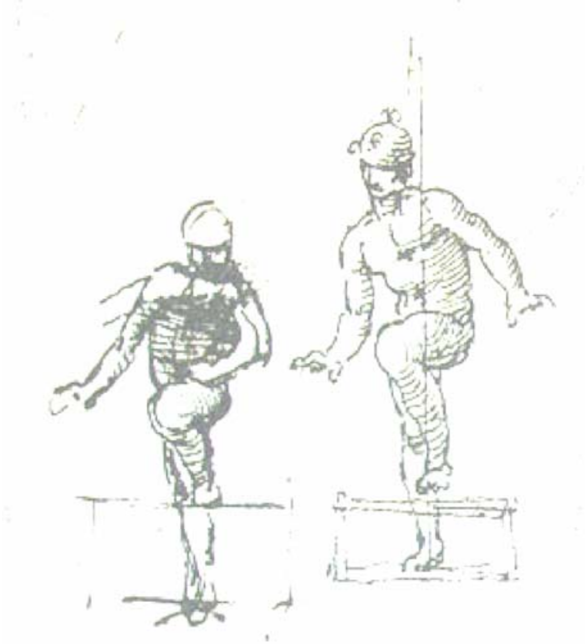


ENS/INRIA Visual Recognition and Machine Learning
Summer School, 25-29 July, Paris, France

Human Action Recognition

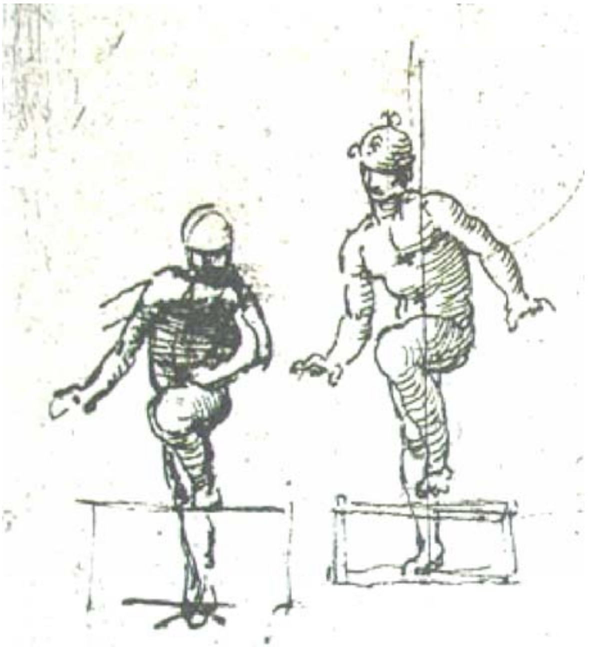
Ivan Laptev

ivan.laptev@inria.fr

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

Includes slides from: Alyosha Efros, Mark Everingham 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

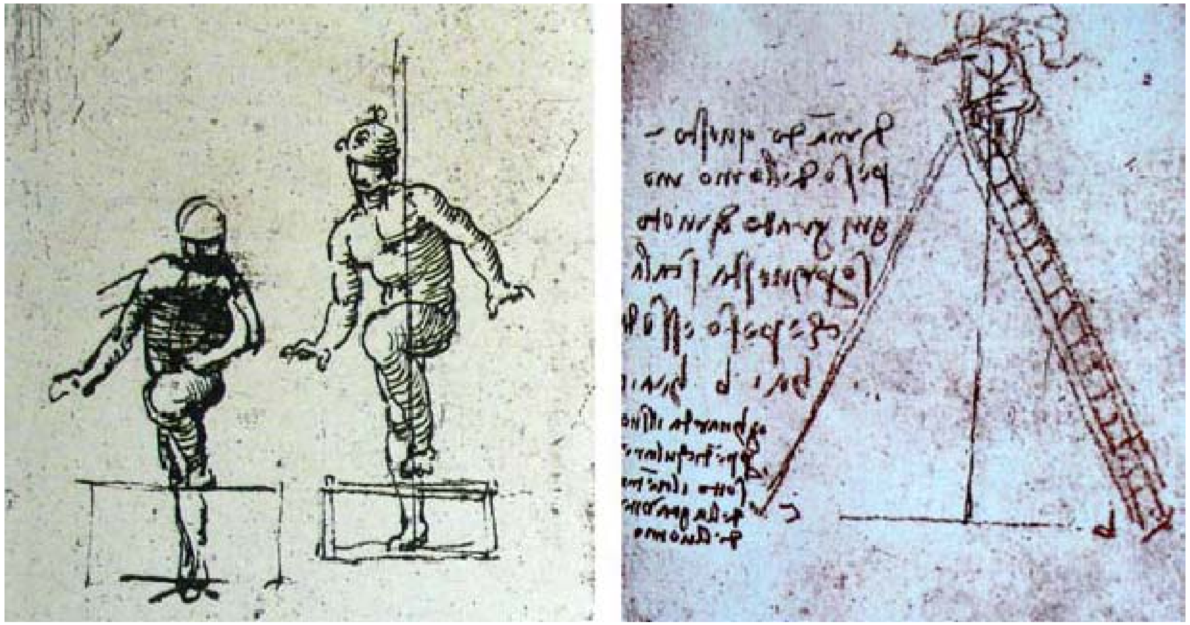
Space-time features
Training with weak supervision

Motivation I: Artistic Representation

Early studies were motivated by human representations in Arts

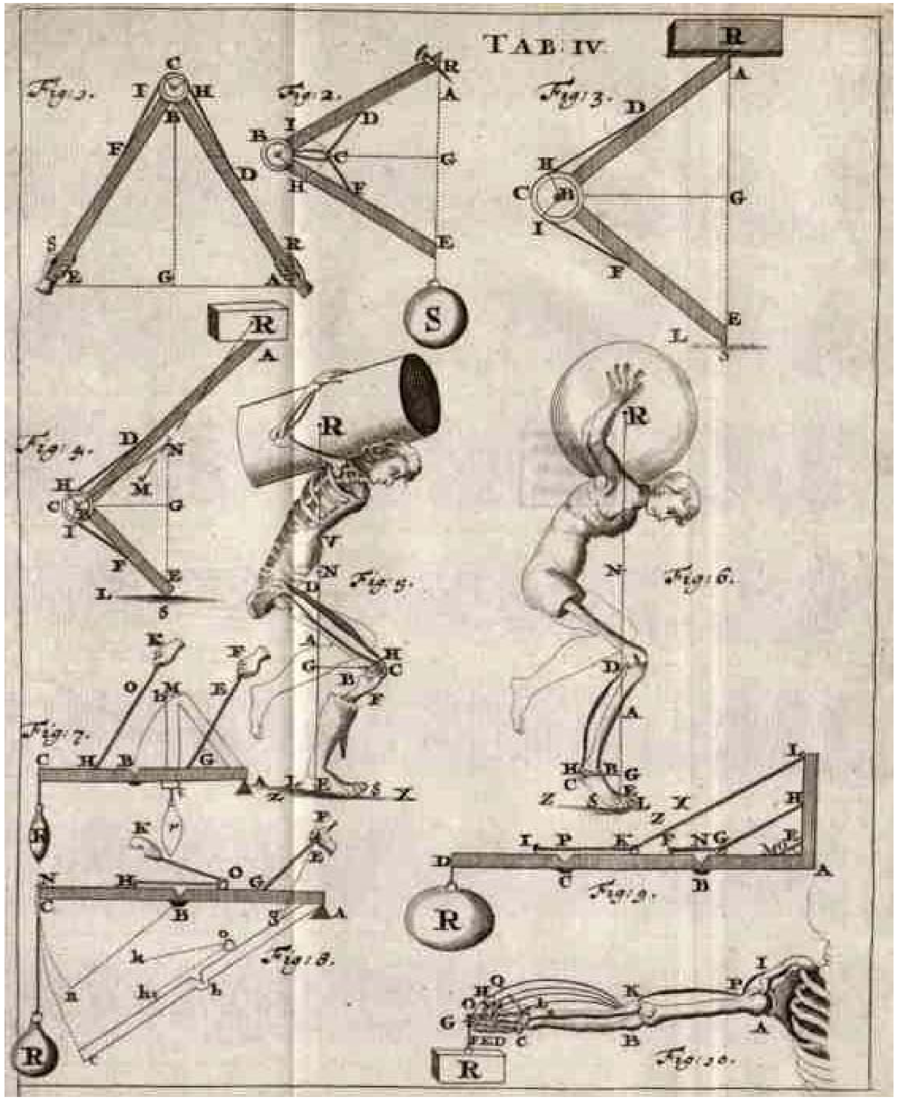
Da Vinci: “it is indispensable for a painter, to become totally familiar with the anatomy of nerves, bones, muscles, and sinews, such that he understands for their various motions and stresses, which sinews or which muscle causes a particular motion”

“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

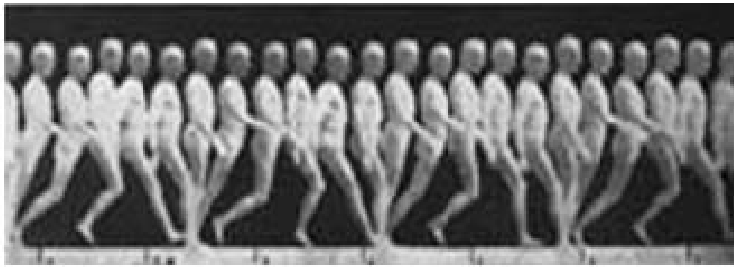


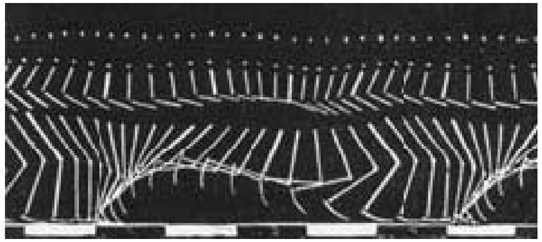
Giovanni Alfonso Borelli (1608–1679)

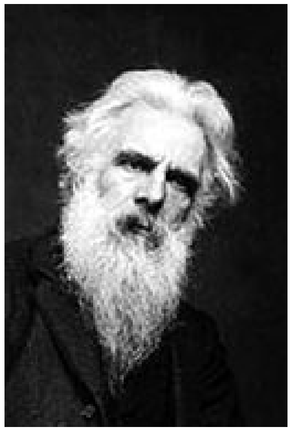
- The emergence of *biomechanics*
- Borelli applied to biology the analytical and geometrical methods, developed by Galileo Galilei
- He was the first to understand that bones serve as levers and muscles function according to mathematical principles
- His physiological studies included muscle analysis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

Motivation III: Motion perception

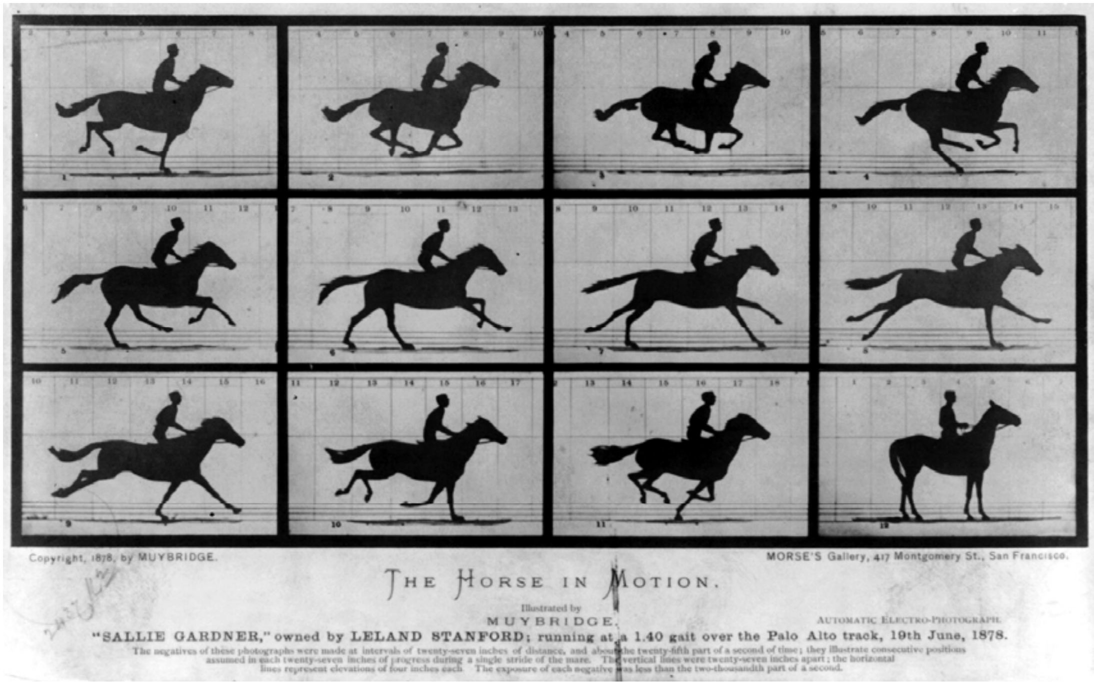
Etienne-Jules Marey: (1830–1904) made Chronophotographic experiments influential for the emerging field of *cinematography*





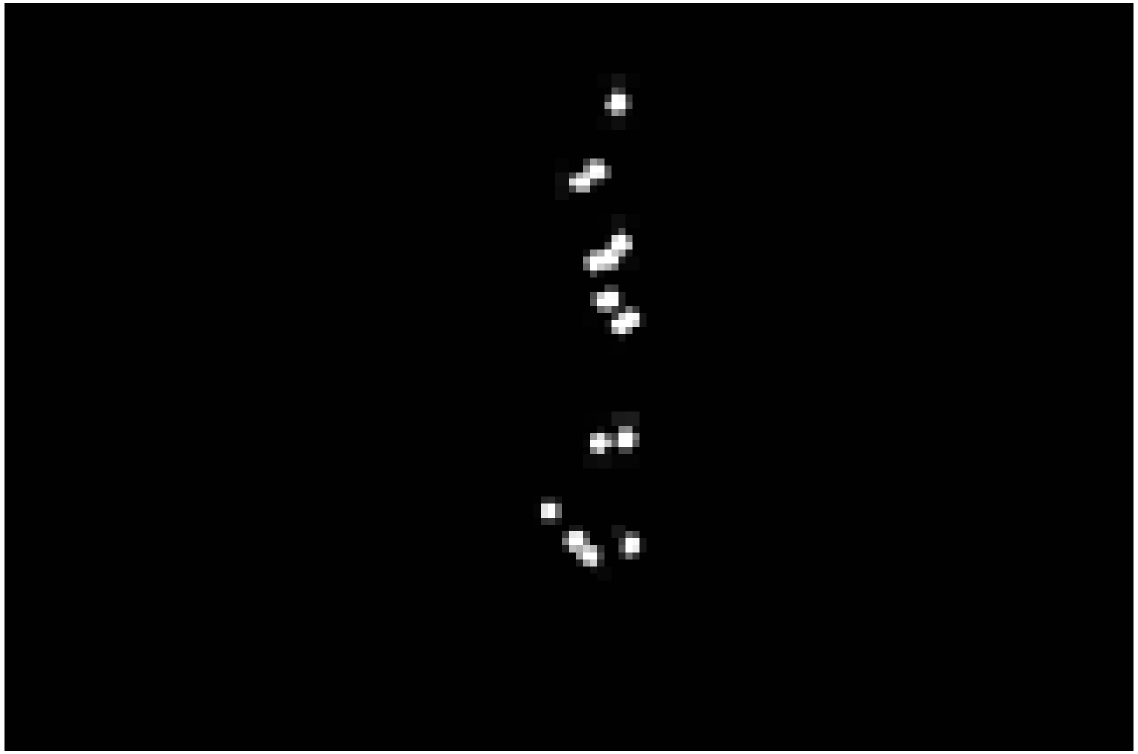


Eadweard Muybridge (1830–1904) invented a machine for displaying the recorded series of images. He pioneered motion pictures and applied his technique to movement studies



Motivation III: Motion perception

- Gunnar Johansson [1973] 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.

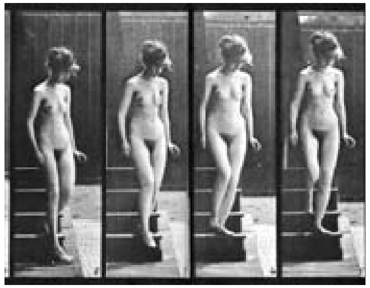


Gunnar Johansson, **Perception and Psychophysics,** 1973

Human actions: Historic overview

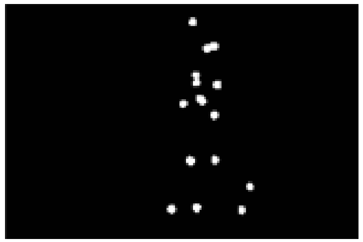
15^{th} century
studies of
anatomy

17^{th} century
emergence of
biomechanics



19^{th} century
emergence of
cinematography

1973
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

Why Action Recognition?

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

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

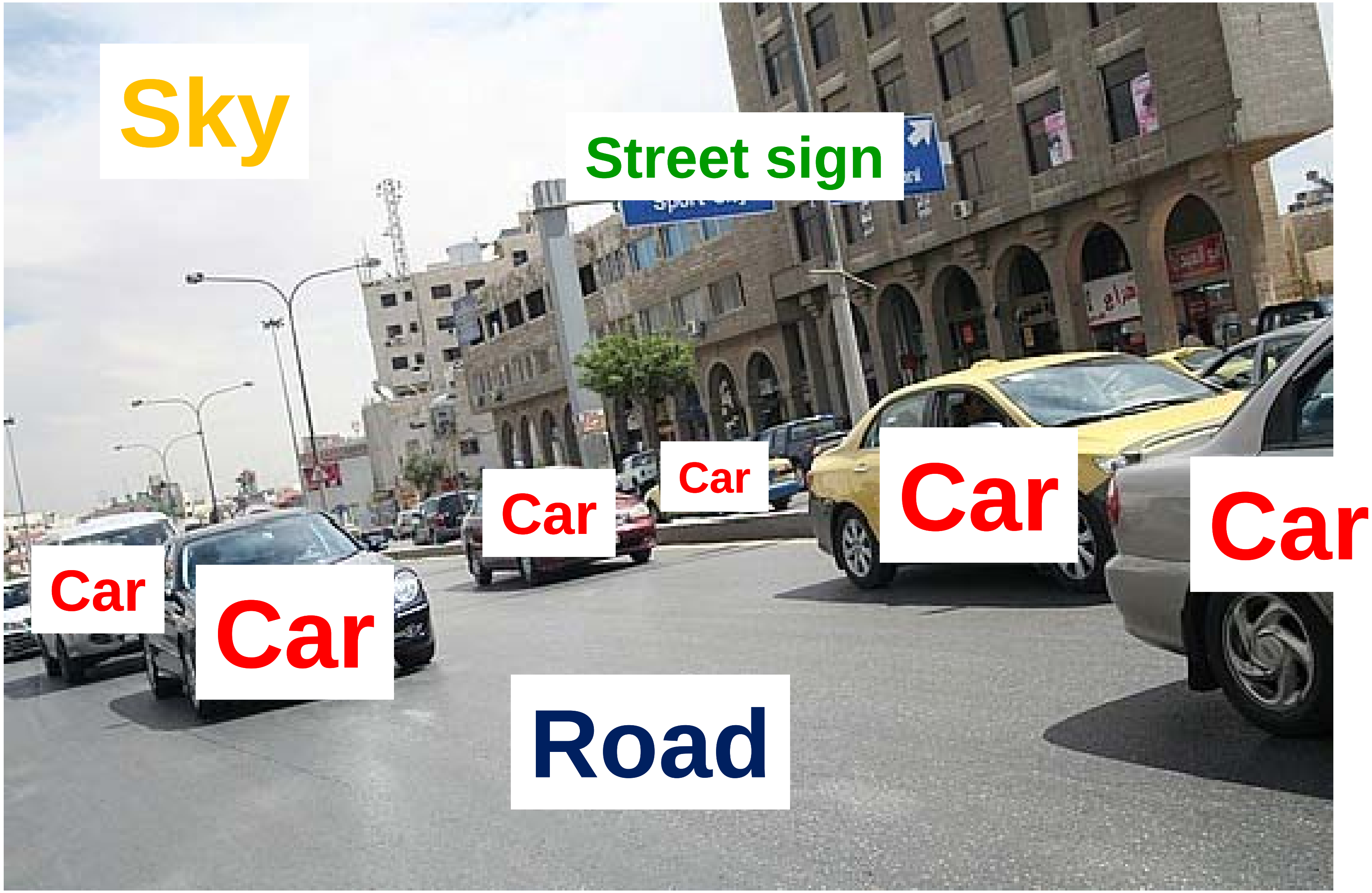
Home videos: e.g. *“My daughter climbing”*

Sociology research:

Manually analyzed smoking actions in 900 movies

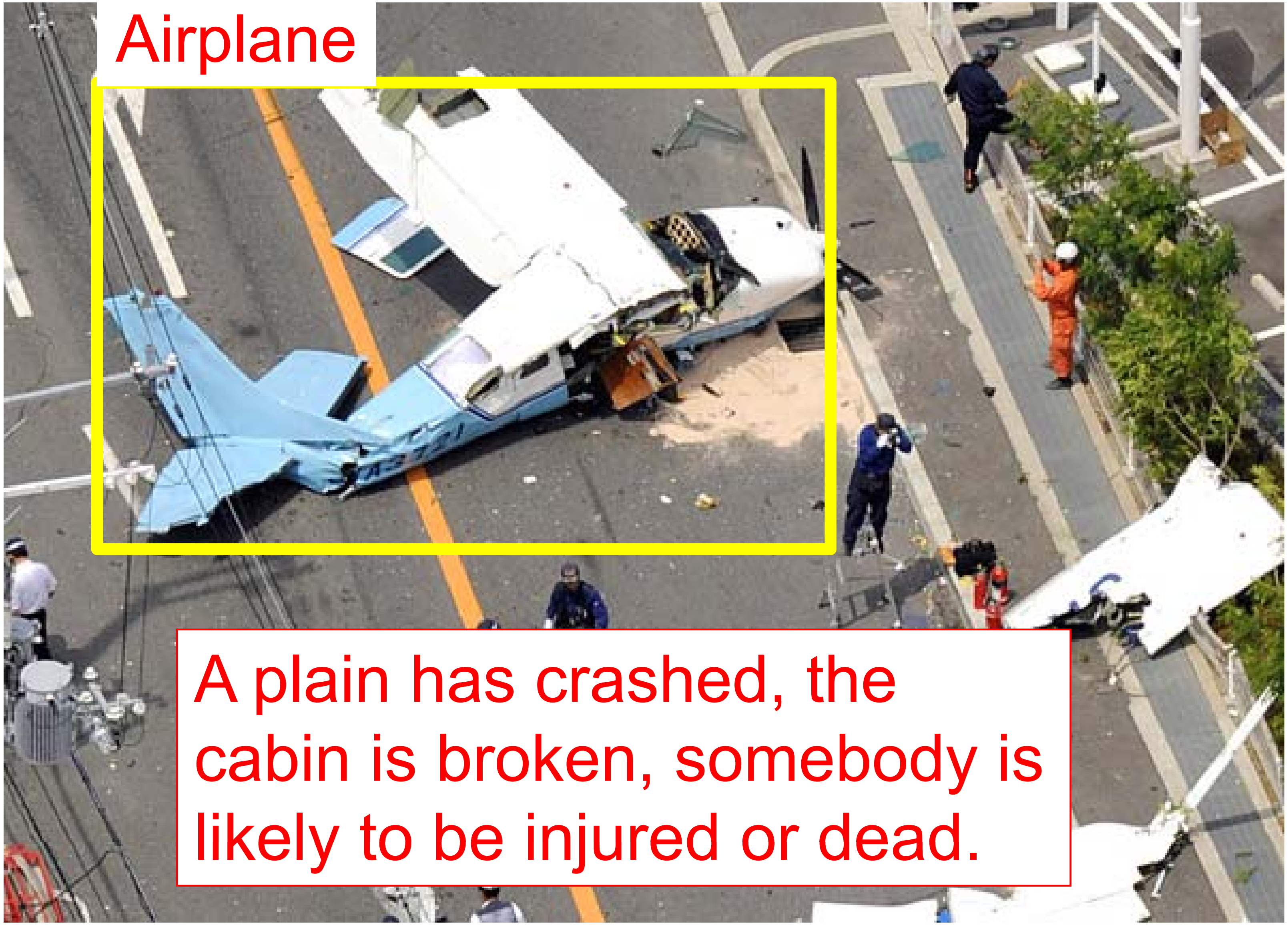
Surveillance: 260K views in 7 days on YouTube

How action recognition is related to computer vision?



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


المدينه الرياضيه
Sport City
Shmeisani


Airplane
A plain has crashed, the cabin is broken, somebody is likely to be injured or dead.


cat
woman
trash bin


ONECO

- Vision is **person-centric**: We mostly care about things which are important to us, people

- Actions of people reveal the function of objects

- Future challenges:

  - Function: What can I do with this and how?
  - Prediction: What can happen if someone does that?
  - Recognizing goals: What this person is trying to do?

How many person-pixels are there?

Movies

TV

YouTube

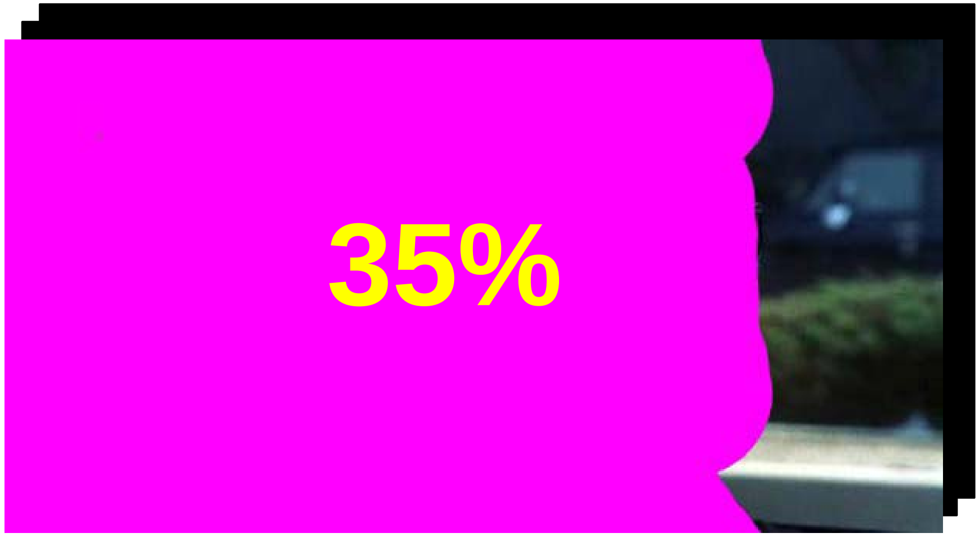
How many person-pixels are there?

Movies

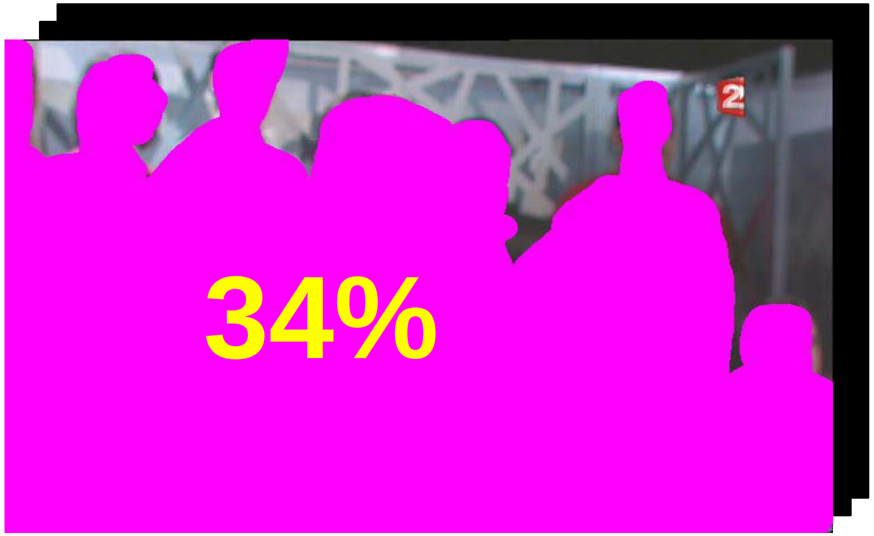
TV

YouTube

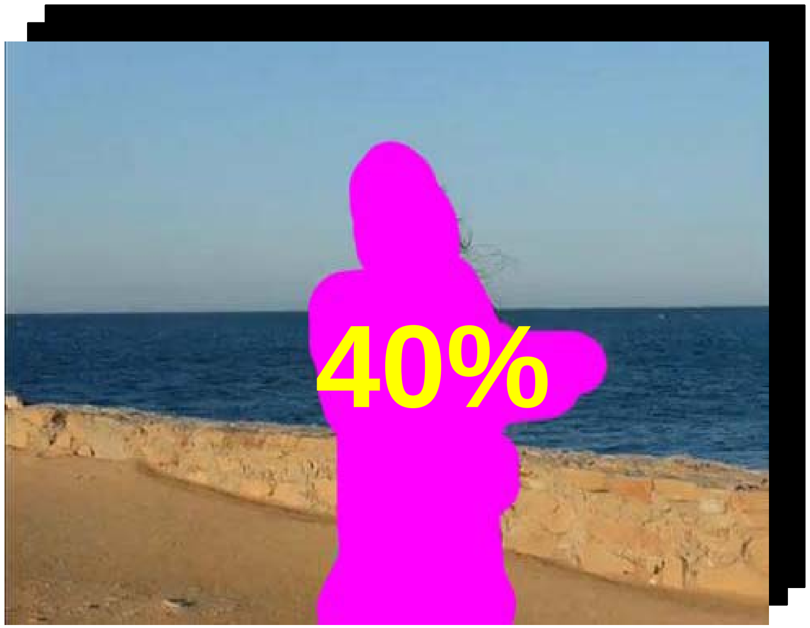
How many person-pixels are there?



Movies



TV



YouTube

How much data do we have?

- Huge amount of video is available and growing

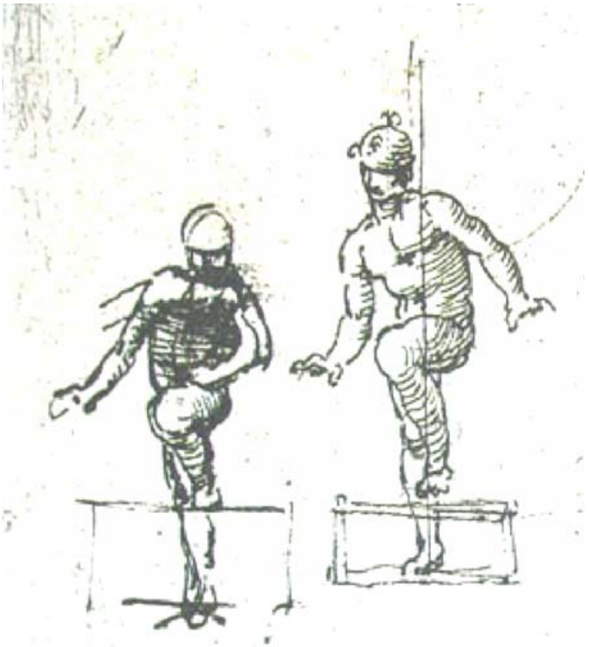
TV-channels recorded since 60's

>34K hours of video upload every day

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

If we want to interpret this data, we should better understand what person-pixels are telling us!

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

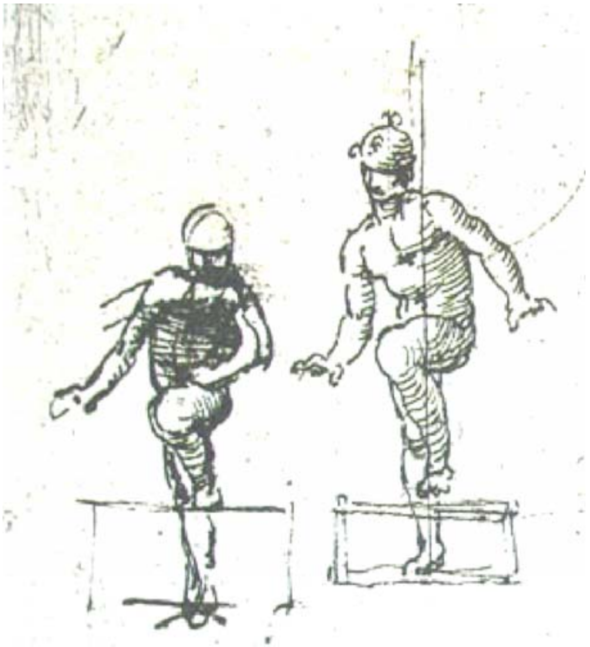
Motion-based methods

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

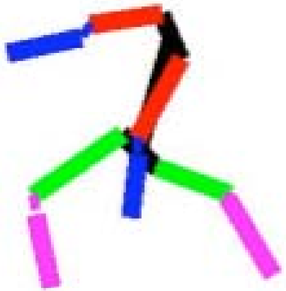
Generic and parametric Optical Flow
Motion templates

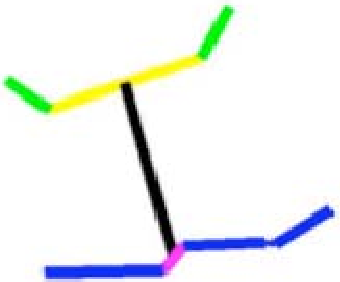
Space-time methods

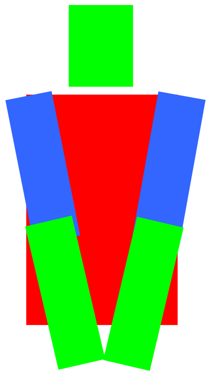
Space-time features
Training with weak supervision

Objective and motivation

Determine human body pose (layout)

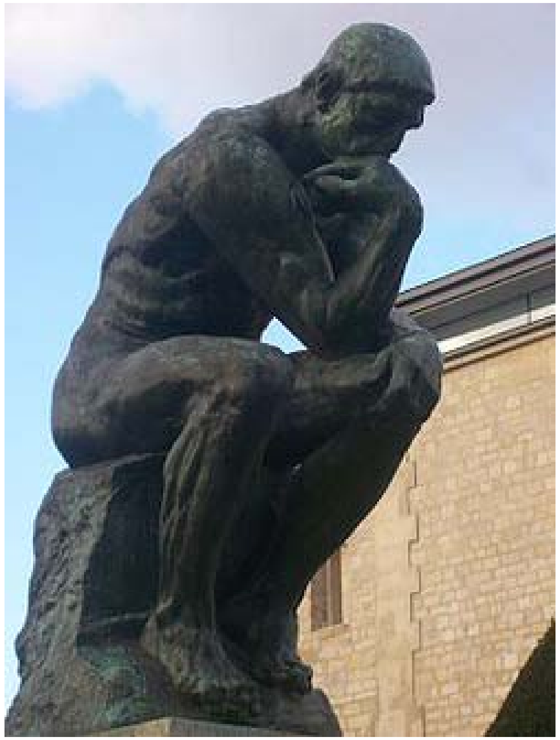






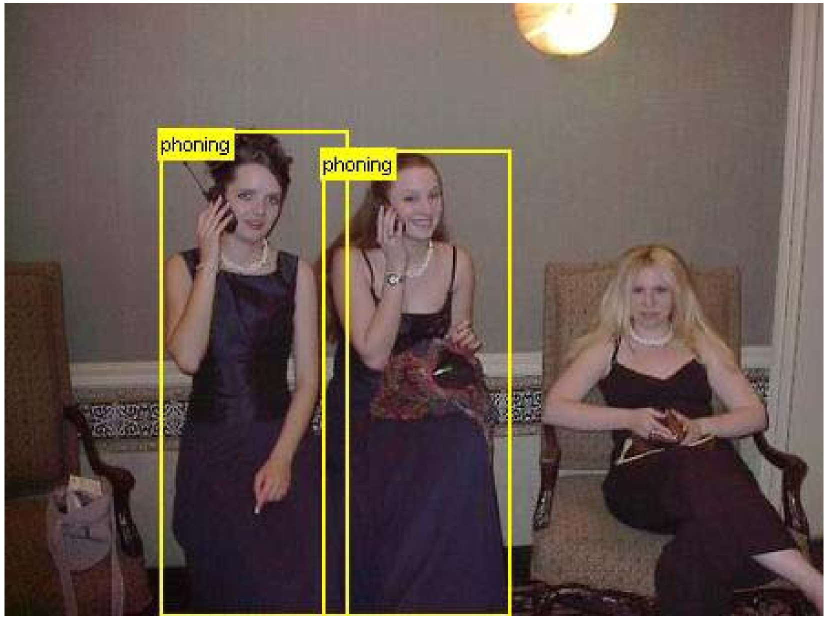
Why? To recognize poses, gestures, actions

Activities characterized by a pose



Activities characterized by a pose

Activities characterized by a pose



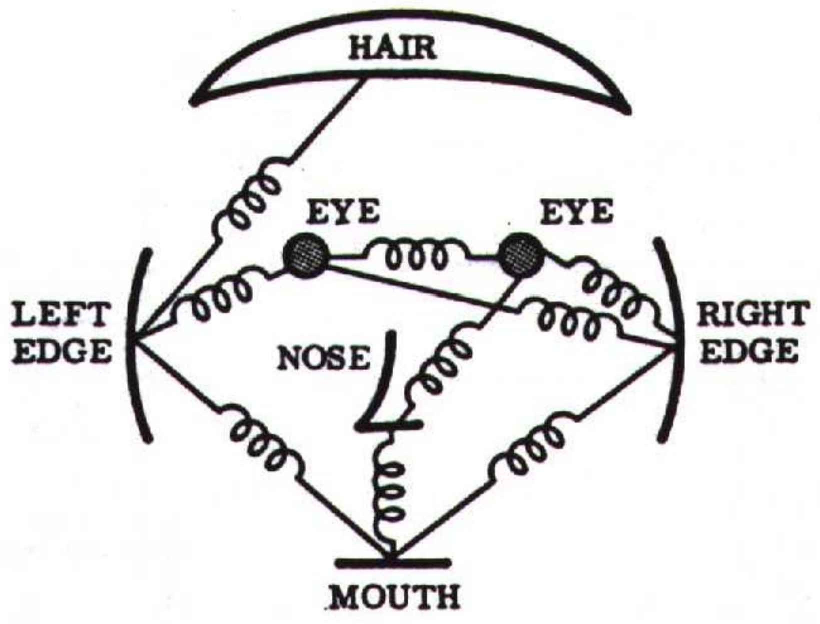
Challenges: articulations and deformations

Challenges: of (almost) unconstrained images

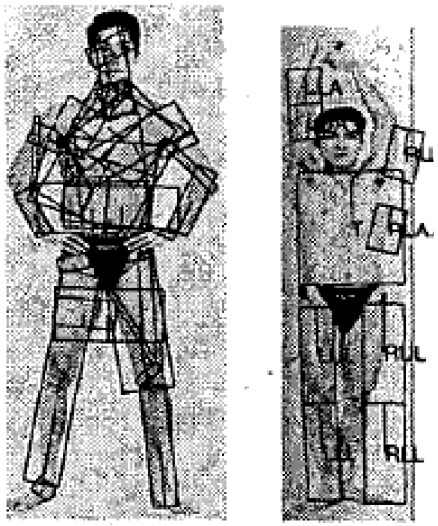
varying illumination and low contrast; moving camera and background; multiple people; scale changes; extensive clutter; any clothing

Pictorial Structures

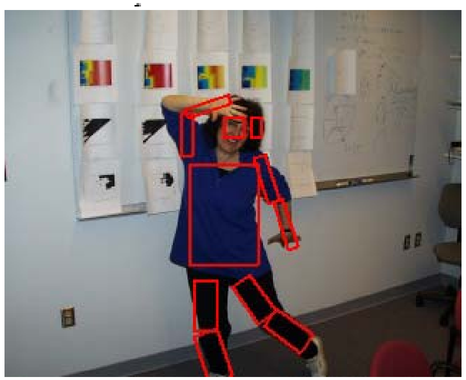
- Intuitive model of an object
- Model has two components
  1. parts (2D image fragments)
  2. structure (configuration of parts)
- Dates back to Fischler & Elschlager 1973



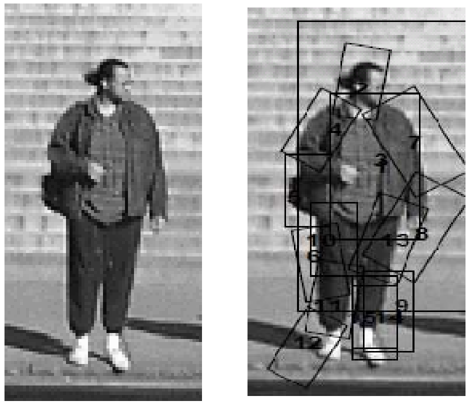
Long tradition of using pictorial structures for humans



Finding People by Sampling
Ioffe & Forsyth, ICCV 1999

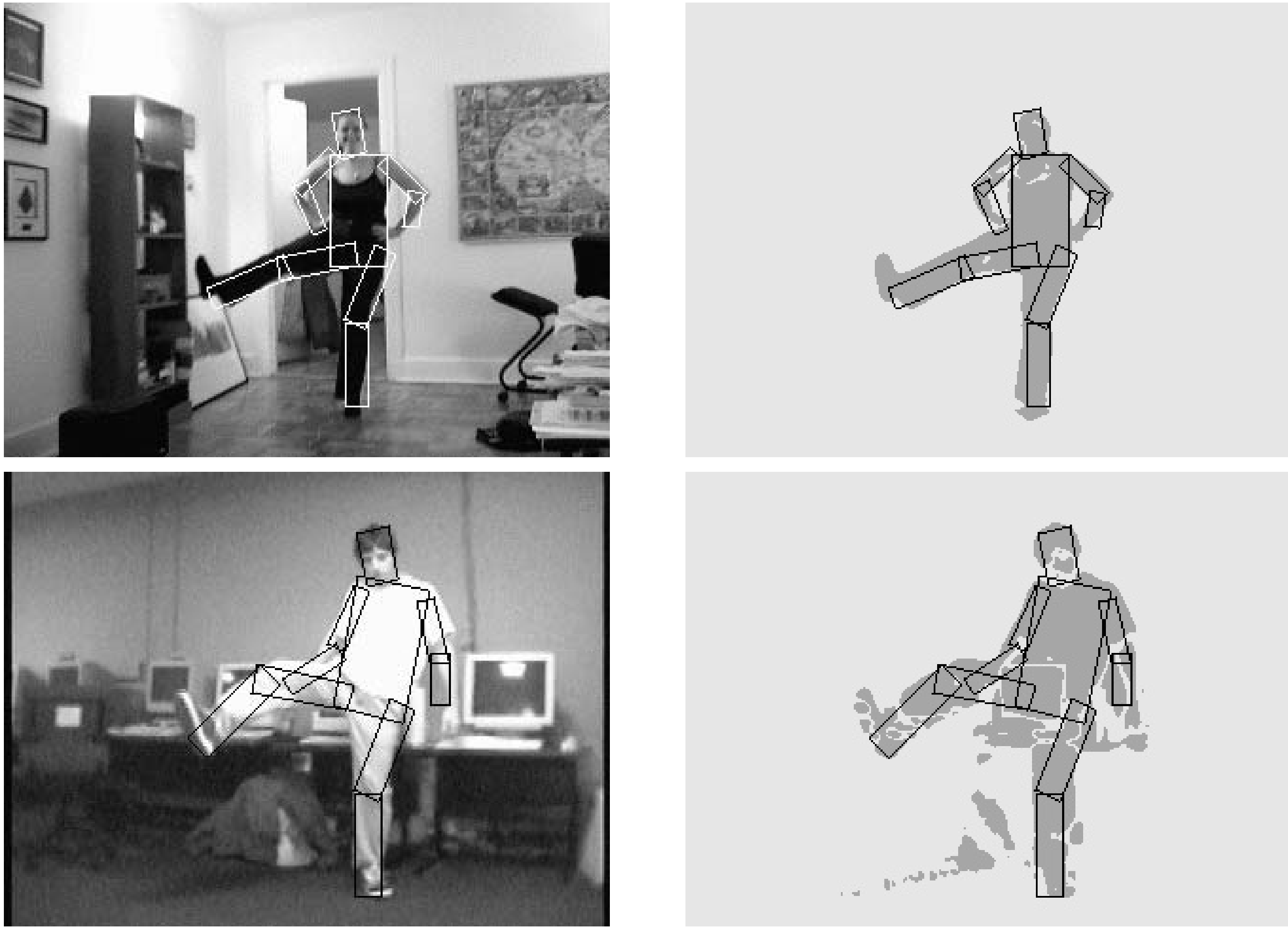


Pictorial Structure Models for Object Recognition
Felzenszwalb & Huttenlocher, 2000



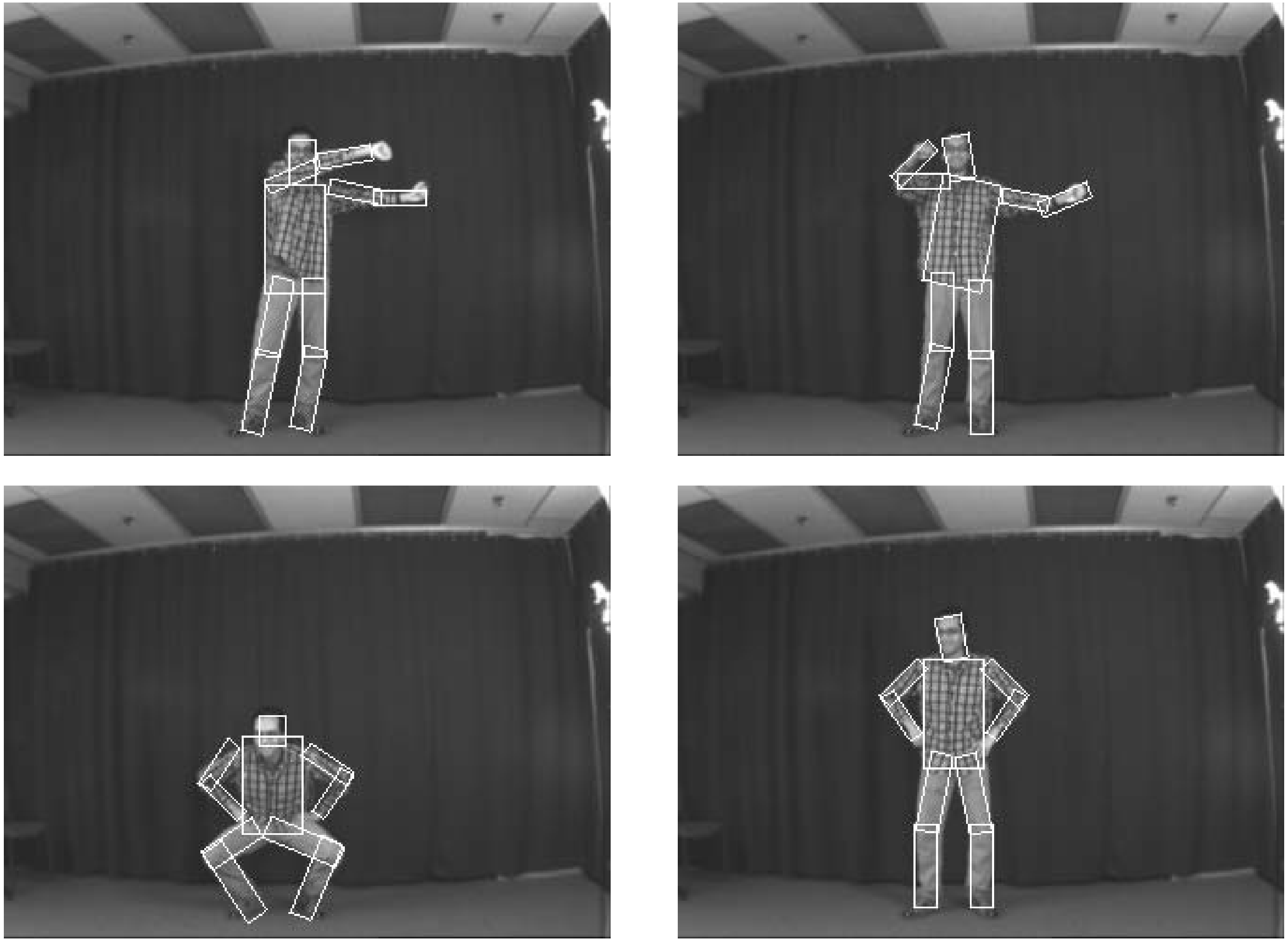
Learning to Parse Pictures of People
Ronfard, Schmid & Triggs, ECCV 2002

Felzenszwalb & Huttenlocher

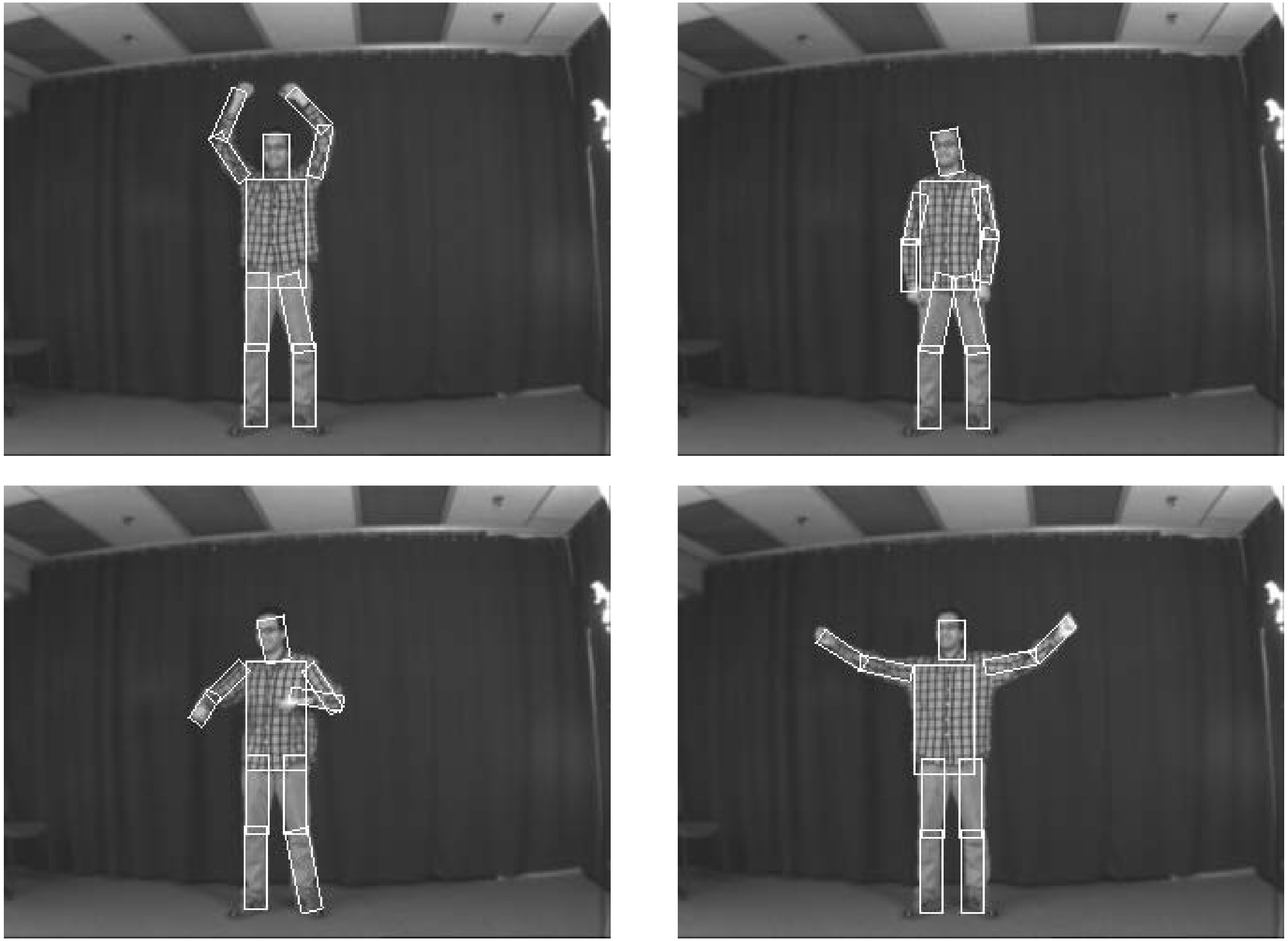


NB: requires background subtraction

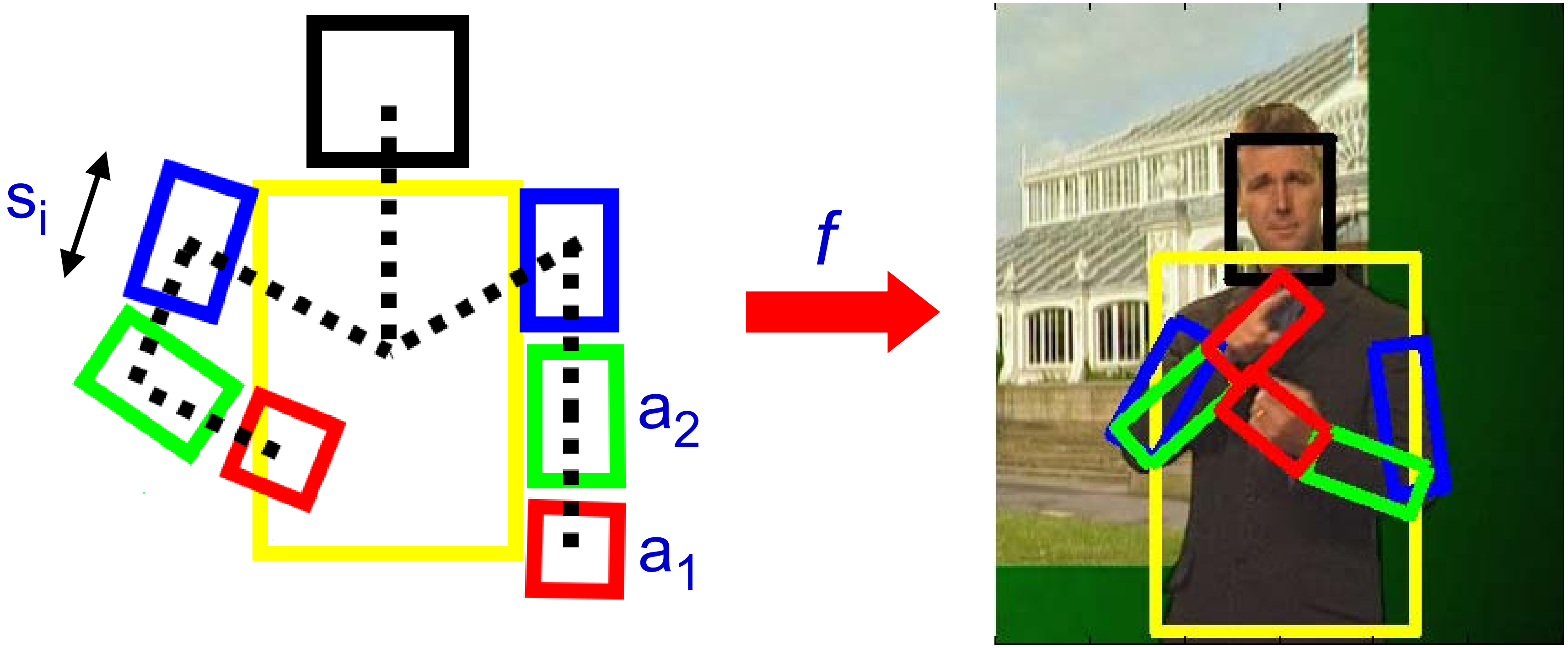
Variety of Poses



Variety of Poses



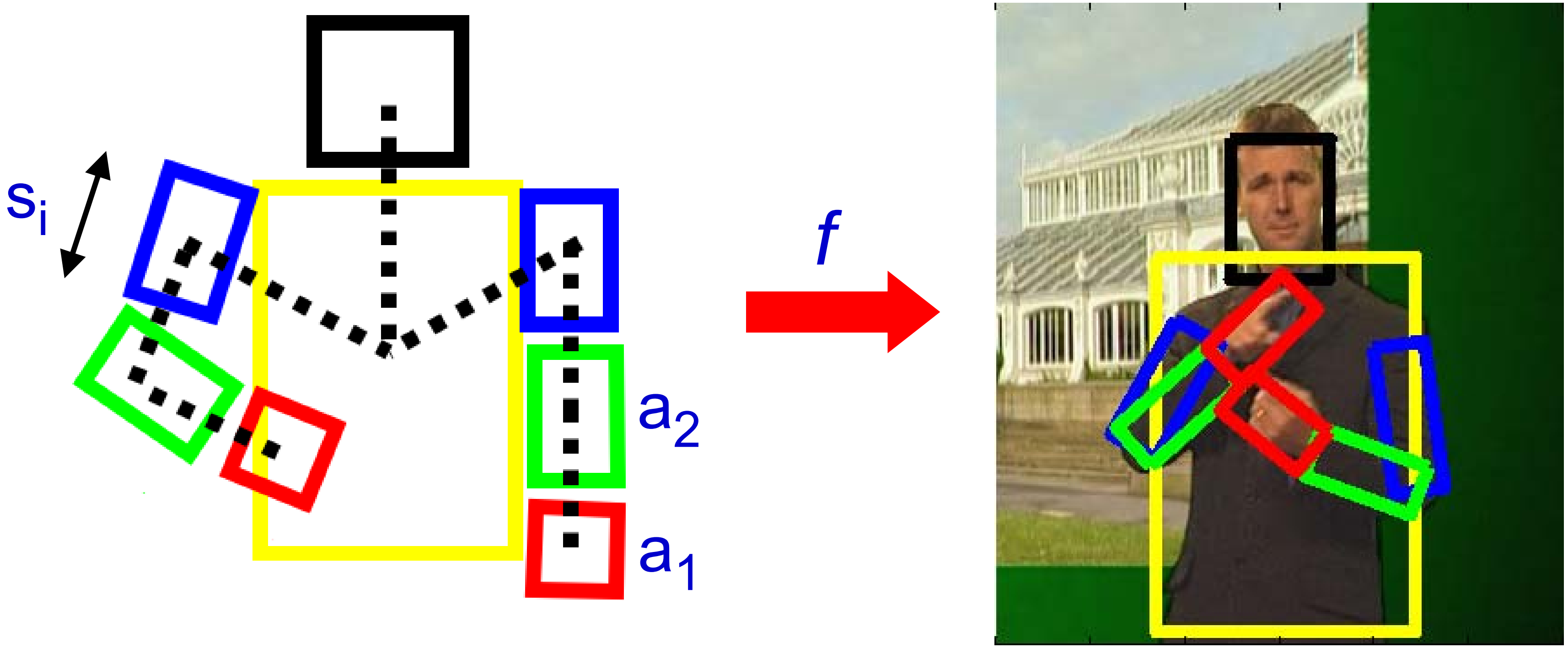
Objective: detect human and determine upper body pose (layout)



Model as a graph labelling problem

- Vertices $\mathcal{V}$ are parts, $a_i, i = 1, \cdots, n$
- Edges $\mathcal{E}$ are pairwise linkages between parts
- For each part there are h possible poses $\mathbf{p}_j = (x_j, y_j, \phi_j, s_j)$
- Label each part by its pose: $f : \mathcal{V} \longrightarrow \{1, \cdots, h\}$, i.e. part a takes pose $\mathbf{p}_{f(a)}$.

Pictorial structure model – CRF



Features for unary:

- colour
- HOG

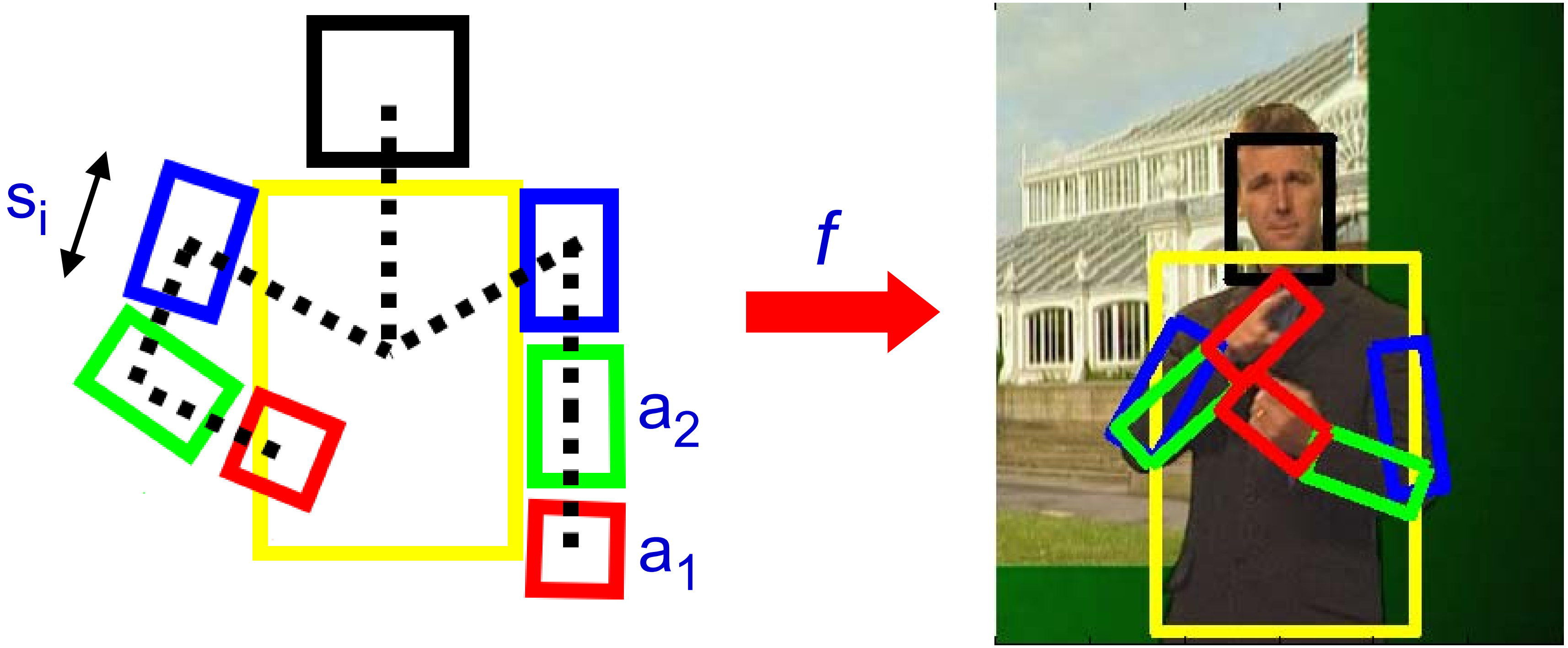
for limbs/torso

- Each labelling has an energy (cost):

$$E(f) = \underbrace{\sum_{a \in \mathcal{V}} \theta_{a;f(a)}}_{\substack{\text{unary terms} \\ \text{(appearance)}}} + \underbrace{\sum_{(a,b) \in \mathcal{E}} \theta_{ab;f(a)f(b)}}_{\substack{\text{pairwise terms} \\ \text{(configuration)}}}$$

- Fit model (inference) as labelling with lowest energy

Complexity

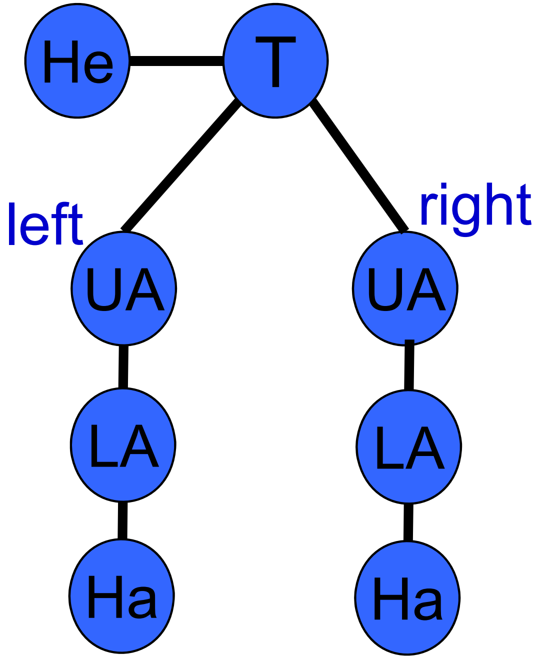


- n parts
- For each part there are h possible poses $\mathbf{p}_j = (x_j, y_j, \phi_j, s_j)$
- There are h^n possible labellings

Problem: any reasonable discretization (e.g. 12 scales and 36 angles for upper and lower arm, etc) gives a number of configurations 10^12 – 10^14

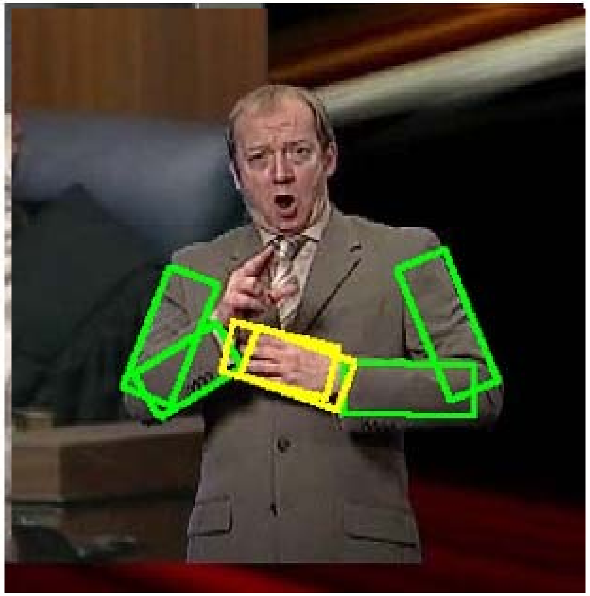
→ Brute force search not feasible

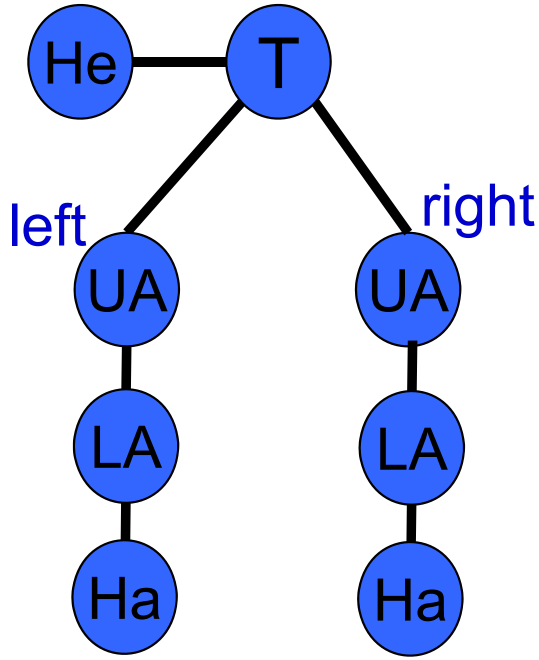
Are trees the answer?



- With n parts and h possible discrete locations per part, $O(h^n)$
- For a tree, using dynamic programming this reduces to $O(nh^2)$
- If model is a tree and has certain edge costs, then complexity reduces to O(nh) using a distance transform [Felzenszwalb & Huttenlocher, 2000, 2005]

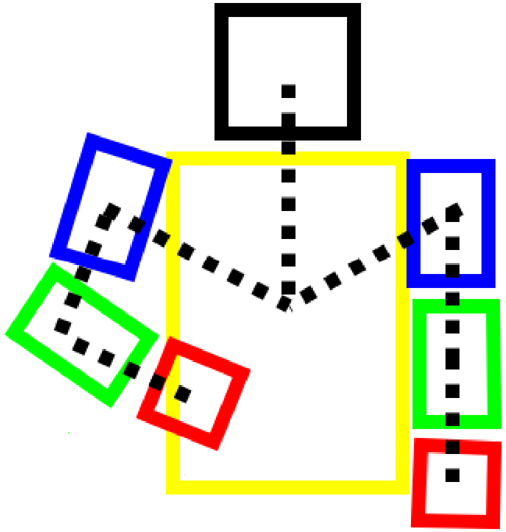
Are trees the answer?



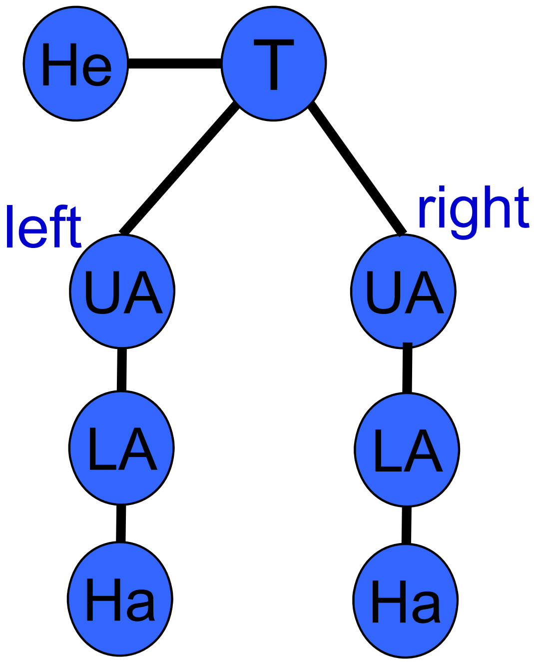


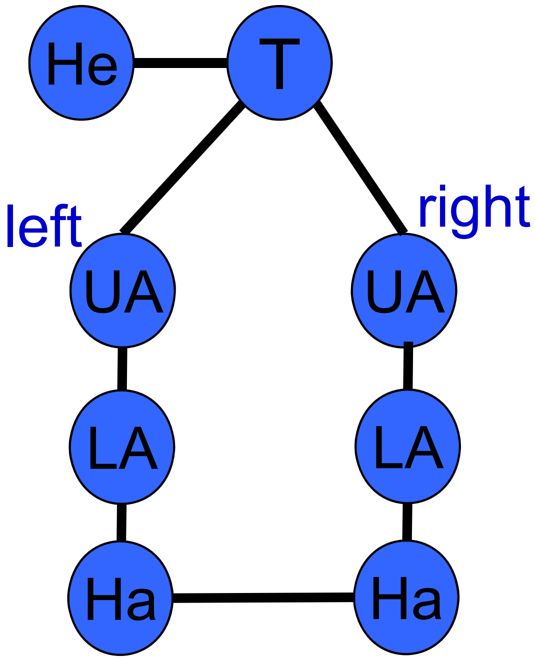
- With n parts and h possible discrete locations per part, $O(h^n)$
- For a tree, using dynamic programming this reduces to $O(nh^2)$
- If model is a tree and has certain edge costs, then complexity reduces to O(nh) using a distance transform [Felzenszwalb & Huttenlocher, 2000, 2005]

Kinematic structure vs graphical (independence) structure



Graph G = (V,E)





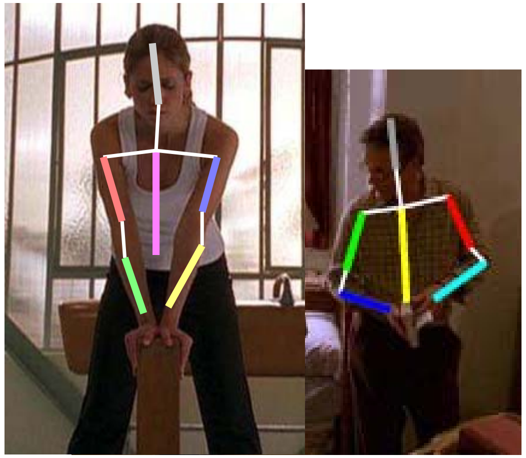
Requires more connections than a tree

More recent work on human pose estimation

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

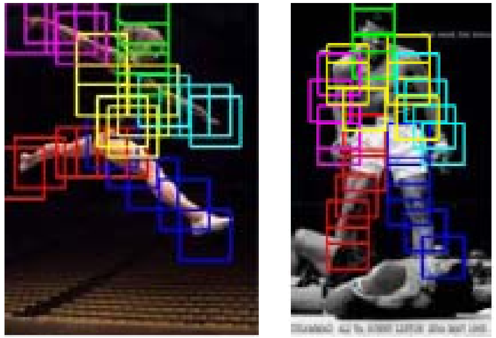
- Person detector & foreground highlighting

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

Learns with weak textual annotation

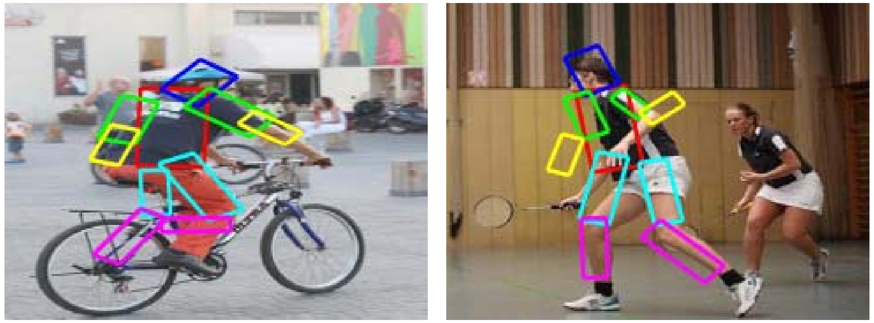
- Multiple instance learning

Pose estimation is a very active research area



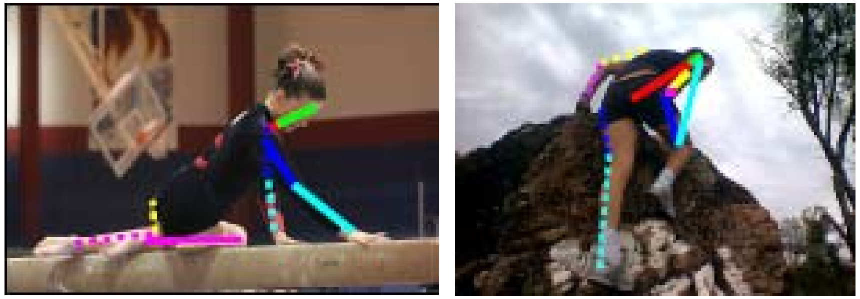
Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In Proc. **CVPR 2011**

Extension of LSVM model of Felzenszwalb et al.



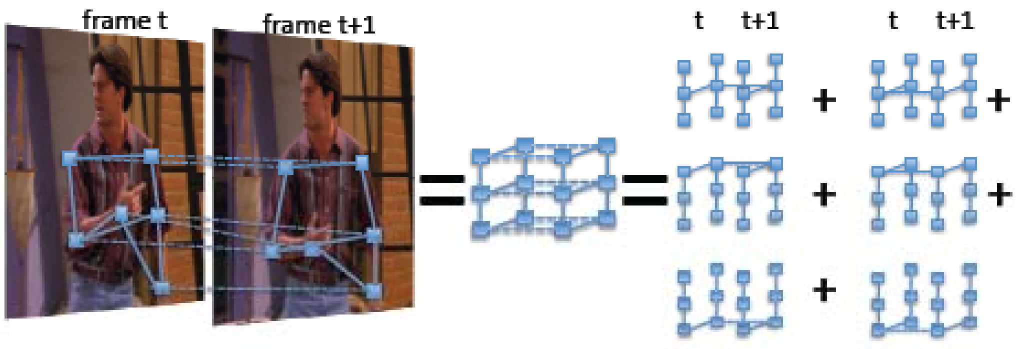
Y. Wang, D. Tran and Z. Liao. Learning Hierarchical Poselets for Human Parsing. In Proc. **CVPR 2011**.

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**.

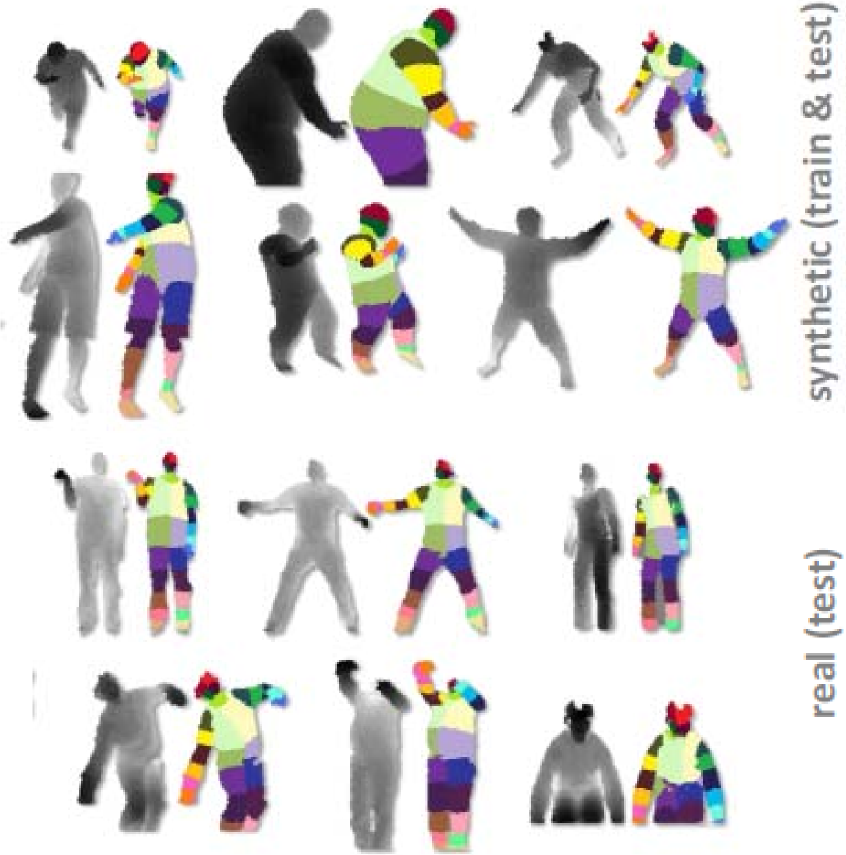
Learns from lots of noisy annotations

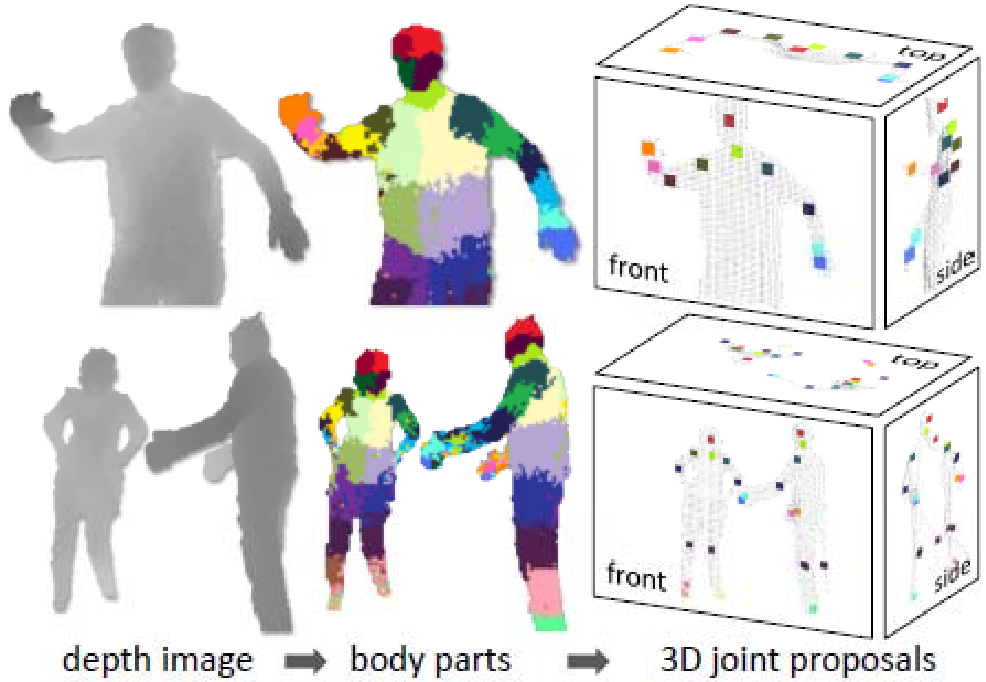


B. Sapp, D.Weiss and B. Taskar. Parsing Human Motion with Stretchable Models. In Proc. **CVPR 2011**.

Explores temporal continuity

Pose estimation is a very active research area

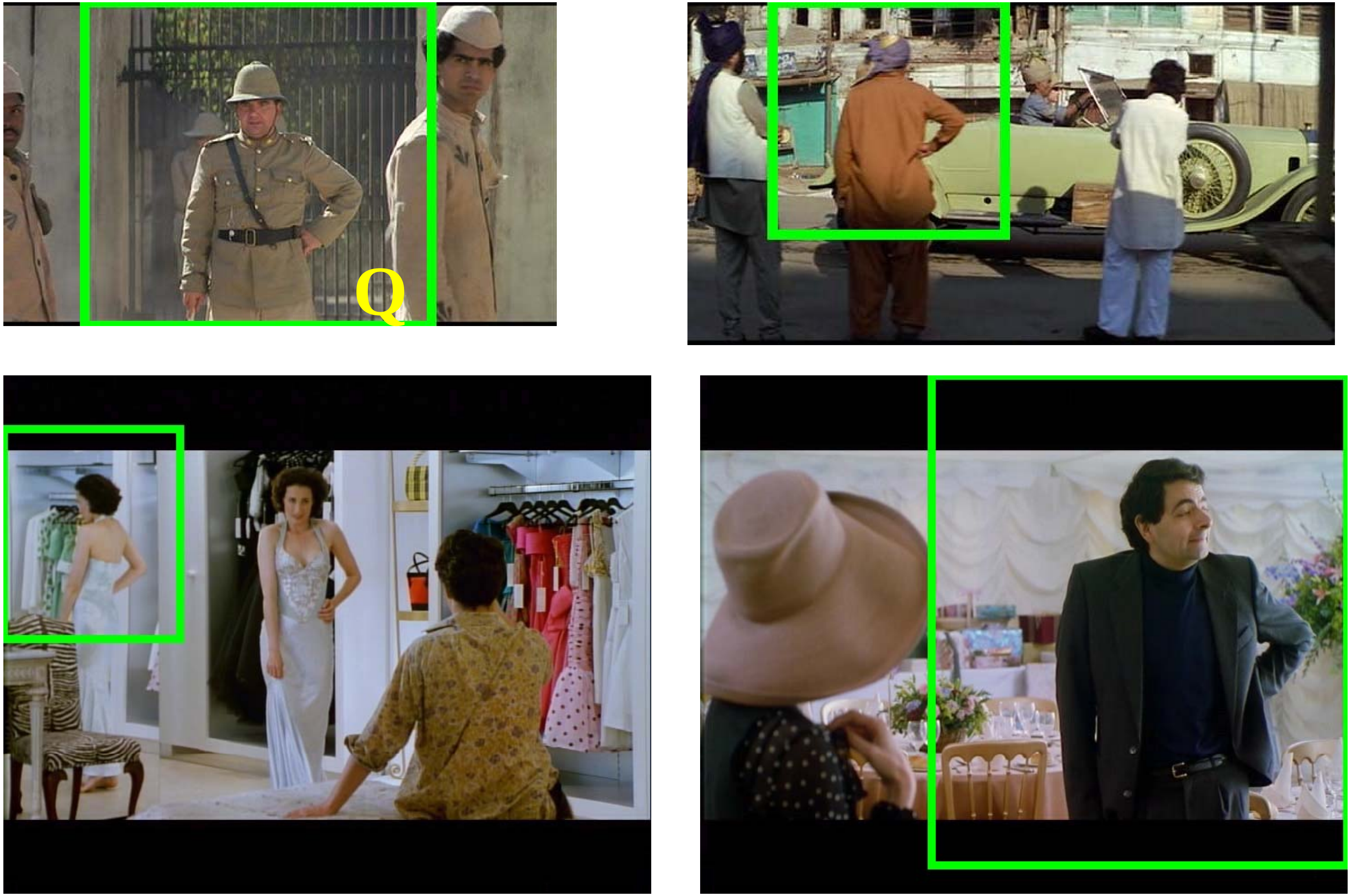




J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. **Best paper award at CVPR 2011**

Exploits lots of synthesized depth images for training

Pose Search



V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In Proc. CVPR2009

Pose Search

V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In Proc. CVPR2009

Application

Learning sign language by watching TV (using weakly aligned subtitles)

Patrick Buehler

Mark Everingham

Andrew Zisserman

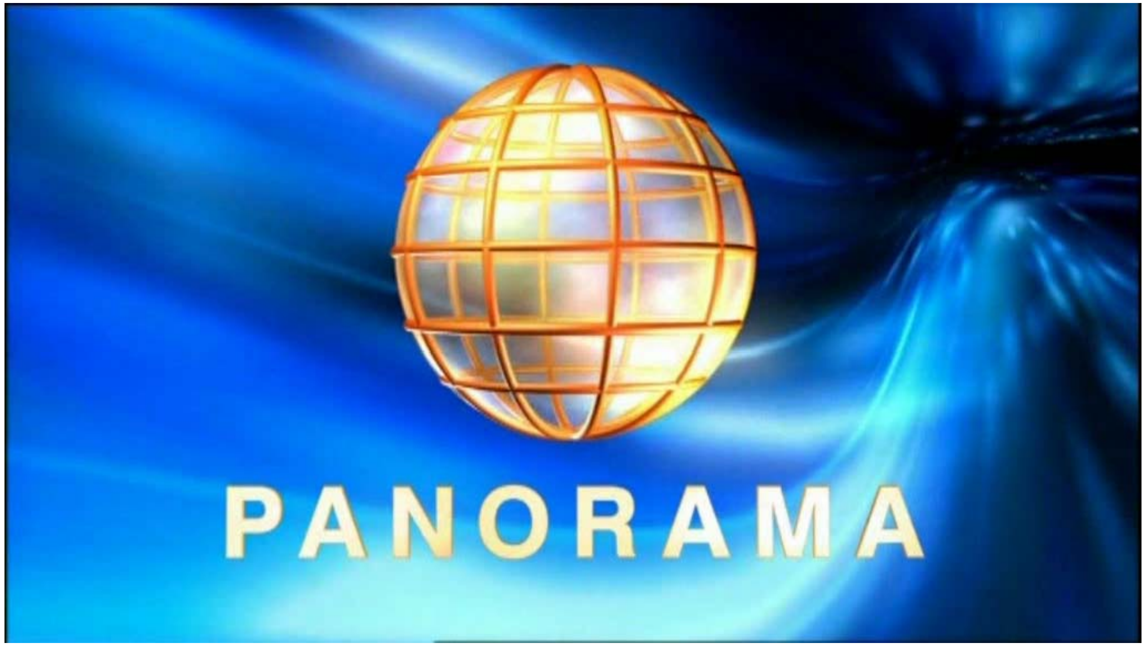
CVPR 2009

Objective

Learn signs in British Sign Language (BSL) corresponding to text words:

- Training data from TV broadcasts with simultaneous signing
- Supervision solely from sub-titles

Input: video + subtitle



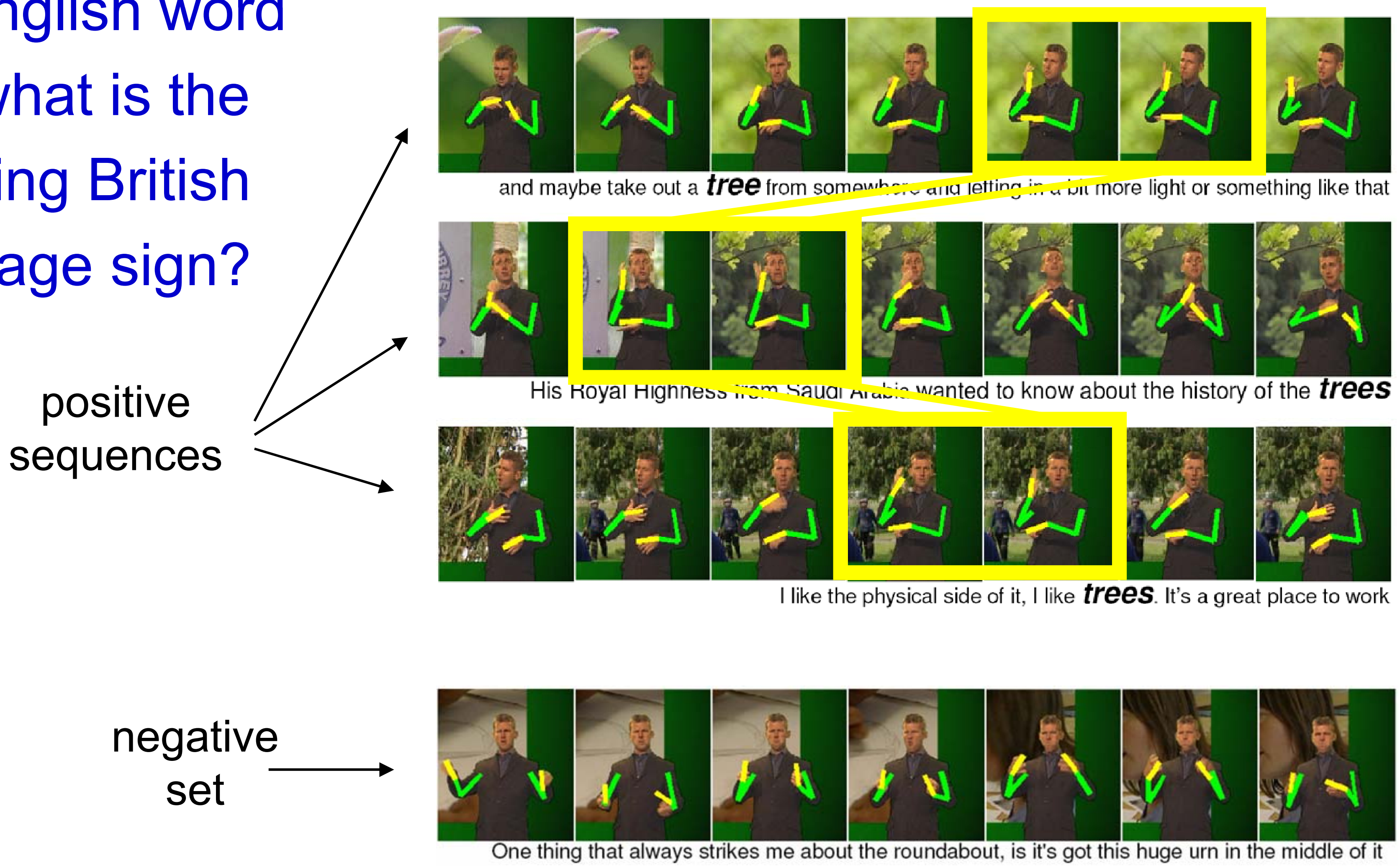
Output: automatically learned signs (4x slow motion)

Office

Government

Use subtitles to find video sequences containing word. These are the positive training sequences. Use other sequences as negative training sequences.

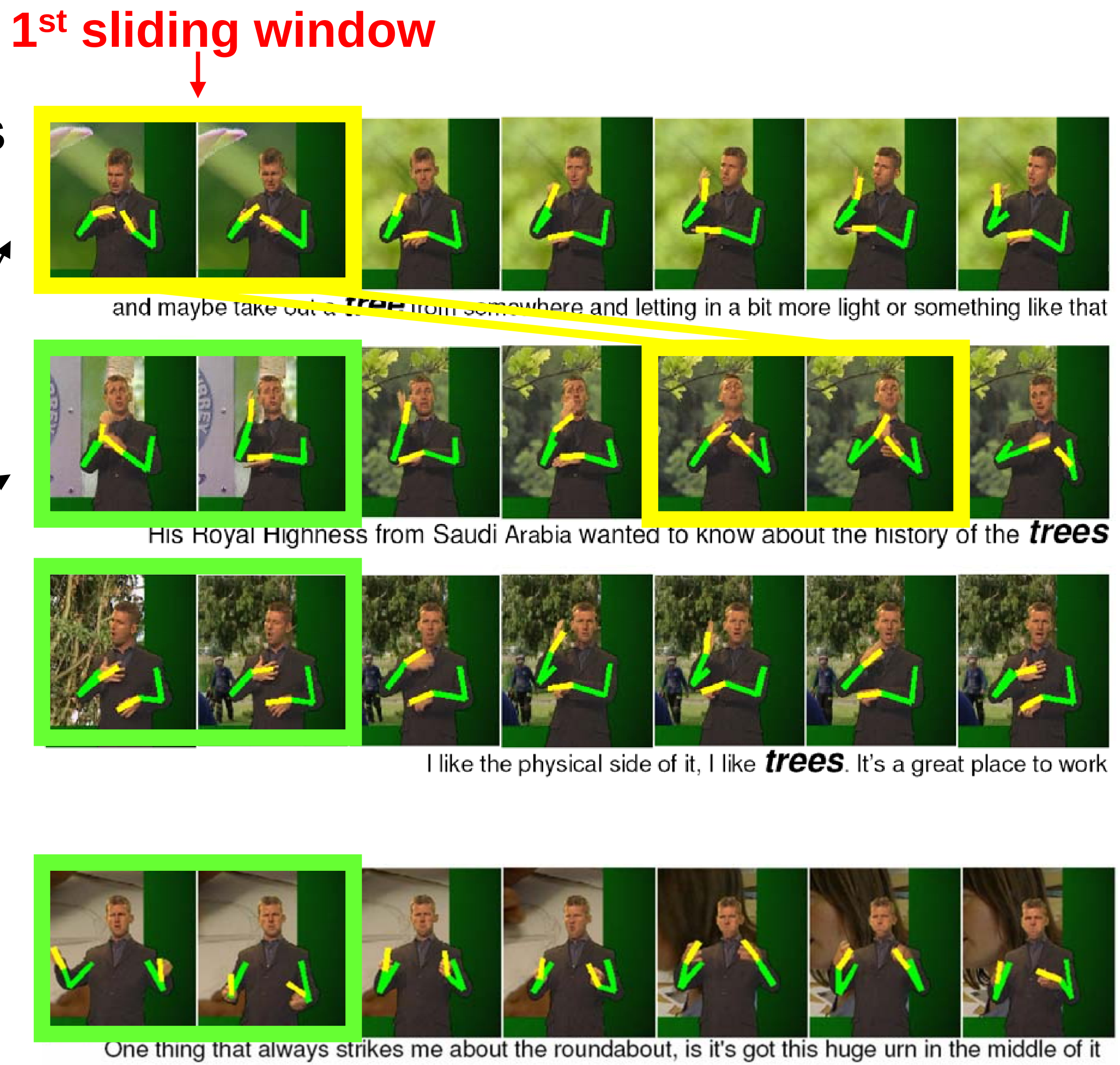
Given an English word e.g. “tree” what is the corresponding British Sign Language sign?



Use sliding window to choose sub-sequence of poses in one positive sequence and determine if

same sub-sequence of poses occurs in other positive sequences somewhere, but

does not occur in the negative set



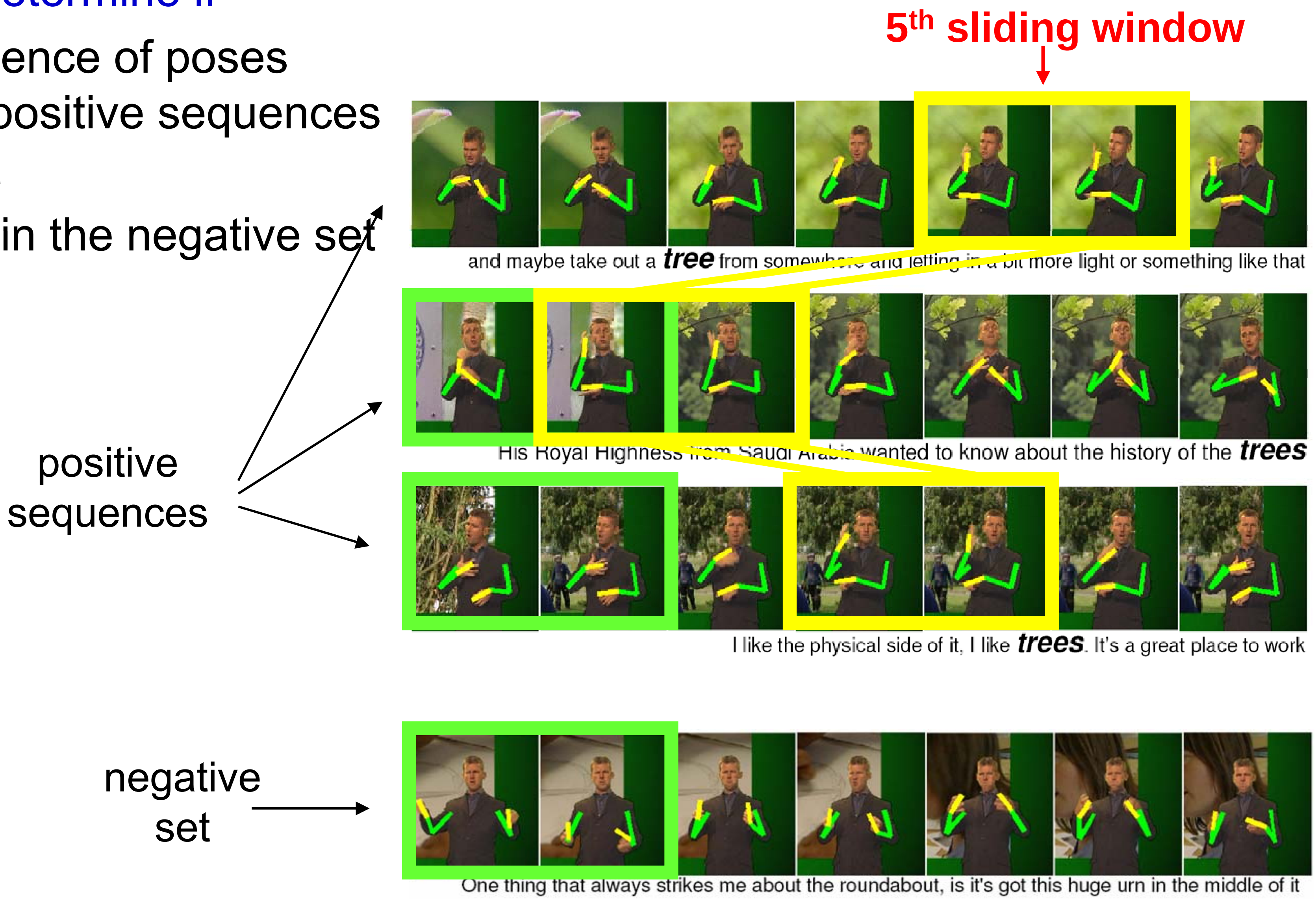
positive sequences

negative set

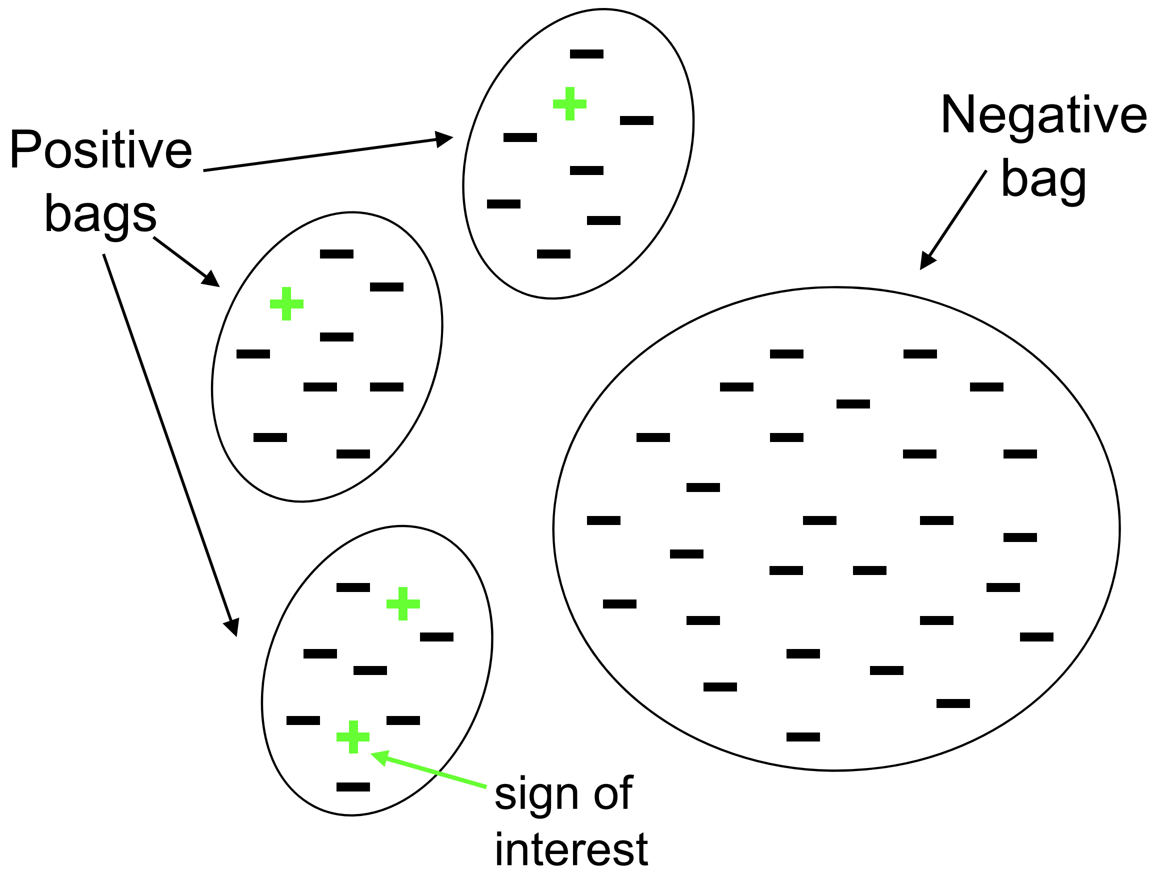
Use sliding window to choose sub-sequence of poses in one positive sequence and determine if

same sub-sequence of poses occurs in other positive sequences somewhere, but

does not occur in the negative set

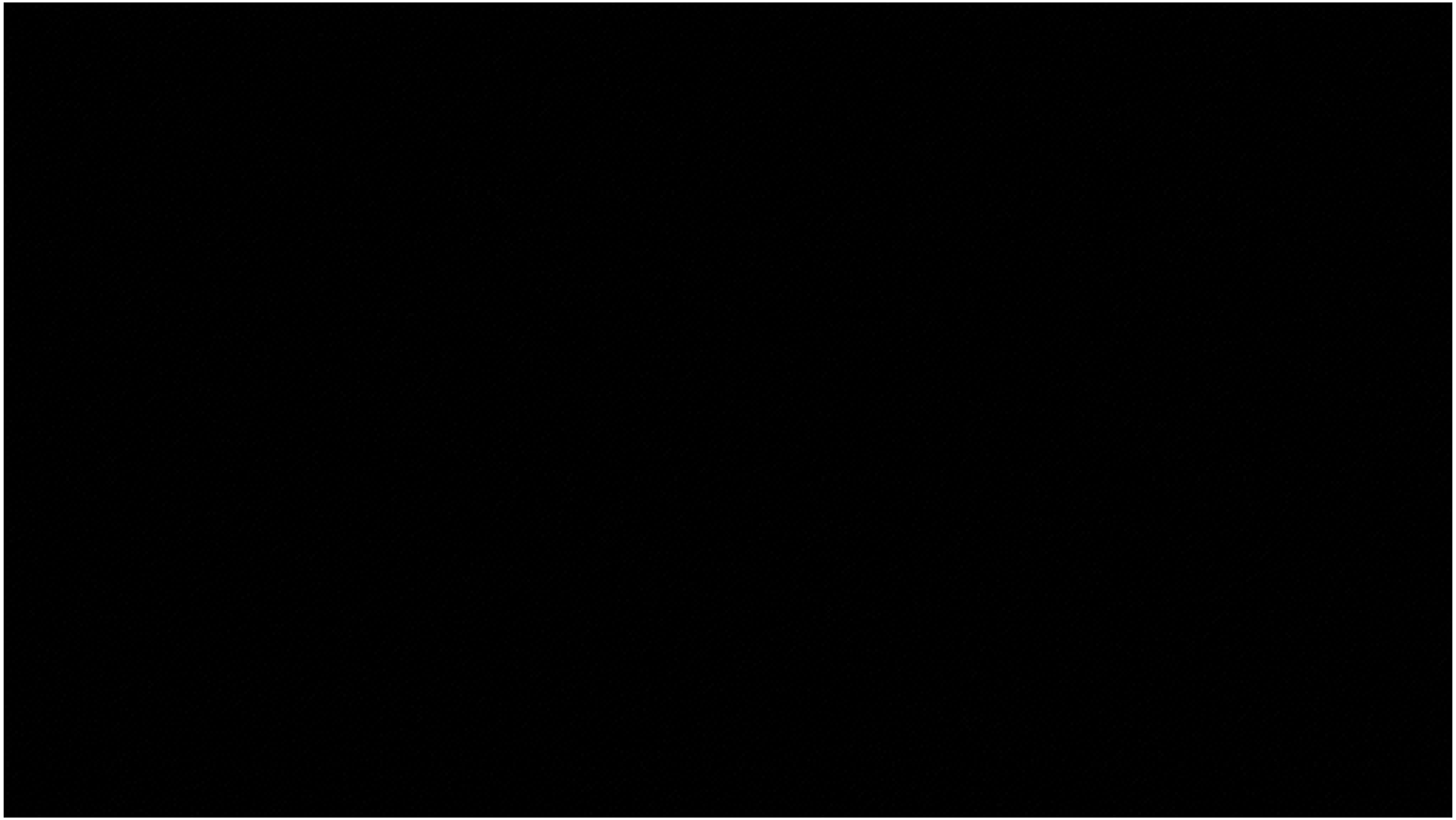


Multiple instance learning



Example

Learn signs in British Sign Language (BSL) corresponding to text words.



Evaluation

Good results for a variety of signs:

Signs where hand movement is important

Navy

Prince

Signs where hand shape is important

Lung

Garden

Signs where both hands are together

Fungi

Golf

Signs which are finger--spelled

Kew

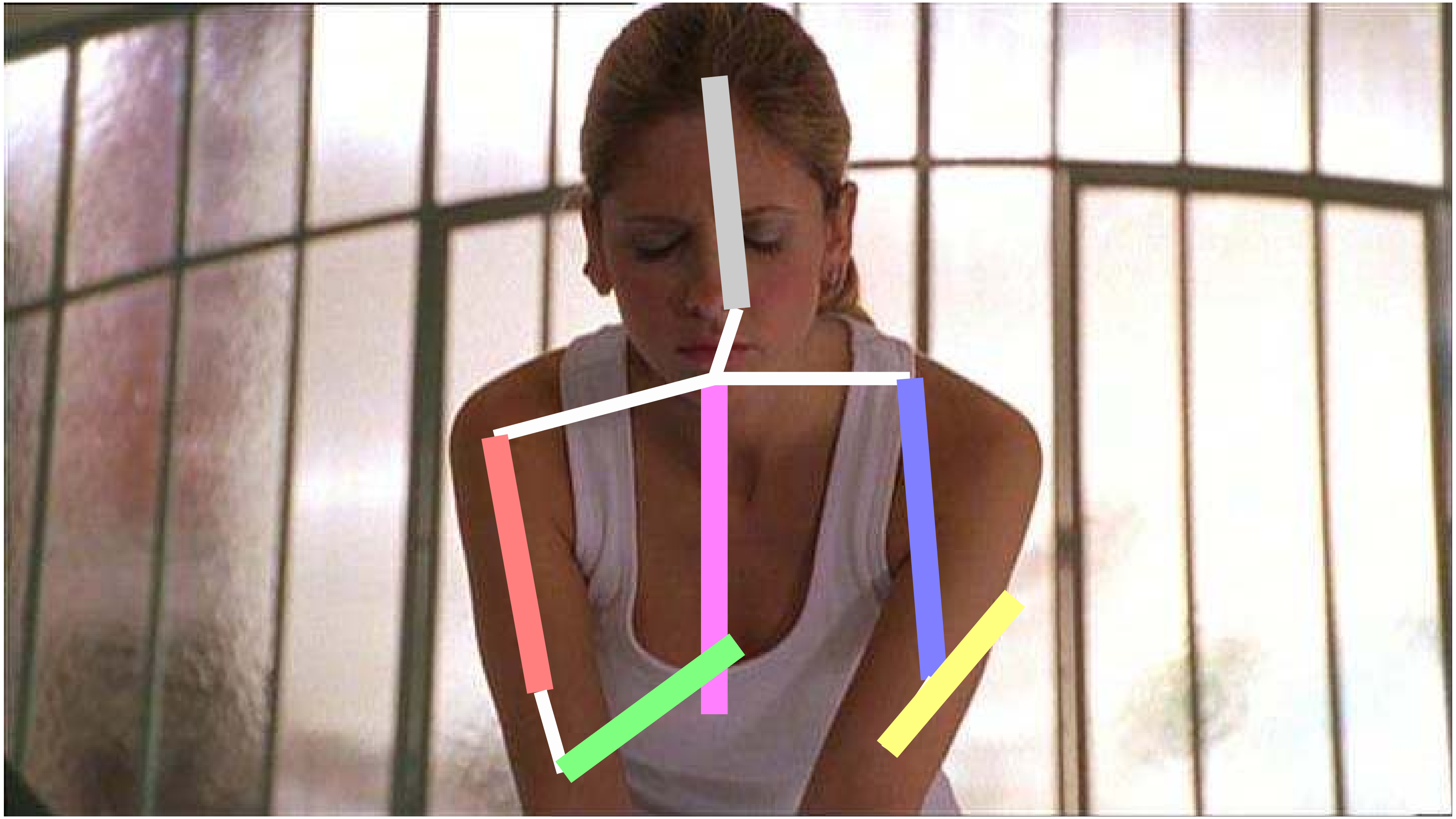
Bob

Signs which are perfomed in front of the face

Whale

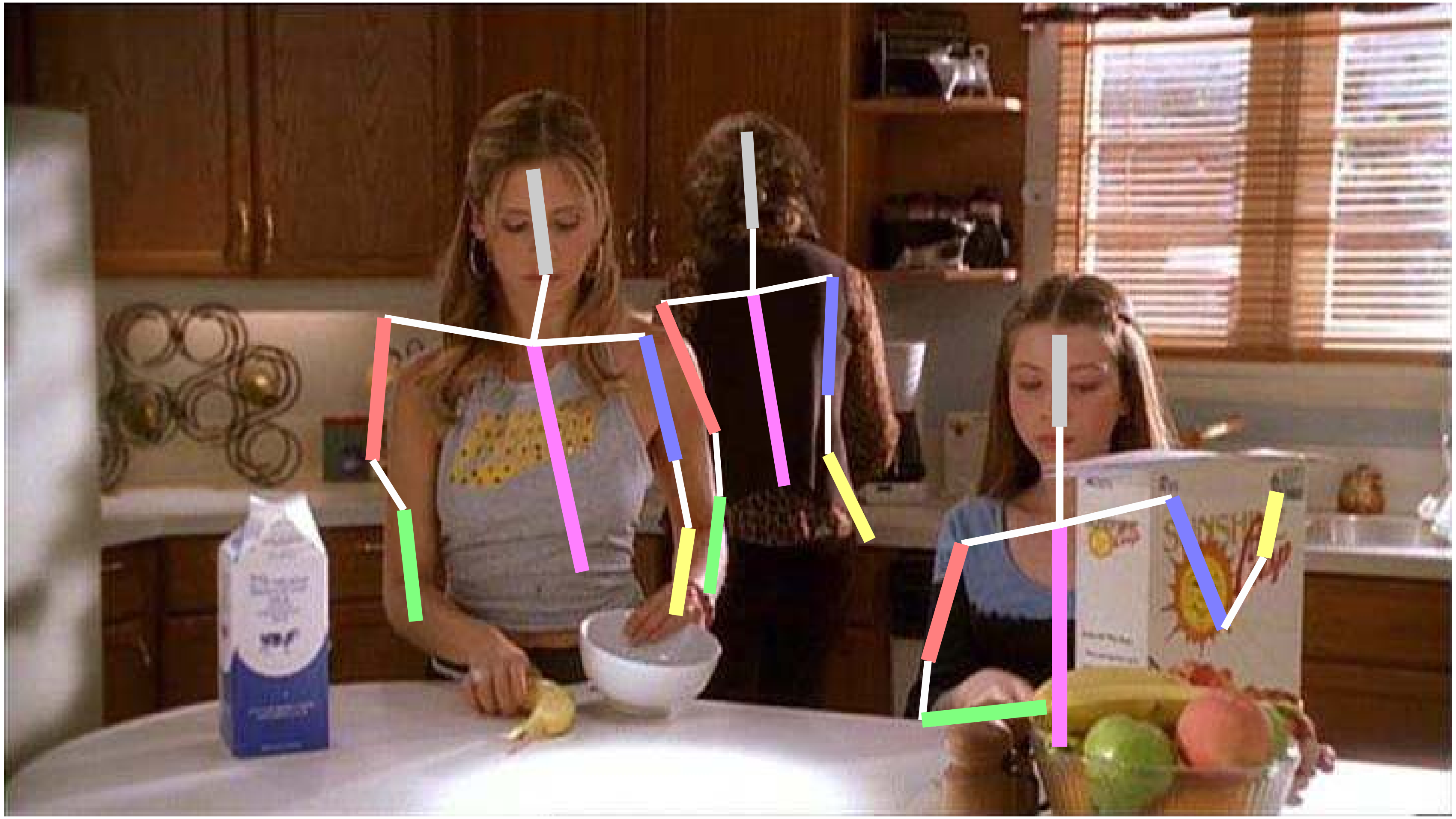
Rose

What is missed?



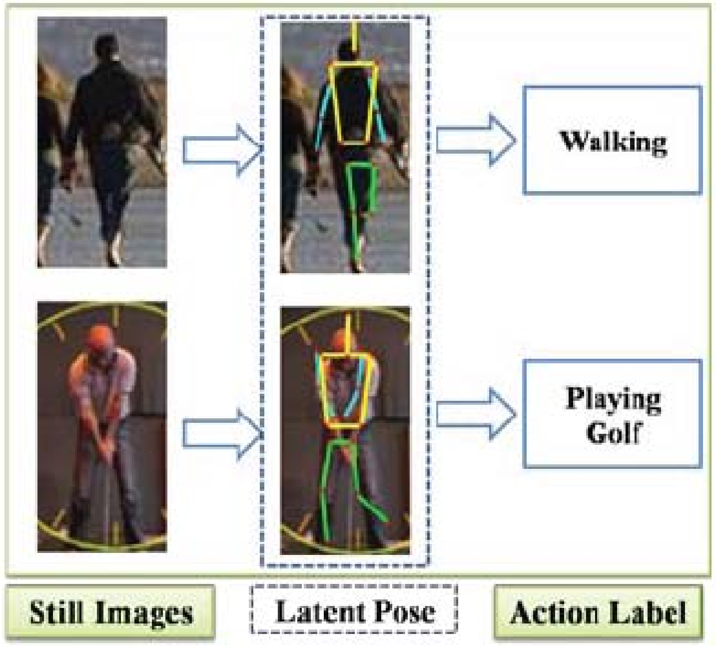
truncation is not modelled

What is missed?



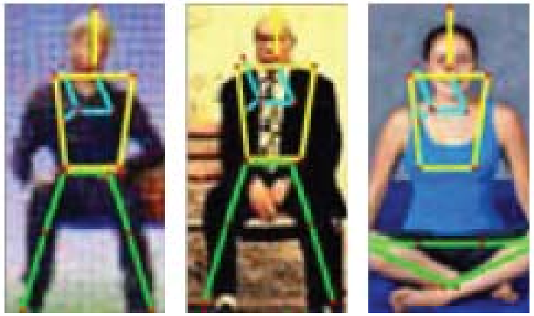
occlusion is not modelled

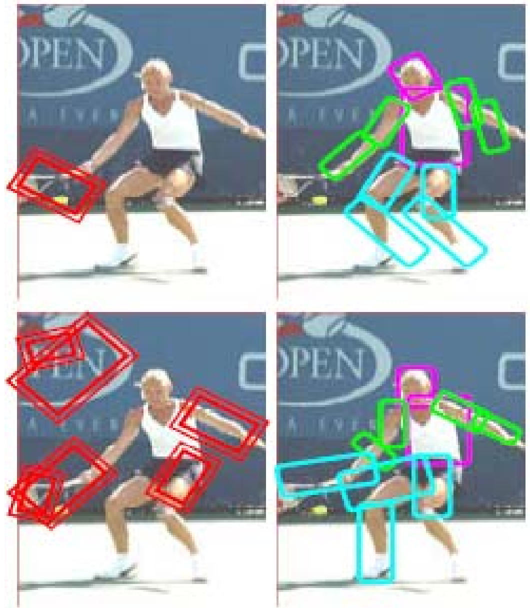
Modelling person-object-pose interactions



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)

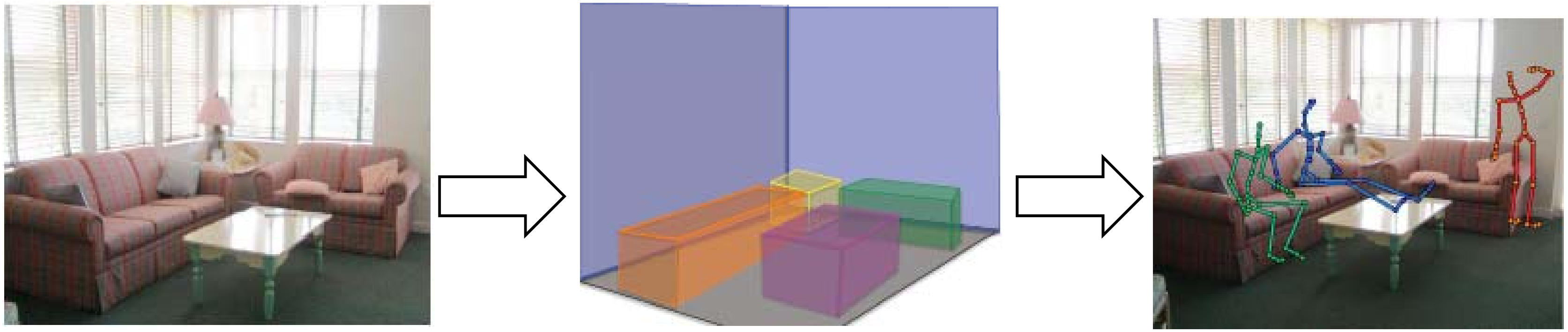




B. Yao and L. Fei-Fei. Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities. In Proc. CVPR 2010.

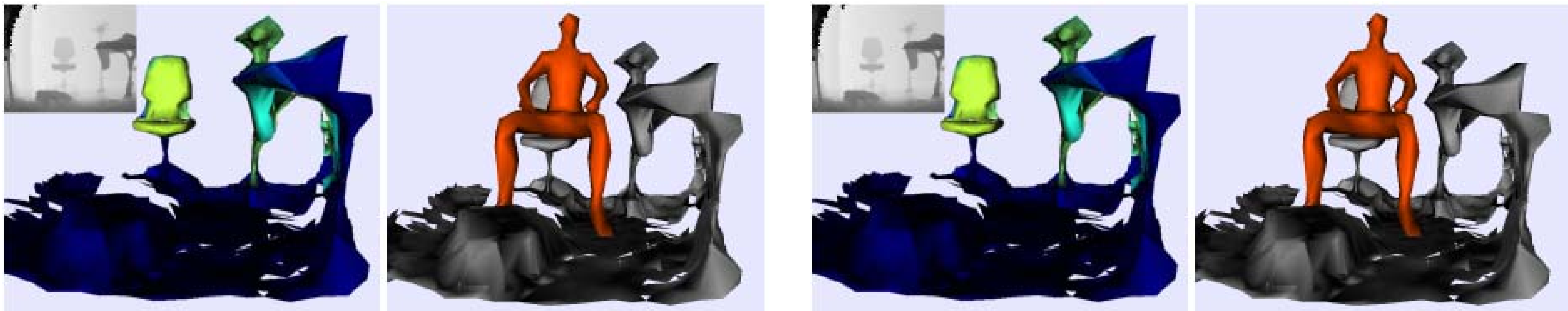
Pose estimation helps object detection and vice versa

Towards functional object understanding



A. Gupta, S. Satkin, A.A. Efros and M. Hebert, From 3D Scene Geometry to HumanWorkspace. In Proc. CVPR 2011

Predicts the “workspace” of a human

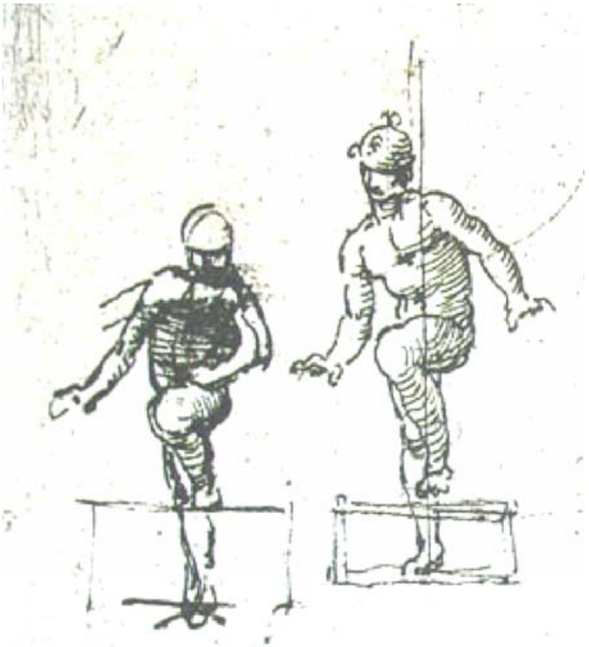


H. Grabner, J. Gall and L. Van Gool. What Makes a Chair a Chair? In Proc. CVPR 2011

Conclusions: 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

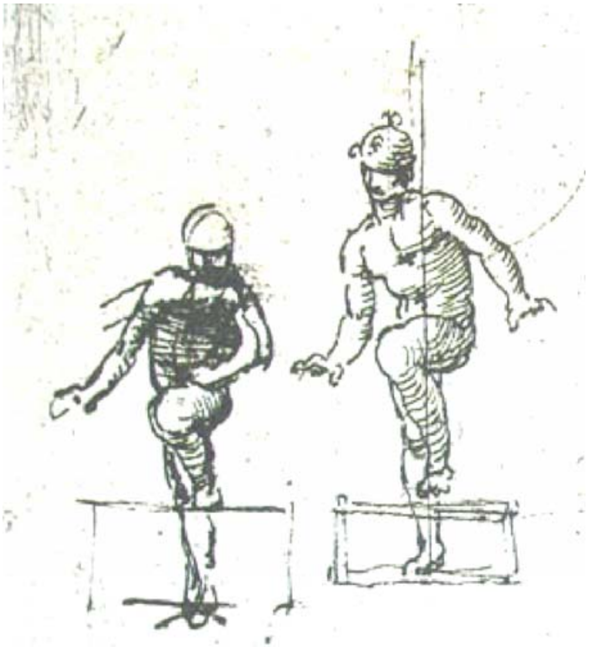
Motion-based methods

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

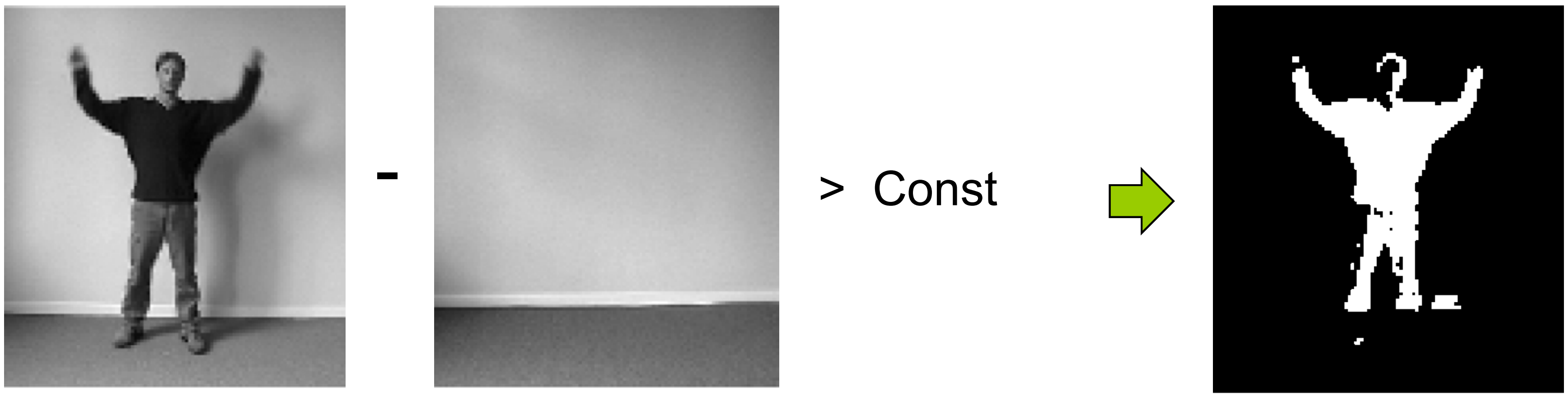
Generic and parametric Optical Flow
Motion templates

Space-time methods

Space-time features
Training with weak supervision

Foreground segmentation

Image differencing: a simple way to measure motion/change

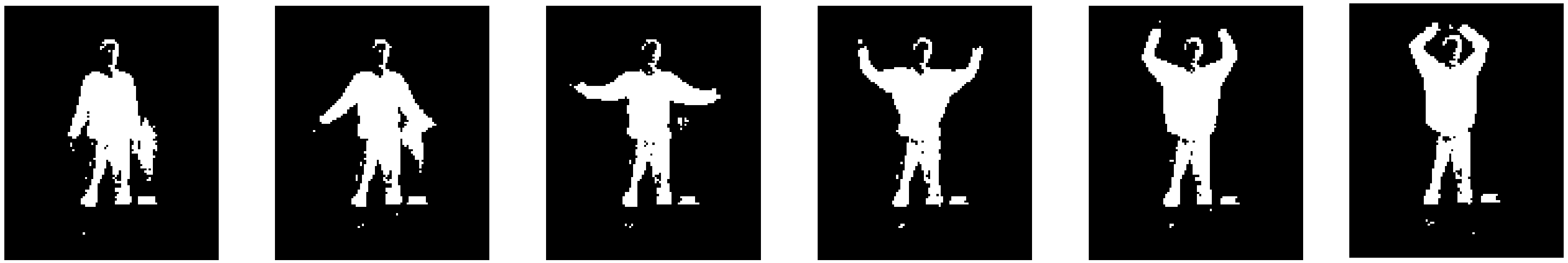


Better Background / Foreground separation methods exist:

- Modeling of color variation at each pixel with Gaussian Mixture
- Dominant motion compensation for sequences with moving camera
- Motion layer separation for scenes with non-static backgrounds

Temporal Templates

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



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

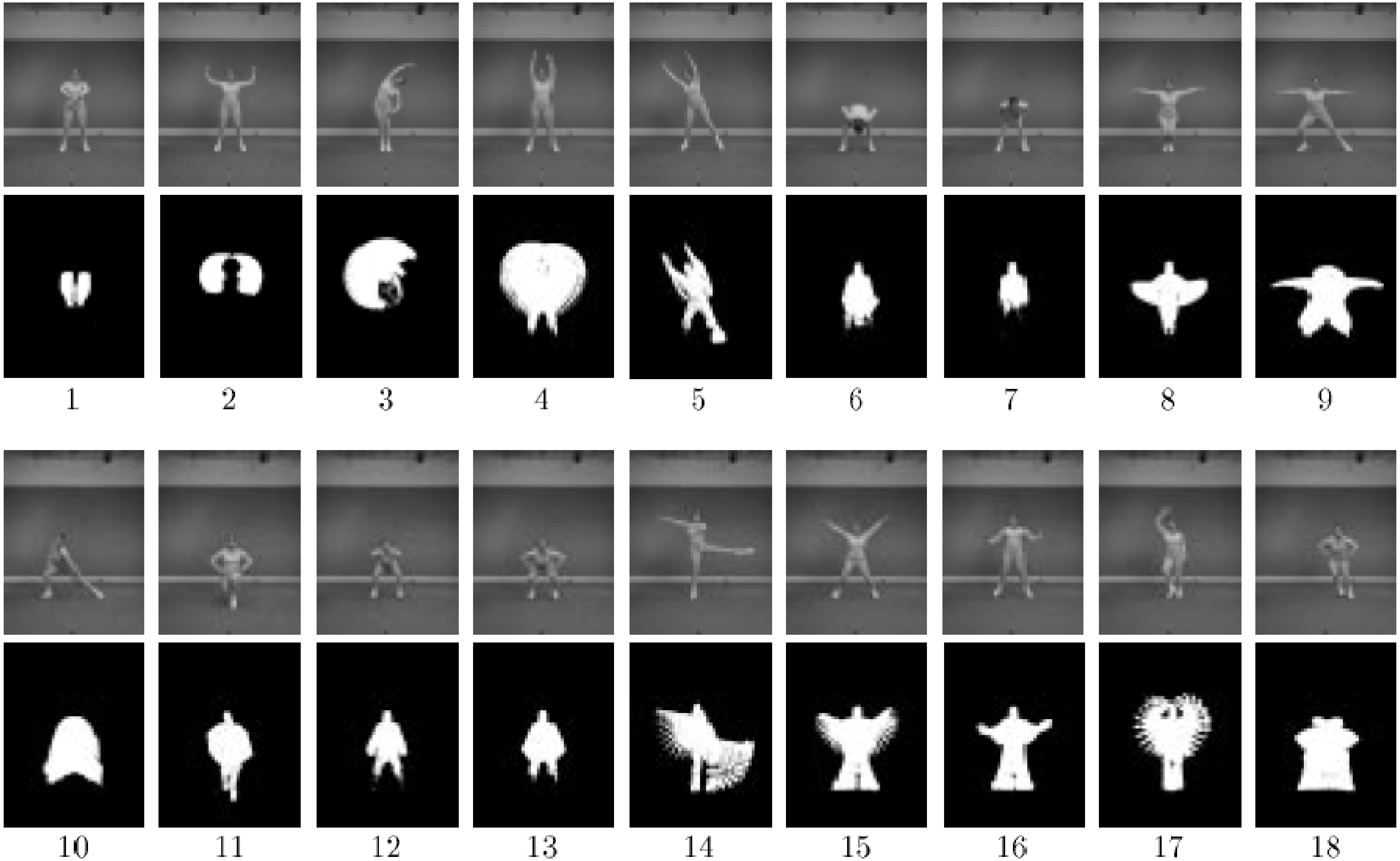
$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ \max\ (0, H_\tau(x, y, t-1) - 1) & \text{otherwise} \end{cases}$$

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

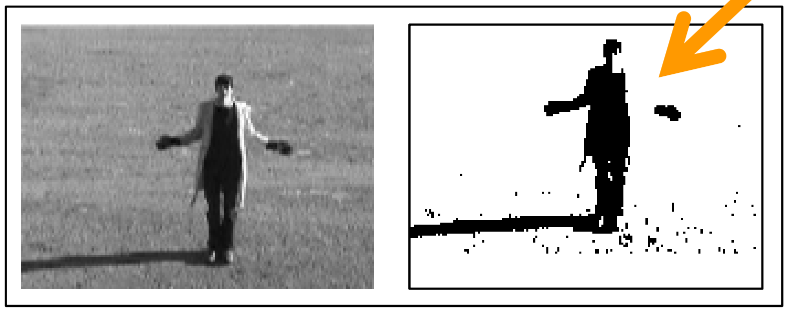
Pros:

+ 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 [Cootes et al.]

- Constrains shape deformation in PCA-projected space

Example: face alignment

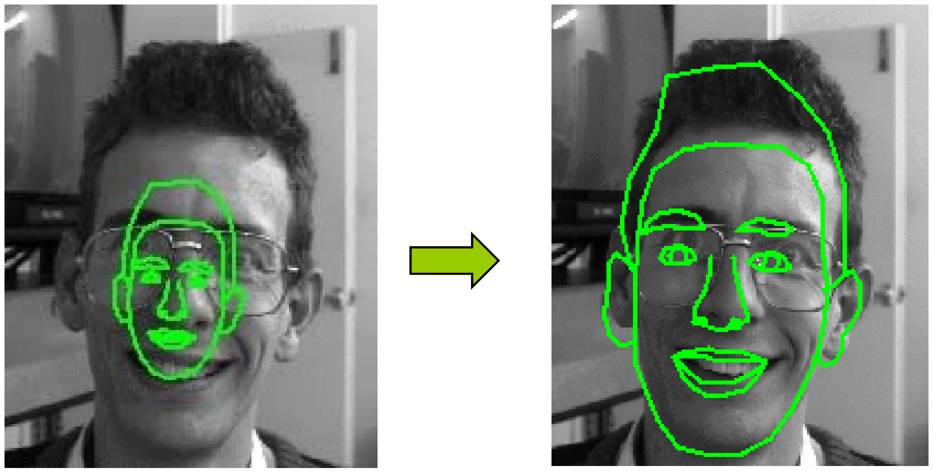
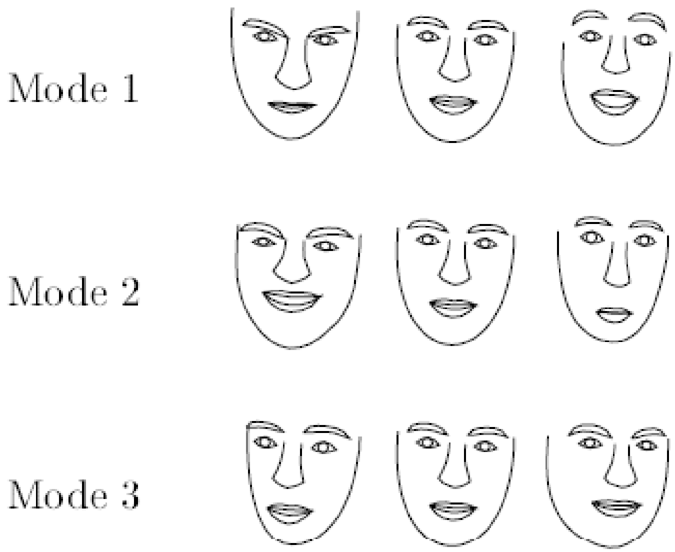


Illustration of face shape space



Active Shape Models: Their Training and Application

T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, **CVIU** 1995

Person Tracking

Learning flexible models from image sequences
A. Baumberg and D. Hogg, **ECCV** 1994

Learning dynamic prior

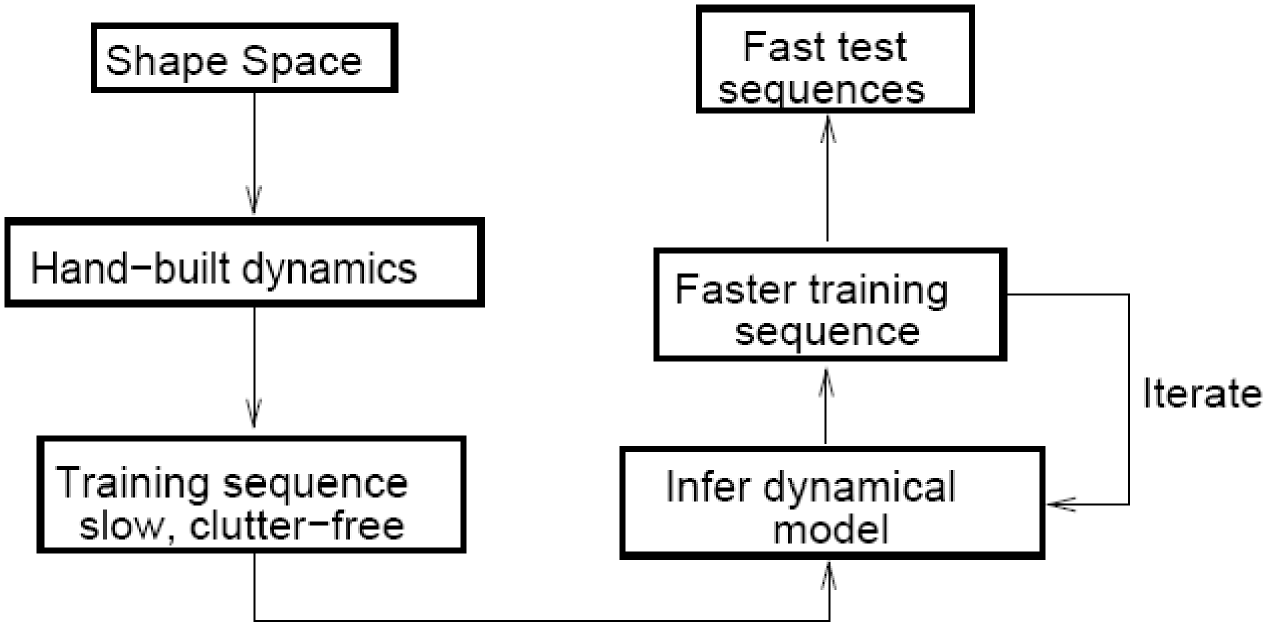
- Dynamic model: 2nd order Auto-Regressive Process

State $$\mathcal{X}_k = \begin{pmatrix} \mathbf{X}_{k-1} \\ \mathbf{X}_k \end{pmatrix}$$

Update rule: $$\mathcal{X}_k - \overline{\mathcal{X}} = A(\mathcal{X}_{k-1} - \overline{\mathcal{X}}) + B\mathbf{w}_k$$

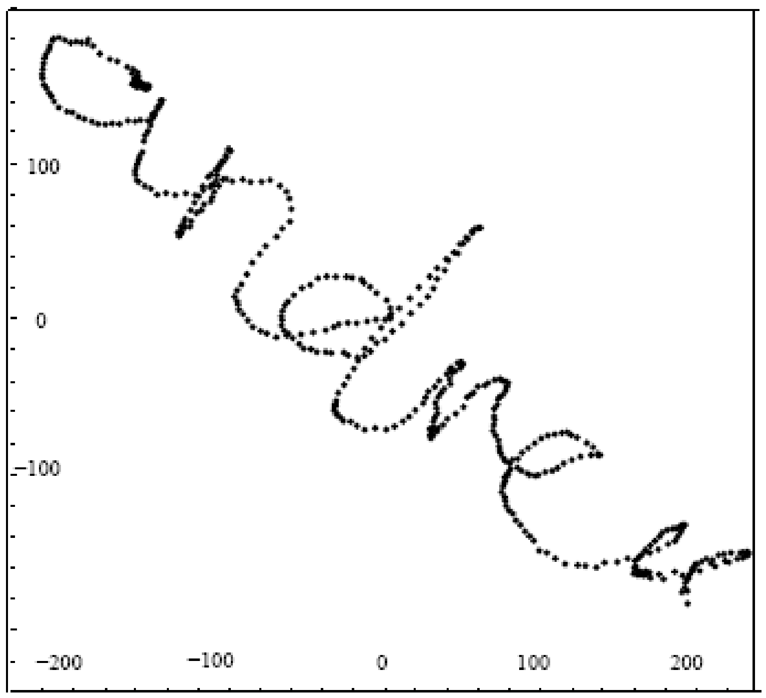
Model parameters: $$A = \begin{pmatrix} 0 & I \\ A_2 & A_1 \end{pmatrix}, \quad \overline{\mathcal{X}} = \begin{pmatrix} \overline{\mathbf{X}} \\ \overline{\mathbf{X}} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 \\ B_0 \end{pmatrix}$$

Learning scheme:

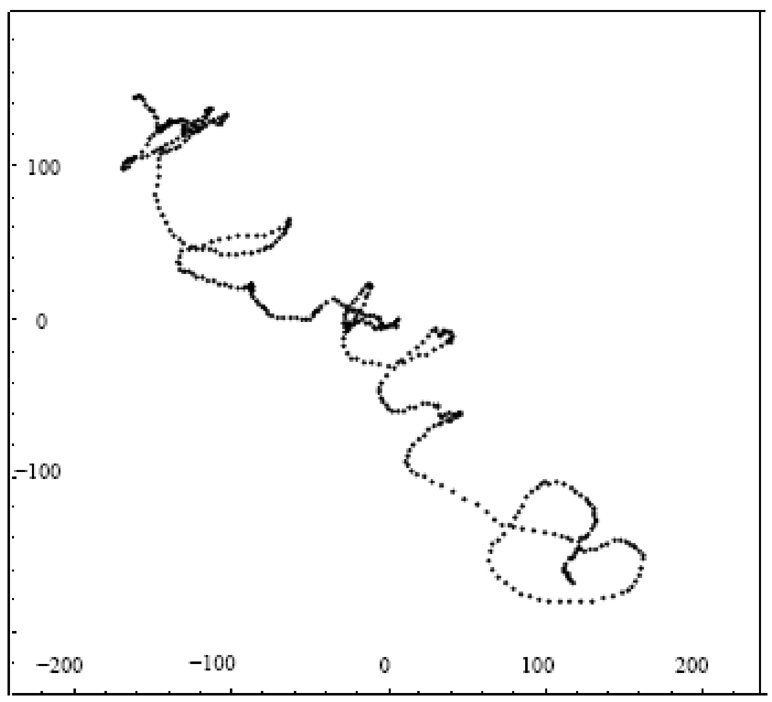


Learning dynamic prior

Learning point sequence



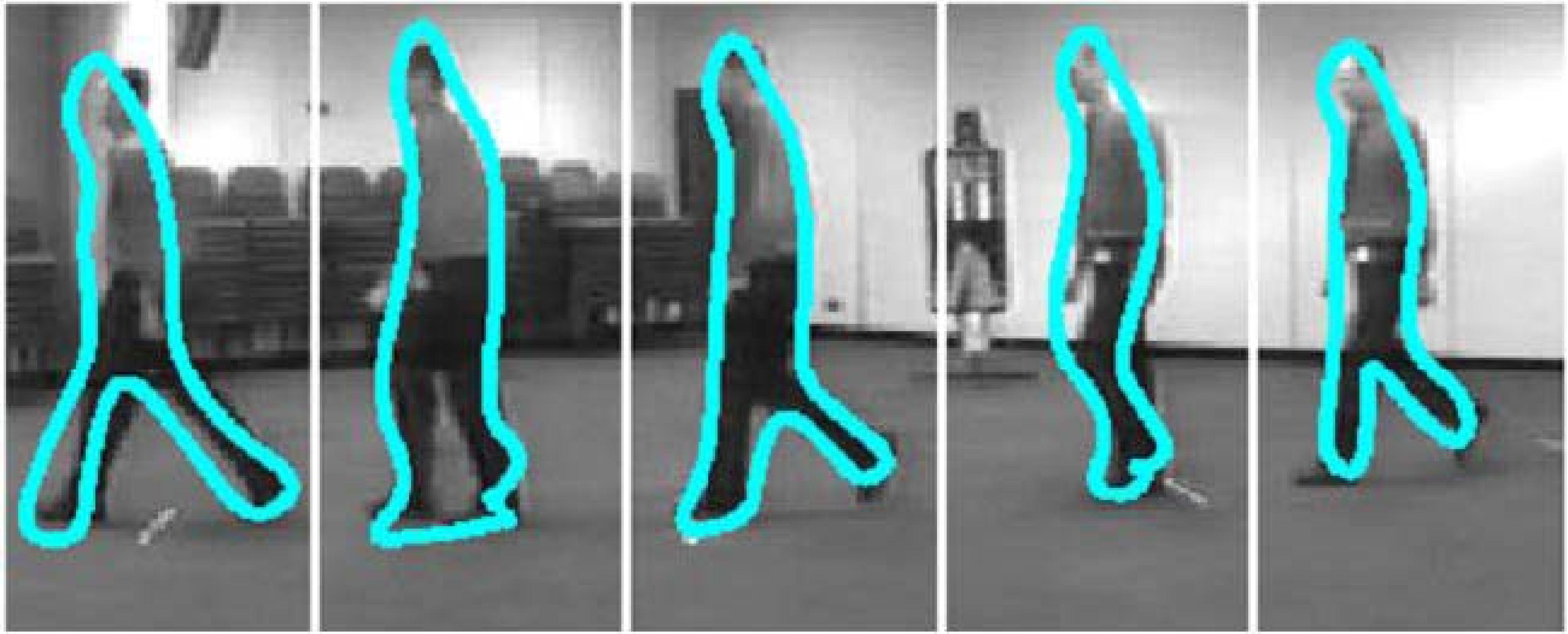
Random simulation of the learned dynamical model



Statistical models of visual shape and motion
A. Blake, B. Bascle, M. Isard and J. MacCormick, **Phil.Trans.R.Soc. 1998**

Learning dynamic prior

Random simulation of the learned gate dynamics



Motion priors

- Constrain temporal evolution of shape
  - ❖ Help accurate tracking
  - ❖ Recognize actions

- Goal: formulate motion models for different types of actions and use such models for action recognition

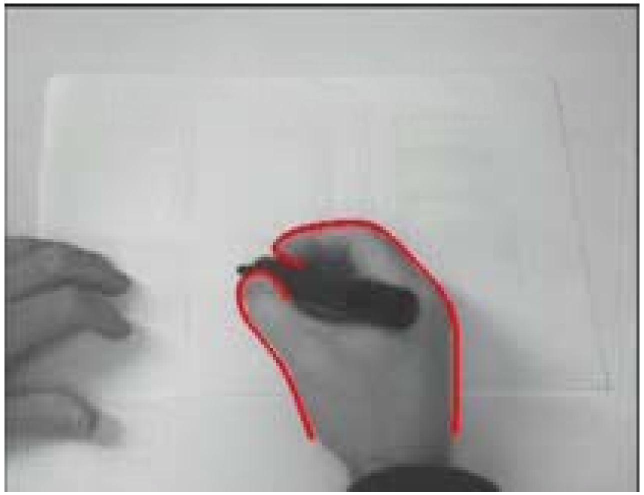
Example:

Drawing with 3 action modes

line drawing

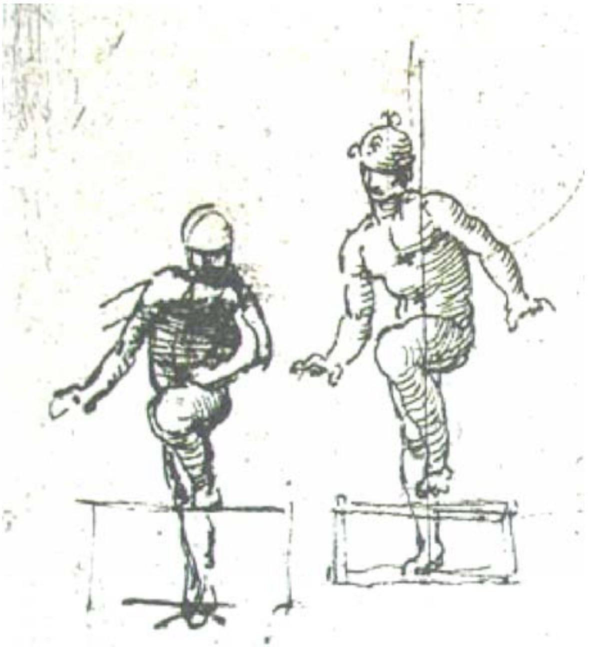
scribbling

idle



[M. Isard and A. Blake, ICCV 1998]

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

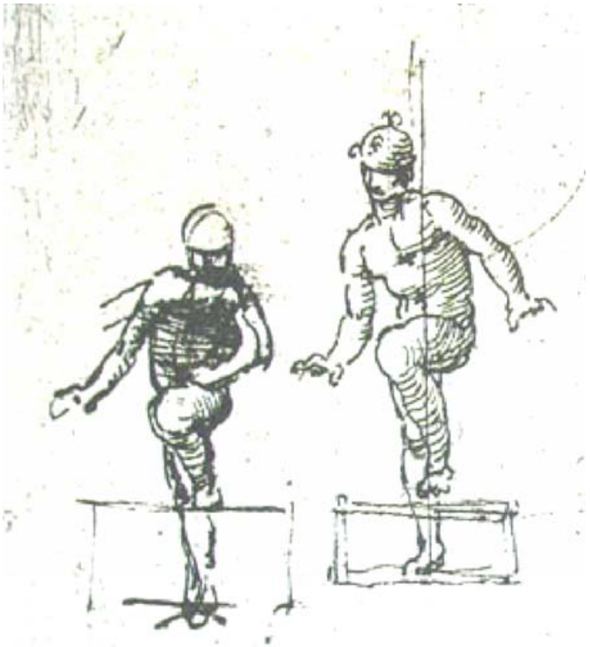
Motion-based methods

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

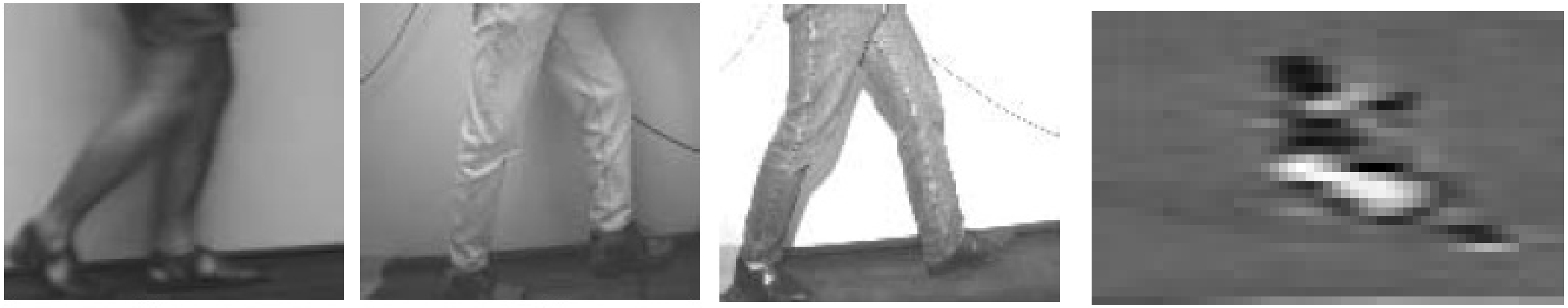
Generic and parametric Optical Flow
Motion templates

Space-time methods

Space-time features
Training with weak supervision

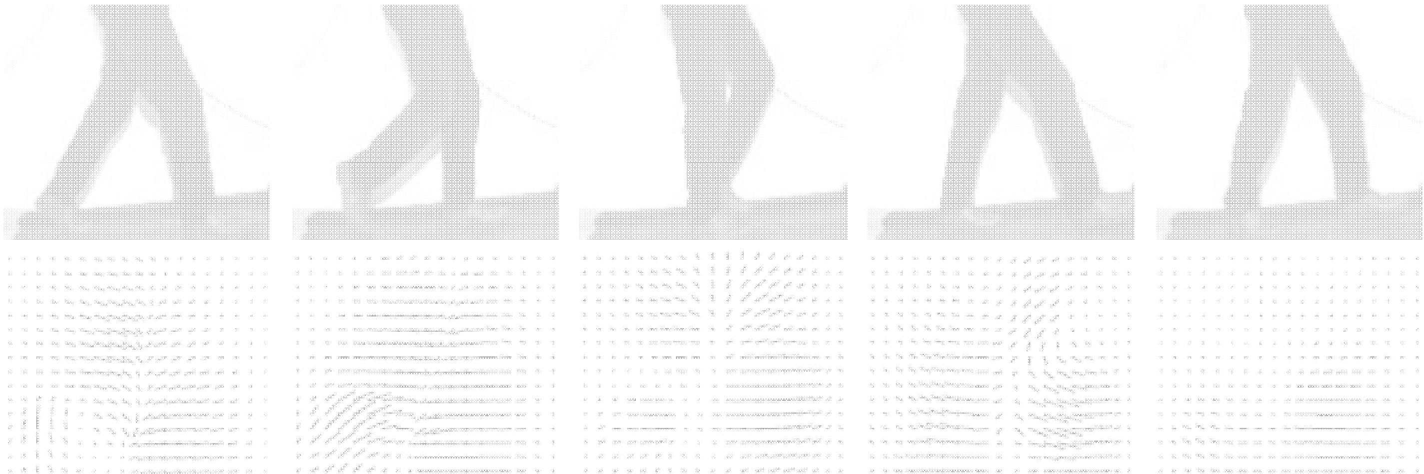
Shape and Appearance vs. Motion

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



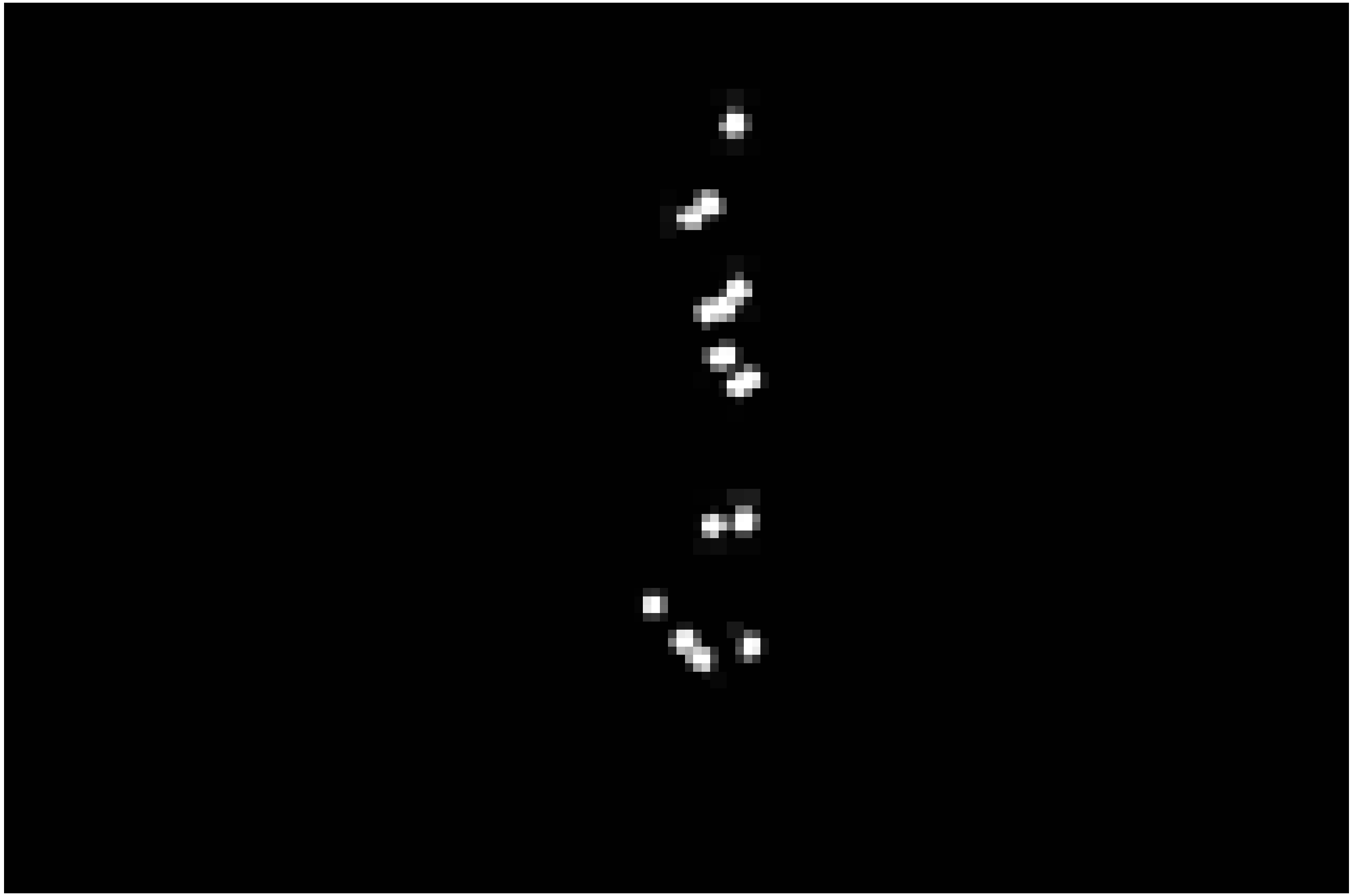
[Efros et al. 2003]

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



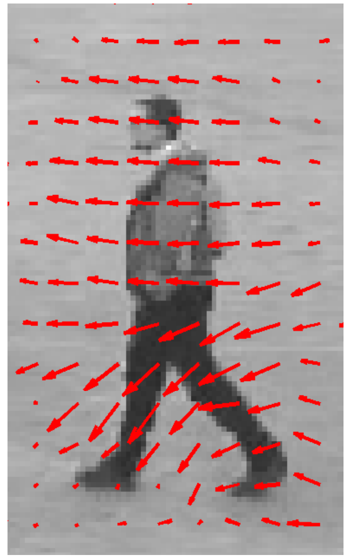
Shape and Appearance vs. Motion

Moving Light Displays



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

Motion estimation: Optical Flow



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

  How? We only have access to image pixels

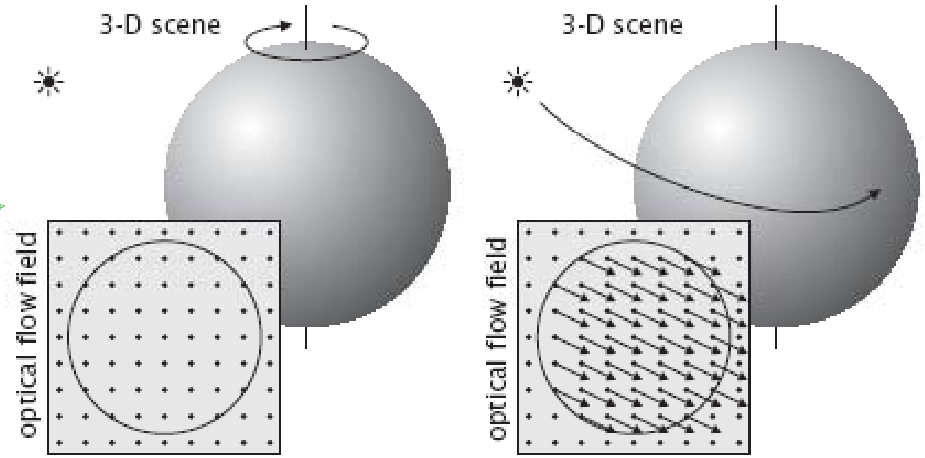
  ⇨ Estimate pixel-wise correspondence between frames = Optical Flow

- 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

Parameterized Optical Flow

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

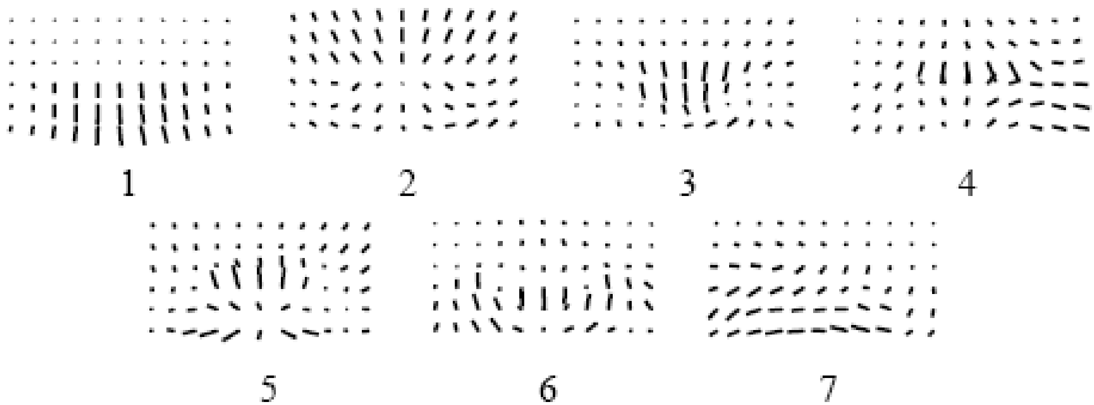
  1. Compute standard Optical Flow for many examples
  2. Put velocity components into one vector

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

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

Training samples

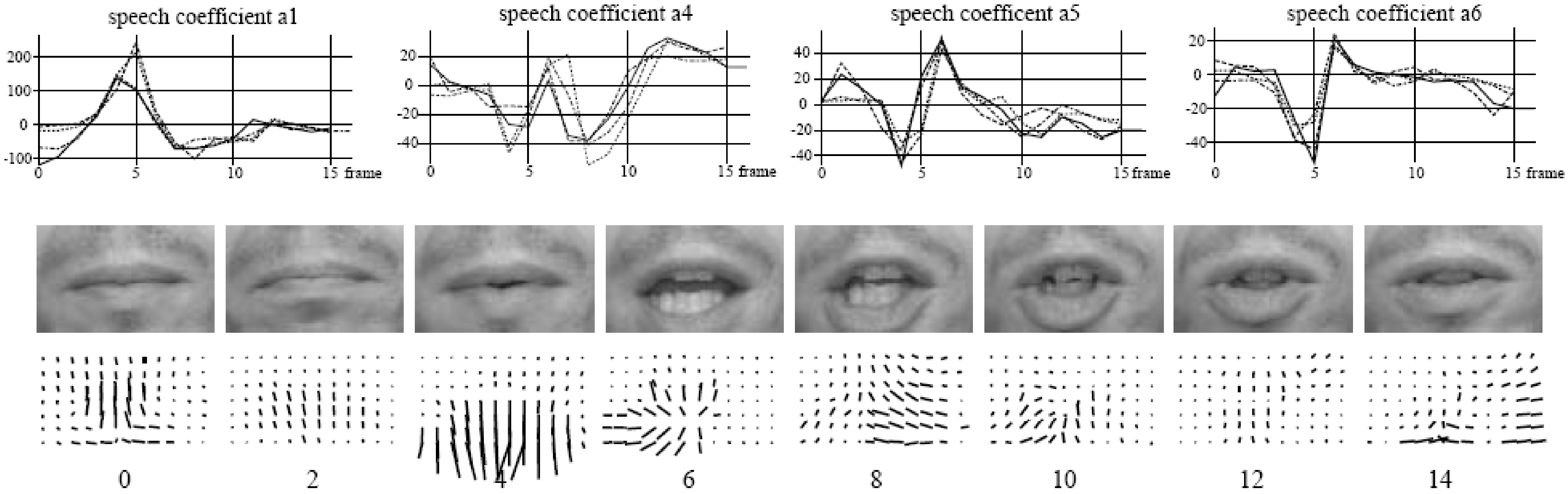
PCA flow bases



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

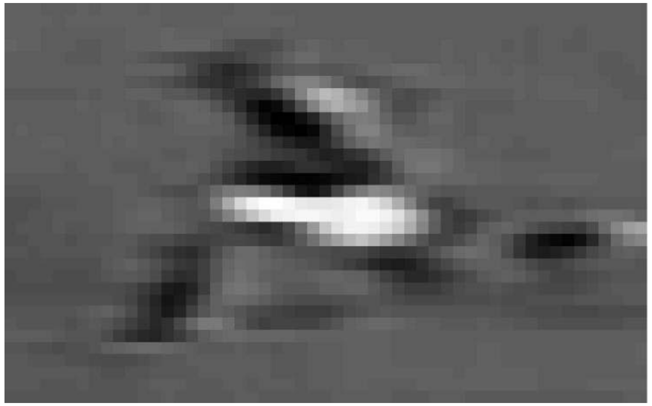
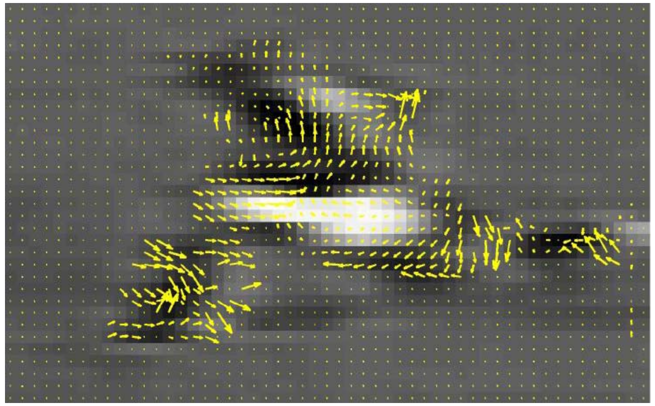
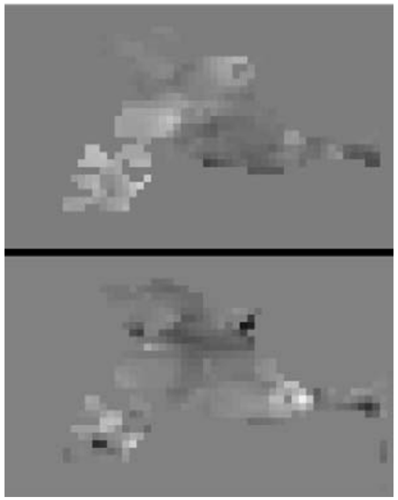


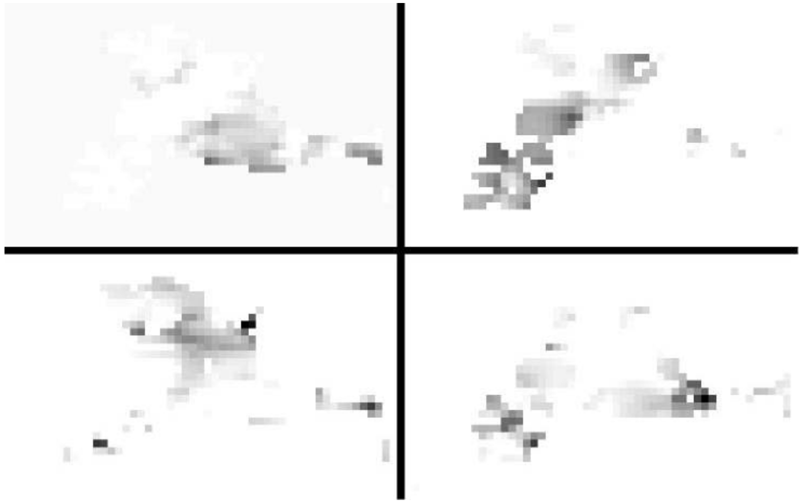
Image frame



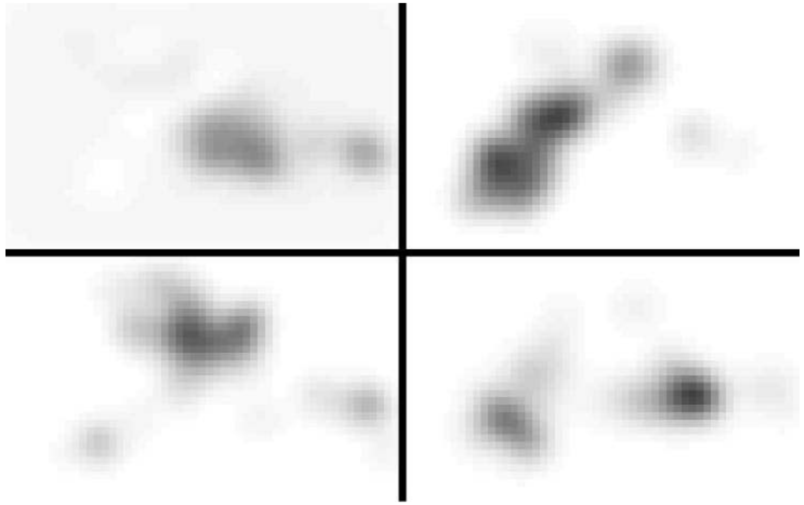
Optical flow $F_{x,y}$



F_x, F_y



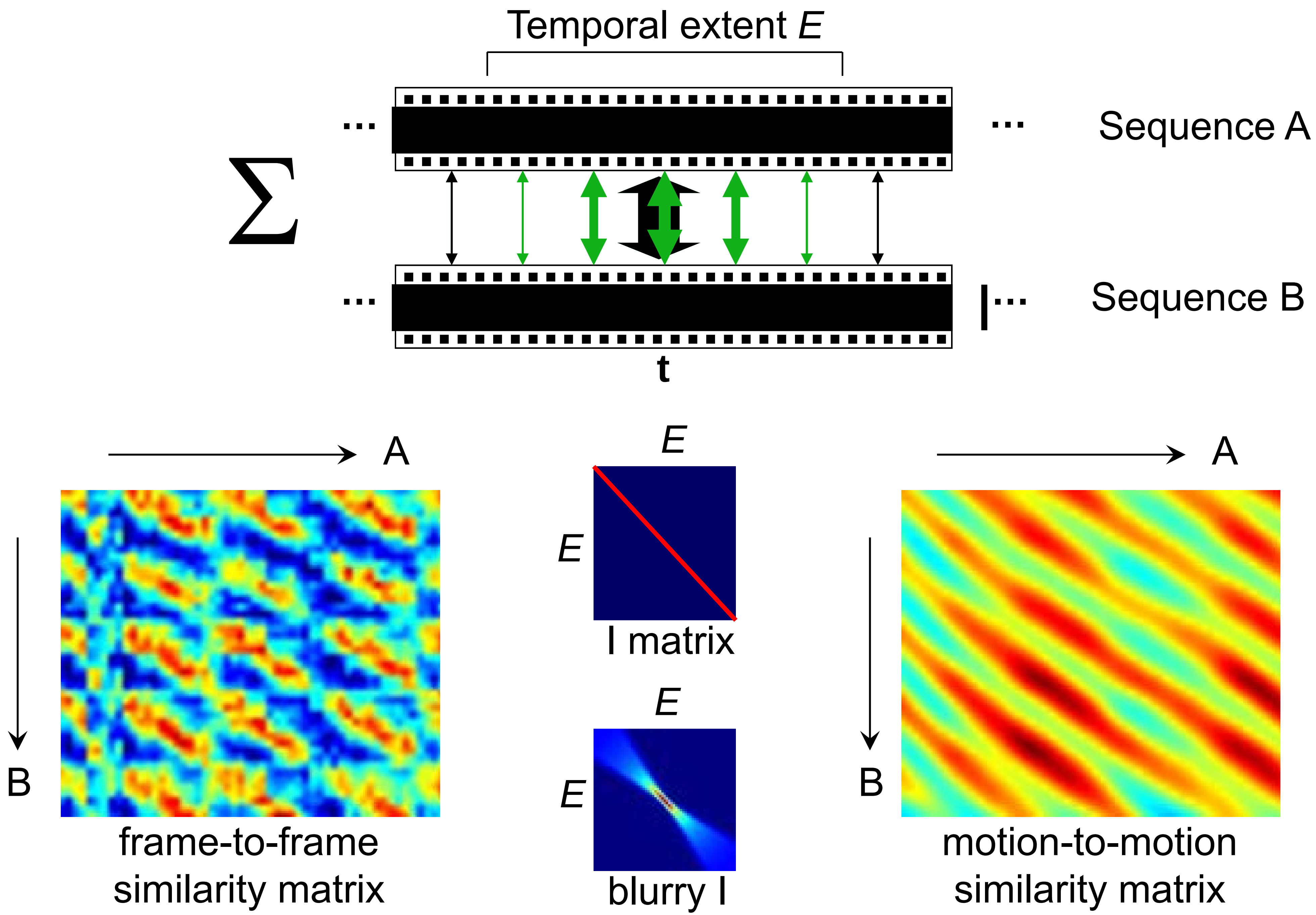
$F_x^-, F_x^+, F_y^-, F_y^+$



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

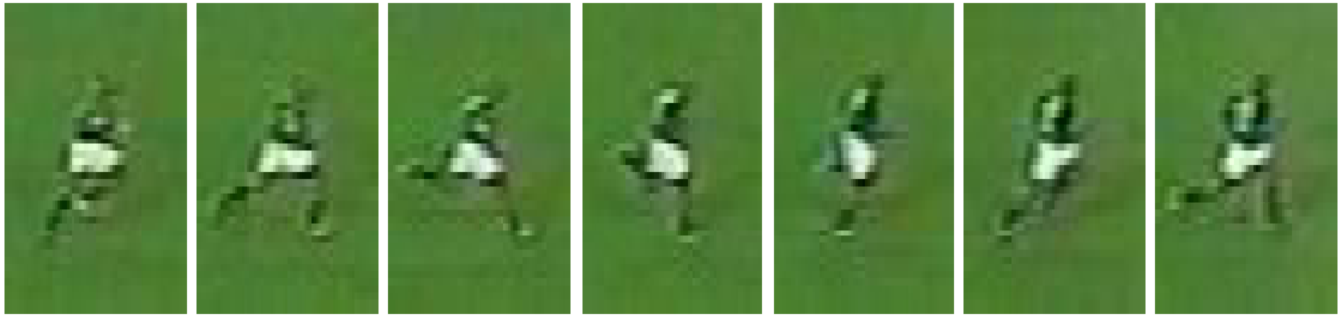
A. A. Efros, A.C. Berg, G. Mori and J. Malik. Recognizing Action at a Distance. In Proc. ICCV 2003

Spatio-Temporal Motion Descriptor

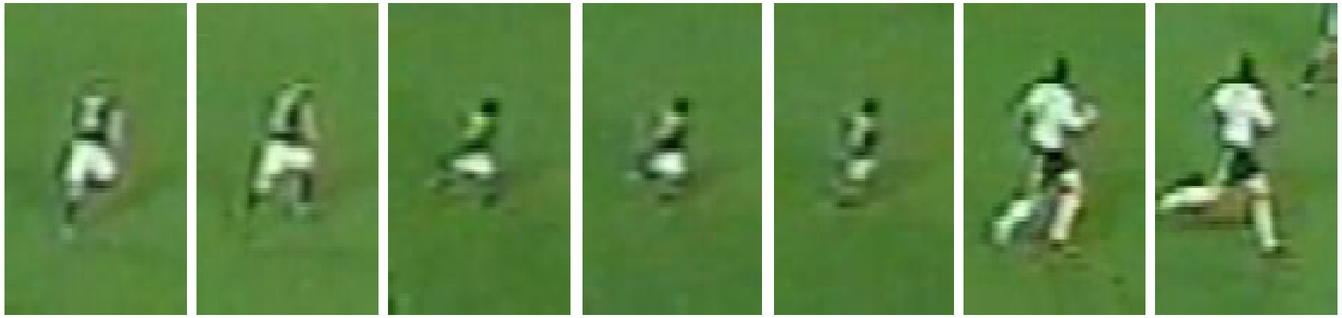


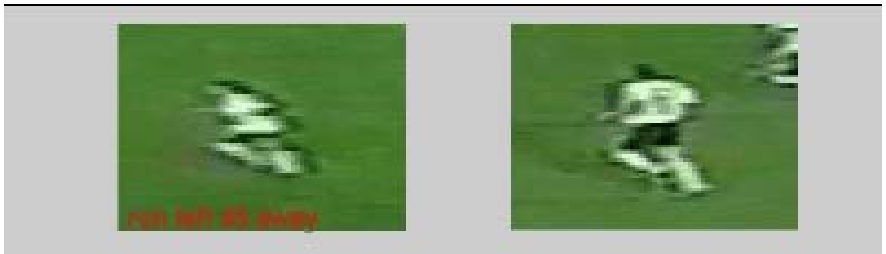
Football Actions: matching

Input Sequence



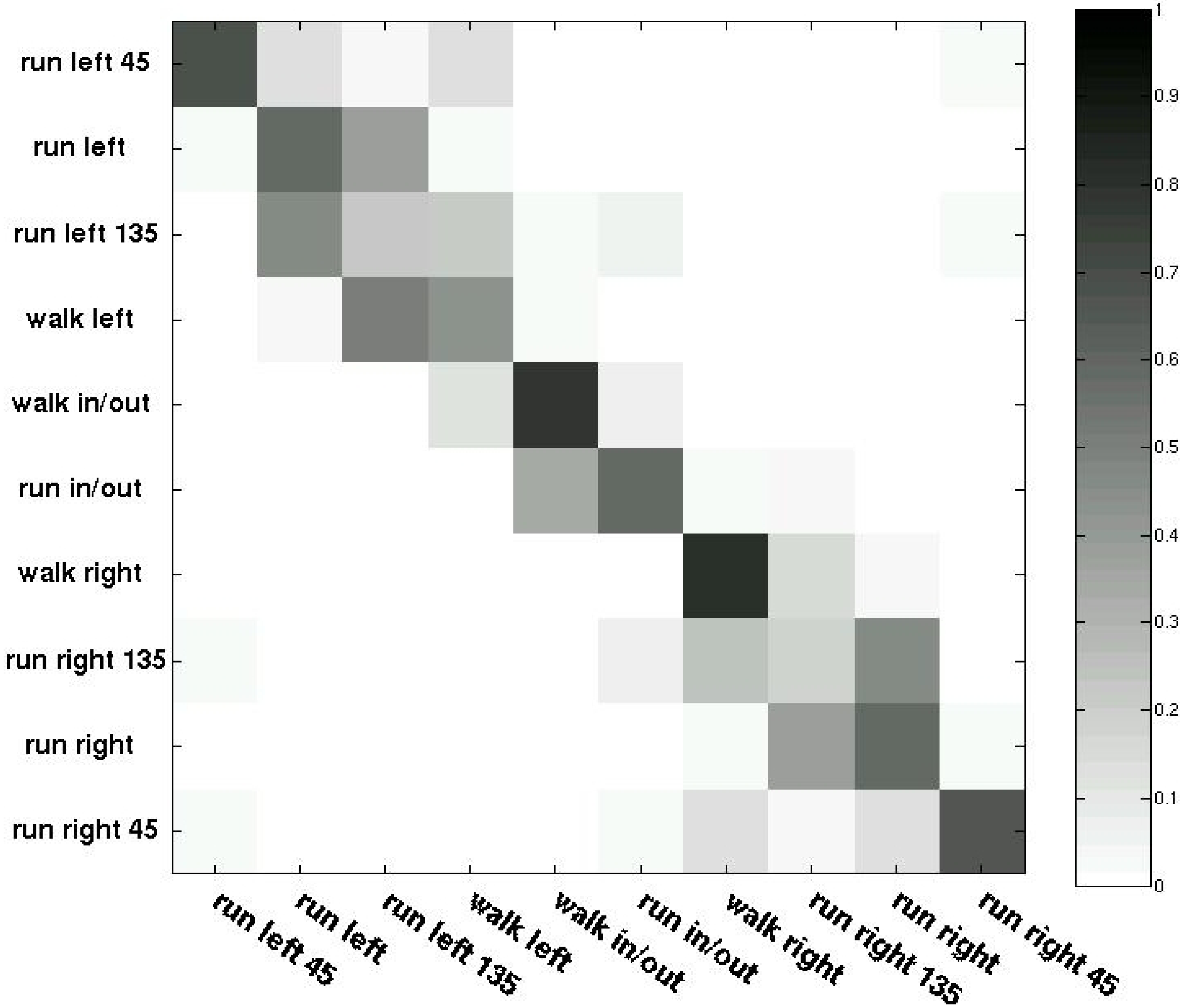
Matched Frames





input matched

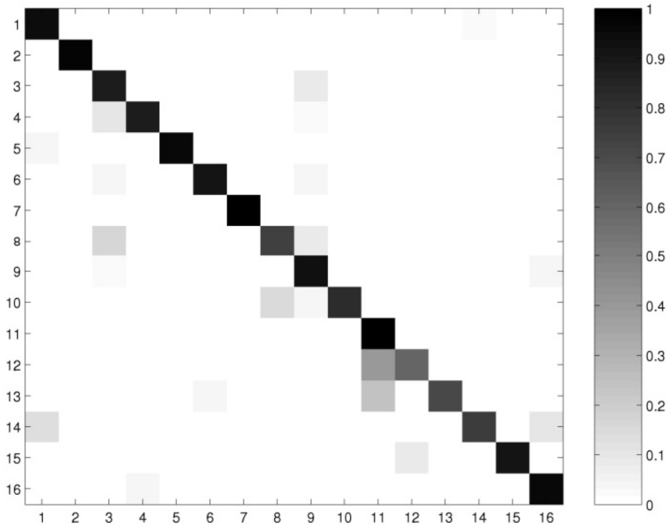
Football Actions: classification



10 actions; 4500 total frames; 13-frame motion descriptor

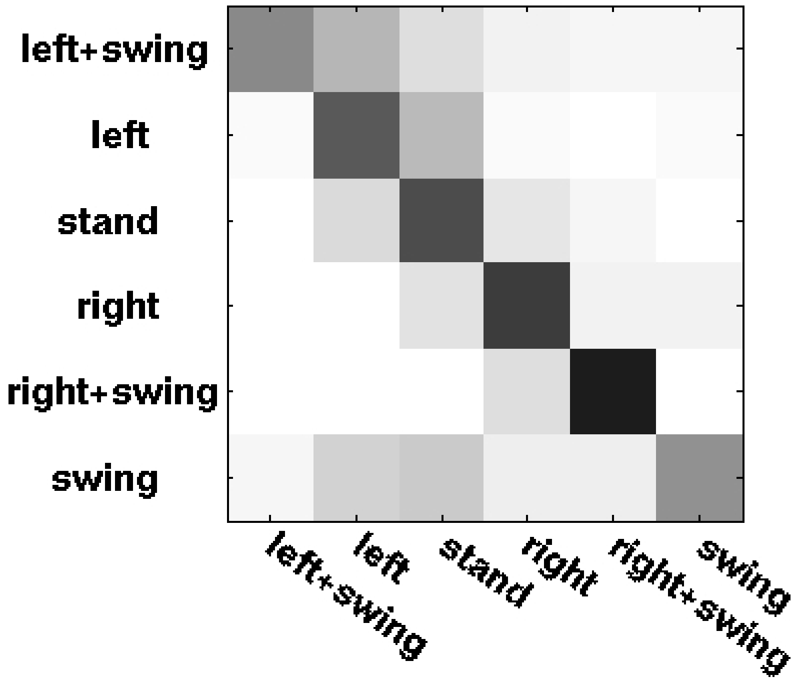
Classifying Ballet Actions

16 Actions; 24800 total frames; 51-frame motion descriptor. Men used to classify women and vice versa.

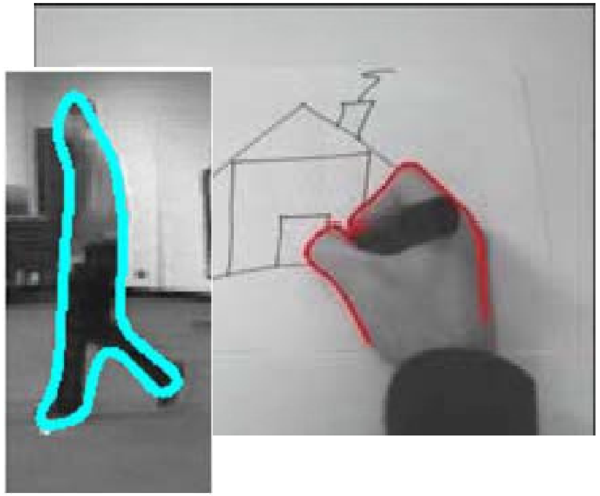


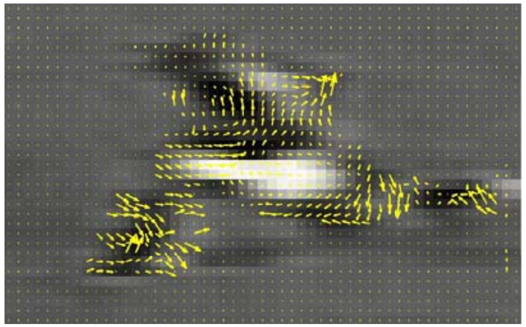
Classifying Tennis Actions

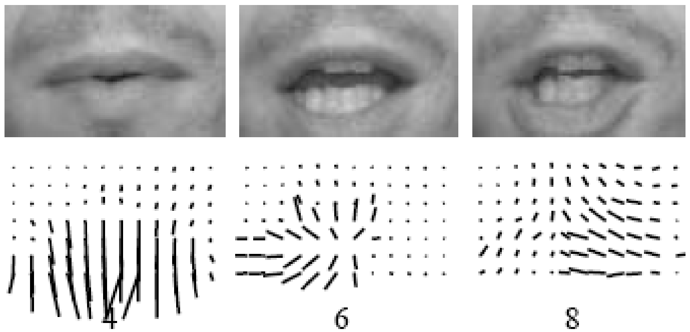
6 actions; 4600 frames; 7-frame motion descriptor
Woman player used as training, man as testing.



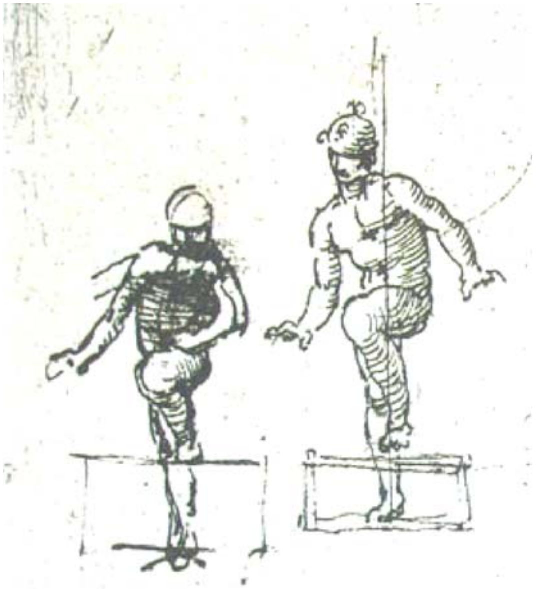
Where are we so far?







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

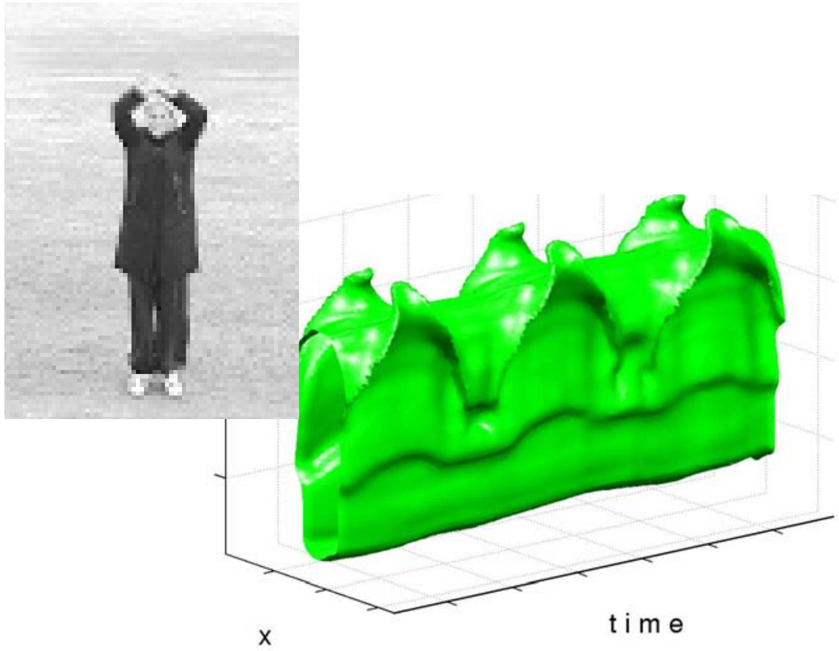
No global assumptions ⇒

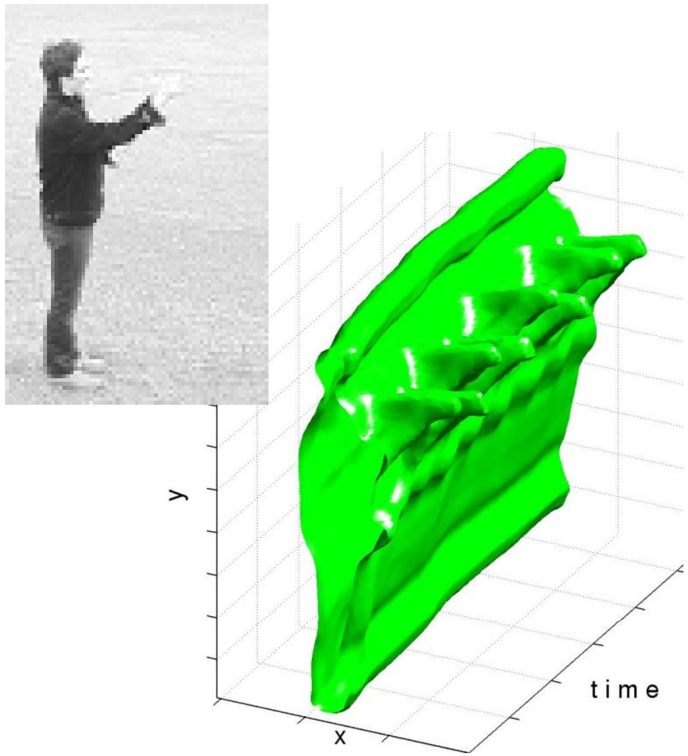
Consider local spatio-temporal neighborhoods

hand waving

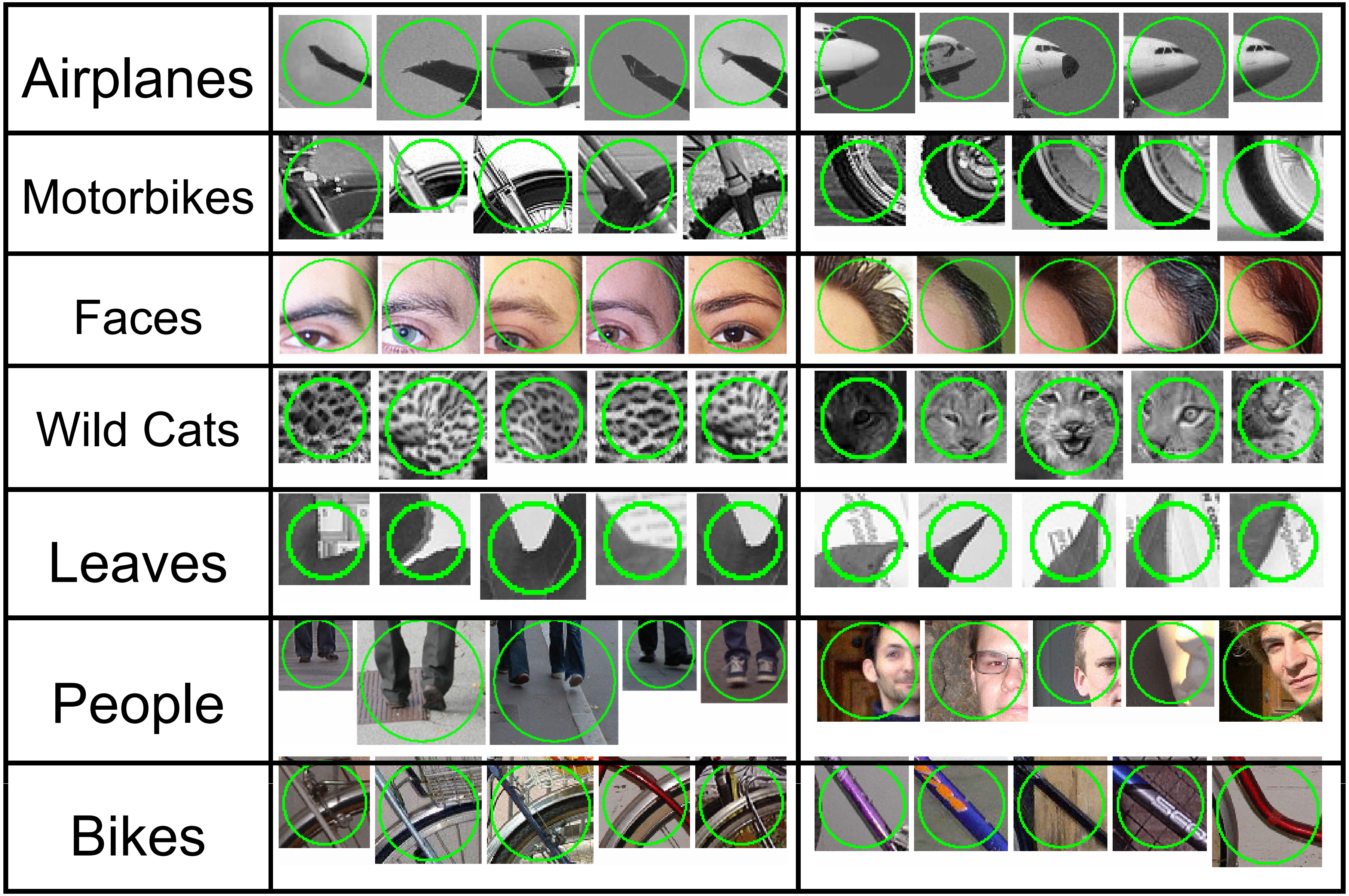
boxing

Actions == Space-time objects?

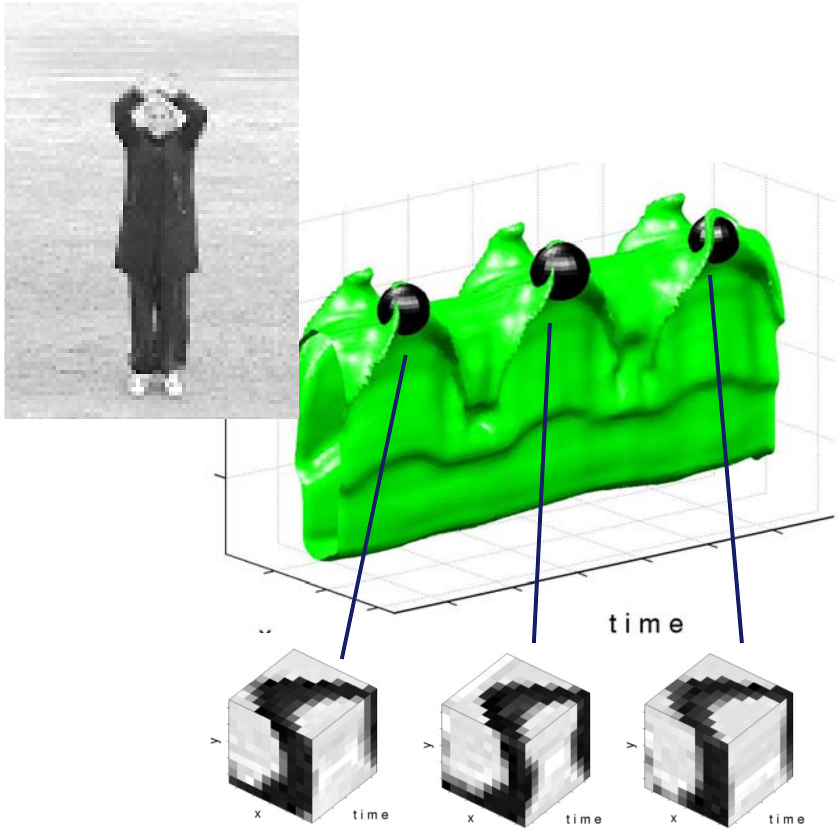


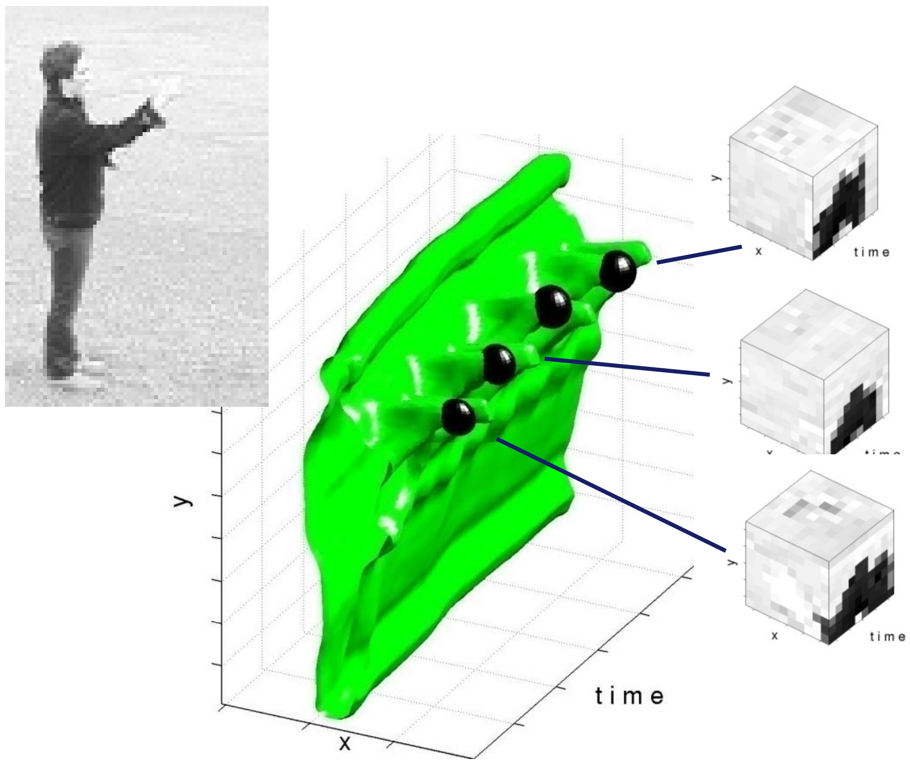


Local approach: Bag of Visual Words



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

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

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

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

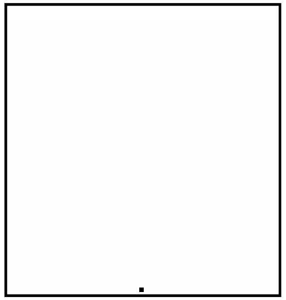
Large eigenvalues of μ can be detected by the local maxima of H over (x,y,t):

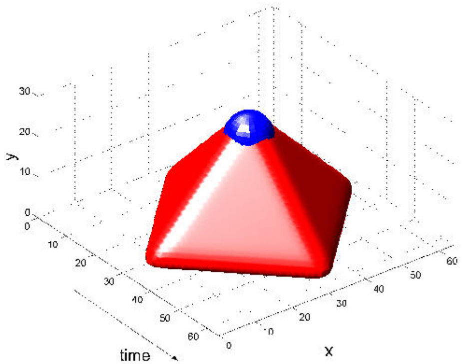
$$
\begin{aligned}
H(p;\,\Sigma) &= \det(\mu(p;\,\Sigma)) + k\,\mathrm{trace}^3(\mu(p;\,\Sigma)) \\
&= \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3
\end{aligned}
$$

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

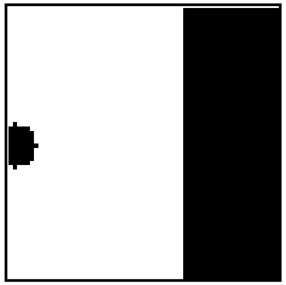
Space-Time interest points

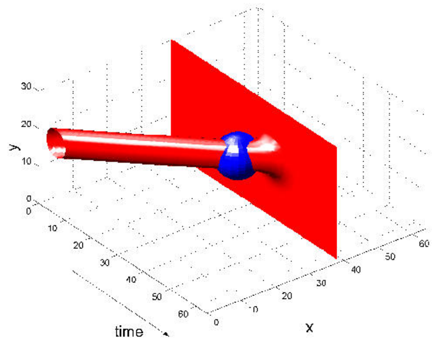
Velocity changes



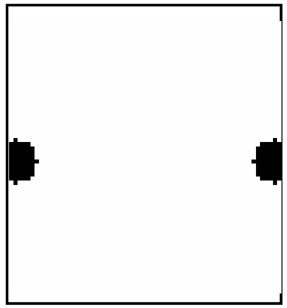


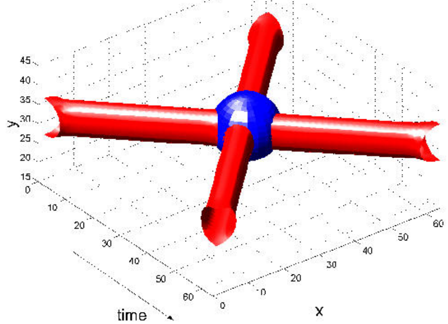
appearance/ disappearance





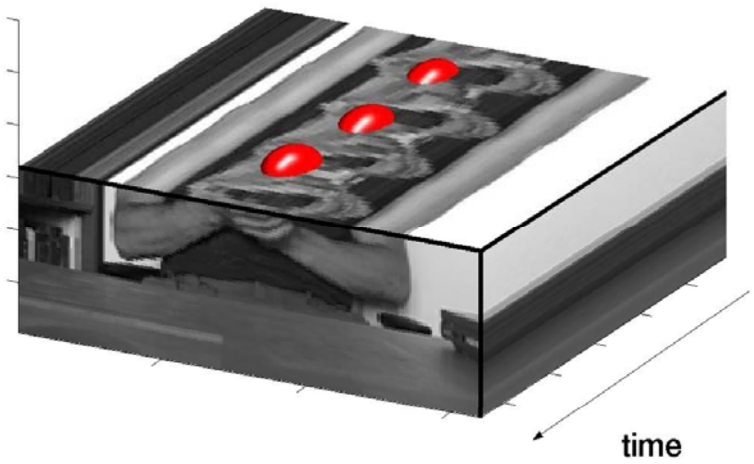
split/merge

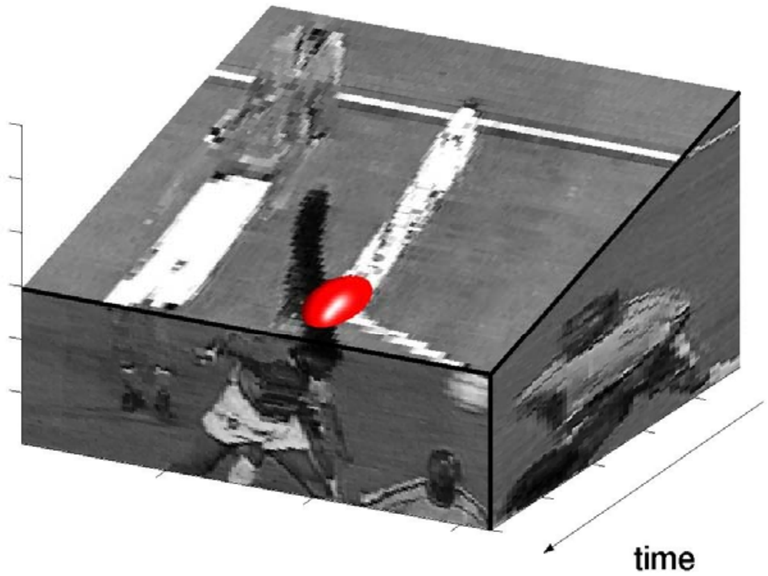




Space-Time Interest Points: Examples

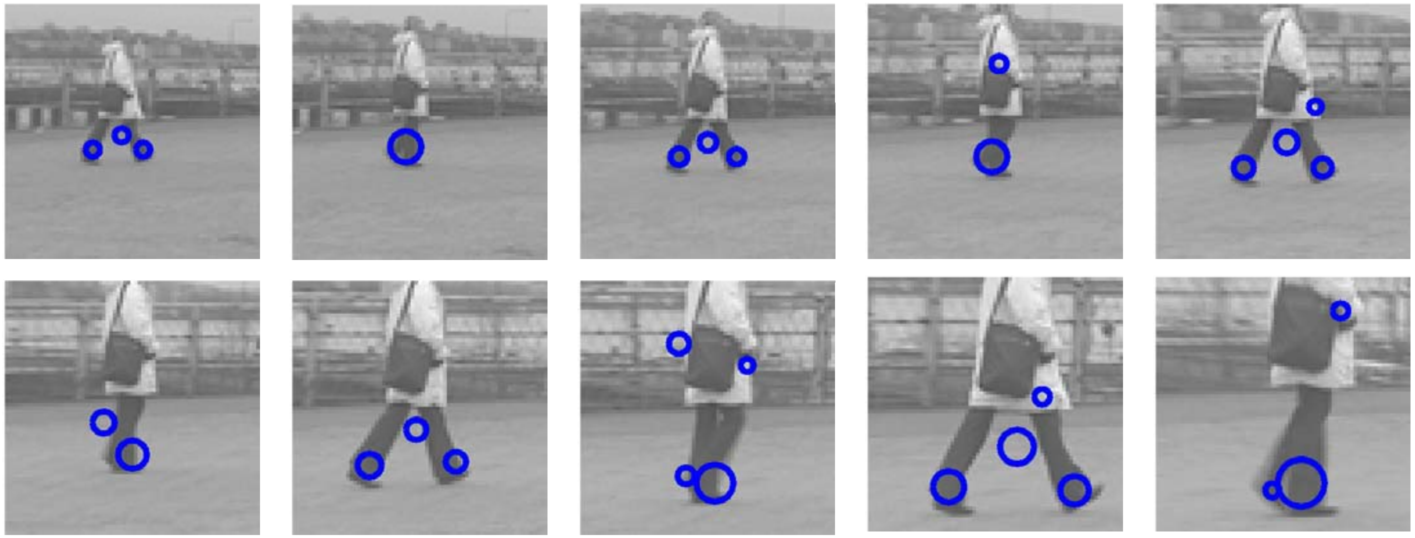
Motion event detection



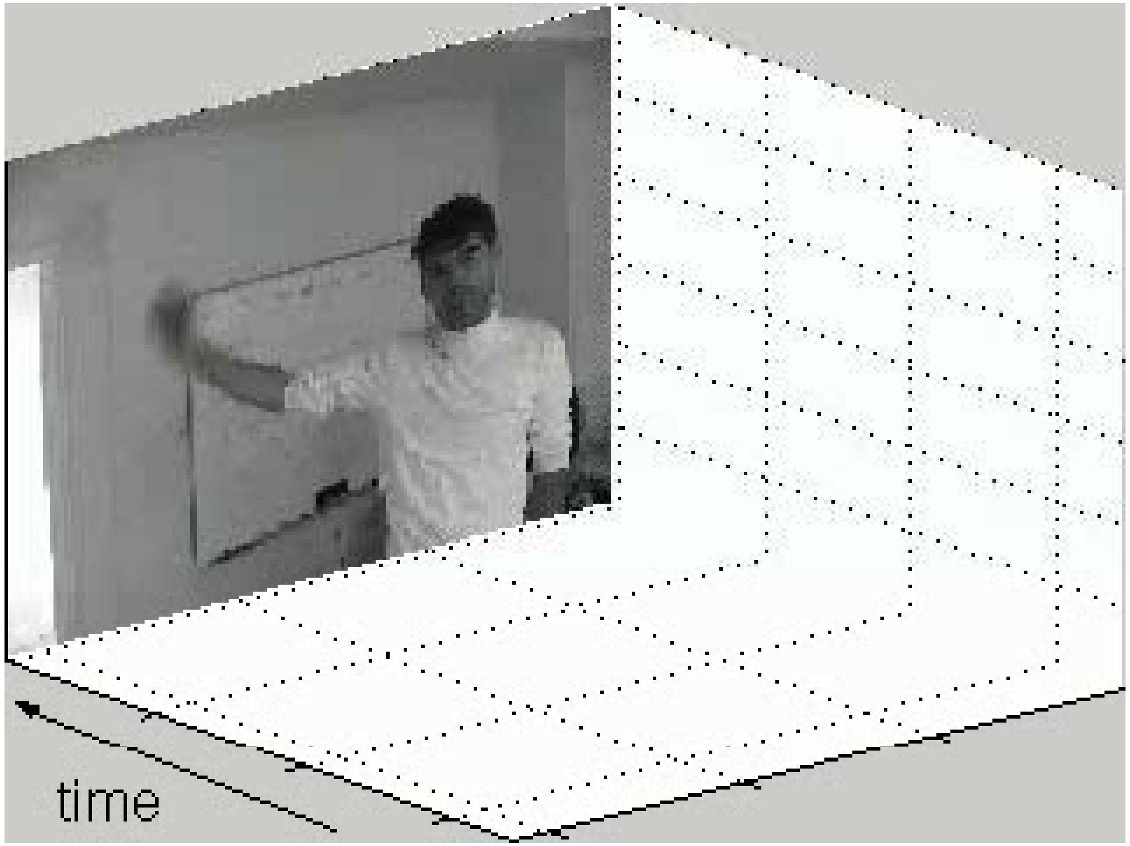


Spatio-temporal scale selection

Stability to size changes,
e.g. camera zoom

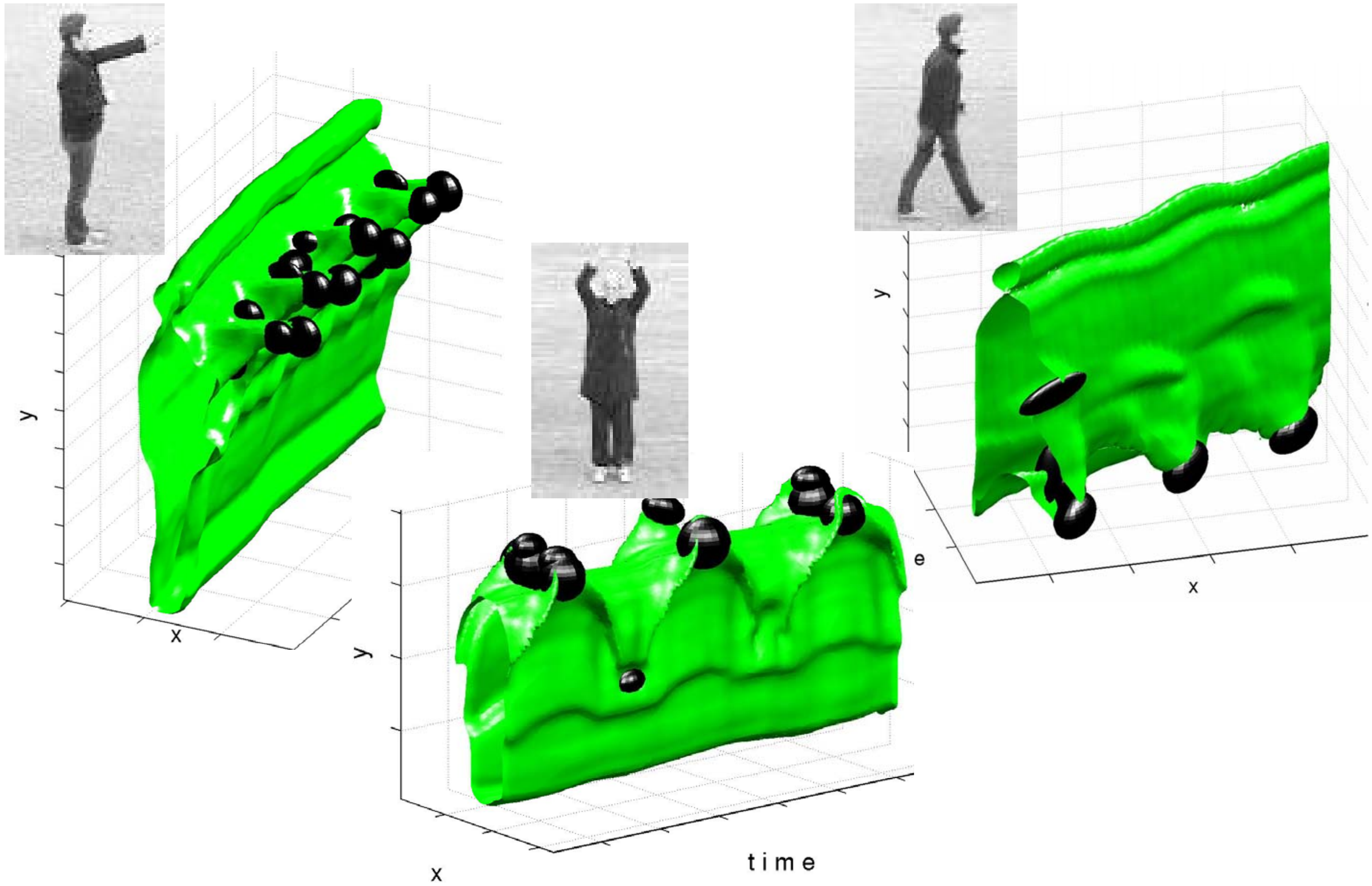


Spatio-temporal scale selection

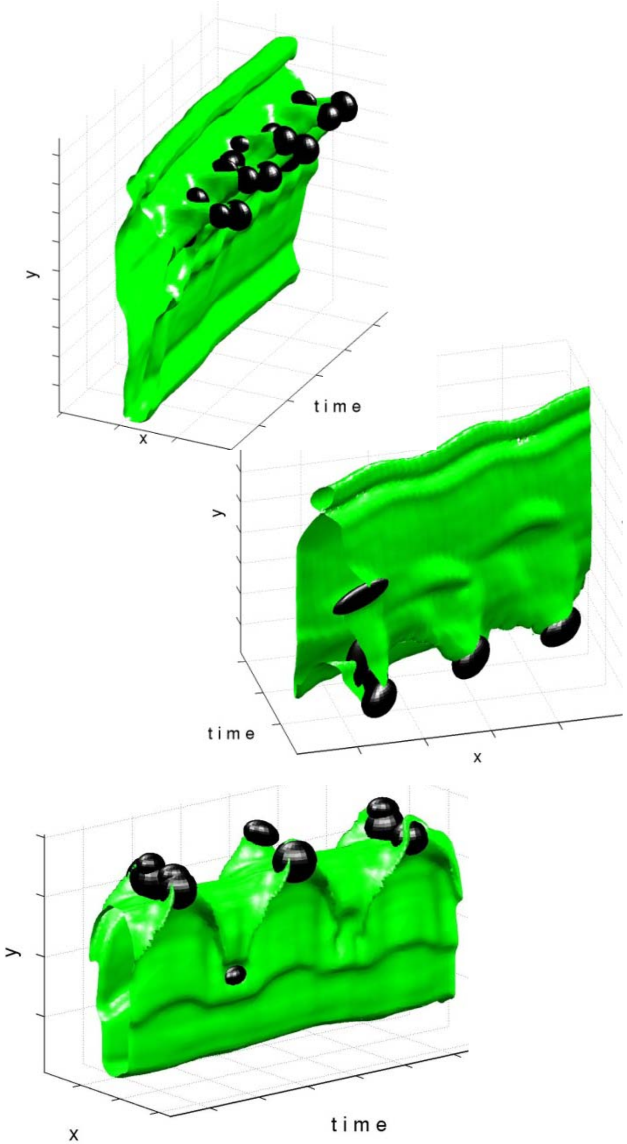


Selection of temporal scales captures the frequency of events

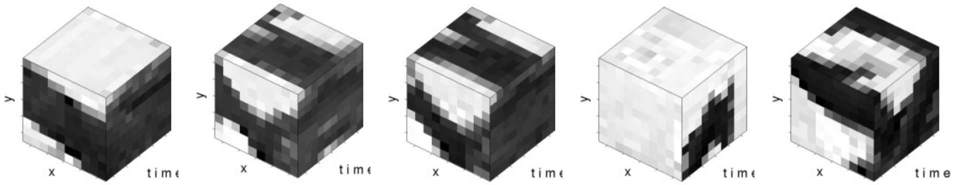
Local features for human actions



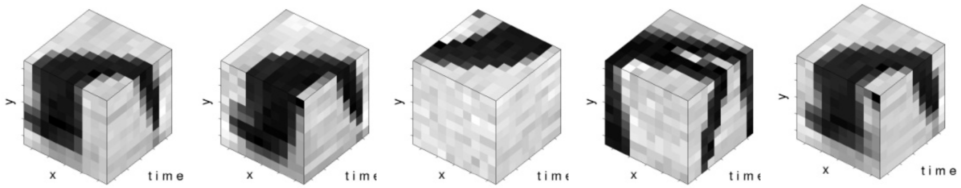
Local features for human actions



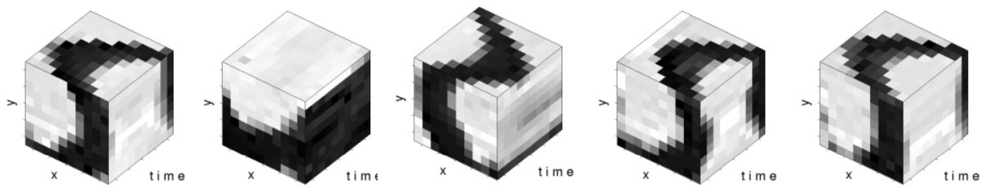
boxing



walking

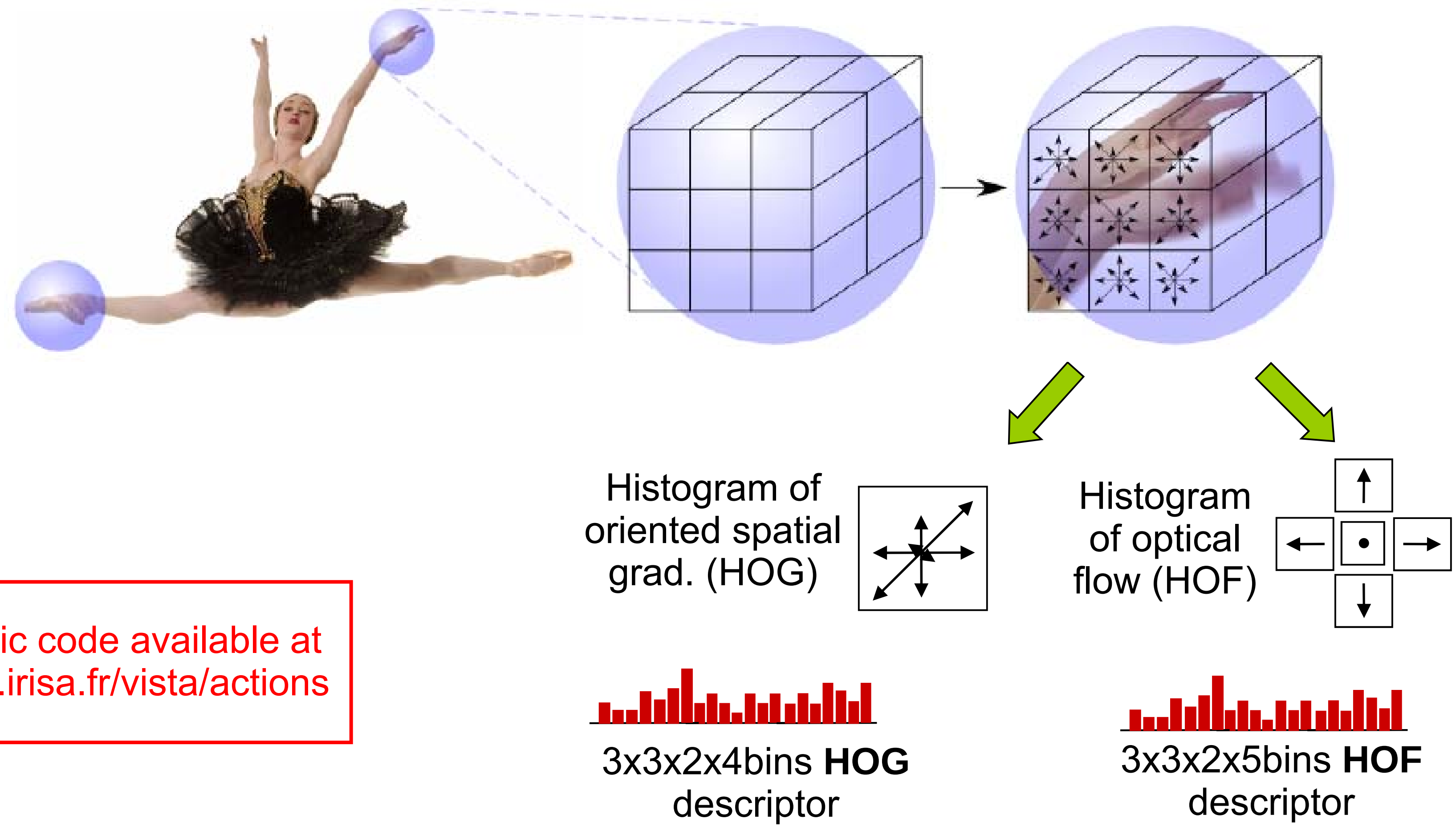


hand waving



Local space-time descriptor: HOG/HOF

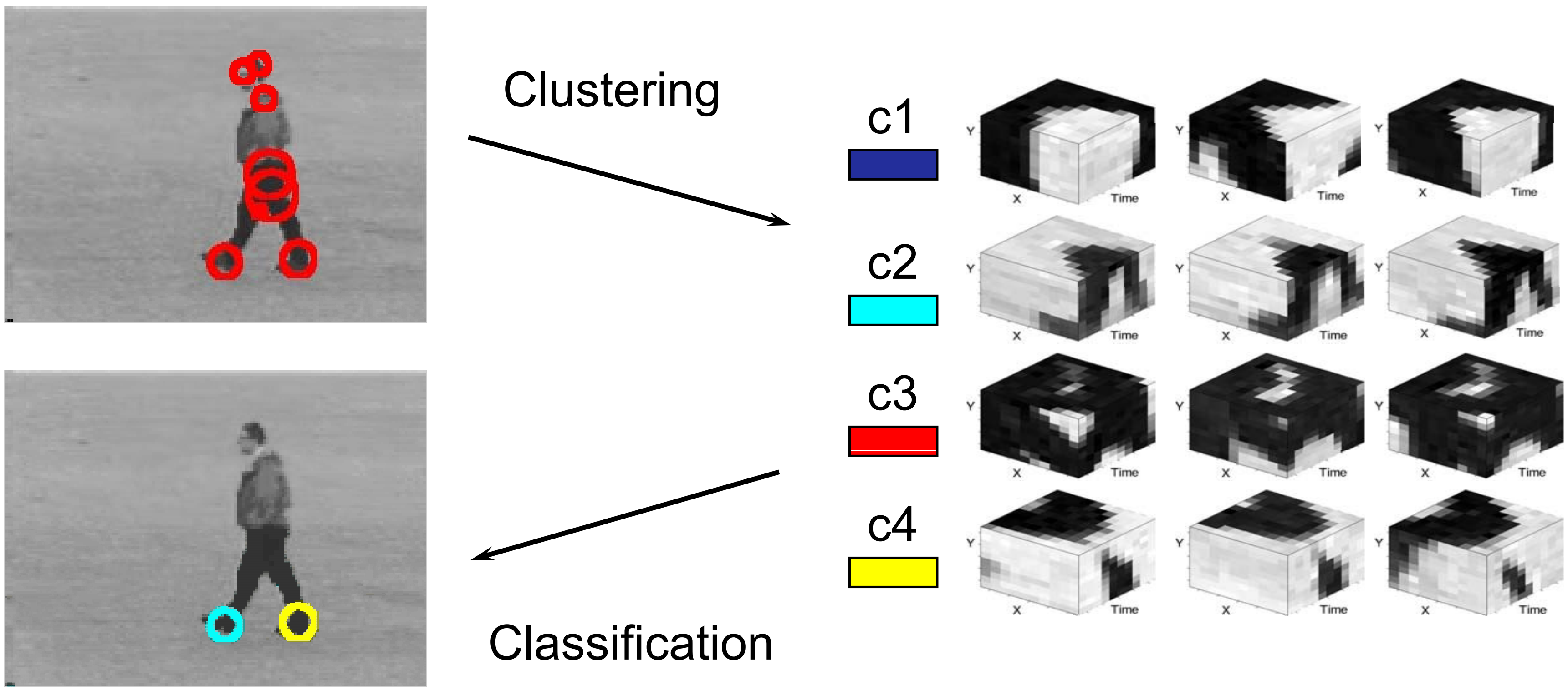
Multi-scale space-time patches



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

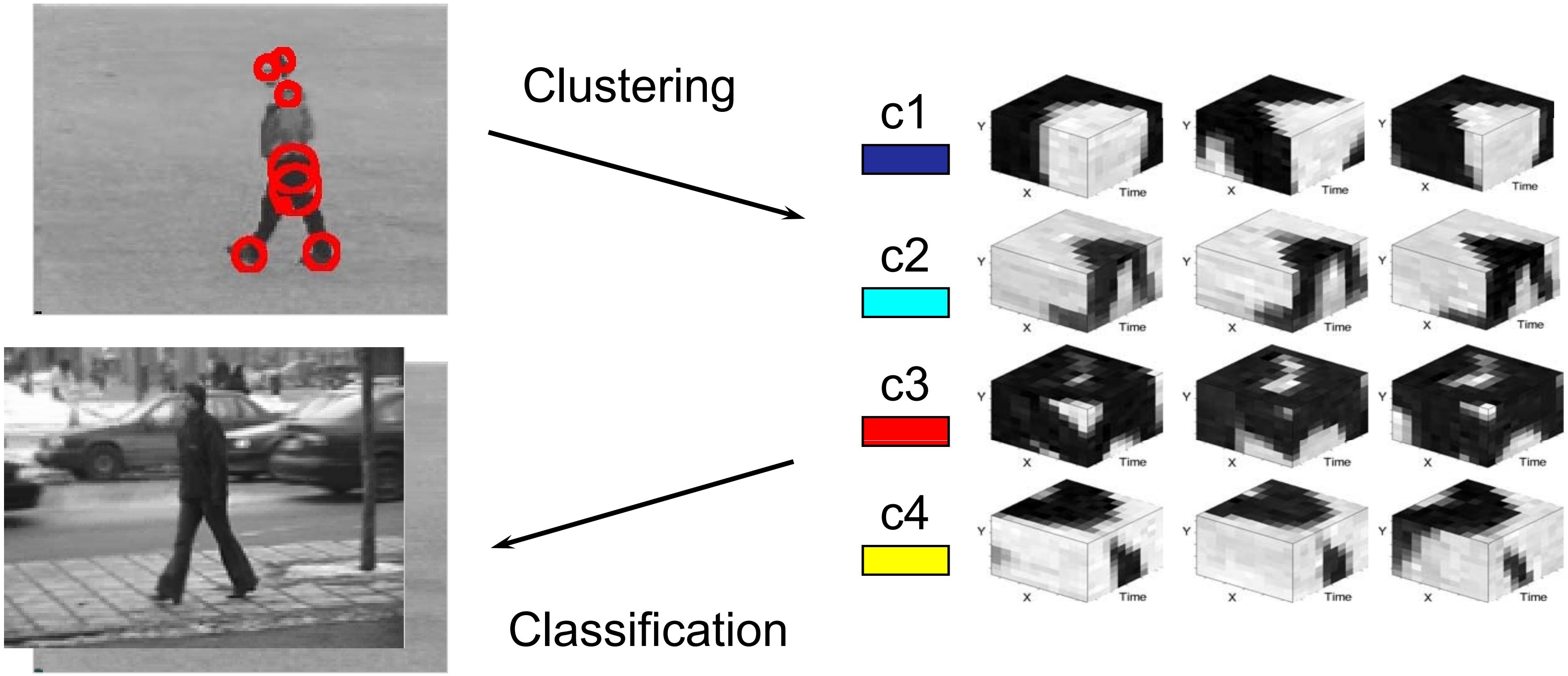
Visual Vocabulary: K-means clustering

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



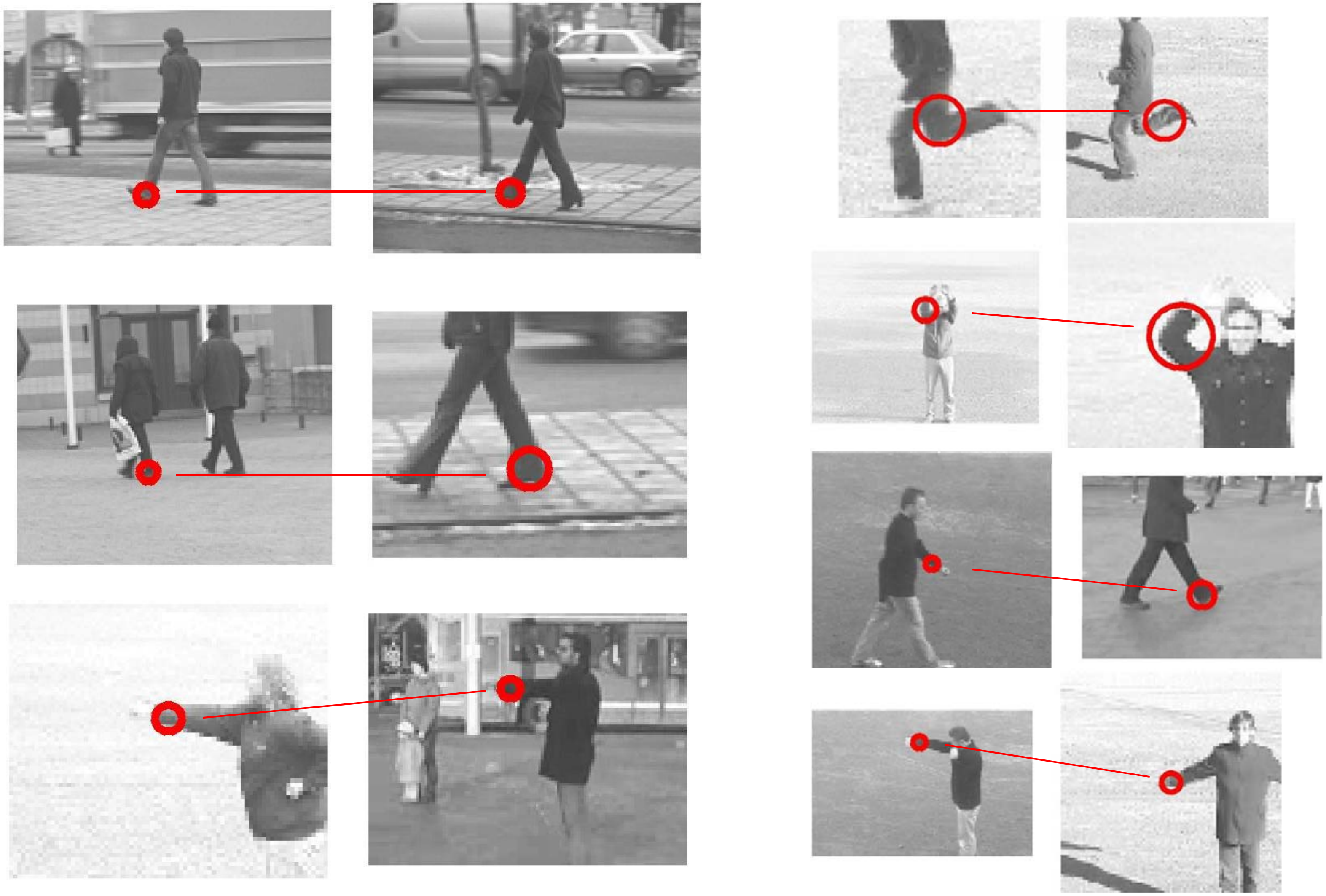
Visual Vocabulary: K-means clustering

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



Local Space-time features: Matching

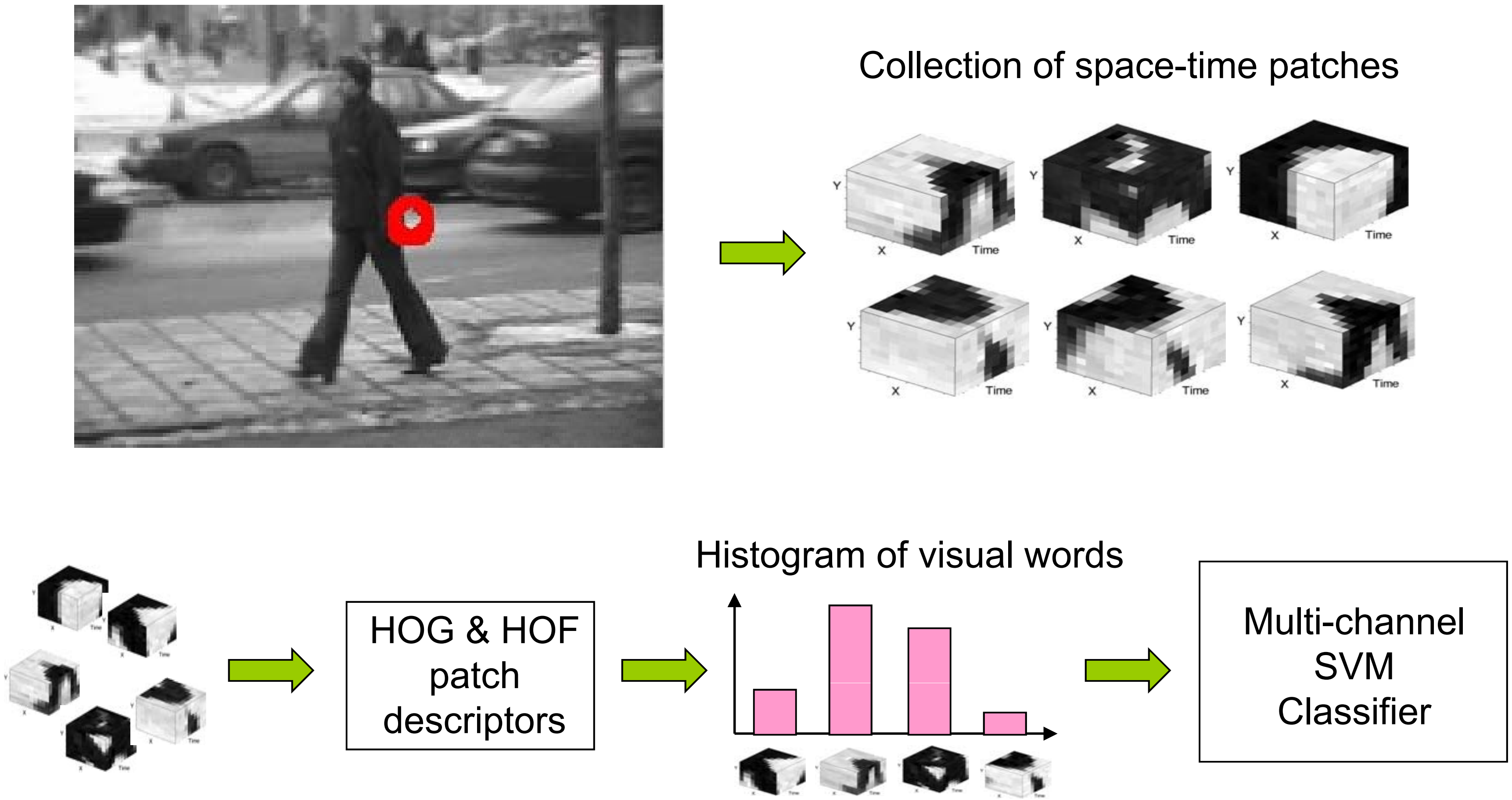
- Find similar events in pairs of video sequences



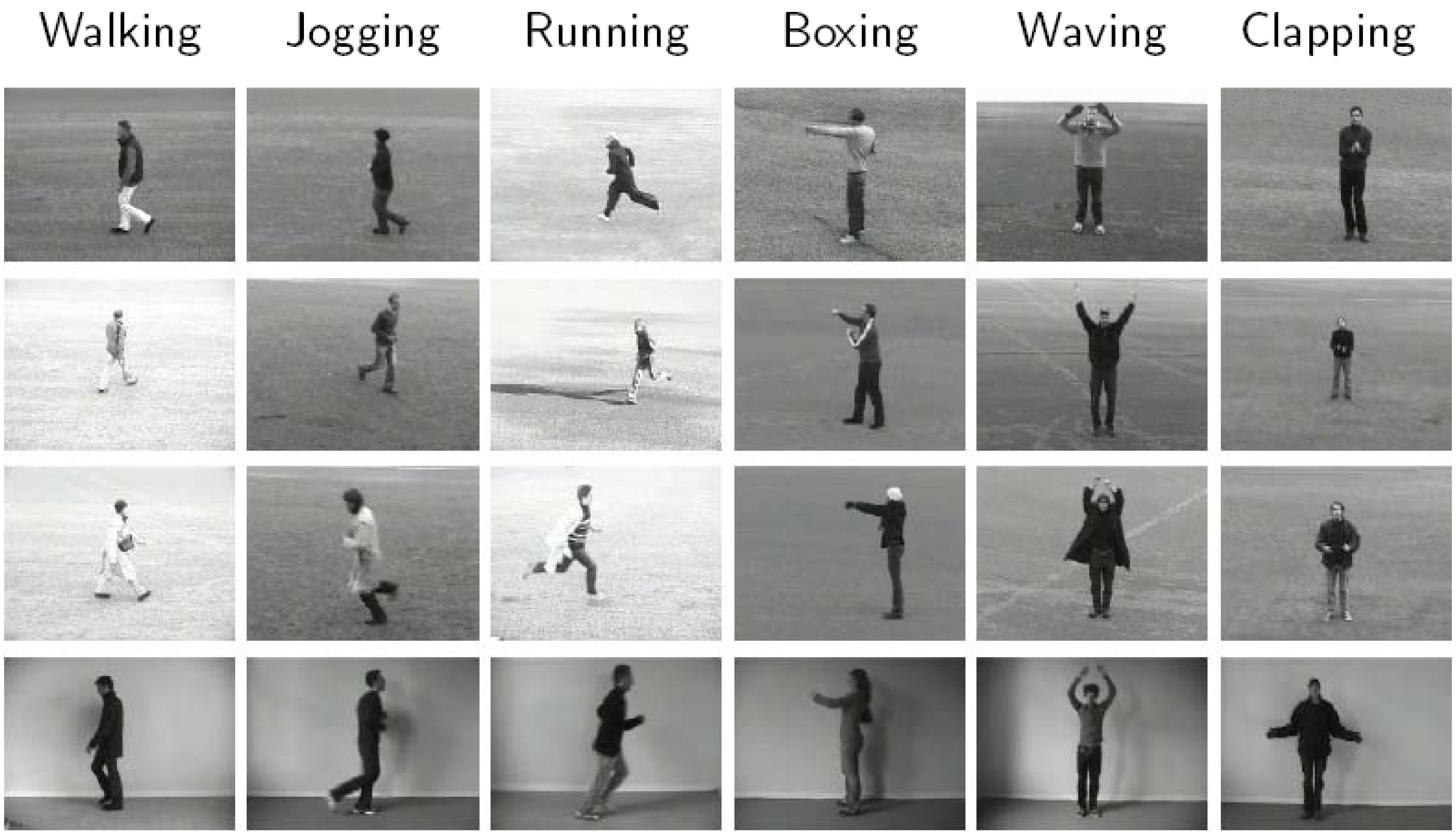
Action Classification: Overview

Bag of space-time features + multi-channel SVM

[Laptev'03, Schuldt'04, Niebles'06, Zhang'07]



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

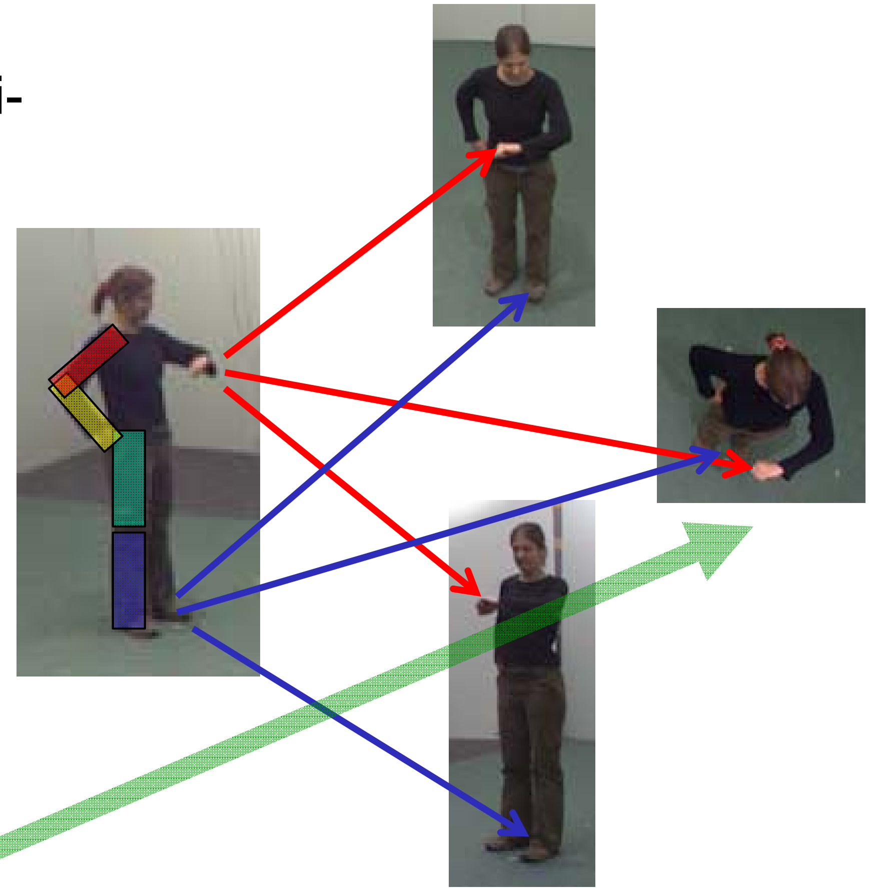
What about 3D?

Local motion and appearance features are not invariant to view changes

Multi-view action recognition

Difficult to apply standard multi-view methods:

- Do not want to search for multi-view point correspondence --- Non-rigid motion, clothing changes, … --> It's Hard!
- Do not want to identify body parts. Current methods are not reliable enough.
- Yet, want to learn actions from one view and recognize actions in very different views

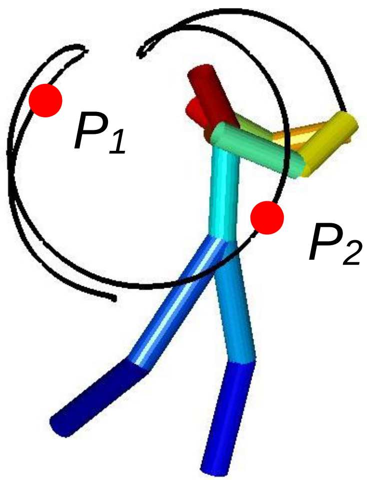


Temporal self-similarities

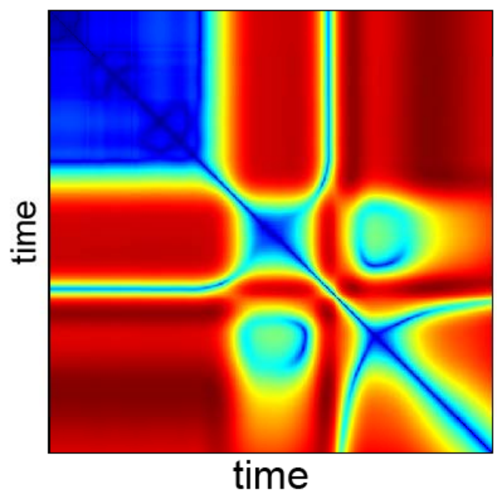
Idea:

- *Cross-view* matching is hard but *cross-time* matching (tracking) is relatively easy.
- Measure self-(dis)similarities across time: $\mathcal{D}(t_1, t_2),\ t_1, t_2 \in (1, ..., T)$

Example: $\mathcal{D}(t_1, t_2) = ||P_1 - P_2||_2$

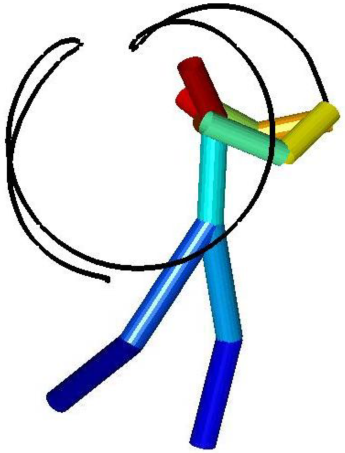


Distance matrix / self-similarity matrix (SSM):

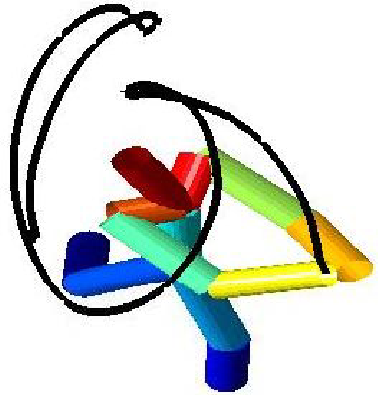


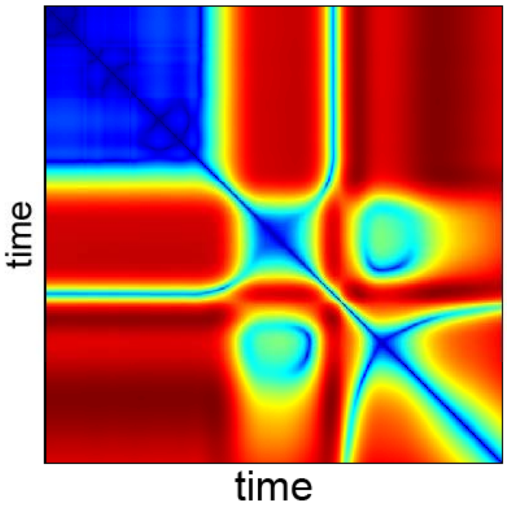
Temporal self-similarities: Multi-views

Side view

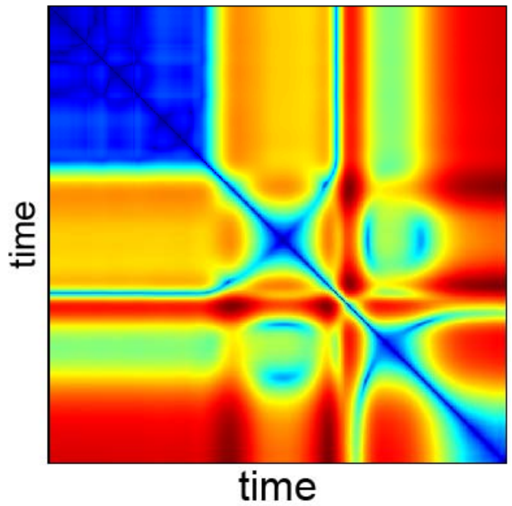


Top view





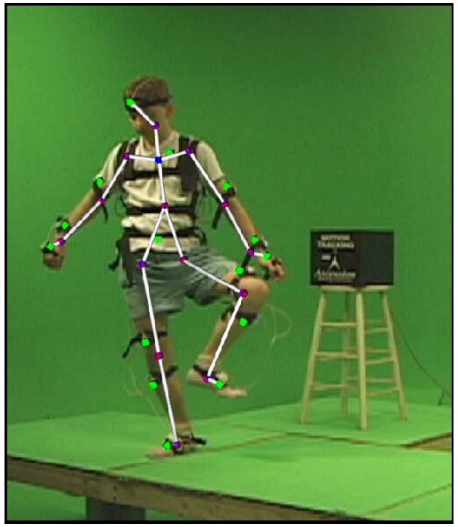
Appear very similar despite the view change!



Intuition: 1. Distance between similar poses is low in any view
2. Distance among different poses is likely to be large in most views

Temporal self-similarities: MoCap

Self-similarities can be measured from Motion Capture (MoCap) data

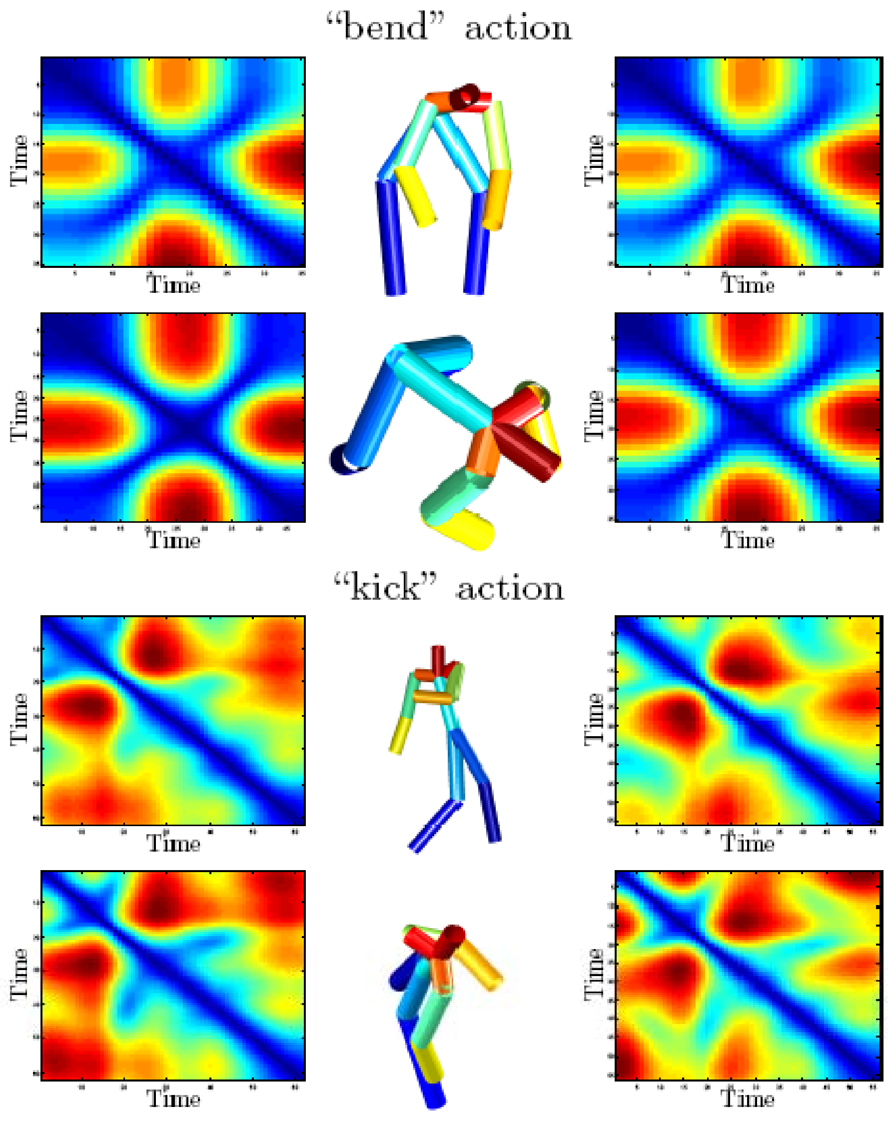


person 1

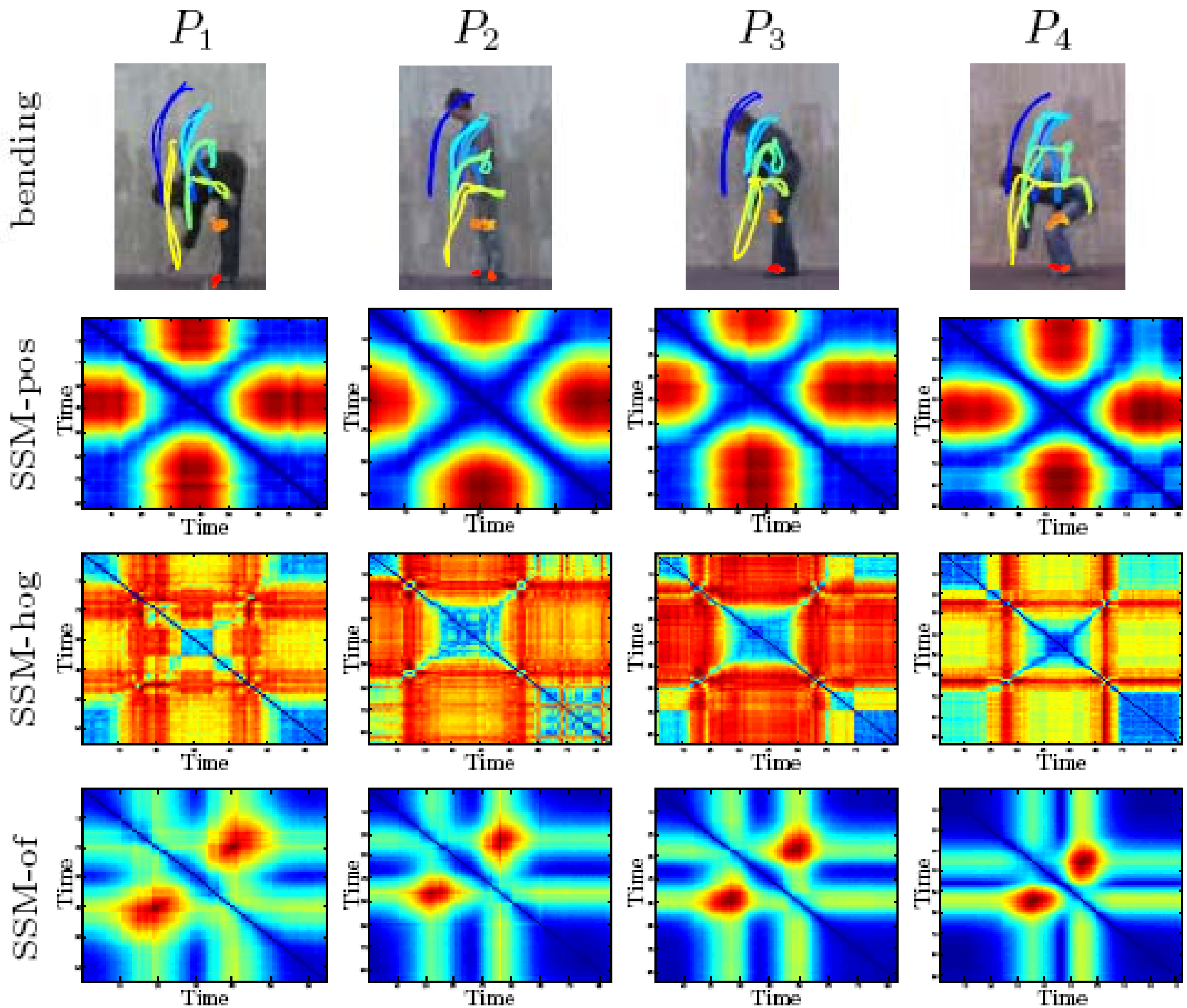
person 2

person 1

person 2



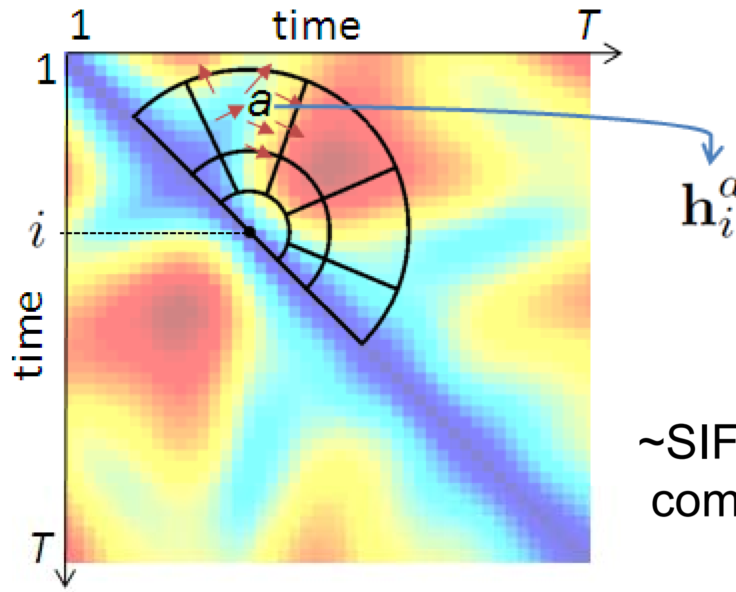
Temporal self-similarities: Video



Self-similarities can be measured directly from video:
HOG or Optical Flow descriptors in image frames

Self-similarity descriptor

Goal:
define a quantitative measure to compare self-similarity matrices



$$\mathbf{h}_i^a = \begin{bmatrix} h_{i,1}^a \\ \vdots \\ h_{i,8}^a \end{bmatrix}$$

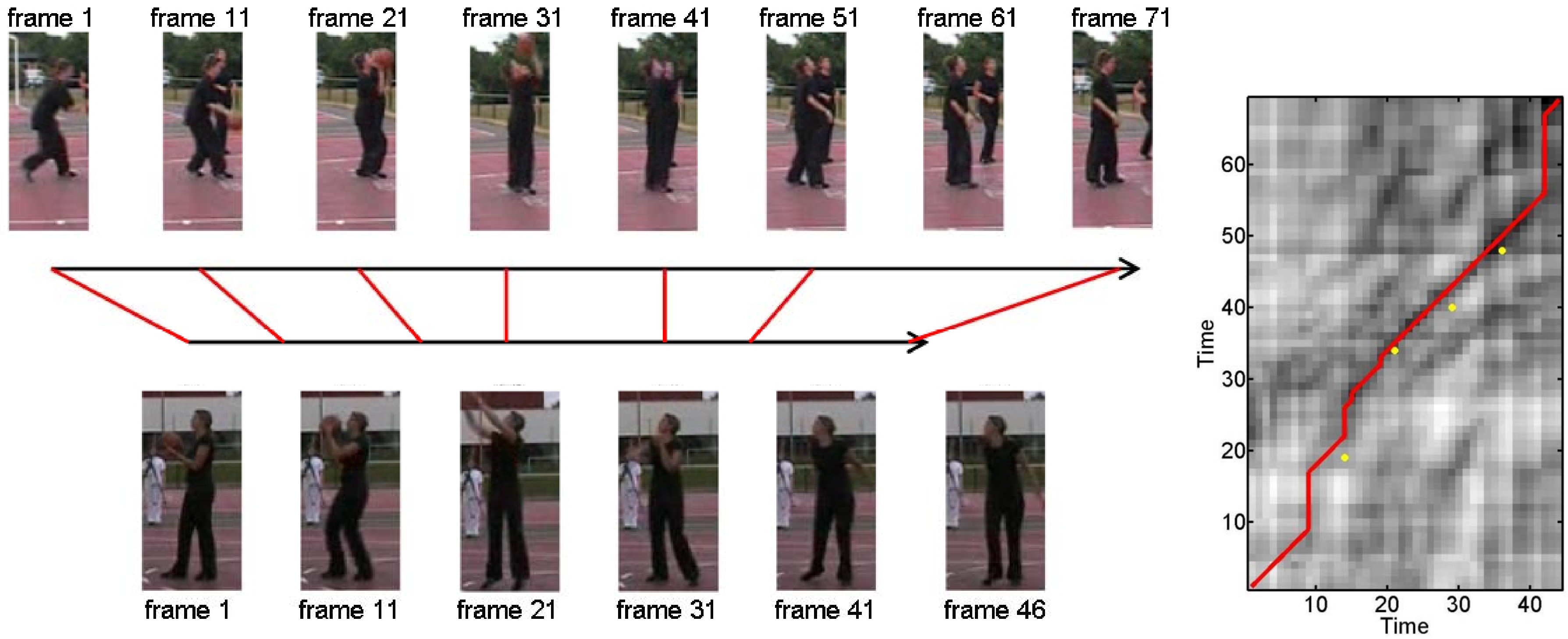
~SIFT descriptor computed on SSM

- Define a local histogram descriptor h_i for each point i on the diagonal.

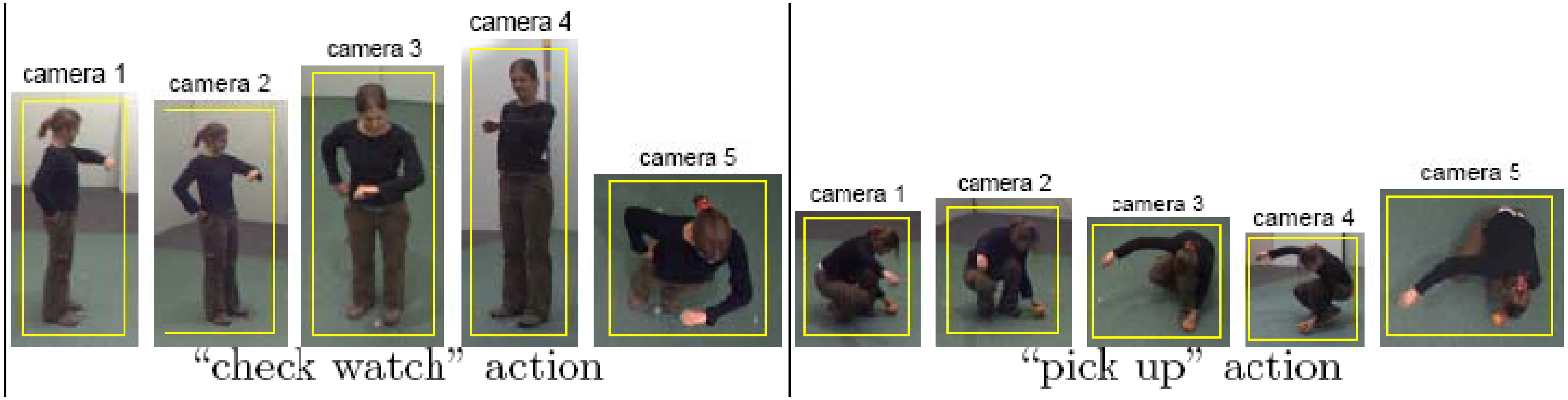
- **Sequence alignment:** Dynamic Programming for two sequences of descriptors $\{h_i\}$, $\{h_j\}$

- **Action recognition:**
  - Visual vocabulary for h
  - BoF representation of $\{h_i\}$
  - SVM

Multi-view alignment



Multi-view action recognition: Video



	Test Cam0	Test Cam1	Test Cam2	Test Cam3	Test Cam4	Test All
Train Cam0	77.0	75.2	69.7	71.8	49.4	68.6
Train Cam1	78.5	77.3	67.9	71.5	48.0	68.6
Train Cam2	70.0	73.0	75.8	68.5	55.2	68.5
Train Cam3	73.6	72.4	67.3	71.2	45.9	66.1
Train Cam4	44.5	41.5	55.2	37.9	68.8	49.6
Train All	77.0	78.8	80.0	73.9	63.3	74.6

cross−camera training/testing   same camera training/testing

SSM-based recognition

	Test Cam0	Test Cam1	Test Cam2	Test Cam3	Test Cam4	Test All
Train Cam0	80.0	75.9	42.3	55.6	21.8	55.6
Train Cam1	74.8	83.9	36.5	58.3	23.6	56.0
Train Cam2	43.6	46.1	80.5	64.7	34.2	53.7
Train Cam3	47.0	50.0	45.8	85.5	18.8	49.5
Train Cam4	19.7	19.4	43.5	26.1	73.3	36.0
Train All	80.3	84.5	79.4	84.8	68.5	79.6

cross−camera training/testing   same camera training/testing

Alternative view-dependent method (STIP)

What are Human Actions?

Actions in recent datasets:

Is it just about kinematics?

Should actions be defined by the *purpose?*

Kinematics + Objects

What are Human Actions?

Actions in recent datasets:

Is it just about kinematics?

Should actions be defined by the *purpose?*

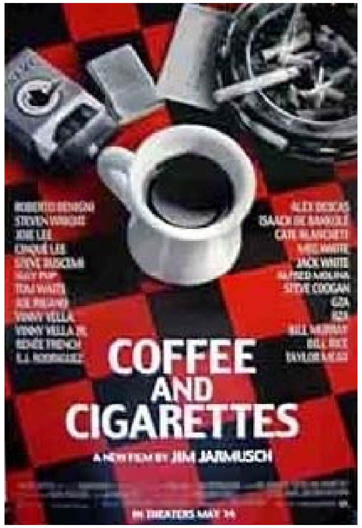
Kinematics + Objects + Scenes

Action recognition in realistic settings

Standard action datasets

Actions “In the Wild”:

Action Dataset and Annotation



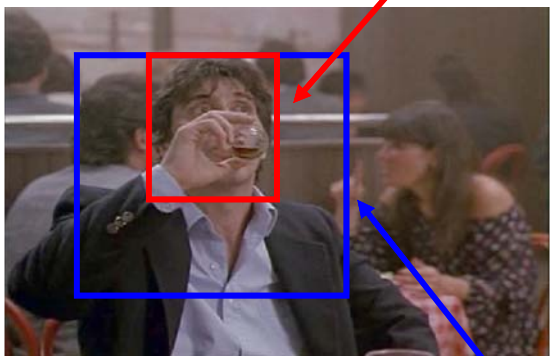
Manual annotation of drinking actions in movies:
“Coffee and Cigarettes”; “Sea of Love”

“*Drinking*”: 159 annotated samples

“*Smoking*”: 149 annotated samples

Spatial annotation

head rectangle



torso rectangle

Temporal annotation

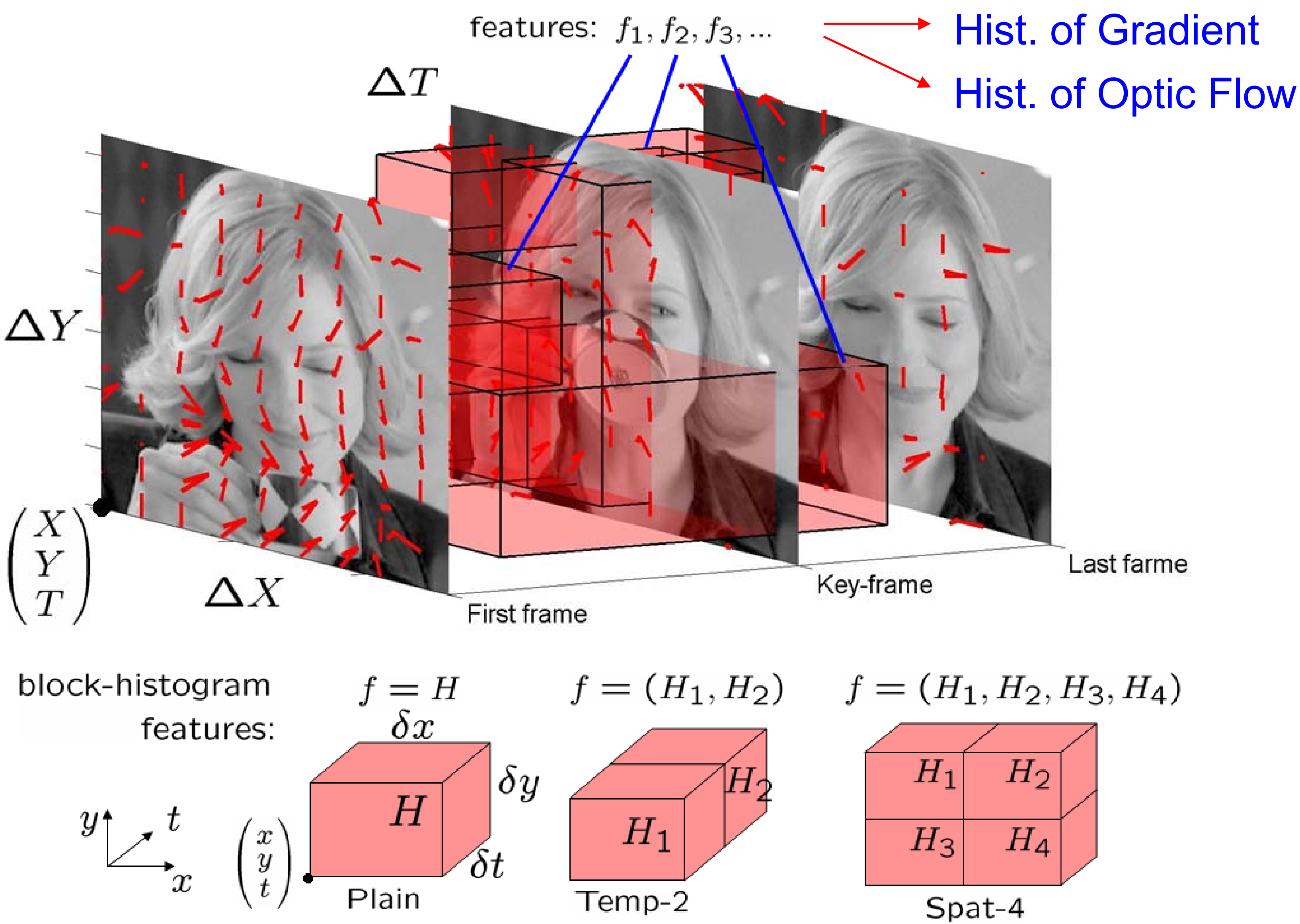
First frame **Keyframe** **Last frame**

“Drinking” action samples

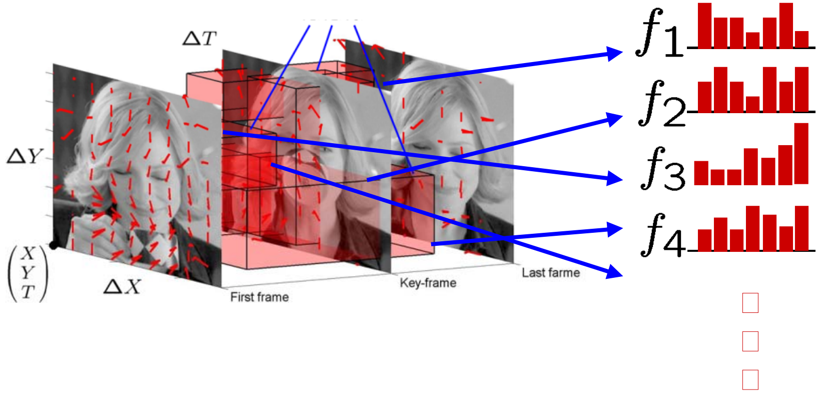
training samples

test samples

Action representation



Action learning



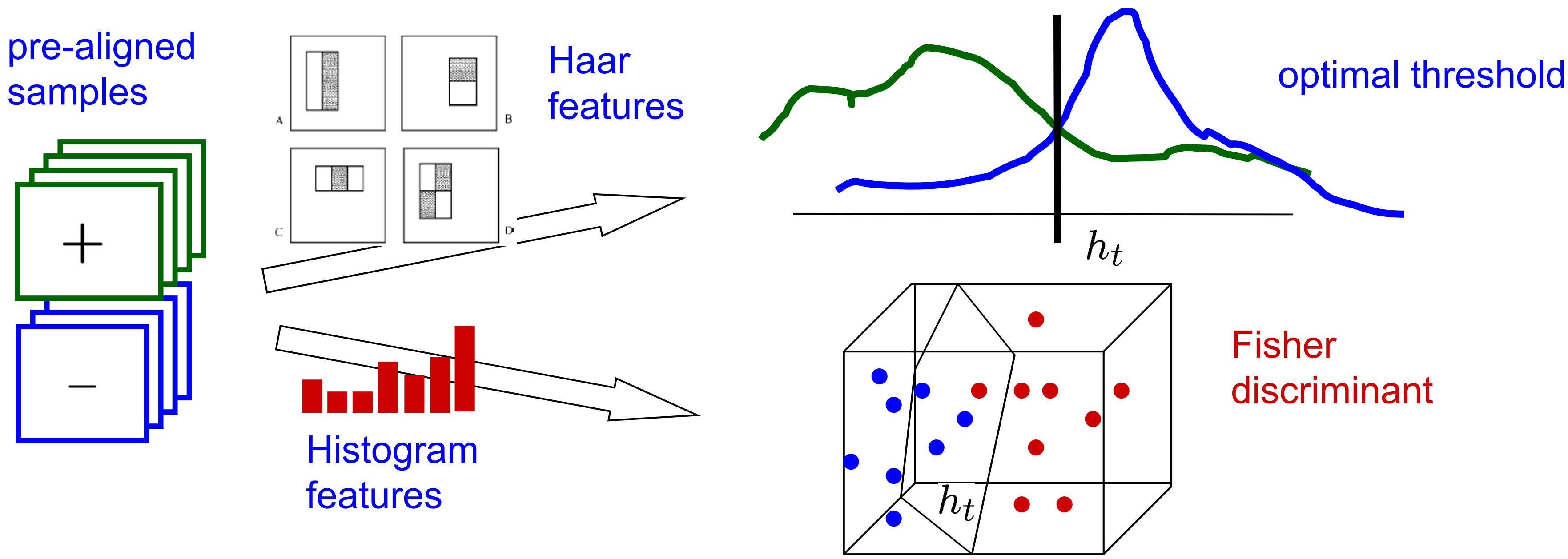
boosting

selected features

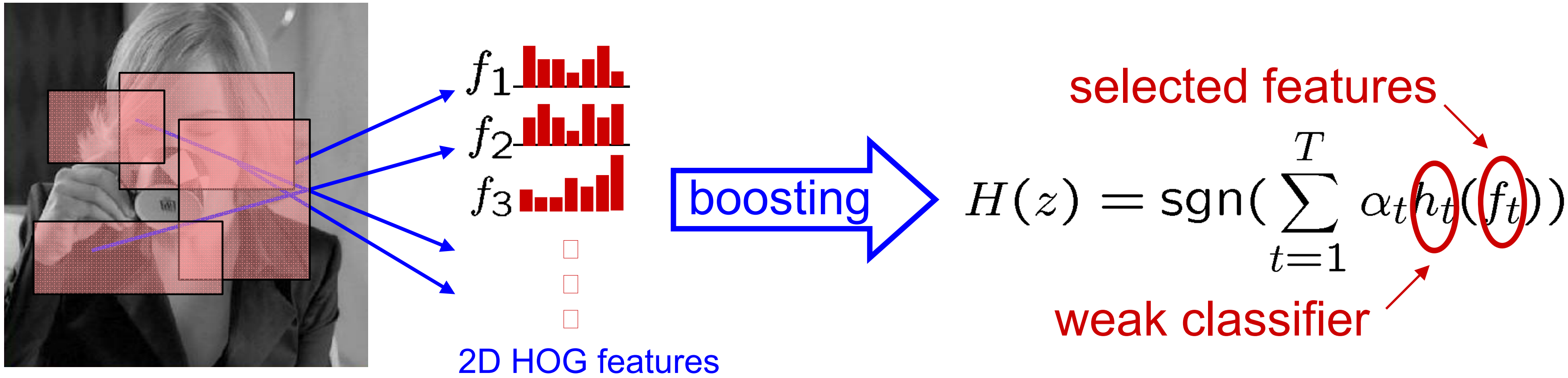
$$H(z) = \mathrm{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(f_t)\right)$$

weak classifier

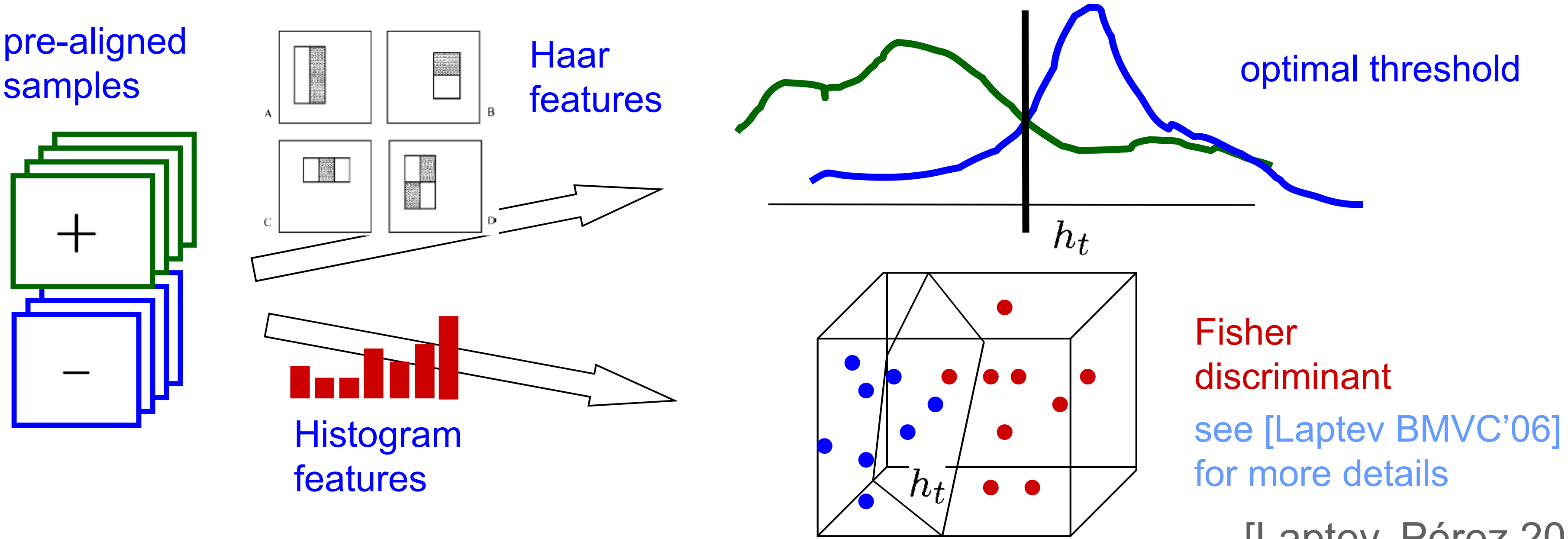
AdaBoost:
- Efficient discriminative classifier [Freund&Schapire'97]
- Good performance for face detection [Viola&Jones'01]



Key-frame action classifier

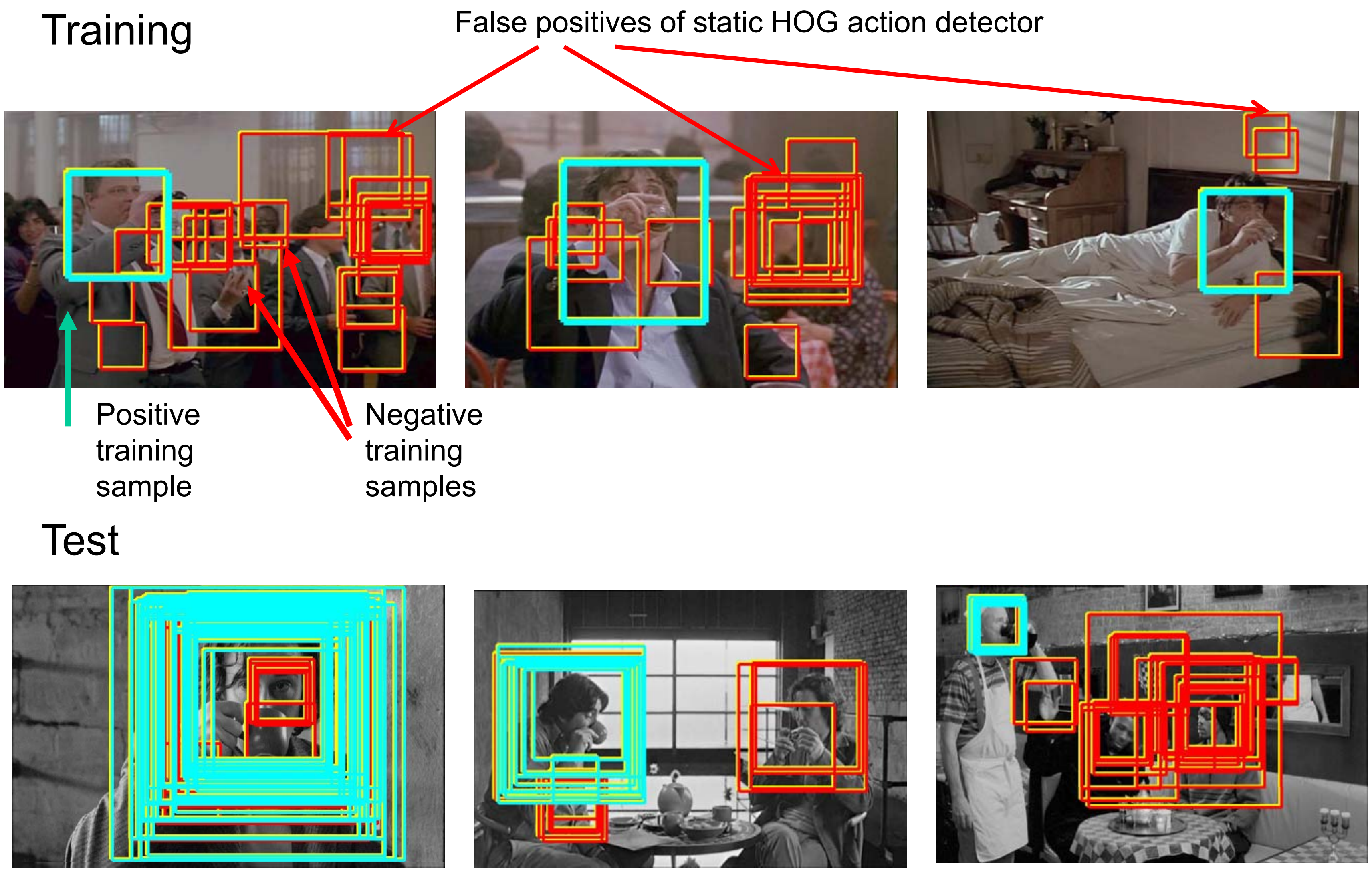


AdaBoost:
- Efficient discriminative classifier [Freund&Schapire'97]
- Good performance for face detection [Viola&Jones'01]



[Laptev, Pérez 2007]

Keyframe priming



Action detection

Test set:

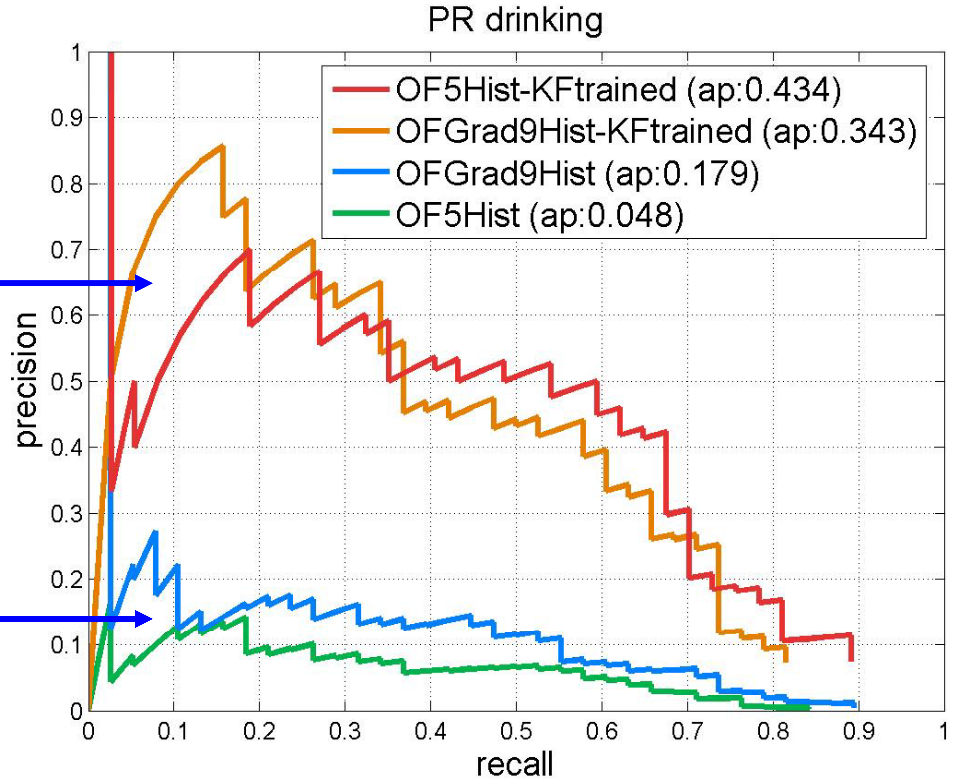
- 25min from “Coffee and Cigarettes” with GT 38 drinking actions
- No overlap with the training set in subjects or scenes

Detection:

- search over all space-time locations and spatio-temporal extents

Keyframe priming

No Keyframe priming



Action Detection (ICCV 2007)

Test episodes from the movie “Coffee and cigarettes”

Video available at http://www.irisa.fr/vista/Equipe/People/Laptev/actiondetection.html

20 most confident detections

Learning Actions from Movies

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

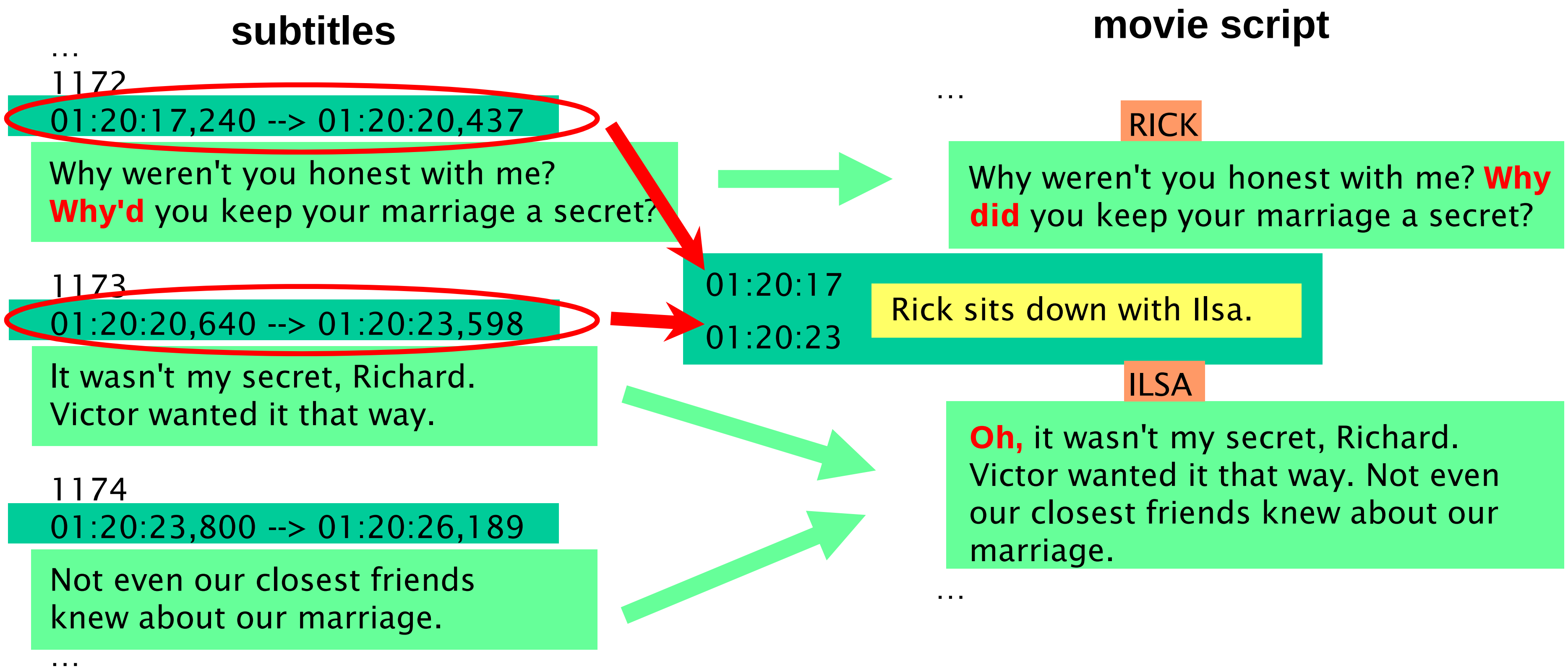
Problems:

- 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



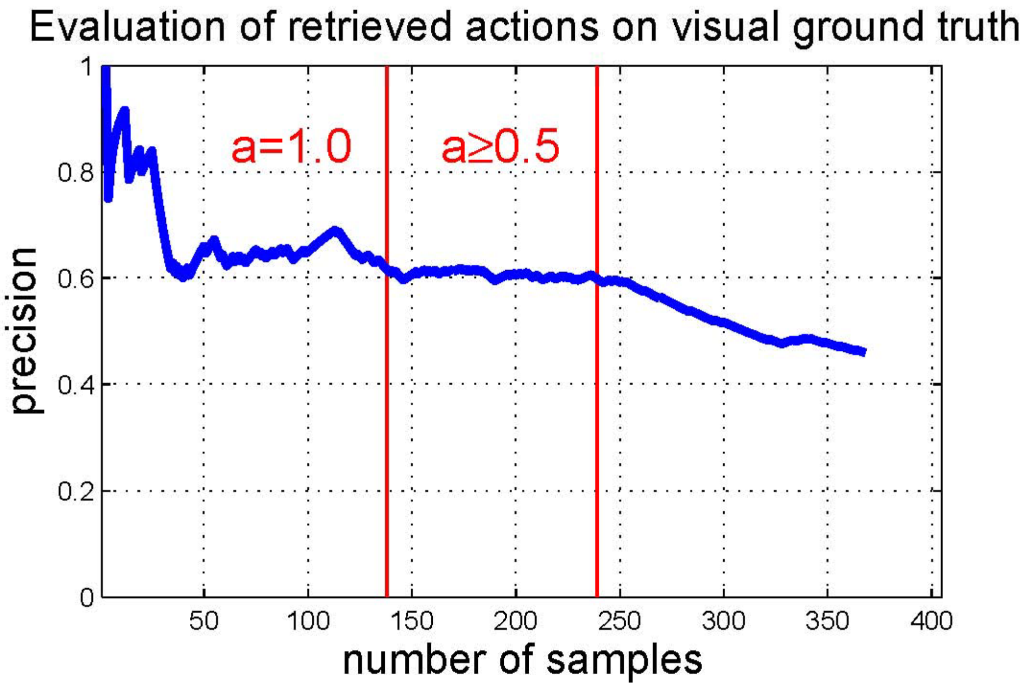
Script-based action annotation

- **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"

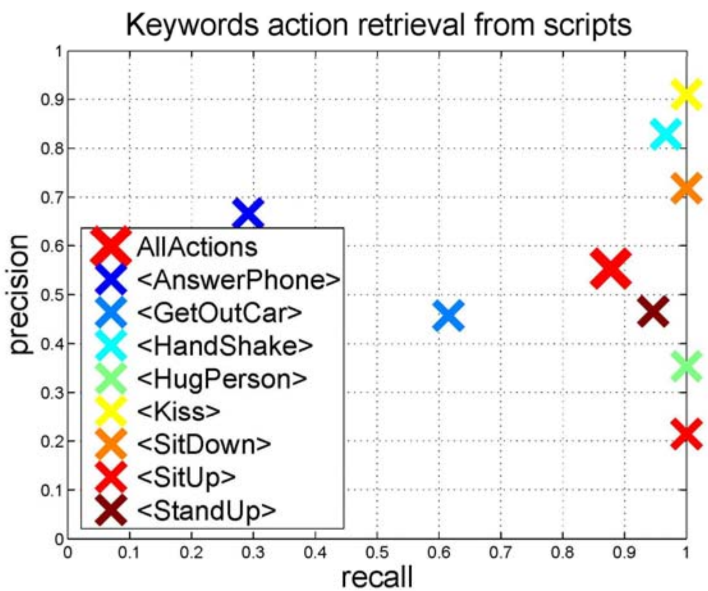
A black car pulls up, two army officers get out.

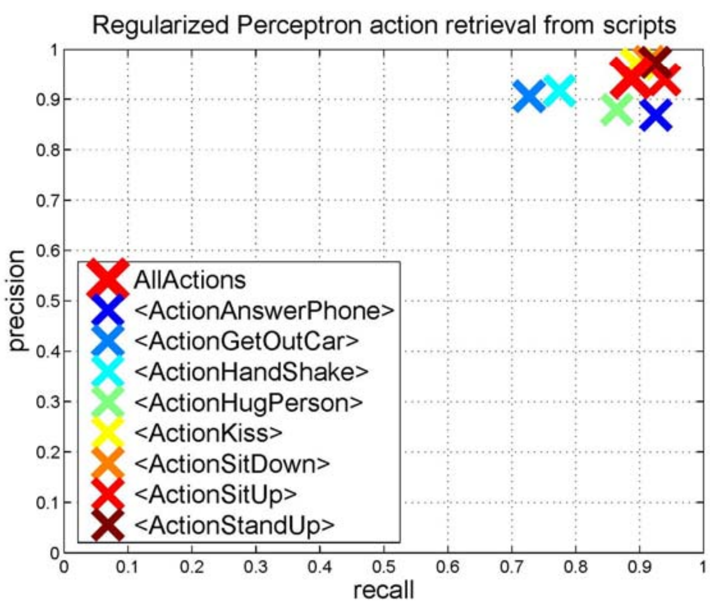
Text-based action retrieval

- 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

AnswerPhone

GetOutCar

HandShake

HugPerson

Kiss

SitDown

SitUp

StandUp

[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

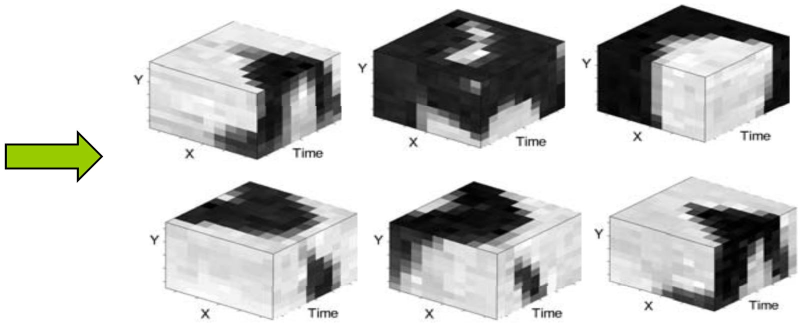
[Laptev, Marszałek, Schmid, Rozenfeld 2008]

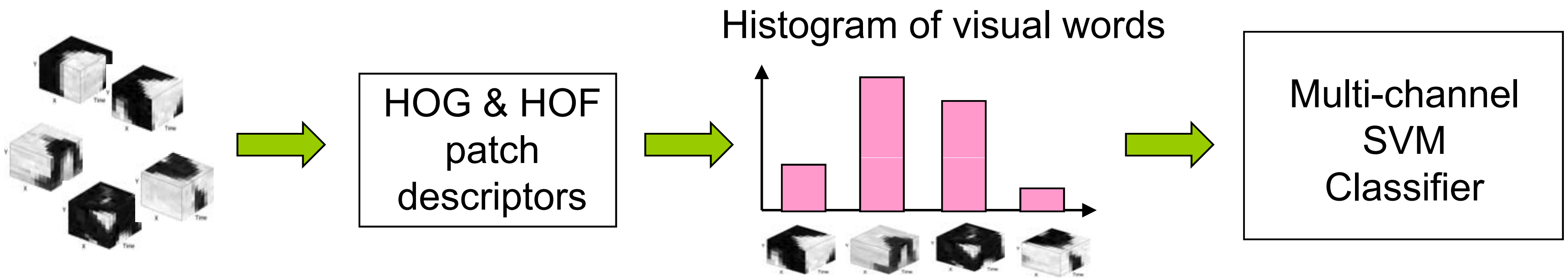
Action Classification: Overview

Bag of space-time features + multi-channel SVM

[Laptev'03, Schuldt'04, Niebles'06, Zhang'07]

Collection of space-time patches





Action classification (CVPR08)

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

Evaluation of local features for action recognition

- Local features provide a popular approach to video description for action recognition:
  - ~50% of recent action recognition methods (cvpr09, iccv09, bmvc09) are based on local features
  - Large variety of feature detectors and descriptors is available
  - Very limited and inconsistent comparison of different features

Goal:

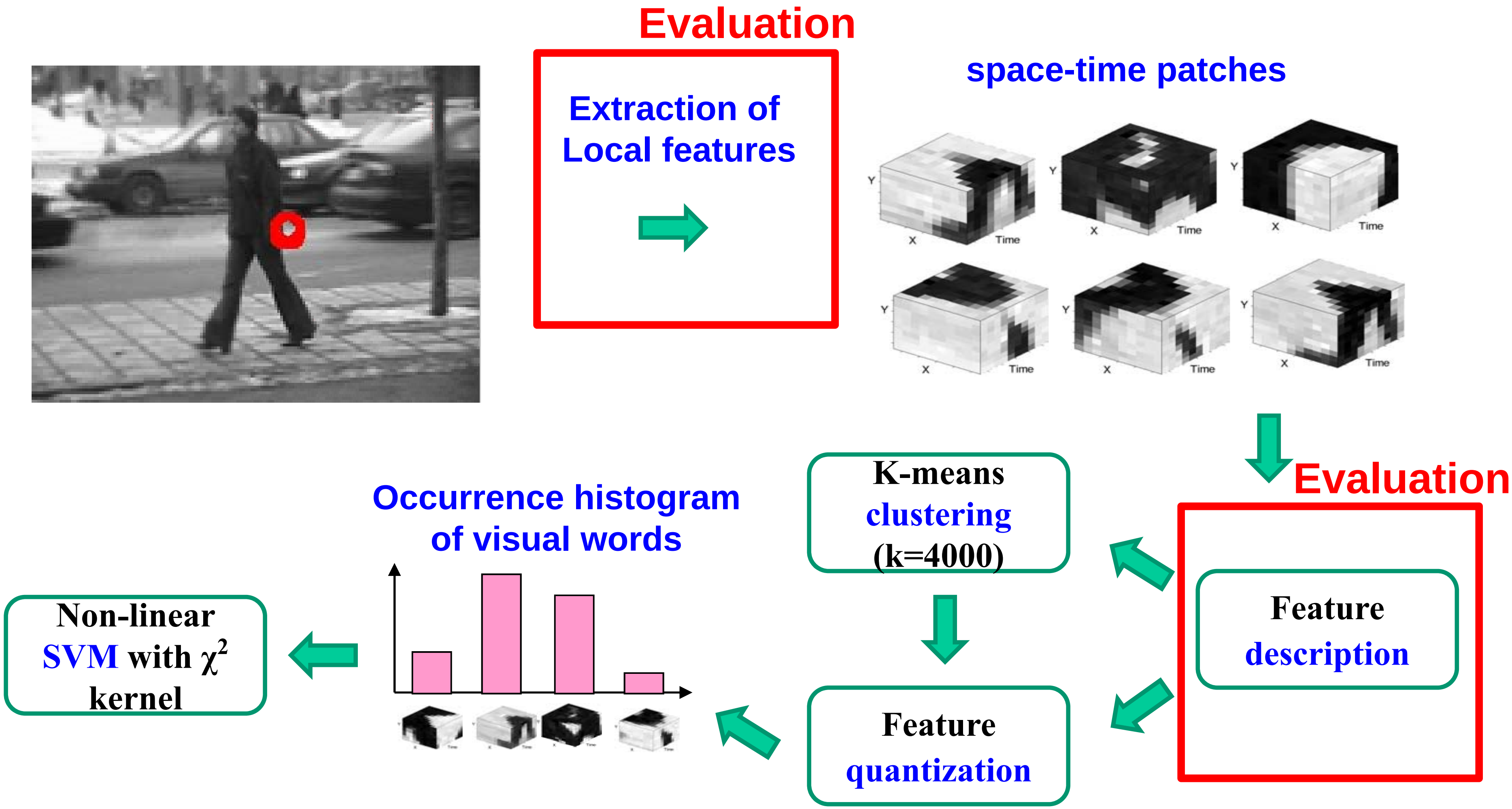
- Systematic evaluation of local *feature-descriptor* combinations
- Compare performance on common datasets
- Propose improvements

Evaluation of local features for action recognition

- Evaluation study [Wang et al. BMVC'09]
  - Common recognition framework
    - Same datasets (varying difficulty): KTH, UCF sports, Hollywood2
    - Same train/test data
    - Same classification method
  - Alternative local feature detectors and descriptors from recent literature
  - Comparison of different detector-descriptor combinations

Action recognition framework

Bag of space-time features + SVM [Schuldt'04, Niebles'06, Zhang'07]

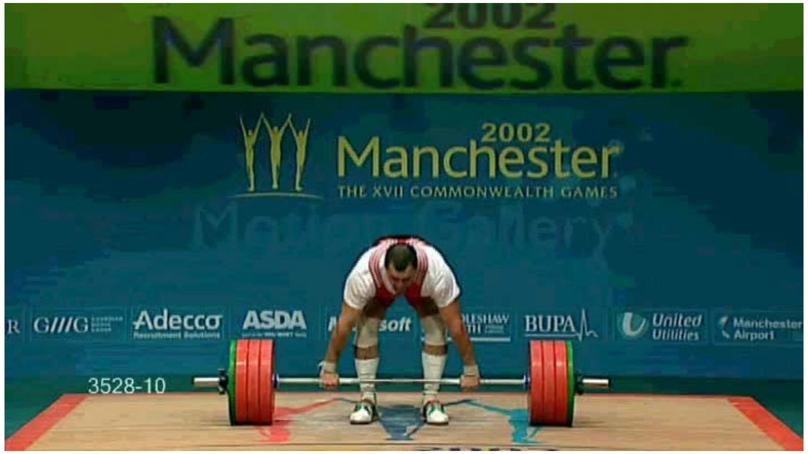


Local feature detectors/descriptors

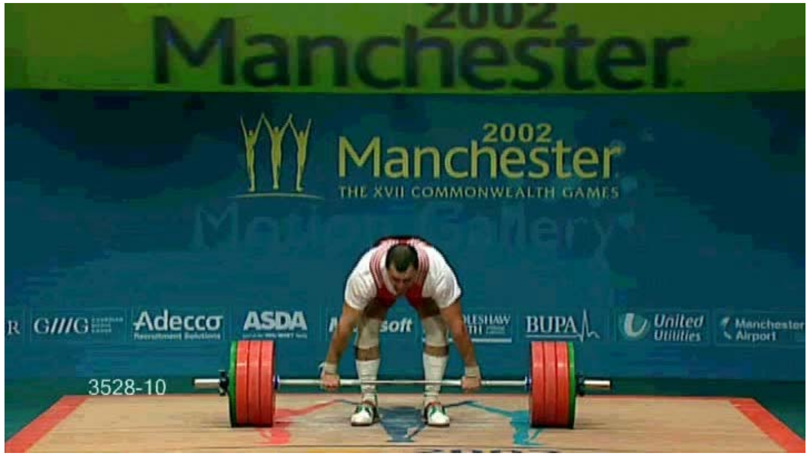
- Four types of detectors:
  - Harris3D [Laptev'05]
  - Cuboids [Dollar'05]
  - Hessian [Willems'08]
  - Regular dense sampling
- Four different types of descriptors:
  - HoG/HoF [Laptev'08]
  - Cuboids [Dollar'05]
  - HoG3D [Kläser'08]
  - Extended SURF [Willems'08]

Illustration of ST detectors

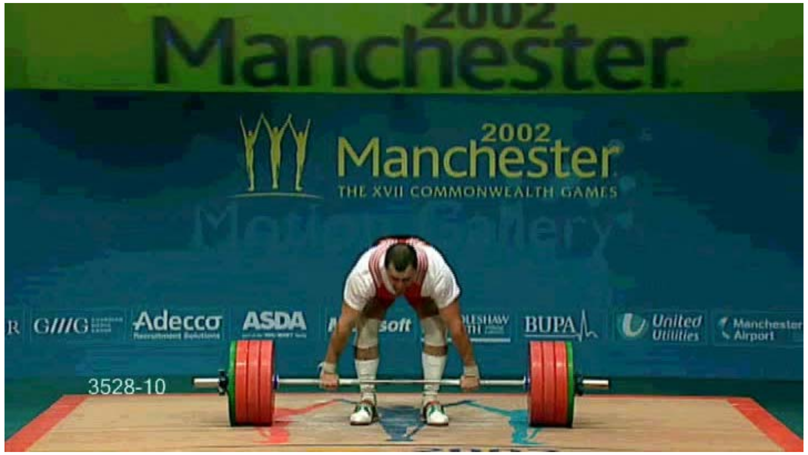
Harris3D



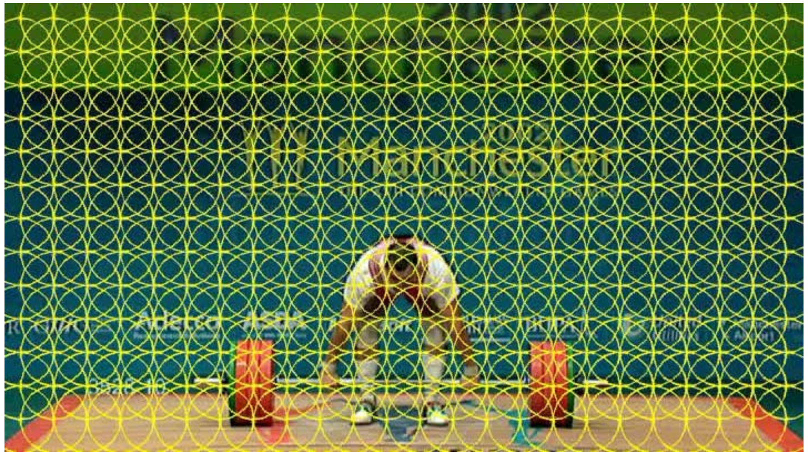
Hessian



Cuboid

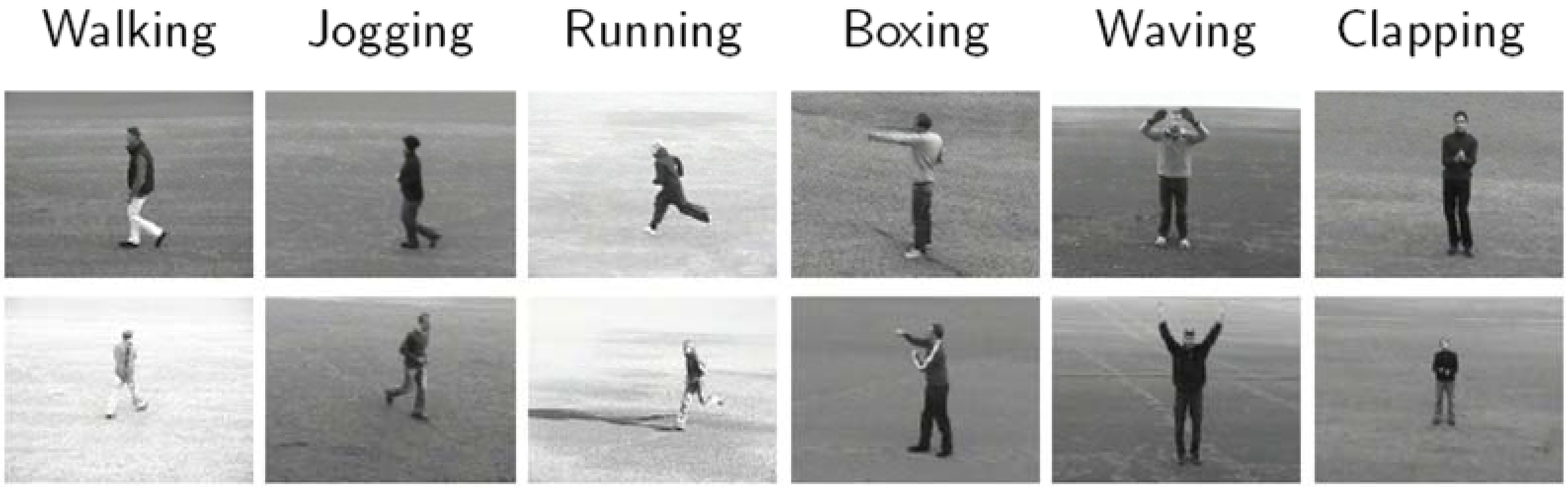


Dense



Dataset: KTH-Actions

- 6 action classes by 25 persons in 4 different scenarios
- Total of 2391 video samples
- Performance measure: average accuracy over all classes



UCF-Sports -- samples

- 10 different action classes
- 150 video samples in total
    - We extend the dataset by flipping videos
- Evaluation method: *leave-one-out*
- Performance measure: *average accuracy* over all classes

Diving

Kicking

Walking

Skateboarding

High-Bar-Swinging

Golf-Swinging

Dataset: Hollywood2

- 12 different action classes from 69 Hollywood movies
- 1707 video sequences in total
- Separate movies for training / testing
- Performance measure: mean average precision (mAP) over all classes

GetOutCar

AnswerPhone

Kiss

HandShake

StandUp

DriveCar

KTH-Actions -- results

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%	-

- Best results for **Sparse** Harris3D + HOF
- Good results for Harris3D and Cuboid detectors with HOG/HOF and HOG3D descriptors
- Dense features perform relatively poor compared to sparse features

UCF-Sports -- results

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%	-

- Best results for **Dense** + HOG3D
- Good results for Dense and HOG/HOF
- Cuboids: good performance with HOG3D

Hollywood2 -- results

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%	-

- Best results for **Dense** + HOG/HOF
- Good results for HOG/HOF

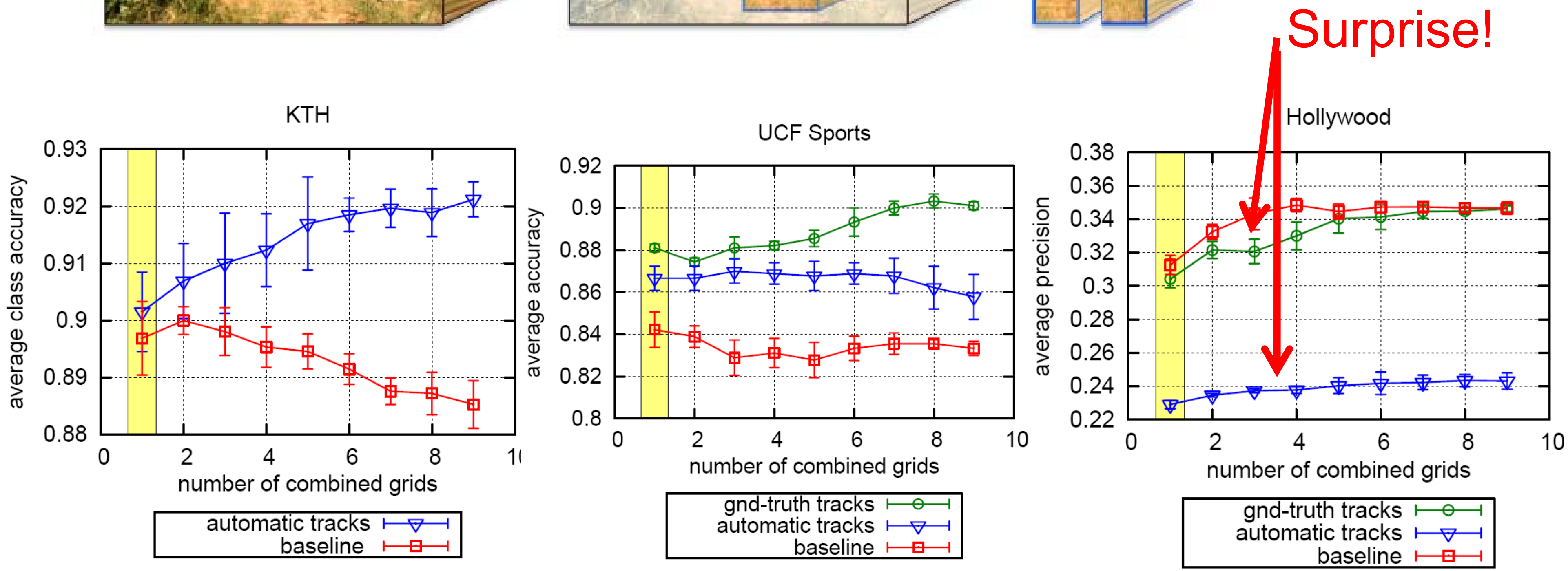
Evaluation summary

- *Dense sampling* consistently outperforms all the tested *sparse features* in realistic settings (UCF + Hollywood2)
  - Importance of realistic video data
  - Limitations of current feature detectors
  - Note: large number of features (15-20 times more)
- Sparse features provide more or less similar results (sparse features better than Dense on KTH)
- Descriptors' performance
  - Combination of gradients + optical flow seems a good choice (HOG/HOF & HOG3D)

How to improve BoF classification?

Actions are about people

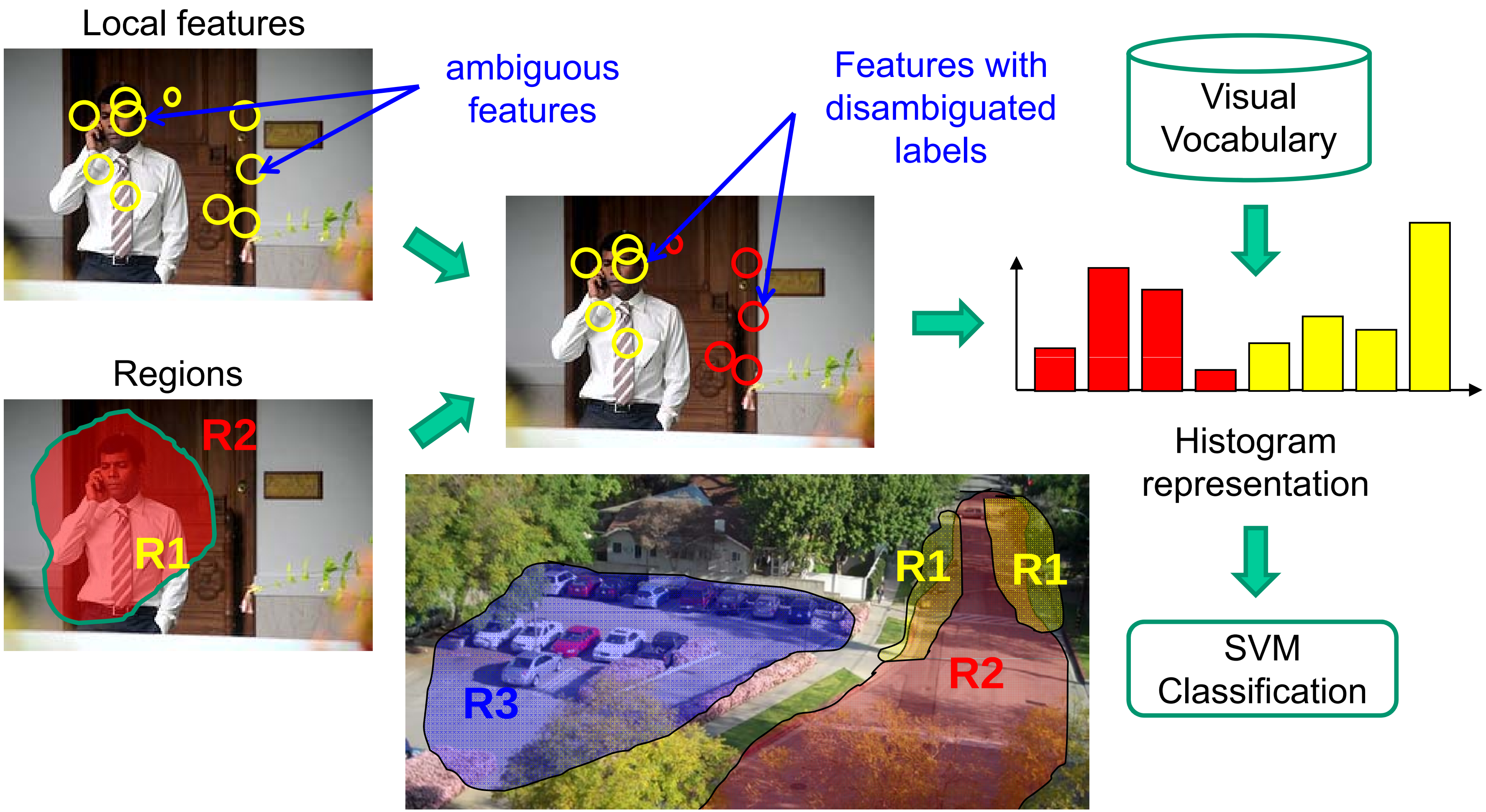
➡ Why not try to combine BoF with person detection?



How to improve BoF classification?

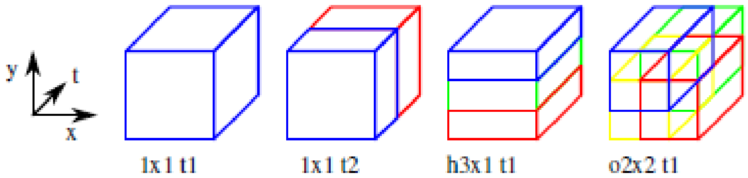
2nd attampt:

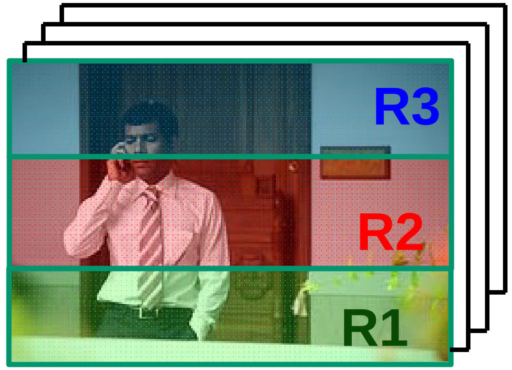
- Do not remove background
- Improve local descriptors with region-level information



Video Segmentation

- Spatio-temporal grids



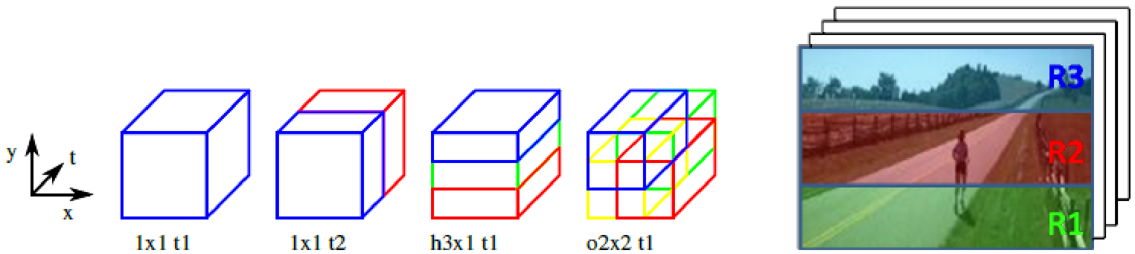


- Static action detectors [Felzenszwalb'08]
  - *Trained from ~100 web-images per class*

AnswerPhone DriveCar Eat HandShake HugPerson Kiss Run Sitting

- Object and Person detectors (Upper body) [Felzenszwalb'08]

Video Segmentation



Hollywood-2 action classification

Attributed feature	Performance (meanAP)
BoF	48.55
Spatiotemoral grid 24 channels	**51.83**
Motion segmentation	50.39
Upper body	49.26
Object detectors	49.89
Action detectors	**52.77**
Spatiotemoral grid + Motion segmentation	53.20
Spatiotemoral grid + Upper body	53.18
Spatiotemoral grid + Object detectors	52.97
Spatiotemoral grid + Action detectors	**55.72**
Spatiotemoral grid + Motion segmentation + Upper body + Object detectors + Action detectors	55.33

Hollywood-2 action classification

Channels	BoF	STG24	AD-class	STG24 + AD-class	STG24 + MS8 + AD-class + UB + OD
mean AP	48.55%	51.83%	52.77%	**55.72%**	55.33%
AnswerPhone	15.71%	25.87%	20.75%	**26.32%**	24.77%
DriveCar	87.61%	85.91%	86.87%	86.48%	**88.11%**
Eat	54.77%	56.39%	57.38%	59.19%	**61.42%**
FightPerson	73.90%	74.93%	75.73%	76.21%	**76.47%**
GetOutCar	33.35%	44.02%	38.26%	45.71%	**47.42%**
HandShake	19.99%	29.68%	45.71%	**49.73%**	38.41%
HugPerson	37.80%	**46.08%**	40.75%	45.41%	44.58%
Kiss	52.12%	54.96%	56.00%	58.96%	**61.47%**
Run	71.13%	69.40%	73.18%	71.97%	**74.31%**
SitDown	59.01%	58.89%	59.59%	**62.43%**	61.26%
SitUp	23.90%	18.40%	24.06%	**27.52%**	25.50%
StandUp	53.30%	57.41%	54.94%	58.76%	**60.41%**

Actions in Context (CVPR 2009)

- 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

ILSA

01:22:00
01:22:03

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.

CARL

01:22:15
01:22:17

I think we lost them.

...

Mining scene captions

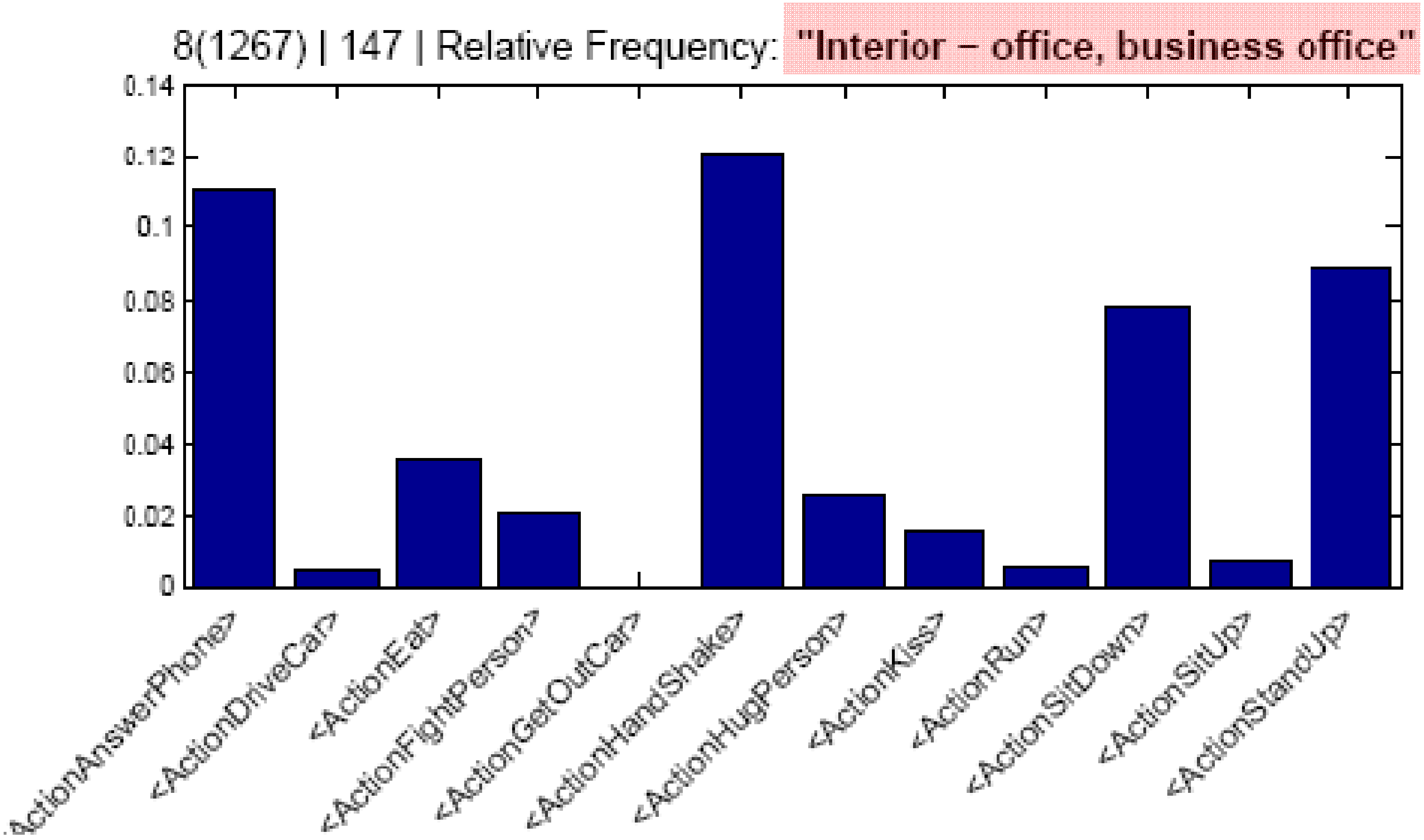
INT. TRENDY RESTAURANT - NIGHT
INT. MARSELLUS WALLACE'S DINING ROOM MORNING
EXT. STREETS BY DORA'S HOUSE - DAY.
INT. MELVIN'S APARTMENT, BATHROOM – NIGHT
EXT. NEW YORK CITY STREET NEAR CAROL'S RESTAURANT – DAY
INT. CRAIG AND LOTTE'S BATHROOM - DAY

- Maximize word frequency ⇒ street, living room, bedroom, car ….
- Merge words with similar senses using WordNet:

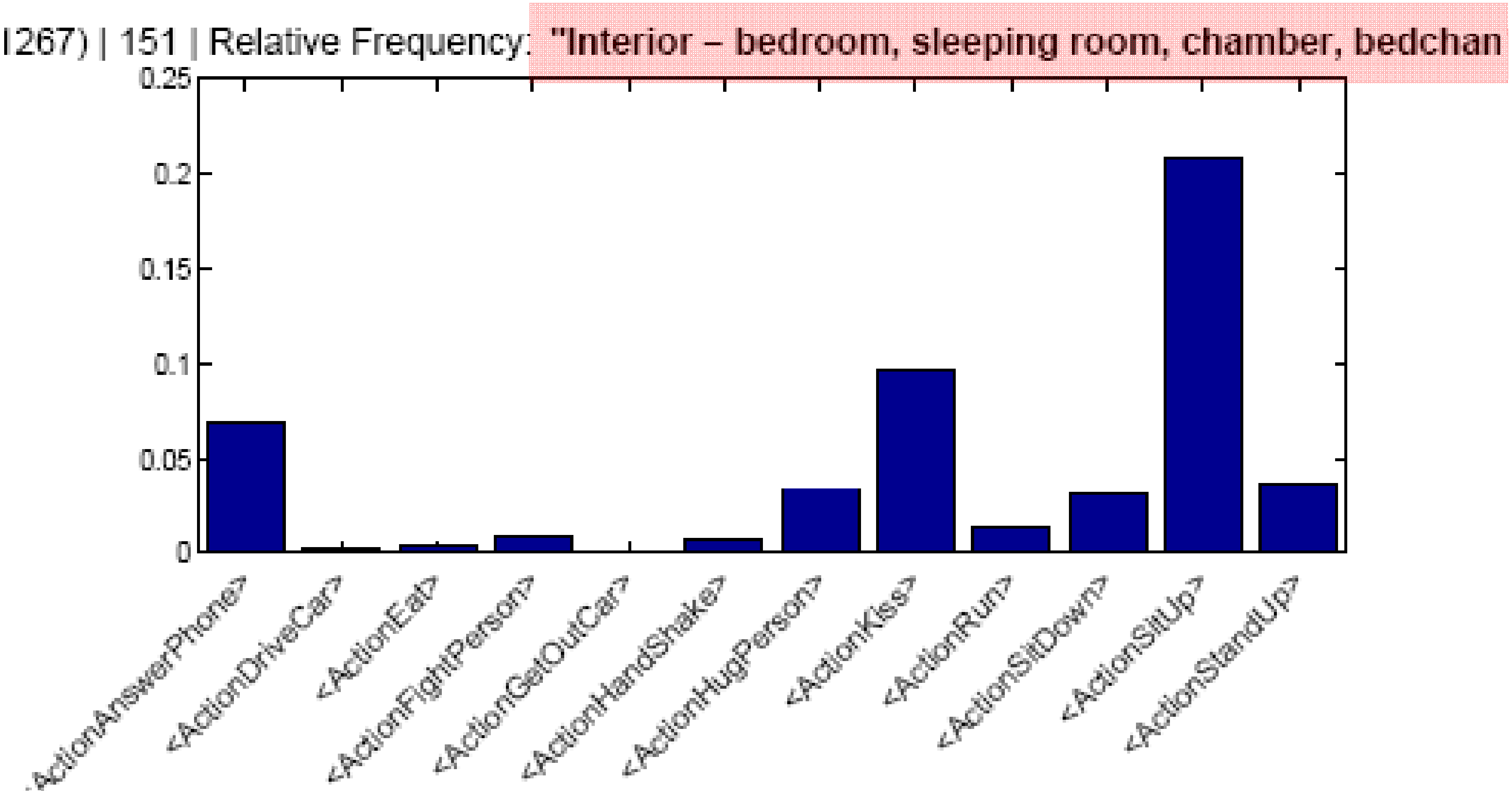
  taxi -> car, cafe -> restaurant

- Measure correlation of words with actions (in scripts) and
- Re-sort words by the entropy $S = -k \sum P_i \ln P_i$
  for P = p(action | word)

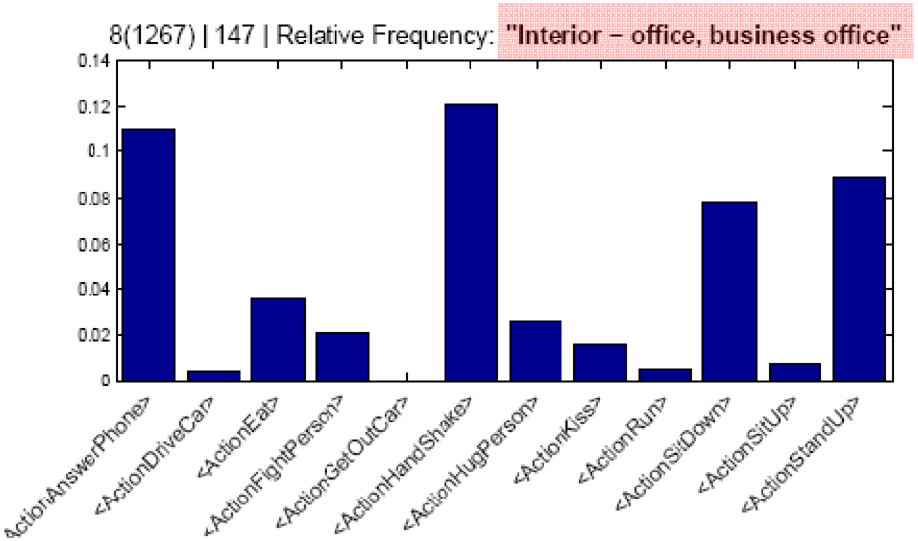
Co-occurrence of actions and scenes in scripts

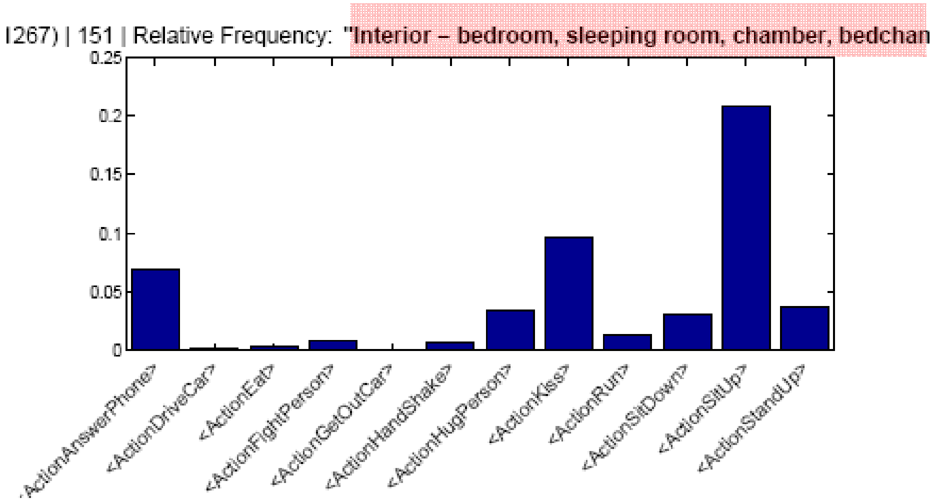


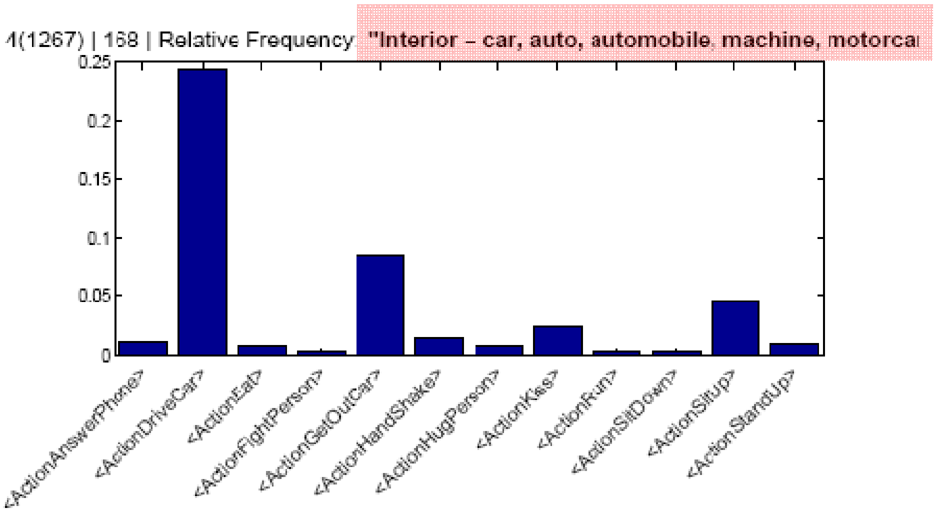
Co-occurrence of actions and scenes in scripts

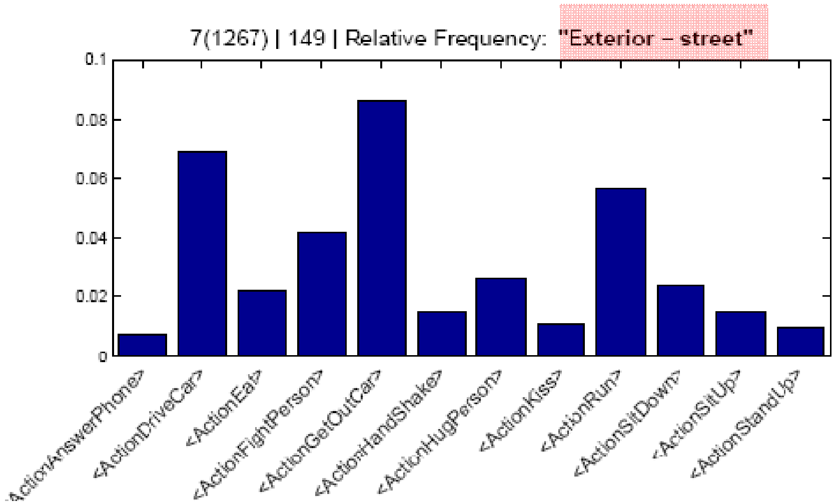


Co-occurrence of actions and scenes in scripts

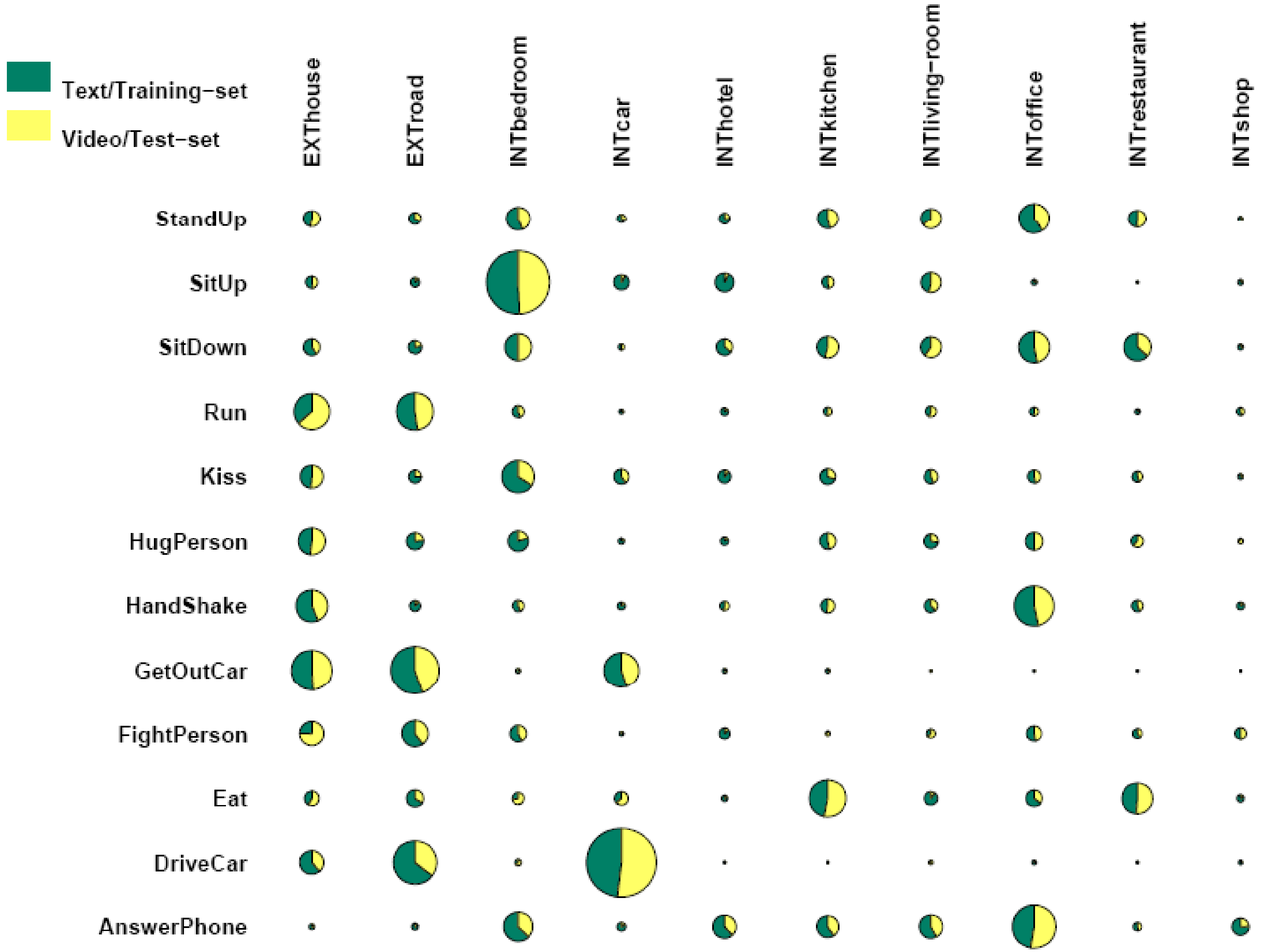








Co-occurrence of actions and scenes in text vs. video



Automatic gathering of relevant scene classes and visual samples

	Auto-Train-Actions	Clean-Test-Actions
AnswerPhone	59	64
DriveCar	90	102
Eat	44	33
FightPerson	33	70
GetOutCar	40	57
HandShake	38	45
HugPerson	27	66
Kiss	125	103
Run	187	141
SitDown	87	108
SitUp	26	37
StandUp	133	146
All Samples	810	884

(a) Actions

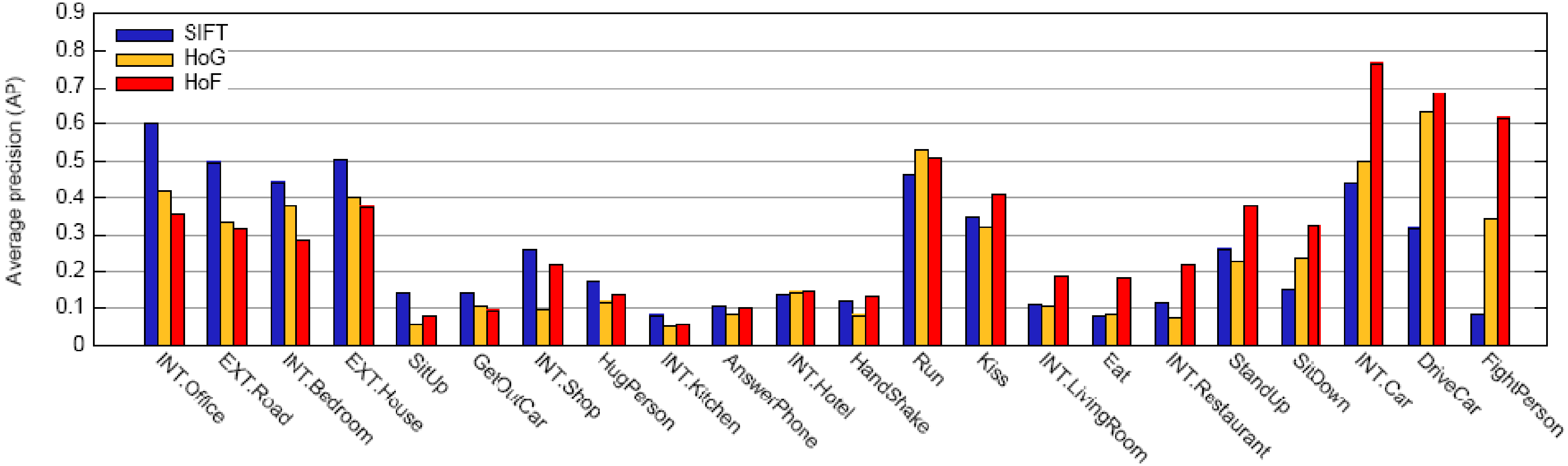
	Auto-Train-Scenes	Clean-Test-Scenes
EXT-house	81	140
EXT-road	81	114
INT-bedroom	67	69
INT-car	44	68
INT-hotel	59	37
INT-kitchen	38	24
INT-living-room	30	51
INT-office	114	110
INT-restaurant	44	36
INT-shop	47	28
All Samples	570	582

(b) Scenes

Source:
69 movies
aligned with
the scripts

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

Results: actions and scenes (separately)



EXT.House	**0.503**	0.363	0.491
EXT.Road	**0.498**	0.372	0.389
INT.Bedroom	**0.445**	0.362	**0.462**
INT.Car	0.444	**0.759**	**0.773**
INT.Hotel	0.141	**0.220**	**0.250**
INT.Kitchen	**0.081**	0.050	0.070
INT.LivingRoom	0.109	**0.128**	**0.152**
INT.Office	**0.602**	0.453	0.574
INT.Restaurant	**0.112**	0.103	0.108
INT.Shop	**0.257**	0.149	0.244
Scene average	*0.319*	*0.296*	*0.351*
Total average	*0.259*	*0.310*	*0.339*

	SIFT	HoG HoF	SIFT HoG HoF
AnswerPhone	**0.105**	0.088	**0.107**
DriveCar	0.313	**0.749**	0.750
Eat	0.082	**0.263**	**0.286**
FightPerson	0.081	**0.675**	0.571
GetOutCar	**0.191**	0.090	**0.116**
HandShake	**0.123**	0.116	**0.141**
HugPerson	0.129	**0.135**	**0.138**
Kiss	0.348	**0.496**	**0.556**
Run	0.458	**0.537**	**0.565**
SitDown	0.161	**0.316**	0.278
SitUp	**0.142**	0.072	**0.078**
StandUp	0.262	**0.350**	0.325
Action average	*0.200*	*0.324*	*0.326*

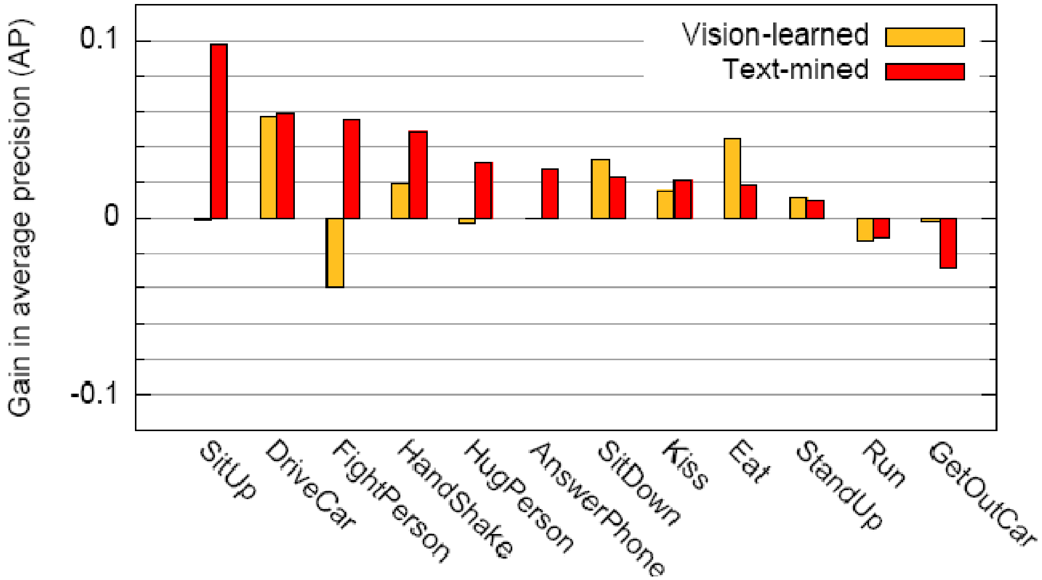
Classification with the help of context

$$a_i'(\boldsymbol{x}) = a_i(\boldsymbol{x}) + \tau \sum_{j \in \mathcal{S}} w_{ij} s_j(\boldsymbol{x})$$

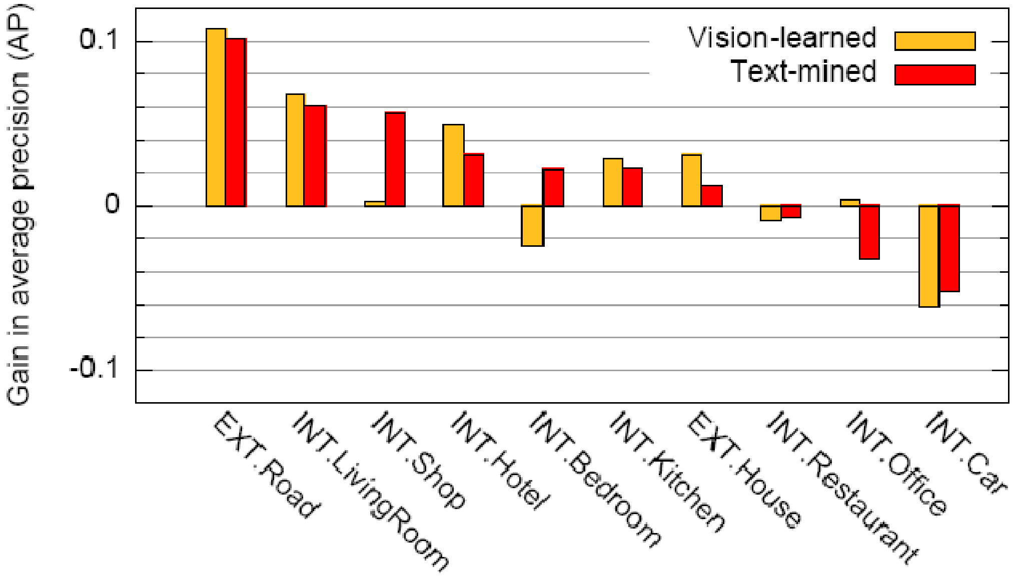
$a_i(\boldsymbol{x})$	Action classification score
$s_j(\boldsymbol{x})$	Scene classification score
w_{ij}	Weight, estimated from text: $p(Scene \mid Action)$
$a_i'(\boldsymbol{x})$	New action score

Results: actions and scenes (jointly)

Actions in the context of Scenes

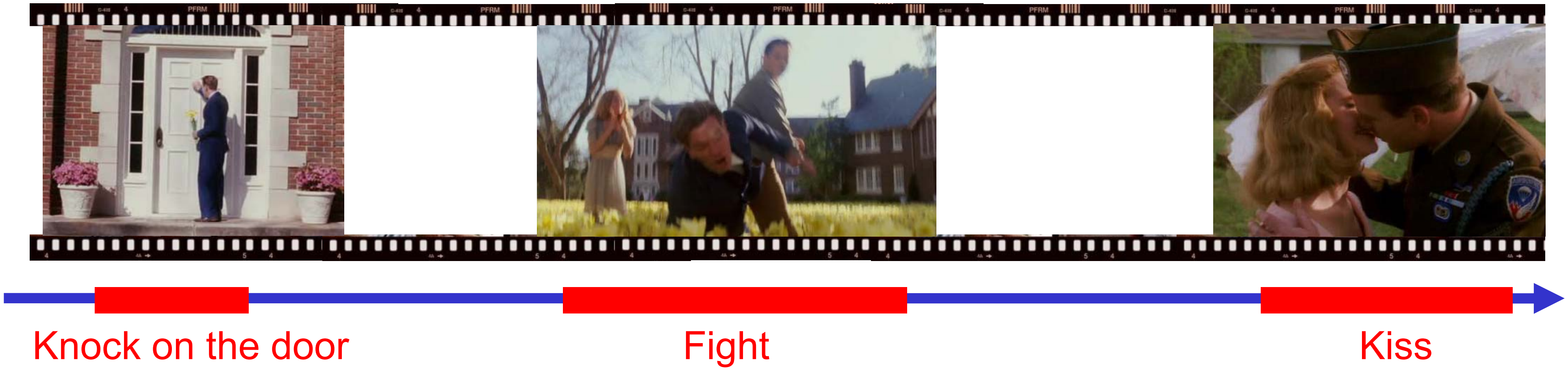


Scenes in the context of Actions



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

WHAT actions?

- Automatic discovery of action classes in text (movie scripts)

  -- Text processing:

  > *Part of Speech (POS) tagging;*
  > *Named Entity Recognition (NER);*
  > *WordNet pruning; Visual Noun filtering*

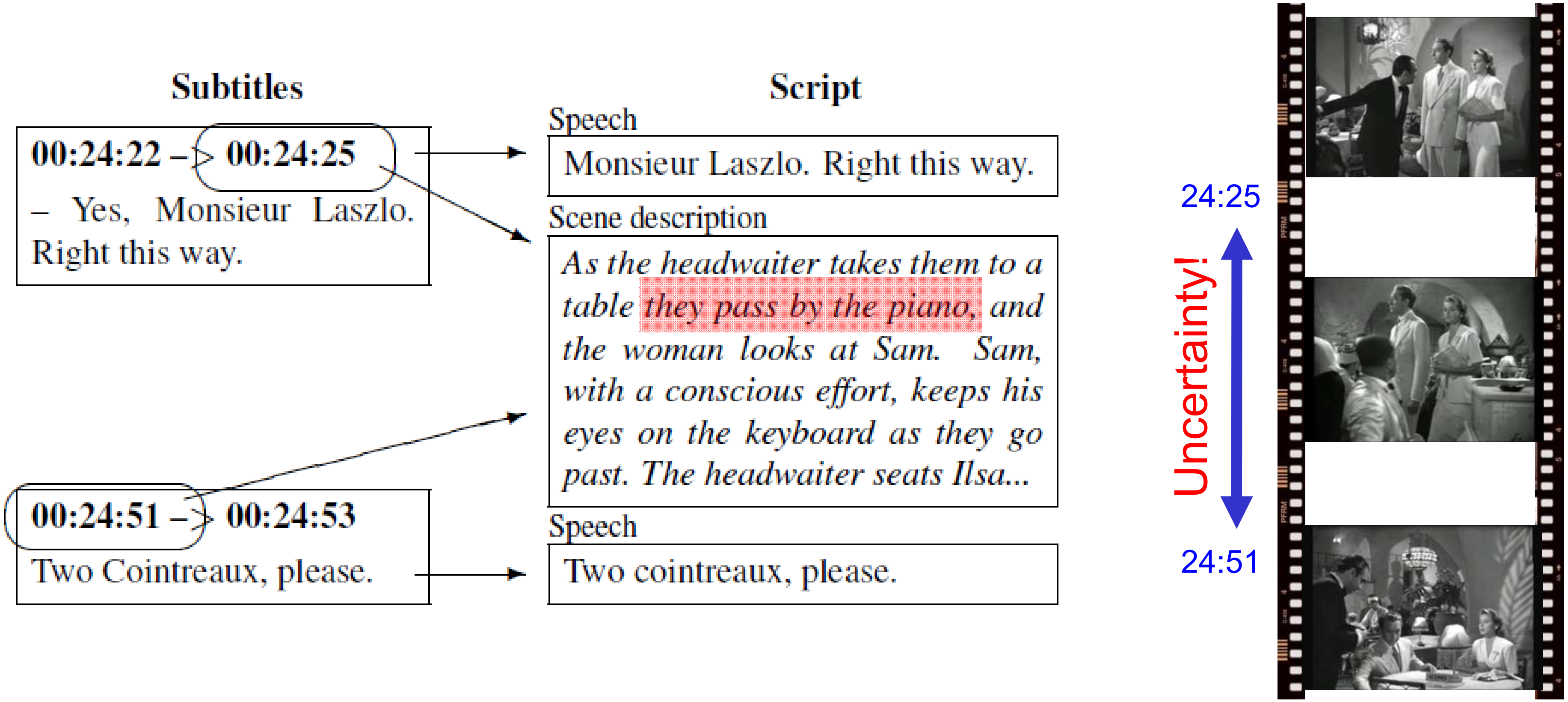
  -- Search action patterns

Person+Verb	Person+Verb+Prep.	Person+Verb+Prep+Vis.Noun
3725 /PERSON .* is	989 /PERSON .* looks .* at	41 /PERSON .* sits .* in .* chair
2644 /PERSON .* looks	384 /PERSON .* is .* in	37 /PERSON .* sits .* at .* table
1300 /PERSON .* turns	363 /PERSON .* looks .* up	31 /PERSON .* sits .* on .* bed
916 /PERSON .* takes	234 /PERSON .* is .* on	29 /PERSON .* sits .* at .* desk
840 /PERSON .* sits	215 /PERSON .* picks .* up	26 /PERSON .* picks .* up .* phone
829 /PERSON .* has	196 /PERSON .* is .* at	23 /PERSON .* gets .* out .* car
807 /PERSON .* walks	139 /PERSON .* sits .* in	23 /PERSON .* looks .* out .* window
701 /PERSON .* stands	138 /PERSON .* is .* with	21 /PERSON .* looks .* around .* room
622 /PERSON .* goes	134 /PERSON .* stares .* at	18 /PERSON .* is .* at .* desk
591 /PERSON .* starts	129 /PERSON .* is .* by	17 /PERSON .* hangs .* up .* phone
585 /PERSON .* does	126 /PERSON .* looks .* down	17 /PERSON .* is .* on .* phone
569 /PERSON .* gets	124 /PERSON .* sits .* on	17 /PERSON .* looks .* at .* watch
552 /PERSON .* pulls	122 /PERSON .* is .* of	16 /PERSON .* sits .* on .* couch
503 /PERSON .* comes	114 /PERSON .* gets .* up	15 /PERSON .* opens .* of .* door
493 /PERSON .* sees	109 /PERSON .* sits .* at	15 /PERSON .* walks .* into .* room
462 /PERSON .* are/VBP	107 /PERSON .* sits .* down	14 /PERSON .* goes .* into .* room

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:

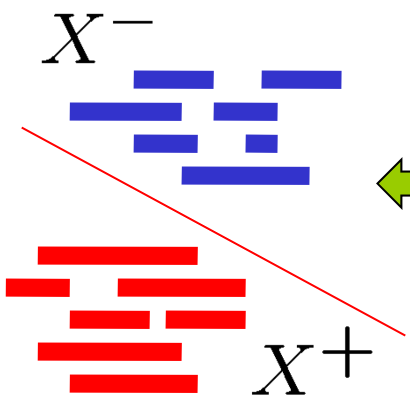
- Action type, e.g.
  Person Opens Door
- Videos + aligned scripts

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

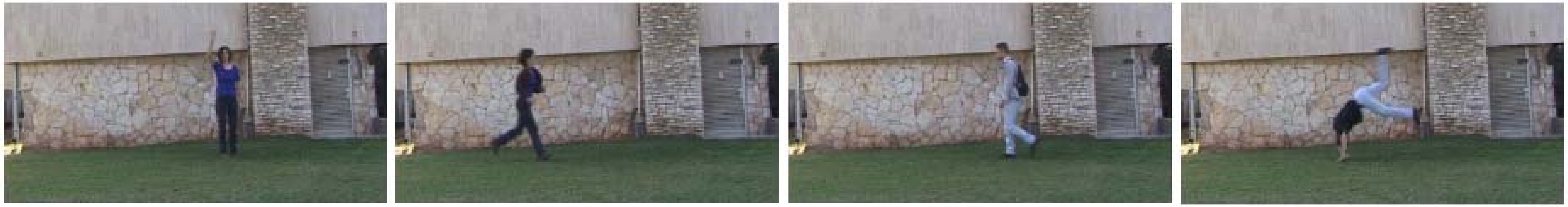


Output:

Sliding-window-style temporal action localization

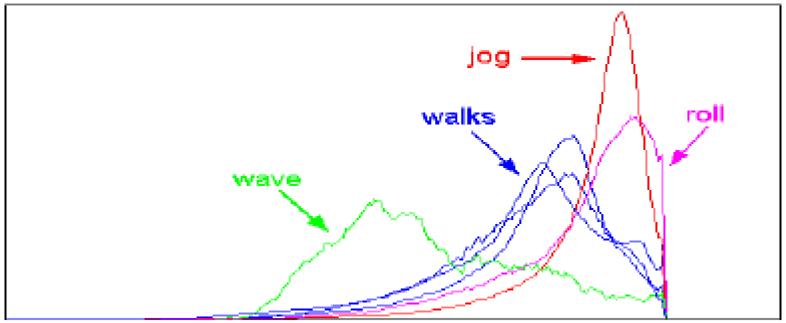
Action clustering

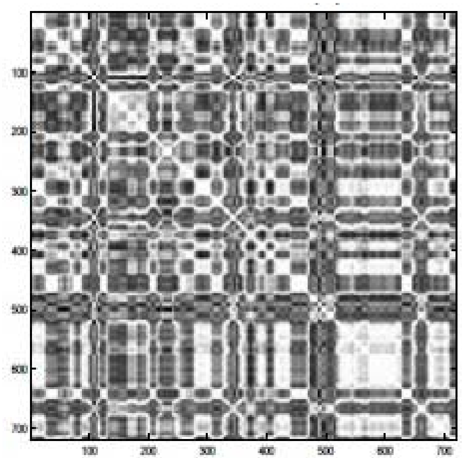
[Lihi Zelnik-Manor and Michal Irani CVPR 2001]

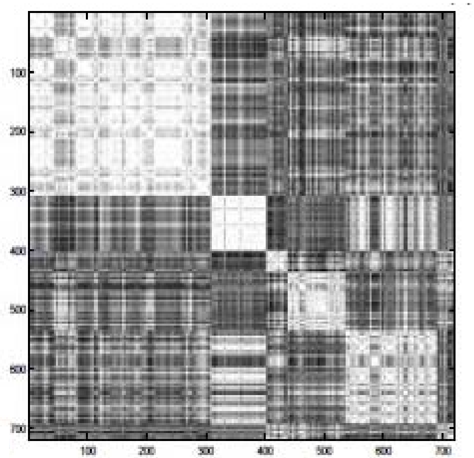


Spectral clustering

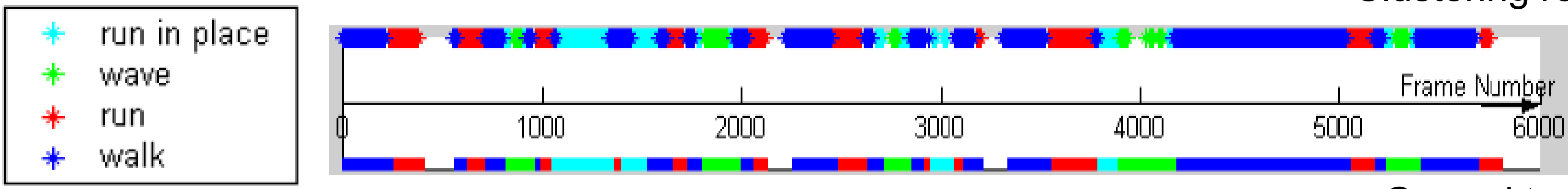
Descriptor space







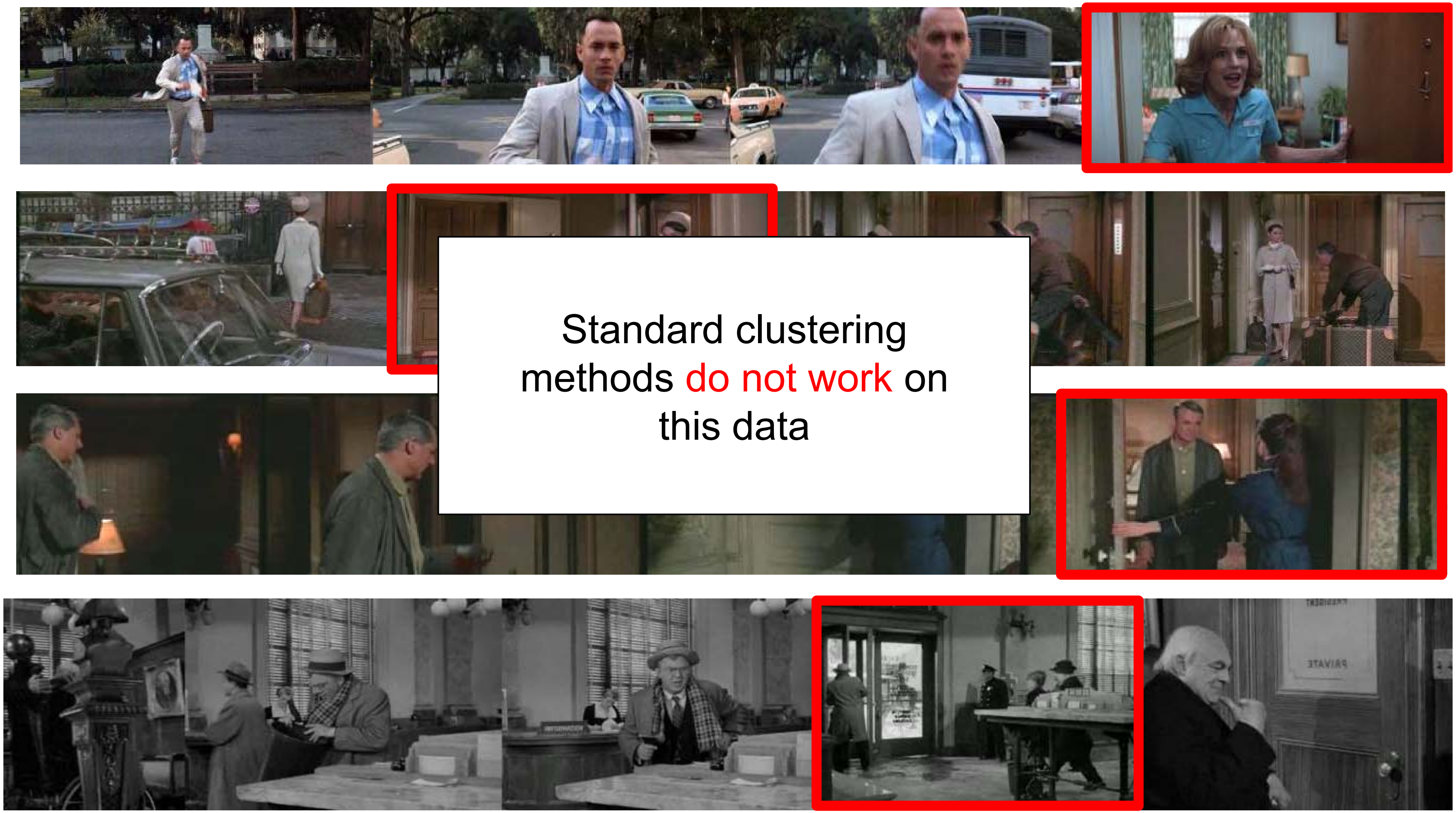
Clustering results



Ground truth

Action clustering

Complex data:

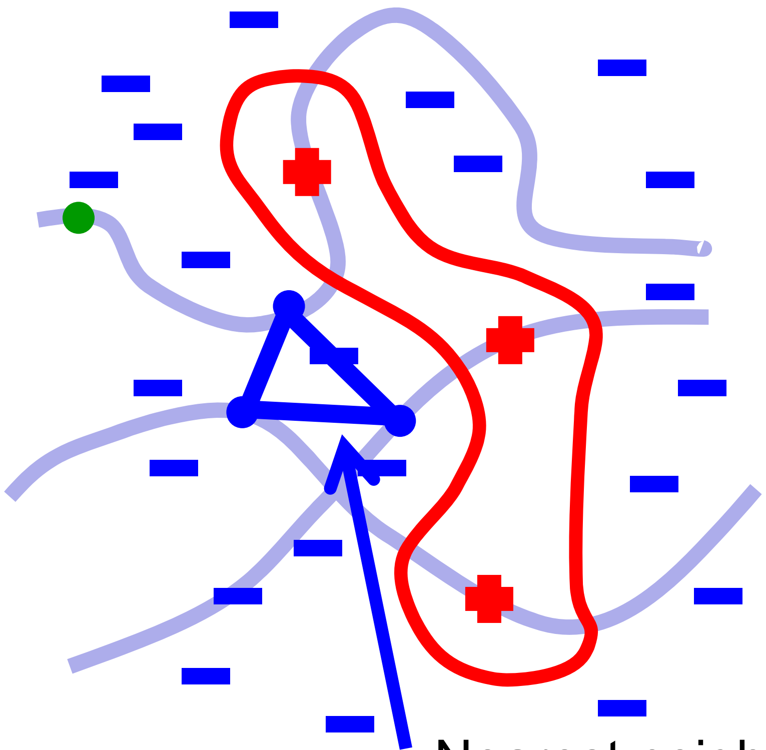


Standard clustering methods do not work on this data

Action clustering

Our view at the problem

Feature space



Nearest neighbor solution: wrong!

Video space

Negative samples!

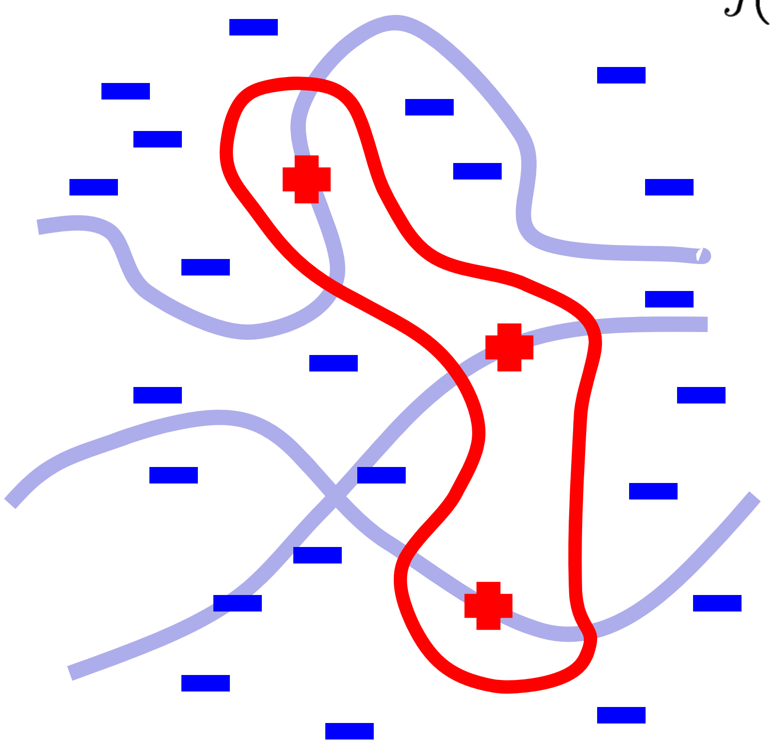
Random video samples: lots of them, very low chance to be positives

Action clustering

Formulation

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

Feature space



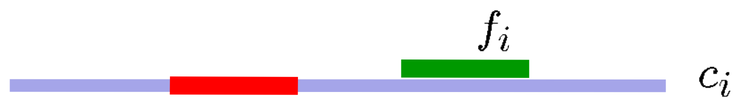
discriminative cost

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

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

x_i^- negative samples

$c_i[f_i]$ parameterized positive samples

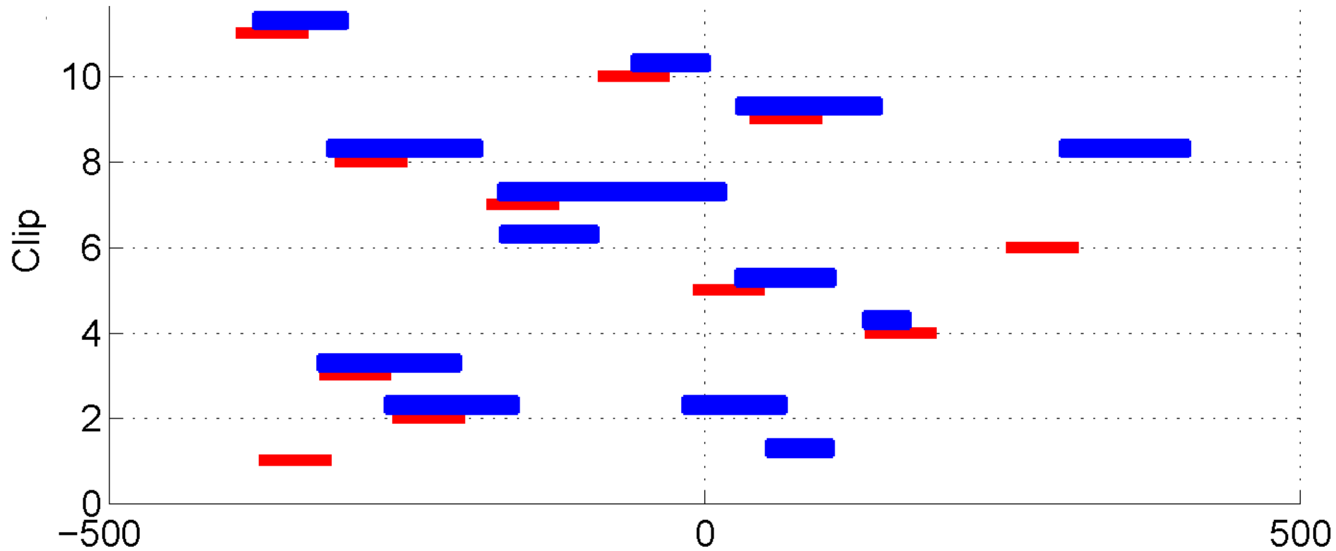


Optimization

SVM solution for w, b
Coordinate descent on f_i

Clustering results

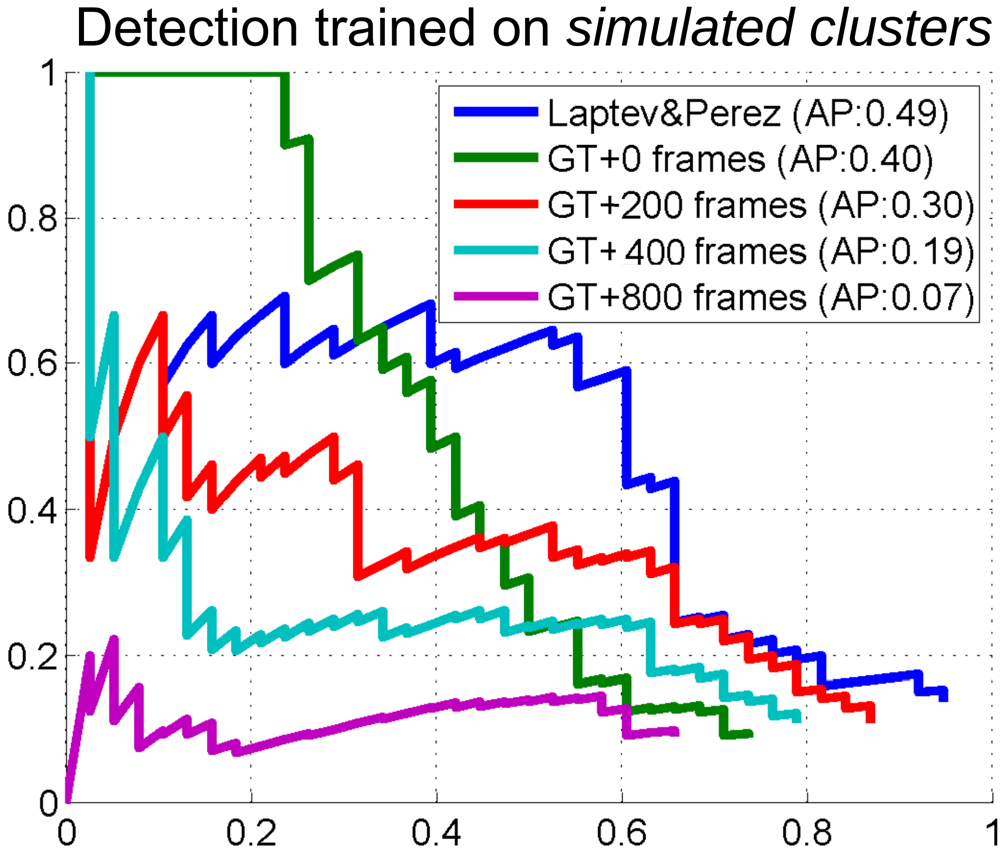
Drinking actions in Coffee and Cigarettes

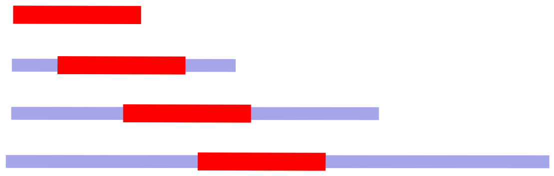


Detection results

Drinking actions in Coffee and Cigarettes

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





Test set:

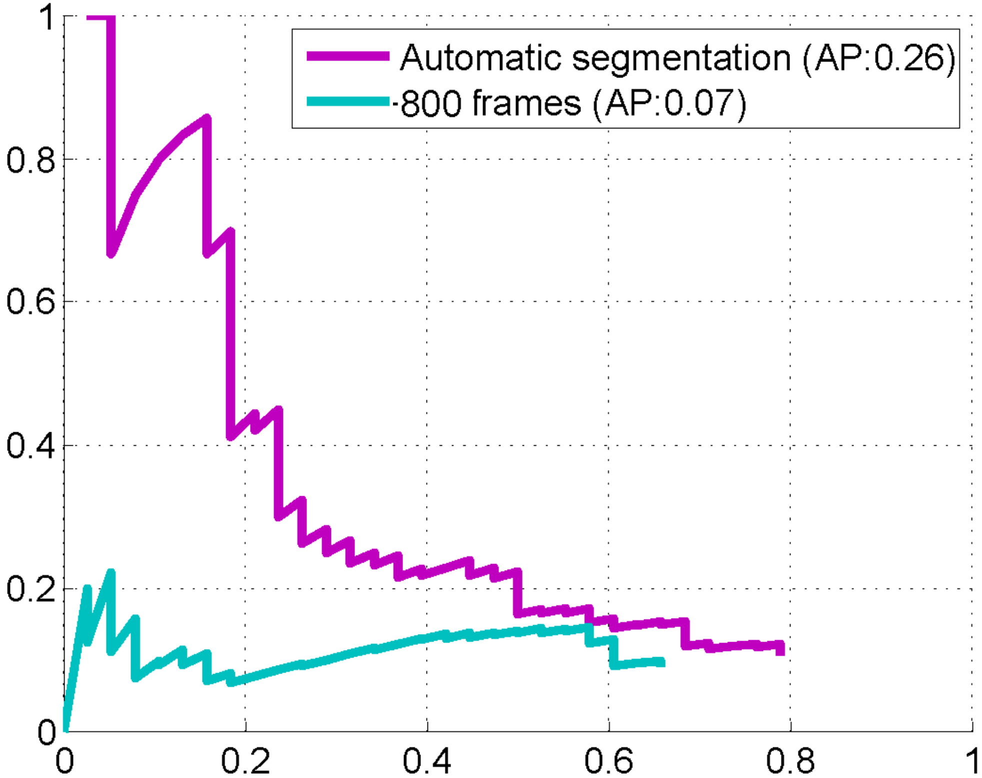
- 25min from "Coffee and Cigarettes" with GT 38 drinking actions

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*

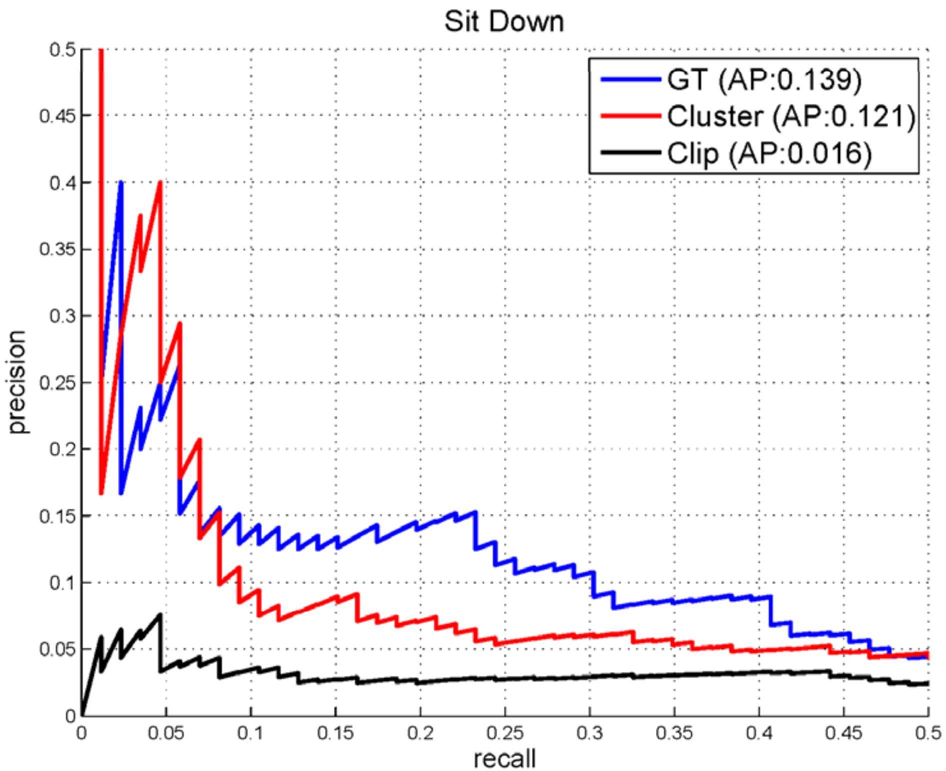


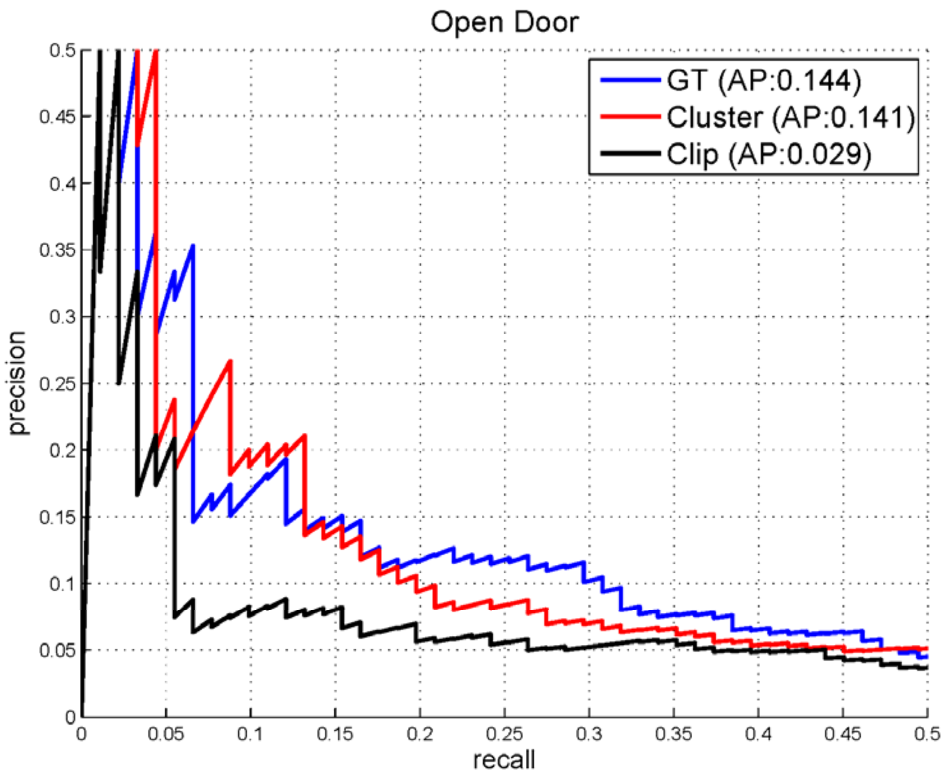
Test set:

- 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

Conclusions

- Bag-of-words models are currently dominant, the structure (human poses, etc.) should be integrated.
- Vocabulary of actions is not well-defined – it depends on the goal and the task
- Actions should be used for the functional interpretation of the visual world