Action modelling and recognition in videos

Antoine Manzanera

U2IS / Robotics and Vision
ENSTA-ParisTech

M2 AIC - D&K
Université Paris-Saclay
Action Modelling and Activity Understanding

Context

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

Focus of this lecture

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

From [Laptev 13]
Action Modelling and Activity Understanding

Applications

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

Challenges

- Huge variability (appearance, geometry)
- Moving camera
1. Introduction

2. Action Features

3. Action Coding and Recognition

4. Results and Evaluation

5. Conclusion and current trends
Presentation Outline

1. Introduction
2. Action Features
3. Action Coding and Recognition
4. Results and Evaluation
5. Conclusion and current trends
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]

![Images of actions with 2D and 3D representations](image1.png)
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
Velocity-based Action Features

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

- \([\text{Efros 03}]\) computes grey level patterns from velocity measures.
- \([\text{Chaudhry 09}]\) uses histograms of optical flow orientations as action descriptor.
[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)

From [Martínez 17]

A. Manzanera (ENSTA/U2IS) Action Modelling
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

**Cons**
- Sparse
- Fragile
- Noisy
- Costly

J.E. Marey *Mouvement*  
*(Chronophographie)* - 1882
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:**
  \[ \Sigma_s(p) = \min_{i=0}^{7} |2I(p) - I(q_i) - I(q_{i+8})| \]
- **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 $\times2$ to $\times15$ (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.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Resolution</th>
<th># points</th>
<th>Freq. (Hz)</th>
<th># Cpp</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU GeForce GTX 460 1.35GHz</td>
<td>640 × 480</td>
<td>8 500</td>
<td>166</td>
<td>957</td>
</tr>
<tr>
<td>CPU quad-core i5 2500k 3.3GHz</td>
<td>640 × 480</td>
<td>8 500</td>
<td>152</td>
<td>2 550</td>
</tr>
<tr>
<td>ARM dual-core STE U8500 1GHz</td>
<td>320 × 240</td>
<td>3 000</td>
<td>11</td>
<td>30 300</td>
</tr>
<tr>
<td>ARM single-core IMX.53 1GHz</td>
<td>720 × 288</td>
<td>2 000</td>
<td>10</td>
<td>50 000</td>
</tr>
</tbody>
</table>

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

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 $\vec{a}_r$ on a trajectory

The temporal scale is related to the standard deviation $\sigma$ of the Gaussian used to smooth the trajectory.
Dominant point detection [Nguyen 13]
Every dominant point is described using a \textit{feature vector} composed of \textit{geometrical} and \textit{statistical} parameters of the trajectory around the dominant point: angle, curvature, directions, average and variance of speed and accelerations...

The size of the support depends on the temporal scale $\sigma$ of the dominant point.
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.
We represent a complex action by concatenating histograms of atomic actions on a hierarchy of space $\times$ time boxes.

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 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
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.
Background Removal: Results
1. Introduction
2. Action Features
3. Action Coding and Recognition
4. Results and Evaluation
5. Conclusion and current trends
Experiments: Data Bases of segmented videos

KTH [Schuldt 04]

UCF Youtube [Liu 09]
Experiments: Data Bases for online recognition

**Visor** [Vezzani 10]

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

**UT-Interaction** [Ryoo 10]

- (e) Hand shaking
- (f) Hugging
- (g) Kicking
- (h) Punching
Experiments: Results of [Nguyen 13] on KTH dataset

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

<table>
<thead>
<tr>
<th></th>
<th>[Nguyen 13]</th>
<th>[Javan 12]</th>
<th>[Yao 10]</th>
<th>[Thi 11]</th>
<th>[Seo 11]</th>
<th>[Wang 11]</th>
<th>[Liu 08]</th>
<th>[Sadanand 12]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>95</td>
<td>94.5</td>
<td>93.5</td>
<td>94.7</td>
<td>95.1</td>
<td>93.8</td>
<td>94.2</td>
<td>98.2</td>
</tr>
<tr>
<td>recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rates for</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>different</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A. Manzanera (ENSTA/U2IS)
Experiments: Results of [Nguyen 13] on UCF Youtube dataset

- 1600 videos: 11 categories.
- Dominant point radial acceleration threshold: 0.25 pixel per frame$^2$.
- Multiple histogram grids: $1 \times 1 \times 1$, $2 \times 2 \times 2$, $4 \times 4 \times 4$.

<table>
<thead>
<tr>
<th>[Nguyen 13]</th>
<th>[Lu 11]</th>
<th>[Bregozio 10]</th>
<th>[Liu 09] v1</th>
<th>[Liu 09] v2</th>
</tr>
</thead>
<tbody>
<tr>
<td>65.1</td>
<td>64</td>
<td>64</td>
<td>65.4</td>
<td>71.2</td>
</tr>
</tbody>
</table>

Average recognition rates for different methods.
Experiments: Results of [Nguyen 13] on UCF Youtube dataset

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

Confusion matrix for [Nguyen 13] on UCF Youtube
Results of [Martínez 17] for online recognition

Confusion matrices on ViSOR:

<table>
<thead>
<tr>
<th>Category</th>
<th>gc</th>
<th>lo</th>
<th>w</th>
<th>r</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>get car</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>leave Object</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>walk</td>
<td>0</td>
<td>0</td>
<td>83.3</td>
<td>16.7</td>
<td>0</td>
</tr>
<tr>
<td>run</td>
<td>0</td>
<td>0</td>
<td>14.29</td>
<td>85.71</td>
<td>0</td>
</tr>
<tr>
<td>hand shake</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>gc</th>
<th>lo</th>
<th>w</th>
<th>r</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>get car</td>
<td>85.76</td>
<td>0</td>
<td>0</td>
<td>14.24</td>
<td>0</td>
</tr>
<tr>
<td>leave Object</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>walk</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>run</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>hand shake</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

3 temporal scales

5 temporal scales

Comparative average accuracies on UT-Interaction:

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Accuracy UT-dataset 1</th>
<th>Accuracy UT-dataset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propagative voting [28]</td>
<td>93</td>
<td>91</td>
</tr>
<tr>
<td><strong>Proposed approach</strong></td>
<td>81.6</td>
<td>78.3</td>
</tr>
<tr>
<td>Daysy [9]</td>
<td>71</td>
<td>51</td>
</tr>
<tr>
<td>SIFT 3D [29]</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>Slimani 2014 [30]</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>Ryoo 2011 [32]</td>
<td>71.7</td>
<td></td>
</tr>
<tr>
<td>Mukherjee [31]</td>
<td>79.17</td>
<td></td>
</tr>
<tr>
<td>Xiaofei [33]</td>
<td>83.33</td>
<td></td>
</tr>
</tbody>
</table>
1. Introduction

2. Action Features

3. Action Coding and Recognition

4. Results and Evaluation

5. Conclusion and current trends
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.
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 computationally 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.
[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]:

  a) LSTM
  
  ![LSTM Diagram](image1)

  b) 3D-ConvNet
  
  ![3D-ConvNet Diagram](image2)

  c) Two-Stream
  
  ![Two-Stream Diagram](image3)

  d) 3D-Fused Two-Stream
  
  ![3D-Fused Two-Stream Diagram](image4)

  e) Two-Stream 3D-ConvNet
  
  ![Two-Stream 3D-ConvNet Diagram](image5)
[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

Actions as Space-Time Shapes

[Laptev 05] I. Laptev
On Space-Time Interest Points
International Journal of Computer Vision 64(2/3), 107-123, Jul. 2005
Recognizing Action at a Distance
Int. Conf. on Computer Vision (ICCV), Vol. 2, pp 726-733, 2003

Histograms of Oriented Optical Flow and Binet-Cauchy Kernels on
Nonlinear Dynamical Systems for the Recognition of Human Actions

[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.

[Martínez 17] F. Martínez, A. Manzanera, and E. Romero
Spatio-temporal multi-scale motion descriptor from a spatially-constrained
decomposition for online action recognition.
[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

Better exploiting motion for better action recognition
Conf. on Computer Vision and Pattern Recognition (CVPR), 2013
[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

Recognizing realistic actions from video “in the wild”

[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

Action Recognition with Improved Trajectories
IEEE Int. Conf. on Comp. Vis.pp.3551-3558 2013
Action Recognition with Image Based CNN Features

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

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