CVML2012
INRIA Visual Recognition and Machine Learning Summer School
Grenoble, France, 9-13 July 2012

# Human Actions: History, Progress, Open problems

Ivan Laptev

*ivan.laptev@inria.fr*

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

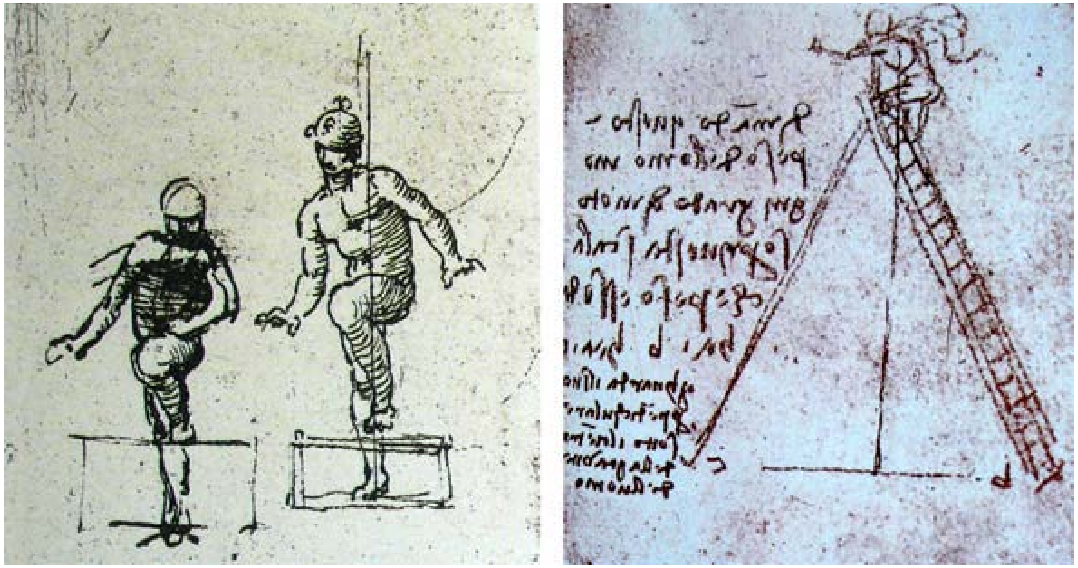
# Human Actions:
# Why do we care?

# Motivation I: Artistic Representation

Early studies were motivated by human representations in Arts

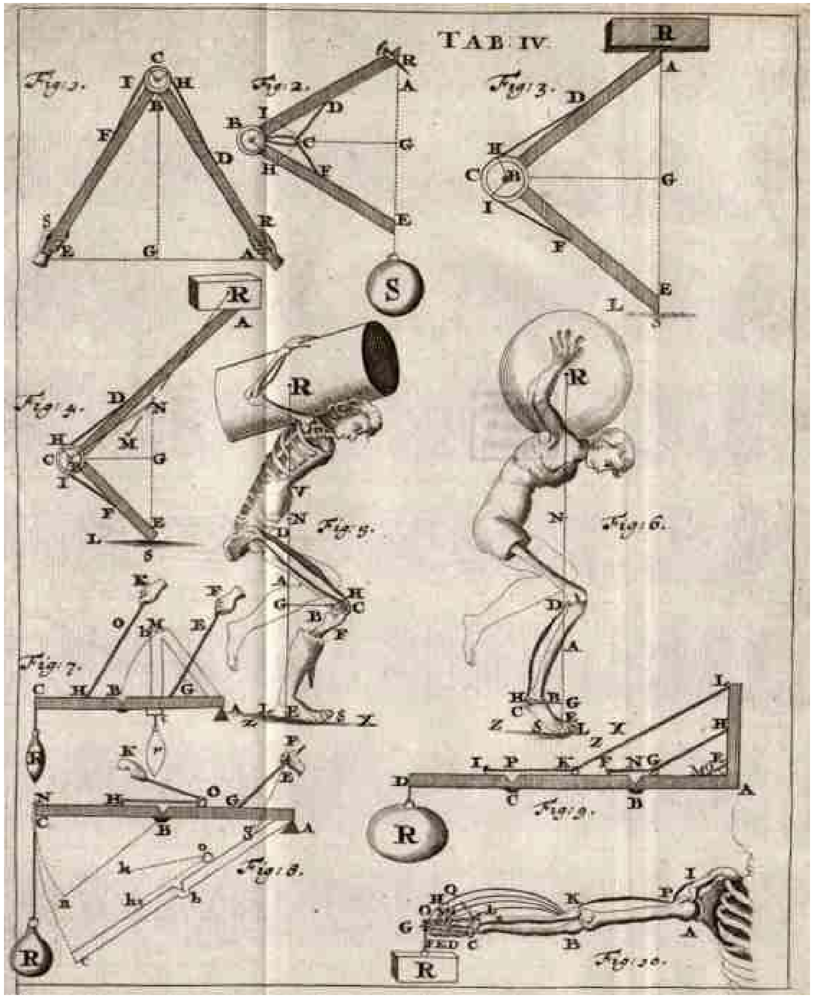
Da Vinci: “it is indispensable for a painter, to become totally familiar with the anatomy of nerves, bones, muscles, and sinews, such that he understands for their various motions and stresses, which sinews or which muscle causes a particular motion”

“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.

# Motivation II: Biomechanics

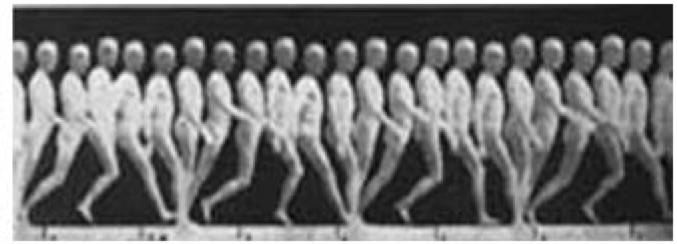


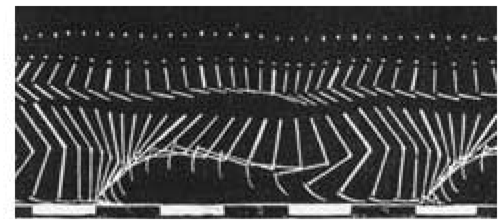

Giovanni Alfonso Borelli (1608–1679)

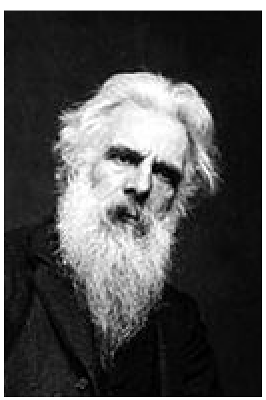
- The emergence of *biomechanics*
- Borelli applied to biology the analytical and geometrical methods, developed by Galileo Galilei
- He was the first to understand that bones serve as levers and muscles function according to mathematical principles
- His physiological studies included muscle analysis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

# Motivation III: Motion perception

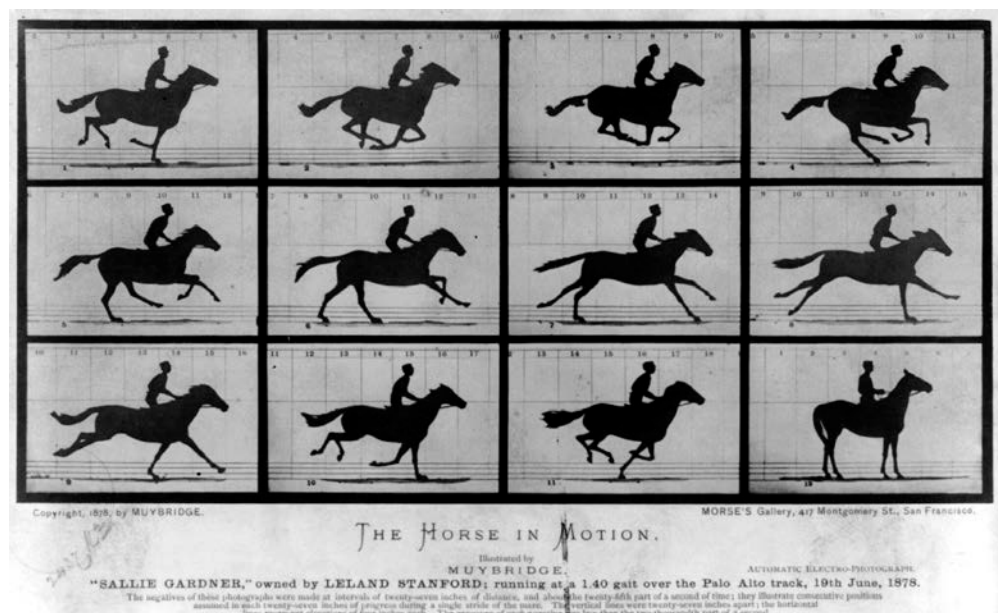
**Etienne-Jules Marey:** (1830–1904) made Chronophotographic experiments influential for the emerging field of *cinematography*





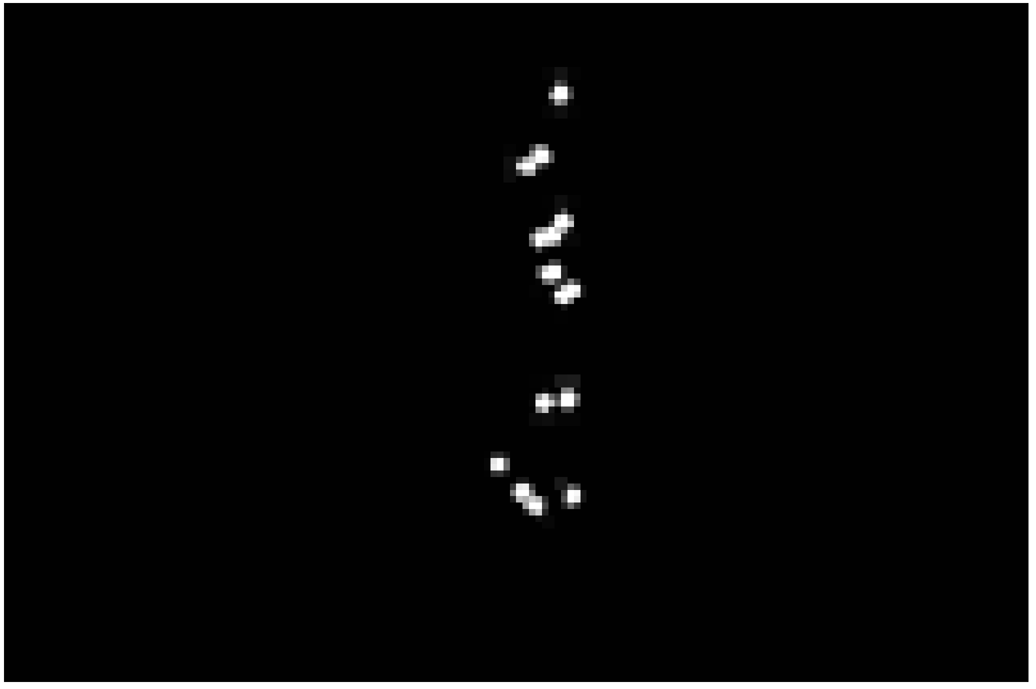


**Eadweard Muybridge** (1830–1904) invented a machine for displaying the recorded series of images. He pioneered motion pictures and applied his technique to movement studies



# Motivation III: 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



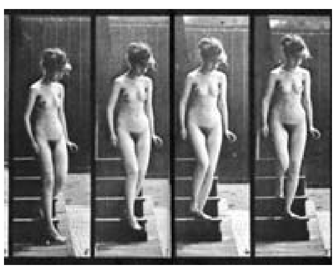
A Teaching Resource

At the Frontiers of Psychological Inquiry

# Human actions: Historic overview

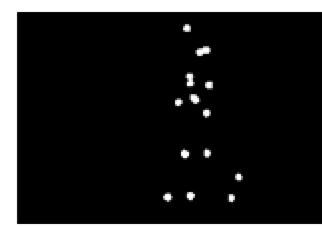
15$^{th}$ century
studies of
anatomy

17$^{th}$ century
emergence of
*biomechanics*



19$^{th}$ century
emergence of
*cinematography*

1971
studies of human
motion perception

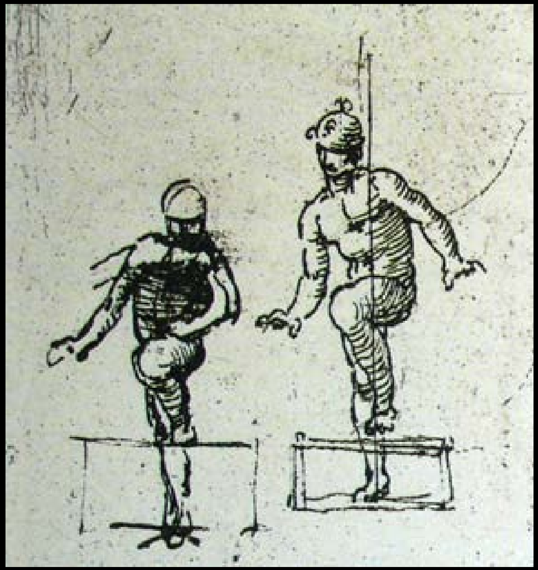


Modern computer vision

# Modern applications: Motion capture and animation

Avatar (2009)

# Modern applications: Motion capture and animation



Leonardo da Vinci (1452–1519)

Avatar (2009)

# Modern applications: Video editing


*Space-Time Video Completion*
Y. Wexler, E. Shechtman and M. Irani, **CVPR** 2004

# Modern applications: Video editing

*Recognizing Action at a Distance*
Alexei A. Efros, Alexander C. Berg, Greg Mori, Jitendra Malik, **ICCV** 2003

# Modern applications: Video editing


*Recognizing Action at a Distance*
Alexei A. Efros, Alexander C. Berg, Greg Mori, Jitendra Malik, **ICCV** 2003

# Technology: Access to lots of data

- Huge amount of video is available and growing

TV-channels recorded since 60's

>34K hours of video uploads every day


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

# Applications

- Video indexing and search is useful for TV production, entertainment, education, social studies, security, special effects…

TV & Web: e.g. *“Fight in a parlament”*

Home videos: e.g. *“My daughter climbing”*

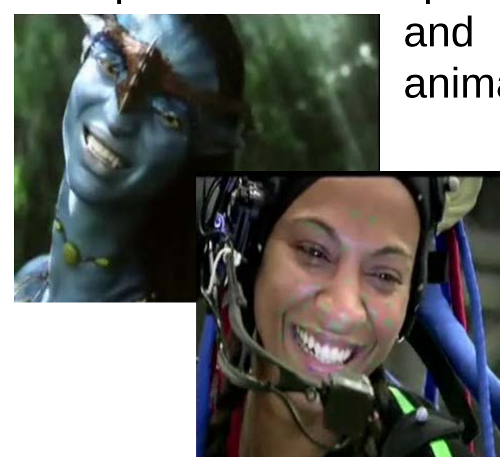
Sociology research:

Manually analyzed smoking actions in 900 movies

Surveillance

suspicious behavior detection

Graphics: motion capture and animation



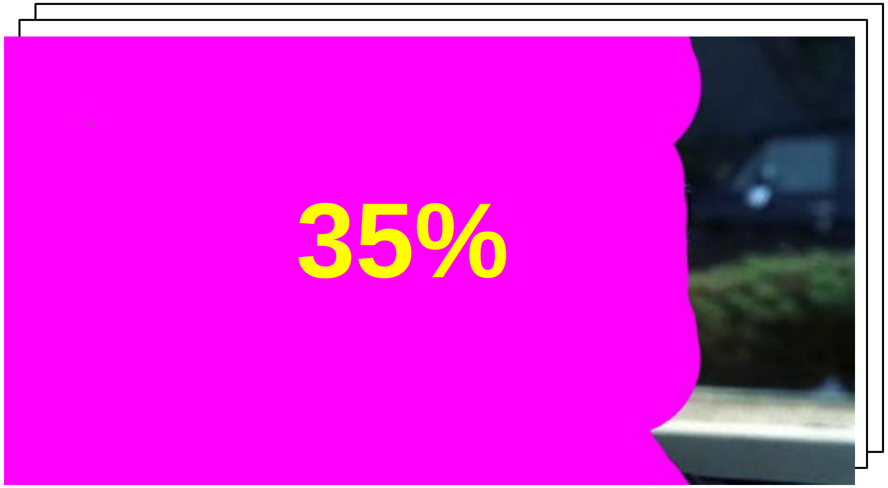
# How many person pixels are in video?

Movies

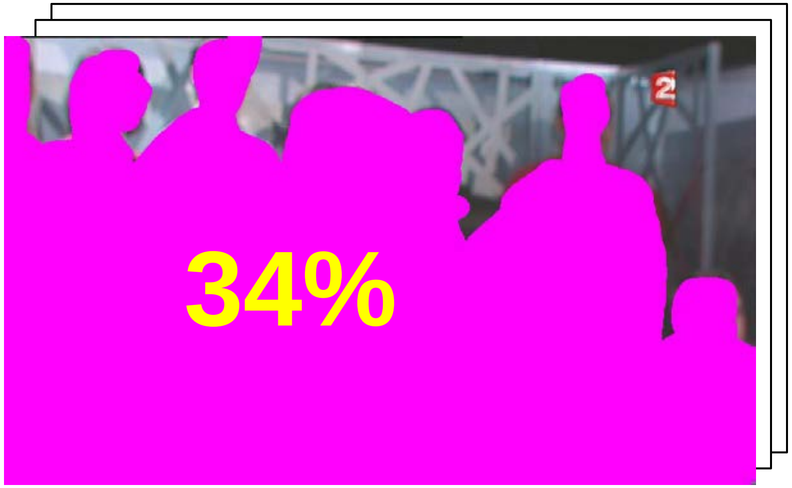

TV

YouTube

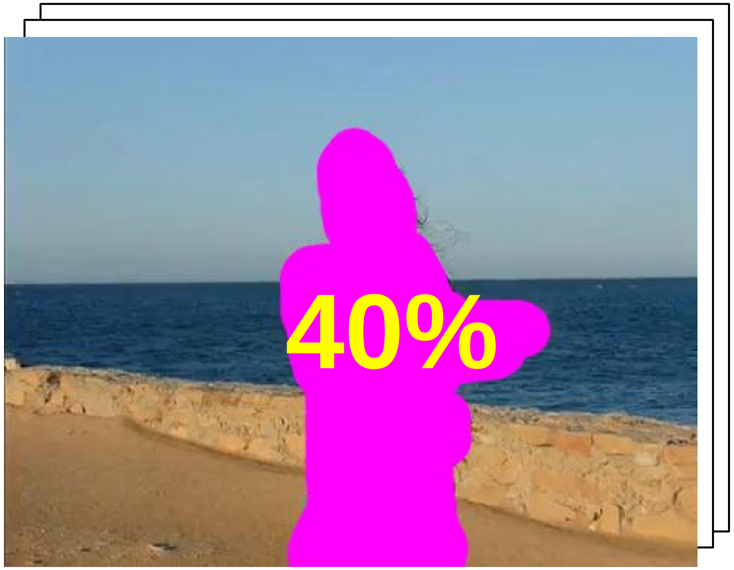
# How many person pixels are in video?




Movies




TV




YouTube

# Why action recognition is difficult?

# Why action recognition is difficult?

- Much diversity in the data (view-points, appearance, motion, lighting…)

Drinking

Smoking

- Many classes and concepts

# How to recognize actions: History

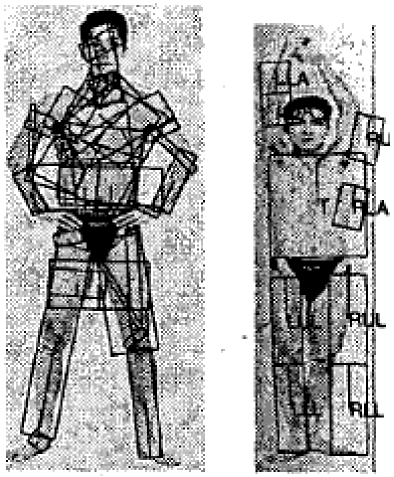
# A HOUGHTON MIFFLIN PRODUCTION



A Teaching Resource

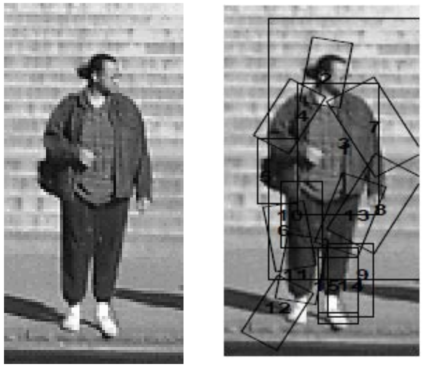
At the Frontiers of Psychological Inquiry

# Human pose estimation (1990-2000)



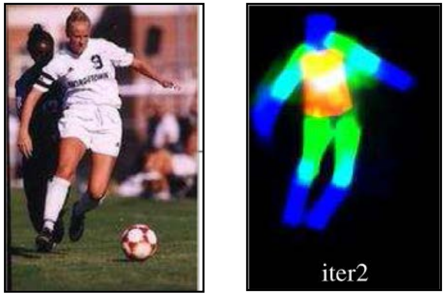
Finding People by Sampling
Ioffe & Forsyth, ICCV 1999

Pictorial Structure Models for Object Recognition
Felzenszwalb & Huttenlocher, 2000



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

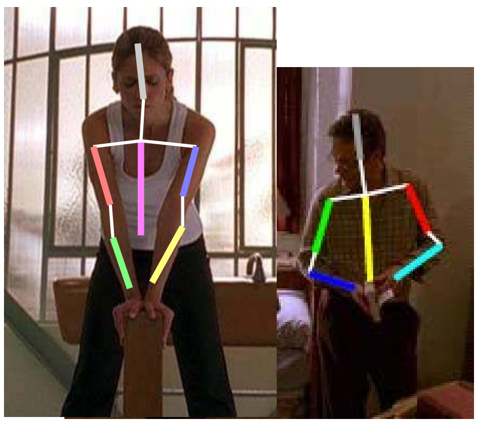
# Human pose estimation (2000-2010)




D. Ramanan. Learning to parse images of articulated bodies. NIPS, 2007

Learn image and person-specific unary terms

- initial iteration → edges
- following iterations → edges & colour



V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In Proc. CVPR, 2008/2009

(Almost) unconstrained images

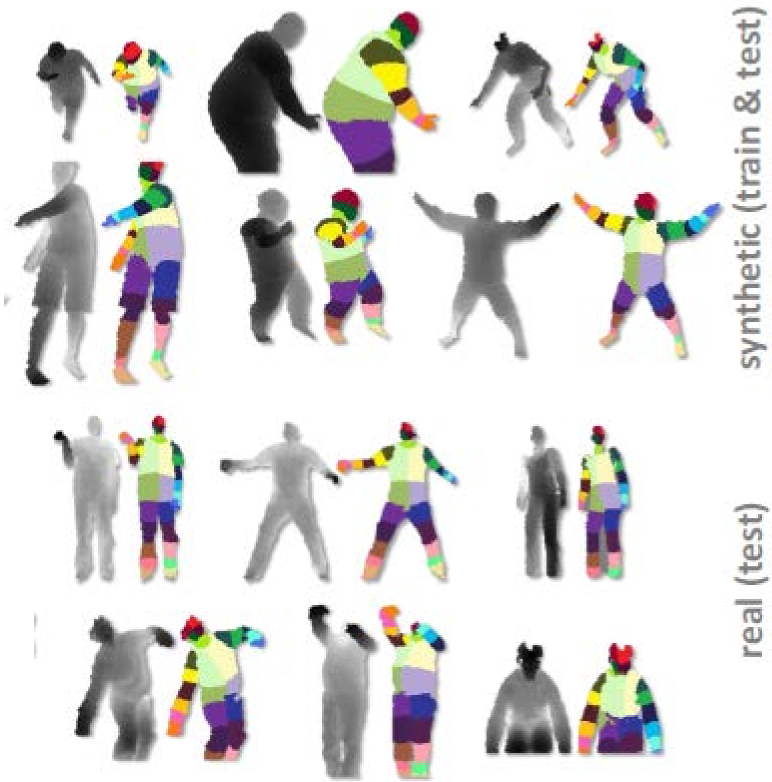
- Person detector & foreground highlighting

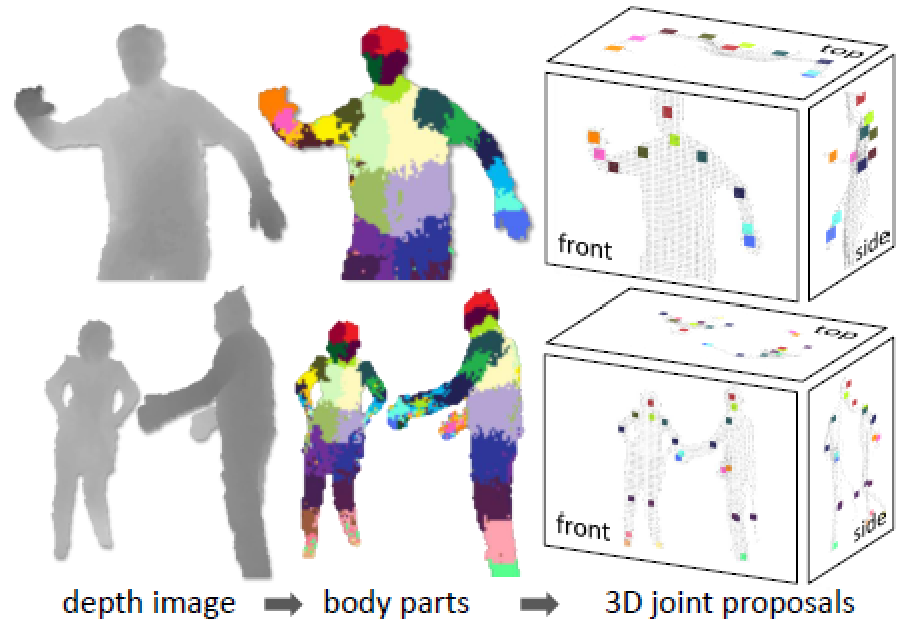

VP. Buehler, M. Everingham and A. Zisserman. Learning sign language by watching TV. In Proc. CVPR 2009

Learns with weak textual annotation

- Multiple instance learning

# Human pose estimation (2011)

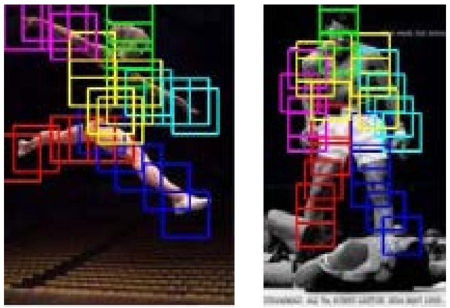



J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. **Best paper award at CVPR 2011**

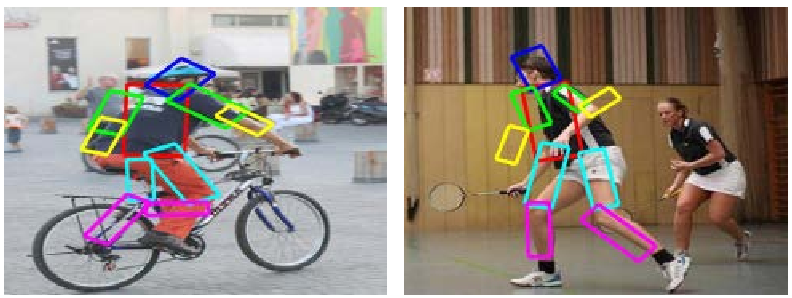
Exploits lots of synthesized depth images for training

# Human pose estimation (2011)



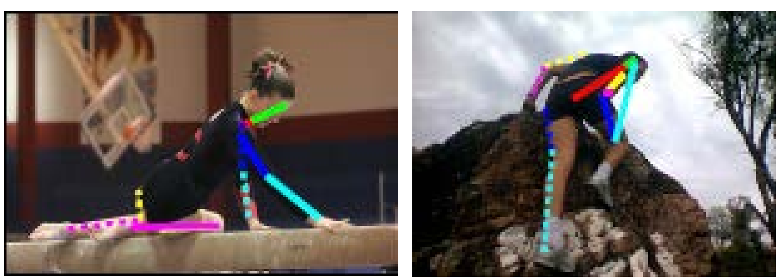
Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In Proc. **CVPR 2011**

Extension of LSVM model of Felzenszwalb et al.



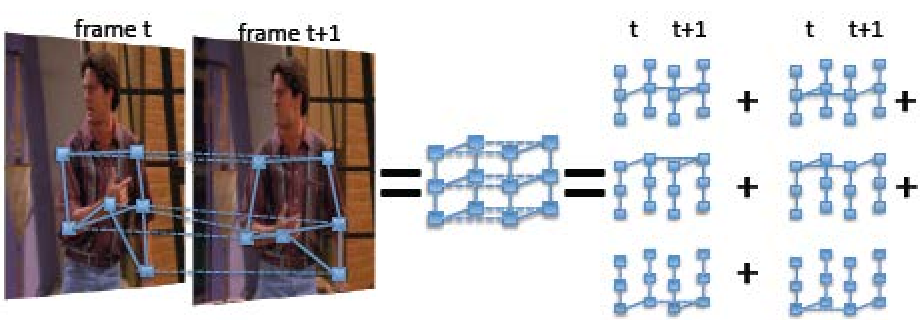
Y. Wang, D. Tran and Z. Liao. Learning Hierarchical Poselets for Human Parsing. In Proc. **CVPR 2011**.

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**.

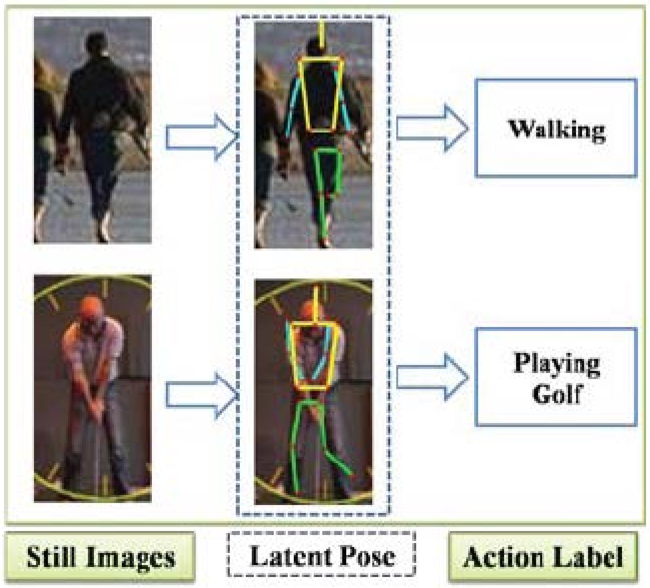
Learns from lots of noisy annotations




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

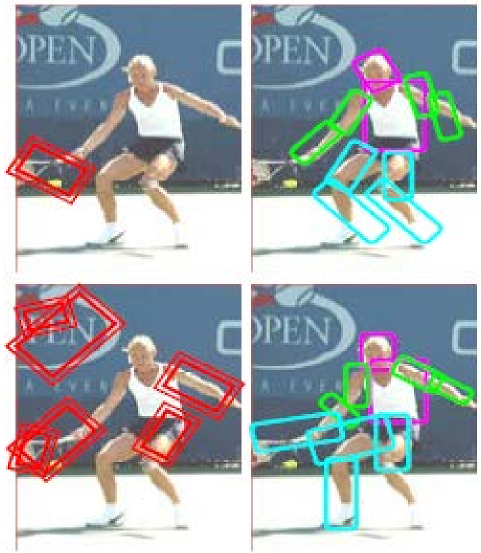
Explores temporal continuity

# 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.

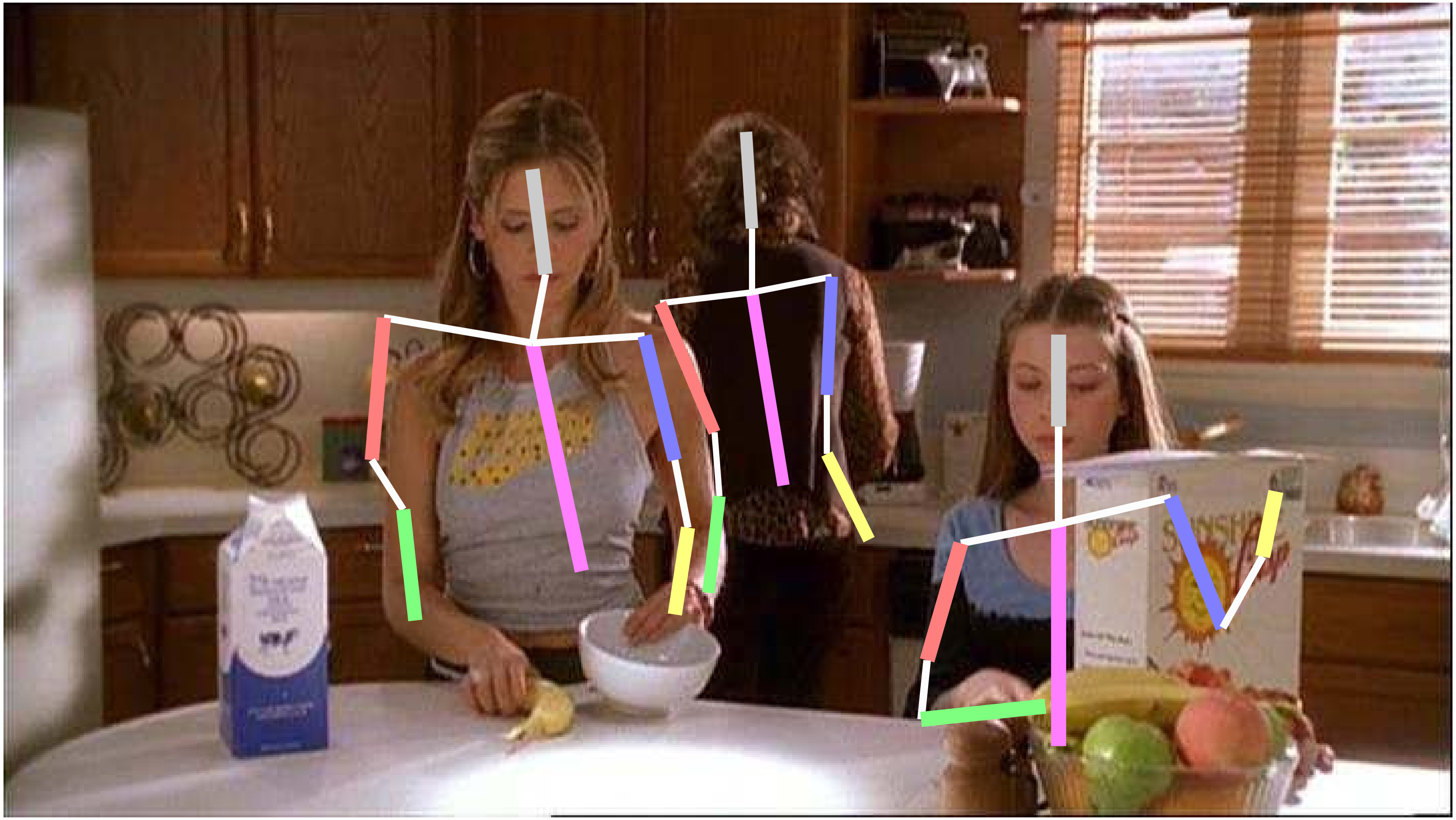
Some limbs may not be important for recognizing a particular action (e.g. sitting)



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

# Pose estimation is still a hard problem

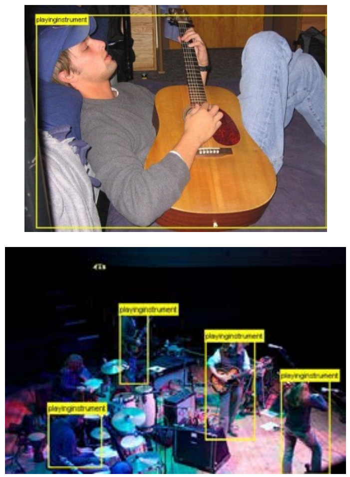


Issues:
- occlusions
- clothing and pose variations

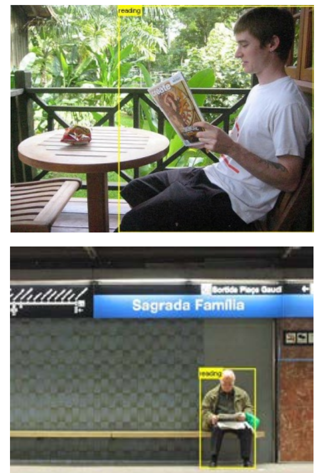
# PASCAL VOC Action classification

Phoning

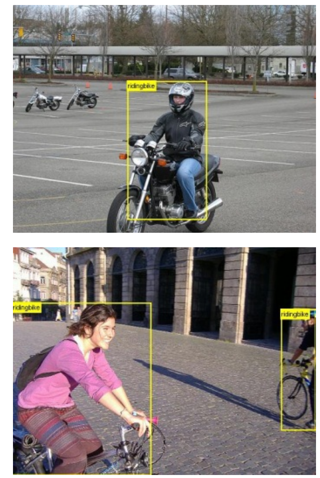
Playing Instrument



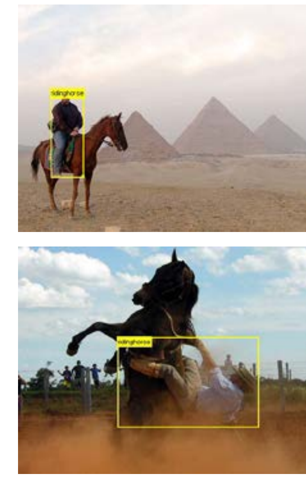
Reading




Riding Bike



Riding Horse

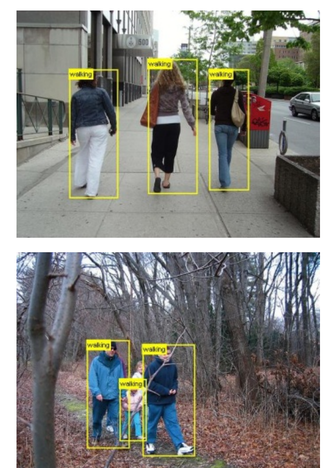


Running

Taking Photo

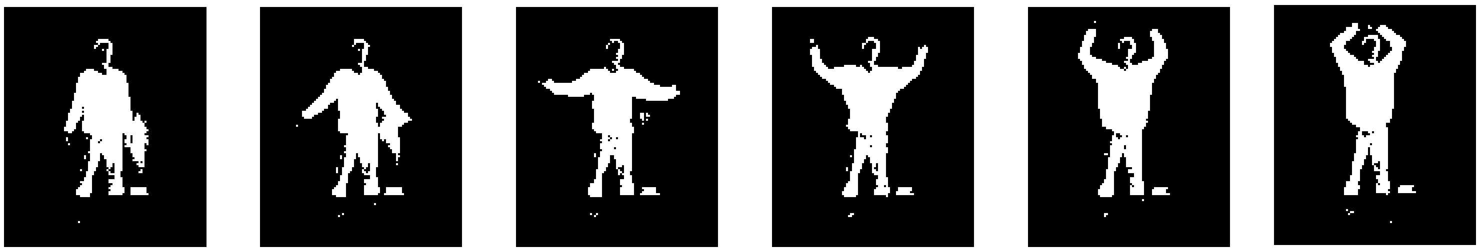
Using Computer

Walking



# Appearance-based methods: background subtraction

$D(x, y, t) \quad t = 1, ..., T$



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

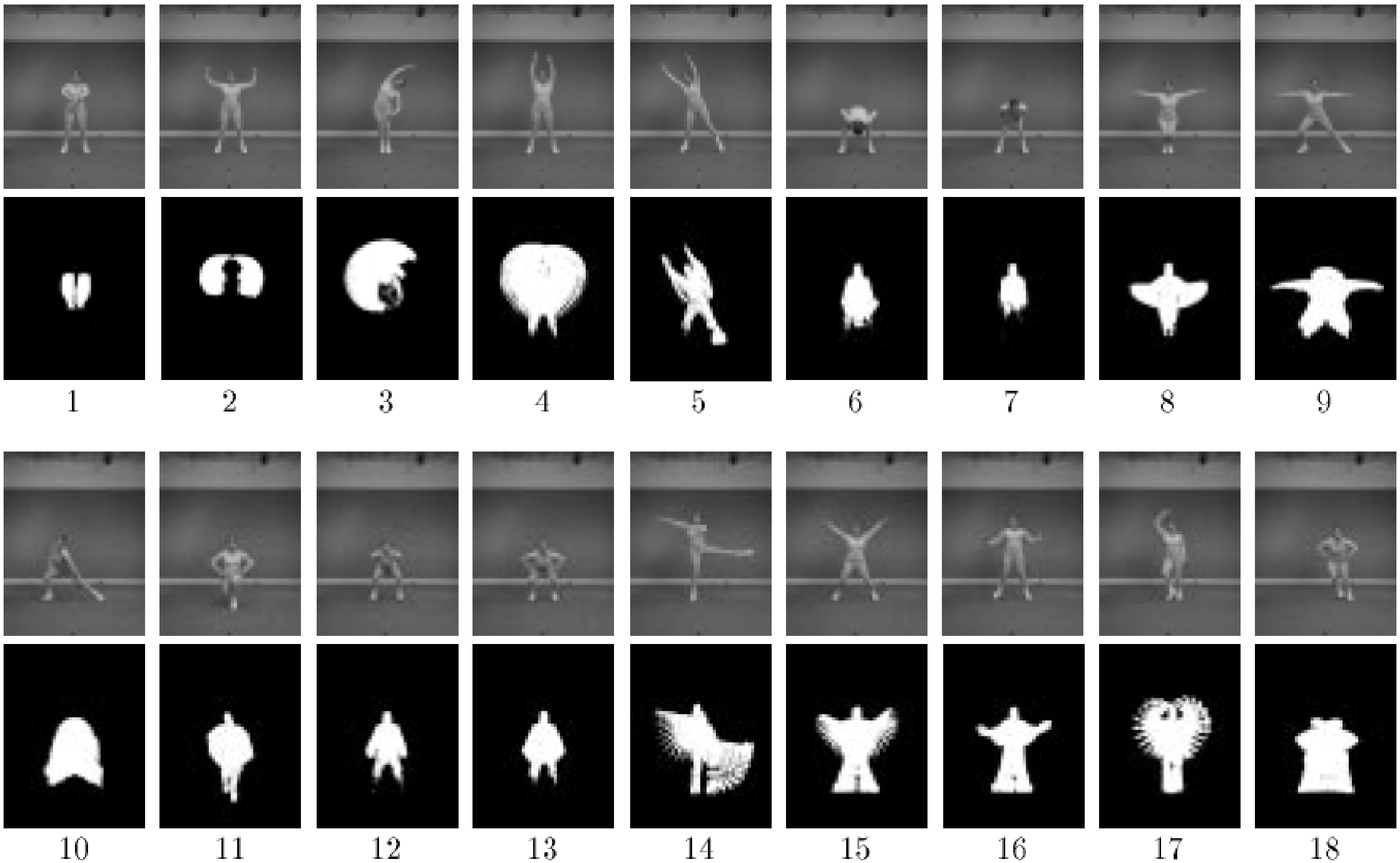
$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ \max\ (0, H_\tau(x, y, t-1) - 1) & \text{otherwise} \end{cases}$$

Descriptor: Hu moments of different orders

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q \rho(x, y) dx dy$$

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

# Aerobics dataset




Nearest Neighbor classifier: 66% accuracy

# Temporal Templates: Summary

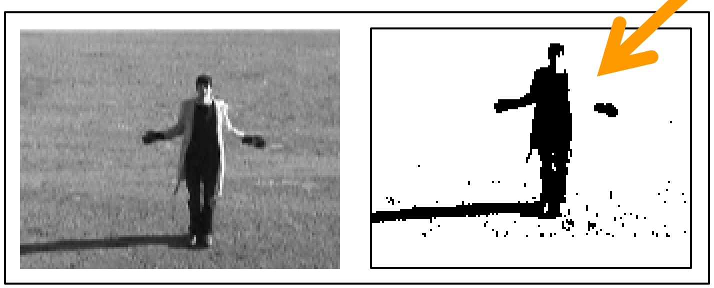
Pros:

+ Simple and fast
+ Works in controlled settings

Not all shapes are valid → Restrict the space of admissible silhouettes

Cons:

- Prone to errors of background subtraction



Variations in light, shadows, clothing…

What is the background here?

- Does not capture *interior* motion and shape

Silhouette tells little about actions

# Active Shape Models of Cootes et al.

**Point Distribution Model**

- Represent the shape of samples by a set of corresponding points or *landmarks*

$$\mathbf{x} = (x_1, \ldots, x_n, y_1, \ldots, y_n)^T$$

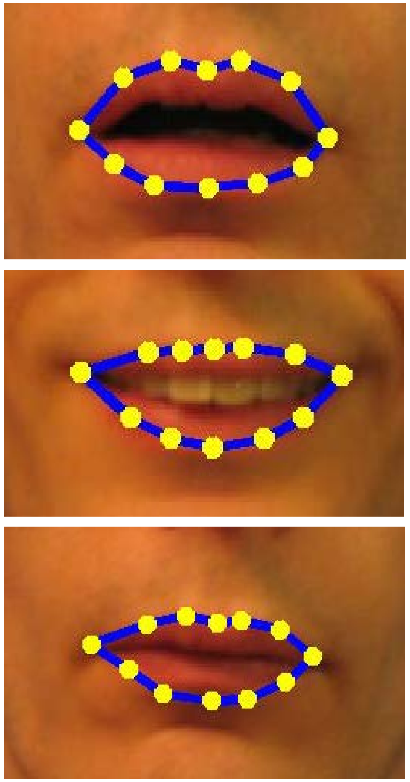
- Assume each shape can be represented by the linear combination of basis shapes

$$\mathbf{\Phi} = (\phi_1 | \phi_2 | \ldots | \phi_t)$$

such that $\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{\Phi}\mathbf{b}$

for mean shape $\bar{\mathbf{x}} = \frac{1}{s}\sum_{i=1}^{s} \mathbf{x}_i$

and some parameters $\mathbf{b}$

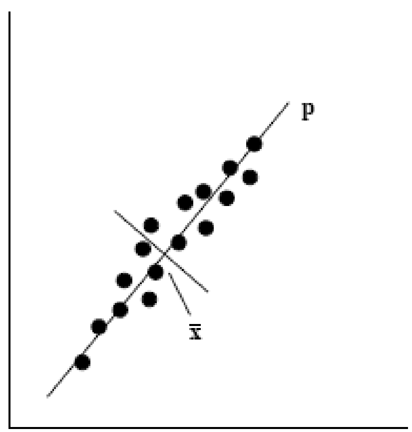


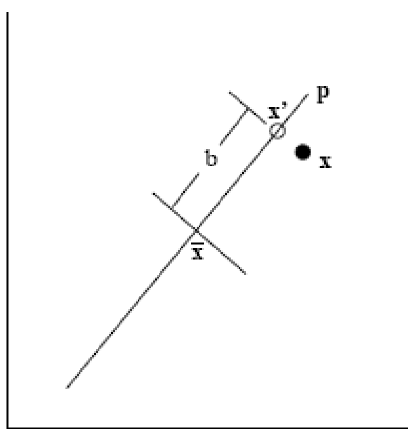
# Active Shape Models of Cootes et al.

- Basis shapes can be found as the main modes of variation of in the training data.

2D Example:

(each point can be thought as a shape in N-Dim space)




Principle Component Analysis (PCA):

Covariance matrix $$\mathbf{S} = \frac{1}{s-1}\sum_{i=1}^{s}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

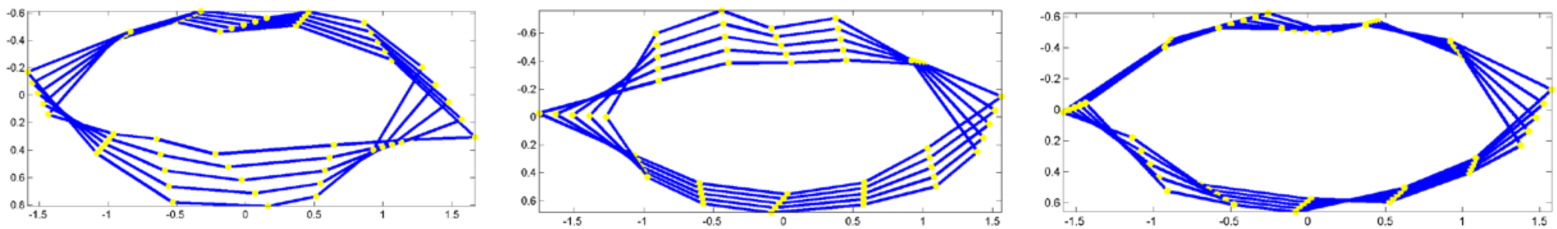
Eigenvectors $\mathbf{\Phi} = (\phi_1|\phi_2|\ldots|\phi_t)$ eigenvalues $\lambda_1, ..., \lambda_t$

# Active Shape Models of Cootes et al.

- Back-project from shape-space $\mathbf{b}$ to image space $\mathbf{x} = \bar{\mathbf{x}} + \Phi\mathbf{b}$

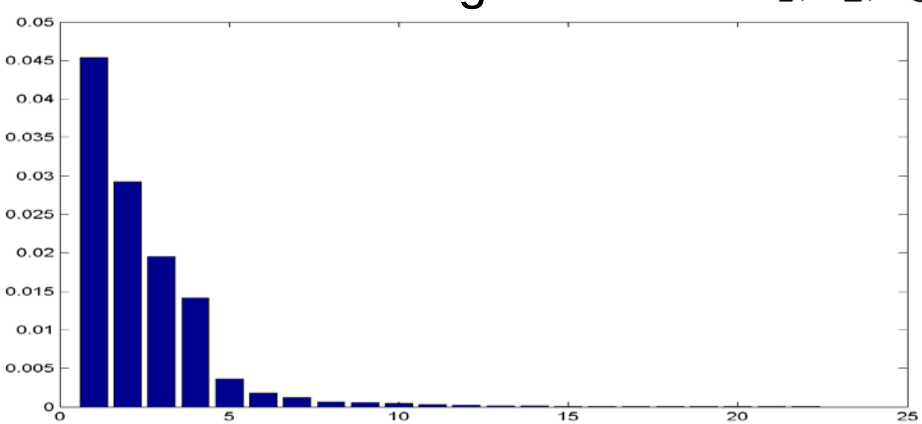
➡ Three main modes of lips-shape variation:

$\mathrm{b} = (\mu\lambda_1, 0, 0, ...)^\top$ $\mathrm{b} = (0, \mu\lambda_2, 0, 0, ...)^\top$ $\mathrm{b} = (0, 0, \mu\lambda_3, 0, 0, ...)^\top$




$\mu = -3, 1.5, 0, 1.5, 3$

Distribution of eigenvalues: $\lambda_1, \lambda_2, \lambda_3, ...$




A small fraction of basis shapes (eigenvecors) accounts for the most of shape variation (=> landmarks are redundant)

# Active Shape Models of Cootes et al.

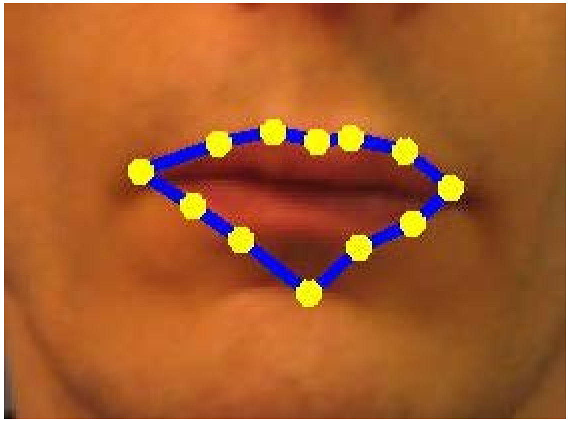
- $\mathbf{\Phi}$ is orthonormal basis, therefore $\mathbf{\Phi}^{-1} = \mathbf{\Phi}^\top$

  ⇒ Given estimate of $\mathbf{x}$ we can recover shape parameters $\mathbf{b}$

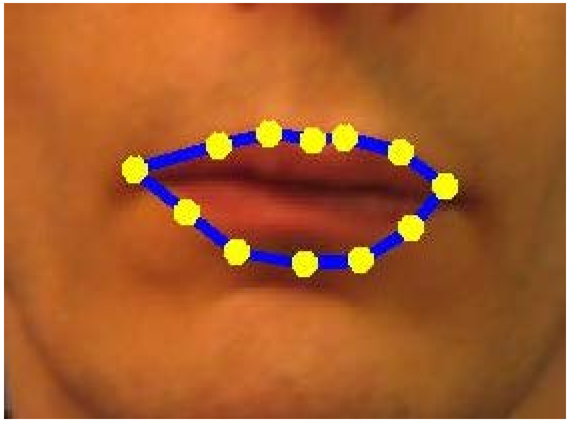
$$\mathbf{b} = \mathbf{\Phi}^\top(\mathbf{x} - \bar{\mathbf{x}})$$

- Projection onto the shape-space serves as a *regularization*

$$\mathbf{x} \quad \Rightarrow \quad \mathbf{b} = \mathbf{\Phi}^\top(\mathbf{x} - \bar{\mathbf{x}}) \quad \Rightarrow \quad \mathbf{x}_{\text{reg}} = \bar{\mathbf{x}} + \mathbf{\Phi}\mathbf{b}$$



⇒



# Active Shape Models [Cootes et al.]

- Constrains shape deformation in PCA-projected space

Example: face alignment

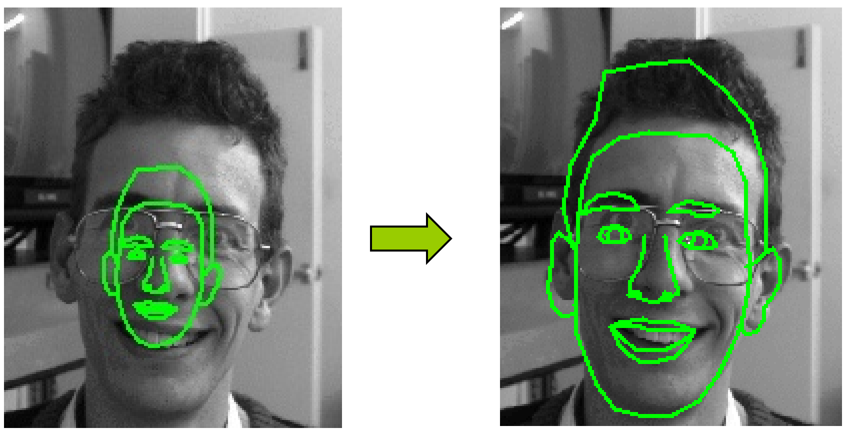
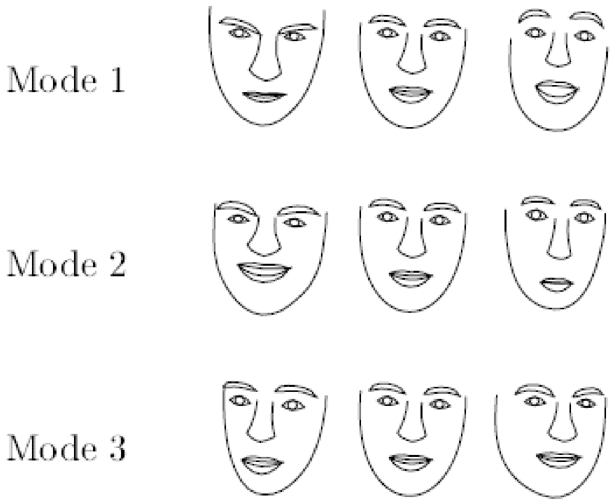


Illustration of face shape space




[Cootes, Taylor, Cooper, and Graham, CVIU 1995]

# Appearance-based methods: shape tracking

[Baumberg and Hogg, ECCV 1994]

# Shape priors & Tracking:

Pros:

+ more accurate tracking using specific shape and motion models
+ Simultaneous tracking and motion recognition with discrete state dynamical models

Cons:

- Local minima is still an issue
- Re-initialization is still an issue

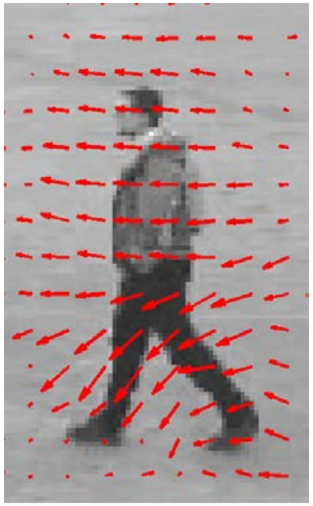
# Shape and Appearance vs. Motion

- 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

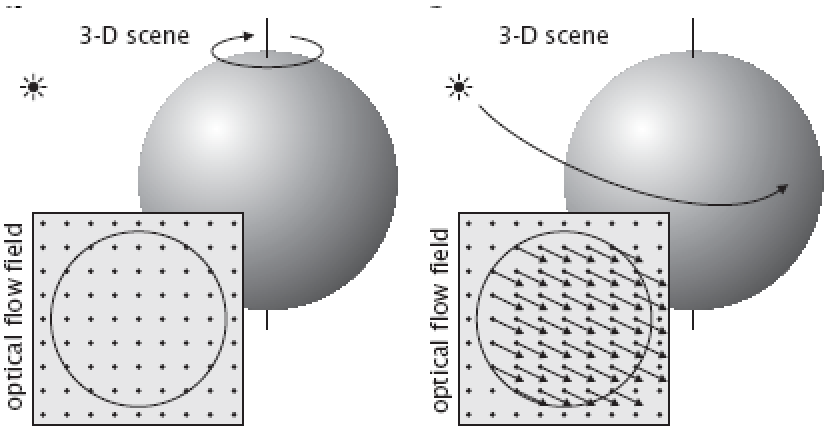
➡ Estimate pixel-wise correspondence between frames = Optical Flow

- 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



# Generic Optical Flow

- Brightness Change Constraint Equation (BCCE)

$$(\nabla I)^\top \mathbf{v} + I_t = 0$$

$\mathbf{v} = (v_x, v_y)^\top$ Optical flow

$\nabla I = (I_x, I_y)^\top$ Image gradient

One equation, two unknowns => cannot be solved directly

➡ Integrate several measurements in the local neighborhood and obtain a *Least Squares Solution* [Lucas & Kanade 1981]

$$< \nabla I (\nabla I)^\top > \mathbf{v} = - < \nabla I I_t >$$

$$\begin{pmatrix} < I_x^2 > & < I_x I_y > \\ < I_x I_y > & < I_y^2 > \end{pmatrix} \mathbf{v} = - \begin{pmatrix} < I_x I_t > \\ < I_y I_t > \end{pmatrix}$$

Second-moment matrix, the same one used to compute Harris interest points!

$< \cdot >$ Denotes integration over a spatial (or spatio-temporal) neighborhood of a point

# Generic Optical Flow

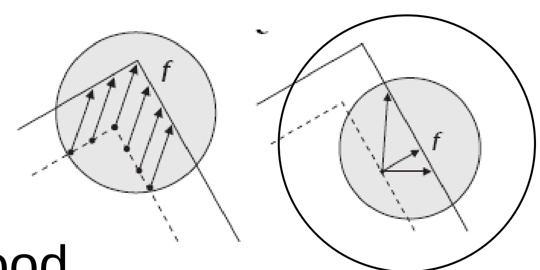
- The solution of $<\nabla I (\nabla I)^\top> \mathbf{v} = -<\nabla I I_t>$ assumes
  1. Brightness change constraint holds in $<\cdot>$
  2. Sufficient variation of image gradient in $<\cdot>$
  3. Approximately constant motion in $<\cdot>$

  Motion estimation becomes *inaccurate* if any of assumptions 1-3 is violated.

- Solutions:

  (2) Insufficient gradient variation known as *aperture problem*

  ➡ Increase integration neighborhood




  (3) Non-constant motion in $<\cdot>$

  ➡ Use more sophisticated motion model

# Parameterized Optical Flow

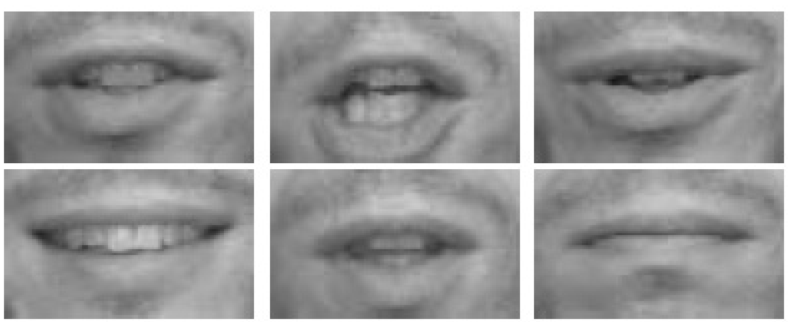
- Another extension of the constant motion model is to compute PCA basis flow fields from training examples

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

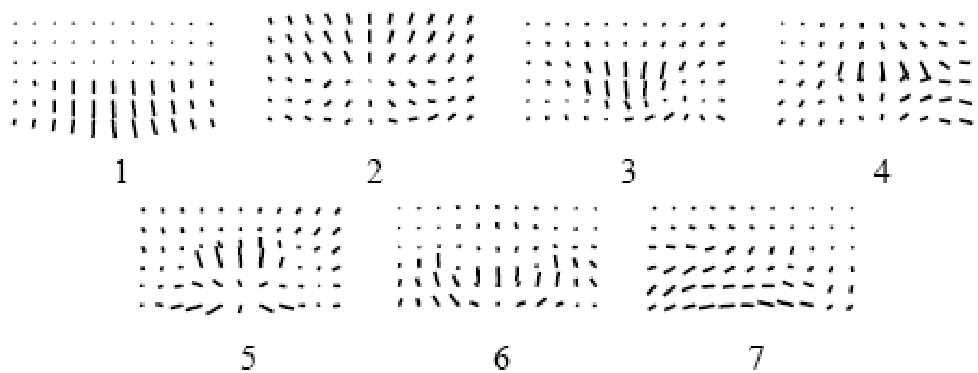
  $$\mathbf{w} = (v_x^1, v_y^1, v_x^2, v_y^2, ..., v_x^n, v_y^n)^\top$$

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

Training samples



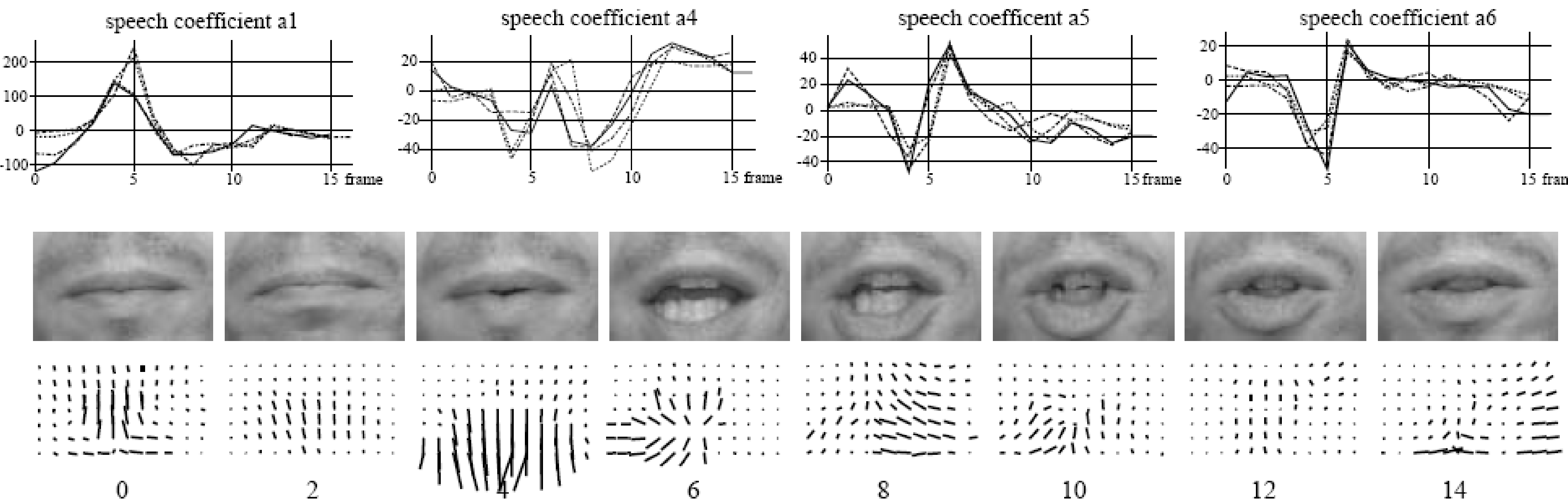
PCA flow bases




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

# Parameterized Optical Flow

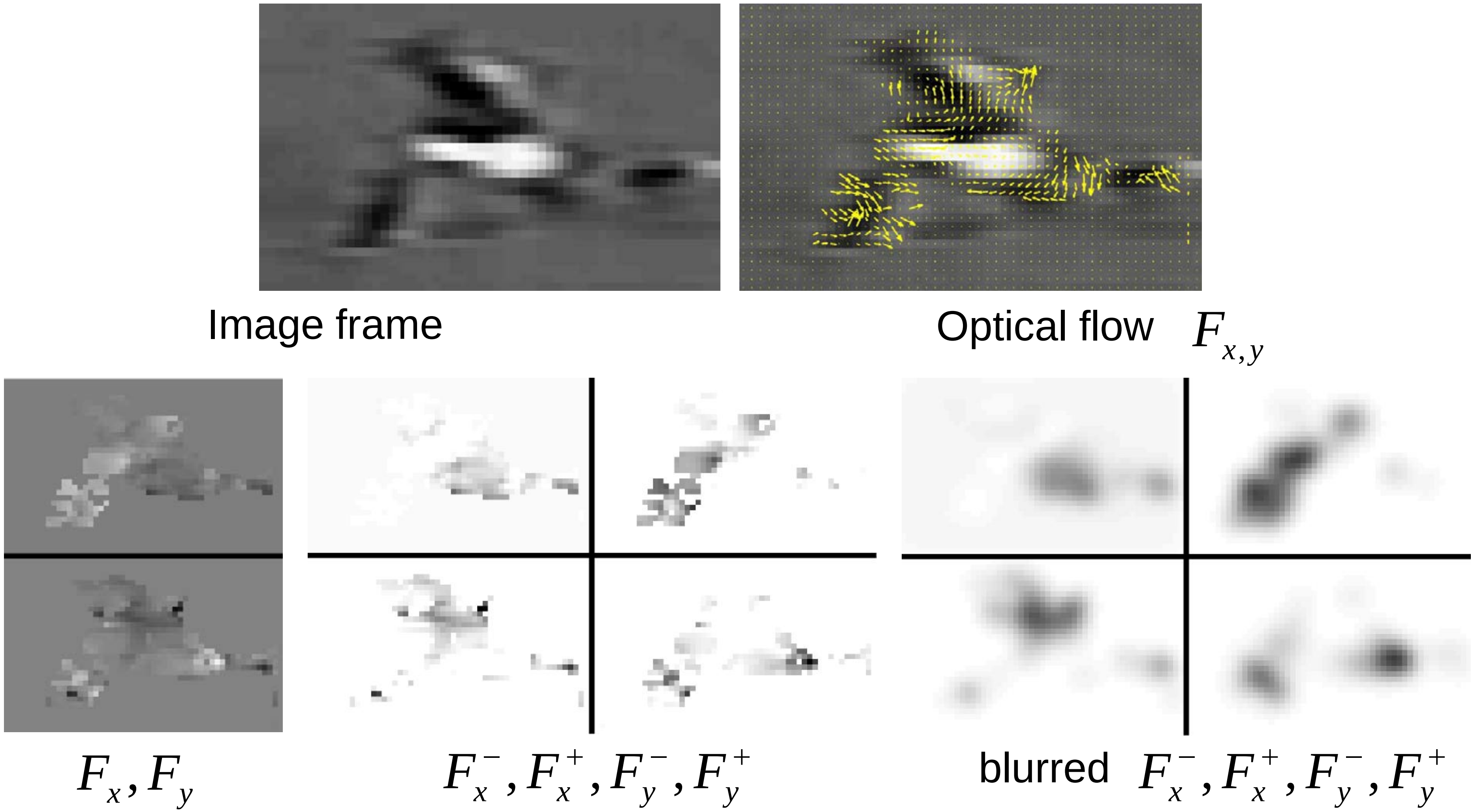
- Estimated coefficients of PCA flow bases can be used as action descriptors




Frame numbers

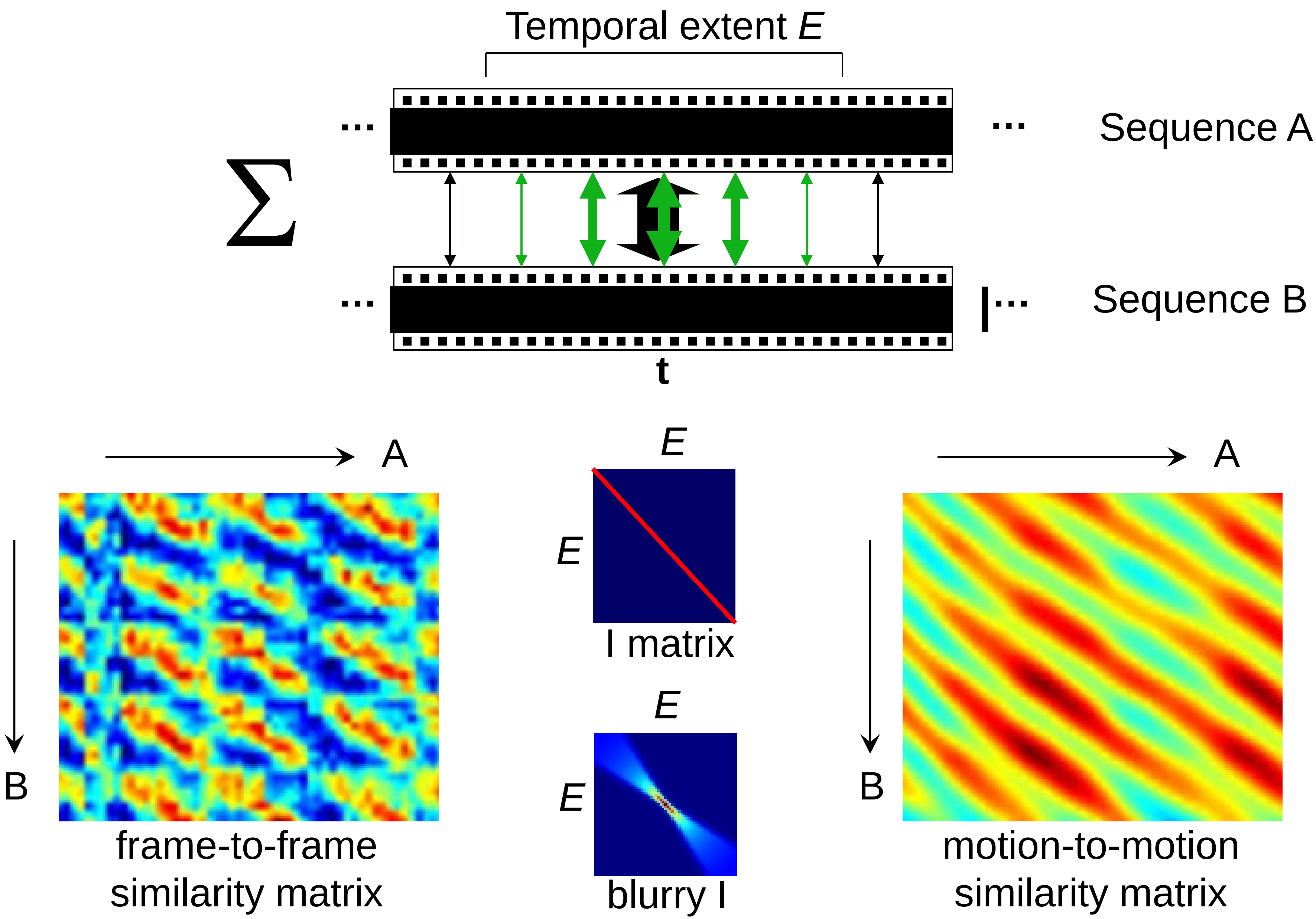
➡ Optical flow seems to be an interesting descriptor for motion/action recognition

# Spatial Motion Descriptor




[Efros, Berg, Mori and Malik, ICCV 2003]

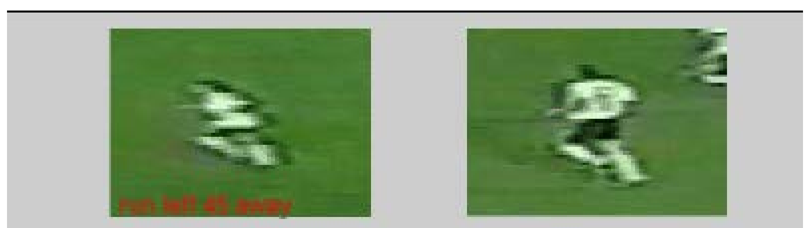
# Spatio-Temporal Motion Descriptor



# Football Actions: matching

Input Sequence

Matched Frames



input matched

Goal:
Interpret complex dynamic scenes


Common methods:

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

Common problems:

- Complex & changing BG
- Changing appearance

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

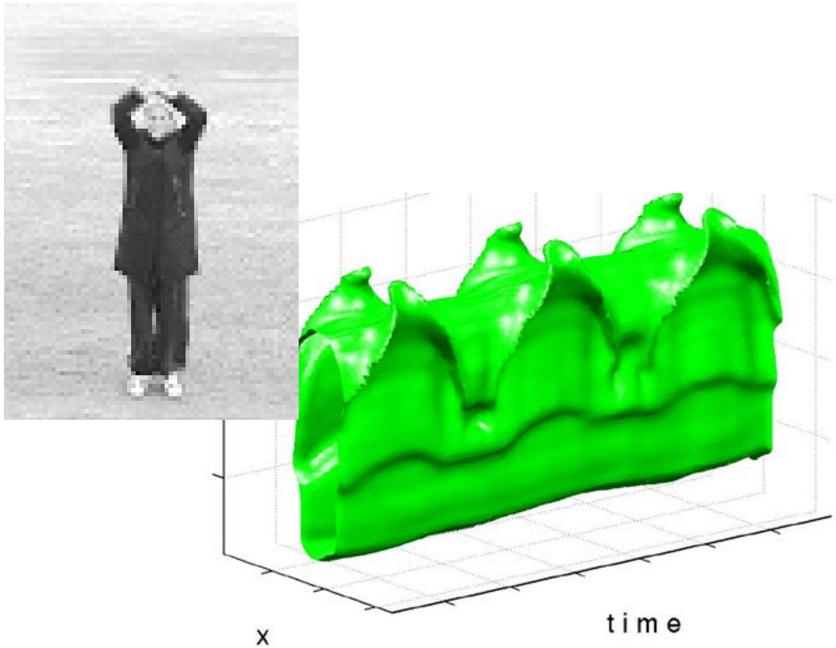
# Space-time

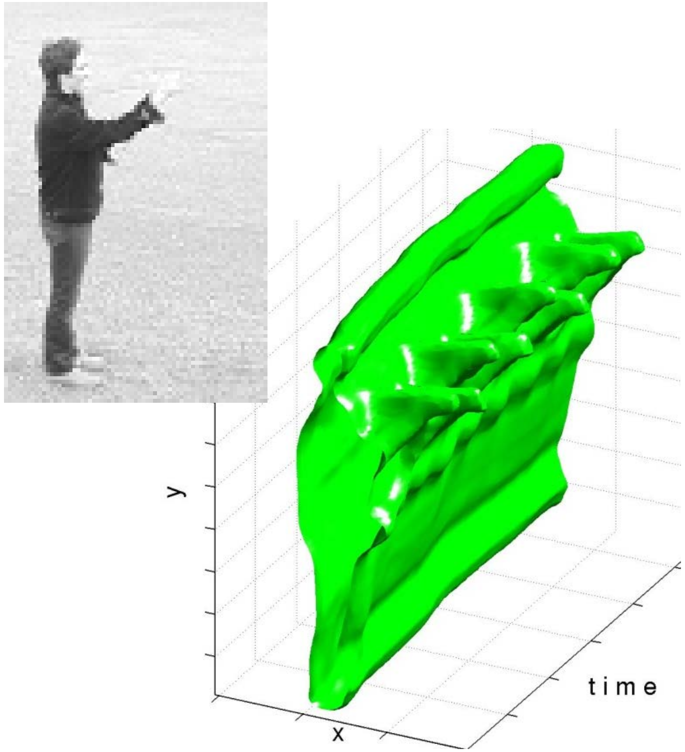
No global assumptions $\Rightarrow$
Consider local spatio-temporal neighborhoods

hand waving

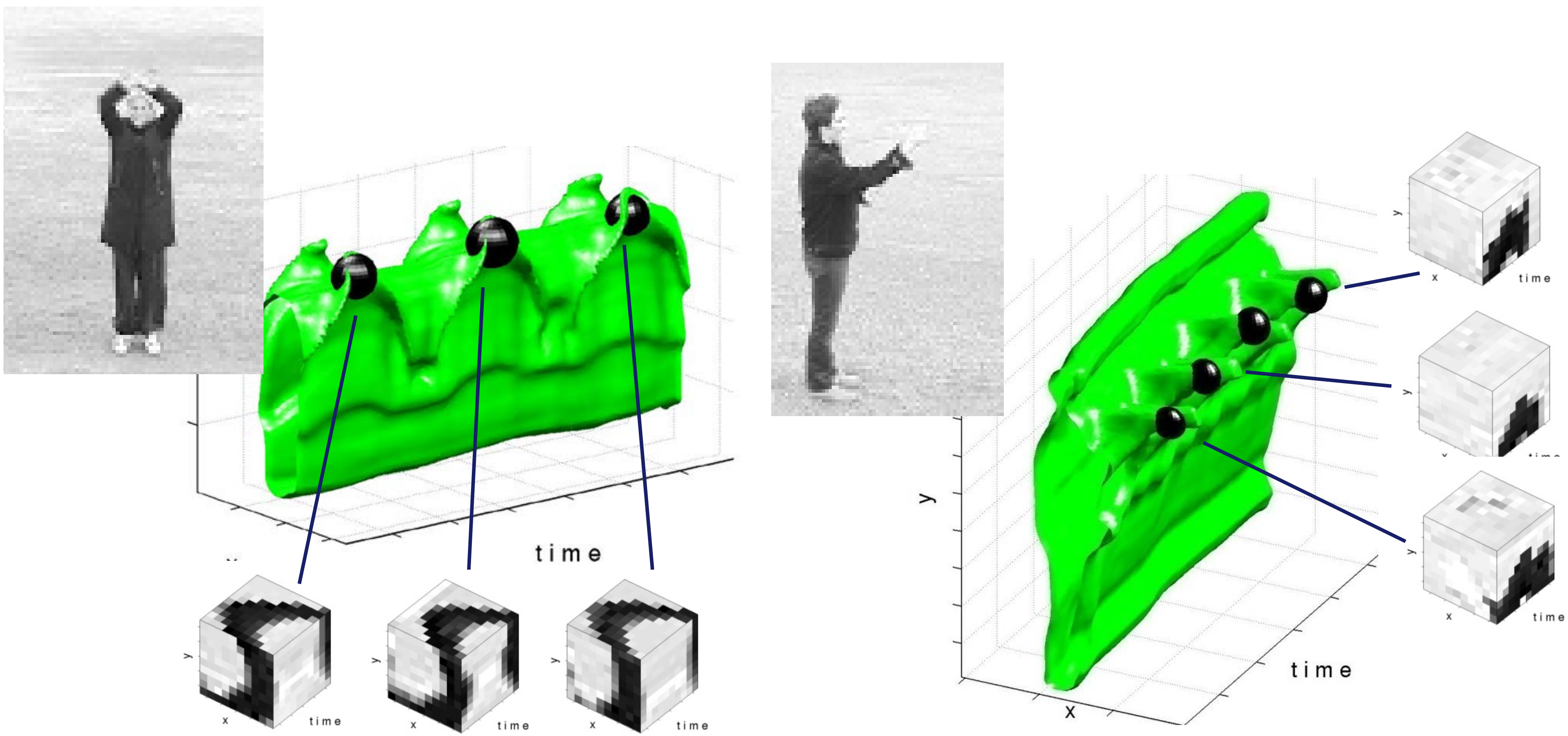
boxing

# Actions == Space-time objects?

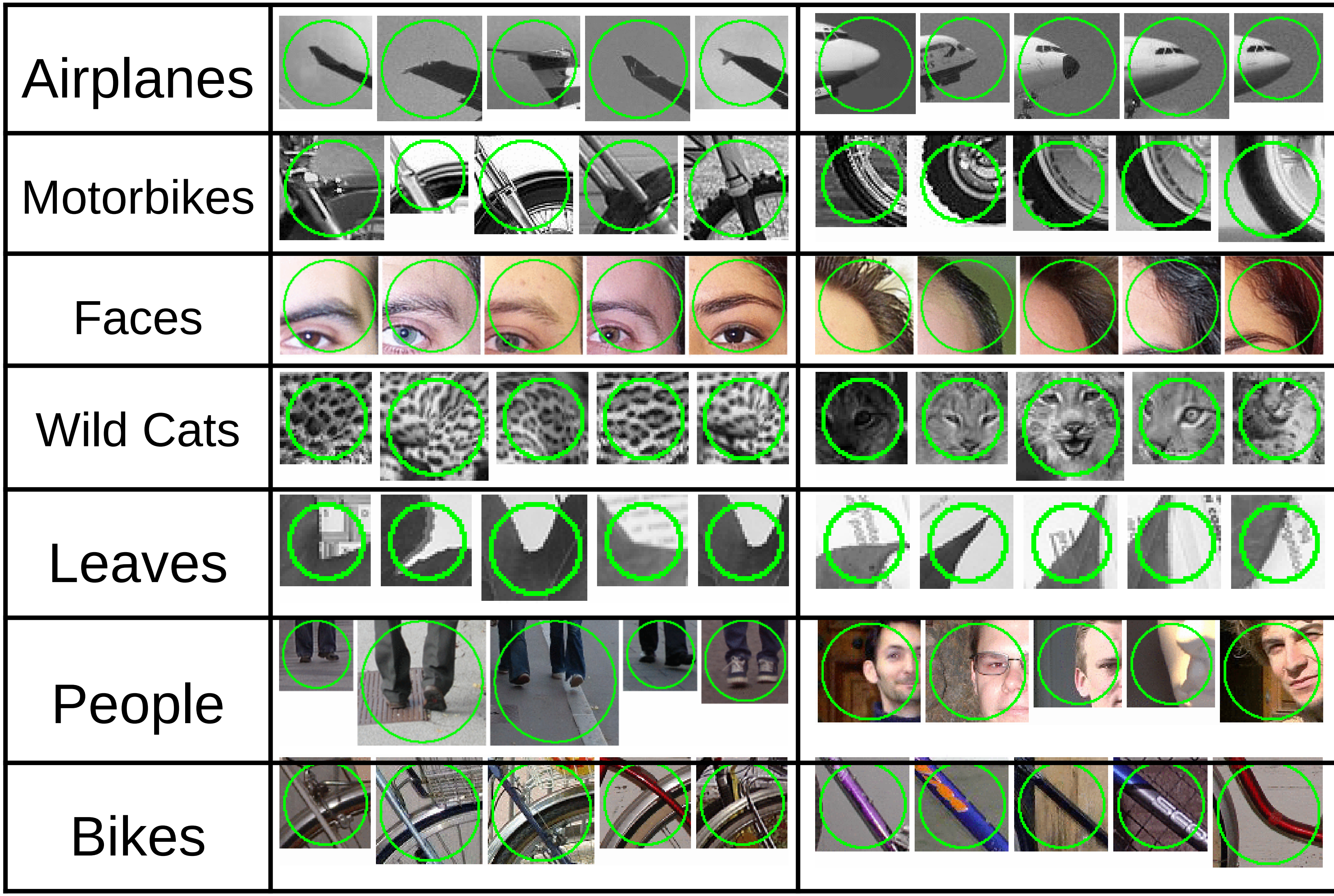


# Space-time local features



# Local approach: Bag of Visual Words



# 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]

# Space-Time Interest Points: Detection

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

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

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

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

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

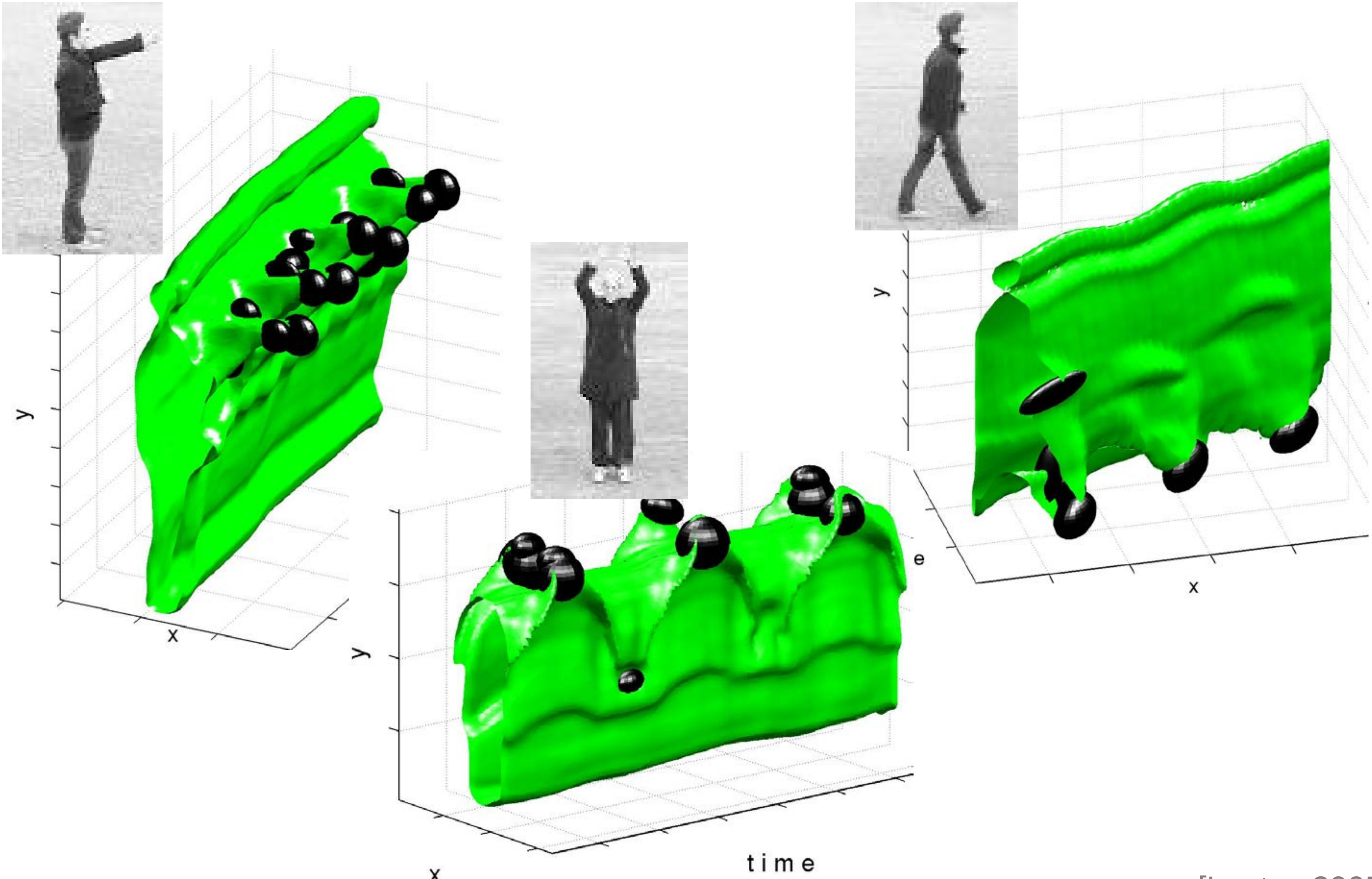
Large eigenvalues of μ can be detected by the local maxima of H over (x,y,t):

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

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

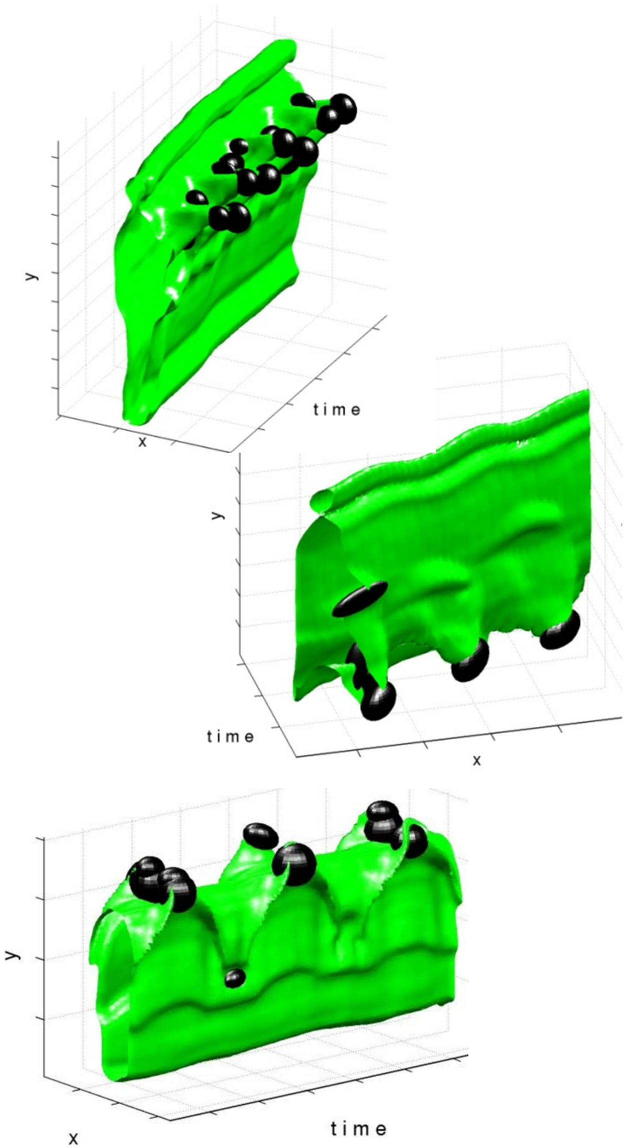
[Laptev 2005]

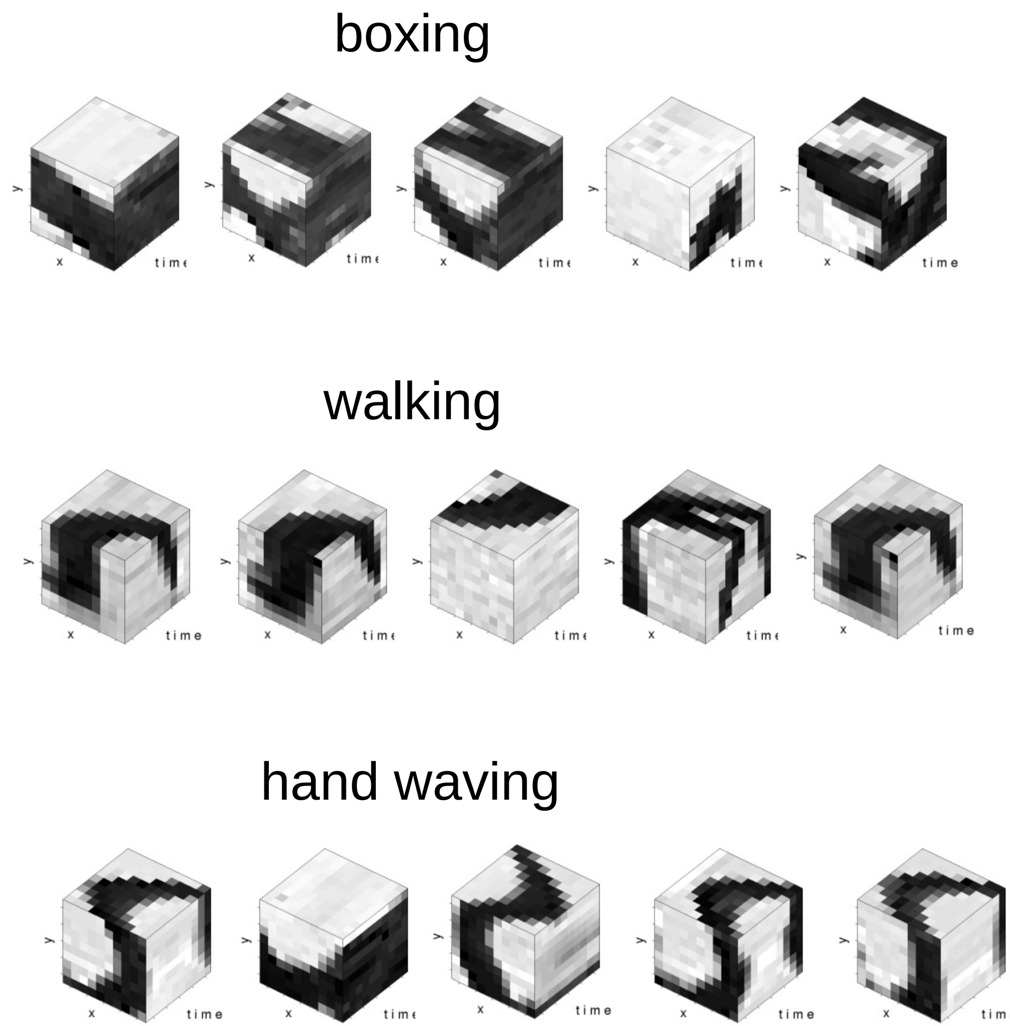
# Local features for human actions




[Laptev 2005]

# Local features for human actions

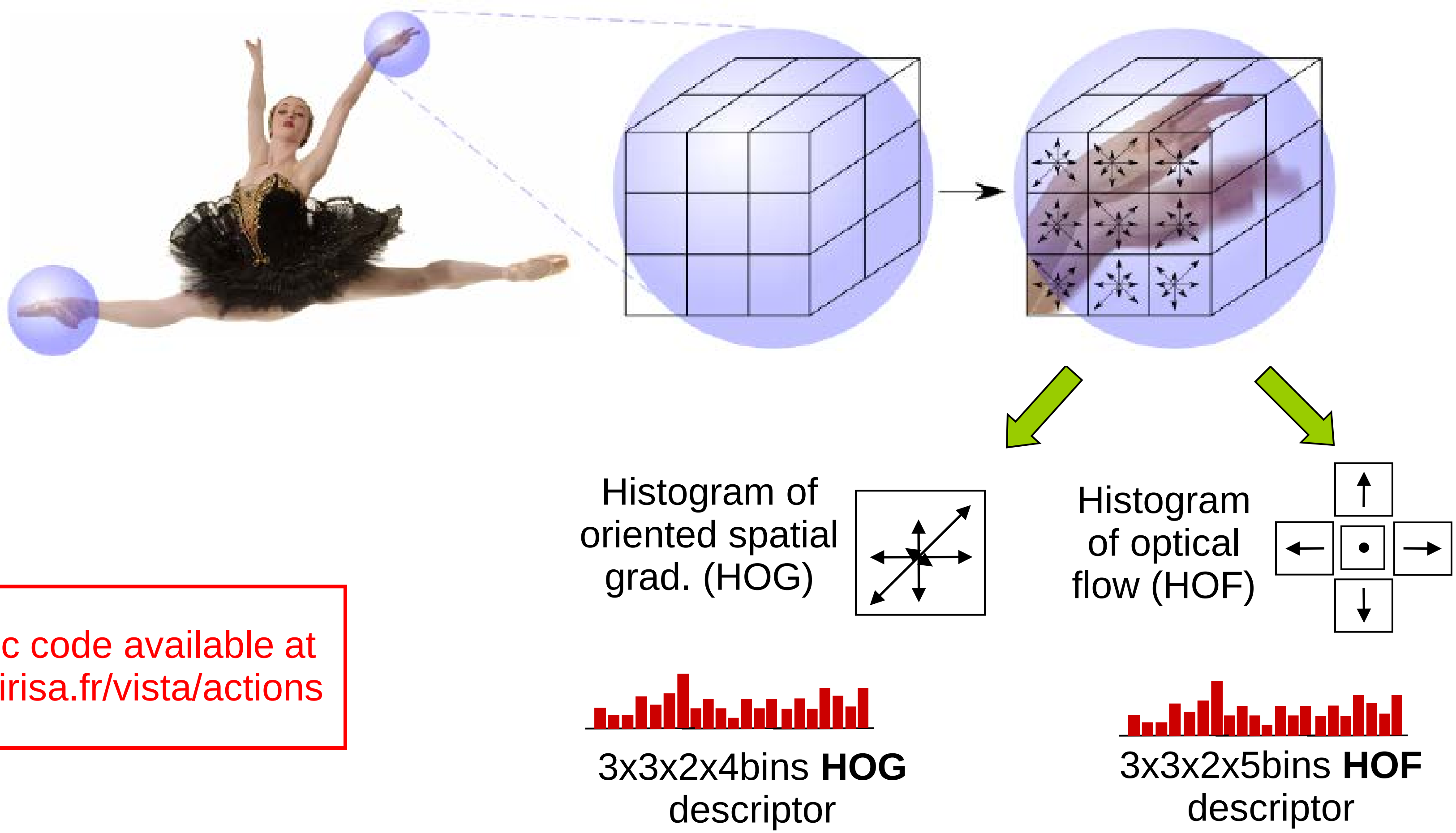



[Laptev 2005]

# Local space-time descriptor: HOG/HOF

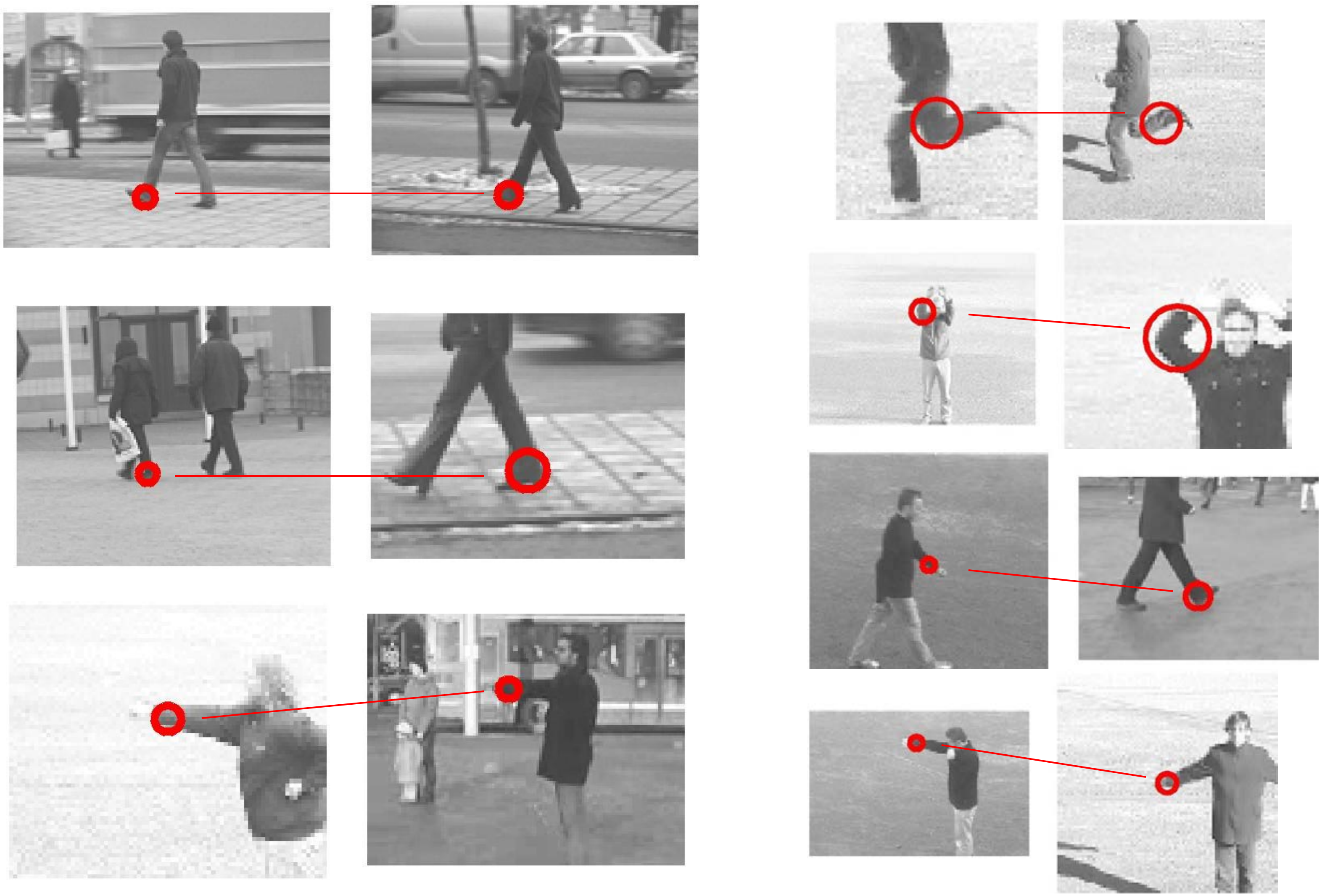
Multi-scale space-time patches




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

# Local Space-time features: Matching

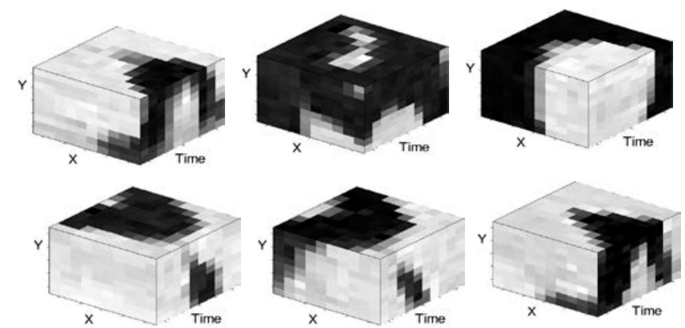
- Find similar events in pairs of video sequences



# Bag-of-Features action recognition

Extraction of Local features

space-time patches

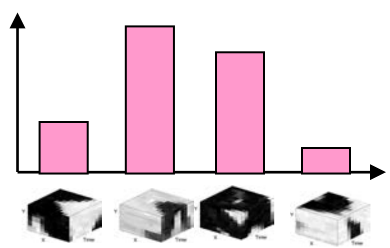



Feature description

K-means clustering (k=4000)

Feature quantization

Occurrence histogram of visual words



Non-linear SVM with $\chi^2$ kernel

[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”

# 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:**

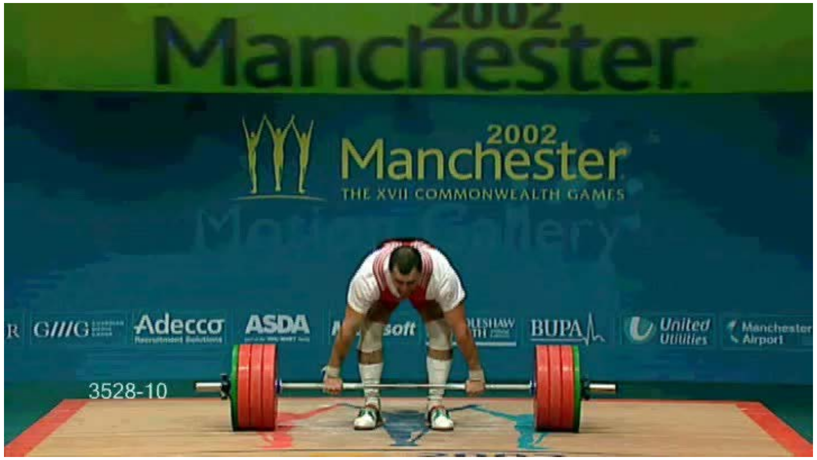
- HoG/HoF [Laptev et al. 2008]
- Cuboids [Dollar et al. 2005]
- HoG3D [Kläser et al. 2008]
- Extended SURF [Willems'et al. 2008]

**Three human actions datasets:**

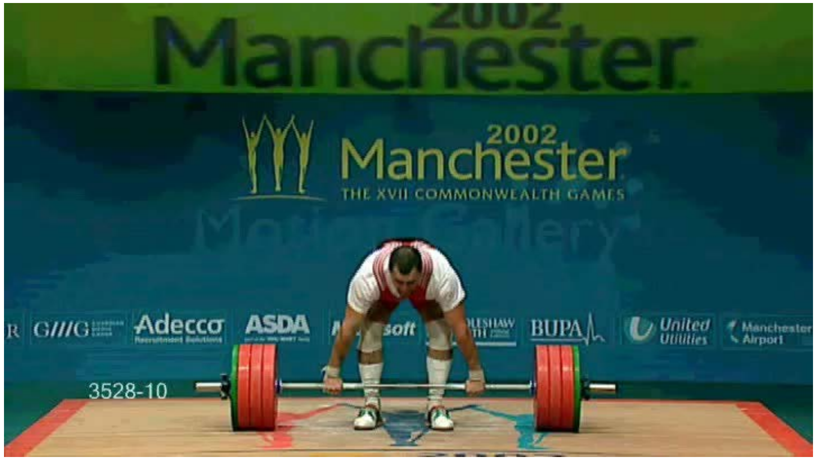
- KTH actions [Schuldt et al. 2004]
- UCF Sports [Rodriguez et al. 2008]
- Hollywood 2 [Marszałek et al. 2009]

# Space-time feature detectors

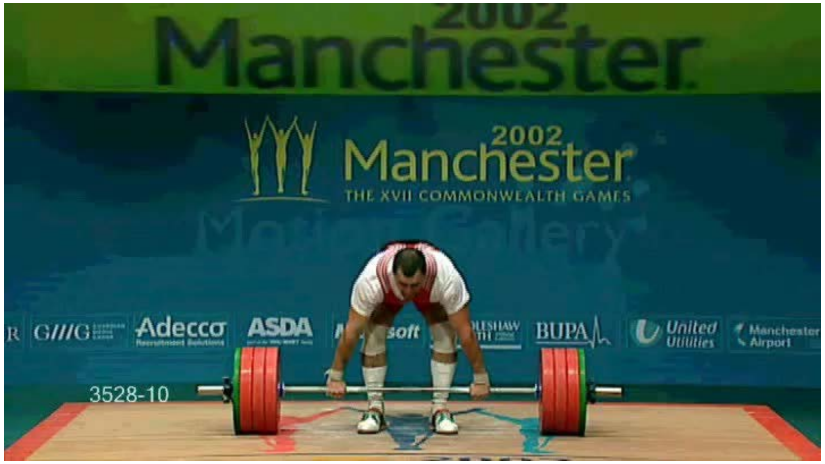
Harris3D



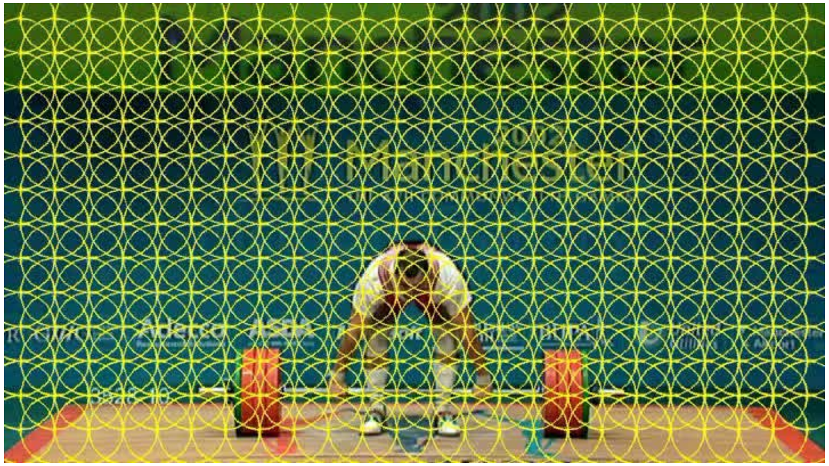

Hessian



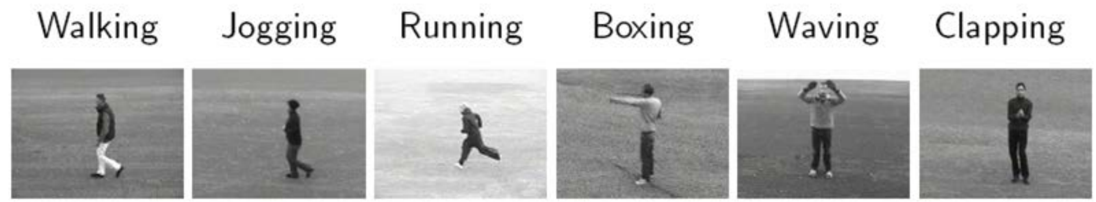

Cuboids




Dense



# Results on KTH Actions




6 action classes, 4 scenarios, staged

Detectors

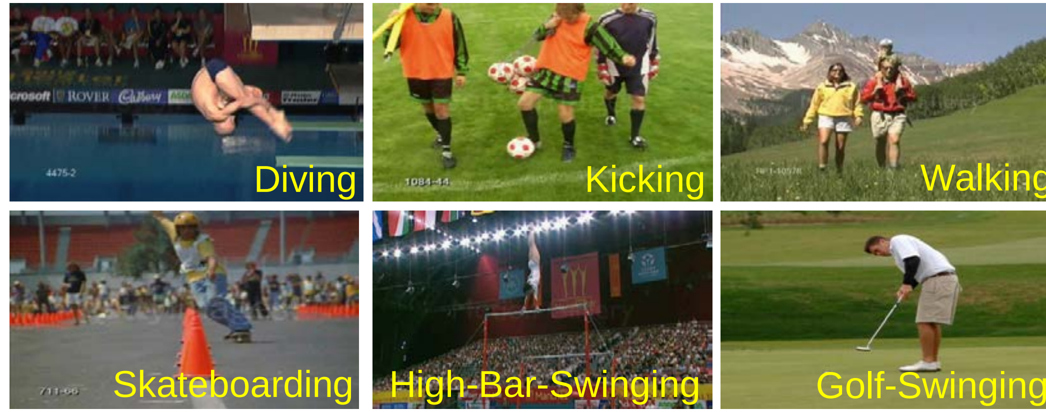
| Descriptors | Harris3D | Cuboids | Hessian | Dense |
|---|---|---|---|---|
| **HOG3D** | 89.0% | 90.0% | 84.6% | 85.3% |
| **HOG/HOF** | 91.8% | 88.7% | 88.7% | 86.1% |
| **HOG** | 80.9% | 82.3% | 77.7% | 79.0% |
| **HOF** | **92.1%** | 88.2% | 88.6% | 88.0% |
| **Cuboids** | - | 89.1% | - | - |
|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

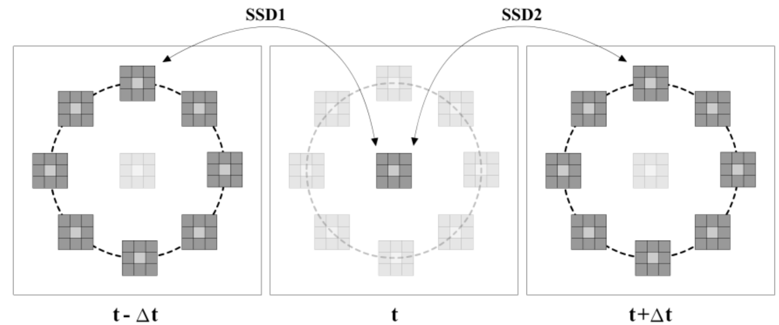
(Average precision scores)

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

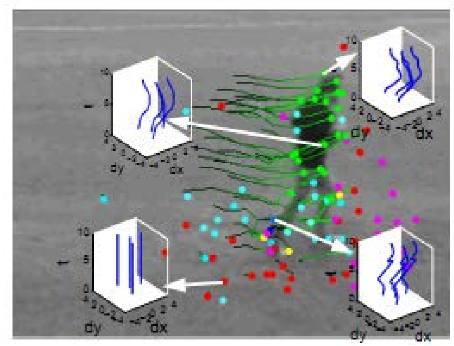
[Wang, Ullah, Kläser, Laptev, Schmid, 2009]

# Other recent local representations

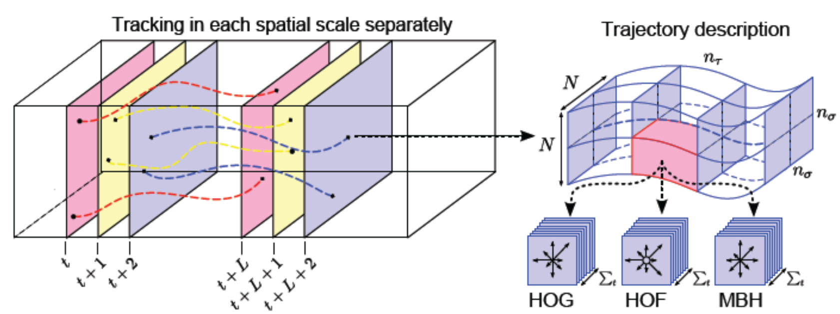
- Y. and L. Wolf, "Local Trinary Patterns for Human Action Recognition ", ICCV 2009




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

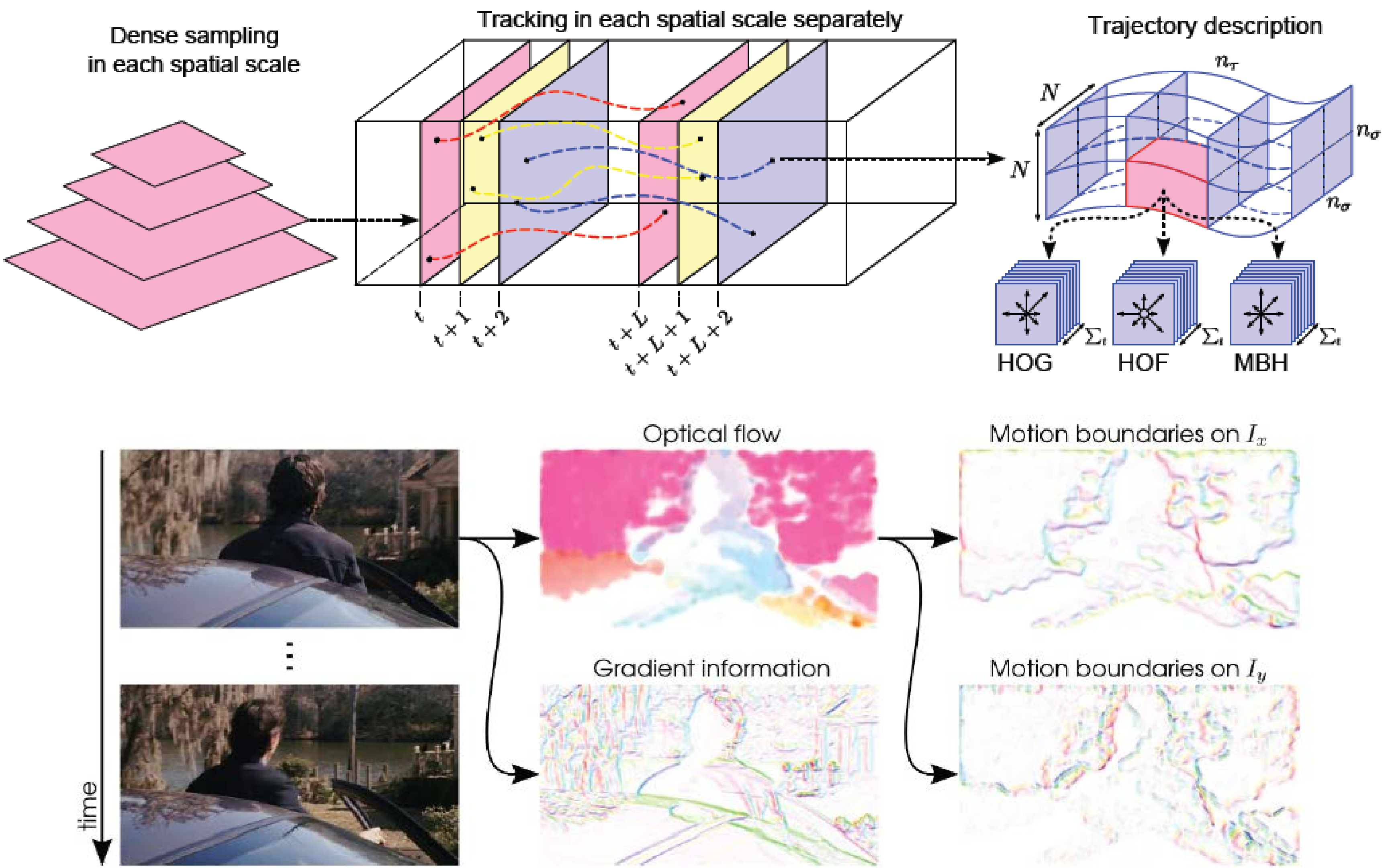


- H. Wang, A. Klaser, C. Schmid, C.-L. Liu, "Action Recognition by Dense Trajectories", CVPR 2011



# Dense trajectory descriptors

[Wang et al. CVPR'11]



# Dense trajectory descriptors

[Wang et al. CVPR'11]

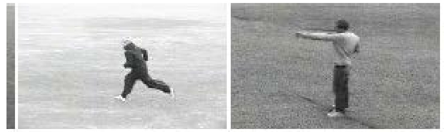
| | KTH | | YouTube | | Hollywood2 | | UCF sports | |
|---|---|---|---|---|---|---|---|---|
| | KLT | Dense trajectories | KLT | Dense trajectories | KLT | Dense trajectories | KLT | Dense trajectories |
| Trajectory | 88.4% | 90.2% | 58.2% | 67.2% | 46.2% | 47.7% | 72.8% | 75.2% |
| HOG | 84.0% | 86.5% | 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%** |

# Where to get training data?

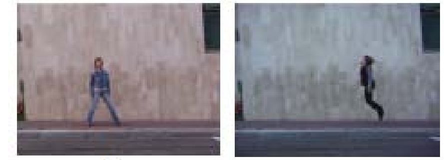
# Action recognition datasets

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



Running Boxing

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



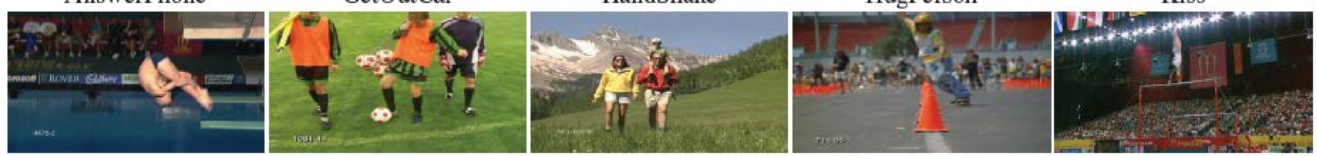
- UCF YouTube, 11 classes, 1168 samples, [Liu et al. 2009]

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]

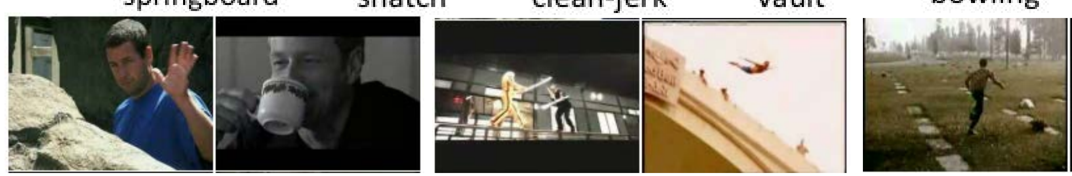


Diving Kicking Walking Skateboarding High-Bar-Swinging

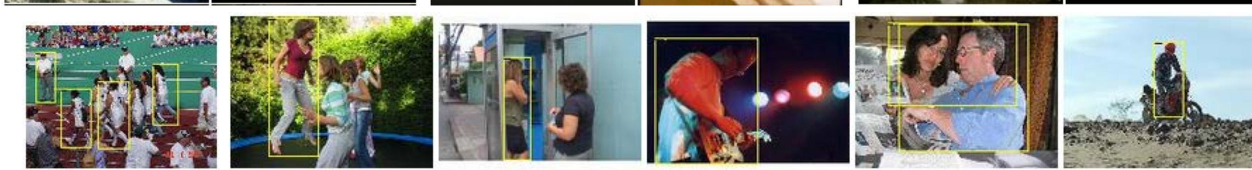
- Olympic Sports, 16 classes, 783 samples, [Niebles et al. 2010]

springboard snatch clean-jerk vault bowling tennis-serve

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

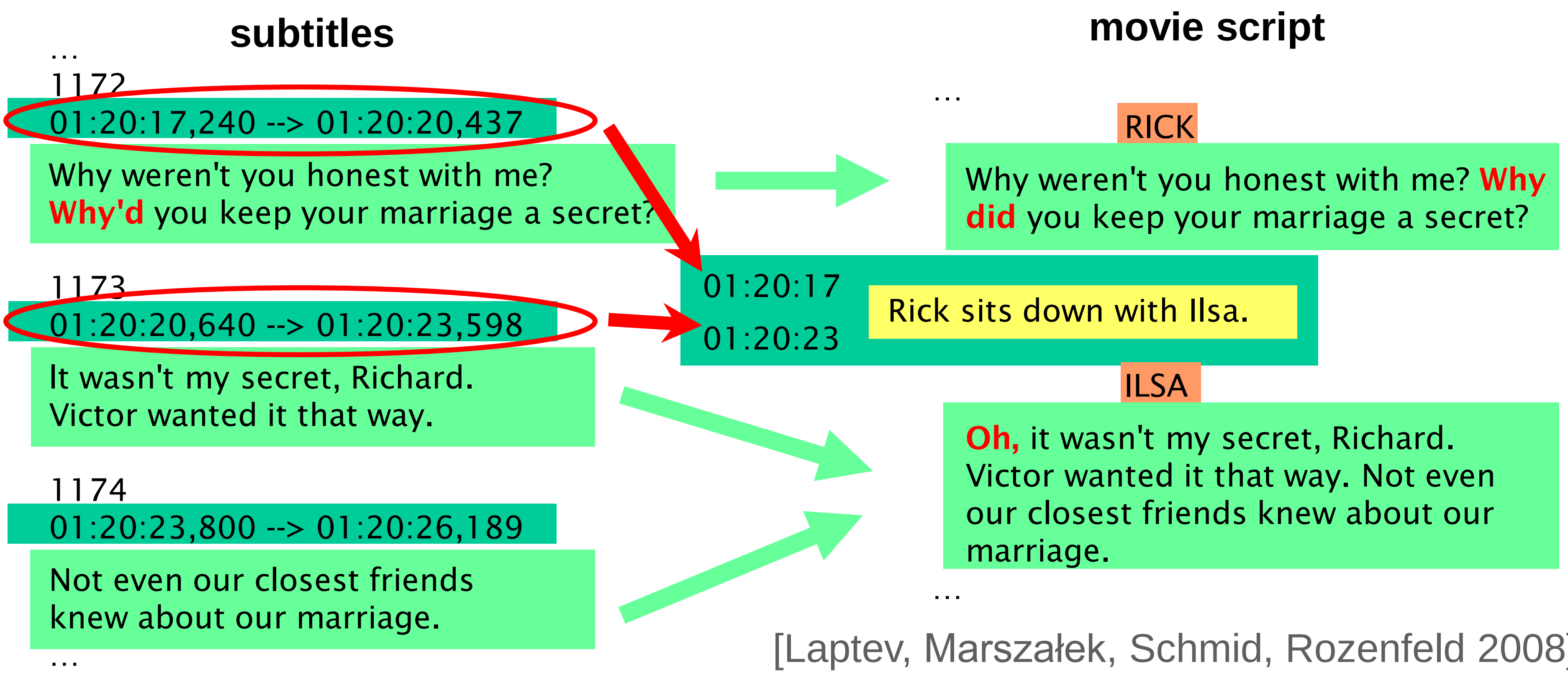


- PASCAL VOC 2011 Action Classification Challenge, 10 classes, 3375 image samples



# Script-based video annotation

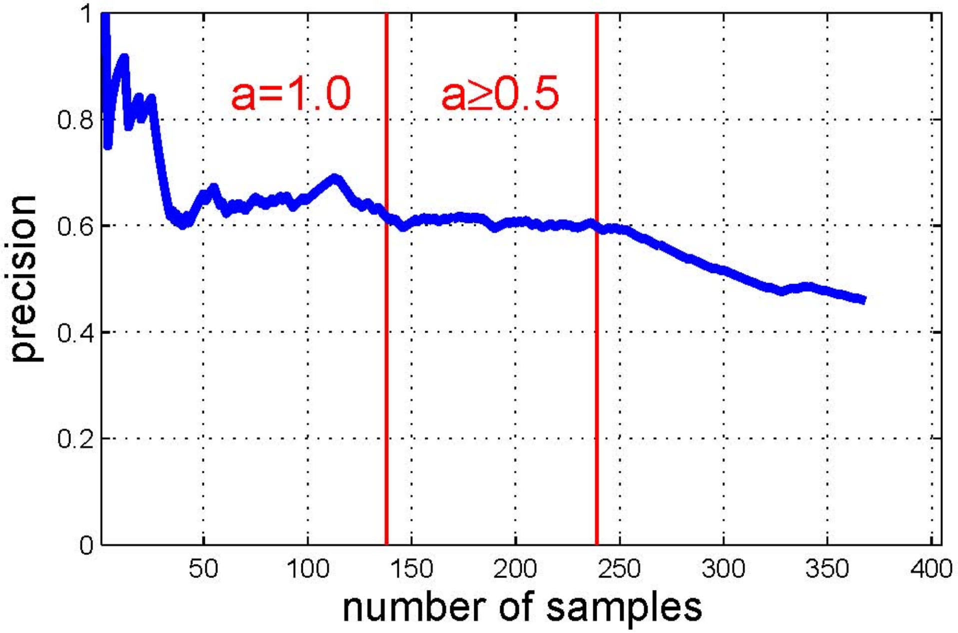
- Scripts available for >500 movies (no time synchronization)
  www.dailyscript.com, www.movie-page.com, www.weeklyscript.com …
- Subtitles (with time info.) are available for the most of movies
- Can transfer time to scripts by text alignment




[Laptev, Marszałek, Schmid, Rozenfeld 2008]

# 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”

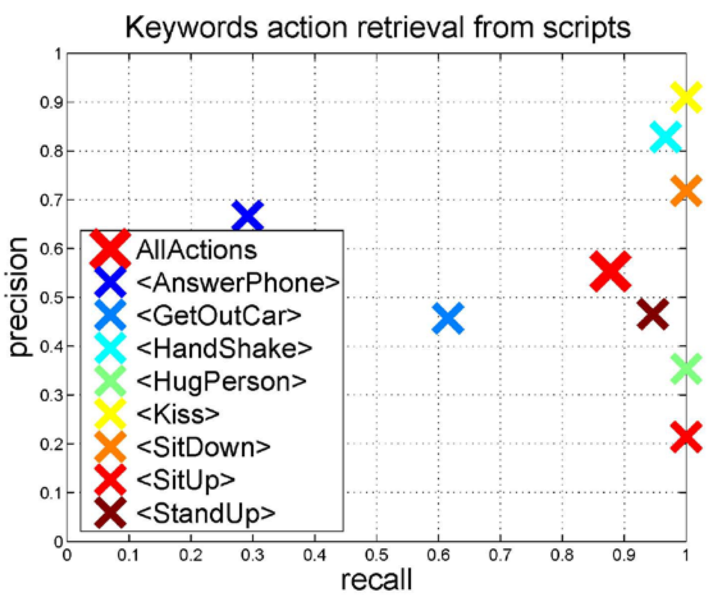
A black car pulls up, two army officers get out.

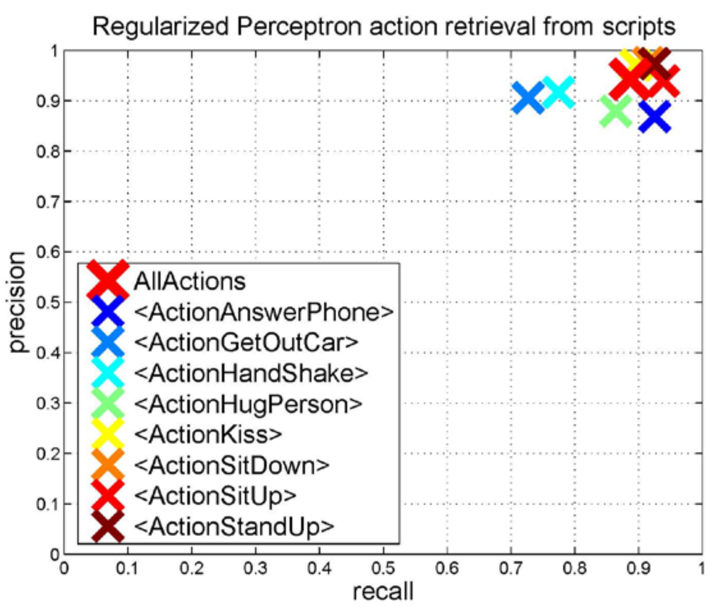
# Text-based action retrieval

- Large variation of action expressions in text:

| | |
|---|---|
| GetOutCar action: | *"… Will gets out of the Chevrolet. …"*<br>*"… 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 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*

# 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.

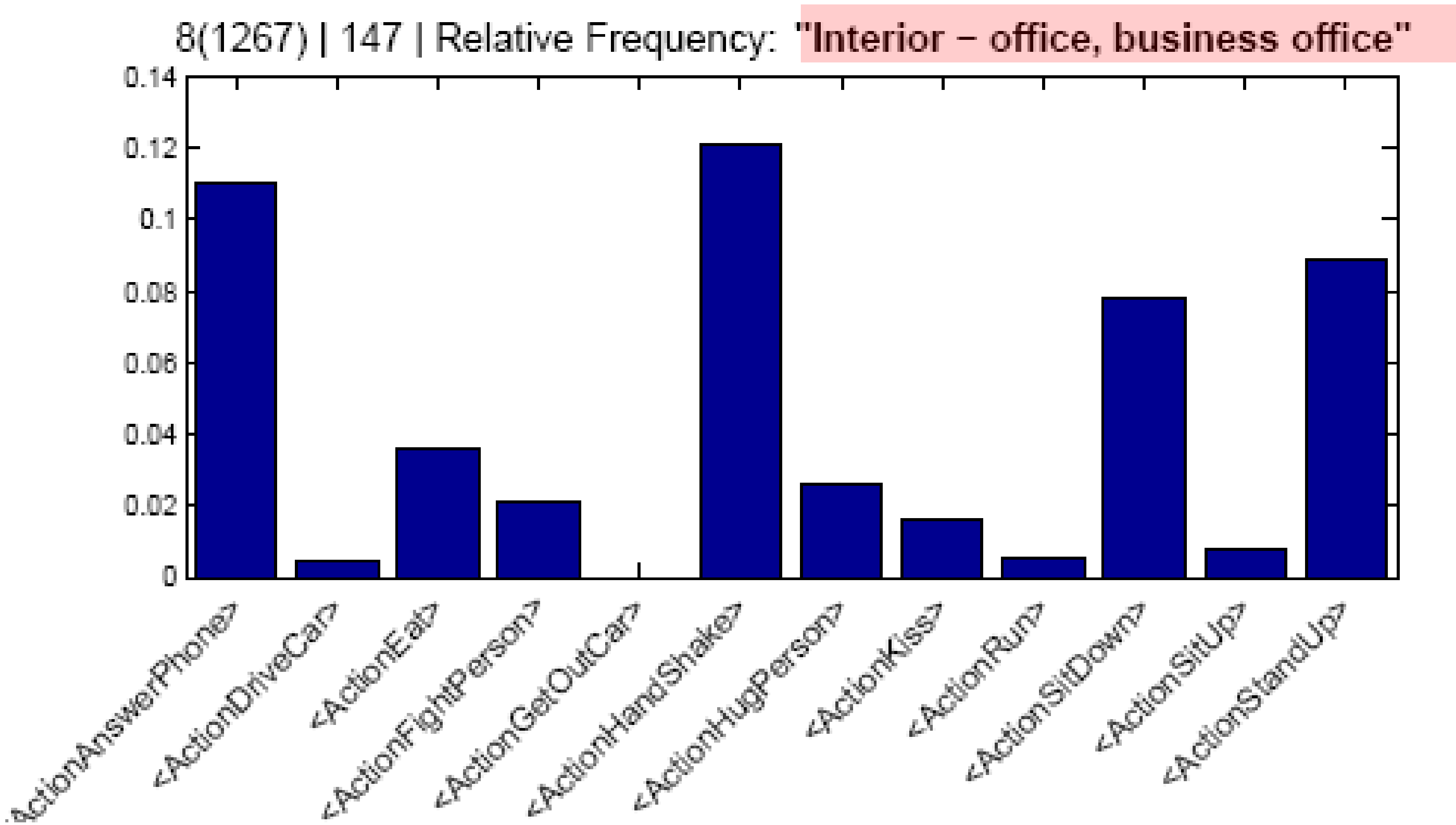
01:22:15
01:22:17

CARL

I think we lost them.

...

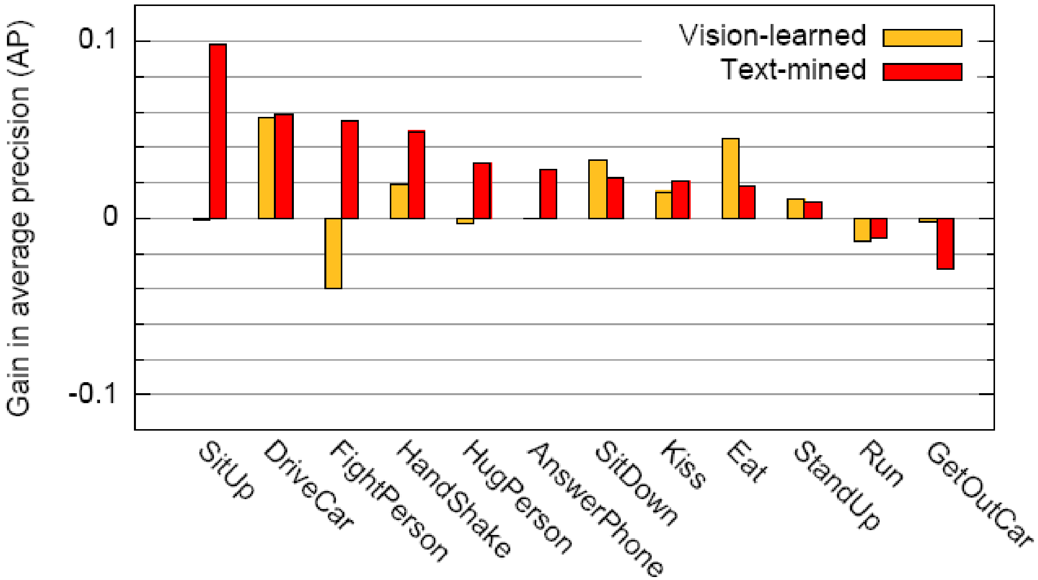
# Co-occurrence of actions and scenes in scripts



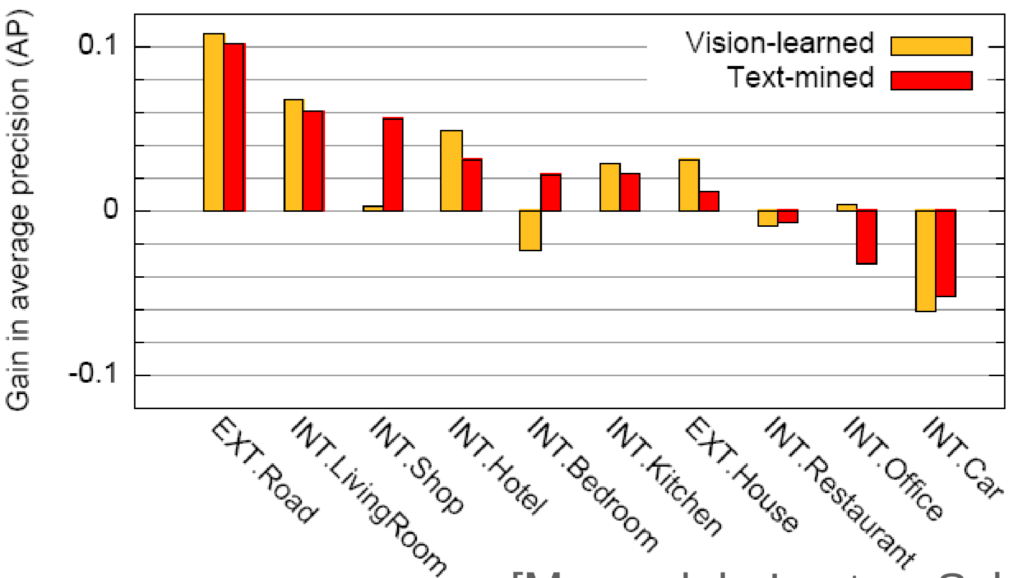

[Marszałek, Laptev, Schmid, 2009]

# Results: actions and scenes (jointly)

Actions in the context of Scenes

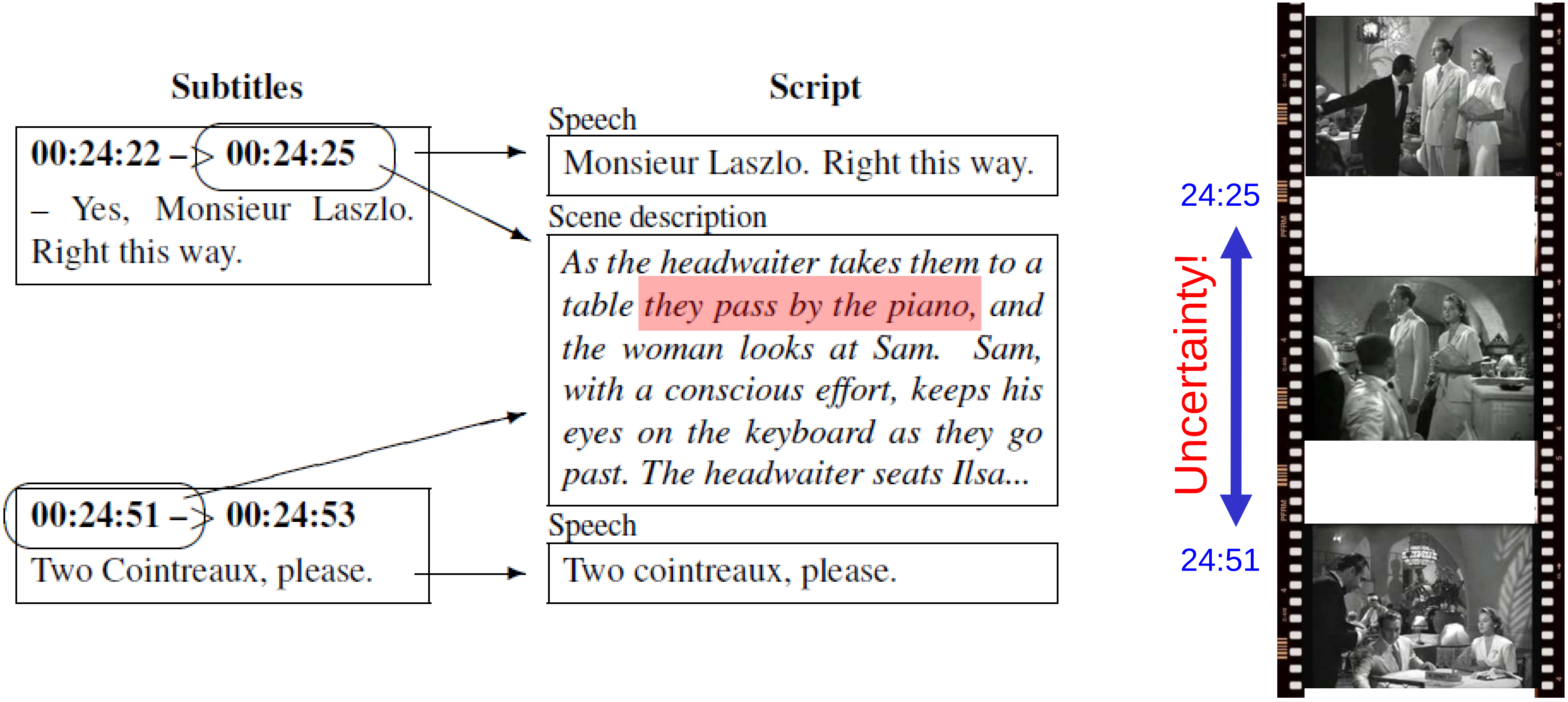



Scenes in the context of Actions




[Marszałek, Laptev, Schmid, 2009]

# Handling temporal uncertainty




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

# Handling temporal uncertainty

Input:

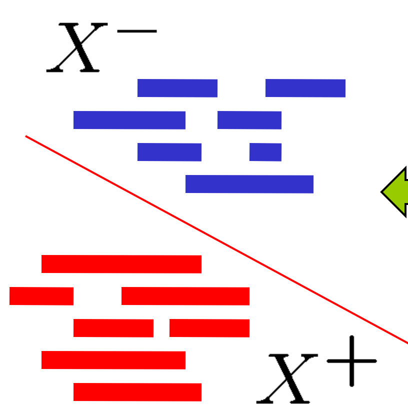
- Action type, e.g.
  Person Opens Door
- Videos + aligned scripts

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



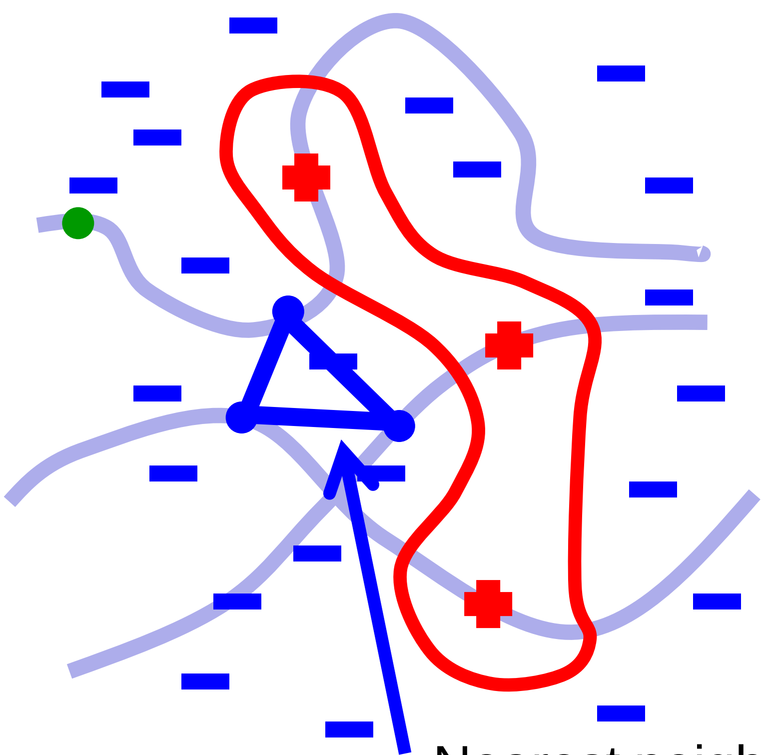

Output:

Sliding-window-style temporal action localization

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

# Discriminative action clustering

Feature space



Nearest neighbor solution: wrong!

Video space

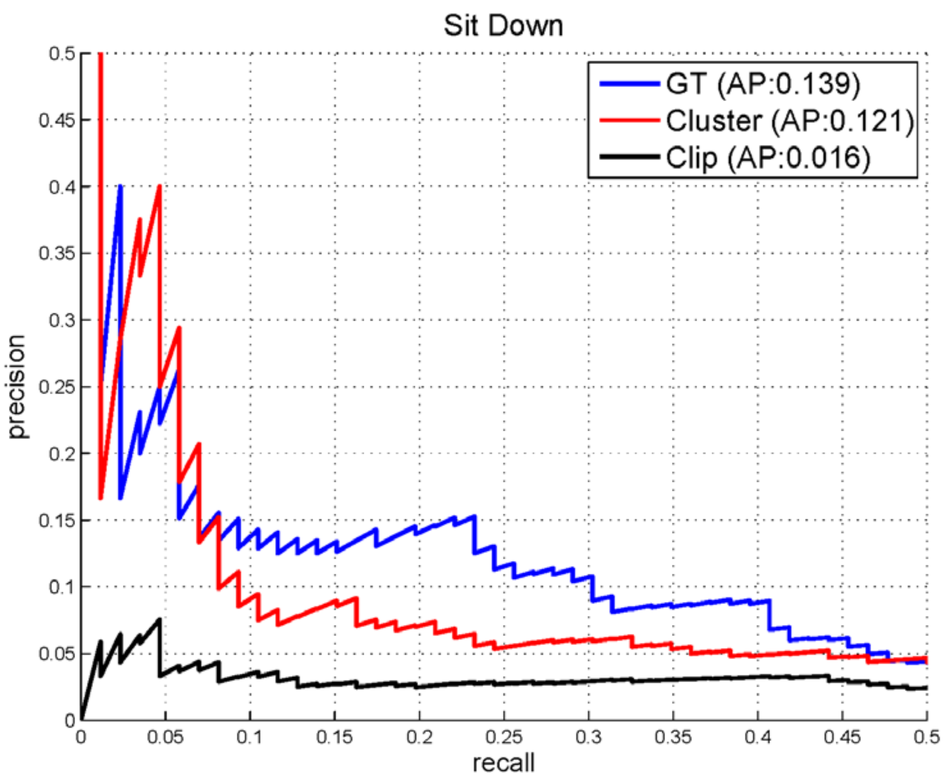
Negative samples

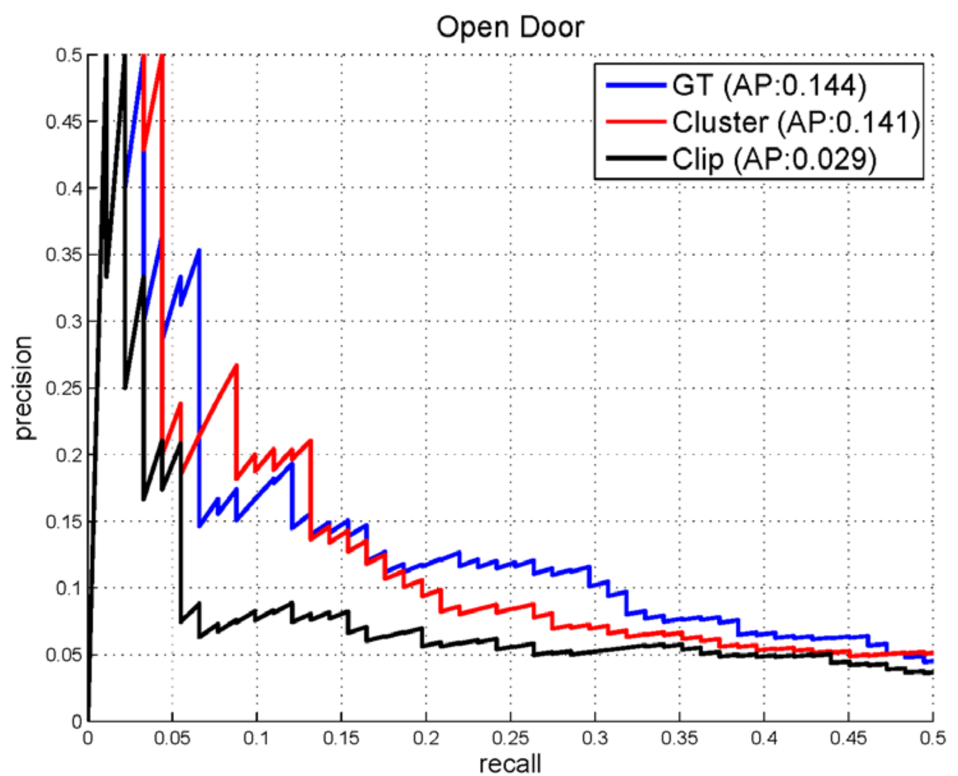
Random video samples: lots of them, very low chance to be positives

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

# Action 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]

# What we have seen so far

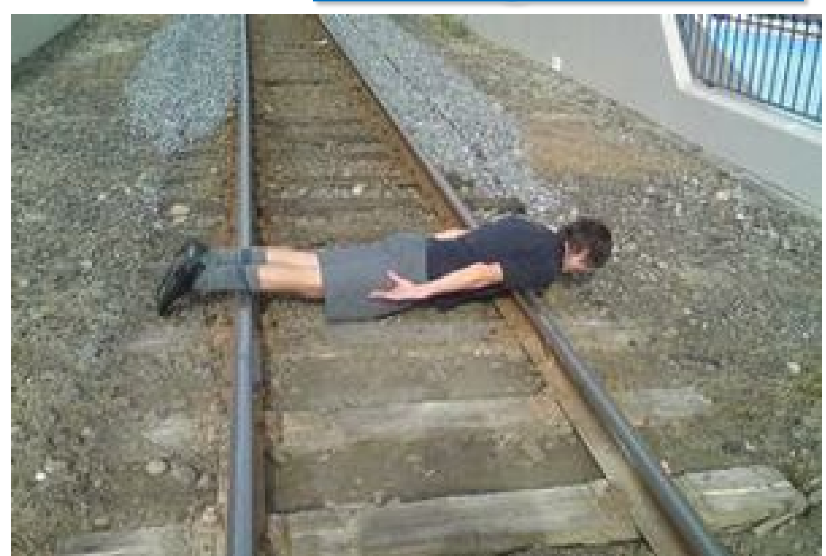
**Actions understanding in realistic settings:**

Action classification


**Is classification the final answer?**

# How to recognize this as unusual?

# How to recognize this as dangerous?



# Is action vocabulary well-defined ?

**Examples of  an action “Open”**

# Is action vocabulary well-defined ?

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

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

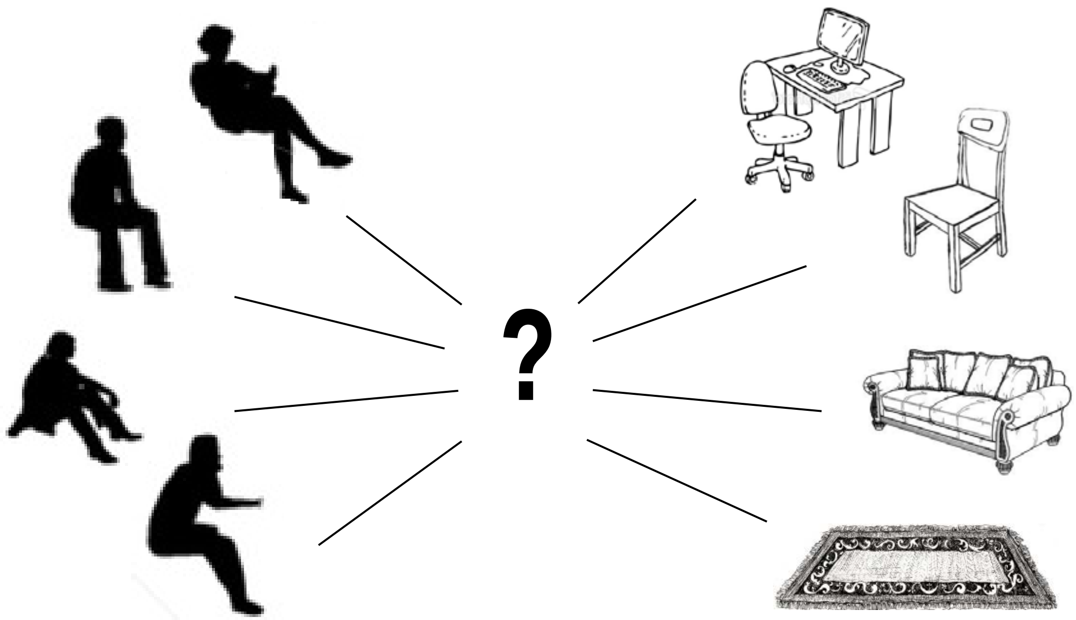
# Scene semantics from long-term observation of people

ECCV 2012

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

# Motivation

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




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

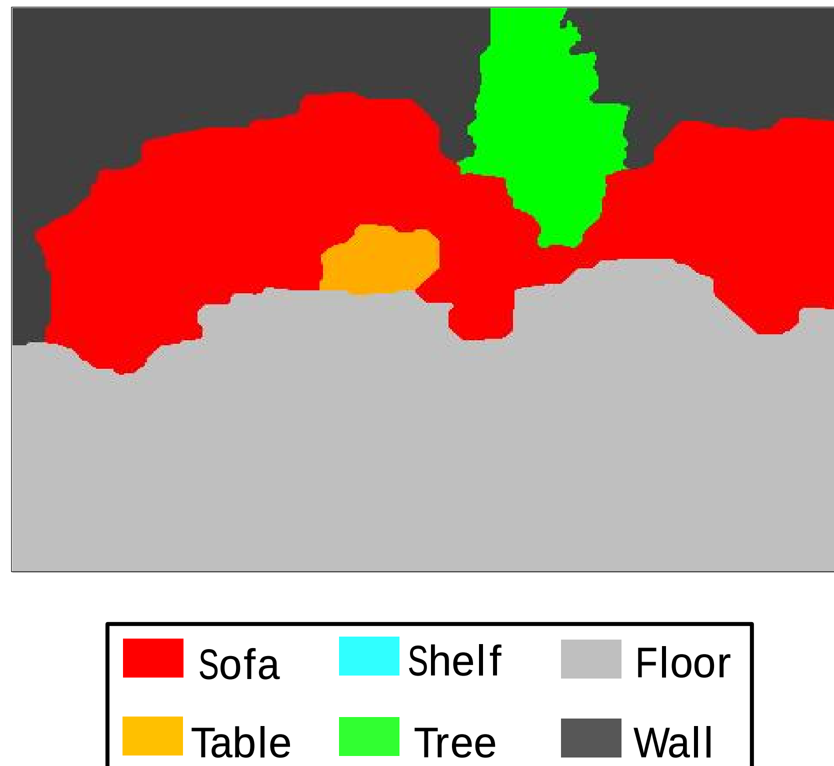
# Goal

Recognize objects by the way people interact with them.

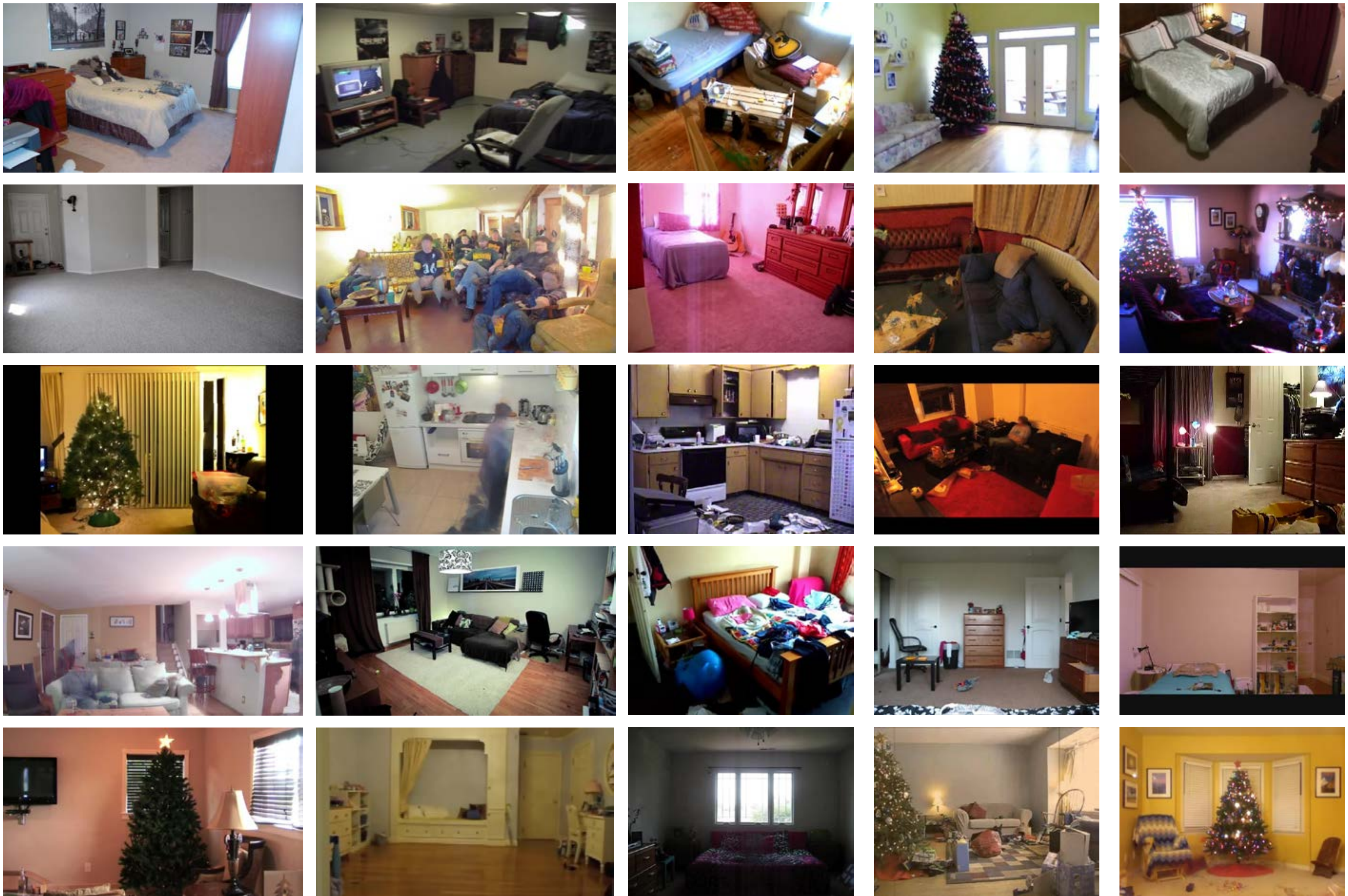
Time-lapse “Party & Cleaning” videos

Lots of person-object interactions, many scenes on YouTube

Semantic object segmentation



# New “Party & Cleaning” dataset

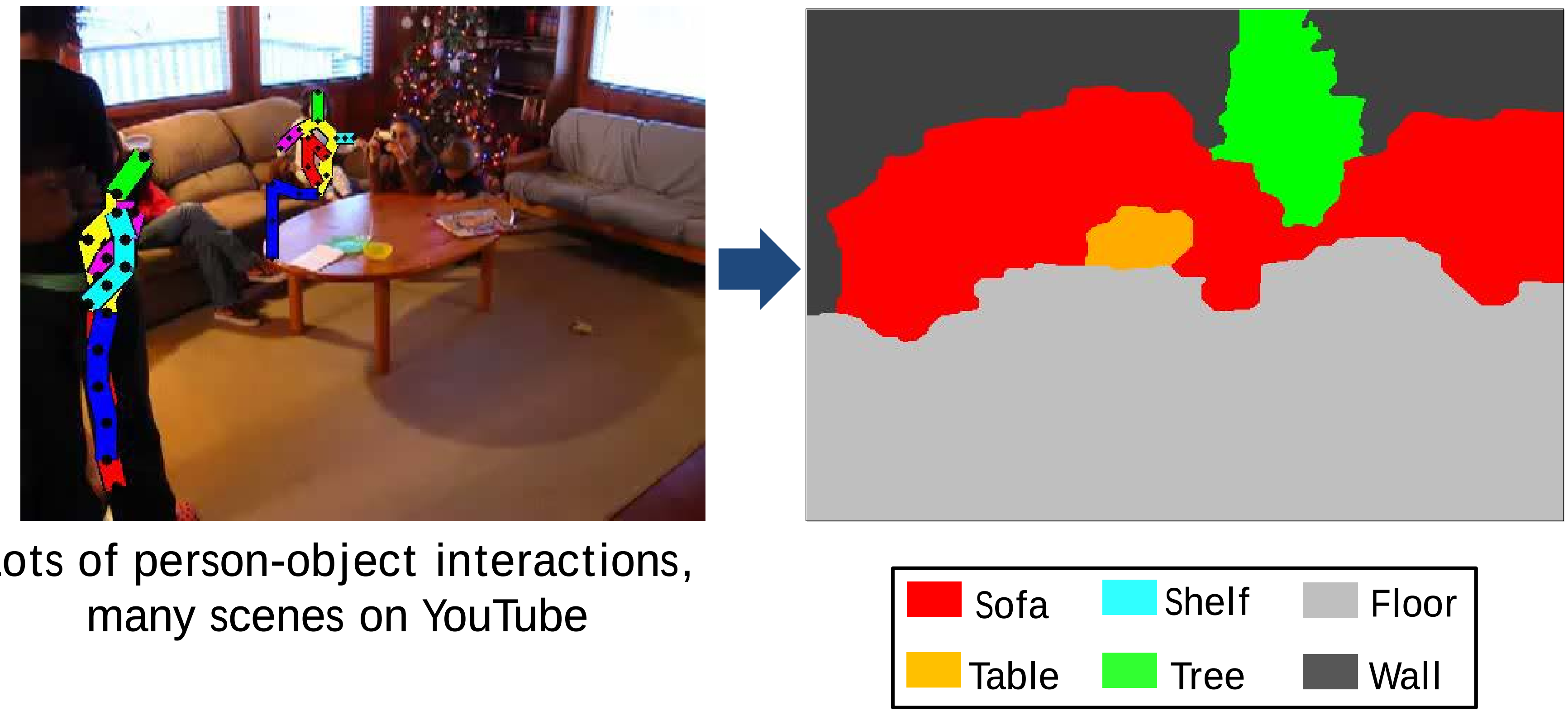


# Goal

Recognize objects by the way people interact with them.

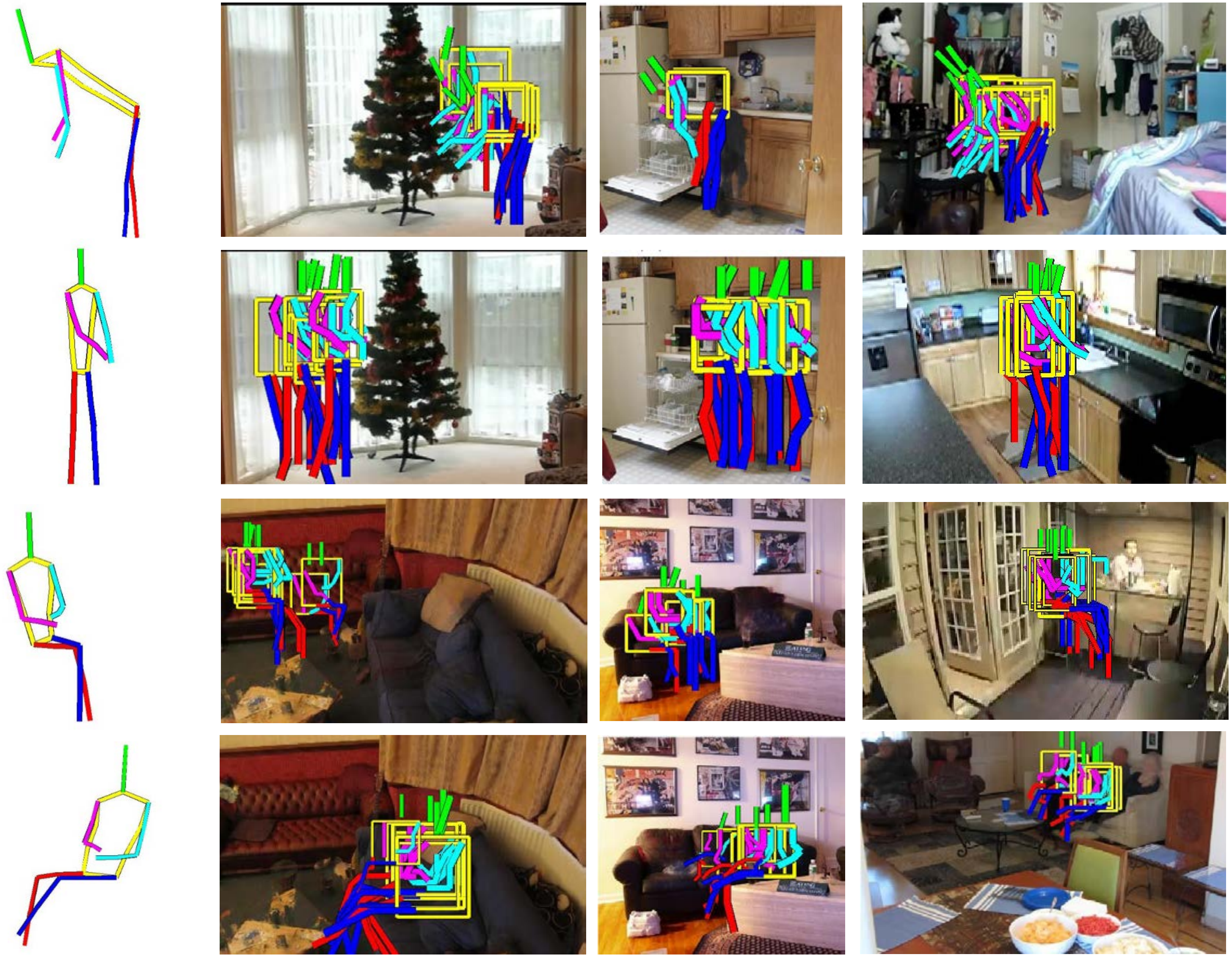
Time-lapse “Party & Cleaning” videos

Semantic object segmentation

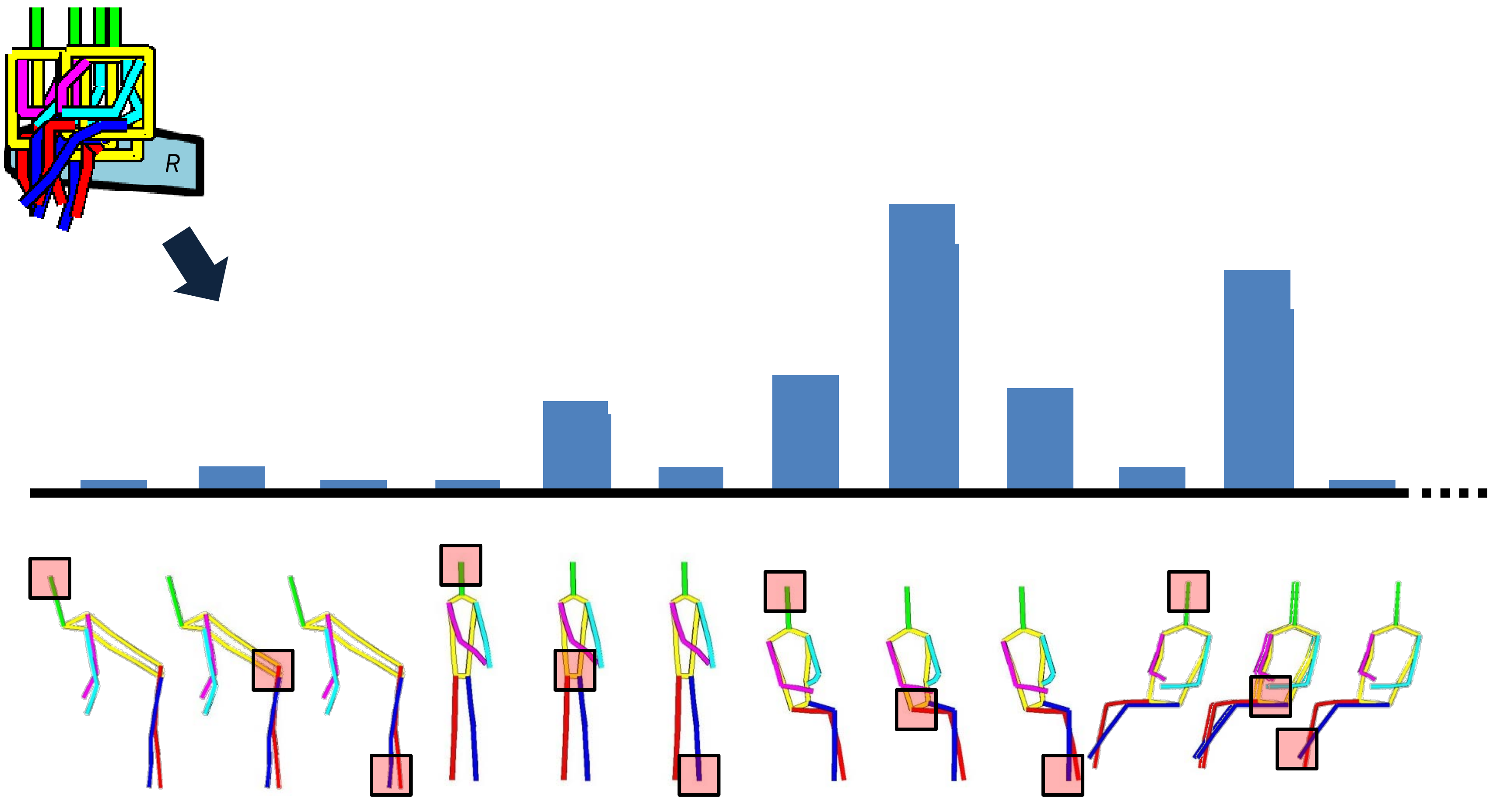



Lots of person-object interactions,
many scenes on YouTube

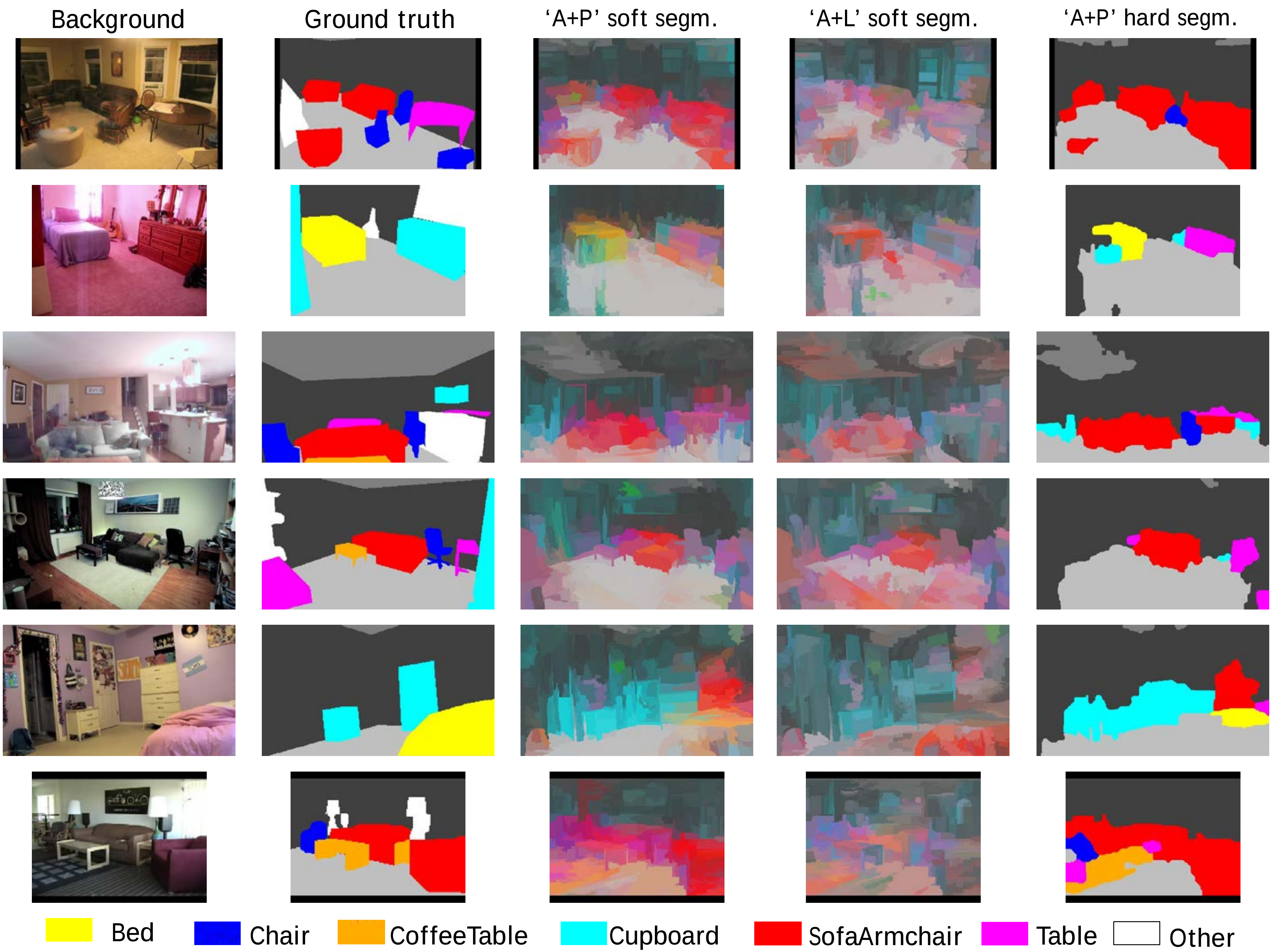
# Pose vocabulary



# Pose histogram

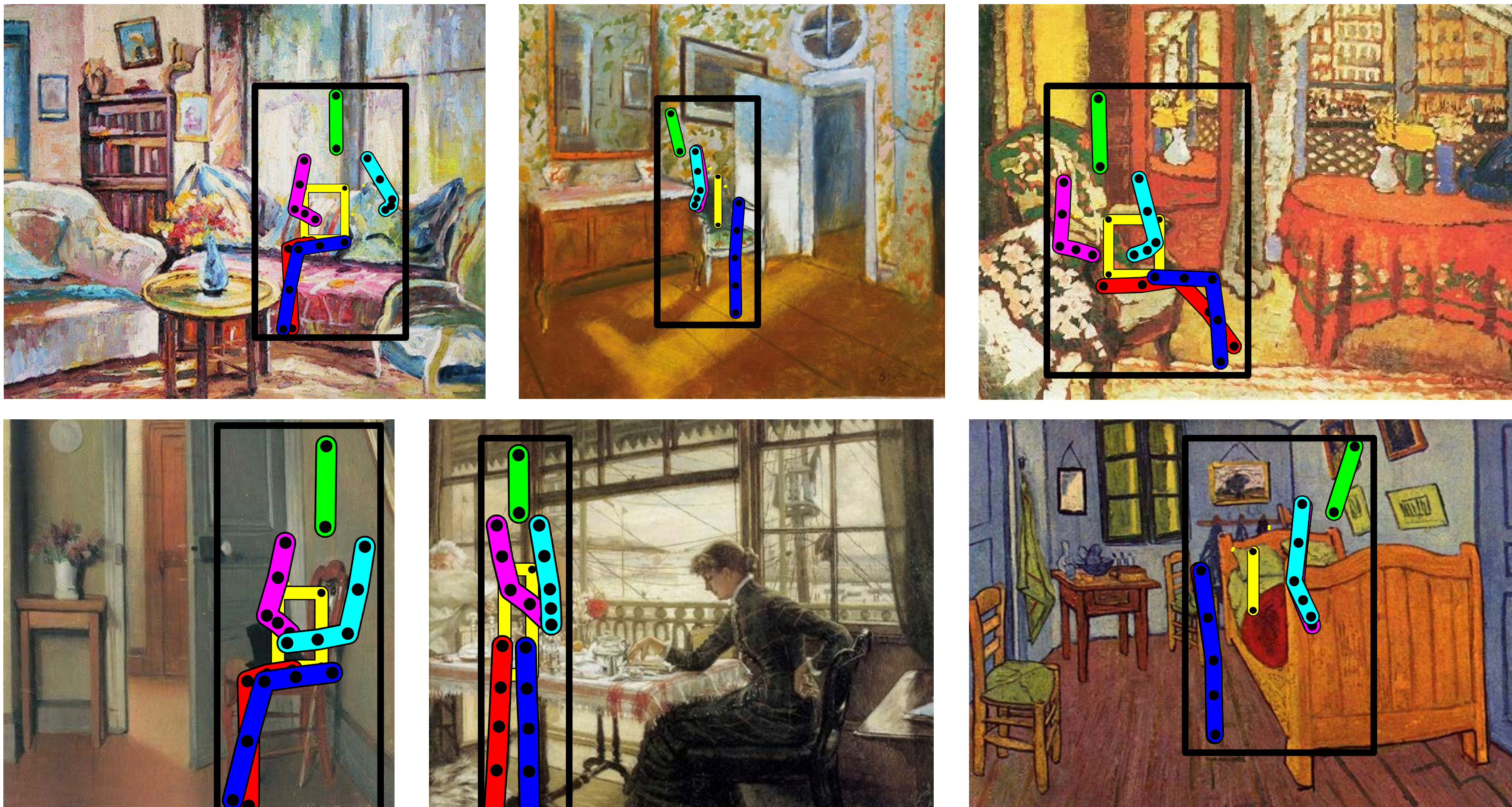


# Some qualitative results


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

- BOF methods give encouraging results for action recognition in realistic data. But better models are needed

- Large-scale readily available annotation provides reach source of supervision for action recognition.

- Action vocabulary is not well-defined. Classifying videos to N labels is not the end of the story. Recognizing object function and human actions should be addressed jointly

informatics mathematics

Willow, Paris