Priors for People Tracking from Small Training Sets

Raquel Urtasun[1] David Fleet[2] Aaron Hertzmann[2] Pascal Fua[1]

[1] EPFL, Switzerland [2] University of Toronto, Canada

Problem

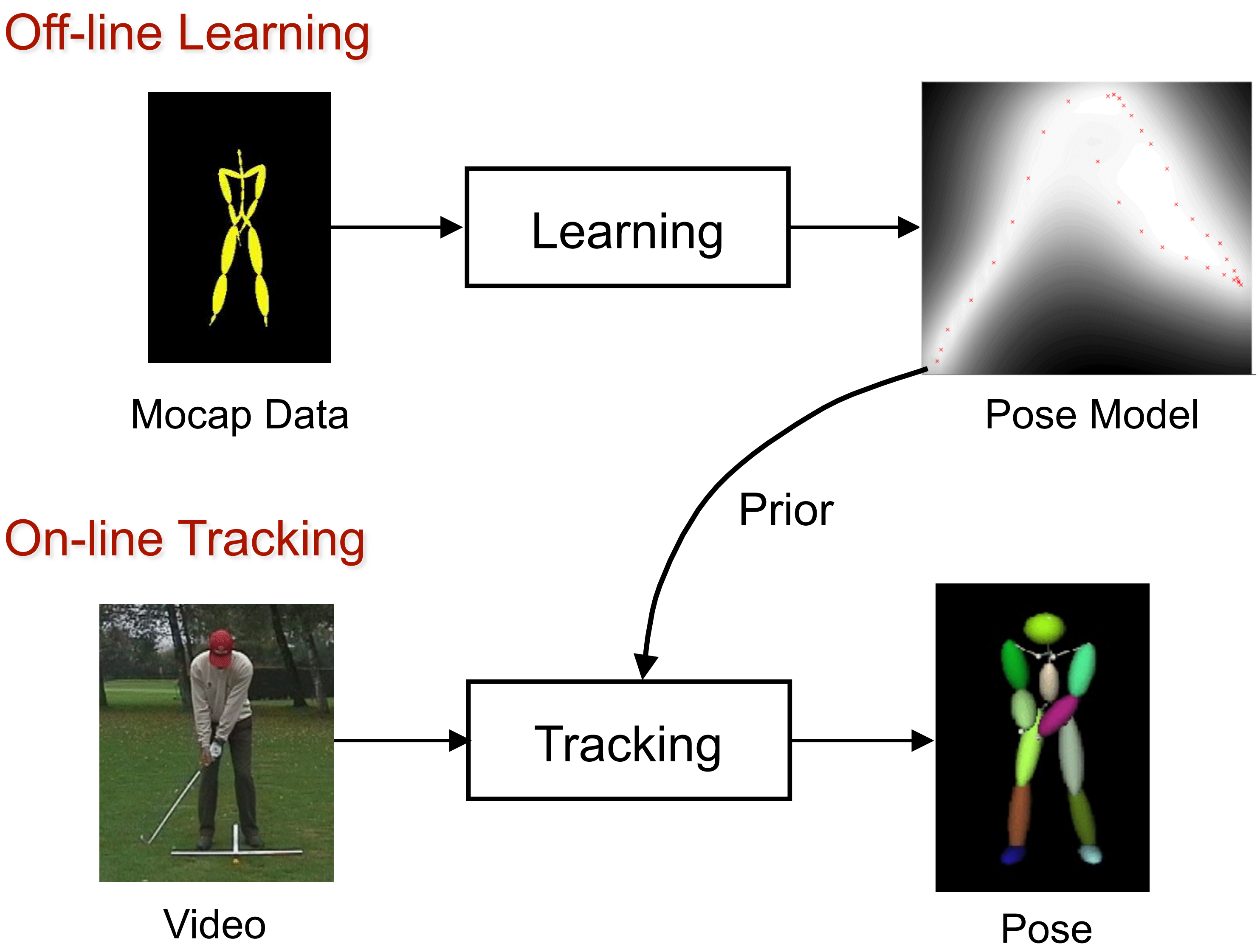
 

Monocular 3D people tracking is usually under-constrained.

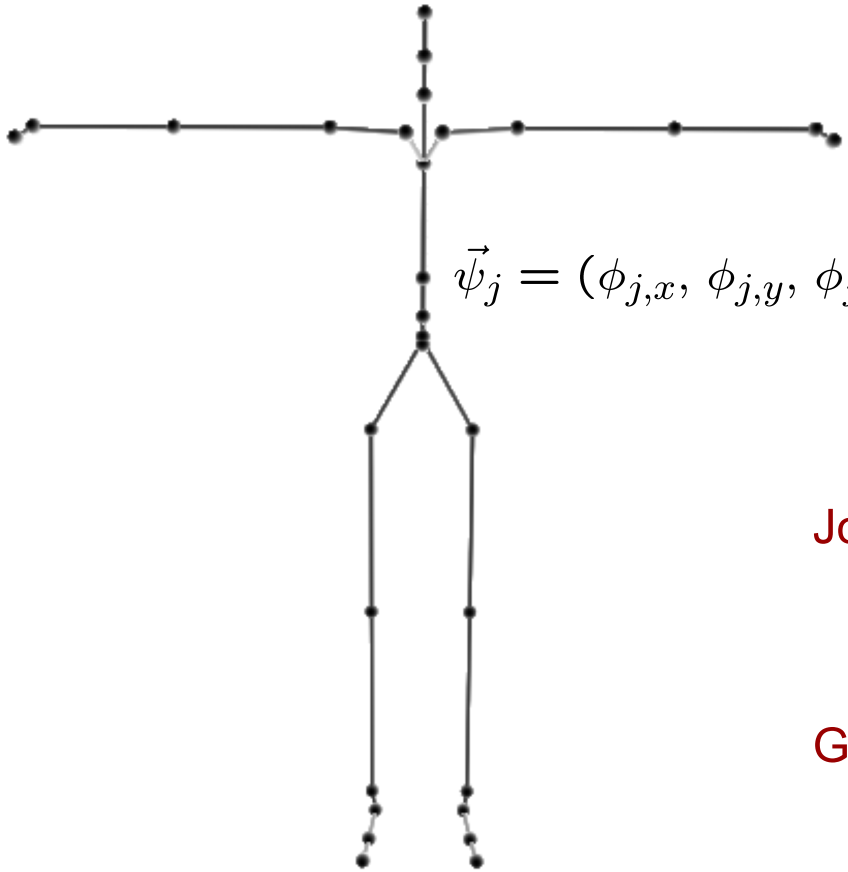
Priors resolve ambiguities but are difficult to learn because:

- human parameterizations are high-dimensional
- training data is hard to acquire

Approach


Off-line Learning
Learning
Mocap Data
Pose Model
Prior
On-line Tracking
Tracking
Video
Pose

Human Parameterization



$$\vec{\psi}_j = (\phi_{j,x},\, \phi_{j,y},\, \phi_{j,z})$$

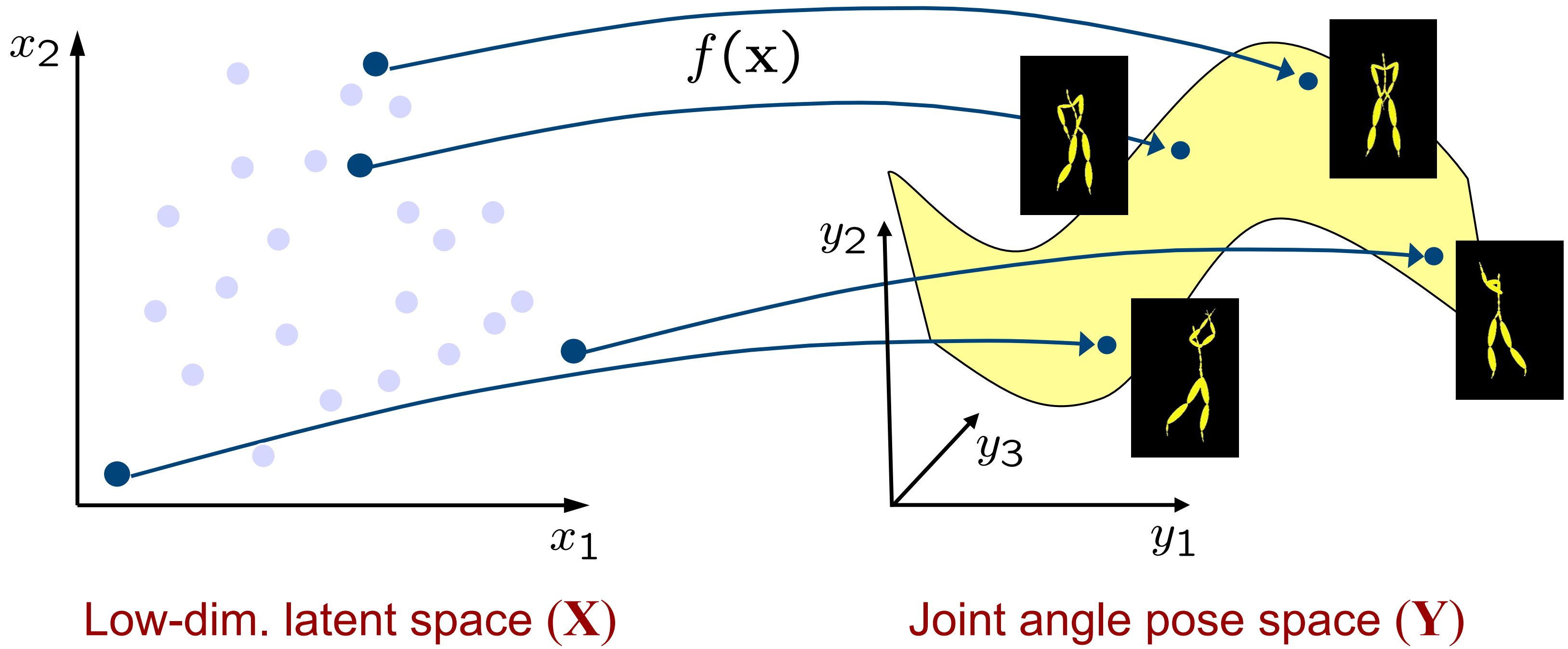
Joint angles:

$$\mathbf{y} = (\vec{\psi}_1,\, ...,\, \vec{\psi}_m) \in \mathcal{R}^d$$

Global pose:

$$\mathbf{G} = (\mathbf{p},\, \vec{\psi}_G)$$

Latent Variable Models / Dimensionality Reduction

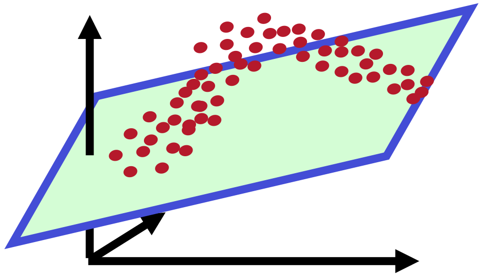


Mapping from latent points to poses, $f(\mathbf{x})$

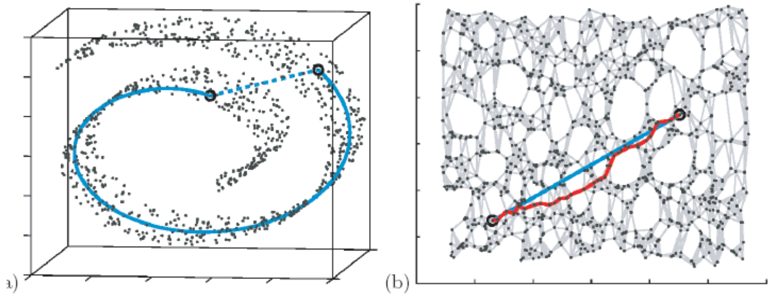
Smooth density function over pose

Latent Variable Models / Dimensionality Reduction

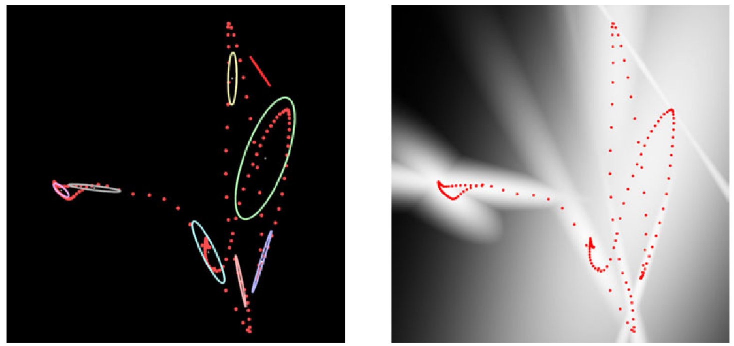
PCA / PPCA
[*Sidenbladh et al '00; Urtasun et al '04; ….*]



Isomap / LLE / Spectral Methods
[*Lee & Elgammal '04; Sminchisescu & Jepson '04; Wang et al '03; …*]

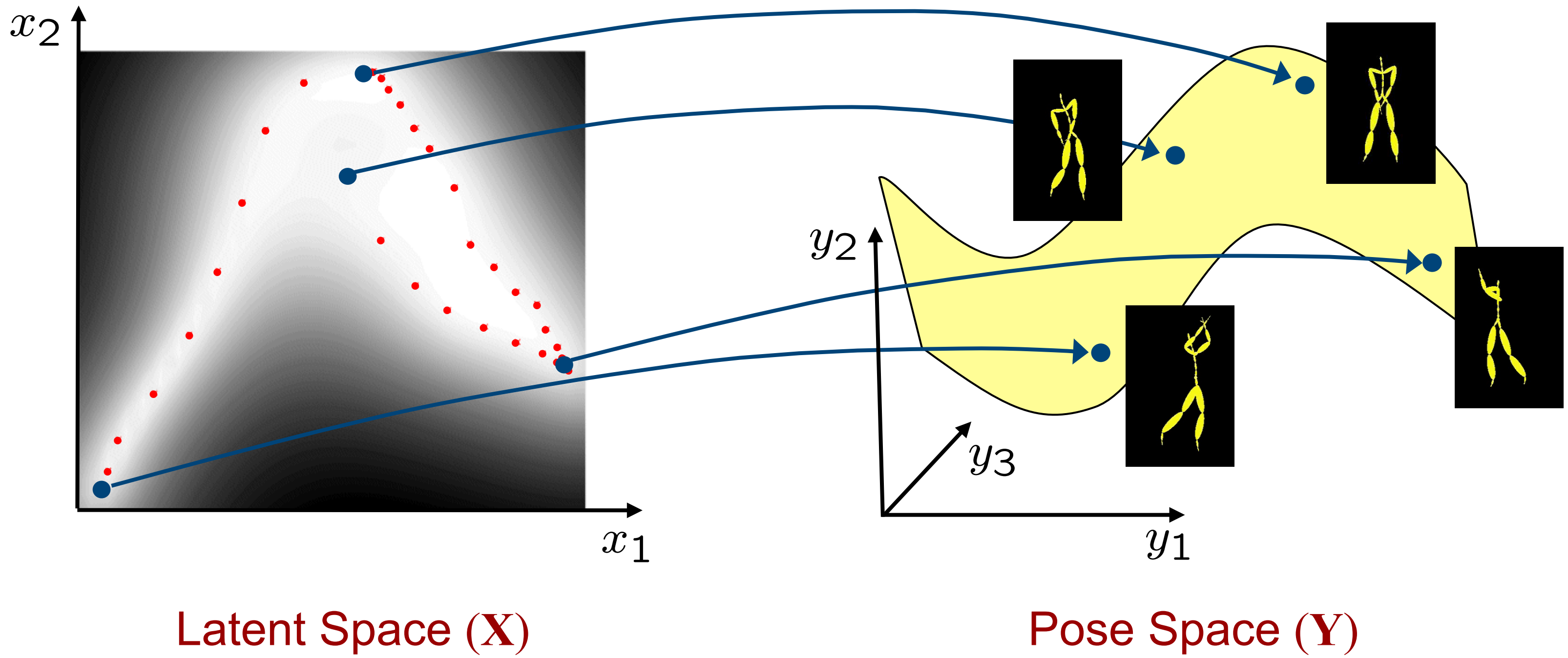


Mixture models
[*Howe et al, 99; Sminchisescu & Jepson '04; …*]



Gaussian components Log-likelihood

Gaussian Process Latent Variable Model (GPLVM)



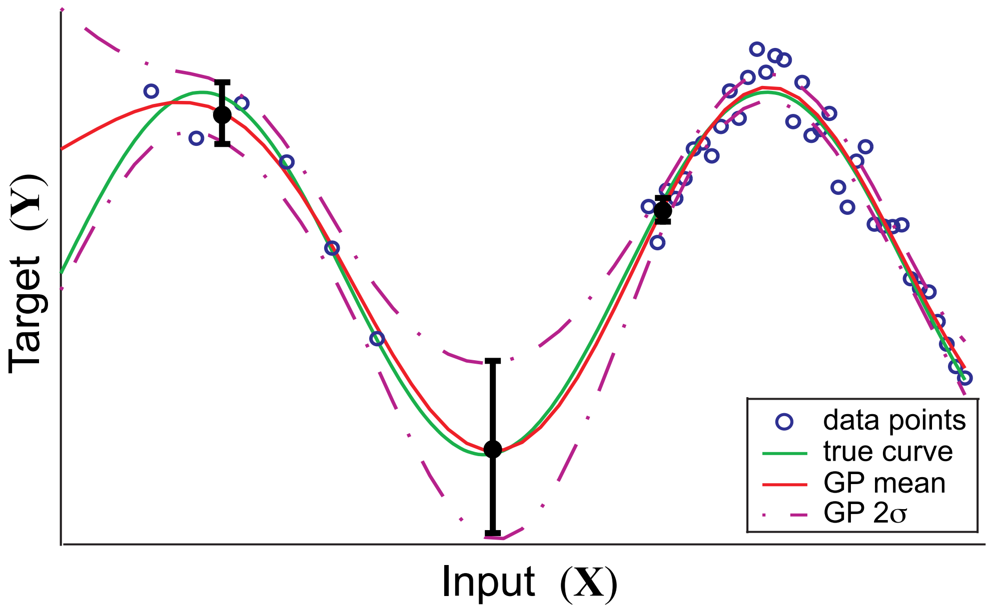
Probabilistic, nonlinear dimensionality reduction [Lawrence 04]

- a nonlinear mapping from latent positions to pose space
- a smooth density function over pose space
- learning based on little data and minimal parameter tuning

GPLVM

The nonlinear manifold is modeled by Gaussian Process regression, with model averaging used to integrate out uncertainty in the model.

GP variance depends on $\mathbf{x}$



For the GPLVM we learn the GP mapping and the latent coordinates of the training poses.

GPLVM Learning

Training poses: $\mathbf{Y} \equiv [\mathbf{y}_1, \cdots, \mathbf{y}_N]^T, \; \mathbf{y}_n \in \mathcal{R}^d$

Model parameters:

- 2D latent coordinates: $\mathbf{X} \equiv \{\mathbf{x}_n\}_{n=1}^N$
- RBF kernel hyperparameters: $\bar{\beta} = \{\beta_j\}$
- weights on output dimensions: $\mathbf{W} = \mathrm{diag}(w_1, ..., w_d)$

Learning: estimate GPLVM parameters by maximizing

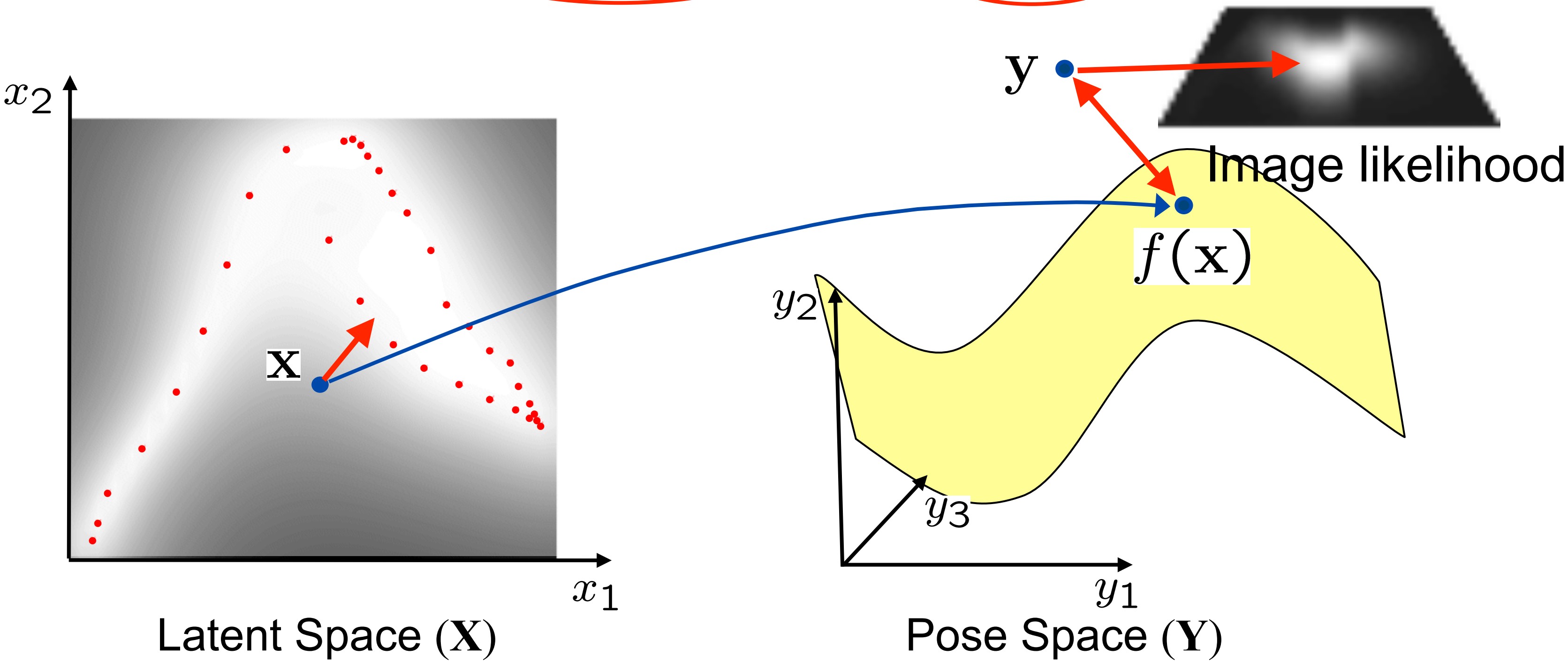
$$p(\mathbf{Y} \mid \mathbf{X}, \bar{\beta}, \mathbf{W}) \; p(\mathbf{X}, \bar{\beta}, \mathbf{W})$$

data likelihood prior

GPLVM prior

The model M then provides a density function over new poses, with negative log likelihood:

$$L(\mathbf{x}, \mathbf{y}; M) = \frac{\|\mathbf{W}(\mathbf{y} - \mathbf{f}(\mathbf{x}))\|^2}{2\sigma^2(\mathbf{x})} + \frac{d}{2} \ln \sigma^2(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}\|^2$$



3D People Tracking

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

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

GPLVM: M

Global pose　Joint angles　Latent coordinates

Posterior Distribution:

$$p(\phi_t \,|\, \mathbf{I}_{1:t}, M) \;\approx\; p(\mathbf{I}_t \,|\, \phi_t)\;\; p(\phi_t \,|\, \phi_{t-1}^{\mathsf{MAP}}, \phi_{t-2}^{\mathsf{MAP}}, M)$$

Likelihood　Dynamics + GPLVM

Online estimation by hill climbing on the negative log posterior:

$$-\ln p(\mathbf{I}_t \,|\, \phi_t) \;+\; D(\phi_t;\, \phi_{t-1}^{\mathsf{MAP}}, \phi_{t-2}^{\mathsf{MAP}}) \;+\; L(\mathbf{x}_t, \mathbf{y}_t;\, M)$$

Measurement Model (WSL 2D Tracker)

2D positions of J joints are tracked (up to IID Gaussian noise):

$$-\ln p(\mathbf{I}_t \,|\, \phi_t) \;=\; \frac{1}{2\sigma_e^2} \sum_{j=1}^{J} ||\, \hat{\mathbf{m}}_t^j - P(\mathbf{p}^j, \phi_t)\, ||^2 \;+\; c$$

$P(\mathbf{p}^j, \phi_t)$ is the perspective projection of point j at time t.

$\hat{\mathbf{m}}_t^j$ is the associated image measurement

Temporal Dynamics

A 2nd-order Markov model is assumed for joint angles and global position / orientation, with IID Gaussian process noise:

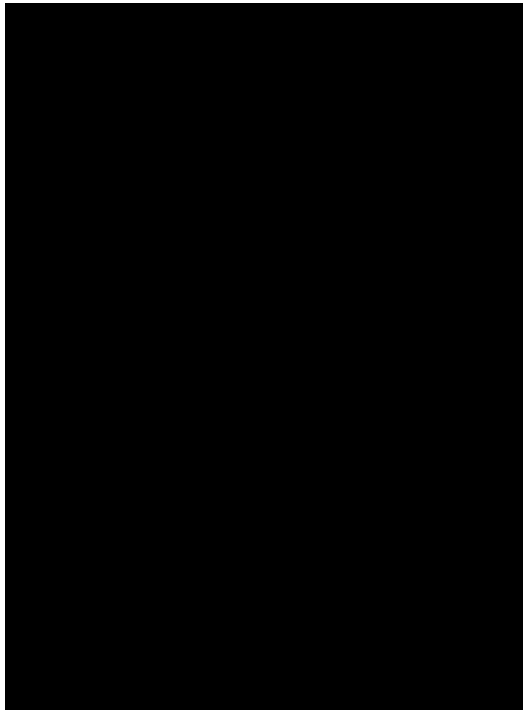
$$D(\phi_t;\, \phi_{t-1}^{\mathsf{MAP}}, \phi_{t-2}^{\mathsf{MAP}}) \;=\; \frac{||\mathbf{y}_t - \hat{\mathbf{y}}_t||^2}{2\sigma_y^2} + \frac{||\mathbf{G}_t - \hat{\mathbf{G}}_t||^2}{2\sigma_G^2}$$

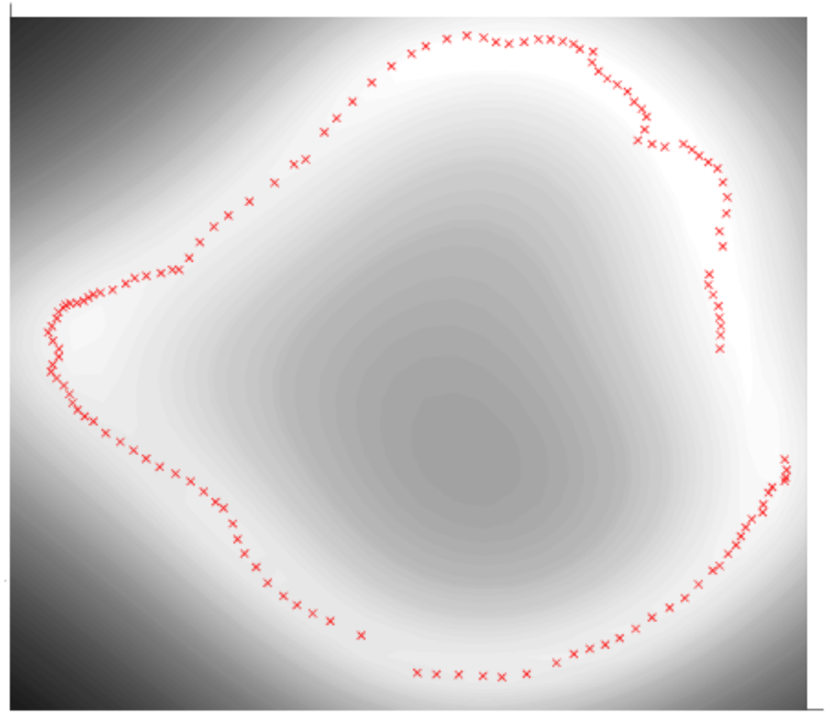
with predictions:

$$\hat{\mathbf{y}}_t \;=\; 2\,\mathbf{y}_{t-1}^{\mathsf{MAP}} - \mathbf{y}_{t-2}^{\mathsf{MAP}}$$

$$\hat{\mathbf{G}}_t \;=\; 2\,\mathbf{G}_{t-1}^{\mathsf{MAP}} - \mathbf{G}_{t-2}^{\mathsf{MAP}}$$

GPLVM Prior: Walking





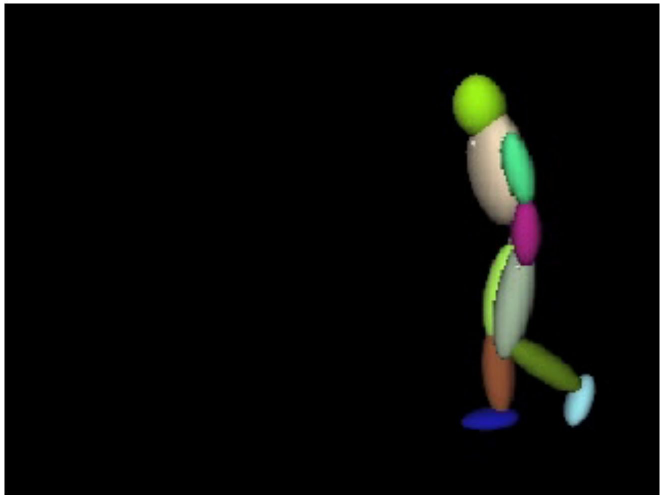
1 gait cycle on a treadmill

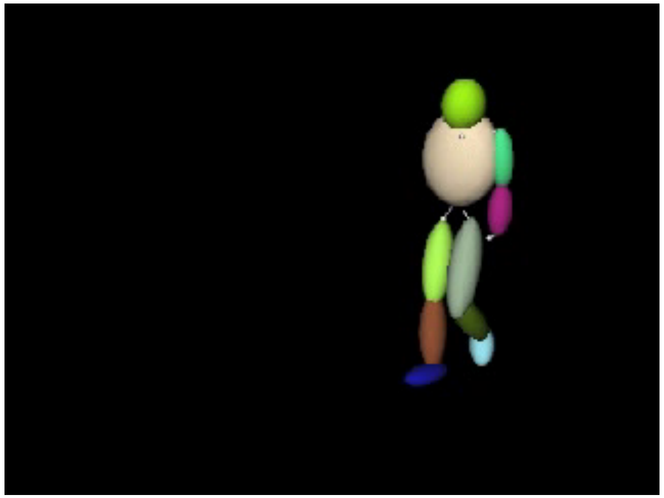
(84 joint angles, 24 active set points)

Tracking: Walking

Tracked 2D Points

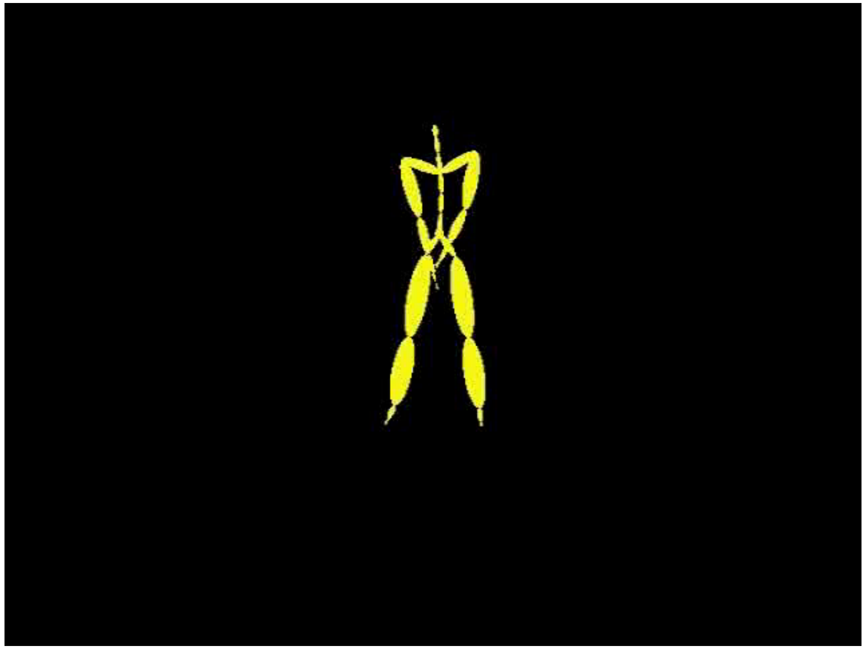
Projected 3D Model

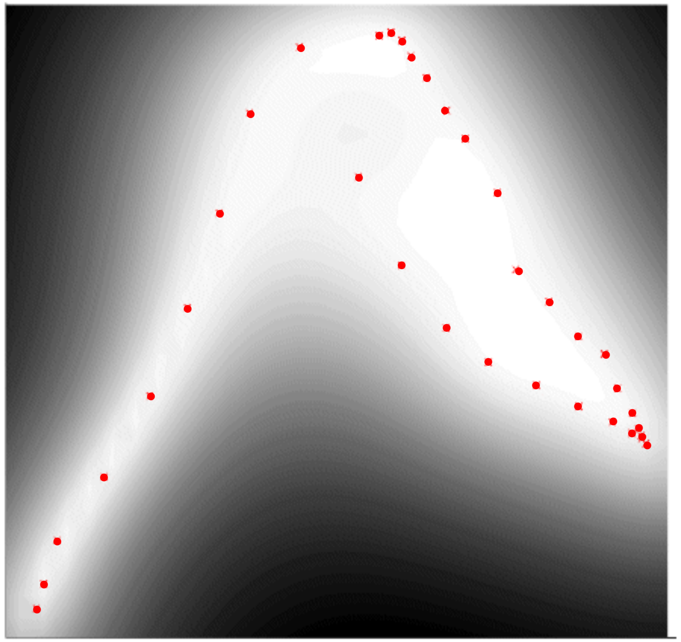




Animations from other viewpoints

SGPLVM Prior: Golf Swing



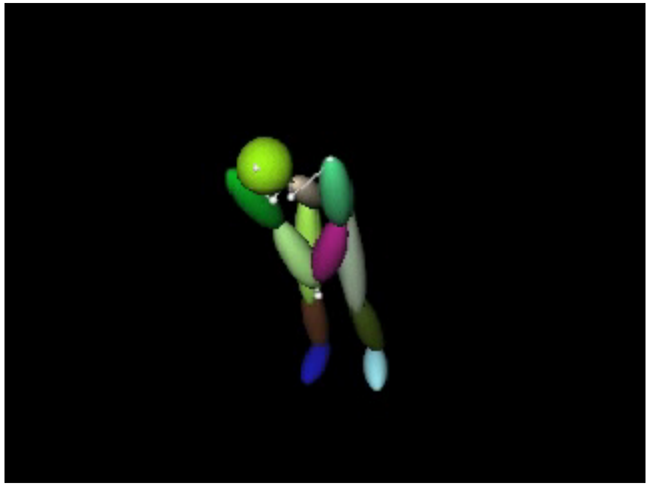


1 swing of golf club from CMU mocap database

(72 joint angles, 19 active set size)

Tracking: Short Swing

Projected 3D Model



Animations from other viewpoints

Tracking: Full Swing

Projected 3D Model

Animations from other viewpoints

Summary

Key Ideas:

- Prior models of human motion learned using the Gaussian Process Latent Variable Model
- Learning from just single training motion
- ML tracking with hill-climbing on log posterior

Limitations / Future work:

- Learning is sensitive to initialization and priors on model parameters
- Works best for small training sets
- Temporal dynamics and appearance models used