3D People Tracking with Gaussian Process Dynamical Models*

Raquel Urtasun
Computer Vision Laboratory
EPFL, Switzerland
raquel.urtasun@epfl.ch

David J. Fleet
Dept. of Computer Science
University of Toronto, Canada
fleet@cs.toronto.edu

Pascal Fua
Computer Vision Laboratory
EPFL, Switzerland
pascal.fua@epfl.ch

Abstract

We advocate the use of Gaussian Process Dynamical Models (GPDMs) for learning human pose and motion priors for 3D people tracking. A GPDM provides a low-dimensional embedding of human motion data, with a density function that gives higher probability to poses and motions close to the training data. With Bayesian model averaging a GPDM can be learned from relatively small amounts of data, and it generalizes gracefully to motions outside the training set. Here we modify the GPDM to permit learning from motions with significant stylistic variation. The resulting priors are effective for tracking a range of human walking styles, despite weak and noisy image measurements and significant occlusions.

1. Introduction

Prior models of pose and motion play a central role in 3D monocular people tracking, mitigating problems caused by ambiguities, occlusions, and image measurement noise. While powerful models of 3D human pose are emerging, sophisticated motion models remain rare. Most state-of-the-art approaches rely on linear-Gaussian Markov models which do not capture the complexities of human dynamics. Learning richer models is challenging because of the high-dimensional variability of human pose, the nonlinearity of human dynamics, and the relative difficulty of acquiring large amounts of training data.

This paper shows that effective models for people tracking can be learned using the Gaussian Process Dynamical Model (GPDM) [22], even when modest amounts of training data are available. The GPDM is a latent variable model with a nonlinear probabilistic mapping from latent positions $\mathbf{x}$ to human poses $\mathbf{y}$, and a nonlinear dynamical mapping on the latent space. It provides a continuous density function over poses and motions that is generally non-Gaussian and multimodal. Given training sequences, one simultaneously learns the latent embedding, the latent dynamics, and the pose reconstruction mapping. Bayesian model averaging is used lessen problems of over-fitting and under-fitting that are otherwise problematic with small training sets [10, 12].

We propose a form of GPDM, the *balanced GPDM*, for learning smooth models from training motions with stylistic diversity, and show that they are effective for monocular people tracking. We formulate the tracking problem as a MAP estimator on short pose sequences in a sliding temporal window. Estimates are obtained with deterministic optimization, and look remarkably good despite very noisy, missing or erroneous image data and significant occlusions.

2. Related Work

The dynamical models used in many tracking algorithms are weak. Most models are linear with Gaussian process noise, including simple first- and second-order Markov models [3, 9], and auto-regressive (AR) models [14]. Such models are often suitable for low-dimensional problems, and admit closed-form analysis, but they apply to a restricted class of systems. For high-dimensional data, the number of parameters that must be manually specified or learned for AR models is untenable. When used for people tracking they usually include large amounts of process noise, and thereby provide very weak temporal predictions.

Switching LDS and hybrid dynamics provide much richer classes of temporal behaviors [8, 14, 15]. Nevertheless, they are computationally challenging to learn, and require large amounts of training data, especially as the dimension of the state space grows. Non-parametric models can also handle complex motions, but they also require very large amounts of training data [11, 17]. Further, they do not produce a density function. Howe et al [7] use mixture model density estimation to learn a distribution of short sequences of poses. Again, with such high-dimensional data, density estimation will have problems of under- and over-fitting unless one has vast amounts of training data.

One way to cope with high-dimensional data is to learn low-dimensional latent variable models. The simplest case involves a linear subspace projection with an AR dynamical process. In [2, 4] a subspace is first identified using PCA, afterwhich a subspace AR model is learned. Linear models are tractable, but they often lack the ability to capture the complexities of human pose and motion.

*This work was supported in part by the Swiss National Science Foundation, NSERC Canada, and the Canadian Institute for Advanced Research. We thank A. Hertzmann and J. Wang for many useful discussions.

Richer parameterizations of human pose and motion can be found through nonlinear dimensionality reduction [5, 16, 18, 21]. Geometrical methods such as Isomap and LLE learn such embeddings, yielding mappings from the pose space to the latent space. But they do not provide a probabilistic density model over poses, a mapping back from pose space to latent space, nor a dynamical model. Thus one requires additional steps to construct an effective model. For example, Sminchisescu and Jepson [18] use spectral embedding, then a Gaussian mixture to model the latent density, an RBF mapping to reconstruct poses from latent positions, and a hand-specified first-order, linear dynamical model. Agarwal and Triggs [1] learn a mapping from silhouettes to poses using relevance vector machines, and then a second-order AR dynamical model.

Rahimi et al [16] learn an embedding through a nonlinear RBF regression with an AR dynamical model to encourage smoothness in the latent space. Our approach is similar in spirit, as this is a natural way to produce well-behaved latent mappings for time-series data. However, our model is probabilistic and allows for nonlinear dynamics.

We use a form of probabilistic dimensionality reduction similar in spirit to the Gaussian Process latent variable model (GPLVM) [10]. The GPLVM has been used to constrain human poses during interactive animation [6], as a prior for 2D upperbody pose estimation [19], and as a prior for 3D monocular people tracking [20]. While powerful, the GPLVM is a static model; it has no intrinsic dynamics and does not produce smooth latent paths from smooth time-series data. Thus, even with an additional dynamical model, our GPLVM-based people tracker often fails due to anomalous jumps in the latent space and to occlusions [20].

3. Gaussian Process Dynamical Model

The GPDM is a latent variable dynamical model, comprising a low-dimensional latent space, a probabilistic mapping from the latent space to the pose space, and a dynamical model in the latent space [22]. The GPDM is derived from a generative model for zero-mean poses $\mathbf{y}_t \in \mathcal{R}^D$ and latent positions $\mathbf{x}_t \in \mathcal{R}^d$, at time t, of the form

$$\mathbf{x}_t = \sum_i \mathbf{a}_i \, \phi_i(\mathbf{x}_{t-1}) + \mathbf{n}_{x,t} \quad (1)$$

$$\mathbf{y}_t = \sum_j \mathbf{b}_j \, \psi_j(\mathbf{x}_t) + \mathbf{n}_{y,t} \quad (2)$$

for weights $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ...]$ and $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, ...]$, basis functions ϕ_i and ψ_j, and additive zero-mean white Gaussian noise $\mathbf{n}_{x,t}$ and $\mathbf{n}_{y,t}$. For linear basis functions, (1) and (2) represent the common subspace AR model (e.g., [4]). With nonlinear basis functions, the model is significantly richer.

In conventional regression (e.g., with AR models) one fixes the number of basis functions and then fits the model parameters, $\mathbf{A}$ and $\mathbf{B}$. From a Bayesian perspective, $\mathbf{A}$ and $\mathbf{B}$ are nuisance parameters and should therefore be marginalized out through model averaging. With an isotropic Gaussian prior on each $\mathbf{b}_j$, one can marginalize over $\mathbf{B}$ in closed form [12, 13] to yield a multivariate Gaussian data likelihood of the form

$$p(\mathbf{Y} \mid \mathbf{X}, \bar{\beta}) = \frac{|\mathbf{W}|^N}{\sqrt{(2\pi)^{ND}|\mathbf{K}_Y|^D}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{K}_Y^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)\right) \quad (3)$$

where $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]^T$ is a matrix of training poses, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^T$ contains the associated latent positions, and $\mathbf{K}_Y$ is a kernel matrix. The elements of kernel matrix are defined by a kernel function, $(\mathbf{K}_Y)_{i,j} = k_Y(\mathbf{x}_i, \mathbf{x}_j)$, which we take to be a common radial basis function (RBF) [12]:

$$k_Y(\mathbf{x}, \mathbf{x}') = \beta_1 \exp\left(-\frac{\beta_2}{2}||\mathbf{x} - \mathbf{x}'||^2\right) + \frac{\delta_{\mathbf{x},\mathbf{x}'}}{\beta_3}. \quad (4)$$

The scaling matrix $\mathbf{W} \equiv \mathrm{diag}(w_1, ..., w_D)$ is used to account for the different variances in the different data dimensions; this is equivalent to a Gaussian Process (GP) with kernel function $k(\mathbf{x}, \mathbf{x}')/w_l^2$ for dimension l. Finally, $\bar{\beta} = \{\beta_1, \beta_2, ..., \mathbf{W}\}$ comprises the kernel hyperparameters that control the output variance, the RBF support width, and the variance of the additive noise $\mathbf{n}_{y,t}$.

The latent dynamics are similar; i.e., we form the joint density over latent positions and weights, $\mathbf{A}$, and then we marginalize out $\mathbf{A}$ [22]. With an isotropic Gaussian prior on the $\mathbf{a}_i$, the density over latent trajectories reduces to

$$p(\mathbf{X} \mid \bar{\alpha}) = \frac{p(\mathbf{x}_1)}{\sqrt{(2\pi)^{(N-1)d}|\mathbf{K}_X|^d}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{K}_X^{-1}\mathbf{X}_{out}\mathbf{X}_{out}^T\right)\right) \quad (5)$$

where $\mathbf{X}_{out} = [\mathbf{x}_2, ..., \mathbf{x}_N]^T$, $\mathbf{K}_X$ is the $(N-1) \times (N-1)$ kernel matrix constructed from $\mathbf{X}_{in} = [\mathbf{x}_1, ..., \mathbf{x}_{N-1}]$, and $\mathbf{x}_1$ is given an isotropic Gaussian prior. For dynamics the GPDM uses a "linear + RBF" kernel, with parameters α_i :

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

The linear term is useful for motion subsequences that are approximately linear.

While the GPDM is defined above for a single input sequence, it is easily extended to multiple sequences $\{\mathbf{Y}_j\}$. One simply concatenates all the input sequences, ignoring temporal transitions from the end of one sequence to the beginning of the next. Each input sequence is then associated with a separate sequence of latent positions, $\{\mathbf{X}_j\}$, all within a shared latent space. Accordingly, in what follows, let $\mathbf{Y} = [\mathbf{Y}_1^T, ..., \mathbf{Y}_m^T]^T$ be the m training motions. Let $\mathbf{X}$ denote the associated latent positions, and for the definition of (5) let $\mathbf{X}_{out}$ comprise all but the first latent position for each sequence. Let $\mathbf{K}_X$ be the kernel matrix computed from all but the last latent position of each sequence.

3.1. Learning

Learning the GPDM entails estimating the latent positions and the kernel hyperparameters. Following [22] we adopt simple prior distributions over the hyperparameters, i.e., $p(\bar{\alpha}) \propto \prod_i \alpha_i^{-1}$, and $p(\bar{\beta}) \propto \prod_i \beta_i^{-1}$,[1] with which the GPDM posterior becomes

$$p(\mathbf{X}, \bar{\alpha}, \bar{\beta} \,|\, \mathbf{Y}) \;\propto\; p(\mathbf{Y} \,|\, \mathbf{X}, \bar{\beta})\, p(\mathbf{X} \,|\, \bar{\alpha})\, p(\bar{\alpha})\, p(\bar{\beta}) \;. \quad (6)$$

The latent positions and hyperparameters are found by minimizing the negative log posterior

$$\begin{aligned} \mathcal{L} \;=\; & \frac{d}{2} \ln |\mathbf{K}_X| \;+\; \frac{1}{2} \mathrm{tr}\left(\mathbf{K}_X^{-1} \mathbf{X}_{out} \mathbf{X}_{out}^T\right) \\ & - N \ln |\mathbf{W}| \;+\; \frac{D}{2} \ln |\mathbf{K}_Y| \;+\; \frac{1}{2} \mathrm{tr}\left(\mathbf{K}_Y^{-1} \mathbf{Y} \mathbf{W}^2 \mathbf{Y}^T\right) \\ & + \sum_i \ln \alpha_i \;+\; \sum_i \ln \beta_i \;+\; C \;, \end{aligned} \quad (7)$$

where C is a constant. The first two terms come from the log dynamics density (5), and the next three terms come from the log reconstruction density (3).

Over-Fitting: While the GPDM has advantages over the GPLVM, usually producing much smoother latent trajectories it can still produce large gaps between the latent positions of consecutive poses; e.g., Fig. 1 shows a GPLVM and a GPDM learned from the same golf swing data (large gaps are shown with red arrows). Such problems tend to occur when the training set includes a relatively large number of individual motions (e.g., from different people or from the same person performing an activity multiple times). The problem arises because of the large number of unknown latent coordinates and the fact that uncertainty in latent positions is not modeled. In practical terms, the GPDM learning estimates the latent positions by simultaneously minimizing squared reconstruction errors in pose space and squared temporal prediction errors in the latent space. In Fig. 1 the pose space is 80D and the latent space is 3D, so it is not surprising that the errors in pose reconstruction dominate the objective function, and thus the latent positions.

3.2. Balanced GPDM:

Ideally one should marginalize out the latent positions to learn hyperparameters, but this is expensive computationally. Instead, we propose a simple but effective GPDM modification to balance the influence of the dynamics and the pose reconstruction in learning. That is, we discount the differences in the pose and latent space dimensions in the two regressions by raising the dynamics density function in (6) to the ratio of their dimensions, i.e., $\lambda = D/d$;

[1] Such priors prefer small output scale (i.e., $\alpha_1, \alpha_3, \beta_1$), large RBF support (i.e., small α_2, β_2), and large noise variances (i.e., small $\alpha_4^{-1}, \beta_3^{-1}$). The fact that the priors are improper is insignificant for optimization.

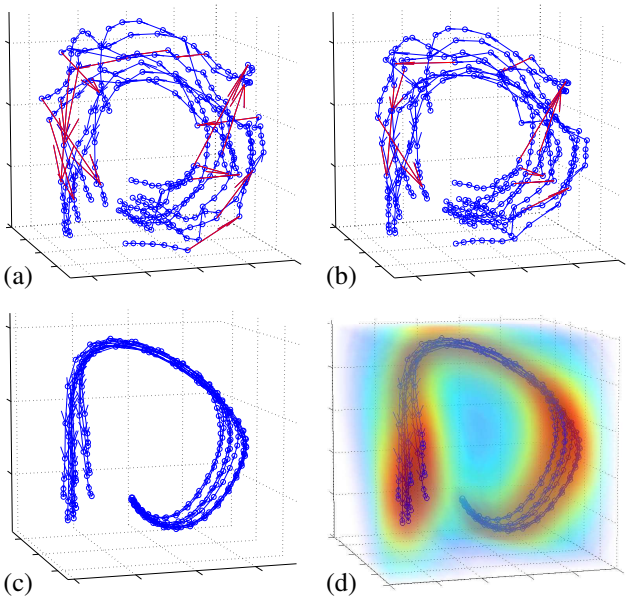

Figure 1. **Golf Swing:** (a) GPLVM, (b) GPDM and (c) balanced GPDM learned from 9 different golf swings performed by the same subject. (d) Volumetric visualization of reconstruction variance; warmer colors (i.e., red) depict lower variance.

for learning this rescales the first two terms in (7) to be

$$\lambda \left(\frac{d}{2} \ln |\mathbf{K}_X| \;+\; \frac{1}{2} \mathrm{tr}\left(\mathbf{K}_X^{-1} \mathbf{X}_{out} \mathbf{X}_{out}^T\right) \right) \;. \quad (8)$$

The resulting models are easily learned and very effective.

3.3. Model Results

Figures 1–4 show models learned from motion capture data. In each case, before minimizing $\mathcal{L}$, the mean pose, μ, was subtracted from the input pose data, and PCA or Isomap were used to obtain an initial latent embedding of the desired dimension. We typically use a 3D latent space as this is the smallest dimension for which we can robustly learn complex motions with stylistic variability. The hyperparameters were initially set to one. The negative log posterior $\mathcal{L}$ was minimized using Scaled Conjugate Gradient.

Golf Swing: Fig. 1 shows models learned from 9 golf swings from one subject (from the CMU database). The body pose was parameterized with 80 joint angles, and the sequence lengths varied by 15 percent. The balanced GPDM (Fig. 1(c)) produces smoother latent trajectories, and hence a more reliable dynamic model, than the original GPDM. Fig. 1(d) shows a volume visualization of the log variance of the reconstruction mapping, $2 \ln \sigma_{\mathbf{y}|\mathbf{x},\mathbf{X},\mathbf{Y},\bar{\beta}}$, as a function of latent position. Warmer colors correspond to lower variances, and thus to latent positions to which the model assigns higher probability; this shows the model's preference for poses close to the training data.

Walking: Figs 2 and 3 show models learned from one gait cycle from each of 6 subjects walking at the same speed on a

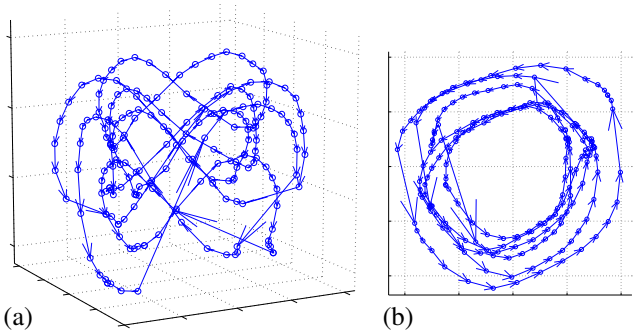

Figure 2. **Walking GPLVM:** Learned from 1 gait cycle from each of 6 subjects. Plots show side and top views of the 3D latent space. Circles and arrows denote latent positions and temporal sequence.

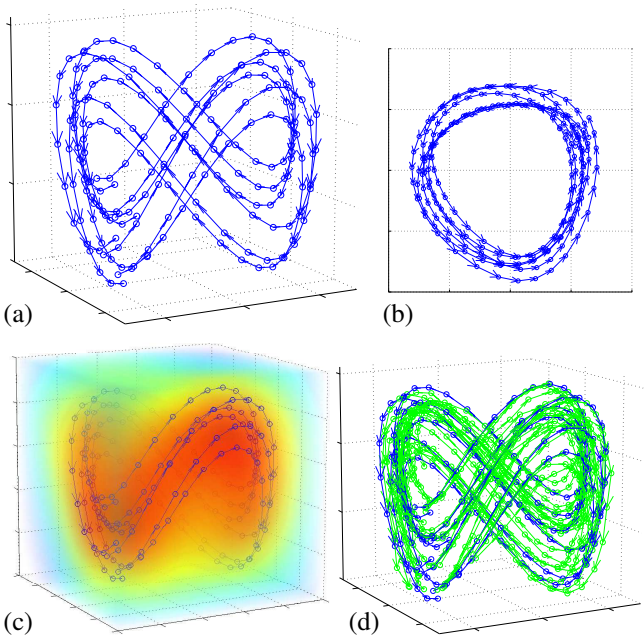

Figure 3. **Walking GPDM:** Balanced GPDM learned from 1 gait cycle from 6 subjects. (a,b) Side and top views of 3D latent space. (c) Volumetric visualization of reconstruction variance. (d) Green trajectories are fair samples from the dynamics model.

treadmill. For each subject the first pose is replicated at the end of the sequence to encourage cyclical paths in the latent space. The body was parameterized with 20 joint angles. With the treadmill we do not have global position data, and hence we cannot learn the coupling between the joint angle times series and global translational velocity.

Fig. 2 shows the large jumps in adjacent poses in the latent trajectories obtained with a GPLVM. By comparison, Fig. 3 (a,b) show the smooth, clustered latent trajectories learned from the training data. Fig. 3(c) shows a volume visualization of the log reconstruction variance. Fig. 3(d) helps to illustrate the model dynamics by plotting 20 latent trajectories drawn at random from the dynamical model. The trajectories are smooth and close to the training data.

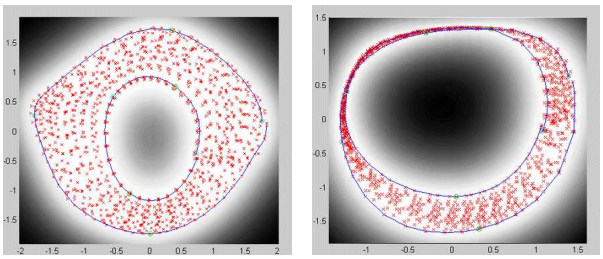

Figure 4. **Speed Variation:** 2D models learned for 2 different subjects. Each one walking at 9 speeds ranging from 3 to 7 km/h. Red points are latent positions of training poses. Intensity is proportional to $-2 \ln \sigma_{\mathbf{y}|\mathbf{x},\mathbf{X},\mathbf{Y},\bar{\beta}}$, so brighter regions have smaller pose reconstruction variance. The subject on the left is healthy while that on the right has a knee pathology and walks asymmetrically.

Speed Variation: Fig. 4 shows 2D GPDMs learned from two subjects, each of which walked four gait cycles at each of 9 speeds between 3 and 7km/h (equispaced). The learned latent trajectories are approximately circular, and organized by speed; the innermost and outermost trajectories correspond to the slowest and fastest speeds respectively. Interestingly, the subject on the left is healthy while the subject on right has a knee pathology. As the treadmill speed increases, the side of the body with the pathology performs the motion at slower speeds to avoid pain, and so the other side of the gait cycle must speed up to maintain the speed. This explains the anisotropy of the latent space.

3.4. Prior over New Motions

The GPDM also defines a smooth probability density over new motions $(\mathbf{Y}', \mathbf{X}')$. That is, just as we did with multiple sequences above, we write the joint density over the concatenation of the sequences. The conditional density of the new sequence is proportional to the joint density, but with the training data and latent positions held fixed:

$$p(\mathbf{X}', \mathbf{Y}' \mid \mathbf{X}, \mathbf{Y}, \bar{\alpha}, \bar{\beta}) \;\propto\; p(\,[\mathbf{X}, \mathbf{X}'], [\mathbf{Y}, \mathbf{Y}'] \mid \bar{\alpha}, \bar{\beta}) \quad (9)$$

This density can also be factored to provide:

$$p(\mathbf{Y}' \mid \mathbf{X}',\, \mathbf{X},\, \mathbf{Y},\, \bar{\beta})\, p(\mathbf{X}' \mid \mathbf{X},\, \bar{\alpha})\,. \quad (10)$$

For tracking we are typically given an initial state $\mathbf{x}'_0$, so instead of (10), we have

$$p(\mathbf{Y}' \mid \mathbf{X}',\, \mathbf{X},\, \mathbf{Y},\, \bar{\beta})\, p(\mathbf{X}' \mid \mathbf{X},\, \bar{\alpha}, \mathbf{x}'_0)\,. \quad (11)$$

4. Tracking

Our tracking formulation is based on a state-space model, with a GPDM prior over pose and motion. The state at time t is defined as $\phi_t = [\mathbf{G}_t, \mathbf{y}_t, \mathbf{x}_t]$, where $\mathbf{G}_t$ denotes the global position and orientation of the body, $\mathbf{y}_t$ denotes the articulated joint angles, and $\mathbf{x}_t$ is a latent position. The goal is to estimate a state sequence, $\phi_{1:T} \equiv (\phi_1, ..., \phi_T)$, given an image sequence, $\mathbf{I}_{1:T} \equiv (\mathbf{I}_1, ..., \mathbf{I}_T)$, and a learned GPDM, $\mathcal{M} \equiv (\mathbf{X},\, \mathbf{Y},\, \bar{\alpha},\, \bar{\beta})$. Toward that end there are

two common approaches: *Online methods* infer ϕ_t given the observation history $\mathbf{I}_{1:t-1}$. The inference is causal, and usually recursive, but suboptimal as it ignores future data. *Batch methods* infer states ϕ_t given all past, present and future data, $\mathbf{I}_{1:T}$. Inference is optimal, but requires all future images which is impossible in many tracking applications.

Here we propose a compromise that allows some use of future data along with predictions from previous times. In particular, at each time t we form the posterior distribution over a (noncausal) sequence of $\tau+1$ states

$$p(\phi_{t:t+\tau} \,|\, \mathbf{I}_{1:t+\tau},\, \mathcal{M}) \;= \\ c\; p(\mathbf{I}_{t:t+\tau} \,|\, \phi_{t:t+\tau})\; p(\phi_{t:t+\tau} \,|\, \mathbf{I}_{1:t-1},\, \mathcal{M})\;. \quad (12)$$

Inference of ϕ_t is improved with the use of future data, but at the cost of a small temporal delay.[2] With a Markov chain model one could use a forward-backward inference algorithm [23] in which separate beliefs about each state from past and future data are propagated forward and backward in time. Here, instead we consider the posterior over the entire window, without requiring the Markov factorization.

With the strength of the GPDM prior, we also assume that we can use hill-climbing to find good state estimates (i.e., MAP estimates). In effect, we assume a form of approximate recursive estimation:

$$p(\phi_{t:t+\tau} \,|\, \mathbf{I}_{1:t+\tau},\, \mathcal{M}) \;\approx \\ c\; p(\mathbf{I}_{t:t+\tau} \,|\, \phi_{t:t+\tau})\; p(\phi_{t:t+\tau} \,|\, \phi^{MAP}_{1:t-1},\, \mathcal{M}) \quad (13)$$

where $\phi^{MAP}_{1:t-1}$ denotes the MAP estimate history. This has the disadvantage that complete beliefs are not propagated forward. But with the temporal window we still exploit data over several frames, yielding smooth tracking.

At each time step we minimize the negative log posterior over states from time t to time $t+\tau$. At this minima we obtain the approximate MAP estimate at time t. The estimate is approximate in two ways. First, we do not represent and propagate uncertainty forward from time $t-1$ in (13). Second, because previous MAP estimates are influenced by future data, the information propagated forward is biased.

Image Likelihood: The current version of our 3D tracker uses a simplistic observation model. That is, the image observations are the approximate 2D image locations of a small number (J) of 3D body points (see Fig. 5). They were obtained with the WSL image-based tracker [9].

While measurement errors in tracking are often correlated over time, as is common we assume that image measurements conditioned on states are independent; i.e.,

$$p(\mathbf{I}_{t:t+\tau} \,|\, \phi_{t:t+\tau}) \;=\; \prod_{i=t}^{t+\tau} p(\mathbf{I}_i \,|\, \phi_i)\;. \quad (14)$$

[2]However an online estimate of $\phi_{t+\tau}$ would still be available at $t+\tau$.

(a) (b) (c) (d) (e)

Figure 5. **WSL Tracks:** The 2D tracked regions for the different tracked sequences (in yellow) are noisy and sometimes missing.

Further, we assume zero-mean Gaussian measurement noise in the 2D image positions provided by the tracker. Let the perspective projection of the j^{th} body point, $\mathbf{p}^j$, in pose ϕ_t, be denoted $P(\mathbf{p}^j(\phi_t))$, and let the associated 2D image measurement from the tracker be $\hat{\mathbf{m}}^j_t$. Then, the negative log likelihood of the observations at time t is

$$-\ln p(\mathbf{I}_t \,|\, \phi_t) \;=\; \frac{1}{2\sigma_e^2} \sum_{j=1}^{J} \left\| \hat{\mathbf{m}}^j_t - P(\mathbf{p}^j(\phi_t)) \right\|^2 . \quad (15)$$

Here we set $\sigma_e = 10$ pixels, based on empirical results.

Prediction Distribution We factor the prediction density $p(\phi_{t:t+\tau} \,|\, \phi^{MAP}_{1:t-1},\, \mathcal{M})$ into a prediction over global motion, and one over poses $\mathbf{y}$ and latent positions $\mathbf{x}$. The reason, as discussed above, is that our training sequences did not contain the global motion. So, we assume that

$$p(\phi_{t:t+\tau} \,|\, \phi^{MAP}_{1:t-1},\, \mathcal{M}) \;= \\ p(\mathbf{X'}_t,\, \mathbf{Y'}_t \,|\, \mathbf{x}^{MAP}_{t-1},\, \mathcal{M})\; p(\mathbf{G}_{t:t+\tau} | \mathbf{G}^{MAP}_{t-1:t-2})\;, \quad (16)$$

where $\mathbf{X'}_t \equiv \mathbf{x}_{t:t+\tau}$ and $\mathbf{Y'}_t \equiv \mathbf{y}_{t:t+\tau}$.

For the global rotation and translation, $\mathbf{G}_t$, we assume a second-order Gauss-Markov model. The negative log transition density is, up to an additive constant,

$$-\ln p(\mathbf{G}_t \,|\, \mathbf{G}^{MAP}_{t-1:t-2}) \;=\; \frac{||\mathbf{G}_t - \hat{\mathbf{G}}_t||^2}{2\sigma_G^2}\;, \quad (17)$$

where the mean prediction is just $\hat{\mathbf{G}}_t = 2\mathbf{G}^{MAP}_{t-1} - \mathbf{G}^{MAP}_{t-2}$.

For the prior over $\mathbf{X'}_t$, $\mathbf{Y'}_t$, we approximate the GPDM in two ways. First we assume that the density over the pose sequence, $p(\mathbf{Y'}_t \mid \mathbf{X'}_t,\, \mathcal{M})$, can be factored into the densities over individual poses. This is convenient computationally since the GPDM density over a single pose, given a latent position, is Gaussian [6, 20]. Thus we obtain

$$-\ln p(\mathbf{Y'}_t | \mathbf{X'}_t, \mathcal{M}) \approx -\sum_{j=t}^{t+\tau} \ln p(\mathbf{y}_j | \mathbf{x}_j, \bar{\beta}, \mathbf{X}, \mathbf{Y}) \\ = \sum_{j=t}^{t+\tau} \frac{\|\mathbf{W}(\mathbf{y}_j - \mu_Y(\mathbf{x}_j))\|^2}{2\sigma^2(\mathbf{x}_j)} + \frac{D}{2}\ln\sigma^2(\mathbf{x}_j) + \frac{1}{2}\|\mathbf{x}_j\|^2 \quad (18)$$

where the mean and variance are given by

$$\mu_Y(\mathbf{x}) \;=\; \mu + \mathbf{Y}^T \mathbf{K}_Y^{-1} \mathbf{k}_Y(\mathbf{x})\;, \quad (19)$$

$$\sigma^2(\mathbf{x}) \;=\; k_Y(\mathbf{x},\mathbf{x}) - \mathbf{k}_Y(\mathbf{x})^T \mathbf{K}_Y^{-1} \mathbf{k}_Y(\mathbf{x})\;, \quad (20)$$

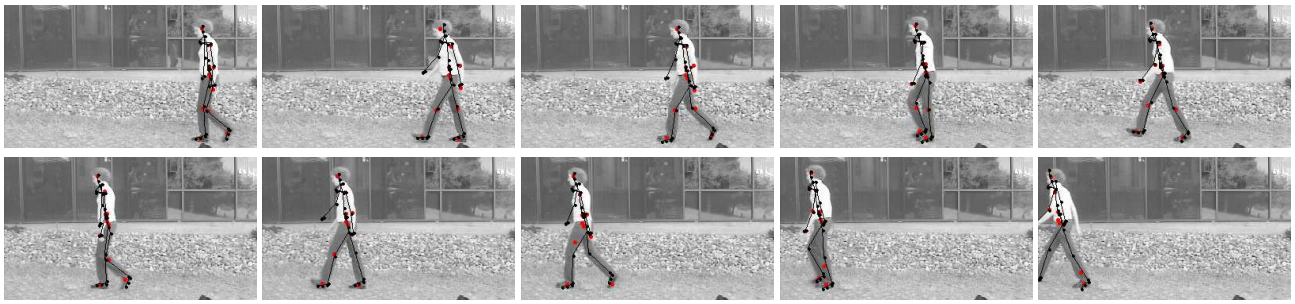

Figure 6. Tracking 63 frames of a walking, with noisy and missing data. The skeleton of the recovered 3D model is projected onto the images. The points tracked by WSL are shown in red.

and $\mathbf{k}_Y(\mathbf{x})$ is the vector with elements $k_Y(\mathbf{x}, \mathbf{x}_j)$ for all other latent positions $\mathbf{x}_j$ in the model.

Second, we anneal the dynamics $p(\mathbf{X'}_t|\mathbf{x}_{t-1}^{MAP}, \mathcal{M})$, because the learned GPDM dynamics often differ in important ways from the video motion. The most common problem occurs when the walking speed in the video differs from the training data. To accommodate this, we effectively blur the dynamics; this is achieved by raising the dynamics density to a small exponent, simply just using a smaller value of λ in (8), for which the kernel matrix must also be updated to include $\mathbf{X'}_t$. For tracking, we fix $\lambda = 0.5$.

Optimization: Tracking is performed by minimizing the approximate negative log posterior in (13). With the approximations above this becomes

$$\begin{aligned}\mathcal{E} \;=\; & -\sum_{j=t}^{t+\tau} \ln p(\mathbf{I}_j \mid \phi_j) \;-\; \sum_{j=t}^{t+\tau} \ln p(\mathbf{G}_j|\mathbf{G}_{j-1:j-2}^{MAP}) \\ & -\ln p(\mathbf{X'}_t|\bar{\alpha}, \mathbf{X}) \;-\; \sum_{j=t}^{t+\tau} \ln p(\mathbf{y}_j|\mathbf{x}_j, \bar{\beta}, \mathbf{X}, \mathbf{Y}) \quad (21)\end{aligned}$$

To minimize $\mathcal{E}$ in (21) with respect to $\phi_{t:t+\tau}$, we find that the following procedure helps to speed up convergence, and to reduce getting trapped in local minima. Each new state is first set to be the mean prediction, and then optimized in a temporal window. For the experiments we use $\tau = 2$.

Algorithm 1 Optimization Strategy (at each time step t)

$\{\mathbf{x}_{t+\tau}\} \leftarrow \mu_X(\mathbf{x}_{t+\tau-1}) = \mathbf{X}_{out}^T \mathbf{K}_X^{-1} \mathbf{k}_X(\mathbf{x}_{t+\tau-1})$
$\{\mathbf{y}_{t+\tau}\} \leftarrow \mu_Y(\mathbf{x}_{t+\tau}) = \mu + \mathbf{Y}^T \mathbf{K}_Y^{-1} \mathbf{k}_Y(\mathbf{x}_{t+\tau})$
$\{\mathbf{G}_{t+\tau}\} \leftarrow 2\mathbf{G}_{t+\tau-1} - \mathbf{G}_{t+\tau-2}$
for $n = 1 \ldots iter$ **do**
  $\{\mathbf{X'}_t\} \leftarrow \min \mathcal{E}$ with respect to $\mathbf{X'}_t$
  $\{\phi_{t:t+\tau}\} \leftarrow \min \mathcal{E}$ with respect to $\phi_{t:t+\tau}$
end for
$\{\mathbf{X'}_t\} \leftarrow \min \mathcal{E}$ with respect to $\mathbf{X'}_t$

One can also significantly speed up the minimization when one knows that the motion of the tracked object is very similar to the training motions. In that case, one can assume that there is negligible uncertainty in the reconstruction mapping, and hence a pose is directly given by $\mathbf{y} = \mu_Y(\mathbf{x})$. This reduces the pose reconstruction likelihood in (18) to $\frac{D}{2} \ln \sigma^2(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}\|^2$, and the state at t to $\phi_t = (\mathbf{G}_t, \mathbf{x}_t)$, which can be optimized straightforwardly.

5. Tracking Results

Here we focus on tracking different styles and speeds for the same activity. We use the Balanced GPDM model shown in Fig. 3 for tracking all walking sequences below. In Fig. 6 we use a well-known sequence to demonstrate the robustness of our algorithm to data loss. In the first frame, we supply nine 2D points—the head, left shoulder, left hand, both knees and feet, and center of the spine (the root). They are then tracked automatically using WSL[9]. As shown in Fig. 5(d) the tracked points are very noisy; the right knee is lost early in the sequence and the left knee is extremely inaccurate. By the end of the sequence the right foot and left hand are also lost. Given such poor input, our algorithm can nevertheless recover the correct 3D motion, as shown by the projections of the skeleton onto the original images.

While better image measurements can be obtained for this sequence, this is not always an option when there are occlusions and image clutter. E.g., Fig. 7 depicts a cluttered scene in which the subject becomes hidden by a shrub; only the head remains tracked by the end of the sequence (see Fig. 5(e)). For these frames only the global translation is effectively constrained by the image data, so the GPDM plays a critical role. In Fig. 7, note how the projected skeleton still appears to walk naturally behind the shrub.

Figure 8 shows a sequence in which the subject is completely occluded for a full gait cycle. When the occlusion begins, the tracking is governed mainly by the prior.[3] The 3D tracker is then switched back on and the global motion during the occlusion is refined by linear interpolation between the 3D tracked poses before and after the occlusion. Before an occlusion, it is very important to have a good estimation of $\mathbf{x}$, as subsequent predictions depend significantly

[3]We manually specify the beginning and end of the occlusion. We use a template matching 2D detector to automatically re-initialize WSL after the occlusion, as shown in Fig 5(c).

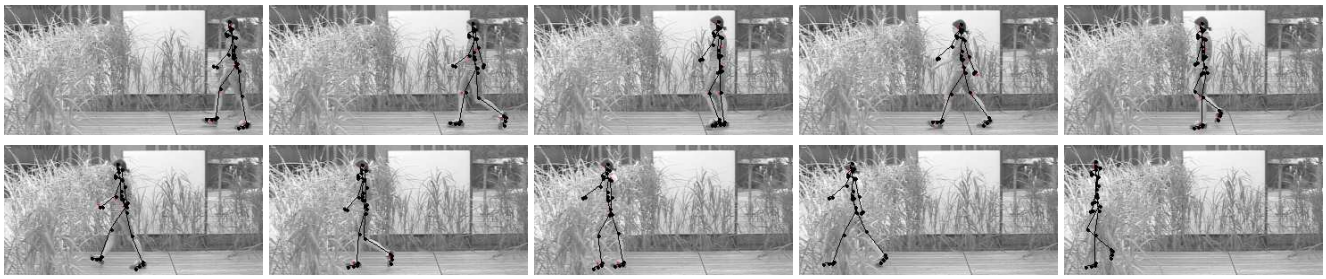

Figure 7. Tracking 56 frames of a walking motion with an almost total occlusion (just the head is visible) in a very clutter and moving background. Note that how the prior encourages realistic motion as occlusion becomes a problem.

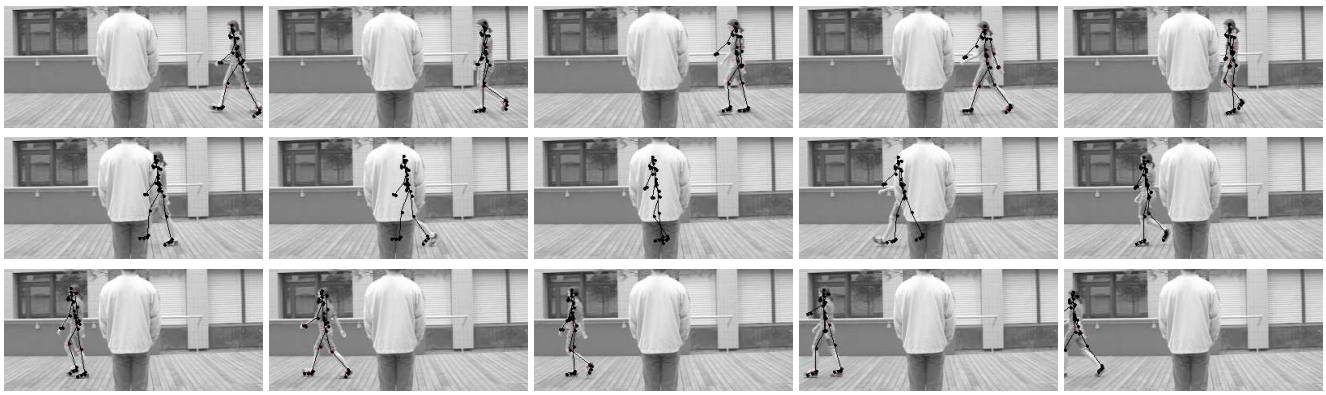

Figure 8. Tracking 72 frames of a walking motion with a total occlusion. During the occlusion the tracker is switch off and the mean prediction is used. Note the quality of the tracking before and after the occlusion and the plausible motion during it.

on the latent position. To reduce the computational cost of estimating the latent positions with great accuracy, we assume perfect reconstruction, i.e., $\mathbf{y} = \mu_Y(\mathbf{x})$, and use the second algorithm described in Section 4.

The latent coordinates obtained by the tracker for all of the above sequences are shown in Fig 10. The trajectories are smooth and reasonably close to the training data. Further, while the training gait period was 32 frames, this three sequences involve gait periods ranging from 22 to 40 frames (by comparison, natural walking gaits span about 1.5 octaves). Thus the prior generalizes well to different speeds.

To demonstrate the ability of the model to generalize to different walking styles, we also track the exaggerated walk shown in Fig. 9. Here, the subject's motion is exaggerated and stylistically unlike the training motions; this includes the stride length, the lack of bending of the limbs, and the rotation of the shoulders and hips. Despite this the 3D tracker does an excellent job. The last two rows of Fig. 9 show the inferred poses with a simple character, shown from two viewpoints, one of which is quite different from that of the camera. The latent coordinates obtained by the tracker are shown in Fig. 10; the distance of the trajectory to the training data is a result of the unusual walking style.

6. Conclusions

We have introduced the balanced GPDM for learning smooth prior models of human pose and motion for 3D people tracking. We showed that these priors can be learned from modest amounts of training motions including stylistic diversity. Further, they are shown to be effective for tracking a range of human walking styles, despite weak and noisy image measurements and significant occlusions. The quality of the results, in light of such a simple measurement model attest to the utility of the GPDM priors.

References

[1] Agarwal, A. and Triggs, B.: Recovering 3D human pose from monocular images. To appear *IEEE Trans. PAMI*, 2005.

[2] Bissacco, A.: Modeling and learning contact dynamics in human motion. *CVPR*, V1, pp. 421-428, San Diego, 2005.

[3] Choo, K., Fleet, D.: People tracking using hybrid Monte Carlo filtering. *Proc. ICCV*, V2, pp. 321-328, Vancouver, 2001.

[4] Doretto, G., Chiuso, A., Wu, Y.N., and Soatto, S.: Dynamic textures *IJCV*, 51(2):91-109, 2003.

[5] Elgammal, A., Lee, C.: Inferring 3D body pose from silhouettes using activity manifold learning. *Proc. CVPR*, V2, pp. 681-688, Washington, 2004.

[6] Grochow, K., Martin, S., Hertzmann, A., Popovic, Z.: Style-based inverse kinematics *SIGGRAPH*, pp. 522-531, 2004

[7] N. R. Howe, M. E. Leventon, and W. T. Freeman. Bayesian reconstructions of 3D human motion from single-camera video. *NIPS 12*, pp. 281-288, MIT Press, 2000.

[8] Isard, M. and Blake, A.: A mixed-state Condensation tracker with automatic model-switching. *Proc. ICCV*, pp. 107-112, Mumbai, 1998.

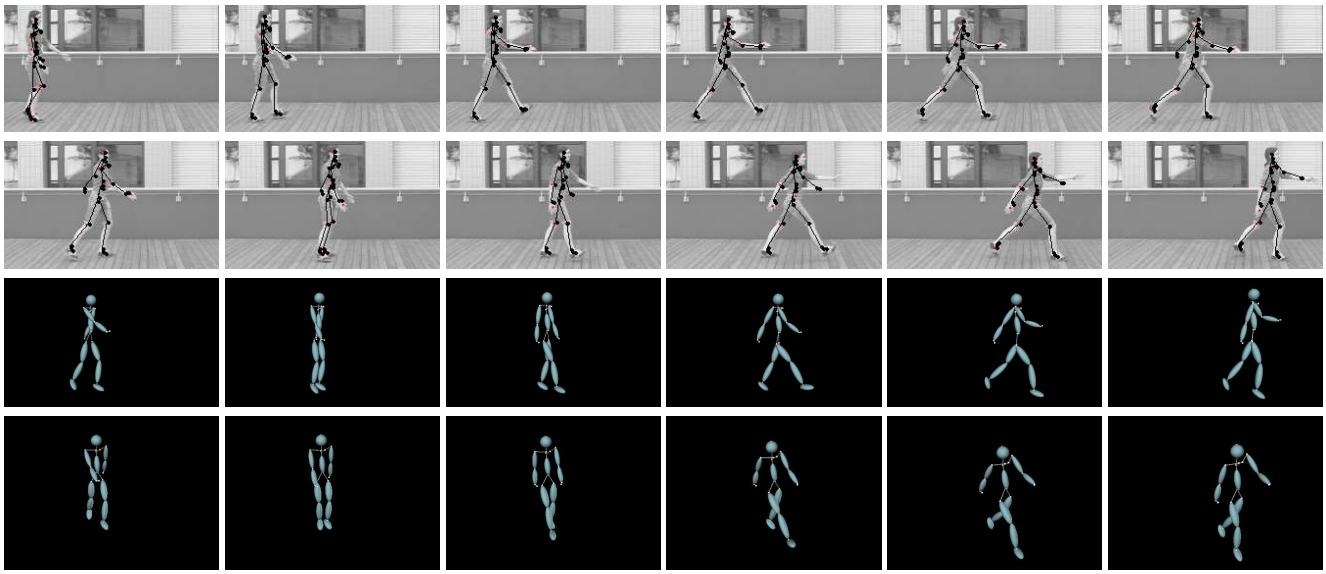

Figure 9. Tracking 37 frames of an exaggerated gait. Note that the results are very accurate even though the style is very different from any of the training motions. The last two rows depict two different views of the 3D inferred poses of the second row.

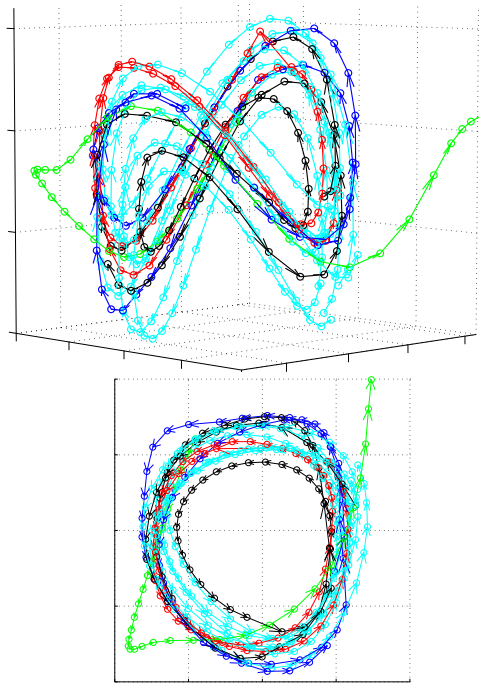

Figure 10. **Tracked Latent Positions:** Side and top views of the 3D latent space show the latent trajectories for the tracking results of Figs. 6, 7, 8, and 9 are shown in red, blue, black, and green. The learned model latent positions are cyan.

[9] Jepson, A.D., Fleet, D.J., El-Maraghi, T.: Robust on-line appearance models for vision tracking. *IEEE Trans. PAMI*, 25(10):1296-1311, 2003.

[10] Lawrence, N.D.: Gaussian process latent variable models for visualisation of high dimensional data. *NIPS 16*, pp. 329-336 MIT Press, 2004.

[11] Lee, J., Chai, J., Reitsma, P., Hodgins, J., Pollard, N.. Interactive control of avatars animated with human motion data. *Proc. SIGGRAPH*, pp. 491-500, 2002.

[12] MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003

[13] Neal R. M.: Bayesian Learning for Neural Networks. Lecture Notes in Statistics No. 118. Springer-Verlag, 1996.

[14] North, B., Blake, A., Isard, M., and Rittscher, J.: Learning and classification of complex dynamics. *IEEE Trans. PAMI*, 25(9):1016-1034, 2000.

[15] Pavolic, J.M., Rehg, J., and MacCormick, J.: Learning switching linear models of human motion. *NIPS 13*, pp. 981-987 MIT Press, 2000.

[16] Rahimi, A., Recht, B., Darrell, T. Learning appearance manifolds from video. *CVPR*, pp868-875, San Diego, 2005

[17] Sidenbladh, H., Black, M.J., Sigal, L.: Implicit probabilistic models of human motion for synthesis and tracking. *Proc. ECCV*, pp. 784-800, Copenhagen, 2002.

[18] Sminchisescu, C., Jepson, A.: Generative modeling for continuous non-linearly embedded visual inference. *Proc. ICML*, Banff, July 2004.

[19] Tian, T., Li, R., Sclaroff, S.: Articulated pose estimation in a learned smooth space of feasible solutions. *CVPR Learning Workshop*, San Diego, 2005

[20] Urtasun, R., Fleet, D.J., Hertzmann, A., Fua, P.: Priors for people tracking from small training sets. *Proc. ICCV*, V1, pp. 403-410, Beijing, 2005.

[21] Wang, Q., Xu, G., Ai, H.: Learning object intrinsic structure for robust visual tracking. *Proc. CVPR*, Vol. 2 pp. 227-233, Madison, 2003.

[22] Wang, J., Fleet, D.J., Hertzmann, A.: Gaussian Process dynamical models. *NIPS 18*, MIT Press, 2005.

[23] Weiss, Y.: Interpreting image by propagating Bayesian beliefs. *NIPS 9*, pp. 908-915, MIT Press, 1997.