Action Recognition

Computer Vision

Jia-Bin Huang, Virginia Tech

Many slides from D. Hoiem

This section: advanced topics

- Convolutional neural networks in vision
- Action recognition
- Vision and Language
- 3D Scenes and Context

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)?

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?

To describe

What is the purpose of action recognition?

- To predict

What is the purpose of action recognition?

- To understand the intention and motivation

Why are they doing that?

to sell ice cream | to commute to work

to answer emergency call | to win race

Predicting Motivations of Actions by Leveraging Text, CVPR 2016

How can we identify actions?

Motion

Pose

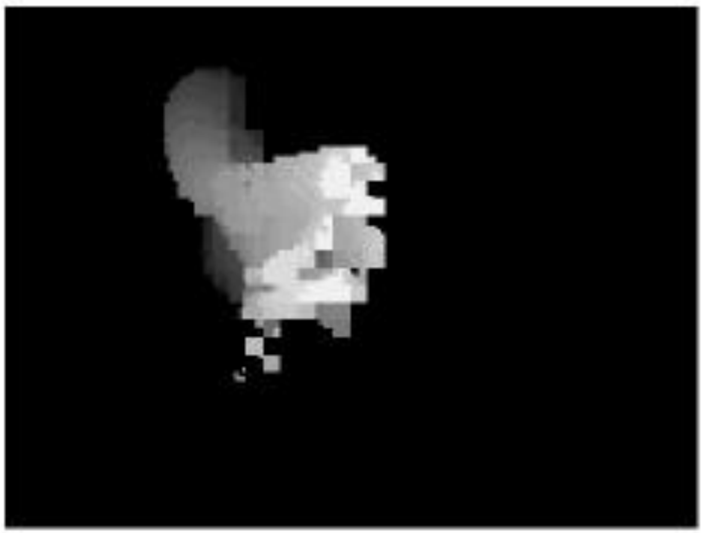
Held Objects

Nearby Objects

Representing Motion

Optical Flow with Motion History

sit-down

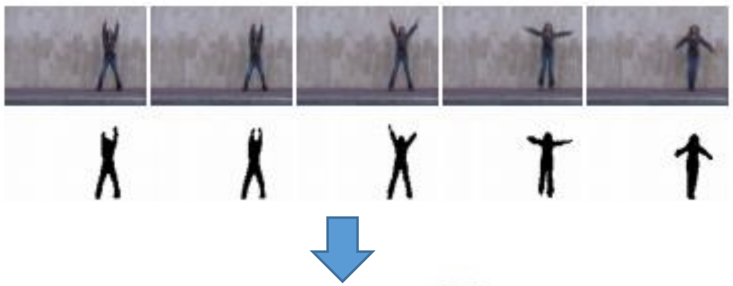


sit-down MHI

[Bobick Davis 2001]

Representing Motion

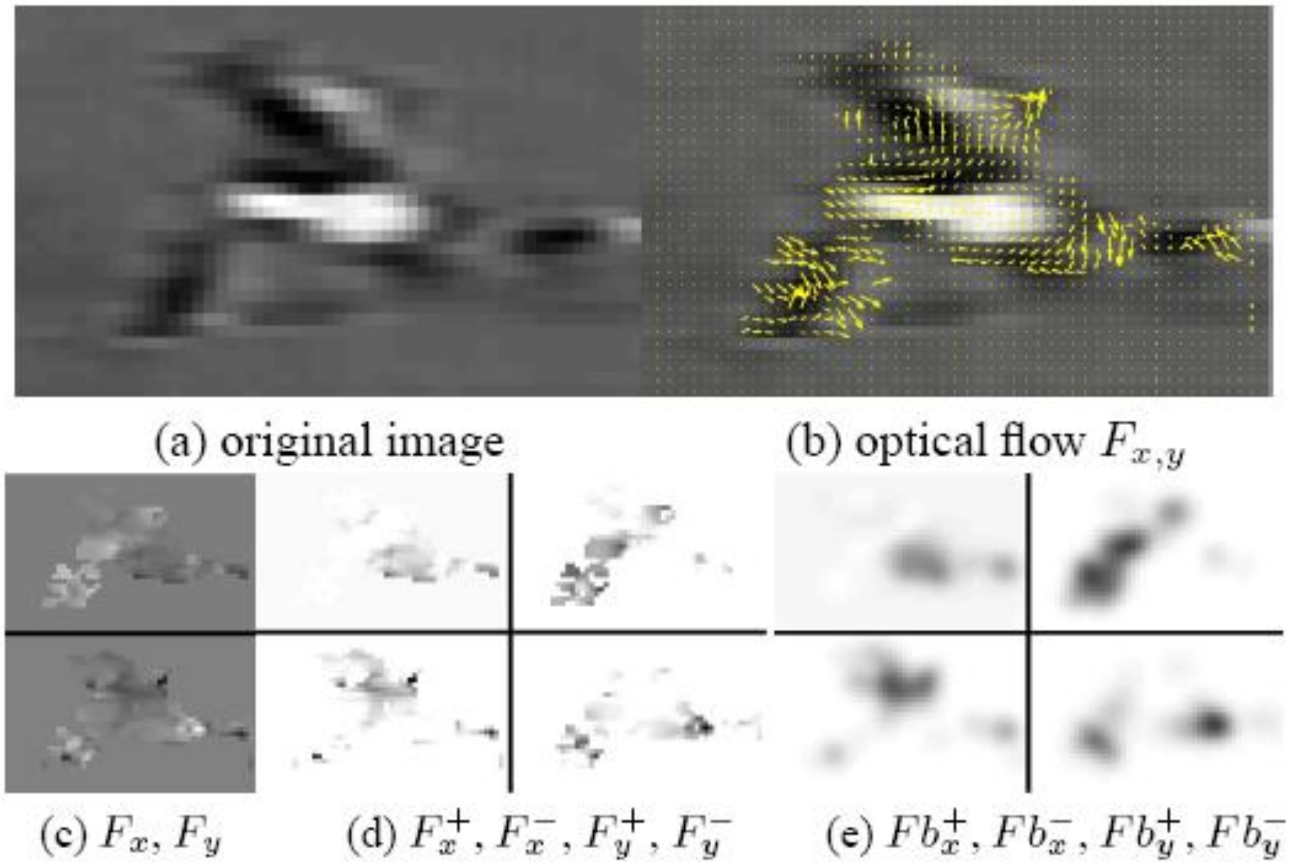
Space-Time Volumes



Blank et al. 2005

Representing Motion

Optical Flow with Split Channels



optical flow

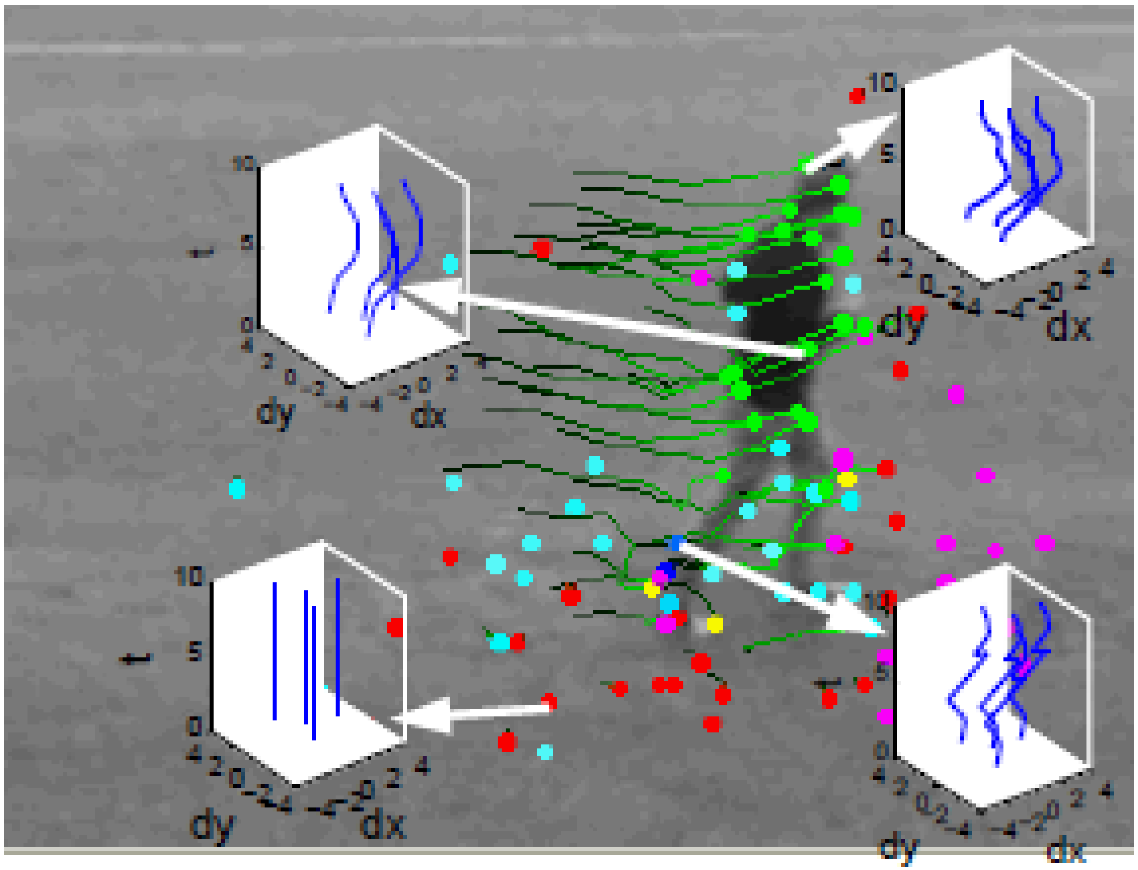
split into pos/neg channels

blurred pos/neg flow

[Efros et al. 2003]

Representing Motion

Tracked Points



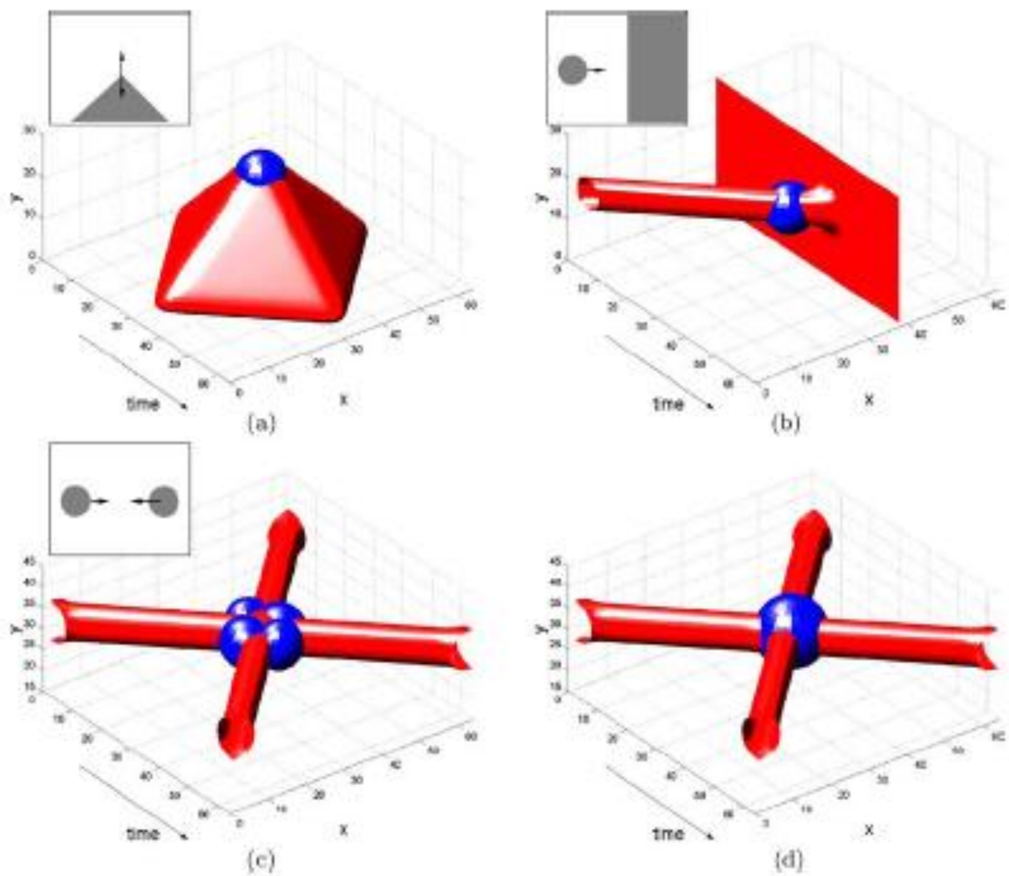
Matikainen et al. 2009

Representing Motion

Space-Time Interest Points

Moving corner

Ball hits wall



Balls collide

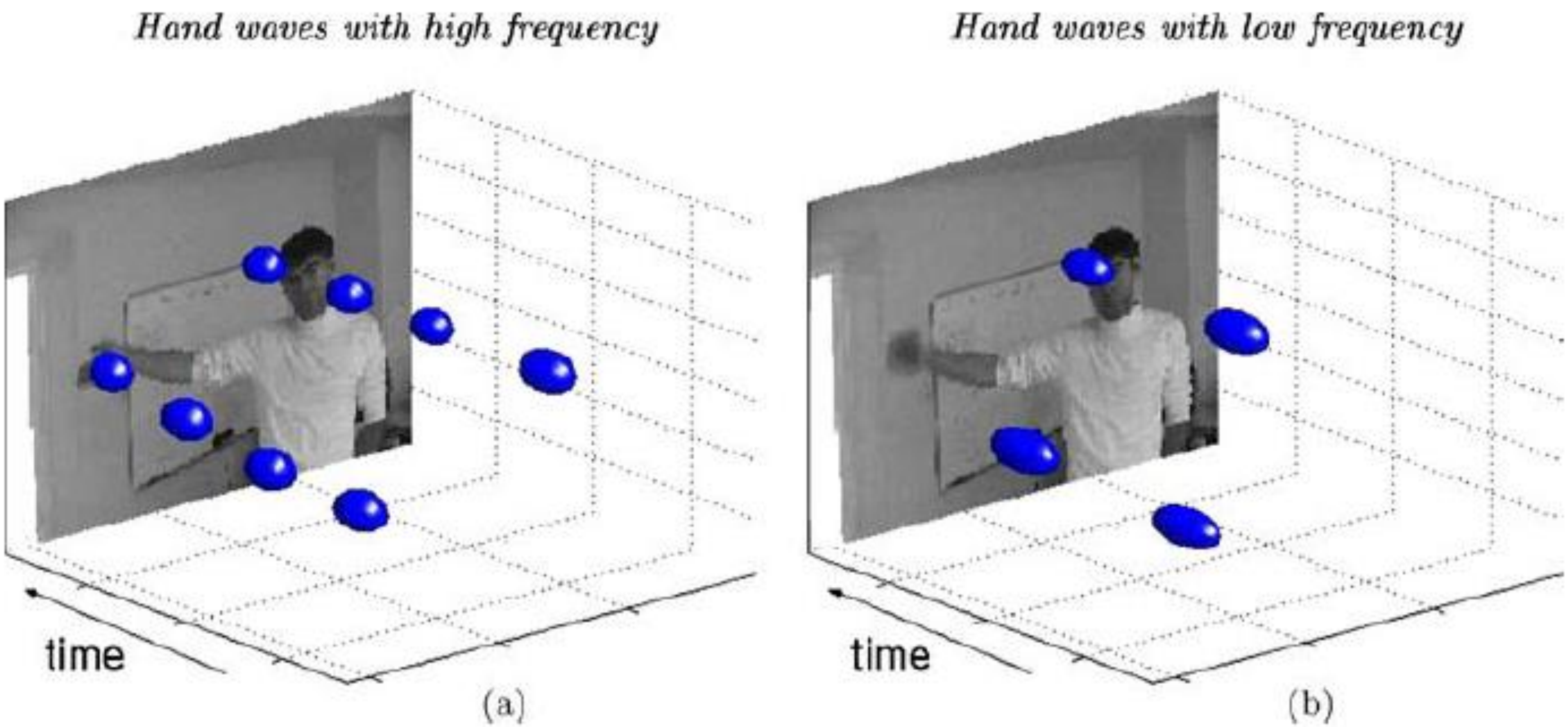
Balls collide (different scale)

Corner detectors in space-time

[Laptev 2005]

Representing Motion

Space-Time Interest Points



Laptev 2005

Examples of Action Recognition Systems

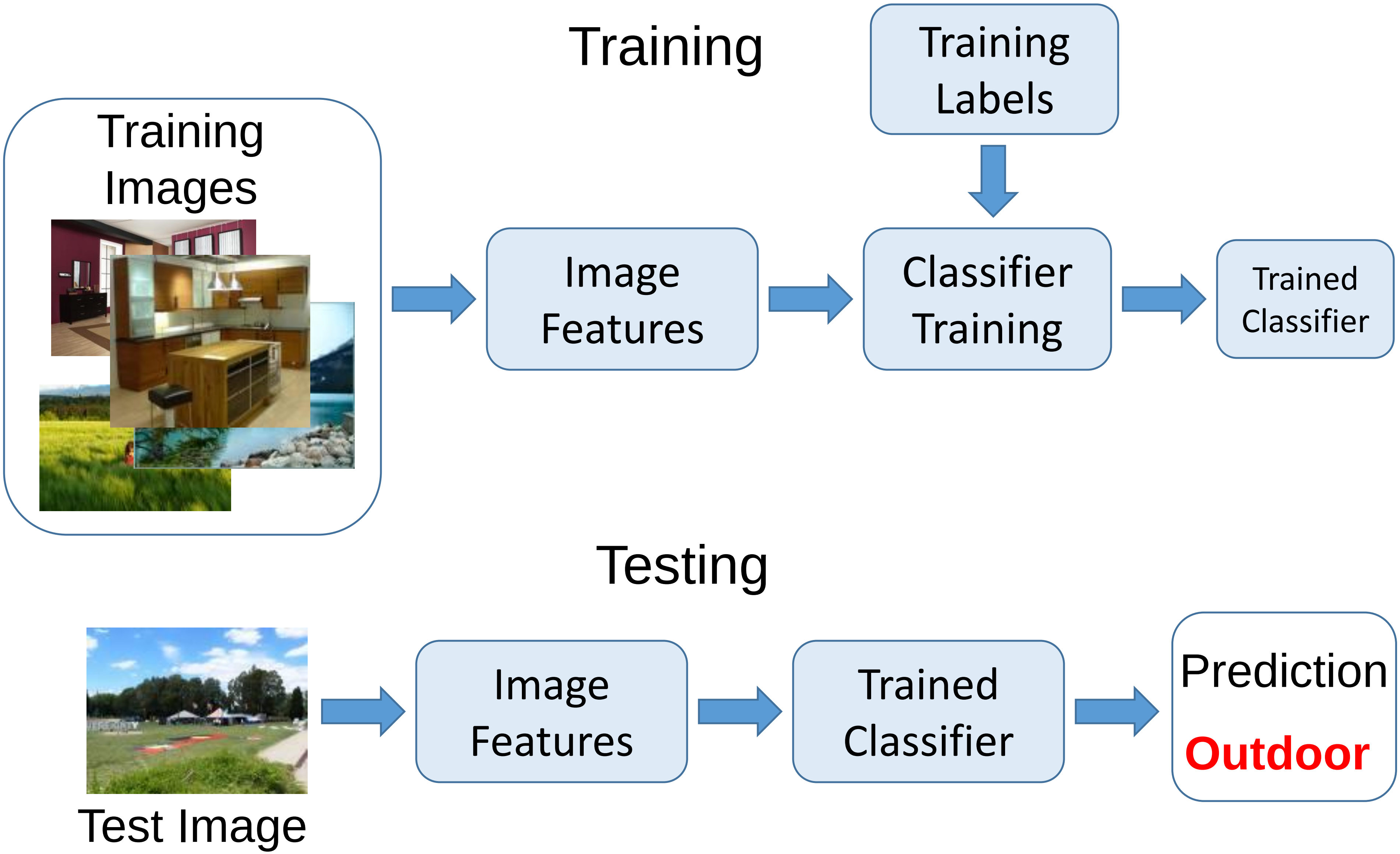
- Feature-based classification

- Recognition using pose and objects

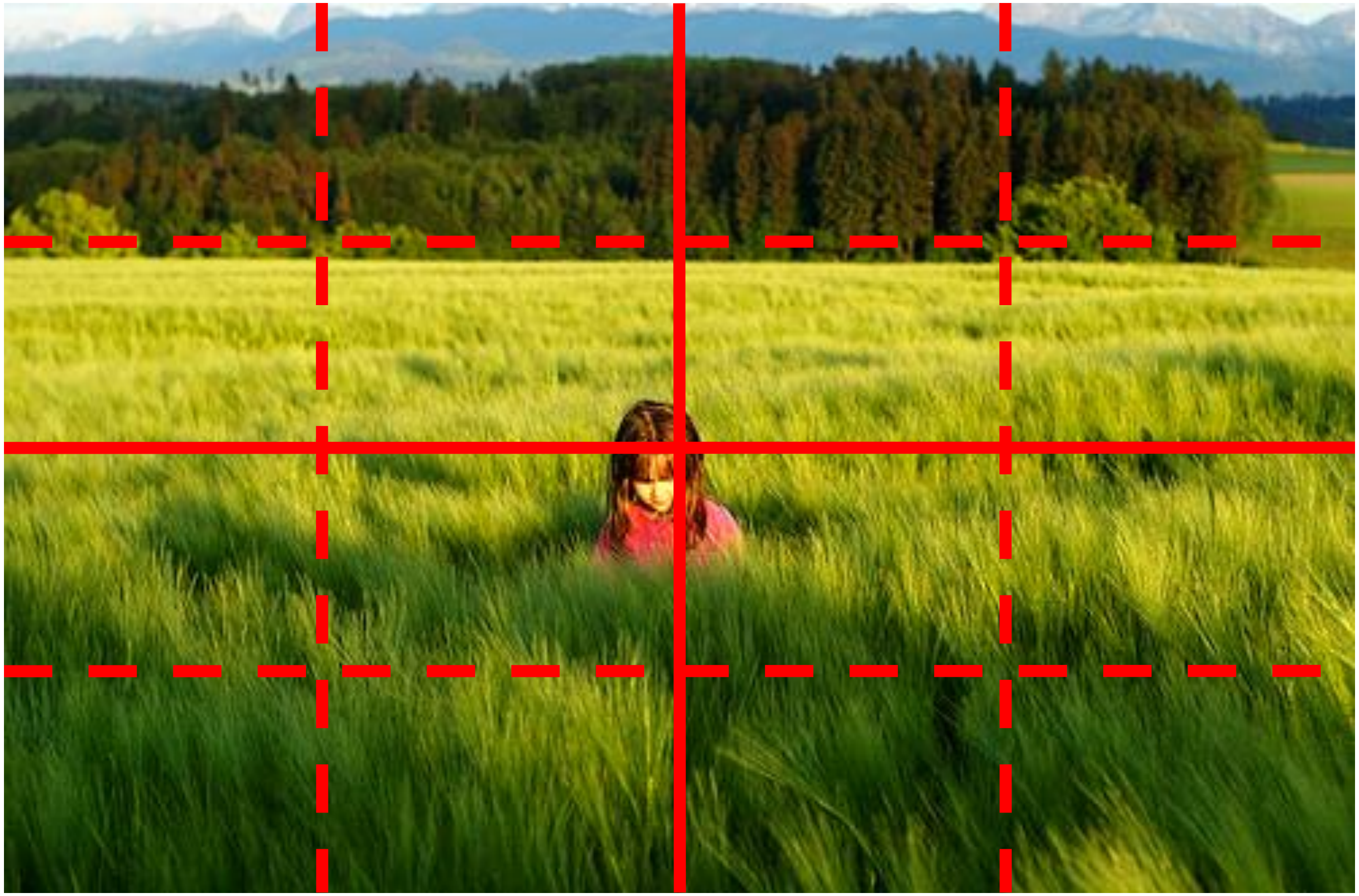
Action recognition as classification

Retrieving actions in movies, Laptev and Perez, 2007

Remember image categorization…



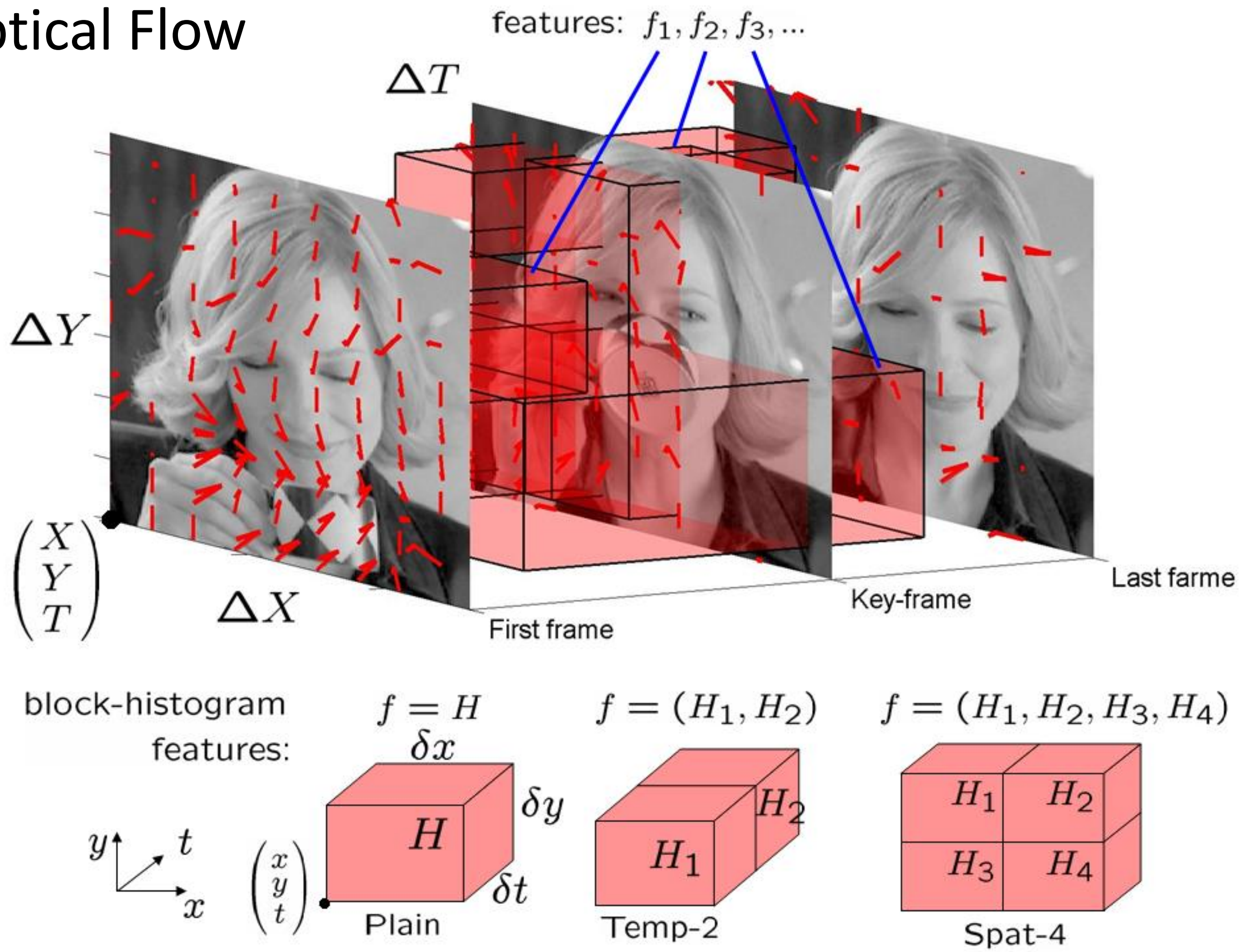
Remember spatial pyramids….



Compute histogram in each spatial bin

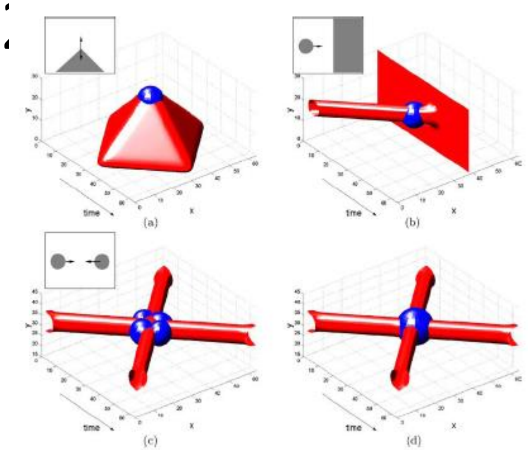
Features for Classifying Actions

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

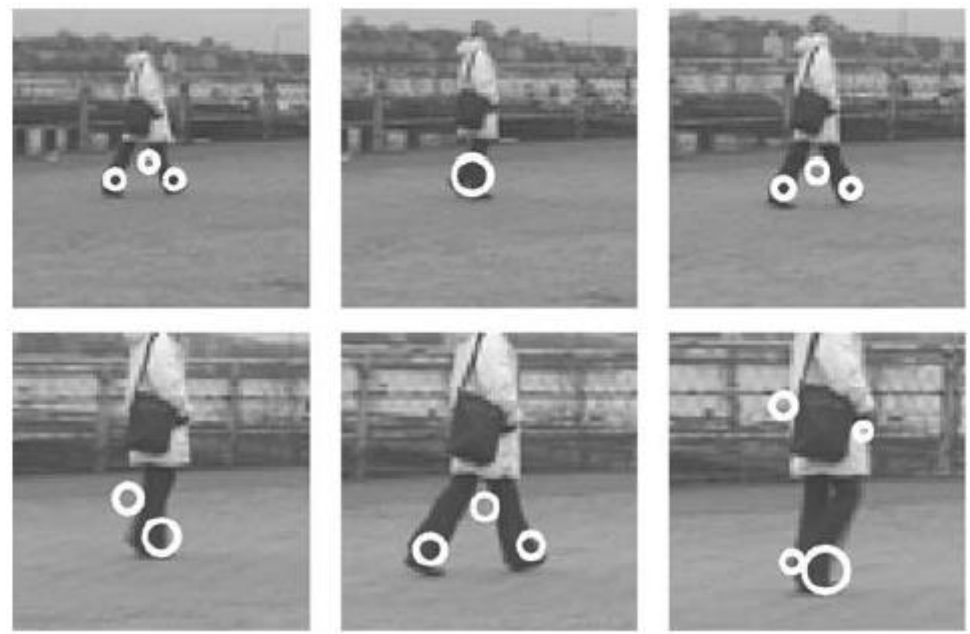


Features for Classifying Actions

l interest points



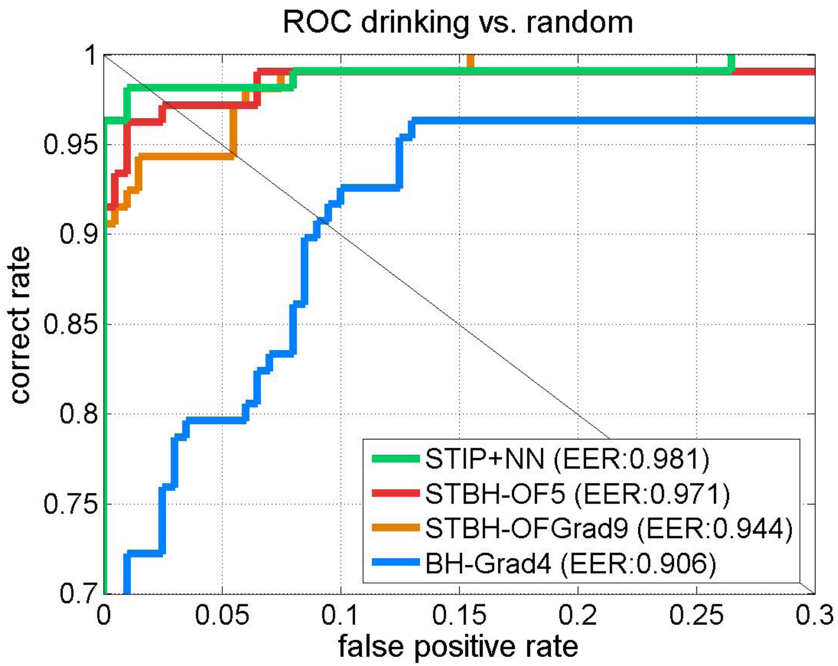
Corner detectors in space-time

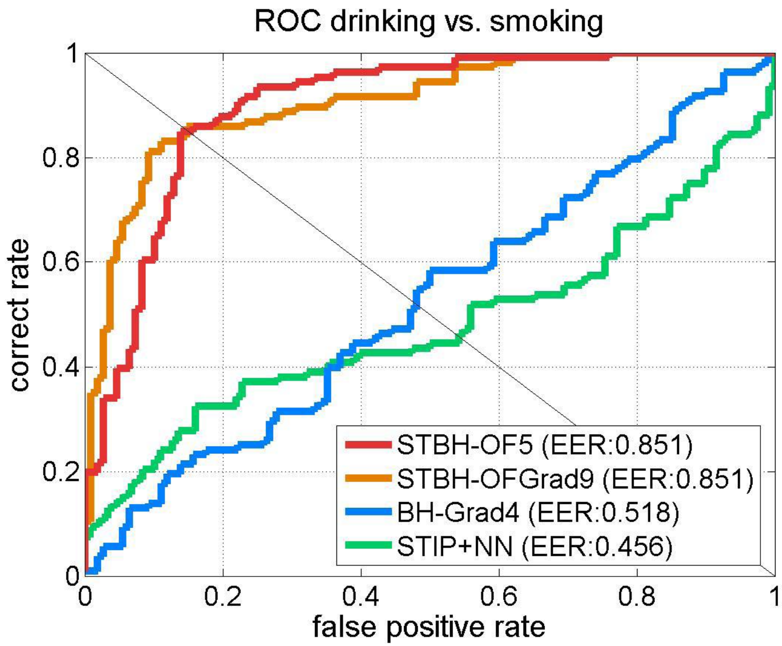


Descriptors based on Gaussian derivative filters over x, y, time

Classification

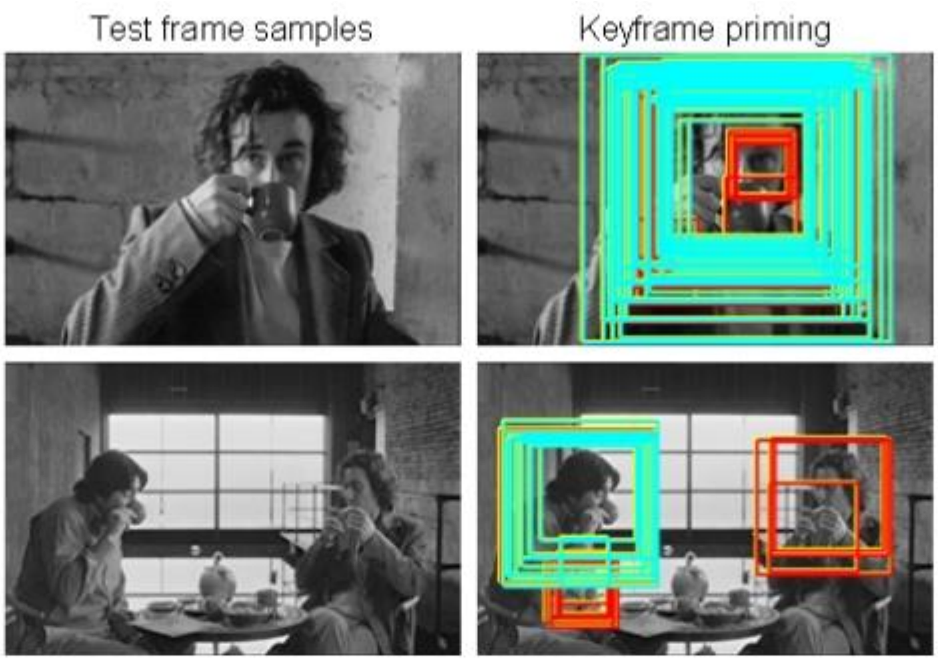
- Boosted stubs for pyramids of optical flow, gradient
- Nearest neighbor for STIP

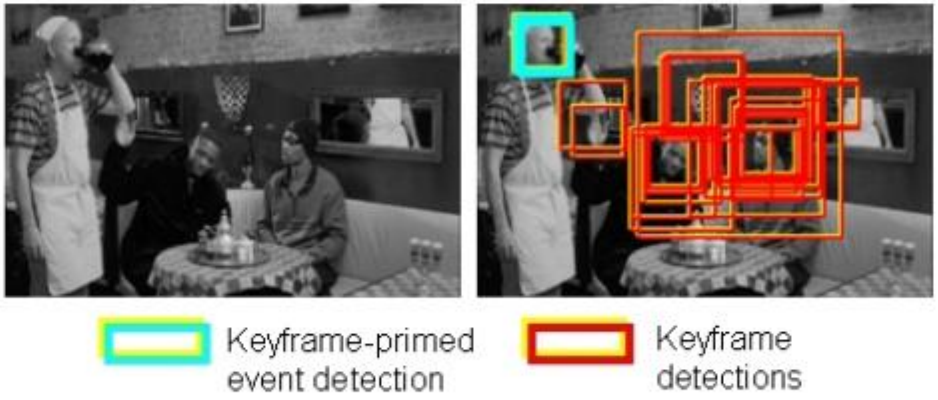




Searching the video for an action

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

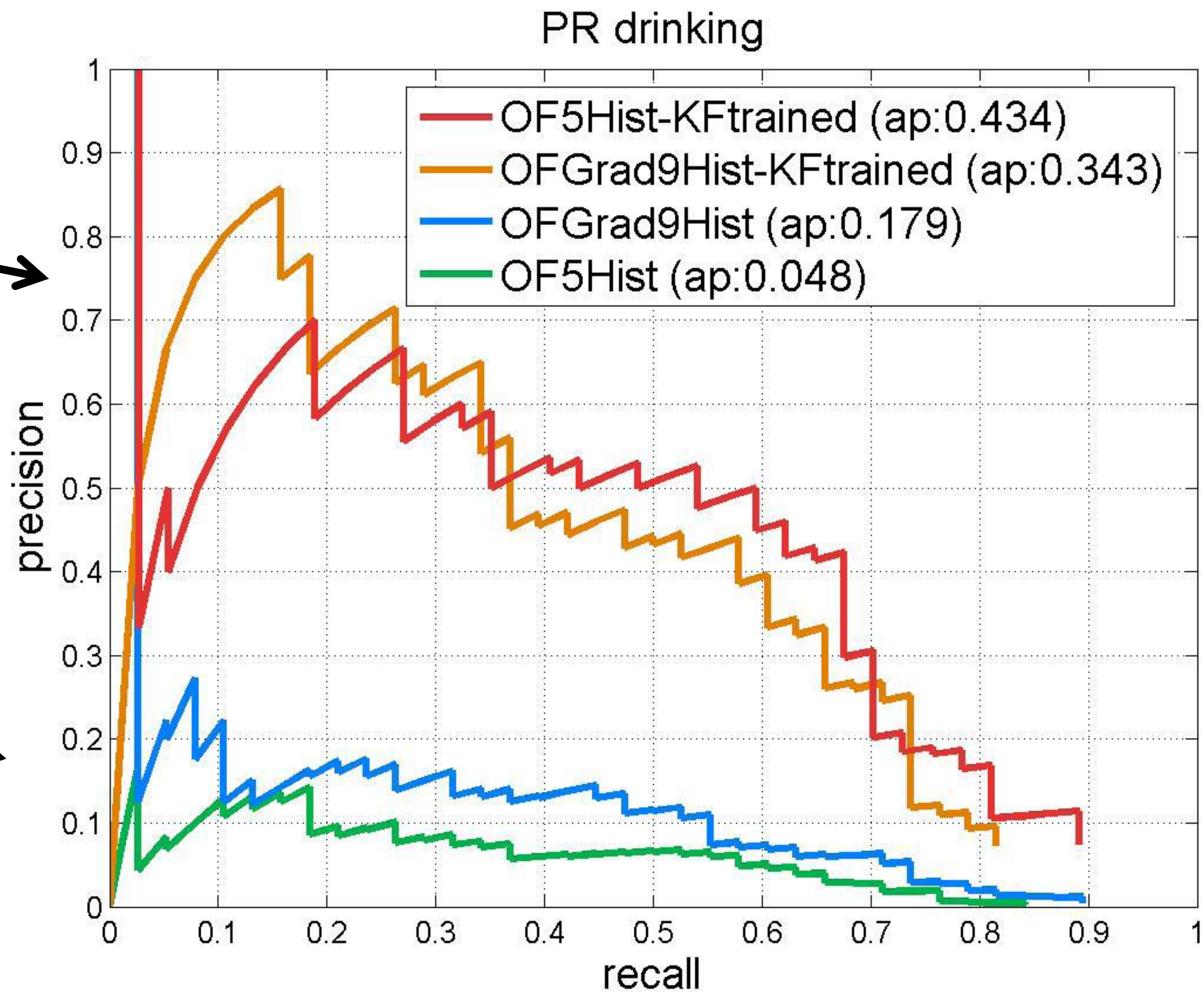




Accuracy in searching video

With keyframe detection

Without keyframe detection



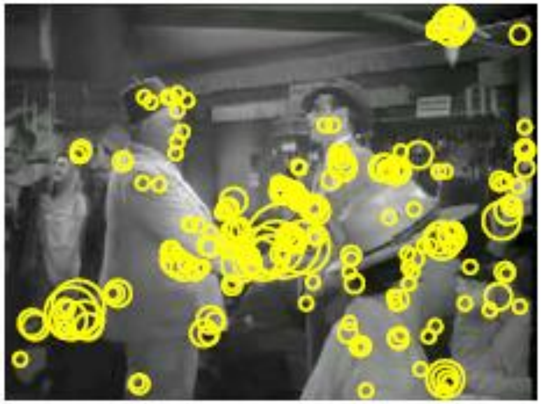
“Talk on phone”

“Get out of car”

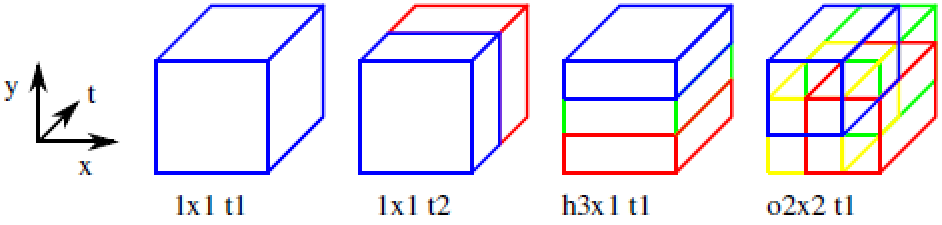
[Learning realistic human actions from movies, Laptev et al. 2008]

Approach

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



Interest Points

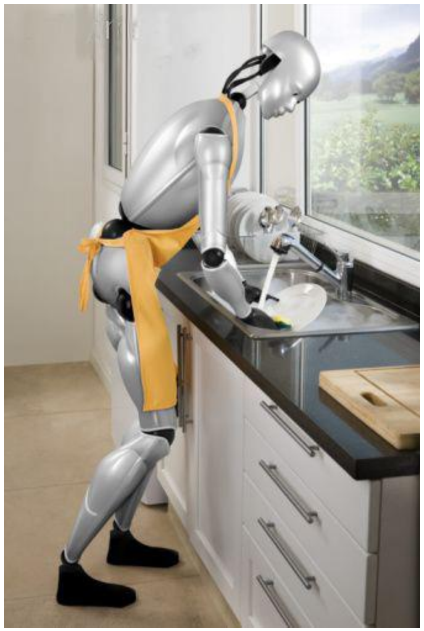


Spatio-Temporal Binning

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)

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**

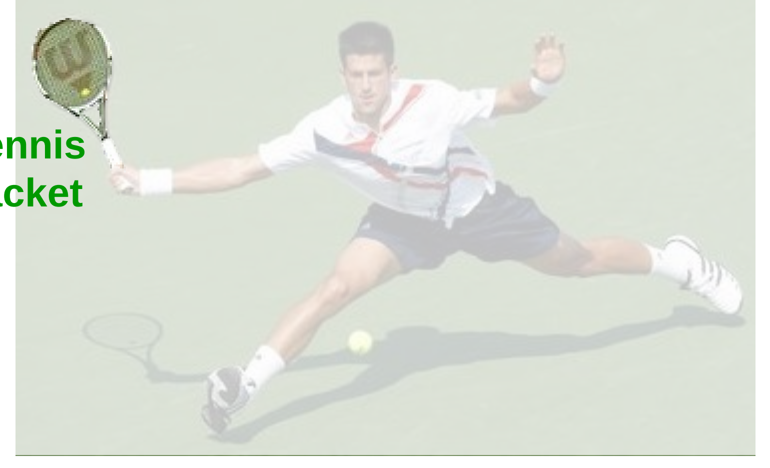
Slide Credit: Yao/Fei-Fei

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

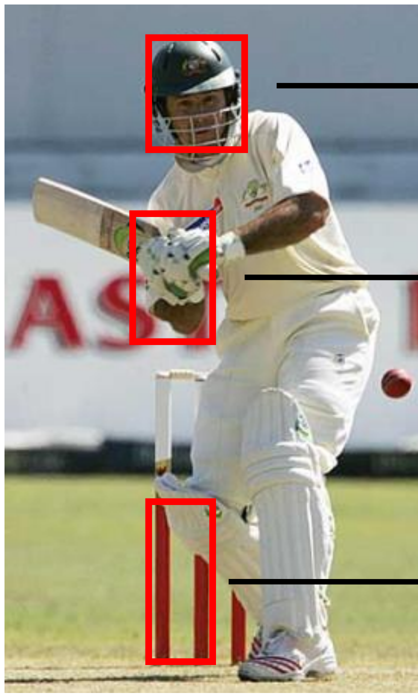
- **Human pose estimation**
- **Object detection**
- **Action categorization**

Activity: Tennis Forehand

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

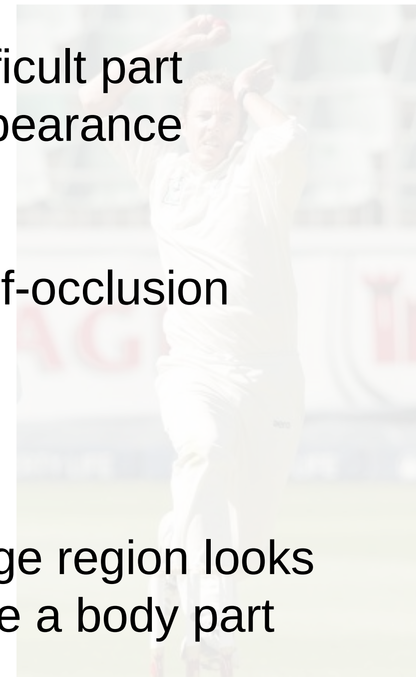
Human pose estimation is challenging.



Difficult part appearance

Self-occlusion

Image region looks like a body part

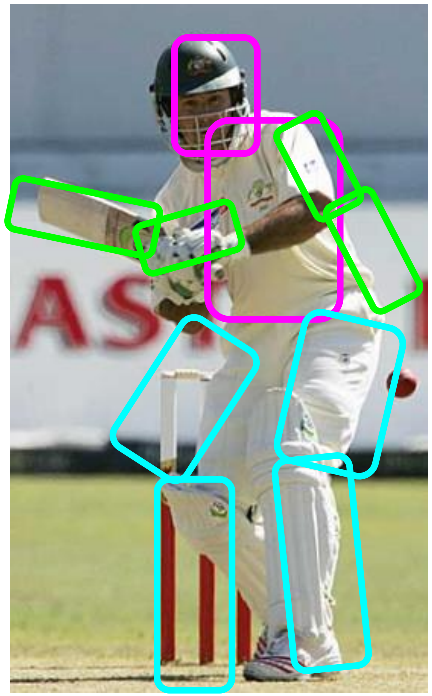


- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Human pose estimation is challenging.



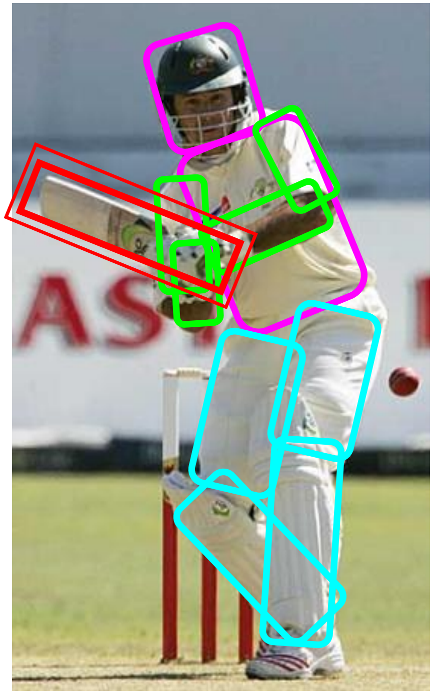
- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

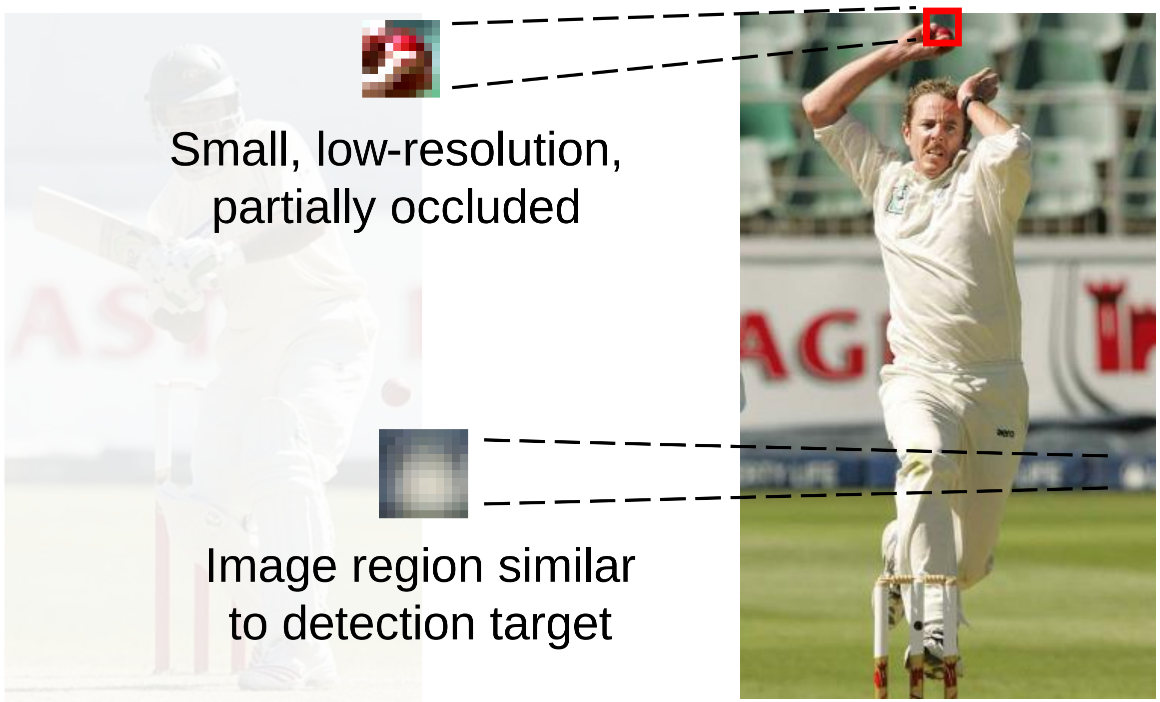
Facilitate

Given the object is detected.



Slide Credit: Yao/Fei-Fei

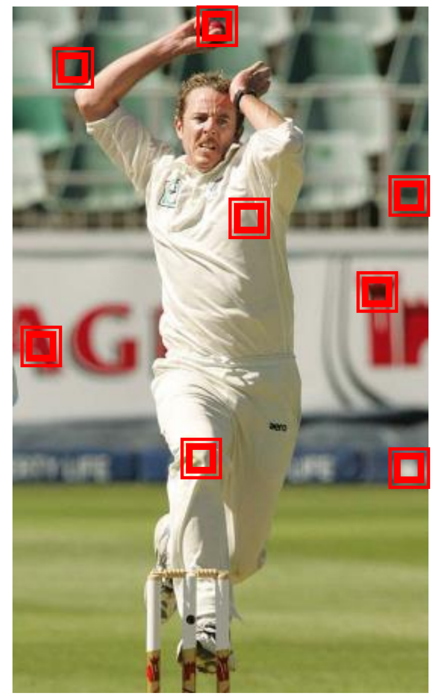
Human pose estimation & Object detection



Object detection is challenging

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

Human pose estimation & Object detection



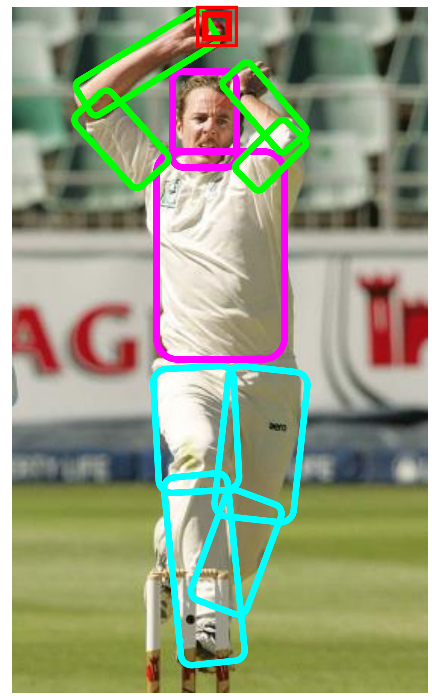
Object detection is challenging

- 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

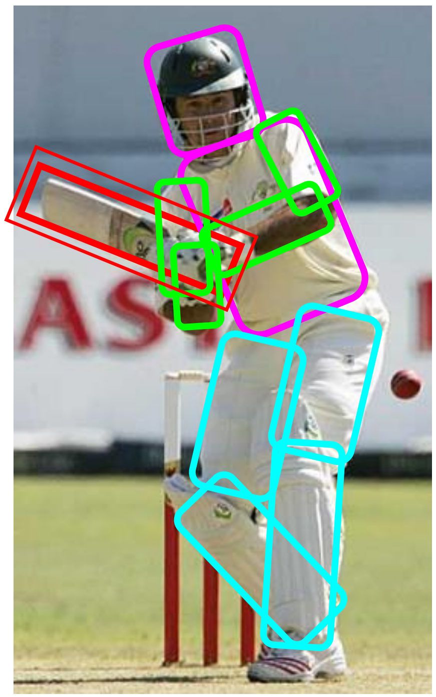


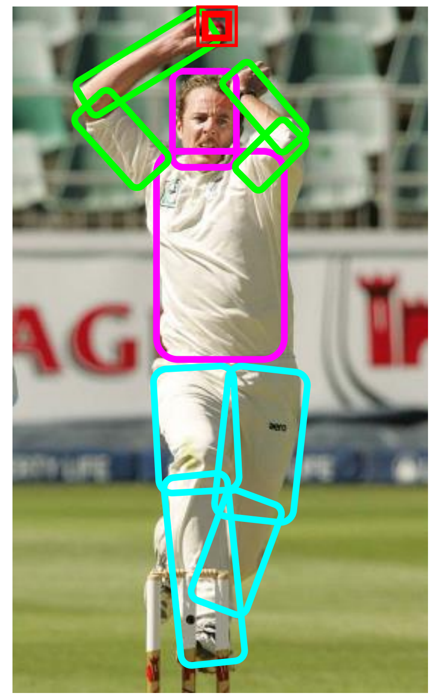
Given the pose is estimated.

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Mutual Context





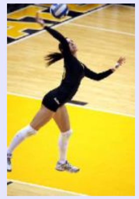
Slide Credit: Yao/Fei-Fei

Mutual Context Model Representation

A:


Tennis forehand


Croquet shot


Volleyball smash

. . .

O:

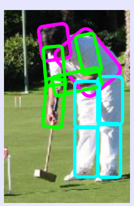
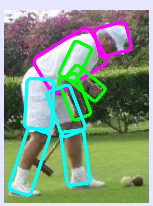
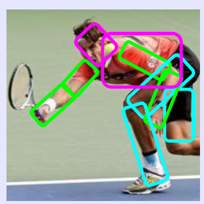
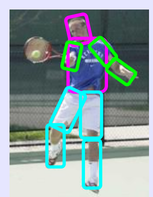

Tennis racket


Croquet mallet

Volleyball

. . .

H:

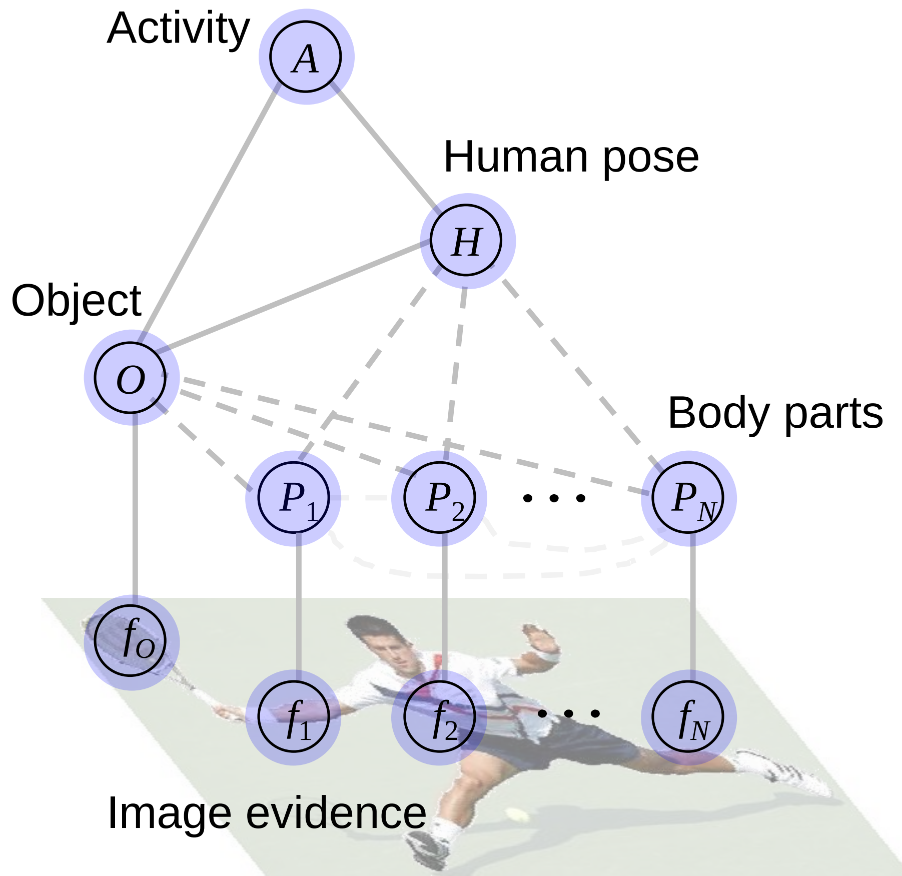





Intra-class variations

- 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

Learning Results

Cricket defensive shot

Cricket bowling

Croquet shot

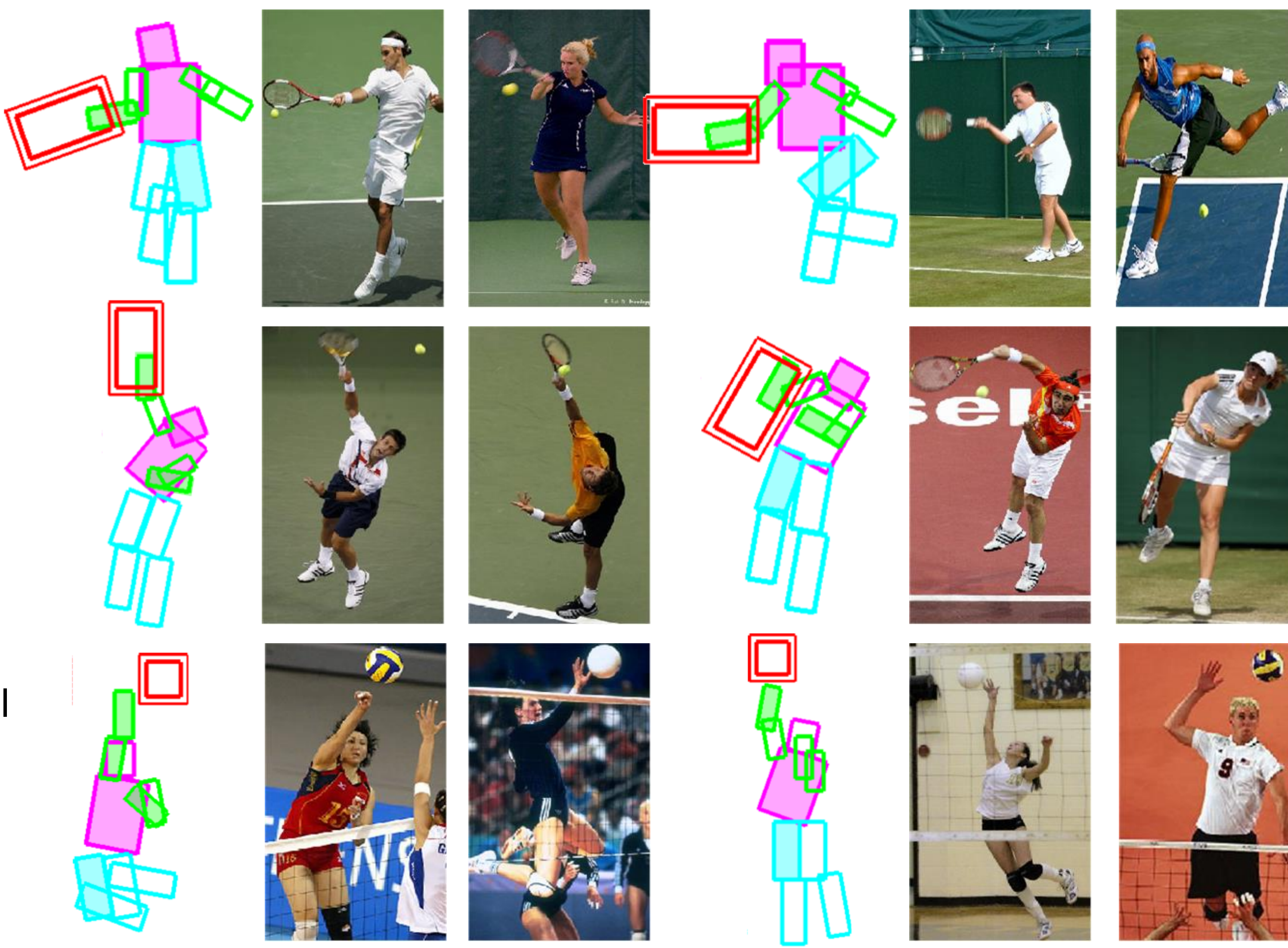
Slide Credit: Yao/Fei-Fei

Learning Results

Tennis forehand

Tennis serve

Volleyball smash

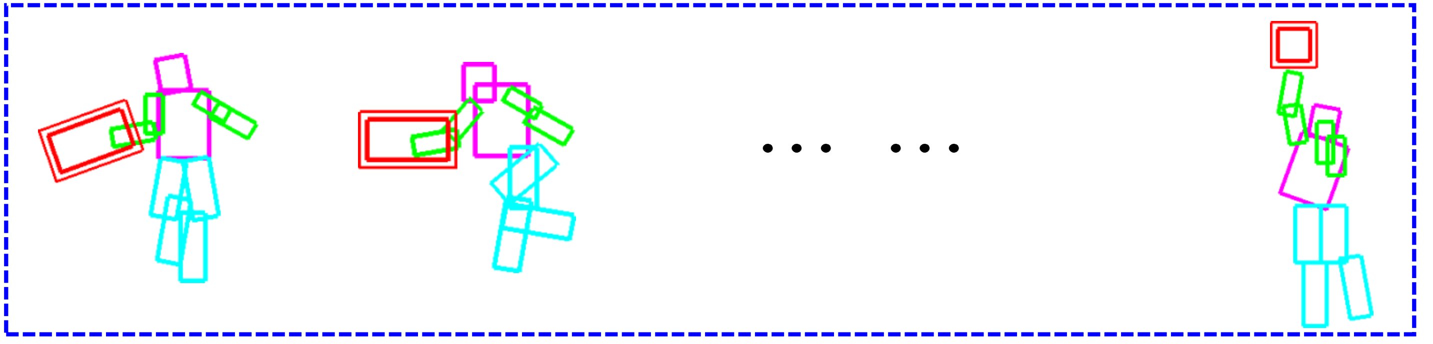


Slide Credit: Yao/Fei-Fei

Model Inference

I

The learned models

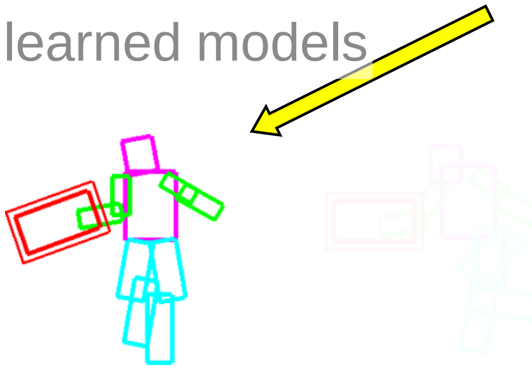


Slide Credit: Yao/Fei-Fei

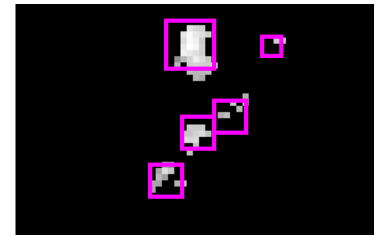
Model Inference

I

The learned models



. . .   . . .



Head detection

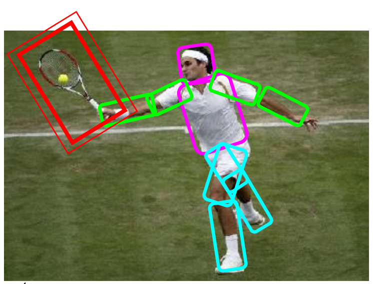
Torso detection

⋮

Tennis racket detection

Compositional Inference

[Chen et al, 2007]



$$\Psi\left(A_1, H_1, O_1^*, \left\{P_{1,n}^*\right\}_n\right)$$

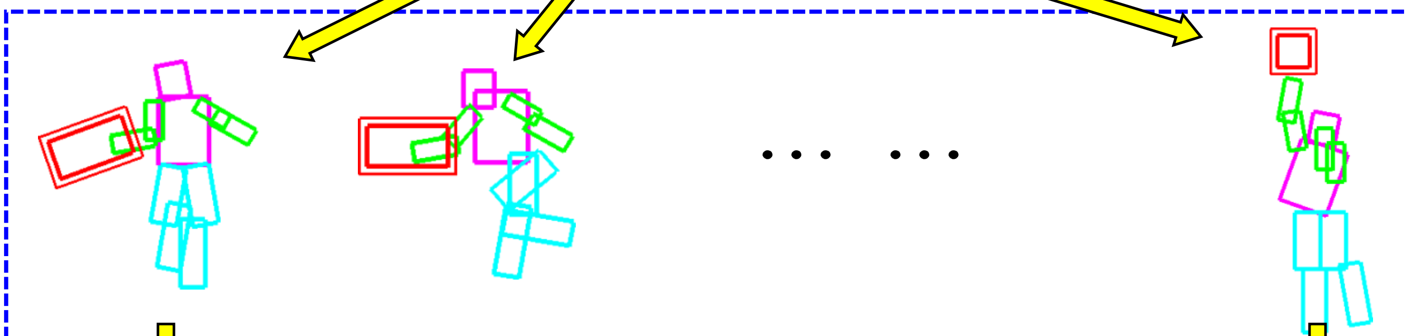
Layout of the **object** and **body parts**.

Slide Credit: Yao/Fei-Fei

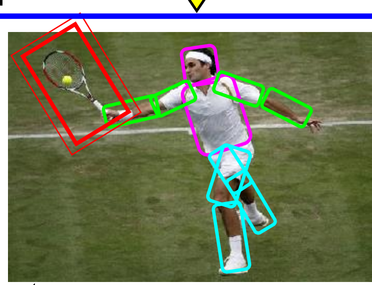
Model Inference

I

The learned models

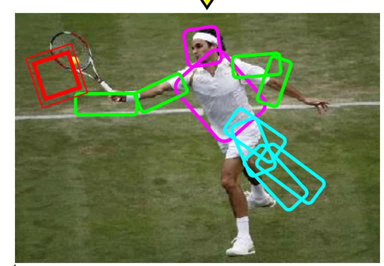


Output



$$\Psi\left(A_1, H_1, O_1^*, \left\{P_{1,n}^*\right\}_n\right)$$

… …



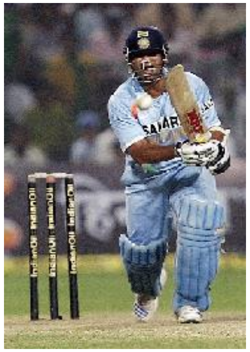
$$\Psi\left(A_K, H_K, O_K^*, \left\{P_{K,n}^*\right\}_n\right)$$

Slide Credit: Yao/Fei-Fei

Dataset and Experiment Setup

Sport data set: 6 classes

180 training (supervised with object and part locations) & 120 testing images



Cricket defensive shot

Cricket bowling

Croquet shot

Tennis forehand

Tennis serve

Volleyball smash

[Gupta et al, 2009]

Tasks:

- Object detection;
- Pose estimation;
- Activity classification.

Slide Credit: Yao/Fei-Fei

Dataset and Experiment Setup

Sport data set: 6 classes

180 training (supervised with object and part locations) & 120 testing images

Cricket defensive shot

Cricket bowling

Croquet shot

Tennis forehand

Tennis serve

Volleyball smash

[Gupta et al, 2009]

Tasks:

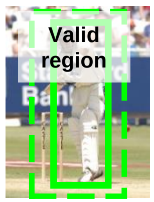
- **Object detection;**
- Pose estimation;
- Activity classification.

Slide Credit: Yao/Fei-Fei

Object Detection Results

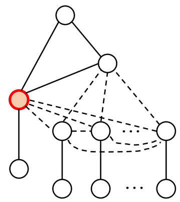
Sliding window

[Andriluka et al, 2009]



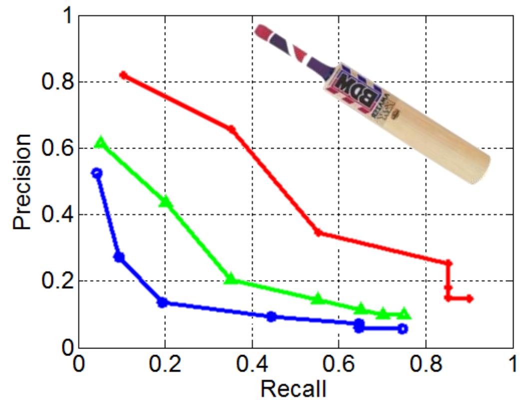
Pedestrian context

[Dalal & Triggs, 2006]

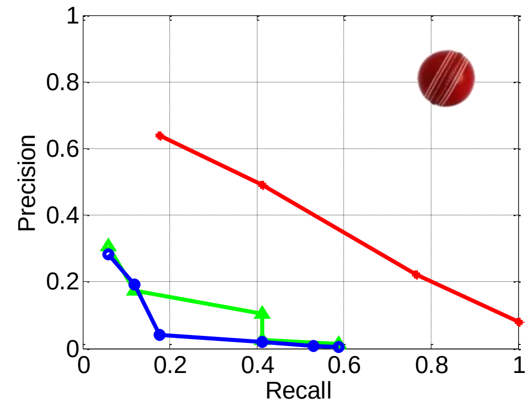


Our Method

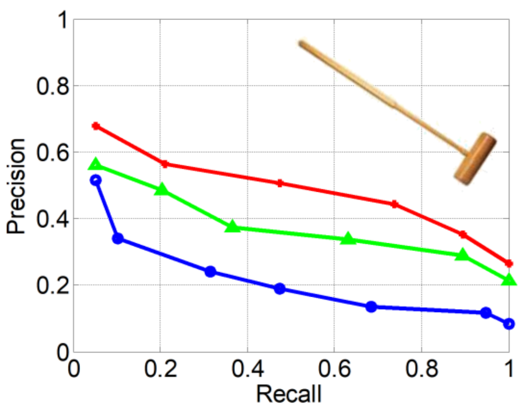
Cricket bat



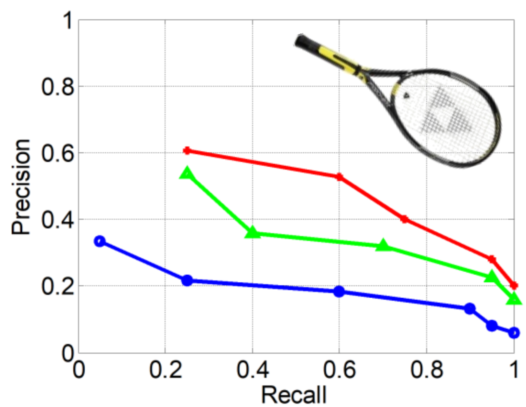
Cricket ball



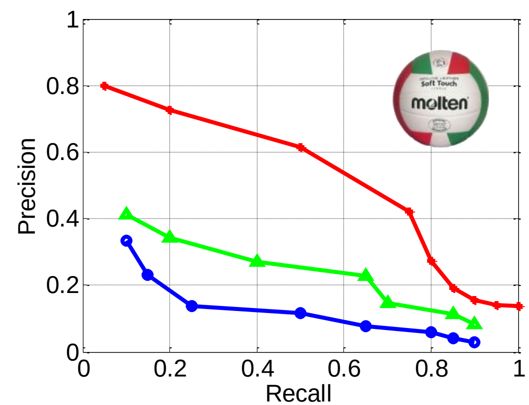
Croquet mallet



Tennis racket



Volleyball



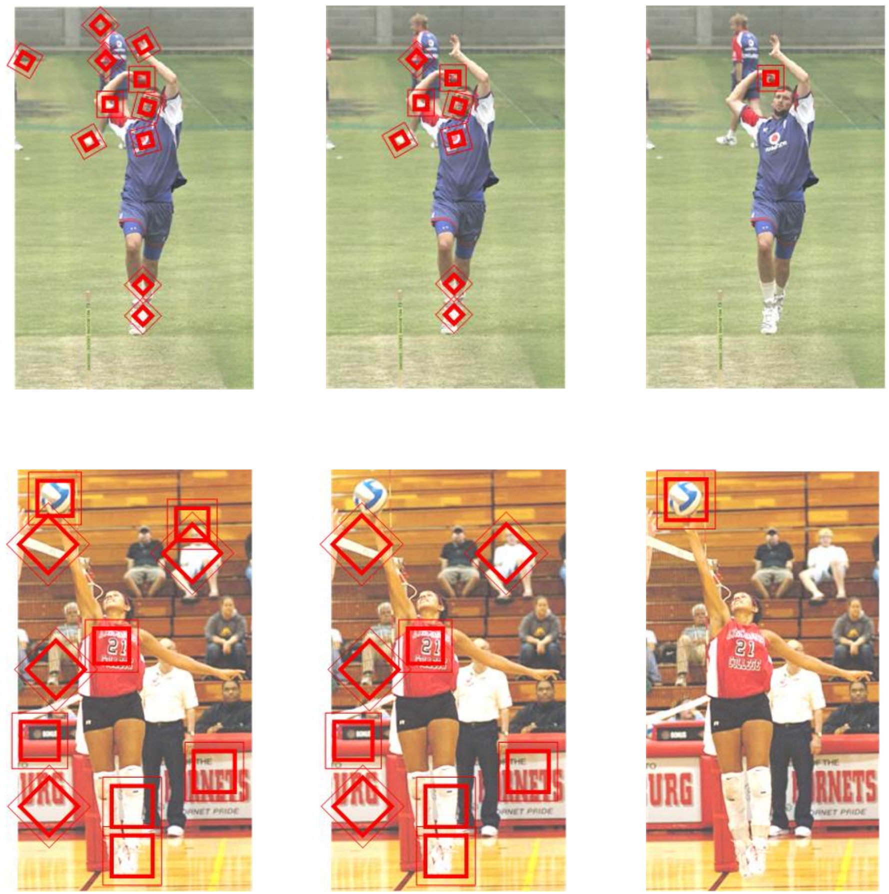
Slide Credit: Yao/Fei-Fei

Object Detection Results

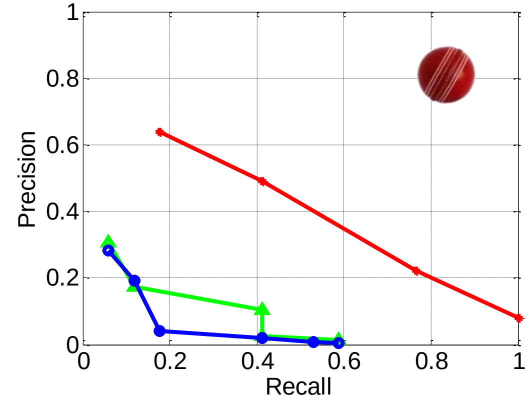
Sliding window | Pedestrian context | Our method

Small object

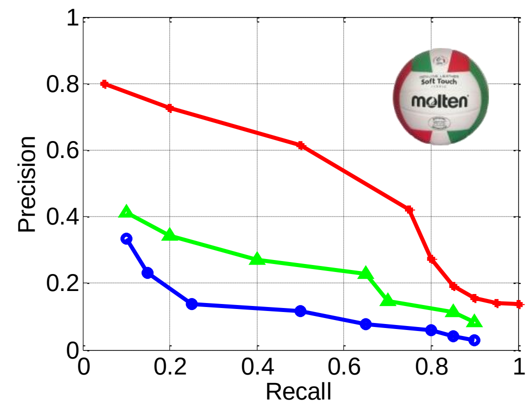
Background clutter



Cricket ball



Volleyball

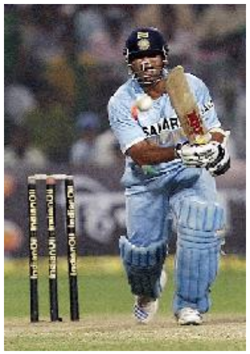


Slide Credit: Yao/Fei-Fei

Dataset and Experiment Setup

Sport data set: 6 classes

180 training & 120 testing images



Cricket defensive shot

Cricket bowling

Croquet shot

Tennis forehand

Tennis serve

Volleyball smash

[Gupta et al, 2009]

Tasks:

- Object detection;
- **Pose estimation;**
- Activity classification.

Slide Credit: Yao/Fei-Fei

Human Pose Estimation Results

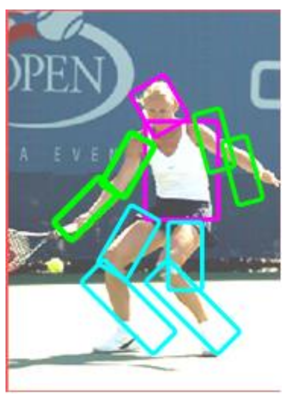
Method	Torso	Upper Leg		Lower Leg		Upper Arm		Lower Arm		Head
Ramanan, 2006	.52	.22	.22	.21	.28	.24	.28	.17	.14	.42
Andriluka et al, 2009	.50	.31	.30	.31	.27	.18	.19	.11	.11	.45
Our full model	**.66**	**.43**	**.39**	**.44**	**.34**	**.44**	**.40**	**.27**	**.29**	**.58**

Slide Credit: Yao/Fei-Fei

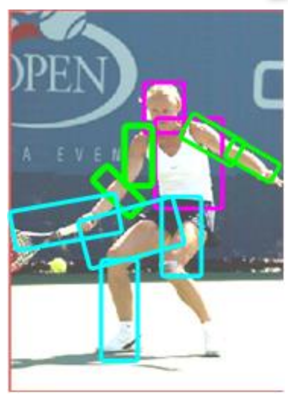
Human Pose Estimation Results

Method	Torso	Upper Leg		Lower Leg		Upper Arm		Lower Arm		Head
Ramanan, 2006	.52	.22	.22	.21	.28	.24	.28	.17	.14	.42
Andriluka et al, 2009	.50	.31	.30	.31	.27	.18	.19	.11	.11	.45
Our full model	**.66**	**.43**	**.39**	**.44**	**.34**	**.44**	**.40**	**.27**	**.29**	**.58**

Tennis serve model

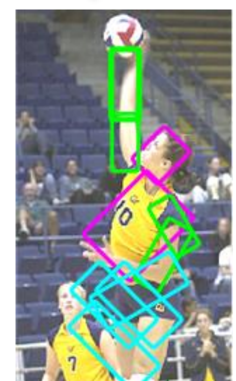


Our estimation result

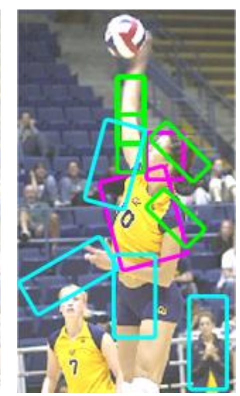


Andriluka et al, 2009

Volleyball smash model



Our estimation result

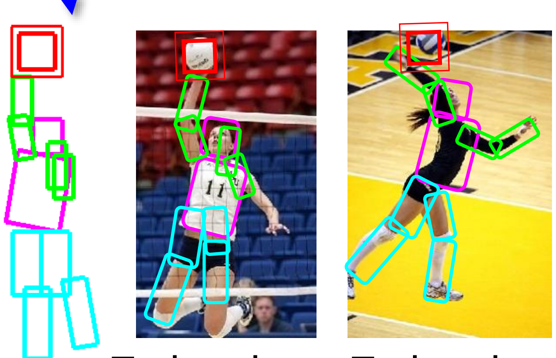


Andriluka et al, 2009

Slide Credit: Yao/Fei-Fei

Human Pose Estimation Results

Method	Torso	Upper Leg		Lower Leg		Upper Arm		Lower Arm		Head
Ramanan, 2006	.52	.22	.22	.21	.28	.24	.28	.17	.14	.42
Andriluka et al, 2009	.50	.31	.30	.31	.27	.18	.19	.11	.11	.45
Our full model	**.66**	**.43**	**.39**	**.44**	**.34**	**.44**	**.40**	**.27**	**.29**	**.58**
One pose per class	.63	.40	.36	.41	.31	.38	.35	.21	.23	.52



Estimation result

Estimation result

Estimation result

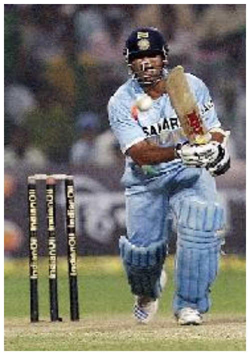
Estimation result

Slide Credit: Yao/Fei-Fei

Dataset and Experiment Setup

Sport data set: 6 classes

180 training & 120 testing images



Cricket defensive shot

Cricket bowling

Croquet shot

Tennis forehand

Tennis serve

Volleyball smash

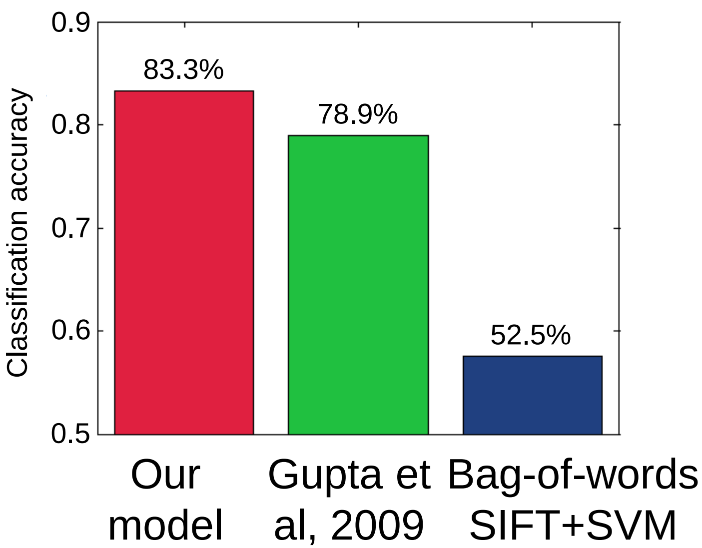
[Gupta et al, 2009]

<u>Tasks:</u>

- Object detection;
- Pose estimation;
- **Activity classification.**

Slide Credit: Yao/Fei-Fei

Activity Classification Results

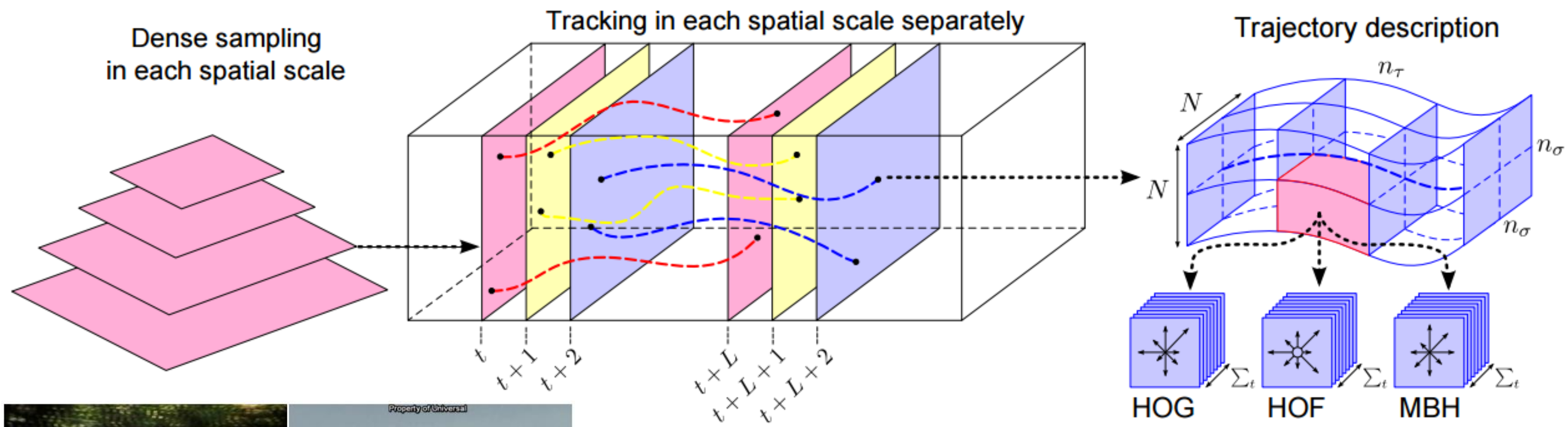


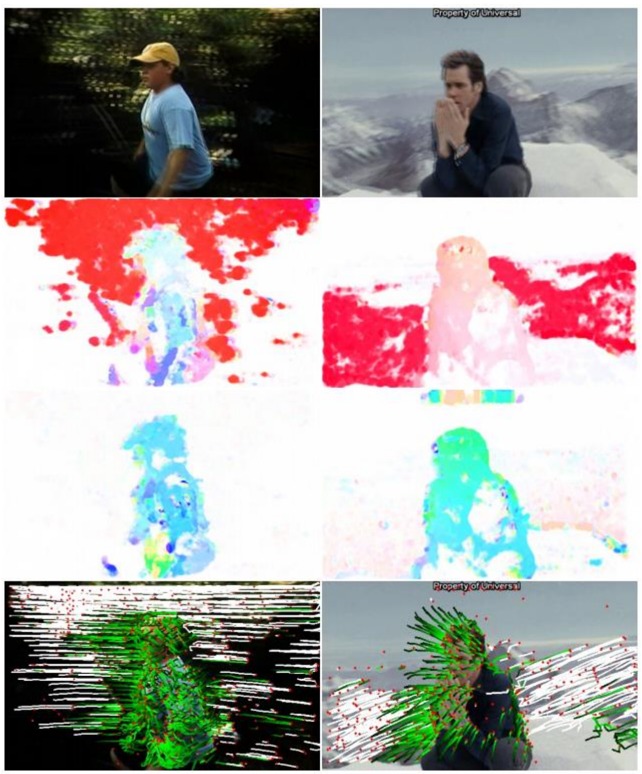
Cricket shot

Tennis forehand

Slide Credit: Yao/Fei-Fei

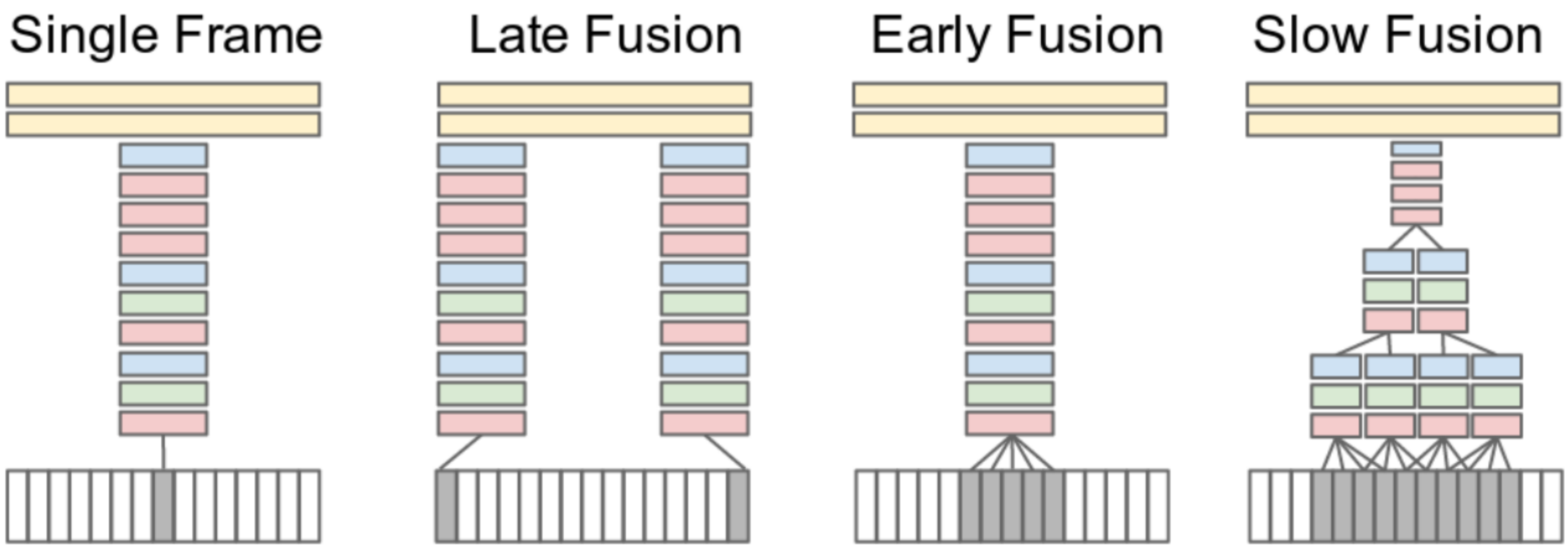
Motion features – Dense Trajectory

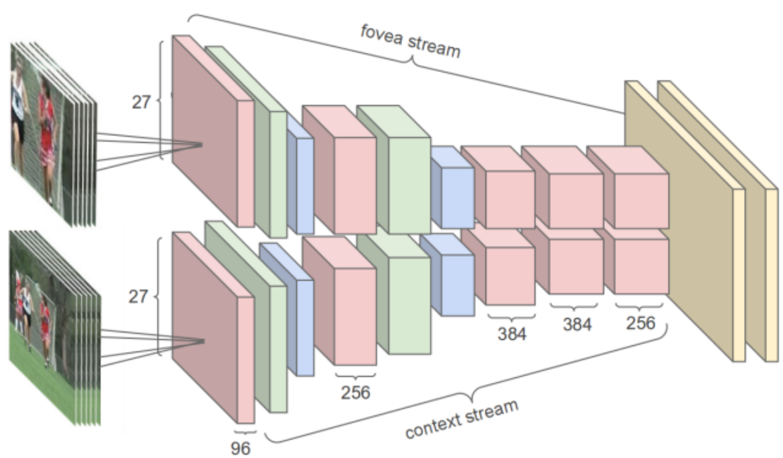




Action Recognition by Dense Trajectories, CVPR 2011
Action Recognition with Improved Trajectories, ICCV 2013

Video classification with CNNs



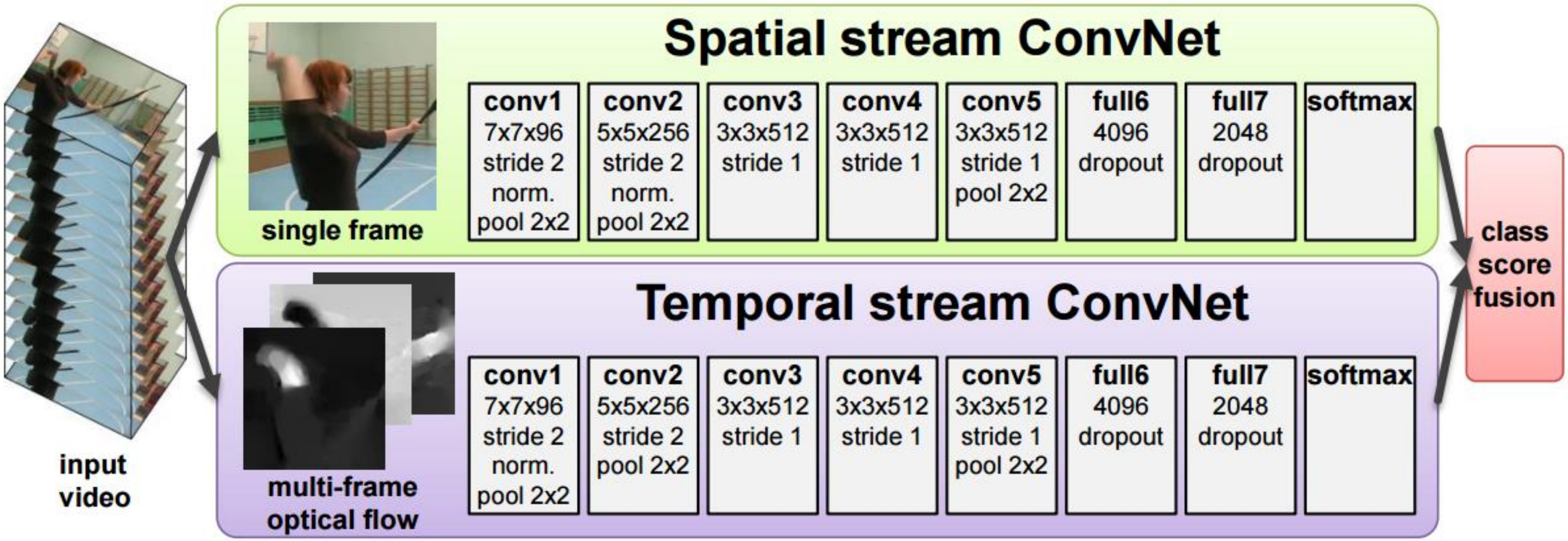


Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

Video classification with CNNs

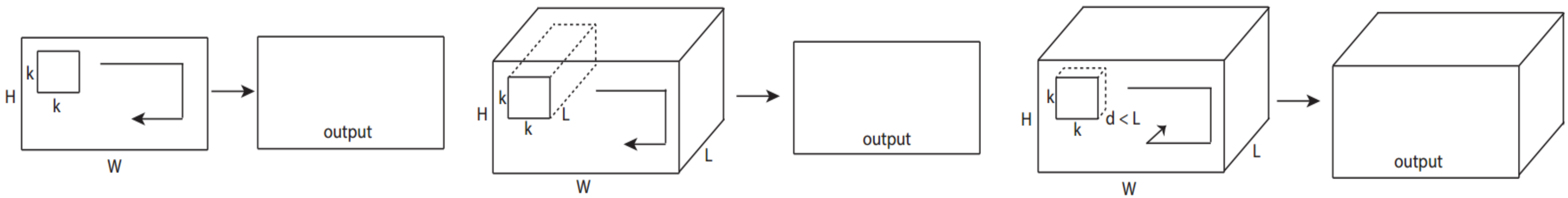
Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

Two-stream CNN



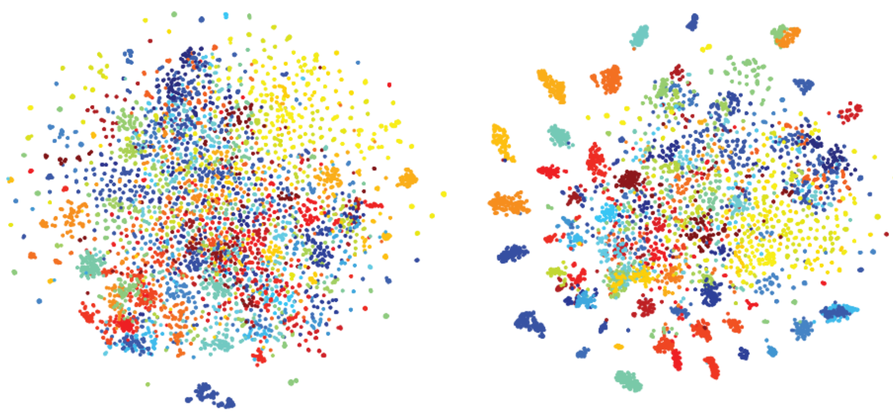
Two-Stream Convolutional Networks for Action Recognition in Videos, NIPS 2014

3D Convolutional Networks



(a) 2D convolution (b) 2D convolution on multiple frames (c) 3D convolution

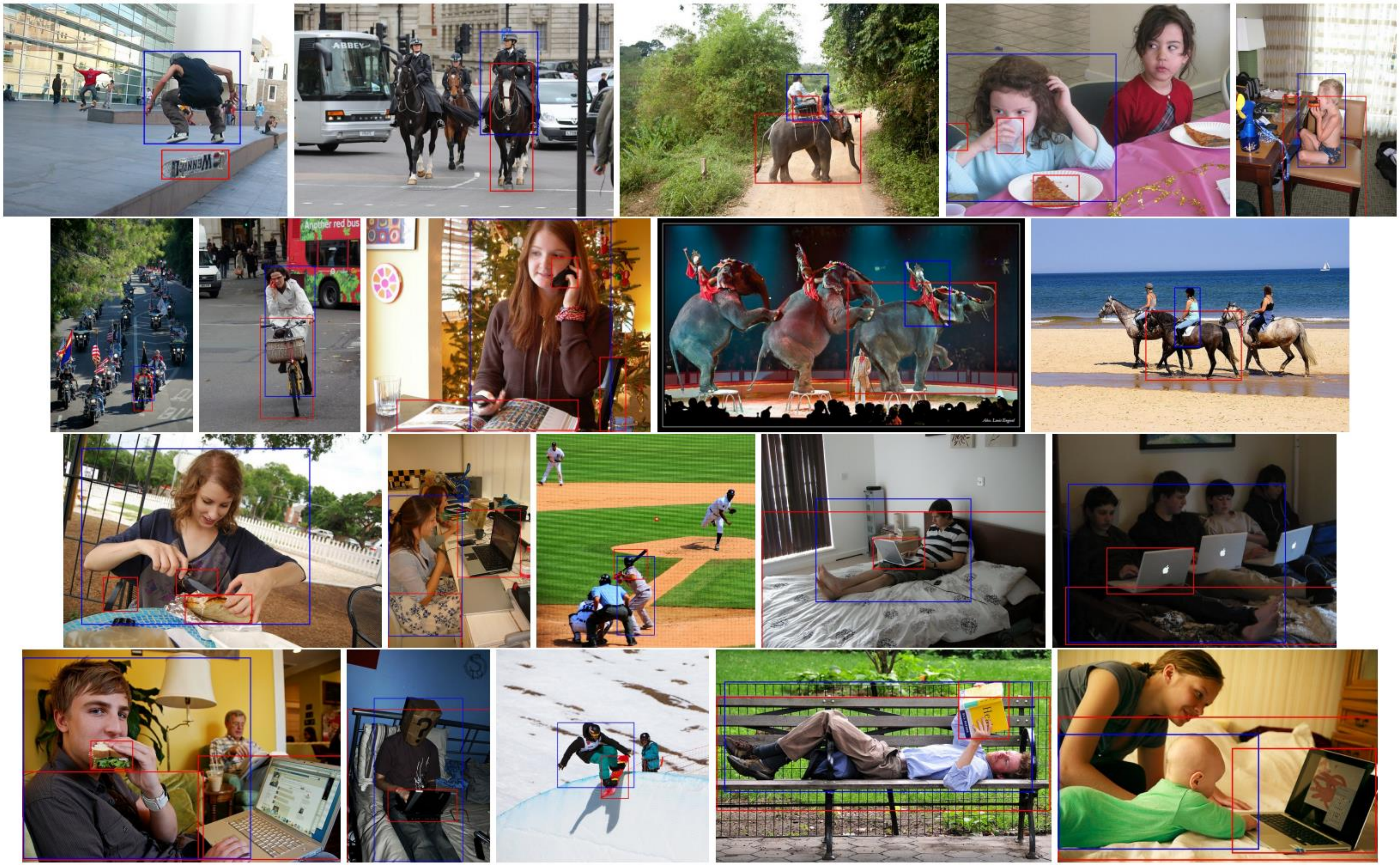
Conv1a 64	Pool1	Conv2a 128	Pool2	Conv3a 256	Conv3b 256	Pool3	Conv4a 512	Conv4b 512	Pool4	Conv5a 512	Conv5b 512	Pool5	fc6 4096	fc7 4096	softmax



Imagenet C3D

Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015

Action recognition -> Semantic role Labeling




ride
motorcycle
drink
cup
talk_on_phone
cell phone
hit
tennis racket
read
book
umbrella
carry
work_on_computer
laptop
throw
sports ball
ride
bicycle
lay
bed
ride
elephant
skateboarding
skateboard
surf
surfboard
cut
knife
eat
pizza
catch
frisbee
hold
cat
kick
sports ball
sports ball
hit
ski
skis
kick
sports ball
hit
sports ball
carry
surfboard
drink
wine glass
hold
wine glass
sit
chair
work_on_computer
laptop
hold
laptop
sit
chair

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