Action recognition in videos

Cordelia Schmid
INRIA Grenoble

Joint work with V. Ferrari, A. Gaidon, Z. Harchaoui,
A. Klaeser, A. Prest, H. Wang

Action recognition - goal

- Short actions, i.e. drinking, sit down

Drinking

Coffee & Cigarettes dataset

Sitting down

Hollywood dataset

Action recognition - goal

- Activities/events, i.e. making a sandwich, feeding an animal

Making sandwich

Feeding an animal

TrecVid Multi-media event detection dataset

Action recognition - tasks

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

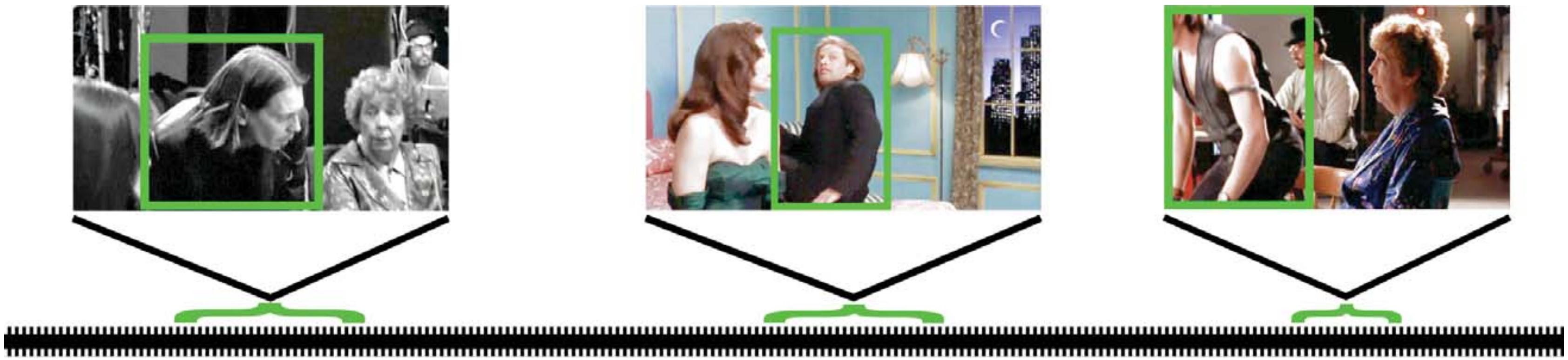
Making sandwich: present
Feeding animal: not present
...

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



Action classification – examples

diving

running

swinging

skateboarding

UCF Sports dataset (9 classes in total)

Actions classification - examples

answer phone

hand shake

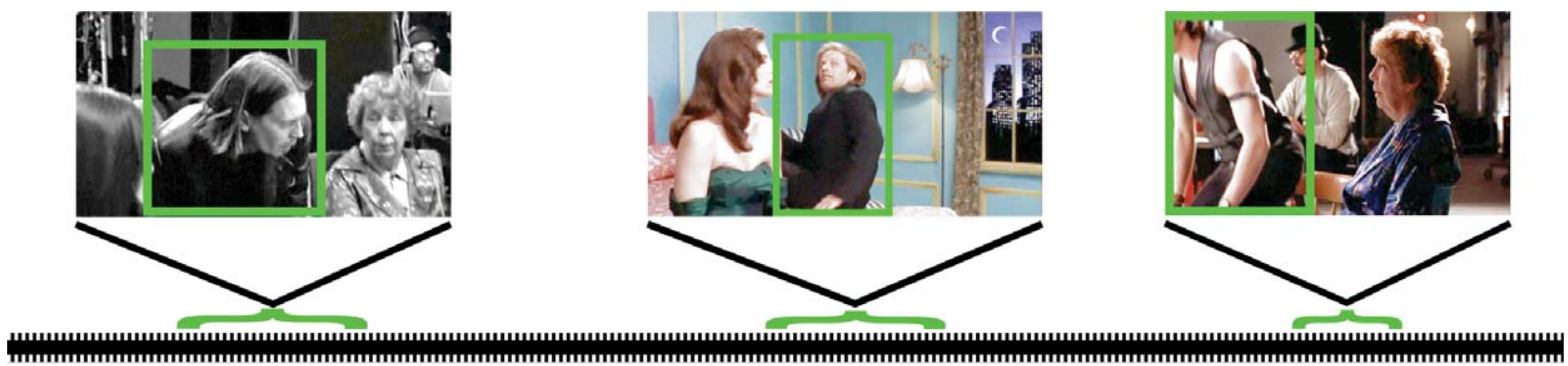
running

hugging

Hollywood2 dataset (12 classes in total)

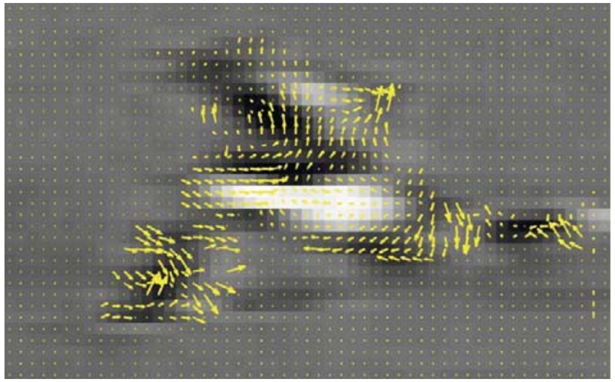
Action localization

- Find if and when an action is performed in a video
- Short human actions (e.g. “sitting down”, a few seconds)
- Long real-world videos for localization (more than an hour)
- Temporal & spatial localization: find clips containing the action and the position of the actor

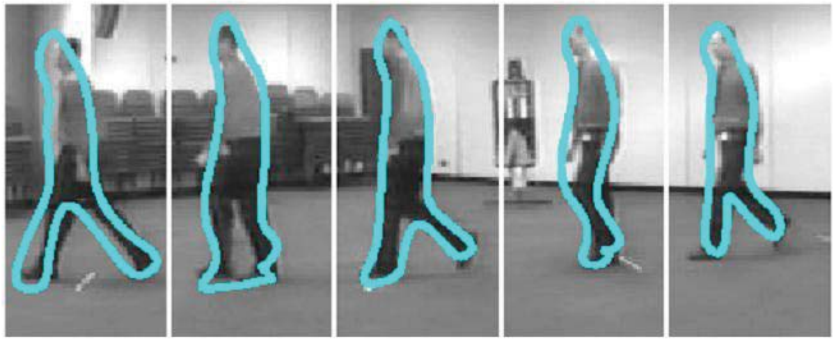


State of the art in action recognition

Motion history image
[Bobick & Davis, 2001]



Spatial motion descriptor
[Efros et al. ICCV 2003]



Learning dynamic prior
[Blake et al. 1998]

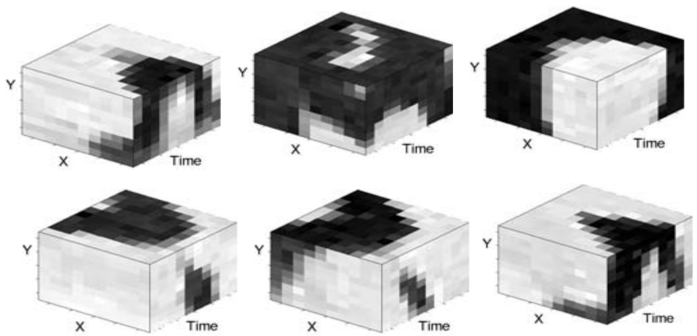
Sign language recognition
[Zisserman et al. 2009]

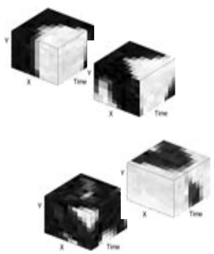
State of the art in action recognition

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

Extraction of space-time features

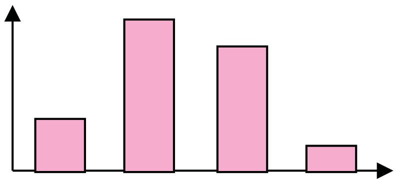
Collection of space-time patches





HOG & HOF patch descriptors

Histogram of visual words



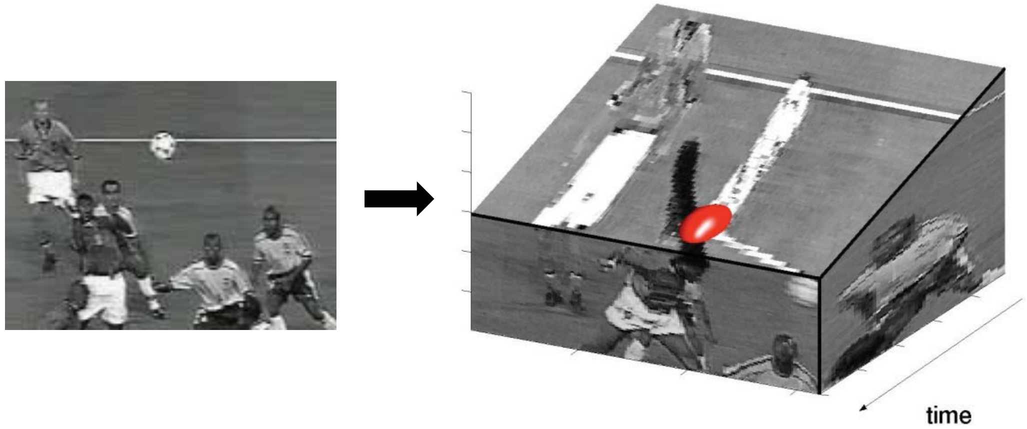
SVM classifier

Space-time features

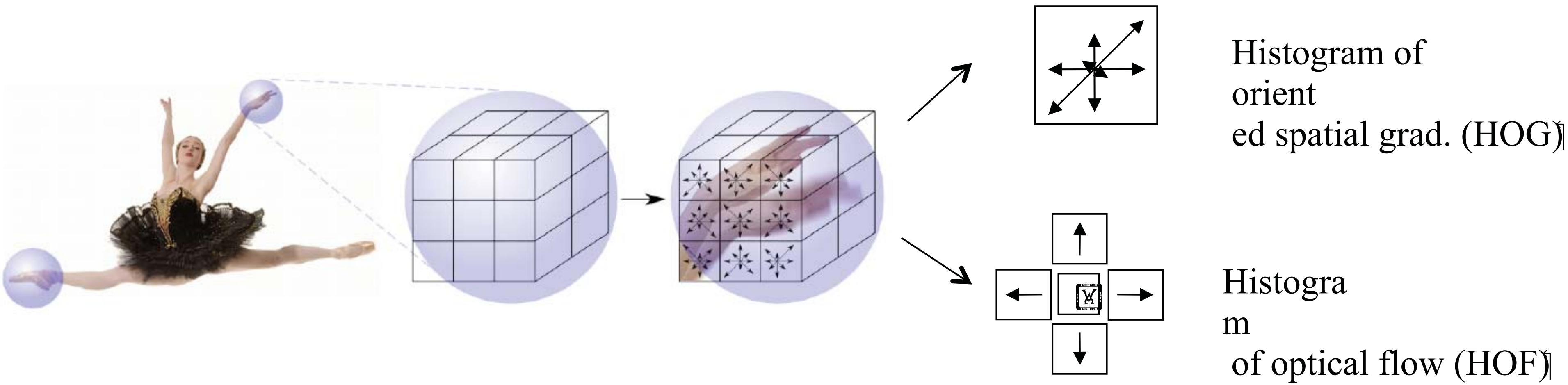
- Detector [Laptev'05]

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

$$\mu = \begin{pmatrix} I_xI_x & I_xI_y & I_xI_t \\ I_xI_y & I_yI_y & I_yI_t \\ I_xI_t & I_yI_t & I_tI_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$

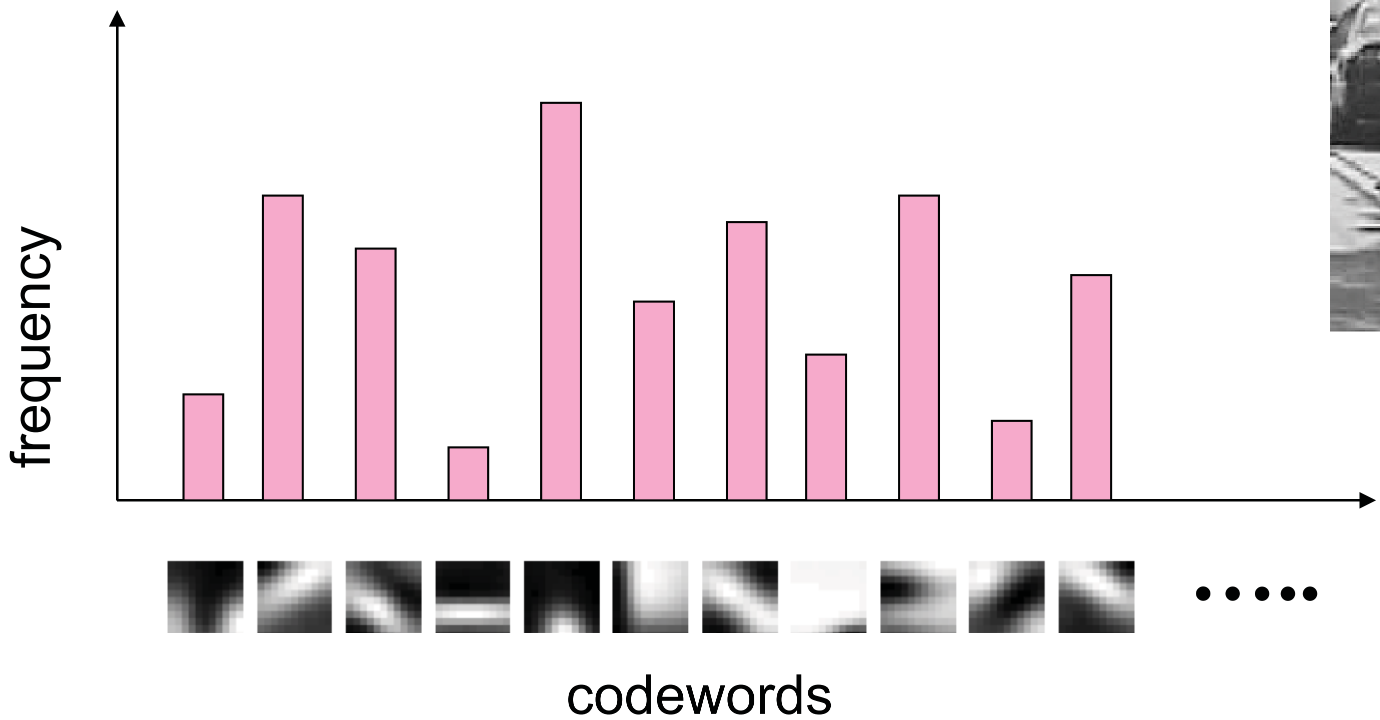


- Descriptor



Bag of features

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



Bag of features

- Advantages
  - Excellent baseline
  - Orderless distribution of local features

- Disadvantages
  - Does not take into account the structure of the action, i.e., does not separate actor and context
  - Does not allow precise localization
  - STIP are sparse features

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

Dense trajectories - motivation

- Dense sampling improves results over sparse interest points for image classification [Fei-Fei'05, Nowak'06]

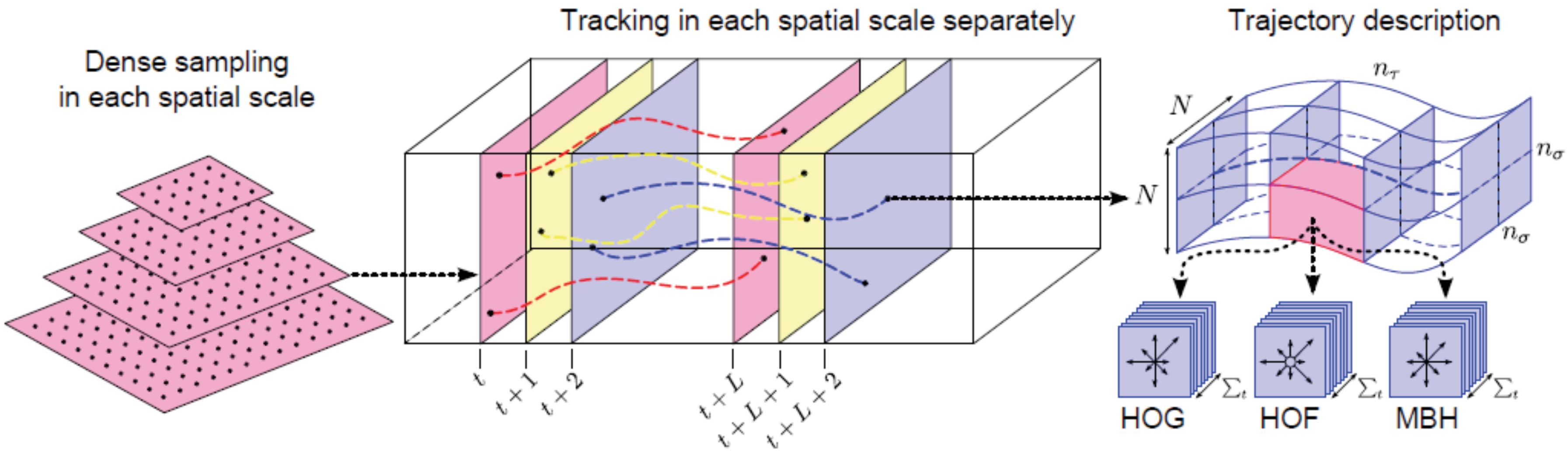
- Recent progress by using feature trajectories for action recognition [Messing'09, Sun'09]

- 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

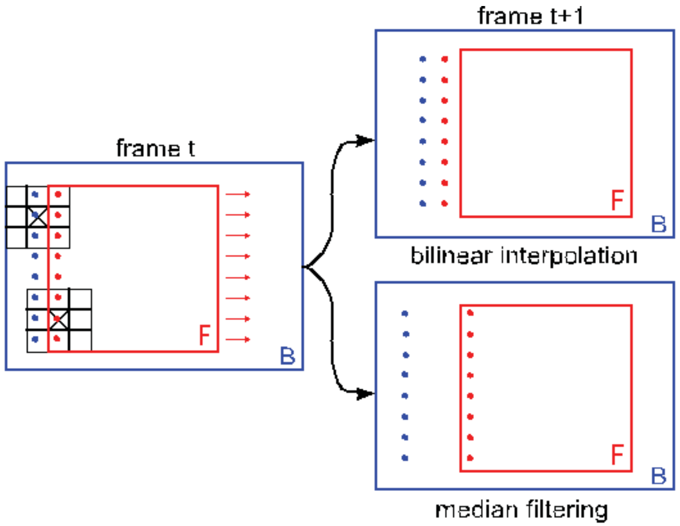
- remove untrackable points
- based on the eigenvalues of the auto-correlation matrix

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

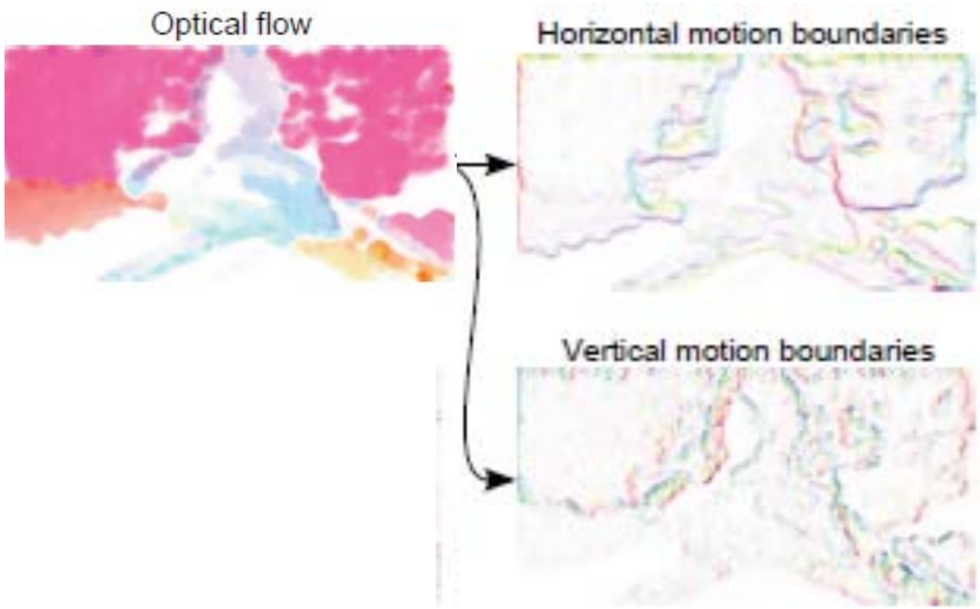
KLT tracks

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

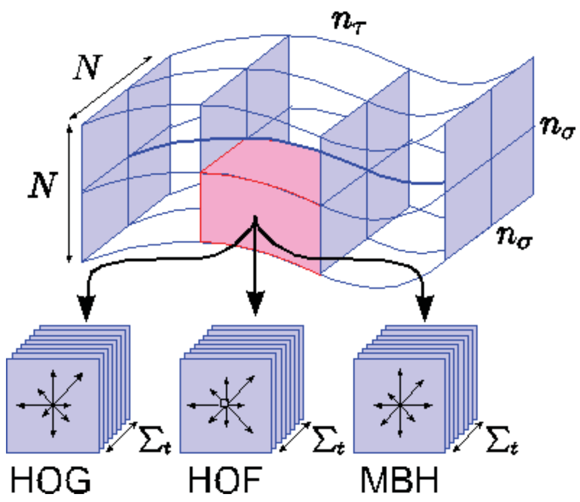


Trajectory descriptors

- 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
- 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

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

Comparison to state of the art

	Hollywood2 (SPM)	UCFSports (SPM)
Our approach (comb.)	58.2% (59.9%)	88.0% (89.1%)
[Le'2011]	53.3%	86.5%
other	53.2% [Ullah'10]	87.3% [Kov'10]

- Improves over the state of the art with a simple BOF model

Conclusion

- Dense trajectory representation for action recognition outperform existing approaches
- Motion boundary histogram descriptors perform very well, they are robust to camera motion
- Efficient algorithm, on-line available at https://lear.inrialpes.fr/people/wang/dense_trajectories

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

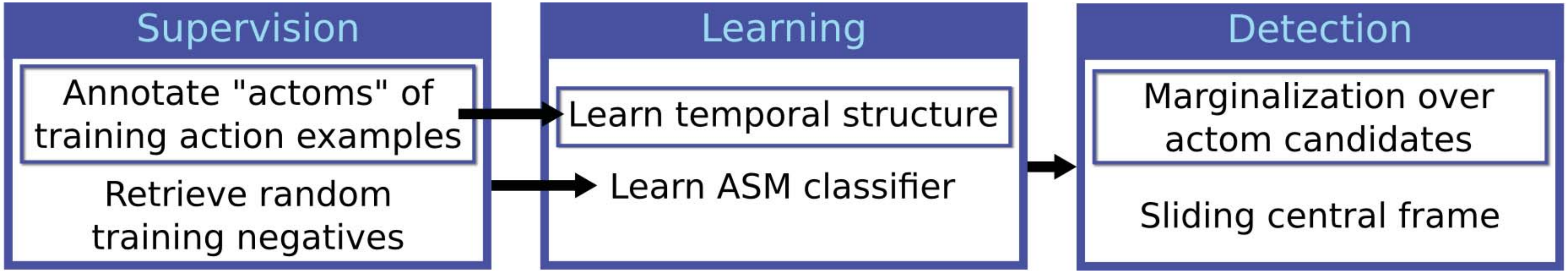
Approach for action modeling

- 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

Related work

- Temporal structuring of video data
  - Bag-of-features with spatio-temporal pyramids [Laptev'08]
  - Loose hierarchical structure of latent motion parts [Niebles'10]
  - Facial action recognition with action unit detection and structured learning of temporal segments [Simon'10]

Approach for action modeling



- Actom Sequence Model (**ASM**):
  histogram of time-anchored visual features

Actom annotation

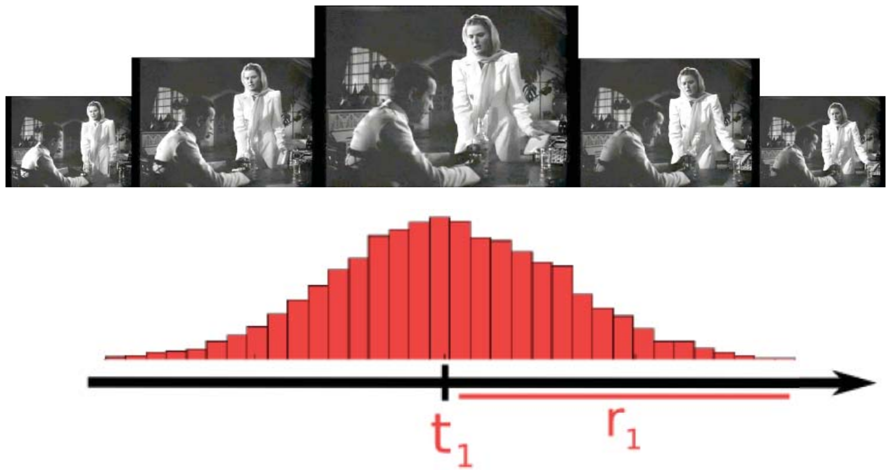
- Actoms for training actions are obtained manually
  (3 actoms per action here)

- Alternative supervision to beginning and end frames with similar cost and smaller annotation variability

- Automatic detection of actoms at test time

Actom descriptor

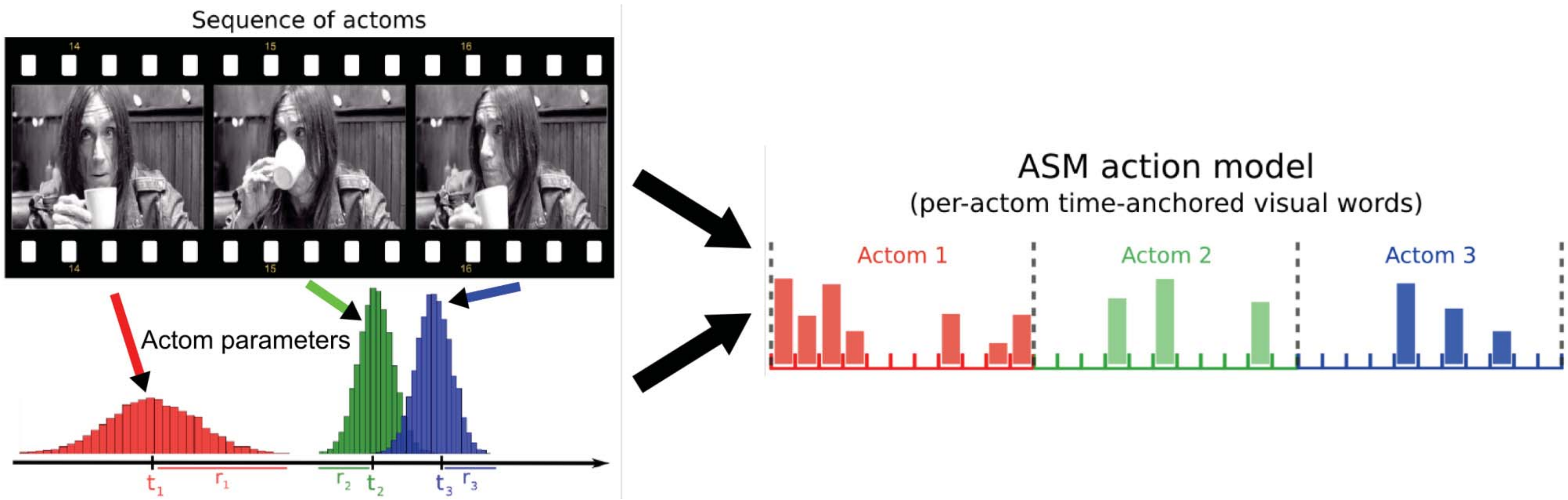
- An actom is parameterized by:
  - central frame location
  - time-span
  - temporally weighted feature assignment mechanism



- Actom descriptor:
  - histogram of quantized visual words in the actom's range
  - contribution depends on temporal distance to actom center (using temporal Gaussian weighting)

Actom sequence model (ASM)

- ASM: concatenation of actom histograms



- ASM model has two parameters: overlap between actoms and soft-voting bandwidth
  ➡ fixed to the same relative value for all actions in our experiments, depends on the distance between actoms

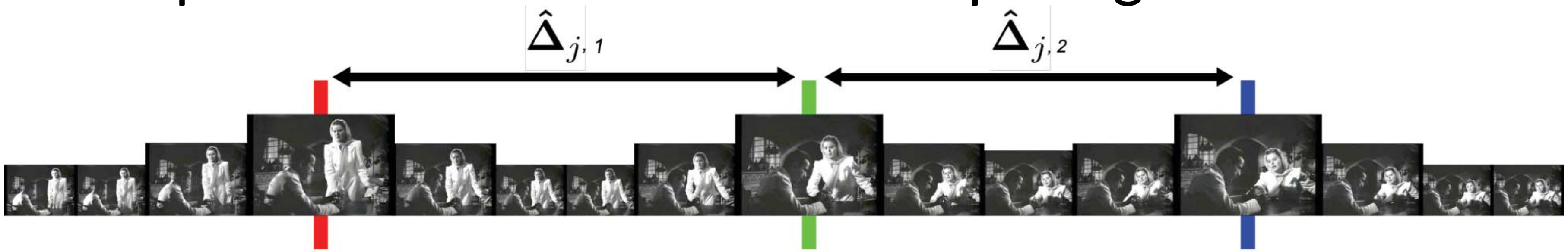
Automatic temporal detection - training

- ASM classifier:
  - non-linear SVM on ASM representations with intersection kernel, random training negatives, probability outputs
  - estimates posterior probability of an action knowing the temporal location of its actoms

- Actoms unknown at test time:
  - use training examples to learn prior on temporal structure of actom candidates

Prior on temporal structure

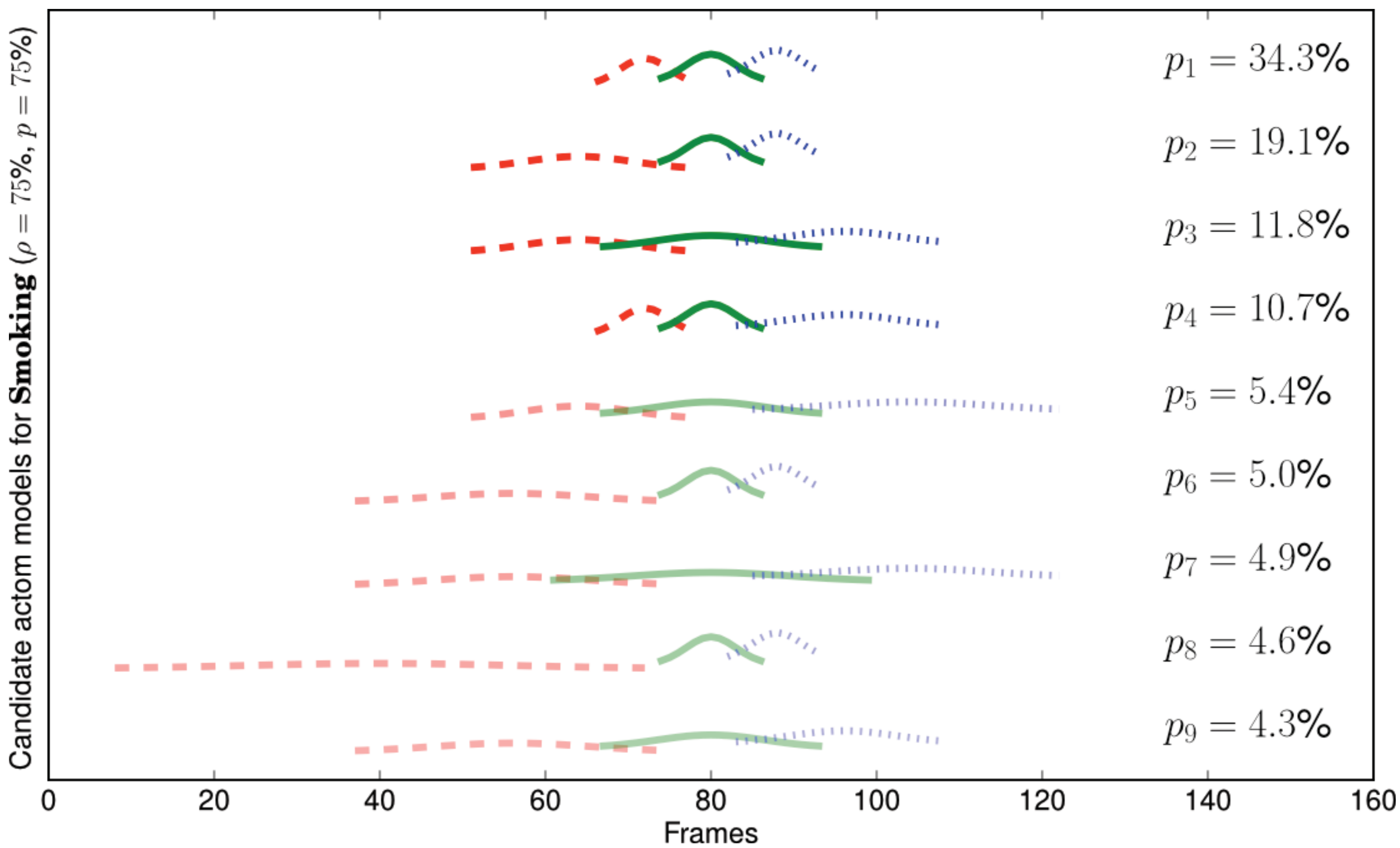
- Temporal structure: inter-actom spacings



- Non-parametric model of the temporal structure
  - kernel density estimation over inter-actom spacings from training action examples
  - discretize it to $\hat{\mathcal{D}} = \{(\hat{\mathbf{\Delta}}_j, \hat{p}_j)\ ,\ \ j = 1 \cdots K\},\ \ \hat{p}_j = \mathbf{P}(\hat{\mathbf{\Delta}}_j)$
    (small support in practice: *K≈10*)
  - use as prior on temporal structure during detection

Example of learned candidates

- Actom models corresponding to the $\hat{\mathcal{D}}$ learned for "smoking"

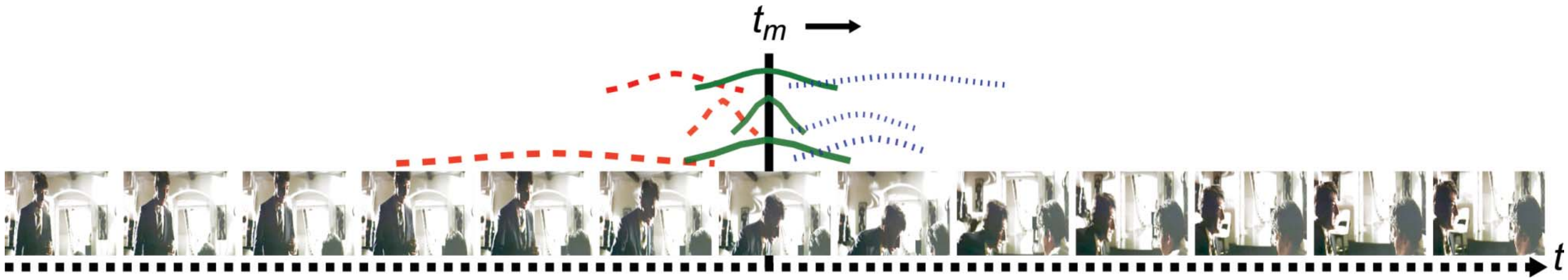


Automatic Temporal Detection

- Probability of action at frame t_m by marginalizing over all learned candidate actom sequences:

$$\mathbf{P}(\text{action at } t_m) = \sum_{j=1}^{K} \mathbf{P}(\text{action at } t_m | \hat{\mathbf{\Delta}}_\mathbf{j})\mathbf{P}(\hat{\mathbf{\Delta}}_\mathbf{j})$$

- Sliding central frame: detection in a long video stream by evaluating the probability every *N* frames (*N=5*)



- Non-maxima suppression post-processing step

Experiments - Datasets

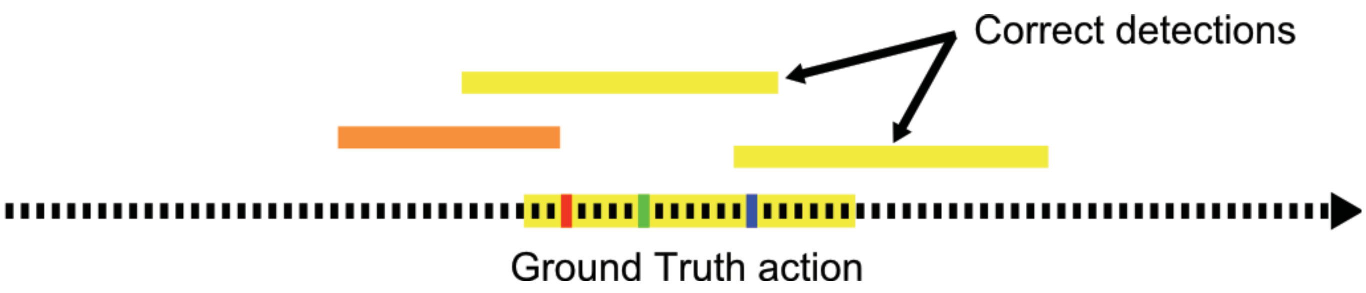
- « Coffee & Cigarettes »: localize drinking and smoking in 36 000 frames [Laptev'07]

- « DLSBP »: localize opening a door and sitting down in 443 000frames [Duchenne'09]

Performance measures

Performance measure: Average Precision (AP) computed w.r.t. overlap with ground truth test actions

•**OV20**: temporal overlap >= 20%



Quantitative Results

Coffee & Cigarettes

Method	"Drinking"	"Smoking"
matching criterion: OV20		
DLSBP [3]	40	NA
LP [12]	49	NA
KMSZ [9]	54.1	24.5
BOF	36 (±1)	19 (±1)
BOF T3	44 (±2)	23 (±3)
ASM	**57** (±3)	**31** (±2)

DLSBP

Method	"Open Door"	"Sit Down"
matching criterion: OV20		
DLSBP [3]	13.9	14.4
BOF	12.2	14.2
BOF T3	11.5	17.7
ASM	**16.4**	**19.8**

- ASM method outperforms BOF
- ASM improves over rigid temporal structure, BOF T3
  (BOF T3: concatenation of 3 BOF: beginning, middle and end of the action)
- More accurate detections with ASM compared to the state of the art

Qualitative Results

Central frames

Frames of the top 5 actions detected with ASM for drinking and opening a door

(only *#2* of opening a door is a false positive)

Qualitative Results

Actoms

Frames of automatically detected actom sequences for 4 actions

Open Door

Drinking

Smoking

Sitting Down

Qualitative Results

ASM

Automatically detected actom sequences

Localization results for action drinking

Localization results for action smoking

Conclusion

- ASM: efficient model of actions with a flexible sequence of key semantic sub-actions (actoms)

- Principled multi-scale action detection using a learned prior on temporal structure

- ASM outperforms bag-of-features, rigid temporal structures and state of the art

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*

Action recognition

- Action recognition is person-centric

Movies

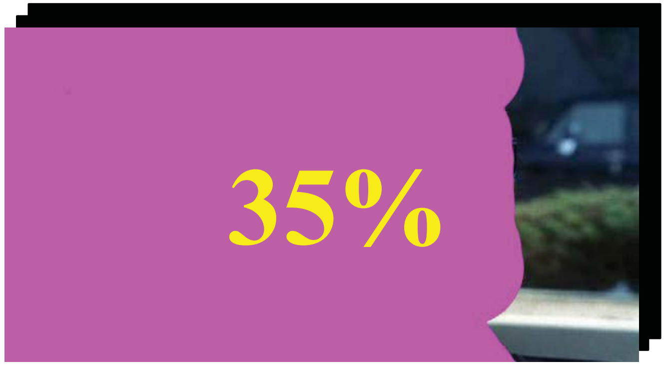
TV

YouTube

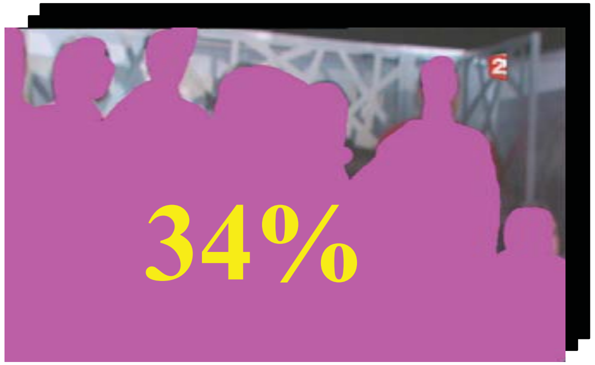
Source I.Laptev

Action recognition

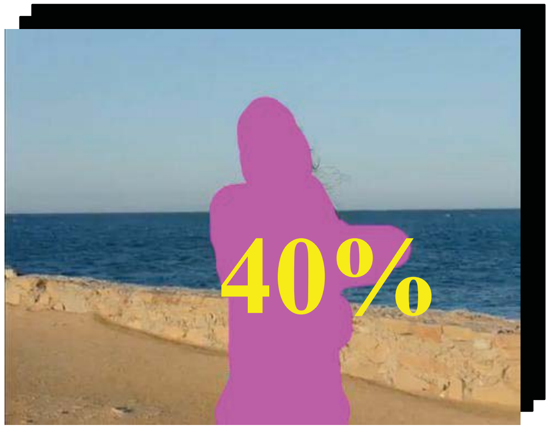
- Action recognition is person-centric



Movies



TV

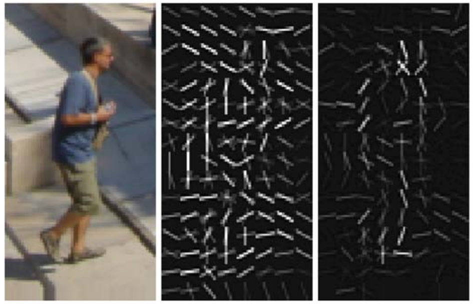


YouTube

Source I.Laptev

Action recognition

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



  - Human body part estimation

Importance of action objects

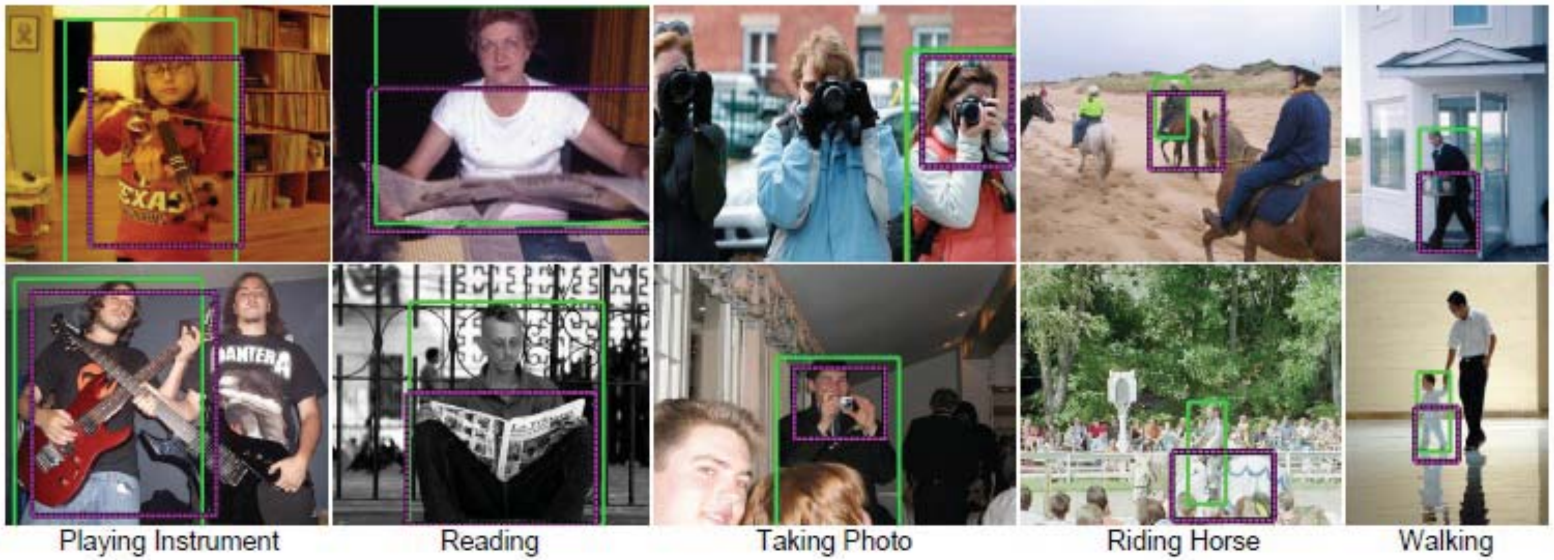
   

- Human pose often not sufficient by itself
- Objects define the actions

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

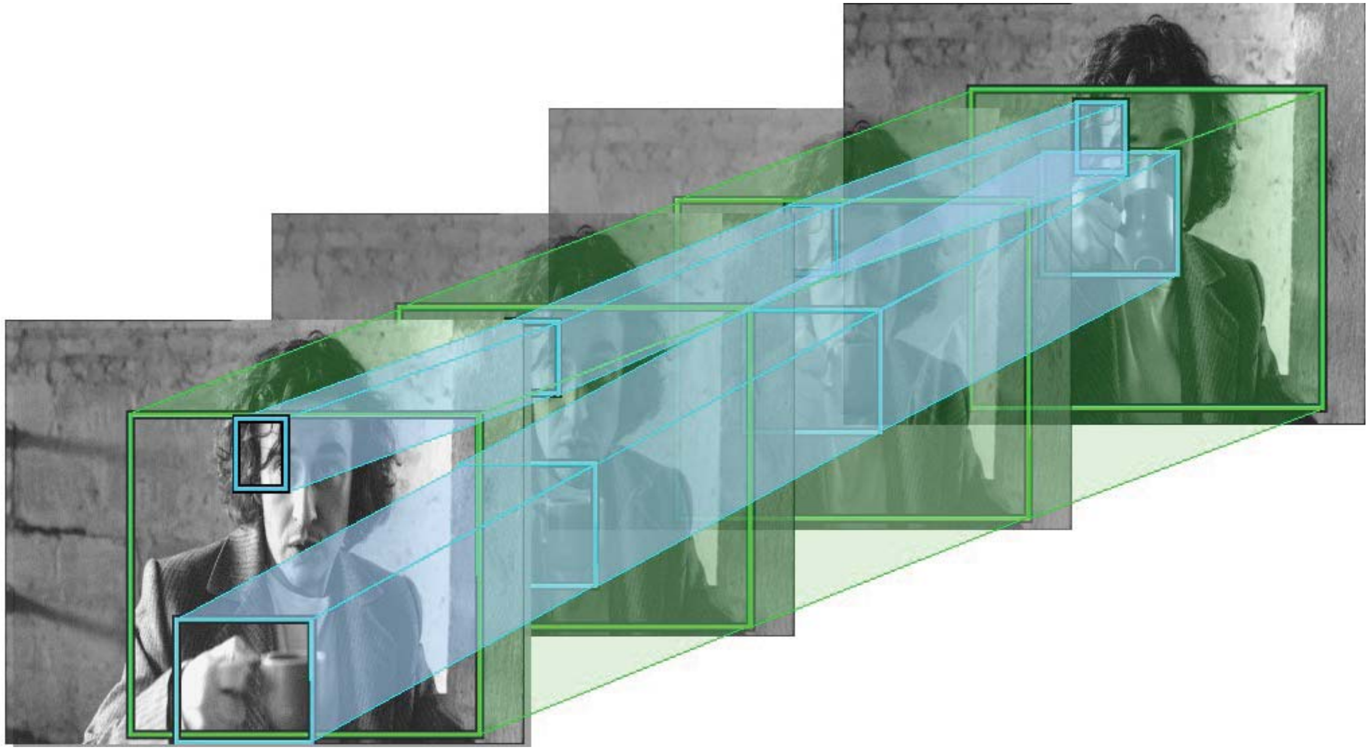



- Video/temporal information necessary to disambiguate actions
- Temporal context describes the action/activity
- Key frames provide significant less information

Modeling temporal human-object interactions

Describing human and object tracks and their relative motion

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

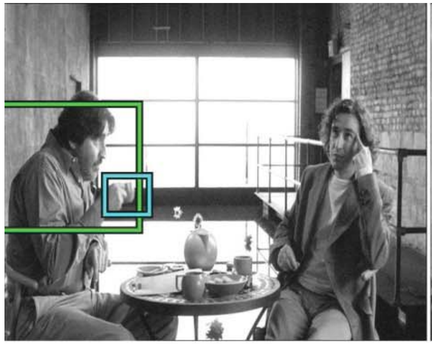
Action descriptors

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

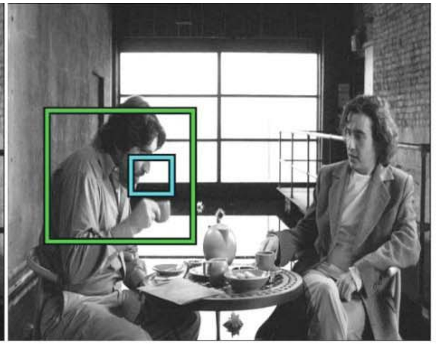
- Human track descriptor: 3DHOG-track [Klaeser et al.'10]

Experimental results on C&C

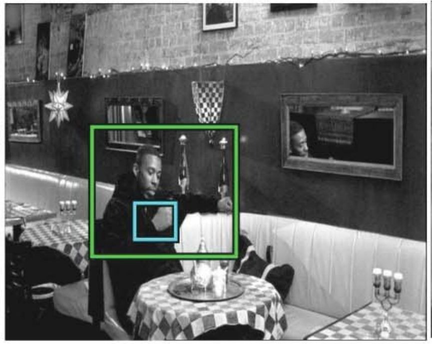
Drinking



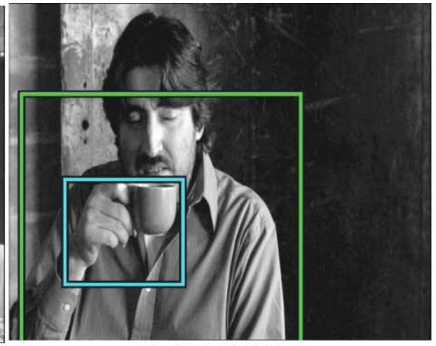
1 (POS)
I: 7 H: 1



2 (POS)
I: 17 H: 2



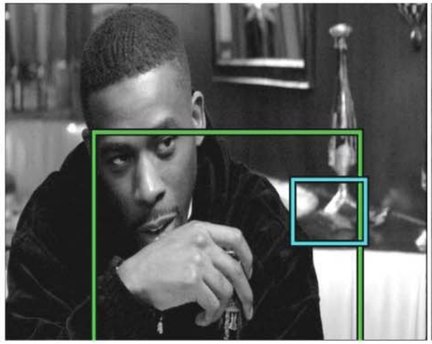
3 (POS)
I: 11 H: 3



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

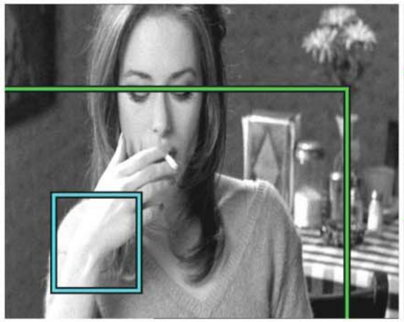

1 (POS)
I: 5 H: 2

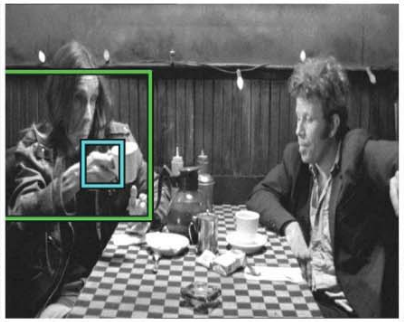

3 (POS)
I: 11 H: 3


4 (POS)
I: 3 H: 6


5 (POS)
I: 7 H: 7

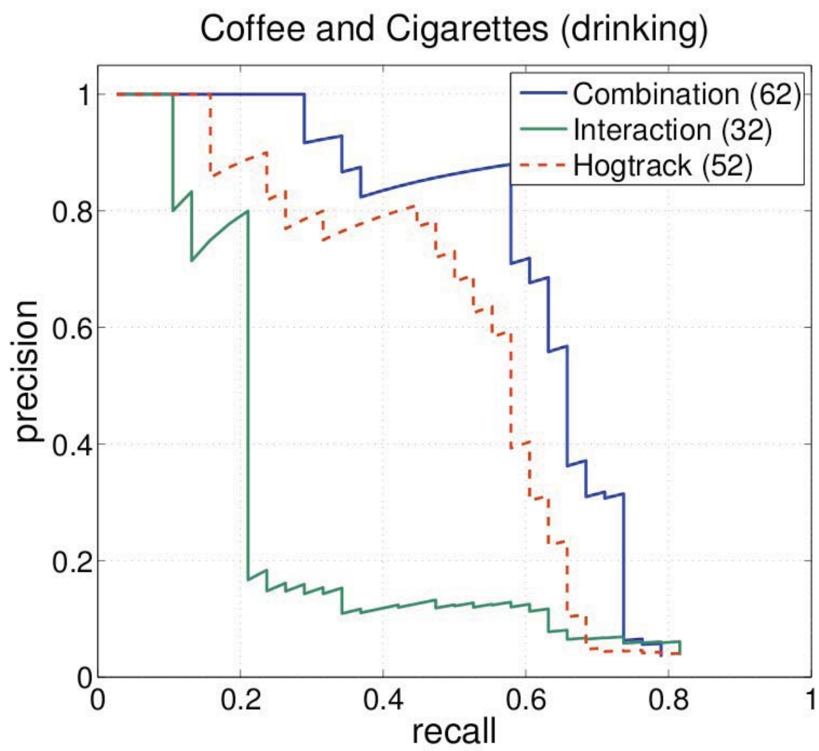

11 (POS)
I: 10 H: 15

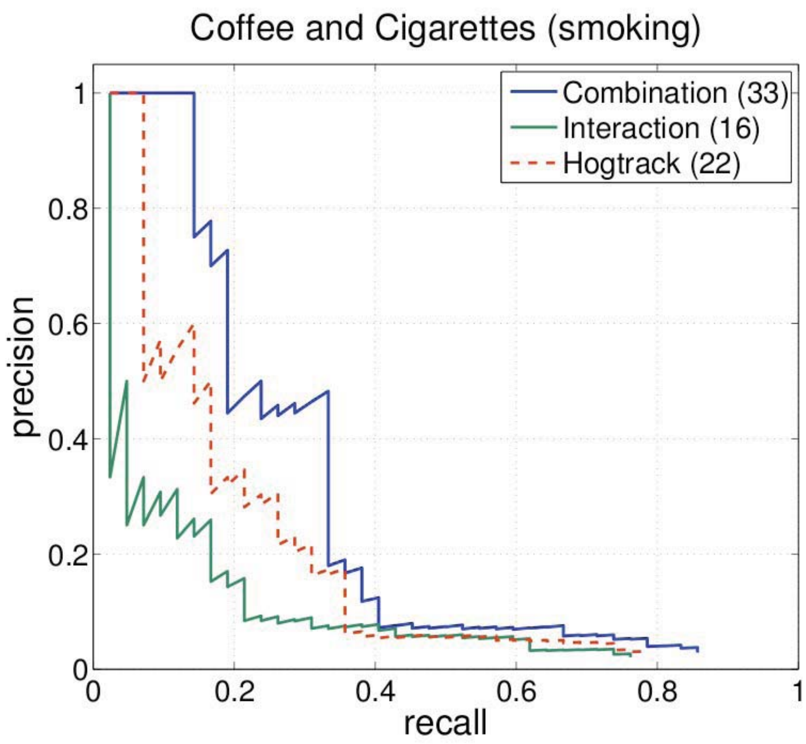

12 (POS)
I: 9 H: 26


13 (NEG)
I: 22 H: 19


16 (NEG)
I: 43 H: 13

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 Gupta dataset

Answering the phone

Making a phone call

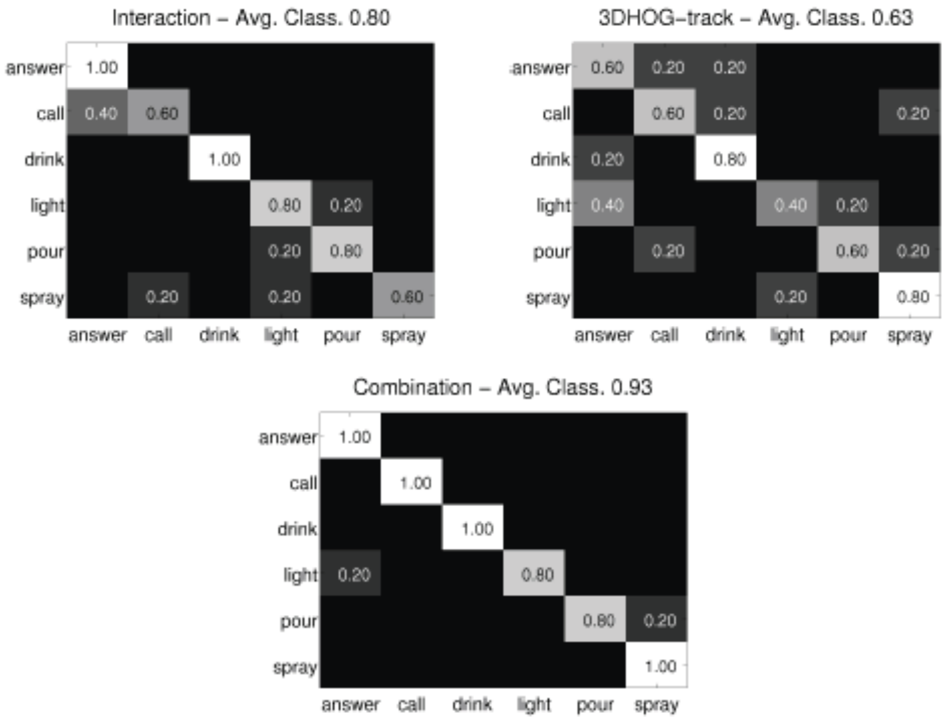
Drinking

Using a light torch

Pouring water from a cup

Using a spray bottle

Experimental results on Gupta dataset



	Gupta video
Interaction classifier	**80.00**
Object classifier	36.60
3DHOG-track classifier	63.30
Combination	**93.30**
Gupta et al. [17]	93.00

- Interactions achieve the best performance alone
- Combination improves results further: only 2 misclassified samples

-Comp. state of the art: Gupta use significantly more training information

Conclusion

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

Discussion

- Need for more challenging datasets
  - Need for realistic datasets

KTH dataset

Hollywood dataset

  - Scale up number of classes (today ~10 actions per dataset)
  - Increase number of examples per class, possibly with weakly supervised learning (the number of examples per videos is low)
  - Define a taxonomy, use redundancy between action classes to improve training
  - Manual exhaustive labeling of all actions impossible

Discussion

- Make better use of the large amount of information inherent in videos
  - automatic collection of additional examples
  - improve models incrementally
  - use weak labels from associated data (text, sound, subtitles)

- Many existing techniques are straightforward extensions of methods for images
  - almost no use of 3D information
  - learn better interaction and temporal models
  - design activity models by decomposition into simple actions

Actom Sequence Model (ASM)

Parameters

- Amount of overlap ρ between closest adjacent actoms
  - defines an adaptive actom time-span $r_i = \frac{d_i}{2-\rho}$
  - robustness to inaccurate temporal localization of actoms while ensuring temporal ordering
  - allows for gaps to represent actions with temporal discontinuities
- "Peaky-ness" *p* of the time-dependent Gaussian soft-voting
  - each feature at frame *t* in the time-span of actom (t_i, r_i) has its contribution weighted by its temporal distance to t_i:

$$w(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{|t-t_i|^2}{2\sigma^2}\right) \qquad \mathbf{P}(|t-t_i| < r_i) \leq p = 1 - \frac{\sigma^2}{r_i^2}$$

  - *p* is the amount of probability mass in an actom's range
  - small p: BOF-like actoms, large p: keyframe-like actoms
- Parameters fixed to ρ *=75%* and *p=75%* for all experiments

Our approach: modeling human-object interactions