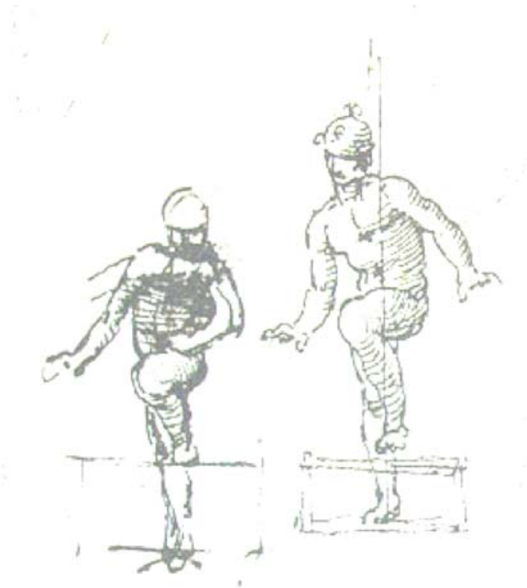


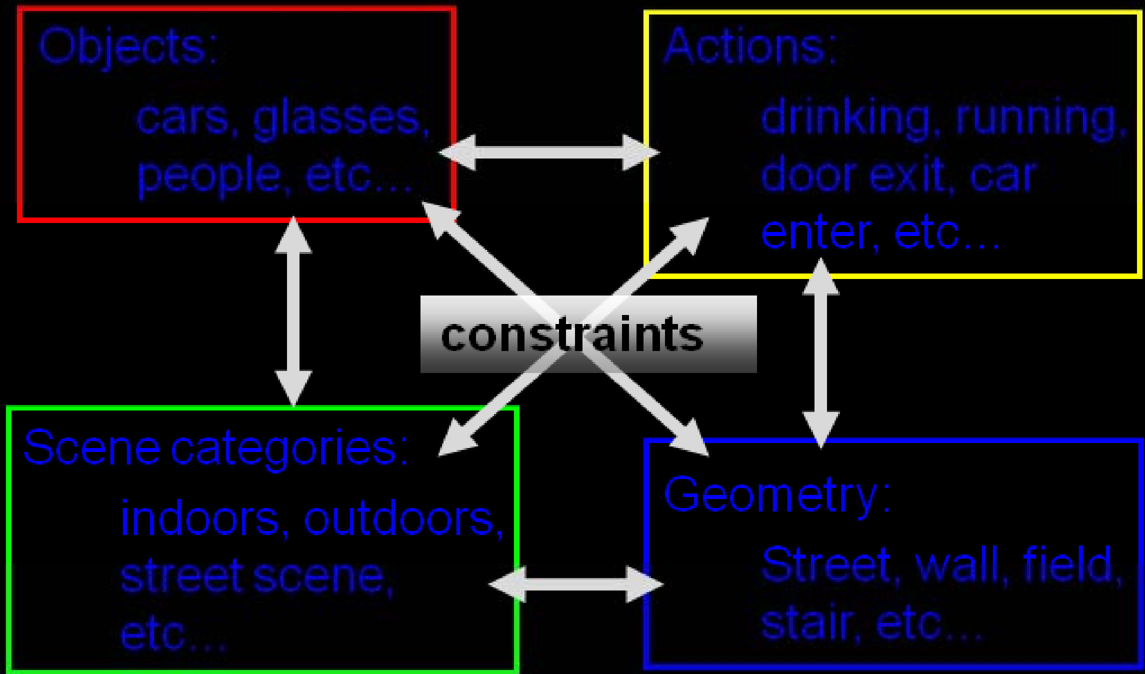
11th European Conference on Computer Vision
Hersonissos, Heraklion, Crete, Greece
September 5, 2010

Tutorial on

Statistical and Structural Recognition of Human Actions

Ivan Laptev and Greg Mori

Computer vision grand challenge: Video understanding

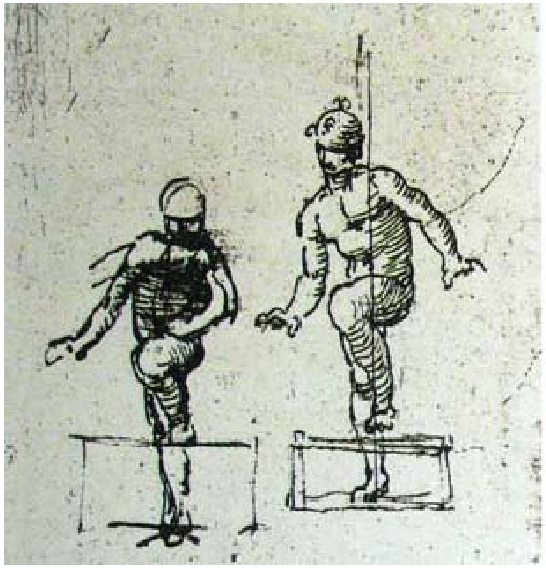


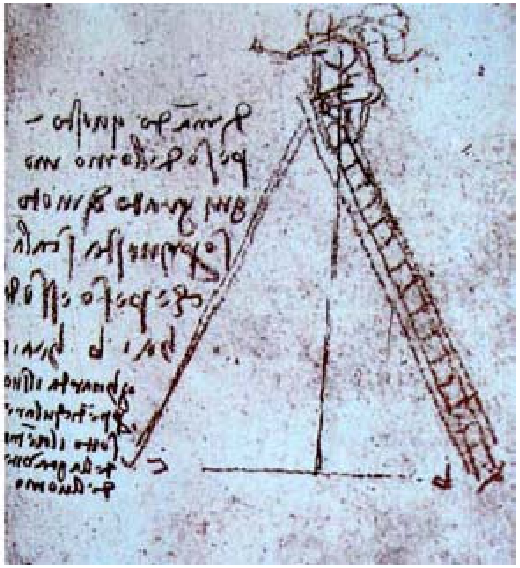
Motivation I: Artistic Representation

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”

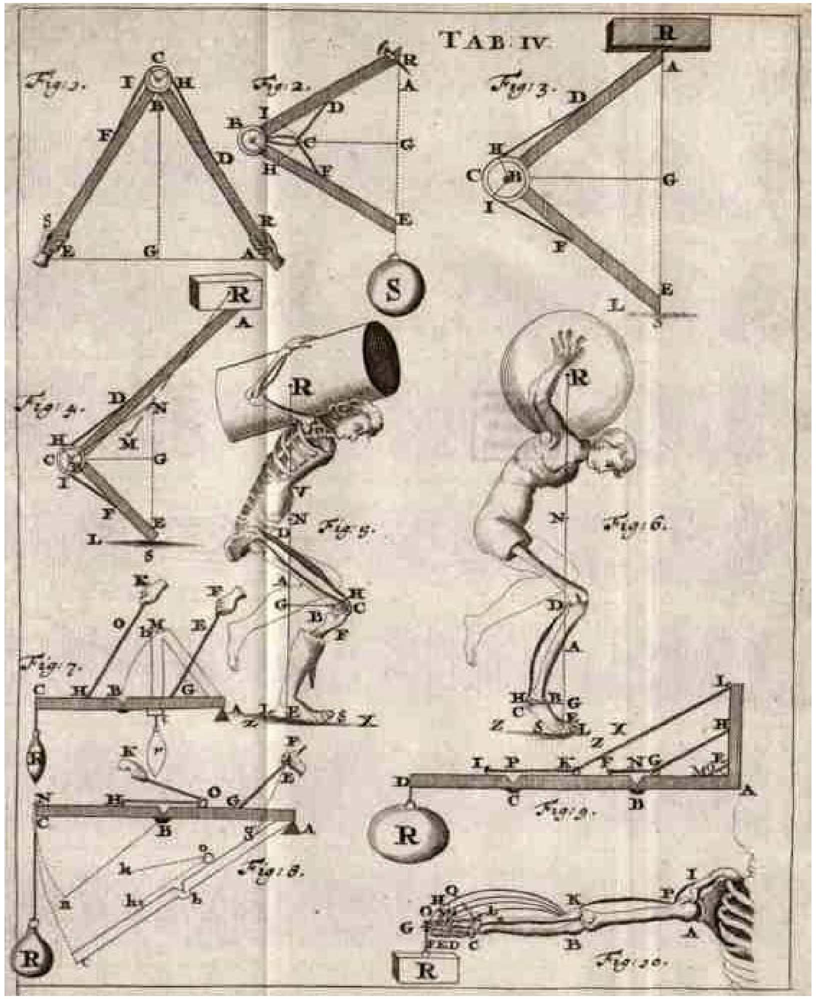
“I ask for the weight [pressure] of this man for every segment of motion when climbing those stairs, and for the weight he places on *b* and on *c*. Note the vertical line below the center of mass of this man.”





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

Motivation II: Biomechanics

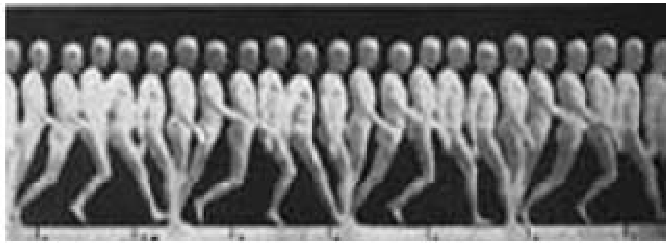


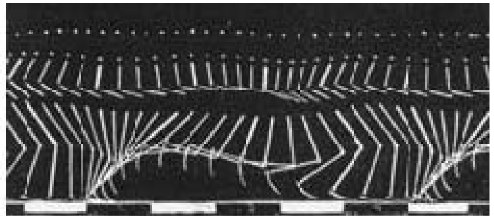
Giovanni Alfonso Borelli (1608–1679)

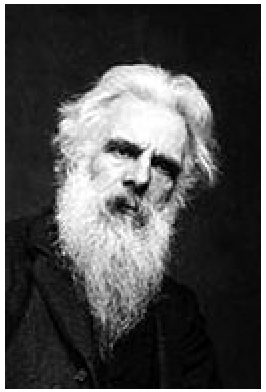
- The emergence of *biomechanics*
- Borelli applied to biology the analytical and geometrical methods, developed by Galileo Galilei
- He was the first to understand that bones serve as levers and muscles function according to mathematical principles
- His physiological studies included muscle analysis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

Motivation III: Motion perception

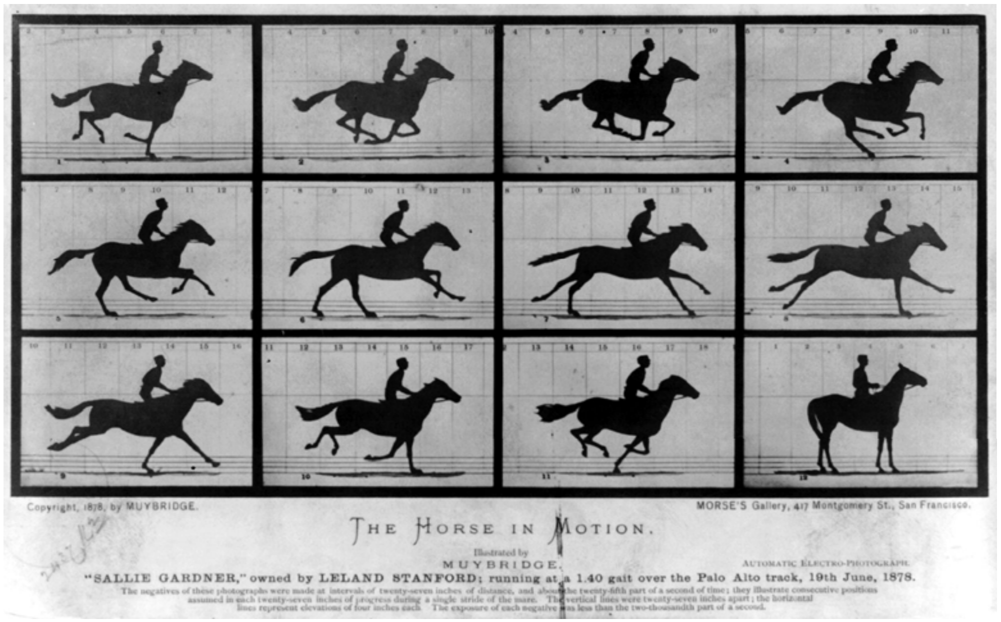
Etienne-Jules Marey: (1830–1904) made Chronophotographic experiments influential for the emerging field of *cinematography*





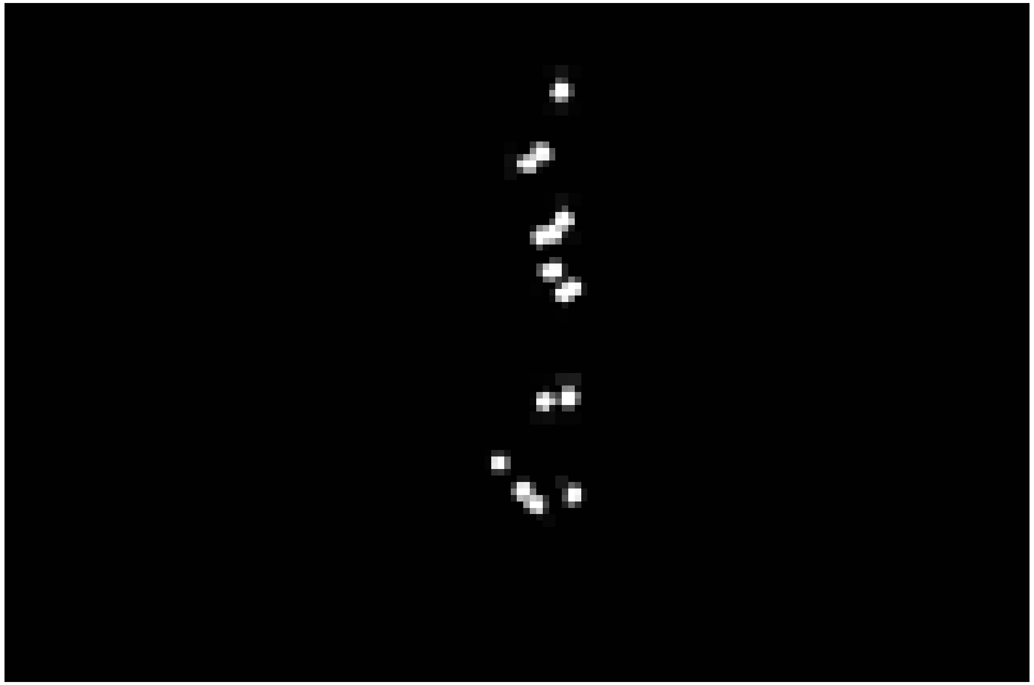


Eadweard Muybridge (1830–1904) invented a machine for displaying the recorded series of images. He pioneered motion pictures and applied his technique to movement studies



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.

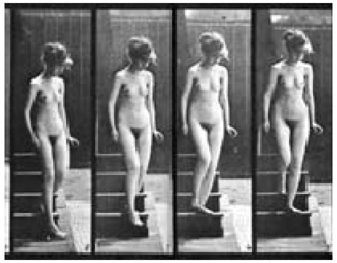


Gunnar Johansson, **Perception and Psychophysics,** 1973

Human actions: Historic overview

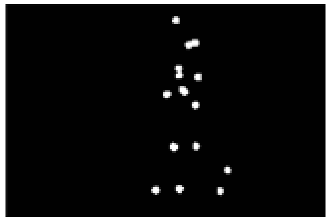
15th century
studies of
anatomy

17th century
emergence of
biomechanics



19th century
emergence of
cinematography

1973
studies of human
motion perception

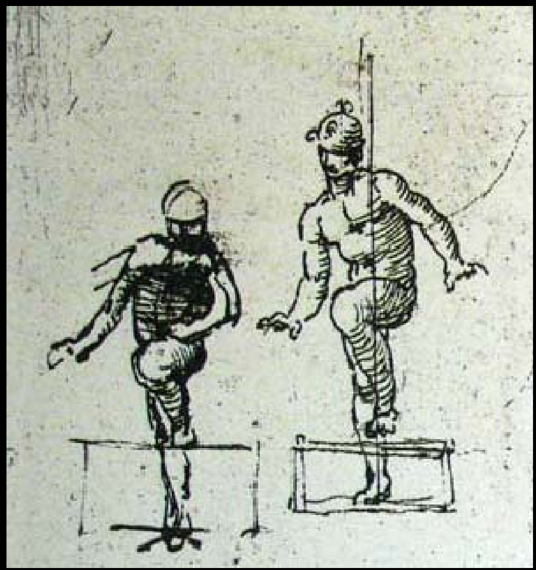


Modern computer vision

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

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

Applications: Unusual Activity Detection

e.g. for surveillance

Detecting Irregularities in Images and in Video
Boiman & Irani, **ICCV** 2005

Applications: Video Search

- 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 Search

- useful for TV production, entertainment, education, social studies, security,…

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

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

Sociology research: e.g.

Manually analyzed smoking actions in 900 movies

Surveillance: e.g. *“Woman throws cat into wheelie bin”* 260K views in 7 days

- … and it’s mainly about people and human actions

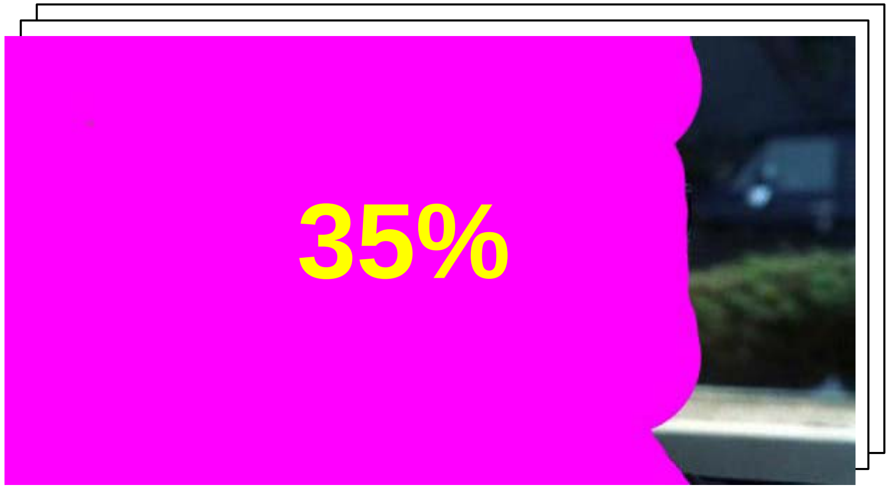
How many person-pixels are in video?


Movies

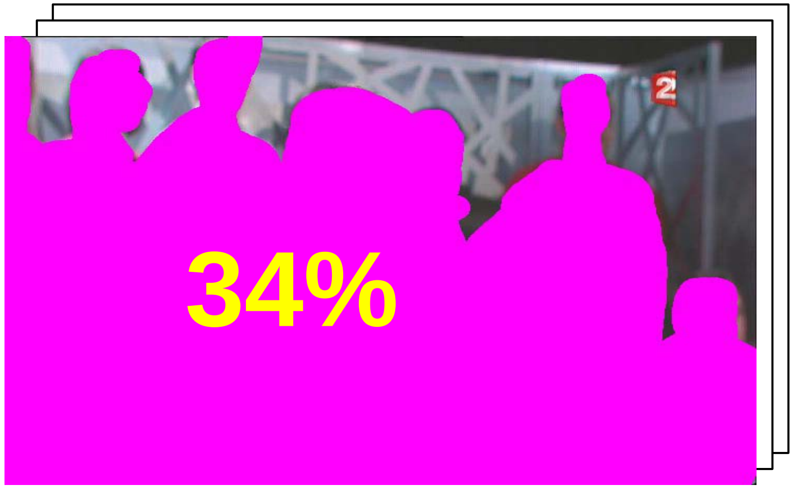
TV


YouTube

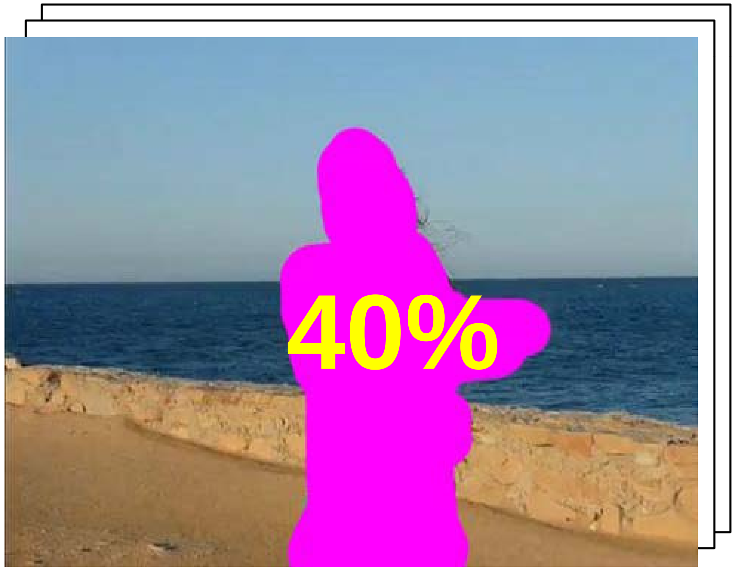
How many person-pixels are in video?



Movies



TV



YouTube

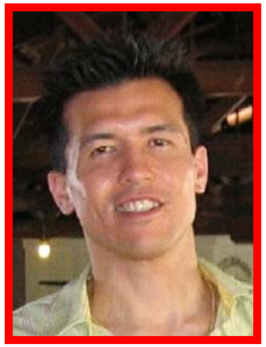
What this course is about?

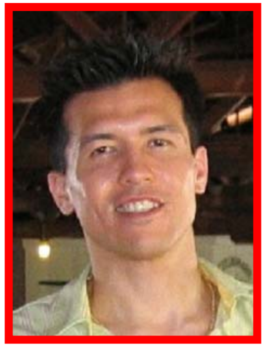
Goal

Get familiar with:

- **Problem formulations**
- **Mainstream approaches**
- **Particular existing techniques**
- **Current benchmarks**
- **Available baseline methods**
- **Promising future directions**

Course overview





- Definitions
- Benchmark datasets
- Early silhouette and tracking-based methods
- Motion-based similarity measures
- Template-based methods
- Local space-time features
- Bag-of-Features action recognition
- Weakly-supervised methods
- Pose estimation and action recognition
- Action recognition in still images
- Human interactions and dynamic scene models
- Conclusions and future directions

What is Action Recognition?

- Terminology
  - What is an “action”?
- Output representation
  - What do we want to say about an image/video?

Unfortunately, neither question has atisfactory answer yet

Terminology

- The terms “action recognition”, “activity recognition”, “event recognition”, are used inconsistently
  - Finding a common language for describing videos is an open problem

Terminology Example

- "Action" is a low-level primitive with semantic meaning
  - E.g. walking, pointing, placing an object
- "Activity" is a higher-level combination with some temporal relations
  - E.g. taking money out from ATM, waiting for a bus
- "Event" is a combination of activities, often involving multiple individuals
  - E.g. a soccer game, a traffic accident
- This is contentious
  - No standard, rigorous definition exists

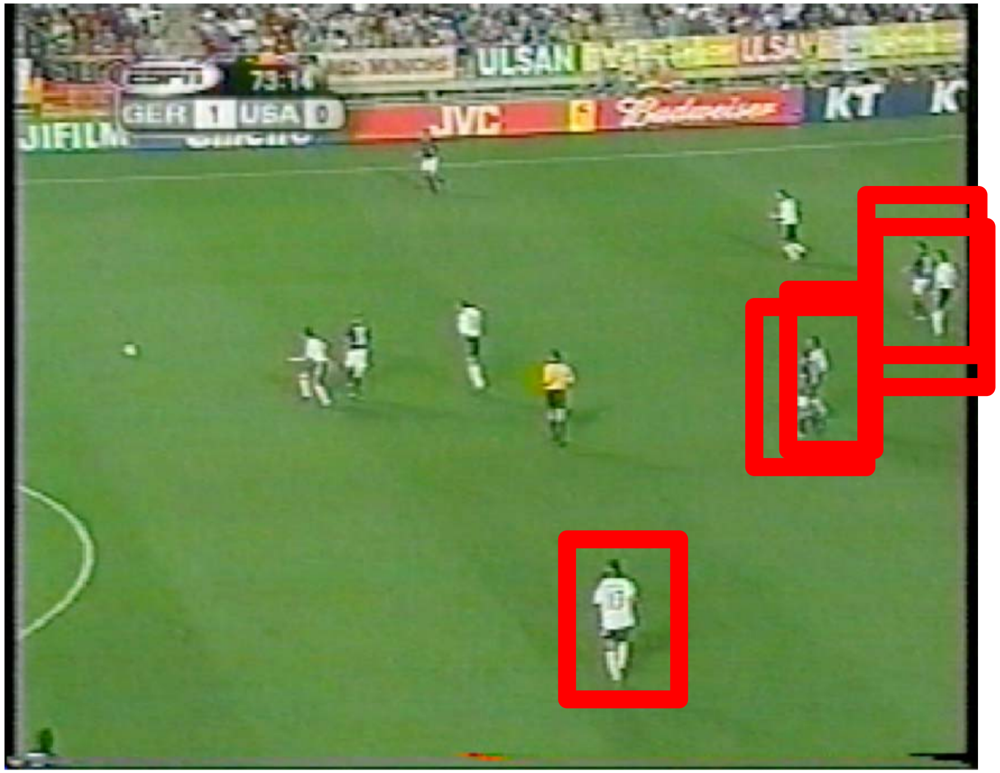
Output Representation

- Given this image what is the desired output?

- This image contains a man walking
  - Action classification / recognition
- The man walking is here
  - Action detection

Output Representation

- Given this image what is the desired output?



- This image contains 5 men walking, 4 jogging, 2 running
- The 5 men walking are here
- This is a soccer game

Output Representation

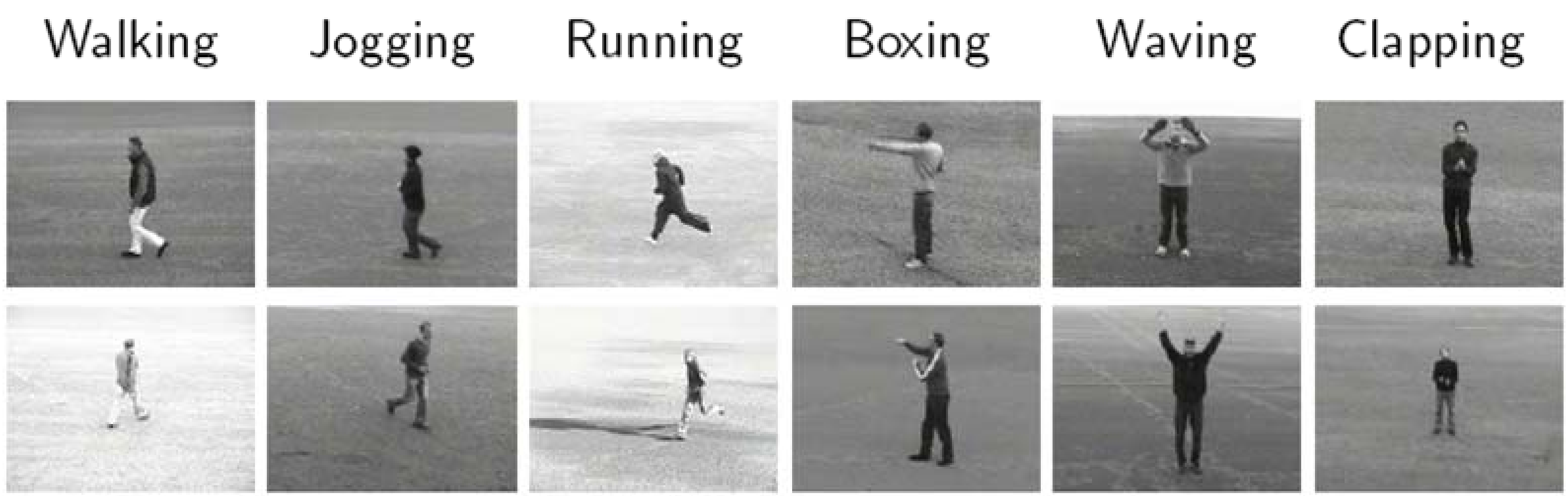
- Given this image what is the desired output?

  - Frames 1-20 the man ran to the left, then frames 21-25 he ran away from the camera
  - Is this an accurate description?
  - Are labels and video frames in 1-1 correspondence?

DATASETS

Dataset: KTH-Actions

- 6 action classes by 25 persons in 4 different scenarios
- Total of 2391 video samples
    - Specified train, validation, test sets
- Performance measure: average accuracy over all classes



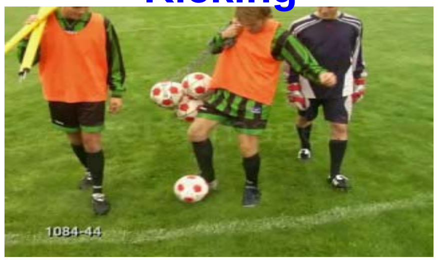
Schuldt, Laptev, Caputo ICPR 2004

UCF-Sports

- 10 different action classes
- 150 video samples in total
- Evaluation method: leave-one-out
- Performance measure: average accuracy over all classes

Diving

Kicking



Walking

Skateboarding

High-Bar-Swinging

Golf-Swinging

Rodriguez, Ahmed, and Shah CVPR 2008

UCF - YouTube Action Dataset

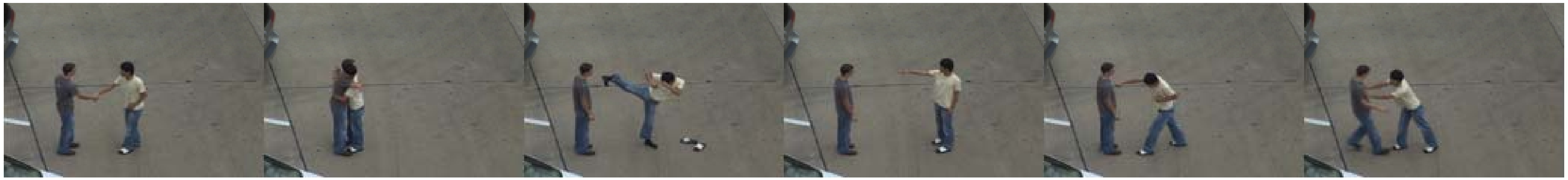
- 11 categories, 1168 videos
- Evaluation method: leave-one-out
- Performance measure: average accuracy over all classes

Liu, Luo and Shah CVPR 2009

Semantic Description of Human Activities (ICPR 2010)

- 3 challenges: interaction, aerial view, wide-area
- Interaction
  - 6 classes, 120 instances over ~20 min. video
  - Classification and detection tasks (+/- bounding boxes)
  - Evaluation method: leave-one-out

Hand Shaking | Hugging | Kicking | Pointing | Punching | Pushing



Ryoo et al. ICPR 2010 challenge

Hollywood2

- 12 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

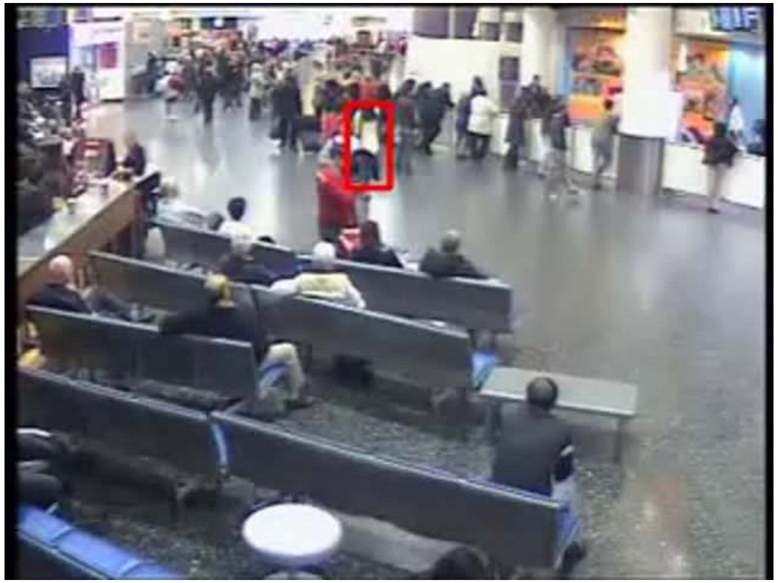
StandUp

DriveCar

Marszałek, Laptev, Schmid CVPR 2009

TRECVid Surveillance Event Detection

- 10 actions: person runs, take picture, cell to ear, …
- 5 cameras, ~100h video from LGW airport
- Detection (in time, not space); multiple detections count as false positives
- Evaluation method: specified training / test videos, evaluation at NIST
- Performance measure: statistics on DET curves



Smeaton, Over, Kraaij, TRECVid

Dataset Desiderata

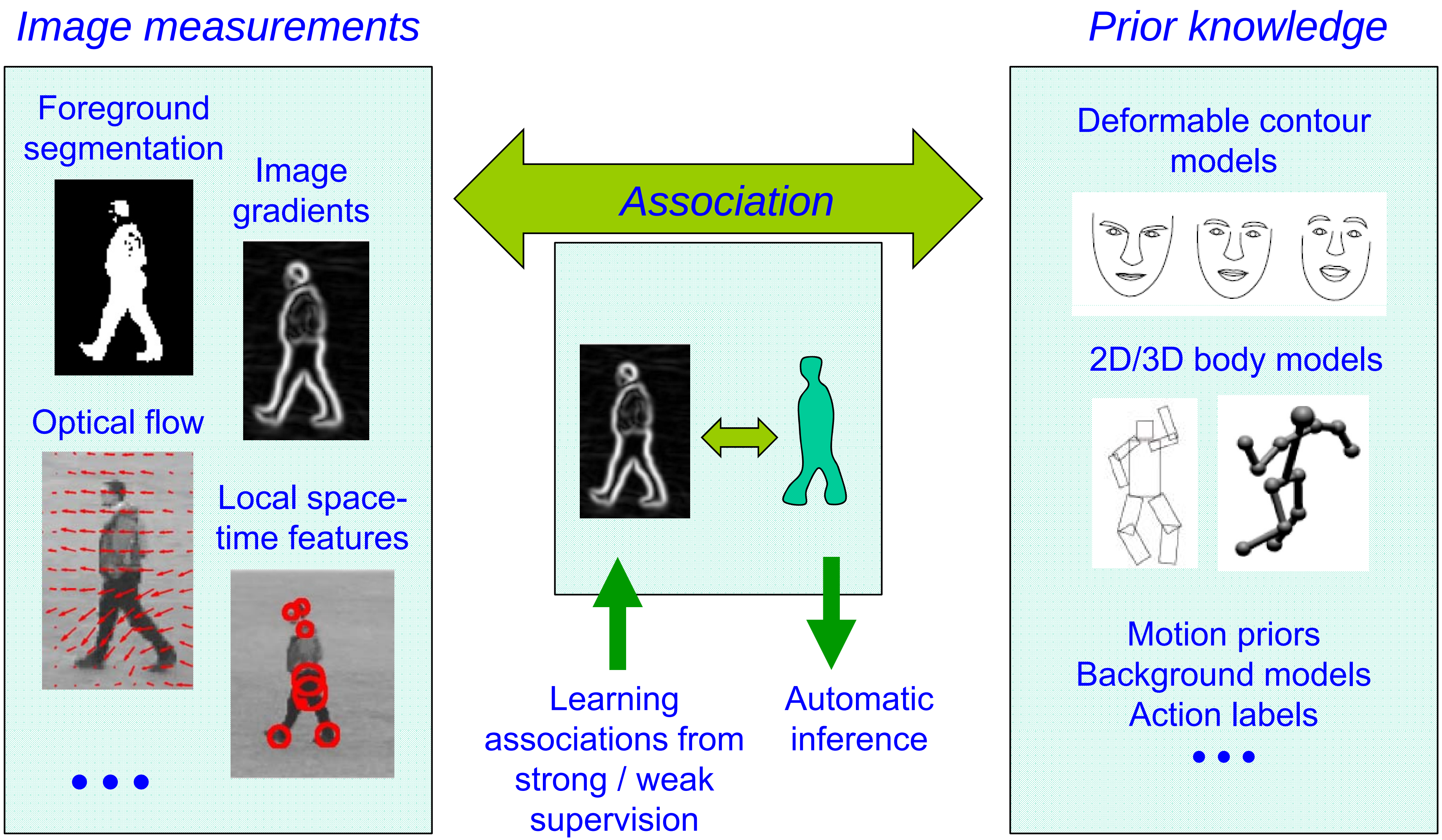
- Clutter
- Not choreographed by dataset collectors
  - Real-world variation
- Scale
  - Large amount of video
- Rarity of actions
  - Detection harder than classification
  - Chance performance should be **very** low
- Clear definition of training/test split
  - Validation set for parameter tuning?
  - Reproducing / comparing to other methods?

Datasets Summary

	Clutter?	Choreographed?	Scale	Rarity of actions	Training/testing split
KTH	No	Yes	2391 videos	Classification - one per video	Defined – by actors
UCF Sports	Yes	No	150 videos	Classification – one per video	Undefined - LOO
UCF Youtube	Yes	No	1168 videos	Classification – one per video	Undefined - LOO
SDHA-ICPR Interaction	No	Yes	20 minutes, 120 instances	Classification / detection	Undefined - LOO
Hollywood2	Yes	No	69 movies, ~1600 instances	Detection, ~xx actions/h	Defined – by videos
TRECVid	Yes	No	~100h	Detection, ~20 actions/h	Defined – by time

How to recognize actions?

Action understanding: Key components



Foreground segmentation

Image differencing: a simple way to measure motion/change

-

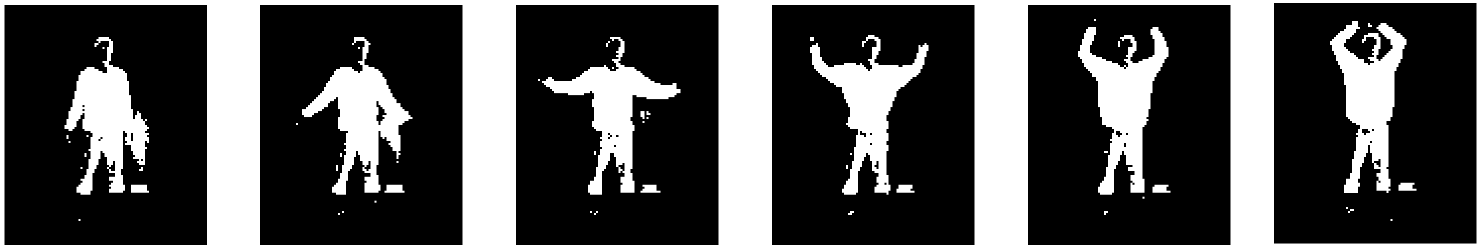
> Const

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):

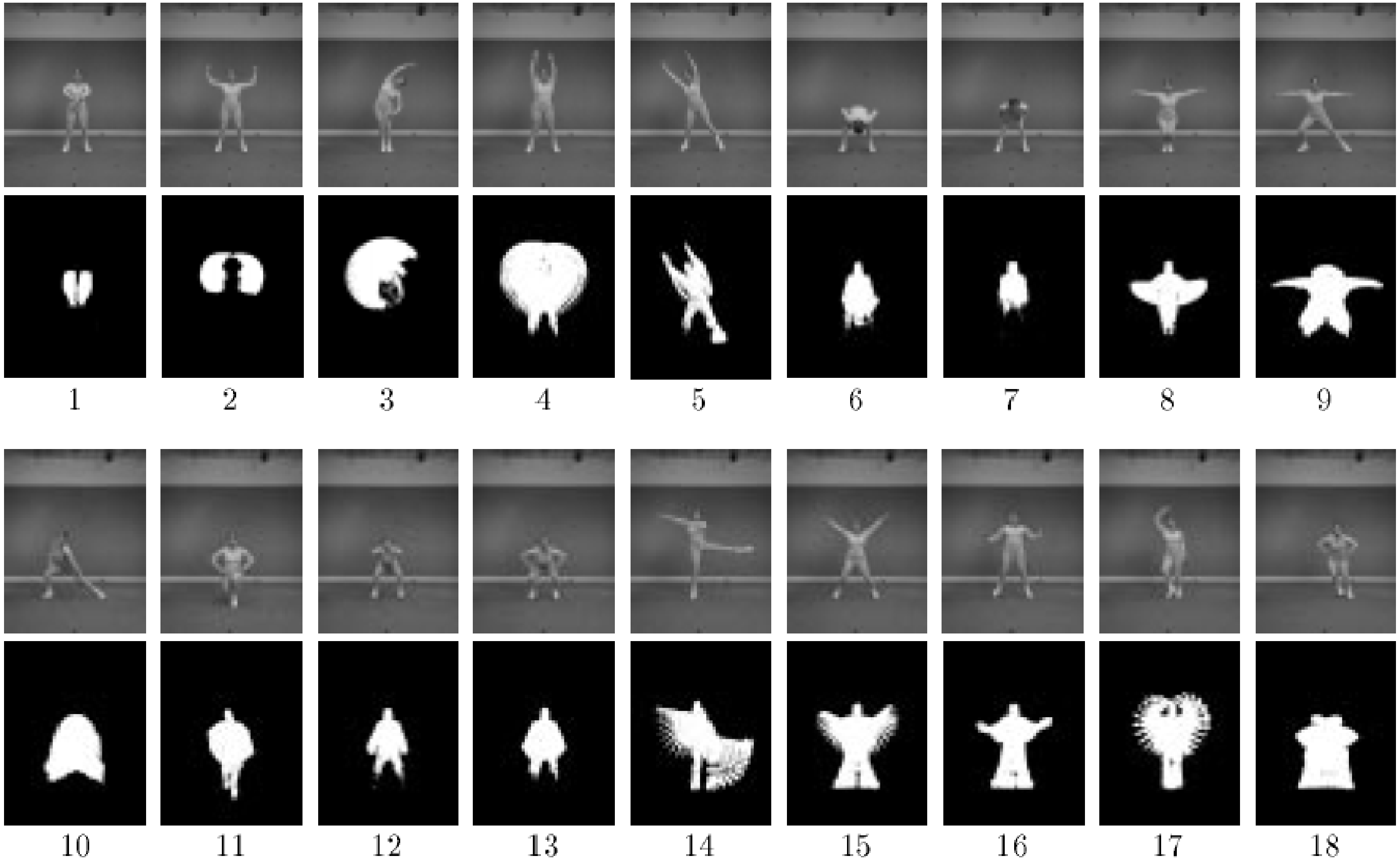
$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ \max\ (0, H_\tau(x, y, t-1) - 1) & \\ \quad \text{otherwise} & \end{cases}$$

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

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

Temporal Templates: Summary

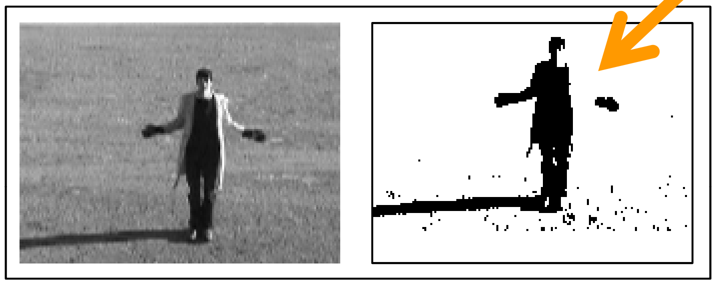
Pros:

- \+ 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

Point Distribution Model

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

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

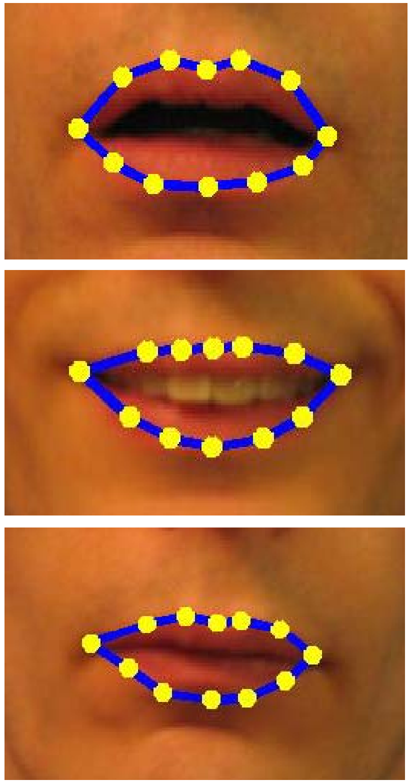
- Assume each shape can be represented by the linear combination of basis shapes

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

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

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

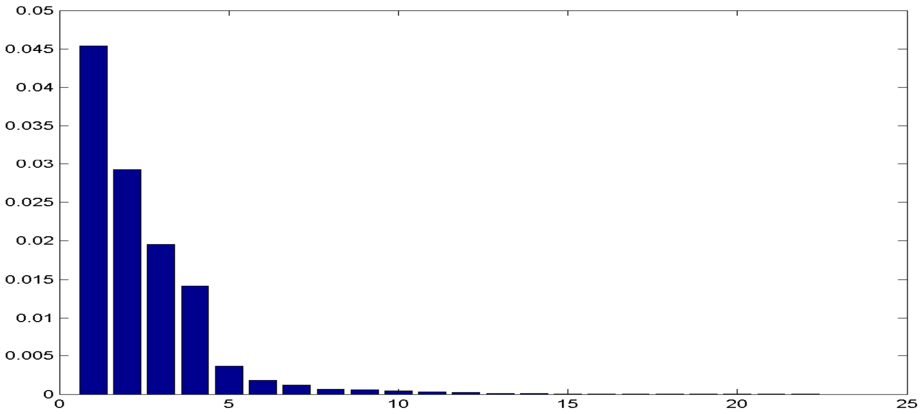
and some parameter vector $\mathbf{b}$



[Cootes et al. 1995]

Active Shape Models

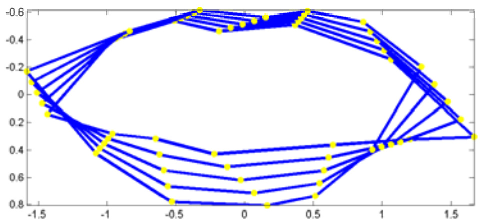
- Distribution of eigenvalues of $\mathbf{S}$: $\lambda_1, \lambda_2, \lambda_3, \ldots$



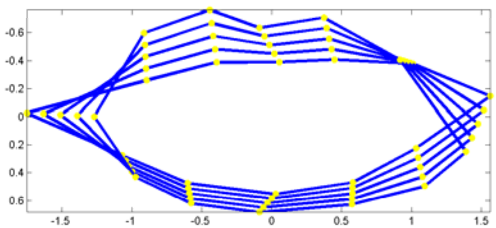
A small fraction of basis shapes (eigenvecors) accounts for the most of shape variation (=> landmarks are redundant)

- Three main modes of lips-shape variation:

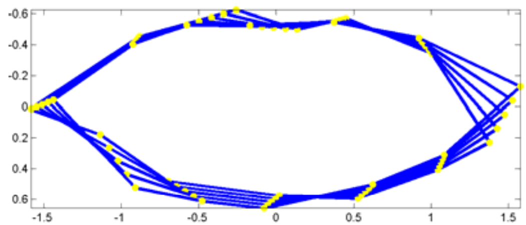
$\mathbf{b} = (\mu\lambda_1, 0, 0, \ldots)^\top$



$\mathbf{b} = (0, \mu\lambda_2, 0, 0, \ldots)^\top$



$\mathbf{b} = (0, 0, \mu\lambda_3, 0, 0, \ldots)^\top$

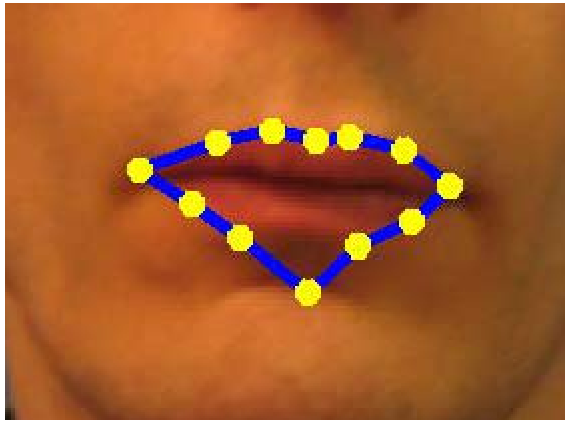


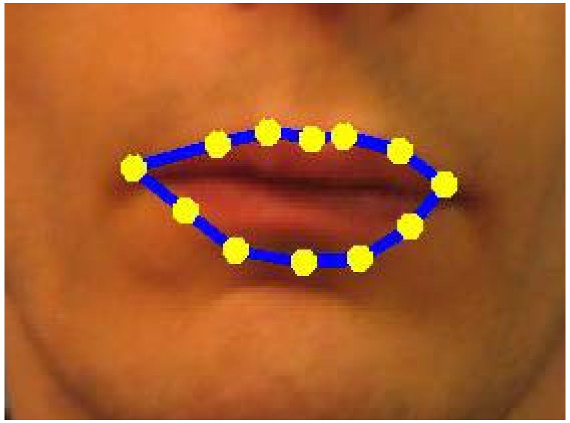
$\mu = -3, 1.5, 0, 1.5, 3$

Active Shape Models: effect of regularization

- Projection onto the shape-space serves as a regularization

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





Person Tracking

Learning flexible models from image sequences

[A. Baumberg and D. Hogg, ECCV 1994]

Active Shape Models: Summary

Pros:

+ Shape prior helps overcoming segmentation errors
+ Fast optimization
+ Can handle interior/exterior dynamics

Cons:

- Optimization gets trapped in local minima
- Re-initialization is problematic

Possible improvements:

- Learn and use motion priors, possibly specific to different actions

Motion priors

- Accurate motion models can be used both to:
  - ❖ Help accurate tracking
  - ❖ Recognize actions

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

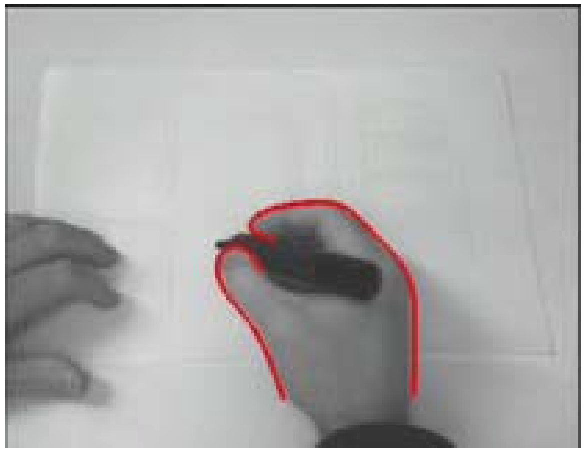
Example:

Drawing with 3 action modes

line drawing

scribbling

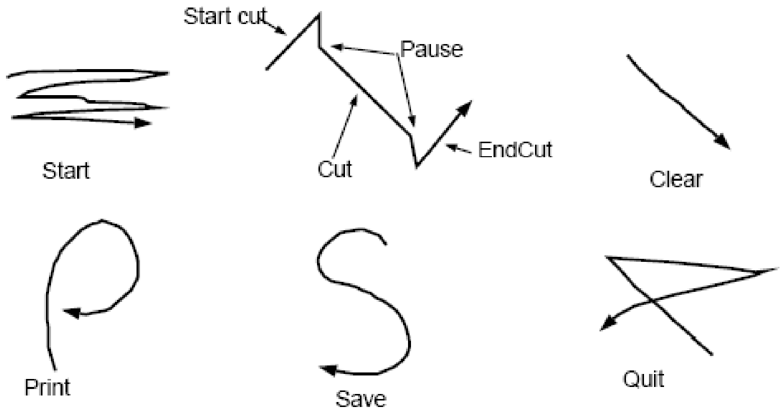
idle

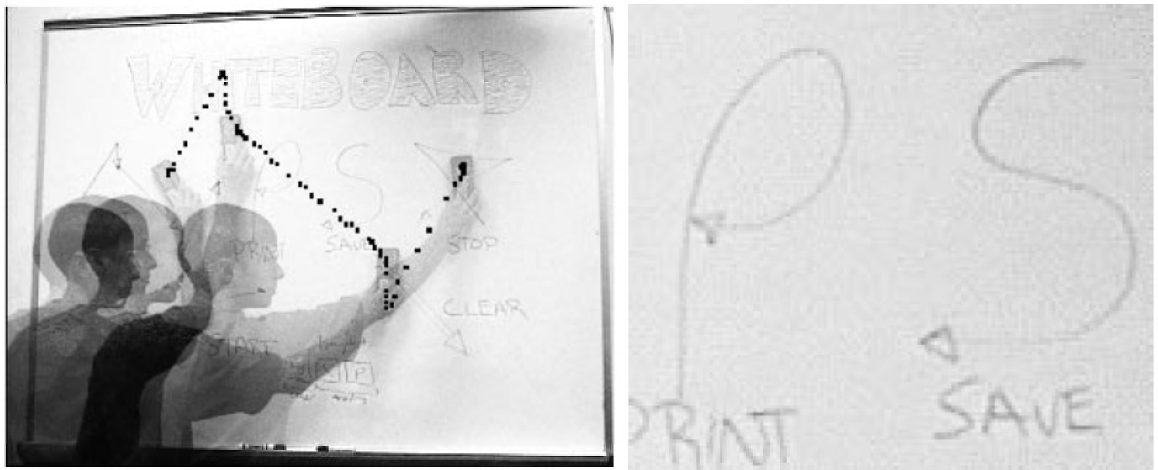


[M. Isard and A. Blake, ICCV 1998]

Dynamics with discrete states

Joint tracking and gesture recognition in the context of a visual white-board interface





[M.J. Black and A.D. Jepson, ECCV 1998]

Motion priors & Trackimg: Summary

Pros:

+ more accurate tracking using specific motion models

+ Simultaneous tracking and motion recognition with discrete state dynamical models

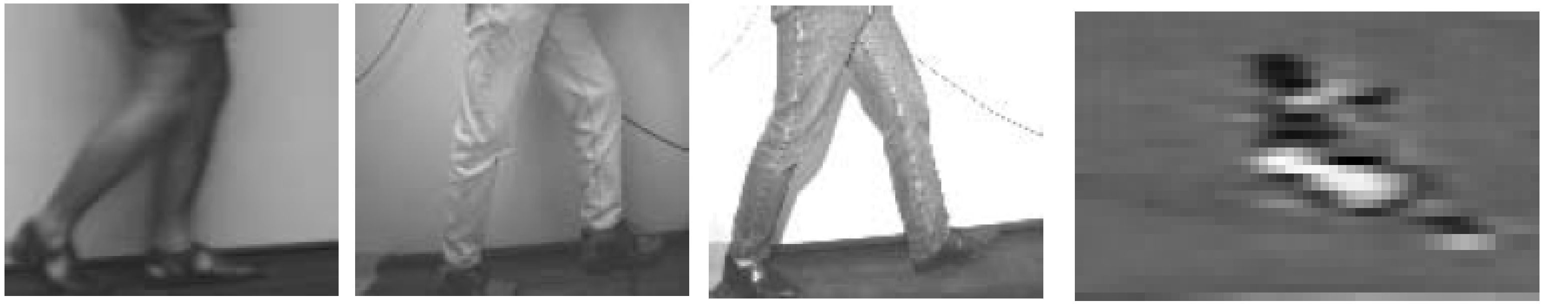
Cons:

- Local minima is still an issue

- Re-initialization is still an issue

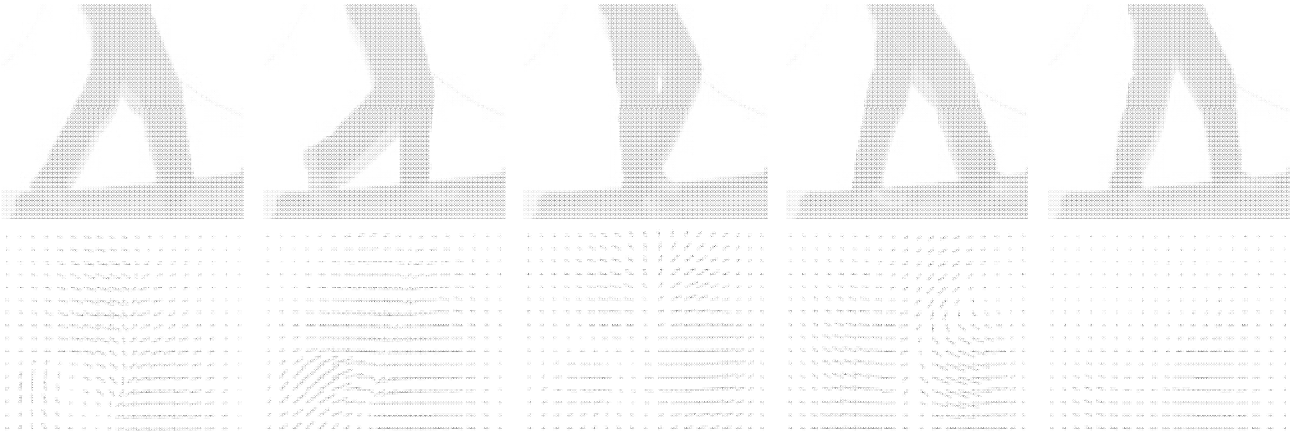
Shape and Appearance vs. Motion

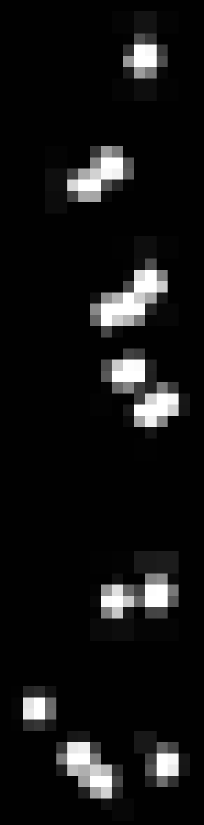
- Shape and appearance in images depends on many factors: clothing, illumination contrast, image resolution, etc…



[Efros et al. 2003]

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





Gunnar Johansson, Moving Light Displays, 1973

Motion estimation: Optical Flow

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

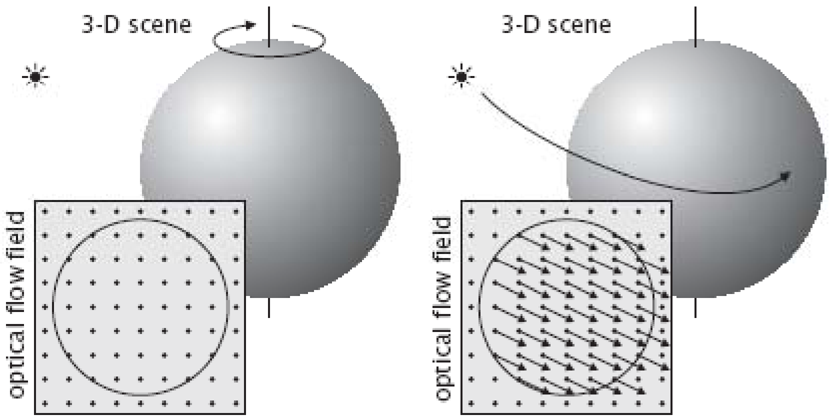
  How? We only have access to image pixels
  ➡ Estimate pixel-wise correspondence between frames = Optical Flow

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

  ❖ Useful assumption in many cases

  ❖ Breaks at occlusions and illumination changes

  ❖ Physical and visual motion may be different



Parameterized Optical Flow

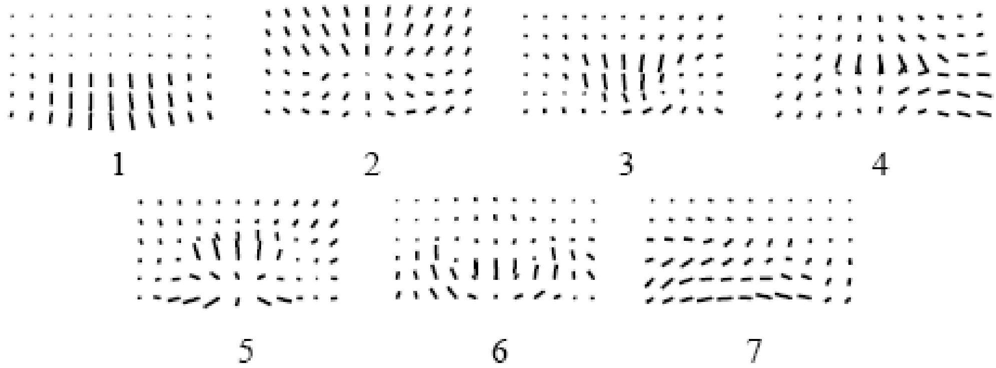
1. Compute standard Optical Flow for many examples
2. Put velocity components into one vector

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

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

Training samples

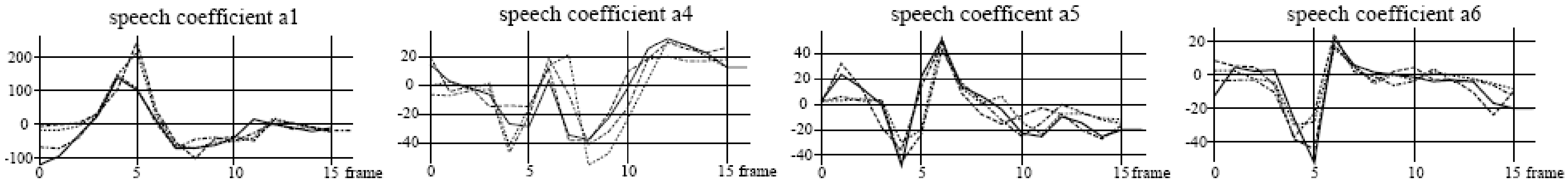
PCA flow bases

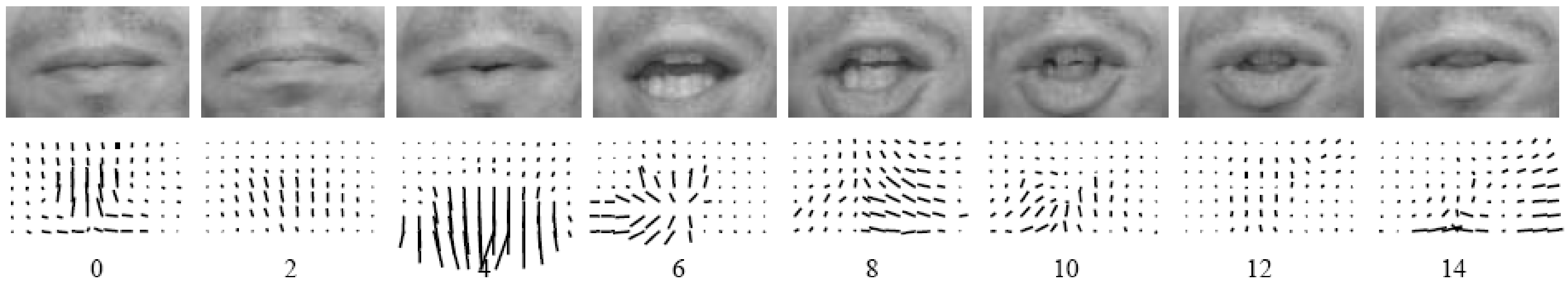


[Black, Yacoob, Jepson, Fleet, CVPR 1997]

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

[Black, Yacoob, Jepson, Fleet, CVPR 1997]

Spatial Motion Descriptor

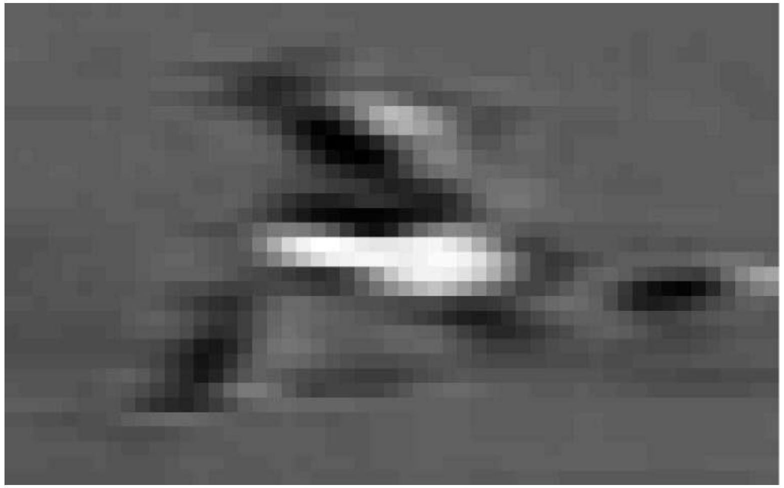
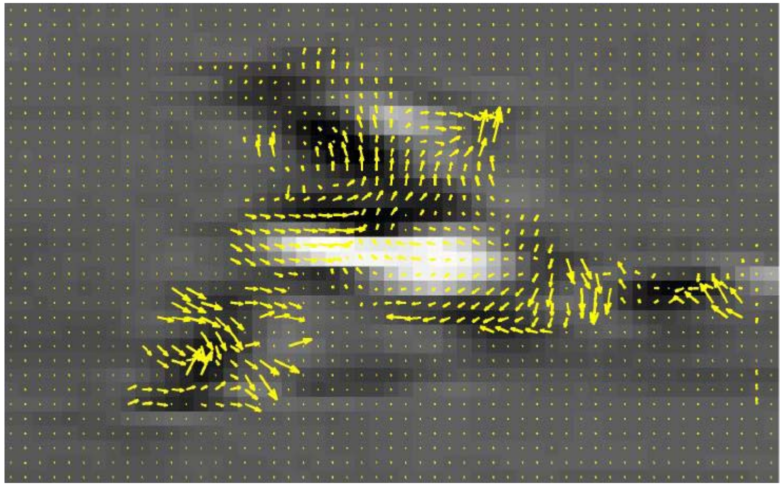
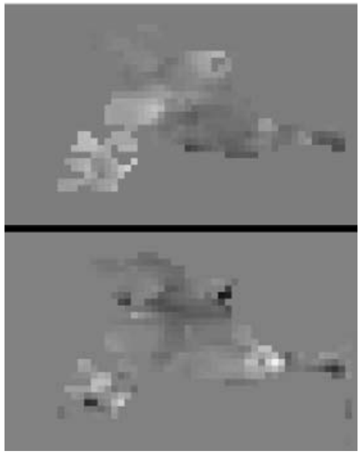


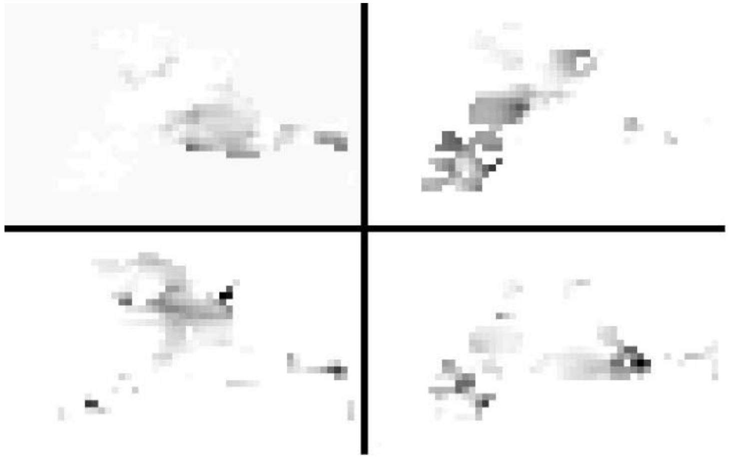
Image frame



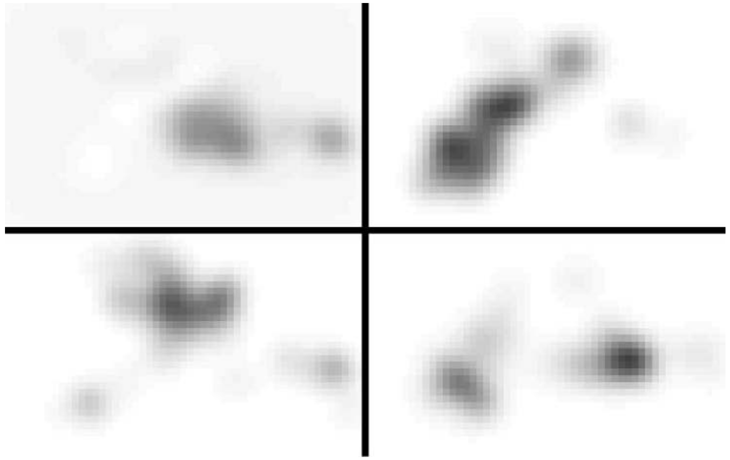
Optical flow $F_{x,y}$



F_x, F_y



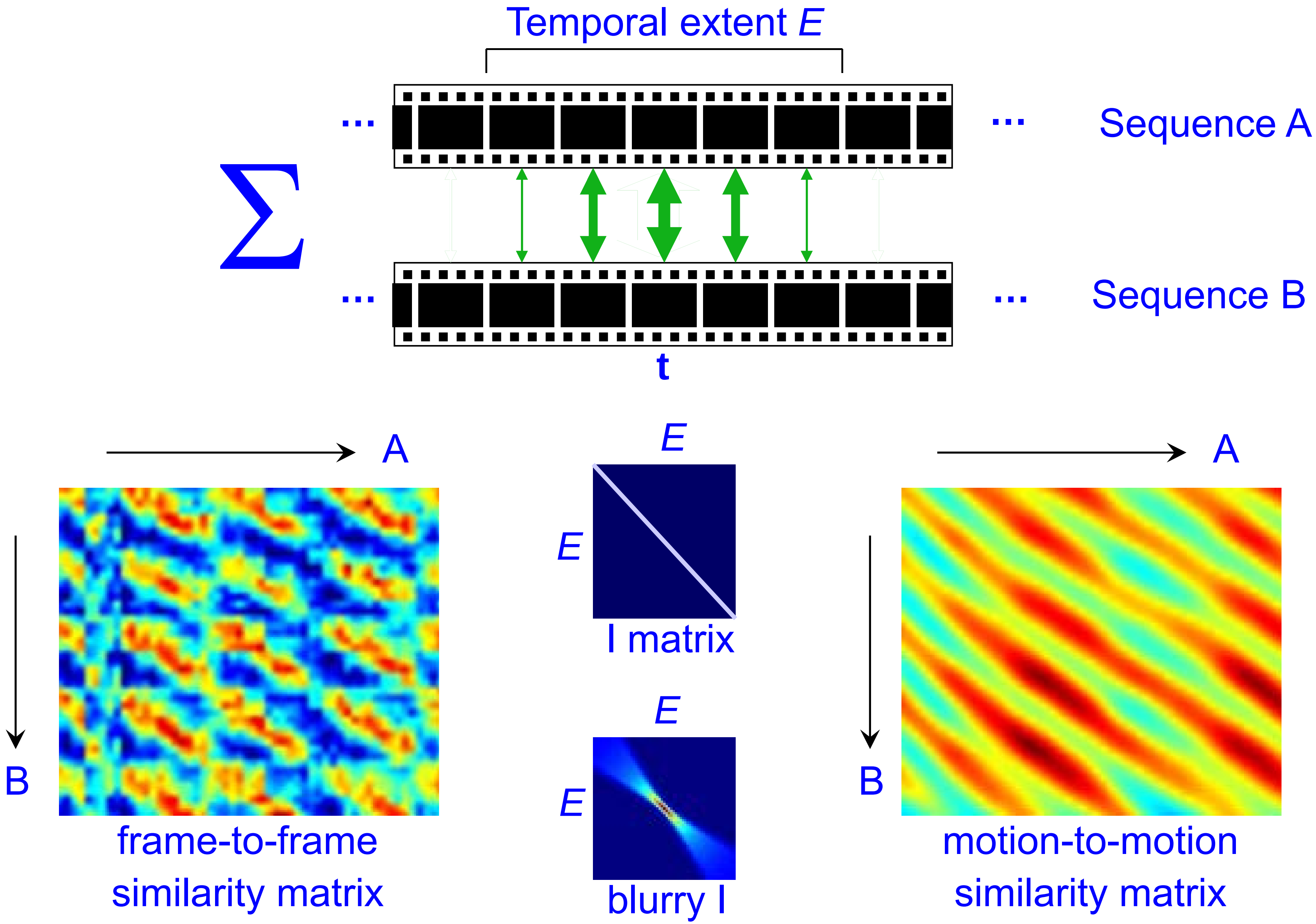
$F_x^-, F_x^+, F_y^-, F_y^+$



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

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

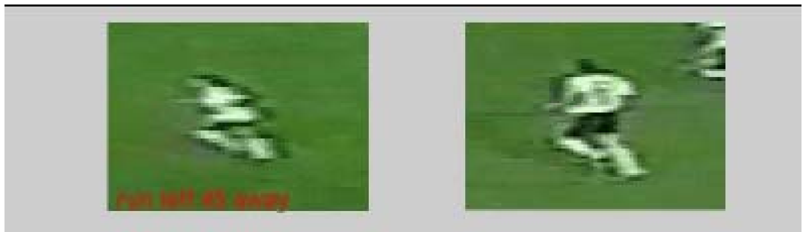
Spatio-Temporal Motion Descriptor



Football Actions: matching

Input Sequence

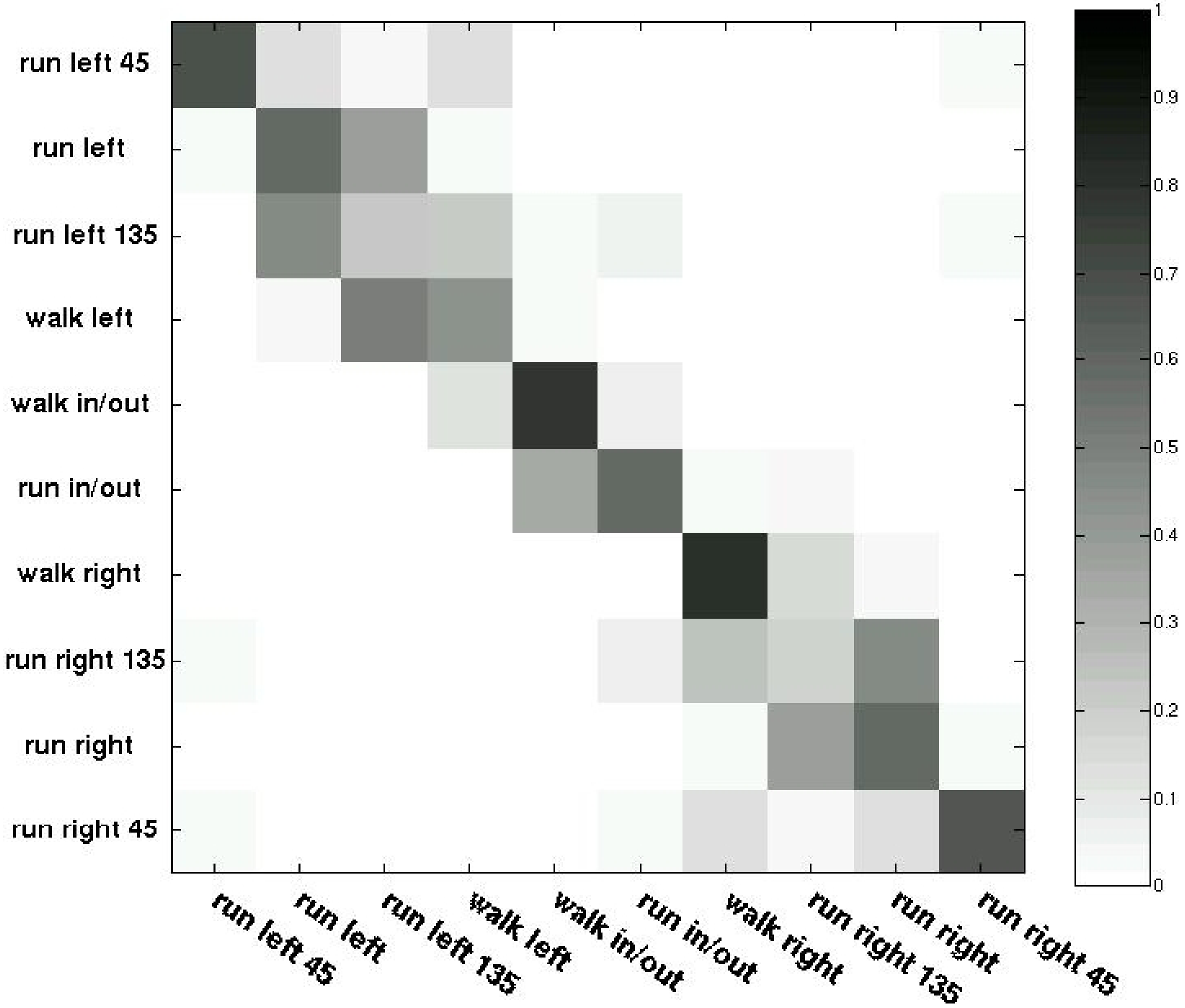
Matched Frames



input matched

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

Football Actions: classification



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

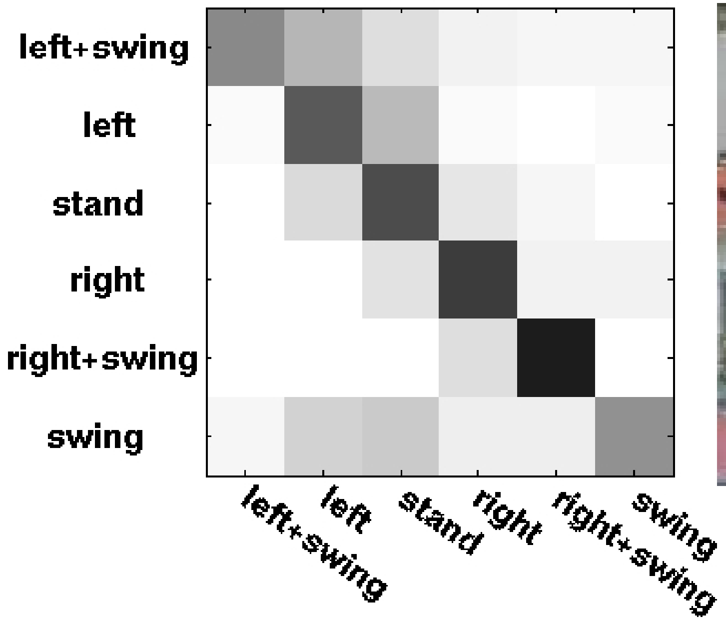
[Efros, Berg, Mori, Malik, ICCV 2003]

Football Actions: Replacement

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

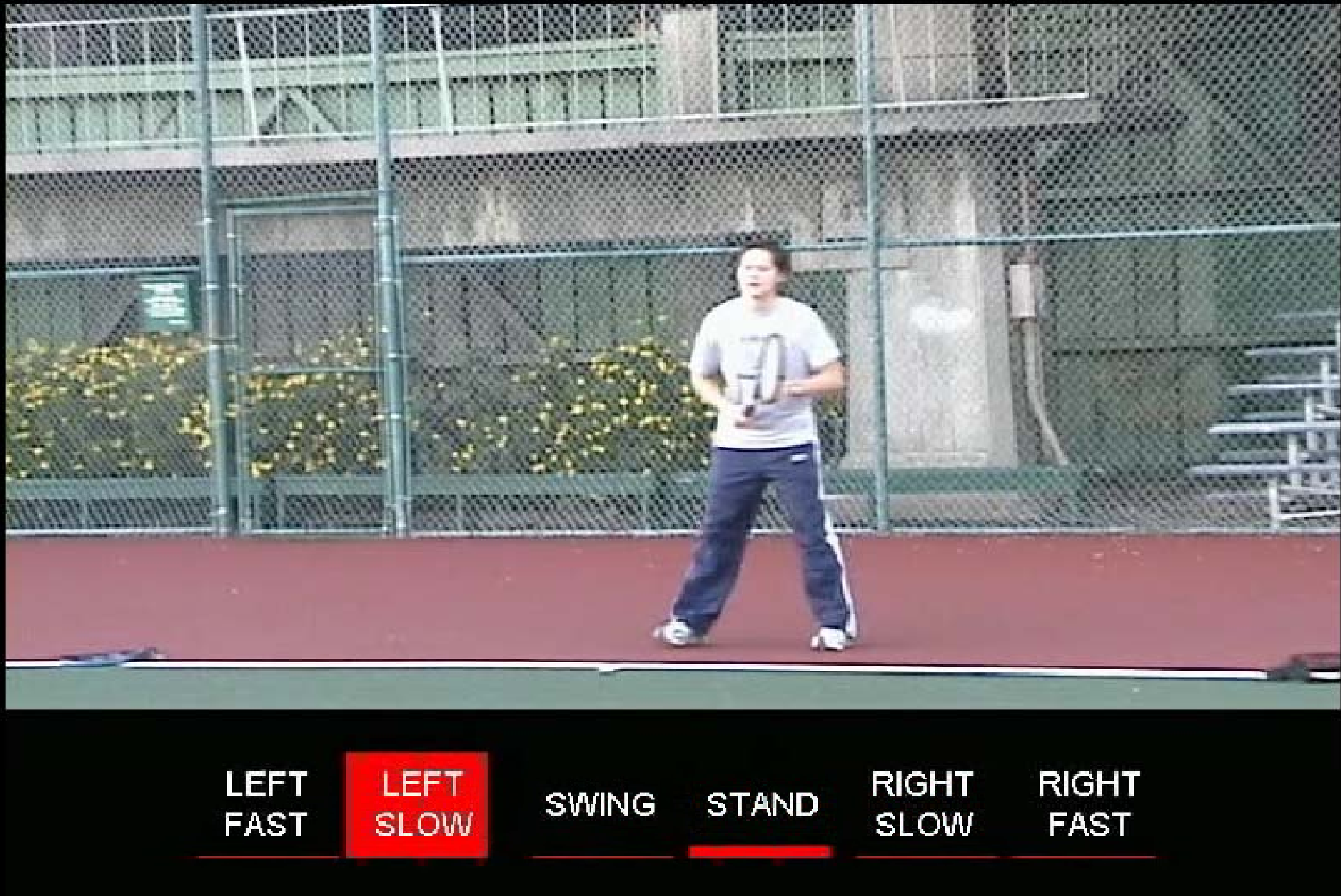
Classifying Tennis Actions

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



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

Classifying Tennis Actions

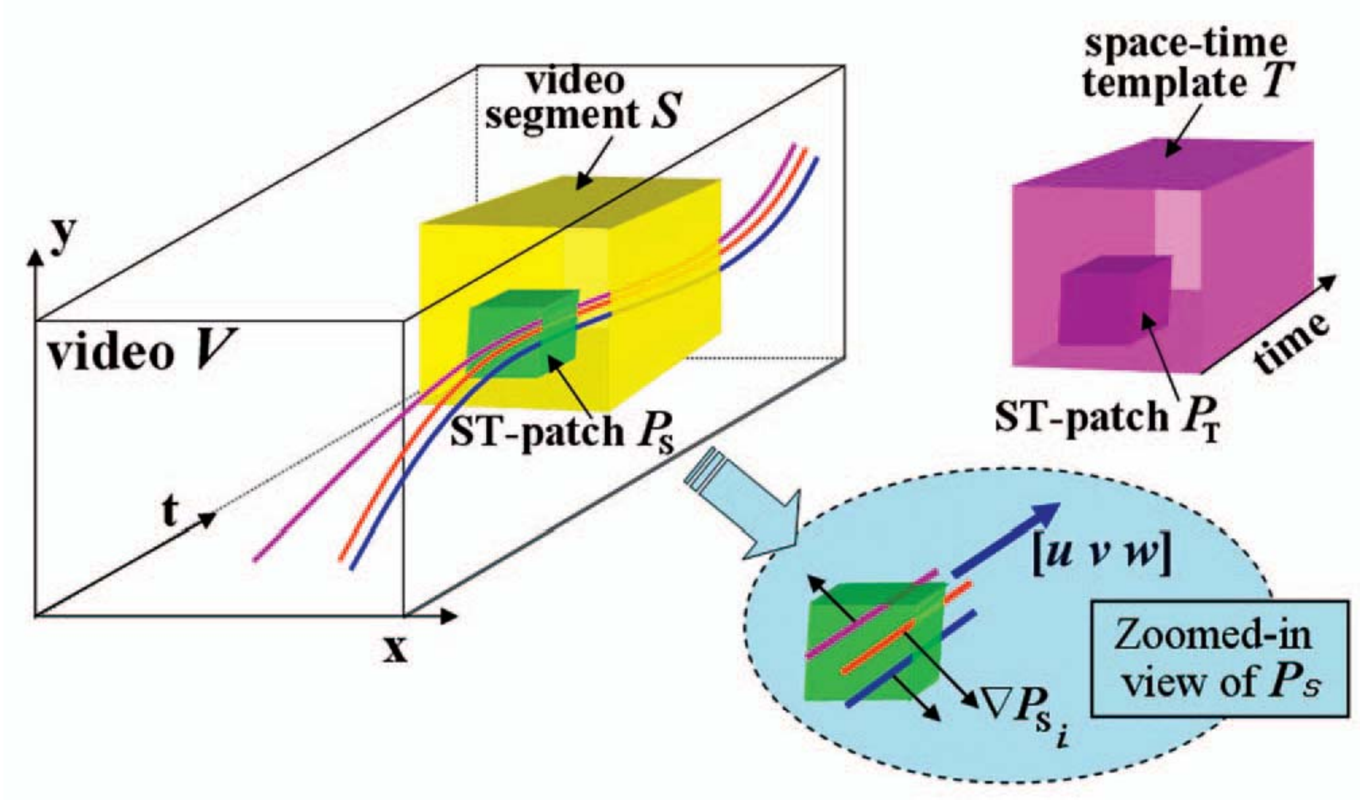


Red bars illustrate classification confidence for each action

[A. A. Efros, A. C. Berg, G. Mori, J. Malik, ICCV 2003]

Motion recognition
<u>without</u> motion estimations

- Motion estimation from video is a often noisy/unreliable
- Measure motion consistency between a template and test video



[Schechtman and Irani, PAMI 2007]

Motion recognition without motion estimations

- Motion estimation from video is a often noisy/unreliable
- Measure motion consistency between a template and test video

Template video

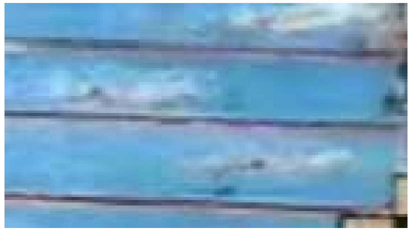
Test video

Correlation result

[Schechtman and Irani, PAMI 2007]

Motion recognition <u>without</u> motion estimations

- Motion estimation from video is a often noisy/unreliable
- Measure motion consistency between a template and test video



Template video

Test video

Correlation result

[Schechtman and Irani, PAMI 2007]

Motion-based template matching

Pros:

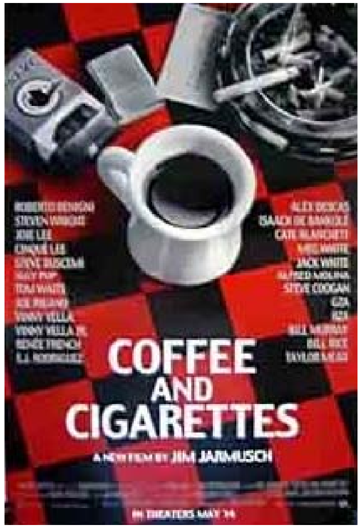
+ Depends less on variations in appearance

Cons:

- Can be slow
- Does not model negatives

Improvements possible using *discriminatively-trained* template-based action classifiers

Action Dataset and Annotation

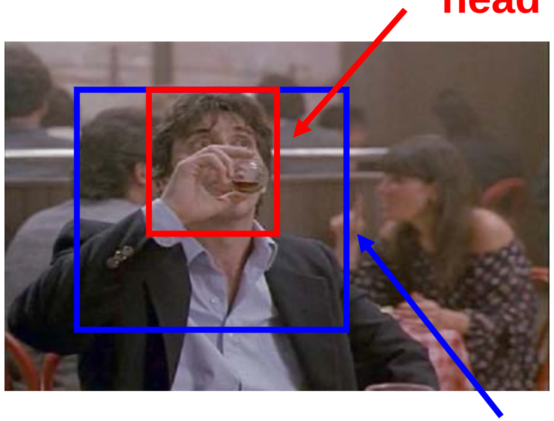


Manual annotation of drinking actions in movies:
“Coffee and Cigarettes”; “Sea of Love”

“*Drinking*”: 159 annotated samples
“*Smoking*”: 149 annotated samples

Spatial annotation

head rectangle



torso rectangle

Temporal annotation

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

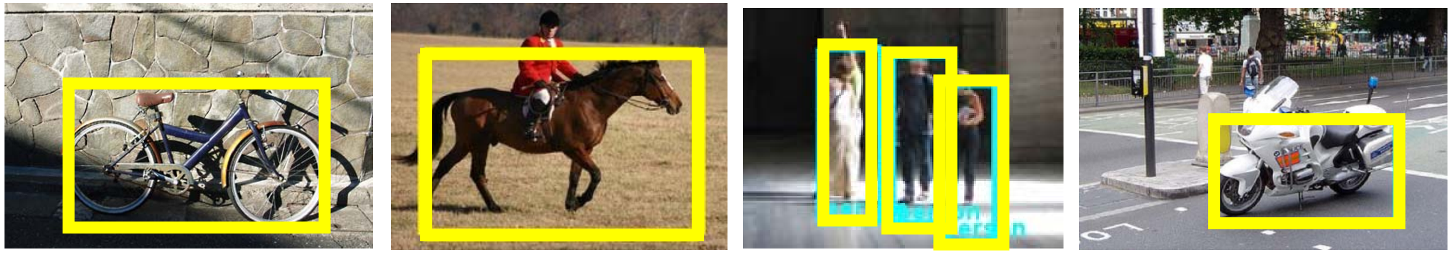
“Drinking” action samples

training samples

test samples

Actions == space-time objects?

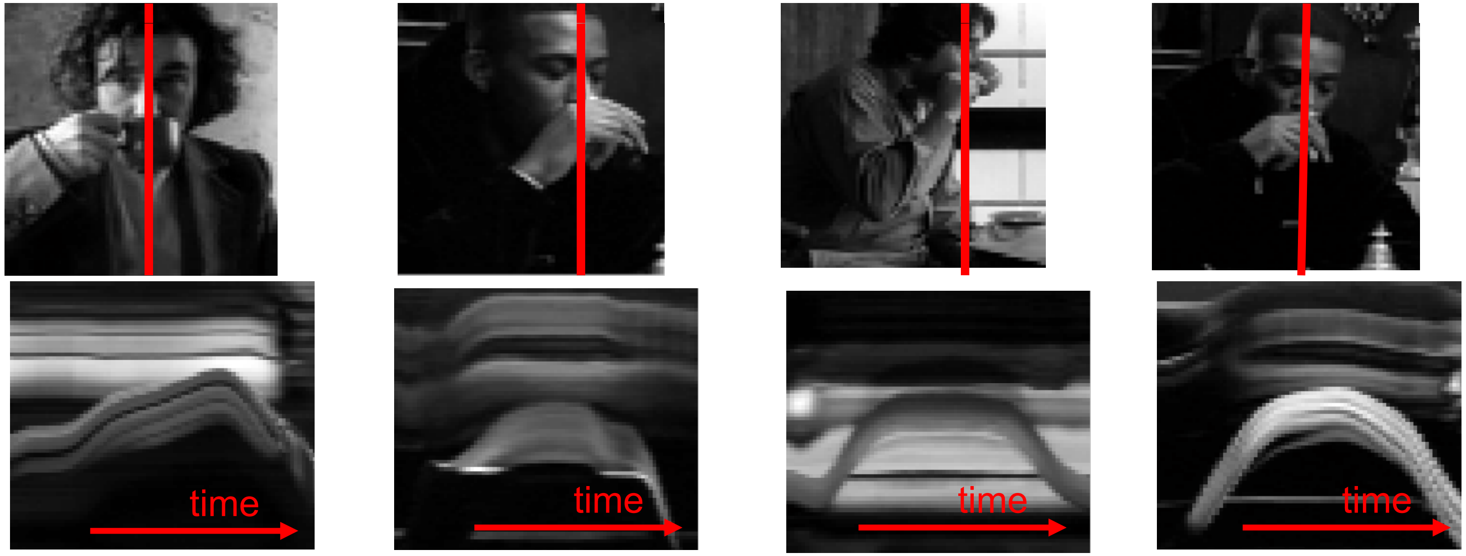
"stable-view" objects



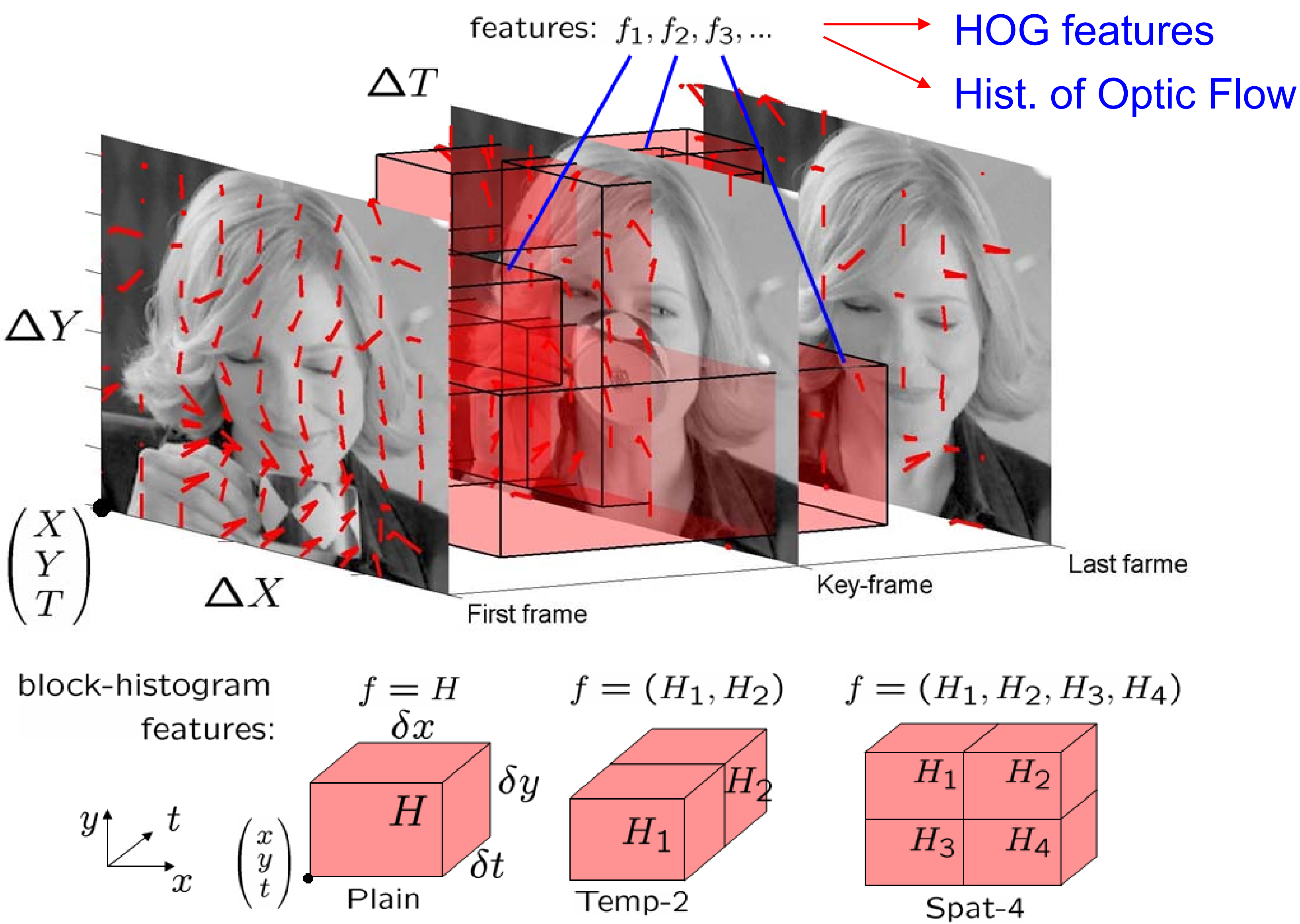
"atomic" actions

car exit phoning smoking hand shaking drinking

Objective: take advantage of space-time shape



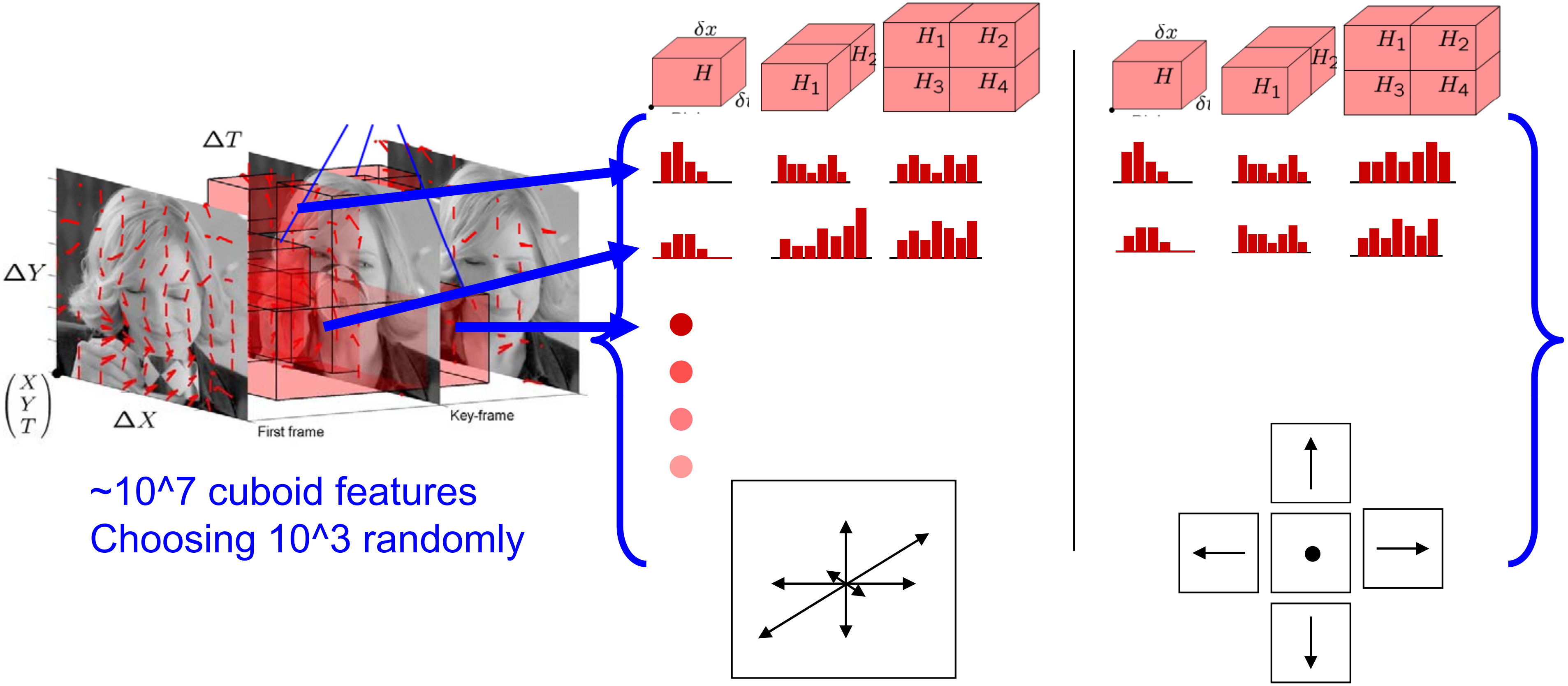
Actions == Space-Time Objects?



Histogram features

HOG: histograms of **oriented gradient**

HOF: histograms of **optic flow**

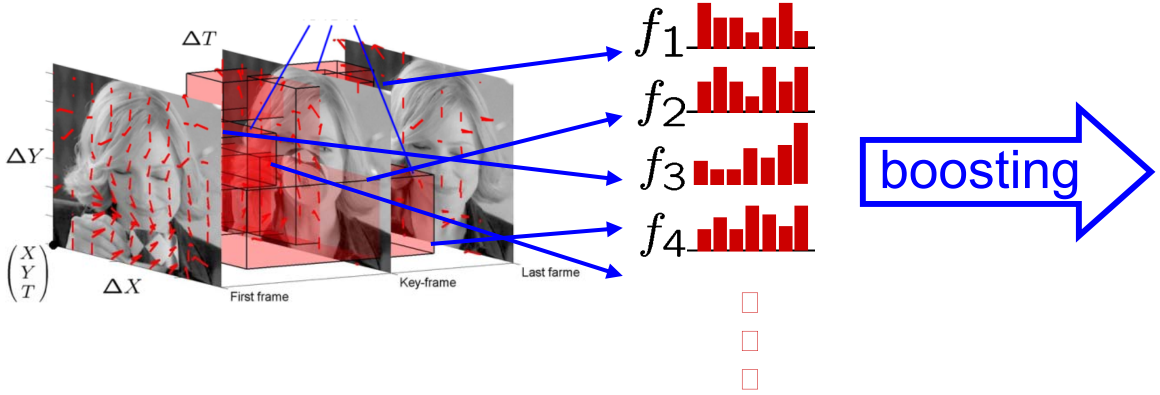


~10^7 cuboid features
Choosing 10^3 randomly

4 grad. orientation bins

4 OF direction bins
+ 1 bin for no motion

Action learning



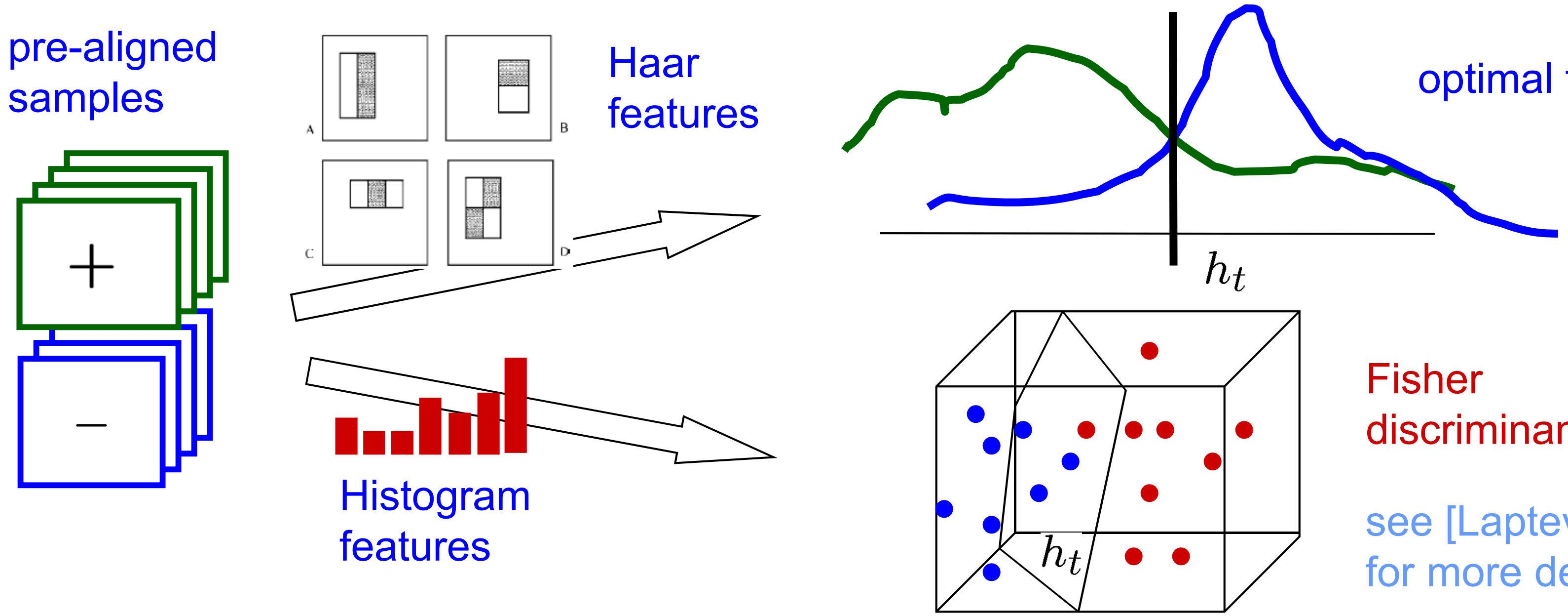
selected features

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

weak classifier

AdaBoost:

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



see [Laptev BMVC'06] for more details

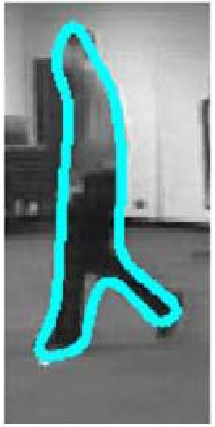
Drinking action detection

Test episodes from the movie “Coffee and cigarettes”

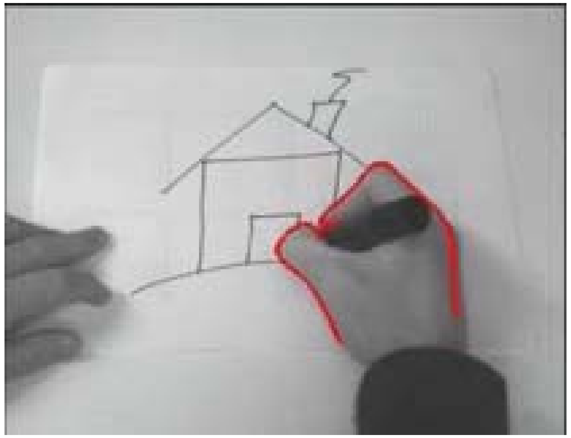
[I. Laptev and P. Pérez, ICCV 2007]

Where are we so far ?

Temporal templates:
+ simple, fast
- sensitive to segmentation errors

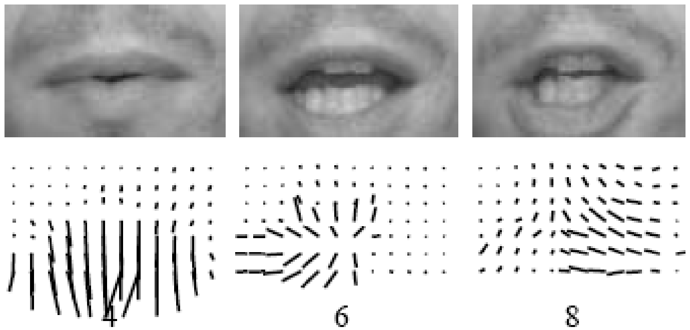


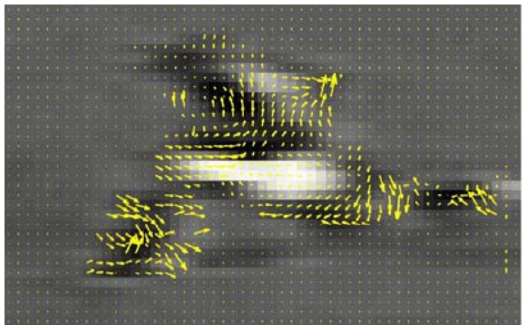
Active shape models:
+ shape regularization
- sensitive to initialization and tracking failures



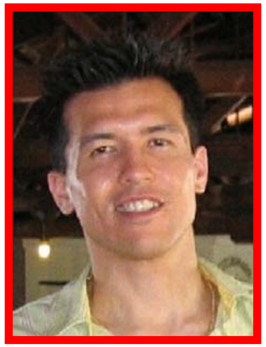
Tracking with motion priors:
+ improved tracking and simultaneous action recognition
- sensitive to initialization and tracking failures

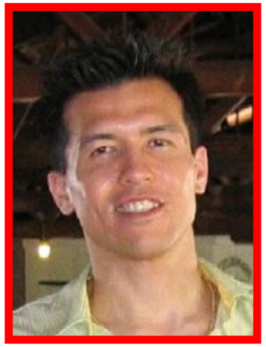
Motion-based recognition:
+ generic descriptors; less depends on appearance
- sensitive to localization/tracking errors





Course overview





- Definitions
- Benchmark datasets
- Early silhouette and tracking-based methods
- Motion-based similarity measures
- Template-based methods
- Local space-time features
- Bag-of-Features action recognition
- Weakly-supervised methods
- Pose estimation and action recognition
- Action recognition in still images
- Human interactions and dynamic scene models
- Conclusions and future directions

How to handle real complexity?

Common methods:

- Camera stabilization
- Segmentation ?

- Tracking ?

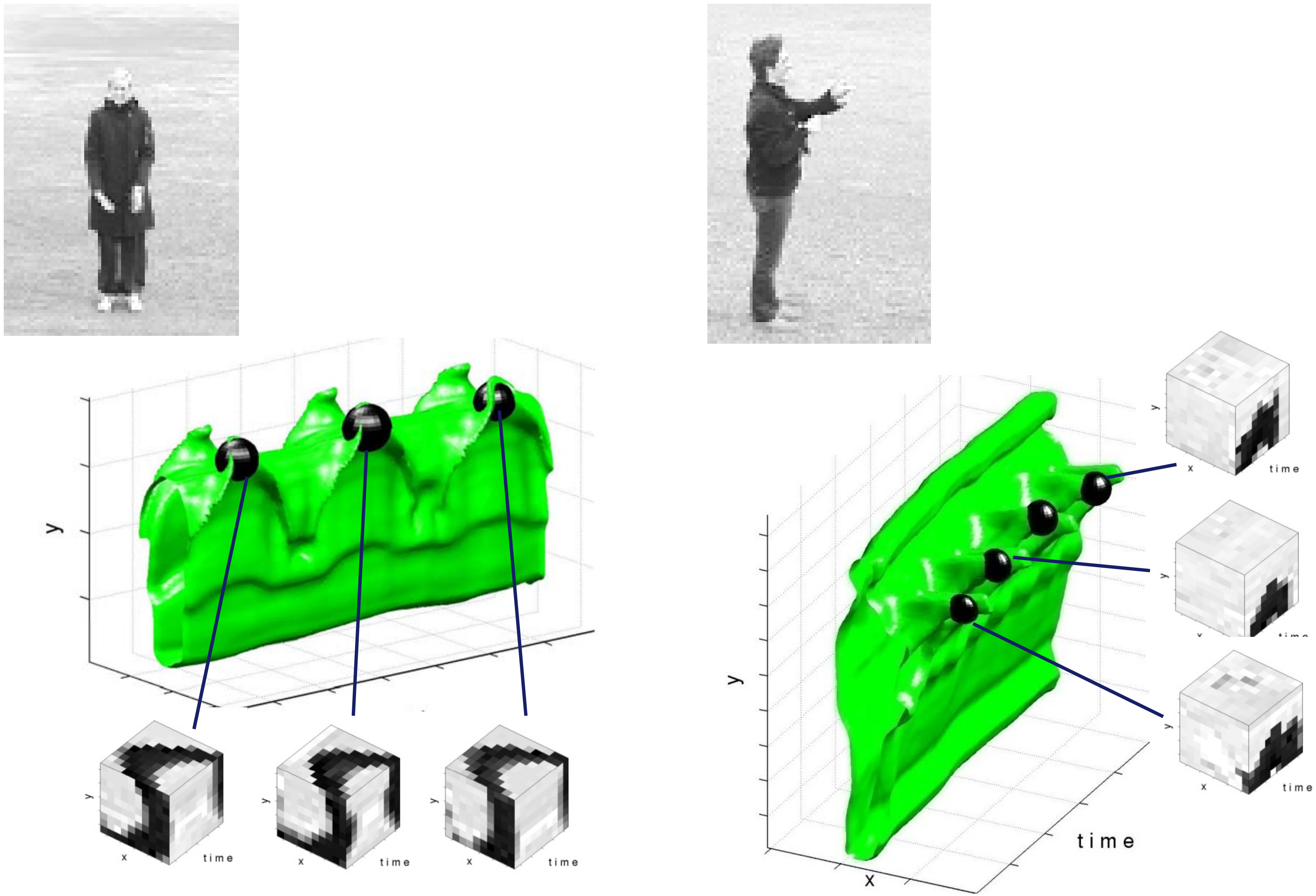
- Template-based methods ?

Common problems:

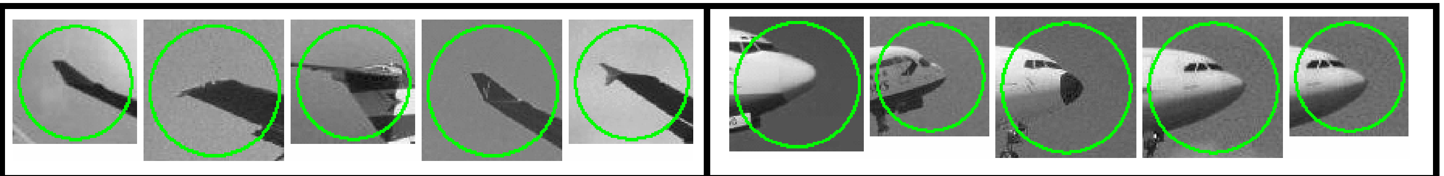
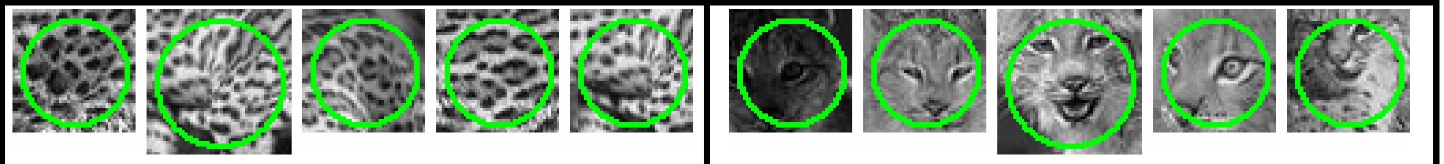
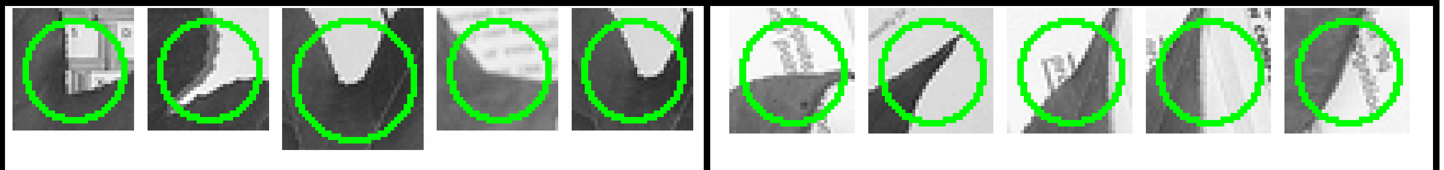
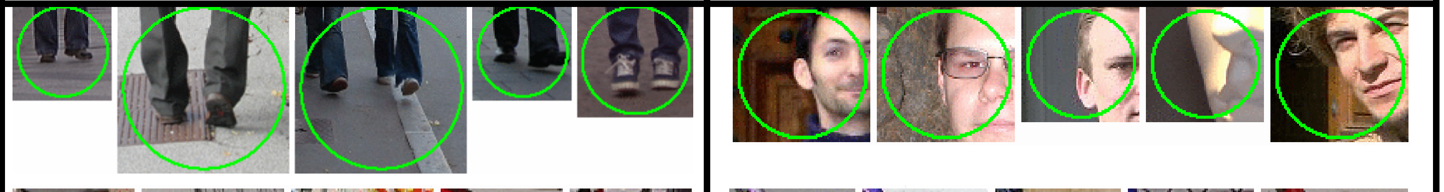
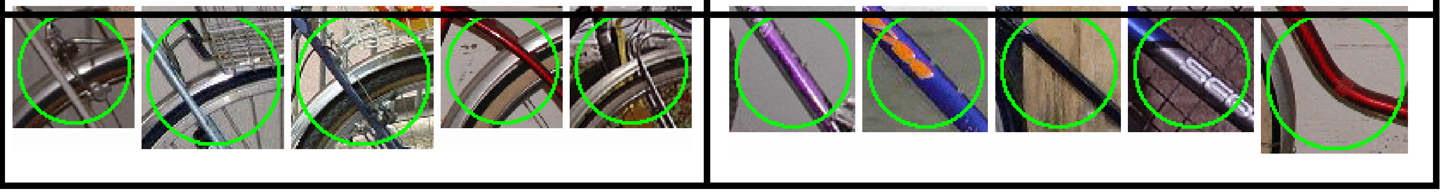
- Complex & changing BG
- Changes in appearance
- Large variations in motion

Avoid global assumptions!

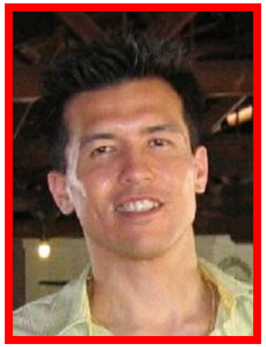
No global assumptions
=> Local measurements

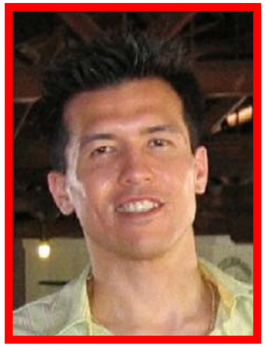


Relation to local image features

Airplanes	
Motorbikes	
Faces	
Wild Cats	
Leaves	
People	
Bikes	

Course overview





- Definitions
- Benchmark datasets
- Early silhouette and tracking-based methods
- Motion-based similarity measures
- Template-based methods
- Local space-time features
- Bag-of-Features action recognition
- Weakly-supervised methods
- Pose estimation and action recognition
- Action recognition in still images
- Human interactions and dynamic scene models
- Conclusions and future directions

Space-Time Interest Points

What neighborhoods to consider?

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

Definitions:

$f: \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ 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, IJCV 2005]

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):

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

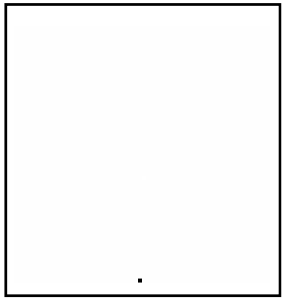
(similar to Harris operator [Harris and Stephens, 1988])

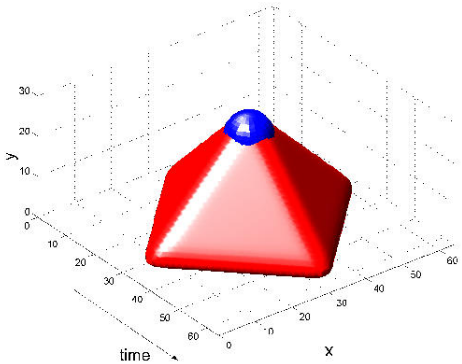
[Laptev, IJCV 2005]

Space-Time Interest Points: Examples

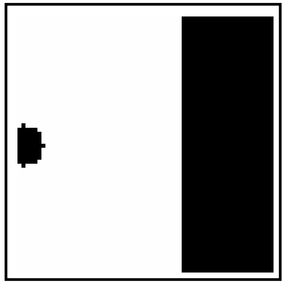
Motion event detection: synthetic sequences

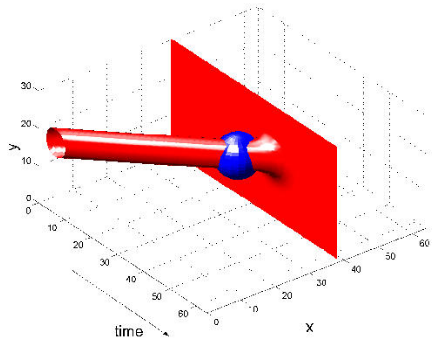
accelerations



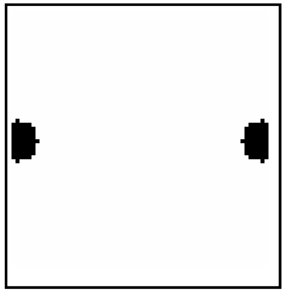


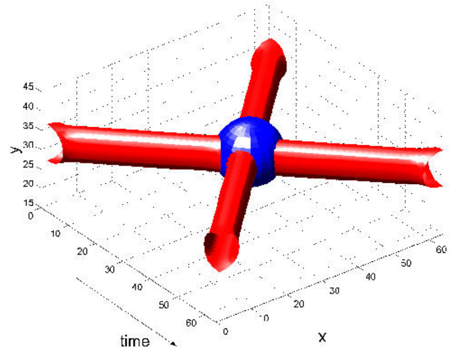
appearance/
disappearance





split/merge

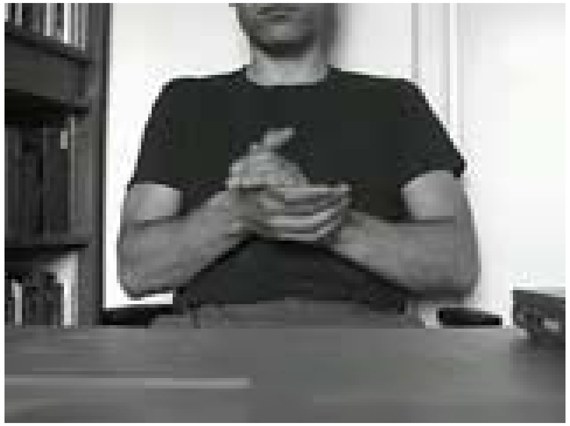


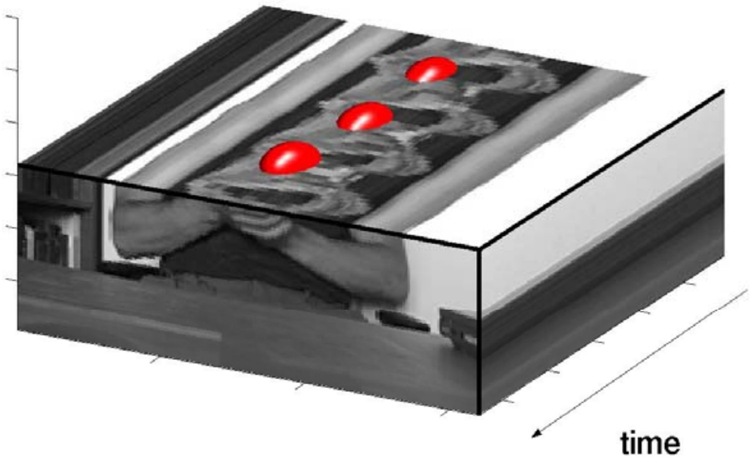


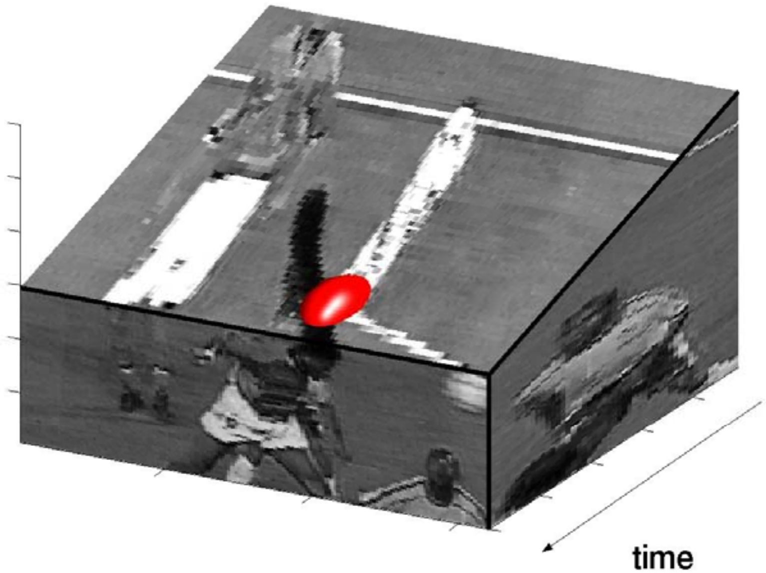
[Laptev, IJCV 2005]

Space-Time Interest Points: Examples

Motion event detection





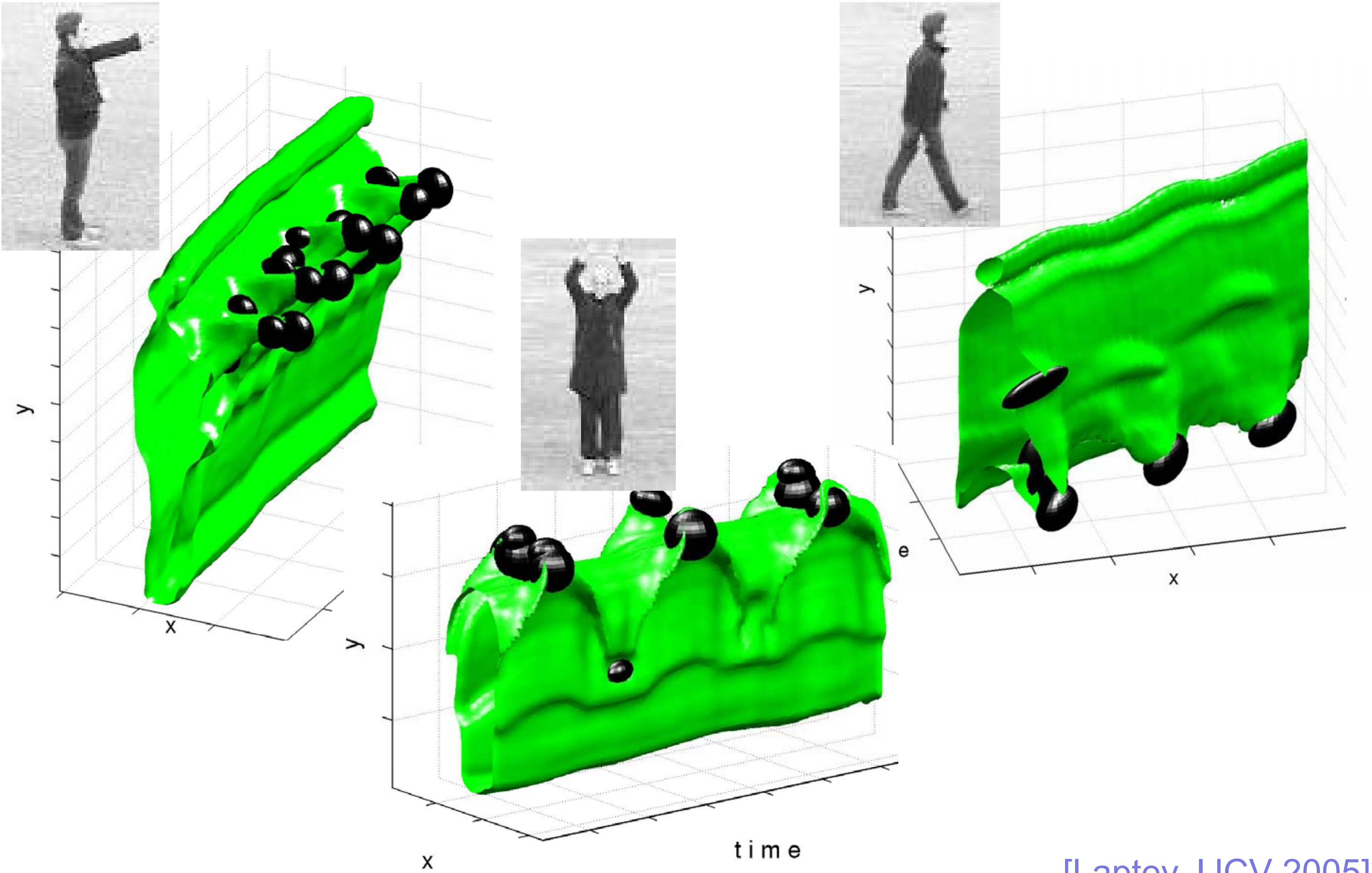


Space-Time Interest Points: Examples

Motion event detection: complex background

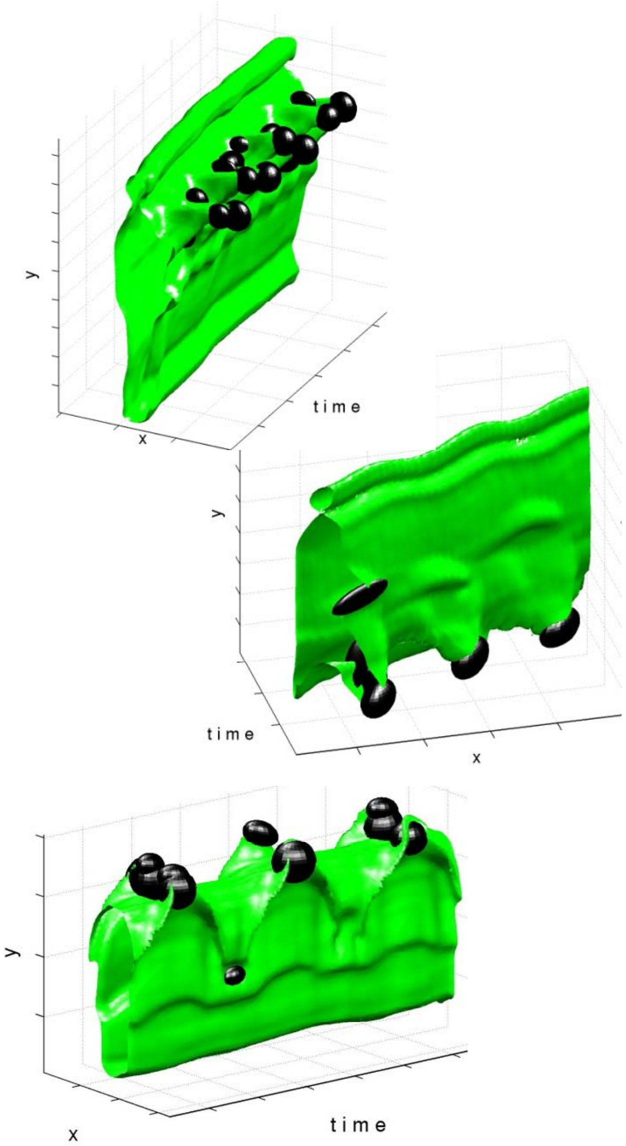
[Laptev, IJCV 2005]

Features from human actions

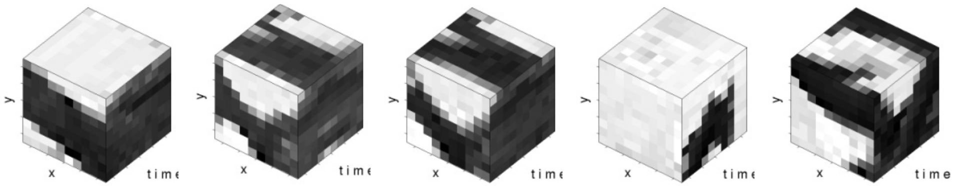


[Laptev, IJCV 2005]

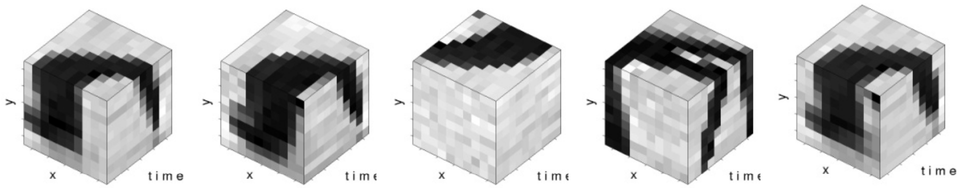
Features from human actions



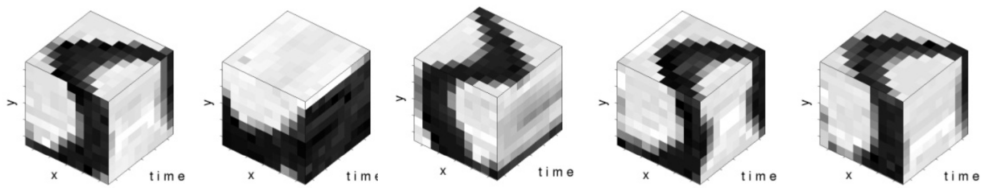
boxing



walking



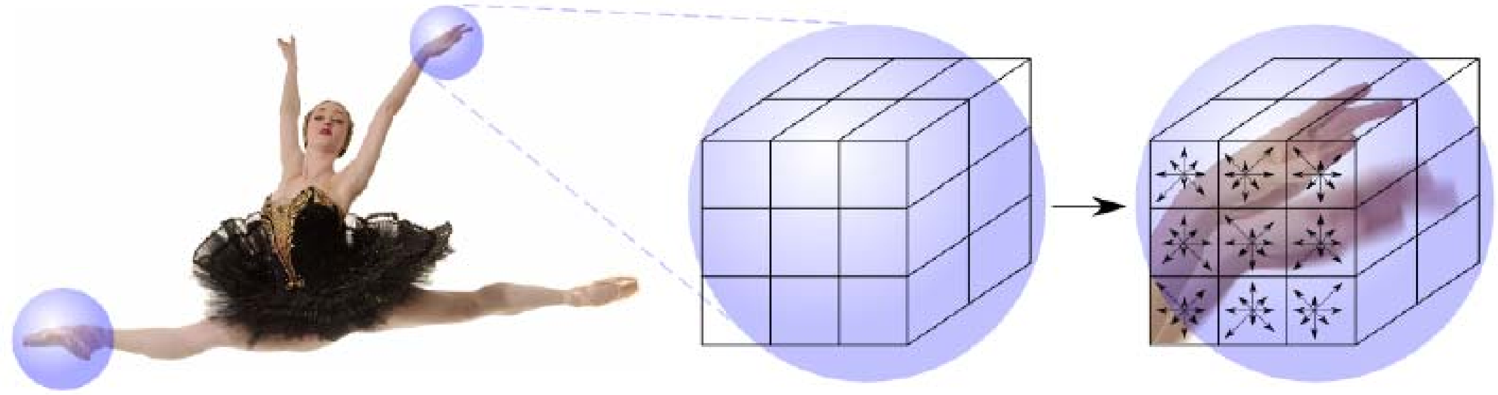
hand waving



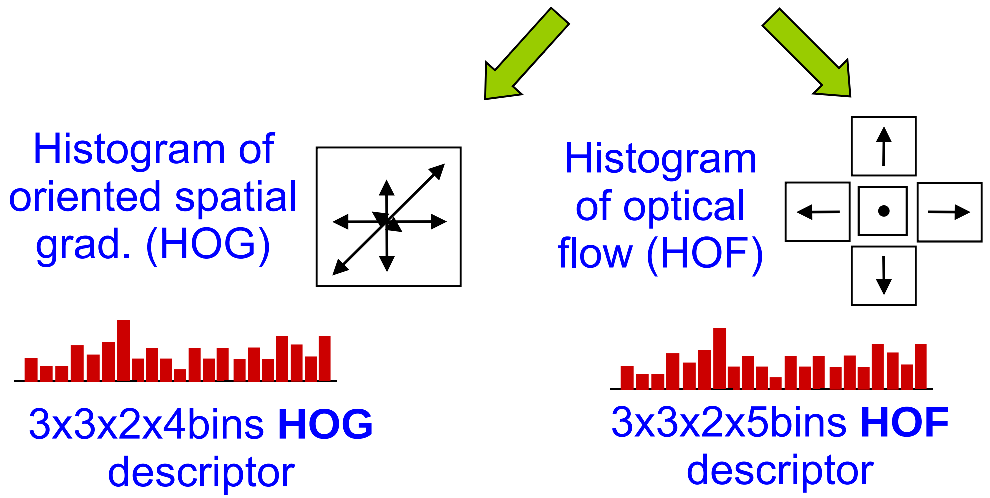
[Laptev, IJCV 2005]

Space-Time Features: Descriptor

Multi-scale space-time patches
from corner detector



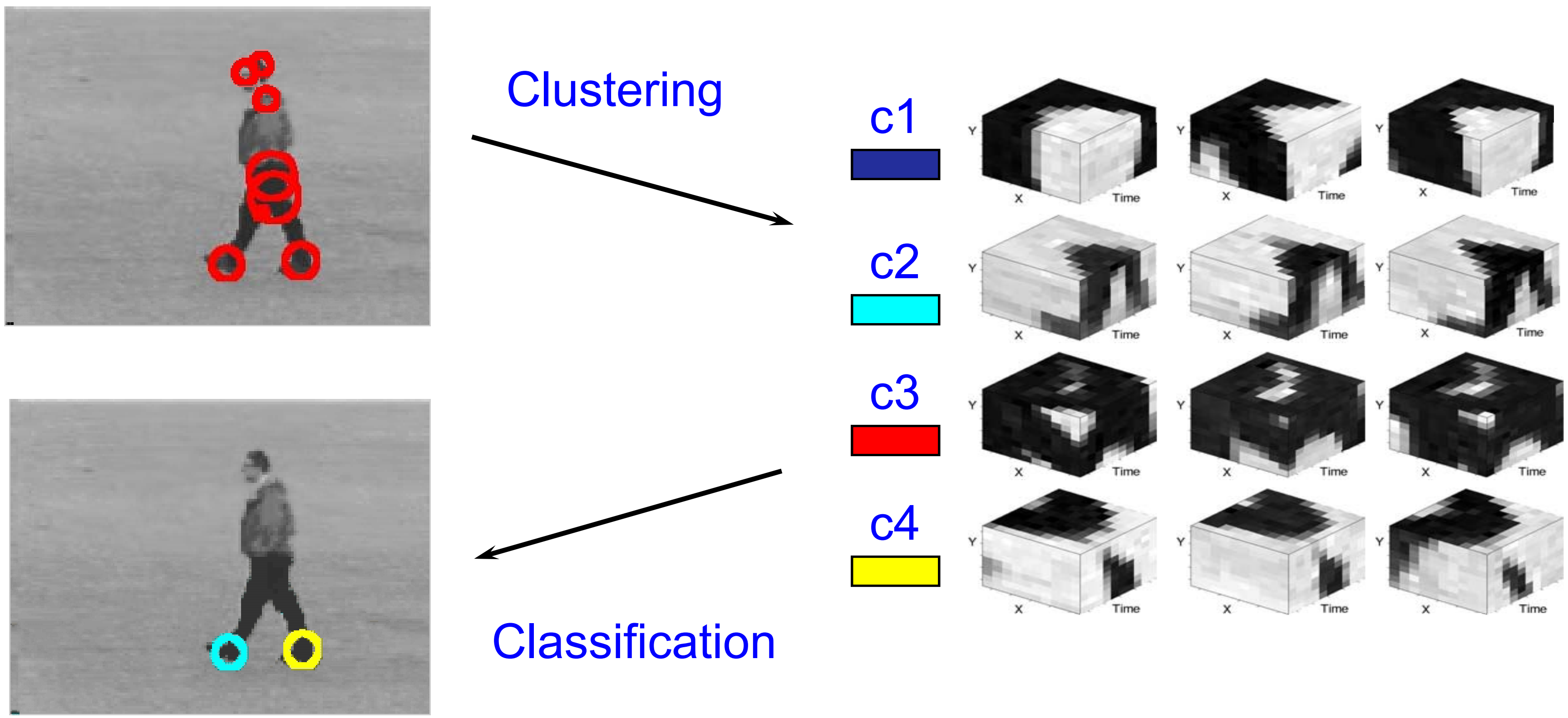
Public code available at
www.irisa.fr/vista/actions



[Laptev, Marszałek, Schmid, Rozenfeld, CVPR 2008]

Visual Vocabulary: K-means clustering

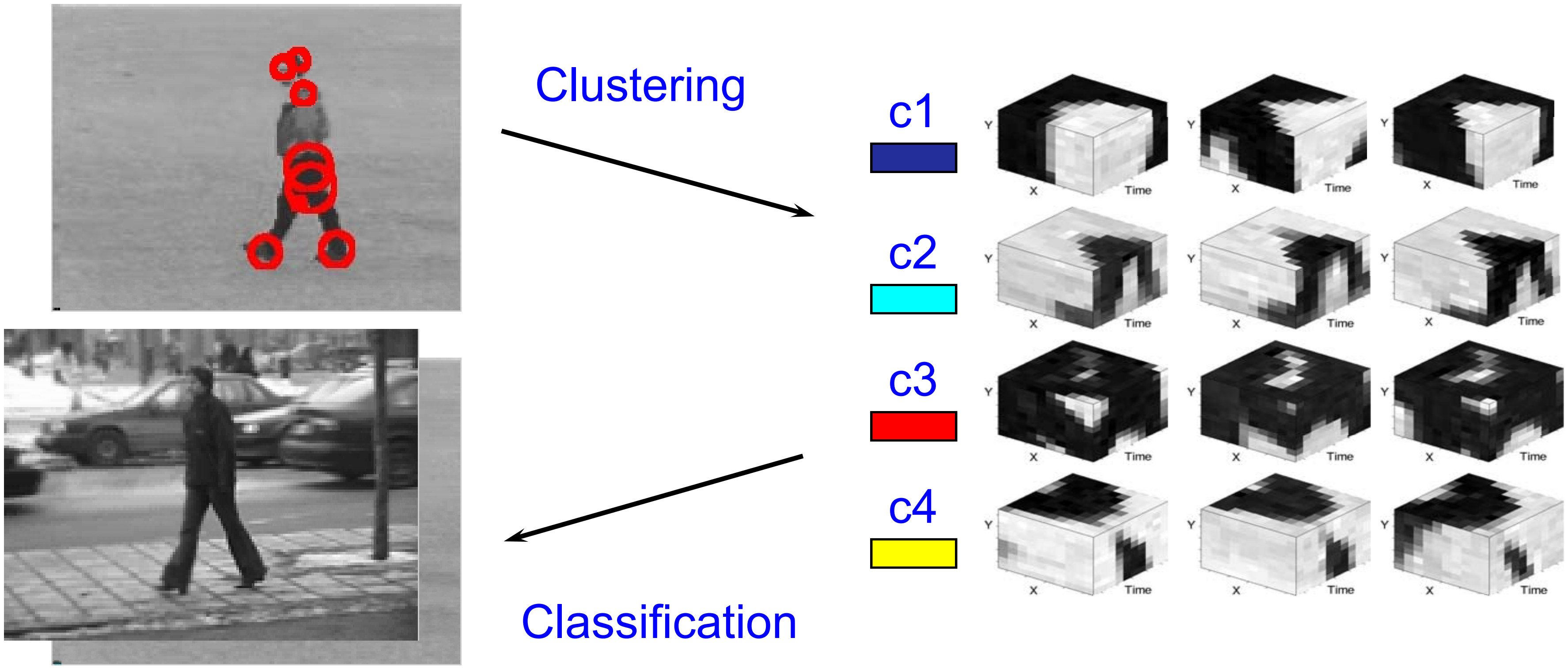
- Group similar points in the space of image descriptors using K-means clustering
- Select significant clusters



[Laptev, IJCV 2005]

Visual Vocabulary: K-means clustering

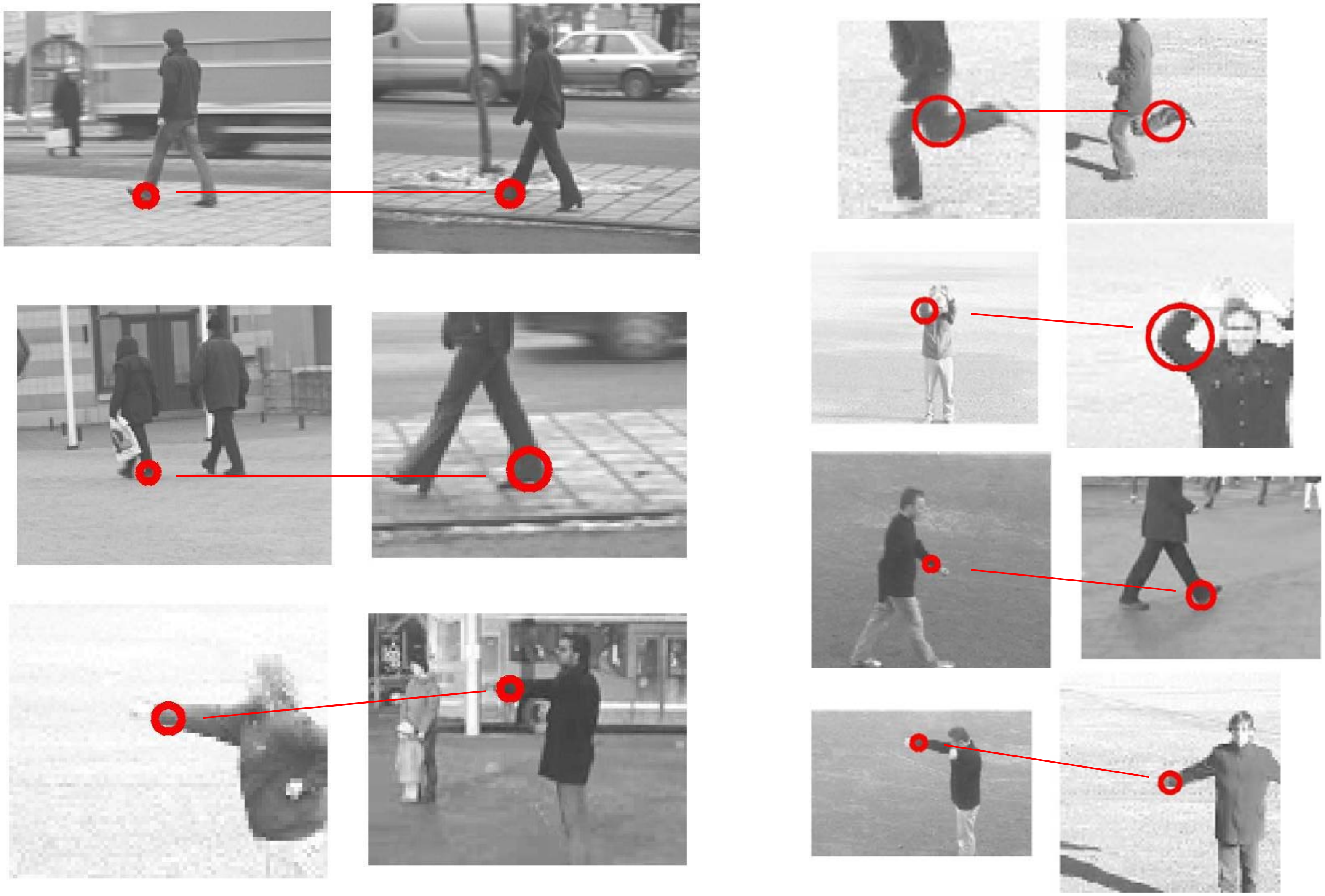
- Group similar points in the space of image descriptors using K-means clustering
- Select significant clusters



[Laptev, IJCV 2005]

Local Space-time features: Matching

- Find similar events in pairs of video sequences



Periodic Motion

- **Periodic views of a sequence can be approximately treated as stereopairs**

$\{s_t, \ldots, s_m\}$

$\{s_{t+p}, \ldots, s_{n+p}\}$
$(p$: period$)$

...

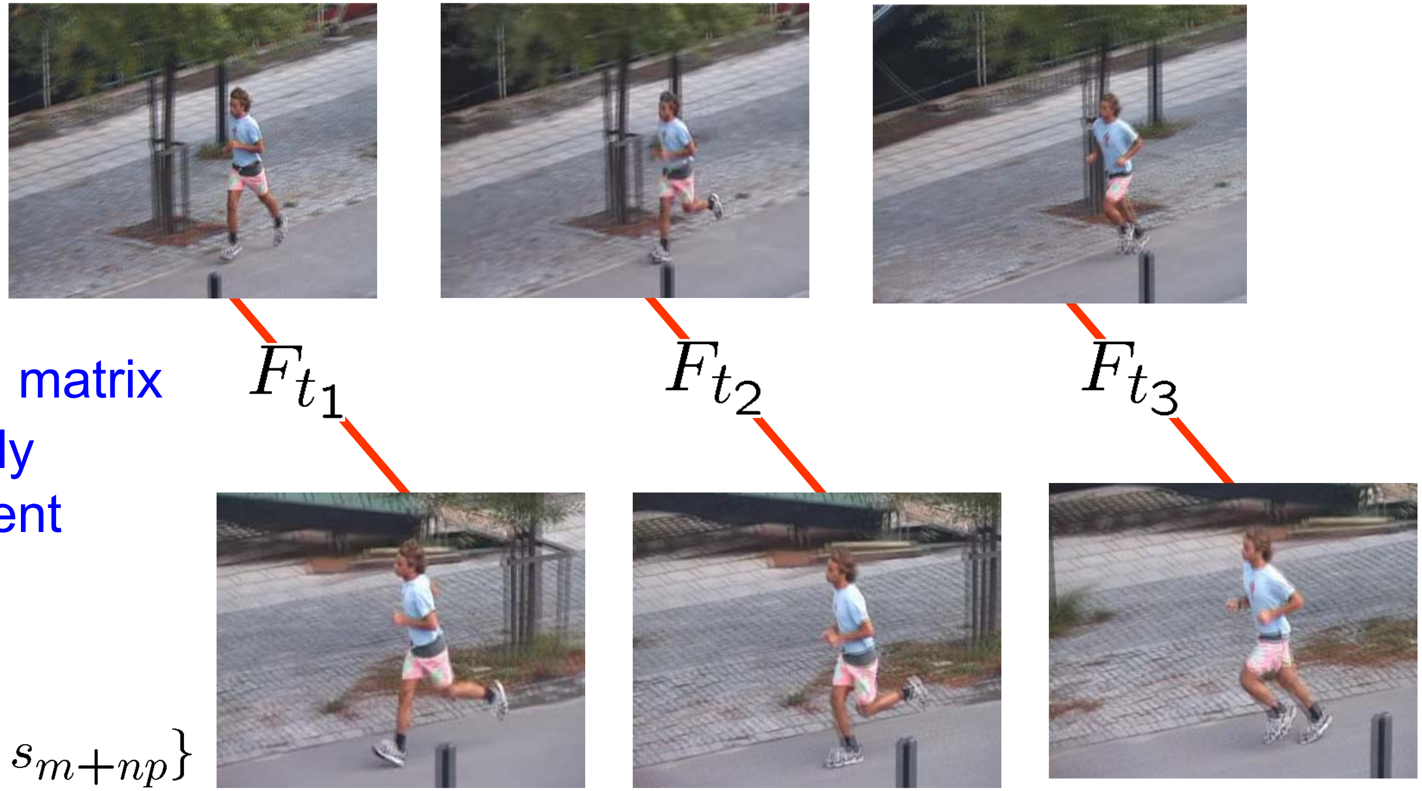
$\{s_{t+np}, \ldots, s_{m+np}\}$

[Laptev, Belongie, Pérez, Wills, ICCV 2005]

Periodic Motion

- **Periodic views of a sequence can be approximately treated as stereopairs**

$\{s_t, \ldots, s_m\}$



Fundamental matrix F is generally time-dependent

$\{s_{t+np}, \ldots, s_{m+np}\}$

⇨ Periodic motion estimation ~ sequence alignment

[Laptev, Belongie, Pérez, Wills, ICCV 2005]

Sequence alignment

Generally hard problem

- Unknown positions and motions of cameras
- Unknown temporal offset
- Possible time warping

Prior work treats special cases

- Caspi and Irani "*Spatio-temporal alignment of sequences*", PAMI 2002
- Rao et.al. "*View-invariant alignment and matching of video sequences*", ICCV 2003
- Tuytelaars and Van Gool "*Synchronizing video sequences*", CVPR 2004

Useful for

- Reconstruction of dynamic scenes
- *Recognition* of dynamic scenes

[Laptev, Belongie, Pérez, Wills, ICCV 2005]

Sequence alignment

Constant translation

- Assume the camera is translating with velocity V relatively to the object

$\Rightarrow$ For sequences $\begin{array}{l} S_a = \{s_t, \ldots, s_m\} \\ S_b = \{s_{t+np}, \ldots, s_{m+np}\} \end{array}$

corresponding points are related by

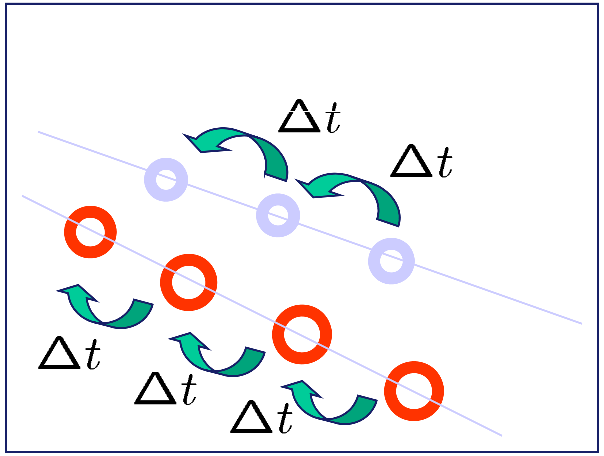
$x_t^\top F x_{t+np} = 0$ with $F = [npV]_\times R \sim [V]_\times$

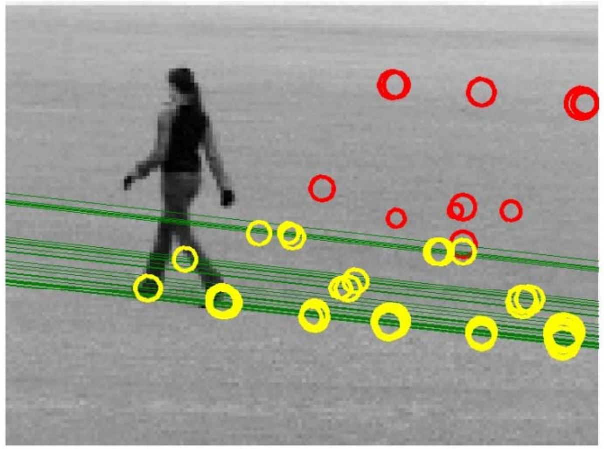
$\Rightarrow$ All corresponding periodic points are on the same epipolar line

[Laptev, Belongie, Pérez, Wills, ICCV 2005]

Periodic motion detection

1. Corresponding points have similar descriptors





2. Same period for all features $p = \Delta t$

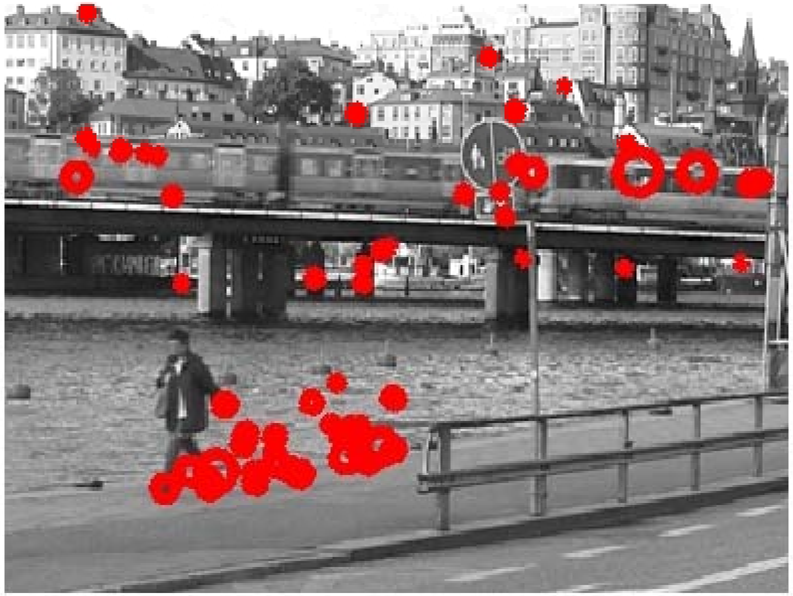
3. Spatial arrangement of features across periods satisfy epipolar constraint: $[x^t]' F x^{t+p} = 0$

➡ Use RANSAC to estimate *F* and *p*

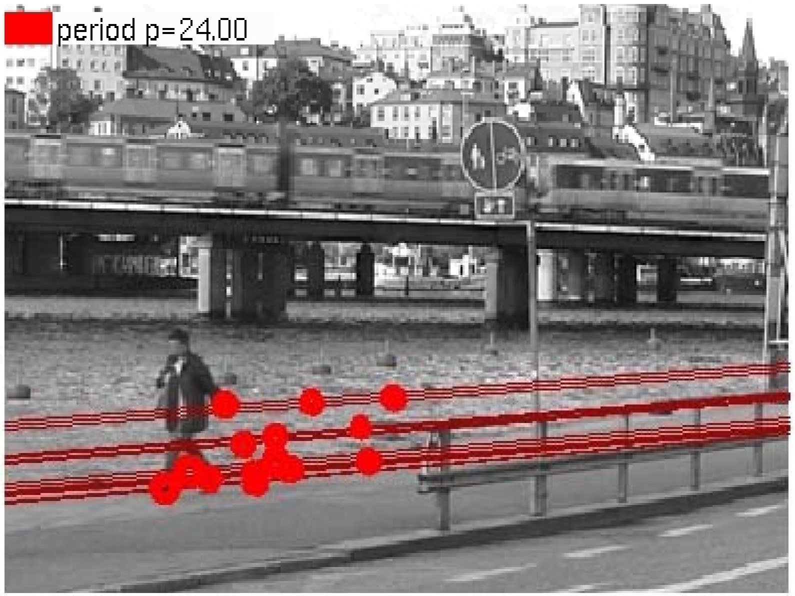
[Laptev, Belongie, Pérez, Wills, ICCV 2005]

Periodic motion detection

Original space-time features



RANSAC estimation of F,p

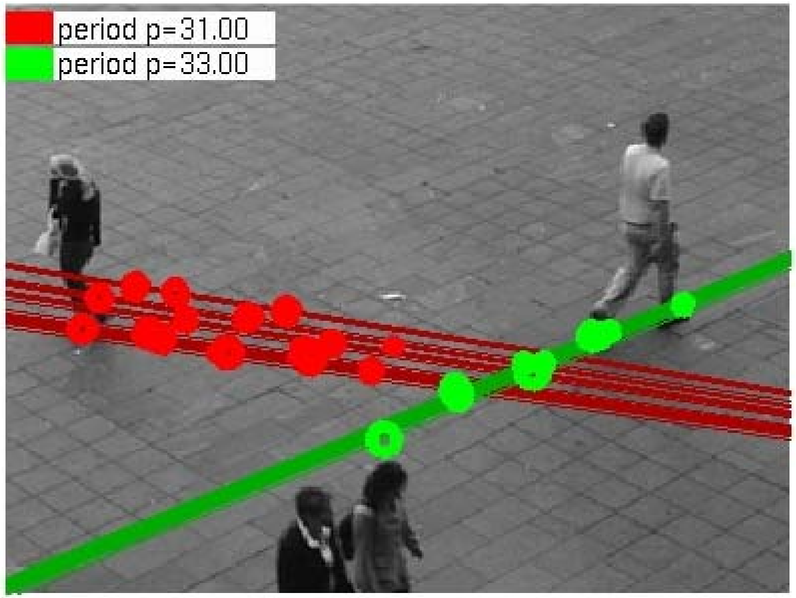


[Laptev, Belongie, Pérez, Wills, ICCV 2005]

Periodic motion detection

Original space-time features

RANSAC estimation of F,p



[Laptev, Belongie, Pérez, Wills, ICCV 2005]

Periodic motion segmentation

- Assume periodic objects are planar

⇨ Periodic points can be related by a *dynamic homography:*

$x_t = Hx_{t+p}$ with

linear in time

$$H(t) = I + p(\mathbf{v}\mathbf{n}^\top - \mathbf{n}^\top\mathbf{v}I)/d - \boxed{t}\,\mathbf{n}^\top\mathbf{v}I/d$$

[Laptev, Belongie, Pérez, Wills, ICCV 2005]

Periodic motion segmentation

- **Assume periodic objects are planar**

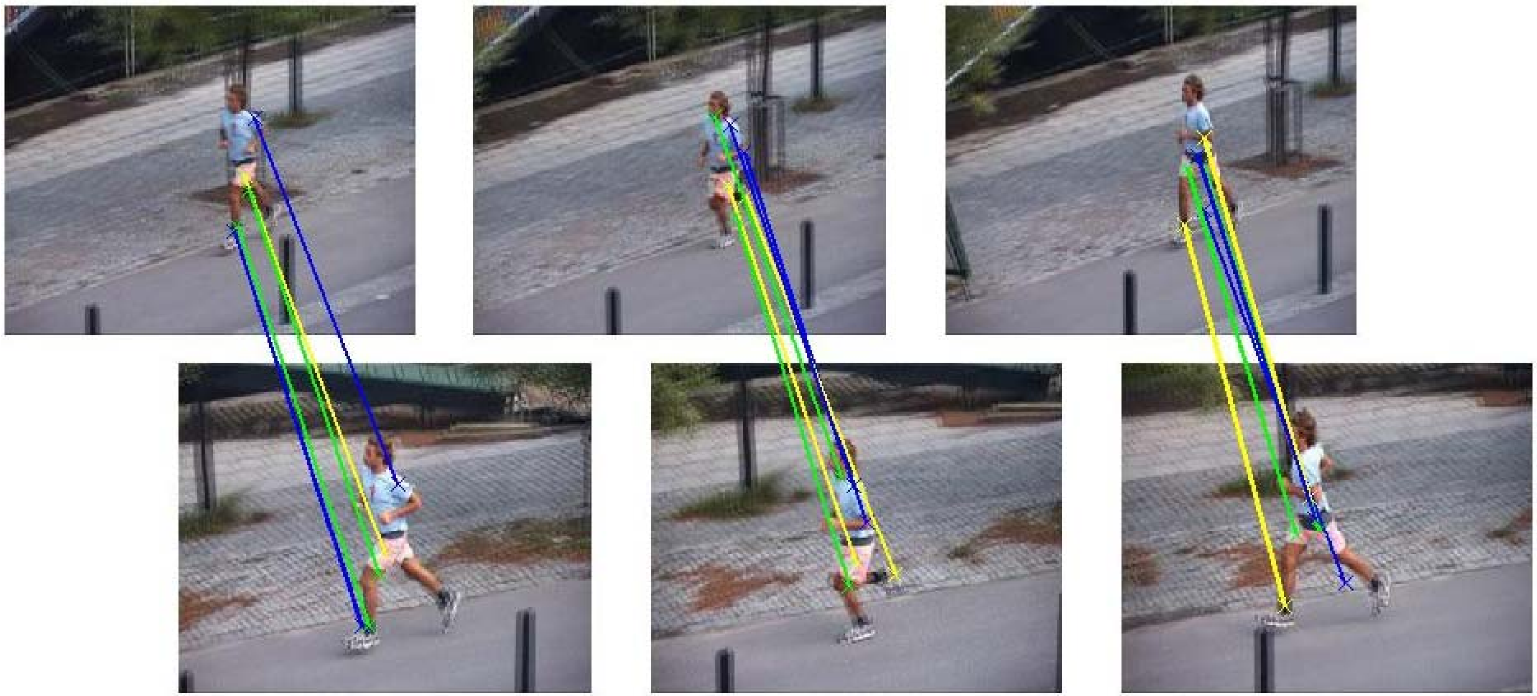
  $\Rightarrow$ Periodic points can be related by a *dynamic homography:*

$$x_t = H x_{t+p} \text{ with}$$

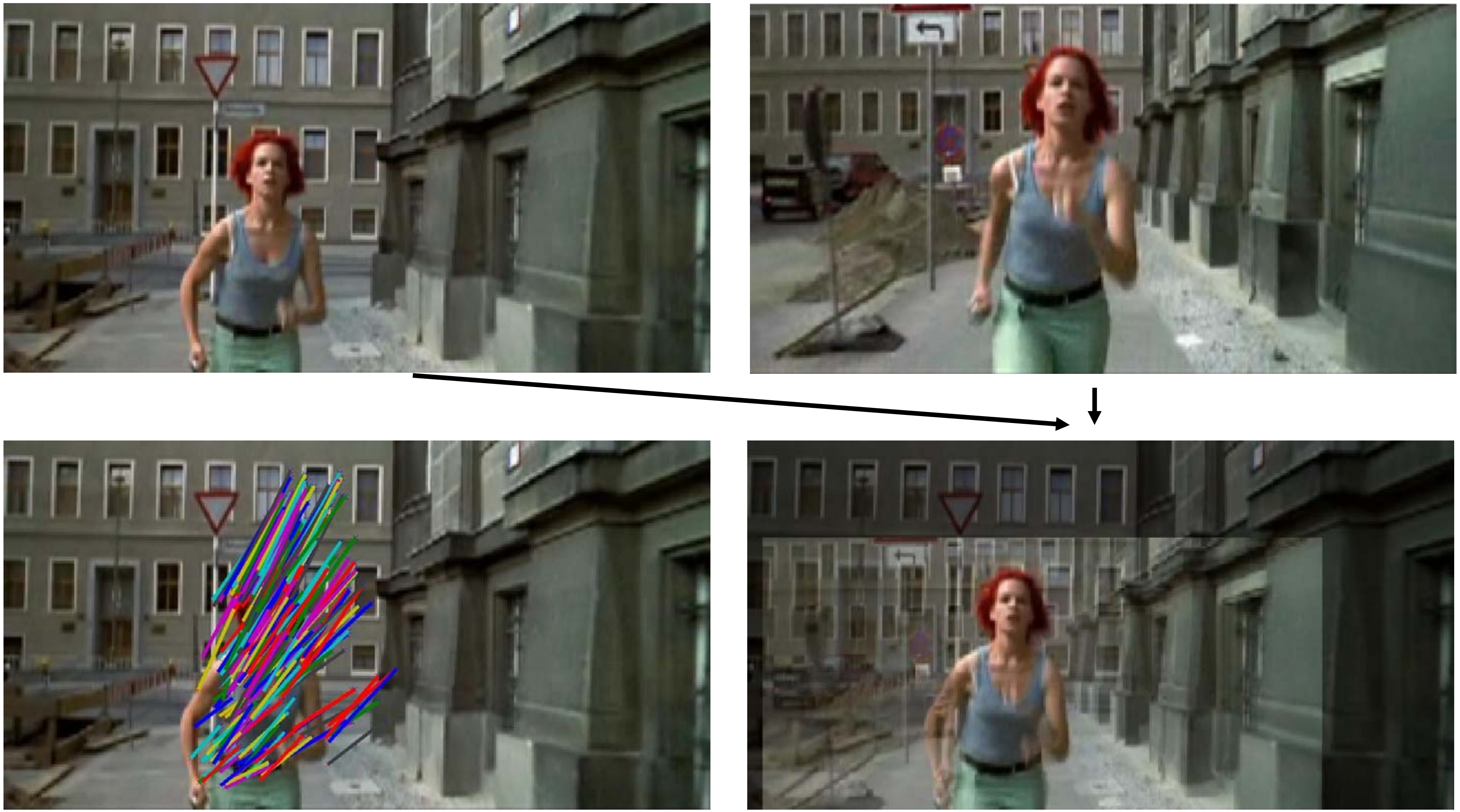
linear in time

$$H(t) = I + p(\mathbf{v}\mathbf{n}^\top - \mathbf{n}^\top\mathbf{v}I)/d - \boxed{t}\,\mathbf{n}^\top\mathbf{v}I/d$$

$\Rightarrow$ RANSAC estimation of *H* and *p*



Object-centered stabilization



[Laptev, Belongie, Pérez, Wills, ICCV 2005]

Segmentation

Disparity estimation

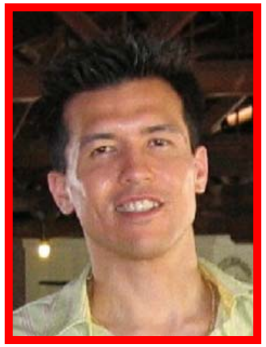
Graph-cut segmentation

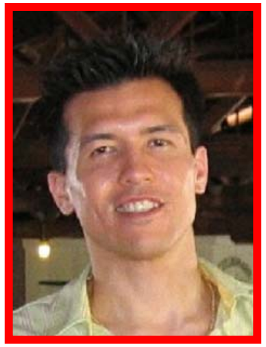
[Laptev, Belongie, Pérez, Wills, ICCV 2005]

Segmentation

[I. Laptev, S.J. Belongie, P. Pérez and J. Wills, ICCV 2005]

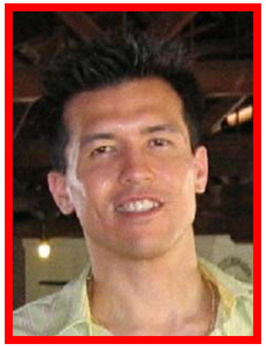
Course overview

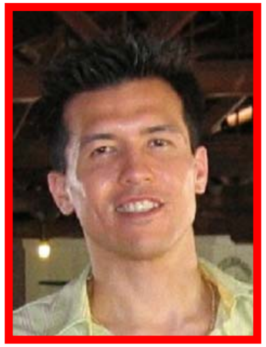




- Definitions
- Benchmark datasets
- Early silhouette and tracking-based methods
- Motion-based similarity measures
- Template-based methods
- Local space-time features
- Bag-of-Features action recognition
- Weakly-supervised methods
- Pose estimation and action recognition
- Action recognition in still images
- Human interactions and dynamic scene models
- Conclusions and future directions

Course overview

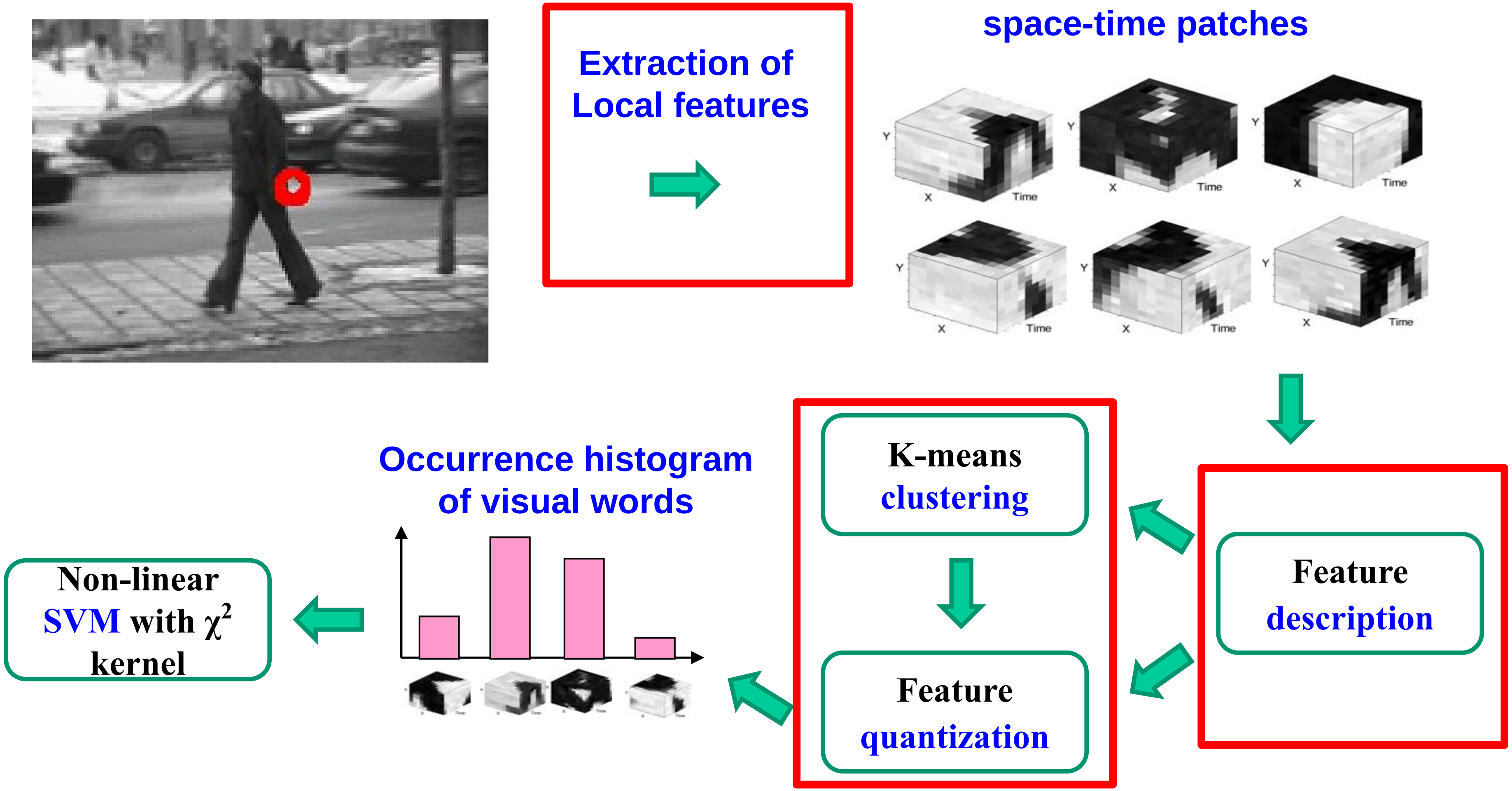




- Definitions
- Benchmark datasets
- Early silhouette and tracking-based methods
- Motion-based similarity measures
- Template-based methods
- Local space-time features
- Bag-of-Features action recognition
- Weakly-supervised methods
- Pose estimation and action recognition
- Action recognition in still images
- Human interactions and dynamic scene models
- Conclusions and future directions

Action recognition framework

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



The spatio-temporal features/descriptors

- **Features: Detectors**
  - Harris3D [I. Laptev, IJCV 2005]
  - Dollar [P. Dollar et al., VS-PETS 2005]
  - Hessian [G. Willems et al, ECCV 2008]
  - Regular sampling [H. Wang et al. BMVC 2009]

- **Descriptors**
  - HoG/HoF [I. Laptev, et al. CVPR 2008]
  - Dollar [P. Dollar et al., VS-PETS 2005]
  - HoG3D [A. Klaeser et al., BMVC 2008]
  - Extended SURF [G. Willems et al., ECCV 2008]

Illustration of ST detectors

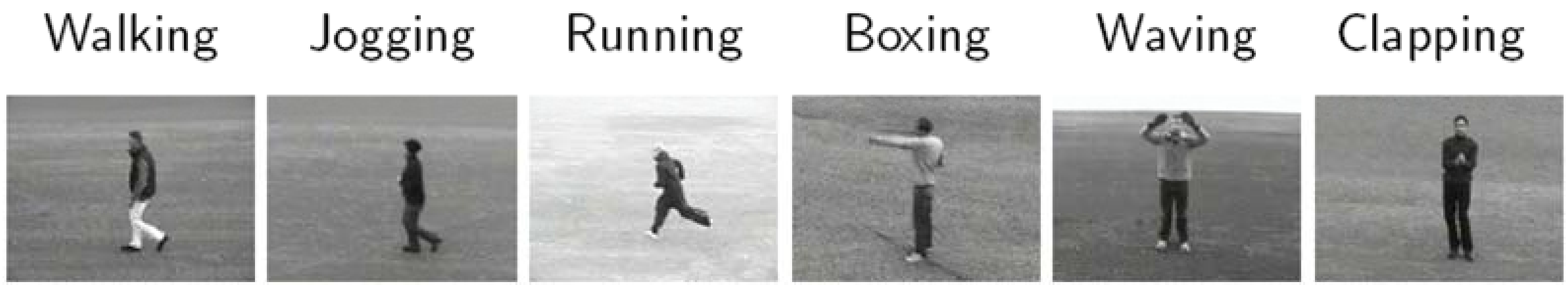
Harris3D

Hessian

Cuboid

Dense

Results: KTH actions



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
- Dense features perform relatively poor compared to sparse features

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

Results: UCF sports

Diving

Walking

Kicking

Skateboarding

High-Bar-Swinging

Golf-Swinging

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
- Cuboids: good performance with HOG3D

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

Results: Hollywood-2

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%	-

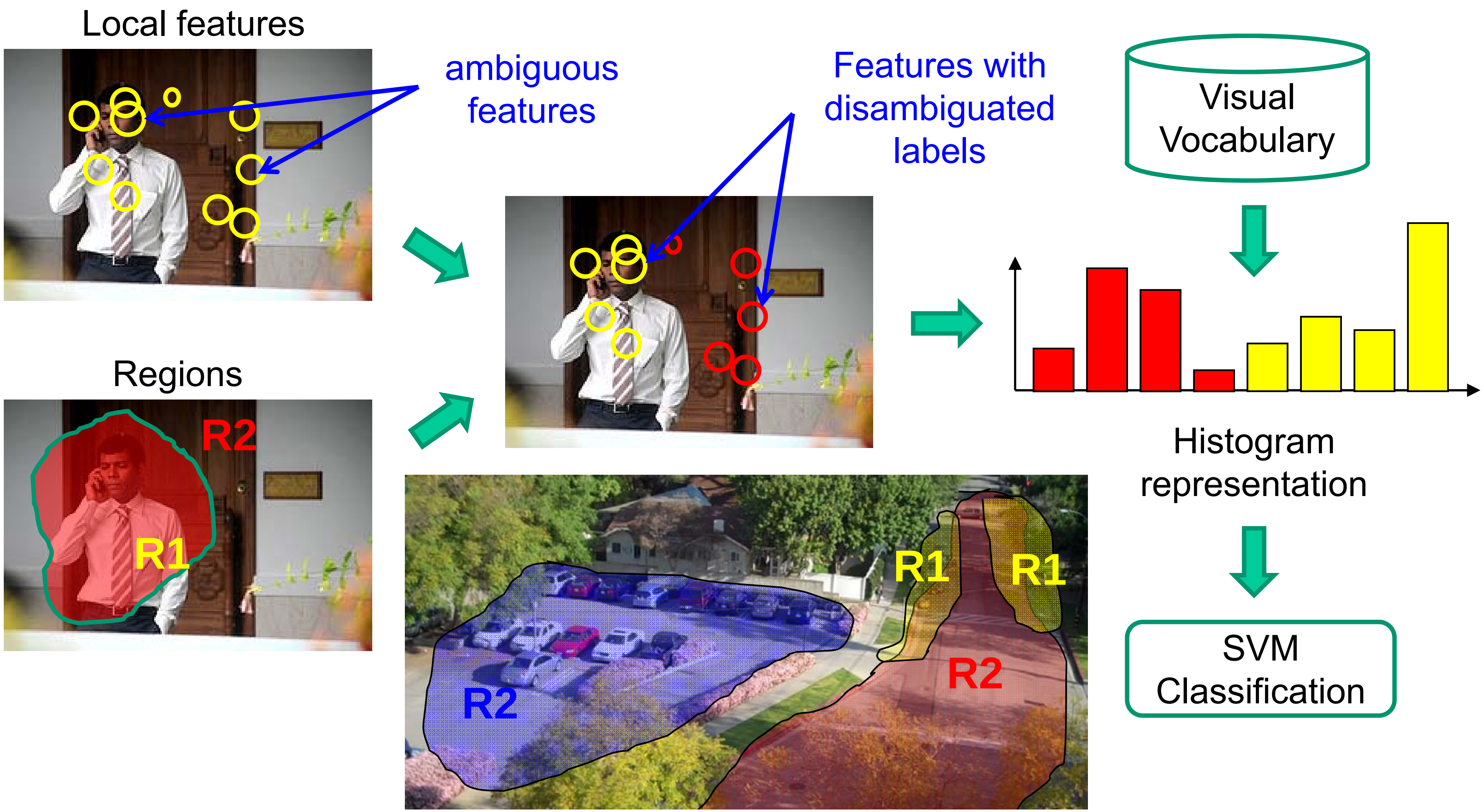
- Best results for **Dense** + HOG/HOF
- Good results for HOG/HOF

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

Improved BoF action classification

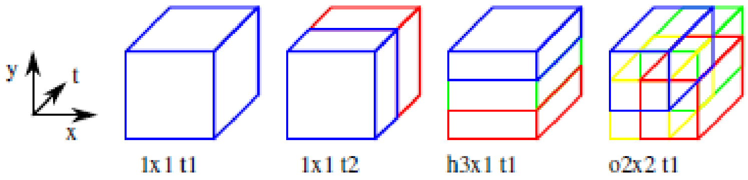
Goals:

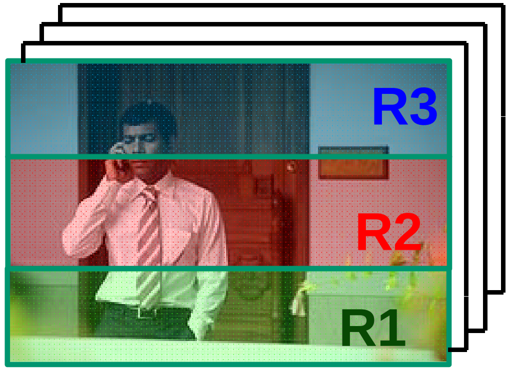
- Inject additional supervision into BoF
- Improve local descriptors with region-level information



Video Segmentation

- Spatio-temporal grids

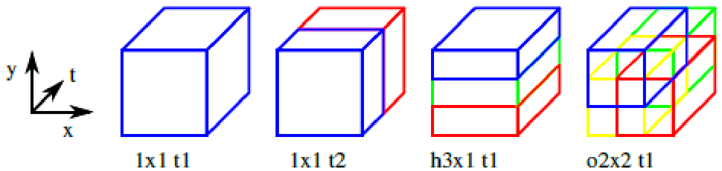




- Static action detectors [Felzenszwalb'08]
  - *Trained from ~100 web-images per class*

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

Video Segmentation



Multi-channel chi-square kernel

Use SVMs with a multi-channel chi-square kernel for classification

$$K(H_i, H_j) = \exp\Big(-\sum_{c \in \mathcal{C}} \frac{1}{A_c} D_c(H_i, H_j)\Big)$$

- Channel *c* corresponds to particular region segmentation
- $D_c(H_i, H_j)$ is the chi-square distance between histograms
- A_c is the mean value of the distances between all training samples
- The best set of channels *C* for a given training set is found based on a greedy approach

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

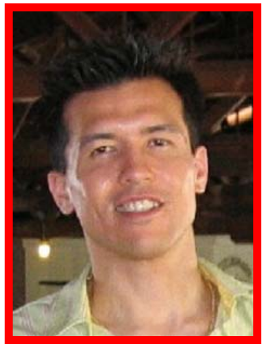
[Ullah, Parizi, Laptev, BMVC 2009]

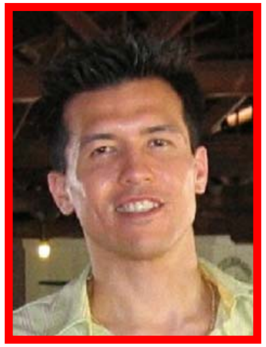
Hollywood-2 action classification

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**%

[Ullah, Parizi, Laptev, BMVC 2009]

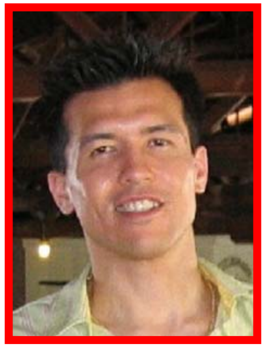
Course overview

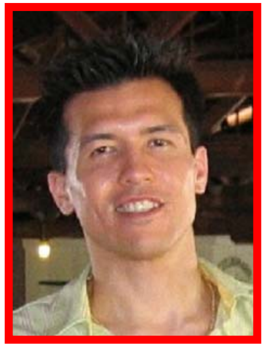




- Definitions
- Benchmark datasets
- Early silhouette and tracking-based methods
- Motion-based similarity measures
- Template-based methods
- Local space-time features
- Bag-of-Features action recognition
- Weakly-supervised methods
- Pose estimation and action recognition
- Action recognition in still images
- Human interactions and dynamic scene models
- Conclusions and future directions

Course overview





- Definitions
- Benchmark datasets
- Early silhouette and tracking-based methods
- Motion-based similarity measures
- Template-based methods
- Local space-time features
- Bag-of-Features action recognition
- Weakly-supervised methods
- Pose estimation and action recognition
- Action recognition in still images
- Human interactions and dynamic scene models
- Conclusions and future directions

Why is action recognition hard?

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

Drinking

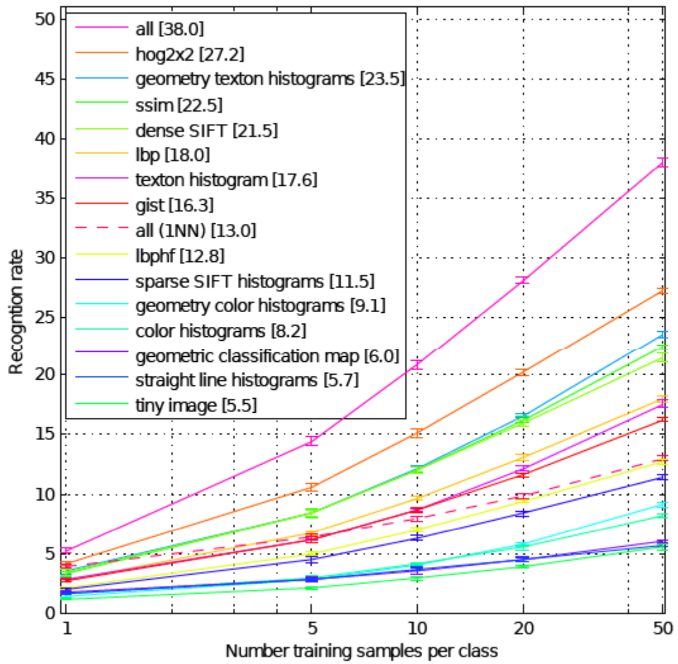
Smoking

- Lots of classes and concepts

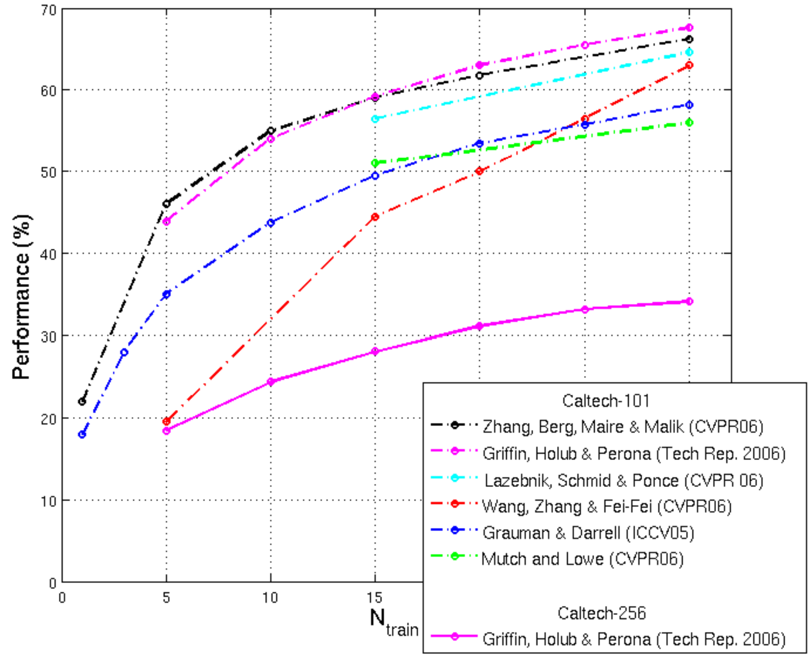
The positive effect of data

- The performance of current visual recognition methods heavily depends on the amount of available training data

Scene recognition: SUN database
[J. Xiao et al CVPR2010]



Object recognition: Caltech 101 / 256
[Griffin et al. Caltech tech. Rep.]



Action recognition:
[Laptev et al. CVPR2008,
Marszałek et al. CVPR2009]

Hollywood (~29 samples / class)	mAP: 38.4 %
Hollywood 2 (~75 samples / class)	mAP: 50.3%

The positive effect of data

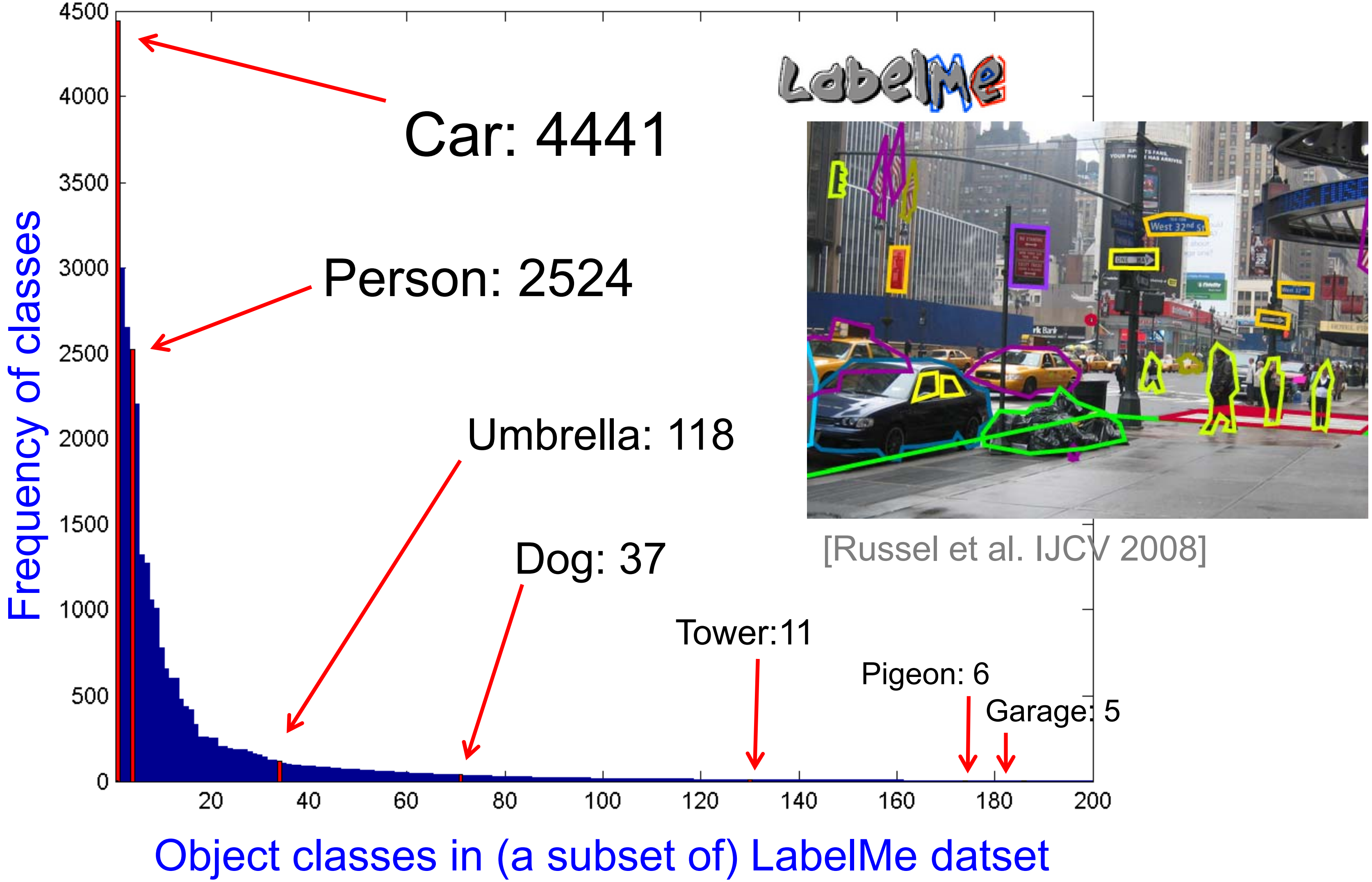
- The performance of current visual recognition methods heavily depends on the amount of available training data

⇨ Need to collect substantial amounts of data for training

⇨ Current algorithms may not scale well / be optimal for large datasets

- See also article "The Unreasonable Effectiveness of Data" by A. Halevy, P. Norvig, and F. Pereira, Google, *IEEE Intelligent Systems*

Why is data collection difficult?



Why is data collection difficult?

- A few classes are very frequent, but most of the classes are very rare

- Similar phenomena have been observed for non-visual data, e.g. word counts in natural language, etc. Such phenomena follow Zipf's empirical law:

  class rank = F(1 / class frequency)

- Manual supervision is very costly *especially for video*

Example: Common actions such as *Kissing*, *Hand Shaking* and *Answering Phone* appear 3-4 times in typical movies

→ ~42 hours of video needs to be inspected to collect 100 samples for each new action class

Learning Actions from Movies

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

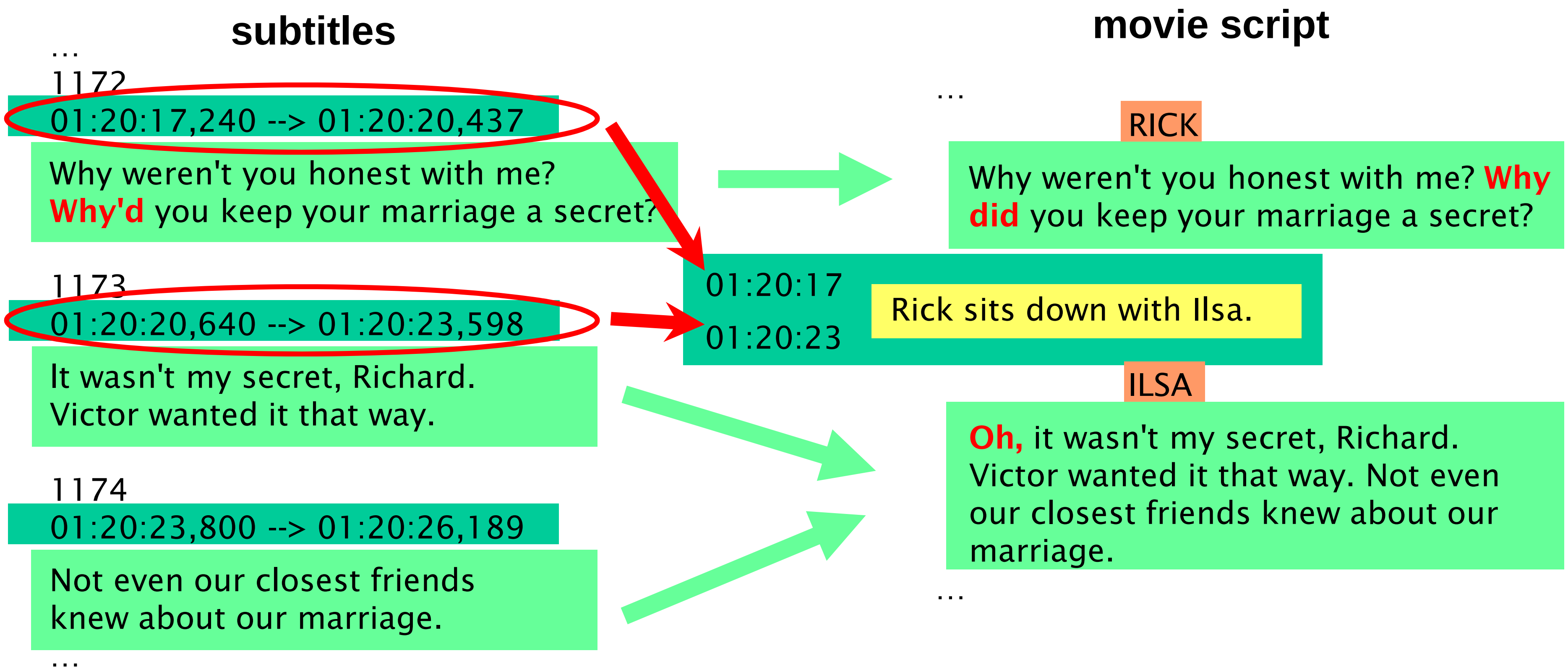
Problems:

- 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 alignment

RICK

All right, I will. Here's looking at
you, kid.

01:21:50
01:21:59

ILSA

I wish I didn't love you so much.

01:22:00
01:22:03

She snuggles closer to Rick.

CUT TO:

EXT. RICK'S CAFE - NIGHT

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

The headlights of a speeding police car sweep toward them.

They flatten themselves against a wall to avoid detection.

The lights move past them.

01:22:15
01:22:17

CARL

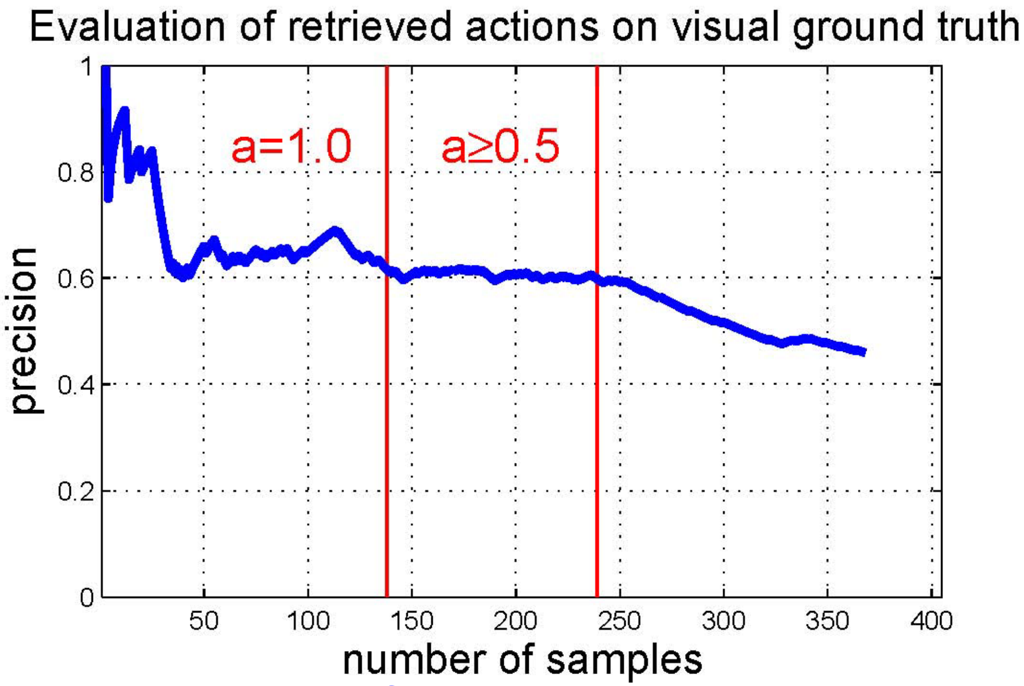
I think we lost them.

...

[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.

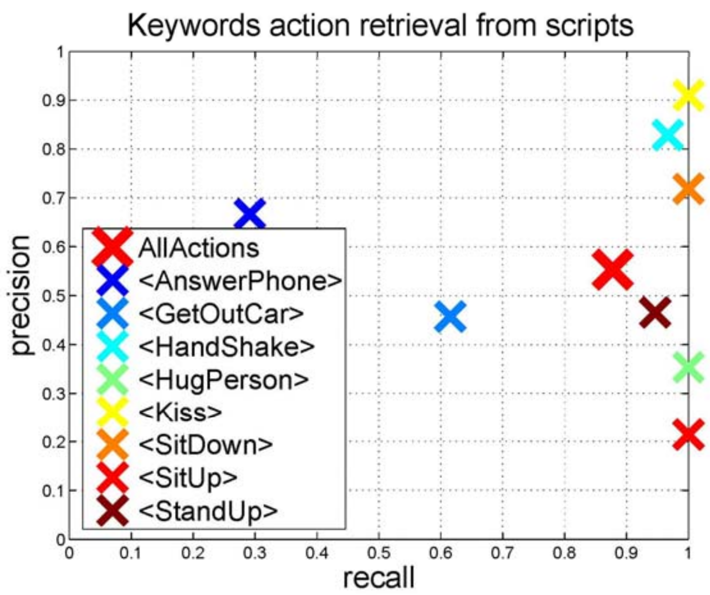
[Laptev, Marszałek, Schmid, Rozenfeld 2008]

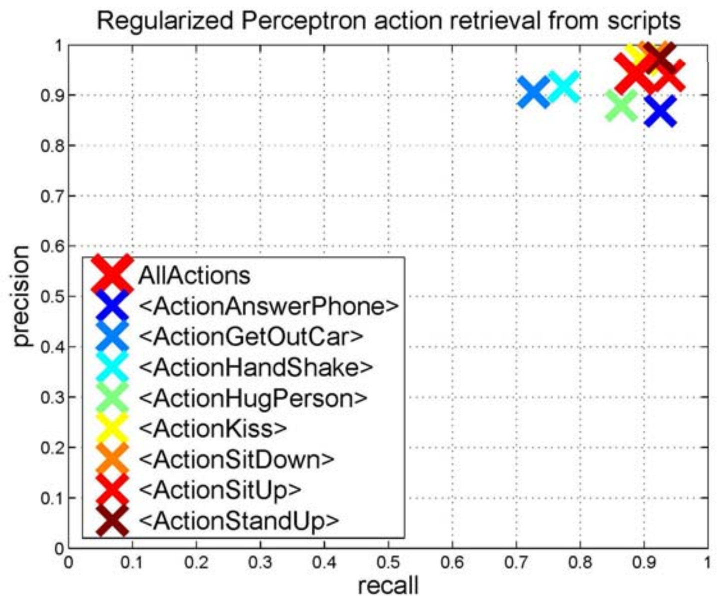
Text-based action retrieval

- Large variation of action expressions in text:

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

- => Supervised text classification approach





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.

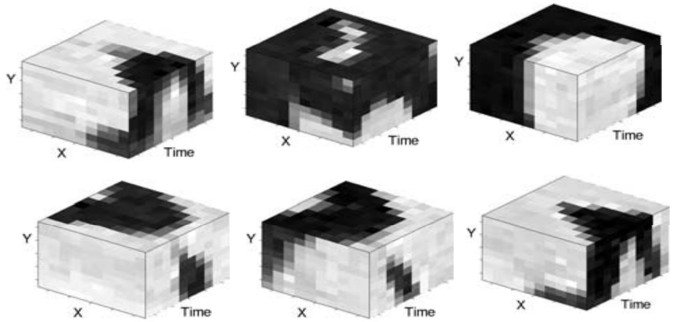
Hollywood-2 dataset is on-line: http://www.irisa.fr/vista/actions/hollywood2

- Learn vision-based classifier from automatic training set
- Compare performance to the manual training set

Bag-of-Features Recognition

Extraction of
Local features

space-time patches

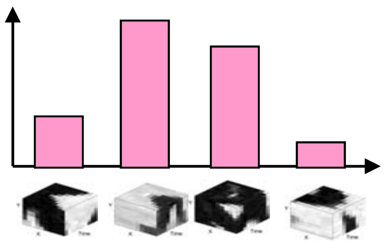


Feature
description

K-means
clustering
(k=4000)

Feature
quantization

Occurrence histogram
of visual words

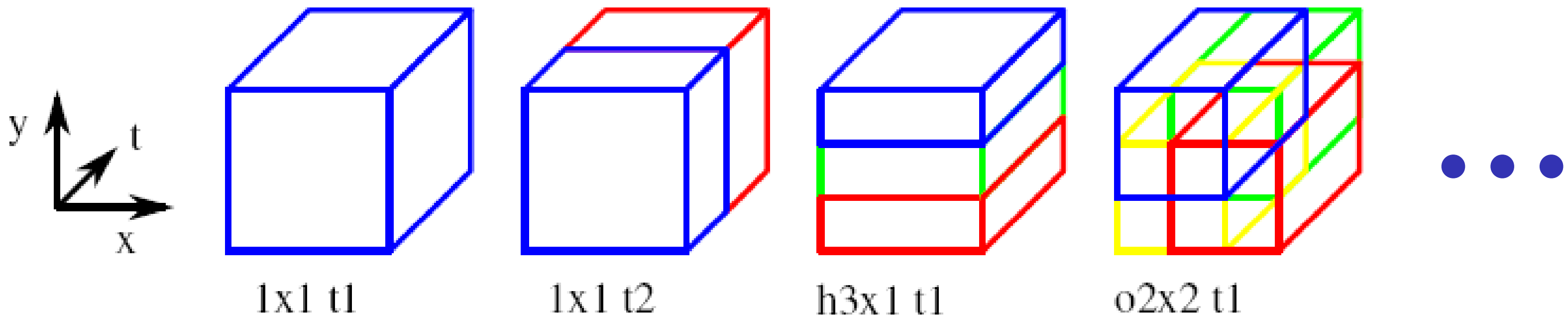


Non-linear
SVM with χ^2
kernel

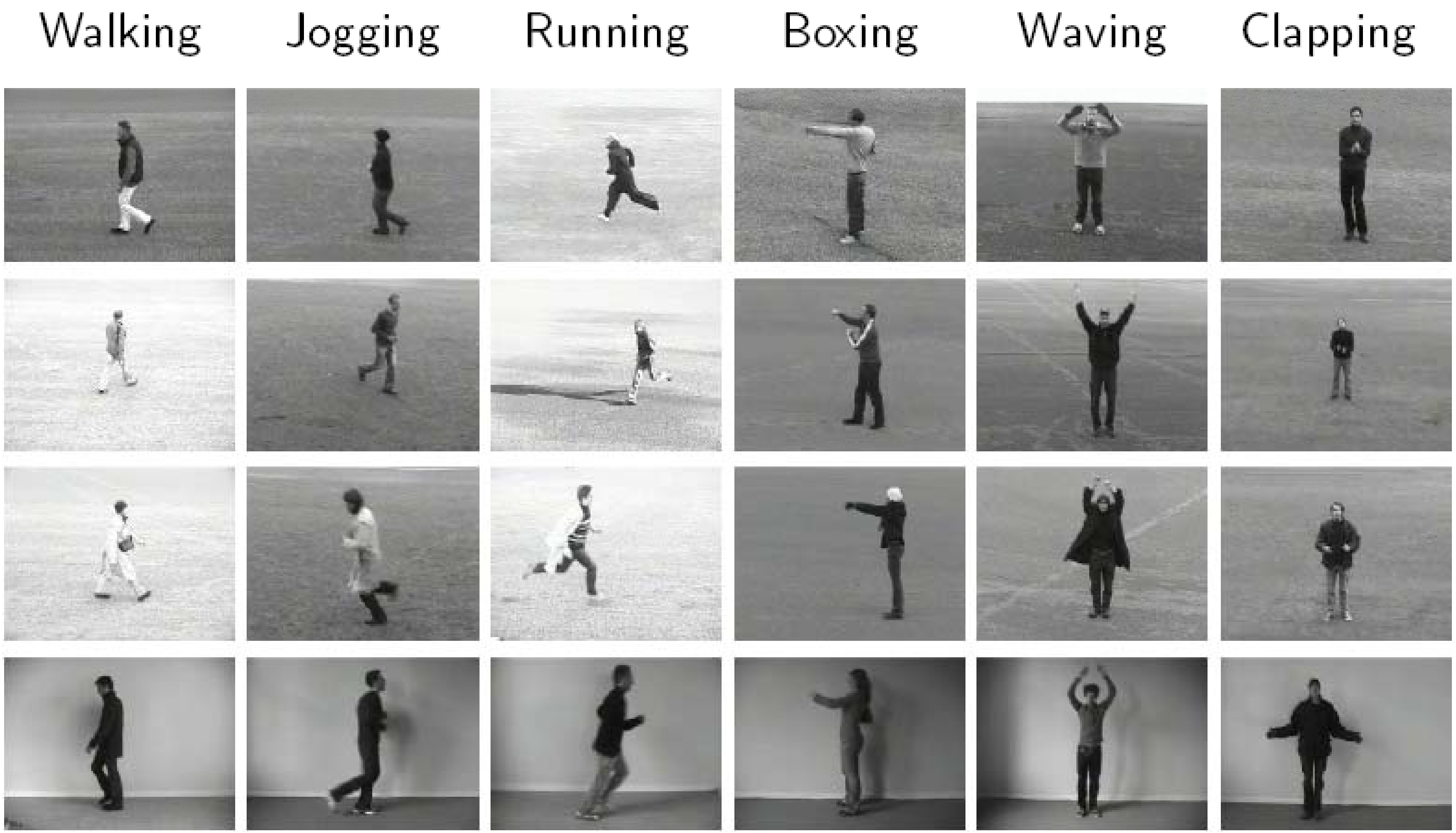
Spatio-temporal bag-of-features

Use global spatio-temporal grids

- In the spatial domain:
  - 1x1 (standard BoF)
  - 2x2, o2x2 (50% overlap)
  - h3x1 (horizontal), v1x3 (vertical)
  - 3x3
- In the temporal domain:
  - t1 (standard BoF), t2, t3

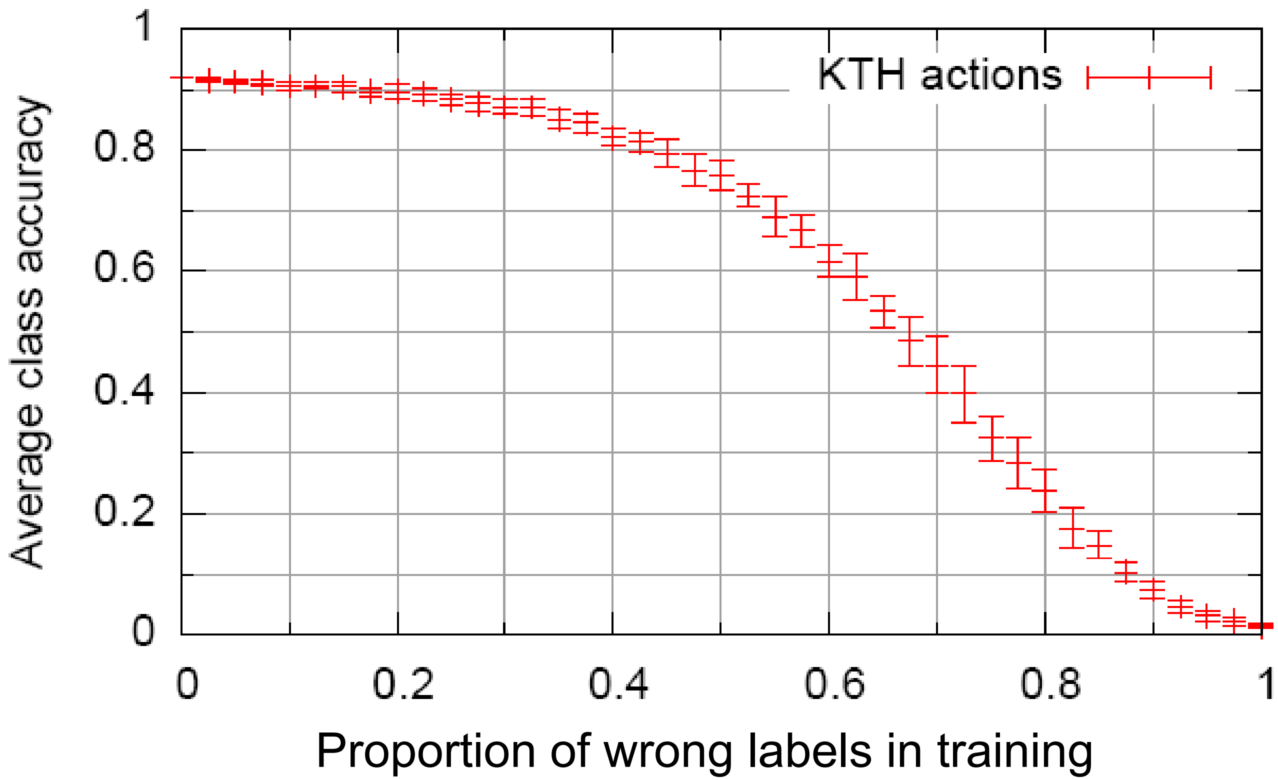


KTH actions dataset



Sample frames from KTH action dataset for six classes (columns) and four scenarios (rows)

Robustness to noise in training



- Up to p=0.2 the performance decreases insignificantly
- At p=0.4 the performance decreases by around 10%

Action recognition in movies

- Real data is hard!
- False Positives (FP) and True Positives (TP) often visually similar
- False Negatives (FN) are often particularly difficult

Results on Hollywood-2 dataset

SetUp	Clean Training		Automatic Training		Chance
Channel	Combination	BoF	Combination	BoF	
mAP	**50.7**	47.3	**34.6**	30.8	9.2
AnswerPhone	20.9	15.7	19.1	17.7	7.2
DriveCar	84.6	86.6	79.1	75.8	11.5
Eat	67.0	59.5	23.5	15.0	3.7
FightPerson	69.8	71.1	59.0	56.3	7.9
GetOutCar	45.7	29.3	25.7	12.3	6.4
HandShake	27.8	21.2	15.2	12.4	5.1
HugPerson	43.2	35.8	14.6	15.6	7.5
Kiss	52.5	51.5	44.4	40.8	11.7
Run	67.8	69.1	50.7	52.6	16.0
SitDown	57.6	58.2	31.4	25.8	12.2
SitUp	17.2	17.5	8.5	8.8	4.2
StandUp	54.3	51.7	44.1	36.8	16.5

Class Average Precision (AP) and mean AP for

- Clean training set
- Automatic training set (with noisy labels)
- Random performance

Action classification

Test episodes from movies “The Graduate”, “It’s a Wonderful Life”, “Indiana Jones and the Last Crusade” [Laptev et al. CVPR 2008]

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

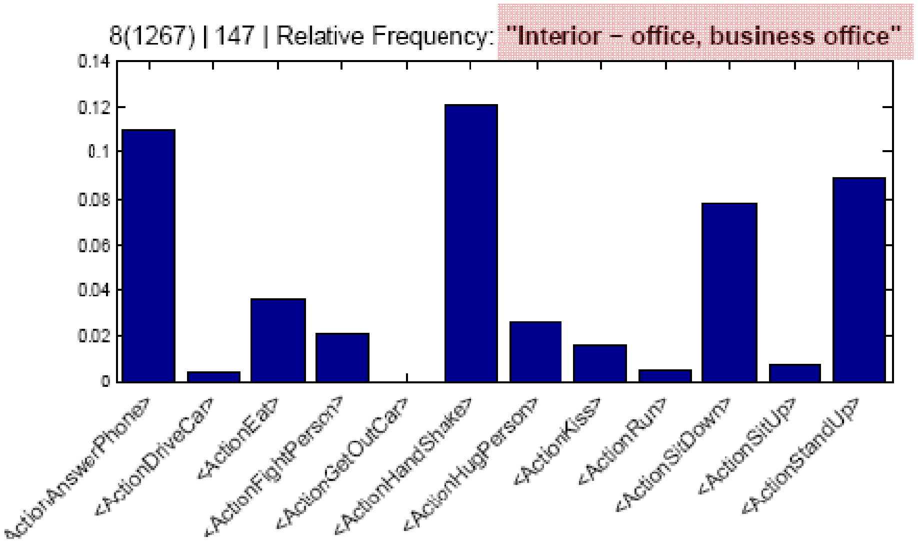
INT. TRENDY RESTAURANT - NIGHT
INT. MARSELLUS WALLACE’S DINING ROOM MORNING
EXT. STREETS BY DORA’S HOUSE - DAY.
INT. MELVIN'S APARTMENT, BATHROOM – NIGHT
EXT. NEW YORK CITY STREET NEAR CAROL'S RESTAURANT – DAY
INT. CRAIG AND LOTTE'S BATHROOM - DAY

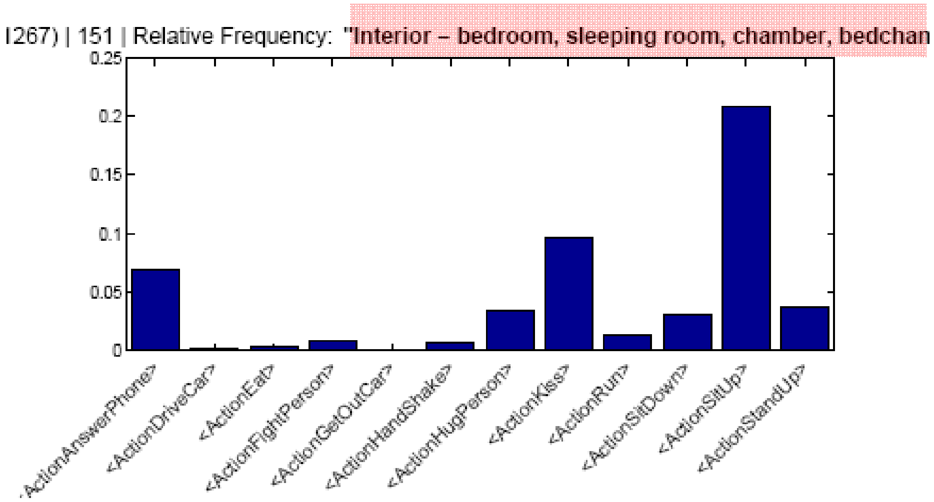
- Maximize word frequency ⇨ street, living room, bedroom, car ….
- Merge words with similar senses using WordNet:

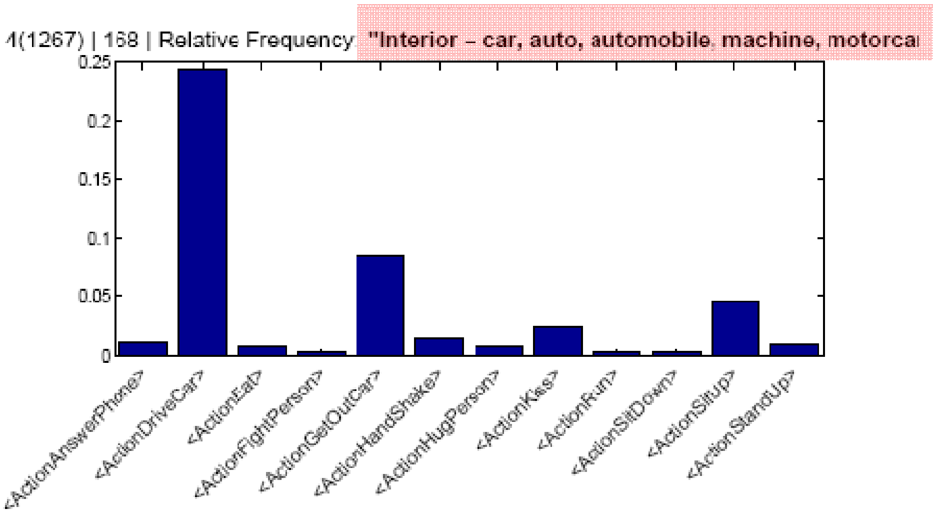
  taxi -> car, cafe -> restaurant

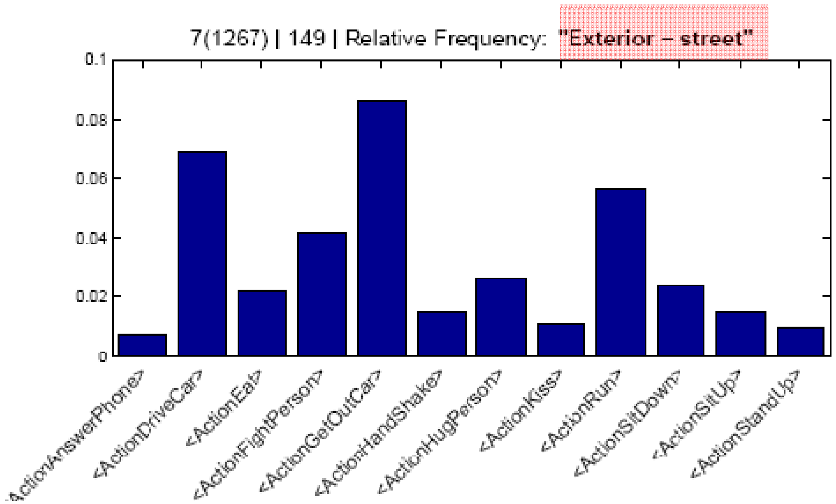
- Measure correlation of words with actions (in scripts) and
- Re-sort words by the entropy $S = -k \sum P_i \ln P_i$ for P = p(action | word)

Co-occurrence of actions and scenes in scripts

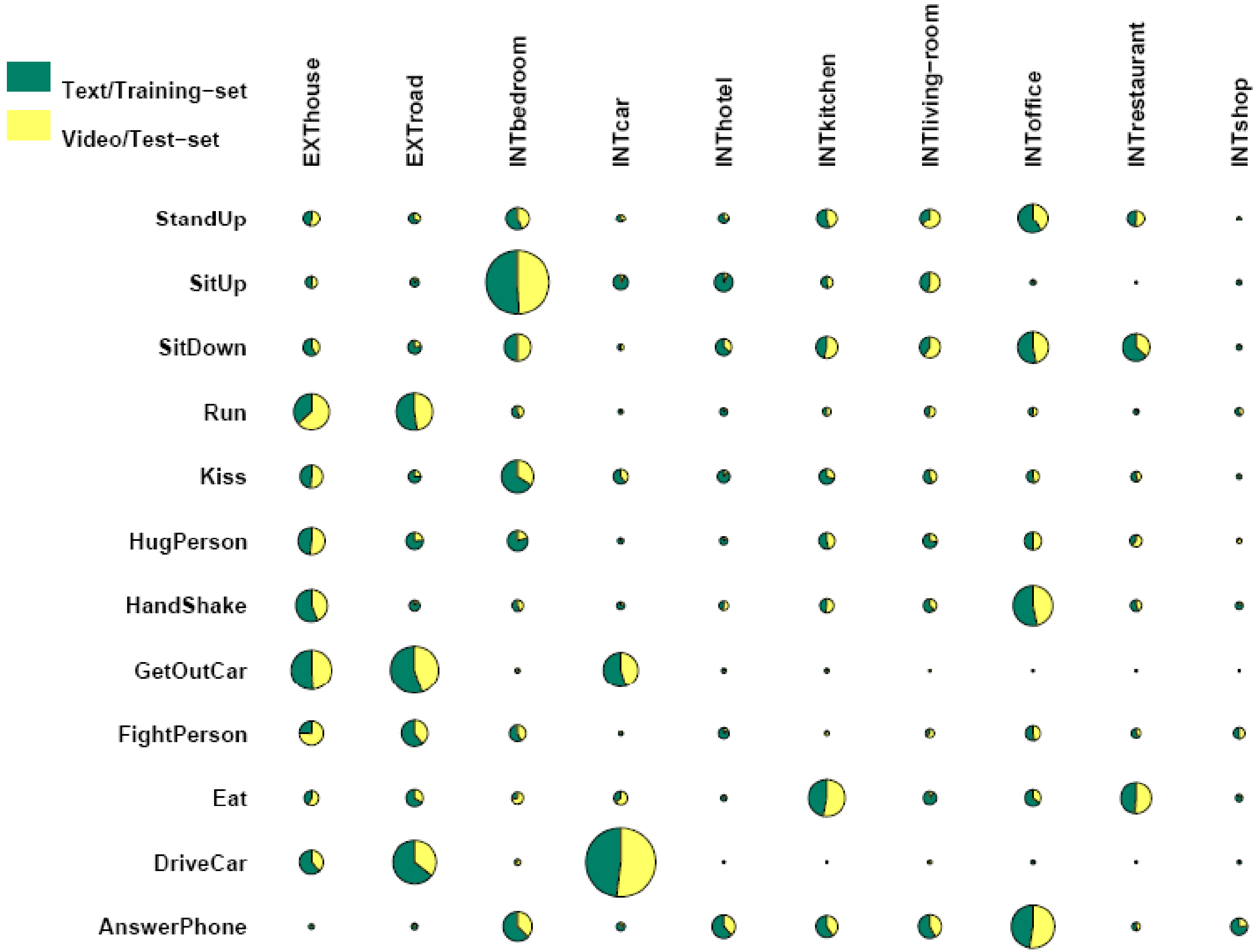








Co-occurrence of actions and scenes in text vs. video



Automatic gathering of relevant scene classes and visual samples

Source:
69 movies
aligned with
the scripts

	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

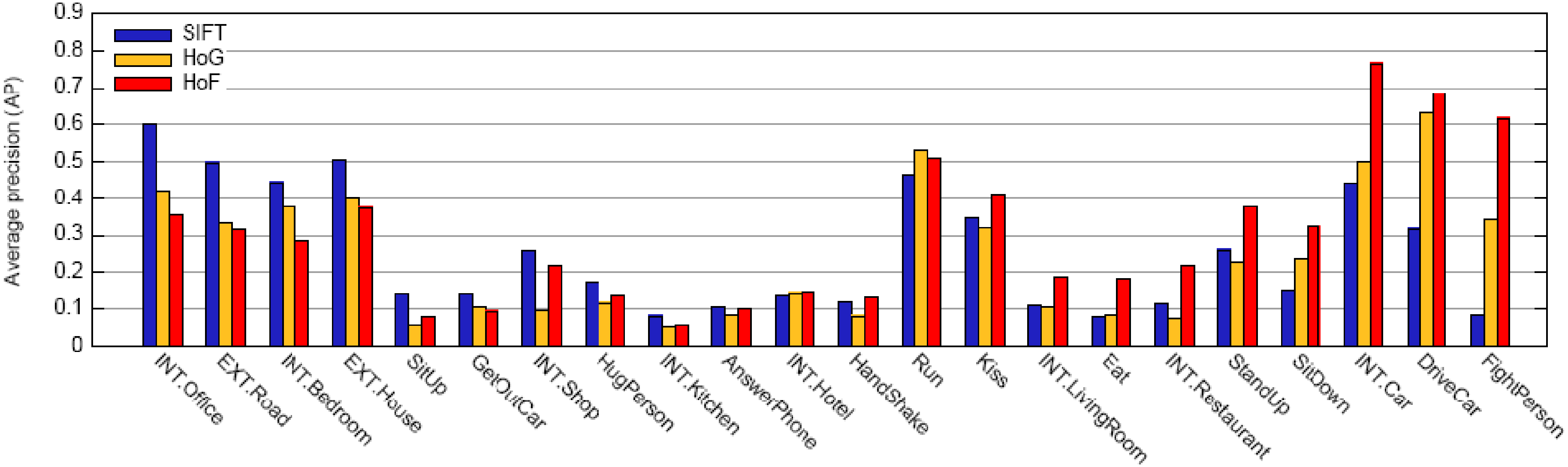
	Auto-Train-Scenes	Clean-Test-Scenes
EXT-house	81	140
EXT-road	81	114
INT-bedroom	67	69
INT-car	44	68
INT-hotel	59	37
INT-kitchen	38	24
INT-living-room	30	51
INT-office	114	110
INT-restaurant	44	36
INT-shop	47	28
All Samples	570	582

(b) Scenes

Hollywood-2 dataset is on-line:

http://www.irisa.fr/vista/actions/hollywood2

Results: actions and scenes (separately)



Actions

	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*

Scenes

	SIFT	HoG HoF	SIFT HoG HoF
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*

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

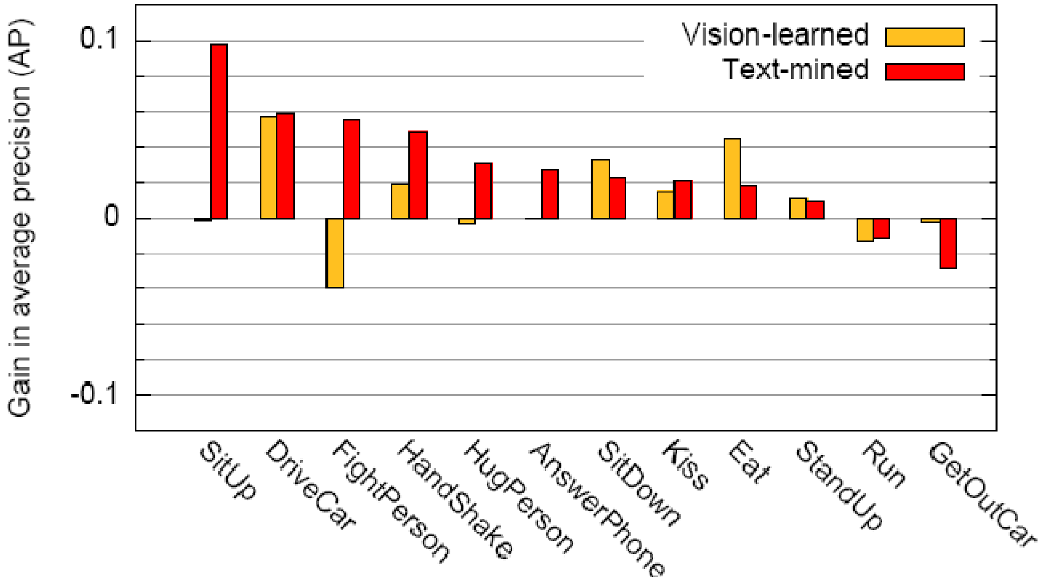
$s_j(\boldsymbol{x})$ Scene classification score

w_{ij} Weight, estimated from text: $p(Scene|Action)$

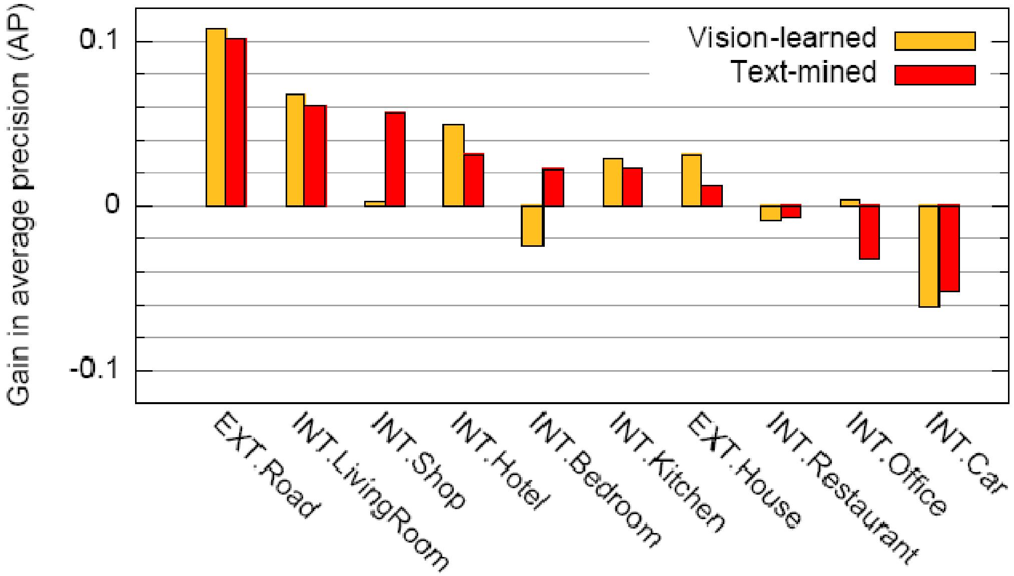
$a'_i(\boldsymbol{x})$ New action score

Results: actions and scenes (jointly)

Actions
in the
context
of
Scenes



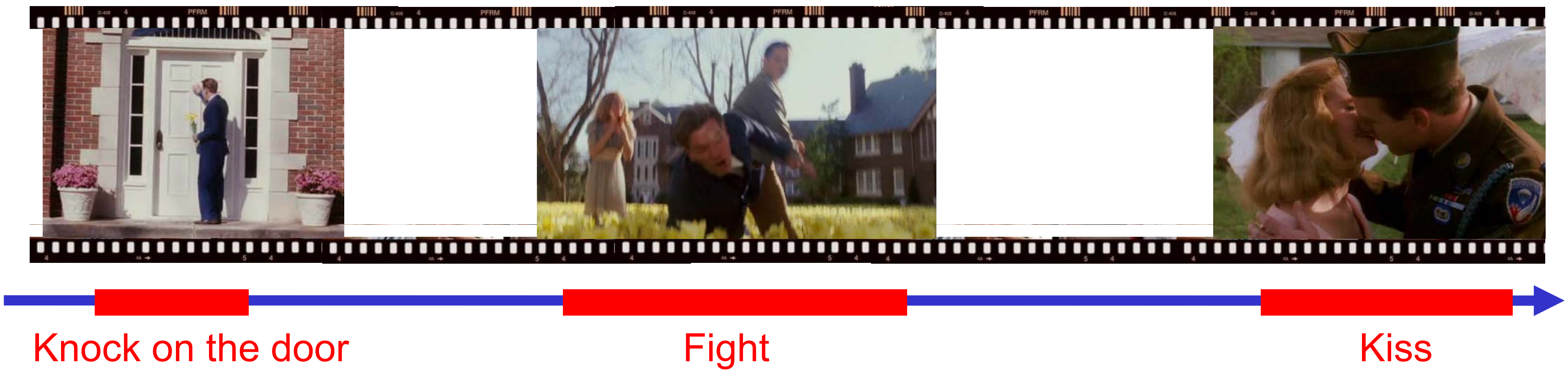
Scenes
in the
context
of
Actions



Weakly-Supervised Temporal Action Annotation

[Duchenne at al. ICCV 2009]

- 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

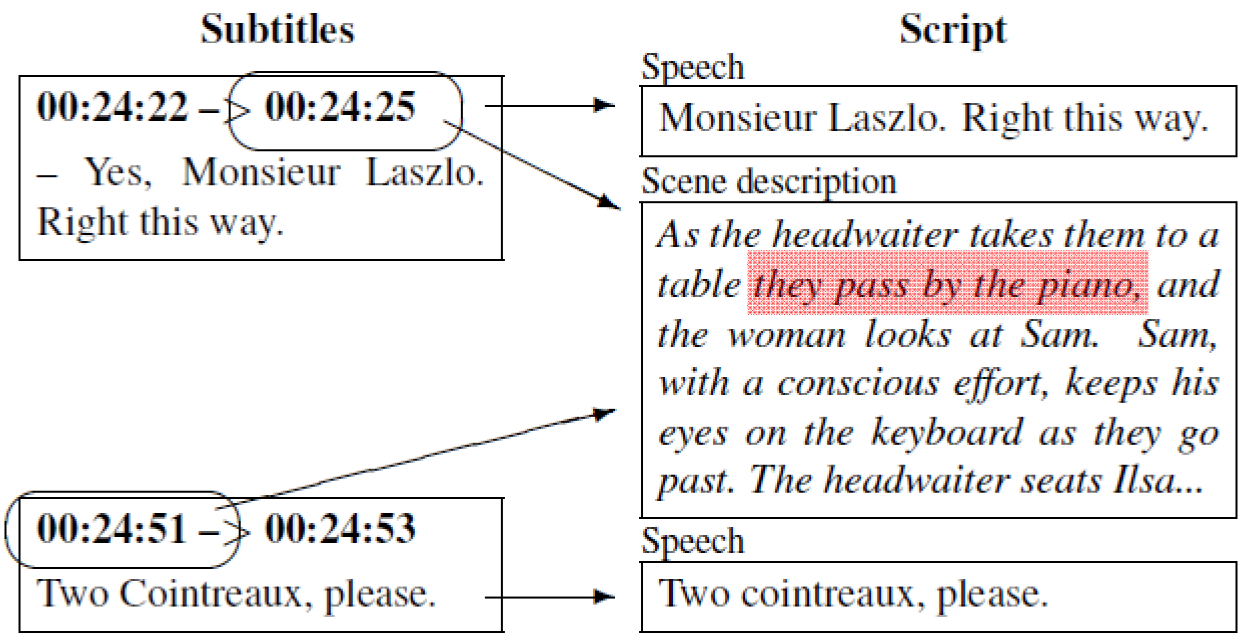
  -- Search action patterns

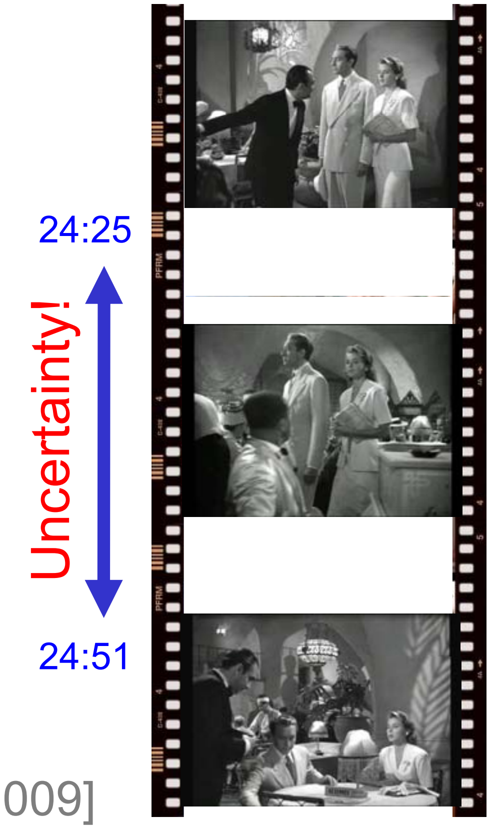
Person+Verb	Person+Verb+Prep.	Person+Verb+Prep+Vis.Noun
3725 /PERSON .* is	989 /PERSON .* looks .* at	41 /PERSON .* sits .* in .* chair
2644 /PERSON .* looks	384 /PERSON .* is .* in	37 /PERSON .* sits .* at .* table
1300 /PERSON .* turns	363 /PERSON .* looks .* up	31 /PERSON .* sits .* on .* bed
916 /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





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

Overview

Input:

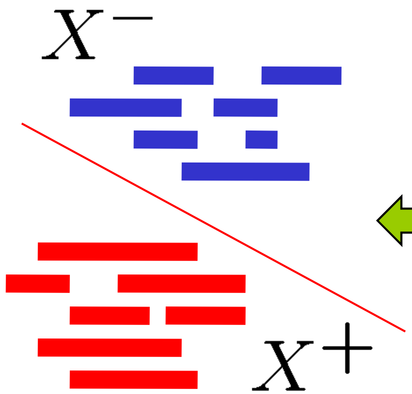
- Action type, e.g.
  Person Opens Door
- Videos + aligned scripts

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



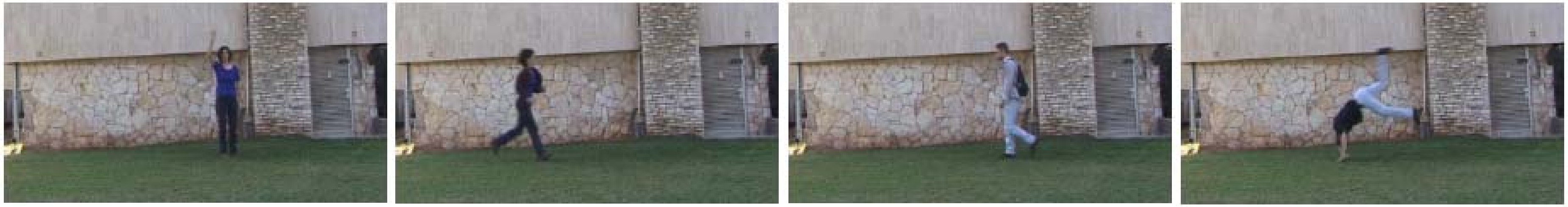
Output:

Sliding-window-style temporal action localization

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

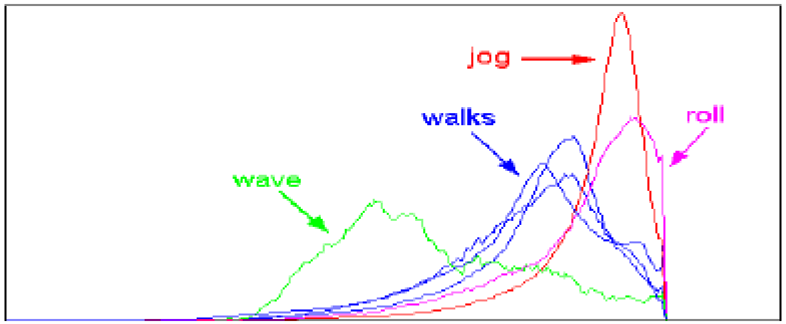
Action clustering

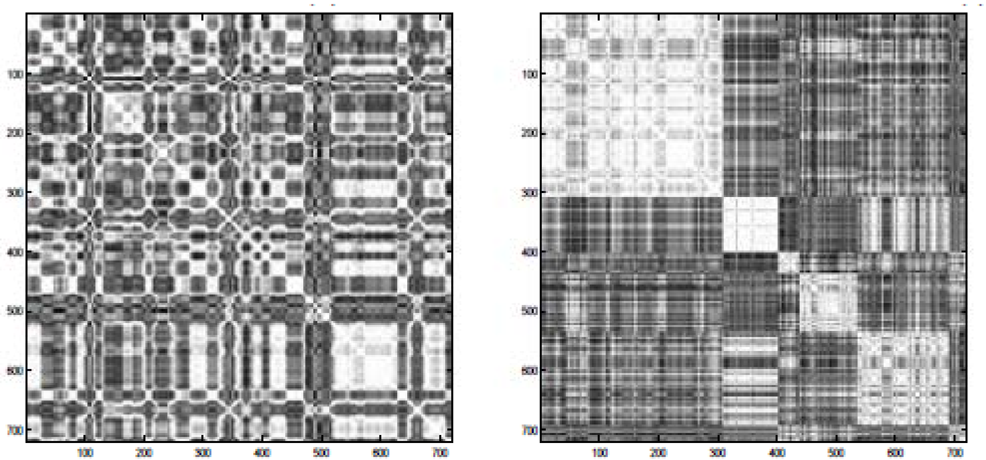
[Lihi Zelnik-Manor and Michal Irani CVPR 2001]



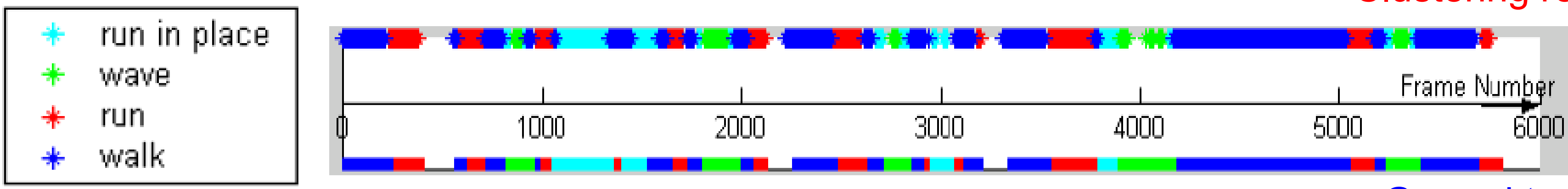
Spectral clustering

Descriptor space





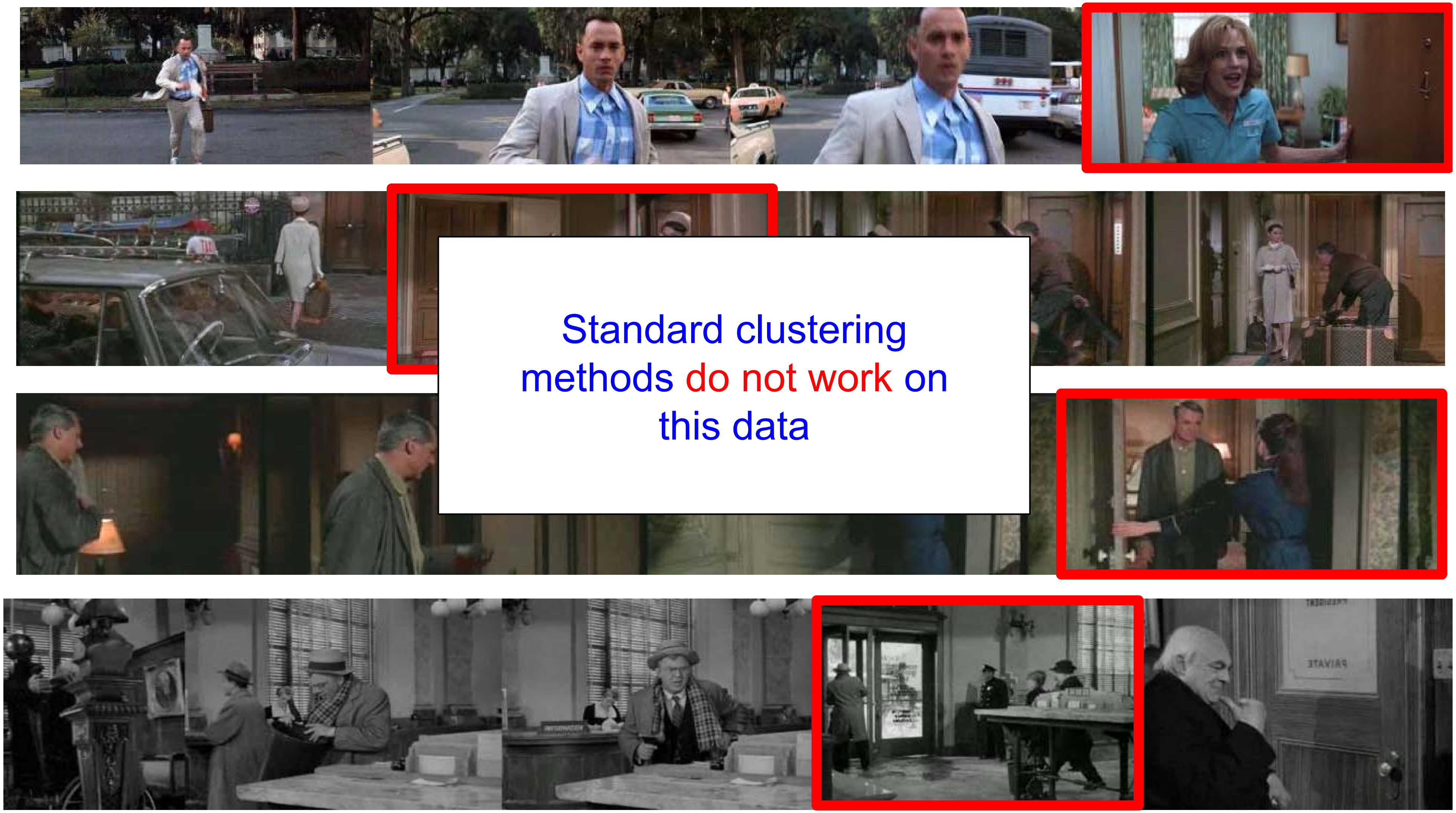
Clustering results



Ground truth

Action clustering

Our data:

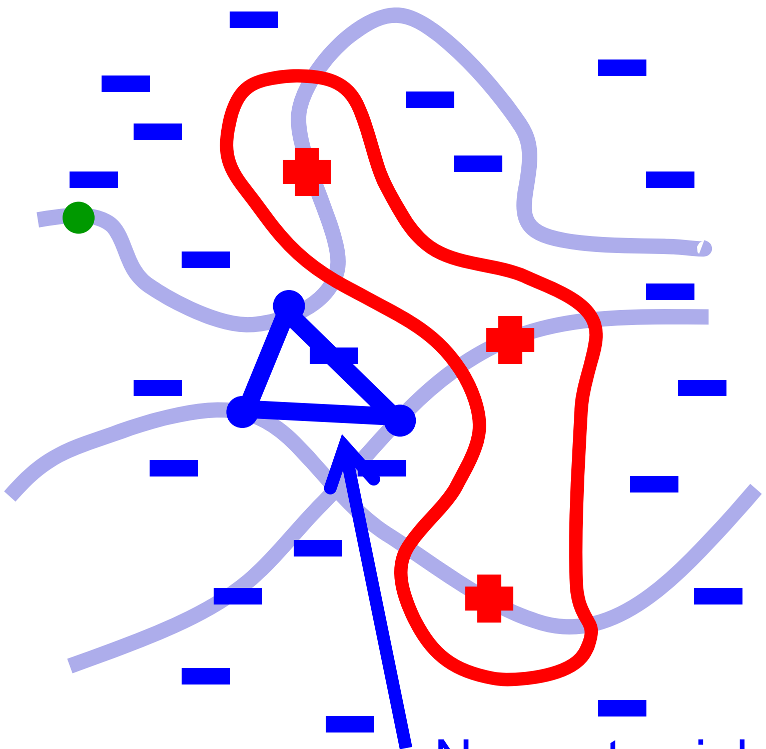


Standard clustering methods do not work on this data

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

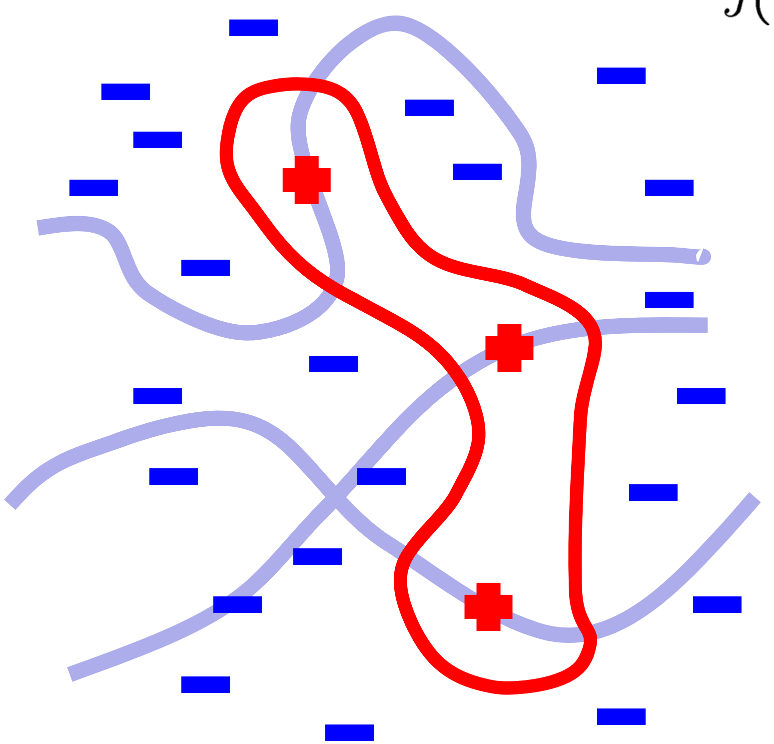
[Duchenne, Laptev, Sivic, Bach, Ponce, ICCV 2009]

Action clustering

Formulation

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

Feature space



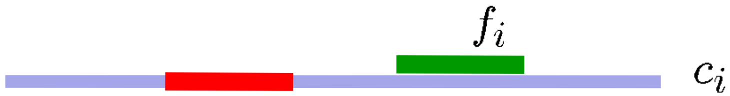
discriminative cost

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

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

x_i^- negative samples

$c_i[f_i]$ parameterized positive samples



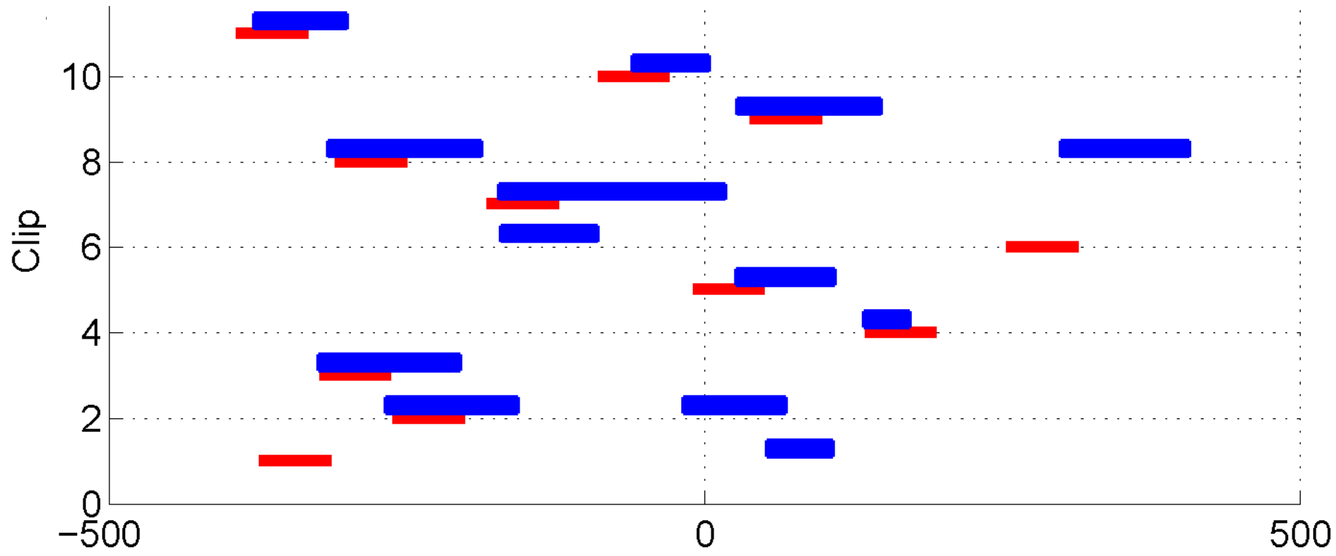
Optimization

SVM solution for w, b
Coordinate descent on f_i

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

Clustering results

Drinking actions in Coffee and Cigarettes

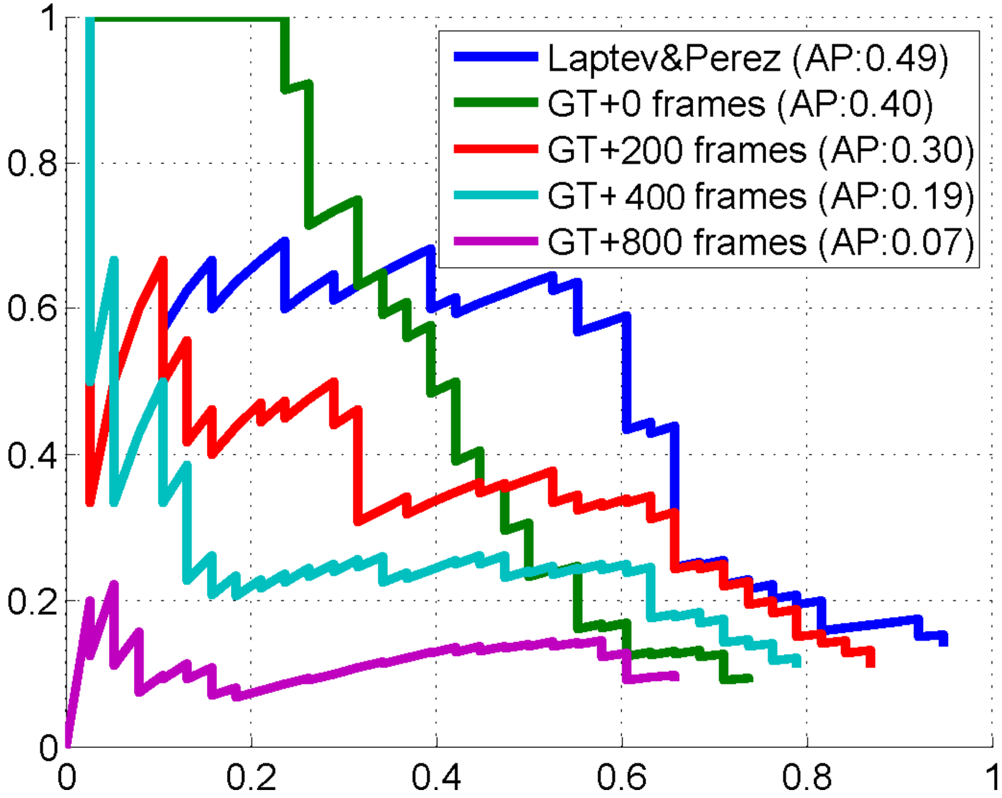


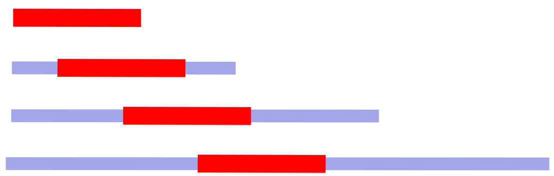
Detection results

Drinking actions in Coffee and Cigarettes

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

Detection trained on *simulated clusters*





Test set:

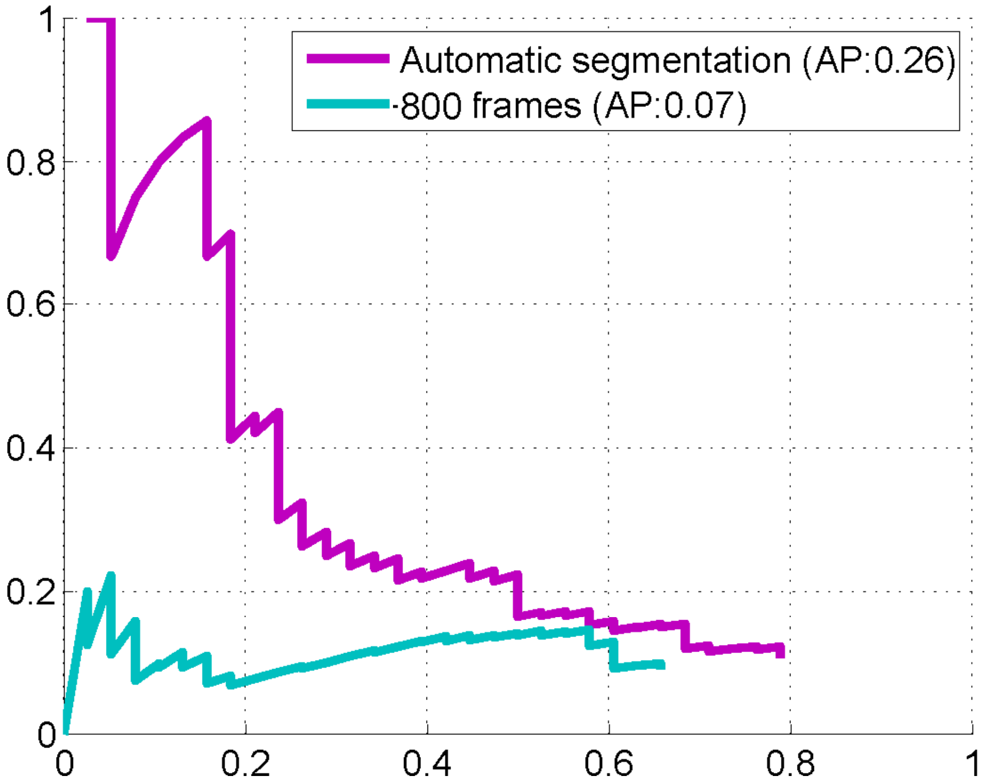
- 25min from “Coffee and Cigarettes” with GT 38 drinking actions

Detection results

Drinking actions in Coffee and Cigarettes

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

Detection trained on *automatic clusters*

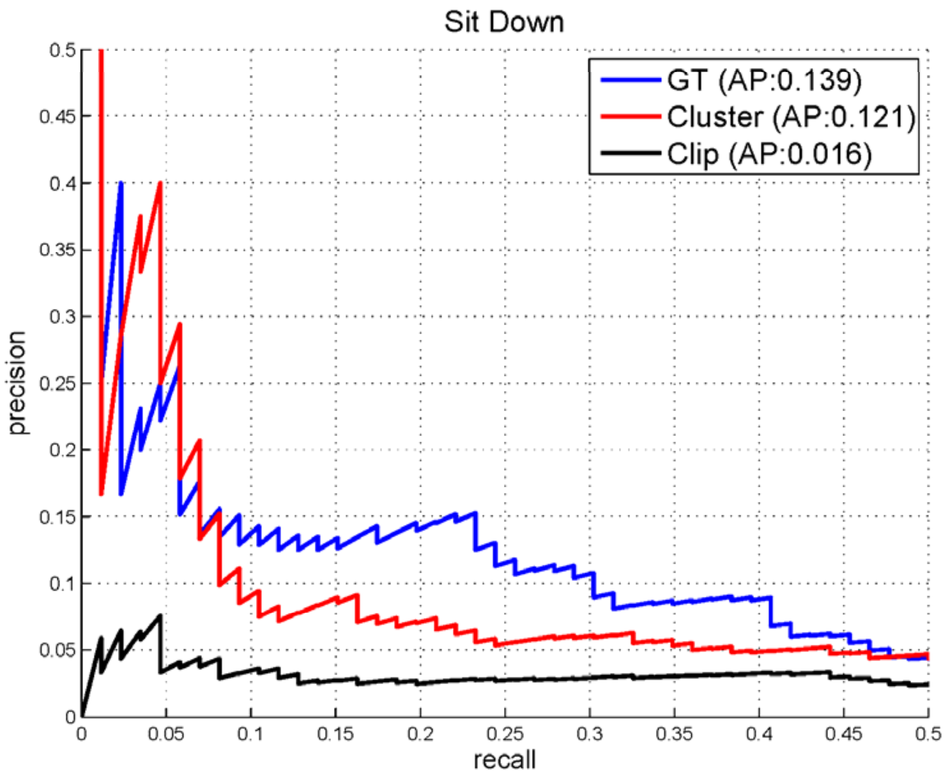


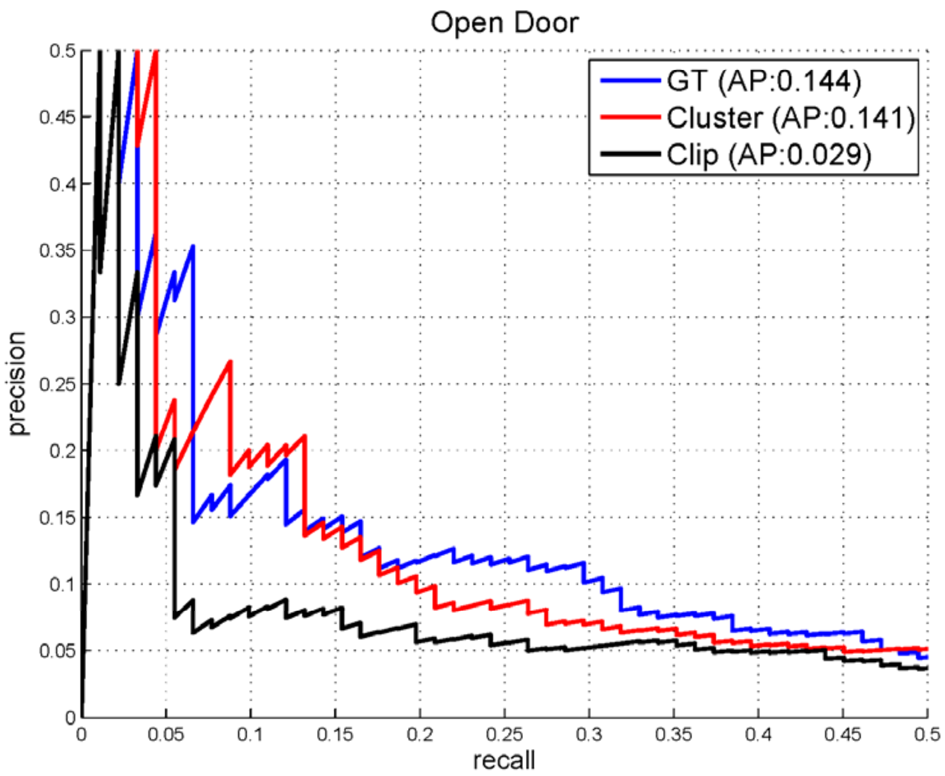
Test set:

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

Detection results

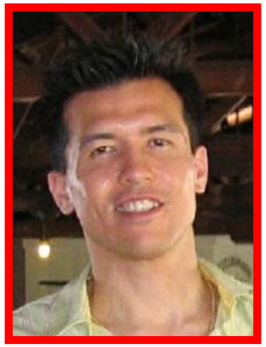
“Sit Down” and “Open Door” actions in ~5 hours of movies

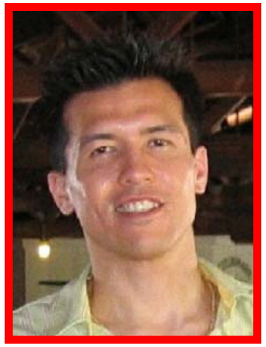




Temporal detection of “Sit Down” and “Open Door” actions in movies:
The Graduate, The Crying Game, Living in Oblivion [Duchenne et al. 09]

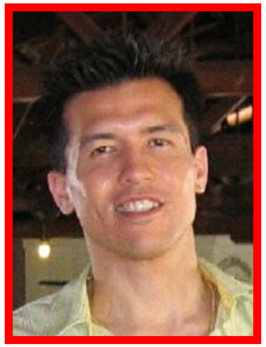
Course overview

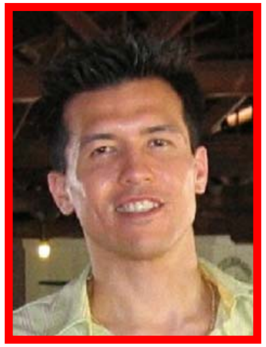




- Definitions
- Benchmark datasets
- Early silhouette and tracking-based methods
- Motion-based similarity measures
- Template-based methods
- Local space-time features
- Bag-of-Features action recognition
- Weakly-supervised methods
- Pose estimation and action recognition
- Action recognition in still images
- Human interactions and dynamic scene models
- Conclusions and future directions

Course overview





- Definitions
- Benchmark datasets
- Early silhouette and tracking-based methods
- Motion-based similarity measures
- Template-based methods
- Local space-time features
- Bag-of-Features action recognition
- Weakly-supervised methods
- Pose estimation and action recognition
- Action recognition in still images
- Human interactions and dynamic scene models
- Conclusions and future directions