Action recognition in videos

Cordelia Schmid

informatics mathematics

Action recognition - goal

- Short actions, i.e. answer phone, shake hands

answer phone

hand shake

Action recognition - goal

- Activities/events, i.e. making a sandwich, doing homework

Making sandwich

Doing homework

TrecVid Multi-media event detection dataset

Action recognition - goal

- Activities/events, i.e. birthday party, parade

Birthday party

Parade

TrecVid Multi-media event detection dataset

Action recognition - tasks

- Action classification: assigning an action label to a video clip

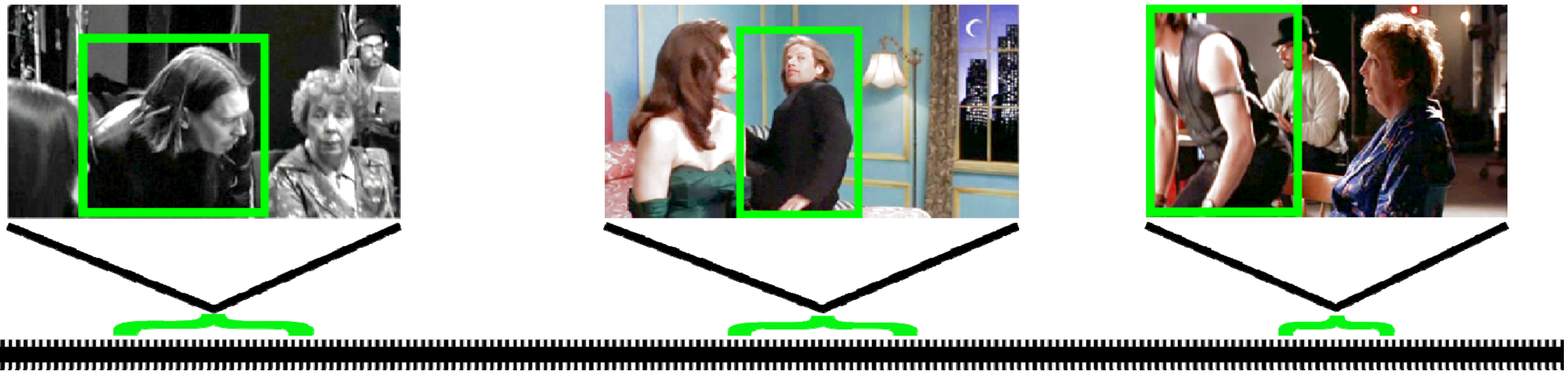
Making sandwich: present
Feeding animal: not present
...

Action recognition - tasks

- Action classification: assigning an action label to a video clip

Making sandwich: present
Feeding animal: not present
...

- Action localization: search locations of an action in a video



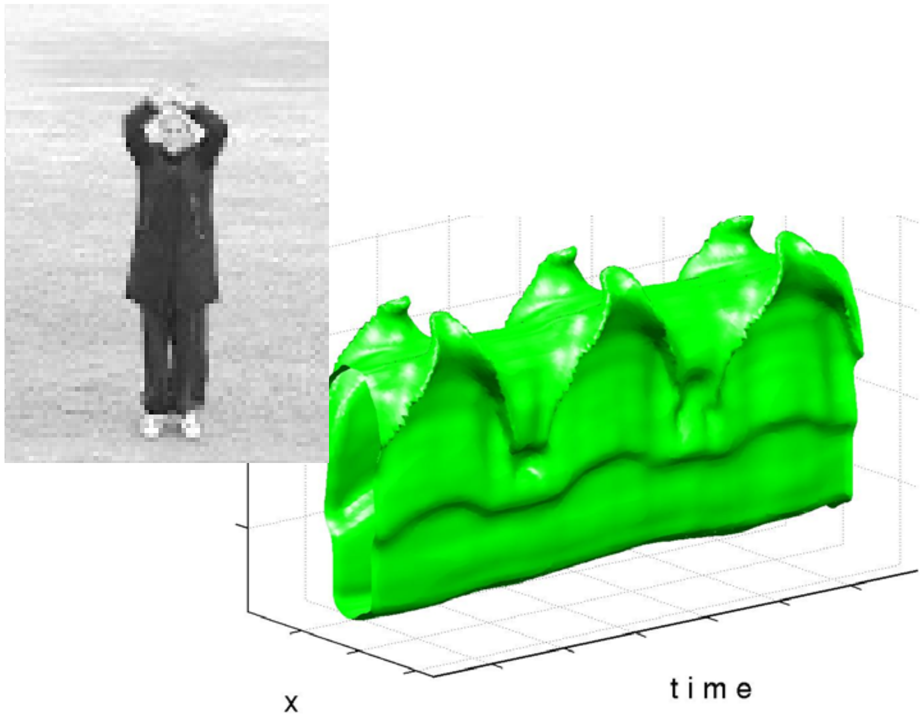
Space-time descriptors

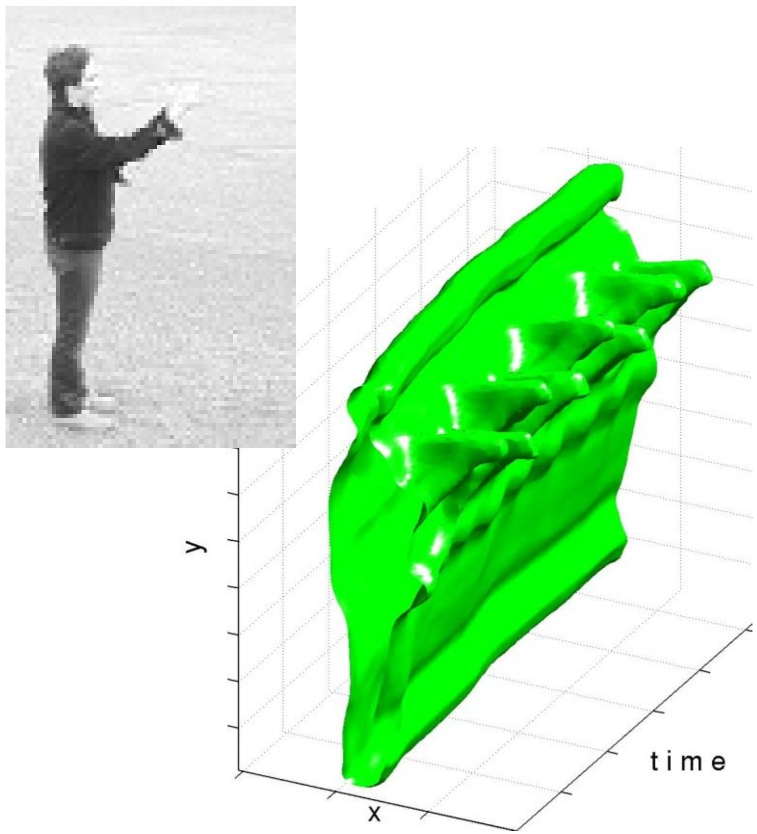
Consider local spatio-temporal neighborhoods

hand waving

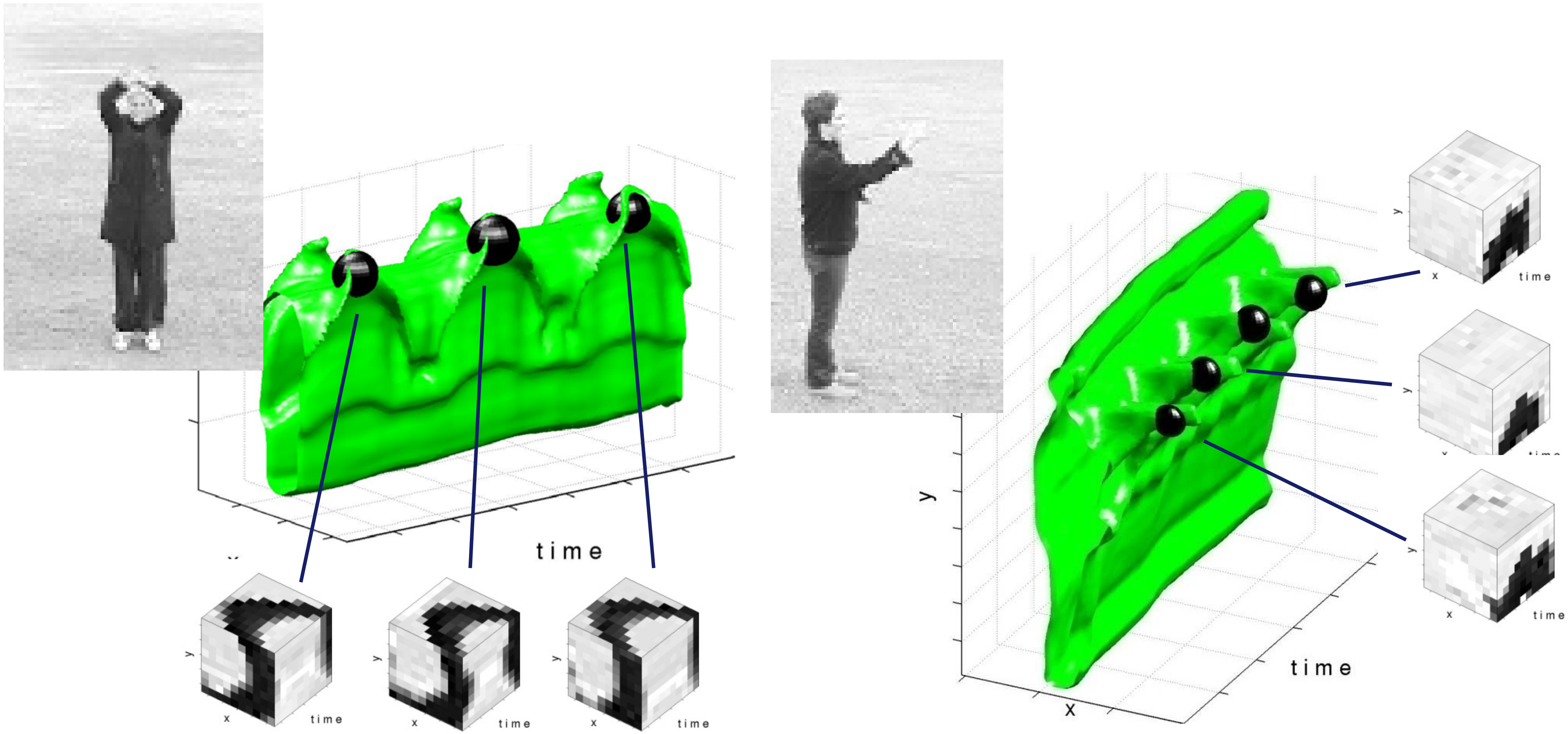
boxing

Actions == Space-time objects?





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

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

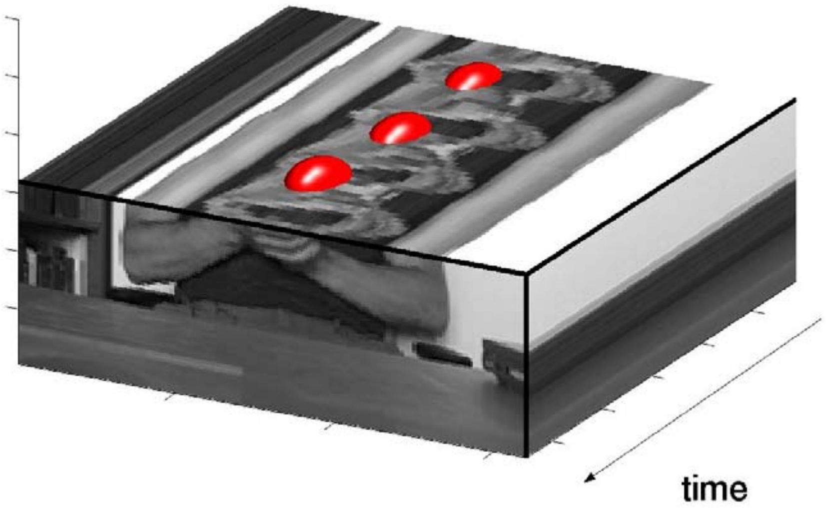
Large eigenvalues of μ can be detected by the local maxima of H over (x,y,t):

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

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

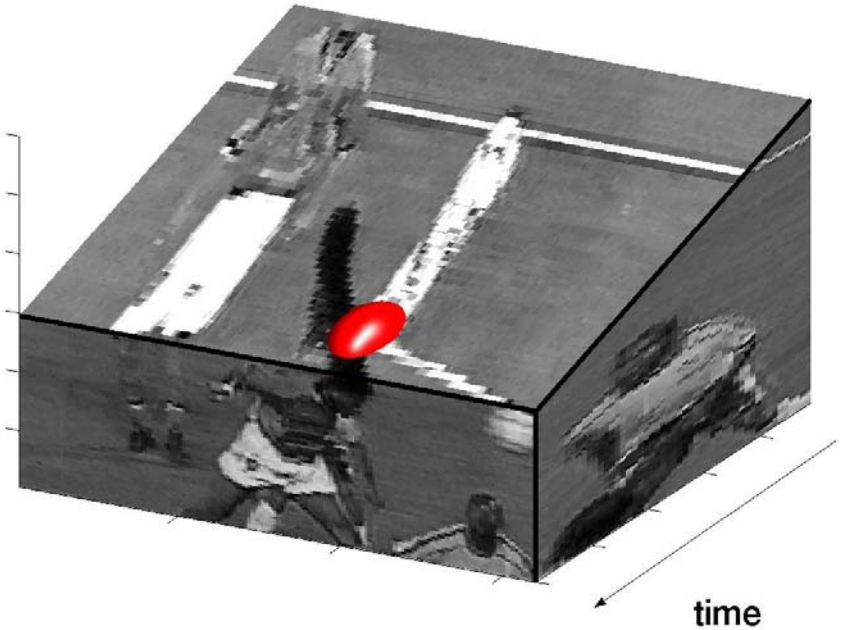
Space-Time Interest Points: Examples

Motion event detection

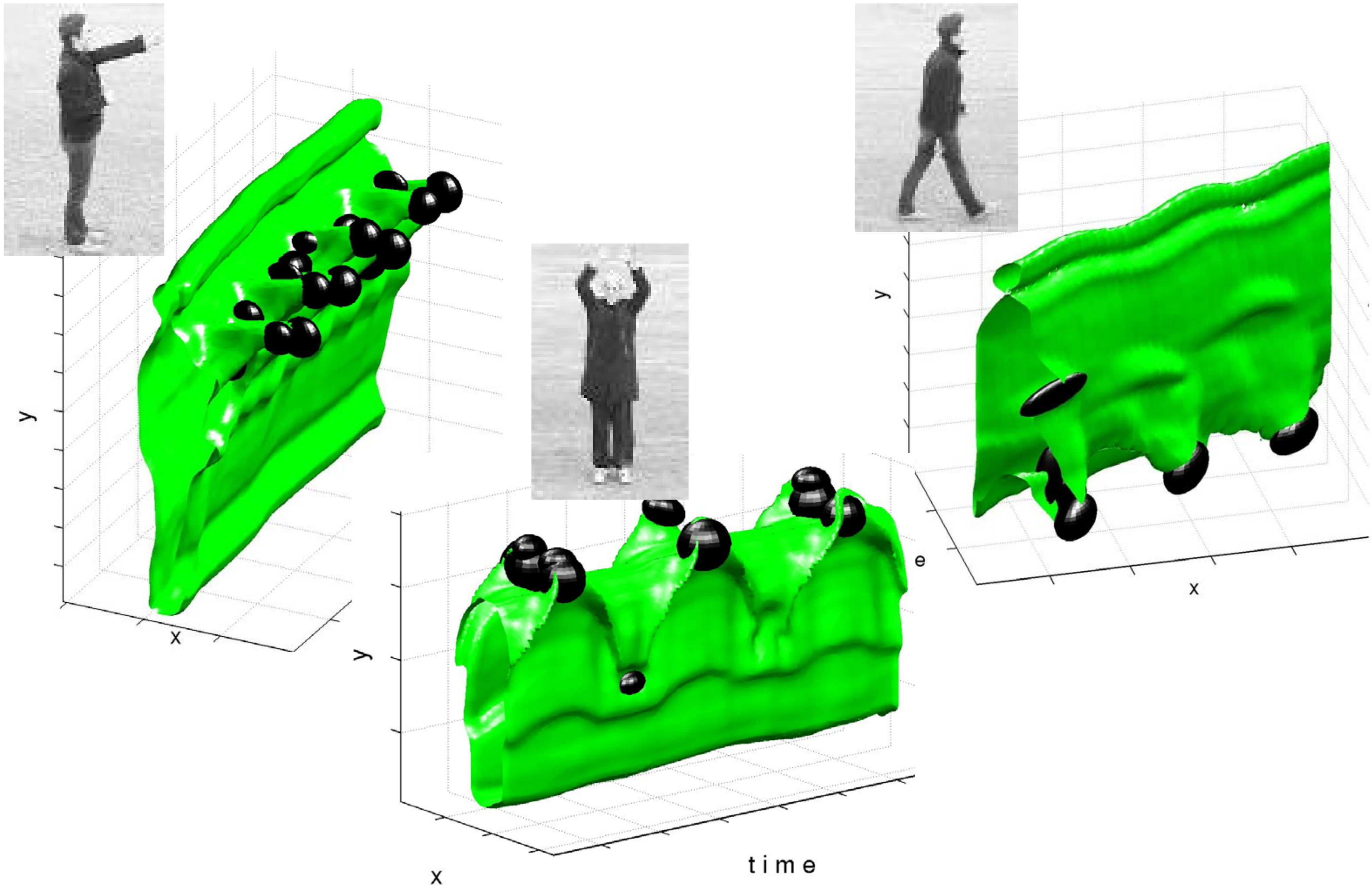


Space-Time Interest Points: Examples

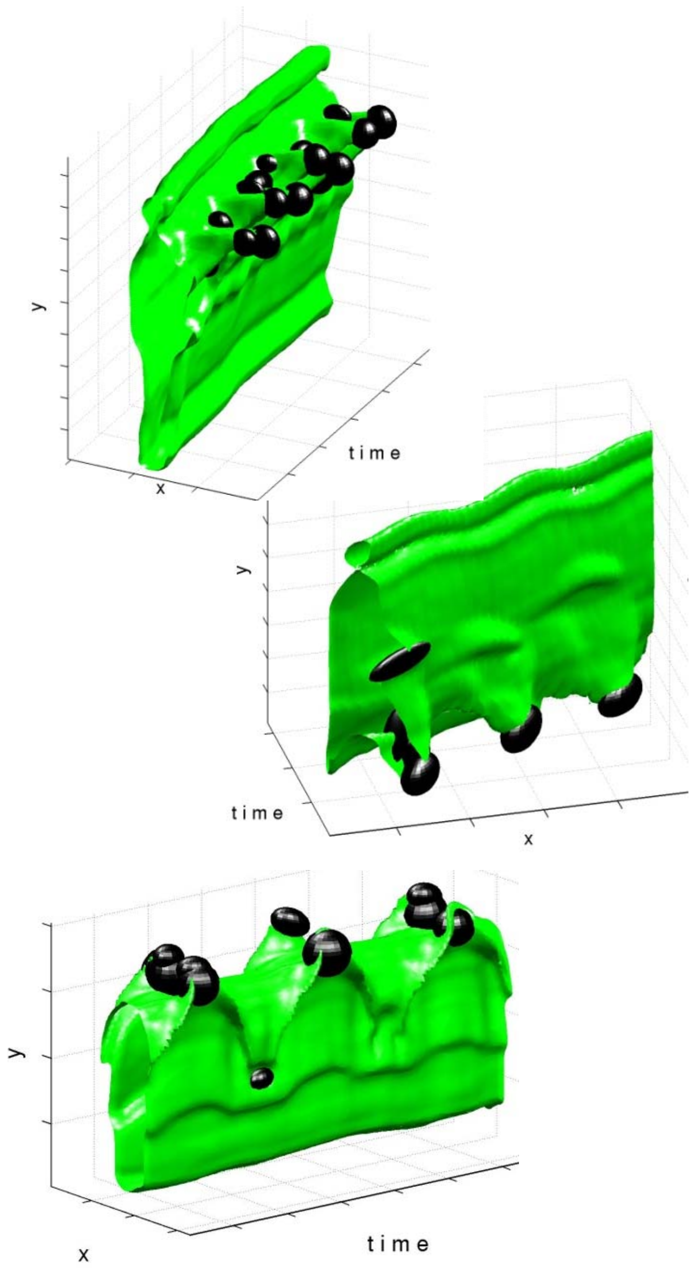
Motion event detection

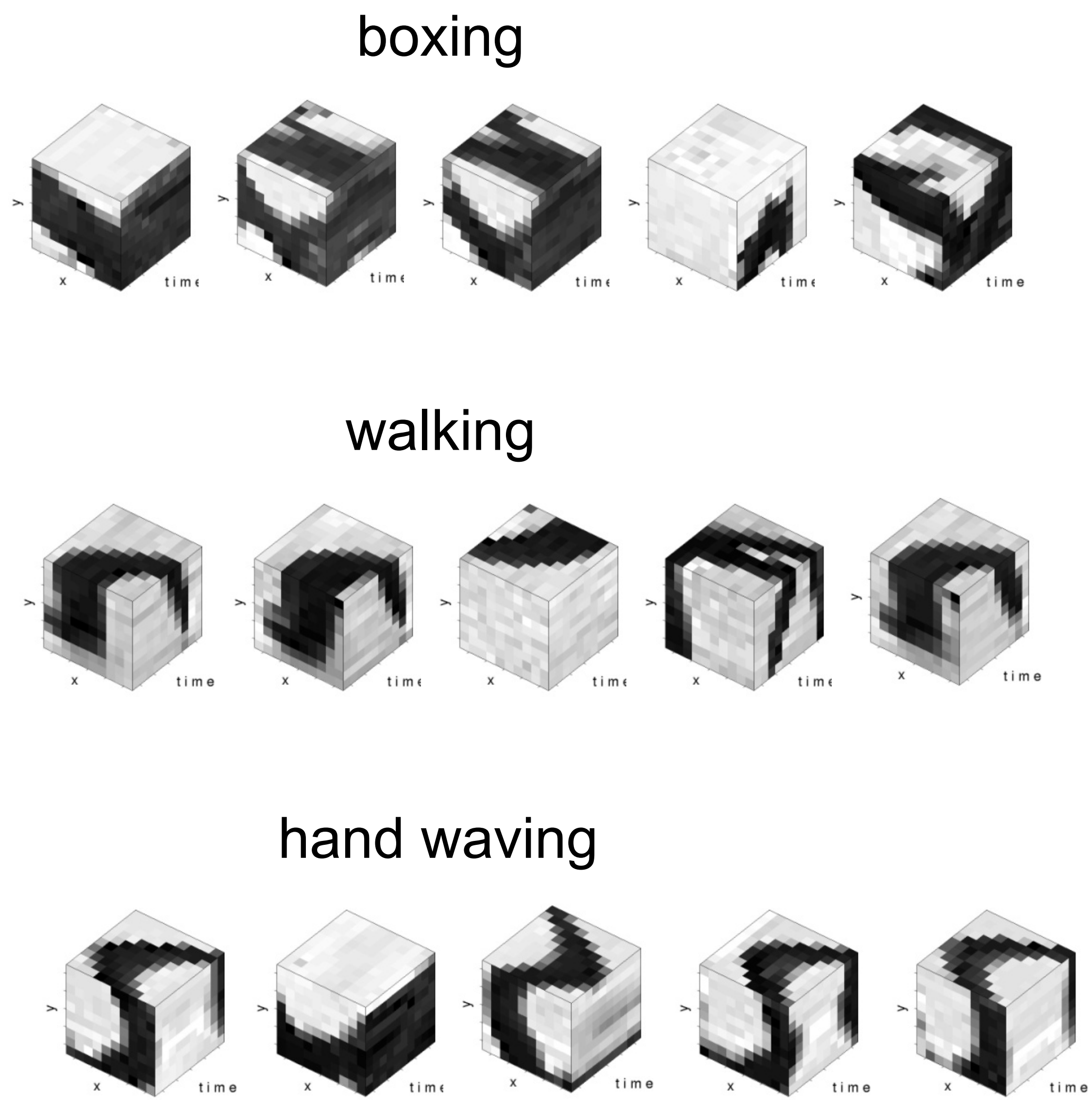


Local features for human actions



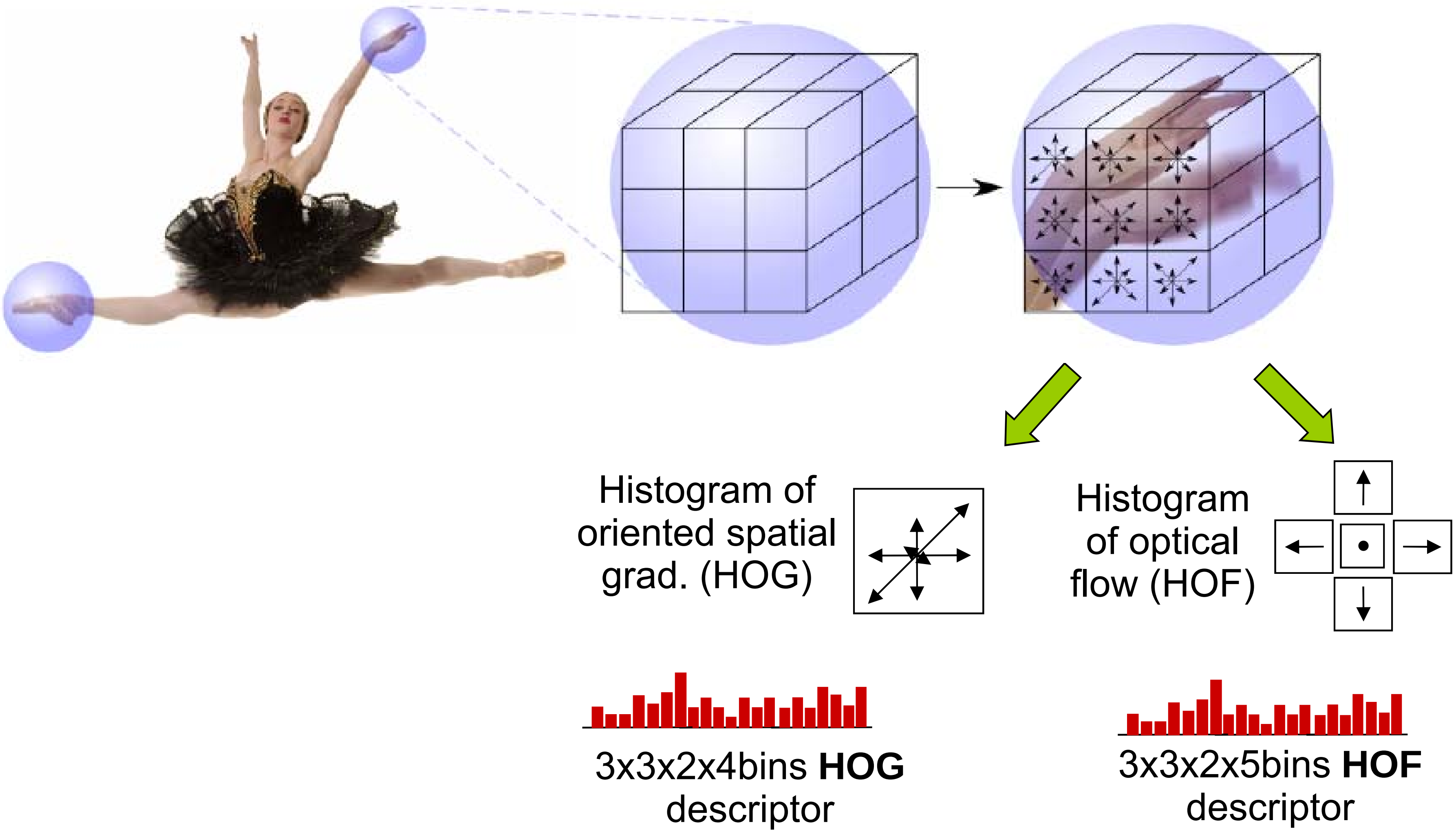
Local features for human actions





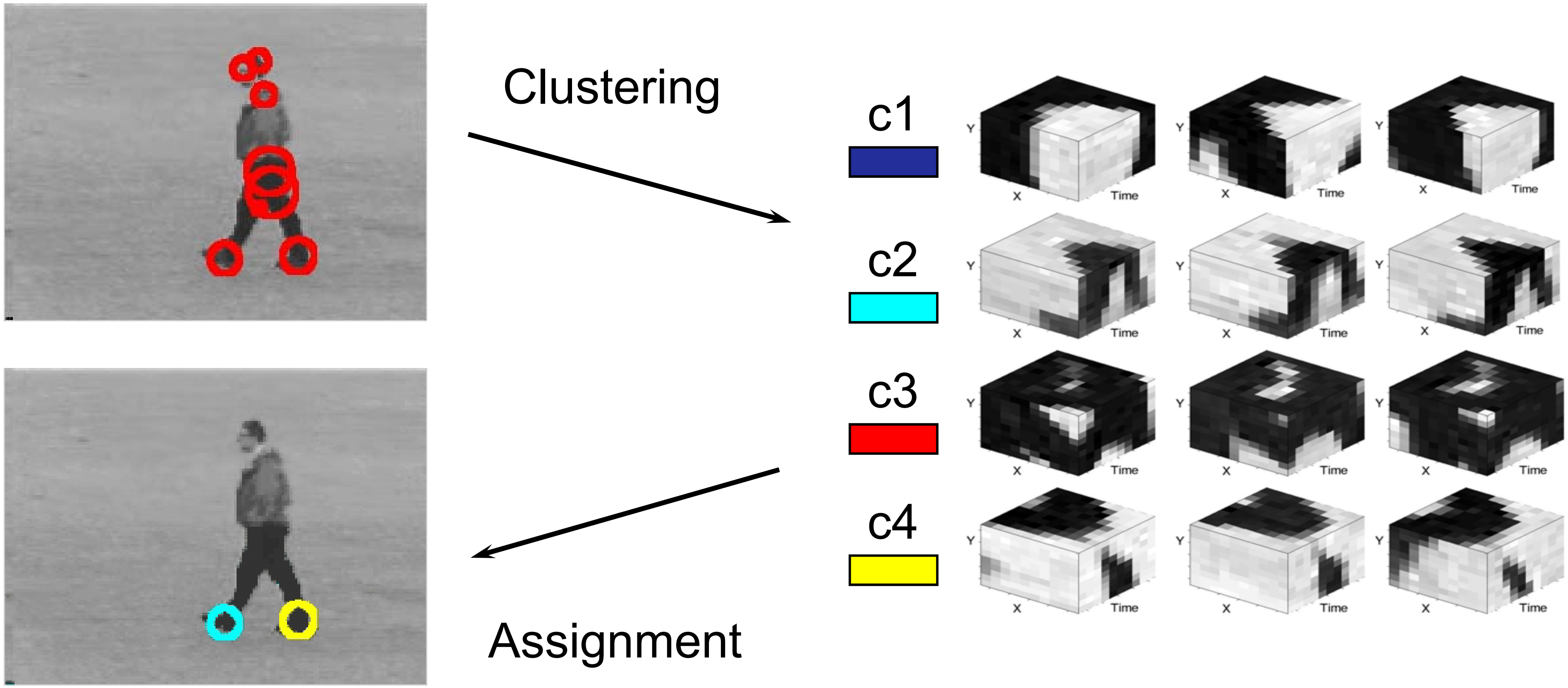
Local space-time descriptor: HOG/HOF

Multi-scale space-time patches



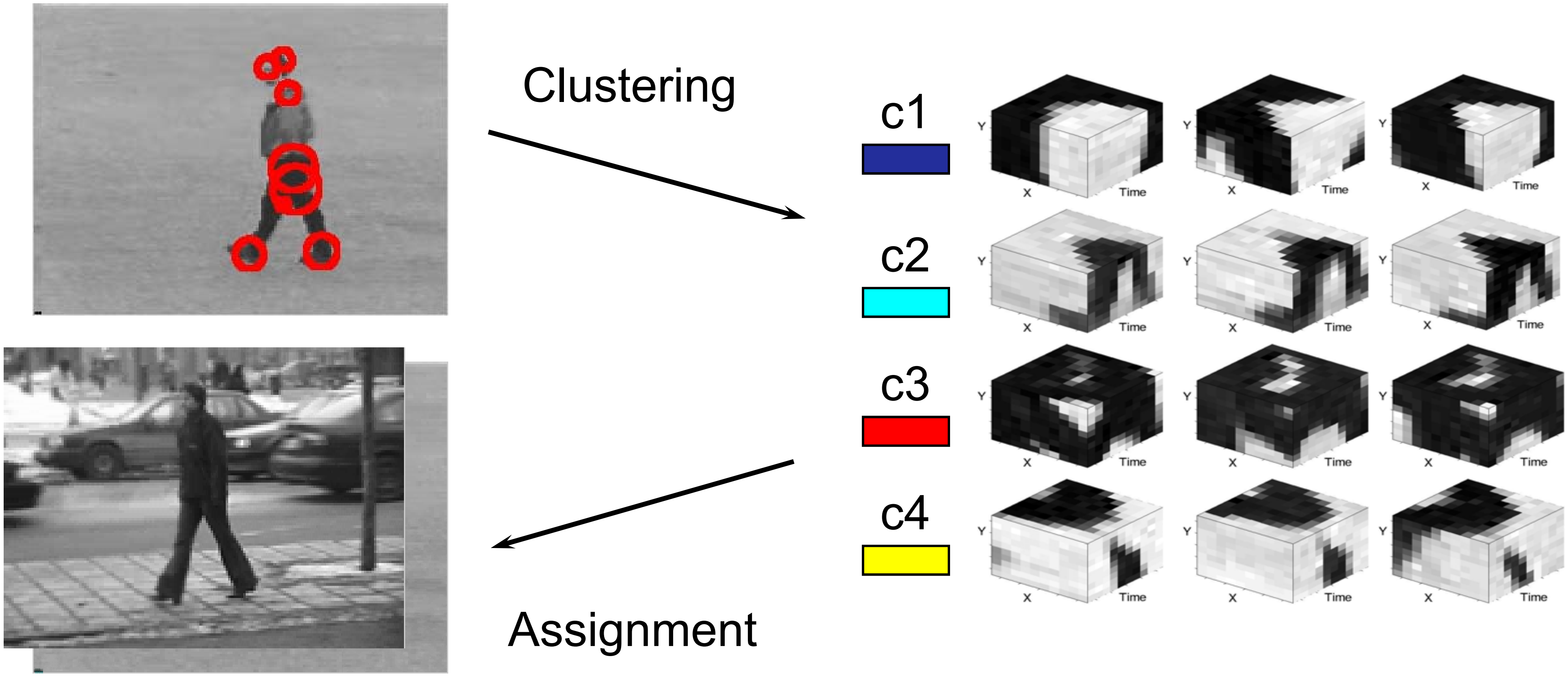
Visual Vocabulary: K-means clustering

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



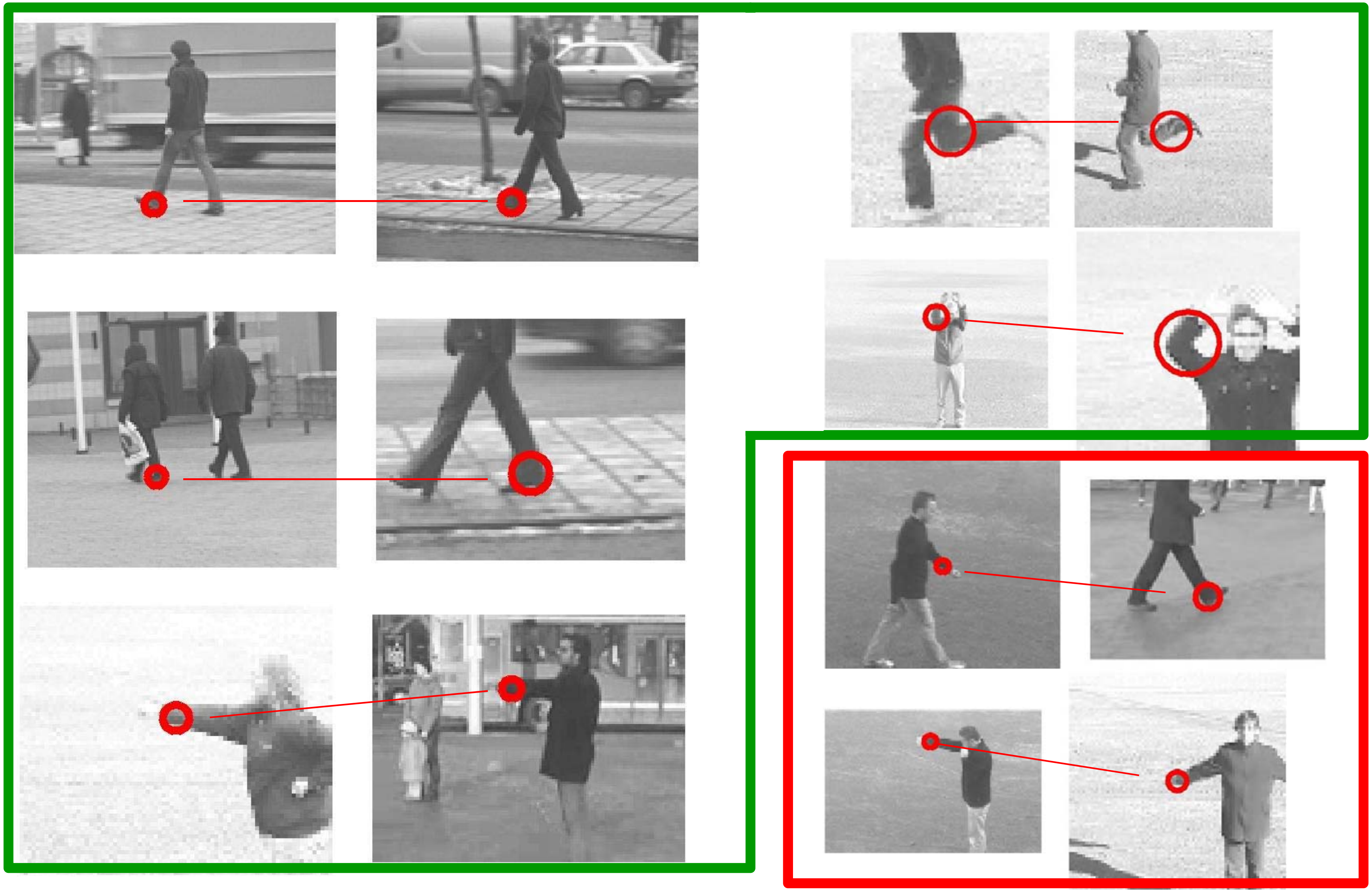
Visual Vocabulary: K-means clustering

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



Local features: Matching

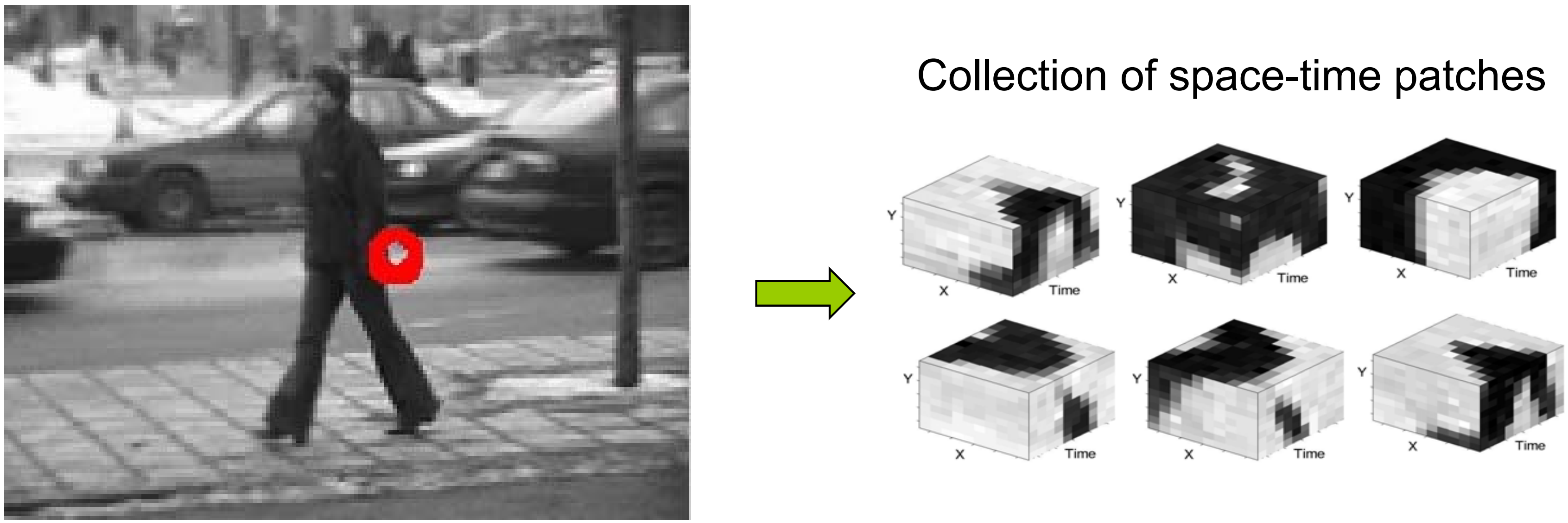
- Finds similar events in pairs of video sequences

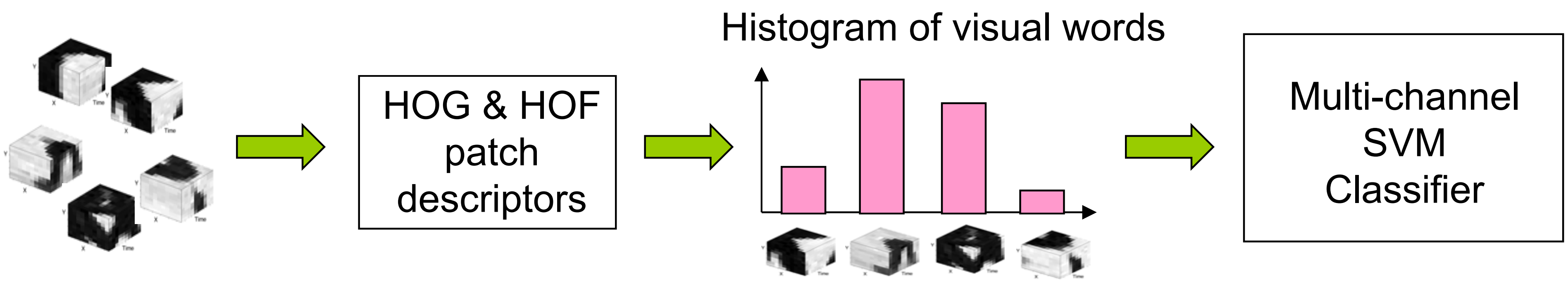


Action Classification

Bag of space-time features + multi-channel SVM

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





Action classification results

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

Hollywood-2 dataset

	hoghof		Chance
Channel	bof	flat	
mAP	47.9	50.3	9.2
AnswerPhone	15.7	20.9	7.2
DriveCar	86.6	84.6	11.5
Eat	59.5	67.0	3.7
FightPerson	71.1	69.8	7.9
GetOutCar	29.3	45.7	6.4
HandShake	21.2	27.8	5.1
HugPerson	35.8	43.2	7.5
Kiss	51.5	52.5	11.7
Run	69.1	67.8	16.0
SitDown	58.2	57.6	12.2
SitUp	17.5	17.2	4.2
StandUp	51.7	54.3	16.5

[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Action classification

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

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]

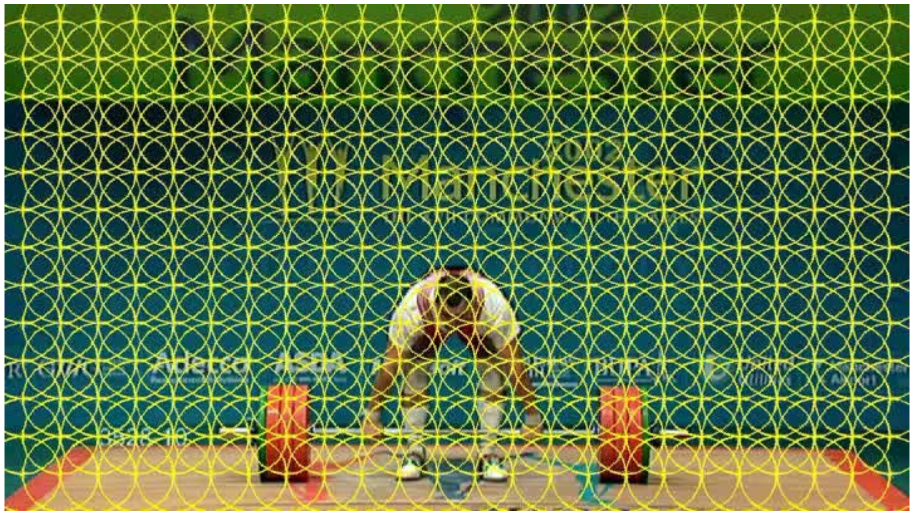
Space-time feature detectors

Harris3D

Hessian

Cuboids

Dense



Results on Hollywood-2

12 action classes collected from 69 movies

Detectors

Descriptors

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

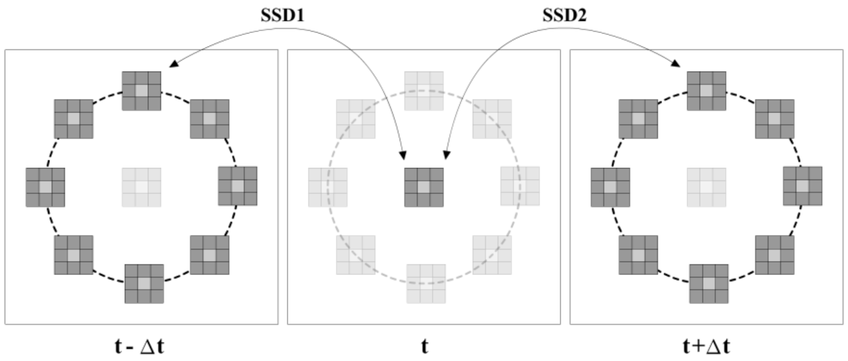
(Average precision scores)

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

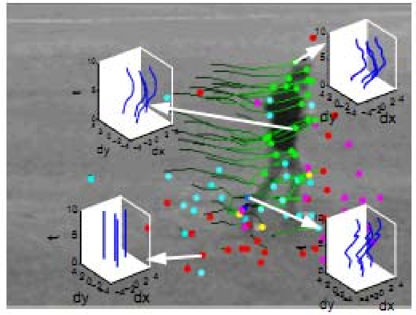
[Wang, Ullah, Kläser, Laptev, Schmid, 2009]

Other recent local representations

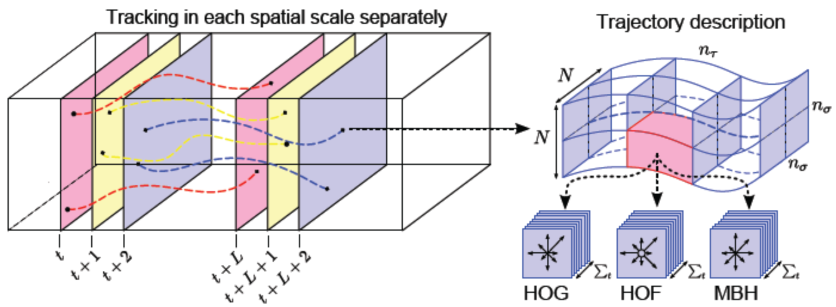
- Y. and L. Wolf, "Local Trinary Patterns for Human Action Recognition ", ICCV 2009



- P. Matikainen, R. Sukthankar and M. Hebert "Trajectons: Action Recognition Through the Motion Analysis of Tracked Features" ICCV VOEC Workshop 2009,

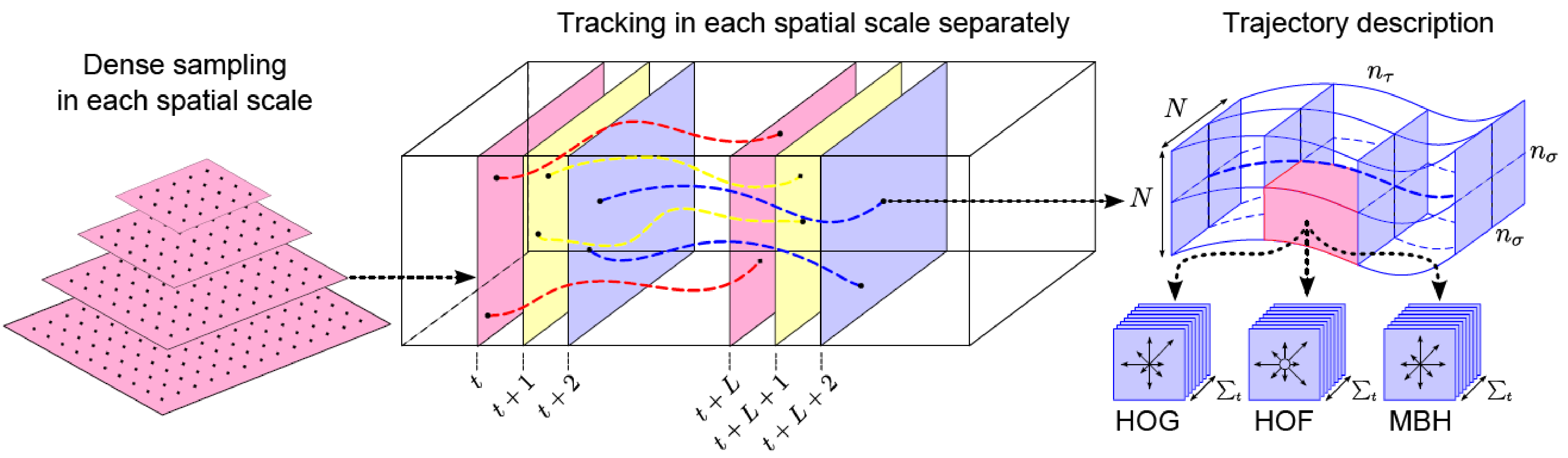


- H. Wang, A. Klaser, C. Schmid, C.-L. Liu, "Action Recognition by Dense Trajectories", CVPR 2011



Dense trajectories [Wang et al. IJCV'13]

- Dense sampling
- Feature tracking based on optical flow
- Trajectory-aligned descriptors



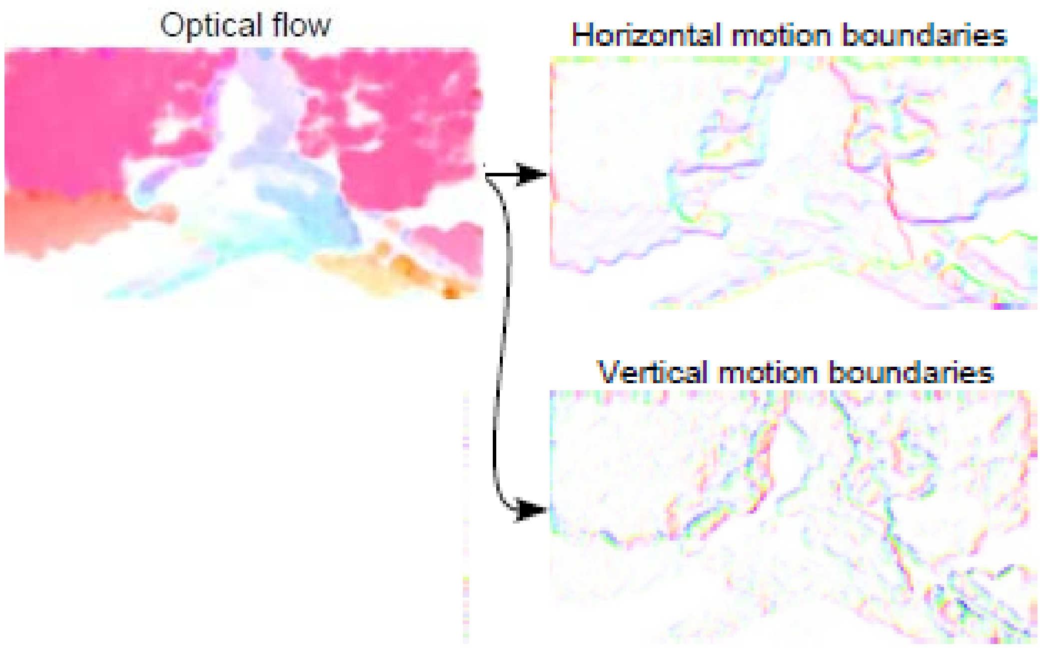
Trajectory descriptors

Motion boundary descriptor

– spatial derivatives are calculated separately for optical flow in x and y, quantized into a histogram

– relative dynamics of different regions

– suppresses constant motions



Dense trajectories

- Advantages:
  - Captures the intrinsic dynamic structures in videos
  - MBH is robust to certain camera motion

- Disadvantages:
  - Generates irrelevant trajectories in background due to camera motion
  - Motion descriptors are modified by camera motion, e.g., HOF, MBH

→ Improved dense trajectories - student presentation

TrecVid MED'13

- 100 positive video clips per event category, 5000 negatives
- Testing on 98000 videos clips, i.e., 4000 hours
- 20 known events, 10 adhoc events
- Videos from publicly available, user-generated content on various Internet sites
- Descriptors: MBH, SIFT, audio, text & speech recognition




informatics mathematics

Quantitative results on TrecVid MED'11

Channel	mAP
Motion	44.65
Static	33.97
Audio	18.15
OCR	10.85
ASR	8.21
Visual=Motion+Static	47.22
Visual+Audio	50.41
Visual+OCR	48.97
Visual+ASR	48.28
Visual+Audio+OCR+ASR	52.28

Quantitative results on TrecVid MED'11

Channel	mAP	Birthday party
Motion	44.65	30.7
Static	33.97	25.9
Audio	18.15	33.3
OCR	10.85	10.1
ASR	8.21	3.6
Visual=Motion+Static	47.22	34.8
Visual+Audio	50.41	47.7
Visual+OCR	48.97	35.8
Visual+ASR	48.28	35.0
Visual+Audio+OCR+ASR	52.28	48.4

informatics mathematics

Quantitative results on TrecVid MED'11

Channel	mAP	Birthday party	Repair appliance
Motion	44.65	30.7	42.6
Static	33.97	25.9	43.6
Audio	18.15	33.3	43.3
OCR	10.85	10.1	32.1
ASR	8.21	3.6	39.2
Visual=Motion+Static	47.22	34.8	47.5
Visual+Audio	50.41	47.7	54.5
Visual+OCR	48.97	35.8	50.8
Visual+ASR	48.28	35.0	54.5
Visual+Audio+OCR+ASR	52.28	48.4	57.2

Quantitative results on TrecVid MED'11

Channel	mAP	Birthday party	Repair appliance	Make sandwich
Motion	44.65	30.7	42.6	22.5
Static	33.97	25.9	43.6	21.5
Audio	18.15	33.3	43.3	11.2
OCR	10.85	10.1	32.1	19.4
ASR	8.21	3.6	39.2	6.7
Visual=Motion+Static	47.22	34.8	47.5	27.8
Visual+Audio	50.41	47.7	54.5	27.3
Visual+OCR	48.97	35.8	50.8	35.7
Visual+ASR	48.28	35.0	54.5	28.8
Visual+Audio+OCR+ASR	52.28	48.4	57.2	35.4

TrecVid MED 2013 – example results

rank 1

rank 2

rank 3

Horse riding competition

TrecVid MED 2013 – example results

rank 1

rank 2

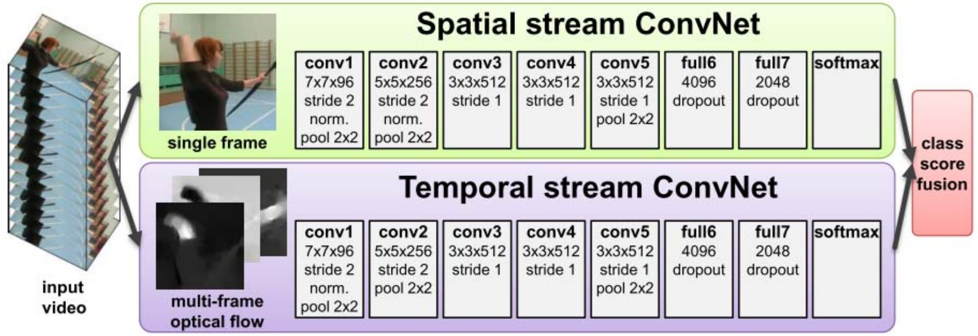
rank 3

Tuning a musical instrument

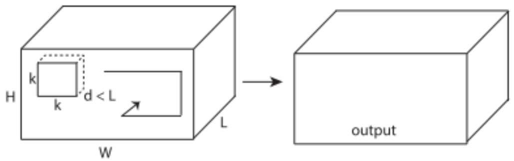
informatics mathematics

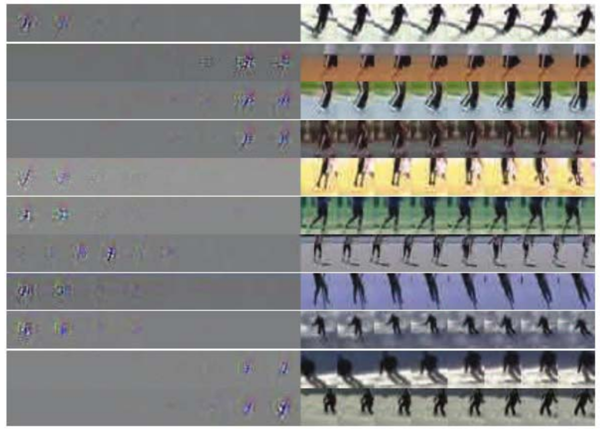
Recent CNN methods

Two-Stream Convolutional Networks for Action Recognition in Videos [Simonyan and Zisserman NIPS14]



Learning Spatiotemporal Features with 3D Convolutional Networks [Tran et al. ICCV15]





Action recognition with trajectory pooled convolutional descriptors [Wang et al. CVPR15]

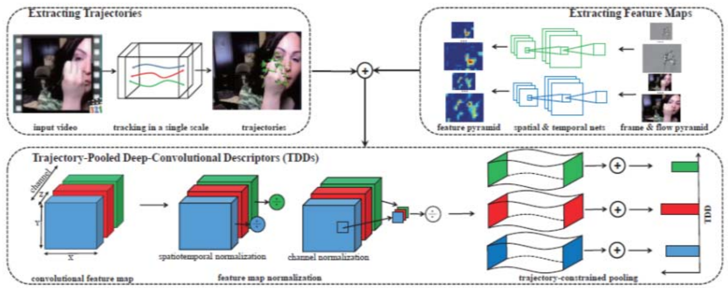
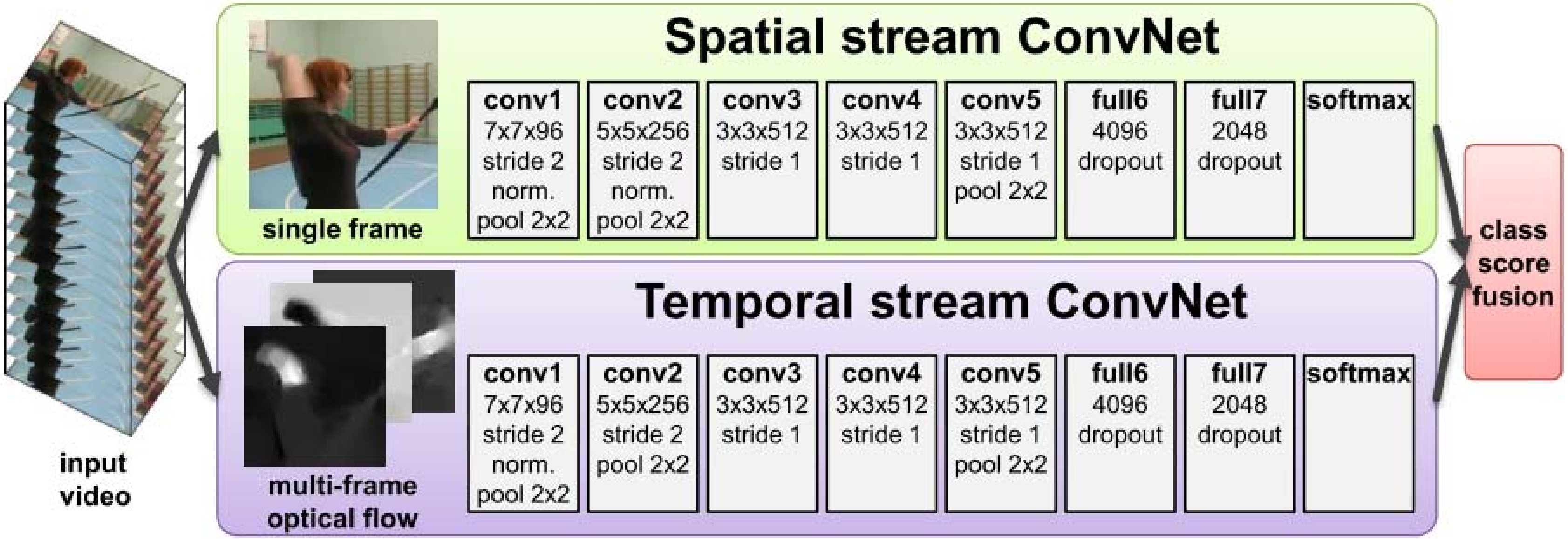


Figure 2. **Pipeline of TDD.** The whole process of extracting TDD is composed of three steps: (i) extracting trajectories, (ii) extracting multi-scale convolutional feature maps, and (iii) calculating TDD. We effectively exploit two available state-of-the-art video representations, namely improved trajectories and two-stream ConvNets. Grounded on them, we conduct trajectory-constrained sampling and pooling over convolutional feature maps to obtain trajectory-pooled deep convolutional descriptors.

Recent CNN methods

Two-Stream Convolutional Networks
for Action Recognition in Videos
[Simonyan and Zisserman NIPS14]



Recent CNN methods

Learning Spatiotemporal Features with
3D Convolutional Networks
[Tran et al. ICCV15]

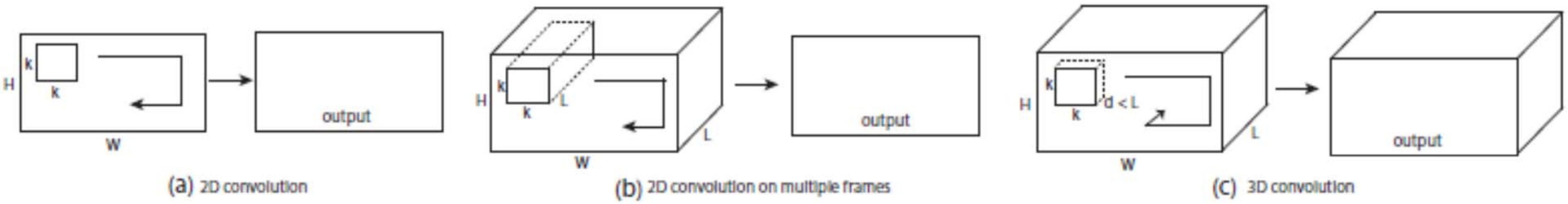
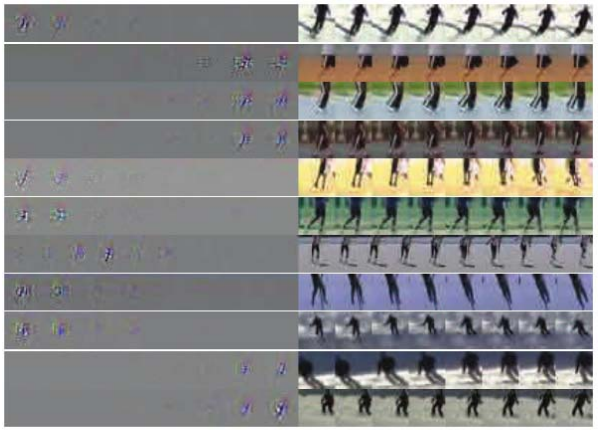


Figure 1. **2D and 3D convolution operations**. a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.



Recent CNN methods

Action recognition with trajectory pooled
convolutional descriptors
[Wang et al. CVPR15]

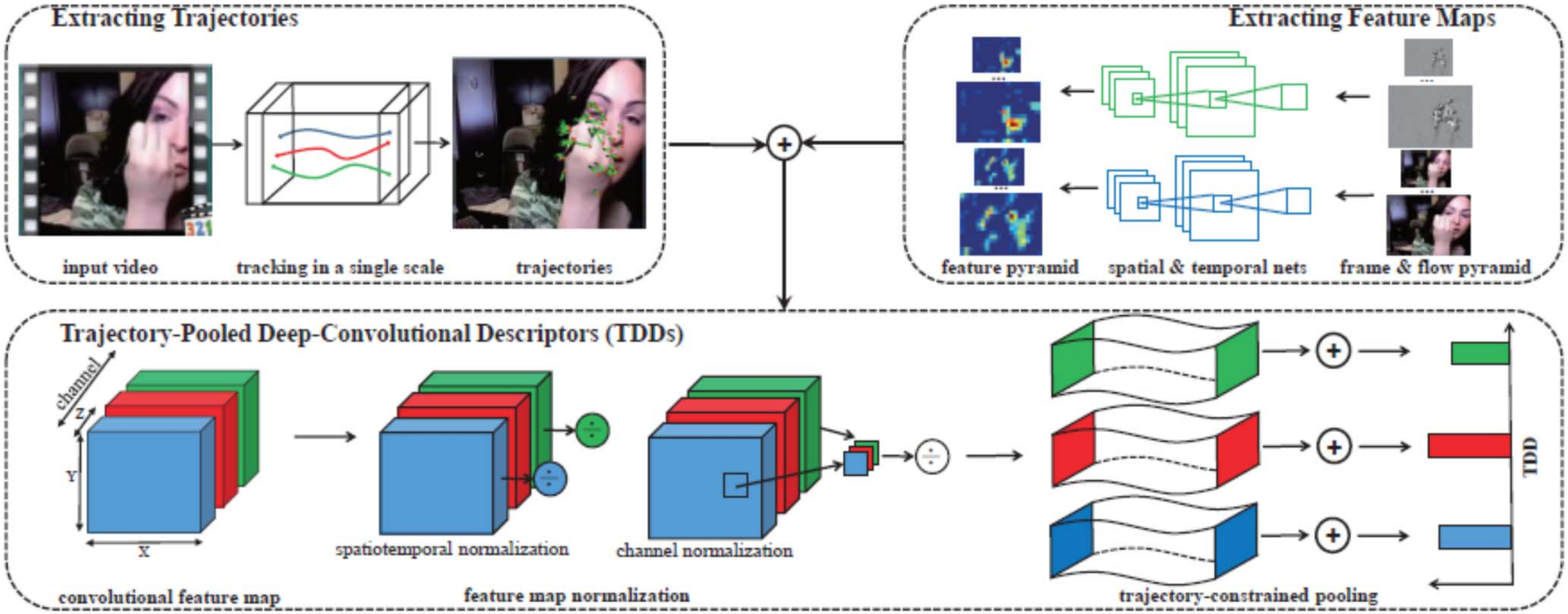


Figure 2. **Pipeline of TDD.** The whole process of extracting TDD is composed of three steps: (i) extracting trajectories, (ii) extracting multi-scale convolutional feature maps, and (iii) calculating TDD. We effectively exploit two available state-of-the-art video representations, namely improved trajectories and two-stream ConvNets. Grounded on them, we conduct trajectory-constrained sampling and pooling over convolutional feature maps to obtain trajectory-pooled deep convolutional descriptors.

Action recognition - tasks

- Action classification: assigning an action label to a video clip

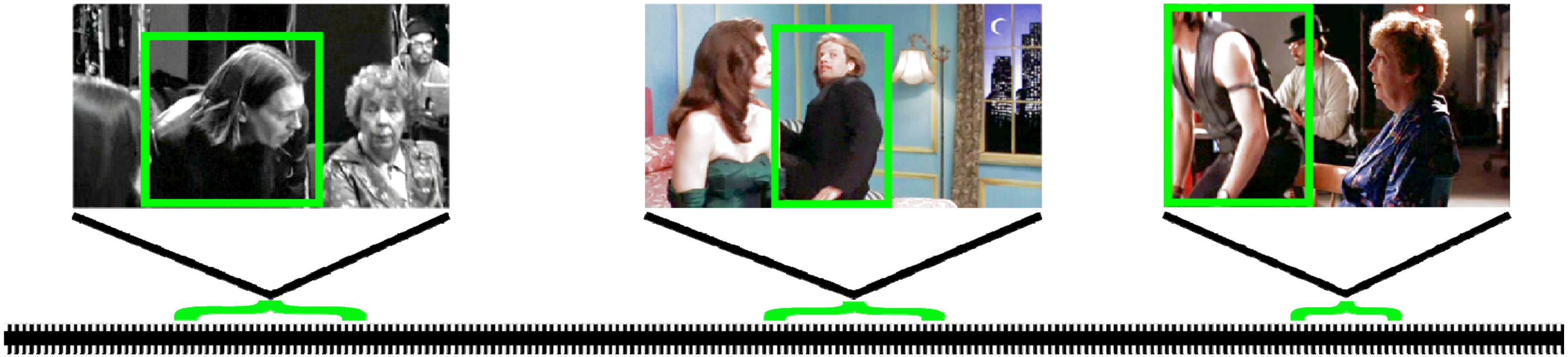
Making sandwich: present
Feeding animal: not present
...

Action recognition - tasks

- Action classification: assigning an action label to a video clip

Making sandwich: present
Feeding animal: not present
...

- Action localization (temporal): search temporal locations of an action in a video



Action recognition - tasks

- Action localization (spatio-temporal) + interaction with an object, human, etc.

[Prest et al., PAMI 13]

Why automatic action localization?

- Query for specific videos in professional Archives and YouTube
- Analyze and describe content of videos
- Produce audio descriptions for visual impaired

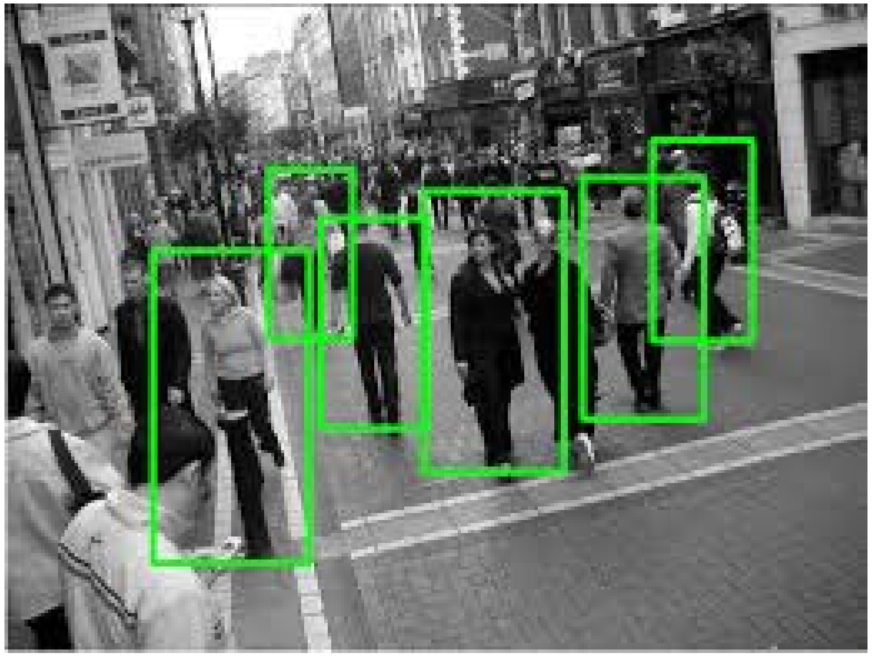
Education: How do I make a pizza?

Sociology research: Influence of character smoking in movies

Why automatic action localization?

- Car safety & self-driving and video surveillance
- Detection of humans (pedestrians) and their motion, detection of unusual behavior

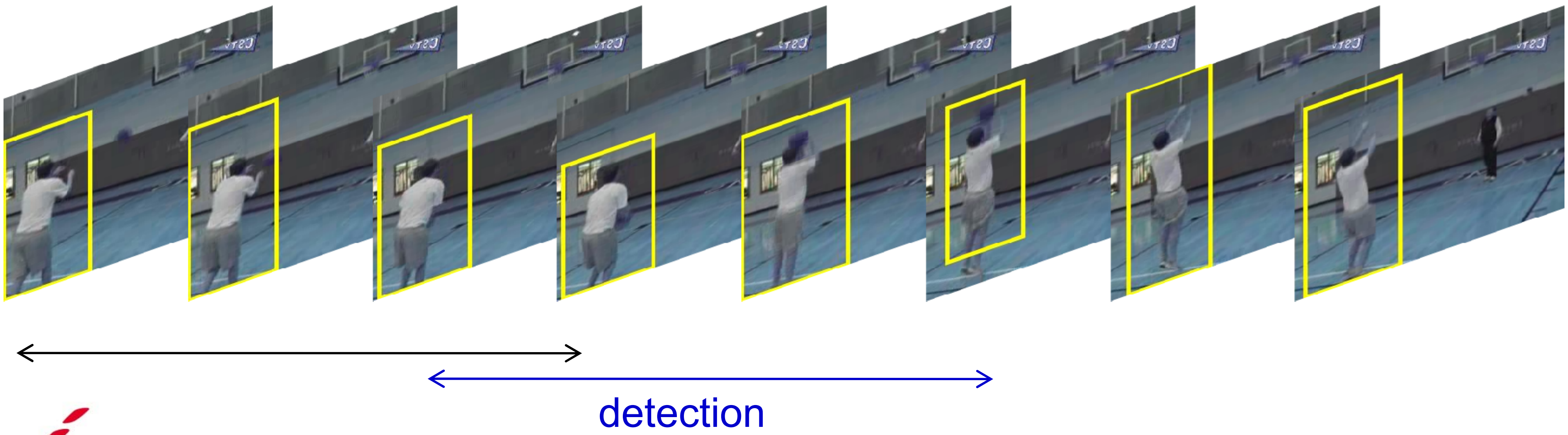
Courtesy Volvo



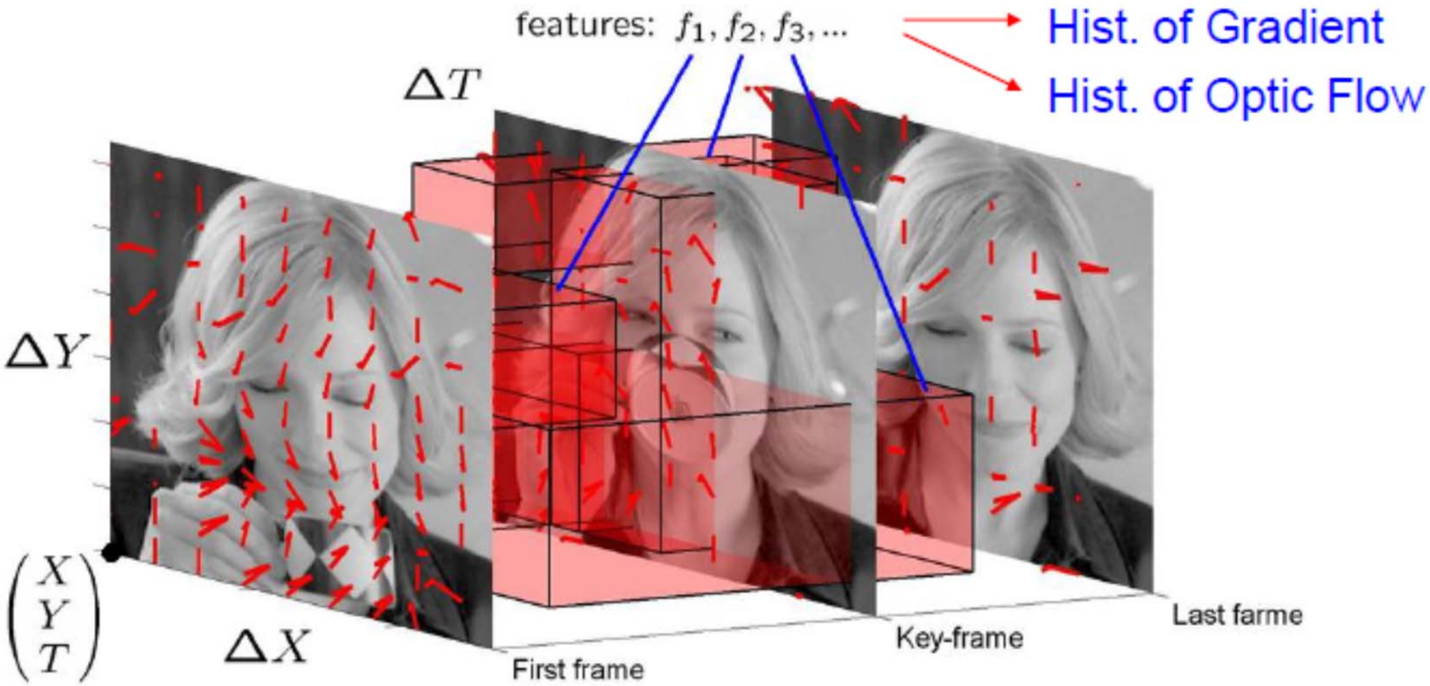
Courtesy Embedded Vision Alliance

Temporal action localization

- Temporal sliding window
  - Robust video repres. for action recognition, Oneata et al., IJCV’15
  - Automatic annotation of actions in video, Duchenne et al., ICCV’09
  - Temporal localization of actions with actoms, Gaidon et al., PAMI’13
- Shot detection
  - ADSC Submission at Thumos Challenge 2015

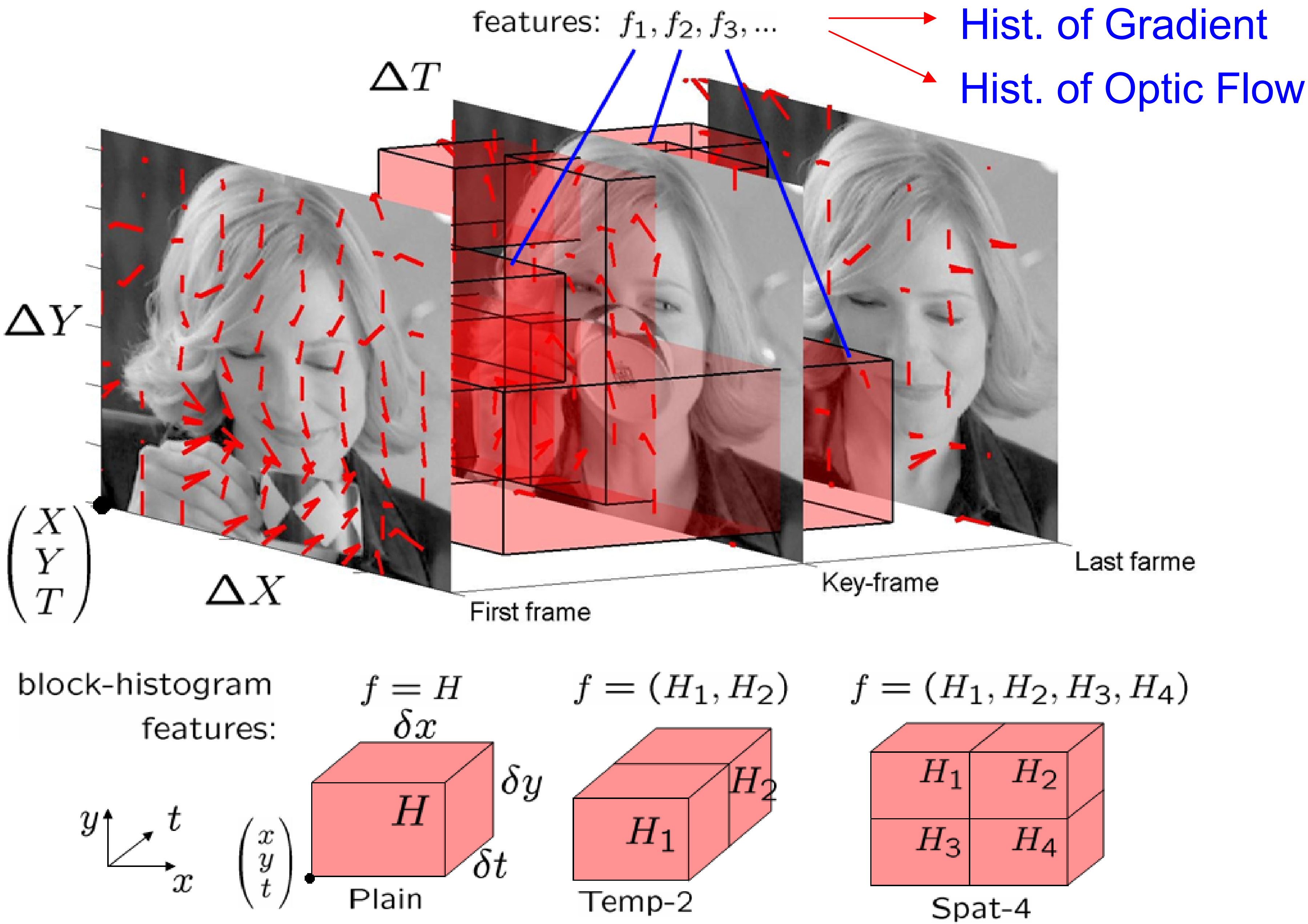


Spatio-temporal action localization

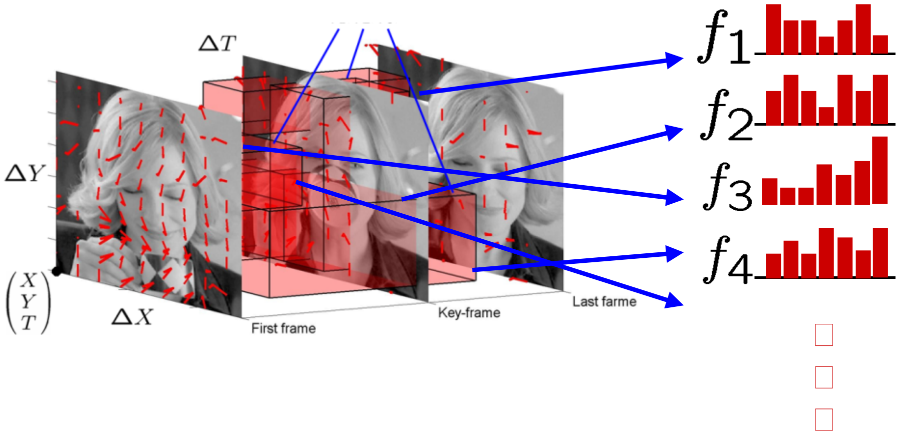


[Retrieving actions in movies, I. Laptev and P. Pérez, ICCV'07]

Action representation



Action learning



boosting

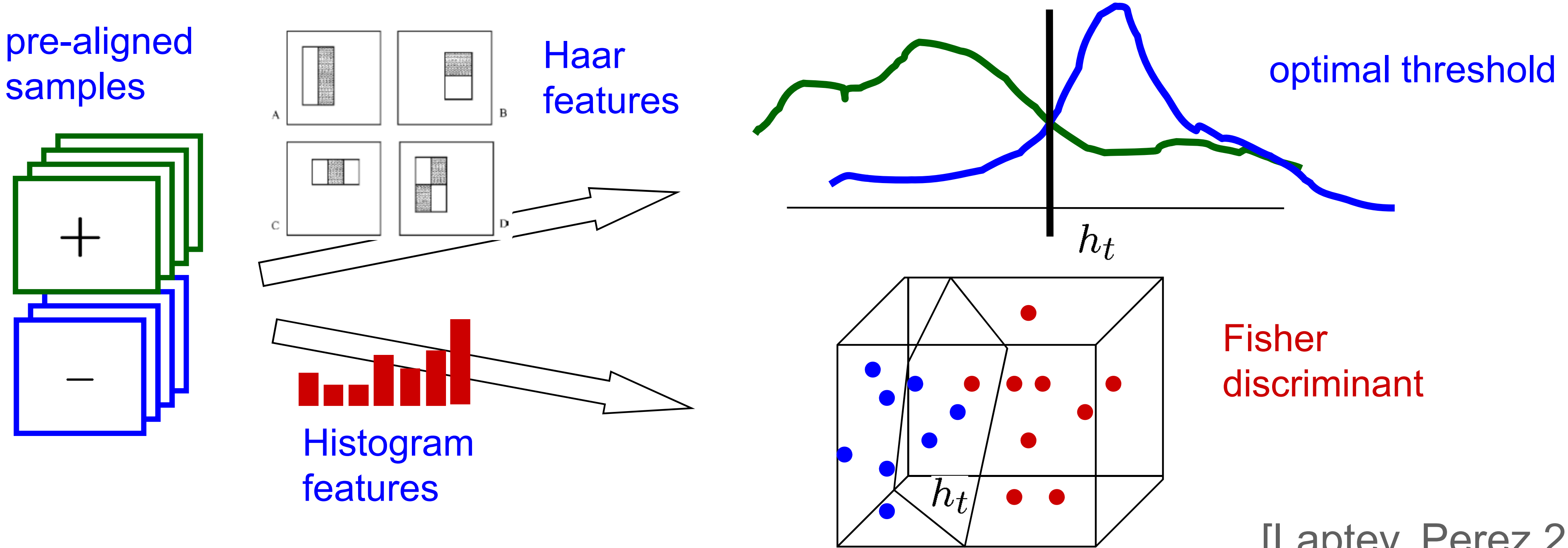
selected features

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

weak classifier

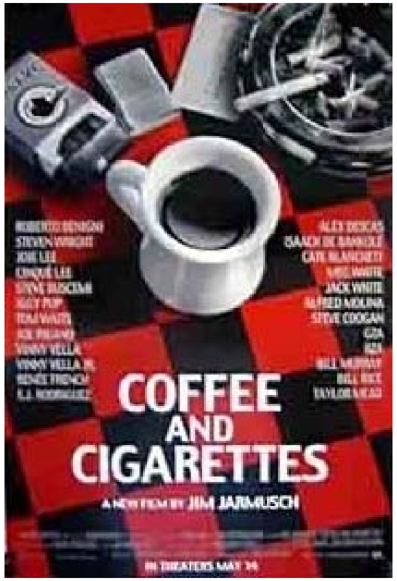
AdaBoost:

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



[Laptev, Perez 2007]

Dataset for action localization



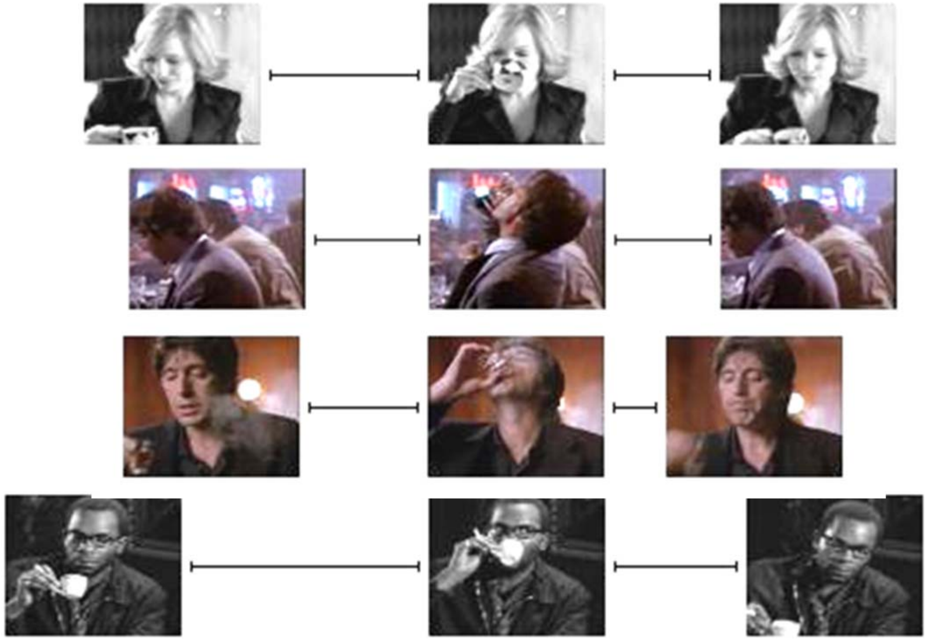
Manual annotation of drinking actions in movies:
"Coffee and Cigarettes"; "Sea of Love"

"*Drinking*": 159 annotated samples

"*Smoking*": 149 annotated samples

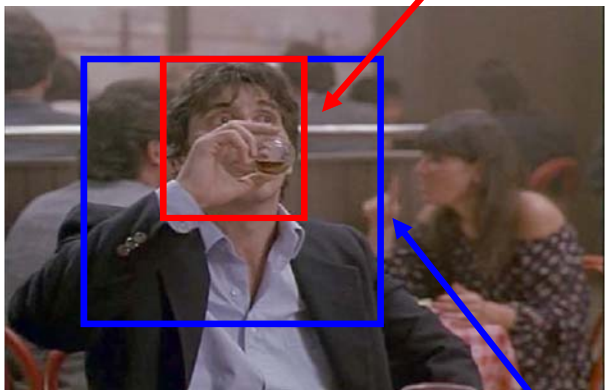
Temporal annotation

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



Spatial annotation

head rectangle



torso rectangle

Action Detection

Test episodes from the movie “Coffee and cigarettes”

[Laptev, Perez 2007]

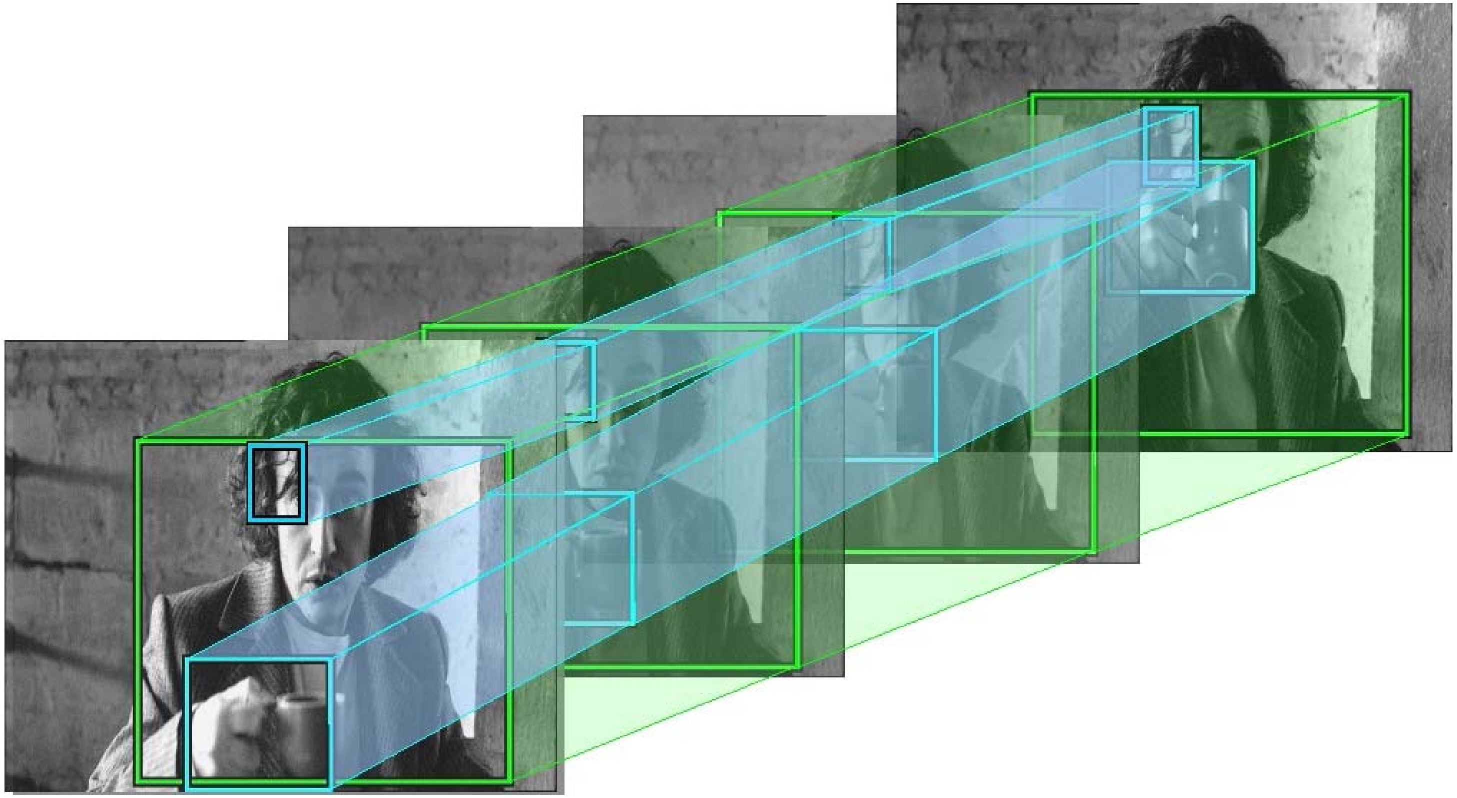
20 most confident detections

Spatio-temporal action localization

- Modeling temporal human-object interaction

[Explicit modeling of human-object interactions in realistic videos, Prest et al., PAMI 13]

Tracking humans and objects

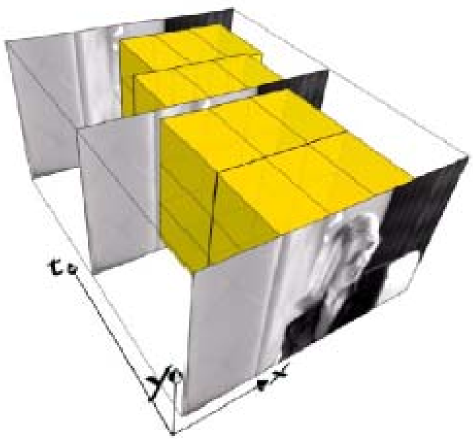


- Fully automatic human tracks: state of the art detector + Brox tracks
- Object tracks: detector learnt from annotated training images + Brox tracks
- Extraction of a large number of human-object track pairs

Action descriptors

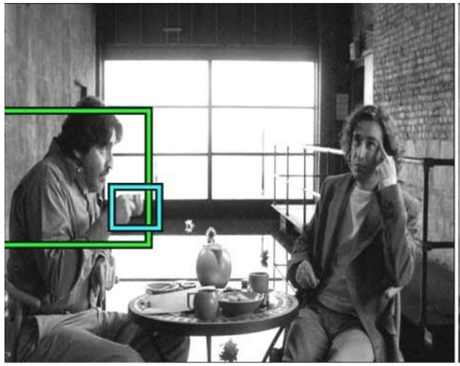
- Interaction descriptor: relative location, area and motion between human and object tracks

- Human track descriptor: 3DHOG-track [Klaeser et al.'10]

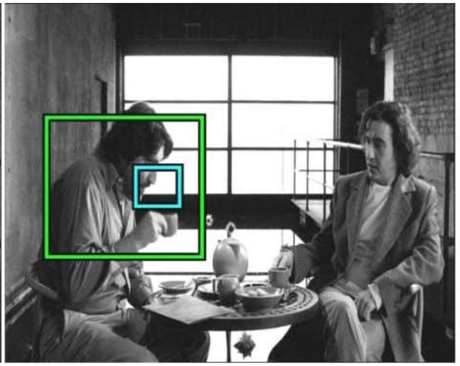


Experimental results on C&C

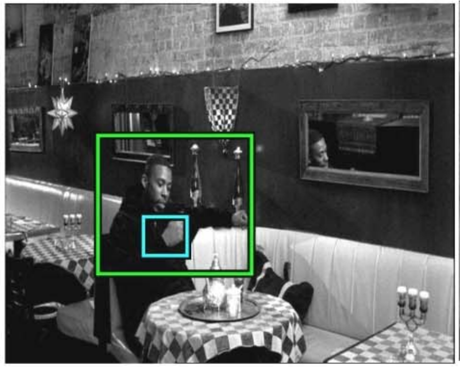
Drinking



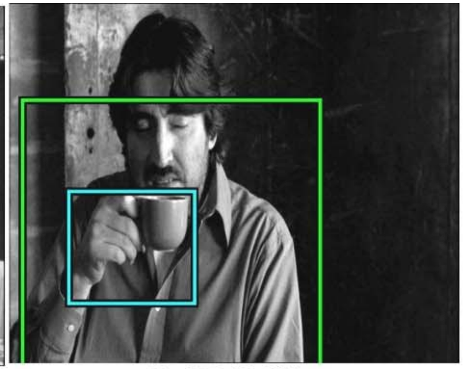
1 (POS)
I: 7 H: 1



2 (POS)
I: 17 H: 2



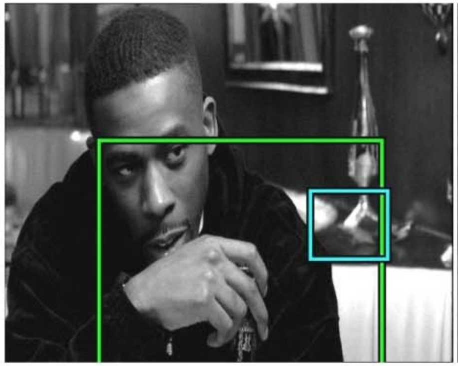
3 (POS)
I: 11 H: 3



6 (POS)
I: 6 H: 4

10 (POS)
I: 21 H: 10

11 (POS)
I: 9 H: 12



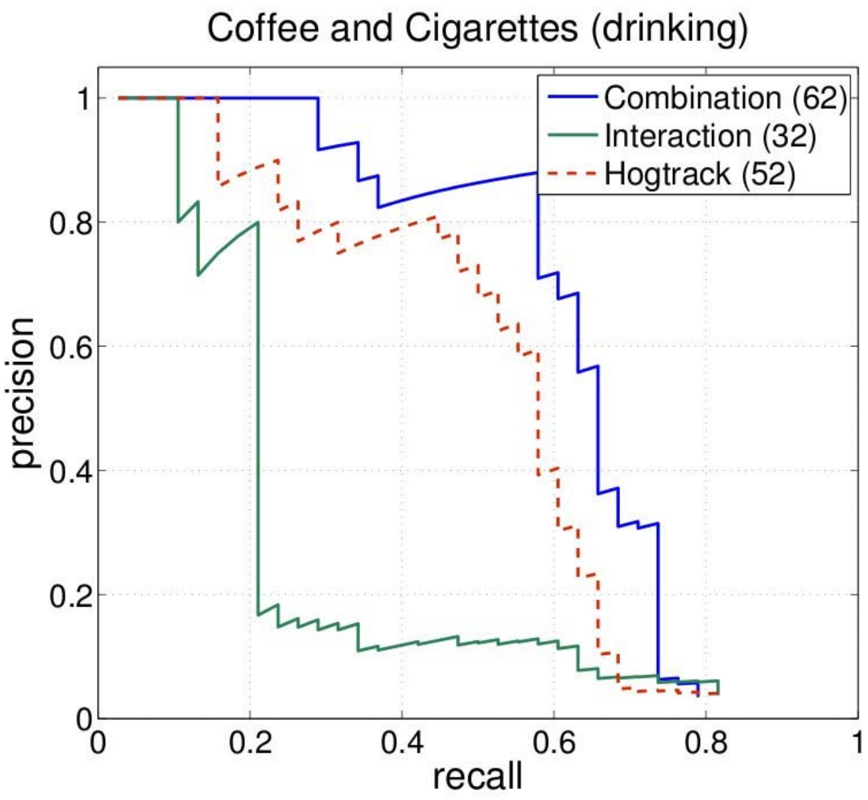
12 (NEG)
I: 33 H: 9

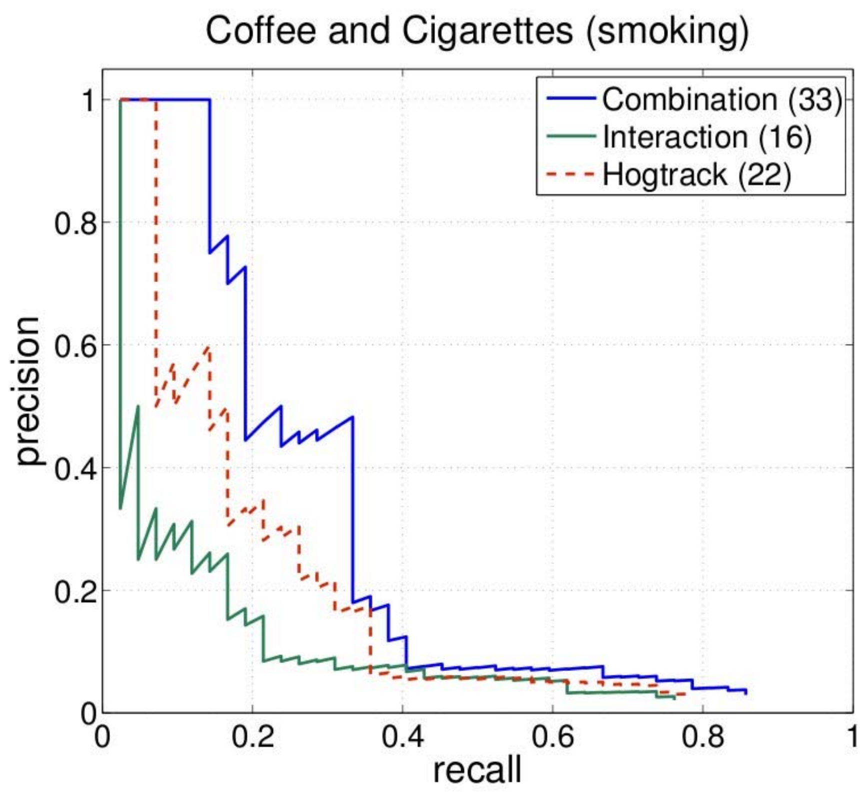
13 (POS)
I: 3 H: 23

Experimental results on C&C

Smoking

Experimental results on C&C





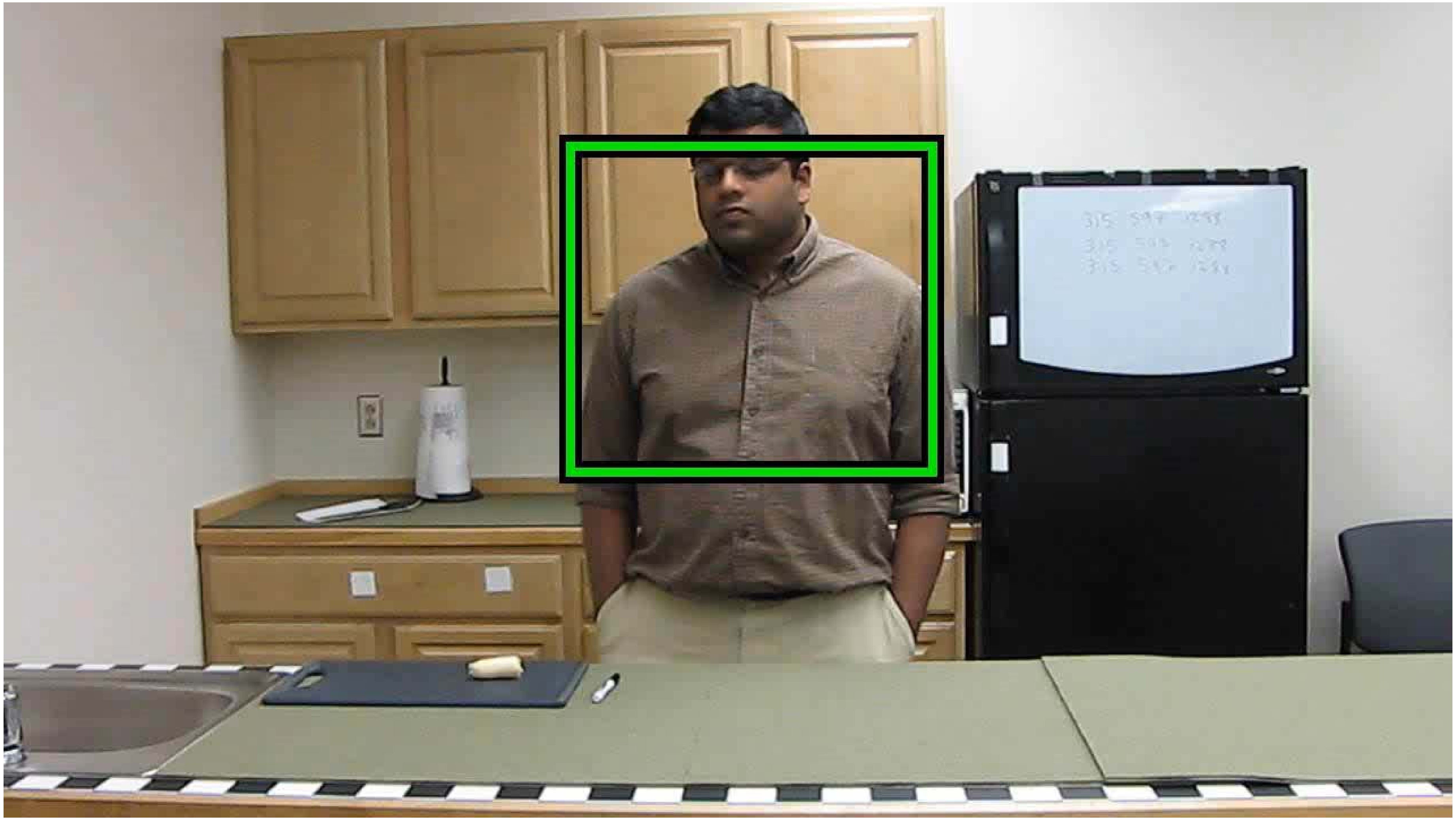
Comparison to the state of the art

	Drinking	Smoking
Interaction classifier	**31.60**	**16.20**
Object classifier	4.30	5.50
3DHOG-track classifier	52.20	21.50
Combination	**62.10**	**32.80**
Laptev et al. [22]	43.40	-
Willems et al. [35]	45.20	-
Klaeser et al. [20]	54.10	24.50

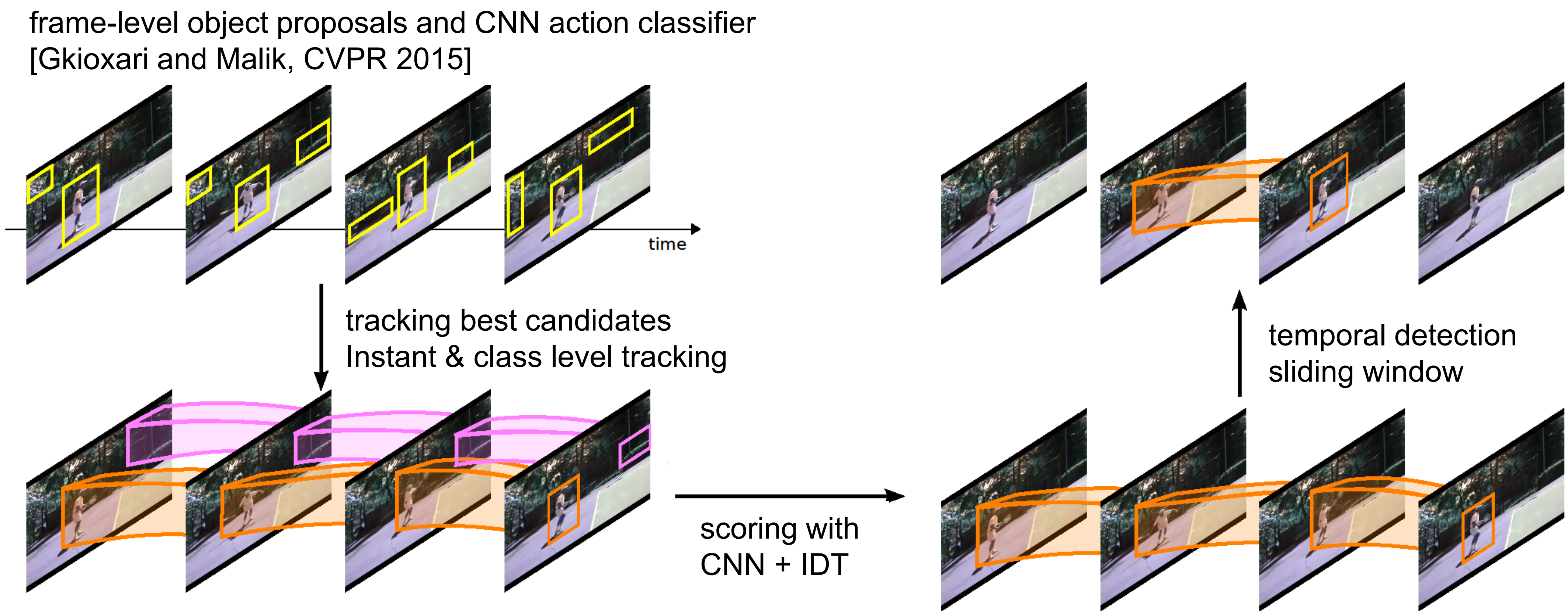
Experimental results on Rochester dataset

- Rochester daily activities dataset
  - 150 videos of 5 persons
  - leave-one-person-out test scenario

Experimental results on Rochester dataset



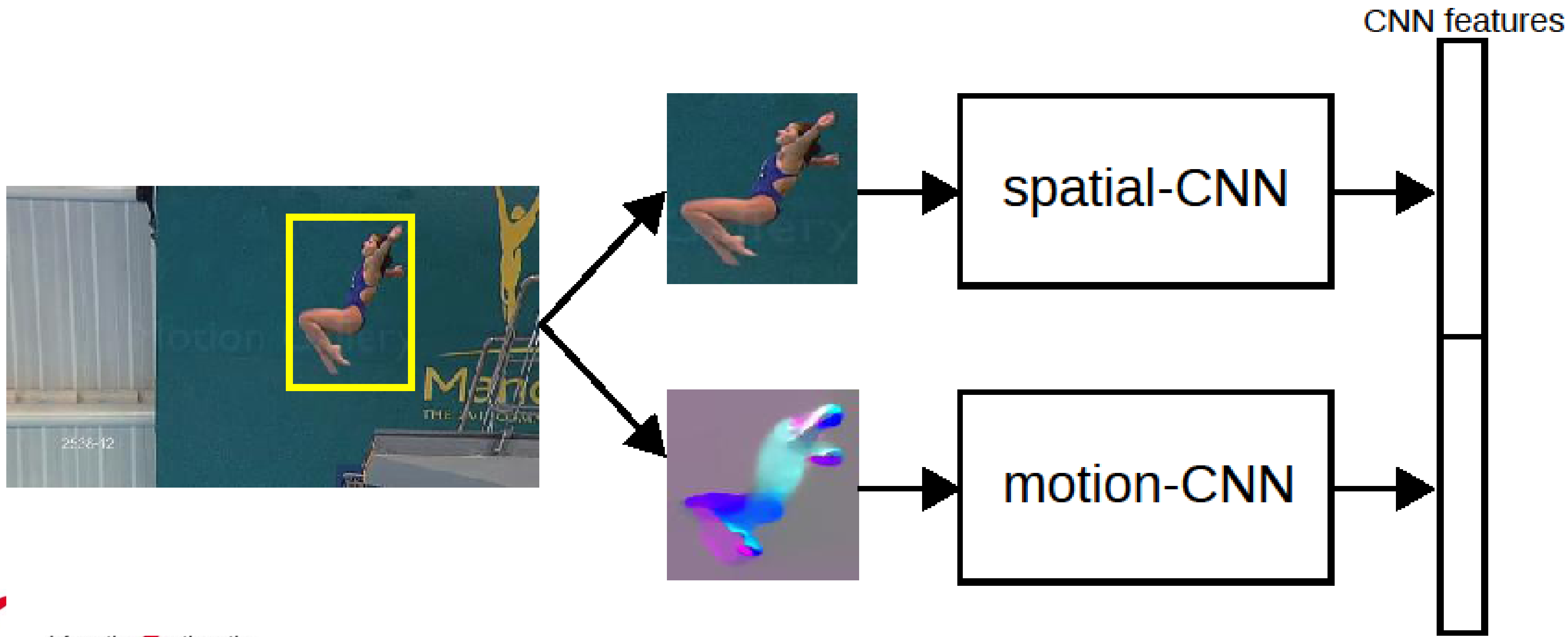
Learning to track for spatio-temporal action localization



[Learning to track for spatio-temporal action localization, P. Weinzaepfel, Z. Harchaoui, C. Schmid, ICCV 2015]

Frame-level candidates

- For each frame
  - ► Compute object proposals (EdgeBoxes [Zitnick et al. 2014])
  - ► Extract CNN features (training similar to R-CNN [Girshicket al. 2014])
  - ► Score each object proposal



[Gkioxari and Malik’15, Simonyan and Zisserman’14]

Tracking best candidates

- Select the top scoring proposals

- For each selected candidate
  - ► Learn an instance-level detector
  - ► For each frame
    - Perform a sliding-window and select the best box according to the class-level detector and the instance-level detector
    - Update instance-level detector

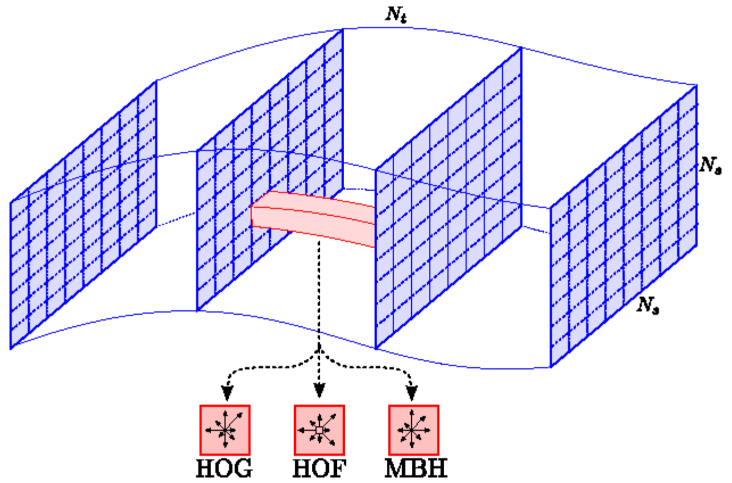
class-level → robustness to drastic change in poses (Diving, Swinging)

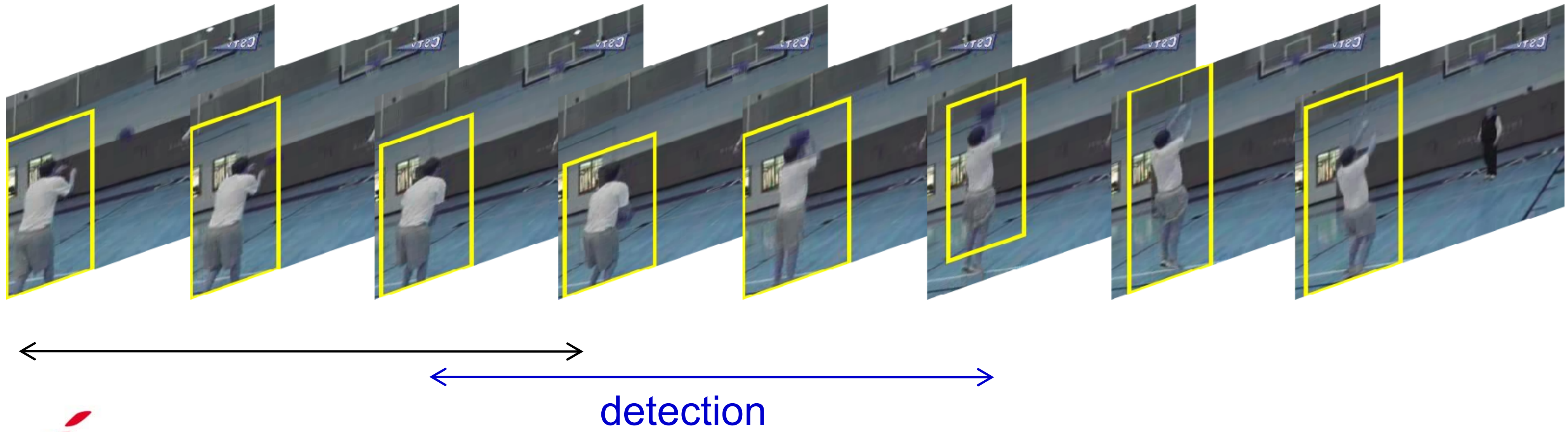
instance-level → sufficiently specific

Rescoring and temporal sliding window

- To capture the dynamics
  - ► Dense trajectories

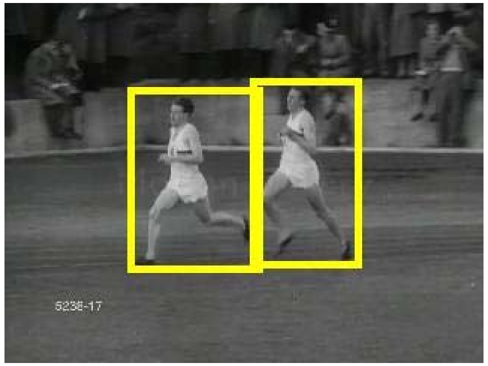
- Temporal sliding window



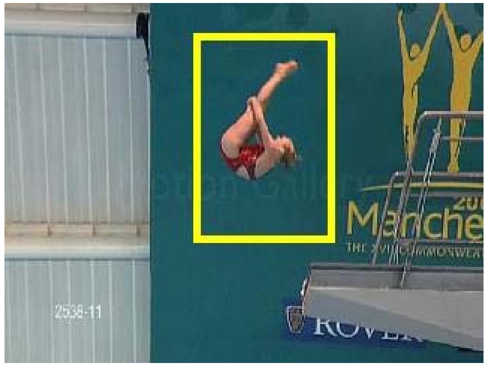


Datasets (spatial localization)

	UCF-Sports [Rodriguez et al. 2008]	J-HMDB [Jhuang et al. 2013]
Number of videos	150	928
Number of classes	10	21
Average length	63 frames	34 frames

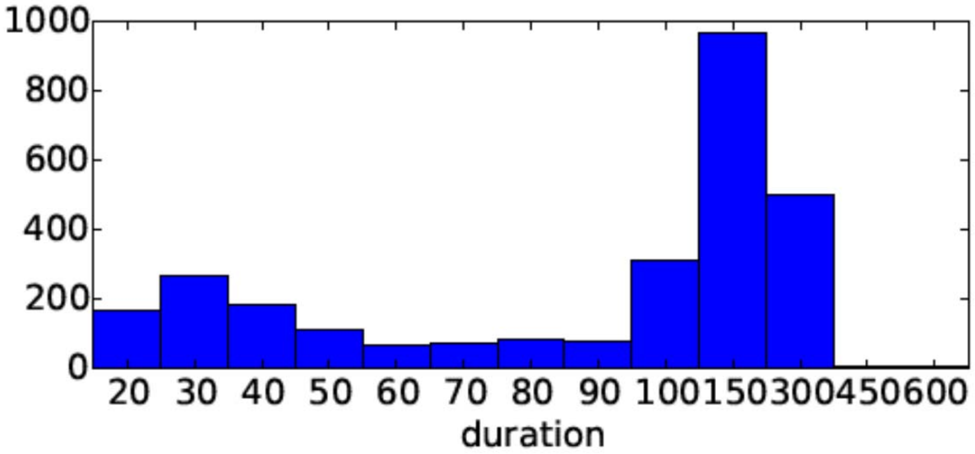



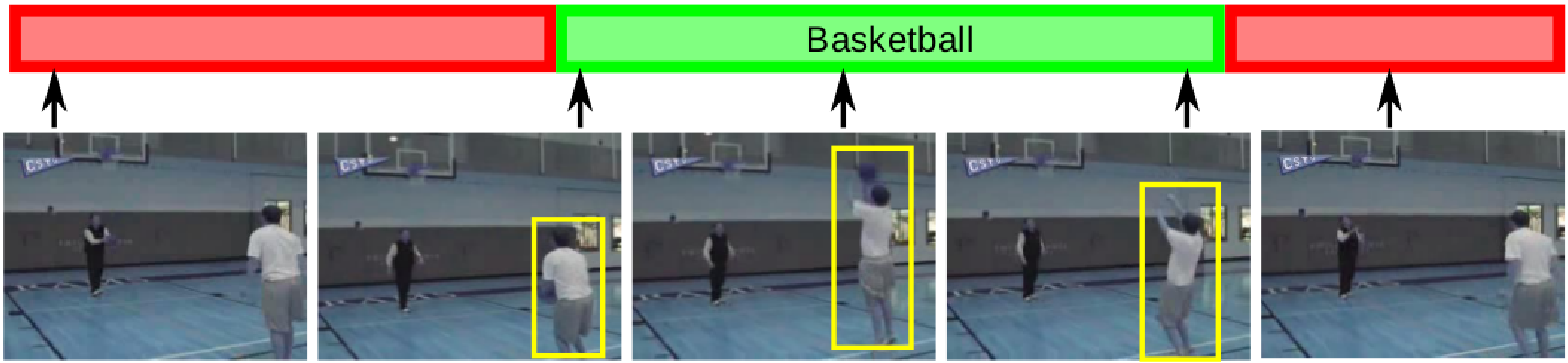





Datasets

- UCF-101 [Soomro et al. 2012]
  - ► Spatio-temporal localization for a subset of the dataset
  - ► 3207 videos, 24 classes
  - ► Average length: 176 frames





Results

Impact of the tracker

Detectors in the tracker	mAP	
	UCF-Sports	J-HMDB
instance-level + class-level	**90.50%**	**59.74%**
instance-level	74.27%	54.32%
class-level	85.67%	53.25%

Comparison to SOA on UCF-Sports

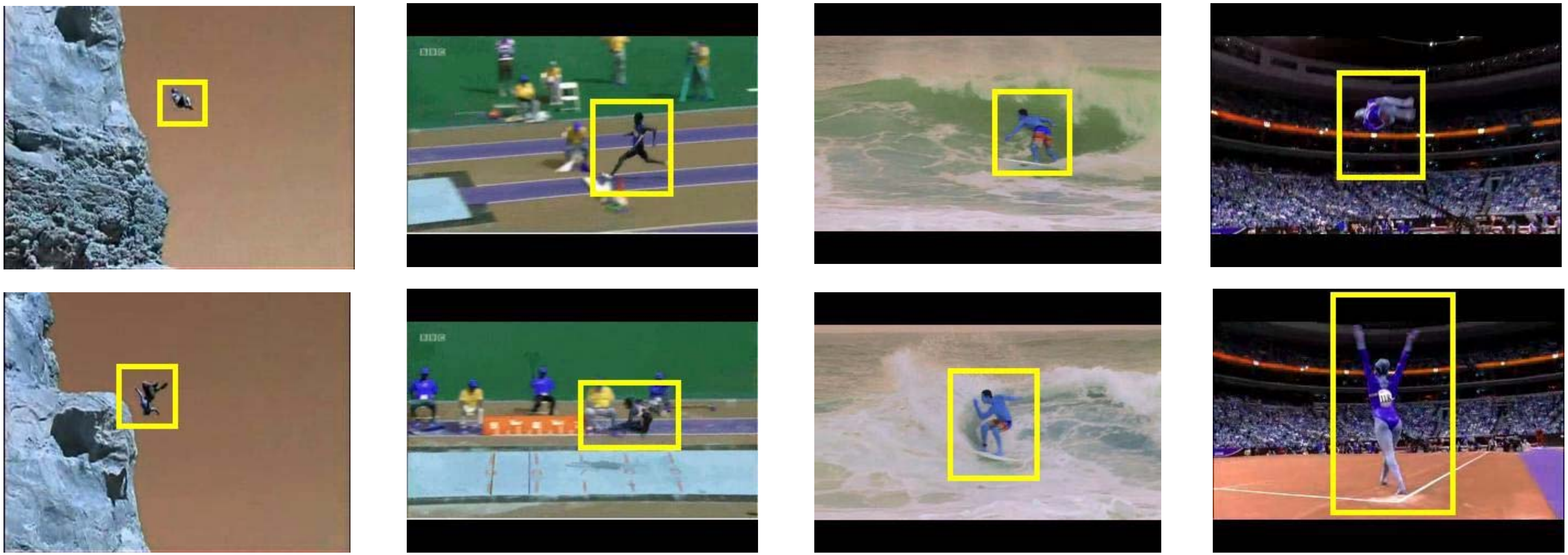
mAP	0.5
Gkioxari and Malik 2015	75.8
Ours	90.5

Comparison to SOA on J-HMDB

mAP	0.5
Gkioxari and Malik 2015	53.3
Ours	59.7

Quantitative evaluation (UCF-101)

mAP	0.05	0.2	0.3
Yu and Yuan'15	42.8		
Ours	54.28	46.7	37.8



Spatio-temporal action localization

Spatio-temporal video tubes

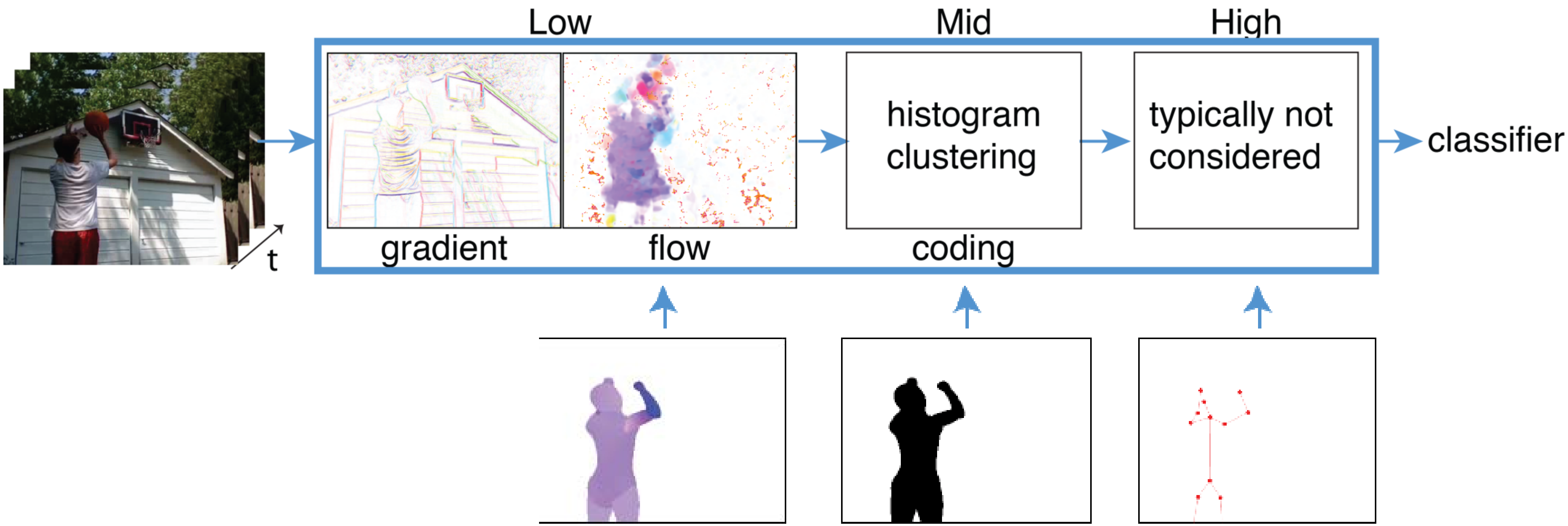
- Brox and Malik, Object segmentation by long term analysis of point trajectories, ECCV'10
- Oneata et al., Spatio-temporal object detection proposals, ECCV'14
- Gemert et al., Action localization proposals from dense trajectories, BMVC'15
- Yu and Yuan, Fast action proposals for human action detection and search, CVPR'15

Human pose estimation + action recognition

- Estimation of body joints in video

Poses in the wild dataset [Cherian'14]   Pose results [Pfister'15]

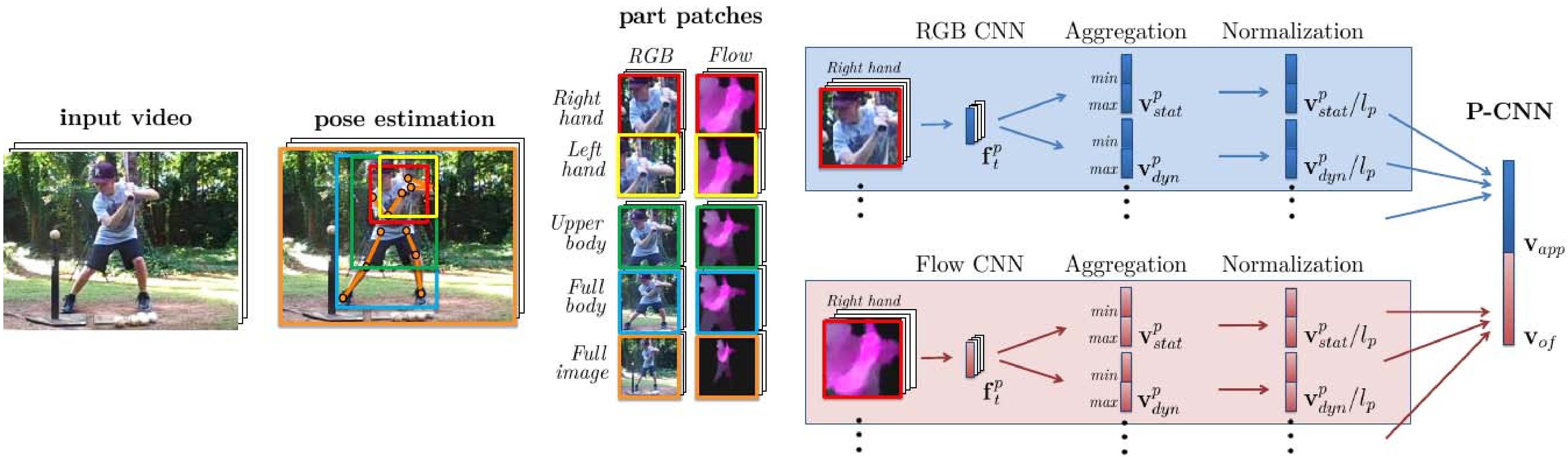
Potential impact of human pose on action classification



- Systematically replace steps of "dense trajectories" with ground truth
- Ground-truth annotations for a subset of HMDB (Joint-HMDB)
- Pose features (joint position and spatio-temporal relations) results in a significant improvement

[H. Jhuang et al.'13]

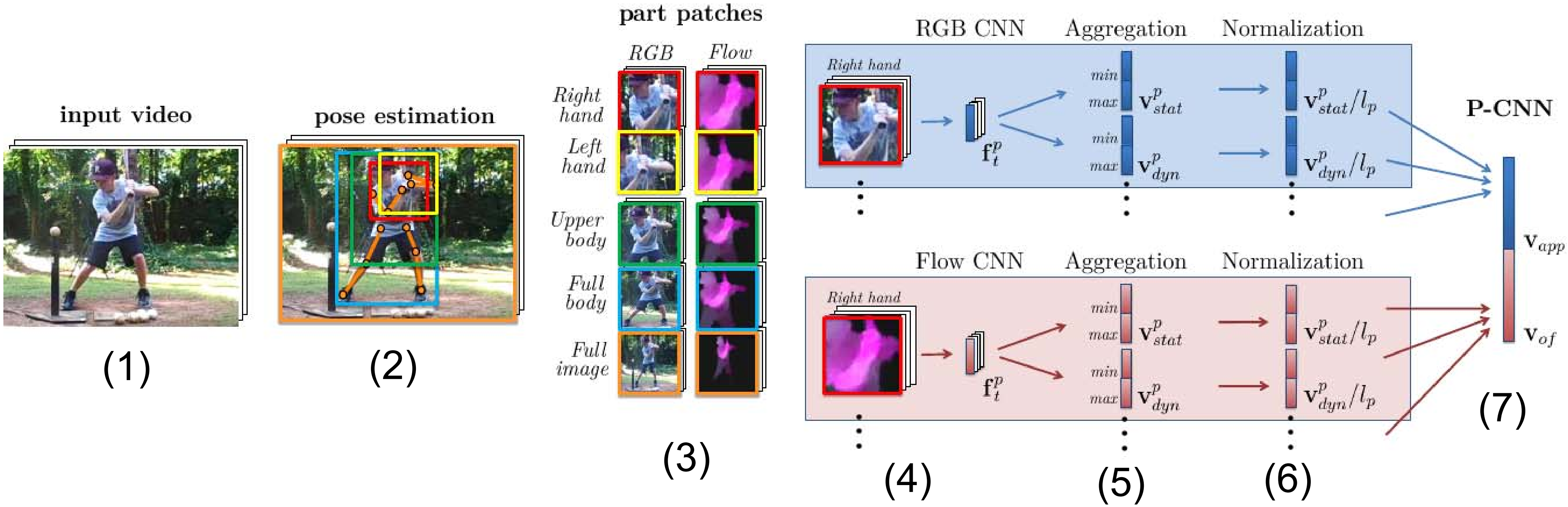
Robust pose features – Pose-CNN



- Track human pose in a video → body part track
- Extract CNN features (appearance and motion) per part-track
- Train SVM classifier

[P-CNN, pose-based CNN features for action recognition, G. Cheron, I. Laptev, C. Schmid, ICCV'15]

Pose-CNN (P-CNN)



1) input video
2) video pose estimation [Cherian'14]
3) crop human body parts
4) extract CNN features (appearance and motion) per part and per frame
5) video descriptors: aggregation of frame features (max/min)
6) P-CNN: concatenation of part features from appearance and flow

Datasets used for evaluation

- JHMB as described previously
- MPI cooking
  - 64 fine grained actions
  - a total of 5609 clips, 7 training/test splits
  - similar action, i.e. cut dice, cut slices, and cut stripes
- Sub-MPI
  - selection of two similar classes
  - wash hands and wash objects with GT pose

Performance of the individual features

	JHMDB-GT			MPII Cooking-Pose [8]		
Parts	App	OF	App + OF	App	OF	App + OF
Hands	46.3	54.9	57.9	39.9	46.9	51.9
Upper body	52.8	60.9	67.1	32.3	47.6	50.1
Full body	52.2	61.6	66.1	-	-	-
Full image	43.3	55.7	61.0	28.8	56.2	56.5
All	**60.4**	**69.1**	**73.4**	**43.6**	**57.4**	**60.8**

- Different body parts are complementary
- Appearance and flow are complementary

Robustness of P-CNN

	JHMDB		
	GT	Pose [7]	Diff
P-CNN	74.6	61.1	13.5
HLPF	77.8	25.3	52.5

	sub-MPII Cooking		
	GT	Pose [7]	Diff
P-CNN	83.6	67.5	16.1
HLPF	76.2	57.4	18.8

	MPII Cooking
	Pose [7]
P-CNN	62.3
HLPF	32.6

- P-CNN on par with HLPF for GT
- P-CNN significantly more robust for real noisy poses

Comparison to state of the art

	JHMDB		MPII Cook.
Method	GT	Pose [7]	Pose [7]
P-CNN	74.6	61.1	62.3
DT-FV	65.9	65.9	67.6
P-CNN + DT-FV	**79.5**	**72.2**	**71.4**

- P-CNN better than IDT on ground-truth
- P-CNN and IDT are complementary

Where to get training data?

➡ **Weakly-supervised learning**

Actions in movies

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






- Typically only a few class-samples per movie
- Manual annotation is <u>very</u> time consuming

Script-based video annotation

- 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

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.
…

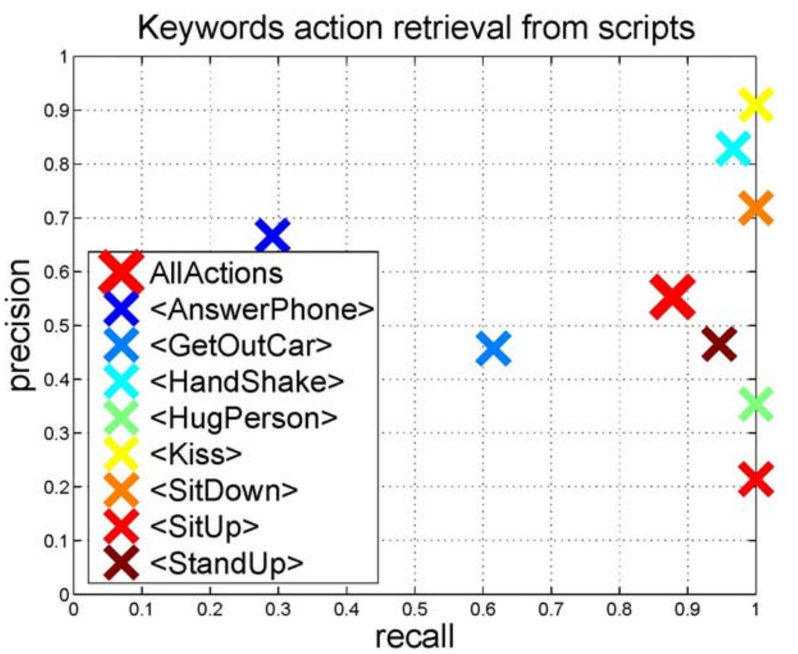
[Laptev, Marszałek, Schmid, Rozenfeld 2008]

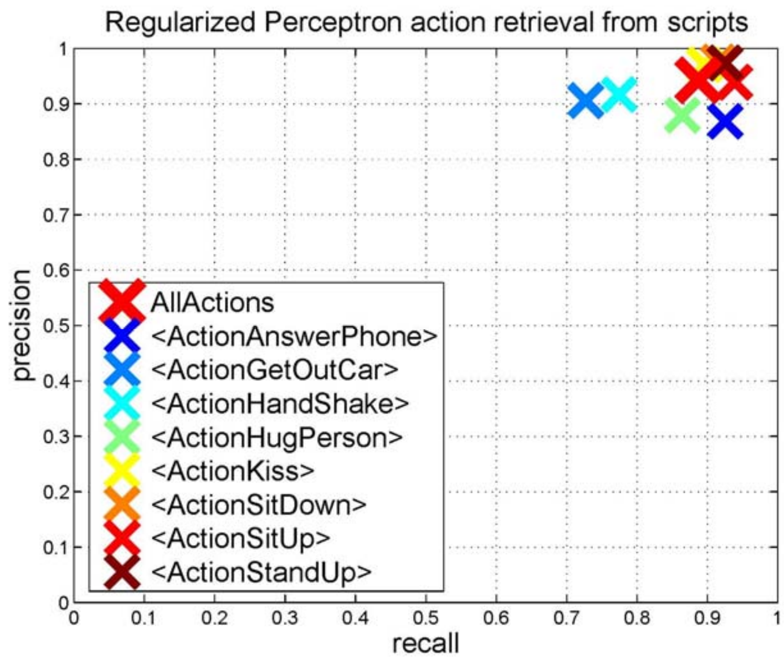
Text-based action retrieval

- Large variation of action expressions in text:

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

- => Supervised text classification approach





[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 results

	Clean		Automatic		
	hoghof		hoghof		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

Scripts as weak supervision

Challenges:

- Imprecise temporal localization
- No explicit spatial localization
- NLP problems, scripts ≠ training labels

"… Will gets out of the Chevrolet. …"
"… Erin exits her new truck…"

vs. Get-out-car