Action modelling and recognition in videos

Antoine Manzanera

U2IS / Robotics and Vision
ENSTA-Paris

Master 2
IP Paris - Univ. Paris-Saclay

Action Modelling and Activity Understanding

Context

Automatic recognition of gesture / action / activity by video analysis.

From [Laptev 13]

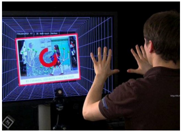
Focus of this lecture

- Action modelling: Extract relevant action features from the video flow.
- Action recognition: Classify a video block with respect to known action classes.

Applications

- Video retrieval (summarization, indexing)
- Video surveillance (assistance)
- Biomedical imaging (gait, flight,...)
- Human machine interaction (gesture control)



Challenges

- Huge variability (appearance, geometry)
- Moving camera

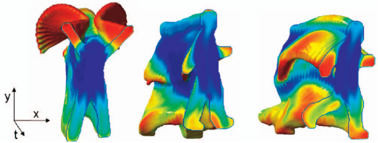
Presentation Outline

Presentation Outline

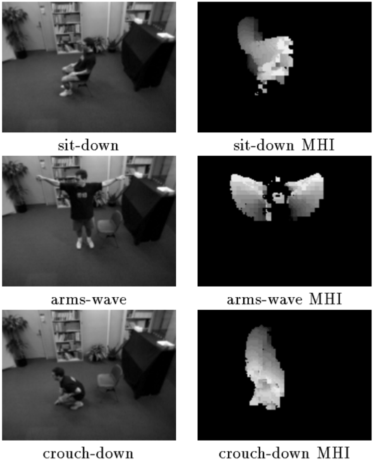
Global Action Features

The action may be modelled using *geometric features* from a *global pattern* obtained by *segmentation* of the moving objects. Examples:

- Action → 2d image **[Bobick 96]**
- Action → 3d shape **[Gorelick 07]**



from [Gorelick 07]

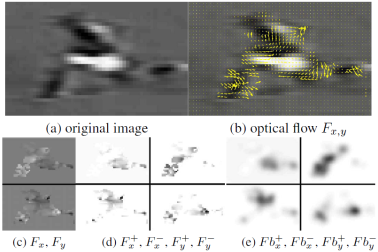


from [Bobick 96]

Velocity-based Action Features

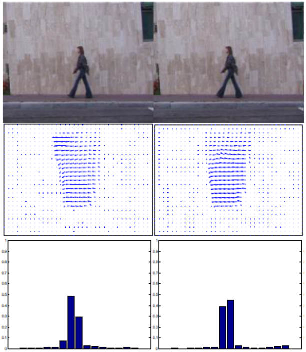
Some models are built from velocity field (optical flow). For example:

- **[Efros 03]** computes grey level patterns from velocity measures.



from **[Efros 03]**

- **[Chaudhry 09]** uses histograms of optical flow orientations as action descriptor.

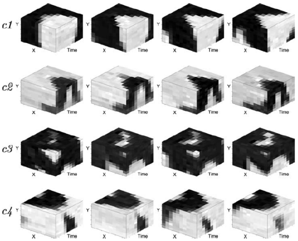


from **[Chaudhry 09]**

Local Action Features - Overview

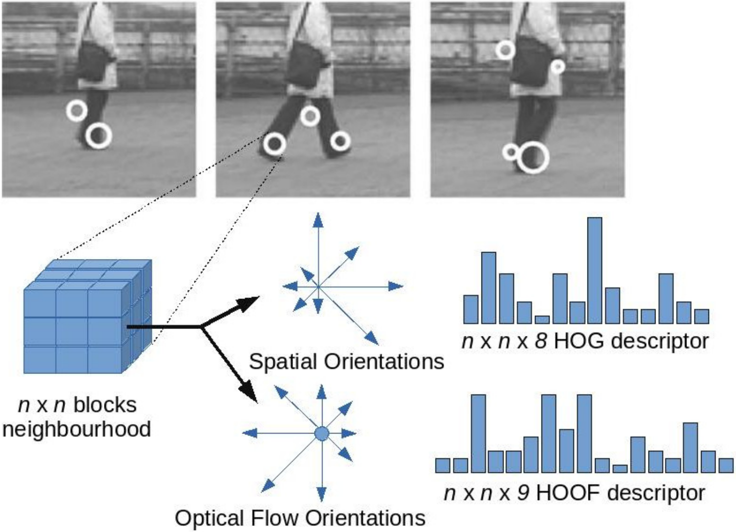
Some action representations are made from a *collection* of features calculated on a set of *space* $\times$ *time salient* points. For example **[Laptev 05]**:

- detects space $\times$ time 3d Harris corner points (STIP)
- describes them using a local (patch) descriptor
- quantises the local descriptor space to form a code book
- describes an action by the code book occurence histogram over the sequence



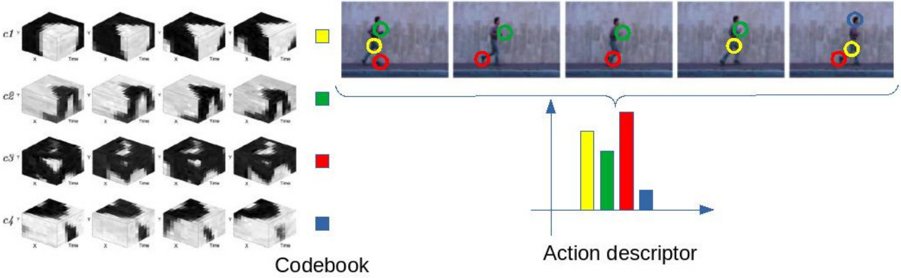
from [Laptev 05]

Local Action Features - Local (STIP) descriptor



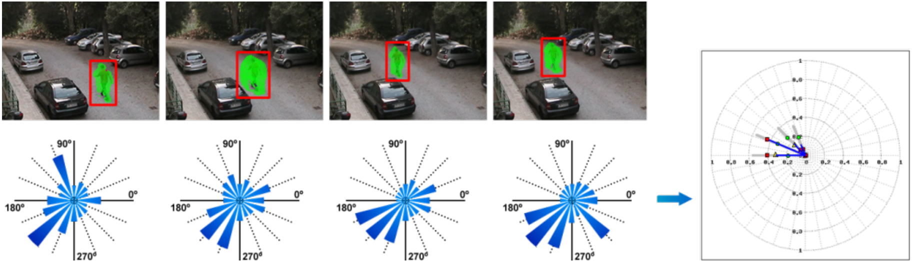
Local Action Features - Global (Action) descriptor

- The space of local descriptors is quantised using a clustering (e.g. K-means) algorithm to form a Code book.
- The action is represented by a Code book Histogram, that counts the number of occurences of each word along the sequence.



Online Action Modelling using Optical Flow Statistics

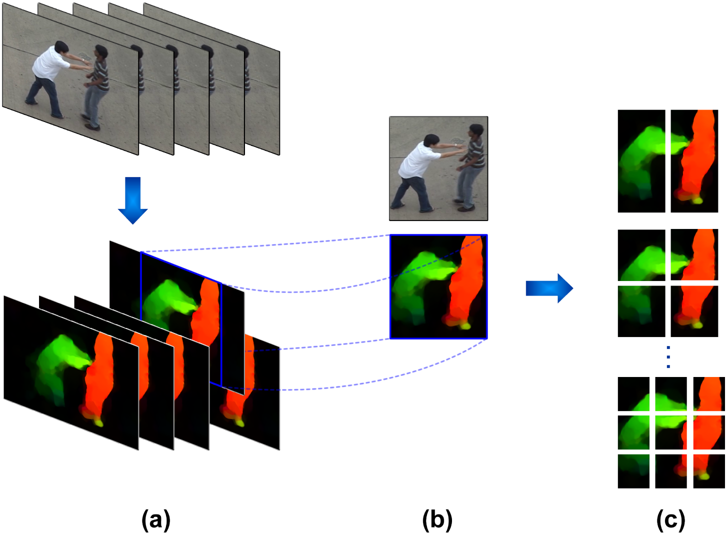
[Martínez 12] computes recursive temporal statistics (average, min, max, variance) of the spatial histogram values of optical flow orientations.



(Left) Orientation histograms, calculated on each frame within the Region of Interest, represent the spatial distribution of instantaneous velocities.
(Right) The action descriptor is obtained by recursively calculating a set of temporal statistics on each bin of the velocity orientation histogram.

See video examples

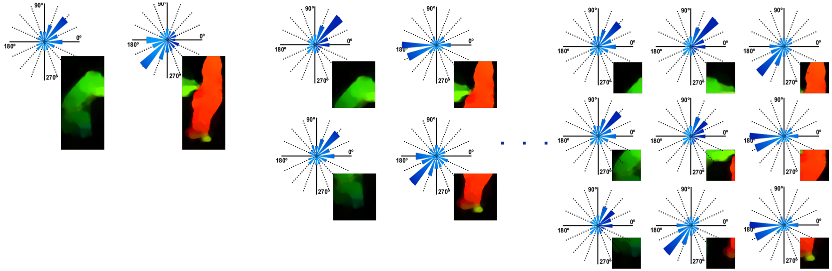
Multiscale version (1: spatial decomposition)



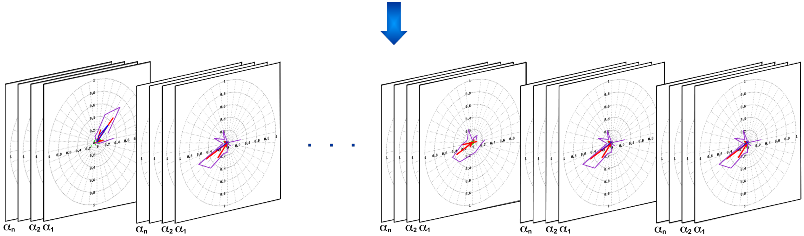
From [Martínez 17]

Multiscale version (2: temporal decomposition)

(a)



(b)



From [Martínez 17]

The apparent trajectory of a moving point can be used to represent gesture, action or activity.

Pros

- Compact
- Large temporal depth
- Appearance invariant
- Facilitates segmentation

J.E. Marey *Mouvement*

(Chronophographie) - 1882

Cons

- Sparse
- Fragile
- Noisy
- Costly

Trajectory beam with semi-dense tracker *Video extruder*

Optical Flow

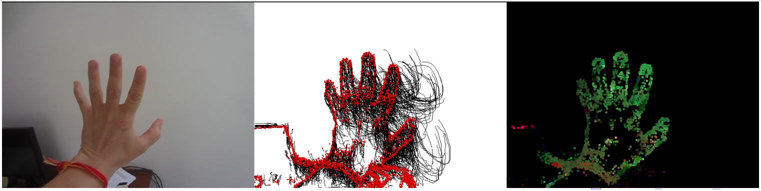
- Temporally *short term*
- Spatially *dense*
- Main computational load: *Spatial regularisation*

Point tracker

- Temporally *long term*
- Spatially *sparse*
- Main computational load: *Spatial characterisation*

Video Extruder

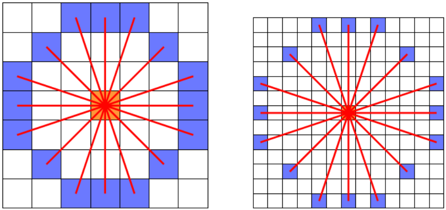
- Temporally *long term*
- Spatially *semi-dense*
- Weak spatial characterisation
- Minimal spatial regularisation



Video extruder: Weak spatial selection

Weak keypoint selection

- Principle: discarding only points whose matching will be ambiguous at all computed scales.
- Saliency measure at one scale:
  $\Sigma_s(\mathbf{p}) = \min_{i=0}^{7} |2I(\mathbf{p}) - I(\mathbf{q}_i) - I(\mathbf{q}_{i+8})|$
- Multi-scale saliency: $\Sigma = \max_{s \in S} \Sigma_s$
- Fast computation of detector and descriptor (Bresenham circles).

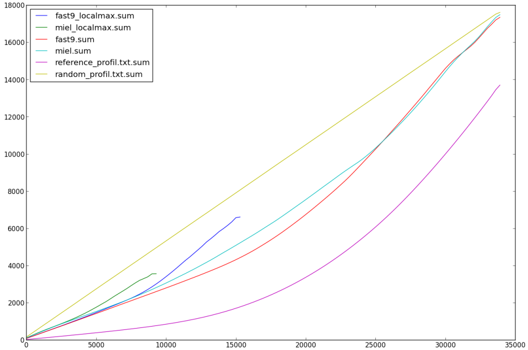


Multi-scale keypoint supports: Bresenham circles

Video extruder: Weak spatial selection

- Block-wise maxima: 2 or 3 times more points than local maxima

  - Geometric selection is better than arbitrary selection (brown curve) up to 10% of the image surface.
  - Different detectors on the same support perform similarly, and far from ideal detector (purple curve).



Keypoint selection evaluation: total error *vs* number of keypoints.

Video extruder: Tracking algorithm

Pyramidal tracking algorithm

- Coarse-to-fine prediction, based on:
  - Point velocity (temporal)
  - Regional dominant motion (spatial)
- Gradient descent based matching.
- Elimination of incoherent points and merging of redundant points.

Comparison with Pyramidal LKT (OpenCV)

- Similar tracking quality.
- Faster from $\times 2$ to $\times 15$ (depending on LKT parameters).

Video extruder: Benchmarking

Thanks to its high level of parallelism and regularity, *Video extruder* can run in real-time on many low-end embedded platforms.

Architecture	Resolution	# points	Freq. (Hz)	# Cpp
GPU Geforce GTX 460 1.35GHz	640 × 480	8 500	166	957
CPU quad-core I5 2500k 3.3GHz	640 × 480	8 500	152	2 550
ARM dual-core STE U8500 1GHz	320 × 240	3 000	11	30 300
ARM single-core IMX.53 1GHz	720 × 288	2 000	10	50 000

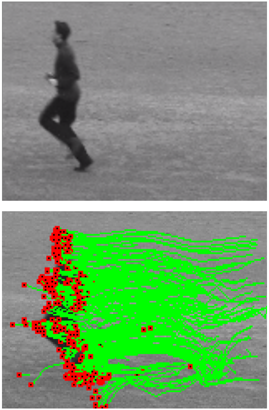
Time performance of *Video extruder* on different architectures.

http://perso.ensta-paristech.fr/~garrigues/video_extruder.html

From Trajectories to Action

hand clapping

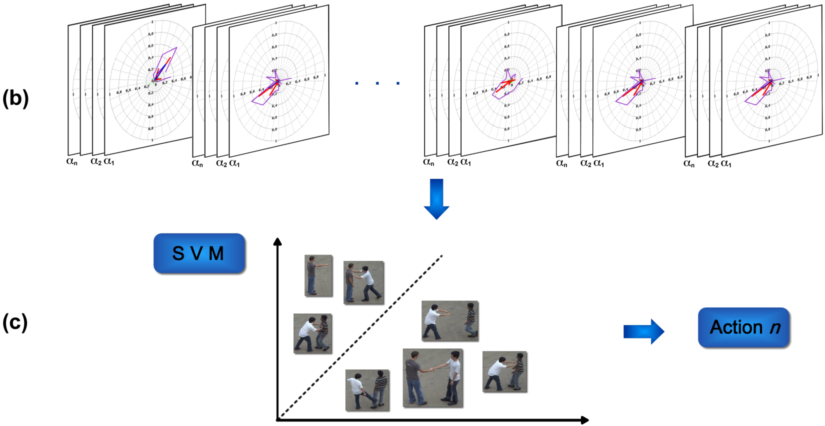
hand waving



running

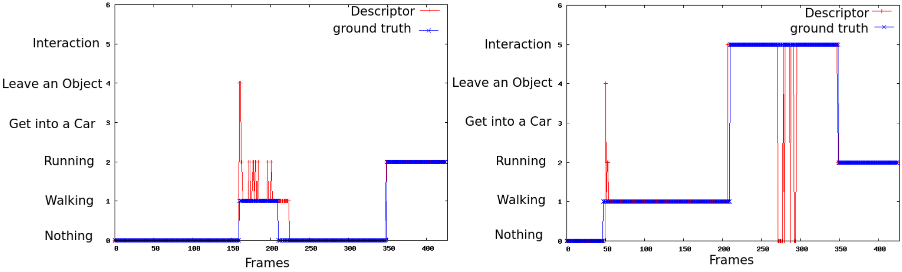
Presentation Outline

Online Multiscale Velocity Orientation Histograms



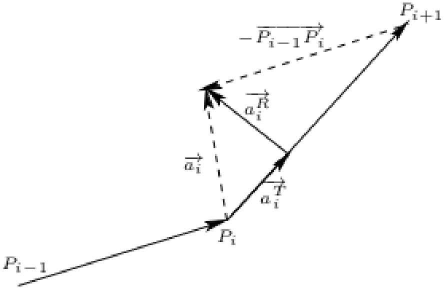
On **[Martínez 17]**, the action descriptor made of the concatenation of VOH, is simply submitted to N linear SVM (with N the number of actions, one-against all method).

Online Multiscale VOH: Classification



The online frame-level classification on **[Martínez 17]** allows temporal filtering of the action labels (example shows two long videos with different actions).

On **[Nguyen 13]**, elementary motion elements (*atomic actions*) are extracted from the trajectories, using *dominant points*, corresponding to *local maxima* of the *radial acceleration* (related to *curvature*), for different temporal scales.

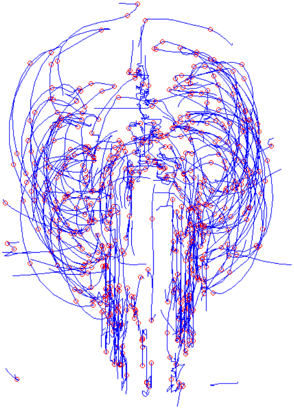


Radial acceleration $\vec{a}_i^R$ on a trajectory

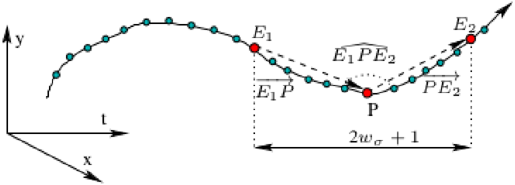
The temporal scale is related to the standard deviation σ of the Gaussian used to smooth the trajectory.

Representation of Atomic Actions on trajectories

Dominant point detection **[Nguyen 13]**



Every dominant point is described using a *feature vector* composed of *geometrical* and *statistical* parameters of the trajectory around the dominant point: angle, curvature, directions, average and variance of speed and accelerations...

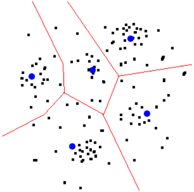


Computation of the feature vector around the dominant point P.

The size of the support depends on the temporal scale σ of the dominant point.

Building a Code Book of Atomic Actions

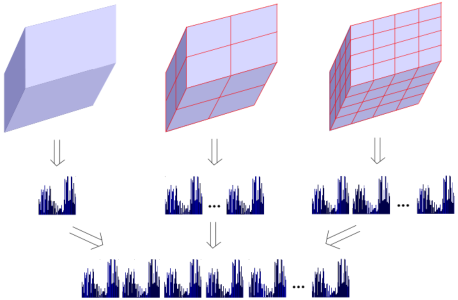
- In a first level (non supervised) learning phase, the feature vectors from a set of actions are vector quantised (K-means algorithm) to form a code book of atomic actions.



- At the run time, every dominant point is classified as an atomic action using a nearest neighbour search.
- The action may then be represented using a classic Bag of Features approach (i.e. distribution of the words from the code book), however the spatiotemporal relations between the atomic actions are crucial to represent a complex action.

Representation of Complex Actions

We represent a complex action by concatenating histograms of atomic actions on a hierarchy of space $\times$ time boxes.

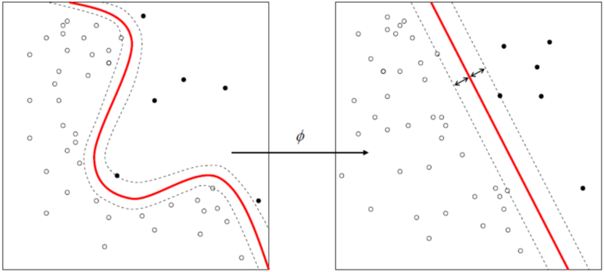


Representation of an action from multiple histograms.

The multiple histogram represents spatiotemporal relations between atomic actions.

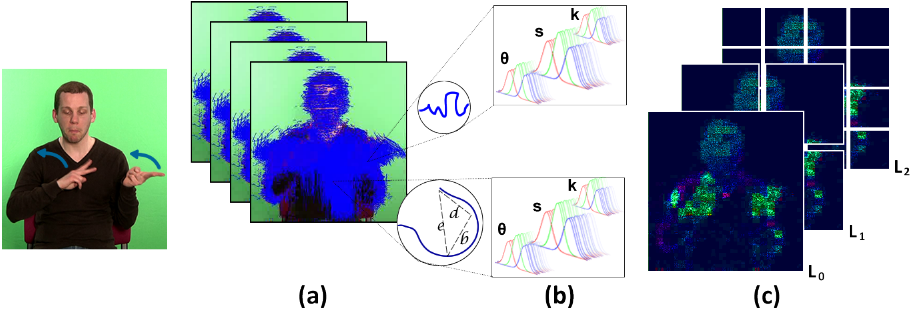
Complex Action Classification

- The second level (supervised) learning phase corresponds to learning a SVM on action descriptors from training sequences.



- At the run time, action classification is performed using *1 vs 1* SVM multiclass classifier.

Online Action Modelling on Trajectories



Online Modelling can be extended to trajectories: On **[Martínez 15]**, a set of kinematic features (orientation, speed, curvature...) is recursively estimated for each trajectory. A kinematic codebook is then used though a multiscale bag-of-words approach.

Background Motion Removal

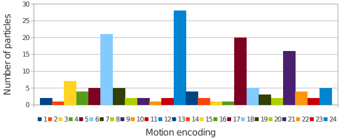
When the camera is moving, many trajectories are due to the relative motion of the background and must be discarded.

Sport sequence from *UCF Youtube dataset*

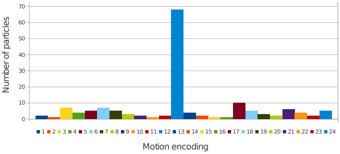
Computed trajectories

Background Removal: Dominant Motion Extraction

- If we suppose the background essentially plane and/or the camera motion is limited to pan/tilt, and if the interest object is not too big, the background motion is associated to the *dominant motion*, calculable by a *cumulative framework* (Figure).
- The framework can be extended to an *affine motion* of the camera $X_{t+1} = A_t X_t + B_t$ **[Jain 13]**.
- The trajectory framework makes the removal *more robust*, by counting the number of times a point has a dominant motion along its trajectory.



No dominant motion



Dominant motion

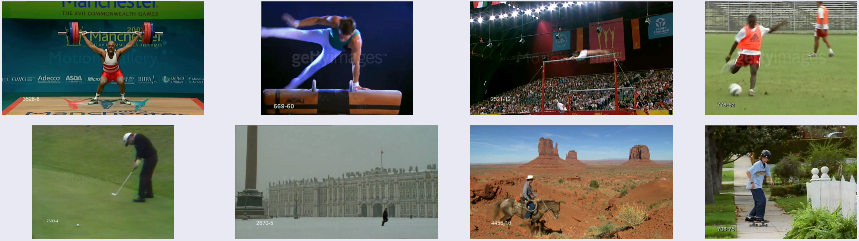
Background Removal: Results

Presentation Outline

Data Bases of segmented videos: from the handcrafted...

KTH [Schuldt 04] 2 391 videos; 6 actions × 25 subjects

UCF Youtube [Liu 09] 800 videos; 11 actions × 25 groups



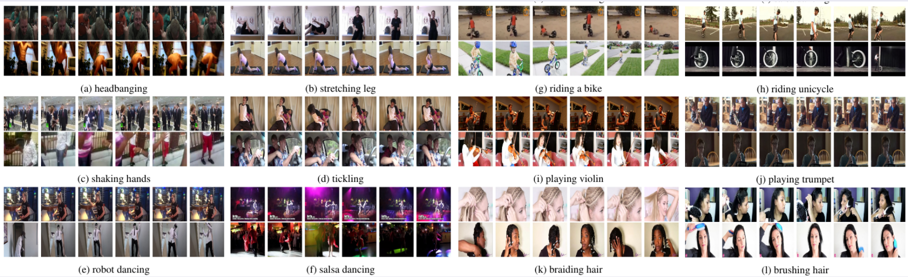
Data Bases of segmented videos

HMDB [Kuehne 11] 6 849 videos; 51 actions

5 types of human actions: (1) General facial actions (smile, laugh, chew, talk...) (2) Facial actions with object manipulation (smoke, eat, drink...) (3) General body movements (cartwheel, clap hands, climb stairs... (4) Body movements with object interaction (brush hair, catch, dribble...) (5) Body movements for human interaction (fencing, hug, kick someone...)

Data Bases of segmented videos: ...to the Big Data era!

Kinetics [Carreira 18] 300 000 videos ($\approx 10s$); 400 actions



Evaluation metrics: Accuracy

The accuracy is the recognition rate, i.e. the number of correct classification divided by the number of predictions made on the test (or validation) set.

Evaluation: UCF101 Eval

Description: Three splits as defined by authors

Results

Show 10 entries Search:

Result	Paper	Description	URL	Peer Reviewed	Year
98.2	PoTion: Pose MoTion Representation for Action Recognition[Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, Cordelia Schmid]	I3D + PoTion	URL	Yes	2018
98.2	Global and Local Knowledge-Aware Attention Network for Action Recognition[Zhenxing Zheng, Gaoyun An, Dapeng Wu, Qiuqi Ruan]	global and local attention + I3D	URL	No	2019
98	Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset[Joao Carreira, Andrew Zisserman]	Two-Stream I3D, Kinetics pre-training	URL	Yes	2017

Best accuracies on UCF 101 Data set (3-fold cross-validation),

from www.actionrecognition.net, University of Bonn

Evaluation metrics: Accuracy

Evaluation: HMDB Eval

Description:

Results

Show 10 entries Search:

Result	Paper	Description	URL	Peer Reviewed	Year
82.48	Hallucinating IDT Descriptors and I3D Optical Flow Features for ActionRecognition with CNNs[Lei Wang, Piotr Koniusz, Du Q. Huynh]	HAF+BoW/FV halluc.	URL	Yes	2019
82.3	Evolving Space-Time Neural Architectures for Videos[AJ Piergiovanni, Anelia Angelova, Alexander Toshev, and Michael Ryoo]		URL	Yes	2019
82.1	End-to-end Video-level Representation Learning for Action Recognition[Jiagang Zhu, Wei Zou, Zheng Zhu, Lin Li]	DTPP (Kinetics pre-training)	URL	No	2017

Best accuracies on HMDB Data set,

from www.actionrecognition.net, University of Bonn

Evaluation metrics: Accuracy

Evaluation: Kinetics-val

Description: Top-1 results for the validation or test set of the Kinetics dataset. Results of the val and test set should be comparable.

Results

Show 10 entries   Search:

Result	Paper	Description	URL	Peer Reviewed	Year
82.8	Large-scale weakly-supervised pre-training for video action recognition[Deepti Ghadiyaram, Matt Feiszli, Du Tran, Xueting Yan, Heng Wang, Dhruv Mahajan]	WeakLargeScale (RGB)	URL	No	2019
82.6	Video Classification with Channel-Separated Convolutional Networks[Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli]	CSN on RGB	URL	Yes	2019
79.4	Attention Clusters: Purely Attention Based Local Feature Integration for Video Classification[Xiang Long , Chuang Gan , Gerard de Melo , Jiajun Wu , Xiao Liu , Shilei Wen]	Attention Cluster (RGB + Flow + Audio)	URL	No	2017

Best accuracies on Kinetics Data set (Validation or Test set),

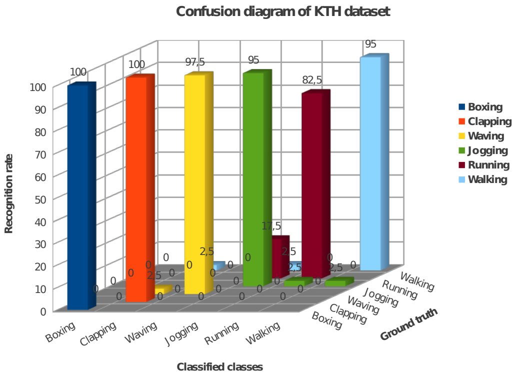
from www.actionrecognition.net, University of Bonn

Evaluation metrics: Confusion Matrices

	bb	bk	dv	gf	rd	sc	sw	tn	tp	vb	wk
basketball	46.2	0	9.6	1.9	0	1.9	0	17.3	5.8	17.3	0
biking	0	51.9	3.7	0	18.5	0	0	7.4	0	0	18.5
diving	0	0	73.3	6.7	0	1.7	5.0	3.3	5.0	3.3	1.7
golf	0	0	4.0	82.0	0	0	2.0	12.0	0	0	0
riding	1.2	2.3	0	0	91.9	0	0	1.2	2.3	1.2	0
soccer	0	3.5	5.3	0	5.3	66.7	14.0	0	5.3	0	0
swing	2.2	0	2.2	6.7	15.6	2.2	57.8	0	6.7	2.2	4.4
tennis	5.1	1.7	1.7	13.6	8.5	6.8	0	59.3	0	1.7	1.7
trampoline	0	0	2.2	0	2.2	0	6.7	0	82.2	0	6.7
volleyball	10.0	0	10.0	5.0	2.5	0	0	12.5	0	60.0	0
walk	0	15.2	4.3	4.3	28.3	8.7	6.5	4.3	4.3	0	23.9

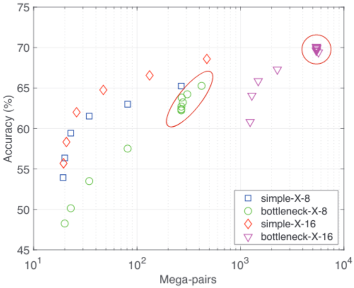
Confusion matrix for **[Nguyen 13]** on UCF Youtube

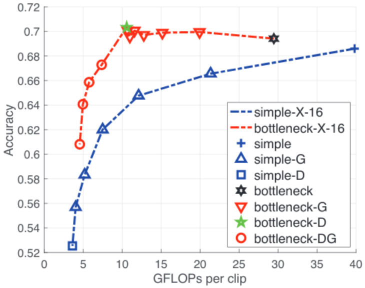
Evaluation metrics: Confusion Matrices



Confusion matrix for **[Nguyen 13]** on KTH

Evaluation: Parametric studies and Computational trade-off

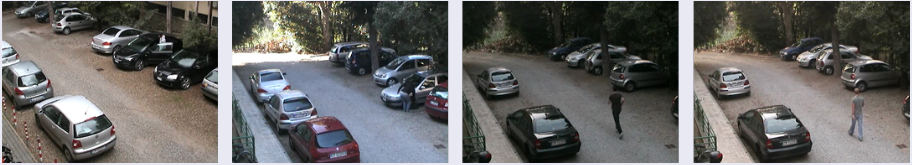




Examples taken from **[Tran 19]**

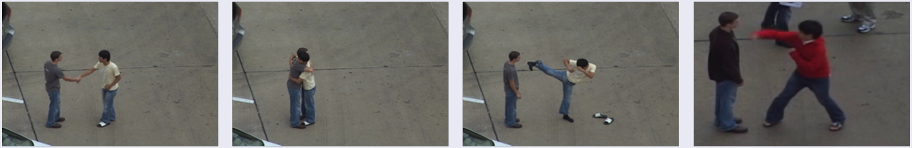
Data Bases for online recognition

Visor [Vezzani 10] 5 actions; 40 short + 1 long videos ($\approx$ 180s)



(a) Getting into a car (b) Leaving an object (c) Running (d) Walking

UT-Interaction [Ryoo 10] 6 actions; 20 videos ($\approx$ 60s and 8 actions per video)



(e) Hand shaking (f) Hugging (g) Kicking (h) Punching

Results of **[Martínez 17]** for online recognition

Accuracies are calculated per frame (on the basis of frame-level action annotations).

Confusion matrices on ViSOR:

Category	gc	lo	w	r	h
get car	100	0	0	0	0
leave Object	0	100	0	0	0
walk	0	0	83.3	16.7	0
run	0	0	14.29	85.71	0
hand shake	0	0	0	0	100

3 temporal scales

Category	gc	lo	w	r	h
get car	85.76	0	0	14.24	0
leave Object	0	100	0	0	0
walk	0	0	100	0	0
run	0	0	0	100	0
hand shake	0	0	0	0	100

5 temporal scales

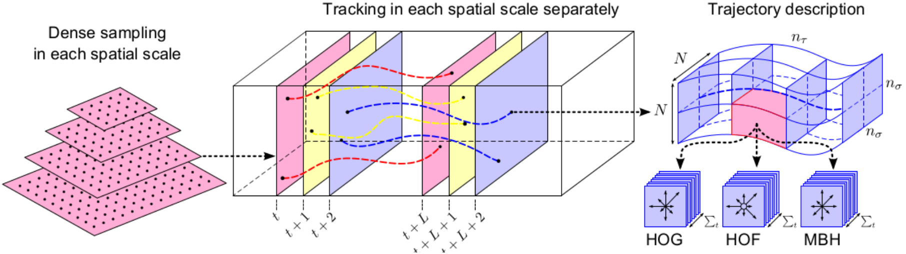
Comparative average accuracies on UT-Interaction:

Approaches	Accuracy UT-dataset 1	Accuracy UT-dataset 2
Propagative voting [28]	93	91
Proposed approach	81.6	78.3
Daysy [9]	71	51
SIFT 3D [29]	63	55
Slimani 2014 [30]	41	
Ryoo 2011 [32]	71.7	
Mukherjee [31]	79.17	
Xiaofei [33]	83.33	

Presentation Outline

Best algorithms of the moment?

Until 2015 (?), the best algorithms in the different action recognition benchmarks were based on dense trajectories:



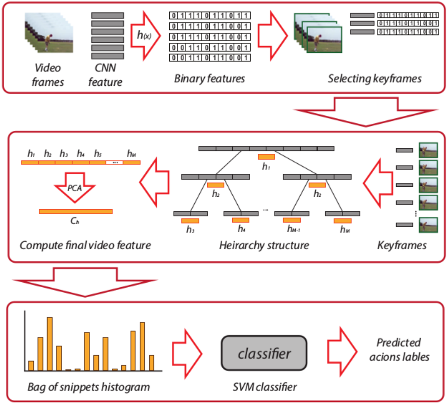
- In **[Wang 13]** dense trajectories are extracted from a fixed block of N frames, by compensating the background (camera) motion (assumed a homography).
- A series of appearance (HOG) and motion (HOF, MBH) histogram based descriptors are calculated within the cuboids centred on each trajectory.
- A codebook is trained for the trajectory features, and then the action descriptors are encoded within a bag-of-feature approach and classified using a SVM.

Best algorithms of the moment?

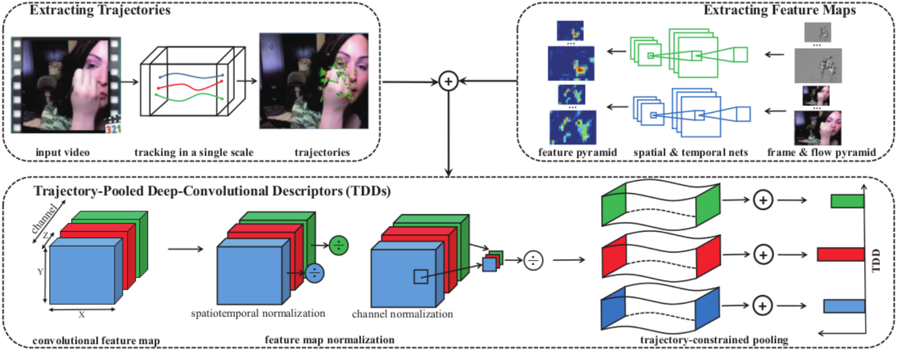
- Unlike still images (recognition, categorisation, detection...) the performance of pure end-to-end deep learning techniques applied on video data hardly reaches state-of-the art algorithms using hand-crafted features.
- End-to-end networks are huge and computationnally very heavy.
- Significant number of training videos is hard to find.
- From 2015, action recognition algorithms using CNN have begun to outperform hand-crafted algorithms, most of them being hybrid, not end-to-end approaches.

DNN action recognition

- For example, **[Ravanbakhsh 15]** use output of the penultimate layer of a CNN pre-trained on ImageNet (still images) as a pose feature.
- Significant changes in the pose features are used as key frames, and composed within a hierarchical descriptor.
- The previous descriptors are quantised (PCA) and used to train a SVM for action classification.



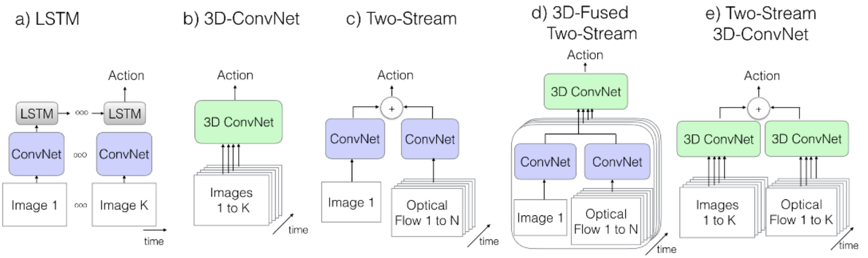
DNN action recognition

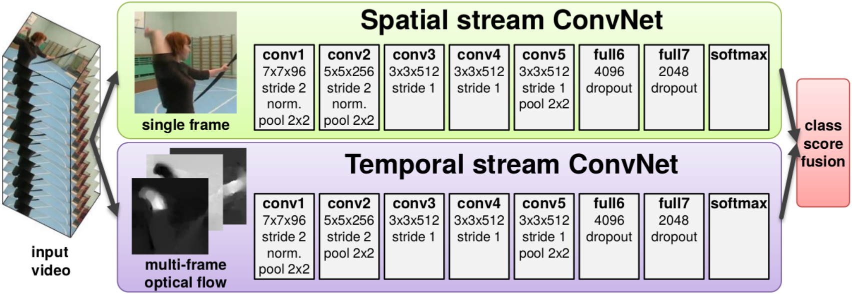


- [Wang 17] calculates, on one hand, the dense trajectories on the video, and on the other hand the feature maps from a CNN using static image and optical flow as inputs.
- Then the features are pooled along the dense trajectories to obtained a trajectory constrained deep CNN.

DNN based Action Recognition in 2019

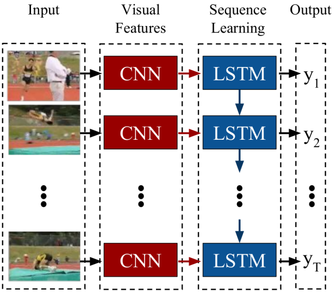
- A new huge action video dataset *Kinetics* **[Carreira 18]** has been proposed with 400 classes of 400 video clips per class.
- Most DNN based action recognition methods have been improved by pre-training on *Kinetics* dataset.
- New deep networks appear every month. Their architecture can be classified as follows **[Carreira 18]**:

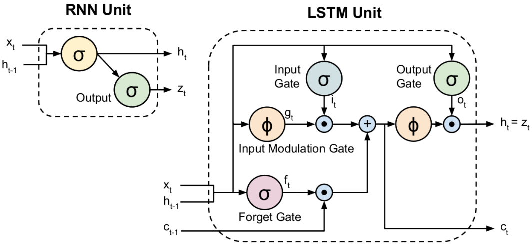




From **[Simonyan 14]**

Example of end-to-end Recurrent NN





From [Donahue 16]

Conclusion

- Global, segmentation or detection based modelling are considered too fragile.
- Local, statistical bag-of-words approaches have better performance for hand-crafted approaches.
- Trajectory seems the most relevant information support.
- Deep CNN techniques have only begun to emerge from 2014.
- Representative datasets and benchmarks are growing fastly, but remain a challenge.
- Online action recognition is still at its infancy.

References (1)

[Vishwakarma 13] S. Vishwakarma and A. Agrawal
A survey on activity recognition and behavior understanding in video surveillance
The Visual Computer 29(10), pp 983-1009, Oct. 2013

[Bobick 96] A.F. Bobick and J.W. Davis
Real-time recognition of activity using temporal templates
Proc. of Workshop on Applications of Computer Vision pp 39-42, 1996

[Gorelick 07] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri
Actions as Space-Time Shapes
IEEE Trans. on Pattern Analysis and Machine Intelligence 29(12), pp 2247-2253, Dec. 2007

[Laptev 05] I. Laptev
On Space-Time Interest Points
International Journal of Computer Vision 64(2/3), 107-123, Jul. 2005

References (2)

[Efros 03] A.A. Efros, A.C. Berg, G. Mori and J. Malik
Recognizing Action at a Distance
Int. Conf. on Computer Vision (ICCV), Vol. 2, pp 726-733, 2003

[Chaudhry 09] R. Chaudhry, A. Ravichandran, G. Hager and R. Vidal
Histograms of Oriented Optical Flow and Binet-Cauchy Kernels on Nonlinear Dynamical Systems for the Recognition of Human Actions
Conf. on Comp. Vision and Pat. Rec. (CVPR), pp.1932-1939, 2009

[Martínez 12] F. Martínez, A. Manzanera, and E. Romero
A motion descriptor based on statistics of optical flow orientations for action classification in video-surveillance.
Conf. on Multimedia and Signal Processing (CMSP). Shanghai, 2012.

[Martínez 17] F. Martínez, A. Manzanera, and E. Romero
Spatio-temporal multi-scale motion descriptor from a spatially-constrained decomposition for online action recognition.
IET Computer Vision. 11(7). 2017. pp.541-549.

References (3)

- [Garrigues 12] M. Garrigues and A. Manzanera
  Real Time Semi-dense Point Tracking
  Int. Conf. on Image Analysis and Recognition (ICIAR), pp245-252, 2012
- [Nguyen 13] T.P. Nguyen and A. Manzanera
  Action Recognition Using Bag of Features extracted from a Beam of Trajectories
  Proc. of Int. Conf. on Image Processing (IEEE-ICIP), Sep. 2013
- [Martínez 15] F. Martínez, A. Manzanera, M. Gouiffès and A. Braffort
  A Gaussian mixture representation of gesture kinematics for on-line Sign Language video annotation
  Int. Symp. on Visual Computing (ISVC). 2015.
- [Jain 13] M. Jain, H. Jégou and P. Bouthémy
  Better exploiting motion for better action recognition
  Conf. on Computer Vision and Pattern Recognition (CVPR), 2013

References (4)

[Schuldt 04] C. Schuldt, I. Laptev and B. Caputo
Recognizing Human Actions: A Local SVM Approach
Int. Conf. on Pattern Recognition (ICPR), pp 32-36, 2004

[Liu 09] J. Liu, J. Luo and M. Shah
Recognizing realistic actions from video "in the wild"
Conf. on Comp. Vis. and Pat. Rec. (CVPR), pp 1996-2003, 2009

[Kuehne 11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre
HMDB: A Large Video Database for Human Motion Recognition
ICCV, 2011

[Tran 19] D. Tran, H. Wang, L. Torresani and M. Feiszli
Video Classification with Channel-Separated Convolutional Networks
arXiv Preprint 1904.02811, 2019

[Ryoo 10] M.S. Ryoo and J.K Aggarwal
UT-Interaction Dataset
ICPR contest on SDHA, 2010

References (5)

[Vezzani 10] Video surveillance online repository (ViSOR): an integrated framework
R. Vezzani and R. Cucchiara
Multimedia Tools and Applications 50 (2), 359-380, 2010

[Sadanand 12] S. Sadanand and J.J. Corso
Action bank: A high-level representation of activity in video
Conf. on Comp. Vis. and Pat. Rec. (CVPR), pp 1234-1241, 2012

[Wang 13] H. Wang and C. Schmid
Action Recognition with Improved Trajectories
IEEE Int. Conf. on Comp. Vis.pp.3551-3558 2013

[Ravanbakhsh 15] M. Ravanbakhsh, H. Mousavi, M. Rastegari, V. Murino and L. S. Davis
Action Recognition with Image Based CNN Features
http://arxiv.org/abs/1512.03980, 2015

References (6)

[Wang 15] L. Wang, Y. Qiao and X. Tang
Action recognition with trajectory-pooled deep-convolutional descriptors
Conf. on Comp. Vis. and Pat. Rec. (CVPR), 2015, pp. 4305-4314

[Donahue 16] J. Donahue et al
Long-term Recurrent Convolutional Networks for Visual Recognition and Description
IEEE Trans. on Pat. Anal. and Mach. Intel 39(4), 677-691, 2017

[Simonyan 14] K. Simonyan and A. Zisserman
Two-Stream Convolutional Networks for Action Recognition in Videos
Advances in Neural Information Processing Systems 27 (NIPS 2014)

[Carreira 18] Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset
J. Carreira and A. Zisserman
https://arxiv.org/abs/1705.07750, 2018