Action Recognition

Slides borrowed from Derek Hoiem

Last classes

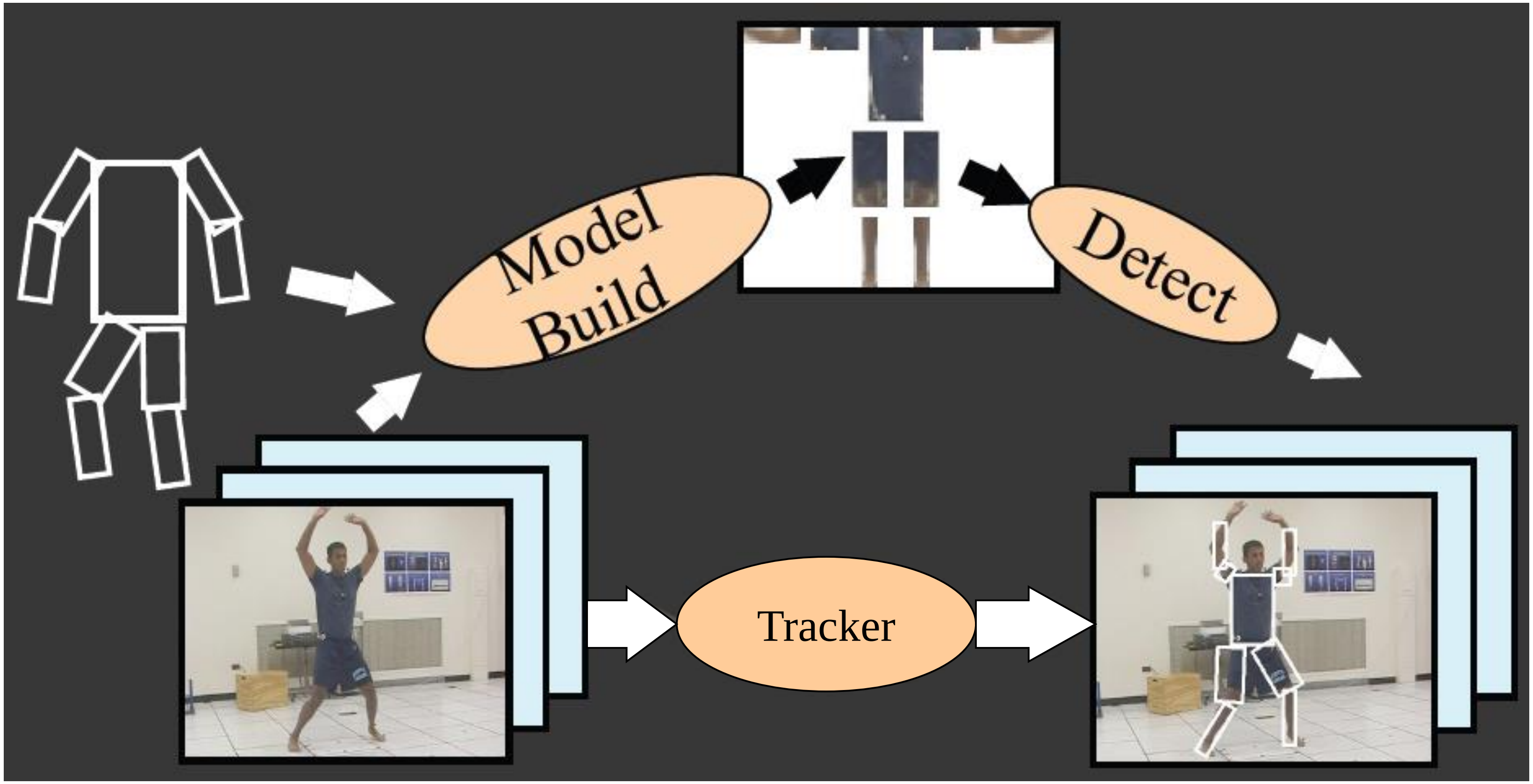
- Parts-based/articulated object models
- Tracking objects

Tracking people

- Person model = appearance + structure (+ dynamics)
- Structure and dynamics are general, appearance is person-specific
- Trying to acquire an appearance model “on the fly” can lead to drift
- Instead, can use the whole sequence to initialize the appearance model and then keep it fixed while tracking
- Given strong structure and appearance models, tracking can essentially be done by repeated detection (with some smoothing)

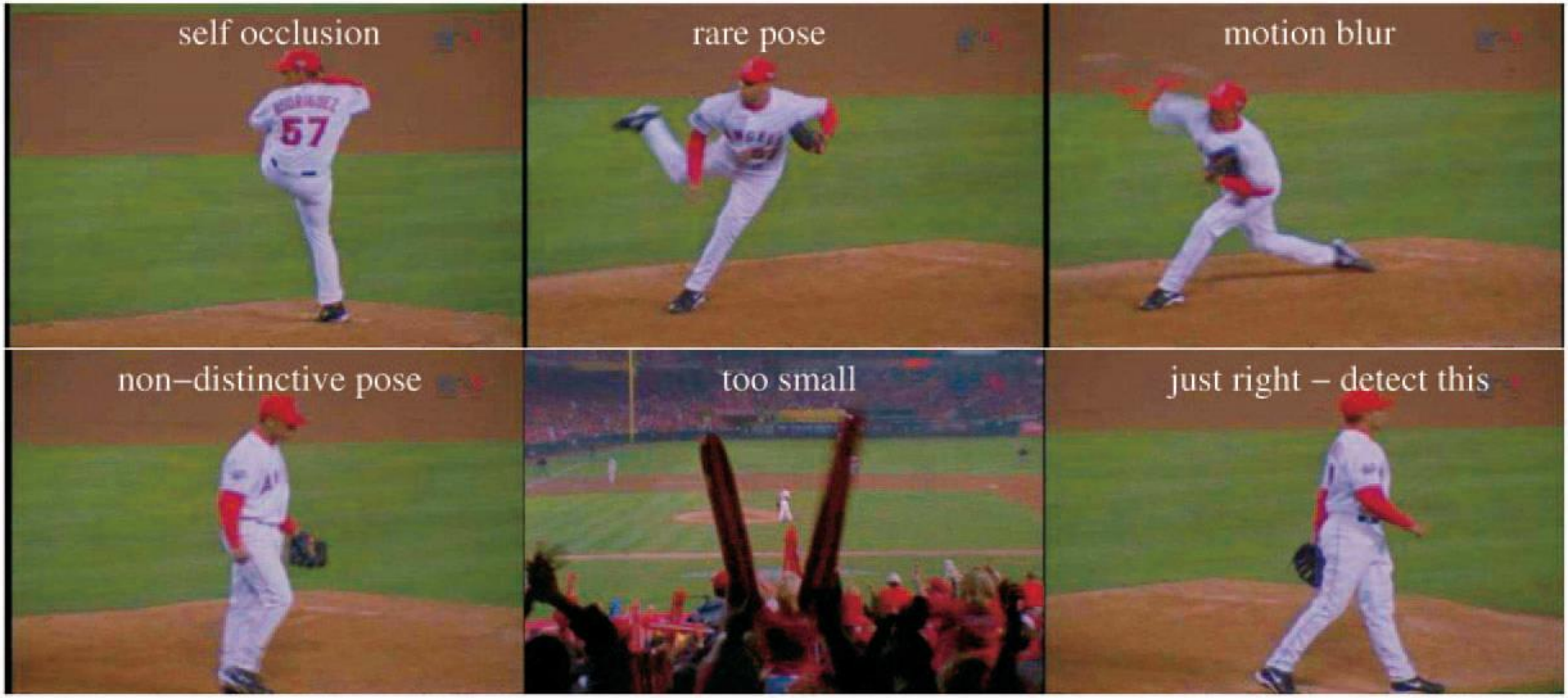
D. Ramanan, D. Forsyth, and A. Zisserman. Tracking People by Learning their Appearance. PAMI 2007.

Tracking people by learning their appearance



D. Ramanan, D. Forsyth, and A. Zisserman. Tracking People by Learning their Appearance. PAMI 2007.

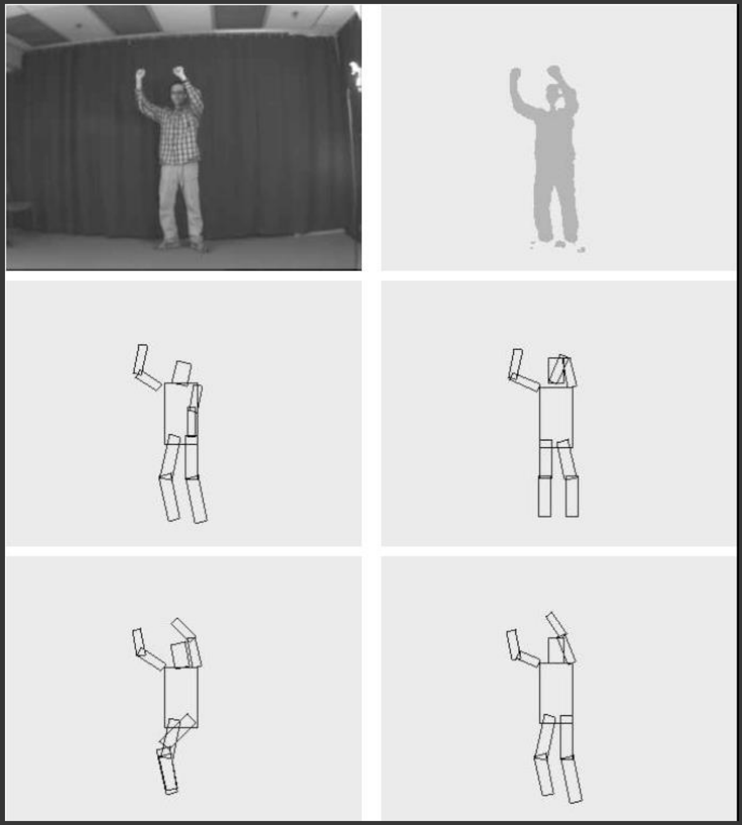
Top-down method to build model: Exploit "easy" poses

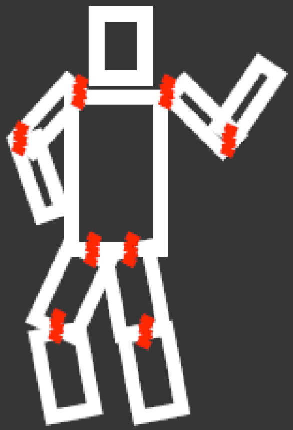


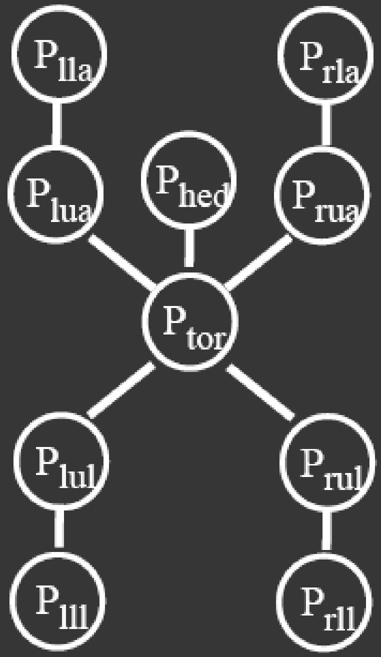
D. Ramanan, D. Forsyth, and A. Zisserman. Tracking People by Learning their Appearance. PAMI 2007.

Pictorial structure model

Fischler and Elschlager(73), Felzenszwalb and Huttenlocher(00)







$$\Pr(P_{tor},P_{arm},\ldots|\mathrm{Im}) \propto \prod_{i,j}\Pr(P_i \mid P_j)\prod_{i}\Pr(\mathrm{Im}(P_i))$$

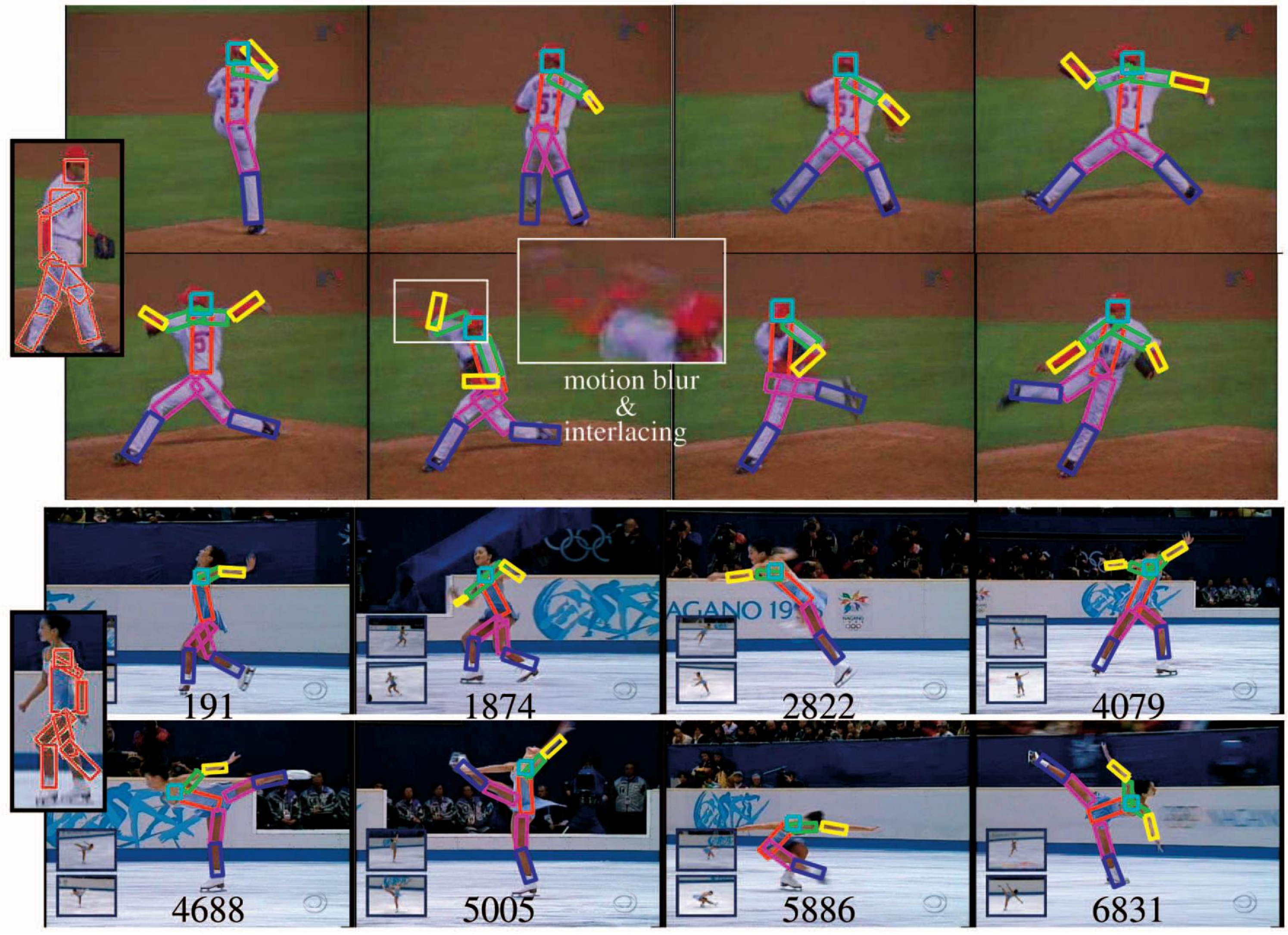
part geometry

part appearance

Temporal model

- Parts cannot move too far

Example results



http://www.ics.uci.edu/~dramanan/papers/pose/index.html

Video

http://www.ics.uci.edu/~dramanan/papers/pose/index.html

How can we identify actions?

Motion

Pose

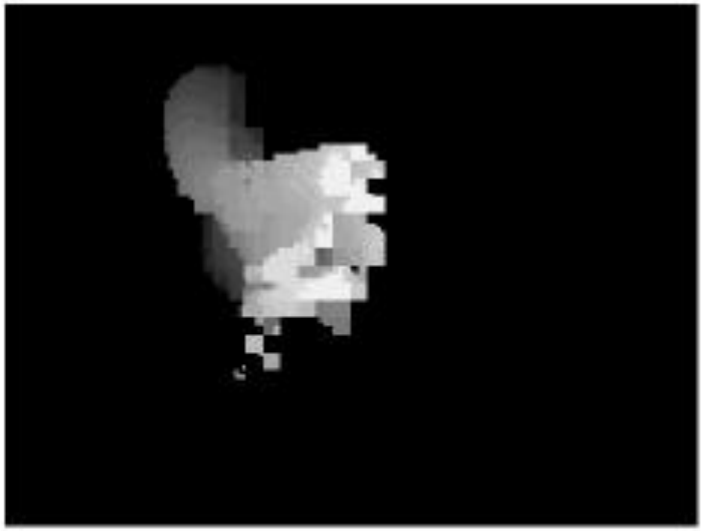
Held Objects

Nearby Objects

Representing Motion

Optical Flow with Motion History

sit-down

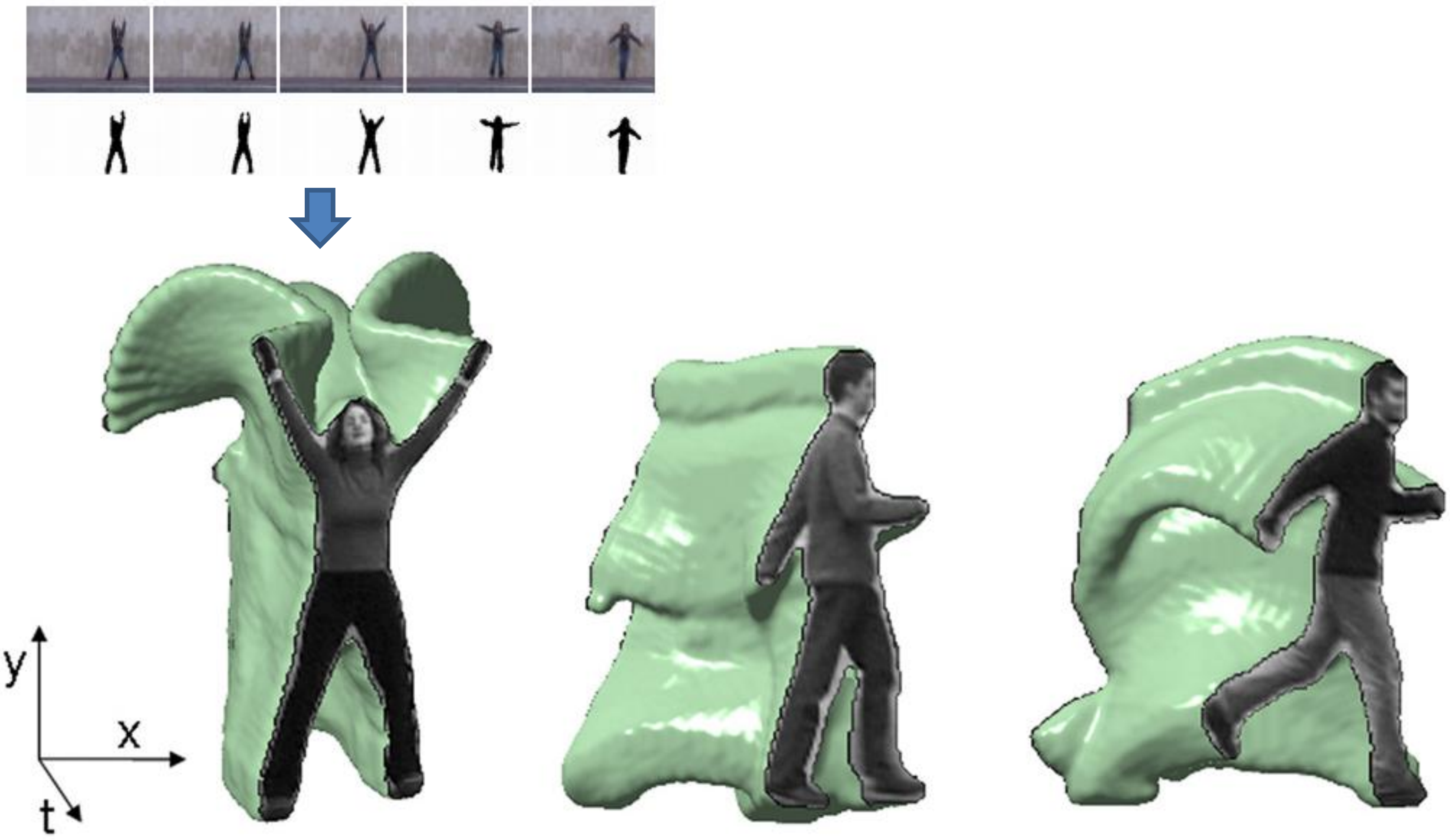


sit-down MHI

Bobick Davis 2001

Representing Motion

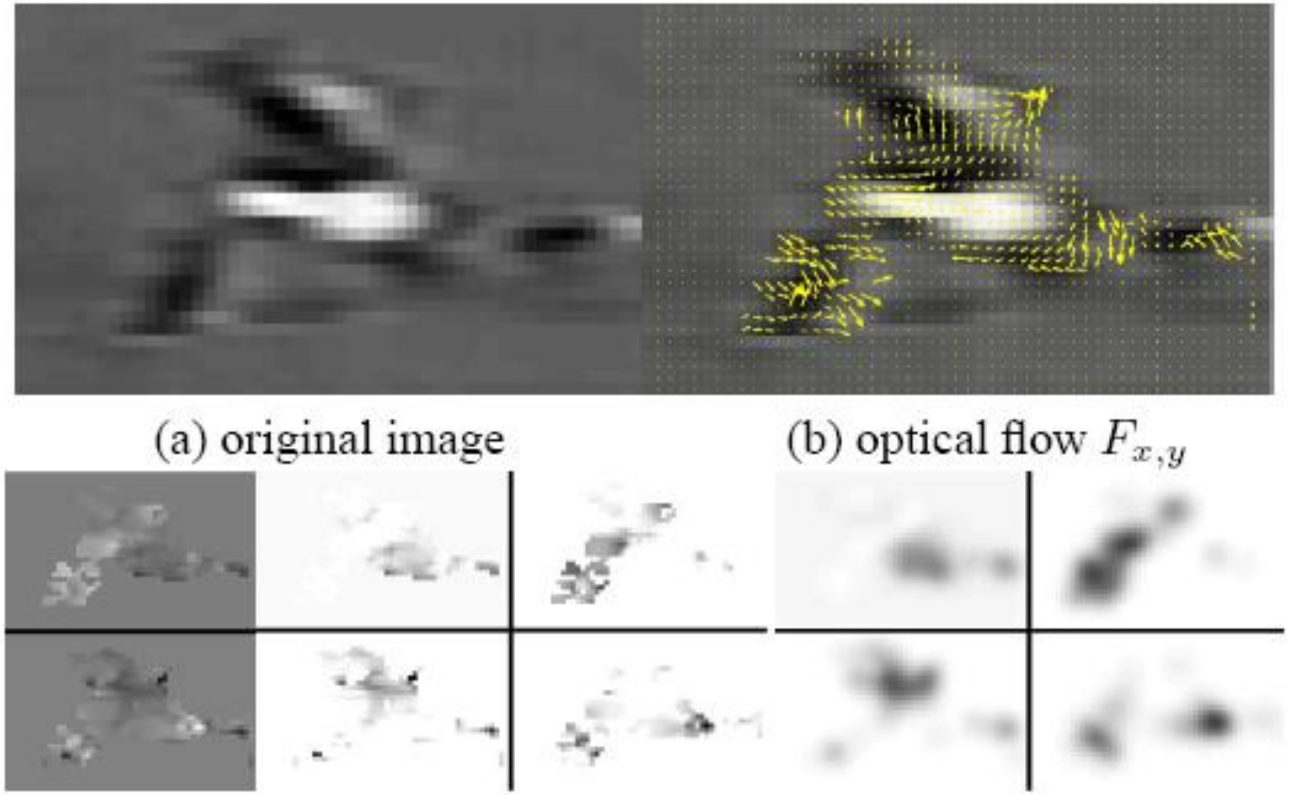
Space-Time Volumes



Blank et al. 2005

Representing Motion

Optical Flow with Split Channels



(a) original image

(b) optical flow $F_{x,y}$

(c) F_x, F_y

optical flow

(d) $F_x^+, F_x^-, F_y^+, F_y^-$

split into pos/neg channels

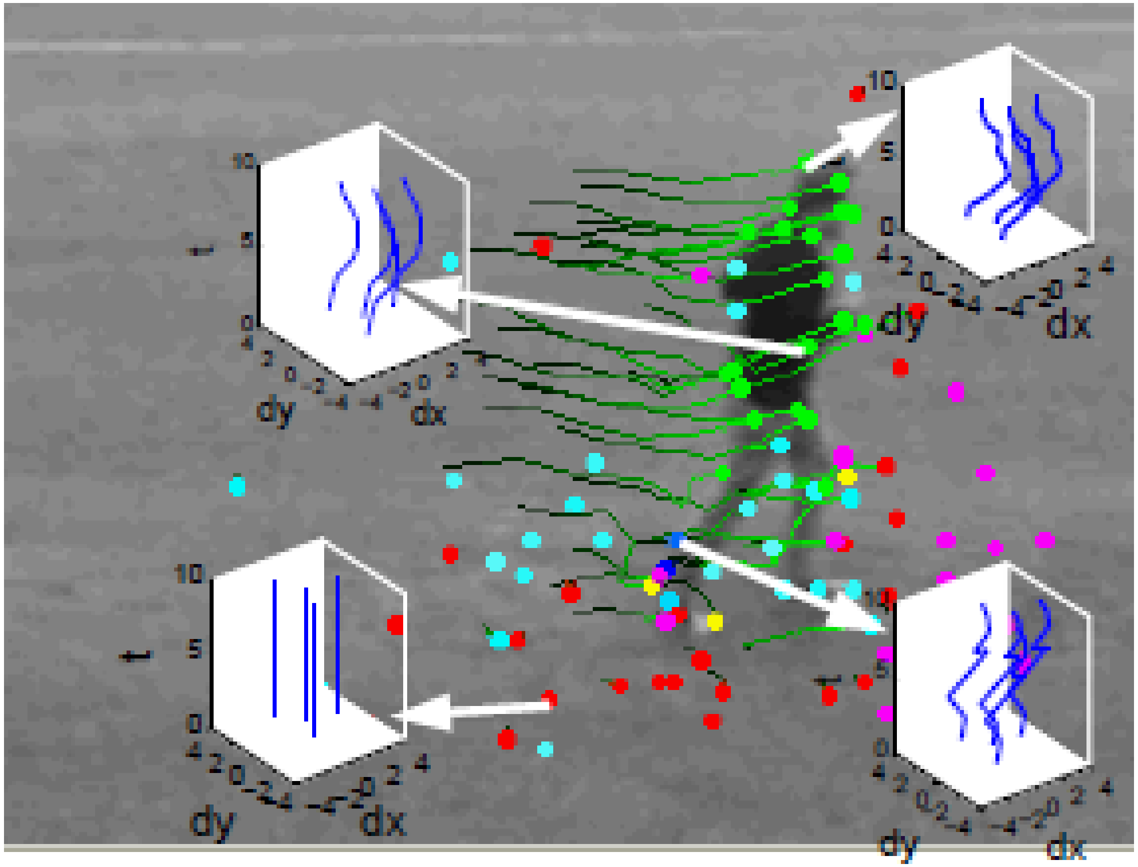
(e) $Fb_x^+, Fb_x^-, Fb_y^+, Fb_y^-$

blurred pos/neg flow

Efros et al. 2003

Representing Motion

Tracked Points



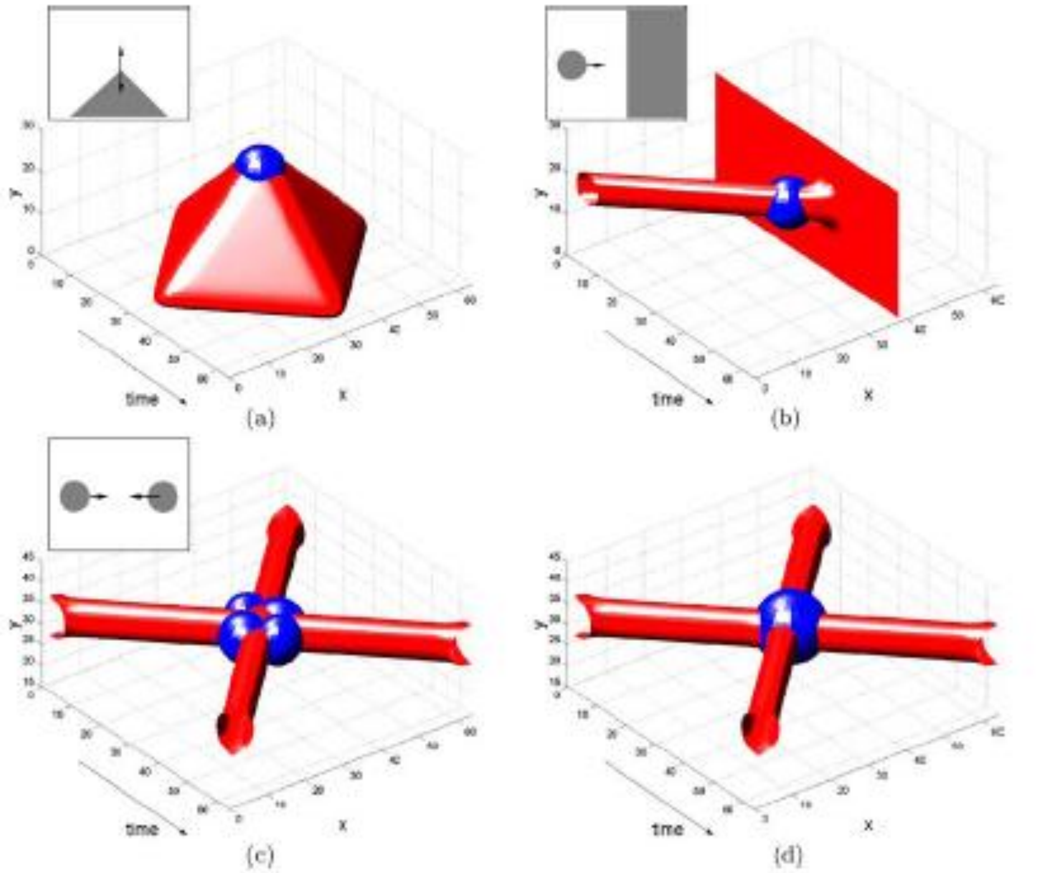
Matikainen et al. 2009

Representing Motion

Space-Time Interest Points

Moving corner

Ball hits wall



Balls collide

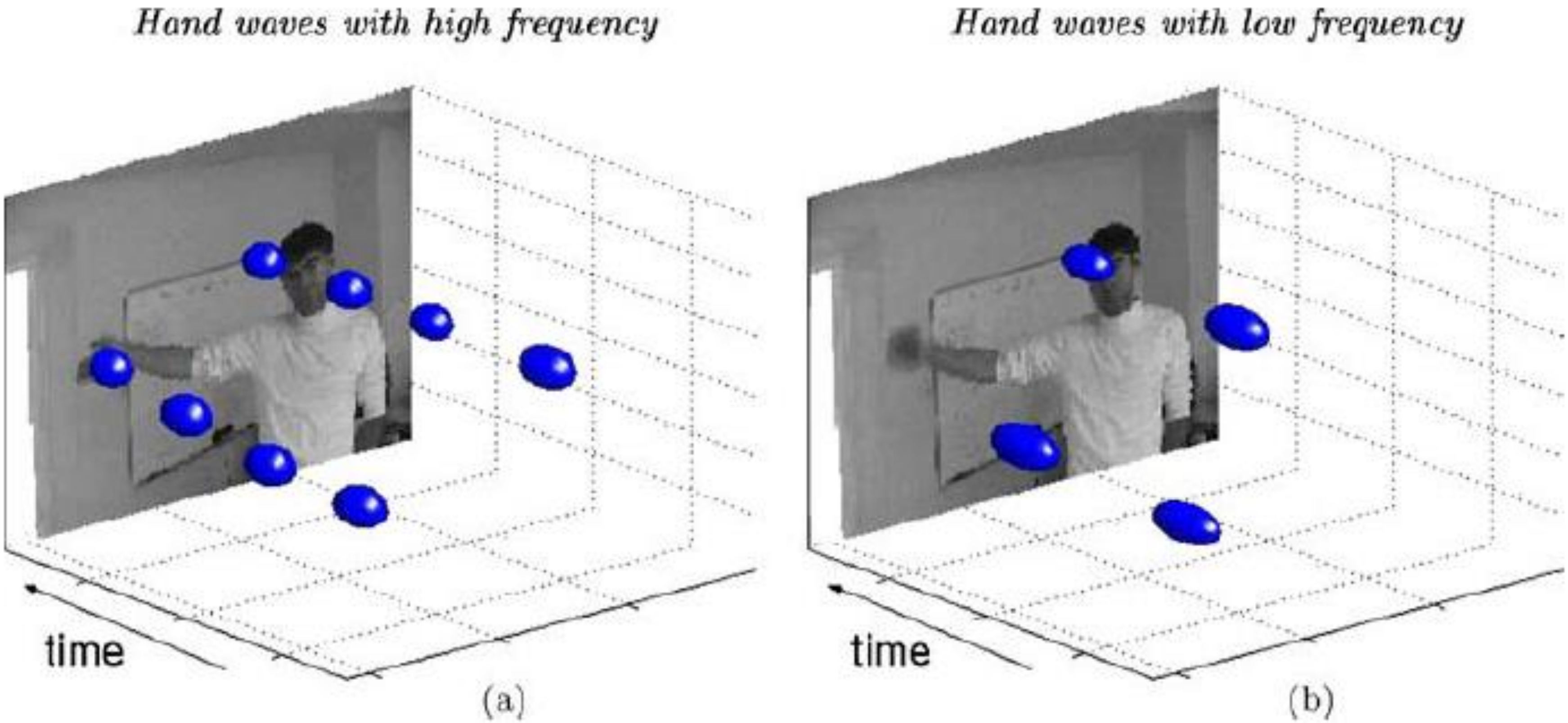
Balls collide (different scale)

Corner detectors in space-time

[Laptev 2005]

Representing Motion

Space-Time Interest Points



Hand waves with high frequency

Hand waves with low frequency

(a)

(b)

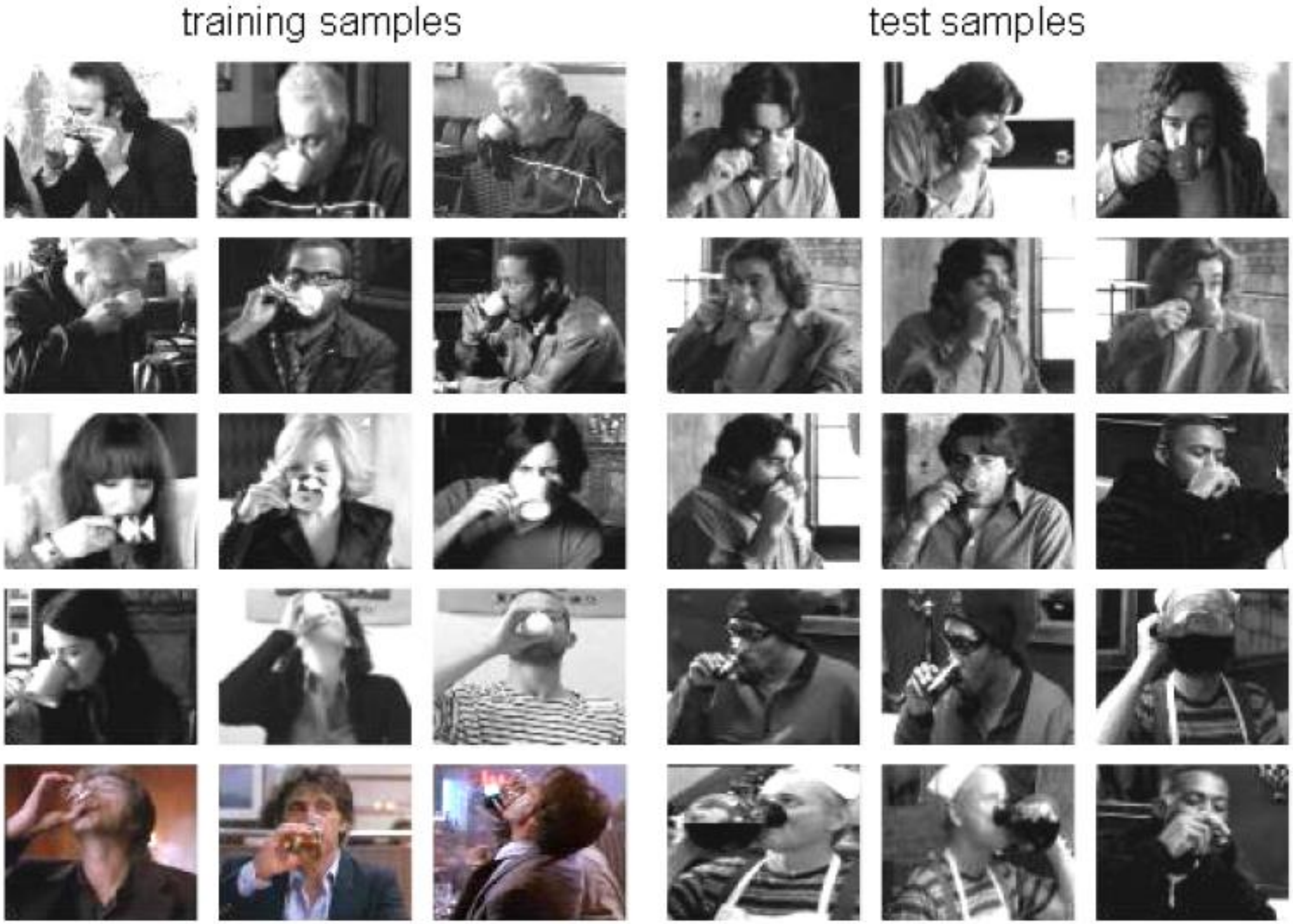
Laptev 2005

Examples of Action Recognition Systems

- Feature-based classification

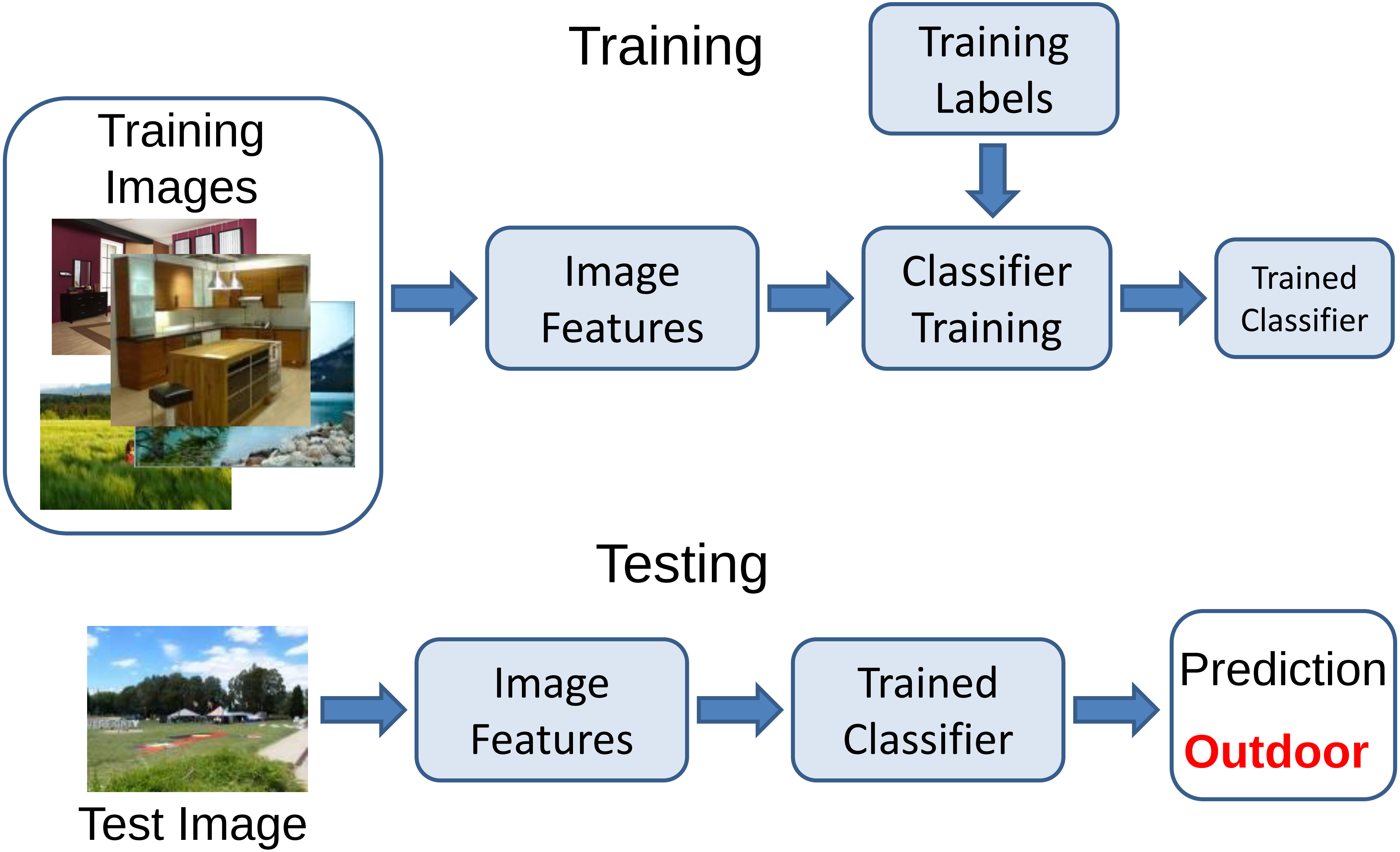
- Recognition using pose and objects

Action recognition as classification



Retrieving actions in movies, Laptev and Perez, 2007

Remember image categorization…



Training

Training Labels

Training Images

Image Features

Classifier Training

Trained Classifier

Testing

Test Image

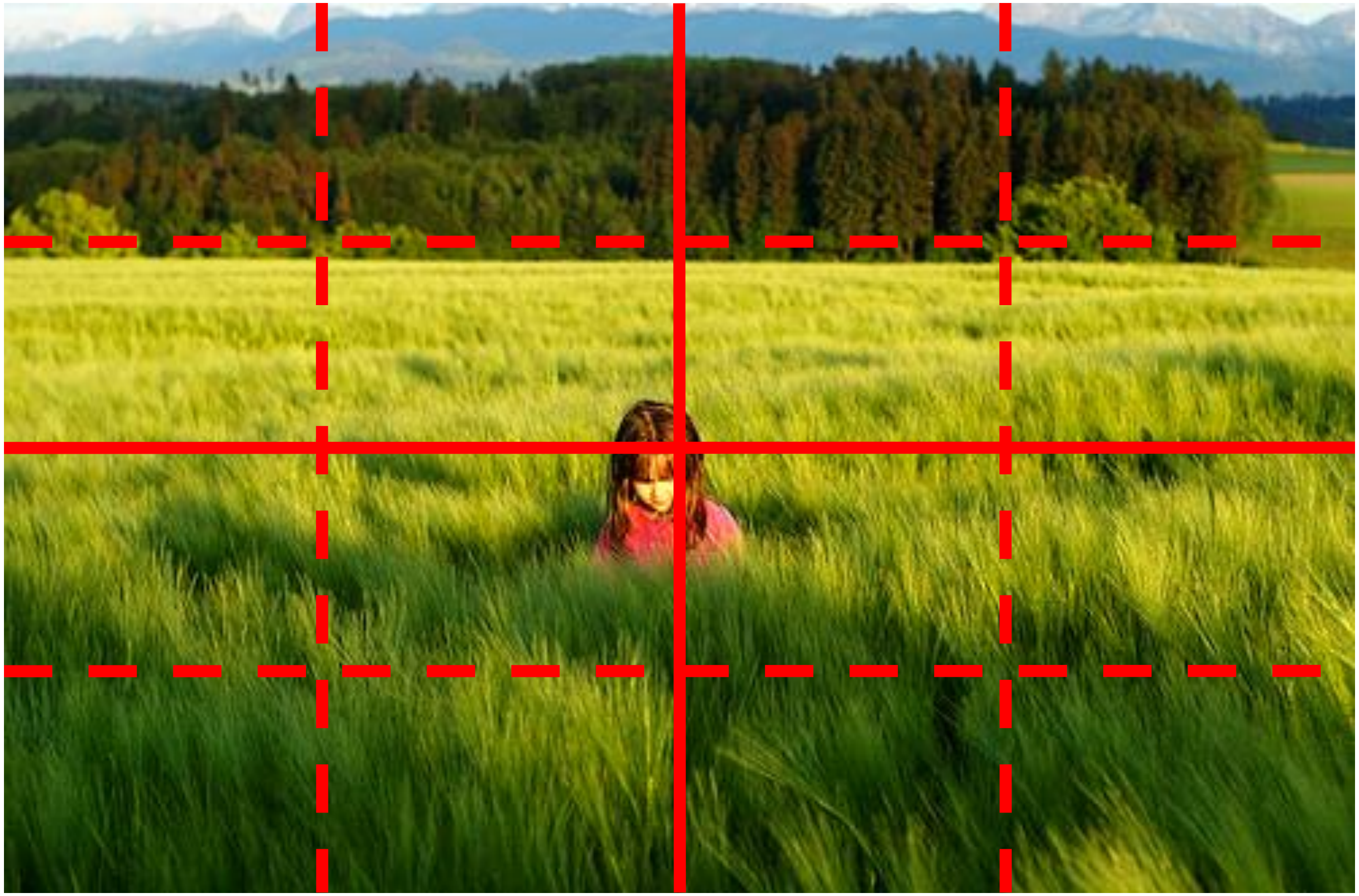
Image Features

Trained Classifier

Prediction

Outdoor

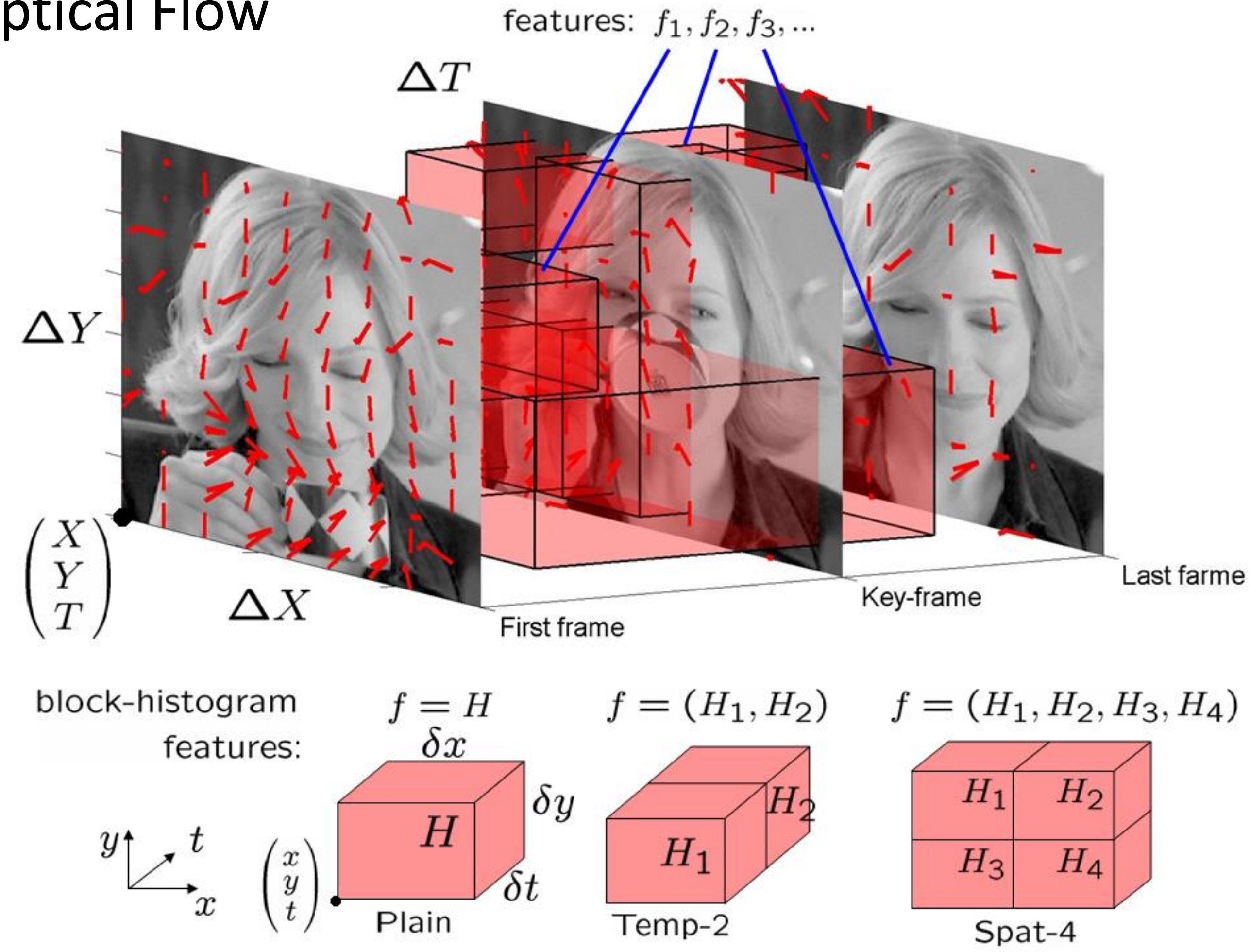
Remember spatial pyramids….



Compute histogram in each spatial bin

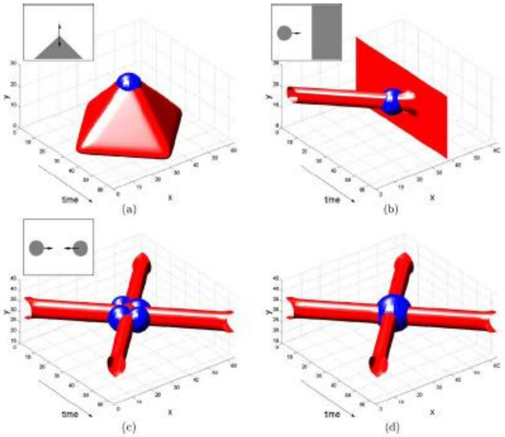
Features for Classifying Actions

1. Spatio-temporal pyramids
   - Image Gradients
   - Optical Flow

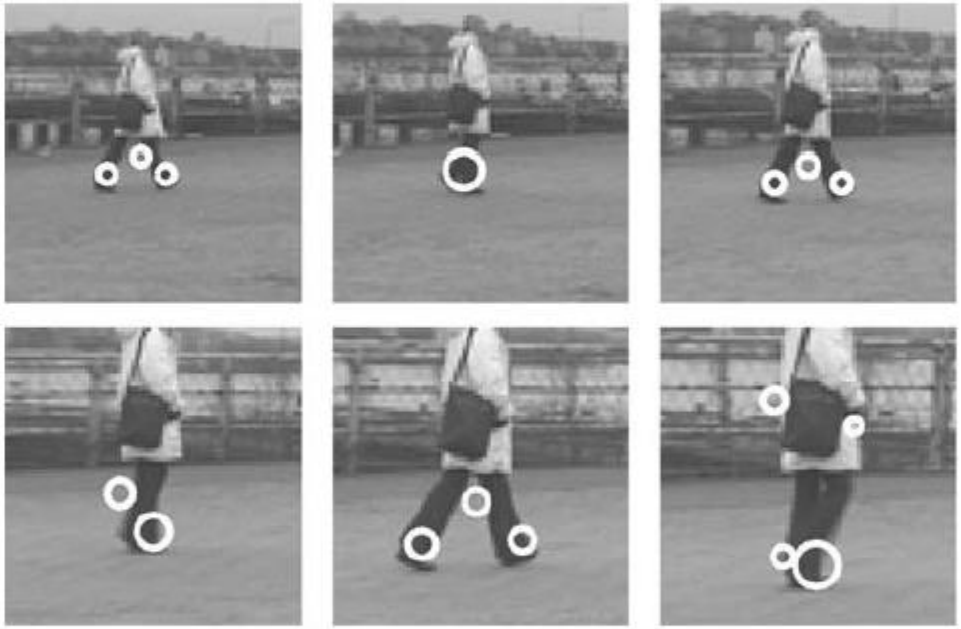


Features for Classifying Actions

2. Spatio-temporal interest points



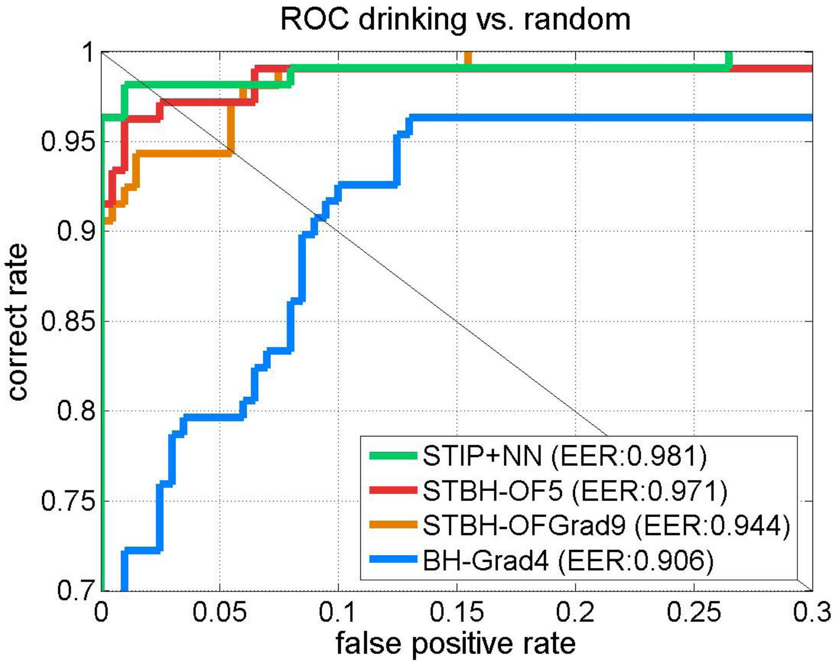
Corner detectors in space-time

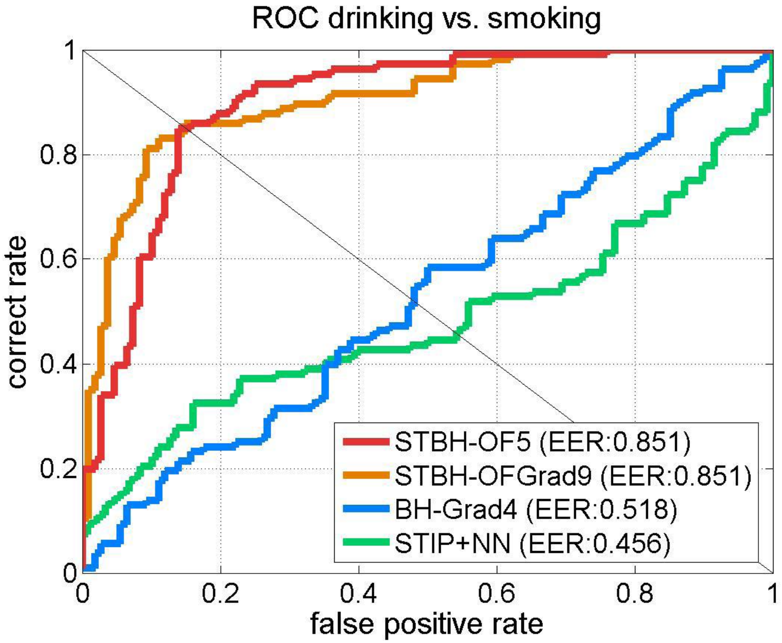


Descriptors based on Gaussian derivative filters over x, y, time

Classification

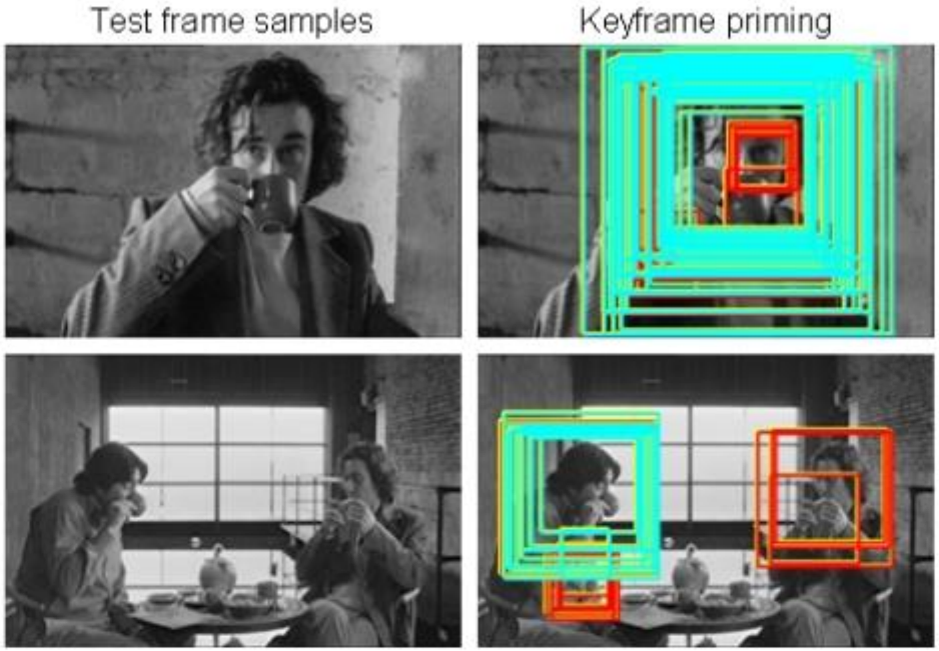
- Boosted stubs for pyramids of optical flow, gradient
- Nearest neighbor for STIP

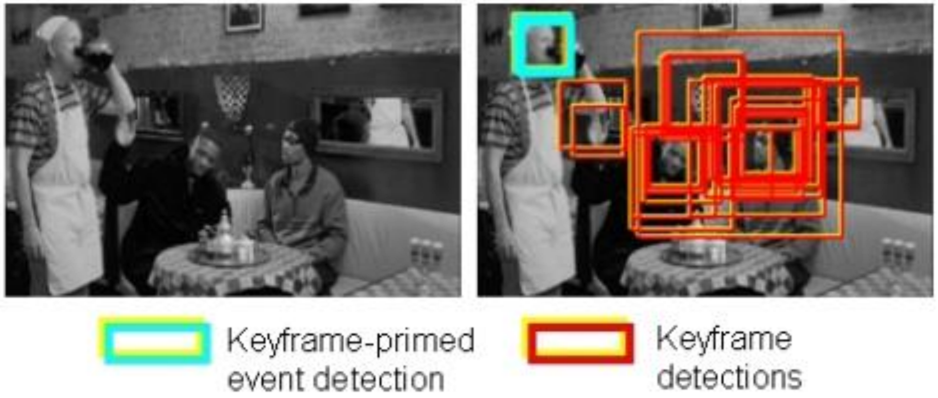




Searching the video for an action

1. Detect keyframes using a trained HOG detector in each frame
2. Classify detected keyframes as positive (e.g., “drinking”) or negative (“other”)

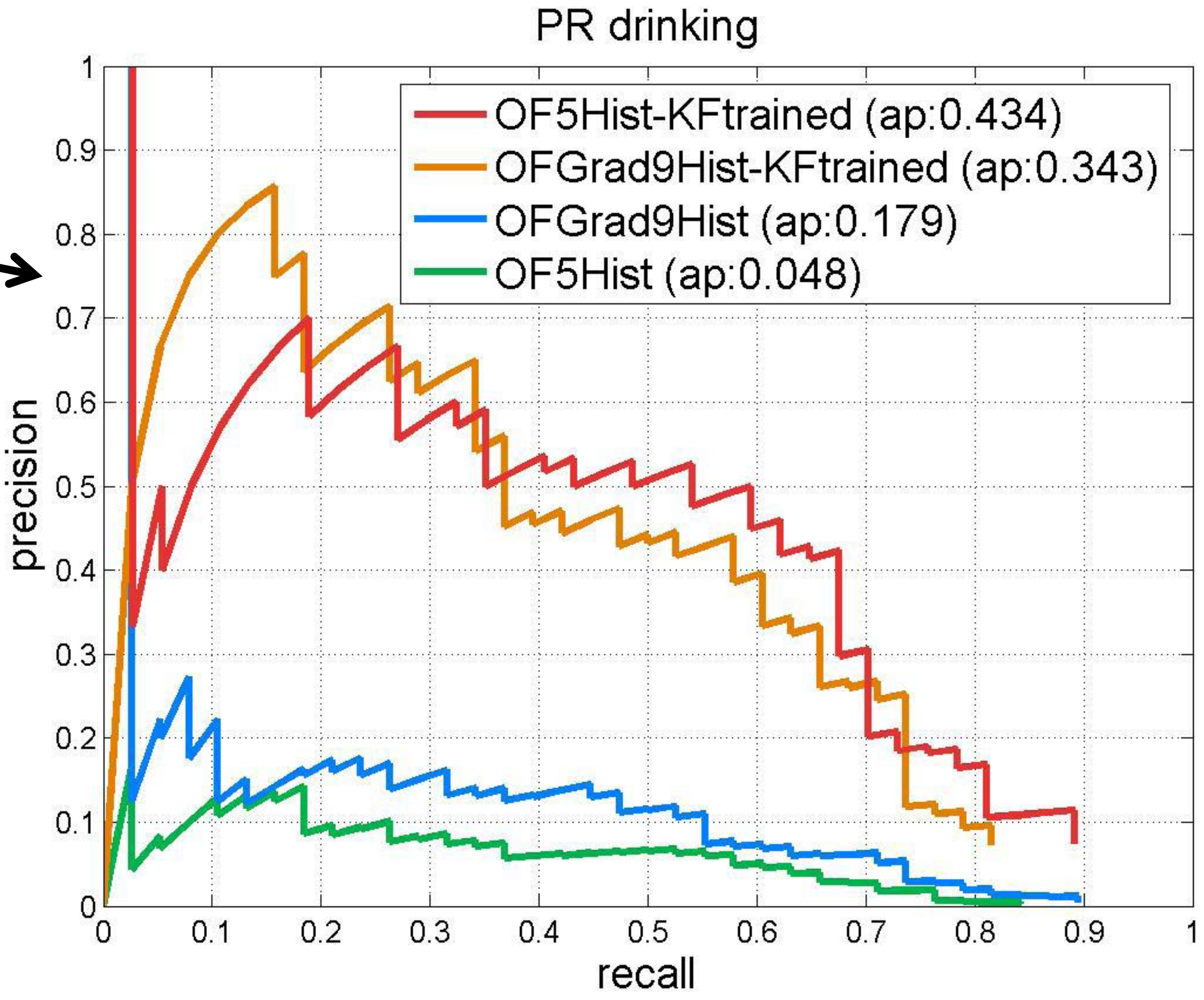


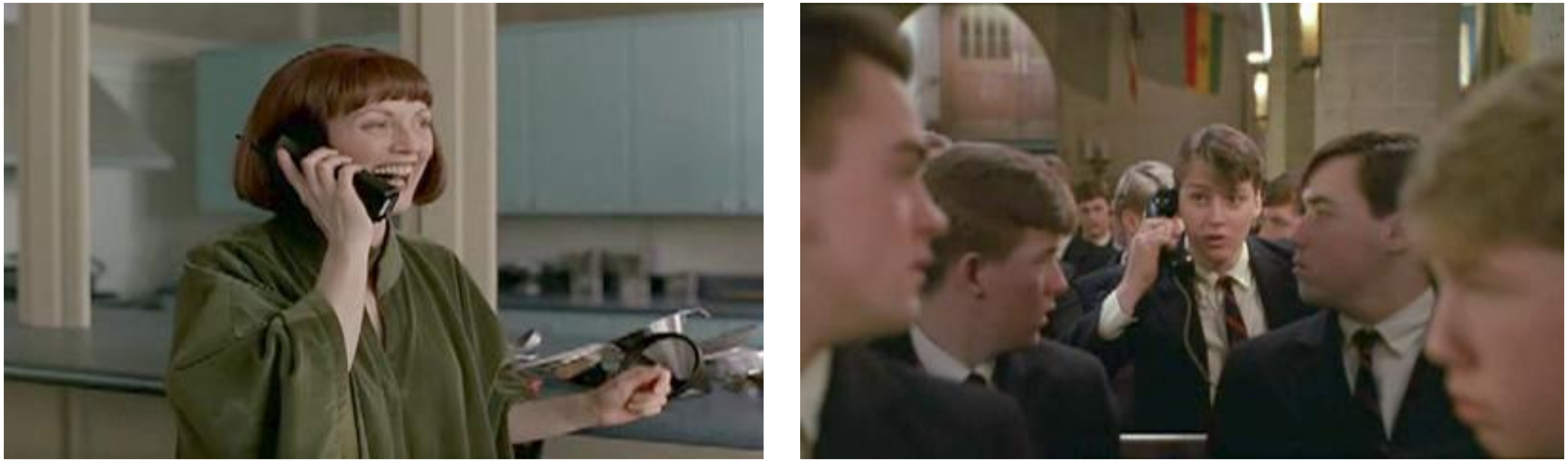


Accuracy in searching video

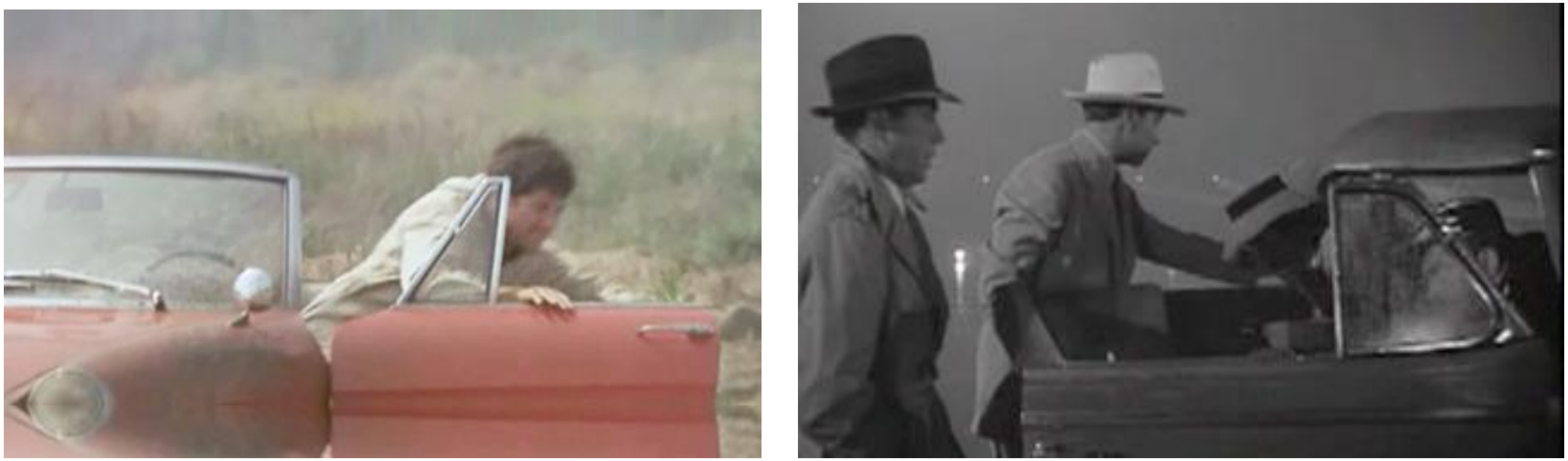
With keyframe detection

Without keyframe detection





“Talk on phone”

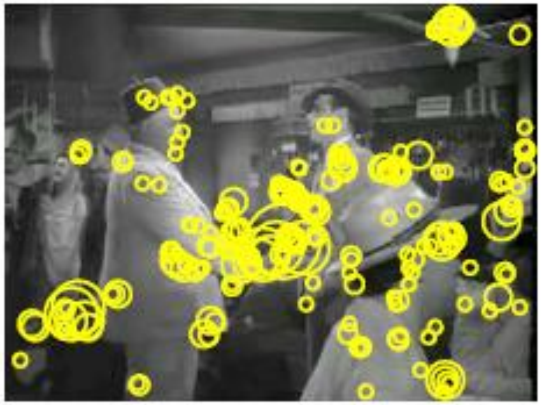


“Get out of car”

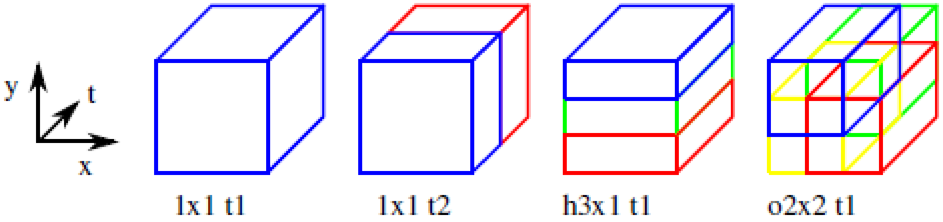
Learning realistic human actions from movies, Laptev et al. 2008

Approach

- Space-time interest point detectors
- Descriptors
  - – HOG, HOF
- Pyramid histograms (3x3x2)
- SVMs with Chi-Squared Kernel

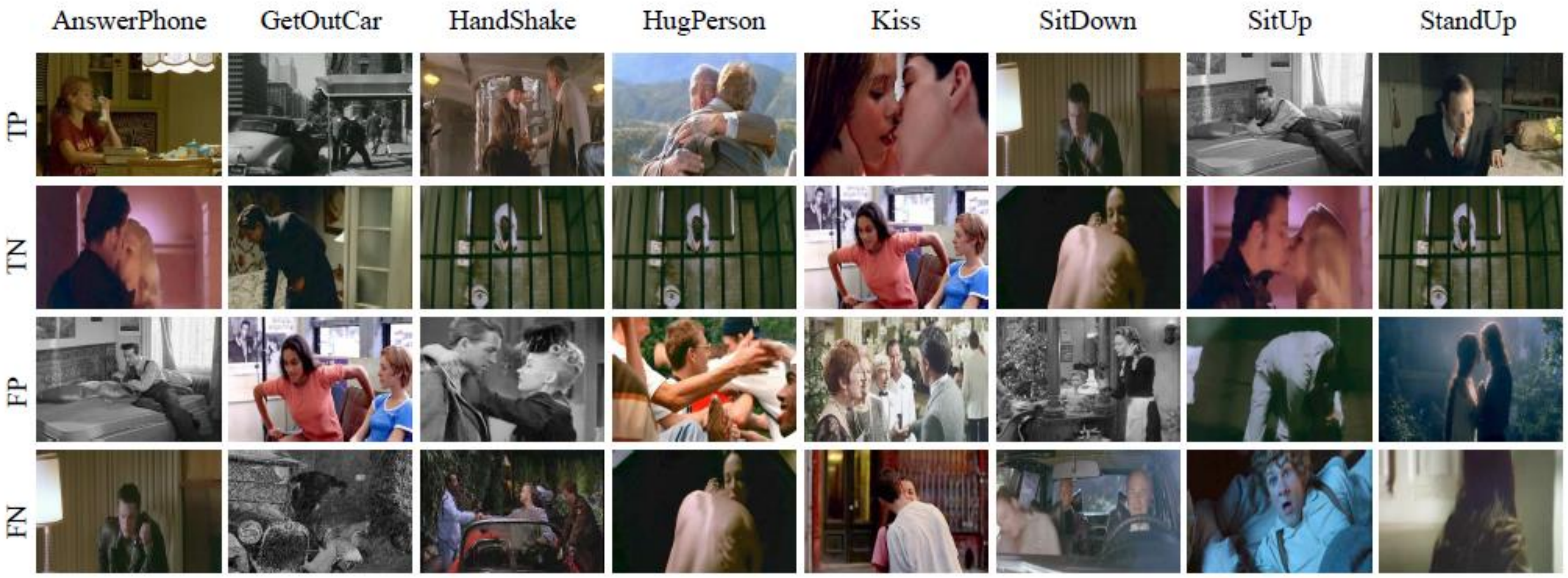


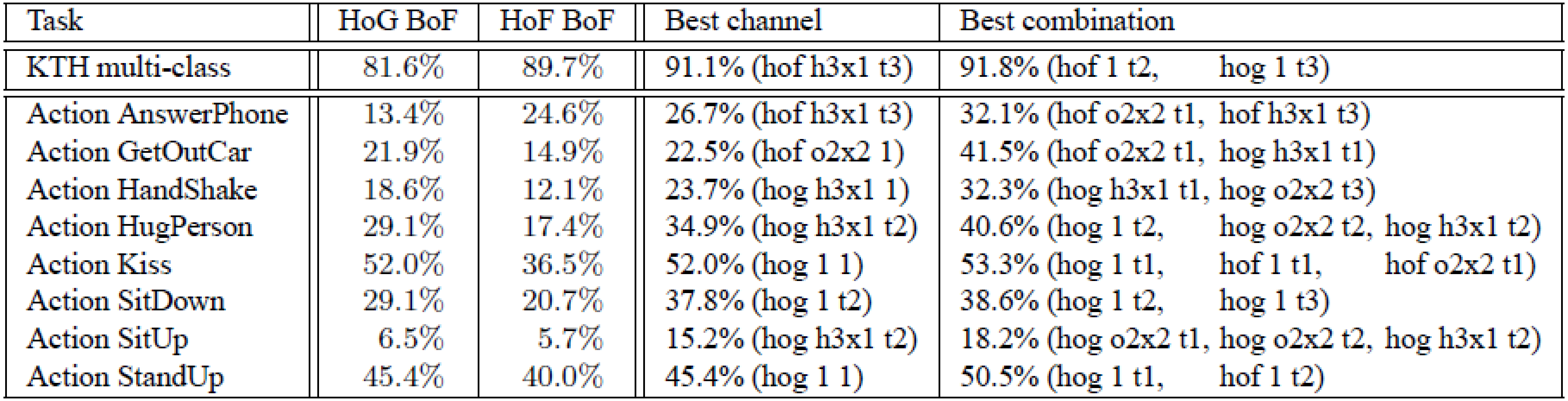
Interest Points



Spatio-Temporal Binning

Results





Task	HoG BoF	HoF BoF	Best channel	Best combination
KTH multi-class	81.6%	89.7%	91.1% (hof h3x1 t3)	91.8% (hof 1 t2, hog 1 t3)
Action AnswerPhone	13.4%	24.6%	26.7% (hof h3x1 t3)	32.1% (hof o2x2 t1, hof h3x1 t3)
Action GetOutCar	21.9%	14.9%	22.5% (hof o2x2 1)	41.5% (hof o2x2 t1, hog h3x1 t1)
Action HandShake	18.6%	12.1%	23.7% (hog h3x1 1)	32.3% (hog h3x1 t1, hog o2x2 t3)
Action HugPerson	29.1%	17.4%	34.9% (hog h3x1 t2)	40.6% (hog 1 t2, hog o2x2 t2, hog h3x1 t2)
Action Kiss	52.0%	36.5%	52.0% (hog 1 1)	53.3% (hog 1 t1, hof 1 t1, hof o2x2 t1)
Action SitDown	29.1%	20.7%	37.8% (hog 1 t2)	38.6% (hog 1 t2, hog 1 t3)
Action SitUp	6.5%	5.7%	15.2% (hog h3x1 t2)	18.2% (hog o2x2 t1, hog o2x2 t2, hog h3x1 t2)
Action StandUp	45.4%	40.0%	45.4% (hog 1 1)	50.5% (hog 1 t1, hof 1 t2)

Take-home messages

- Action recognition is an open problem.
  - How to define actions?
  - How to infer them?
  - What are good visual cues?
  - How do we incorporate higher level reasoning?

Take-home messages

- Some work done, but it is just the beginning of exploring the problem. So far…
  - Actions are mainly categorical (could be framed in terms of effect or intent)
  - Most approaches are classification using simple features (spatial-temporal histograms of gradients or flow, s-t interest points, SIFT in images)
  - Just a couple works on how to incorporate pose and objects
  - Not much idea of how to reason about long-term activities or to describe video sequences