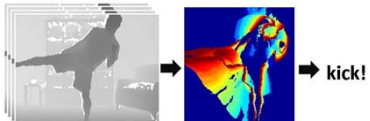


Actions in video

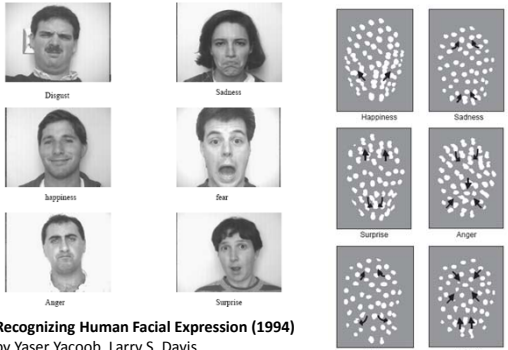
Monday, April 25
Kristen Grauman
UT-Austin



Today

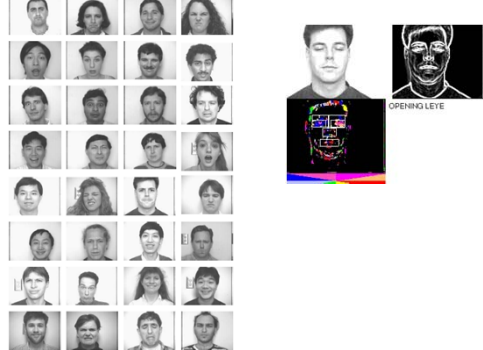
- Optical flow wrapup
- Activity in video
 - Background subtraction
 - Recognition of actions based on motion patterns
 - Example applications

Using optical flow: recognizing facial expressions

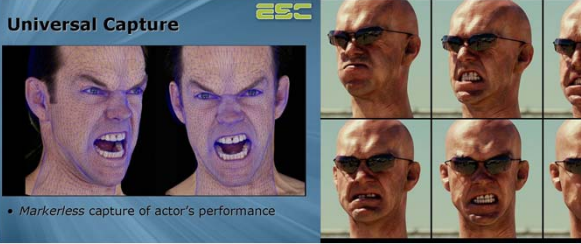


Recognizing Human Facial Expression (1994)
by Yaser Yacoob, Larry S. Davis

Using optical flow: recognizing facial expressions




Example use of optical flow: facial animation



<http://www.fxguide.com/article333.html>

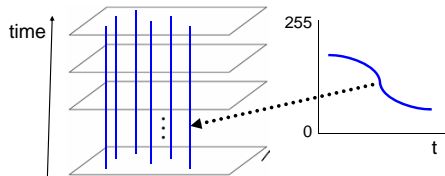
Example use of optical flow: Motion Paint

Use optical flow to track brush strokes, in order to animate them to follow underlying scene motion.



<http://www.fxguide.com/article333.html>

Video as an "Image Stack"

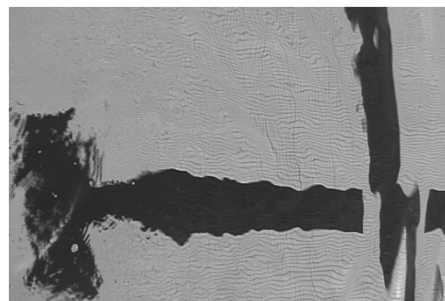


Can look at video data as a spatio-temporal volume

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

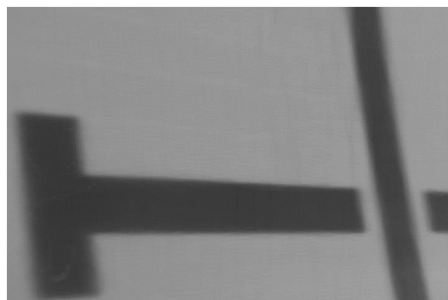
Alvisha Efros, CMU

Input Video



Alvisha Efros, CMU

Average Image



Alvisha Efros, CMU

Background Subtraction

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



Motivation

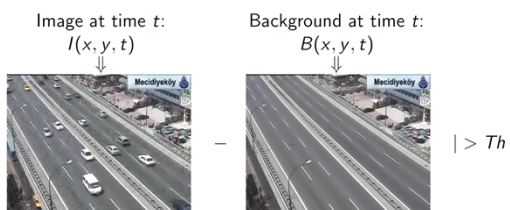
- ▶ In most cases, objects are of interest, not the scene.
- ▶ Makes our life easier: less processing costs, and less room for error.

Slide credit: Birgi Tamersov

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 frame.
3. Apply a threshold, Th , to the absolute difference to get the **foreground mask**.

Slide credit: Birgi Tamersov

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 Tamersov

Frame Differencing

Slide credit: Birgi Tamersov

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 Tamersov

Frame differences vs. background subtraction

Test Image							
Ideal Foreground							
Adjacent Frame Difference							
Mean & Threshold							

- 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:

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

$$\downarrow$$

$$|I(x, y, t) - \text{median}\{I(x, y, t - i)\}| > Th \text{ where } i \in \{0, \dots, n - 1\}.$$
- For $n = 10$:

Estimated Background

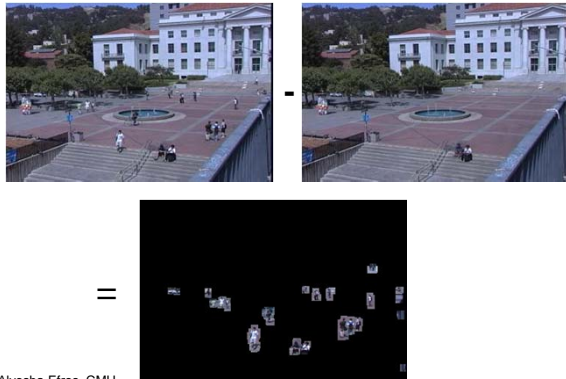
Foreground Mask

Slide credit: Birgi Tamersov

Average/Median Image

Alvisha Efros, CMU

Background Subtraction



Alvisha Efron, 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.

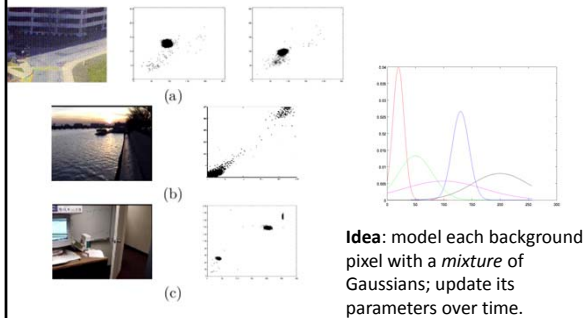
Disadvantages:

- 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 Tamersov

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?

Today

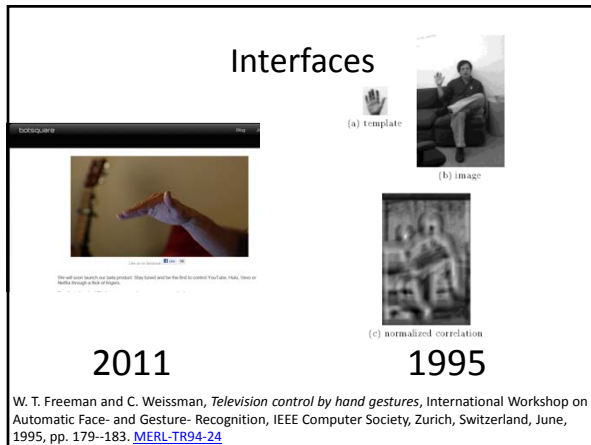
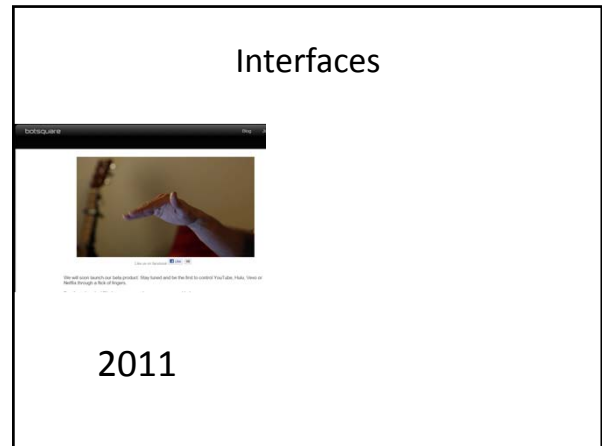
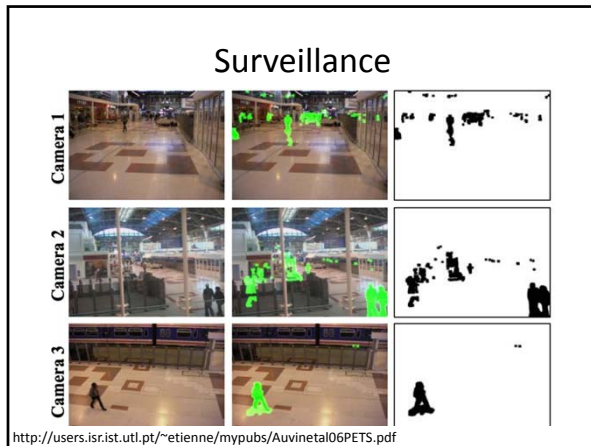
- Optical flow wrapup
- Activity in video
 - Background subtraction
 - Recognition of action based on motion patterns
 - Example applications

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



- ### 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

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



Video from Davis & Bobick

Using optical flow: action recognition at a distance

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

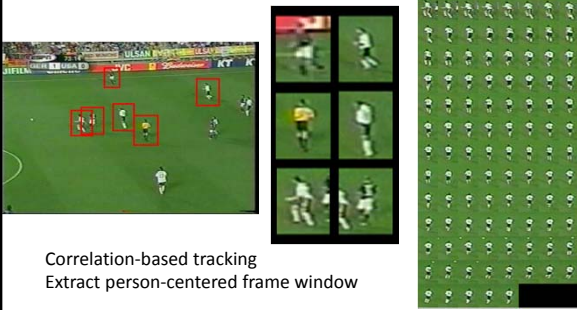


Challenge: low-res data, not going to be able to track each limb.

The 30-Pixel Man

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

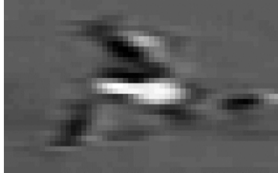
Using optical flow: action recognition at a distance



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/>

Motivation

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



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

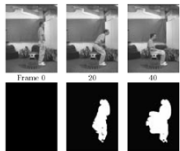


Figure 2: Example of someone sitting. Top row contains original frames, bottom row is cumulative motion images starting from Frame 0.

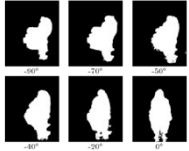
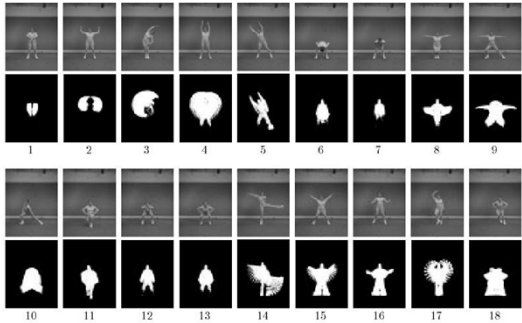


Figure 3: MEIs of sitting action over 90° viewing angle. The smooth change implies only a coarse sampling of viewing direction is necessary to recognize the action from all angles.

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

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



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:

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

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

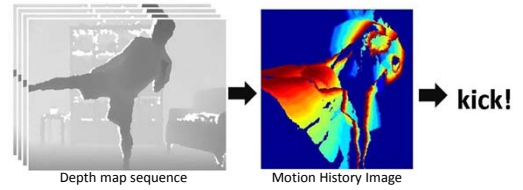
Hu moments

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



→ $[h_1, h_2, h_3, h_4, h_5, h_6, h_7]$

Pset 5



Nearest neighbor action classification with Motion History Images + Hu moments

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

Hu moments

$$\begin{aligned}
 h_1 &= \mu_{20} + \mu_{02}, \\
 h_2 &= (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2, \\
 h_3 &= (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2, \\
 h_4 &= (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2, \\
 h_5 &= (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\
 &\quad + (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) \\
 &\quad \cdot [3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2], \\
 h_6 &= (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\
 &\quad + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}),
 \end{aligned}$$

$$\begin{aligned}
 h_7 &= (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\
 &\quad - (\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]
 \end{aligned}$$