Learning human actions in video

Alexander Kläser
INRIA Grenoble, LJK

PhD thesis defense, 31 July 2010

Thesis advisor: Cordelia Schmid

Goal of this thesis

Recognizing actions in realistic videos

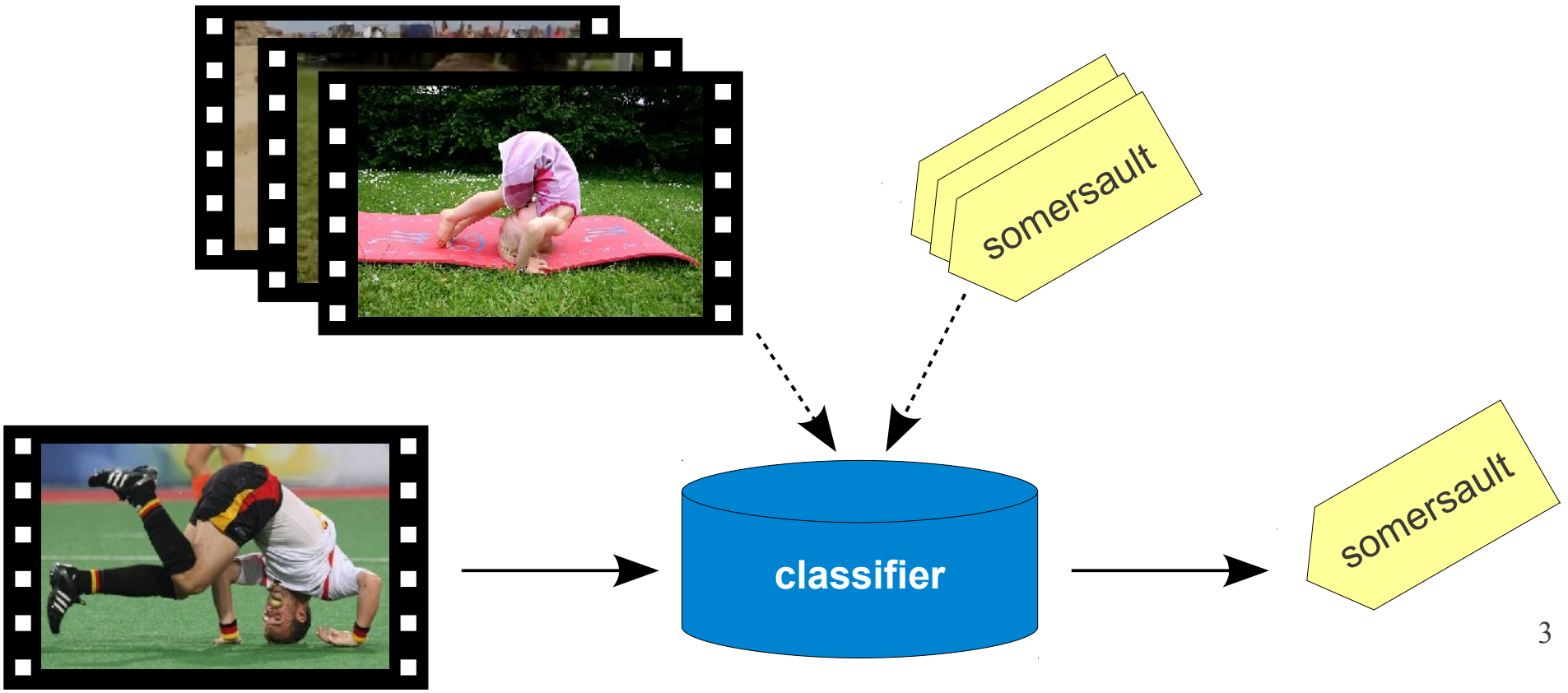
- Actions: focus on visible low-level action primitives and actions of a rather generic type
  - e.g.: running, drinking, smoking, answering phone, standing up, hugging, shaking hand, punching, ...
- Realistic video: uncontrolled video data, such as movies or internet videos



Goal of this thesis

Task 1: Action Classification

- Label a given video sequence as belonging to a particular action or not



Goal of this thesis

Task 2: Action Localization

- Determine the beginning, end, and spatial extent of an action in a video sequence
- Much more challenging !
  - ... as for for object localization in images (VOC)
  - Certain type of actions are rare

Why it is challenging ?

Video-specific:

camera ego-motion, shot boundaries, motion blur, interlacing, compression artifacts etc.

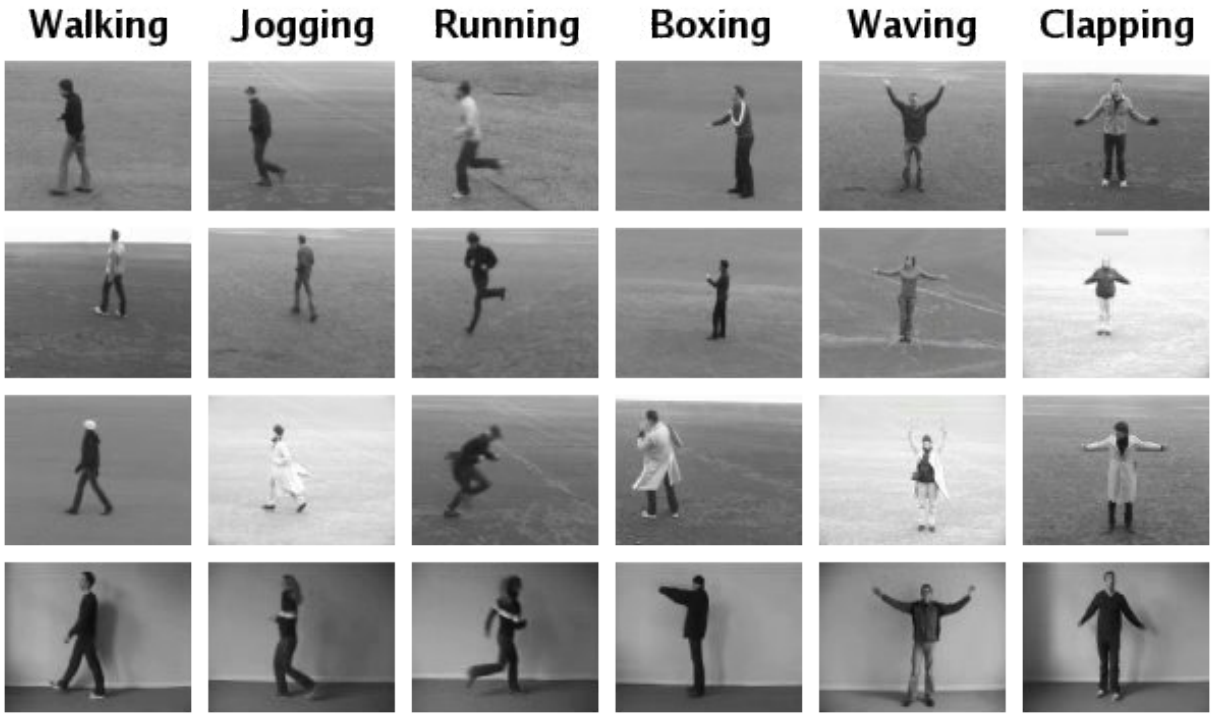
Typical problems:

intra/inter class variations, pose variations, background clutter, occlusions/cropping, illumination conditions, rareness etc.

Motivation & applications

- More and more growing amount of video data...
  - Videos uploaded **per minute** on YouTube increased from **6h** in 2007 to **20h** in 2009 (+330%)
- Many works still use simplistic video data (no clutter, simple background, artificial actions etc.)
- Applications
  - Video search + indexing (e.g., for film archives, websites), commonly based on text (e.g., YouTube)
  - Surveillance applications
  - Human-computer interfaces, computer games (e.g., Microsofts Project Natal)
  - Film industry (animation, special effects, video editting)
  - Analysis of sport athletics, dance choreography

Simplistic vs. realistic data: KTH



- 6 action classes, 2391 video samples in total
- Homogeneous background, artificial actions
- State-of-the-art: 94.5% [Gilberts09], 94.1% [Han09]

Simplistic vs. realistic data: HW2

- 12 action classes, 1707 samples from 69 Hollywood movies
- Large intra-class variations, clutter, camer ego-motion etc.
- State-of-the-art: 50.9% [Gilbert10], 42.1% [Han09]

Outline

- Bag-of-features
- Local spatio-temporal HOG3D descriptor
- Evaluation of local feature detectors & descriptors
- Human focused action localization in space-time
- Summary & conclusion

Outline

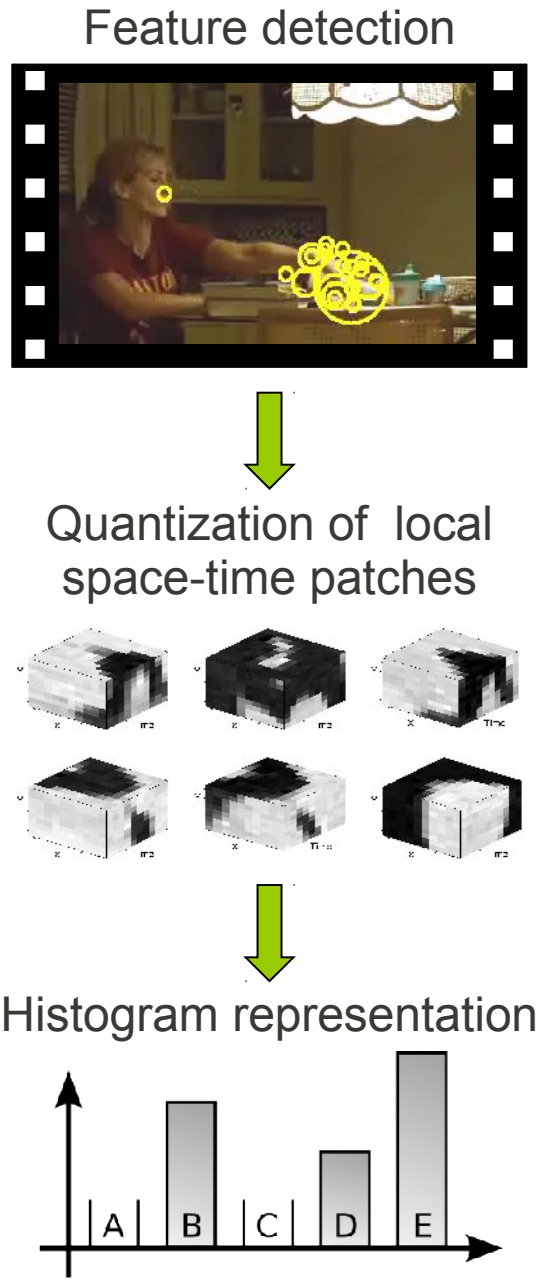
- **Bag-of-features**
  - History, overview, motivation
- Local spatio-temporal HOG3D descriptor
- Evaluation of local feature detectors & descriptors
- Human focused action localization in space-time
- Summary & conclusion

Short history of bag-of-features

- Origin from text documents [Salton68]
  - Bag-of-words counts the occurrence of words
  - Common representation for text documents
- Application to images [Culana01,Sivic03,Csurka04,Sivic05]
  - Local image feature descriptors replace “words” => bag-of-features (BoF)
  - Current state-of-the-art for image classification [VOC09]
- Application to action classification in videos [Schüldt04,Dollár05,Niebles06]

Bag-of-features overview

- Detection and description of local space-time features
- Codebook generation via clustering of training features (e.g., k-means, k=4000)
- Representation with occurrence histogram
  - Each feature is assigned to its closest cluster center (visual word)
- Classification of histograms (e.g., SVM with χ^2-kernel)



Motivation

- No prior knowledge needed (human position, body parts, clutter) ☺
  - Depending on the video data, human / limb detection might not be feasible
- BoF can be applied to challenging data ☺
- Straightforward approach ☺
- Works well in practice ☺
- No separation between background and foreground ☹
- No notion of geometry (extensions exist) ☹

Outline

- Bag-of-features
- **Local spatio-temporal HOG3D descriptor**
  - Motivation, approach, parameter optimization
- Evaluation of local feature detectors & descriptors
- Human focused action localization in space-time
- Summary & conclusion

[A. Kläser, M. Marszalek, and C. Schmid. *A spatio-temporal descriptor based on 3D-gradients*. In BMVC, 2008]

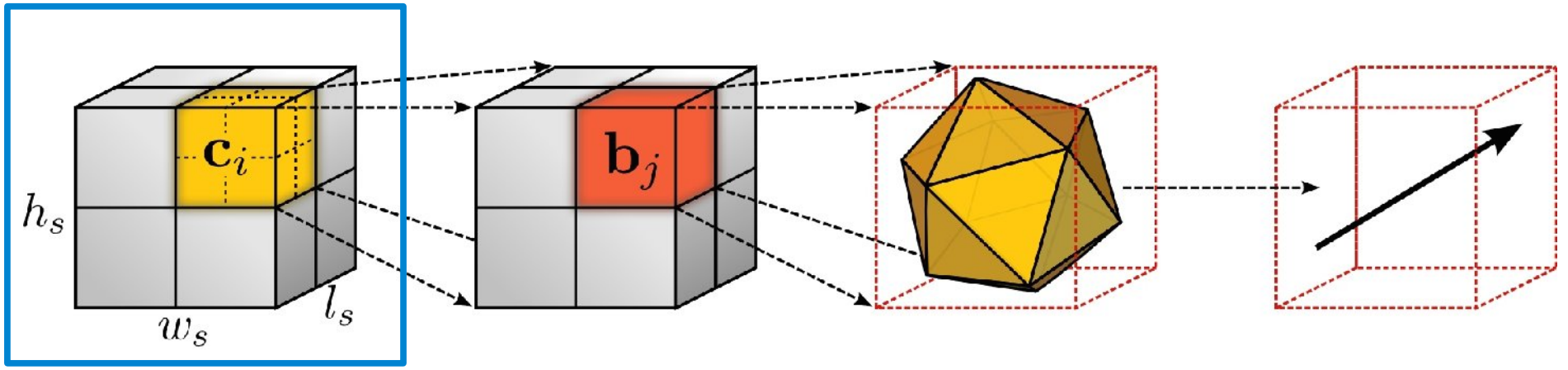
Motivation

- Many concepts have been sucessfully extended from static images to videos
  - Feature detectors/descriptors, BoF representations, voting approaches etc.
- Few spatio-temporal descriptors exist that combine spatial with temporal information:
  - Optical flow and (spatial) gradient orientations [Laptev08]
  - Spatio-temporal gradient magnitudes [Laptev04,Dollár05]
  - Spatio-temporal SIFT [Scovanner07]
  - Extended SURF descriptor [Willems08]
- Histograms of Oriented Gradients (HOG) work well for images [Dalal06, Lowe04]

Overview HOG3D

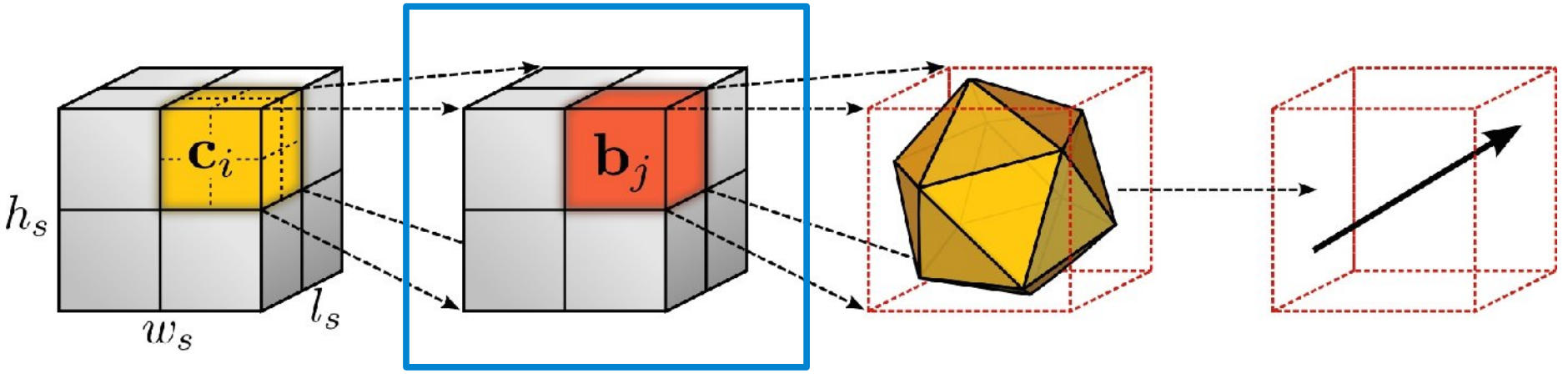
- **Main idea:** extension of HOG using 3D gradients
  - Spatial orientation captures appearance information and temporal orientation captures velocity
  - Gradients are straight forward to compute
- **What is new ?**
  - Quantization of 3D gradients with regular polyhedra
  - Gradient computation using integral videos [Ke05, Willems08]
    - Efficient gradient computation for arbitrary scales
  - Optimization of descriptor parameters
  - Extensive evaluation on different datasets

(1) Local descriptor



- Describes local neighborhood around sampling position
  - Sampling point is given by x, t, y position and characteristic spatial and temporal scale σ, τ
  - Spatial and temporal scales need to be separated
- Width/height and length given by: $h = w = \sigma_0 \sigma$, $l = \tau_0 \tau$
- Local neighborhood is divided into M x M x N cells
- For each cell, histograms are computed, normalized, and concatenated

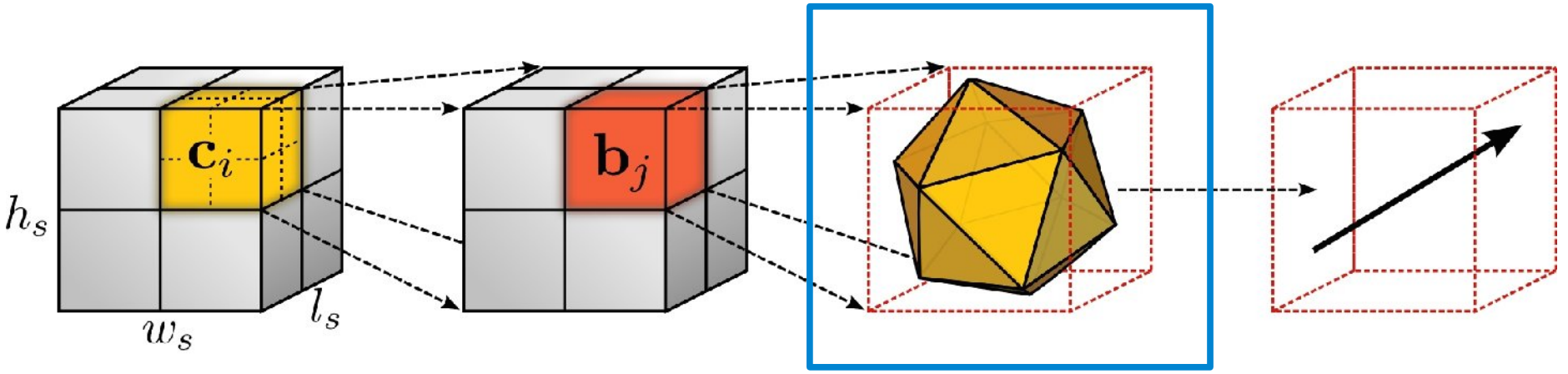
(2) Histogram of oriented gradients



- A cell is divided into S x S x S sub-blocks
- For each sub-block, mean gradients are computed and quantized
- All votes are summed up for final histogram

$$\mathbf{h}_i = \quad = \sum_{j=1}^{S^3} \mathbf{q}_j \qquad \mathbf{q}_j =$$

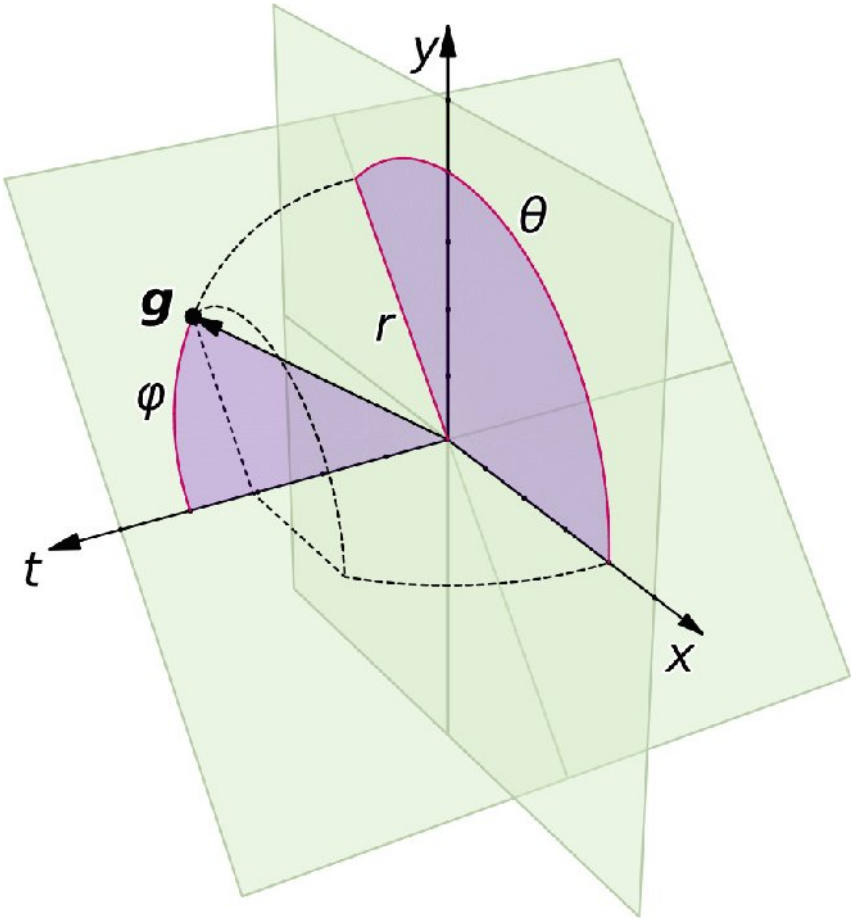
(3) Orientation quantization



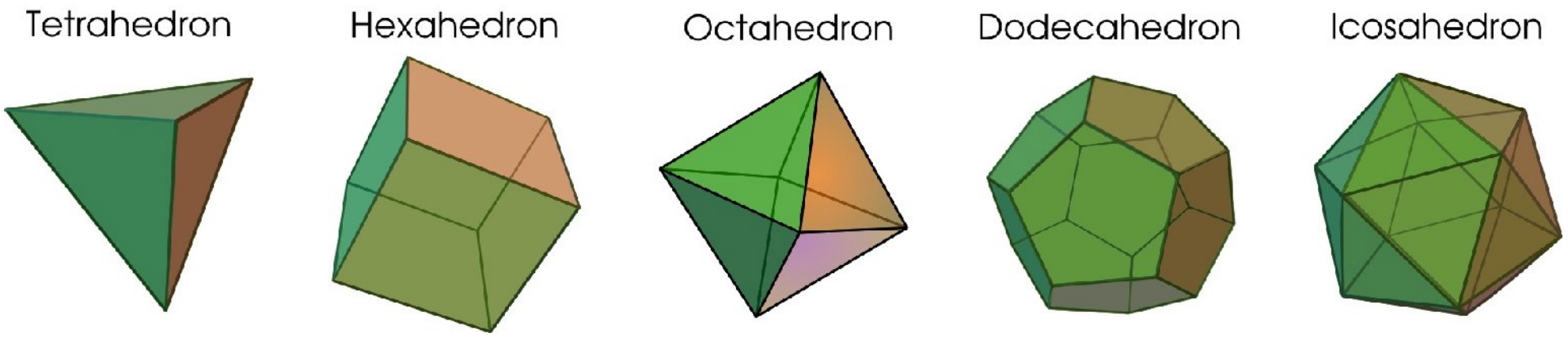
- Gradient orientation is more robust to illumination changes than magnitude [Freeman95]
- Quantization for 3D gradients
  - Spherical coordinates (longitude and latitude)
  - Regular polyhedra (each face is one bin)

(3) Spherical coordinates

- Azimuth (θ) and elevation angle (φ) are quantized into a regular grid
- Spatial and temporal resolution can be controlled separately ☺
- Leads to singularities at poles ☹
- Size of bins varies ☹

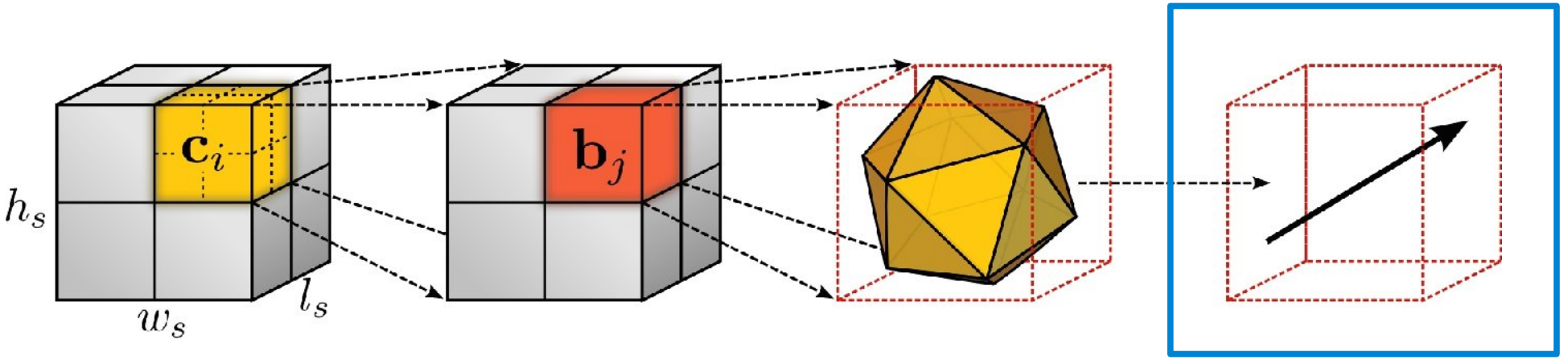


(3) Regular polyhedra



- Also called Platonic solids, there exist only 5
- Faces (used as bins) are congruent and evenly distributed ☺
- Quantization by projecting gradients on axes through polyhedron center and face center
- We use dodecahedron (12 bins) and icosahedron (20 bins) in our experiments

(4) Gradient computation



- Gradients need to be computed for different spatial and temporal scales
- Approximate gradients via integral videos
  - $\nabla L_{\sigma,\tau} = \nabla(G_{\sigma,\tau} * v) = G_{\sigma,\tau} * \nabla v \approx B_{\sigma,\tau} * \nabla v$
  - Constant computation time for gradients at arbitrary scale

Parameter optimization

- Descriptor parameters are optimized via cross-fold-validation on the training set
  - Spatial/temporal support, number of spatial/temporal cells, number of sub-blocks, full/half orientation
  - For spherical coordinates: number of spatial/temporal bins
- Optimization via gradient descent
  - Division of parameter space into rough grid
  - Caching of results, optimization on mean
- Separate optimization on two datasets
  - Simple dataset with uncluttered background (KTH)
  - Realistic dataset based on Hollywood movies (Hollywood2)

Outline

- Bag-of-features
- Local spatio-temporal HOG3D descriptor
- **Evaluation of local feature detectors & descriptors**
  - Motivation & goal, detectors, descriptors, results, conclusion
- Local feature trajectory descriptor
- Human focused action localization in space-time
- Summary & conclusion

[H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. *Evaluation of local spatio-temporal features for action recognition*. In BMVC, 2009]

Motivation & goal

- Motivation
  - Local features have become popular for action recognition
  - Several methods exist for detection/description of local features
  - Existing comparisons are limited [Laptev04, Dollar05, Scovanner07, Jhuang07, Kläser08, Laptev08, Willems08]
    - Different experimental settings and datasets
    - Evaluations limited to only few descriptors
- **Main idea:** thorough evaluation of local video features
  - Systematic evaluation of detector-descriptor combinations
  - Same datasets (varying difficulty): KTH, UCF sports, Hollywood2
  - Same classification method

Evaluated feature detectors

- Harris 3D [Laptev03]
  - Space-time corner detector
  - Based on Harris cornerness criterion
- Gabor [Dollár05]
  - Combination of spatial Gaussian filter and temporal Gabor filters
  - Detection of salient regions undergoing a complex motion
- Hessian 3D [Willems08]
  - Spatio-temporal extension of Hessian saliency measure
  - Approximation with integral videos
  - Detection of spatio-temporal “blobs”
- Dense sampling (in x, y, t and σ, τ)

Evaluated feature descriptors

- HOG/HOF [Laptev08]
  - Based on histograms of oriented (spatial) gradients (HOG) + histograms of optical flow (HOF)
- Gradient [Dollár05]
  - PCA on concatenated pixel gradient values (i.e., spatio-temporal magnitudes)
- Extended SURF [Willems08]
  - Extension of SURF descriptor to videos
  - Weighted sums of axis-aligned 3D Haar Wavelets
- HOG3D (as presented earlier)

Evaluated datasets

- KTH actions
  - 6 action classes, 2391 video samples
  - Homogenous background, artifial actions
  - State-of-the-art: 94.5% [Gilberts09], 94.1% [Han09] (accuracy)
- UCF sport actions
  - 10 action classes, 150 video samples
  - State-of-the-art: 69.2% [Rodriguez'08] (accuracy)
- Hollywood human actions (2)
  - 12 action classes, 1707 samples from 69 different Hollywood movies (spatially subsampled)
  - State-of-the-art: 50.9% [Gilbert10], 42.1% [Han09] (mAP)

KTH actions – results

Descriptors	*Detectors*			
	Harris3D	**Gabor**	**Hessian**	**Dense**
HOG3D	**92.4%**	91.4%	88.1%	88.5%
HOG/HOF	**91.8%**	88.7%	88.7%	86.1%
HOG	80.9%	82.3%	77.7%	79.0%
HOF	**92.1%**	88.2%	88.6%	88.0%
Gradient	-	89.1%	-	-
ESURF	-	-	81.4%	-

- Best results for Harris3D + HOG3D
  - HOG3D parameters learned on KTH training set
- Good results for Harris3D & Gabor detector and HOG/HOF & HOG3D descriptor
- Dense features worse than interest points
  - Large number of features on static background

UCF sports – results

Detectors

Descriptors	Harris3D	Gabor	Hessian	Dense
HOG3D	77.6%	**85.0%**	78.9%	**84.8%**
HOG/HOF	78.1%	77.7%	79.3%	81.6%
HOG	71.4%	72.7%	66.0%	77.4%
HOF	75.4%	76.7%	75.3%	**82.6%**
Gradient	-	76.6%	-	-
ESURF	-	-	77.3%	-

- Best results for Dense / Gabor + HOG3D
  - HOG3D parameter set learned on KTH
- Good results for Dense and HOG/HOF

Hollywood2 actions – results

Detectors

Descriptors	Harris3D	Gabor	Hessian	Dense
HOG3D	44.3%	**46.1%**	43.5%	44.8%
HOG/HOF	45.2%	**46.2%**	**46.0%**	**47.4%**
HOG	32.8%	39.4%	36.2%	39.4%
HOF	43.3%	42.9%	43.0%	45.5%
Gradient	-	45.0%	-	-
ESURF	-	-	38.2%	-

- Best results for Dense + HOG/HOF
- Good results for HOG/HOF and Gabor in general
- HOG3D + Gabor performs well
  - Parameters learned on HW2 train set in **full resolution**
  - For full resolution videos HOG3D + Harris3D yield **48.8%** and HOG/HOF + Harris3D **47.6%**

Conclusion

- Dense sampling outperforms tested detectors in realistic settings (UCF + Hollywood2)
  - Importance of realistic video data
  - Limitations of current feature detectors
  - Note: large number of features (15-20 times more)
- Detectors: Harris3D, Gabor, and Hessian provide comparable results (interest points better than Dense on KTH)
- Descriptors overall ranking:
  - HOG3D & HOG/HOF > Gradient > ESURF & HOG
  - Combination of gradients + optical flow seems good choice

Outline

- Bag-of-features
- Local spatio-temporal HOG3D descriptor
- Evaluation of local feature detectors & descriptors
- **Human focused action localization in space-time**
  - Overview, tracking, action description, results
- Summary & conclusion

[A. Kläser, M. Marszalek, C. Schmid, and A. Zisserman. *Human focused action localization in video*. SGA 2010]

Overview

- **Goal:** action localization in realistic video
- **Main idea:** actions are performed by actors
  - Actor's position generically determines spatial location of action
  - Determine temporal extent after spatial location
  - More efficient and more accurate
- **What is new ?**
  - We develop a robust actor detector and tracker
    - Good human detector and tracker is crucial
  - We propose a track-aligned action descriptor
    - Action localization via sliding window on tracks
  - New localization dataset based on Hollywood movies

Related work

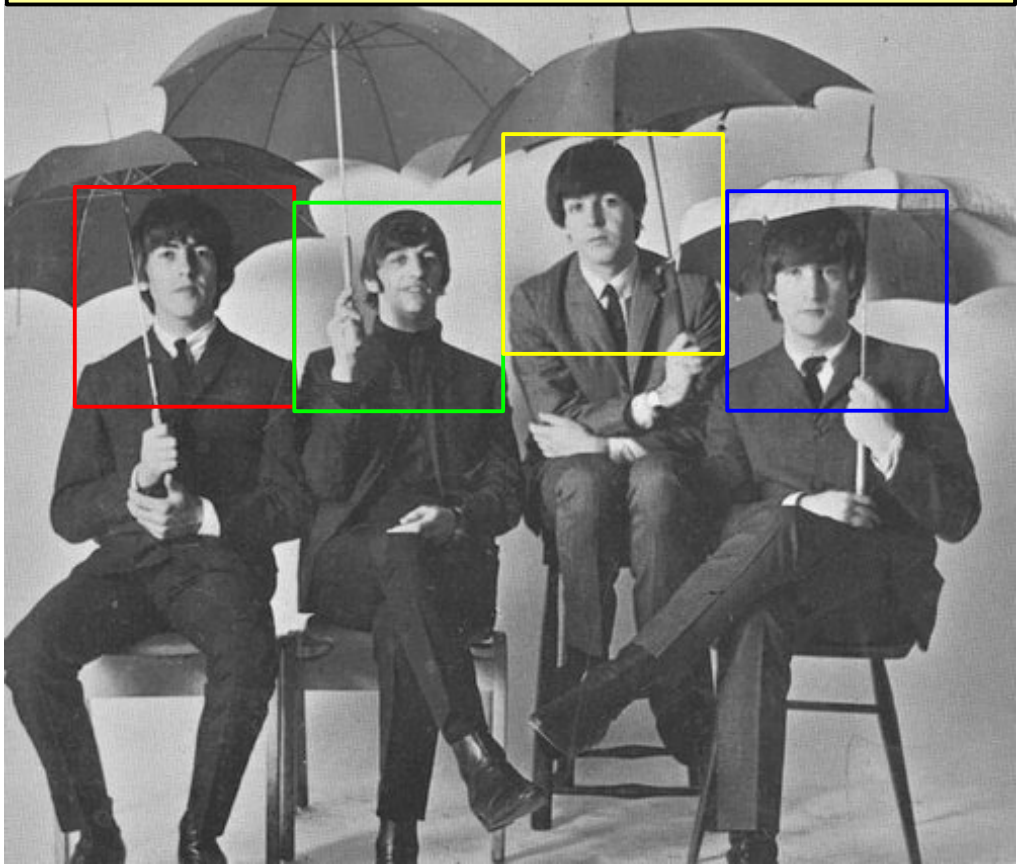
- Keyframe priming [Laptev07]
  - Cuboid classifier
  - Adaboost learns combinations of HOG and HOF features within cuboid region
  - Pre-filtering by action specific pose detector
- Local features based voting [Willems09]
  - Strongest video words vote for action hypotheses
  - Strong hypotheses are evaluated with full BoF cuboid representation
- Shape matching [Ke07]
  - Shape templates are matched to over-segmented videos
  - Combination of shape and optical flow
- Other works concentrate on static cameras [Hu09, Yuan09] or simplified settings [Efros03, Lu06]

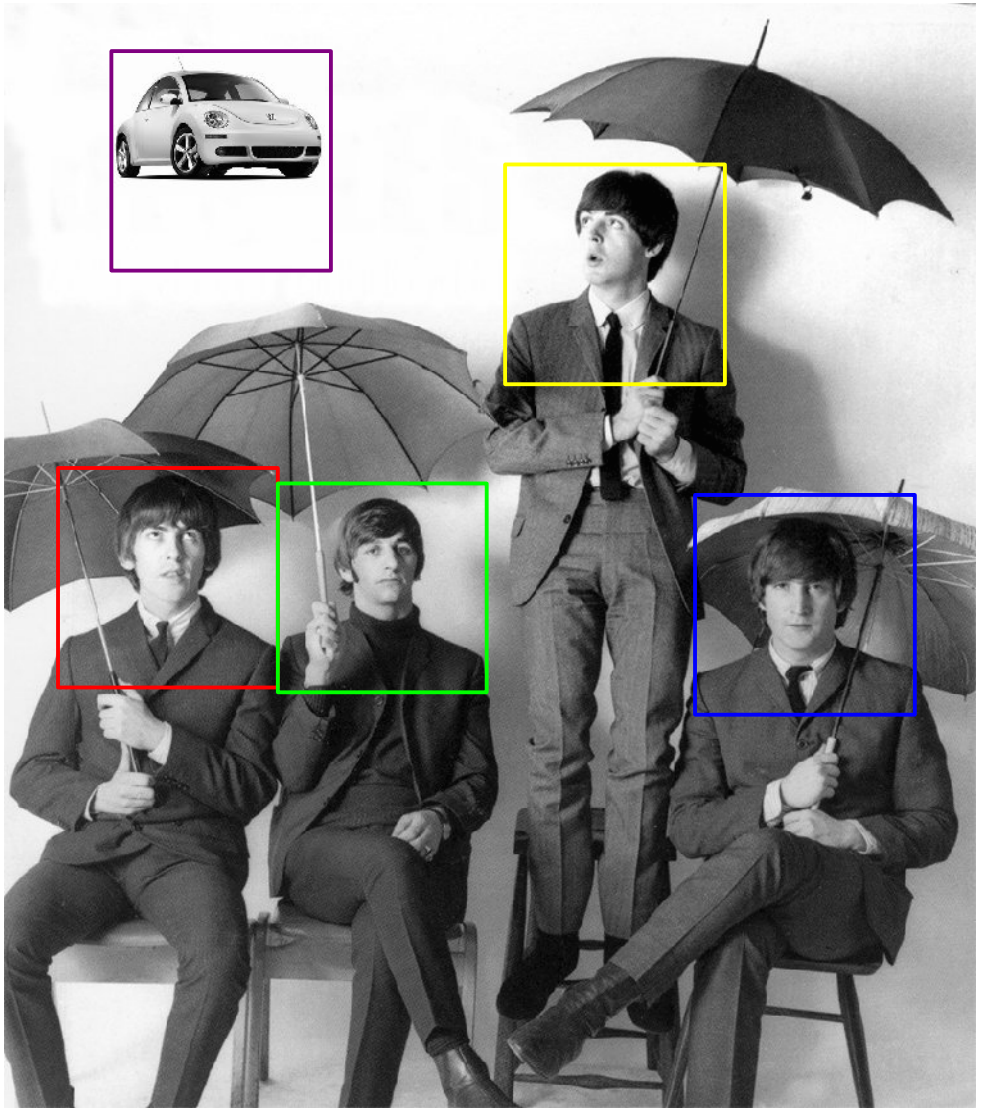
Our approach (illustration)

1. Load the Beatles video

Our approach (illustration)

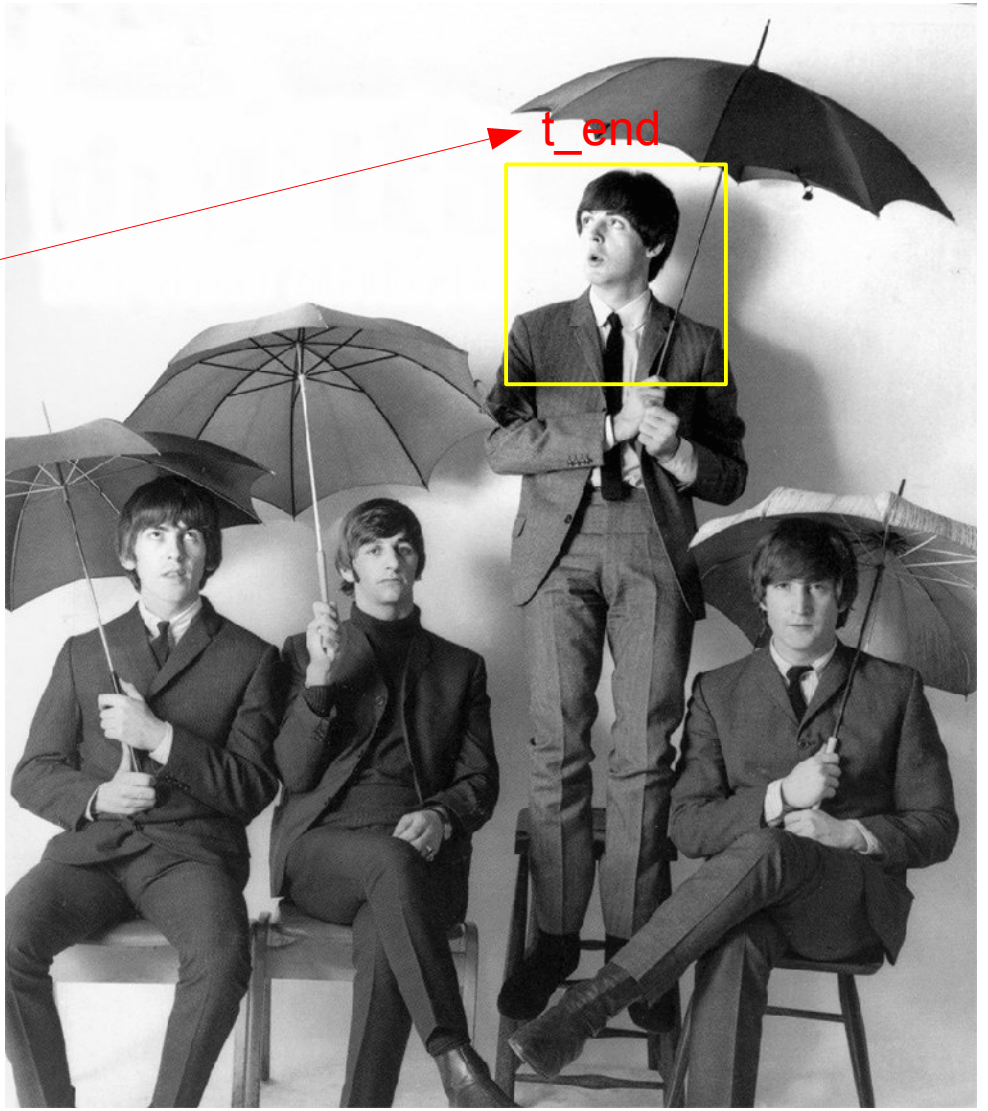
2. Detect and track Beatles





Our approach (illustration)

3. Find Paul standing up



Human detection & tracking

- HOG detector [Dalal05] trained for upper bodies
  - Training samples from Hollywood movies
- Tracking-by-detection [Everingham09]
  - KLT tracker yields feature trajectories
  - Detections are clustered together (agglomerative clustering) based on connectivity score
  - Smoothing + interpolation for continuous tracks

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

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

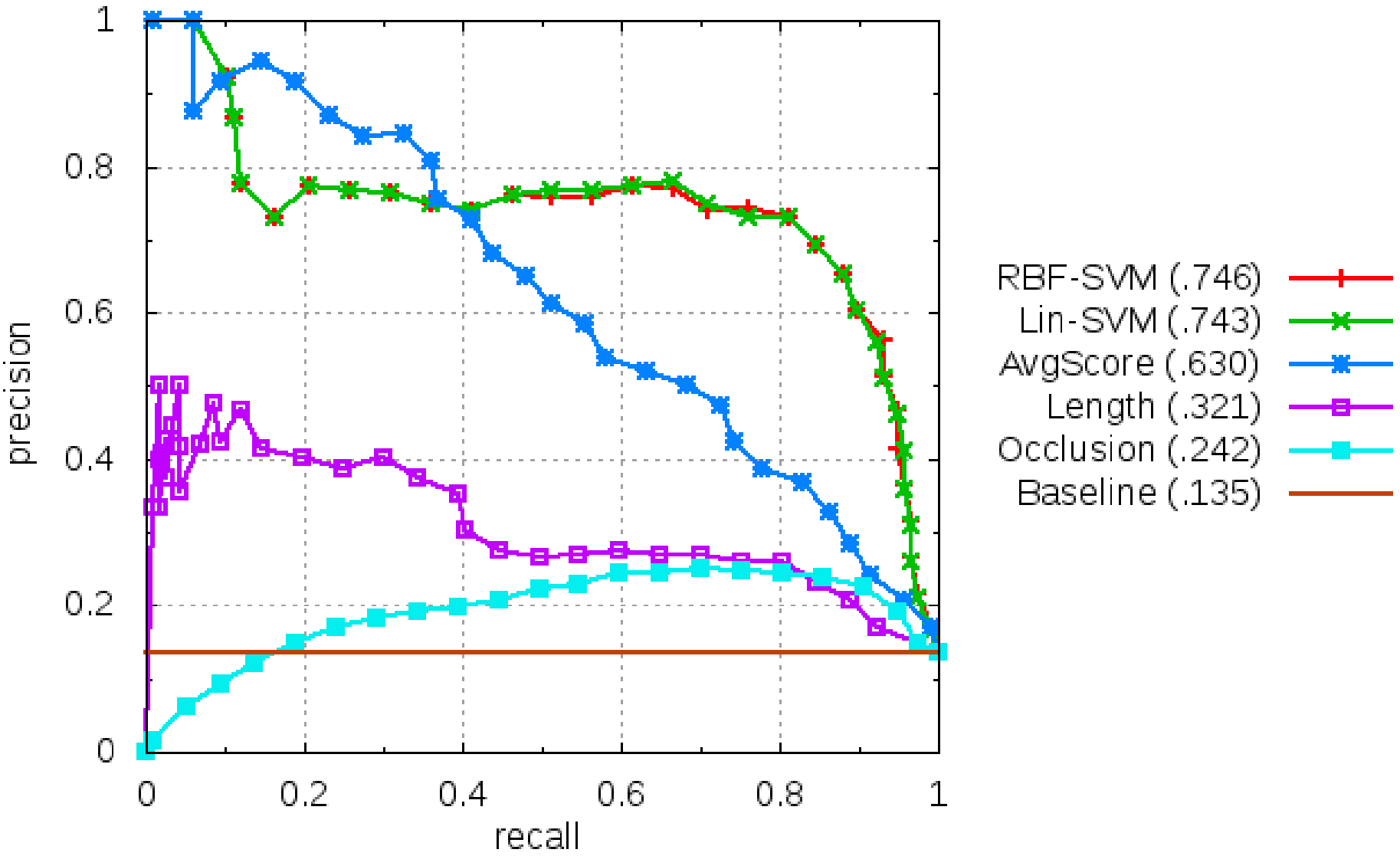
Human detection & tracking

- The procedure works well despite articulations, viewpoint and lighting changes, occlusions

Tracks post-processing

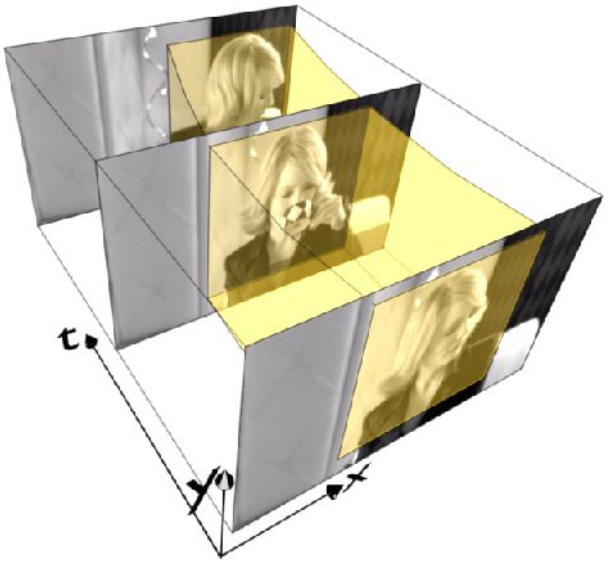
- **Improve precision at high recall with final classification stage of tracks**
- SVM classifier is learned on 12 different measures characterizing a track
- For each measure we compute (if applicable min, max, average)
  - Track length (false tracks are often short)
  - Upper body SVM detection score
  - Scale and position variability (those often reveal artificial detections)
  - Occlusion by other tracks (patterns in the background often generate a number of overlapping detections)

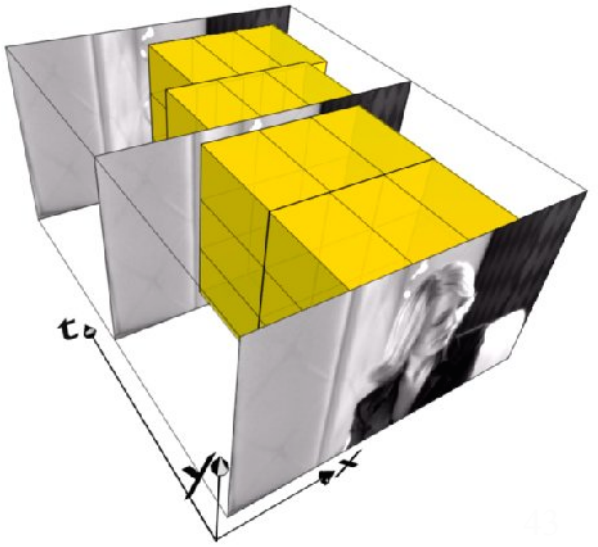
Tracks post-processing



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





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
- 2-stage classification [Harzallah09]
  - Linear SVM as first classifier, generate hypotheses via non-maxima suppression
  - Re-evaluation of final hypotheses with non-linear SVM (RBF)

Dataset: Coffee and Cigarettes

- We use the original split by stories [Laptev07]
  - Annotation via bounding box at key frame + start/end position

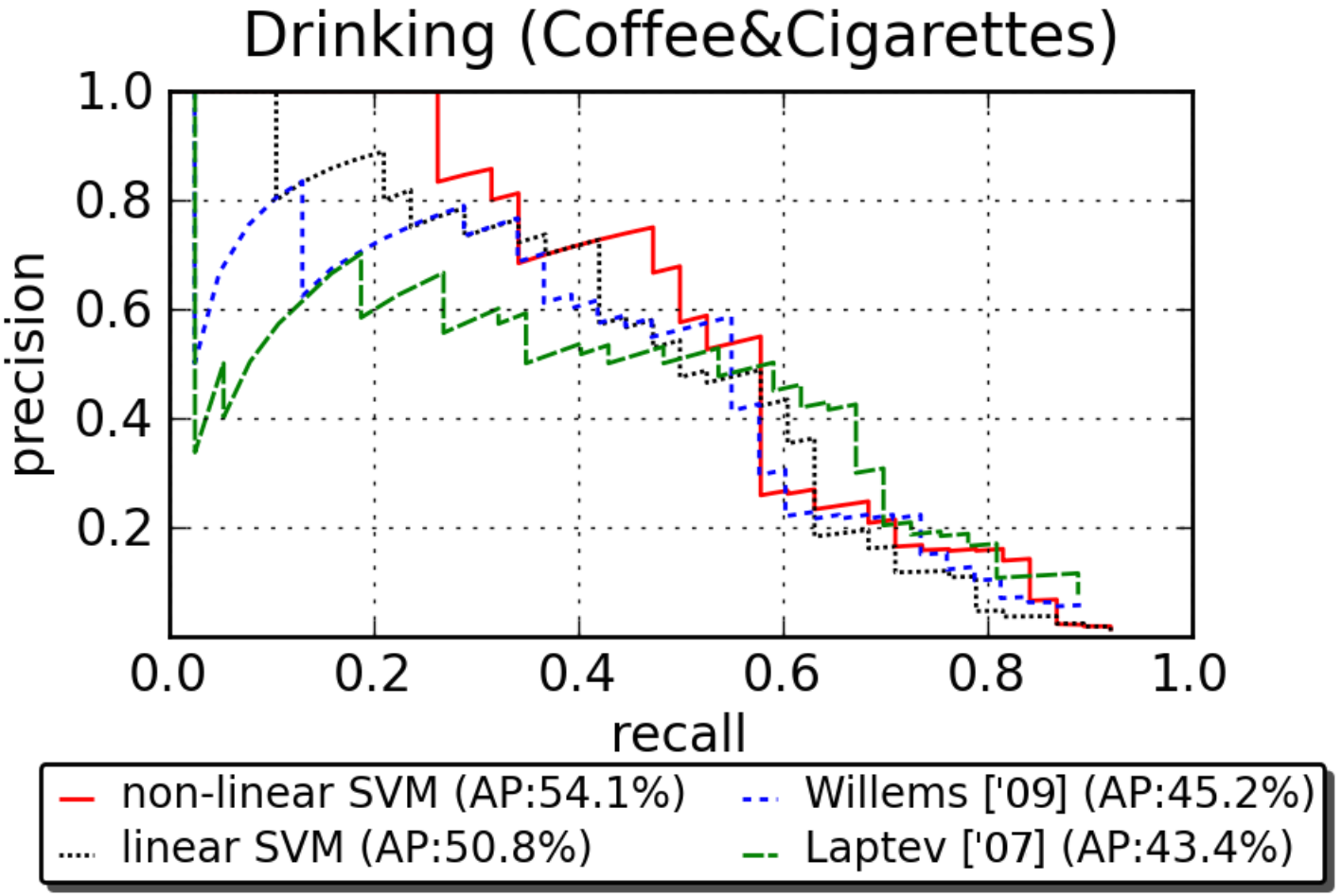
    Ttraining: 6 stories, 40min, 106 drinking, 90 smoking actions (+"Sea of Love" and "Lab" videos)
  - Test-drinking: 2 stories, 24min, 38 drinking actions
  - Test-smoking: 3 stories, 21min 46 smoking actions (originally validation set)
- Average Precision is used for evaluation

training

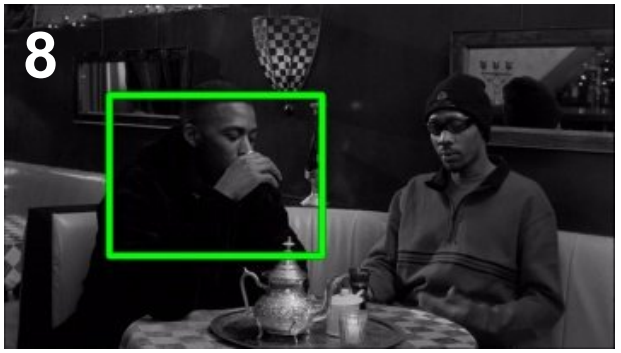
test-smoking

test-drinking

Results for drinking

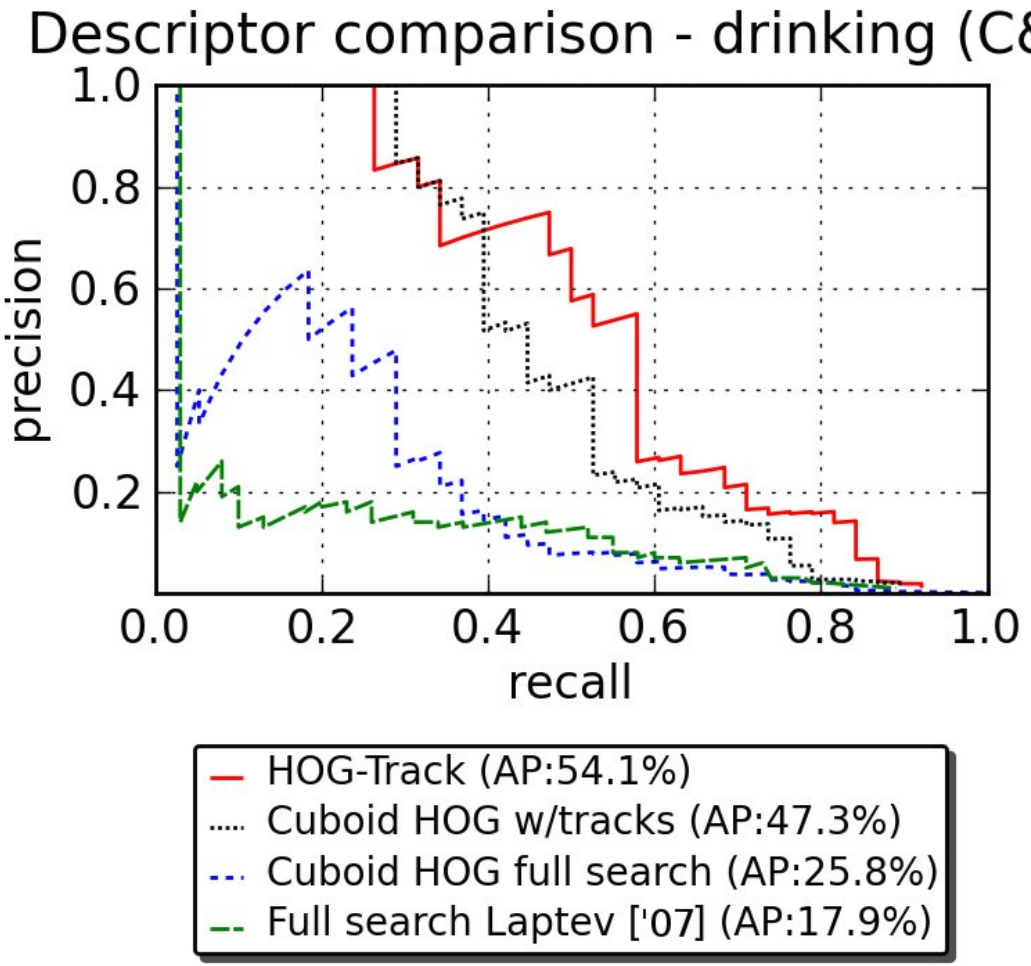


Top 9 results for drinking



Results for drinking

Do tracks really help?



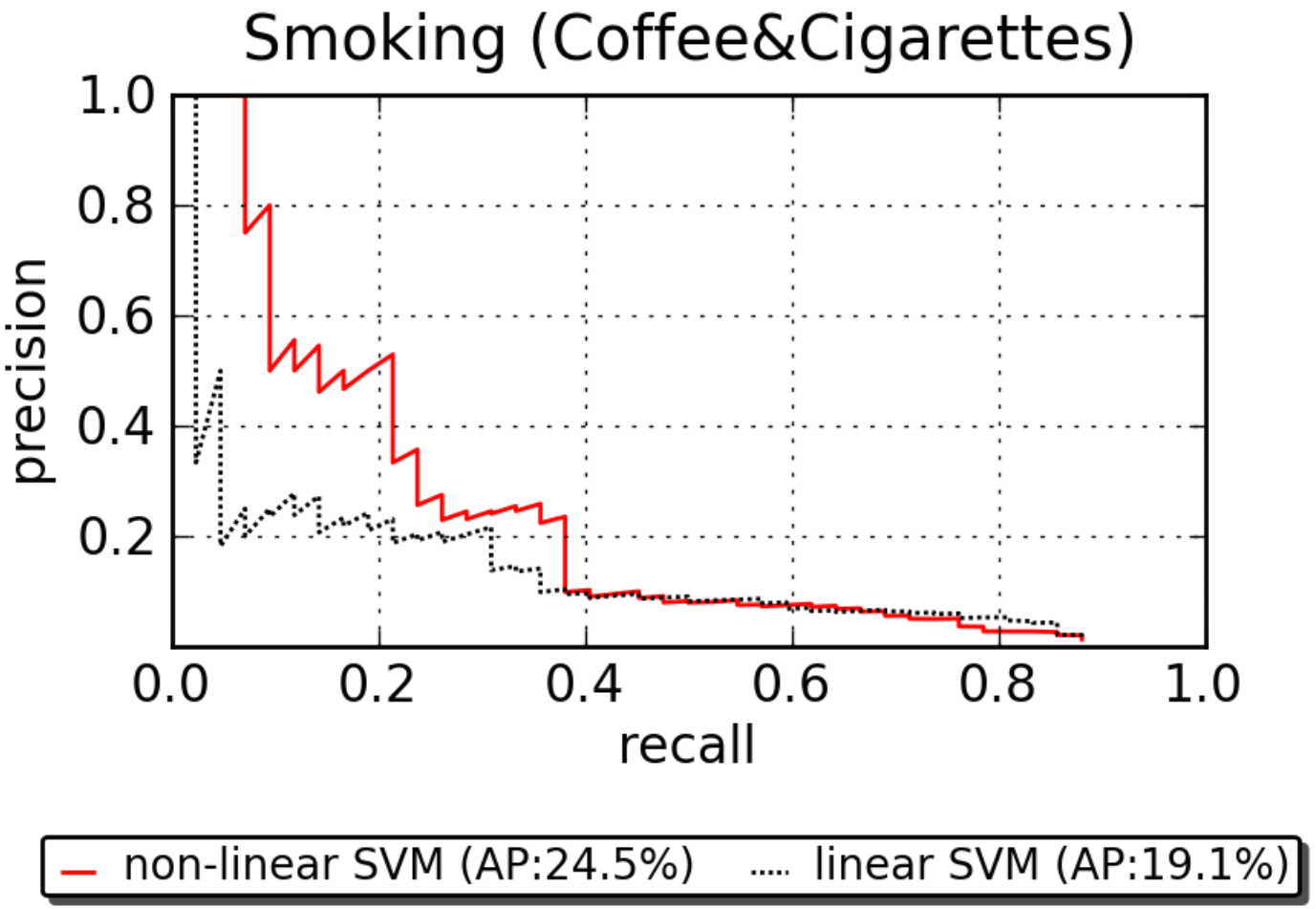
Baselines:

- Laptev's baseline, exhaustive search
- Cuboid classifier, exhaustive search in video
- Cuboid classifier, centered on tracks

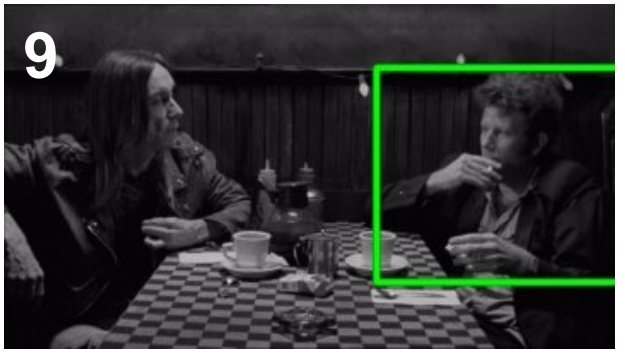
Why tracks help

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

Results for smoking



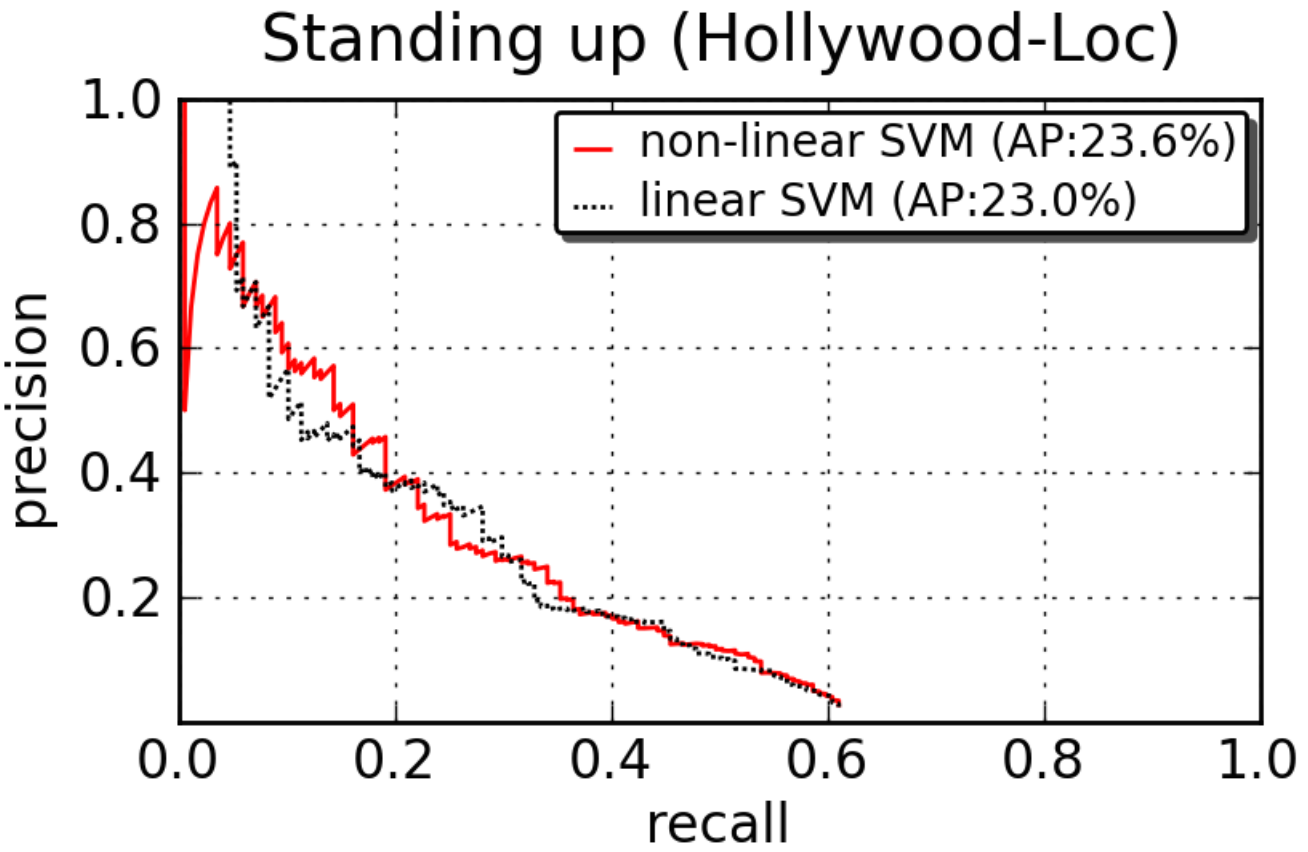
Top 9 results for smoking



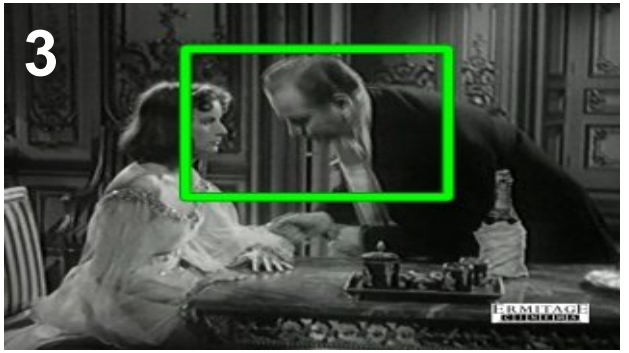
Dataset: Hollywood localization

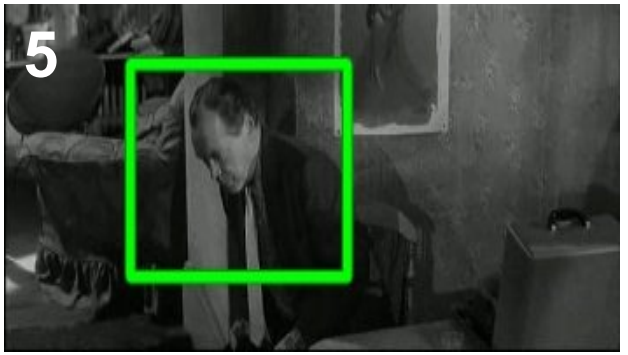
- Dataset based on Hollywood2 data and split
  - ~2h of video data in total (~1h training, ~1h test)
- We annotate the spatial and temporal extent of "**phoning**" and "**standing up**" actions
  - Annotation via bounding box at key frame + start/end position
  - 153 "phoning" actions (73 training, 80 test)
  - 274 "standing up" actions (129 training, 145 test)
- Average Precision is used for evaluation

Results for standing up

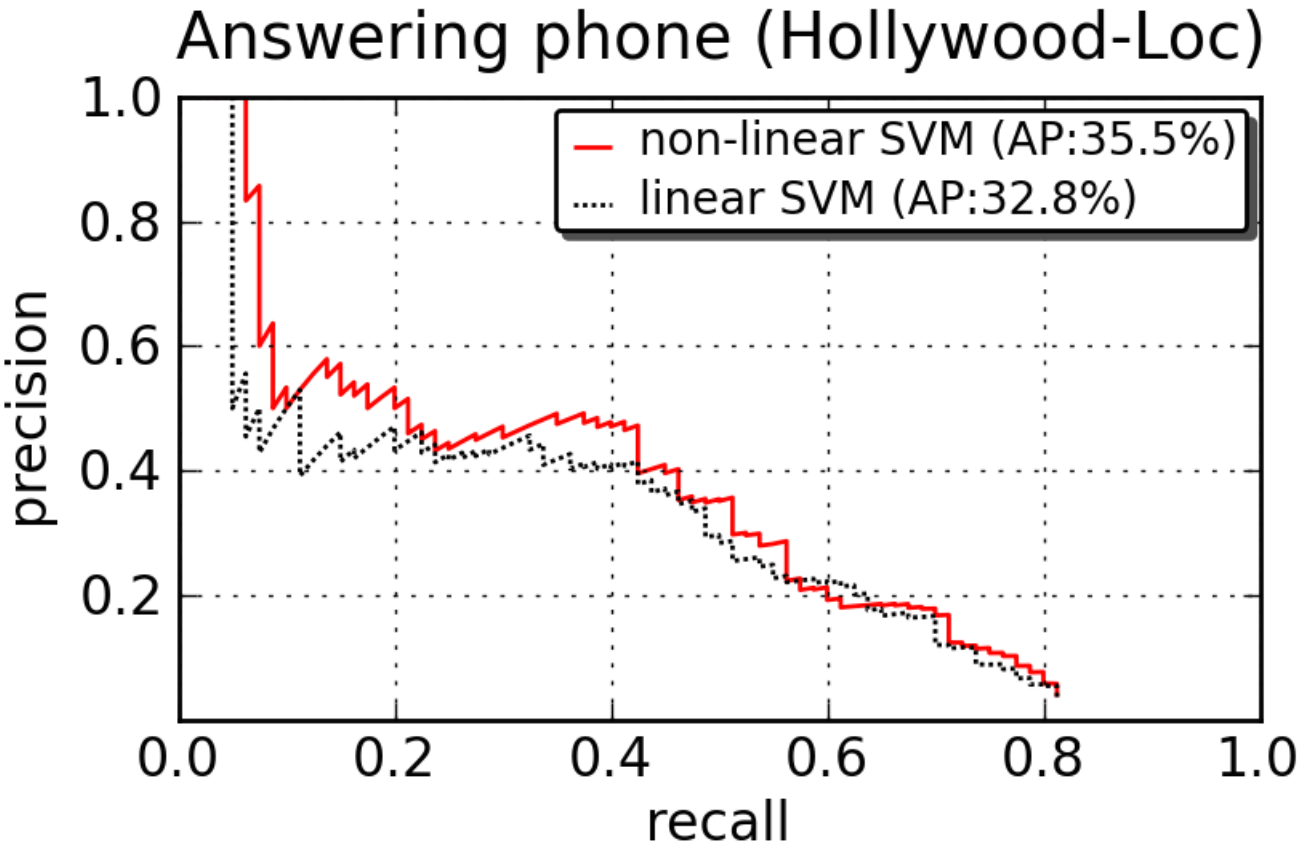


Top 9 results for standing up





Results for phoning



Top 9 results for phoning

Results for phoning

Action detections

Human detections

Human tracks

Outline

- Bag-of-features
- Local spatio-temporal HOG3D descriptor
- Evaluation of local feature detectors & descriptors
- Human focused action localization in space-time
- **Summary & conclusion**

Summary

- Several contributions to action recognition (classification and localization) in realistic video settings have been presented
  - Local spatio-temporal HOG3D descriptor
  - Evaluation of local feature detectors/descriptors
  - Action localization based on generic human tracks
  - Local feature trajectory descriptor
    - Improved performance with combination of trajectory and motion / appearance information
    - Novel descriptor based on motion boundary histograms
  - Combination of BoF with human tracks
    - Improved performance for foreground features (class dependent)
    - Spatial layout information can be incorporated

Future work

- Adapted representation for each action
  - Adapting descriptor parameters has been investigated in this thesis
  - Combination of different type of information (MKL, early fusion)
  - Explicit learning of context information
- Action modeling using feature trajectories
  - Since trajectories follow local motion, they offer interesting possibilities for more principled representations [Matikainen09]
  - Trajectories of similar shape can be grouped together to model relations between/within local regions

Future work

- Motion boundary histograms
  - Have shown promising results
  - Invariance to camera ego-motion is important
  - Application to other problems (e.g., action localization)
- Action localization
  - Multiple body parts model can improve robustness of tracking [Mikolajczyk04, Felzenszwalb10]
  - Possibility to model actions at different levels (head, upper body, legs, full body)
  - Incorporate multiple view information in action representation

Thank you for your attention

I will be glad to answer your questions

Local feature trajectory descriptor

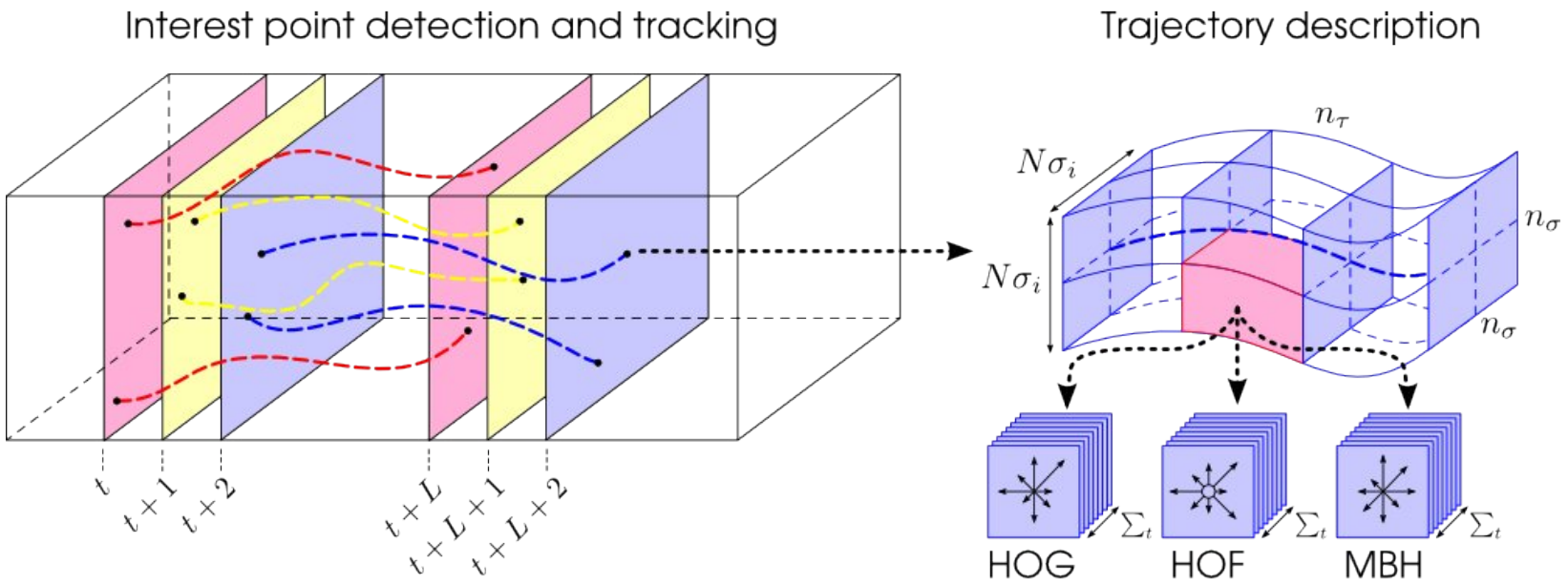
Motivation & goal, approach, current results

[H. Wang, A. Kläser, C.-L. Liu, and C. Schmid. *Action recognition with feature trajectories*. Unpublished]

Motivation & goal

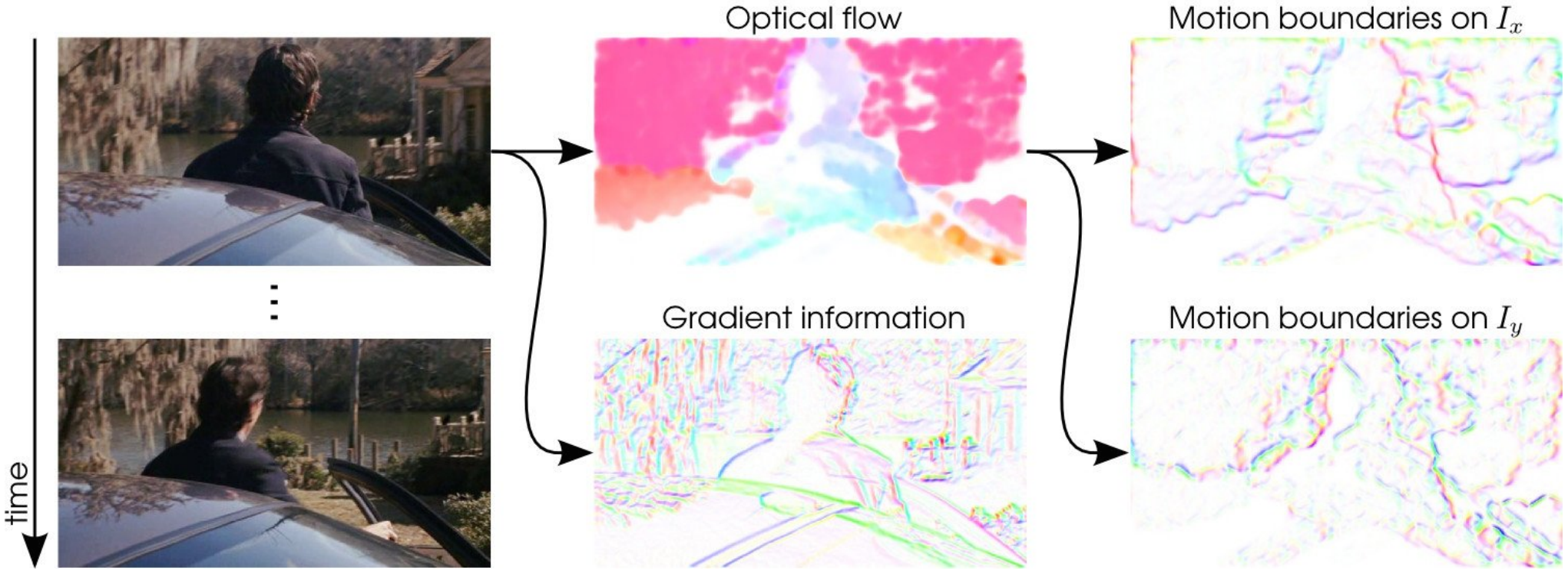
- Common feature detectors: detection of 3D positions based on saliency criterion
- In video, however, image structures move over time
  - Tracking is naturally used to capture motion
  - Objects are difficult to track in realistic videos
- **Main idea:** Local feature trajectories for action recognition [Messing09, Matikainen09, Sun09]
  - Combination of local features + tracking approaches
- **What is new ?**
  - Combined trajectory descriptor: trajectory shape, HOG, HOF, MBH
  - Novel motion boundary descriptor (MBH) for action recognition
  - Extensive evaluation, also on realistic data
  - Learn parameters from training data

Descriptor computation



- Spatial interest points are detected and tracked
  - Trajectory length limited (15 frames) to cope with drifting
- Descriptors computed for a grid in local neighborhood of trajectory
  - Combination of appearance and motion information
  - Histograms of oriented gradients (HOG) and optical flow (HOF), motion boundary histograms (MBH)
  - MBH computes HOG representations on x/y optical flow components

Descriptors: HOF vs. HOG vs. MBH



- MBH is able to capture complementary information ☺
  - Static clutter vanishes
- Motion boundaries are invariant to camera ego-motion (very important for realistic movies) ☺

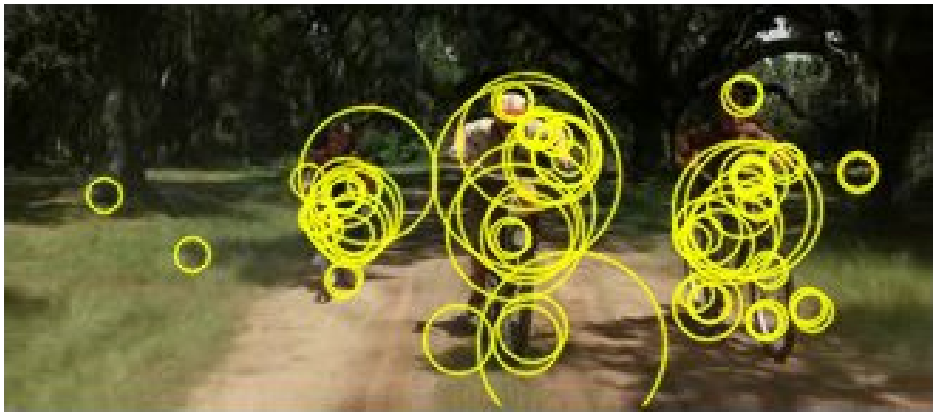
Results

	Dataset		
	KTH	**YouTube**	**Hollywood2**
Ours	**94.2%**	**79.8%**	**52.5%**
Harris3D+HOG/HOF	92.0%	68.7%	47.3%
State-of-the-art	**94.5%**	71.2%	50.9%
	[Gilbert09]	*[Liu09]*	*[Gilbert10]*

- Very promising results, our method outperforms more advanced ones ☺
- MBH helps to improve results considerably ☺

Illustration of feature detectors

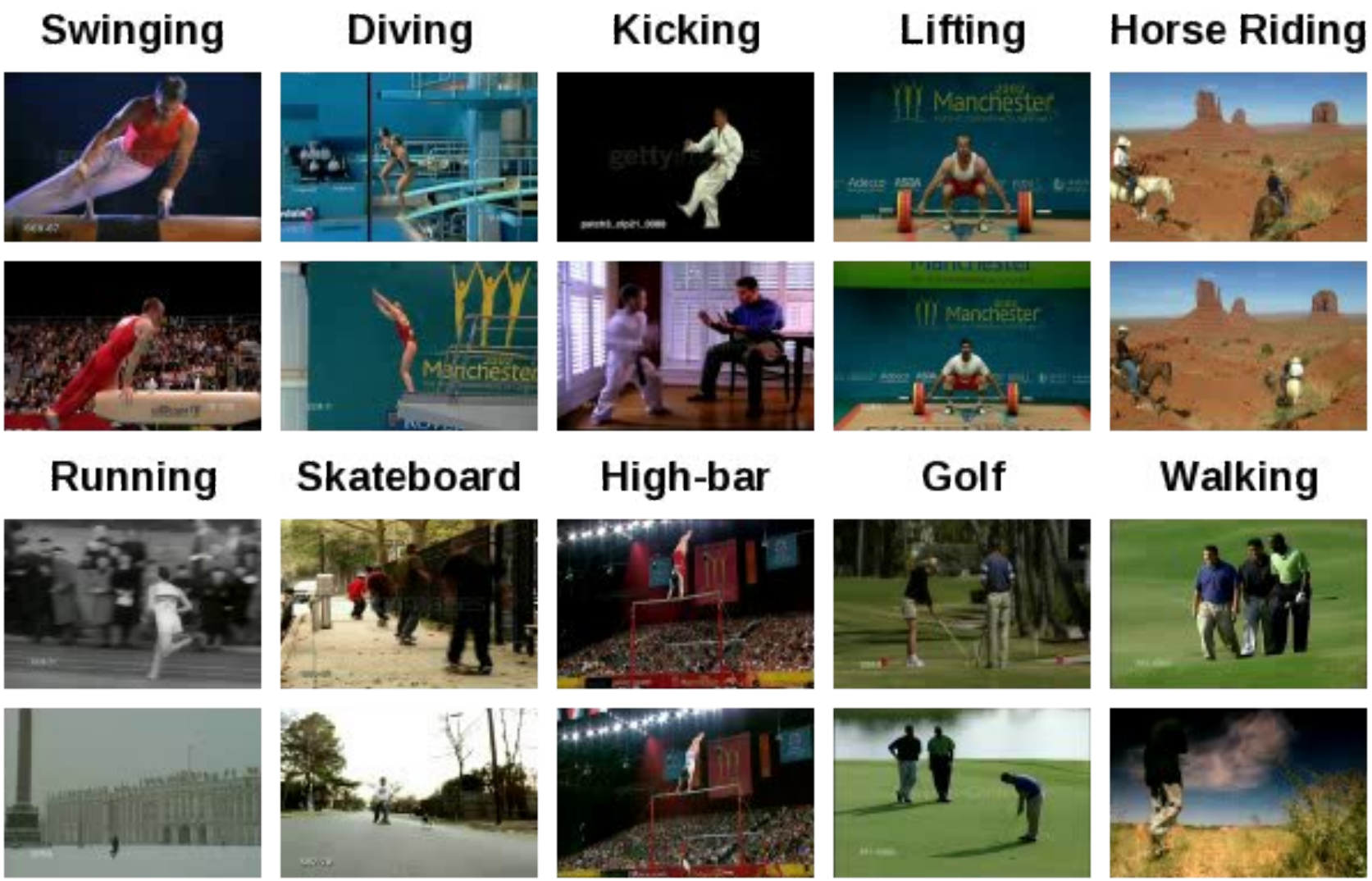
Harris



Cuboid

Hessian

UCF sports – samples



Human tracking in realistic videos