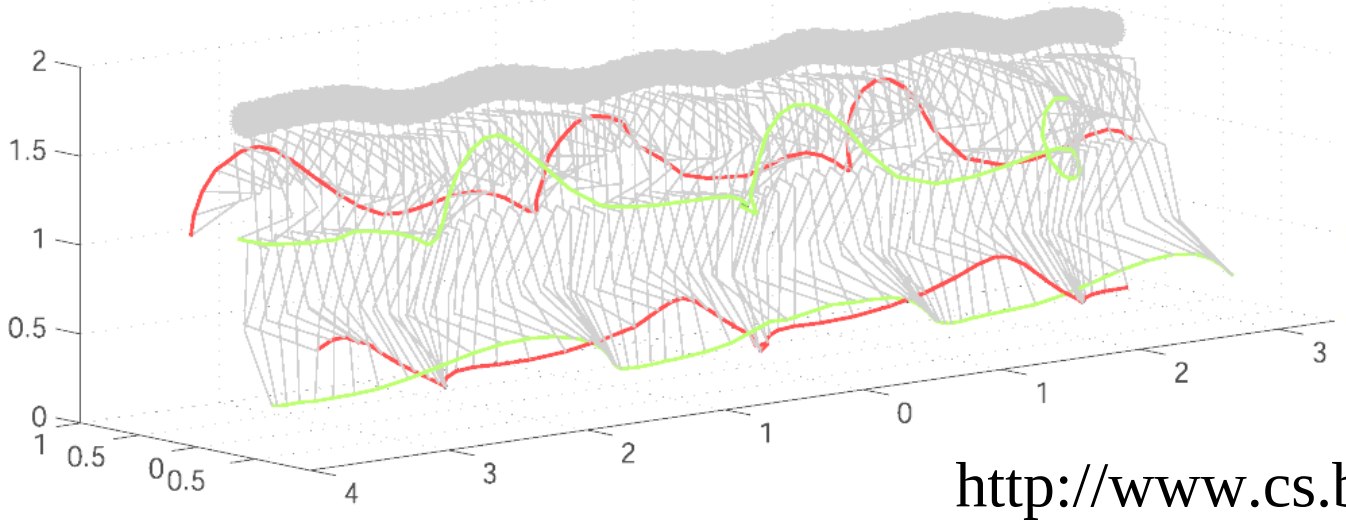
# Learning to See People


**Michael J. Black**

*Department of Computer Science*
Brown University


http://www.cs.brown.edu/~black

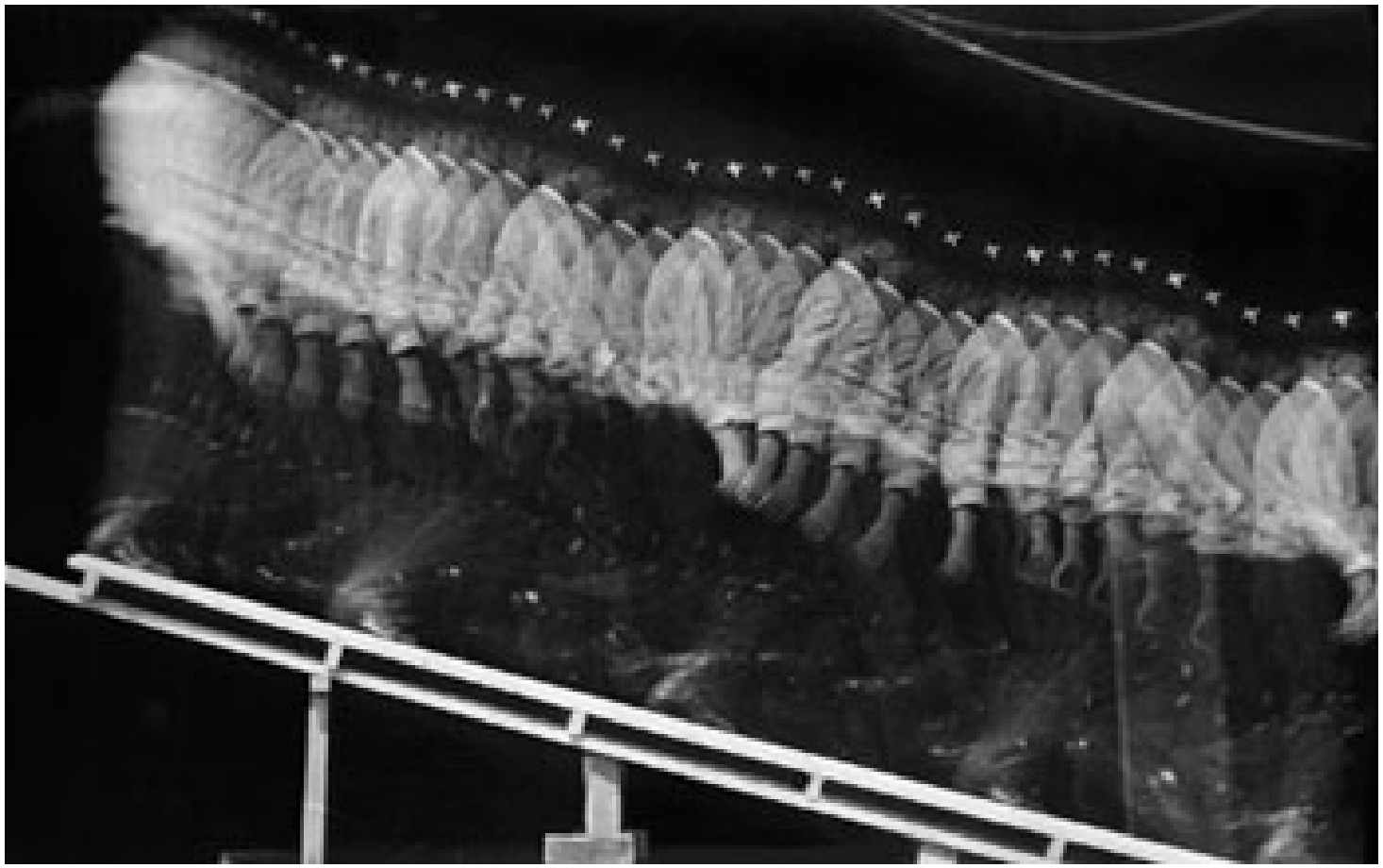
# Capturing Humans in Motion

*If a shadow is a two dimensional projection of the three- dimensional world, then the three-dimensional world as we know it is the projection of the four dimensional universe.*

*Marcel Duchamp*

MARCEL DUCHAMP (1912)
Nude Descending A Staircase

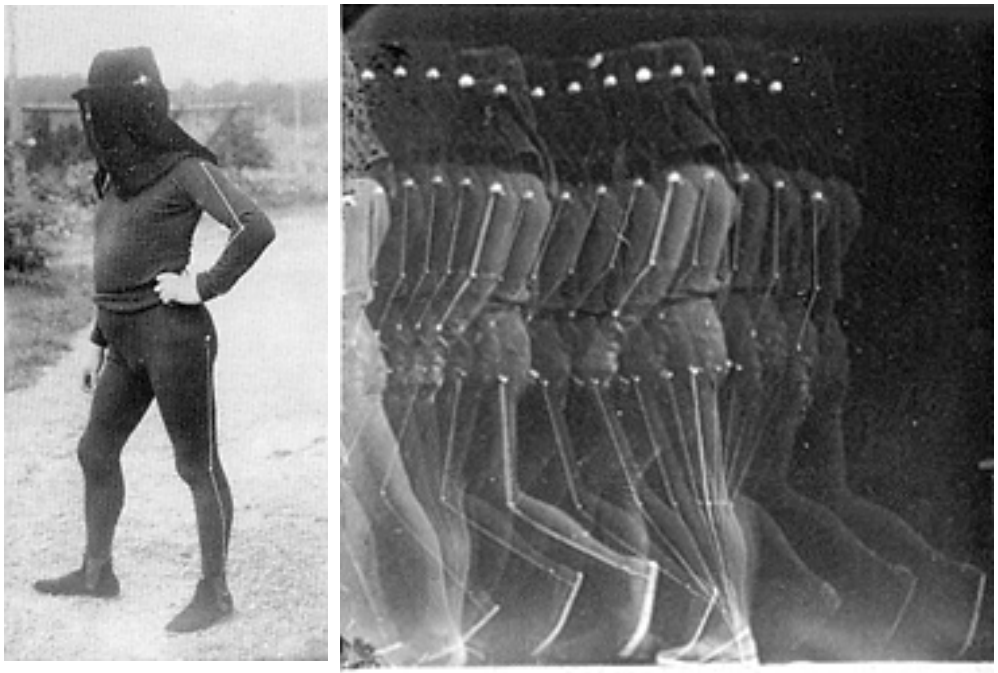
# Capturing Humans in Motion



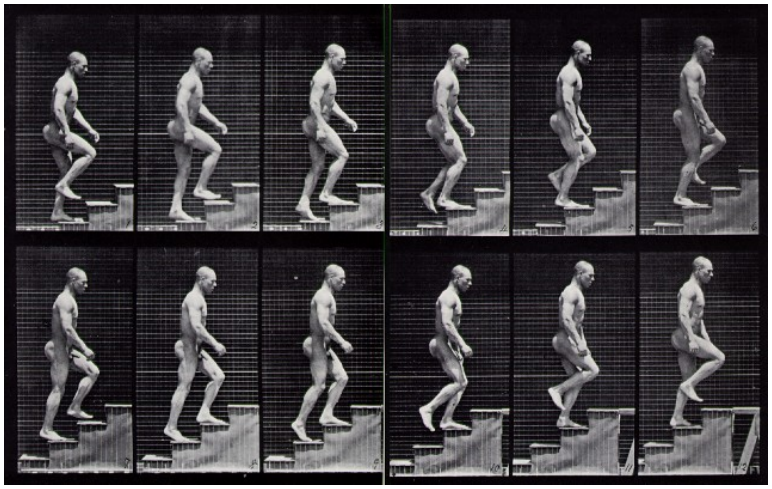
ETIENNE-JULES MAREY, 1882 chronophotograph.

# Capturing Humans in Motion

Loss of *depth* and *motion* in projection to 2D images.



ETIENNE-JULES MAREY, 1882.
Marker-based tracking

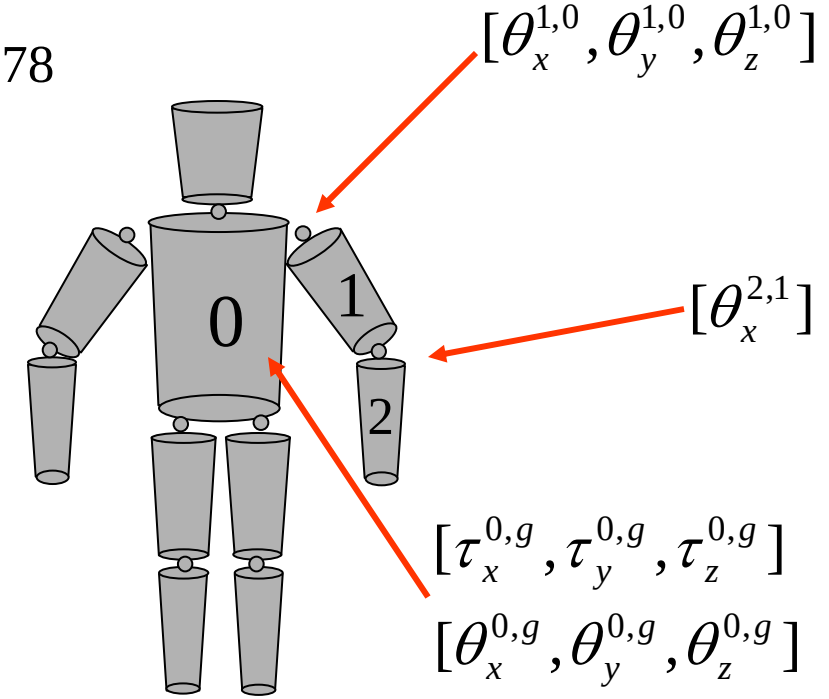


EADWEARD MUYBRIDGE, 1884-5.
Multiple cameras.

# Articulated Body Model

Kinematic tree:

Marr&Nishihara ‘78

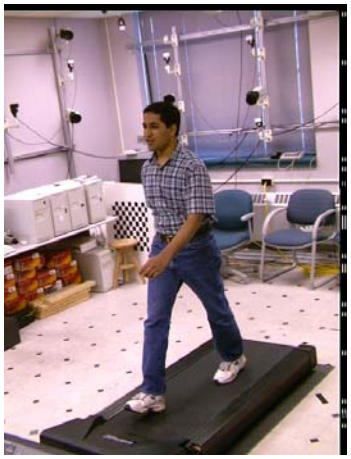


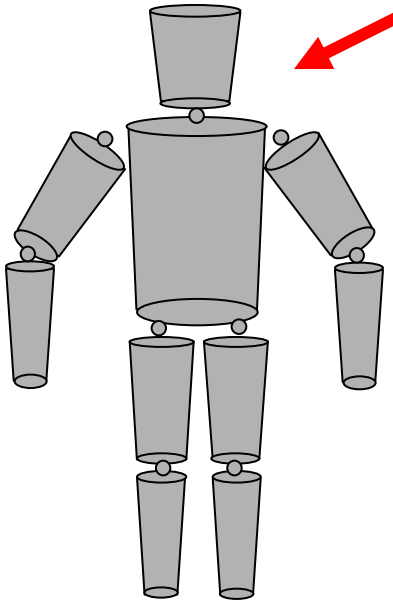

> 30D space

Represent a “pose” at time *t* by a vector of these parameters: $\mathbf{X}_t$

# Motion Capture

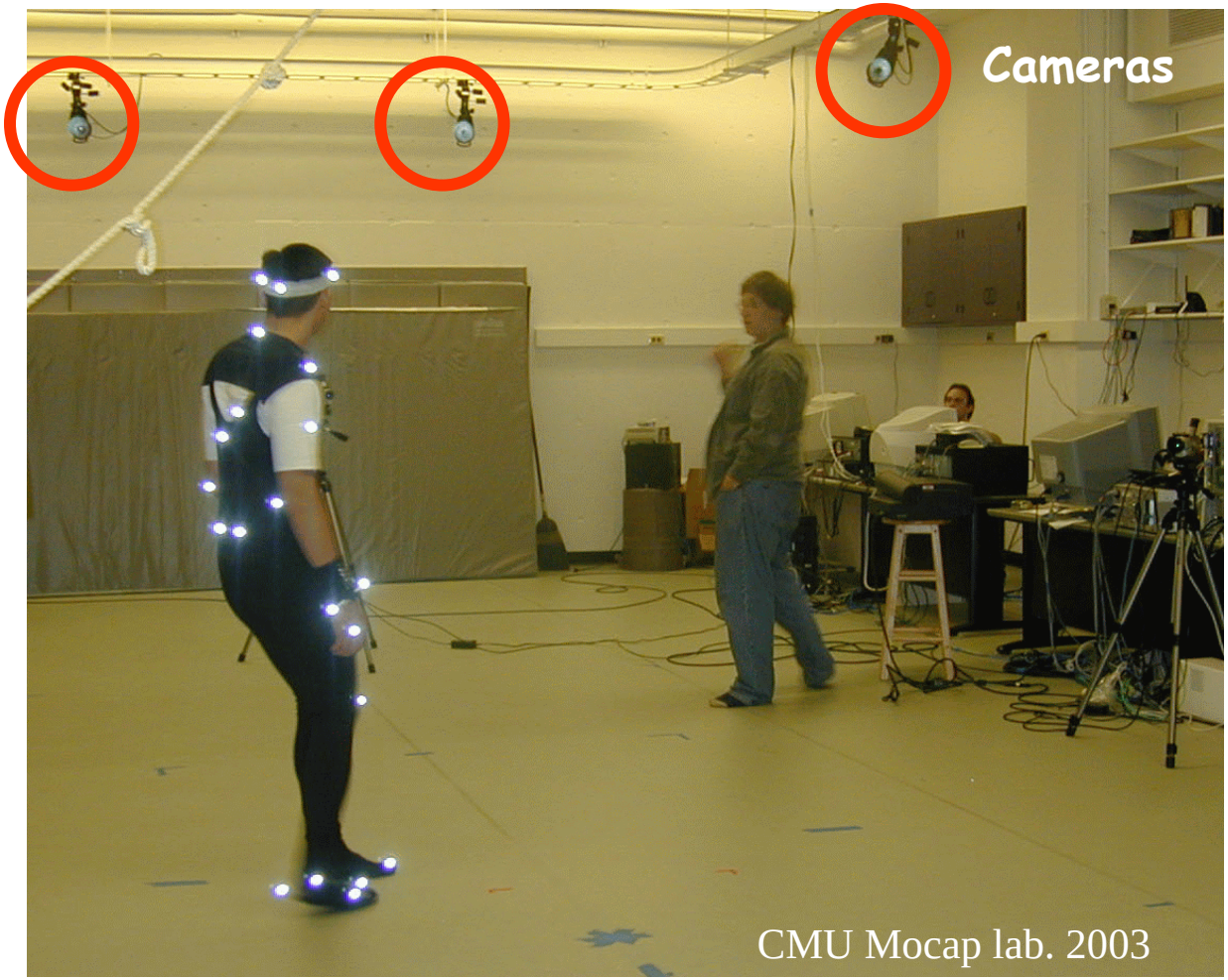
From images to "models" that support reasoning.



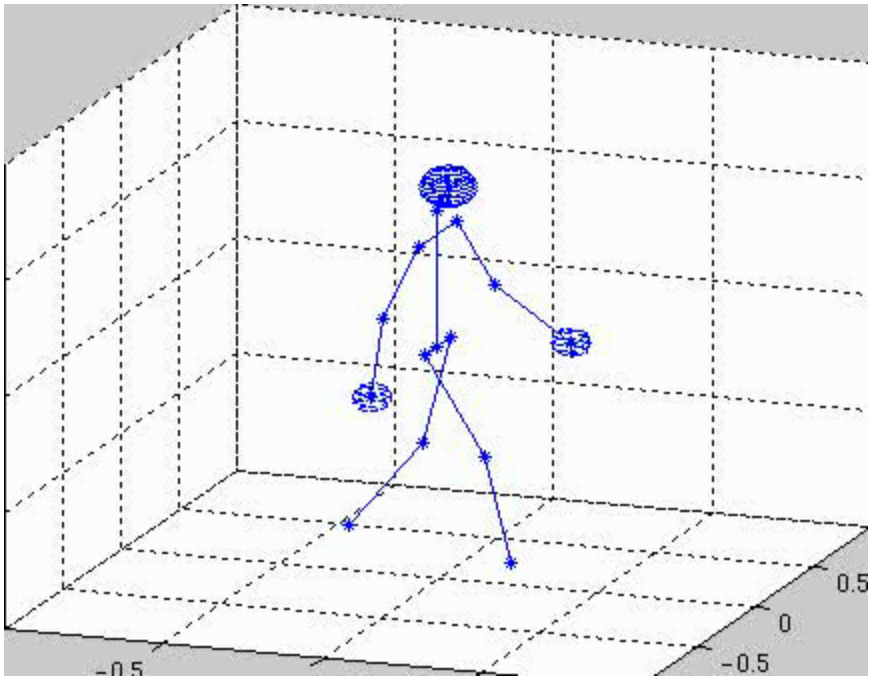


Recover 3D pose and motion.

# “Mocap” Today



# “Mocap” Today




Walking motion learned from mocap.

# “Mocap” Today

# Mocap in the Wild

Humans in captivity

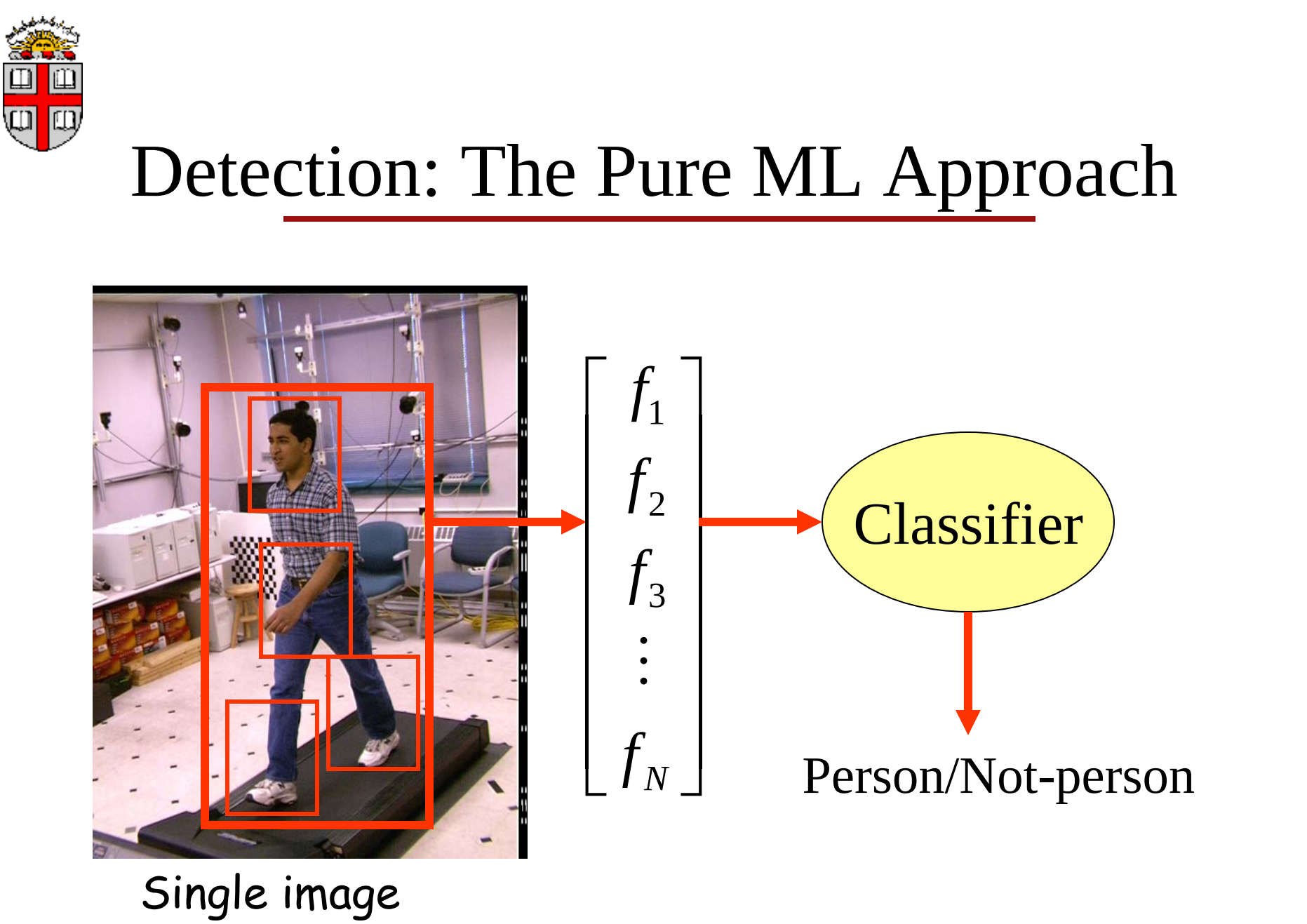
Humans in their natural habitat

# Detecting and Tracking People

* Where are the people?
* What are their poses?
* How are they moving?
* What are they doing?

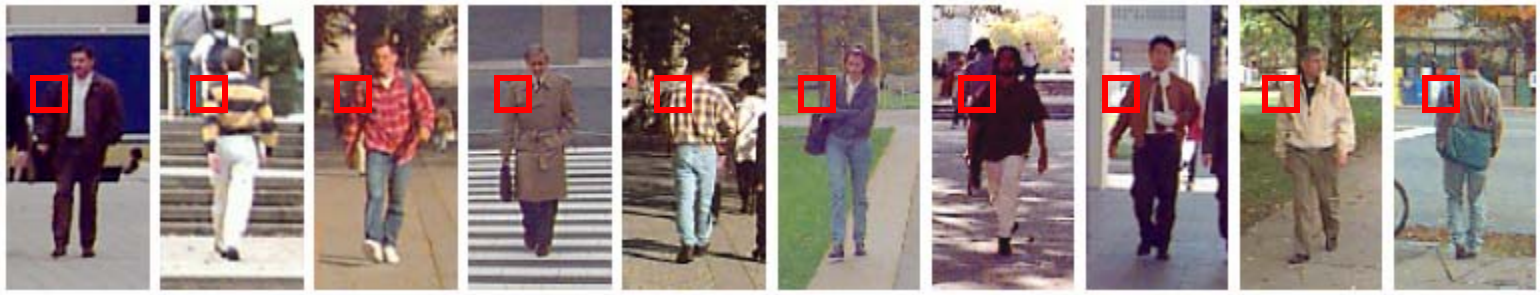
# Markerless Mocap

* Mocap
  - animation, film, games, archival footage
  - sports and rehabilitation medicine
* Tracking
  - surveillance
* Understanding
  - HCI/gesture recognition (cars, elder care, games, …)
  - video search/annotation



# Detection: The Pure ML Approach

# Support Vector Machines



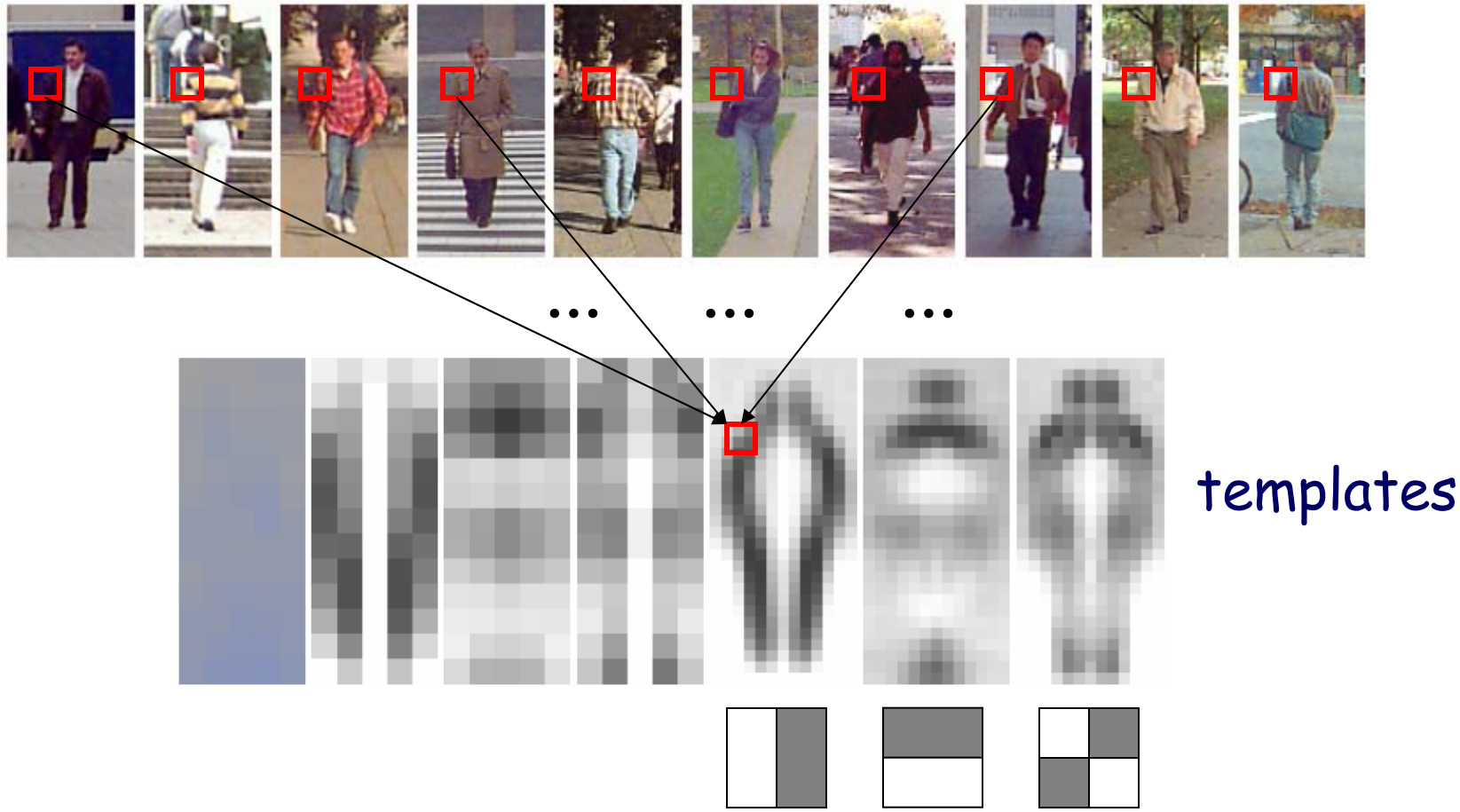
Multiply the pixel values in the region by this “mask” or “filter”:

| | |
|---|---|
| 1 | -1 |

Average the resulting absolute responses.

“Pedestrian detection using wavelet templates,” Oren *et al* CVPR’97.

# Support Vector Machines

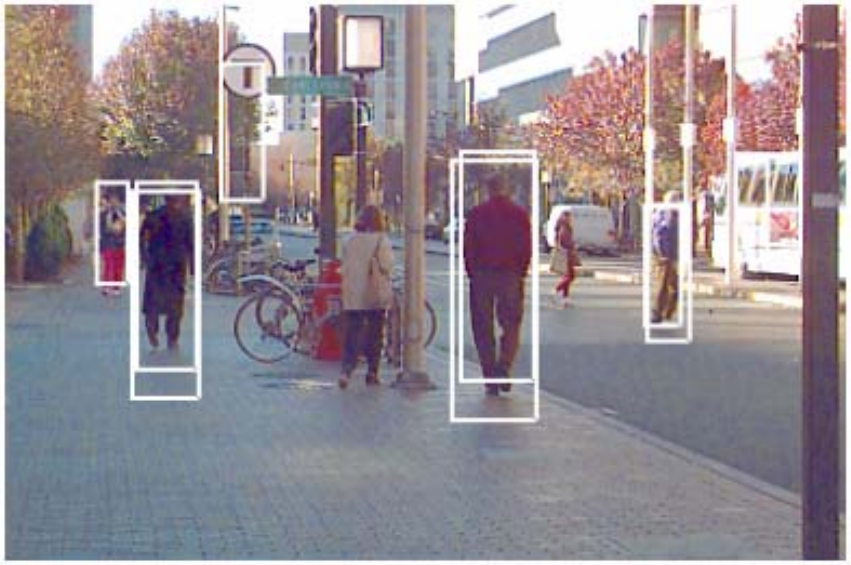



"Pedestrian detection using wavelet templates," Oren *et al* CVPR'97.

# Support Vector Machines

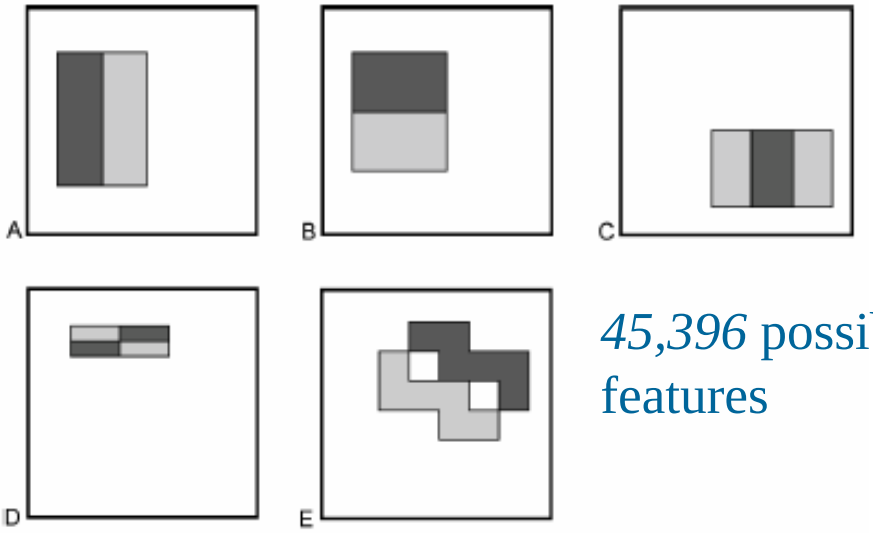
Product of wavelet templates and filtered image regions gives a vector of responses for each region.

Bootstrapped SVM learns the classify pedestrian/background.

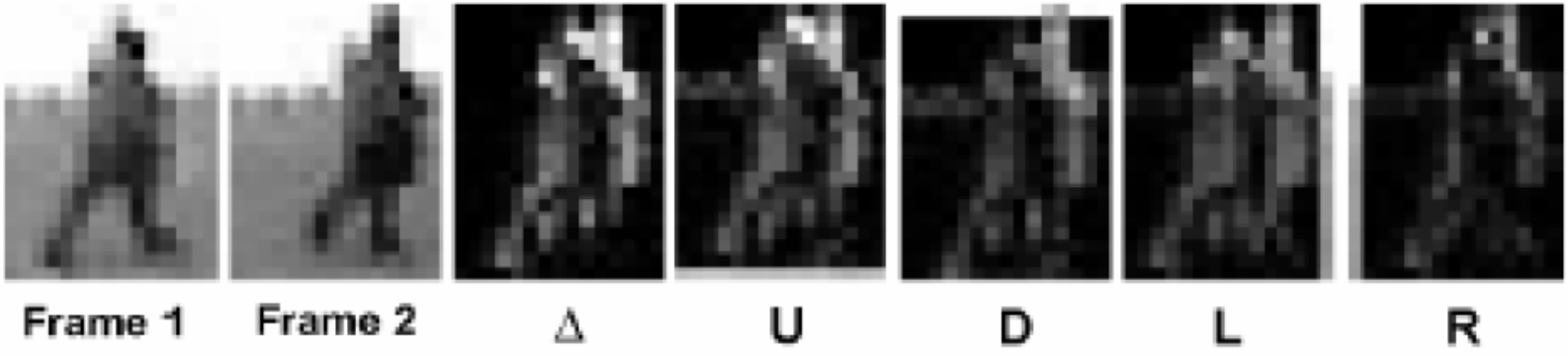


"Pedestrian detection using wavelet templates," Oren *et al* CVPR'97.

# AdaBoost




*45,396* possible features

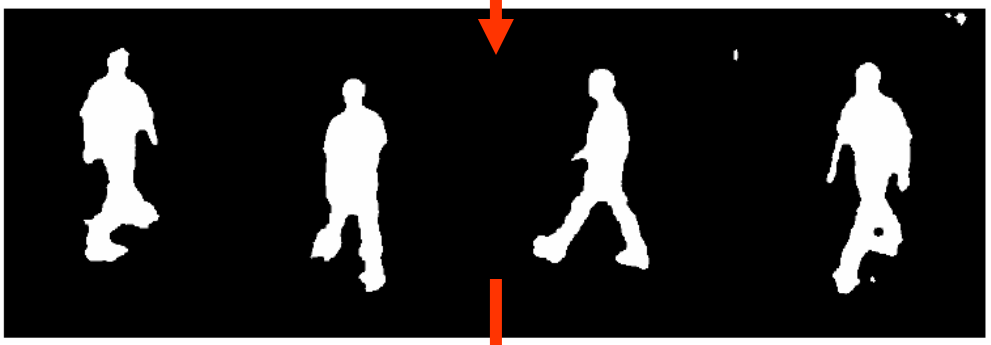



Viola, Jones and Snow, ICCV'03

# Pedestrian Detection

Viola, Jones and Snow, ICCV'03

# What about 3D Pose?



Contour Points / Shape Model

*Learned mapping*

30+ dimensions


K. Grauman, G. Shakhnarovich, T. Darrell, ICCV'03

# Single View to 3D Pose

Given synthetic training data, learn the mapping from silhouette contours to 3D pose.

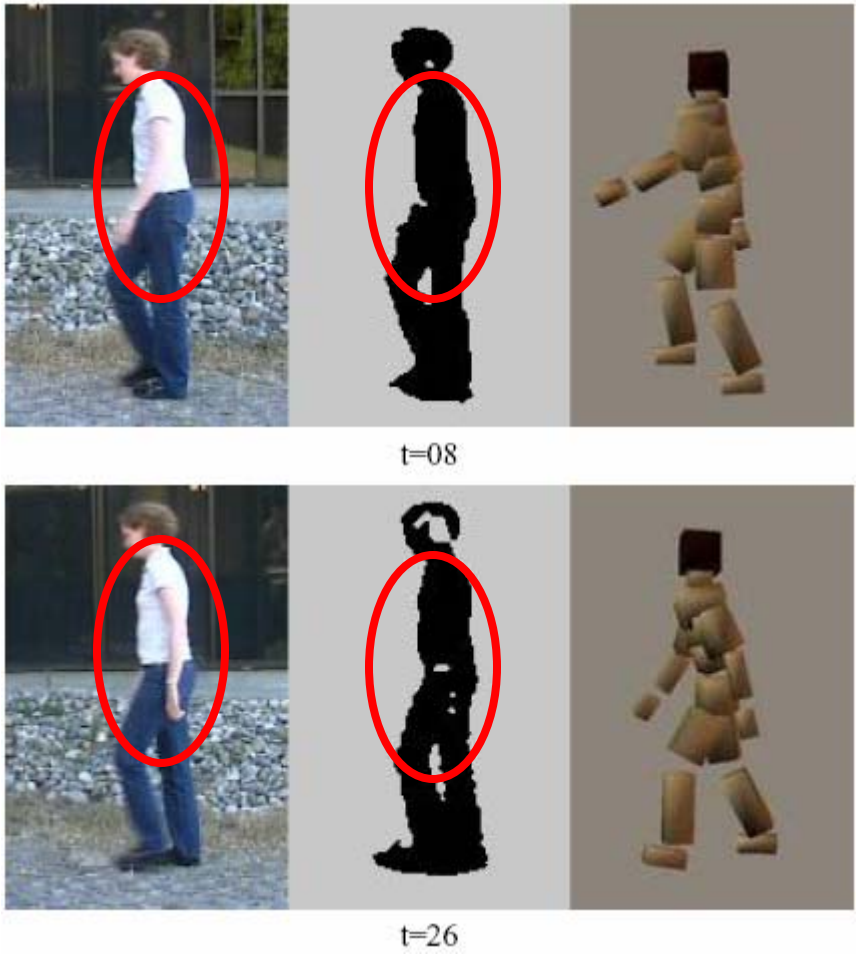
*"Gaussian kernel RVM"*, Agarwal and Triggs CVPR04

*"Fast Pose Estimation with Parameter Sensitive Hashing"*, Shakhnarovich, G., Viola, P., & Darrell, T. CVPR'03.

# Monocular Silhouettes

* lose internal structure.

* difficult to recover with complex moving backgrounds.




Agarwal & Triggs, ICML'04

# Pure ML – State of the Art

* Detection of people
  - Directly from image measurements
  - Canonical poses
  - Tiny people – i.e. no detailed 3D pose
* Recovery of pose
  - Requires good quality, isolated, silhouettes
  - Accurate pose requires multiple views

# Problems

The appearance/size/shape of people can vary dramatically (high-D space).

Underlying structure (bones and joints) is *unobservable* (obscured by muscle, skin, clothing).

Occlusion and partial views.

# Problems

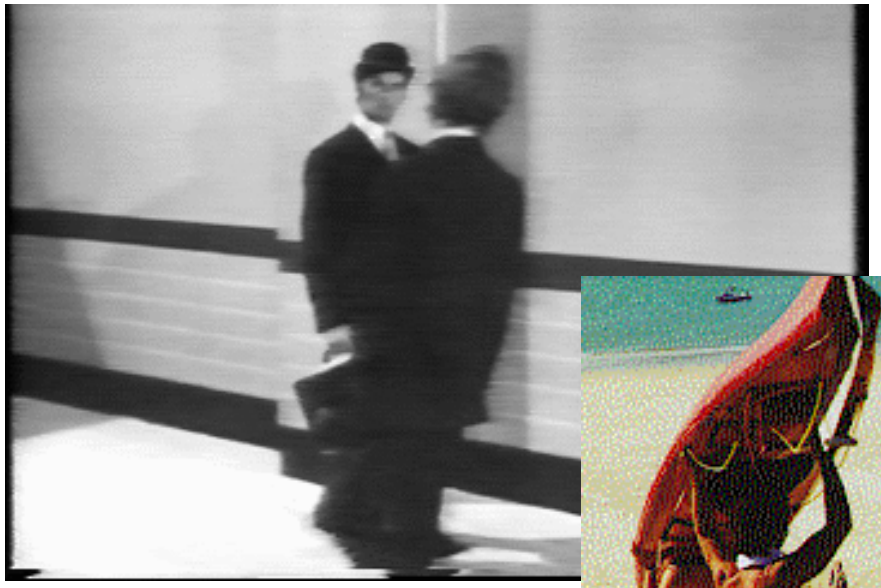
Loss of 3D in 2D projection

Unusual poses

Self occlusion

Low contrast

# Problems



Multiple people and occlusion leads to ambiguity.

Moving cameras & complex changing backgrounds.

# Problems

Accidental alignment

Motion blur.

(nothing to match)

# Requirements

1. Represent uncertainty and multiple hypotheses.
2. Model complex kinematics of the body.

   Correlations between joints and over time.
3. Exploit multiple image cues in a *robust* fashion without relying on background subtraction.
4. Integrate information over time.

*The recovery of human pose/motion is fundamentally a problem of inference from ambiguous and uncertain measurements.*

# Bayesian Approach

$$p(\text{model} \mid \text{cues}) = \frac{p(\text{cues} \mid \text{model})\, p(\text{model})}{p(\text{cues})}$$

1. **Model**: Kinematic tree.
2. **Likelihood**:
   combine various cues – *learn* from examples.
3. **Prior**: statistical model, *learned* from mocap data.
4. **Search**: discretize intelligently using factored sampling and search using a *particle filter*.

# Makerless Motion Capture

<movie>

# Problems

* Search in a huge-dimensional space: 30+ dimensions.

  -Non-Gaussian, multi-modal, posterior.

* Multiple cameras for reliable tracking.
* **Manual initialization**.
* Still relying on good silhouettes.
* Brittle – can't recover when it gets lost.

# So Where Are We?

* Learning mapping from raw images to models is asking too much.
* Matching a complex model directly to the image is asking too much.

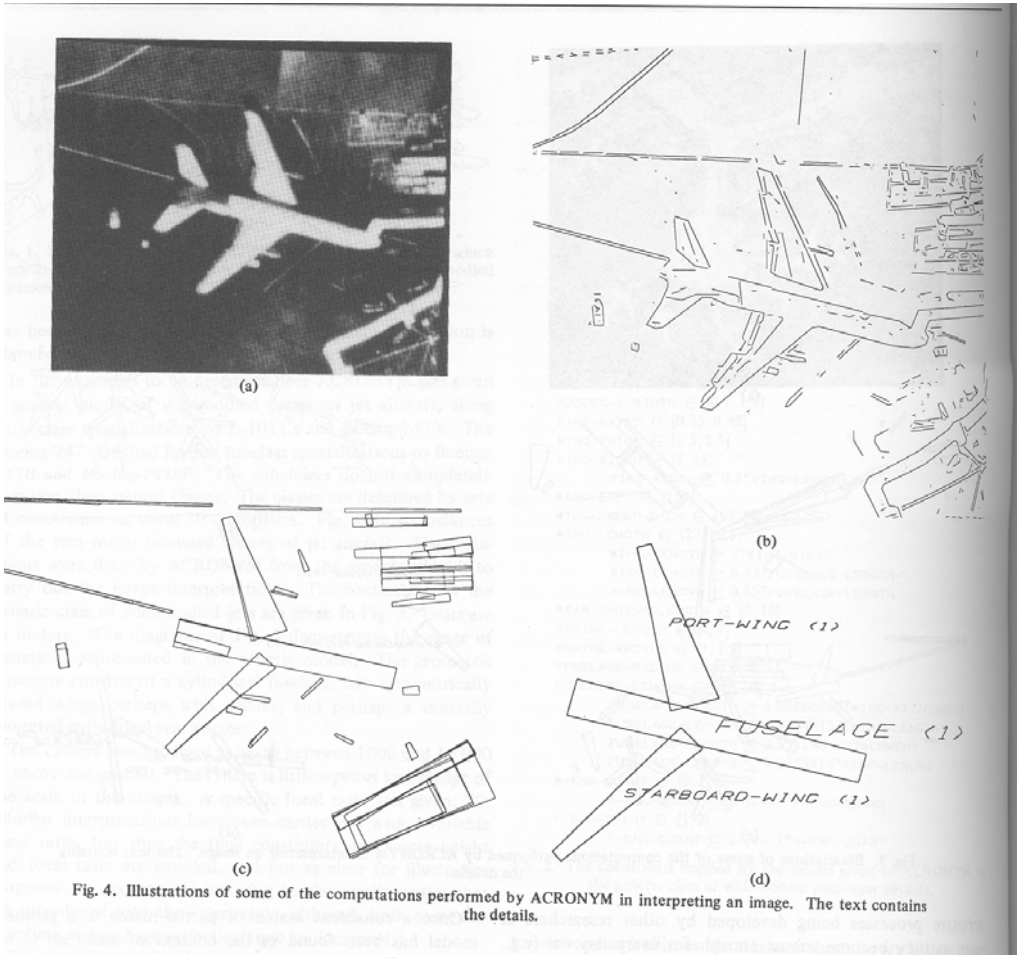



Fig. 4. Illustrations of some of the computations performed by ACRONYM in interpreting an image. The text contains the details.

# Hierarchy of Representations

Brooks, ACRONYM, 1981.

# Approach

Stuff: Rich set of filters applied to images

Shouters: Simple feature (part) detectors

* faces, limbs, feet, hands…

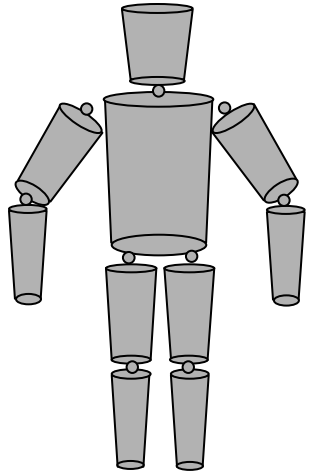
Glue: Loose-limbed model

* local constraints in space and time

Things: Articulated body model

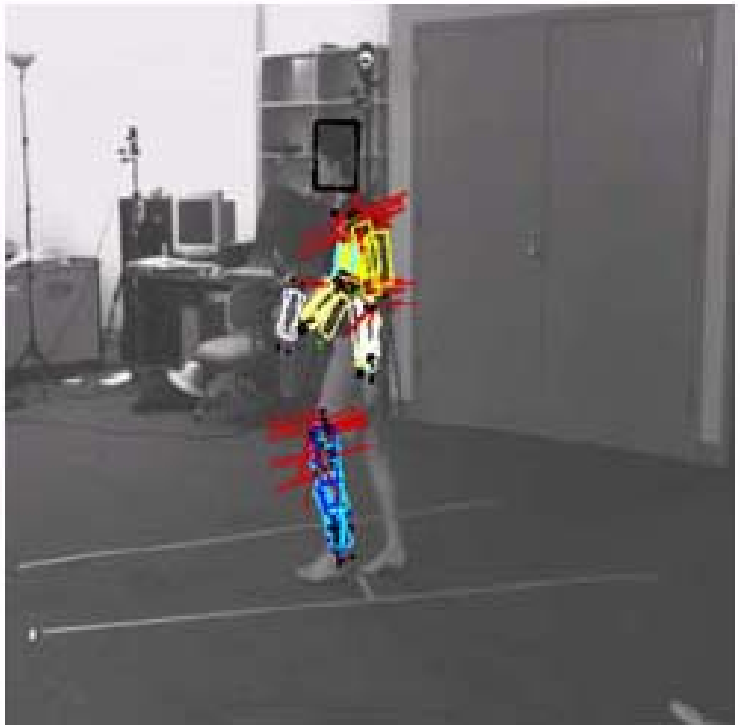
* expresses model knowledge about humans and how they move

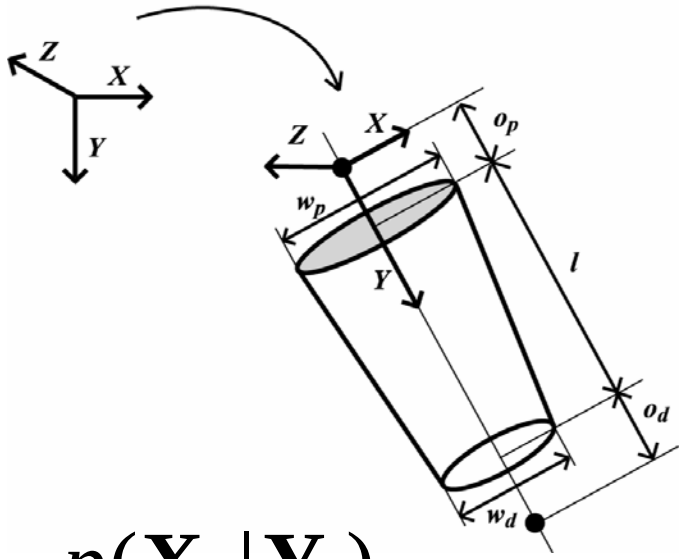
inference



*cf recent work in object recognition*

# “Shouters”

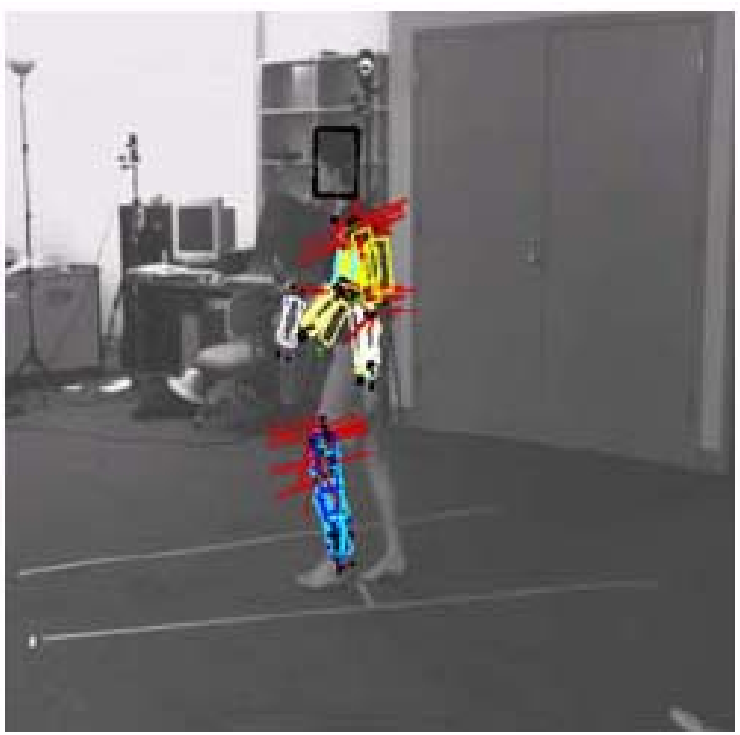


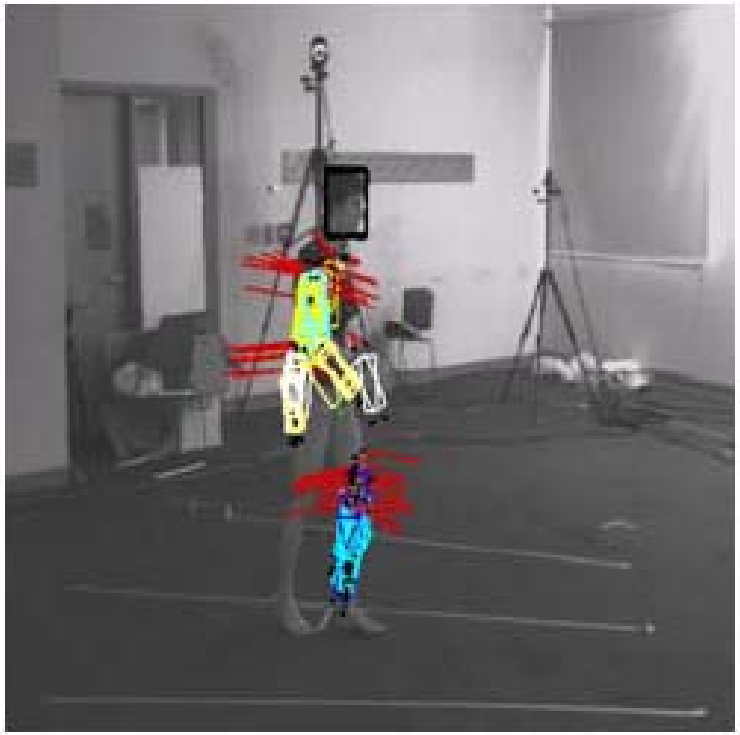



$$p(\mathbf{X}_i \mid \mathbf{Y}_t)$$

Finding a person is hard but finding plausible limbs is much easier.

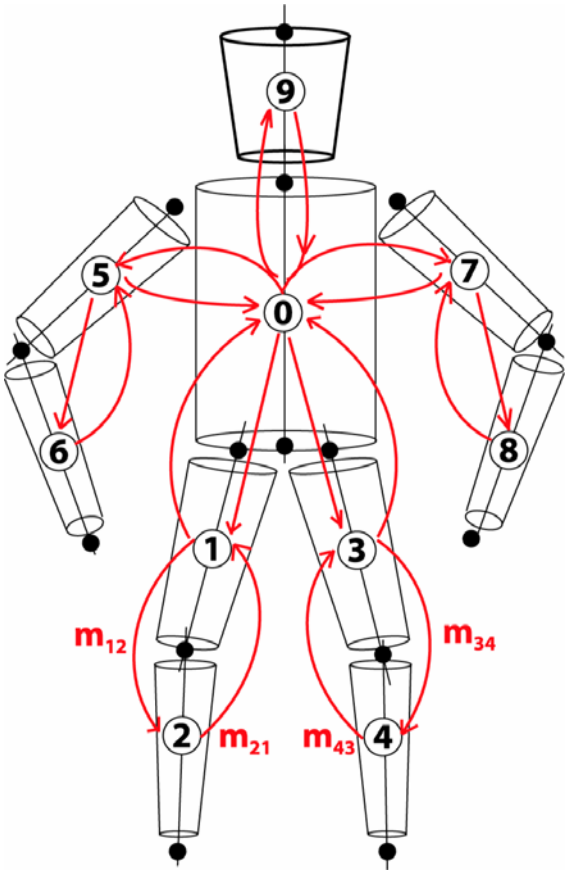
# “Shouters”





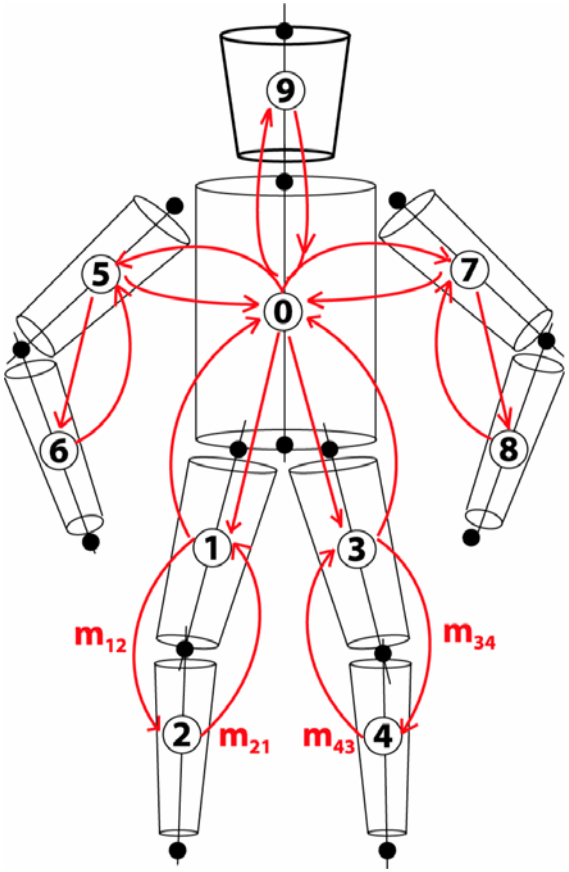
* Various ML techniques (eigen-X, SVM, RVM, Adaboost, etc).
* Relatively low dimensional problem.
* Accuracy not critical.

# Glue: Loose-limbed People




Loose-limbed body
(graphical model)

"Push puppet" toy




# Glue: Loose-limbed People

Soft constraints between limbs (messages).

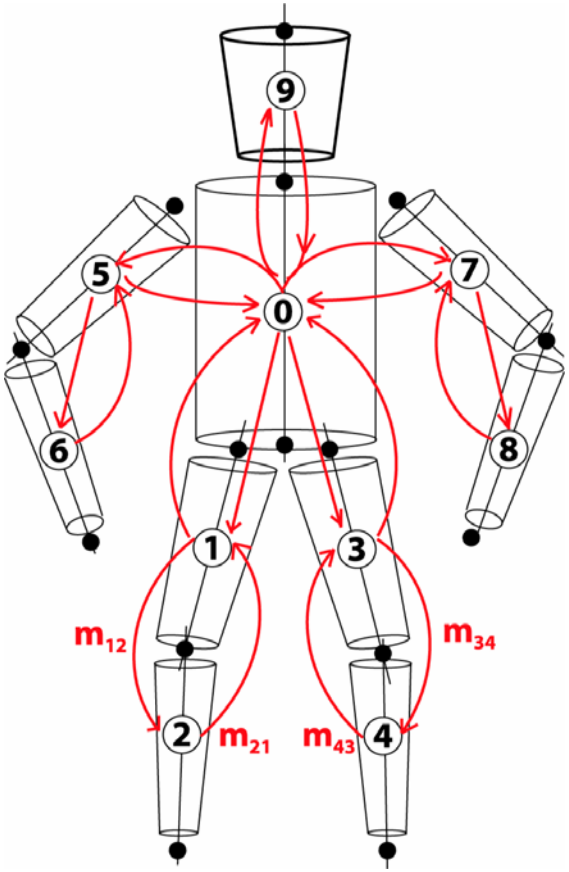
Pose estimation as inference in a graphical model (Belief Propagation).

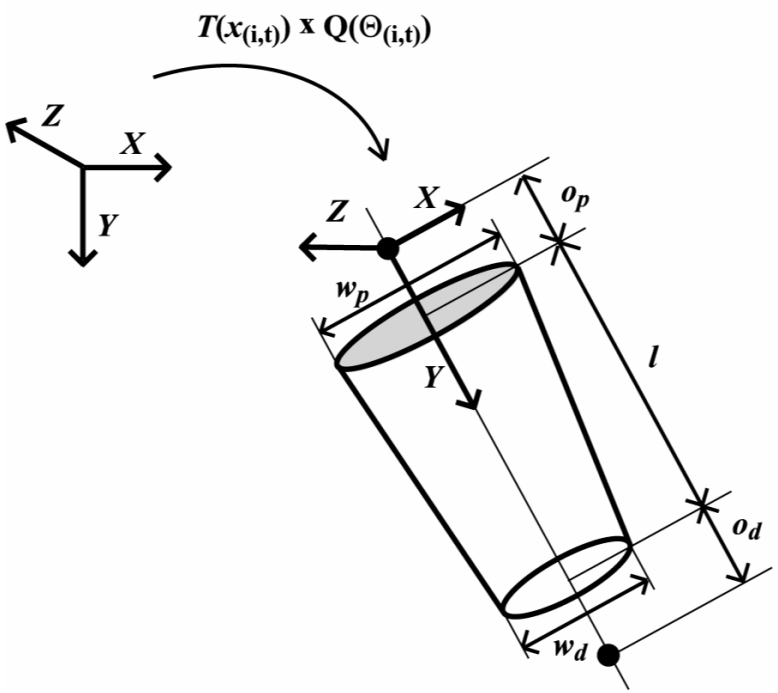
Allows bottom up initialization.

Deals well with unobserved limbs.

Pictorial structures – Fischler and Elschlager '73
More recently (2D, discretized):
* Felzenszwalb & Huttenlocher '00, Ronfard et al '02, Forsyth et al '00-03.

# Glue: Loose-limbed People

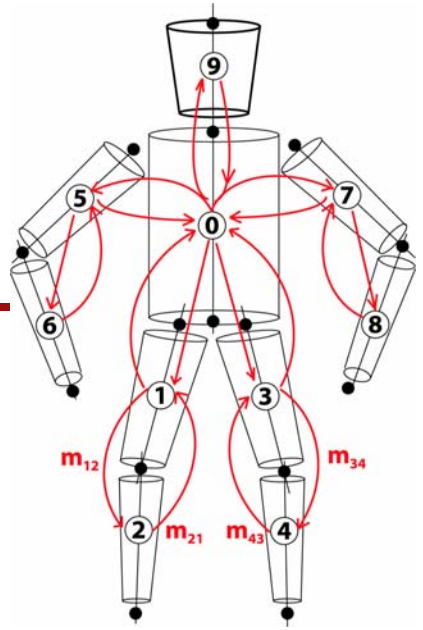



* $\mathbf{X}_i$ = position & orientation

* 6D – discretization not practical

# Inference: Belief




random vector (position and orientation of node *i*)

local evidence

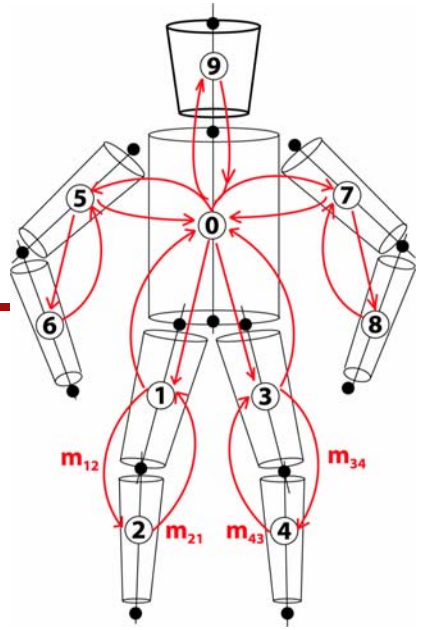
$$p(\mathbf{X}_i \mid \mathbf{Y}) = \alpha \lambda(\mathbf{Y}_i \mid \mathbf{X}_i) \prod_{k \in A_i} m_{ki}(\mathbf{X}_i)$$

observations (image and filter responses)

neighbors of node *i*

incoming messages

# Messages




message from node $i$ to node $j$

local evidence

$$m_{ij}(\mathbf{X}_j) = \alpha \int \psi_{ij}(\mathbf{X}_i, \mathbf{X}_j)\lambda(\mathbf{X}_i) \prod_{k \in A_i \setminus j} m_{ki}(\mathbf{X}_i)\, d\mathbf{X}_i$$

probability of $\mathbf{X}_j$ conditioned on $\mathbf{X}_i$

("spatial" or temporal prior)

neighbors of $i$, not including $j$

incoming messages at $i$

this is the *hard* part if things aren't Gaussian

# Recipe for Finding People

1. Learn local evidence
   * Model appearance.
2. Learn spatial and temporal priors
   * How joints connect and move.
3. Search
   * Develop an effective inference algorithm for message passing.
   * Exploit shouters as a proposal distribution.

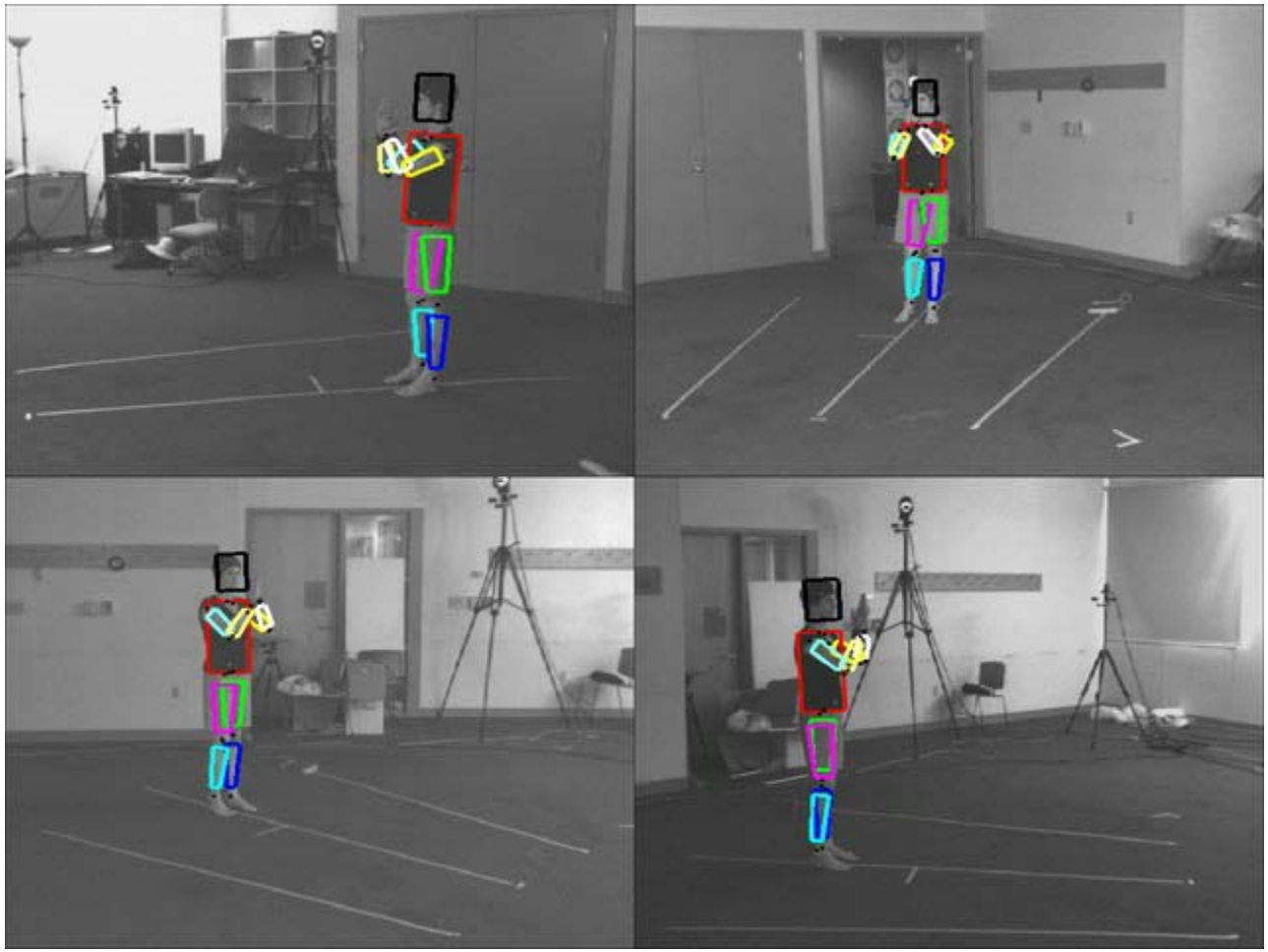
# Recipe for Finding People

1. Learn local evidence
   * Model appearance.

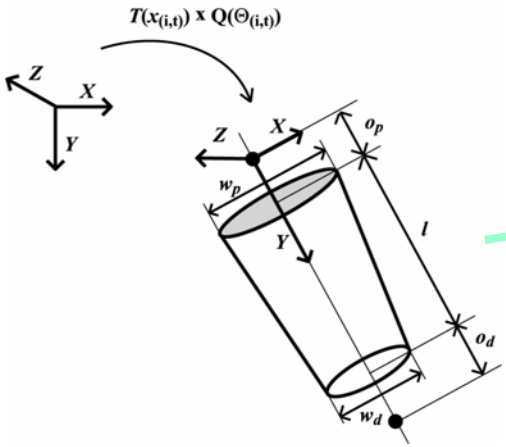
local evidence

$$p(\mathbf{X}_i \mid \mathbf{Y}) = \alpha \, \lambda(\mathbf{Y}_i \mid \mathbf{X}_i) \prod_{k \in A_i} m_{ki}(\mathbf{X}_i)$$

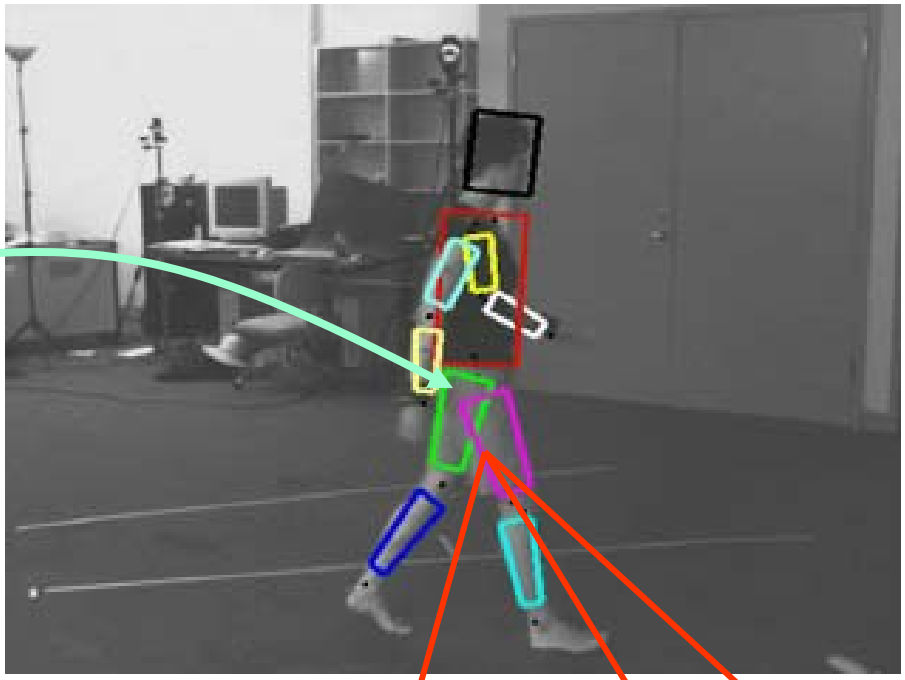
# Ground Truth (MoVid)



3D Vicon data projected onto video streams.

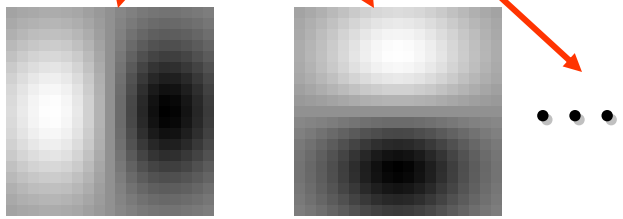



# Towards a Rigorous Likelihood

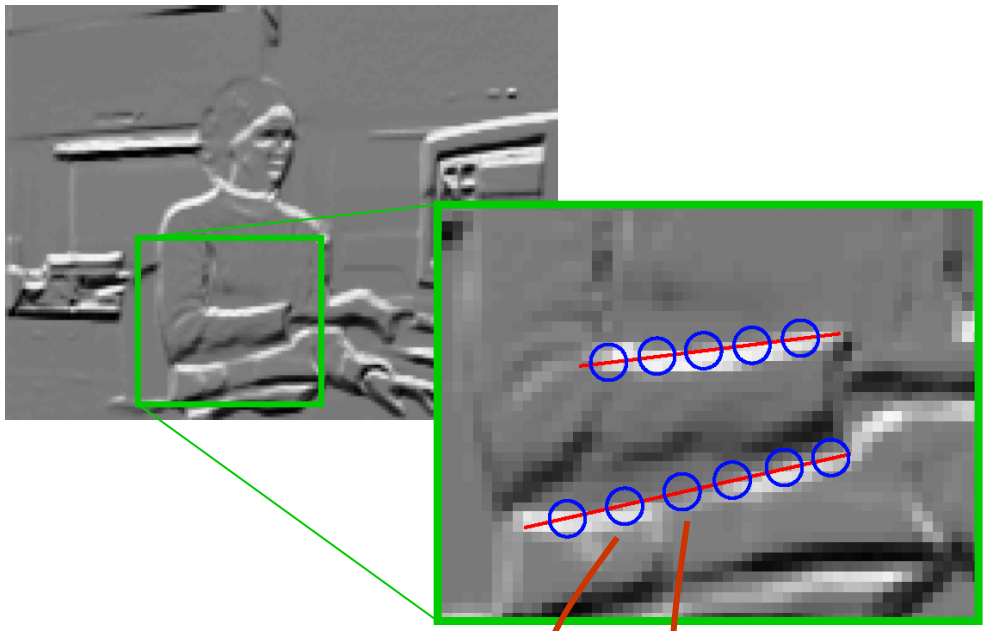


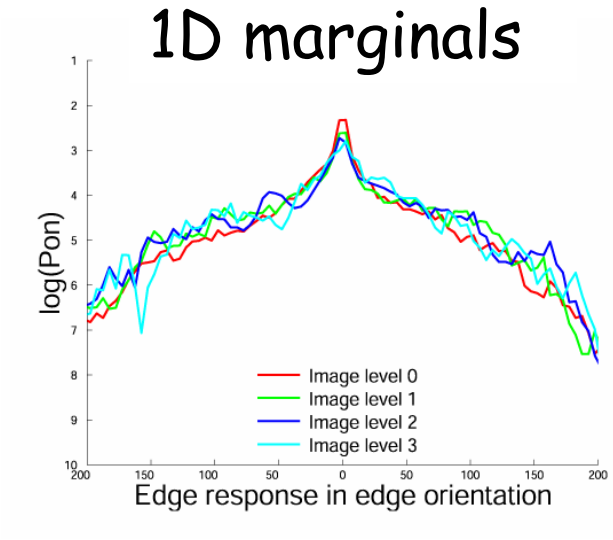
Learn from examples:

$$\lambda(\mathbf{Y} \mid \mathbf{X}_i) = p(f_1, f_2, \ldots, f_n \mid \mathbf{X}_i)$$



# Naïve Likelihood






$$\lambda(\mathbf{Y} \mid \mathbf{X}_j) = p(f_1, \ldots, f_k \mid \mathbf{X}_j) = \prod_i p(f_i \mid \mathbf{X}_j)$$

# Problems

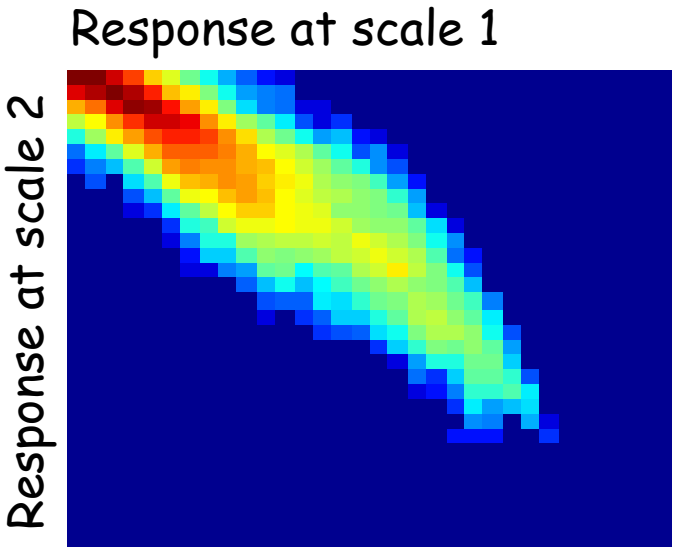
* Filter responses are not conditionally independent.

* Naïve model results in "peaked" likelihoods – makes tracking brittle.

* High dimensional likelihood.

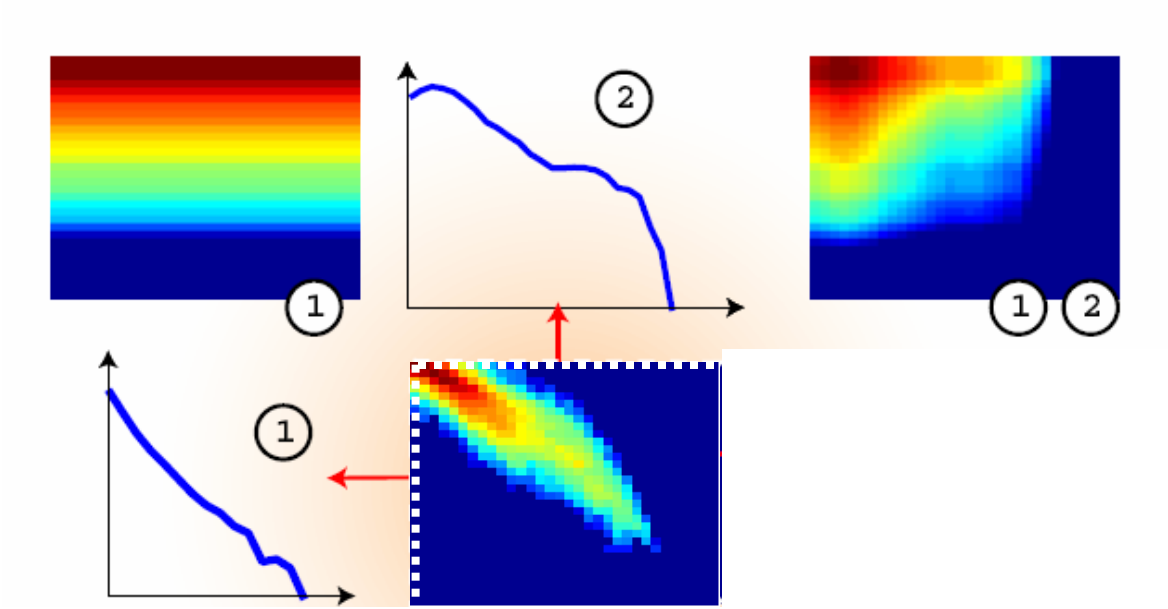
* Non-Gaussian (image statistics).

* Limited training data.

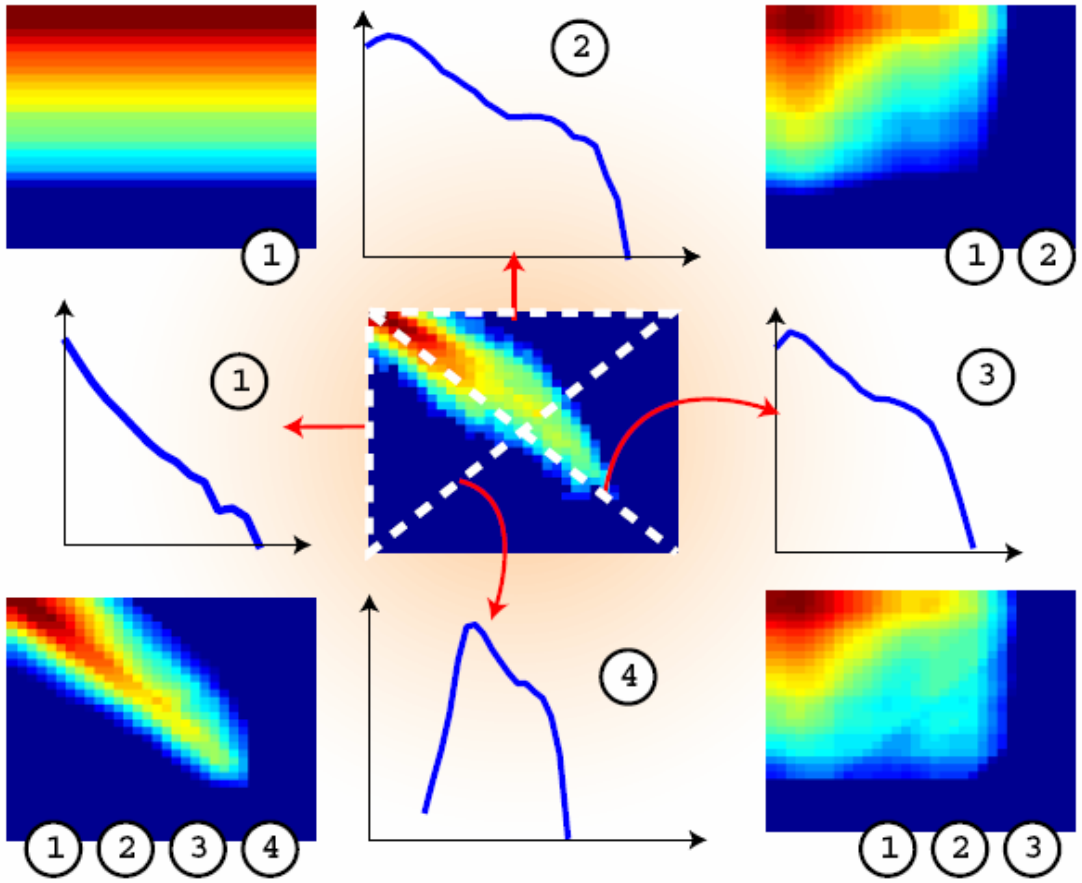



Observation: marginals are easy to learn

# Maximum Entropy Learning



# Maximum Entropy Learning



# Maximum Entropy Learning

Learn a Gibbs model:

$$p(\mathbf{Y} \mid \mathbf{X}_j) = \frac{1}{Z} \exp\left( -\sum_i < \omega^{(i)}, \phi^{(i)}(\mathbf{Y}, \mathbf{X}_j) > \right)$$

* the marginals match those of the training data.

* entropy is maximized.

Zhu & Mumford, PAMI '96

# Recipe for Finding People

2. Learn spatial and temporal priors

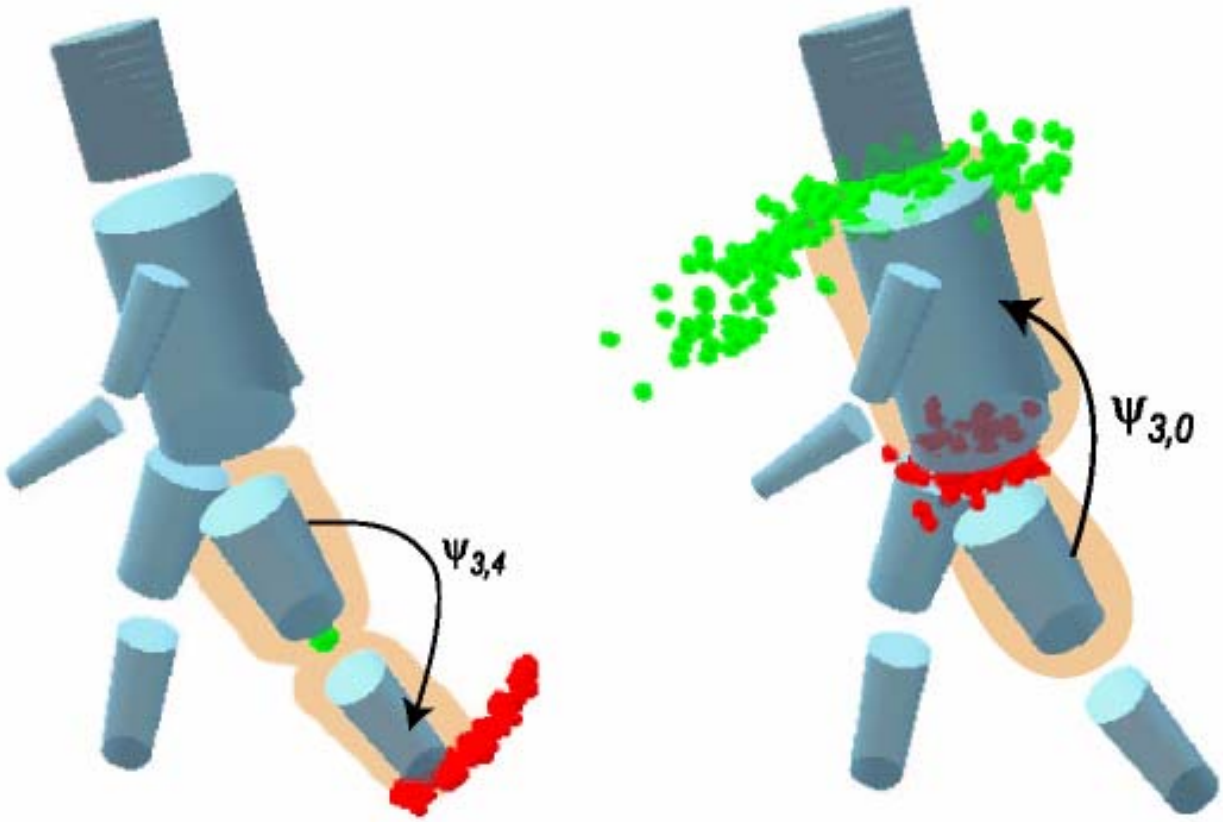
   * How joints connect and move.

$$m_{ij}(\mathbf{X}_j) = \alpha \int \psi_{ij}(\mathbf{X}_i, \mathbf{X}_j) \lambda(\mathbf{X}_i) \prod_{k \in A_i \setminus j} m_{ki}(\mathbf{X}_i)\, d\mathbf{X}_i$$

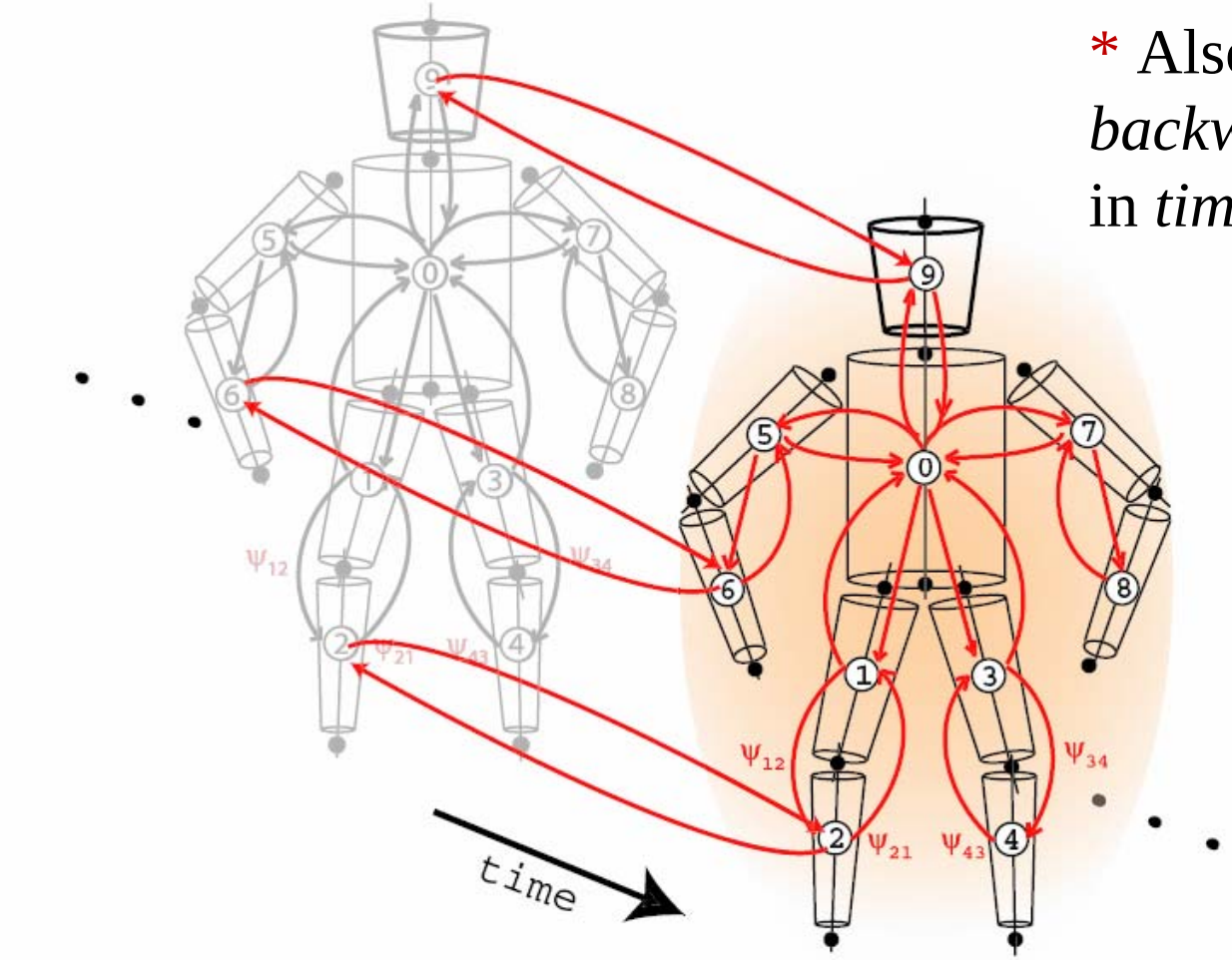
probability of $\mathbf{X}_j$ conditioned on $\mathbf{X}_i$

("spatial" or temporal prior)

# Learned Conditionals




* represented by a mixture of Gaussians (learned from mocap data).

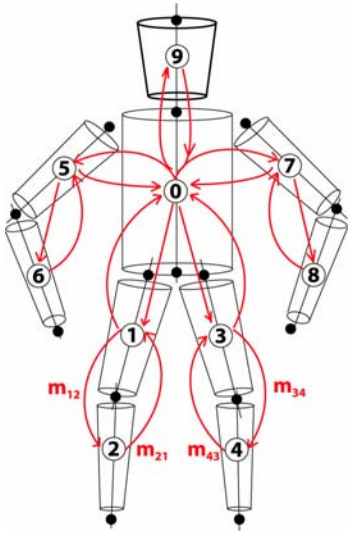



# Spatial and Temporal

* Also learn conditionals *backwards* and *forwards* in *time.*

* Introduces loops

# Recipe for Finding People




## 3. Search

* Develop an effective inference algorithm for message passing.

Non-Gaussian

6D continuous

$$m_{ij}(\mathbf{X}_j) = \alpha \int \psi_{ij}(\mathbf{X}_i, \mathbf{X}_j) \lambda(\mathbf{X}_i) \prod_{k \in A_i \setminus j} m_{ki}(\mathbf{X}_i)\, d\mathbf{X}_i$$

Product of mixtures

* Gibbs sampler

Mixture of Gausians

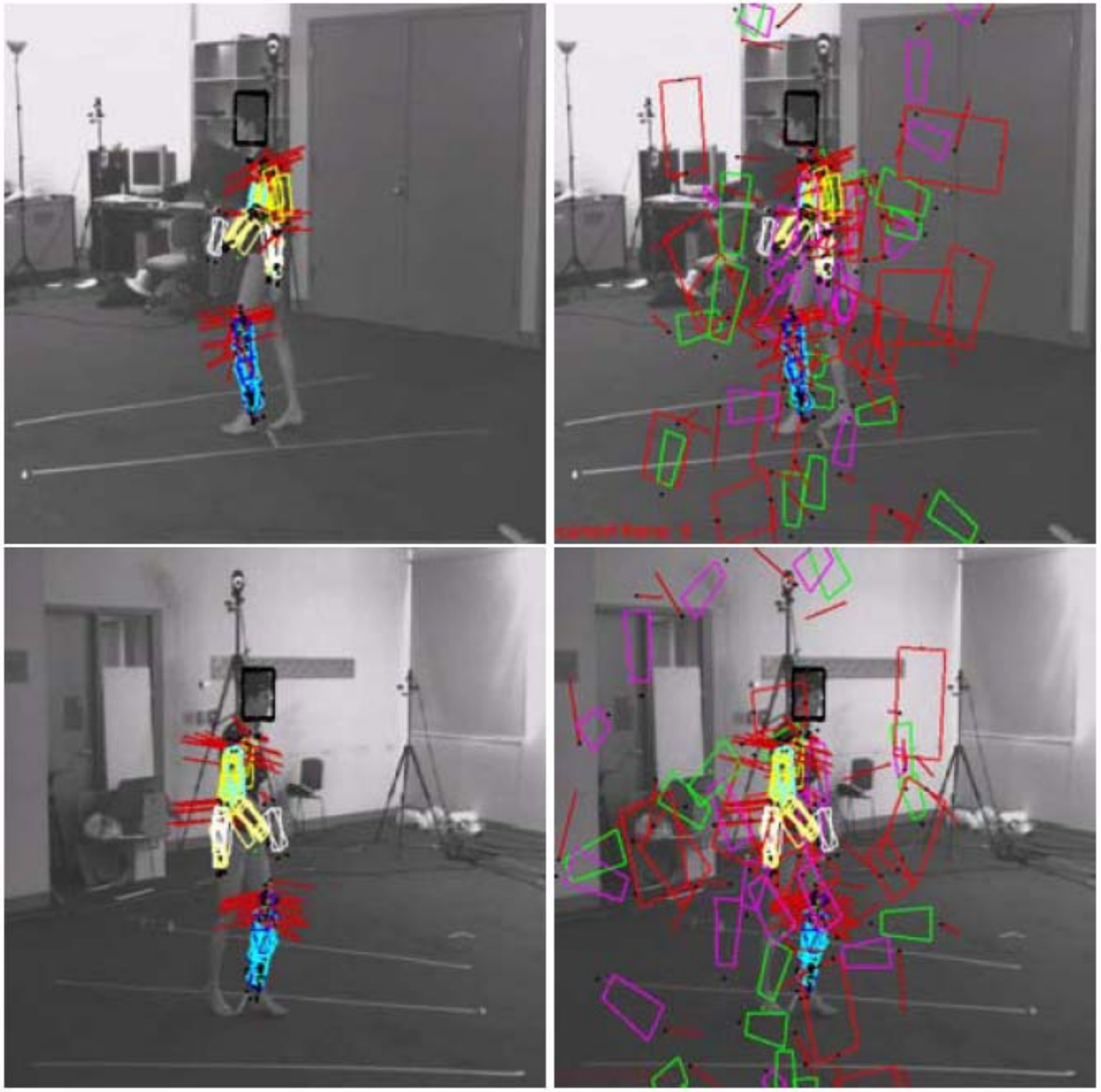
# Algorithm Highlights

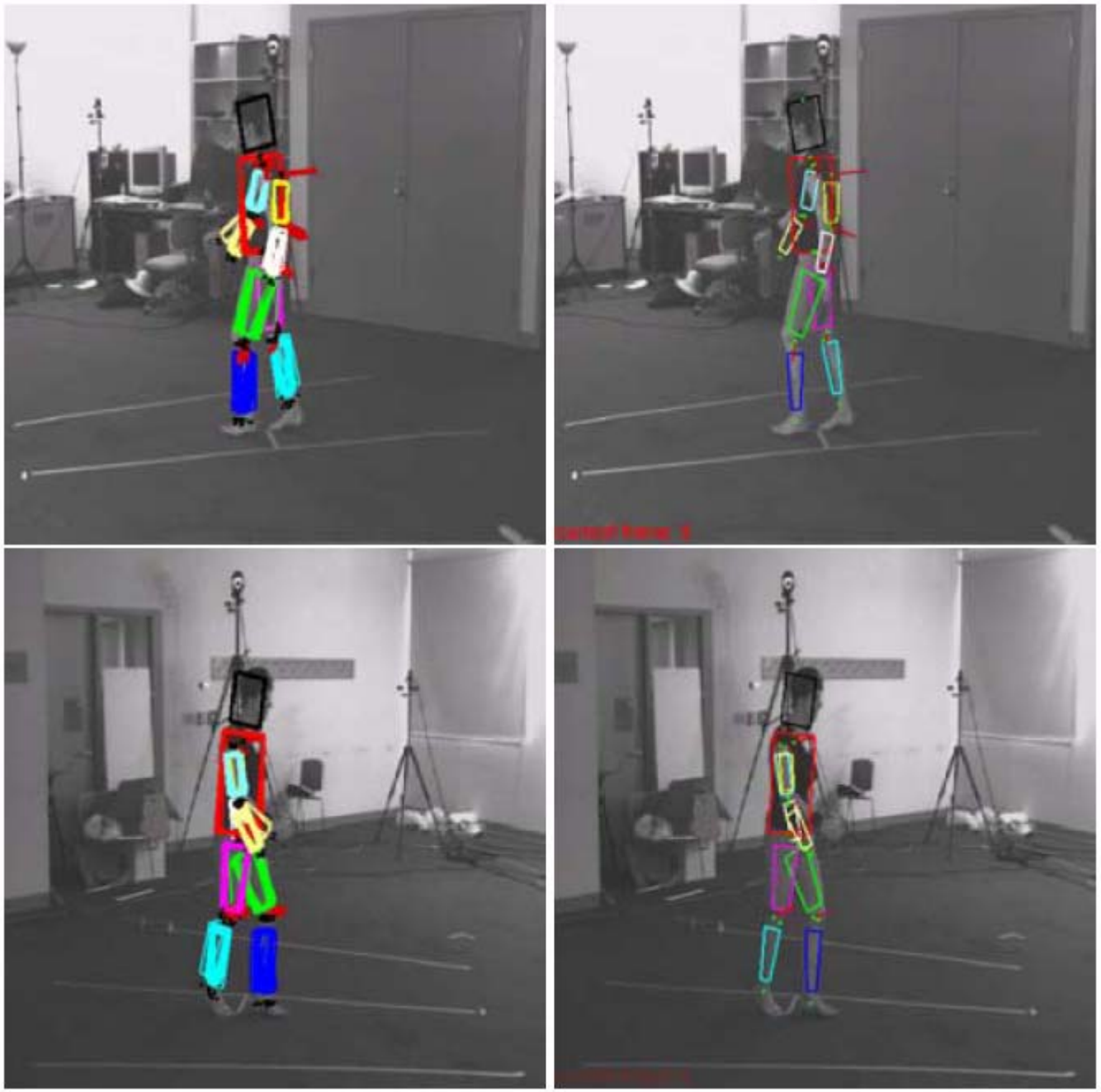
$$m_{ij}(\mathbf{X}_j) = \alpha \int \psi_{ij}(\mathbf{X}_i, \mathbf{X}_j) \lambda(\mathbf{X}_i) \prod_{k \in A_i \setminus j} m_{ki}(\mathbf{X}_i) \, d\mathbf{X}_i$$

Non-parametric Belief Propagation:

(Isard CVPR '03, Sudderth et al CVPR '03)

* Gibbs sample from message product (Sudderth et al)
* Monte Carlo integration (like particle filtering)
* importance sample from a proposal distribution
    * including bottom-up shouters
* propagate through potential function





# Summary

We have tackled four important parts of the problem:

likelihood

1. Probabilistically modeling human appearance.

prior

2. Learn relationships between limbs from training data – weak model.

search

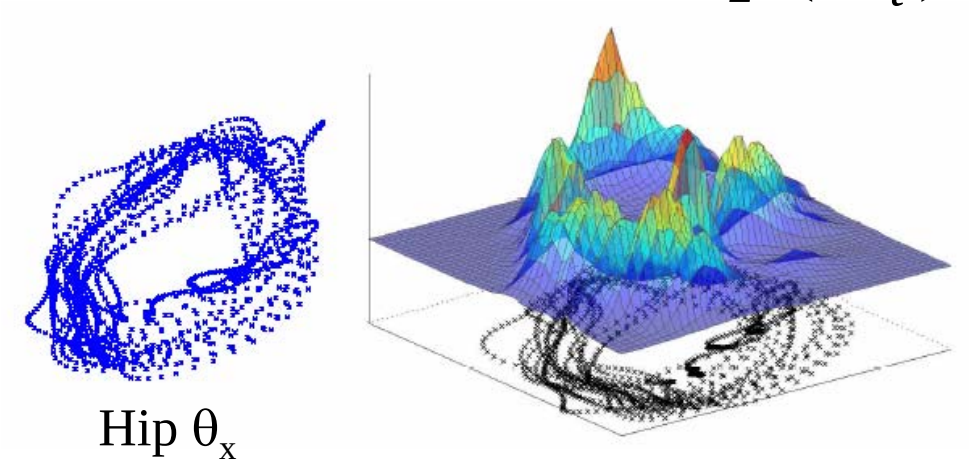
3. Bayesian inference using non-parametric belief propagation.

These general tools are widely applicable.

# Challenge for ML

Tools for modeling high dimensional joint densities from limited data

- Prior models of human motion $p(\mathbf{X}_t)$




- Non-Gaussian likelihoods $p(\mathbf{Y}_t \mid \mathbf{X}_t)$

- Data available – just contact me.

# Outlook

5 years:

- Accurate mocap with multiple cameras.
- Automatic initialization.

10 years:

- Reliable monocular tracking in complex scenes.
- Pedestrian detection in cars.
- Surveillance and tracking in public spaces.

10+ years:

- Analysis and understanding of human action.

# Collaborators

**Graduate Students at Brown:**

**Leonid Sigal,** *Computer Science*
**Stefan Roth,** *Computer Science*
**Alex Balan,** *Computer Science*
**Sidharth Bhatia,** *Engineering*

**Undergrads***:*

**Jonathan Bankard,** & **David Erickson** – *Mocap lab*

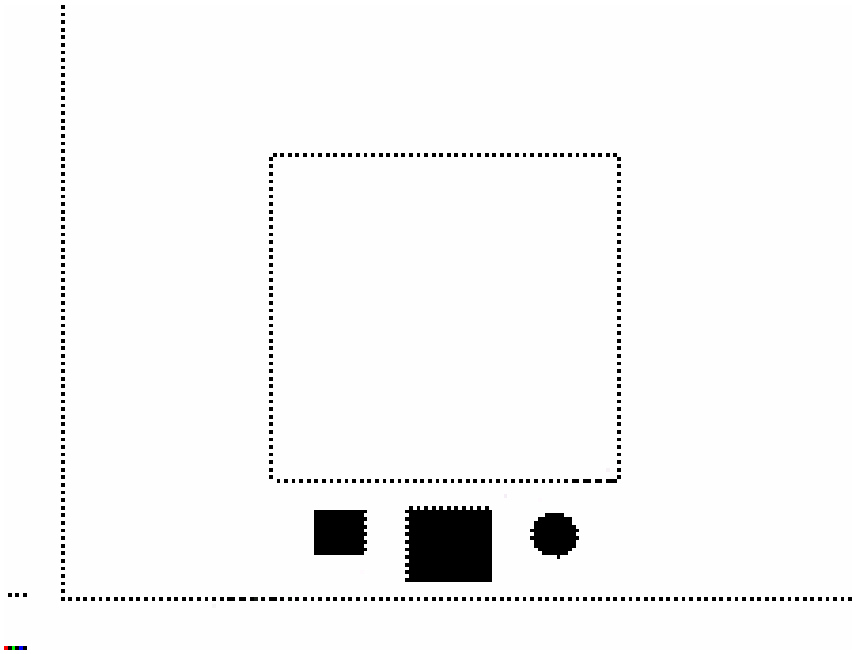
**Michael Isard,** Microsoft Research

**Alumni:**

**Hedvig Sidenbladh**, *Swedish Defense Research Inst.*
**David Fleet**, *University of Toronto*
**Ben Sigelman**, *Brown University (now Google).*

# What is still far off?

Motion interpretation.



Heider&Simmel, 1944

* Here *estimation* problem is trivial but *explanation* is hard.

Movie by Emre Yilmaz