



# Efficient feature extraction, encoding and classification for action recognition

Vadim Kantorov, Ivan Laptev  
INRIA – WILLOW / École Normale Supérieure, Paris, France

## Goal

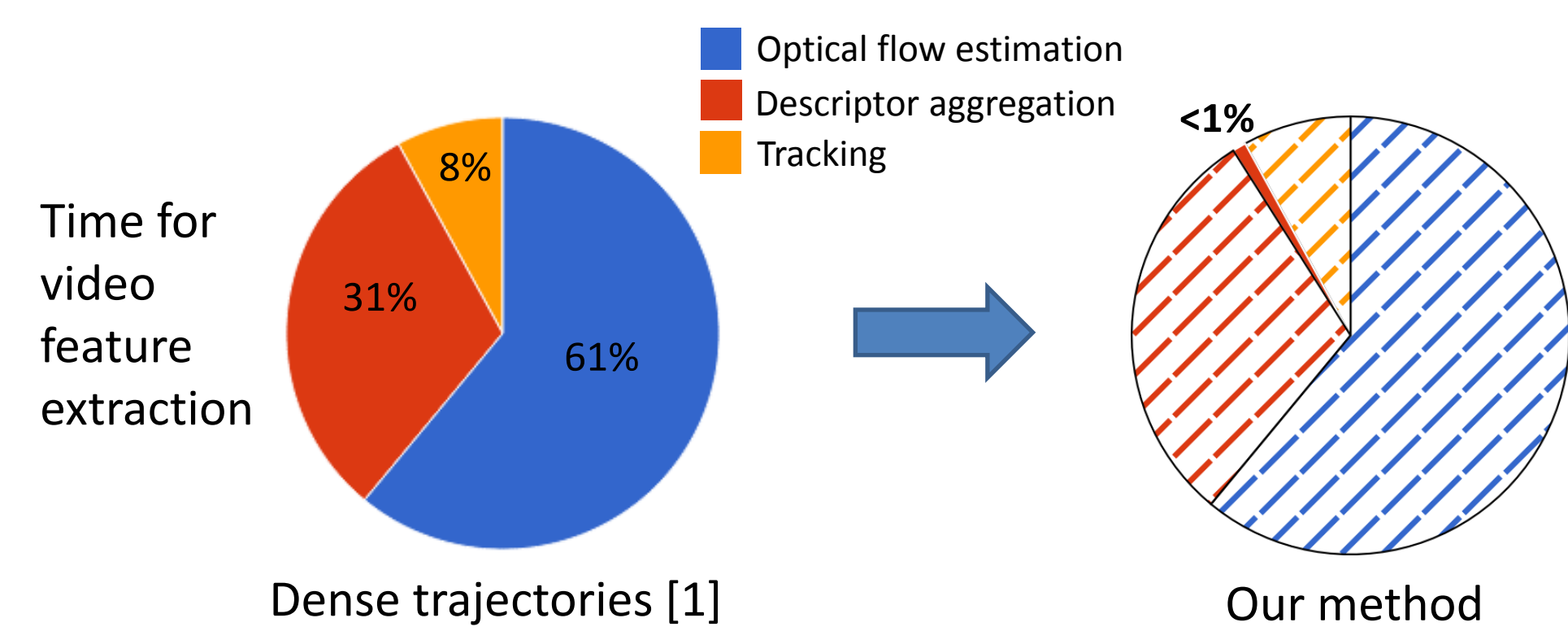
- Fast action recognition.
- State-of-the-art performance.

## Motivation

- Huge amounts of video:
  - Decades of TV channels
  - 6000 years of new video each year
  - 5M years of video transfer per month in 2018
- Large-scale applications:
  - Video indexing
  - Surveillance
  - Augmented reality
- Current state-of-the-art methods for action recognition typically process  $\approx 1$  frame per second

## Contributions

- >100x speed-up of video feature extraction.



- 4x real-time action recognition (CPU).
- Minor decrease in recognition accuracy.
- Publicly available implementation  
<http://www.di.ens.fr/willow/research/fastvideofeat>



## Related work

- H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. IJCV, 2013.
- F. Shi, E. Petriu, and R. Laganier. Sampling strategies for real-time action recognition. In CVPR, pages 2595–2602, 2013.
- F. Perronnin and J. Sanchez. High-dimensional signature compression for large-scale image classification. In CVPR, 2012.
- M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In VISSAPP, pp. 331–340, 2009.

## MPEG flow

- Estimated **motion vectors** are part of the most compressed video representations: MPEG, H-264, VP9.
- MPEG motion vectors are sparse, typically defined on a 16x16 pixel grid.
- The quality of MPEG flow is comparable to motion estimation by standard Optical Flow algorithms.

### Motion in the synthetic MPI Sintel Flow dataset:



Quantized ground truth flow

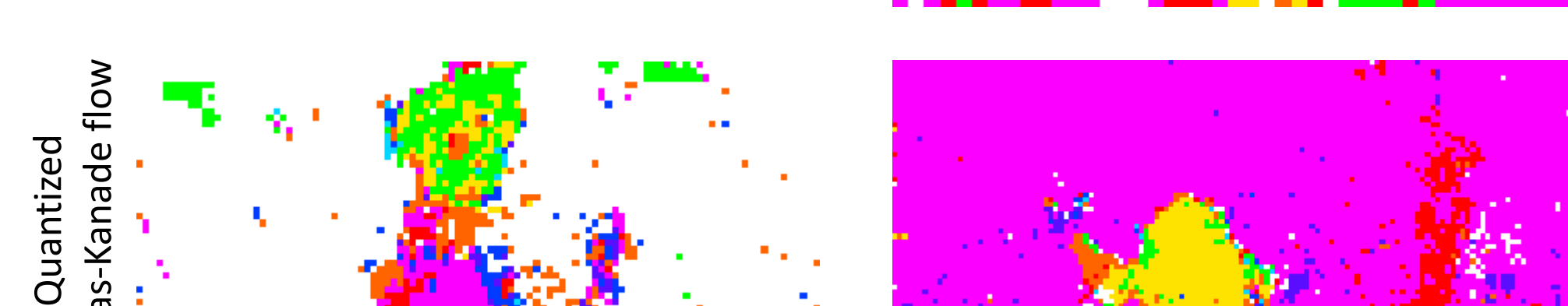
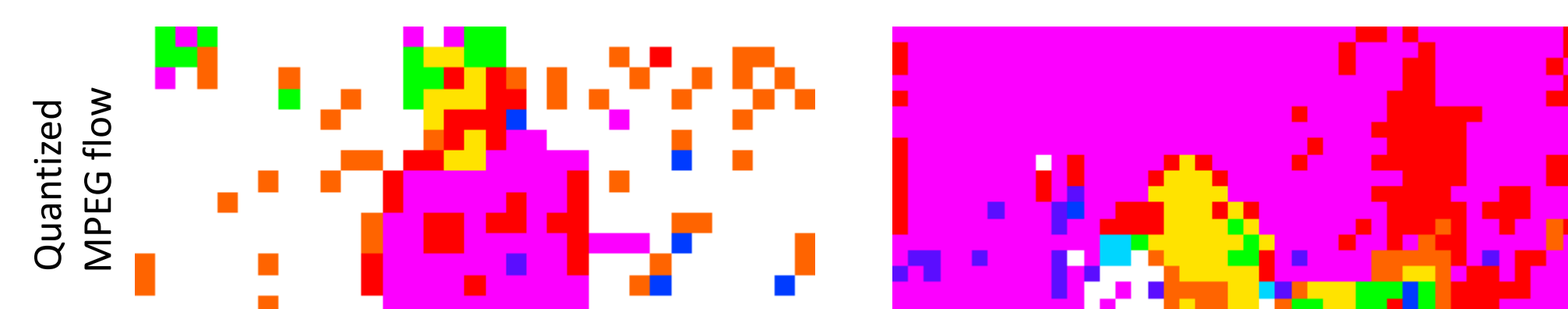
Quantized MPEG flow, err=0.283



Quant. LK flow, err=0.334

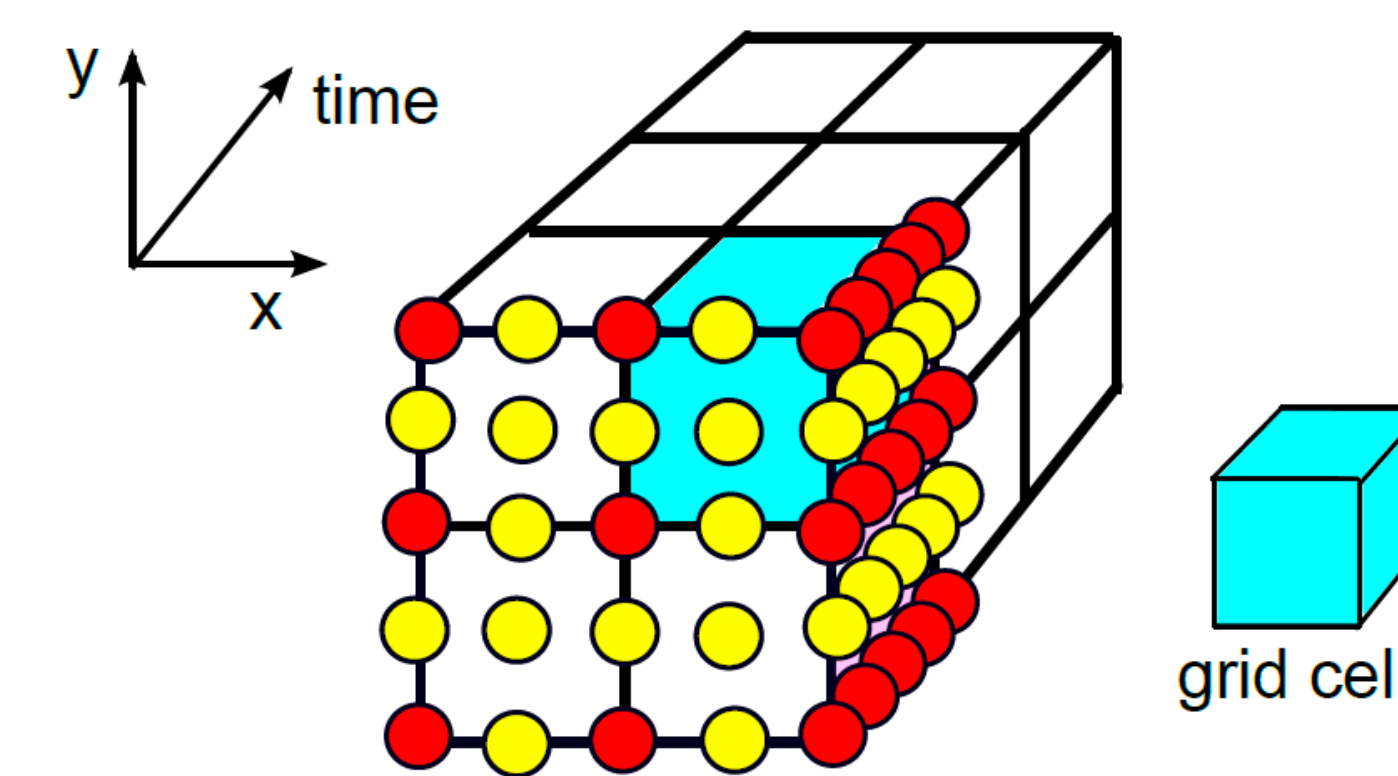
Quant. Farnebäck flow, err=0.286

### Motion in movie frames:



## Approach

### Local motion descriptor



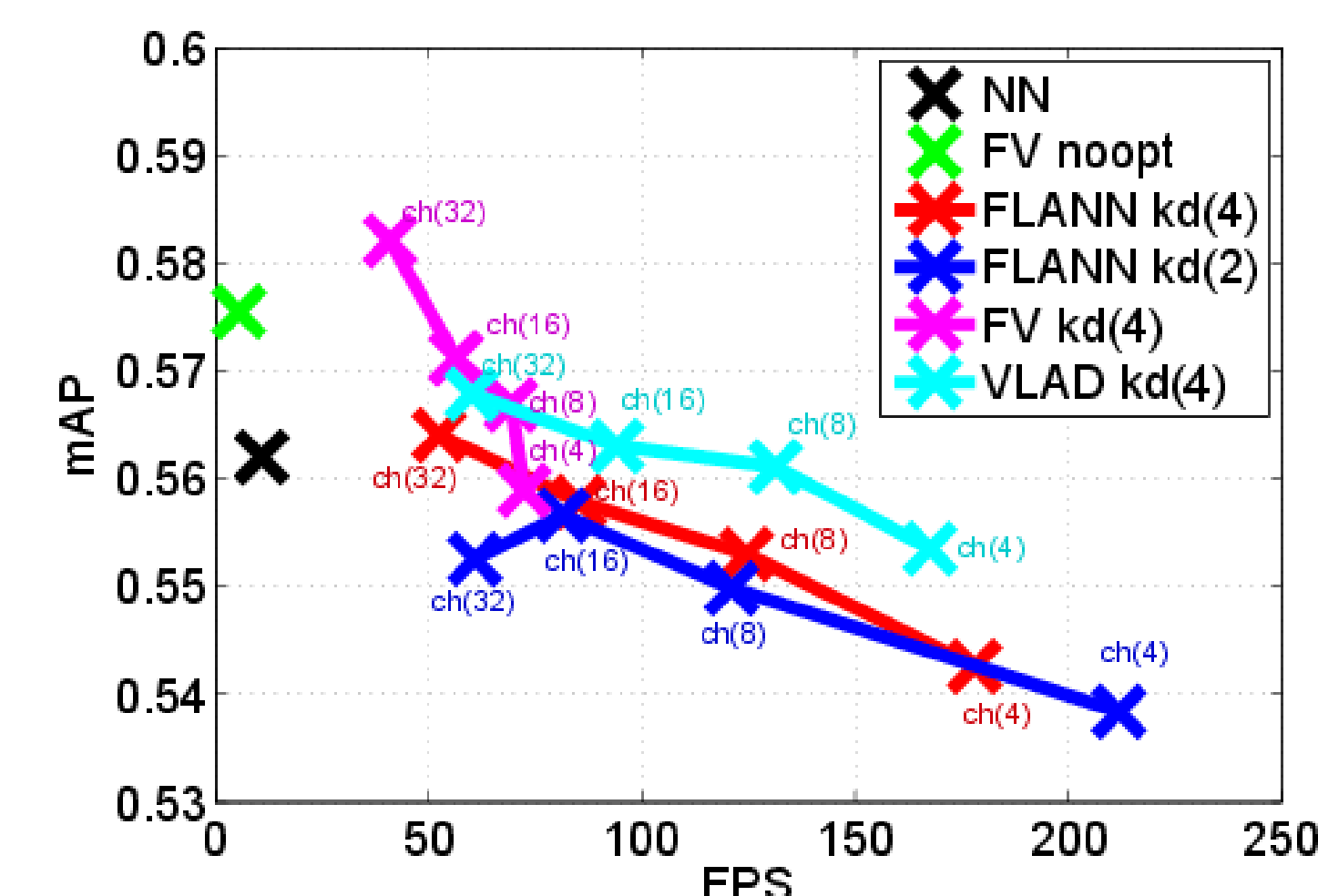
- MPEG flow vectors ( $v_x, v_y$ )
- Interpolated flow vectors ( $\tilde{v}_x, \tilde{v}_y$ )

- Use sparse MPEG flow vectors to compute
  - HOF**: Histograms of flow
  - MBH**: Motion boundary histograms

- Grid cells of two scales:
  - 16x16 pixels, 5 frames
  - 24x24 pixels, 5 frames
- Dense descriptor sampling with
  - 16 pixels spatial stride
  - 5 frames temporal stride

### Descriptor aggregation

- Feature encoding and classification schemes:
  - Histogram** encoding +  $\chi^2$  kernel SVM
  - VLAD** + linear SVM
  - Fisher Vector** [3] + linear SVM
- Descriptor assignment using approximate Nearest Neighbor search (FLANN) [4].
- Approximate FV aggregation with updates of five nearest centroids only.

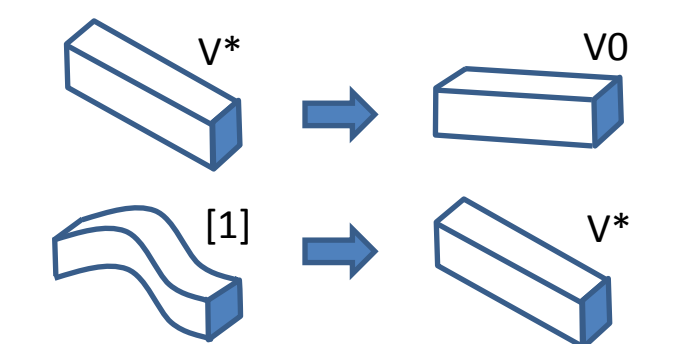


## Results

### Descriptor evaluation

Hollywood2 Histogram encoding	Classification (mAP)		Speed (fps)	
	MF (our)	DT [1]	MF (our)	DT [1]
HOF	47.2%	52.9%	346.8	
MBHx	49.0%	52.0%	330.3	
MBHy	50.4%	56.1%	330.3	
HOF+MBHx+MBHy	53.9%	58.9%	218.7	
HOF+MBHx+MBHy+HOG	56.2%	60.0%	168.4	1.2

	mAP
HOF+MBHx+MBHy+HOG (V0)	58.0%
HOF+MBHx+MBHy+HOG (V*)	58.9%
HOF+MBHx+MBHy+HOG [1]	60.0%
HOF+MBHx+MBHy+HOG+TRAJ [1]	60.3%



➔ Trajectory information has limited influence on results

### Parameter sensitivity

Sampling stride	mAP	fps
16	58.3%	35.2
8	58.6%	24.1
4	59.2%	13.7

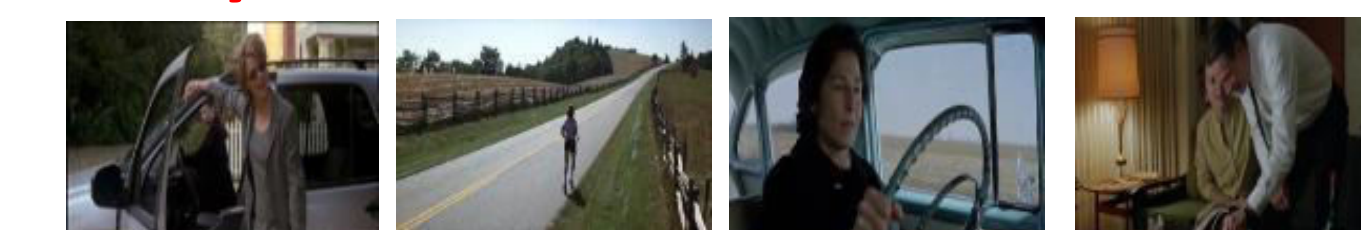
➔ OF stride marginally affects accuracy

	xvid	x264
5000 kbit/s	58.9%	57.5%
1000 kbit/s	58.2%	57.4%
500 kbit/s	57.7%	57.1%
250 kbit/s	57.7%	57.0%

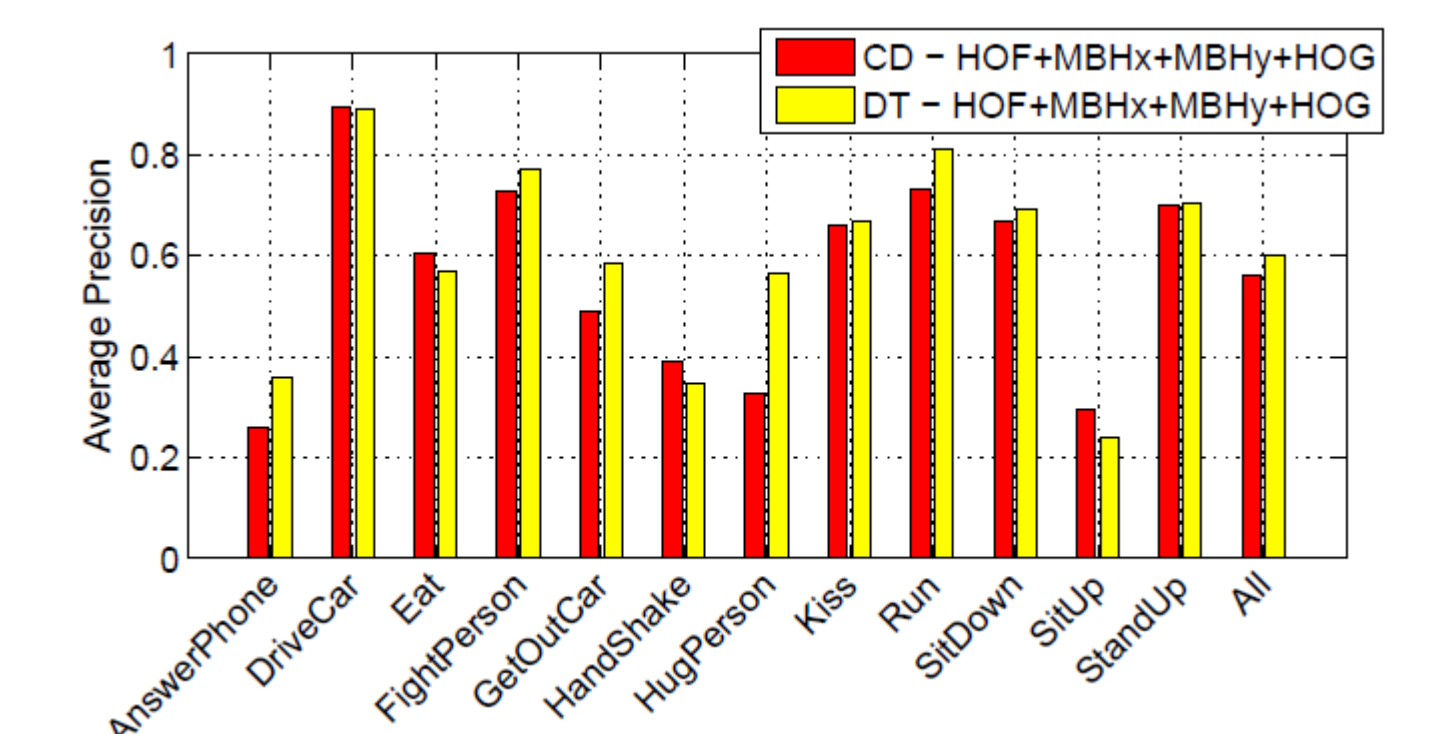
➔ Stable recognition across codecs and bit-rates

### Comparison to the state of the art

#### Hollywood 2



	Acc.	Feat. (fps)	Quant. (fps)	Total (fps)
MF FLANN(4-32)	55.8%		52.4	40.0
MF VLAD(4)	56.7%	168.4	167.5	84.0
MF FV(32)	58.2%		40.9	32.9
DT [1]	59.9%	1.2	5.1	1.0

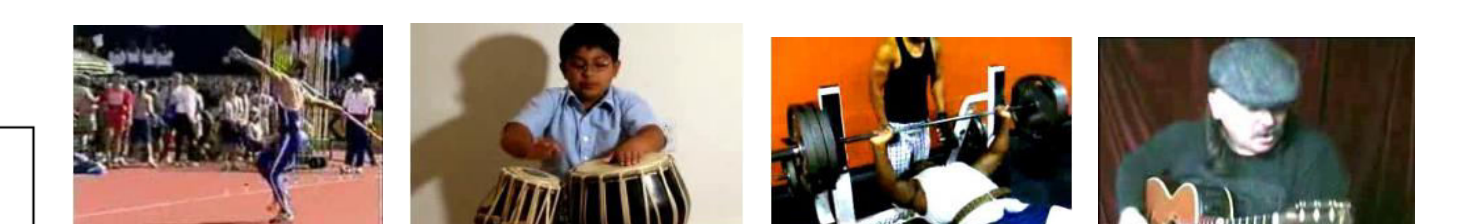


#### HMDB 51



	Acc.	Feat. (fps)	Quant. (fps)	Total (fps)
MF ALL FV(32)	46.7%	455.6	129.7	101.0
MF MBH FV(32)	45.4%	683.3	268.0	192.5
MF ALL VLAD(32)	46.3%	455.6	455.6	227.8
MBH [2]	41.1%	33.9	267.1	30.8
HOG3D [2]	33.3%	49.6	290.8	42.2
DT [1]	48.3%	3.1		

#### UCF 50



	Acc.	Feat. (fps)	Quant. (fps)	Total (fps)
MF FLANN(4-32)	81.6%		52.4	48.1
MF VLAD(4)	80.6%	591.8	671.4	314.6
MF FV(32)	82.2%		171.3	132.8
DT [1]	85.6%	2.8	5.1	1.8

➔ Code available <http://www.di.ens.fr/willow/research/fastvideofeat>