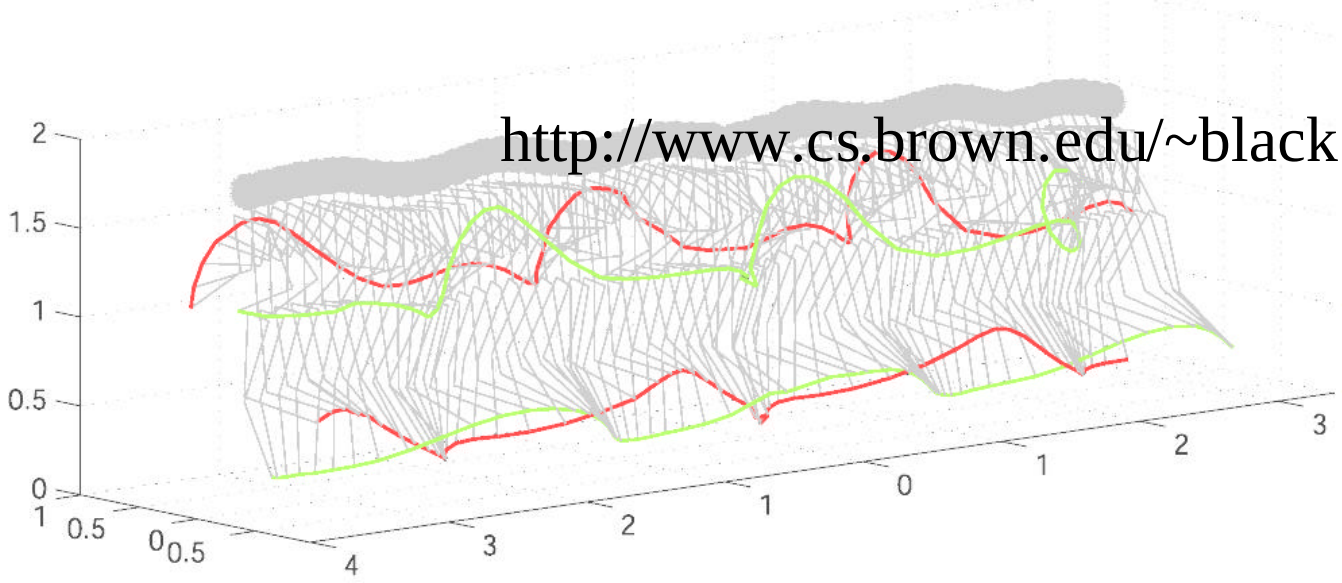
Inferring 3D People from 2D Images

Michael J. Black

Department of Computer Science
Brown University

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



Collaborators

Hedvig Sidenbladh, *Swedish Defense Research Inst.*
Leon Sigal, *Brown University*
Michael Isard, *Microsoft Research*
Ben Sigelman, *Brown University*
David Fleet, *PARC Inc. (*soon: *U. Toronto)*

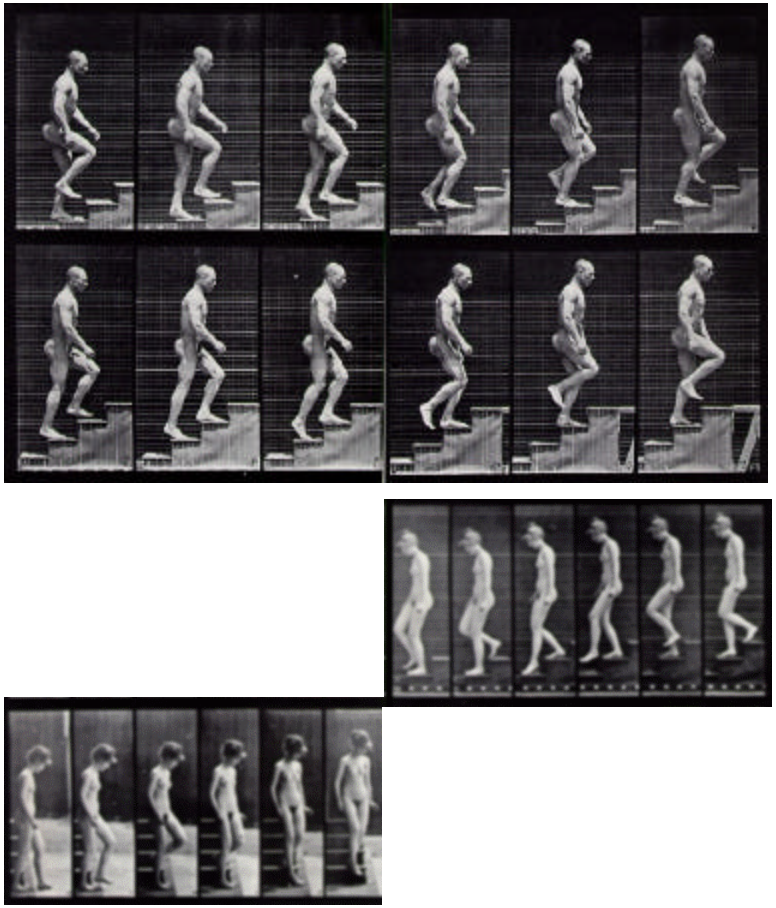
Funding:

DARPA HumanID project.

Capturing Humans in Motion

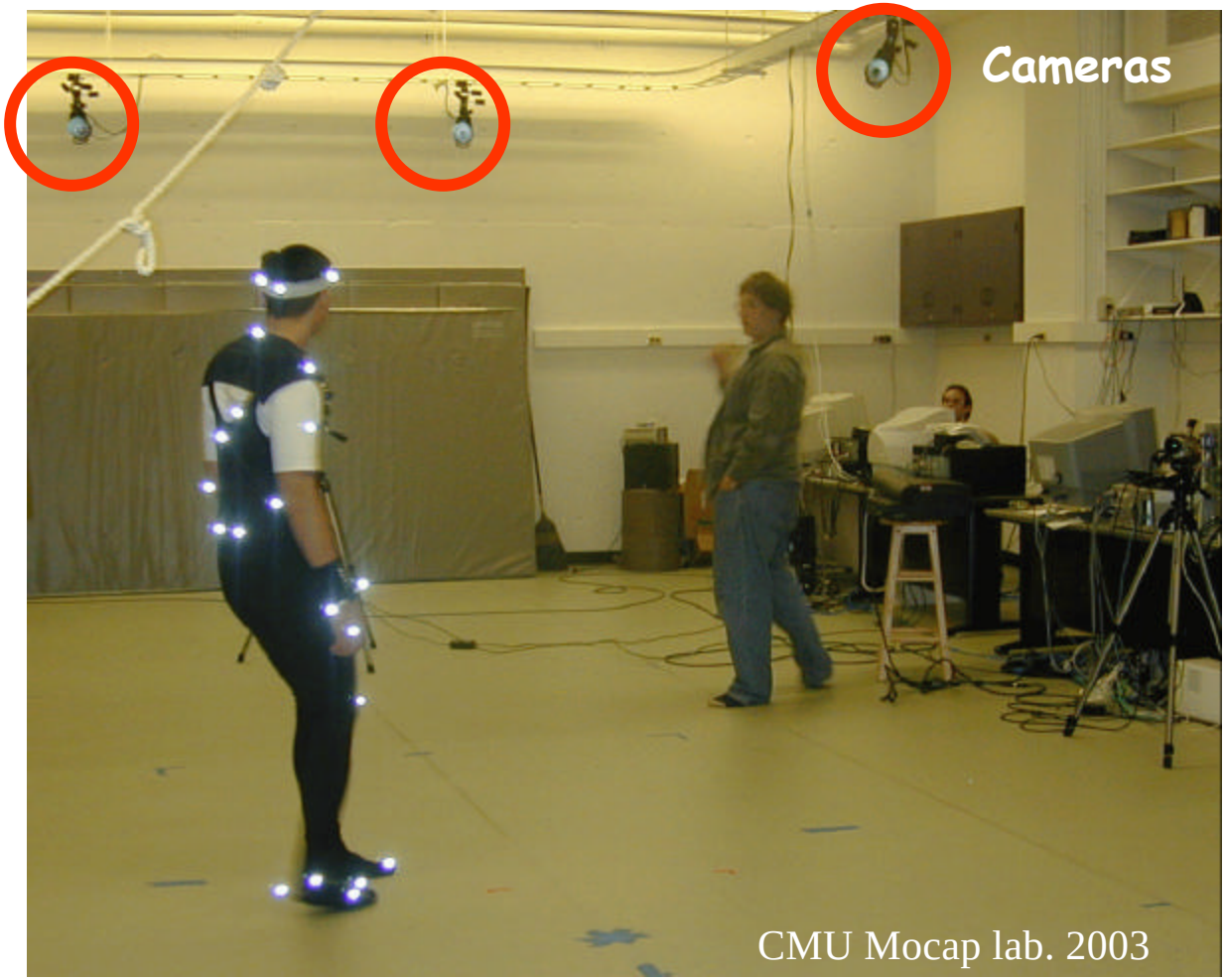
Loss of depth and motion in projection to 2D images.

ETIENNE-JULES MAREY, 1882.
Marker-based tracking



EADWEARD MUYBRIDGE, 1884-5.
Multiple cameras.

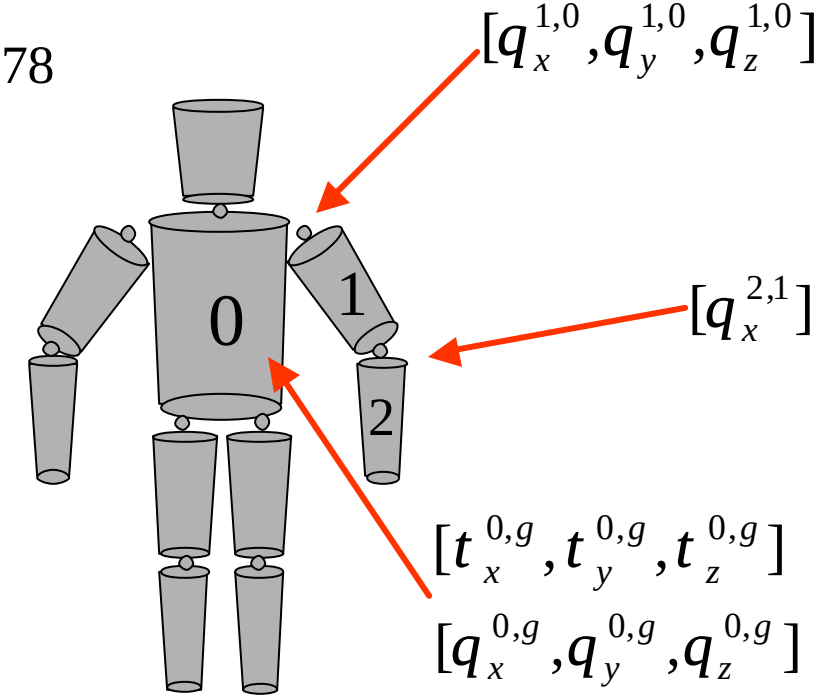
“Mocap” Today



Markerless Motion Capture

Kinematic tree:

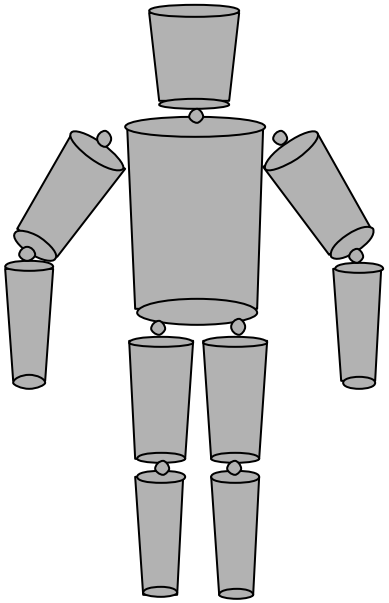
Marr&Nishihara ‘78



Also model angular velocities ~ 50D space

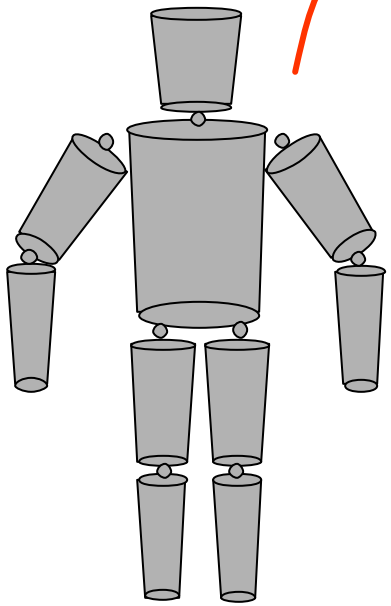
Represent a “pose” at time *t* by a vector of these parameters: f_t

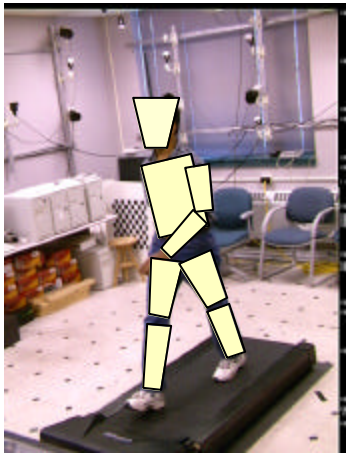
Markerless Motion Capture



Markerless Motion Capture

Find the pose f_t



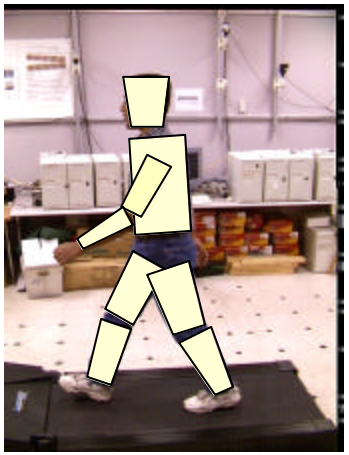


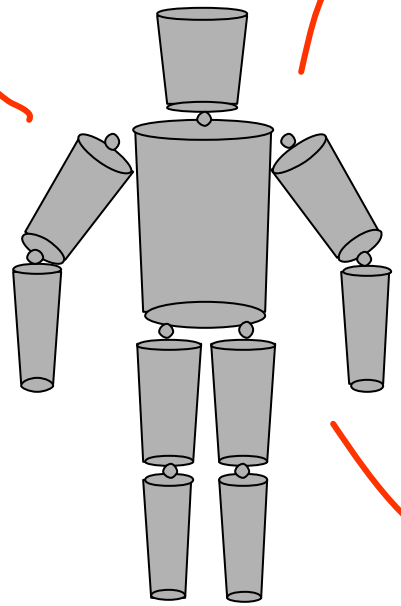
* No special clothing
* Unknown, cluttered, environment
* Incremental estimation

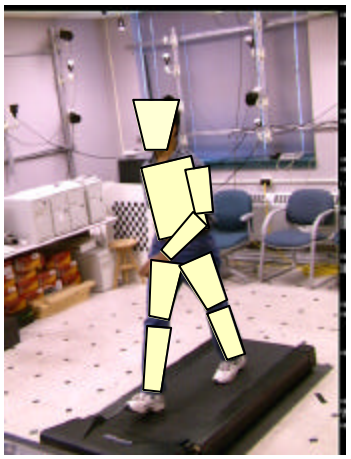
such that the projection “matches”
the image data.

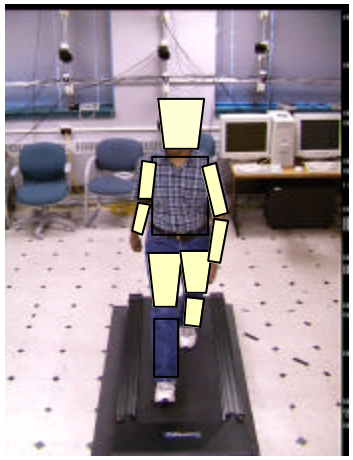
Markerless Motion Capture

Find the pose f_t









such that the projection “matches” the image data.

Motivation

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

Why is it Difficult?

The appearance of people can vary dramatically.

Bones and joints are *unobservable* (muscle, skin, clothing hide the underlying structure).

Why is it Difficult?

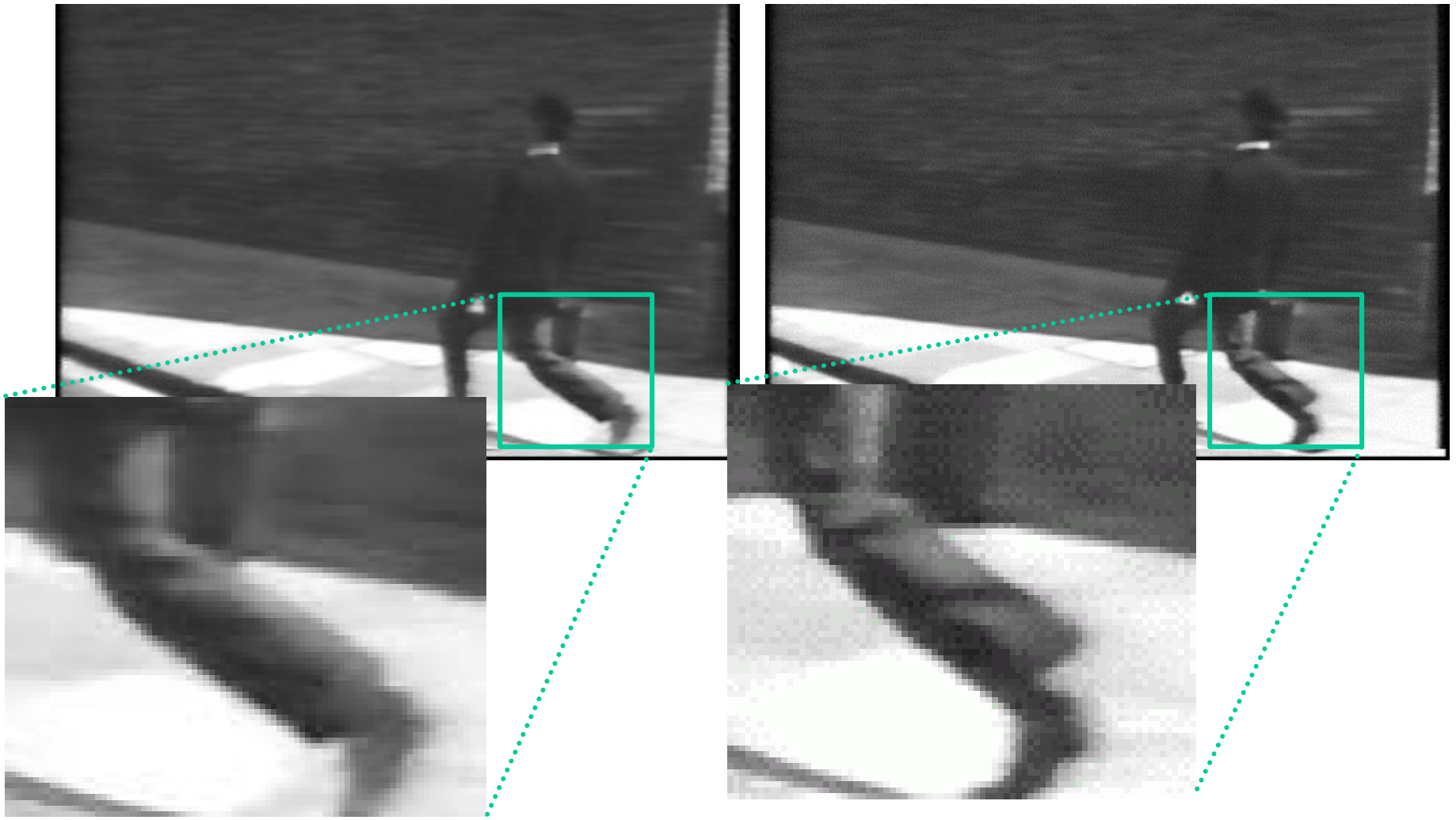
Loss of 3D in 2D projection

Unusual poses

Self occlusion

Low contrast

Clothing and Lighting



Large Motions

Long-range motions.

(makes search and matching hard)

Motion blur.

(nothing to match)

Ambiguities

Ambiguous matches

Self occlusion

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.
4. Integrate information over time.

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

Approach

Bayesian formulation

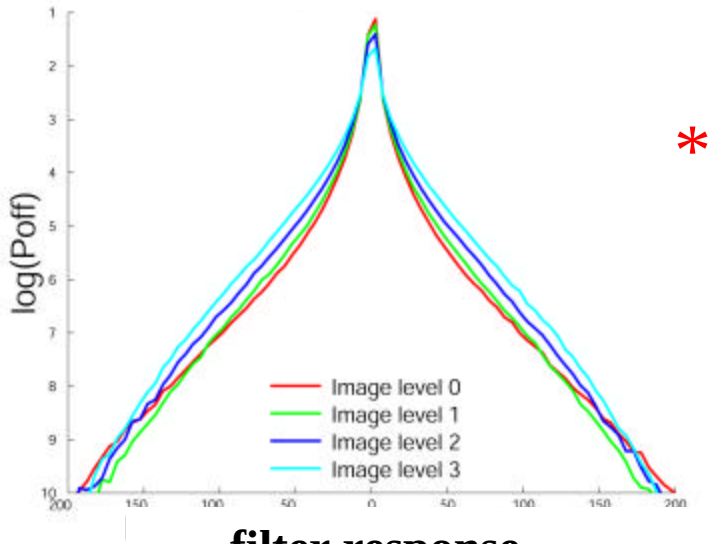
$$p(\text{model} \mid \text{cues}) = \frac{p(\text{cues} \mid \text{model})\, p(\text{model})}{p(\text{cues})}$$

1. **Model**: Kinematic tree. **Cues:** filter responses.
2. **Likelihood**: exploit *learned* likelihood of filter responses conditioned on the projected model.
3. **Prior**: statistical model, *learned* from examples.
4. **Search**: discretize intelligently using factored sampling and search using a particle filter.

Towards a Rigorous Likelihood

1. Project 3D model into image to predict the location of limb regions in the scene.
2. Compute rich set of spatial and temporal filter responses conditioned on the predicted orientation of the limb.
3. Compute likelihood of filter responses using a statistical model *learned from examples*.

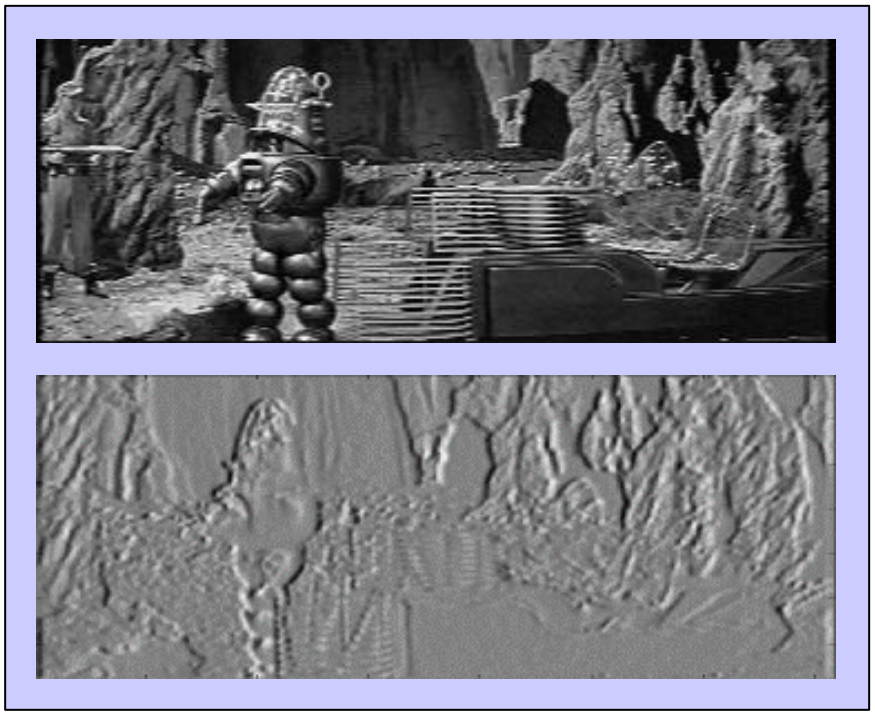
Natural Image Statistics



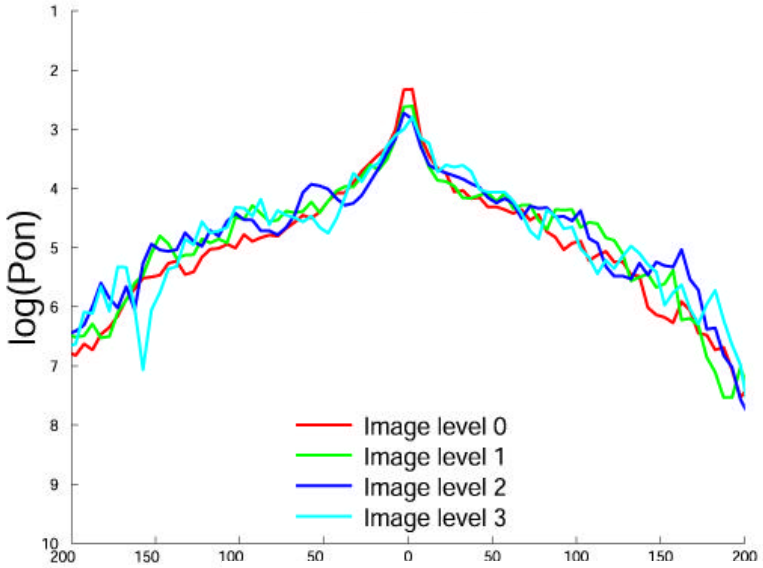
filter response

Ruderman. Lee, Mumford, Huang. Portilla and Simoncelli. Olshausen & Field. Xu, Wu, & Mumford. …

* Marginal statistics of image derivatives are non-Gaussian.
* Consistent across scale.



Human-Specific Statistics

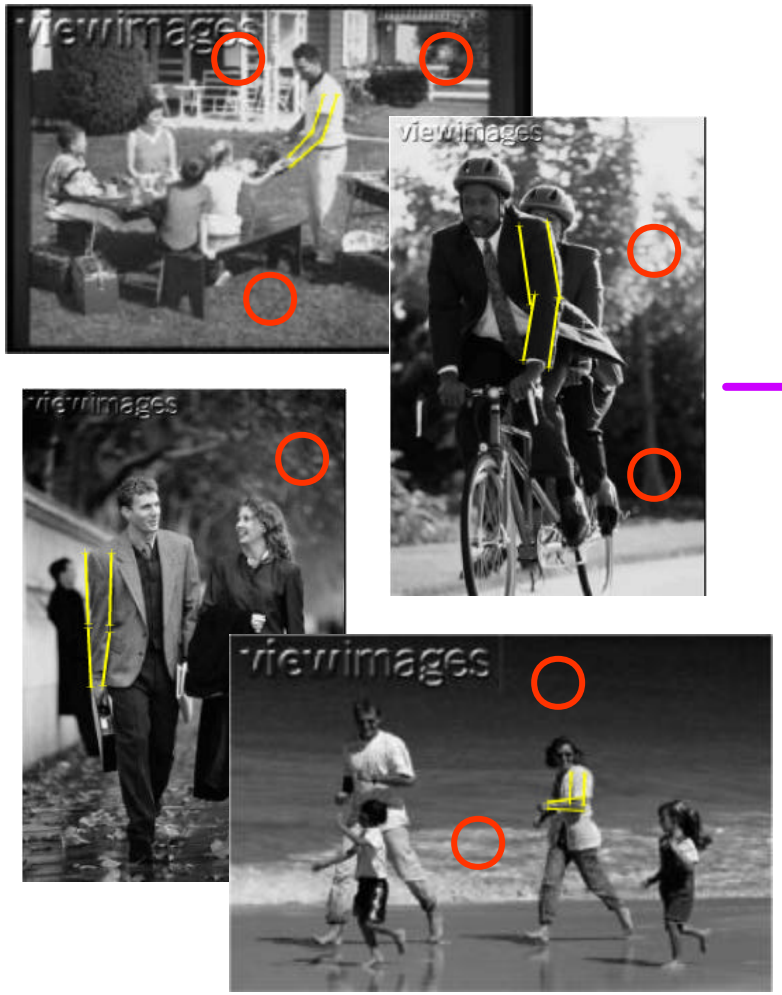


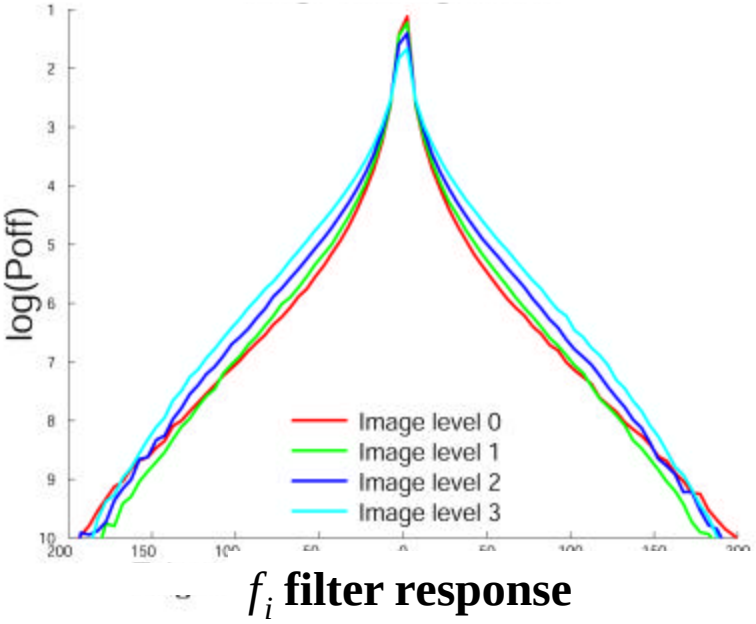
f_i **filter response**

$$p_{on}(f_i \mid f_t)$$

Learn marginal statistics.

Generic Background Statistics



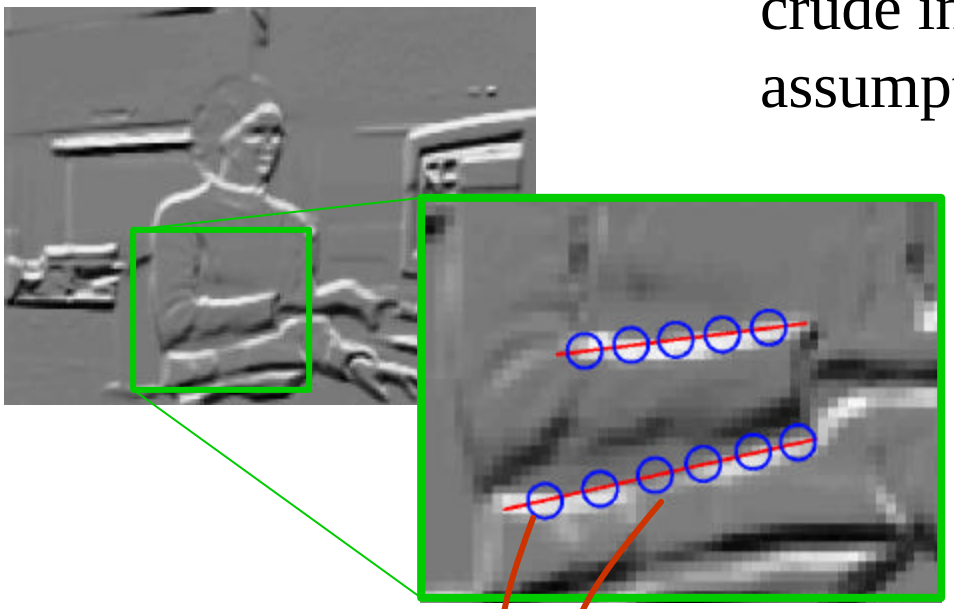


f_i **filter response**

$$p_{off}(f_i)$$

Likelihood

crude independence assumptions



$$p(f_1,\ldots,f_k \mid f_t) \propto \prod_{\text{limbs}} \prod_{i=1}^{\text{filters}} \frac{p_{on}(f_i \mid f_t)}{p_{off}(f_i)}$$

Prior

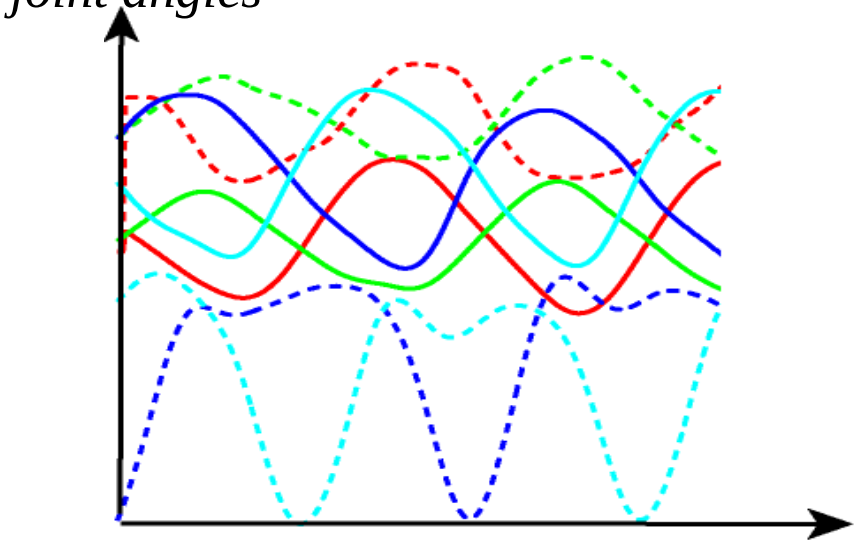
Bayesian formulation

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

1. **Model**: Kinematic tree. **Cues:** filter responses.
2. **Likelihood**: exploit *learned* likelihood of filter responses conditioned on the projected model.
3. **Prior**: statistical model, *learned* from examples.

Learning Human Motion

* constrain the posterior to likely & valid poses/motions
* What we want: $p(f_t \mid f_{t-1})$



(from M. Gleicher)

3D motion-capture data.

* Database with multiple actors and a variety of motions.

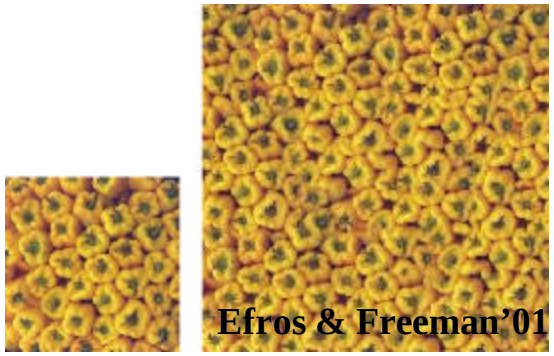
Implicit Probabilistic Prior

Problem:

* insufficient data to ***learn*** an explicit prior probabilistic model of human motion.

Alternative:

* the *data* represents all we know



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

Texture Synthesis

"Database"

Synthetic Texture

* e.g. De Bonnet&Viola, Efros&Leung, Efros&Freeman, Paztor&Freeman, Hertzmann et al.

* Image(s) as an *implicit probabilistic model*.

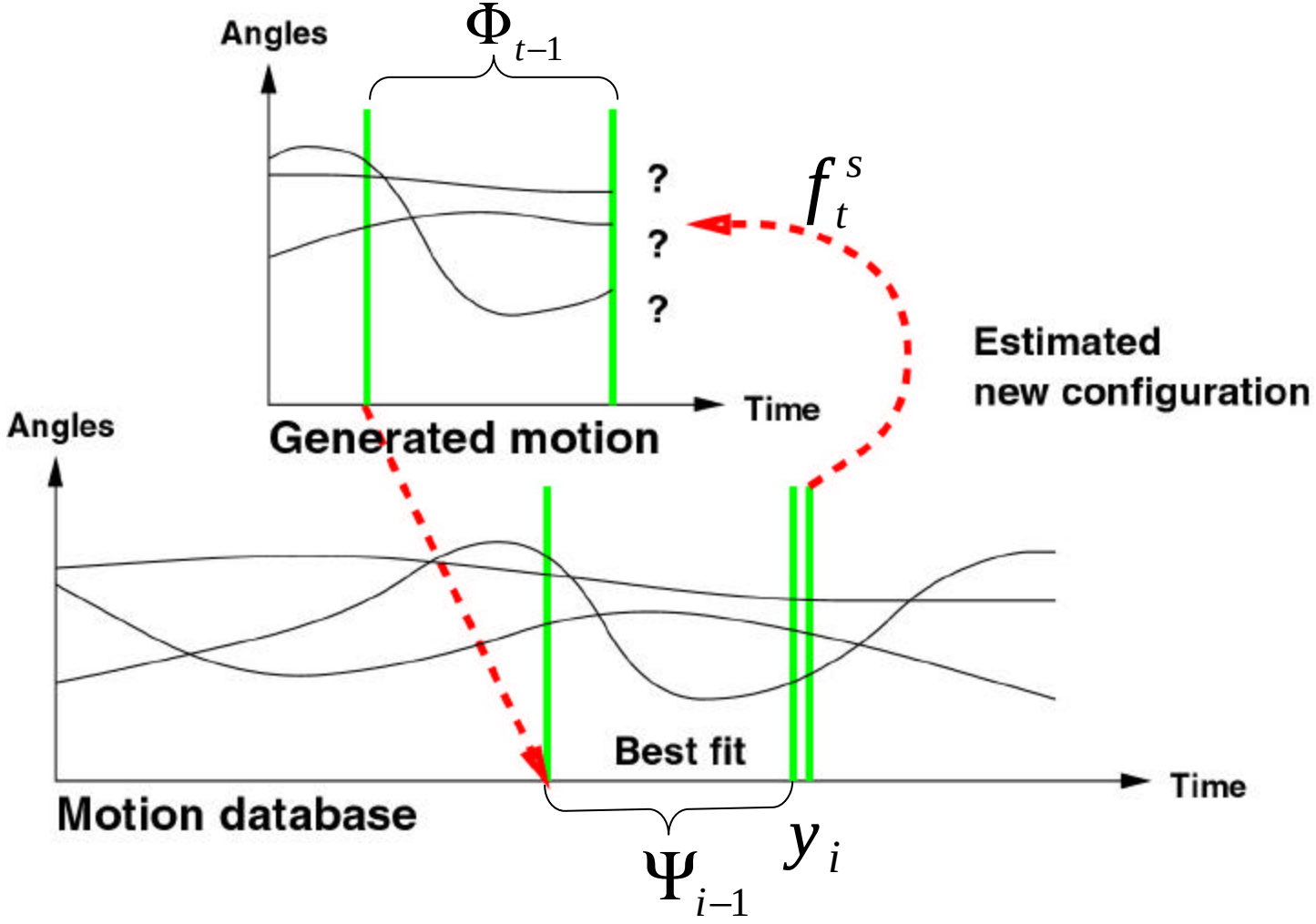
"Mocap Soup" [Cohen]

SIGGRAPH'2002:

- Arikan & Forsyth. *Interactive motion generation from examples*
- Li et al. *Motion textures: A two-level statistical model for character motion synthesis*
- Lee et al. *Interactive control of avatars animated with human motion data*
- Kovar et al. *Motion graphs*
- Pullen & Bregler. *Motion capture assisted animation: Texturing and synthesis*

Here we formulate a probabilistic model suitable for stochastic search and Bayesian tracking.

Motion Texture



Probabilistic Formulation

Want samples from the temporal prior

$$p(f_t^s \mid \Phi_{t-1})$$

pose at time t — history of poses up to t-1

Instead, sample from

$$p(\Psi_{i-1} \mid \Phi_{t-1})$$

Database index i-1

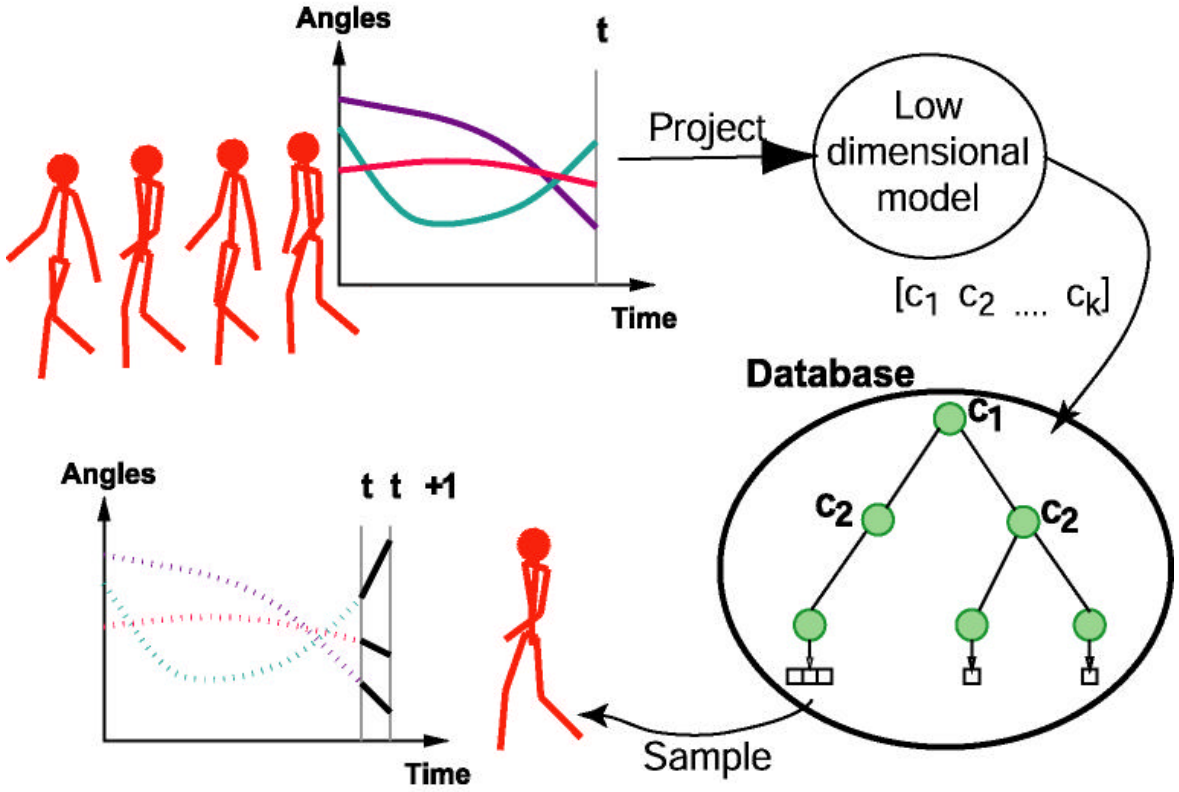
and take $f_t^s = y_i$

Problem:

Compute $p(\Psi_i \mid \Phi_t)$ for all motions i in the database?

Approximate Probabilistic Model

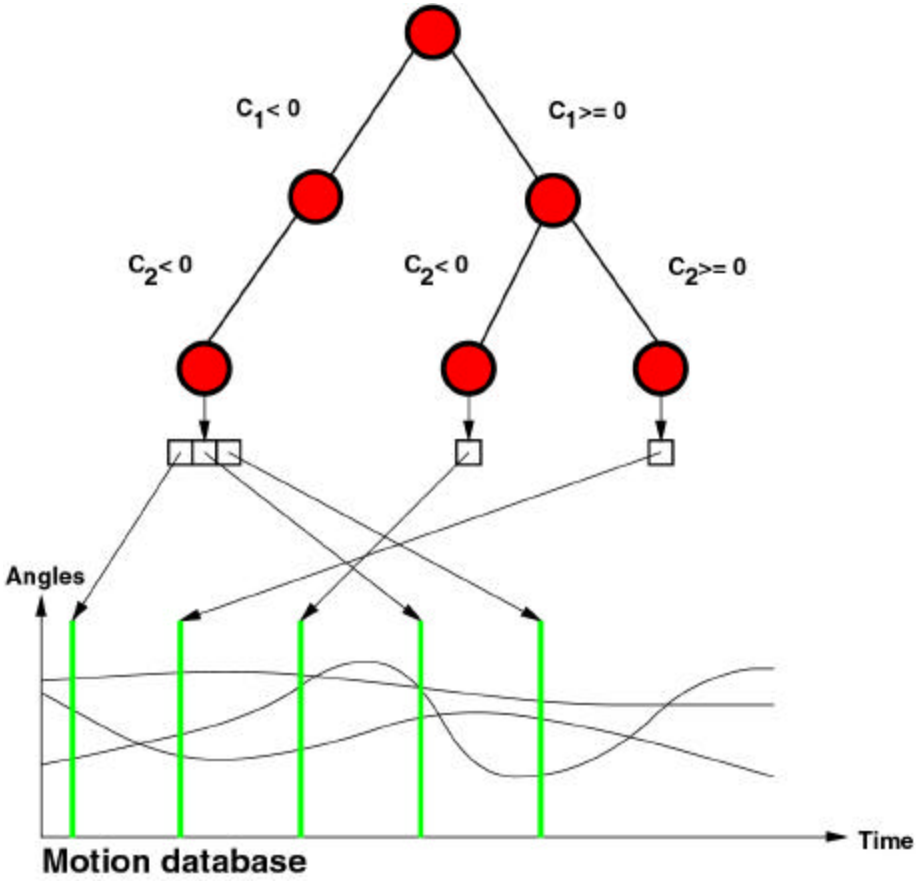
Approach:



Only need *samples*.

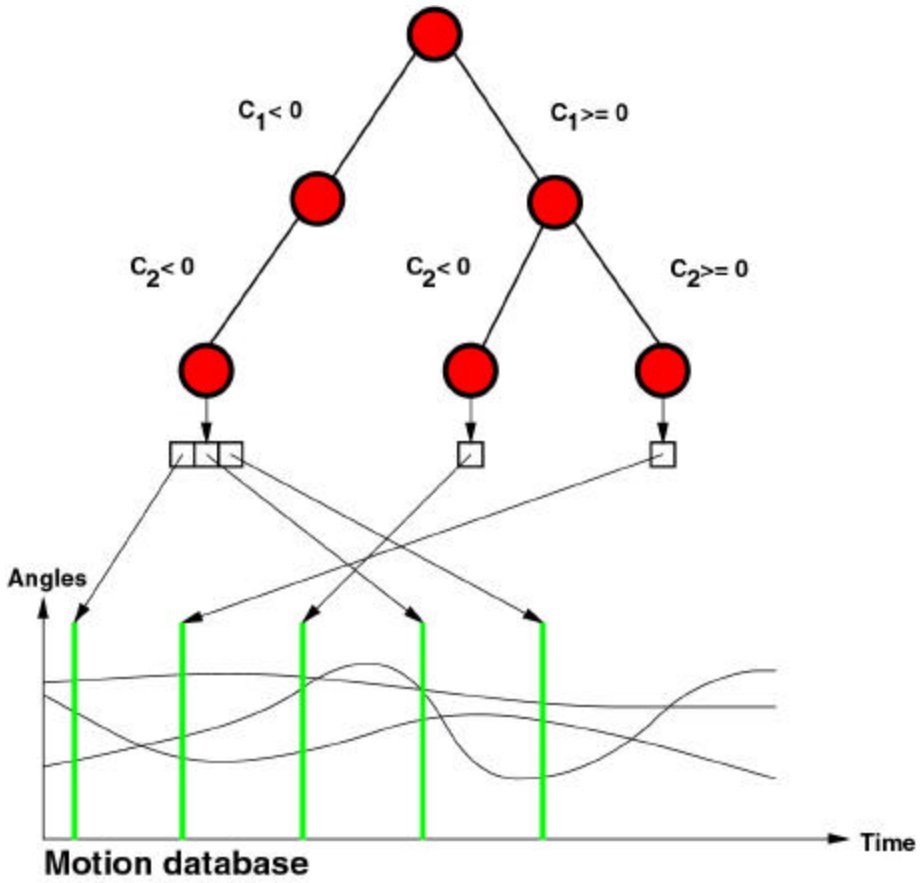
Trade accuracy for speed of sampling.

Probabilistic Database Search



Each level in the tree corresponds to one PCA coefficient *l*.

Sort each motion example *i* into a tree: Left subtree for negative value of $c_{l,i}$, right for positive value.



Probabilistic Database Search

$$p(\Psi_i \mid \Phi_t) \approx p(\mathbf{c}_i \mid \mathbf{c}_t)$$

Approximated by sampling from tree iteratively:

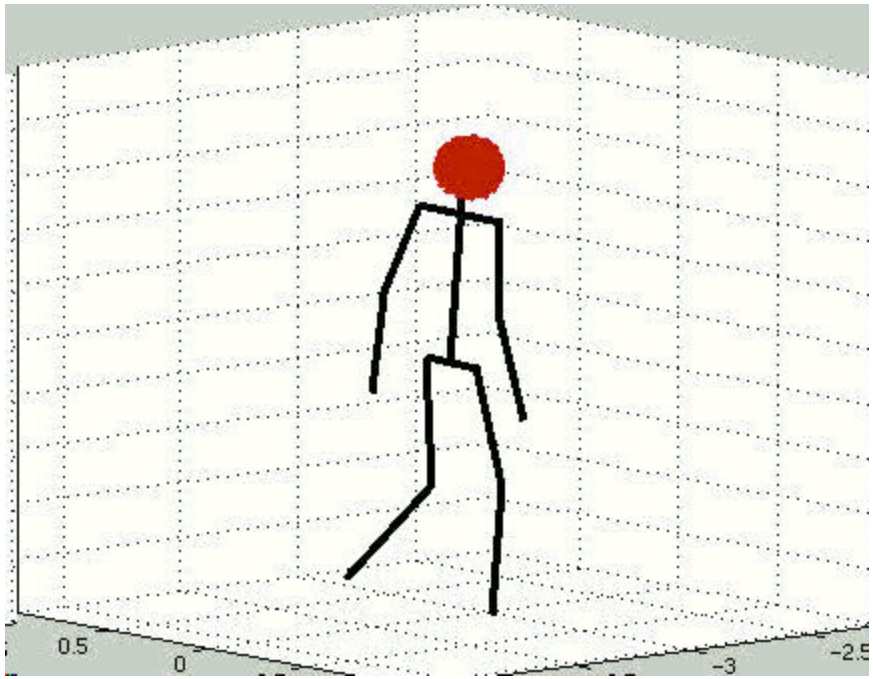
$$p_{l,right} = p(c_{l,i} \geq 0 \mid c_{l,t})$$

$$= \frac{1}{\sqrt{2pb}\, s_l} \int_{z=-\infty}^{c_{l,t}} \exp\left(-\frac{z^2}{2bs_l^2}\right) dz$$

$$p_{l,left} = p(c_{l,i} < 0 \mid c_{l,t})$$

Motion "Texture" Synthesis

Start with a small motion "chunk," sample to generate a new sequence of poses.



Walking

Changing color indicates new example sequence.

Bayesian Formulation

Posterior over model parameters given an image sequence.

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

Temporal model (prior)

Likelihood of observing the image features (filters responses) given the model parameters

Posterior from previous time instant

Monte Carlo integration

Inference/Search

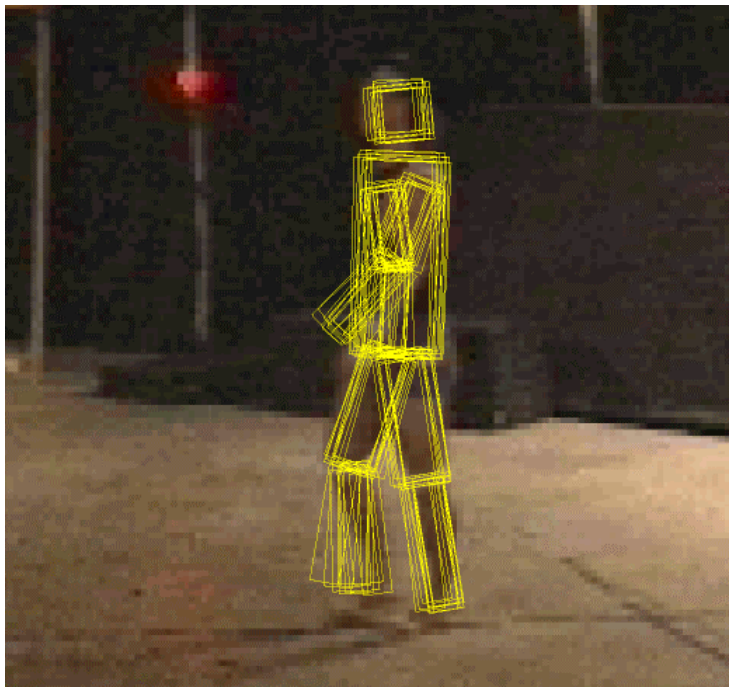
Bayesian formulation

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

1. **Model**: Kinematic tree. **Cues:** filter responses.
2. **Likelihood**: exploit *learned* likelihood of filter responses conditioned on the projected model.
3. **Prior**: statistical model, *learned* from examples.
4. **Search**: discretize intelligently using factored sampling and search using a particle filter.

Key Idea: Represent Ambiguity

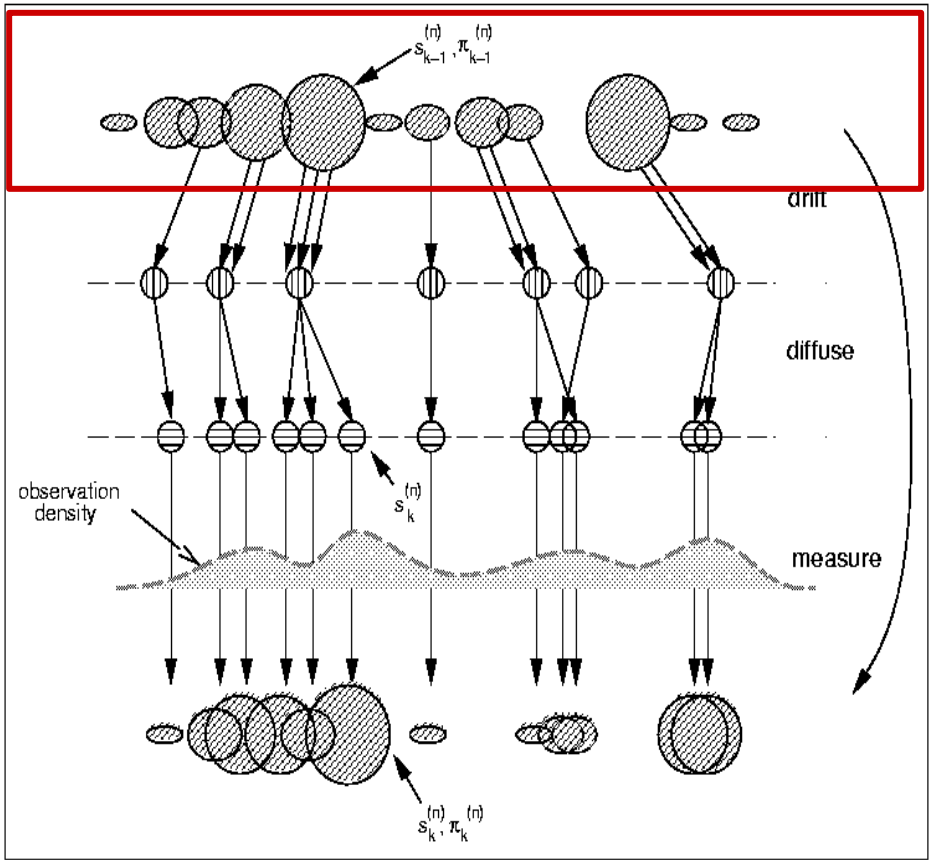
* Represent a multi-modal posterior probability distribution over model parameters
  - sampled representation
  - each sample is a pose and its probability (likelihood weighting)
  - predict over time using a *particle filter*.



Samples from a distribution over 3D poses.

Particle Filter

Posterior $p(f_{t-1} \mid \vec{\mathrm{I}}_{t-1})$

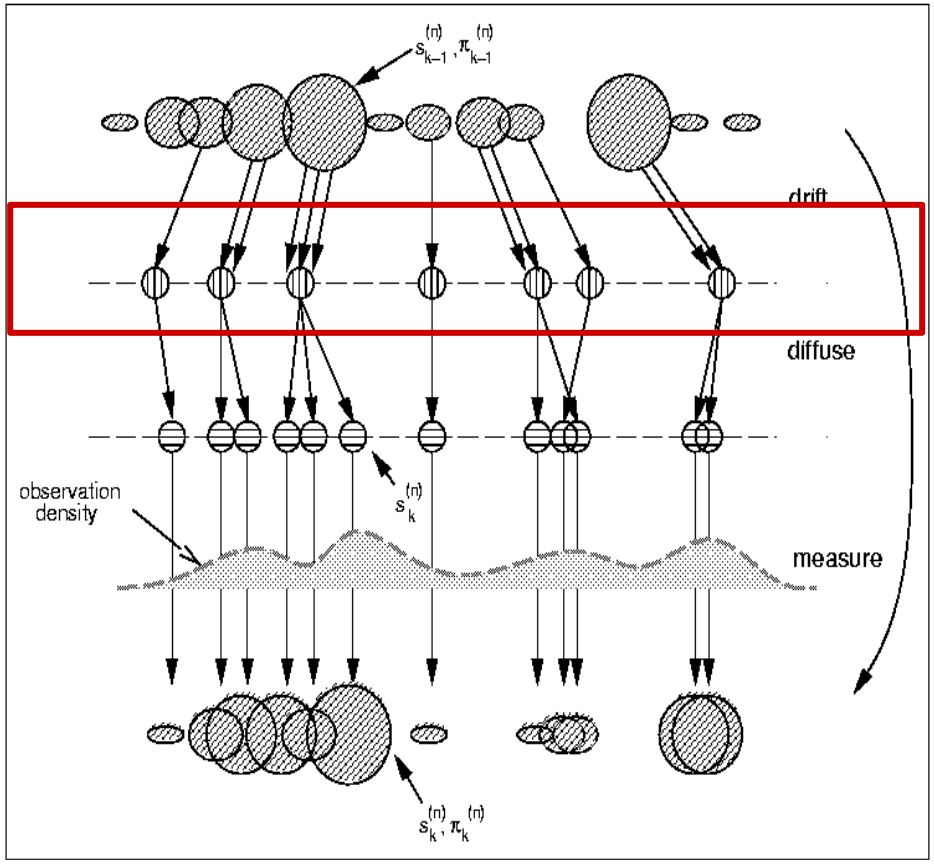


Isard & Blake ‘96

Particle Filter

Posterior $p(f_{t-1} \mid \vec{\mathrm{I}}_{t-1})$

sample



Isard & Blake '96

Particle Filter

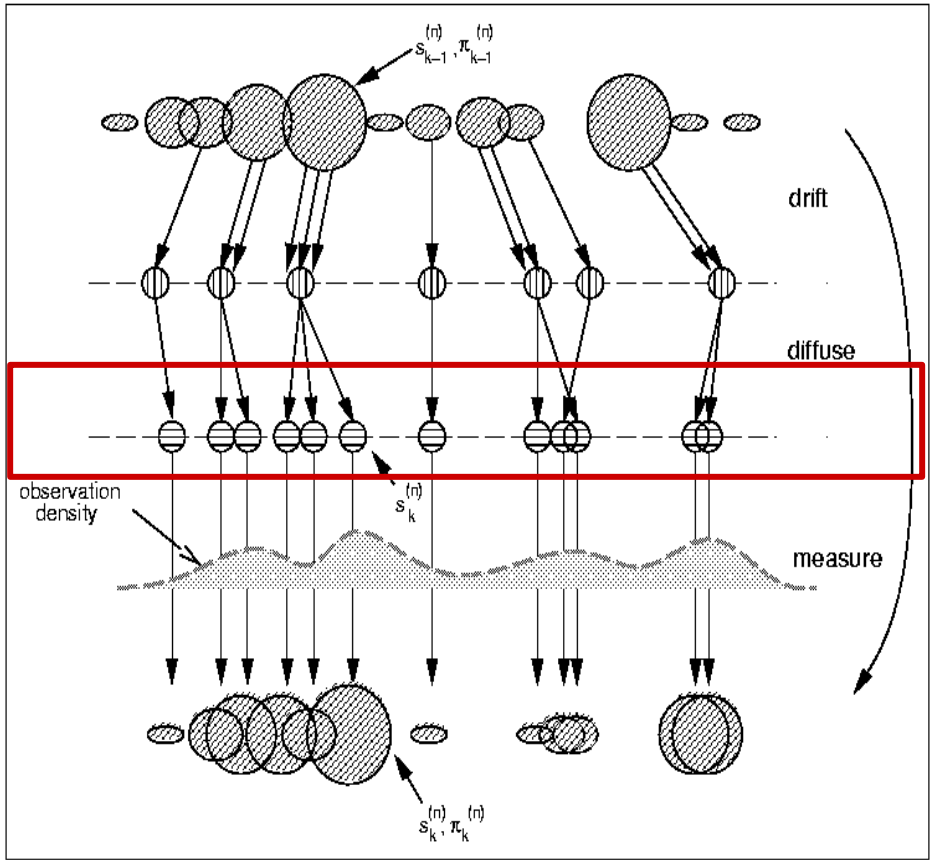
Posterior $p(f_{t-1} \mid \vec{\mathrm{I}}_{t-1})$

sample

Temporal dynamics

$p(f_t \mid f_{t-1})$

sample



Isard & Blake '96

Particle Filter

Posterior $p(f_{t-1} \mid \vec{\mathrm{I}}_{t-1})$

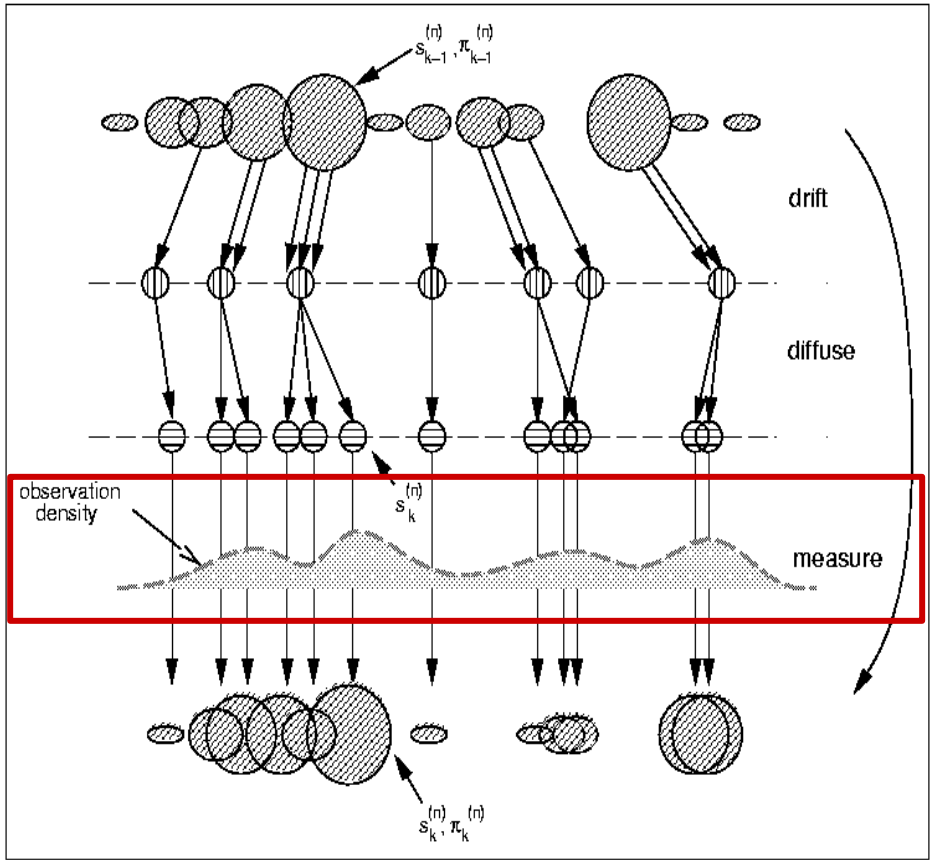
sample

Temporal dynamics

$p(f_t \mid f_{t-1})$

sample

Likelihood $p(\mathrm{I}_t \mid f_t)$



Isard & Blake '96

Particle Filter

Posterior $p(f_{t-1} \mid \vec{\mathrm{I}}_{t-1})$

sample

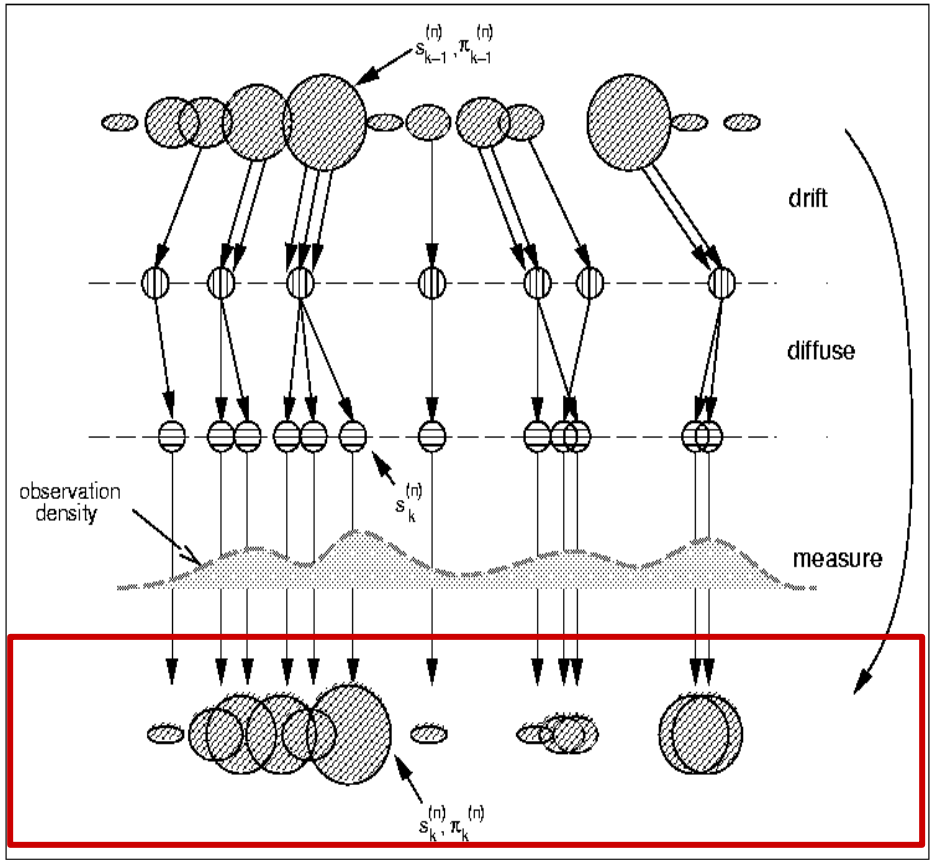
Temporal dynamics $p(f_t \mid f_{t-1})$

sample

Likelihood $p(\mathrm{I}_t \mid f_t)$

normalize

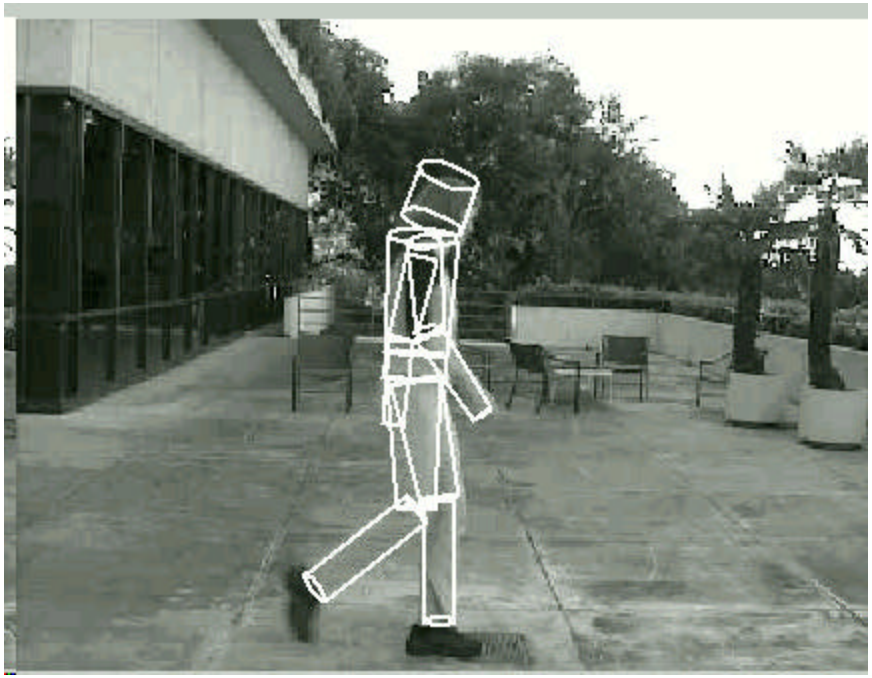
Posterior $p(f_t \mid \vec{\mathrm{I}}_t)$



Isard & Blake '96

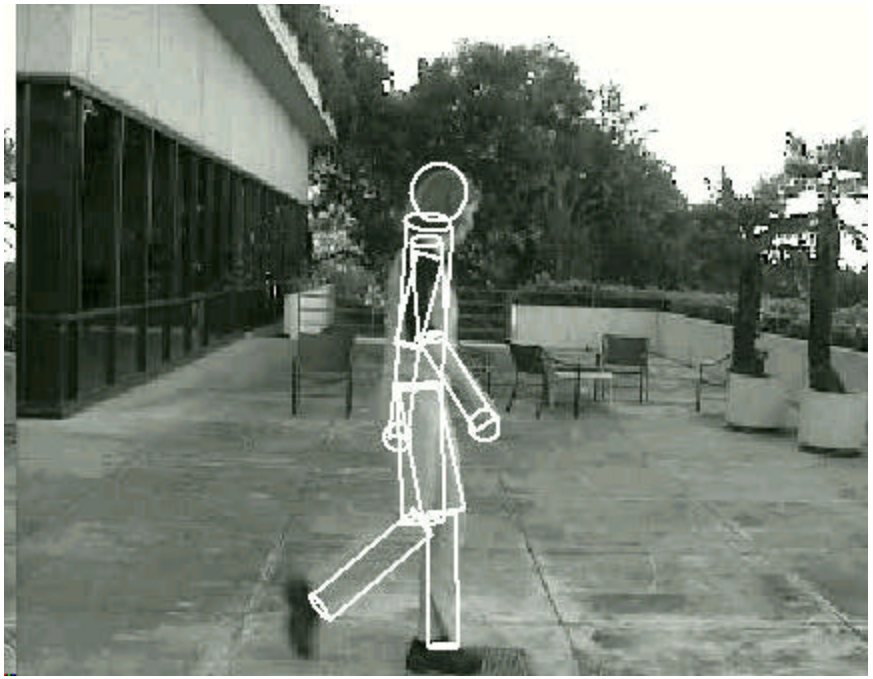
Stochastic 3D Tracking

monocular sequence



* 2500 samples.
* circa 2000.

How Strong is the Prior?



* Learned walking prior.
* No likelihood = hallucination.

Related Work

Cham & Rehg ‘99

* Single camera, multiple hypotheses.
* 2D templates (no change in viewpoint)
* Manual initialization.

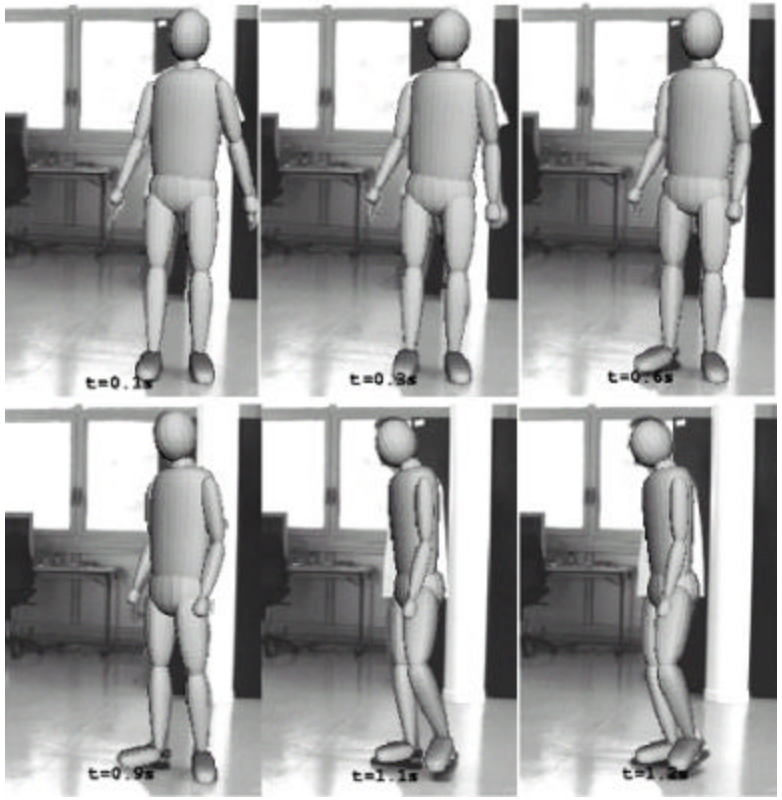
Related Work

Deutscher, North, Bascle, & Blake ‘99

* multiple cameras
* simplified clothing and light
* manual initialization
* weak prior
* “annealed” particle filter

Related Work

* manual initialization
* monocular
* complex background
* multiple cues (motion, edges, motion discontinuities)
* weak prior
* careful attention paid to the optimization problem.



Sminchisescu &Triggs ‘01

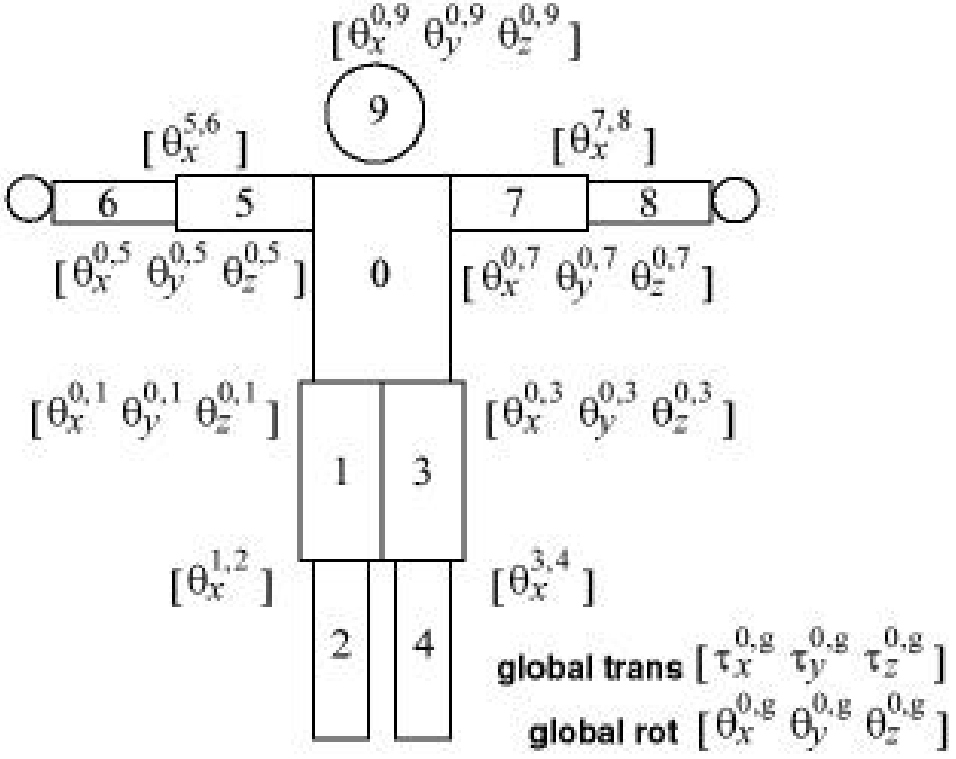
Are we done?

Current systems:

* require manual initialization
* are brittle (can't re-initialize)
* can't easily exploit robust, bottom-up, detectors

The search space is huge.

Particle filtering on the whole space requires strong priors or huge numbers of samples.



Kinematic Tree

Traditional kinematic tree

(dogma)

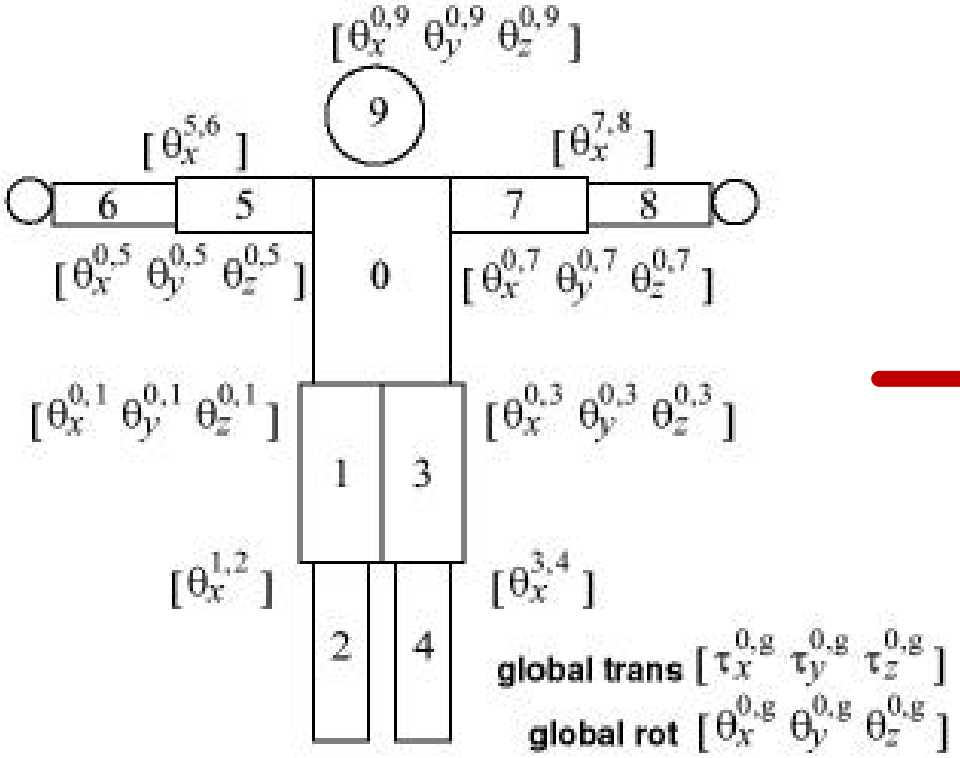
* brittle if it does not fit perfectly.

* starts with torso which is hardest to find.

* faces, hands, and feet may be easier to find.

* these are defined wrt the other limbs – results in a full high-dimensional search to fit bottom up measurements.

Attractive People



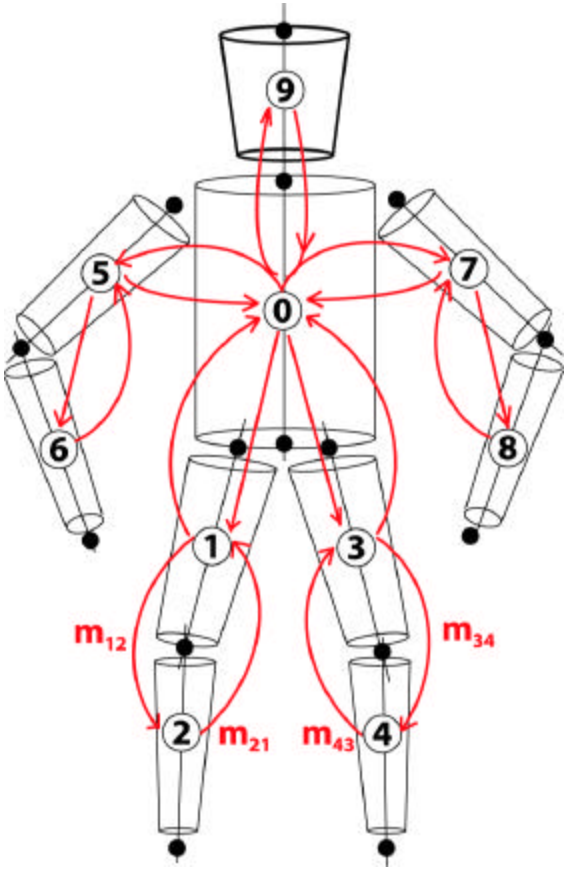
Traditional kinematic tree

(dogma)

"Push puppet" toy

(dog)

Attractive People

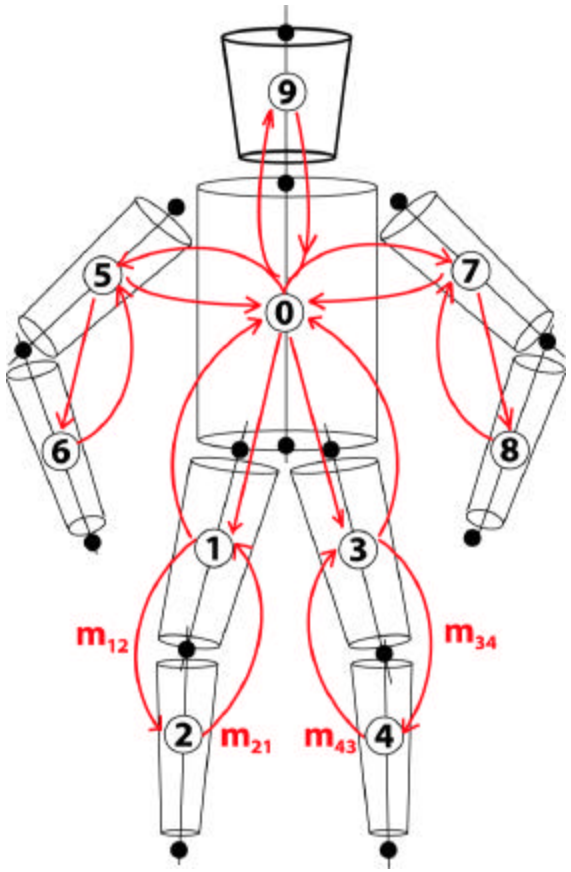


Loose-limbed body
(graphical model)

"Push puppet" toy

Loosely-Jointed Bodies

(with Michael Isard and Leon Sigal)



Soft constraints (messages) between limbs.

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

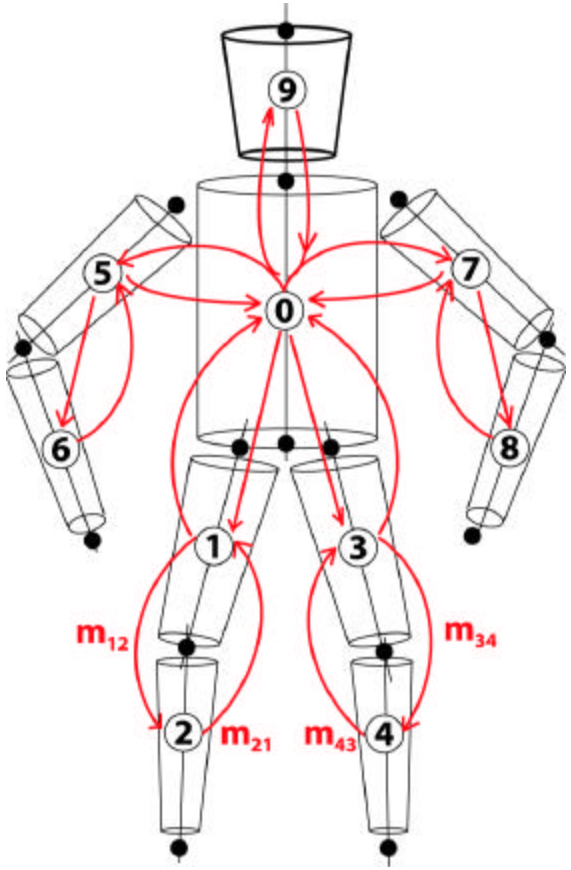
Allows bottom up initialization.

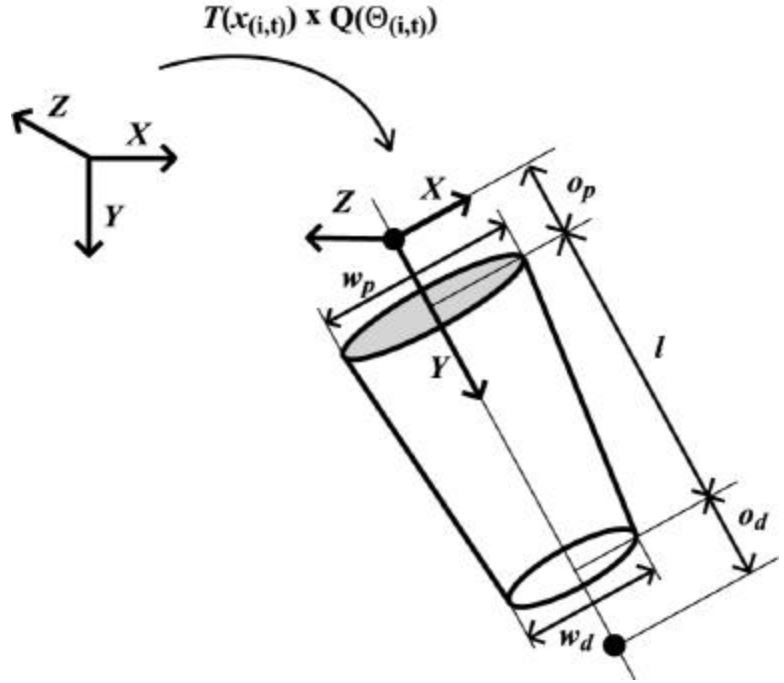
Deals well with unobserved limbs.

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

Loosely-Jointed Bodies

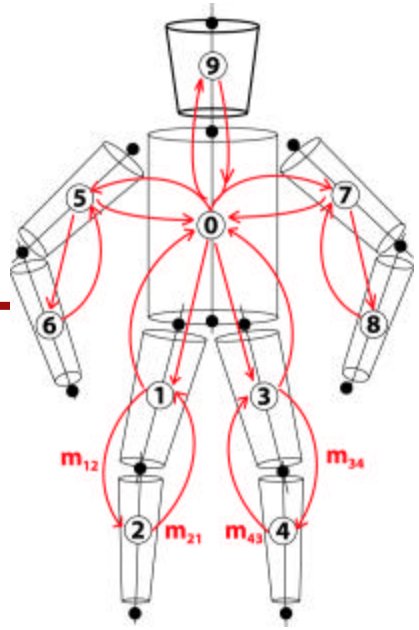
(with Michael Isard and Leon Sigal)





* 6D (position & orientation) discretization not practical

Belief



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

local evidence

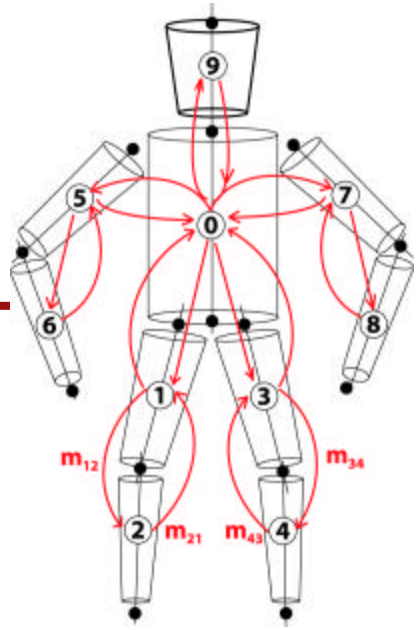
$$p(\mathbf{X}_i \mid \mathbf{Y}) = a\, l(\mathbf{Y}_i \mid \mathbf{X}_i) \prod_{k \in A_i} m_{ki}(\mathbf{X}_i)$$

observations (image and filter responses)

neighbors of node *i*

incoming messages

Messages



message from node i to node j

local evidence

$$m_{ij}(\mathbf{X}_j) = a \int y_{ij}(\mathbf{X}_i, \mathbf{X}_j)\, l(\mathbf{X}_i) \prod_{k \in A_i \setminus j} m_{ki}(\mathbf{X}_i)\, d\mathbf{X}_i$$

potential from i to j (prob of $\mathbf{X}_j$ conditioned on $\mathbf{X}_i$)

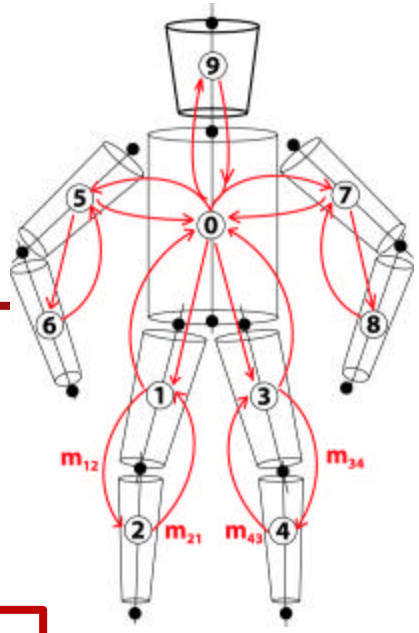
("spatial" or temporal prior)

neighbors of i, not including j

incoming messages at i

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

Messages



message "foundation"

$$m_{ij}(\mathbf{X}_j) = a \int y_{ij}(\mathbf{X}_i, \mathbf{X}_j) \boxed{l(\mathbf{X}_i) \prod_{k \in A_i \setminus j} m_{ki}(\mathbf{X}_i)} d\mathbf{X}_i$$

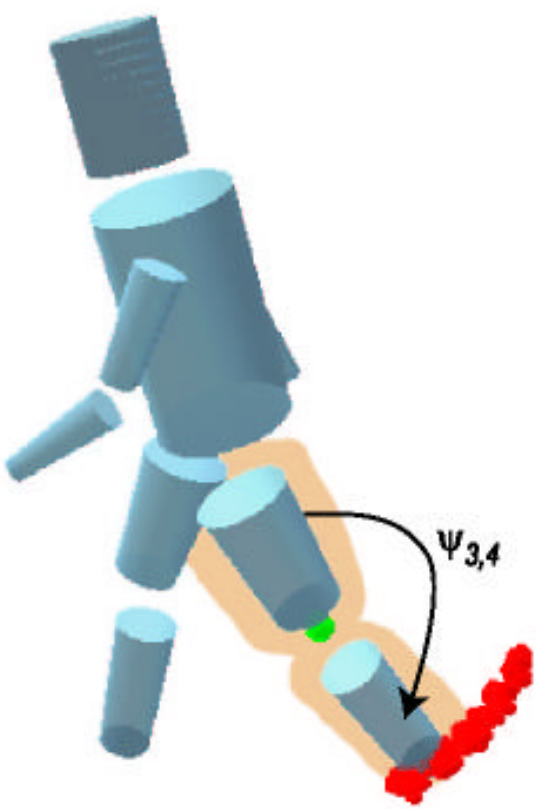
Monte Carlo integration.

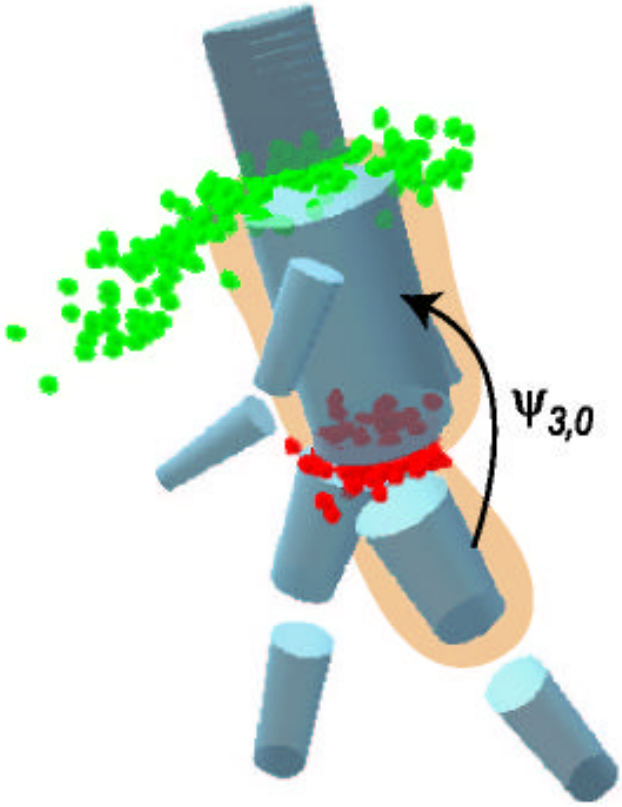
* draw samples from normalized foundation

* propagate through potential function

Potential Functions

"forward" "backward"

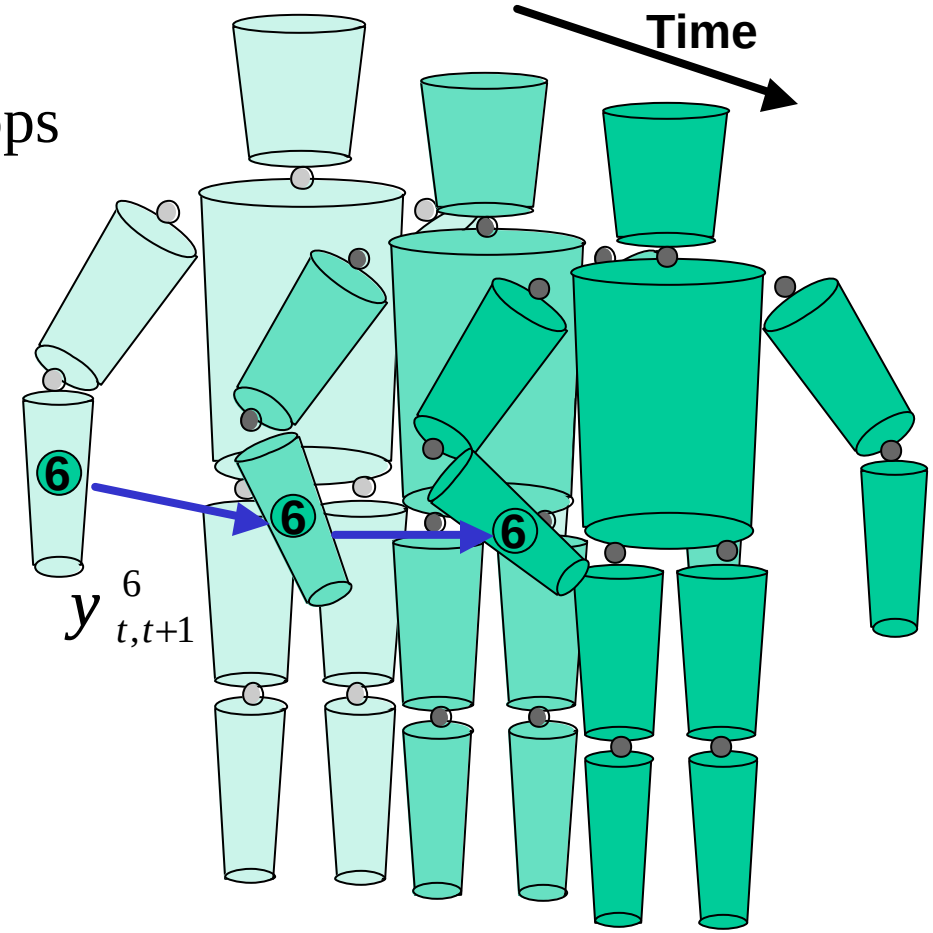




* represented by a mixture of Gaussians (fit to mocap data).

Temporal Potentials

Introduces loops



* we also define potentials backwards/forwards in *time*.

Approach

Problems:

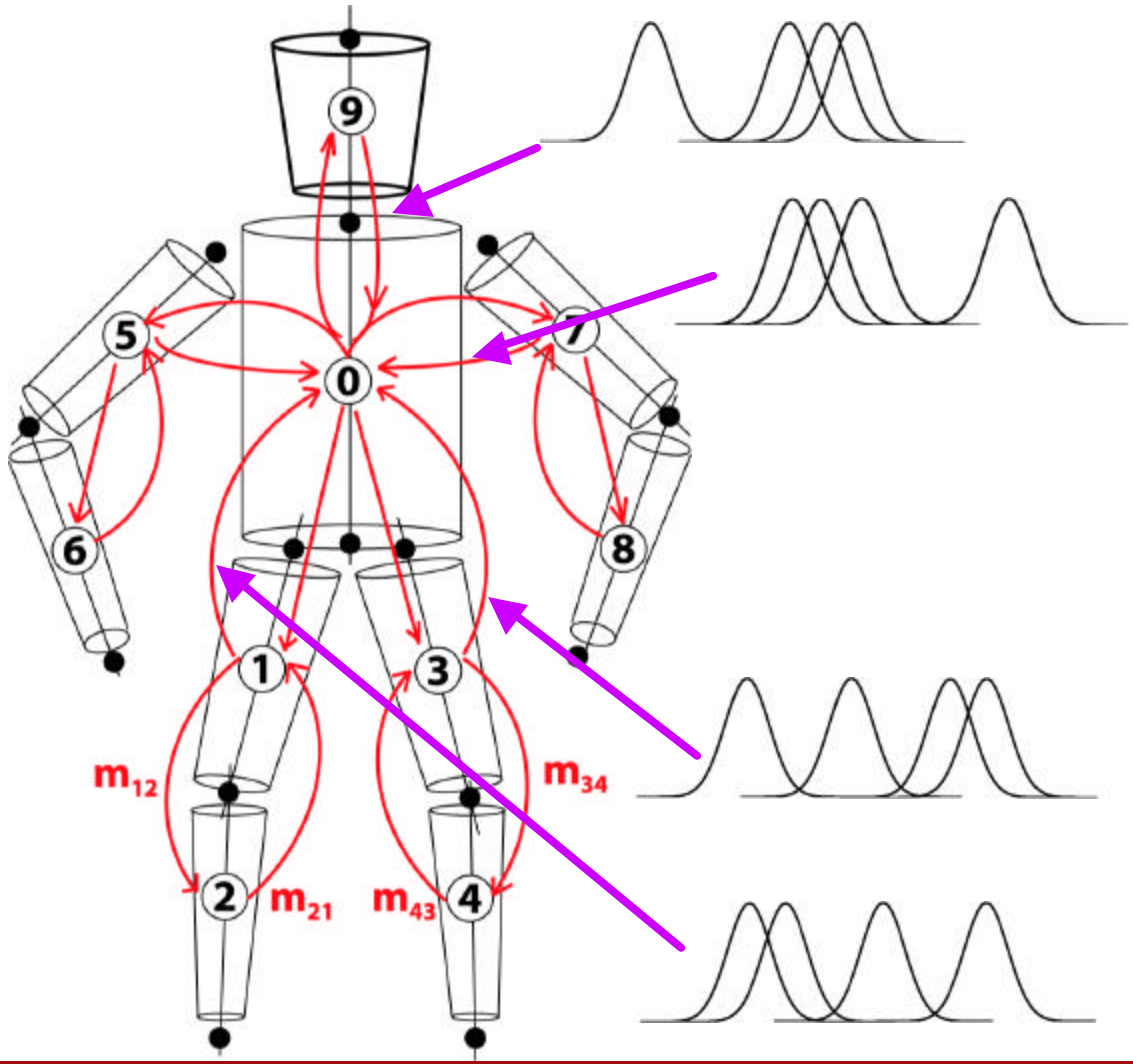
* state space is continuous and relatively high dimensional
* likelihoods are non-Gaussian and multi-modal
* relationships between limbs are complex (not Gaussian)

Approach:

* exploit particle set idea to represent messages
* Algorithm: *Non-parametric Belief Propagation* (Isard'03, Sudderth et al '03)

Algorithm Sketch

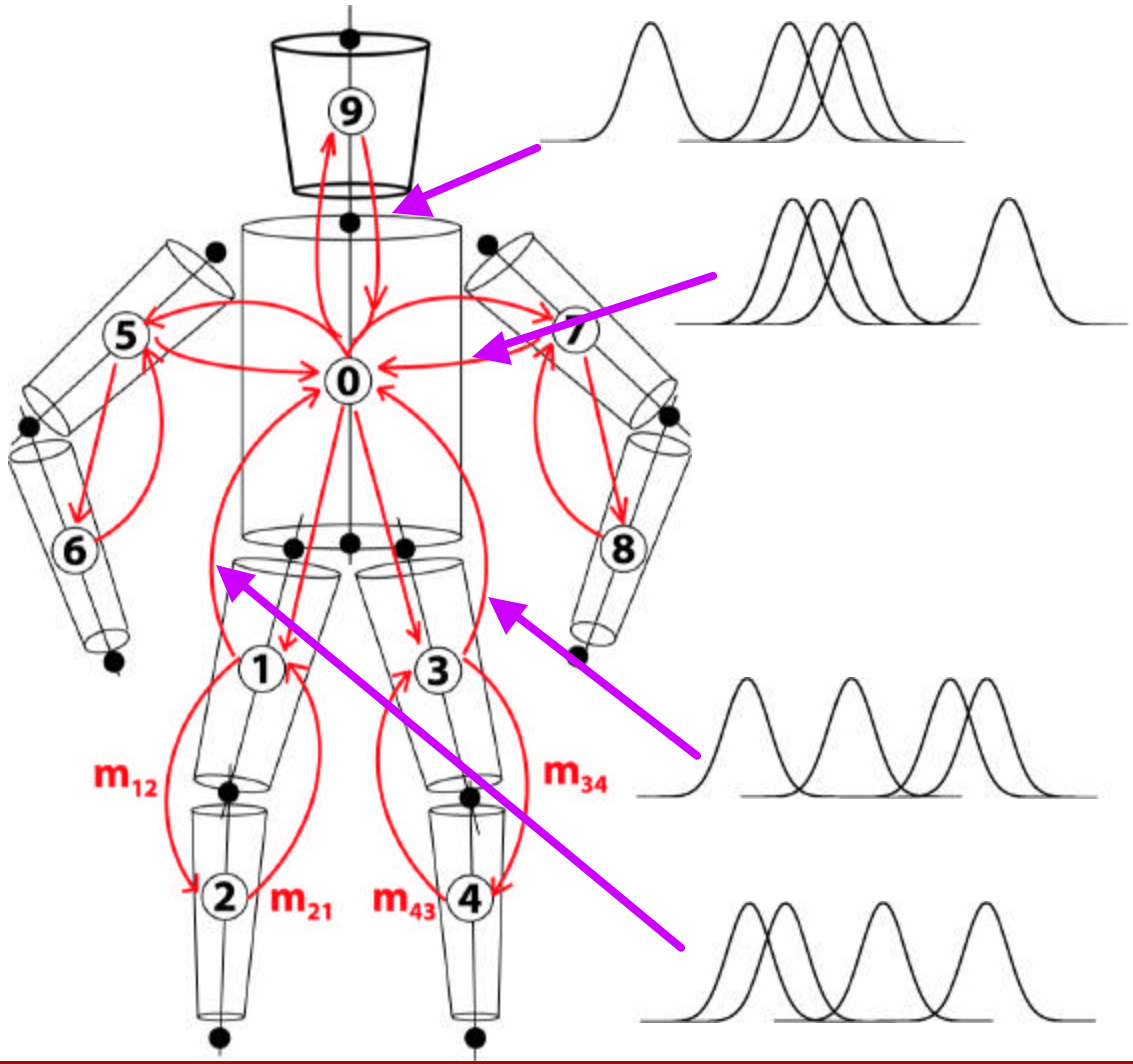
1. represent messages and beliefs by a discrete set of weighted samples (ie. mixture of narrow Gaussians).



Algorithm Sketch

2. compute product of incoming messages (also a mixture of Gaussians).

Product of d mixtures of n Gaussians: n^d

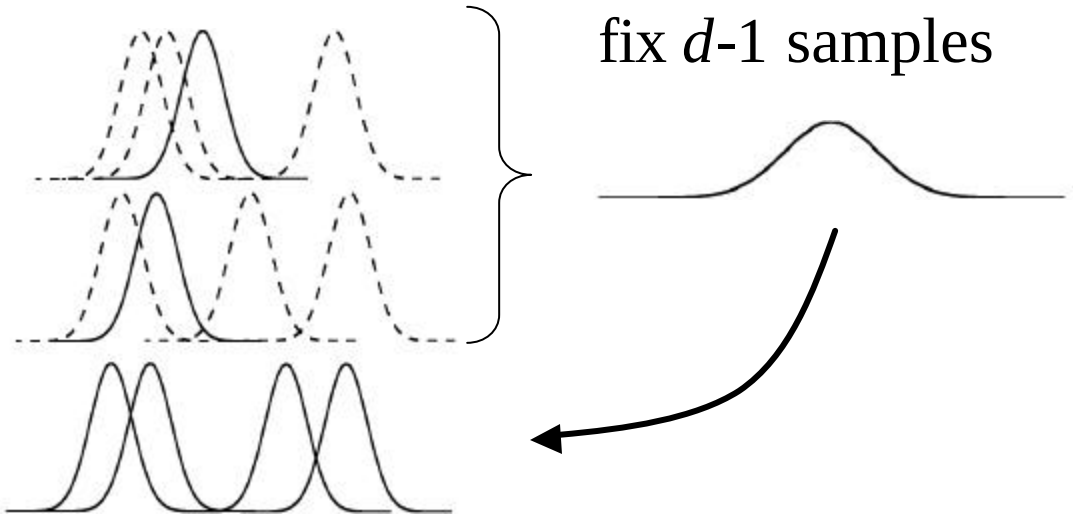


Algorithm Sketch

2. compute product of incoming messages (also a mixture of Gaussians).

Product of d mixtures of n Gaussians: n^d

Gibbs sampler (Sudderth et al):



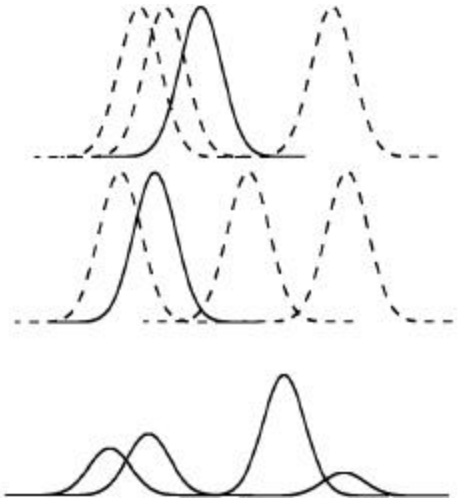
take the product with n Gaussians.

Algorithm Sketch

2. compute product of incoming messages (also a mixture of Gaussians).

Product of d mixtures of n Gaussians: n^d

Gibbs sampler (Sudderth et al):



weight the samples

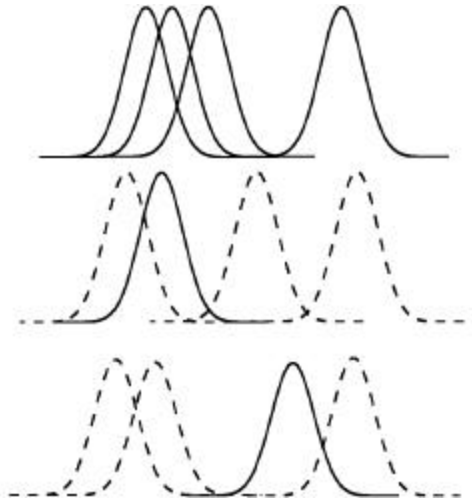
Algorithm Sketch

2. compute product of incoming messages (also a mixture of Gaussians).

Repeat until you've drawn n samples.

Cost: $O(dkn^2)$

Gibbs sampler (Sudderth et al):

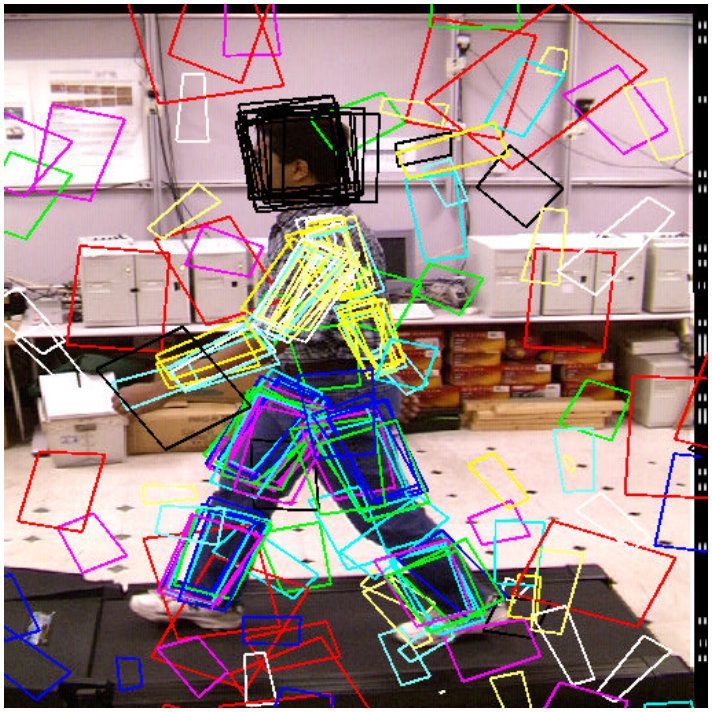


Sample to select a new Gaussian.

* Repeat with each message.
* Repeat the process k times.
* Take the product of the selected Gaussians and draw a sample.

Algorithm Sketch

3. to construct a message, we draw samples from a *proposal distribution* (including incoming message product, belief, and bottom up processes); importance re-weight.

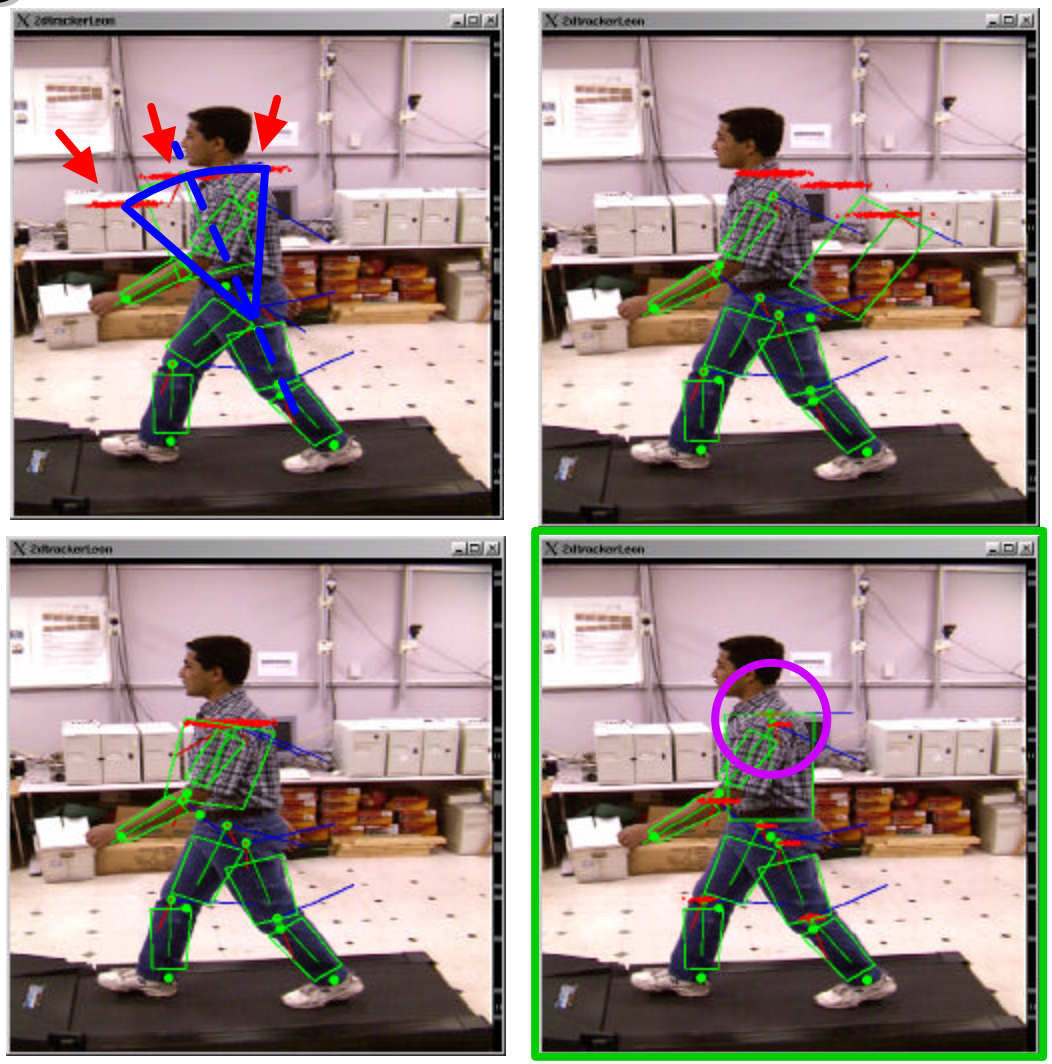


Noisy bottom up process (limb detector)

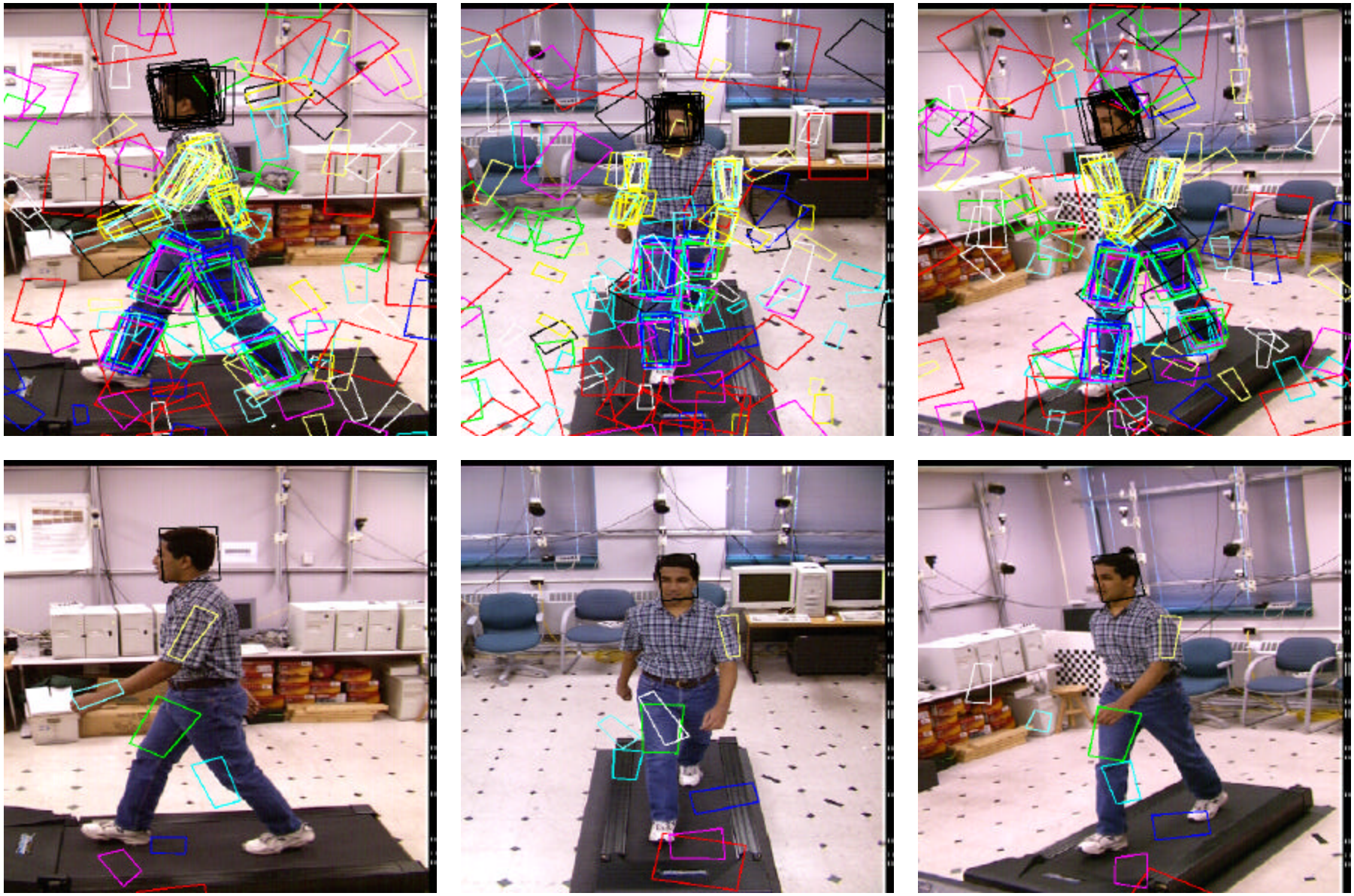
Algorithm Sketch

4. propagate samples through the potential to get new message.

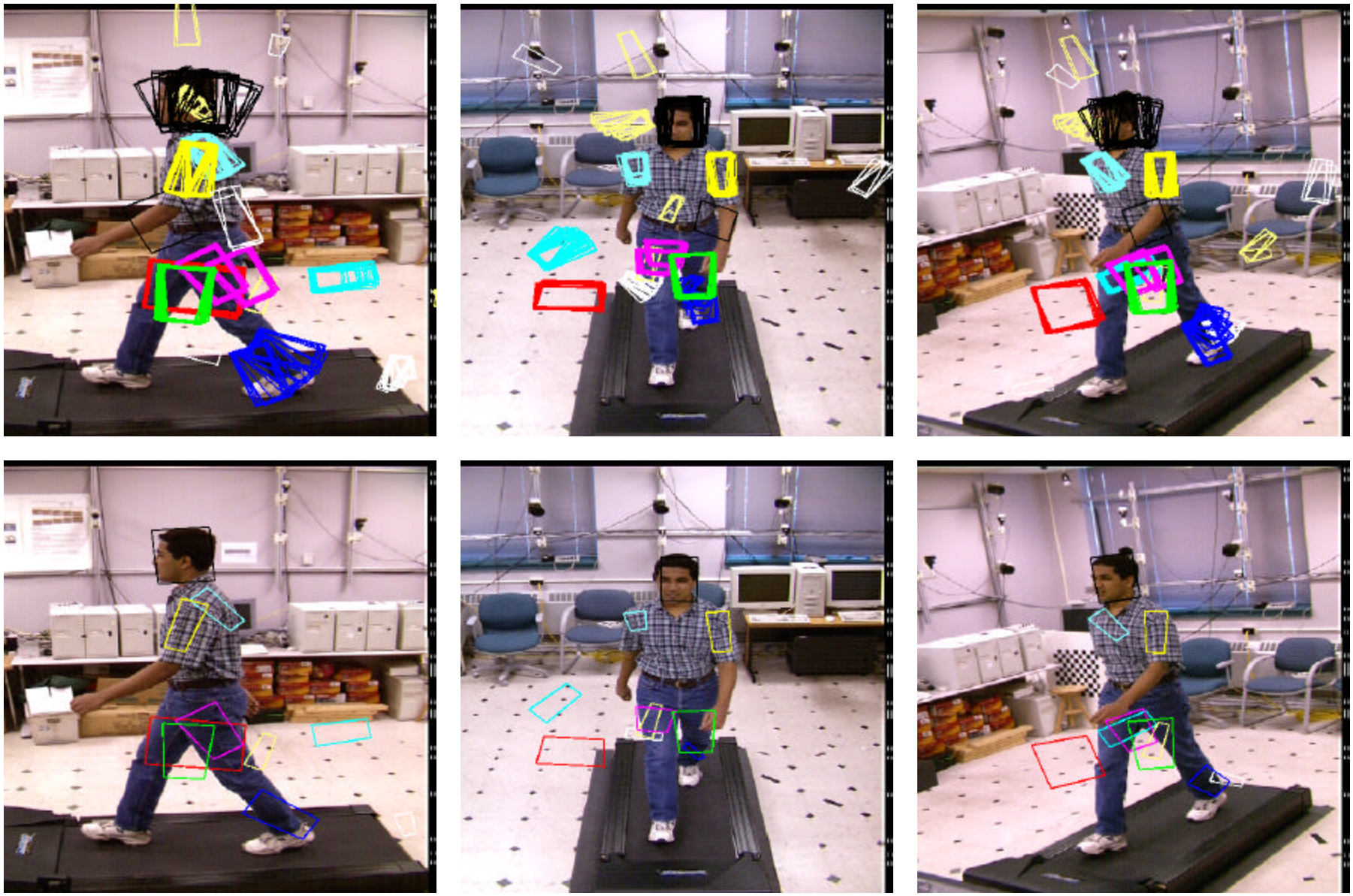
Repeat.



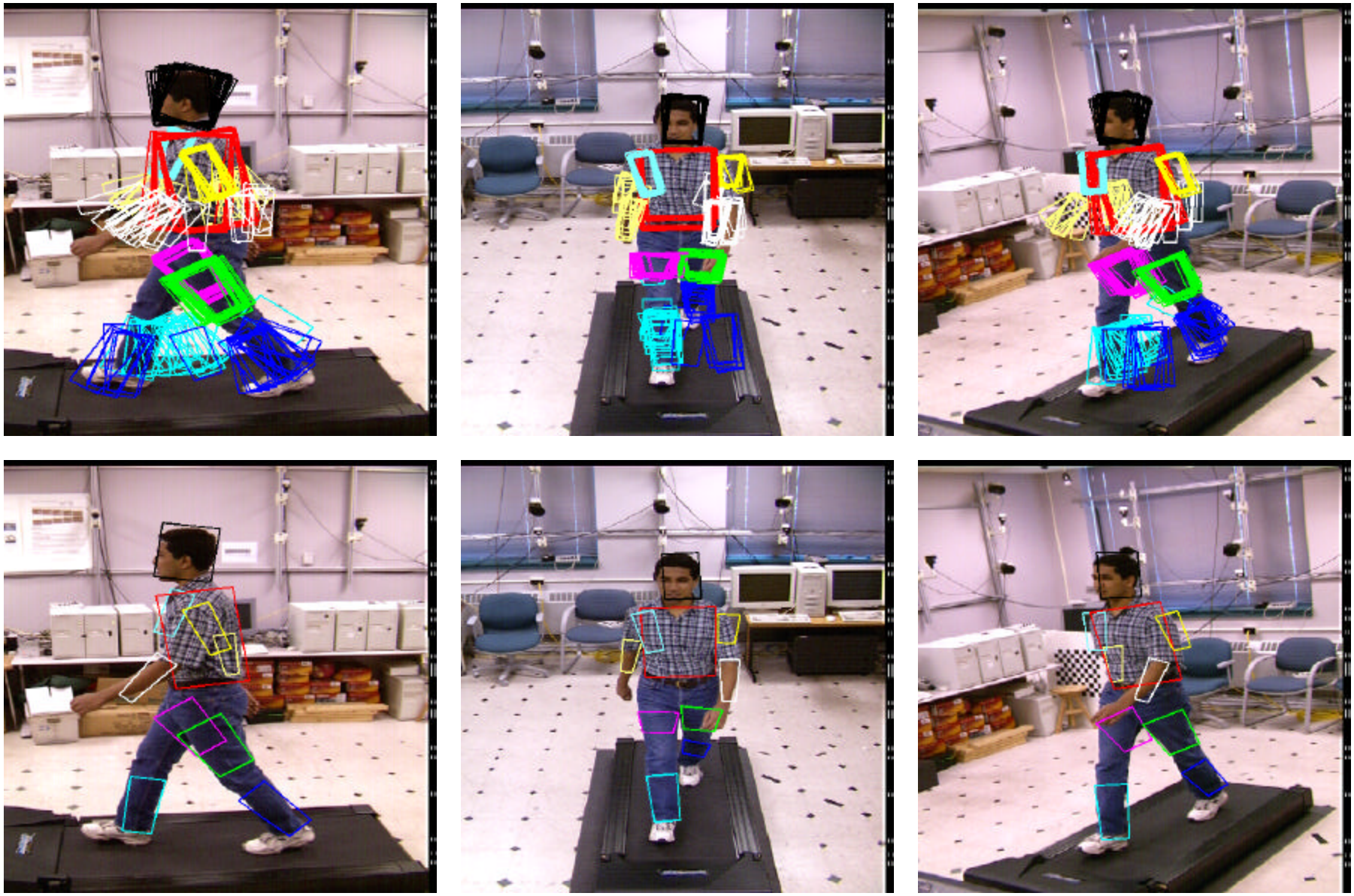
Iteration 0



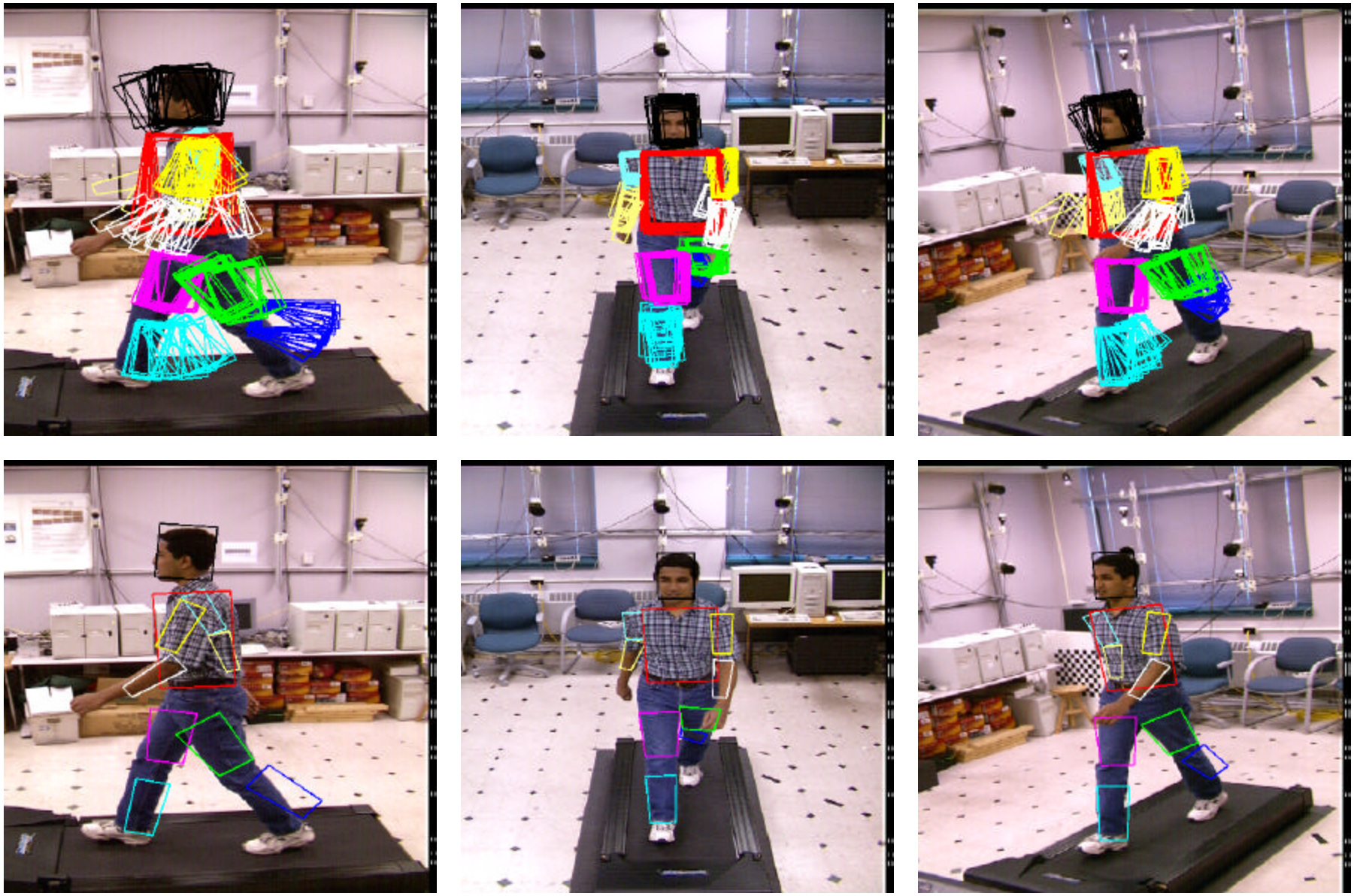
Iteration 1



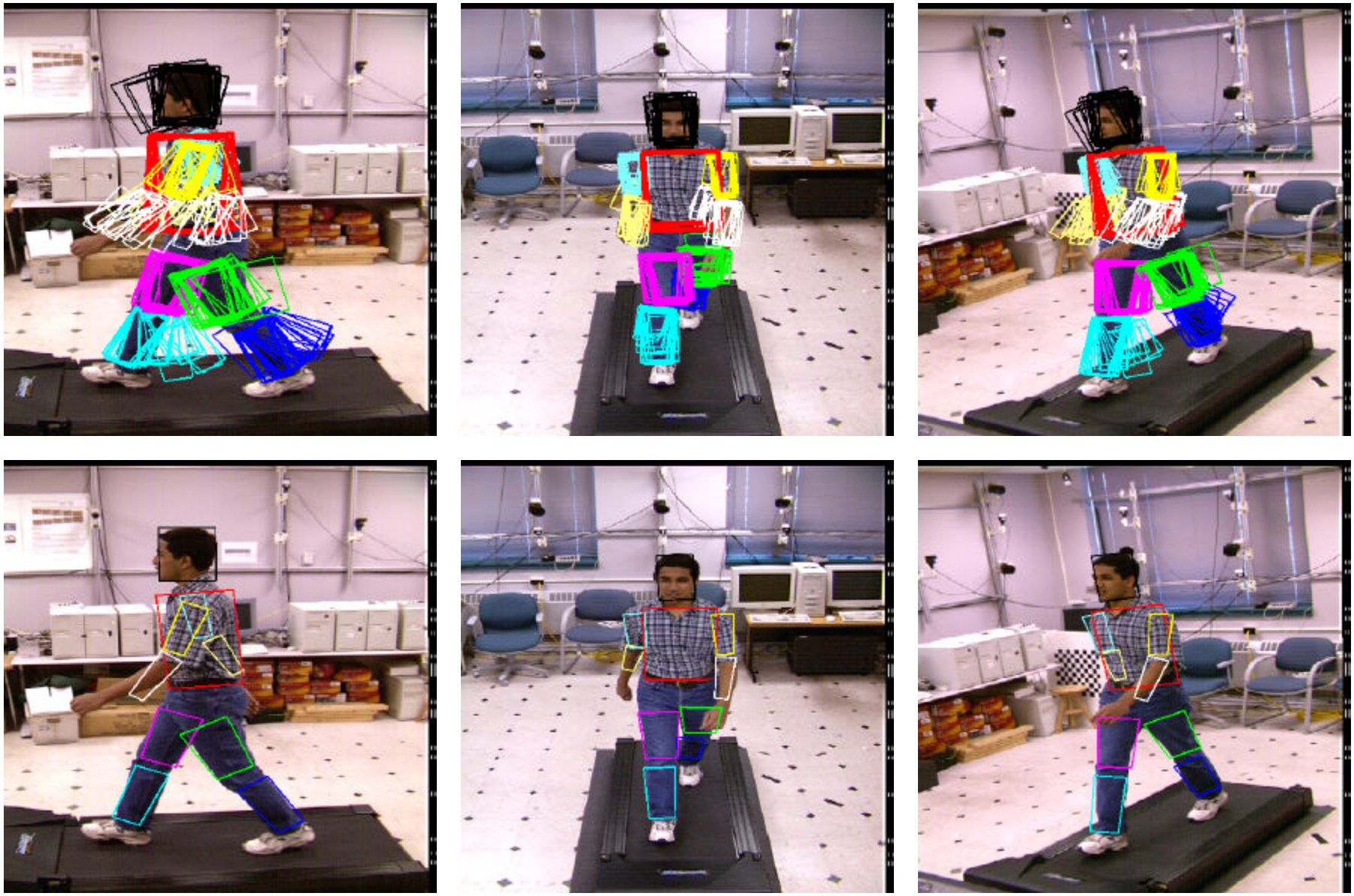
Iteration 2



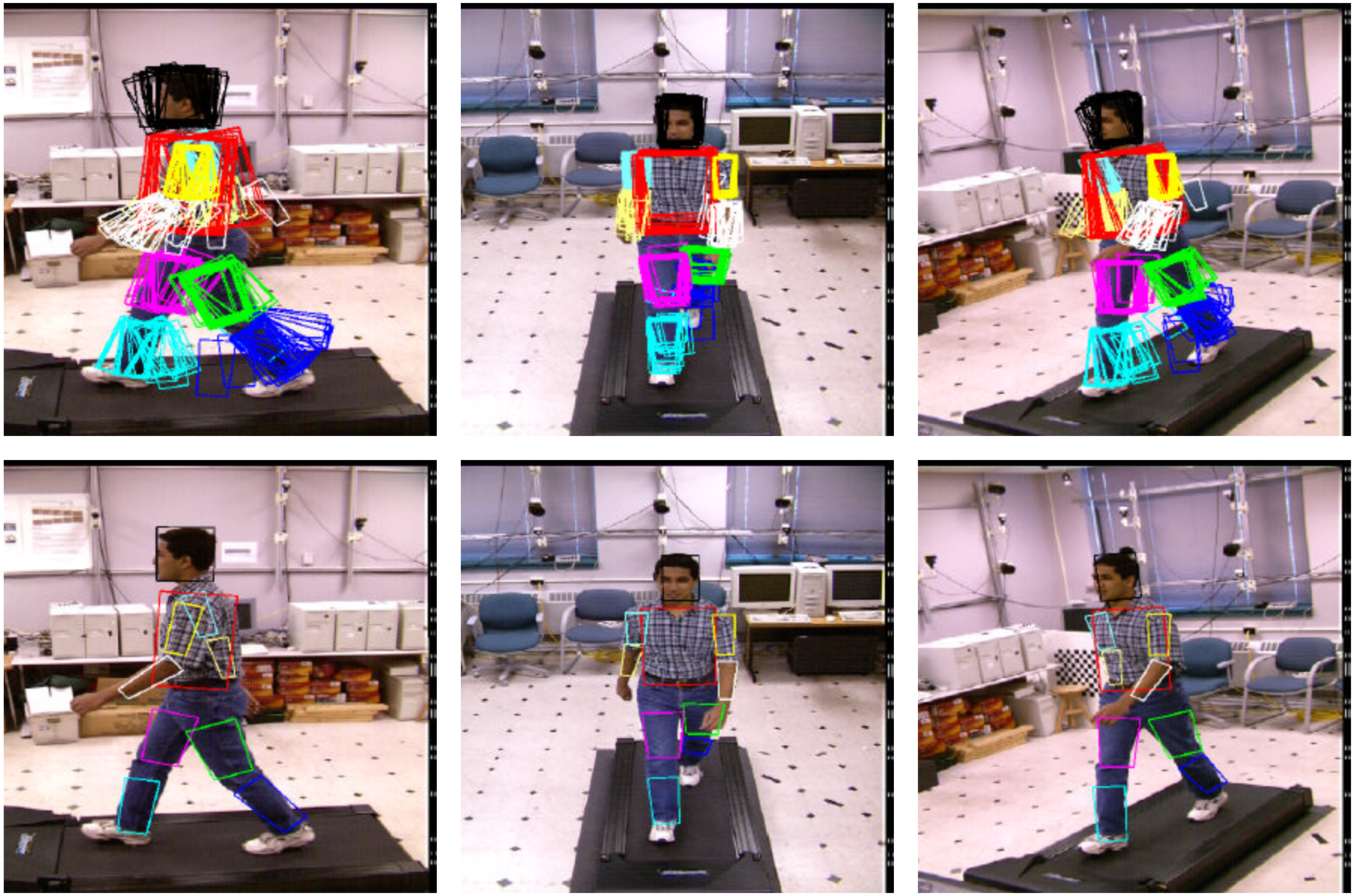
Iteration 3



Iteration 7



Iteration 10



Summary

We have tackled four important parts of the problem:

likelihood

1. Probabilistically modeling human appearance in a generic, yet constraining, way.

prior

2. Representing the range of possible motions using techniques from texture modeling.

search

3. Dealing with ambiguities and non-linearities using particle filtering for Bayesian inference.

initialization

4. Automatic initialization using Belief Propagation.

What Next?

* part detectors (faces, limbs, hands, feet, etc)
* interactions between non-adjacent limbs
    - introduces new nodes and *loops* to the graphical model
* new techniques for learning probabilistic models in high dimensional spaces with limited training data.

Outlook

5-10 years:

- Relatively reliable people tracking in monocular video
- Accurate with multiple cameras
- Path is pretty clear.

Next step: *Beyond person-centric*

- people interacting with object/world

… solve the vision problem.

Beyond that: Recognizing action

- goals, intentions, ...

… solve the AI problem.