11/21/11

Activity Recognition

Computer Vision

CS 143, Brown

James Hays

With slides by
Derek Hoiem and
Kristen Grauman

What is an action?

Action: a transition from one state to another

- Who is the actor?
- How is the state of the actor changing?
- What (if anything) is being acted on?
- How is that thing changing?
- What is the purpose of the action (if any)?

Human activity in video

No universal terminology, but approximately:

- "**Actions**": atomic motion patterns -- often gesture-like, single clear-cut trajectory, single nameable behavior (e.g., sit, wave arms)

- "**Activity**": series or composition of actions (e.g., interactions between people)

- "**Event**": combination of activities or actions (e.g., a football game, a traffic accident)

Adapted from Venu Govindaraju

How do we represent actions?

Categories

Walking, hammering, dancing, skiing, sitting down, standing up, jumping

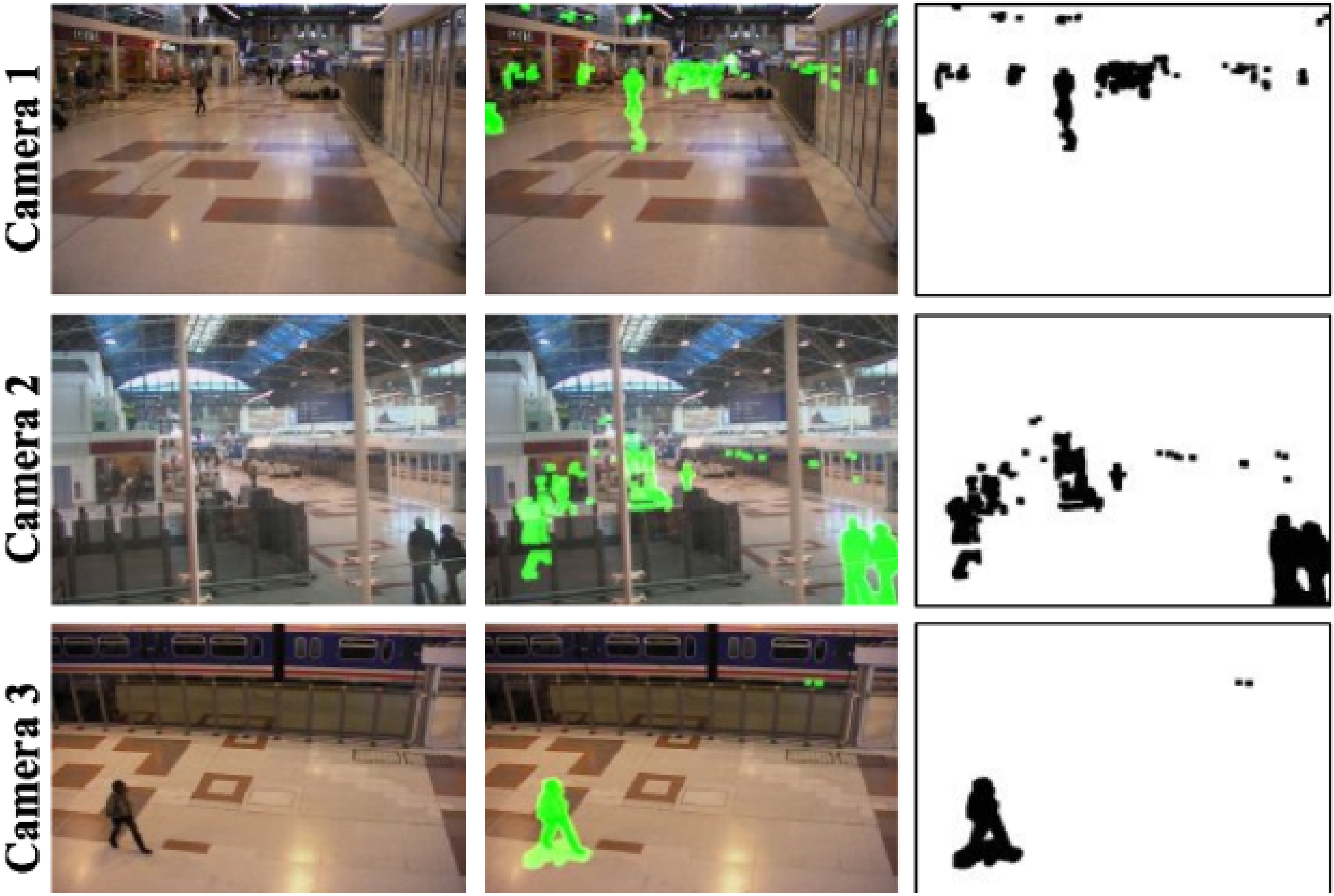
Poses

Nouns and Predicates

<man, swings, hammer>
<man, hits, nail, w/ hammer>

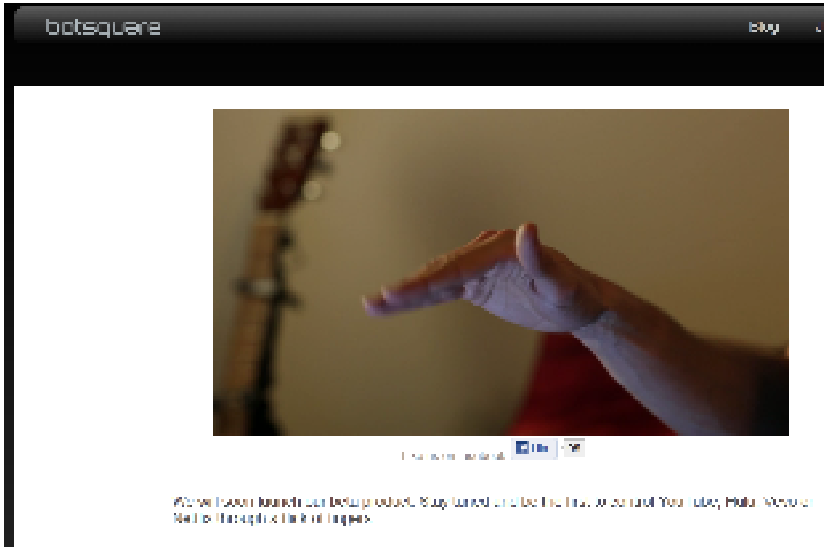
What is the purpose of action recognition?

Surveillance



http://users.isr.ist.utl.pt/~etienne/mypubs/Auvinetal06PETS.pdf

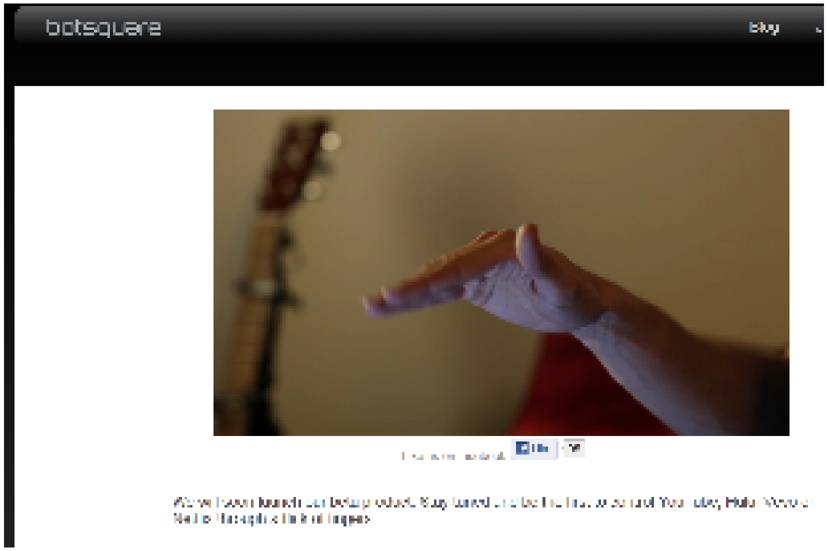
Interfaces



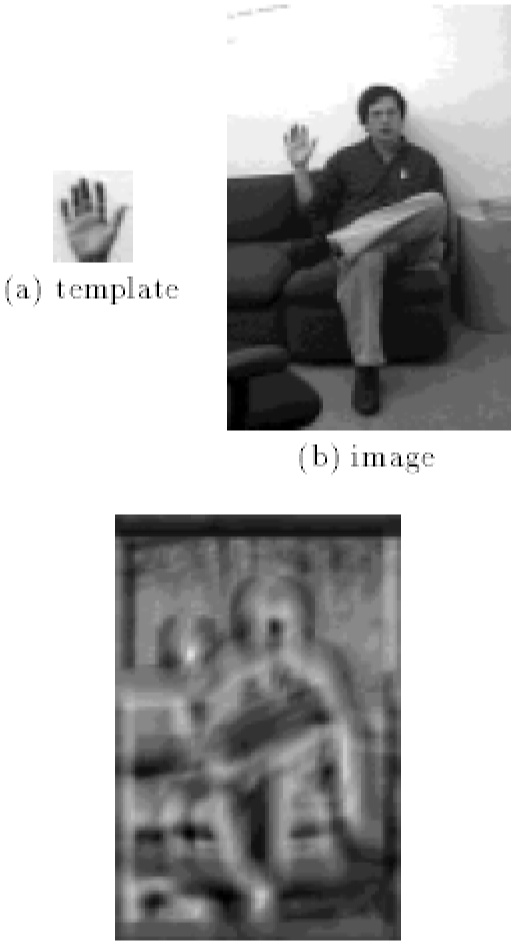
We've been launch our beta product. Stay tuned and be the first to control YouTube, Hulu, Vimeo or Netflix through a flick of fingers.

2011

Interfaces



We've just launched our beta product. Stay tuned and be the first to control YouTube, Hulu, Vimeo or Netflix through a flick of fingers.


(a) template

(b) image

(c) normalized correlation

2011 1995

W. T. Freeman and C. Weissman, *Television control by hand gestures*, International Workshop on Automatic Face- and Gesture- Recognition, IEEE Computer Society, Zurich, Switzerland, June, 1995, pp. 179--183. MERL-TR94-24

How can we identify actions?

Motion

Pose

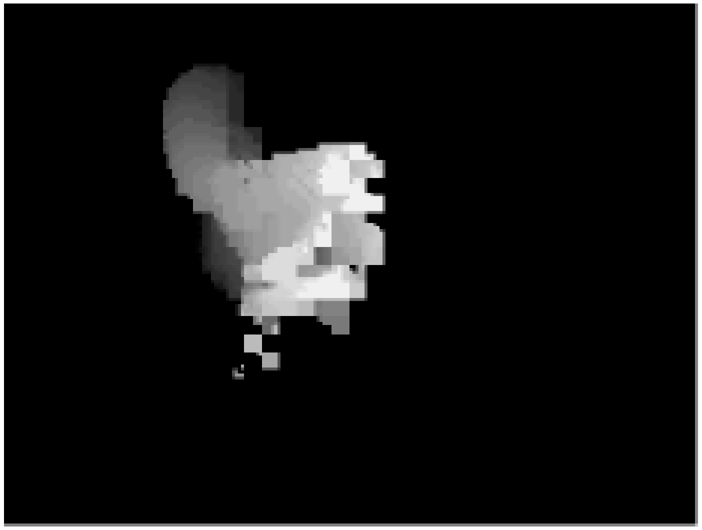
Held Objects

Nearby Objects

Representing Motion

Optical Flow with Motion History

sit-down

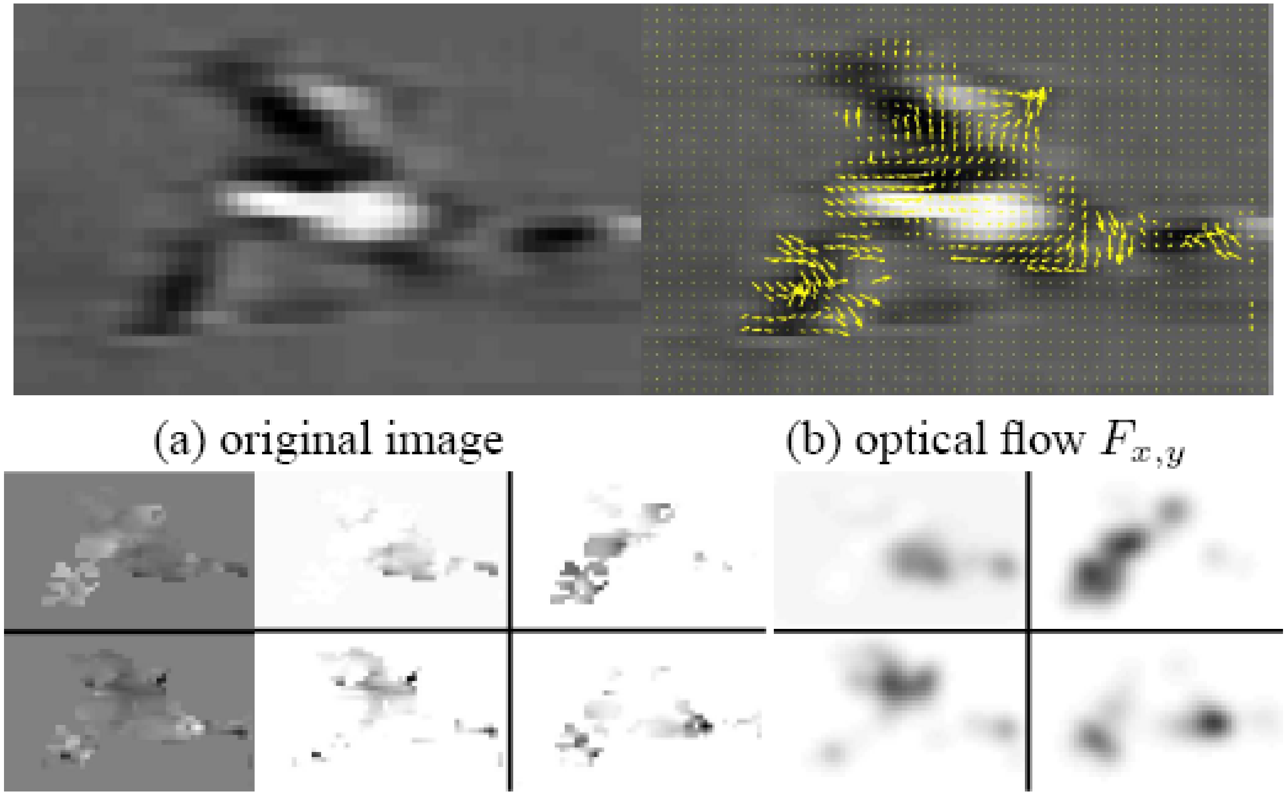


sit-down MHI

Bobick Davis 2001

Representing Motion

Optical Flow with Split Channels



(a) original image

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

(c) F_x, F_y

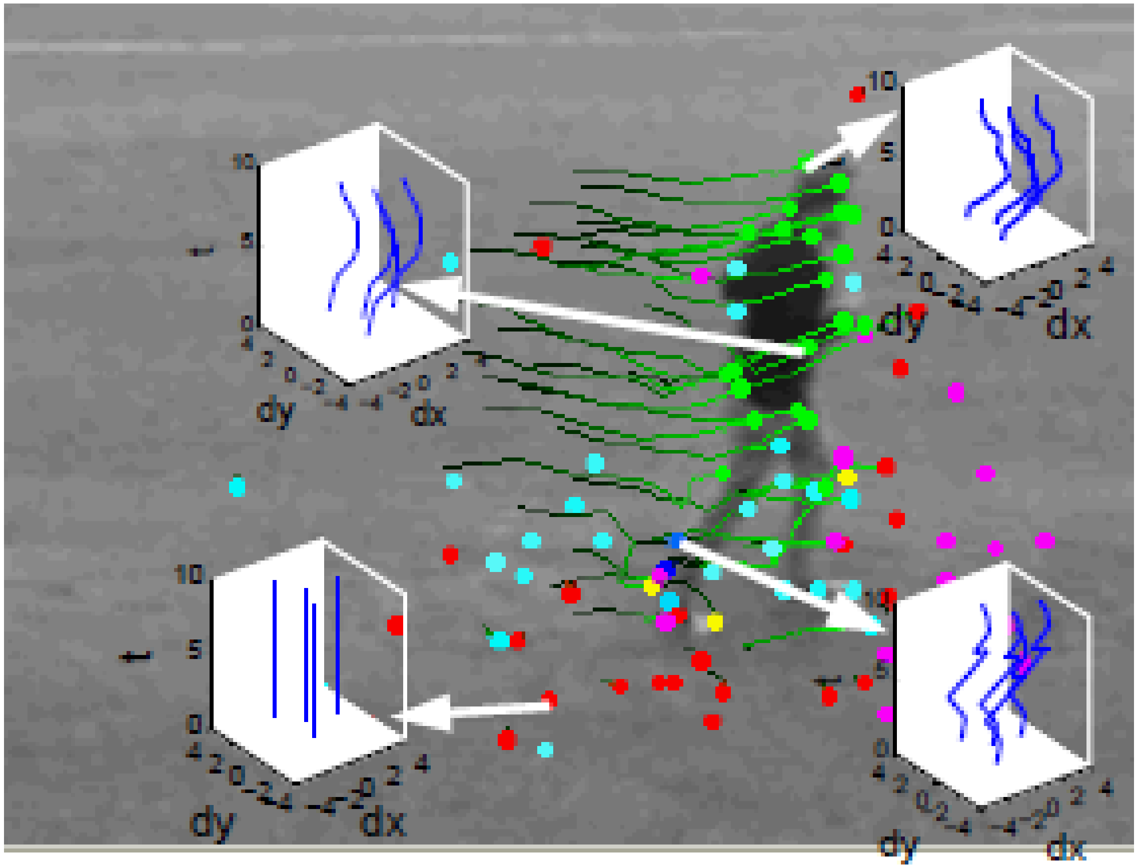
(d) $F_x^+, F_x^-, F_y^+, F_y^-$

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

Efros et al. 2003

Representing Motion

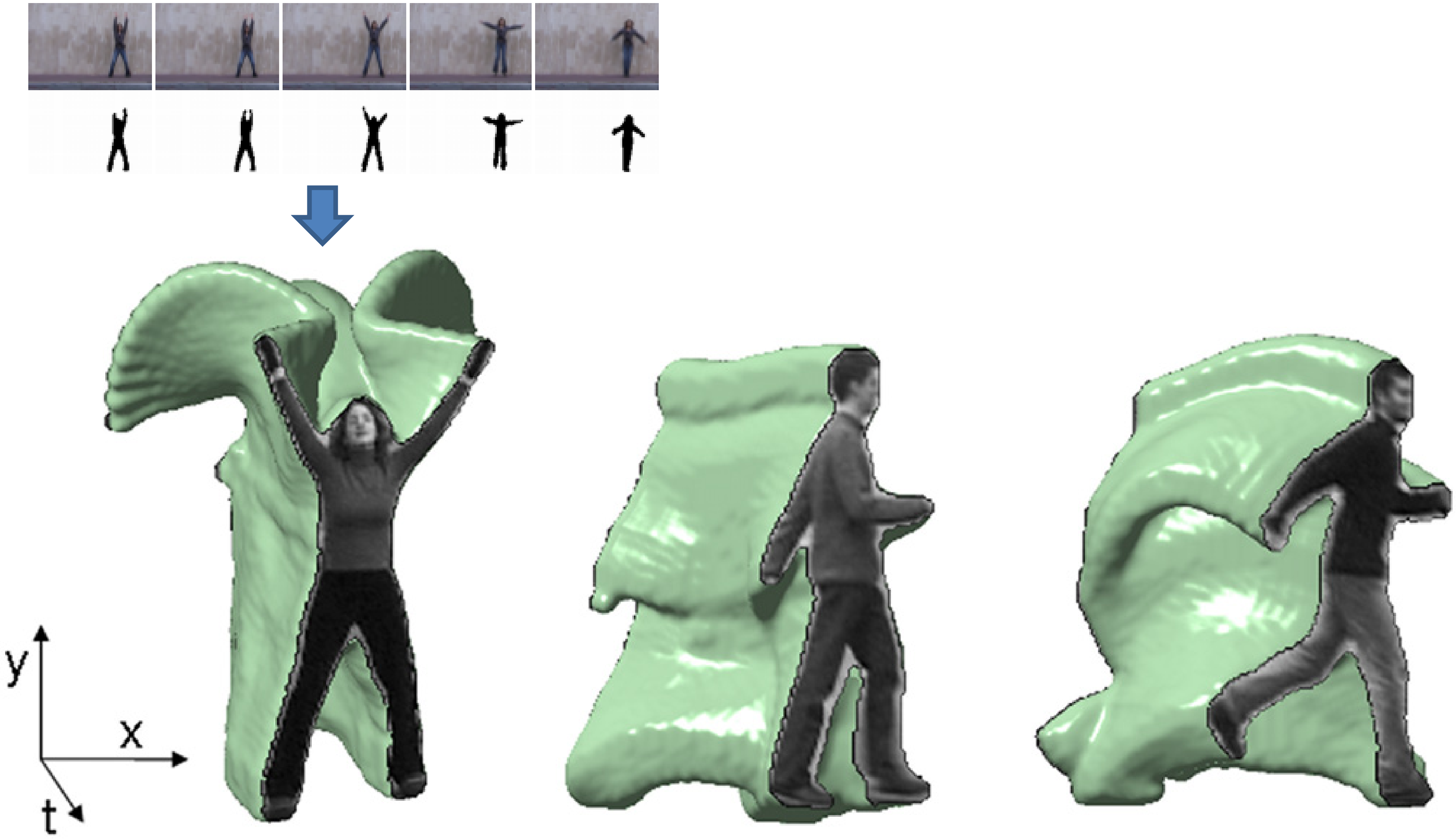
Tracked Points



Matikainen et al. 2009

Representing Motion

Space-Time Volumes



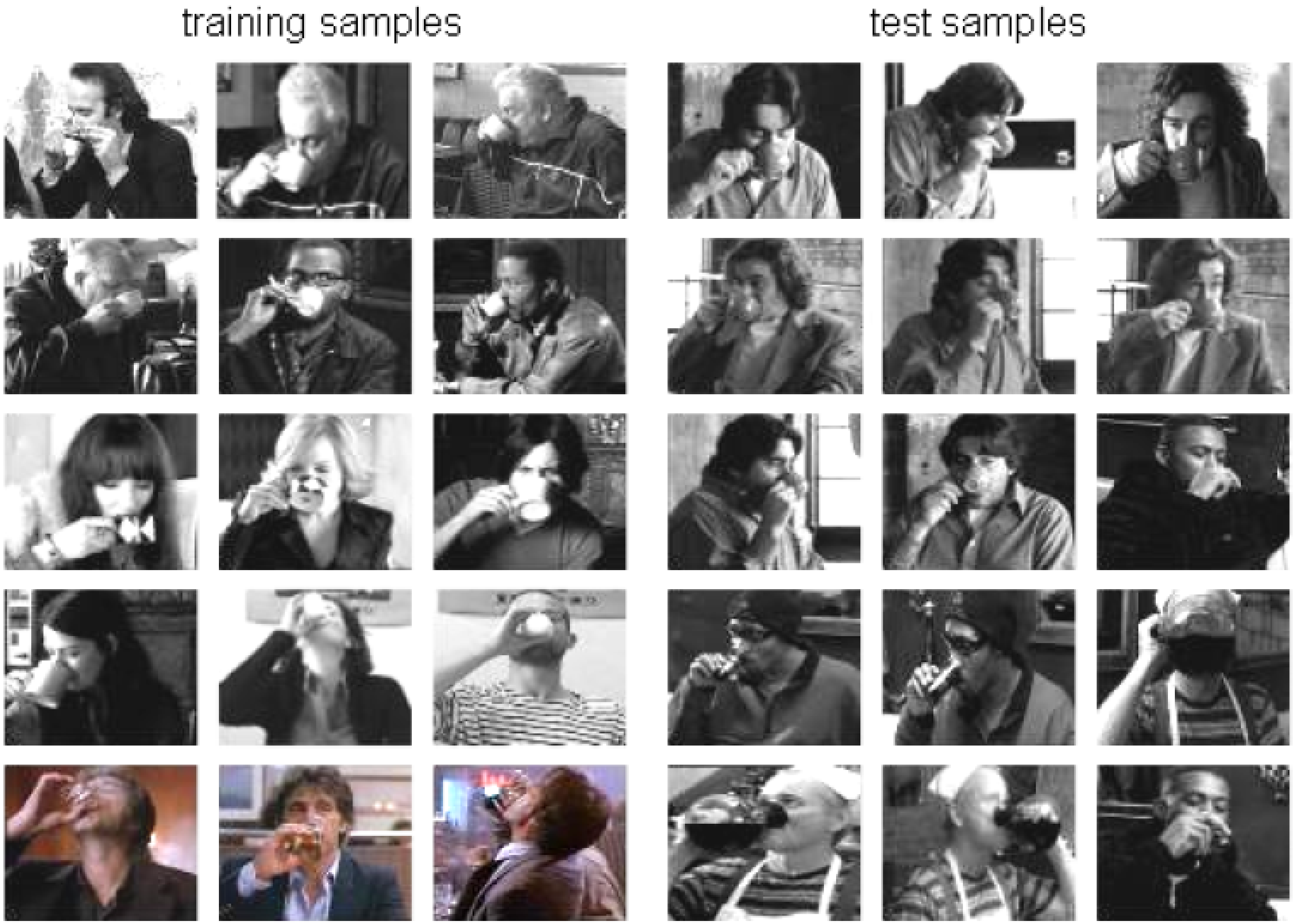
Blank et al. 2005

Examples of Action Recognition Systems

- Feature-based classification

- Recognition using pose and objects

Action recognition as classification



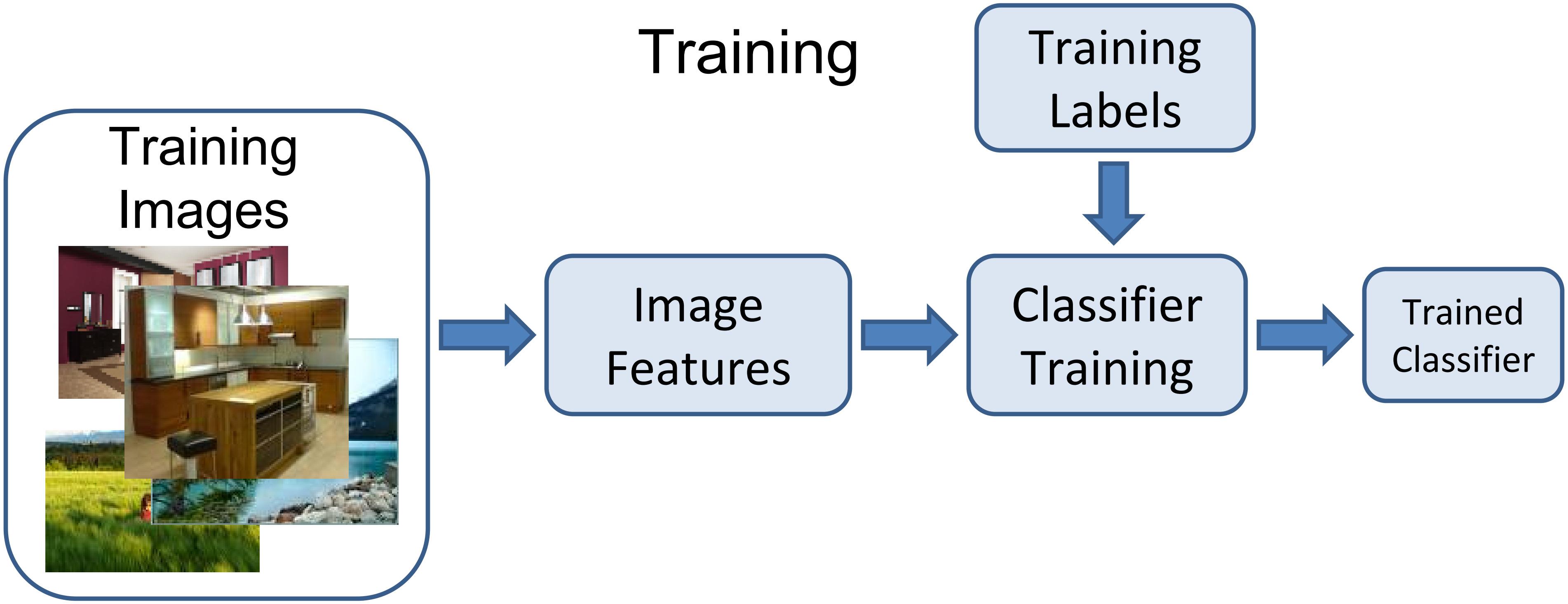
training samples

test samples

Retrieving actions in movies, Laptev and Perez, 2007

Remember image categorization…

Training



Training Images

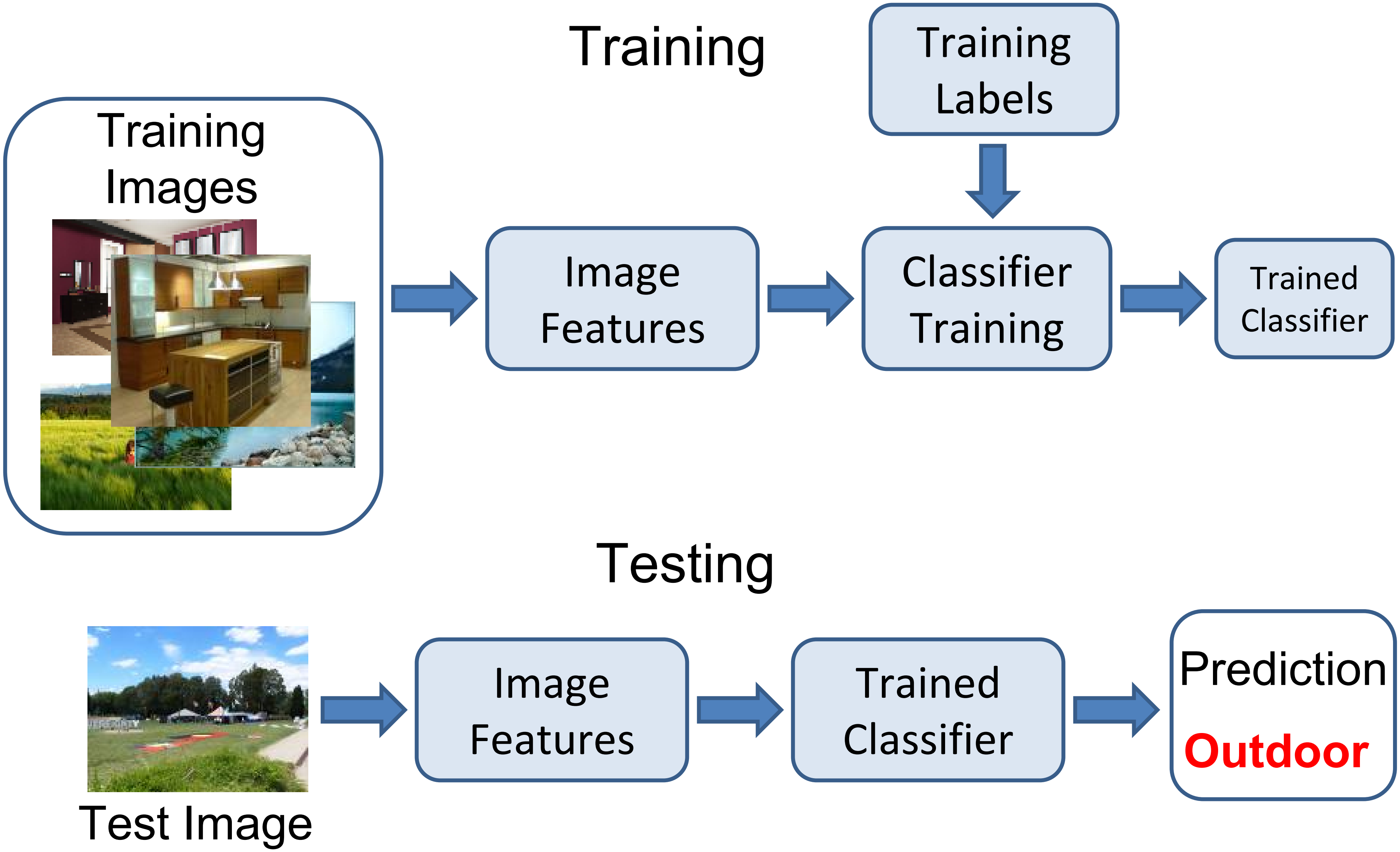
Image Features

Training Labels

Classifier Training

Trained Classifier

Remember image categorization…



Training

Training Images

Image Features

Training Labels

Classifier Training

Trained Classifier

Testing

Test Image

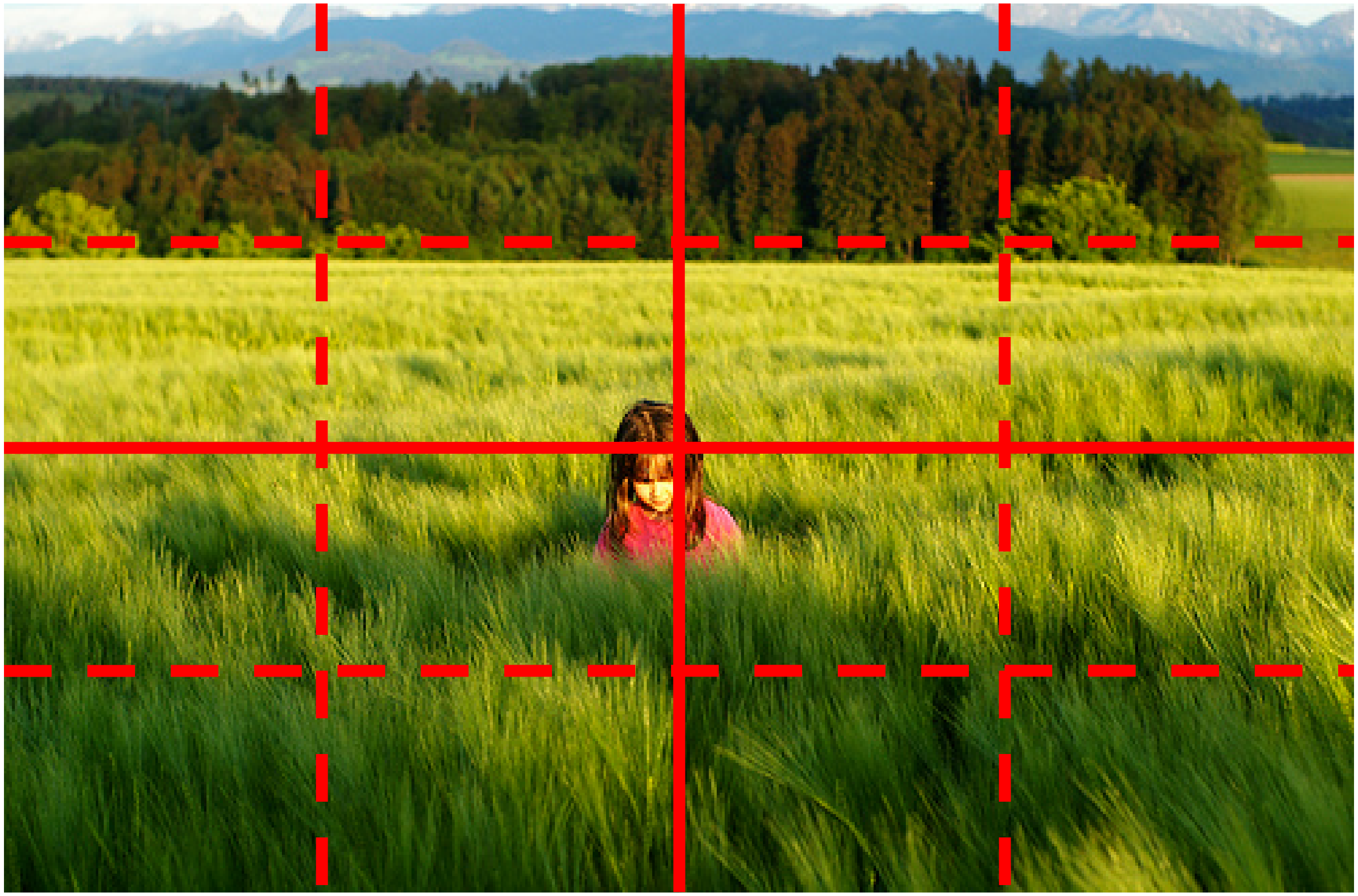
Image Features

Trained Classifier

Prediction

Outdoor

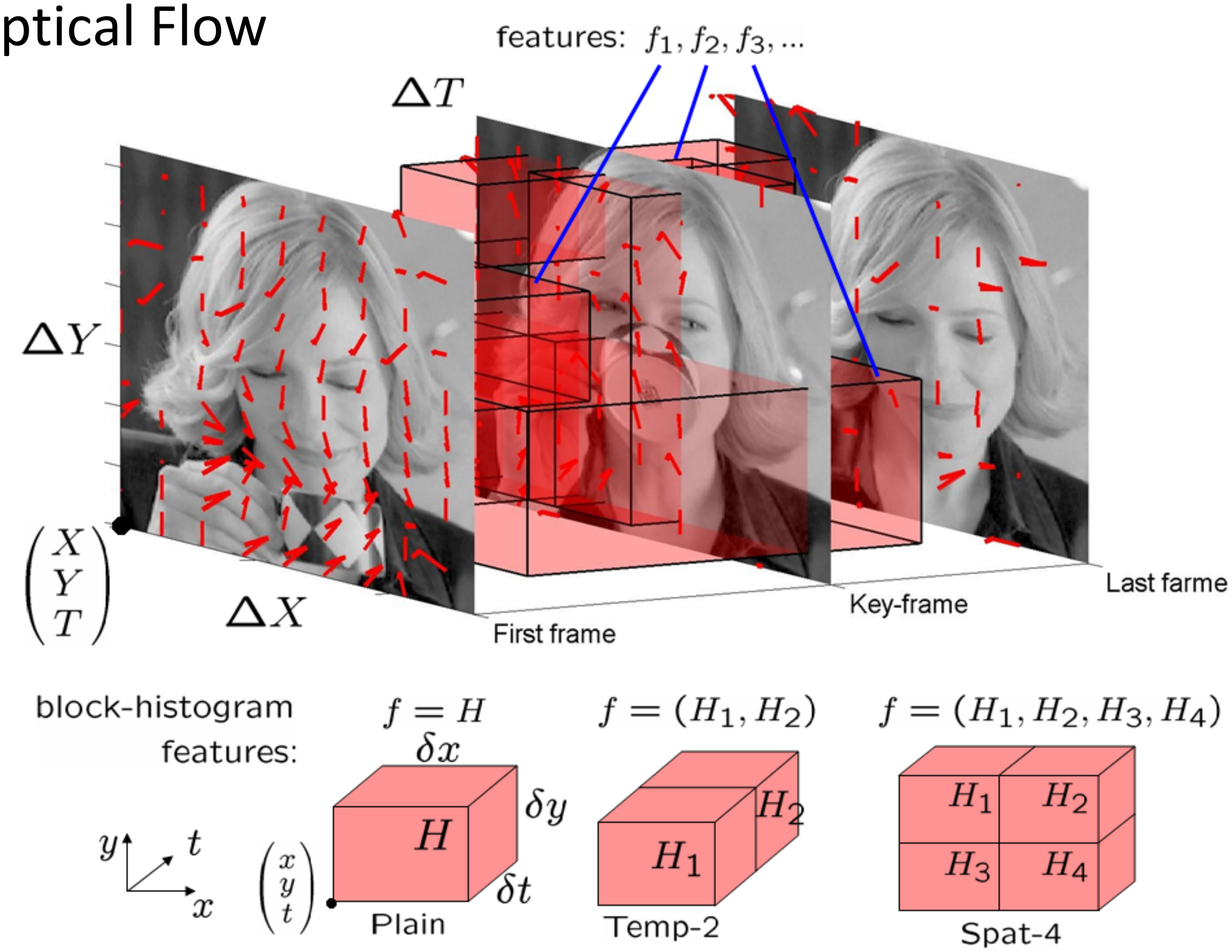
Remember spatial pyramids….



Compute histogram in each spatial bin

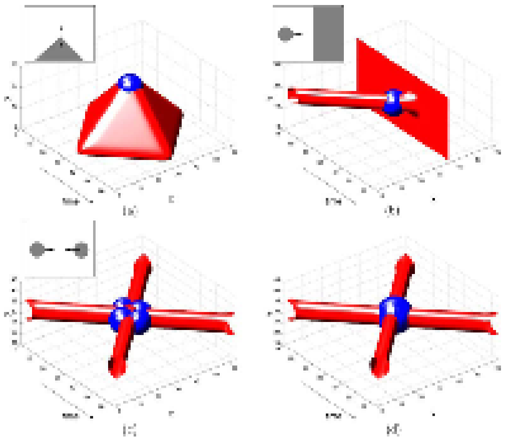
Features for Classifying Actions

1. Spatio-temporal pyramids (14x14x8 bins)
   - Image Gradients
   - Optical Flow

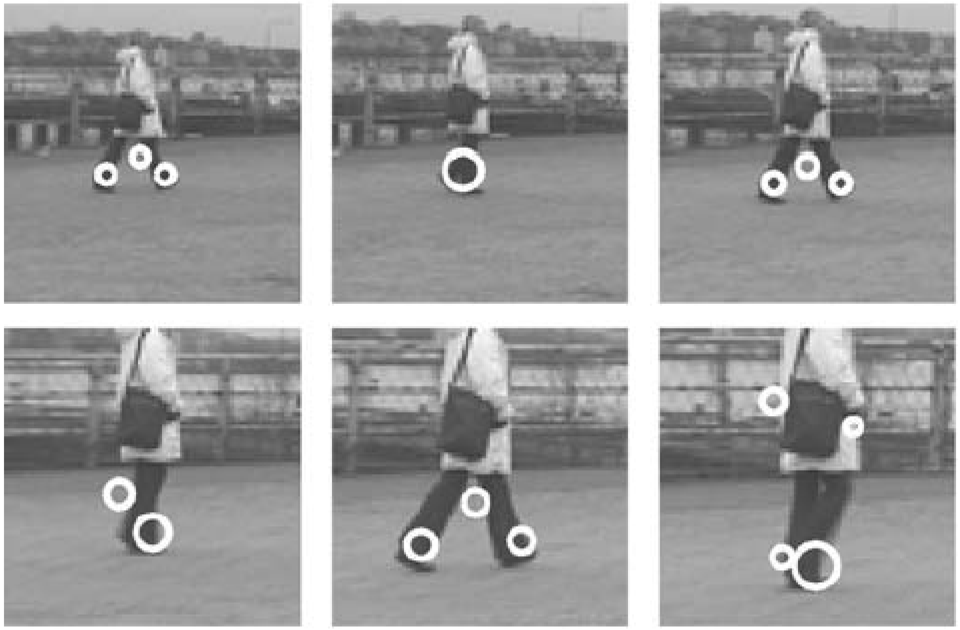


Features for Classifying Actions

2. Spatio-temporal interest points



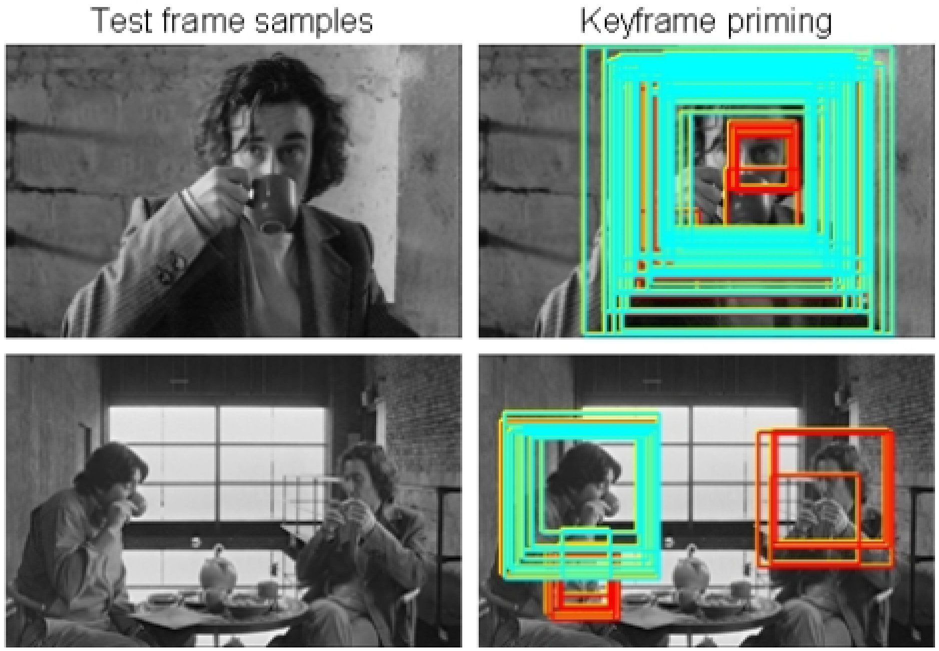
Corner detectors in space-time

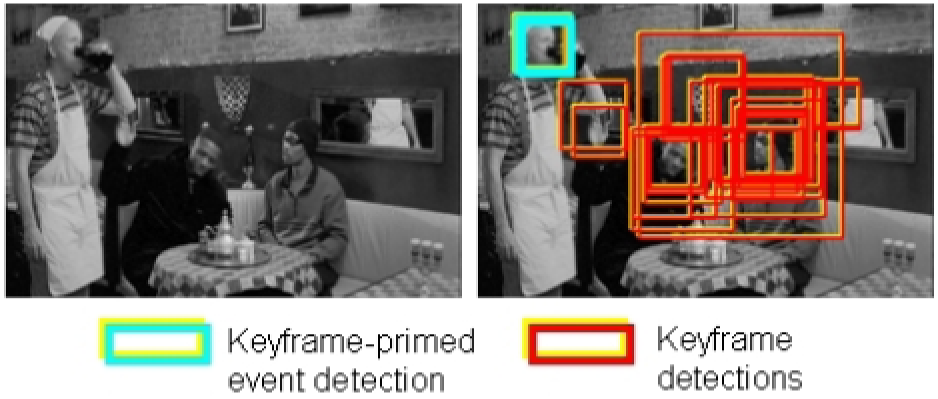


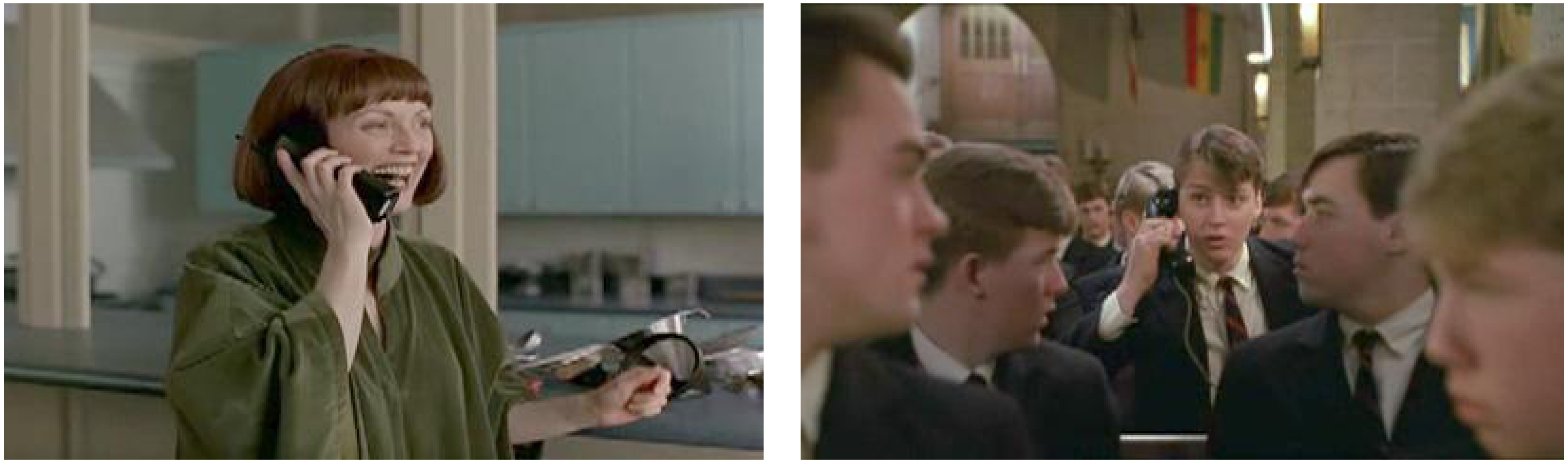
Descriptors based on Gaussian derivative filters over x, y, time

Searching the video for an action

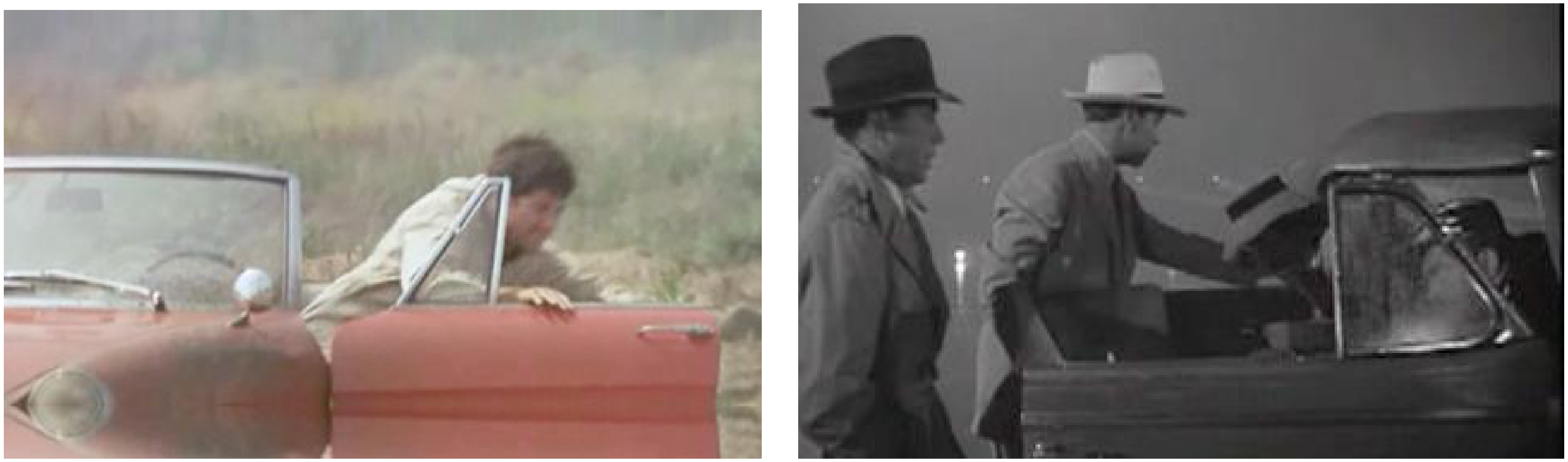
1. Detect keyframes using a trained HOG detector in each frame
2. Classify detected keyframes as positive (e.g., “drinking”) or negative (“other”)







“Talk on phone”



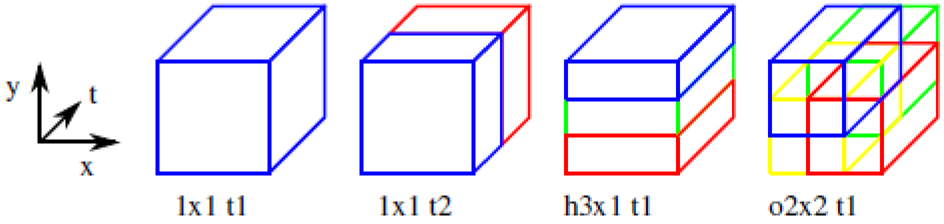
“Get out of car”

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

Approach

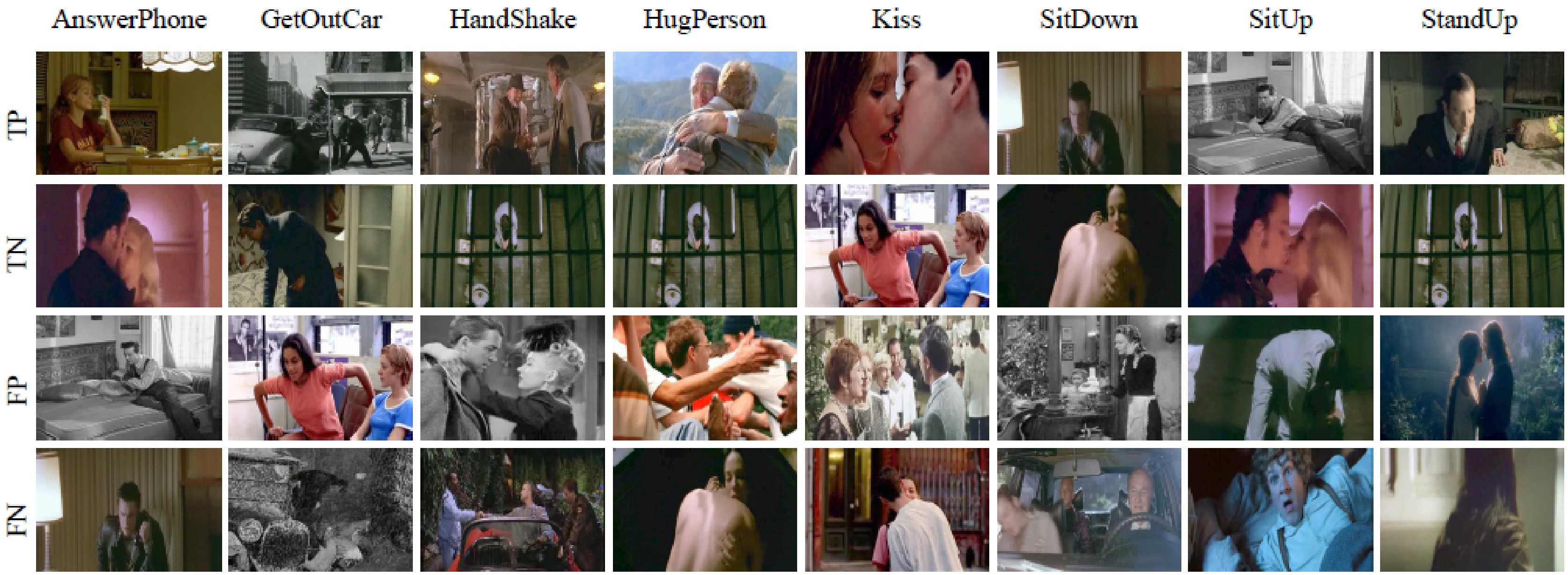
- Space-time interest point detectors
- Descriptors
  - HOG, HOF
- Pyramid histograms (3x3x2)
- SVMs with Chi-Squared Kernel

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)

Action Recognition using Pose and Objects

Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities, B. Yao and Li Fei-Fei, 2010

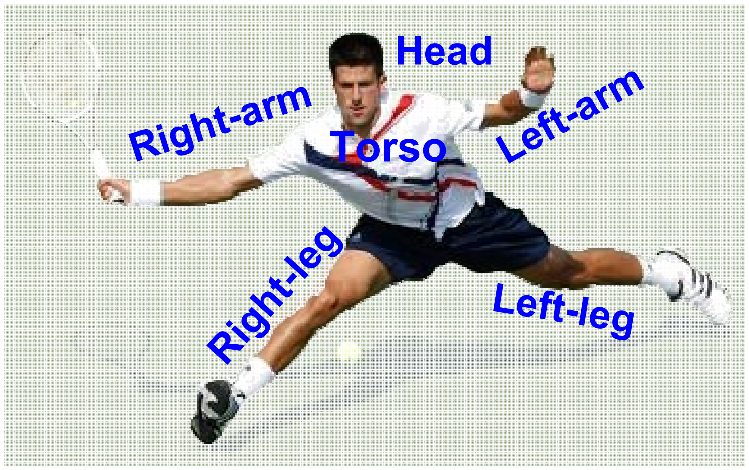
Slide Credit: Yao/Fei-Fei

Human-Object Interaction

Holistic image based classification

Integrated reasoning

- **Human pose estimation**



Slide Credit: Yao/Fei-Fei

Human-Object Interaction

Holistic image based classification

Integrated reasoning

- Human pose estimation
- **Object detection**

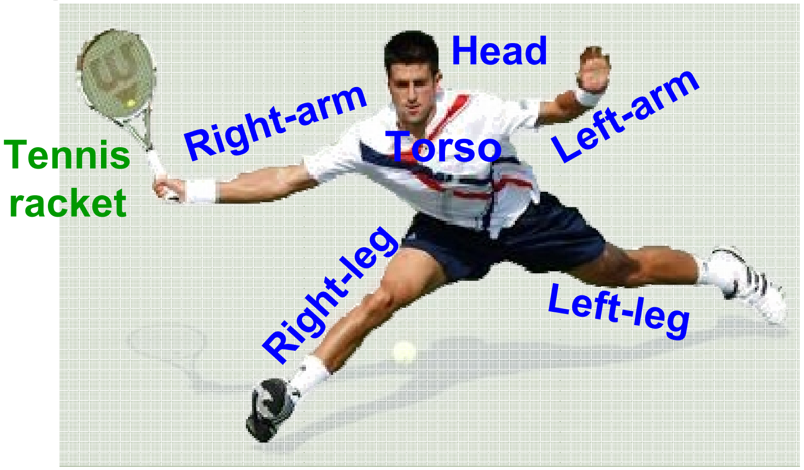
Slide Credit: Yao/Fei-Fei

Human-Object Interaction

Holistic image based classification

Integrated reasoning

- **Human pose estimation**
- **Object detection**
- **Action categorization**

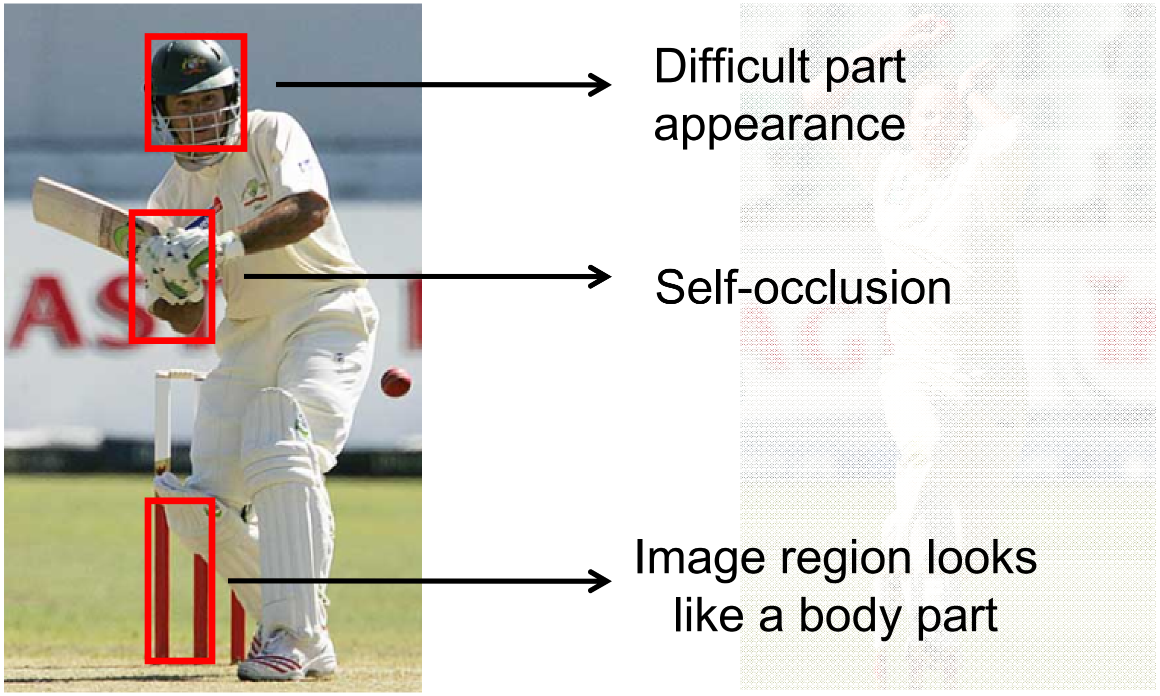


HOI activity: Tennis Forehand

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Human pose estimation is challenging.



Difficult part appearance

Self-occlusion

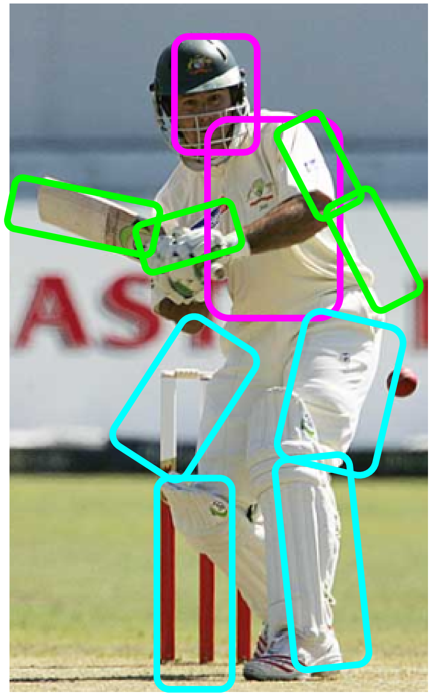
Image region looks like a body part

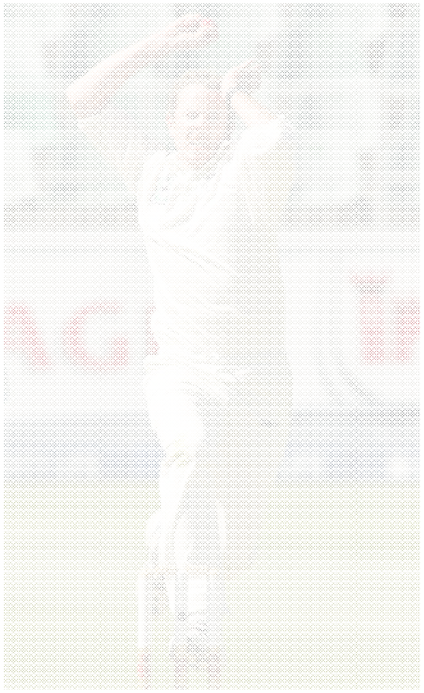
- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Human pose estimation is challenging.





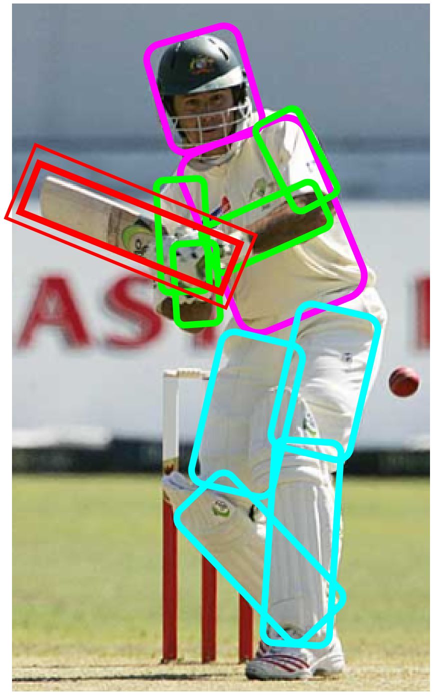
- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

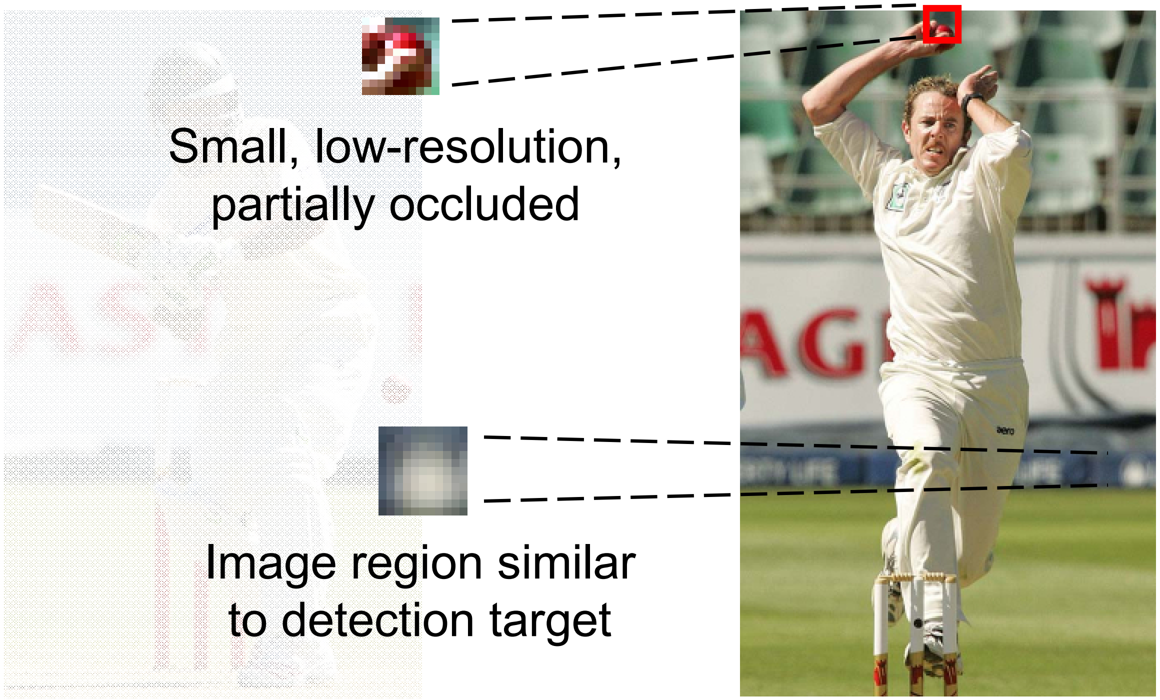
Facilitate

Given the object is detected.



Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection



Small, low-resolution, partially occluded

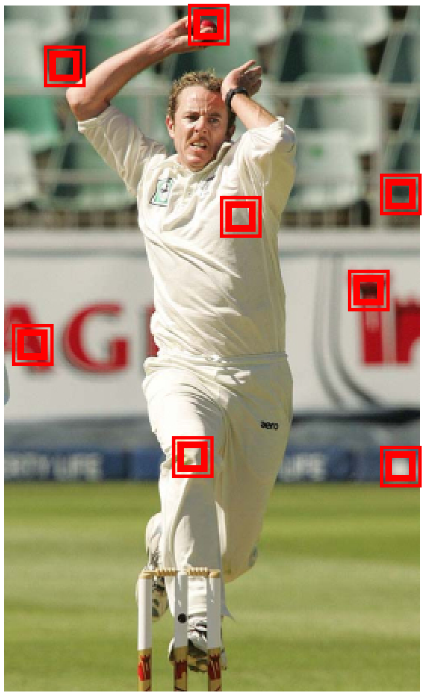
Image region similar to detection target

Object detection is challenging

- Viola & Jones, 2001
- Lampert et al, 2008
- Divvala et al, 2009
- Vedaldi et al, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection



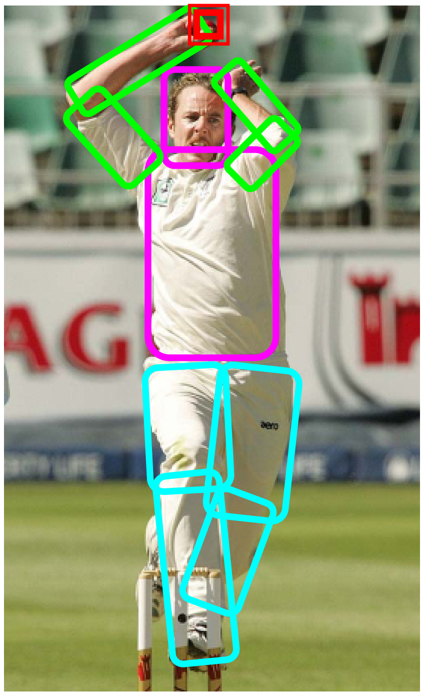
Object detection is challenging

- Viola & Jones, 2001
- Lampert et al, 2008
- Divvala et al, 2009
- Vedaldi et al, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Facilitate

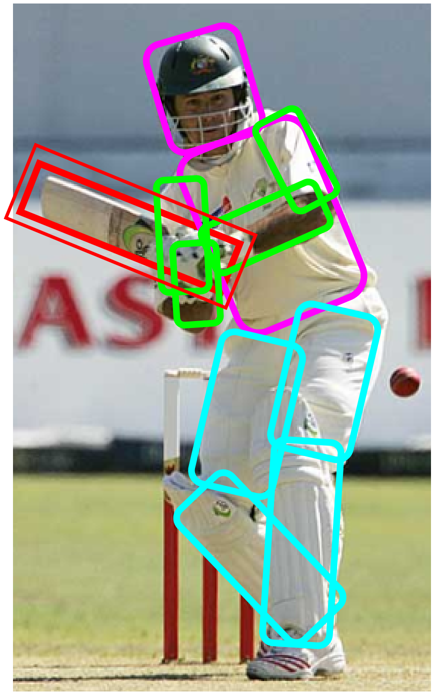


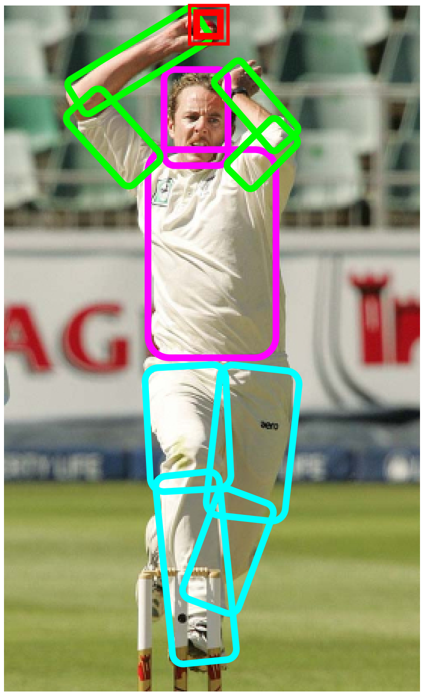
Given the pose is estimated.

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Mutual Context





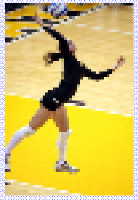
Slide Credit: Yao/Fei-Fei

Mutual Context Model Representation

A:

Tennis forehand

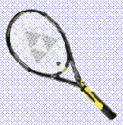
Croquet shot



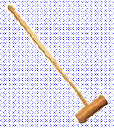
Volleyball smash

…

O:



Tennis racket

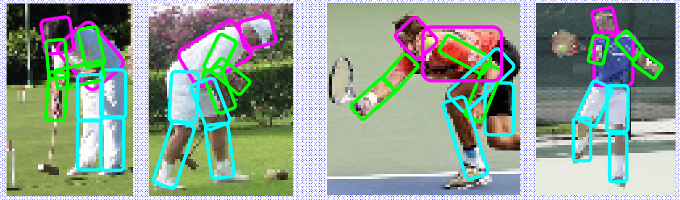


Croquet mallet

Volleyball

…

H:

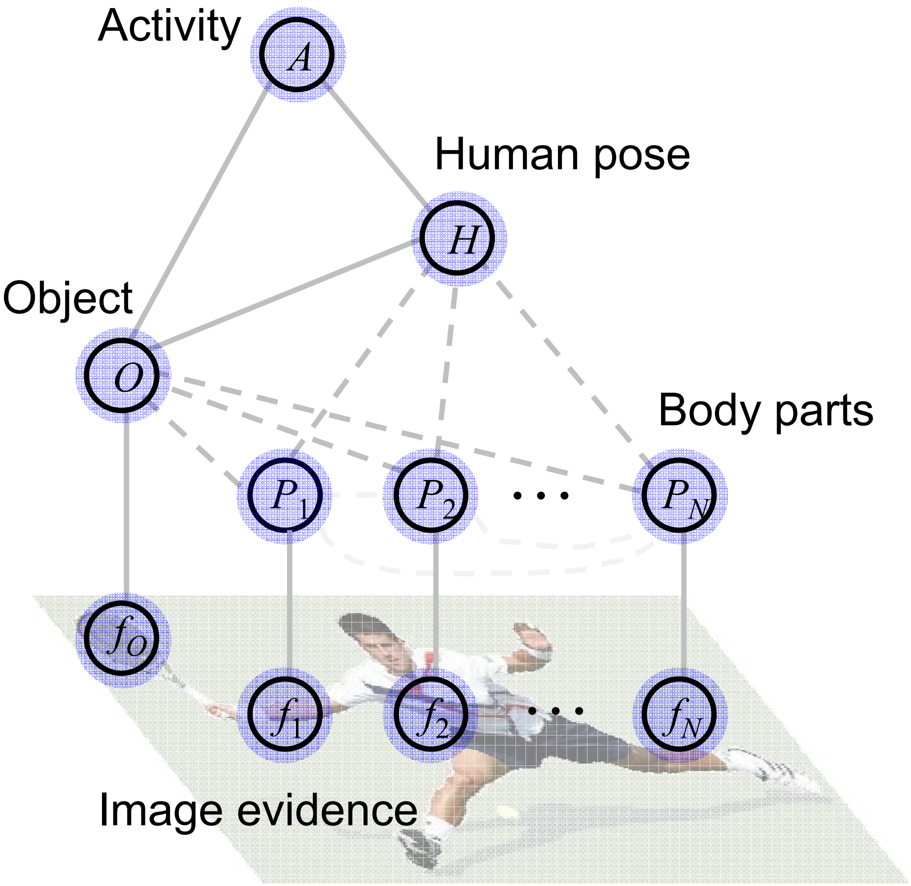


Intra-class variations

- More than one H for each A;
- **Unobserved** during training.

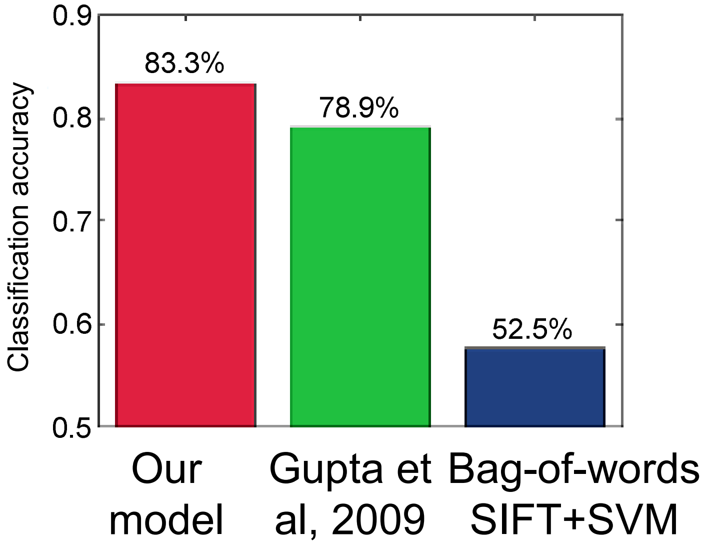
P: l_P: location; θ_P: orientation; s_P: scale.

f: Shape context. [Belongie et al, 2002]

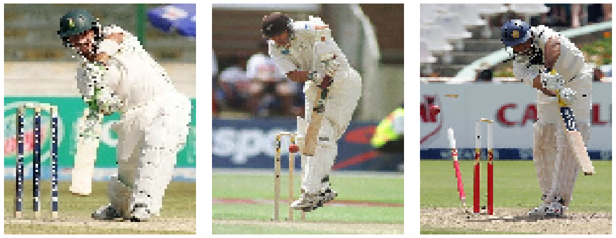


Slide Credit: Yao/Fei-Fei

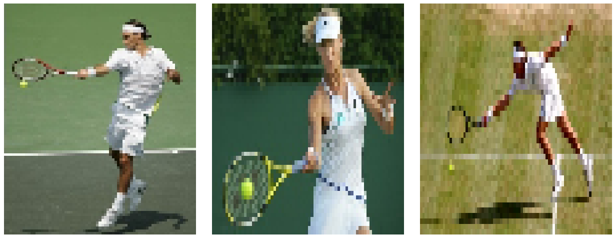
Activity Classification Results

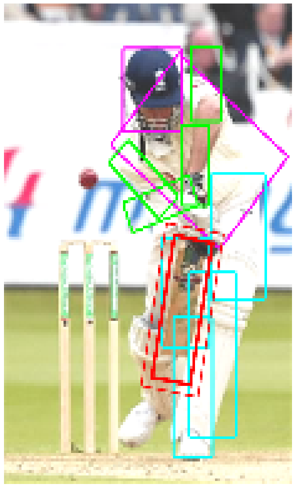


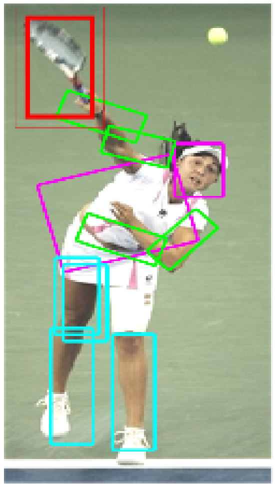
Cricket shot

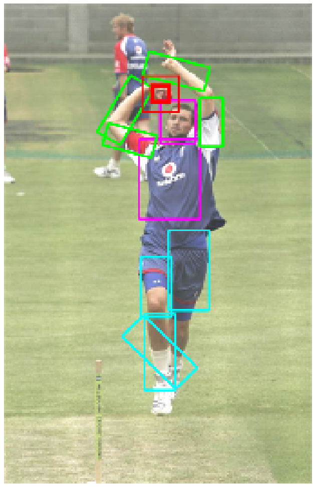


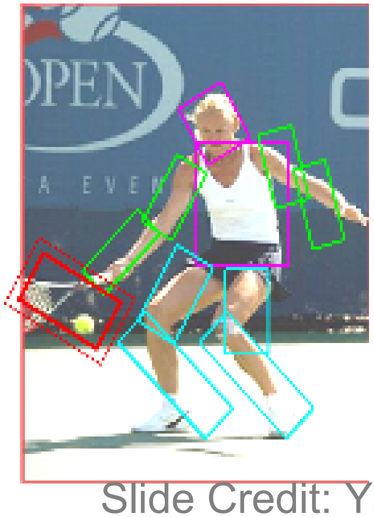
Tennis forehand











Slide Credit: Yao/Fei-Fei

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
  - 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