IEEE Workshop on Evaluation of Articulated Human Motion and Pose Estimation, June 2007, Minneapolis *[Keynote Presentation]*

Model-Based Human Pose Tracking

David J Fleet

Department of Computer Science

University of Toronto

with

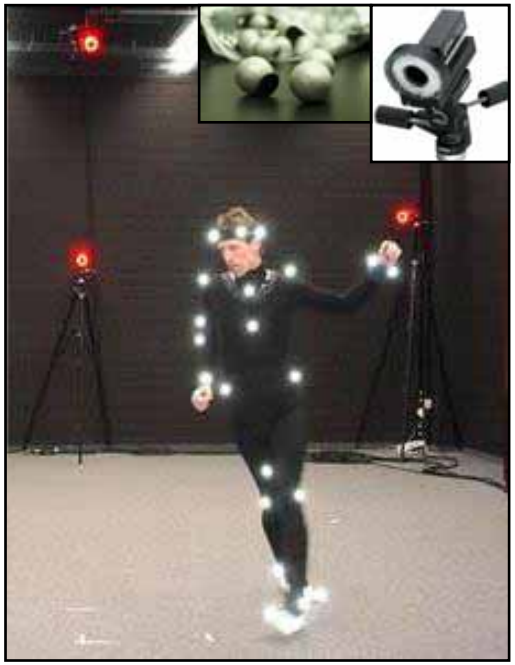
- M. Brubaker, Toronto
- P. Fua, EPFL
- A. Hertzmann, Toronto
- N. Lawrence, Sheffield
- R. Urtasun, MIT
- J. Wang, Toronto

Introduction

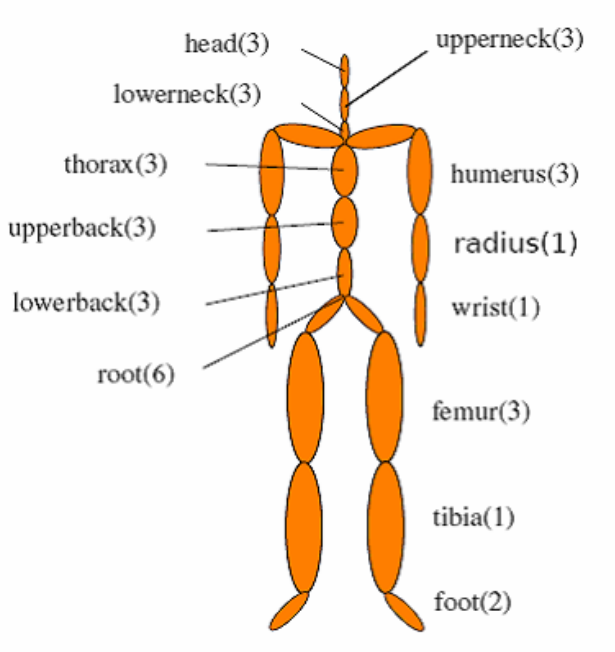
We need to get a lot of things right to successfully infer 3D pose and motion from monocular video:

- body size and shape
- pose and motion
- appearance (foreground and background)
- lighting and occlusion
- image measurement
- search and detection
- …

Introduction



motion capture

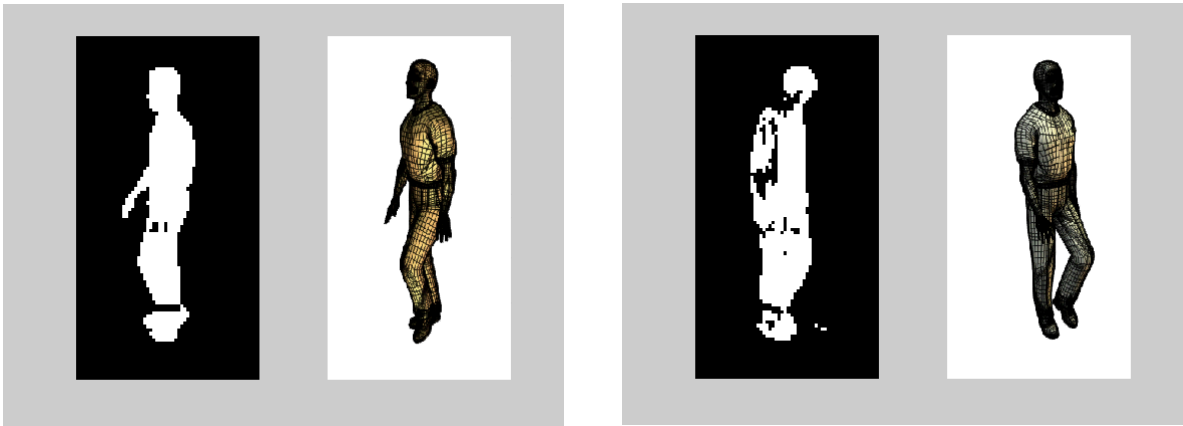


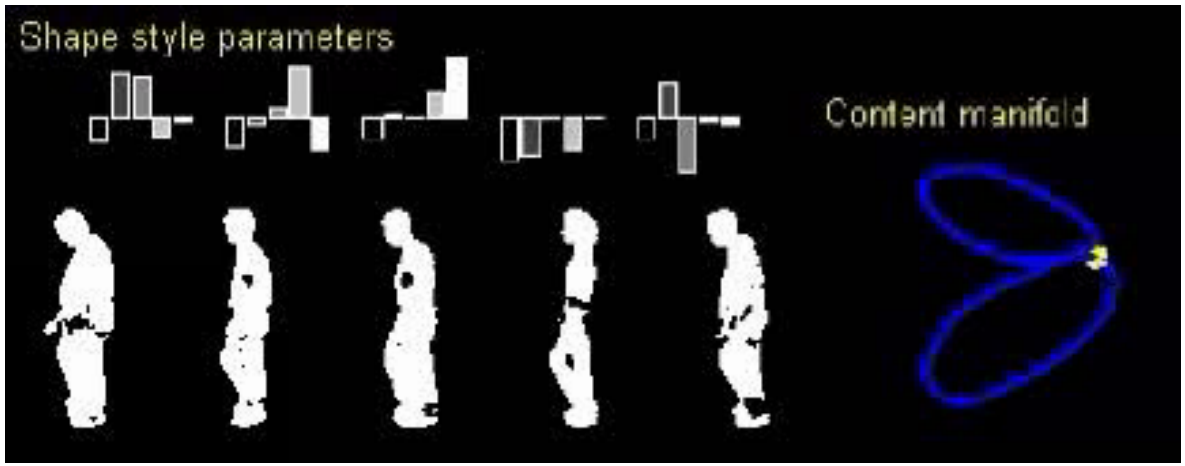
articulated model

Human pose and motion data are high-dimensional, and difficult to obtain.

Sparseness of data, over-fitting and generalization are significant issues.

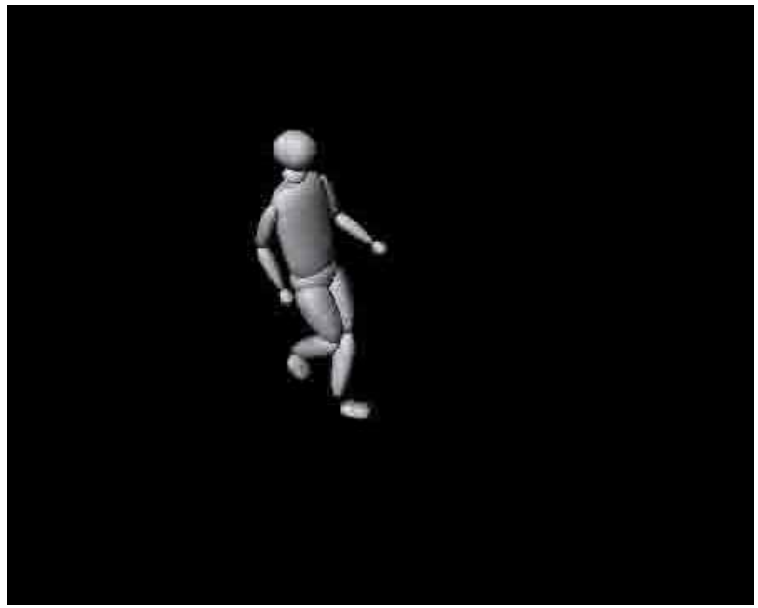
Introduction





[Elgammal and Lee `04]

Introduction

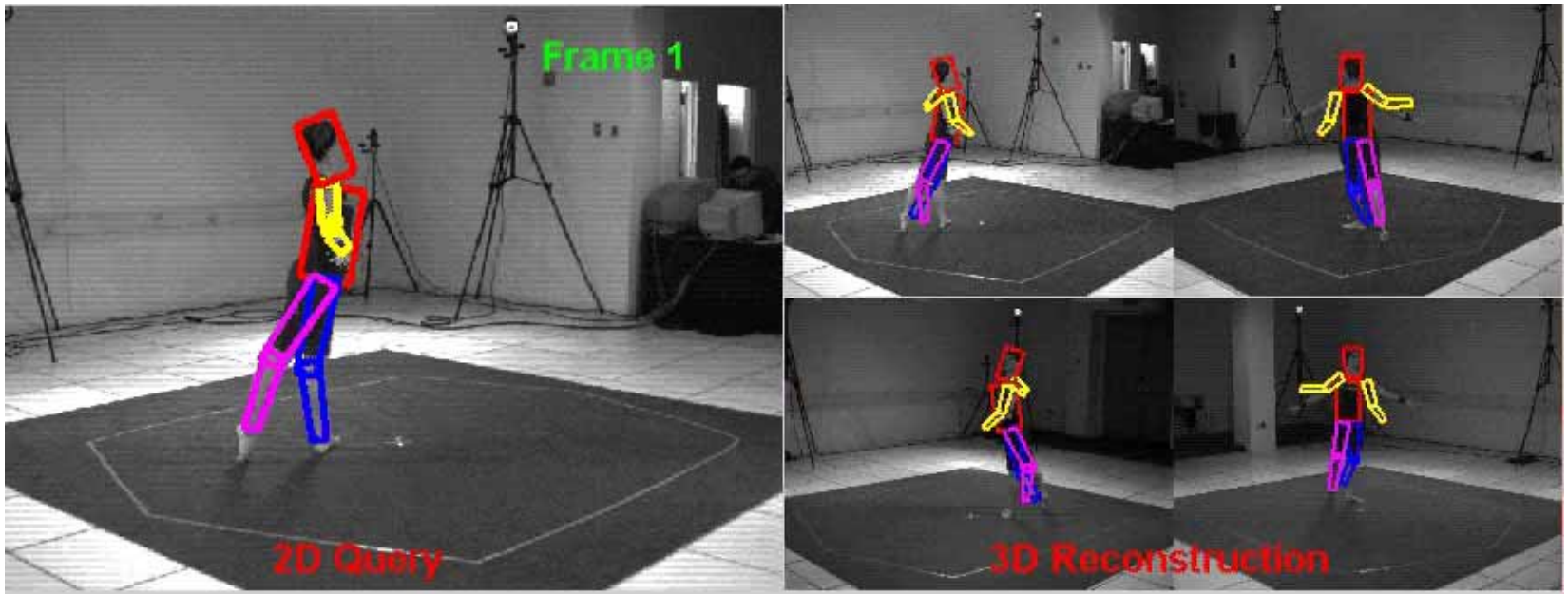


[Sminchisescu and Jepson `04]

Introduction

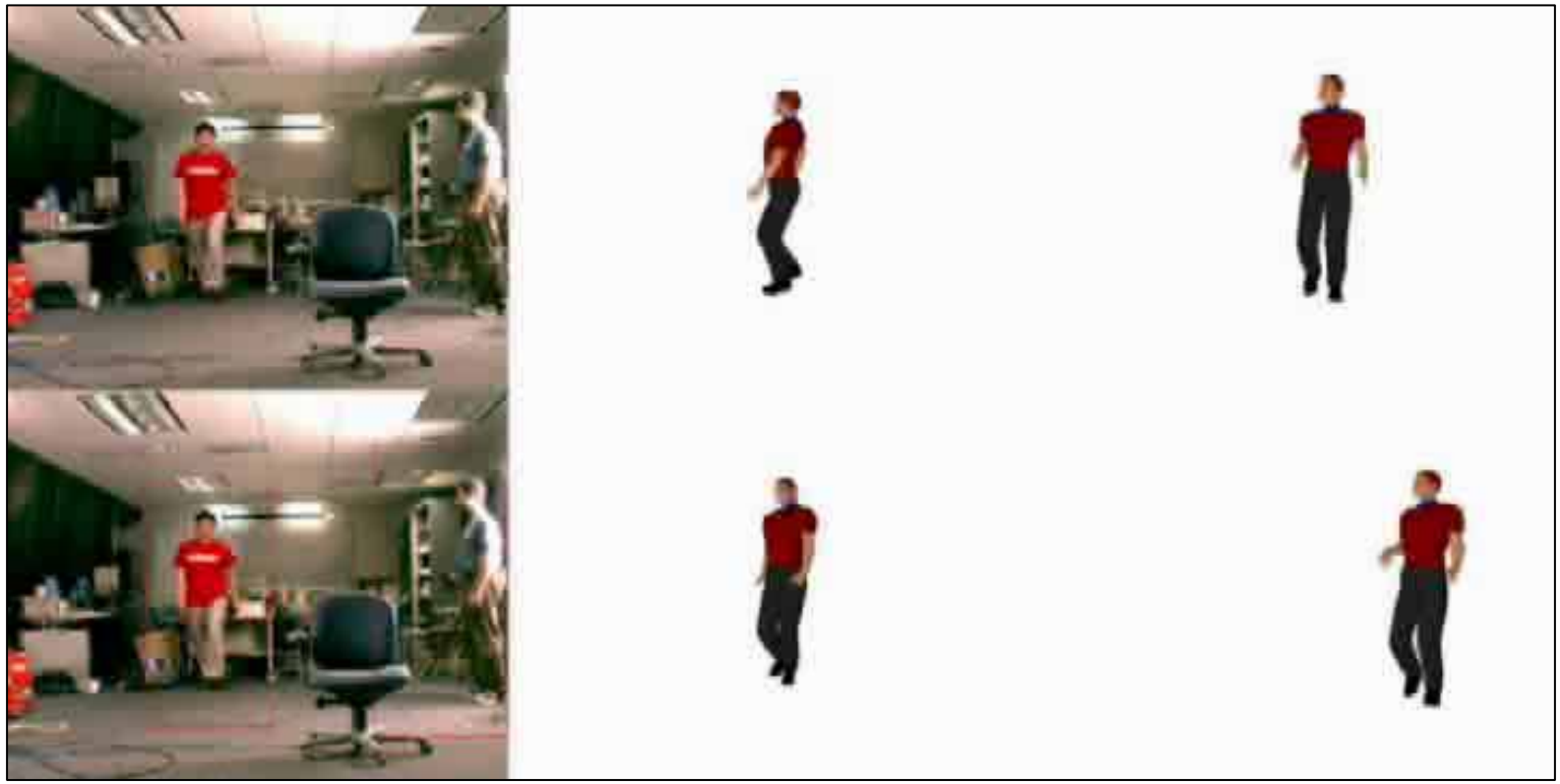
[Agarwal and Triggs `06]

Introduction



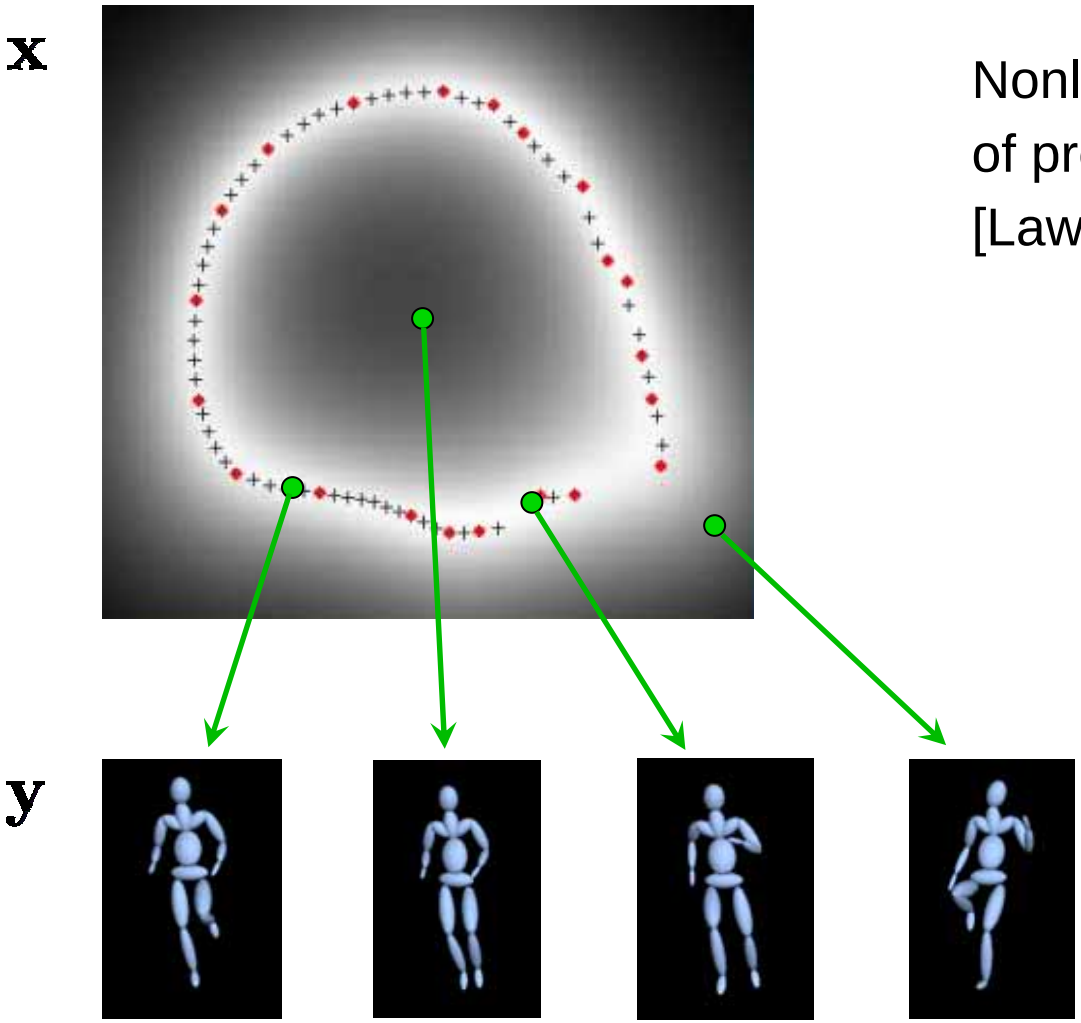
[Sigal et al, `06]

Introduction



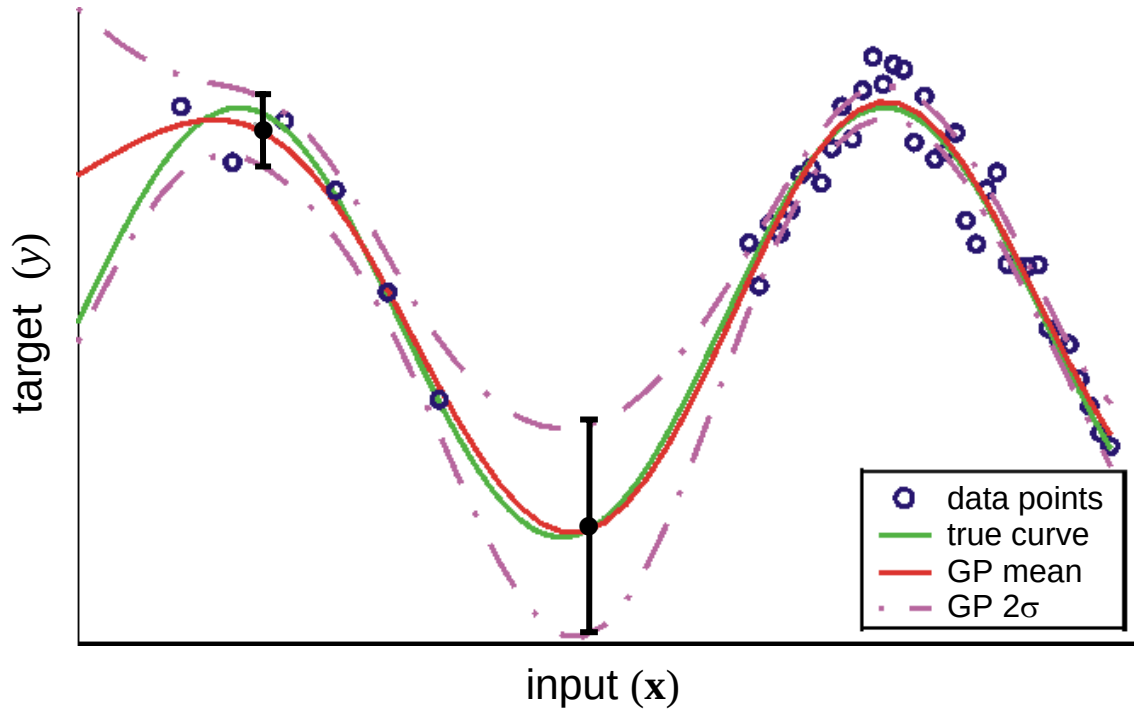
[Sminchisescu et al. `06]

Gaussian process latent variable models (GPLVM)



Nonlinear generalization of probabilistic PCA [Lawrence `05].

Gaussian processes



Model averaging (marginalization of the parameters) helps to avoid problems due to over-fitting and under-fitting with small data sets.

Gaussian processes

Output y is modelled as a function of input $\mathbf{x}$:

$$y \;=\; g(\mathbf{x}) \;=\; \sum_j w_j\,\phi_j(\mathbf{x}) \;=\; \mathbf{w}^T\mathbf{\Phi}(\mathbf{x})$$

If $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then $y \,|\, \mathbf{x}$ is zero-mean Gaussian with covariance

$$k(\mathbf{x}, \mathbf{x}') \;\equiv\; E[\,yy'\,] \;=\; \mathbf{\Phi}(\mathbf{x})^T\mathbf{\Phi}(\mathbf{x}')$$

A Gaussian process is fully specified by a covariance function $k(\mathbf{x}, \mathbf{x}')$ and its hyper-parameters

Linear: $k(\mathbf{x}, \mathbf{x}') \;=\; \theta\,\mathbf{x}^T\mathbf{x}'$

RBF: $k(\mathbf{x}, \mathbf{x}') \;=\; \theta\,\exp(-\frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}'\|^2)$

Gaussian process latent variable models

Joint likelihood of vector-valued data $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]^T,\ \mathbf{y}_n \in \mathcal{R}^D$, given the latent positions $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^T$:

$$p(\mathbf{Y} \mid \mathbf{X}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{Y}_d; \mathbf{0}, \mathbf{K})$$

where $\mathbf{Y}_d$ denotes the d^{th} dimension of the training data, and the kernel matrix has elements $(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, and is shared by all data dimensions.

GPDM: For time-series data one can include a Gaussian process prior on latent state sequences $p(\mathbf{X})$ [Wang et al '06].

Learning: Maximize log likelihood (or MAP) to find latent positions and kernel hyper-parameters, given an initial guess (e.g., using PCA).

Conditional (predictive) distribution

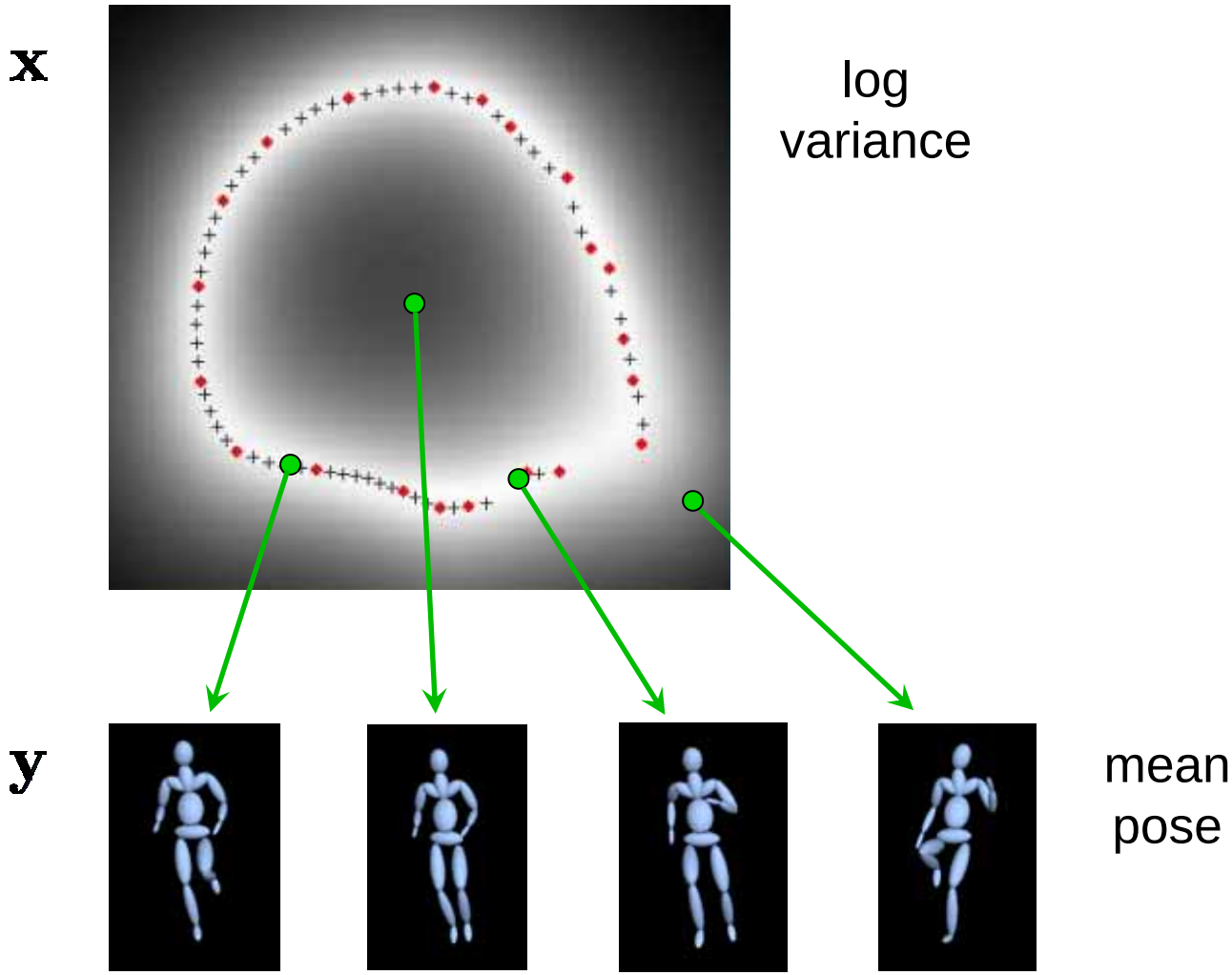
Given a model $\mathcal{M} = (\mathbf{Y}, \mathbf{X})$, the distribution over the data $\mathbf{y}_*$ conditioned on a latent position, $\mathbf{x}_*$, is Gaussian:

$$\mathbf{y}_* \,|\, \mathbf{x}_*, \mathcal{M} \;\sim\; \mathcal{N}(\mathbf{m}(\mathbf{x}_*),\, \sigma^2(\mathbf{x}_*))$$

where

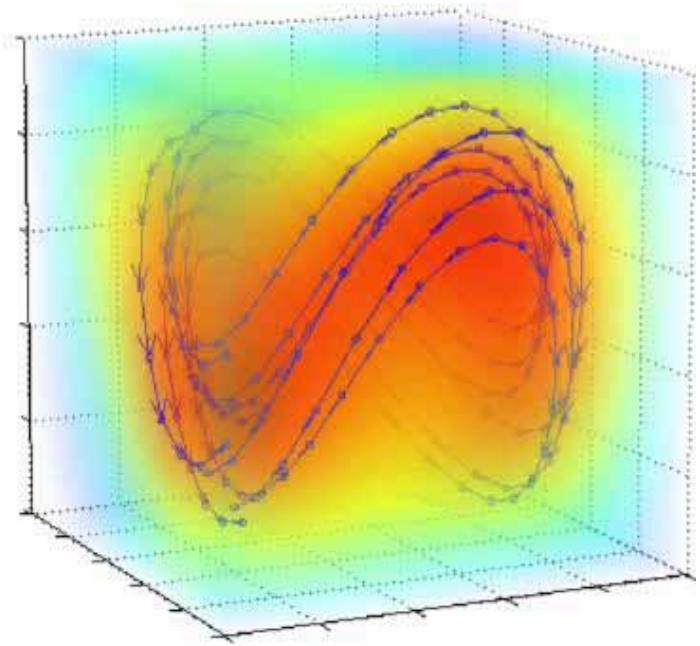
$$\begin{aligned}
\mathbf{m}(\mathbf{x}_*) &= \mathbf{Y}\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}_*) \\
\sigma^2(\mathbf{x}_*) &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}(\mathbf{x}_*)^T \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}_*) \\
\mathbf{k}(\mathbf{x}_*) &= [k(\mathbf{x}_*, \mathbf{x}_1), ..., k(\mathbf{x}_*, \mathbf{x}_N)]^T
\end{aligned}$$

Conditional (predictive) distribution

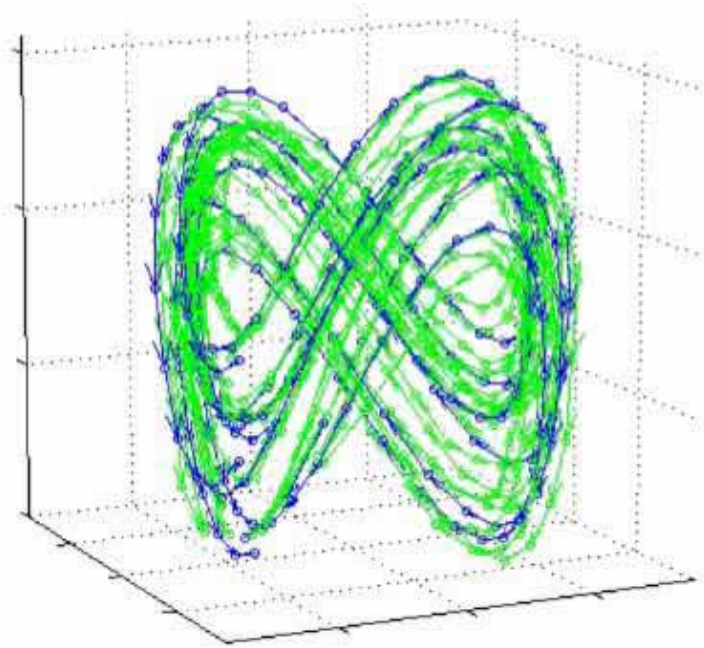


3D B-GPDM for walking

6 walking subjects,1 gait cycle each, on treadmill at same speed with a 20 DOF joint parameterization.



GPDM: log reconstruction variance $\ln \sigma^2_{\mathbf{y}|\mathbf{x},\mathbf{X},\mathbf{Y}}$



GPDM: sample trajectories

[Urtasun et al, `06]

People tracking with GPDM

Image Observations: $\mathbf{I}_{1:t} \equiv (\mathbf{I}_1, \ldots, \mathbf{I}_t)$

State: $\phi_t = [\mathbf{G}_t, \mathbf{y}_t, \mathbf{x}_t]$

GPDM: $\mathcal{M}$

global pose

joint angles

latent coordinates

Inference: MAP estimation by hill climbing on the negative log posterior over windowed state sequences

$$p(\phi_{t:t+\tau} \,|\, \mathbf{I}_{1:t+\tau}, \mathcal{M}) \;\propto\; p(\mathbf{I}_{t:t+\tau} \,|\, \phi_{t:t+\tau})\; p(\phi_{t:t+\tau} \,|\, \mathbf{I}_{1:t-1}, \mathcal{M})$$

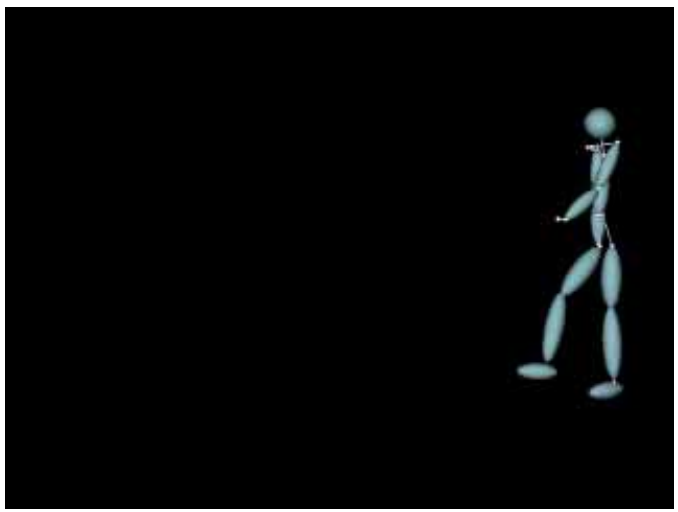
posterior — likelihood — prediction

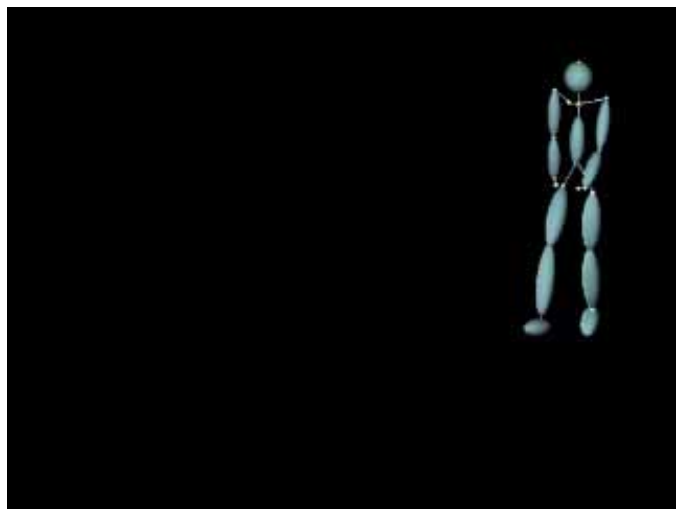
Image Measurements: trajectories of 2D patches, estimated with the WSL tracker [Jepson et al. `03]

[Urtasun et al, `06]

Occlusion

3D model
overlaid
on video

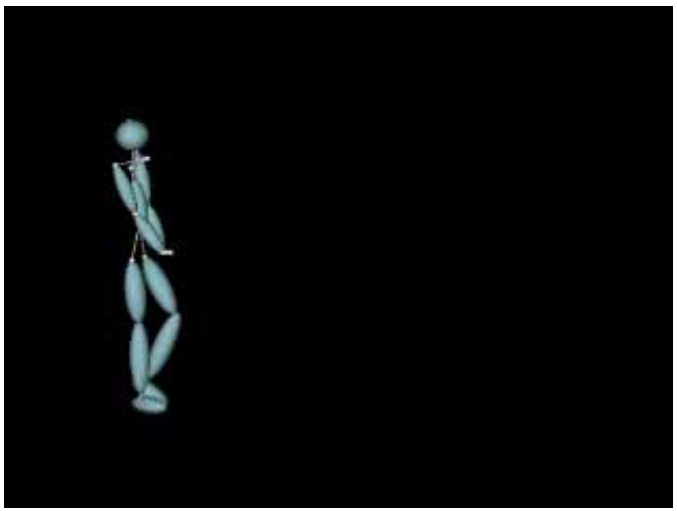


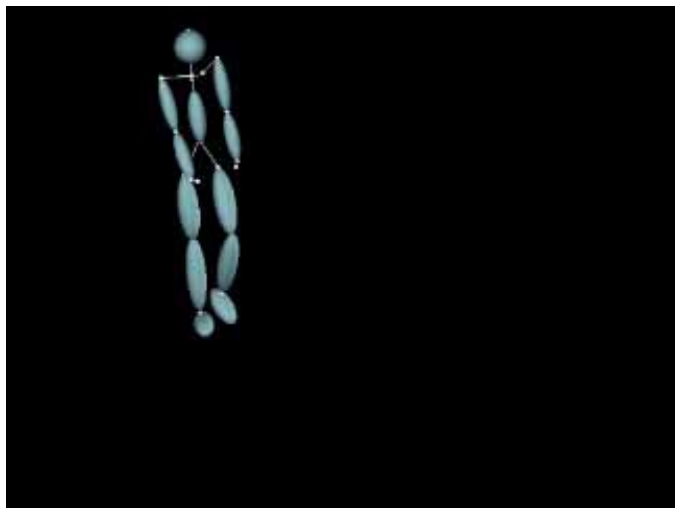


3D animated characters

Exaggerated gait

3D model
overlaid
on video





3D animated characters

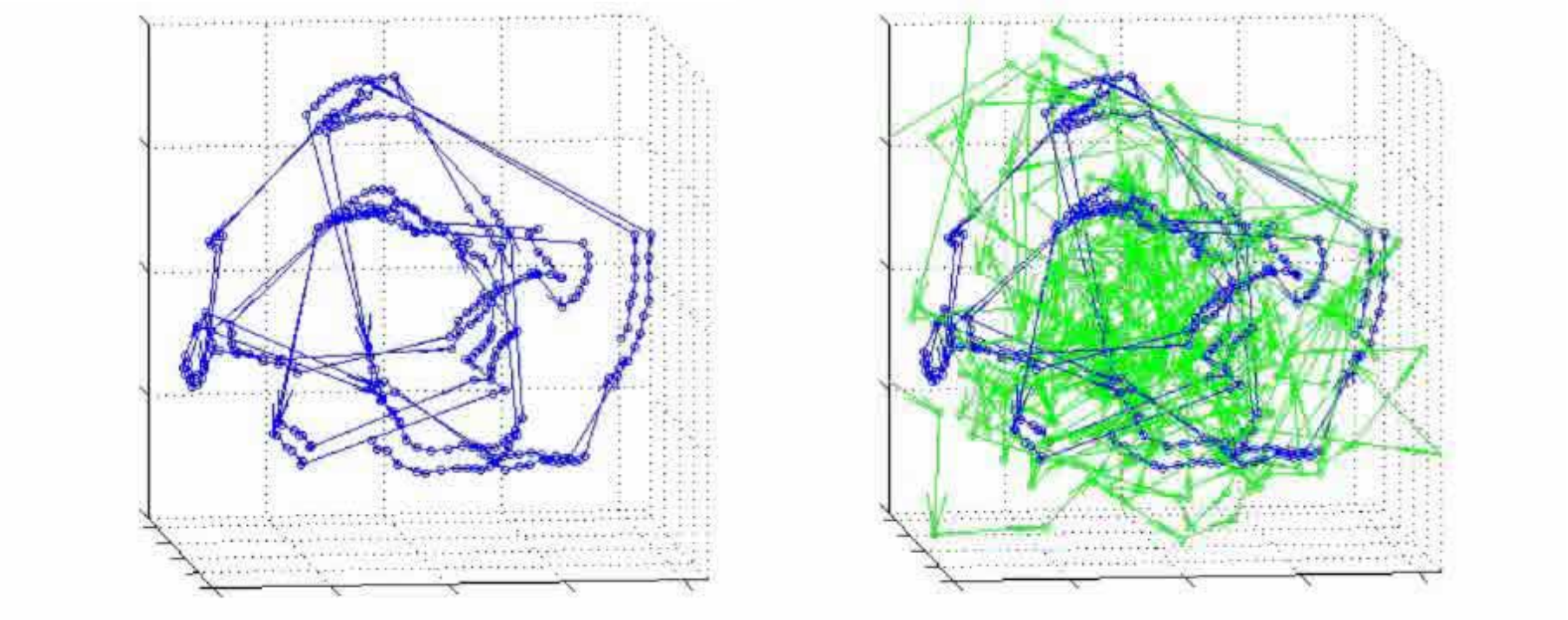
… but wait …

you're thinking …

- these models won't scale
- they won't handle different styles of motion
- efficiency is a major issue
- the amount of data required for training is daunting

Multiple motions often produce poor models

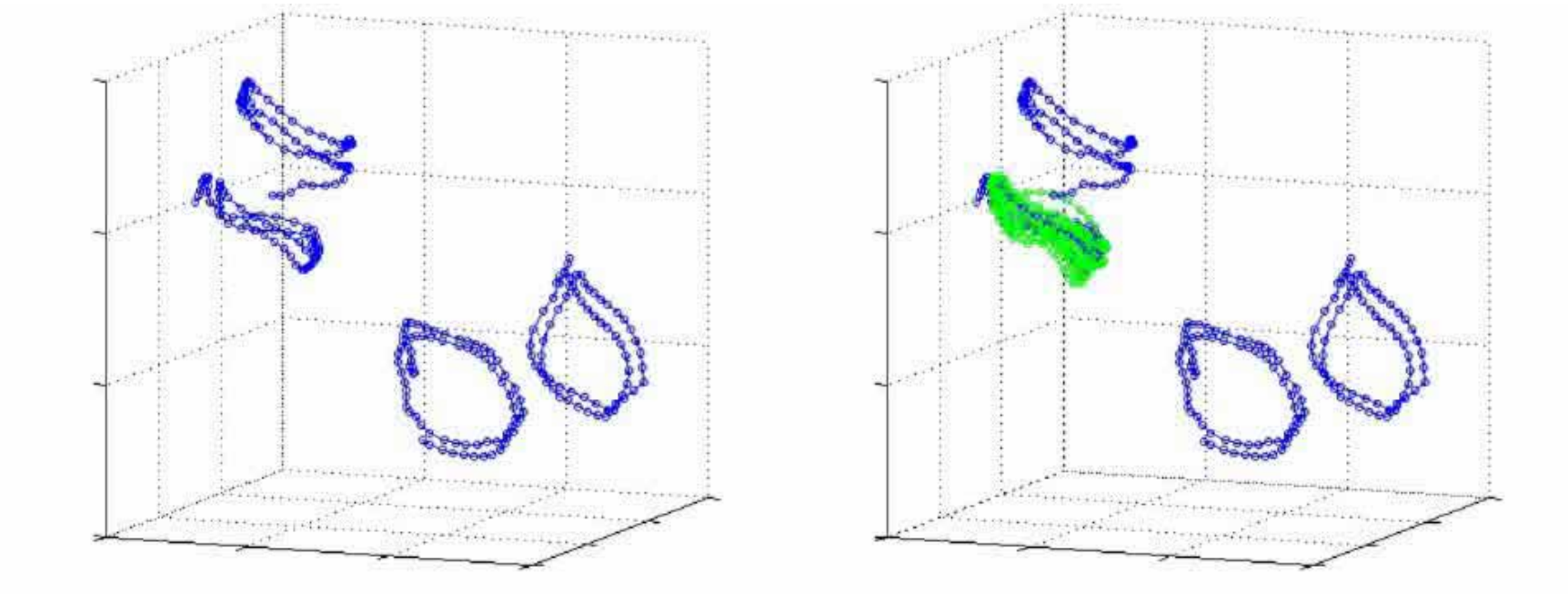
4 walking subjects, 2 gait cycles each, 50 DOFs



GPDM with MAP learning

Multiple motions often produce poor models

4 walking subjects, 2 gait cycles each, 50 DOFs



Marginalize latent positions, and solve with HMC-EM [Wang et al, ‘06]

Multiple motions often produce poor models

GPLVMs do not ensure that the map from pose to the latent space is smooth, i.e., that nearby poses map to nearby latent positions.

With sparse mocap data, different motions may be modeled as arising from different distributions.

But there is more valuable information in the training data, and prior knowledge about human motion that can be used to influence the structure of the model.

Topological constraints

Exploit prior knowledge to control the topology of the latent space, and promote smoothness.

Topological constraints and smoothness can be encouraged with back-constraints (parameterized latent coordinates):

$$\mathbf{x}_j = g(\mathbf{y}_j, \mathbf{a})$$

Smoothness “priors” in terms of latent positions of “similar” poses:

$$-\log p(\mathbf{X}) = \kappa \sum_j ||\mathbf{x}_j - \sum_{i \in \mathcal{N}(j)} w_{ij} \mathbf{x}_i||^2 + c$$

where weights are chosen (or optimized) to represent likely latent positions in terms of neighboring positions (or desired predictors).

[Urtasun et al. ’07]

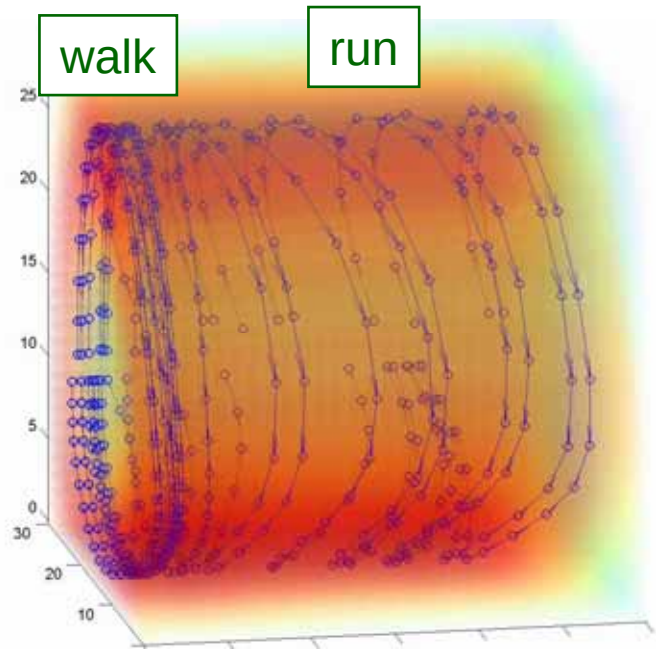
“Cylindrical” topology

Back-constraints: first 2 latent coordinates map to vicinity of unit circle

Locally-linear prior: similar poses map to nearby latent positions

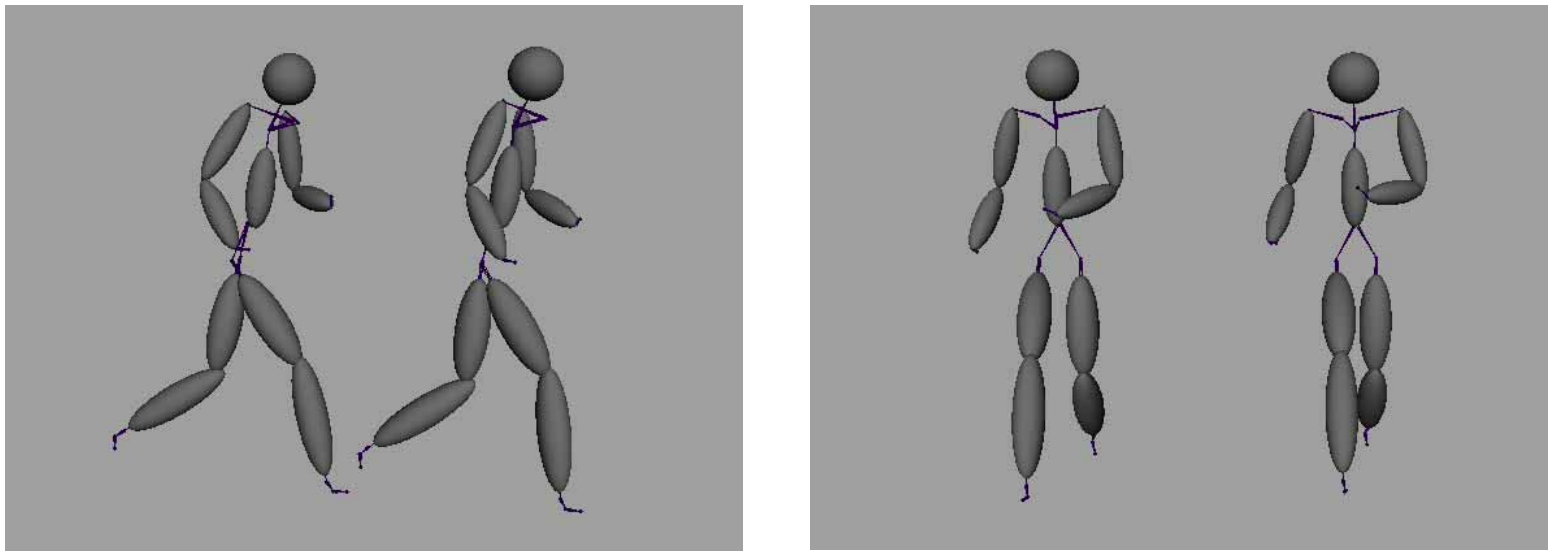
Transitions: poses with left (right) foot on the ground map to similar phases (around unit circle)

9 walk cycles
10 run cycles
different speeds
different subjects



[Urtasun et al. ’07]

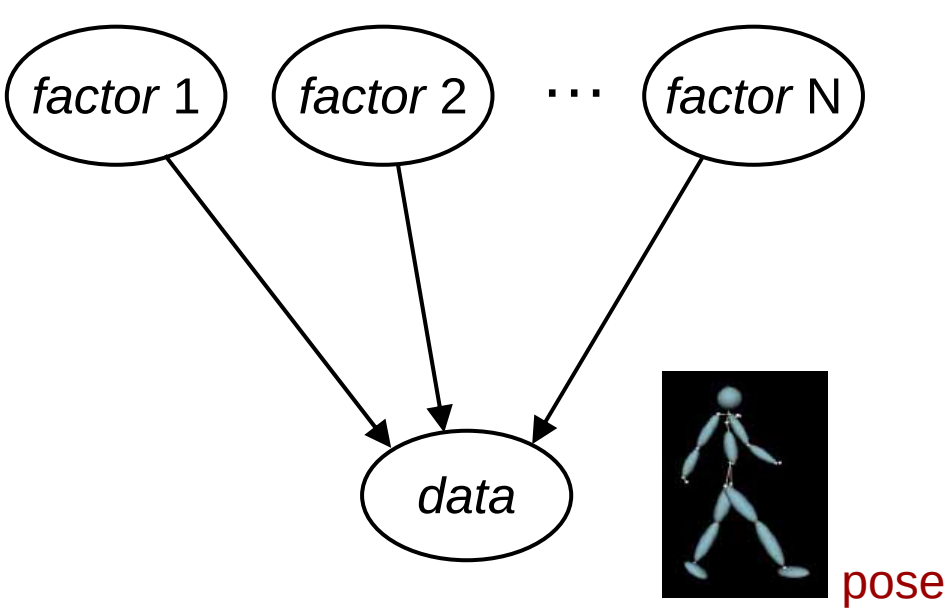
“Cylindrical” topology



Simulations that transition from running to walking

[Urtasun et al. ’07]

Style-content separation



gait,
phase,
identity,
gender,
...

pose

$$y = \sum_{i,j,k,\ldots} w_{ijk\ldots}\, a_i\, b_j\, c_k \cdots + \varepsilon$$

Multilinear style-content models
[Tenenbaum and Freeman ’00;
Vasilescu and Terzopoulos ‘02]

$$y = \sum_{i,j} w_{ij}\, a_i\, \phi_j(\mathbf{b}) + \varepsilon$$

Nonlinear basis functions
[Elgammal and Lee ‘04]

Multifactor GPLVM

Suppose y depends linearly on latent style parameters $s_1, s_2, \ldots$, and nonlinearly on $\mathbf{x}$:

$$y = \sum_i s_i \, g_i(\mathbf{x}) + \varepsilon = \sum_i s_i \, \mathbf{w}_i^T \mathbf{\Phi}(\mathbf{x}) + \varepsilon$$

where $\mathbf{\Phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_{N_x}(\mathbf{x})]^T$

If $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\varepsilon \sim \mathcal{N}(0, \beta^{-1})$, then $y \,|\, \mathbf{x}$ is zero-mean Gaussian, with covariance

$$E[\, yy' \,] = \mathbf{s}^T \mathbf{s}' \; \mathbf{\Phi}(\mathbf{x})^T \mathbf{\Phi}(\mathbf{x}') + \beta^{-1} \delta$$

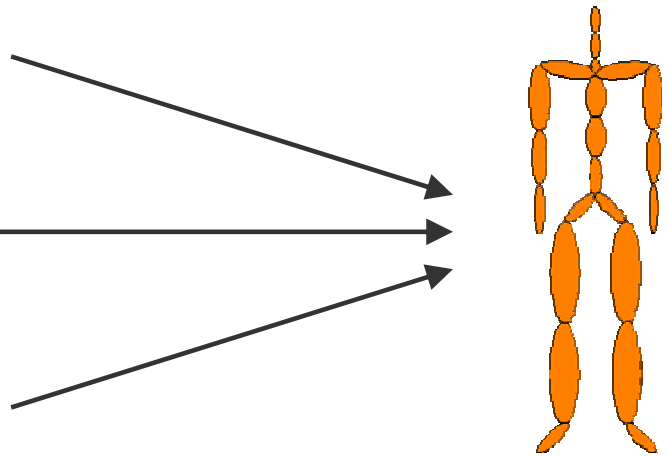
where $\mathbf{s} = [s_1, \ldots, s_{N_s}]^T$

[Wang et al. ’07]

Multifactor locomotion model

Three factor latent model with $\mathcal{X} = \{\mathbf{s}, \mathbf{g}, \mathbf{x}\}$:

- $\mathbf{s}$: identity of the subject performing the motion
- $\mathbf{g}$: gait of the motion (walk, run, stride)
- $\mathbf{x}$: current state of motion (evolves w.r.t. time)

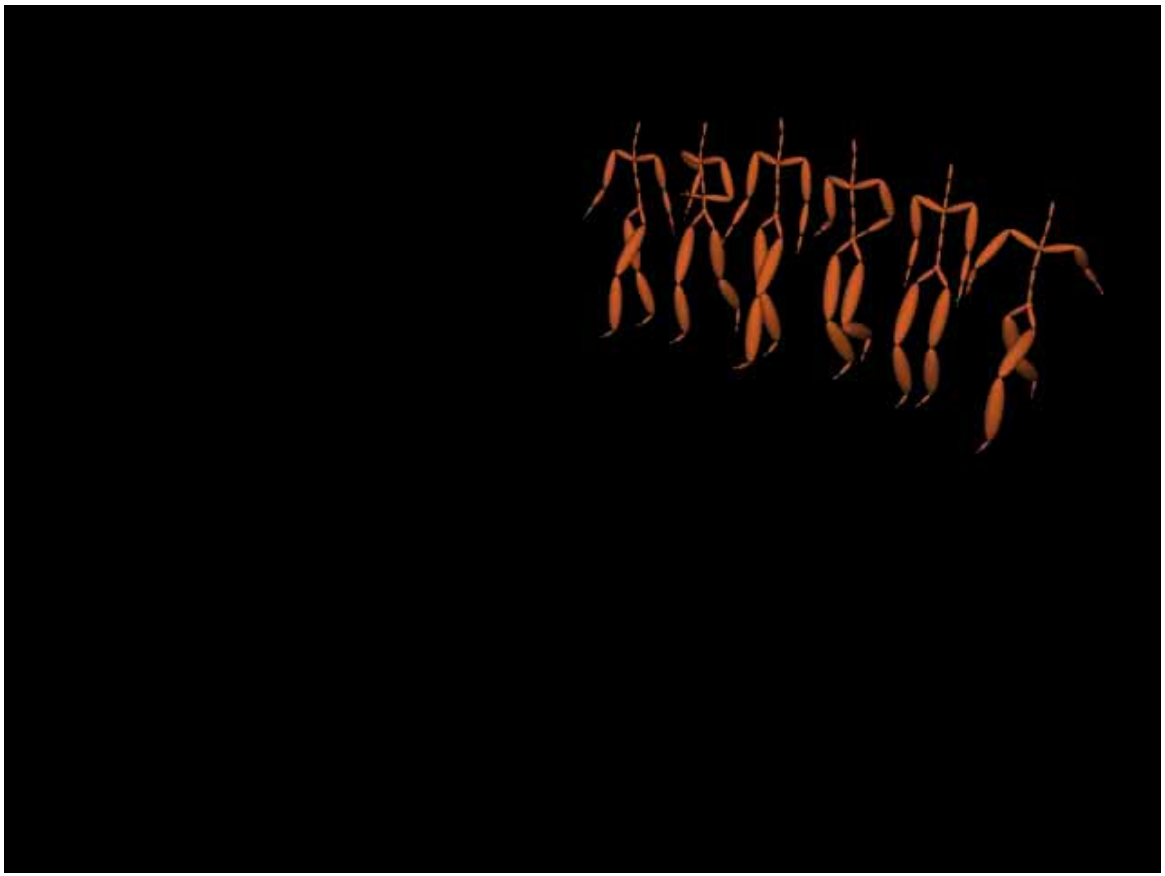


Covariance function:

$$k_d(\mathcal{X}, \mathcal{X}') = \theta_d \, \mathbf{s}^T\mathbf{s}' \, \mathbf{g}^T\mathbf{g}' \, e^{-\frac{\gamma}{2}||\mathbf{x}-\mathbf{x}'||^2} + \beta^{-1}\delta$$

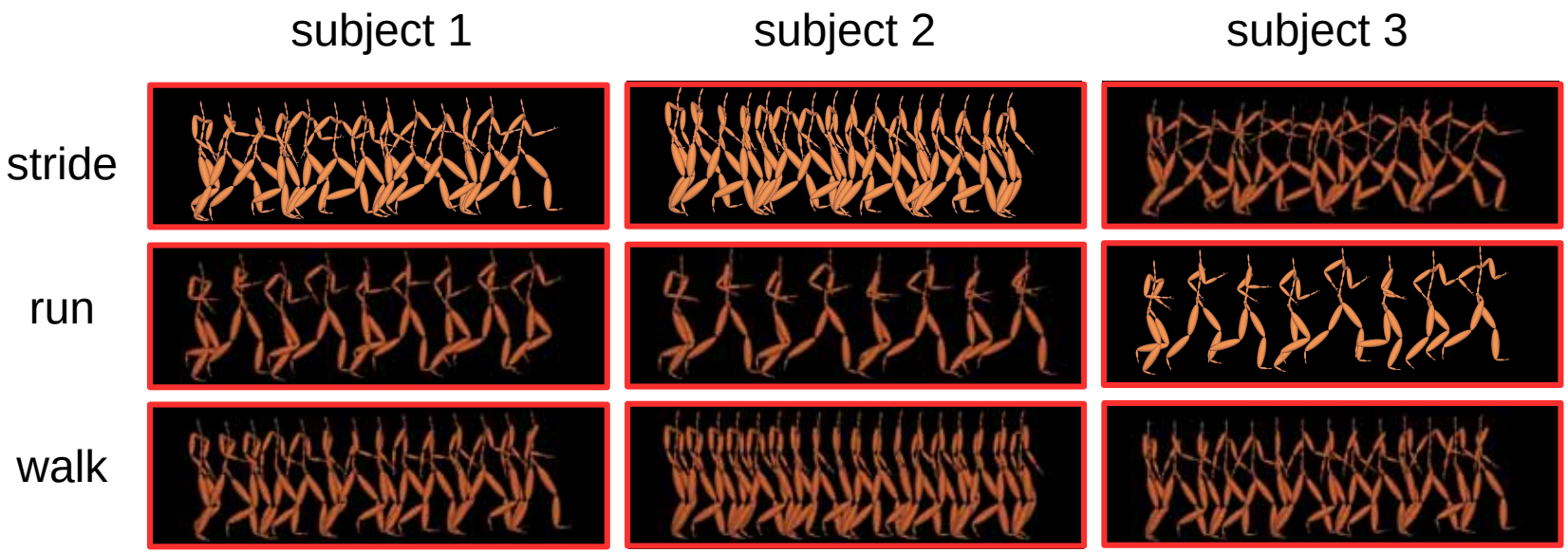
[Wang et al. ’07]

Multifactor locomotion model



Training data: 6 motions, 314 poses in total, $\mathbf{y} \in \mathcal{R}^{89}$

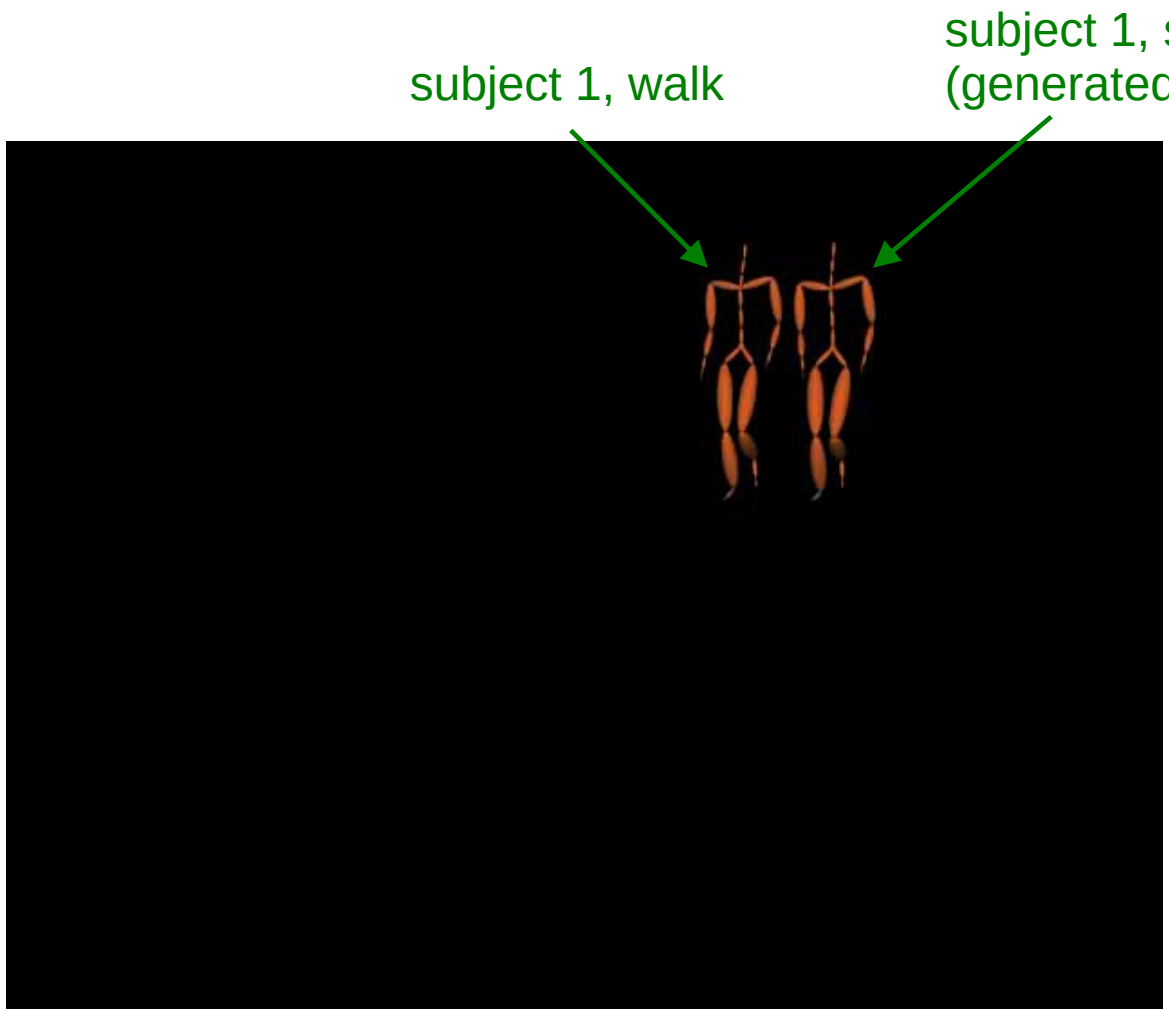
Generating new motions



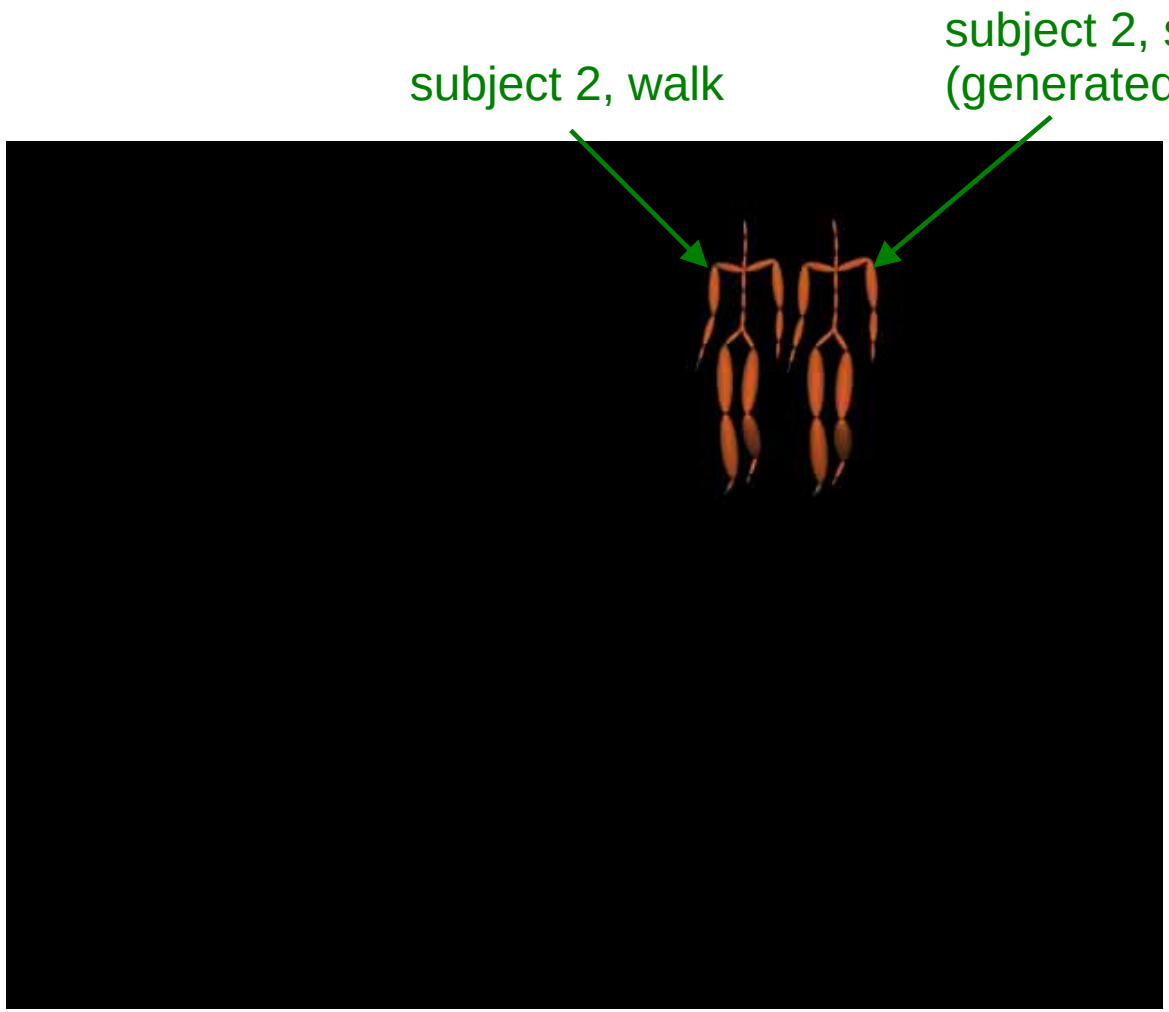
Each training motion is a sequence of poses, sharing the same combination of subject ($\mathbf{s}$) and gait ($\mathbf{g}$).

The GP model provides a Gaussian prediction for new motions. We use the mean to generate motions with different styles.

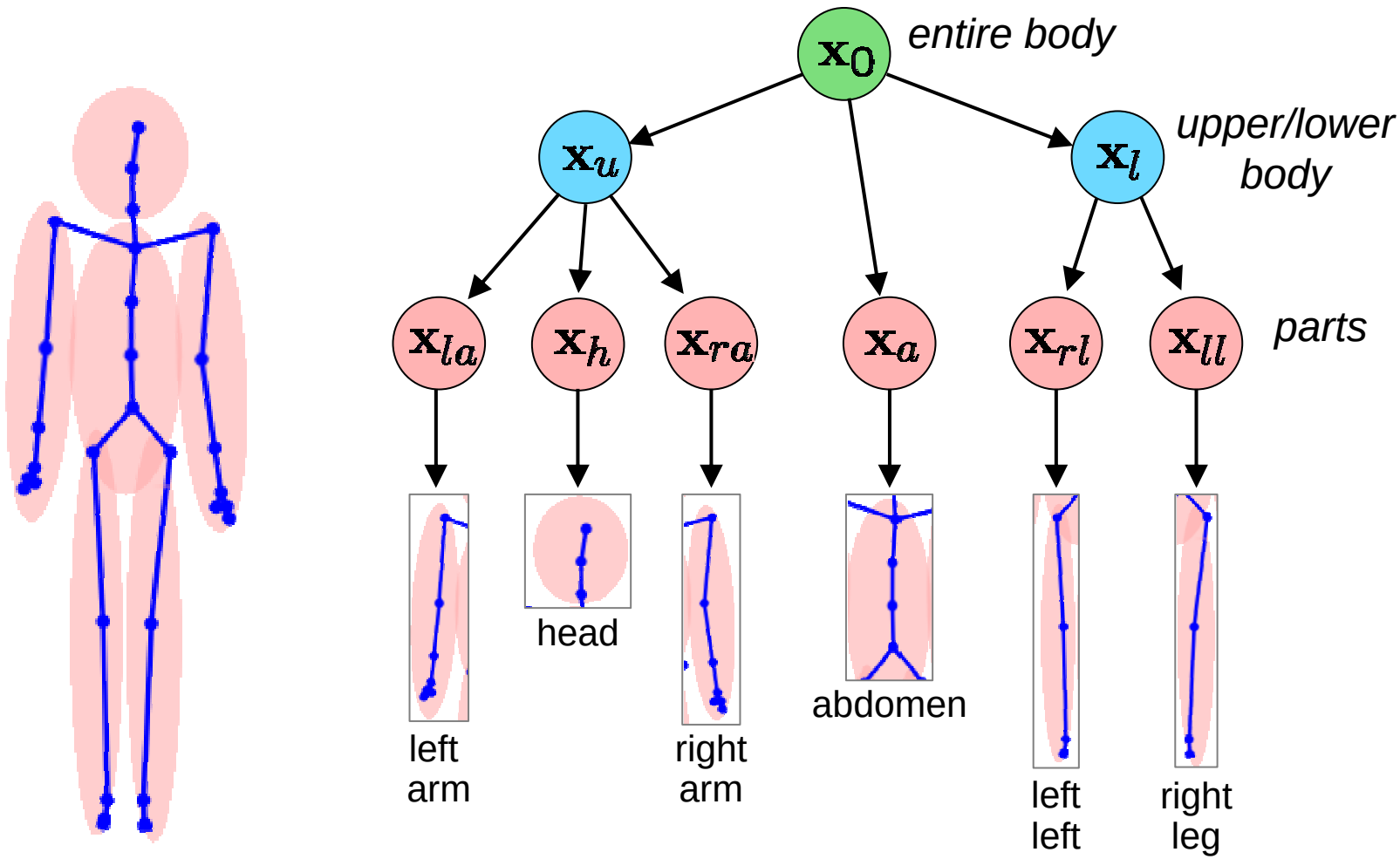
Generating new motions



Generating new motions



Compositionality



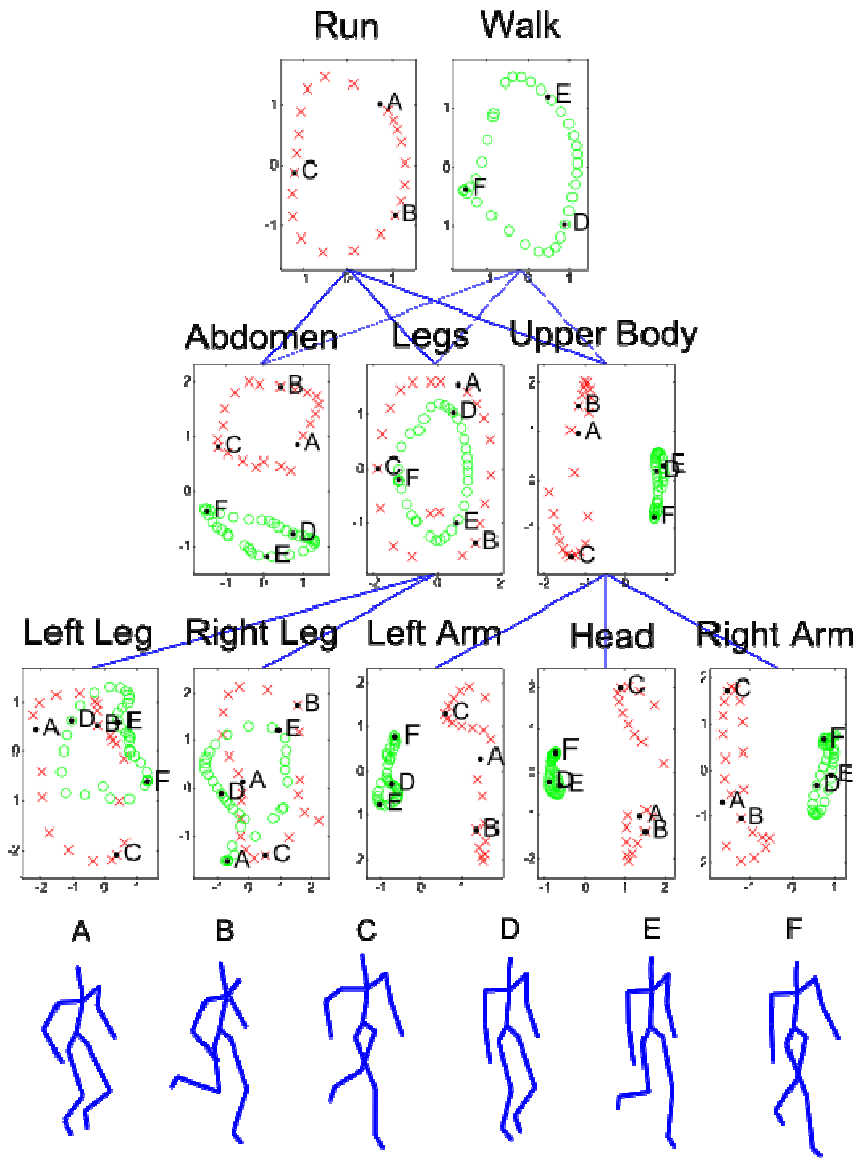
Hierarchical GPLVM [Lawrence and Moore '07]

Compositionality

Data: 1 walk cycle, 1 run cycle

Initialization: PCA

Learning: joint ML optimization of latent coordinates and hyper-parameters at all layers.

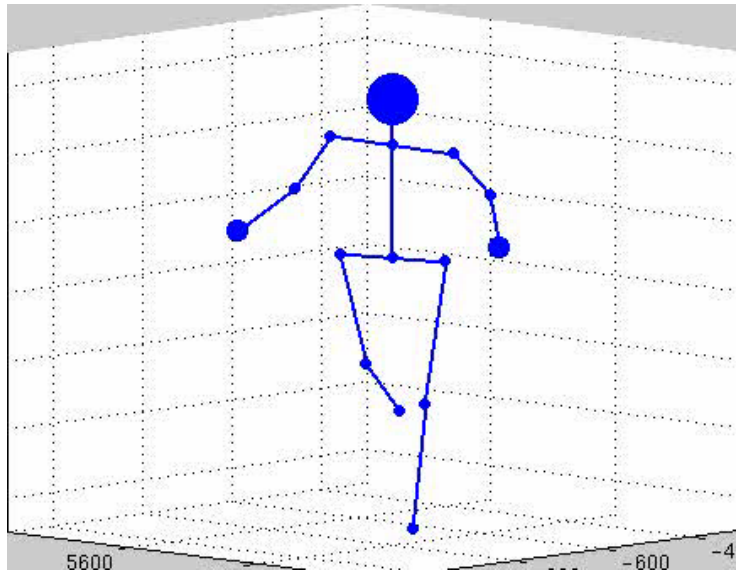


… and beyond

- Scaling / Efficiency
  (e.g.,[Quinonero-Candela and Rasmussen `05])
- Switching models for modeling activity transitions
  (e.g., [Pavlovic et al `00; Li et al `02; Oh et al `05; Li et al. `06])
- …
- Tracking applications

Interactions are important

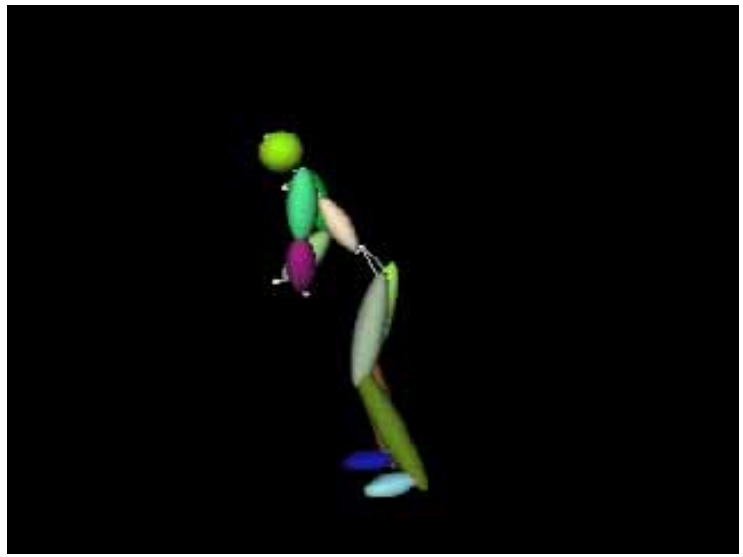
Implausible motion



[Poon and Fleet, 01]

- Kinematic Model: damped second-order Markov model with Beta process noise and joint angle limits
- Observations: steerable pyramid coefficients (image edges)
- Inference: hybrid Monte Carlo particle filter

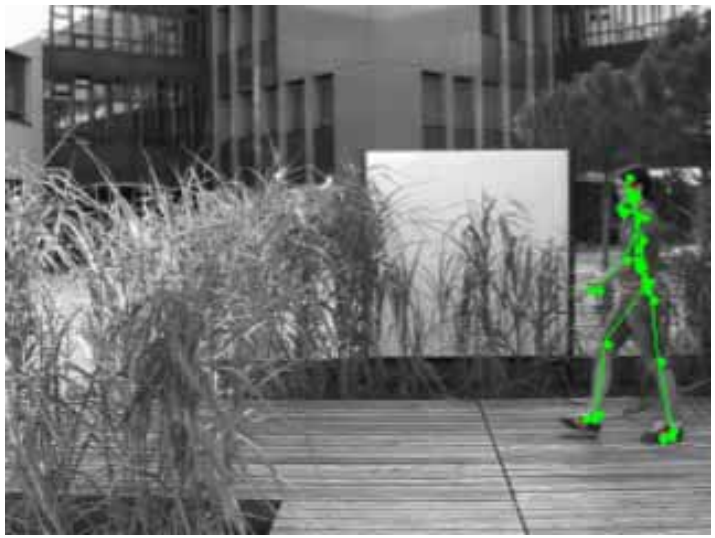
Implausible motion

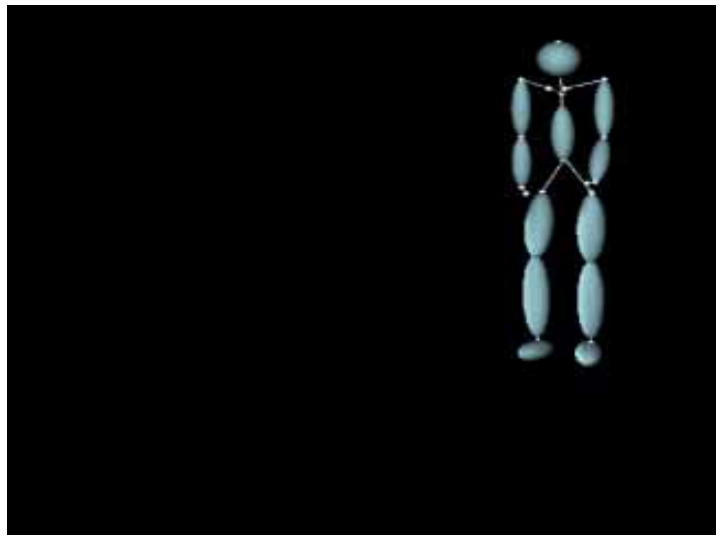


[Urtasun et al. `05]

- Kinematic Model: non-linear latent model of the pose manifold, with second-order Gauss-Markov model for temporal evolution
- Observations: tracked 2D patches on body (WSL tracker)
- Inference: MAP estimation (hill-climbing)

Implausible motion





[Urtasun et al. `06]

- Kinematic Model: Gaussian process dynamical model (GPDM)
- Observations: tracked 2D patches on body (WSL tracker)
- Inference: MAP estimation (hill climbing) with sliding window

Can learning scale?

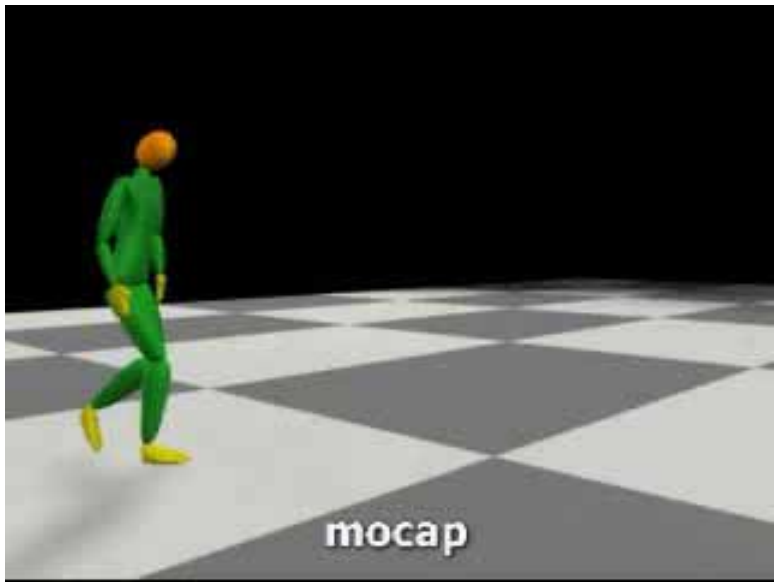
Problem: Learning kinematic models that incorporate dependence on the environment and other bodies from motion capture data may be untenable.

Physics-based models

Physics specifies the motions of bodies and their interactions in terms of inertial descriptions and forces, and generalize naturally to account for:

- balance and body lean (e.g., on hills)
- sudden accelerations (e.g., collisions)
- static contact (e.g., avoiding footskate)
- variations in style due to speed and mass distribution (e.g., carrying an object)
- …

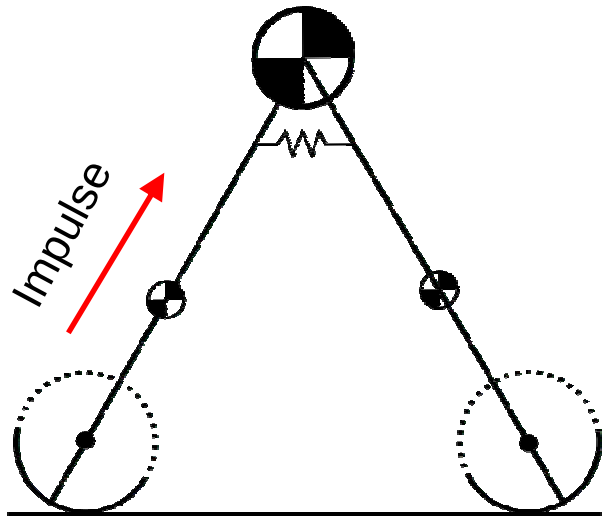
Modeling full-body dynamics is difficult



[Liu et al. `06]

[Kawada Ind. HRP-2, Robodex 2003]

Biomechanics



[McGeer `90; Kuo `01,`02]

Anthropomorphic Walker

- 2D model with rigid bodies for the torso and each leg
- forces can be added with a spring between the legs and an impulsive toe-off

Properties:

- walks efficiently (passively down an incline)
- when powered, exhibits a human-like preferred speed-step length relationship
- invariant to total mass and leg length (approximately)

Physics-based model of lower-body dynamics

Development of physics-based motion models for tracking:

- equations of motions
- prior distribution over forces that model natural 2D locomotion with different speeds, step lengths, …
- a 3D pose model consistent with underlying dynamics

control parameters

2D dynamics

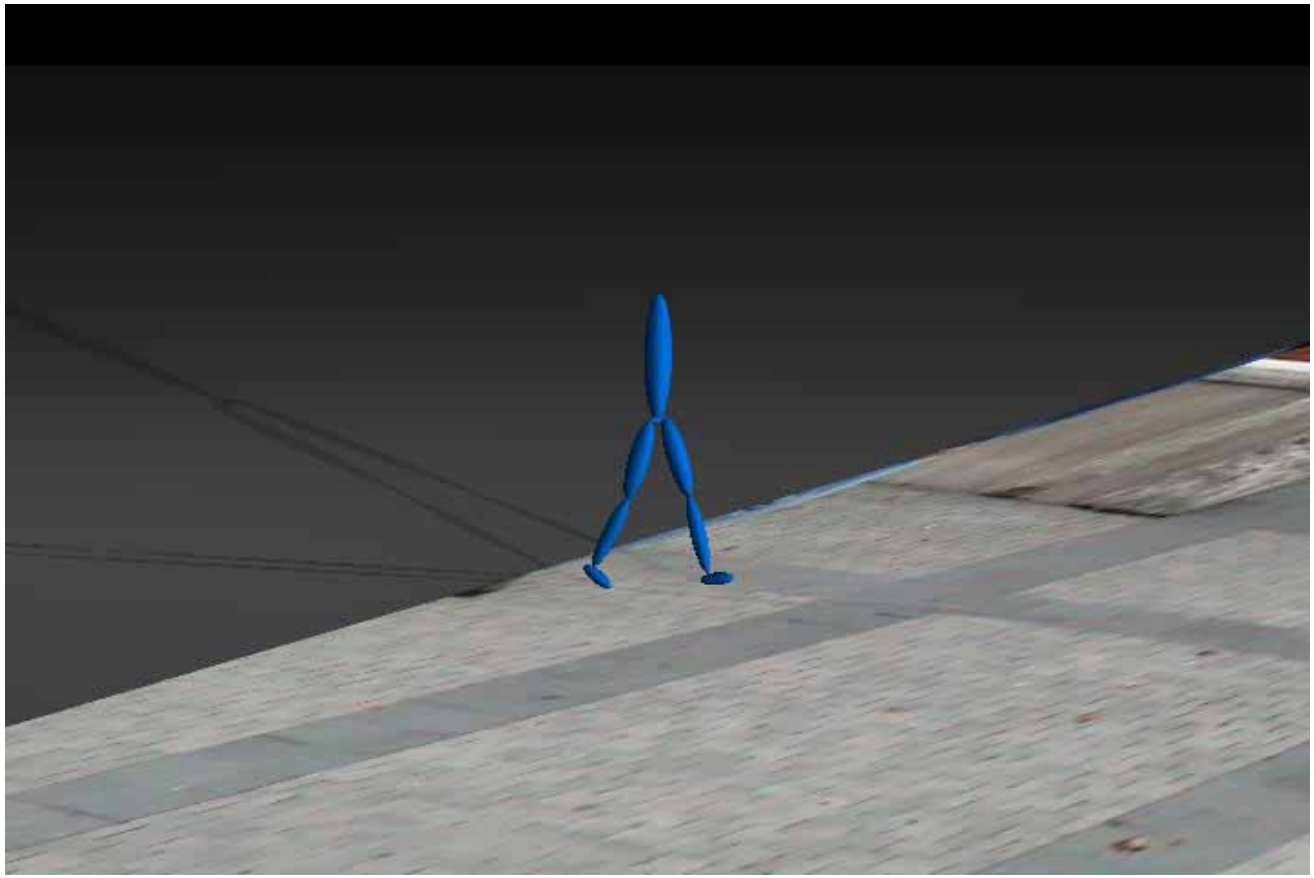
3D kinematics, given dynamics

[Brubaker et al. `07]

Tracking results

Approximate MAP trajectory

Tracking results



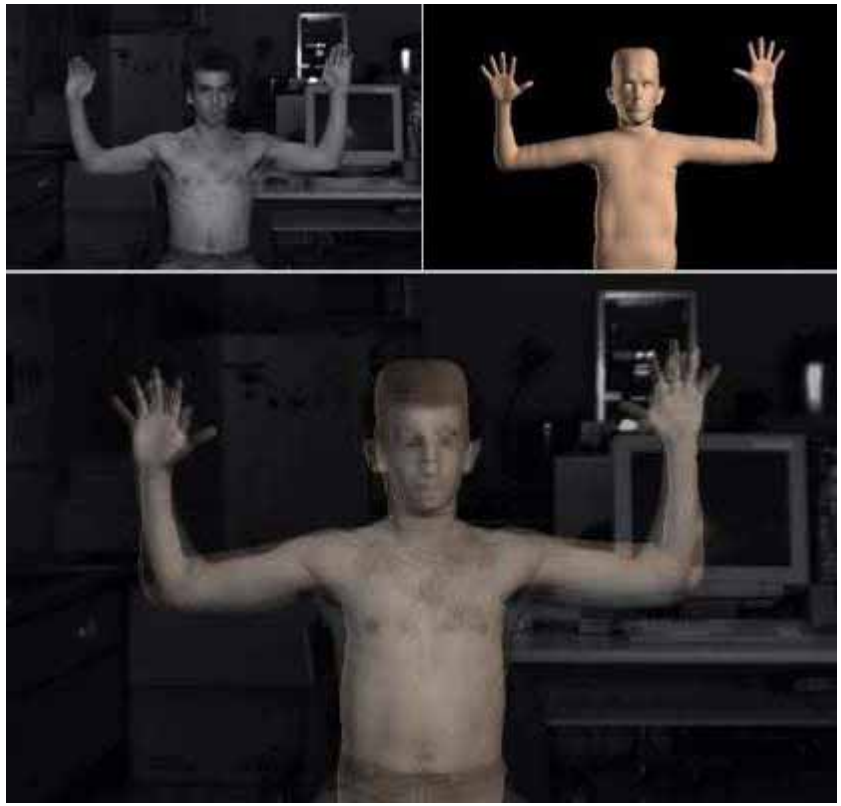
Approximate MAP trajectory in 3D

Limitations (future work)

Our work thus far has just scratched the surface

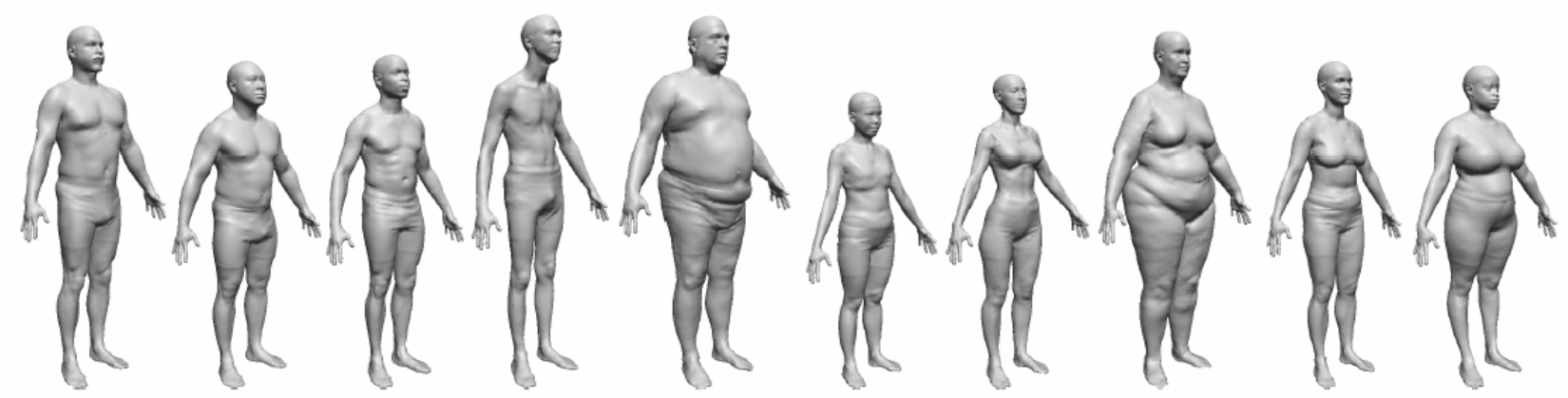
- Knees and torso are needed to help account for bipedal locomotion on stairs, hills, etc.
- 3D models would allow body lean and foot placement in turning, and variations in upper-body moments of inertia
- Extend dynamics to capture standing (both feet in contact) and running (no contact during flight)
- Learning
  - parameters of physics-based models from mocap
  - conditional kinematics.

Modeling appearance: Shape

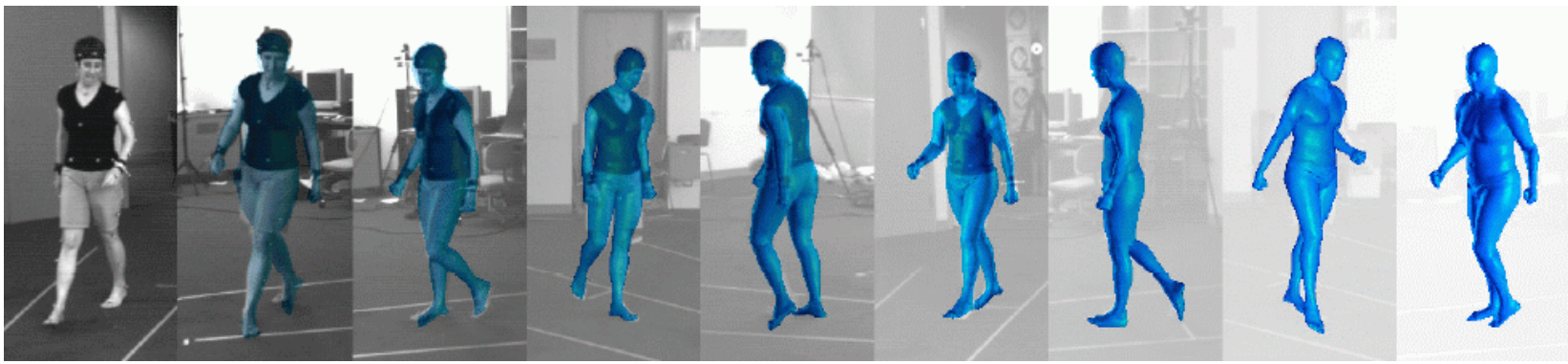


[Plankers and Fua, 2003]

Modeling appearance: Shape

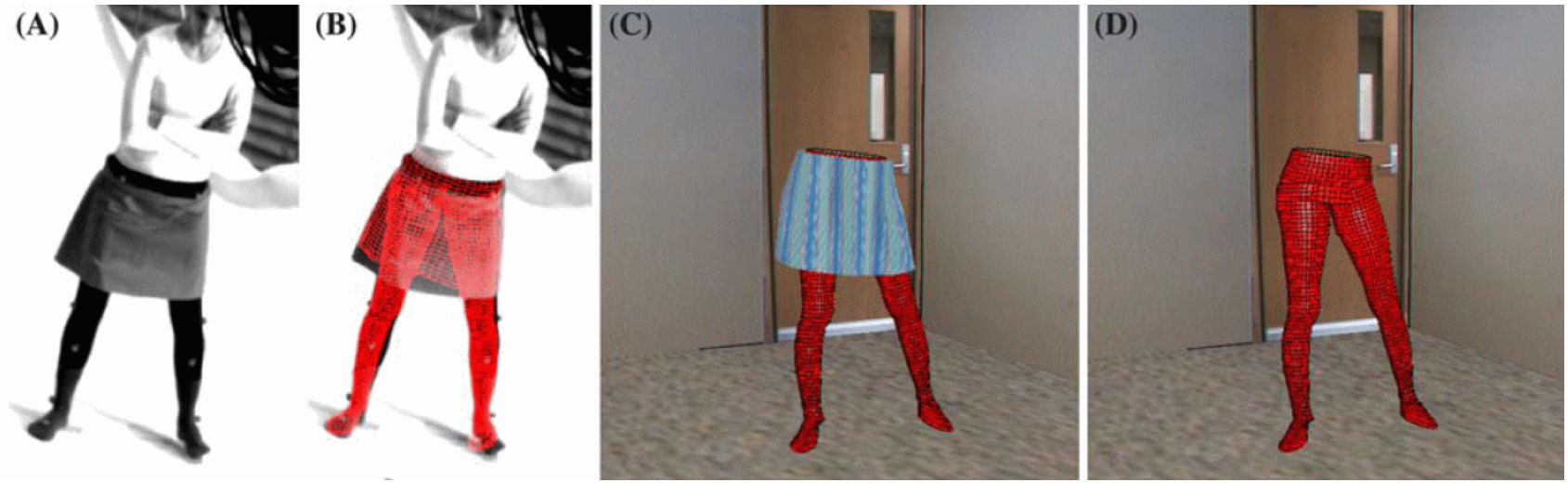


[Allen et al. 2003]



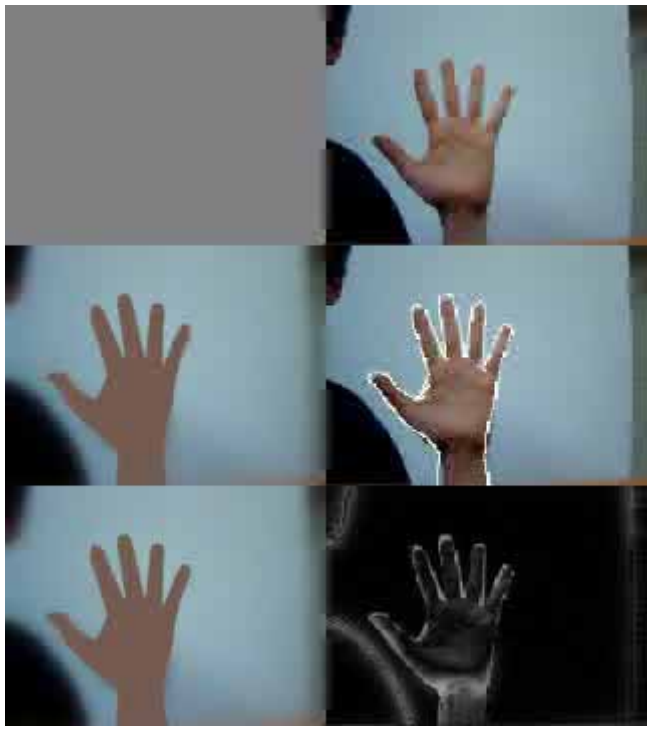
[Balan et al. 2007]

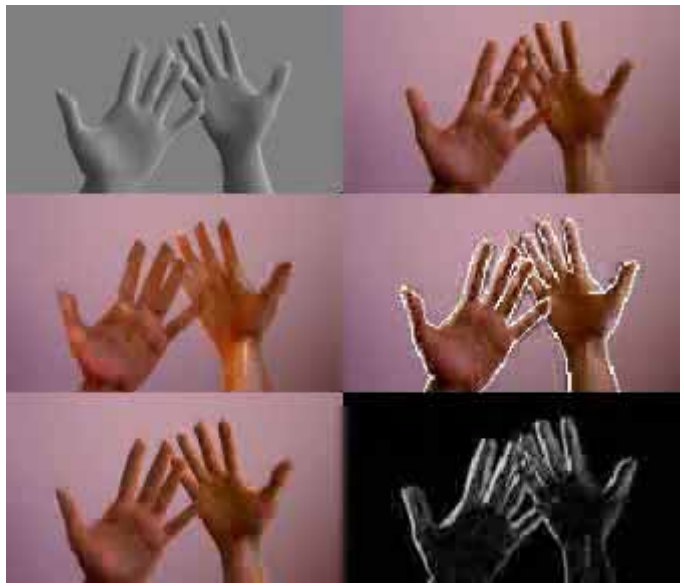
Modeling appearance: Clothing



[Rosenhahn et al. `07]

Modeling appearance: Lighting





[de la Gorce et al. `07]

Conclusions

We need to get a lot of things right to successfully infer 3D pose and motion from monocular video:

- body size and shape
- pose and motion
- appearance (foreground and background)
- lighting and occlusion
- image measurement
- search and detection
- …