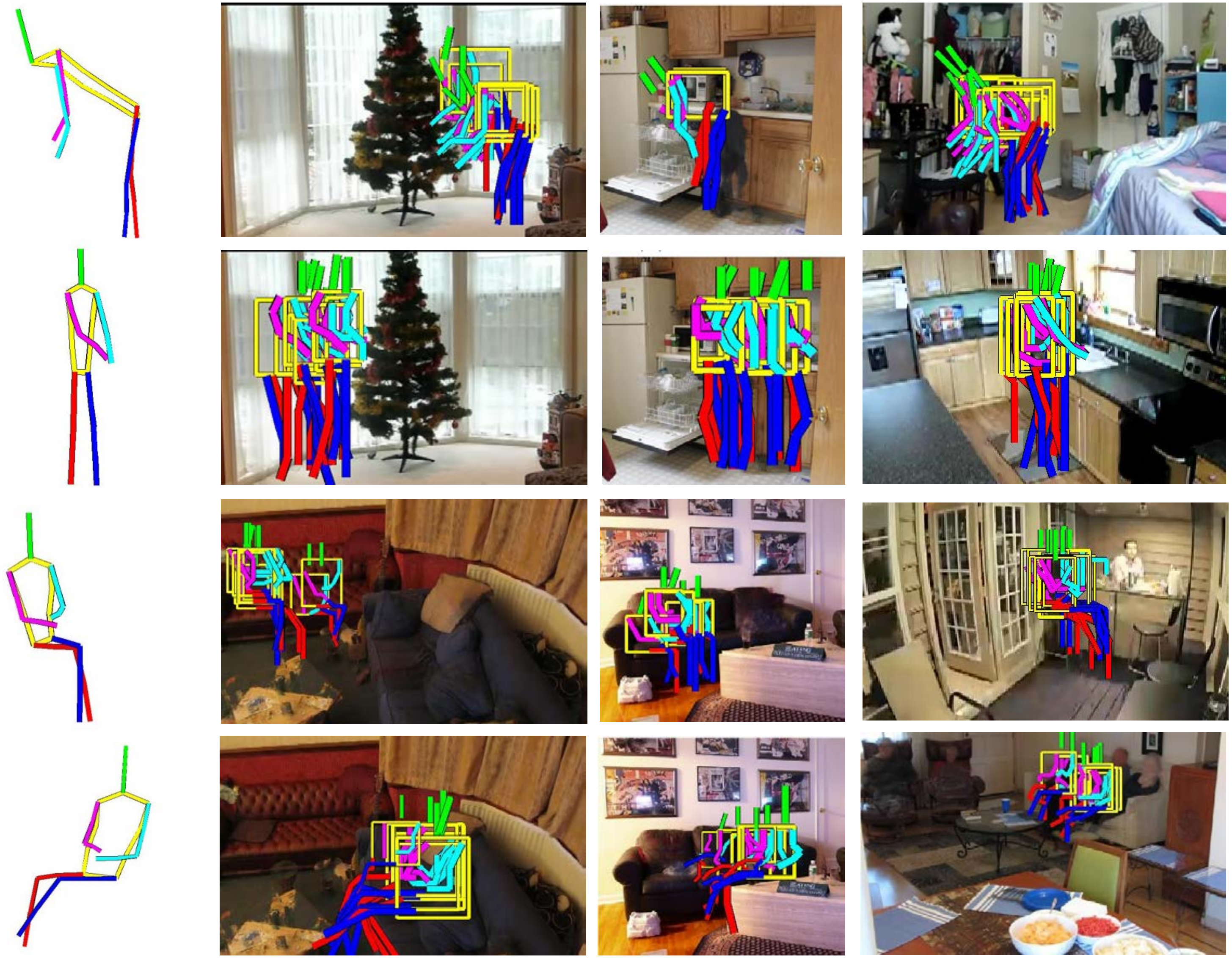
Time-lapse "Party & Cleaning" videos

Lots of person-object interactions, many scenes on YouTube

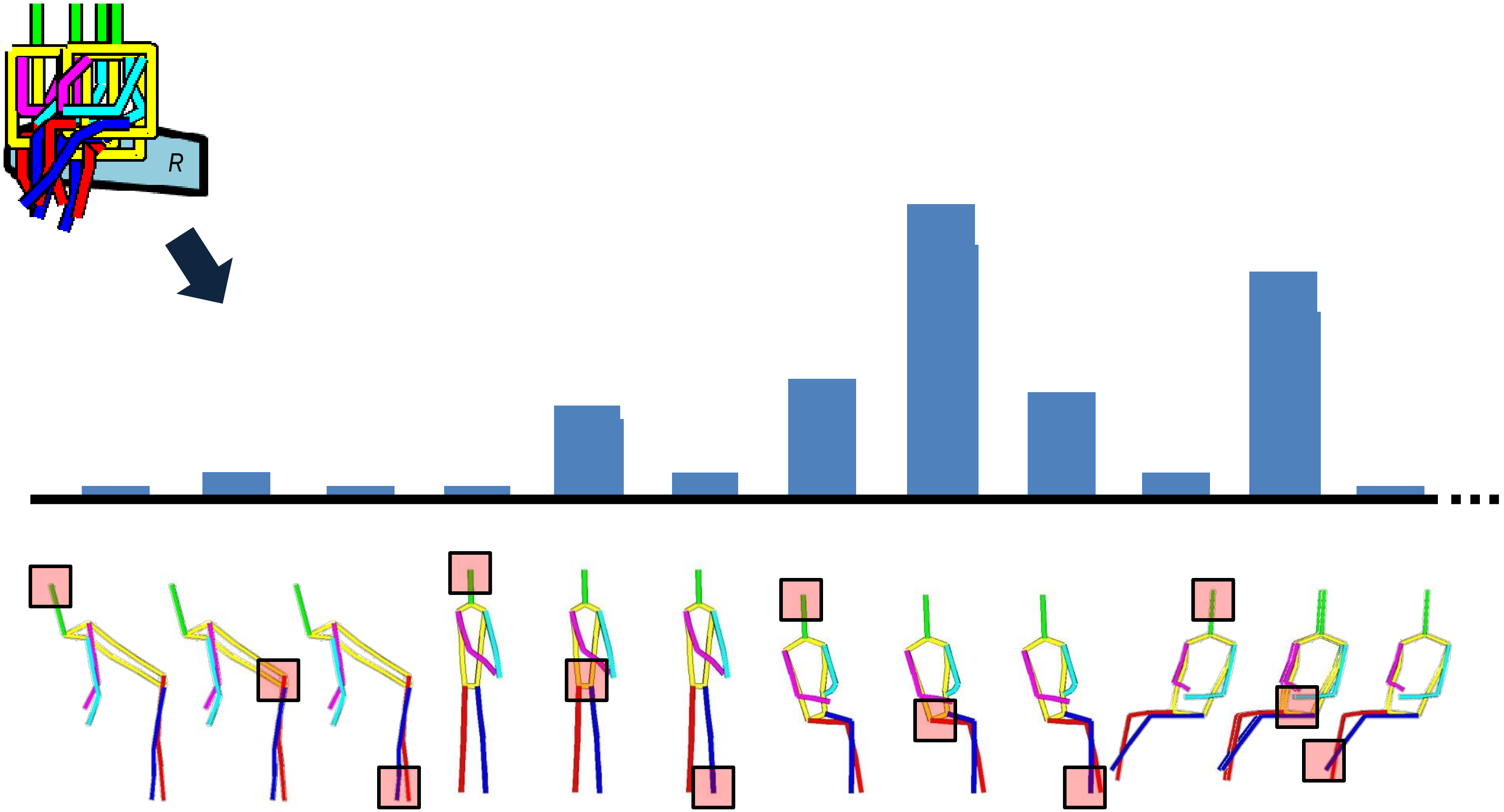
Semantic object segmentation



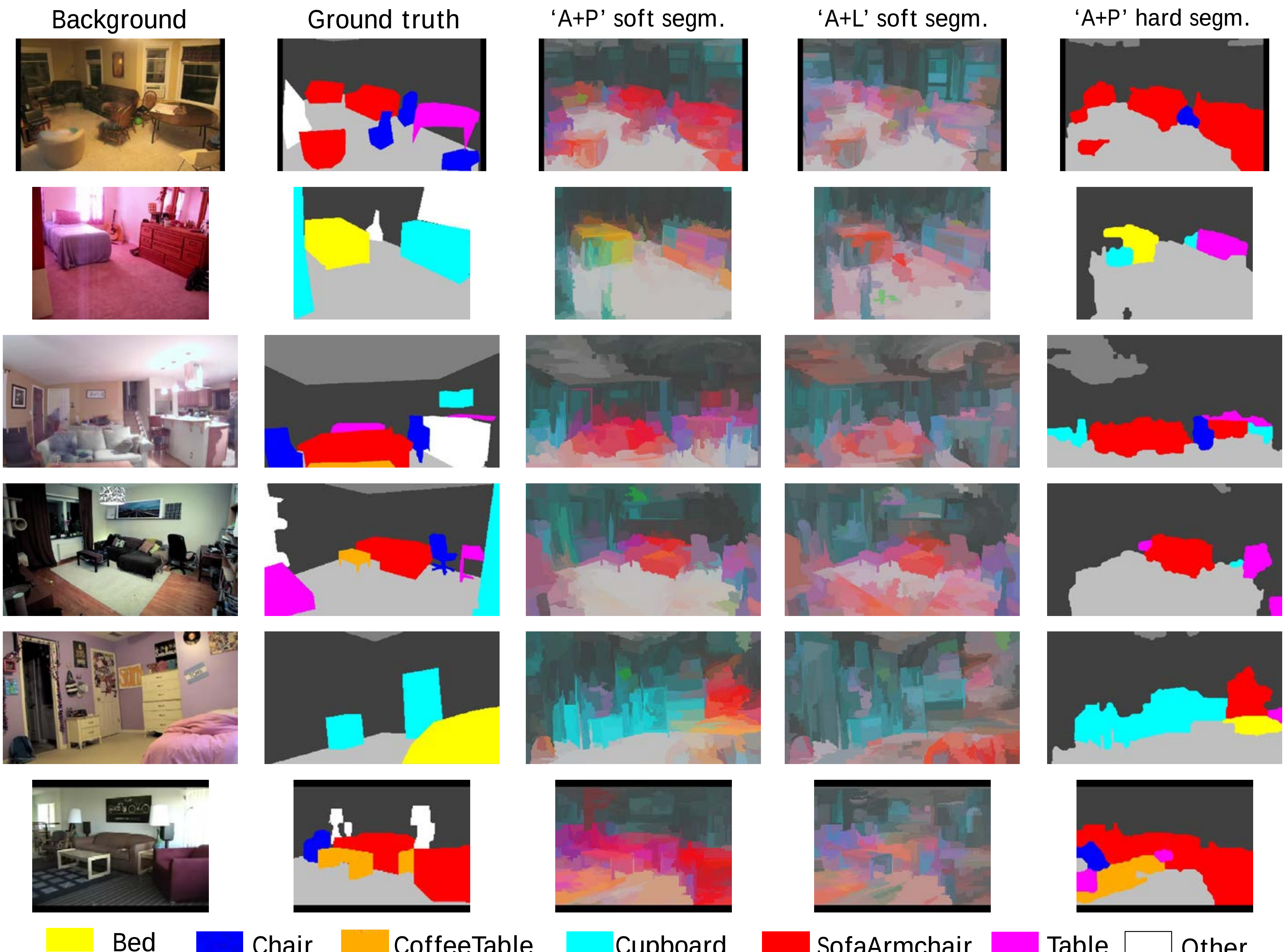
Pose vocabulary



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	$\mathbf{82} \pm 1.2$	81 ± 1.3
Ceiling	—	47 ± 20	53 ± 8.0	52 ± 7.4	$\mathbf{69} \pm 6.7$	$\mathbf{69} \pm 6.6$
Floor	—	59 ± 3.1	64 ± 5.5	65 ± 3.6	$\mathbf{76} \pm 3.2$	$\mathbf{76} \pm 2.9$
Bed	$\mathbf{31} \pm \mathbf{20}$	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	$\mathbf{44} \pm 5.4$	43 ± 5.8
Coffee Table	11 ± 5.4	11 ± 5.2	11 ± 4.4	12 ± 4.3	$\mathbf{17} \pm 10$	$\mathbf{17} \pm 9.6$
Chair	9.5 ± 3.9	6.3 ± 2.8	8.3 ± 2.7	5.8 ± 1.4	11 ± 5.4	$\mathbf{12} \pm 5.9$
Table	15 ± 6.4	18 ± 3.8	17 ± 3.9	16 ± 7.1	$\mathbf{22} \pm 6.2$	$\mathbf{22} \pm 6.4$
Wardrobe/Cupboard	27 ± 10	27 ± 8.2	28 ± 6.4	22 ± 1.1	$\mathbf{36} \pm 7.4$	$\mathbf{36} \pm 7.2$
Christmas tree	50 ± 3.3	55 ± 12	72 ± 1.8	20 ± 6.0	76 ± 6.2	$\mathbf{77} \pm 5.5$
Other Object	12 ± 6.4	11 ± 1.2	7.9 ± 1.9	13 ± 4.2	$\mathbf{16} \pm 8.3$	$\mathbf{16} \pm 8.2$
Average	23 ± 1.8	31 ± 2.0	35 ± 2.4	30 ± 1.7	$\mathbf{43} \pm 4.4$	$\mathbf{43} \pm 4.3$

A: Appearance (SIFT) histograms;

L: Location;

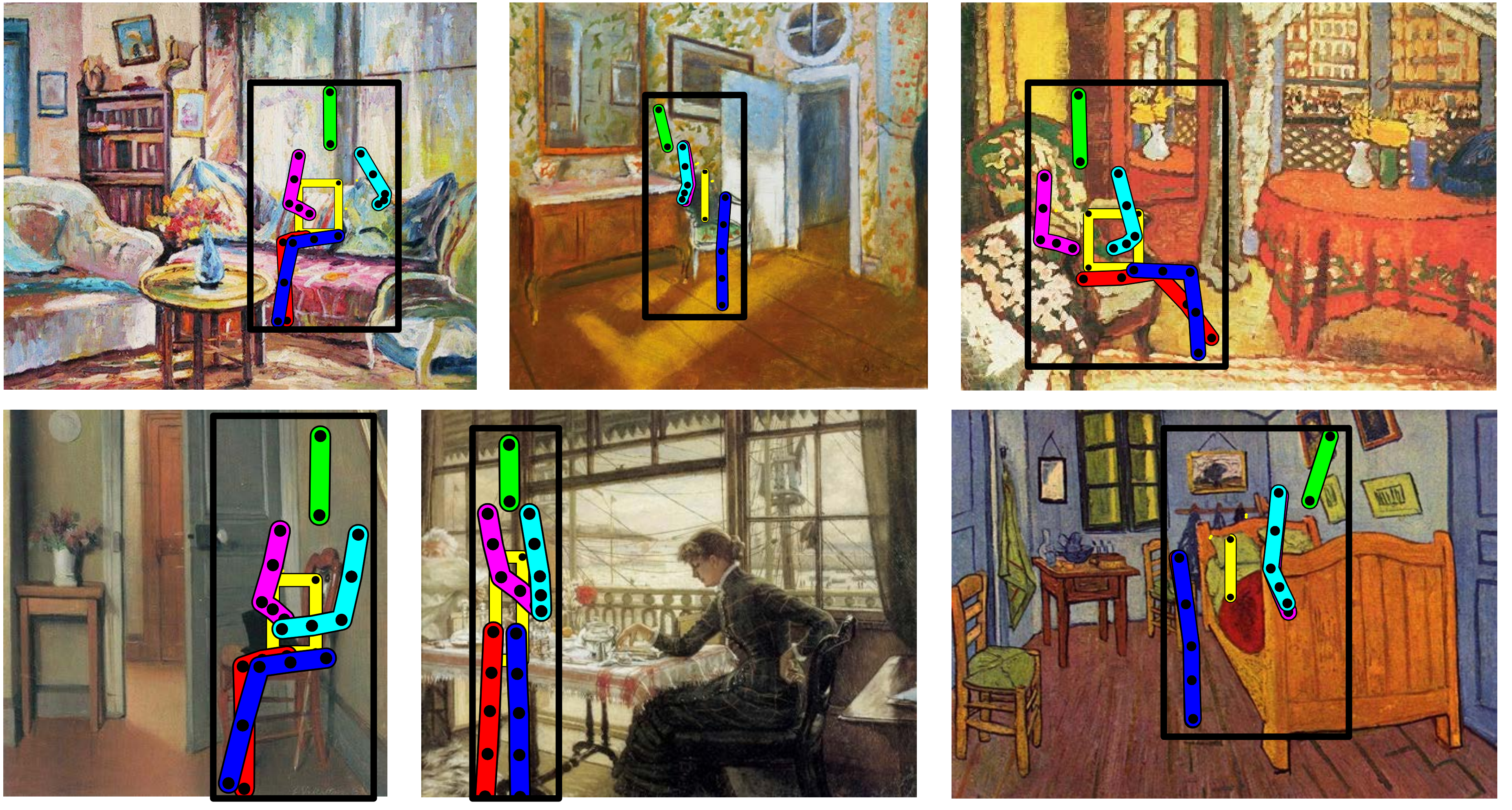
P: Pose histograms

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