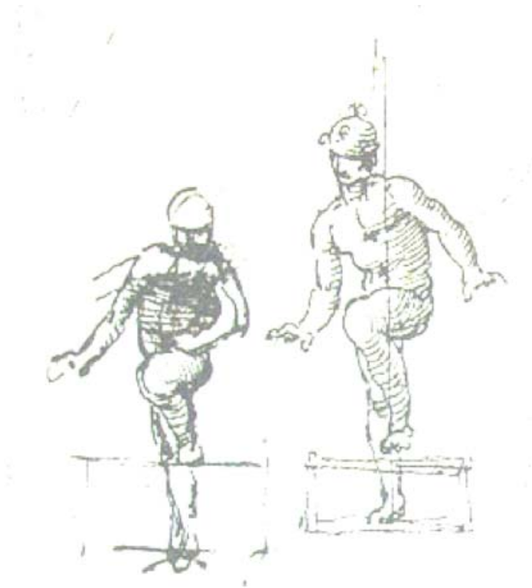


Reconnaissance d'objets et vision artificielle 2010

Motion and Human Actions II

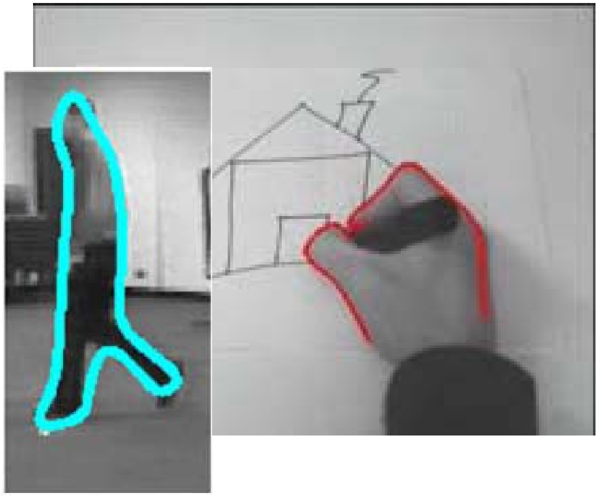
Ivan Laptev

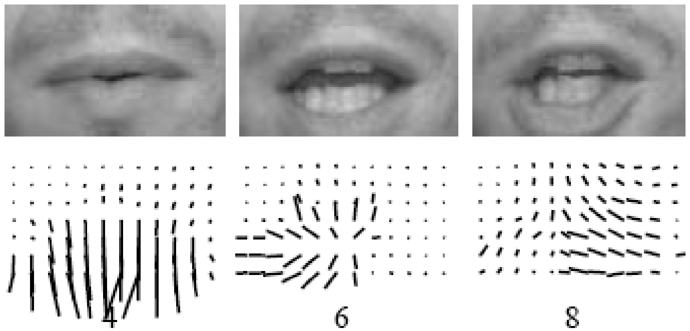
ivan.laptev@inria.fr

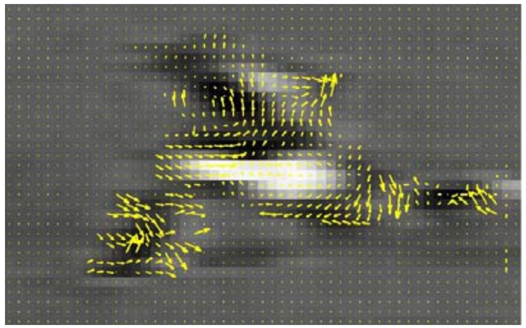
INRIA, WILLOW, ENS/INRIA/CNRS UMR 8548
Laboratoire d'Informatique, Ecole Normale Supérieure, Paris

Includes slides from: Mark Everingham,Josef Sivic, Andrew Zisserman

Poses and actions so far:







Motivation

Goal:
Interpreting complex dynamic scenes

Common methods:

- Segmentation ?
- Tracking ?

Common problems:

- Complex & changing BG
- Changing appearance

⇒ No global assumptions about the scene

Space-time

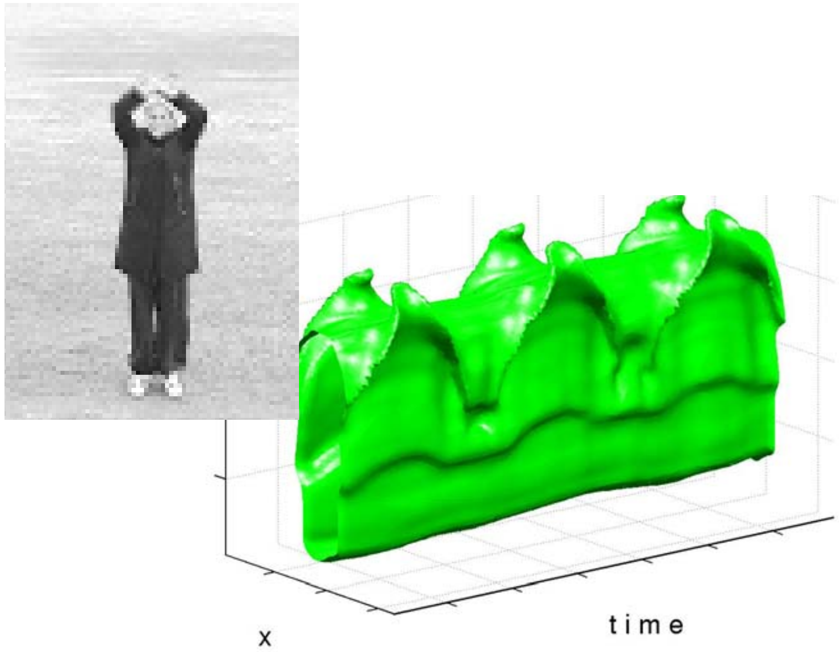
No global assumptions $\Rightarrow$

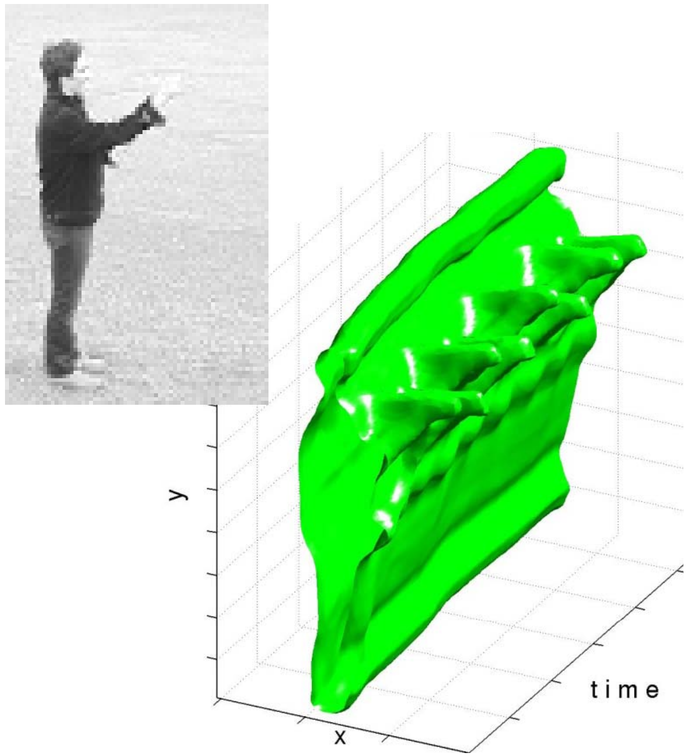
Consider local spatio-temporal neighborhoods

hand waving

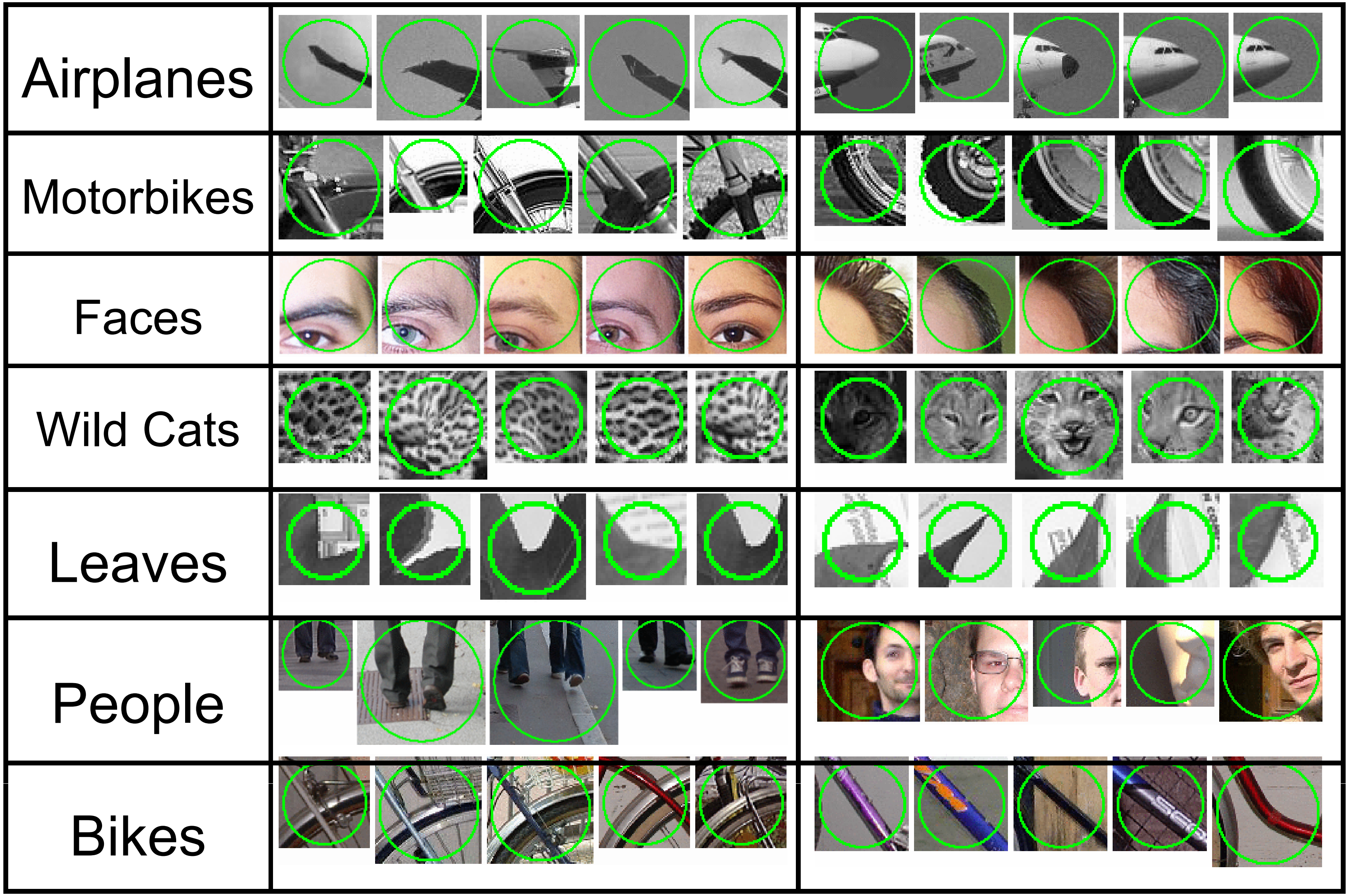
boxing

Actions == Space-time objects?

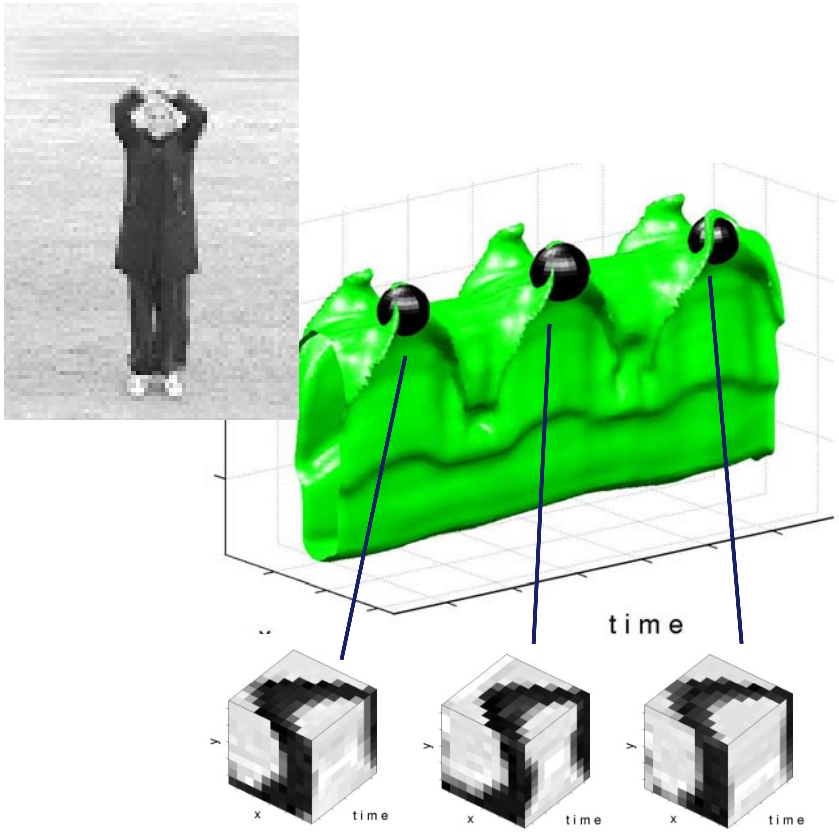


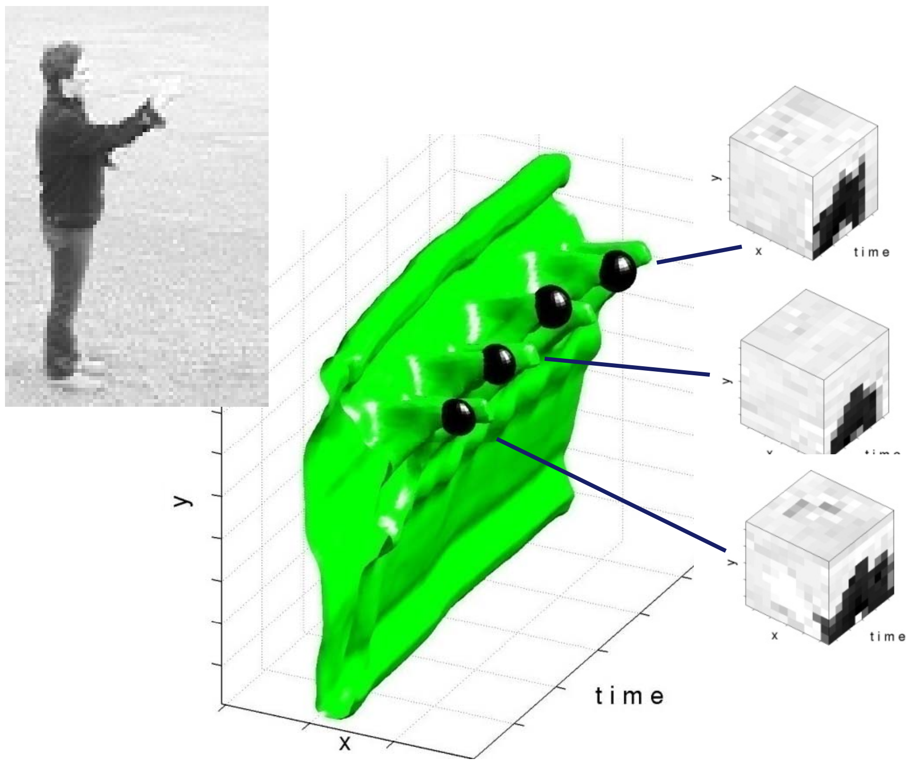


Local approach: Bag of Visual Words



Space-time local features





Space-Time Interest Points: Detection

What neighborhoods to consider?

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

Definitions:

$f: \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ 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

Space-Time Interest Points: Detection

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

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

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

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

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

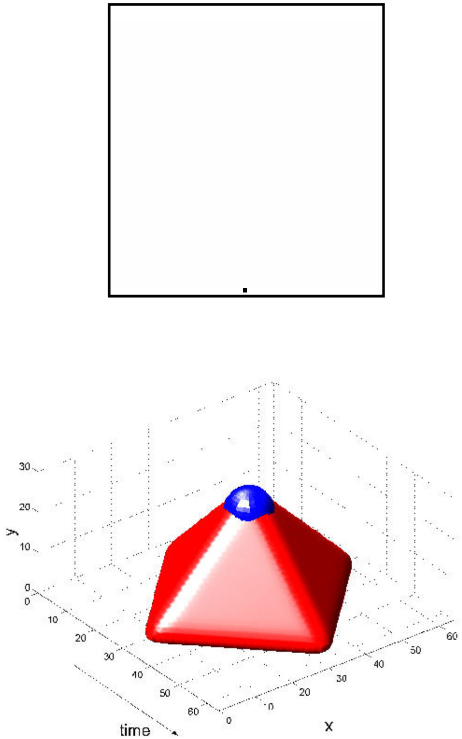
Large eigenvalues of μ can be detected by the local maxima of H over (x,y,t):

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

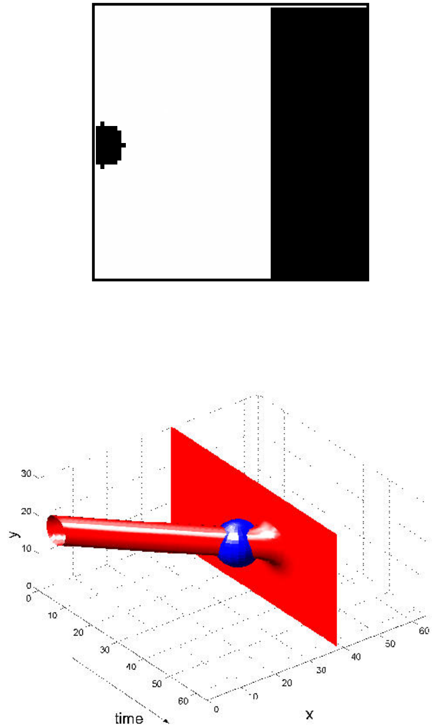
(similar to Harris operator [Harris and Stephens, 1988])

Space-Time interest points

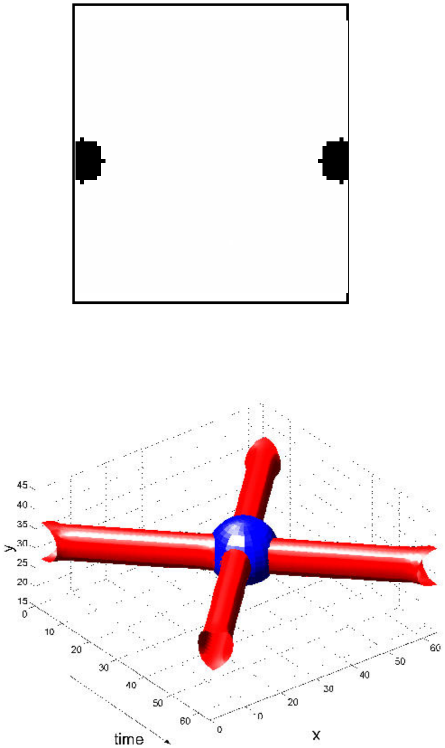
Velocity changes



appearance/ disappearance

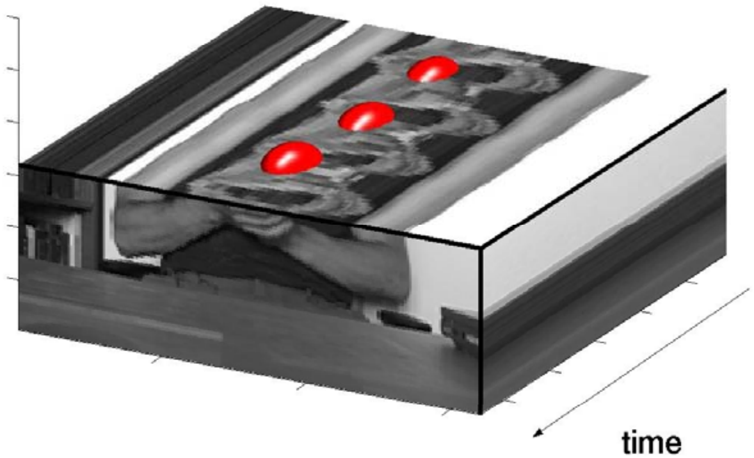


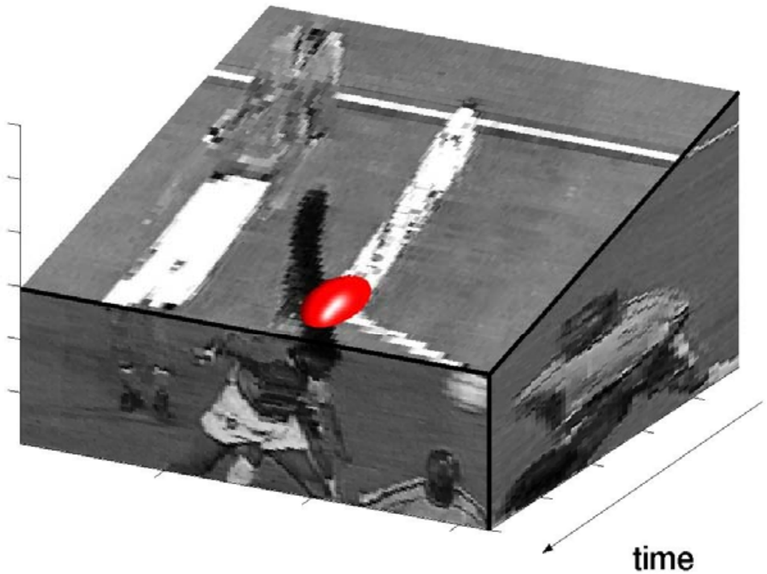
split/merge



Space-Time Interest Points: Examples

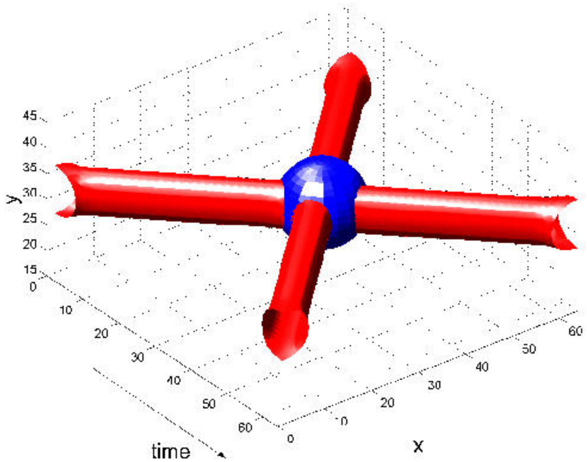
Motion event detection

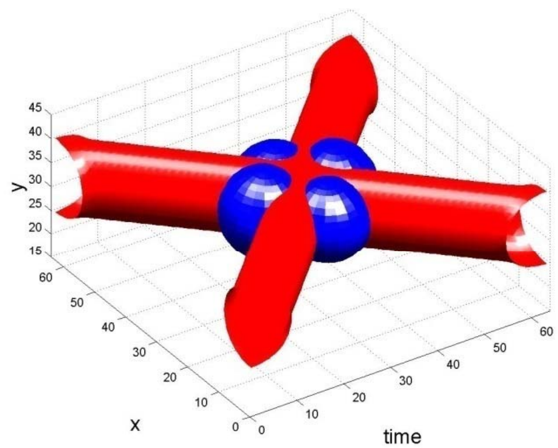




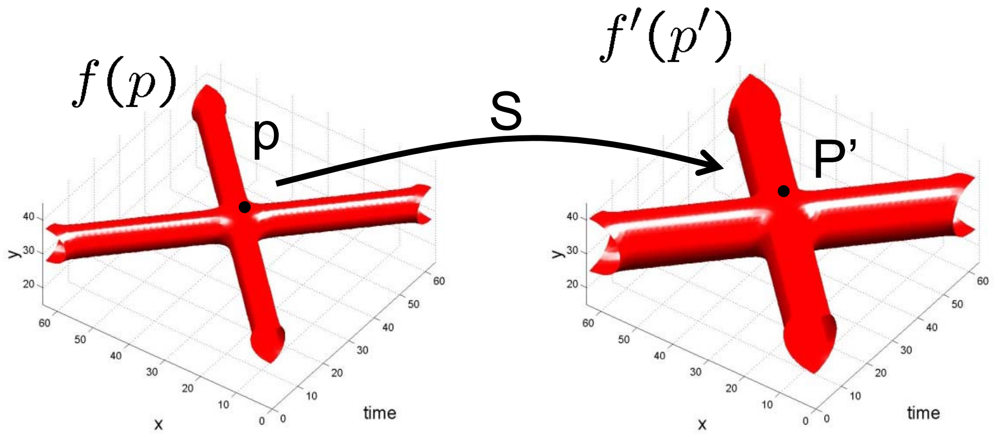
Spatio-temporal scale

What if the spatial and/or temporal resolution changes?





Spatio-temporal scale selection



point transformation

$$p = S^{-1}p', \quad S = \begin{pmatrix} s_\sigma & 0 & \\ 0 & s_\sigma & 0 \\ 0 & 0 & s_\tau \end{pmatrix}, \quad p = \begin{pmatrix} x \\ y \\ t \end{pmatrix}$$

covariance transformation

$$\Sigma = pp^T = S^{-2}\Sigma' = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \tau^2 \end{pmatrix}$$

Spatio-temporal scale selection

point transformation

$$p = S^{-1}p', \quad S = \begin{pmatrix} s_\sigma & 0 & \\ 0 & s_\sigma & 0 \\ 0 & 0 & s_\tau \end{pmatrix}, \quad p = \begin{pmatrix} x \\ y \\ t \end{pmatrix}$$

covariance transformation

$$\Sigma = pp^T = S^{-2}\Sigma' = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \tau^2 \end{pmatrix}$$

To be invariant to scale transformations we need to change filter covariance:

$$\begin{aligned} L_\xi(\cdot;\,\Sigma) &= f(\cdot) * g_\xi(\cdot;\,\Sigma) \\ &= f'(\cdot) * g_\xi(\cdot;\,\Sigma') \end{aligned}$$

Q: how to estimate the right filer size Σ ?

=>

Scale selection problem

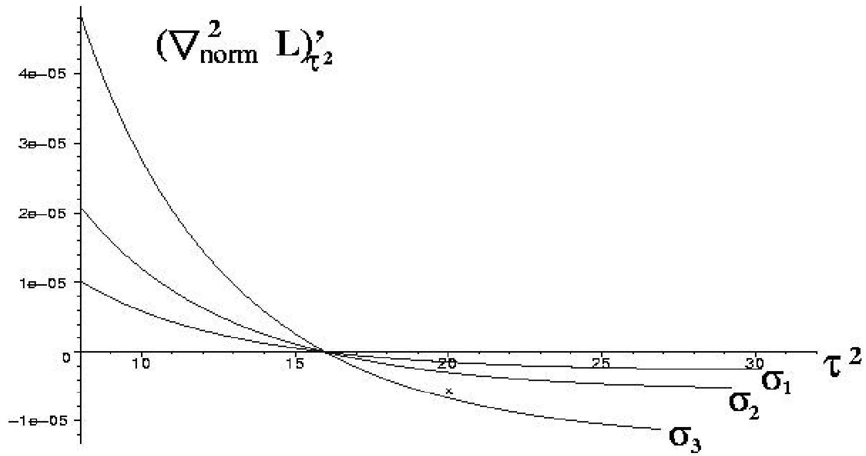
Spatio-temporal scale selection

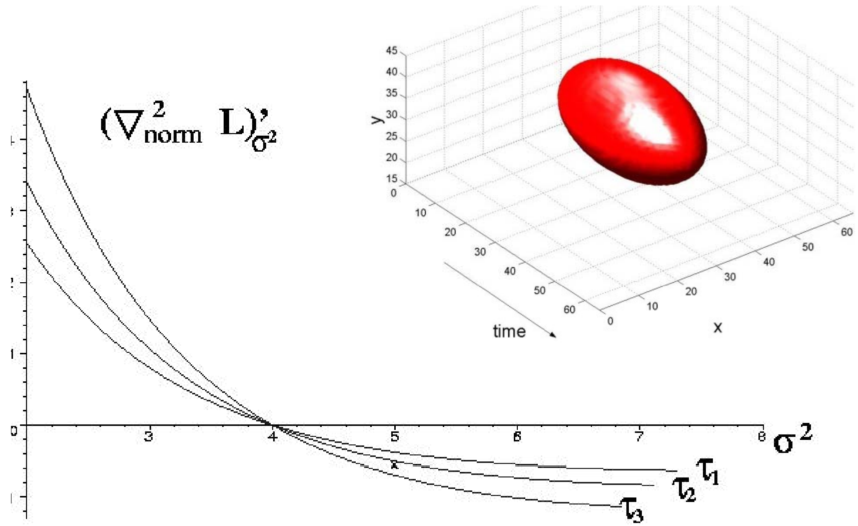
The normalized spatio-temporal Laplacian operator

$$\nabla^2_{norm} L = \sigma^2 \tau^{1/2} (L_{xx} + L_{yy}) + \sigma \tau^{3/2} L_{tt}$$

assumes scale-extrema values at the scale parameters of a spatio-temporal of a Gaussian blob

$\Rightarrow$ Estimate scale by maximizing $\quad (\nabla^2_{norm} L)^2 \quad \sigma, \tau$





(similar to scale selection is space [Lindeberg, 1998])

Space-Time interest points

H depends on μ and, hence, on Σ and scale transformation S

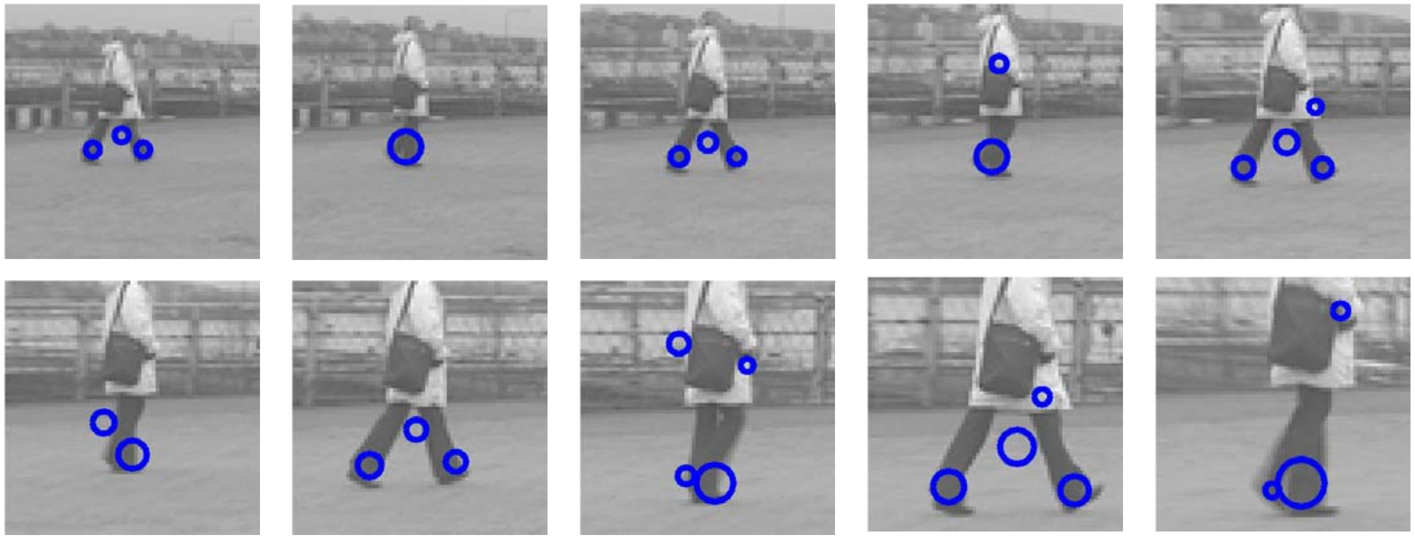
$\Rightarrow$ Adapt interest points by iteratively computing:

- Interest point detection $\quad H(p;\,\Sigma) = \det(\mu(p;\,\Sigma)) + k\,\mathrm{trace}^3(\mu(p;\,\Sigma))$ (*)
- Scale estimation $\quad (\sigma_0, \tau_0) = \mathrm{argmax}_{\sigma,\tau}(\nabla^2_{norm} L(p;\,\Sigma))^2$ (**)

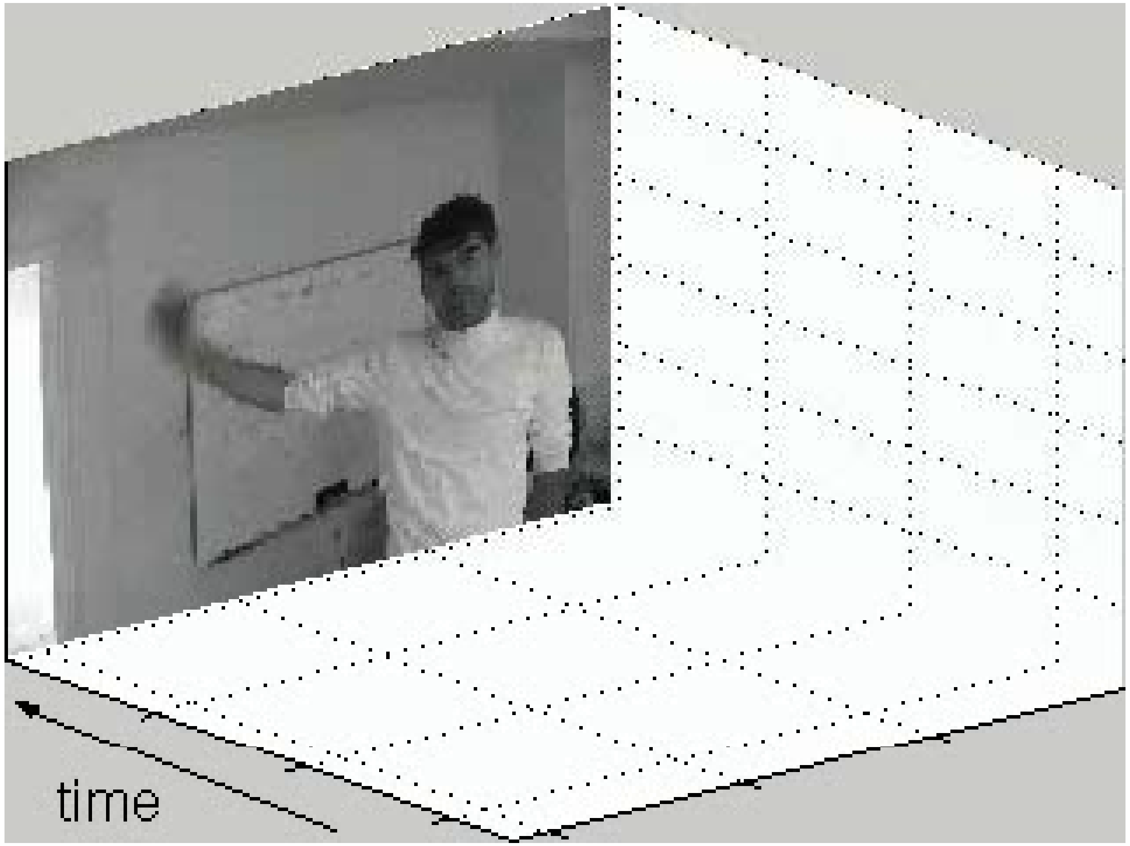
1. Fix Σ
2. For each detected interest point p_i (*)
3. Estimate scale $S(\sigma, \tau)$ (**)
4. Update covariance $\Sigma' = S^2$
5. Re-detect p_i using Σ'
6. Iterate 3-6 until convergence of σ, τ and p_i

Spatio-temporal scale selection

Stability to size changes,
e.g. camera zoom



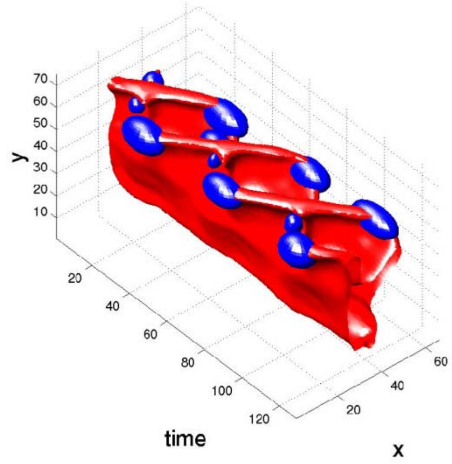
Spatio-temporal scale selection

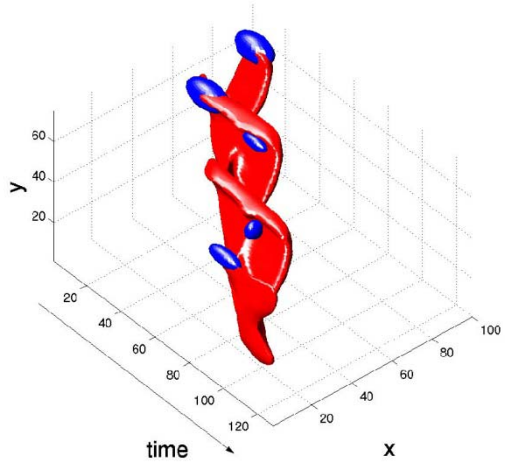


Selection of temporal scales captures the frequency of events

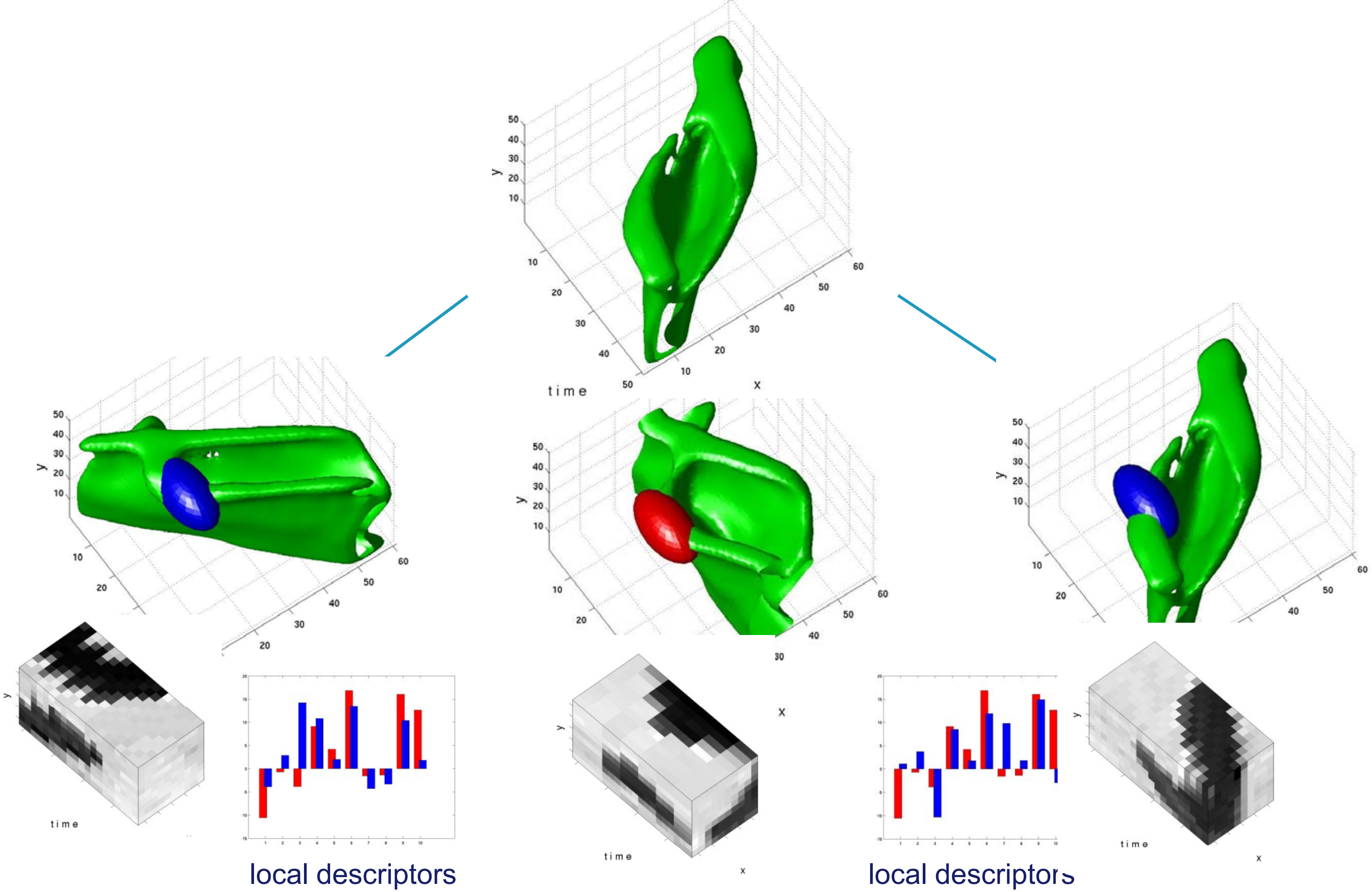
Relative camera motion

Space-time signal and its derivatives will change when if camera moves



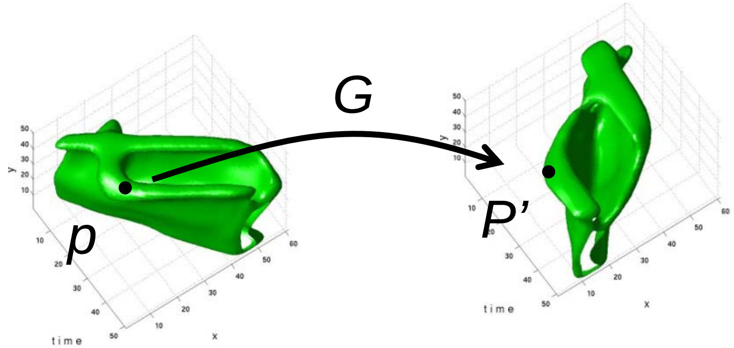


Effect of camera motion



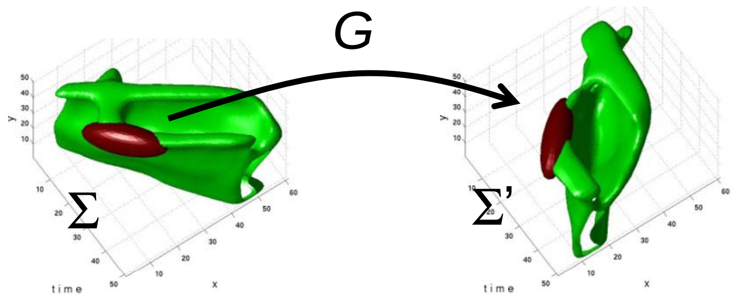
Galilean transformation

point
transformation



$$p = G^{-1}p' \qquad G = \begin{pmatrix} 1 & 0 & v_x \\ 0 & 1 & v_y \\ 0 & 0 & 1 \end{pmatrix}, \qquad p = \begin{pmatrix} x \\ y \\ t \end{pmatrix}$$

covariance
transformation



$$\Sigma = pp^T = G^{-1}\Sigma' G^{-T} \qquad \Sigma = \begin{pmatrix} c_{xx} & c_{xy} & c_{xt} \\ c_{xy} & c_{yy} & c_{yt} \\ c_{xt} & c_{yt} & c_{tt} \end{pmatrix}$$

Estimation of *G*

Want to "undo" the effect of *G*

$$\left.\begin{array}{rcl} p & = & G^{-1}p' \\ \Sigma & = & G^{-1}\Sigma' G^{-T} \end{array}\right\}$$

Consider local measurements:

$$L_\xi(\cdot;\,\Sigma) = f(\cdot) * g_\xi(\cdot;\,\Sigma)$$

Space-time gradient

$$\nabla L = (L_x, L_y, L_t)^T$$

$$g_\xi(\bar{x};\,\Sigma) = \partial_\xi \left(\frac{e^{-\frac{1}{2}p^T \Sigma^{-1} p}}{2\pi\sqrt{\det \Sigma}} \right)$$

Second-moment matrix

$$\mu(\cdot;\,\Sigma) = \nabla L(\cdot;\,\Sigma)(\nabla L(\cdot;\,\Sigma))^T * g(\cdot;\,s\Sigma)$$

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

Estimation of *G*

Transformations of ∇L and μ

$$\left.\begin{array}{rcl} p & = & G^{-1}p' \\ \Sigma & = & G^{-1}\Sigma' G^{-T} \end{array}\right\} \Rightarrow \begin{array}{l} \nabla L(p;\,\Sigma) = G^T \nabla L'(p';\,\Sigma') \\ \mu(p;\,\Sigma) = G^T \mu'(p';\,\Sigma')G \end{array}$$

Idea: Fix the "normal" form of μ and estimate *G by normalizing* μ.

- Let $\mu = \begin{pmatrix} \mu_{xx} & \mu_{xy} & 0 \\ \mu_{xy} & \mu_{yy} & 0 \\ 0 & 0 & \mu_{tt} \end{pmatrix}$

$$\begin{pmatrix} \mu'_{xt}(\cdot;\,\Sigma') \\ \mu'_{yt}(\cdot;\,\Sigma') \end{pmatrix} = \begin{pmatrix} \mu'_{xx}(\cdot;\,\Sigma') & \mu'_{xy}(\cdot;\,\Sigma') \\ \mu'_{xy}(\cdot;\,\Sigma') & \mu'_{yy}(\cdot;\,\Sigma') \end{pmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix}$$

Estimation of *G*

1. Fix Σ let $\Sigma' = \Sigma$
2. Estimate v_x, v_yording to (*)
3. Update $\Sigma = G^{-1}\Sigma' G^{-T}$
4. Iterate 2-3-4 until convergence of v_x, v_y

Iterative method for estimating v_x, v_y and Σ'

$\Uparrow$

Can solve for v_x, v_y *from* μ'*! (similar to Lucas&Kanade OF)*

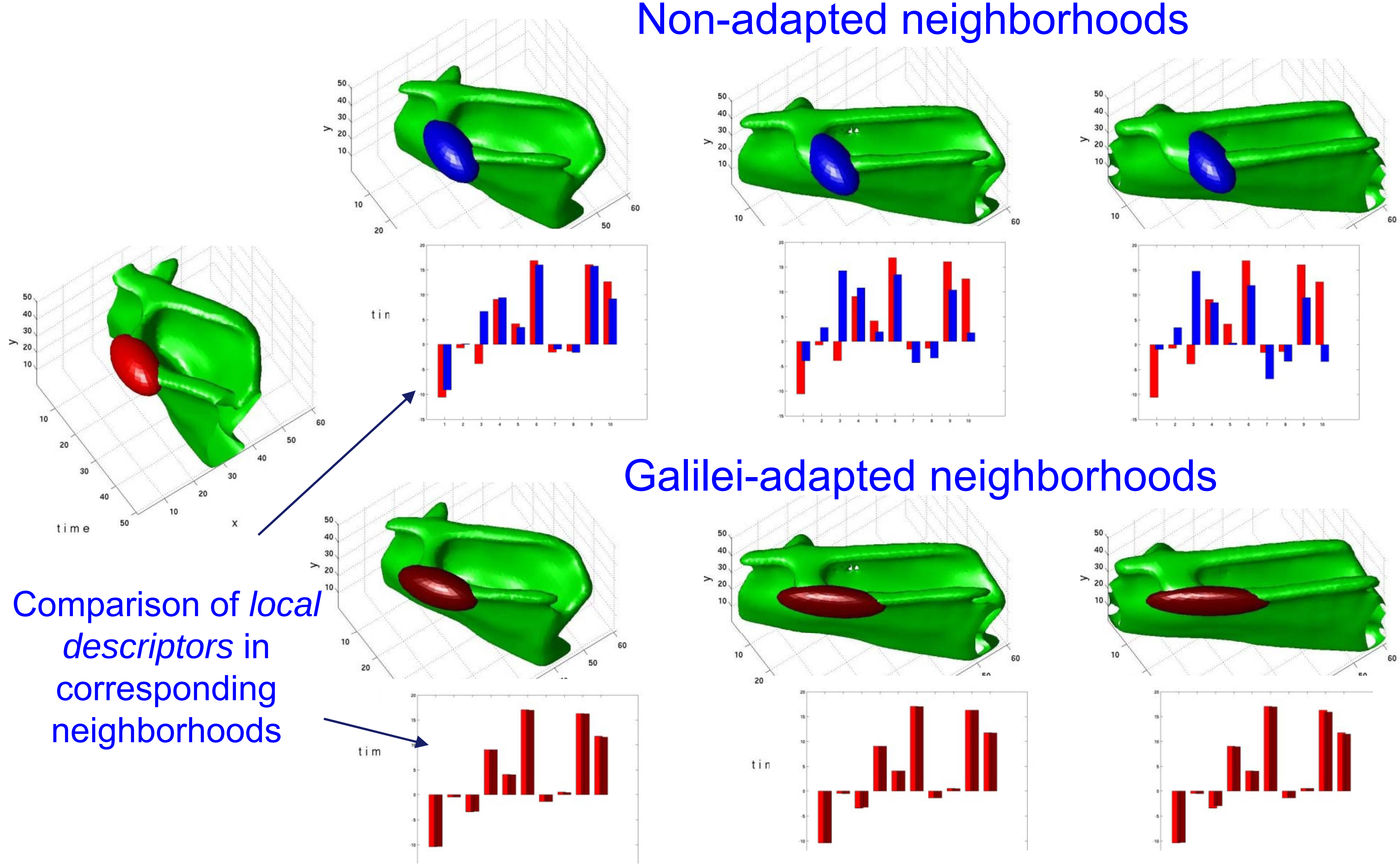
$\Uparrow$ ... however $(v_x, v_y)^T = \mathcal{F}_1(\Sigma') = \mathcal{F}_2(G)$

(*)
$$\begin{pmatrix} \mu'_{xt}(\cdot;\ \Sigma') \\ \mu'_{yt}(\cdot;\ \Sigma') \end{pmatrix} = -\begin{pmatrix} \mu'_{xx}(\cdot;\ \Sigma') & \mu'_{xy}(\cdot;\ \Sigma') \\ \mu'_{xy}(\cdot;\ \Sigma') & \mu'_{yy}(\cdot;\ \Sigma') \end{pmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix}$$

Estimation of *G*: experiments

Non-adapted neighborhoods

Galilei-adapted neighborhoods

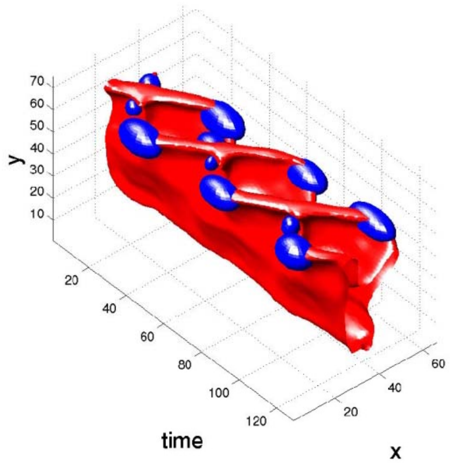


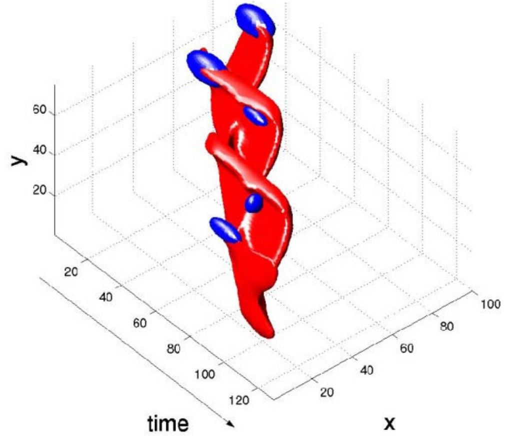
Comparison of *local descriptors* in corresponding neighborhoods

Adapted interest points

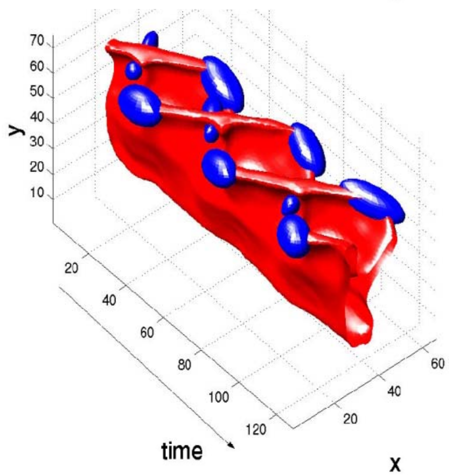
Stabilized camera | Stationary camera

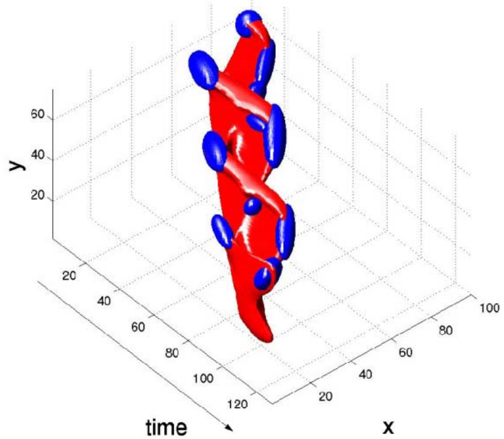
Interest points



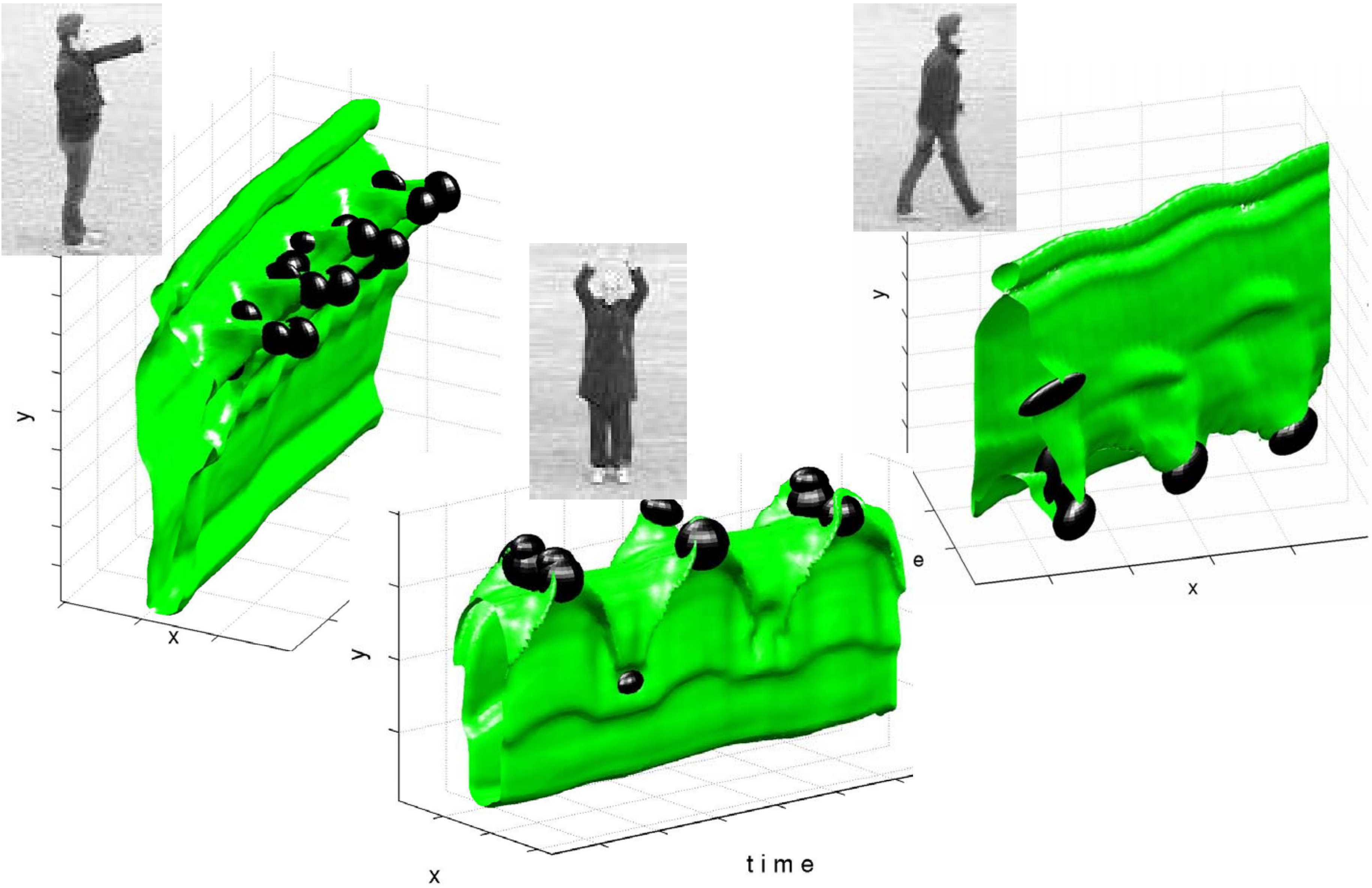


Velocity-adapted interest points

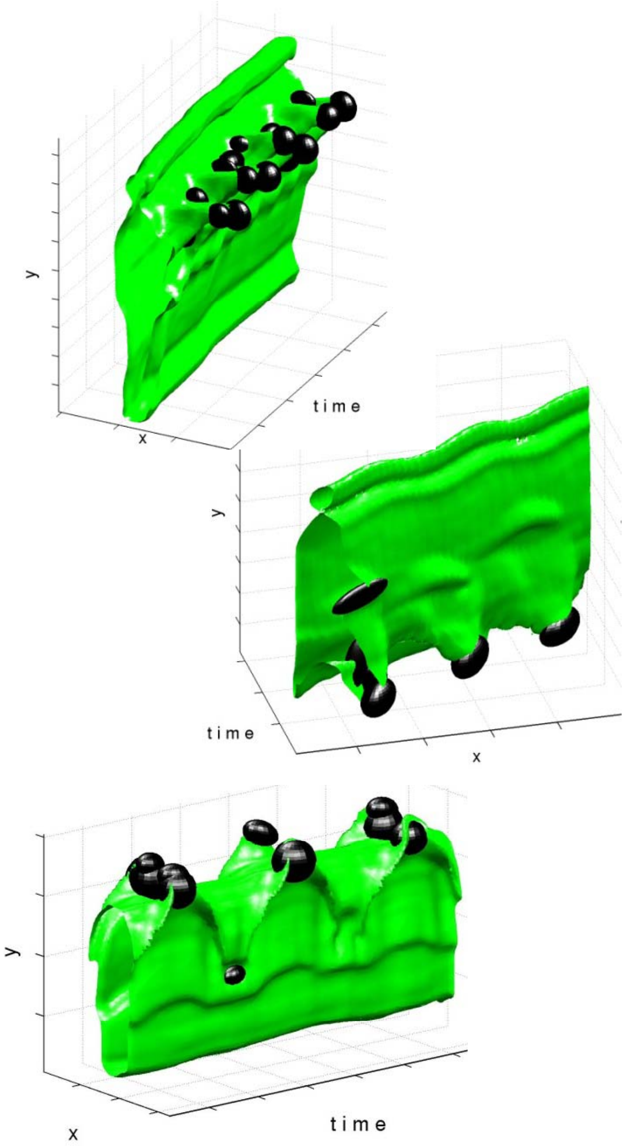




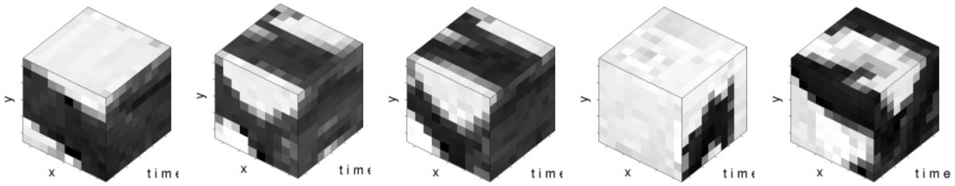
Local features for human actions



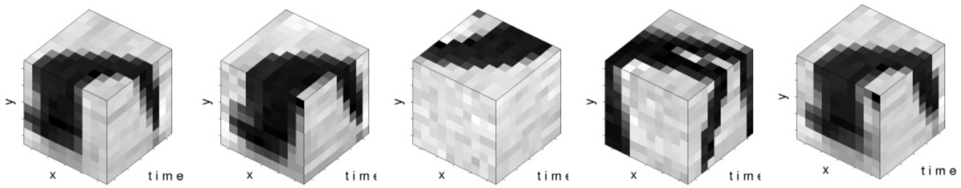
Local features for human actions



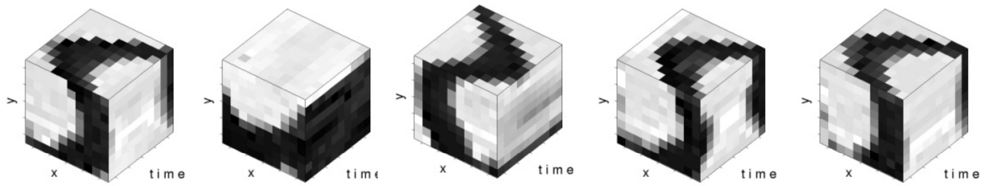
boxing



walking



hand waving



Local space-time descriptor: Jet

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

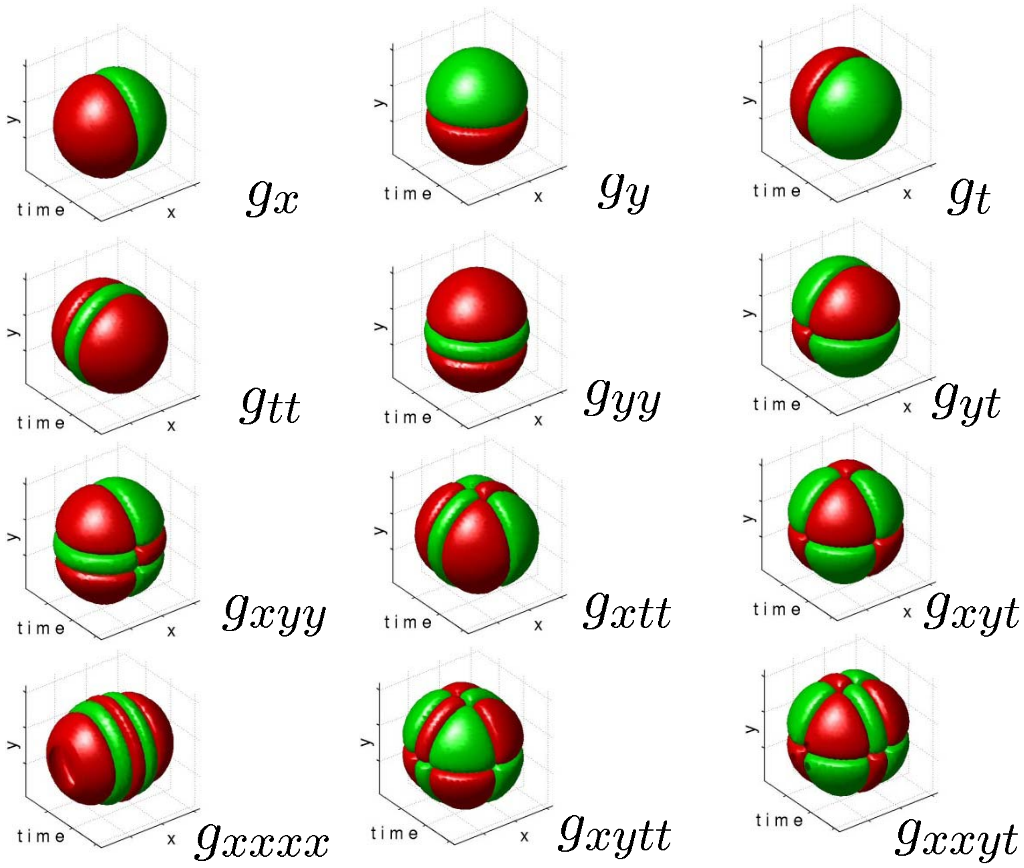
$$D(p) = (L_x(p), L_y(p), L_t(p), L_{xx}(p), ..., L_{tttt}(p))$$

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

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

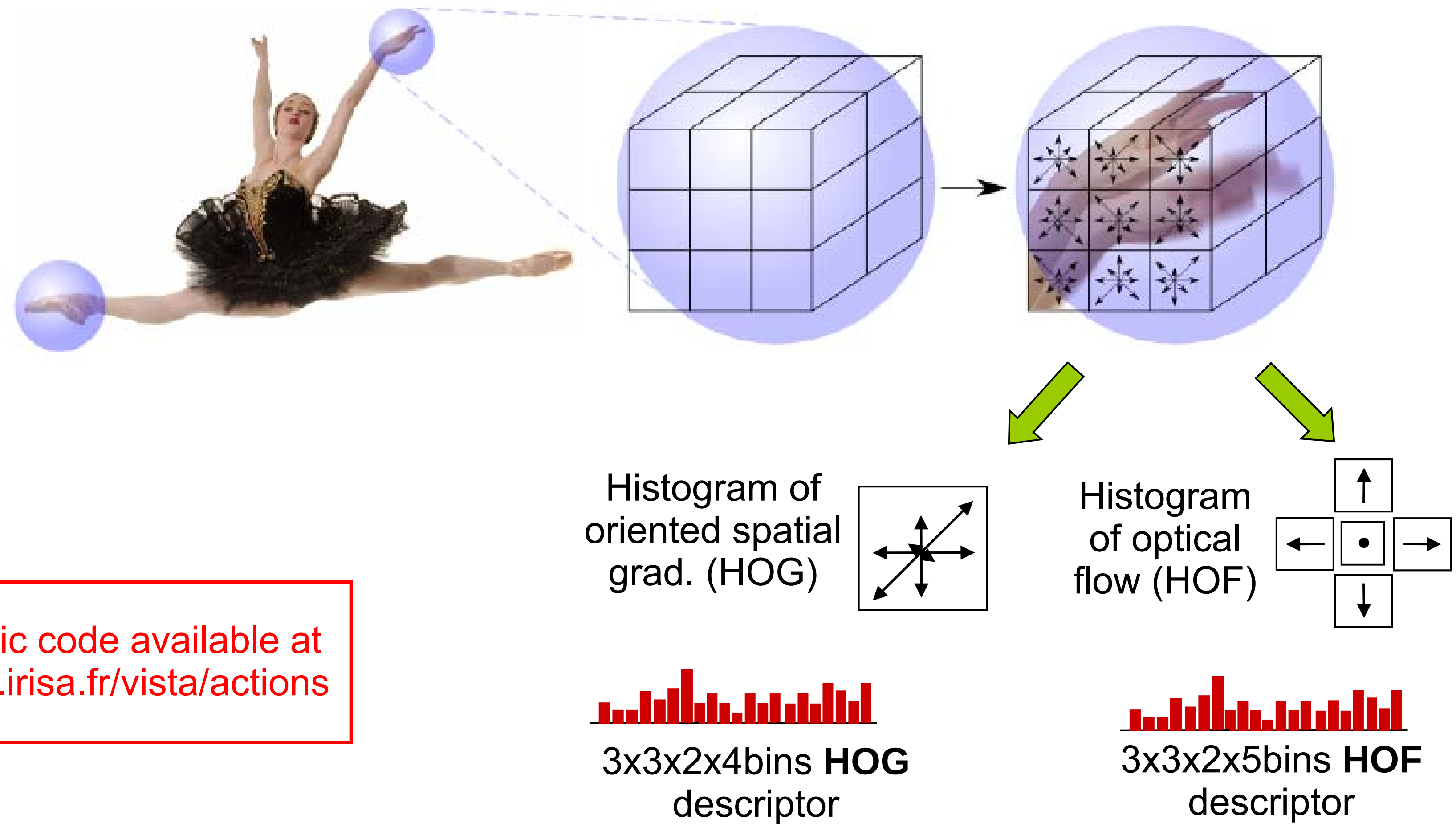
$$\vdots$$

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



Local space-time descriptor: HOG/HOF

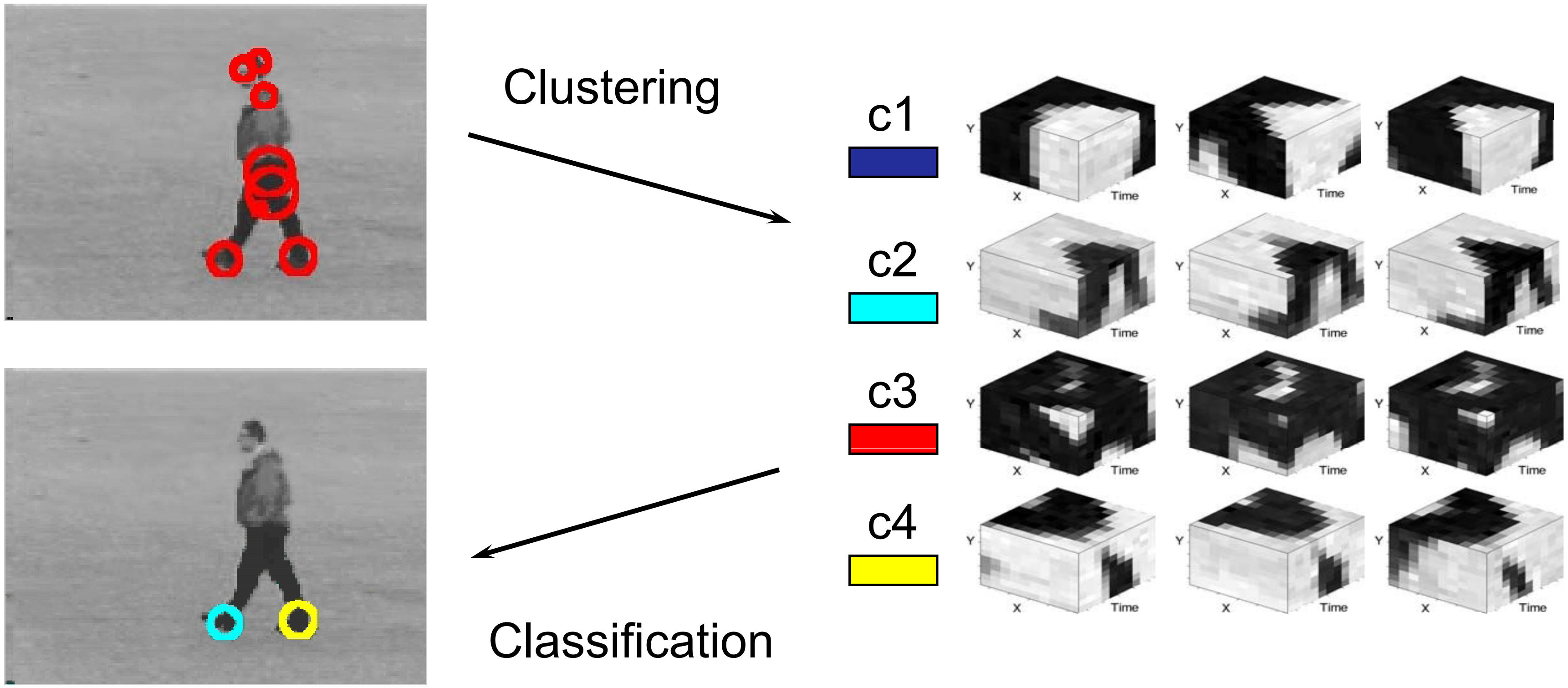
Multi-scale space-time patches



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

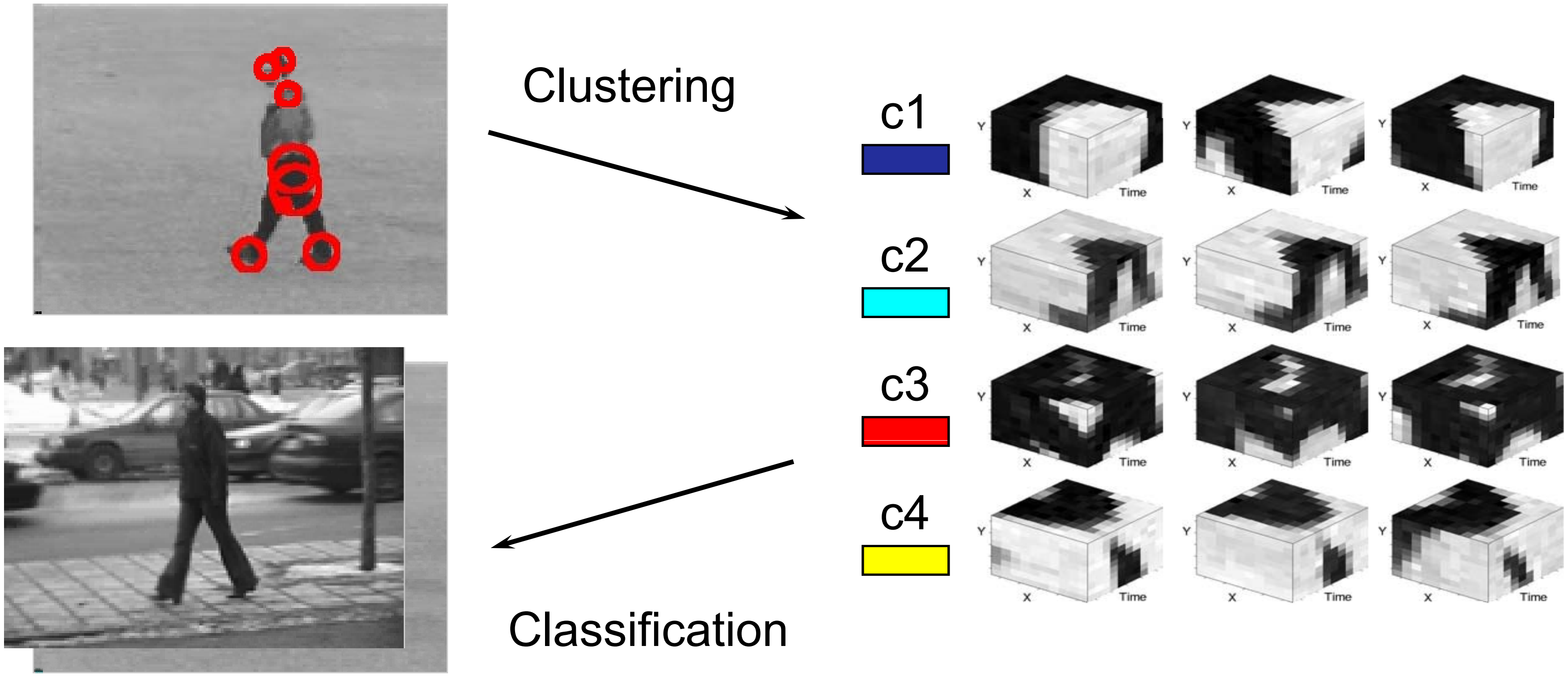
Visual Vocabulary: K-means clustering

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



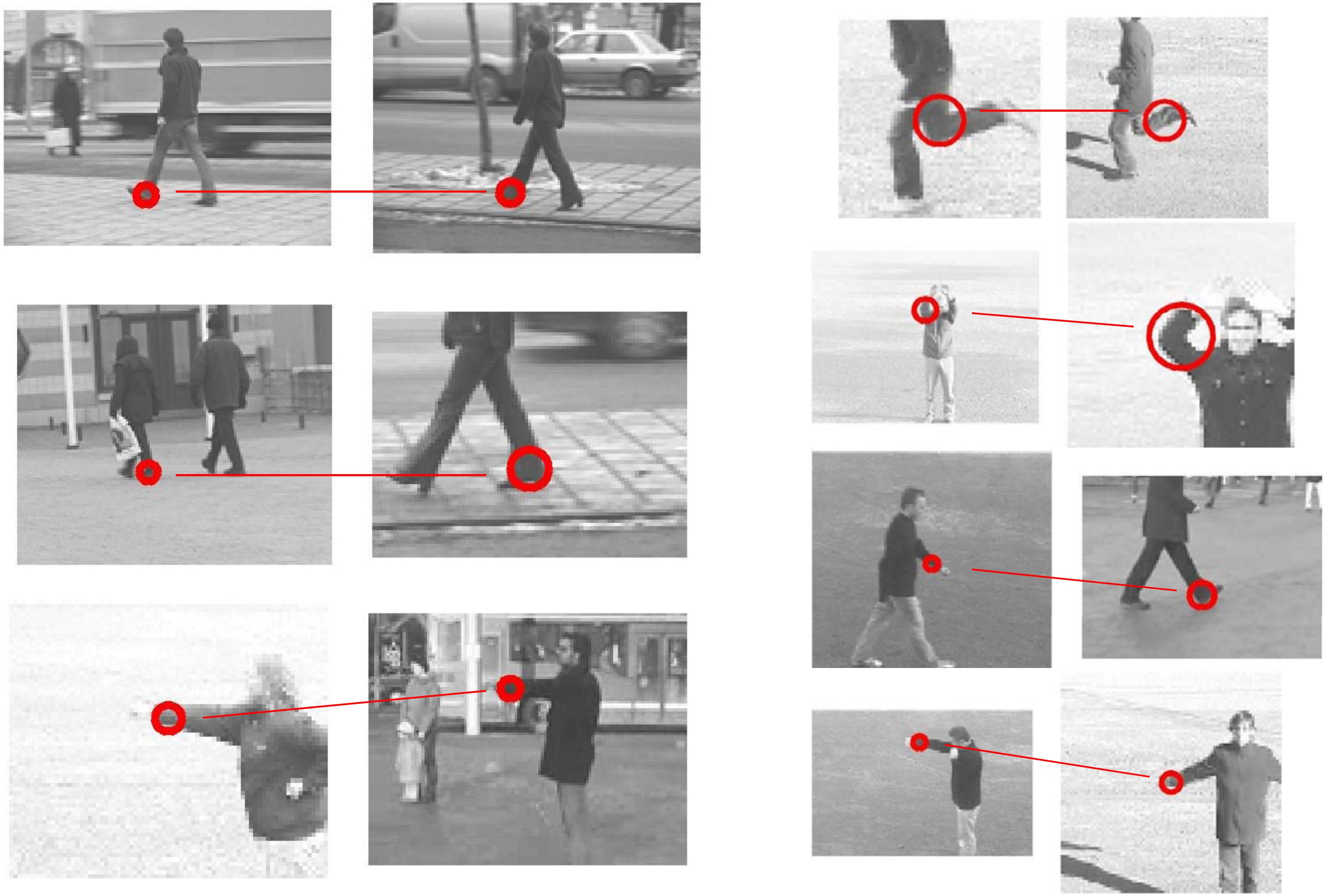
Visual Vocabulary: K-means clustering

- 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

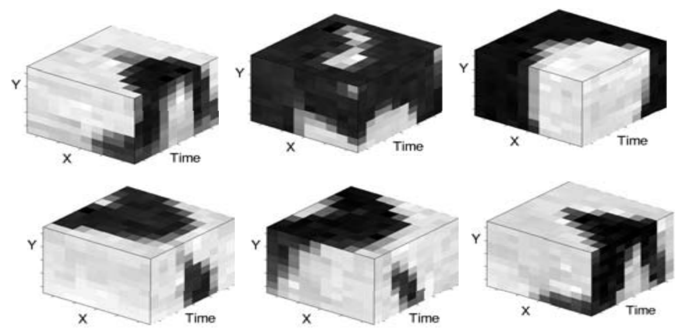


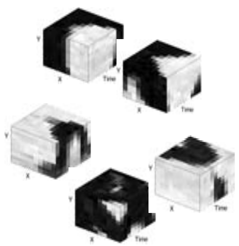
Action Classification: Overview

Bag of space-time features + multi-channel SVM

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

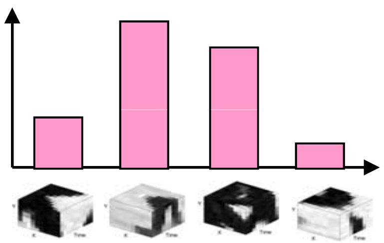
Collection of space-time patches





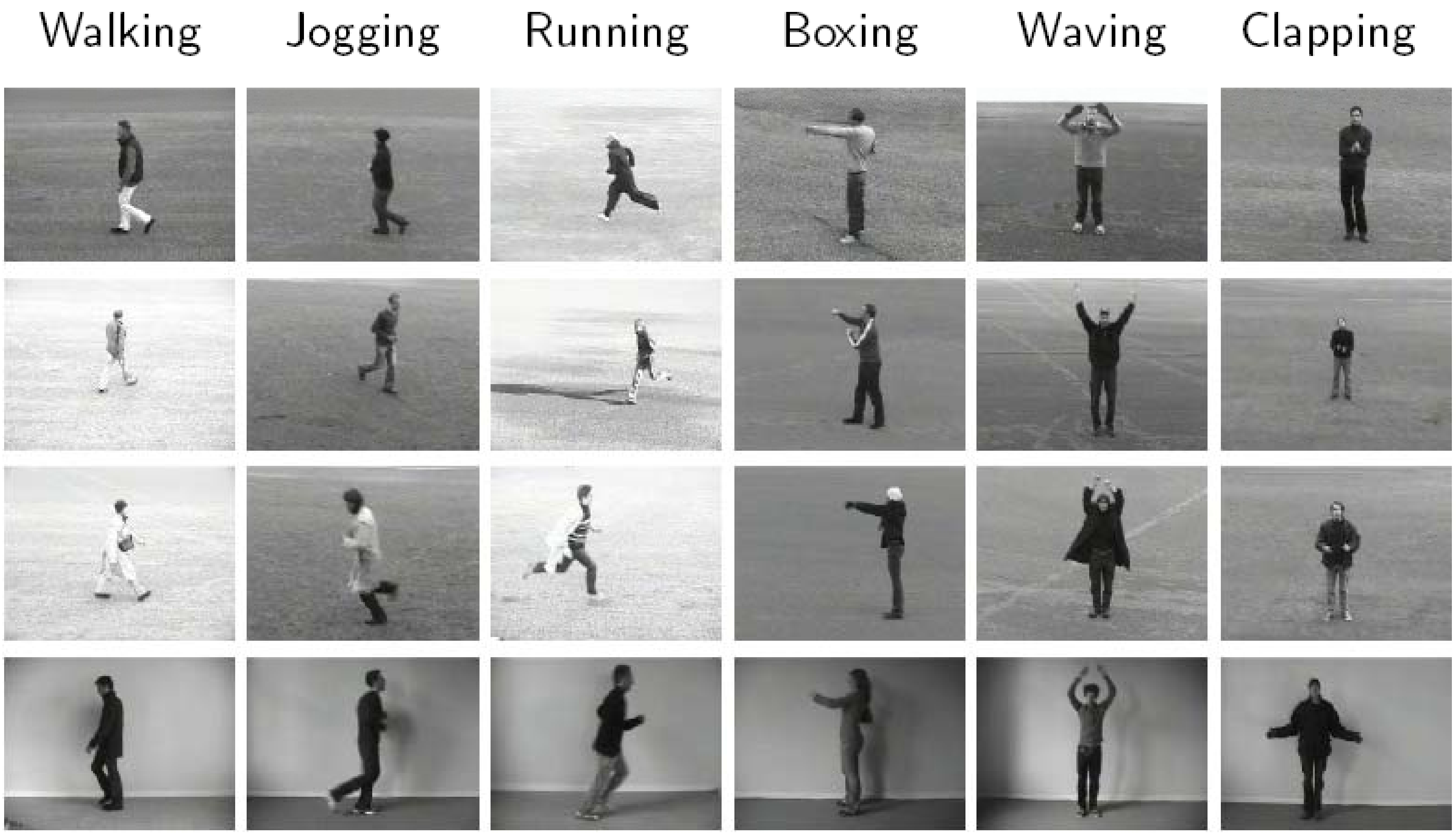
HOG & HOF patch descriptors

Histogram of visual words



Multi-channel SVM Classifier

Action recognition in KTH dataset



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

Classification results on KTH dataset

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

Confusion matrix for KTH actions

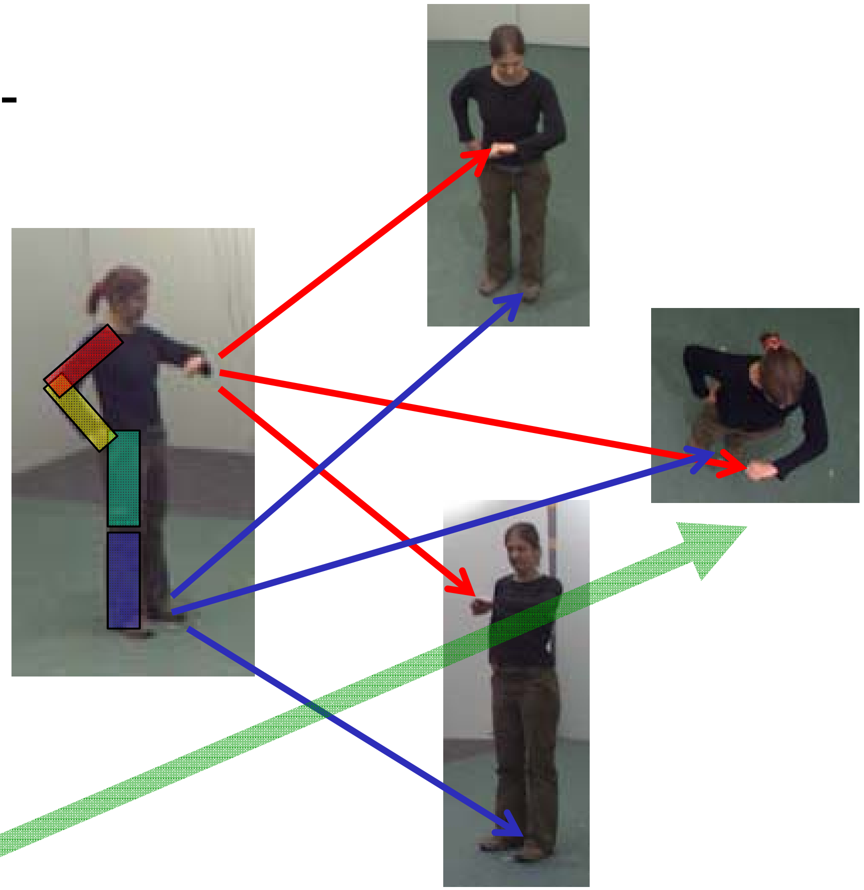
What about 3D?

Local motion and appearance features are not invariant to view changes

Multi-view action recognition

Difficult to apply standard multi-view methods:

- Do not want to search for multi-view point correspondence --- Non-rigid motion, clothing changes, … --> It's Hard!



- Do not want to identify body parts. Current methods are not reliable enough.

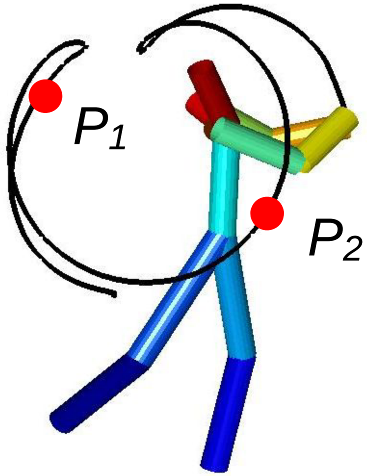
- Yet, want to learn actions from one view
  and recognize actions in very different views

Temporal self-similarities

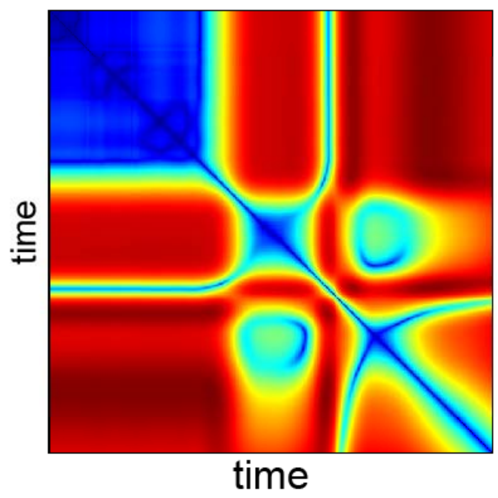
Idea:

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

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

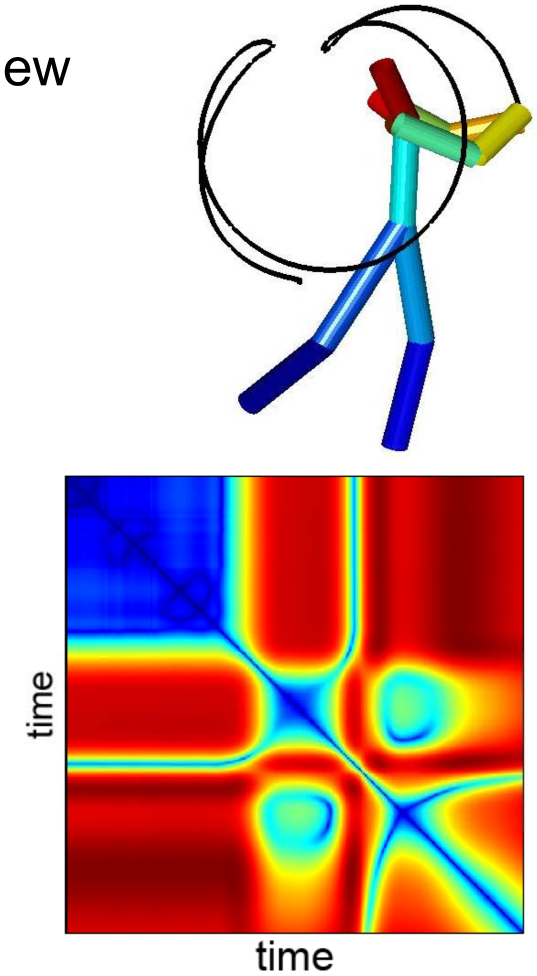


Distance matrix / self-similarity matrix (SSM):

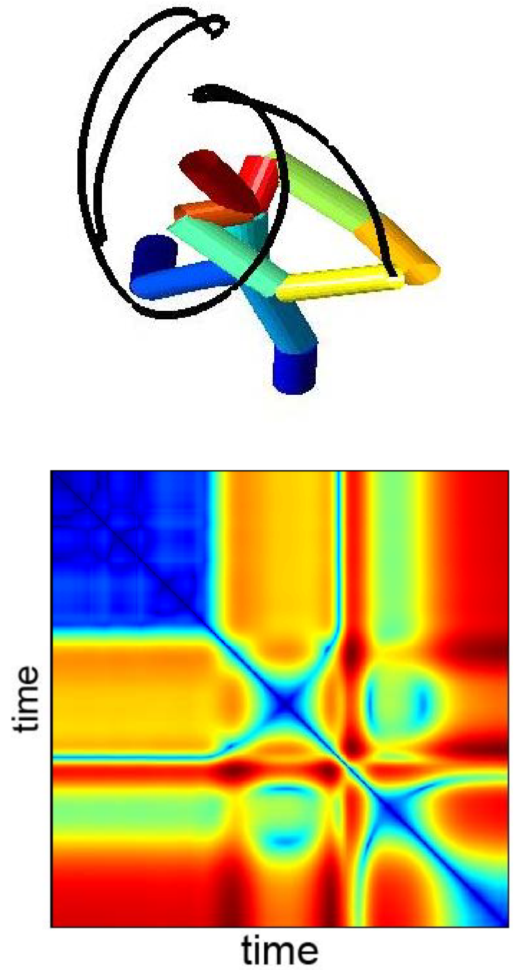


Temporal self-similarities: Multi-views

Side view



Top view

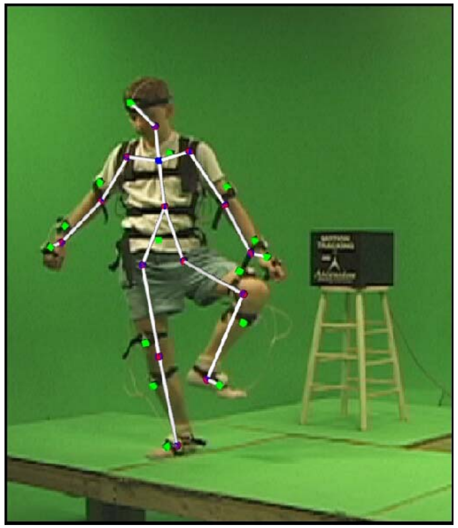


Appear very similar despite the view change!

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

Temporal self-similarities: MoCap

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

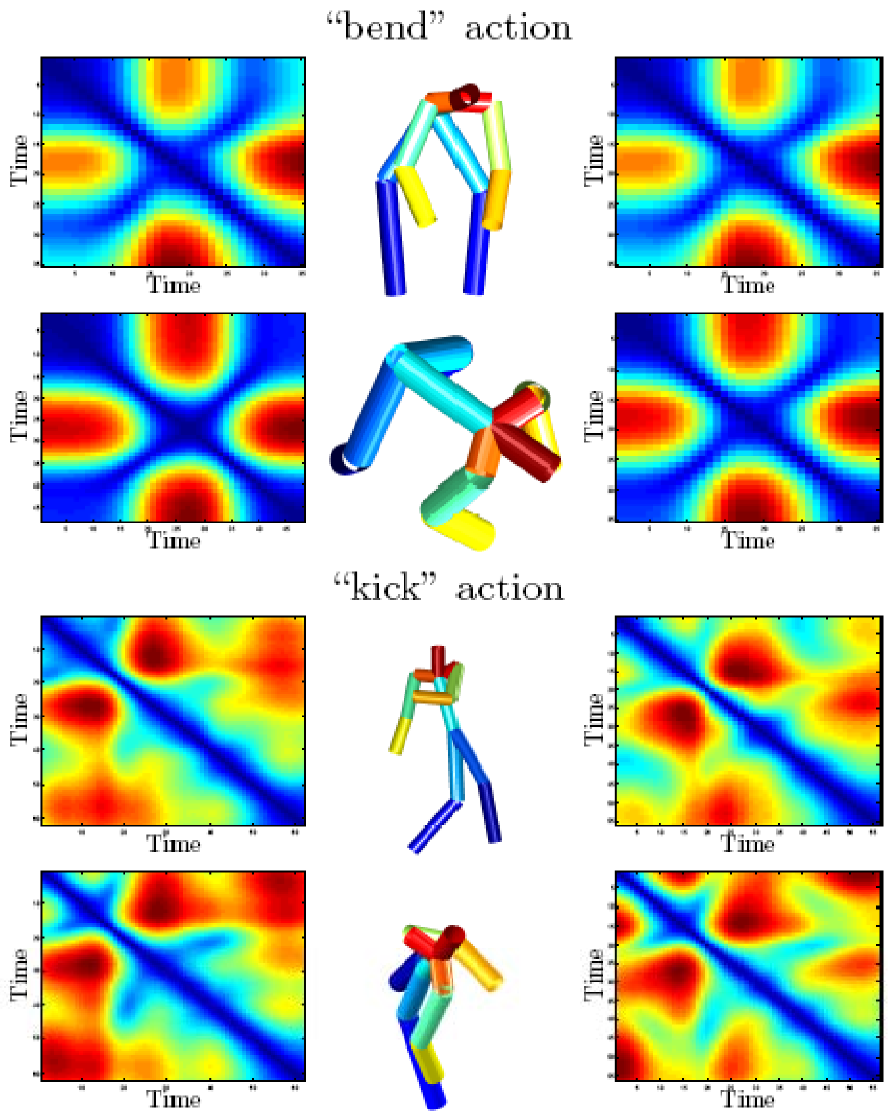


person 1

person 2

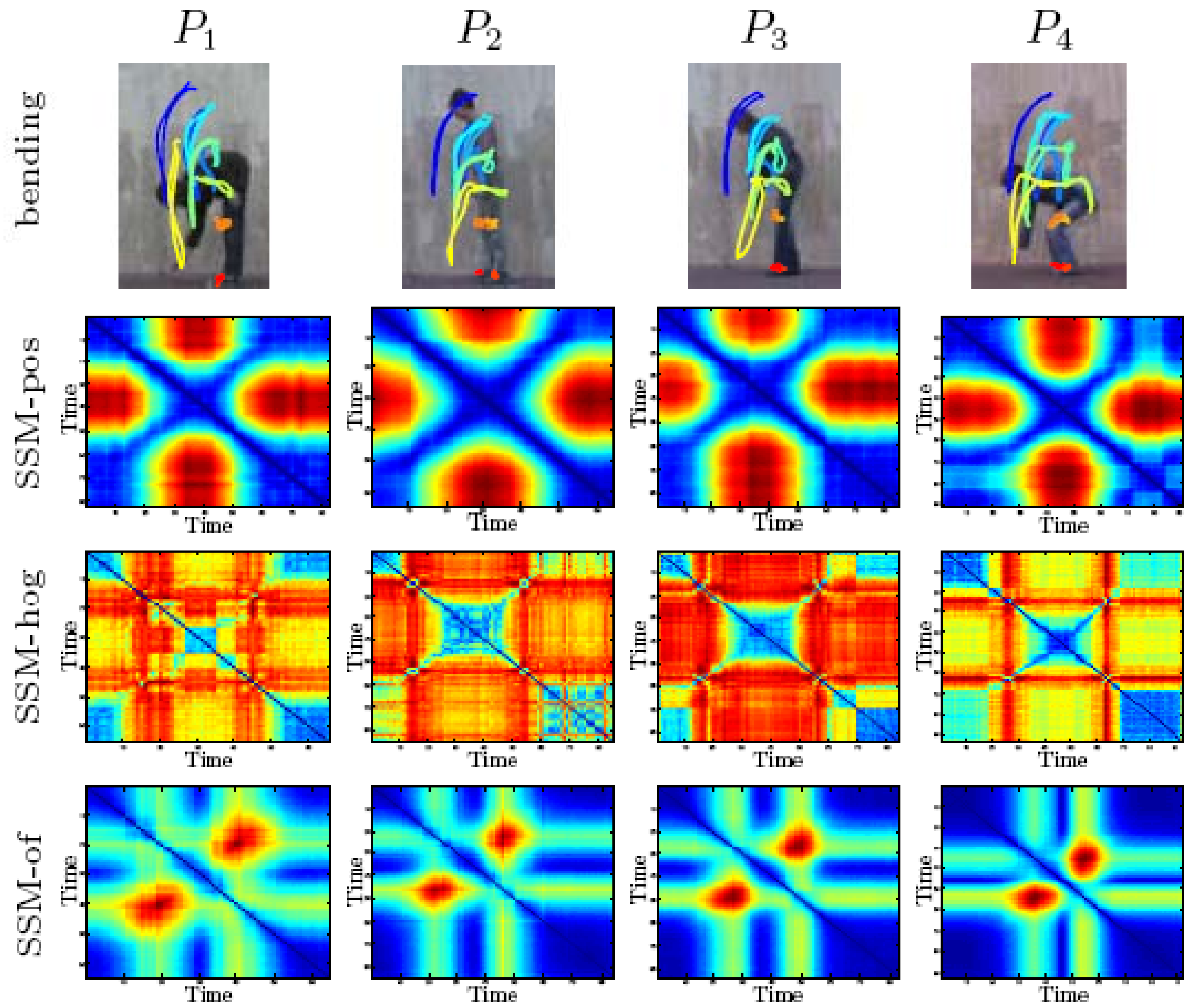
person 1

person 2



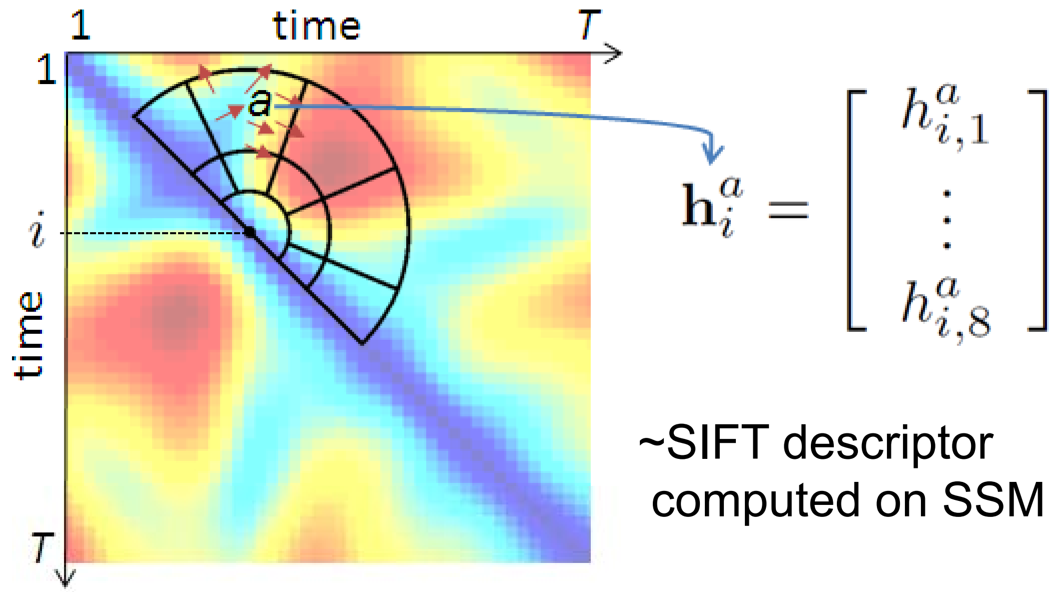
Temporal self-similarities: Video

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



Self-similarity descriptor

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



~SIFT descriptor computed on SSM

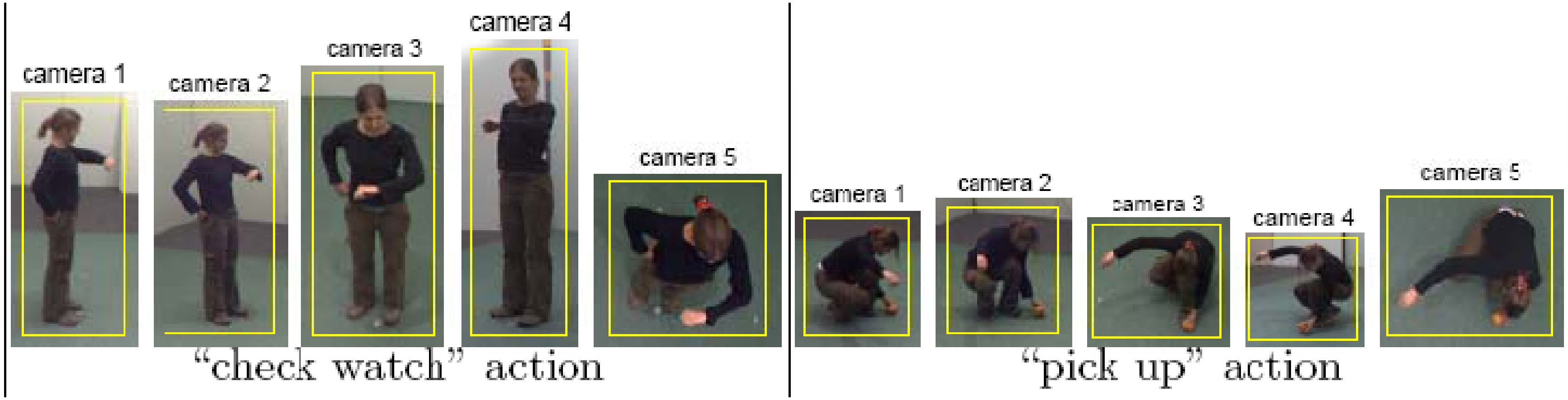
- Define a local histogram descriptor h_i for each point i on the diagonal.

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

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

Multi-view alignment

Multi-view action recognition: Video



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

cross−camera training/testing   same camera training/testing

SSM-based recognition

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

cross−camera training/testing   same camera training/testing

Alternative view-dependent method (STIP)

What are Human Actions?

Actions in recent datasets:

Is it just about kinematics?

Should actions be defined by the *purpose?*

Kinematics + Objects

What are Human Actions?

Actions in recent datasets:

Is it just about kinematics?

Should actions be defined by the *purpose?*

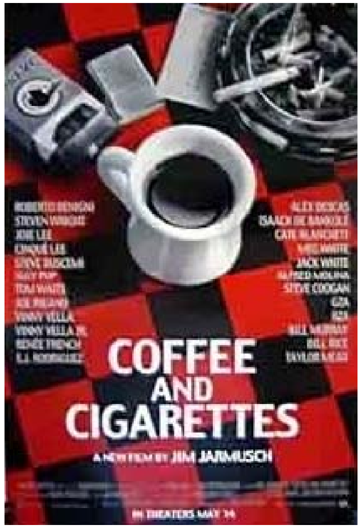
Kinematics + Objects + Scenes

Action recognition in realistic settings

Standard action datasets

Actions “In the Wild”:

Action Dataset and Annotation



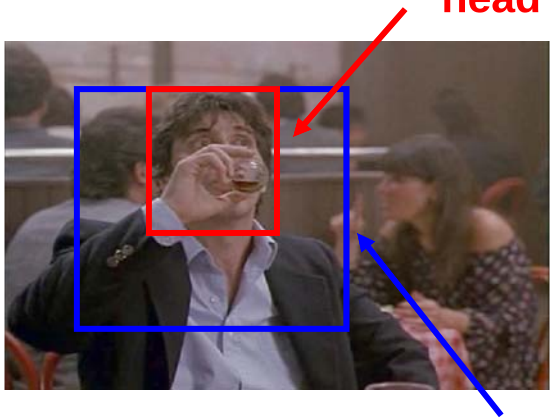
Manual annotation of drinking actions in movies:
"Coffee and Cigarettes"; "Sea of Love"

"*Drinking*": 159 annotated samples

"*Smoking*": 149 annotated samples

Spatial annotation

head rectangle



torso rectangle

Temporal annotation

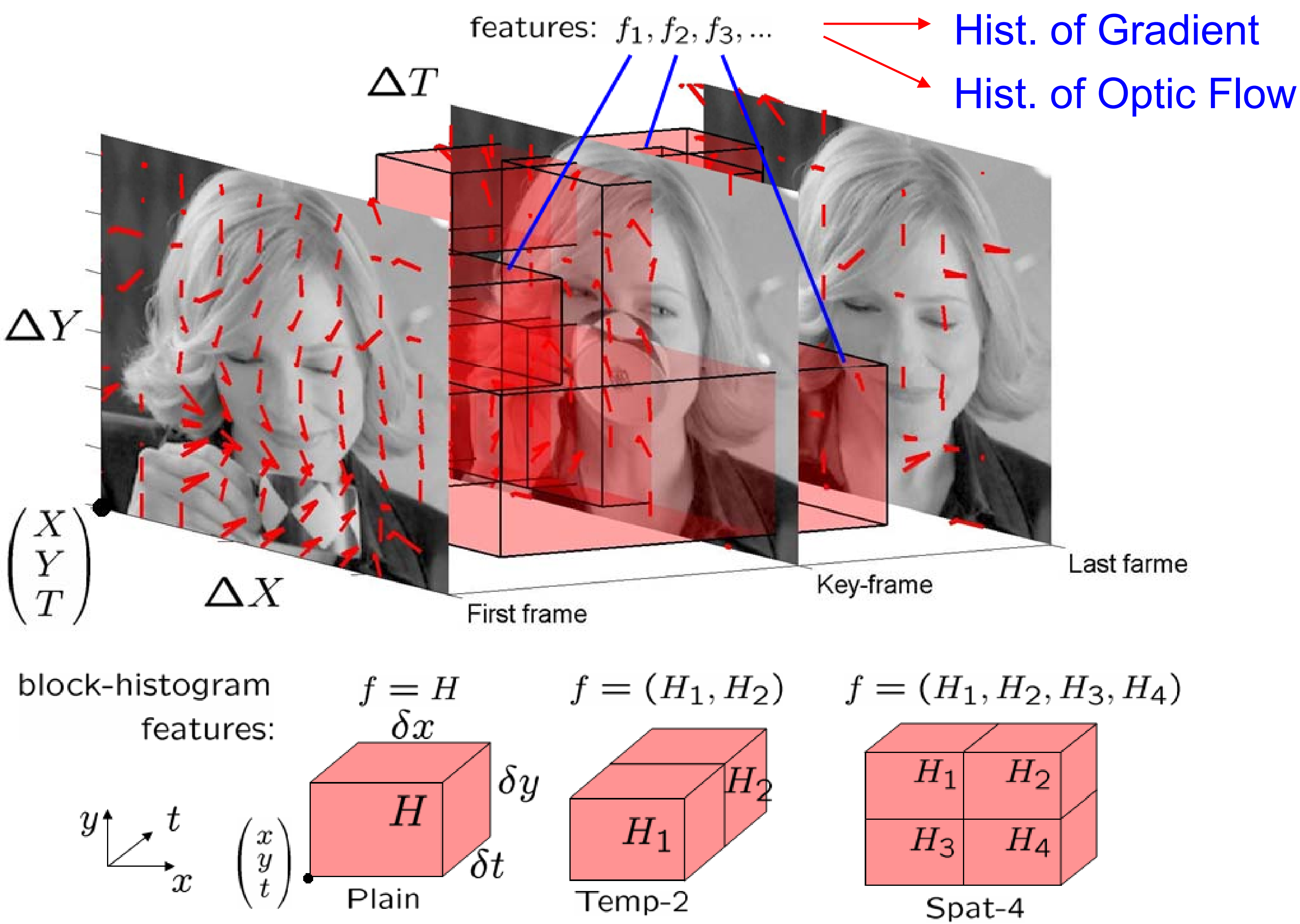
First frame **Keyframe** **Last frame**

“Drinking” action samples

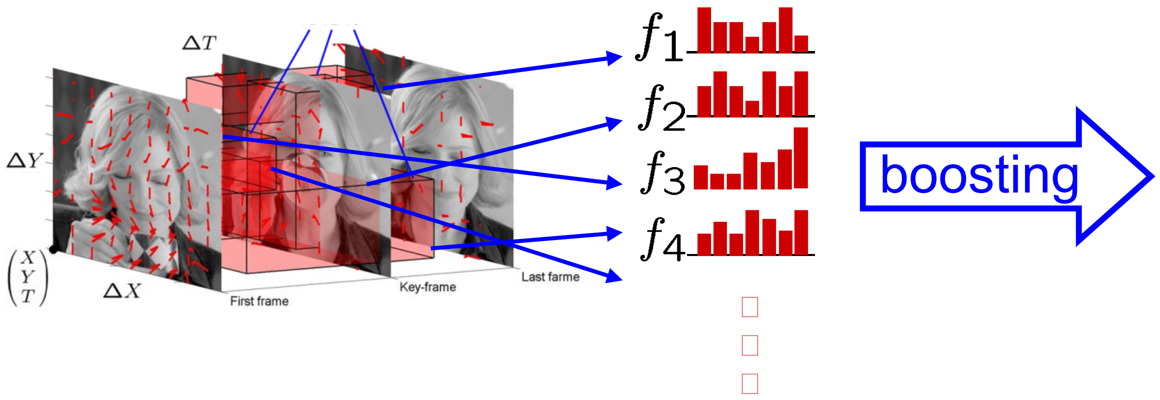
training samples

test samples

Action representation



Action learning



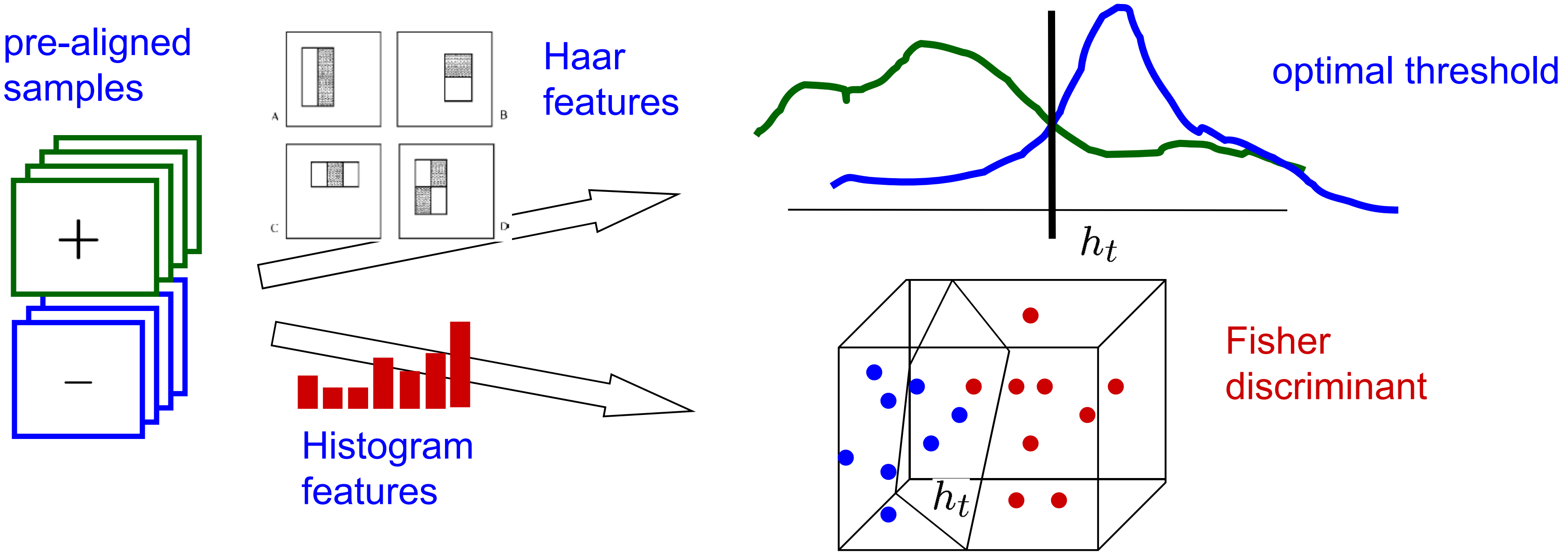
selected features

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

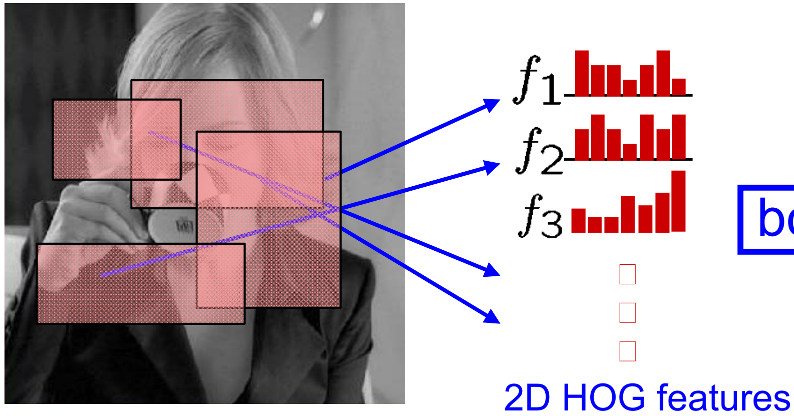
weak classifier

AdaBoost:

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



Key-frame action classifier



boosting

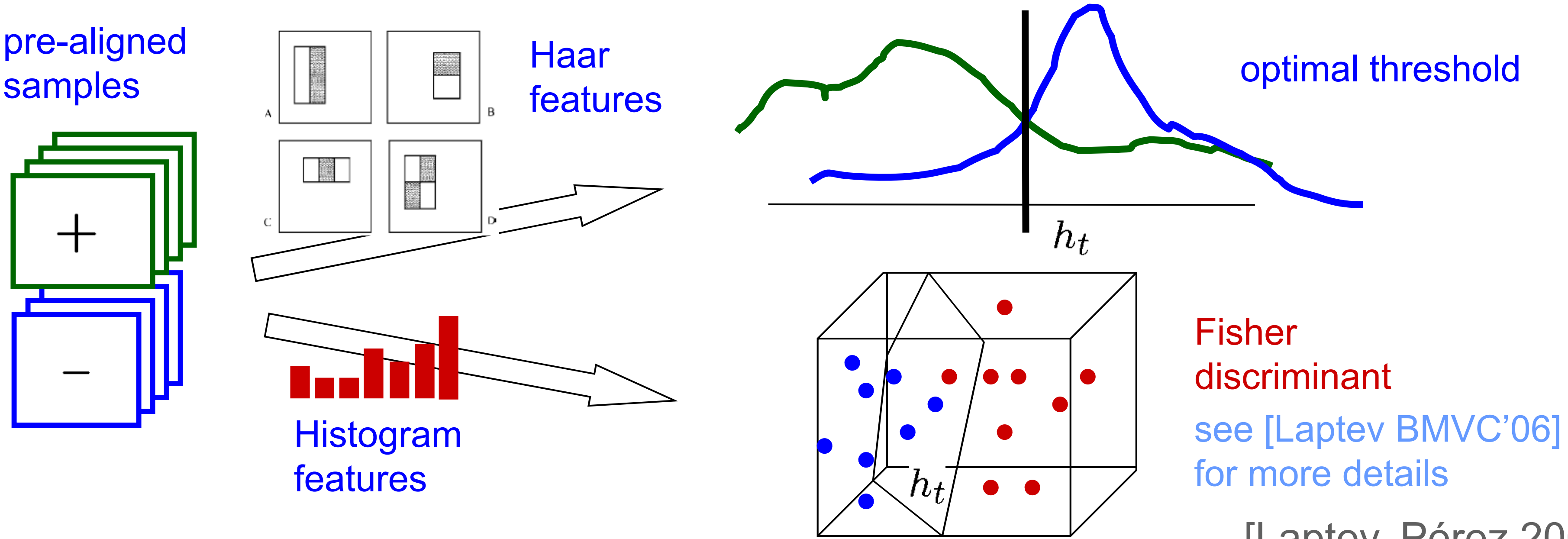
selected features

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

weak classifier

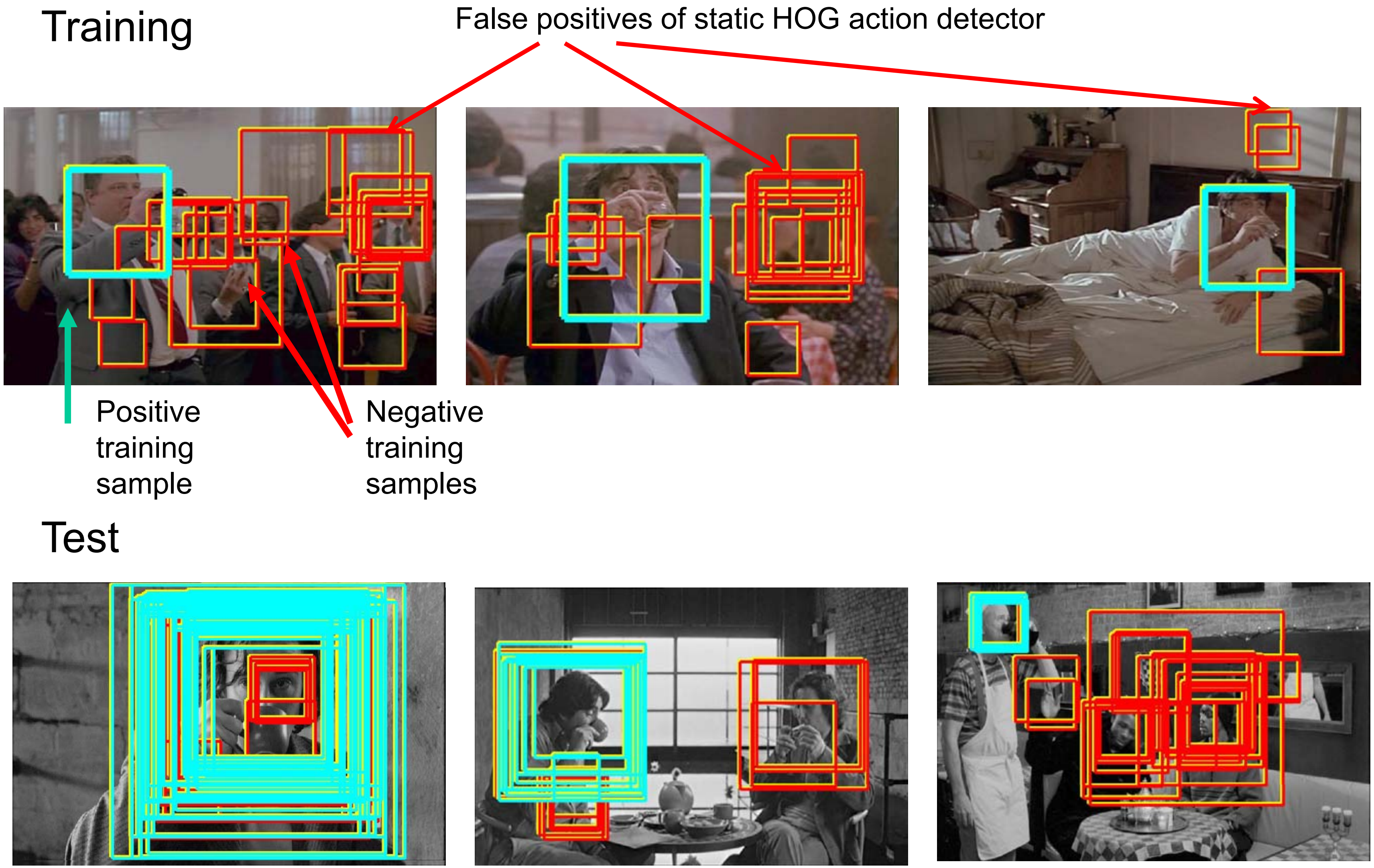
AdaBoost:

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



[Laptev, Pérez 2007]

Keyframe priming



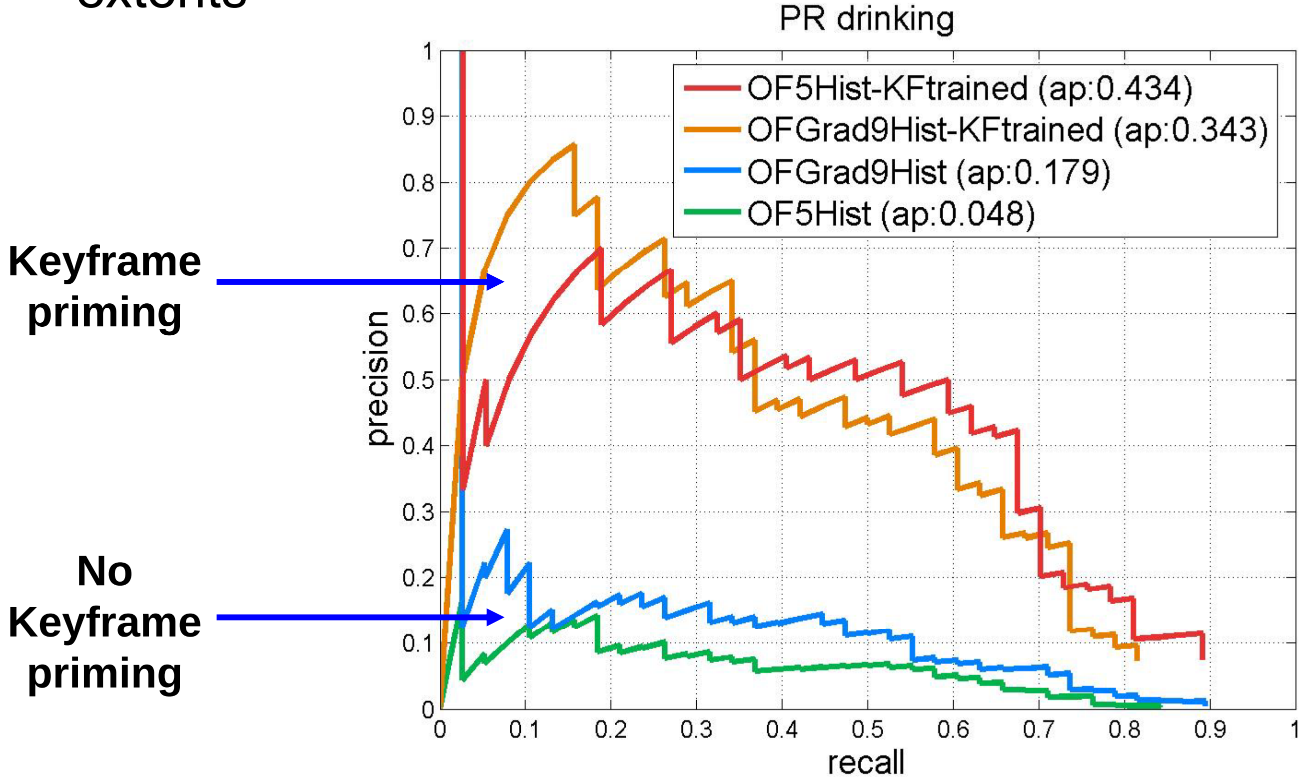
Action detection

Test set:

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

Detection:

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



Action Detection (ICCV 2007)

Test episodes from the movie “Coffee and cigarettes”

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

20 most confident detections

Learning Actions from Movies

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

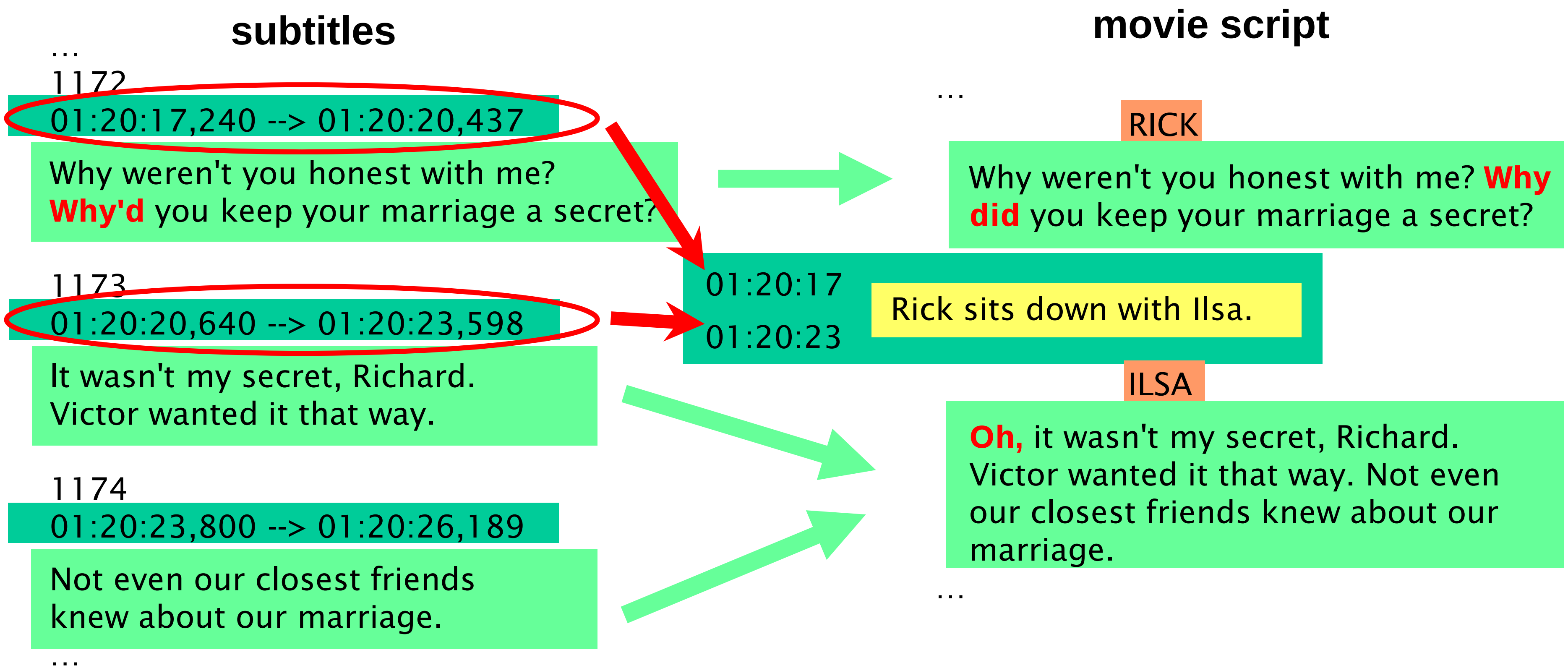
Problems:

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

Automatic video annotation with scripts

- Scripts available for >500 movies (no time synchronization)

  www.dailyscript.com, www.movie-page.com, www.weeklyscript.com …
- Subtitles (with time info.) are available for the most of movies
- Can transfer time to scripts by text alignment



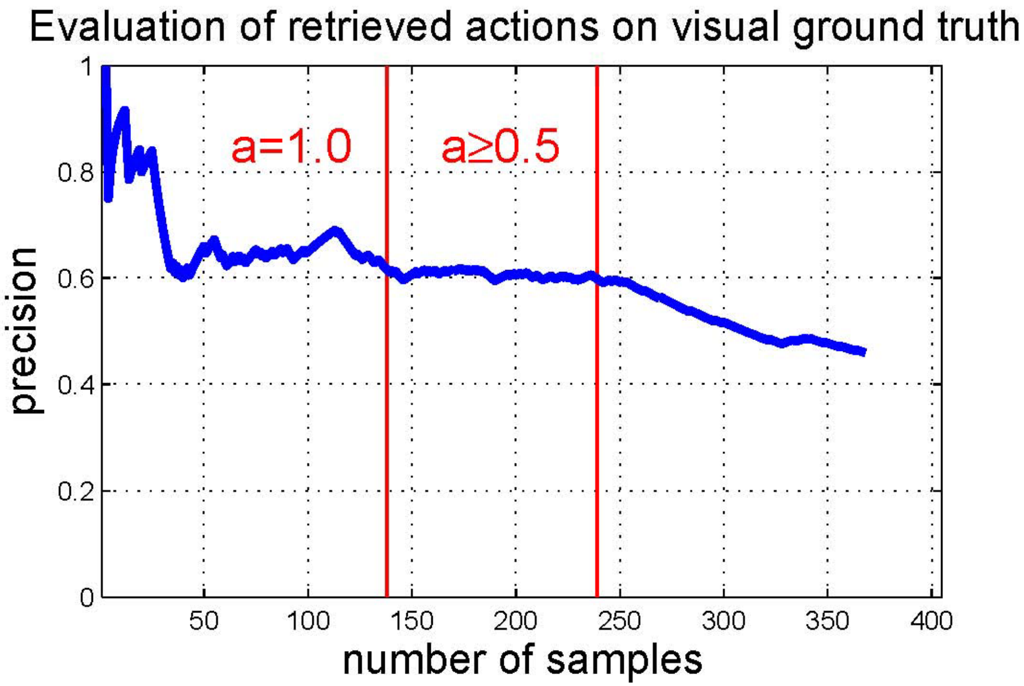
Script-based action annotation

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

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

Script alignment: Evaluation

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



a: quality of subtitle-script matching

Example of a “visual false positive”

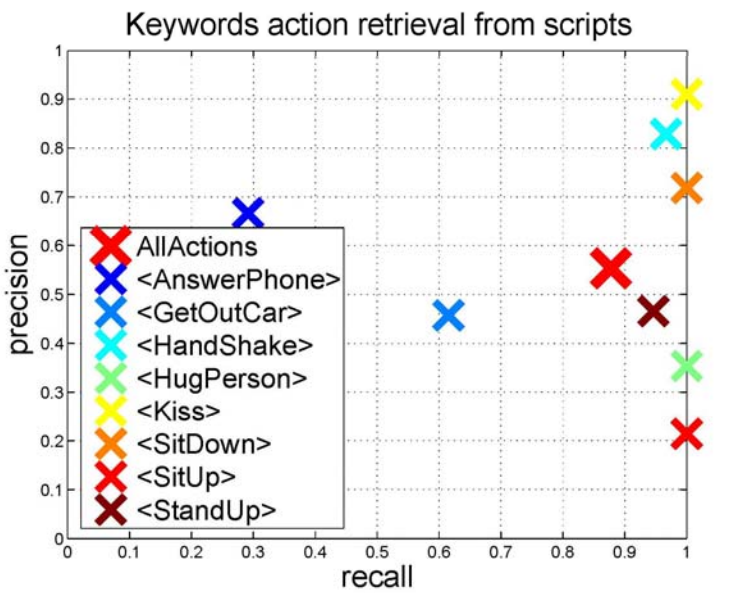
A black car pulls up, two army officers get out.

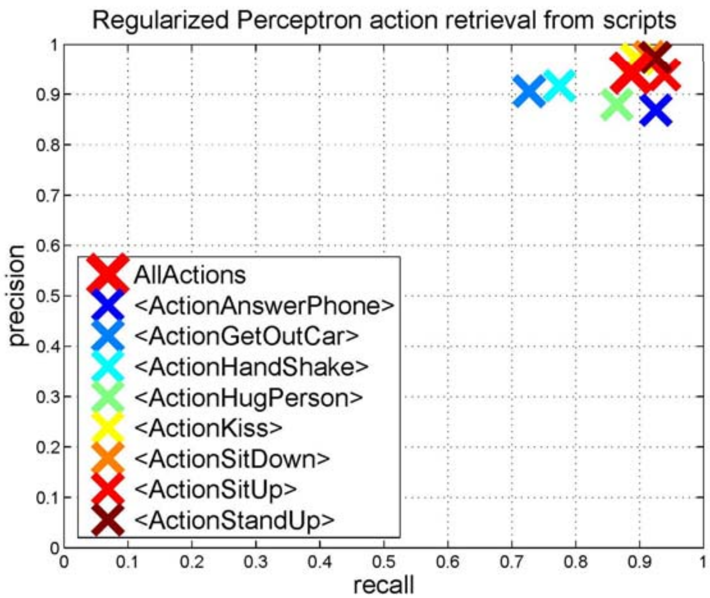
Text-based action retrieval

- Large variation of action expressions in text:

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

- => Supervised text classification approach





Automatically annotated action samples

AnswerPhone

GetOutCar

HandShake

HugPerson

Kiss

SitDown

SitUp

StandUp

[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Hollywood-2 actions dataset

Actions			
	Training subset (clean)	Training subset (automatic)	Test subset (clean)
AnswerPhone	66	59	64
DriveCar	85	90	102
Eat	40	44	33
FightPerson	54	33	70
GetOutCar	51	40	57
HandShake	32	38	45
HugPerson	64	27	66
Kiss	114	125	103
Run	135	187	141
SitDown	104	87	108
SitUp	24	26	37
StandUp	132	133	146
All Samples	**823**	**810**	**884**

Training and test samples are obtained from 33 and 36 distinct movies respectively.

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

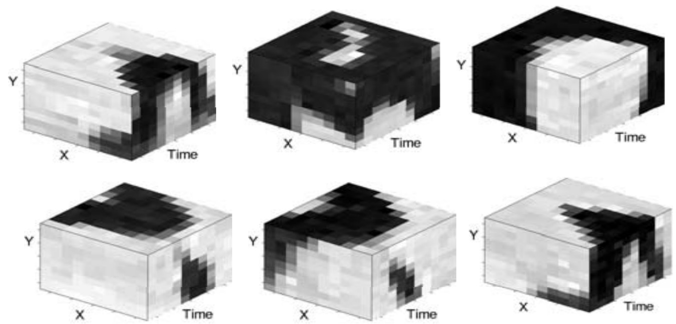
[Laptev, Marszałek, Schmid, Rozenfeld 2008]

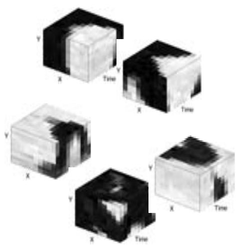
Action Classification: Overview

Bag of space-time features + multi-channel SVM

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

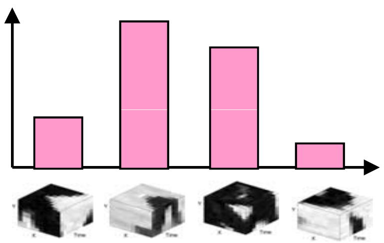
Collection of space-time patches





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

01:22:00
01:22:03

ILSA

I wish I didn't love you so much.

She snuggles closer to Rick.

CUT TO:

EXT. RICK'S CAFE - NIGHT

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

The headlights of a speeding police car sweep toward them.

They flatten themselves against a wall to avoid detection.

The lights move past them.

01:22:15
01:22:17

CARL

I think we lost them.

...

Mining scene captions

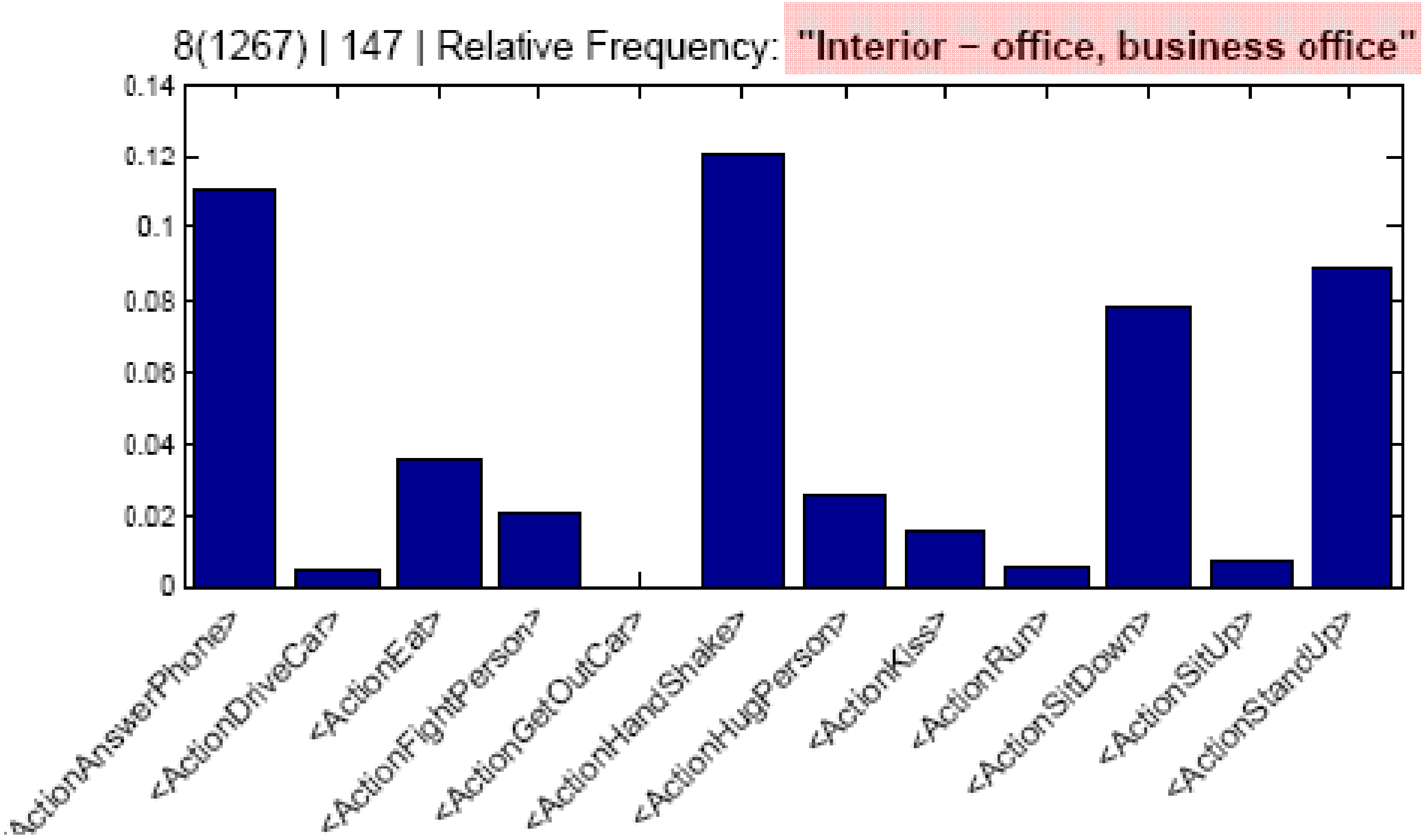
INT. TRENDY RESTAURANT - NIGHT
INT. MARSELLUS WALLACE'S DINING ROOM MORNING
EXT. STREETS BY DORA'S HOUSE - DAY.
INT. MELVIN'S APARTMENT, BATHROOM – NIGHT
EXT. NEW YORK CITY STREET NEAR CAROL'S RESTAURANT – DAY
INT. CRAIG AND LOTTE'S BATHROOM - DAY

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

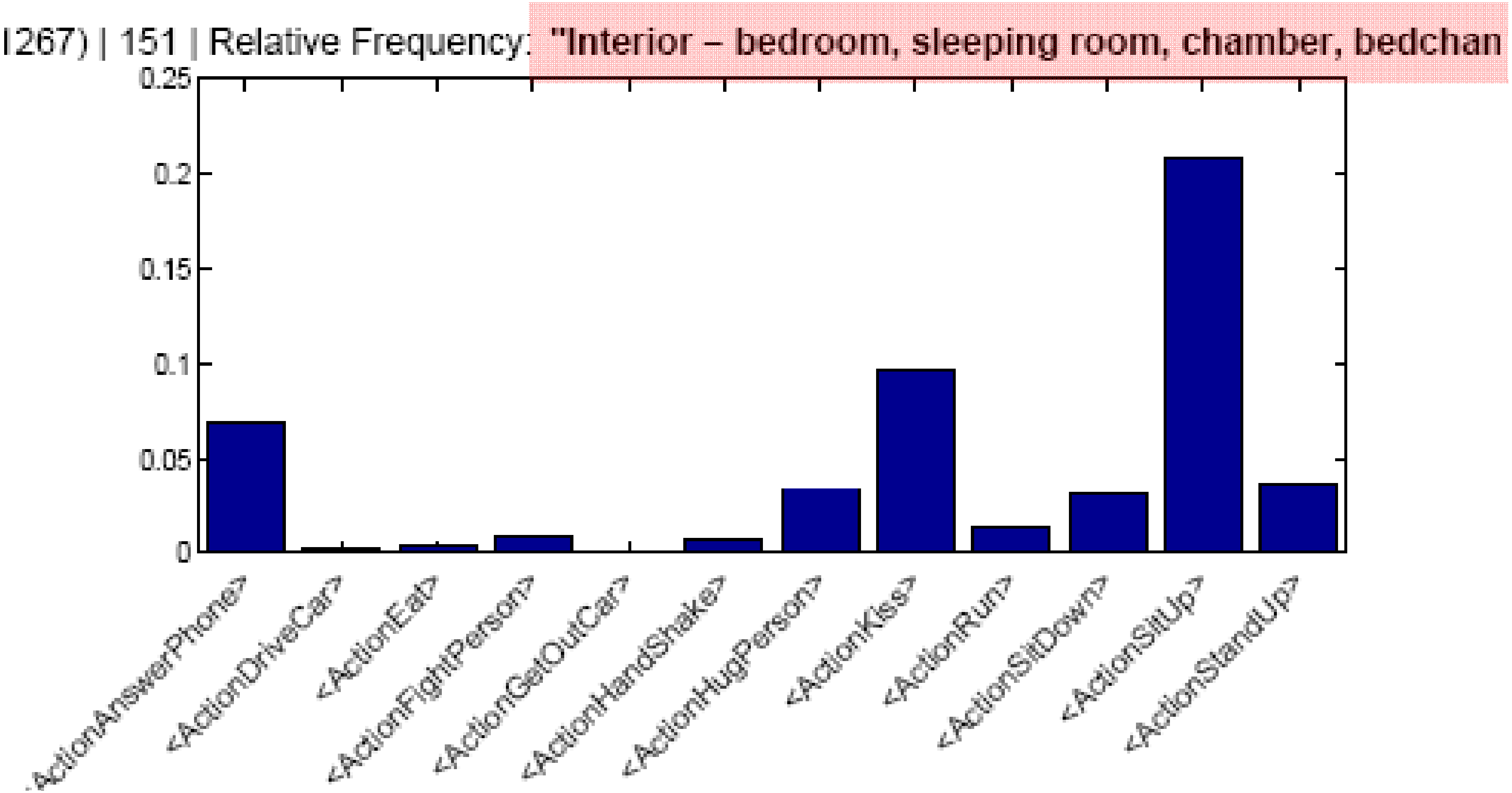
  taxi -> car, cafe -> restaurant

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

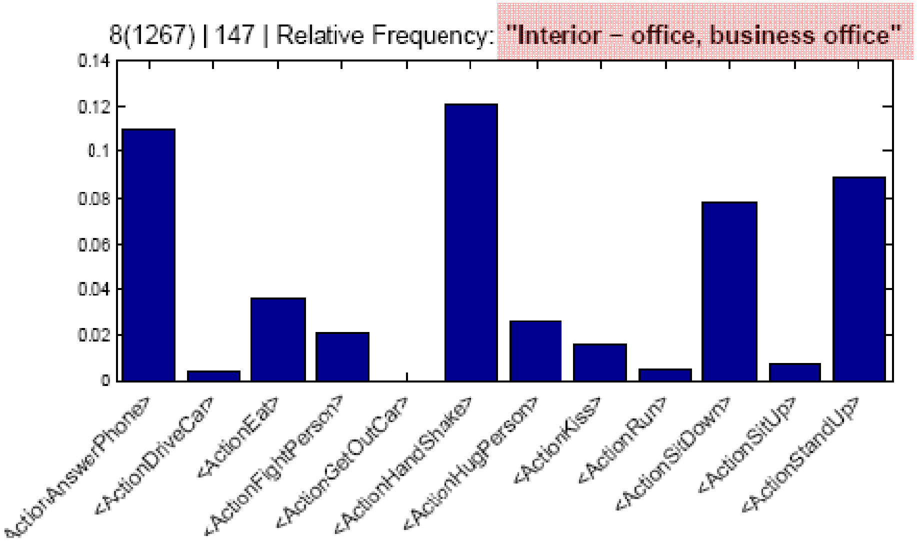
Co-occurrence of actions and scenes in scripts

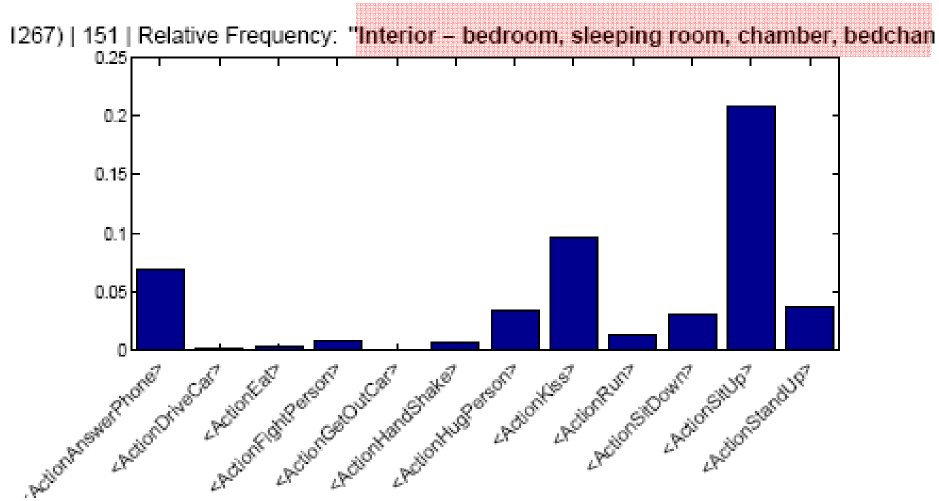


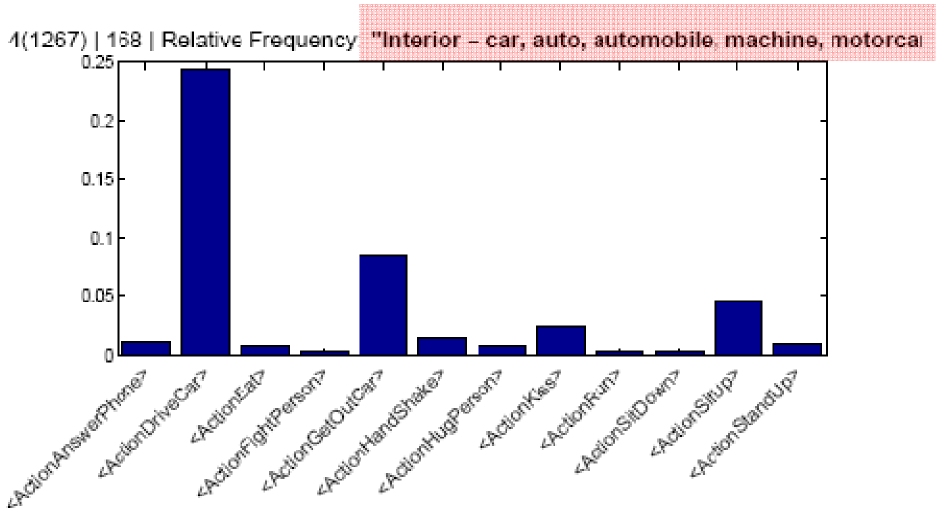
Co-occurrence of actions and scenes in scripts

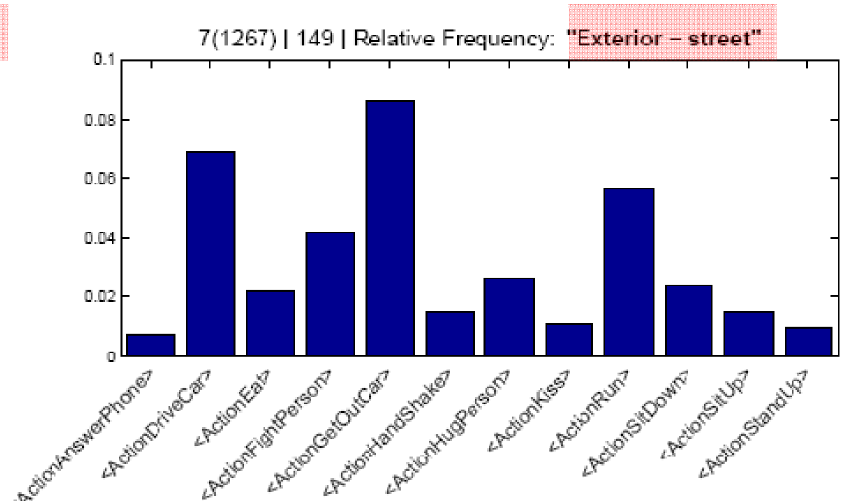


Co-occurrence of actions and scenes in scripts

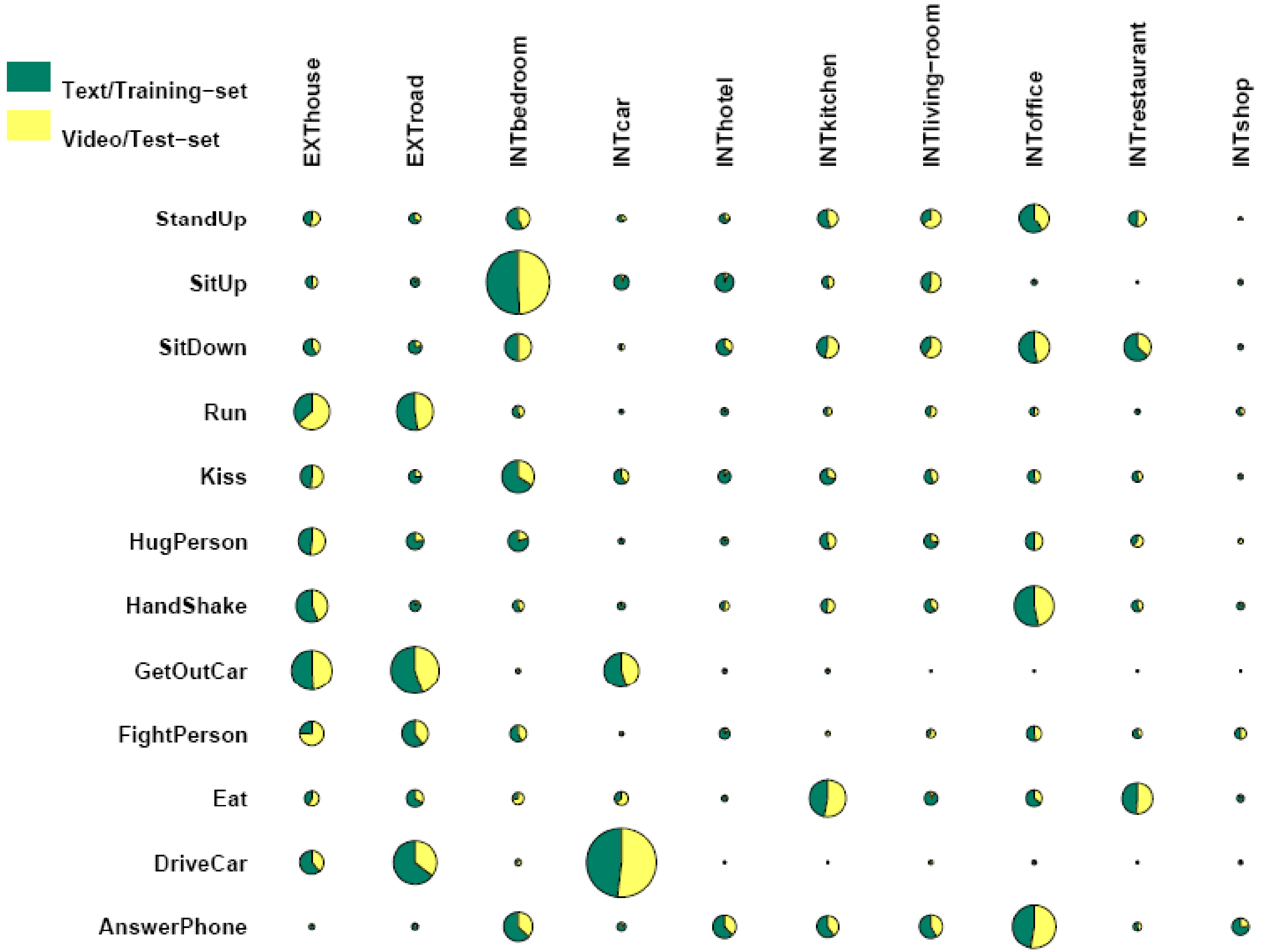








Co-occurrence of actions and scenes in text vs. video



Automatic gathering of relevant scene classes and visual samples

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

(a) Actions

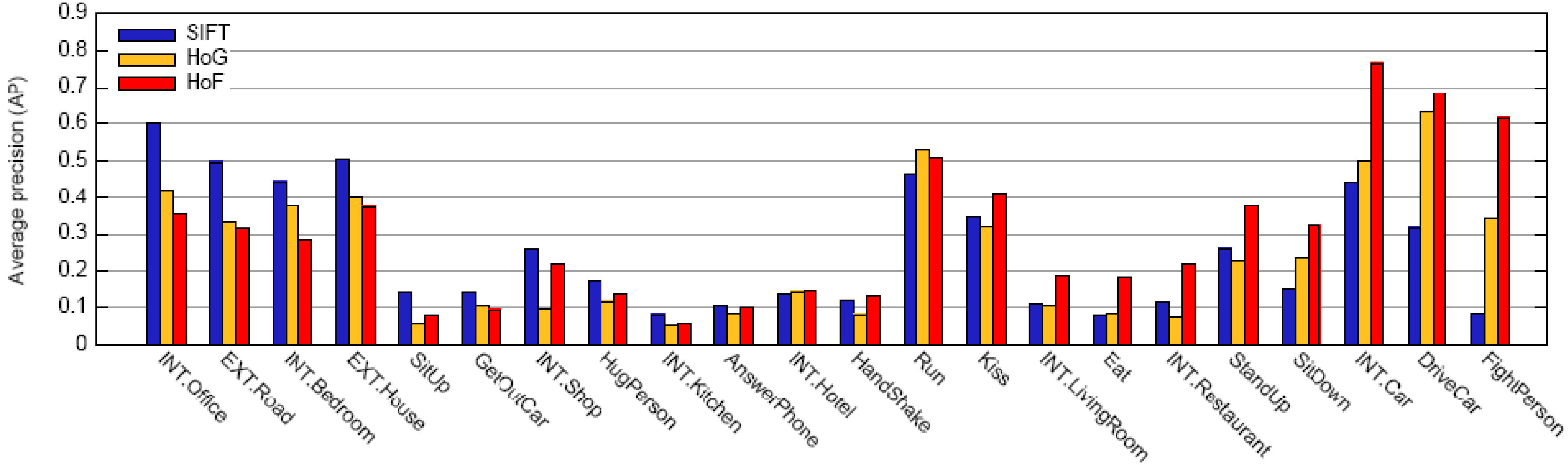
	Auto-Train-Scenes	Clean-Test-Scenes
EXT-house	81	140
EXT-road	81	114
INT-bedroom	67	69
INT-car	44	68
INT-hotel	59	37
INT-kitchen	38	24
INT-living-room	30	51
INT-office	114	110
INT-restaurant	44	36
INT-shop	47	28
All Samples	570	582

(b) Scenes

Source:
69 movies
aligned with
the scripts

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

Results: actions and scenes (separately)



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

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

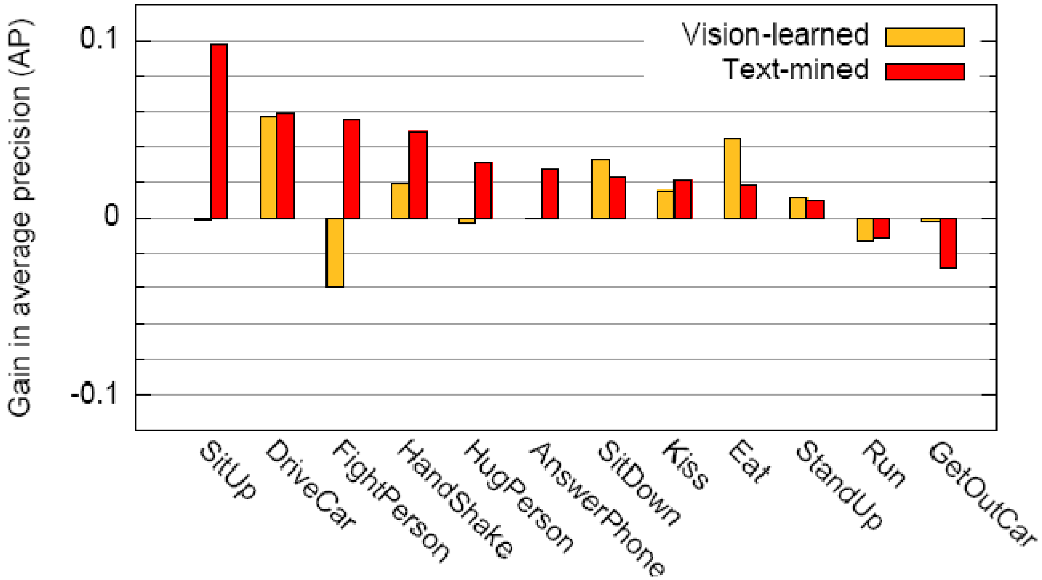
Classification with the help of context

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

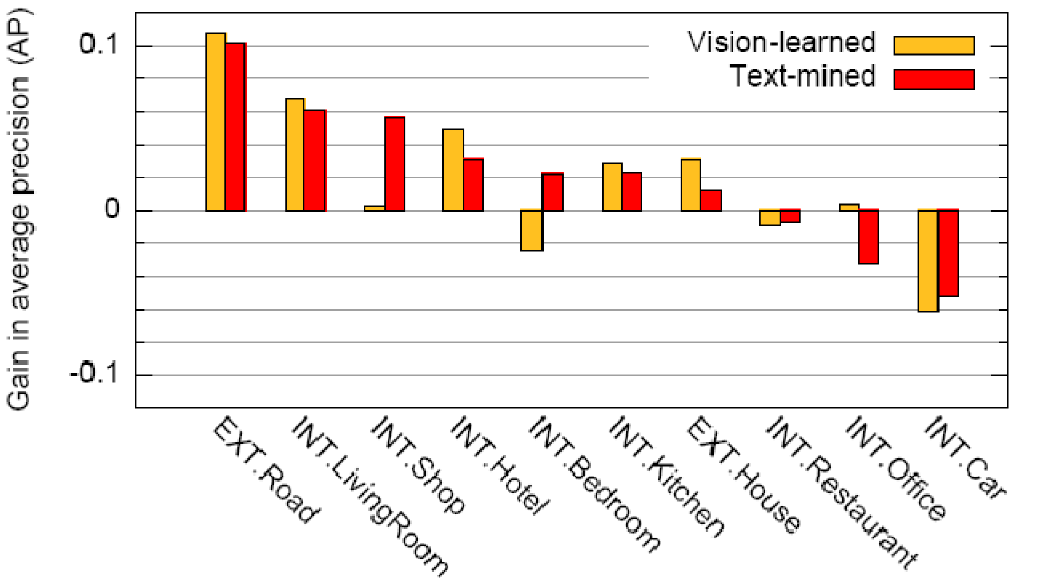
$a_i(\boldsymbol{x})$	Action classification score
$s_j(\boldsymbol{x})$	Scene classification score
w_{ij}	Weight, estimated from text: $p(Scene \mid Action)$
$a'_i(\boldsymbol{x})$	New action score

Results: actions and scenes (jointly)

Actions in the context of Scenes

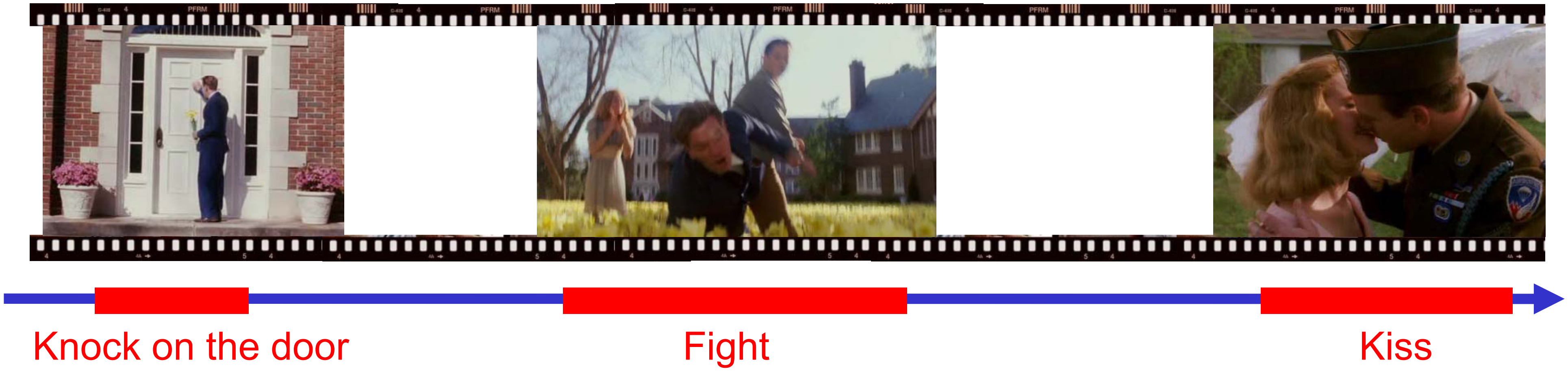


Scenes in the context of Actions



Weakly-Supervised
Temporal Action Annotation

- Answer questions: WHAT *actions and* WHEN *they happened* ?



- Train visual action detectors and annotate actions with the minimal manual supervision

WHAT actions?

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

  -- Text processing:

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

  -- Search action patterns

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

WHEN: Video Data and Annotation

- Want to target realistic video data
- Want to avoid manual video annotation for training

➡ Use movies + scripts for automatic annotation of training samples

Subtitles

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

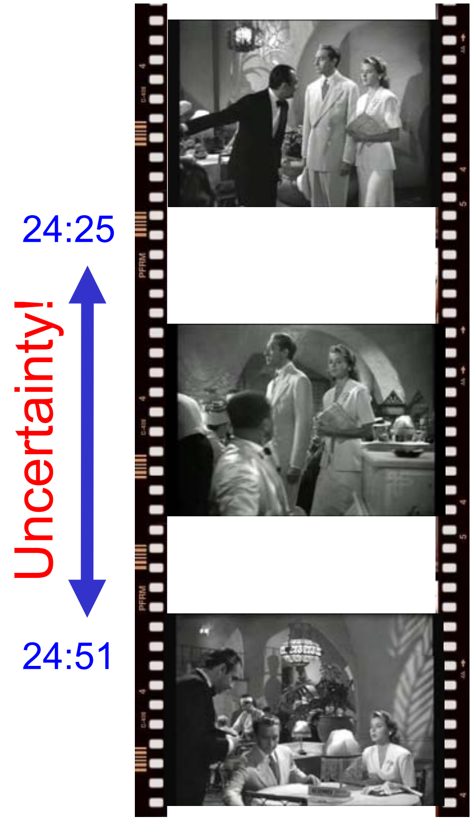
00:24:51 –> 00:24:53
Two Cointreaux, please.

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.



Overview

Input:

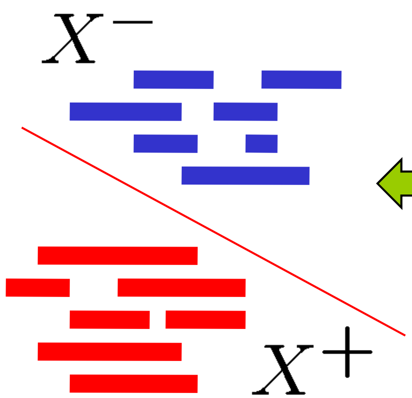
- Action type, e.g.
  Person Opens Door
- Videos + aligned scripts

Automatic collection of training clips

... **Jane** jumps up and **opens** the **door** ...
... **Carolyn opens** the front **door** ...
... **Jane opens** her bedroom **door** ...

Clustering of positive segments

Training classifier

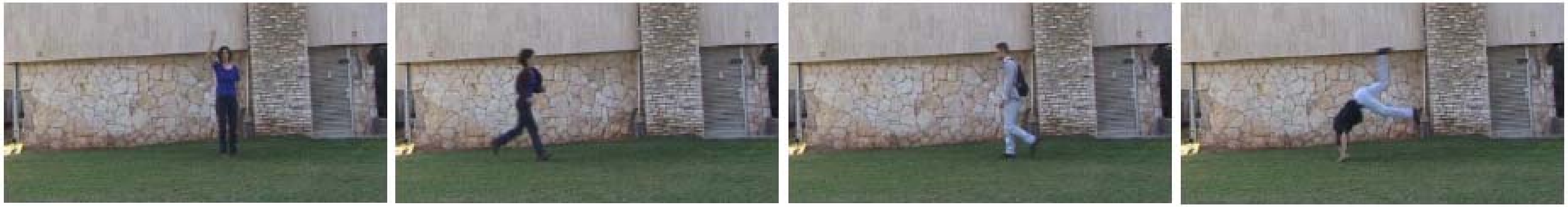


Output:

Sliding-window-style temporal action localization

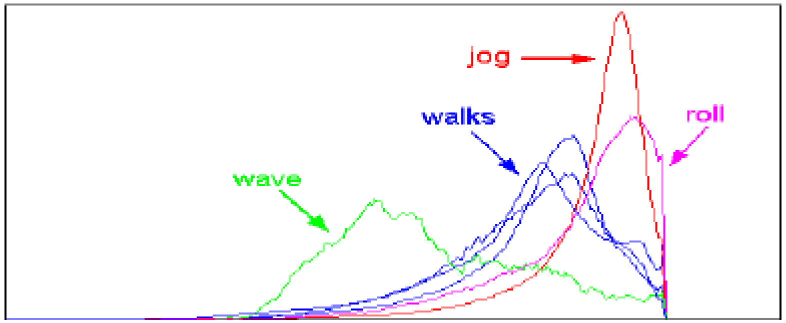
Action clustering

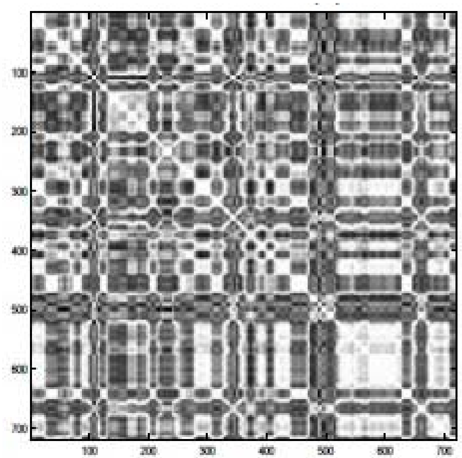
[Lihi Zelnik-Manor and Michal Irani CVPR 2001]

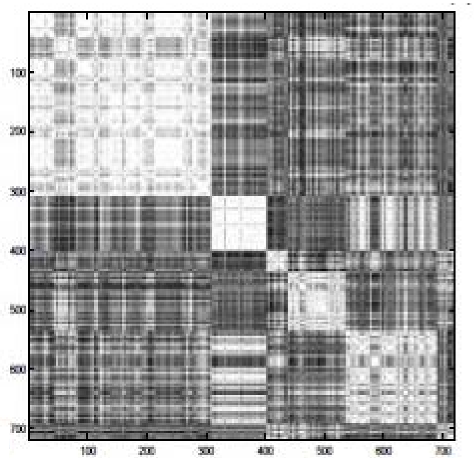


Spectral clustering

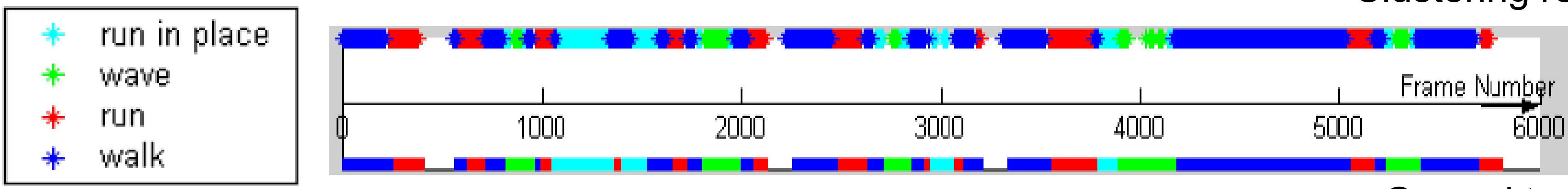
Descriptor space







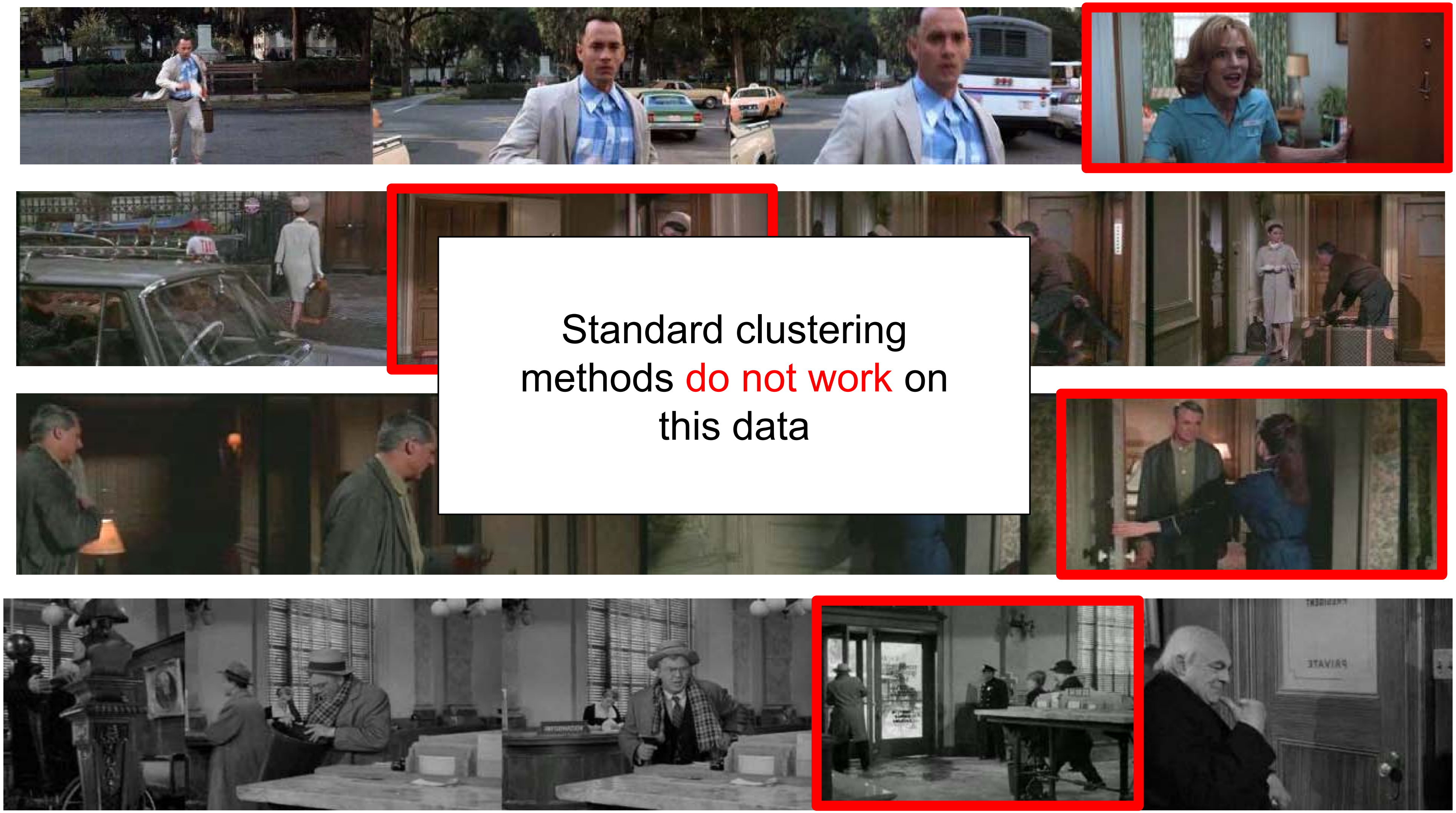
Clustering results



Ground truth

Action clustering

Complex data:

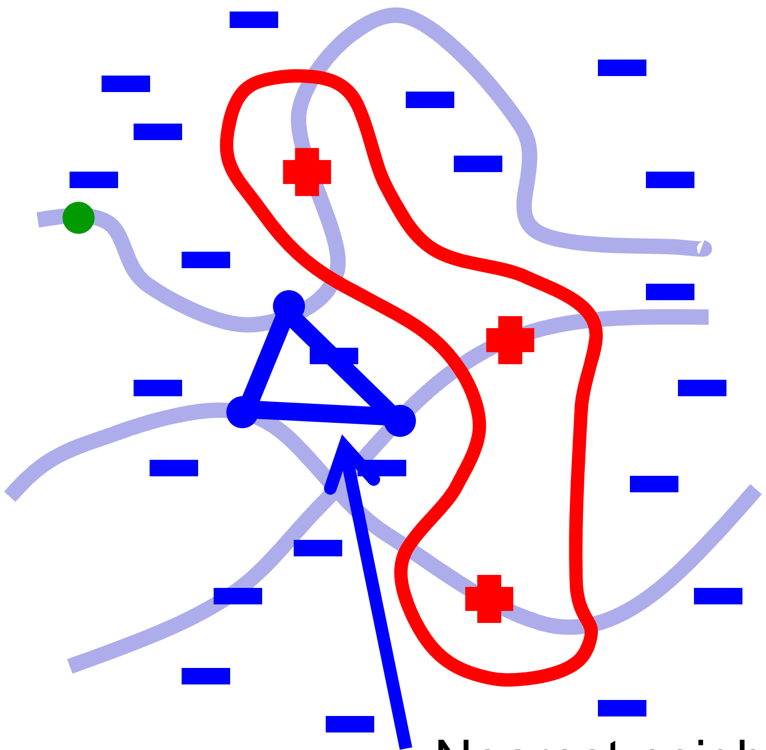


Standard clustering methods do not work on this data

Action clustering

Our view at the problem

Feature space



Nearest neighbor solution: wrong!

Video space

Negative samples!

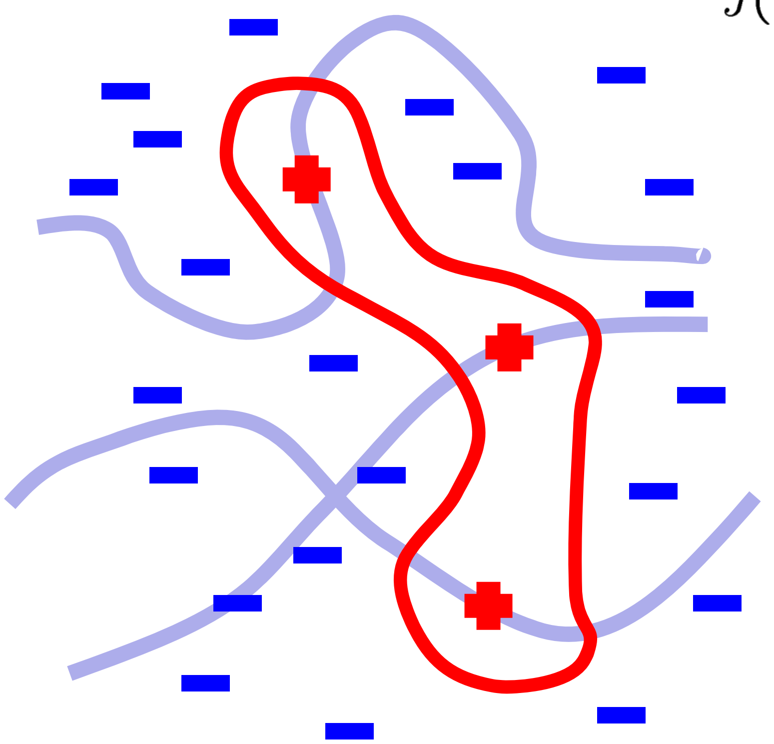
Random video samples: lots of them, very low chance to be positives

Action clustering

Formulation

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

Feature space



discriminative cost

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

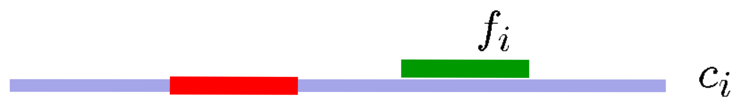
Loss on positive samples

$$+ C_- \boxed{\sum_{i=1}^{P} \max\{0, 1 + w^\top \Phi(x_i^-) + b\}} + \|w\|^2$$

Loss on negative samples

x_i^- negative samples

$c_i[f_i]$ parameterized positive samples

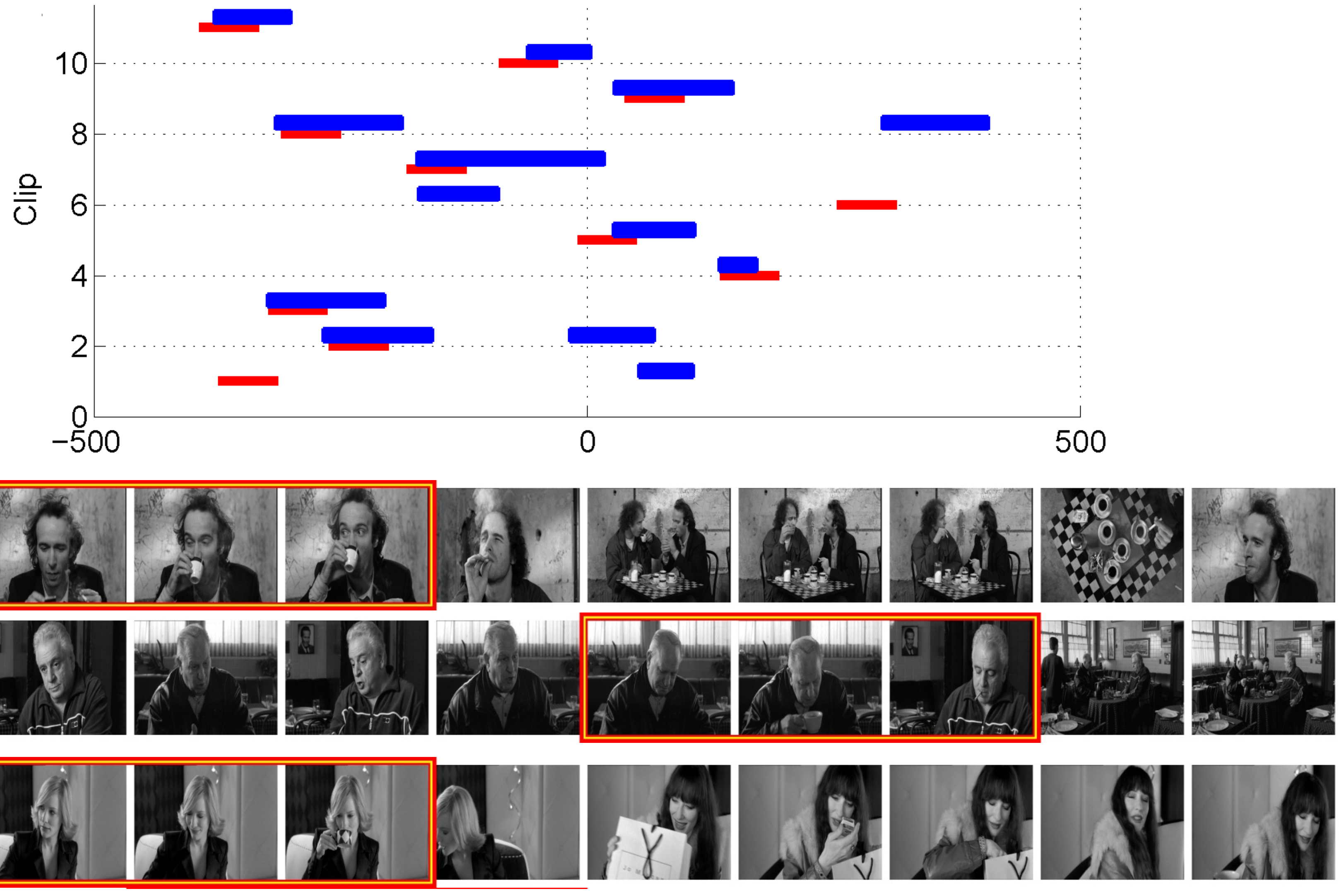


Optimization

SVM solution for w, b
Coordinate descent on f_i

Clustering results

Drinking actions in Coffee and Cigarettes

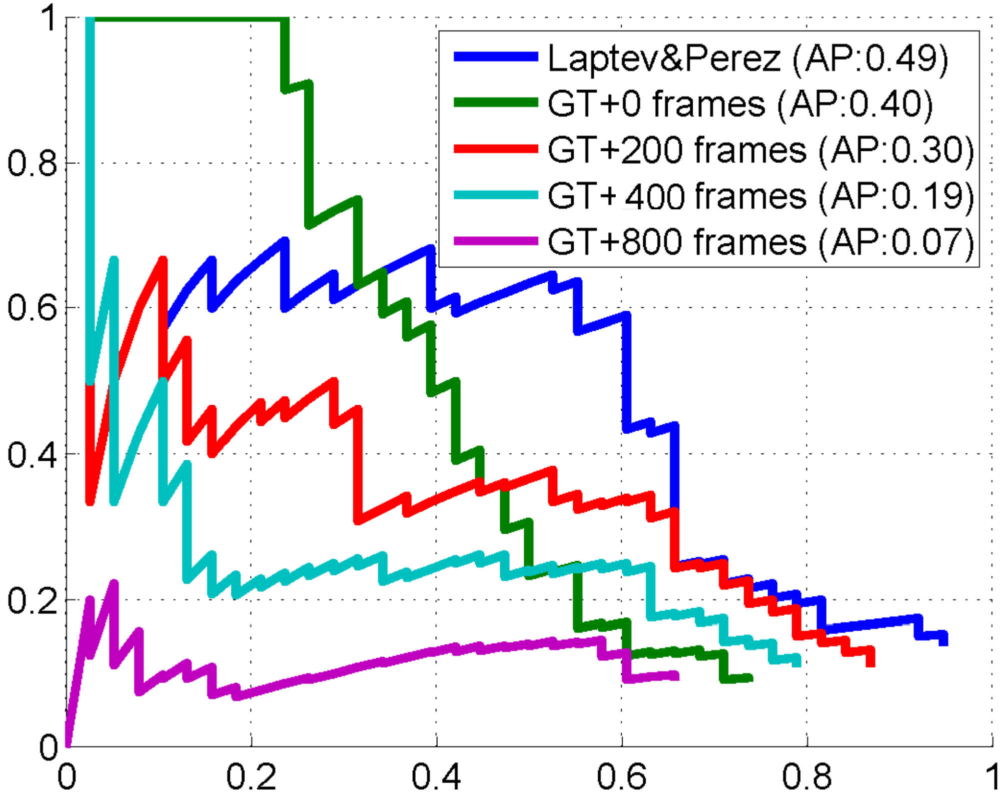


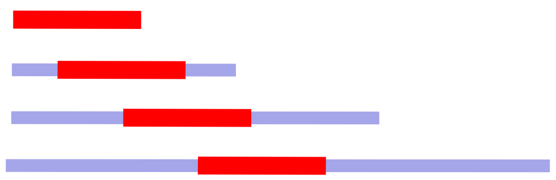
Detection results

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:

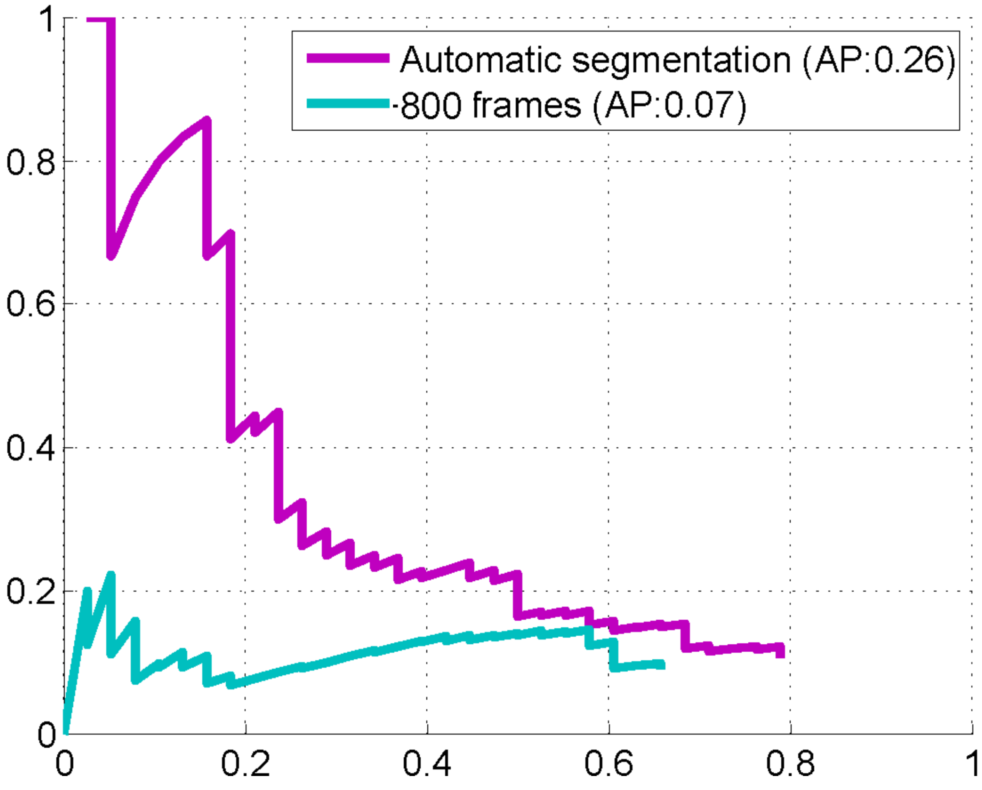
- 25min from “Coffee and Cigarettes” with GT 38 drinking actions

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*

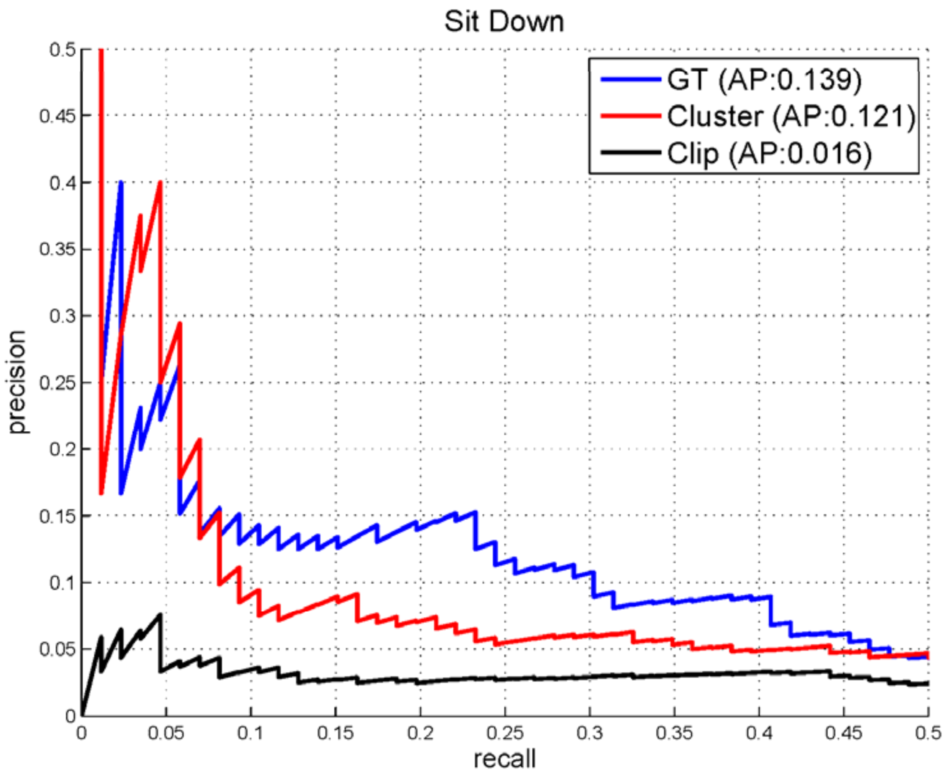


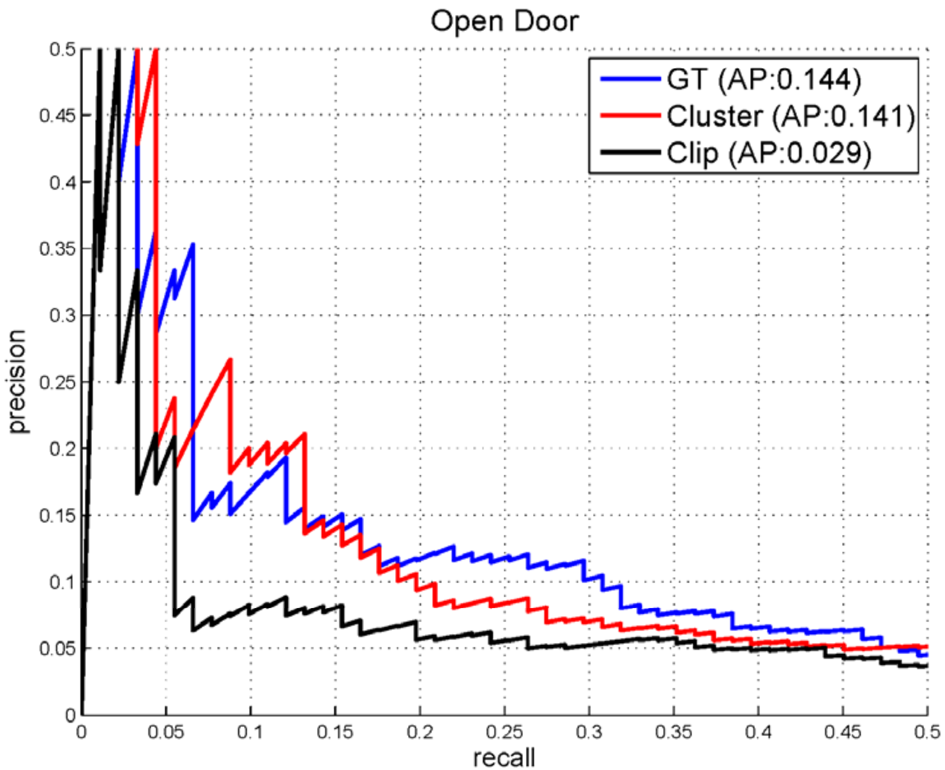
Test set:

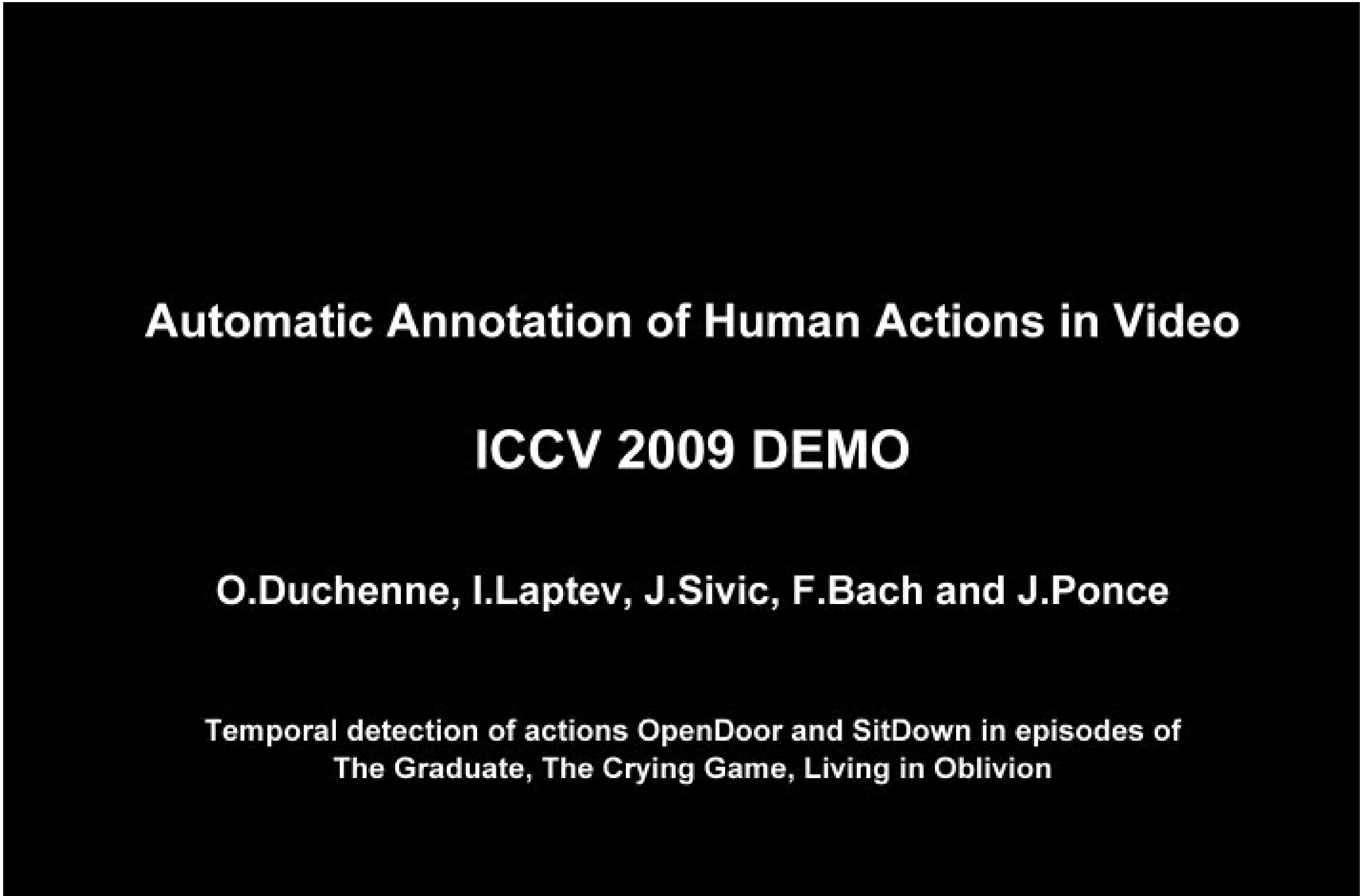
- 25min from "Coffee and Cigarettes" with GT 38 drinking actions

Detection results

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







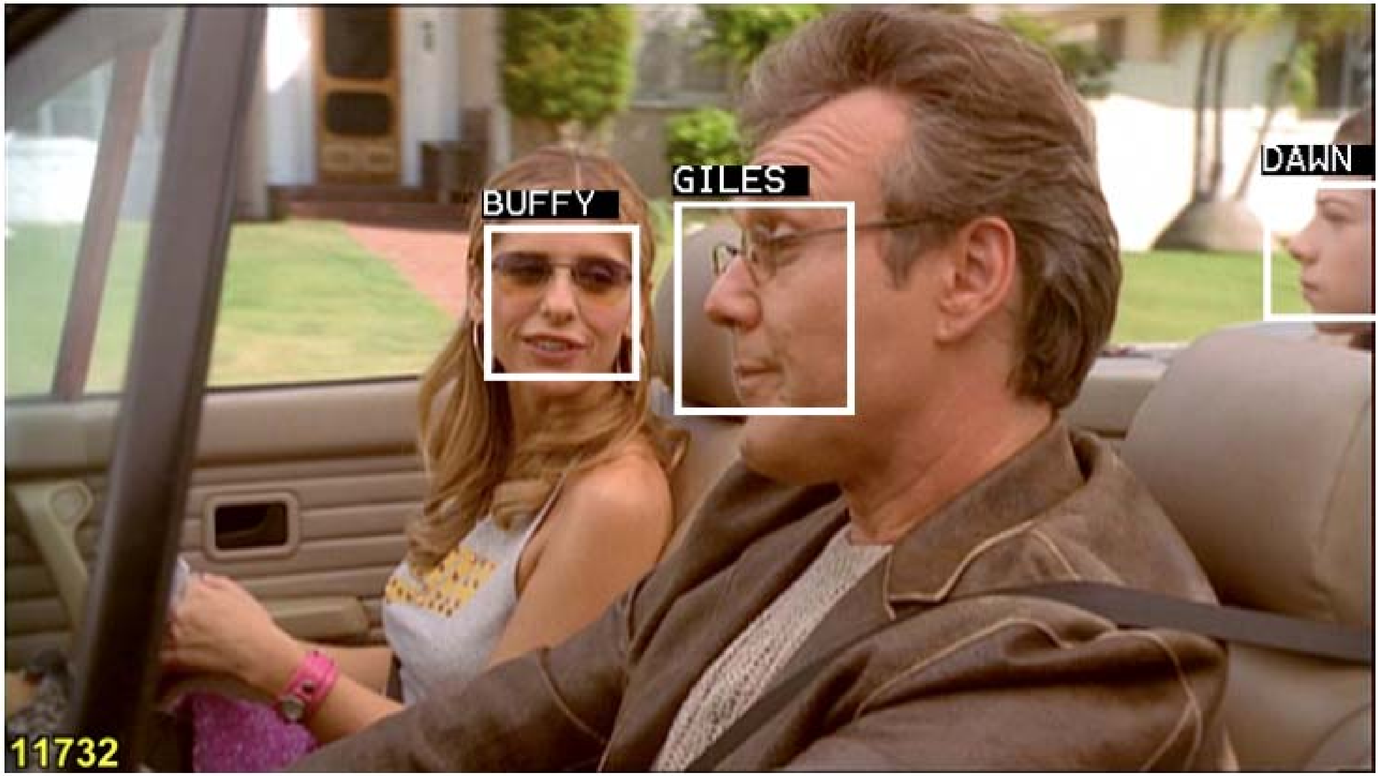
Temporal detection of “Sit Down” and “Open Door” actions in movies:
The Graduate, The Crying Game, Living in Oblivion

“Who are you?”: Learning person specific classifiers from video

[Sivic, Everingham, Zisserman]

The objective

- Automatically annotate characters in video with their identity
- Recognize characters whenever they appear in the video

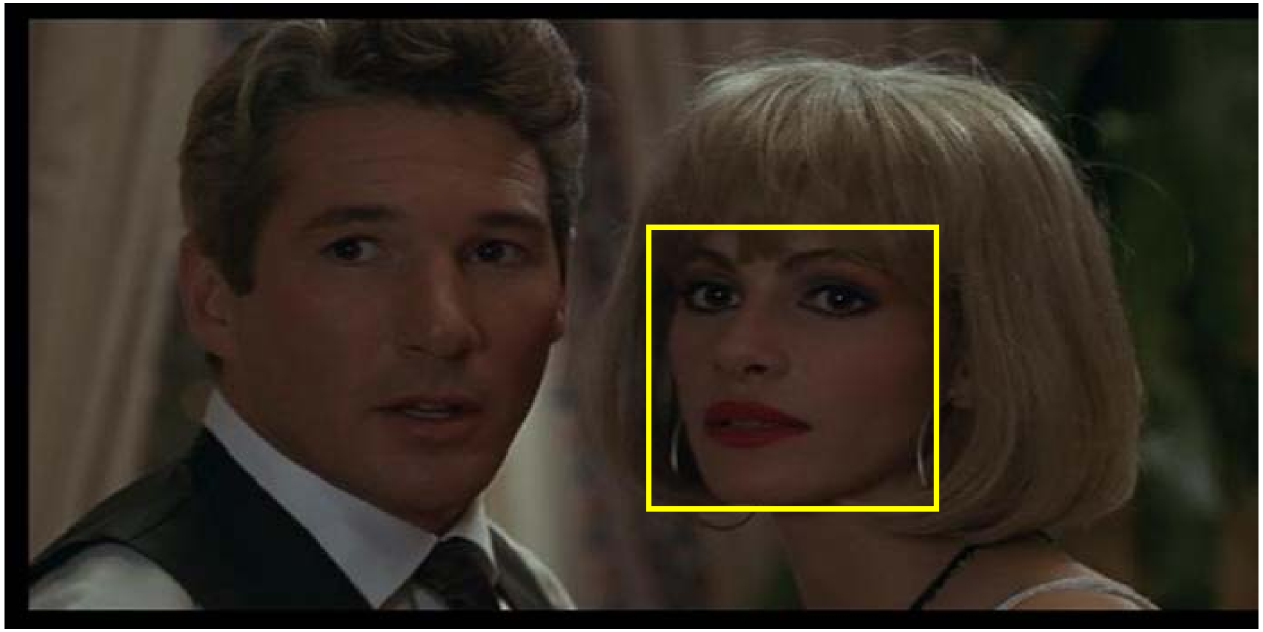


Visual search and automatic annotation of objects in video

[Sivic and Zisserman, ICCV'2003, CVPR'2004]

Visually defined search – on faces

Retrieve all shots in a video, e.g. a feature length film, containing a particular person



"Pretty Woman"

[Marshall, 1990]

Applications:

- intelligent fast forward on characters
- pull out all videos of "x" from 1000s of digital camera mpegs

[Sivic, Everingham and Zisserman, CIVR'05]

Matching faces in video

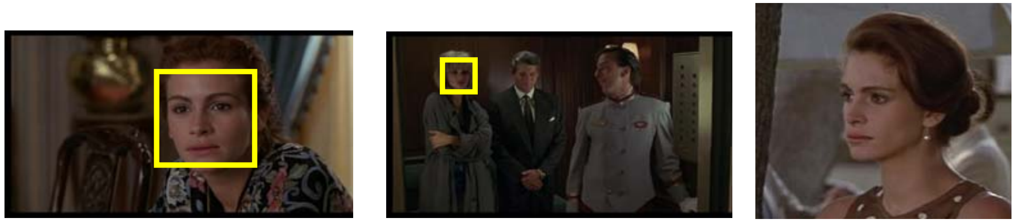
"Pretty Woman" (Marshall, 1990)

Are these faces of the same person?

Uncontrolled viewing conditions

Image variations due to:

- pose/scale



- lighting

- partial occlusion

- expression

c.f. Standard face databases

Matching Faces

Are these images of the same person ?

Can be difficult for individual examples …

Matching Faces

Are these images of the same person ?

But easier for sets of faces

The benefits of video

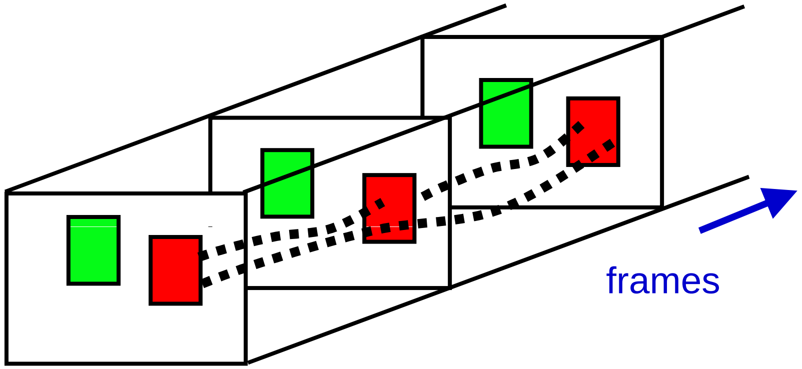
Automatically associate face examples

Obtaining sets of faces from video: Tracking by detection

Face detection - example

Operate at high precision (90%) point – few false positives

Need to associate detections with the same identity

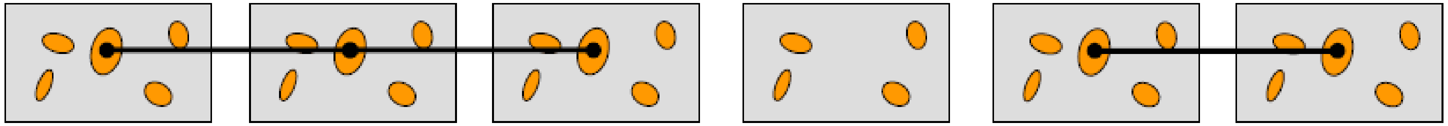


Example – tracked regions

Tracking covariant regions – two stages

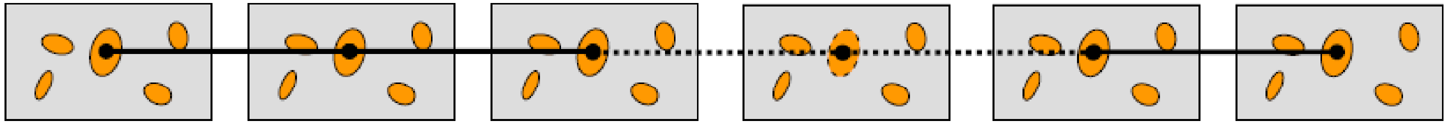
Goal: develop very long and good quality tracks

- Stage I – match regions detected in neighbouring frames



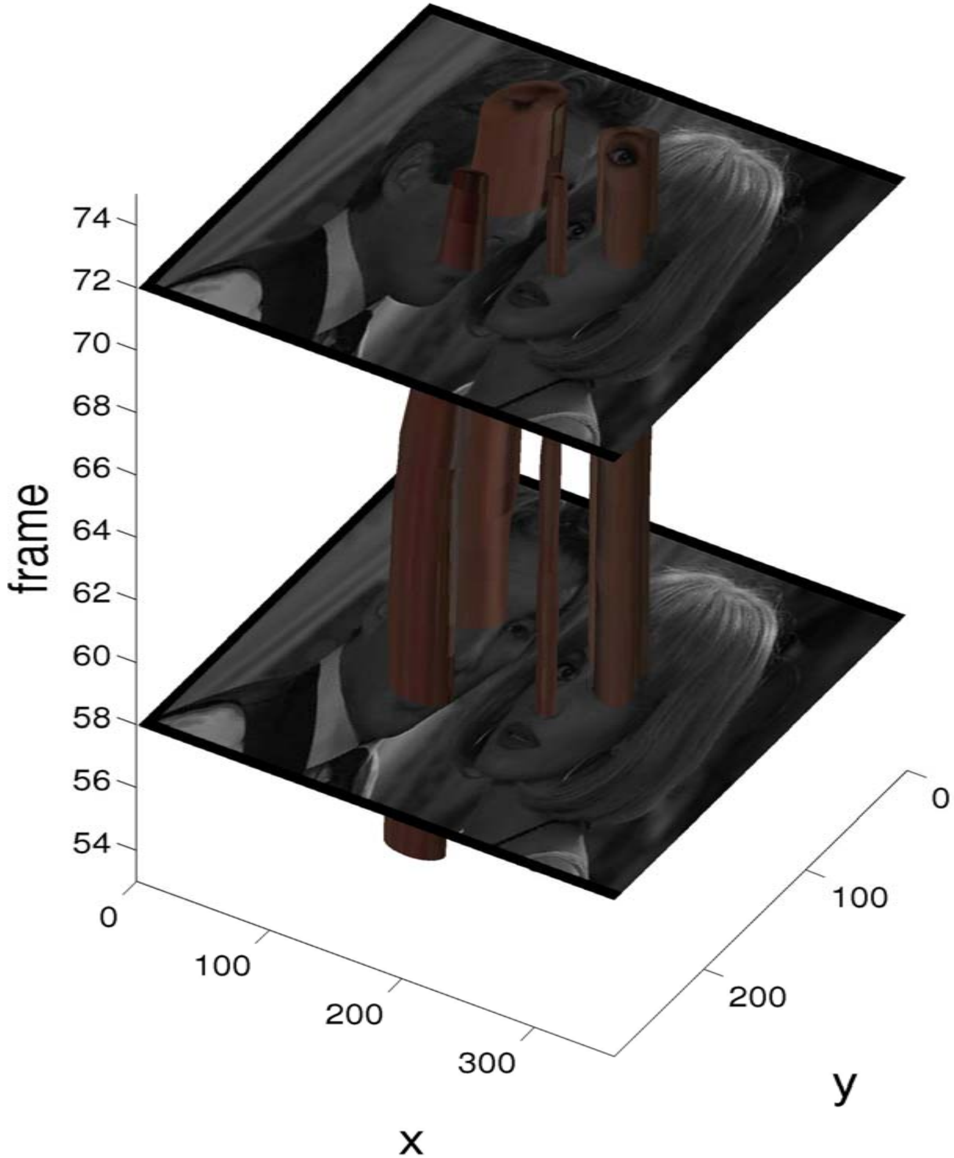
Problems: e.g. missing detections

- Stage II – repair tracks by region propagation



[Ferrari et al. 2004, Sivic et al. 2004]

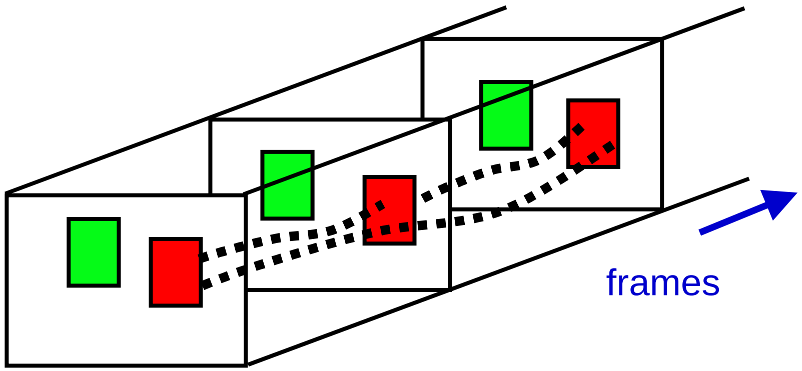
Region tubes



Connecting face detections temporally

Goal: associate face detections of each character within a shot

Approach: Agglomeratively merge face detections based on connecting ‘tubes’



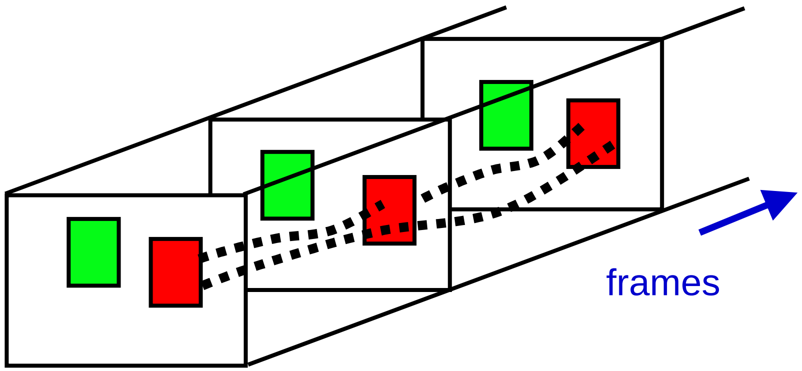
Measure connectivity score of a pair of faces by number of tracks intersecting both detections

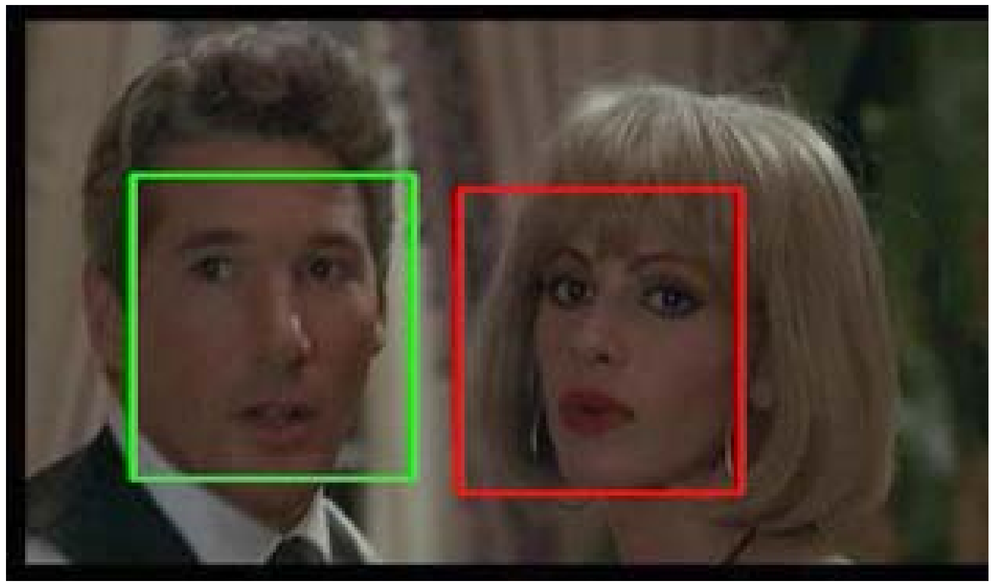
require a minimum number of region tubes to overlap face detections

Connecting face detections temporally

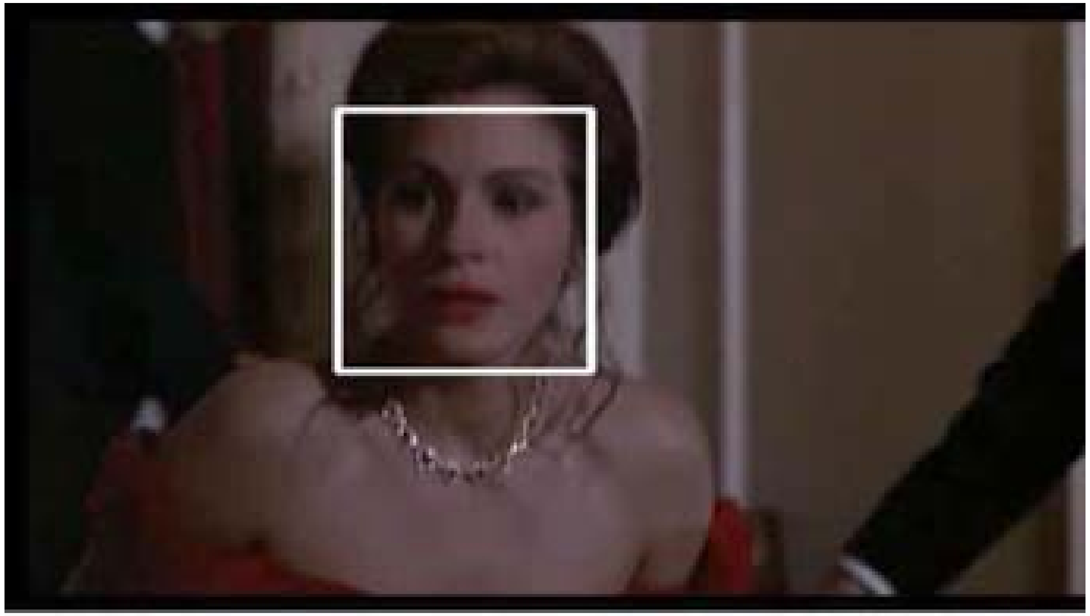
Goal: associate face detections of each character within a shot

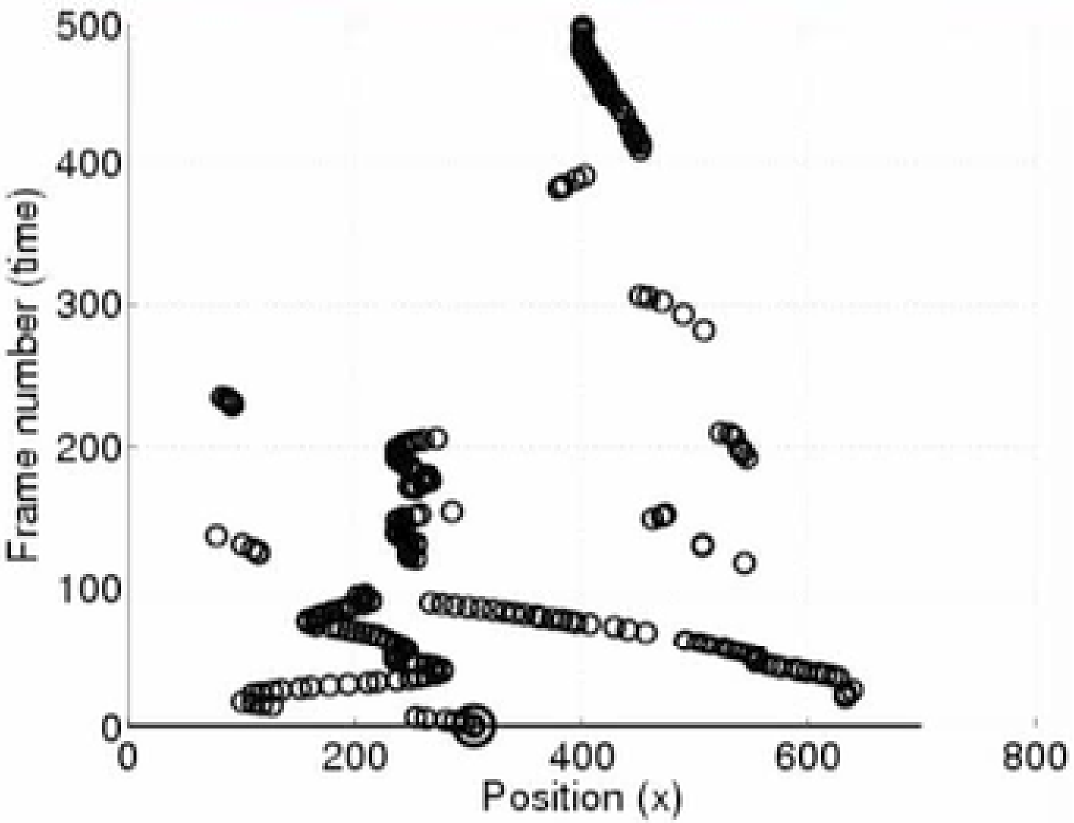
Approach: Agglomeratively merge face detections based on connecting ‘tubes’



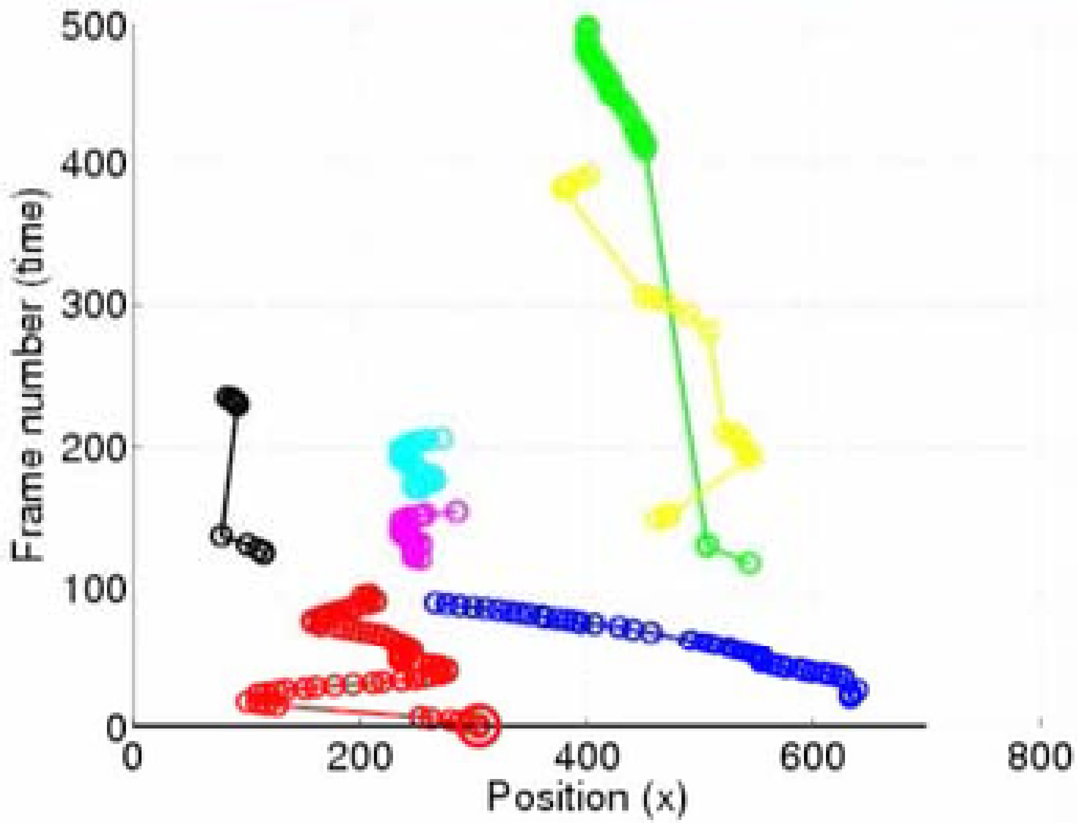


Alternatives: Avidan CVPR 01, Williams *et al* ICCV 03





raw face detections

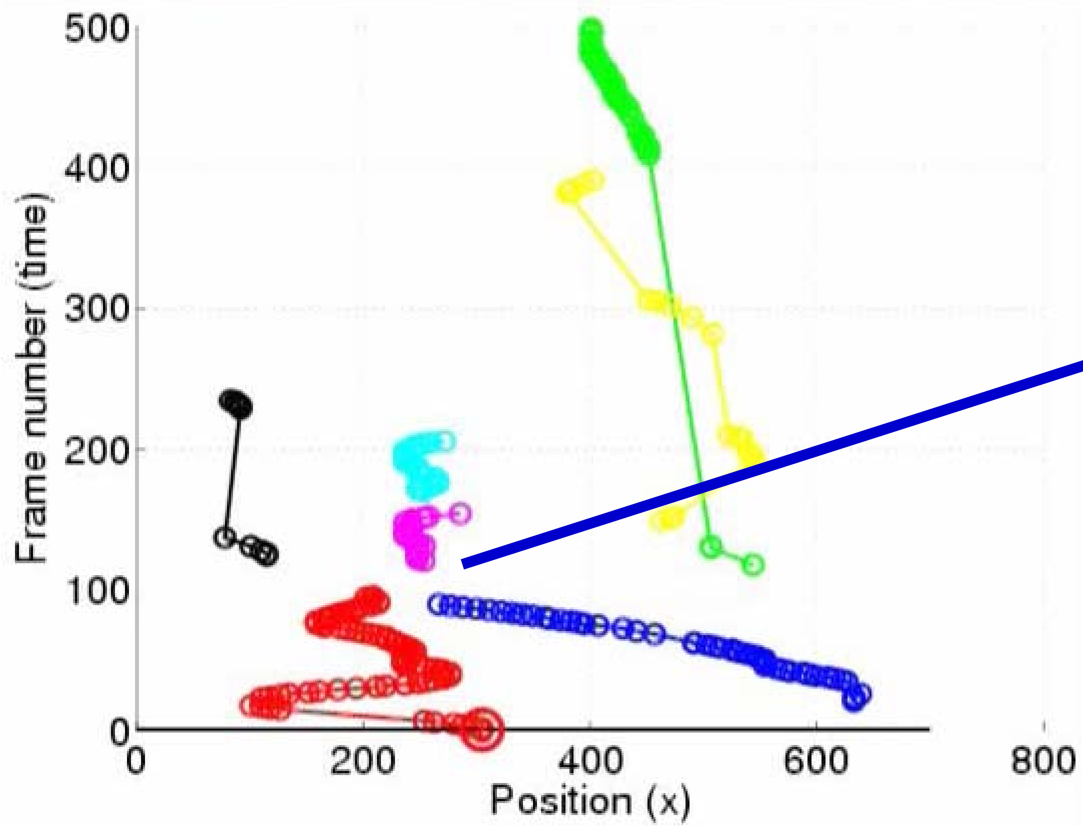


500
400
300
200
100
0
0
200
400
600
800
Frame number (time)
Position (x)

Face tracks

Face tracks

Tracking by recognition

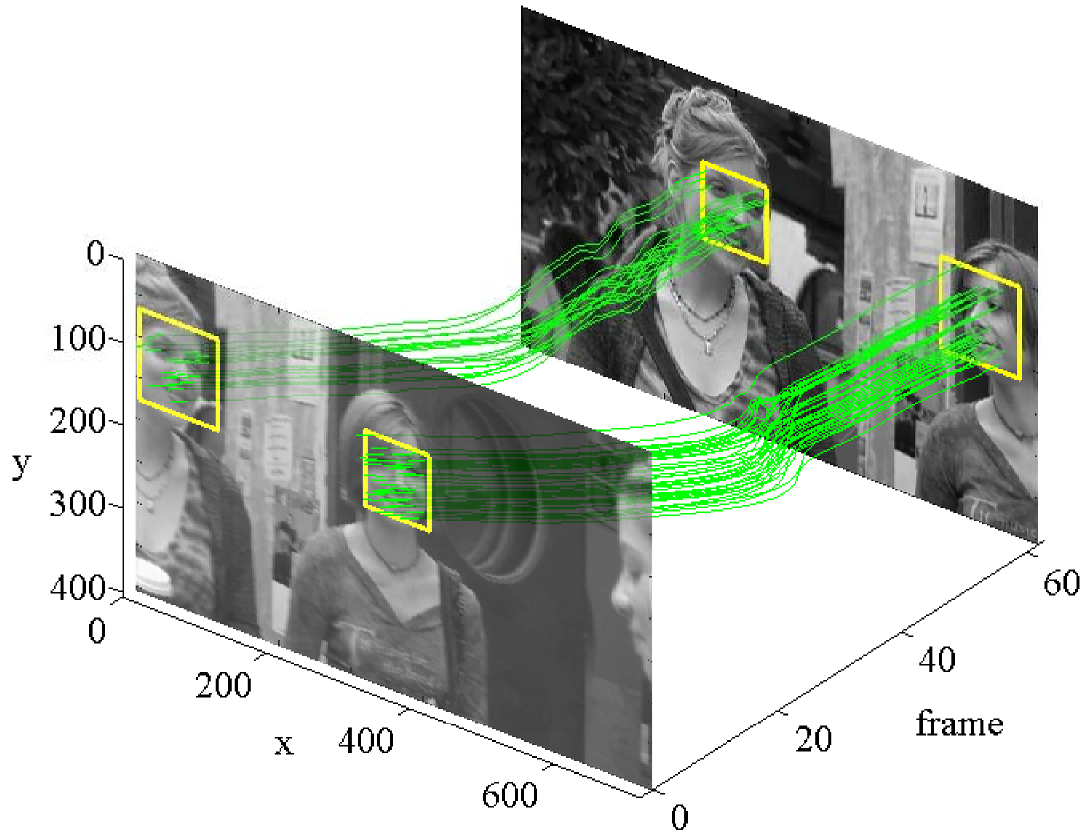


Tracking by recognition

Connected face tracks

Connecting face detections temporally

+ Does not require contiguous detections
+ Independent evidence – no drift
- Tracking affine covariant regions is expensive



Tracking faces in spatio-temporal video volume

- Use “light-weight” KLT tracker (3fps)
- Fix occasional broken tracks later:
  tracking by recognition

Face representation and matching

Matching faces

Easier if faces aligned to remove pose variation

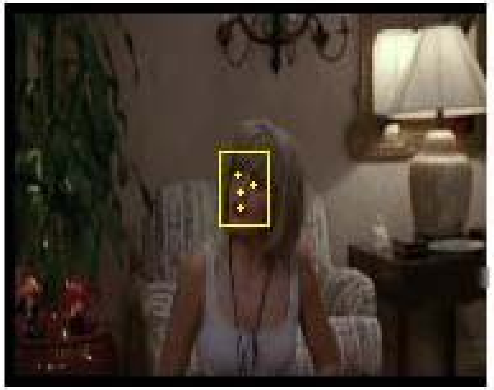
face detector

eyes/nose/mouth

Rectified face

Face normalization - example

- affine transform face using detected features

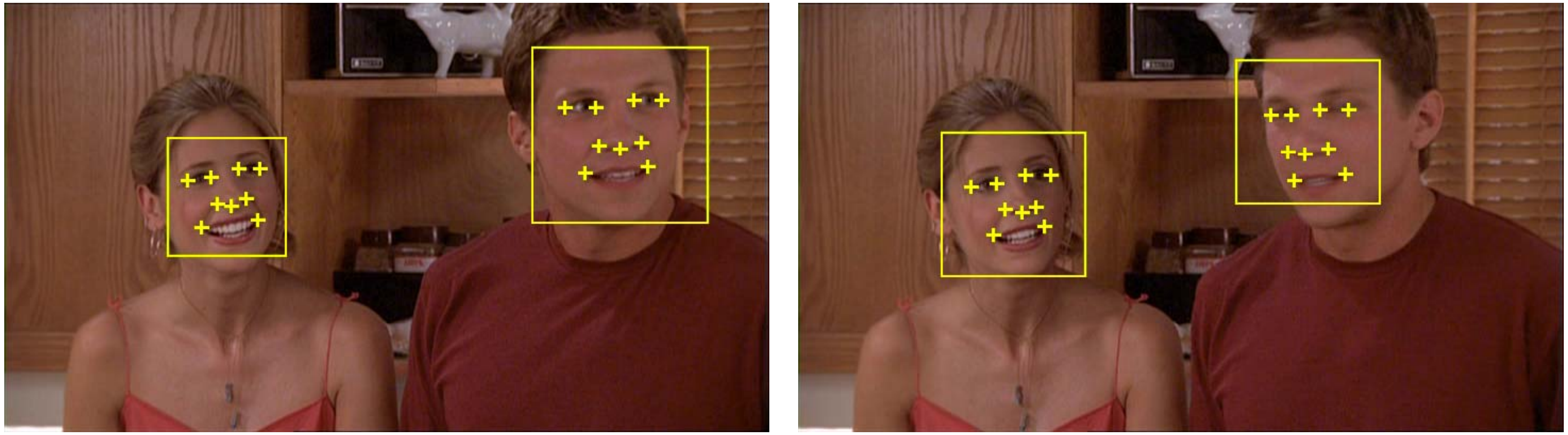


original detection

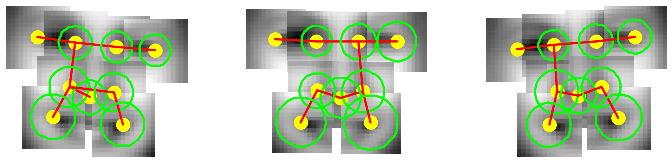
rectified

Facial feature localization using a pictorial structure model

- Stabilize representation by localizing features
  - Pose of face varies and face detector is noisy



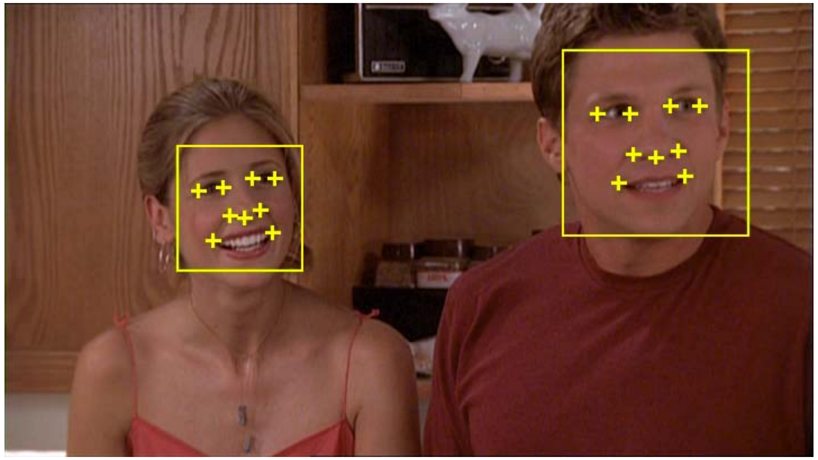
- Extended “pictorial structure” model
  - Joint model of feature appearance and position

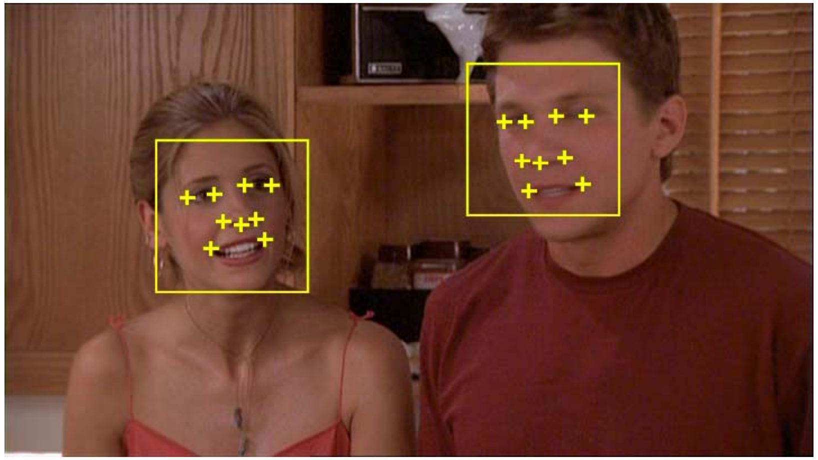


[Felzenszwalb and Huttenlocher’2004]

Facial feature localization using a pictorial structure model

- Stabilize representation by localizing features
  - Pose of face varies and face detector is noisy

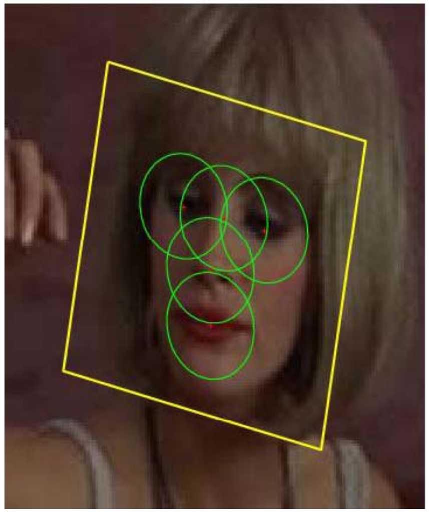




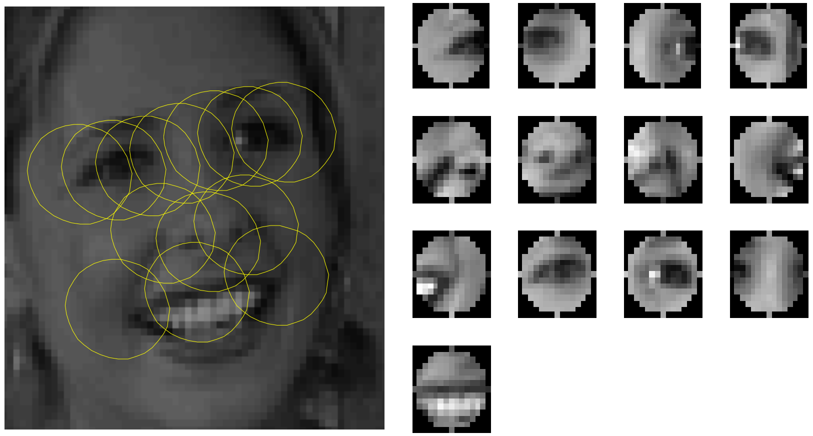
- Matlab code available online:

http://www.robots.ox.ac.uk/~vgg/research/nface/

Face representation – local descriptors: from sparse to dense

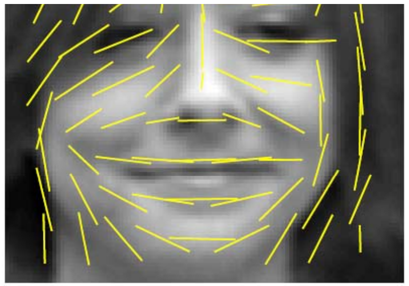


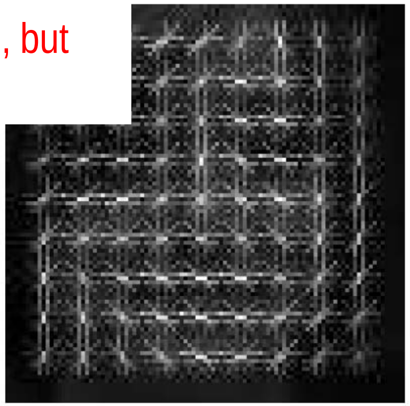
[Sivic, Everingham, Zisserman, 2005]



[Everingham, Sivic, Zisserman, 2006]

Dense representation is beneficial, but faces need to be well aligned!

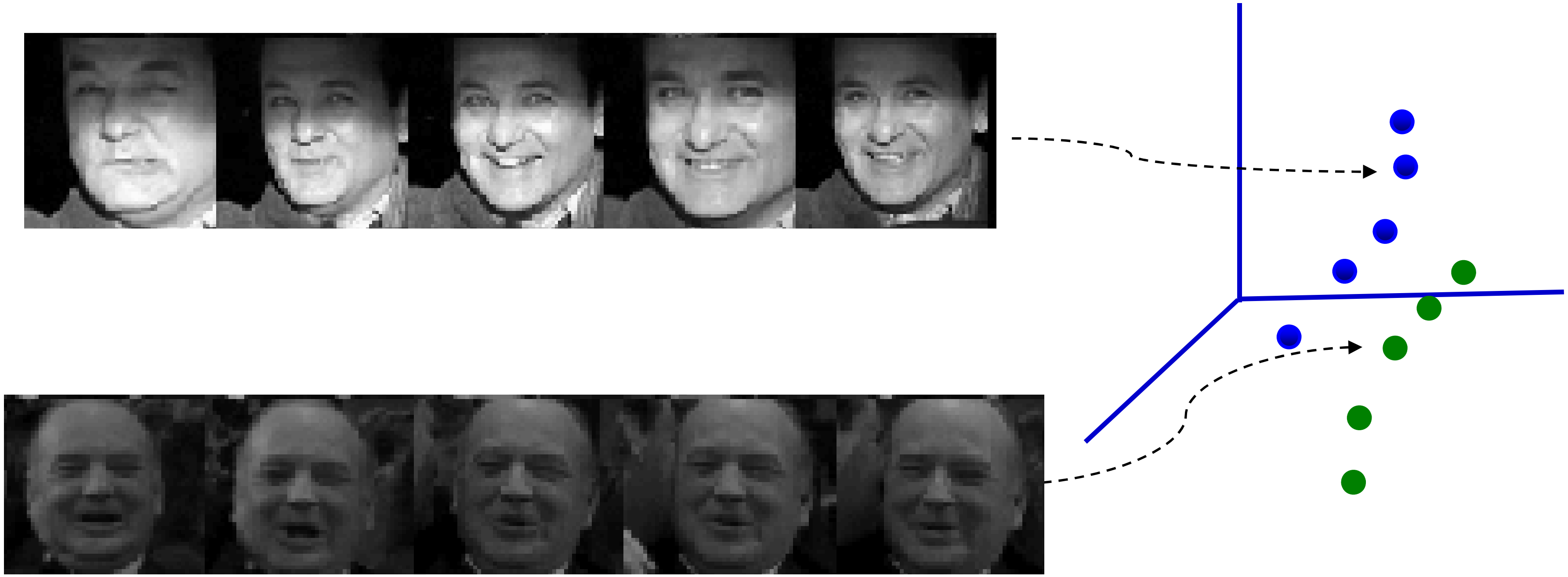




[Sivic, Everingham, Zisserman, 2009]

[Heisele et al., 2003]

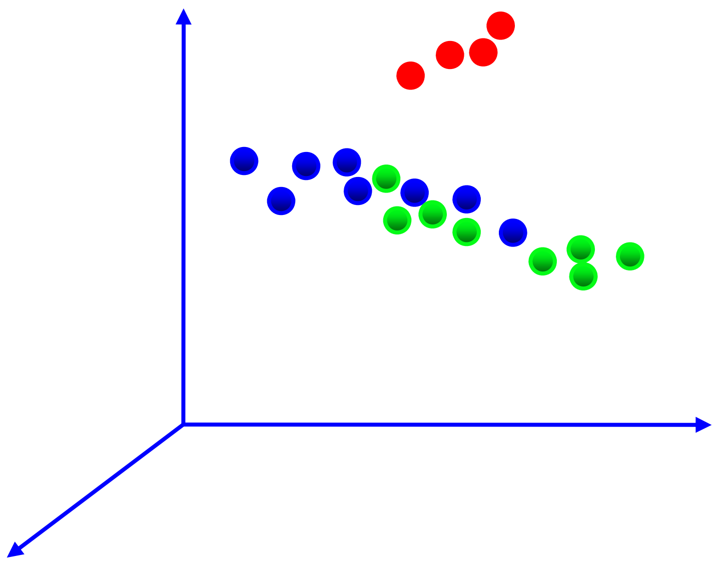
Matching face sets



Matching face sets

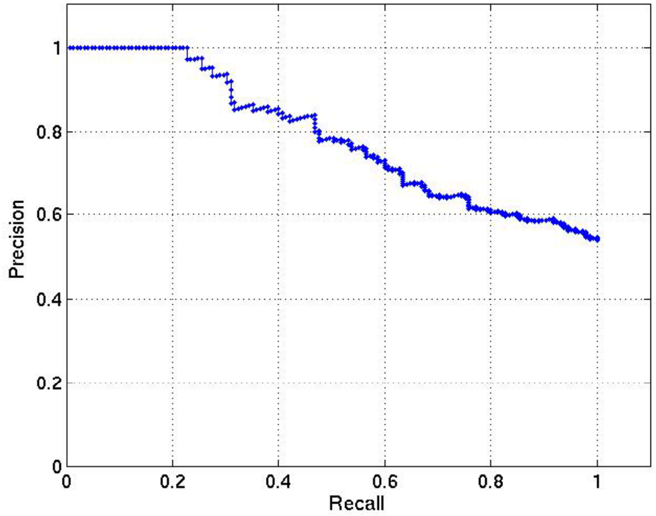
min-min distance: $$d(A,B) = \min_{\mathbf{a}\in A,\mathbf{b}\in B} d(\mathbf{a},\mathbf{b})$$

A, *B* ... sets of face descriptors



Face retrieval – example

Query sequence



Retrieved sequences (shown by first detection)

Example sequence

Face retrieval in movies - demo

http://www.robots.ox.ac.uk/~vgg/research/fgoogle/

Training person specific classifiers: from retrieval to classification

Aims

- Automatically label appearances of characters with names

- Requires additional information
- No supervision from the user, use only readily-available annotation

[Everingham, Sivic, Zisserman, 2006]

Textual Annotation: Subtitles/Closed-captions

- DVD contains timed subtitles as bitmaps
  - Automatically convert to text using simple OCR

00:18:55,453 --> 00:18:56,086
Get out!

00:18:56,093 --> 00:19:00,044
- But, babe, this is where I belong.
- Out! I mean it.

00:19:00,133 --> 00:19:03,808
I've been doing a lot of reading,
and I'm in control of my own power now,...

- What is said, and when, but not who says it

[Everingham, Sivic, Zisserman, 2006]

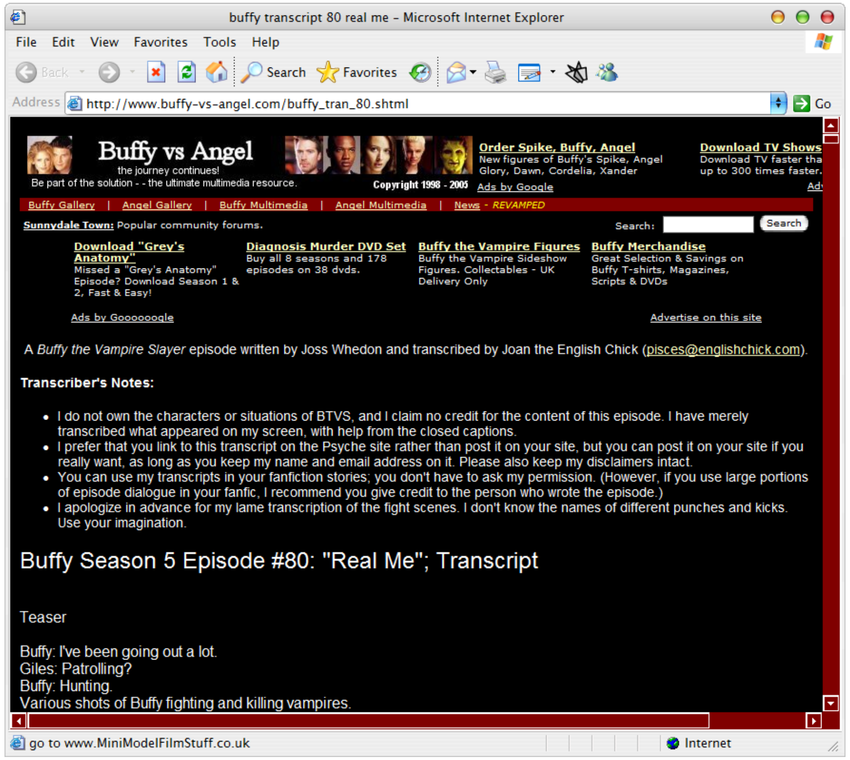
Textual Annotation: Script

- Many fan websites publish transcripts

HARMONY
Get out.

SPIKE
But, baby... This is where I belong.

HARMONY
Out! I mean it. I've done a lot of reading, and, and I'm in control of my own power now.



- What is said, and who says it, but not when

[Everingham, Sivic, Zisserman, 2006]

Subtitle/Script Alignment

- Alignment of what allows subtitles to be tagged with identity giving who and when
  - “Dynamic Time Warping” algorithm

00:18:55,453 --> 00:18:56,086
Get out!

00:18:56,093 --> 00:19:00,044
- But, babe, this is where I belong.
- Out! I mean it.

00:19:00,133 --> 00:19:03,808
I've been doing a lot of reading,
and I'm in control of my own power now,...

00:19:03,893 --> 00:19:05,884
..so we're through.

HARMONY
Get out.

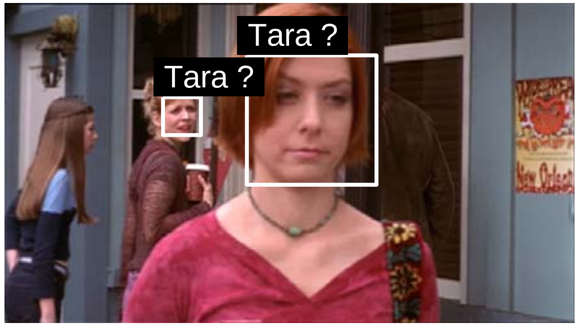
SPIKE
But, baby... This is where I belong.

HARMONY
Out! I mean it. I've done a lot of reading, and, and I'm in control of my own power now. So we're through.

[Everingham, Sivic, Zisserman, 2006]

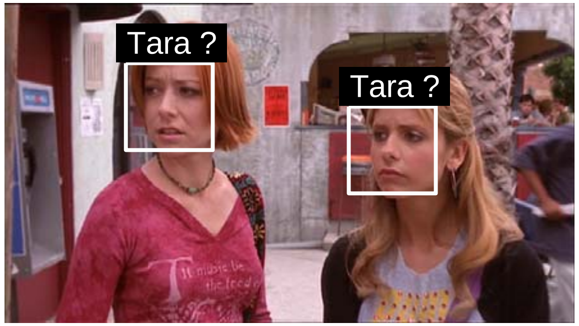
Ambiguity

- Knowledge of speaker is a <u>weak</u> cue that the character is visible



Multiple characters

Speaker not detected



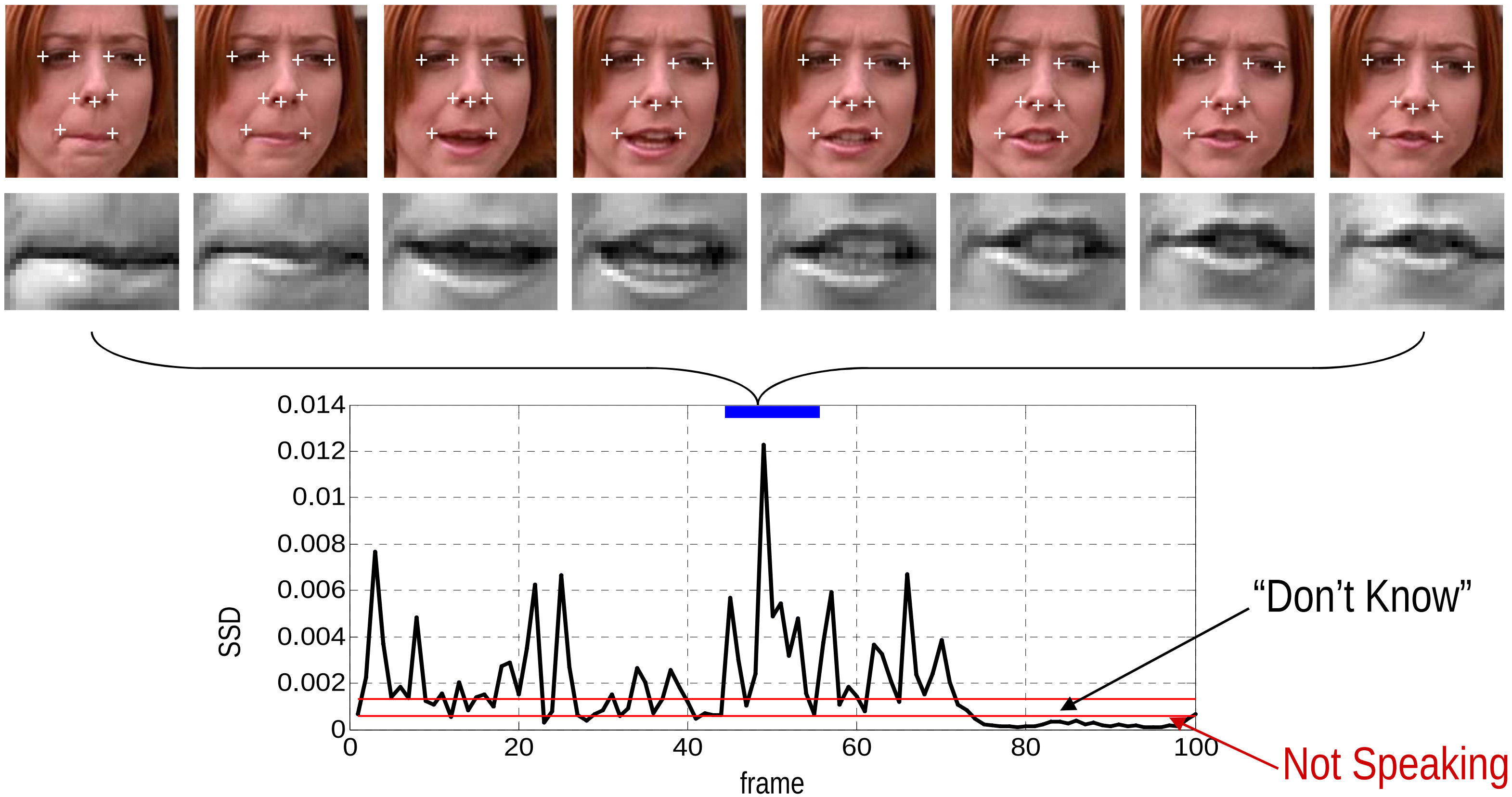
Speaker not visible

- Ambiguities will be resolved using vision-based speaker detection

[Everingham, Sivic, Zisserman, 2006]

Speaker Detection

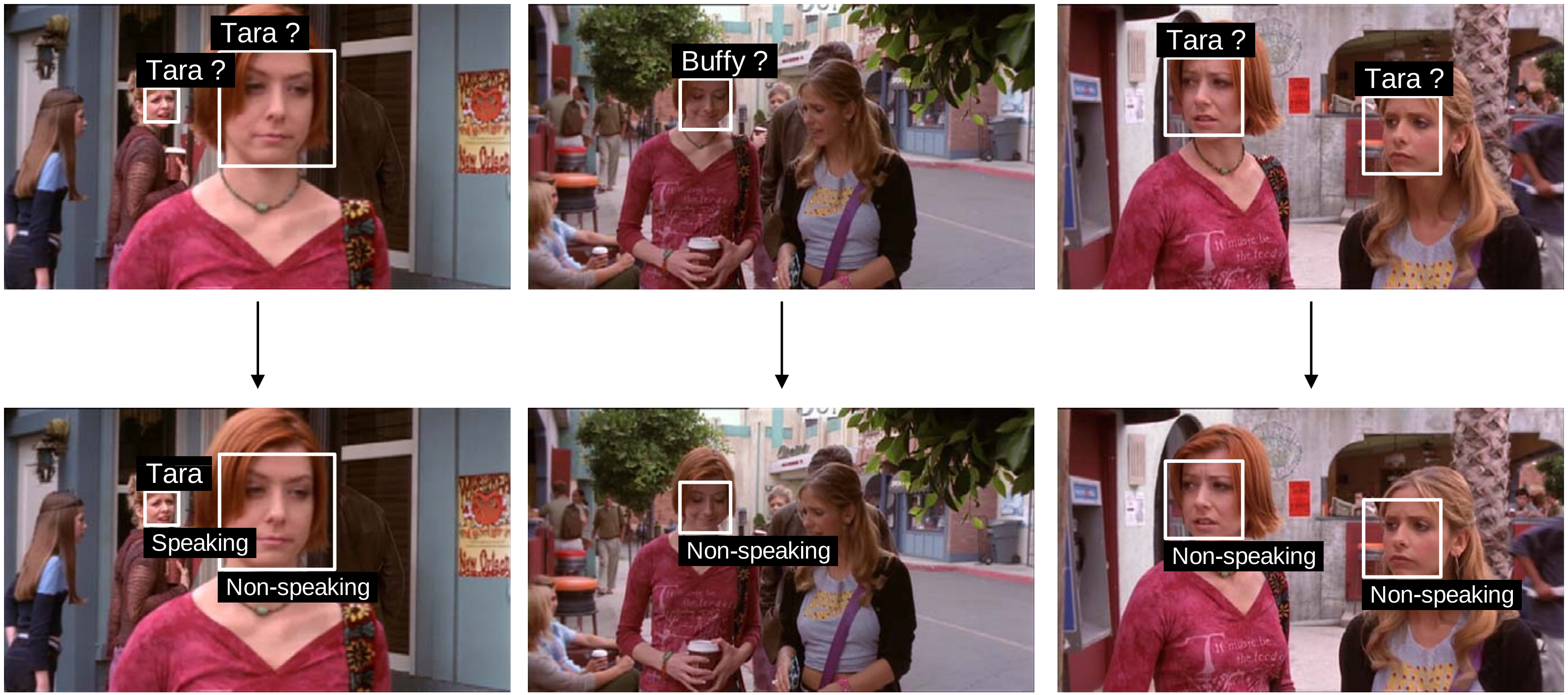
- Measure the amount of motion of the mouth
  - Search across frames around detected mouth points



[Everingham, Sivic, Zisserman, 2006]

Resolved Ambiguity

- When the speaker (if any) is identified, the ambiguity in the textual annotation is resolved



[Everingham, Sivic, Zisserman, 2006]

Exemplar Extraction

- Face tracks detected as speaking and with a single proposed name give exemplars

Buffy

2,300 faces

Willow

1,222 faces

Xander

425 faces

[Everingham, Sivic, Zisserman, 2006]

Annotation as classification

- Use extracted exemplars to train a classifier for each character (Nearest Neighbour or SVM)
- Need to deal with noise in the training data (~10% errors)
- Assign names to unlabelled faces by classification based on extracted exemplars

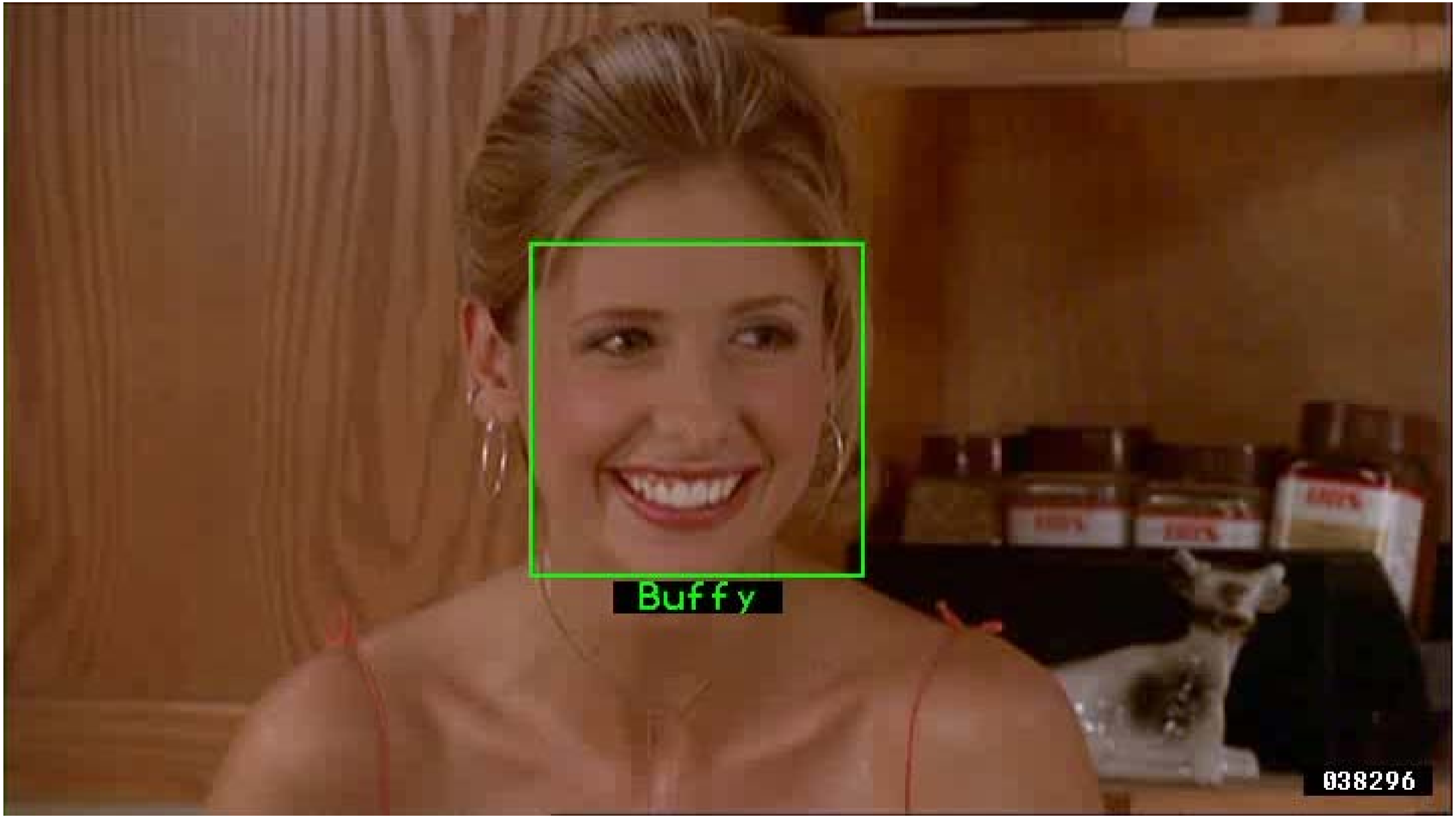
[Everingham, Sivic, Zisserman, 2006]

Example Results

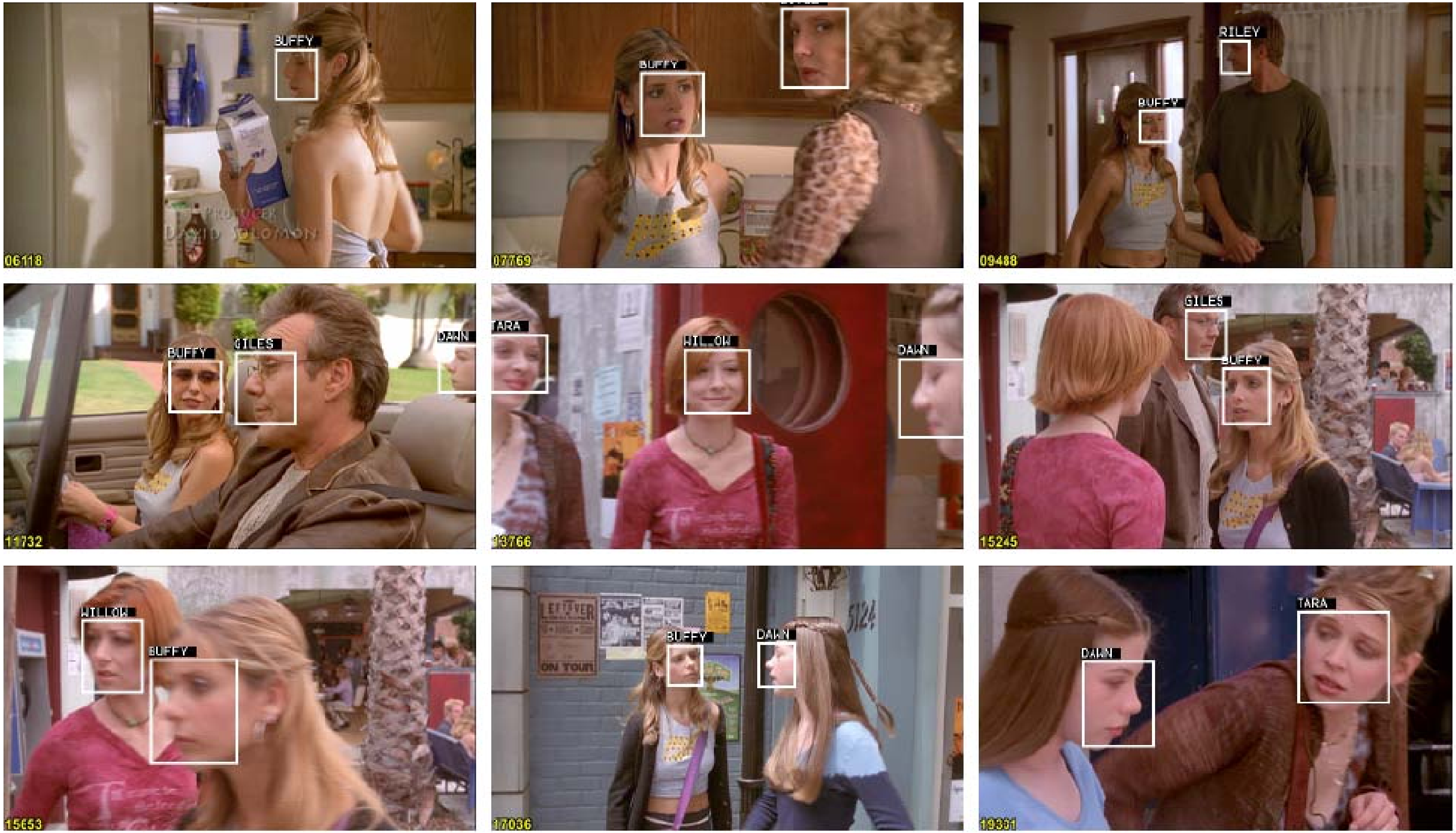
- No user involvement, just hit "go"…

[Everingham, Sivic, Zisserman, 2006]

Example Results



Examples of correct classification




BUFFY
WILLOW
21915
BUFFY
JOYCE
25475
TARA
29596
RILEY
BUFFY
30846
XANDER
ANYA
38856
BUFFY
RILEY
39137
SPIKE
41003
SPIKE
44649
BUFFY
XANDER
46640
RILEY
BUFFY
12808
ANYA
XANDER
28691
BUFFY
DAWN
JOYCE
09301

Example Video

Conclusions – benefits of video

- Additional signal – visual speaker detection
- Temporal association provide additional generalization

  > Detect characters whenever they are visible in video.

  > Match face tracks rather than individual faces

  > Use video as a source of additional training data.