Human Pose Tracking I: Basics

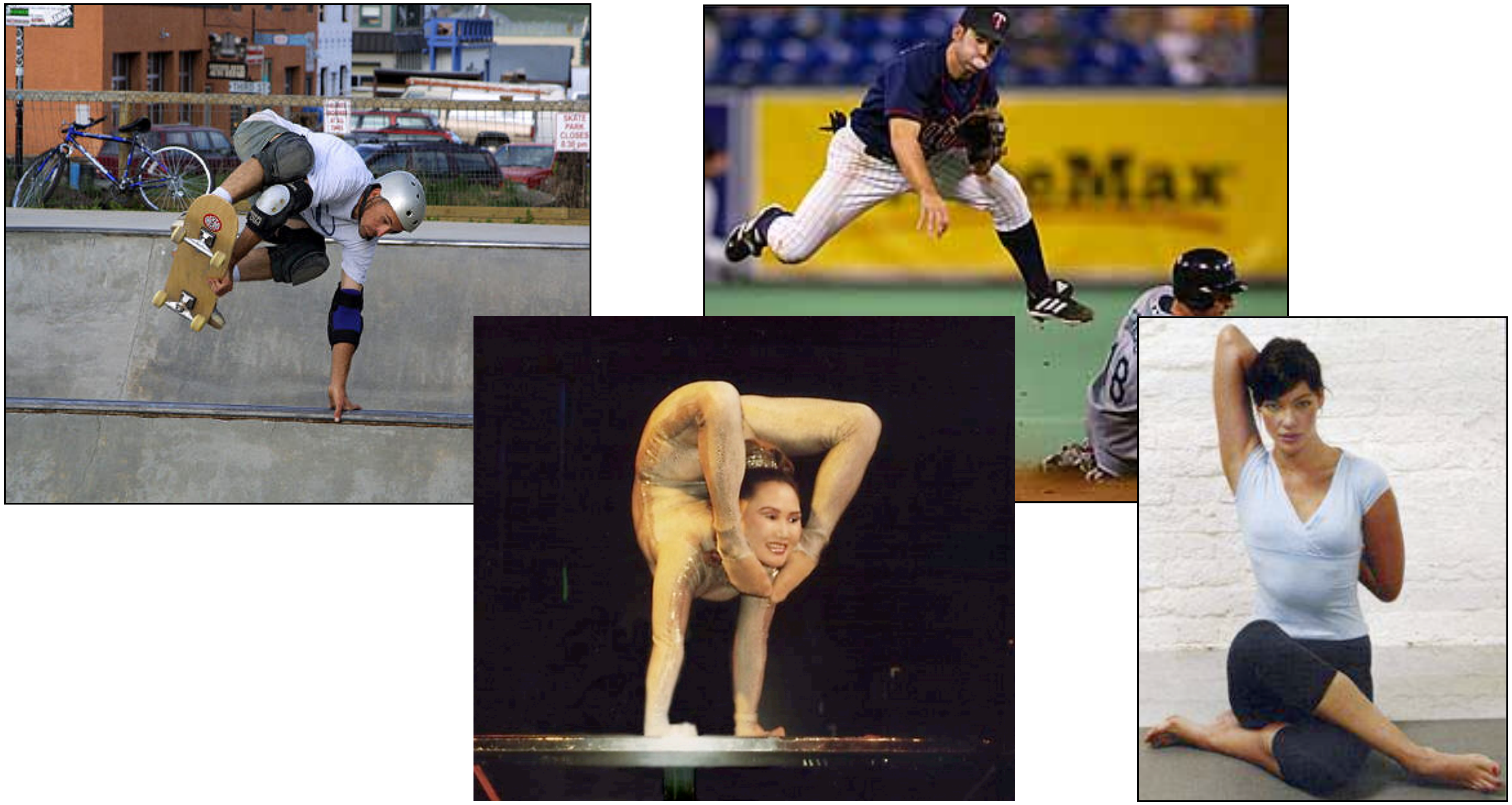
David Fleet

University of Toronto

CIFAR Summer School, 2009

Looking at People

Challenges: Complex pose / motion

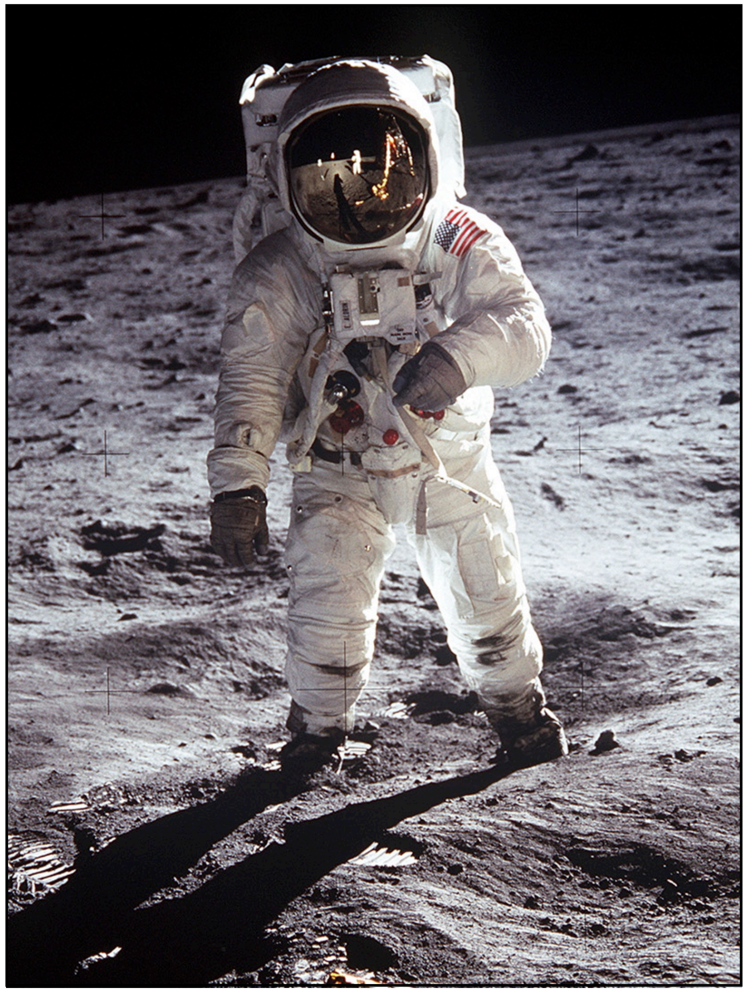


People have many degrees of freedom, comprising an articulated skeleton overlaid with soft tissue and deformable clothing.

Challenges: Complex movements

People move in complex ways, often communicating with subtle gestures

Challenges: Appearance, size and shape



People come in all shapes and sizes, with highly variable appearance.

Challenges: Appearance variability

Image appearance changes dramatically over time
due to non-rigidity of body and clothing and lighting.

Challenges: Appearance variability

Image appearance changes dramatically over time
due to non-rigidity of body and clothing and lighting.

Challenges: Context dependence

Perceived scene context influences object recognition.

[Courtesy of Antonio Torralba]

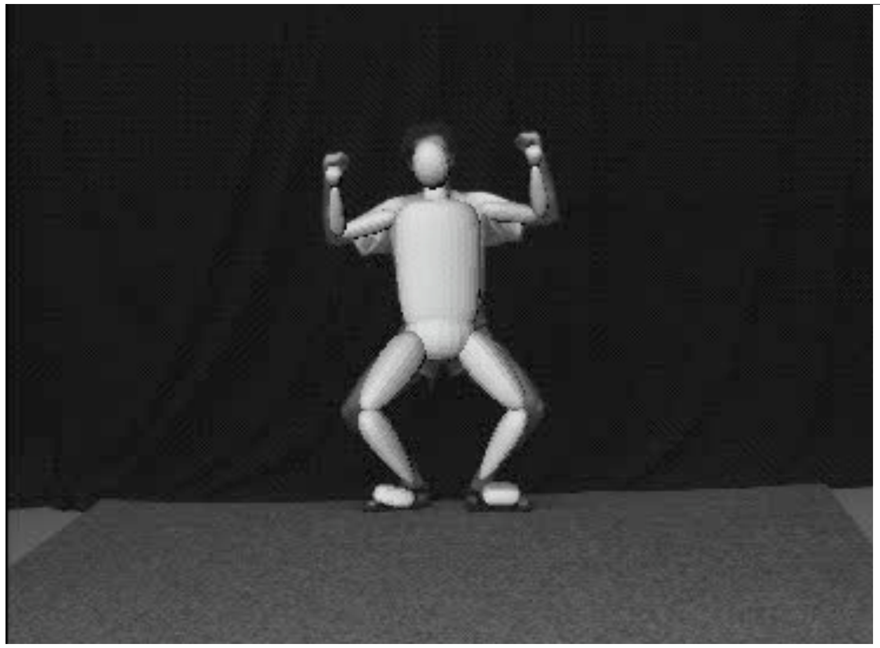
Challenges: Noisy and missing measurements

Ambiguities in pose are commonplace, due to

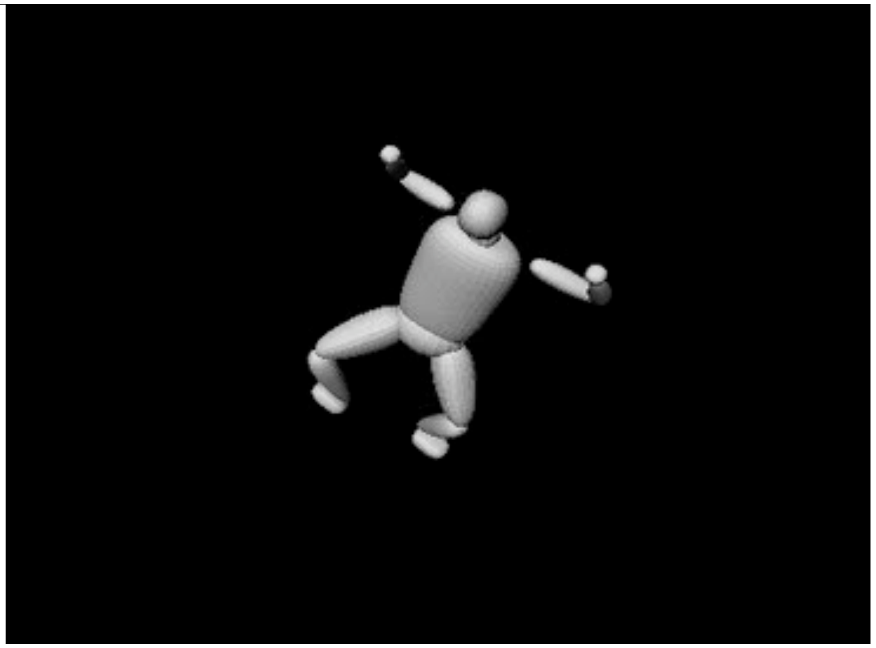
- background clutter
- apparent similarity of parts
- occlusions
- loose clothing
- ...

Challenges: Depth and reflection ambiguities

image



3D model
(camera view)



3D model
(top view)

Multiple 3D poses may be consistent with a given image.

[courtesy of Cristian Sminchisescu]

Model-based pose tracking

Video input

3D articulated model

Outline

- Introduction
- Bayesian Filtering
- Kinematic Motion Models
- Discriminative Pose Estimation
- Physics-Based Motion Models
- Concluding remarks

Bayesian Filtering

State: n-vector comprising variables to be inferred: $\mathbf{s}_t$

- continuous variables [eg., position, velocity, shape, size, …]
- discrete state variables [eg., # objects, gender, activity, …]
- state history: $\mathbf{s}_{1:t} = (\mathbf{s}_1, ..., \mathbf{s}_t)$

Observations: data from which we estimate state: $\mathbf{z}_t = f(\mathbf{s}_t)$

- observation history: $\mathbf{z}_{1:t} = (\mathbf{z}_1, ..., \mathbf{z}_t)$

Bayesian Filtering

Posterior distribution over states conditioned on observations

$$p(\mathbf{s}_{1:t} \,|\, \mathbf{z}_{1:t})$$

Bayes' rule:

likelihood prior

$$p(\mathbf{s}_{1:t} \,|\, \mathbf{z}_{1:t}) = \frac{p(\mathbf{z}_{1:t} \,|\, \mathbf{s}_{1:t})\, p(\mathbf{s}_{1:t})}{p(\mathbf{z}_{1:t})}$$

independent of state

Filtering distribution: marginal posterior at current time

$$p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t}) = \int_{\mathbf{s}_1} \int_{\mathbf{s}_{t-1}} p(\mathbf{s}_{1:t} \,|\, \mathbf{z}_{1:t})$$

Simplifying model assumptions

1st-order Markov model for state dynamics:

$$p(\mathbf{s}_t \,|\, \mathbf{s}_{1:t-1}) \;=\; p(\mathbf{s}_t \,|\, \mathbf{s}_{t-1})$$

so

$$p(\mathbf{s}_{1:t}) \;=\; \left(\prod_{j=2}^{t} p(\mathbf{s}_j \,|\, \mathbf{s}_{j-1}) \right) p(\mathbf{s}_1)$$

sequence prior

one-step transition prior

Conditional independence of observations

joint likelihood

$$p(\mathbf{z}_{1:t} \,|\, \mathbf{s}_{1:t}) \;=\; p(\mathbf{z}_t \,|\, \mathbf{s}_t)\, p(\mathbf{z}_{1:t-1} \,|\, \mathbf{s}_{1:t-1})$$

$$= \prod_{\tau=1}^{t} p(\mathbf{z}_\tau \,|\, \mathbf{s}_\tau)$$

likelihood at time τ

Recursive form of filtering/posterior distribution

Filtering distribution:

$$p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t}) \;=\; \int_{\mathbf{s}_1} \cdots \int_{\mathbf{s}_{t-1}} p(\mathbf{s}_{1:t} \,|\, \mathbf{z}_{1:t})$$

$$=\; c\; \underbrace{p(\mathbf{z}_t \,|\, \mathbf{s}_t)}_{\text{likelihood}}\; p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t-1})$$

Prediction distribution (temporal prior):

$$p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t-1}) \;=\; \int_{\mathbf{s}_{t-1}} p(\mathbf{s}_t \,|\, \mathbf{s}_{t-1})\, p(\mathbf{s}_{t-1} \,|\, \mathbf{z}_{1:t-1})$$

Bayesian smoothing

Inverting the dynamics permits inference backwards in time:

$$\begin{aligned} p(\mathbf{s}_\tau \,|\, \mathbf{z}_{\tau:t}) &= c\, p(\mathbf{z}_\tau \,|\, \mathbf{s}_\tau) \int_{\mathbf{s}_{\tau+1}} p(\mathbf{s}_\tau \,|\, \mathbf{s}_{\tau+1})\; p(\mathbf{s}_{\tau+1} \,|\, \mathbf{z}_{\tau+1:t}) \\ &= c\, p(\mathbf{z}_\tau \,|\, \mathbf{s}_\tau)\; p(\mathbf{s}_\tau \,|\, \mathbf{z}_{\tau+1:t}) \end{aligned}$$

Smoothing distribution (forward-backward belief propagation):

$$p(\mathbf{s}_\tau \,|\, \mathbf{z}_{1:t}) = \frac{c}{p(\mathbf{s}_\tau)}\, p(\mathbf{z}_\tau \,|\, \mathbf{s}_\tau)\; p(\mathbf{s}_\tau \,|\, \mathbf{z}_{1:\tau-1})\; p(\mathbf{s}_\tau \,|\, \mathbf{z}_{\tau+1:t})$$

Batch Algorithms (smoothing): Estimation of state sequences using the entire observation sequence (i.e., all past, present & future data):

- optimal and efficient but not always applicable

Online Algorithms (filtering): Casual estimation of $\mathbf{x}_t$ occurs as observations become available, using present and past data only.

Kalman filter

Assume linearity and Gaussianity for the observation and dynamical models:

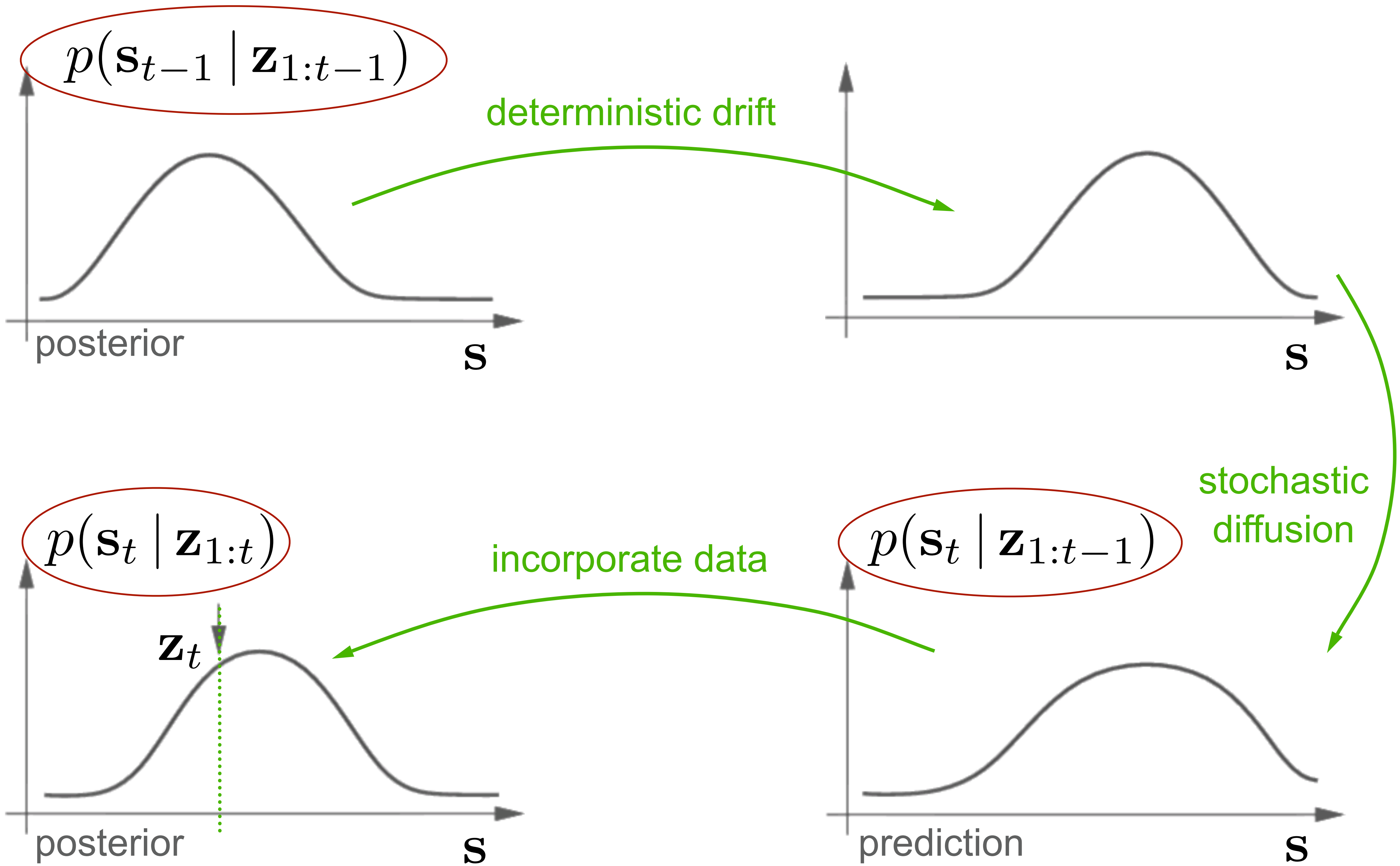
$$\begin{aligned} \mathbf{s}_t &= A\,\mathbf{s}_{t-1} + \eta_d \qquad & \eta_d &\sim \mathcal{N}(0,\, C_d) \\ \mathbf{z}_t &= M\,\mathbf{s}_t + \eta_m \qquad & \eta_m &\sim \mathcal{N}(0,\, C_m) \end{aligned}$$

Key Result: Prediction and filtering distributions are Gaussian, so they may be represented by sufficient statistics:

$$p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t-1}) = \int_{\mathbf{s}_{t-1}} p(\mathbf{s}_t \,|\, \mathbf{s}_{t-1})\, p(\mathbf{s}_{t-1} \,|\, \mathbf{z}_{1:t-1}) \;\sim\; \mathcal{N}(\mathbf{s}_t^-,\, C_t^-)$$

$$p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t}) = c\, p(\mathbf{z}_t \,|\, \mathbf{s}_t)\, p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t-1}) \;\sim\; \mathcal{N}(\mathbf{s}_t^+,\, C_t^+)$$

Depiction of filtering



Kalman filter

First well-known uses in computer vision:

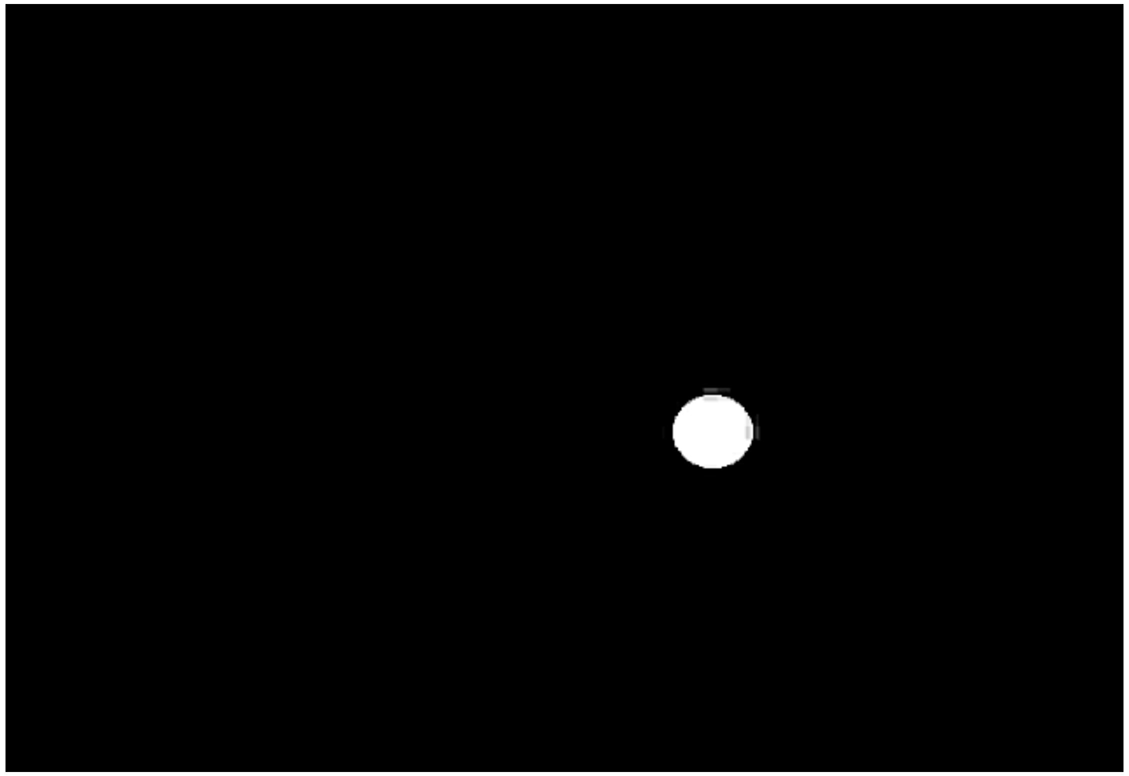
- Road following by tracking lane markers
  [Dickmanns & Graefe, "Dynamic monocular machine vision." Machine Vision and Applications, 1988]
- Rigid structure from feature tracks under perspective projection
  [Broida et al., "Recursive estimation of 3D motion from monocular image sequence." IEEE Trans. Aerosp. & Elec. Sys., 1990]

Multimodal likelihood functions

[Khan et al, CVPR '04]

Measurement clutter and occlusion often cause multimodal likelihoods.

Nonlinear dynamics



[Jepson et al, WSL Tracker, PAMI, 2001]

Object motion and interactions between objects often produce complex nonlinear dynamics (so Gaussianity is not preserved)

Approximate inference

Coping with multimodal, non-Gaussian distributions

- Optimization (to find MAP solution)
  - e.g., WSL tracker
- Monte Carlo approximations

WSL tracker

Goal: Tracking with precise alignment over long times

Problem: Changing appearance and unmodeled deformations

Key: Use ‘stable’ properties of appearance for tracking

WSL tracker

Monte Carlo inference (Particle filters)

Approximate the filtering distribution using point samples:

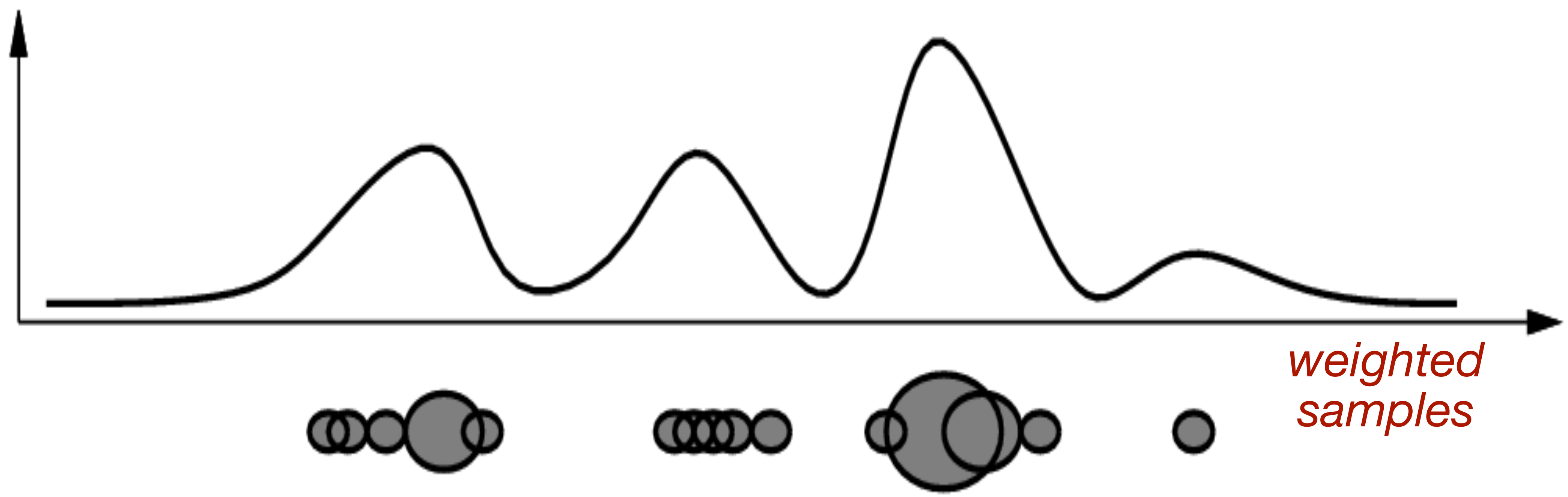
- By drawing a set of random samples from the filtering distribution, we could use samples statistics to approximate expectations

  Let $\mathcal{S} = \{\mathbf{s}^{(j)}\}$ be a set of N fair samples from distribution $\mathcal{P}(\mathbf{s})$, then for functions $f(\mathbf{s})$

$$E_{\mathcal{S}}\left[f(\mathbf{s})\right] \equiv \frac{1}{N}\sum_{j=1}^{N} f(\mathbf{s}^{(j)}) \xrightarrow{N\to\infty} E_{\mathcal{P}}\left[f(\mathbf{s})\right]$$

Problem: we don’t know how to draw samples from $p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t})$

Importance sampling



Importance sampling

Weighted sample set $\mathcal{S} = \{\mathbf{s}^{(j)},\, w^{(j)}\}$

- draw samples $\mathbf{s}^{(j)}$ from a *proposal distribution* $\mathcal{Q}(\mathbf{s})$, with weights $w^{(j)} = w(\mathbf{s}^{(j)})$, then

$$E_{\mathcal{S}}\left[f(\mathbf{s})\right] \;\equiv\; \sum_{j=1}^{N} w^{(j)} f(\mathbf{s}^{(j)}) \;\xrightarrow{N\to\infty}\; E_{\mathcal{Q}}\left[w(\mathbf{s})\, f(\mathbf{s})\right]$$

- If $w(\mathbf{s}) = \mathcal{P}(\mathbf{s})/\mathcal{Q}(\mathbf{s})$ then weighted sample statistics approximate expectations under $\mathcal{P}(\mathbf{s})$, i.e.,

$$\begin{aligned} E_{\mathcal{Q}}\left[w(\mathbf{s})\, f(\mathbf{s})\right] &= \int w(\mathbf{s})\, f(\mathbf{s})\, \mathcal{Q}(\mathbf{s})\, d\mathbf{s} \\ &= \int f(\mathbf{s})\, \mathcal{P}(\mathbf{s})\, d\mathbf{s} \\ &= E_{\mathcal{P}}[f(\mathbf{s})] \end{aligned}$$

Particle filter

Simple particle filter approximates the filtering distribution by drawing samples from the prediction distribution:

$$p(\mathbf{s}_t \mid \mathbf{z}_{1:t}) = c\ p(\mathbf{z}_t \mid \mathbf{s}_t)\, p(\mathbf{s}_t \mid \mathbf{z}_{1:t-1})$$

$$\mathcal{P}(\mathbf{s}) \qquad\qquad \mathcal{Q}(\mathbf{s})$$

$$w(\mathbf{s}) = \frac{\mathcal{P}(\mathbf{s})}{\mathcal{Q}(\mathbf{s})}$$

Particle filter

Simple particle filter approximates the filtering distribution by drawing samples from the prediction distribution:

$$p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t}) \;=\; c\; p(\mathbf{z}_t \,|\, \mathbf{s}_t)\; p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t-1})$$

With resampling at each time step:

$$\underset{\text{posterior}}{p(\mathbf{s}_{t-1} \,|\, \mathbf{z}_{1:t-1})} \xrightarrow{\text{sample}} \underset{\text{temporal dynamics}}{p(\mathbf{s}_t \,|\, \mathbf{s}_{t-1})} \xrightarrow{\text{sample}} \underset{\text{likelihood}}{p(\mathbf{z}_t \,|\, \mathbf{s}_t)} \xrightarrow{\text{normalize}} \underset{\text{posterior}}{p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t})}$$

[Gordon et al '93; Isard & Blake '98; Liu & Chen '98, ...]

Particle filter

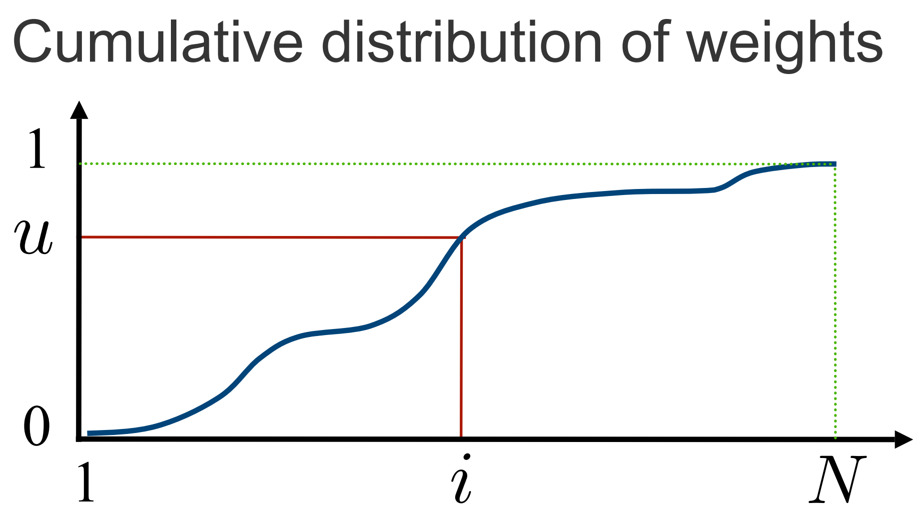
Given a weighted sample set $\mathcal{S} = \{\mathbf{s}_{t-1}^{(j)},\ w_{t-1}^{(j)}\}$, the prediction distribution is a mixture model

$$p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t-1}) \;=\; \sum_{j=1}^{N} w^{(j)}\, p(\mathbf{s}_t \,|\, \mathbf{s}_{t-1}^{(j)})$$

To draw samples from it:

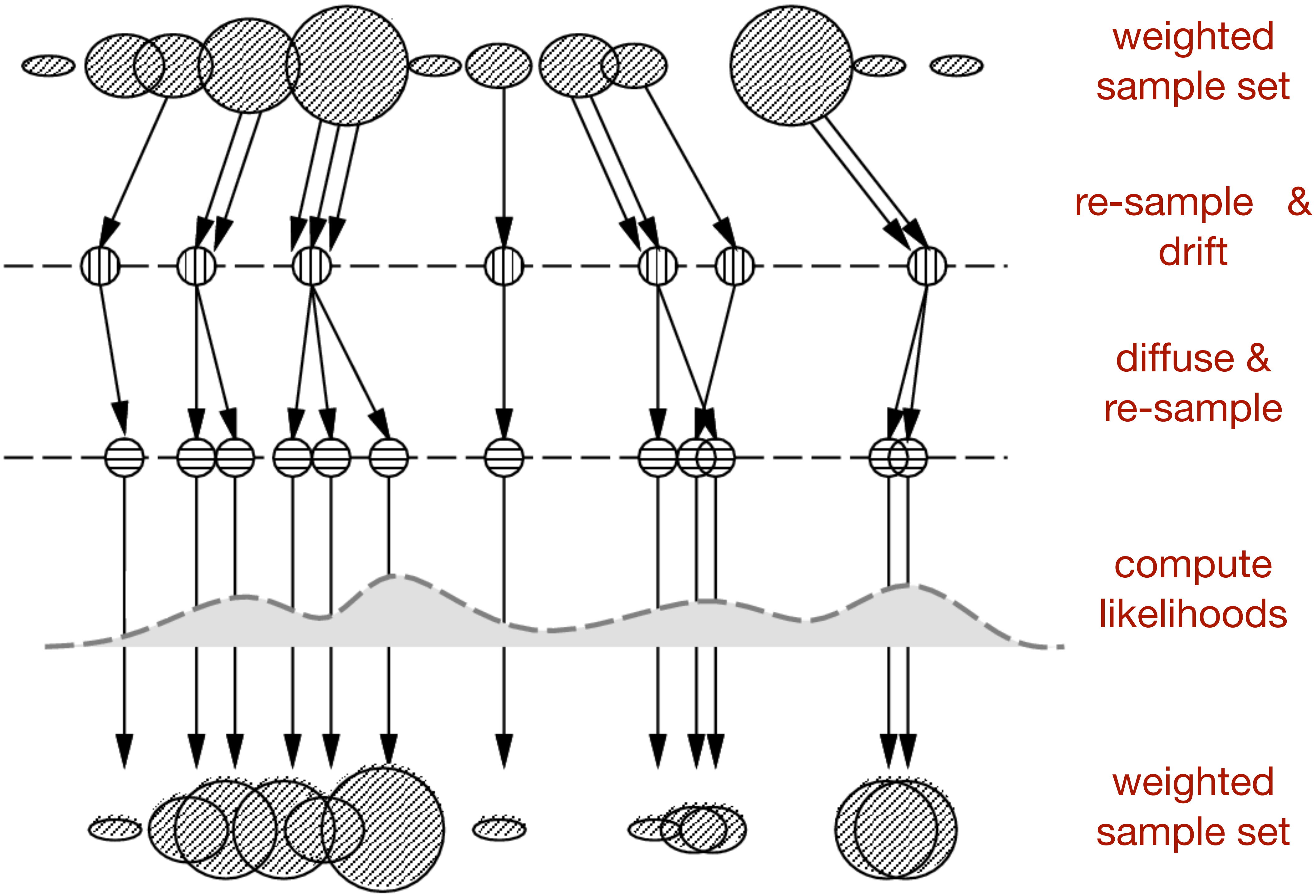
- sample a component of the mixture by the treating weights as mixing probabilities

sample
$u \sim U(0,1)$



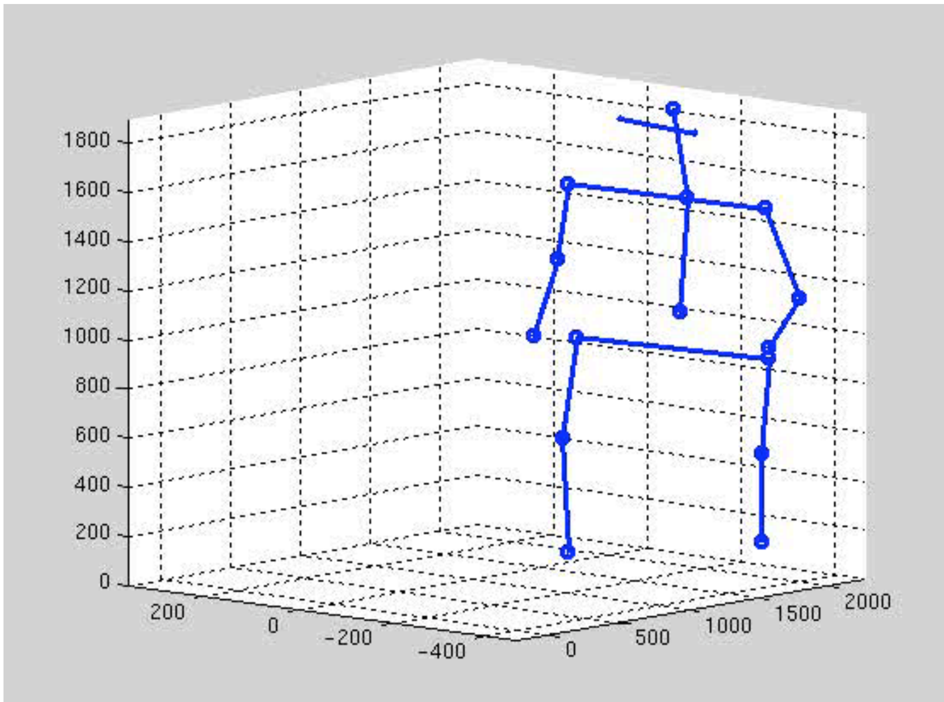
- then sample from the associated dynamics pdf $p(\mathbf{s}_t \,|\, \mathbf{s}_{t-1}^{(i)})$

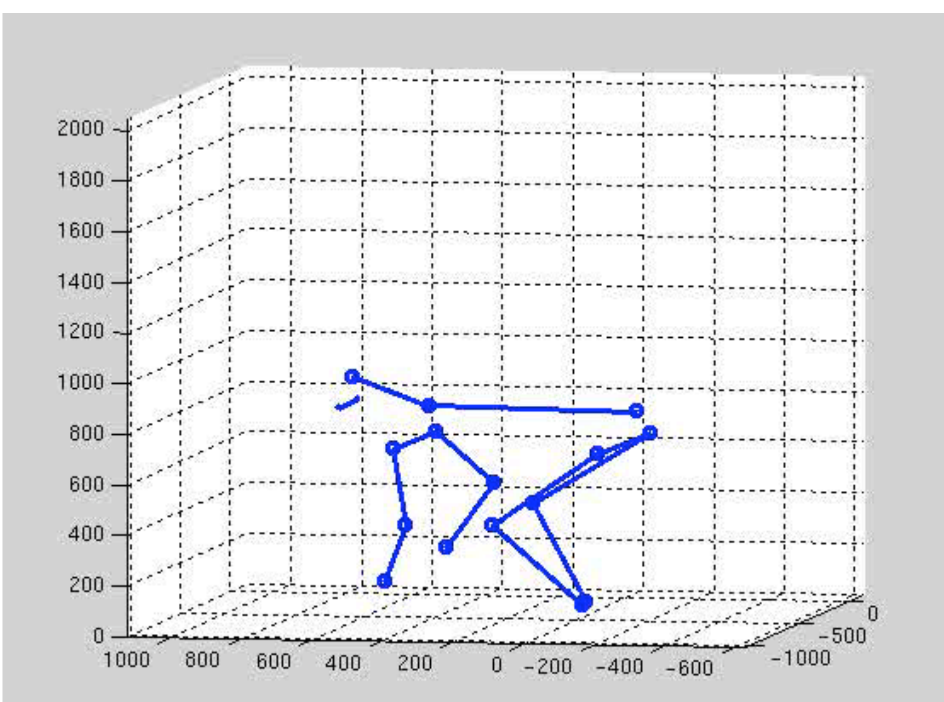
Particle filter

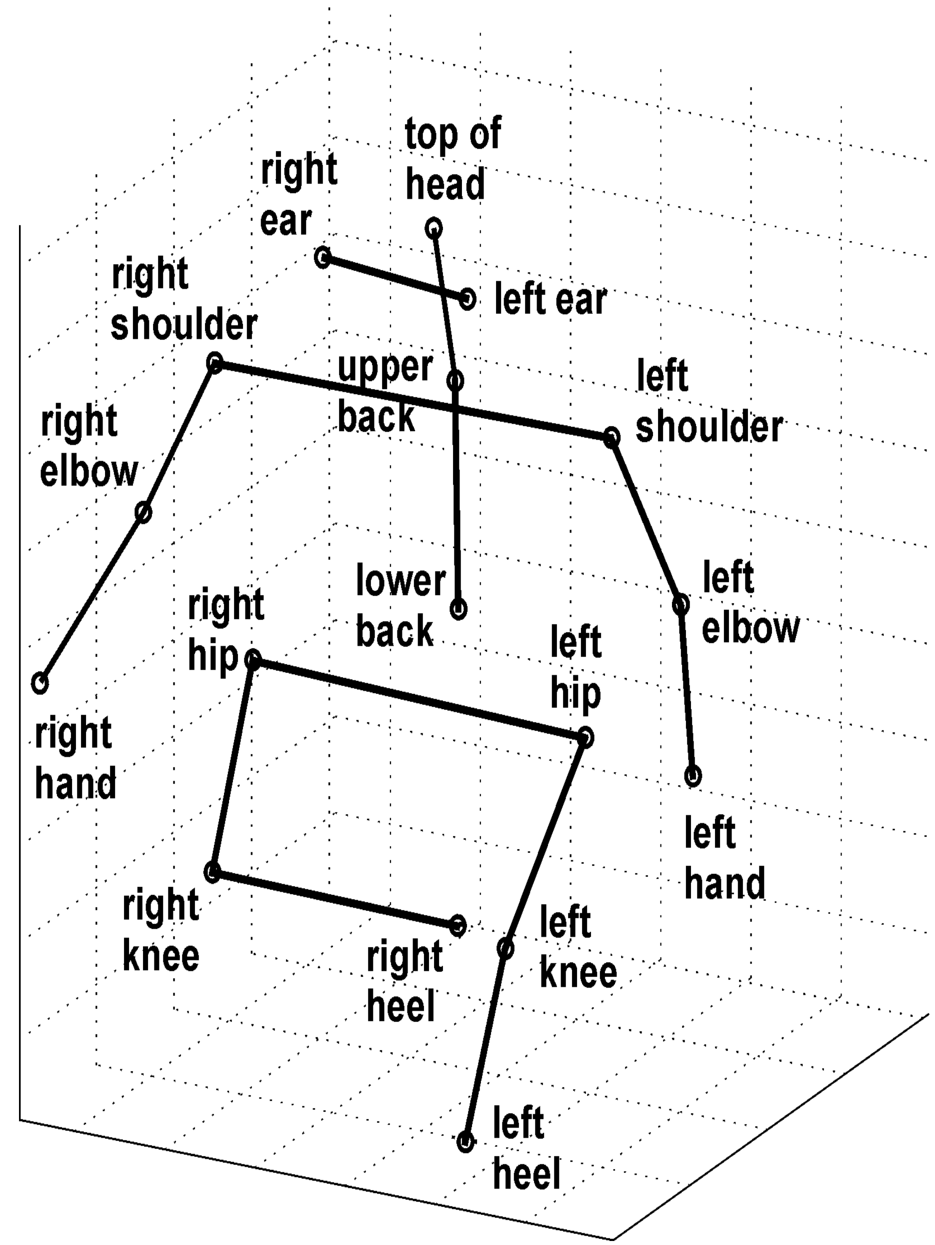


[Isard and Blake, IJCV '98]

Lessons learned: Sampling efficiency







3D Kinematic Model

(28D state: 22 joint angles, 6 global DOFs)

[Choo & Fleet, ICCV '01]

Likelihood and dynamics

Given the state, $\mathbf{s}$, and the articulated model, the 3D marker positions $\mathbf{X}_j$ onto the 2D image plane:

$$\mathbf{d}_j(\mathbf{s}) = T_j(\mathbf{X}_j; \mathbf{s})$$

Observation model:

$$\hat{\mathbf{d}}_j \; = \; \mathbf{d}_j + \eta_j \; , \qquad \eta_j \; \sim \; \mathcal{N}(0; \sigma_m^2 \mathbf{I}_2)$$

Likelihood of observed 2D locations, $\mathbf{D} = \{\hat{\mathbf{d}}_j\}$:

$$p(\mathbf{D} \,|\, \mathbf{s}) \; \propto \; \exp\left(-\frac{1}{2\sigma_m^2} \sum_j || \, \hat{\mathbf{d}}_j - \mathbf{d}_j(\mathbf{s}) \, ||^2 \right)$$

Smooth dynamics:

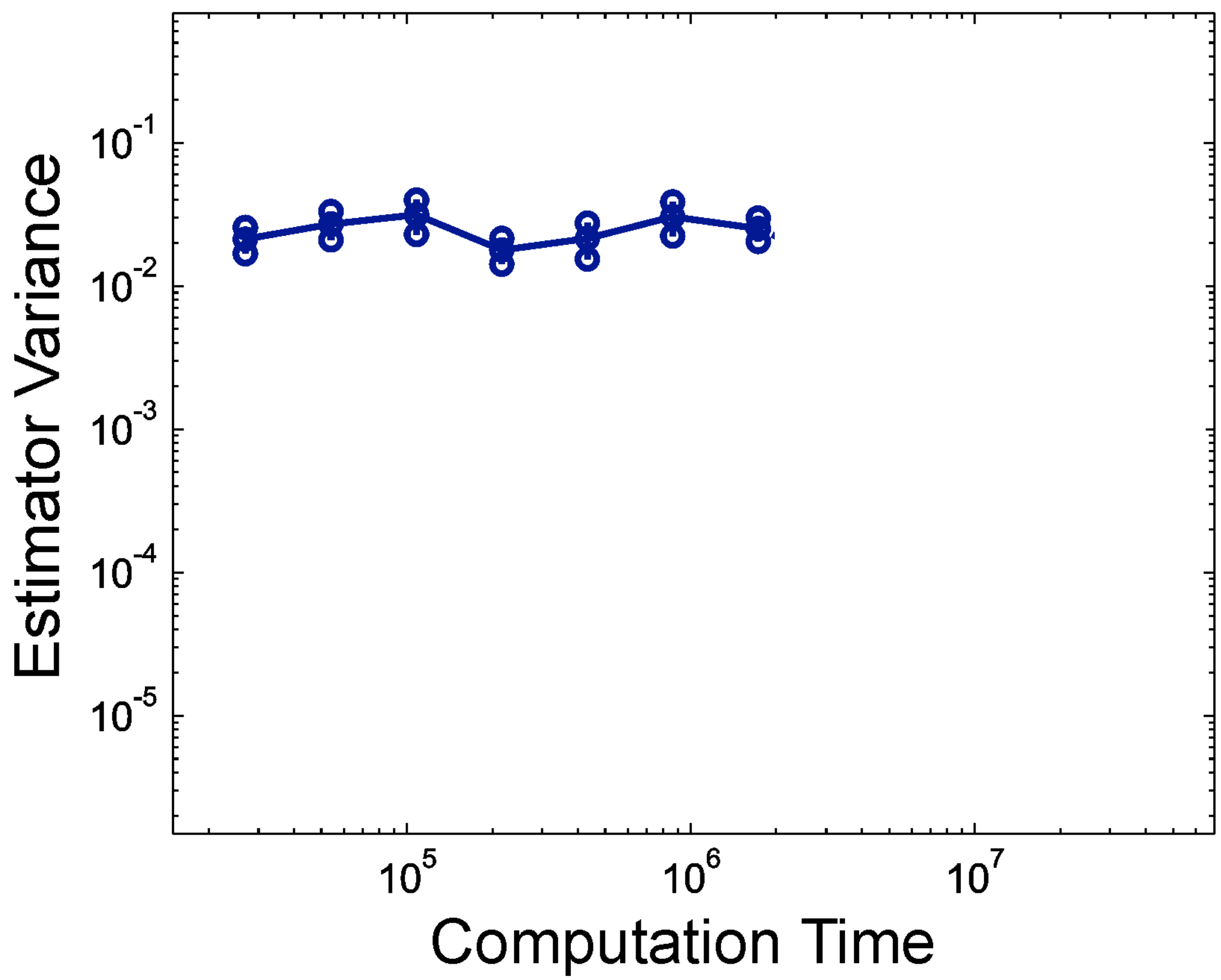
$$\mathbf{s}_t \; = \; \mathbf{s}_{t-1} + \epsilon_t$$

where ϵ_t is isotropic Gaussian for translational & angular variables

Performance

Estimator Variance:

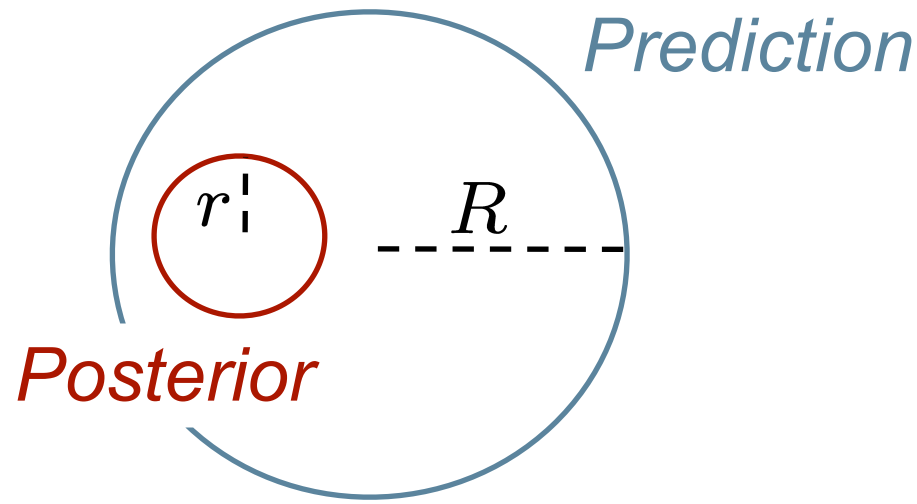
- multiple runs with independent noise & sampling
- variance measured as MSE from ground truth (from MCMC)



Problem: Sampling efficiency

Number of samples needed depends on the effective volumes (entropies) of the prediction and posterior distributions.

- With random sampling from the prediction density, the number of particles grow exponentially for samples to fall on states with high posterior. E.g., for *D*-dim spheres, with radii *R* and *r*,



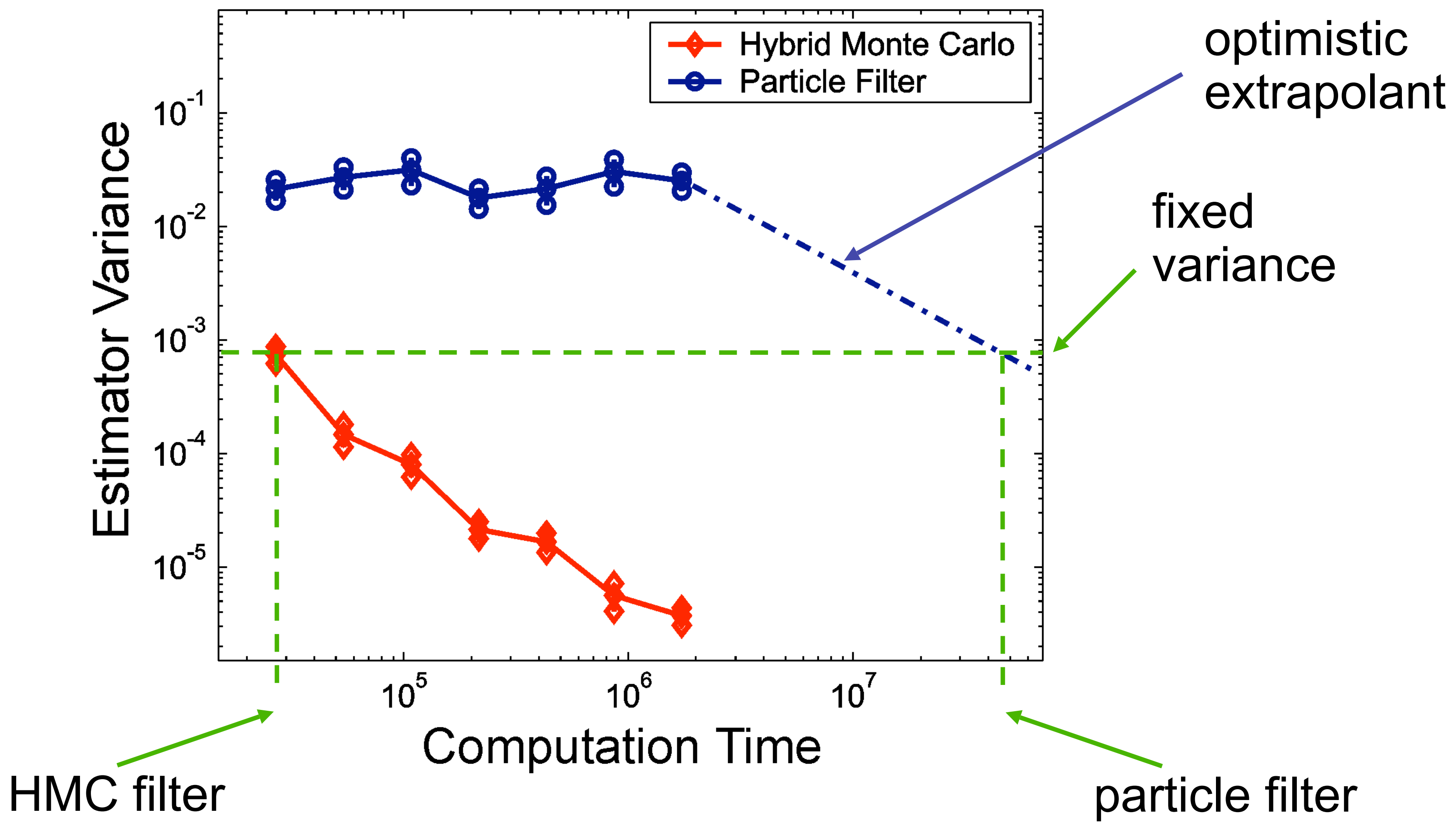
$$N > \left(\frac{R}{r}\right)^{D}$$

- effective number of 'independent' samples: $N_e = 1/\sum_j (w^{(j)})^2$

Hybrid Monte Carlo filter

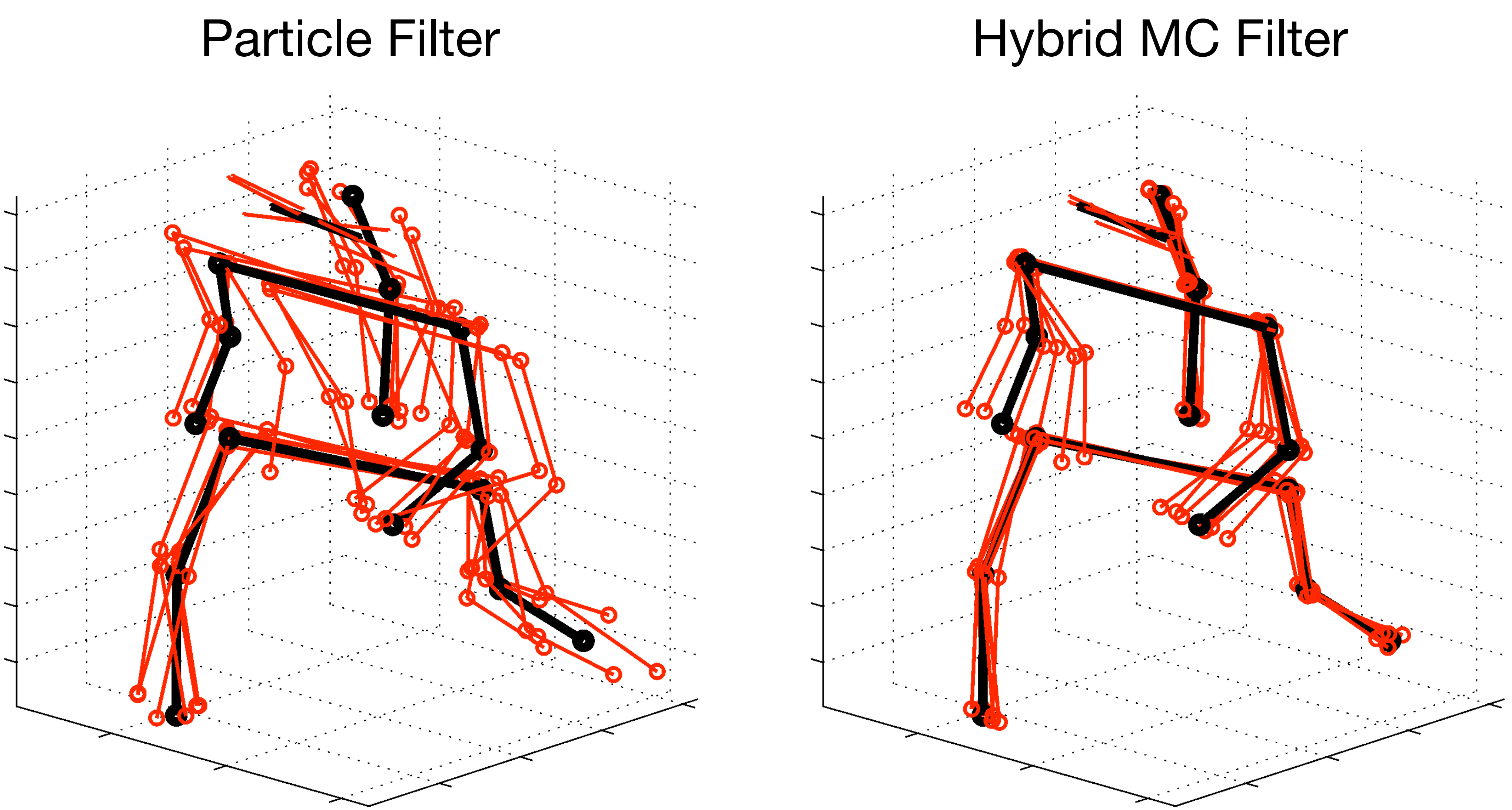
Improved sampling through MCMC:

- initialize a set of particles from a particle filter
- select a subset from which to initiate MCMC with stochastic gradient search (hybrid Monte Carlo)



[Choo & Fleet, ICCV '01]

Mean estimates on independent trials



Black: Ground truth (at frame 10)

Red: Mean state from 6 random trials

Lessons learned: Effective proposals

If proposal and target distributions differ significantly, then:

- most particle weights are near zero, and some modes get no samples, so the normalization constant c can be wildly wrong.

Prediction distributions $\mathcal{Q} = p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t-1})$ make poor proposals: dynamics are often uncertain, and likelihoods are often peaked.

Use the current observation to improve proposals.

- Let $\mathcal{D}(\mathbf{z}_t)$ be a continuous distribution obtained from some detector that yields target locations (e.g., a Gaussian mixture).
- Then, just modify the proposal density and importance weights:

$$\mathcal{Q} \;=\; \mathcal{D}(\mathbf{z}_t)\; p(\mathbf{s}_t \,|\, \mathbf{z}_{1:t-1}) \;\text{ with }\; w(\mathbf{s}_t) \;=\; \frac{c\; p(\mathbf{z}_t \,|\, \mathbf{s}_t)}{\mathcal{D}(\mathbf{z}_t)}$$

Lessons learned: Proper likelihoods

Do not compare states using different sets of observations.
Explain the entire image or use likelihood ratios.

E.g., let pixels intensities, conditioned on state be independent where D_f and D_b are disjoint sets of foreground and background pixels, and p_f and p_b are the respective likelihood functions.

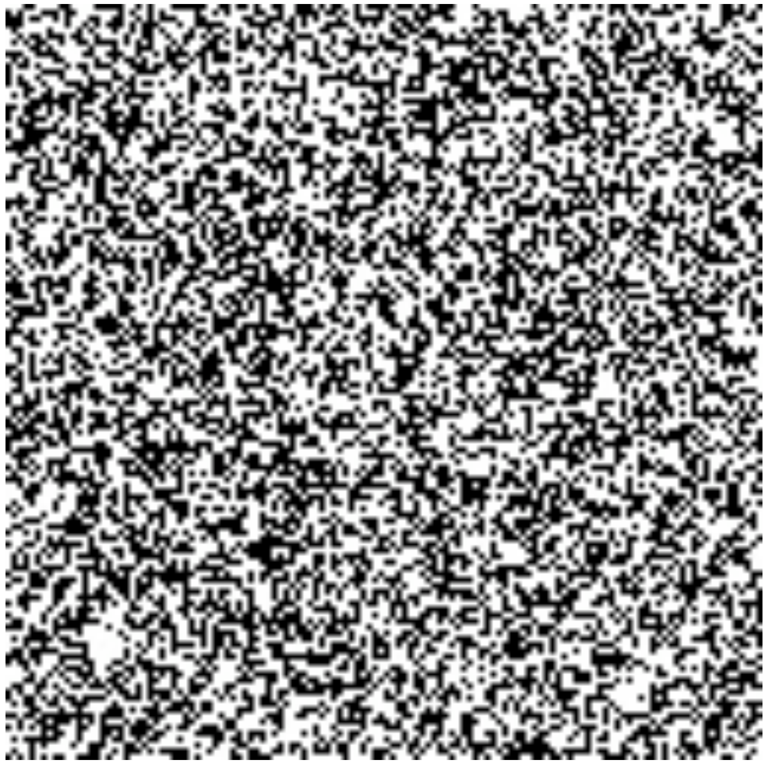
Divide $p(I \,|\, \mathbf{s})$ by the background likelihood of all pixels (i.e., as if no target is present):

$$
\begin{aligned}
p(I \,|\, \mathbf{s}) \quad &\propto \quad \frac{\prod_{\mathbf{y} \in D_f} p_f(I(\mathbf{y}) \,|\, \mathbf{s}) \;\; \prod_{\mathbf{y} \in D_b} p_b(I(\mathbf{y}))}{\prod_{\mathbf{y}} p_b(I(\mathbf{y}))} \\
&= \quad \frac{\prod_{\mathbf{y} \in D_f} p_f(I(\mathbf{y}) \,|\, \mathbf{s}) \;\; \prod_{\mathbf{y} \in D_b} p_b(I(\mathbf{y}))}{\prod_{\mathbf{y} \in D_f} p_b(I(\mathbf{y}) \,|\, \mathbf{s}) \;\; \prod_{\mathbf{y} \in D_b} p_b(I(\mathbf{y}))} \\
&= \quad \prod_{\mathbf{y} \in D_f} \frac{p_f(I(\mathbf{y}) \,|\, \mathbf{s})}{p_b(I(\mathbf{y}))}
\end{aligned}
$$

Lessons Learned: Use the right state space

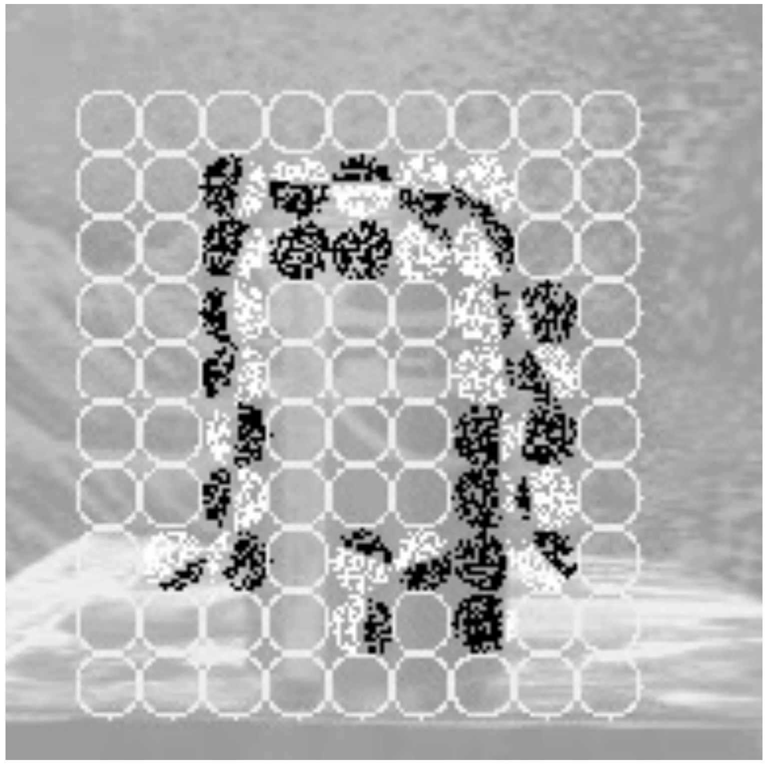
Despite the potential to approximate multimodal posteriors, tracking multi targets with a single target state space is unadvised.

Finding occlusion boundaries



Motion boundaries yield information about position and orientation of surface boundaries, and about relative surface depths

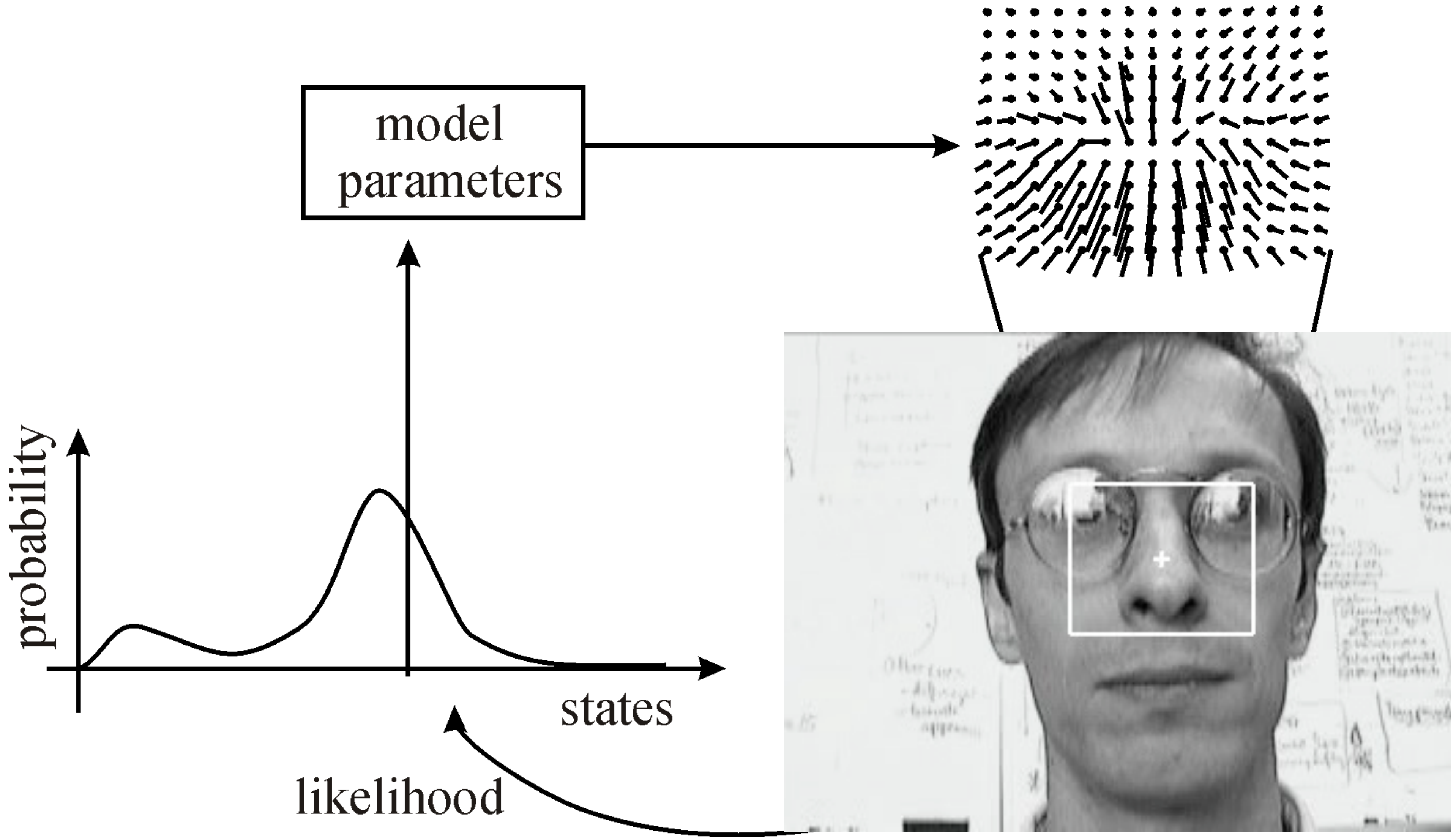
Finding occlusion boundaries



Estimation of smooth motion and occlusion boundaries on hybrid random fields with non-parametric Bayesian inference.

[Nestares and Fleet, CVPR '01]

Finding lips and lip reading

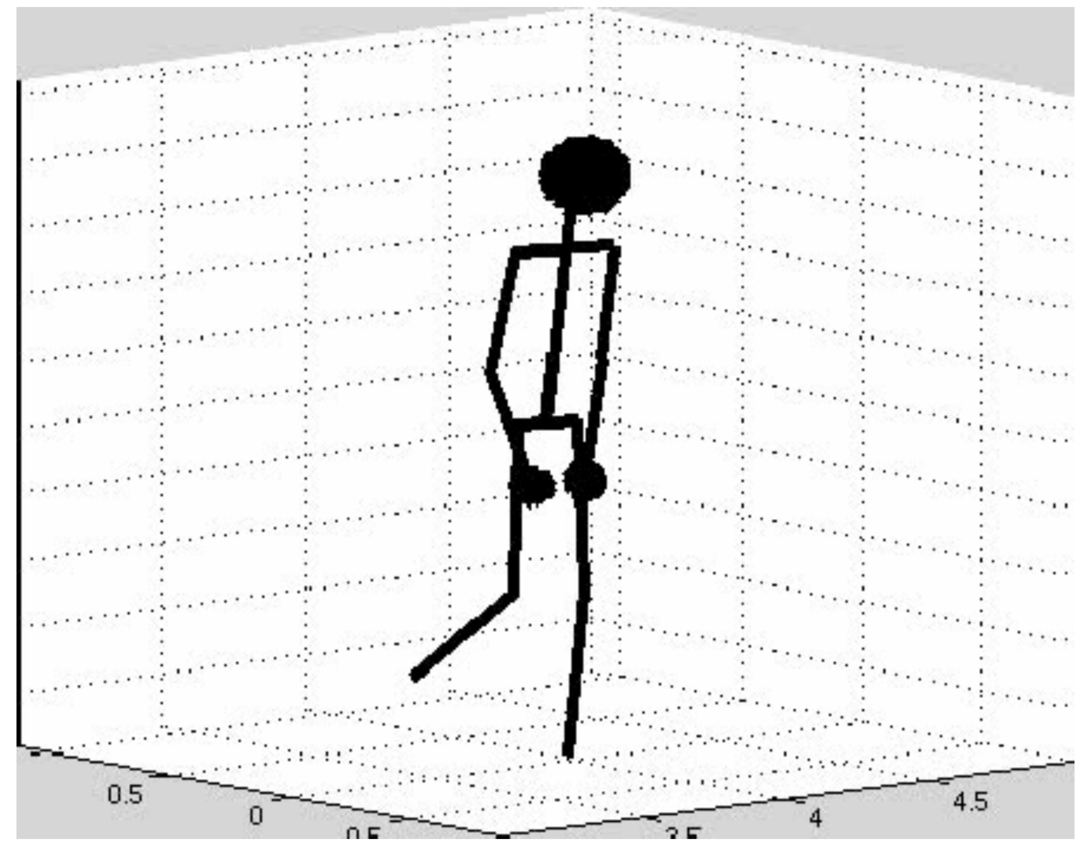


Probabilistic detection, tracking, and recognition of motion events in video, with learned models of image motion.

[Fleet, Black, Yacoob and Jepson, IJCV 2000]

Human pose tracking

Estimate the three-dimensional structure of people from video, with constraints on their shape, size, & motion.



[Sidenbladh, Black and Fleet, ECCV 2000]

Selected references

Arulampalam, M. et al., A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans Signal Processing 50(2),* 2002.

Blake, A., Visual tracking. In *Mathematical Models for Computer Vision*, Paragios, Chen and Faugeras (ed), Springer 2005

Choo K. and Fleet, D., People tracking using Hybrid Monte Carlo filtering. *Proc* IEEE *ICCV,* 2001

Doucet, A. et al, On Sequential Monte Carlo sampling methods for Bayesian filtering. *Stats and Computing* 10, 2000

Gordon N. et al., Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F,* 140(2), 1993

Isard M. & Blake A., Condensation: Conditional density propagation. *IJCV* 29, 1998

Jepson A. et al, Robust online appearance models for visual tracking. *IEEE Trans PAMI* 25(10), 2003

Khan Z. et al, A Rao-Blackwellized particle filter for Eigentracking. *Proc IEEE CVPR* 2004

Liu J and Chen, R. Sequential Monte Carlo methods for dynamic systems. *JASA 93,* 1998