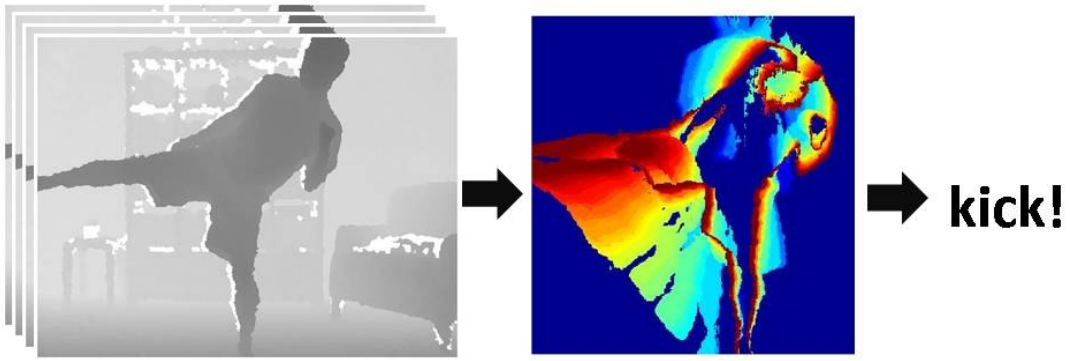
CS 4495 Computer Vision
Activity Recognition

Aaron Bobick

School of Interactive Computing

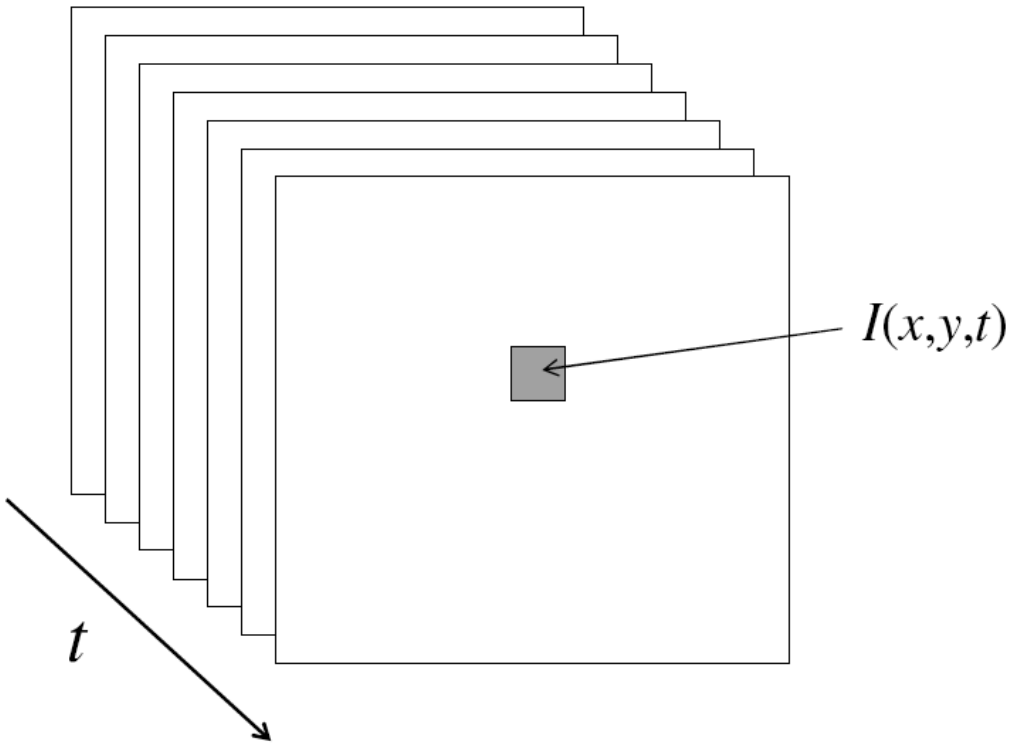


Administrivia

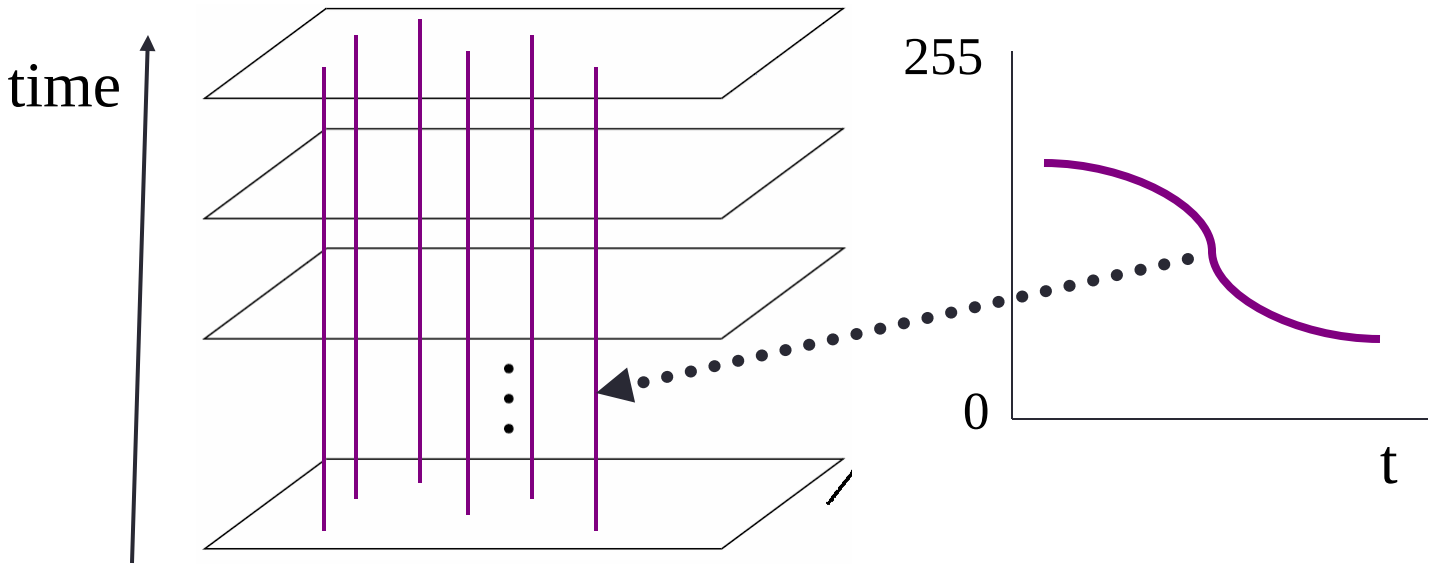
- PS6 – should be working on it! Due Sunday Nov 24th.
- Exam: Tues November 26th .
  - Short answer and multiple choice (mostly short answer)
  - Study guide is posted in calendar.
- PS7 – we hope to have out by 11/26. Will be straight forward implementation of Motion History Images

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)

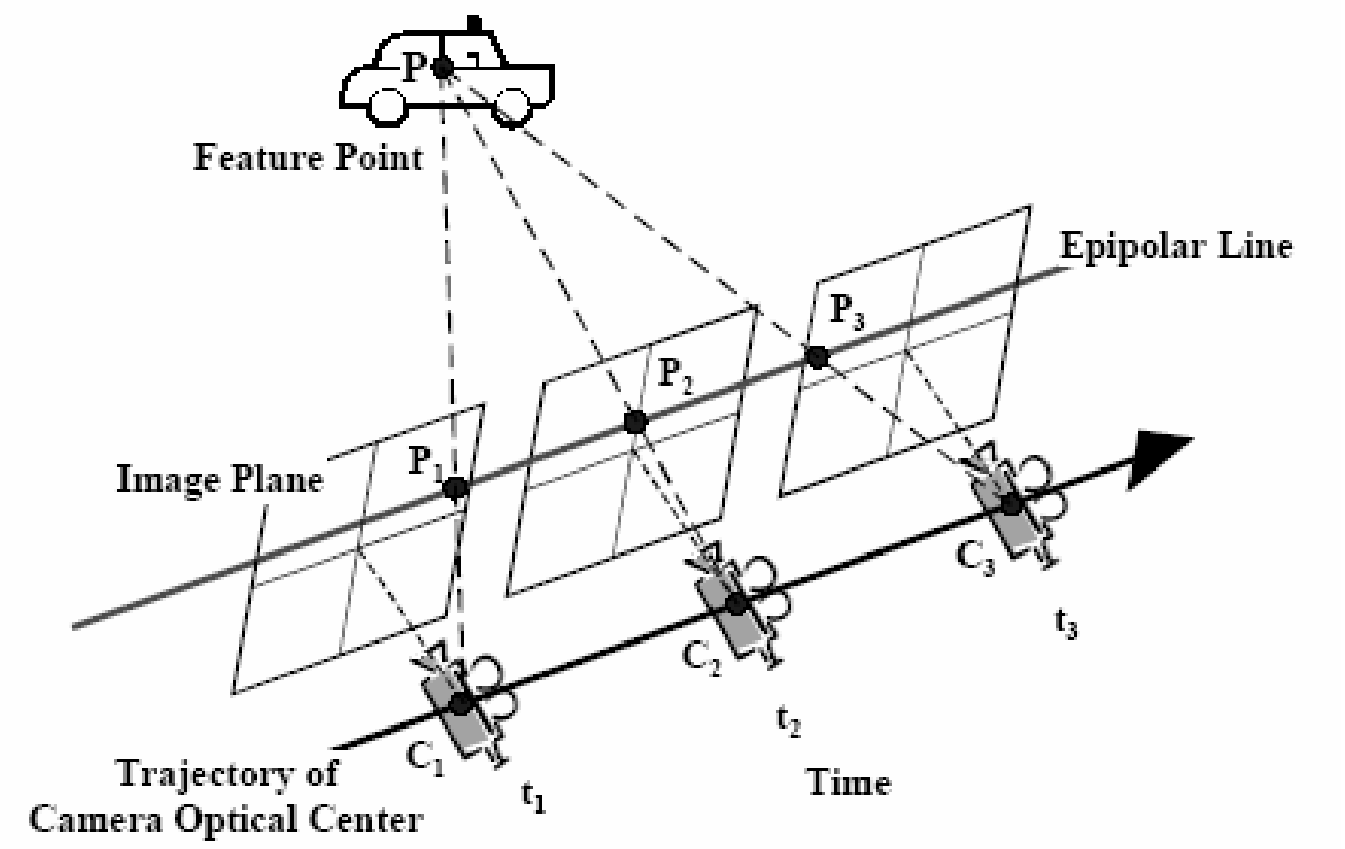


Video as an "Image Stack"

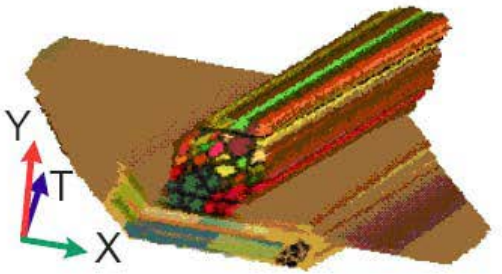


- Can look at video data as a spatio-temporal volume
  - If camera is stationary, each line through time corresponds to a single ray in space

Aside: Epipolar Plane ("EPI") images



Aside: Epipolar Plane (“EPI”) images



EPI images and activity

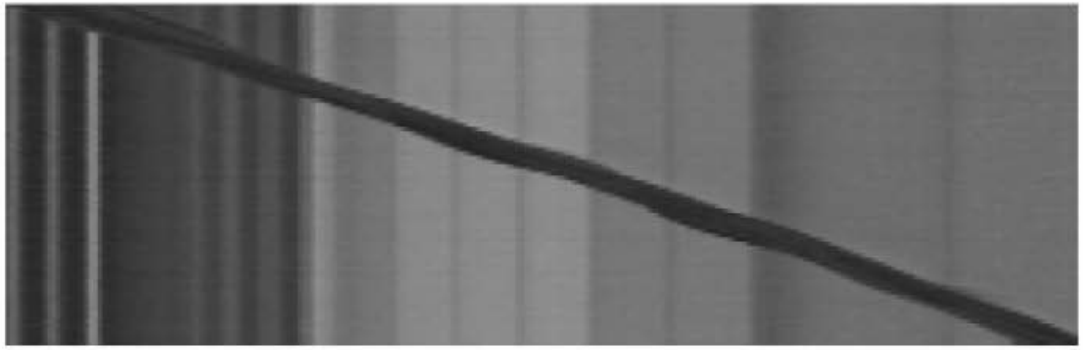


Figure 9 An XT-slice taken at the walker's head height, indicating the head mostly only undergoes translational movement during walking.

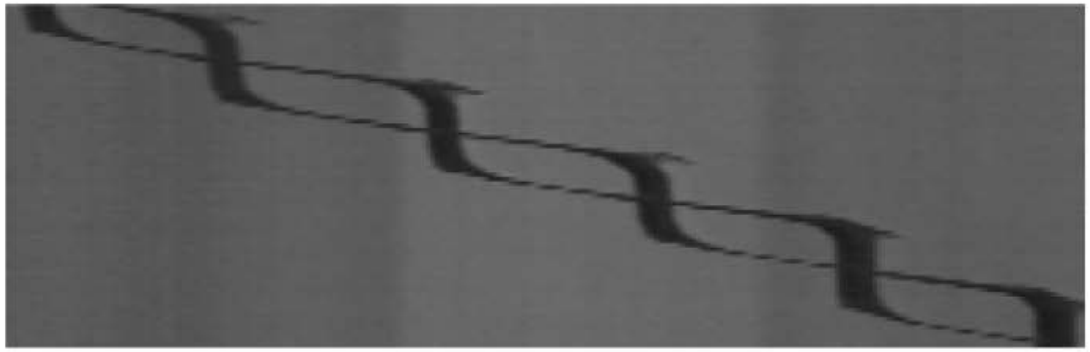
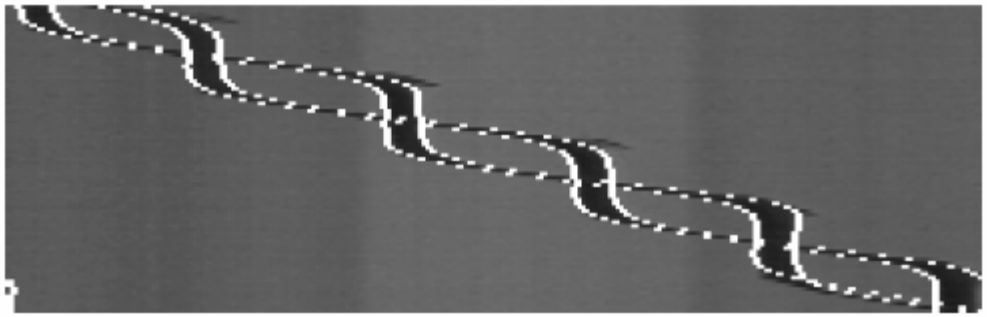
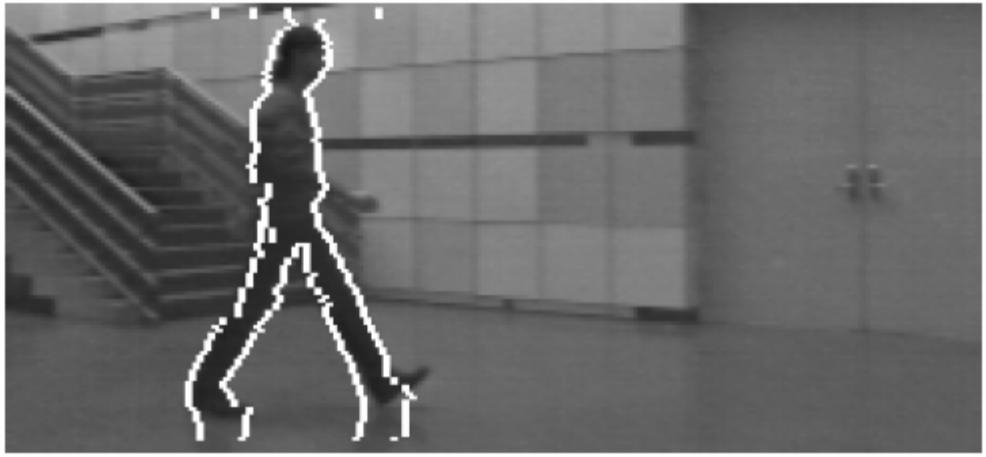


Figure 10 A slice taken at the height of the walker's ankles. The criss-crossing of the walker's legs as the walker moves from left to right is given as a unique braded signature for walking patterns

EPI images and activity





Processing video: object detection

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

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

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 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: background subtraction

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

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

$$| \quad - \quad | > Th$$

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 filtering

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

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

$$\Downarrow$$

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

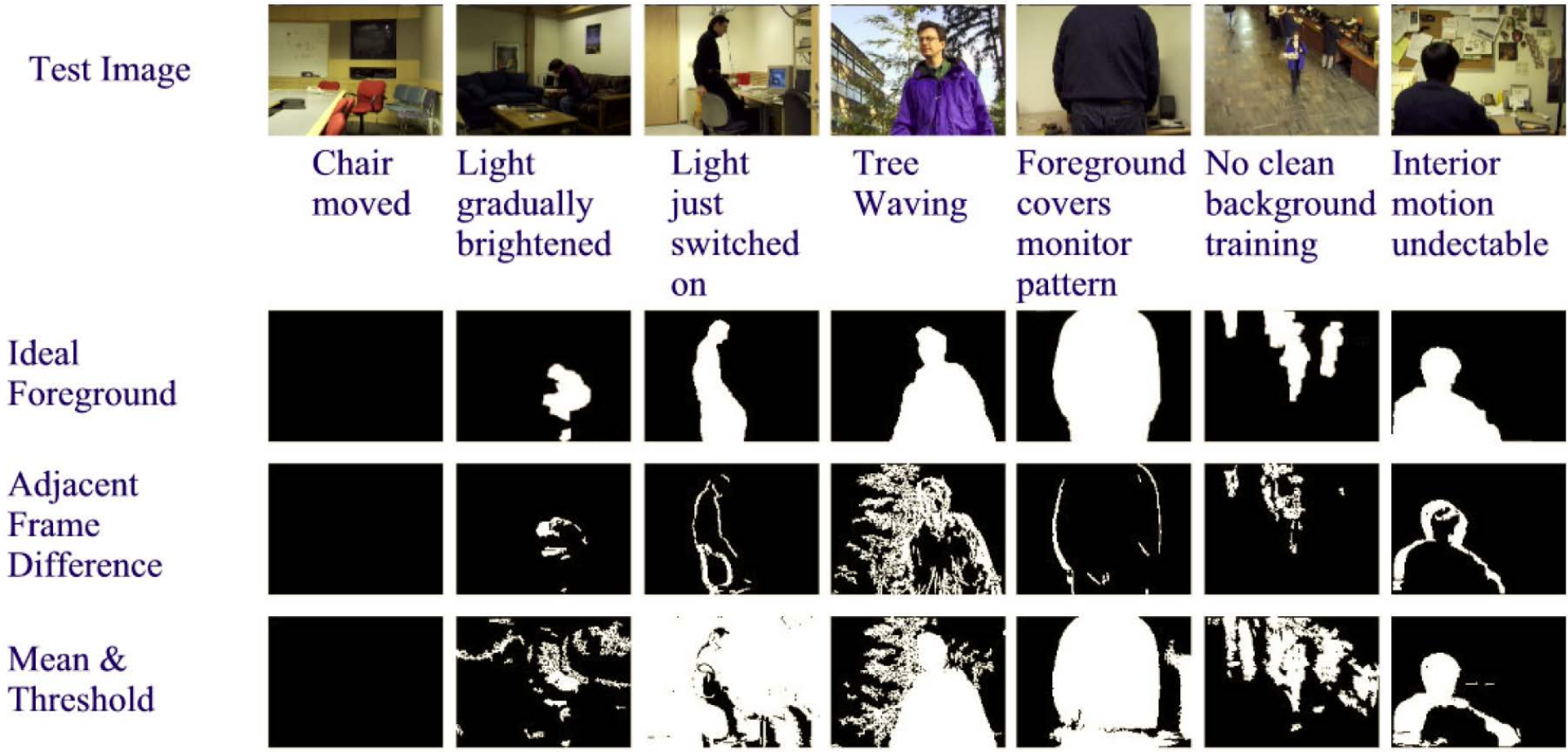
- For $n = 10$:

Estimated Background

Foreground Mask

Slide credit: Birgi Tamersoy

Frame differences vs. background subtraction



- Toyama et al. 1999

Median Filtering

- 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, CMU

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.

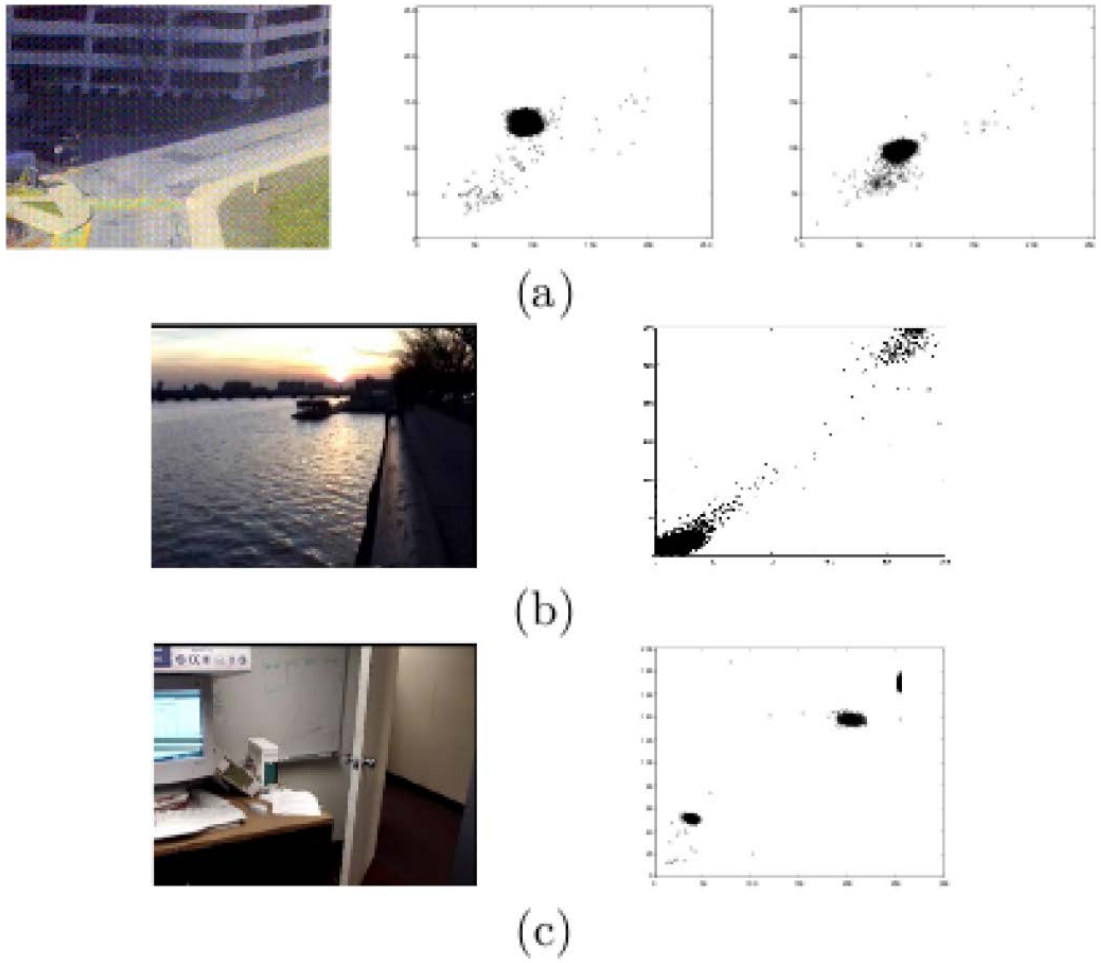
Disadvantages:

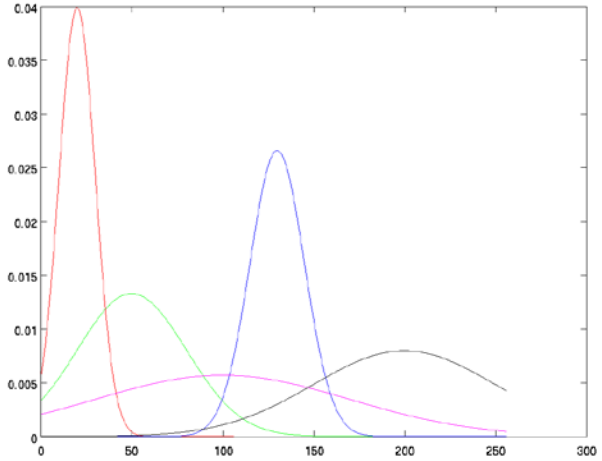
- Accuracy of frame differencing depends on object speed and frame rate
- Median background model: relatively high memory requirements.
- Setting global threshold Th…

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:

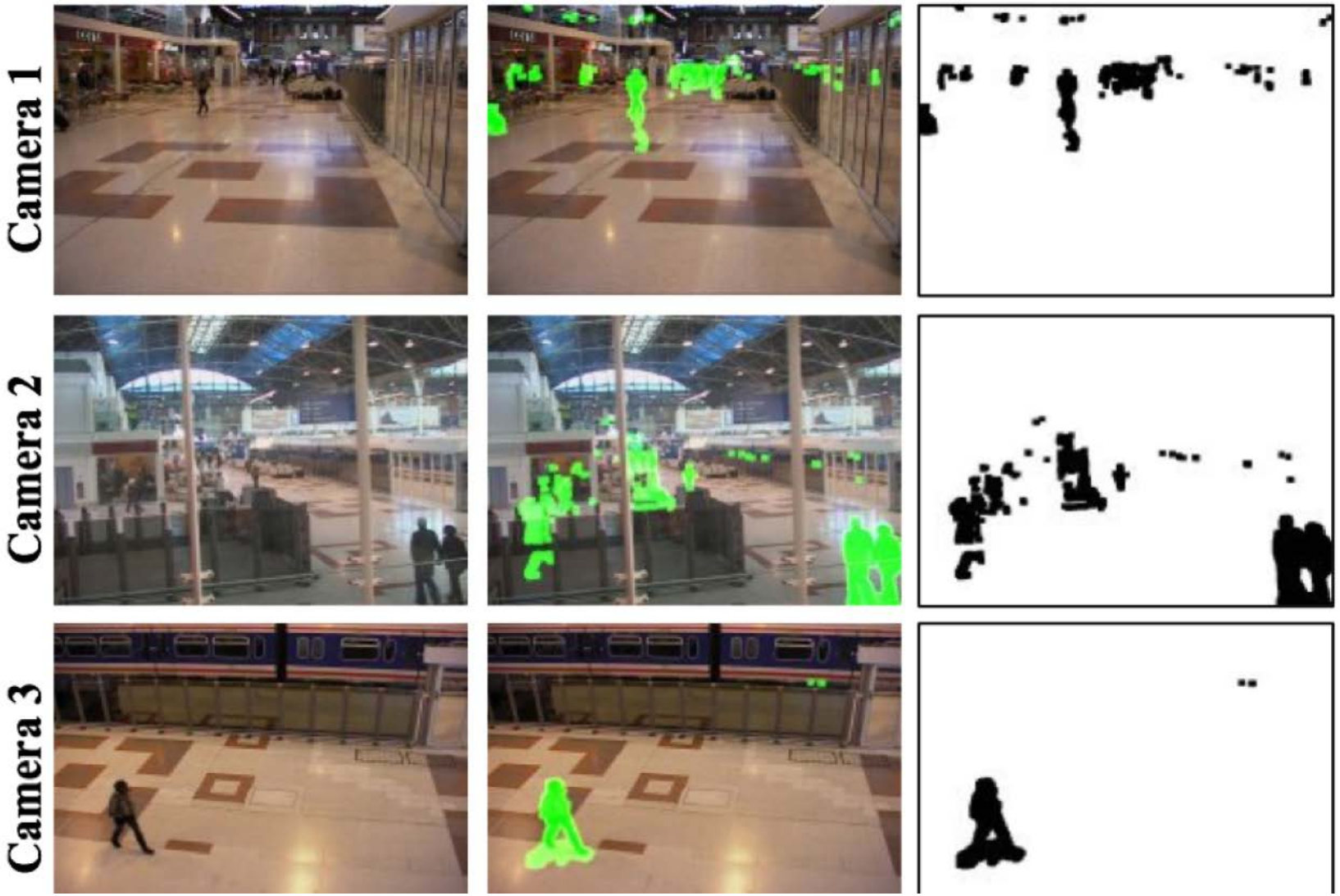
- "**Event**": a single instant in time detection.

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

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

Adapted from Venu Govindaraju and A.Bobick

Surveillance



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

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: accurate tracks in spite of occlusion, ambiguity, low resolution

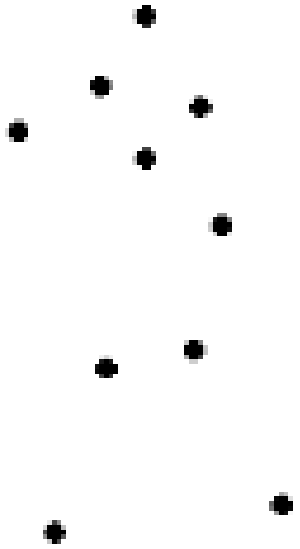
- **Model-based activity recognition:**
  - Given some lower level detection of actions (or events) recognize the activity by comparing to some structural representation of the activity
  - Needs to handle uncertainty.

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

- Recently: **"Activity-recognition" from static image**
  - Imagine a picture of a person holding a flute. What are they doing?

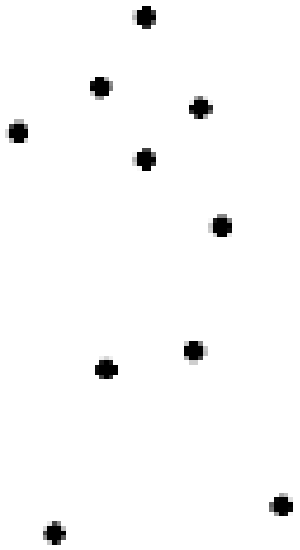
Motion and perceptual organization

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



Motion and perceptual organization

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



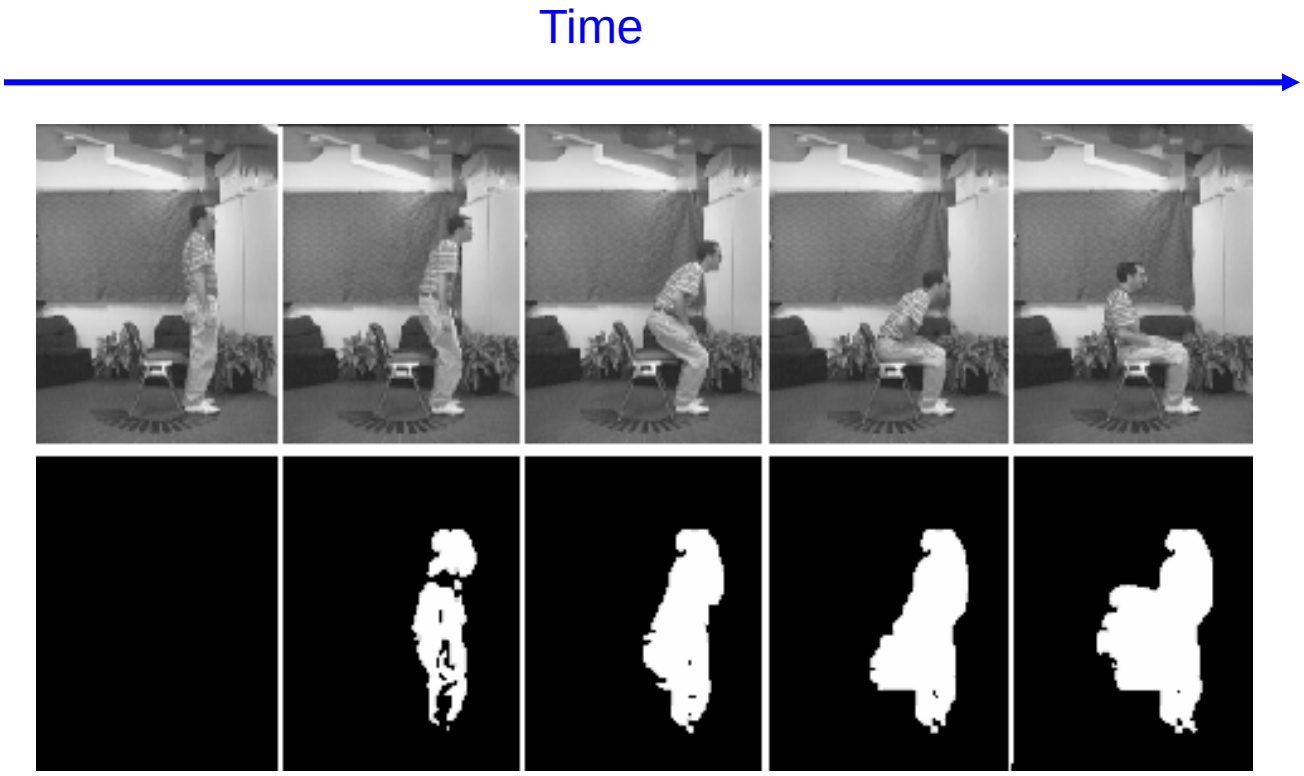
Example

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

Video from Davis & Bobick

Motion energy images

- Spatial accumulation of motion.
- Collapse over specific time window.
- Motion measurement method not critical (e.g. motion differencing).

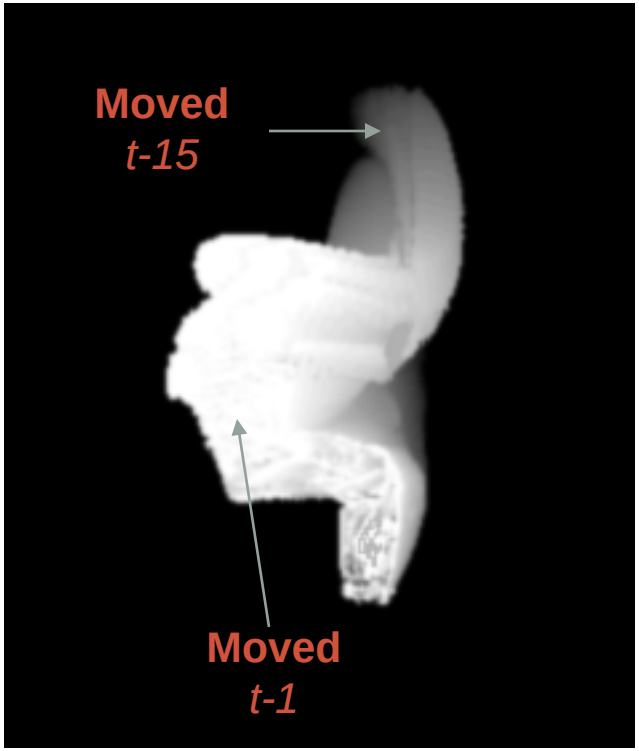


Motion history images

- Motion history images are a different function of temporal volume.
- Pixel operator is replacement decay:

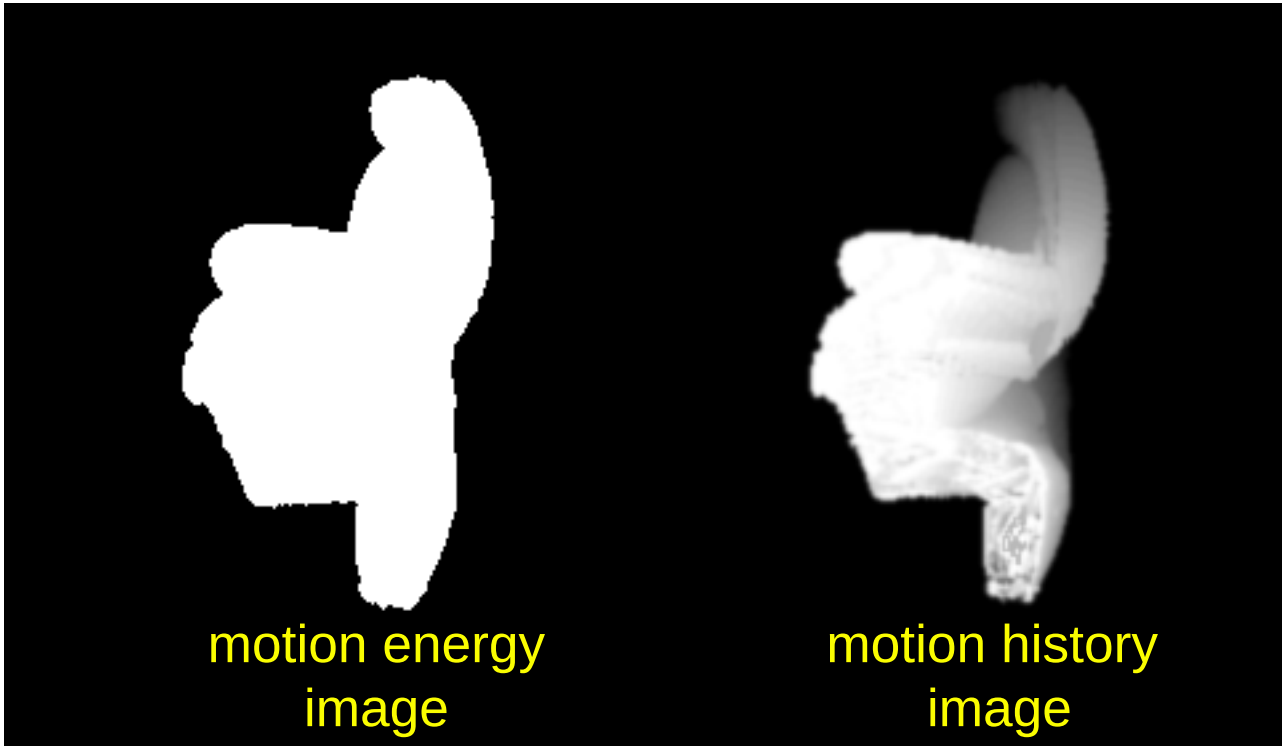
  `if moving` $I_\tau(x,y,t) = \tau$
  `otherwise`
  $I_\tau(x,y,t) = \max(I_\tau(x,y,t-1)-1, 0)$

- Trivial to construct $I_{\tau-k}(x,y,t)$ from $I_\tau(x,y,t)$ so can process multiple time window lengths without more search.
- MEI is thresholded MHI

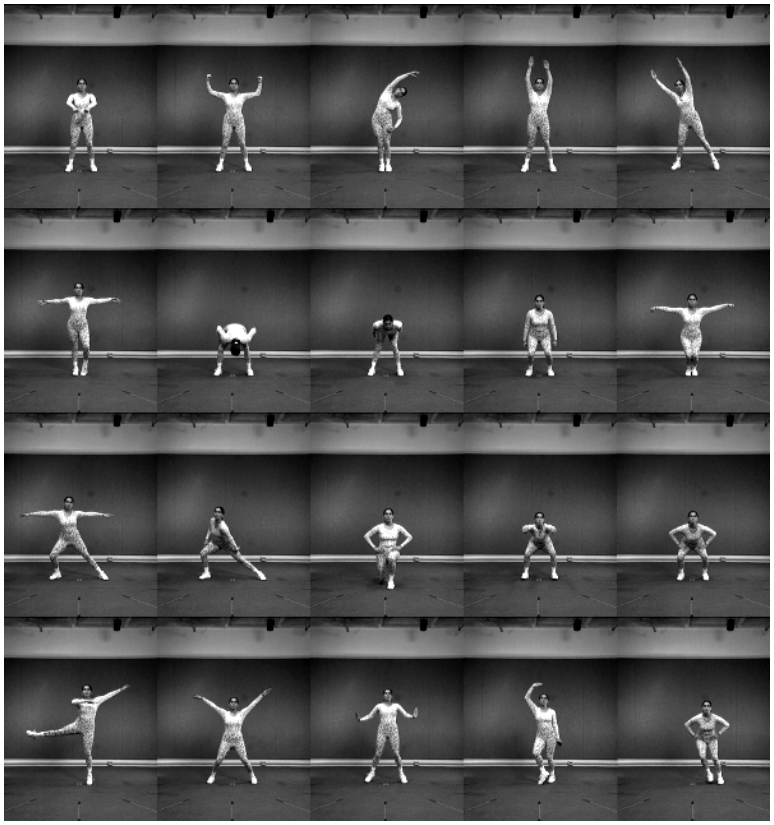


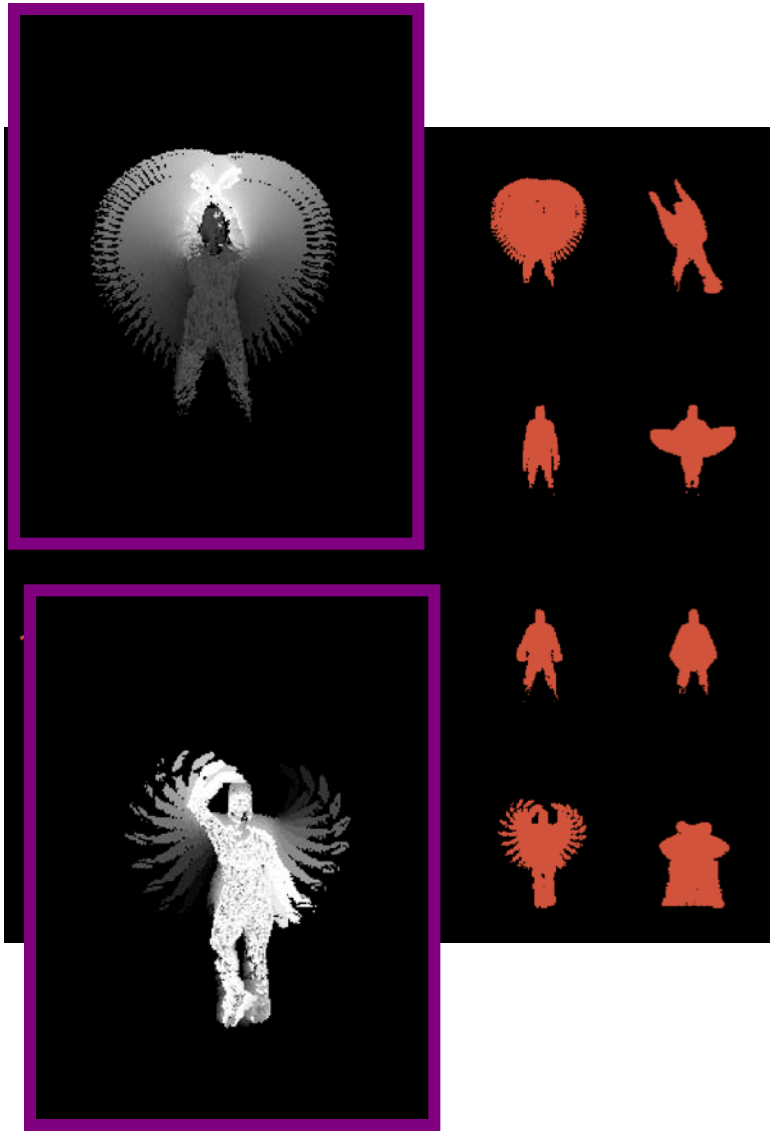
Temporal-templates

- *MEI+ MHI = Temporal template*

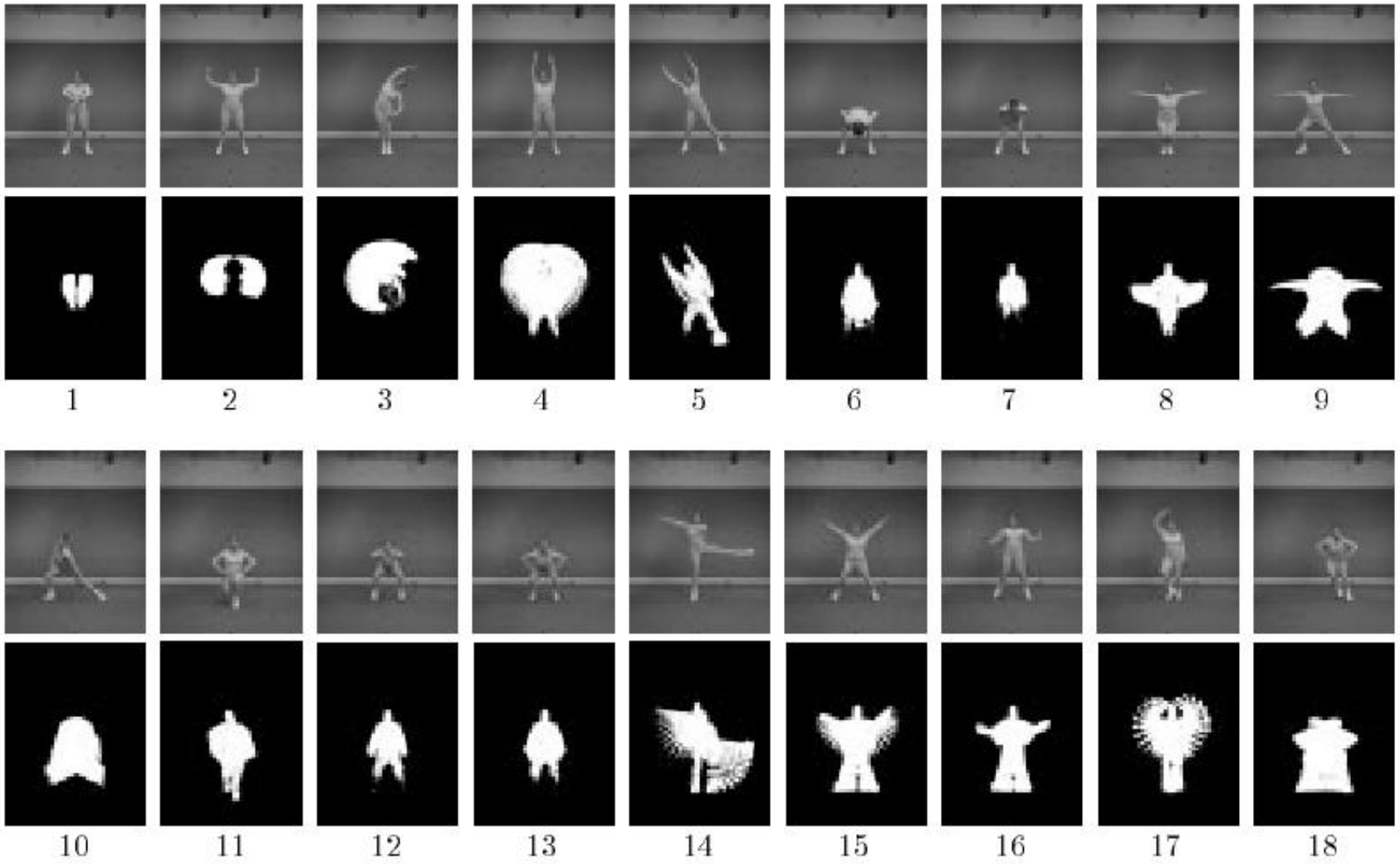


Aerobics examples





Motion Energy Images



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

How to recognize these images?

- These are gray scale blob like images.
- 100 years of computer vision for recognizing gray blobs (for small values of a hundred).
- 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 $I(x,y)$

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

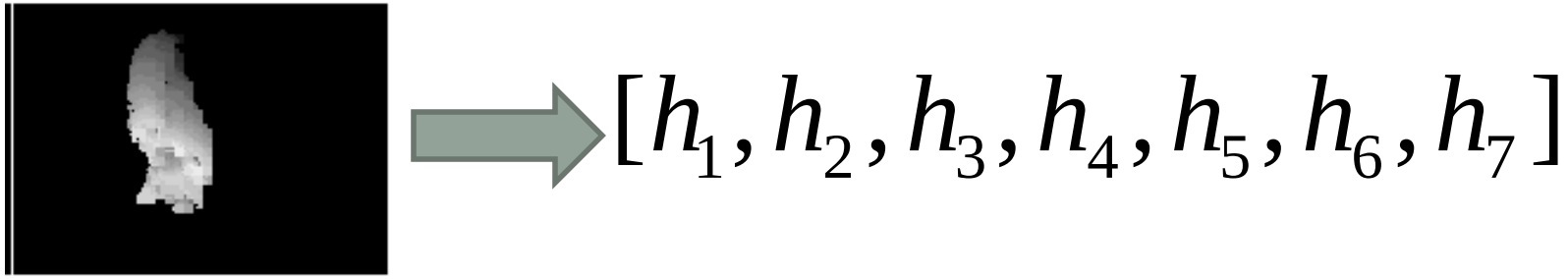
Central moments are translation invariant:

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

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

Hu moments

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



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

Hu moments

$$h_1 = \mu_{20} + \mu_{02},$$

$$h_2 = (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2,$$

$$h_3 = (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2,$$

$$h_4 = (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2,$$

$$\begin{aligned} h_5 = {} & (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\ & + (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) \\ & \cdot [3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2], \end{aligned}$$

$$\begin{aligned} h_6 = {} & (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\ & + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}), \end{aligned}$$

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

Build a classifier

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

- 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 use NN? (Problem Set!)
- More on classification on Dec 3

Recognizing temporal templates

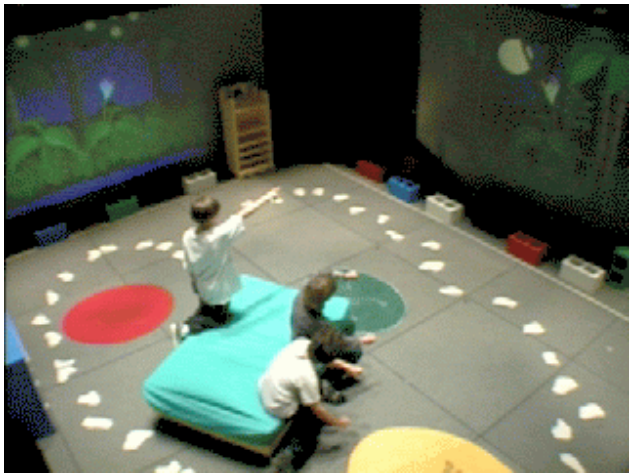
- For MEI and MHI compute global properties (e.g. Hu moments). Treat both as grayscale images.
- Collect statistics on distribution of those properties over people for each movement.
- At run time, construct MEIs and MHIs backwards in time.
  - Recognizing movements as soon as they complete.
- Linear time scaling.
  - Compute range of τ using the min and max of training data.
- Simple recursive fomulation so very fast.
- Filter implementation obvious so biologically “relevant”.
- Best reference is *PAMI 2001, Bobick and Davis*

Virtual PAT (Personal Aerobics Trainer)

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

The KidsRoom

- A narrative, interactive children's playspace.
- Demonstrates computer vision "action" recognition.
- Sometimes, possible because the machine knows the context.
- A kinder, gentler C3I interface
- Ported to the Millenium Dome, London, 2001
- Summary and critique in Presence, August 1999.



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.*

So far…

- **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

A little philosophy…

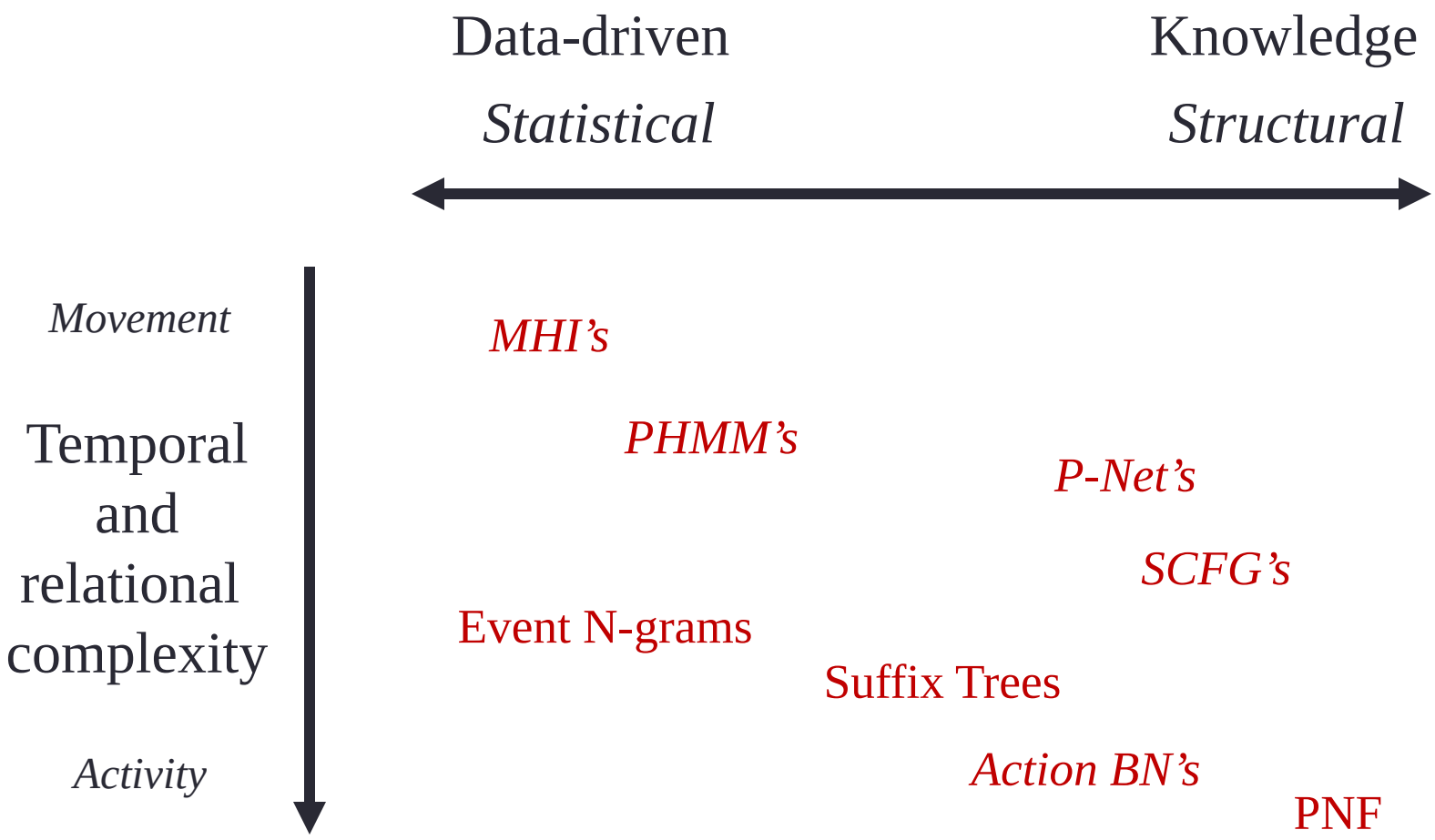
What is the goal of a representation of activity/behaviors?

- Recognition implies representation

- Representations can talk about what events *ARE*:
  - Definitional – but sometimes not “real” because primitives not grounded
  - Permits specification of reasoning mechanism
  - Context can be made explicit (but is not usually)
  - Hard to learn

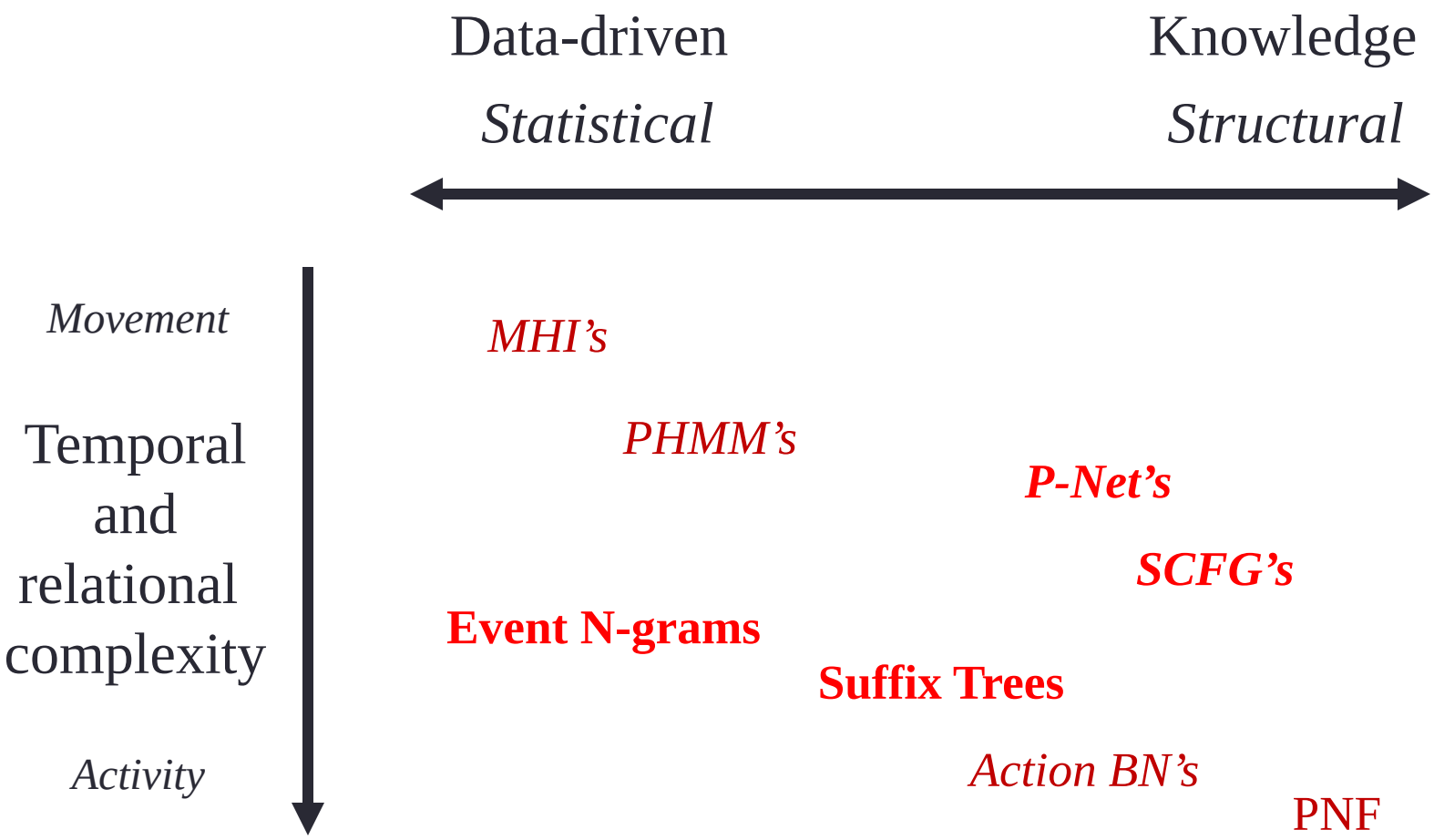
- Representations can talk about what events *LOOK LIKE*:
  - Sometimes learnable, always well defined primitives
  - Typically not guaranteed to be complete
  - Have no explanatory power
  - Often leverages (ie is wholly dependent upon) context – makes it learnable from specific data

Data-driven vs Knowledge-taught



Skip to P-Nets?

Data-driven vs Knowledge-taught



Structure and Statistics

- Grammar-based representation and parsing
  - Highly expressive for activity description
  - Easy to build higher level activity from reused low level vocabulary.
- P-Net (Propagation nets) – really stochastic Petri nets
  - Specify the structure – with some annotation can learn detectors and triggering probabilities
- Statistics of events
  - Low level events are statistically sequenced – too hard to learn full model.
  - N-grams or suffix trees

"Higher-level" Activities: Known structure, uncertain elements

- Many activities are comprised of a priori defined sequences of primitive elements.
  - Dancing, conducting, pitching, stealing a car from a parking lot.
  - The states are not hidden.
- The activities can be described by a set of grammar-like rules; often ad hoc approaches taken.
- But, the sequences are uncertain:
  - Uncertain performance of elements
  - Uncertain observation of elements

The basic idea and approach

Idea: split the problem into:

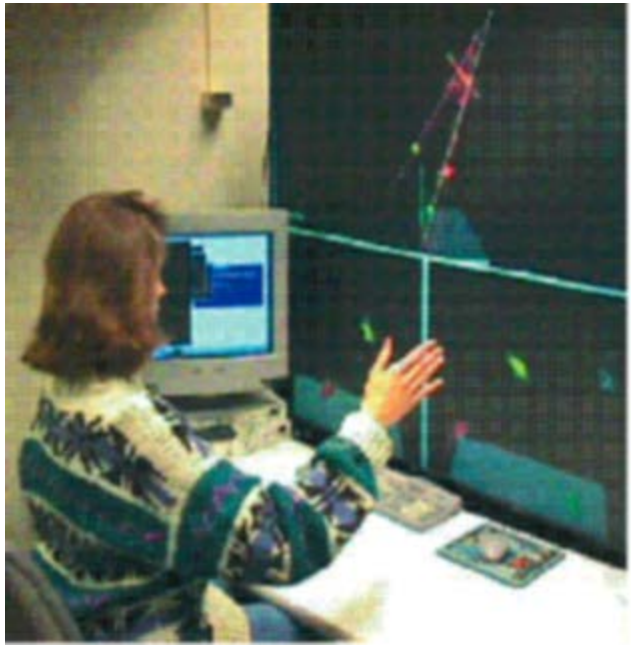
- Low-level primitives with uncertain feature detection (individual elements might be HMMs)
- High-level description found by parsing input stream of uncertain primitives.

Approach:

- Extend Stochastic Context Free Grammars to handle perceptually relevant uncertainty.

Stochastic CFGs

- Traditional SCFGs have probabilities associated with the production rules. Traditional parsing yields most likely parse given a known set of input symbols.
  - PIECE -> BAR PIECE | [0.5]
  - BAR [0.5]
  - BAR -> TWO | [0.5]
  - THREE [0.5]
  - THREE -> down3 right3 up3 [1.0]
  - TWO -> down2 up2 [1.0]
- Thanks to Andreas Stolcke's priori work on parsing SCFGs using efficient Earley parser.



Extending SCFGs *(Ivanov and Bobick, PAMI)*

- Within the parser we handle:
  - Uncertainty about input symbols
    - Input is multi-valued string (vector of likelihoods)
  - Deletion, substitution, and insertion errors
    - Introduce error rules
  - Individually recognized primitives typically temporally inconsistent
    - Introduce penalty for overlap.
    - Spatial and temporal consistency enforced.
- Need to define when a symbol has been generated. We have some level primitives or even HMMs.
- How do we learn production probabilities? (Not many examples.) Make sure not too sensitive to them.

Video Sample

Event Grammar and Parsing

- Tracker generates events: ENTER, LOST, FOUND, EXIT, STOP. Tracks have properties (e.g. size) and trajectories.
- Tracker assigns class to each event, though only probabilistically.
- Parser parses single stream that contains interleaved events: (CAR-ENTER, CAR-STOP, PERSON-FOUND, CAR-EXIT, PERSON-EXIT)
- Parser enforces spatial and temporal consistency for each object class and interactions (e.g. to be a PICK-UP, the PERSON-FOUND event must be close to CAR-STOP)
- Spatial and temporal consistency eliminates symbolic ambiguity.

Advantages of SCFGs

- What grammar can do (simplified):

  CAR_PASS -> CAR_ENTER CAR_EXIT **|**
  CAR_ENTER CAR_HIDDEN CAR_EXIT

  CAR_HIDDEN -> CAR_LOST CAR_FOUND **|**
  CAR_LOST CAR_FOUND CAR_HIDDEN

- Skip allows concurrency (and junk):

  PERSON_LOST -> person_lost **|** SKIP person_lost

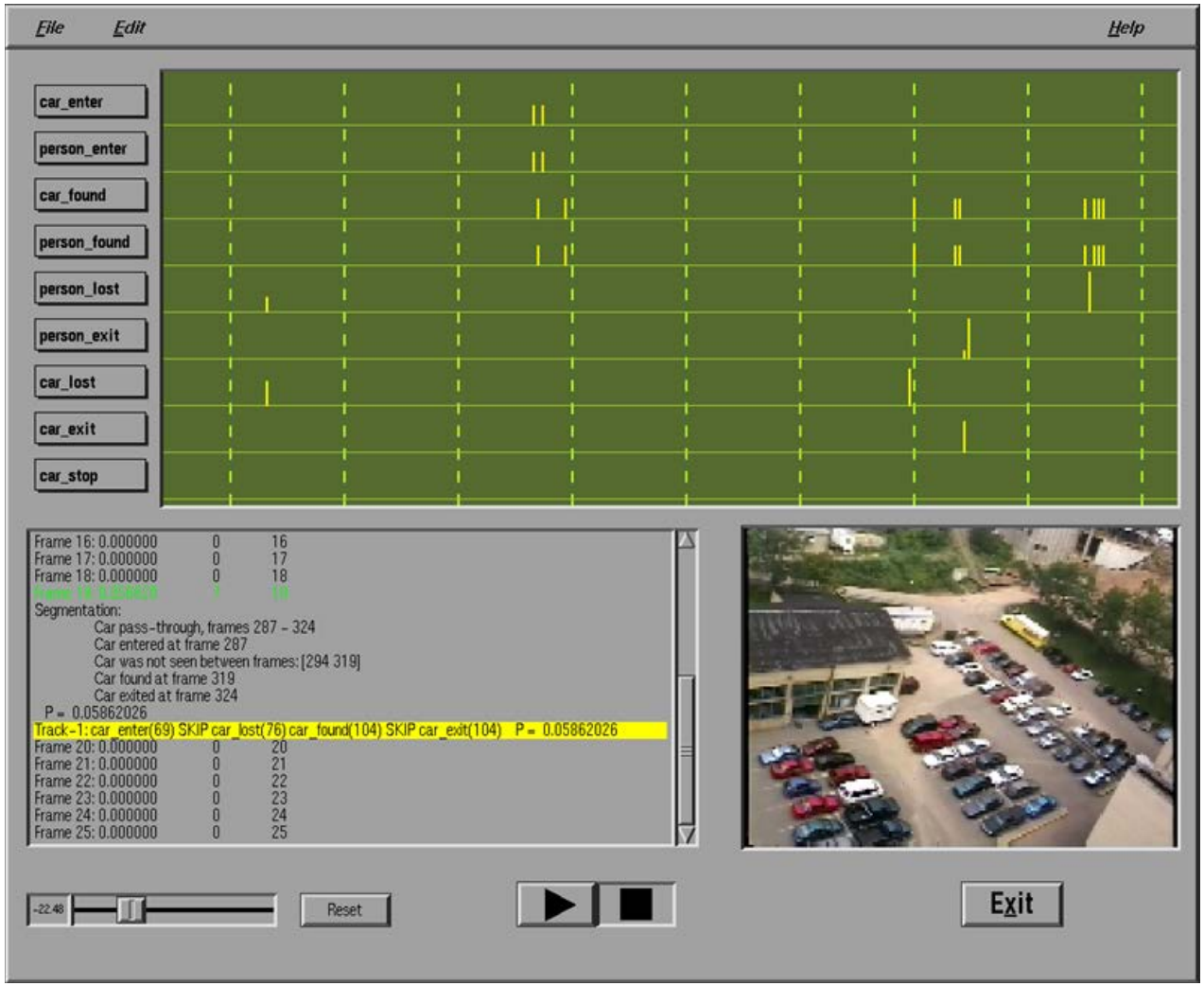
- Concurrent parse:

  Events: ce pe cl cf cs px pl cx

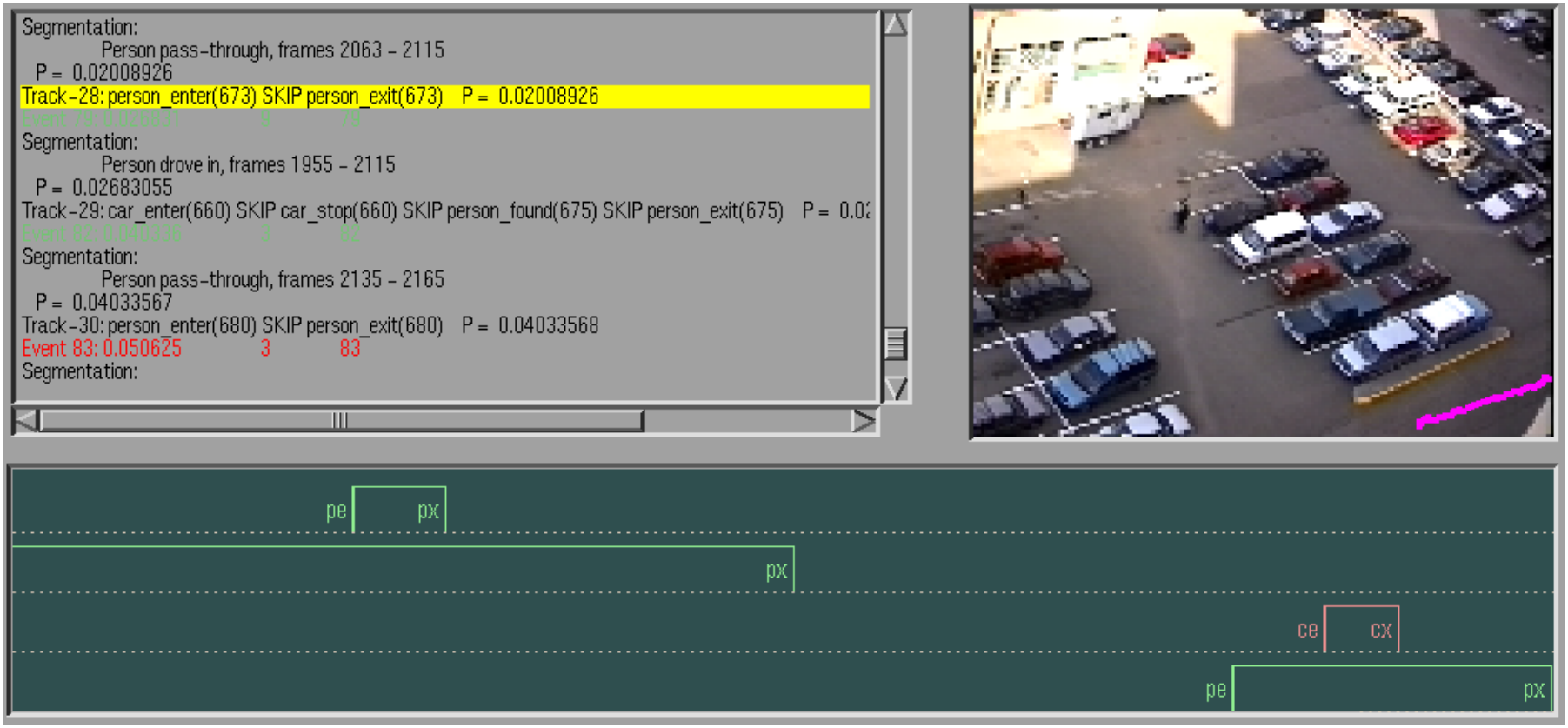
  PICKUP -> **ce** pe **cl cf cs** px **pl cx**

  P_PASS -> ce **pe** cl cf cs **px** pl cx

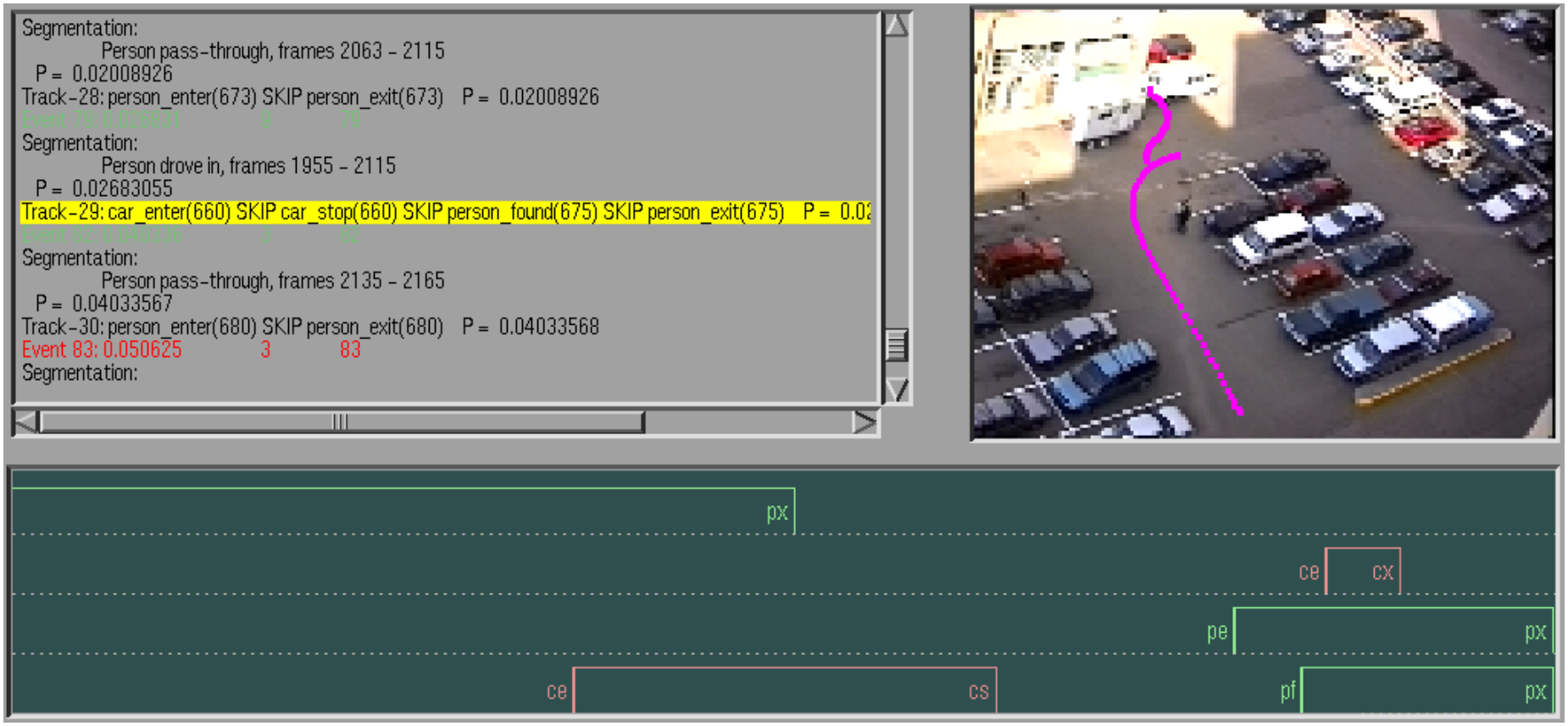
Parsing System



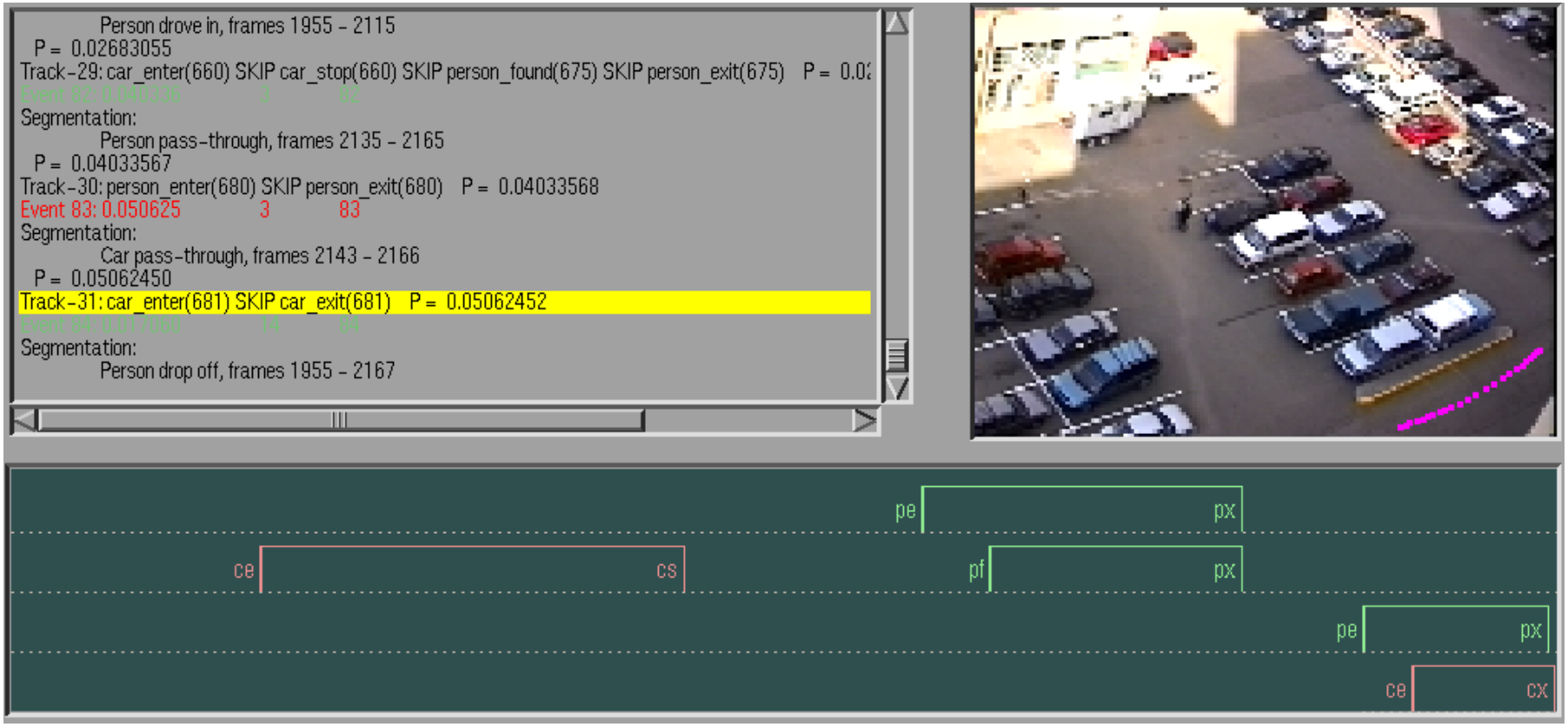
Parse 1: Person-pass- through



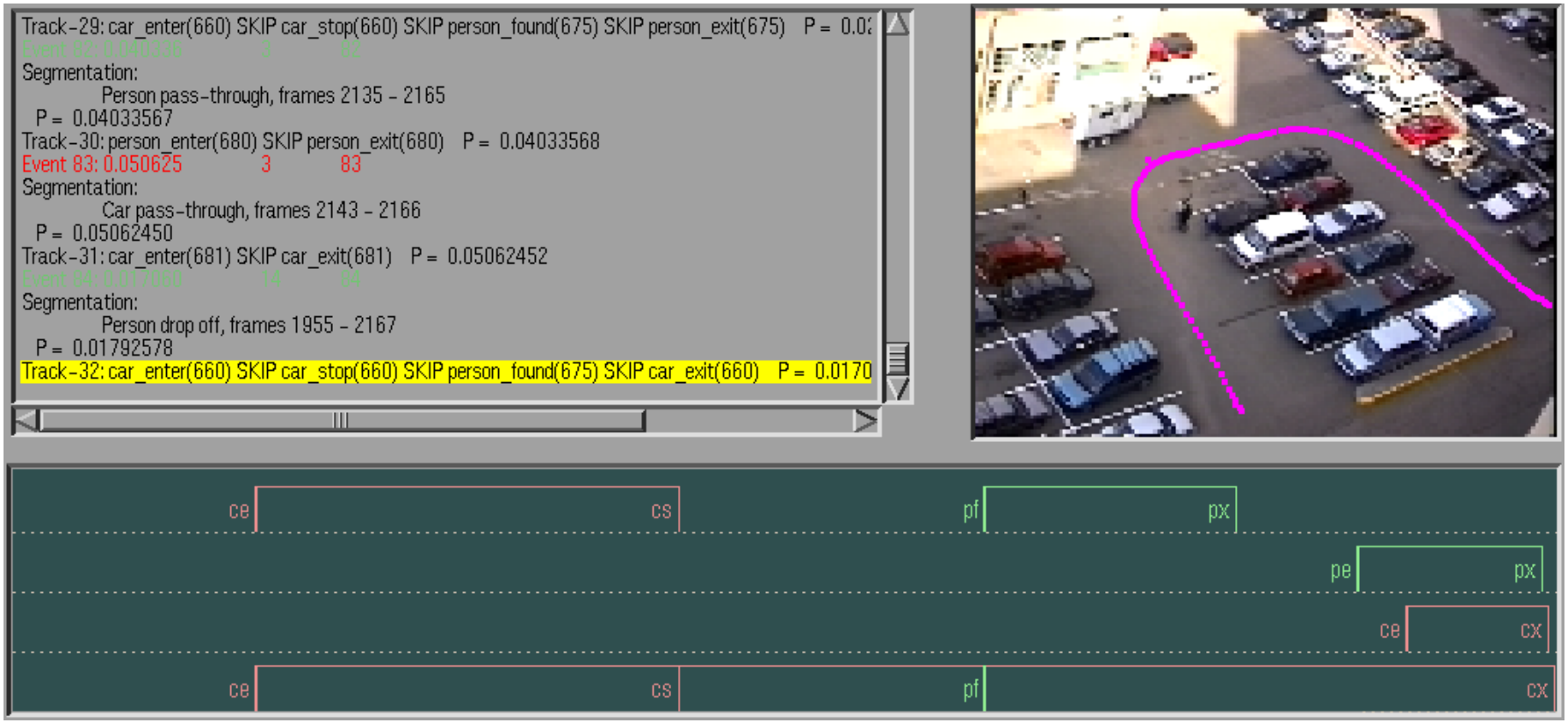
Parse 2: Drive-in



Parse 3: Car-pass-through



Parse 4: Drop-off



Advantages of STCFG approach

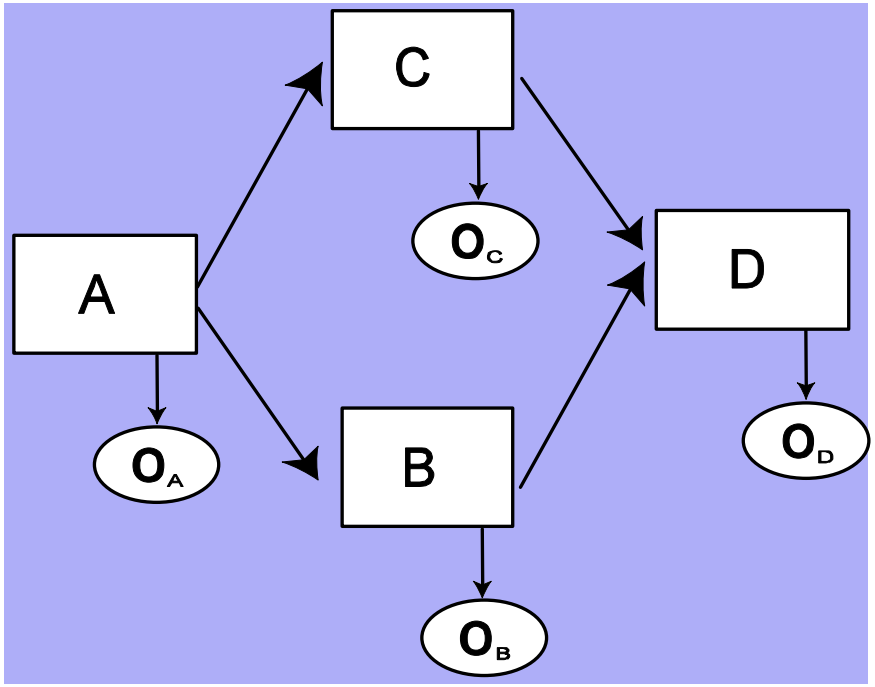
- Structure and components of activities defined a priori and are the right levels of annotation to recover (compare to HMMs).
- FSM vs CFG is not the point. Rather explicit representation of structural elements and uncertainties.
- Often many (enough) examples of each primitive to support training, but not of higher level activity.
- Allows for integration of heterogeneous primitive detectors; only assumes likelihood generation.
- More robust than ad-hoc rule based techniques: handles errors through probability.
- No notion of causality, or anything other than (multi-stream) sequencing.

Advantages of STCFG approach

- Structure and components of activities defined a priori and are the right levels of annotation to recover (compare to HMMs).
- FSM vs CFG is *not* the point. Rather explicit representation of structural elements and uncertainties.
- Often many (enough) examples of each primitive to support training, but not of higher level activity.
- Allows for integration of heterogeneous primitive detectors; only assumes likelihood generation.
- More robust than ad-hoc rule based techniques: *handles errors through probability.*
- *No notion of causality, or anything other than (multi-stream) sequencing.*

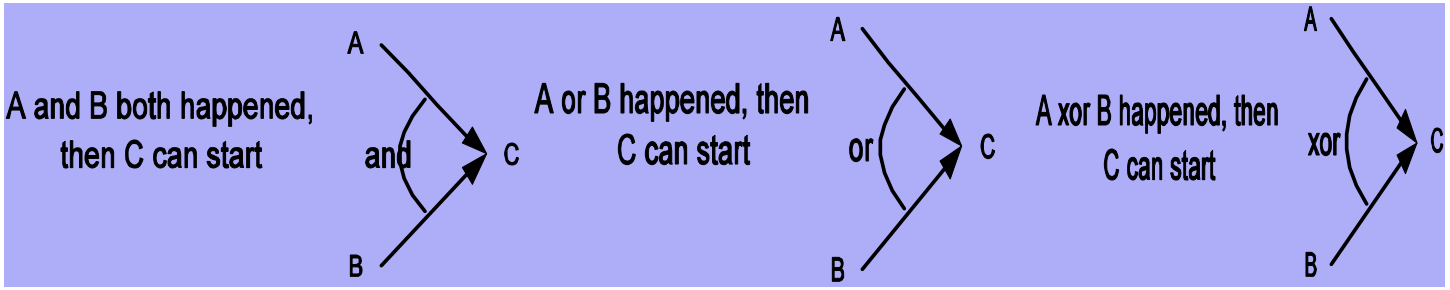
P-Nets (Propagation Networks) (Shi and Bobick, '04 and '06)

- Nodes represent ***activation intervals***
  - Active vs. inactive: Token propagation
- ***More than one node can be active at a time!***
- Links represent partial order as well logical constraint
- ***Duration model*** on each link and node:
  - Explicit model on length of activation
  - Explicit model on length between successive intervals
- Observation model on each node



Conceptual Schema

- Logical relation
  - Autonomous assumption: logic constraint only exists at start/end points of any intervals
  - Condition probability function can represent any logical function

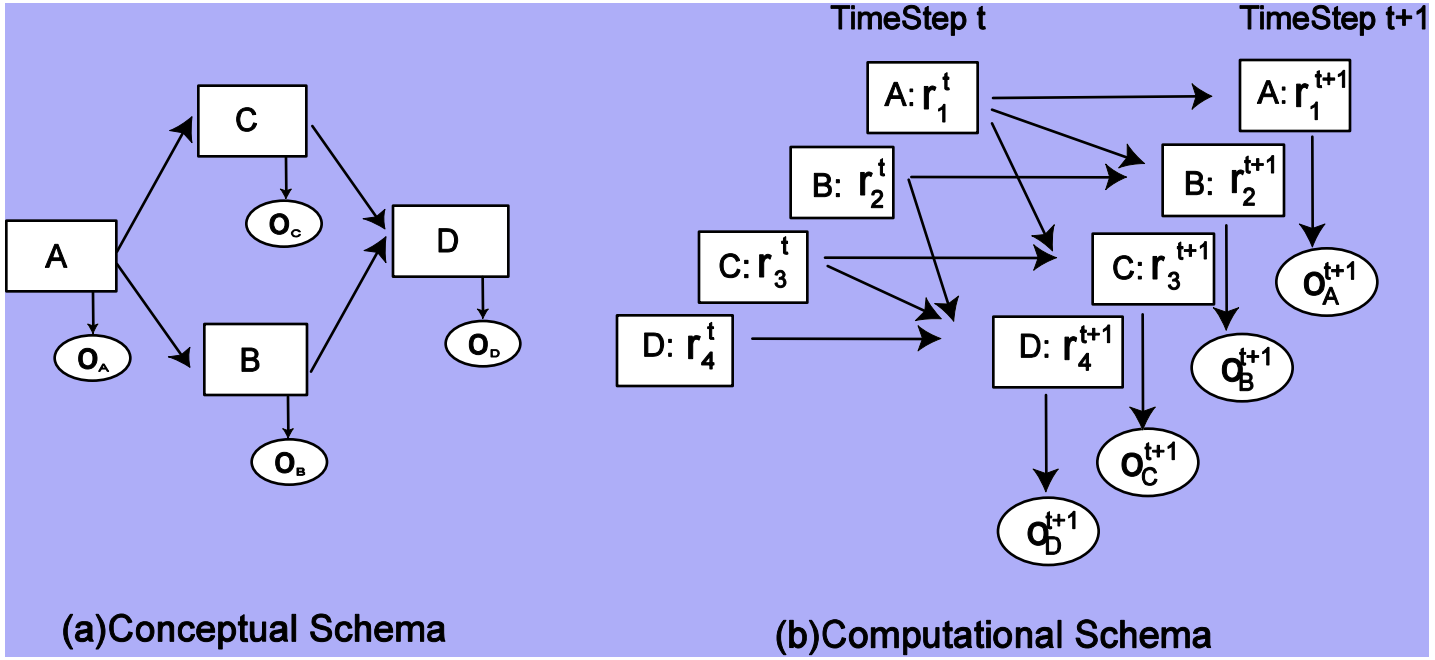


Examples of logic constraint

Propagation Net – Computing

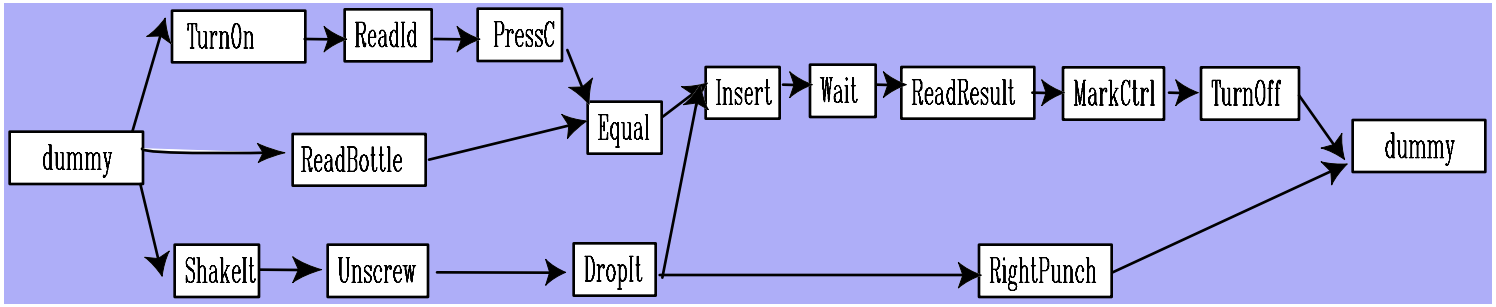
- Computational Schema

  A DBN style rollout to compute corresponding conceptual schema



Experiment: Glucose Project

- Task: monitor an user to calibrate a glucose meter and point out operating error as feedback.
- Constructed 16 node P-Net as representation
- 3 subjects with total of 21 perfect sequences, 10 missing_1_step sequences and 10 missing_6_steps sequences



D-Condensation

Initiate 1 particle at dummy starting node

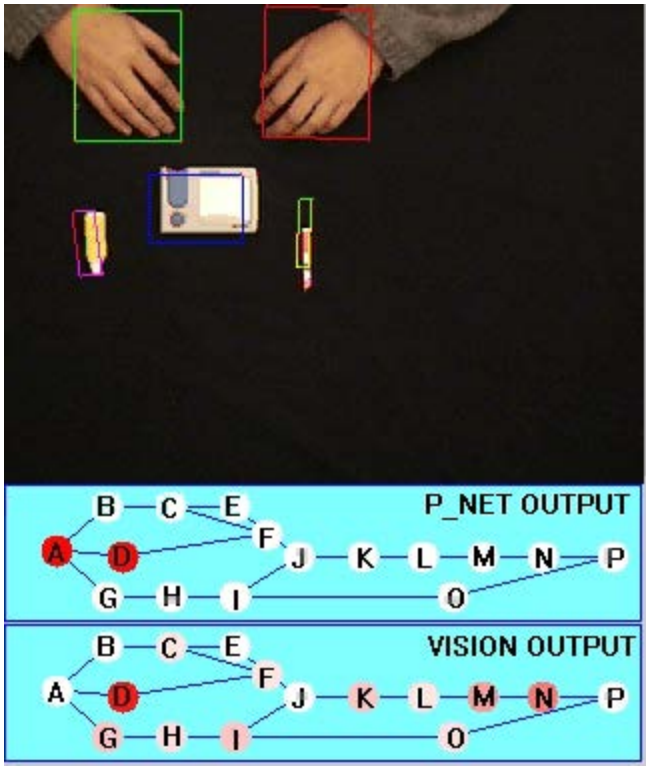
Repeat

- For each particle
  - generate all possible consequent states
  - calculate the probability for each states
- End
- Select n particles to survive

Until the final time steps is reached

Output the path represented by the particle with highest probability

Experiment: Glucose Meter Calibration



Experiment: Classification Performance

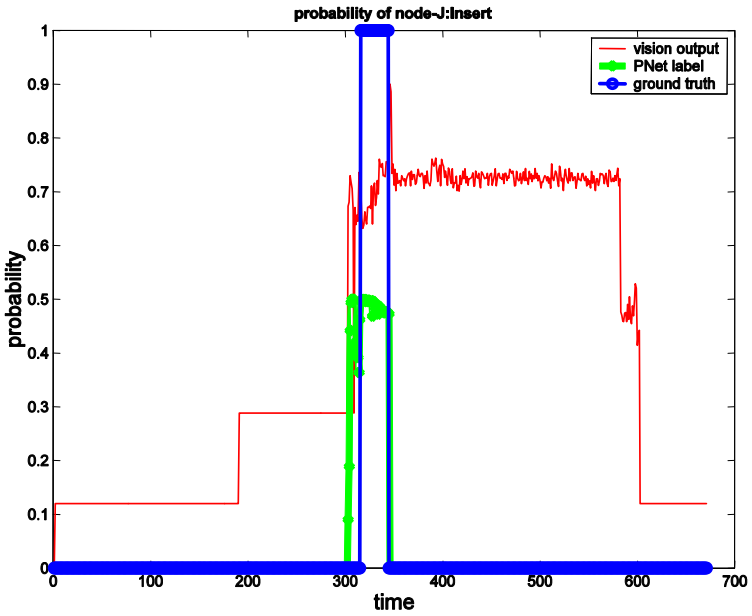
Overall Evaluation.

Sequence Category	Total	Perfect	Almost Right	Negative
Training	6	100%	0%	0%
Perfect	15	100%	0%	0%
Missing One	10	20%	80%†	0%
Missing Six	10	0%	50%‡	50%

Experiment: Label individual frames

Table 2: Labelling individual node.

Individual Node	Overall Success †	Correct Positive ‡	Correct Negative ∗
B:TurnOn	0.9999	1.0000	0.9999
C:RdIdScreen	0.9901	0.9956	0.9897
D:RdIdSstrip	0.9893	0.9333	0.9909
E:PressC	0.9787	0.2344	0.9998
F:Equal	0.9847	0.9267	0.9908
G:ShakeIt	0.9590	0.6003	0.9738
H:Unscrew	0.9563	0.5041	0.9857
I:DropIt	0.9827	0.8584	0.9941
J:Insert	0.9878	0.8643	0.9961
K:Wait	0.9964	0.9987	0.9958
L:ReadResult	0.9966	0.9847	0.9991
M:MarkCtrl	0.9983	0.9720	0.9993
N:TurnOff	0.9967	0.8997	0.9997
O:screw	0.9476	0.6629	0.9617
Average	0.9839	0.8709	0.9914



Labeling individual nodes

Labels on Node J: Insert

Now finish with most recent work…