Action recognition in videos

Cordelia Schmid

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

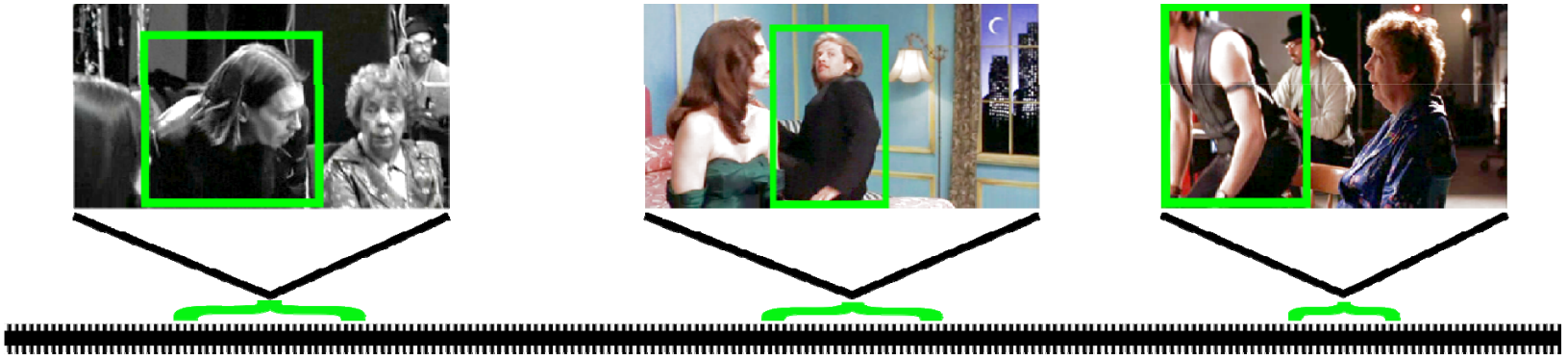
Making sandwich: present
Feeding animal: not present
...

Action recognition - tasks

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

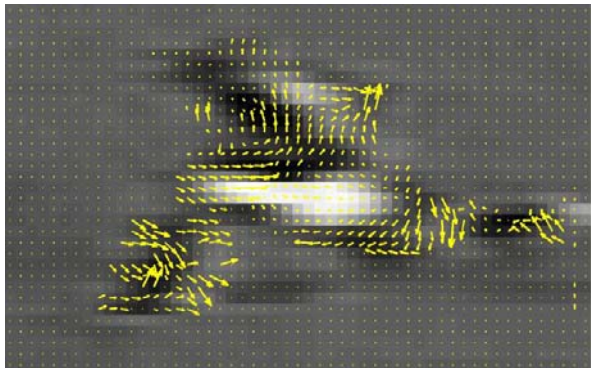
Making sandwich: present
Feeding animal: not present
...

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

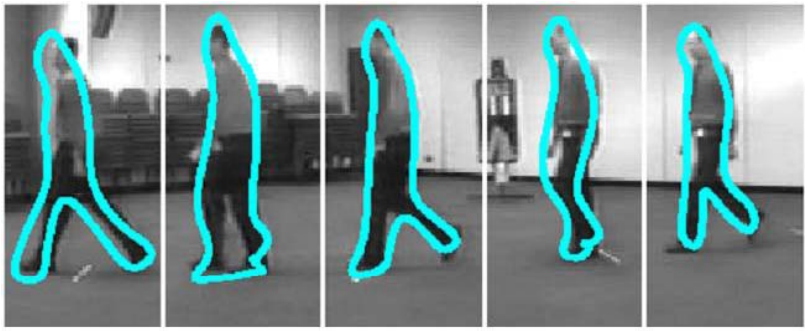


State of the art in action recognition

Motion history image
[Bobick & Davis, 2001]



Spatial motion descriptor
[Efros et al. ICCV 2003]



Learning dynamic prior
[Blake et al. 1998]

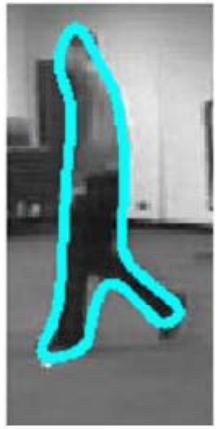
Sign language recognition
[Zisserman et al. 2009]

Advantages/disadvantages

Temporal templates:

+ simple, fast

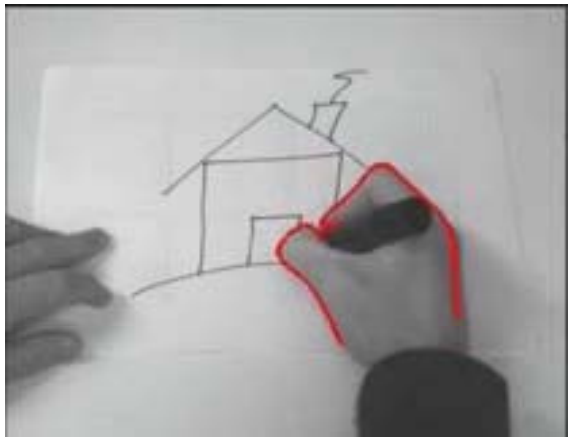
- sensitive to segmentation errors



Active shape models:

+ shape regularization

- sensitive to initialization and tracking failures



Tracking with motion priors:

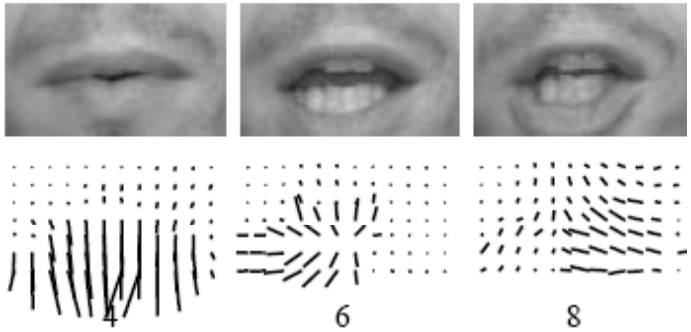
+ improved tracking and simultaneous action recognition

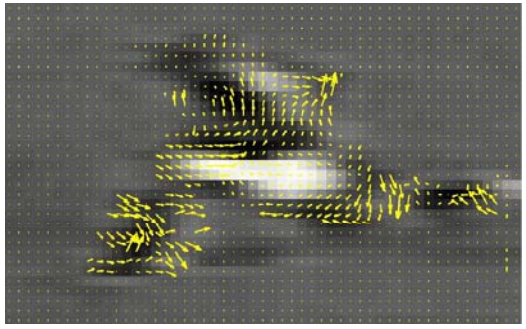
- sensitive to initialization and tracking failures

Motion-based recognition:

+ generic descriptors; less depends on appearance

- sensitive to localization/tracking errors



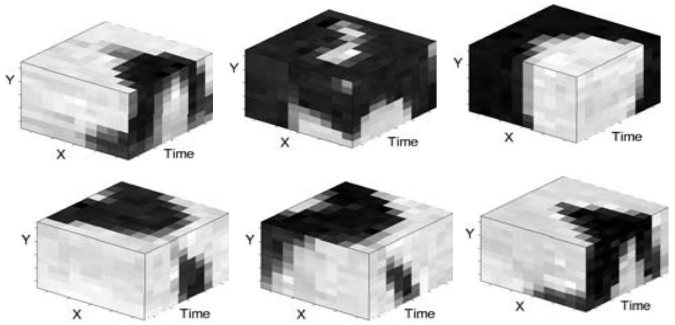


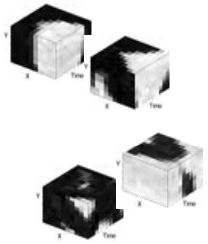
State of the art in action recognition

- Bag of space-time features [Laptev'03, Schuldt'04, Niebles'06, Zhang'07]

Extraction of space-time features

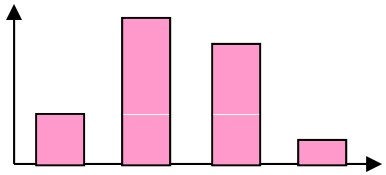
Collection of space-time patches





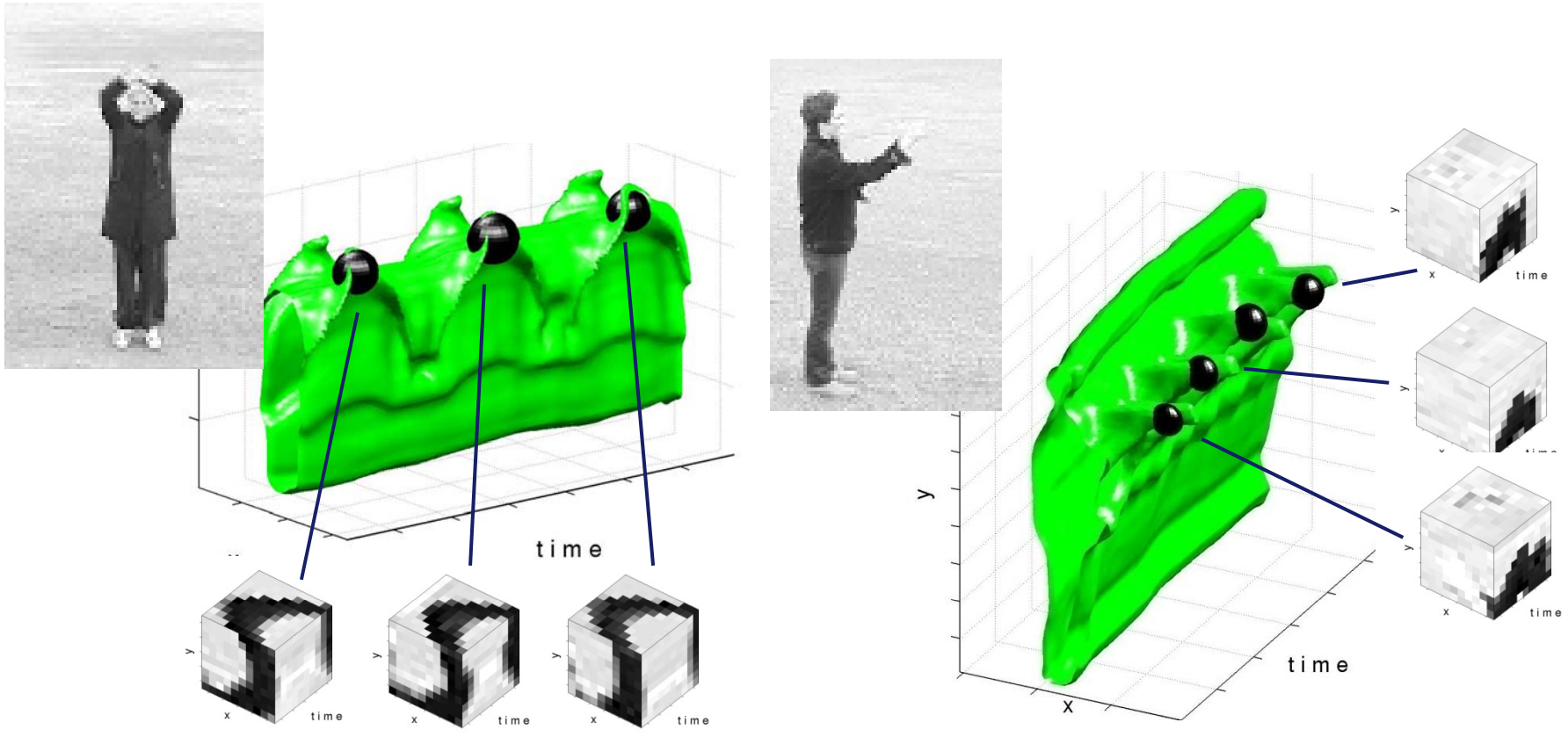
HOG & HOF
patch descriptors

Histogram of visual words



SVM classifier

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

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

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

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

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

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

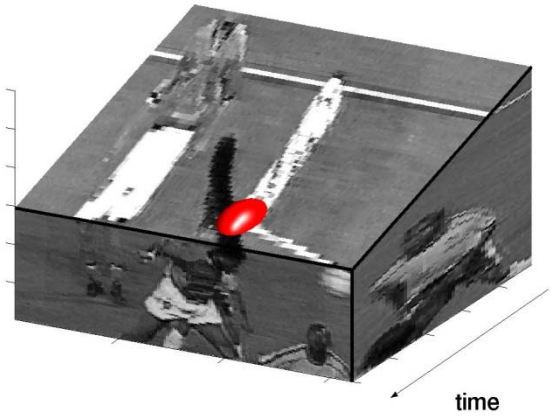
(similar to Harris operator [Harris and Stephens, 1988])

Space-time features

- Detector [Laptev'05]

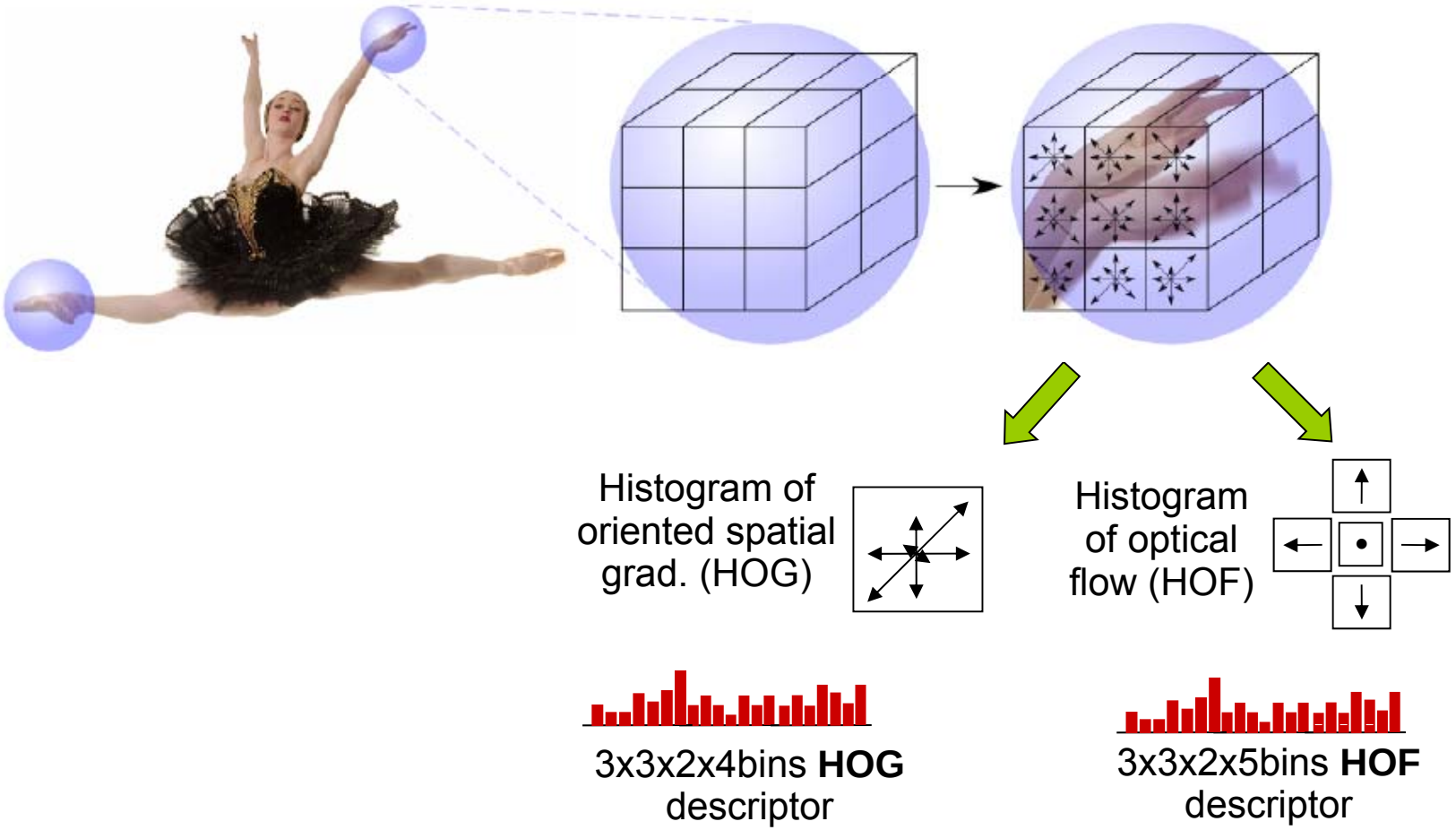
$$H = \det(\mu) + k\,\mathrm{tr}^3(\mu)$$

$$\mu = \begin{pmatrix} I_x I_x & I_x I_y & I_x I_t \\ I_x I_y & I_y I_y & I_y I_t \\ I_x I_t & I_y I_t & I_t I_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$



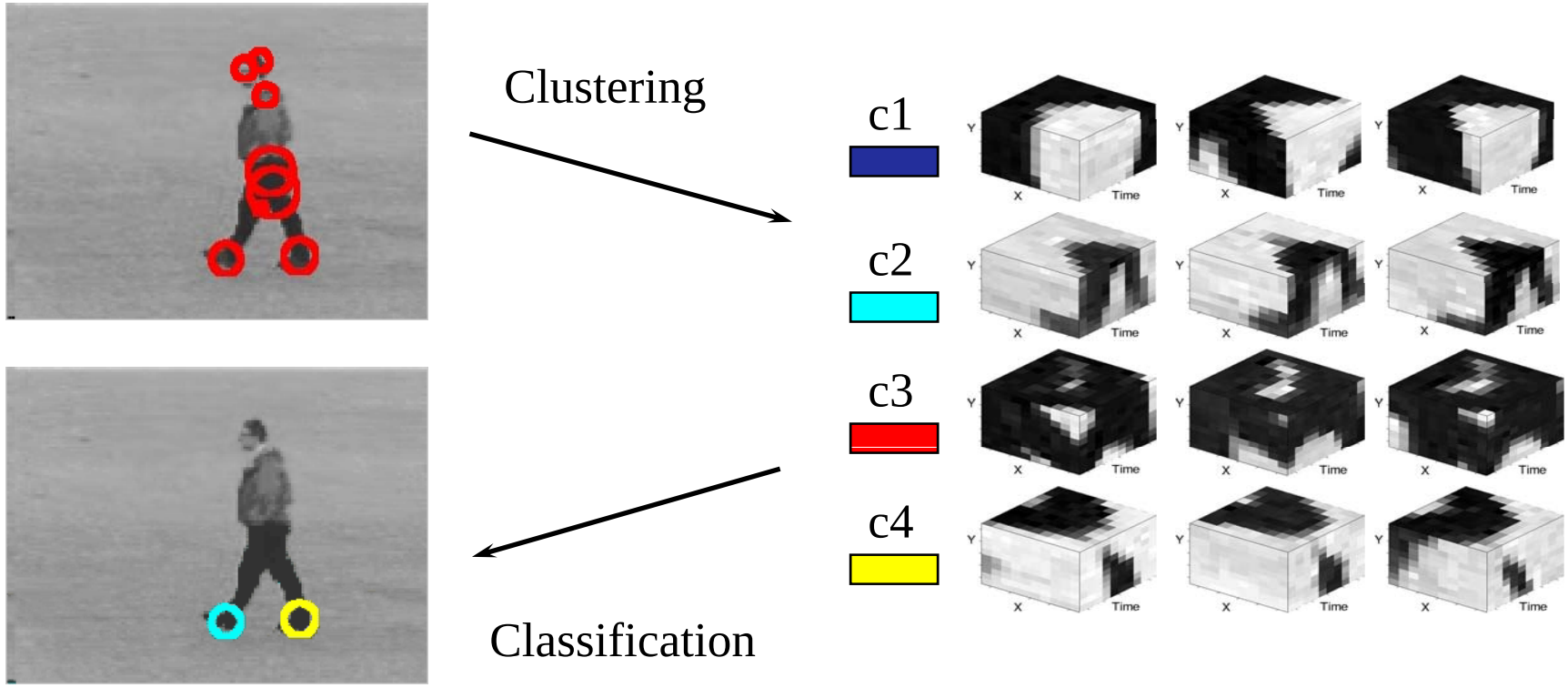
Space-time features

- Descriptors: HOG / HOF



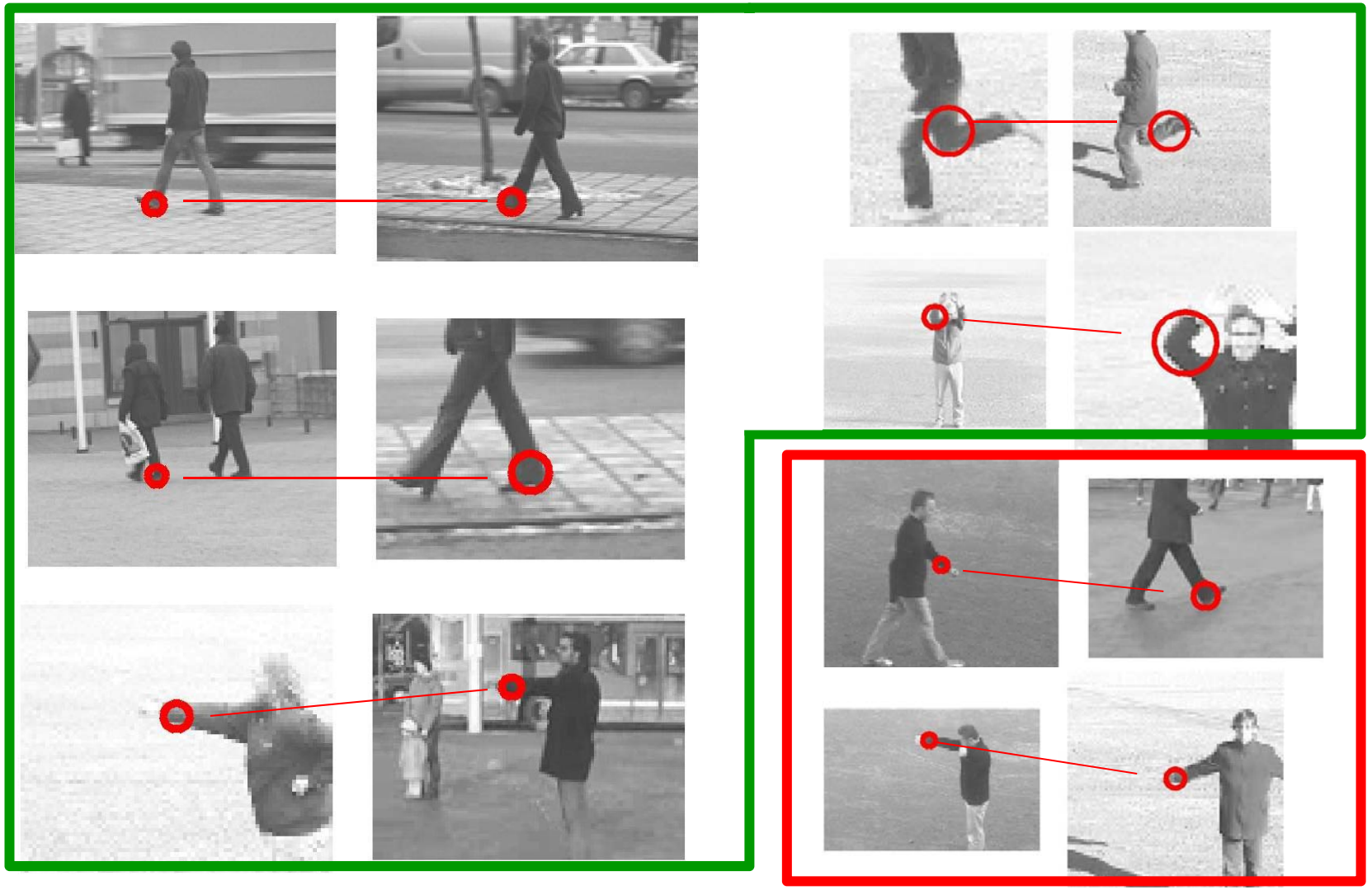
Visual Vocabulary: K-means clustering

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



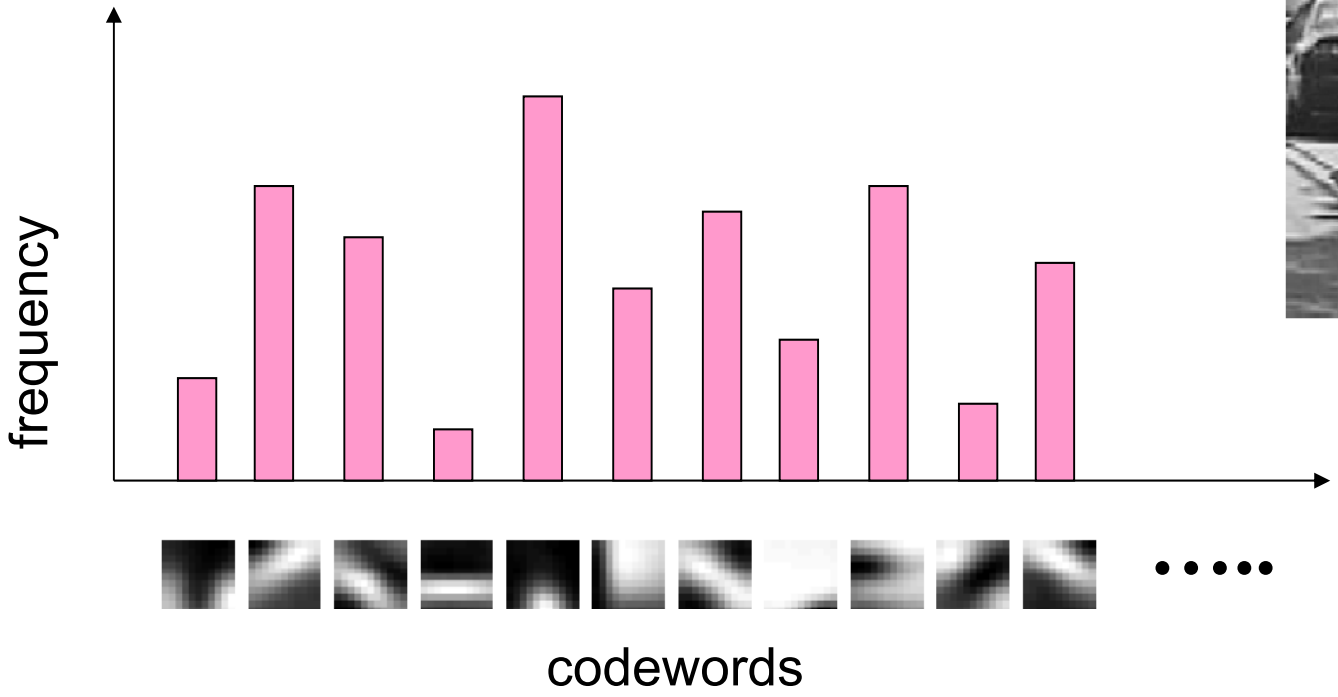
Local features: Matching

- Finds similar events in pairs of video sequences



Bag of features

- Cluster descriptors with k-means (~4000 clusters)
- Assign each descriptor to the closest center
- Measure frequency



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”

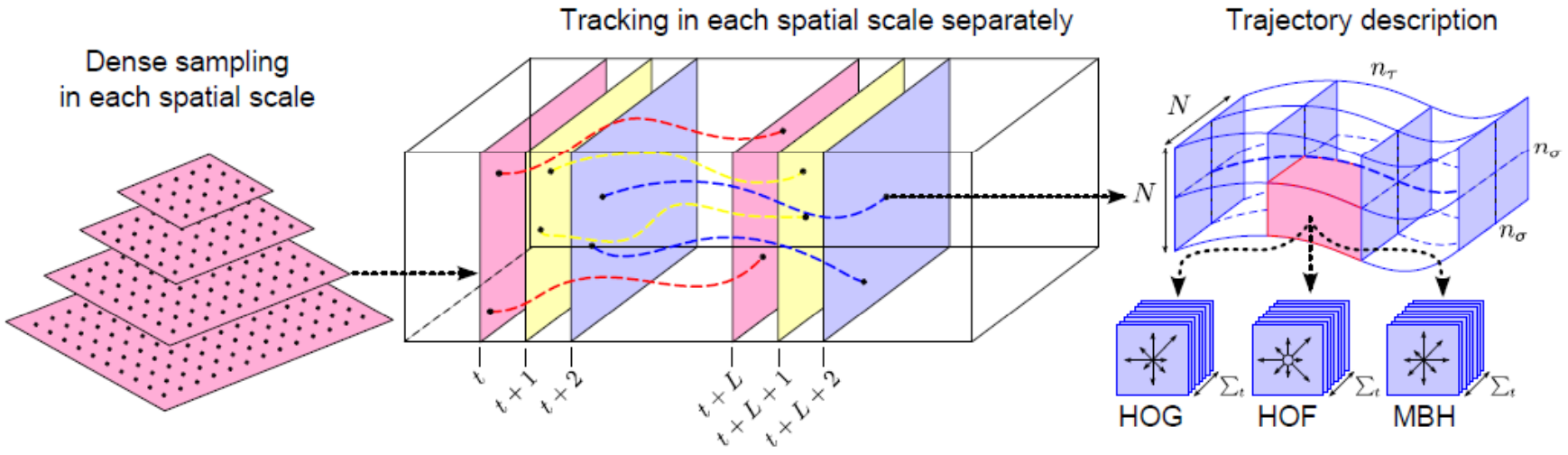
Improved descriptors: Dense trajectories

- Dense sampling improves results over sparse interest points for image classification [Fei-Fei'05, Nowak'06]
- Recent progress by using feature trajectories for action recognition [Messing'09, Sun'09]
- The 2D space domain and 1D time domain in videos have very different characteristics

➔ Dense trajectories: a combination of dense sampling with feature trajectories [Wang, Klaeser, Schmid & Lui, CVPR'11]

Approach

- Dense multi-scale sampling
- Feature tracking over L frames with optical flow
- Trajectory-aligned descriptors with a spatio-temporal grid



Approach

Dense sampling

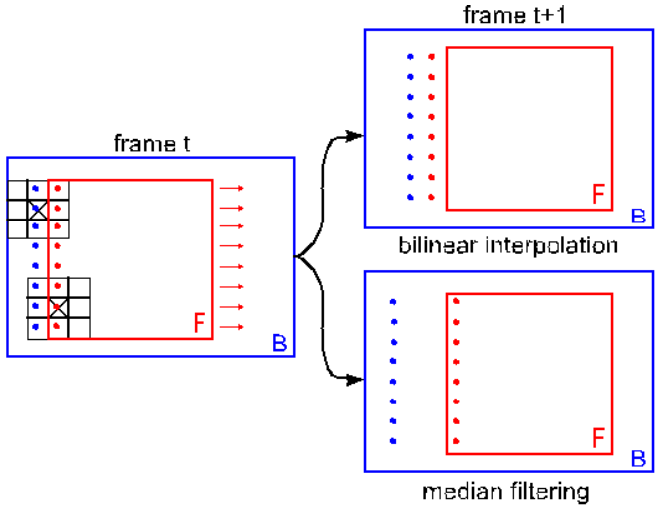
- remove untrackable points
- based on the eigenvalues of the auto-correlation matrix

Feature tracking

- by median filtering in dense optical flow field

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega_t)|_{(\bar{x}_t, \bar{y}_t)}$$

- length is limited to avoid drifting



Feature tracking

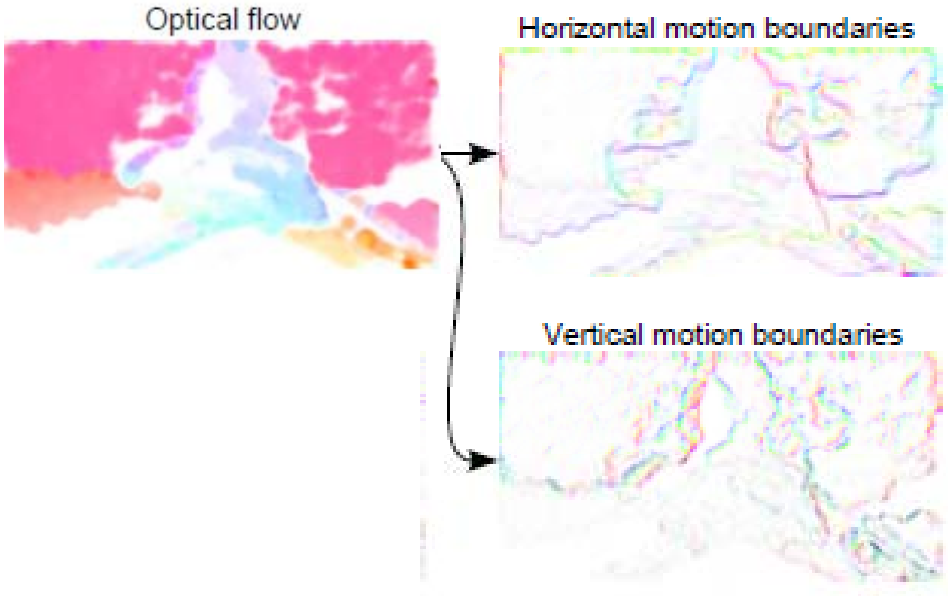
KLT tracks

SIFT tracks

Dense tracks

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 as appears for example due to background camera motion

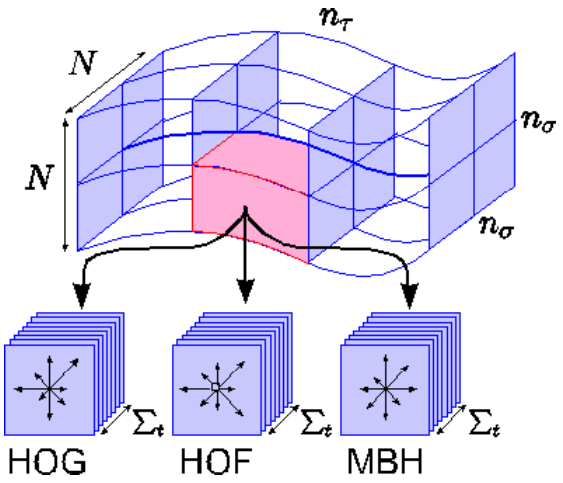


Trajectory descriptors

- Trajectory shape described by normalized relative point coordinates

$$S = \frac{(\Delta P_t, \ldots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} ||\Delta P_j||}$$

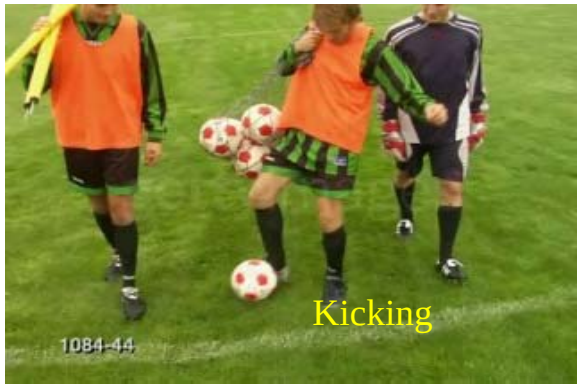
- HOG, HOF and MBH are encoded along each trajectory



Experimental setup

- Bag-of-features with 4000 clusters obtained by k-means, classification by non-linear SVM with RBF + chi-square kernel
  - Ialso possible to use Fisher vector + linear SVM
- Descriptors are combined by addition of distances
- Evaluation on two datasets: UCFSport (classification accuracy) and Hollywood2 (mean average precision)
- Two baseline trajectories: KLT and SIFT

UCF Sports



10 action classes, videos from TV broadcasts

Comparison of descriptors

	Hollywood2	UCFSports
Trajectory	47.8%	75.4%
HOG	41.2%	84.3%
HOF	50.3%	76.8%
MBH	55.1%	84.2%
Combined	58.2%	88.0%

- Trajectory descriptor performs well
- HOF >> HOG for Hollywood2, dynamic information is relevant
- HOG >> HOF for sports datasets, spatial context is relevant
- MBH consistently outperforms HOF, robust to camera motion

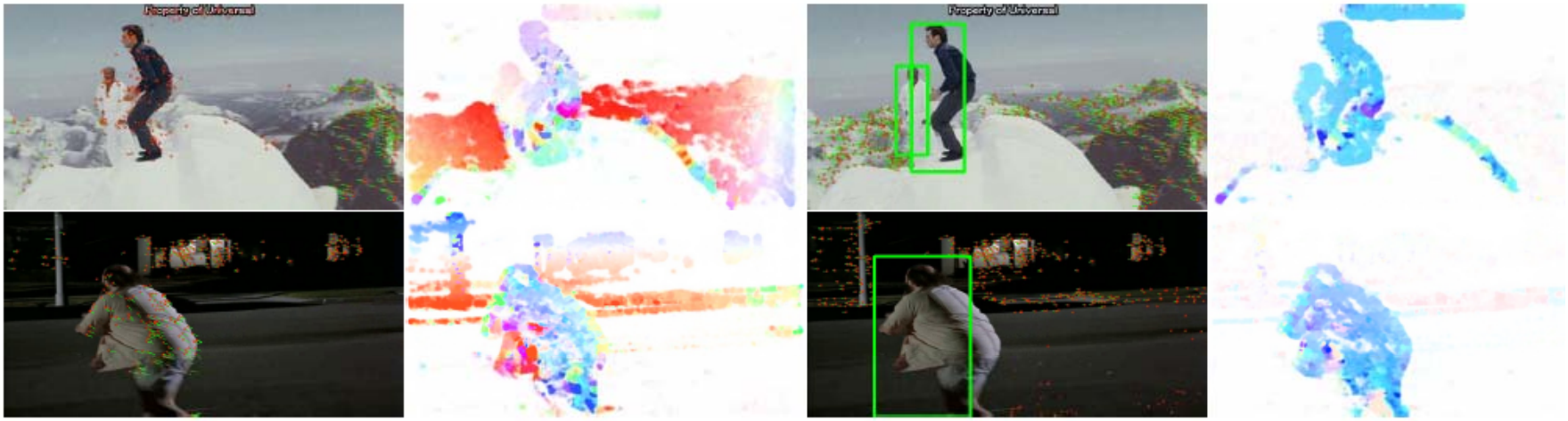
Comparison of trajectories

	Hollywood2	UCFSports
Dense trajectory + MBH	55.1%	84.2%
KLT trajectory + MBH	48.6%	78.4%
SIFT trajectory + MBH	40.6%	72.1%

- Dense >> KLT >> SIFT trajectories

Improved trajectories (Wang & Schmid ICCV'13)

- Dense trajectories impacted by camera motion
  - Stabilize camera motion before computing optical flow
  - Use human detector and robust homography estimation
  - Wrap optical flow and remove background trajectories



Inlier feature matches and warped flow, without or with HD

student presentation

Results

Datasets	Bag of features		Fisher vector	
	DTF	ITF	DTF	ITF
Hollywood2	58.5%	62.2%	60.1%	64.3%
HMDB51	47.2%	52.1%	52.2%	57.2%
Olympic Sport	75.4%	83.3%	84.7%	91.1%
UCF50	84.8%	87.2%	88.6%	91.2%

Compare DTF and ITF using different feature encoding

- Standard bag of features: train a codebook of 4000 visual words with k-means for each descriptor; RBF-χ^2 kernel SVM for classification
- We observe similar improvement of ITF over DTF when using BOF or FV for feature encoding
- The improvement of FV over BOF varies on different datasets, from 2% to 7%

Results

Hollywood2		HMDB51		Olympic Sports		UCF50	
Jain CVPR'13	62.5%	Jain CVPR'13	52.1%	Jain CVPR'13	83.2%	Shi CVPR'13	83.3%
With HD	64.3%	With HD	57.2%	With HD	91.1%	With HD	91.2%
Without HD	63.0%	Without HD	55.9%	Without HD	90.2%	Without HD	90.5%

HD stands for human detection

- Human detection always helps. For Hollywood2 and HMDB51, it's more significant as they are more dominated by humans
- Significantly outperforms the state of the art on all four datasets,

Excellent results in TrecVid MED'13

- Combination of MBH SIFT, audio, text & speech recognition
- First in the know event challenge, first in the adhoc event challenge

Making sandwich – results

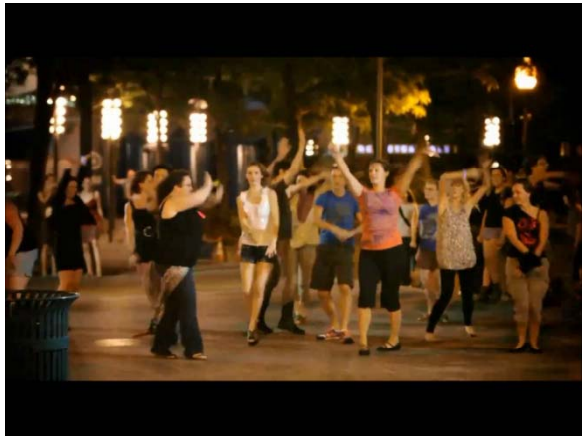
Rank 1 (pos)

Rank 20 (pos)

Rank 21 (neg)

Excellent results in TrecVid MED'13

FlashMob gathering – results



Rank 1 (pos)

Rank 18 (pos)

Rank 19 (neg)

Impact of different channels

		Birthday party	Changing a vehicle tire	Flash mob gathering	Getting vehicle unstuck	Grooming an animal	Making a sandwich	Parade	Parkour	Repairing an appliance	Sewing project	Mean
AP	MBH	20.40	15.50	54.76	30.43	18.33	13.10	41.21	71.03	34.56	29.15	32.84
	SIFT	23.10	28.88	48.49	31.76	17.12	17.09	30.27	37.50	33.20	22.95	29.04
	MBH+SIFT	27.37	34.59	61.94	**41.64**	21.37	20.21	47.88	71.43	43.55	34.57	40.45
	MBH+SIFT+MFCC	**45.33**	**41.77**	**63.90**	39.16	**27.24**	**21.64**	**53.22**	**71.91**	**50.85**	**38.20**	**45.32**

Conclusion

- Dense trajectory representation for action recognition outperforms existing approaches
- Motion boundary histogram descriptors perform very well, they are robust to camera motion
- Motion stabilization improves results
- Software available on-line at https://lear.inrialpes.fr/software
- Recent excellent results in the TrecVID MED 2013 challenge

Outline

- Improved video description
  - Dense trajectories and motion-boundary descriptors
- *Adding temporal information to the bag of features*
  - *Actom sequence model for efficient action detection*
- Modeling human-object interaction

Adding temporal information to the BOF

- Model of the temporal structure of an action with a sequence of “action atoms” (actoms)
- Action atoms are action specific short key events, whose sequence is characteristic of the action

student presentation

Modeling human-object interaction

- Action recognition is person-centric

Movies

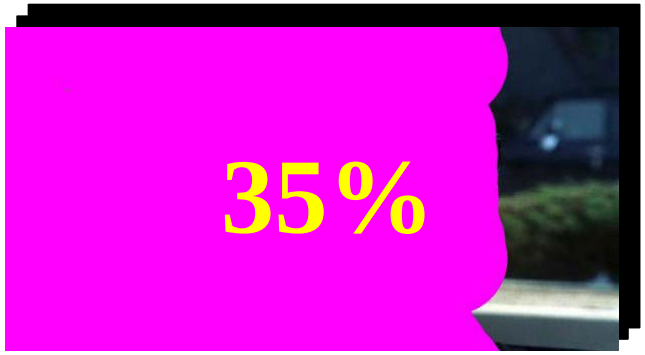
TV

YouTube

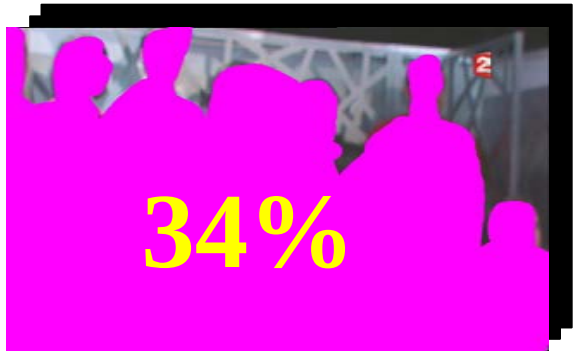
Source I.Laptev

Modeling human-object interaction

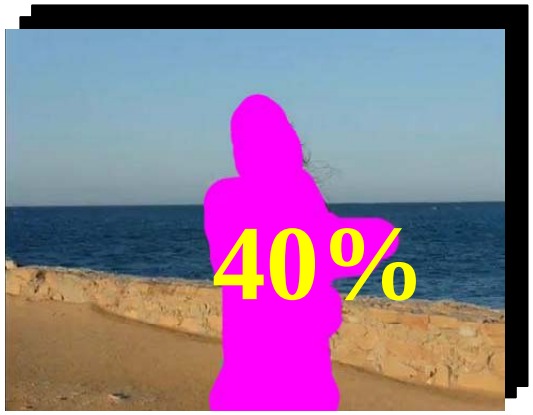
- Action recognition is person-centric



Movies



TV



YouTube

Source I.Laptev

Modeling human pose

- Description of the human pose
  - Silhouette description [Sullivan & Carlsson, 2002]
  - Histogram of gradients (HOG) [Dalal & Triggs 2005]

  - Human body part estimation [Felzenzswalb & Huttenlocher 2005]

Importance of action objects

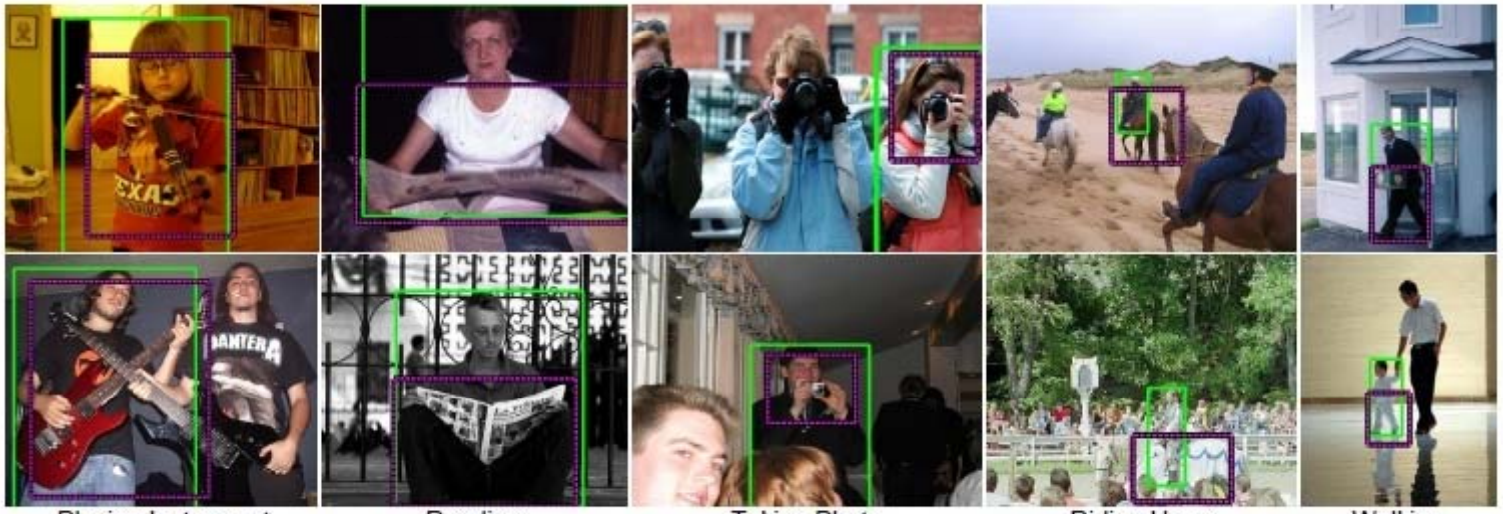



- Human pose often not sufficient by itself
- Objects define the actions

Action recognition from still images

- Supervised modeling interaction between human & object
  [Gupta et al. 2009, Yao & Fei-Fei 2009]

- Weakly-supervised learning of objects [Prest, Schmid & Ferrari 2011]



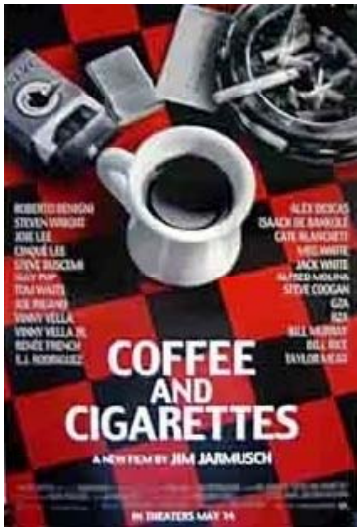
Results on PASCAL VOC 2010 Human action classification dataset

Importance of temporal information



- Video/temporal information necessary to disambiguate actions
- Temporal context describes the action/activity
- Key frames provide significant less information

Beyond BOF: Action localization

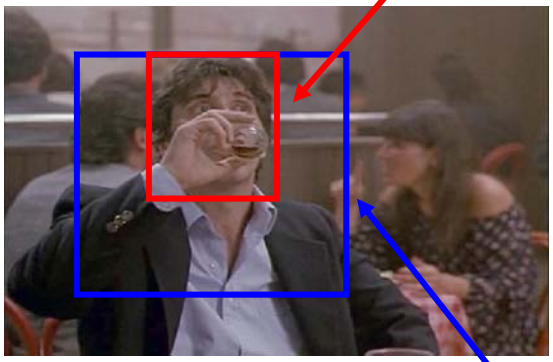


Manual annotation of drinking actions in movies:
"Coffee and Cigarettes"; "Sea of Love"

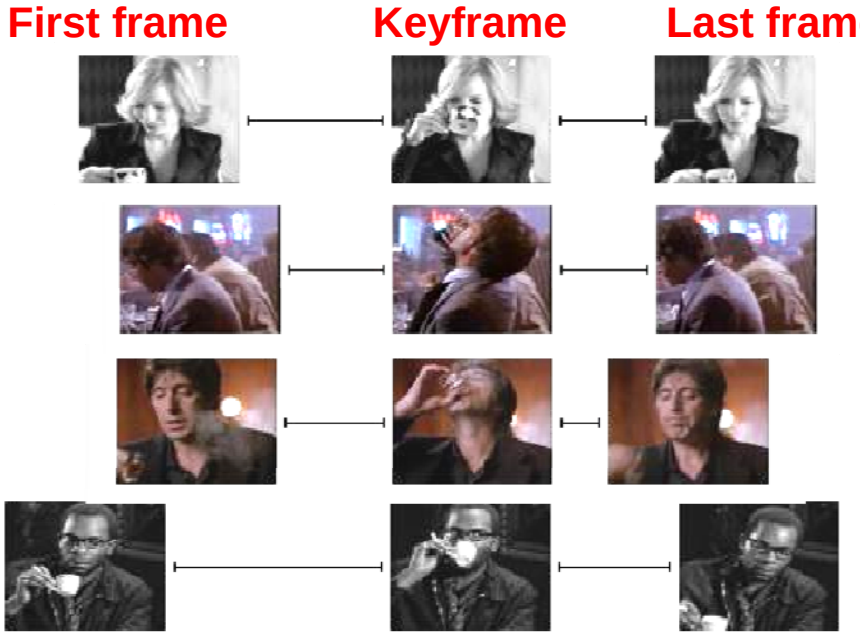
"*Drinking*": 159 annotated samples

"*Smoking*": 149 annotated samples

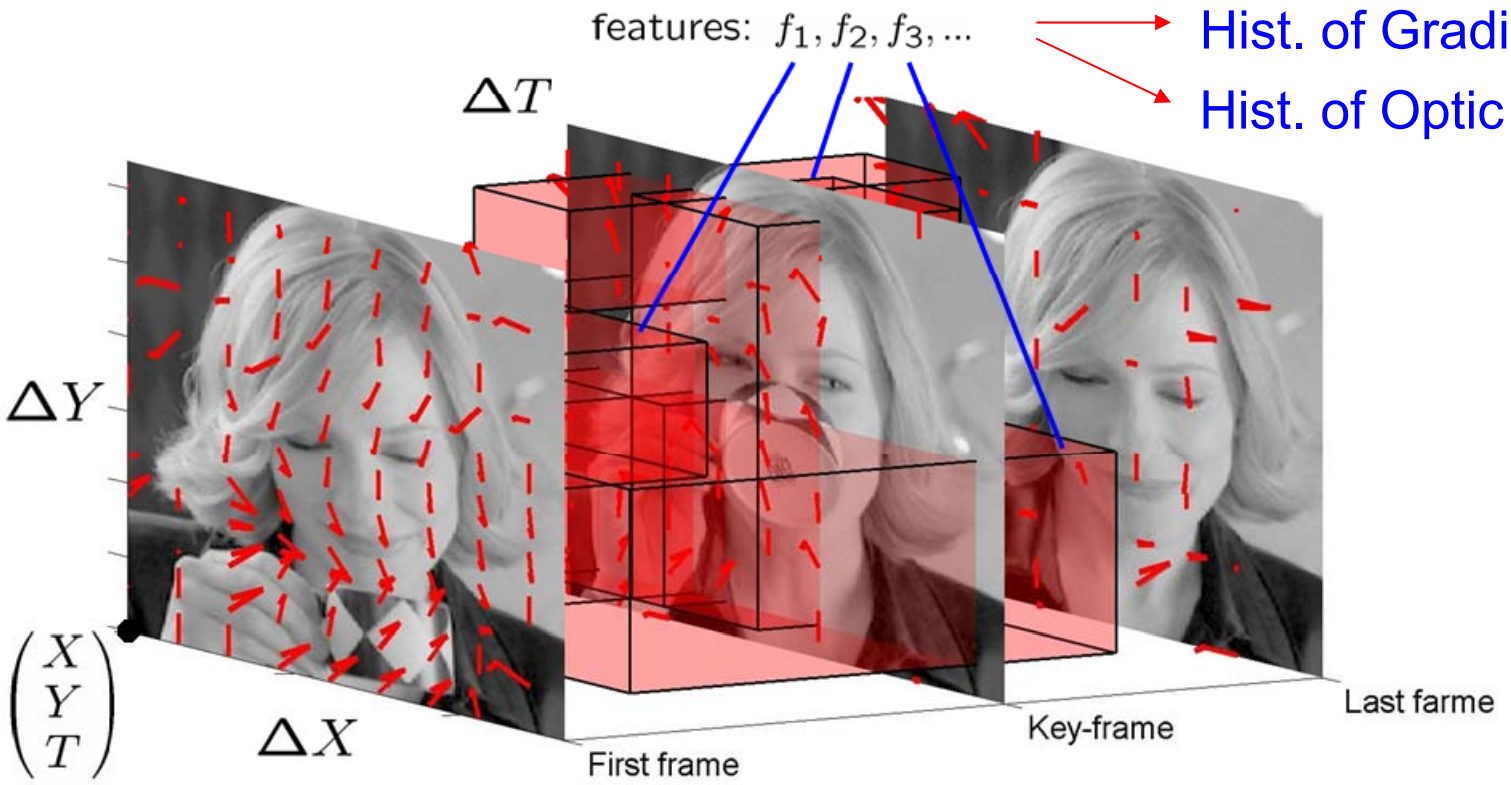
Spatial annotation

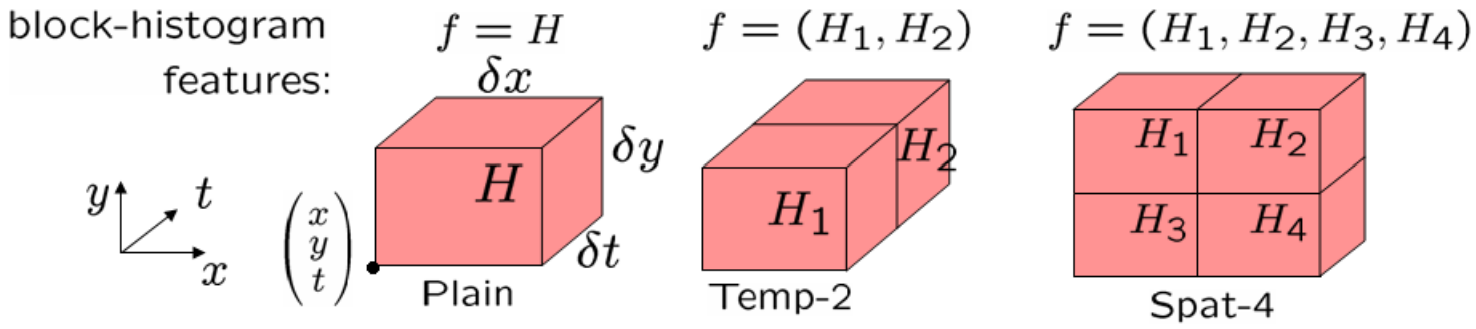


Temporal annotation

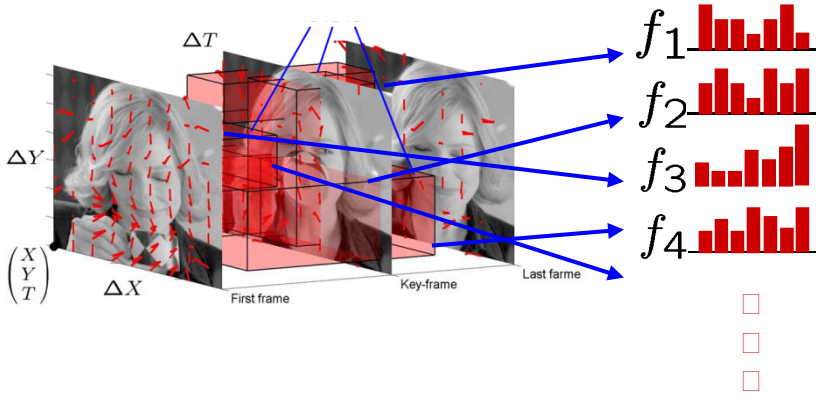


Action representation





Action learning



f_1 f_2 f_3 f_4 ⋯

boosting

selected features

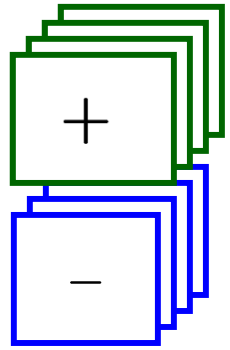
$$H(z) = \mathrm{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(f_t)\right)$$

weak classifier

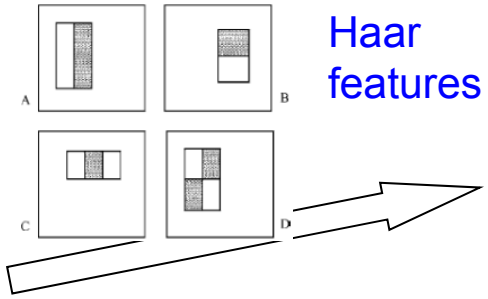
AdaBoost:

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

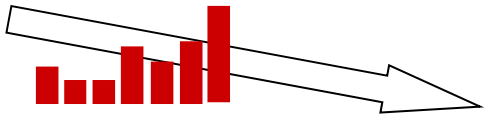
pre-aligned samples

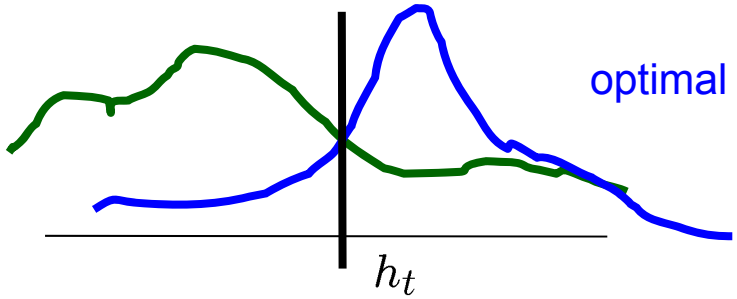


Haar features

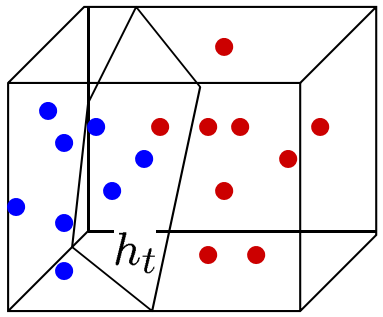


Histogram features





optimal threshold



Fisher discriminant

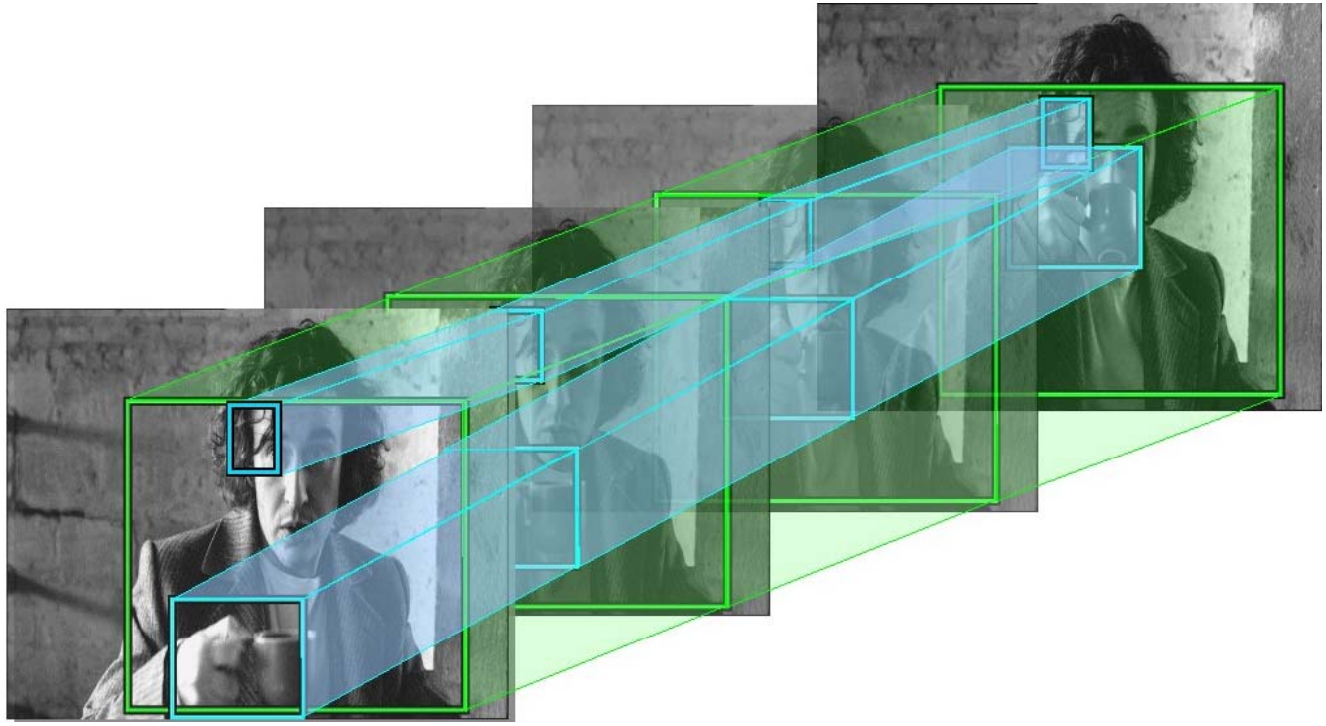
[Laptev, Perez 2007]

Modeling temporal human-object interactions

Describing human and object tracks and their relative motion

[Explicit modeling of human-object interactions in realistic videos,
A. Prest, V. Ferrari, C.Schmid, PAMI'13]

Tracking humans and objects



Fully automatic human tracks: state of the art detector + Brox tracks

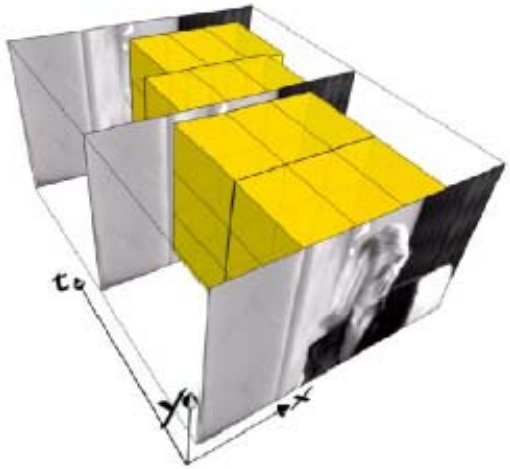
Object tracks: detector learnt from annotated training examples + Brox tracks

Extraction of a large number of human-object track pairs

Action descriptors

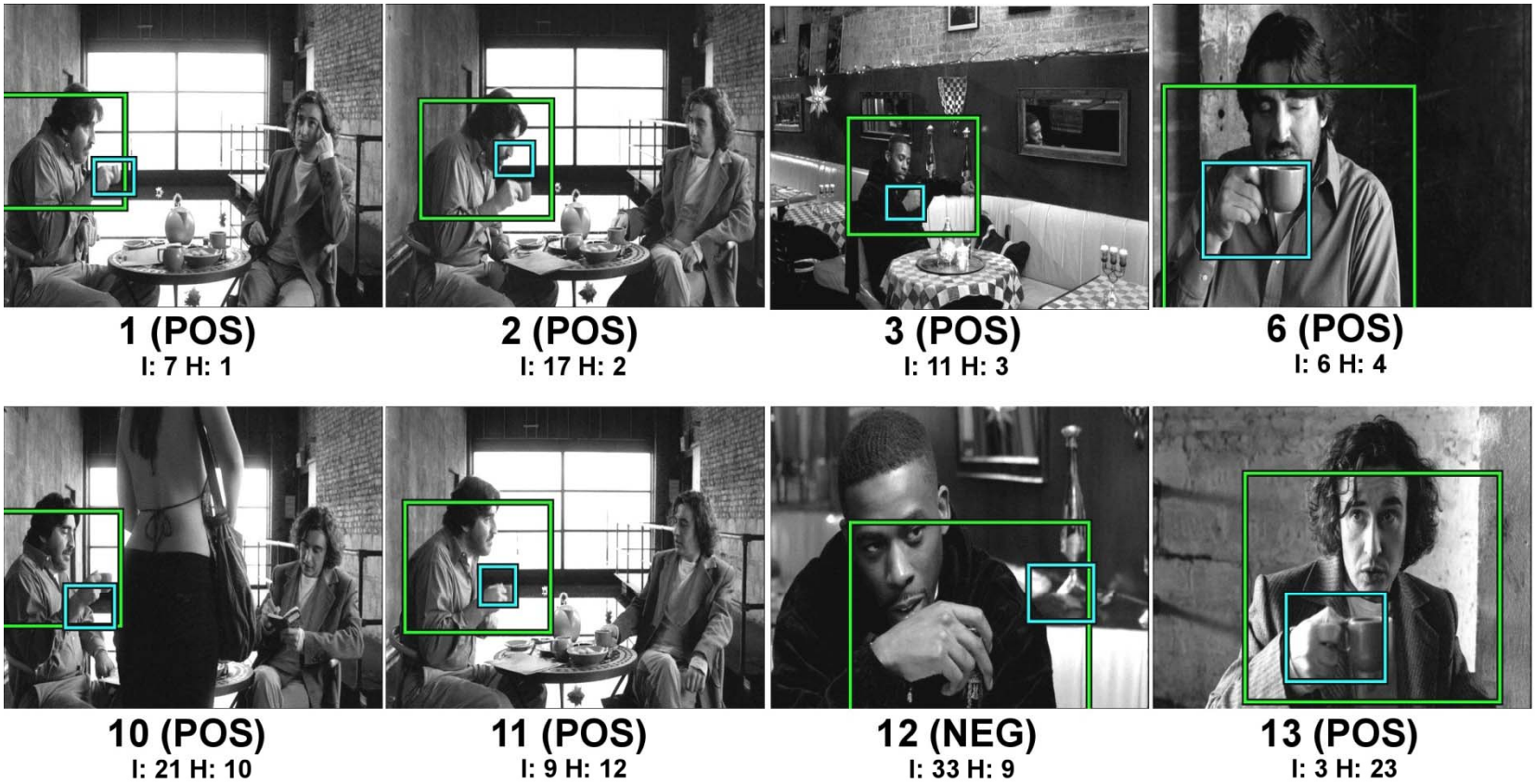
- Interaction descriptor: relative location, area and motion between human and object tracks

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



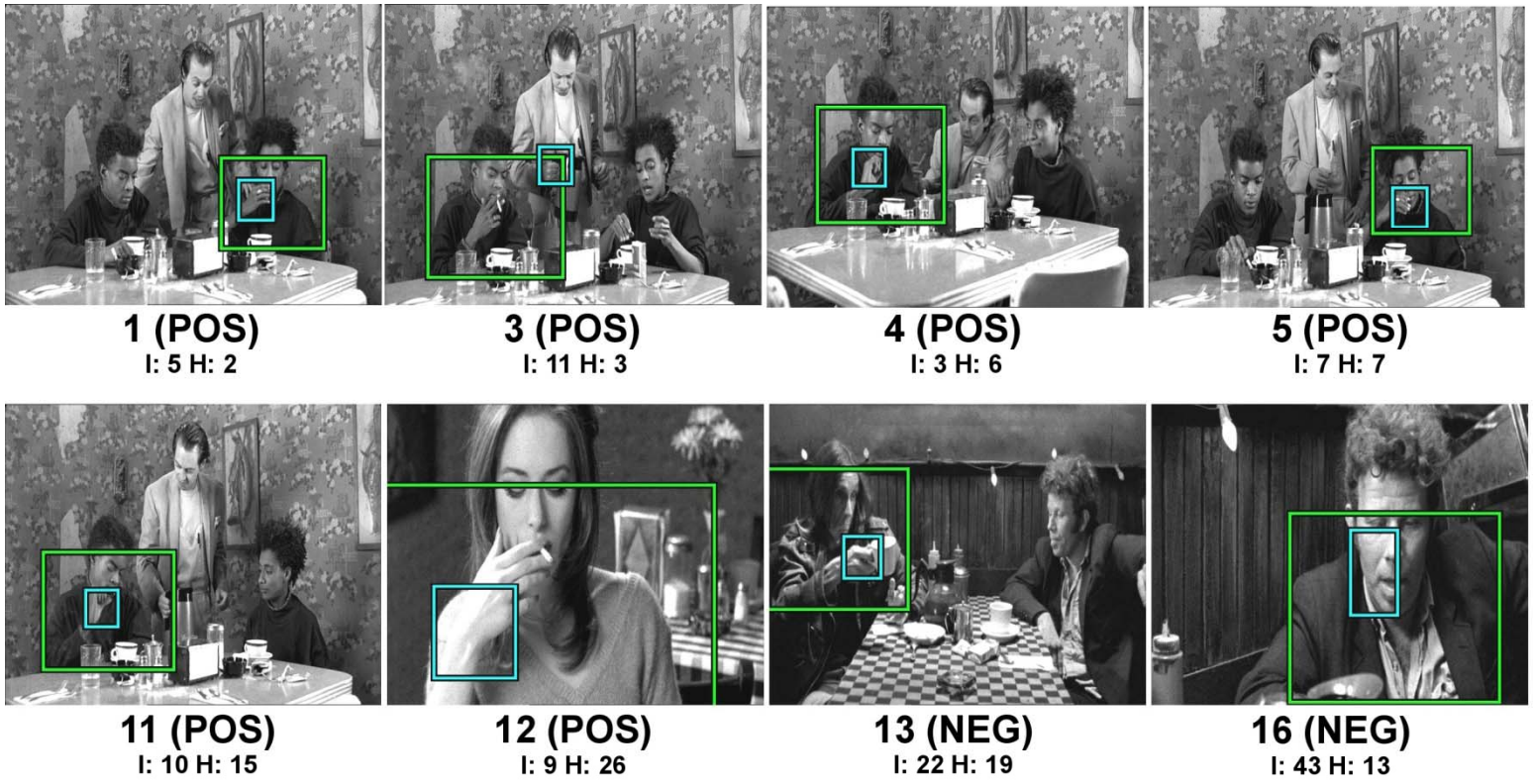
Experimental results on C&C

Drinking

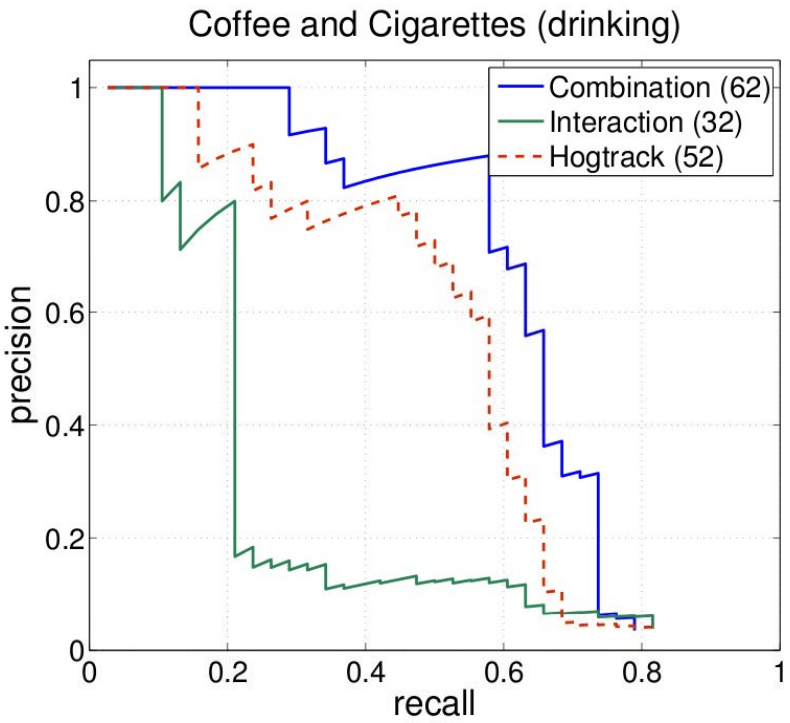


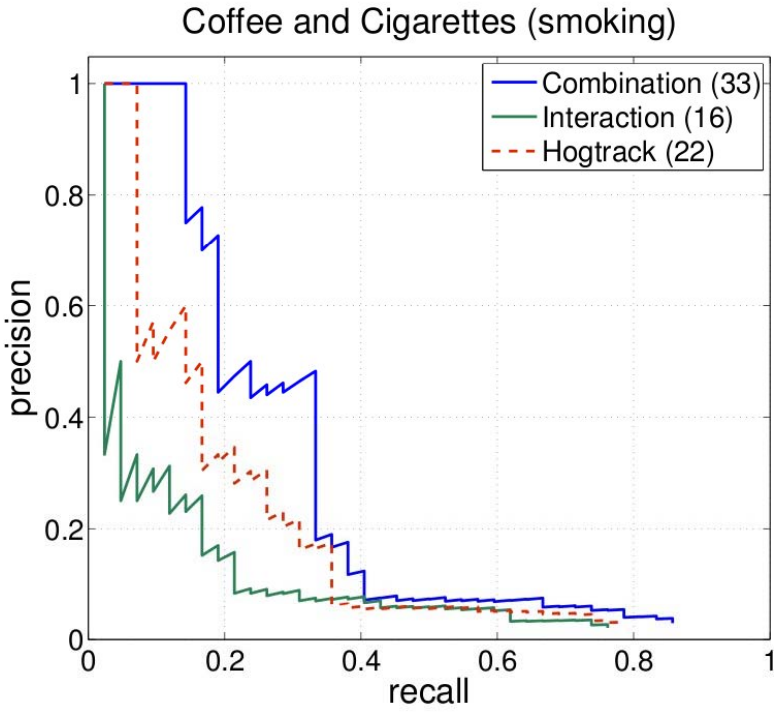
Experimental results on C&C

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

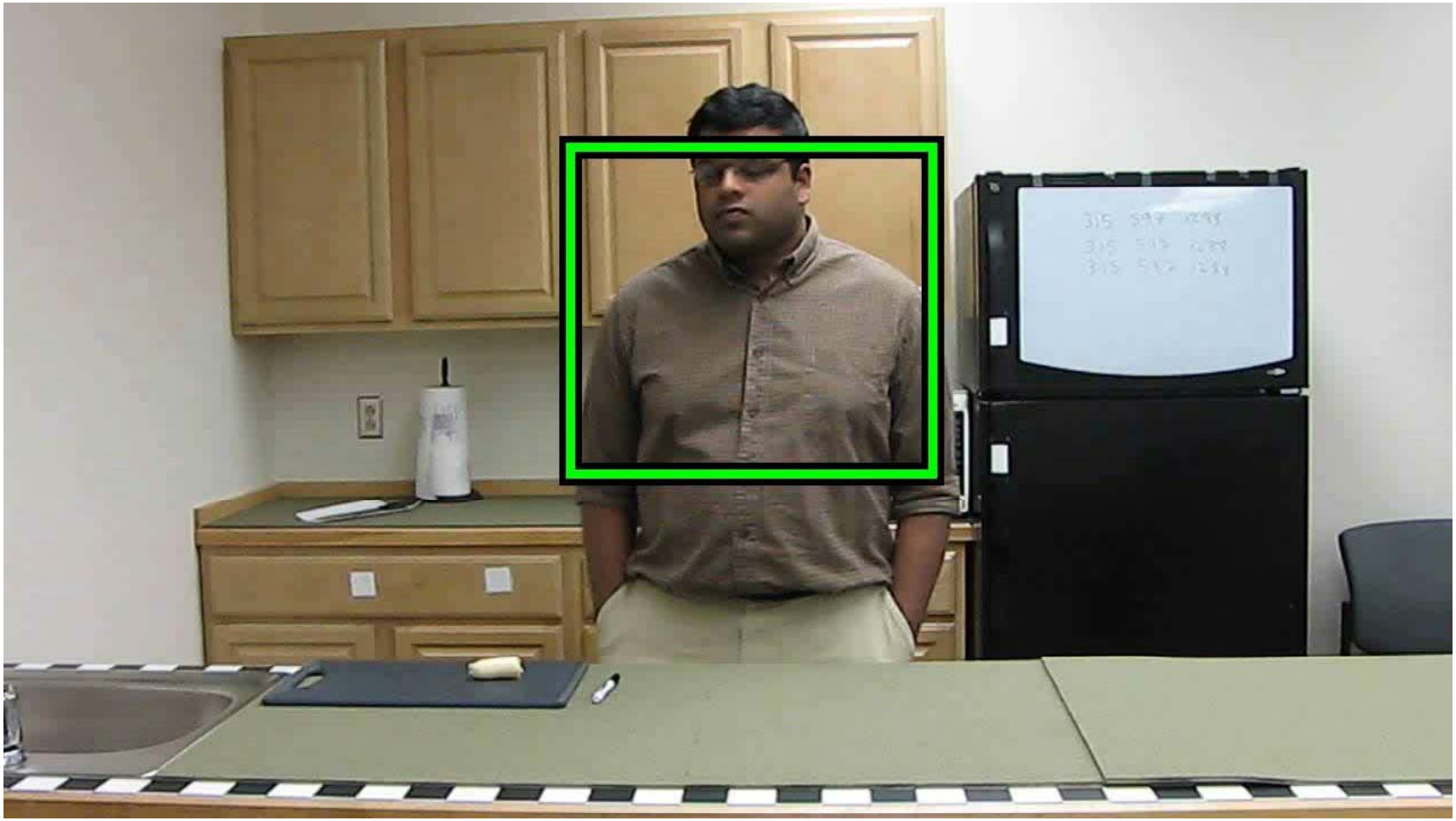
Experimental results on Rochester dataset

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

Experimental results on Rochester dataset

	Rochester Daily Activities
Interaction classifier	74
Combination (our full method)	92
Messing et al. (full method) [29]	89
Messing et al. (point tracks) [29]	67
Matikainen et al. (point tracks) [28]	70

Experimental results on Rochester dataset



Conclusion

- Human-object interaction descriptor obtains state-of-the-art performance
- Complementary to 3DHOG-track descriptor
- Combination obtains excellent performance
- Automatic extraction of objects