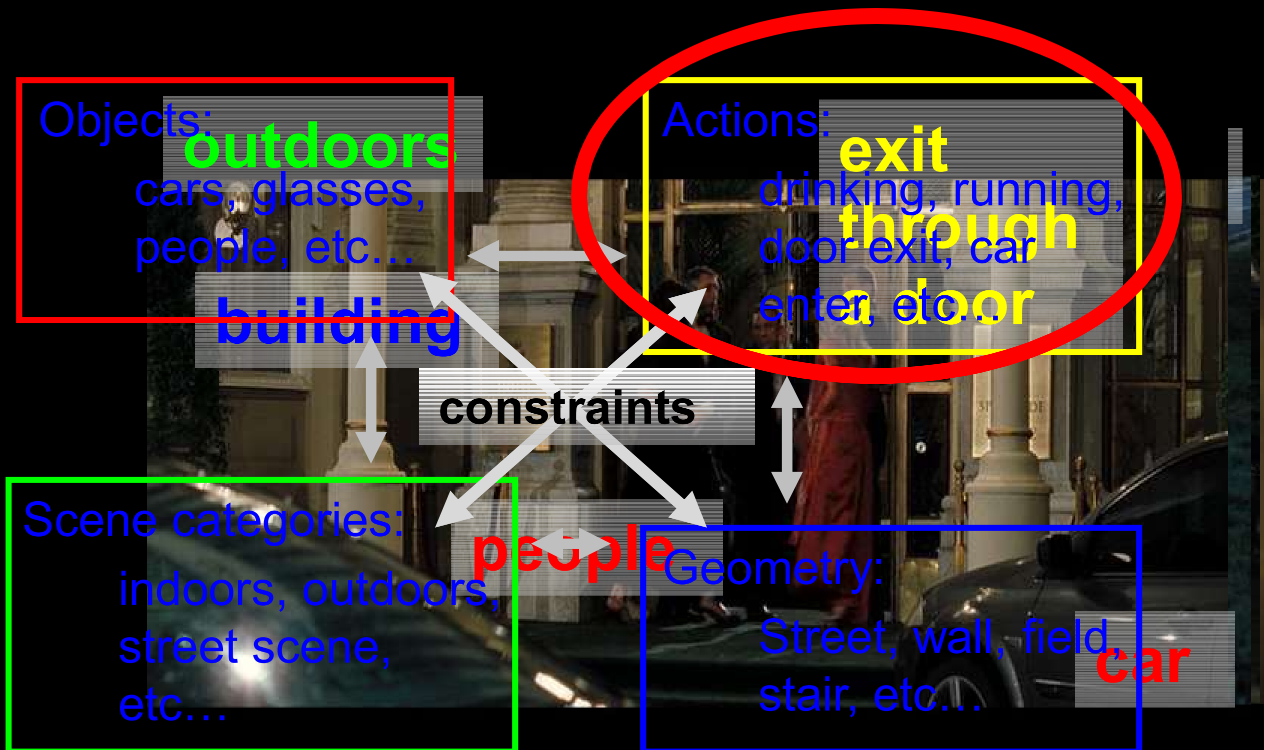
Action Recognition

CS 554 – Computer Vision
Pinar Duygulu
Bilkent University

(Slide credit: Nazli Ikizler-Cinbis)

Computer vision grand challenge: Video understanding



Slide credit I.Laptev

Why analyzing people and human actions?

How many person pixels are in video?

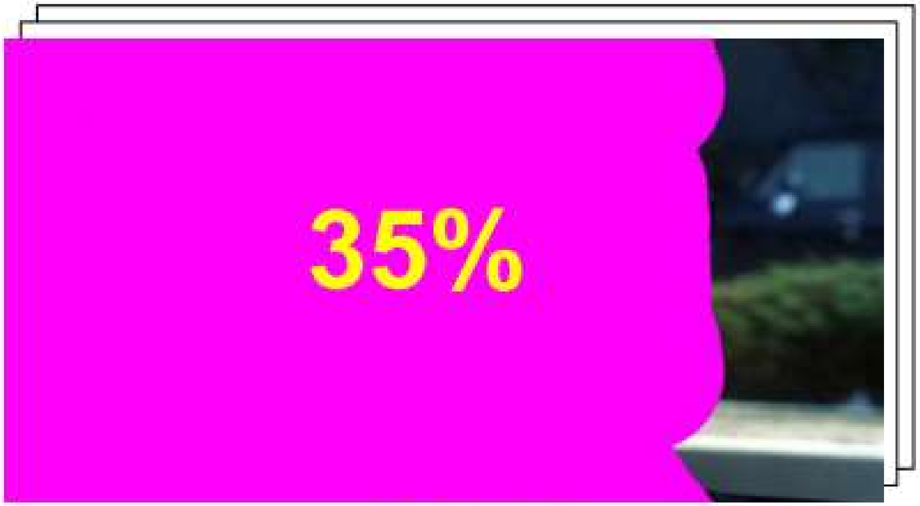
Movies

TV

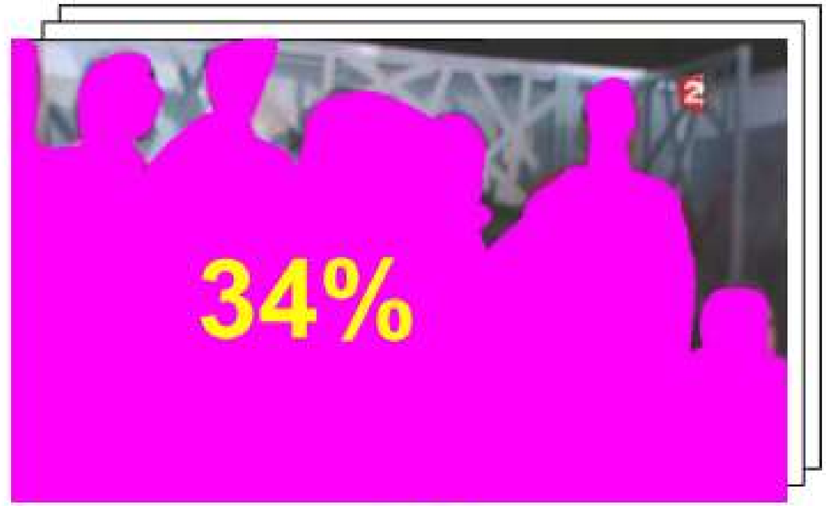
YouTube

Slide credit I.Laptev

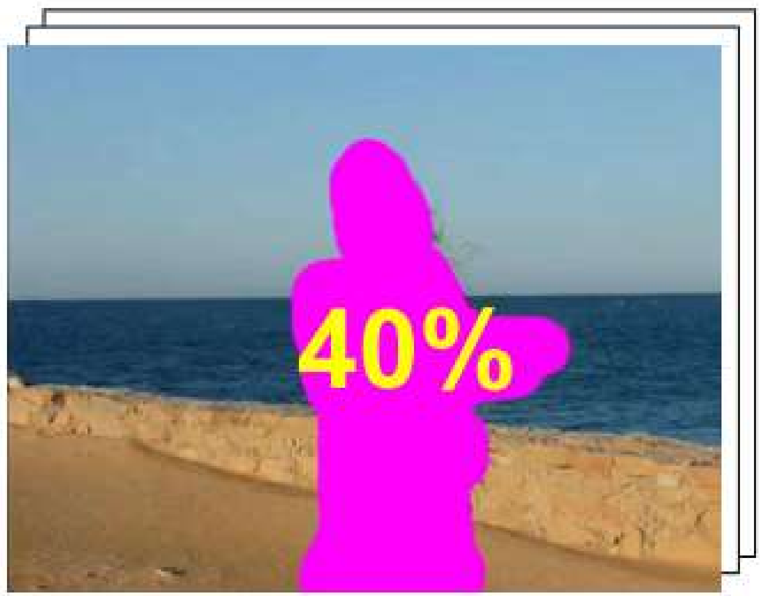
How many person pixels are in video?



Movies



TV



YouTube

Slide credit I.Laptev

Applications: Video editing

Alexei A. Efros, Alexander C. Berg, Greg Mori, Jitendra Malik, “Recognizing Action at a Distance”
ICCV 2003

Applications

- Analyzing video archives

First appearance of
N. Sarkozy on TV

Sociology research:
Influence of character
smoking in movies

Education: How do I
make a pizza?

- Surveillance

Where is my cat?

Predicting crowd behavior
Counting people

- Graphics

Motion capture and animation

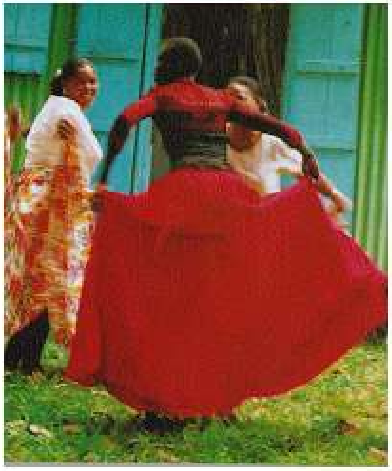
Slide credit I.Laptev

Definition: Act, Action and Activity

- Act: Short-timescale movements like a *forward-step* or a *hand-raise*
- Action: Medium timescale movements like *walking*, *running*, *jumping*
  - Typically composites of multiple acts
- Activity: Long timescale movements (e.g., interactions between people)
  - Complex composites of actions
  - Composition can be
    - across time
    - across body
- Event: combination of activities or actions (e.g., a football game, a traffic accident)

Problems

The appearance/size/shape of people can vary dramatically (high-D space).



Underlying structure (bones and joints) is *unobservable* (obscured by muscle, skin, clothing).

Occlusion and partial views.

Problems

Loss of 3D in 2D projection

Unusual poses

Self occlusion

Low contrast

Problems

Multiple people and occlusion leads to ambiguity.

Moving cameras & complex changing backgrounds.

Problems

Accidental alignment

Motion blur.

(nothing to match)

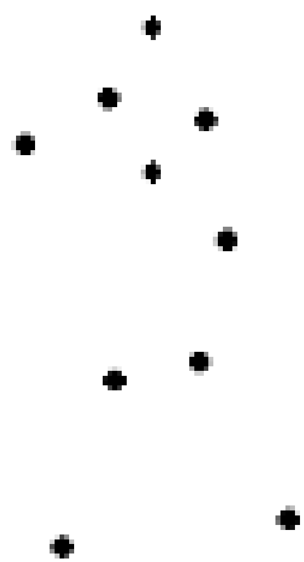
Human activity in video: basic approaches

- **Model-based action/activity recognition**:
  - Use human body tracking and pose estimation techniques, relate to action descriptions (or learn)
  - Major challenge: accurate tracks in spite of occlusion, ambiguity, low resolution

- **Activity as motion, space-time appearance patterns**
  - Describe overall patterns, but no explicit body tracking
  - Typically learn a classifier
  - *We'll look at some specific instances…*

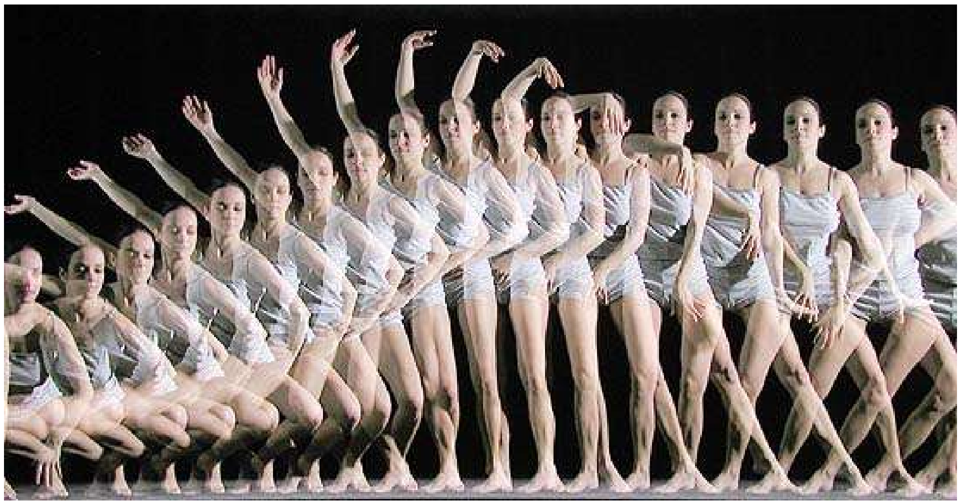
Motion and perceptual organization

- Even “impoverished” motion data can evoke a strong percept



How can we identify actions?

Motion



Pose

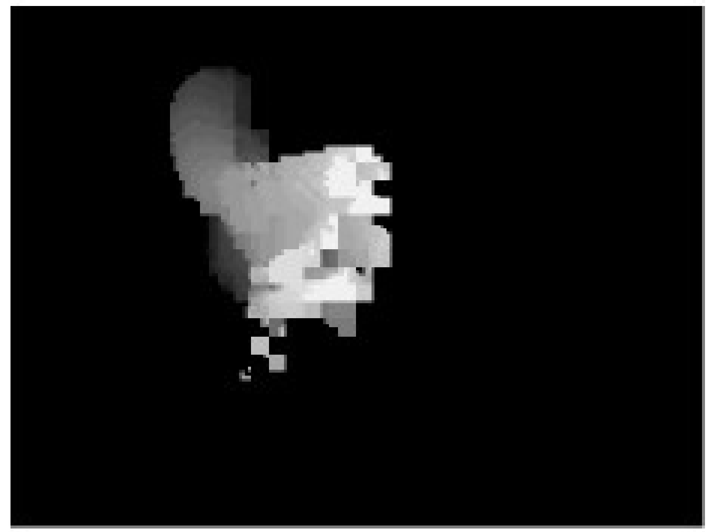
Held Objects

Nearby Objects

Representing Motion

Optical Flow with Motion History

sit-down

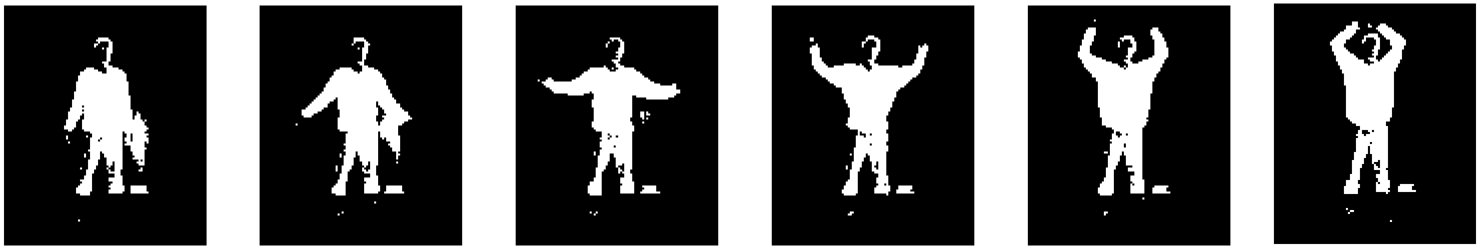


sit-down MHI

Bobick Davis 2001

Appearance based methods: Global Shape

$D(x, y, t) \quad t = 1, ..., T$



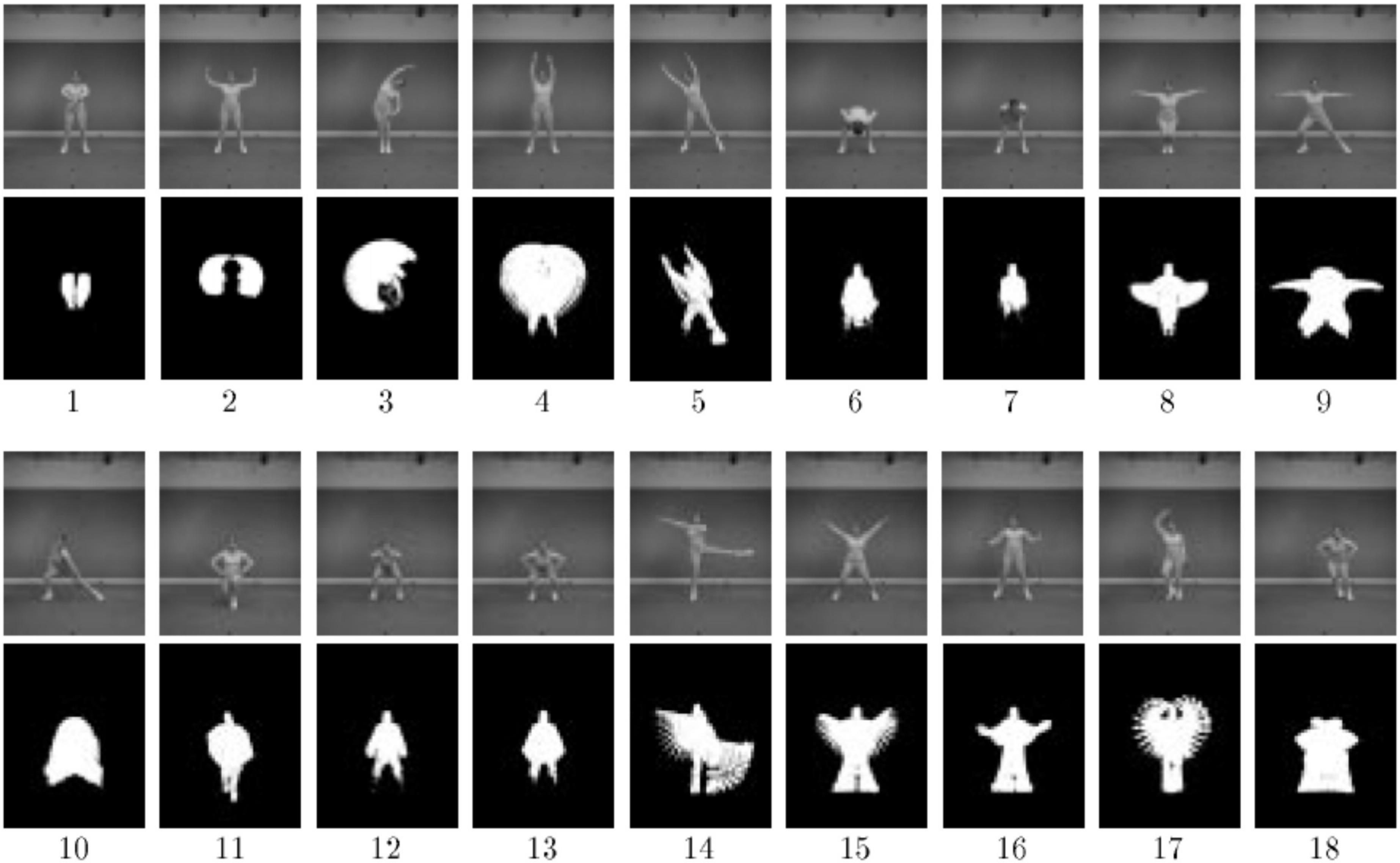
Idea: summarize motion in video in a *Motion History Image (MHI)*:

$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ \max\ (0, H_\tau(x, y, t-1) - 1) & \text{otherwise} \end{cases}$$

Nearest Neighbor action classification with Mahalanobis distance between training and test descriptors *d*.

Aaron F. Bobick and James W. Davis, "The Recognition of Human Movement Using Temporal Templates", PAMI 2001

Appearance Templates at Aerobics Dataset



Temporal Global Templates

Pros:

+ Simple
+ Fast

Cons:

- Assumes static camera, static background
- Sensitive to segmentation errors
- Silhouettes do not capture interior motion/shape
- Needs lots of examples for each variation

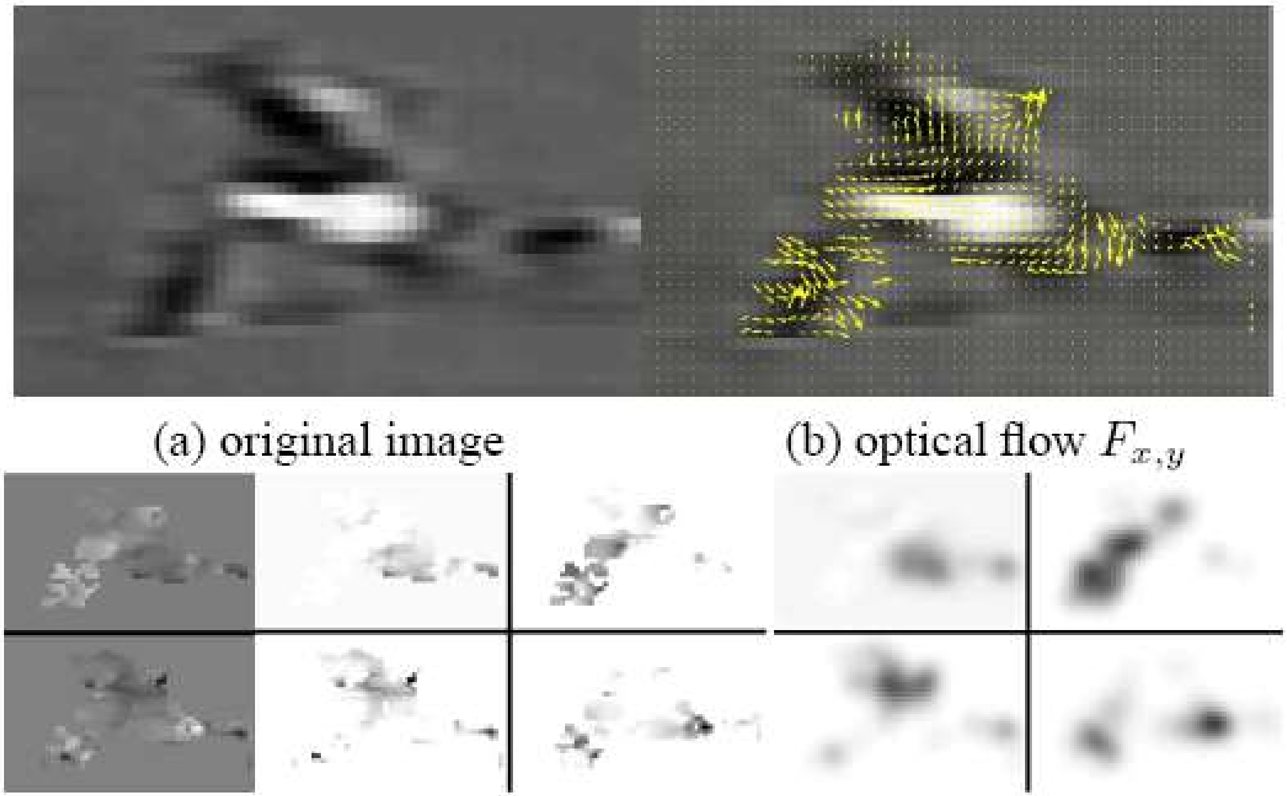
Possible improvements:

- Not all shapes are valid ⇨ Restrict the space of admissible shapes to overcome segmentation errors

Slide credit I.Laptev

Representing Motion

Optical Flow with Split Channels



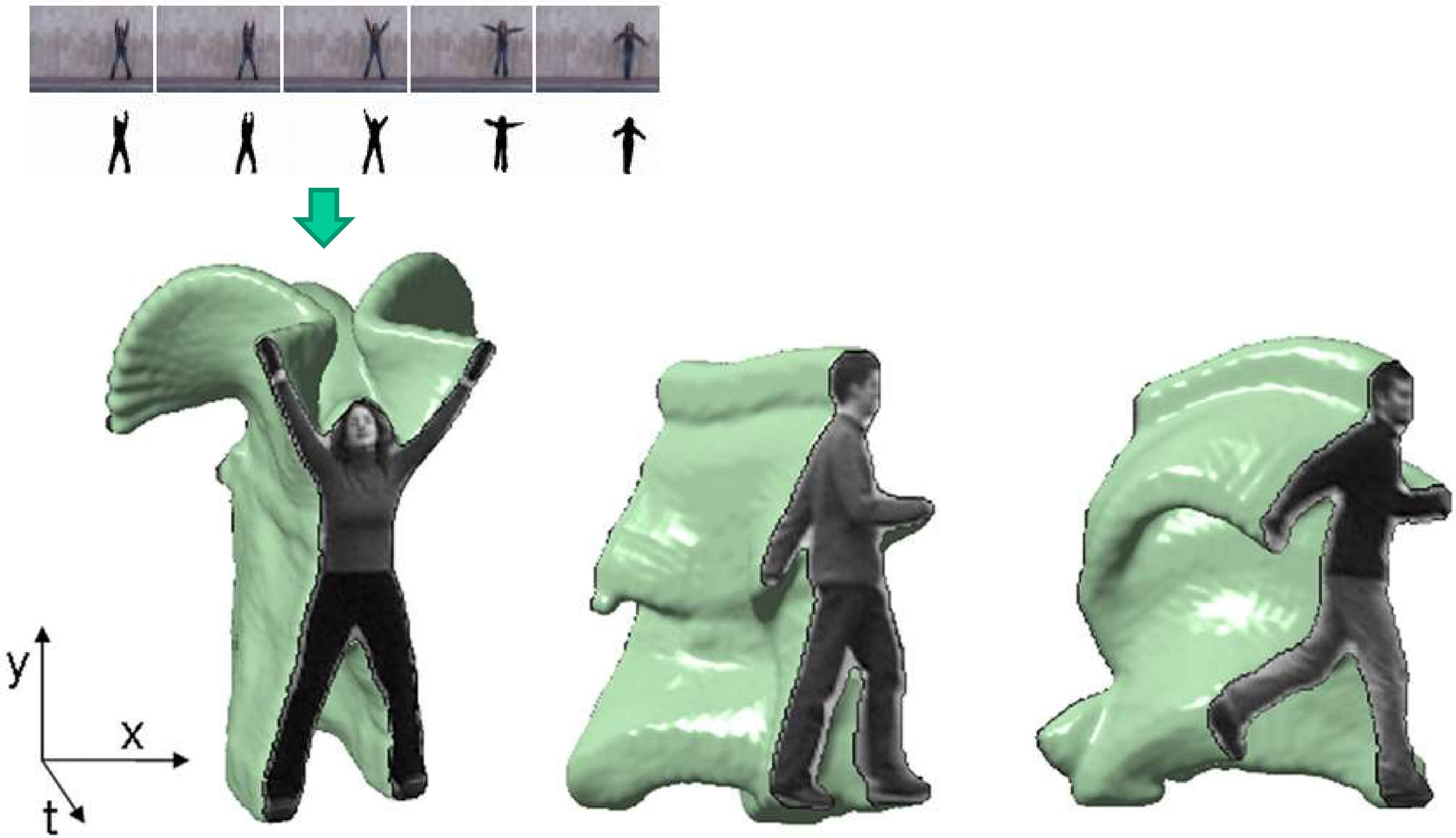
(a) original image (b) optical flow $F_{x,y}$

(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

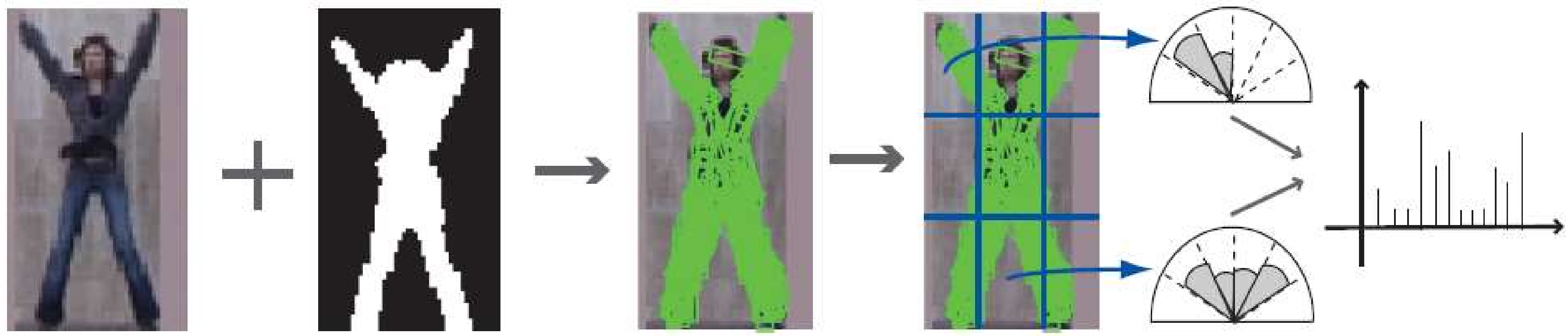
Representing Motion

Space-Time Volumes



Blank et al. 2005

Histogram of Oriented Rectangles (HoR)

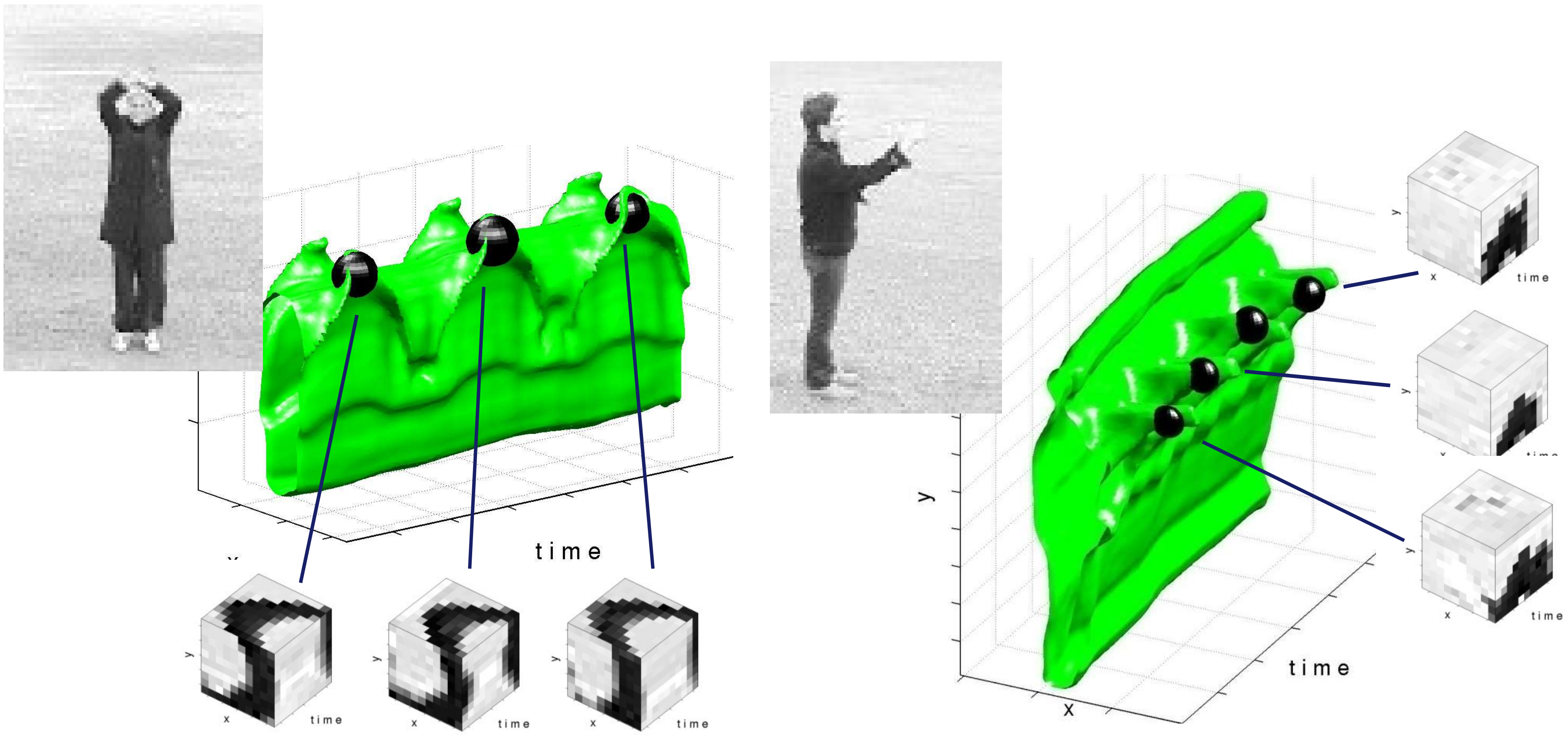


- Body can be thought as a collection of rectangular regions
- We can represent the pose based on the orientation of these rectangles
  - Tracker finds the human subject
  - Extract the silhouettes
  - Rectangular regions are extracted using convolution of a zero-padded rectangular 2D Gaussian on different orientations and scales
    - 12 angles 15° apart

joint work with P. Duygulu, Human Motion Workshop, ICCV 2007

Space-time local features

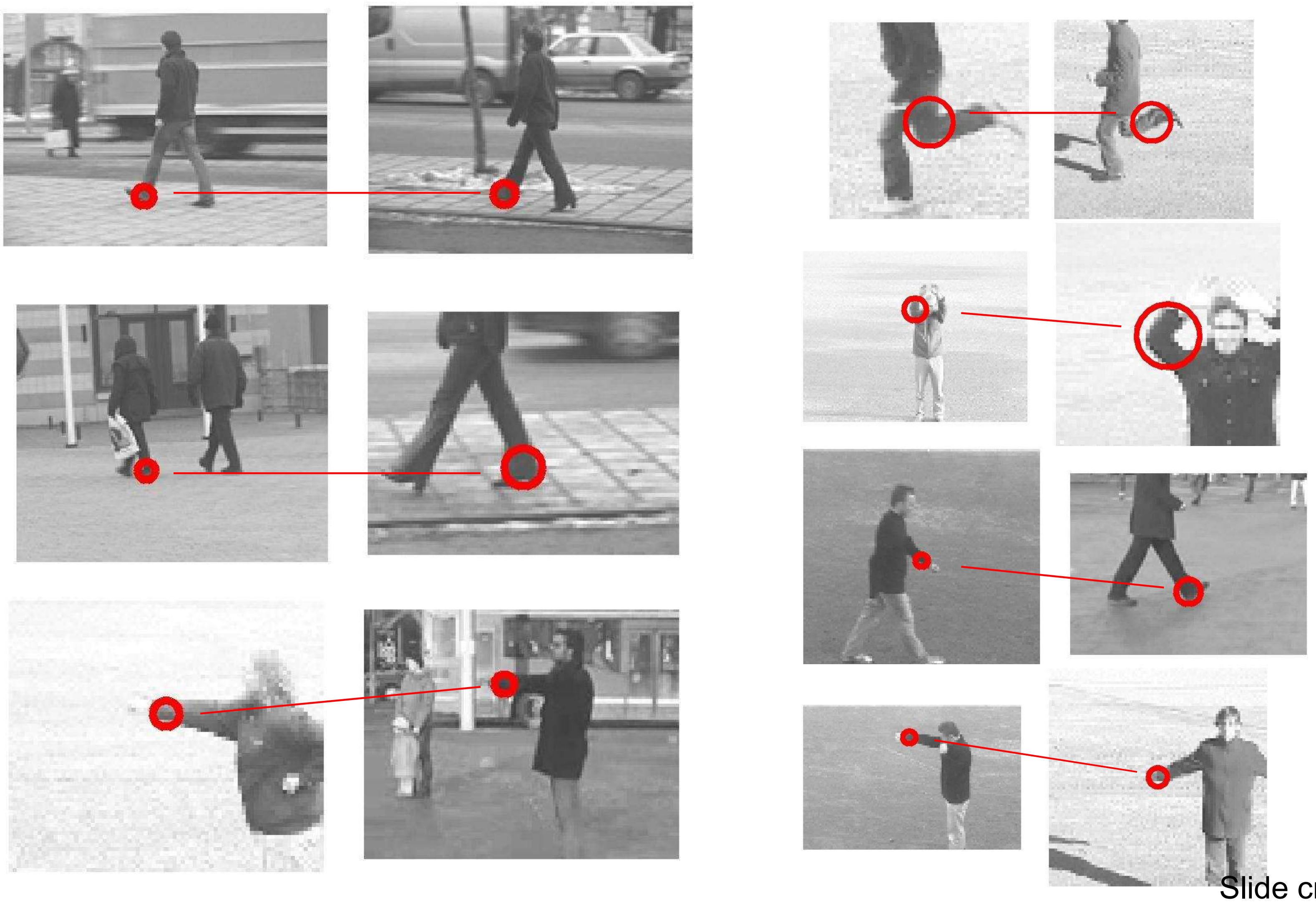
No Global assumptions => Consider local spatio-temporal neighborhoods



Slide credit I.Laptev

Local Space-time features: Matching

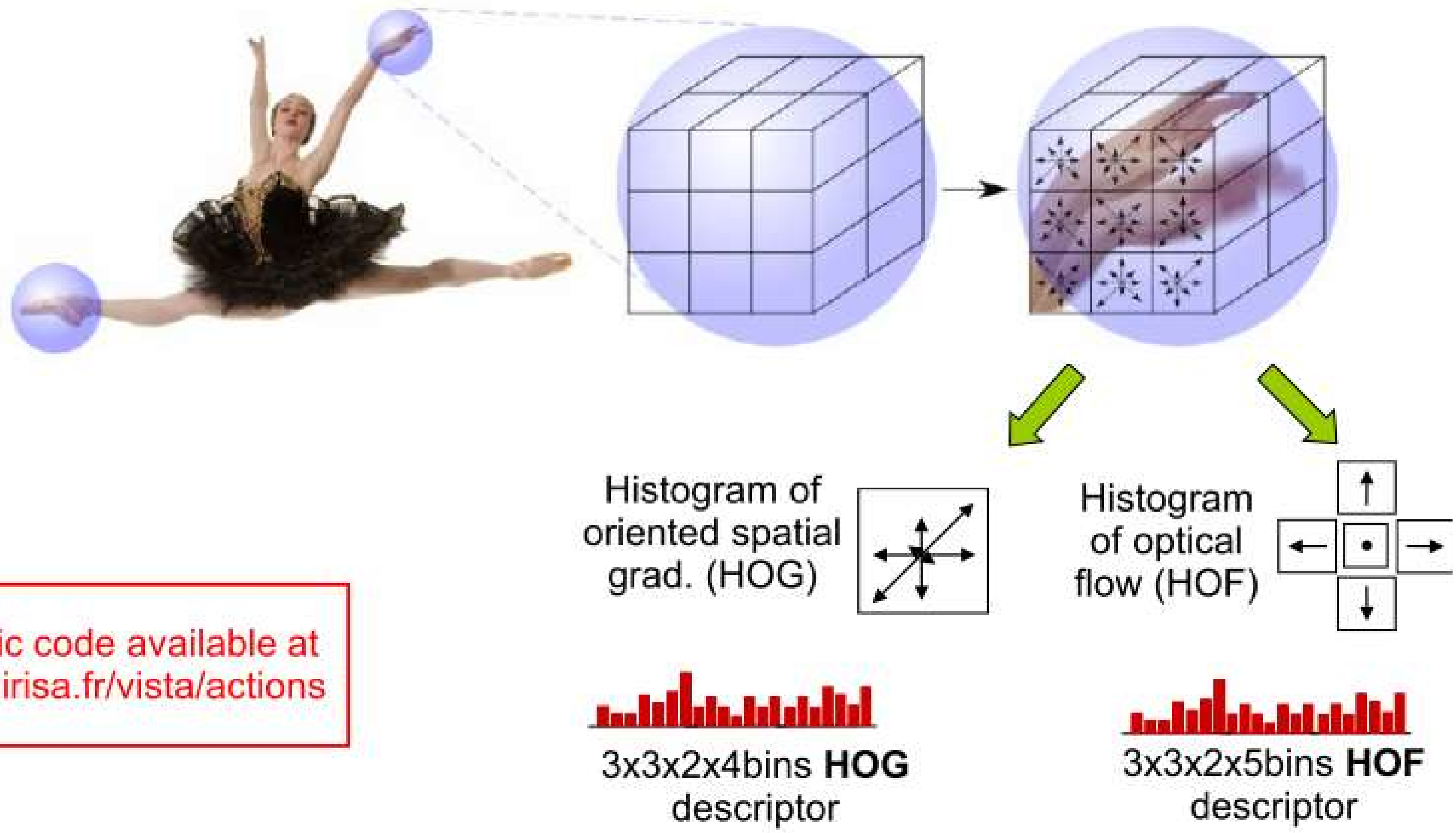
- Find similar events in pairs of video sequences



Slide credit I.Laptev

Local space-time descriptor: HOG/HOF

Multi-scale space-time patches



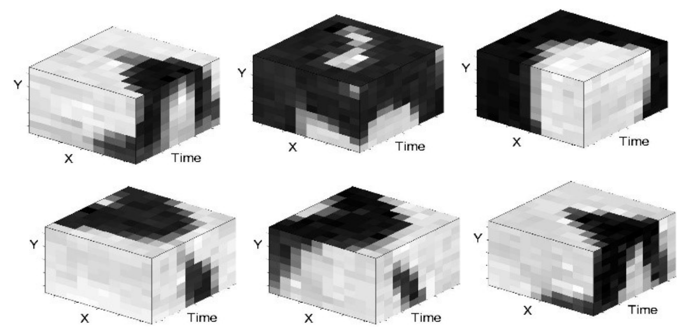
Public code available at
www.irisa.fr/vista/actions

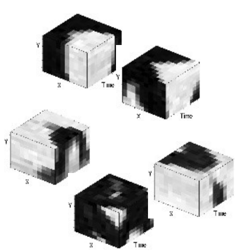
Action Classification with Spatio-temporal Words

Bag of space-time features + multi-channel SVM

[Laptev'03, Schuldt'04, Niebles'06, Zhang'07]

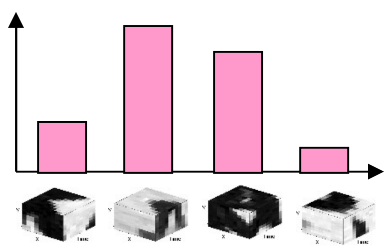
Collection of space-time patches





HOG & HOF patch descriptors

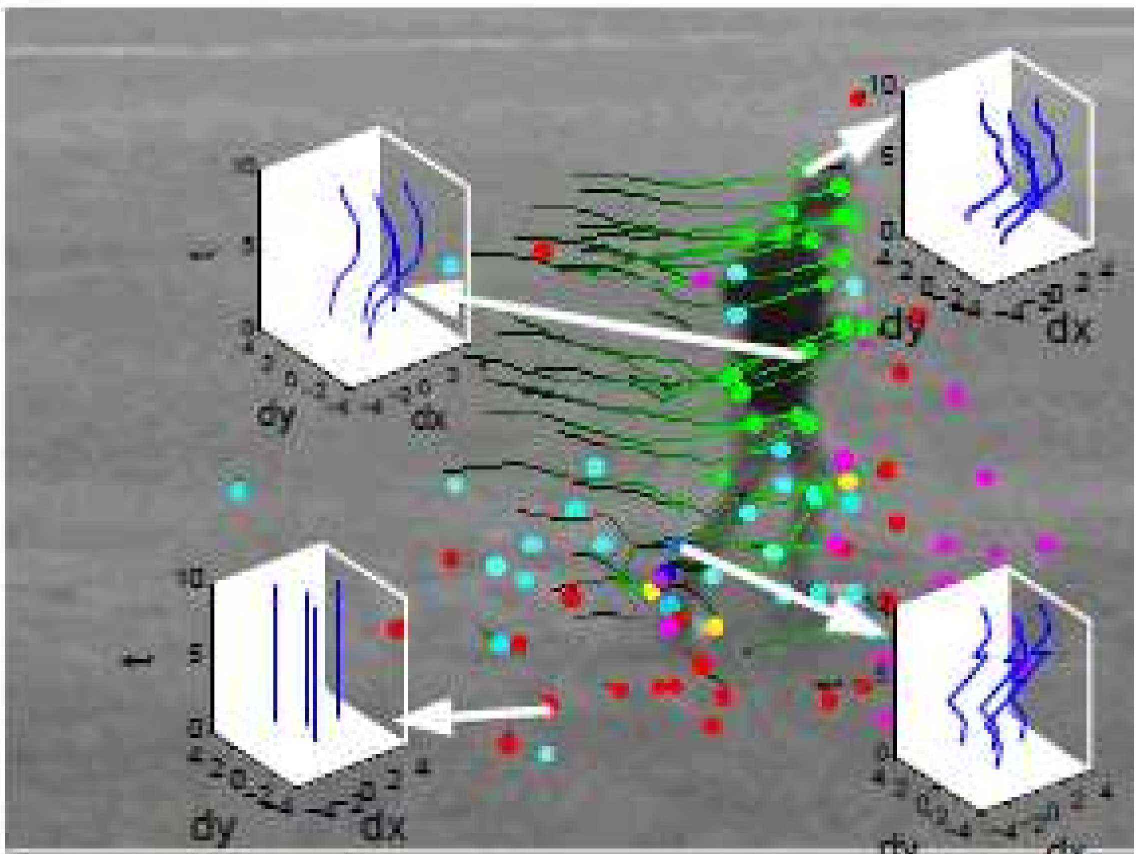
Histogram of visual words



Multi-channel SVM Classifier

Slide credit I.Laptev

Representing Motion: Tracked Points



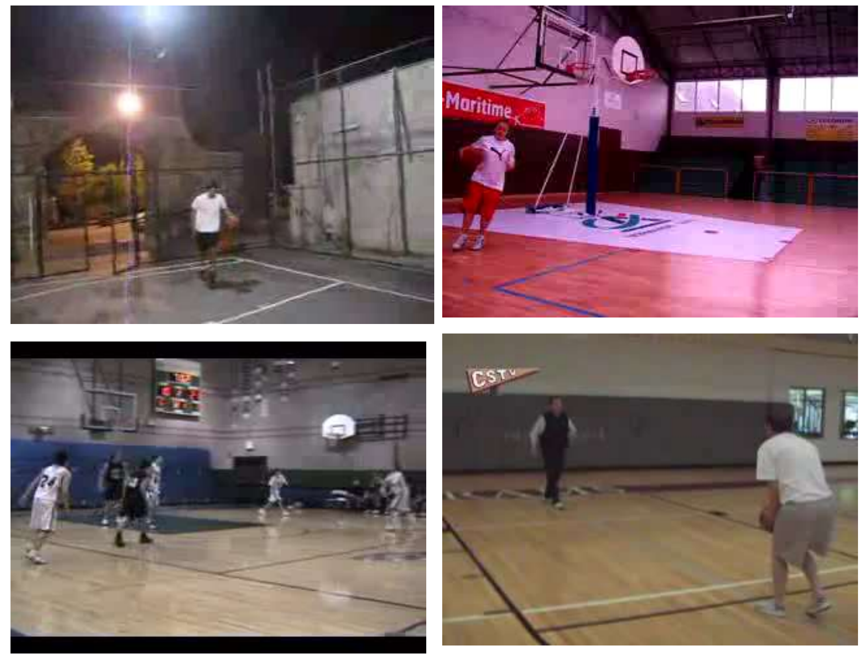
Matikainen et al. 2009

Things are much complex in real world:
Action recognition “*in the wild*”

- Complex activities
- Multiple people
- Cluttered backgrounds

Why is action recognition in uncontrolled videos difficult?

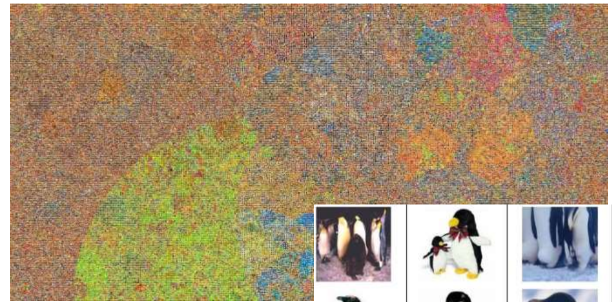
- Various challenges
  - Moving camera
  - Low resolution
  - Diverse appearance, viewpoints
  - Diverse dynamics



- Need for lots of training video
  - Different styles of action
  - Different viewpoints
  - Lots of different actions

Internet Vision

80 million tiny images – Fergus et al.



Schroff et al 2007

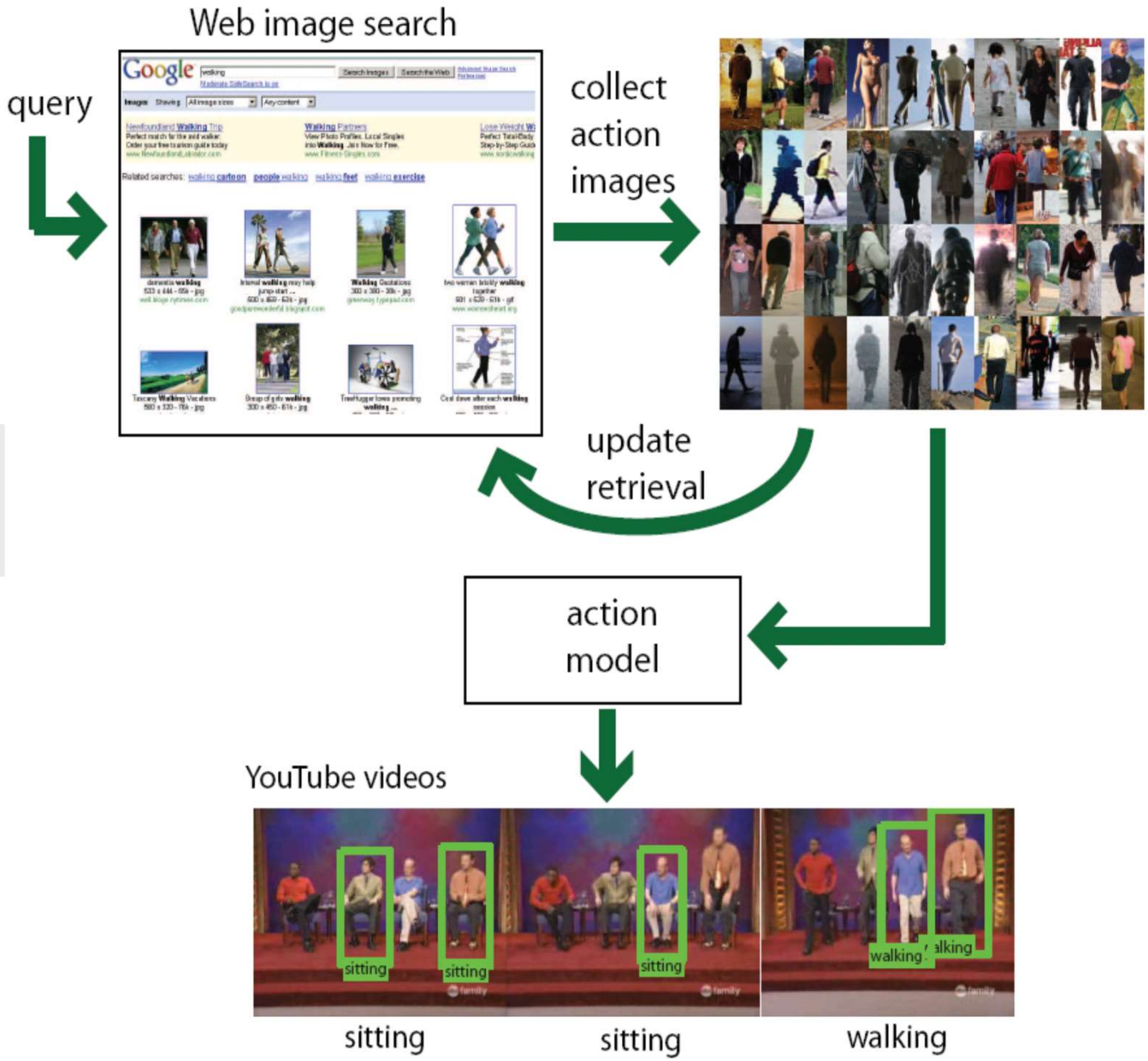
- Web is an enormous source of information
  - Recently used widely by object recognition community

- There are lots of "action images" – untouched!
  - Lots of data can help to capture the diverse nature of actions
  - Overcomes the training bias
    - Uncontrolled poses
    - Various people, clothing, body proportions, etc.

Idea

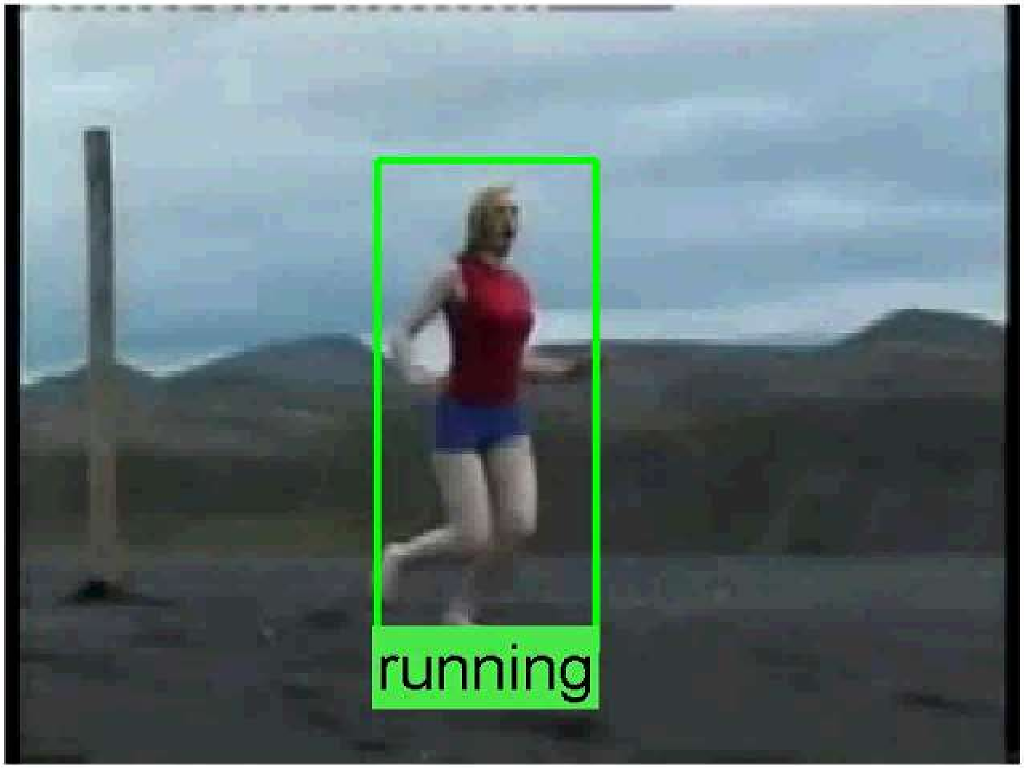
- Collect action images from the web
- Learn action pose models
- Use these models to annotate actions in videos
  - Classification by pose

Overall System



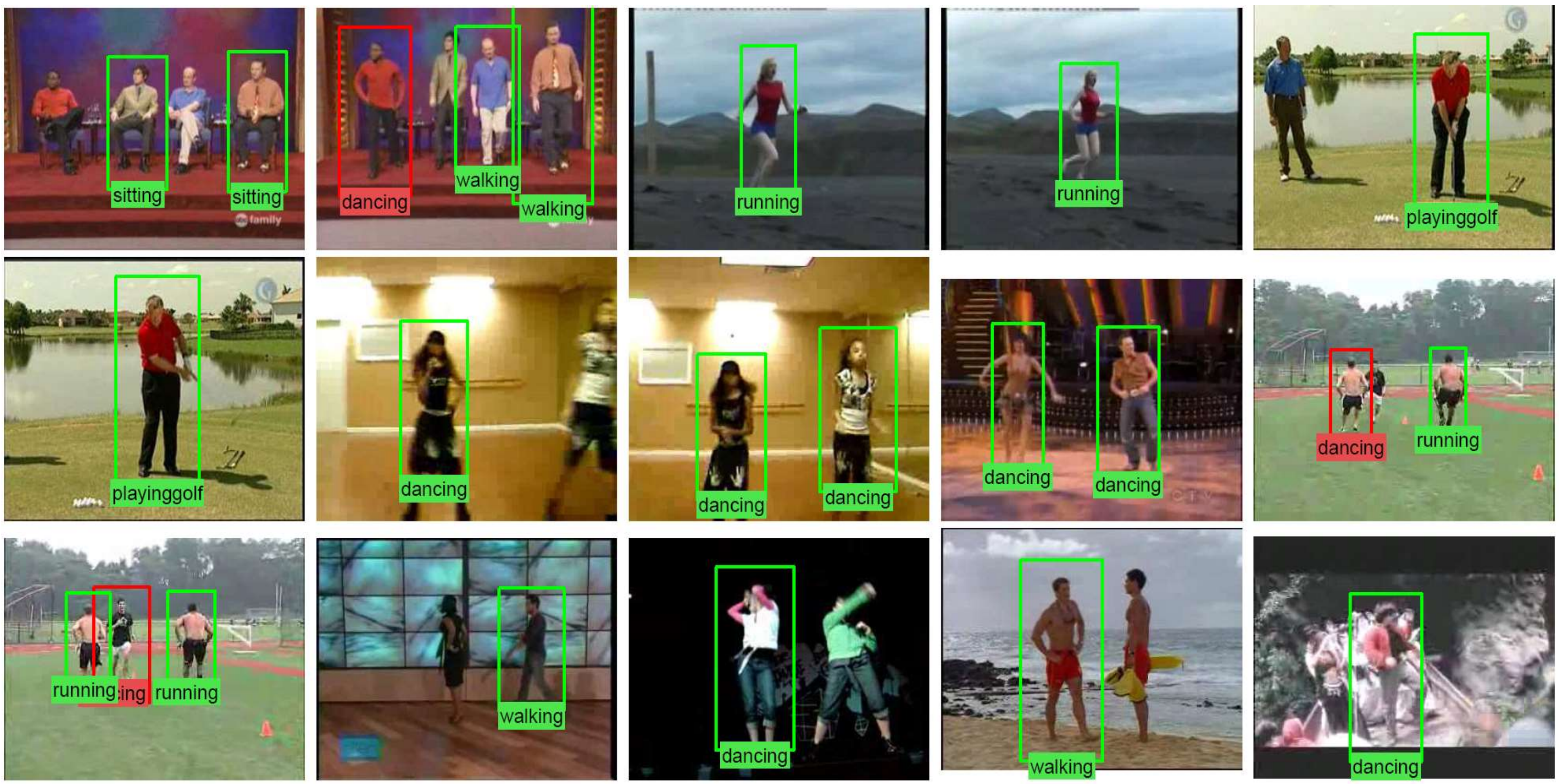
Some Results - I

Some Results - II

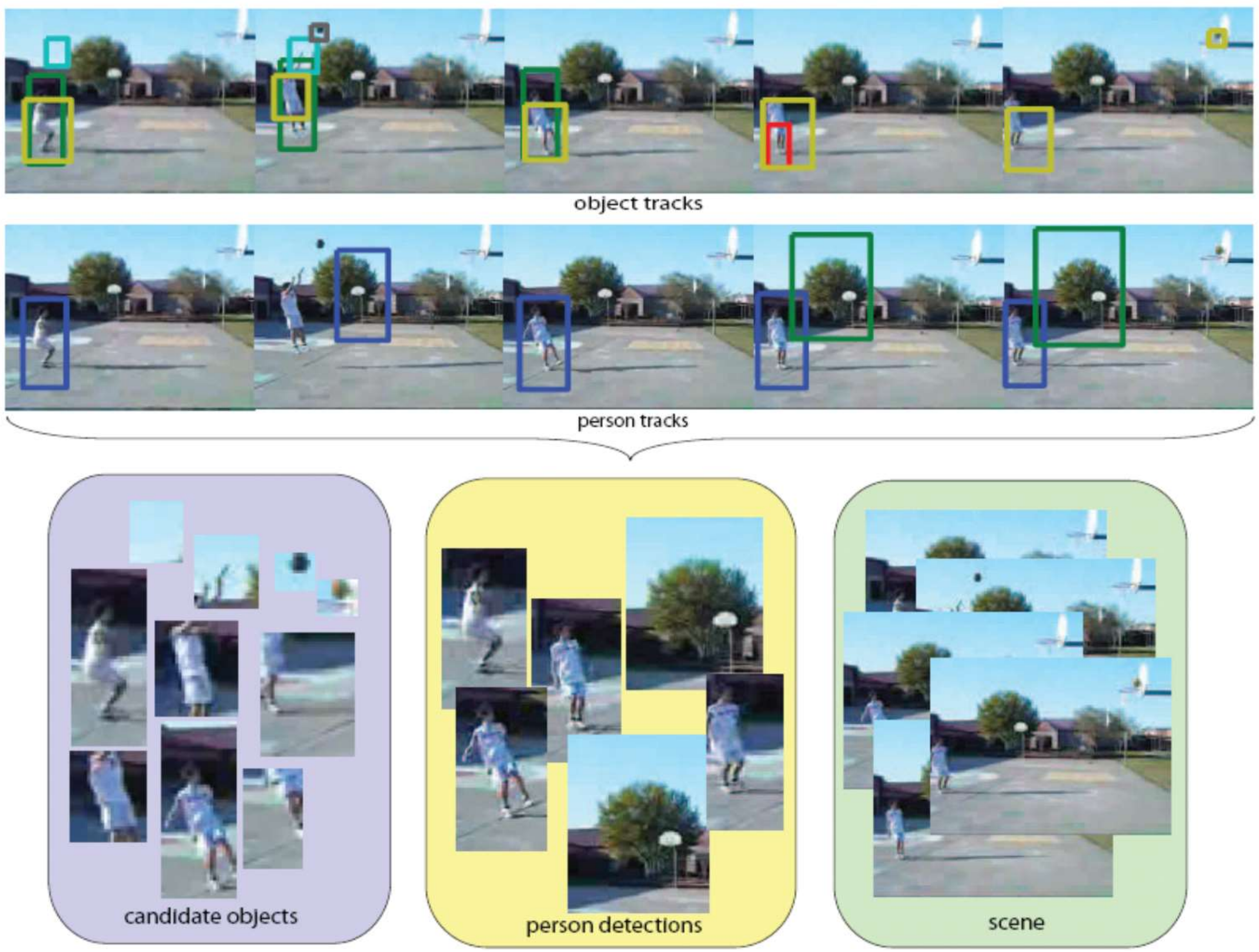


Some Results - III

Action Recognition In YouTube Videos



Objects, Scene and Actions



Joint work with Stan Sclaroff, ECCV 2010

Motivation

- The presence (or absence) of particular objects or scene properties can often be used to infer the possible subset of actions that can take place.
  - if there is a pool in the scene, then "diving" becomes a possible action.
  - if there is no pool, but a court, then the probability of the "diving" action reduces
  - if there is a basketball moving towards the hoop, there can be someone playing basketball

Problem/Approach

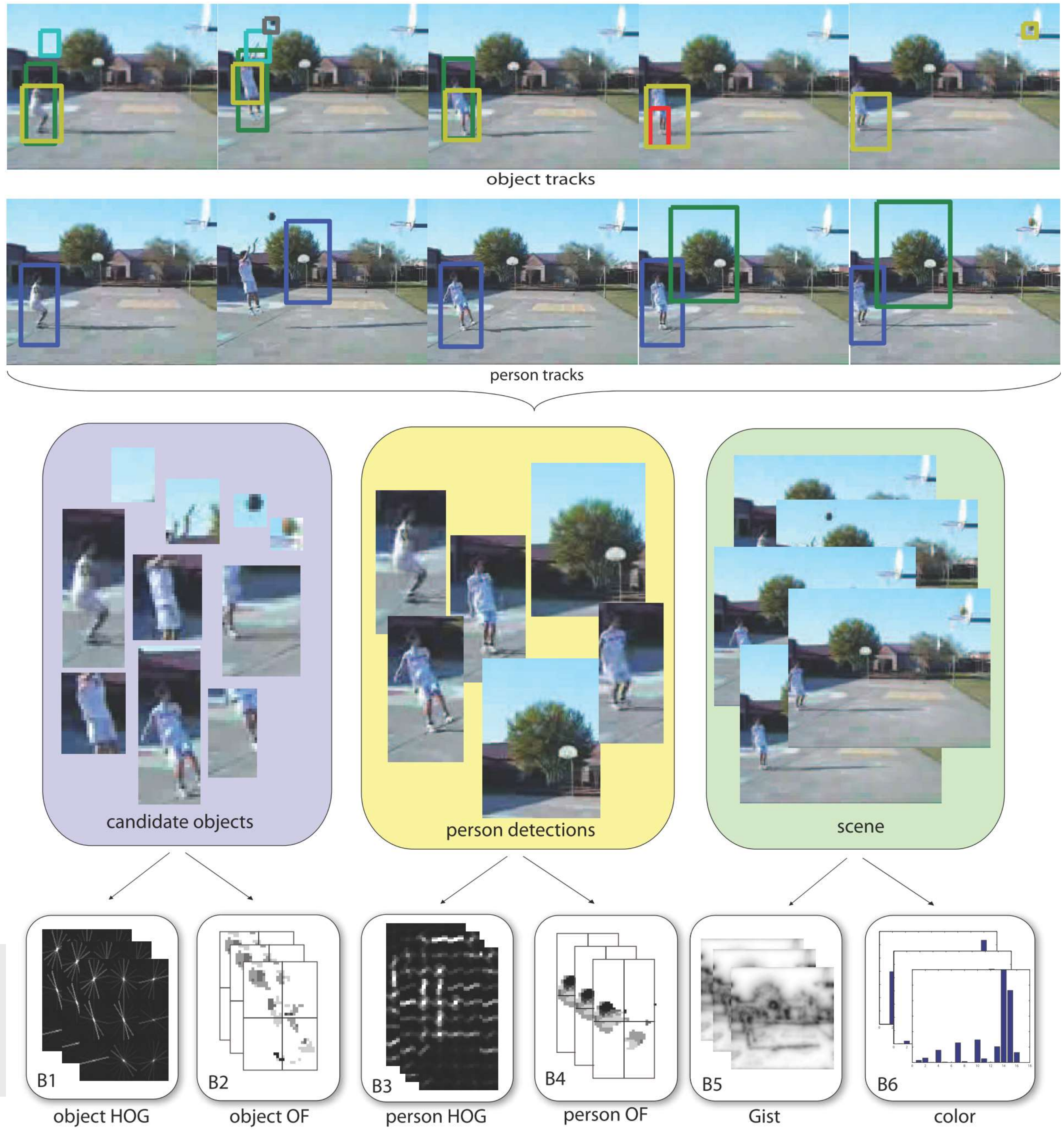
- P: Single features may not be solely reliable / discriminative
  - A: Extract many different (noisy) features complementary to each other
- P: Many non-relevant tracks, including other people not performing that action
  - A: Formulate the problem as Multiple Instance Learning and extend the positivity constraint of MIL to multiple bags

Extract moving object tracks

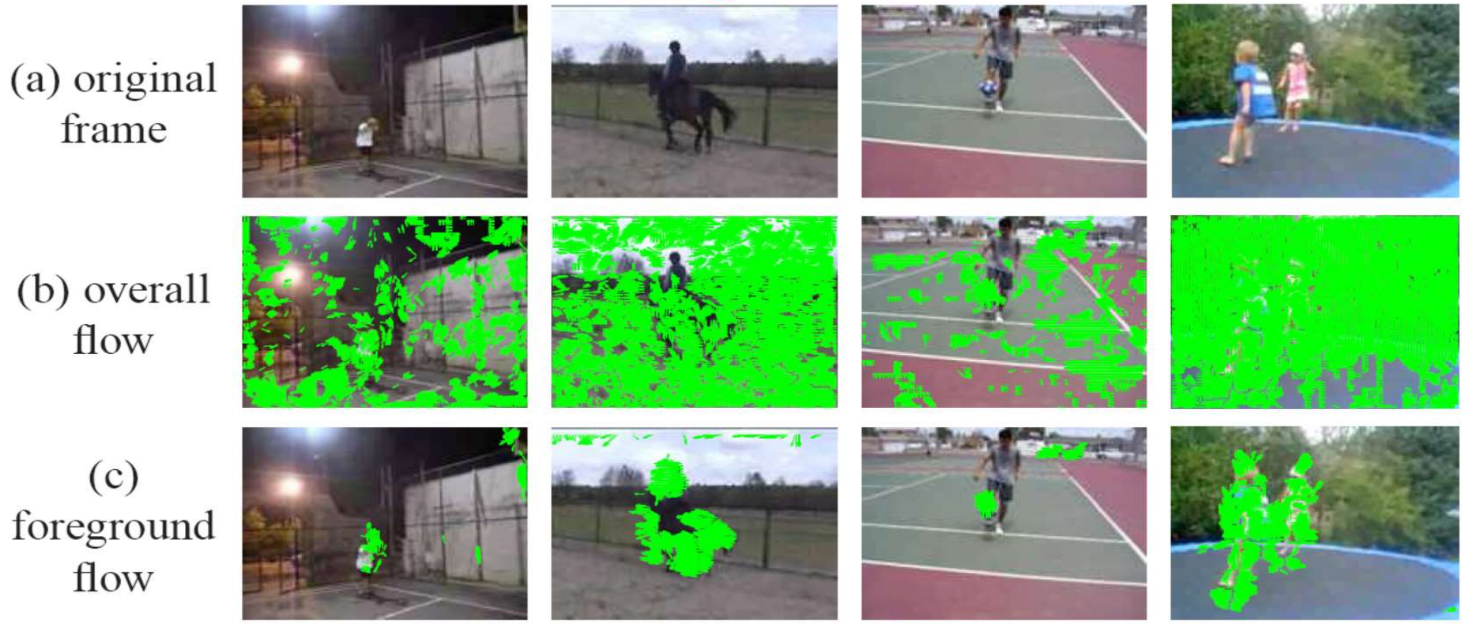
Extract person tracks

Each video consist of multiple (noisy) feature bags

Extract features from object and person tracks and the scene



Stabilizing the Videos



Dominant motion compensation (Liu and Gleicher, 2009)

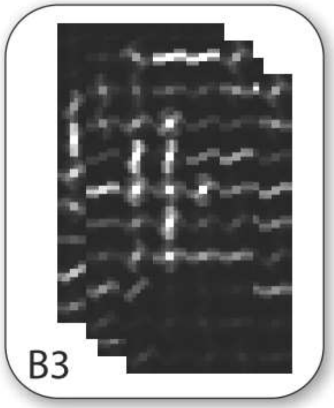
- Assuming the background is relatively dominant,
  - extract Harris corner features from each frame
  - estimate homography between consecutive frames
  - use homography to compute background flow $\mathbf{m}_b$ and as a prior to the block-based optical flow algorithm to compute overall flow $\mathbf{m}_o$

$$\mathbf{m}_f(x, y) = (\mathbf{m}_o(x, y) - \mathbf{m}_b(x, y))$$

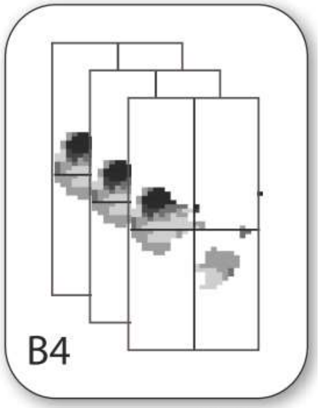
Liu, F., Gleicher, M.: Learning color and locality cues for moving object detection and segmentation. In: CVPR (2009)

Person-centric features

- Extract person tracks
  - run Felzenswalb’s person detector
  - apply mean-shift tracker in between where there is no detection
  - eliminate short tracks
- Extract features from tracks
  - Person-motion: HOF from snippets over temporal windows
  - Person-shape: HOG from snippets over temporal windows



person HOG



person OF

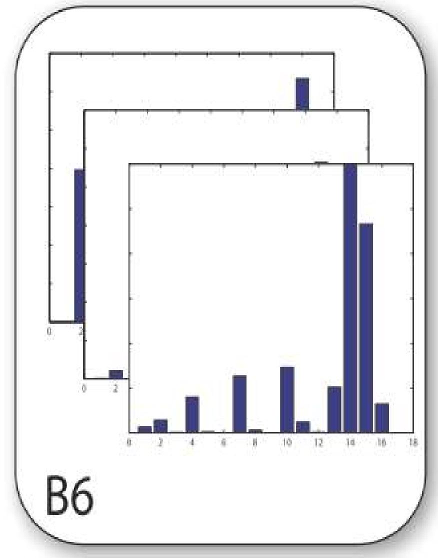
Object-centric Features

- ***Object candidate:*** moving region that has sufficient temporal and spatial coherence
- Extract object candidate tracks
  - connected components of the flow field
  - agglomerative clustering of the object regions
    - spatial coherence
    - appearance similarity
  - generate tracks using mean-shift tracking
  - eliminate short tracks

Scene Features

- Scene-shape: GIST features from random frames

- Scene-color: 3x1 color histograms from random frames



Multiple Instance Learning (MIL)

- There may be many object and/or person tracks extracted from each video.
- Some of these tracks may be relevant to the action
  - the track of a basketball
  - a jumping person
- Some of the tracks may be irrelevant or caused by noise
  - wrong person detections
  - Tracks caused by excessive camera motion
- Particular suitability of MIL => The given class label is associated with bags, rather than instances

Experimental Evaluation

- Experimented over the UCF YouTube dataset
  - 1168 videos and 11 action classes like basketball shooting, diving, horse riding, playing tennis, etc.
  - Leave-one-out cross validation

Results

	b_shoot	bike	dive	golf	h_ride	s_juggle	swing	t_swing	t_jump	v_spike	walk	**Avg**
% correct classification using single feature channels												
perOF	20.20	44.83	51.0	69.0	45.0	44.0	36. 0	32. 0	64. 0	29. 0	29.27	42.72
perHOG	28.28	57.93	56.0	40.0	51.0	36.0	43. 0	45. 0	34. 0	49. 0	39.84	43.64
objOF	14.14	45.52	24.0	36.0	51.0	20.0	42. 0	14. 0	59. 0	25. 0	33.33	33.09
objHOG	21.21	44.14	62. 0	55. 0	38. 0	22. 0	42. 0	44. 0	42. 0	45. 0	21.95	39.75
gist	38.38	60.69	69. 0	61. 0	66. 0	9. 0	42. 0	61. 0	54. 0	81. 0	43.09	53.20
color	33.33	44.83	86. 0	65. 0	43. 0	22. 0	27. 0	47. 0	57. 0	73. 0	43.90	49.28
% correct classification using combinations of channels												
p+s	44.44	70.34	92. 0	87. 0	63. 0	35. 0	56. 0	75. 0	84. 0	84. 0	56.91	67.97
p+o	40.40	70.34	84. 0	91. 0	63. 0	**54. 0**	63. 0	60. 0	84. 0	78. 0	50.41	67.11
o+s	47.47	73.79	91. 0	90. 0	73. 0	35. 0	64. 0	75. 0	83. 0	**89. 0**	56.10	70.67
% correct classification using all feature channels												
p+o+s	48.48	**75.17**	95. 0	**95. 0**	**73. 0**	53. 0	**66. 0**	77. 0	**93. 0**	85. 0	66.67	**75.21**
w[p+o+s]	43.43	**75.17**	**96. 0**	94. 0	72. 0	47. 0	65. 0	74. 0	**93. 0**	85. 0	67.48	73.83
Liu [22]	**53.0**	73.0	81.0	86.0	72.0	**54.0**	57.0	**80.0**	79.0	73.3	**75.0**	71.2

Best classification accuracy per action

Best classification accuracy using single feature channels

Best classification accuracy using multiple feature channels

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

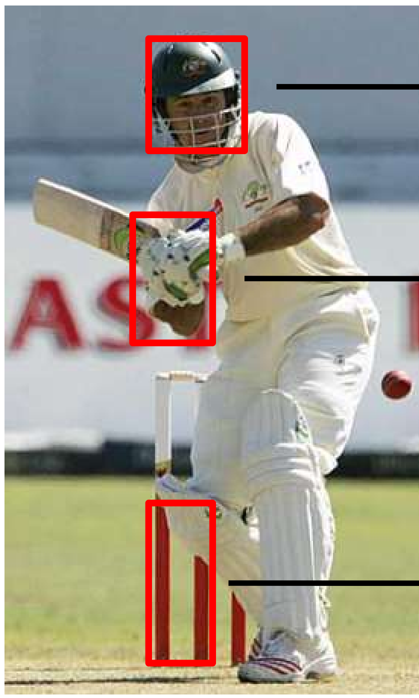
- **Human pose estimation**
- **Object detection**
- **Action categorization**

HOI activity: Tennis Forehand

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

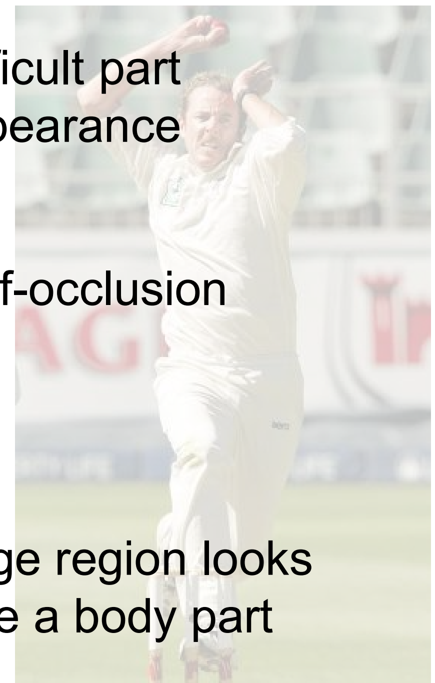
Human pose estimation is challenging.



Difficult part appearance

Self-occlusion

Image region looks like a body part

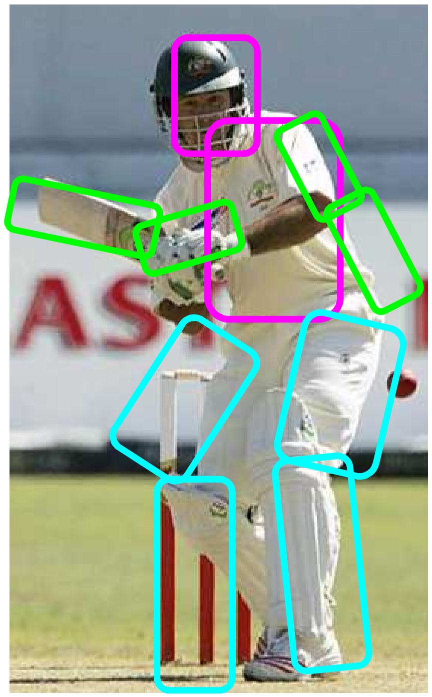


- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Human pose estimation is challenging.



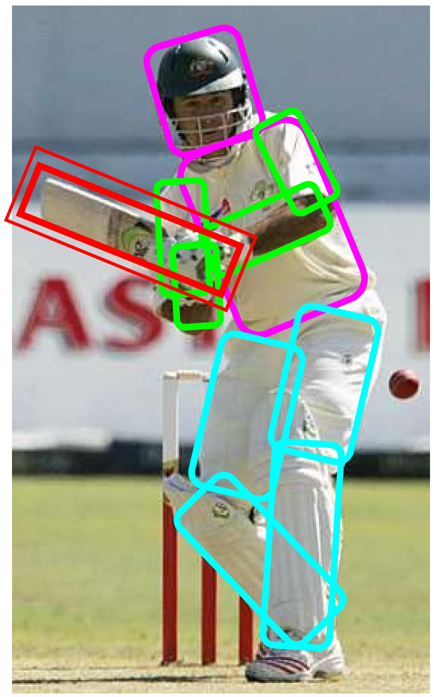
- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

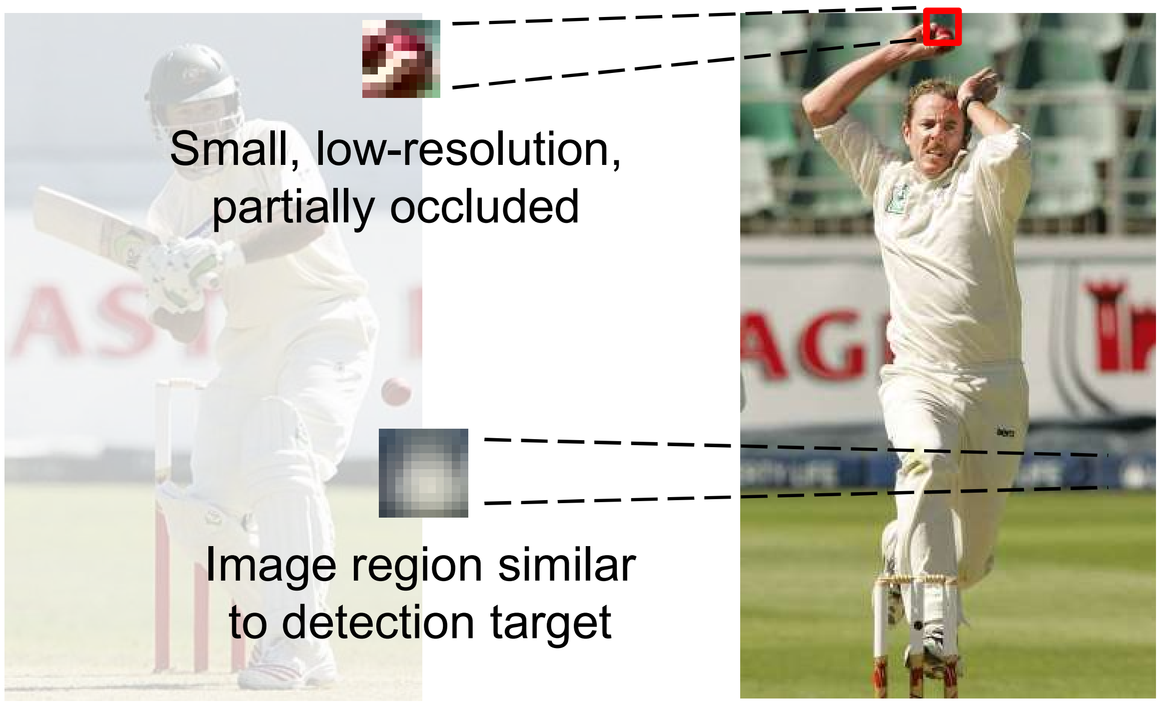
Facilitate

Given the object is detected.



Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

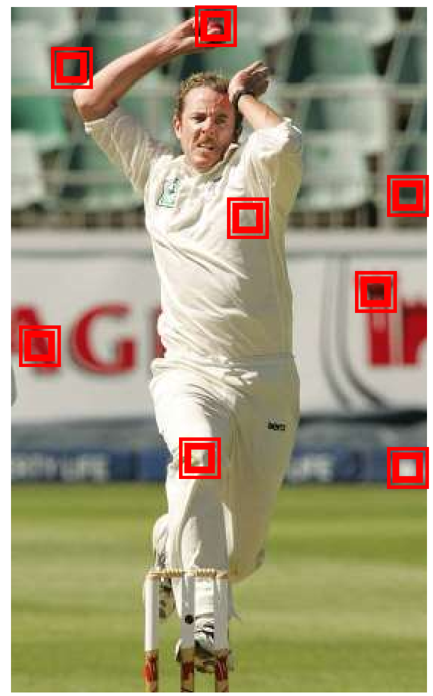


Object detection is challenging

- 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



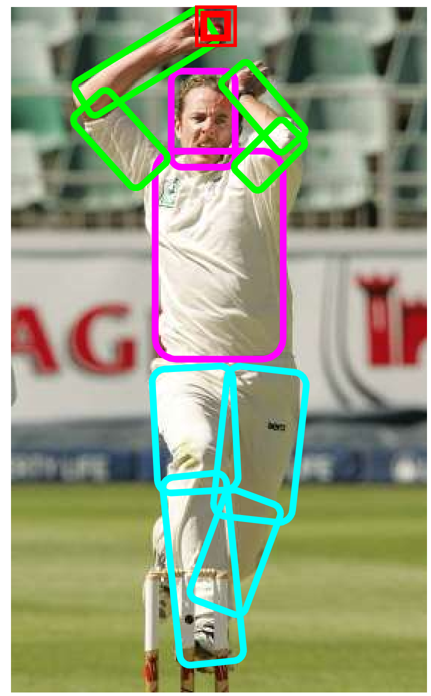
Object detection is challenging

- 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

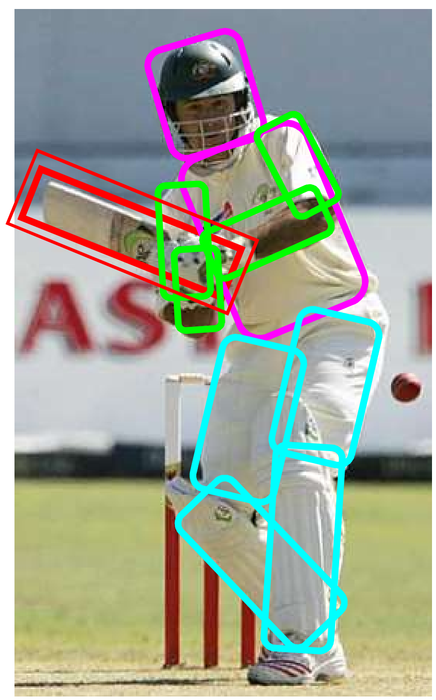


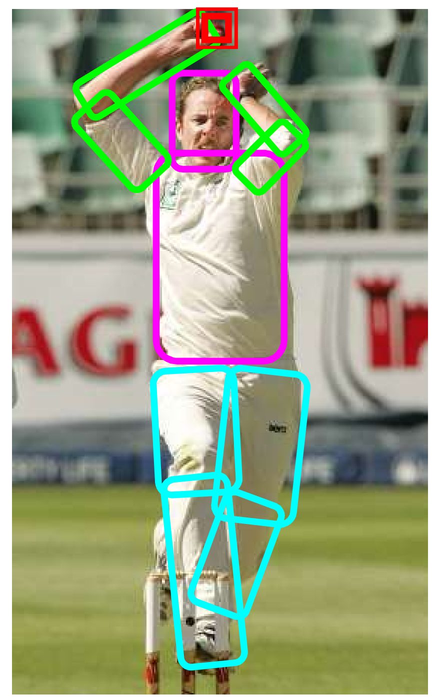
Given the pose is estimated.

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Mutual Context





Slide Credit: Yao/Fei-Fei

Mutual Context Model Representation

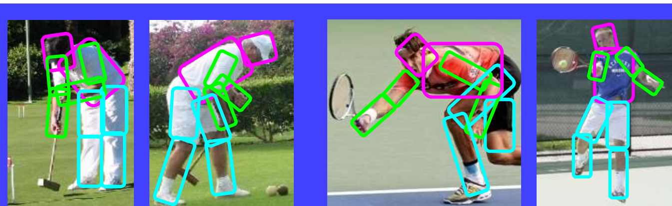
A:

Tennis forehand | Croquet shot | Volleyball smash | …

O:

Tennis racket | Croquet mallet | Volleyball | …

H:

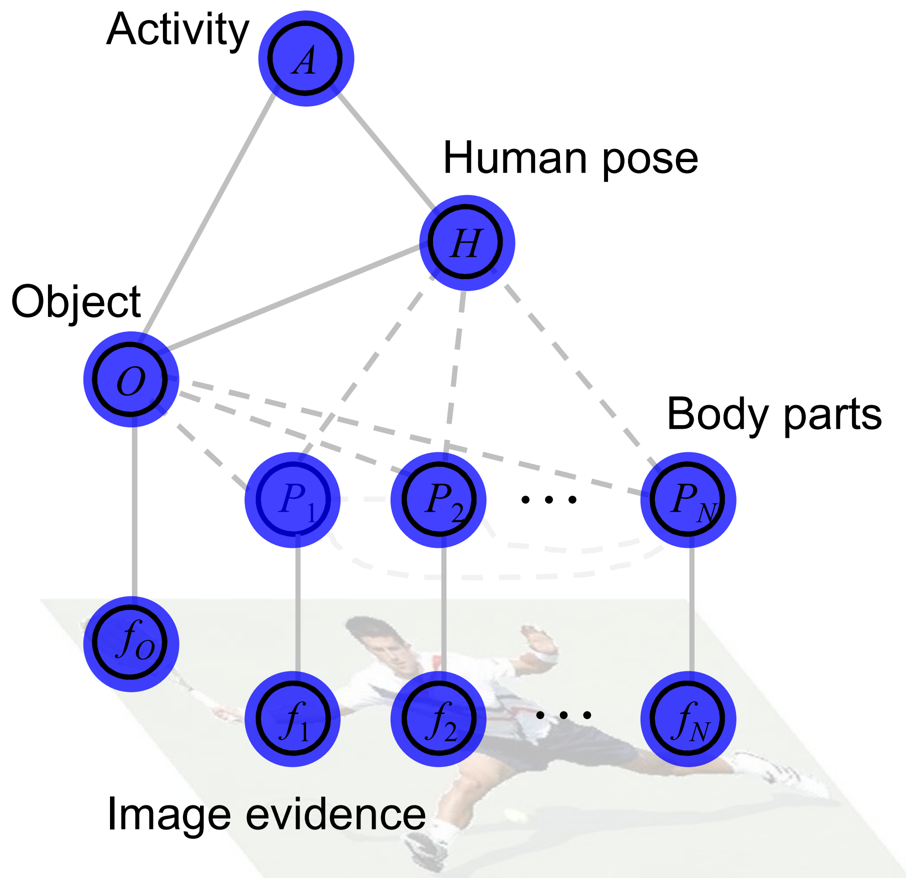


Intra-class variations

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

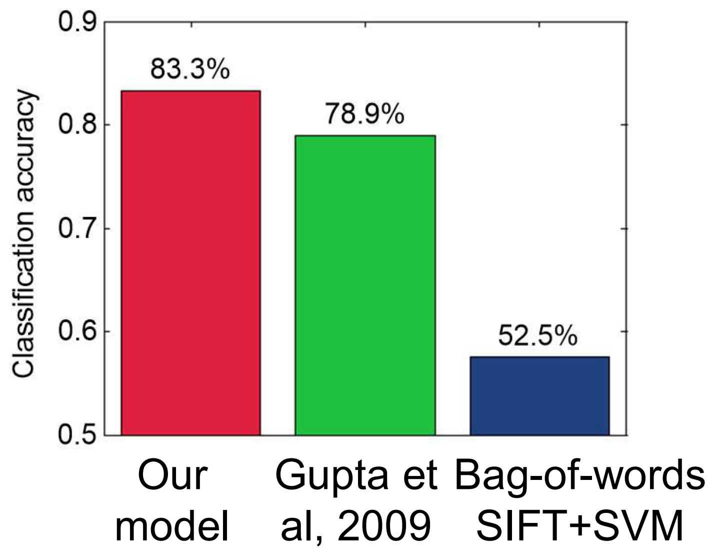
P: l_P: location; θ_P: orientation; s_P: scale.

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



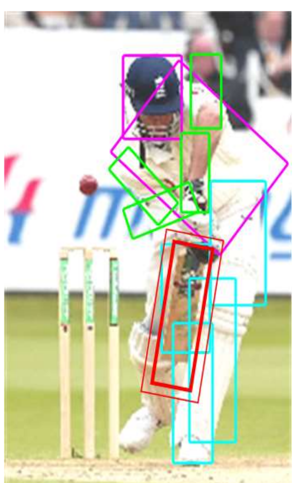
Slide Credit: Yao/Fei-Fei

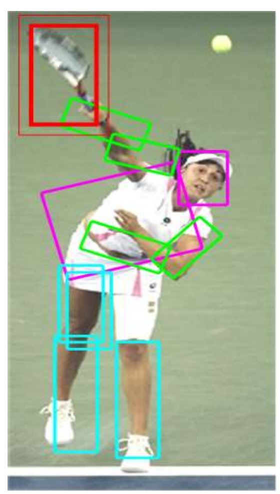
Activity Classification Results

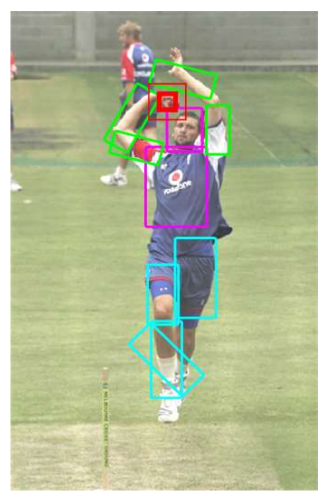


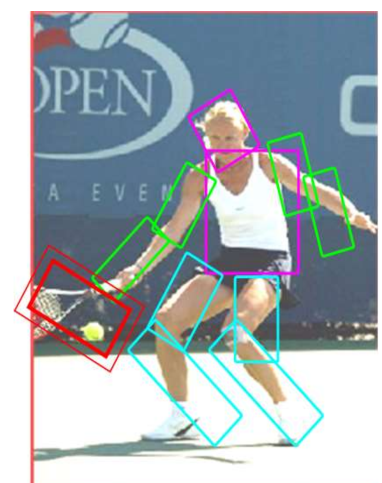
Cricket shot

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

Many more subjects and research directions

-Structure from Motion
-Tracking
-Video object Segmentation
-Context
-Attributes
- And more..