Overview

- Optical flow
- *Video classification*
  - *Bag of spatio-temporal features*
- Action localization
  - Spatio-temporal human localization

State of the art for video classification

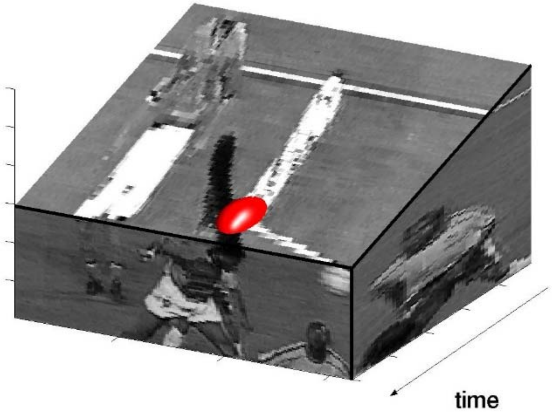
- Space-time interest points [Laptev, IJCV’05]
- Dense trajectories [Wang and Schmid, ICCV’13]
- Video-level CNN features

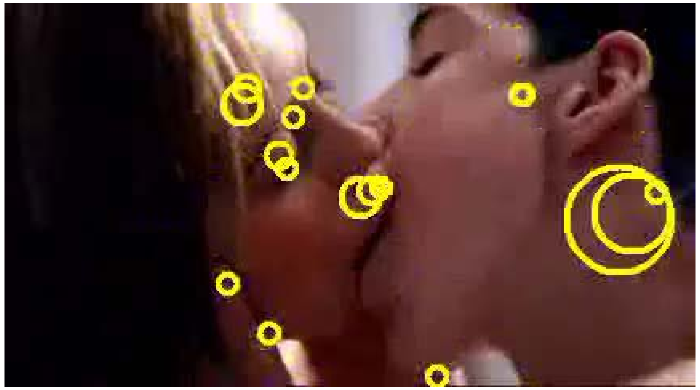
Space-time interest points (STIP)

- Space-time corner detector
  [Laptev, IJCV 2005]

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

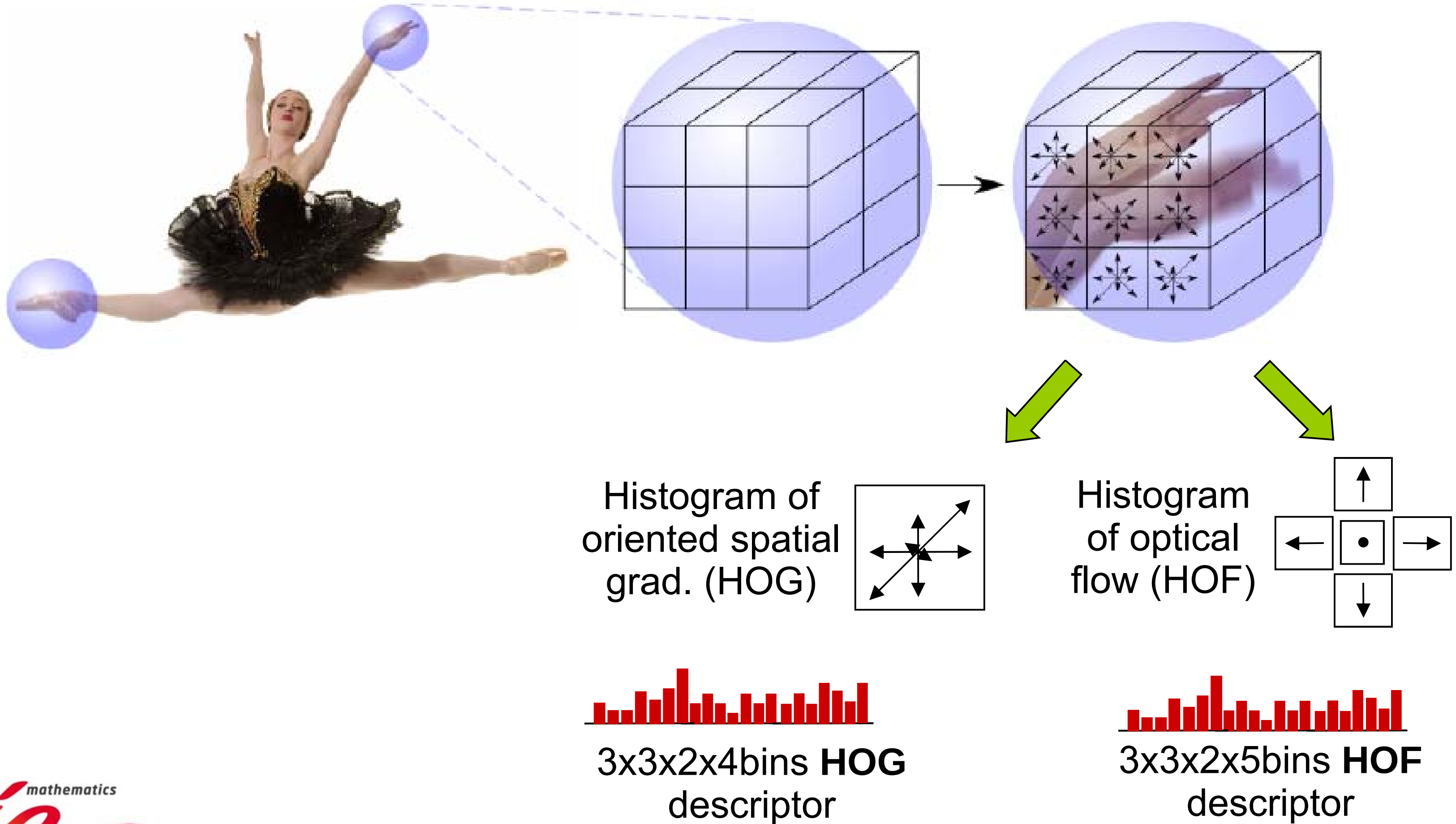
$$\mu = \begin{pmatrix} I_x I_x & I_x I_y & I_x I_t \\ I_x I_y & I_y I_y & I_y I_t \\ I_x I_t & I_y I_t & I_t I_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$





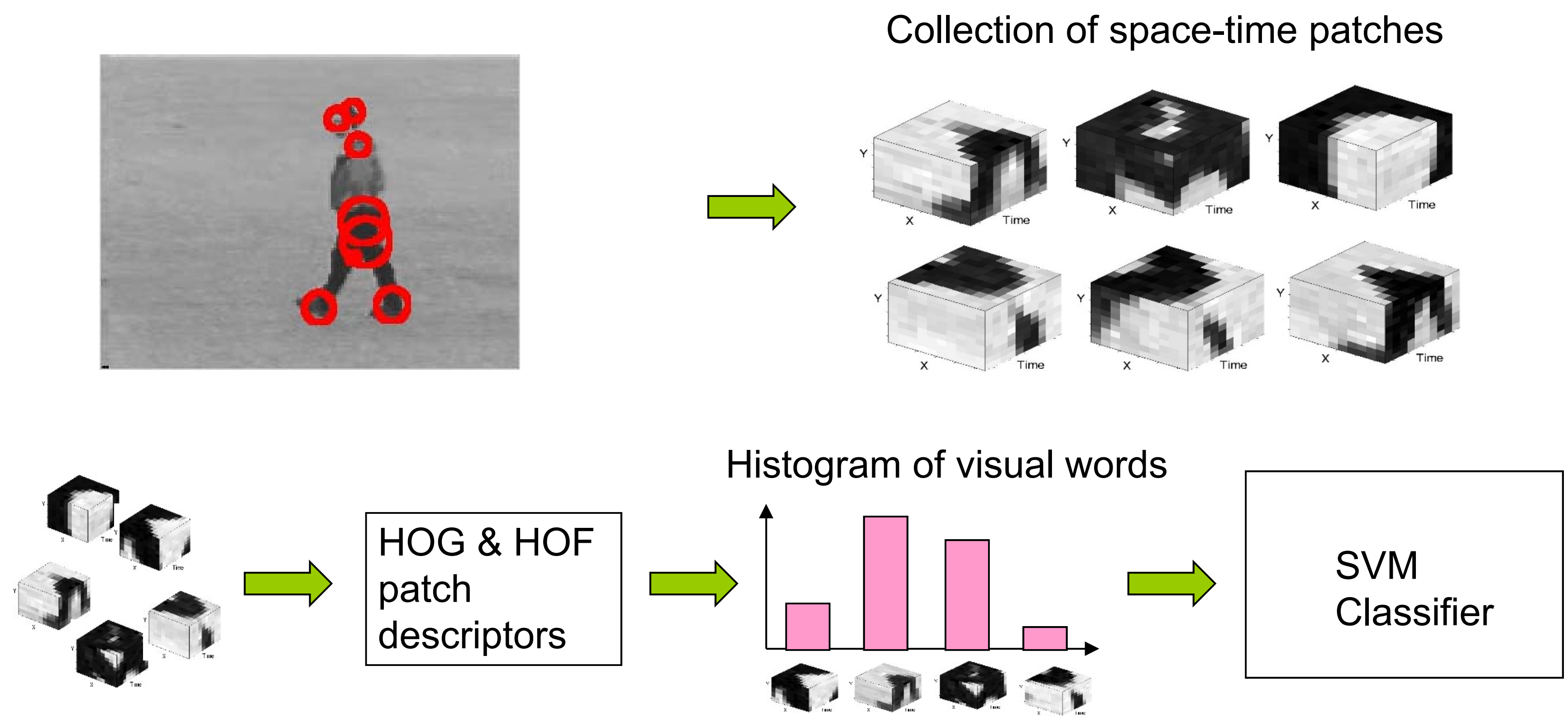
STIP descriptors

Space-time interest points



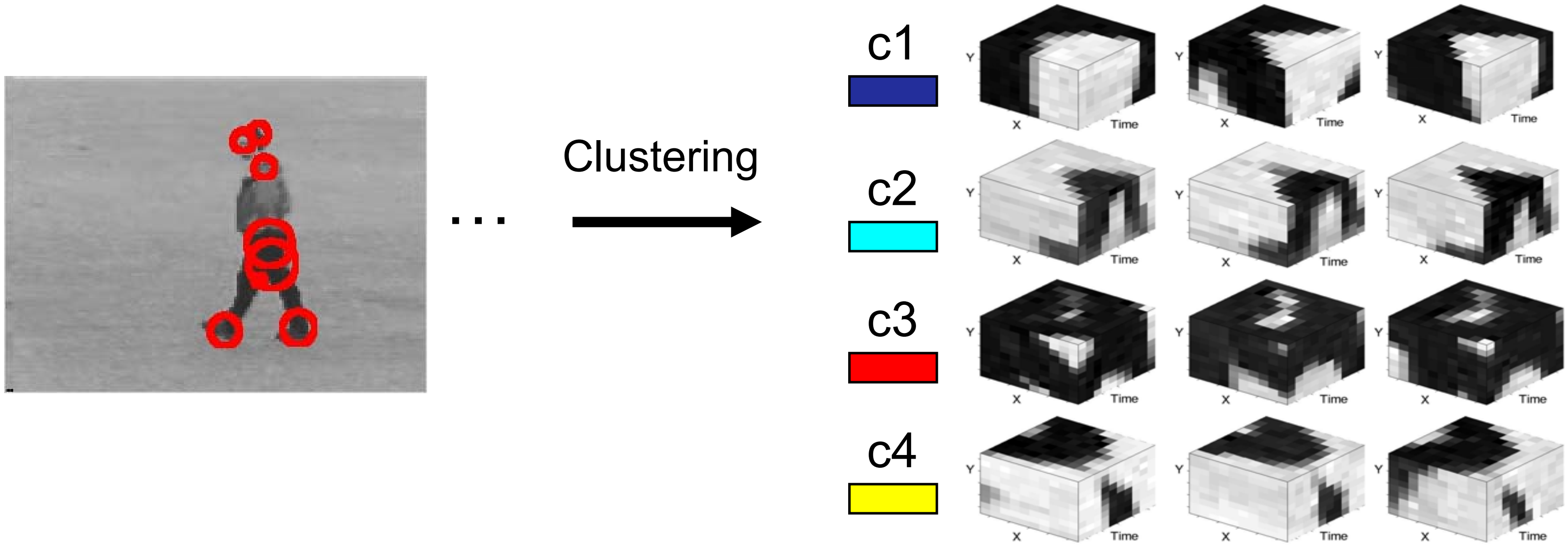
Action classification

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



Visual words: k-means clustering

- Group similar STIP descriptors together with k-means

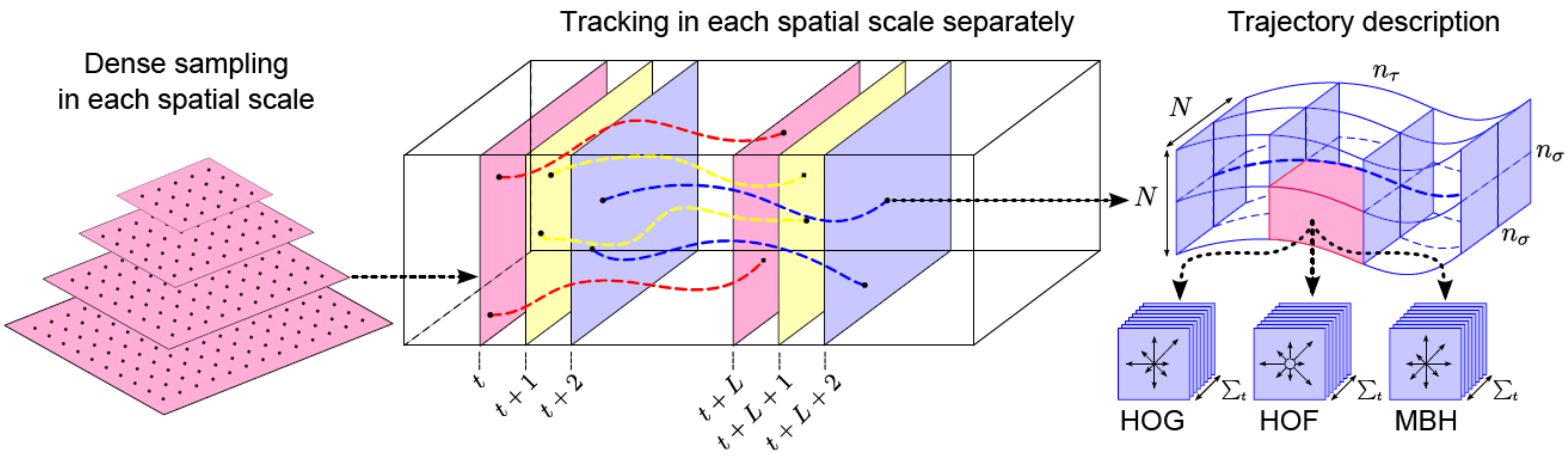


Action classification

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

State of the art for video description

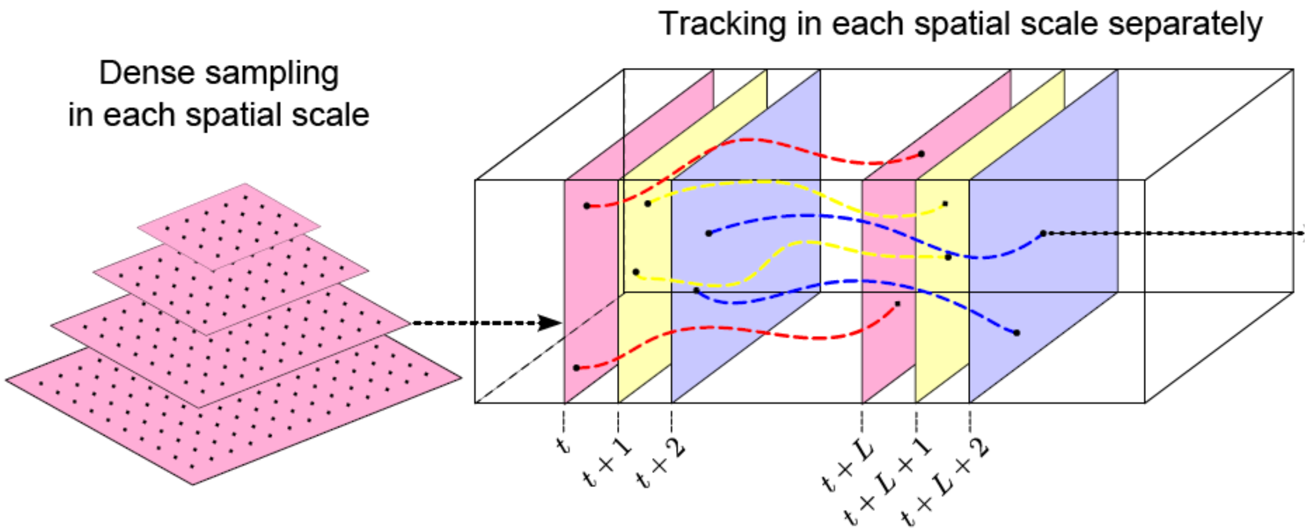
- Dense trajectories [Wang et al., IJCV'13] and Fisher vector encoding [Perronnin et al. ECCV'10]



- Orderless representation

Dense trajectories [Wang et al., IJCV'13]

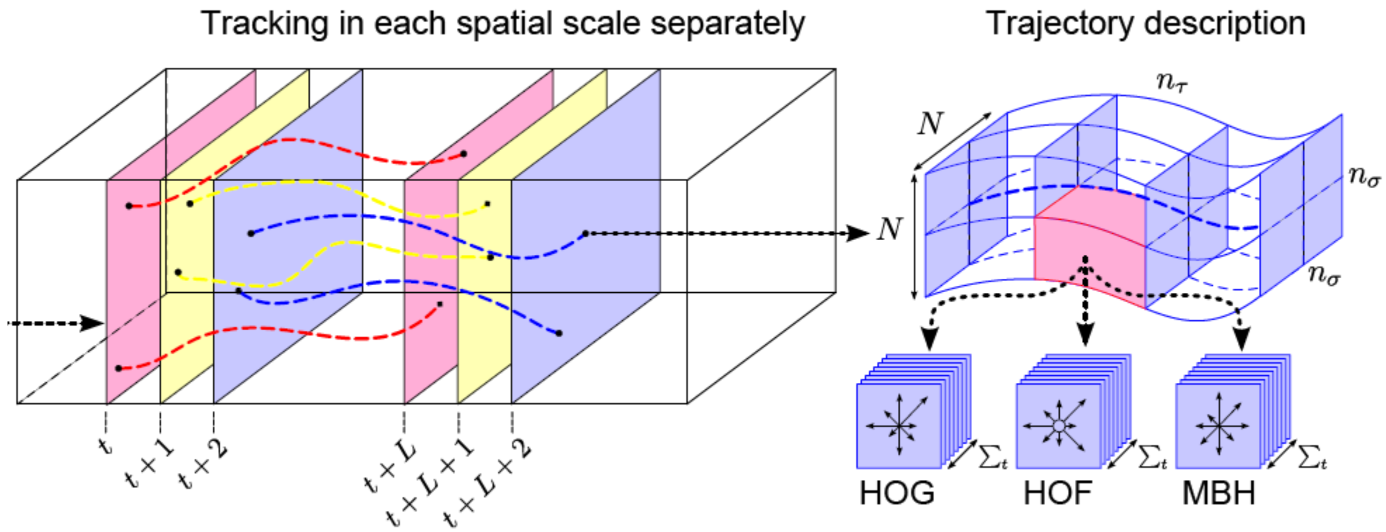
- Dense sampling at several scales
- Feature tracking based on optical flow for several scales
- Length 15 frames, to avoid drift



Example for dense trajectories

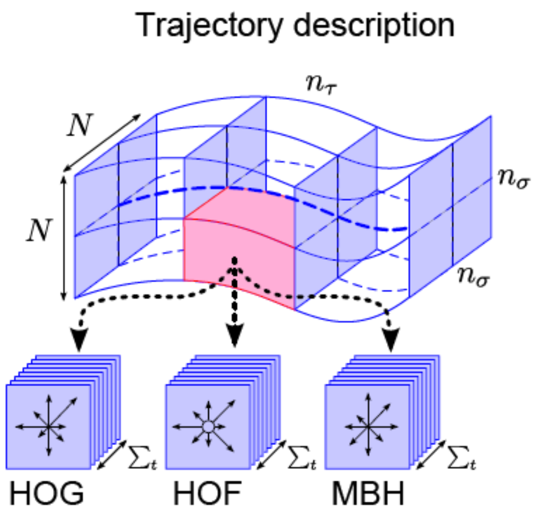
Descriptors for dense trajectory

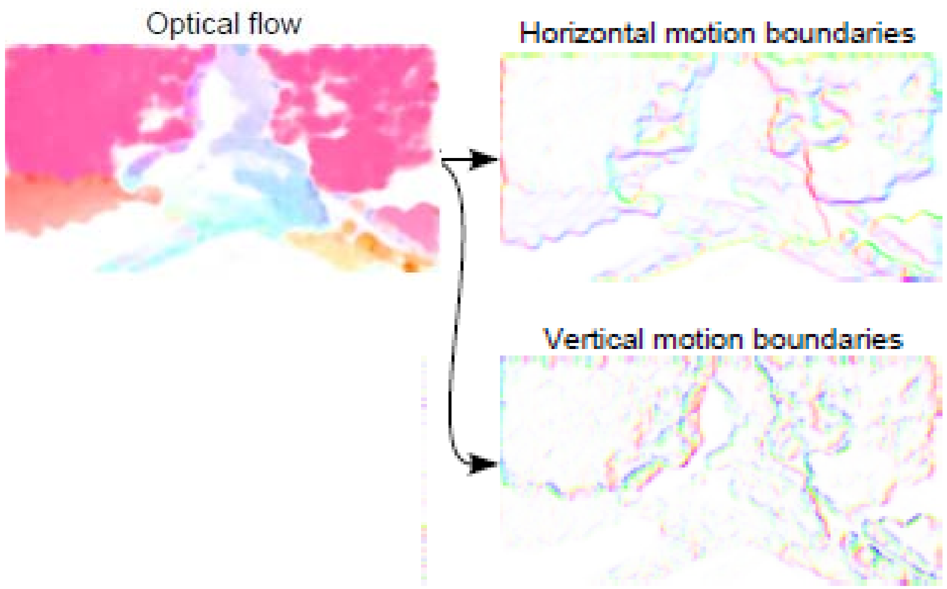
- Histogram of gradients (HOG: 2x2x3x8)
- Histogram of optical flow (HOF: 2x2x3x9)



Descriptors for dense trajectory

- Motion-boundary histogram (MBHx + MBHy: 2x2x3x8)
  - spatial derivatives are calculated separately for optical flow in x and y, quantized into a histogram
  - captures relative dynamics of different regions
  - suppresses constant motions





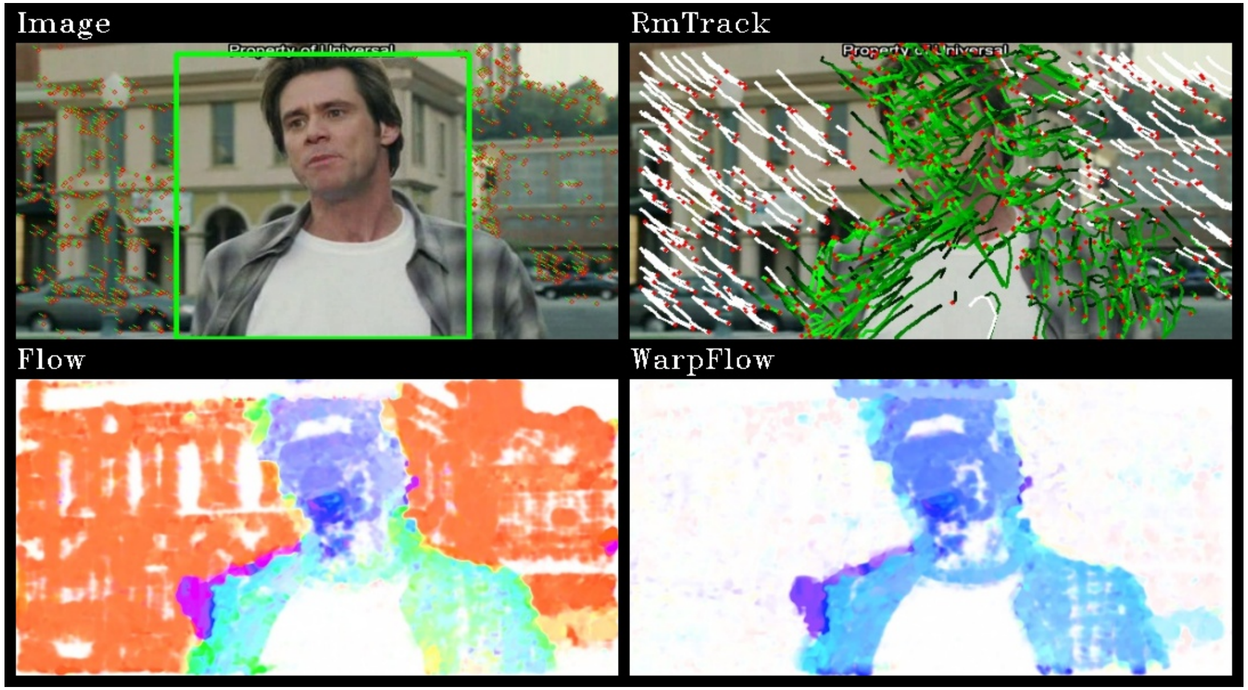
Dense trajectories

- 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

- Improve dense trajectories by explicit camera motion estimation
- Detect humans to remove outlier matches for homography estimation
- Stabilize optical flow to eliminate camera motion



[Wang and Schmid. Action recognition with improved trajectories. ICCV'13]

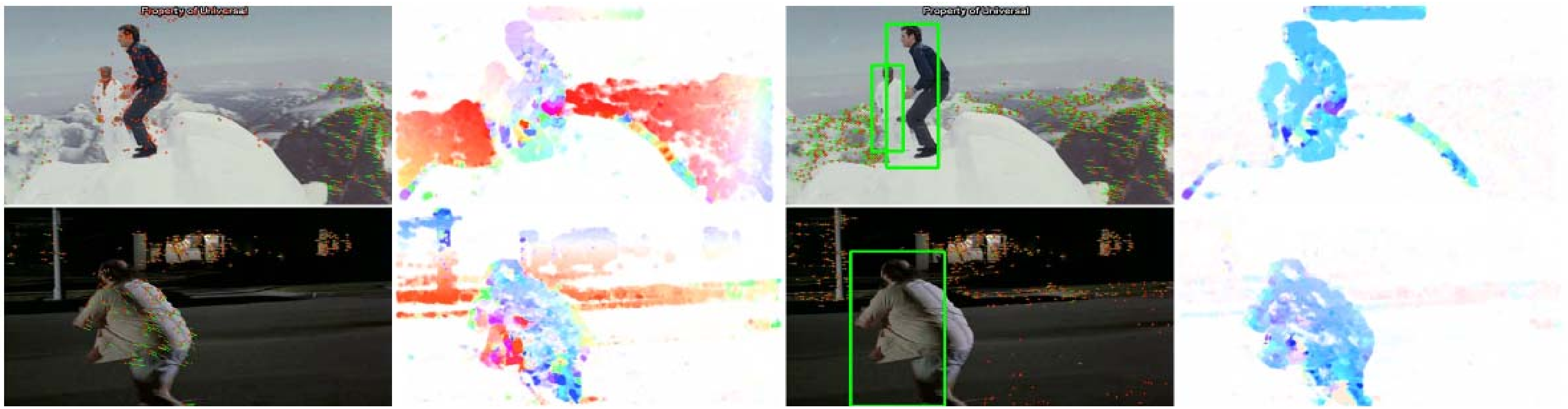
Camera motion estimation

- Find the correspondences between two consecutive frames:
  - Extract and match SURF features (robust to motion blur)
  - Use optical flow, remove uninformative points
- Combine SURF (green) and optical flow (red) results in a more balanced distribution
- Use RANSAC to estimate a homography from all feature matches

Inlier matches of the homography

Remove inconsistent matches due to humans

- Human motion is not constrained by camera motion, thus generates outlier matches
- Apply a human detector in each frame, and track the human bounding box forward and backward to join detections
- Remove feature matches inside the human bounding box during homography estimation



Inlier matches and warped flow, without or with HD

Remove background trajectories

- Remove trajectories by thresholding the maximal magnitude of stabilized motion vectors

- Our method works well under various camera motions, such as pan, zoom, tilt

Successful examples

Failure cases

Removed trajectories (white) and foreground ones (green)

- Failure due to severe motion blur; the homography is not correctly estimated due to unreliable feature matches

Experimental setting

- Motion stabilized trajectories and features (HOG, HOF, MBH)
- Normalization for each descriptor, then PCA to reduce its dimension by a factor of two
- Use Fisher vector to encode each descriptor separately, set the number of Gaussians to K=256
- Use Power+L2 normalization for FV, and linear SVM with one-against-rest for multi-class classification

Datasets

- Hollywood2: 12 classes from 69 movies, report mAP
- HMDB51: 51 classes, report accuracy on three splits
- UCF101: 101 classes, report accuracy on three splits

Datasets

Hollywood dataset [Marszalek et al.'09]

answer phone

get out of car

fight person

Hollywood2: 12 classes from 69 movies, report mAP

Datasets

HMDB 51 dataset [Kuehne et al.'11]

push-up cartwheel sword-exercice

HMDB51: 51 classes, report accuracy on three splits

Datasets

UCF 101 dataset [Soomro et al.'12]

haircut

archery

ice-dancing

UCF101: 101 classes, report accuracy on three splits

Impact of feature encoding on improved trajectories

Datasets	Fisher vector		
	DTF	ITF wo human	ITF w human
Hollywood2	63.6%	66.1%	66.8%
HMDB51	55.9%	59.3%	60.1%
UCF101	83.5%	85.7%	86.0%

Compare DTF and ITF with and without human detection using HOG+HOF+MBH and Fisher encoding

- IDT significantly improvement over DT
- Human detection always helps. For Hollywood2 and HMDB51, the difference is more significant, as there are more humans present.
- Source code: http://lear.inrialpes.fr/~wang/improved_trajectories

TrecVid MED 2011

- 15 categories

Attempt a board trick

Feed an animal

Landing a fish

…

Wedding ceremony

Working on a wood project

Birthday party

TrecVid MED 2011

- 15 categories
- ~100 positive video clips per event category, 9600 negative video clips
- Testing on 32000 videos clips, i.e., 1000 hours
- Videos come from publicly available, user-generated content on various Internet sites

- Descriptors: MBH, SIFT, audio, text & speech recognition

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

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

Performance of all channels (mAP)

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

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

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

Performance of all channels (mAP)

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

Experimental results

- Example results

rank 1

rank 2

rank 3

Highest ranked results for the event «horse riding competition»

Experimental results

- Example results

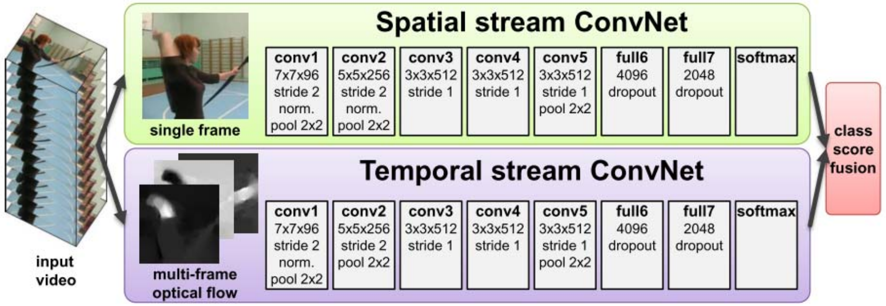
rank 1 rank 2 rank 3

Highest ranked results for the event «tuning a musical instrument»

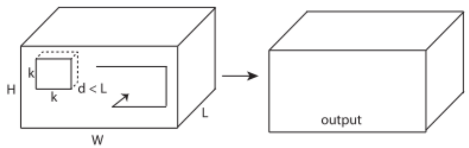
informatics mathematics
inria

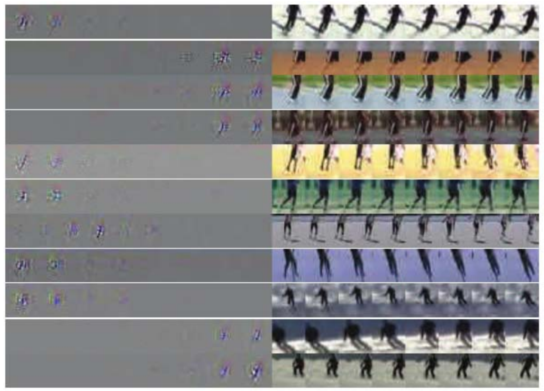
Recent CNN methods

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

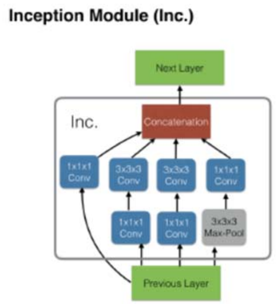


Learning Spatiotemporal Features with 3D Convolutional Networks [Tran et al. ICCV15]





Quo vadis action recognition? A new model and the Kinetics dataset [Carreira et al. CVPR17]



Recent CNN methods

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

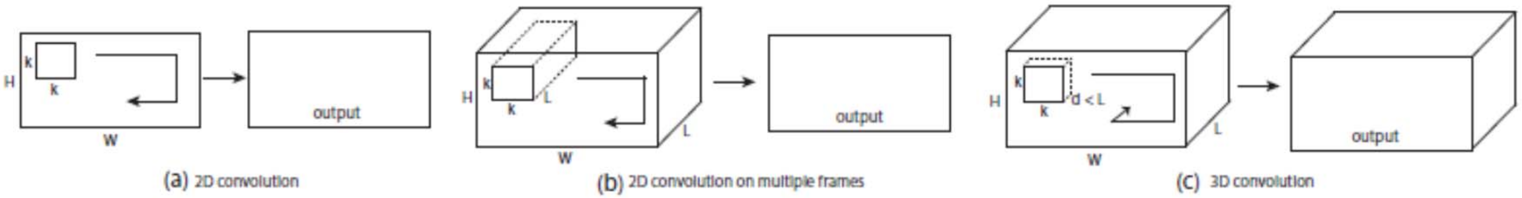
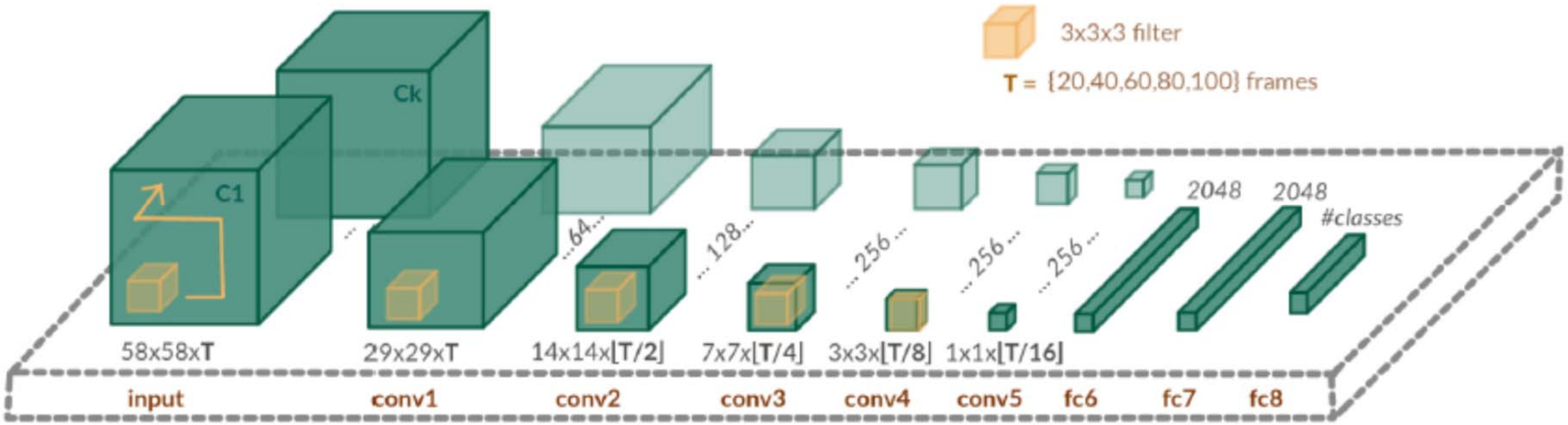


Figure 1. **2D and 3D convolution operations.** a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving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

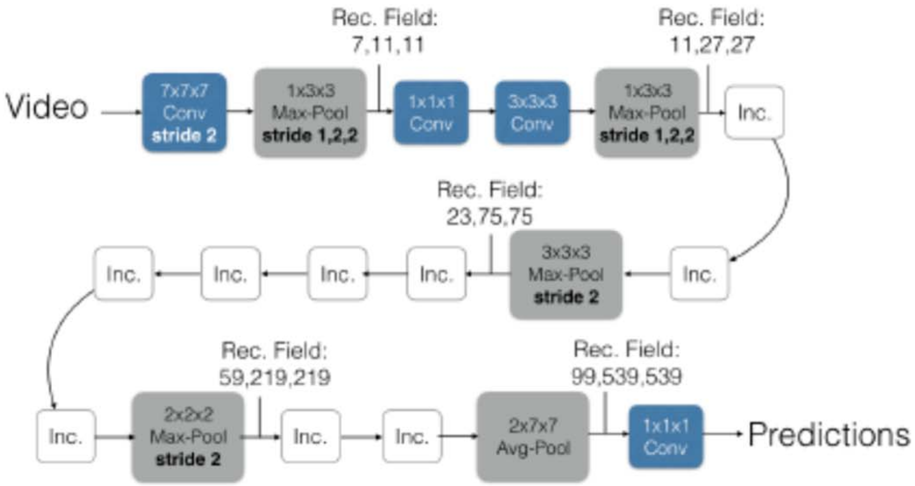


[redacted]

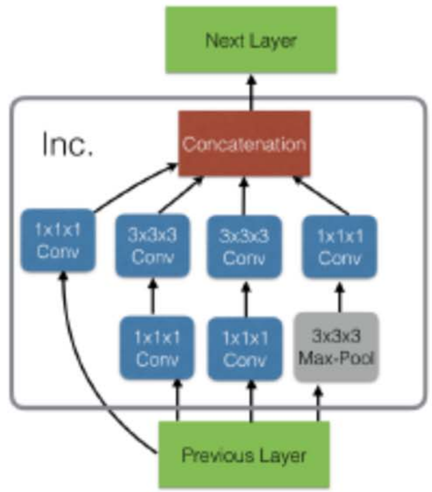
Recent CNN methods

Quo vadis, action recognition? A new model and the Kinetics dataset
[Carreira et al. CVPR17]

Inflated Inception-V1



Inception Module (Inc.)

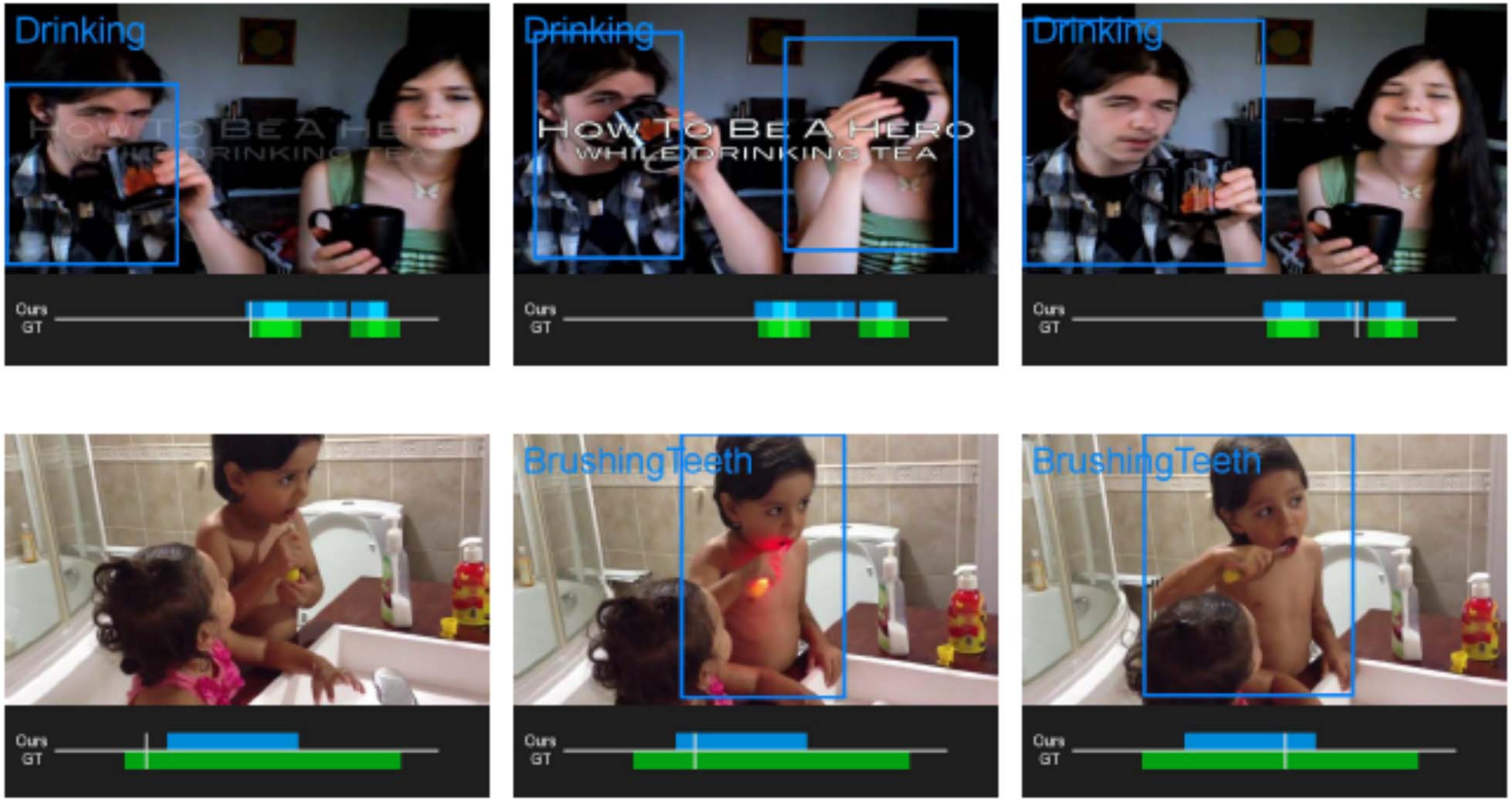


Pre-training on the large-scale Kinetics dataset 240k training videos
→ significant performance grain

Overview

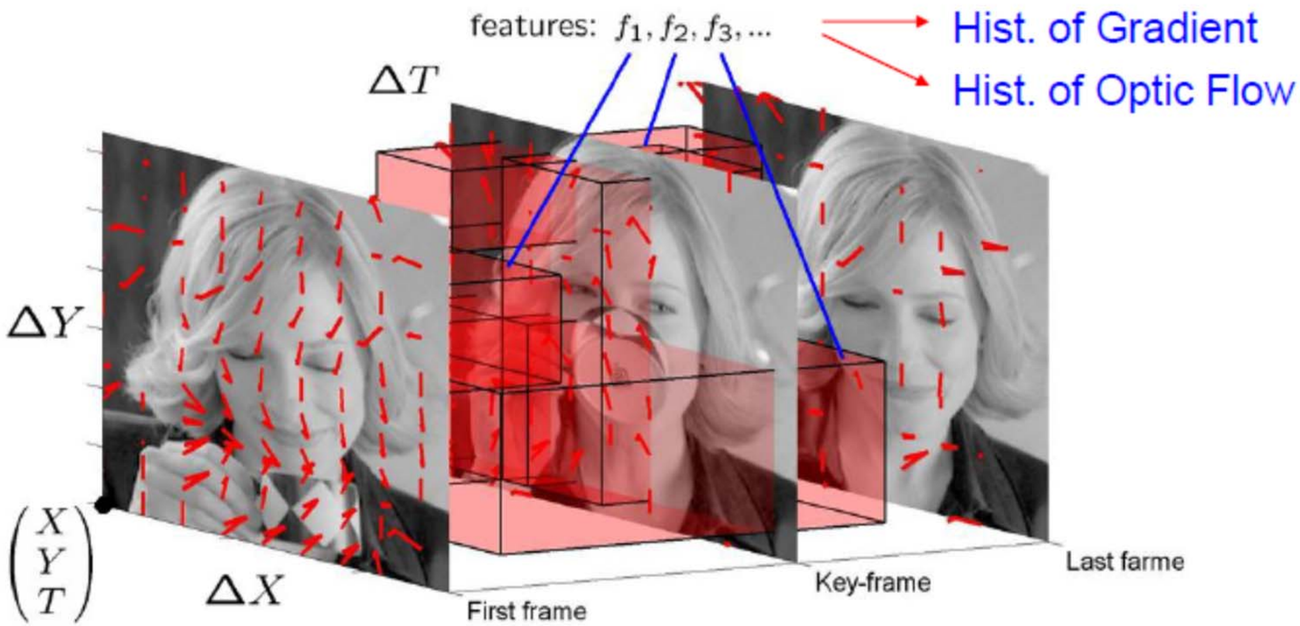
- Optical flow
- Video classification
  - Bag of spatio-temporal features
- *Action localization*
  - *Spatio-temporal human localization*

Spatio-temporal action localization

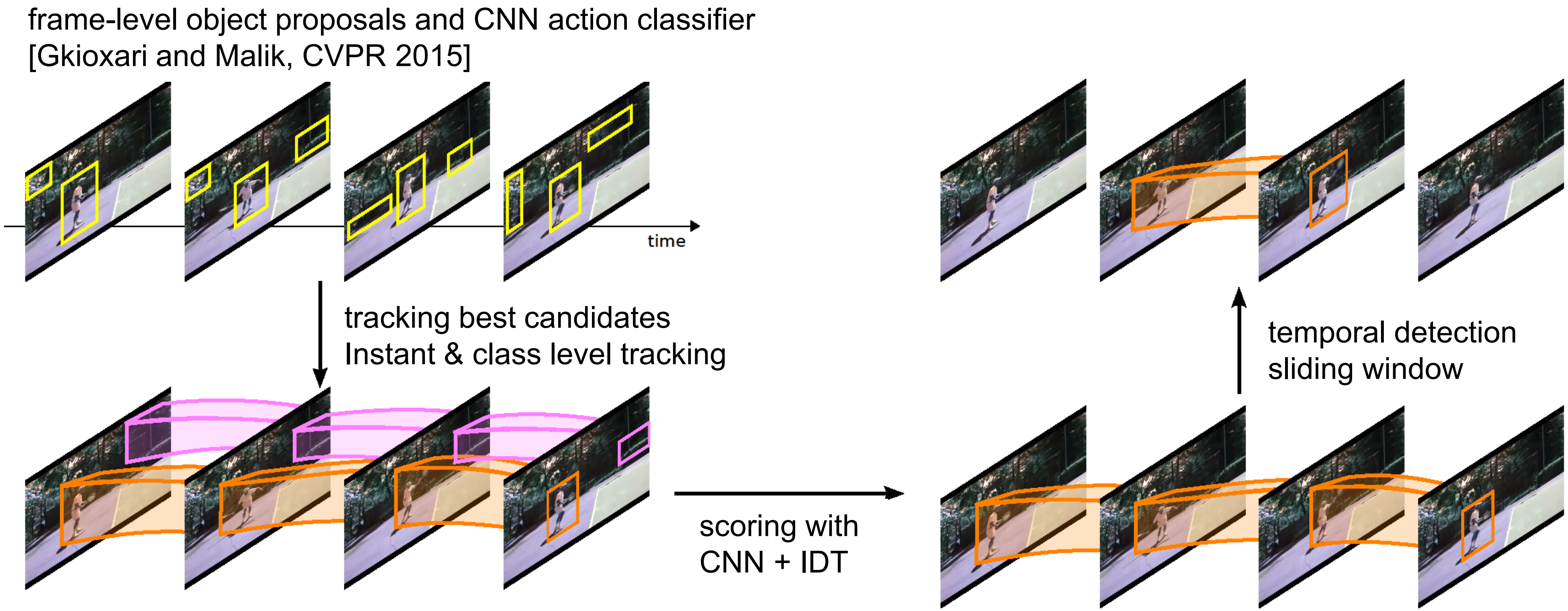


Initial approach: space-time sliding window

- Spatio-temporal features selection with a cascade [Laptev & Perez, ICCV'07]



Learning to track for spatio-temporal action localization

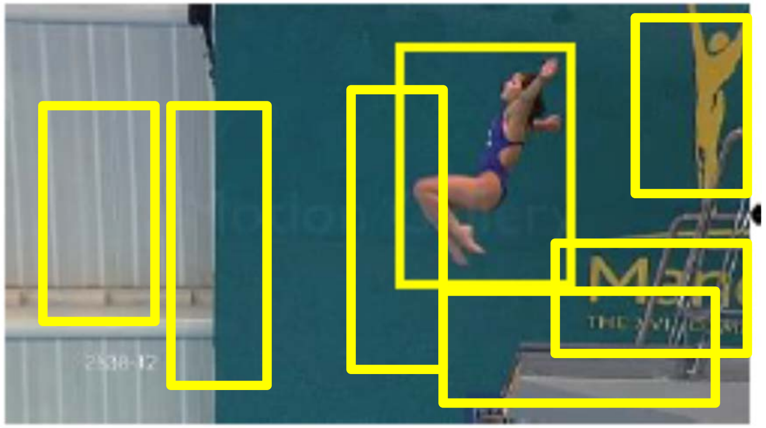


informatics mathematics
inria

[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]

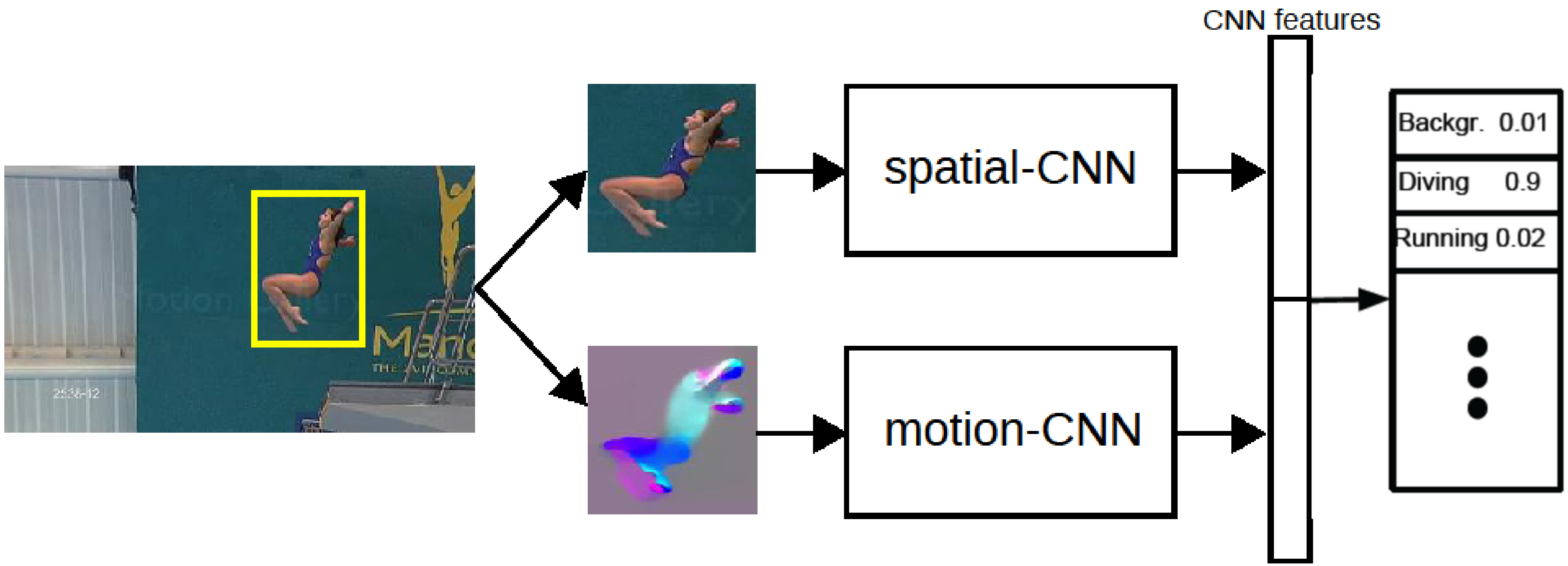


Frame-level candidates

- For each frame
  - Compute object proposals: EdgeBoxes [Zitnick et al. 2014]
  - Extraction of salient boxes based on edgeness

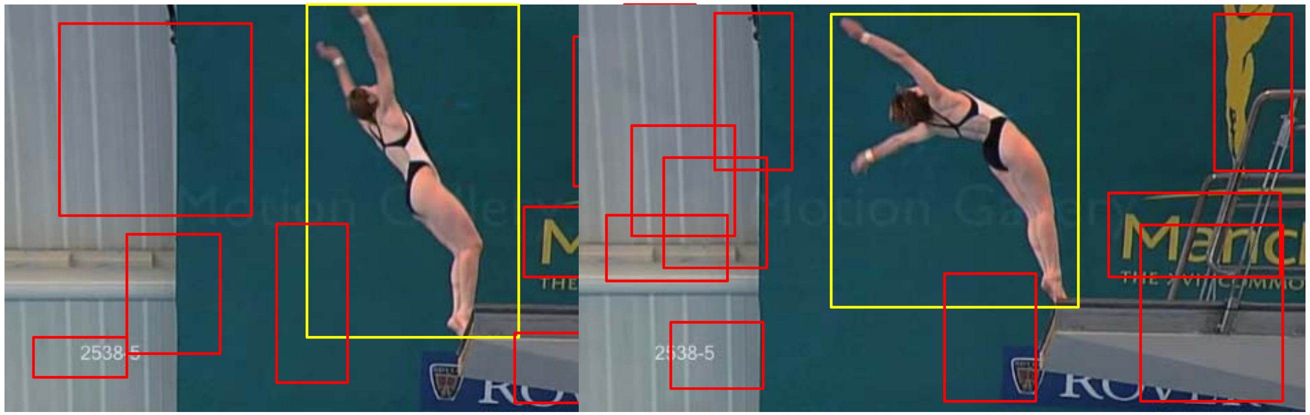
Frame-level candidates

- 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]

Extracting action tubes - tracking



- Tracking an action detection (select highest scoring proposal)
  - Learn an instance-level detector
    mining negatives in the same frame
  - 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

Extracting action tubes

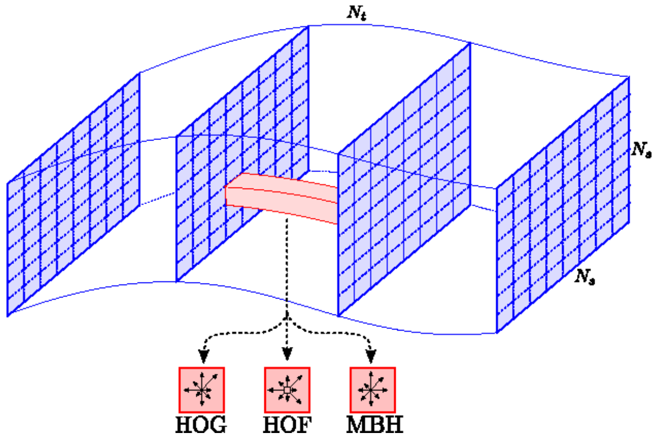
- Start with the highest scored action detection in the video
- Track forward and the backward
- Once tracking is done, delete detections with high overlap
- Restart from the highest scored remaining action detection

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

- Instance-level → models specific appearance

Rescoring and temporal sliding window

- To capture the dynamics
  - ► Dense trajectories [Wang et Schmid, ICCV'13]

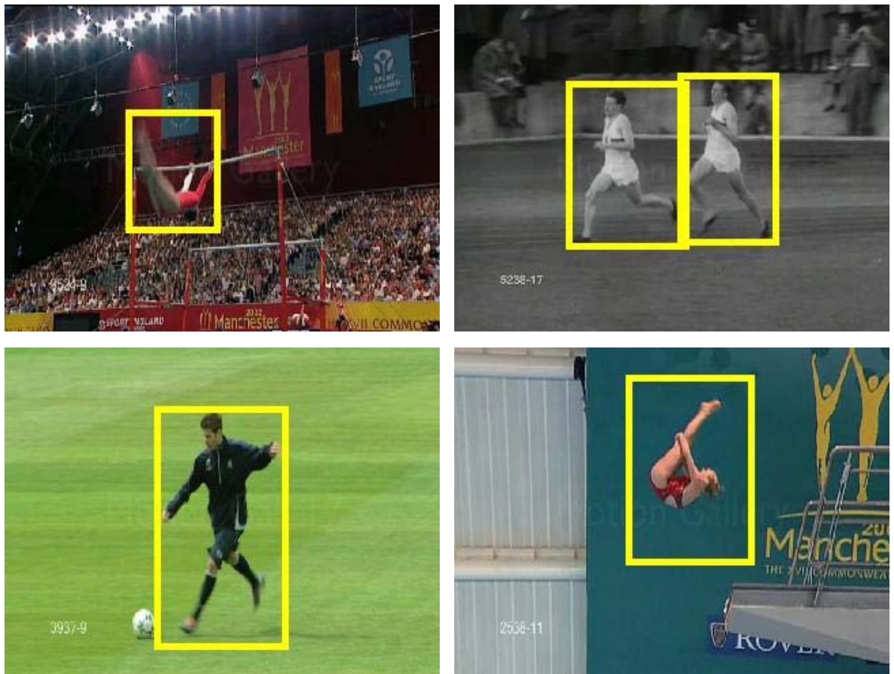


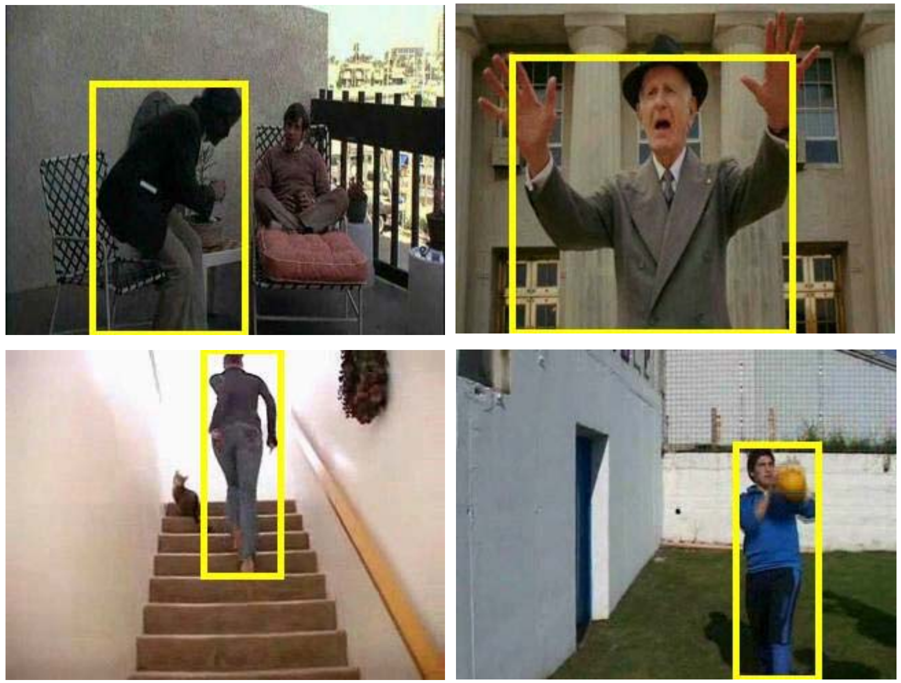
- Temporal sliding window

detection

Datasets (spatial localization)

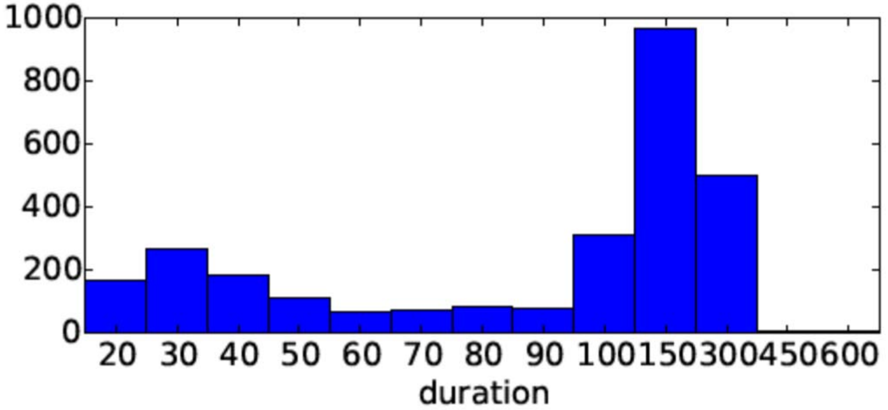
	UCF-Sports [Rodriguez et al. 2008]	J-HMDB [Jhuang et al. 2013]
Number of videos	150	928
Number of classes	10	21
Average length	63 frames	34 frames

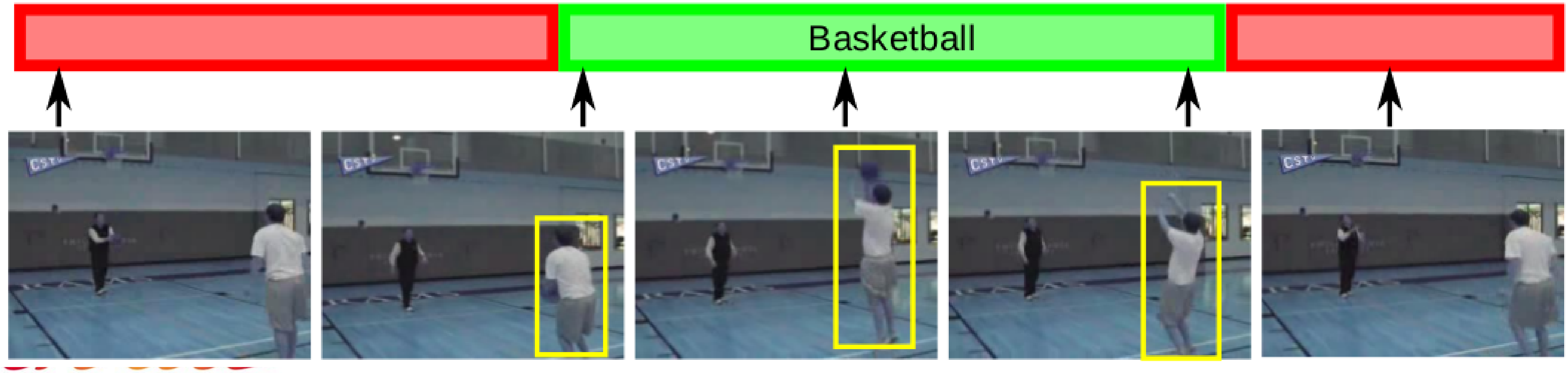




Datasets

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





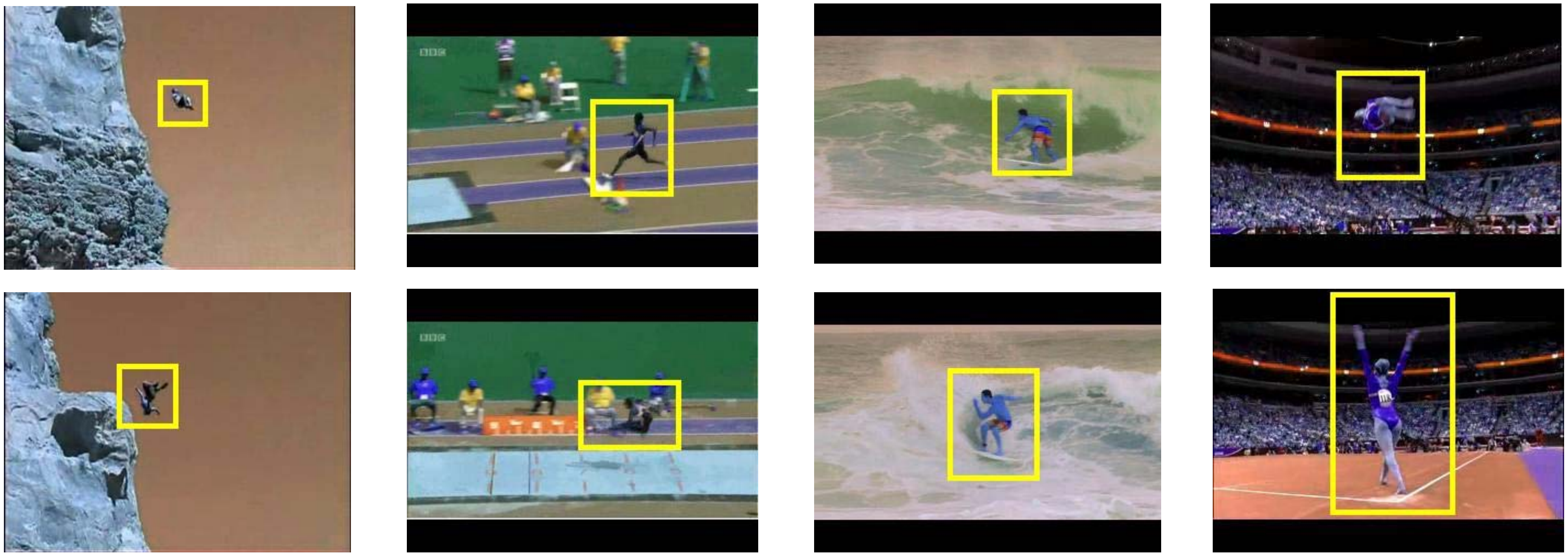
Experimental results

Impact of the tracker

Detectors in the tracker	mAP	
	UCF-Sports	J-HMDB (split 1)
instance-level + class-level	**95.1%**	**65.0%**
instance-level	77.5%	61.1%
class-level	91.0%	60.6%
Comparison to the state of the art		
Gkioxari & Malik, 15	75.8%	53.3%

Quantitative evaluation on UCF-101

mAP	0.2	0.3
Ours	46.7	37.8



Spatio-temporal action localization