Action Recognition

Computer Vision

CS 543 / ECE 549

University of Illinois

Derek Hoiem

Last classes

- Parts-based/articulated object models

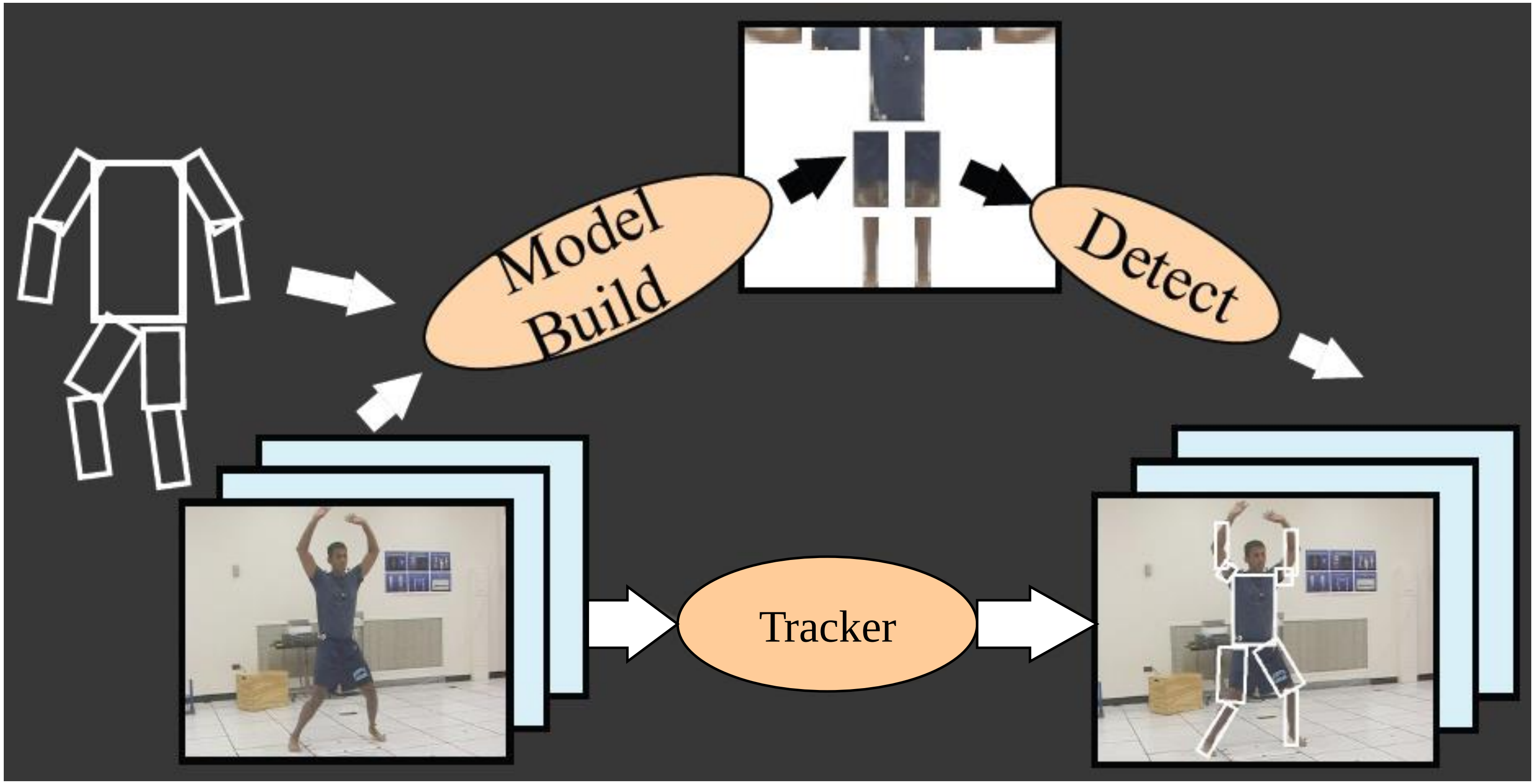
- Tracking objects

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



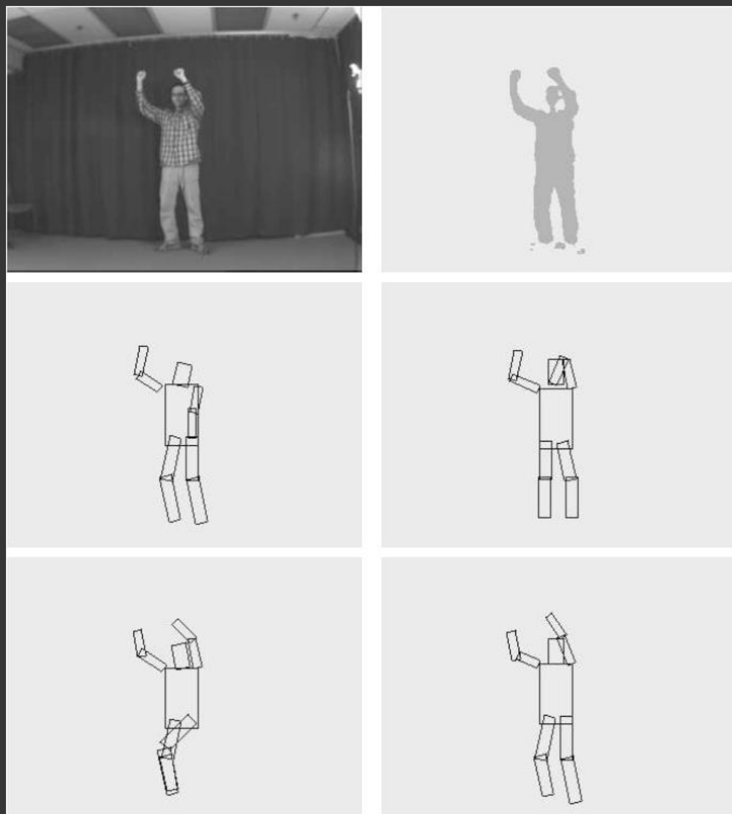
D. Ramanan, D. Forsyth, and A. Zisserman. [Tracking People by Learning their Appearance](). PAMI 2007.

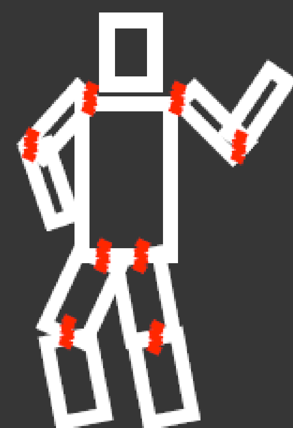
Top-down method to build model: Exploit "easy" poses

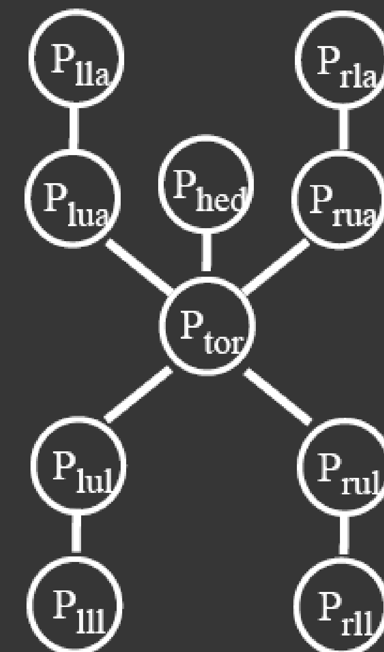
D. Ramanan, D. Forsyth, and A. Zisserman. Tracking People by Learning their Appearance. PAMI 2007.

Pictorial structure model

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







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

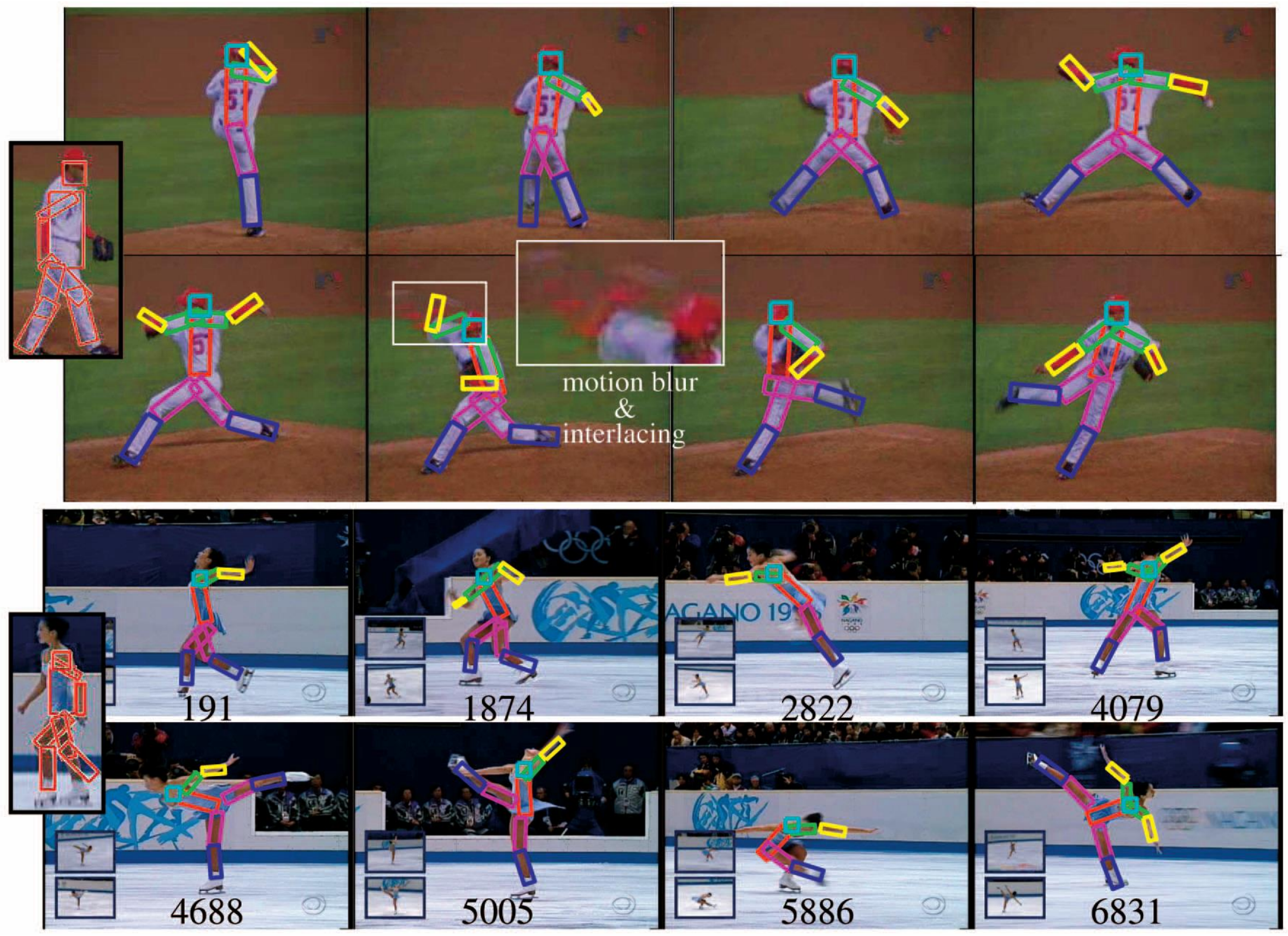
part geometry

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

This section: advanced topics

- Action recognition
- 3D Scenes and Context
- Visual Question Answering (Tanmay)

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

  https://www.youtube.com/watch?v=bcgXAQcvxdc

- To predict

  http://www.youtube.com/watch?v=LQm25nW6aZw

How can we identify actions?

Motion

Pose

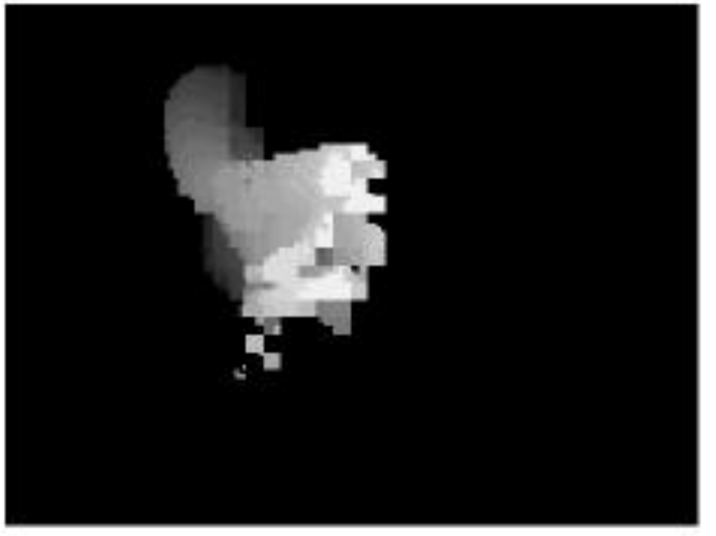
Held Objects

Nearby Objects

Representing Motion

Optical Flow with Motion History

sit-down

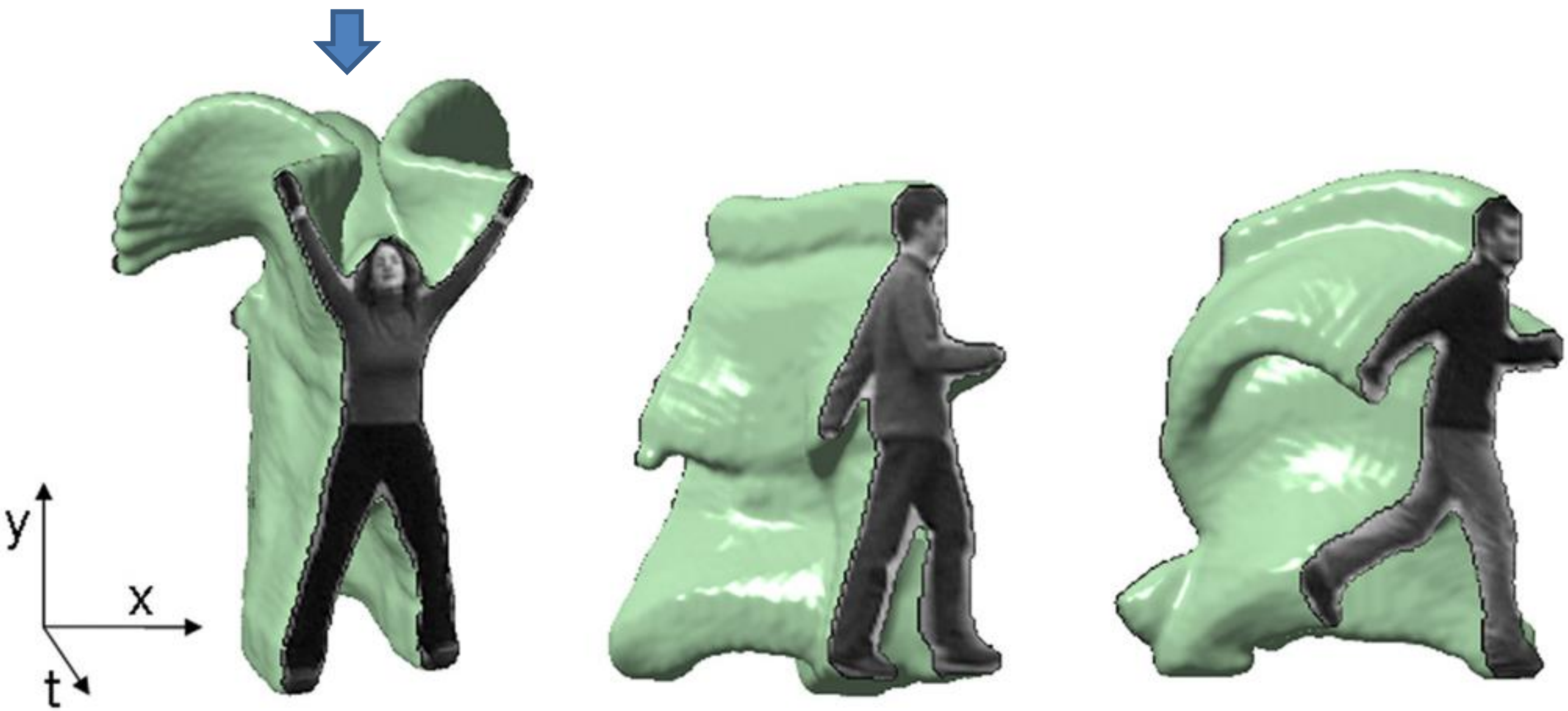


sit-down MHI

Bobick Davis 2001

Representing Motion

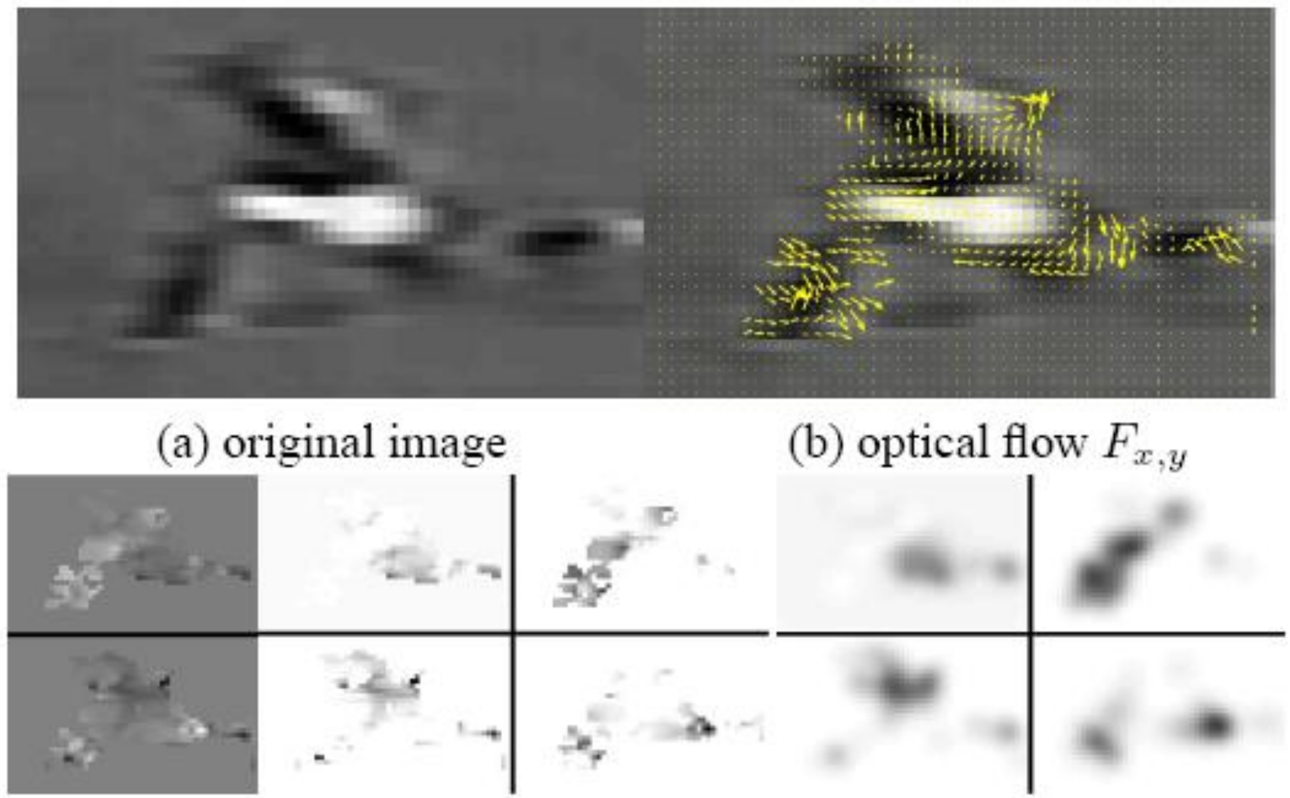
Space-Time Volumes



Blank et al. 2005

Representing Motion

Optical Flow with Split Channels



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

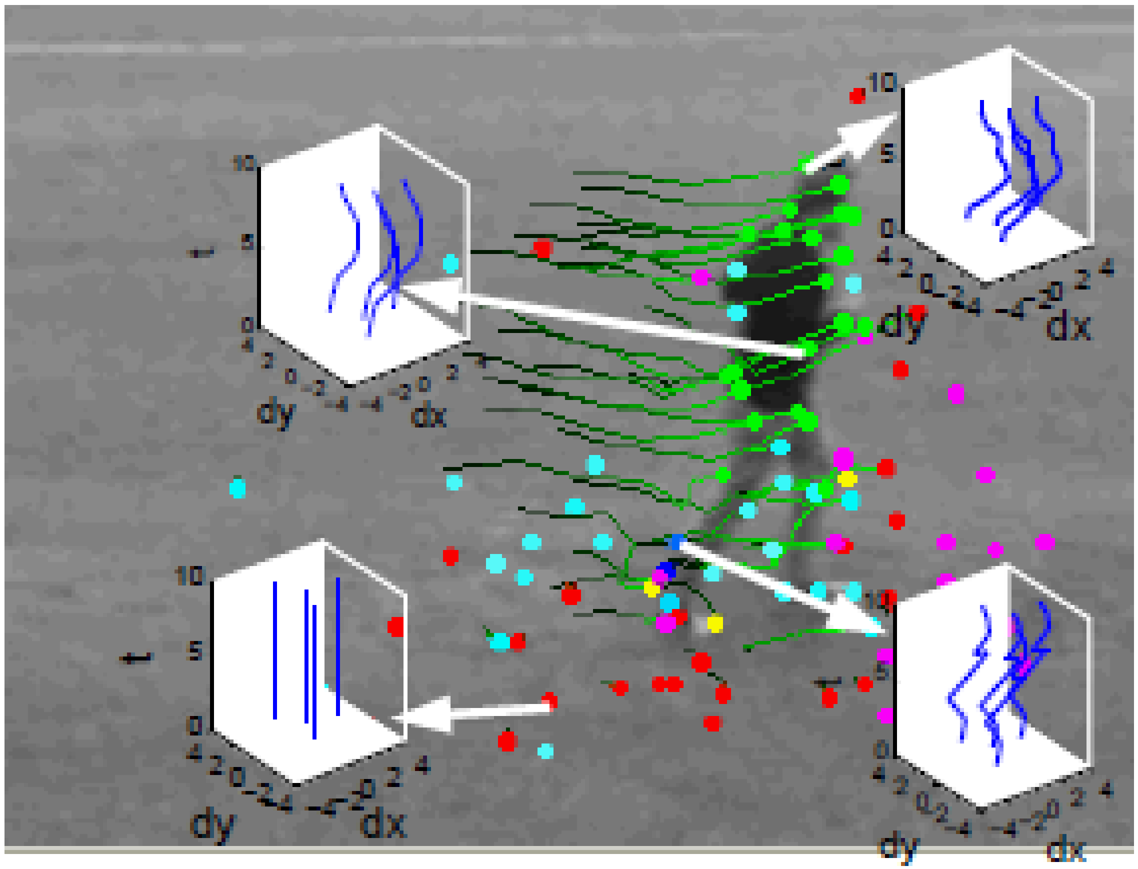
(c) F_x, F_y (d) $F_x^+, F_x^-, F_y^+, F_y^-$ (e) $Fb_x^+, Fb_x^-, Fb_y^+, Fb_y^-$

optical flow split into pos/neg channels blurred pos/neg flow

[Efros et al. 2003](#)

Representing Motion

Tracked Points



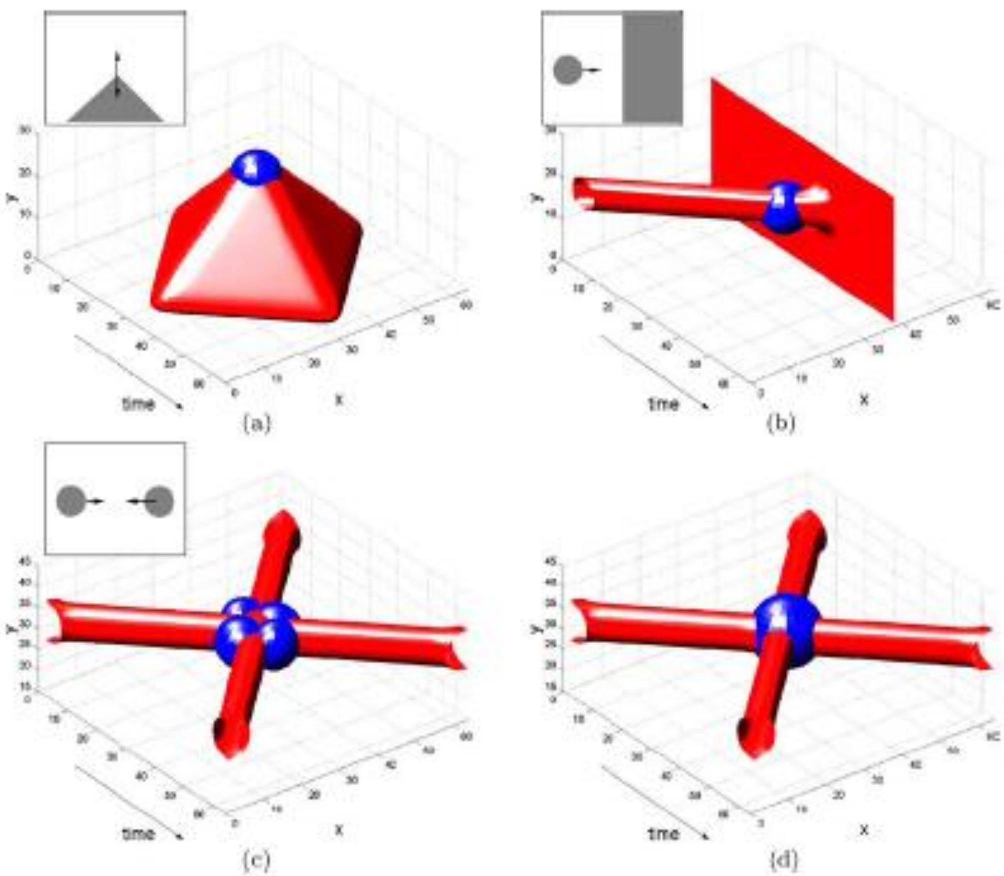
Matikainen et al. 2009

Representing Motion

Space-Time Interest Points

Moving corner

Ball hits wall



Balls collide

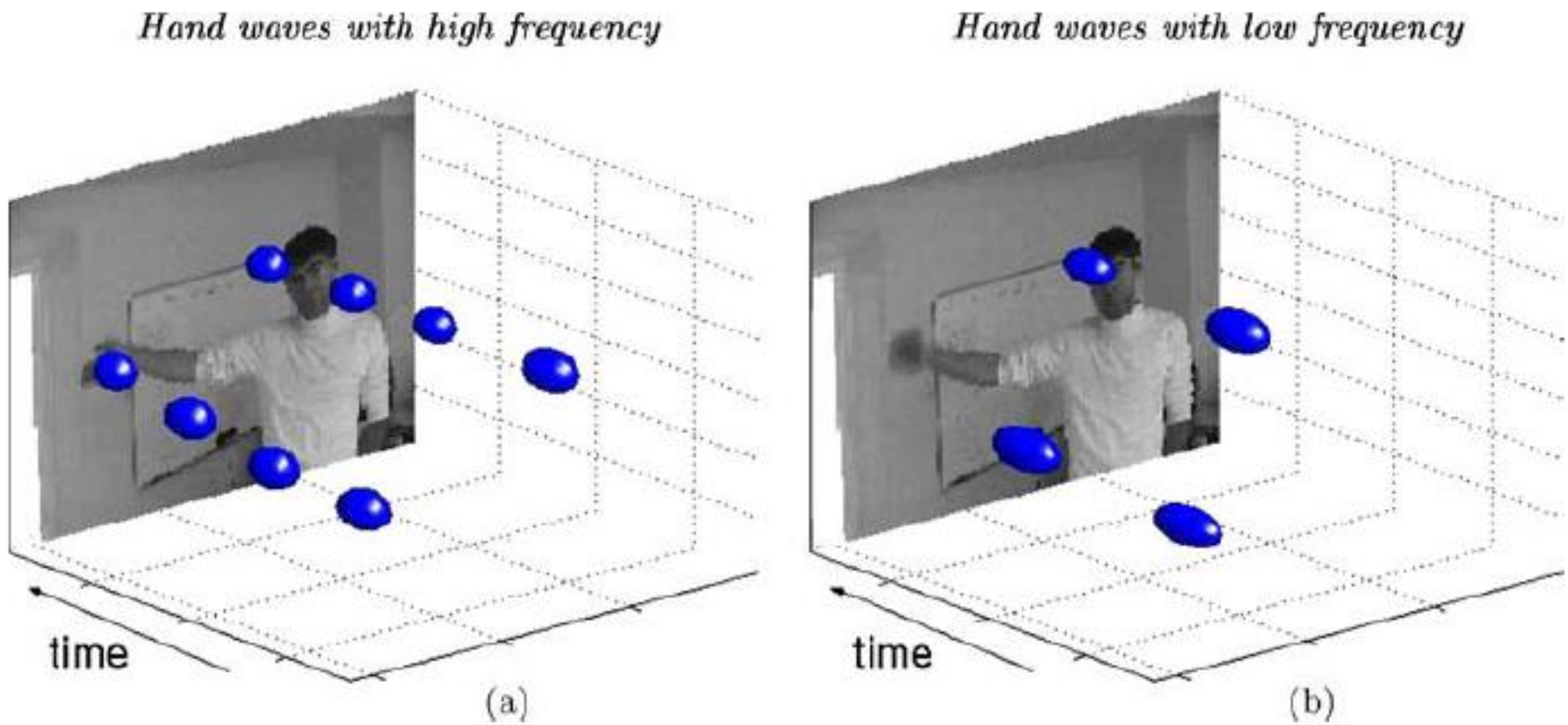
Balls collide (different scale)

Corner detectors in space-time

[Laptev 2005](#)

Representing Motion

Space-Time Interest Points



[Laptev 2005]

Examples of Action Recognition Systems

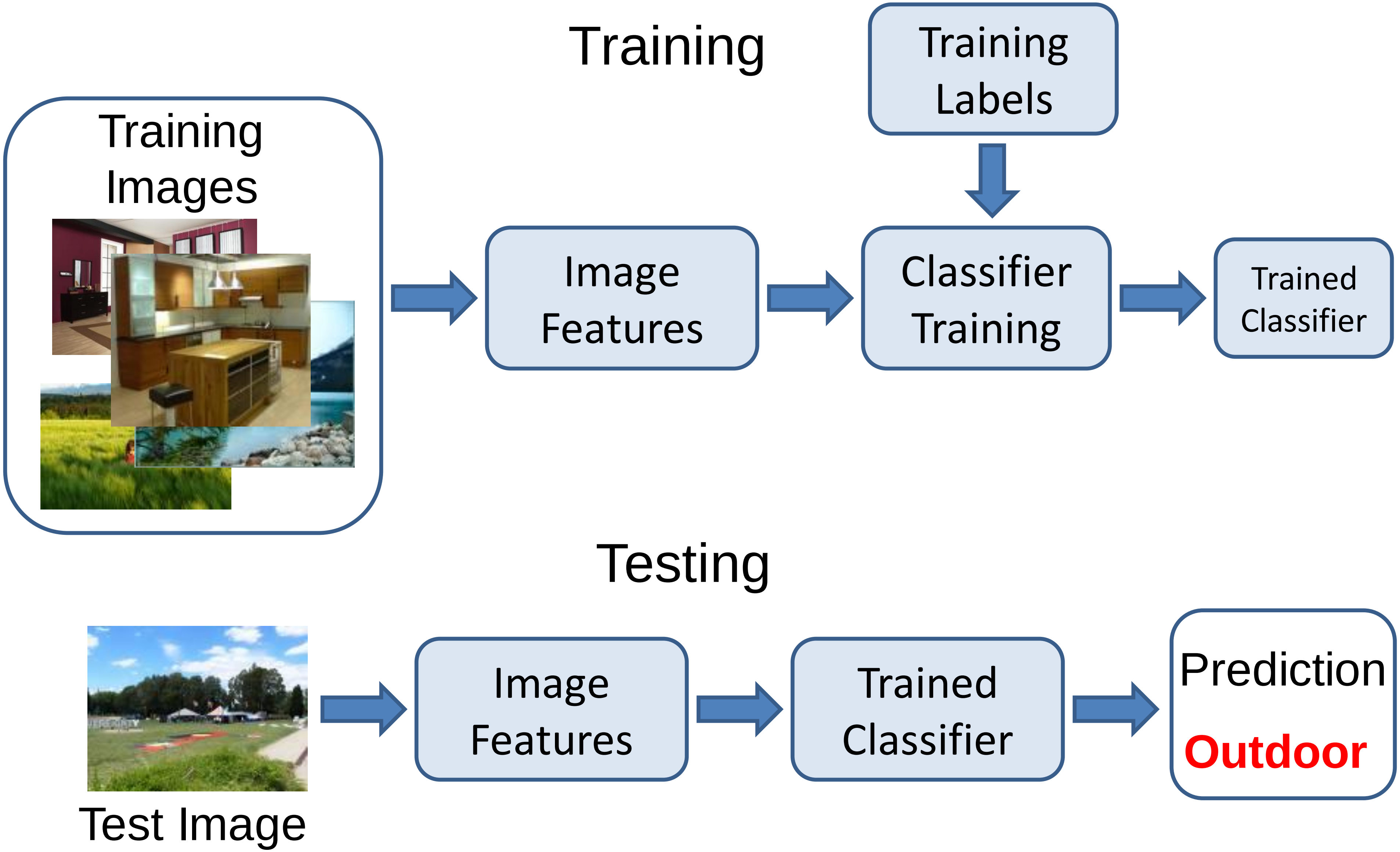
- Feature-based classification

- Recognition using pose and objects

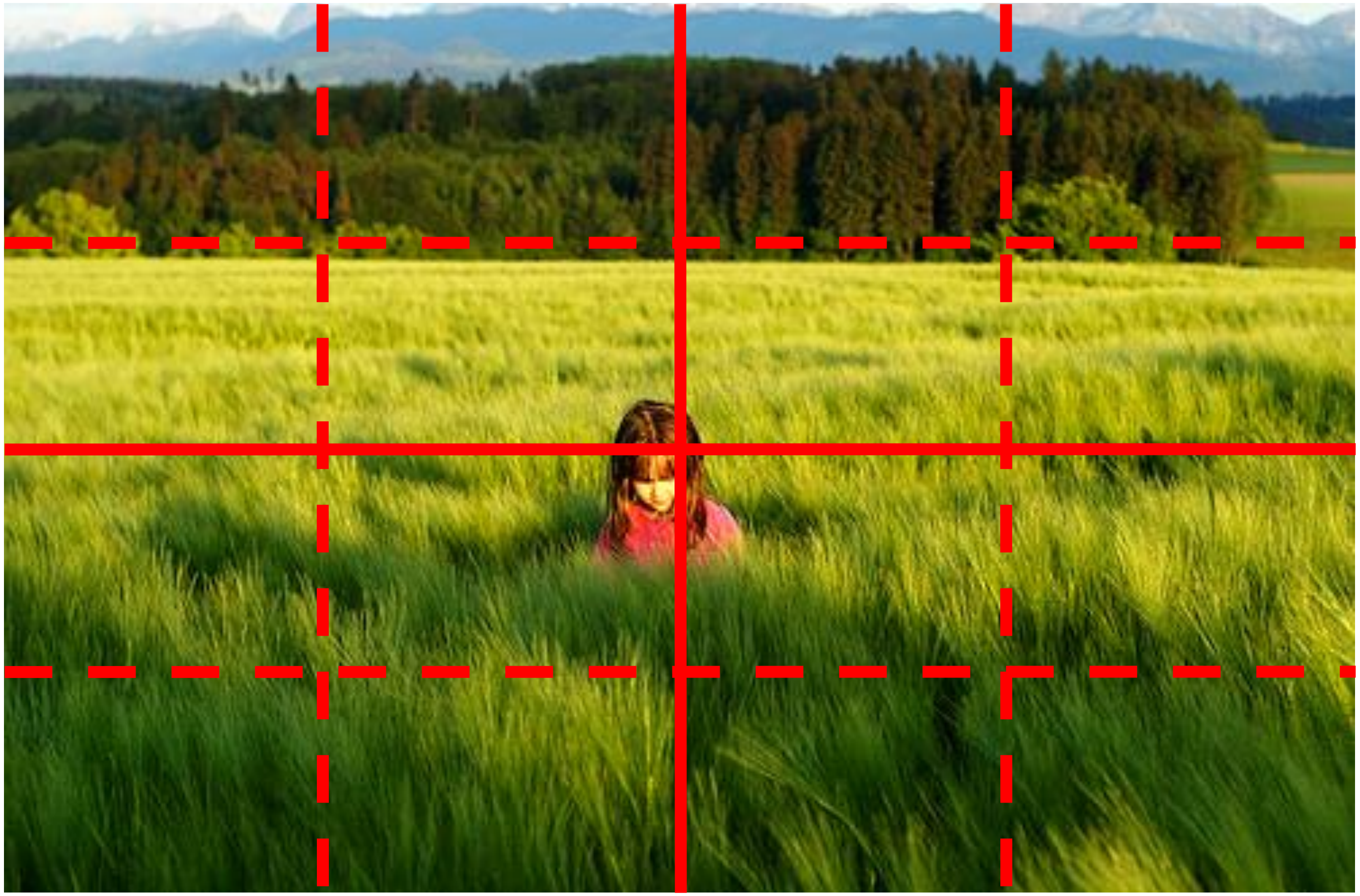
Action recognition as classification

[Retrieving actions in movies](), Laptev and Perez, 2007

Remember image categorization…



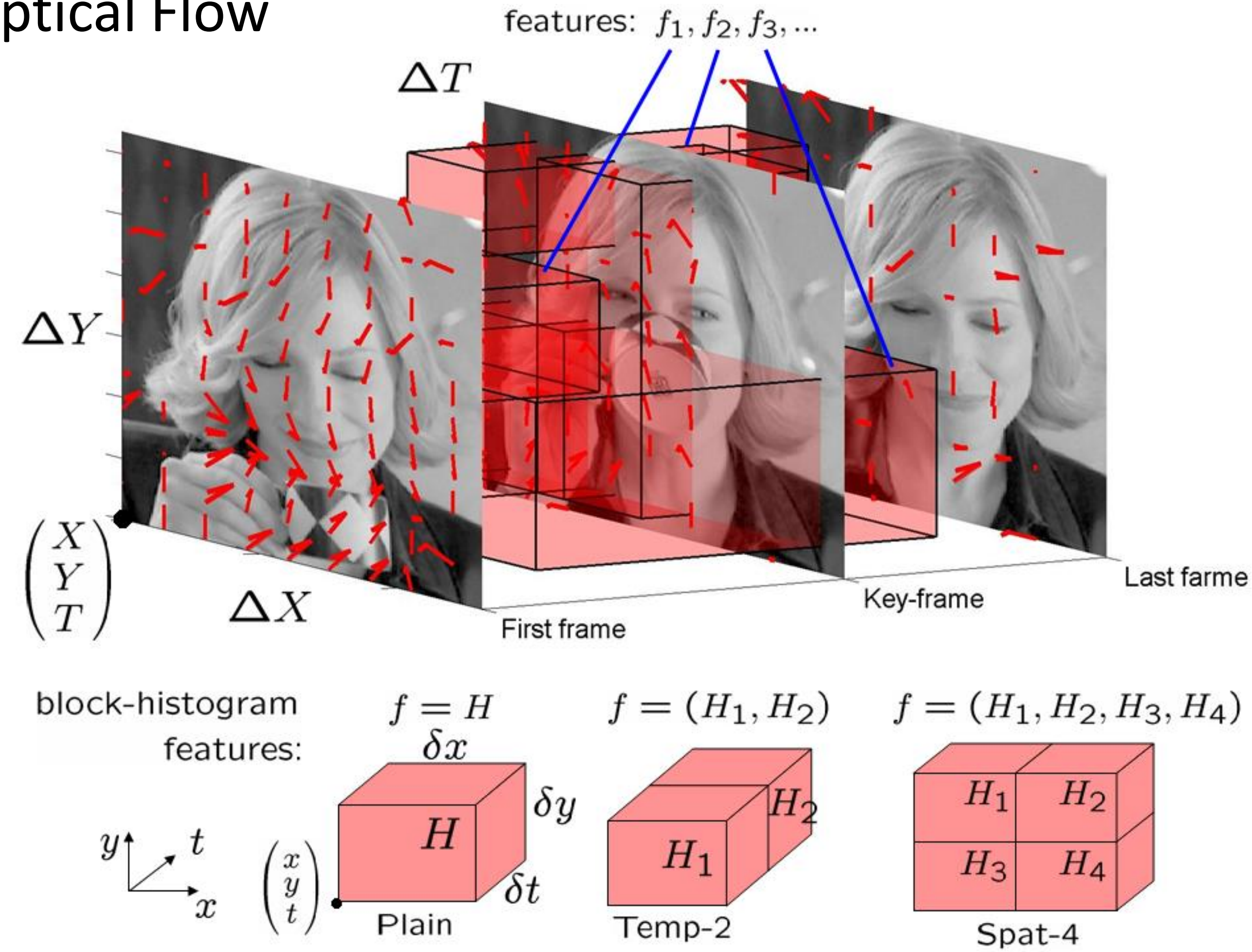
Remember spatial pyramids….



Compute histogram in each spatial bin

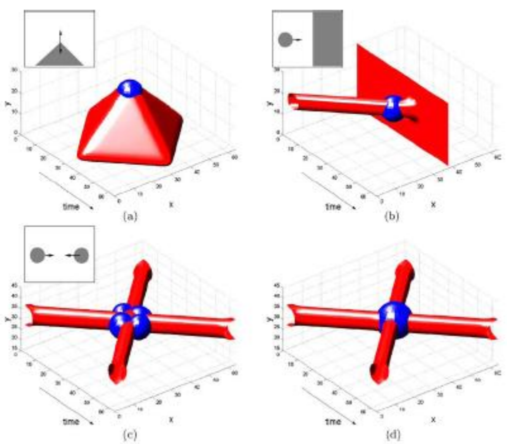
Features for Classifying Actions

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

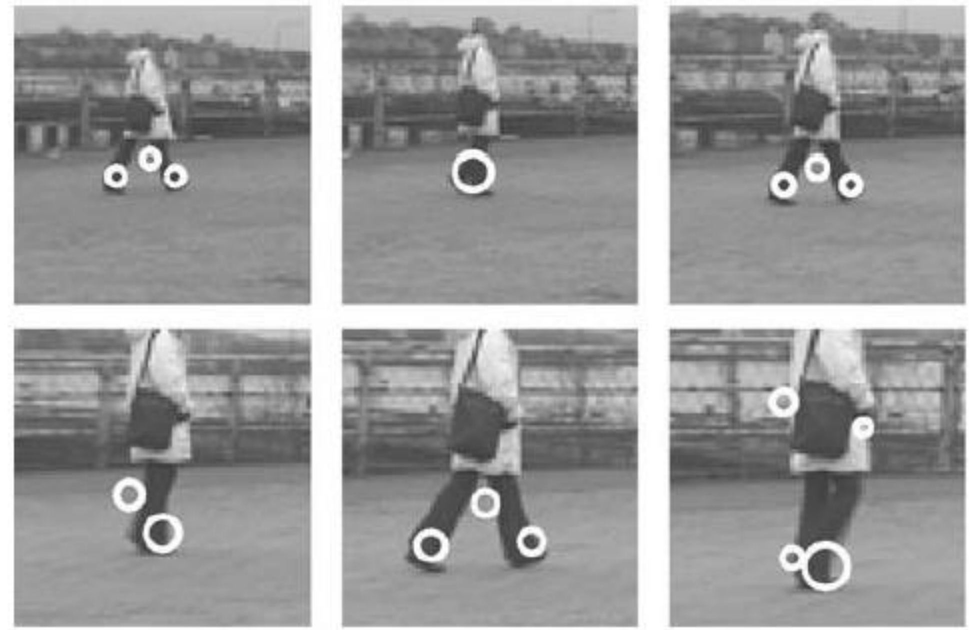


Features for Classifying Actions

2. Spatio-temporal interest points



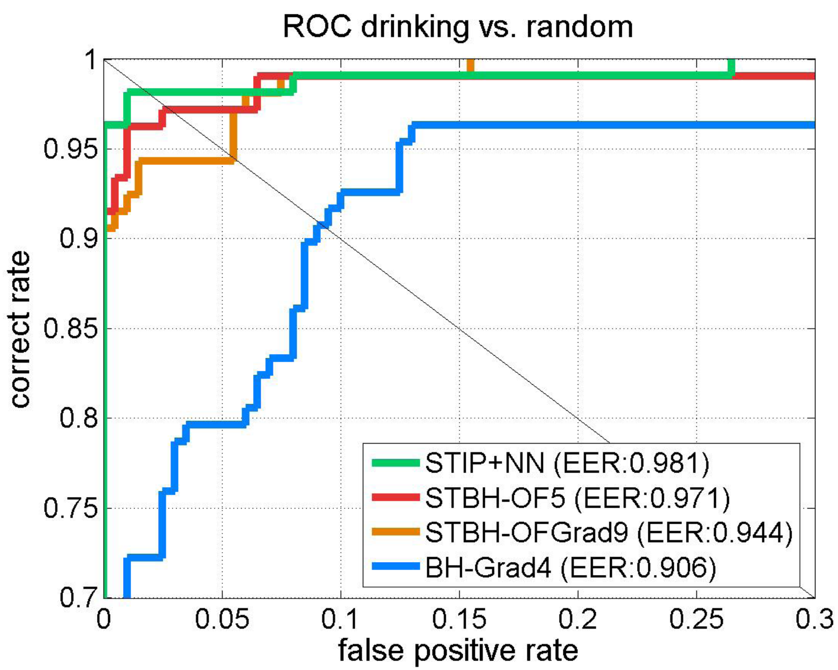
Corner detectors in space-time

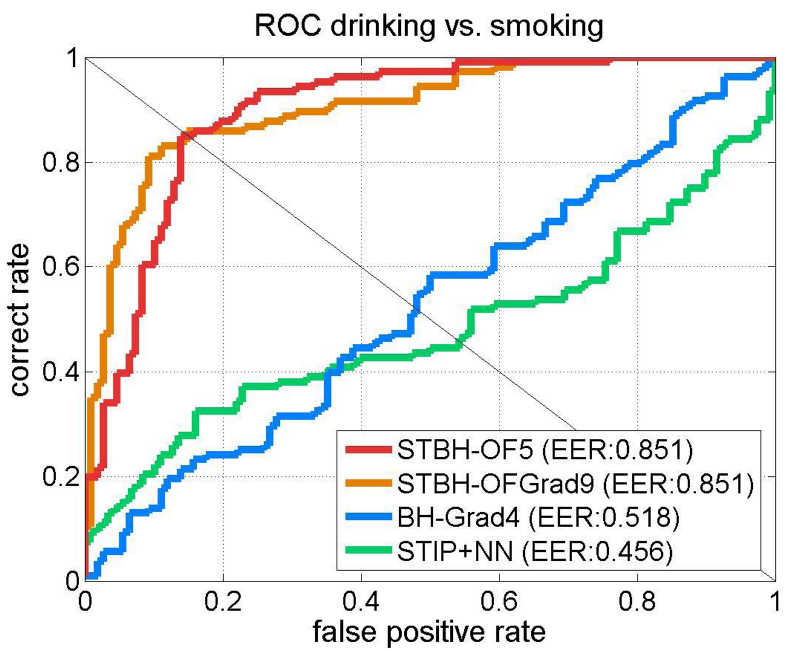


Descriptors based on Gaussian derivative filters over x, y, time

Classification

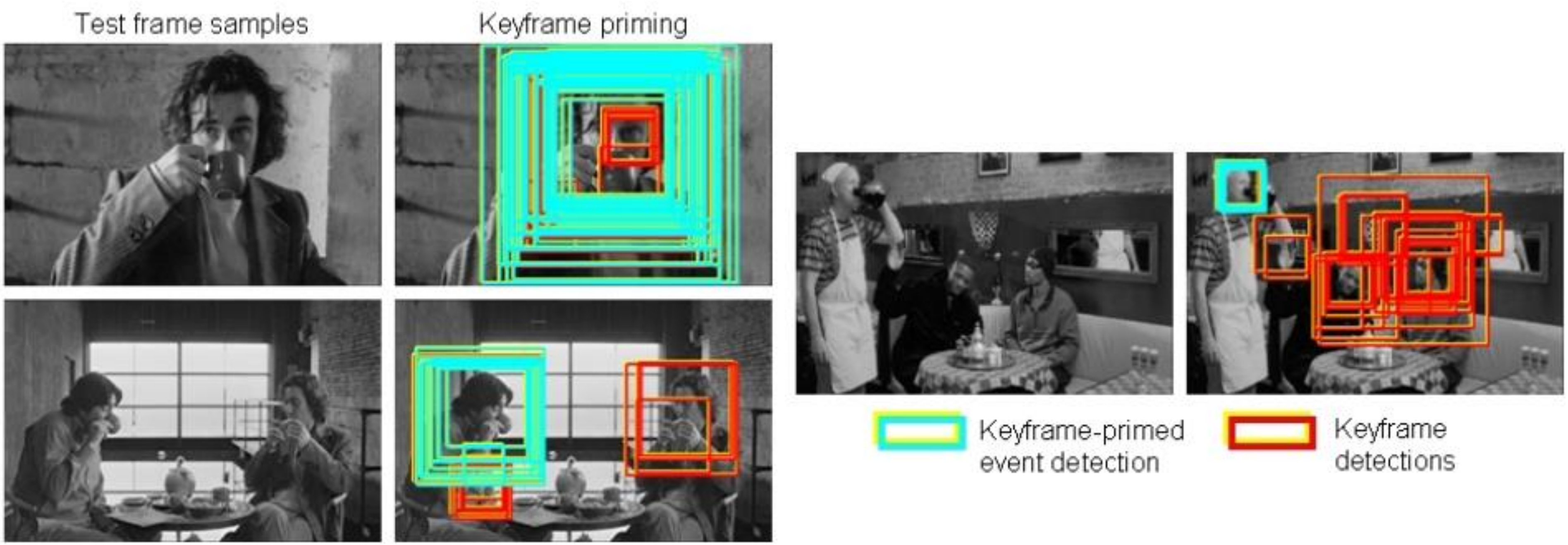
- Boosted stubs for pyramids of optical flow, gradient
- Nearest neighbor for STIP





Searching the video for an action

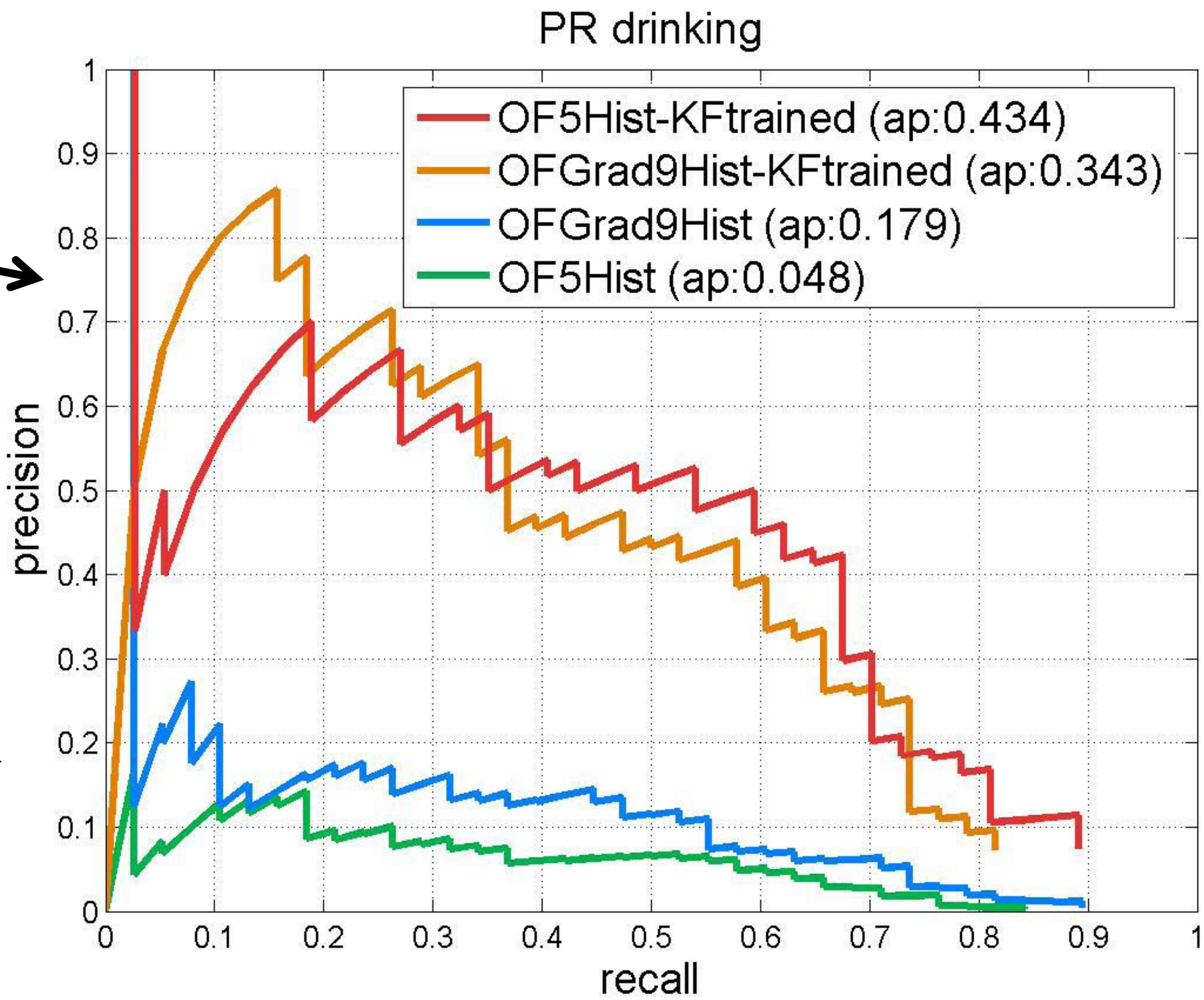
1. Detect keyframes using a trained HOG detector in each frame
2. Classify detected keyframes as positive (e.g., “drinking”) or negative (“other”)



Accuracy in searching video

With keyframe detection

Without keyframe detection



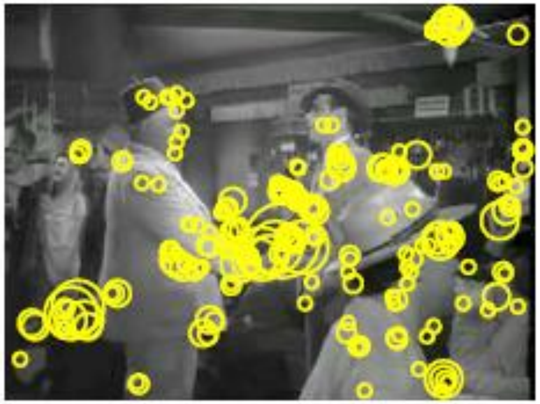
“Talk on phone”

“Get out of car”

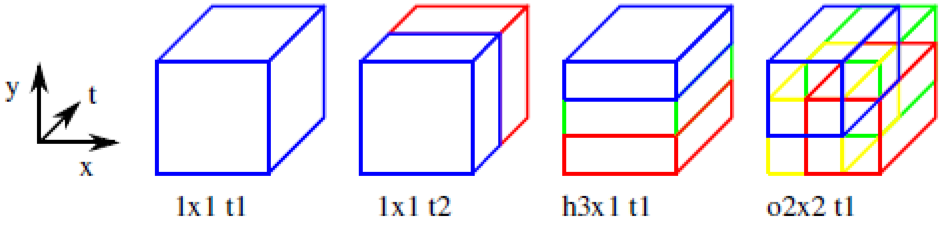
[Learning realistic human actions from movies, Laptev et al. 2008]

Approach

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



Interest Points



Spatio-Temporal Binning

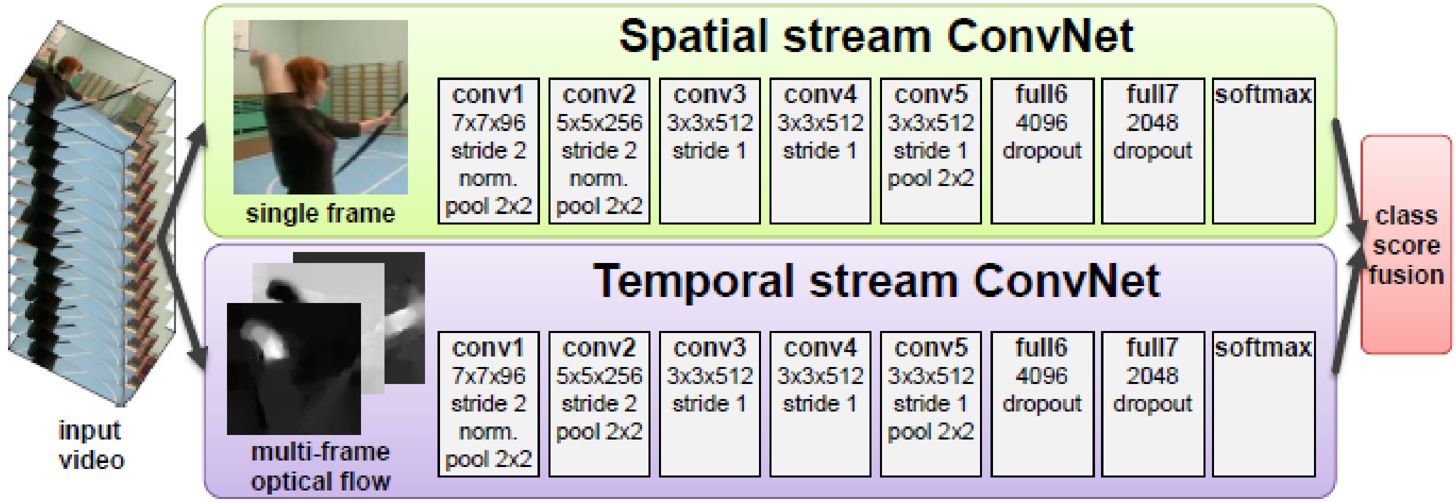
Results

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)

Bring on the conv-nets

Two-Stream Convolutional Networks for Action Recognition in Videos

Karen Simonyan Andrew Zisserman
Visual Geometry Group, University of Oxford
{karen, az}@robots.ox.ac.uk



IDT: improved dense trajectories + HOG/motion descriptor

Table 4: **Mean accuracy (over three splits) on UCF-101 and HMDB-51.**

Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

Note: No big win over hand-crafted features, perhaps due to more limited training data.
Latest (Wang et al. ECCV 2016) does better (94%, 69%)

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**

Human-Object Interaction

Holistic image based classification

Integrated reasoning

- Human pose estimation
- **Object detection**

Human-Object Interaction

Holistic image based classification

Integrated reasoning

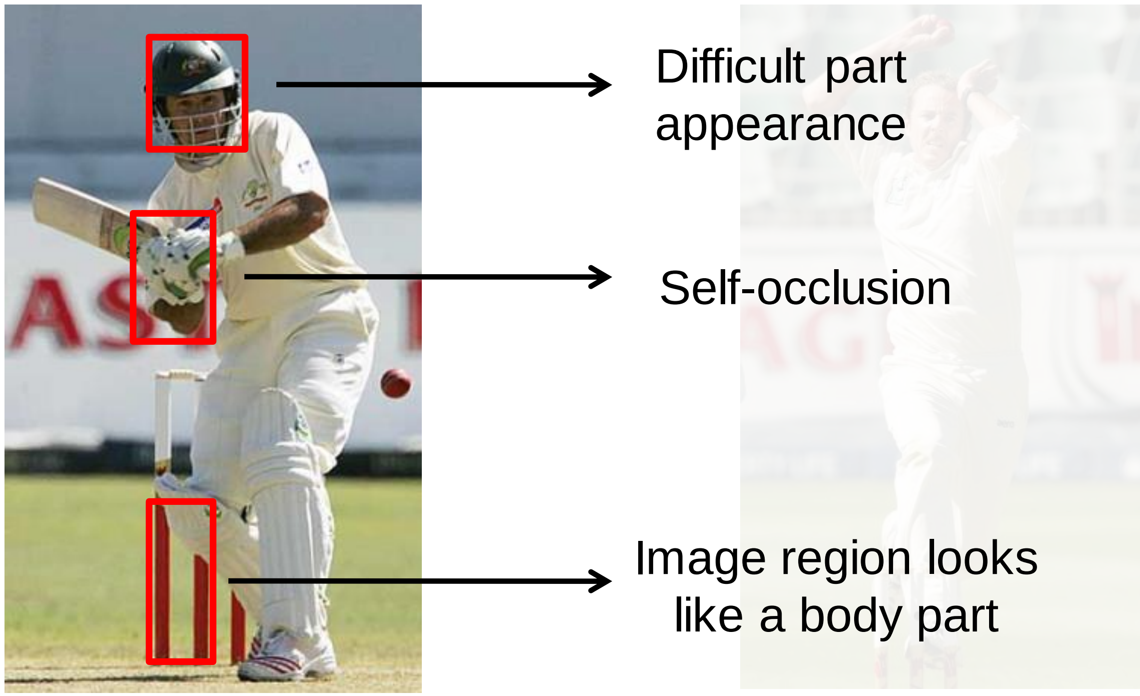
- **Human pose estimation**
- **Object detection**
- **Action categorization**

Activity: Tennis Forehand

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Human pose estimation is challenging.

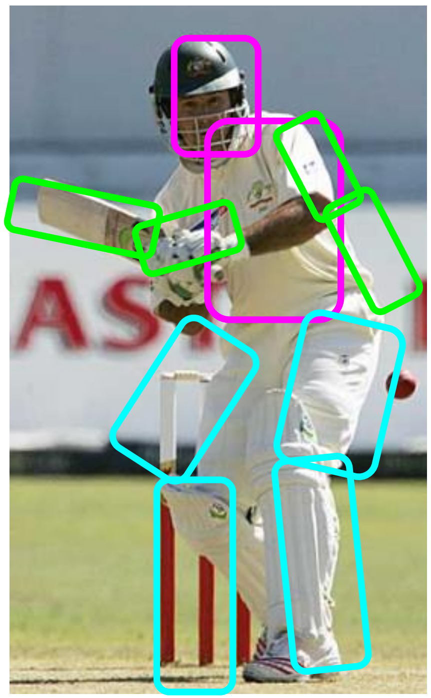


- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Human pose estimation is challenging.



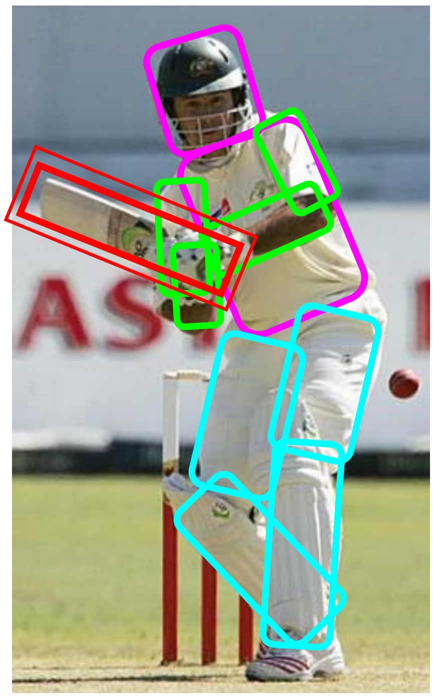
- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

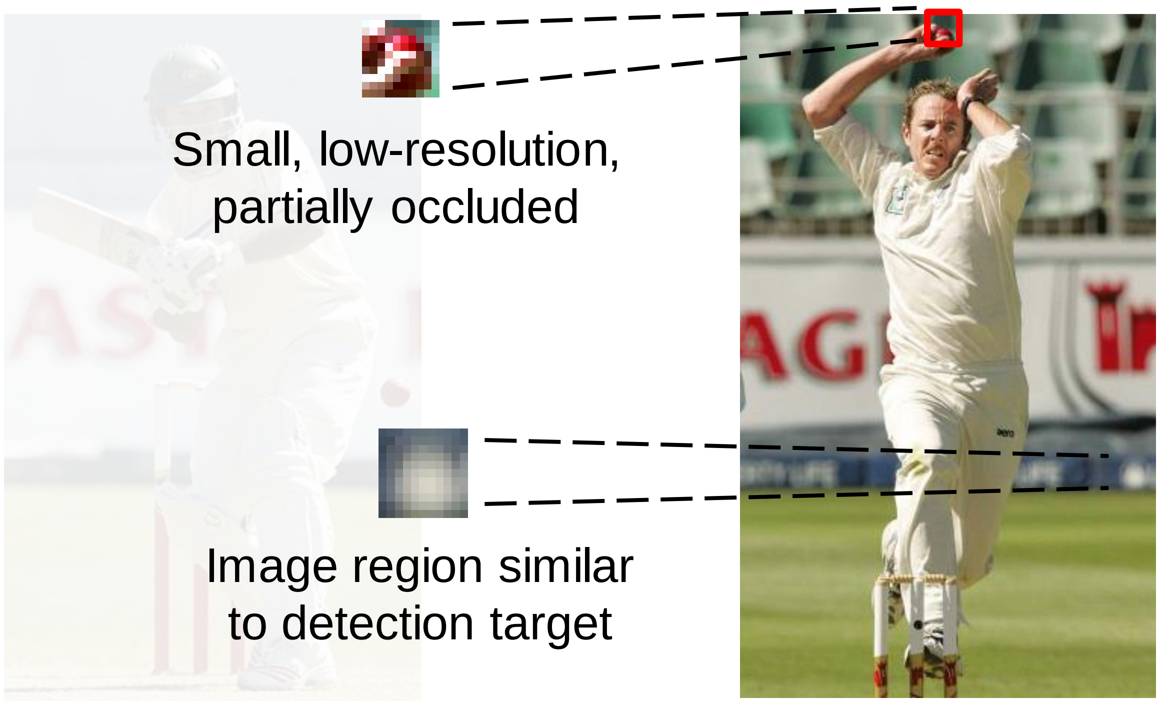
Facilitate

Given the object is detected.



Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

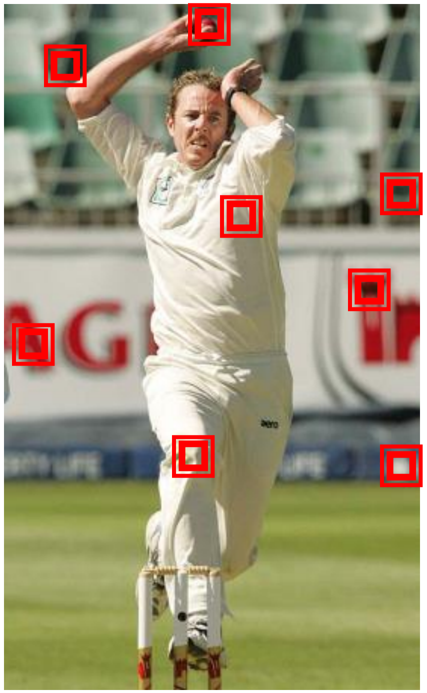


Object detection is challenging

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

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection



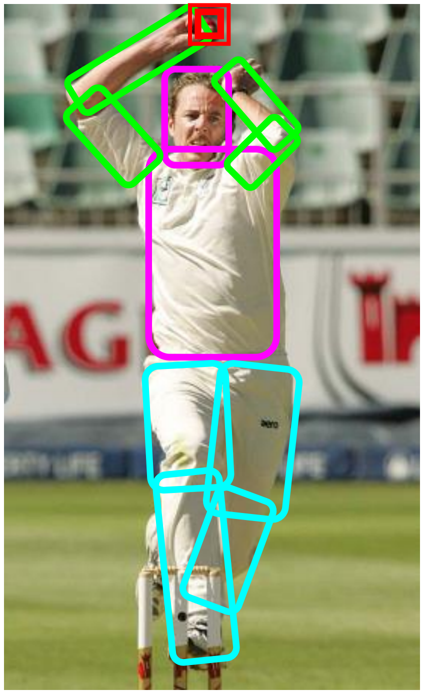
Object detection is challenging

- 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

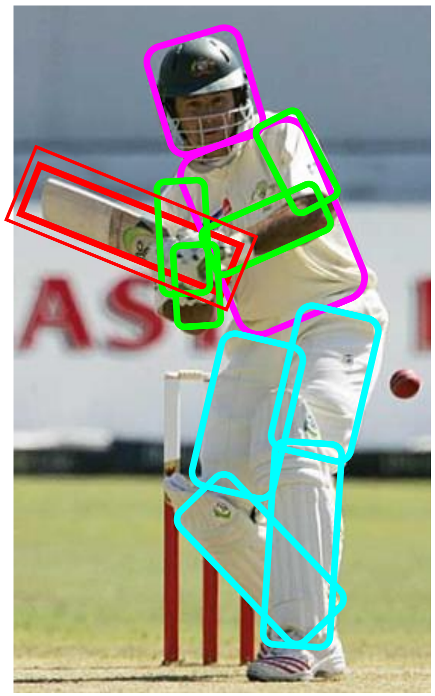
Facilitate

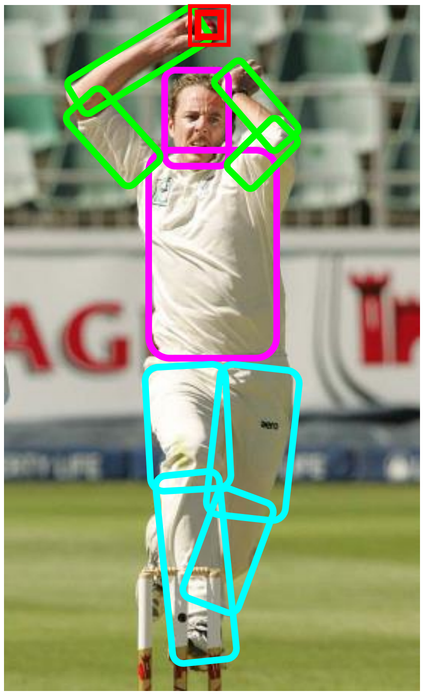


Given the pose is estimated.

Human pose estimation & Object detection

Mutual Context





Slide Credit: Yao/Fei-Fei

Mutual Context Model Representation

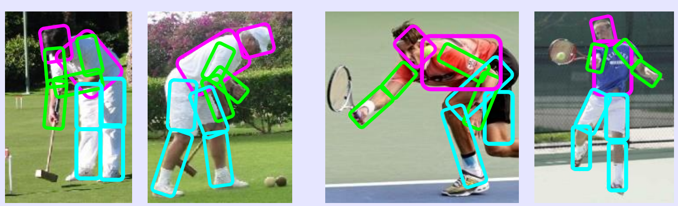
A:

Tennis forehand | Croquet shot | Volleyball smash | . . .

O:

Tennis racket | Croquet mallet | Volleyball | . . .

H:

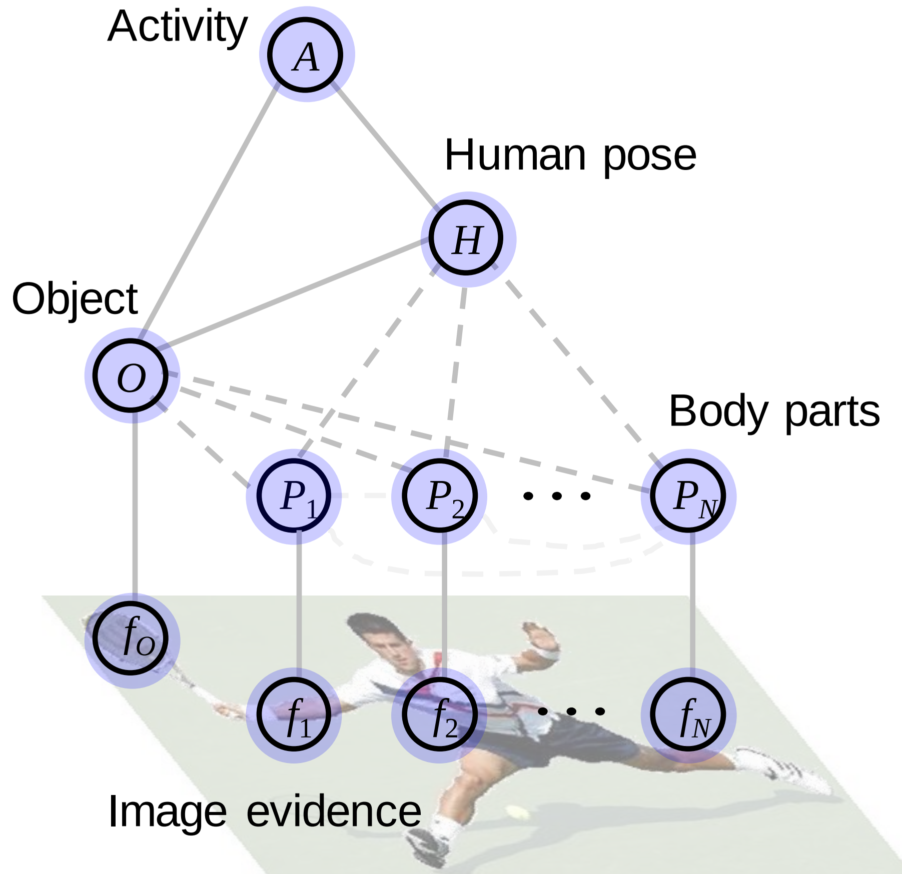


Intra-class variations

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

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

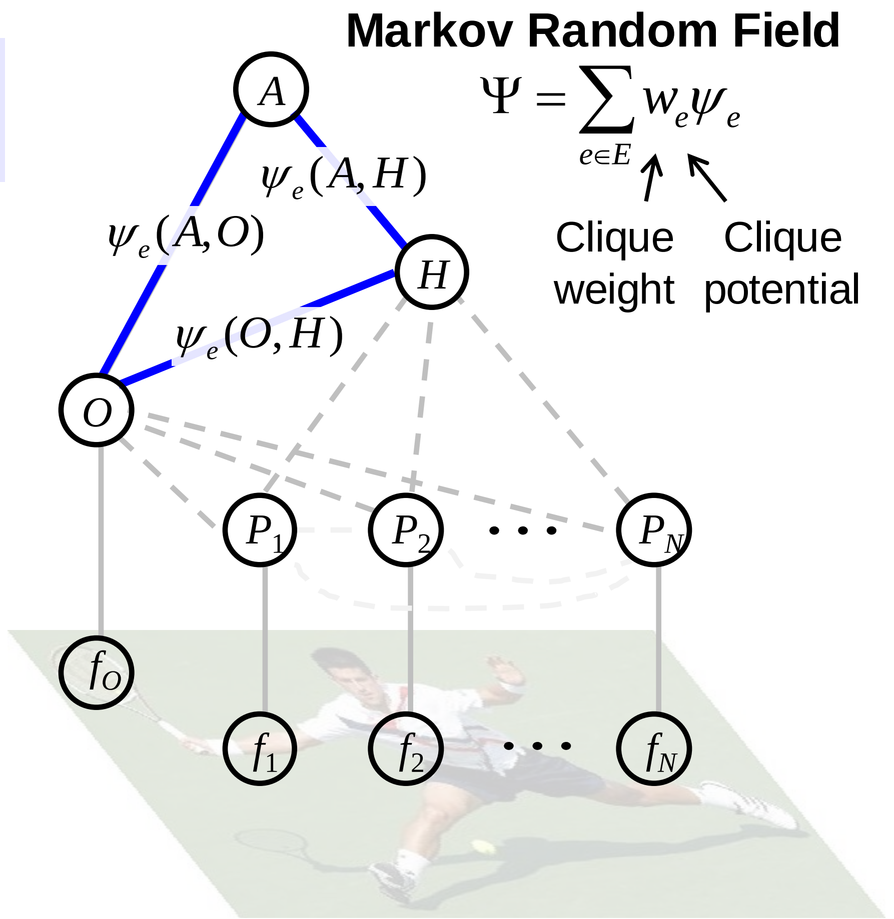
f: Shape context. [Belongie et al, 2002]



Slide Credit: Yao/Fei-Fei

Mutual Context Model Representation

- $\psi_e(A,O)$, $\psi_e(A,H)$, $\psi_e(O,H)$: Frequency of co-occurrence between A, O, and H.



Slide Credit: Yao/Fei-Fei

Mutual Context Model Representation

- $\psi_e(A,O)$, $\psi_e(A,H)$, $\psi_e(O,H)$: Frequency of co-occurrence between A, O, and H.

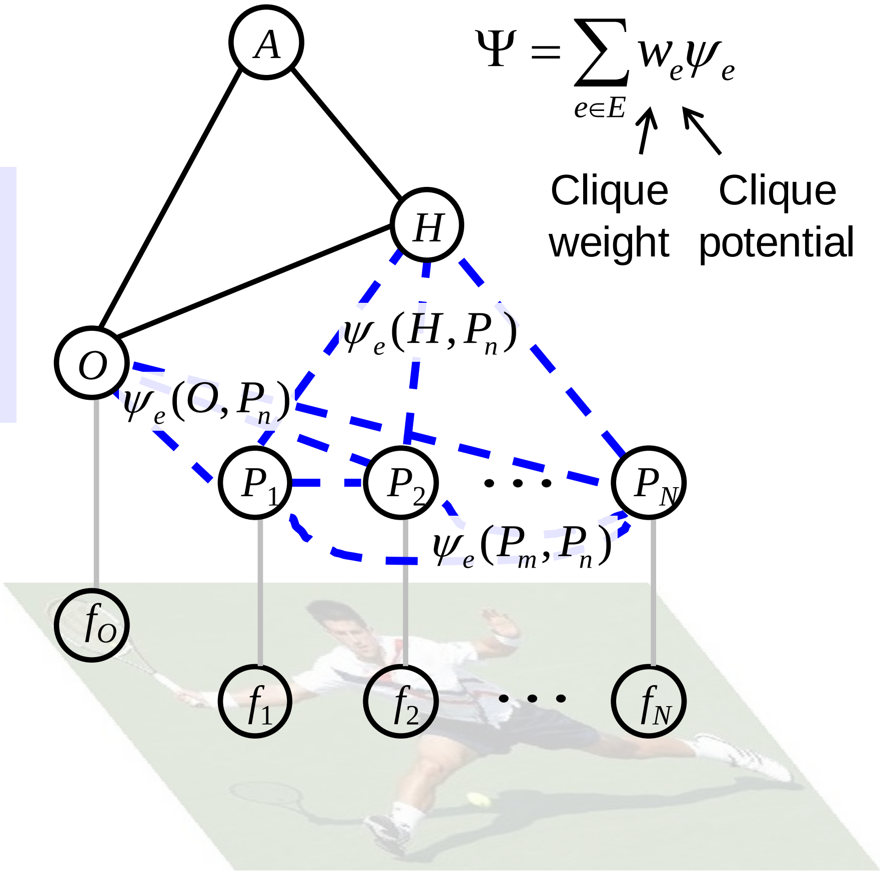
- $\psi_e(O,P_n)$, $\psi_e(H,P_n)$, $\psi_e(P_m,P_n)$: Spatial relationship among object and body parts.

$$\underbrace{\mathrm{bin}\left(l_O - l_{P_n}\right)}_{\textbf{location}} \cdot \underbrace{\mathrm{bin}\left(\theta_O - \theta_{P_n}\right)}_{\textbf{orientation}} \cdot \underbrace{\mathrm{N}\left(s_O / s_{P_n}\right)}_{\textbf{size}}$$

Markov Random Field

$$\Psi = \sum_{e \in E} w_e \psi_e$$

Clique weight (w_e) Clique potential (ψ_e)

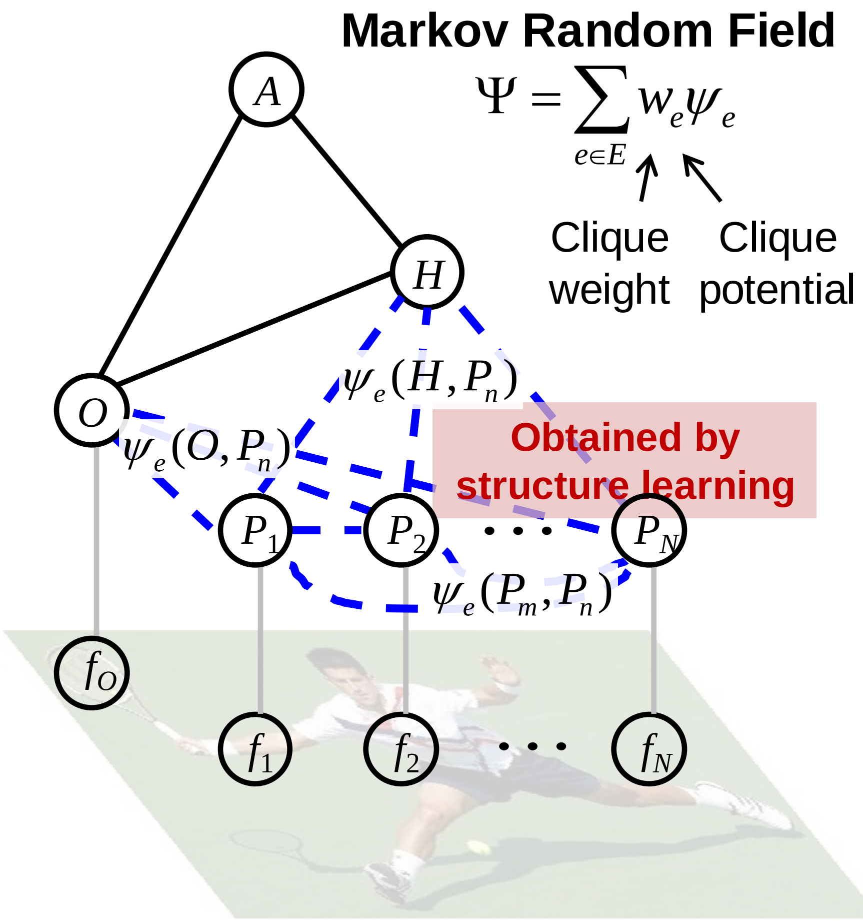


Mutual Context Model Representation

- $\psi_e(A,O)$, $\psi_e(A,H)$, $\psi_e(O,H)$: Frequency of co-occurrence between A, O, and H.
- $\psi_e(O,P_n)$, $\psi_e(H,P_n)$, $\psi_e(P_m,P_n)$: Spatial relationship among object and body parts.

$$\underbrace{\mathrm{bin}\left(l_O - l_{P_n}\right)}_{\text{location}} \cdot \underbrace{\mathrm{bin}\left(\theta_O - \theta_{P_n}\right)}_{\text{orientation}} \cdot \underbrace{\mathrm{N}\left(s_O / s_{P_n}\right)}_{\text{size}}$$

- Learn structural connectivity among the body parts and the object.



Slide Credit: Yao/Fei-Fei

Mutual Context Model Representation

- $\psi_e(A,O)$, $\psi_e(A,H)$, $\psi_e(O,H)$: Frequency of co-occurrence between A, O, and H.

- $\psi_e(O,P_n)$, $\psi_e(H,P_n)$, $\psi_e(P_m,P_n)$: Spatial relationship among object and body parts.

$$\underbrace{\mathrm{bin}\left(l_O - l_{P_n}\right)}_{\textbf{location}} \cdot \underbrace{\mathrm{bin}\left(\theta_O - \theta_{P_n}\right)}_{\textbf{orientation}} \cdot \underbrace{\mathrm{N}\left(s_O / s_{P_n}\right)}_{\textbf{size}}$$

- Learn structural connectivity among the body parts and the object.

- $\psi_e(O, f_O)$ and $\psi_e(P_n, f_{P_n})$: Discriminative part detection scores.

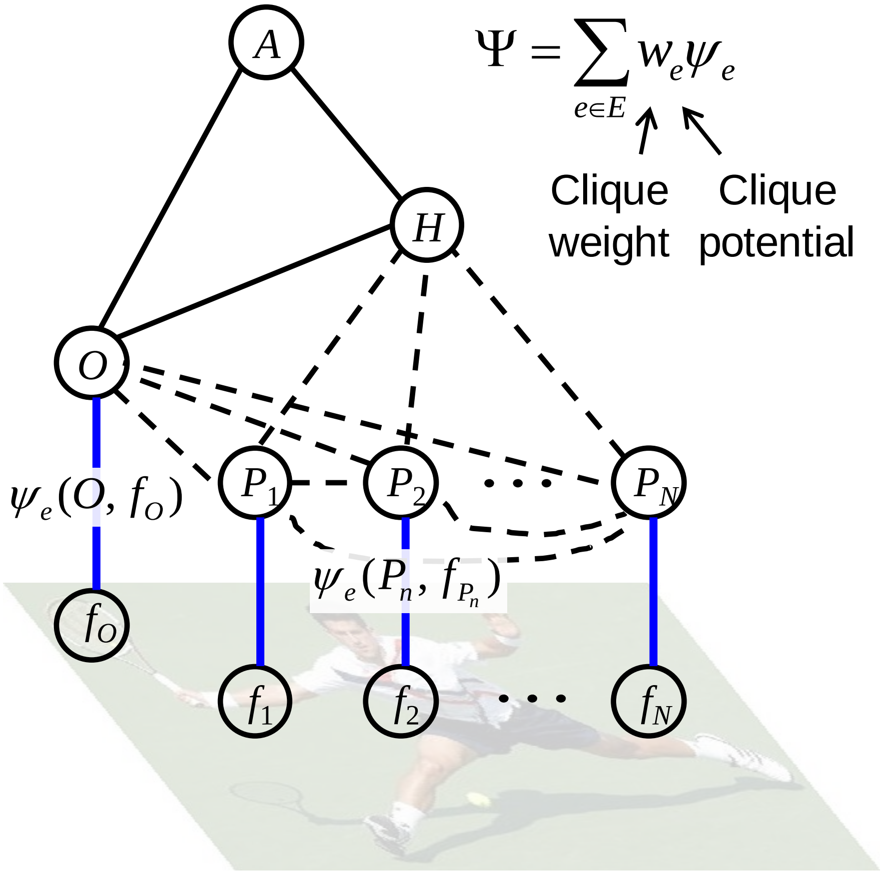
  Shape context + AdaBoost

  [Andriluka et al, 2009]
  [Belongie et al, 2002]
  [Viola & Jones, 2001]

Markov Random Field

$$\Psi = \sum_{e \in E} w_e \psi_e$$

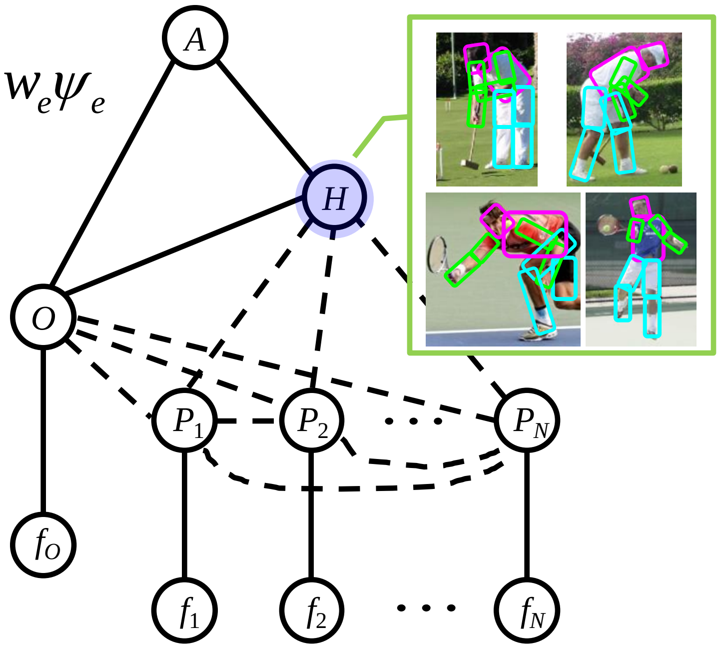
Clique weight — Clique potential



Slide Credit: Yao/Fei-Fei

Model Learning

$$\Psi = \sum_{e \in E} w_e \psi_e$$



Input:

cricket shot

cricket bowling

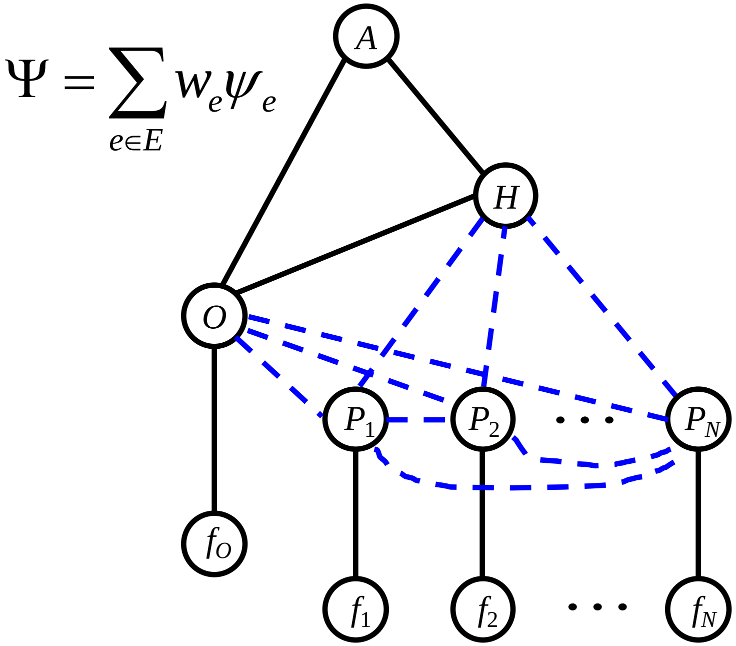
. . .

Goals:

Hidden human poses

Slide Credit: Yao/Fei-Fei

Model Learning



Input:

cricket shot

cricket bowling

. . .

Goals:

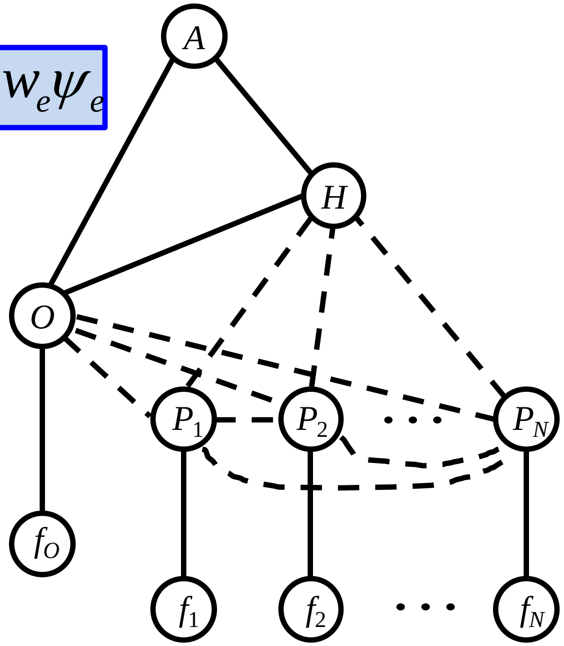
Hidden human poses

Structural connectivity

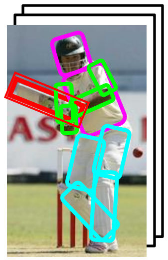
Slide Credit: Yao/Fei-Fei

Model Learning

$$\Psi = \sum_{e \in E} \boxed{w_e \psi_e}$$



Input:



cricket
shot

cricket
bowling

. . .

Goals:

Hidden human poses

Structural connectivity

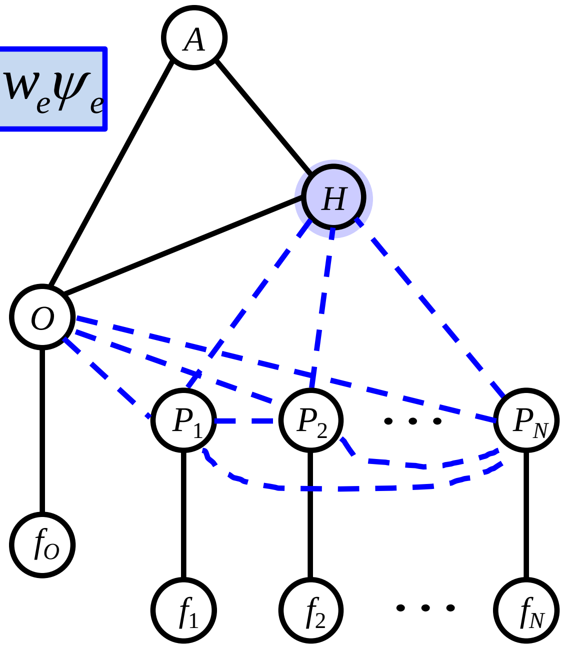
Potential parameters

Potential weights

Slide Credit: Yao/Fei-Fei

Model Learning

$$\Psi = \sum_{e \in E} \boxed{w_e \psi_e}$$



Input:

cricket shot

cricket bowling

. . .

Goals:

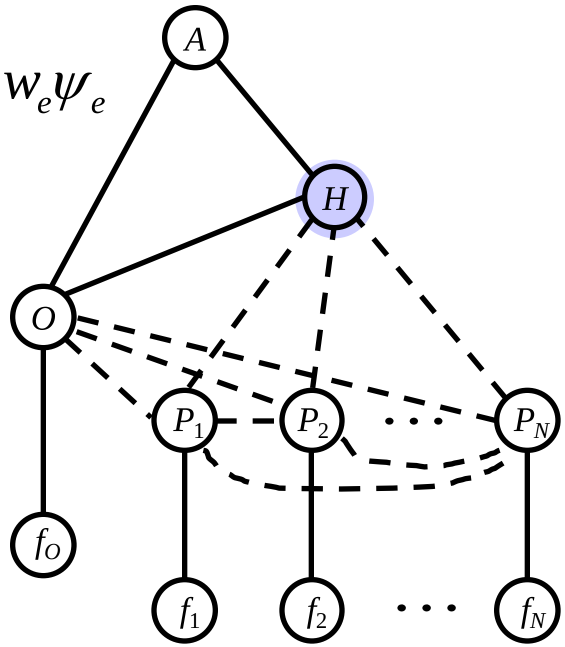
Hidden human poses → **Hidden variables**

Structural connectivity → **Structure learning**

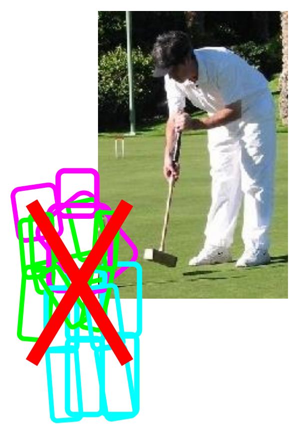
Potential parameters
Potential weights } **Parameter estimation**

Model Learning

$$\Psi = \sum_{e \in E} w_e \psi_e$$



Approach:



Goals:

Hidden human poses

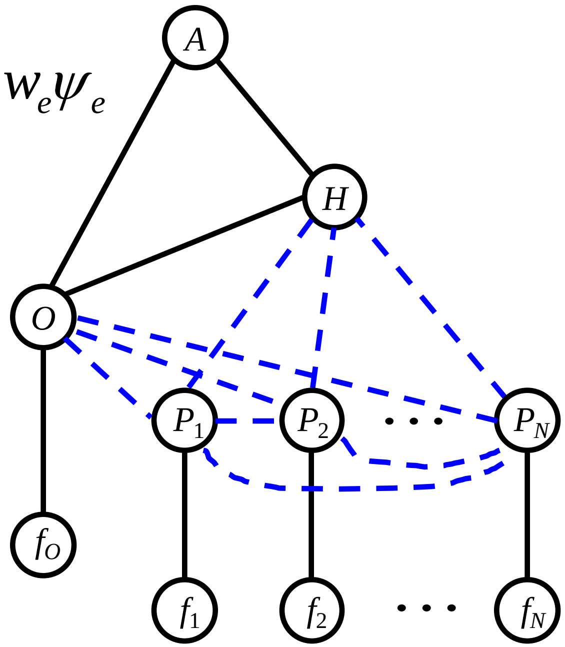
Structural connectivity

Potential parameters

Potential weights

Model Learning

$$\Psi = \sum_{e \in E} w_e \psi_e$$



Goals:

Hidden human poses

Structural connectivity

Potential parameters

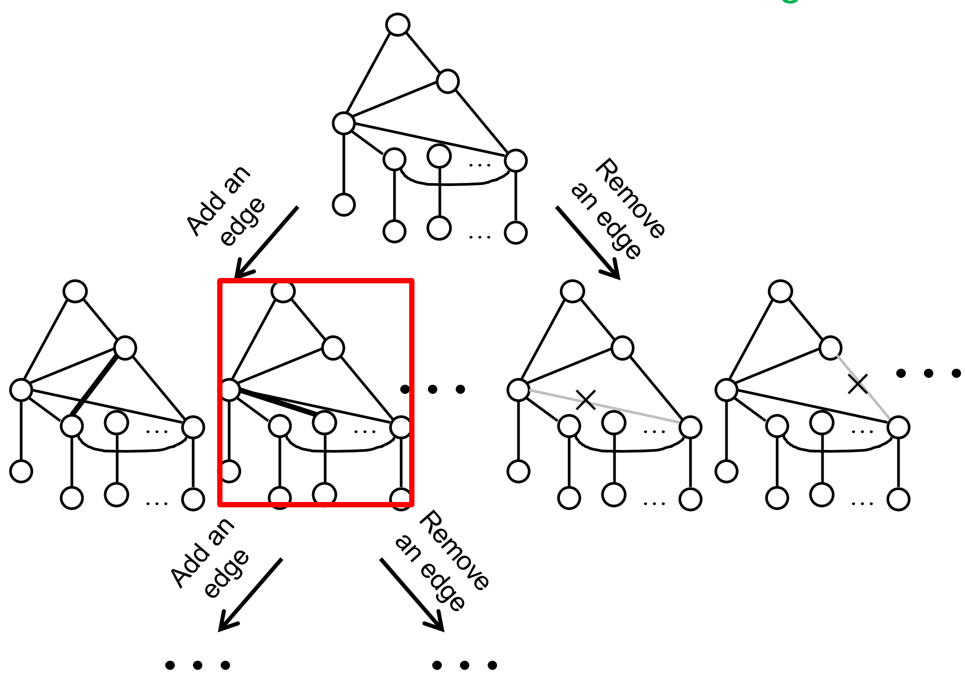
Potential weights

Approach:

Hill-climbing $\max_{E=\{e\}} \left\{ \sum_e w_e \psi_e - \frac{\left(|E| - \mu\right)^2}{2\sigma^2} \right\}$

Joint density of the model

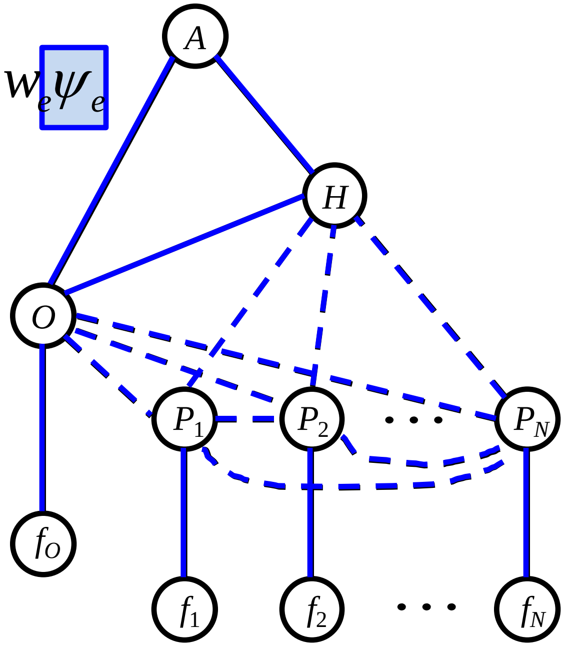
Gaussian priori of the edge number



Slide Credit: Yao/Fei-Fei

Model Learning

$$\Psi = \sum_{e \in E} w_e \boxed{\psi_e}$$



Goals:

Hidden human poses

Structural connectivity

Potential parameters

Potential weights

Approach:

- Maximum likelihood

$\psi_e(A,O)$ $\psi_e(A,H)$ $\psi_e(O,H)$

$\psi_e(H,P_n)$ $\psi_e(O,P_n)$ $\psi_e(P_m,P_n)$

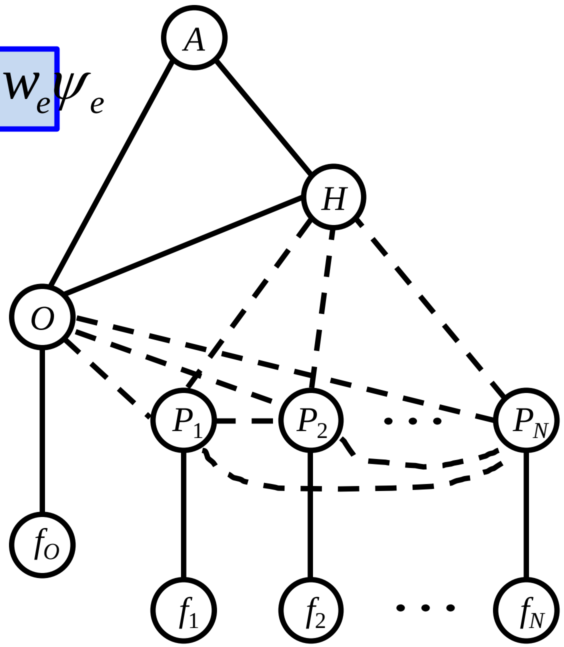
- Standard AdaBoost

$\psi_e(O, f_O)$ $\psi_e(P_n, f_{P_n})$

Slide Credit: Yao/Fei-Fei

Model Learning

$$\Psi = \sum_{e \in E} \boxed{w_e} \psi_e$$



Goals:

Hidden human poses

Structural connectivity

Potential parameters

Potential weights

Approach:

Max-margin learning

$$\min_{\mathbf{w},\xi} \frac{1}{2} \sum_r \|\mathbf{w}_r\|_2^2 + \beta \sum_i \xi_i$$

$$\text{s.t. } \forall i, r \text{ where } y(r) \neq y(c_i),$$

$$\mathbf{w}_{c_i} \cdot \mathbf{x}_i - \mathbf{w}_r \cdot \mathbf{x}_i \geq 1 - \xi_i$$

$$\forall i, \xi_i \geq 0$$

Notations

- $\mathbf{x}_i$: Potential values of the i-th image.
- $\mathbf{w}_r$: Potential weights of the r-th pose.
- $y(r)$: Activity of the r-th pose.
- ξ_i: A slack variable for the i-th image.

Slide Credit: Yao/Fei-Fei

Learning Results

Cricket defensive shot

Cricket bowling

Croquet shot

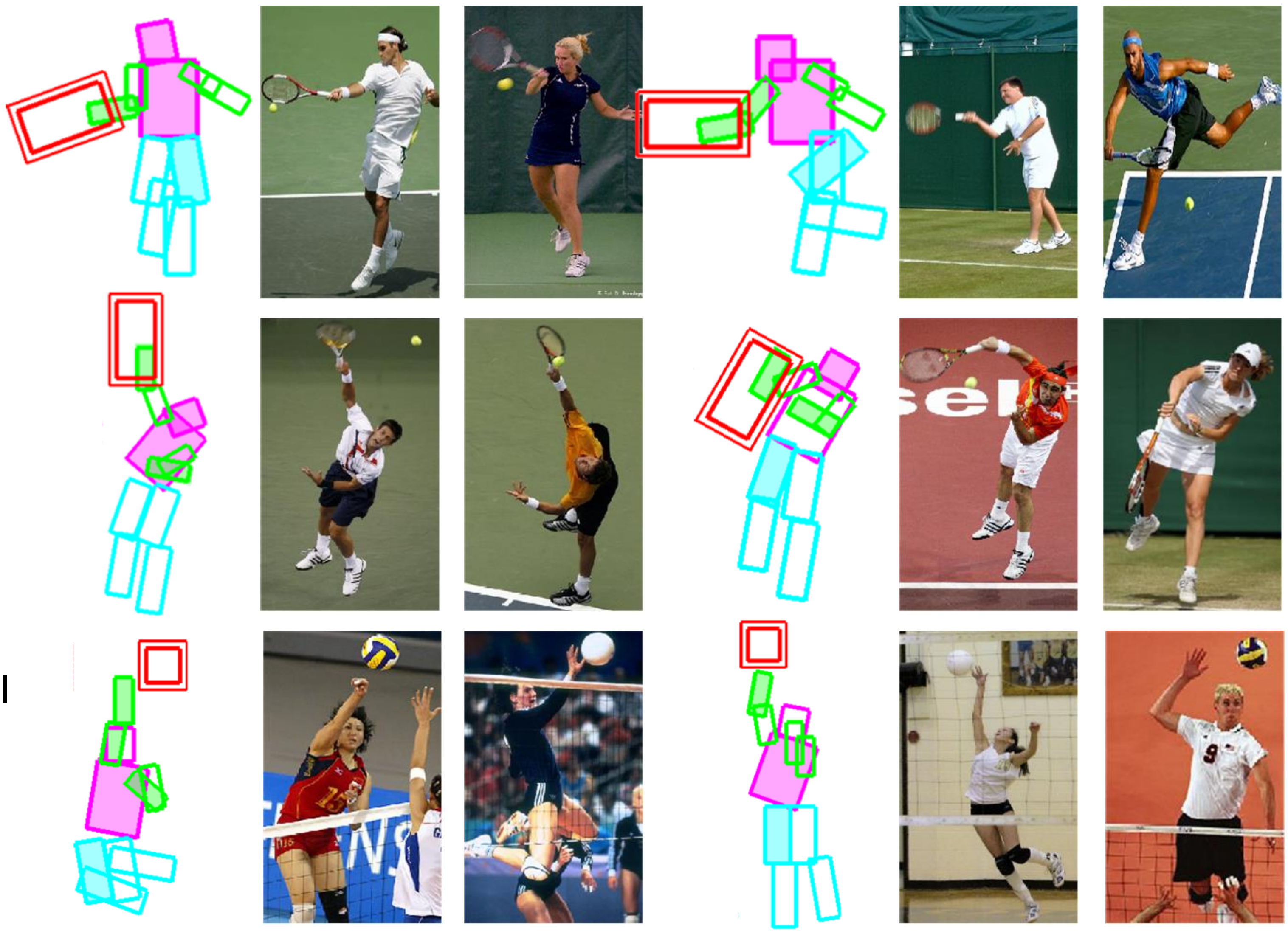
Slide Credit: Yao/Fei-Fei

Learning Results

Tennis forehand

Tennis serve

Volleyball smash

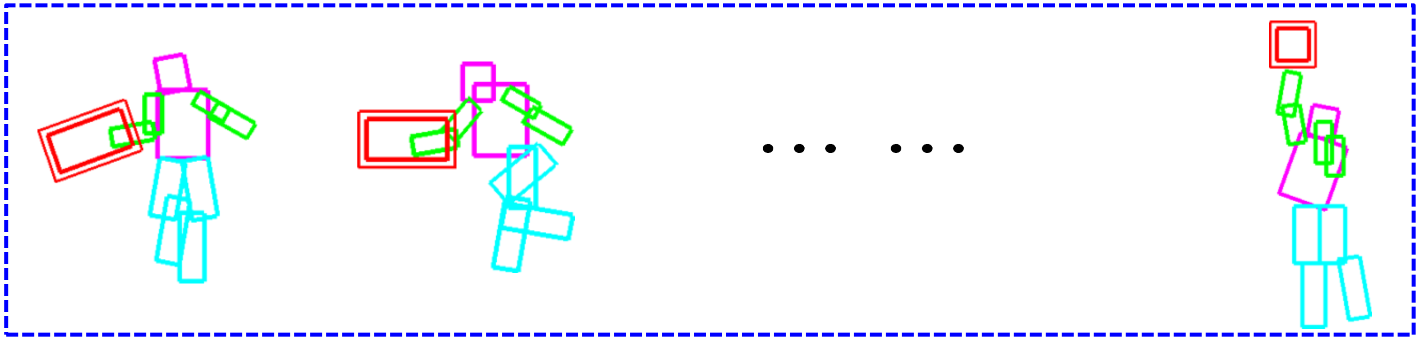


Slide Credit: Yao/Fei-Fei

Model Inference

I

The learned models

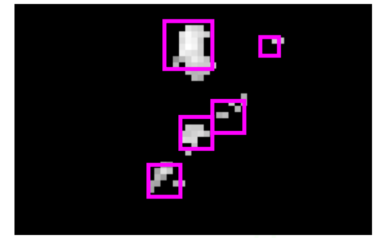


Slide Credit: Yao/Fei-Fei

Model Inference

I

The learned models



Head detection

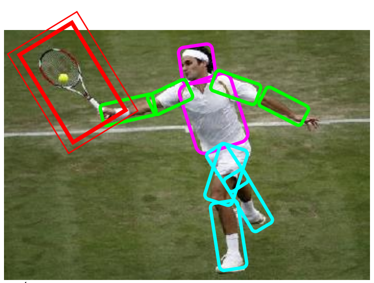
Torso detection

⋮

Tennis racket detection

Compositional Inference

[Chen et al, 2007]

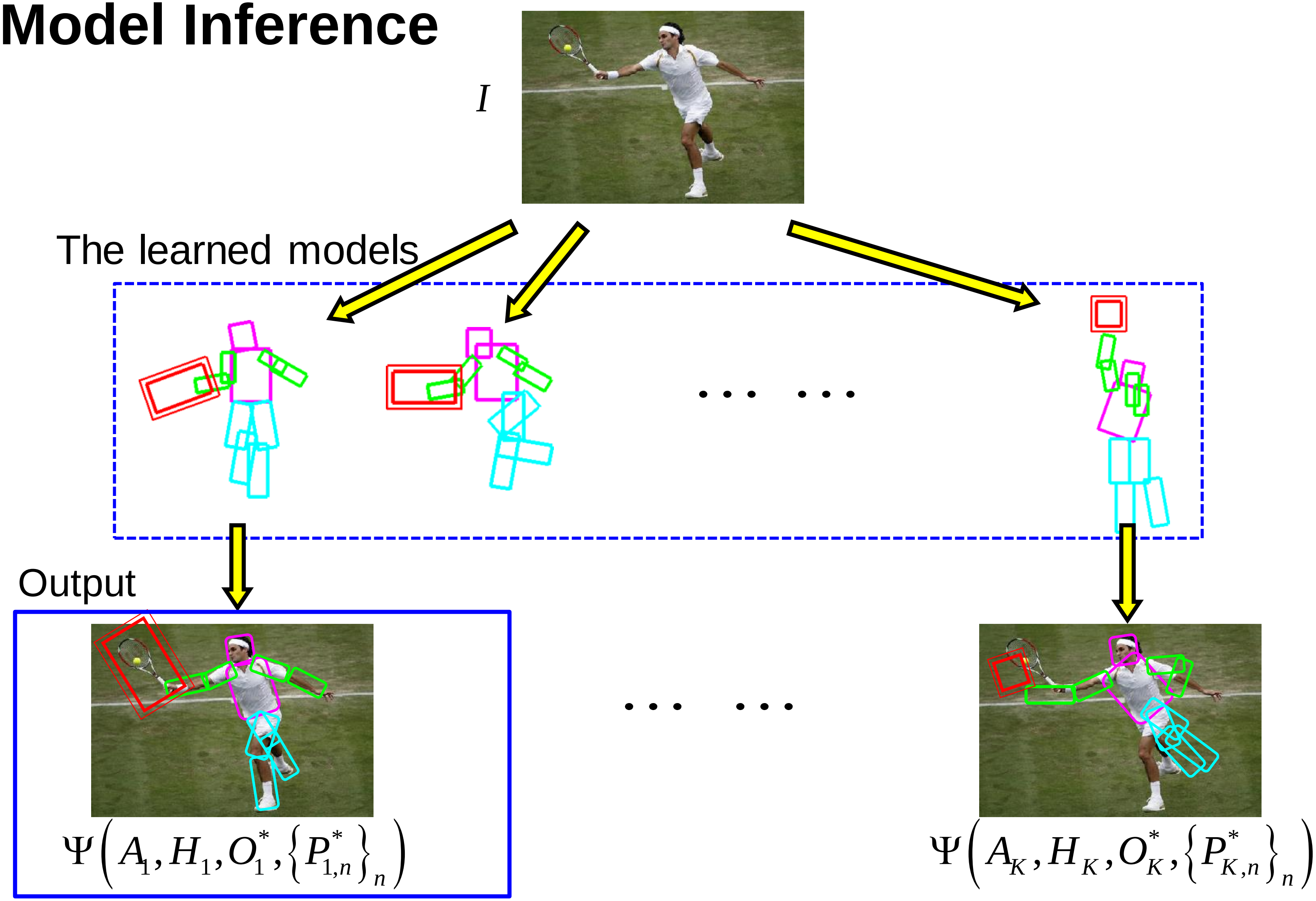


$$\Psi\left(A_1, H_1, O_1^*, \left\{P_{1,n}^*\right\}_n\right)$$

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

Slide Credit: Yao/Fei-Fei

Model Inference

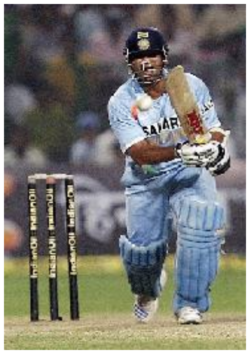


Slide Credit: Yao/Fei-Fei

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]

<u>Tasks:</u>

- 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

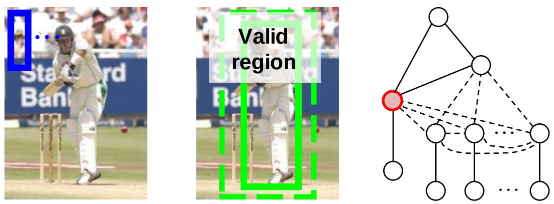
[Gupta et al, 2009]

Tasks:

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

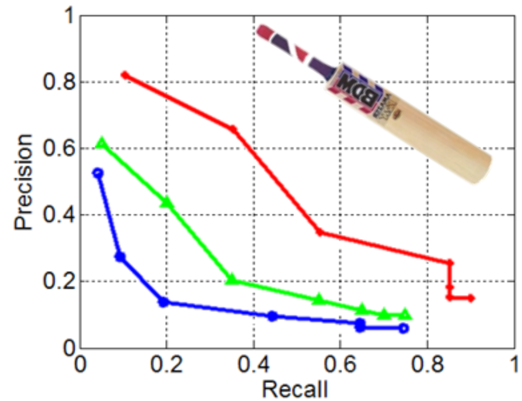
Slide Credit: Yao/Fei-Fei

Object Detection Results

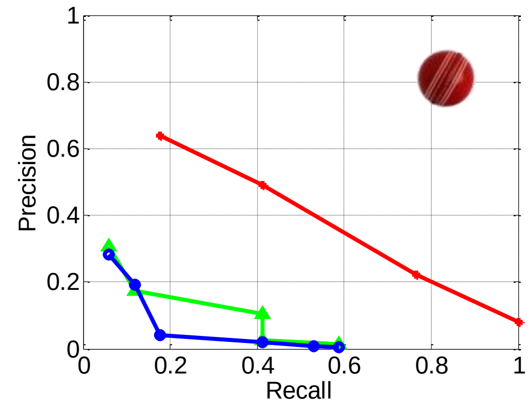


Sliding window	Pedestrian context	Our Method
[Andriluka et al, 2009]	[Dalal & Triggs, 2006]	

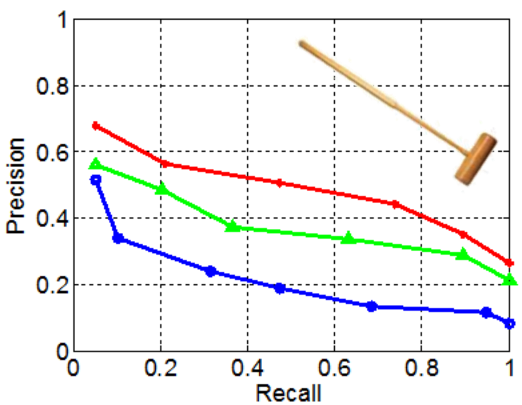
Cricket bat



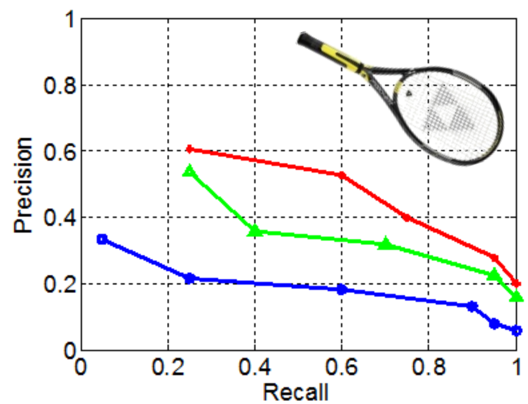
Cricket ball



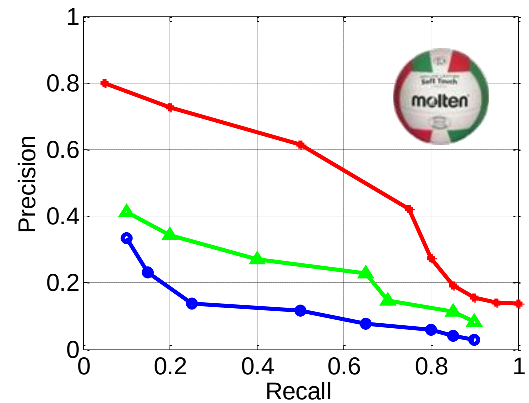
Croquet mallet



Tennis racket



Volleyball



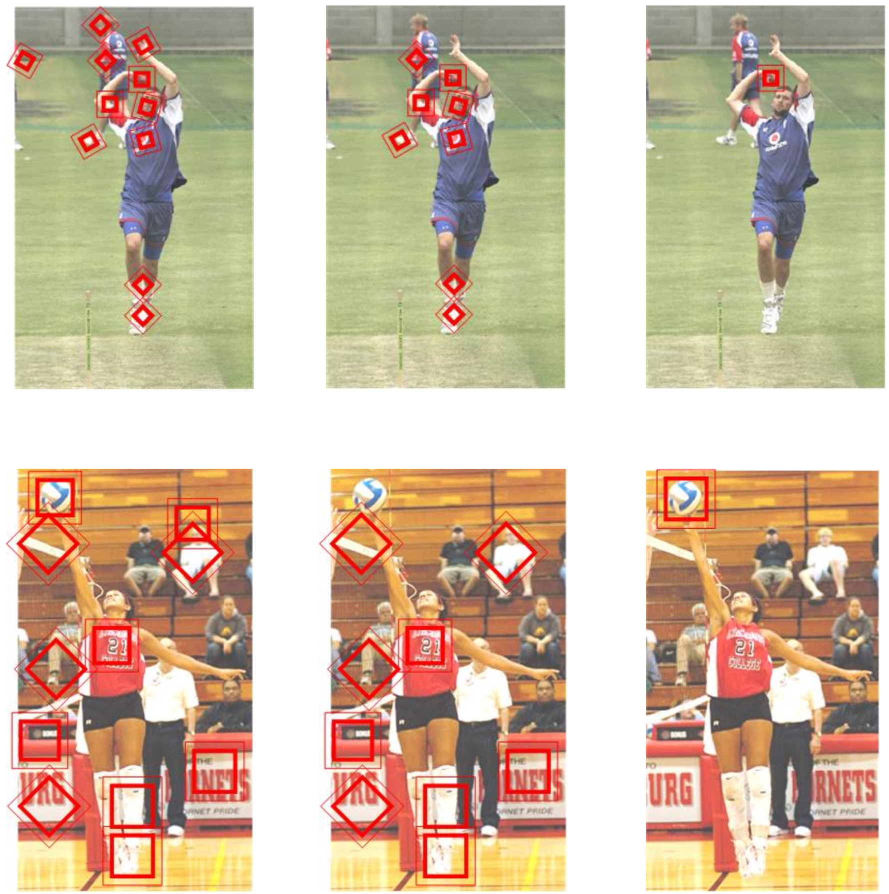
Slide Credit: Yao/Fei-Fei

Object Detection Results

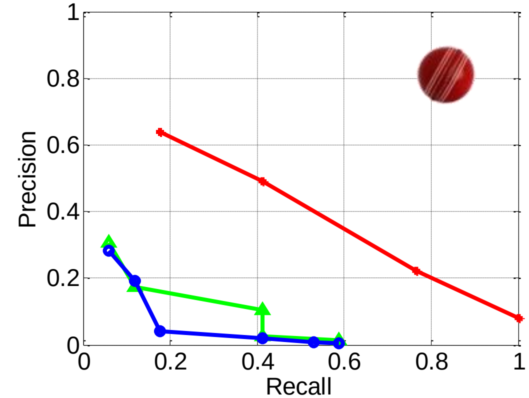
Sliding window | Pedestrian context | Our method

Small object

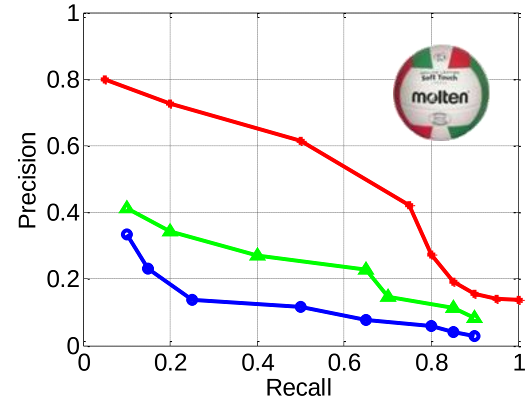
Background clutter



Cricket ball



Volleyball

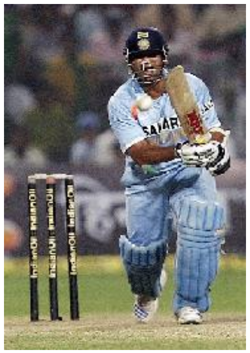


Slide Credit: Yao/Fei-Fei

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

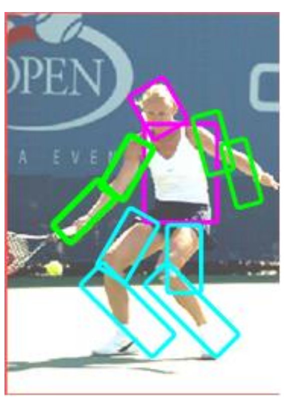
Method	Torso	Upper Leg		Lower Leg		Upper Arm		Lower Arm		Head
Ramanan, 2006	.52	.22	.22	.21	.28	.24	.28	.17	.14	.42
Andriluka et al, 2009	.50	.31	.30	.31	.27	.18	.19	.11	.11	.45
Our full model	**.66**	**.43**	**.39**	**.44**	**.34**	**.44**	**.40**	**.27**	**.29**	**.58**

Slide Credit: Yao/Fei-Fei

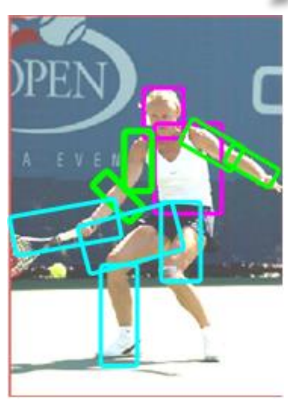
Human Pose Estimation Results

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

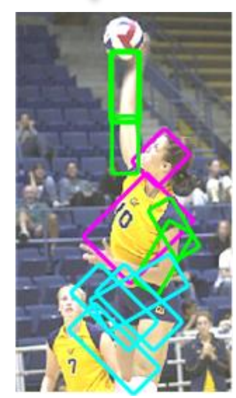


Our estimation result



Andriluka et al, 2009

Volleyball smash model



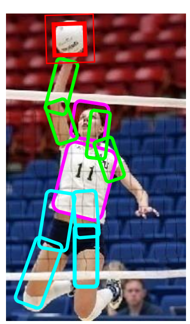
Our estimation result

Andriluka et al, 2009

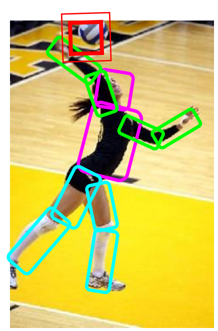
Slide Credit: Yao/Fei-Fei

Human Pose Estimation Results

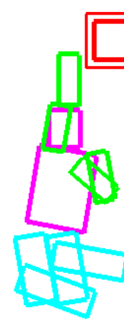
Method	Torso	Upper Leg		Lower Leg		Upper Arm		Lower Arm		Head
Ramanan, 2006	.52	.22	.22	.21	.28	.24	.28	.17	.14	.42
Andriluka et al, 2009	.50	.31	.30	.31	.27	.18	.19	.11	.11	.45
Our full model	**.66**	**.43**	**.39**	**.44**	**.34**	**.44**	**.40**	**.27**	**.29**	**.58**
One pose per class	.63	.40	.36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

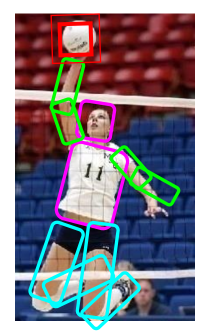


Estimation result

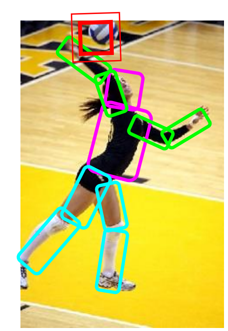


Estimation result





Estimation result



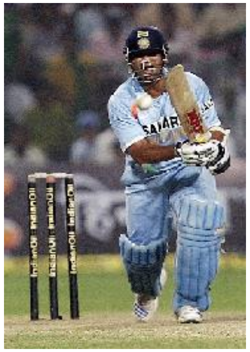
Estimation result

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

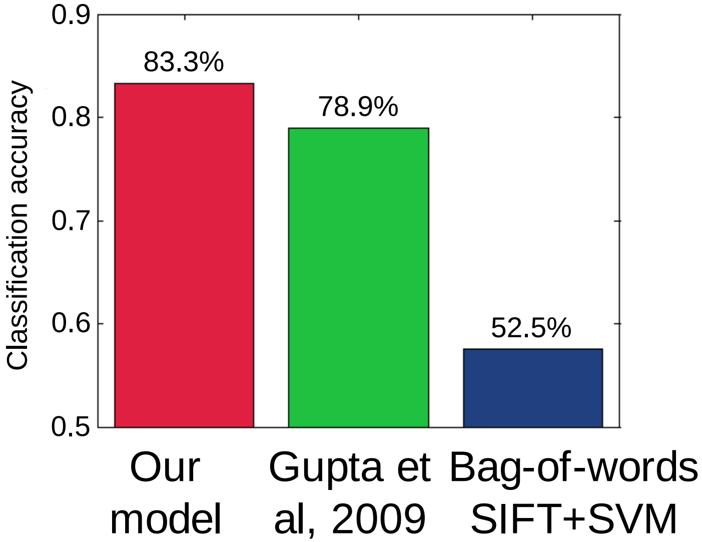
[Gupta et al, 2009]

Tasks:

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

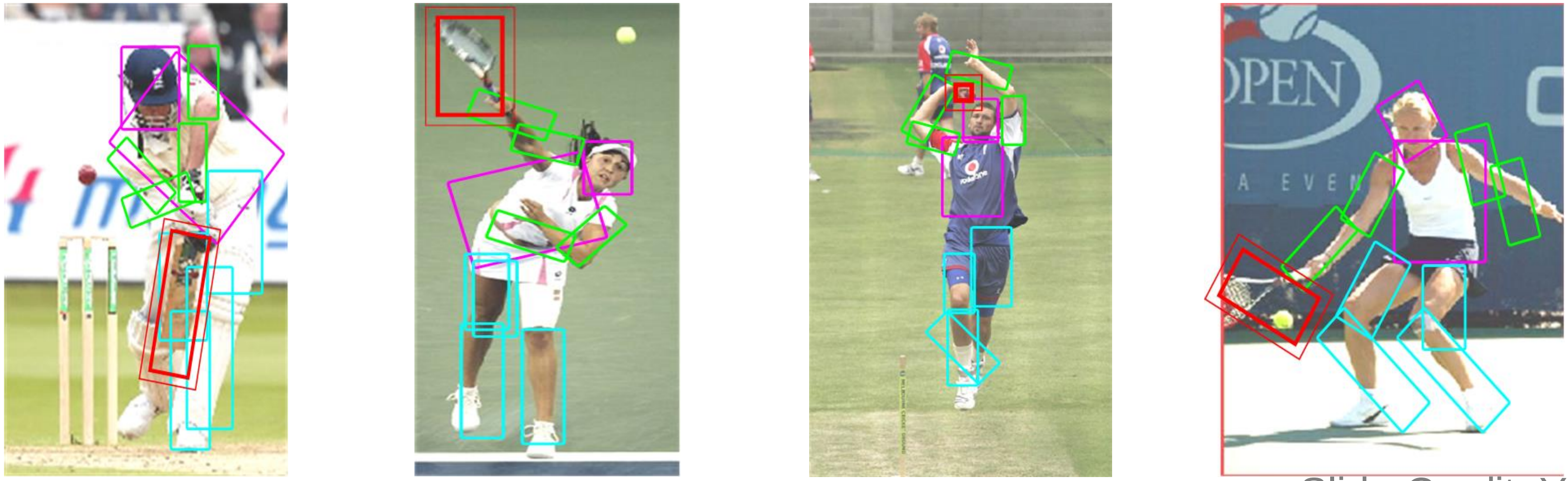
Slide Credit: Yao/Fei-Fei

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 (could be framed in terms of effect or intent)
  - Just a couple works on how to incorporate pose and objects (though recently more datasets getting into this)
  - Not much idea of how to reason about long-term activities or to describe video sequences

Next class: 3D Scenes and Context