Learning and Tracking Human Motion Using Functional Analysis

D. Ormoneit*, H. Sidenbladh†, M. J. Black‡, T. Hastie*, D. J. Fleet‡

*Dept. of Statistics, Stanford University, Stanford, CA 94305
{ormoneit,hastie}@stat.stanford.edu, http://www-stat.stanford.edu/

†Royal Institute of Technology (KTH), S–100 44 Stockholm, Sweden
hedvig@nada.kth.se, http://www.nada.kth.se/~hedvig/

‡Xerox Palo Alto Research Center, Palo Alto, CA 94304
black@parc.xerox.com, http://www.parc.xerox.com/black/

Abstract

We present a method for the modeling and tracking of human motion using a sequence of 2D video images. Our analysis is divided in two parts: statistical learning and Bayesian tracking. First, we estimate a statistical model of typical activities from a large set of 3D human motion data. For this purpose, the human body is represented as a set of articulated cylinders and the evolution of a particular joint angle is described by a time-series. Specifically, we consider periodic motion such as "walking" in this work, and we develop a new set of tools that allows for the automatic segmentation of the training data into a sequence of identical "motion cycles". Then we compute the mean and the principal components of these cycles using a new algorithm to account for missing information and to enforce smooth transitions between different cycles. The learned temporal model provides a prior probability distribution over human motions which is used for tracking. We adopt a Bayesian perspective and approximate the posterior distribution of the body parameters using a particle filter. The resulting algorithm is able to track human subjects in monocular video sequences and to recover their 3D motion in complex unknown environments.

1 Introduction

The modeling and tracking of human motion in video is important for problems as varied as animation, video database search, sports medicine, and human-computer interaction. Technically, the human body can be approximated by a collection of articulated limbs (Figure 1) and its motion can be thought of as a collection of time-series describing the joint angles as they evolve over time. A key difficulty for the modeling of these body angles is that each time-series has to be decomposed into suitable temporal primitives prior to statistical analysis. For example, in the case of repetitive human motion such as walking, motion sequences decompose naturally into a sequence of identical "motion cycles". Of course, the exact nature of this decomposition is unknown to the modeler and needs to be estimated from the motion data. In this work, we present a new set of tools that carry out this identification automatically. In detail, we suggest an iterative procedure that generates the best segmentation with respect to the signal-to-noise ratio of the data in an aligned reference domain. This procedure allows us to use the mean and the principal components of the individual cycles in the reference domain as a statistical model. Technical difficulties include missing information in the motion time-series and the necessity of enforcing smooth transitions between different cycles. To deal with these problems, we develop a new iterative method for functional Principal Component Analysis (PCA) that is based on a truncated Fourier transform.

The learned temporal model provides a prior probability distribution over human motions which can be used in a Bayesian framework for tracking. For this purpose, we specify a generative model of image appearance and the likelihood of observing image data given the model. The non-linearity of this generative model results in a posterior distribution that cannot be represented in closed form. Hence, the posterior is represented using a discrete set of samples and is propagated over time using particle filtering. Here the prior distribution based on the PCA serves to improve the efficiency of the particle filter by constraining the samples to the most likely regions of a low-dimensional subspace of the parameter space. The resulting algorithm is able to track human subjects in monocular video sequences and to recover their 3D motion under changes in their pose and against complex unknown backgrounds.

The Bayesian tracking is described in detail in [18] and is summarized here. Unlike that previous work which used hand-segmented and aligned training data, this paper de-

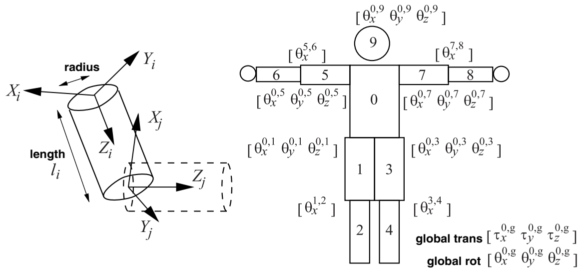

Figure 1: Human body model, consisting of a set of articulated cylinders with 25 degrees of freedom (DOF). Each limb, i, has a local coordinate system with the Z_i axis directed along the limb. Joints have up to 3 angular DOF, expressed as relative rotations $(\theta_x^{j,i}, \theta_y^{j,i}, \theta_z^{j,i})$ between body parts i and j.

tails an automated method for learning temporal models of cyclic data which form a limited but important class of human motions. The learning methods developed here may be extensible from cyclic to more general human motions.

2 Related Work

The modeling of human motion has attracted a great deal of attention in both the computer graphics and computer vision communities. Much of the work in the computer vision community has focused on recognition of activities and models of human motion appropriate to this task. For example, Hidden Markov Models (HMM's) have been used extensively for recognition of human action (e.g. [3, 4, 20]). HMM's typically provide a coarse approximation to motion data and hence are not as good for synthesis as for recognition (though Brand's recent work [3] suggests that may not be a fundamental limitation).

The weakness of HMM's for modeling is that they do not well capture some of the intrinsic properties of biological motion such as smoothness. The same can be said for linear dynamical models [13]. Instead, human motions are often represented by explicit temporal curves that describe the change over time of 3D joint angles [5, 6, 12, 14, 15].

Here we focus on the modeling of 3D joint angles for tracking of human motion. The simplest models place constraints on the smooth change in joint angles [10, 19]. More detailed models represent explicit curves corresponding to the variation in joint angles over time. These models are derived from biometric studies of human motion [5, 15] or learned from 3D motion-capture data [6, 12, 18].

Statistical representations of time-series data using functional analysis are described in detail in [14]. A common approach to model 3D motion curves is to hand-segment and align the data corresponding to particular activities. The variation across subjects is modeled by principal component analysis (PCA) of the curve data. Here the first few principal components capture most of the variation in the training set. This approach has been used for representing 2D image changes [2], optical flow [21], and 3D joint angles [18]. A related approach uses vector quantization [12] rather than PCA.

The primary use of such detailed models is in tracking using 3D articulated models of people. Given the high dimensionality of the human body, the temporal curves are used to constrain the possible motions to lie on a far lower dimensional manifold. Recently, Bayesian methods have become popular for tracking 3D human motion [7, 12, 17, 18]. In these models, the temporal curves can be thought of as providing a prior probability distribution over valid human motions. The PCA representation of temporal curves provides a statistical model of the variation present in the training set that can be used to construct such a probabilistic prior.

Most work on modeling 3D human motion curves has focused on cyclic motions such as walking and running. While cyclic motions are particularly simple they are also an important class of human activities and have been extensively studied [1, 16]. Likewise, in the current paper, we focus on cyclic motion and provide a thorough statistical treatment.

Our ultimate goal in modeling 3D human motion is to automatically learn probabilistic models from training data. To do so involves automatically segmenting the data into individual activities, aligning activities from different examples, modeling the statistical variation in the data, and deriving a probabilistic model. Complicating matters is the fact that training data tends to be imperfect and, with commercial motion capture systems, contains missing data that must be accounted for in modeling. Additionally, with cyclic motions, the learning method must enforce smooth transitions between cycles. In previous work, some of these issues were either ignored or dealt with via manual intervention. Here, in the case of cyclic motions, we provide a complete treatment that automatically copes with missing data and enforces smoothness.

3 Learning

In the first part of our analysis, we develop a modeling procedure for periodic motion sequences. By definition, periodic motion is composed of repetitive "cycles" which constitute a natural unit of statistical modeling and which must be identified in the training data prior to building a model. Frequently, this segmentation is carried out manually in an error-prone and burdensome procedure (see, for example, [14, 21]). In this section, we present alignment algorithms that segment the data automatically. Based on the estimated alignment parameters, the cycles of different motion sequences are then transformed into a common reference domain, and the mean and the principal components of the transformed cycle data are computed as a statistical model. Here the mean cycle can be interpreted as a prototype of a specific motion class, e.g. walking, and the principal components characterize the main sources of deviation of sequences in the data set from the mean cycle. Below we use these statistics to construct a prior distribution for Bayesian tracking.

3.1 The Motion Data

Training data, in the form of 3D joint angles, is provided by a commercial motion-capture system. For each "motion sequence", there are 19 such angle time-series in our case, and we use the term "motion class" to indicate the type of motion rendered by the subject during the observation (walking, running, etc.). Altogether our data set consists of eight motion sequences rendered by four individuals. The length of the motion sequences ranges from about 500 to 5000 frames.

Formally, we let T_i denote the length of the i-th motion sequence and we use $t = 1, \ldots, T_i$ as a time index. Similarly, $m = 19$ is the number of angles in each motion sequence and $a = 1, \ldots, m$ indicates a particular angle. The i-th motion sequence is written formally as[1]

$$Z_i(t) \equiv \{z_{a,i}(t) | a = 1, \ldots, m\} \text{ for } t = 1, \ldots, T_i.$$

There are $n = 20$ motion sequences in our training data set and associated with each sequence we have the indicator set $I_{a,i} \equiv \{t \in \{1, \ldots, T_i\} \,|\, z_{a,i}(t) \text{ is not missing}\}$. Missing observations occur frequently in our data set because some markers may be occluded during parts of a motion. The capturing system reports an angle of zero for some of the position coordinates in this case. Typically occlusion lasts for several frames which prevents the imputation of interpolated values using neighboring observations. Below we spend considerable effort to design our algorithms in a manner insensitive to this artifact.

3.2 Sequence Alignment

First, we describe a procedure to estimate alignment parameters that segment motion sequences into cycles. In detail, for each motion sequence, we estimate its "cycle length", p, and an "offset parameter", o. Based on these parameters, the individual motion sequences can be transformed into a common "reference domain" for further analysis.

To estimate the cycle length, we simply try a large number of candidate values for p and we assess the quality of the alignment resulting from p using a simple score function. Formally, let the projection index associated with p be defined according to $\xi_p(t) \equiv \lceil t/p \rceil$, where $\lceil a \rceil$ denotes the smallest integer greater or equal a. In other words, ξ_p "folds" the original sequence into the domain $\{1, \ldots, p\}$. Also, let $I_{i,a}(k)$ denote the index-set of non-missing values projected onto k. Then the mean of the observations mapped onto k can be written as

$$\bar{z}_{i,a}(k) \equiv \frac{1}{|I_{i,a}(k)|} \sum_{j \in I_{i,a}(k)} z_{i,a}(j),$$

and the magnitudes

$$\begin{aligned} noise_{i,a}(p) &\equiv \frac{\sum_{j \in I_{i,a}} (z_{i,a}(j) - \bar{z}_{i,a}(\xi_p(j)))^2}{|I_{i,a}| - p} \\ signal_{i,a}(p) &\equiv \frac{\sum_{k=1}^{p} (\bar{z}_{i,a}(k) - \sum_{j \in I_a} z_{i,a}(j)/|I_{i,a}|)^2}{(p-1)p/|I_{i,a}|} \end{aligned}$$

measure the signal- and the noise-content of the projected sequence. Combined into a single value, we can define the "signal-to-noise ratio"

$$stn\_ratio_i(p) \equiv \sum_a \frac{signal_{i,a}(p)}{noise_{i,a}(p)}. \tag{1}$$

Specifically, $noise_{i,a}$ can be interpreted as the variation in the data that is not explained by the mean cycle; $signal_{i,a}$ measures the signal intensity.[2] Therefore, it is natural to prefer values of p producing a large signal-to-noise ratio. In our algorithm we try candidate values from the set $\{50, 51, \ldots, 250\}$ and choose the maximum with respect to (1) as our estimate of the cycle-length.

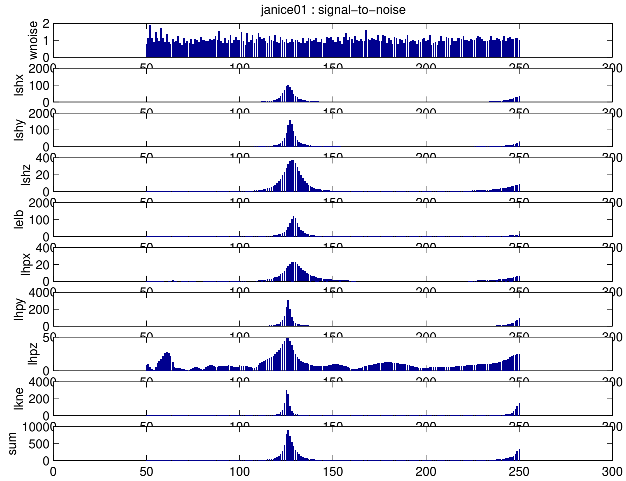

Figure 2: Signal-to-noise ratio of a representative set of angles as a function of the candidate period length. We show left shoulder (lshx,lshy,lshz), elbow (lelb), hip (lhpx,lhpy,lhpz), and knee angles (lkne). The top series corresponds to a white noise signal. The bottom series shows the accumulated signal-to-noise ratio (1). The minimum overall variation was obtained for $p = 126$, ($stn\_ratio = 8.986505e+02$).

[1]The data Z_i is not to be confused with the reference coordinate Z_j for limb j in Figure 1.

[2]Note that both of these magnitudes are normalized so as to produce unbiased estimates.

In Figure 2 we show the individual signal-to-noise ratios for a subset of the angles as well as the accumulated signal-to-noise ratio (1) as functions of p. Note the sharp peak of these values around the optimal cycle length $p = 126$. Note also that the signal-to-noise ratio of the artificially generated white noise series in the first row is approximately constant, warranting the unbiasedness of our approach with respect to changing values of p.

The described folding procedure computes an estimate of the optimal cycle length $p(i)$ for each sequence and stores these values in an array of length n. In our second step, we use this array to align multiple sequences in a common domain by rescaling. In detail, we construct offset estimates $o(1), o(2), \ldots, o(n)$ so that the shifted motion sequences minimize the deviation from a common prototype model by analogy to the noise-criterion of the previous paragraph. An exhaustive search for the optimal offset combination is clearly infeasible due to its high computational complexity of $O\left(\prod_{i=1}^{n} p(i)\right)$ Instead, we suggest the iterative procedure illustrated in Figure 3 to compute an approximate solution: We initialize the offset values to zero in Step 1,

1. Initialize offset values. For $i = 1, \ldots, n$, let $o(i) := 0$.

2. From a given function class $\mathcal{R}$, choose the minimum least-squares fit with respect to the aligned data. For $a = 1, \ldots, m$:

$$r_a := \arg\min_{r \in \mathcal{R}} \sum_{i=1}^{n} \sum_{j \in I_{i,a}} \left[z_{i,a}(j) - r\left(\frac{j - o(i)}{p(i)}\right)\right]^2 .$$

3. Update the offset parameters. For $i = 1, \ldots, n$:

$$o(i) := \arg\min_{o} \sum_{a} \sum_{j \in I_{i,a}} \left[z_{i,a}(j) - r_a\left(\frac{j - o}{p(i)}\right)\right]^2 .$$

4. Stop, if the performance improvement is below 10^{-6}. Otherwise, goto Step 2.

Figure 3: Iterative algorithm for the computation of the optimal offset parameters.

and we define a *reference signal* r_a in Step 2 so as to minimize the deviation with respect to the aligned data. Next, we choose the offsets of all sequences so that they minimize the prediction error with respect to the reference signal (Step 3). By contrast to the exhaustive search, this operation requires $O\left(\sum_{i=1}^{n} p(i)\right)$ comparisons only. Because the solution of the first iteration may well be suboptimal, we construct an improved reference signal using the current offset estimates, and use this signal in turn to improve the offset estimates. Repeating these steps, we obtain an iterative optimization algorithm that is terminated if the improvement falls below a given threshold (Step 4). Because Steps 2 and 3 both decrease the prediction error, it is clear that the algorithm converges.

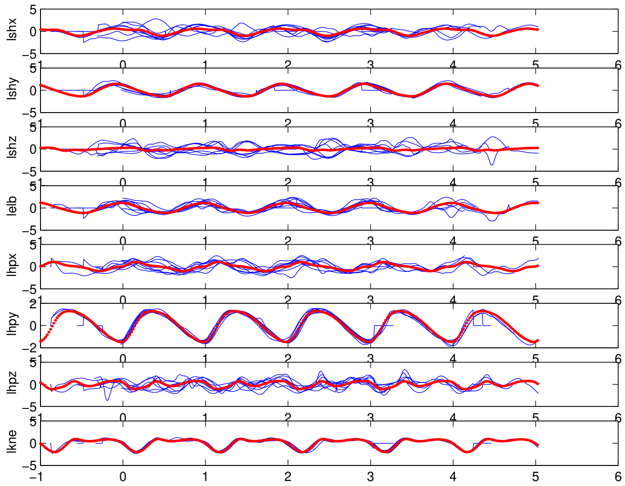

Figure 4: Aligned representation of eight walking sequences (blue). The red curve denotes repetitions of the periodic regression spline estimate. (final error: 6.2536e+04, convergence after five steps).

Figure 4 shows eight sequences of a walking motion, aligned using this procedure. As a function class for the reference signal, $\mathcal{R}$, we chose periodically constrained regression splines in our implementation; i.e. the zeroth, first, and second derivatives of the spline are designed so as to coincide at the boundaries (for details on splines, see [22]). Otherwise, the concatenated reference signal in Figure 4 would be discontinuous at the transitions between cycles.

3.3 Principal Component Analysis

Next, we break down the aligned data representation of Figure 4 into individual cycles and we compute the principal components of the resulting slices. The principal components can be interpreted as the major sources of variation in the data. Below we use them to formulate a probability model for walking. The algorithm developed for this purpose is illustrated in Figure 5.

Our algorithm addresses several difficulties: First, even though the individual motion sequences are aligned in Figure (4), they are still sampled at different frequencies in the common domain due to the different alignment parameters. This problem is accommodated in Step 1c by resampling after computing a functional estimate in continuous time in Step 1b. As function estimates, $\mathcal{S}$, we choose (non-constrained) regression splines in this case. A second difficulty consists of missing data in the design matrix X. Therefore, we cannot simply use the Singular Value Decomposition (SVD) of $X^{(1)}$ to obtain the principal com-

1. For $a = 1, \ldots, m$ and $i = 1, \ldots, n$:
   (a) Dissect $z_{i,a}$ into K_i cycles of length $p(i)$, marking missing values at both ends. This gives a new set of time series $z_{k,a}^{(1)}$ for $k = 1, \ldots, K_i$ where $K_i = \lceil \frac{T_i - o(i)}{p(i)} \rceil + 1$. Let $\bar{I}_{k,a}$ be the new index set for this series.
   (b) Compete functional estimates in the domain $[0, 1]$. For $k = 1, \ldots, K_i$:
   $$f_{k,a} := \arg\min_{f \in \mathcal{S}} \sum_{j \in \bar{I}_{k,a}} \left[z_{k,a}^{(1)}(j) - f\left(\frac{j}{p(i)}\right) \right]^2 .$$
   (c) Resample the data in a common reference domain, imputing missing observations. This gives yet another time-series
   $$z_{k,a}^{(2)}(j) := f_{k,a}\left(\frac{j}{\mathcal{T}}\right) \text{ for } j = 0, 1, \ldots, \mathcal{T}.$$
2. Stack the "slices" $z_{k,a}^{(2)}$ obtained from all sequences row-wise into a $\sum_i K_i \times m\mathcal{T}$ design matrix X.
3. Compute the row-mean μ of X, and let $X^{(1)} := X - 1'\mu$. Here 1 is a vector of ones.
4. Slice by slice, compute the Fourier coefficients of $X^{(1)}$, and store them in a new matrix, $X^{(2)}$. Use the first 20 coefficients only.
5. Compute the Singular Value Decomposition of $X^{(2)}$:
   $$X^{(2)} = USV'.$$
6. Reconstruct the design matrix, using the rank q approximation to S:
   $$X^{(3)} = US^q V'. \quad (2)$$
7. Apply the Inverse Fourier Transform and add $1'\mu$ to obtain $X^{(4)}$.
8. Impute the missing values in X using the corresponding values in $X^{(4)}$.
9. Evaluate $||X - X^{(4)}||$. Stop, if the performance improvement is below 10^{-6}. Otherwise, goto Step 3.

Figure 5: Functional PCA algorithm with data imputation.

ponents. An appropriate alternative is to use an iterative approximation scheme suggested recently by Hastie et al. in the context of analyzing Gene Expression Arrays [23]. In detail, we alternate between an SVD step (4 through 7) and a data imputation step (8), where each update is designed so as to decrease the matrix distance between X and its reconstruction, $X^{(4)}$. As an additional complication, we cannot compute the SVD directly because the principal components obtained in this manner could be non-periodic or even discontinuous. This is due to the data imputations in Steps 1c and 8 which do not explicitly enforce these constraints. We accommodate this problem by translating the cycles into a Fourier domain and by truncating high-frequency coefficients (Step 4). Then we compute the SVD in the Fourier domain in Step 5, and we reconstruct the design matrix using a rank-q approximation in Steps 6 and 7. In Step 8 we use the reconstructed values as improved estimates for the missing data in X, and then we repeat Steps 4 through 7 using these improved estimates. This iterative process is continued until the performance improvement falls below a given threshold. The convergence of the algorithm can be proven using an argument similar to the one in Section 3.2; for brevity, we do not provide details here.

The above algorithm generates the imputed design matrix, X, as well as its singular value decomposition as its output. These serve as a prior distribution for the tracking algorithm below. Specifically, we use $q = 5$ in our experiments and we assume that all motions are essentially superpositions of these five components in the next section.

4 Bayesian Tracking

In tracking, our goal is to estimate joint angles of the body and its 3D pose given a sequence of image measurements, $\vec{\mathbf{I}}_t$, up to time t. Below, we outline a Bayesian framework in which we define a generative model of image appearance to obtain a likelihood term that specifies the probability of observing an image given the parameters of the body model. Also, we use the Singular Value Decomposition of $X^{(2)}$ to formulate a Bayesian prior distribution. In this framework, tracking can be reformulated as an inference problem where we compute a posterior distribution given the prior and the likelihood. This computation is typically very difficult given the high dimensionality of the body model. However, by approximating motion in terms of a few principal components, we effectively reduce the dimensionality and arrive at a computationally feasible algorithm.

Formally, let $\theta(t) \equiv (\theta_a(t) | a = 1, \ldots, m)$ be a random vector summarizing the relative joint angles at time t; in other words, the value of a motion sequence, $Z_i(t)$, at time t is now interpreted as the i-th realization of $\theta(t)$. Under the modeling assumptions of the SVD in Figure 5, $\theta(t)$ can be written in the form

$$\theta(t) = \tilde{\mu}(\psi_t) + \sum_{k=1}^{q} c_{t,k} v_k(\psi_t), \quad (3)$$

where v_k is the Fourier inverse of the k-th column of V, rearranged as an $\mathcal{T} \times m$-matrix; similarly, $\tilde{\mu}$ denotes the rearranged mean vector μ. $v_k(\psi)$ is the ψ-th column of v_k, and the $c_{t,k}$ are time-varying coefficients. $\psi_t \in \{0, \mathcal{T}-1\}$ maps absolute time onto relative cycle positions or phases, and ρ_t denotes the speed of the motion such that $\psi_{t+1} = (\psi_t + \rho_t) \bmod \mathcal{T}$.

Given this representation (3), body positions are characterized by the low-dimensional state-vector $\boldsymbol{\phi}_t = (\mathbf{c}_t, \psi_t, \rho_t, \boldsymbol{\tau}_t^g, \boldsymbol{\theta}_t^g)'$, where $\mathbf{c}_t = (c_{t,1}, \ldots, c_{t,q})$ is a vector of the q linear coefficients and where $\boldsymbol{\tau}_t^g$ and $\boldsymbol{\theta}_t^g$ represent the global 3D translation and rotation of the torso.

The tracking of a person in a monocular video sequence entails estimating a distribution over $\boldsymbol{\phi}_t$ at each time t. We adopt a Bayesian perspective in which the posterior probability over the parameters $\boldsymbol{\phi}_t$ given all observations, $\vec{\mathbf{I}}_t$, up to time t can be updated recursively according to:

$$p(\boldsymbol{\phi}_t \mid \vec{\mathbf{I}}_t) \propto \qquad (4)$$
$$p(\mathbf{I}_t \mid \boldsymbol{\phi}_t) \int p(\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1}) p(\boldsymbol{\phi}_{t-1} \mid \vec{\mathbf{I}}_{t-1}) \, d\boldsymbol{\phi}_{t-1} \,.$$

Here $p(\mathbf{I}_t \mid \boldsymbol{\phi}_t)$ is the likelihood of observing the image $\mathbf{I}_t$ given the parameters and $p(\boldsymbol{\phi}_{t-1} \mid \vec{\mathbf{I}}_{t-1})$ is the posterior probability from the previous instant. $p(\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1})$ is a temporal prior probability distribution that encodes how the parameters $\boldsymbol{\phi}_t$ change over time. The elements of the Bayesian approach are summarized below; for details the reader is referred to [18].

4.1 Generative Image Model

The geometrical optics are modeled as a pinhole camera and we define a mapping from 3D scene coordinates to a 3D camera-centered coordinate system. The body is modeled as a kinematic tree of articulated cylinders with the body as the root (see [18] for details). The global translation and rotation $(\boldsymbol{\tau}_t^g, \boldsymbol{\theta}_t^g)$ map the torso into scene coordinates. Rigid transformations specify the relative positions and orientations between connected limbs.

Given specific values for the parameter vector $\boldsymbol{\phi}_t$, the values $\mathbf{c}_t$ and ψ_t define a set of relative joint angles as specified by Equation (3). Combining these joint angles with the global translation and rotation, $\boldsymbol{\tau}_t^g$, $\boldsymbol{\theta}_t^g$, defines the configuration of the body at time t. The camera model then specifies how this 3D model is projected into the image.

We must now specify how this geometric formulation can be used to predict the image appearance at time $t+1$. Let $M(\mathbf{I}_t, \boldsymbol{\phi}_t)$ be a function that takes image texture at time t and, given the model parameters, maps it onto the surfaces of the 3D model. Similarly, let $M^{-1}(\cdot)$ take a 3D model and project its texture back into the image.

Given these functions, the generative model of images at time $t+1$ can be viewed as a mapping from the image at time t to the model using the parameters at time t and then the projection of this model into the image using the parameters at time $t+1$:

$$\mathbf{I}_{t+1} = M^{-1}(M(\mathbf{I}_t, \boldsymbol{\phi}_t), \boldsymbol{\phi}_{t+1}) + \eta, \quad \eta \sim G(0, \sigma),$$

where $G(x, \sigma)$ denotes a zero mean Gaussian distribution where the standard deviation, σ depends on the viewing angle of the limb with respect to the camera and increases as the limb is viewed more obliquely (see [18] for more details).

4.2 Temporal Prior

The temporal prior, $p(\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1})$, models how the parameters describing the body configuration are expected to vary over time. It is expressed formally as a collection of distributions of the individual components of $\boldsymbol{\phi}$:

$$p(\mathbf{c}_t|\mathbf{c}_{t-1}) = G(\mathbf{c}_t - \mathbf{c}_{t-1}, \boldsymbol{\sigma}^c) \quad (5)$$
$$p(\psi_t|\psi_{t-1}) = G(\psi_t - \psi_{t-1}, \sigma^\psi) \quad (6)$$
$$p(\rho_t|\rho_{t-1}) = G(\rho_t - \rho_{t-1}, \sigma^\rho) \quad (7)$$
$$p(\boldsymbol{\tau}_t^g|\mathbf{T}_{t-1}, \boldsymbol{\tau}_{t-1}^g, \rho_{t-1}) = G([\boldsymbol{\tau}_{t-1}^g, 1]' - \mathbf{T}_{t-1}^{-1}[\rho_{t-1}\ 0\ 0\ 1]', \boldsymbol{\sigma}^{\boldsymbol{\tau}}) \quad (8)$$
$$p(\boldsymbol{\theta}_t^g|\boldsymbol{\theta}_{t-1}^g) = G(\boldsymbol{\theta}_t^g - \boldsymbol{\theta}_{t-1}^g, \boldsymbol{\sigma}^{\boldsymbol{\theta}}) \quad (9)$$

where σ^ψ, σ^ρ, $\boldsymbol{\sigma}^{\boldsymbol{\tau}}$ and $\boldsymbol{\sigma}^{\boldsymbol{\theta}}$ are empirically determined standard deviations while $\boldsymbol{\sigma}^c = \varepsilon \boldsymbol{\lambda}$ where ε is a small scalar and $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_q]'$ are the singular values in S^q in Equation (2). ε is expected to be small since $\mathbf{c}$ varies little throughout the walking cycle for each individual. Finally, $\mathbf{T}_{t-1}$ is a homogeneous transformation matrix containing the the global body parameters, $\boldsymbol{\tau}_t^g$ and $\boldsymbol{\theta}_t^g$.

The Gaussian distribution over ψ and $\mathbf{c}$ implies a Gaussian distribution over joint angles. Thus, samples from the distribution $p(\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1})$ correspond to possible body configurations. This provides a lower-dimensional model for the distribution over the relative joint angles of the body.

4.3 Likelihood Model

Finally, to complete the Bayesian formulation, we define the likelihood, $p(\mathbf{I}_t|\boldsymbol{\phi}_t)$, of observing image $\mathbf{I}_t$ given the human model has configuration $\boldsymbol{\phi}_t$ at time t. Based on the generative model we predict the image at time t and compare the generated image with the observed image $\mathbf{I}_t$. In detail, we independently evaluate the generative model for each limb and compute the likelihood of observing the image I_t as the product of the resulting likelihood terms. For a given limb j, we take n_s pixel locations, $\mathbf{x}_{j,i}$, uniformly sampled from the projected limb region and compute:

$$E_j \equiv \sum_{i=1}^{n} (\mathbf{I}_t(\mathbf{x}_{j,i}) - M^{-1}(M(\mathbf{I}_{t-1}, \boldsymbol{\phi}_{t-1}), \boldsymbol{\phi}_t)(\mathbf{x}_{j,i}))^2 . \quad (10)$$

Then we define the (limb-specific) likelihood of observing the image as

$$p_j^{image} = \frac{1}{\sqrt{2\pi}\sigma(\alpha_j)} \exp(-E_j/(2\sigma(\alpha_j)^2 n_s)), \quad (11)$$

where α_j is the angle between the limb j principal axis and the image plane of the camera and $\sigma(\alpha_j)$ is a function that increases with narrow viewing angles. While this simple model works well in practice, the formulation of robust likelihood models remains an area of our current research.

When a limb is completely occluded, $\sigma(\alpha_j)$ is large and the likelihood will be low. Similarly, as the limb is viewed at narrow angles (all visible surface normals are roughly perpendicular to the viewing direction) the true texture pattern may be highly distorted. The limb can be thought of as occluded and the probability of the viewing it goes to zero. To model occluded regions we introduce the constant probability, $p_j^{occluded}$, that a limb is occluded.

We express the likelihood as a mixture between p_j^{image} and the likelihood of occlusion, $p_j^{occluded}$, which acts as a "penalty term." The visibility q, i.e. the influence of the actual image measurement, decreases with the increase of the angle α_j between the limb j principal axis and the image plane. The likelihood for the image likelihood of limb j is defined as:

$$p_j = q(\alpha_j) p_{image} + (1 - q(\alpha_j)) p_{occluded} \quad (12)$$

where $q(\alpha_j) = cos(\alpha_j)$ if limb j is non-occluded, or 0 if limb j is occluded. The likelihood of observing the image given a particular body pose is given by the product:

$$p(\mathbf{I}_t | \boldsymbol{\phi}_t) = \prod_j p_j \,. \quad (13)$$

4.4 Stochastic Optimization

The posterior distribution may well be multi-modal due to the nonlinearity of the likelihood function which results from self-occlusions, viewpoint singularities, and matching ambiguities. Representation of the posterior is further complicated by the use of a (moderately) high-dimensional dynamical model of the state evolution as embodied by the temporal prior. For these reasons we represent the posterior as a weighted set of state samples, which are propagated in time using a particle filtering approach. Here we briefly describe the method (see [11, 9, 18] for details.).

A state, $\mathbf{s}_t$, is represented by a vector of parameters assignments, $\mathbf{s}_t = [\boldsymbol{\phi}_t^s]$. The posterior at time t is represented by N_s samples ($N_s \approx 10^4$ in our experiments). To compute the posterior (5) at time t we first draw N_s samples from the posterior at time $t-1$. Similarly, the shape parameters are propagated by sampling from $p(\boldsymbol{\phi}_t | \boldsymbol{\phi}_{t-1})$. At this point we have new values of $\boldsymbol{\phi}_t$ which can be used to compute the likelihood $p(\mathbf{I}_t | \boldsymbol{\phi}_t)$. The N likelihoods are normalized to sum to one and the resulting set of samples approximates the posterior distribution $p(\boldsymbol{\phi}_t, |\vec{\mathbf{I}}_t)$ at time t.

5 Experiments

To illustrate the method we show an example of tracking a walking person in a cluttered scene. On an Ultra 1 Sparcstation the C++ implementation ran at a rate of approximately 1 frames/minute. To visualize the posterior distribution we display the projection of the 3D model corresponding to the expected value of the model parameters: $\frac{1}{N_s} \sum_{i=1}^{N_s} p_i \boldsymbol{\phi}_i^s$ where p_i is the likelihood of sample $\boldsymbol{\phi}_i^s$. All parameters were initialized with a Gaussian prior at time $t = 0$.

Figure 6 shows the tracking results for frames 0 to 50 of a sequence showing a walking person. Note that the legs of the model are better aligned with the image data than the arms. This is probably due to the fact that the arms are more often occluded by the torso, and thus more prior driven than the legs. In parts of the cycle where large occlusion occurs (frame 30) the model has little image information, and starts to drift off the person. However, it recovers when a larger part of the body is visible (frame 40).

6 Conclusions

This paper describes a fully automated method for learning periodic human motions from training data. Statistical methods are presented for detecting the length of the periods in the data, segmenting it into cycles, and optimally aligning the cycles. We also presented a novel principal component analysis technique for building a statisical eigenmodel of the motion curves. The method copes with missing data and enforces smoothness between the beginning and ending of a motion cycle. The learned eigencurves are used as prior probability distributions in a Bayesian tracking framework. Tracking in monocular image sequences is performed using a particle filtering techinque and we have demonstrated results for tracking a person in a cluttered image sequence.

Acknowledgements. We are grateful to Michael Gleicher for generously providing the 3D motion-capture data used in our experiments. We thank Manolis Kamvysselis for discussions about human motion.

References

[1] M. Allmen and C.R. Dyer. Cyclic motion detection using spatiotemporal surfaces and curves. *ICPR*, pp. 365–370, 1990.

[2] A. Bobick and J. Davis. An appearance-based representation of action. *ICPR*, 1996.

[3] M. Brand. Shadow puppetry. *ICCV*, pp. 1237–1244, 1999.

[4] C. Bregler. Learning and recognizing human dynamics in video sequences. *CVPR*, pp. 568–574, 1997.

[5] C. Bregler and J. Malik. Tracking people with twists and exponential maps. *CVPR*, 1998.

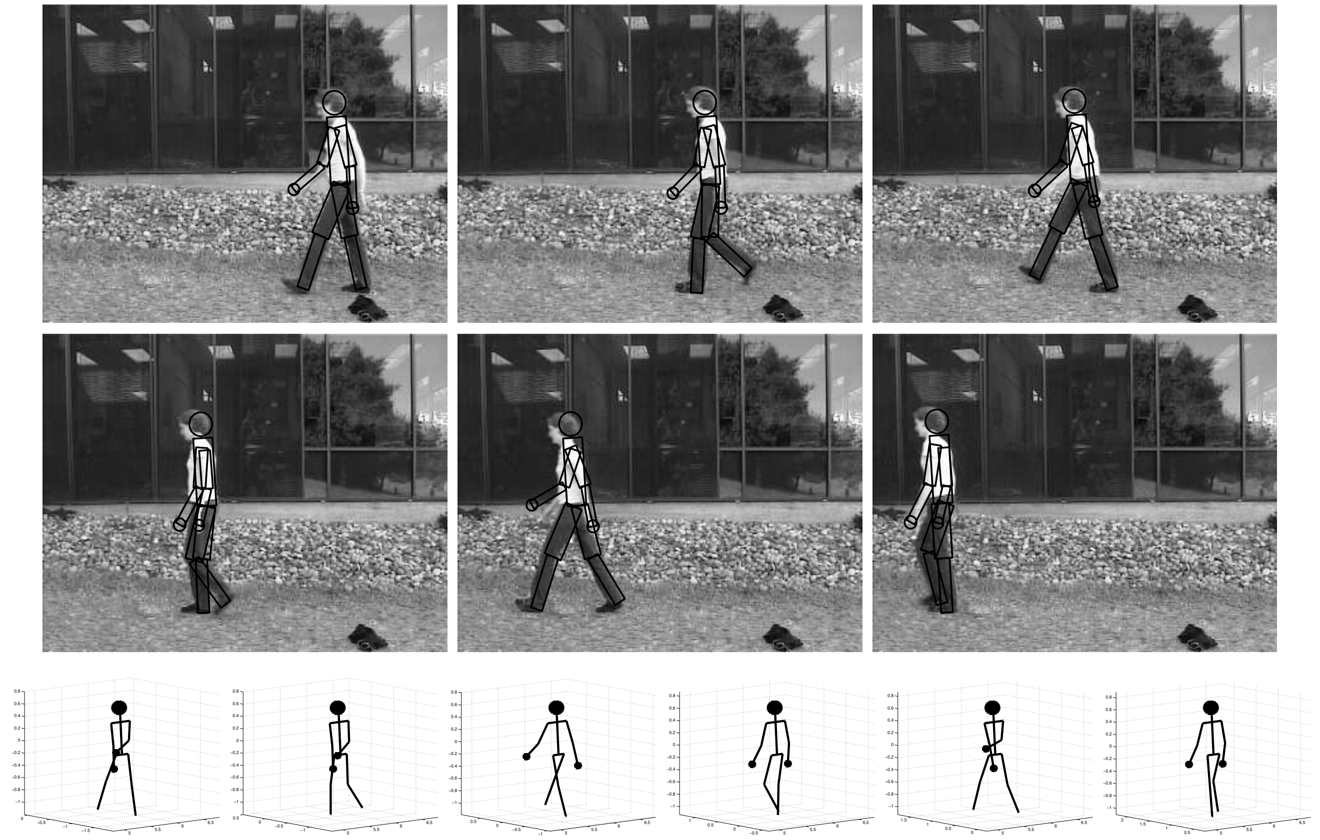

Figure 6: Tracking of person walking, 10000 samples. The two upper rows show frames 0, 10, 20, 30, 40, 50 in the sequence with the projection of the expected model configuration overlayed. The lower row shows the expected 3D configuration in the same frames.

[6] L.W. Campbell and A.F. Bobick. Recognition of human body motion using phase space constraints. *ICCV*, pp. 624–630, 1995.

[7] T-J. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. *CVPR*, pp. 239–245, 1999.

[8] J. Deutscher, B. North, B. Bascle, and A. Blake. Tracking through singularities and discontinuities by random sampling. *ICCV*, pp. 1144–1149, 1999.

[9] N. Gordon. A novel approach to nonlinear/non-gaussian Bayesian state estimation. *IEE Proceedings on Radar, Sonar and Navigation*, 140(2):107–113, 1996.

[10] D. Hogg, Model-based vision: A program to see a walking person. *IVC*, 1(1), pp. 5–20, 1983.

[11] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. *ECCV*, pp. 343–356, 1996.

[12] M. E. Leventon and W. T. Freeman. Bayesian estimation of 3-D human motion from an image sequence. Technical Report TR–98–06, Mitsubishi Electric Research Lab, 1998.

[13] V. Pavolvić, J. Rehg, T-J. Cham, and K. Murphy, A dynamic Bayesian network approach to figure tracking using learned dynamic models. *ICCV*, pp. 94–101, 1999.

[14] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. New York: Springer Verlag, 1997.

[15] K. Rohr. Human movement analysis based on explicit motion models. In M. Shah and R. Jain, eds., *Motion-Based Recognition*, pp. 171–198, 1997. Kluwer Academic Pub.

[16] S.M. Seitz and C.R. Dyer. Affine invariant detection of periodic motion. *CVPR*, pp. 970–975, 1994.

[17] H. Sidenbladh, F. de la Torre and M. J. Black. A framework for modeling the appearance of 3D articulated figures. *Int. Conf. on Autom. Face and Gesture Recog.*, 2000, pp. 368–375.

[18] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. *ECCV-2000*, Dublin, Ireland.

[19] S. Wachter and H. H. Nagel. Tracking persons in monocular image sequences. *CVIU*, 74(3):174–192, 1999.

[20] A. D. Wilson and A. F. Bobick. Parametric Hidden Markov Models for gesture recognition. *PAMI*, 21(9):884–900, Sept. 1999.

[21] Y. Yacoob and M. Black. Parameterized modeling and recognition of activities in temporal surfaces. *CVIU*, 73(2):232–247, 1999.

[22] T. Hastie and R. Tibshirani, *Generalized additive models*, Chapman and Hall,, 1990.

[23] G. Sherlock, M. Eisen, O. Alter, D. Botstein, P. Brown, T. Hastie, and R. Tibshirani. “Imputing missing data for gene expression arrays,” 2000, Working Paper, Department of Statistics, Stanford University.