CS 558: Computer Vision
14th Set of Notes

Instructor: Philippos Mordohai
Webpage: www.cs.stevens.edu/~mordohai
E-mail: Philippos.Mordohai@stevens.edu
Office: Lieb 215

Overview

- Action and activity recognition

- Background subtraction
- Tracking
- Tracking people

  – Based on slides by K. Grauman and D. Hoiem

Action and 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)

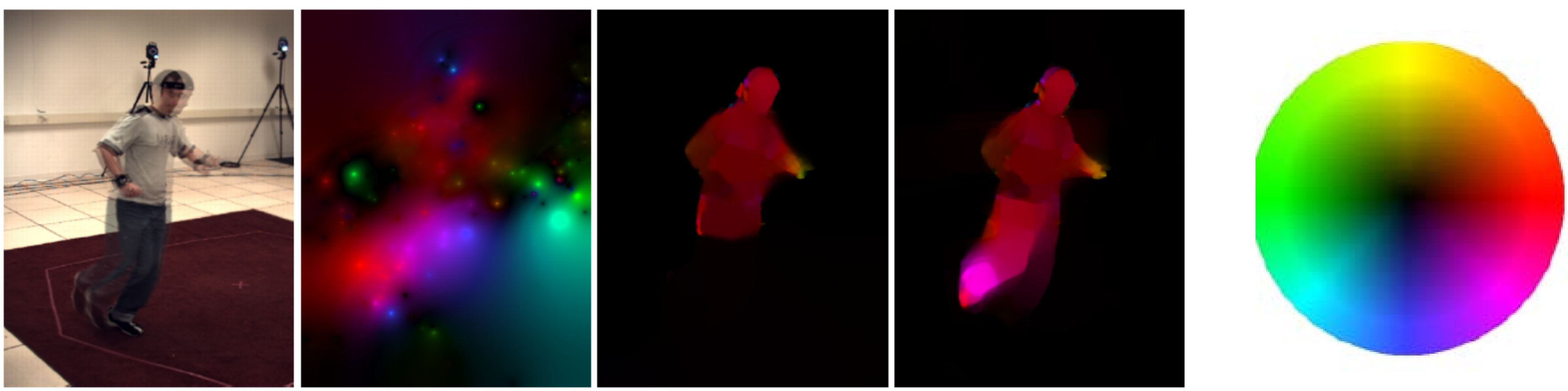
- "**Activity**": series or composition of actions (e.g., interactions between people)

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

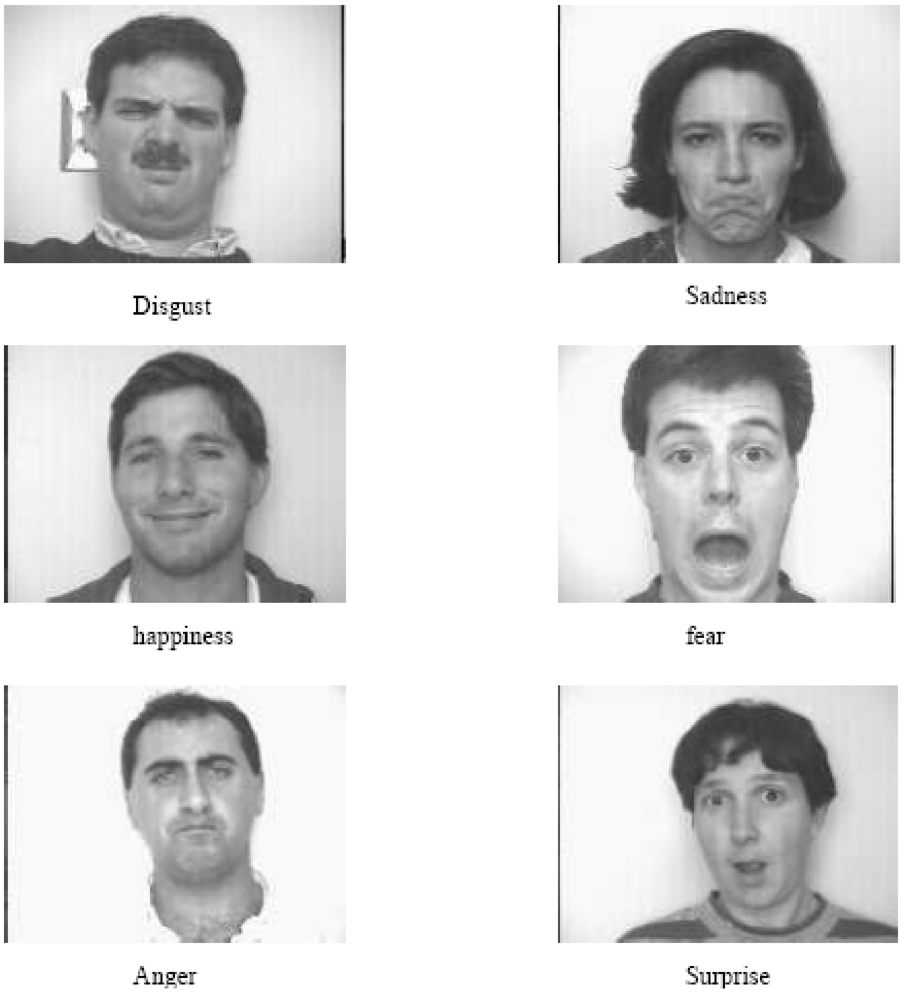
Adapted from Venu Govindaraju

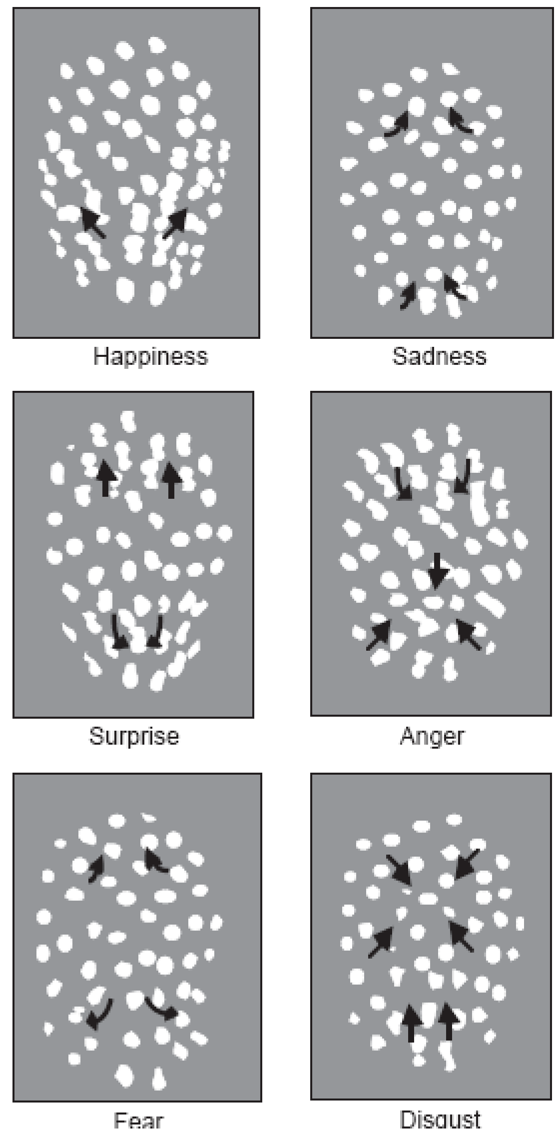
Reminder: Optical Flow

- Definition: optical flow is the *apparent* motion of brightness patterns in the image



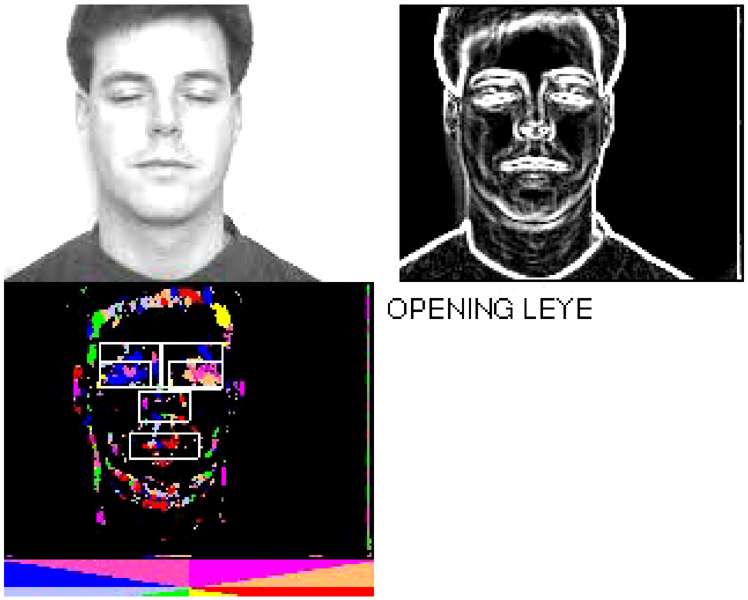
Using optical flow: recognizing facial expressions



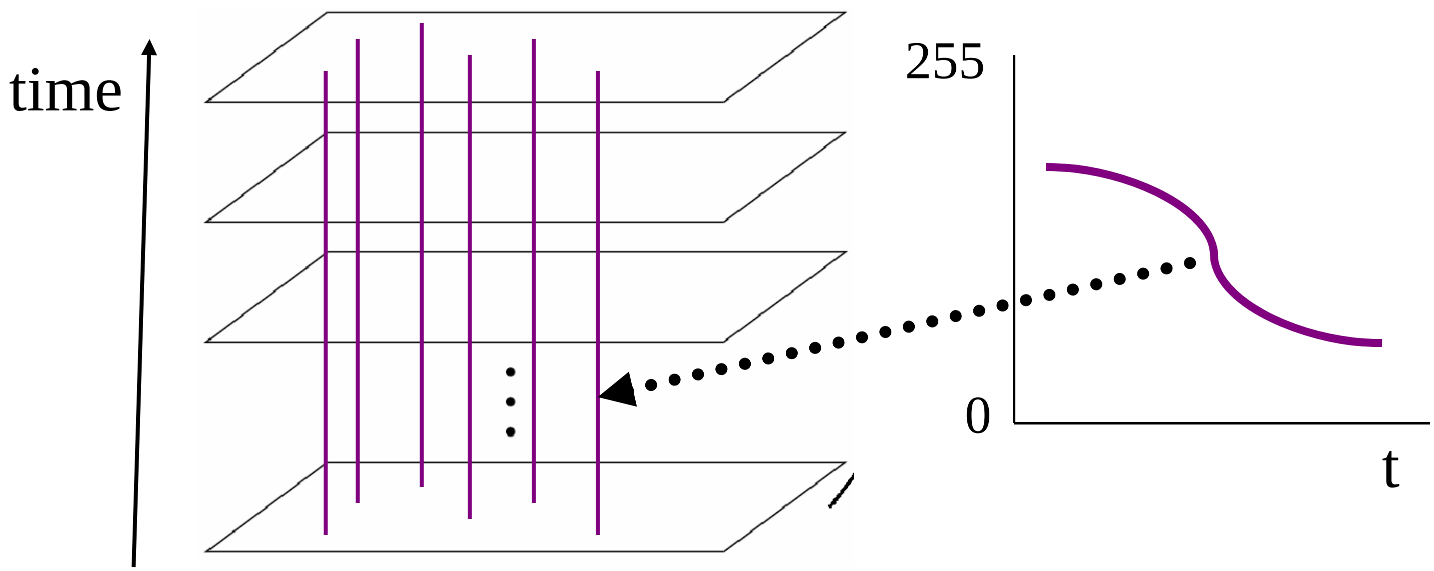


Recognizing Human Facial Expression (1994)
by Yaser Yacoob, Larry S. Davis

Using optical flow: recognizing facial expressions



Video as an “Image Stack”



- Can look at video data as a spatio-temporal volume
  - If camera is stationary, each line through time corresponds to a single ray in space

Alyosha Efros, CMU

Background Subtraction

- Given an image (a video frame), we want to identify the foreground objects

$\Rightarrow$

- In most cases, objects are of interest, not the static part of the scene
- Makes our life easier: lower processing costs and less room for error

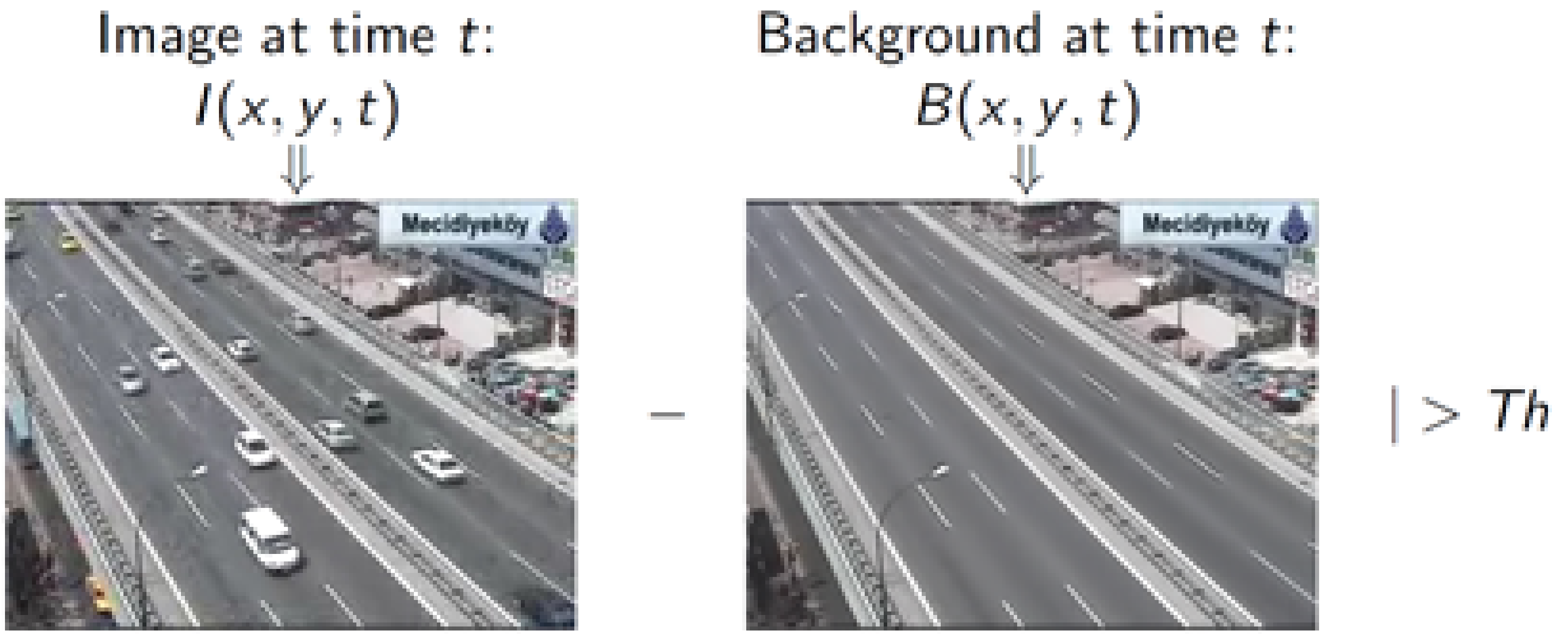
Slide credit: Birgi Tamersoy

Background subtraction

- Simple techniques can do ok with static camera
- …But hard to do perfectly

- Widely used:
  - Traffic monitoring (counting vehicles, detecting & tracking vehicles, pedestrians),
  - Human action recognition (run, walk, jump, squat),
  - Human-computer interaction
  - Object tracking

Simple Approach



1. Estimate the background for time t
2. Subtract the estimated background from the input image
3. Apply a threshold *Th* to the absolute difference to get the foreground mask

Slide credit: Birgi Tamersoy

Frame Differencing

- Background is estimated to be the previous frame. Background subtraction equation becomes:

$$B(x,y,t) = I(x,y,t-1)$$

$$|I(x,y,t)=I(x,y,t-1)| > Th$$

- Depending on the object structure, speed, frame rate and global threshold, this approach may or may **not** be useful (usually not)

Slide credit: Birgi Tamersoy

Frame Differencing

$Th = 25$

$Th = 50$

$Th = 100$

$Th = 200$

Slide credit: Birgi Tamersoy

Mean Filter

- The background is the mean of the previous n frames

$$B(x, y, t) = \frac{1}{n}\sum_{i=0}^{n-1} I(x, y, t-i)$$

$$\Downarrow$$

$$|I(x, y, t) - \frac{1}{n}\sum_{i=0}^{n-1} I(x, y, t-i)| > Th$$

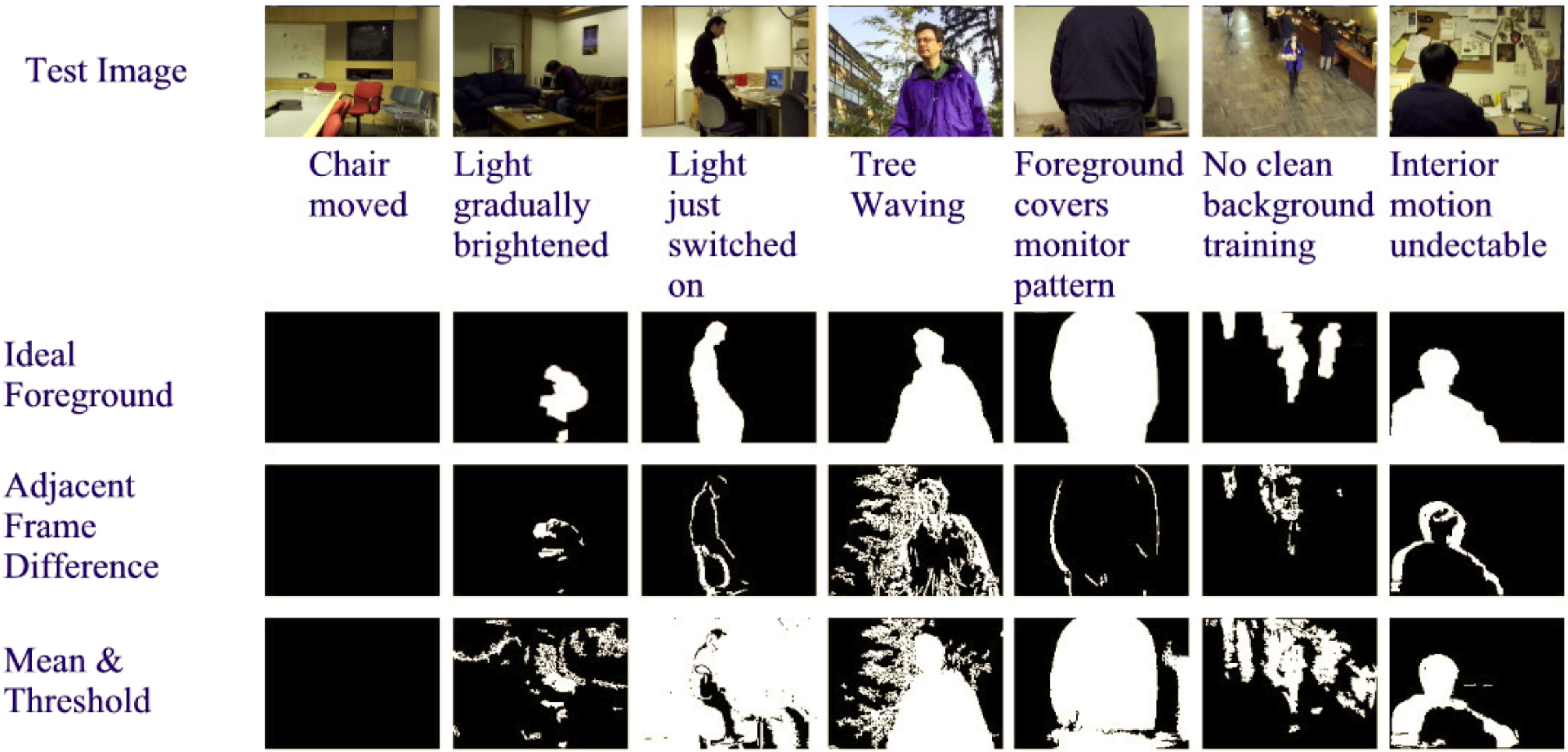
- For n=10

Estimated Background

Foreground Mask

Slide credit: Birgi Tamersoy

Frame differences vs. background subtraction



- Toyama et al. 1999

Median Filter

- Assuming that the background is more likely to appear in a scene, we can use the median of the previous n frames as the background model:

$$B(x,y,t) = median\{I(x,y,t-i)\}$$
$$\Downarrow$$
$$|I(x,y,t) - median\{I(x,y,t-i)\}| > Th \text{ where}$$
$$i \in \{0,\ldots,n-1\}.$$

- For n=10

Estimated Background

Foreground Mask

Slide credit: Birgi Tamersoy

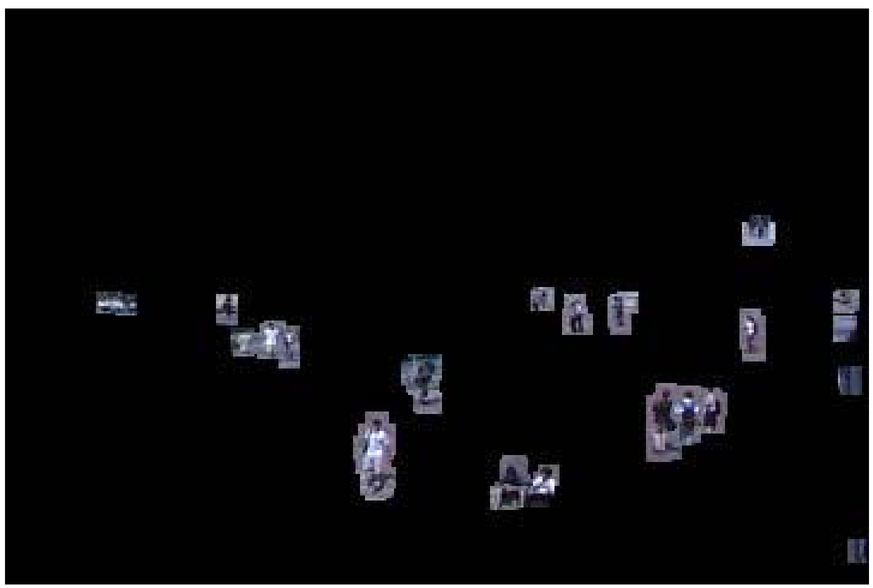
Average/Median Image

Alyosha Efros

Background Subtraction

\-

=



Alyosha Efros

Pros and cons

Advantages:

- Extremely easy to implement and use!
- All pretty fast.
- Corresponding background models need not be constant, they change over time.

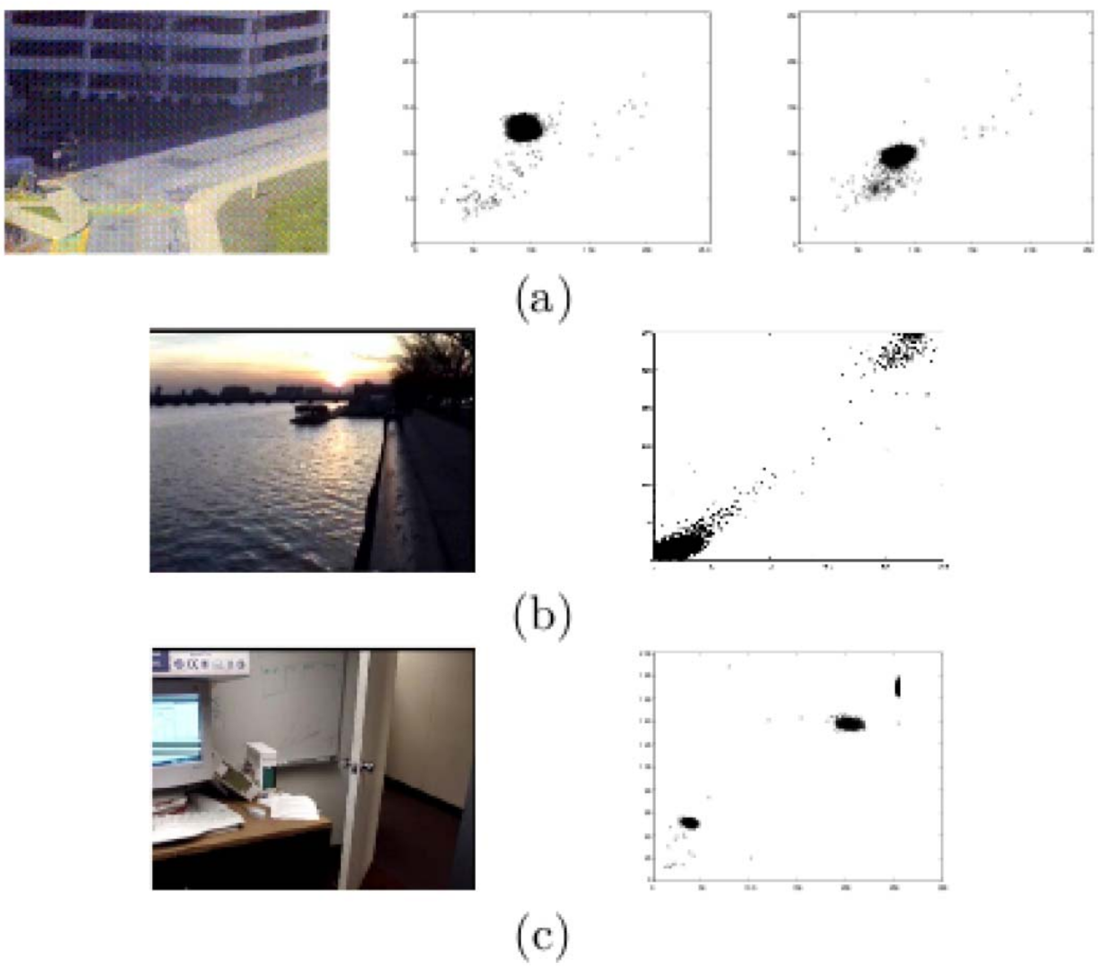
Disadvantages:

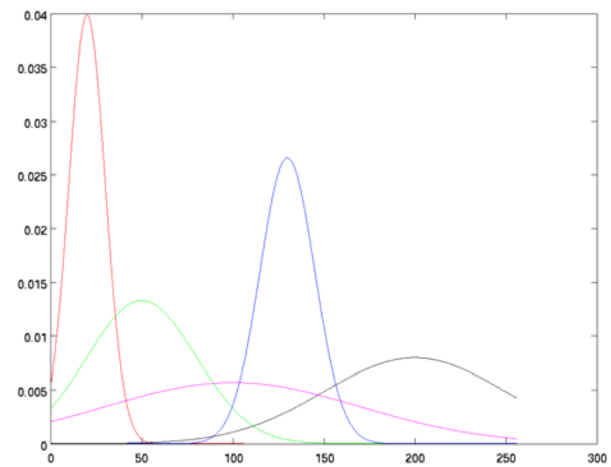
- Accuracy of frame differencing depends on object speed and frame rate
- Median background model: relatively high memory requirements.
- Setting global threshold Th…

When will this basic approach fail?

Slide credit: Birgi Tamersoy

Background mixture models





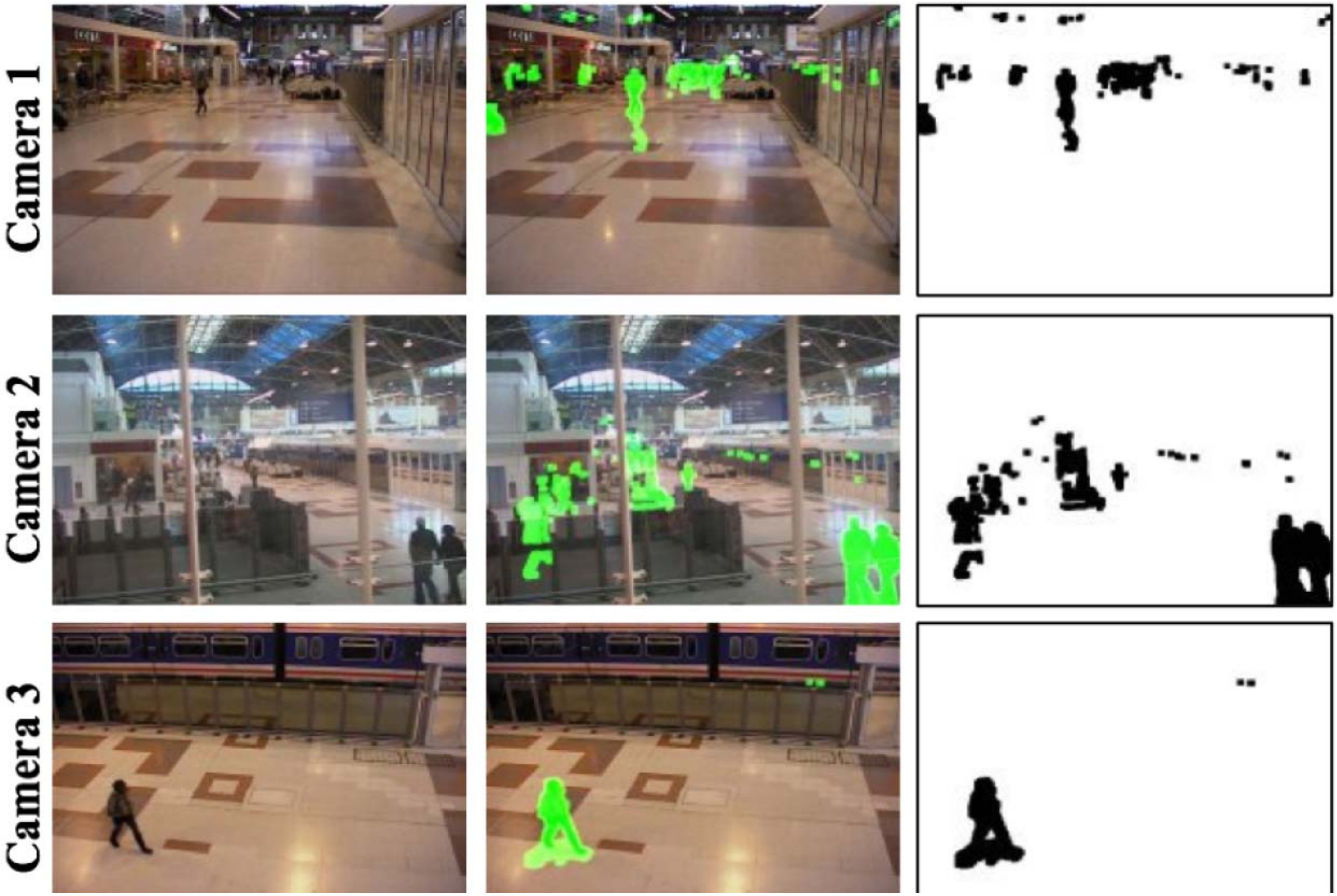
Idea: model each background pixel with a *mixture* of Gaussians; update its parameters over time.

Adaptive Background Mixture Models for Real-Time Tracking, Chris Stauer & W.E.L. Grimson

Background subtraction with **depth**

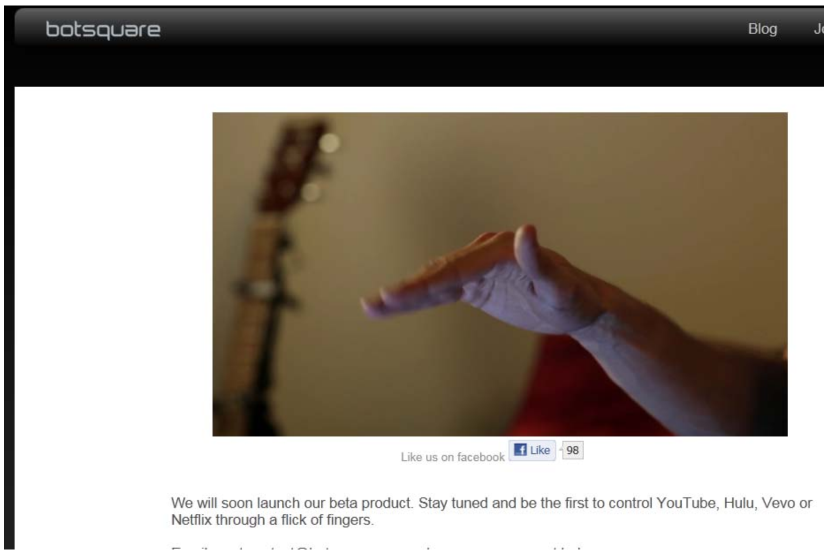
How can we select foreground pixels based on depth information?

Surveillance



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

Interfaces



2011

(a) template

(b) image

(c) normalized correlation

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

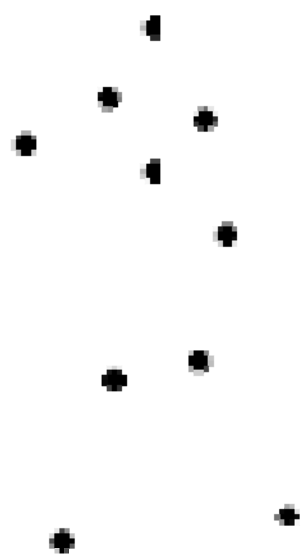
Human activity in video: basic approaches

- **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

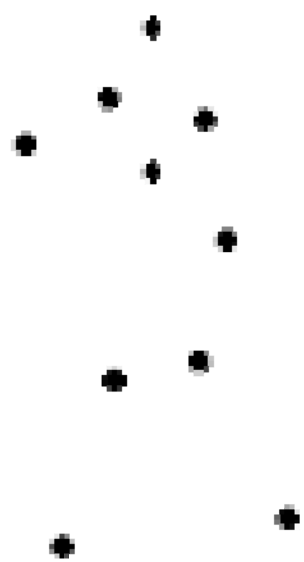
Motion and perceptual organization

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



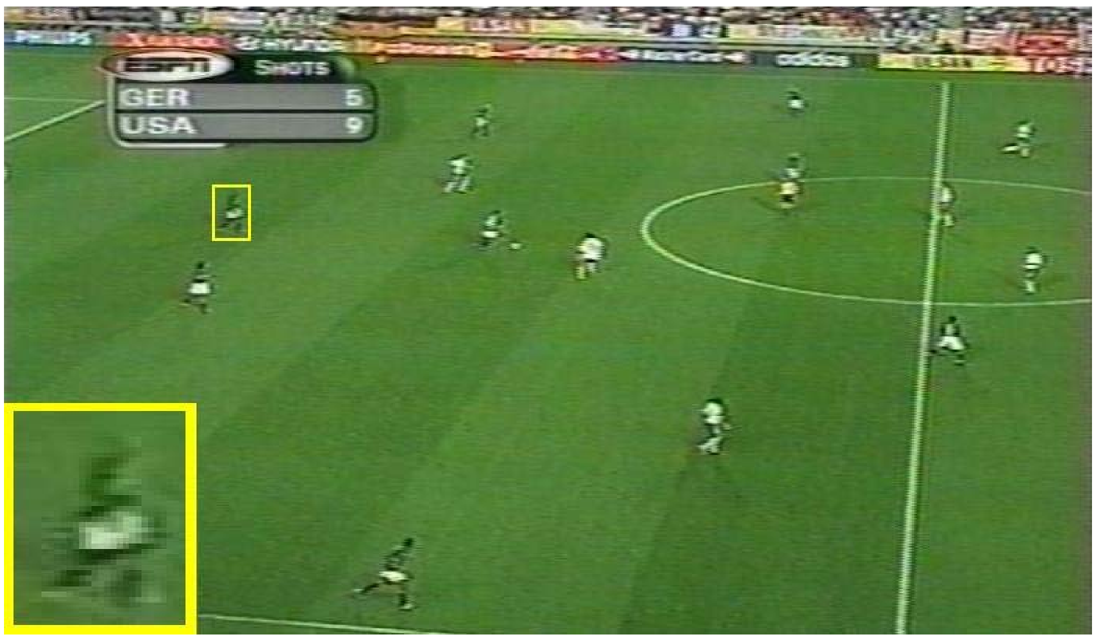
Motion and perceptual organization

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



Using optical flow: action recognition at a distance

- Features = optical flow within a region of interest
- Classifier = nearest neighbors

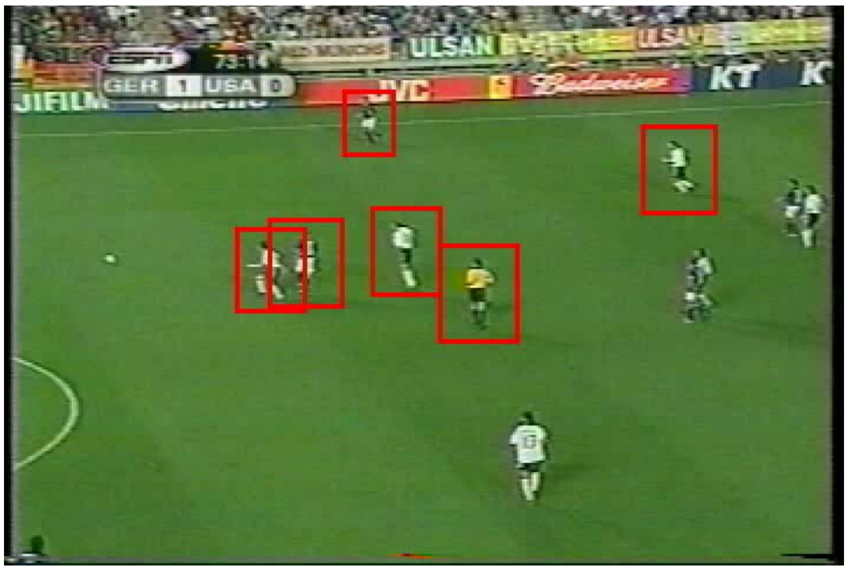


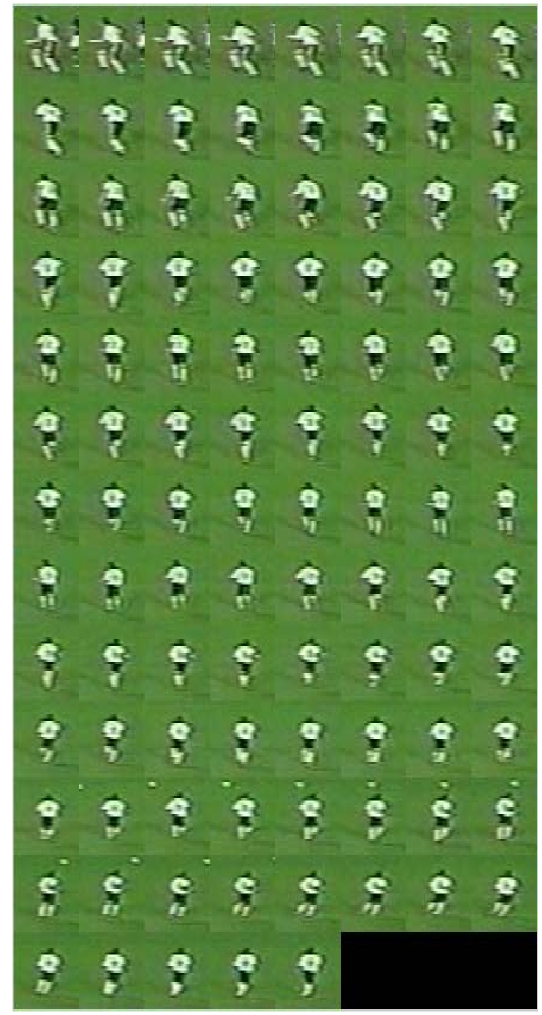
The 30-Pixel Man

Challenge: low-res data, not going to be able to track each limb.

[Efros, Berg, Mori, & Malik 2003]
http://graphics.cs.cmu.edu/people/efros/research/action/

Using optical flow: action recognition at a distance

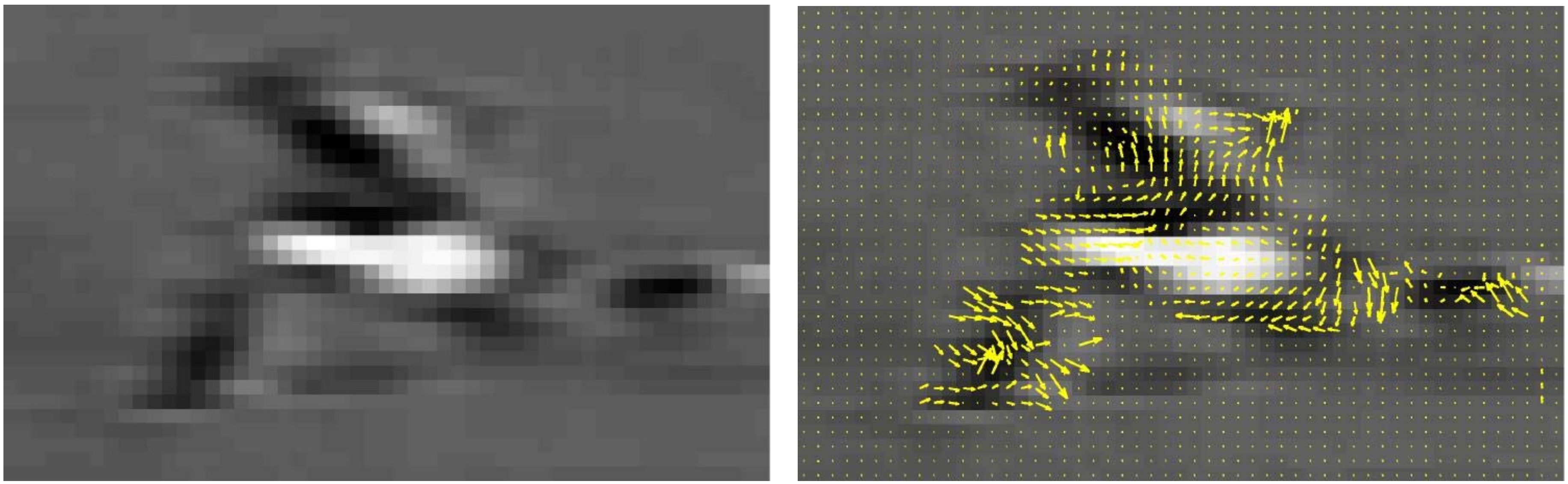




Correlation-based tracking
Extract person-centered frame window

[Efros, Berg, Mori, & Malik 2003]
http://graphics.cs.cmu.edu/people/efros/research/action/

Using optical flow: action recognition at a distance

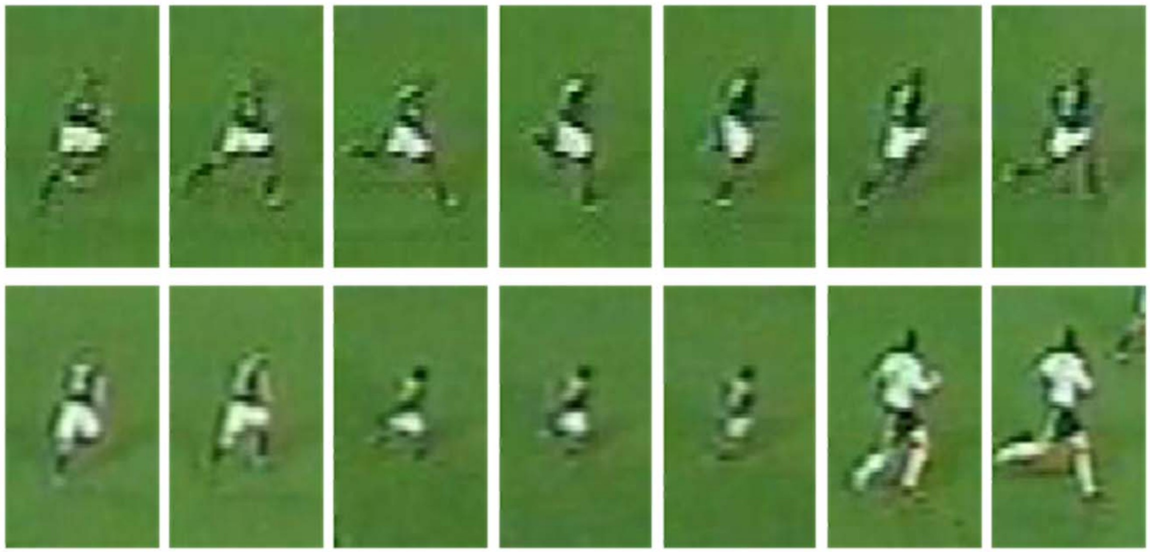


Extract optical flow to describe the region's motion.

[Efros, Berg, Mori, & Malik 2003]
http://graphics.cs.cmu.edu/people/efros/research/action/

Using optical flow: action recognition at a distance

Input Sequence

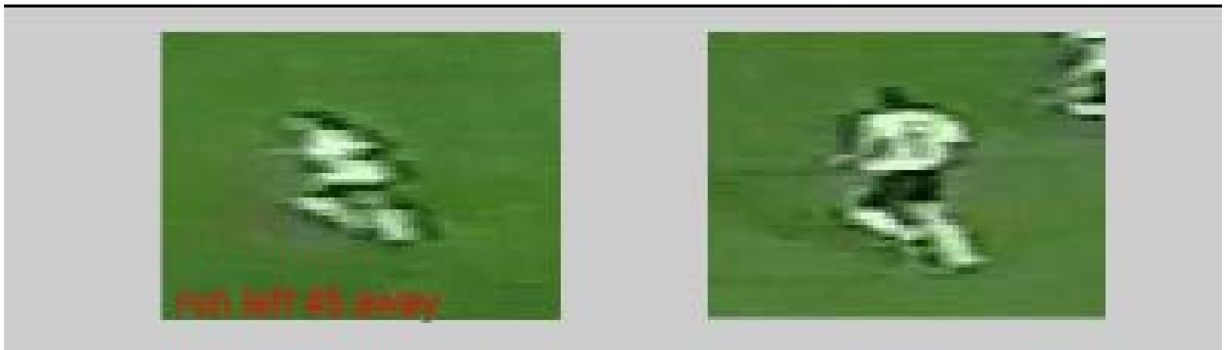


Matched Frames

Use **nearest neighbor** classifier to name the actions occurring in new video frames.

[Efros, Berg, Mori, & Malik 2003]
http://graphics.cs.cmu.edu/people/efros/research/action/

Using optical flow: action recognition at a distance



Input Sequence

Matched NN Frame

Use **nearest neighbor** classifier to name the actions occurring in new video frames.

[Efros, Berg, Mori, & Malik 2003]
http://graphics.cs.cmu.edu/people/efros/research/action/

Do as I do: motion retargeting

[Efros, Berg, Mori, & Malik 2003]
http://graphics.cs.cmu.edu/people/efros/research/action/

Motion Energy Images

$$E_\tau(x,y,t) = \bigcup_{i=0}^{\tau-1} D(x,y,t-i)$$

D(x,y,t): Binary image sequence indicating motion locations

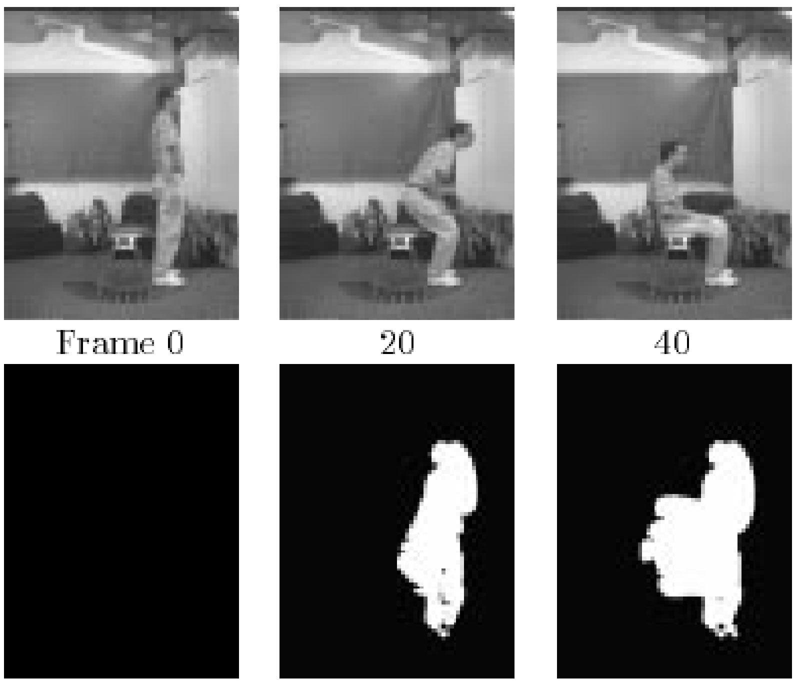


Figure 2: Example of someone sitting. Top row contains key frames; bottom row is cumulative motion images starting from Frame 0.

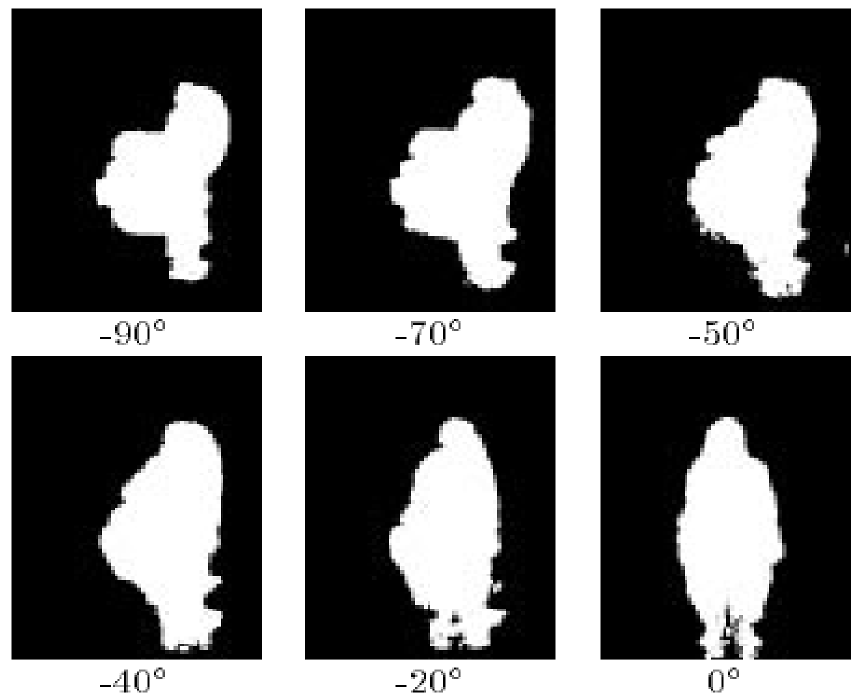
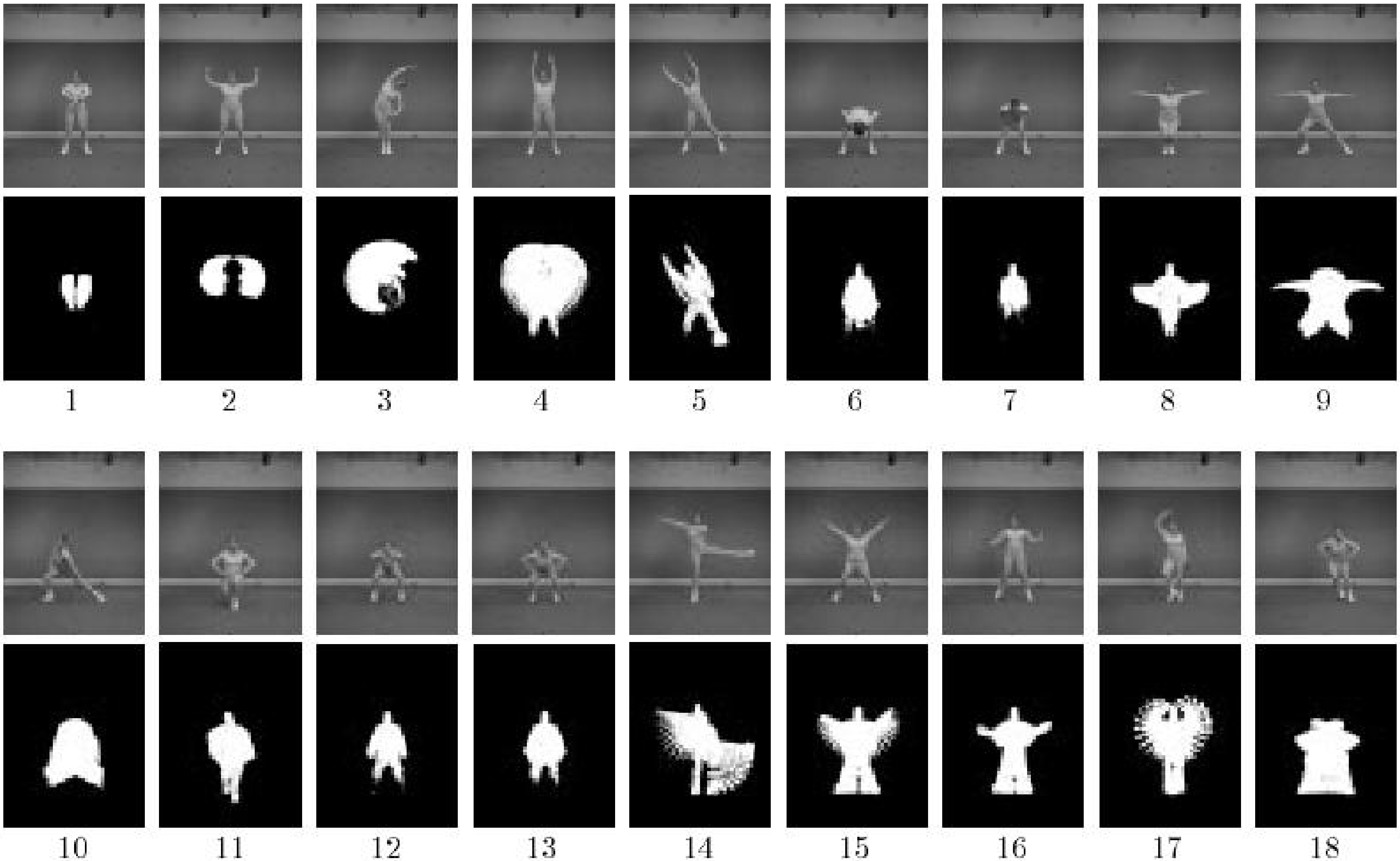


Figure 3: MEIs of sitting action over 90° viewing angle. The smooth change implies only a coarse sampling of viewing direction is necessary to recognize the action from all angles.

Davis & Bobick 1999: The Representation and Recognition of Action Using Temporal Templates

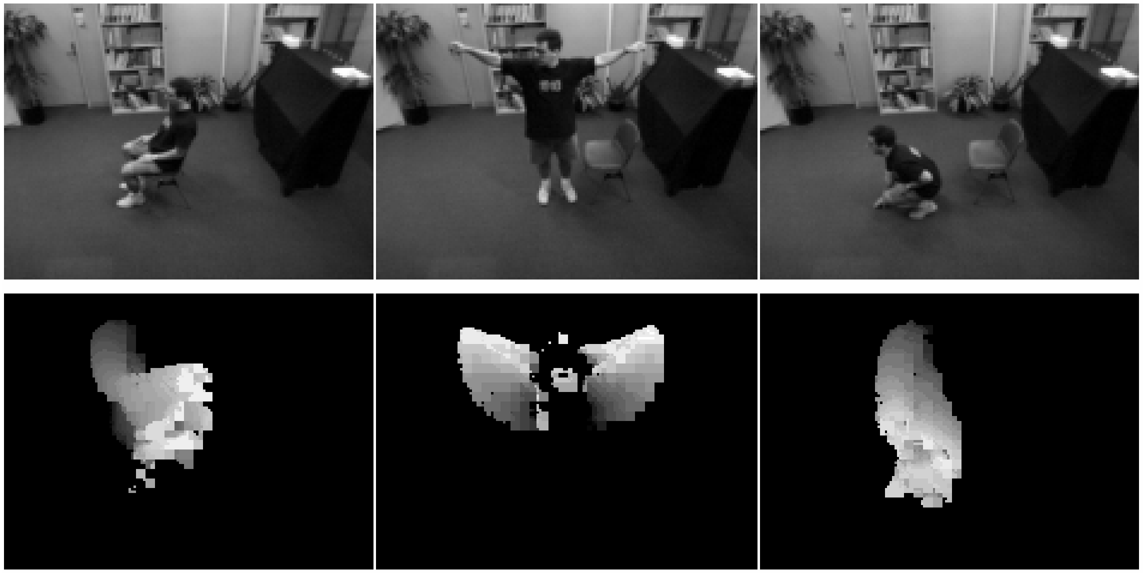
Motion Energy Images



Davis & Bobick 1999: The Representation and Recognition of Action Using Temporal Templates

Motion History Images

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



Davis & Bobick 1999: The Representation and Recognition of Action Using Temporal Templates

Image moments

Use to summarize shape given image *I(x,y)*

$$M_{ij} = \sum_x \sum_y x^i y^j I(x, y)$$

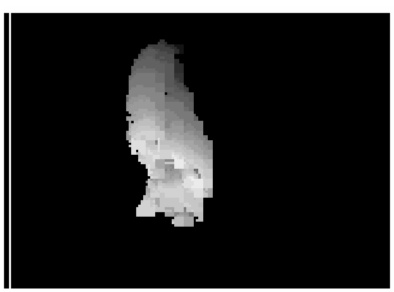
Central moments are translation invariant:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q I(x, y)$$

$$\bar{x} = \frac{M_{10}}{M_{00}} \qquad \bar{y} = \frac{M_{01}}{M_{00}}$$

Hu moments

- Set of 7 moments
- Apply to Motion History Image for global space-time “shape” descriptor
- Translation and rotation invariant



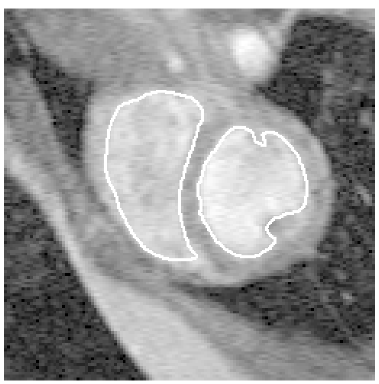
$$\Rightarrow [h_1, h_2, h_3, h_4, h_5, h_6, h_7]$$

Tracking: some applications

Body pose tracking, activity recognition

Censusing a bat population

Video-based interfaces



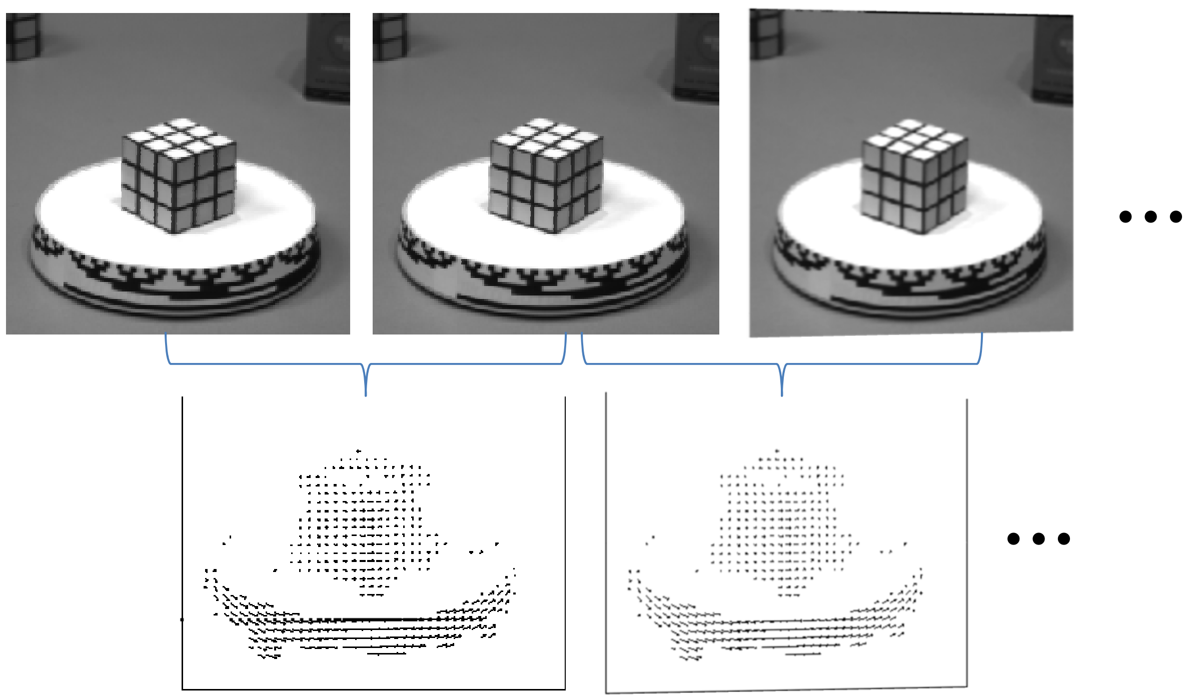
Medical

Surveillance

Kristen Grauman

Optical flow for tracking?

If we have more than just a pair of frames, we could compute flow from one to the next:



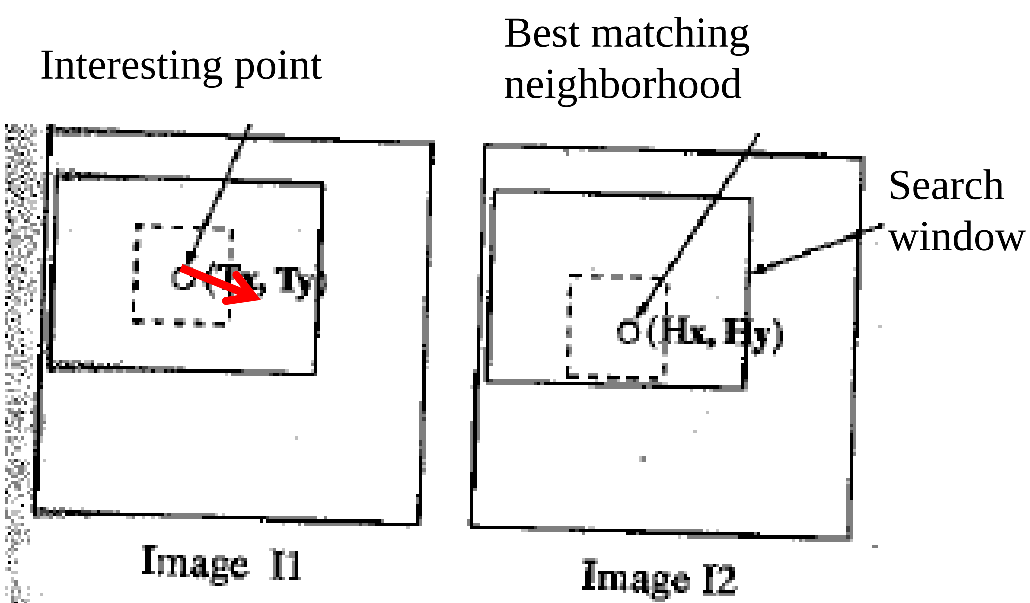
But flow only reliable for small motions, and we may have occlusions, textureless regions that yield bad estimates anyway…

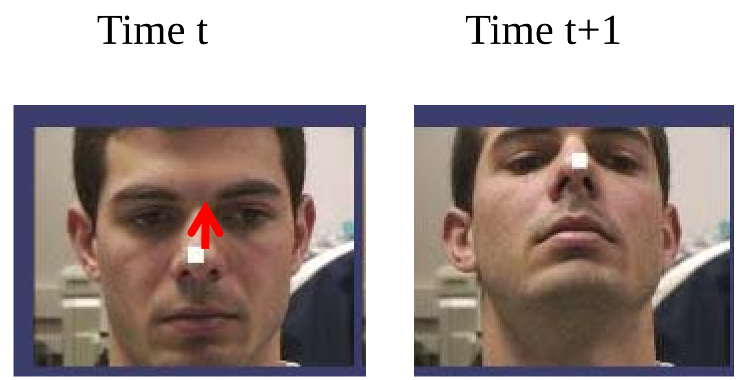
Motion estimation techniques

- Direct methods
  - Directly recover image motion at each pixel from spatio-temporal image brightness variations
  - Dense motion fields, but sensitive to appearance variations
  - Suitable for video and when image motion is small

- **Feature-based methods**
  - Extract visual features (corners, textured areas) and track them over multiple frames
  - Sparse motion fields, but more robust tracking
  - Suitable when image motion is large (10s of pixels)

Feature-based matching for motion





Search window is centered at the point where we last saw the feature, in image I1.

Best match = position where we have the highest normalized cross-correlation value.

Kristen Grauman

Example: A Camera Mouse

- Video interface: use feature tracking as mouse replacement

- User clicks on the feature to be tracked
- Take the 15x15 pixel square of the feature
- In the next image do a search to find the 15x15 region with the highest correlation
- Move the mouse pointer accordingly
- Repeat in the background every 1/30th of a second

James Gips and Margrit Betke
http://www.bc.edu/schools/csom/eagleeyes/

Kristen Grauman

Example: A Camera Mouse

- Specialized software for communication, games

James Gips and Margrit Betke
http://www.bc.edu/schools/csom/eagleeyes/

Kristen Grauman

Detection vs. tracking

t=1

t=2

…

t=20

t=21

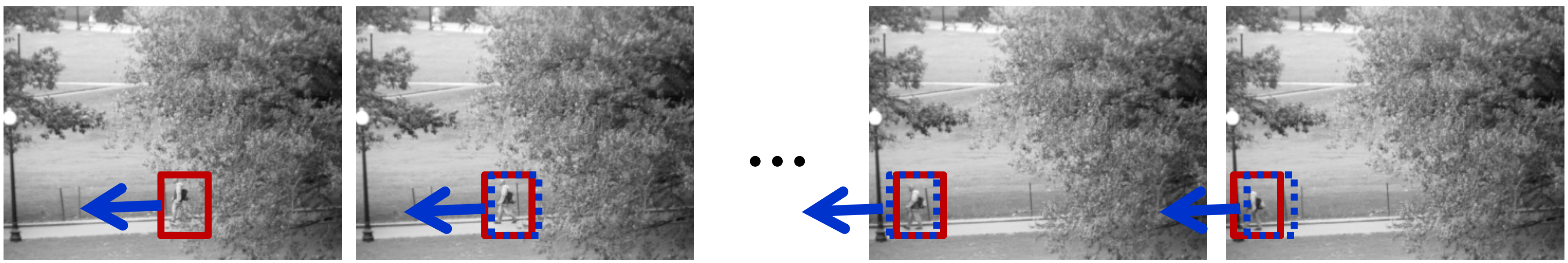
Kristen Grauman

Detection vs. tracking

Detection: We detect the object independently in each frame and can record its position over time, e.g., based on blob's centroid or detection window coordinates

Kristen Grauman

Detection vs. tracking



Tracking with *dynamics*: We use image measurements to estimate position of object, but also incorporate position predicted by dynamics, i.e., our expectation of object's motion pattern.

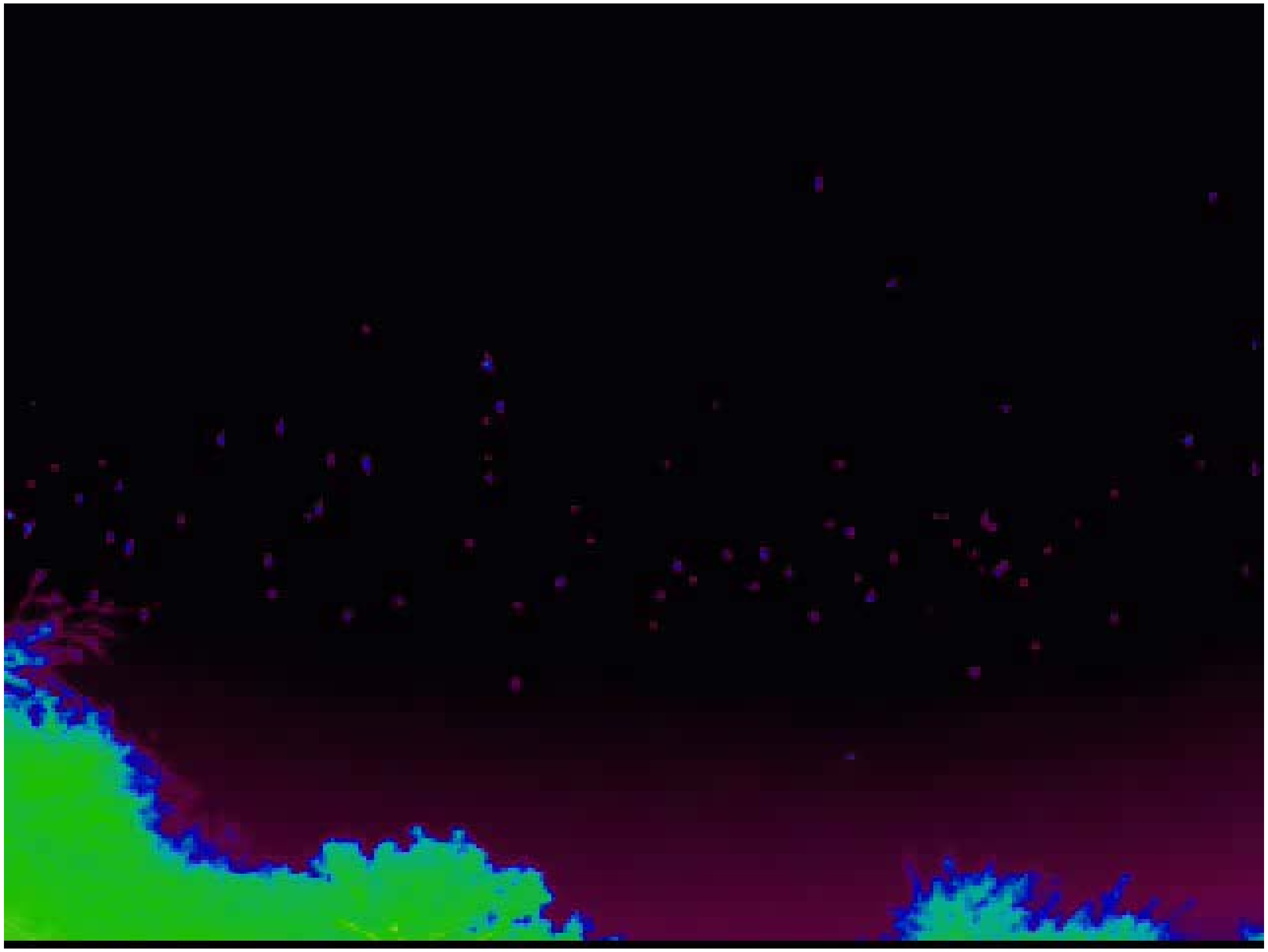
Kristen Grauman

Tracking with dynamics

- Use model of expected motion to *predict* where objects will occur in next frame, even before seeing the image
- **Intent**:
  - Do less work looking for the object, restrict the search
  - Get improved estimates since measurement noise is tempered by smoothness, dynamics priors
- **Assumption**: continuous motion patterns:
  - Camera is not moving instantly to new viewpoint
  - Objects do not disappear and reappear in different places in the scene
  - Gradual change in pose between camera and scene

Kristen Grauman

A bat census



http://www.cs.bu.edu/~betke/research/bats/

Tracking: issues

- **Initialization**
  - Often done manually
  - Background subtraction, detection can also be used
- **Data association**, multiple tracked objects
  - Occlusions, clutter

Tracking: issues

- **Initialization**
  - Often done manually
  - Background subtraction, detection can also be used
- **Data association**, multiple tracked objects
  - Occlusions, clutter
  - Which measurements go with which tracks?

Tracking: issues

- **Initialization**
  - Often done manually
  - Background subtraction, detection can also be used
- **Data association**, multiple tracked objects
  - Occlusions, clutter
- **Deformable** and articulated objects

Tracking via deformable contours

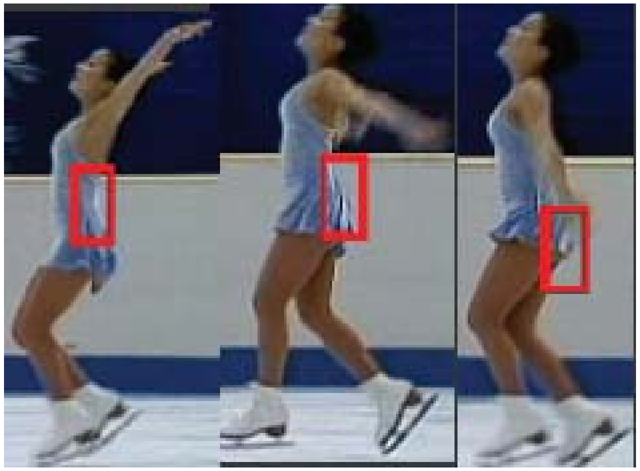
1. Use final contour/model extracted at frame *t* as an initial solution for frame *t+1*
2. Evolve initial contour to fit exact object boundary at frame *t+1*
3. Repeat, initializing with most recent frame.

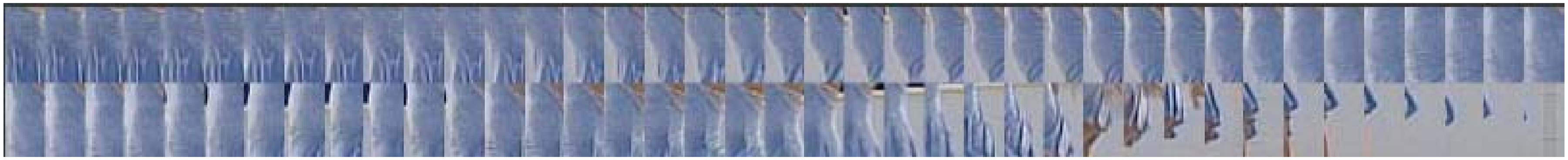
Visual Dynamics Group, Dept. Engineering Science, University of Oxford.

Tracking: issues

- **Initialization**
  - Often done manually
  - Background subtraction, detection can also be used
- **Data association**, multiple tracked objects
  - Occlusions, clutter
- **Deformable** and articulated objects
- **Constructing accurate models** of dynamics
  - E.g., Fitting parameters for a linear dynamics model
- **Drift**
  - Accumulation of errors over time

Drift





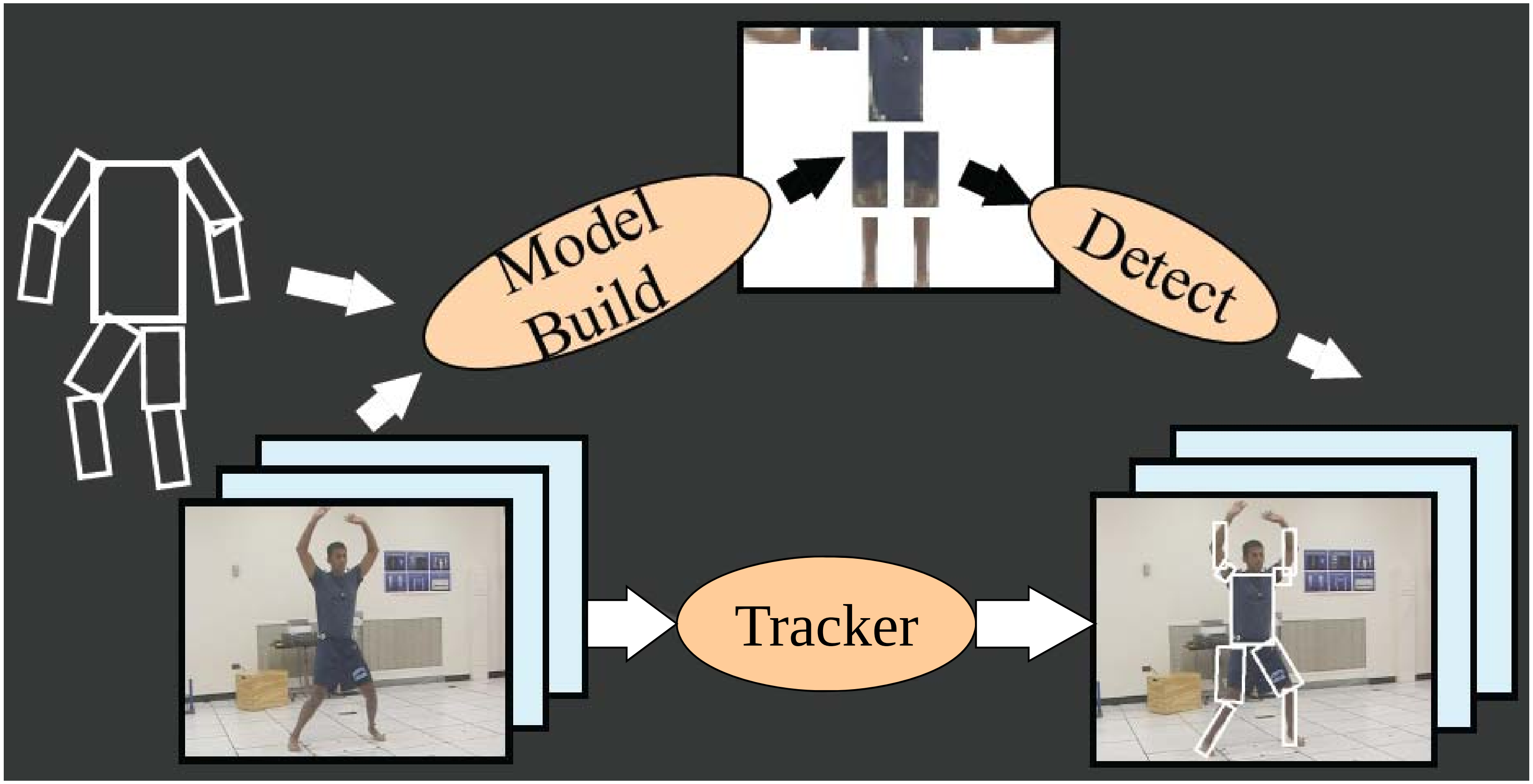
D. Ramanan, D. Forsyth, and A. Zisserman. [Tracking People by Learning their Appearance](). PAMI 2007.

Tracking people

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

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

Tracking people by learning their appearance



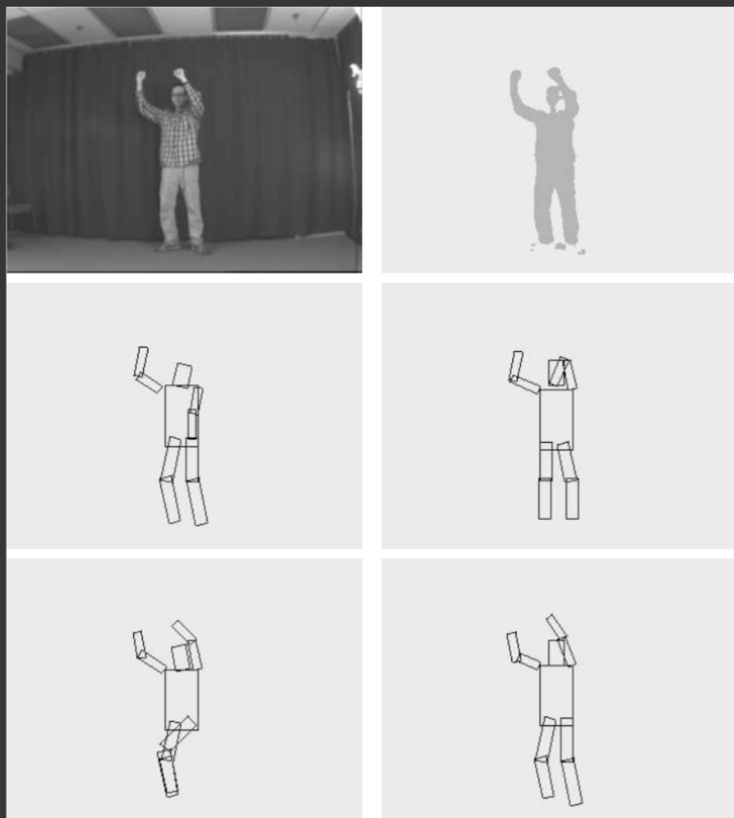
D. Ramanan, D. Forsyth, and A. Zisserman. [Tracking People by Learning their Appearance](). PAMI 2007.

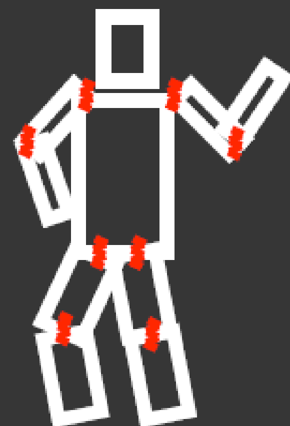
Top-down method to build model: Exploit "easy" poses

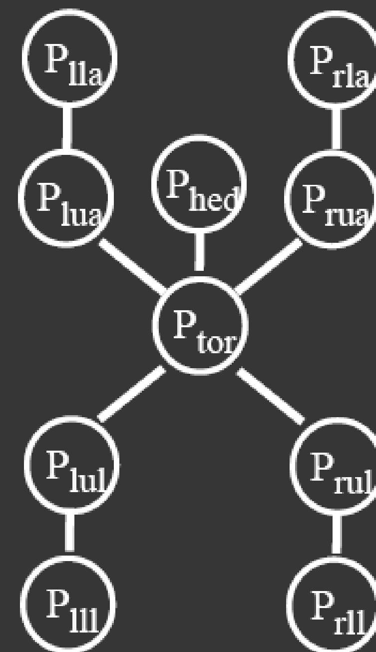
D. Ramanan, D. Forsyth, and A. Zisserman. [Tracking People by Learning their Appearance](). PAMI 2007.

Pictorial structure model

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







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

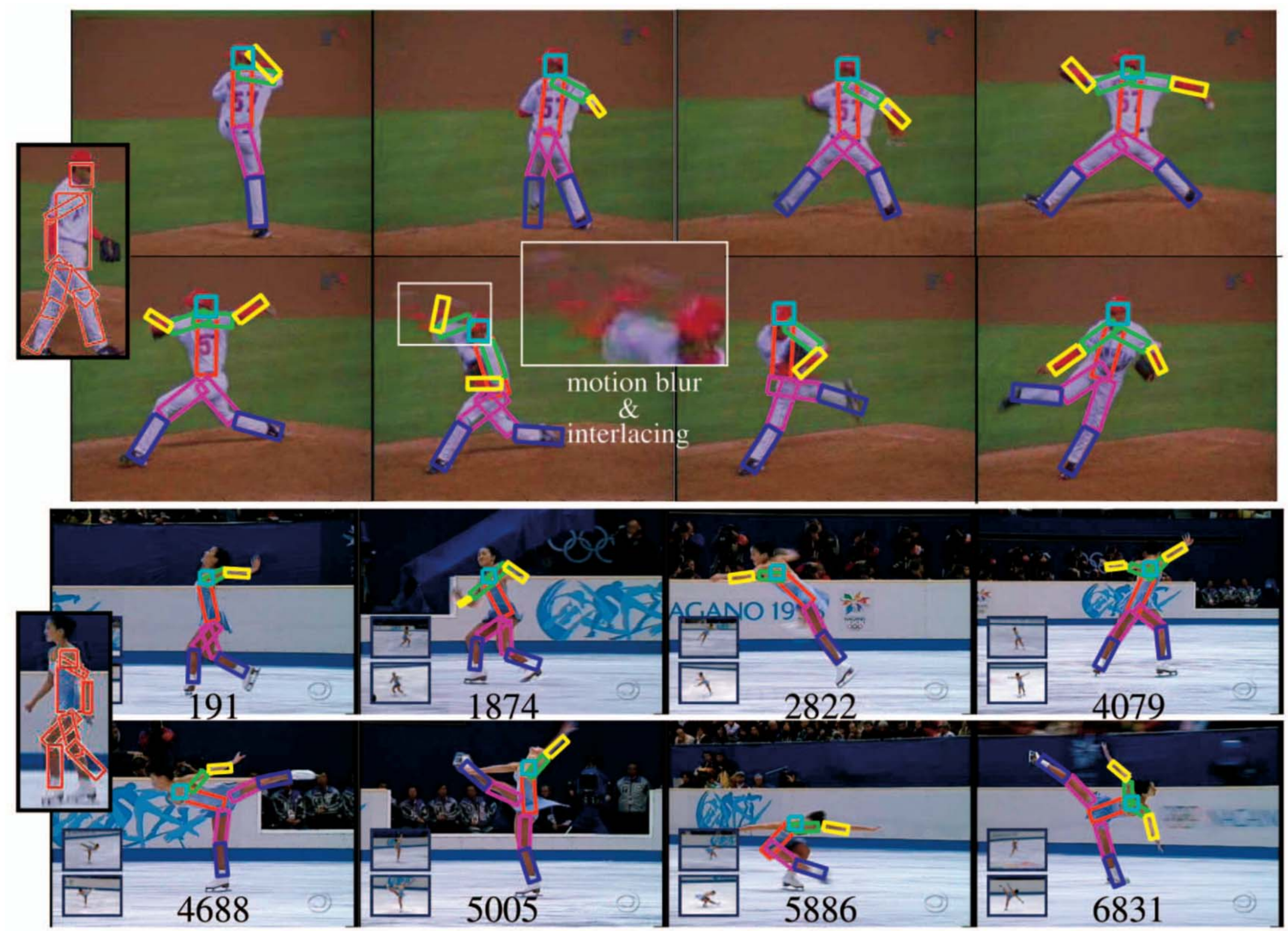
part geometry

part appearance

Temporal model

- Parts cannot move too far

Example results



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

- To predict

How can we identify actions?

Motion

Pose

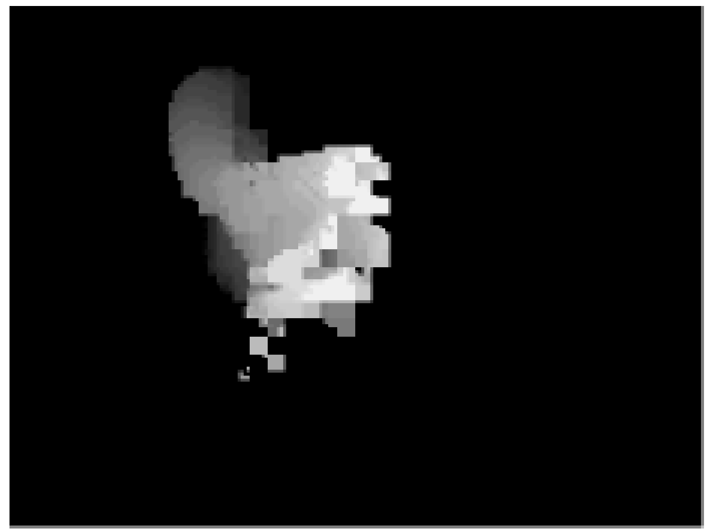
Held Objects

Nearby Objects

Representing Motion

Optical Flow with Motion History

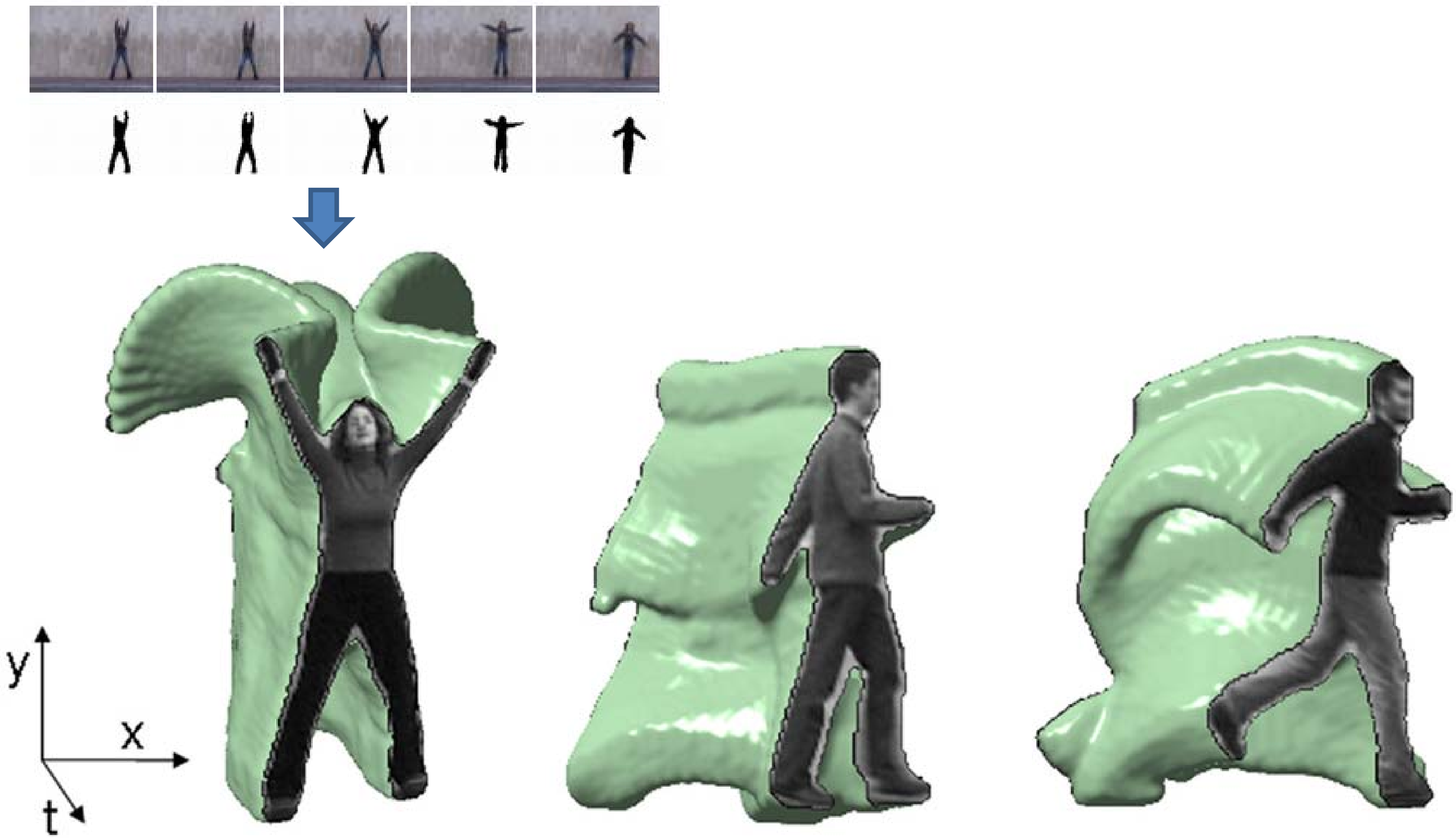
sit-down



sit-down MHI

Bobick Davis 2001

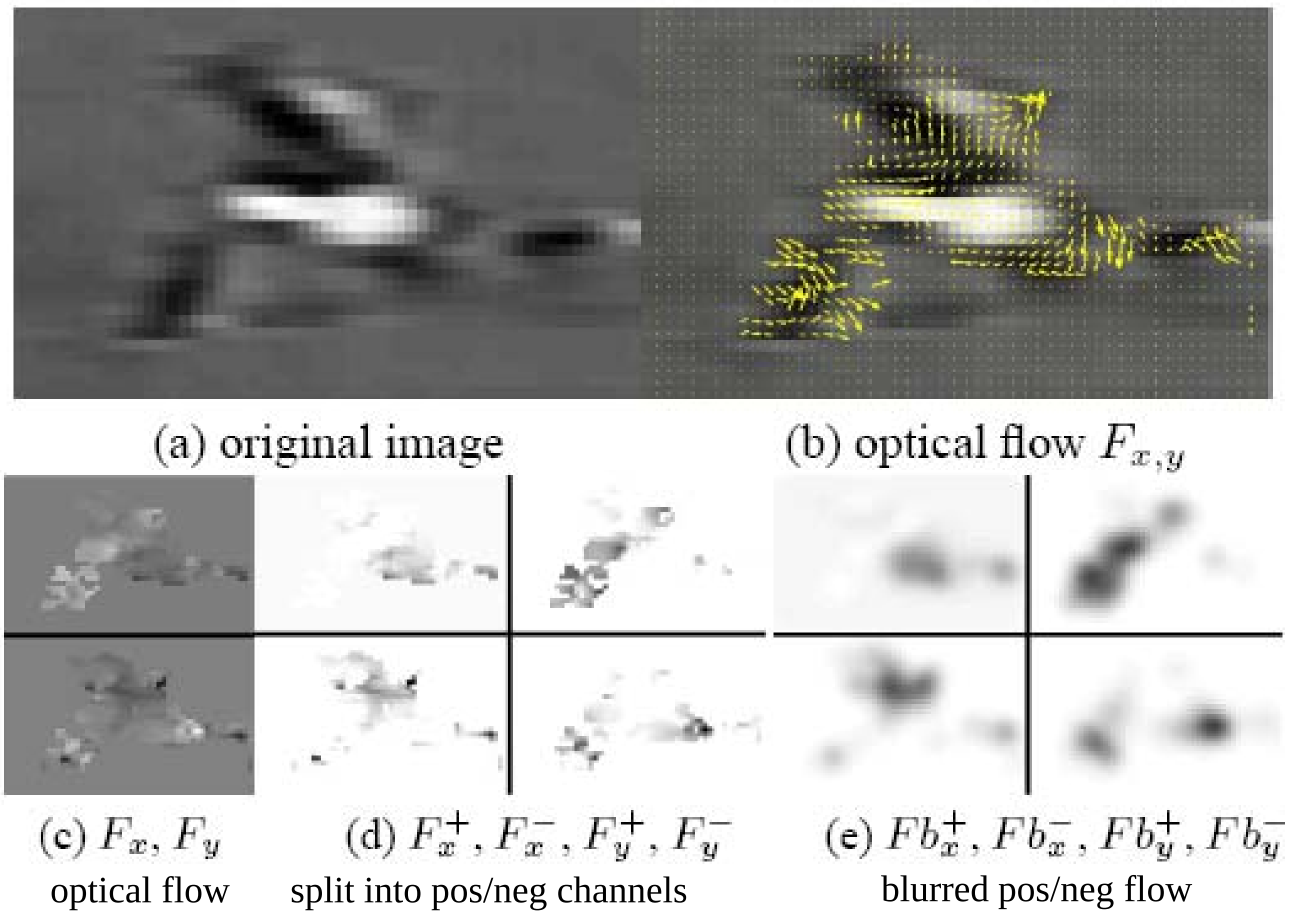
Representing Motion
Space-Time Volumes



Blank et al. 2005

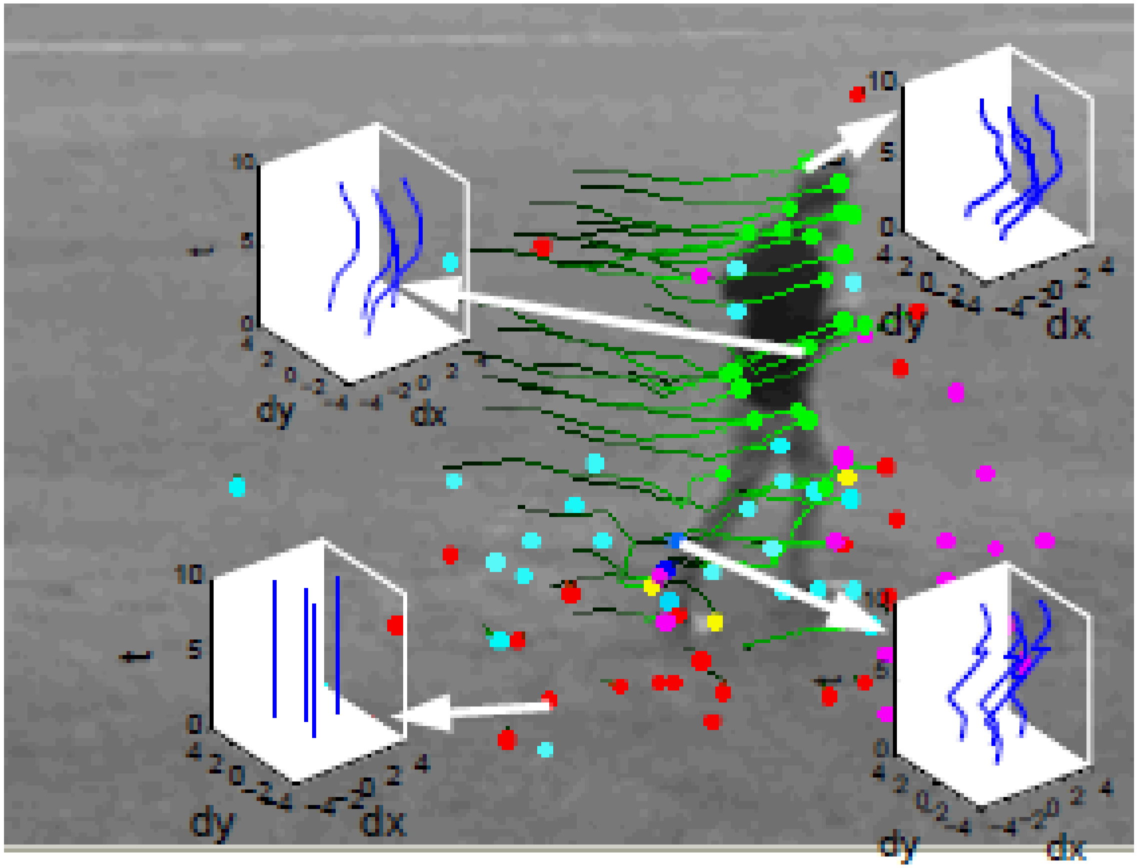
Representing Motion

Optical Flow with Split Channels



[Efros et al. 2003]

Representing Motion
Tracked Points



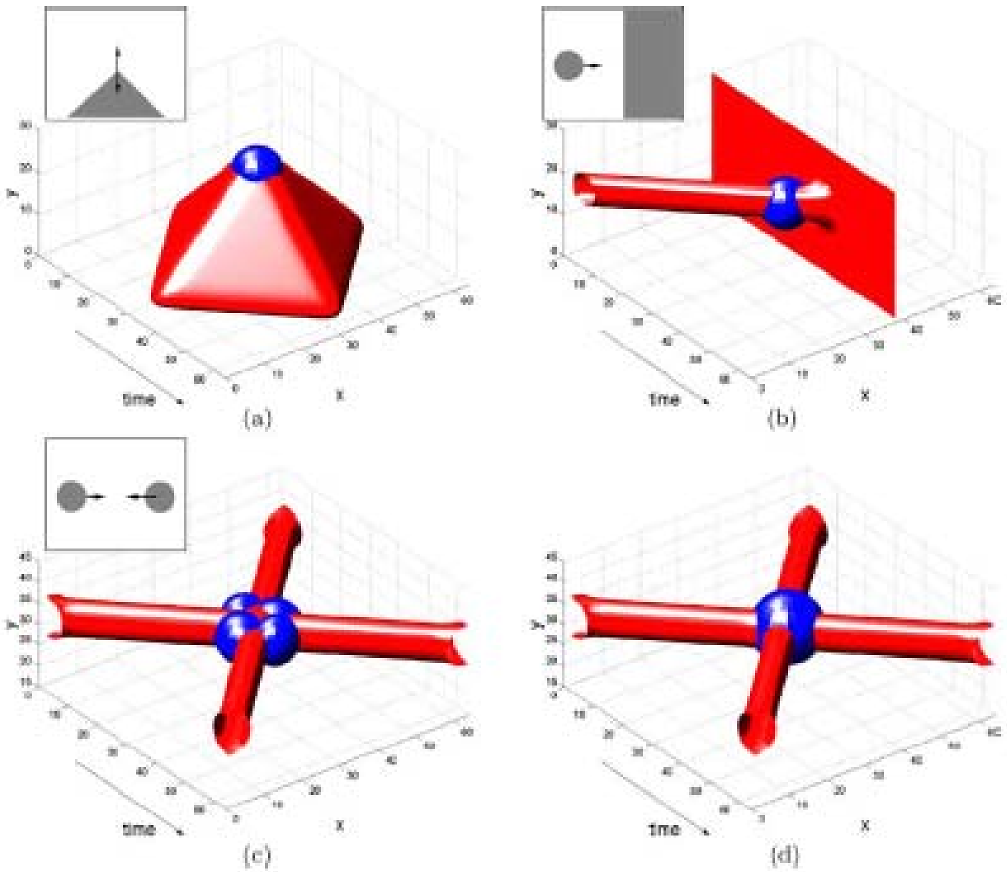
[Matikainen et al. 2009]

Representing Motion

Space-Time Interest Points

Moving corner

Ball hits wall



Balls collide

Balls collide (different scale)

Corner detectors in space-time

[Laptev 2005]

Examples of Action Recognition Systems

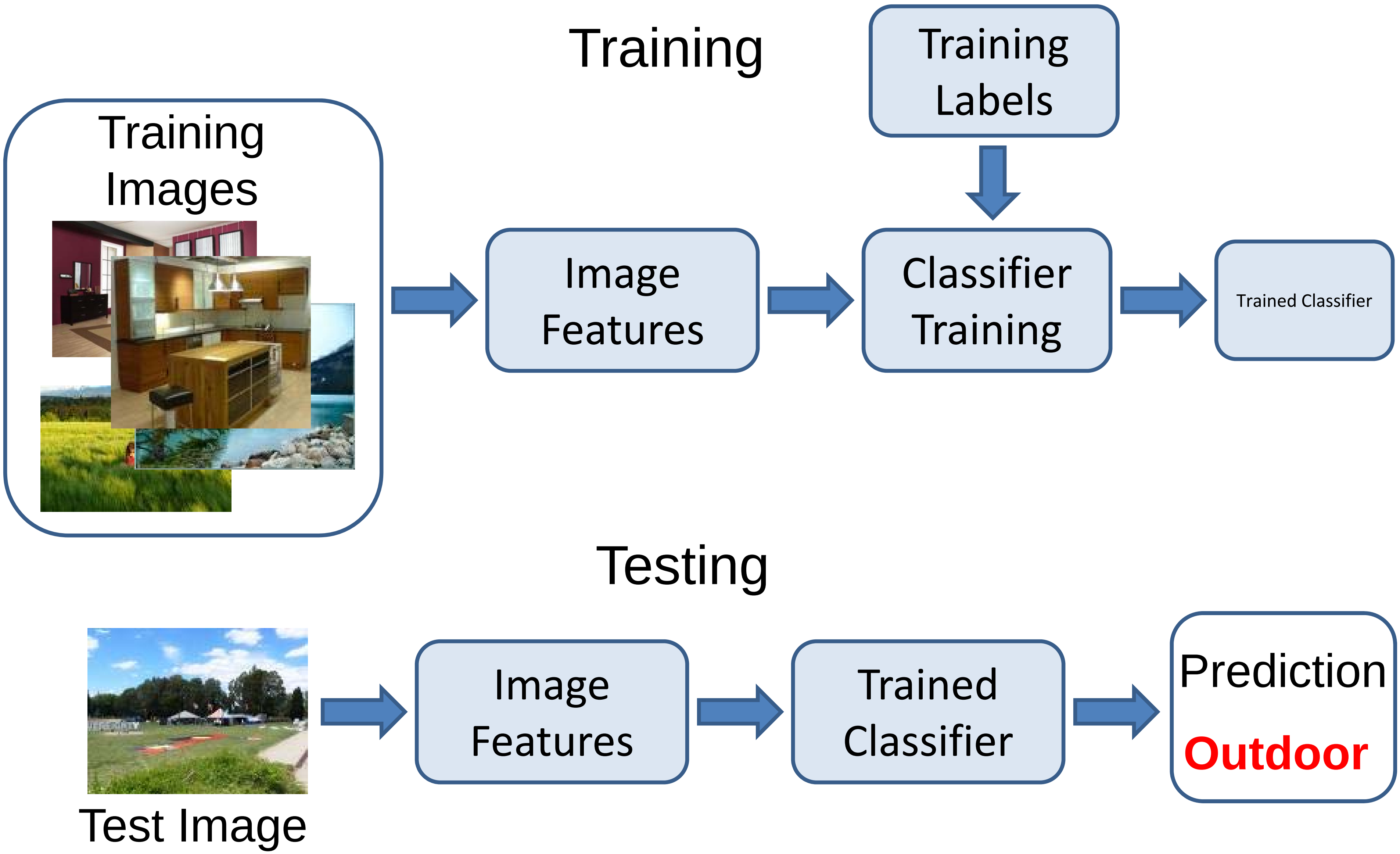
- Feature-based classification

- Recognition using pose and objects

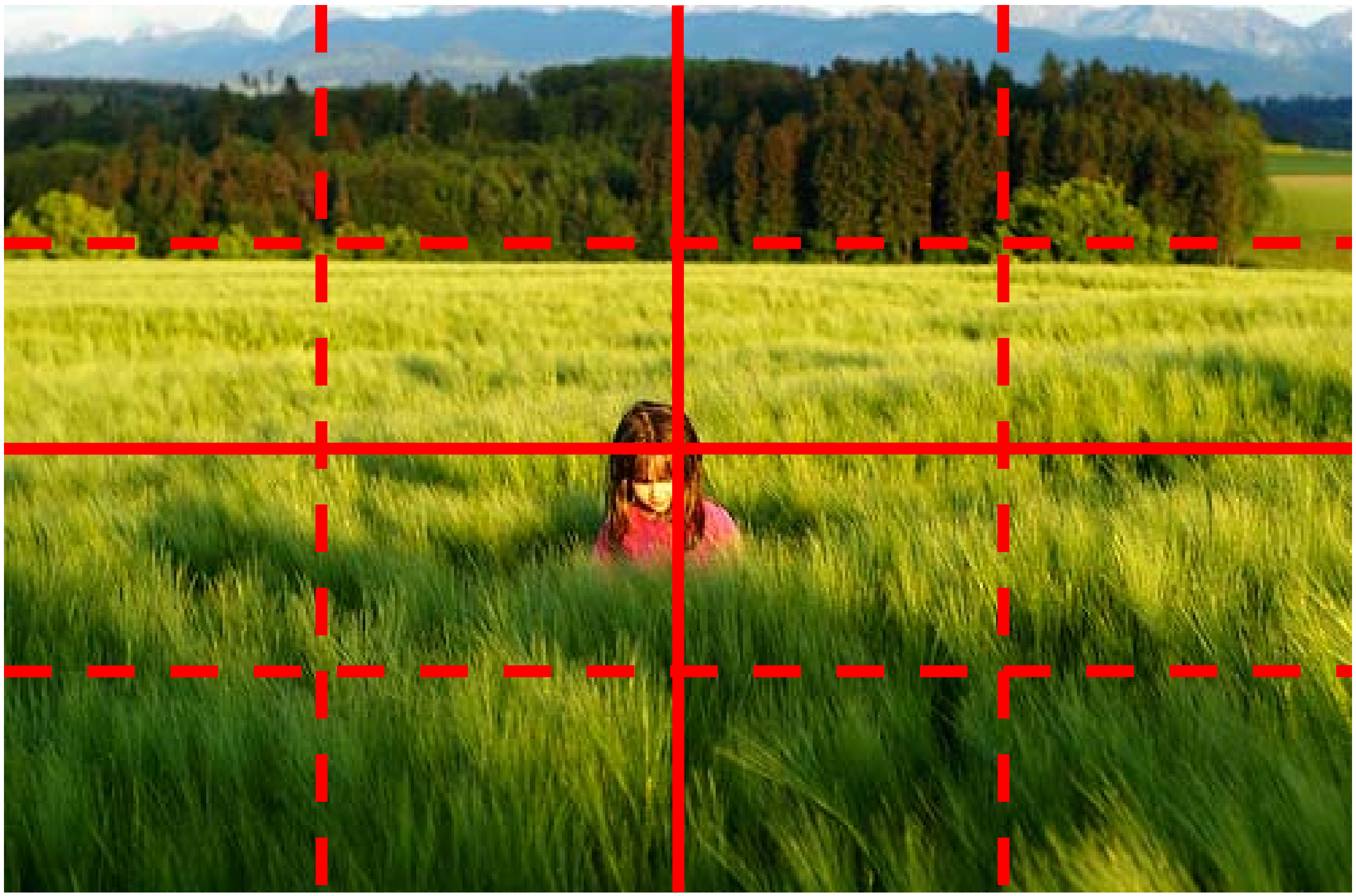
Action recognition as classification

Retrieving actions in movies, Laptev and Perez, 2007

Remember image categorization…



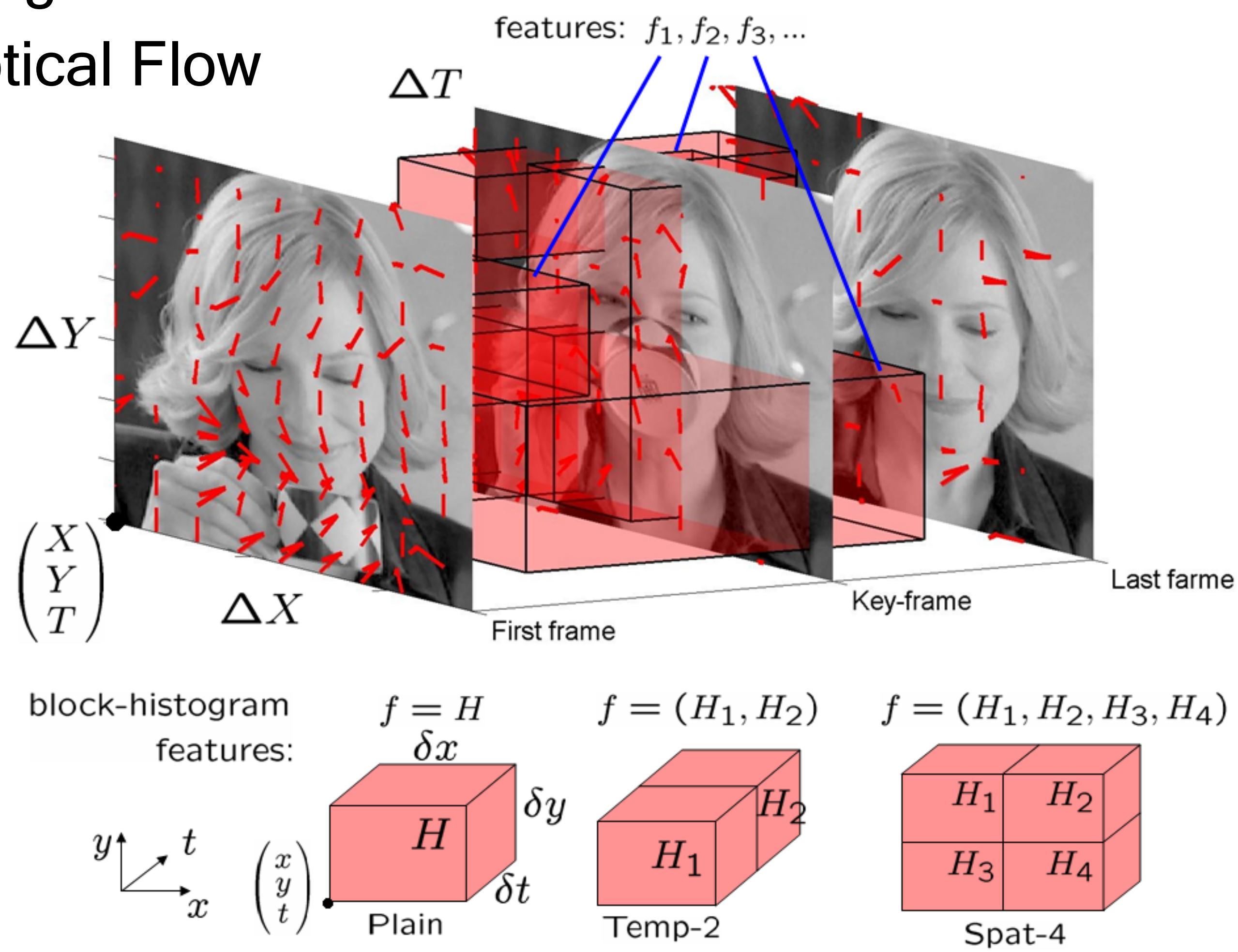
Remember spatial pyramids….



Compute histogram in each spatial bin

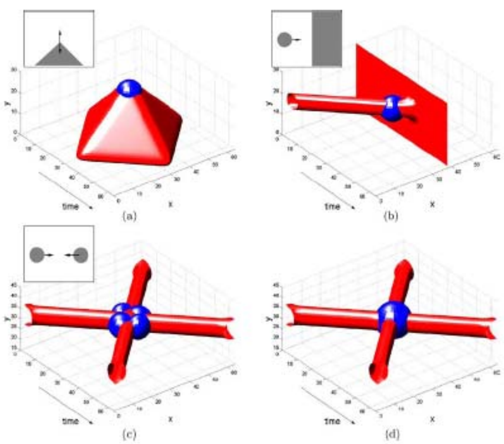
Features for Classifying Actions

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

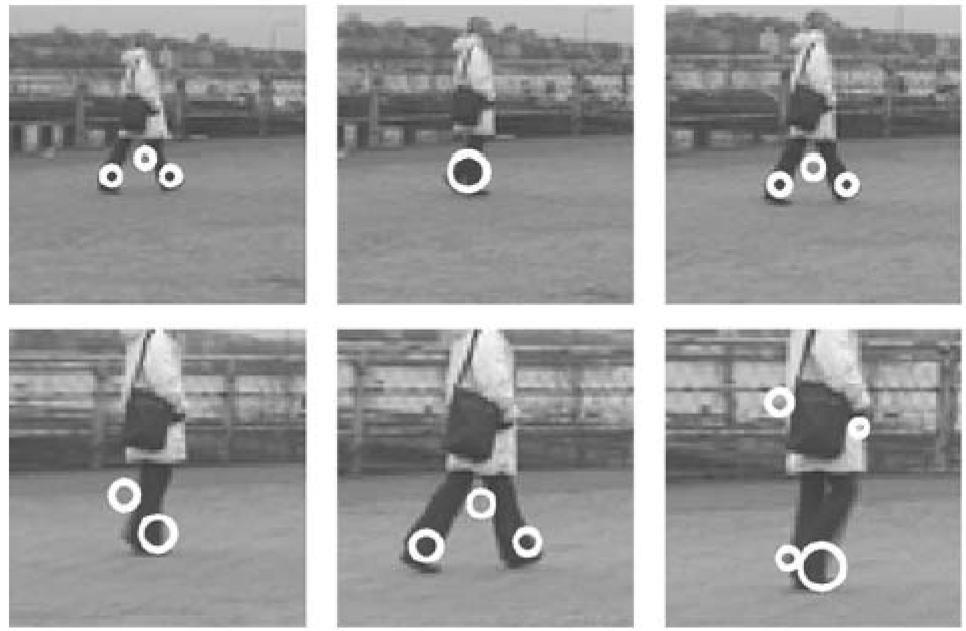


Features for Classifying Actions

2. Spatio-temporal interest points



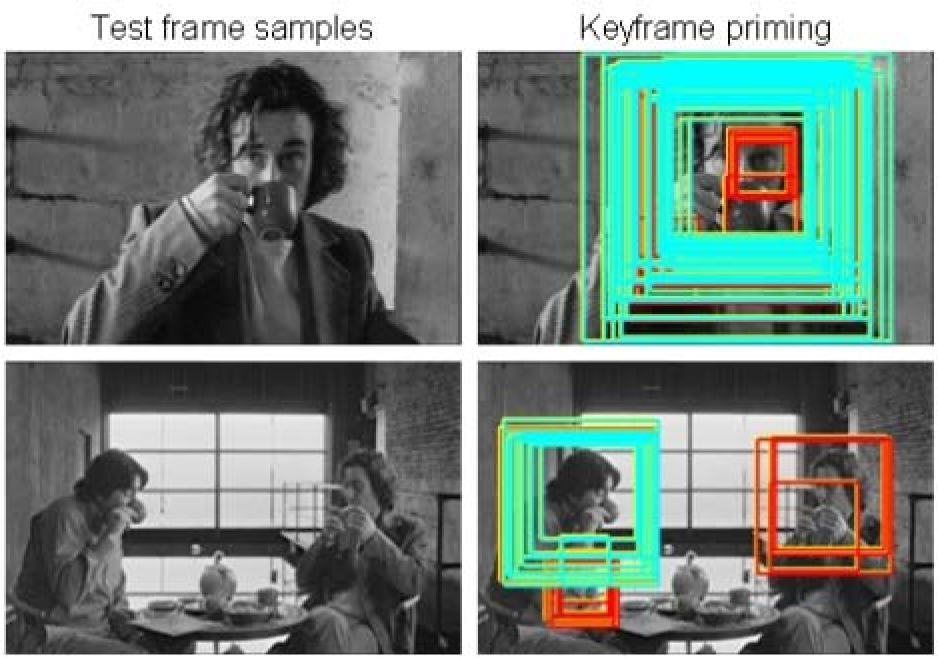
Corner detectors in space-time

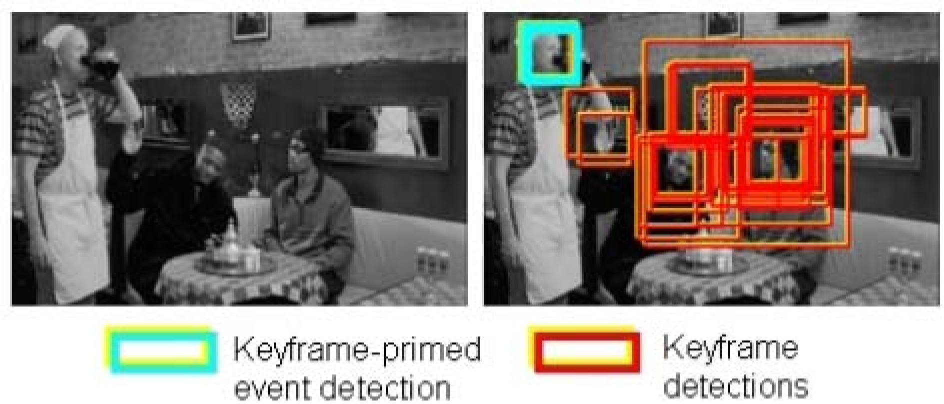


Descriptors based on Gaussian derivative filters over x, y, time

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”)

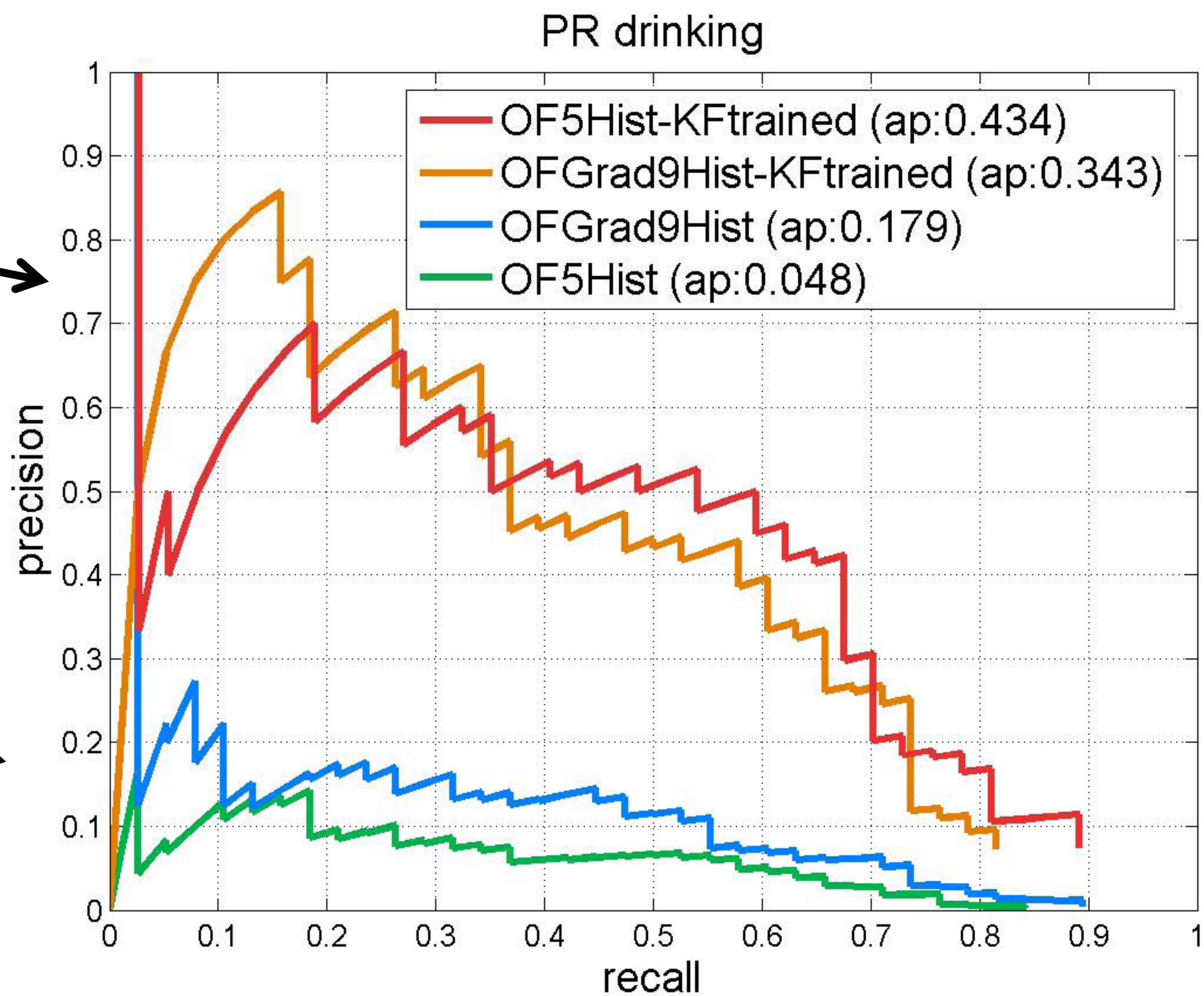




Accuracy in searching video

With keyframe detection

Without keyframe detection



“Talk on phone”

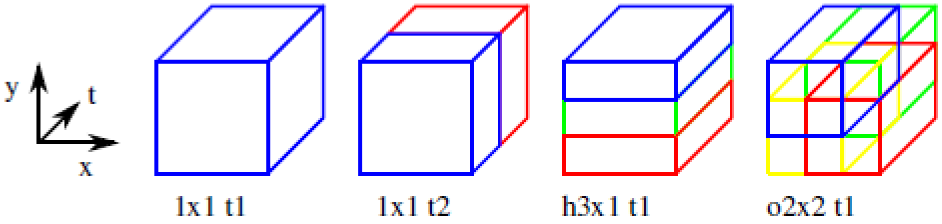
“Get out of car”

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

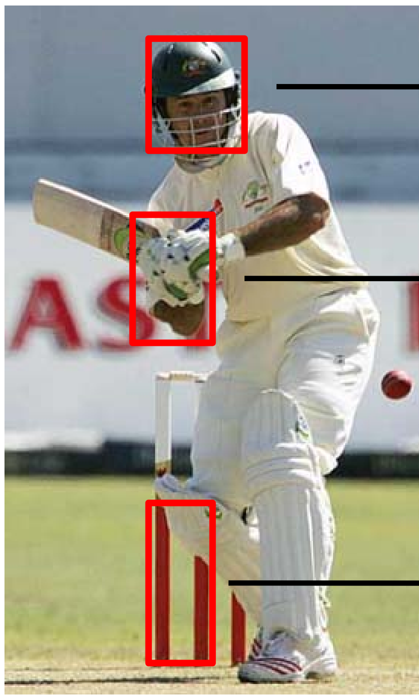
- **Human pose estimation**
- **Object detection**
- **Action categorization**

Activity: Tennis Forehand

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

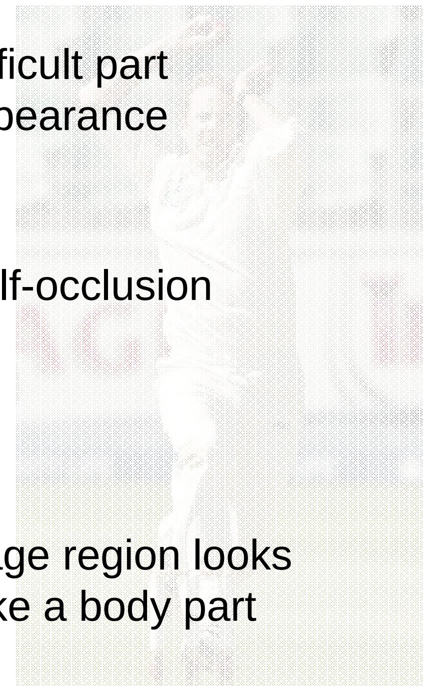
Human pose estimation is challenging.



Difficult part appearance

Self-occlusion

Image region looks like a body part

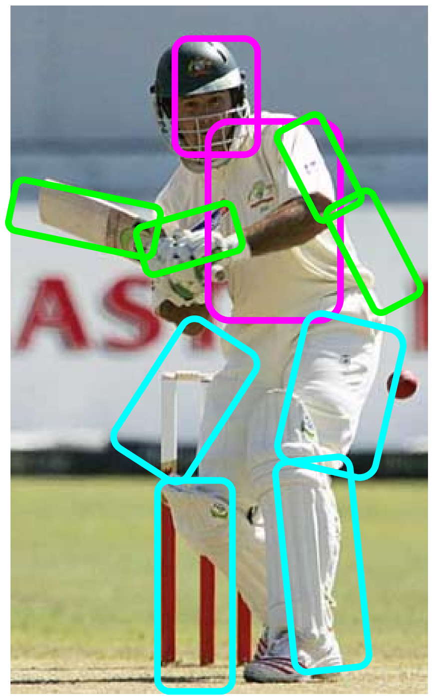


- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Human pose estimation is challenging.



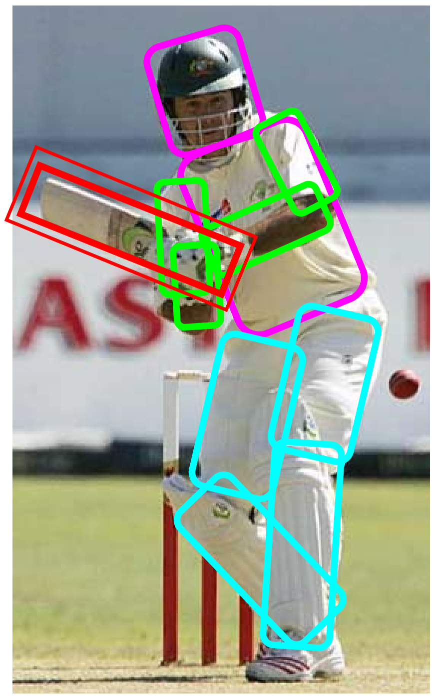
- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

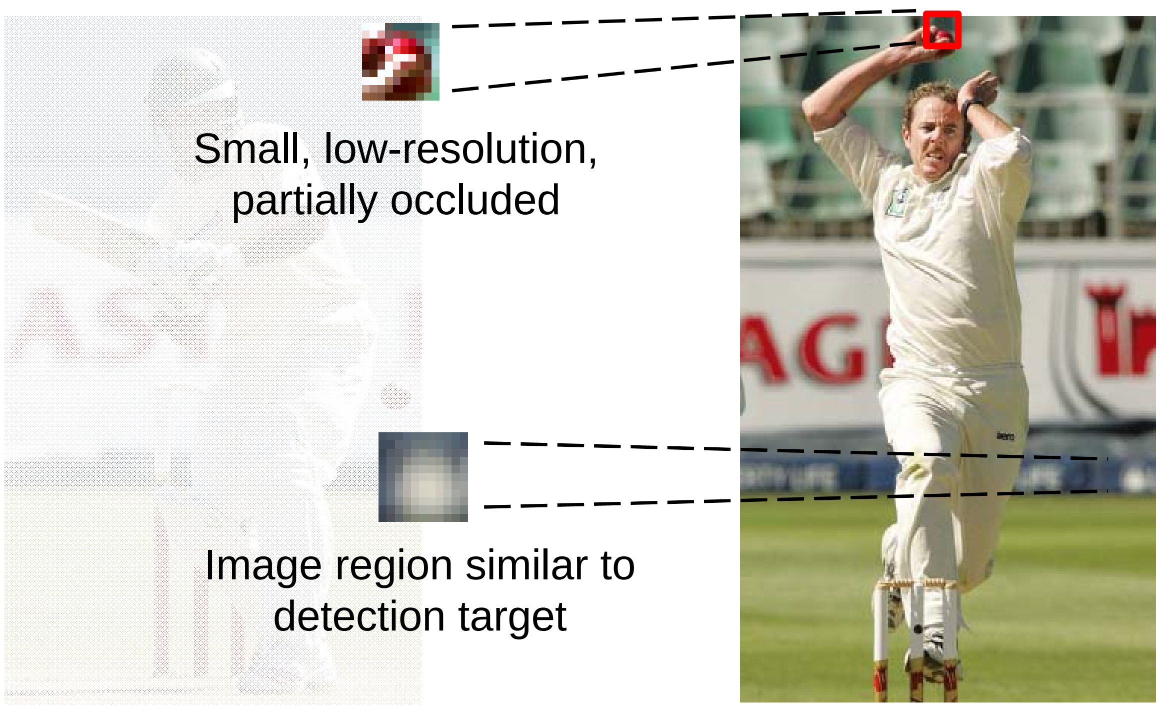
Facilitate

Given the object is detected



Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

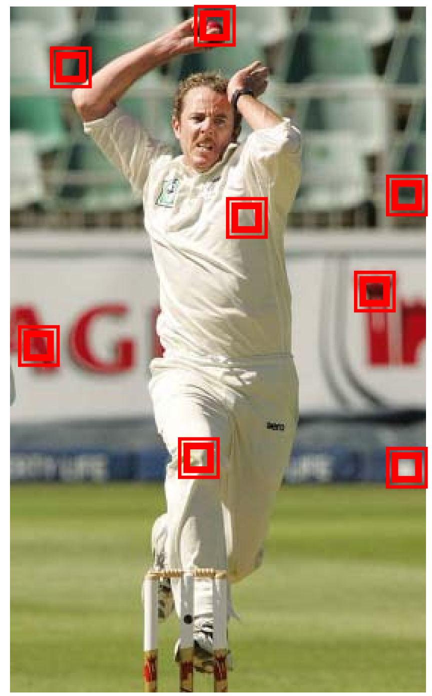


Object detection is challenging

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

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection



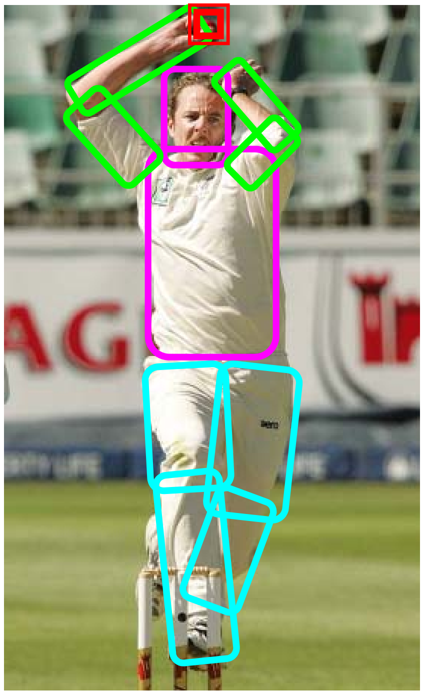
Object detection is challenging

- 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

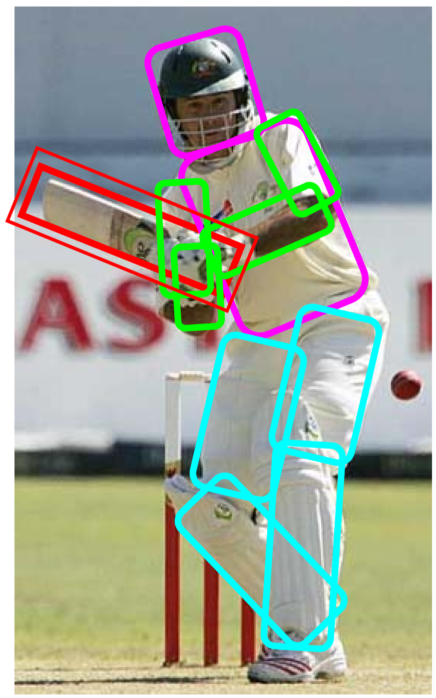


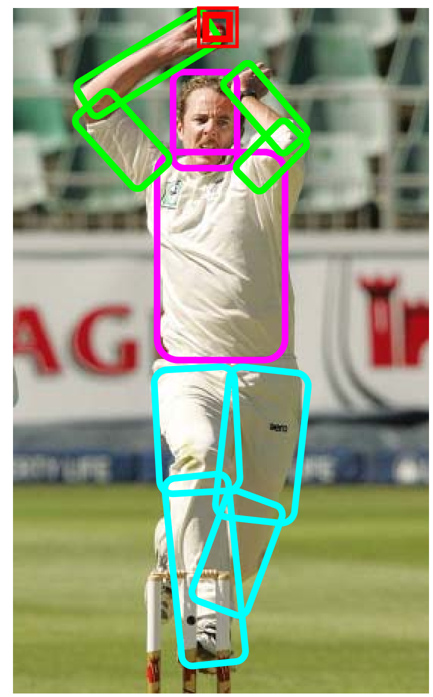
Given the pose is estimated

Slide Credit: Yao/Fei-Fei

Human pose estimation & Object detection

Mutual Context





Slide Credit: Yao/Fei-Fei

Mutual Context Model Representation

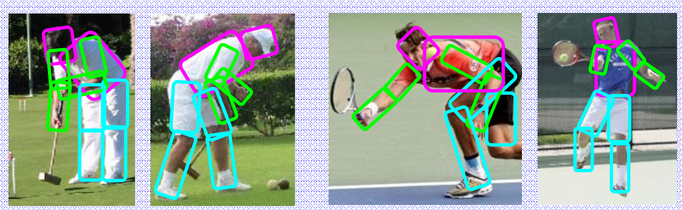
A:

Tennis forehand | Croquet shot | Volleyball smash | …

O:

Tennis racket | Croquet mallet | Volleyball | …

H:

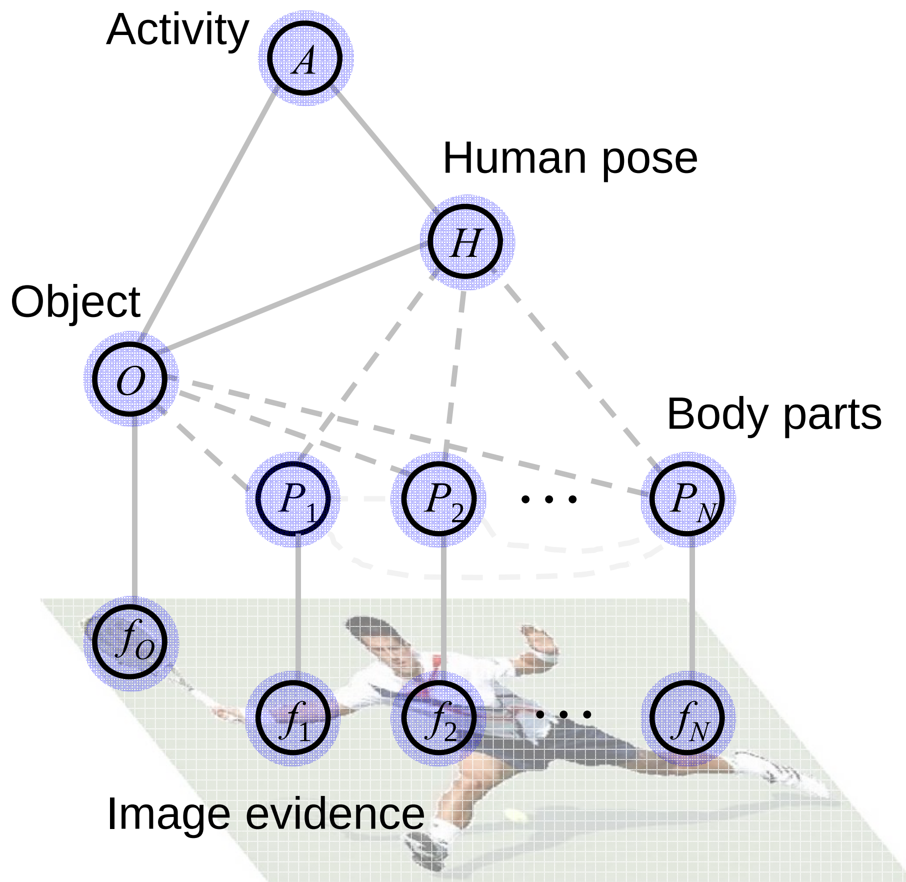


Intra-class variations

- 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

Learning Results

Cricket defensive shot

Cricket bowling

Croquet shot

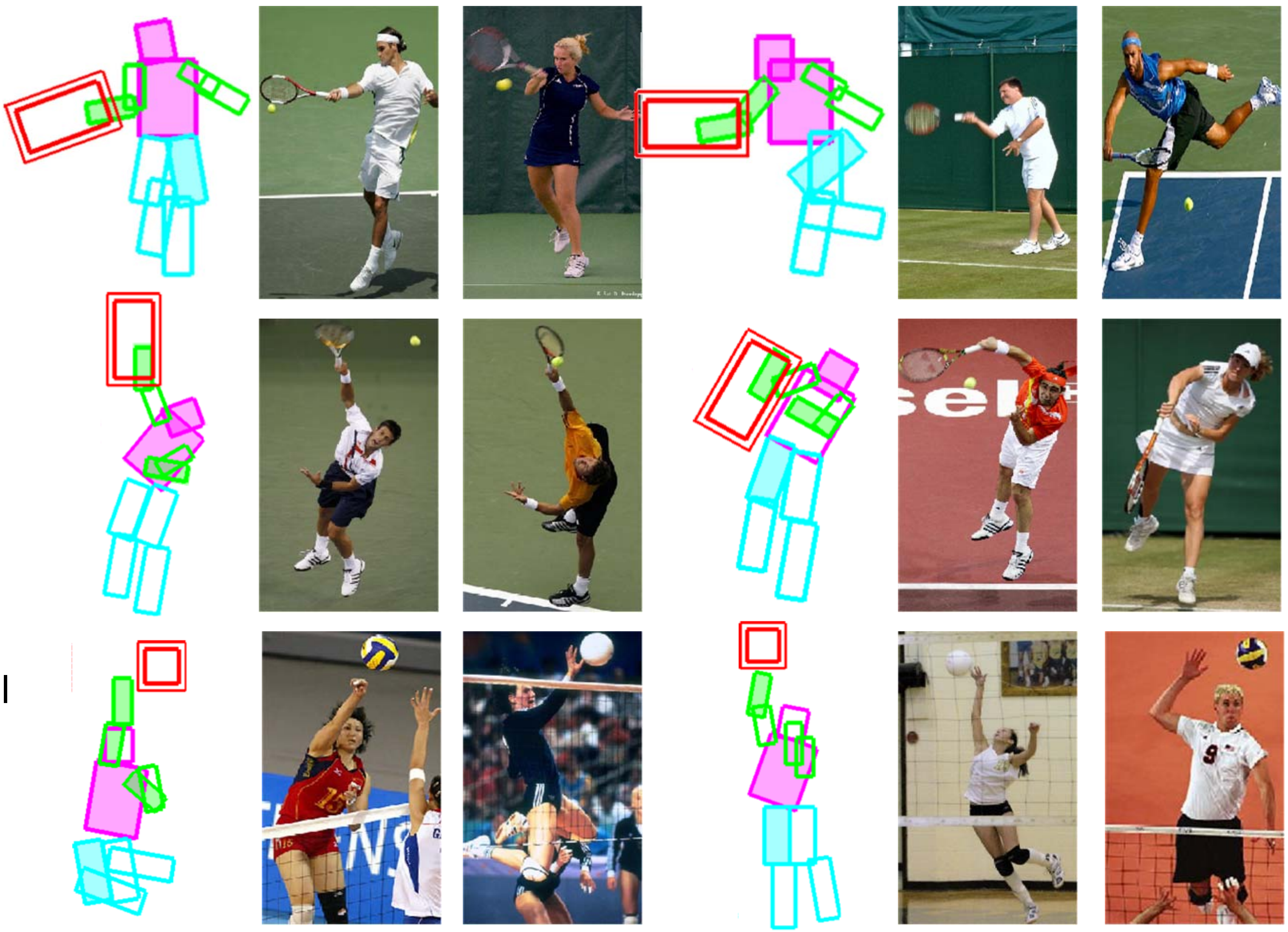
Slide Credit: Yao/Fei-Fei

Learning Results

Tennis forehand

Tennis serve

Volleyball smash



Slide Credit: Yao/Fei-Fei

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

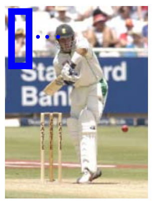
[Gupta et al, 2009]

Tasks:

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

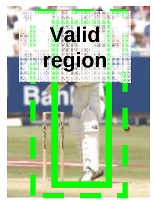
Slide Credit: Yao/Fei-Fei

Object Detection Results



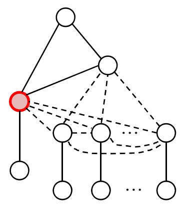
Sliding window

[Andriluka et al, 2009]



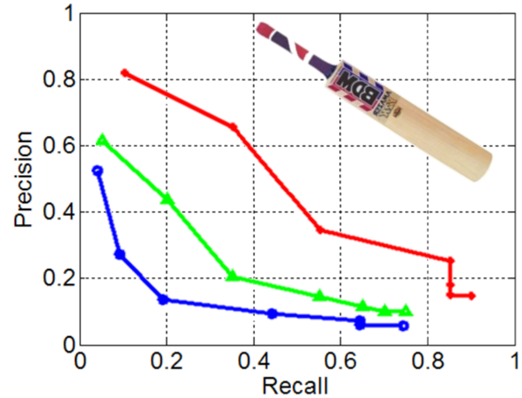
Pedestrian context

[Dalal & Triggs, 2006]

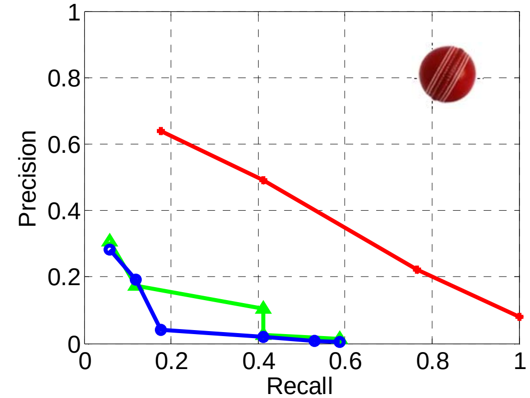


Our Method

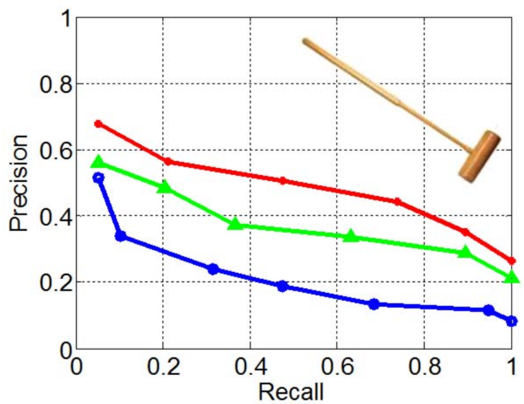
Cricket bat



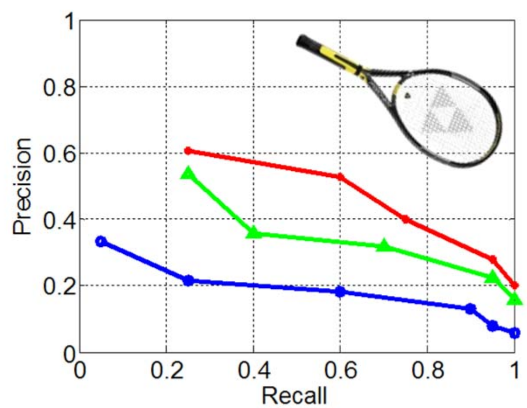
Cricket ball



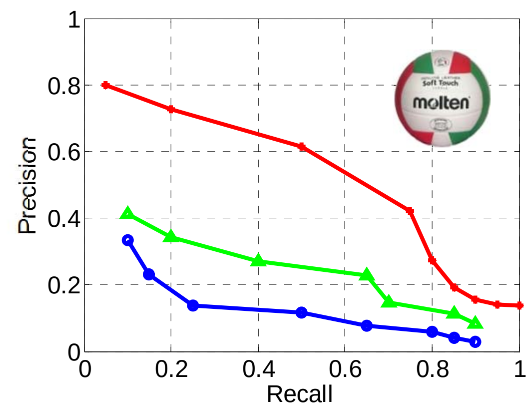
Croquet mallet



Tennis racket

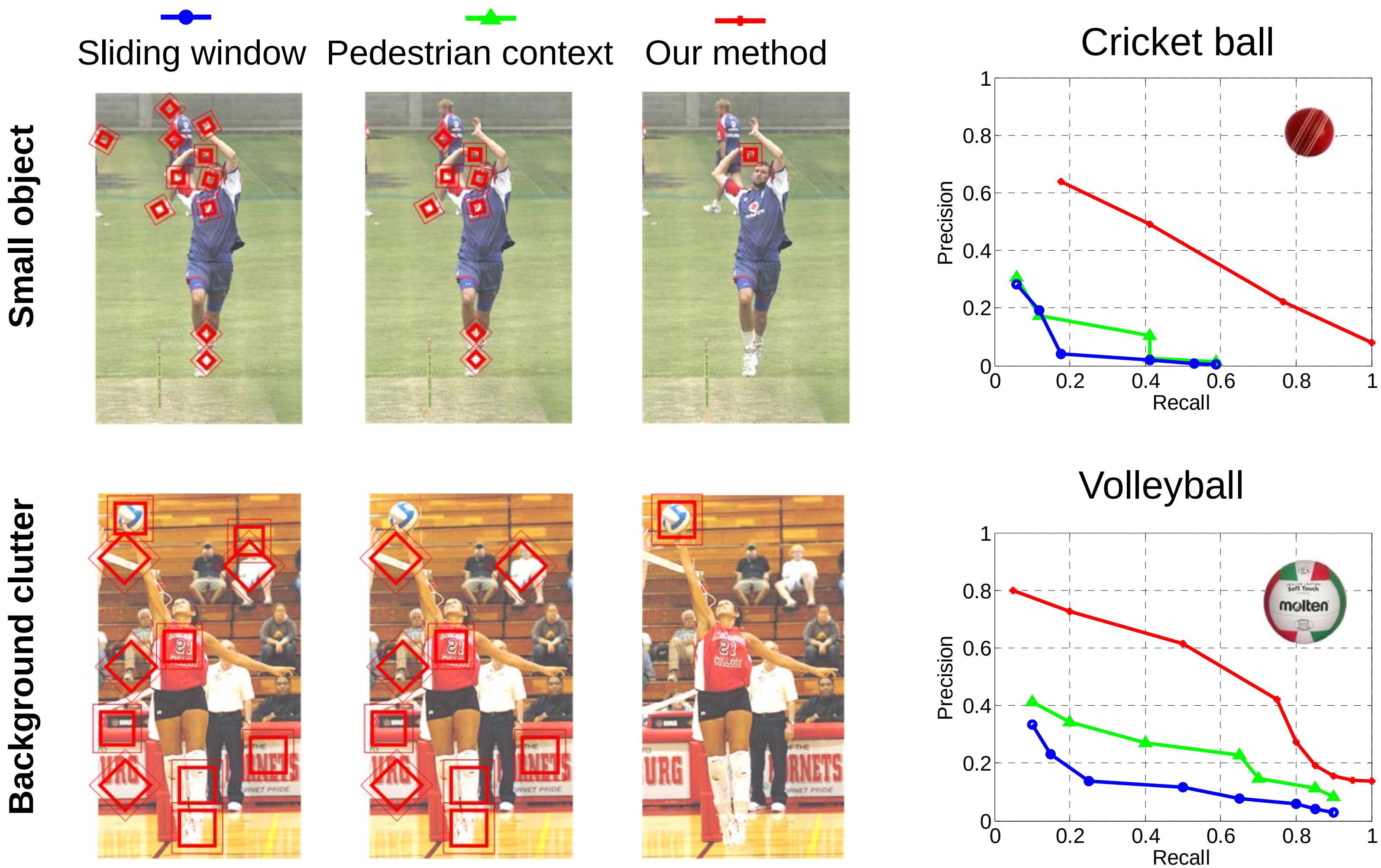


Volleyball



Slide Credit: Yao/Fei-Fei

Object Detection Results



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

Human Pose Estimation Results

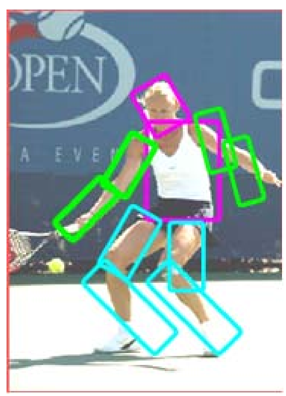
Method	Torso	Upper Leg		Lower Leg		Upper Arm		Lower Arm		Head
Ramanan, 2006	.52	.22	.22	.21	.28	.24	.28	.17	.14	.42
Andriluka et al, 2009	.50	.31	.30	.31	.27	.18	.19	.11	.11	.45
Our full model	**.66**	**.43**	**.39**	**.44**	**.34**	**.44**	**.40**	**.27**	**.29**	**.58**

Slide Credit: Yao/Fei-Fei

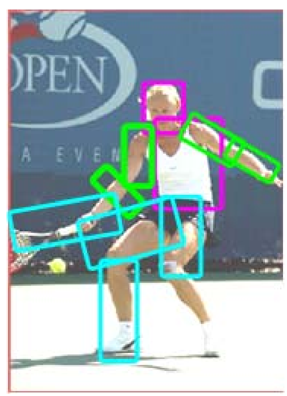
Human Pose Estimation Results

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

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

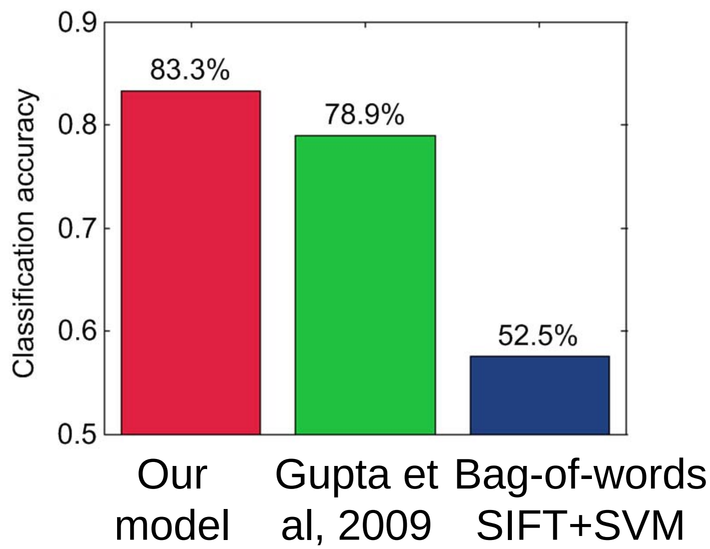
[Gupta et al, 2009]

Tasks:

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

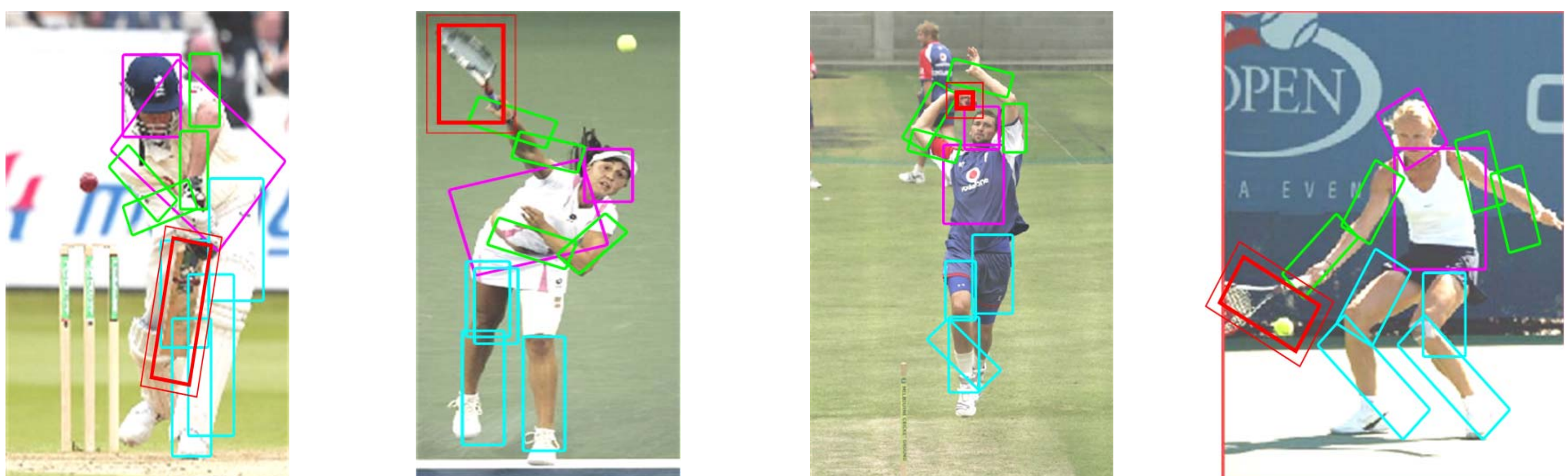
Slide Credit: Yao/Fei-Fei

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)
  - 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

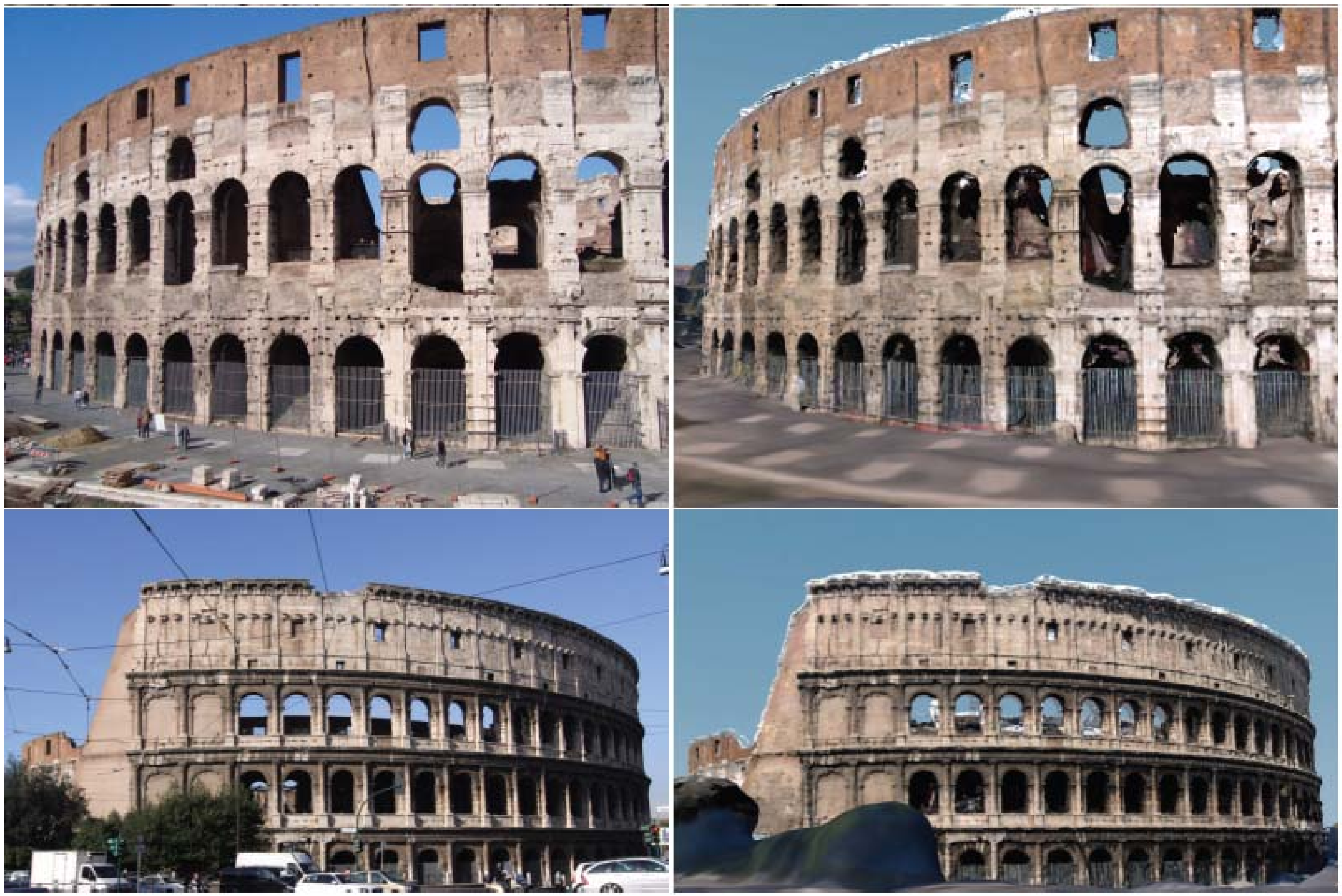
Sources of inspiration

… at least for me

The Campanile Movie

Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik
University of California at Berkeley

The Visual Turing Test for Scene Reconstruction



Qi Shan, Riley Adams, Brian Curless, Yasutaka Furukawa, and Steven M. Seitz
University of Washington and Google Inc.

High-Quality Streamable Free-Viewpoint Video

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese,
Hugues Hoppe, Adam Kirk, Steve Sullivan
MICROSOFT CORPORATION