The Science of Silly Walks

Hedvig Sidenbladh

Royal Inst. of Technology, KTH
Stockholm Sweden

http://www.nada.kth.se/~hedvig

Michael J. Black

Department of Computer Science
Brown University

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

Collaborators

David Fleet, *Xerox PARC*

Nancy Pollard, *Brown University*

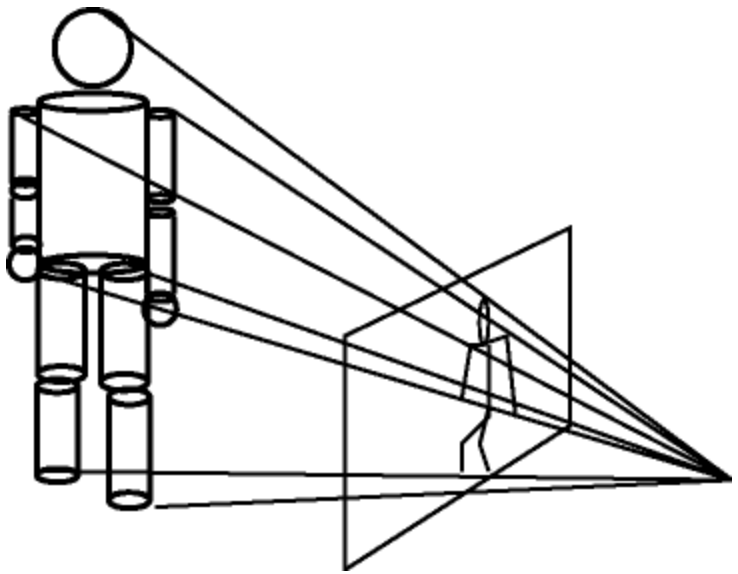
Dirk Ormoneit and **Trevor Hastie**
Dept. of Statistics, Stanford University

Allan Jepson, *University of Toronto*

The (Silly) Problem

Inferring 3D Human Motion

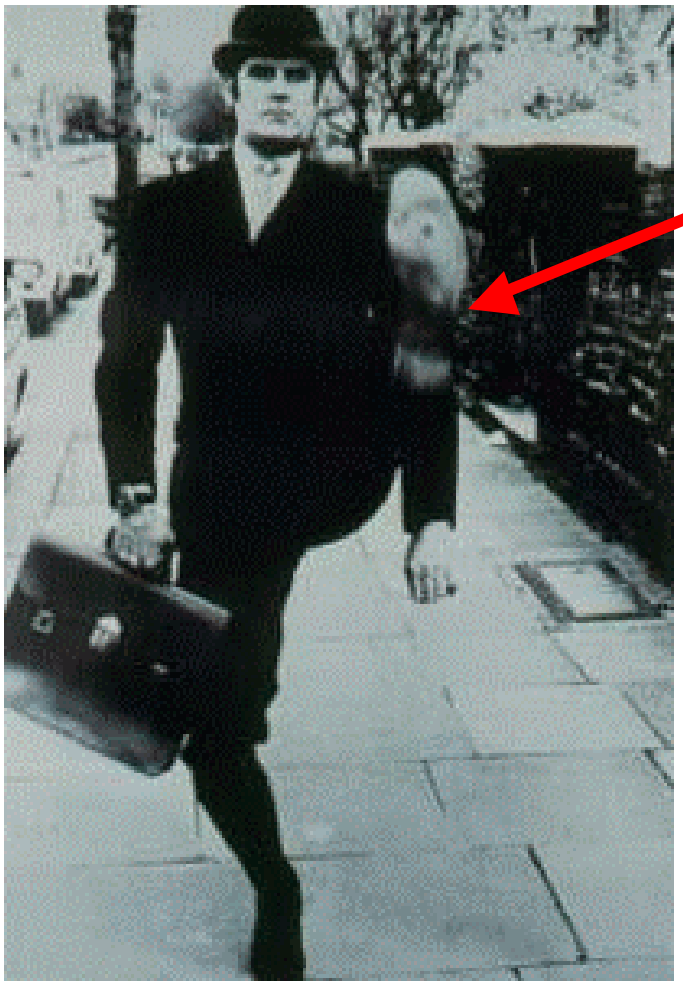
* Infer 3D human motion from 2D image properties.



* No special clothing
* Monocular, grayscale, sequences (archival data)
* Unknown, cluttered, environment
* Incremental estimation

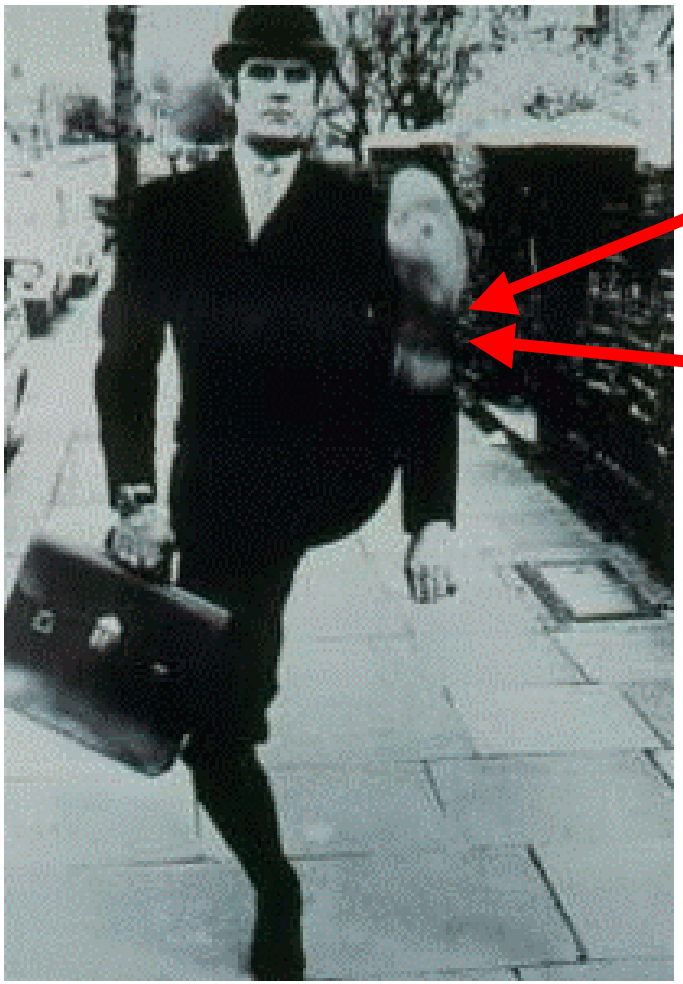
Why is it Hard?

Why is it Hard?



Singularities in viewing direction

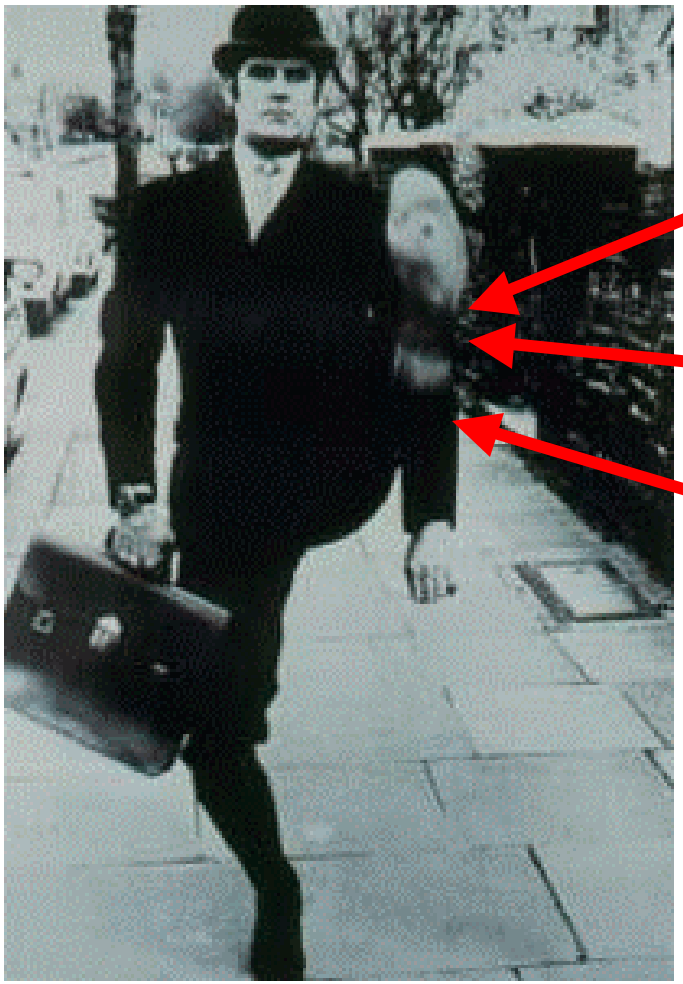
Why is it Hard?



Singularities in viewing direction

Unusual viewpoints

Why is it Hard?



Singularities in viewing direction

Unusual viewpoints

Self occlusion

Why is it Hard?

Singularities in viewing direction

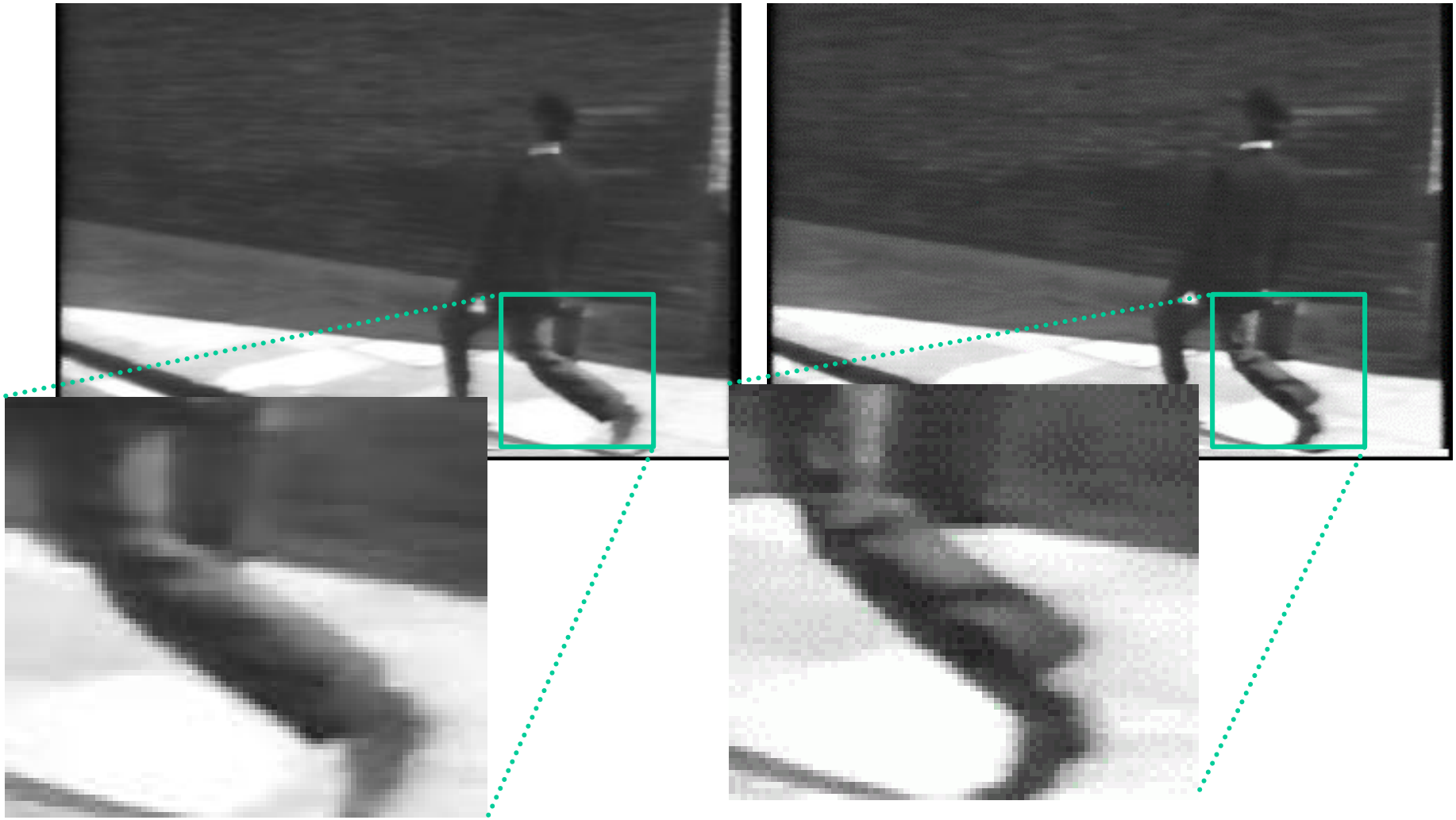
Unusual viewpoints

Self occlusion

Low contrast

Clothing and Lighting

Clothing and Lighting



Large Motions

Limbs move rapidly with respect to their width.

Non-linear dynamics.

Motion blur.

Large Motions

Limbs move rapidly with respect to their width.

Non-linear dynamics.

Motion blur.

Large Motions

Limbs move rapidly with respect to their width.

Non-linear dynamics.

Motion blur.

Ambiguities

Where is the leg?

Which leg is in front?

Ambiguities

Where is the leg?

Which leg is in front?

Ambiguities

Where is the leg?

Which leg is in front?

Ambiguities

Where is the leg?

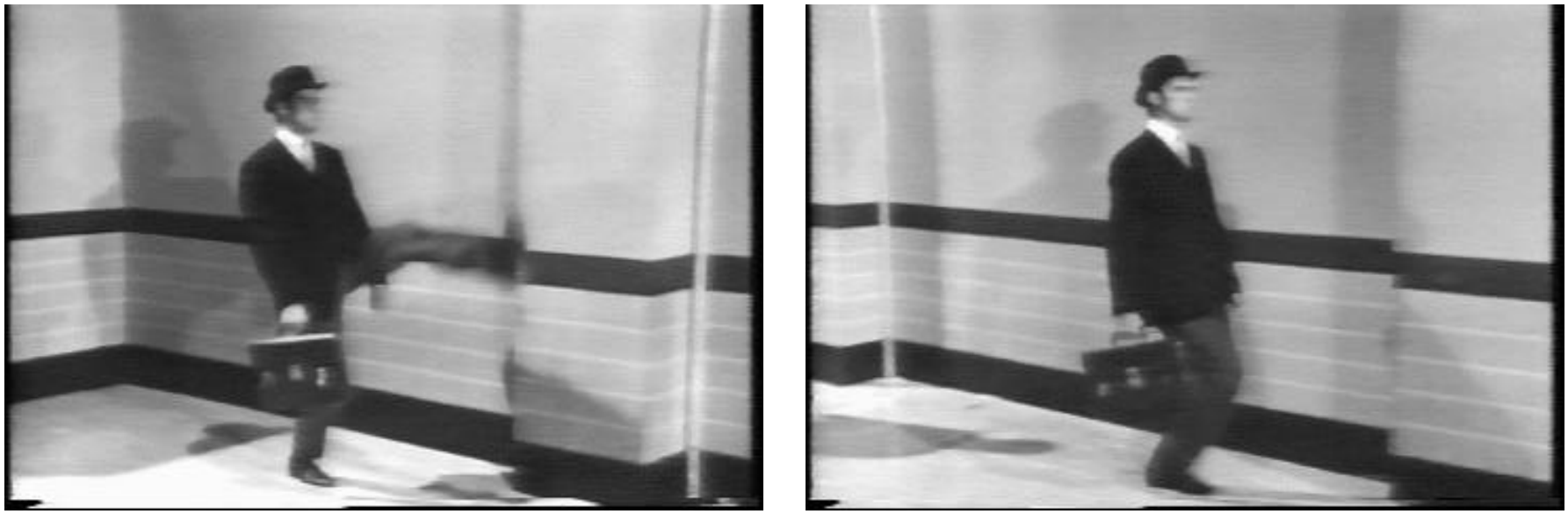
Which leg is in front?

Ambiguities

Where is the leg?

Which leg is in front?

Ambiguities



Accidental alignment

Ambiguities

Occlusion

Whose legs are whose?

Inference/Issues

Bayesian formulation

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

Inference/Issues

Bayesian formulation

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

1. Need a constraining *likelihood* model that is also invariant to variations in human appearance.

Inference/Issues

Bayesian formulation

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

1. Need a constraining *likelihood* model that is also invariant to variations in human appearance.

2. Need a *prior* model of how people move.

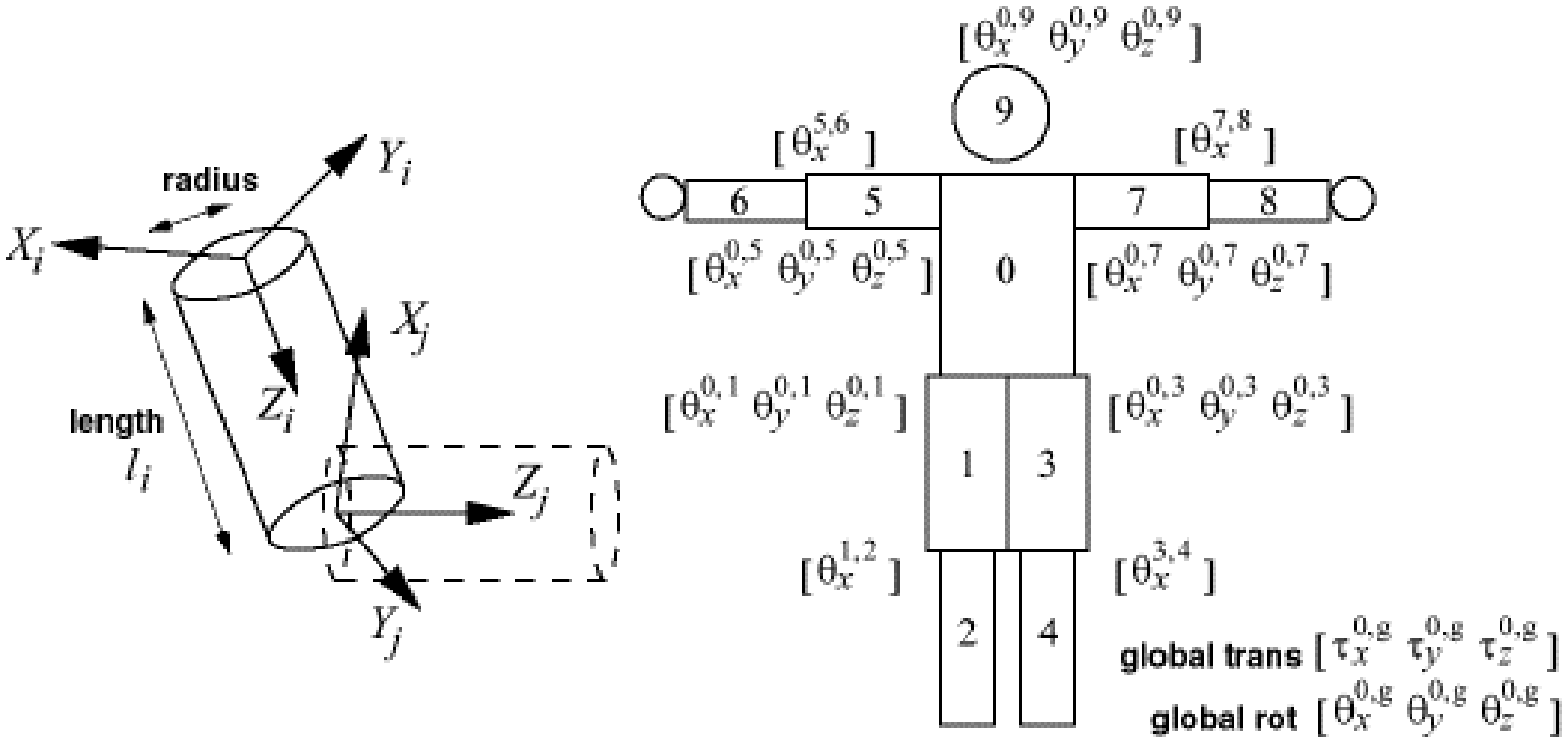
Inference/Issues

Bayesian formulation

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

1. Need a constraining *likelihood* model that is also invariant to variations in human appearance.
2. Need a *prior* model of how people move.
3. Need an effective way to explore the model space (very high dimensional) and represent ambiguities.

Simple Body Model



* Limbs are truncated cones
* Parameter vector of joint angles and angular velocities = f

Key Idea #1 (Likelihood)

1. Use the 3D model to predict the location of limb boundaries (not necessarily features) in the scene.
2. Compute various filter responses *steered* to the predicted orientation of the limb.
3. Compute likelihood of filter responses using a statistical model *learned from examples*.

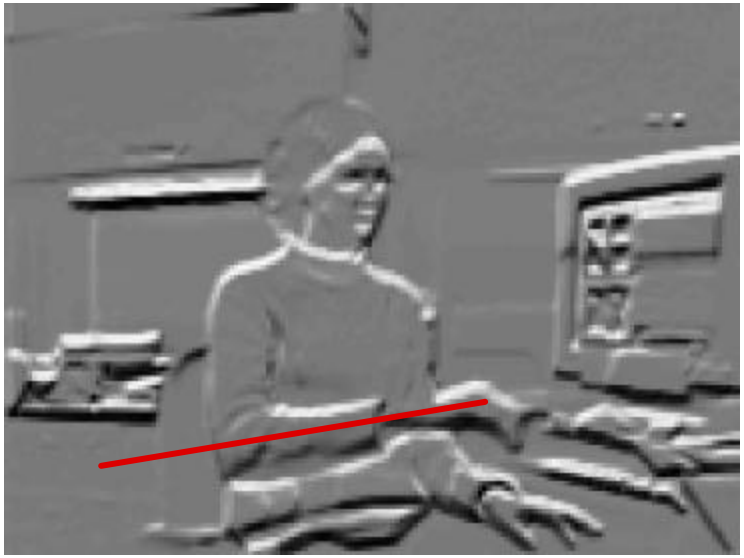
Example Training Images

Edge Filters

Normalized derivatives of Gaussians (Lindeberg, Granlund and Knutsson, Perona, Freeman&Adelson, …)

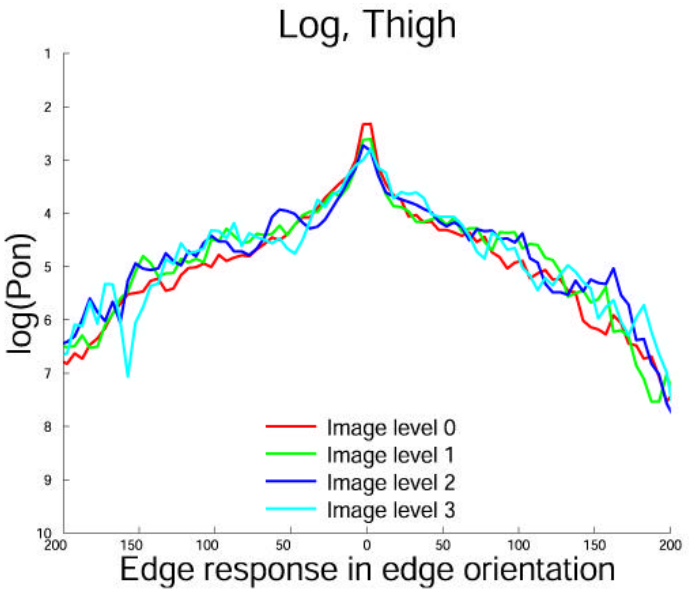
Edge filter response steered to limb orientation:

$$f^{e}(\mathbf{x},q,s) = \sin q\, f_x(\mathbf{x},s) + \cos q\, f_y(\mathbf{x},s)$$

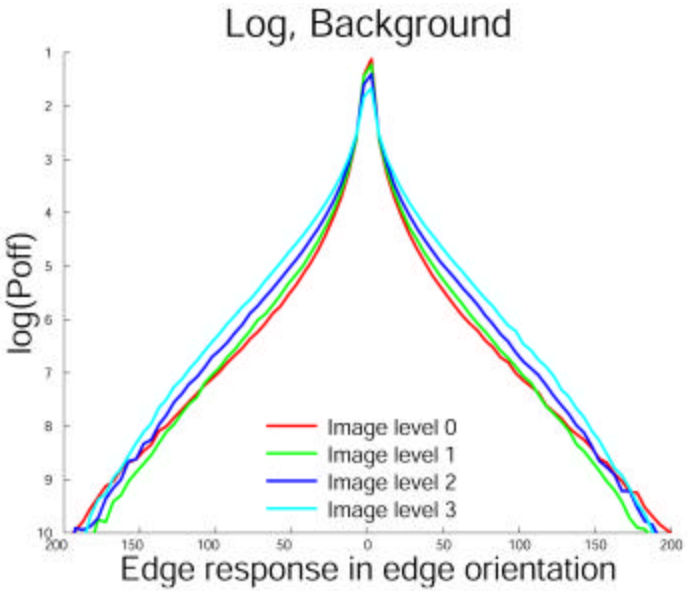


Filter responses steered to arm orientation.

Distribution of Edge Filter Responses



p_{on}(F)



p_{off} (F)

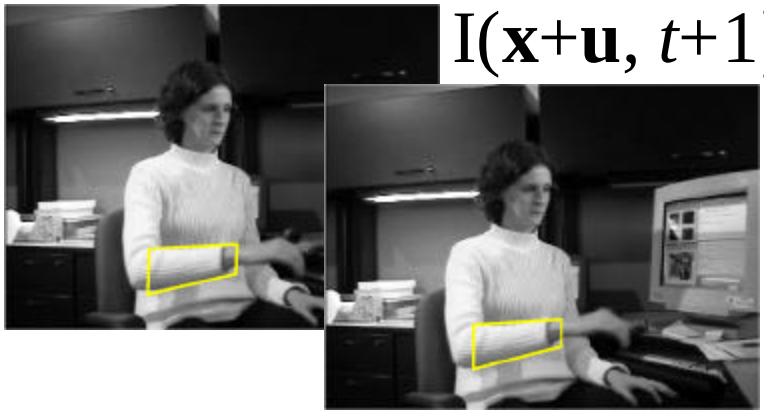
Likelihood ratio, $p_{on}/\ p_{off}$, used for edge detection

Geman & Jednyak and Konishi, Yuille, & Coughlan

Object specific statistics

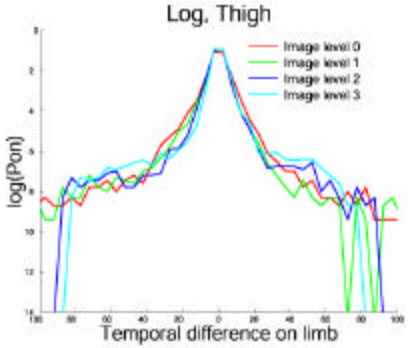
Other Cues

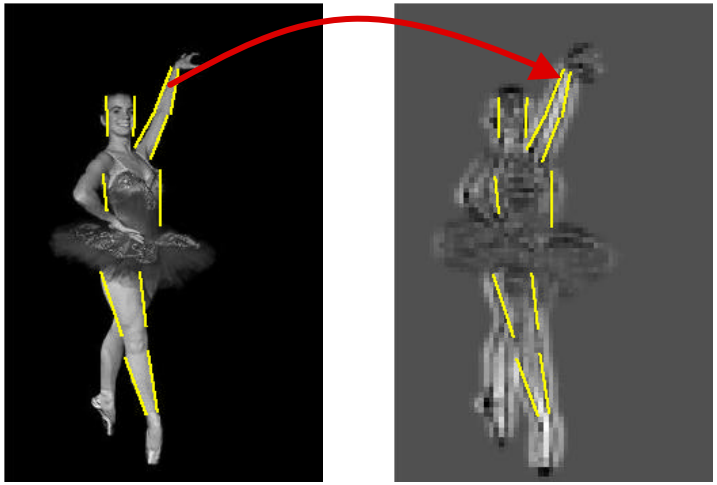
I(**x**, *t*)



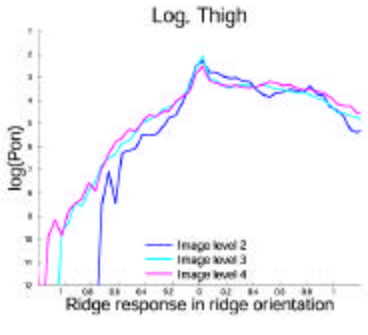
I(**x**+**u**, *t*+1)

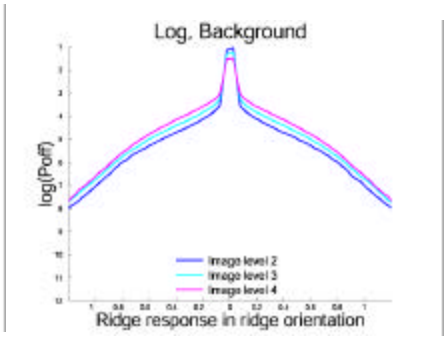
Motion





Ridges



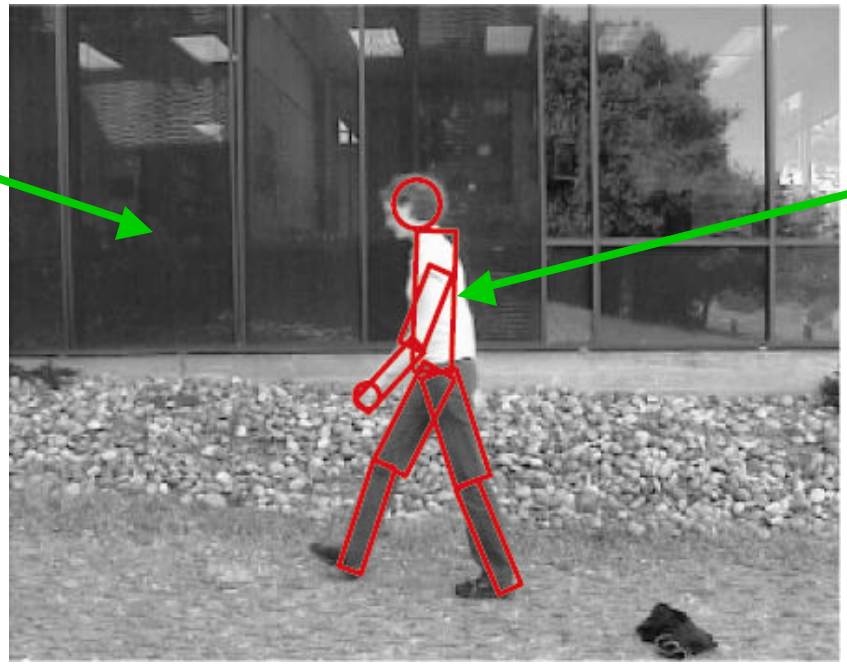


Key Idea #2 (Likelihood)

"Explain" the entire image.

$$p(\text{image} \mid \text{foreground, background}) = \frac{const \prod_{\text{fore pixels}} p(image \mid fore)}{\prod_{\text{fore pixels}} p(image \mid back)}$$

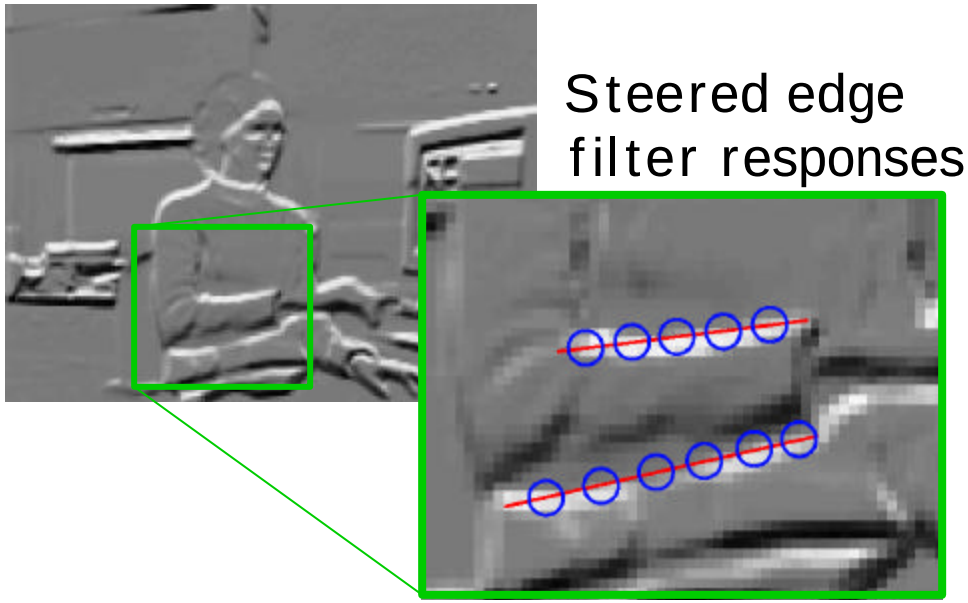
Generic, unknown, background



Foreground person

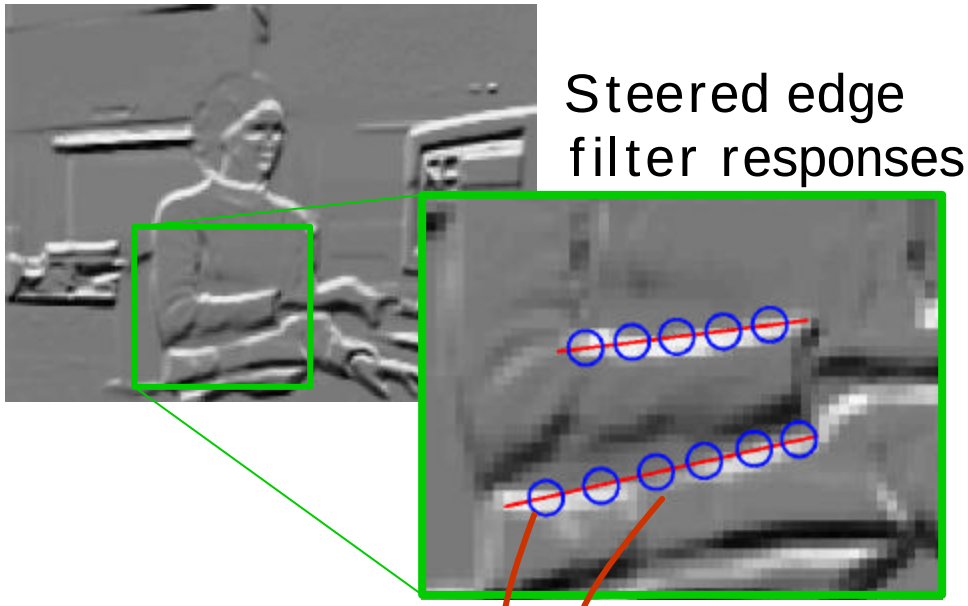
Foreground should explain what the background can't.

Likelihood

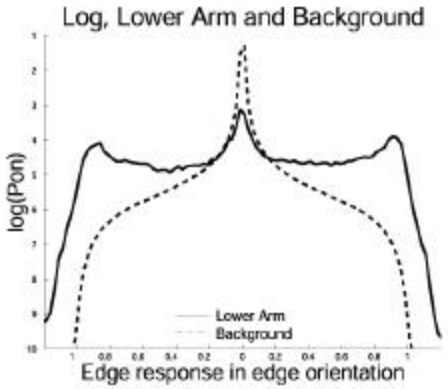


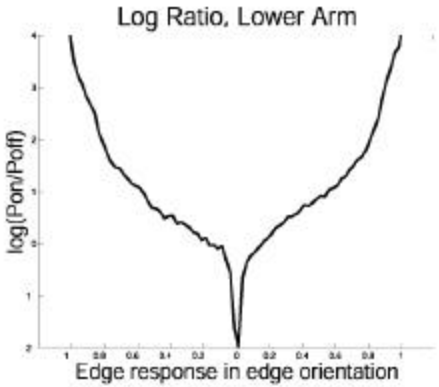
Steered edge
filter responses

Likelihood



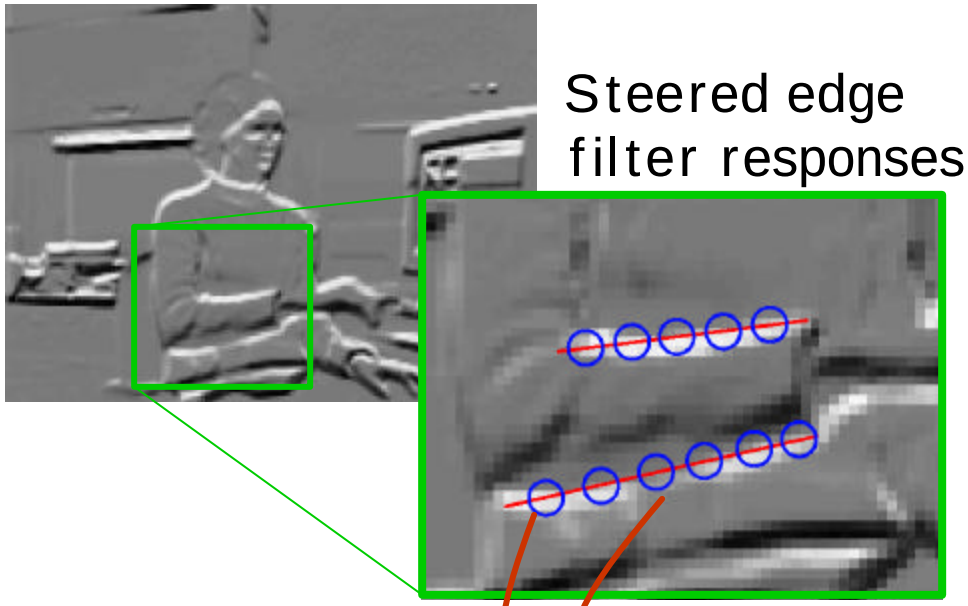
Steered edge filter responses



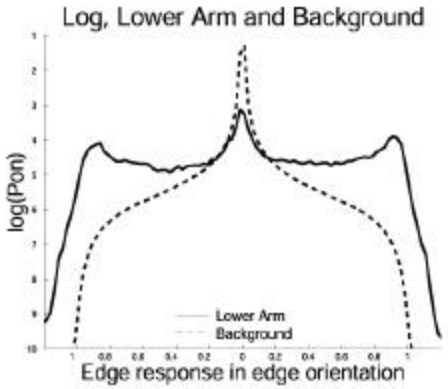


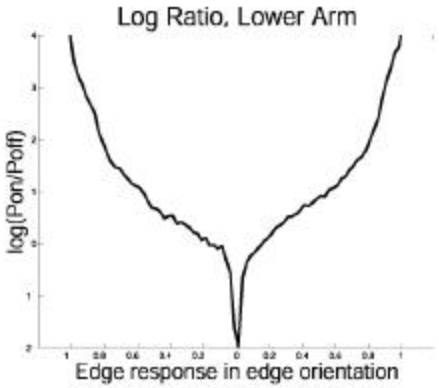
$$\prod_{\text{limbs cues}} \frac{p(\text{filter response} \mid \text{person})}{p(\text{filter response} \mid \text{background})}$$

Likelihood



Steered edge filter responses



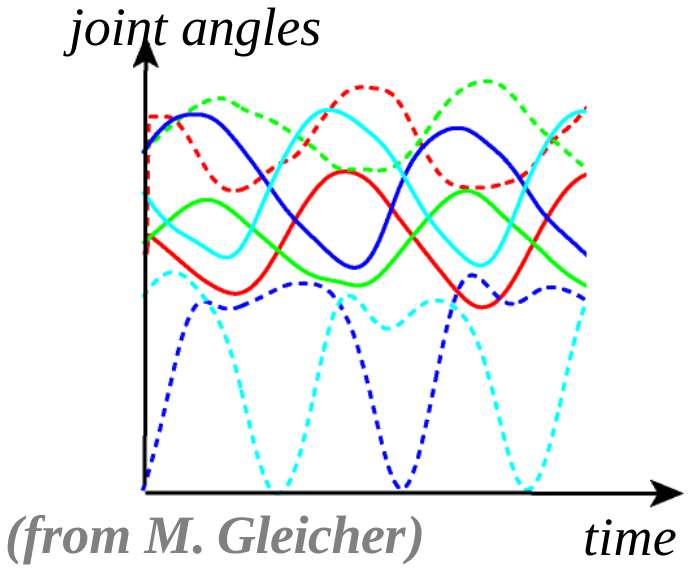


$$\prod_{\text{limbs cues}} \frac{p(\text{filter response} \mid \text{person})}{p(\text{filter response} \mid \text{background})}$$

crude assumption: filter responses independent across scale.

Learning Human Motion

* constrain the posterior to likely & valid poses/motions
* model the variability



(from M. Gleicher)

3D motion-capture data.

* Database with multiple actors and a variety of motions.

Key Idea #3 (Prior)

Problem:

* insufficient data to learn probabilistic model of human motion.

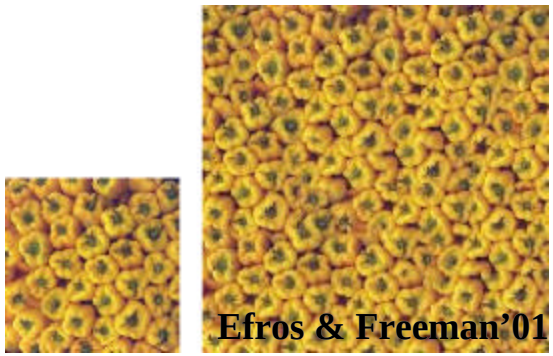
Key Idea #3 (Prior)

Problem:

* insufficient data to learn probabilistic model of human motion.

Alternative:

* the *data* represents all we know



* replace *representation* and *learning* with *search.* (search has to be fast)

* De Bonnet & Viola, Efros & Leung, Efros & Freeman, Paztor & Freeman, Hertzmann et al, …

Implicit Empirical Distribution

Off-line:

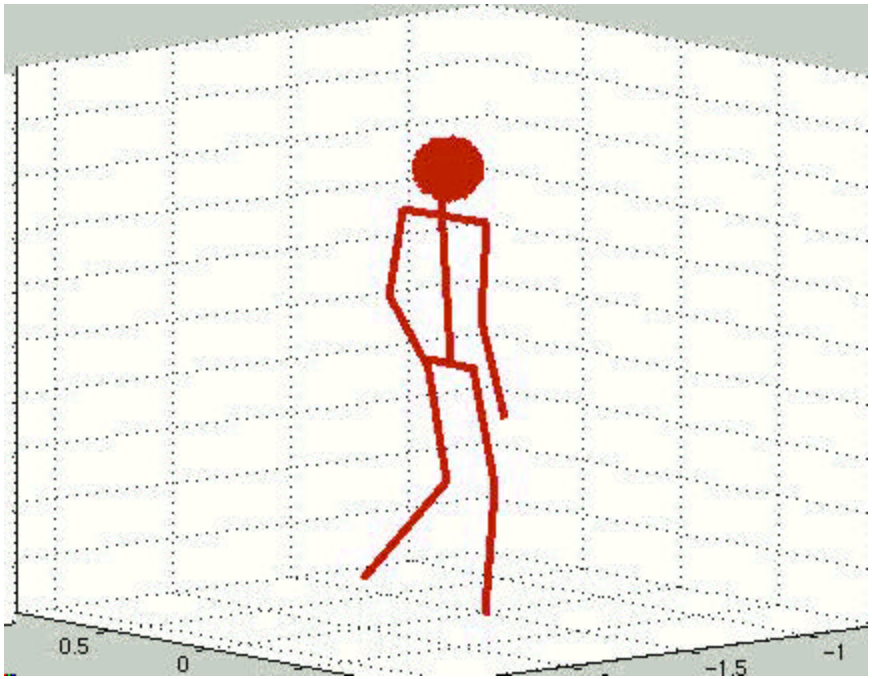
- learn a low-dimensional model of every *n*-frame sequence of joint angles and angular velocities (Leventon & Freeman, Ormoneit et al, …)
- project training data onto model to get small number of coefficients describing each time instant
- build a tree structured representation

“Textural” Model

On-line: Given an *n*-frame input motion

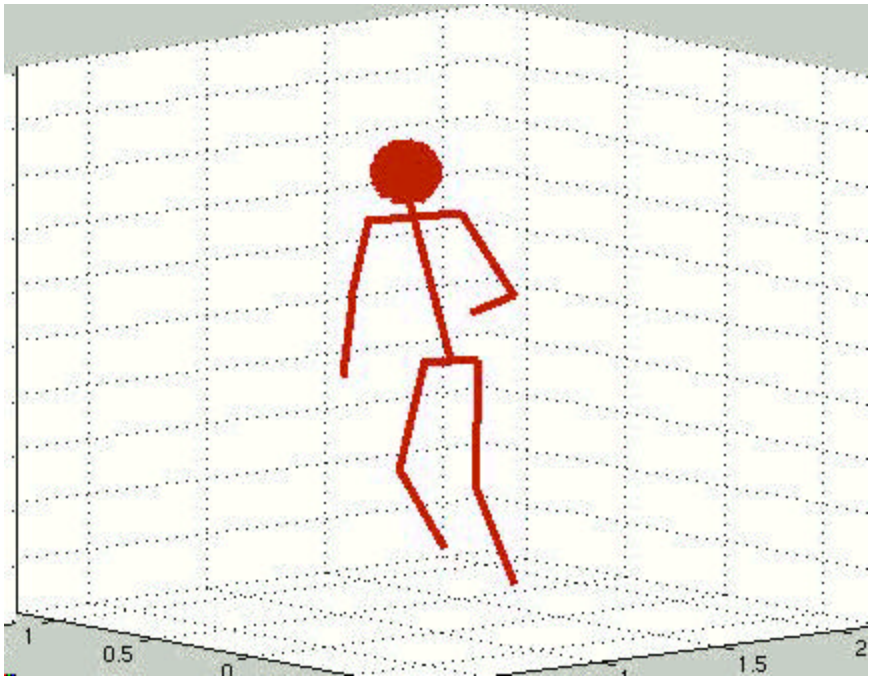
- project onto low-dimensional model.
- index in *log* time using the coefficients.
- return the best *k* approximate matches (and form a “proposal” distribution).
- sample from them and return the $n+1^{st}$ pose.

Synthetic Walker



* Colors indicate different training sequences.

Synthetic Swing Dancer



Bayesian Formulation

$$p(f_t \mid \vec{\mathbf{I}}_t) =$$

$$k \; p(\mathbf{I}_t \mid f_t) \int (p(f_t \mid f_{t-1}) \, p(f_{t-1} \mid \vec{\mathbf{I}}_{t-1})) \, df_{t-1}$$

Bayesian Formulation

Posterior over model parameters given an image sequence.

$$p(f_t \mid \vec{\mathbf{I}}_t) =$$

$$k \; p(\mathbf{I}_t \mid f_t) \int (p(f_t \mid f_{t-1}) \, p(f_{t-1} \mid \vec{\mathbf{I}}_{t-1})) \, d f_{t-1}$$

Bayesian Formulation

$$p(f_t \mid \vec{\mathbf{I}}_t) =$$
$$k \, p(\mathbf{I}_t \mid f_t) \int (p(f_t \mid f_{t-1}) \, p(f_{t-1} \mid \vec{\mathbf{I}}_{t-1})) \, df_{t-1}$$

Likelihood of
observing the image
given the model parameters

Bayesian Formulation

$$p(f_t \mid \vec{\mathbf{I}}_t) = \qquad \text{Temporal model (prior)}$$

$$k\ p(\mathbf{I}_t \mid f_t) \int \left(p(f_t \mid f_{t-1})\ p(f_{t-1} \mid \vec{\mathbf{I}}_{t-1}) \right) df_{t-1}$$

Bayesian Formulation

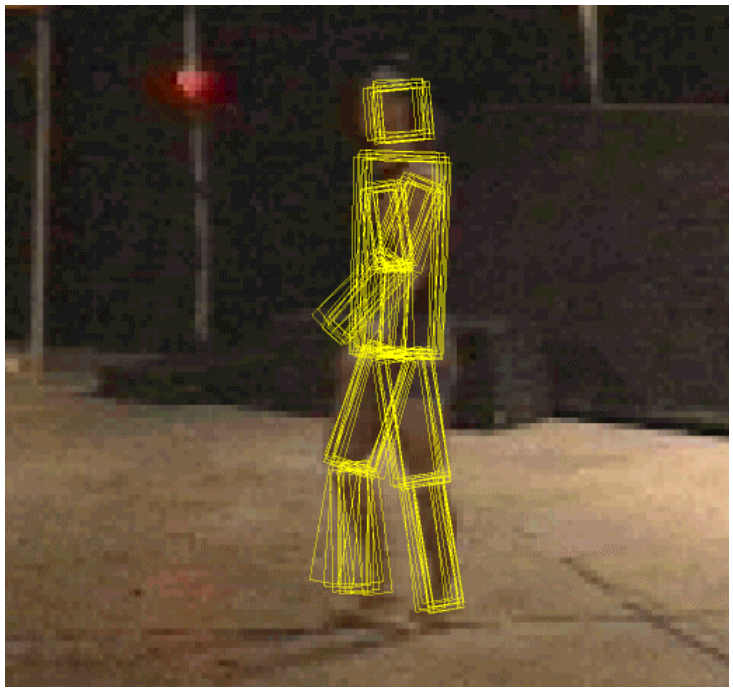
$$p(f_t \mid \vec{\mathbf{I}}_t) =$$

$$k \; p(\mathbf{I}_t \mid f_t) \int \left(p(f_t \mid f_{t-1}) \; p(f_{t-1} \mid \vec{\mathbf{I}}_{t-1}) \right) df_{t-1}$$

Posterior from previous time instant

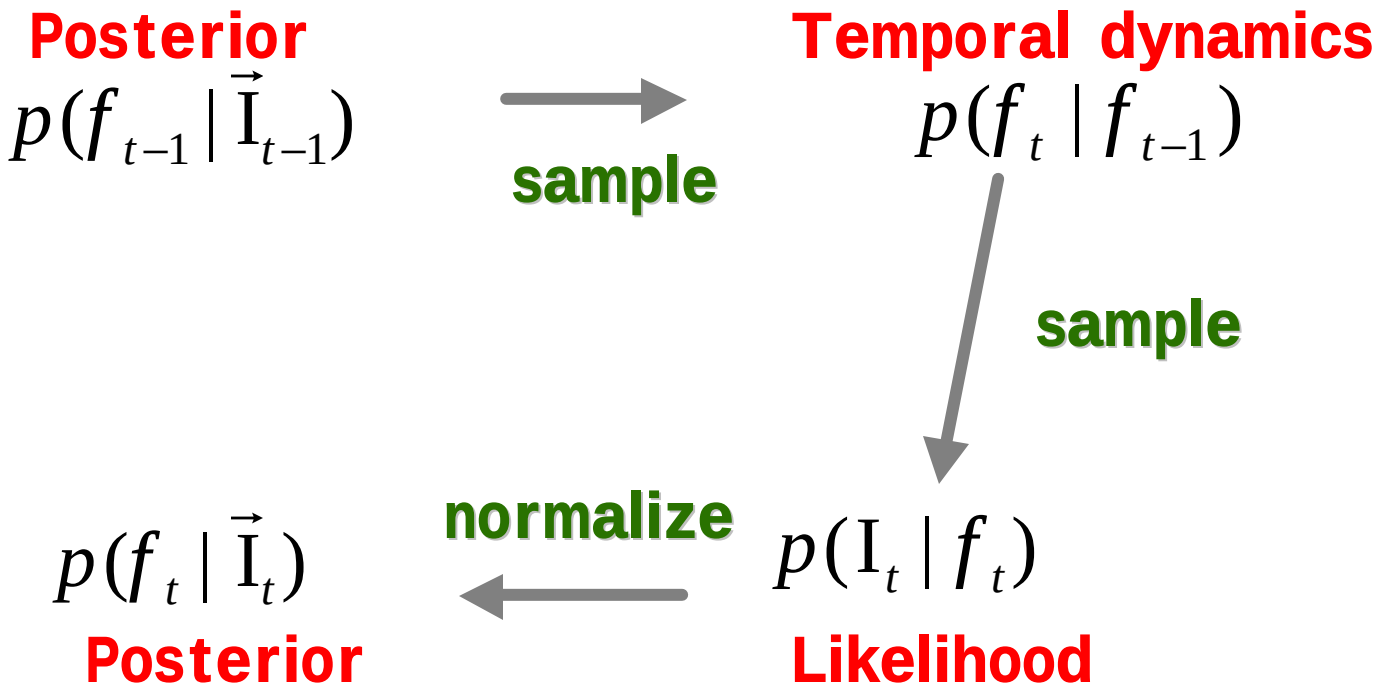
Key Idea #4 (Ambiguity)

* Represent a multi-modal posterior probability distribution over model parameters
  - sampled representation
  - each sample is a pose and its probability
  - predict over time using a *particle filtering* approach.

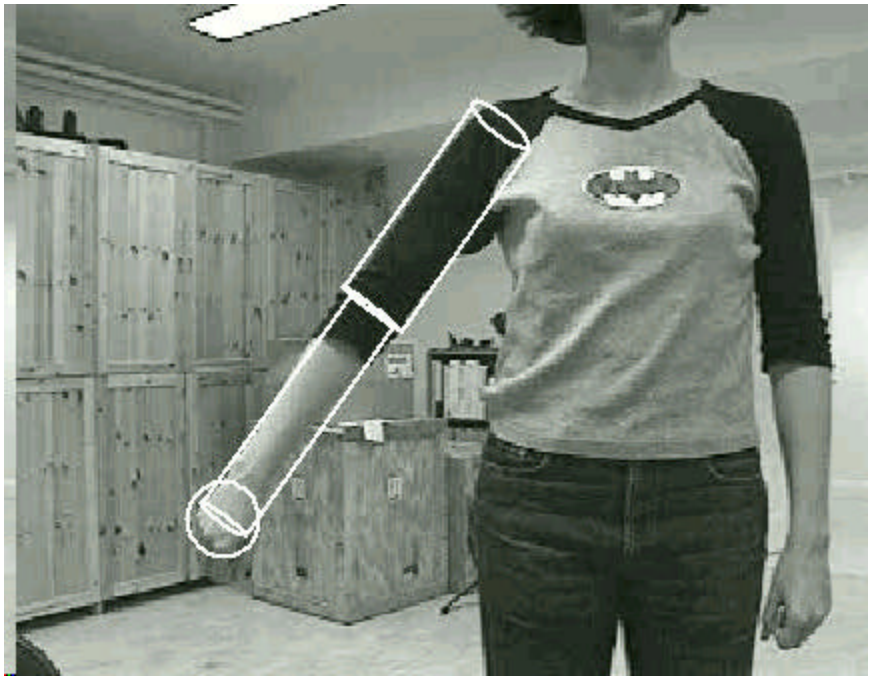


Samples from a distribution over 3D poses.

Particle Filter



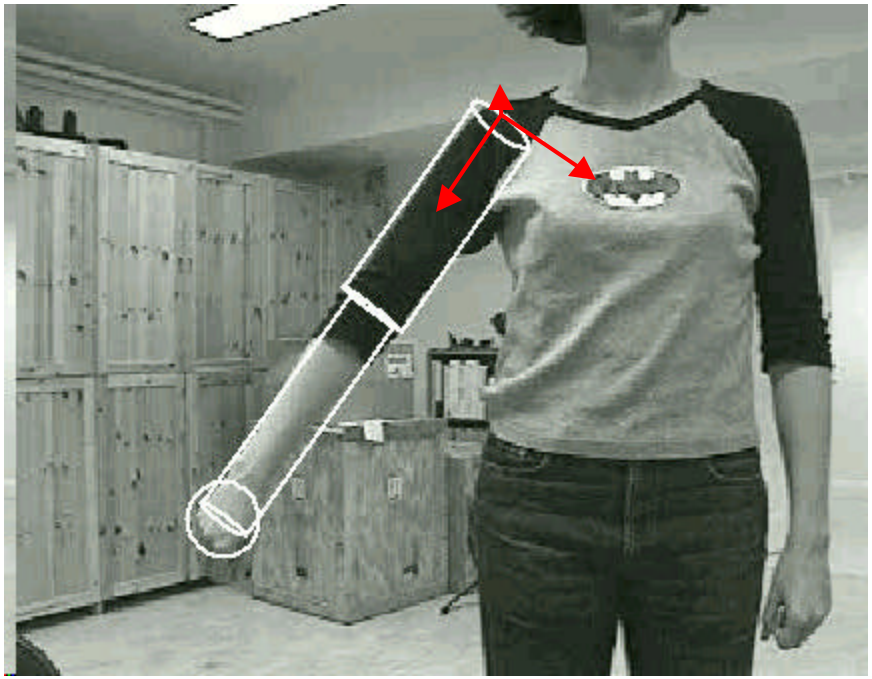
What does the posterior look like?



Shoulder: 3dof
Elbow: 1dof

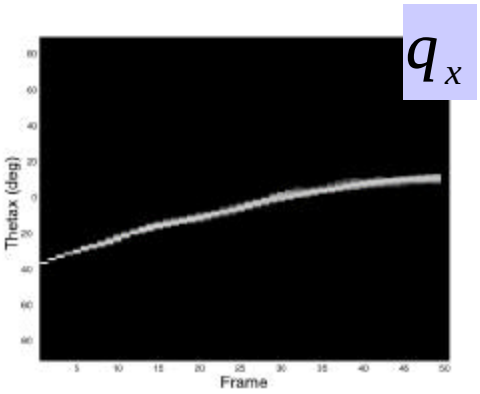
Elbow bends

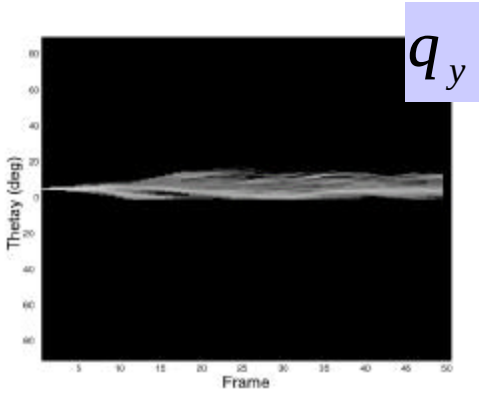
What does the posterior look like?

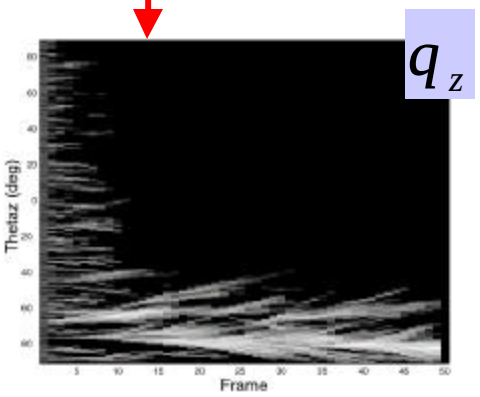


Shoulder: 3dof
Elbow: 1dof

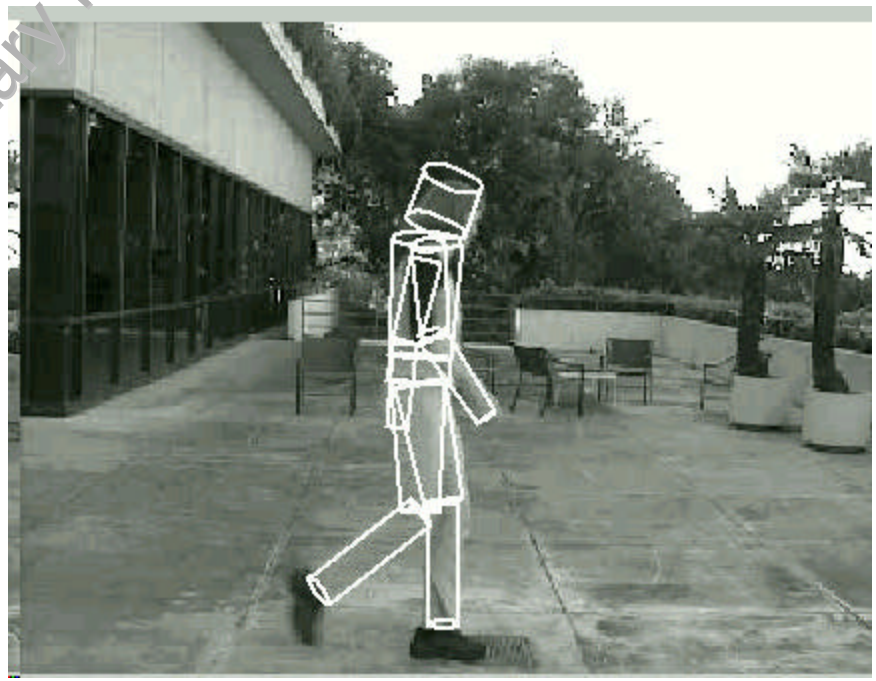
Elbow bends







Stochastic 3D Tracking



* 2500 samples, multiple cues.

Conclusions

Inferring human motion, silly or not, from video is challenging.

We have tackled three important parts of the problem:

1. Probabilistically modeling human appearance in a generic, yet useful, way.
2. Representing the range of possible motions using techniques from texture modeling.
3. Dealing with ambiguities and non-linearities using particle filtering for Bayesian inference.