Human Pose Tracking II: Kinematic Models

David Fleet
University of Toronto

CIFAR Summer School, 2009

Pose tracking as Bayesian filtering

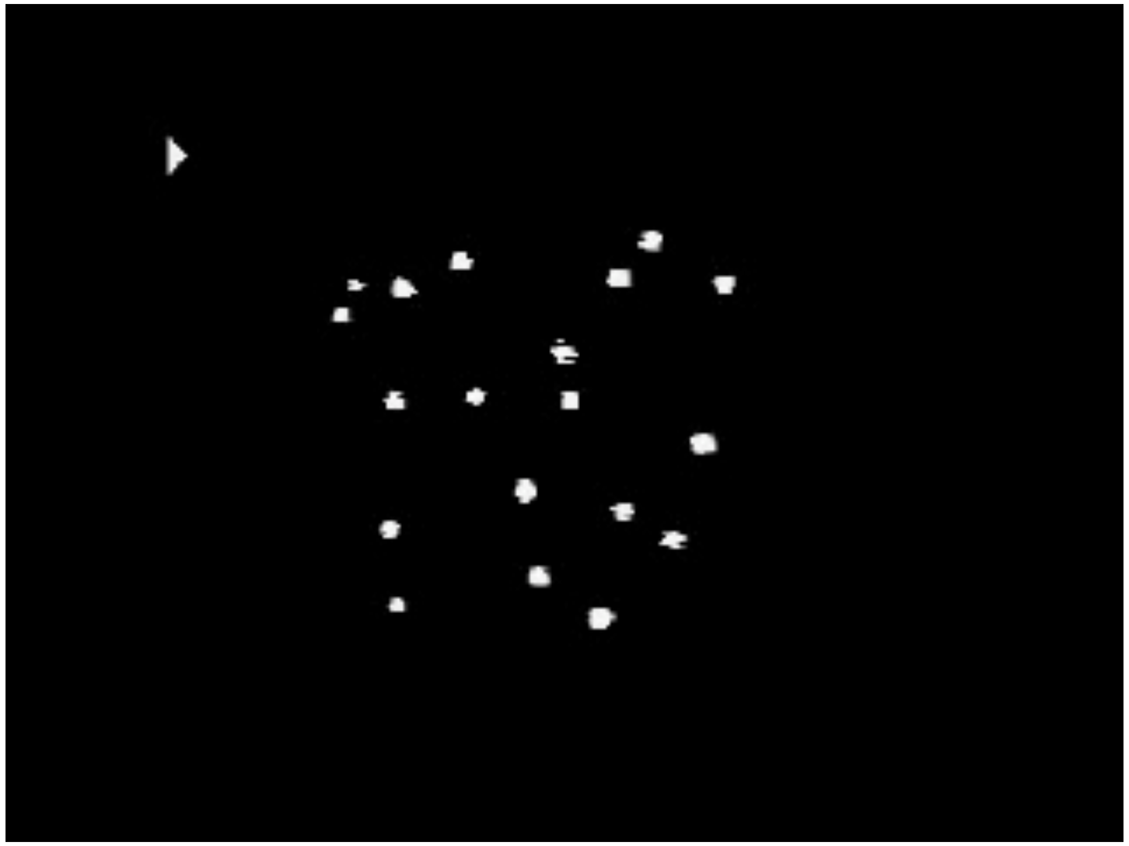
Posterior distribution

$$p(motion \,|\, video) \;=\; \frac{p(video \,|\, motion)\; p(motion)}{p(video)}$$

Filtering distribution

$$p(pose_t \,|\, images_{1:t}) \;=\; \frac{p(image_t \,|\, pose_t)\;\; p(pose_t \,|\, images_{1:t-1})}{p(images_{1:t})}$$

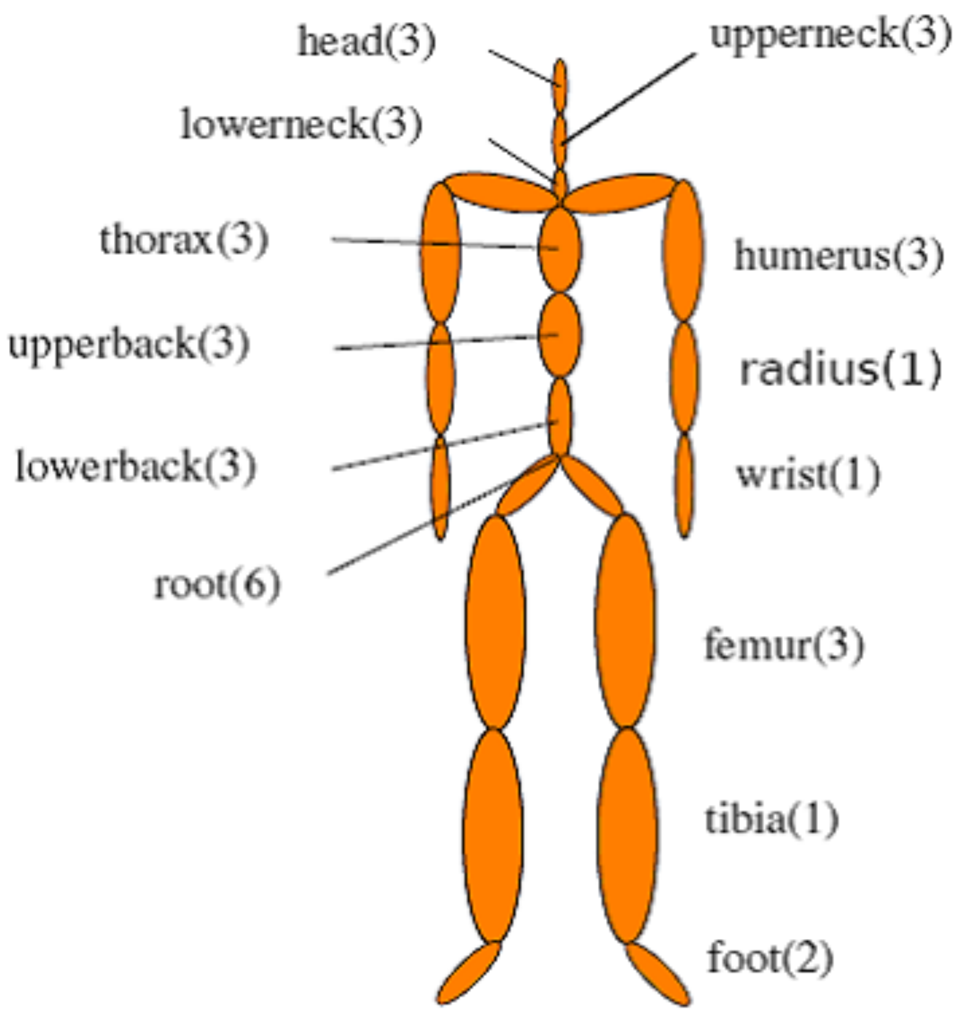
Motion capture data



[Johansson, ’73]

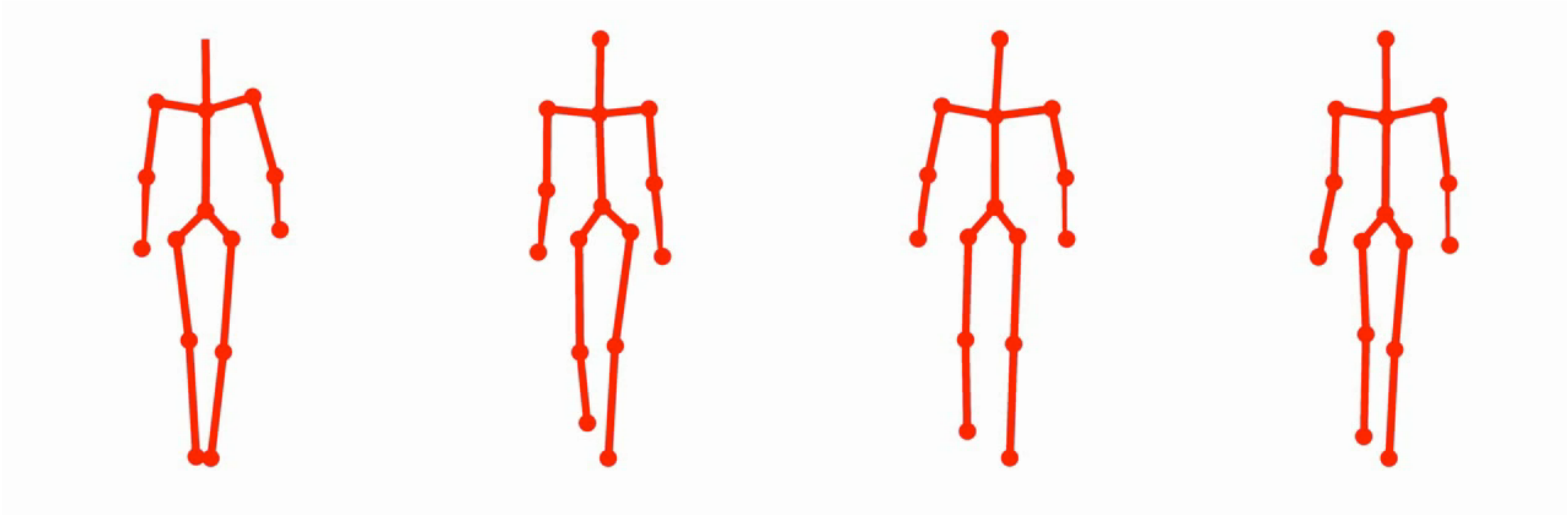
Motion capture data

motion capture

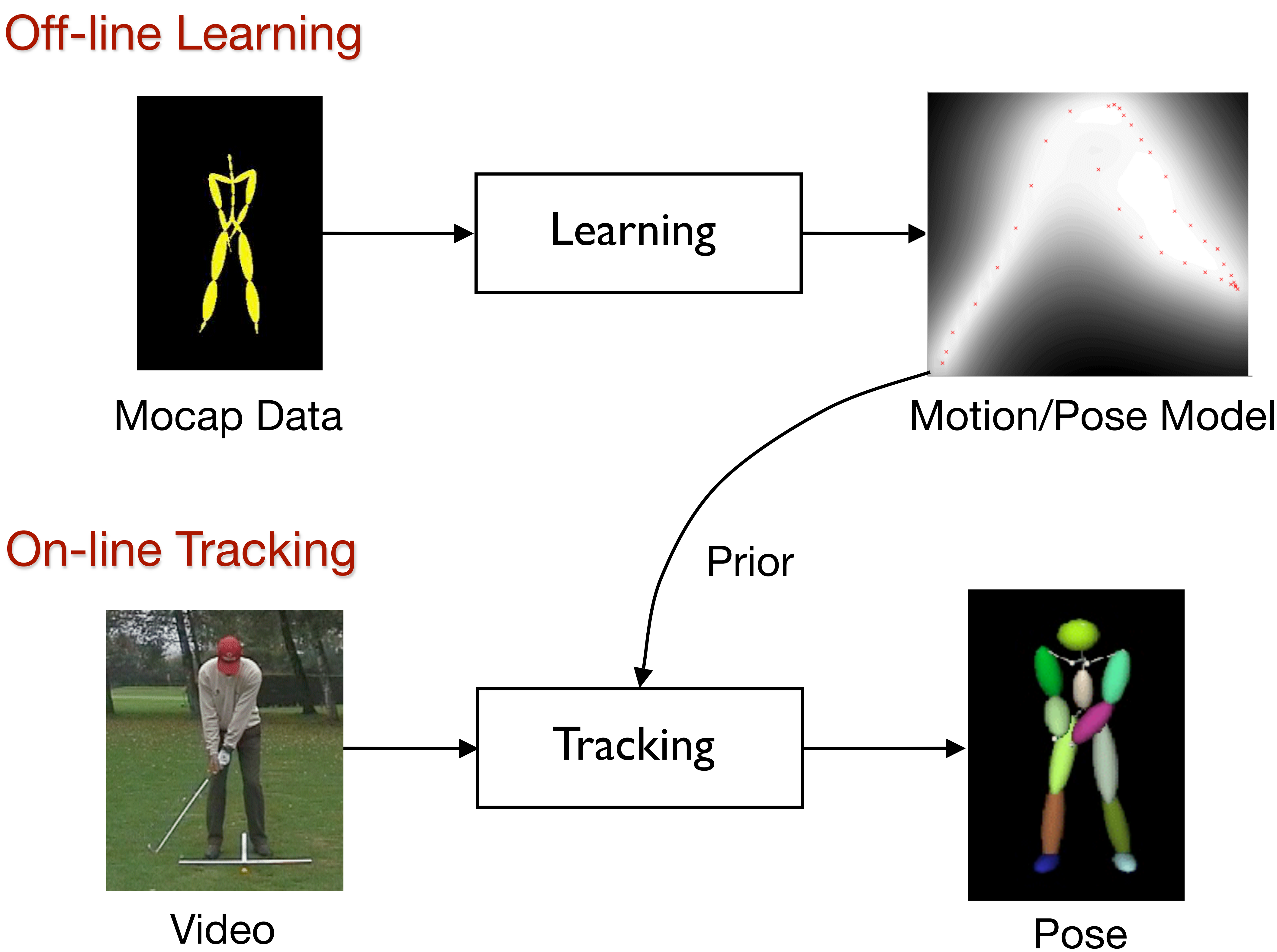


3D articulated model

Motion capture data

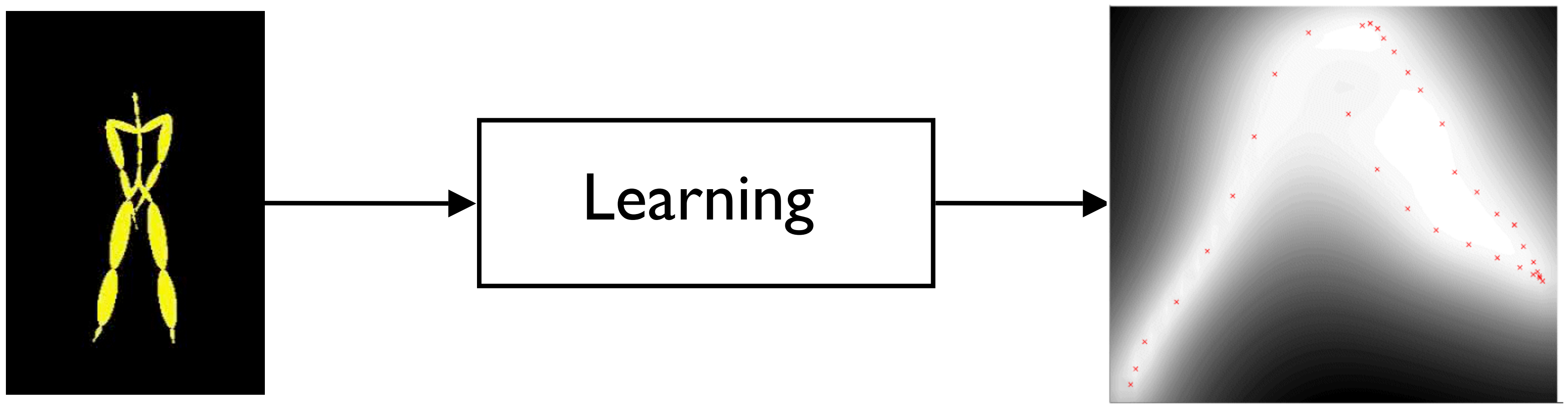


Model-based pose tracking



Model-based pose tracking

Off-line Learning

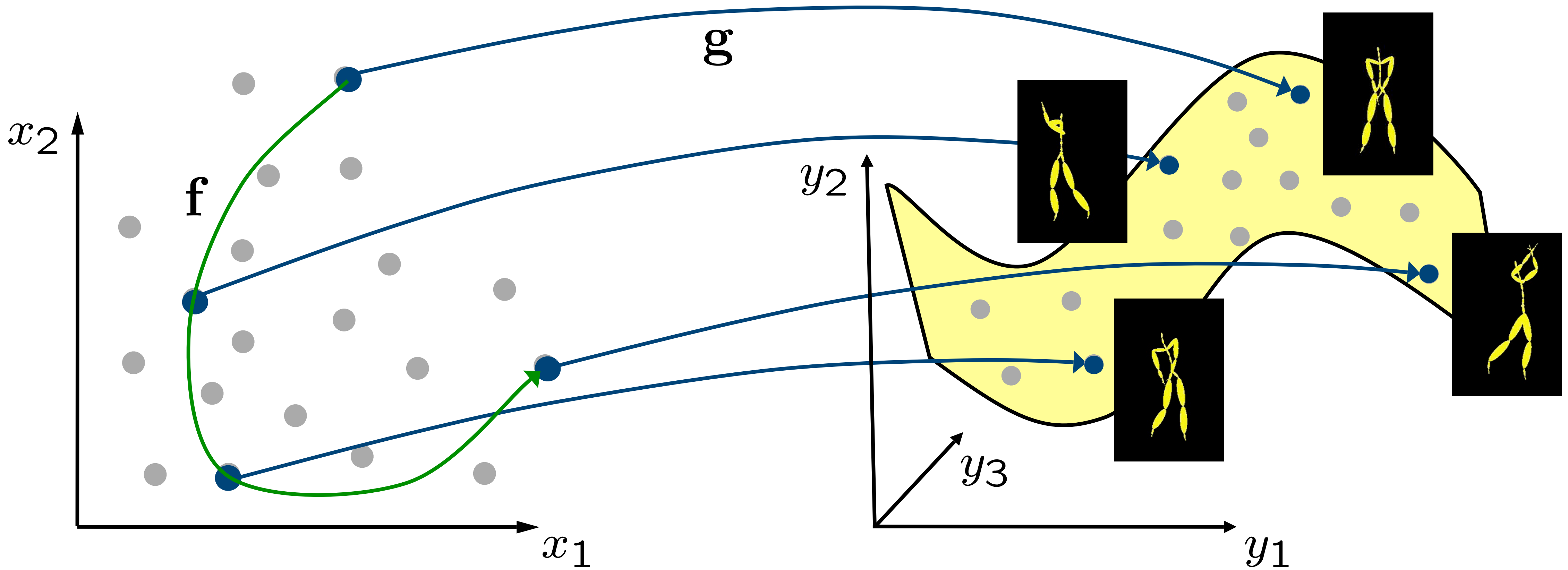


Mocap Data

Motion/Pose Model

Problem: Human pose data are high-dimensional, and difficult to obtain, so over-fitting and generalization are significant issues in learning useful models.

Latent variable models



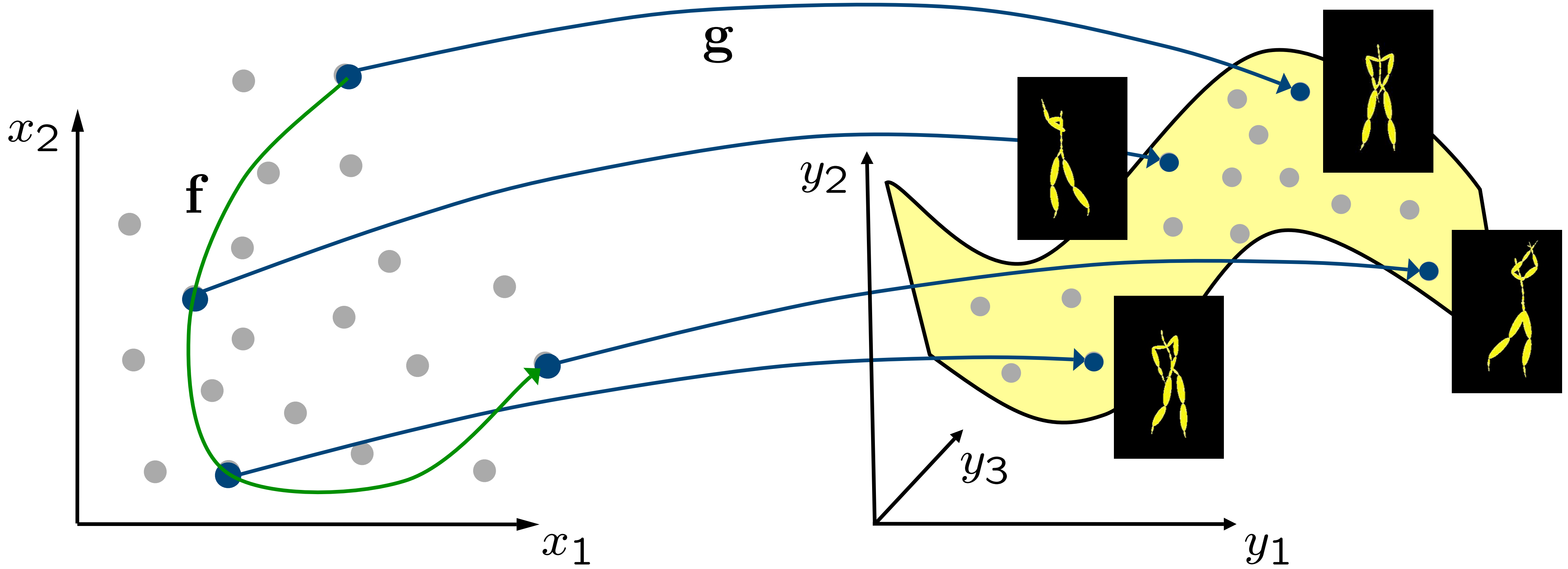
Low-dim. latent space $(\mathbf{x})$ Joint angle pose space $(\mathbf{y})$

Mapping from latent positions to poses, $\mathbf{g}$

Latent dynamical model, $\mathbf{f}$

Density function over pose and motion (latent trajectories)

Latent variable models



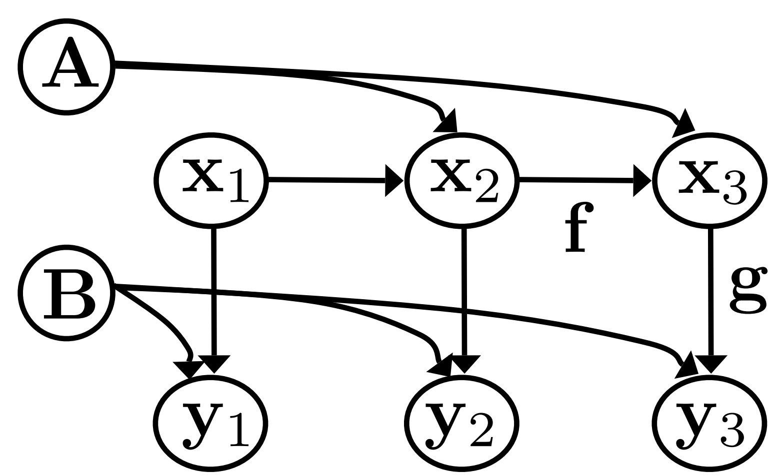
Low-dim. latent space $(\mathbf{x})$

Joint angle pose space $(\mathbf{y})$

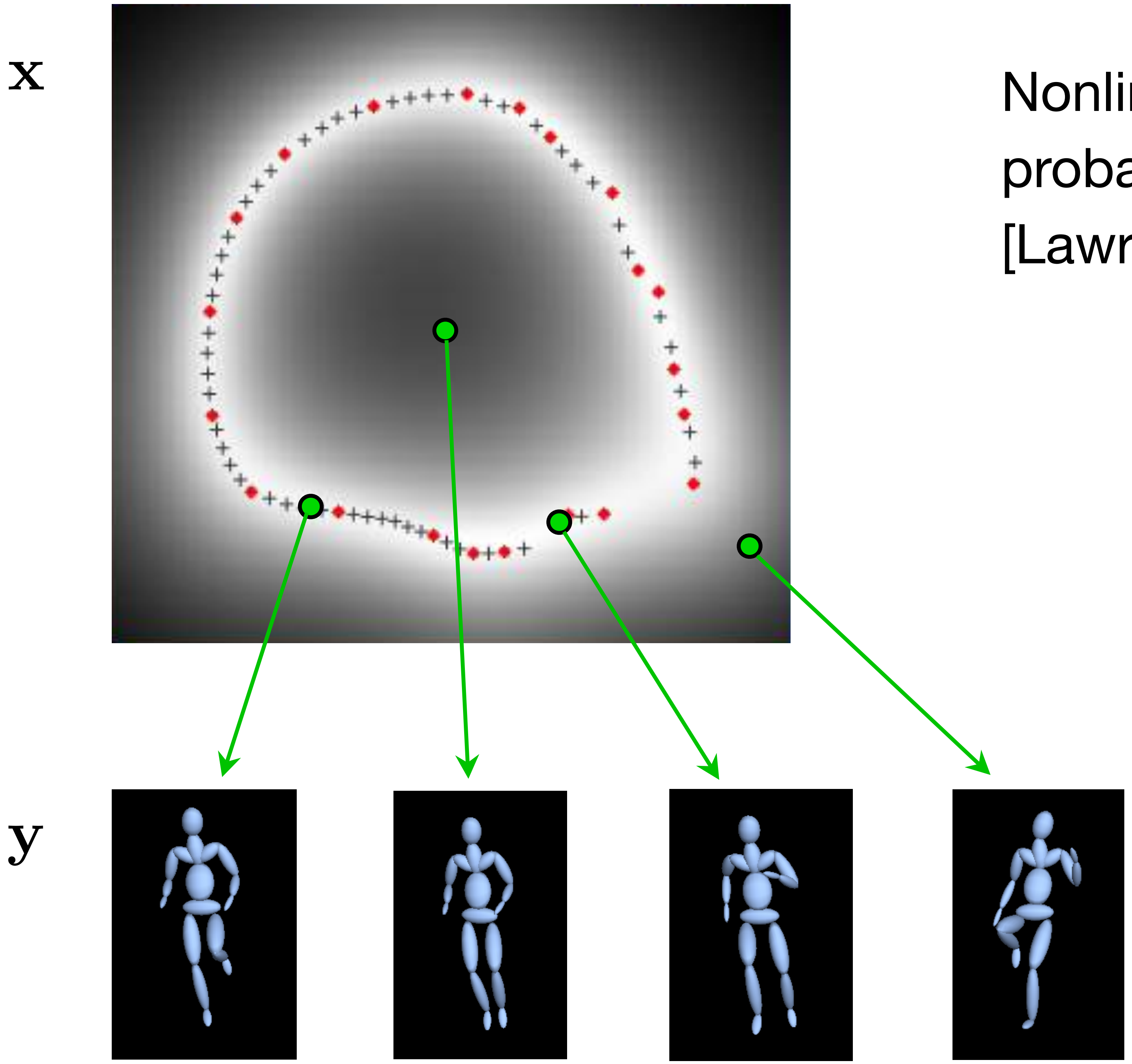
Linear dynamical system:

$$\mathbf{x}_t = \mathbf{f}(\mathbf{x}_{t-1}; \mathbf{A}) + \mathbf{n}_{x,t}$$

$$\mathbf{y}_t = \mathbf{g}(\mathbf{x}_t; \mathbf{B}) + \mathbf{n}_{y,t}$$

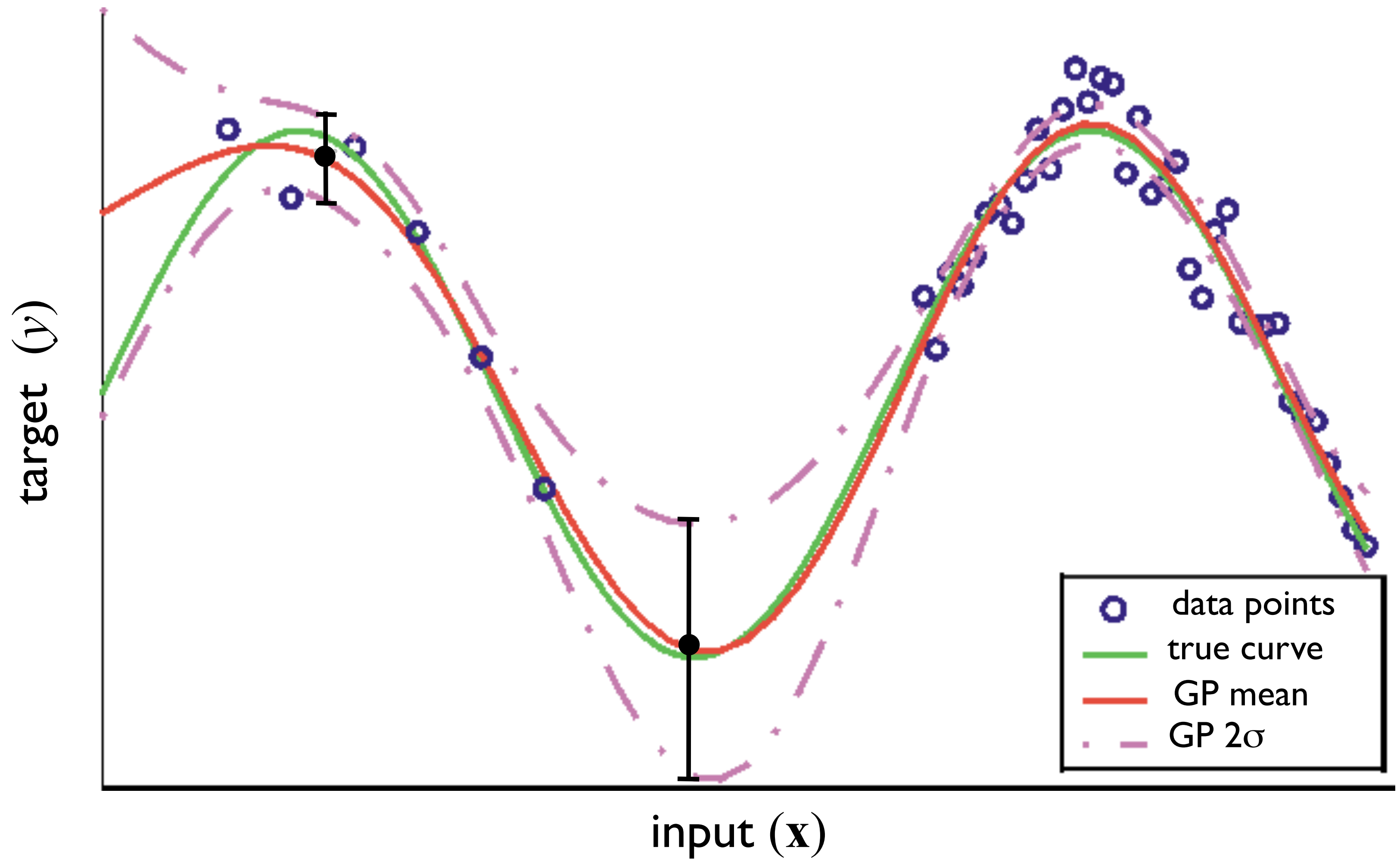


Gaussian Process Latent Variable Model



Nonlinear generalization of probabilistic PCA [Lawrence `05].

Gaussian Process



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

Gaussian Process

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

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

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

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

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

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 Model (GPLVM)

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 $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^T$:

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

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.

Learning: Maximize log likelihood of the data to find latent positions and kernel hyper-parameters, given an initial guess (e.g., use 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:

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

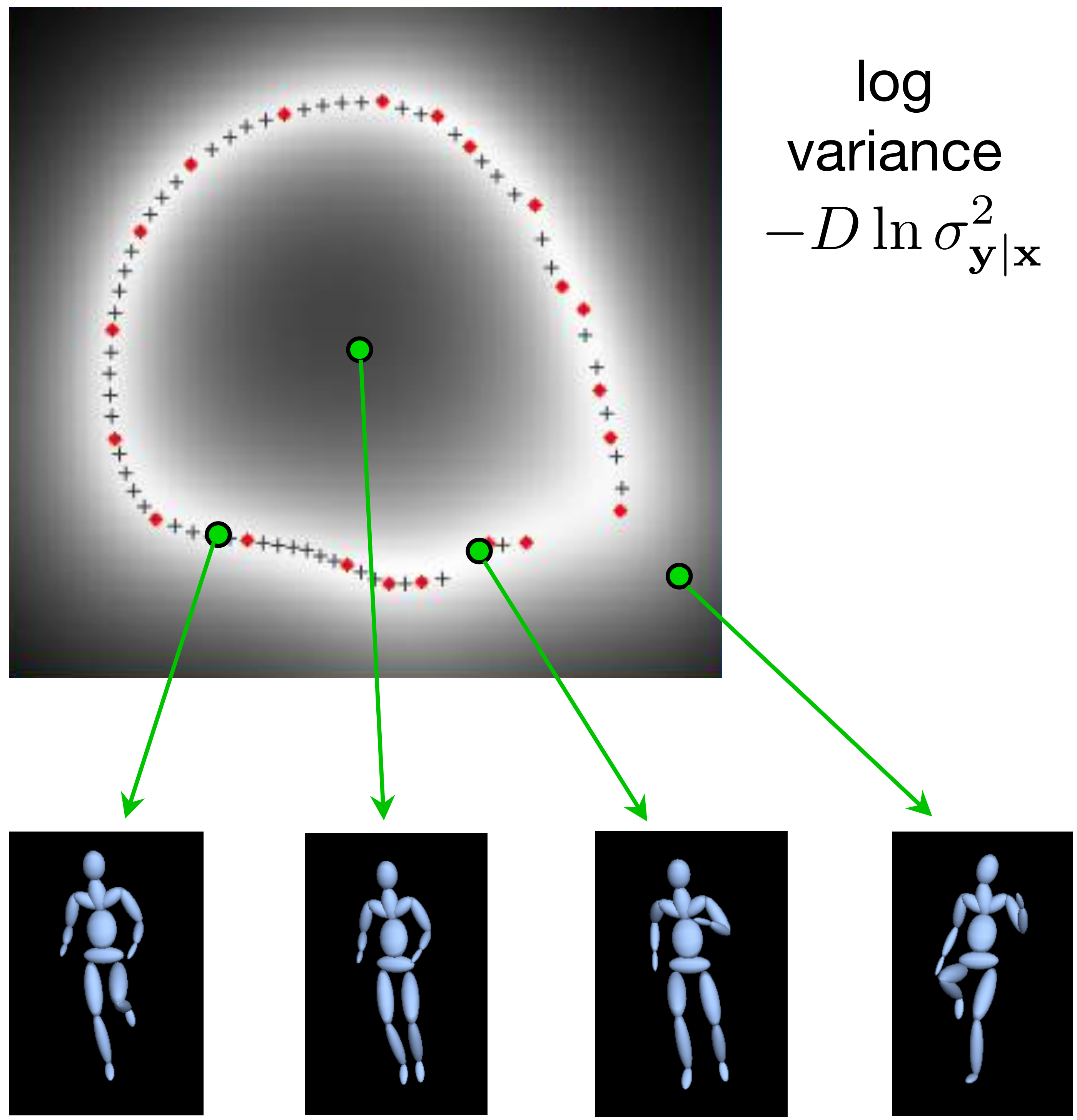
where

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

Gaussian Process Latent Variable Model

$\mathbf{x}$

log variance $-D \ln \sigma^2_{\mathbf{y}|\mathbf{x}}$



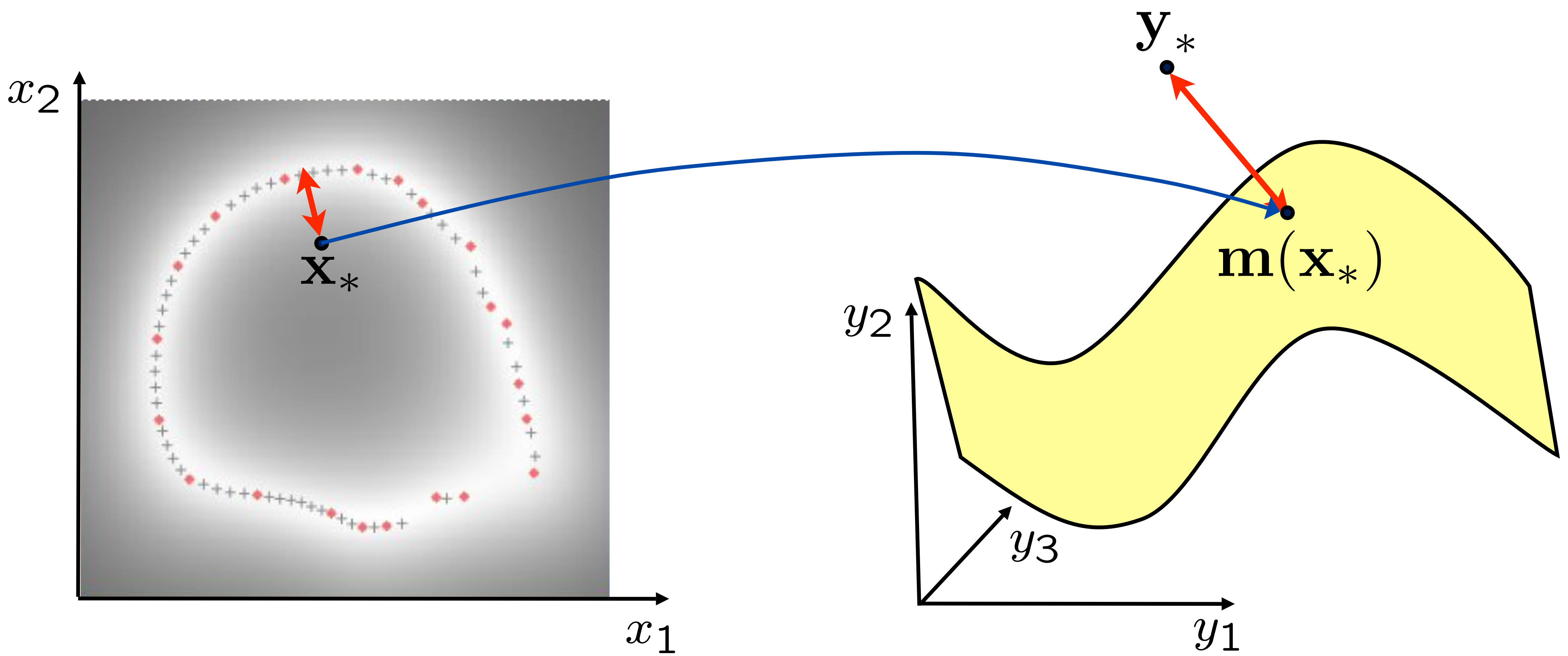
$\mathbf{y}$

mean pose $\mathbf{m}(\mathbf{x})$

Conditional (predictive) distribution

The negative log density for a new pose, given $\mathcal{M} \equiv (\mathbf{Y}, \mathbf{X})$, has a simple form:

$$L(\mathbf{x}_*, \mathbf{y}_*; \mathcal{M}) = \frac{\|\mathbf{y}_* - \mathbf{m}(\mathbf{x}_*)\|^2}{2\sigma^2(\mathbf{x}_*)} + \frac{D}{2} \ln \sigma^2(\mathbf{x}_*)$$

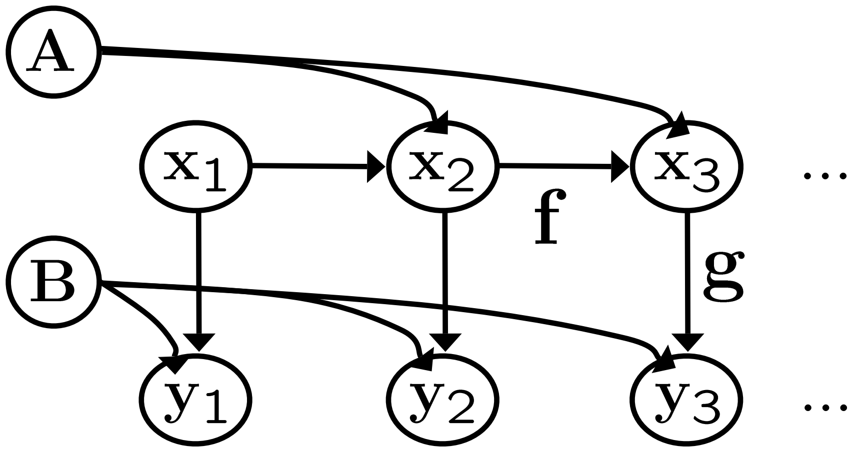


Latent Space ($\mathbf{X}$)

Pose Space ($\mathbf{Y}$)

Gaussian Process Dynamical Model (GPDM)

Latent dynamical model *[Wang et al 05]*:



$$\begin{aligned}\mathbf{x}_t &= \mathbf{f}(\mathbf{x}_{t-1};\mathbf{A}) + \mathbf{n}_{x,t} \\ \mathbf{y}_t &= \mathbf{g}(\mathbf{x}_t;\mathbf{B}) + \mathbf{n}_{y,t}\end{aligned}$$

Assume IID Gaussian noise, and

$$\begin{aligned}\mathbf{f}(\mathbf{x};\mathbf{A}) &= \sum_i \mathbf{a}_i\,\phi_i(\mathbf{x}) \\ \mathbf{g}(\mathbf{x};\mathbf{B}) &= \sum_j \mathbf{b}_j\,\psi_j(\mathbf{x})\end{aligned}$$

with Gaussian priors on $\mathbf{A} \equiv \{\mathbf{a}_i\}$ and $\mathbf{B} \equiv \{\mathbf{b}_j\}$

Marginalize out $\{\mathbf{a}_i, \mathbf{b}_j\}$, and then optimize the latent positions, $\{\mathbf{x}, ..., \mathbf{x}_N\}$, to simultaneously minimize pose reconstruction error and prediction error on training sequence $\{\mathbf{y}, ..., \mathbf{y}_N\}$.

Reconstruction

The data likelihood for the reconstruction mapping, given centered inputs $\mathbf{Y} \equiv [\mathbf{y}, ..., \mathbf{y}_N]^T, \ \mathbf{y}_n \in \mathcal{R}^D$ has the form:

$$p(\mathbf{Y} \,|\, \mathbf{X}, \vec{\beta}, \mathbf{W}) \ = \ \frac{|\mathbf{W}|^N}{\sqrt{(2\pi)^{ND} |\mathbf{K}_Y|^D}} \exp\left(-\frac{1}{2} tr(\mathbf{K}_Y^{-1} \mathbf{Y} \mathbf{W}^2 \mathbf{Y}^T) \right)$$

where

$\mathbf{K}_Y$ is a kernel matrix shared across pose outputs, with entries $(\mathbf{K}_Y)_{ij} = k_Y(\mathbf{x}_i, \mathbf{x}_j)$ for kernel function

$$k_Y(\mathbf{x}, \mathbf{x}') \ = \ \beta_1 \exp\left(-\frac{\beta_2}{2} ||\mathbf{x} - \mathbf{x}'||^2 \right) + \beta_3^{-1} \delta_{\mathbf{x},\mathbf{x}'}$$

with hyperparameters $\vec{\beta} = \{\beta_1, \beta_2, \beta_3\}$

$\mathbf{W} \equiv \mathrm{diag}(w_1, ..., w_D)$ scales the different pose parameters

Dynamics

The latent dynamic process on $\mathbf{X} \equiv [\mathbf{x}, ..., \mathbf{x}_N]^T, \;\; \mathbf{x}_n \in \mathcal{R}^d$ has a similar form:

$$p(\mathbf{X} \,|\, \vec{\alpha}) \;\;=\;\; \frac{\mathcal{N}(\mathbf{x}_1; \mathbf{0}, \mathbf{I}_d)}{\sqrt{(2\pi)^{(N-1)\,d}\,|\mathbf{K}_X|^d}} \exp\left(-\frac{1}{2} tr(\mathbf{K}_X^{-1} \hat{\mathbf{X}} \hat{\mathbf{X}}^T)\right)$$

where

$\hat{\mathbf{X}} = [\mathbf{x}_2, ..., \mathbf{x}_N]^T$

$\mathbf{K}_X$ is a kernel matrix defined by kernel function

$$k_X(\mathbf{x}, \mathbf{x}') \;=\; \alpha_1 \exp\left(-\frac{\alpha_2}{2} ||\mathbf{x} - \mathbf{x}'||^2\right) + \alpha_3 \mathbf{x}^T \mathbf{x}' + \alpha_4^{-1} \delta_{\mathbf{x}'}$$

with hyperparameters $\vec{\alpha}$

Learning

GPDM posterior:

$$p(\mathbf{Y}, \mathbf{X}, \underbrace{\bar{\alpha}, \bar{\beta}, \mathbf{W}}_{\text{kernel hyperparameters}}) \;=\; \underset{\text{reconstruction likelihood}}{p(\mathbf{Y} \,|\, \mathbf{X}, \bar{\beta}, \mathbf{W})} \; \underset{\text{dynamics likelihood}}{p(\mathbf{X} \,|\, \bar{\alpha})} \; \underset{\text{priors}}{p(\bar{\alpha})\; p(\bar{\beta})}$$

training motions   latent trajectories   kernel hyperparameters

To estimate the latent coordinates & kernel parameters we minimize

$$\mathcal{L} \;=\; -\ln p(\mathbf{X}, \bar{\alpha}, \bar{\beta}, \mathbf{W} \,|\, \mathbf{Y})$$

with respect to $\mathbf{X}$, $\bar{\alpha}$, $\bar{\beta}$ and $\mathbf{W}$.

GPDM prior over new poses and motions

The model $\mathcal{M} \equiv (\mathbf{Y}, \mathbf{X}, \vec{\alpha}, \vec{\beta}, \mathbf{W})$ then provides a density function over new poses, with negative log likelihood:

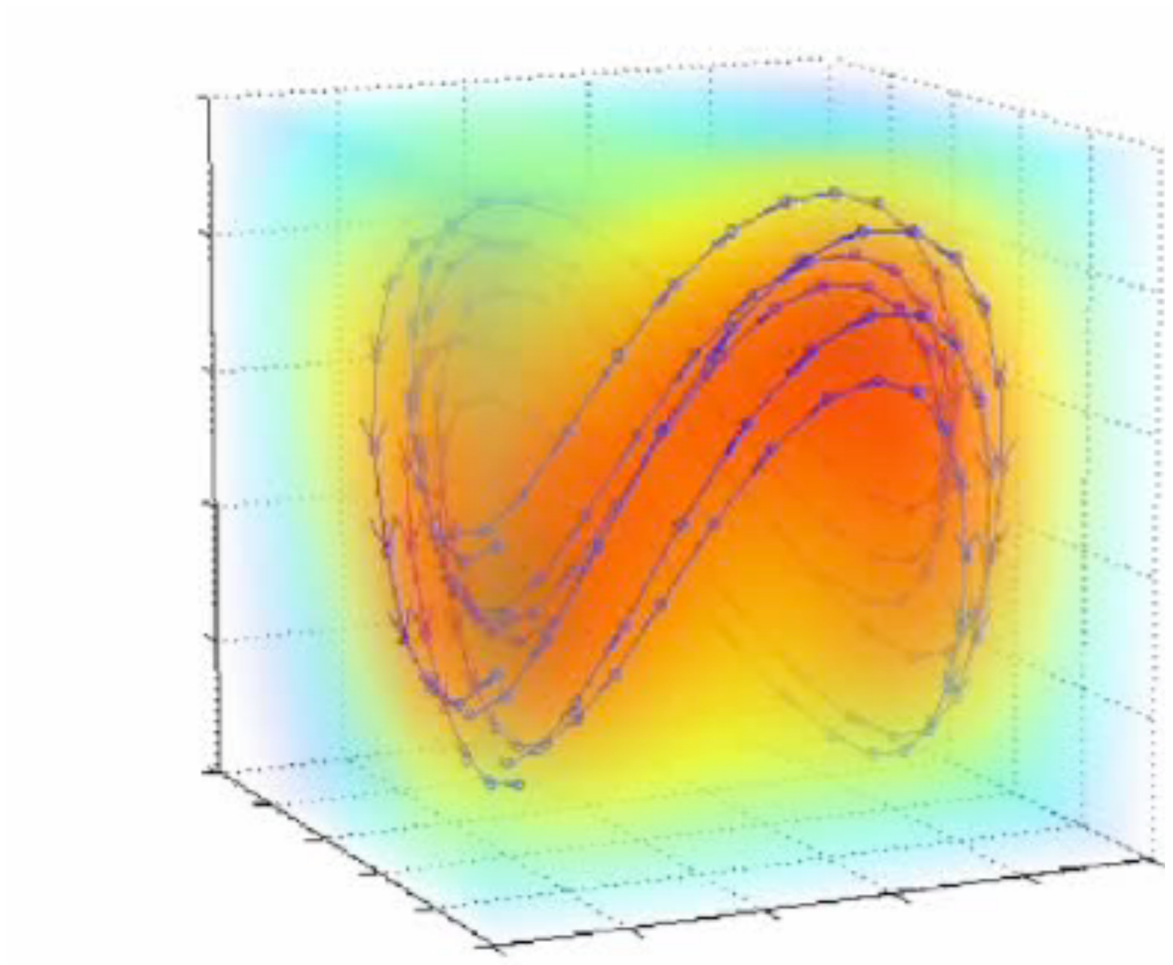
$$L(\mathbf{x}, \mathbf{y}; M) \;=\; \frac{\|\mathbf{W}(\mathbf{y} - f(\mathbf{x}))\|^2}{2\sigma_Y^2(\mathbf{x})} \;+\; \frac{D}{2} \ln \sigma_Y^2(\mathbf{x})$$

and a density over latent trajectories, with negative log likelihood:

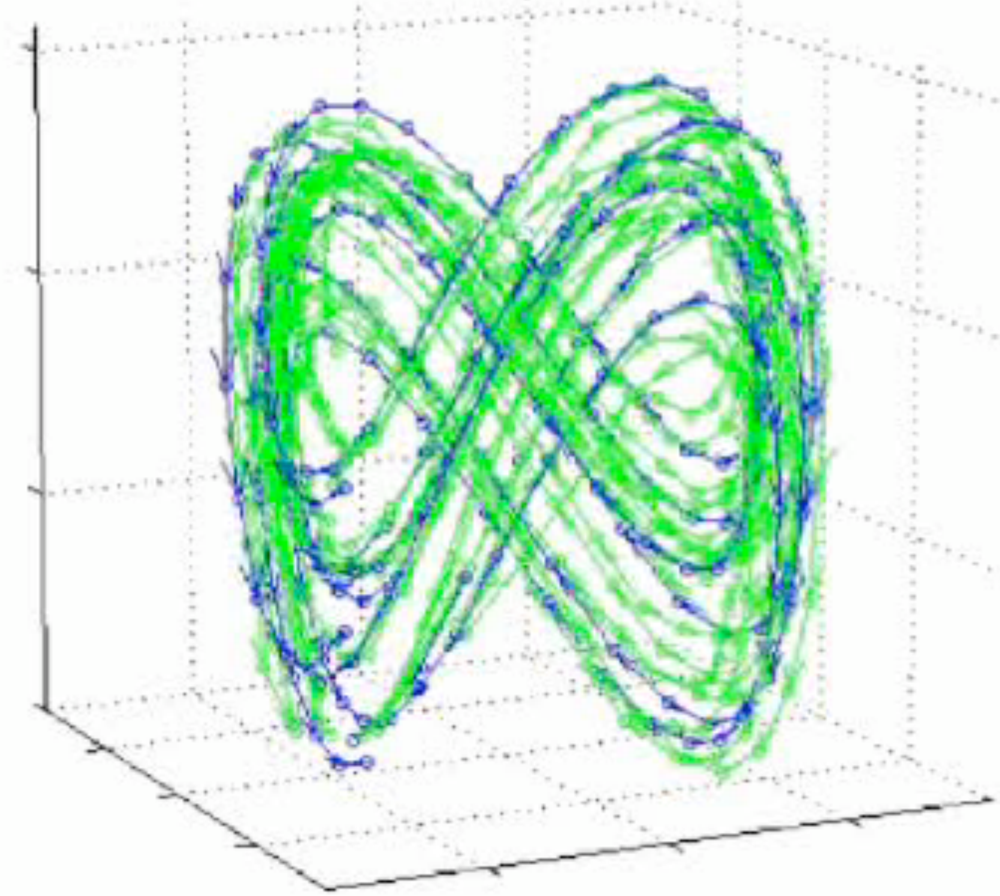
$$L_D(\bar{\mathbf{X}}; \bar{\mathbf{x}}_0, \mathcal{M}) \;=\; \frac{1}{2} tr\left(\bar{\mathbf{K}}_X^{-1} \bar{\mathbf{X}} \bar{\mathbf{X}}^T\right) \;+\; \frac{d}{2} \ln |\bar{\mathbf{K}}_X|$$

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 $\ln \sigma_{\mathbf{y}}^2 \,|\, \mathbf{x}, \mathbf{X}, \mathbf{Y}$

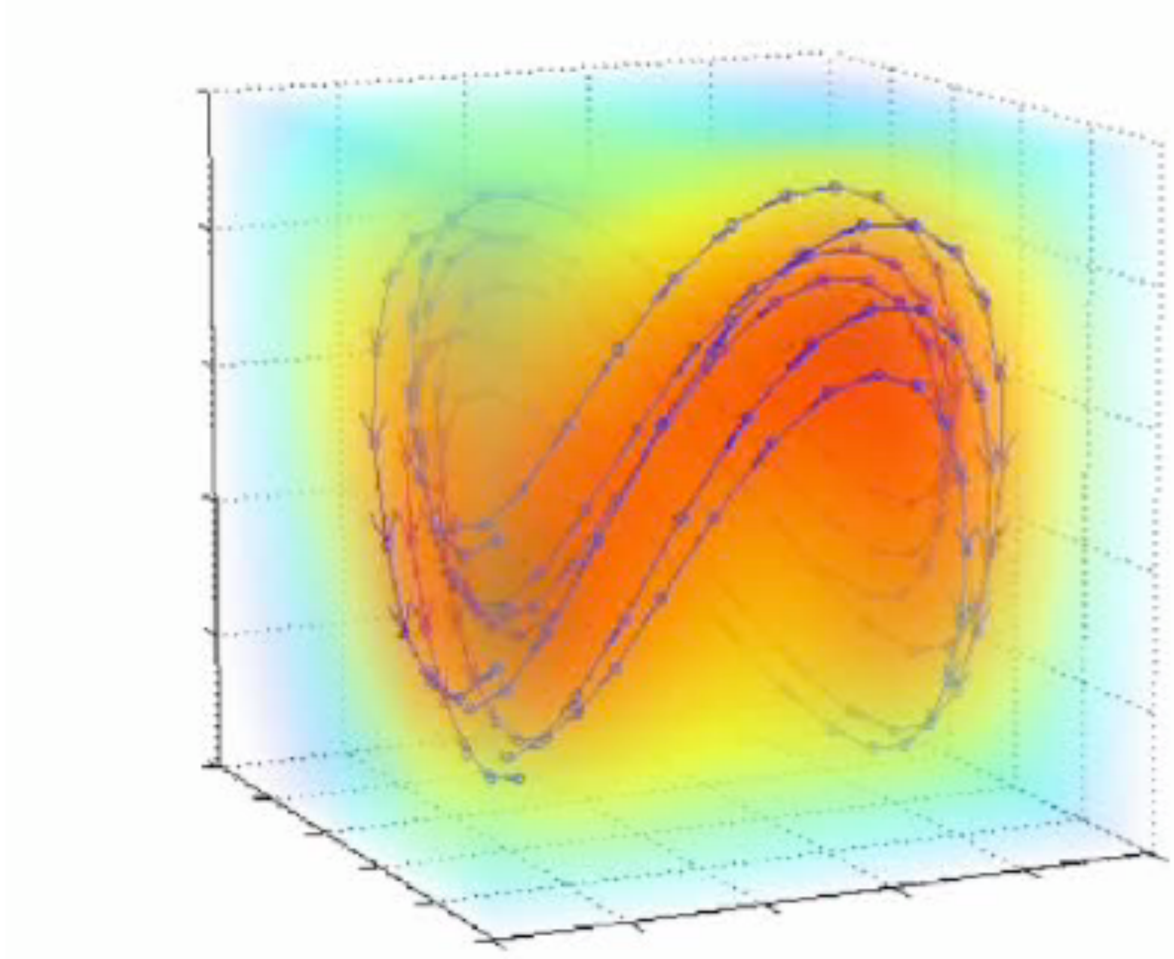


GPDM: sample trajectories

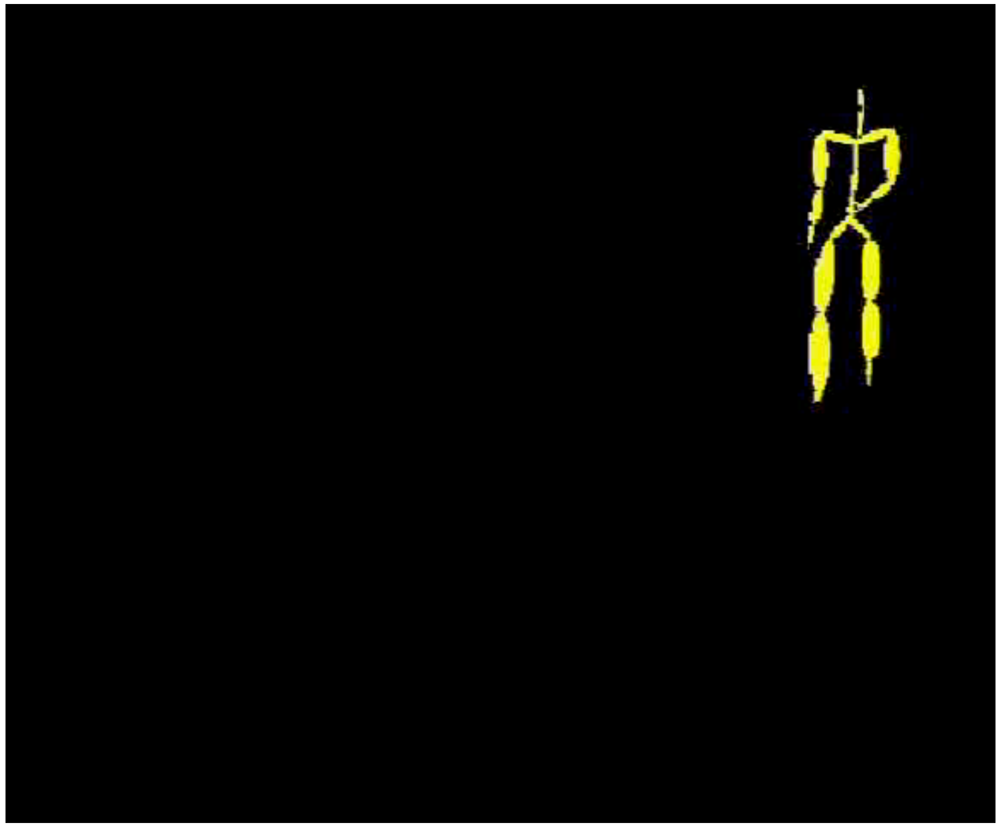
[Urtasun et al, `06]

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: mean tracjectory

[Urtasun et al, `06]

People tracking with GPDM

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

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 gradient ascent on the posterior:

$$\underbrace{p(\phi_t \,|\, \mathbf{I}_{1:t}, \mathcal{M})}_{\text{posterior}} \;\propto\; \underbrace{p(\mathbf{I}_t \,|\, \phi_t)}_{\text{likelihood}}\; \underbrace{p(\phi_t \,|\, \mathbf{I}_{1:t-1}, \mathcal{M})}_{\text{prediction}}$$

Temporal predictions for the global DOFs based on a damped second-order Markov model.

[Urtasun et al, `06]

Measurement model

Measurements are the 2D image positions for several locations on the body, obtained with a 2D patch-based tracker *[Jepson et al 03]*.

Assume the measurements are corrupted with IID Gaussian noise.

Tracking experiments

Input videos:

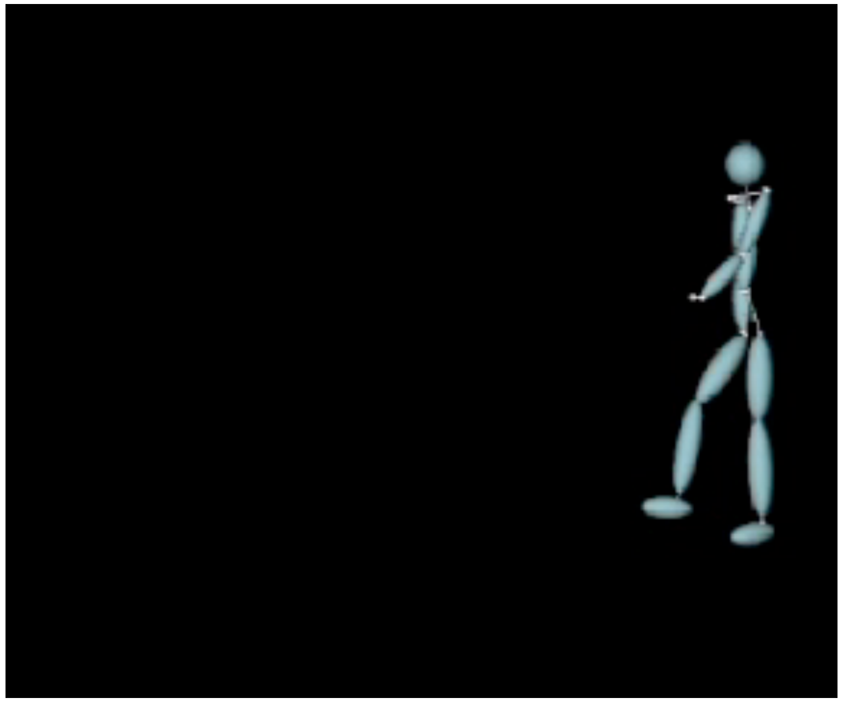
- noisy measurements
- occlusion (measurement loss)
- speed change (1 octave)
- stylistic variation

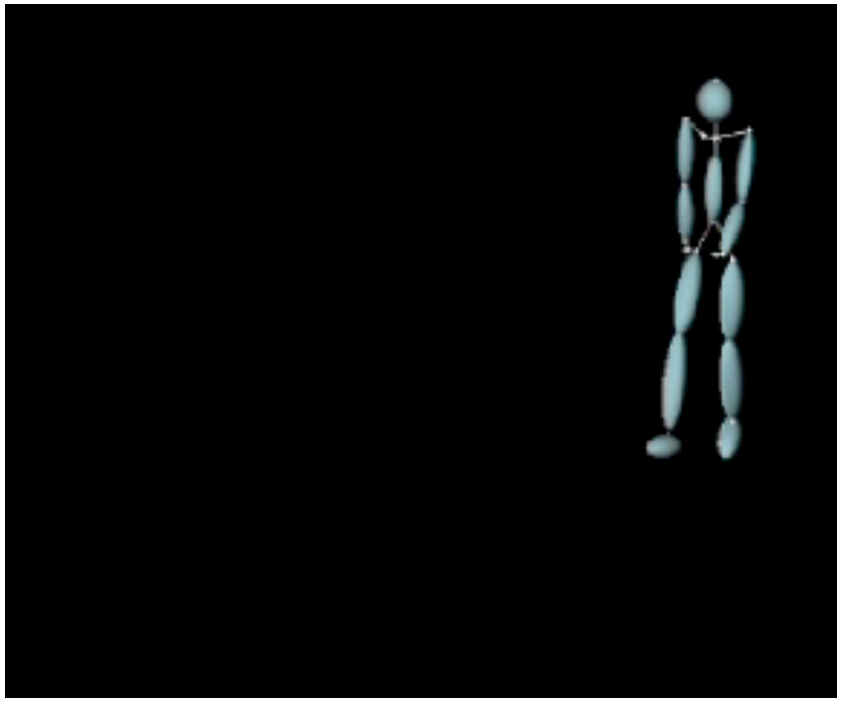
Initialization:

- 2D WSL points and 3D model are initialized manually in the first frame

Occlusion

3D
model
overlaid
on video

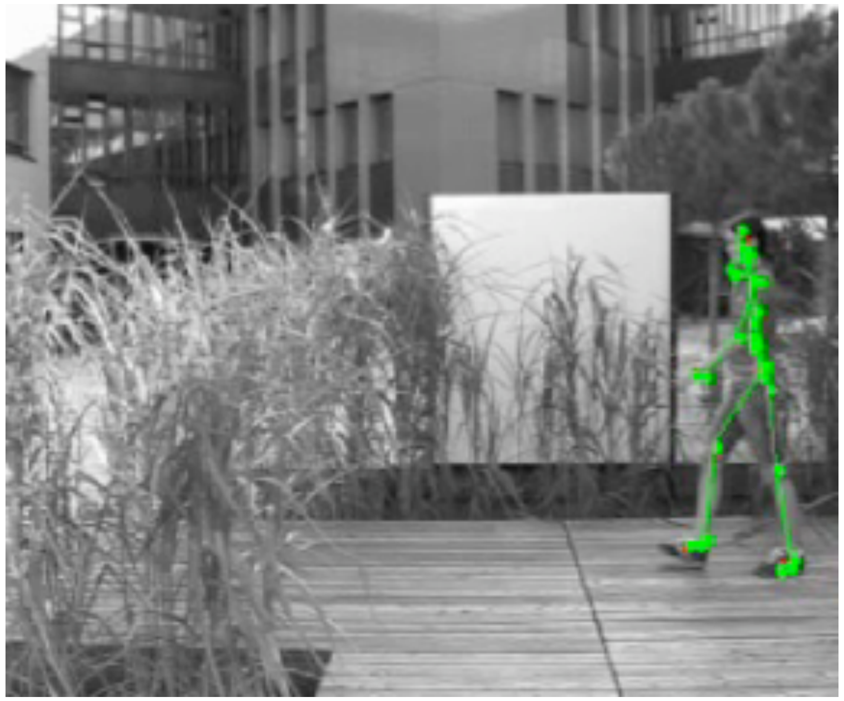


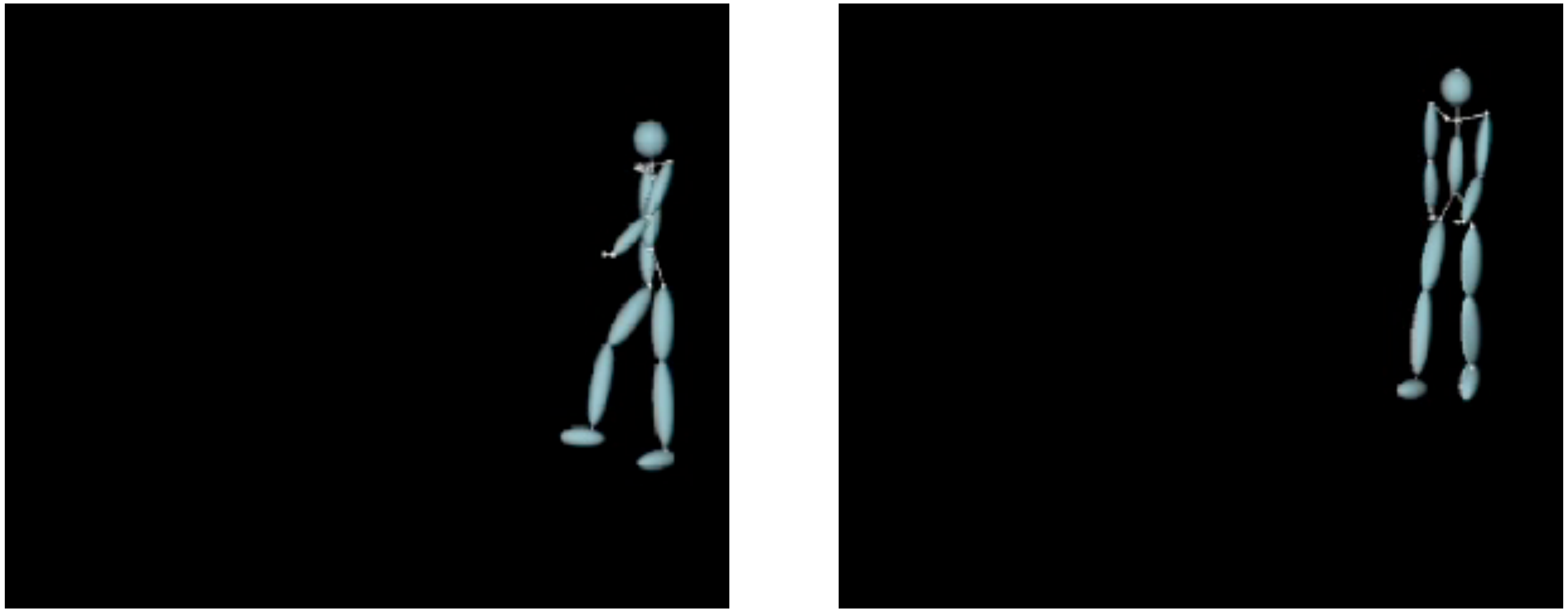


3D animated characters

Occlusion

3D model overlaid on video

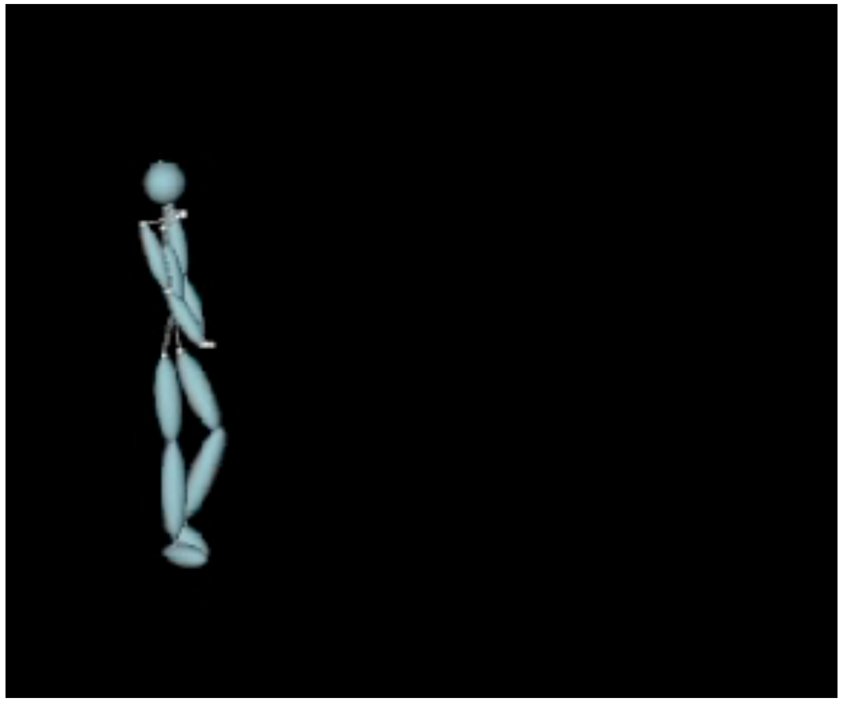


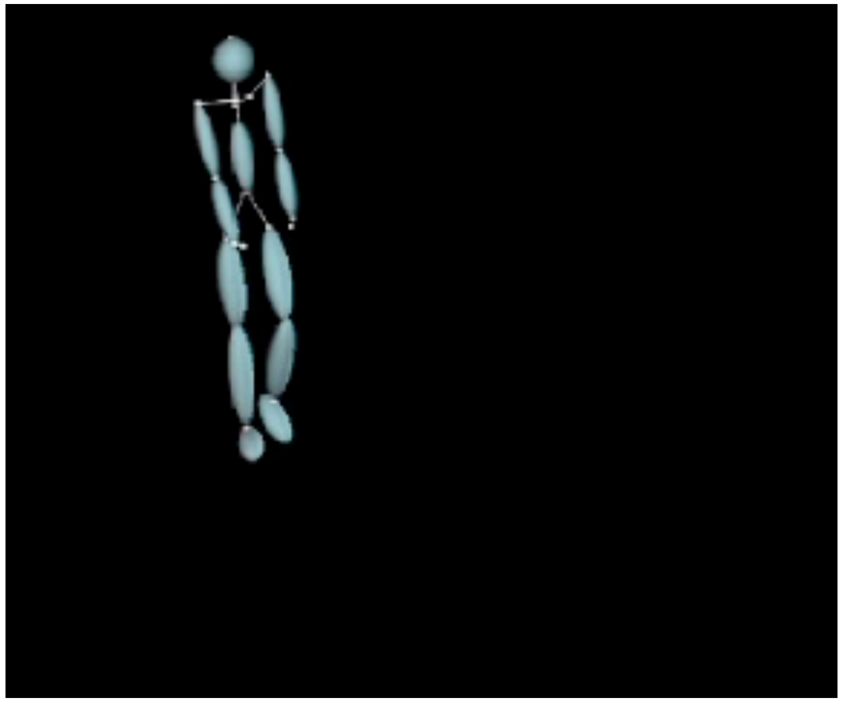


3D animated characters

Exaggerated gait

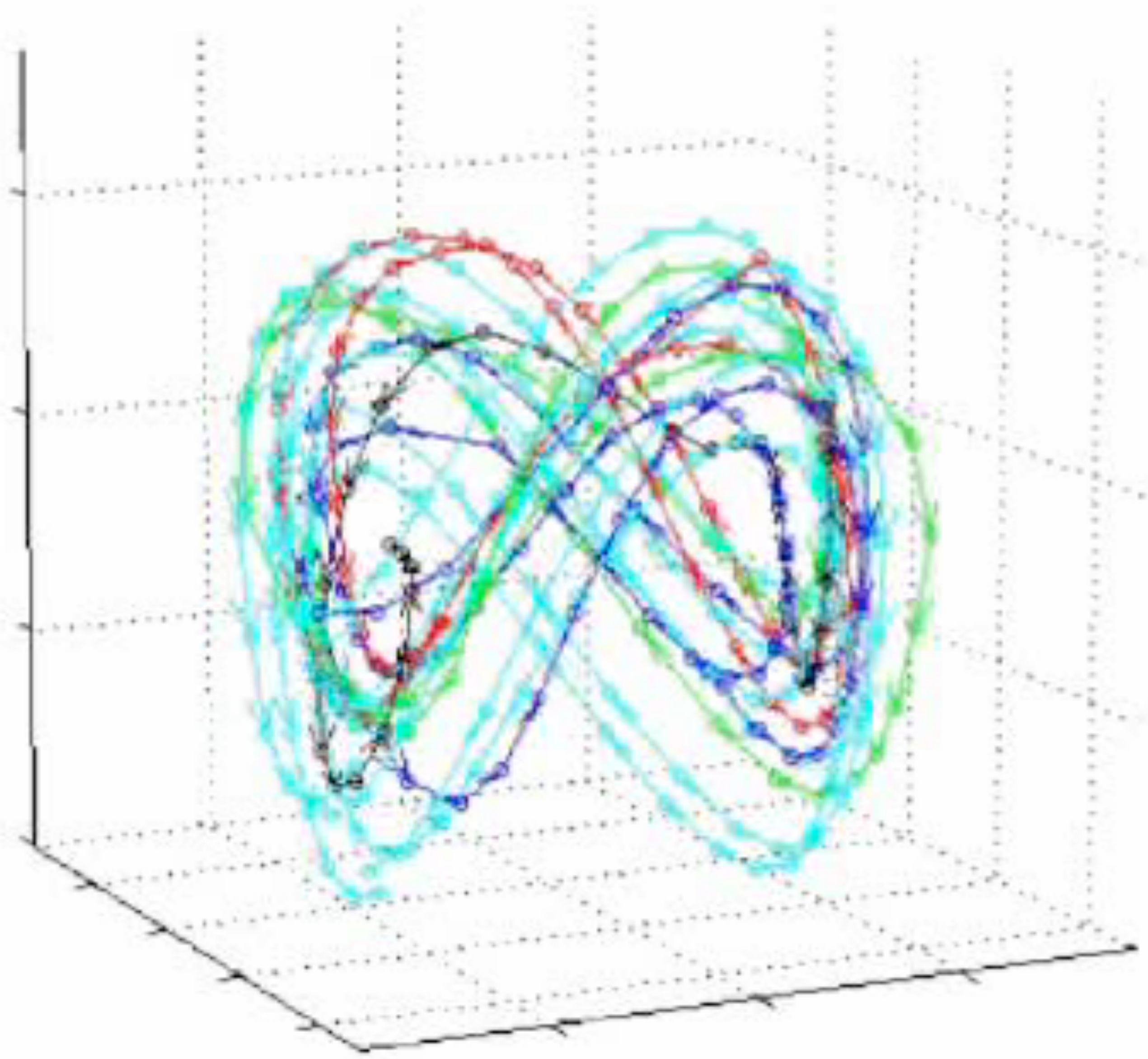
3D
model
overlaid
on video





3D animated characters

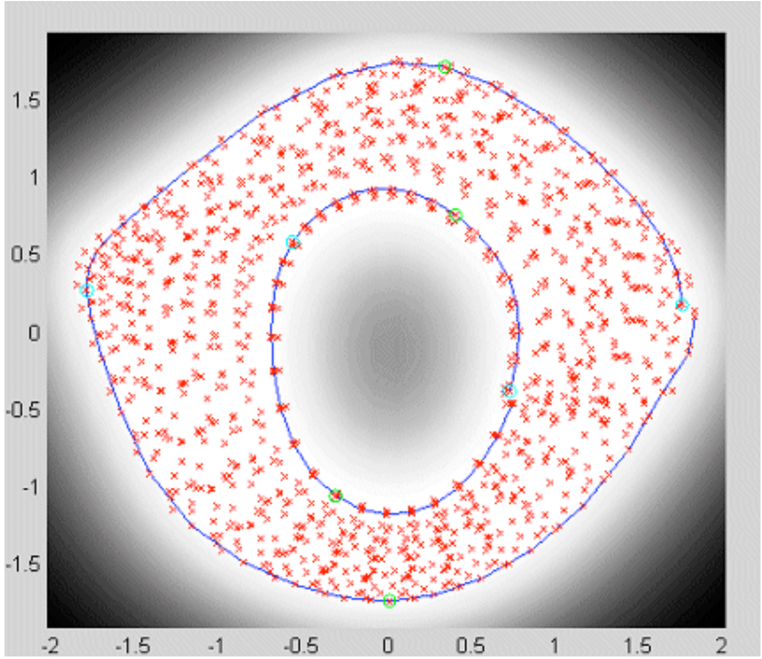
Latent trajectories

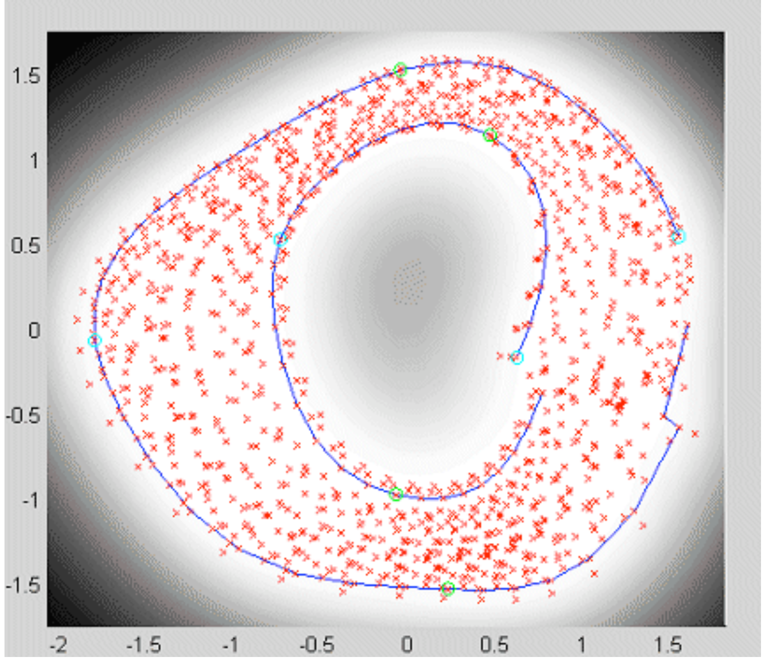


Hedvig
Shrub
Occlusion
Exaggerated
Training Data

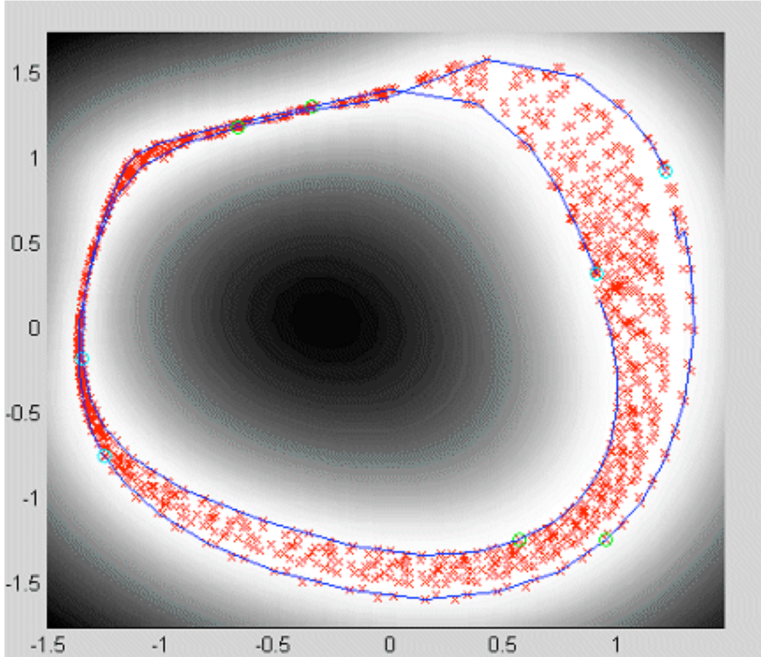
Multiple speeds and visualization of pathologies

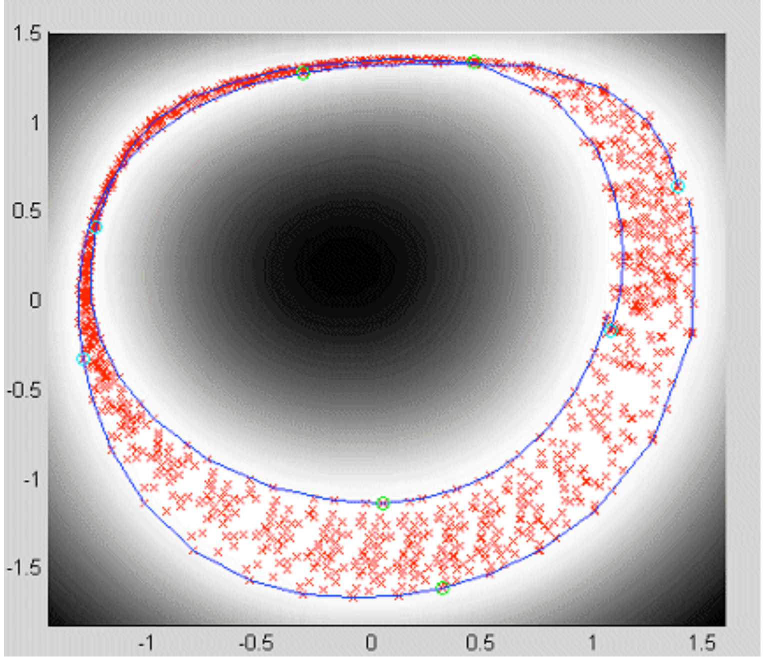
Two subjects, four walk gait cycles at each of 9 speeds (3-7 km/hr)





Two subjects with a knee pathology.





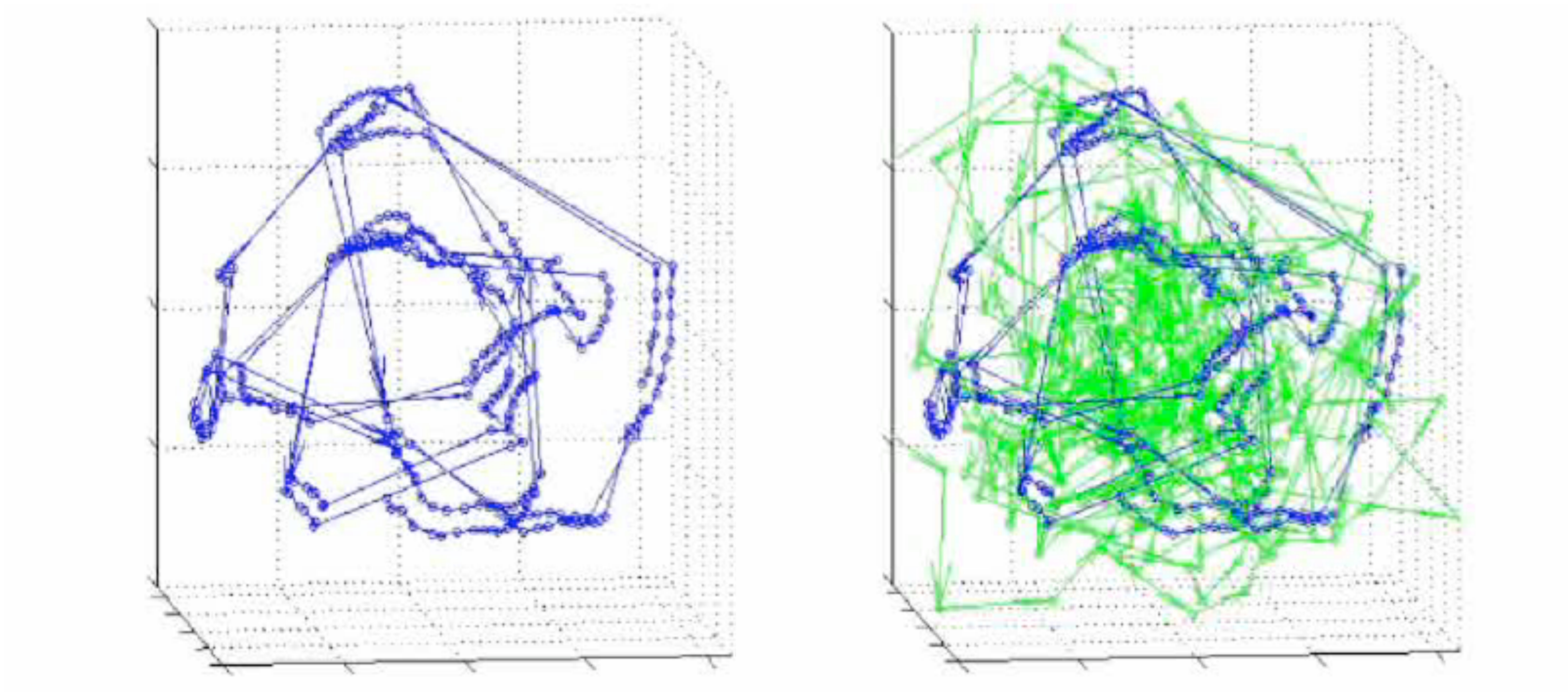
But

GPLVM has its limits …

- models don’t scale
- they don’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

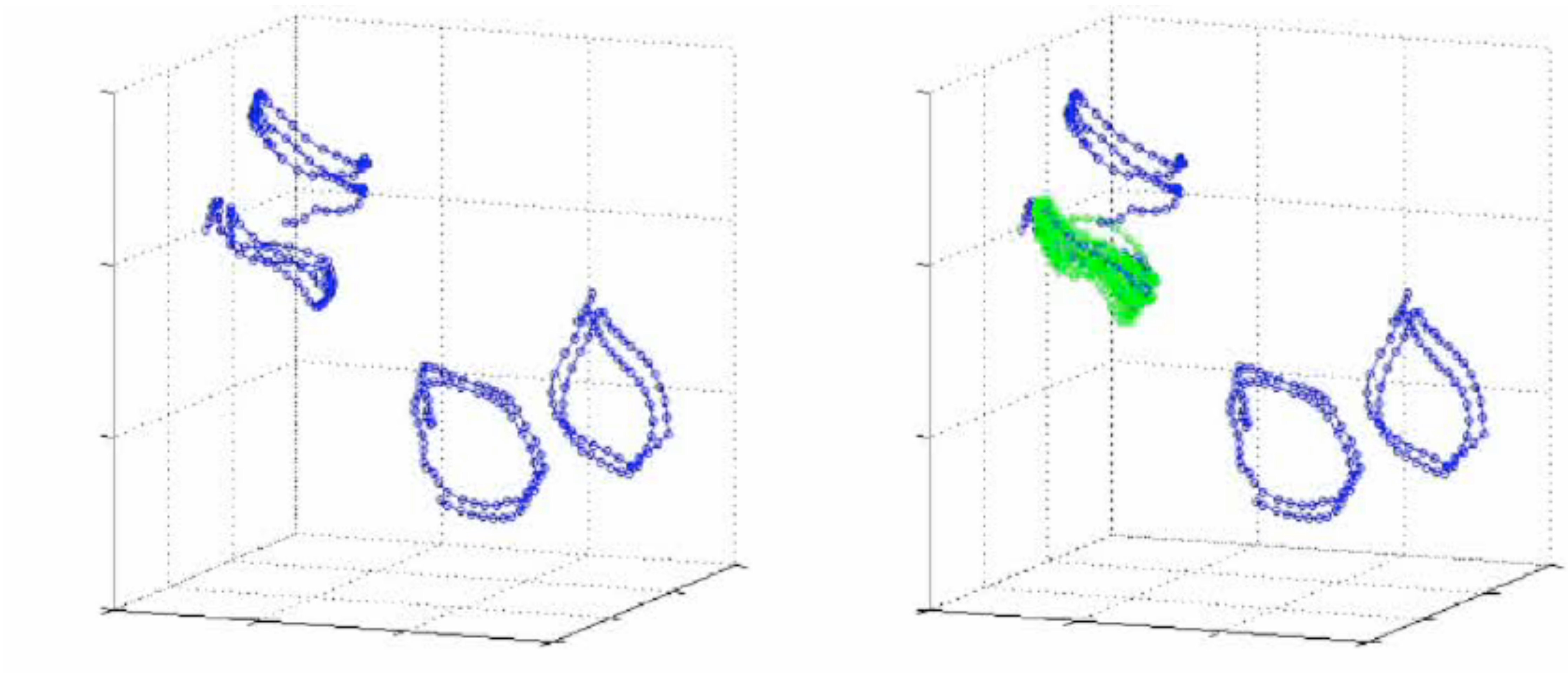
4 walking subjects, 2 gait cycles each, 50 DOFs



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]

Problems with multiple motions / styles

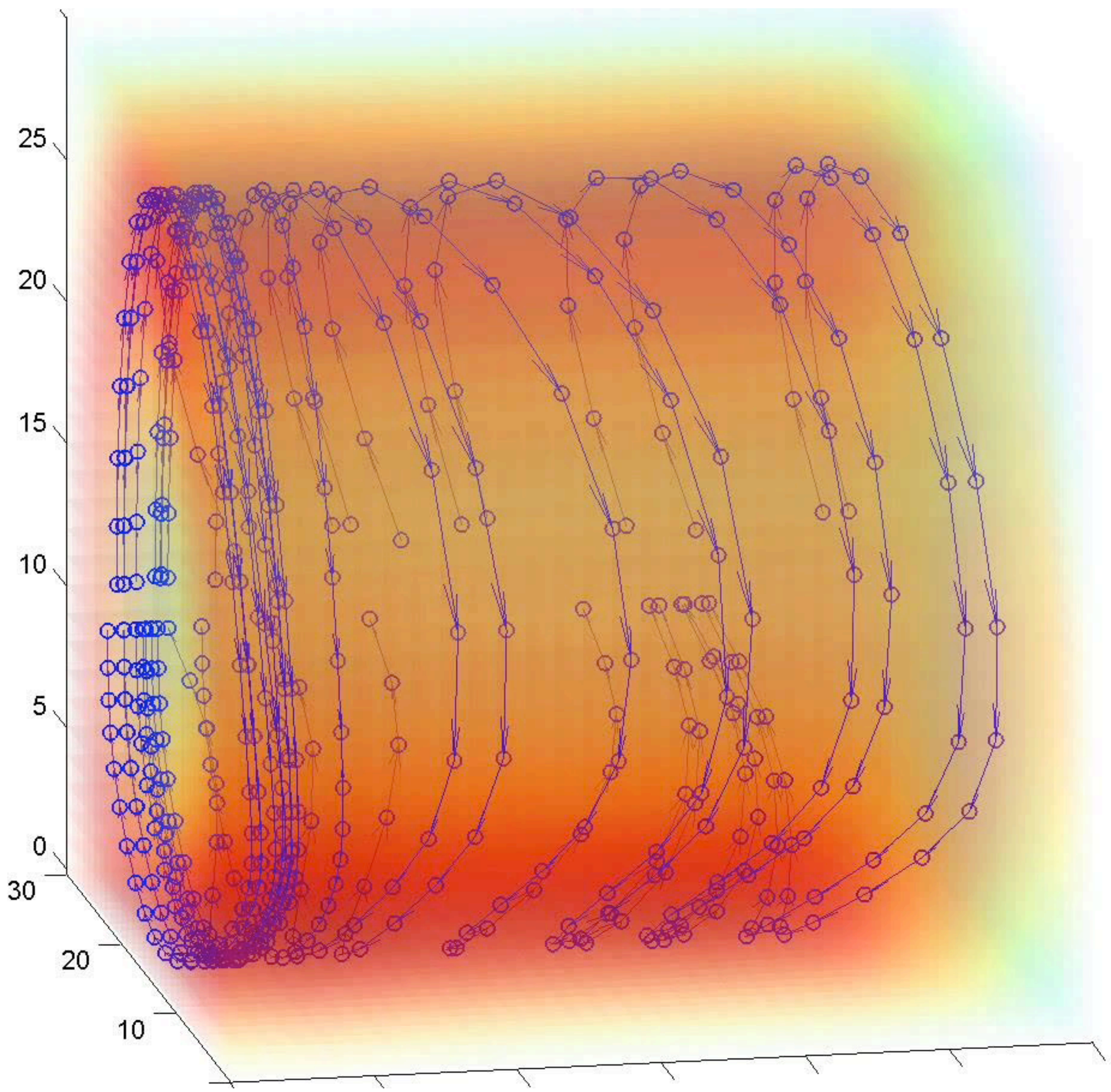
GPLVMs do not ensure that the map from the pose space $\mathbf{y}$ to the latent space $\mathbf{x}$ is smooth, i.e., that nearby poses map to nearby latent positions.

With sparse mocap data, it is often hard to generalize well from the motions of a few individuals with different styles.

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

Topologically-constrained GPLVM

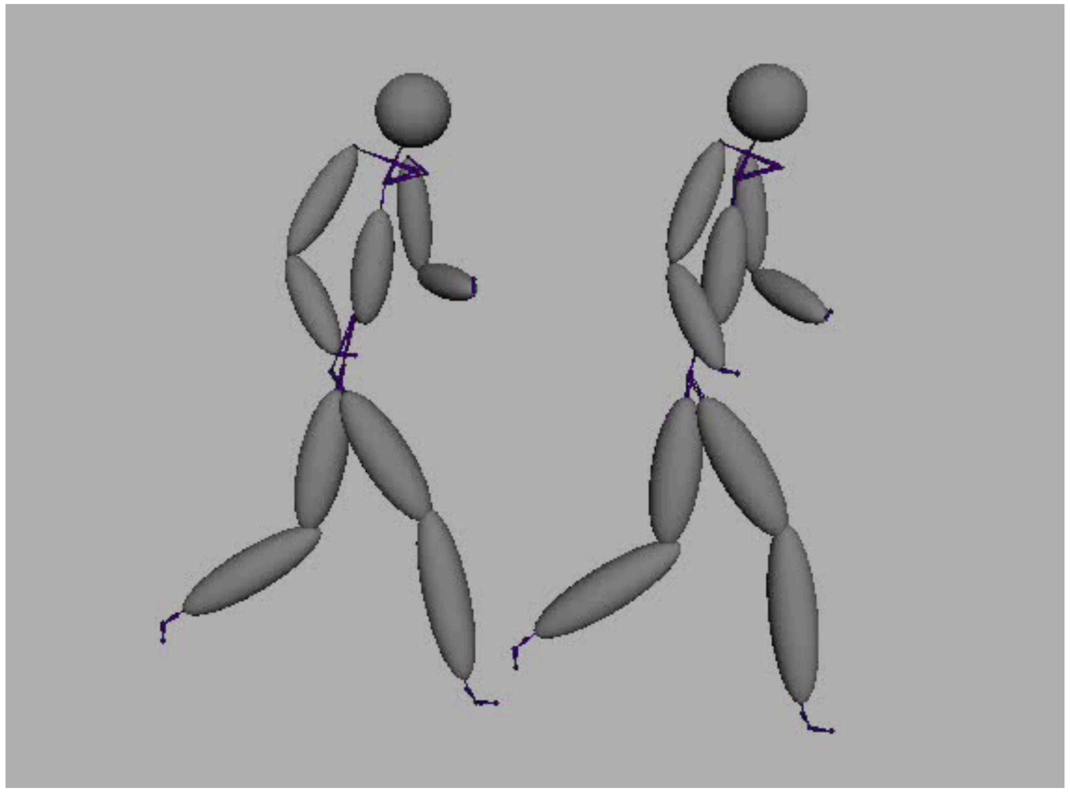
Global constraints on latent space topology (e.g., for periodic motions), and local topological constraints to preserve pose neighborhoods.

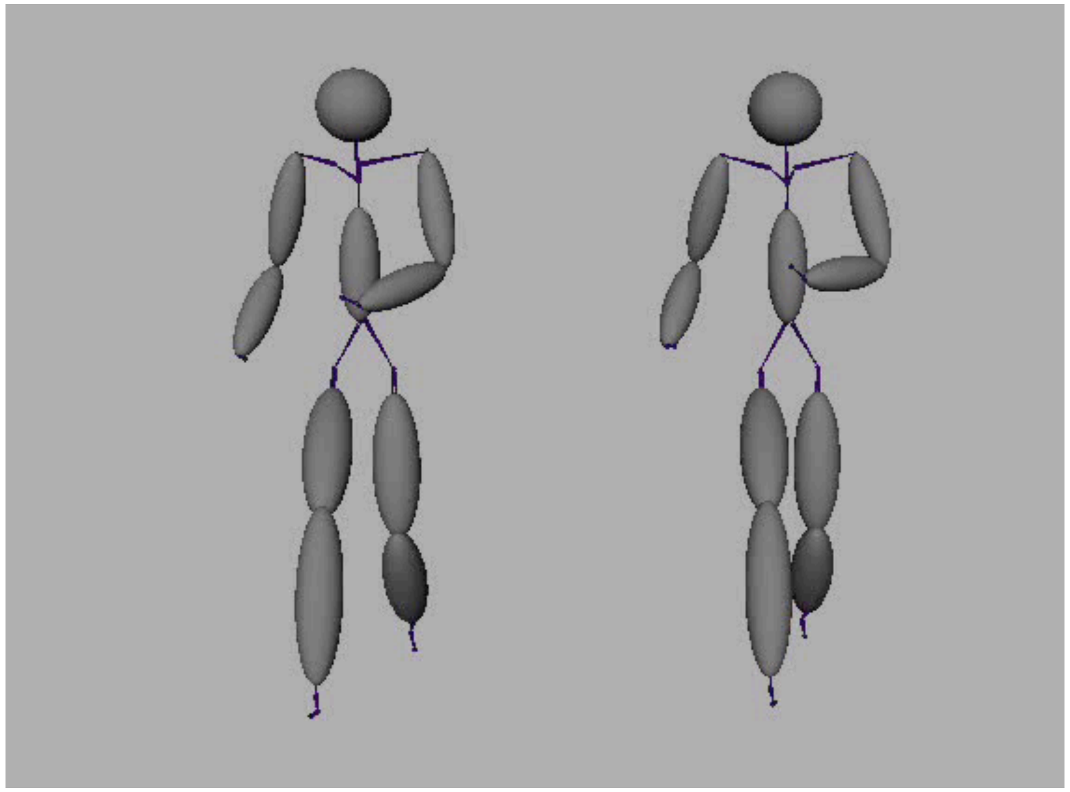


9 walk cycles and
10 jog cycles, with
different speeds
and subjects

[Urtasun et al. ICML '08]

Topologically-constrained GPLVM





Simulation with transitions.

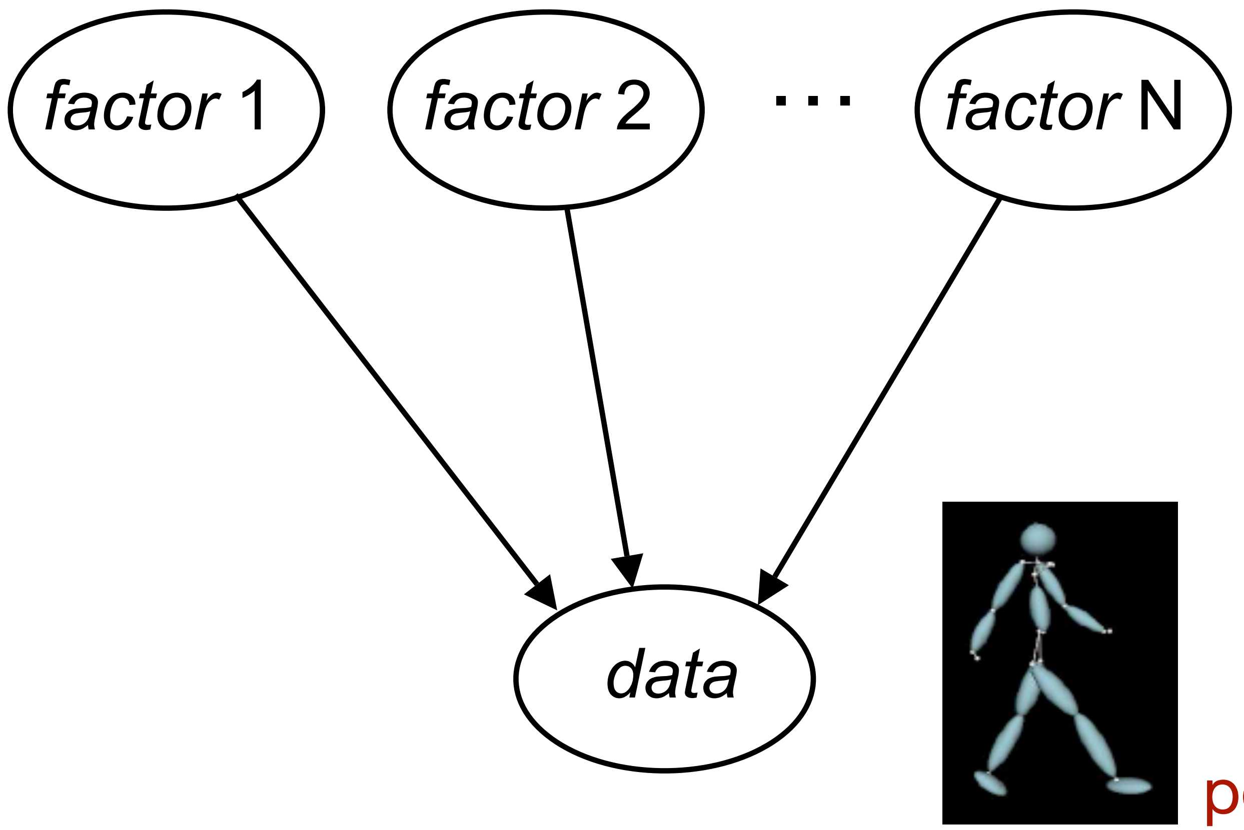
[Urtasun et al. ICML '08]

Style-content separation

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

[Wang et al. ICML '07]

Style-content separation



gait,
phase,
identity,
gender,
...

pose

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

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

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

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 $\mathbf{x}$:

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

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

If $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \beta^{-1})$, 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 $\mathbf{s} = [s_1, ..., s_{N_s}]^T$

$k_s(\mathbf{s}, \mathbf{s}')$ $k_x(\mathbf{x}, \mathbf{x}')$

[Wang et al. ICML '07]

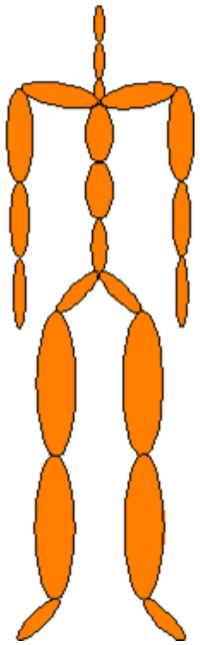
Multifactor locomotion model

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:

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

scale of variance
dimension

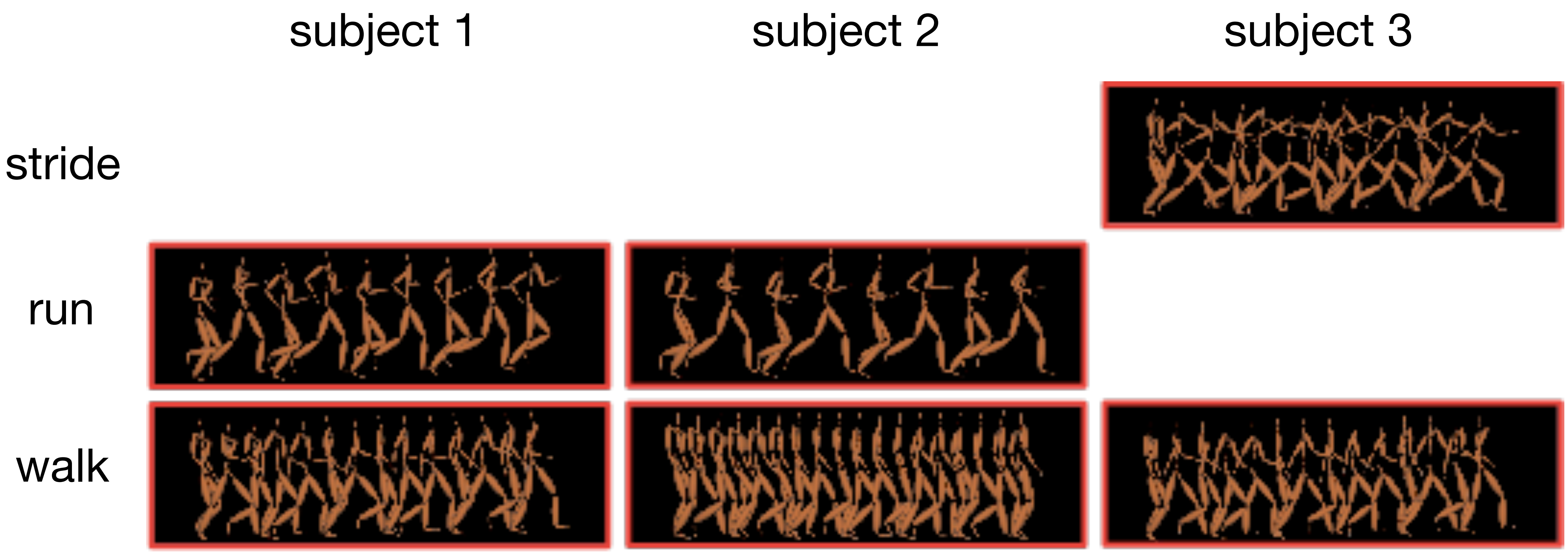
linear kernels for identity
and gait (style)

RBF kernel for state
(content)

additive white process
noise

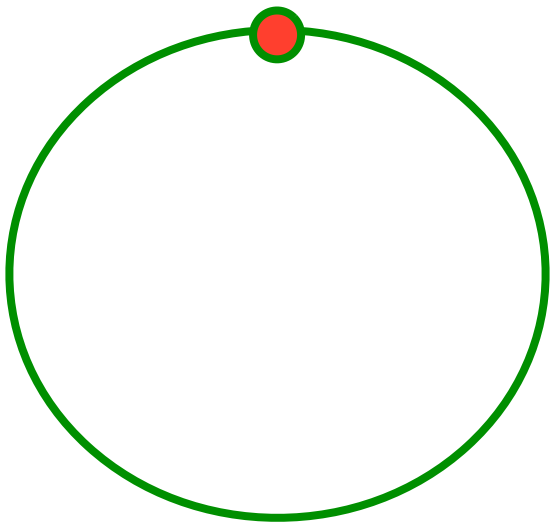
[Wang et al. ICML '07]

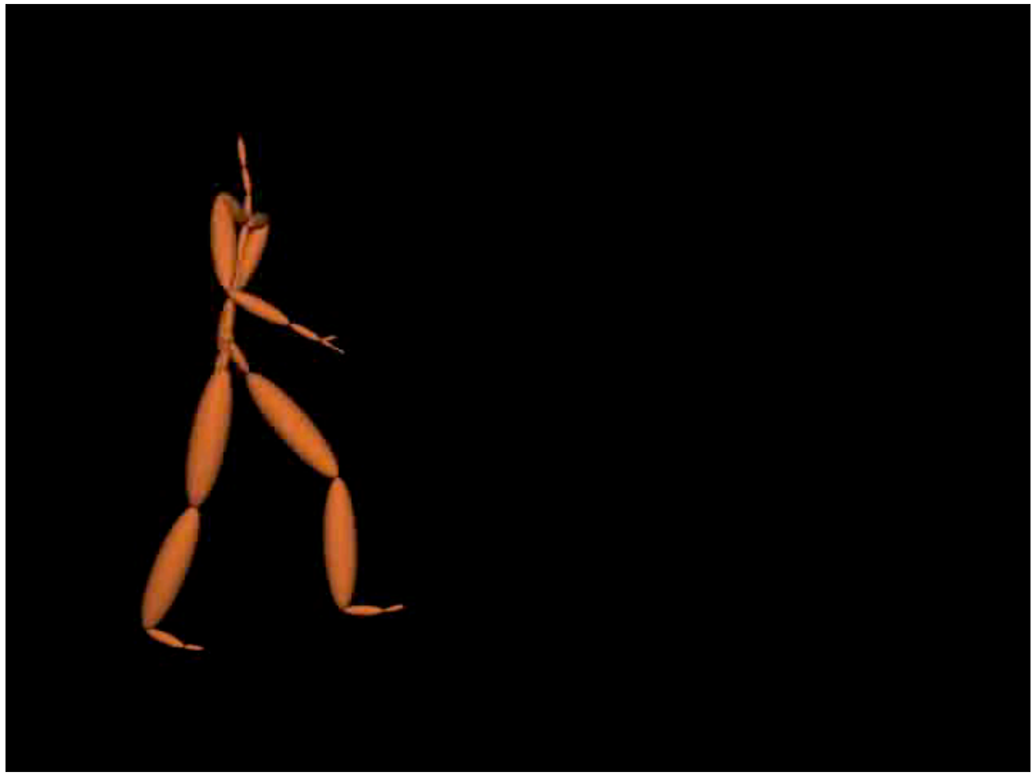
Training data



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

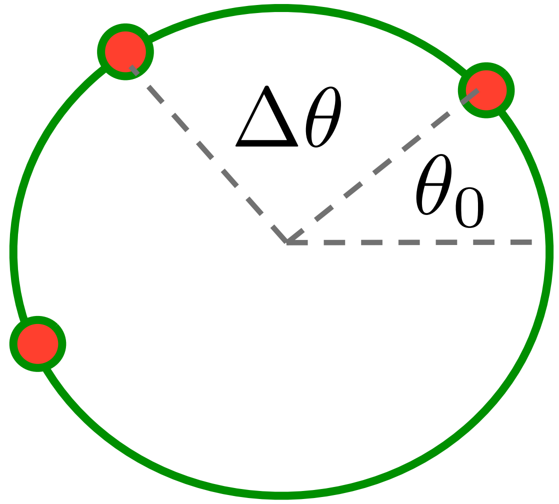
A locomotion model





The state of the motion ($\mathbf{x}$) is assumed to lie on the unit circle, which is shared by all motions.

A locomotion model



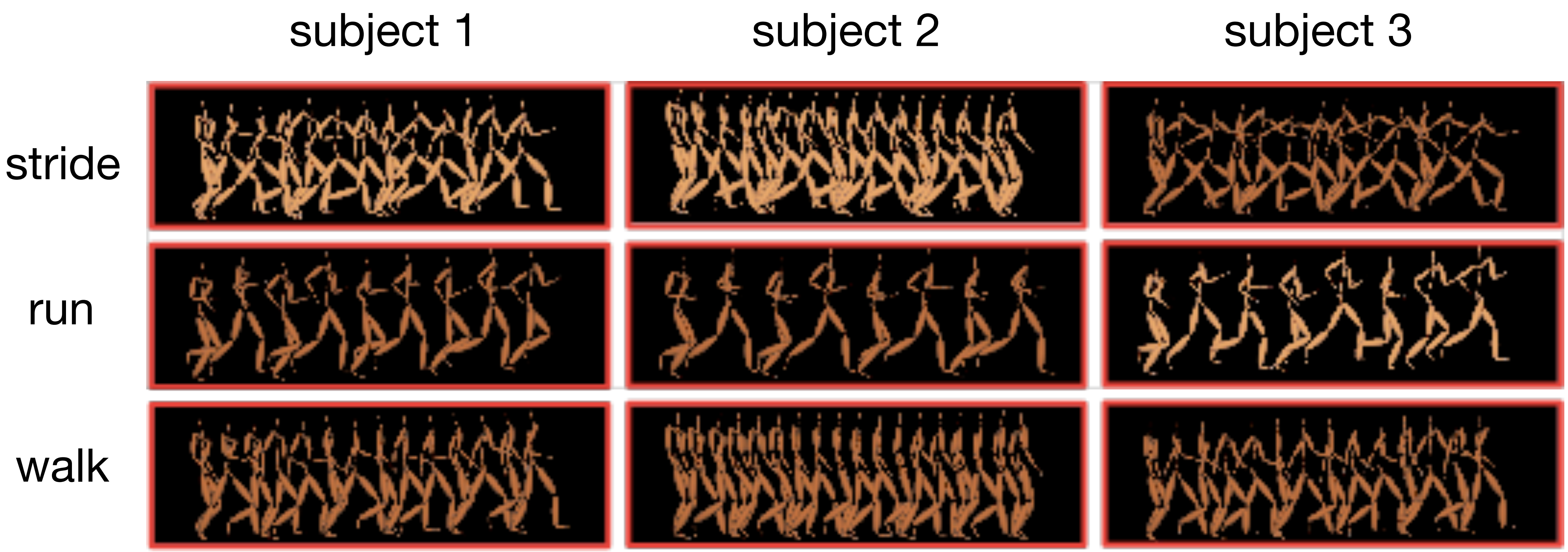
We assume no knowledge of correspondence between poses (i.e., no “time-warping”).

Each sequence is parameterized by θ_0 and $\Delta\theta$, which are learned.

$$\theta_t = \theta_0 + t\,\Delta\theta$$

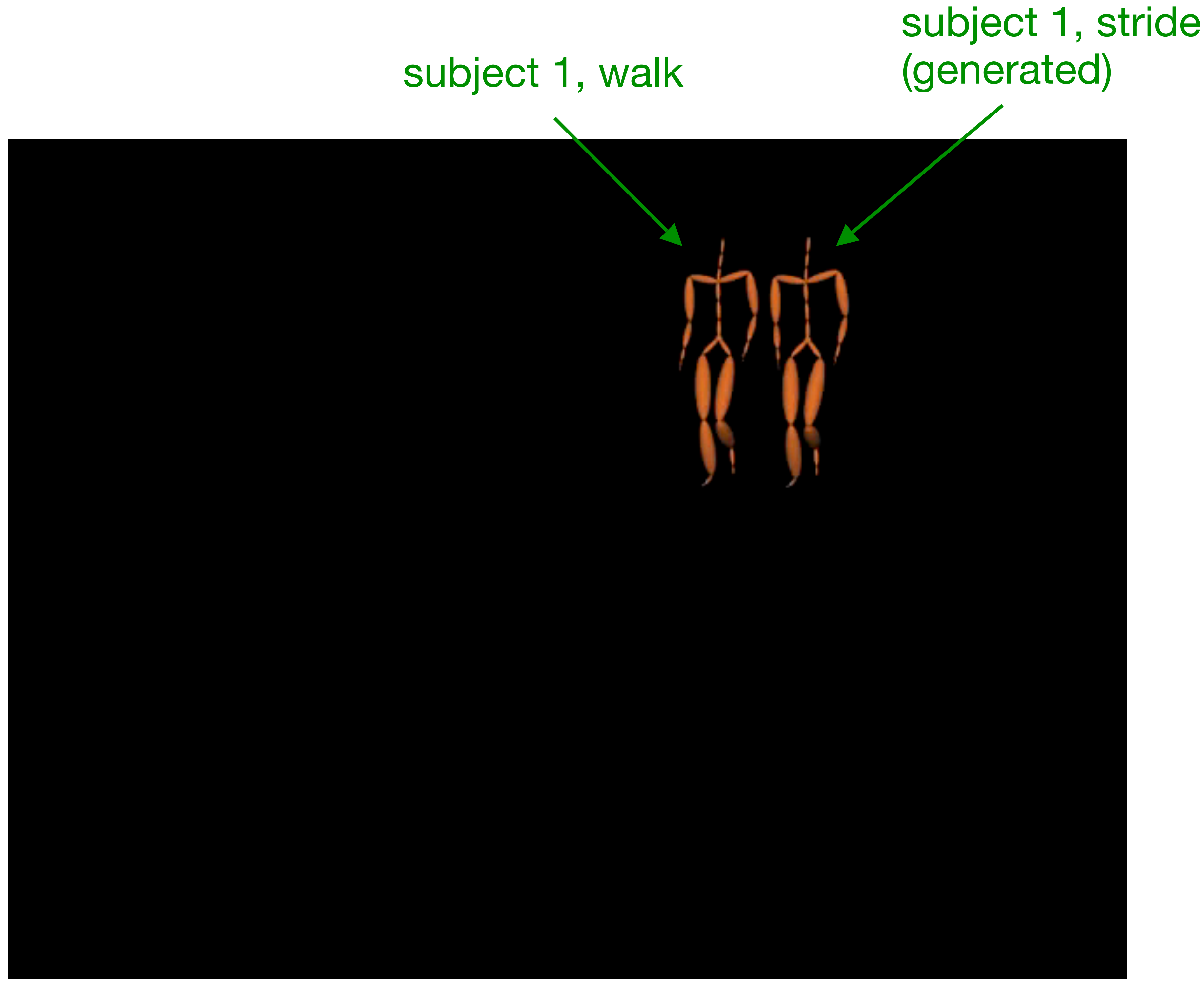
$$\mathbf{x}^t = [\cos\theta_t, \sin\theta_t]^T$$

Generating new motions



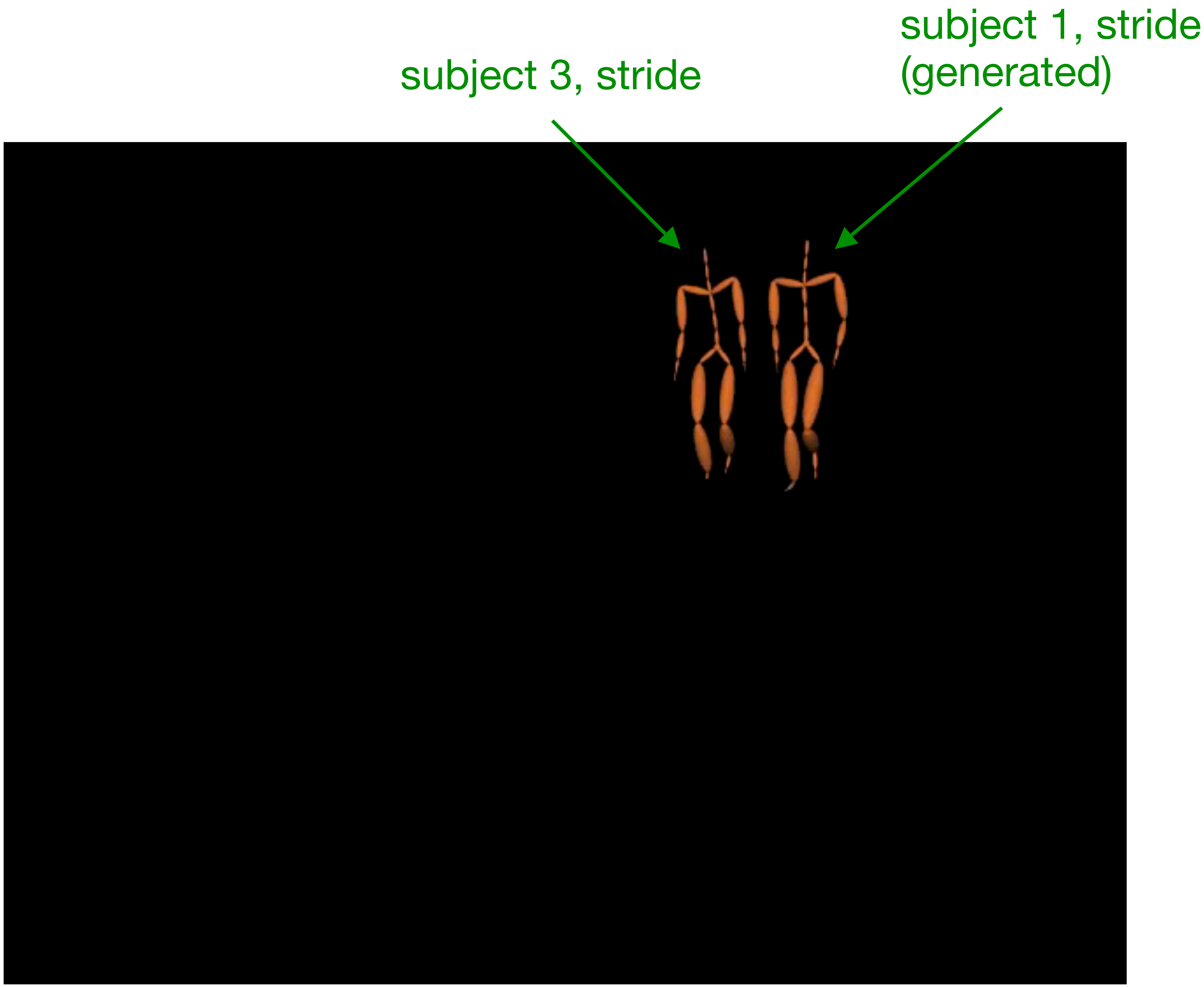
The GP model provides a Gaussian prediction for new motions.
We use the mean to generate motions with different styles.

Generating new motions



[Wang et al. ICML ’07]

Generating new motions

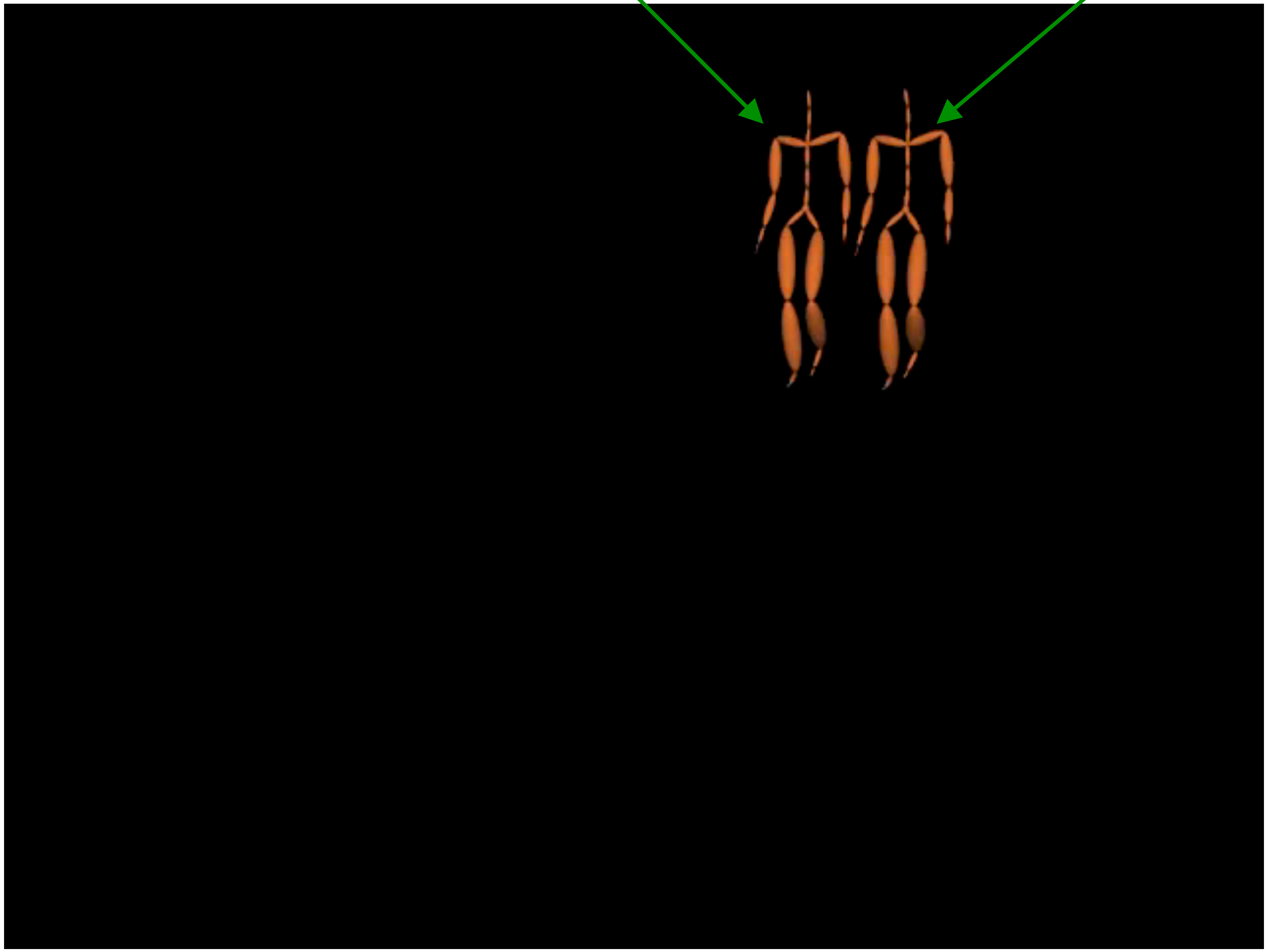


[Wang et al. ICML '07]

Generating new motions

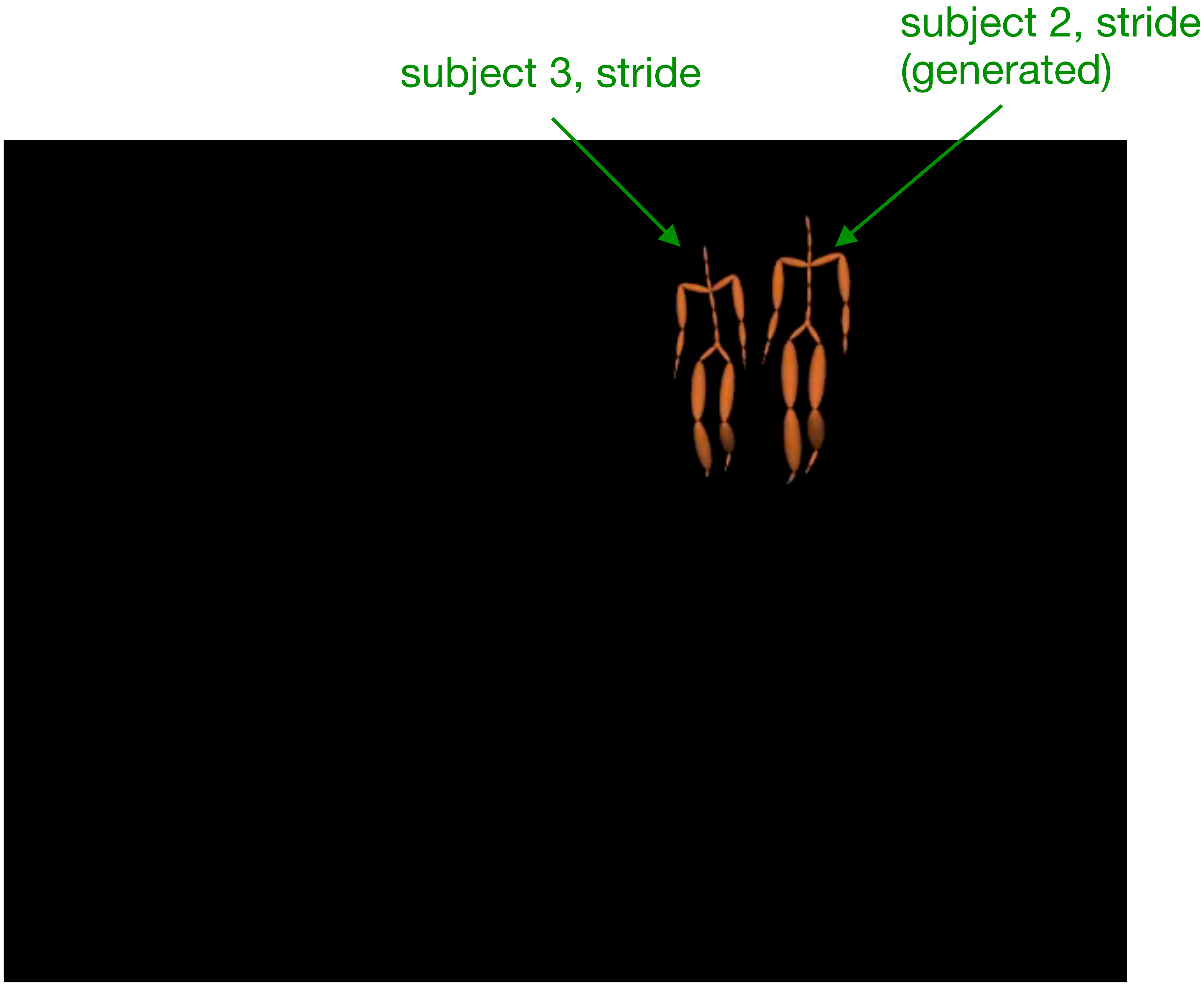
subject 2, walk

subject 2, stride (generated)



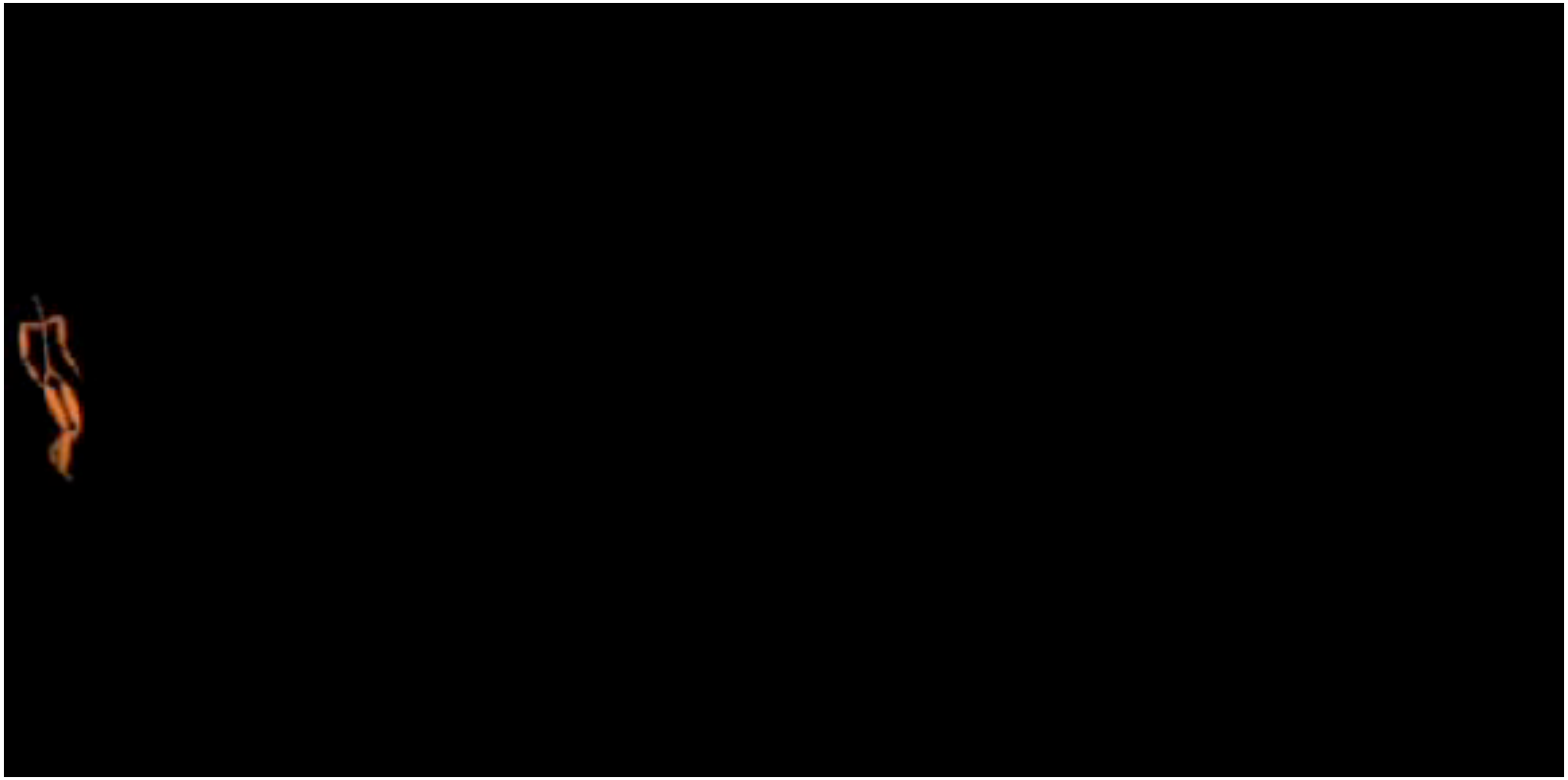
[Wang et al. ICML '07]

Generating new motions



[Wang et al. ICML '07]

Generating new motions



Transitions

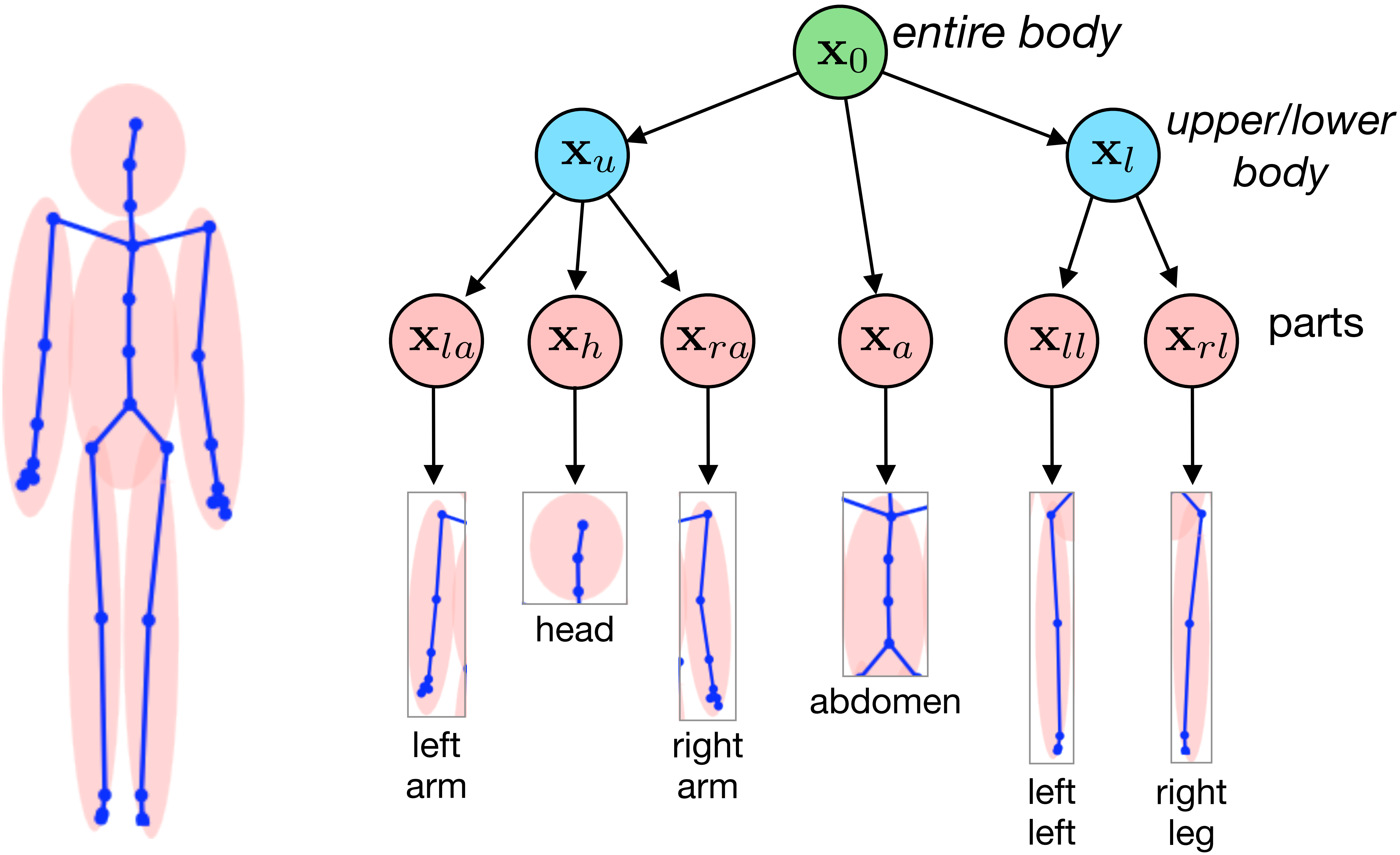
[Wang et al. ICML ’07]

Generating new motions

Random motions

[Wang et al. ICML '07]

Hierarchical GPLVM



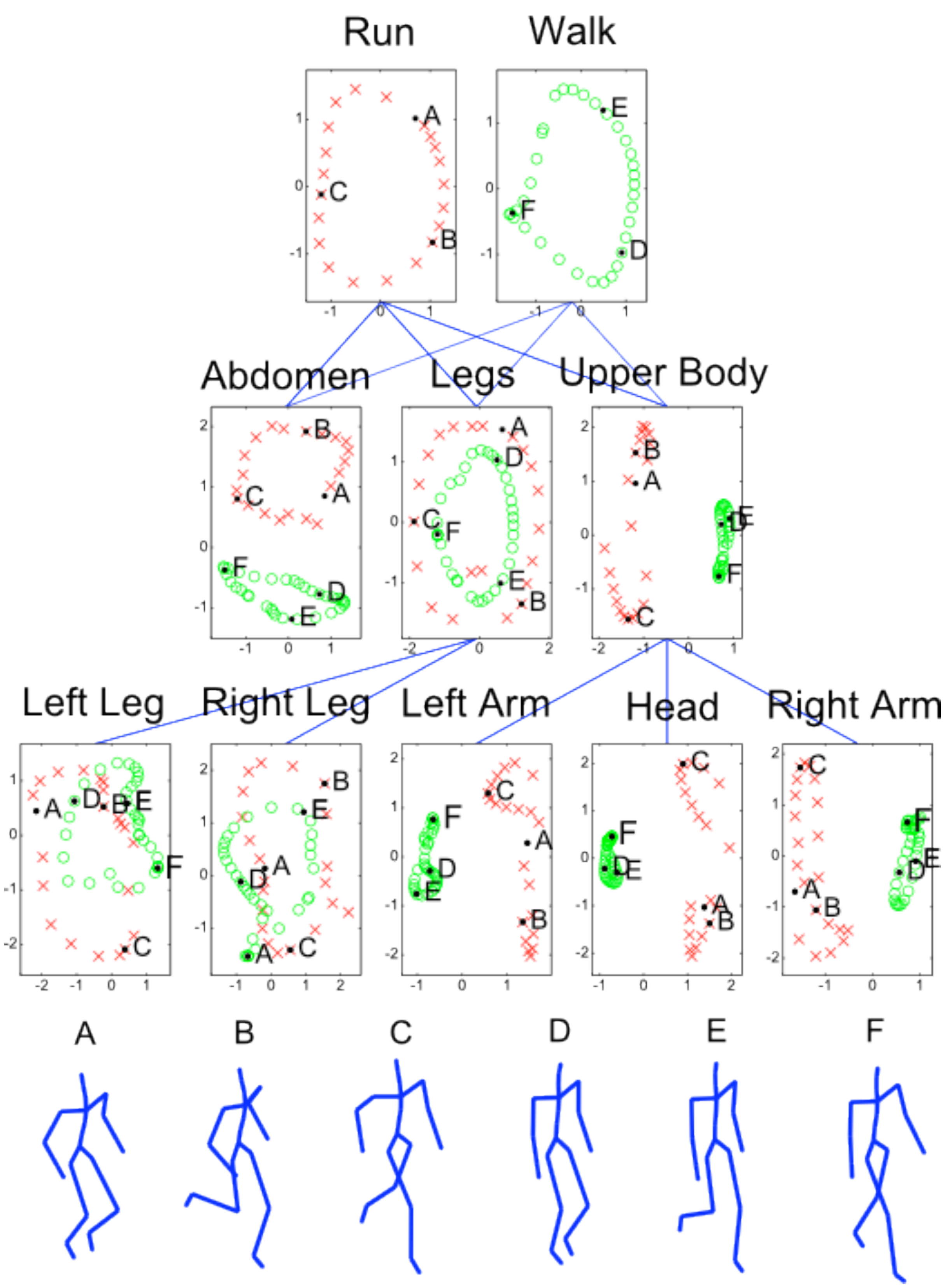
Hierarchical GPLVM *[Lawrence and Moore ICML '07]*

Hierarchical GPLVM

Data: 1 walk cycle, 1 run cycle

Initialization: PCA

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



Hierarchical GPLVM

[Darby et al., BMVC '09]

Shared latent variable models

Done on whiteboard *[Sigal et al, CVPR '09]*

Selected references

Lawrence N, Probabilistic nonlinear principal components analysis with Gaussian Process latent variable models. *JMLR* 6, 2005 (also see NIPS 2004)

Moore A and Lawrence N, Hierarchical Gaussian process latent variable models. *Proc ICML*, 2007

Navaratnam et al., The joint manifold model for semi-supervised multi-valued regression. *Proc ICCV,* 2007

Quinonero-Candela & Rasmussen, A unifying view of sparse approximate Gaussian Process regression. *JMLR* 6, 2006

Sigal L et al., Shared kernel information embedding for discriminative inference. *Proc IEEE CVPR*, 2009

Urtasun R and Darrell T, Local Probabilistic regression for activity-independent human pose inference, *Proc CVPR 2008*

Urtasun R et al., People tracking with the Gaussian process dynamical model. *Proc IEEE CVPR,* 2006

Urtasun R et al., Topologically constrained latent variable models. *Proc ICML* 2008

Wang J et al ., Multifactor Gaussian process models for style-content separation. *Proc ICML,* 2007.

Wang J et al, Gaussian Process dynamical models for human motion. *IEEE Trans PAMI* 30(2), 2008 (also see NIPS 2005)