3D Reconstruction with Computer Vision
Meeting 15: Motion

CS 378 Fall 2014, UT Austin, Bryan Klingner, 16 October
Slides by Kristen Grauman, Alexei Efros, and others

Using optical flow: recognizing facial expressions

Disgust

Sadness

happiness

fear

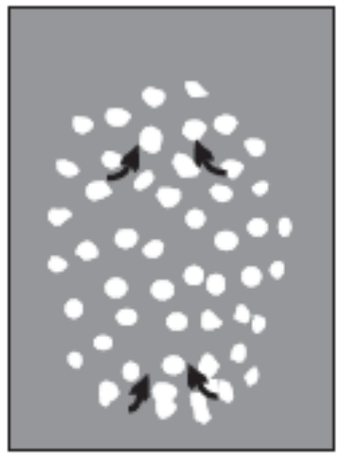
Anger

Surprise

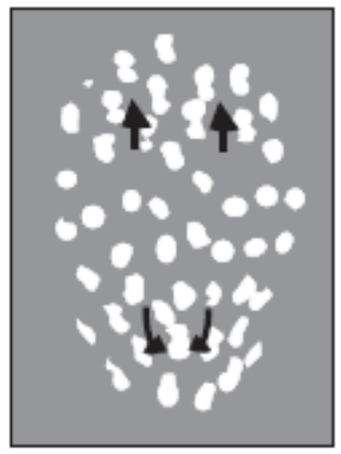
Recognizing Human Facial Expression (1994)

by Yaser Yacoob, Larry S. Davis

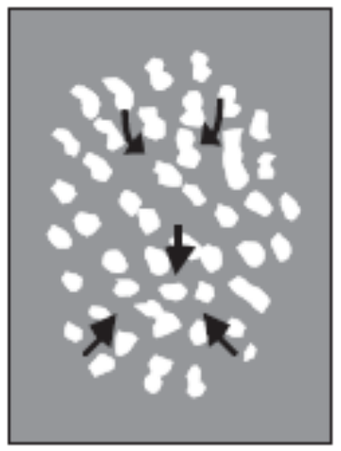
Happiness



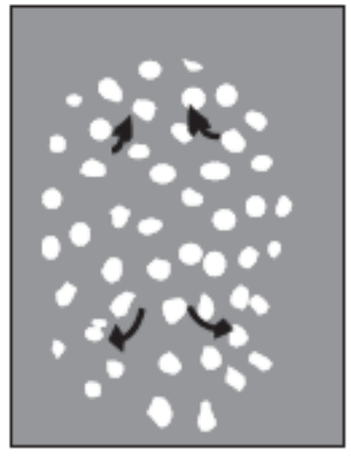
Sadness



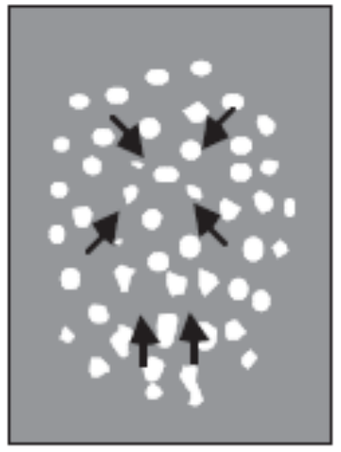
Surprise



Anger



Fear



Disgust

Applying optical flow: video stabilization

http://googleresearch.blogspot.com/2012/05/video-stabilization-on-youtube.html

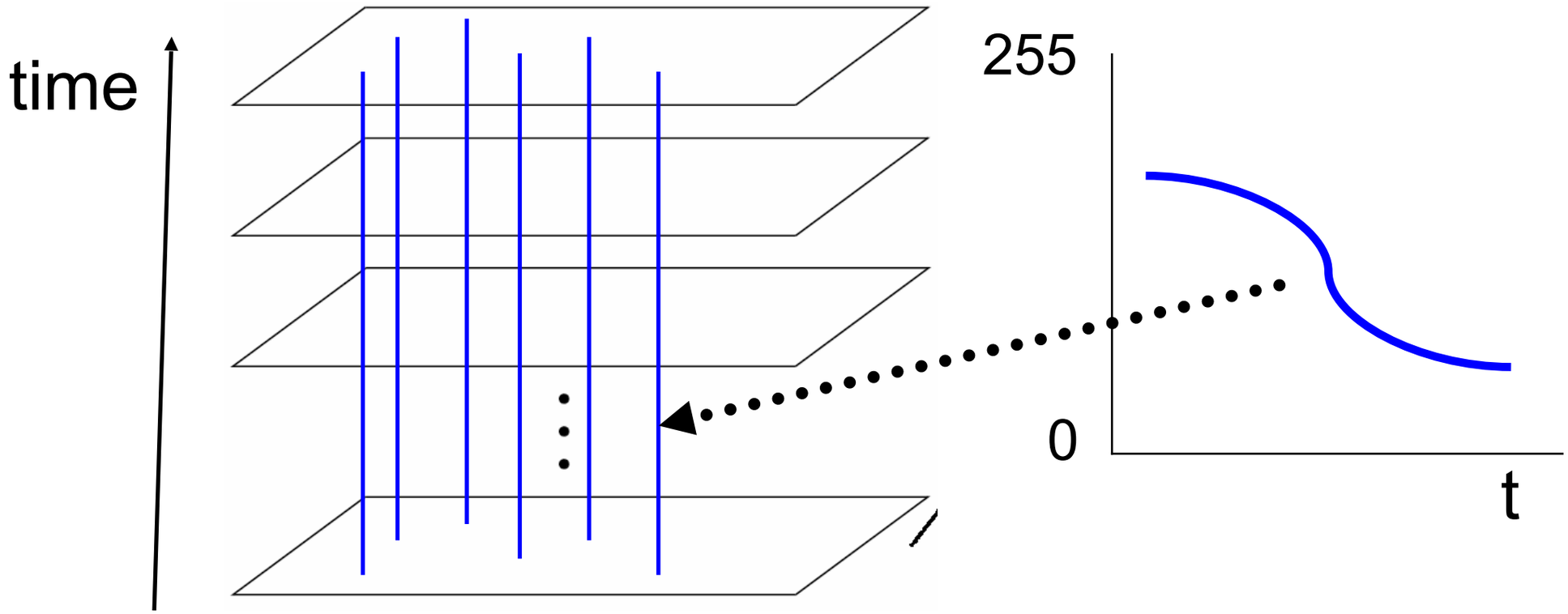
Applying optical flow: video stabilization

http://googleresearch.blogspot.com/2012/05/video-stabilization-on-youtube.html

Applying optical flow: video stabilization

Applying optical flow: video stabilization

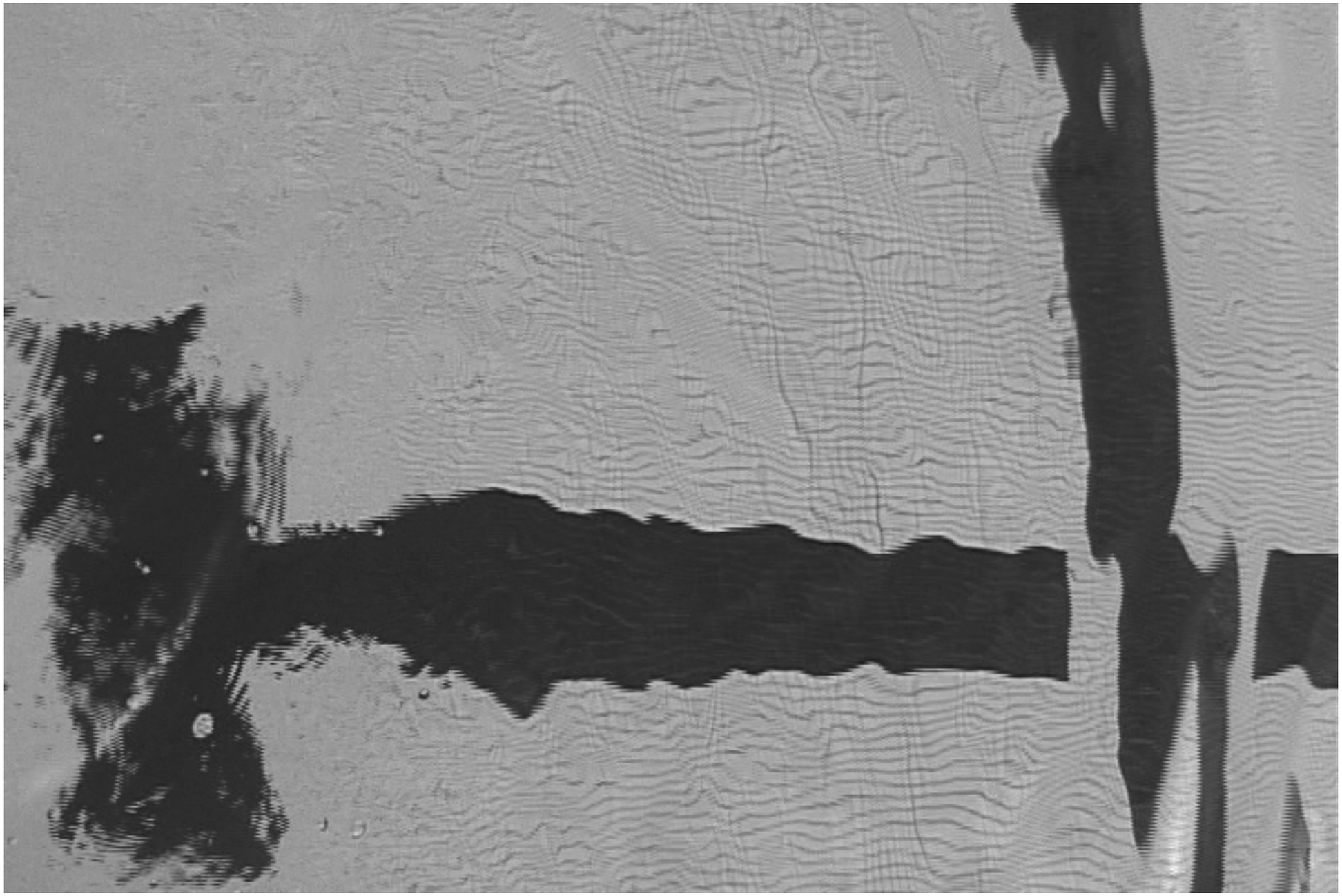
Video as an “Image Stack”



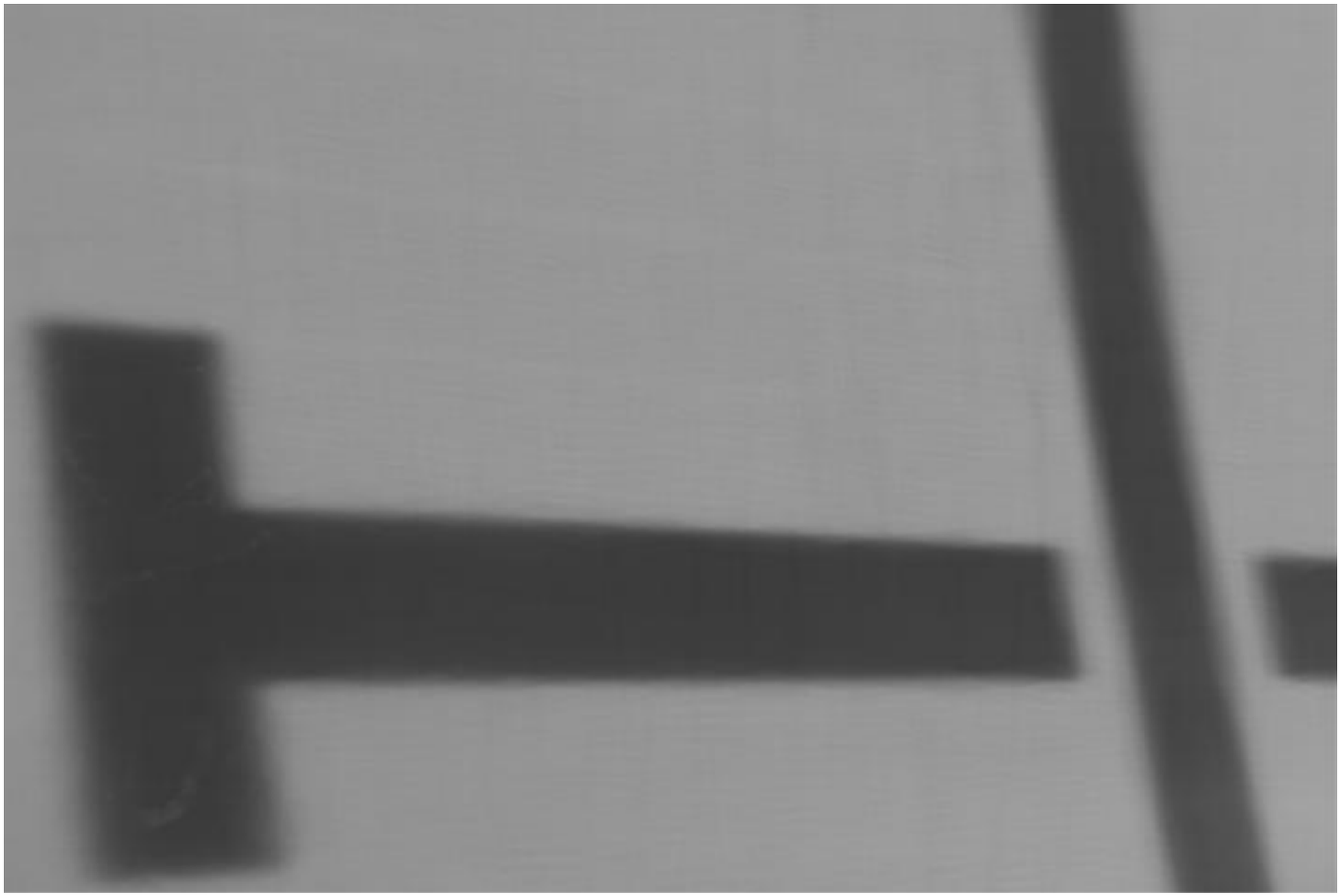
Can look at video data as a spatio-temporal volume

- If camera is stationary, each line through time corresponds to a single ray in space

Input Video



Average Image



Background Subtraction

- Given an image (mostly likely to be a video frame), we want to identify the **foreground objects** in that image!

$\Rightarrow$

Motivation

- In most cases, objects are of interest, not the scene.
- Makes our life easier: less 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

Image at time t:

$I(x, y, t)$

$\Downarrow$

$|$ $-$

Background at time t:

$B(x, y, t)$

$\Downarrow$

$| > Th$

1. Estimate the background for time t.
2. Subtract the estimated background from the input frame.
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 then becomes:

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

$$\Downarrow$$

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

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

$|$  $-$  $| > Th$

Slide credit: Birgi Tamersoy

Frame Differencing

$Th = 25$

$Th = 50$

$Th = 100$

$Th = 200$

Slide credit: Birgi Tamersoy

Mean Filter

- In this case the background is the mean of the previous n frames:

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

$$\Downarrow$$

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

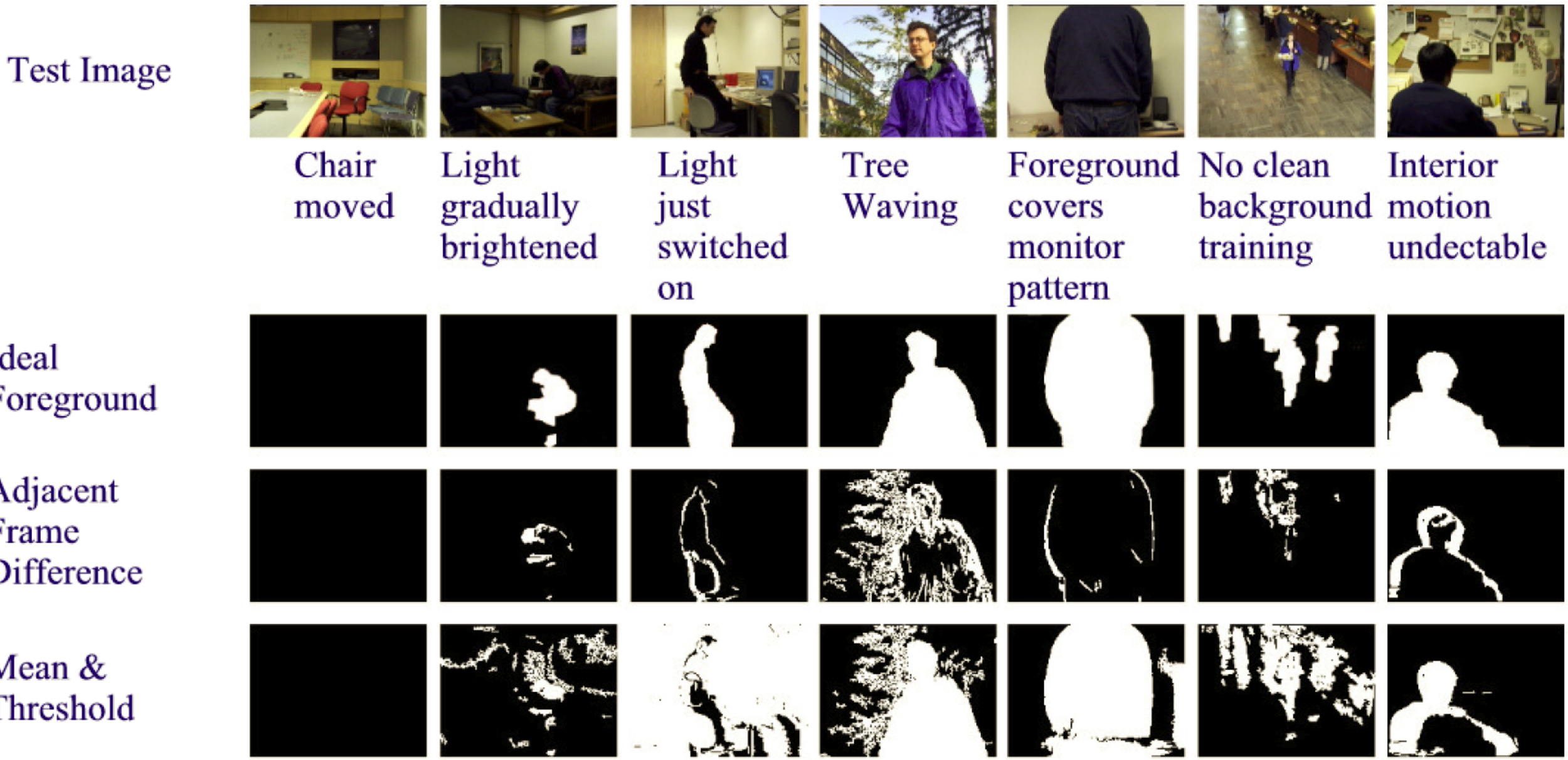
- For $n = 10$:

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 $n = 10$:

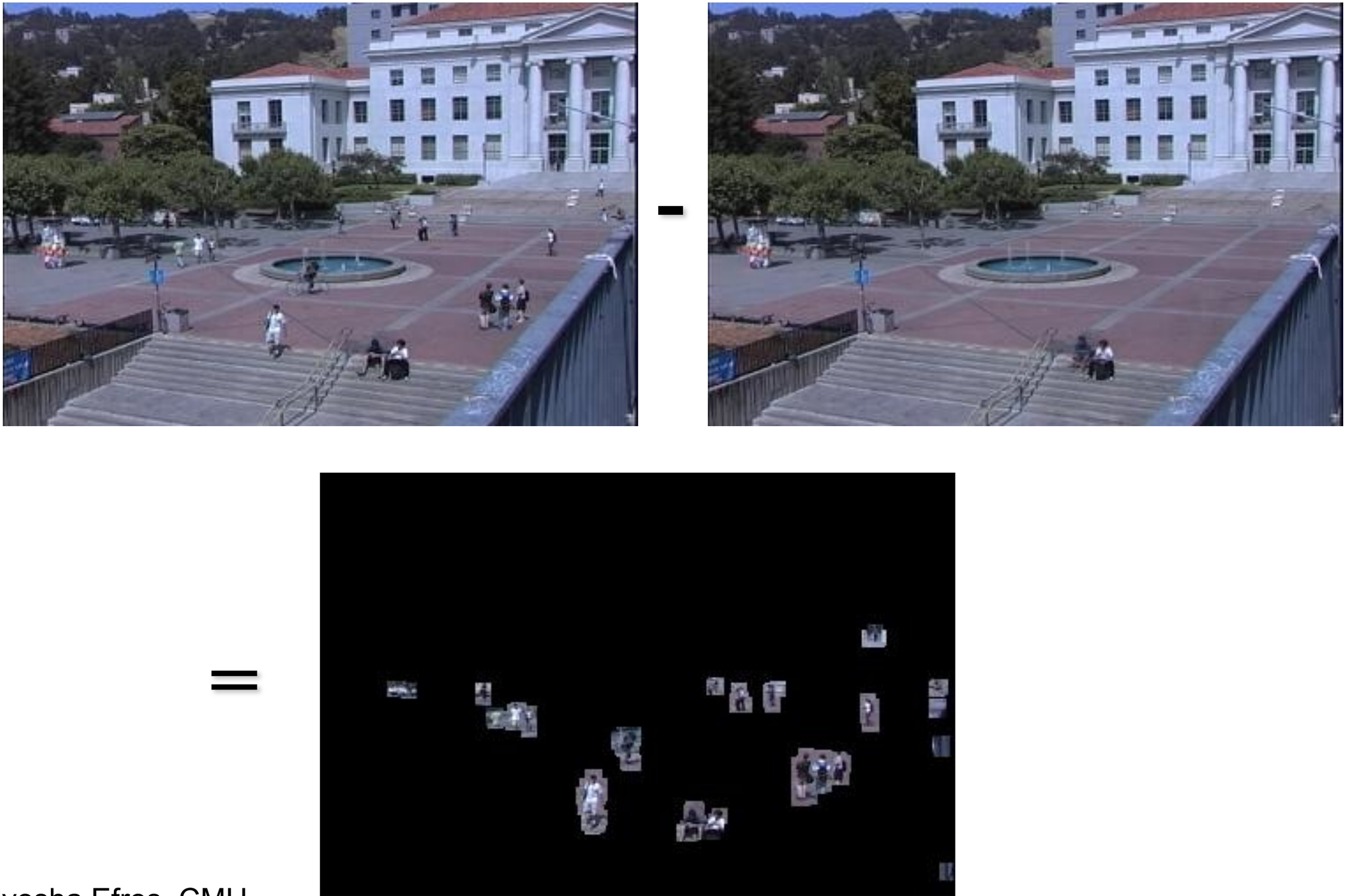
Estimated Background

Foreground Mask

Slide credit: Birgi Tamersoy

Average/Median Image

Background Subtraction



Alyosha Efros, CMU

Pros and cons

Advantages:

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

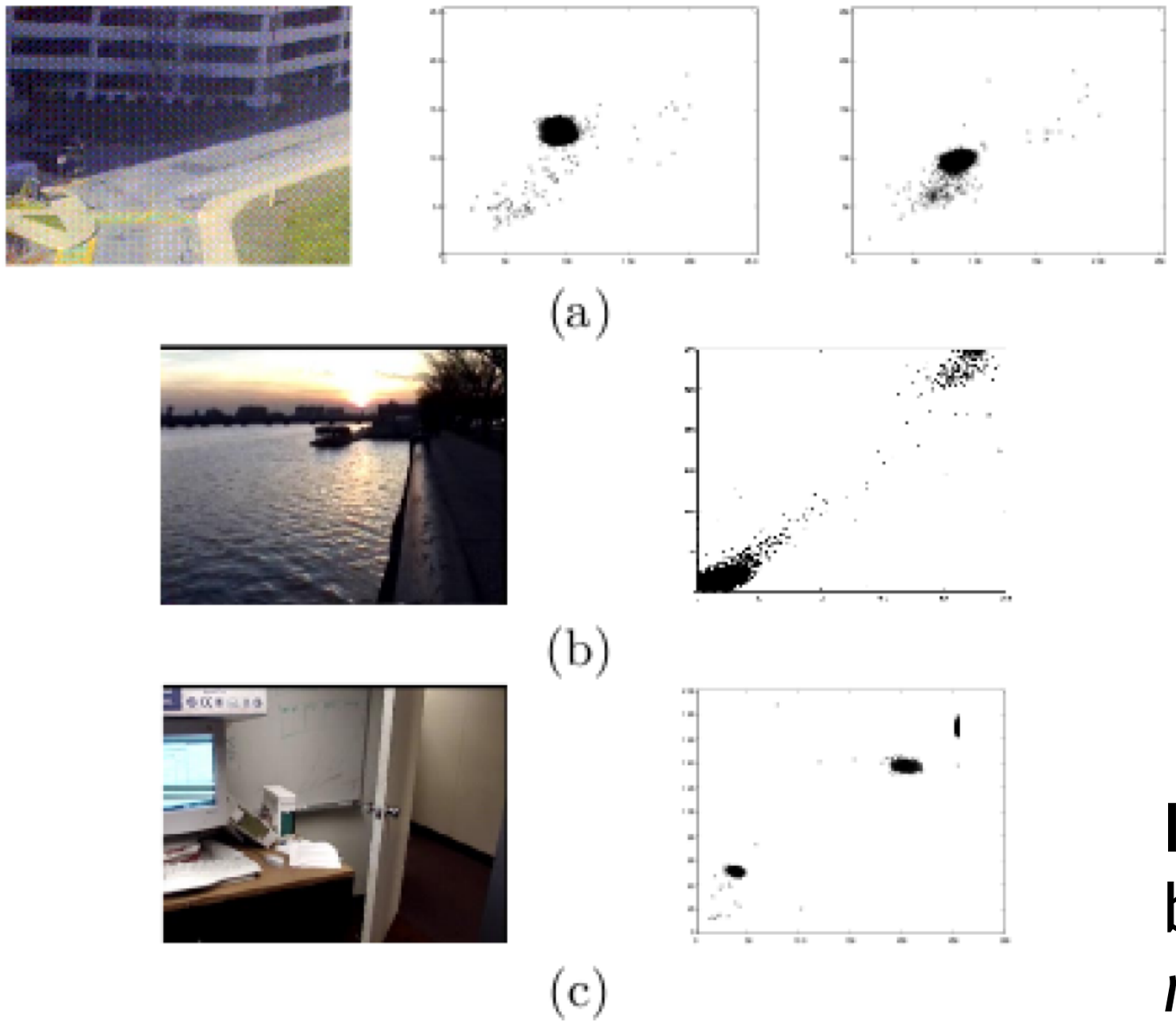
Disadvantages:

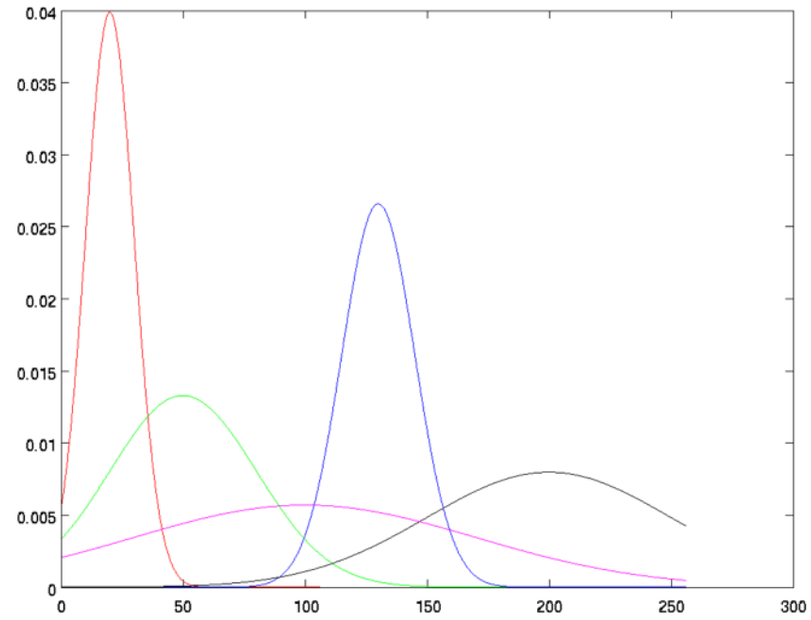
- Accuracy of frame differencing depends on object speed and frame rate
- Median background model: relatively high memory requirements.
- Setting global threshold Th…

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?

Human activity in video

No universal terminology, but approximately:

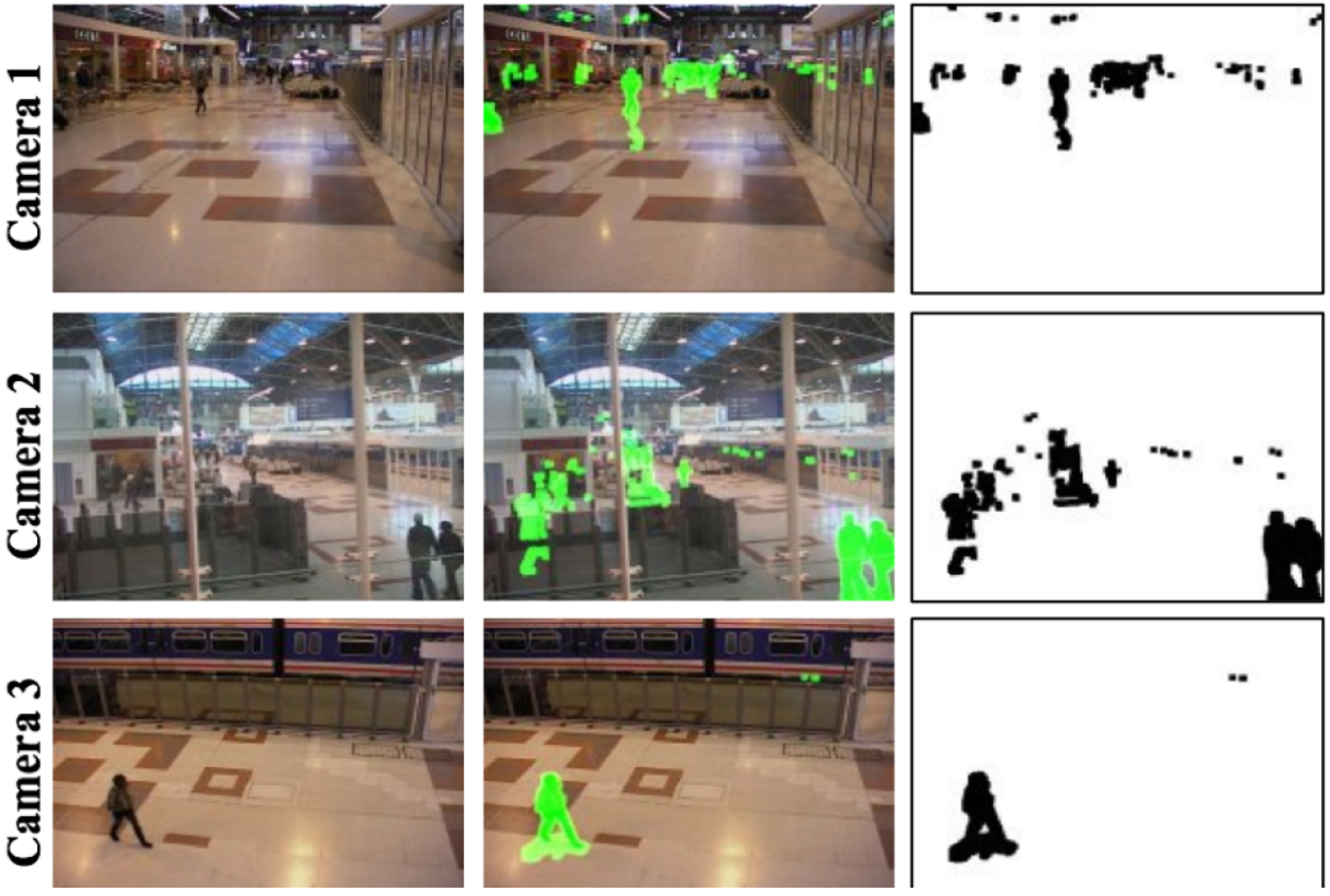
- "**Actions**": atomic motion patterns -- often gesture-like, single clear-cut trajectory, single nameable behavior (e.g., sit, wave arms)

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

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

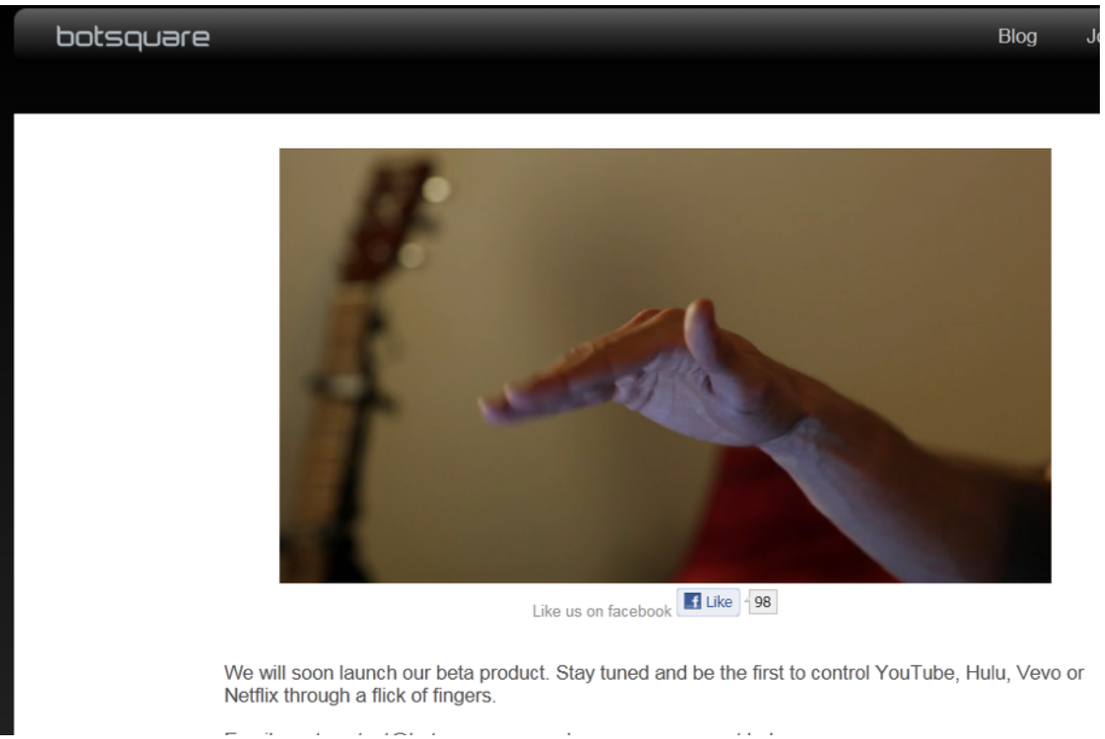
Adapted from Venu Govindaraju

Surveillance



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

Interfaces



2011

(a) template

(b) image

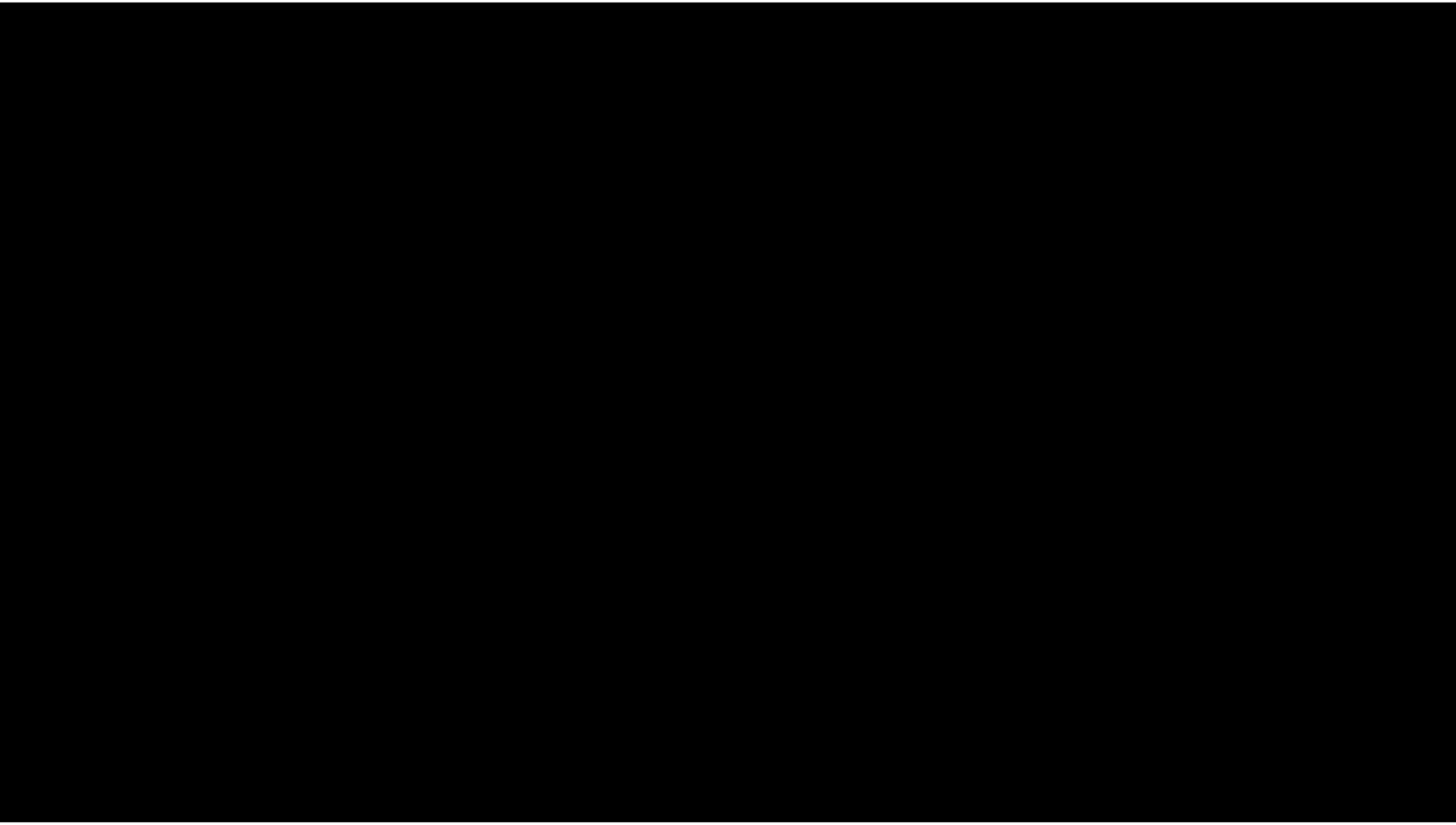
(c) normalized correlation

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

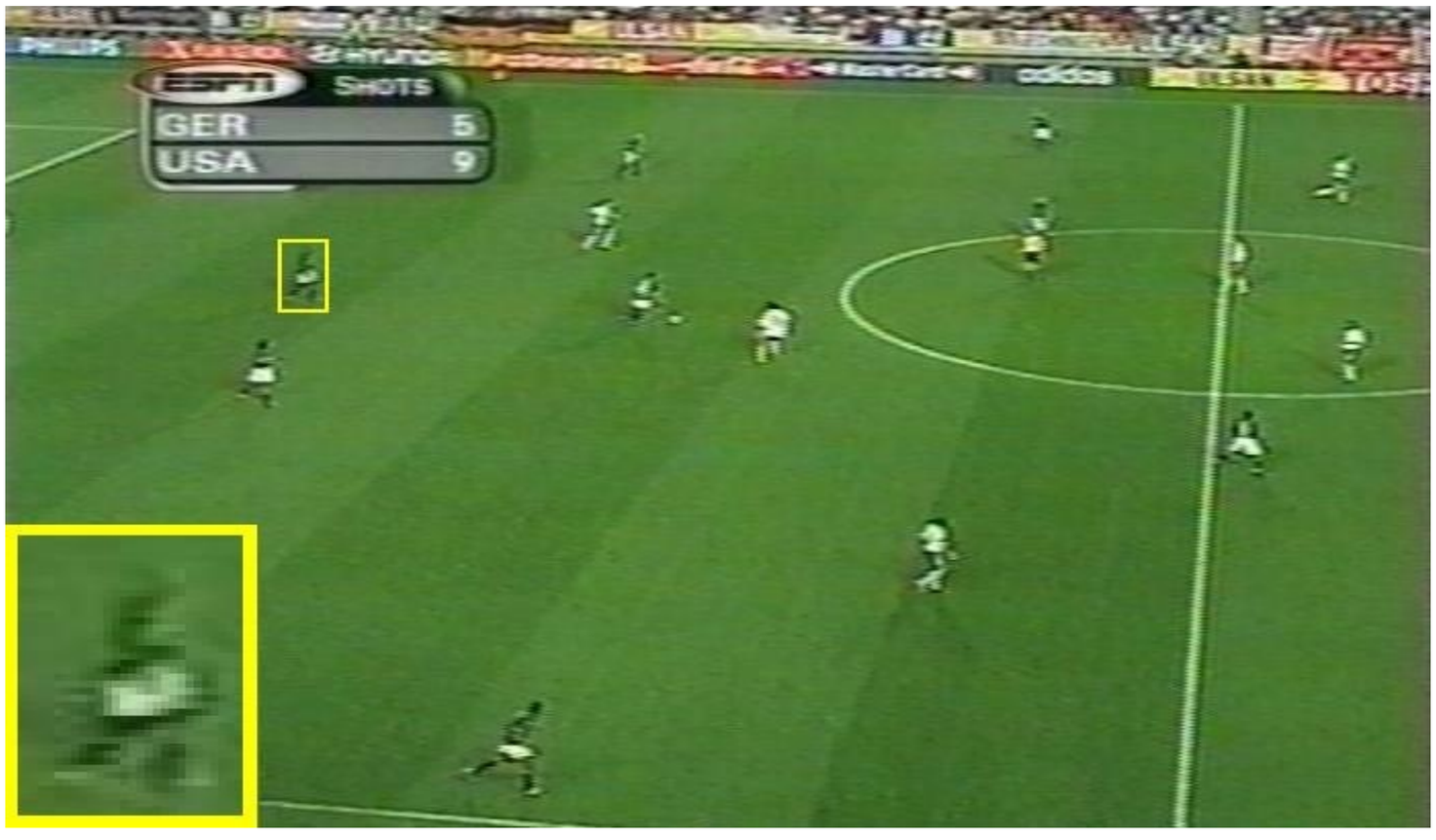
2008: Leap motion

2008: Leap motion



Using optical flow: action recognition at a distance

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

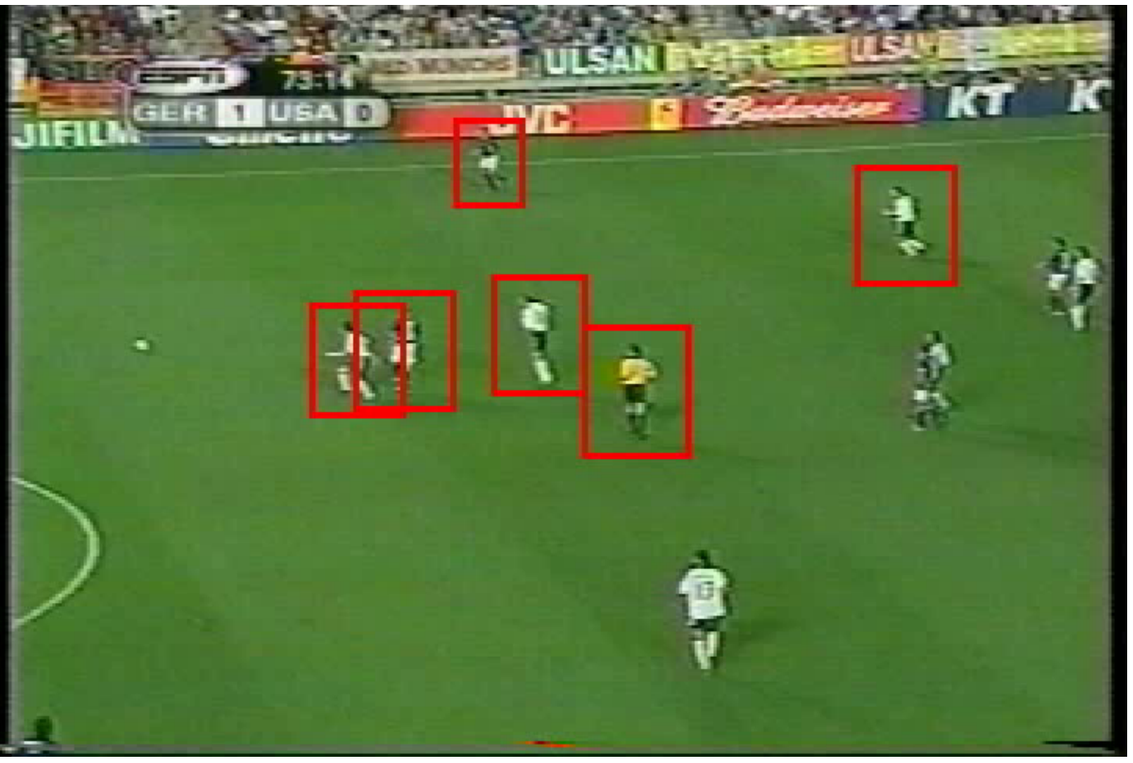


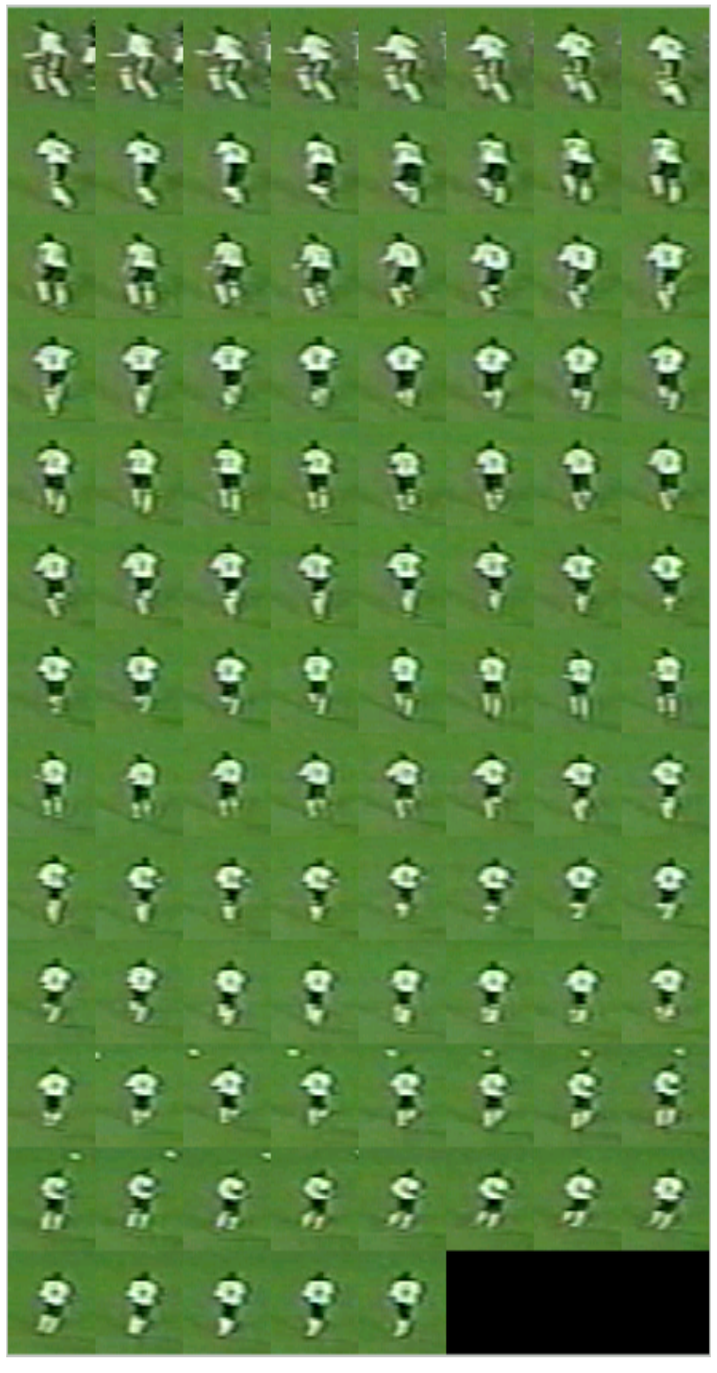
The 30-Pixel Man

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

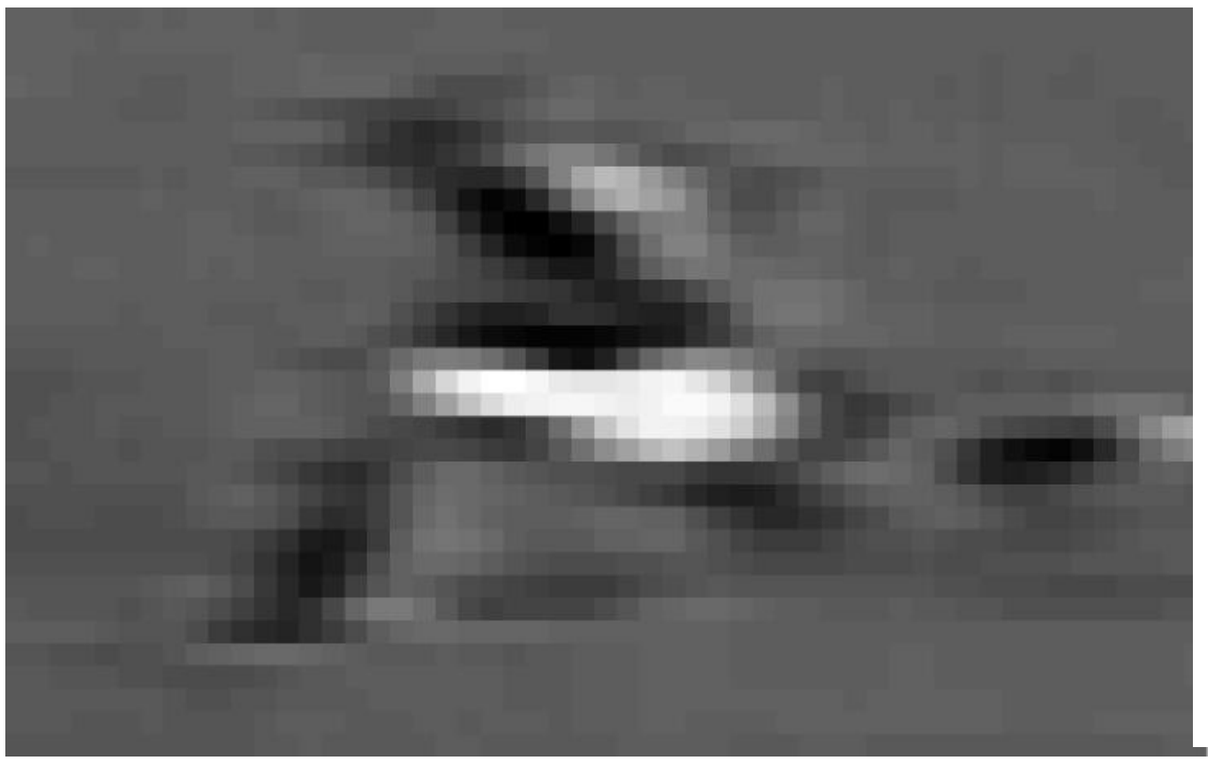




Correlation-based tracking
Extract person-centered frame window

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

Using optical flow: action recognition at a distance



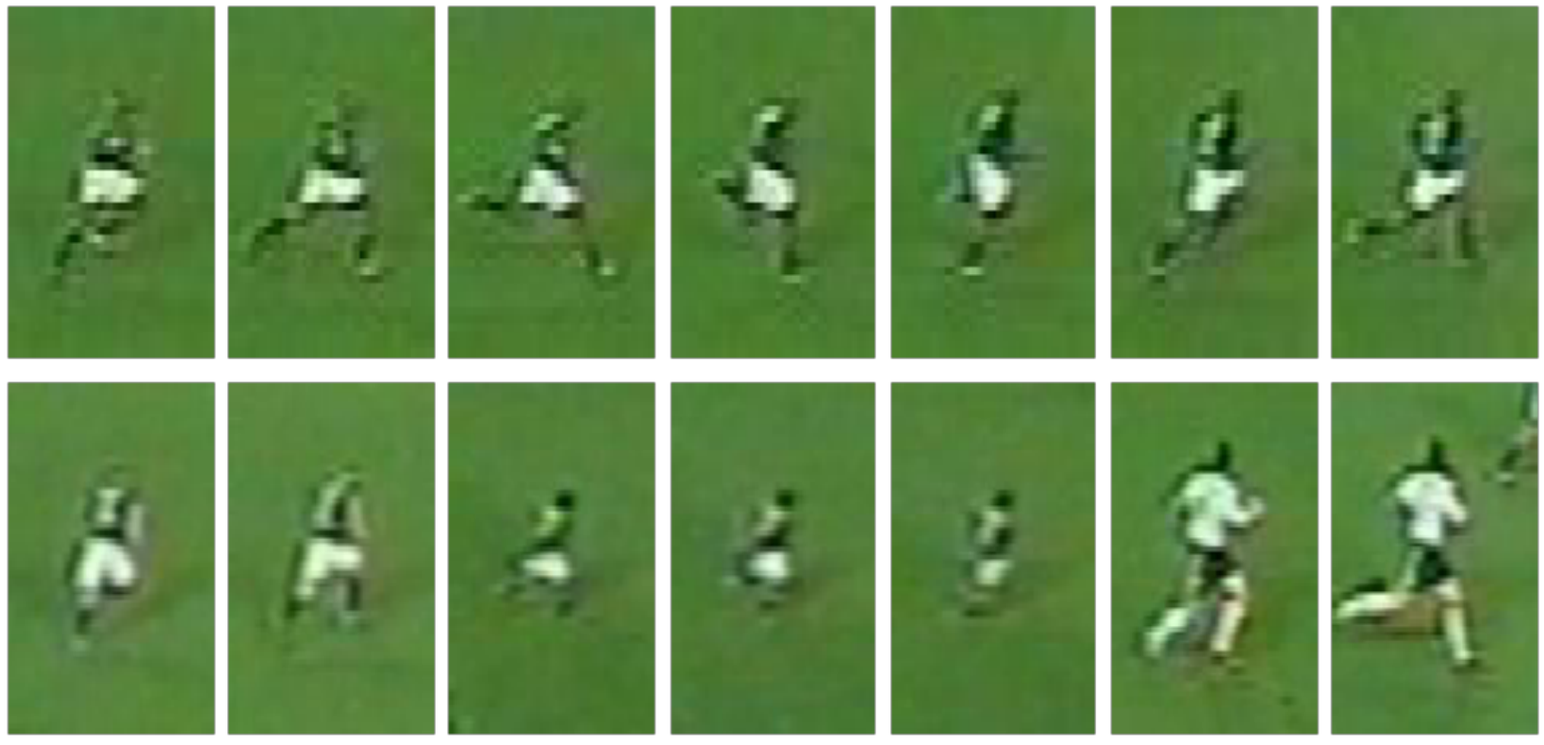
Extract optical flow to describe the region's motion.

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

Using optical flow: action recognition at a distance

Input Sequence

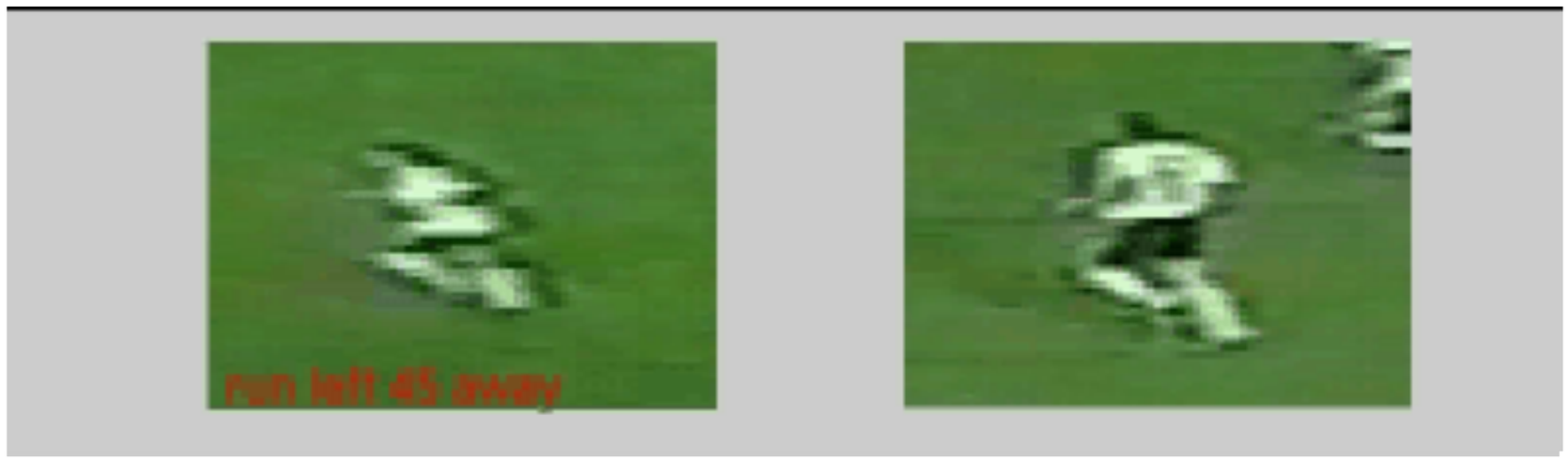
Matched Frames



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

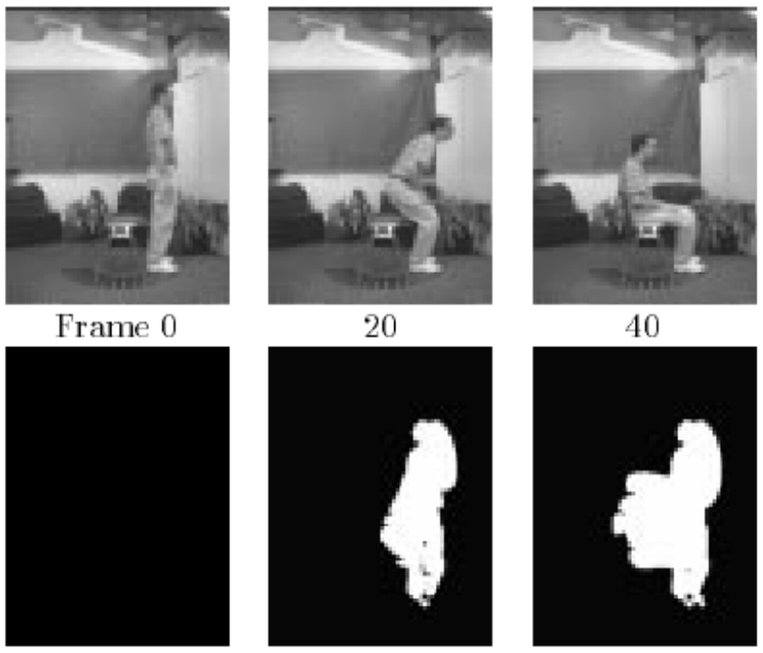


Figure 2: Example of someone sitting. Top row contains key frames; bottom row is cumulative motion images starting from Frame 0.

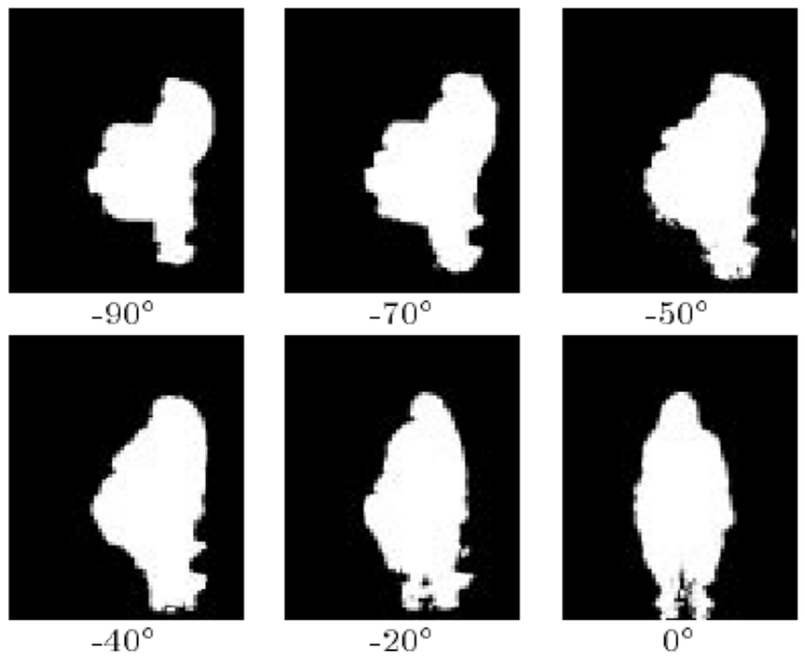
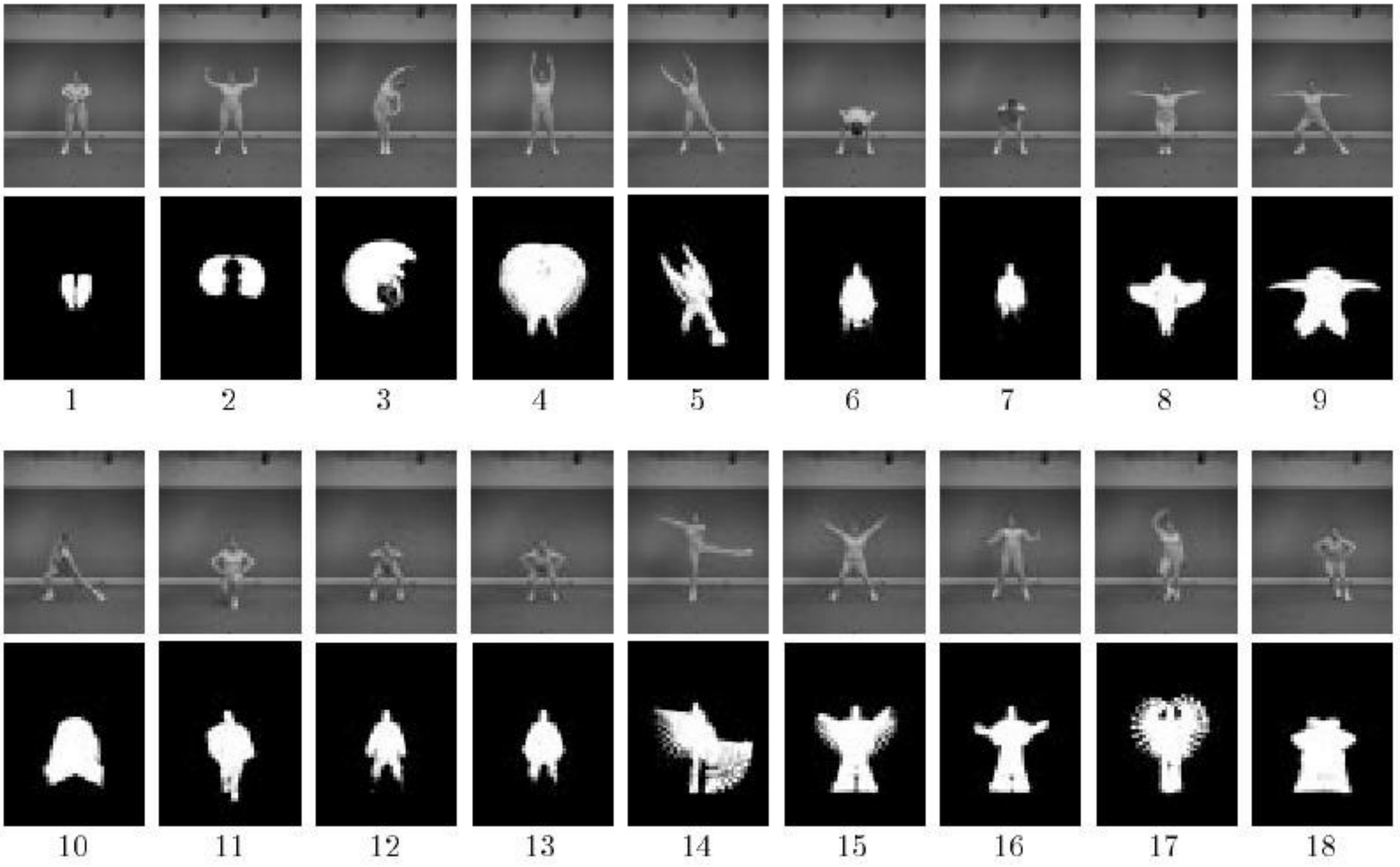


Figure 3: MEIs of sitting action over 90° viewing angle. The smooth change implies only a coarse sampling of viewing direction is necessary to recognize the action from all angles.

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

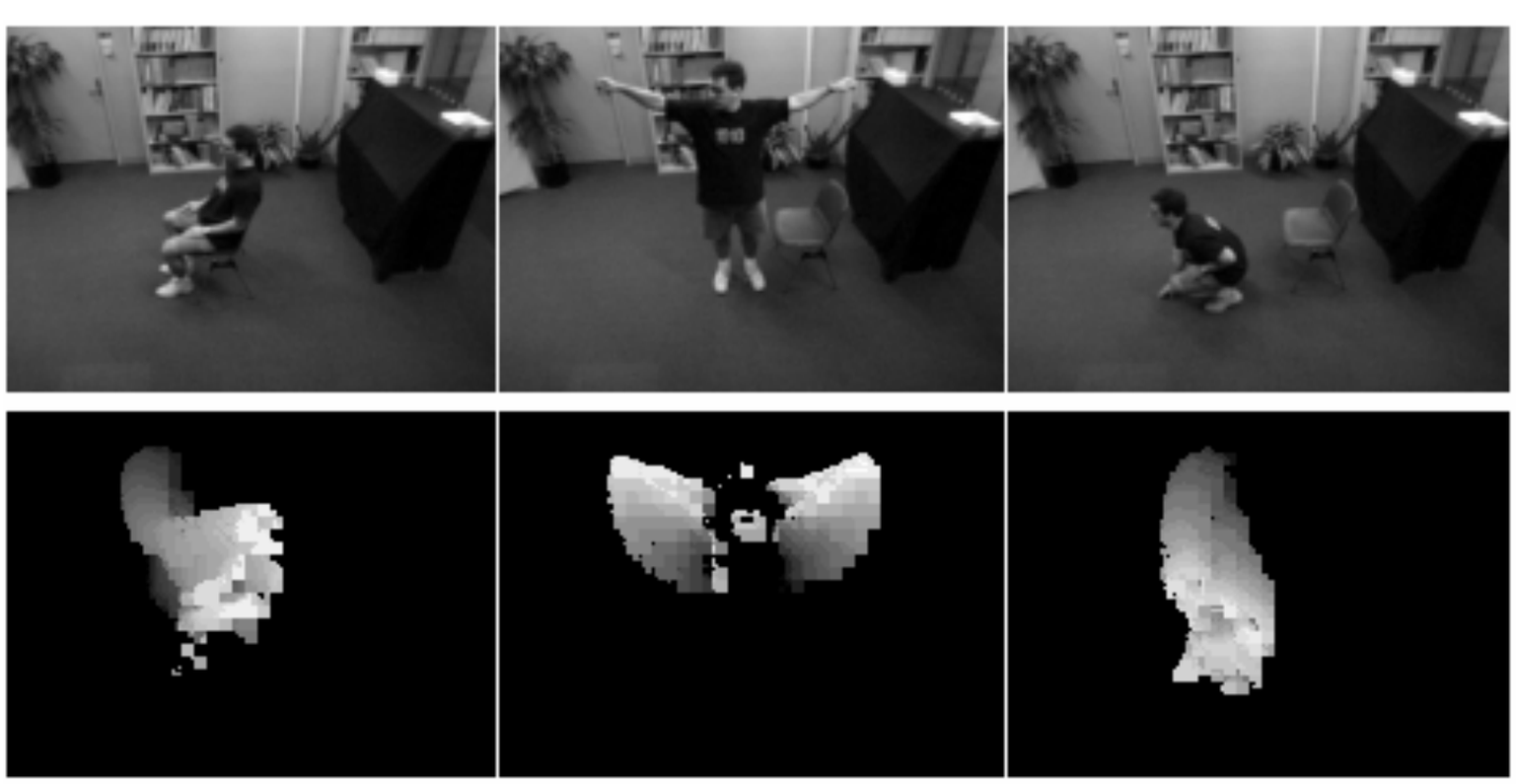
Motion Energy Images



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

Summary

- **Background subtraction**:
  - Essential low-level processing tool to segment moving objects from static camera's video
- **Action recognition:**
  - Increasing attention to actions as motion and appearance patterns
  - For instrumented/constrained environments, relatively simple techniques allow effective gesture or action recognition

Code review/refactor demo!

Project 2: Stereo

- **Code reviews due this Friday, 17 Oct**
- **Project 3 out in one week: 21 Oct**
- **Start thinking about final project groups and topics!**