



# Motion and Human Actions II

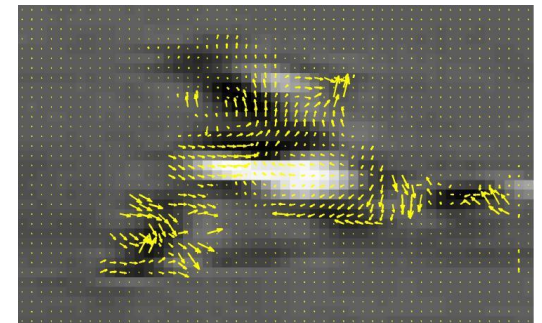
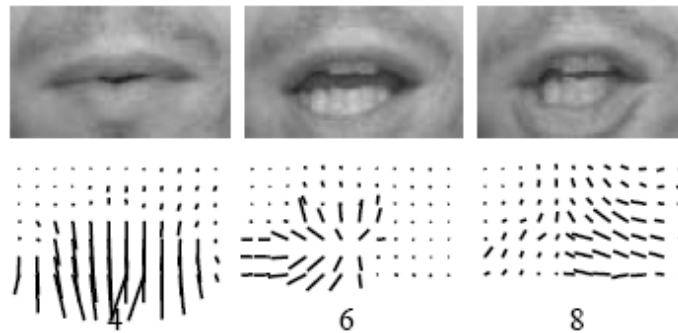
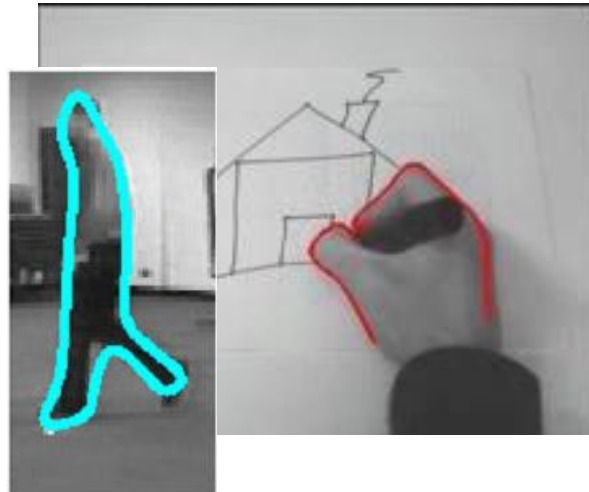
Ivan Laptev

*ivan.laptev@inria.fr*

INRIA, WILLOW, ENS/INRIA/CNRS UMR 8548

Laboratoire d'Informatique, Ecole Normale Supérieure, Paris

# Poses and actions so far:



# Space-time

No **global** assumptions  $\Rightarrow$

Consider **local** spatio-temporal neighborhoods

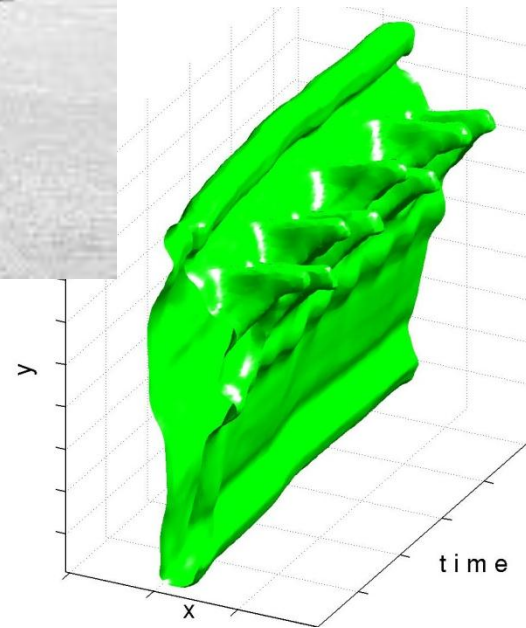
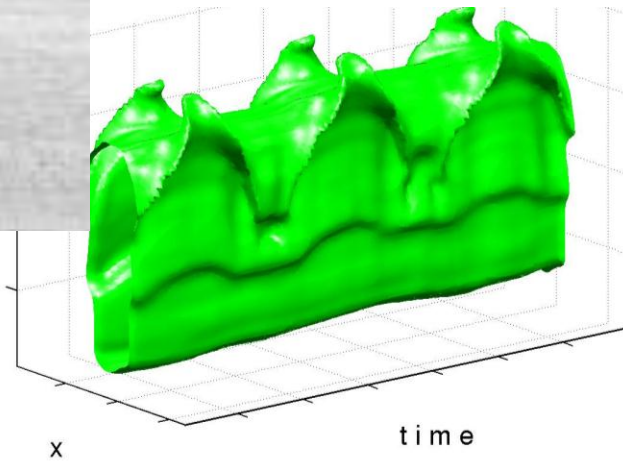


hand waving




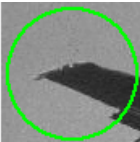

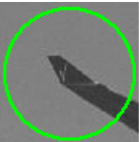









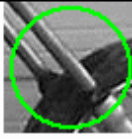





















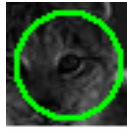




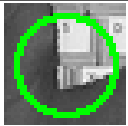


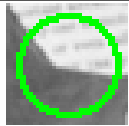
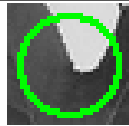



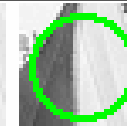





















boxing

# Actions == Space-time objects?

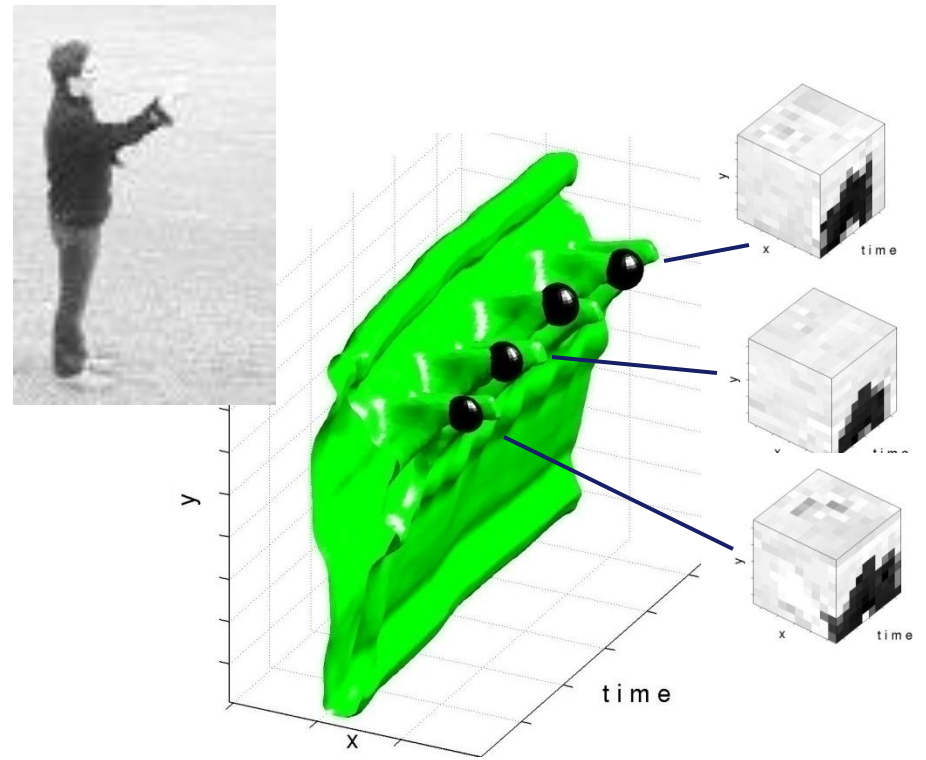
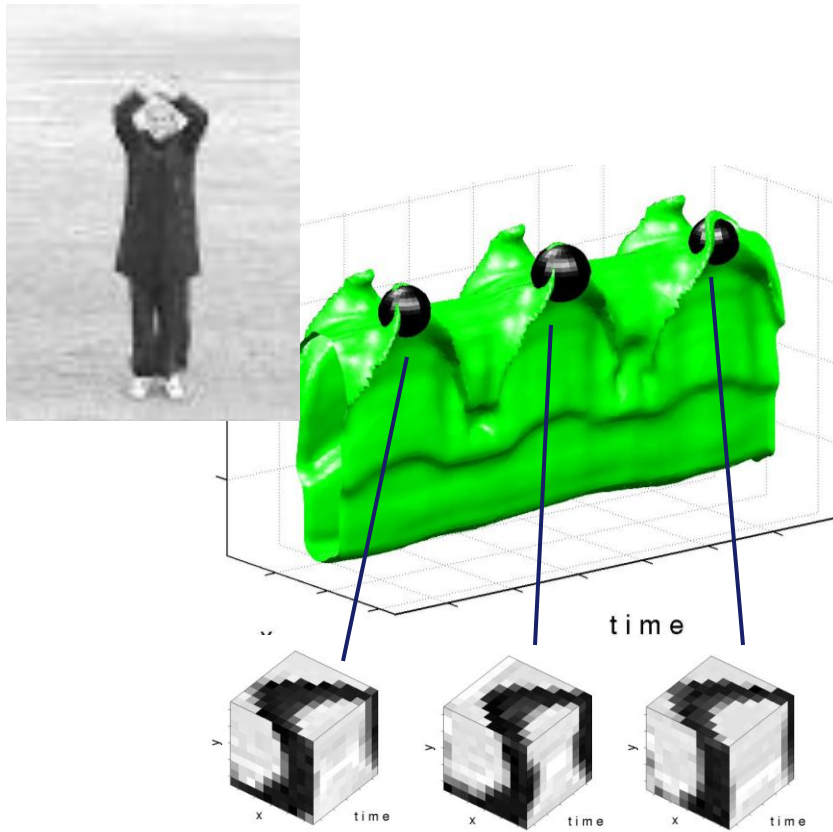




# Local approach: Bag of Visual Words

Airplanes	         
Motorbikes	         
Faces	         
Wild Cats	         
Leaves	         
People	         
Bikes	         

# 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  $\Sigma \in \text{SPSD}(3)$

$L_\xi(\cdot; \Sigma) = f(\cdot) * g_\xi(\cdot; \Sigma)$  Gaussian derivative of  $f$

$\nabla L = (L_x, L_y, L_t)^T$  Space-time gradient

$\mu(\cdot; \Sigma) = \nabla L(\cdot; \Sigma)(\nabla L(\cdot; \Sigma))^T * g(\cdot; s\Sigma) =$   
Second-moment matrix

$$\begin{pmatrix} \mu_{xx} & \mu_{xy} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix}$$

# Space-Time Interest Points: Detection

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

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

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

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

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

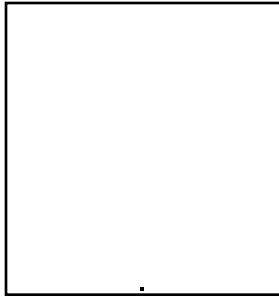
Large eigenvalues of  $\mu$  can be detected by the local maxima of  $H$  over  $(x,y,t)$ :

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

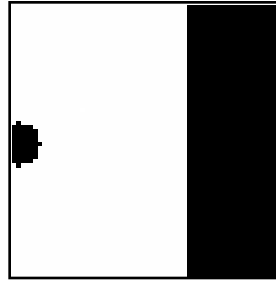
(similar to Harris operator [Harris and Stephens, 1988])

# Space-Time interest points

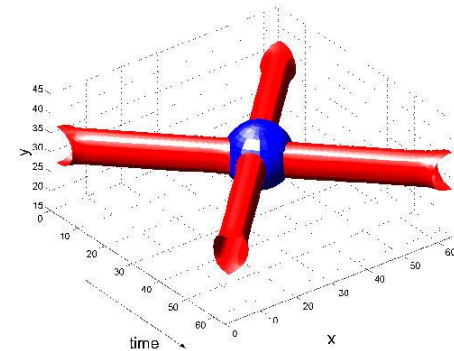
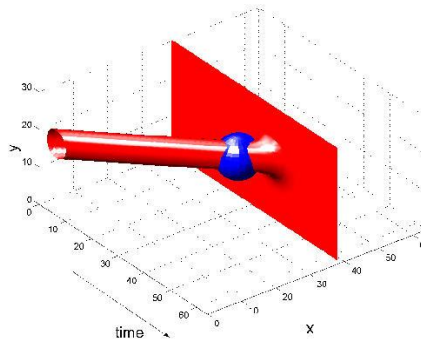
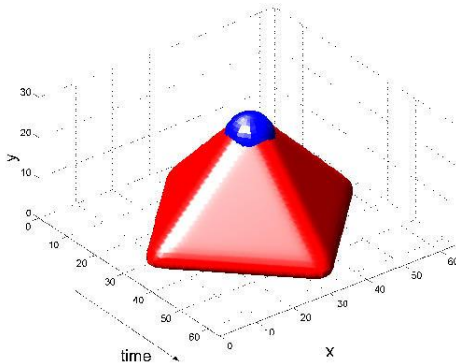
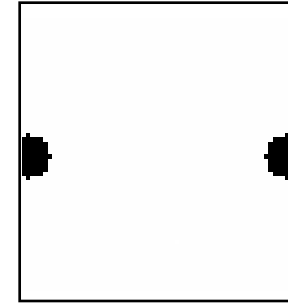
Velocity  
changes



appearance/  
disappearance



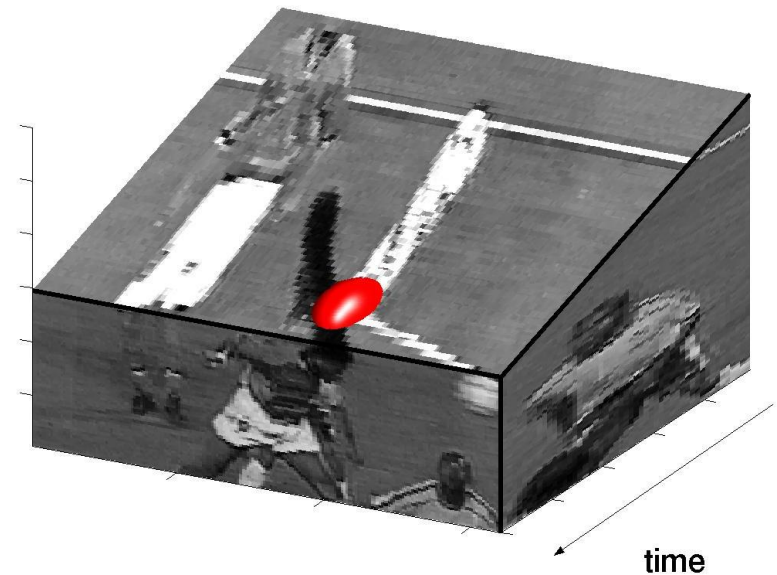
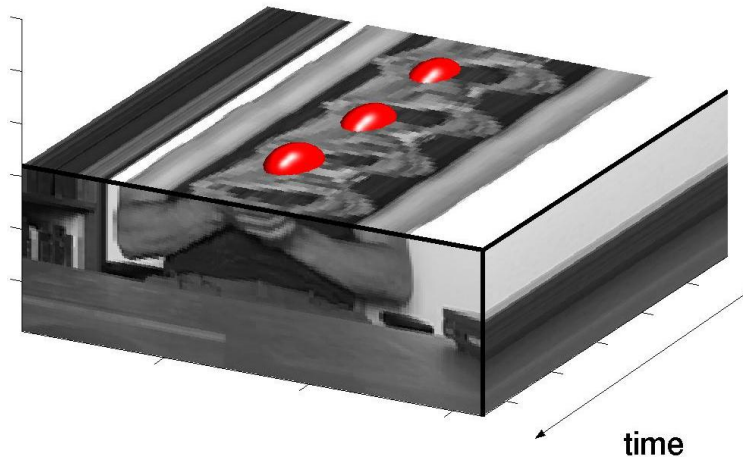
split/merge





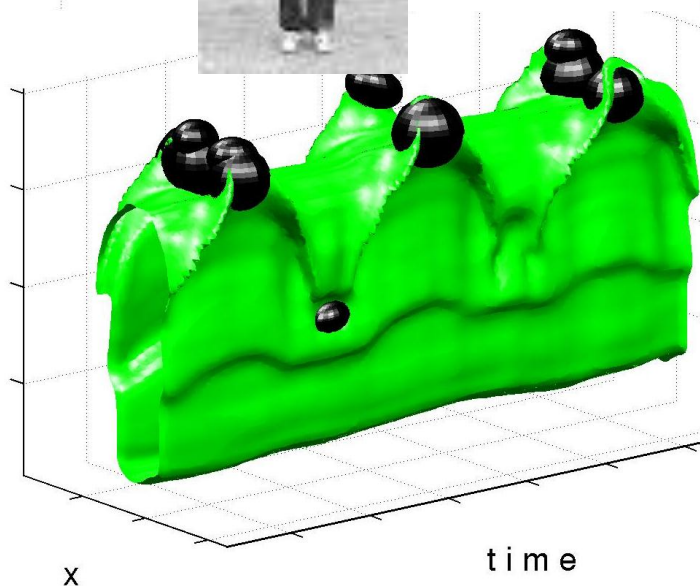
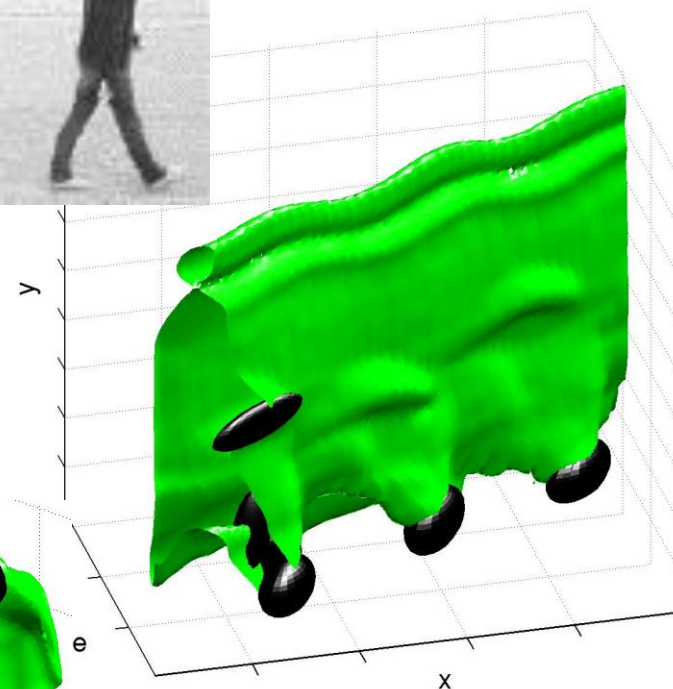
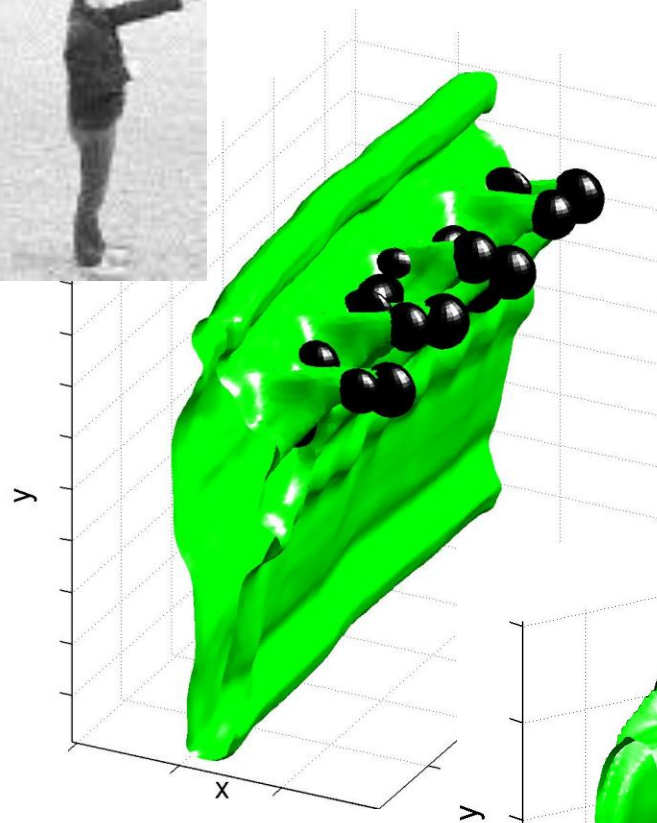
# Space-Time Interest Points: Examples

Motion event detection

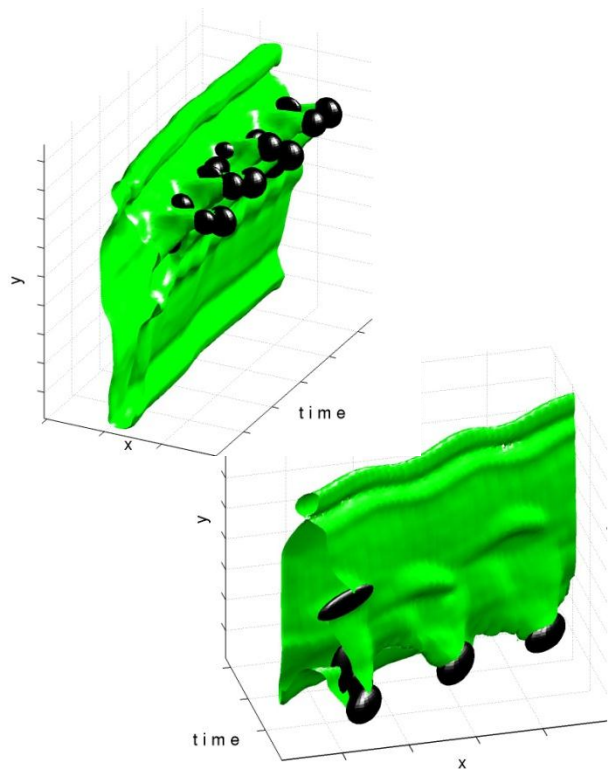




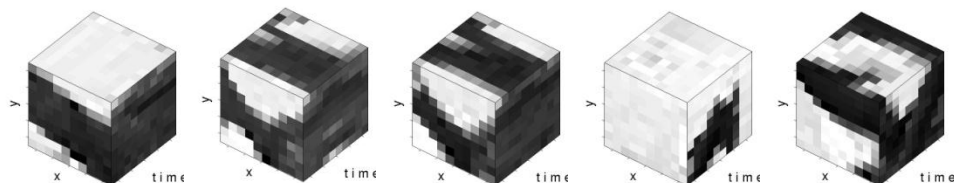
# Local features for human actions



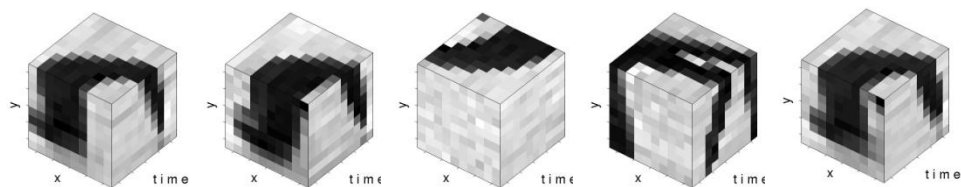
# Local features for human actions



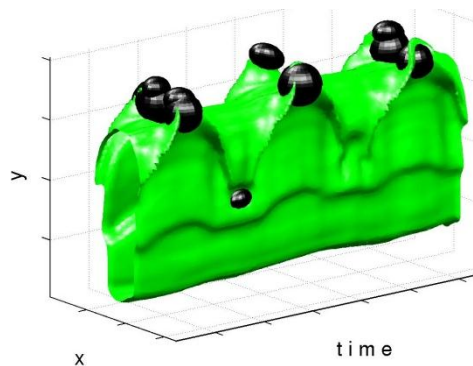
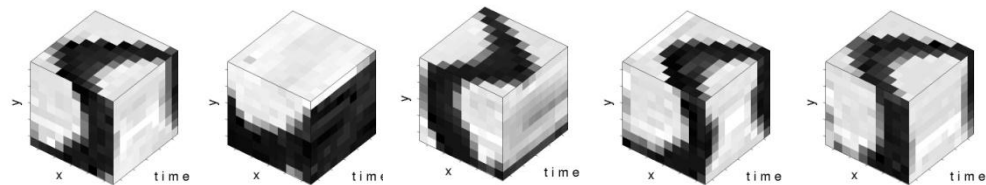
boxing



walking



hand waving



# Local space-time descriptor: Jet

**Local jet** descriptor [Koenderink and van Doorn, 1987]:  
spatio-temporal Gaussian derivatives at interest points  $p$ :

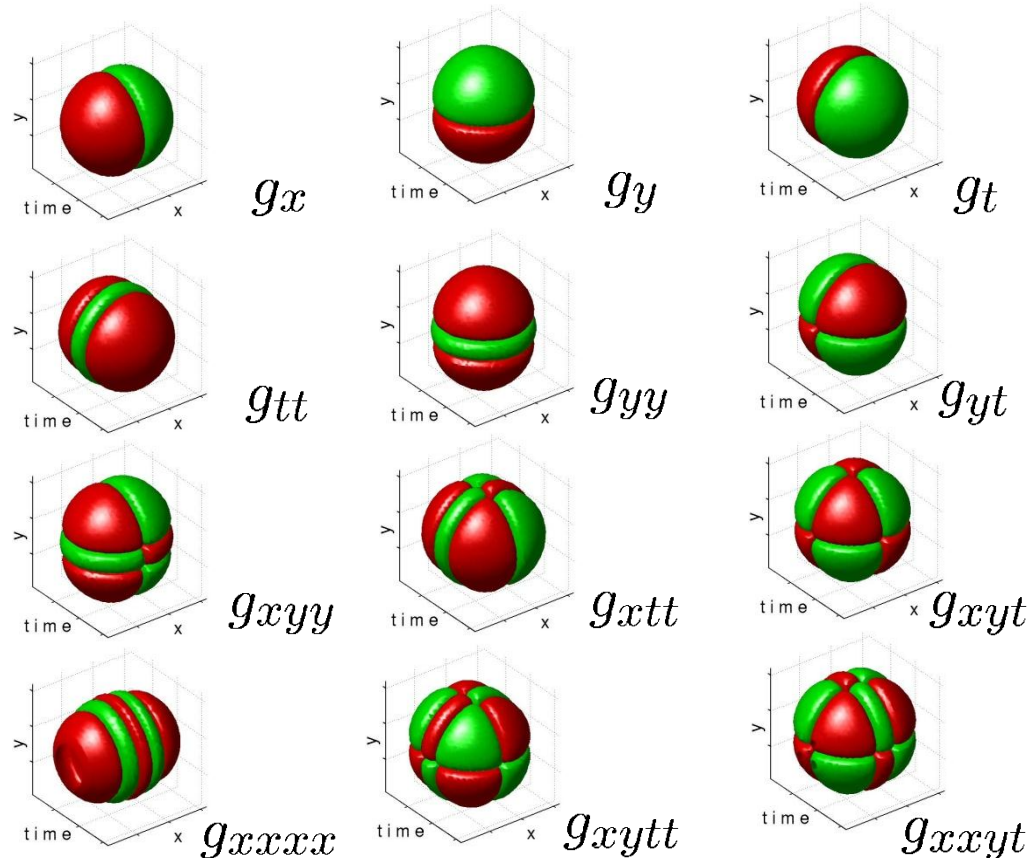
$$D(p) = (L_x(p), L_y(p), L_t(p), L_{xx}(p), \dots, L_{tttt}(p))$$

$$L_x(p) = \sum_q f(p - q)g_x(q)$$

$$L_y(p) = \sum_q f(p - q)g_y(q)$$

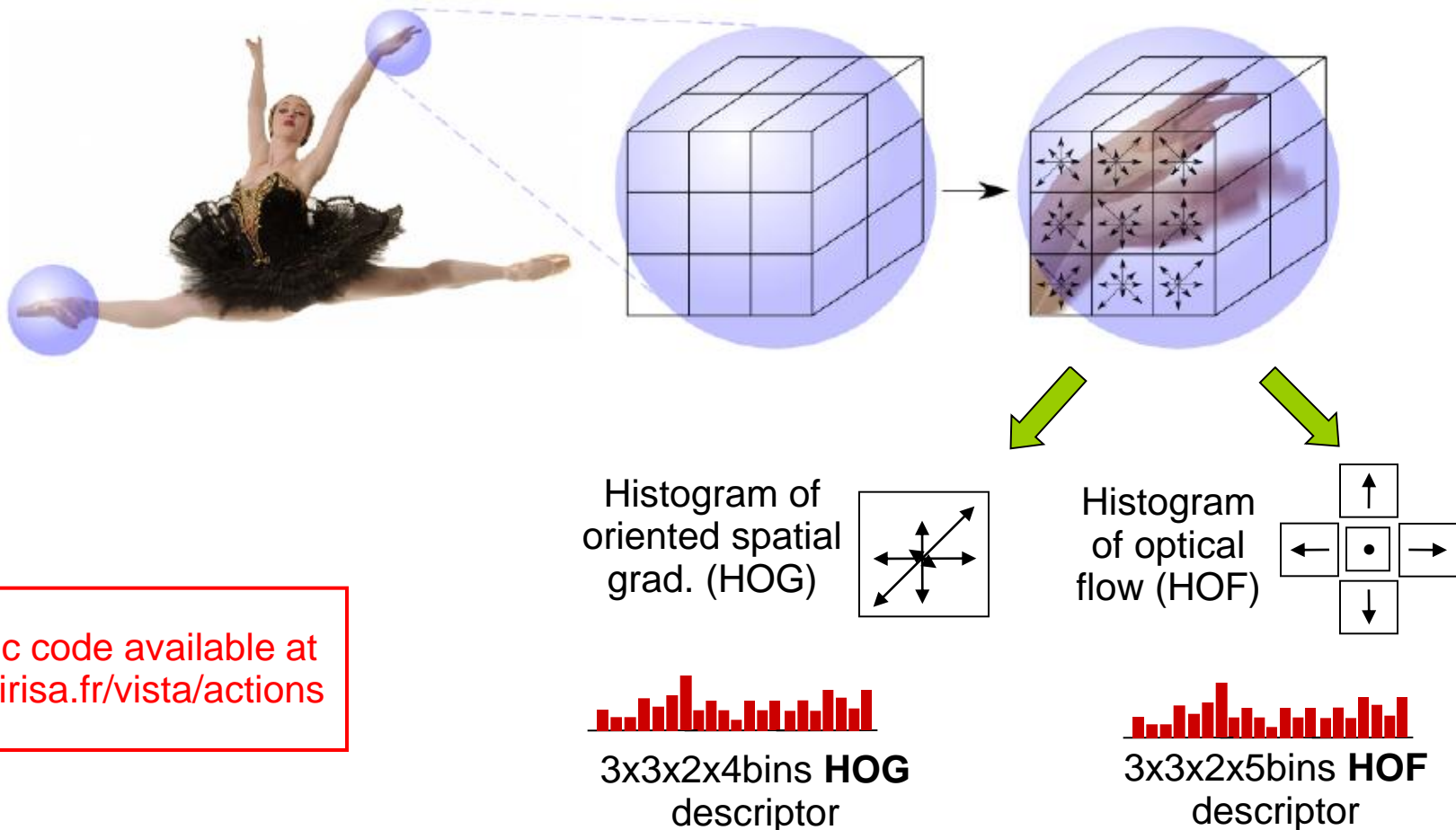
•  
•  
•

$$L_{tttt}(p) = \sum_q f(p - q)g_{tttt}(q)$$



# Local space-time descriptor: HOG/HOF

Multi-scale space-time patches

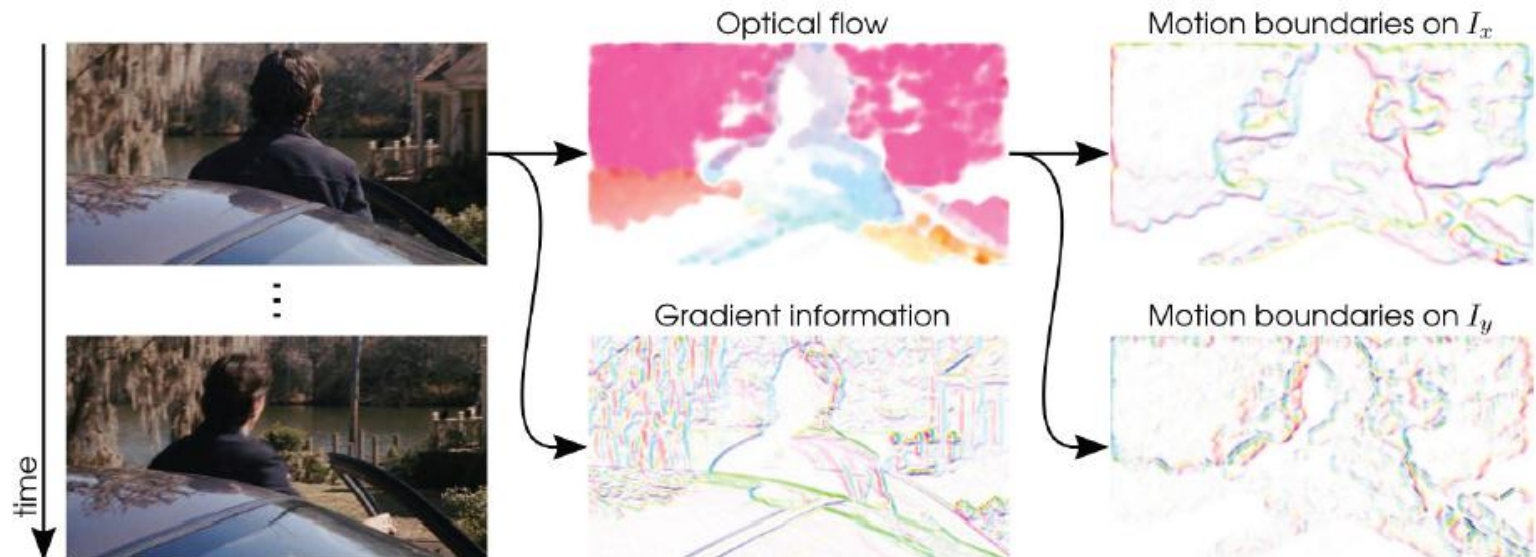
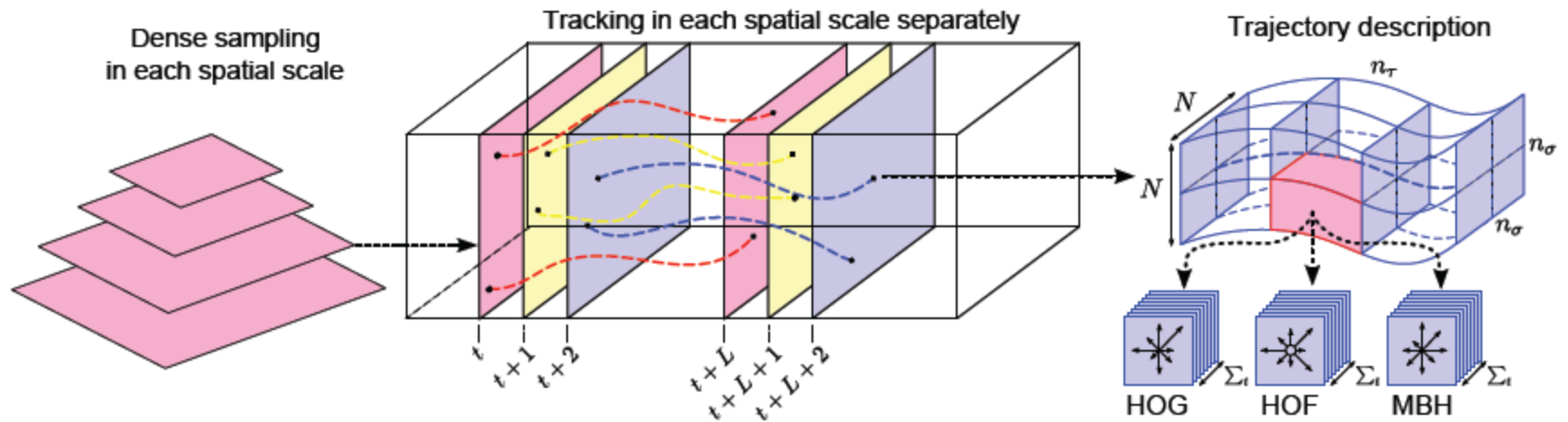


Public code available at  
[www.irisa.fr/vista/actions](http://www.irisa.fr/vista/actions)



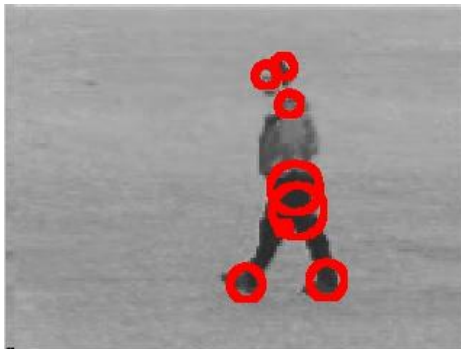
# Dense trajectory descriptors

[Wang et al. CVPR'11]

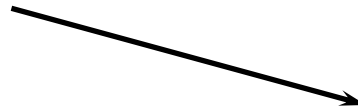



# Visual Vocabulary: K-means clustering


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




Clustering

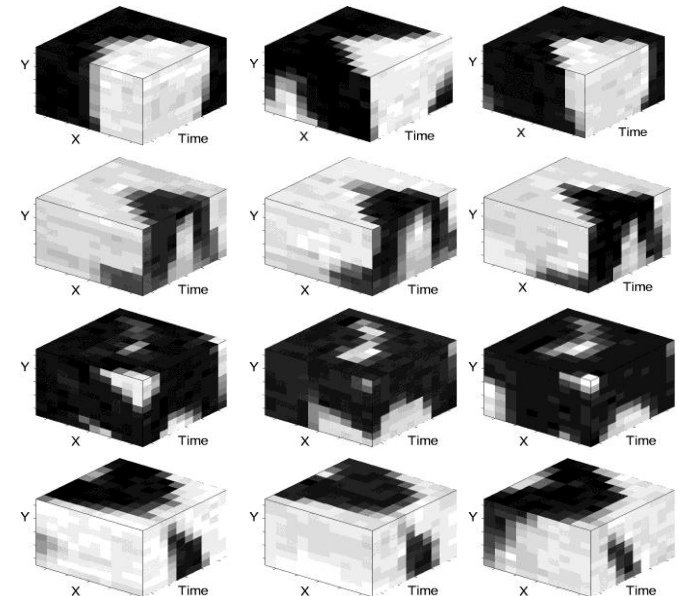


c1  


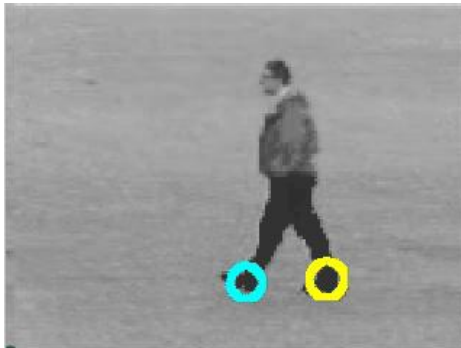
c2  


c3  


c4  

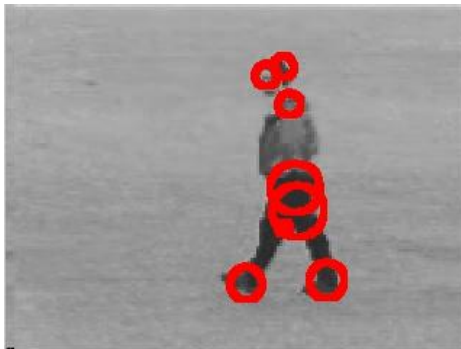
Classification





# Visual Vocabulary: K-means clustering

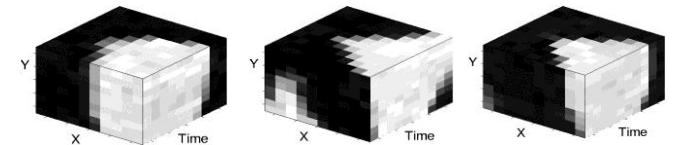
- Group similar points in the space of image descriptors using K-means clustering
- Select significant clusters



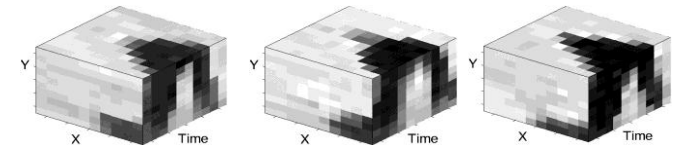
Clustering



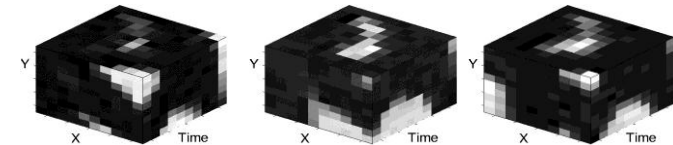
c1



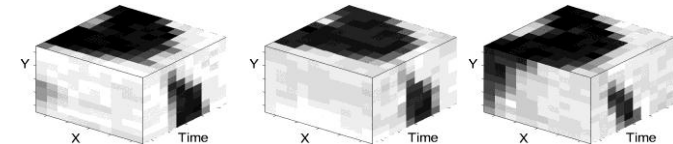
c2



c3



c4

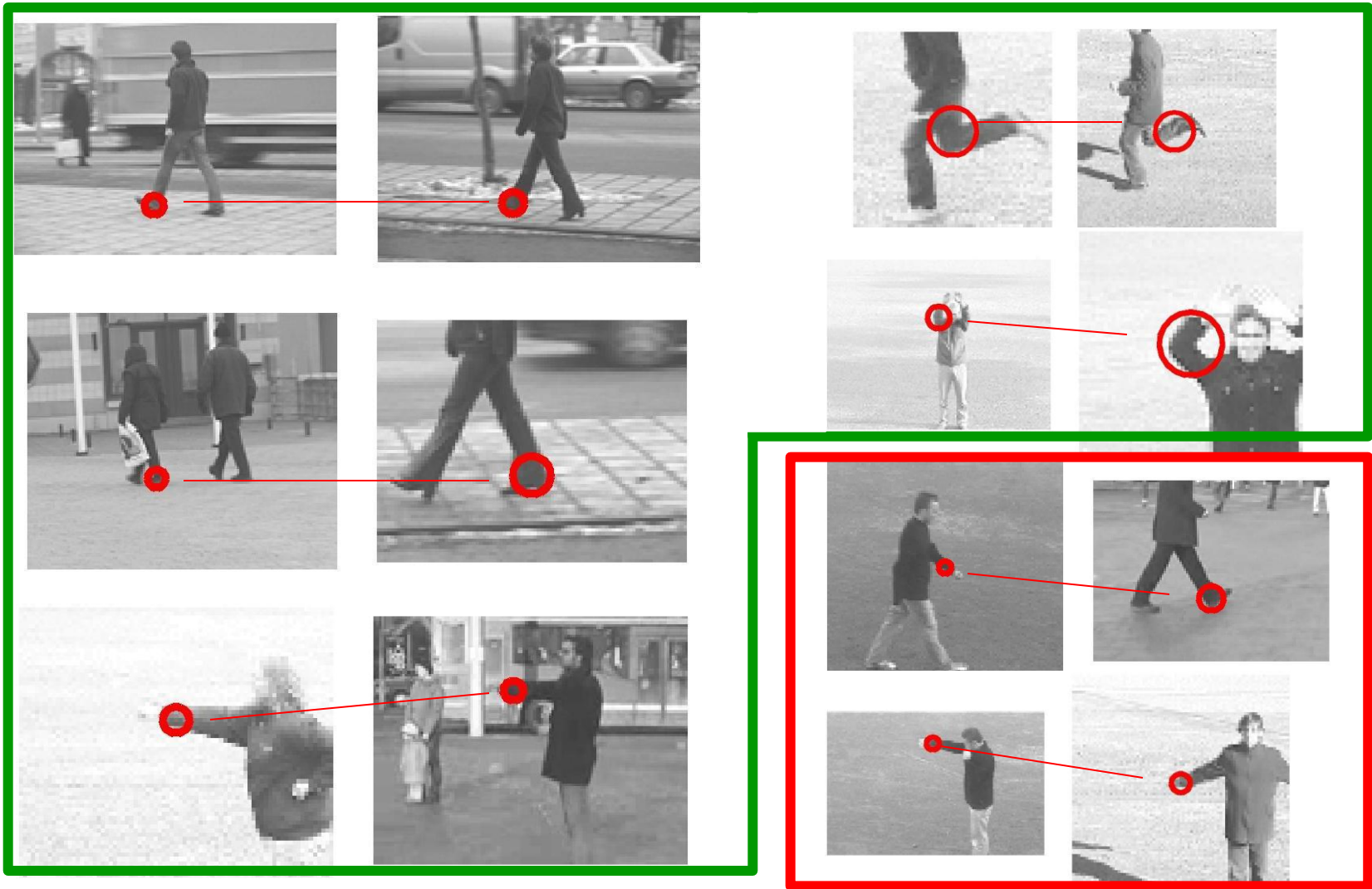


Classification



# Local feature methods: Matching

- Finds similar events in pairs of video sequences



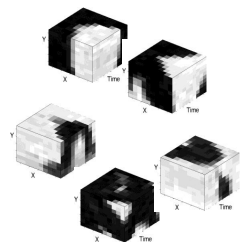
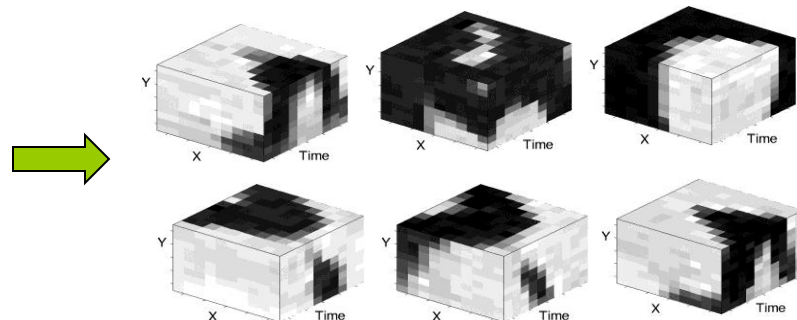
# Action Classification: Overview

Bag of space-time features + multi-channel SVM

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

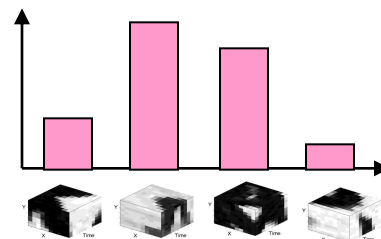


Collection of space-time patches



HOG & HOF  
patch  
descriptors

Histogram of visual words



Multi-channel  
SVM  
Classifier

# Action recognition in KTH dataset

Walking

Jogging

Running

Boxing

Waving

Clapping



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

# Classification results on KTH dataset

	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	.99	.01	.00	.00	.00	.00
Jogging	.04	.89	.07	.00	.00	.00
Running	.01	.19	.80	.00	.00	.00
Boxing	.00	.00	.00	.97	.00	.03
Waving	.00	.00	.00	.00	.91	.09
Clapping	.00	.00	.00	.05	.00	.95

Confusion matrix for KTH actions

# Evaluation of local feature detectors and descriptors

## Four types of detectors:

- Harris3D [Laptev 2003]
- Cuboids [Dollar et al. 2005]
- Hessian [Willems et al. 2008]
- Regular dense sampling

## Four types of descriptors:

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

## Three human actions datasets:

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



# Space-time feature detectors

Harris3D



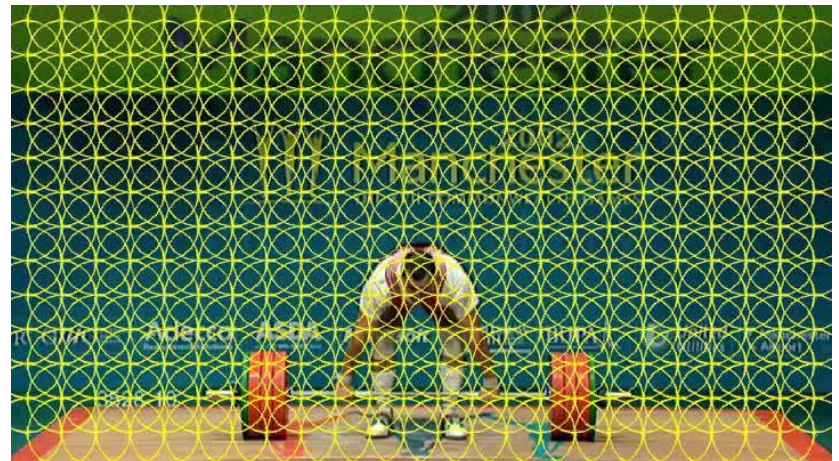
Hessian



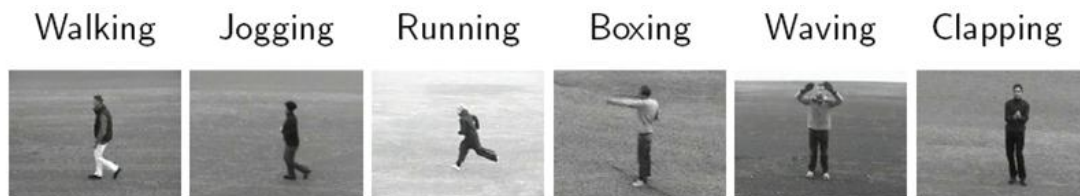
Cuboids



Dense



# Results on KTH Actions



6 action classes, 4 scenarios, staged

## Detectors

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

(Average accuracy scores)

- Best results for **sparse** Harris3D + HOF
- Dense features perform relatively poor compared to sparse features

# Results on UCF Sports



10 action classes, videos from TV broadcasts

## Detectors

Descriptors

	Harris3D	Cuboids	Hessian	Dense
HOG3D	79.7%	82.9%	79.0%	<b>85.6%</b>
HOG/HOF	78.1%	77.7%	79.3%	81.6%
HOG	71.4%	72.7%	66.0%	77.4%
HOF	75.4%	76.7%	75.3%	82.6%
Cuboids	-	76.6%	-	-
E-SURF	-	-	77.3%	-

(Average precision scores)

- Best results for **dense + HOG3D**

# Results on Hollywood-2



12 action classes collected from 69 movies

## Detectors

Descriptors

	Harris3D	Cuboids	Hessian	Dense
<b>HOG3D</b>	43.7%	45.7%	41.3%	45.3%
<b>HOG/HOF</b>	45.2%	46.2%	46.0%	<b>47.4%</b>
<b>HOG</b>	32.8%	39.4%	36.2%	39.4%
<b>HOF</b>	43.3%	42.9%	43.0%	45.5%
<b>Cuboids</b>	-	45.0%	-	-
<b>E-SURF</b>	-	-	38.2%	-

(Average precision scores)

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



# What about 3D?

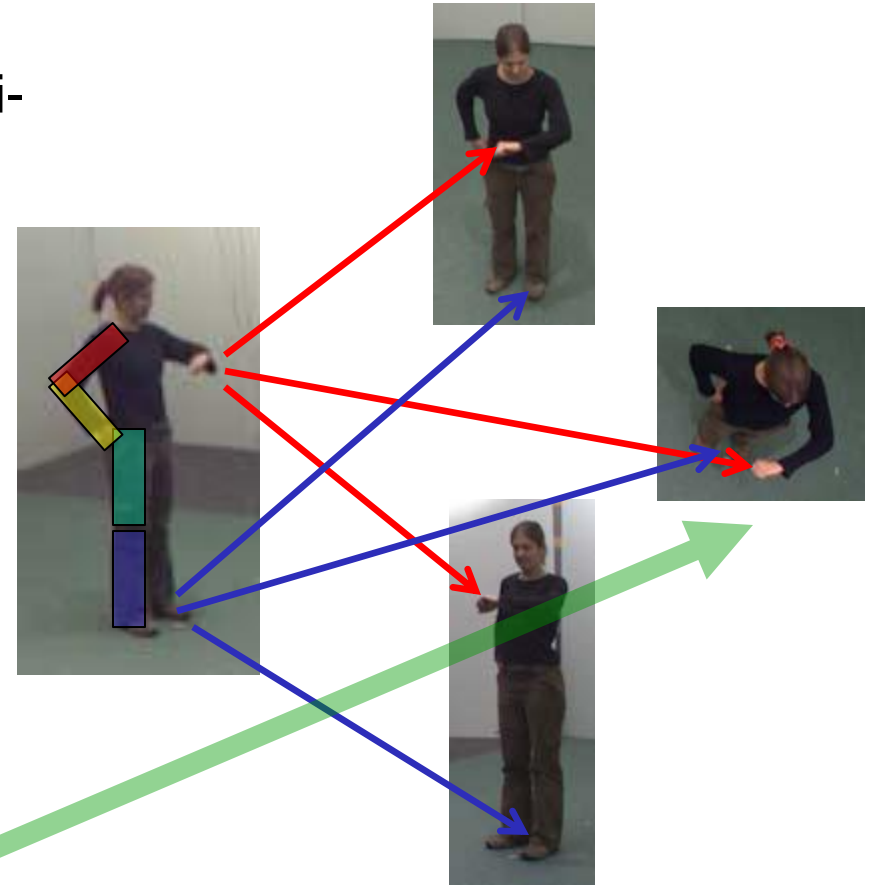
Local motion and appearance features are **not invariant** to view changes



# Multi-view action recognition

**Difficult to apply standard multi-view methods:**

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



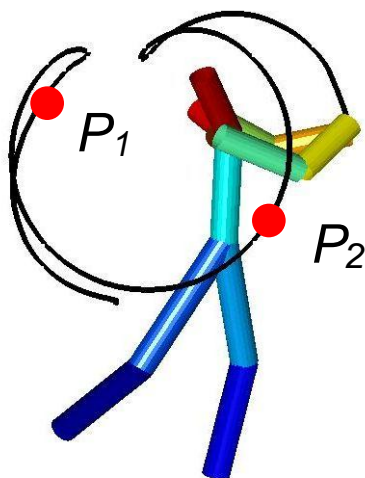


# Temporal self-similarities

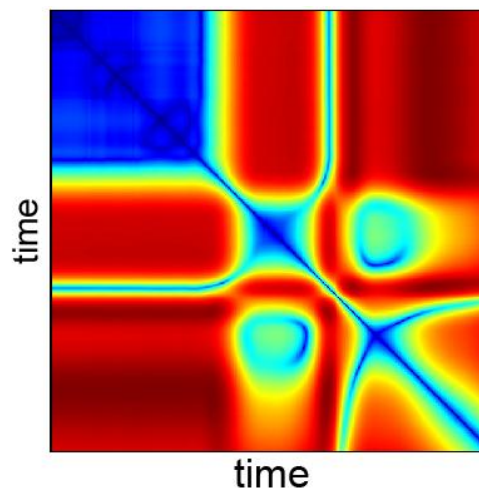
## Idea:

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

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

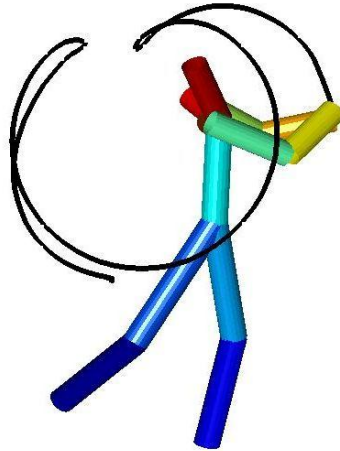


Distance matrix / self-similarity matrix (SSM):

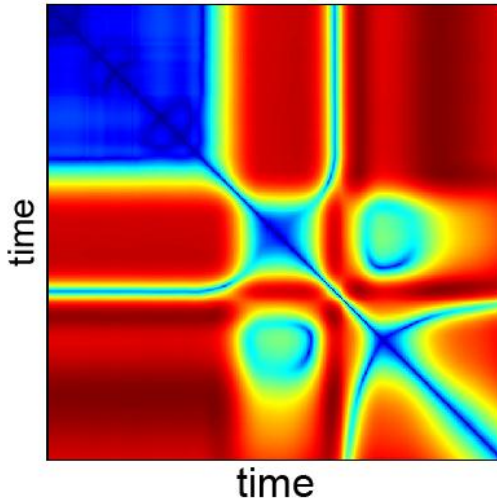
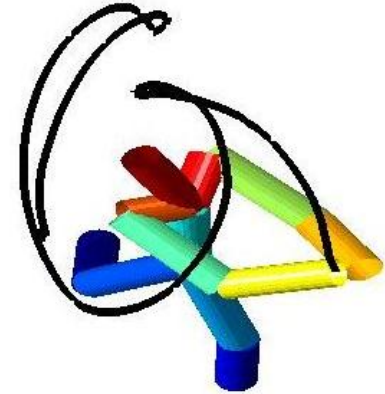


# Temporal self-similarities: Multi-views

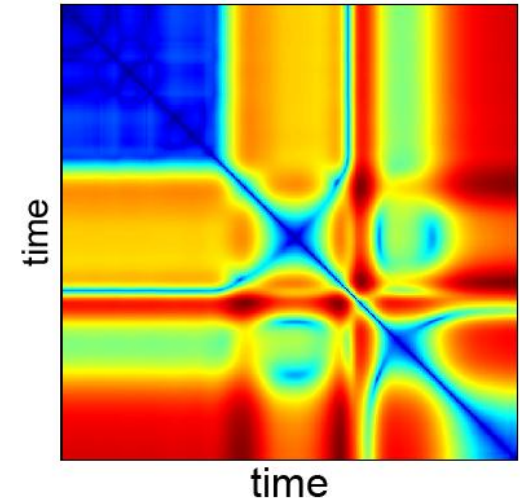
Side view



Top view



Appear  
very  
similar  
despite  
the view  
change!



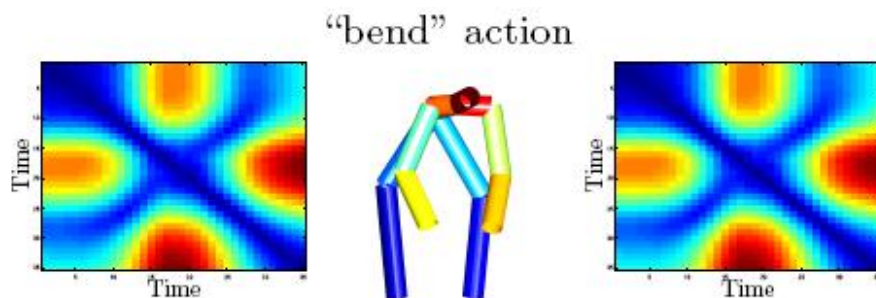
- Intuition:
1. Distance between similar poses is low in any view
  2. Distance among different poses is likely to be large in most views

# Temporal self-similarities: MoCap

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



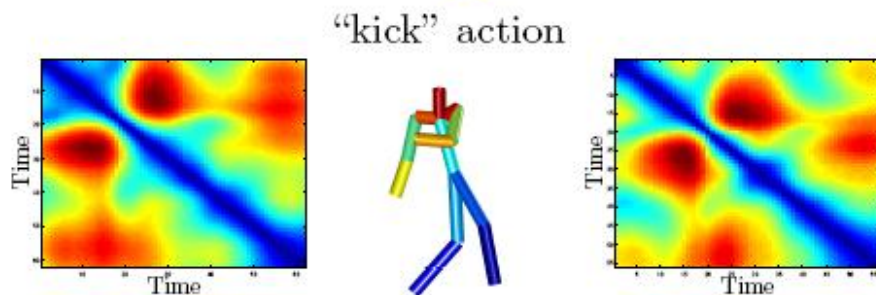
person 1



person 2



person 1

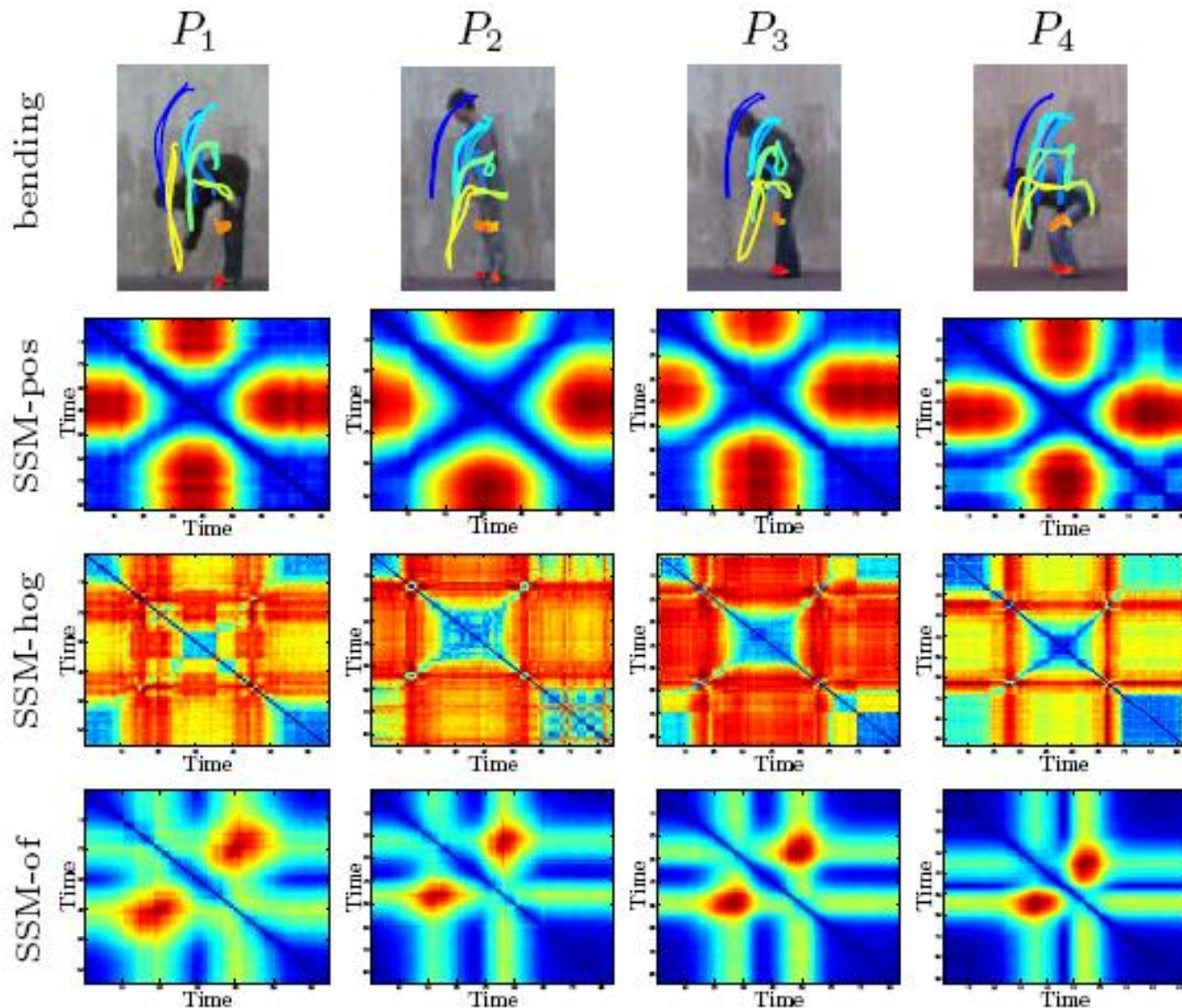


person 2





# Temporal self-similarities: Video



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

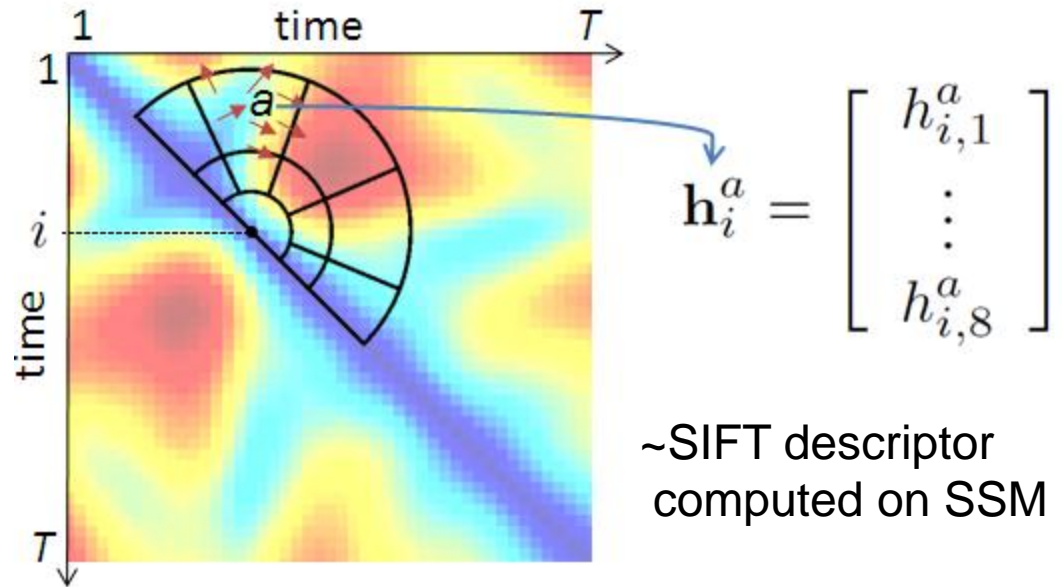
# Self-similarity descriptor

## Goal:

define a quantitative measure to compare self-similarity matrices

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

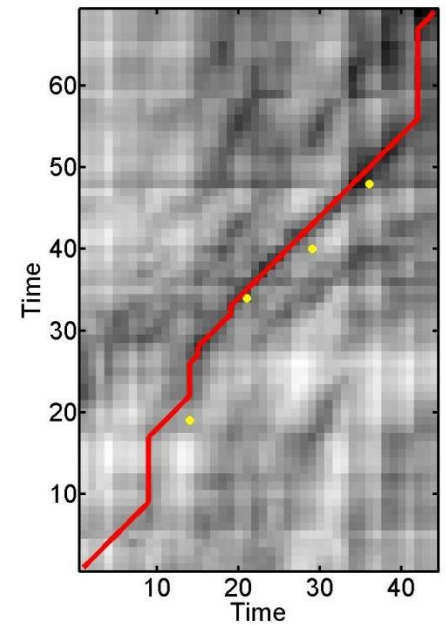
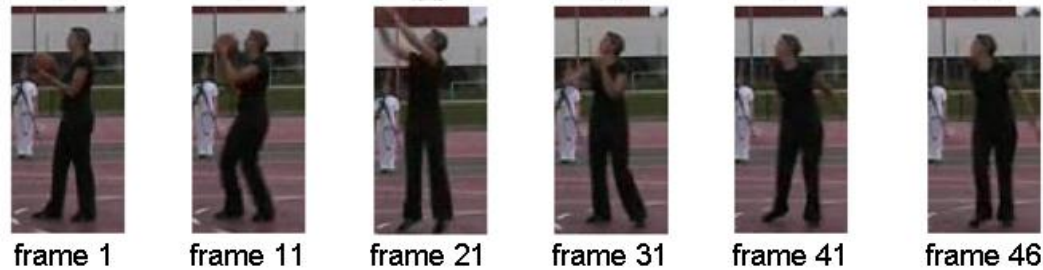
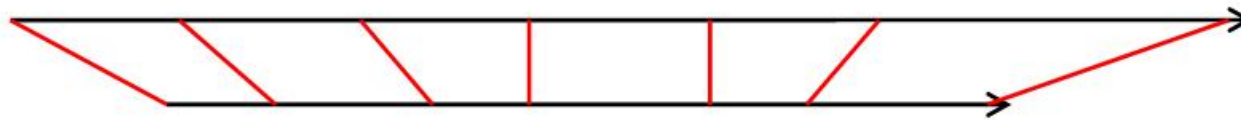
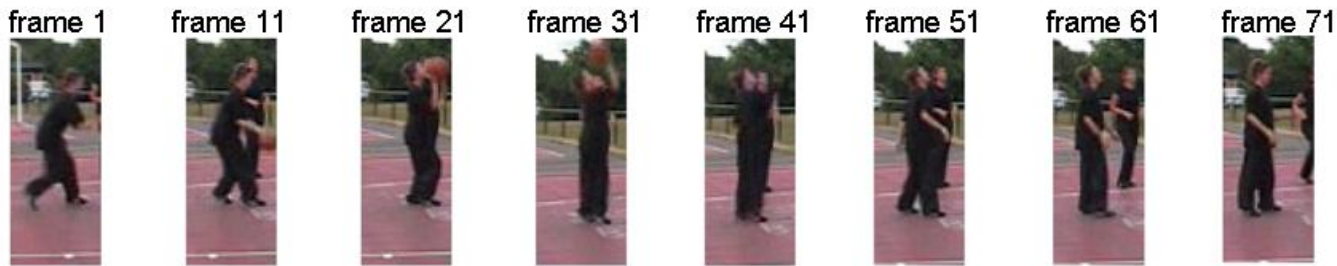
- **Sequence alignment:**  
Dynamic Programming for two sequences of descriptors  $\{h_i\}$ ,  $\{h_j\}$



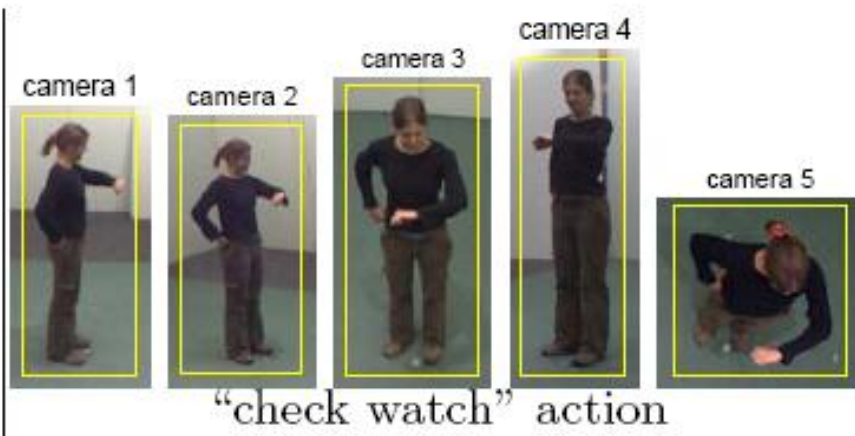
- **Action recognition:**
  - Visual vocabulary for  $h$
  - BoF representation of  $\{h_i\}$
  - SVM



# Multi-view alignment



# Multi-view action recognition: Video



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

cross-camera training/testing 
  same camera training/testing

SSM-based recognition

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

cross-camera training/testing 
  same camera training/testing

Alternative **view-dependent** method (STIP)

# Space-time action detection

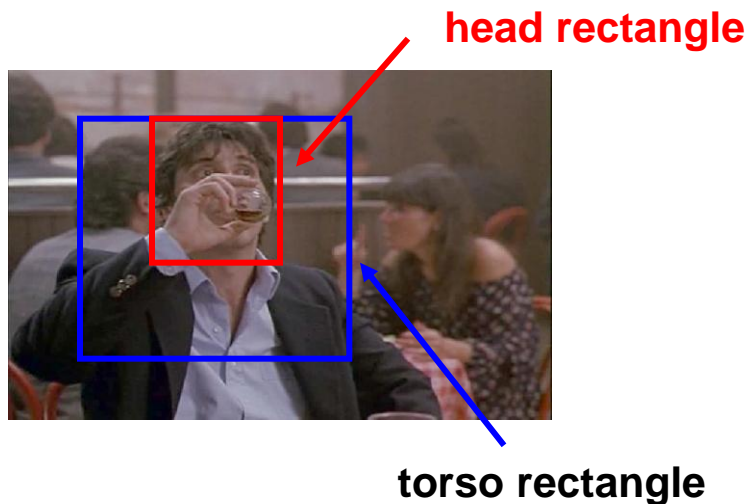


Manual annotation of drinking actions in movies:  
“Coffee and Cigarettes”; “Sea of Love”

“*Drinking*”: 159 annotated samples

“*Smoking*”: 149 annotated samples

Spatial annotation



Temporal annotation

First frame      Keyframe      Last frame





# “Drinking” action samples

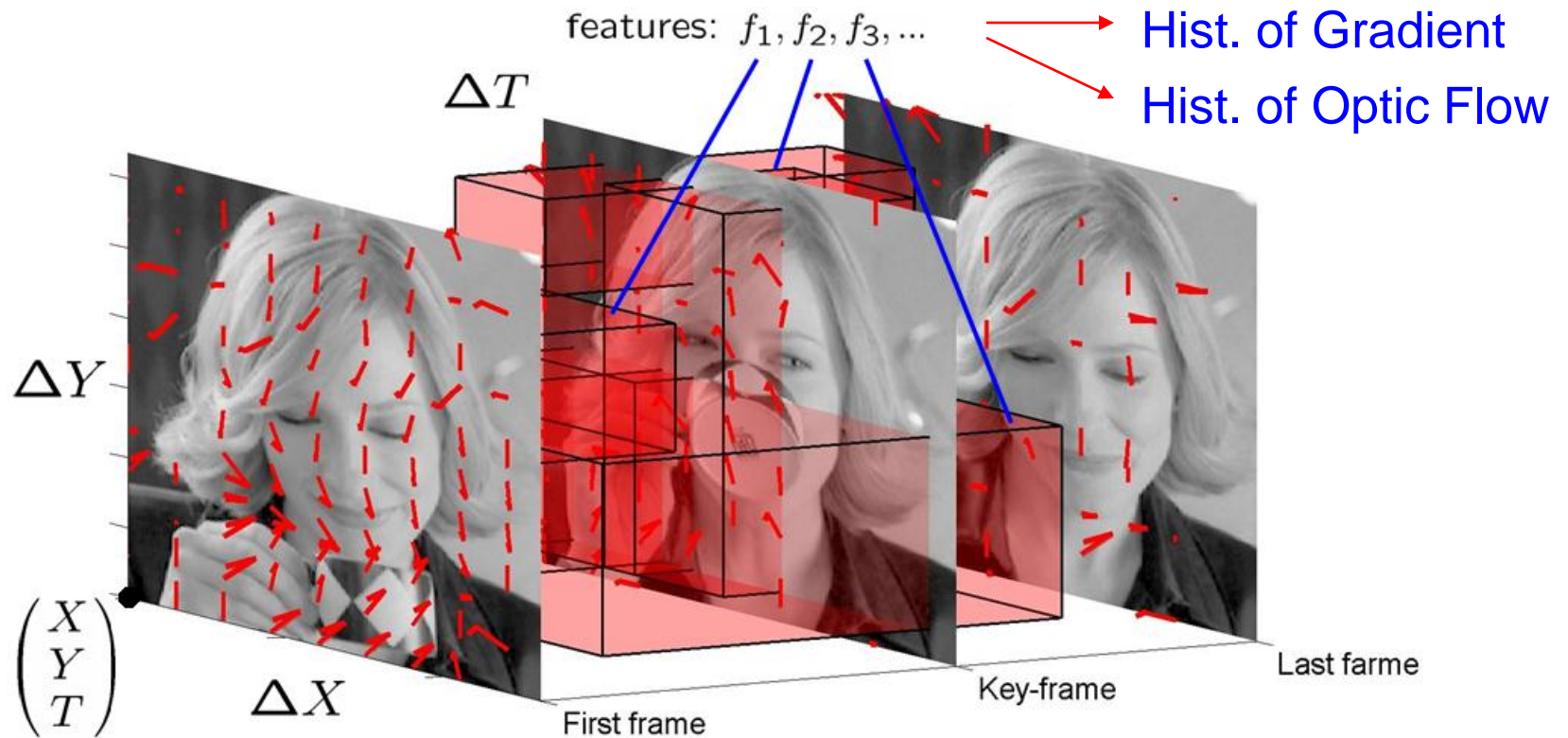
training samples



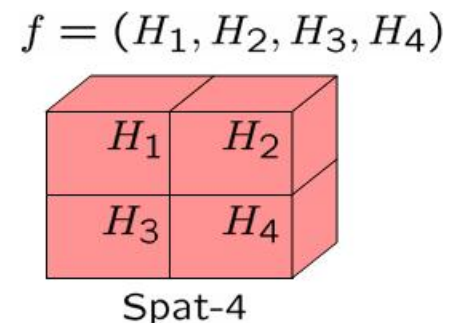
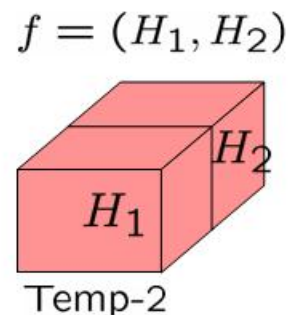
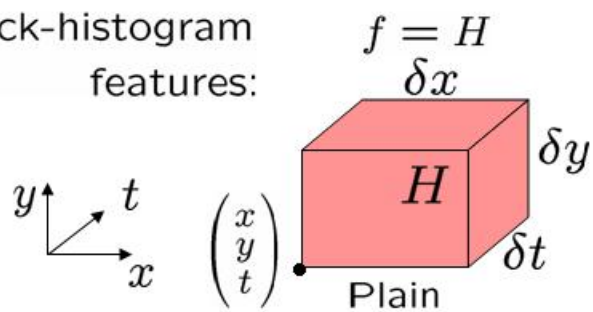
test samples



# Action representation

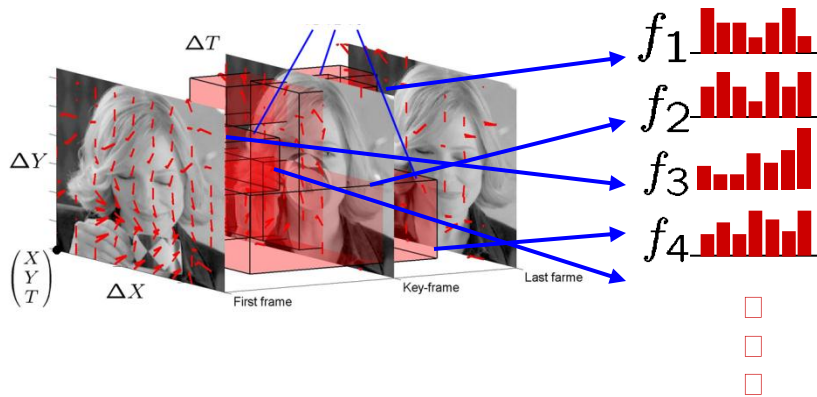


block-histogram  
features:





# Action learning



selected features

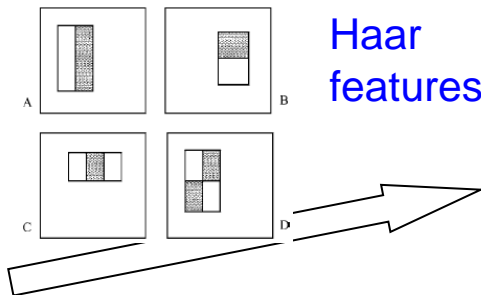
$$H(z) = \text{sgn}\left(\sum_{t=1}^T \alpha_t h_t(f_t)\right)$$

weak classifier

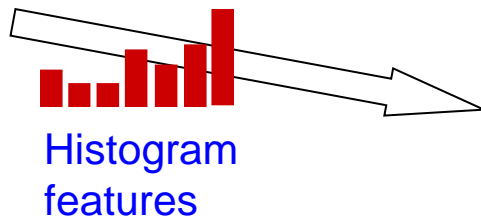
AdaBoost:

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

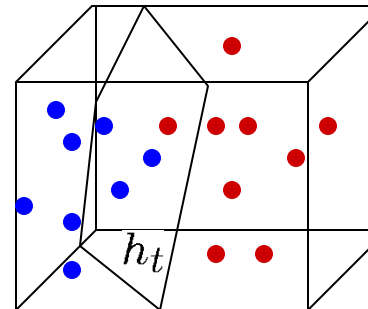
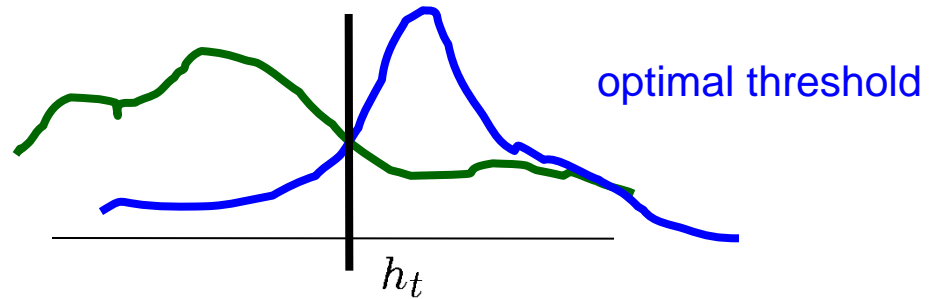
pre-aligned samples



Haar features

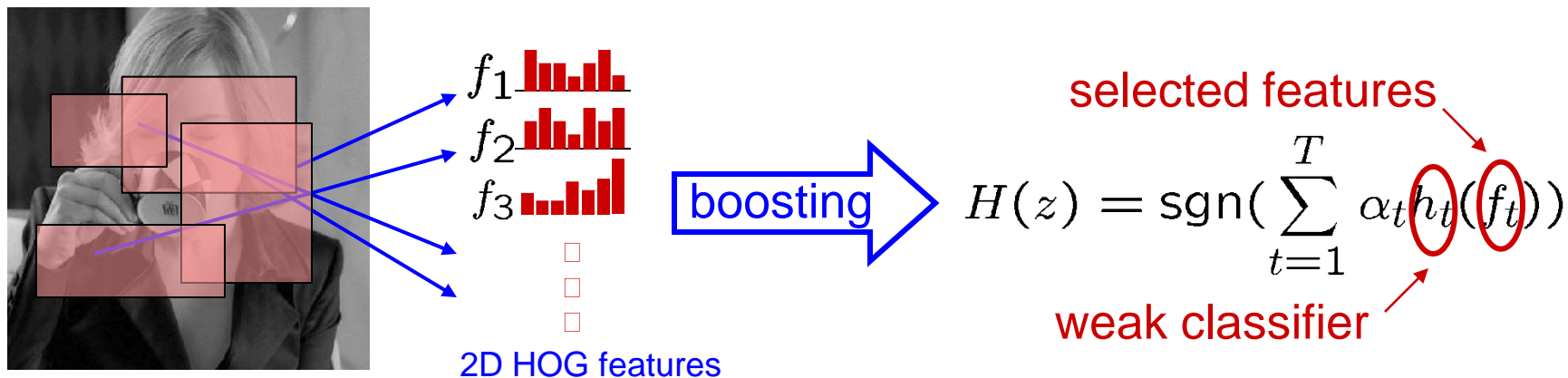


Histogram features



Fisher discriminant

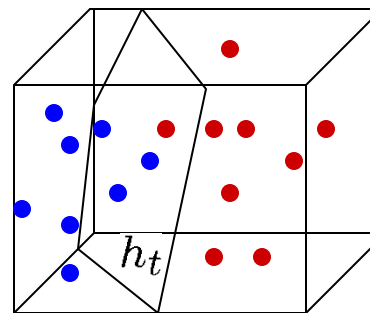
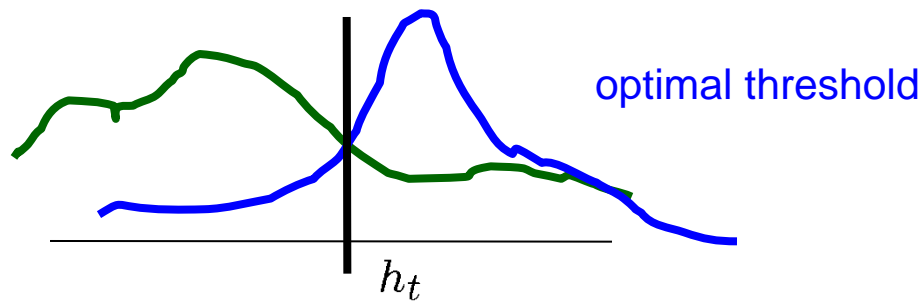
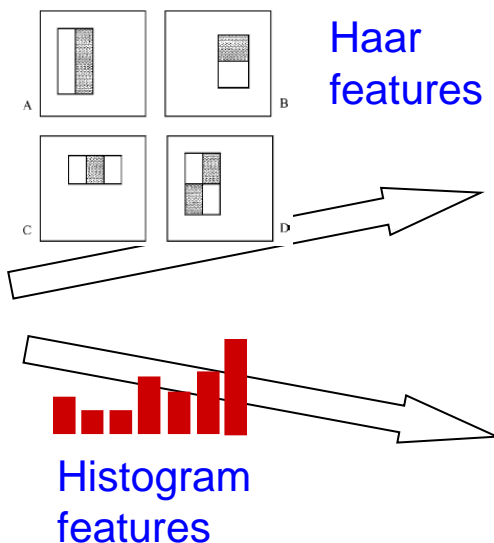
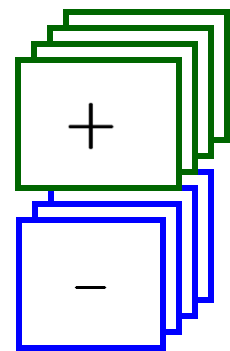
# Key-frame action classifier



AdaBoost:

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

pre-aligned samples



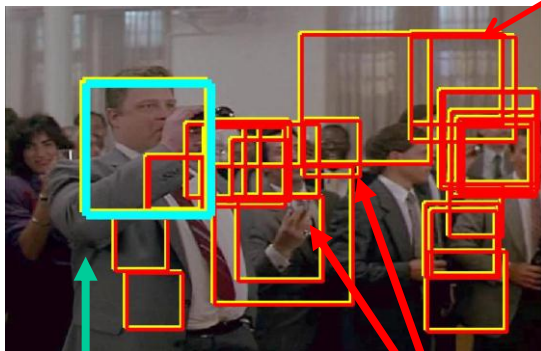
Fisher discriminant  
see [Laptev BMVC'06]  
for more details

[Laptev, Pérez 2007]

# Keyframe priming

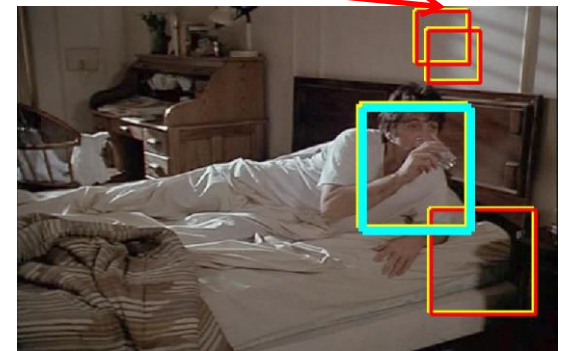
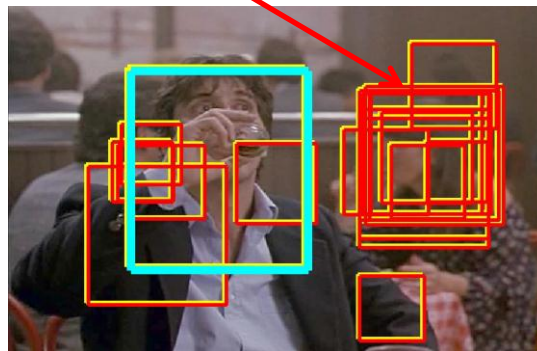
Training

False positives of static HOG action detector

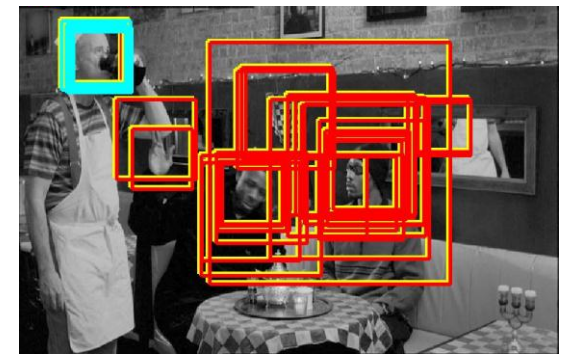
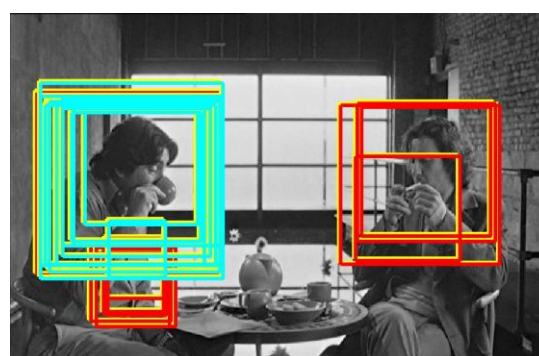
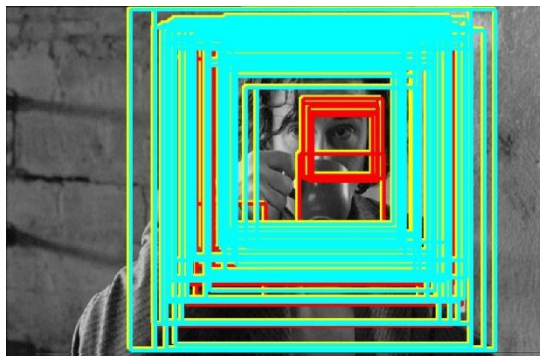


Positive training sample

Negative training samples



Test



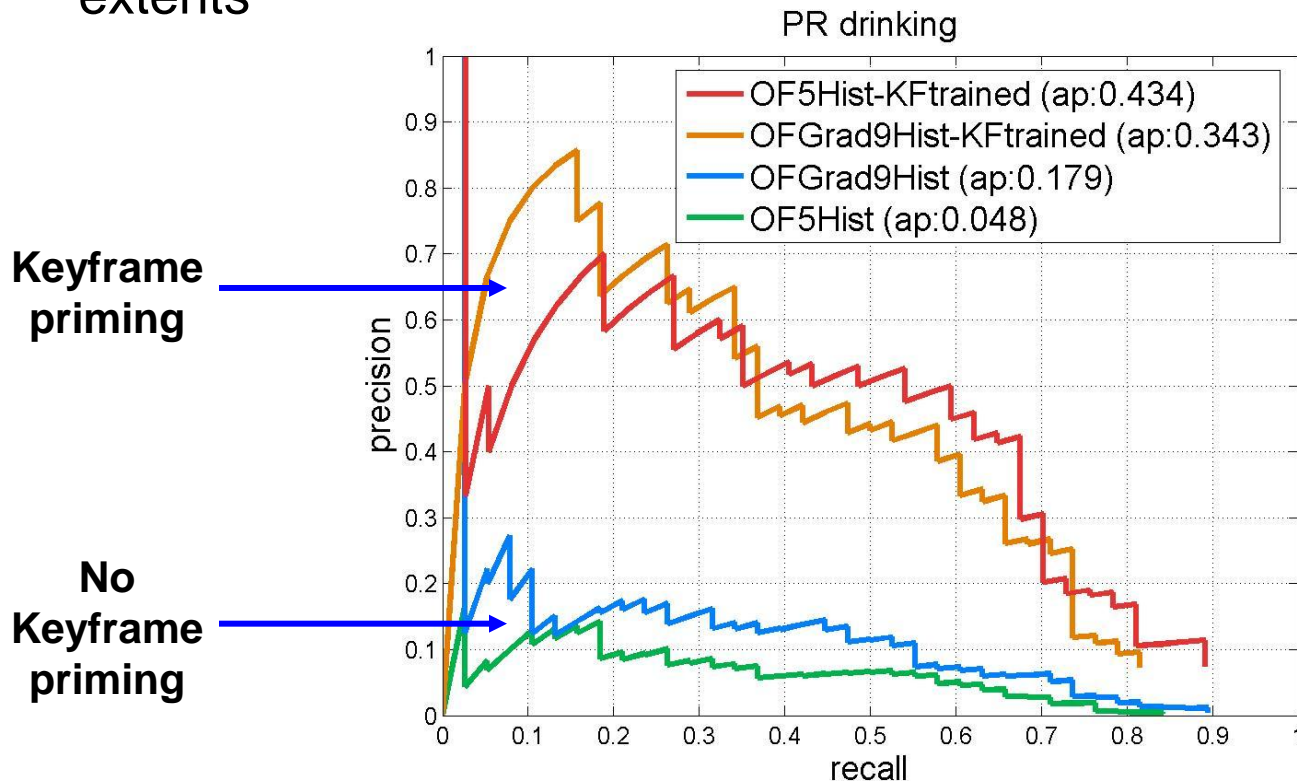
# Action detection

Test set:

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

Detection:

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



# Action Detection (ICCV 2007)



Test episodes from the movie “Coffee and cigarettes”

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



# **20 most confident detections**

# Learning Actions from Movies

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



Problems:

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

# Automatic video annotation with scripts

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

## subtitles

...  
1172  
01:20:17,240 --> 01:20:20,437

Why weren't you honest with me?  
**Why'd** you keep your marriage a secret?

1173  
01:20:20,640 --> 01:20:23,598

It wasn't my secret, Richard.  
Victor wanted it that way.

1174  
01:20:23,800 --> 01:20:26,189

Not even our closest friends  
knew about our marriage.

## movie script

...  
RICK  
Why weren't you honest with me? **Why**  
**did** you keep your marriage a secret?

01:20:17  
01:20:23

Rick sits down with Ilsa.

ILSA

**Oh**, it wasn't my secret, Richard.  
Victor wanted it that way. Not even  
our closest friends knew about our  
marriage.

...

# Script-based action annotation

## – On the good side:

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

## – Problems:

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

# Script alignment: Evaluation

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

Example of a “visual false positive”



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



# Text-based action retrieval

- Large variation of action expressions in text:

GetOutCar  
action:

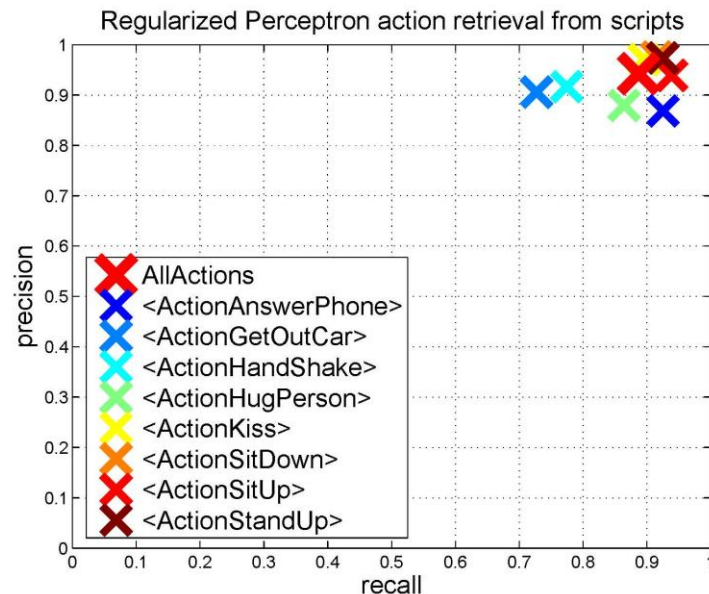
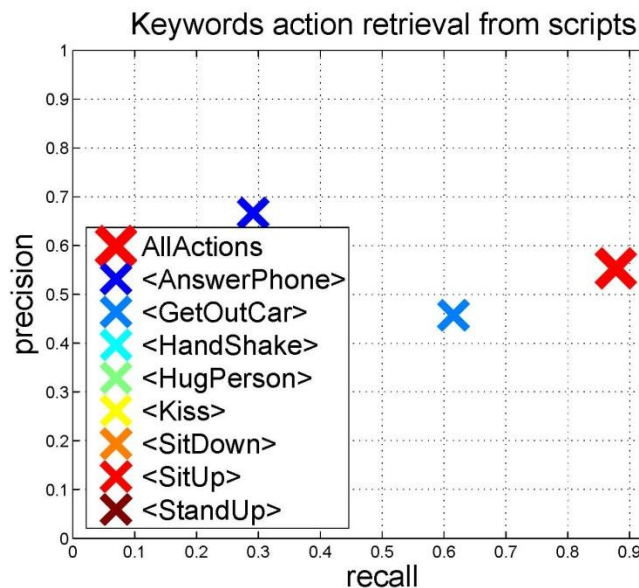
*“... Will gets out of the Chevrolet. ...”*

*“... Erin exits her new truck...”*

Potential false  
positives:

*“...About to sit down, he freezes...”*

- => Supervised text classification approach



# Automatically annotated action samples

AnswerPhone



GetOutCar



HandShake



HugPerson



Kiss



SitDown



SitUp



StandUp



# 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>

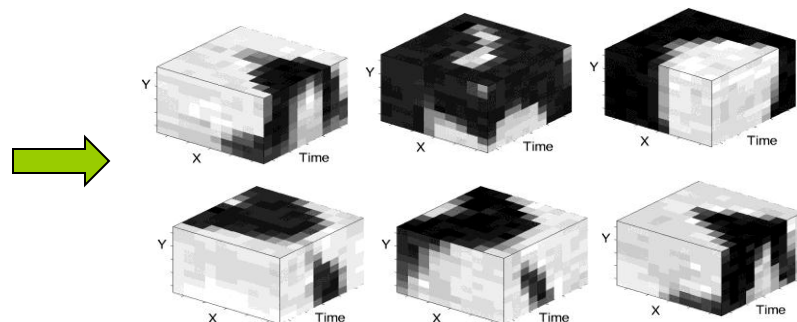
# Action Classification: Overview

Bag of space-time features + multi-channel SVM

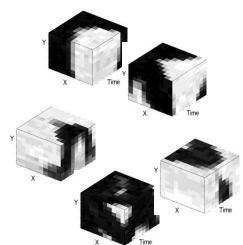
[Laptev'03, Schuldt'04, Niebles'06, Zhang'07]



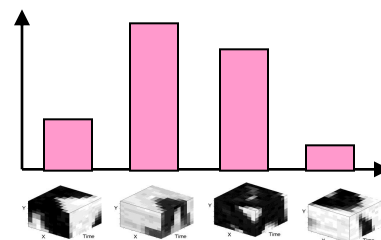
Collection of space-time patches



Histogram of visual words  
With space-time grids



HOG & HOF  
patch  
descriptors



Multi-channel  
SVM  
Classifier

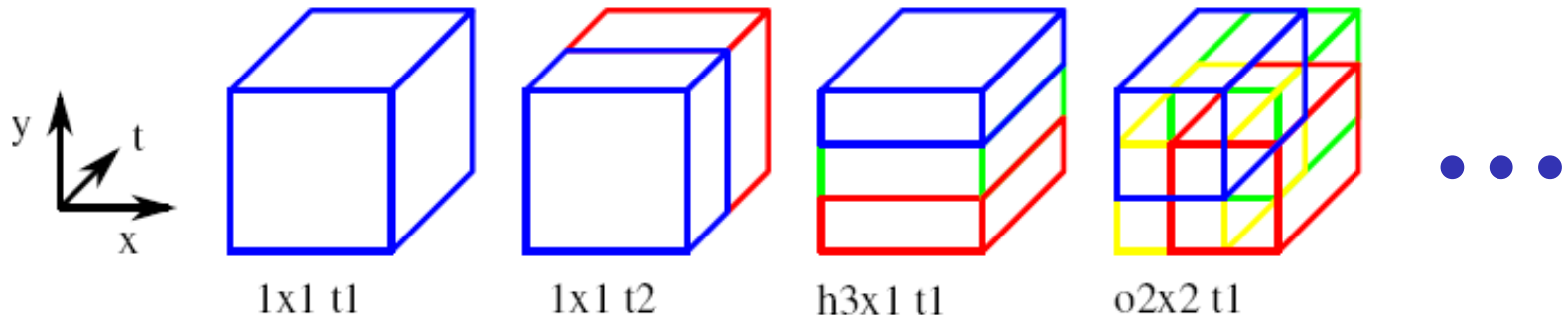
# Spatio-temporal grids

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



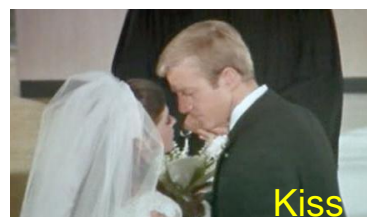


# Multi-channel chi-square kernel

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

- Channel  $c$  is a combination of a detector, descriptor and a grid
- $D_c(H_i, H_j)$  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  $\mathcal{C}$  for a given training set is found based on a greedy approach

# Action classification results



Clean  
training

Automatic  
training

	hohof		hohof		Chance
Channel	bof	flat	bof	flat	
mAP	47.9	50.3	31.9	36.0	9.2
AnswerPhone	15.7	20.9	18.2	19.1	7.2
DriveCar	86.6	84.6	78.2	80.1	11.5
Eat	59.5	67.0	13.0	22.3	3.7
FightPerson	71.1	69.8	52.9	57.6	7.9
GetOutCar	29.3	45.7	13.8	27.7	6.4
HandShake	21.2	27.8	12.8	18.9	5.1
HugPerson	35.8	43.2	15.2	20.4	7.5
Kiss	51.5	52.5	43.2	48.6	11.7
Run	69.1	67.8	54.2	49.1	16.0
SitDown	58.2	57.6	28.6	34.1	12.2
SitUp	17.5	17.2	11.8	10.8	4.2
StandUp	51.7	54.3	40.5	43.6	16.5

Average precision (AP) for Hollywood-2 dataset

# 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

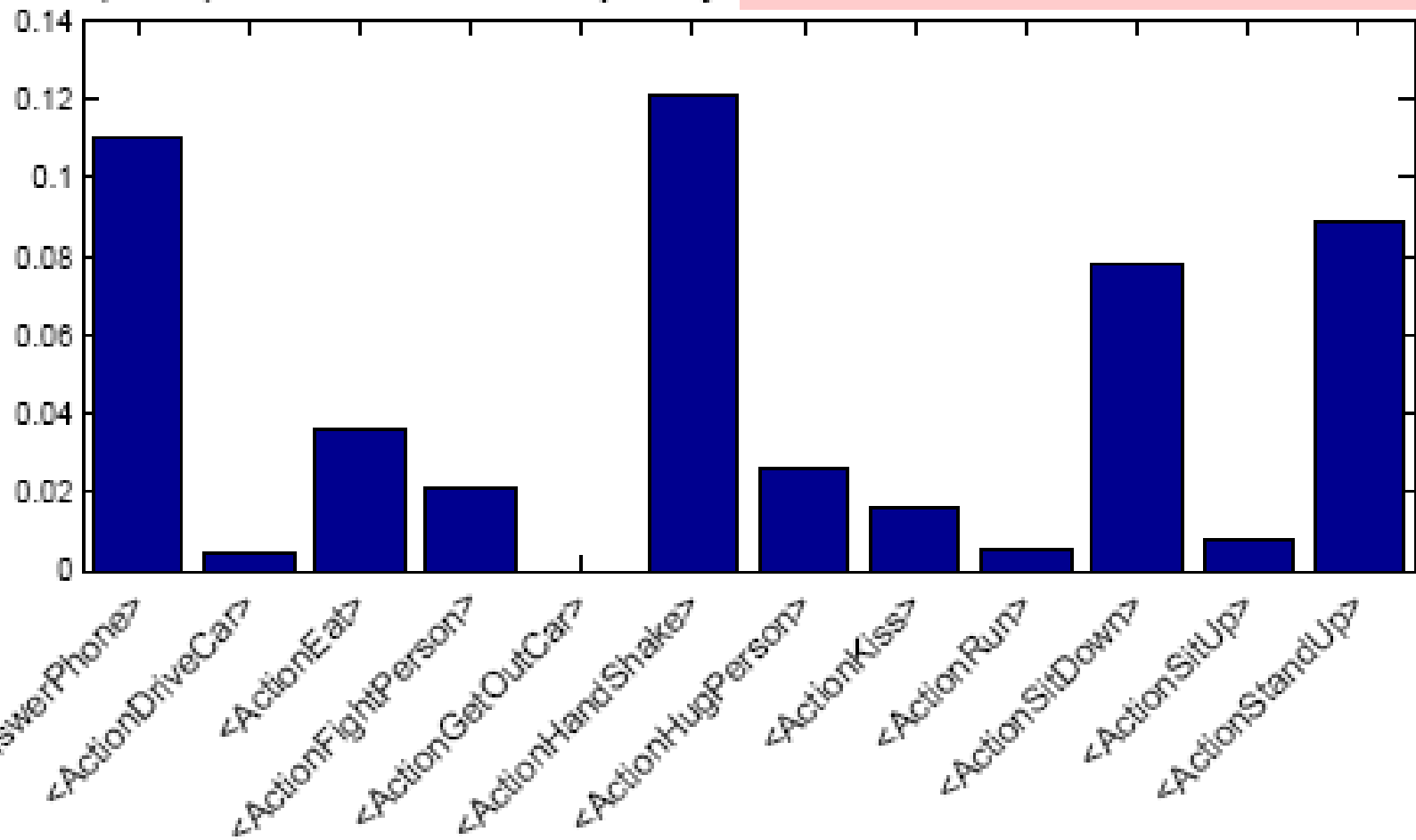
- Maximize word frequency ➡ street, living room, bedroom, car ....
- Merge words with similar senses using WordNet:

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  $P = p(\text{action} \mid \text{word})$

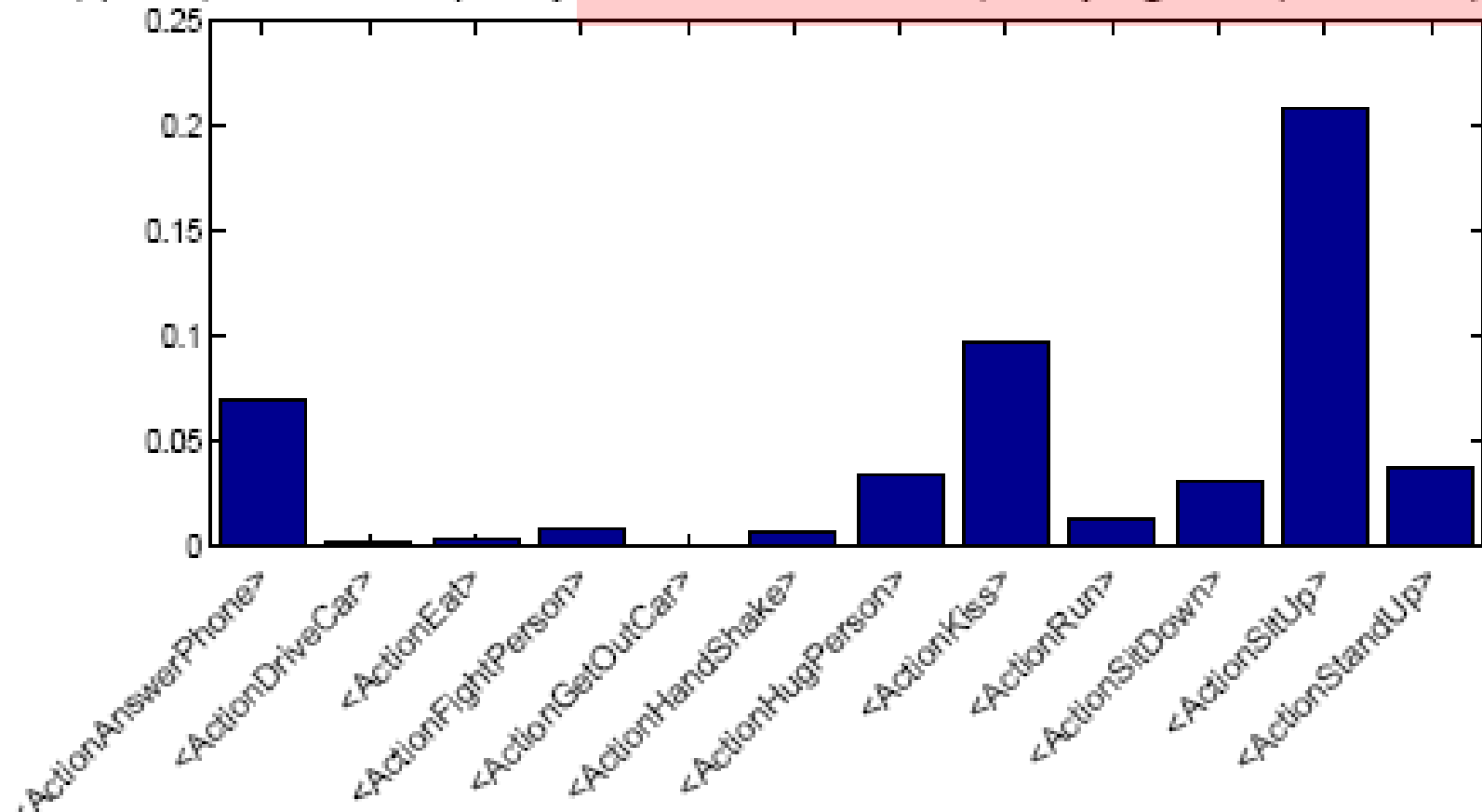
# Co-occurrence of actions and scenes in scripts

8(1267) | 147 | Relative Frequency: "Interior – office, business office"



# Co-occurrence of actions and scenes in scripts

1267) | 151 | Relative Frequency: "Interior – bedroom, sleeping room, chamber, bedchan



# 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

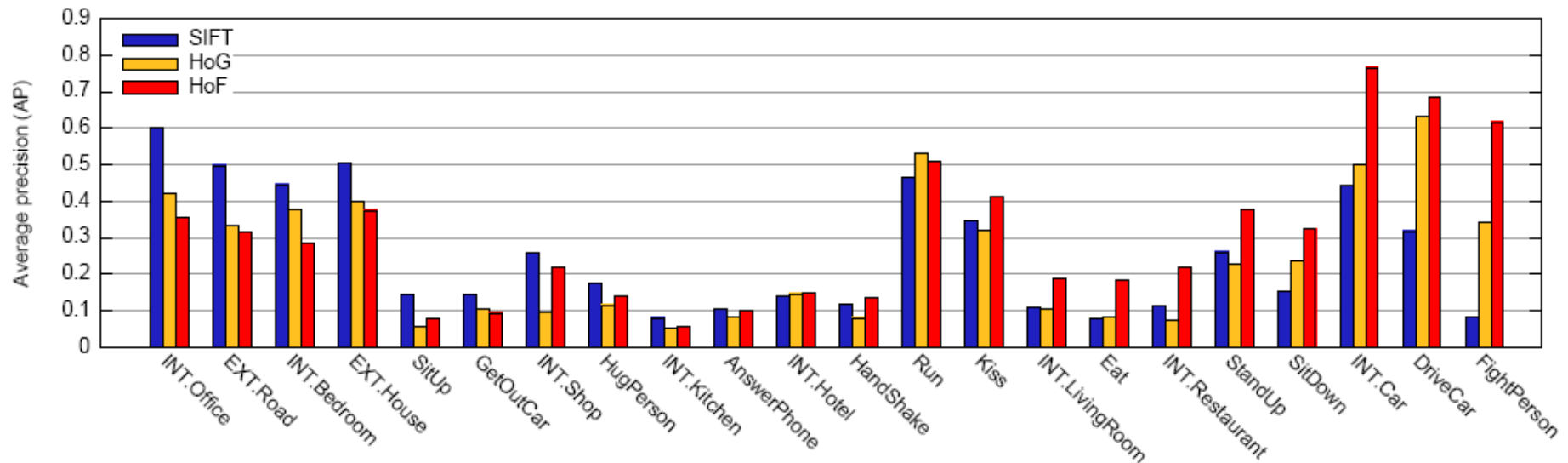
	Auto-Train-Scenes	Clean-Test-Scenes
EXT-house	81	140
EXT-road	81	114
INT-bedroom	67	69
INT-car	44	68
INT-hotel	59	37
INT-kitchen	38	24
INT-living-room	30	51
INT-office	114	110
INT-restaurant	44	36
INT-shop	47	28
All Samples	570	582

(b) Scenes

Source:  
69 movies  
aligned with  
the scripts

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

# Results: actions and scenes (separately)



EXT.House	<b>0.503</b>	0.363	0.491
EXT.Road	<b>0.498</b>	0.372	0.389
INT.Bedroom	<b>0.445</b>	0.362	<b>0.462</b>
INT.Car	0.444	<b>0.759</b>	<b>0.773</b>
INT.Hotel	0.141	<b>0.220</b>	<b>0.250</b>
INT.Kitchen	<b>0.081</b>	0.050	0.070
INT.LivingRoom	0.109	<b>0.128</b>	<b>0.152</b>
INT.Office	<b>0.602</b>	0.453	0.574
INT.Restaurant	<b>0.112</b>	0.103	0.108
INT.Shop	<b>0.257</b>	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	<b>0.105</b>	0.088	<b>0.107</b>
DriveCar	0.313	<b>0.749</b>	0.750
Eat	0.082	<b>0.263</b>	<b>0.286</b>
FightPerson	0.081	<b>0.675</b>	0.571
GetOutCar	<b>0.191</b>	0.090	<b>0.116</b>
HandShake	<b>0.123</b>	0.116	<b>0.141</b>
HugPerson	0.129	<b>0.135</b>	<b>0.138</b>
Kiss	0.348	<b>0.496</b>	<b>0.556</b>
Run	0.458	<b>0.537</b>	<b>0.565</b>
SitDown	0.161	<b>0.316</b>	0.278
SitUp	<b>0.142</b>	0.072	<b>0.078</b>
StandUp	0.262	<b>0.350</b>	0.325
Action average	0.200	0.324	0.326



# Classification with the help of context

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

$a_i(\mathbf{x})$       Action classification score

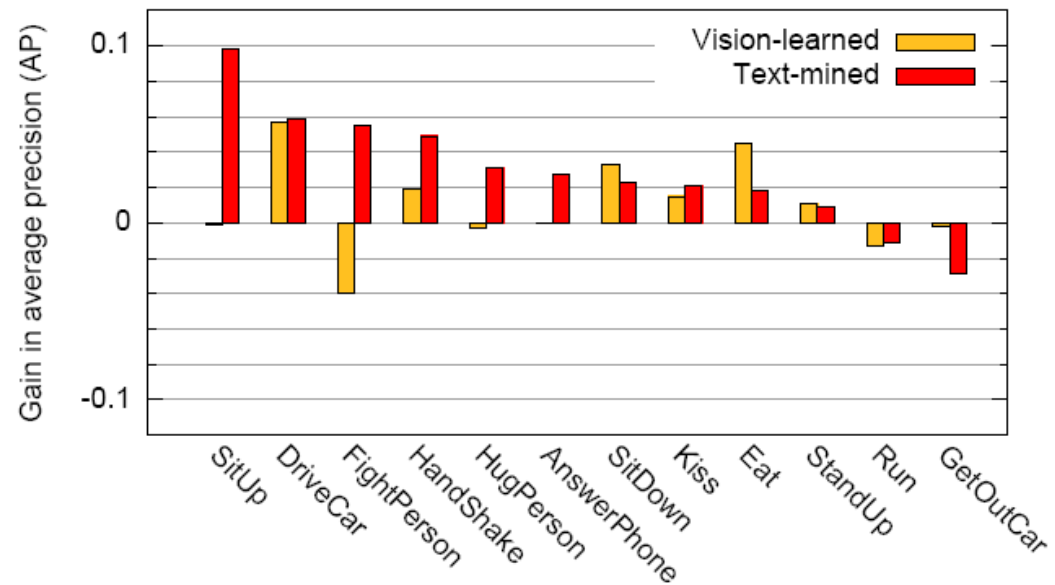
$s_j(\mathbf{x})$       Scene classification score

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

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

# Results: actions and scenes (jointly)

Actions  
in the  
context  
of  
Scenes



Scenes  
in the  
context  
of  
Actions

