Action recognition in videos

Cordelia Schmid

Action recognition - goal

- Short actions, i.e. drinking, sit down

Drinking

Coffee & Cigarettes dataset

Sitting down

Hollywood dataset

Action recognition - goal

- Activities/events, i.e. making a sandwich, feeding an animal

Making sandwich

Feeding an animal

TrecVid Multi-media event detection dataset

Action recognition - tasks

- Action classification: assigning an action label to a video clip

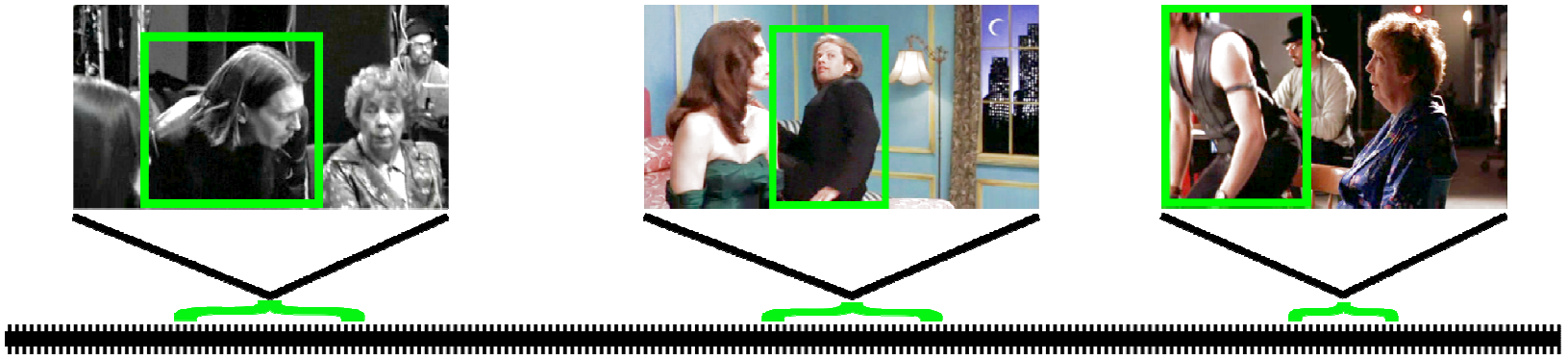
Making sandwich: present
Feeding animal: not present
…

Action recognition - tasks

- Action classification: assigning an action label to a video clip

Making sandwich: present
Feeding animal: not present
…

- Action localization: search locations of an action in a video



Action classification – examples

diving

running

swinging

skateboarding

UCF Sports dataset (9 classes in total)

Actions classification - examples

answer phone

hand shake

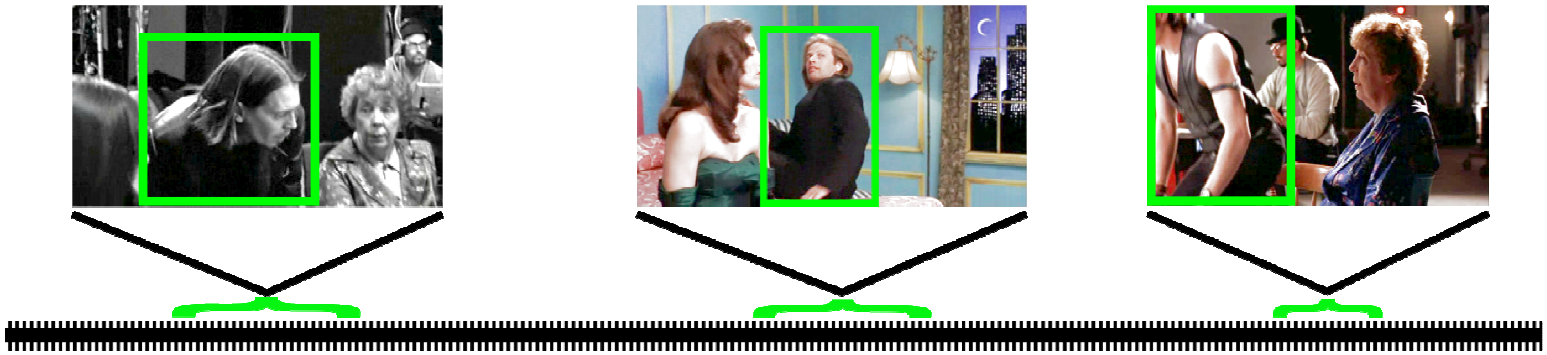
running

hugging

Hollywood2 dataset (12 classes in total)

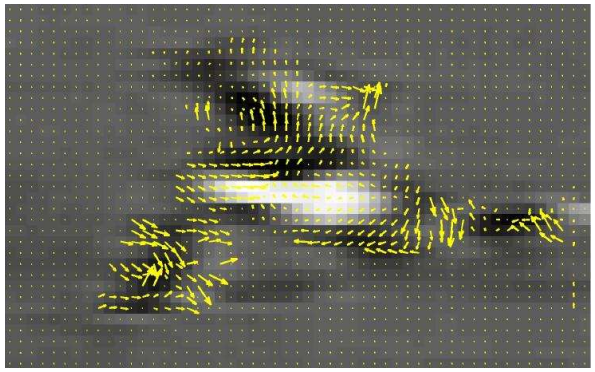
Action localization

- Find if and when an action is performed in a video
- Short human actions (e.g. “sitting down”, a few seconds)
- Long real-world videos for localization (more than an hour)
- Temporal & spatial localization: find clips containing the action and the position of the actor

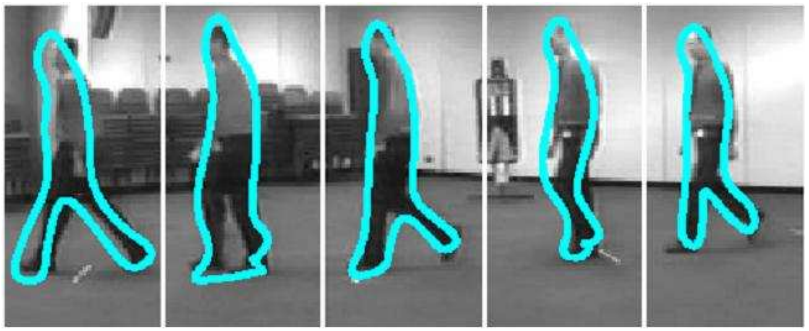


State of the art in action recognition

Motion history image
[Bobick & Davis, 2001]



Spatial motion descriptor
[Efros et al. ICCV 2003]

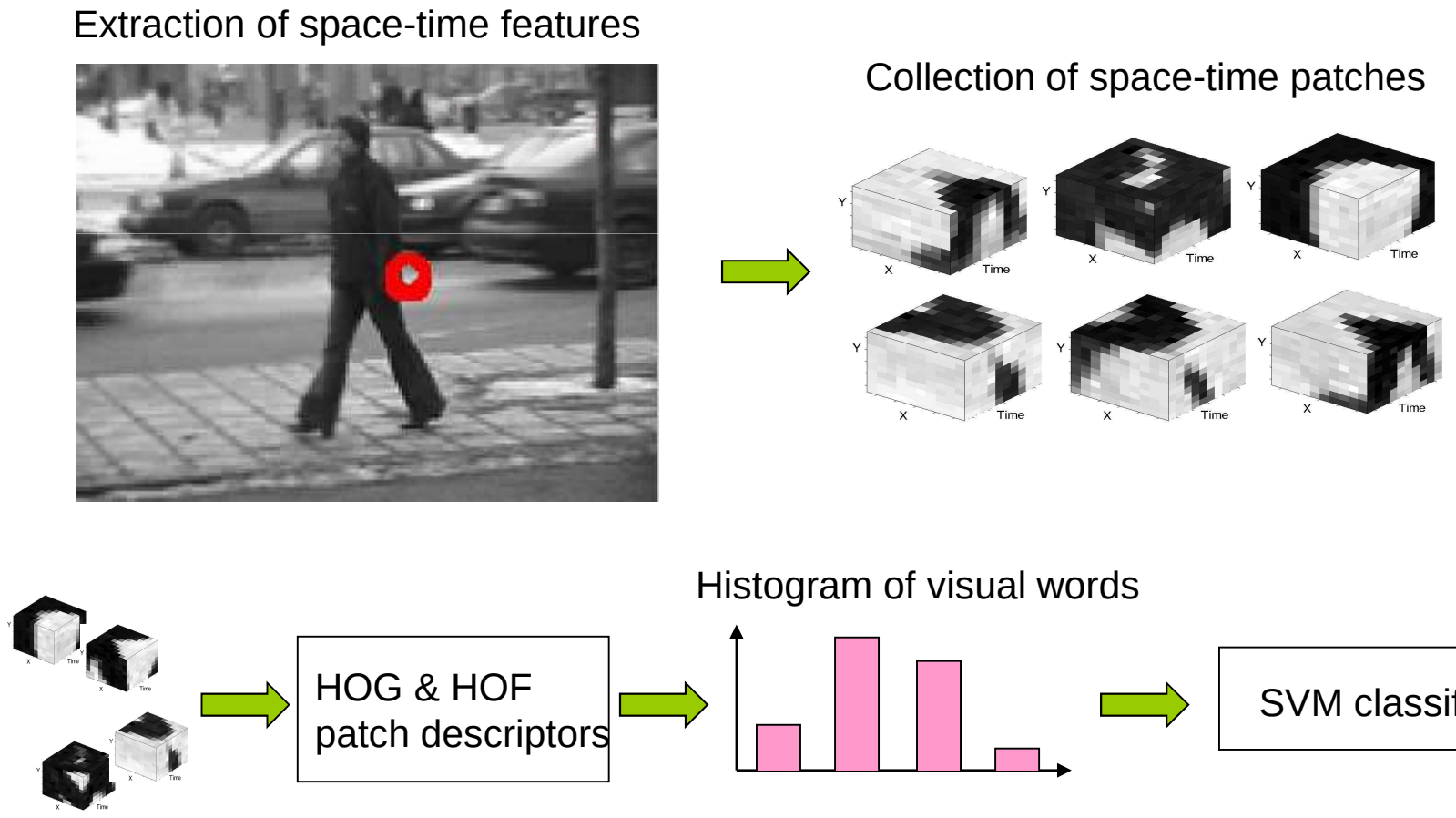


Learning dynamic prior
[Blake et al. 1998]

Sign language recognition
[Zisserman et al. 2009]

State of the art in action recognition

- Bag of space-time features [Laptev'03, Schuldt'04, Niebles'06, Zhang'07]

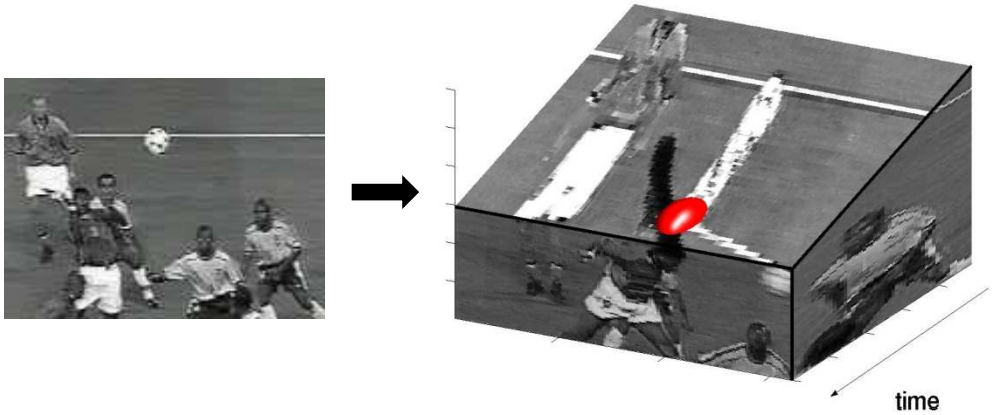


Space-time features

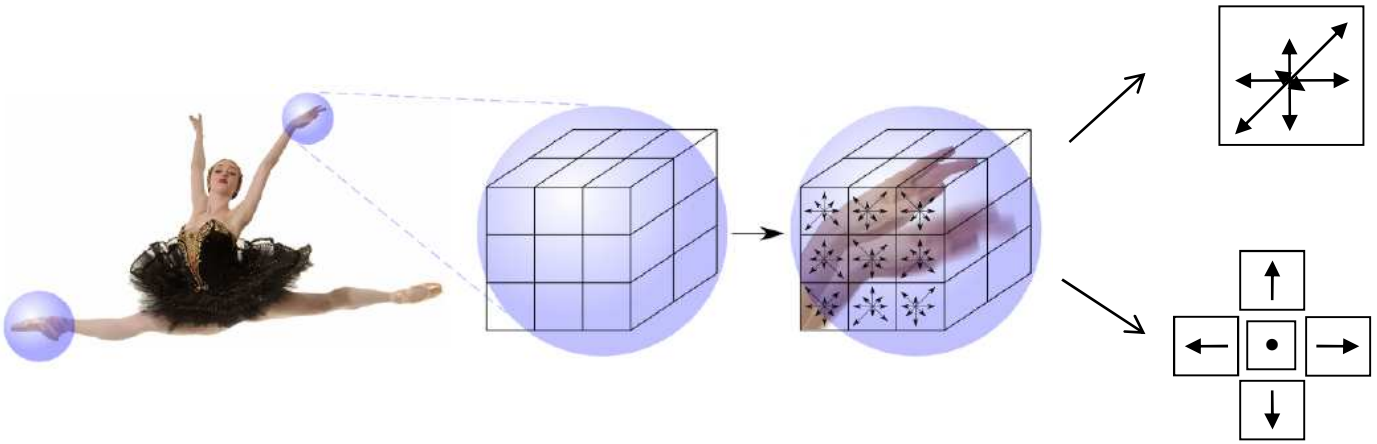
- Detector [Laptev'05]

$$H = \det(\mu) + k\,\mathrm{tr}^3(\mu)$$

$$\mu = \begin{pmatrix} I_x I_x & I_x I_y & I_x I_t \\ I_x I_y & I_y I_y & I_y I_t \\ I_x I_t & I_y I_t & I_t I_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$



- Descriptor

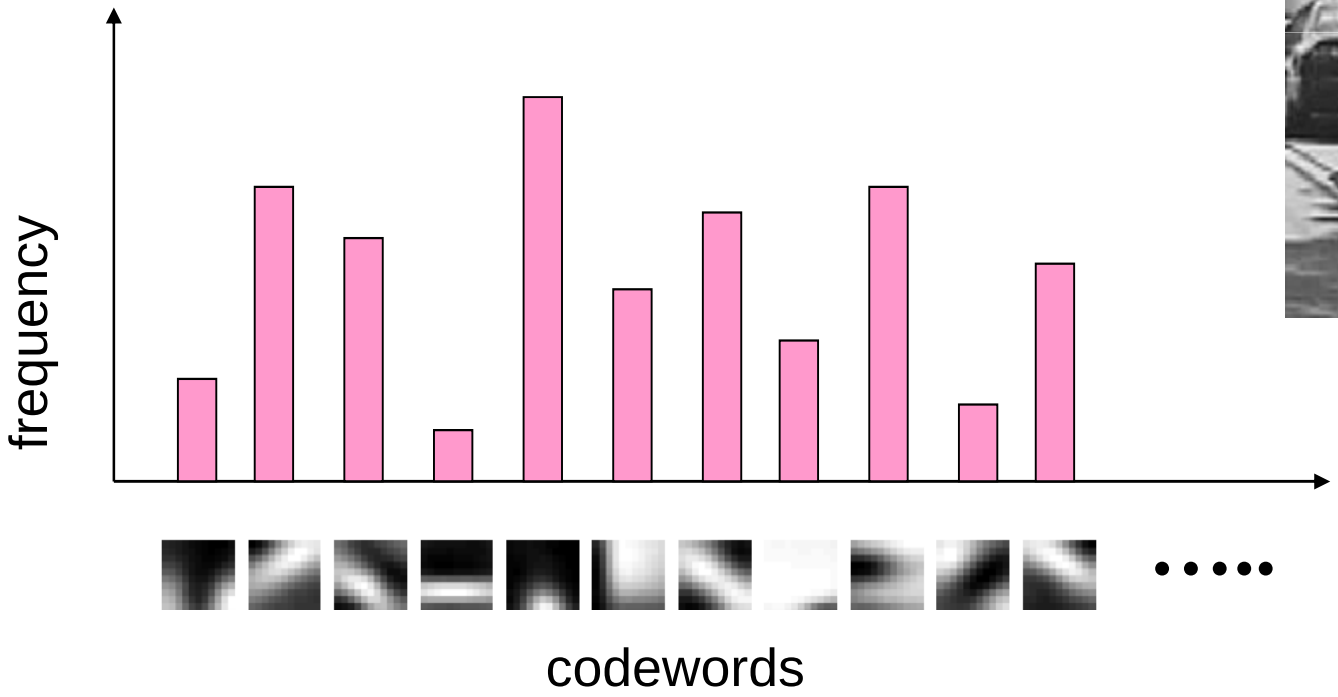


Histogram of oriented spatial grad. (HOG)

Histogram of optical flow (HOF)

Bag of features

- Cluster descriptors with k-means (~4000 clusters)
- Assign each descriptor to the closest center
- Measure frequency



Bag of features

- Advantages
  - Excellent baseline
  - Orderless distribution of local features

- Disadvantages
  - Does not take into account the structure of the action, i.e., does not separate actor and context
  - Does not allow precise localization
  - STIP are sparse features