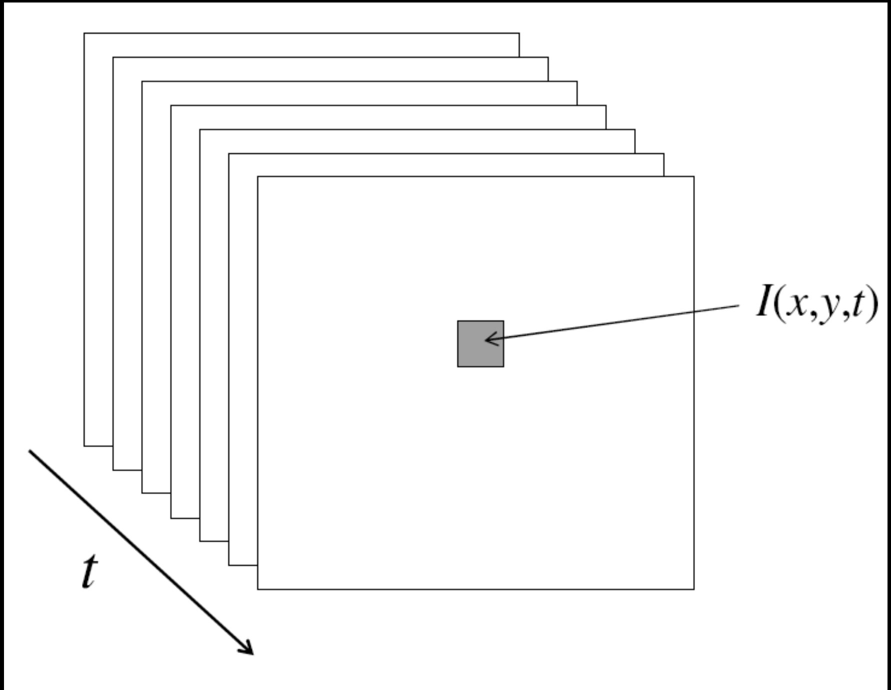
CS4495/6495
Introduction to Computer Vision

8D-L1 *Intro to video analysis*

Video

A video is a sequence of frames captured over time

- Now our image data is a function of space (x, y) and time (t)

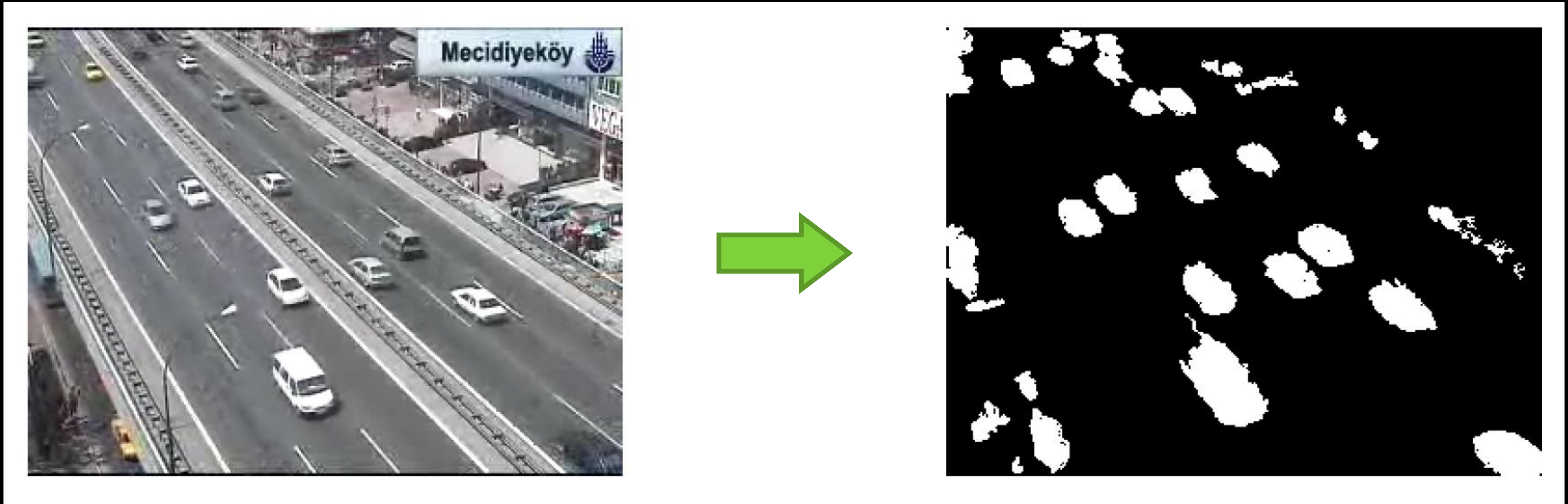


Processing video: Object detection

If the goal of “activity recognition” is to recognize the activity of objects…

… you (may) have to *find* the objects….

Background subtraction



Slides: Birgi Tamersoy

Background subtraction

- Needs static camera – still hard!
- Widely used:
  - Traffic monitoring (counting vehicles, detecting & tracking vehicles, pedestrians)
  - Human action recognition (run, walk, jump, squat)
  - Human-computer interaction
  - Object tracking

Simple approach: Background subtraction

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

Image at time t:
$I(x,y,t)$

Background at time t:
$B(x,y,t)$

$|$ 

$-$ 

$| > Th$

But, how can we estimate the background?

Frame differencing

Background is estimated to be the previous frame:

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

Background subtraction then becomes:

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

Image at time t:
$I(x, y, t)$

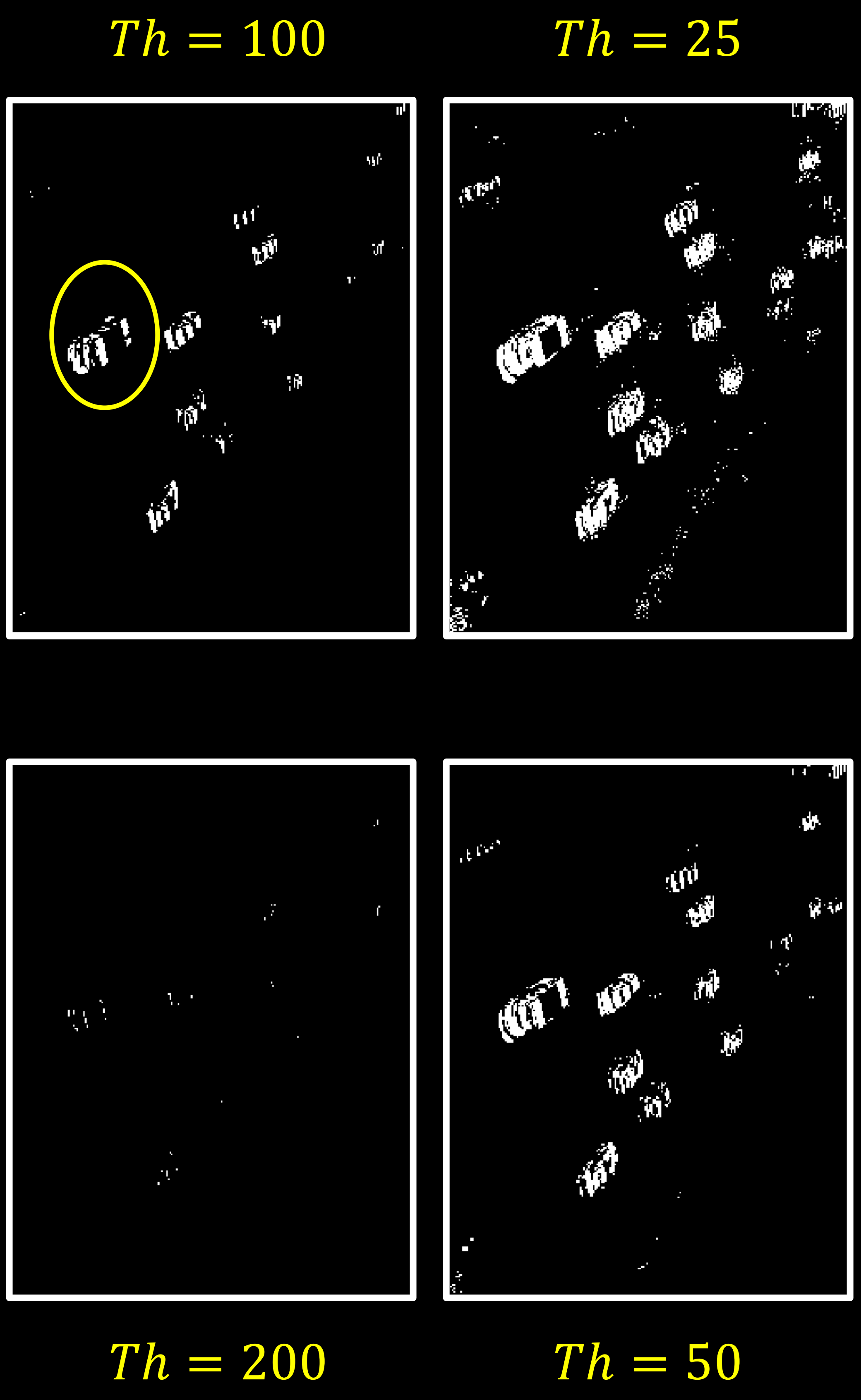
Background at time t:
$B(x, y, t) = I(x, y, t - 1)$

|

−

$| > Th$

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


$Th = 100$
$Th = 25$
$Th = 200$
$Th = 50$

Mean filtering

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

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

Therefore, foreground mask is computed as:

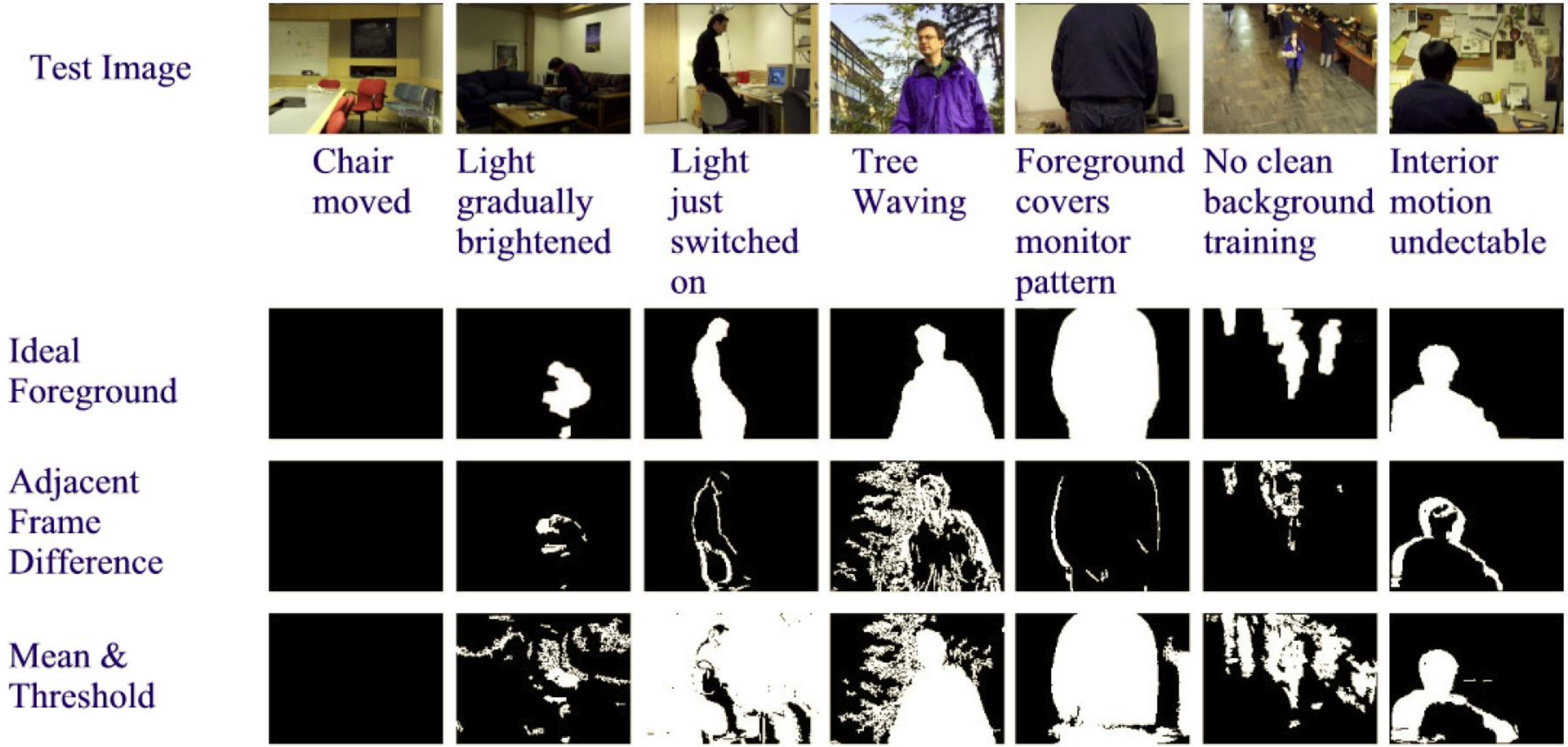
$$\left| I(x,y,t) - \frac{1}{n}\sum_{i=1}^{n} I(x,y,t-i) \right| > Th$$

Mean filtering

Estimated background

Foreground mask

Time window: $n = 10$



Frame difference vs. mean filtering [Toyama et al. 1999]

Median Filtering

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

$$B(x, y, t) = median\{I(x, y, t - i)\}$$

Median Filtering

Therefore, foreground mask is computed as:

$$|I(x,y,t) - median\{I(x,y,t-i)\}| > Th$$

where $i \in \{1, \dots, n\}$

Median Filtering

Estimated background

Foreground mask

Time window: $n = 10$

Median image computation

Input frames

Background model

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

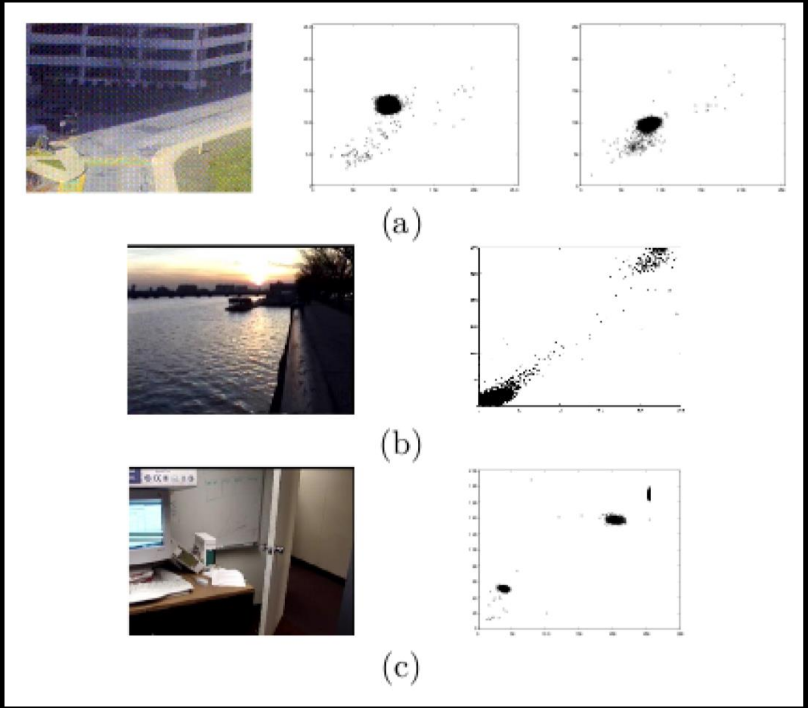
Pros and cons

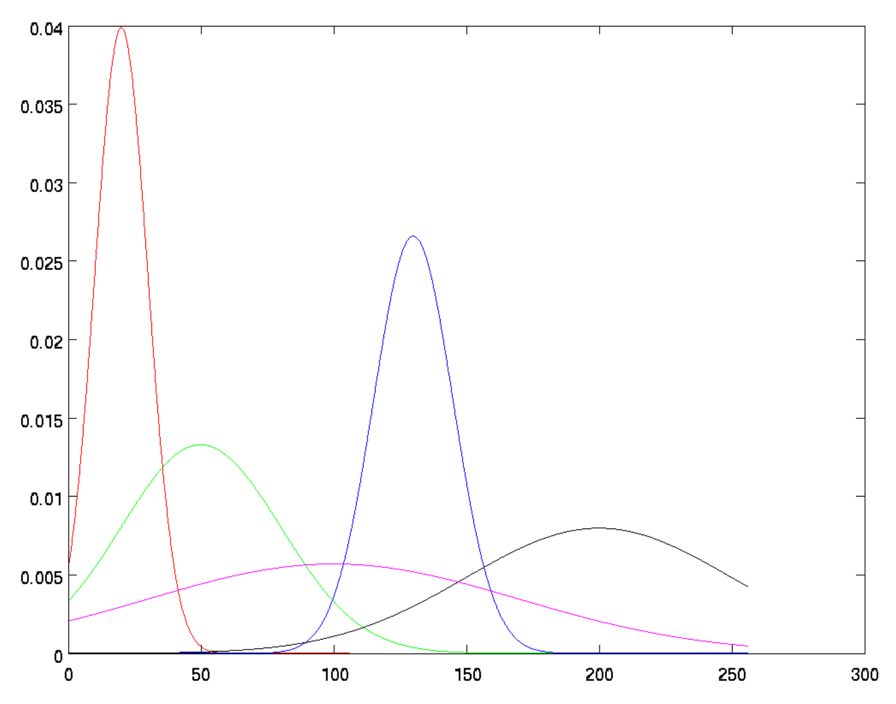
Disadvantages:

- 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?

Birgi Tamersoy





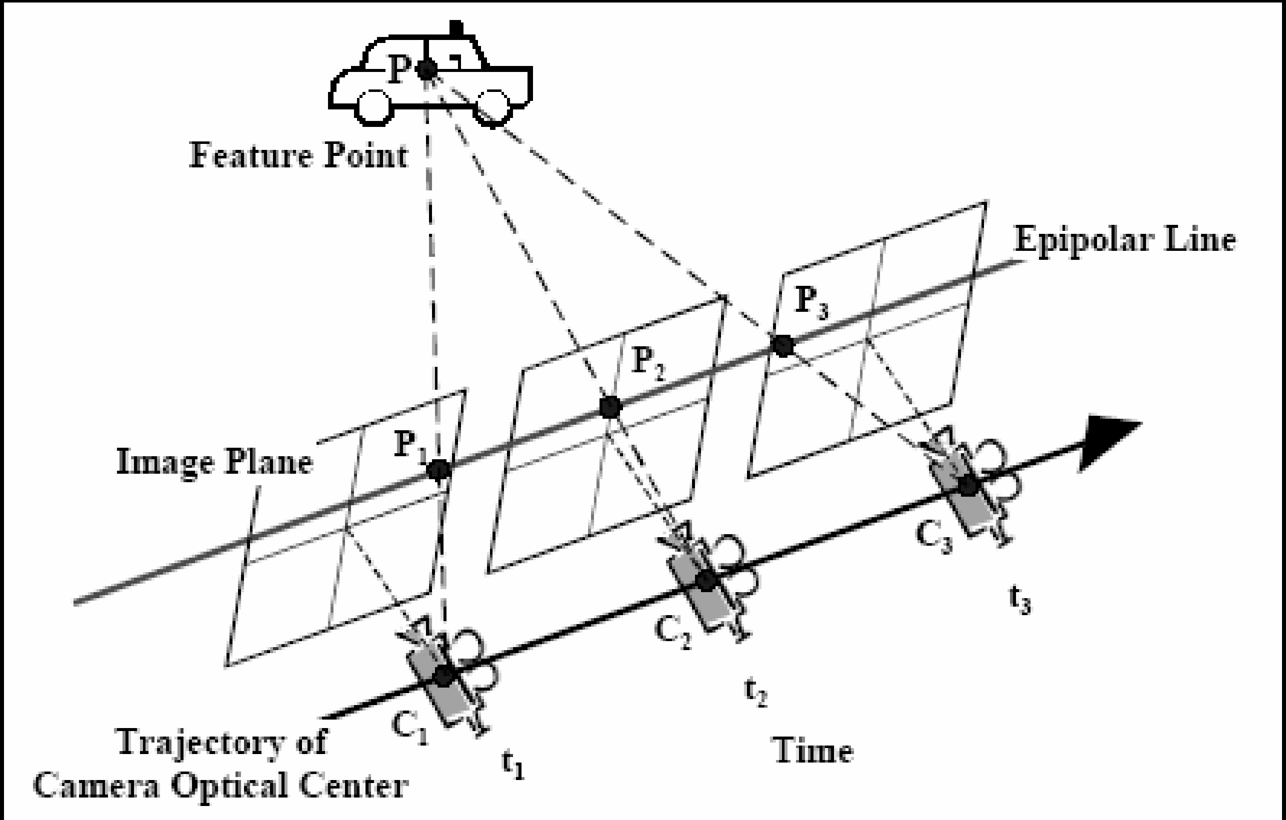
Adaptive Background Mixture Models for Real-Time Tracking, Chris Stauer & W.E.L. Grimson

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

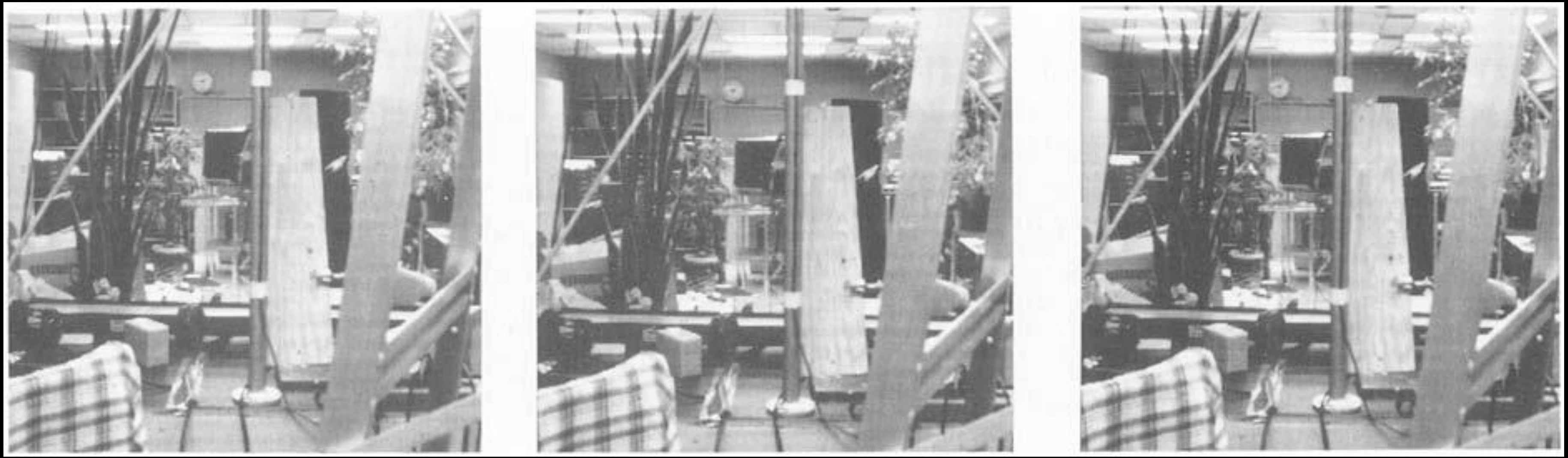
Background subtraction with *depth*

We can select foreground pixels based on depth information – RGBD

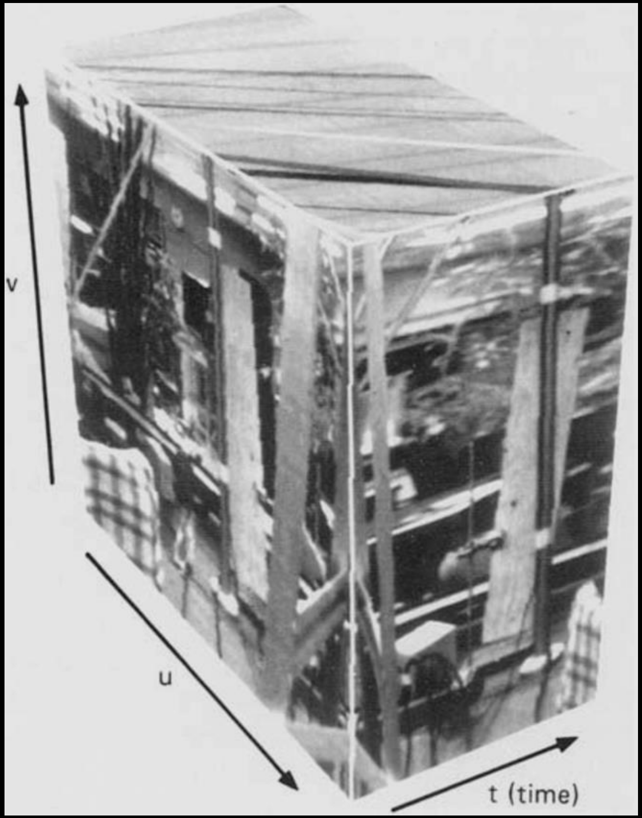
Aside: Epipolar Plane ("EPI") images

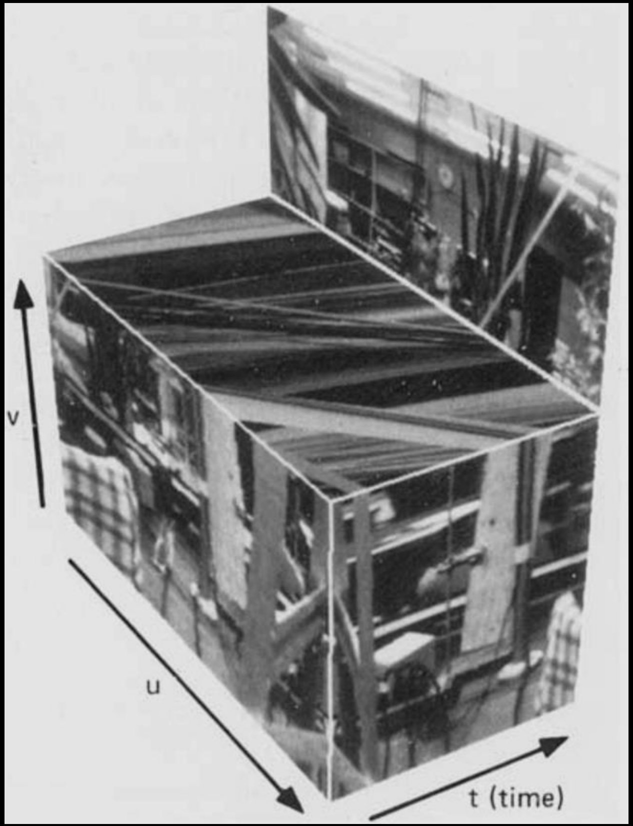


Individual images



Volume of data

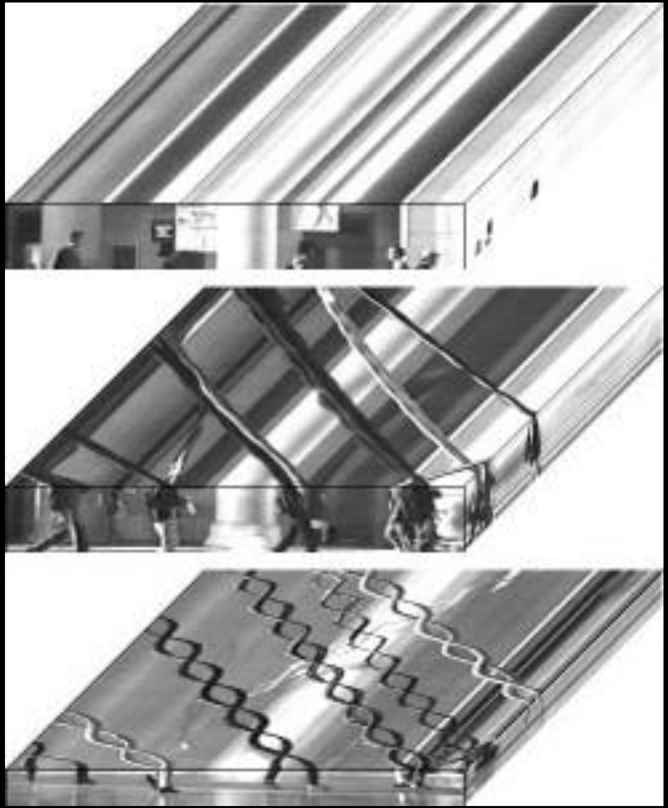





Y
T
X


T
X

EPI-gait



Niyogi and Adelson, 1994

EPI-gait

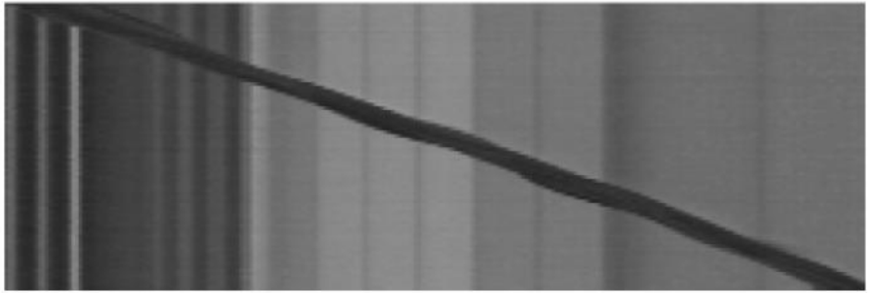


Figure 9 An XT-slice taken at the walker's head height, indicating the head mostly only undergoes translational movement during walking.

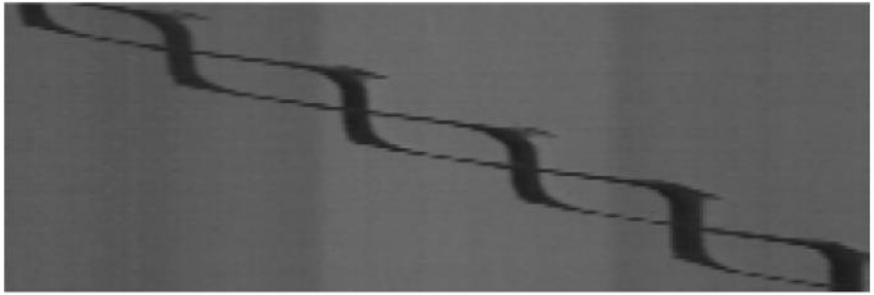


Figure 10 A slice taken at the height of the walker's ankles. The criss-crossing of the walker's legs as the walker moves from left to right is given as a unique braded signature for walking patterns