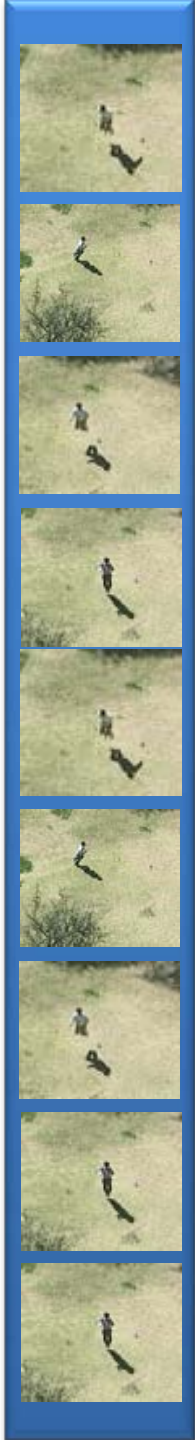
ICPR 2010 Contest on
Semantic Description of Human Activities
Aerial View Activity Classification Challenge

HMM Based Action Recognition with Projection Histogram Features

Roberto Vezzani, Davide Baltieri, Rita Cucchiara

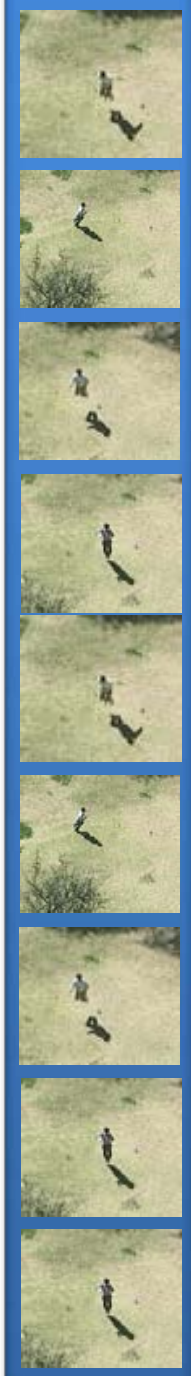
University of Modena and Reggio Emilia (Italy)



Contest goal

- Classify human actions in low-resolution videos
- Types of Actions in the Aerial View Challenge: Pointing Standing Digging Walking Carrying Running Wave1 Wave2 Jumping
- The average height of human figures in this dataset is about 20 pixels.

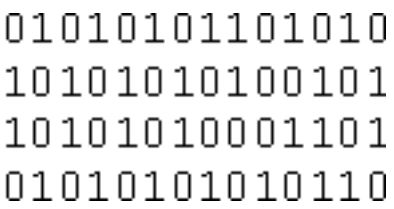
- **We propose to use a classical HMM framework with projection histogram features**



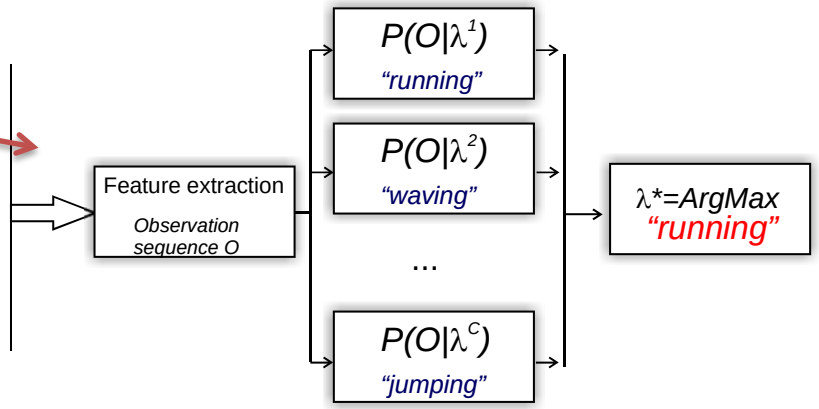
Overall schema

frames

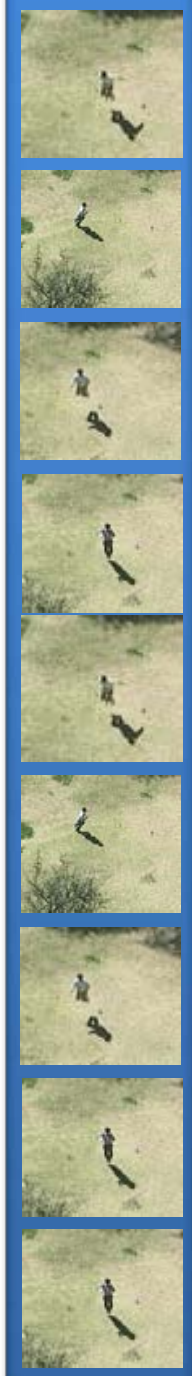
Foreground detection



Feature extraction

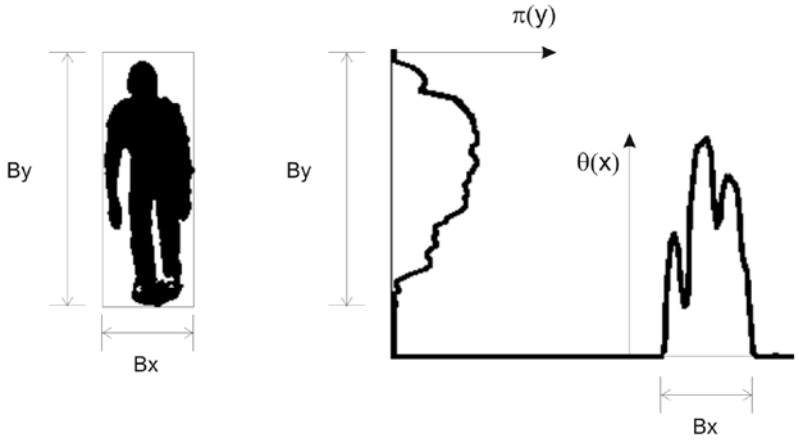


HMM Classifier



Feature set 1: Projection histograms

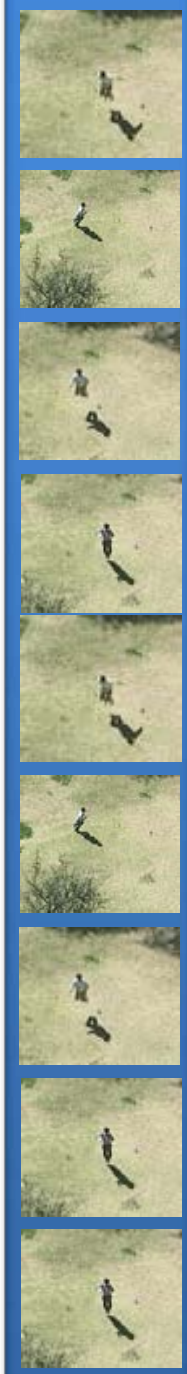
- Projections of the person's silhouette onto the principal axes *x* and *y*
- Given the boolean foreground mask F(x, y):



$$\theta(x) = \sum_{y=0}^{F_y} \phi(F(x,y)) \; ; \; \pi(y) = \sum_{x=0}^{F_x} \phi(F(x,y))$$

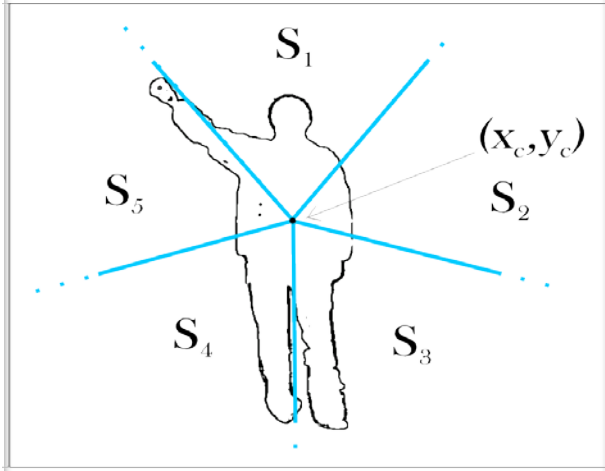
Projection histograms

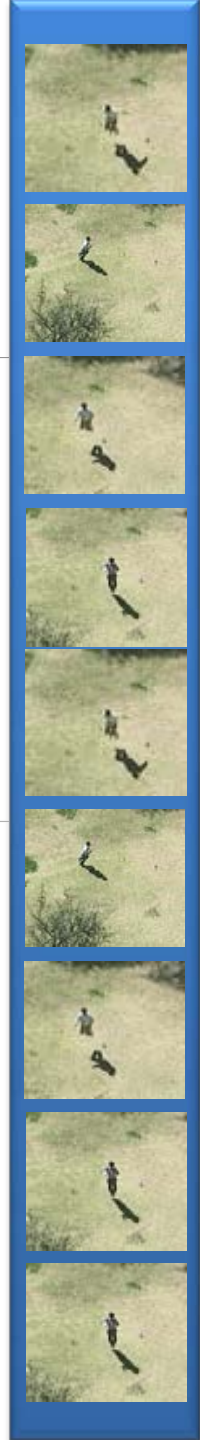
- **Pros**:
  - Very fast
  - Low sensitivity to pixel noise
  - Generic (no assumption on the human shape)
- **Cons**
  - Strongly view dependent
  - High sensitivity to occlusions



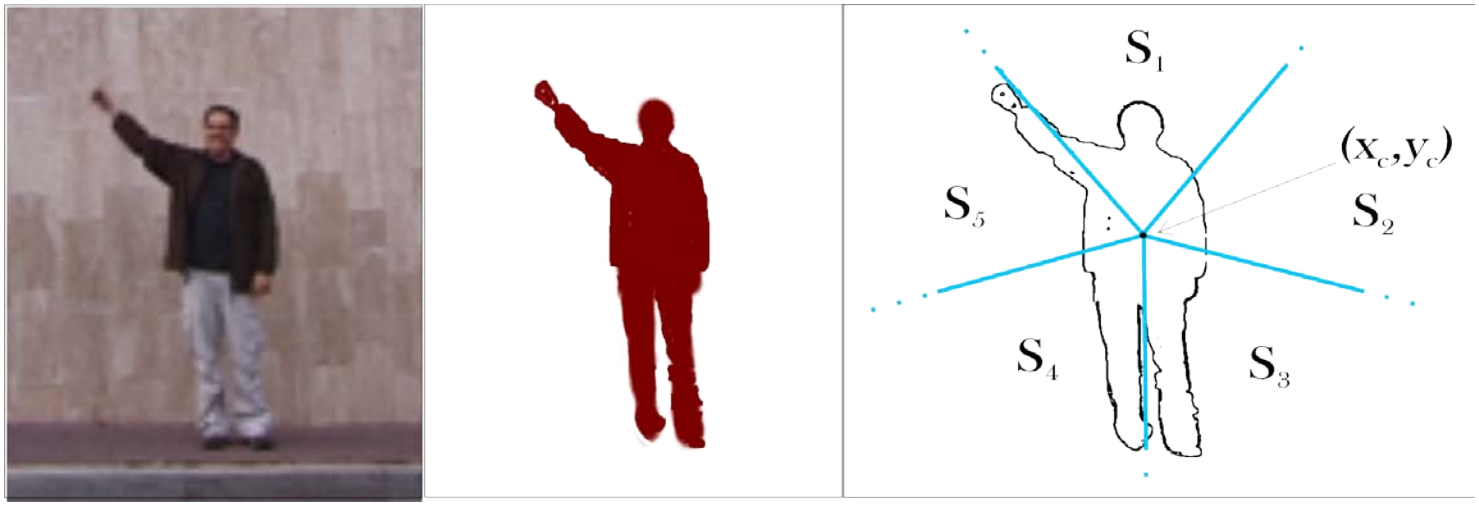
An alternative feature set: a model based approach

- The foreground silhouettes are divided into five slices S1 . . . S5 using a radial partitioning centered in the gravity center
- These slices ideally correspond to the head, the arms and the legs
- 17-dimensional feature set containing both motion and shape information





Feature set 2: model based



$$o_t = \left\{ o_t^1 \dots o_t^{17} \right\},$$

$$\begin{cases} o_t^1 = x_c(t) - x_c(t-1) \\ o_t^2 = y_c(t) - y_c(t-1) \\ o_t^{3\dots7} = A_t^i / A_t \\ o_t^{8\dots12} = \max\limits_{S_i} \left((x - x_c)/\sqrt{A_t^i} \right) \\ o_t^{13\dots17} = \max\limits_{S_i} \left((y - y_c)/\sqrt{A_t^i} \right) \end{cases}$$

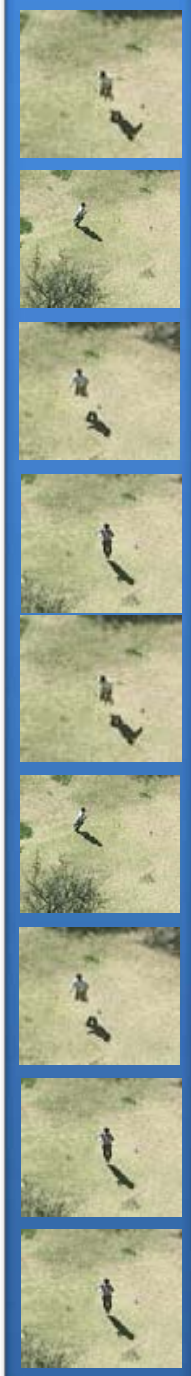
← *motion*

← *Slice areas*

← *Extremal points*

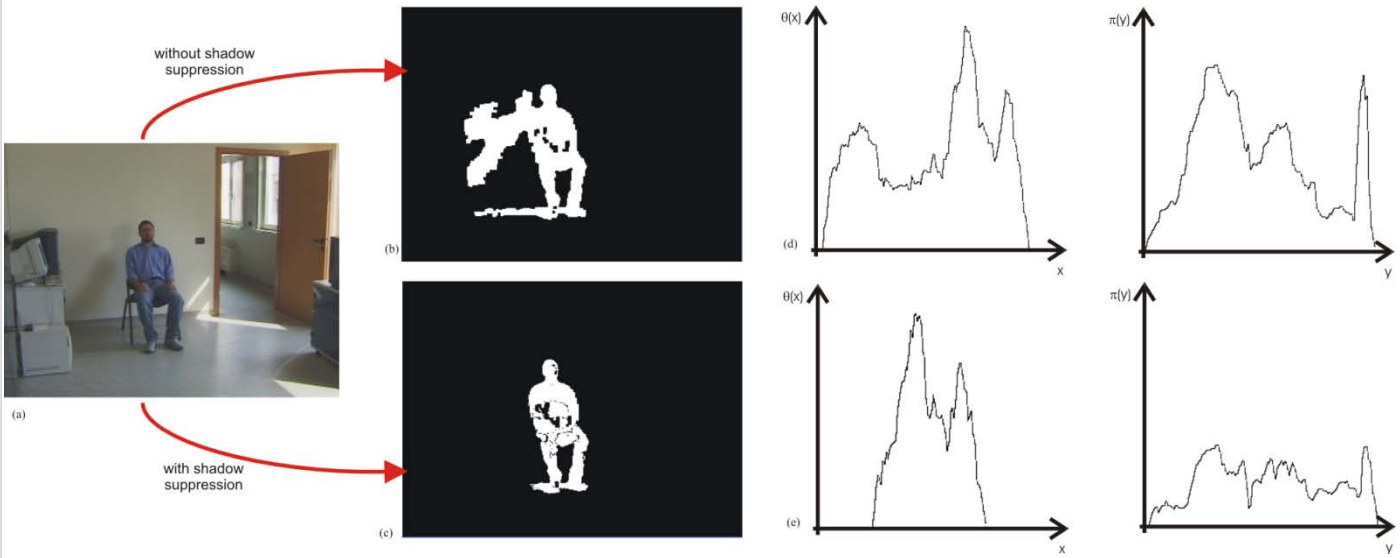
Pros and cons

- **Pros**:
  - Higher Information content
  - Motion features allows to easily recognize walks and jumps
- **Cons**
  - Some features require more computation time then PH
  - Speed features depend on the previous observation
  - High sensitivity to center estimation



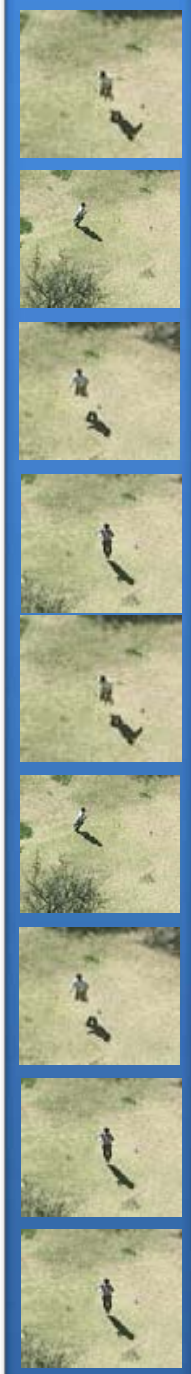
Shadow removal

- Shadows are usually enemies to fight againts
- Shapes and features are negatively affected by shadows
- Projection histograms changes, gravity center is wrongly estimated…



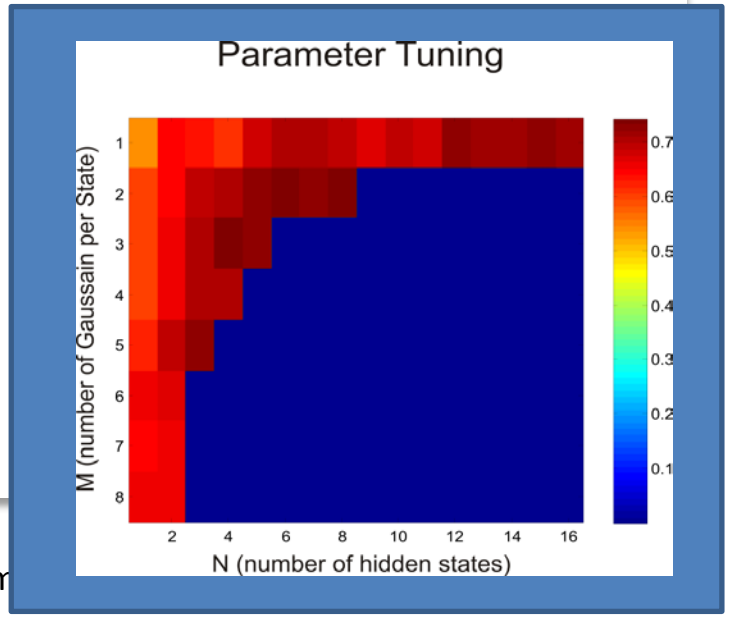
Shadows in aerial view contest

- If the shadow characteristics (i.e., size, position, direction) are not changing among sequences, they can be leaved;
- Information about the performed action are also embedded in the shadow mask.
- Thus, we can avoid any shadow removal step if the shadows are always in the same direction (as in the contest videos) and if the adopted feature set is not model based (such as the projection histograms).



HMM training

- We adopt a Gaussian Mixture Model, which simplifies the learning phase
- One HMM trained for each action, leaving out the video to classify
- Simultaneous estimation of both the HMM and the Mixtures parameters using the Baum-Welch algorithm
- The numbers N and M of hidden states and Gaussians per state have been empirically estimated



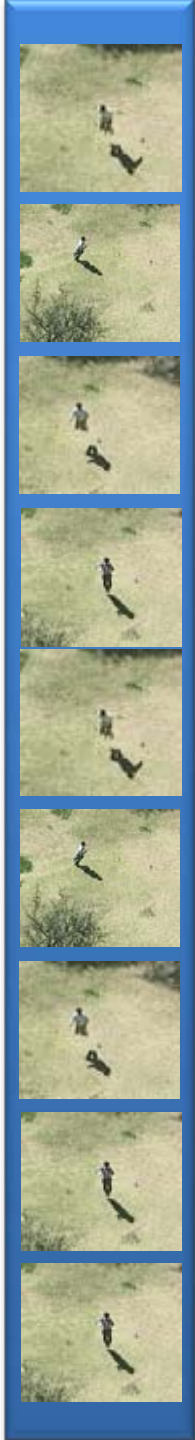
HMM Action Classification

- Using the recursive forward algorithm

$$\alpha_1(j) = \pi_i b_j(o_1), 1 \leq i \leq N$$
$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j\left(o_{t+1}\right)$$
$$P\left(O|\lambda\right) = \sum_{j=1}^{N} \alpha_T\left(j\right)$$

- Then

$$\lambda^* = \arg\max_{1 \leq c \leq C} \left[P\left(O|\lambda^c\right)\right]$$

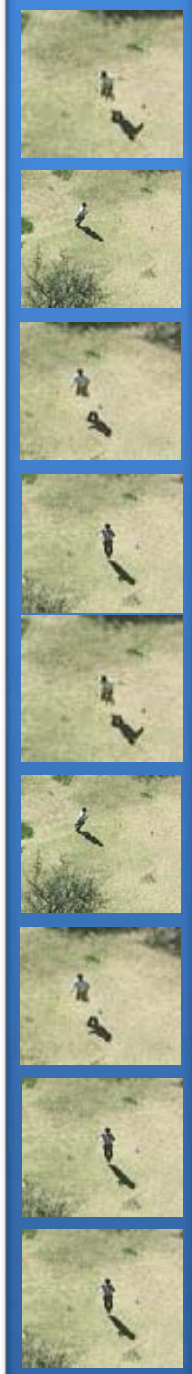


Results

- Confusion matrix using the projection histogram feature set
- leave-one-out cross-validation, where one video sequence is used for testing at a time

Ground thruth - Action ID

Recognized Action ID	1	2	3	4	5	6	7	8	9
1	**10**	0	0	0	0	0	2	0	0
2	0	**10**	0	1	0	0	1	0	0
3	0	0	**12**	0	0	0	0	0	0
4	0	0	0	**12**	0	0	0	0	0
5	0	0	0	0	**12**	0	0	0	0
6	0	0	0	0	0	**12**	0	0	0
7	0	0	0	0	0	0	**12**	0	0
8	0	0	0	0	0	0	0	**12**	0
9	0	0	0	0	0	0	0	0	**12**

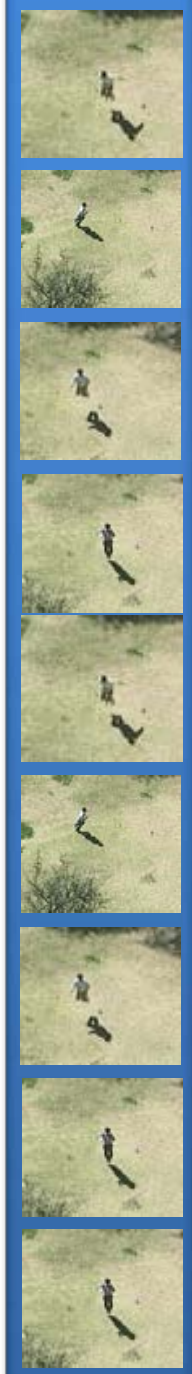


Weizmann dataset

- Confusion matrix of the Model Based Feature set on the Weizmann dataset

Ground thruth - Action ID

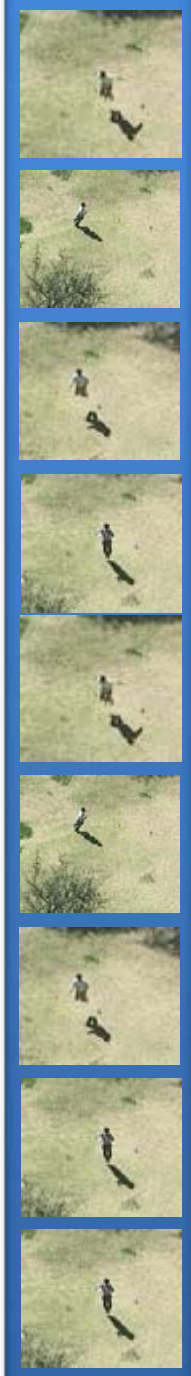
Recognized Action ID	1	2	3	4	5	6	7	8	9	10
1	**100**	0	0	0	0	0	0	0	0	0
2	0	**99**	0	0	0	0	0	0	1	0
3	0	0	**68**	0	4	0	27	1	0	0
4	0	12	0	**87**	0	0	0	0	1	0
5	0	0	0	0	**81**	0	19	0	0	0
6	0	0	0	0	5	**95**	0	0	0	0
7	0	0	12	0	31	0	**57**	0	0	0
8	0	0	0	0	0	0	0	**100**	0	0
9	0	0	0	0	0	0	0	0	**86**	14
10	0	0	0	0	0	0	0	0	6	**94**



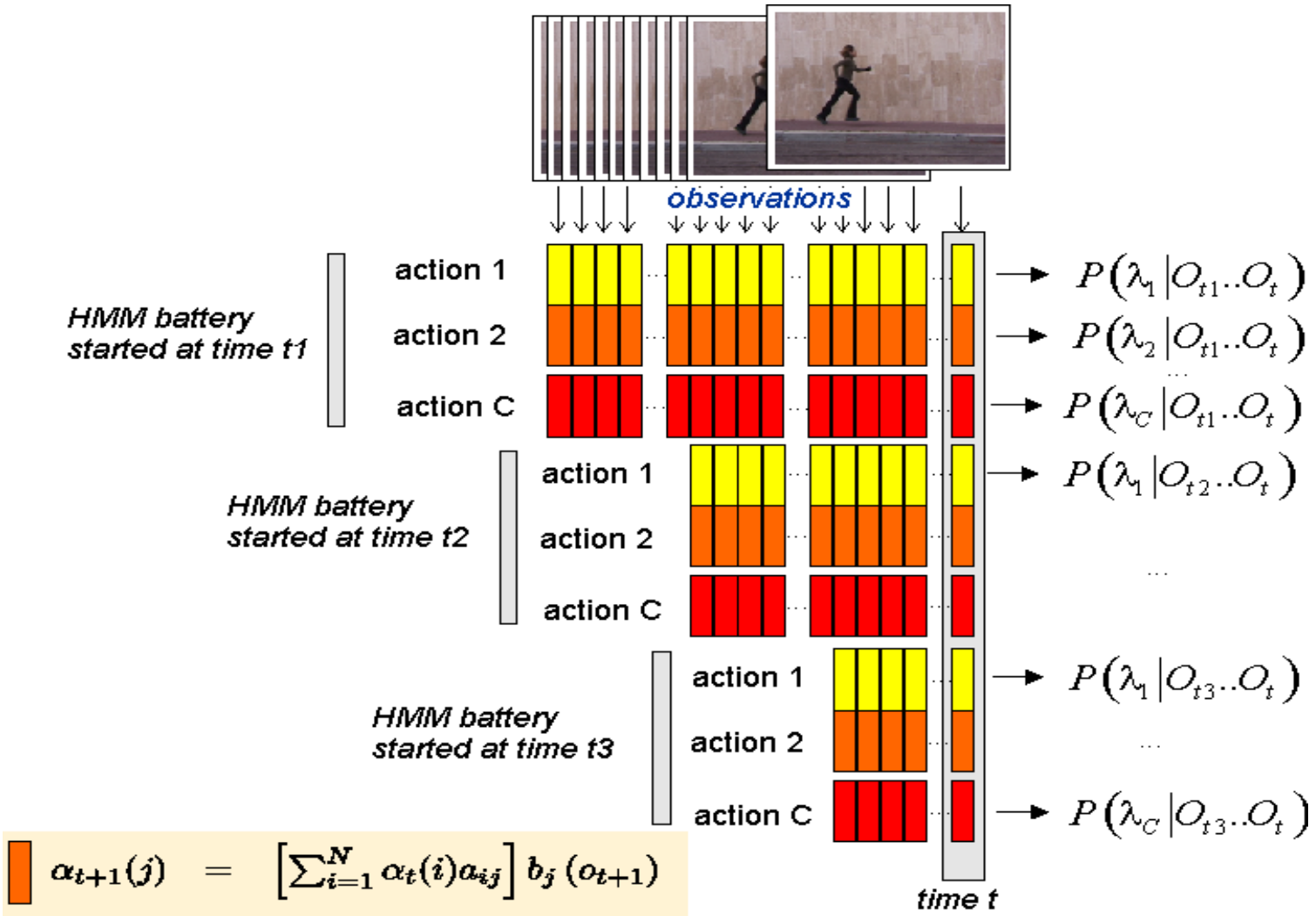
Framework extension

- This HMM framework is part of an extended system for **simultaneous action classification & time segmentation**
- Streams of HMM are triggered and updated to provide, for each frame, the most likely current action and the most likely temporal segmentation of the current action

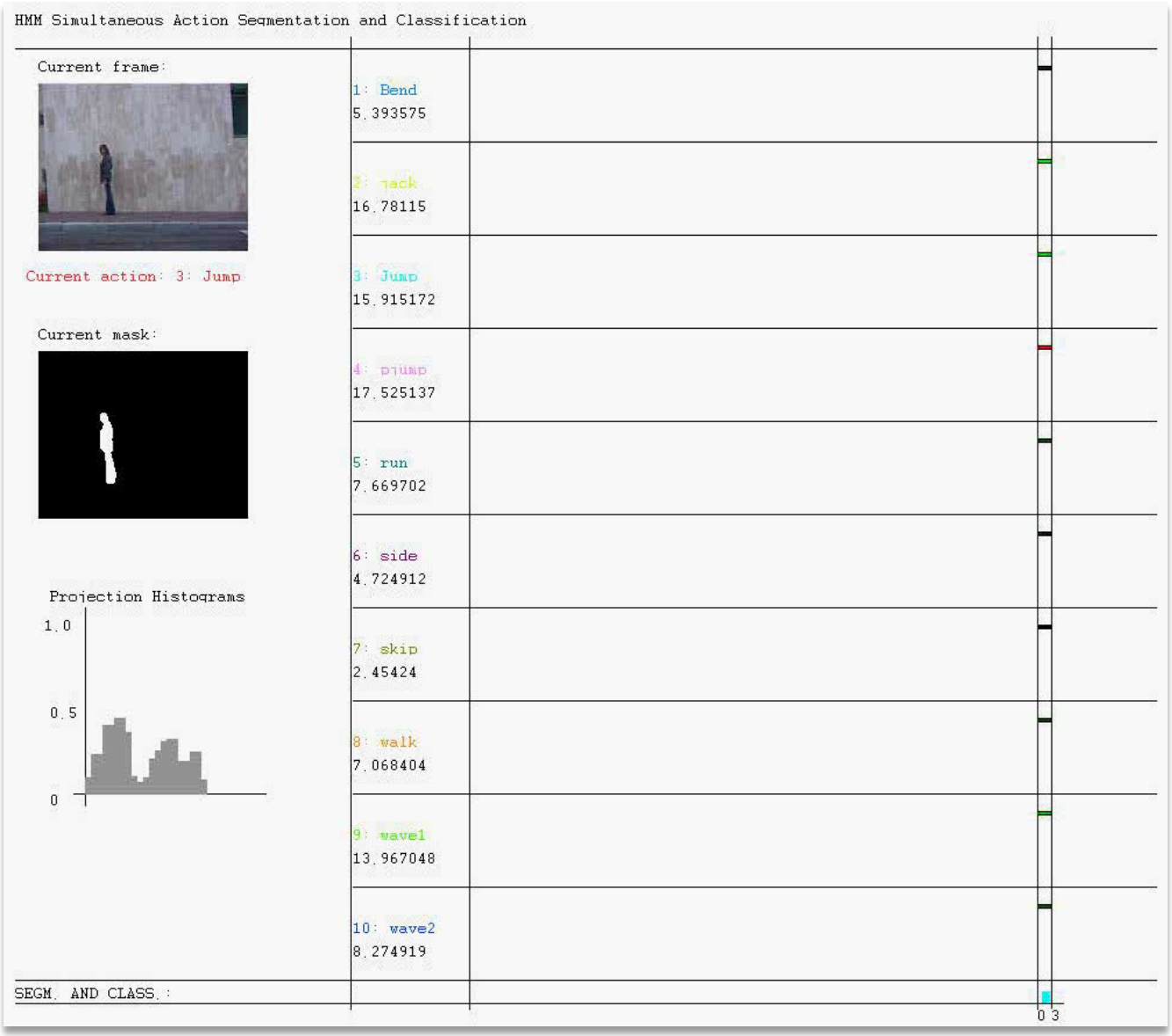
R. Vezzani, M. Piccardi, R. Cucchiara, "An efficient Bayesian framework for on-line action recognition" in *Proceedings of the 16th International Conference on Image Processing (ICIP 2009)*, Cairo, Egypt, Nov. 7-11, 2009



Streams of HMMs



Preliminary result



Acknowledgments

The framework is currently under development and improvement **within the project**

with the support of the Prevention, Preparedness and Consequence Management of Terrorism and other Security-related Risks Programme European Commission - Directorate-General Justice, Freedom and Security