Action modelling and recognition in videos using beams of trajectories

Thanh Phuong Nguyen, Matthieu Garrigues, Antoine Manzanera

ENSTA-ParisTech

ENSTA ParisTech

SIVA 2013
Guelma, November 2013

Context of our work

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

From [Laptev 13]

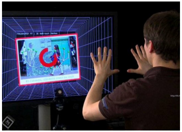
Focus of our work

Action modelling: Extract relevant action features from the video flow.

Applications

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



Challenges

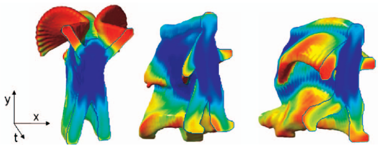
- Huge variability (appearance, geometry)
- Moving camera

Presentation Outline

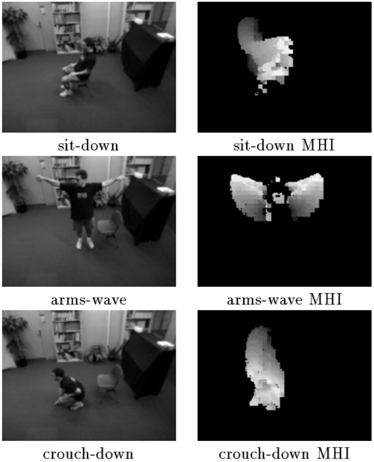
Presentation Outline

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]**



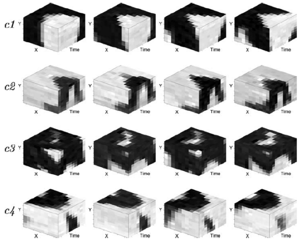
from [Gorelick 07]



from [Bobick 96]

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 scale space 3d Harris corner points
- quantises their local appearance to form a code book
- describes them using space $\times$ time partial derivatives

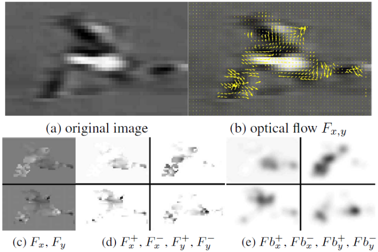


from [Laptev 05]

Action Modelling State-of-the-Art: Velocity based methods

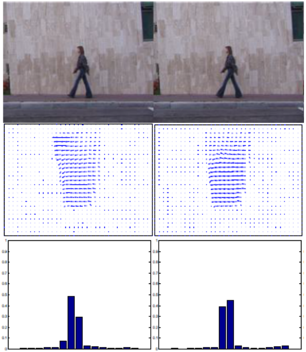
Some models are built from velocity field (optical flow). For example:

- **[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]

Presentation Outline

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

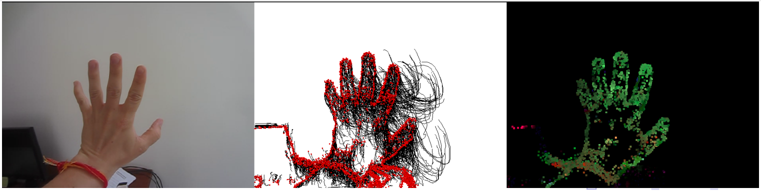
- Temporally *short term*
- Spatially *dense*
- Main computational load: *Spatial regularisation*

Point tracker

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

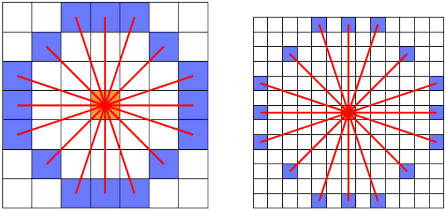
Video Extruder

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



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).

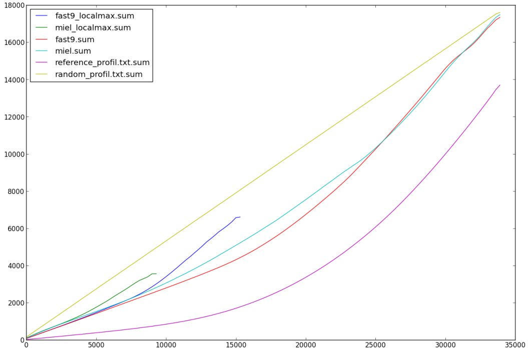


Multi-scale keypoint supports: Bresenham circles

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.

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

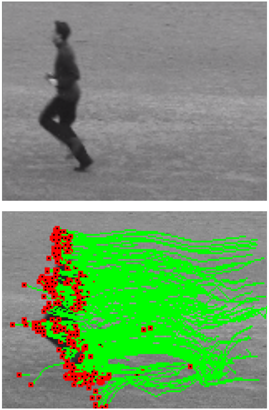
Time performance of *Video extruder* on different architectures.

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

From Trajectories to Action

hand clapping

hand waving

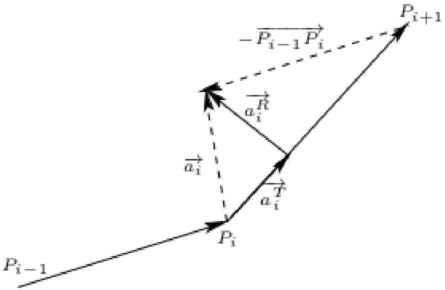


running

Presentation Outline

Representation of Atomic Actions

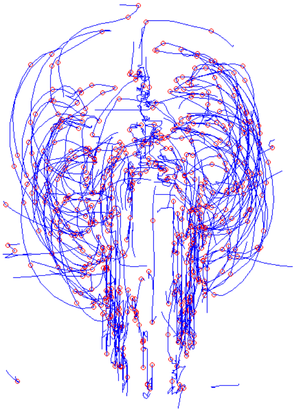
Elementary motion elements (*atomic actions*) are extracted from the trajectories, using *dominant points*, corresponding to *local maxima* of the *radial acceleration* (related to *curvature*), for different temporal scales.



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

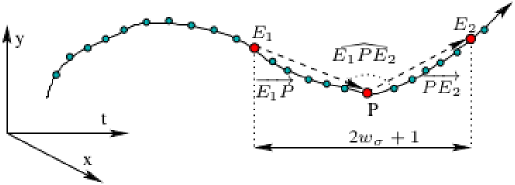
The temporal scale is related to the standard deviation σ of the Gaussian used to smooth the trajectory.

Dominant point detection



Representation of Atomic Actions

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...

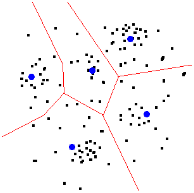


Computation of the feature vector around the dominant point P.

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

Building a Code Book of Atomic Actions

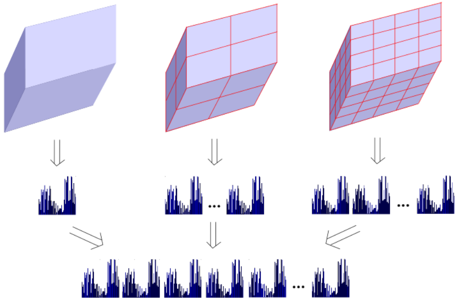
- In a first level (non supervised) learning phase, the feature vectors from a set of actions are vector quantised (K-means algorithm) to form a code book of atomic actions.



- 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.

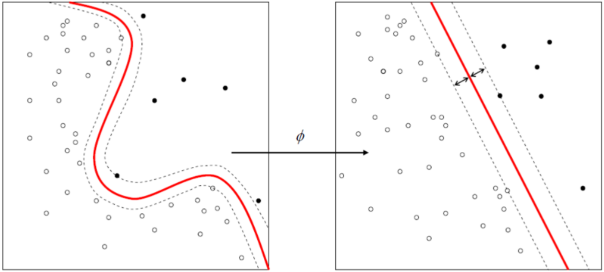


Representation of an action from multiple histograms.

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.

Presentation Outline

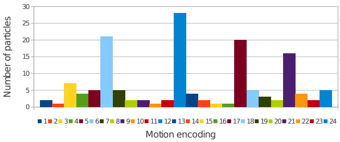
When the camera is moving, many trajectories are due to the relative motion of the background and must be discarded.

Sport sequence from *UCF Youtube dataset*

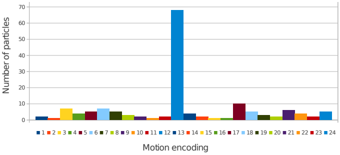
Computed trajectories

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

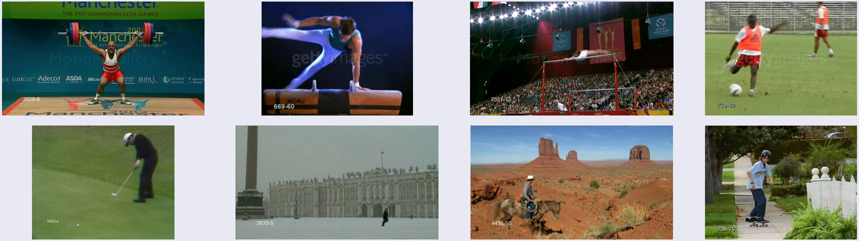
Background Removal: Results

Presentation Outline

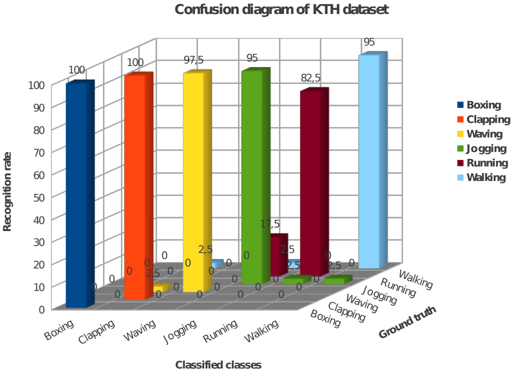
Experiments: Data Bases

KTH [Schuldt 04]

UCF Youtube [Liu 09]



Experiments: Results on KTH dataset



- 600 videos: 6 actions for 25 people.
- Dominant point radial acceleration threshold: 0.25 pixel per frame^2.
- Code book: 70 atomic actions.
- Multiple histogram grids: $1 \times 1 \times 1$, $2 \times 2 \times 2$, $4 \times 4 \times 4$.

Ours 95	[Javan 12] 94.5	[Yao 10] 93.5	[Thi 11] 94.7
[Seo 11] 95.1	[Wang 11] 93.8	[Liu 08] 94.2	[Sadanand 12] 98.2

Average recognition rates for different methods.

Experiments: Results on UCF Youtube dataset

- 1 600 videos: 11 categories.
- Dominant point radial acceleration threshold: 0.25 pixel per frame^2.
- Code book: 1 000 atomic actions.
- Multiple histogram grids: $1 \times 1 \times 1$, $2 \times 2 \times 2$, $4 \times 4 \times 4$.

Ours	[Lu 11]	[Bregonzio 10]	[Liu 09] v1	[Liu 09] v2
65.1	64	64	65.4	71.2

Average recognition rates for different methods.

Experiments: Results on UCF Youtube dataset

	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 our method on UCF Youtube

Presentation Outline

Contribution Outline

- Action recognition using features extracted from *trajectories*.
- *High density* of trajectories $\longrightarrow$ *Statistical* bag-of-feature approach.
- *Low computational* load of the semi-dense tracker $\longrightarrow$ potentially *real-time* method.
- Run time cost of recognition $\longleftrightarrow$
  - size of the code book
  - dimension of action descriptor
  - number of actions
- Cumulative estimation of dominant motion $\longrightarrow$ Recognition with moving camera.

Perspectives

- Other action representations based on semi dense trajectories are investigated.
- Semantics of the atomic actions $\leftrightarrow$ Managing the code book size.
- Selection of the most relevant features for the atomic action descriptors.
- Individual / hierarchical classification of the trajectories $\longrightarrow$ More reactive, more parallel...

Acknowledgements

This work is part of an ITEA2 European project, and is supported by French Ministry of Economy (DGCIS).

Thanh Phuong Nguyen

Action Modelling

Learning & Classification

Matthieu Garrigues

Semi dense Tracking

Embedded Implementations

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

References (2)

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

[Efros 03] A.A. Efros, A.C. Berg, G. Mori and J. Malik
Recognizing Action at a Distance
Prof. of 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
Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1932 - 1939, 2009

References (3)

[Garrigues 12] M. Garrigues and A. Manzanera
Real Time Semi-dense Point Tracking
Proc. of Int. Conf. on Image Analysis and Recognition (ICIAR), LNCS vol. 7324, 245-252, Jun. 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

[Jain 13] M. Jain, H. Jégou and P. Bouthémy
Better exploiting motion for better action recognition
Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR), Apr. 2013

References (4)

[Schuldt 04] C. Schuldt, I. Laptev and B. Caputo
Recognizing Human Actions: A Local SVM Approach
Proc. of 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"
Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR), pp 1996-2003, 2009

[Sadanand 12] S. Sadanand and J.J. Corso
Action bank: A high-level representation of activity in video
Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR), pp 1234-1241 (2012)