Actom Sequence Models for Efficient Action Detection

Adrien Gaidon\textsuperscript{1,2} Zaid Harchaoui\textsuperscript{2} Cordelia Schmid\textsuperscript{2}
\textsuperscript{1} Microsoft Research - INRIA joint center  \textsuperscript{2} LEAR, INRIA Grenoble - LJK

Overview

- **Supervision**: Annotate “actoms” of training action examples
- **Learning**: Learn temporal structure
- **Marginalization over actom candidates**: Learn ASM classifier
- **Detection**: Sliding central frame

**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
- 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 $\rho$ between adjacent actoms (defines a flexible time-span) and “peakiness” $p$ of the time-dependent soft-voting (small $p$: keyframe-like actoms)

**ASM classifier**
- Non-linear binary SVM with random training negatives, intersection kernel and probability outputs

**Prior on temporal structure**
- Learn a non-parametric generative model of the temporal structure from the training examples
- Normal density estimation over inter-actom spacings followed by a discretization step
- Kernel density estimation over inter-actom spacings $\Delta_s$

**Sliding central frame**
- Evaluate probability of an action centered on frame $t_m$ by marginalizing over the candidate actom sequences
- In a long video stream, evaluate every $N$ frames (usually $N = 5$)
- 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 three Hollywood movies
    - OV20: detection if temporal overlap is greater than 20% (loose criterion)
    - OVAA: detection if it contains all ground-truth actions (strict criterion)
- **Quantitative results**
  - ASM improves over state-of-the-art methods, bag-of-features (BOF) and its temporally structured extension (BOFT3) on challenging Hollywood and DLSBP datasets
  - ASM outperforms bag-of-features, rigid temporal structures and state of the art

- **Qualitative results**
  - Central frames of the top 5 actions detected with ASM for “drinking” and “open door”
  - Automatically detected actom sequences for an “open door” and a “sit down” action

**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/