Human Focused Action Localization in Video

Alexander Klaeser[1], Marcin Marszalek[2],
Cordelia Schmid[1], Andrew Zisserman[2]

[1] LEAR, INRIA Grenoble
[2] Visual Geometry Group, University of Oxford

Workshop on Sign, Gesture, Activity
ECCV 2010

The problem

- Goal: **localization** of actions in **realistic** video
  - localization in space (where)
  - localization in time (when)
  - uncontrolled environment (movies)

The challenge

- Why is it **hard**?
  - typical problems: intra/inter class variations, background clutter, occlusions, compression, etc.
  - movie/video-specific: cropping, camera ego-motion, motion blur, interlacing, shot boundaries

Related work

- Localization by tracking and classification
  - No background clutter [Efros ICCV'03, Lu CRV'06]
  - Static camera [Hu ICCV'09, Yuan CVPR'09]
- Action localization in space or in time
  - Periodic actions [Niebles BMVC'06]
  - Temporal alignment [Duchenne ICCV'09]
- Action localization in space-time
  - Keyframe priming [Laptev ICCV'07]
  - Hypothesis generation [Willems BMVC'09]

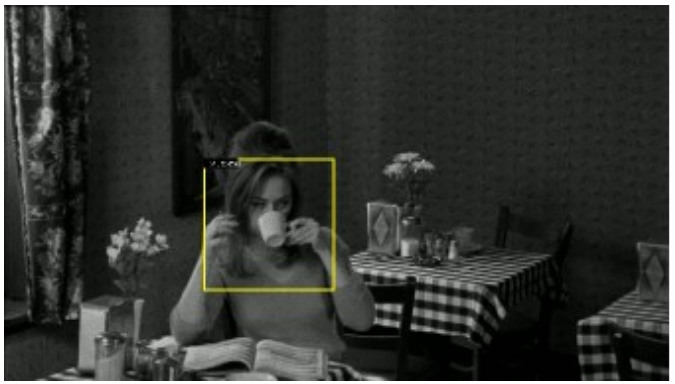
Our approach

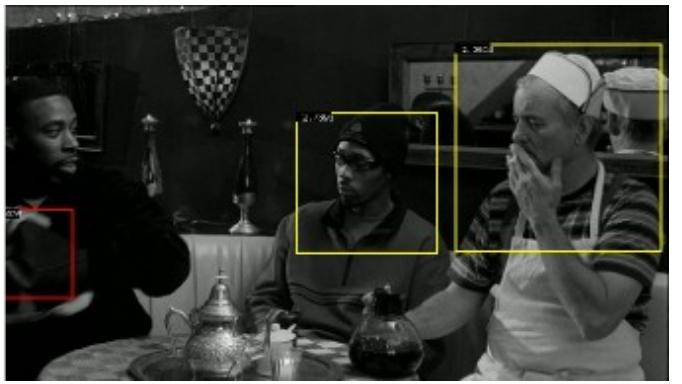
- Stems from the simple observation that **actions** are performed by **actors**
  - spatial location is determined by actor's position and does not depend on the type of action
  - temporal extent can be found efficiently and more accurately after the spatial location is already known
- We develop a robust actor detector and tracker
- We propose a track-aligned action descriptor
  - efficient action localization via sliding window on tracks

Human detection and tracking

Robust human detection

- HOG detector [Dalal05] trained for upper bodies
  - 1122 frames from Hollywood2 training movies
    - 1607 annotations jittered to 32k positive samples
    - 55k negatives sampled from the same set of frames
    - 150k hard negatives
  - 193 frames from Coffee&Cigarettes training stories
    - additional jittered 6k positives and 9k hard negatives





Smoothing and interpolation

- Tracking-by-detection [Everingham09]
  - KLT tracker yields feature trajectories
  - detections are clustered together (agglomerative clustering) based on connectivity score
- Smoothing + interpolation for continuous tracks can be done very efficiently

$$\min_{\{\mathbf{p}_t\}} \sum_{t \in T} \left(||\mathbf{p}_t - \bar{\mathbf{p}}_t||^2 + \lambda^2 ||\mathbf{p}_t - \mathbf{p}_{t+1}||^2 \right)$$

$\mathbf{p}_t = (x_t, y_t, w_t, h_t)$ denotes the position

$\bar{\mathbf{p}}_t = (\bar{x}_t, \bar{y}_t, \bar{w}_t, \bar{h}_t)$ are the detections

λ is a temporal smoothing parameter

Tracks post-processing

- **Final classification of tracks to improve precision at high recall**
- SVM classifier is learned on 12 different measures characterizing a track – those are min, max and averages (as applicable) of:
  - track length (false tracks are often short)
  - upper body SVM detection score
  - scale and position variability (those often reveal artificial detections)
  - occlusion by other tracks (patterns in the background often generate a number of overlapping detections)

Detected human tracks

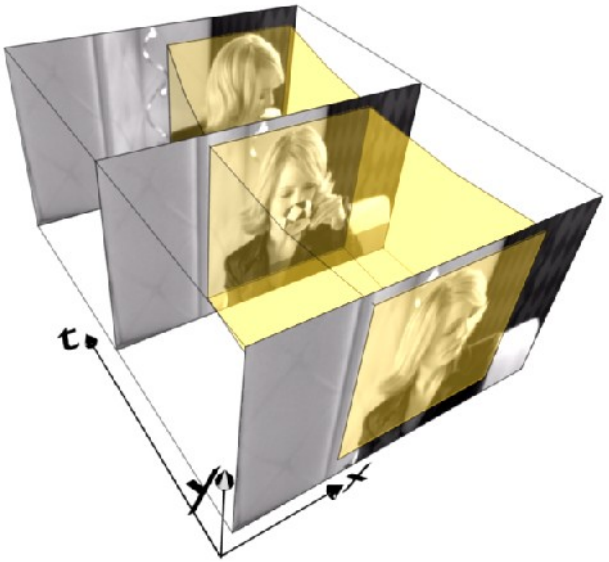
Action localization in tracks

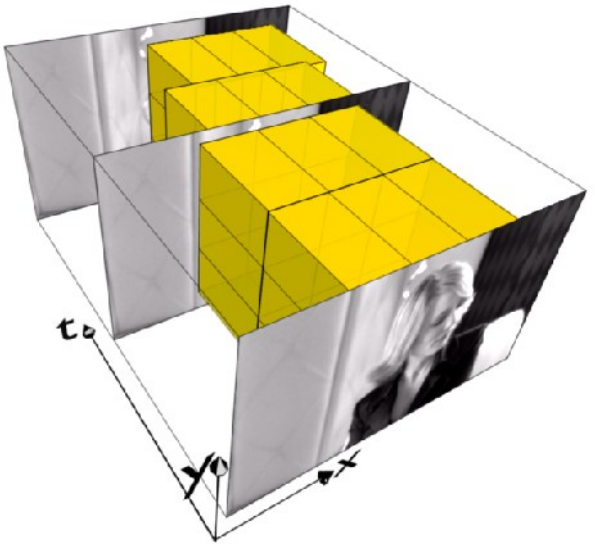
Why tracks-based descriptor?

- Brings focus to an object of interest
  - Background clutter can be reduced
  - Geometrically stronger models can be built
  - See our **technical report** for more details
- Adapts to human motion
  - Invariant and discriminative at the same time
- Allows efficient action localization
  - Human tracks can be reused for multiple actions
  - Temporal search is linear in tracks

Action descriptor

- Grid layout of N x N x M cells
- Cells overlap spatially with 50%
- Each temporal slice is aligned to the track (follow movement)
- Each cell 3D HOG histogram
  - Icosahedron for orientation quantization (half orientation)
- Layout optimization to 5x5x5
- Descriptor size: 1250





Action localization

- Sliding window approach
  - Exhaustive search over all tracks, track positions and action lengths
  - Very efficient in fact, in practice linear in video time
- Further speedup: 2-stage classification
  - Linear SVM as first classifier, generate hypotheses via non-maxima suppression
  - Re-evaluation of final hypotheses with non-linear SVM (RBF)

Results

Coffee and Cigarettes

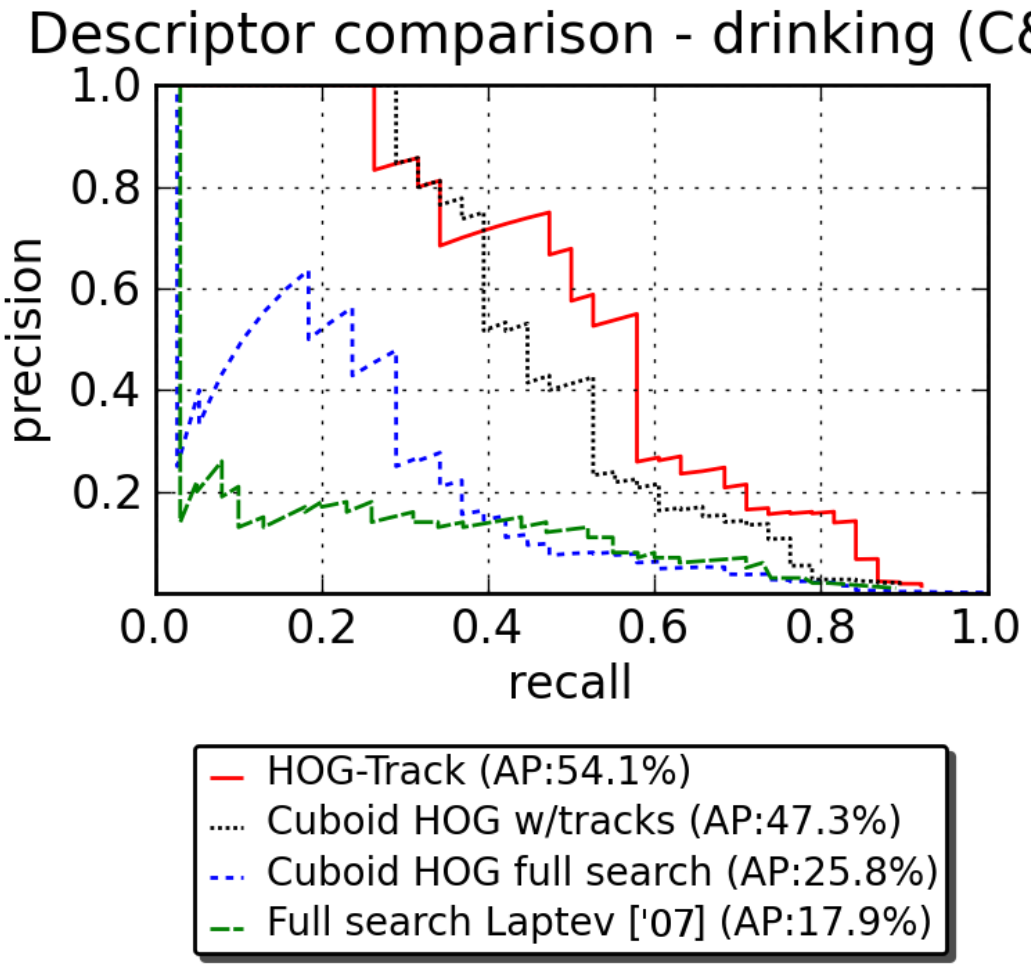
- We use the original split by stories [Laptev'07]
  - training: 6 stories, 40min, 106 drinking, 90 smoking actions (+"Sea of Love" and "Lab" videos)
  - test-drinking: 2 stories, 24min, 38 drinking actions
  - test-smoking: 3 stories, 21min 46 smoking actions (originally validation set)
- Average Precision is used for evaluation

training

test-smoking

test-drinking

Do tracks really help?



Baselines:

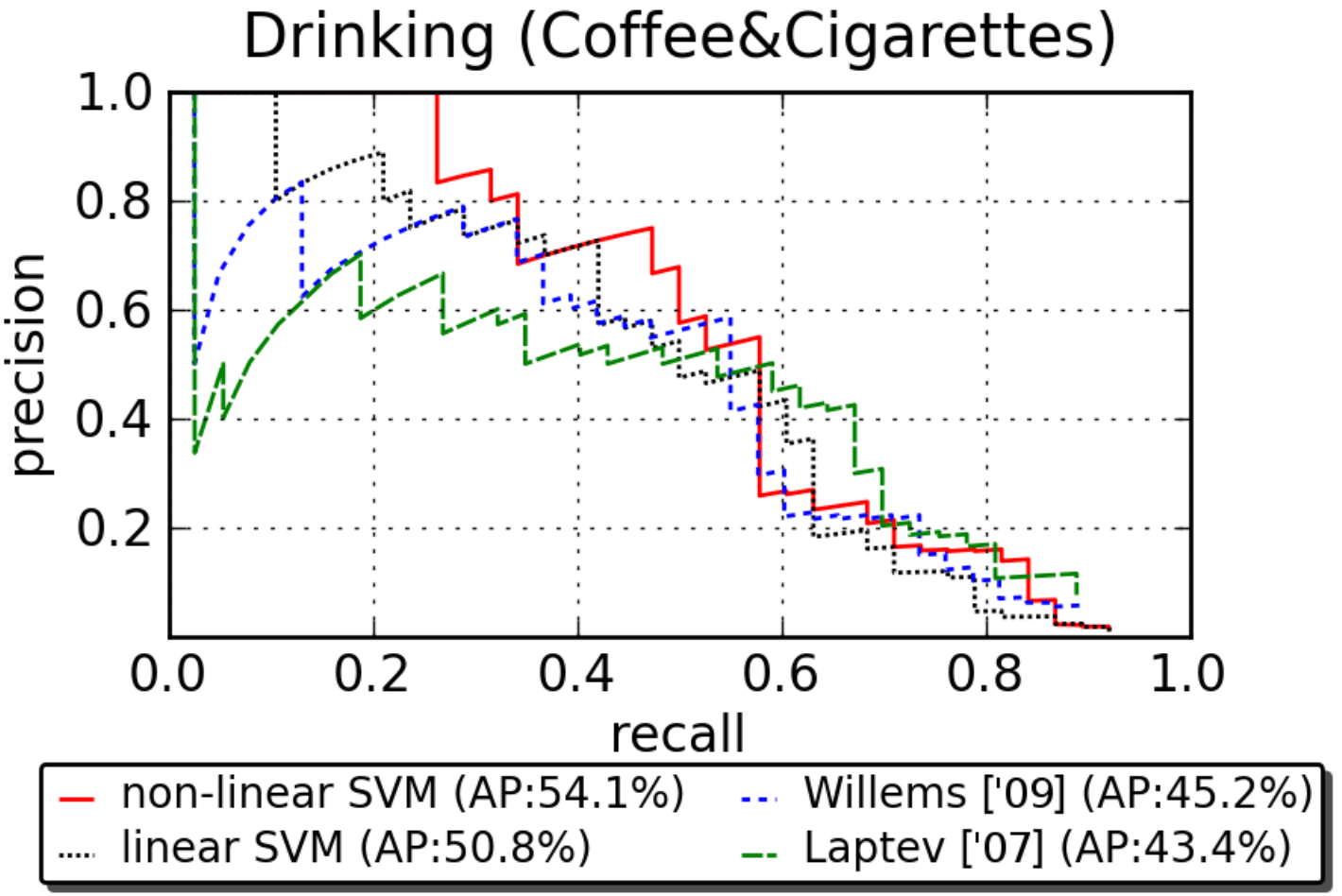
1) Cuboid classifier, full search in video

2) Cuboid classifier, centered on tracks

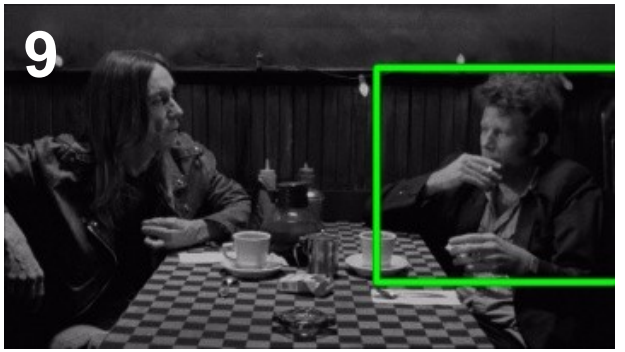
3) Laptev's baseline, full search

Results for drinking

Results for drinking



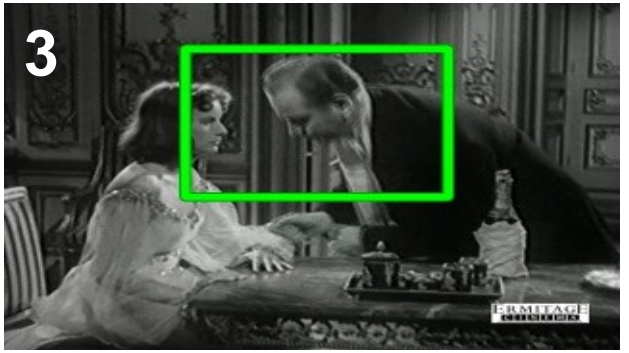
Top 9 results for smoking

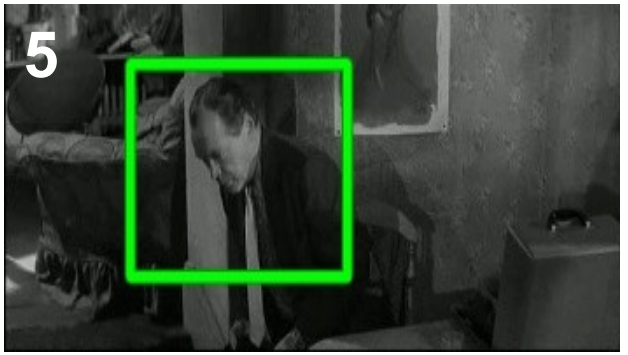


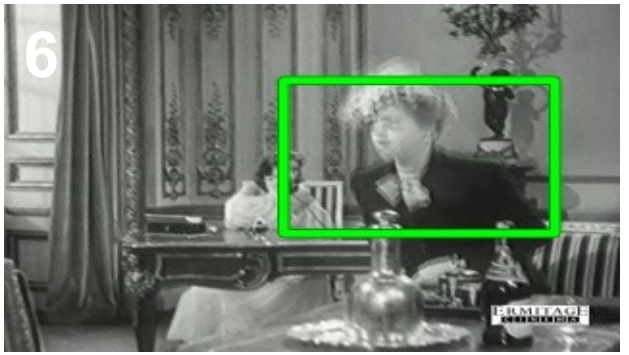
Hollywood localization

- Dataset based on Hollywood2 data and split
  - ~2h of video data in total (~1h training, ~1h test)
- We annotate the spatial and temporal extent of "**phoning**" and "**standing up**" actions
  - 153 "phoning" actions (73 training, 80 test)
  - 274 "standing up" actions (129 training, 145 test)
- Average Precision is used for evaluation

Top 9 results for standing up







Top 9 results for phoning

Why tracks help

- Classification task is simplified
  - the “action world” gets restricted to actors
- Better modeling capability
  - descriptor follows actor movements
- Search space is reduced heavily
  - less false positives

Complexity

- Exhaustive search
  - **5D** search (x,y,t position, x/y,t scale) with rigid **3D** action representation
  - 25min video: 100h processing time **per action**
- Our approach
  - Human detection: **4D** search (x,y,t position, x/y scale) with **2D** representation
  - Action localization: 2D search (t position, t scale) with flexible action representation
  - 25min video: 13h per video + 10min per action

Thank you

Action detections

Human detections

Human tracks