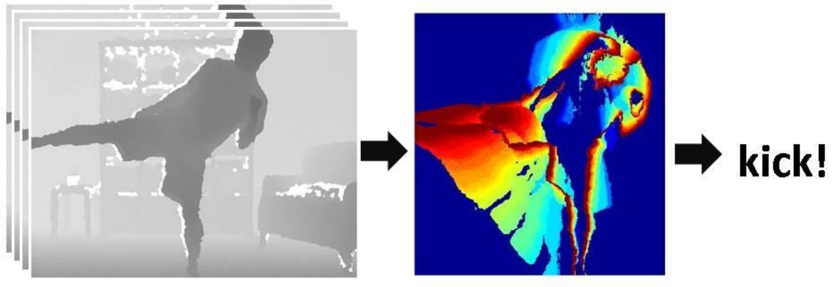
CS4495/6495
Introduction to Computer Vision

8D-L2 *Activity recognition*



Human activity in video

No universal terminology, but approximately:

- Event: A single instant in time detection
- Actions or Movements: Atomic motion patterns
  - Often gesture-like
  - Single clear-cut trajectory
  - Single nameable behavior (e.g., sit, wave arms)

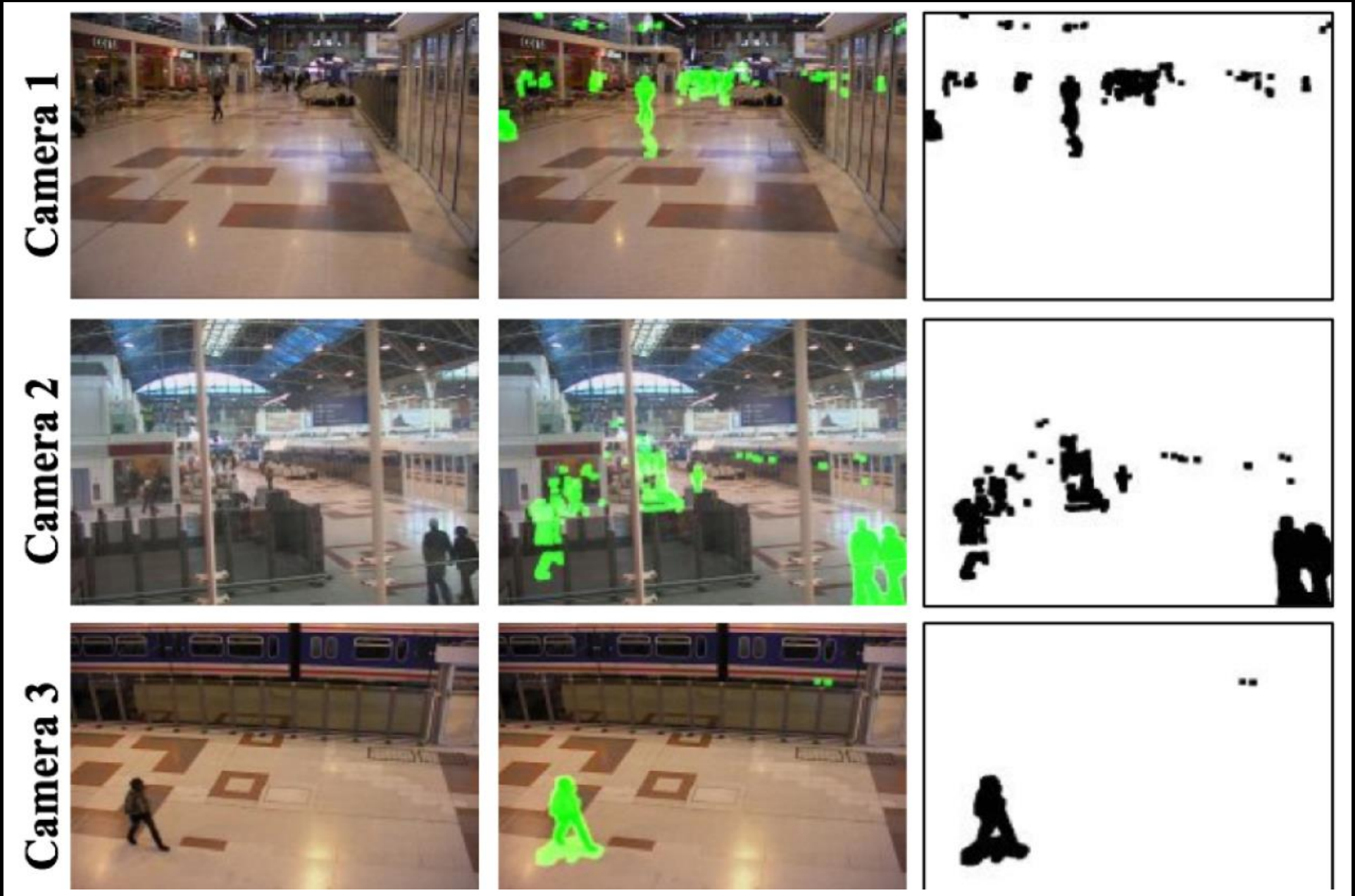
Adapted from Venu Govindaraju and A.Bobick

Human activity in video

- Activity: Series or composition of actions
  - E.g., interactions between people

Adapted from Venu Govindaraju and A.Bobick

Surveillance



Human activity in video: Basic approaches

- Model-based *action* recognition
  - Use human body tracking and pose estimation techniques, relate to action descriptions (or learn)
  - Major challenge: training data from different context than testing

Human activity in video: Basic approaches

- Model-based *activity* recognition
  - Given some lower level detection of actions (or events) recognize the activity by comparing to some structural representation of the activity
  - Needs to handle uncertainty
  - Major challenge: Accurate tracks in spite of occlusion, ambiguity, low resolution

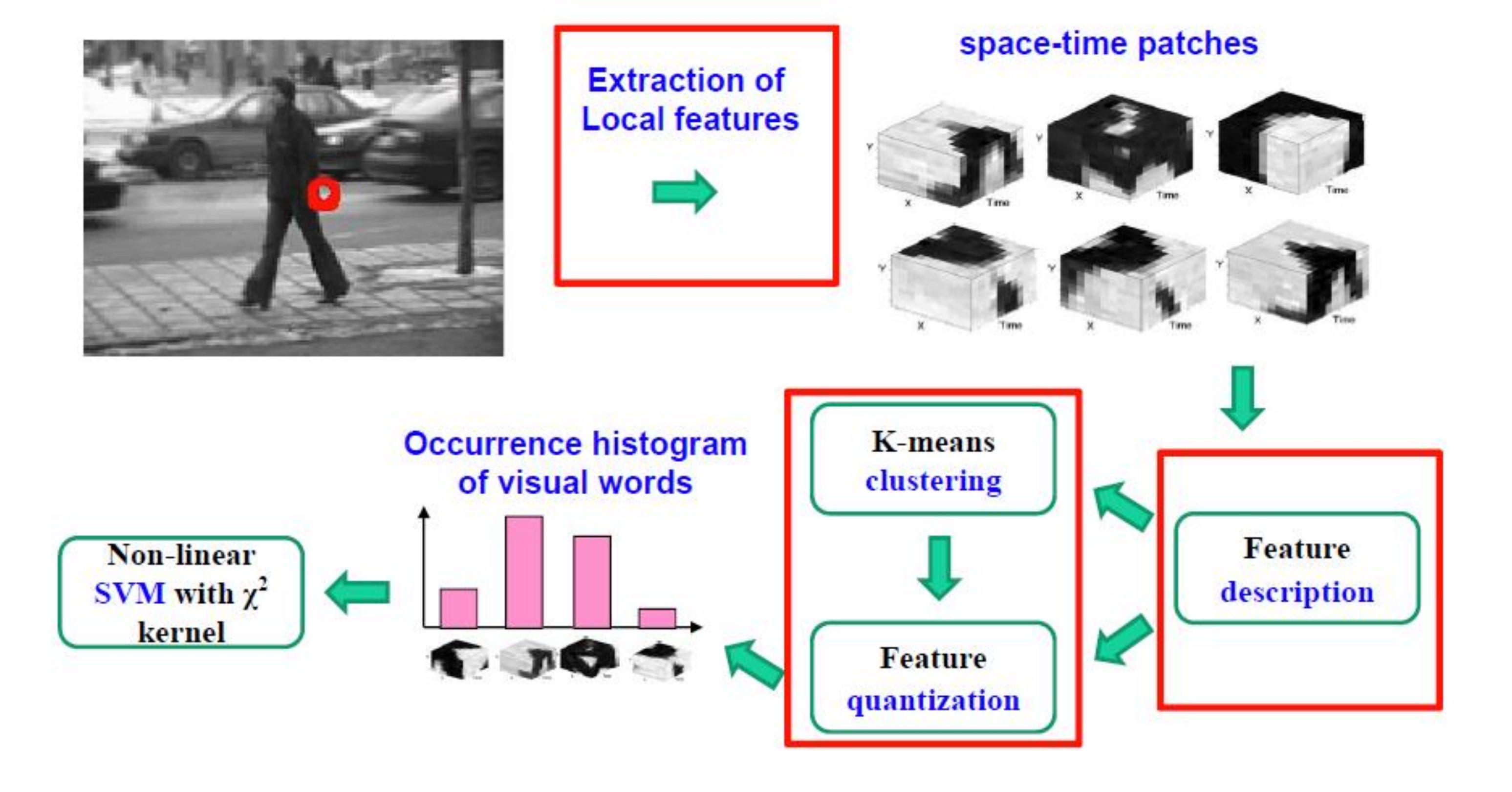
Human activity in video: Basic approaches

- Recently activity as motion, space-time appearance patterns
- Describe overall patterns, but no explicit body tracking
- Typically learn a classifier

Human activity in video: Basic approaches

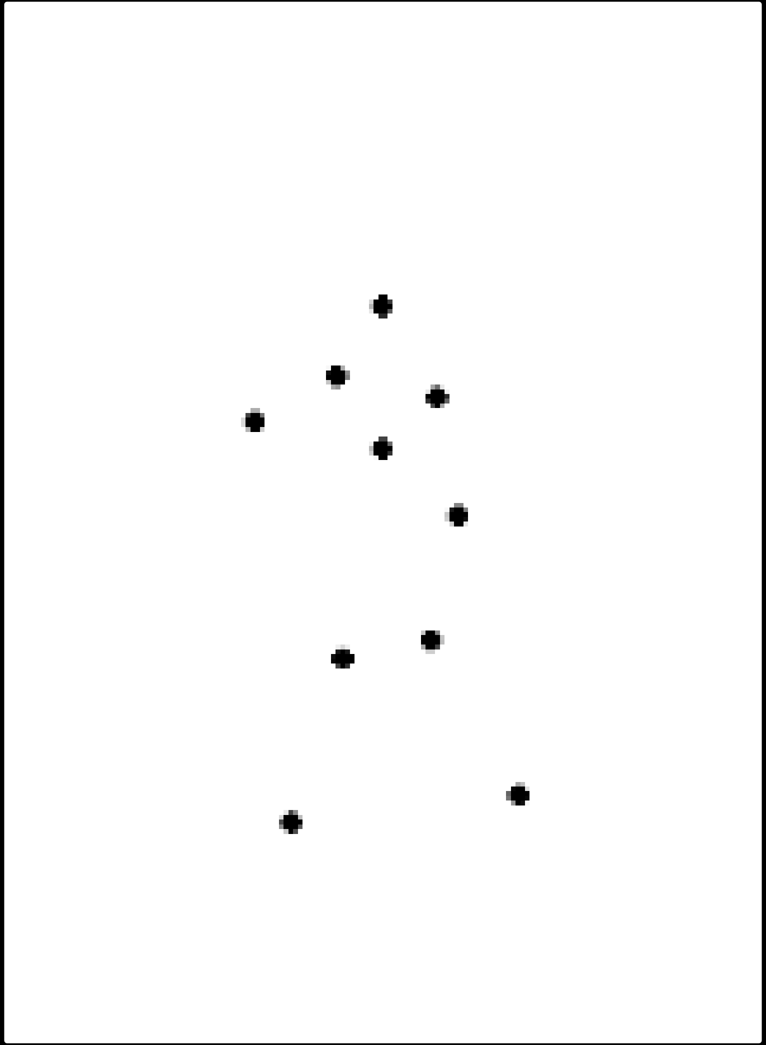
- Also recently: “Activity-recognition” from a static image
- Imagine a picture of a person holding a flute – what are they doing?

What we’re not going to cover?


Extraction of Local features
space-time patches
Y
X
Time
Feature description
K-means clustering
Feature quantization
Occurrence histogram of visual words
Non-linear SVM with χ^2 kernel

Motion and perceptual organization

Even “impoverished” motion data can evoke a strong percept



Motion and perceptual organization

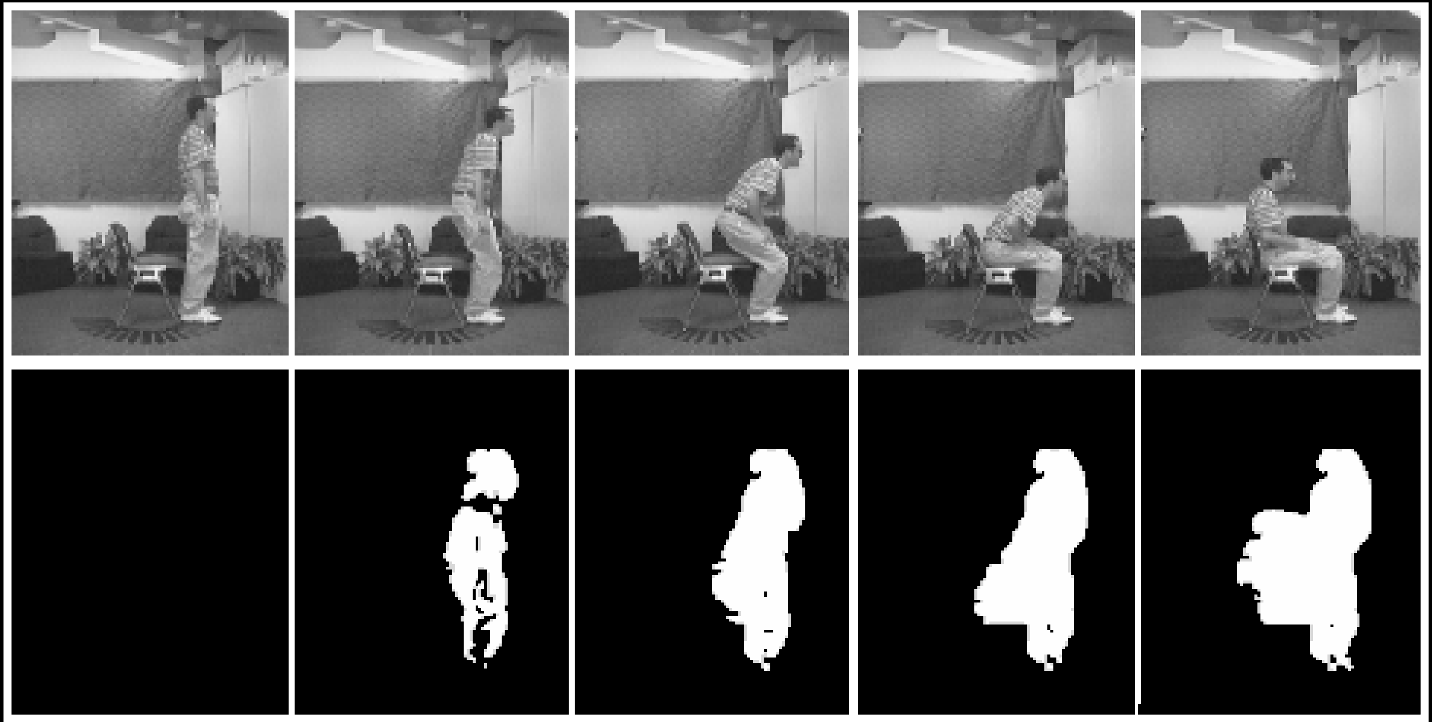
Even “impoverished” motion data can evoke a strong percept

Davis & Bobick, 1999

The Representation and Recognition of Action Using Temporal Templates

Motion Energy Images

time

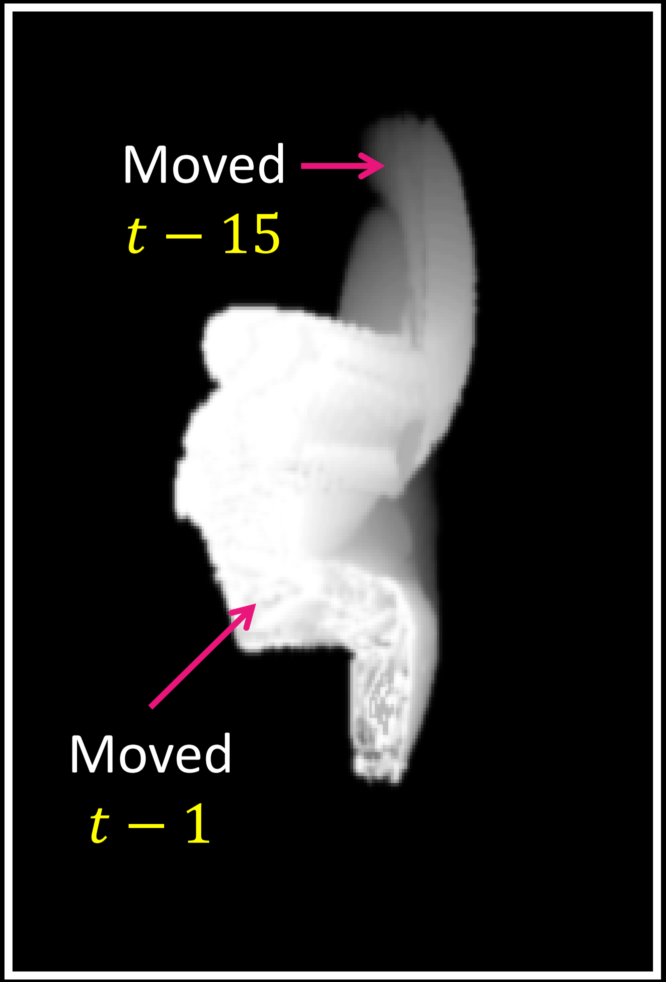


Motion History Images

MHIs are a different function of temporal volume

- Pixel operator is replacement decay:

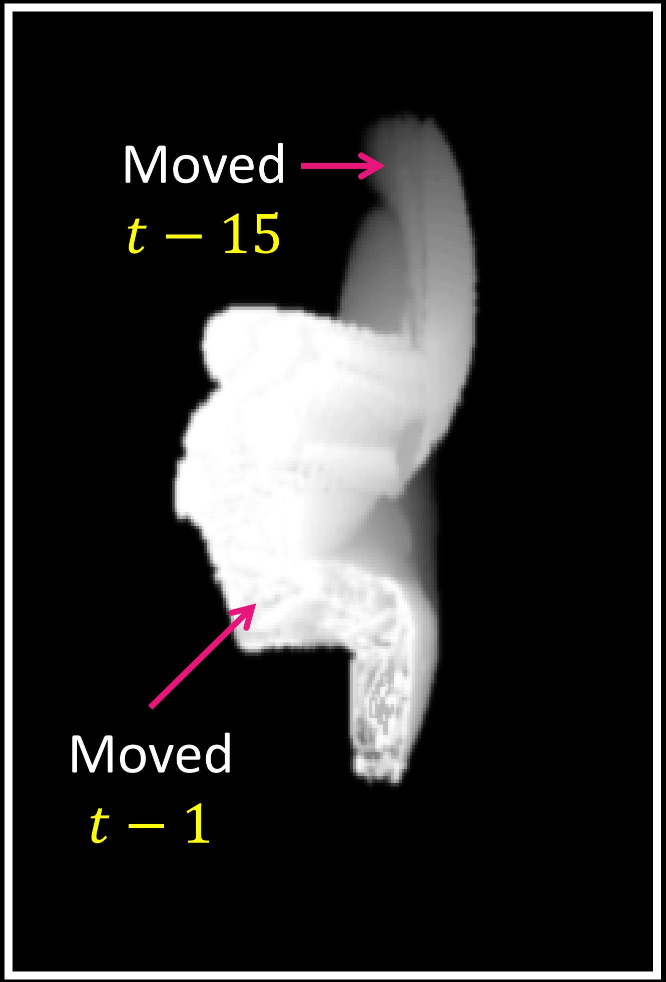
```
if moving:
  $I_\tau(x, y, t) = \tau$
otherwise:
  $I_\tau(x, y, t) = \max(I_\tau(x, y, t-1) - 1, 0)$
```



Motion History Images

- Trivial to construct $I_{\tau-k}(x, y, t)$ from $I_{\tau}(x, y, t)$ – so we can process multiple time window lengths without additional image analysis.

- MEI is thresholded MHI

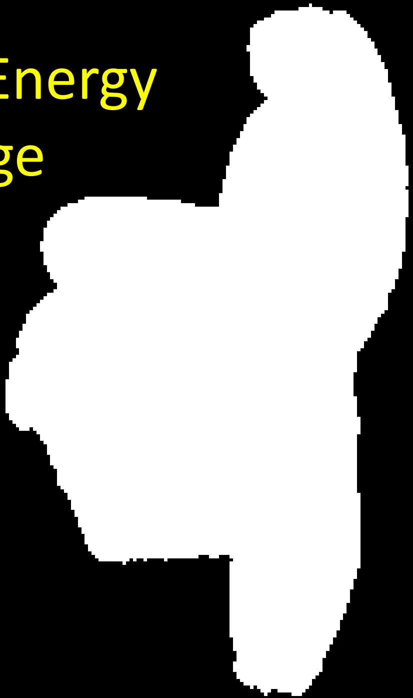


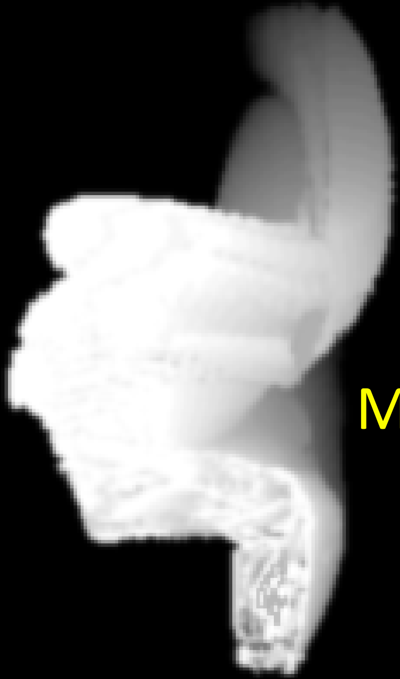
Moved $t - 15$

Moved $t - 1$

Temporal templates

Motion Energy Image

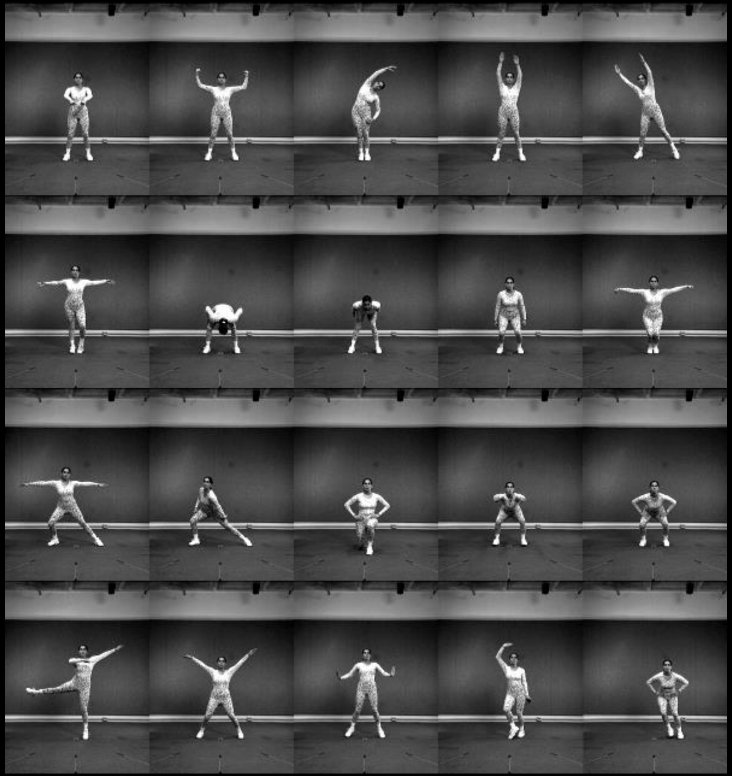


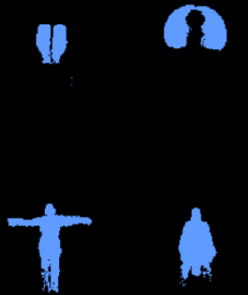


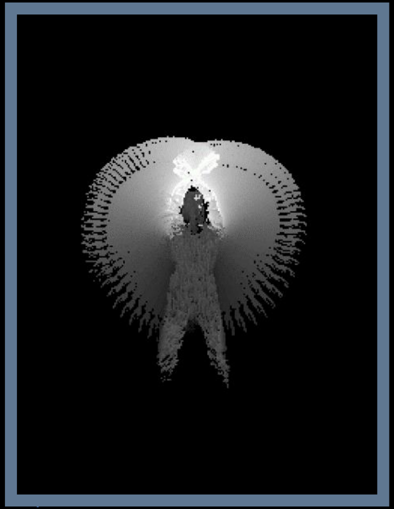
Motion History Image

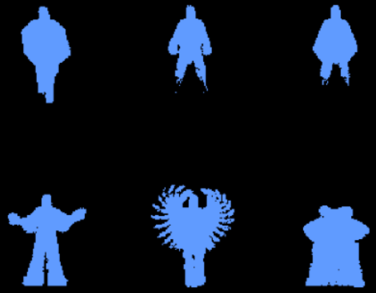
MEI + MHI = Temporal template

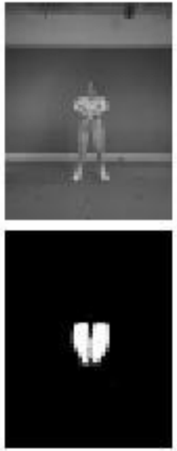
Aerobics

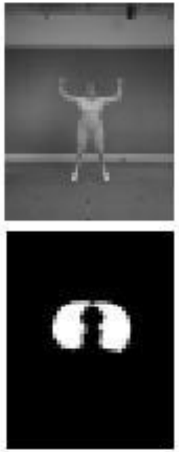


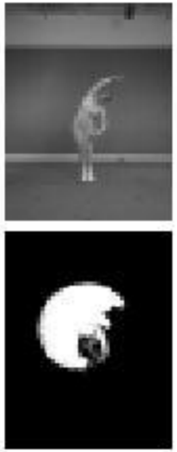


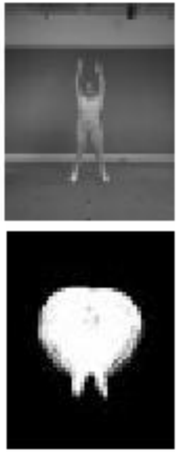


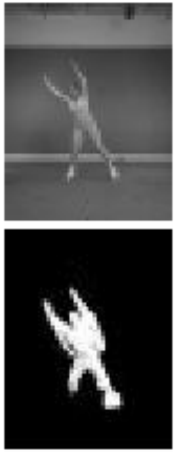


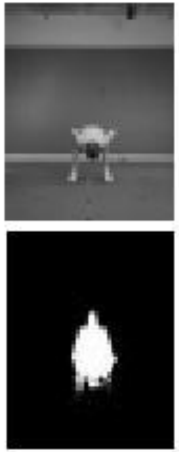

1

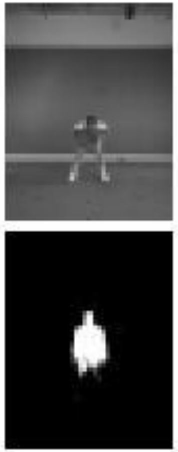

2

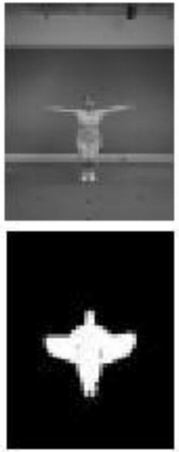

3

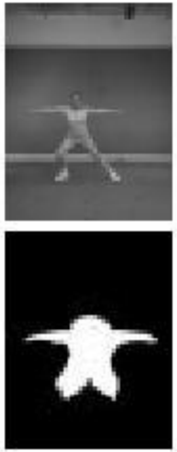

4


5

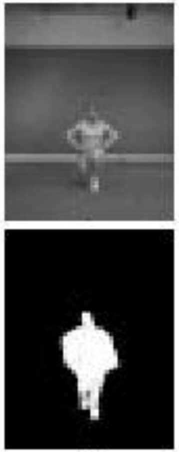

6

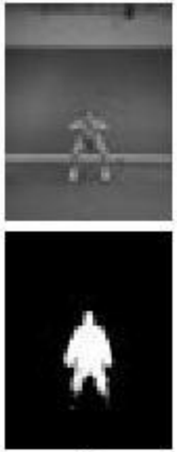

7

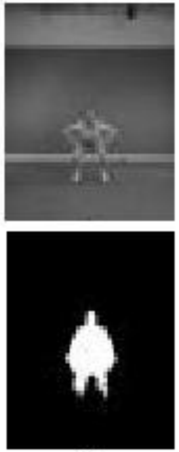

8

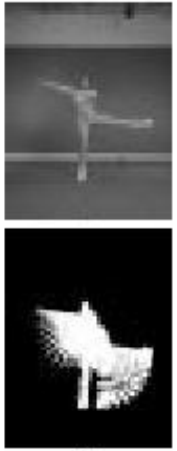

9

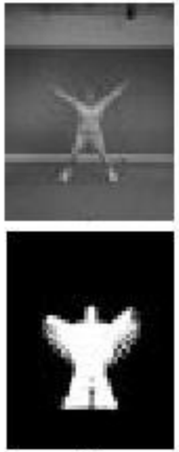
10

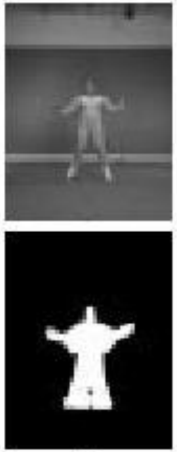

11

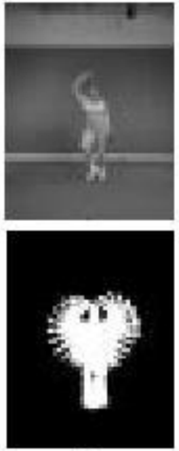

12


13


14


15


16


17

18

How to recognize these images?

- In 1999, old style computer vision:
  1. compute some summarization statistics of the pattern
  2. construct generative model
  3. recognize based upon those statistics.

Image moments

Moments summarize a shape given image $I(x, y)$:

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

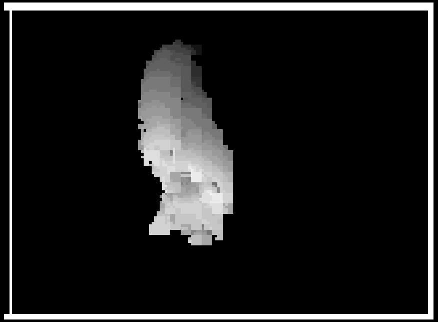
Central moments are translation invariant:

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

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

Hu moments

- Translation and rotation *and scale* invariant
- We chose 7 moments
- Apply to Motion History Image for global space-time "shape" descriptor



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

Hu Moments $(h_1 \ldots h_6)$

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

Hu Moments (h_7)

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

Build a classifier

Remember Generative vs Discriminative?

- Generative – builds model of each class; compare all
- Discriminative – builds model of the *boundary* between classes

Build a classifier

How would you build decent generative models of each class of action?

- Use a Gaussian in Hu-moment feature space
- Compare *likelihoods*: *p(data | model of action i)*
- If have priors, use them by Bayes rule

$$p(\text{model}_i \mid \text{data}) \propto \text{p}(\text{data} \mid \text{model}_i)\, \text{p}(\text{model}_i)$$

- Otherwise just use likelihood. Or even NN.

Recognizing temporal templates

- For MEI and MHI compute global properties (e.g. Hu moments)
    - Treat both as grayscale images.
- Collect statistics on distribution of those properties over people for each movement.
- At run time, construct MEIs & MHIs backwards in time
    - Recognizing movements as soon as they complete.

Recognizing temporal templates: Pros

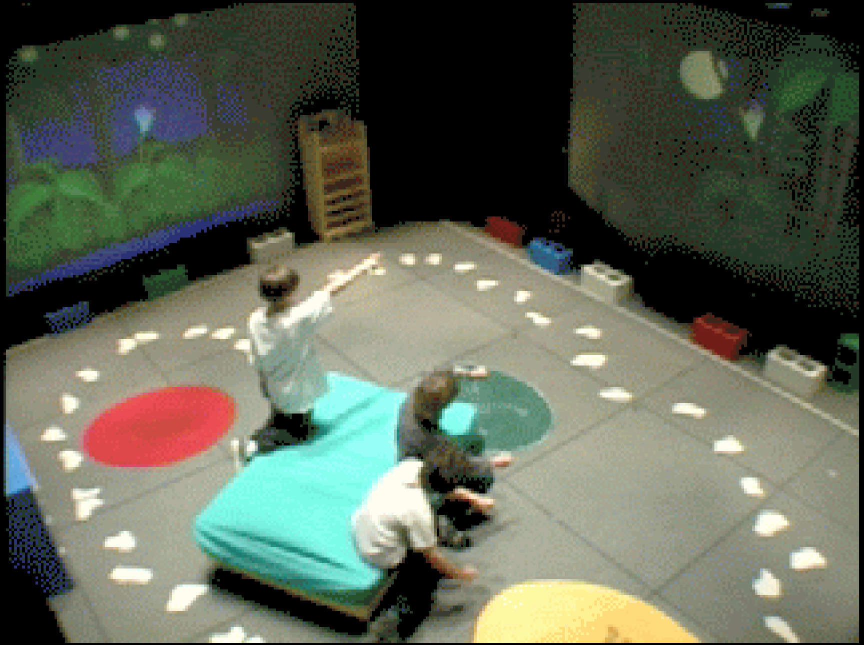
- Linear time scaling
  - Compute range of τ using the min and max of training data.
- Simple recursive formulation, so very fast.
- Filter implementation obvious, so biologically “relevant”.

Best reference is *Bobick and Davis, PAMI 2001*

Virtual PAT (Personal Aerobics Trainer)

- Uses MHI recognition
- Portable IR background subtraction system (CAPTECH ’98)

The KidsRoom



Recognizing Movement in the KidsRoom

- First teach the kids, then observe
- Temporal templates “plus” (but in paper)
- Monsters always do something, *but only speak it when sure*