Actom Sequence Models for Efficient Action Detection

Adrien Gaidon[1,2] Zaid Harchaoui[2] Cordelia Schmid[2]
[1] Microsoft Research - INRIA joint center [2] LEAR, INRIA Grenoble - LJK

CVPR 11 20 COLORADO 11

Problem

- Find ***if*** and ***when*** an action is performed in a video
- Temporal detection of ***short*** actions (a few seconds, e.g. "sitting down")
- Search in ***long*** un-segmented video sequences (several hours)
- Large ***real-world*** video databases (e.g. movies)

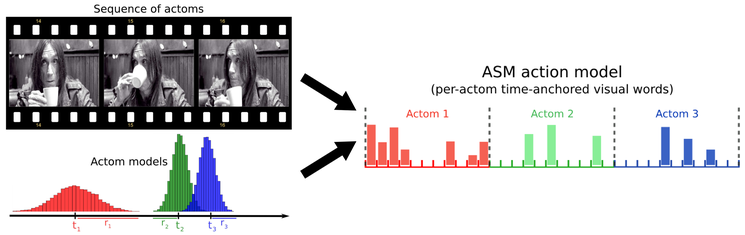
Proposed approach

Main idea

- Improve detection using the global **temporal structure of actions**
- Model as sequences of **"action atoms"**

ASM: Actom Sequence Model

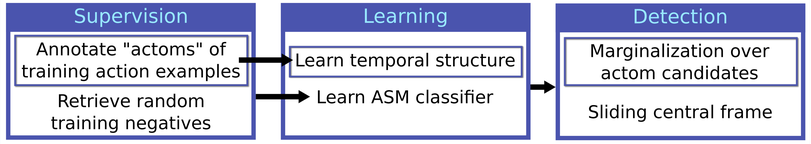
- **ASM**: sparse, efficient and flexible model of actions as **sequences** of histograms of time-anchored visual features



Automatic Temporal Detection with Actoms

- Learn the temporal distribution of actoms from the training data
  Use as a **prior on the global temporal structure of actions**
- Evaluate probability of an action centered on a particular frame by **marginalizing over learned candidate actoms**
- Detection by **sliding central frame** and non-maxima suppression

Overview



Training

Actoms

- Atomic action units, *i.e.* action specific short key events, whose *sequence* is characteristic of the action
- Actoms for training examples are obtained manually by annotating a few key frames (3 in practice)
- Actoms are automatically detected at test time

ASM

- Concatenation of actom descriptors: aggregation of actom-anchored spatio-temporal visual words (temporally structured extension of bag-of-features, using quantized HOG-HOF descriptors at spatio-temporal interest points)
- Two parameters: overlap ρ between adjacent actoms (defines a flexible time-span) and "peaky-ness" p of the time-dependent soft-voting (small p: actom models similar to bag-of-features, high p: keyframe-like actoms)

ASM classifier

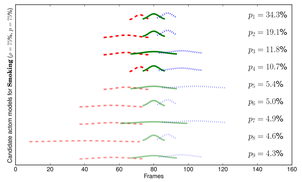
- Non-linear binary SVM with random training negatives, intersection kernel and probability outputs
- Estimates posterior probability of an action knowing its actoms

Temporal Detection

Prior on temporal structure

- Learn a non-parametric generative model of the temporal structure from the training examples
- Kernel density estimation over inter-actom spacings followed by a discretization step
- Result: discrete distribution $\hat{\mathcal{D}}$ with a small support of few candidate actom spacings $\hat{\mathbf{\Delta}}_j$

$$\hat{\mathcal{D}} = \{(\hat{\mathbf{\Delta}}_j, \hat{p}_j)\ ,\ \ j = 1 \cdots K\}, \quad \hat{p}_j = \mathbf{P}(\hat{\mathbf{\Delta}}_j)$$

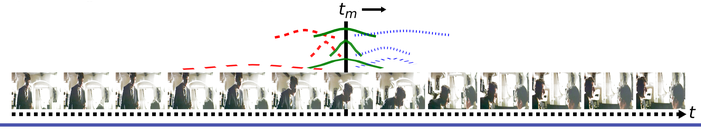


Sliding central frame

- Evaluate probability of action centered on frame t_m by marginalizing over the candidate actom sequences

$$\mathbf{P}(\text{action at } t_m) = \sum_{j=1}^{K} \mathbf{P}(\text{action at } t_m | \hat{\mathbf{\Delta}}_\mathbf{j}) \mathbf{P}(\hat{\mathbf{\Delta}}_\mathbf{j})$$

- In a long video stream, evaluate every N frames (*N=5* in practice)
- Post-processing with a non-maxima suppression step to filter out close detections



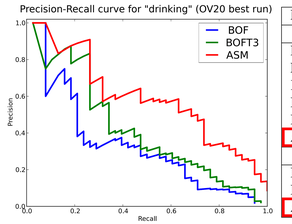
Experiments

Datasets

- "Coffee & Cigarettes": detecting "drinking" and "smoking" actions in approx. 36 000 frames from the movie "Coffee & Cigarettes" [I. Laptev and P. Perez, Retrieving actions in movies. In ICCV 2007]
- "DLSBP": detecting "opening a door" and "sitting down" actions in approx. 443 000 frames from three Hollywood movies [O. Duchenne et al., Automatic annotation of human actions in video. In ICCV 2009]

Quantitative results

- Performance measure: Average Precision (AP) computed using two detection criteria w.r.t. ground truth
  - OV20: detection if temporal overlap is greater than 20% (loose criterion)
  - OVAA: detection if it contains all ground truth actoms (strict criterion)
- ASM improves over state of the art methods, bag-of-features (BOF) and its temporally structured extension (BOFT3: sequence of start, middle and end BOFs [I. Laptev et al., Learning realistic human actions from movies. In CVPR 2008])
- **ASM detections are more accurate** (results of BOF and BOFT3 drop significantly from OV20 to OVAA)



Method	"Drinking"	"Smoking"
matching criterion: OV20		
DLSBP [3]	40	NA
LP [12]	49	NA
KMSZ [9]	54.1	24.5
BOF	36 (±1)	19 (±1)
BOF T3	44 (±2)	23 (±3)
ASM	**57** (±3)	**31** (±2)
matching criterion: OVAA		
BOF	11 (±2)	1 (±0)
BOF T3	18 (±3)	4 (±1)
ASM	**50** (±5)	**22** (±2)

Method	"Open Door"	"Sit Down"
matching criterion: OV20		
DLSBP [3]	13.9	14.4
BOF	12.2	14.2
BOF T3	11.5	17.7
ASM	**16.4**	**19.8**
matching criterion: OVAA		
BOF	9.9	5.8
BOF T3	5.1	13.1
ASM	**14.9**	**16.7**

Qualitative results

- Central frames of the top 5 actions detected with ASM for "drinking" and "open door" (only #2 of "open door" is a false positive)

- Automatically detected actom sequences for an "open door" and a "sit down" action (frames are subsampled, size is proportional to the vote in the predicted ASM model, using the latest actom for overlaps)

Conclusion

- **ASM**: efficient model of actions with a **flexible sequence of key semantic sub-actions (actoms)**
- Principled multi-scale detection using a **prior on temporal structure**
- **ASM outperforms bag-of-features, rigid temporal structures and state of the art**

Data and more information at http://lear.inrialpes.fr/people/gaidon/