Object recognition and computer vision 2009/2010

Lecture 11, December 15

Motion and Human Actions

Ivan Laptev

ivan.laptev@ens.fr

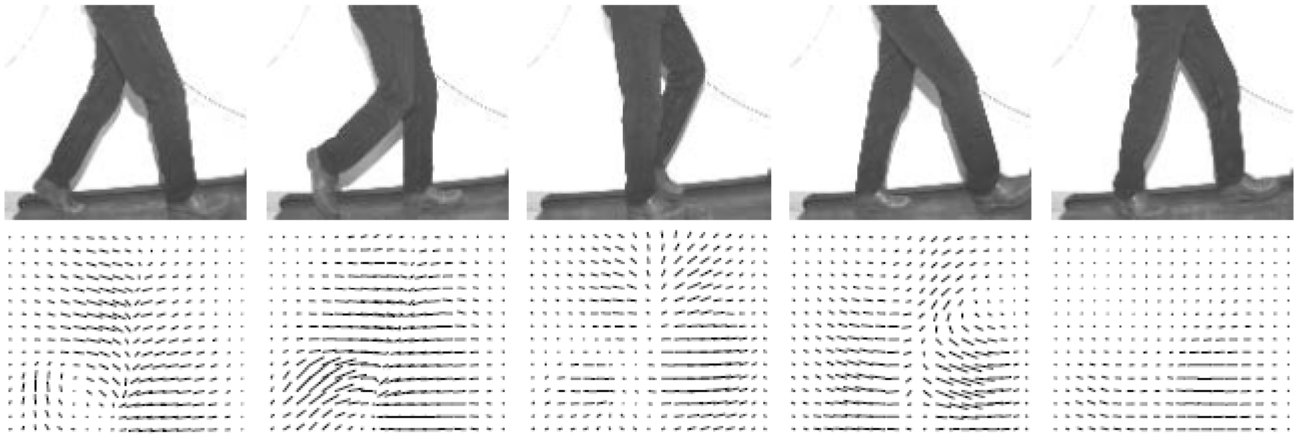
Equipe-projet WILLOW, ENS/INRIA/CNRS UMR 8548

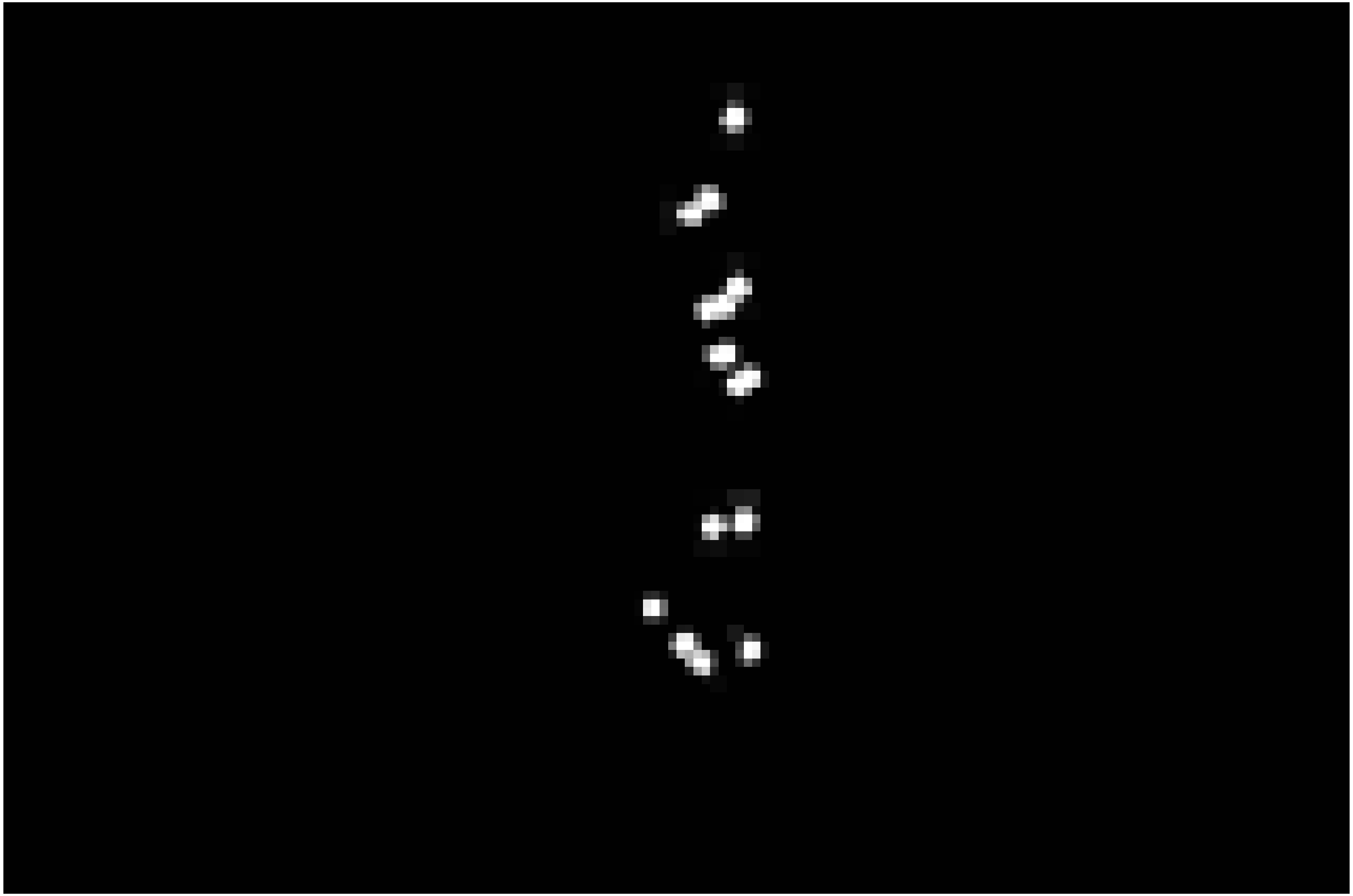
Laboratoire d'Informatique, Ecole Normale Supérieure, Paris

Shape versus Motion

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

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





Gunnar Johansson, **Moving Light Displays,** 1973

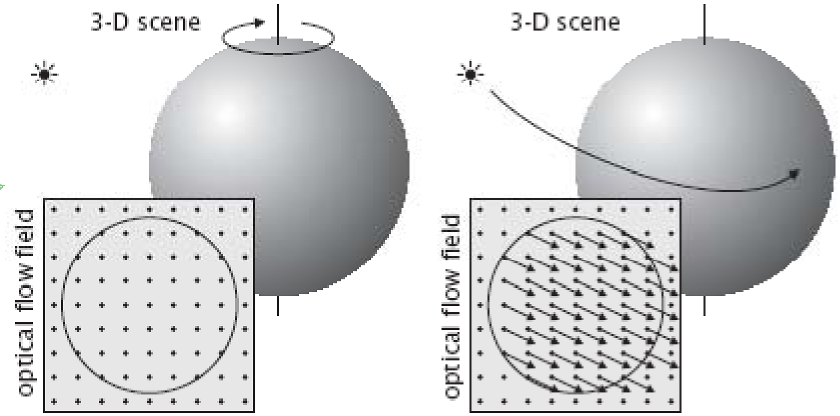
Generic Optical Flow

- Classic problem of computer vision [Gibson 1955]
- Goal: estimate motion field
  How? We only have access to image pixels
  ⇨ Estimate pixel-wise correspondence between frames = Optical Flow
- Brightness Change assumption: corresponding pixels preserve their intensity (color)

  Useful assumption in many cases

  Breaks at occlusions and illumination changes

  Physical and visual motion may be different



Generic Optical Flow

- Brightness Change Constraint Equation (BCCE)

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

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

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

One equation, two unknowns => cannot be solved directly

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

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

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

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

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

Generic Optical Flow

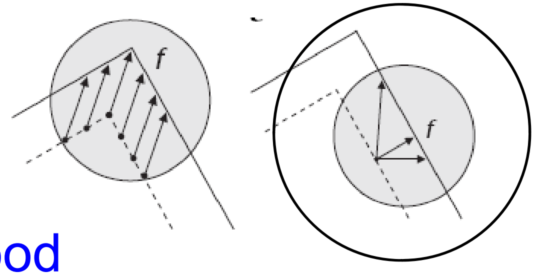
- The solution of $<\nabla I (\nabla I)^\top> \mathbf{v} = -<\nabla I I_t>$ assumes

  1. Brightness change constraint holds in $<\cdot>$
  2. Sufficient variation of image gradient in $<\cdot>$
  3. Approximately constant motion in $<\cdot>$

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

- Solutions:

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

  ⇨ Increase integration neighborhood

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

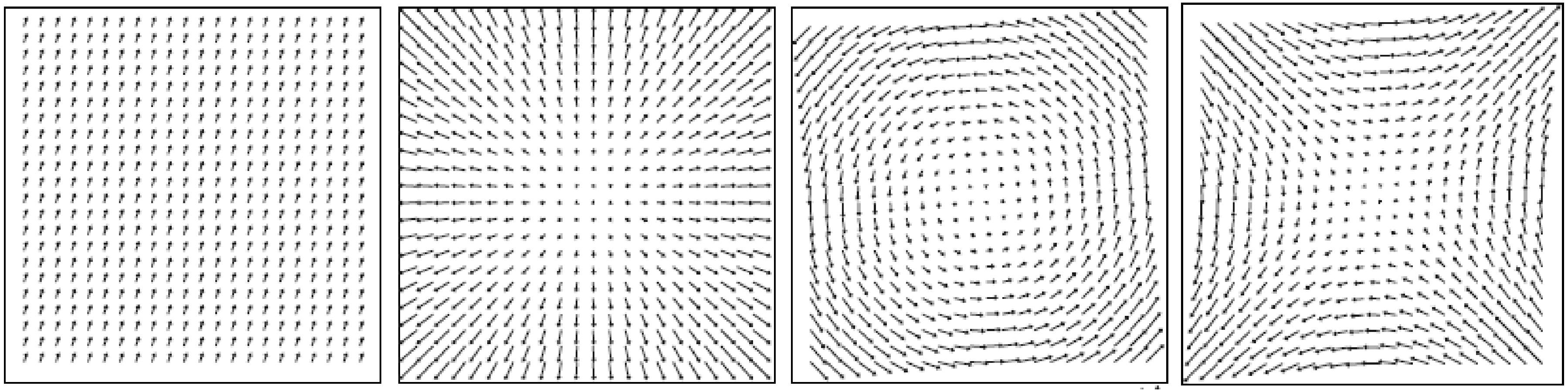
  ⇨ Use more sophisticated motion model

Parameterized Optical Flow

- Constant velocity model: $\mathbf{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix}$
- Upgrade to affine motion model: $\mathbf{v} = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} v_x \\ v_y \end{pmatrix}$

  Now motion depends on the position $(x, y)^\top$ inside the neighborhood

Examples of Affine motion models for different parameters:



- Can be formulated as Least Squares approach to estimate $\mathbf{v}$ as before!

Parameterized Optical Flow

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

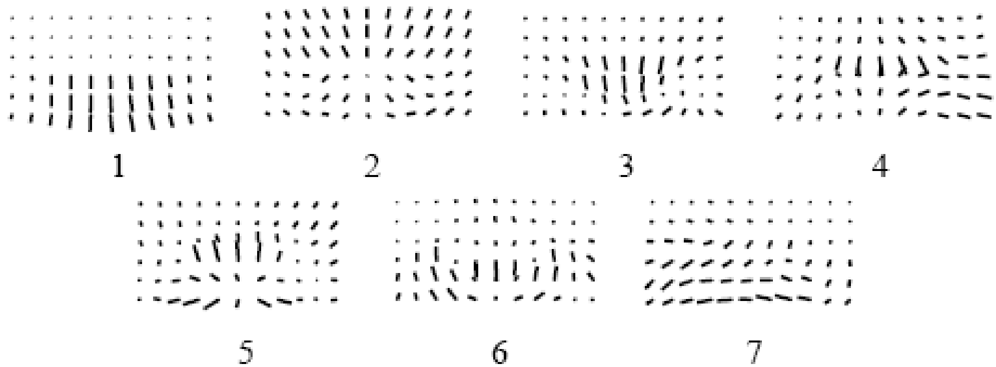
  1. Compute standard Optical Flow for many examples
  2. Put velocity components into one vector

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

  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

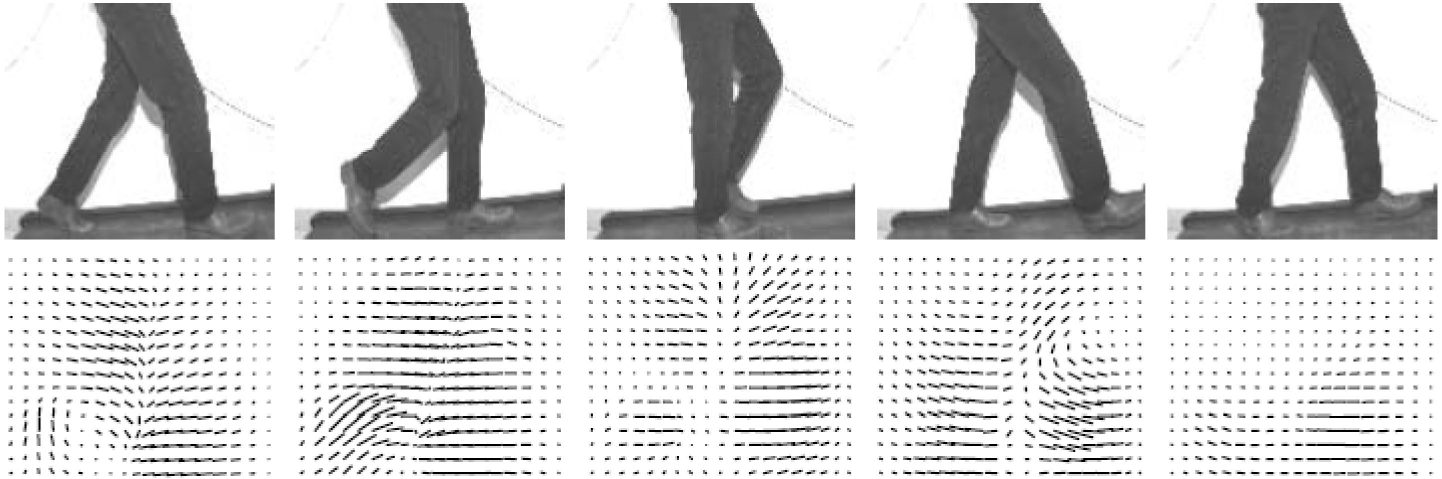
M.J. Black, Y. Yacoob, A.D. Jepson and D.J. Fleet, **CVPR 1997**

Parameterized Optical Flow

- Use PCA flow bases to *regularize* solution of motion estimation
- Motion estimation for test samples can be computed *without* explicit computation of optical flow!

Solution formulation e.g. in terms of Least Squares

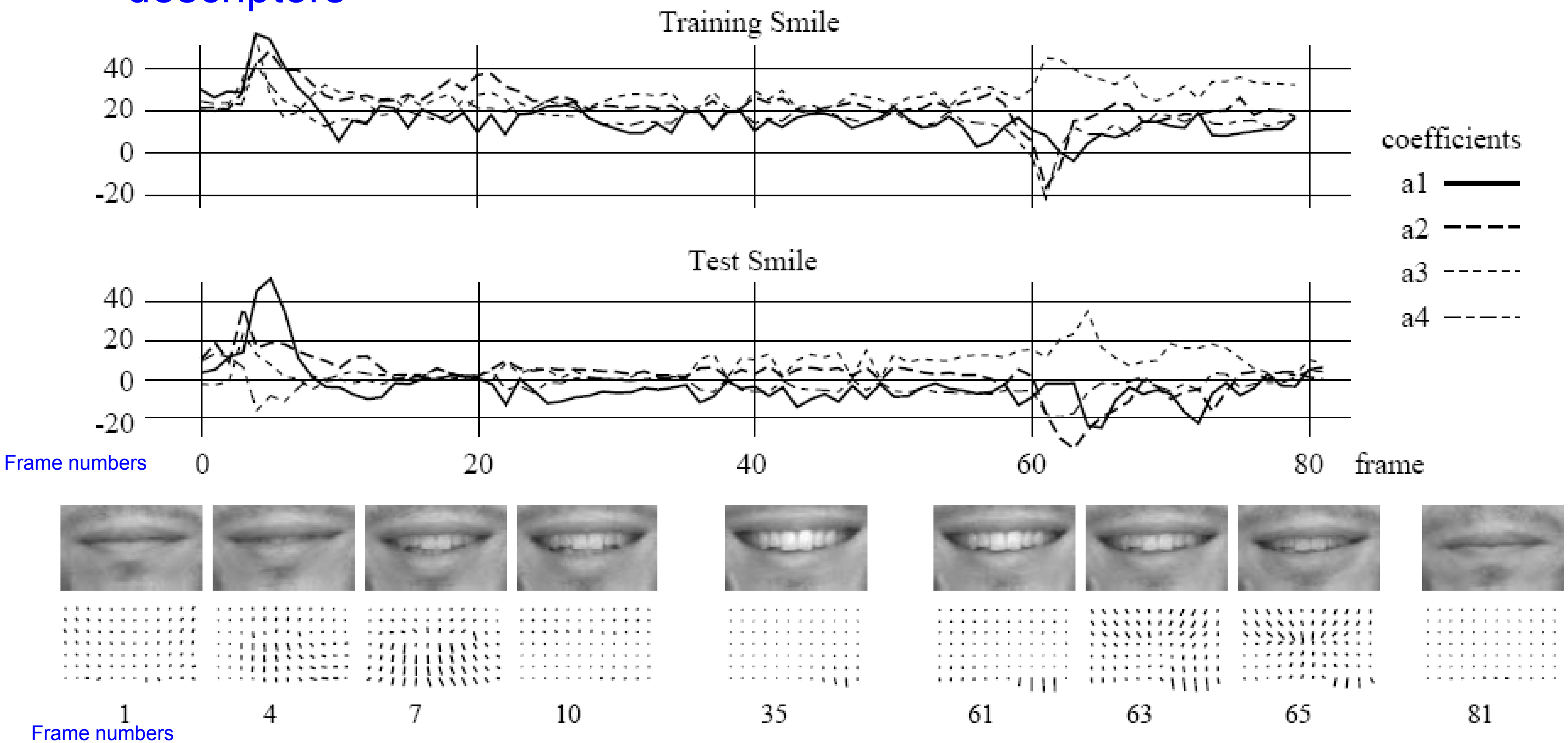
Direct flow recovery:



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

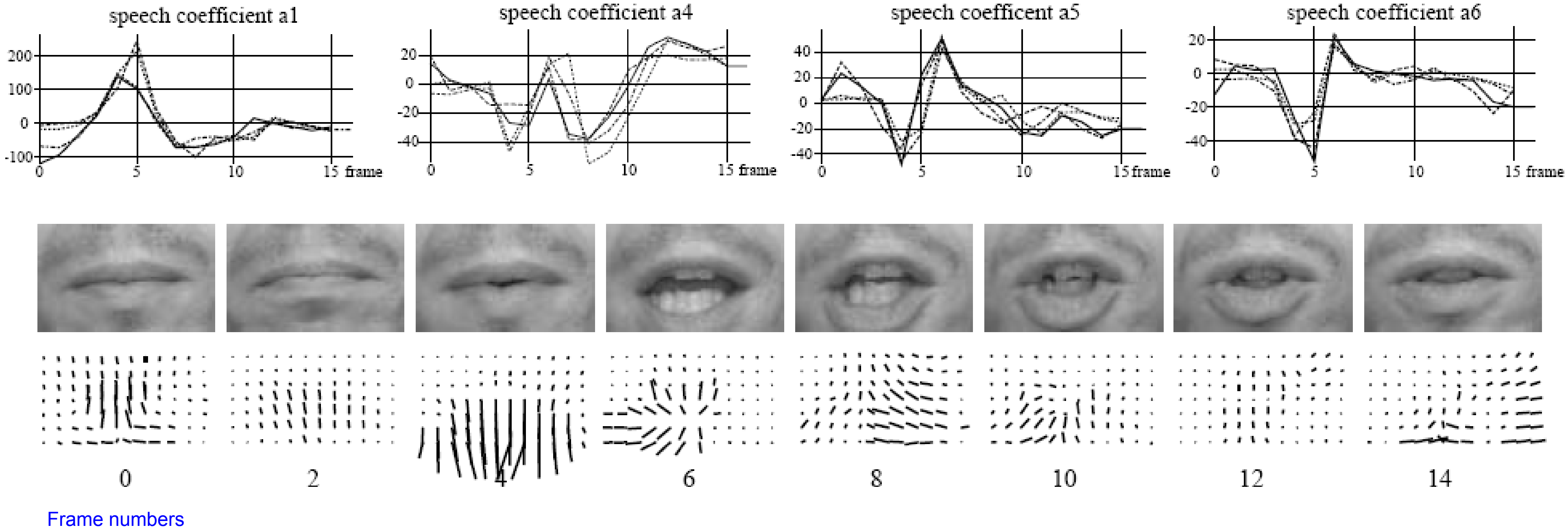


Learning Parameterized Models of Image Motion

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



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

Spatial Motion Descriptor

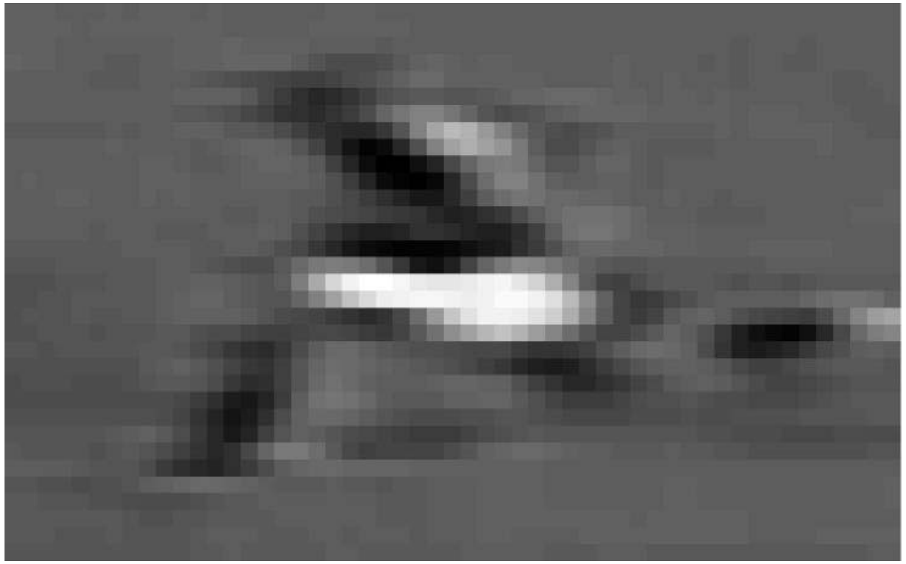
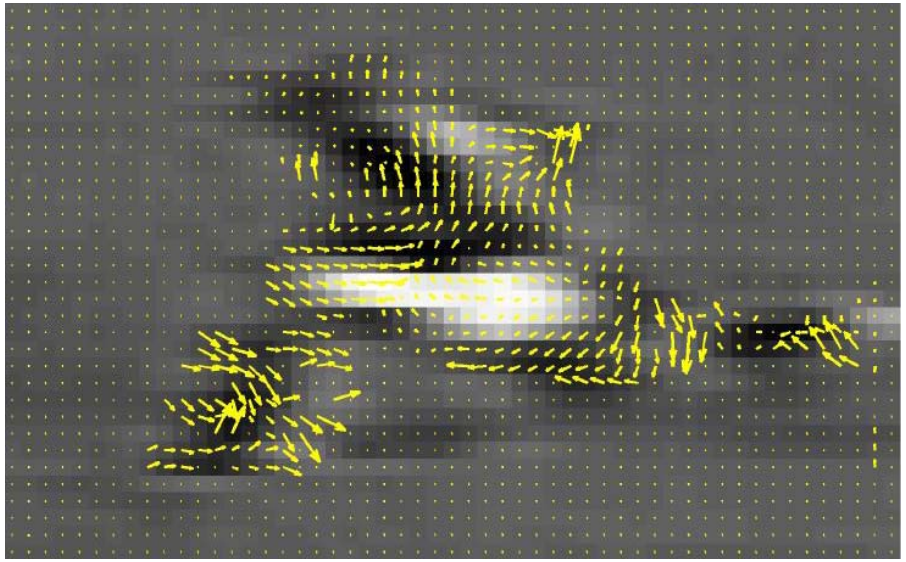
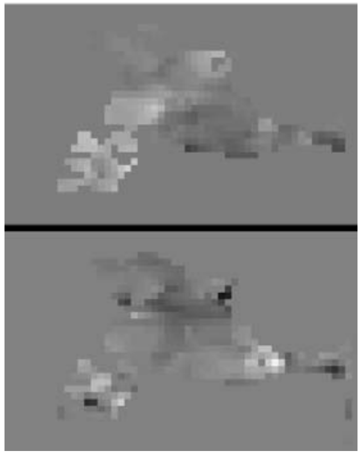


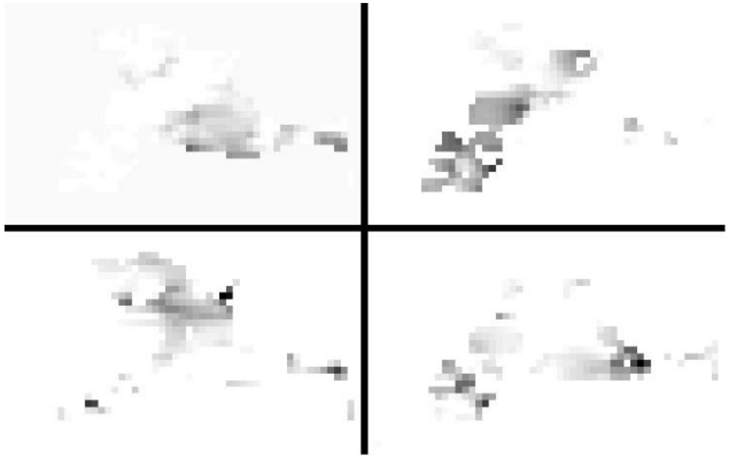
Image frame



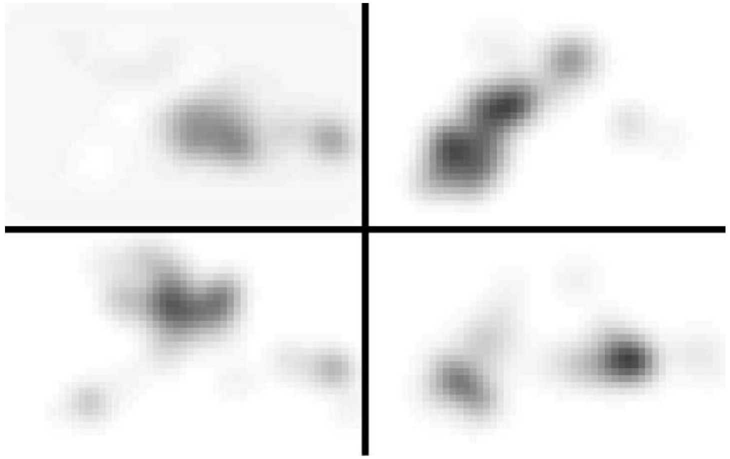
Optical flow $F_{x,y}$



F_x, F_y

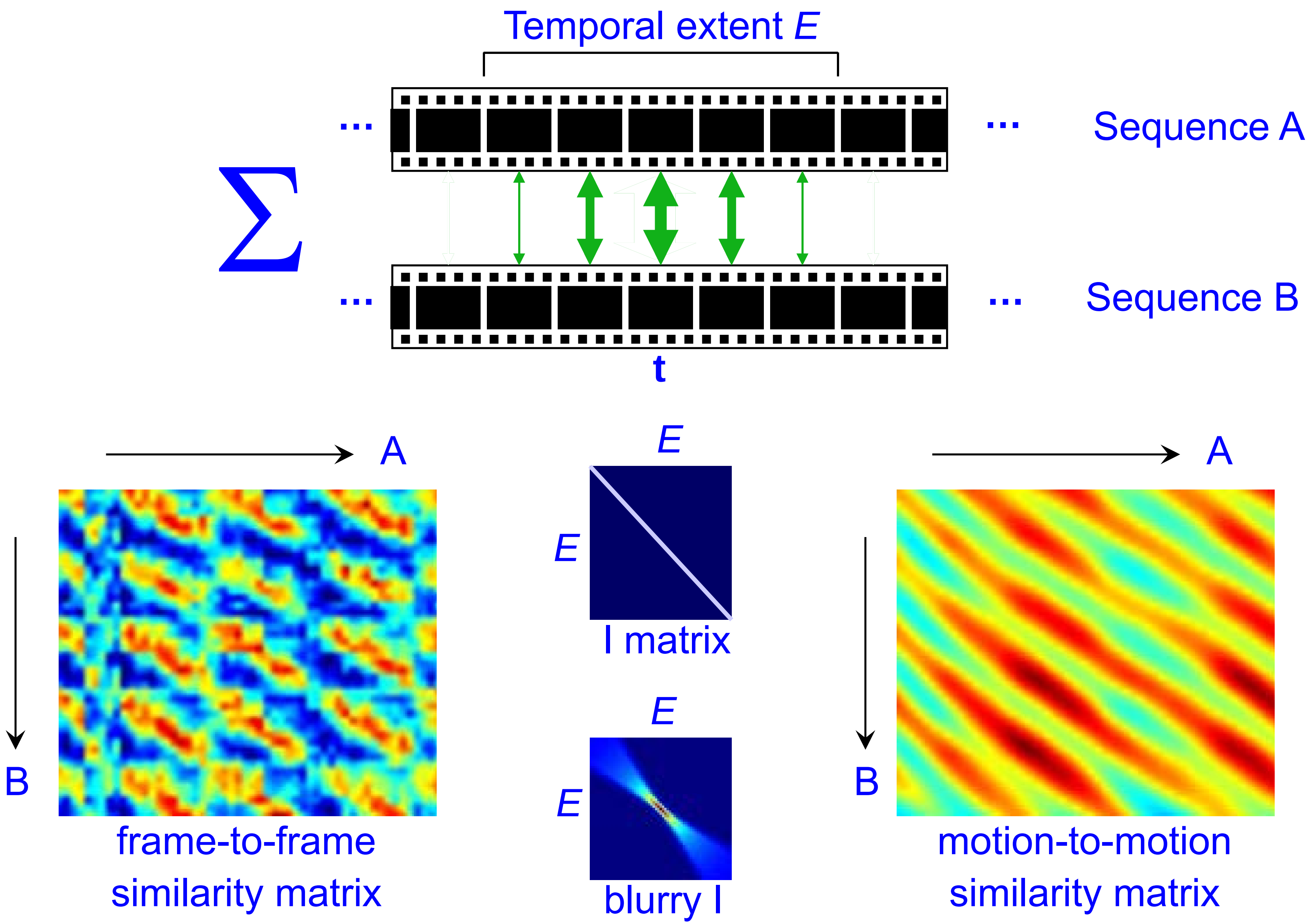


$F_x^-, F_x^+, F_y^-, F_y^+$



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

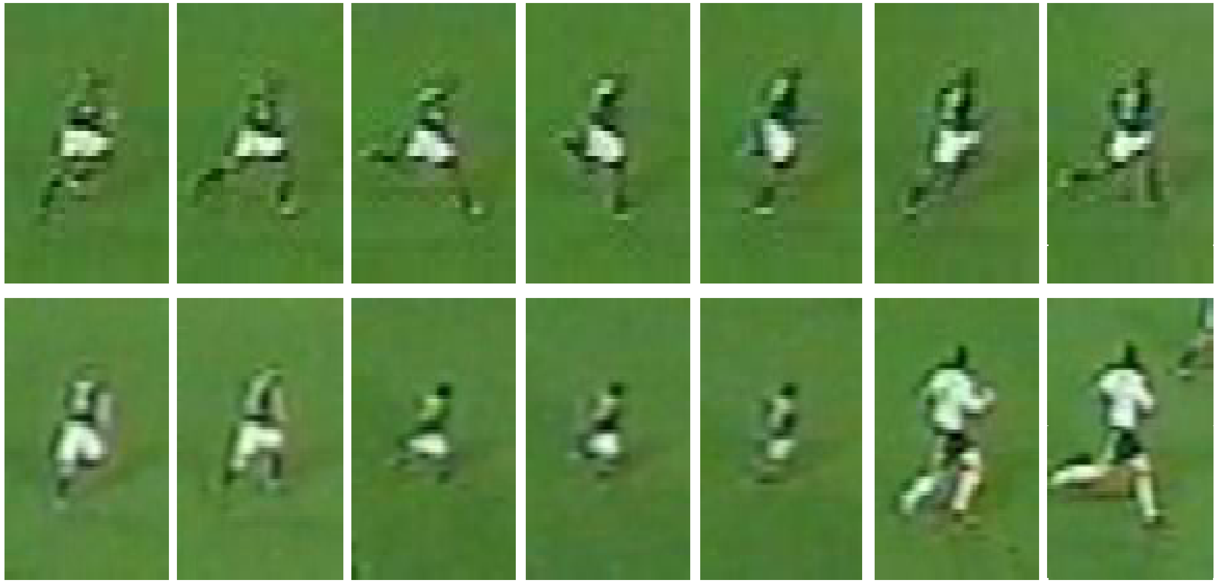
Spatio-Temporal Motion Descriptor

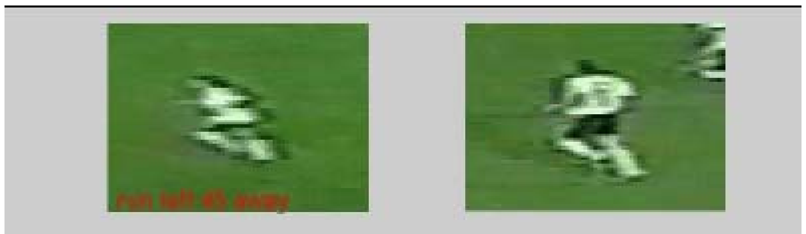


Football Actions: matching

Input Sequence

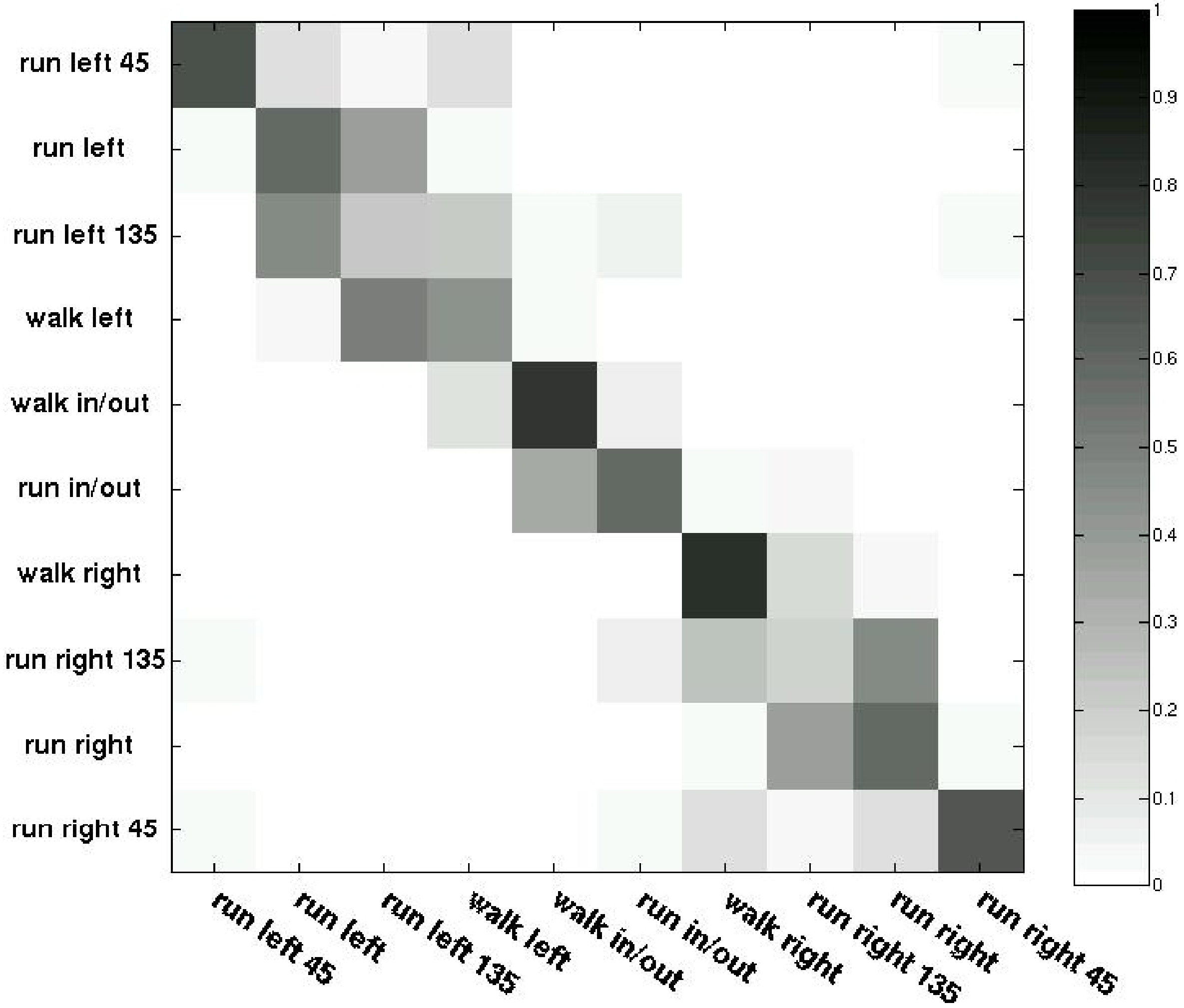
Matched Frames





input matched

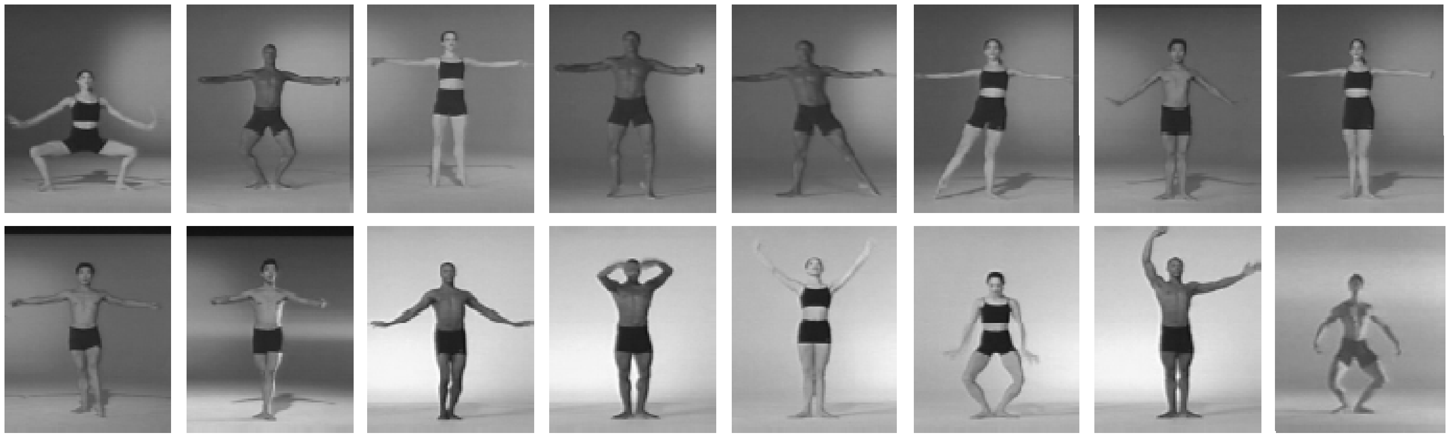
Football Actions: classification

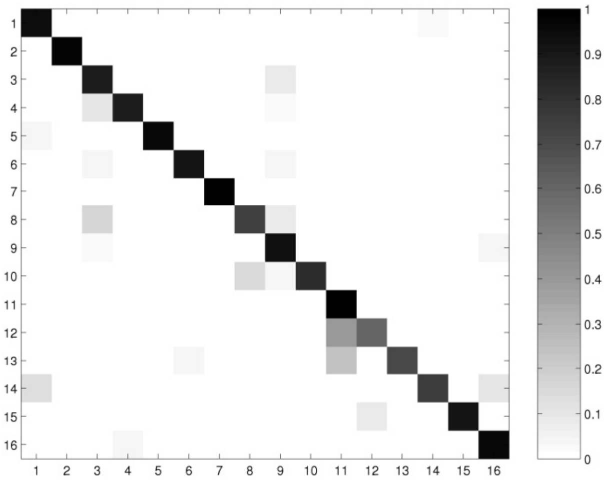


10 actions; 4500 total frames; 13-frame motion descriptor

Classifying Ballet Actions

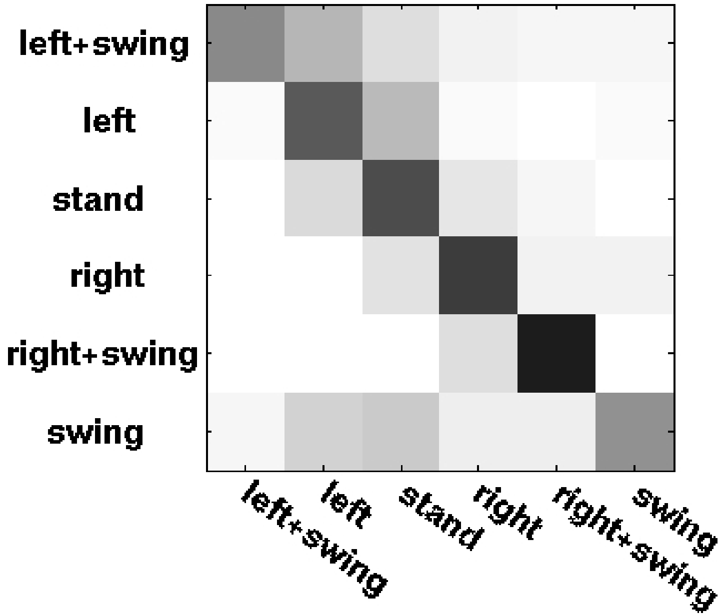
16 Actions; 24800 total frames; 51-frame motion descriptor. Men used to classify women and vice versa.



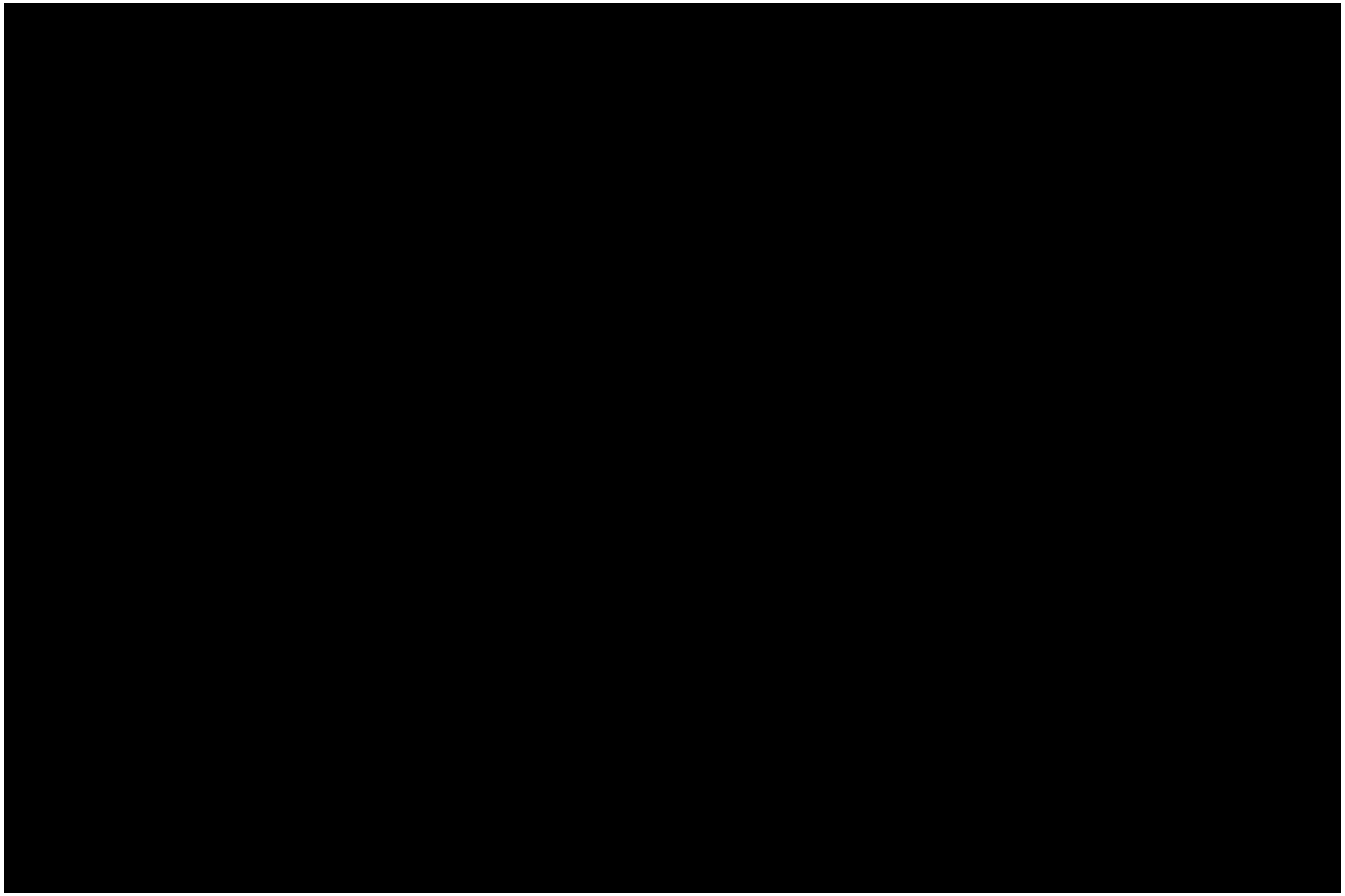


Classifying Tennis Actions

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

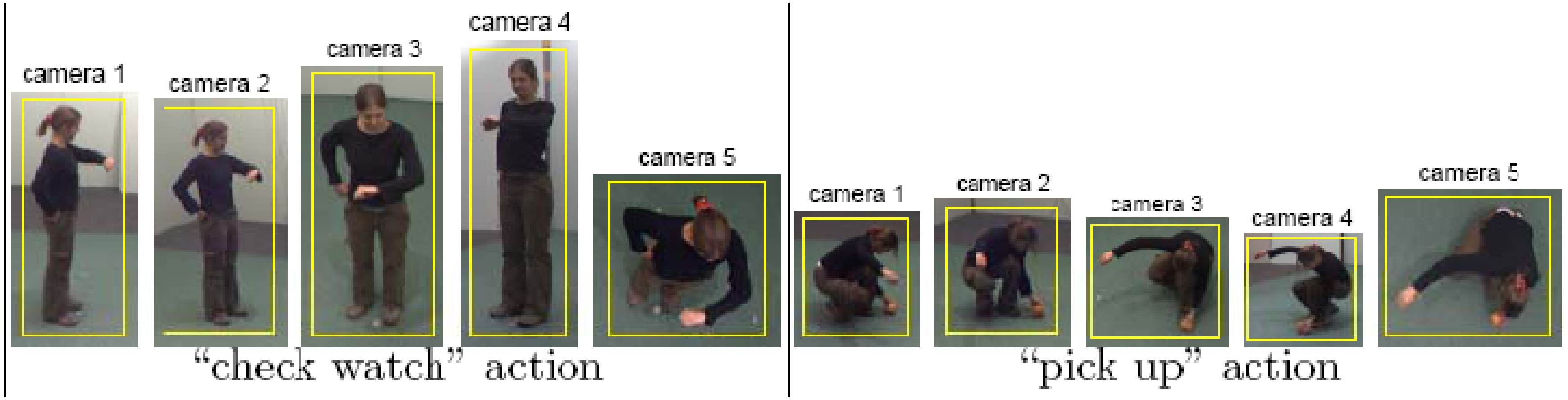


Classifying Tennis Actions



Red bars illustrate classification confidence for each action

What about 3D?



Motion and appearance descriptors are not invariant to view changes

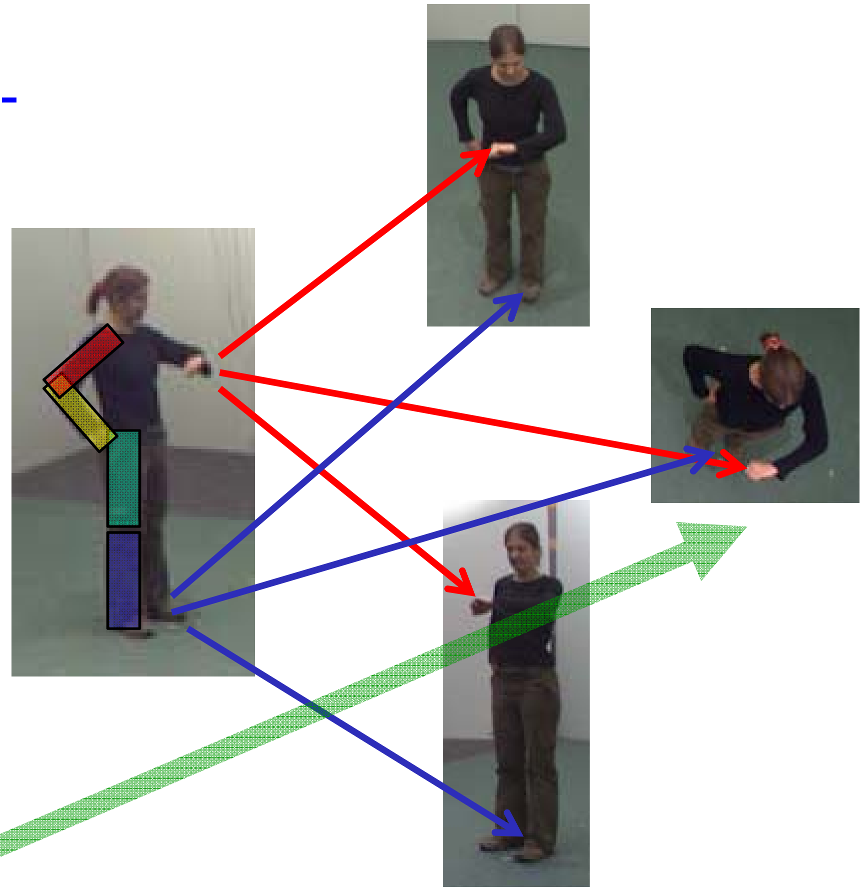
Multi-view action recognition

Difficult to apply standard multi-view methods:

- Do not want to search for multi-view point correspondence --- Non-rigid motion, cloth 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 to recognize actions in different views

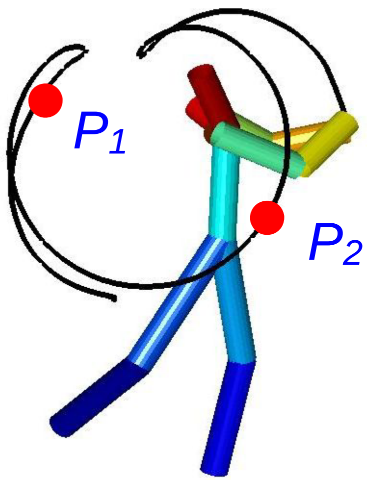


Temporal self-similarities

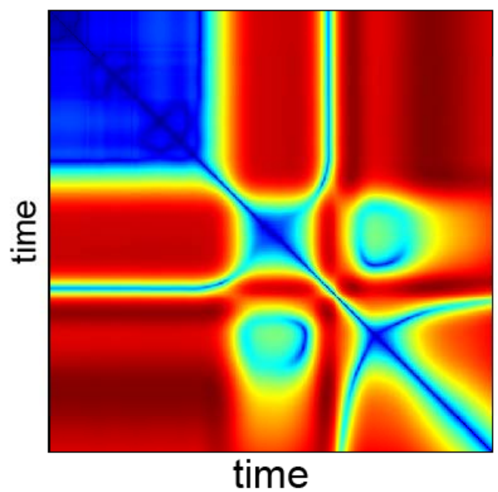
Ideas:

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

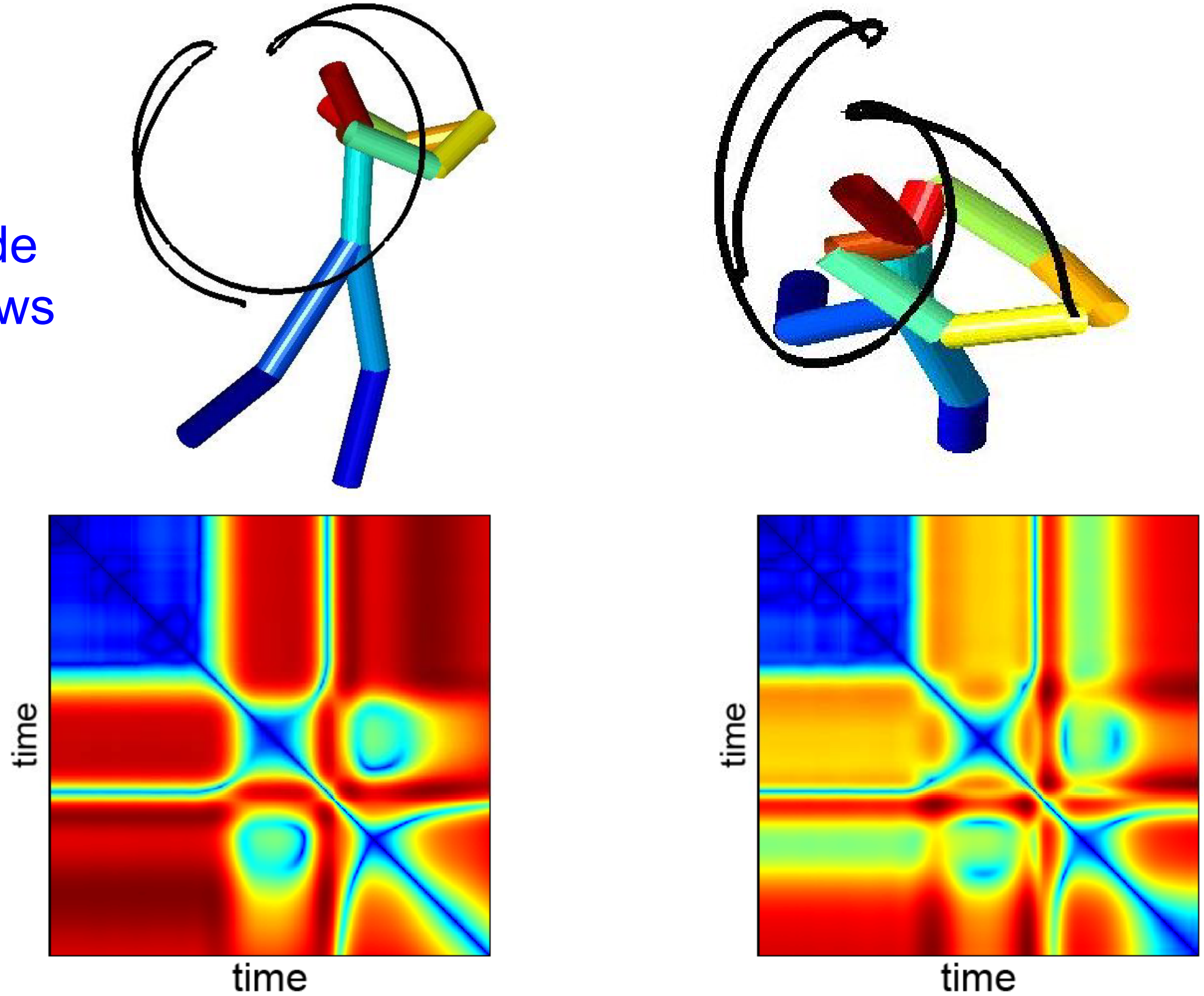


Distance matrix / self-similarity matrix (SSM):



Temporal self-similarities: Multi-views

Example:
Golf swing
from the side
and top views



Cross-View Action Recognition from Temporal Self-Similarities

I. Junejo, E. Dexter, I. Laptev, and P. Perez, **ECCV 2008**

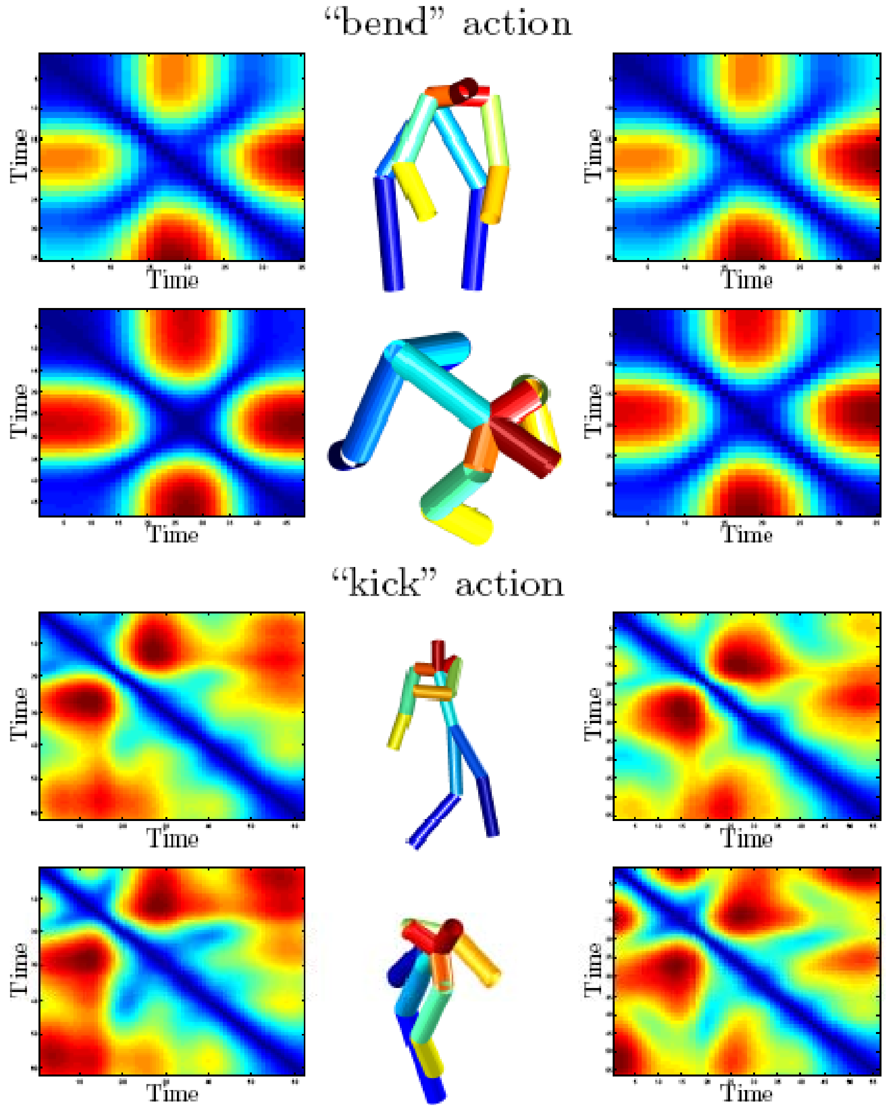
Temporal self-similarities: MoCap

person 1

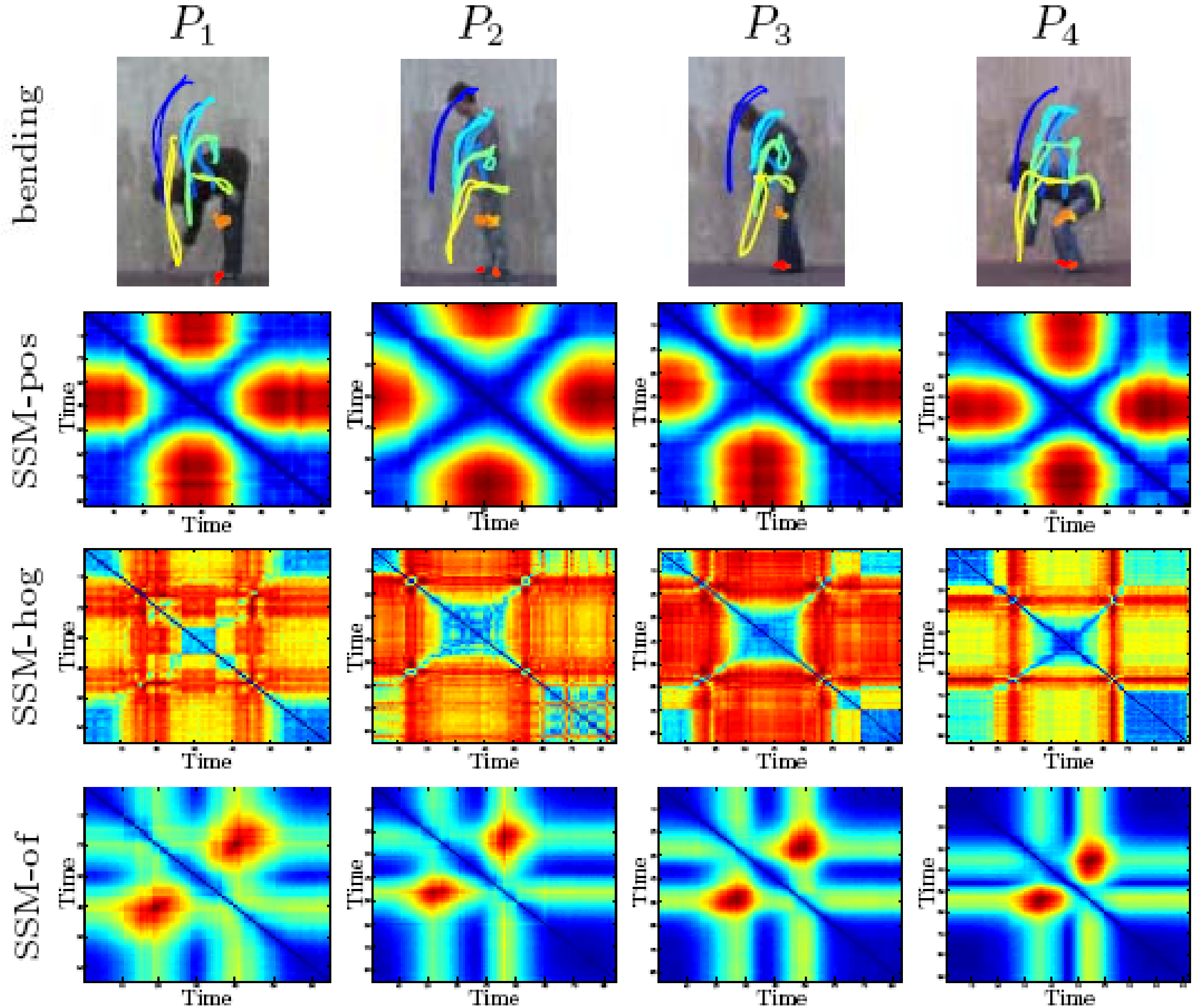
person 2

person 1

person 2



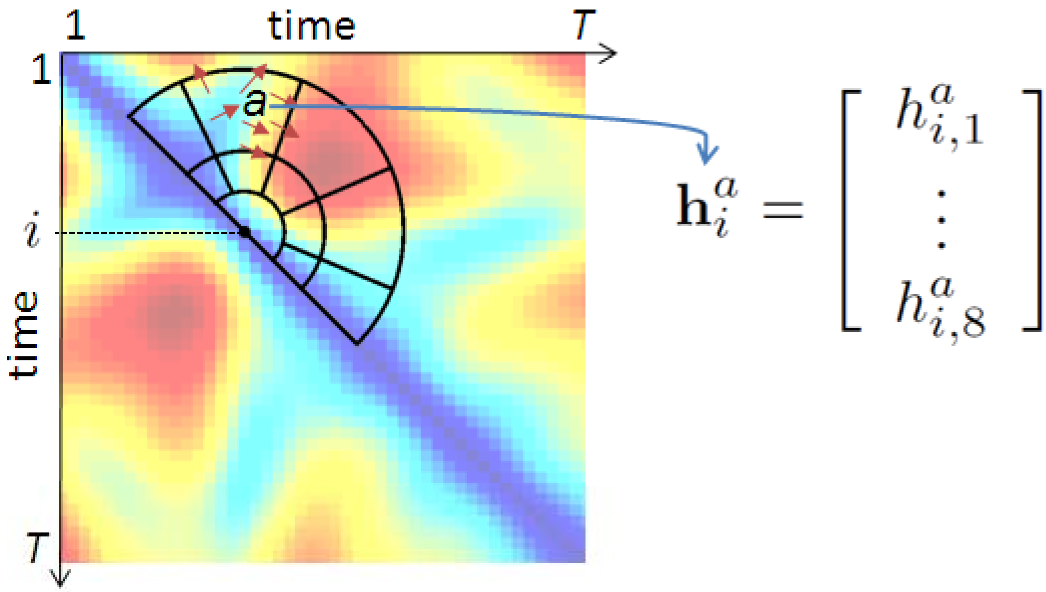
Temporal self-similarities: Video



Self-similarity descriptor

Properties of SSM:

- SPSD
- 0-valuaed diagonal
- uncertainty increases with the distance from the diagonal $\Delta t = t_2 - t_1$

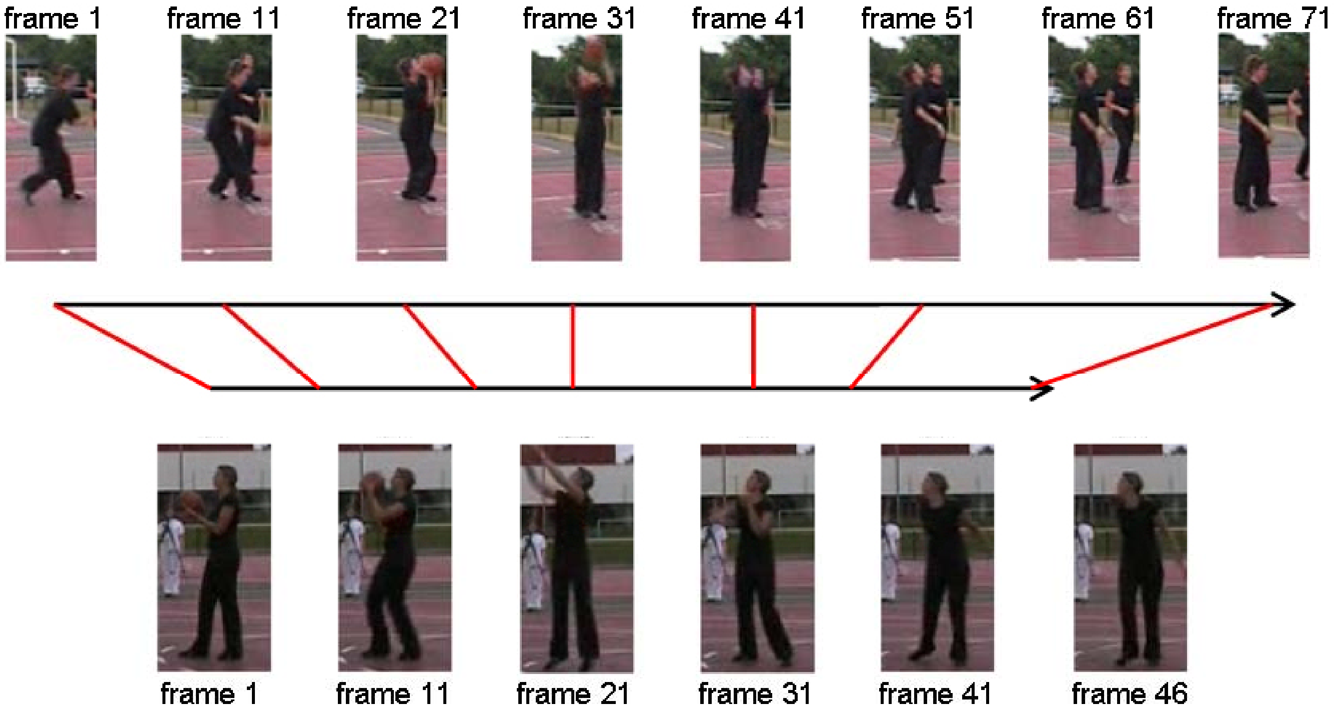


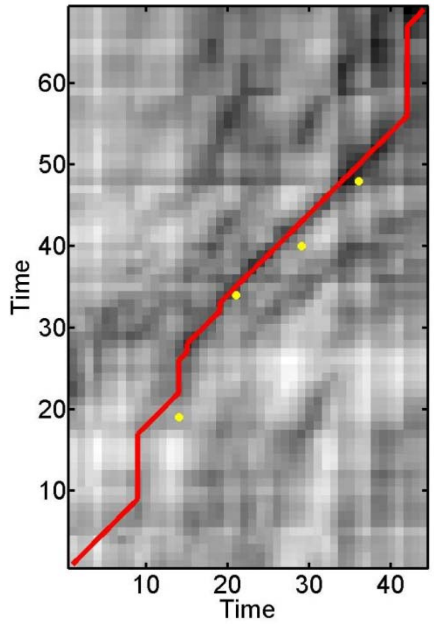
- Define a local histogram descriptor h_i for each point i on the diagonal.

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

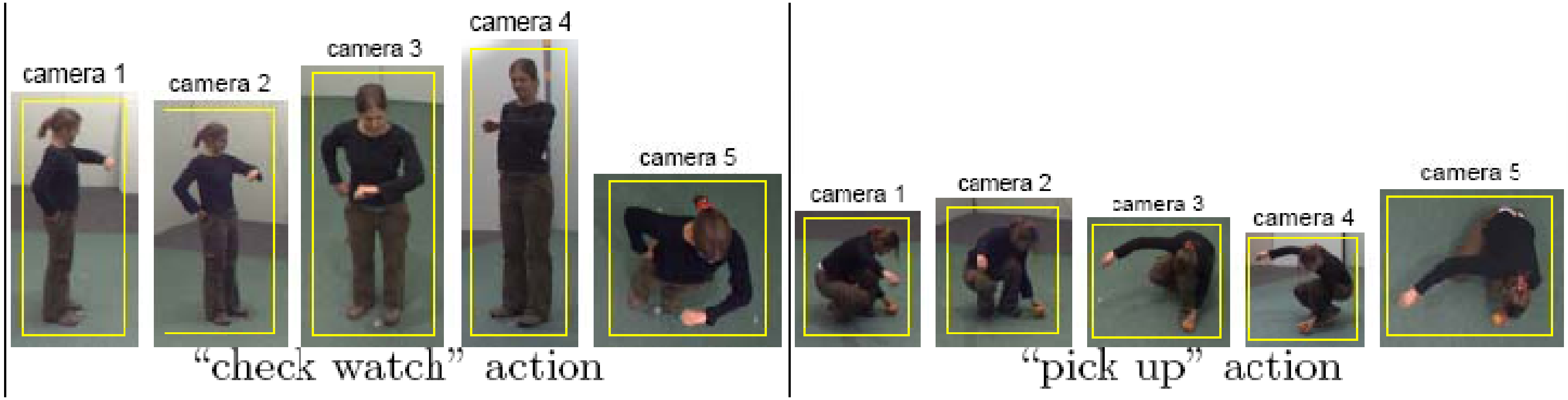
- **Action recognition:**
  - Visual vocabulary for h
  - BoF representation of $\{h_i\}$
  - SVM

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

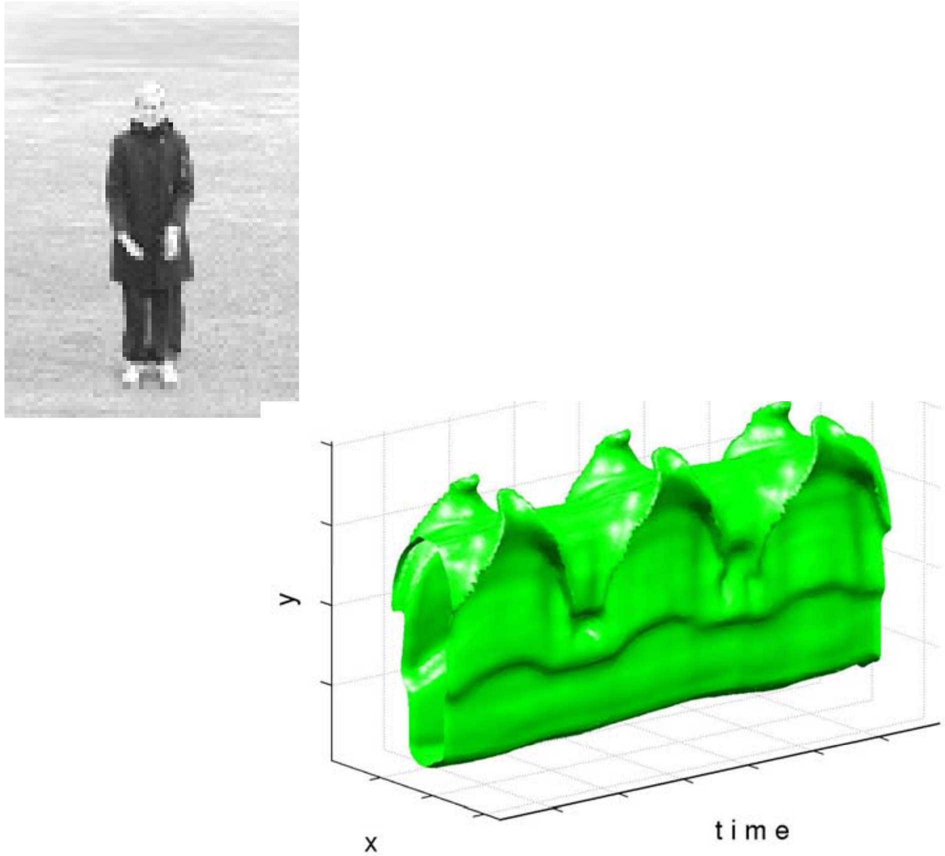
SSM-based recognition

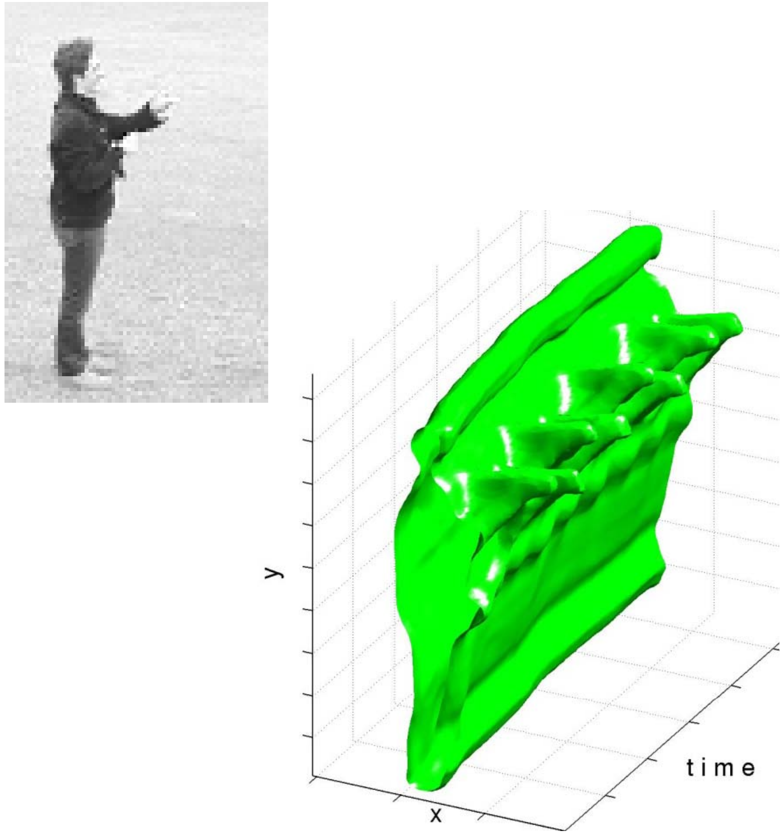
	Test Cam0	Test Cam1	Test Cam2	Test Cam3	Test Cam4	Test All
Train Cam0	80.0	75.9	42.3	55.6	21.8	55.6
Train Cam1	74.8	83.9	36.5	58.3	23.6	56.0
Train Cam2	43.6	46.1	80.5	64.7	34.2	53.7
Train Cam3	47.0	50.0	45.8	85.5	18.8	49.5
Train Cam4	19.7	19.4	43.5	26.1	73.3	36.0
Train All	80.3	84.5	79.4	84.8	68.5	79.6

cross−camera training/testing   same camera training/testing

Alternative view-variant method (STIP)

Actions == Space-time objects?

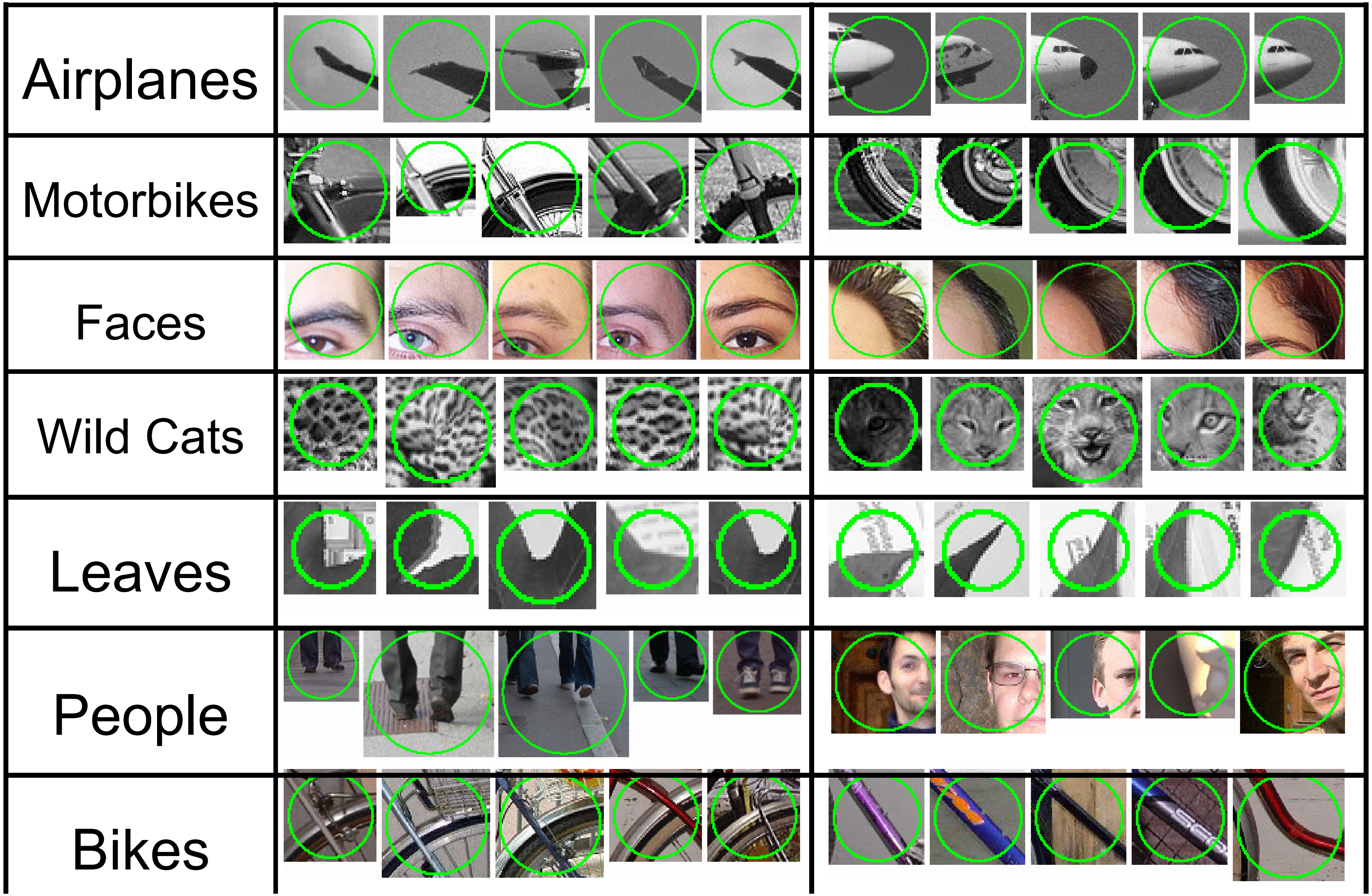




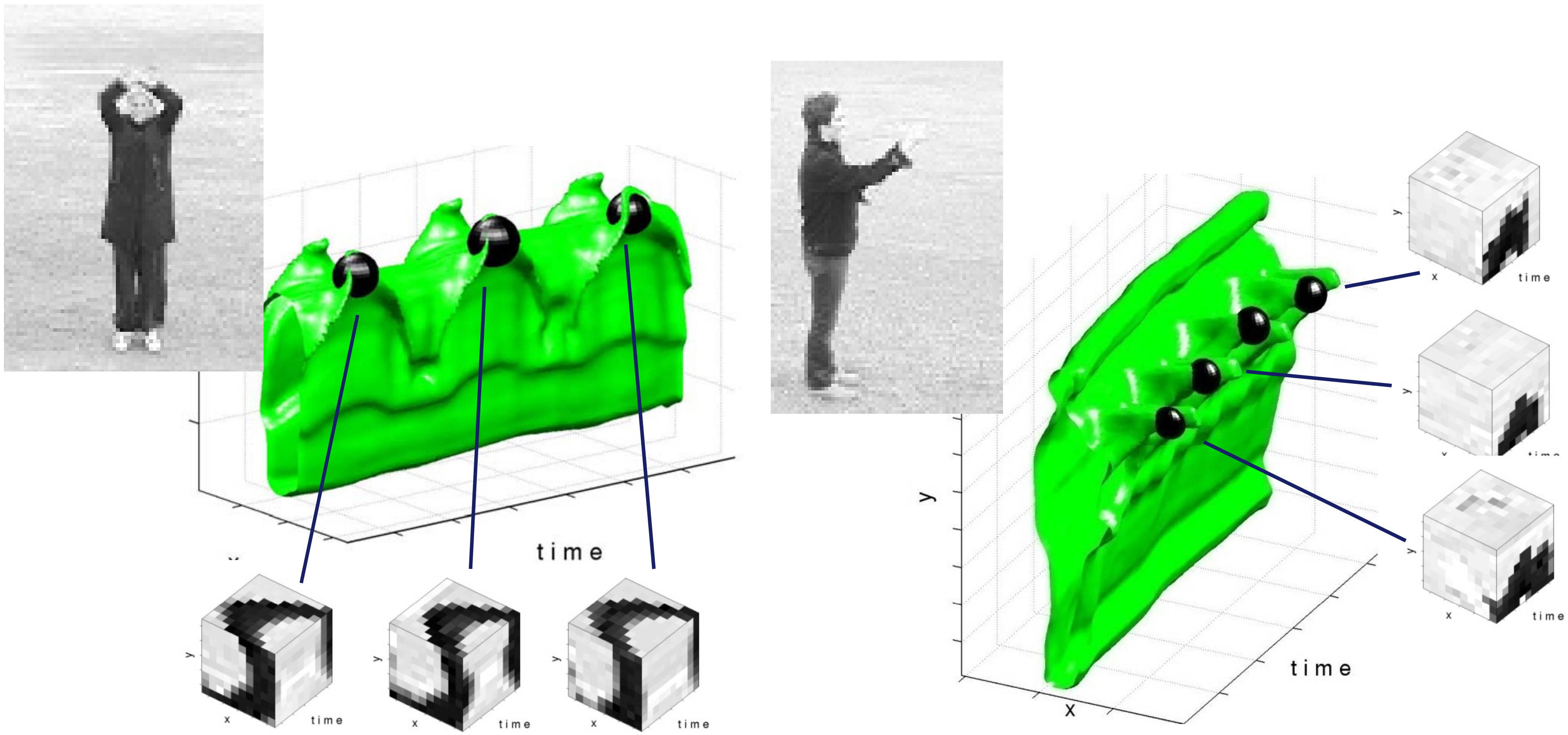
Can we treat actions as space-time objects and apply object recognition methods used in static images?

Local approach: Bag of Visual Words

(Lecture 5)



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 $\mu(\cdot;\, \Sigma)$

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

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

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

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

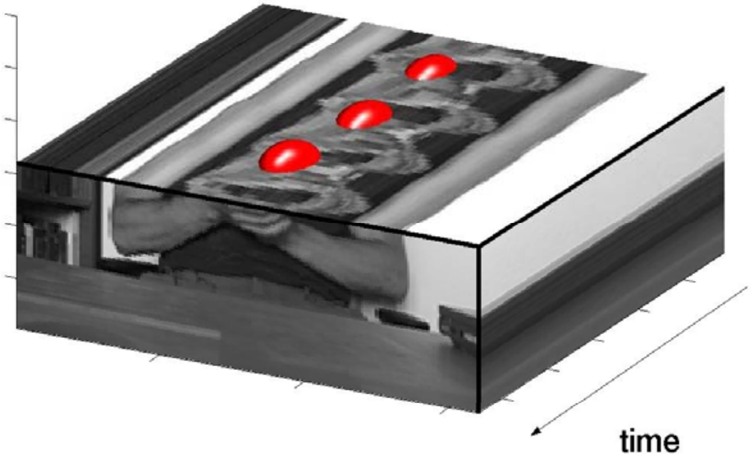
Large eigenvalues of μ can be detected by the local maxima of H over (x,y,t):

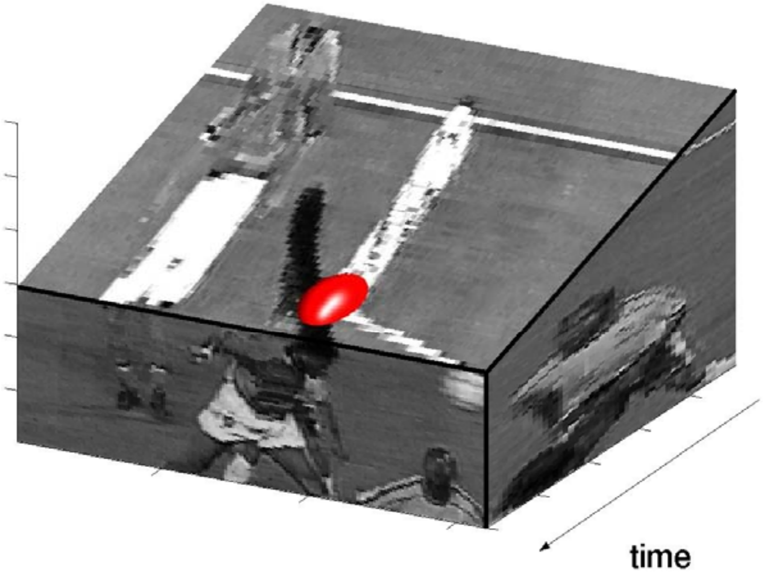
$$\begin{aligned} H(p;\, \Sigma) &= \det(\mu(p;\, \Sigma)) + k\,\text{trace}^3(\mu(p;\, \Sigma)) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3 \end{aligned}$$

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

Space-Time Interest Points: Examples

Motion event detection

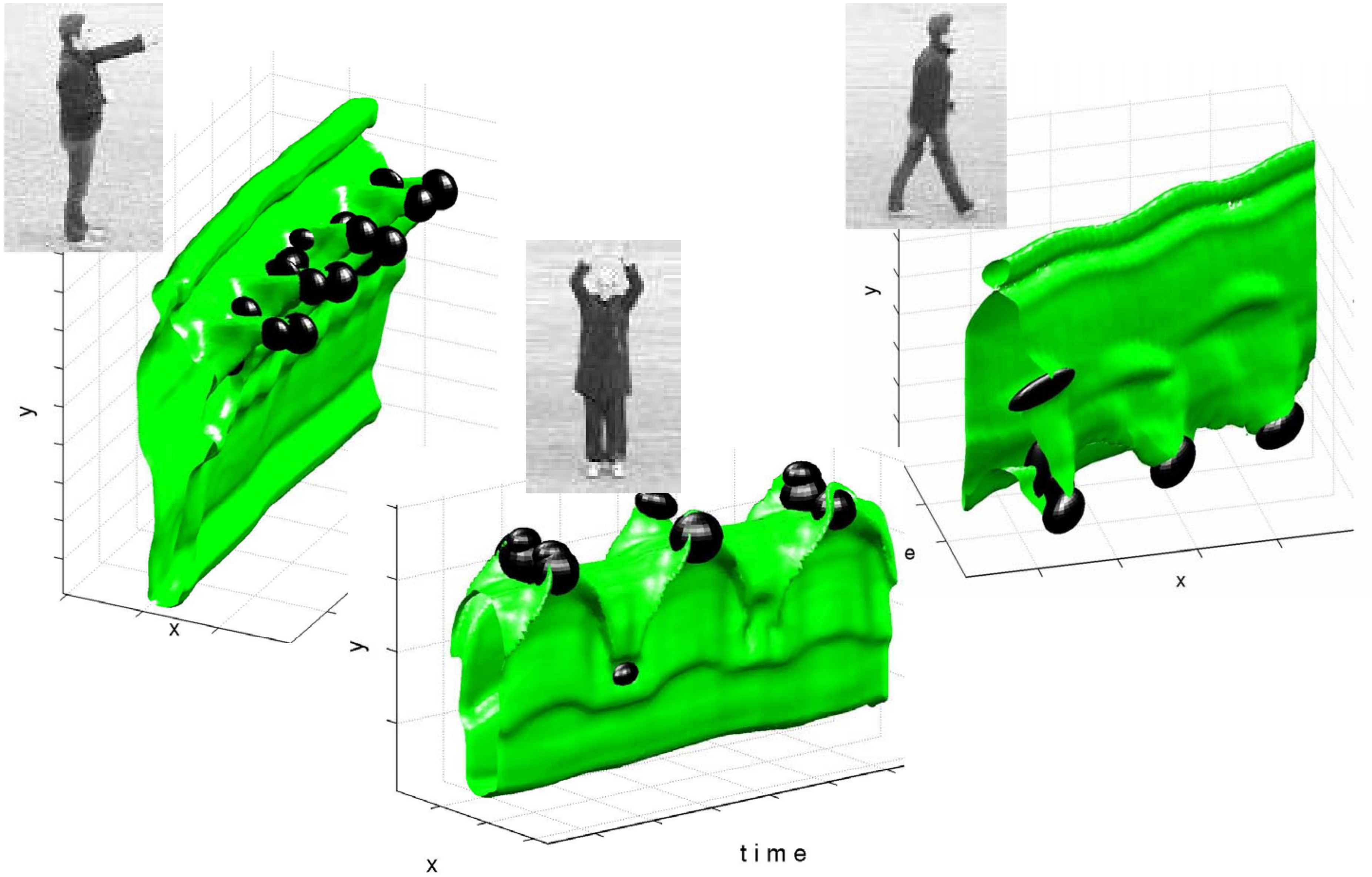




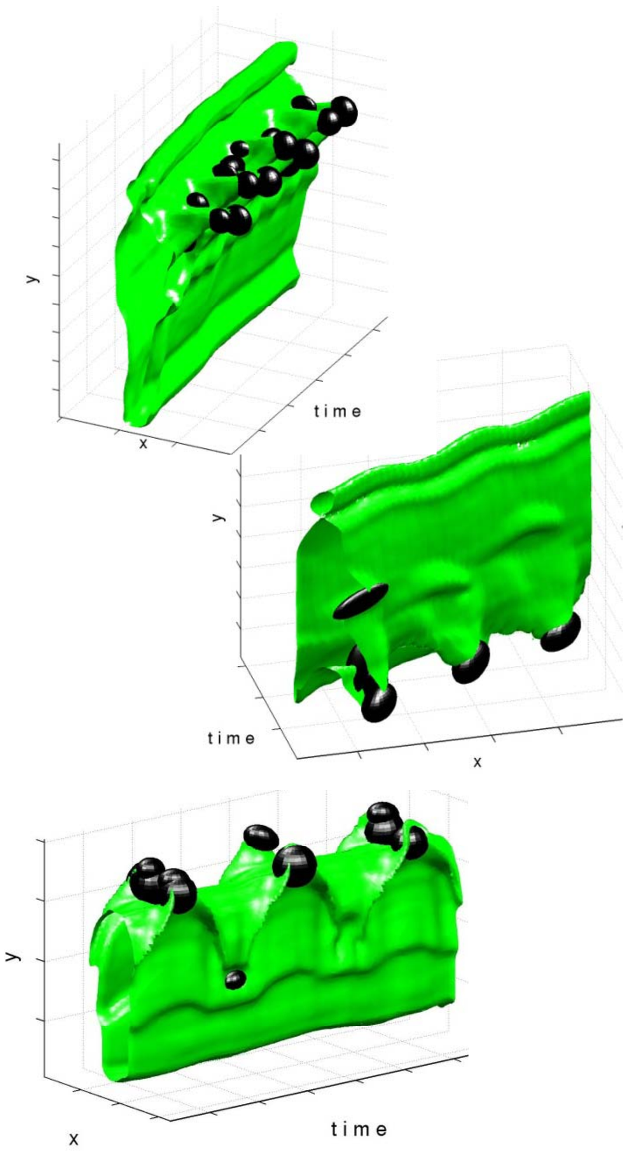
Space-Time Interest Points: Examples

Motion event detection: complex background

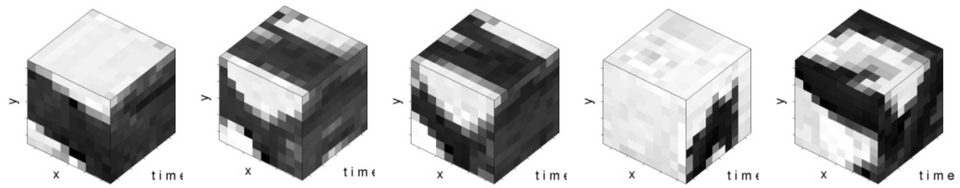
Features from human actions



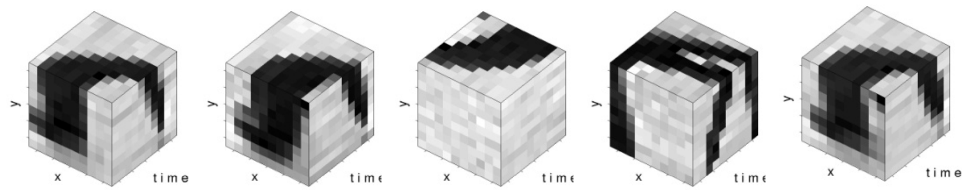
Features from human actions



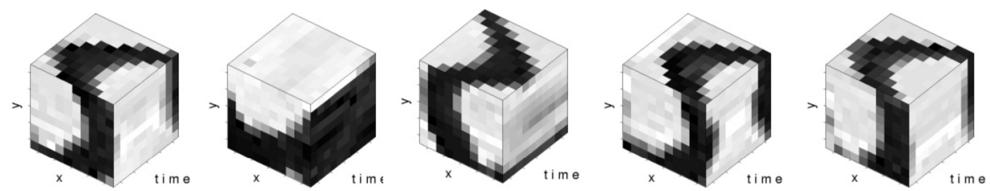
boxing



walking



hand waving



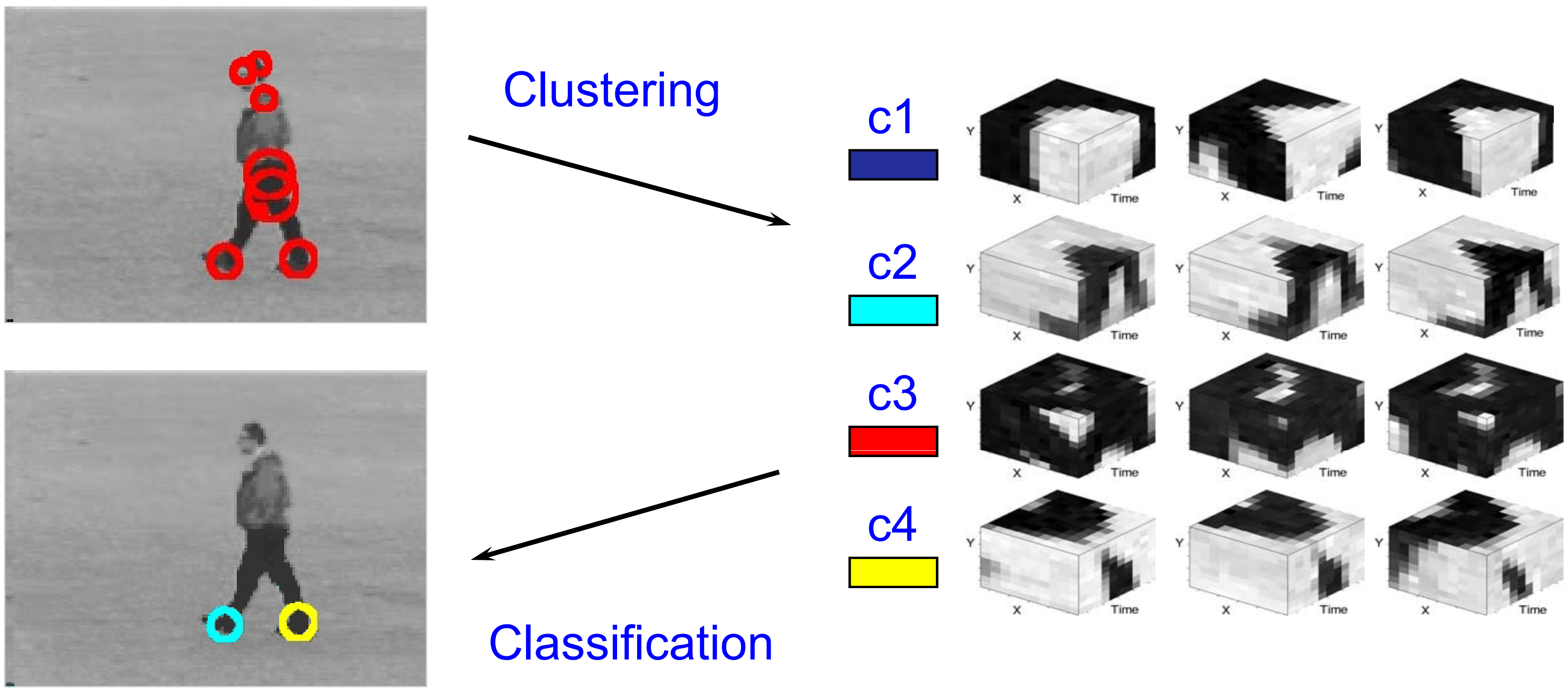
Local space-time descriptors

A common choice for local descriptors is a local jet (Koenderink and van Doorn, 1987) computed from spatio-temporal Gaussian derivatives (here at interest points p_i)

$$d_i = (L_{x'}, L_{y'}, L_{t'}, L_{x'x'}, L_{x'y'}, L_{x't'}, \ldots, L_{t't't't'})$$

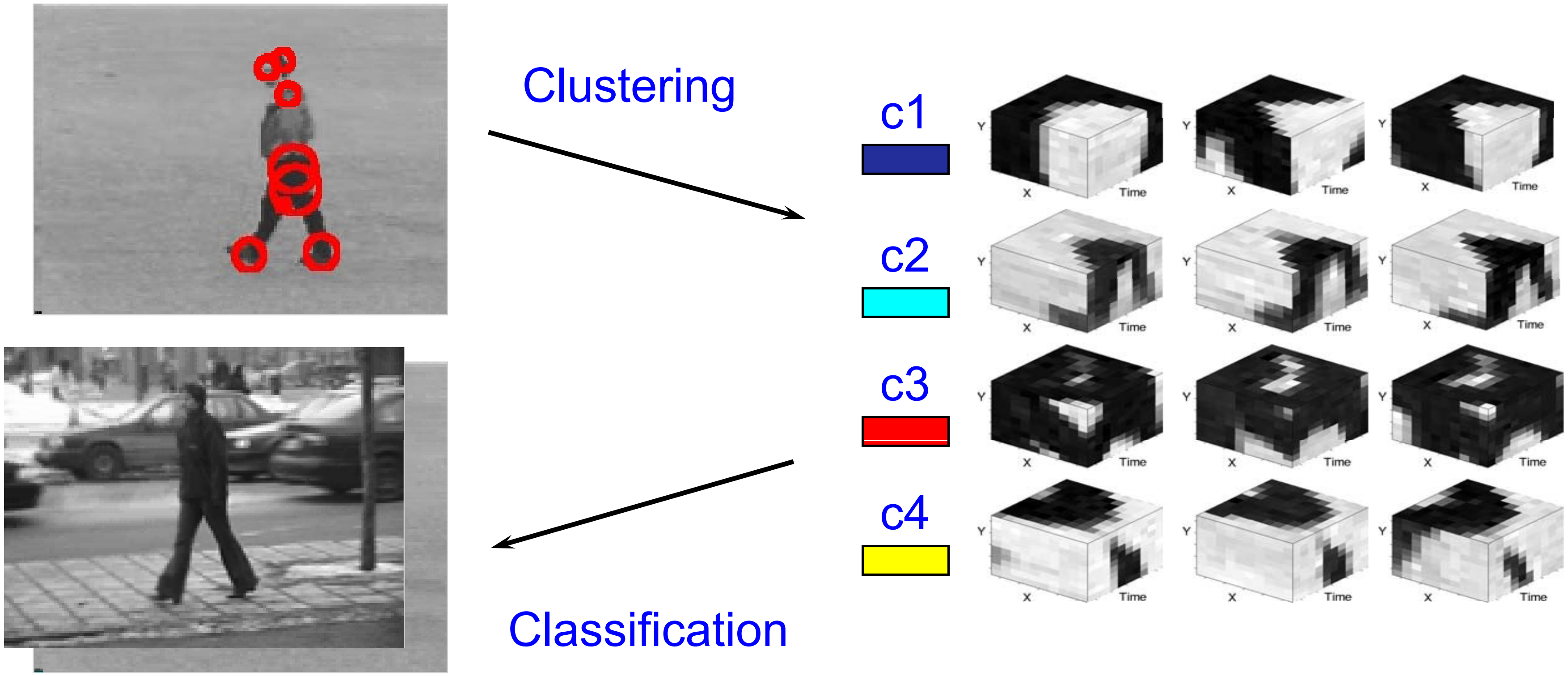
Visual Vocabulary: K-means clustering

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



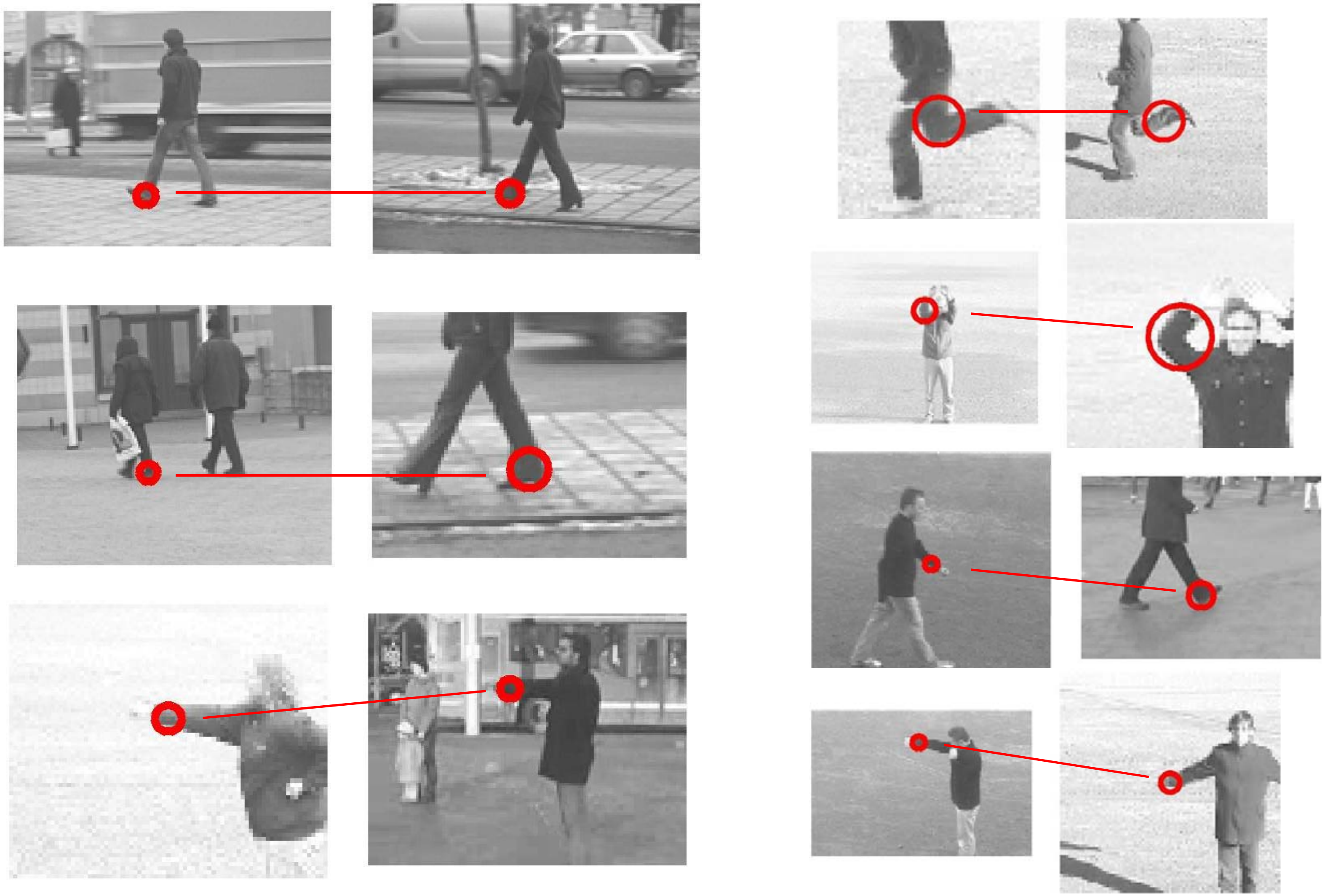
Visual Vocabulary: K-means clustering

- 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

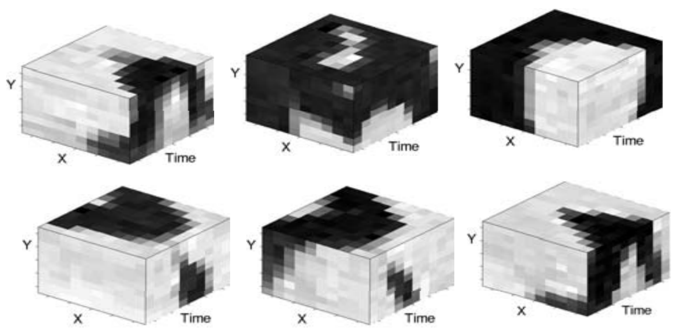


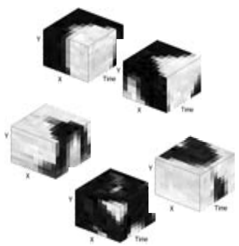
Action Classification: Overview

Bag of space-time features + multi-channel SVM

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

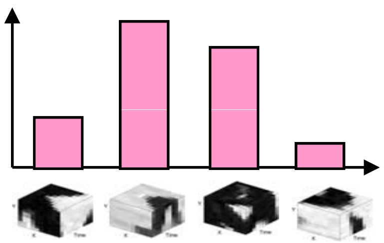
Collection of space-time patches





HOG & HOF
patch
descriptors

Histogram of visual words



Multi-channel
SVM
Classifier

Action recognition in KTH dataset

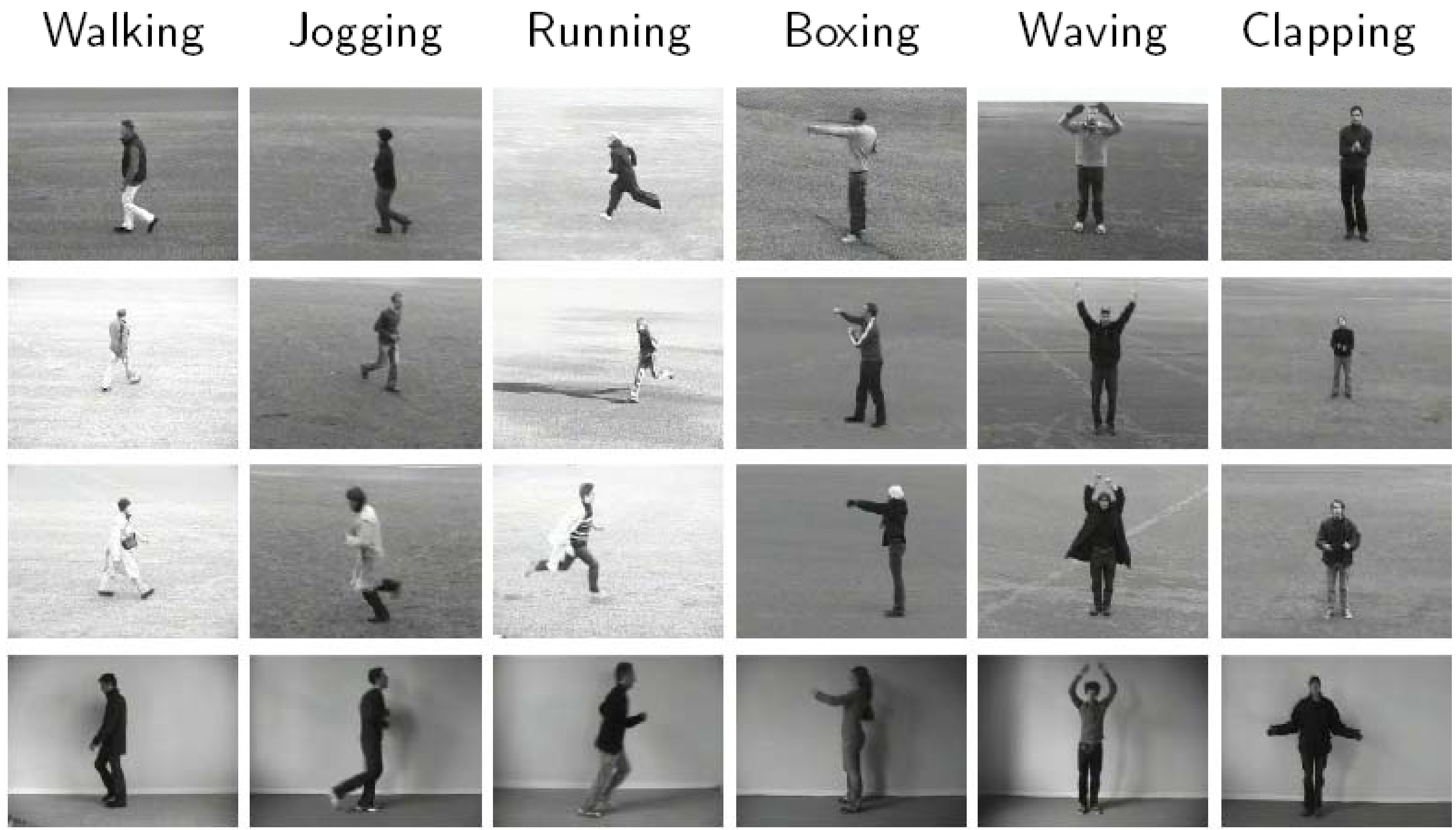


Figure: Sample frames from the KTH actions sequences, all six classes (columns) and scenarios (rows) are presented

Classification results on KTH dataset

Method	Schuldt et al.	Niebles et al.	Wong et al.	Nowozin et al.	ours
Accuracy	71.7%	81.5%	86.7%	87.0%	**91.8%**

Table: Average class accuracy on the KTH actions 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

Table: Confusion matrix for the KTH actions

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”:






Learning Actions from Movies

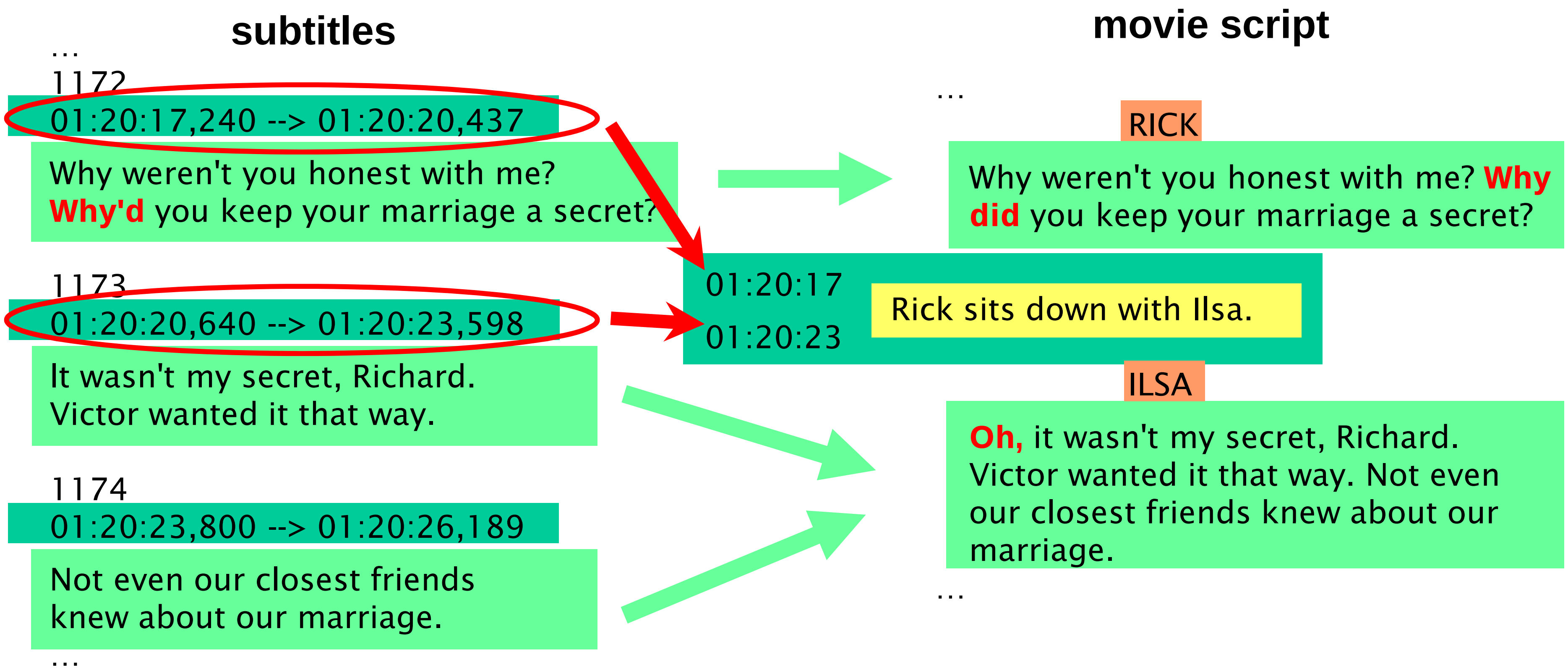
- Realistic variation of human actions
- Many classes and many examples per class

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

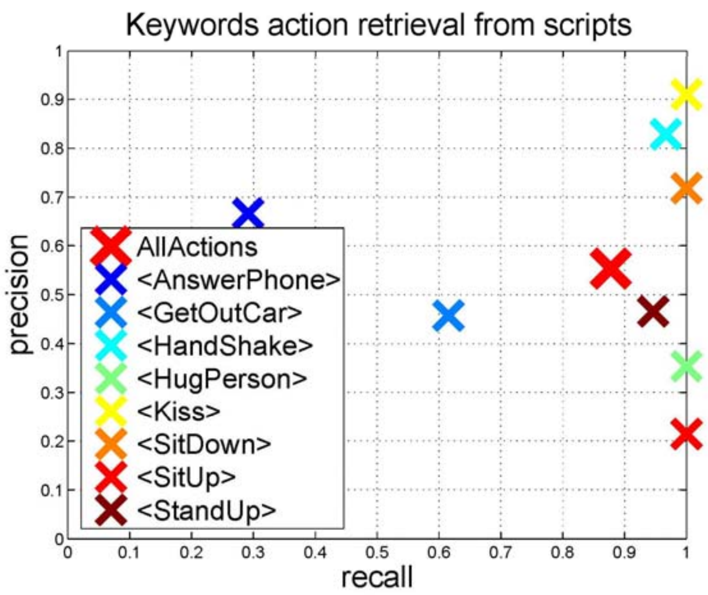


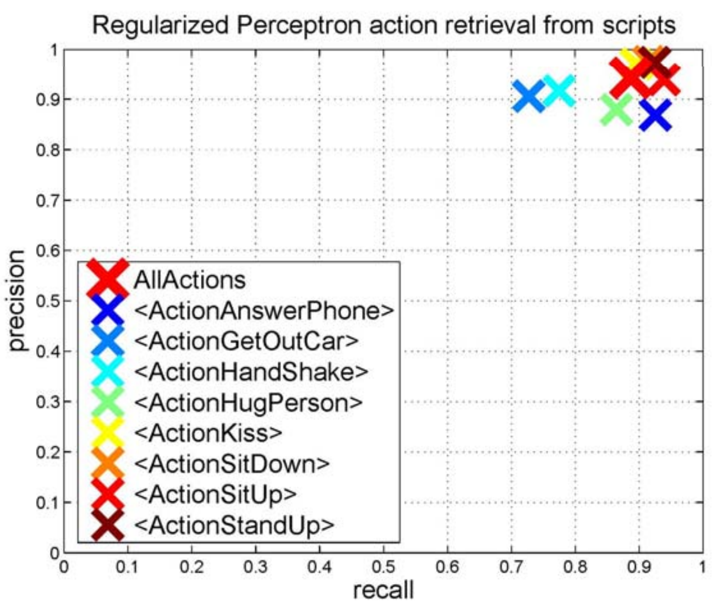
Text-based action retrieval

- 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





Movie actions dataset

		<AnswerPhone>	<GetOutCar>	<HandShake>	<HugPerson>	<Kiss>	<SitDown>	<SitUp>	<StandUp>	Total
12 movies	automatically labeled training set: False	5	6	9	7	10	21	5	33	96
	automatically labeled training set: Correct	15	6	14	8	34	30	7	29	143
	automatically labeled training set: All	20	12	23	15	44	51	12	62	239
	manually labeled training set	22	13	20	22	49	47	11	48	232
20 different movies	test set	23	13	19	22	51	30	10	49	217

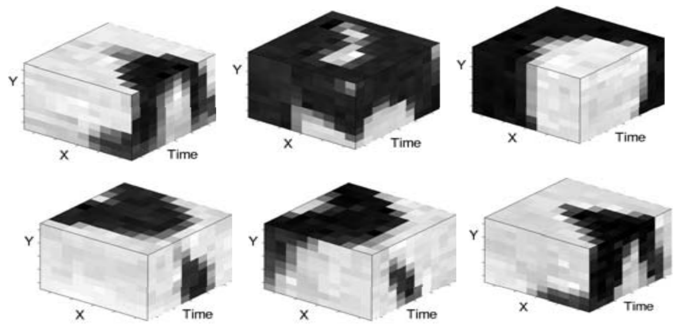
- Learn vision-based classifier from automatic training set
- Compare performance to the manual training set

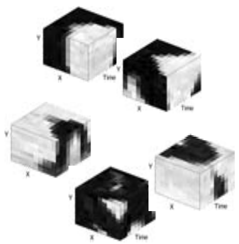
Action Classification: Overview

Bag of space-time features + multi-channel SVM

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

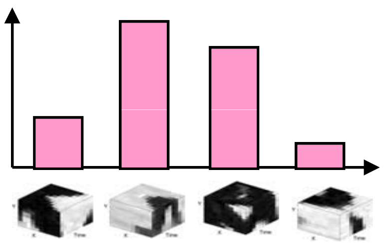
Collection of space-time patches





HOG & HOF
patch
descriptors

Histogram of visual words



Multi-channel
SVM
Classifier

Action classification (CVPR08)

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

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

I wish I didn't love you so much.

01:22:00
01:22:03

She snuggles closer to Rick.

CUT TO:

EXT. RICK'S CAFE - NIGHT

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

The headlights of a speeding police car sweep toward them.

They flatten themselves against a wall to avoid detection.

The lights move past them.

CARL

01:22:15
01:22:17

I think we lost them.

...

Mining scene captions

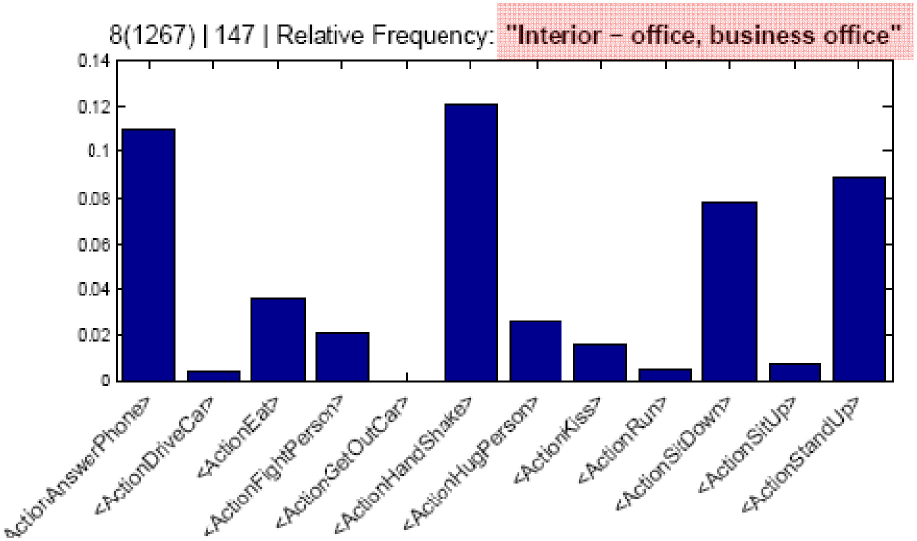
INT. TRENDY RESTAURANT - NIGHT
INT. MARSELLUS WALLACE'S DINING ROOM MORNING
EXT. STREETS BY DORA'S HOUSE - DAY.
INT. MELVIN'S APARTMENT, BATHROOM – NIGHT
EXT. NEW YORK CITY STREET NEAR CAROL'S RESTAURANT – DAY
INT. CRAIG AND LOTTE'S BATHROOM - DAY

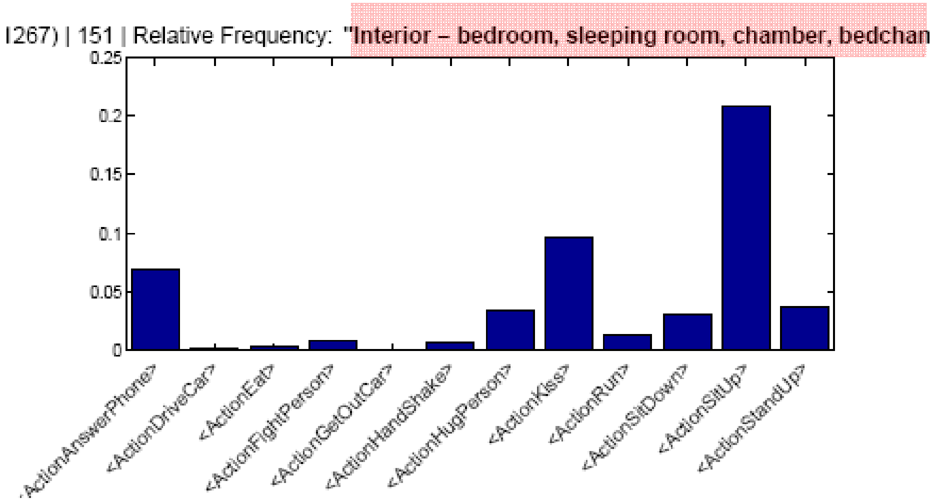
- Maximize word frequency ⇨ street, living room, bedroom, car ….
- Merge words with similar senses using WordNet:

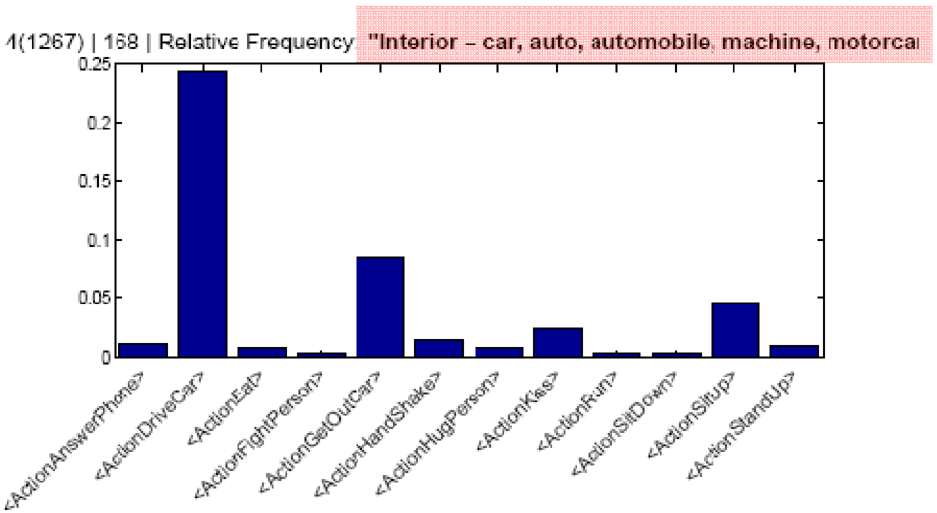
  taxi -> car, cafe -> restaurant

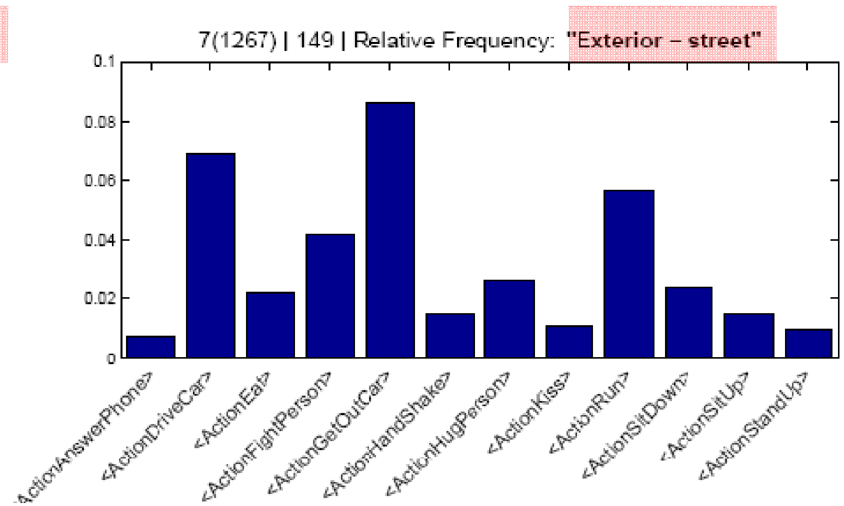
- Measure correlation of words with actions (in scripts) and
- Re-sort words by the entropy $S = -k \sum P_i \ln P_i$
  for P = p(action | word)

Co-occurrence of actions and scenes in scripts

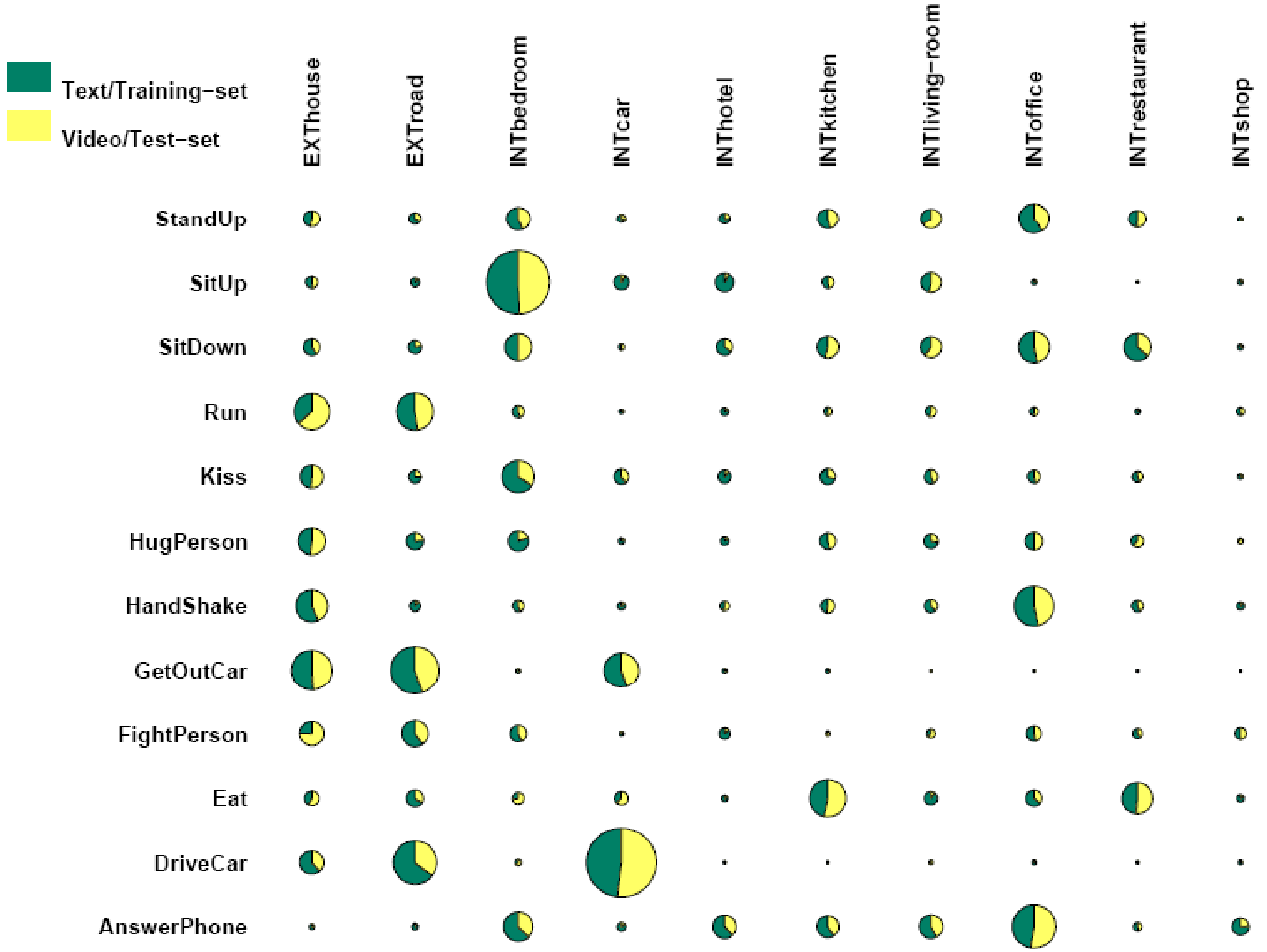








Co-occurrence of actions and scenes in text vs. video



Automatic gathering of relevant scene classes and visual samples

	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

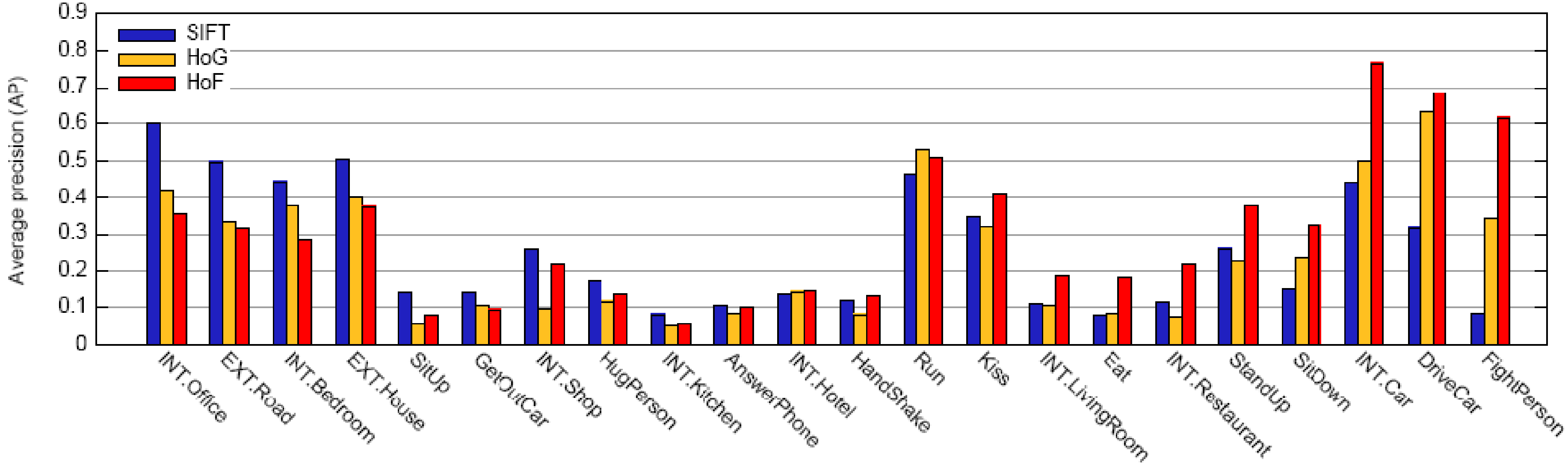
	Auto-Train-Scenes	Clean-Test-Scenes
EXT-house	81	140
EXT-road	81	114
INT-bedroom	67	69
INT-car	44	68
INT-hotel	59	37
INT-kitchen	38	24
INT-living-room	30	51
INT-office	114	110
INT-restaurant	44	36
INT-shop	47	28
All Samples	570	582

(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

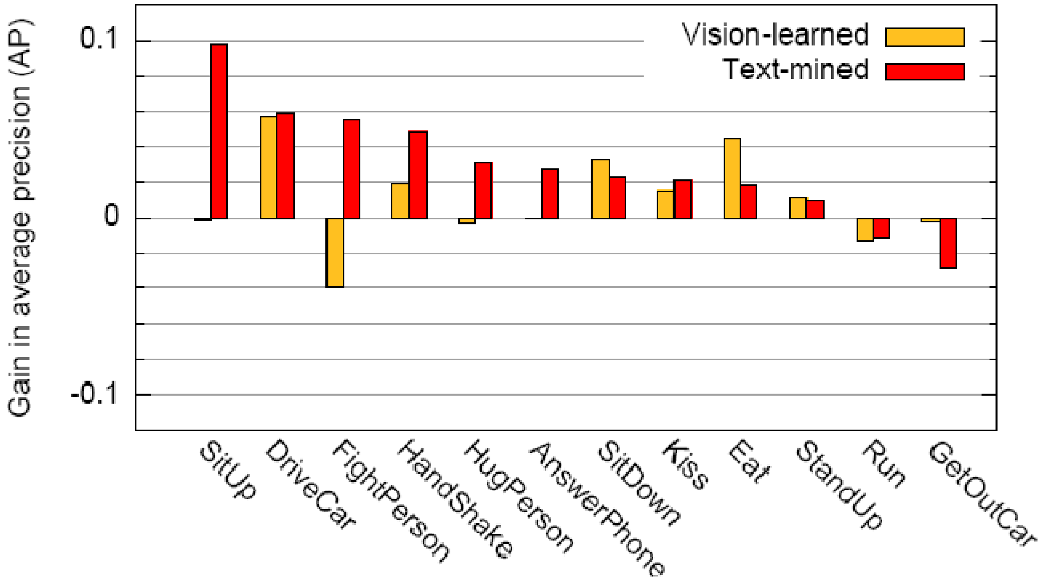
$s_j(\boldsymbol{x})$ Scene classification score

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

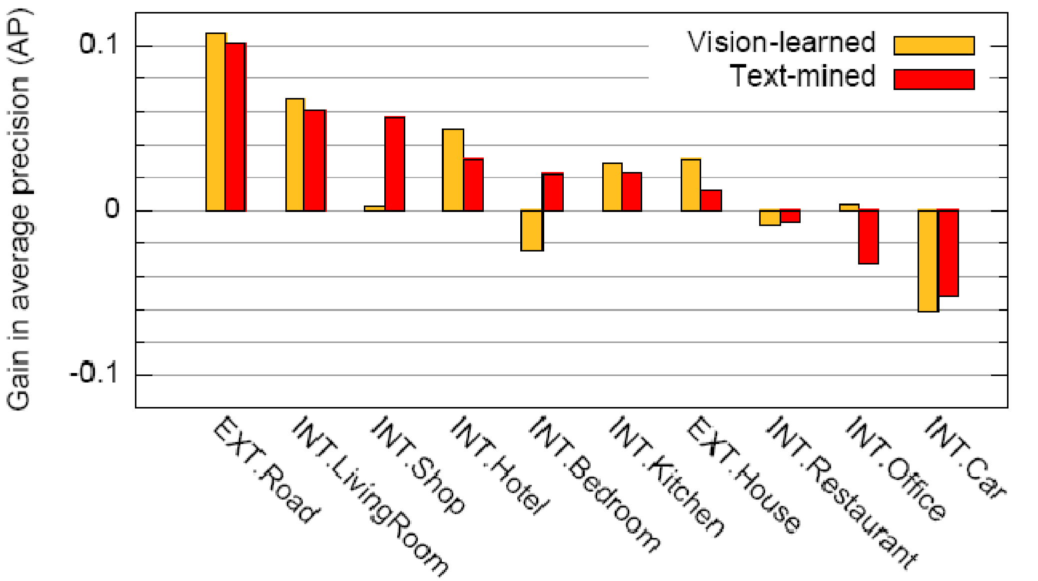
$a'_i(\boldsymbol{x})$ New action score

Results: actions and scenes (jointly)

Actions
in the
context
of
Scenes

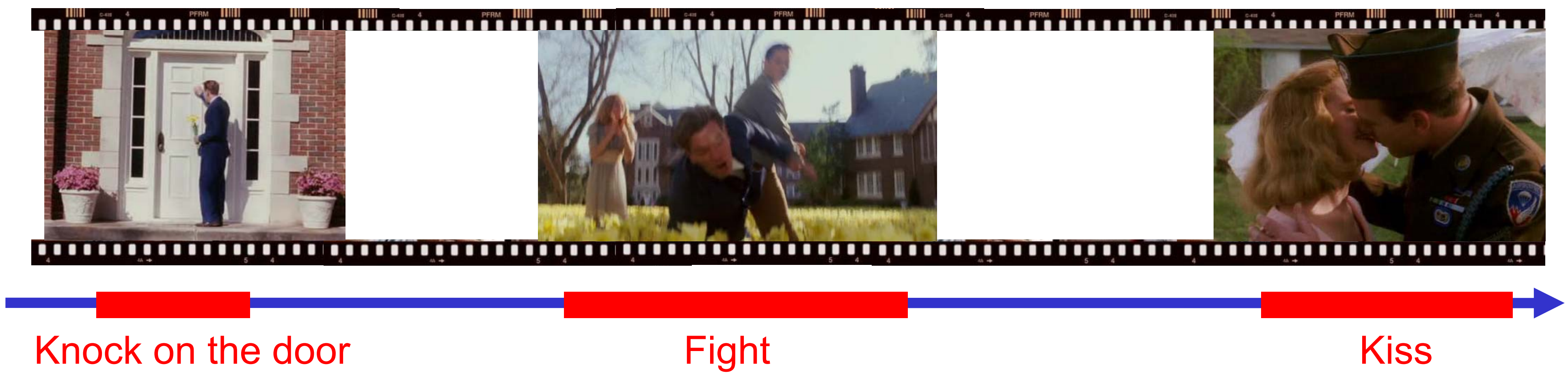


Scenes
in the
context
of
Actions



Weakly-Supervised Temporal Action Annotation (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

Subtitles

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

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

Script

Speech
Monsieur Laszlo. Right this way.

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

Speech
Two cointreaux, please.

24:25

Uncertainty!

24:51

Overview

Input:

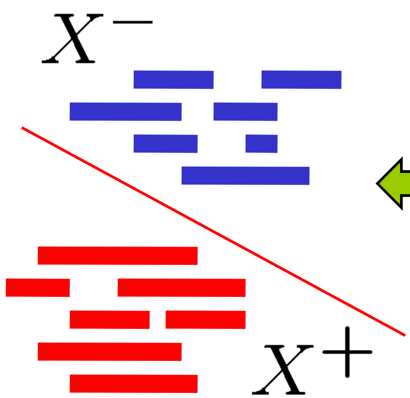
- Action type, e.g.
  Person Opens Door
- Videos + aligned scripts

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

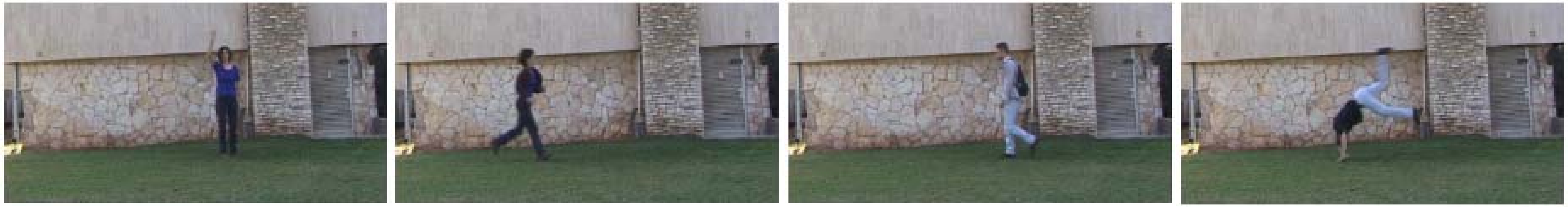


Output:

Sliding-window-style temporal action localization

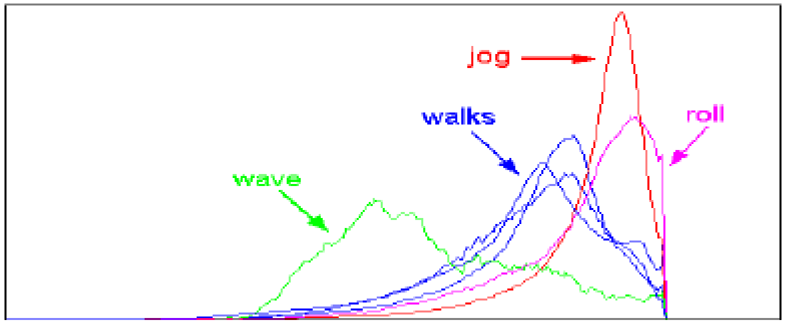
Action clustering

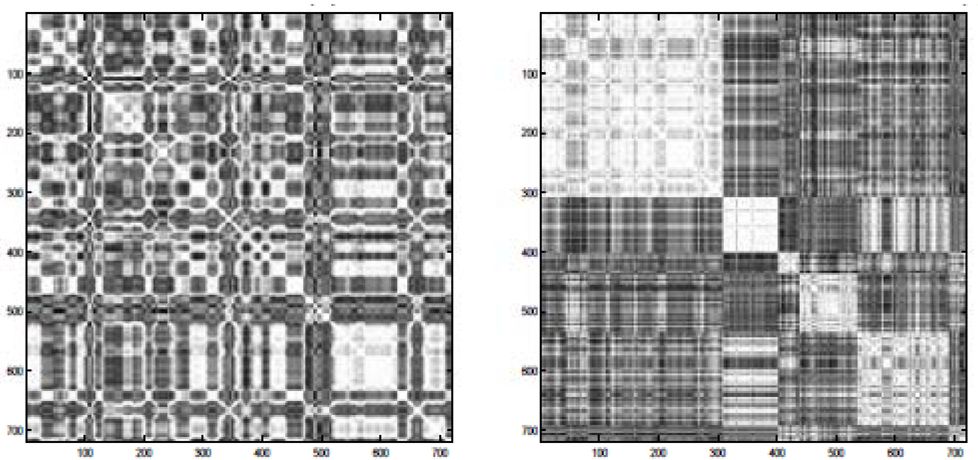
[Lihi Zelnik-Manor and Michal Irani CVPR 2001]



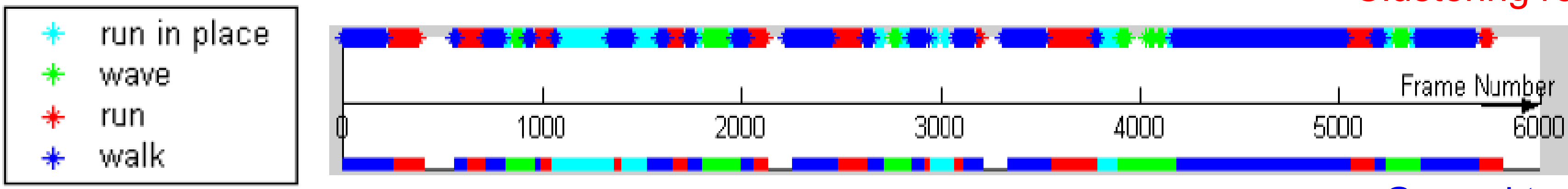
Spectral clustering

Descriptor space





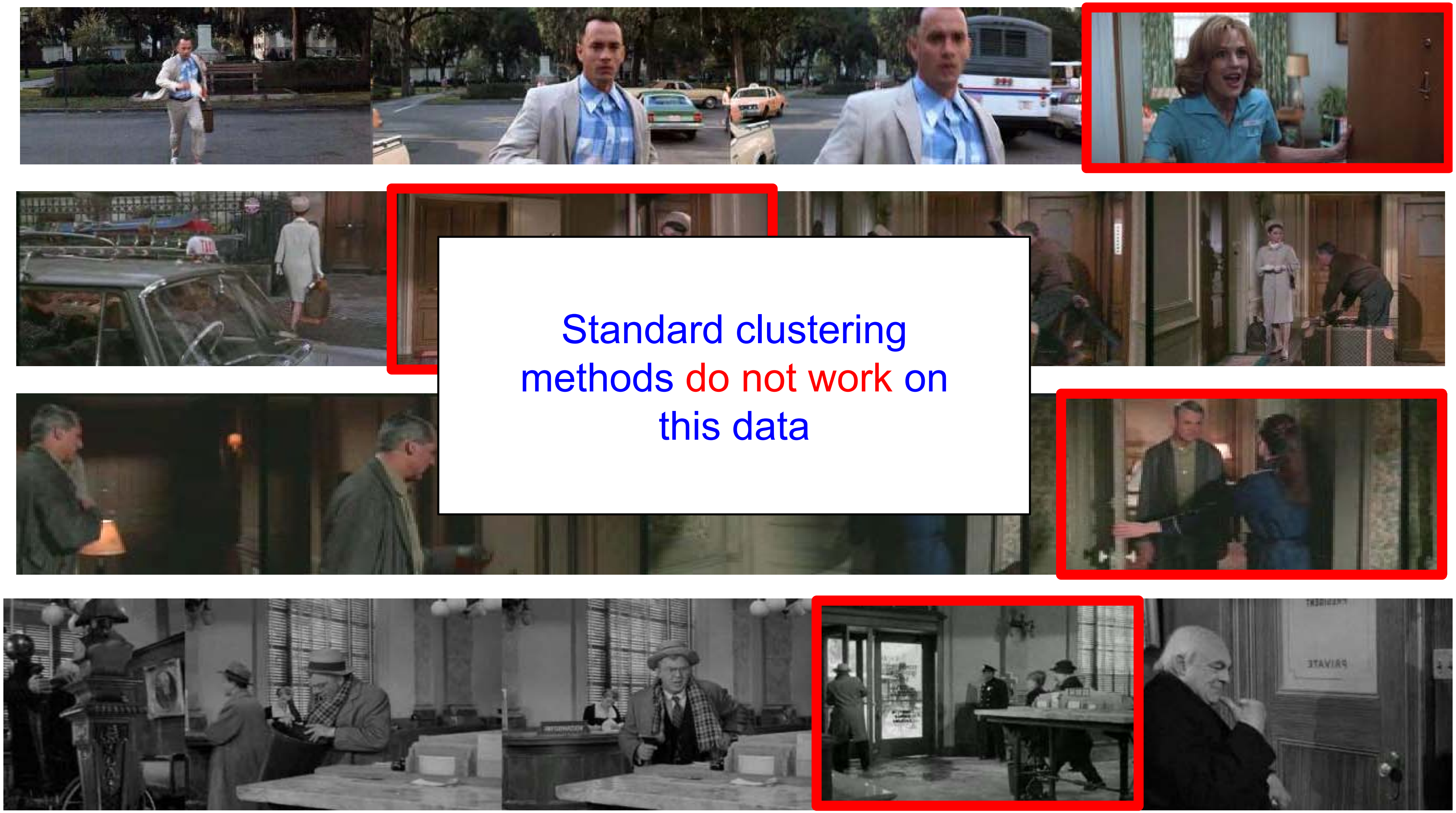
Clustering results



Ground truth

Action clustering

Our data:

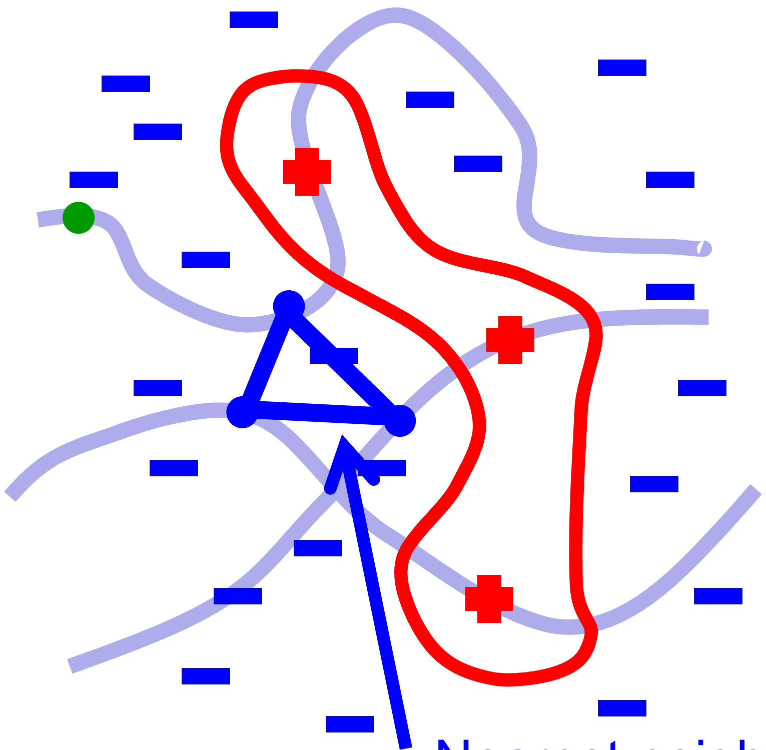


Standard clustering methods do not work on this data

Action clustering

Our view at the problem

Feature space



Nearest neighbor solution: wrong!

Video space

Negative samples!

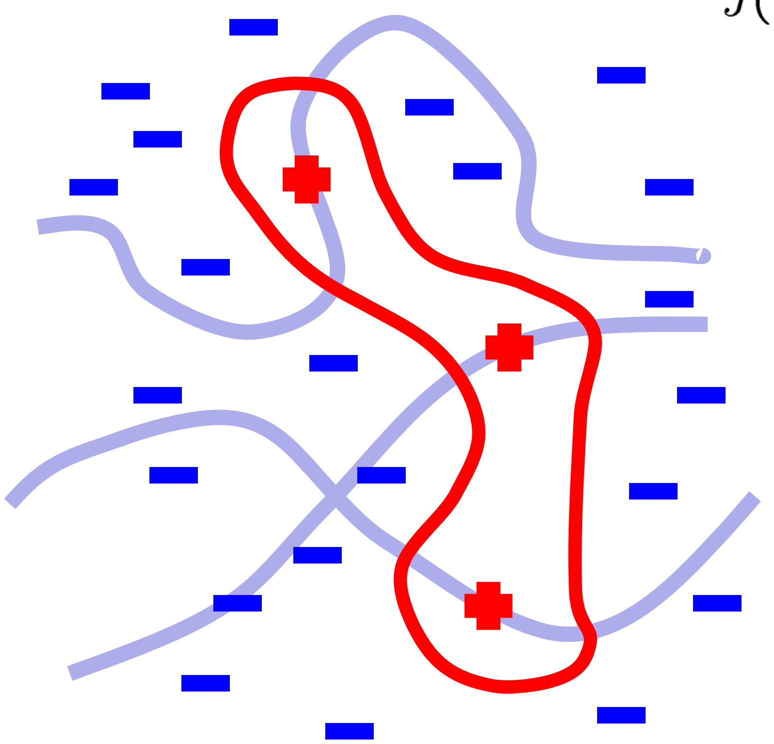
Random video samples: lots of them, very low chance to be positives

Action clustering

Formulation

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

Feature space



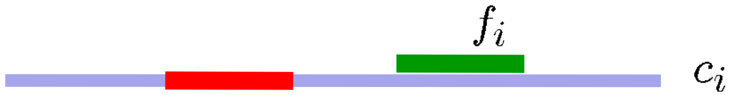
discriminative cost

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

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

x_i^- negative samples

$c_i[f_i]$ parameterized positive samples

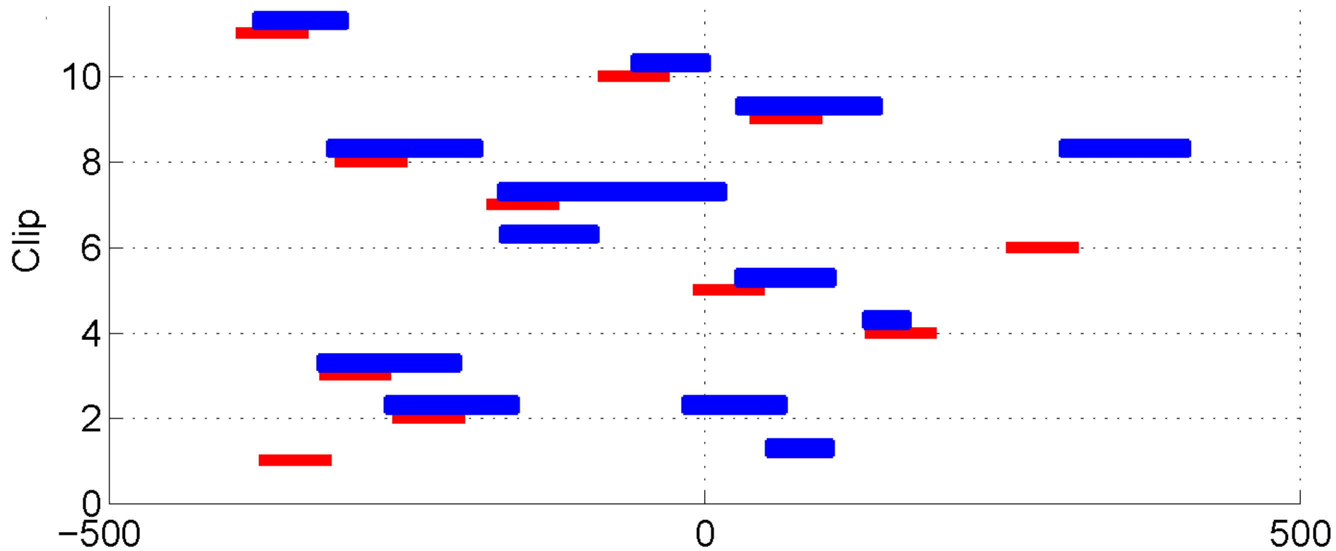


Optimization

SVM solution for w, b
Coordinate descent on f_i

Clustering results

Drinking actions in Coffee and Cigarettes

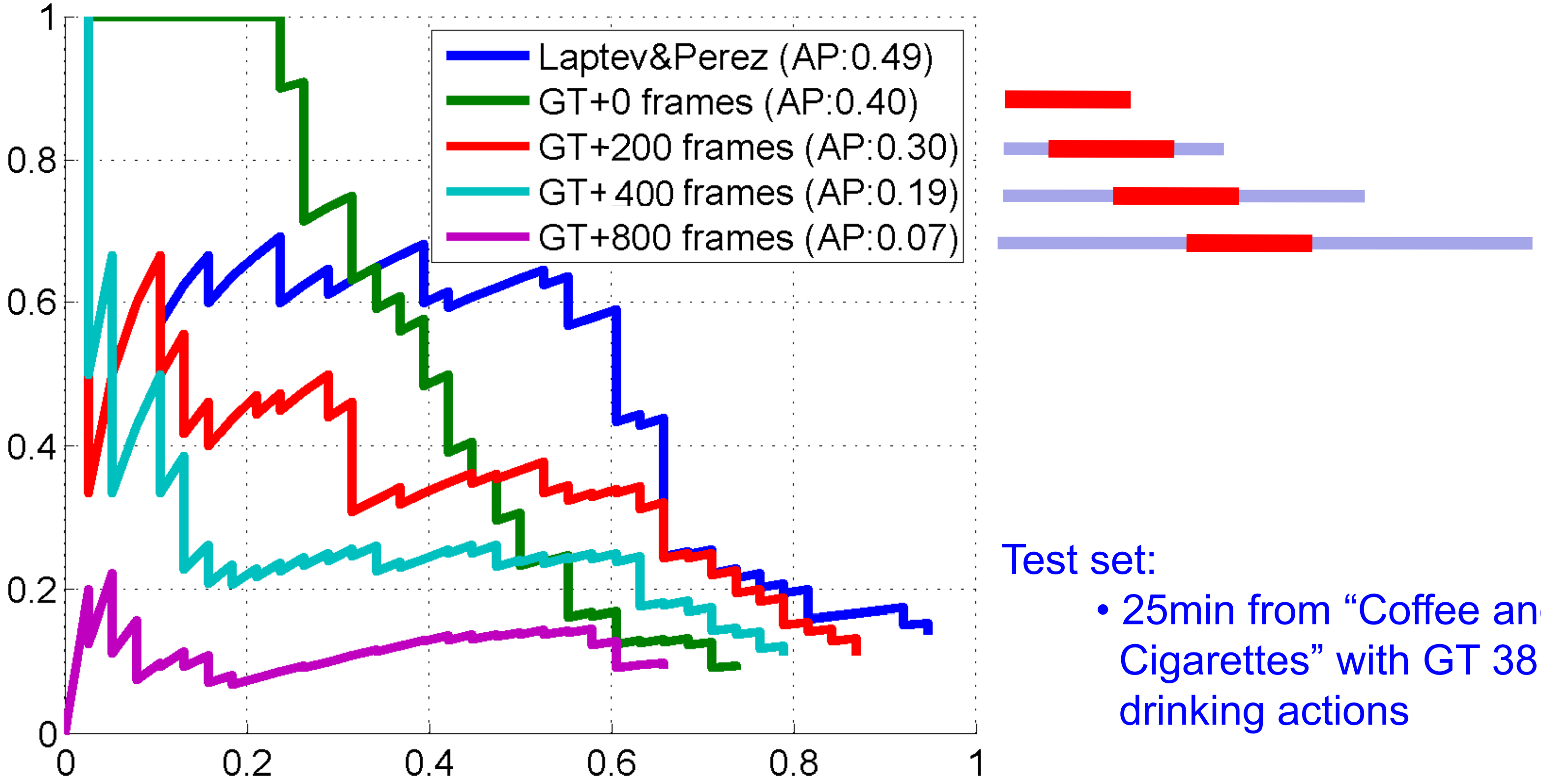


Detection results

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:

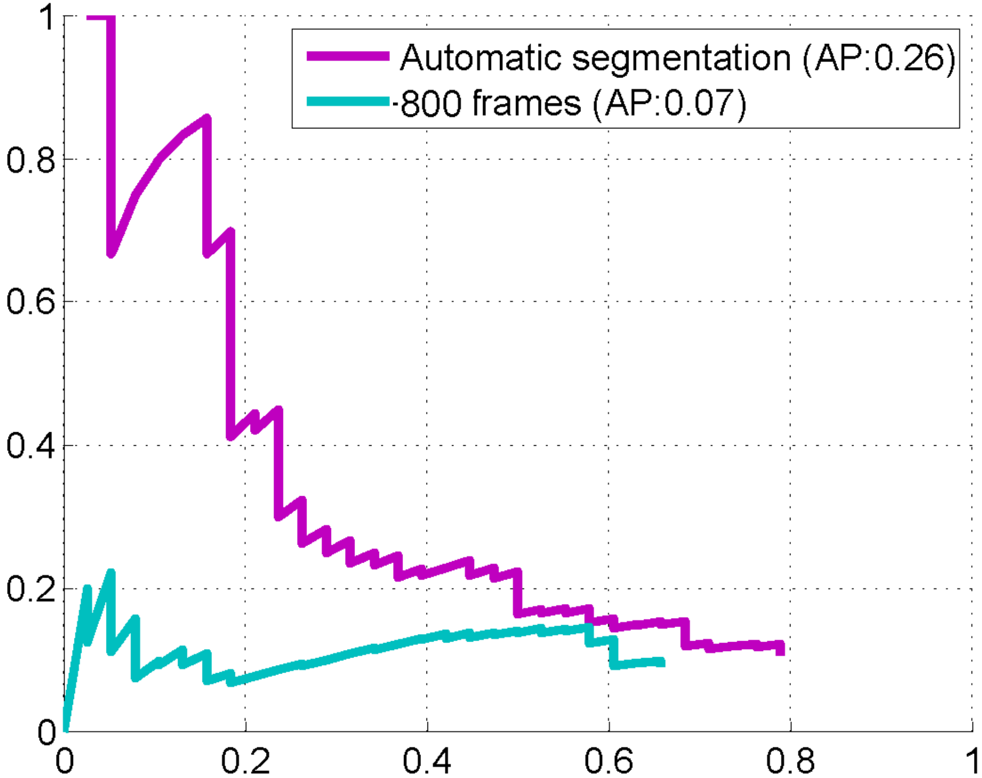
- 25min from "Coffee and Cigarettes" with GT 38 drinking actions

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*

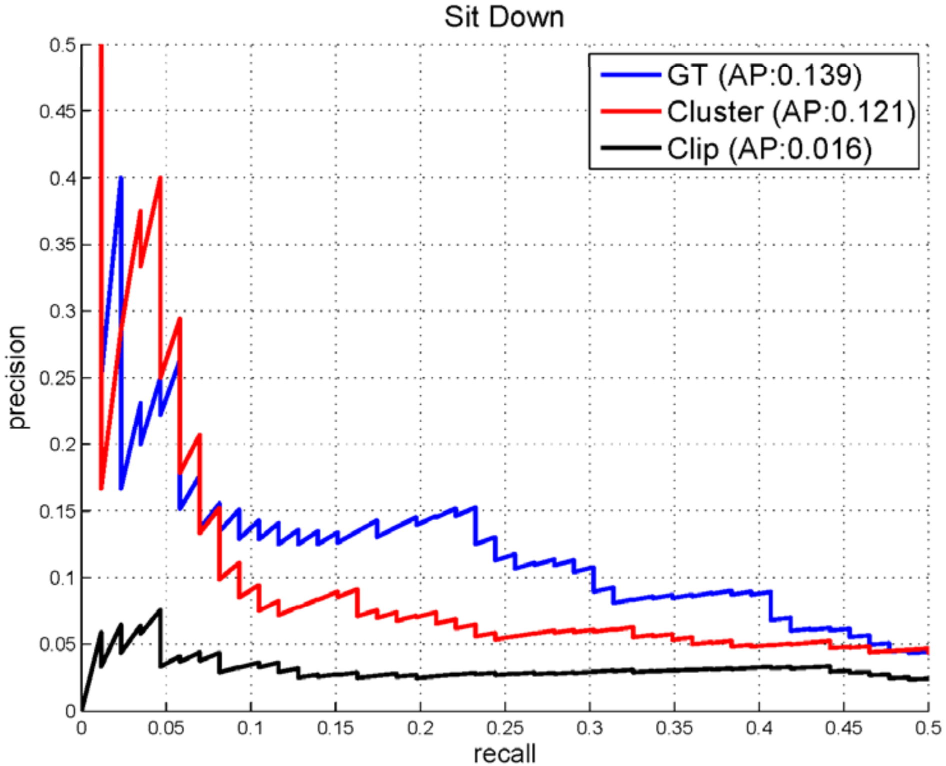


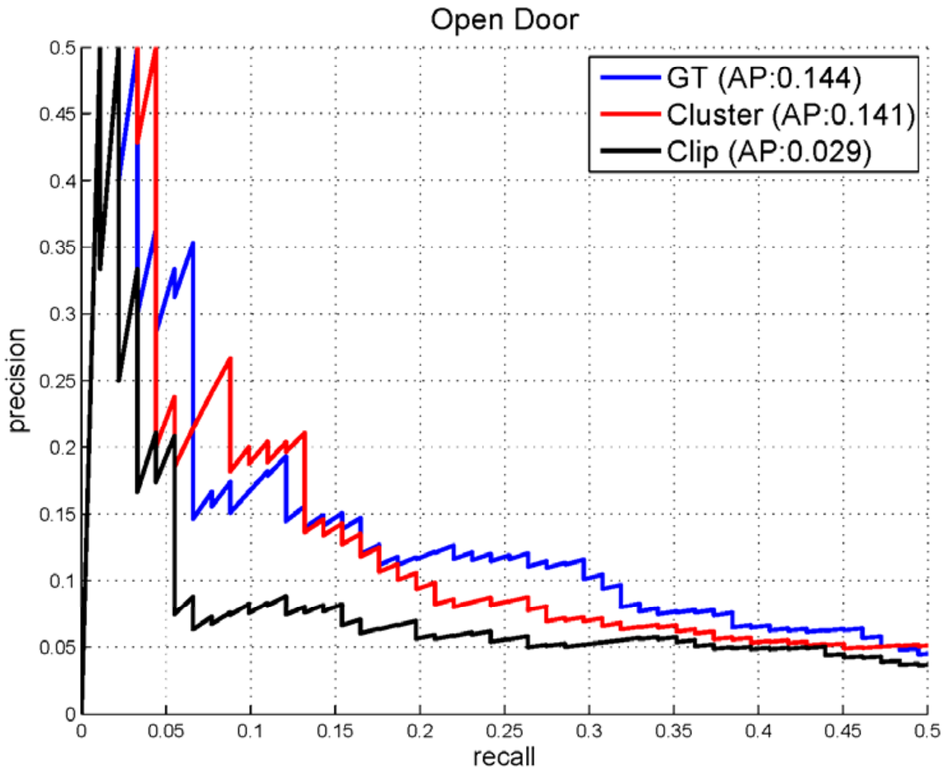
Test set:

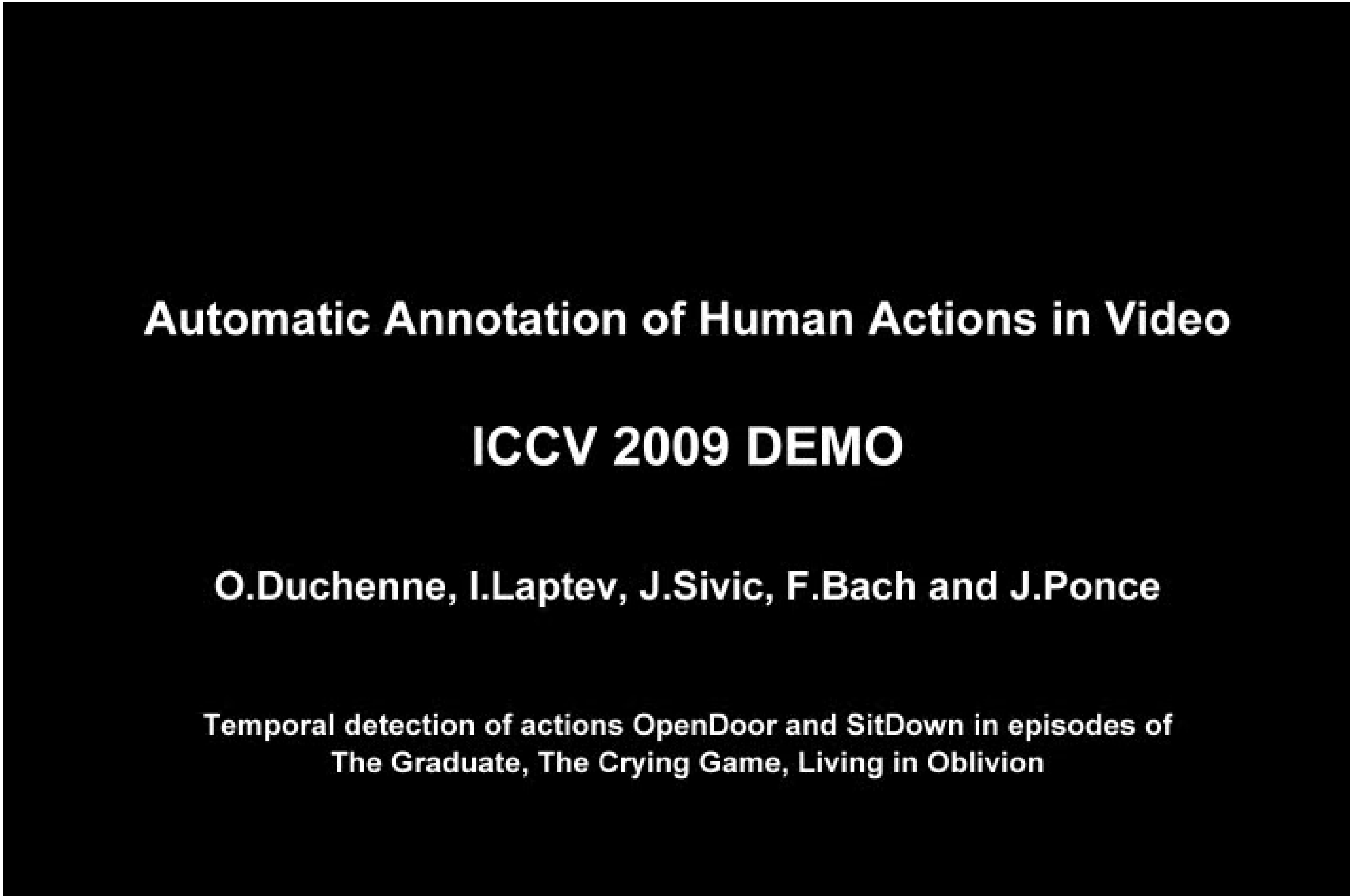
- 25min from "Coffee and Cigarettes" with GT 38 drinking actions

Detection results

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

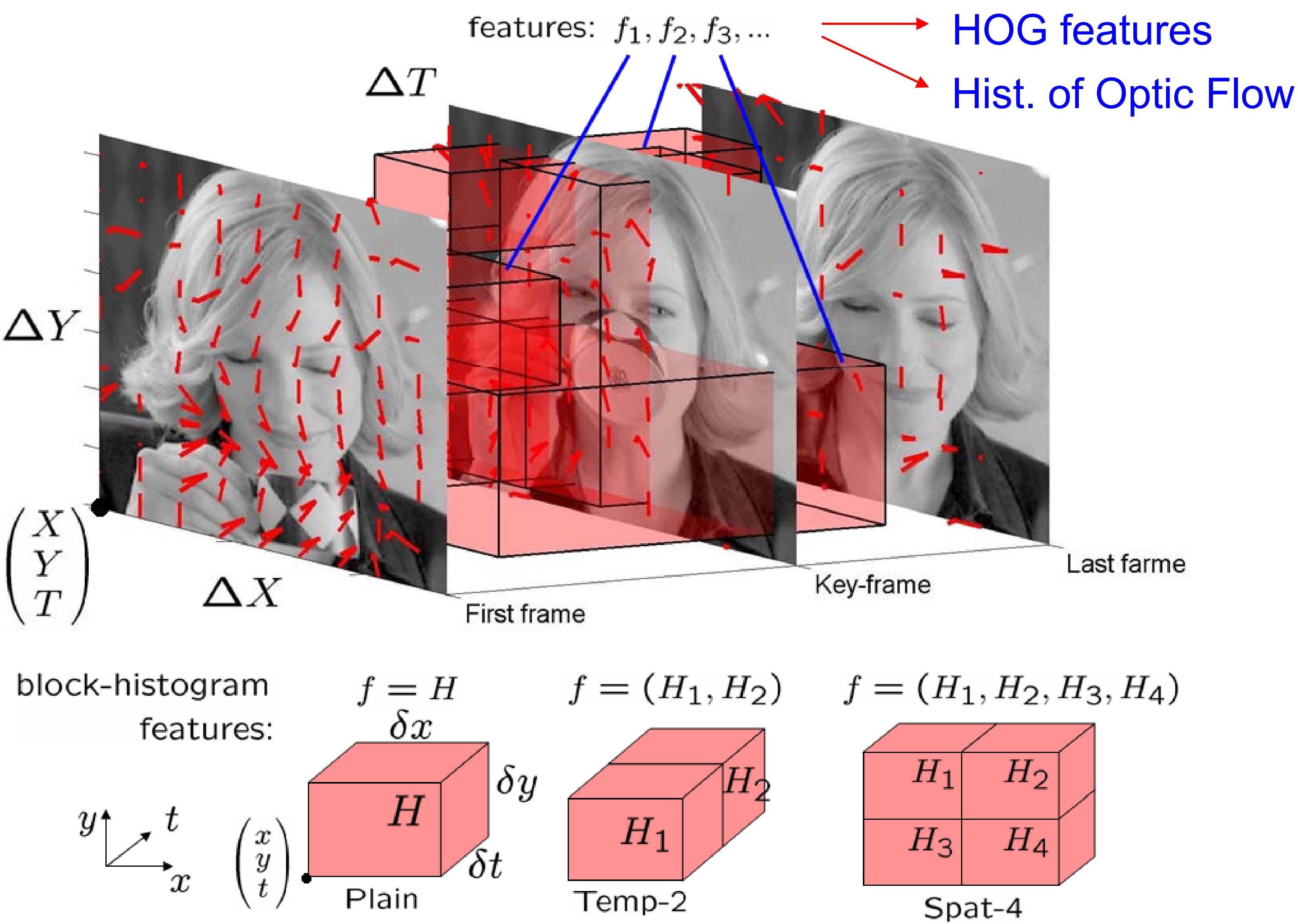




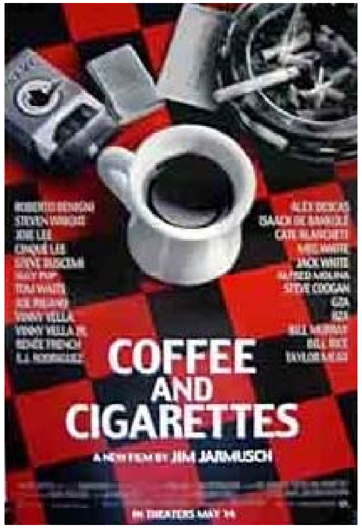


Temporal detection of “Sit Down” and “Open Door” actions in movies:
The Graduate, The Crying Game, Living in Oblivion

Actions as Space-Time Objects II



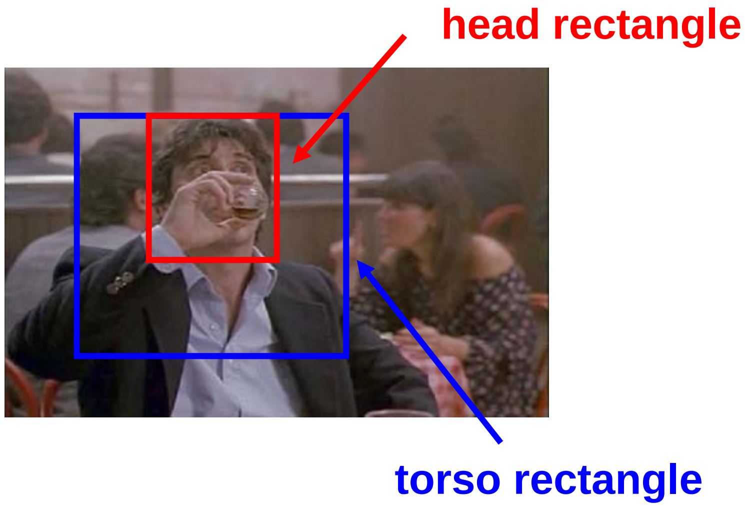
Action Dataset and Annotation



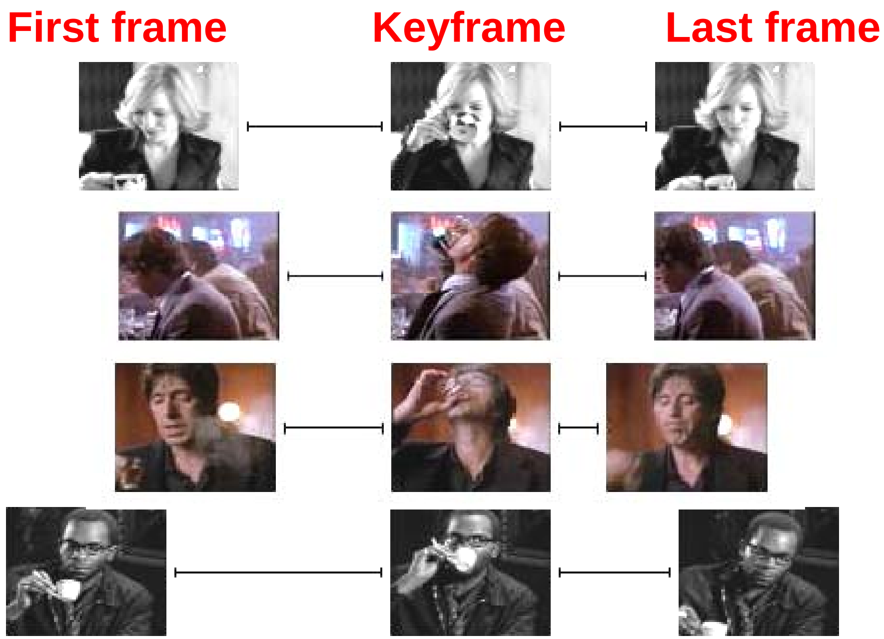
Manual annotation of drinking actions in movies:
"Coffee and Cigarettes"; "Sea of Love"

"*Drinking*": 159 annotated samples
"*Smoking*": 149 annotated samples

Spatial annotation



Temporal annotation

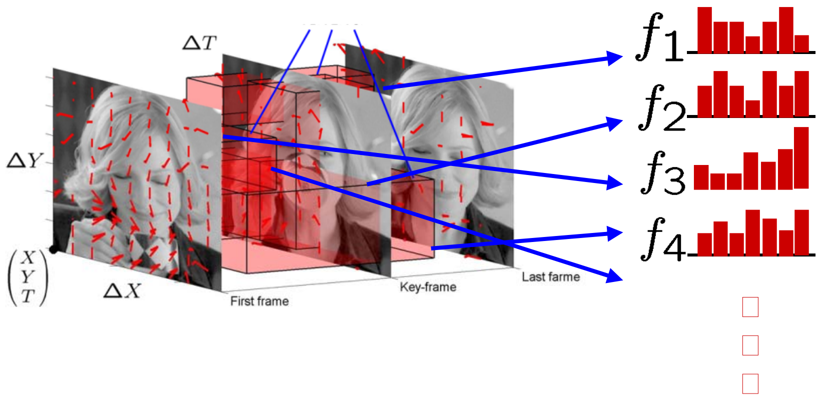


“Drinking” action samples

training samples

test samples

Action learning



boosting

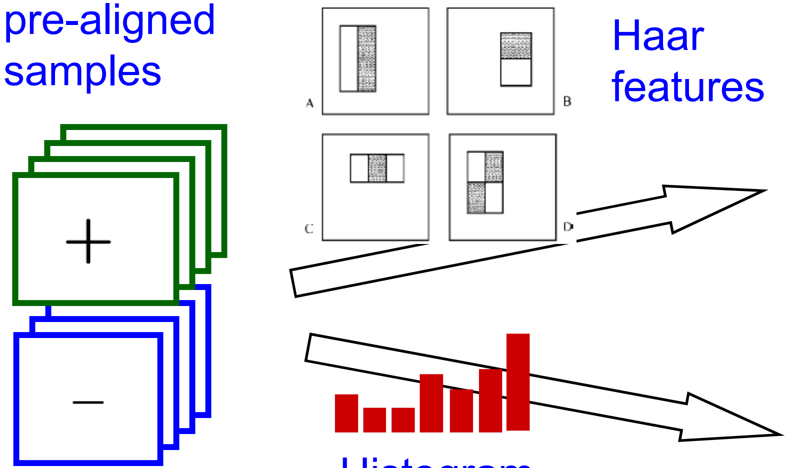
selected features

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

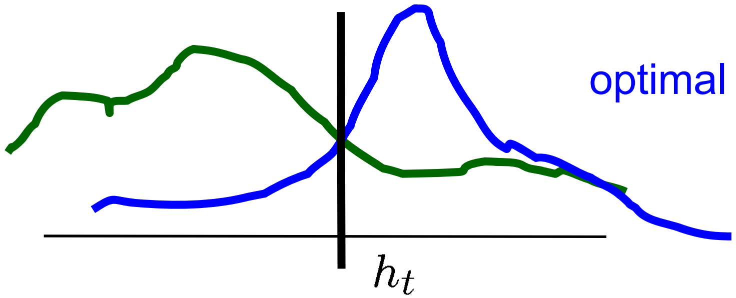
weak classifier

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

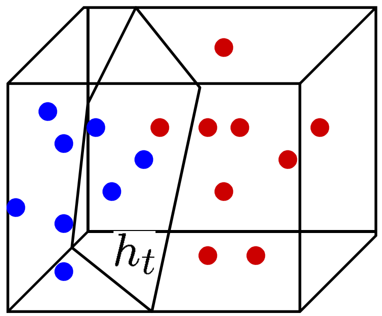
pre-aligned samples



Histogram features



optimal threshold

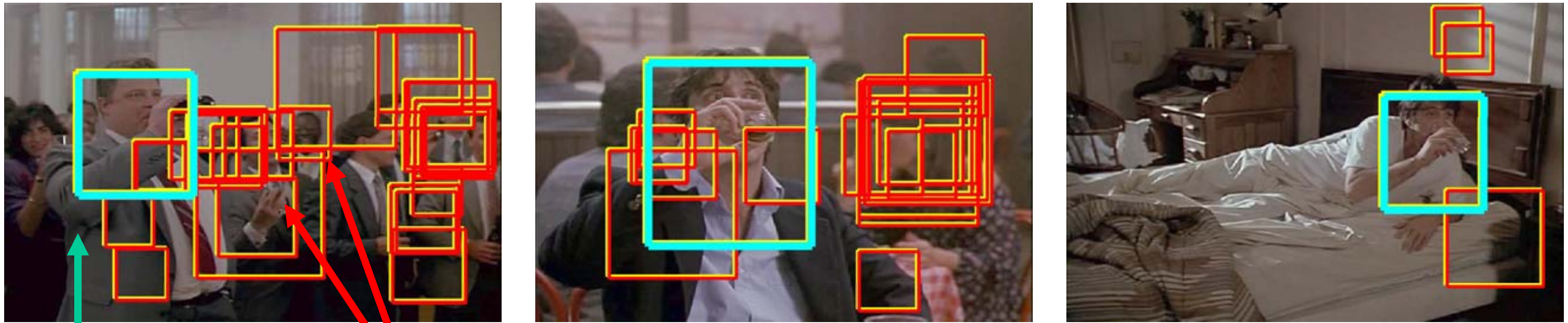


Fisher discriminant

see [Laptev BMVC'06] for more details

Keyframe priming

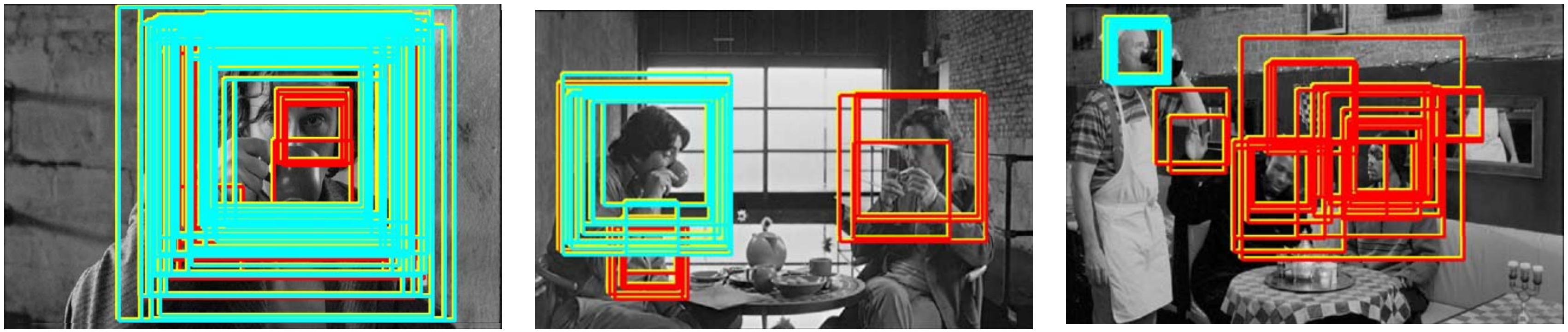
Training



Positive training sample

Negative training samples

Test



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