Stochastic Tracking of 3D Human Figures Using 2D Image Motion

Hedvig Sidenbladh[1] Michael J. Black[2] David J. Fleet[2]

[1] Royal Institute of Technology (KTH), CVAP/NADA, S–100 44 Stockholm, Sweden
hedvig@nada.kth.se
[2] Xerox Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304 USA
{black,fleet}@parc.xerox.com

Abstract. A probabilistic method for tracking 3D articulated human figures in monocular image sequences is presented. Within a Bayesian framework, we define a generative model of image appearance, a robust likelihood function based on image graylevel differences, and a prior probability distribution over pose and joint angles that models how humans move. The posterior probability distribution over model parameters is represented using a discrete set of samples and is propagated over time using particle filtering. The approach extends previous work on parameterized optical flow estimation to exploit a complex 3D articulated motion model. It also extends previous work on human motion tracking by including a perspective camera model, by modeling limb self occlusion, and by recovering 3D motion from a monocular sequence. The explicit posterior probability distribution represents ambiguities due to image matching, model singularities, and perspective projection. The method relies only on a frame-to-frame assumption of brightness constancy and hence is able to track people under changing viewpoints, in grayscale image sequences, and with complex unknown backgrounds.

1 Introduction

We present a Bayesian approach to tracking 3D articulated human figures in monocular video sequences. The human body is represented by articulated cylinders viewed under perspective projection. A *generative model* is defined in terms of the shape, appearance, and motion of the body, and a model of noise in the pixel intensities. This leads to a likelihood function that specifies the probability of observing an image given the model parameters. A prior probability distribution over model parameters depends on the temporal dynamics of the body and the history of body shapes and motions. With this likelihood function and temporal prior, we formulate the posterior distribution over model parameters at each time instant, given the observation history.

The estimation of 3D human motion from a monocular sequence of 2D images is challenging for a variety of reasons. These include the non-linear dynamics of the limbs, ambiguities in the mapping from the 2D image to the 3D model, the similarity of the appearance of different limbs, self occlusions, kinematic singularities, and image noise. One consequence of these difficulties is that, in general, we expect the posterior probability distribution over model parameters to be multi-modal. Also, we cannot expect to find an analytic, closed-form, expression for the likelihood function over model parameters. For these two reasons, we represent the posterior distribution non-parametrically using a discrete set of samples (i.e., states), where each sample corresponds to some hypothesized set of model parameters. Figure $1(a)$ illustrates this by showing a few samples from such a distribution over 3D model parameters projected into an image. This distribution is propagated in time using a particle filter [11, 13].

The detection and tracking of human motion in video has wide potential for application in domains as diverse as animation and human-computer interaction. For this reason there has been a remarkable growth in research on this problem. The majority of proposed methods rely on sources of information such as skin color or known backgrounds which may not always be available. Such cues, while useful, are not intrinsic to 3D human motion. We focus, instead, on the 3D motion of the figure and its projection into the image plane of the camera. This formulation, in terms of image motion, gives the tracker some measure of independence with respect to clothing, background clutter, and ambient lighting. Additionally, the approach does not require color images, nor does it require multiple cameras with different viewpoints. As a consequence, it may be used with archival movie footage and inexpensive video surveillance equipment. The use of perspective projection allows the model to handle significant changes in depth. Finally, unlike template tracking methods [6], the use of image motion allows tracking under changing viewpoint. These properties are illustrated with examples that include tracking people walking in cluttered images while their depth and orientation with respect to the camera changes significantly.

2 Related Work

Estimation of human motion is an active and growing research area [8]. We briefly review previous work on image cues, body representations, temporal models, and estimation techniques.

Image Cues. Methods for full body tracking typically use simple cues such as background difference images [4], color [22] or edges [7, 9, 10, 15]. However robust, these cues provide sparse information about the features in the image. Image motion (optical flow) [5, 14, 24] provides a dense cue but, since it only exploits relative motion between frames, it is sensitive to the accumulation of errors over multiple frames. The result is that these techniques are prone to "drift" from the correct solution over time. The use of image templates [6] can avoid this problem, but such approaches are sensitive to changes in view and illumination. Some of the most interesting work to date has combined multiple cues such as edges and optical flow [21]. The Bayesian approach we describe may provide a framework for the principled combination of such cues.

The approach here focuses on the estimation of 3D articulated motion from 2D image changes. In so doing we exploit recent work on the probabilistic estimation of optical flow using particle filtering [1, 2]. The method has been applied to non-linear spatial and temporal models of optical flow, and is extended here to model the motion of articulated 3D objects.

Body and Camera Models. Models of the human body vary widely in their level of detail. At one extreme are methods that crudely model the body as a collection of articulated planar patches [14, 24]. At the other extreme are 3D models in which the limb shapes are deformable [9, 15]. Additionally, assumptions about the viewing conditions vary from scaled orthographic projection [5] to full perspective [21, 25]. To account for large variations in depth, we model the body in terms of articulated 3D cylinders [12] viewed under perspective projection.

Temporal Models. Temporal models of body limb or joint motion also vary in complexity; they include smooth motion [7], linear dynamical models [18], non-linear models learned from training data using dimensionality reduction [3, 16, 23], and probabilistic Hidden Markov Models (HMM's) (e.g., [4]). In many of these methods, image measurements are first computed and then the temporal models are applied to either smooth or interpret the results. For example, Leventon and Freeman [16] proposed a Bayesian framework for recovering 3D human motion from the motion of a 2D stick figure. They learned a prior distribution over human motions using vector quantization. Given the 2D motion of a set of joints, the most plausible 3D motion could be found. They required a pre-processing step to determine the 2D stick figure motion and did not tie the 3D motion directly to the image. Their Bayesian framework did not represent multi-modal distributions and therefore did not maintain multiple interpretations.

Brand [4] learned a more sophisticated HMM from the same 3D training data used in [16]. Brand's method used binary silhouette images to compute a feature vector of image moments. The hidden states of the HMM represented 3D body configurations and the method could recover 3D models from a sequence of feature vectors. These weak image cues meant that the tracking results were heavily dependent on the prior temporal model.

Unlike the above methods, we explore the use of complex non-linear temporal models early in the process to constrain the estimation of low-level image measurements. In related work Yacoob and Davis [24] used a learned "eigen-curve" model of image motion [23] to constrain estimation of a 2D articulated model. Black [1] used similar non-linear temporal models within a probabilistic framework to constrain the estimation of optical flow.

Estimation. Problems with articulated 3D tracking arise due to kinematic singularities [17], depth ambiguities, and occlusion. Multiple camera views, special clothing, and simplified backgrounds have been used to ameliorate some of these problems [5, 9, 15]. In the case of monocular tracking, body parts with low visibility (e.g. one arm and one leg) are often excluded from the tracking to avoid occlusion effects and also to lower the dimensionality of the model [5]. Cham and Rehg [6] avoid kinematic singularities and depth ambiguities by using a 2D model with limb foreshortening [17]. They also employ a multi-modal tracking approach related to particle filtering.

Bregler and Malik [5] assumed scaled orthographic projection and posed the articulated motion problem as a linear estimation problem. Yamamoto *et al.* [25] also formulated a linear estimation problem and relied on multiple camera views. These approaches elegantly modeled the image motion but did not account for imaging ambiguities and multiple matches.

Recently, Deutscher *et al.* [7] showed promising results in 3D tracking of body parts using a particle filtering method (the Condensation [13] algorithm). They successfully tracked an arm through kinematic singularities. We address the singularity problems in the same way but focus on image motion rather than edge tracking. We also employ learned temporal models to compensate for depth ambiguities and occlusion effects, and we show tracking results with more complex full-body motions.

3 Generative Model

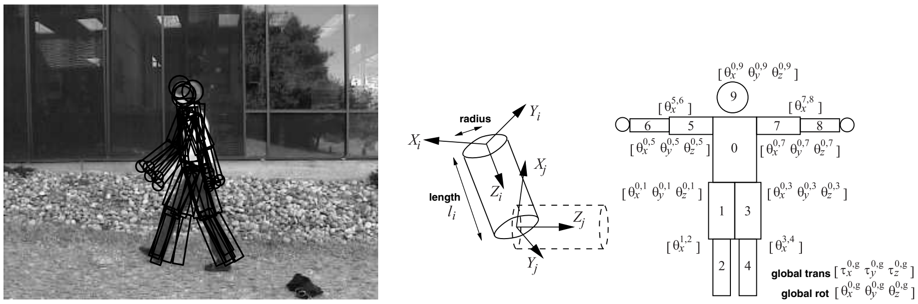
A Bayesian approach to human motion estimation requires that we formulate a generative model of image appearance and motion. This model defines the state space representation for humans and their motion and specifies the probabilistic relationship between these states and observations. The generative model of human appearance described below has three main components, namely, shape, appearance, and motion. The human body is modeled as an articulated object, parameterized by a set of joint angles and an appearance function for each of the rigid parts. Given the camera parameters and the position and orientation of the body in the scene, we can render images of how the body is likely to appear. The probabilistic formulation of the generative model provides the basis for evaluating the likelihood of observing image measurements, $\mathbf{I}_t$ at time t, given the model parameters.

3.1 Shape: Human Body Model

As shown in Figure 1, the body is modeled as a configuration of 9 cylinders and 3 spheres, numbered for ease of identification. All cylinders are right-circular, except for the torso which has an elliptical cross-section. More sophisticated tapered cylinders [7, 21] or superquadrics [8] could be employed. Each part is defined in a part-centric coordinate frame with the origin at the base of the cylinder (or sphere). Each part is connected to others at joints, the angles of which are represented as Euler angles. The origin in each part's coordinate frame corresponds to the center of rotation (the joint position).

Rigid transformations, $\mathbf{T}$, are used to specify relative positions and orientations of parts and to change coordinate frames. We express them as a homogeneous transformation matrices:

$$\mathbf{T} = \begin{bmatrix} \mathbf{R}_z\mathbf{R}_y\mathbf{R}_x & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \tag{1}$$



a *b*

Fig. 1. *(a)* A few samples from a probability distribution over 3D model parameters projected into the image coordinate system. *(b)* Human body model. 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 rotations $(\theta_x, \theta_y, \theta_z)$.

where $\mathbf{R}_x$, $\mathbf{R}_y$ and $\mathbf{R}_z$ denote 3×3 rotation matrices about the coordinate axes, with angles θ_x, θ_y and θ_z, and $\mathbf{t} = [\tau_x, \tau_y, \tau_z]^T$ denotes the translation.

A kinematic tree, with the torso at its root, is used to order the transformations between the coordinate frames of different limbs. For example, in Figure 1*b*, the point $\mathbf{P}_1$ in the local coordinate system of limb 1 (the right thigh) can be transformed to the corresponding point $\mathbf{P}_g$ in the global coordinate system as $\mathbf{P}_g = \mathbf{T}_{0,g}\mathbf{T}_{1,0}\mathbf{P}_1$. The global translation and rotation of the torso are represented by $\mathbf{T}_{0,g}$, while the translation and rotation of the right thigh with respect to the torso are represented by $\mathbf{T}_{1,0}$.

With these definitions, as shown in Figure 1*b*, the entire pose and shape of the body is given by 25 parameters, that is, angles at the shoulders, elbows, hips and knees, and the position and orientation of the torso in the scene. Let $\boldsymbol{\phi}$ be the vector containing these 25 parameters.

Camera Model. The geometrical optics are modeled as a pinhole camera, with a transformation matrix $\mathbf{T}_c$ defining the 3D orientation and position of a 3D camera-centered coordinate system with a focal length f and an image center $\mathbf{c} = [x_\mathbf{c}, y_\mathbf{c}]^T$. The matrix maps points in scene coordinates to points in camera coordinates. Finally, points in 3D camera coordinates are projected onto the image at locations, $\mathbf{x} = [x, y]^T$, given by $\mathbf{x} = \mathbf{c} - f[\frac{Z_c}{X_c}, \frac{Y_c}{X_c}]^T$.

3.2 Appearance Model

For generality, we assume that each limb is textured mapped with an appearance model, $\mathbf{R}$. There are many ways in which one might specify such a model, including the use of low-dimensional linear subspaces [20]. Moreover, it is desirable, in general, to estimate the appearance parameters through time to reflect the changing appearance of the object in the image. Here we use a particularly simple approach in which the appearance function at time t is taken to be the mapping, $M(\cdot)$, of the image at time $t-1$ onto the 3D shape given by the shape parameters at time $t-1$:

$$\mathbf{R}_t = M(\mathbf{I}_{t-1}, \boldsymbol{\phi}_{t-1}) \ .$$

In probabilistic terms, this means that the probability distribution over appearance functions at time t, conditioned on past shapes $\bar{\boldsymbol{\phi}}_{t-1} = [\boldsymbol{\phi}_{t-1}, \ldots, \boldsymbol{\phi}_0]$, past image observations, $\bar{\mathbf{I}}_{t-1} = [\mathbf{I}_{t-1}, \ldots, \mathbf{I}_0]$, and past appearance functions $\bar{\mathbf{R}}_{t-1} = [\mathbf{R}_{t-1}, \ldots, \mathbf{R}_0]$, is given by

$$p(\mathbf{R}_t \,|\, \bar{\mathbf{I}}_{t-1}, \bar{\boldsymbol{\phi}}_{t-1}, \bar{\mathbf{R}}_{t-1}) = p(\mathbf{R}_t \,|\, \mathbf{I}_{t-1}, \boldsymbol{\phi}_{t-1}) = \delta(\mathbf{R}_t - M(\mathbf{I}_{t-1}, \boldsymbol{\phi}_{t-1})) \ , \quad (2)$$

where $\delta(\cdot)$ is a Dirac delta function.

Our generative model of the image, $\mathbf{I}_t$, at time t is then the projection of the human model (shape and appearance) corrupted by noise:

$$\mathbf{I}_t(\mathbf{x}_j) = M^{-1}(\mathbf{R}_t, \boldsymbol{\phi}_t, \mathbf{x}_j) + \eta \quad (3)$$

where $M^{-1}(\mathbf{R}_t, \boldsymbol{\phi}_t, \mathbf{x}_j)$ maps the 3D model of limb j to image location $\mathbf{x}_j$ and $\mathbf{I}_t(\mathbf{x}_j)$ is the image brightness at pixel location $\mathbf{x}_j$. To account for "outliers", the noise, η, is taken to be a mixture of a Gaussian and a uniform distribution

$$p_\eta(\eta; \mathbf{x}_j, \boldsymbol{\phi}_t) = (1-\epsilon)\, G(\sigma(\alpha(\mathbf{x}_j, \boldsymbol{\phi}_t))) + \epsilon\, c,$$

where $0 \le \epsilon \le 1$ and $c = 1/256$. The uniform noise is bounded over a finite interval of intensity values while $G(\cdot)$ is zero-mean normal distribution the variance of which may change with spatial position. In general, the variance is sufficiently small that the area of the Gaussian outside the bounded interval may be ignored.

The prediction of image structure, $\mathbf{I}_t$, given an appearance model, $\mathbf{R}_t$, estimated from the image at time $t-1$ will be less reliable in limbs, or regions of limbs, that are viewed obliquely compared with those that are nearly frontoparallel. In these regions, the image structure can change greatly from one frame to the next due to perspective distortions and self occlusion. This is captured by allowing the variance to depend on the orientation of the model surface.

Let $\alpha(\mathbf{x}_j, \boldsymbol{\phi}_t)$ be a function that takes an image location, $\mathbf{x}_j$, and projects it onto a 3D limb position $\mathbf{P}$ and returns the angle between the surface normal at the point $\mathbf{P}$ and the vector from $\mathbf{P}$ to the focal point of the camera. The variance of the Gaussian component of the noise is then defined with respect to the expected image noise, σ_I, which is assumed constant, and $\alpha(\mathbf{x}_j, \boldsymbol{\phi}_t)$:

$$\sigma^2(\alpha(\mathbf{x}_j, \boldsymbol{\phi}_t)) = (\sigma_I / \cos(\alpha(\mathbf{x}_j, \boldsymbol{\phi}_t)))^2 \ . \quad (4)$$

3.3 Temporal Dynamics

Finally we must specify the temporal dynamics as part the generative model. Towards this end we parameterize the motion of the shape in terms of a vector of velocities, $\mathbf{V}_t$, whose elements correspond to temporal derivatives of the shape and pose parameters in $\boldsymbol{\phi}$. Furthermore, we assume a first-order Markov model on shape and velocity. Let the entire history of the shape and motion parameters up to time t be denoted by $\bar{\boldsymbol{\phi}}_t = [\boldsymbol{\phi}_t, \ldots, \boldsymbol{\phi}_0]$ and $\bar{\mathbf{V}}_t = [\mathbf{V}_t, \ldots, \mathbf{V}_0]$. Then, the temporal dynamics of the model are given by

$$p(\boldsymbol{\phi}_t \mid \bar{\boldsymbol{\phi}}_{t-1}, \bar{\mathbf{V}}_{t-1}) = p(\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1}, \mathbf{V}_{t-1}) \; , \tag{5}$$

$$p(\mathbf{V}_t \mid \bar{\boldsymbol{\phi}}_{t-1}, \bar{\mathbf{V}}_{t-1}) = p(\mathbf{V}_t \mid \mathbf{V}_{t-1}) \; . \tag{6}$$

Humans move is a variety of complex ways, depending on the activity or gestures being made. Despite this complexity, the movements are often predictable. In Section 6, we explore two specific models of human motion. The first is a simple, general model of constant angular velocity. The second is an activity-specific model of walking.

4 Bayesian Formulation

The goal of tracking a human figure can now be formulated as the computation of the posterior probability distribution over the parameters of the generative model at time t, given a sequence of images, $\bar{\mathbf{I}}_t$; i.e., $p(\boldsymbol{\phi}_t, \mathbf{V}_t, \mathbf{R}_t \mid \bar{\mathbf{I}}_t)$. This can be expressed as a marginalization of the joint posterior over all states up to time t given all images up to time t:

$$p(\boldsymbol{\phi}_t, \mathbf{V}_t, \mathbf{R}_t \mid \bar{\mathbf{I}}_t) = \int p(\bar{\boldsymbol{\phi}}_t, \bar{\mathbf{V}}_t, \bar{\mathbf{R}}_t \mid \bar{\mathbf{I}}_t) d\bar{\boldsymbol{\phi}}_{t-1} d\bar{\mathbf{V}}_{t-1} d\bar{\mathbf{R}}_{t-1} \; . \tag{7}$$

Using Bayes' rule and the Markov assumptions above, it can be shown that the dependence on states at times before time $t-1$ can be removed, to give

$$\begin{aligned} & p(\boldsymbol{\phi}_t, \mathbf{V}_t, \mathbf{R}_t \mid \bar{\mathbf{I}}_t) = \\ & \kappa\, p(\mathbf{I}_t \mid \boldsymbol{\phi}_t, \mathbf{V}_t, \mathbf{R}_t) \int \big[p(\boldsymbol{\phi}_t, \mathbf{V}_t, \mathbf{R}_t \mid \boldsymbol{\phi}_{t-1}, \mathbf{V}_{t-1}, \mathbf{R}_{t-1}, \mathbf{I}_{t-1}) \\ & \qquad\qquad p(\boldsymbol{\phi}_{t-1}, \mathbf{V}_{t-1}, \mathbf{R}_{t-1} \mid \bar{\mathbf{I}}_{t-1}) \big] \, d\boldsymbol{\phi}_{t-1} d\mathbf{V}_{t-1} d\mathbf{R}_{t-1} \end{aligned} \tag{8}$$

where κ is a normalizing constant that does not depend on the state variables. Here, $p(\mathbf{I}_t \mid \boldsymbol{\phi}_t, \mathbf{V}_t, \mathbf{R}_t)$, which we refer to as the "likelihood," is the probability of observing the image at time t, given the shape, motion and appearance states at time t. The integral in (8) is referred to as a temporal prior, or a prediction, as it is equivalent to the probability over states at time t given the image measurement history; i.e., $p(\boldsymbol{\phi}_t, \mathbf{V}_t, \mathbf{R}_t \mid \bar{\mathbf{I}}_{t-1})$. It is useful to understand the integrand as the product of two terms; these are the posterior probability distribution over states at the previous time, $p(\boldsymbol{\phi}_{t-1}, \mathbf{V}_{t-1}, \mathbf{R}_{t-1} \mid \bar{\mathbf{I}}_{t-1})$, and the dynamical process that propagates this distribution over states from time $t-1$ to time t.

Before turning to the computation of the posterior in (8), it is useful to simplify it using the generative model described above. For example, the likelihood of observing the image at time t does not depend on the velocity $\mathbf{V}_t$, and therefore $p(\mathbf{I}_t \mid \boldsymbol{\phi}_t, \mathbf{V}_t, \mathbf{R}_t) = p(\mathbf{I}_t \mid \boldsymbol{\phi}_t, \mathbf{R}_t)$. Also, the probability distribution over the state variables at time t, conditioned on those at time $t-1$, can be factored further. This is based on the generative model, and the assumption that the evolution of velocity and shape from time $t-1$ to t is independent of the evolution of appearance. This produces the following factorization

$$p(\boldsymbol{\phi}_t, \mathbf{V}_t, \mathbf{R}_t \,|\, \boldsymbol{\phi}_{t-1}, \mathbf{V}_{t-1}, \mathbf{R}_{t-1}, \mathbf{I}_{t-1}) = \\ p(\boldsymbol{\phi}_t \,|\, \boldsymbol{\phi}_{t-1}, \mathbf{V}_{t-1}) p(\mathbf{V}_t \,|\, \mathbf{V}_{t-1}) p(\mathbf{R}_t \,|\, \mathbf{I}_{t-1}, \boldsymbol{\phi}_{t-1}) \ .$$

Finally, these simplifications, taken together, produce the posterior distribution

$$\begin{aligned} & p(\boldsymbol{\phi}_t, \mathbf{V}_t, \mathbf{R}_t \,|\, \bar{\mathbf{I}}_t) = \\ & \kappa\, p(\mathbf{I}_t \,|\, \boldsymbol{\phi}_t, \mathbf{R}_t) \int \big[p(\boldsymbol{\phi}_t \,|\, \boldsymbol{\phi}_{t-1}, \mathbf{V}_{t-1}) p(\mathbf{V}_t \,|\, \mathbf{V}_{t-1}) p(\mathbf{R}_t \,|\, \mathbf{I}_{t-1}, \boldsymbol{\phi}_{t-1}) \\ & \qquad p(\boldsymbol{\phi}_{t-1}, \mathbf{V}_{t-1}, \mathbf{R}_{t-1} \,|\, \bar{\mathbf{I}}_{t-1}) \big] \, d\boldsymbol{\phi}_{t-1} d\mathbf{V}_{t-1} d\mathbf{R}_{t-1} \ . \end{aligned} \tag{9}$$

4.1 Stochastic Optimization

Computation of the posterior distribution is difficult due to the nonlinearity of the likelihood function over model parameters. This is a consequence of self-occlusions, viewpoint singularities, and matching ambiguities. While we cannot derive an analytic expression for the likelihood function over the parameters of the entire state space, we can evaluate the likelihood of observing the image given a particular state $(\boldsymbol{\phi}_t^s, \mathbf{V}_t^s, \mathbf{R}_t^s)$; the computation of this likelihood is described in Section 5.

Representation of the posterior is further complicated by the use of a non-linear dynamical model of the state evolution as embodied by the temporal prior. While we cannot assume that the posterior distribution will be Gaussian, or even unimodal, robust tracking requires that we maintain a representation of the entire distribution and propagate it through time. For these reasons we represent the posterior as a weighted set of state samples, which are propagated using a particle filter with sequential importance sampling. Here we briefly describe the method (for foundations see [11, 13], and for applications to 2D image tracking with non-linear temporal models see [1, 2]).

Each state, $\mathbf{s}_t$, is represented by a vector of parameter assignments, $\mathbf{s}_t = [\boldsymbol{\phi}_t^s, \mathbf{V}_t^s]$. Note that in the current formulation we can drop the appearance model $\mathbf{R}_t^s$ from the state as it is completely determined by the shape parameters and the images. The posterior at time $t-1$ is represented by N state samples ($N \approx 10^4$ in our experiments). To compute the posterior (9) at time t we first draw N samples according to the posterior probability distribution at time $t-1$. For each state sample from time $t-1$, we compute $\mathbf{R}_t$ given the generative model. We propagate the angular velocities forward in time by sampling from the prior $p(\mathbf{V}_t \,|\, \mathbf{V}_{t-1})$. Similarly, the shape parameters are propagated by sampling from $p(\boldsymbol{\phi}_t \,|\, \boldsymbol{\phi}_{t-1}, \mathbf{V}_{t-1})$. At this point we have new values of $\boldsymbol{\phi}_t$ and $\mathbf{R}_t$ which can be used to compute the likelihood $p(\mathbf{I}_t \,|\, \boldsymbol{\phi}_t, \mathbf{R}_t)$. The N likelihoods are normalized to sum to one and the resulting set of samples approximates the posterior distribution $p(\boldsymbol{\phi}_t, \mathbf{V}_t, \mathbf{R}_t \,|\, \bar{\mathbf{I}}_t)$ at time t.

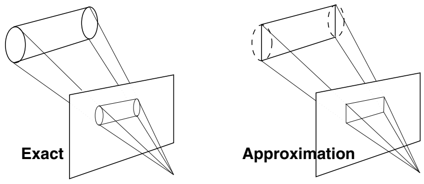


Fig. 2. Planar approximation of limbs improves efficiency.

5 Likelihood Computation

The likelihood $p(\mathbf{I}_t \mid \boldsymbol{\phi}_t, \mathbf{R}_t)$ is the probability of observing image $\mathbf{I}_t$ given that the human model has configuration $\boldsymbol{\phi}_t$ and appearance $\mathbf{R}_t$ at time t. To compare the image, $\mathbf{I}_t$, with the generative model, the model must be projected into the image plane of the camera as described in Section 3. To reduce the influence of camera noise on the matching, the images, $\mathbf{I}_t$, are smoothed by a Gaussian filter with a standard deviation of $\sqrt{2}$. This has the effect of smoothing the likelihood function over model parameters and hence the posterior distribution.

Projection. The projection of limb surface points into the image plane and vice versa is computationally expensive. Given the stochastic sampling framework, this operation is performed many times and hence we seek a efficient approximation. To simplify the projection onto the image, we first project the the visible portion of the cylindrical surface onto a planar patch that bisects the cylinder, as shown in Figure 2. The projection of the appearance of a planar patch into the image can be performed by first projecting the corners of the patch via perspective projection. The projection of other limb points is given by interpolation. This approximation speeds up the likelihood computation significantly.

Recall that the variance in the generative model (3) depends on the angle, $\alpha(\mathbf{x}_j, \boldsymbol{\phi}_t)$, between of the surface normal and the optical axis of the camera. With the planar approximation, α_j becomes the angle between the image plane and the Z axis of limb j.

Likelihood Model. Given the generative model we define the likelihood of each limb j independently. We sample, with replacement, $i = 1 \ldots n$ pixel locations, $\mathbf{x}_{j,i}$, uniformly from the projected region of limb j. According to (3), the grayvalue differences between points on the appearance model and the corresponding image values are independent and are modeled as a mixture of a zero-mean normal distribution and a uniform outlier distribution. We expect outliers, or unmatched pixels, to result from occlusion, shadowing, and wrinkled clothing.

The image likelihood of limb j is then expressed as:

$$p_{image} = \frac{\epsilon}{256} + \frac{1-\epsilon}{\sqrt{2\pi}\sigma(\alpha_j)} \exp\left(-\sum_{i=1}^{n} \frac{(\mathbf{I}_t(\mathbf{x}_{j,i}) - \hat{\mathbf{I}}_t(\mathbf{x}_{j,i}))^2}{2\sigma^2(\alpha_j)}\right) \tag{10}$$

where $\hat{\mathbf{I}}_t(\mathbf{x}_{j,i}) = M^{-1}(M(\mathbf{I}_{t-1}, \boldsymbol{\phi}_{t-1}), \boldsymbol{\phi}_t, \mathbf{x}_{j,i})$.

The likelihood must also account for occlusion which results from the depth ordering of the limbs or from the surface orientation. To model occluded regions we introduce the constant probability, $p_{occluded}$, that a limb is occluded. $p_{occluded}$ is currently determined empirically.

To determine self occlusion in the model configuration $\boldsymbol{\phi}_t$, the limbs are ordered according to the shortest distance from the limb surface to the image plane, using the camera parameters and $\boldsymbol{\phi}_t$. Limbs that are totally or partly covered by other limbs with lower depth are defined as occluded. This occlusion detection is sub-optimal and could be refined so that portions of limbs can be defined as occluded (cf. [21]).

Similarly as the limb is viewed at narrow angles (all visible surface normals are roughly perpendicular to the viewing direction) the linearized limb shape formulation makes the appearance pattern highly distorted. In this case, the limb can be thought of as occluding itself.

We then express the likelihood as a mixture between p_{image} and the constant probability of occlusion, $p_{occluded}$. 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. When the limb is exactly perpendicular to the image plane, it is by this definition considered occluded. The expression for the image likelihood of limb j is defined as:

$$p_j = q(\alpha_j) p_{image} + (1 - q(\alpha_j)) p_{occluded} \tag{11}$$

where $q(\alpha_j) = cos(\alpha_j)$ if limb j is non-occluded, or 0 if limb j is occluded.

According to the generative model, the appearance of the limbs are independent and the likelihood of observing the image given a particular body pose is given by the product of the limb likelihoods:

$$p(\mathbf{I}_t \mid \boldsymbol{\phi}_t, \mathbf{R}_t) = \prod_j p_j \ . \tag{12}$$

6 Temporal Model

The temporal model encodes information about the dynamics of the human body. Here it is formulated as a prior probability distribution and is used to constrain the sampling to portions of the parameter space that are likely to correspond to human motions. General models such as constant acceleration can account for arbitrary motions but do not constrain the parameter space greatly. For a constrained activity such as walking or running we can construct a temporal model with many fewer degrees of freedom which makes the computational problem more tractable. Both types of models are explored below.

6.1 Generic Model: Smooth Motion

The smooth motion model assumes that the angular velocity of the joints and the velocity of the body are constant over time. Recall that the shape parameters

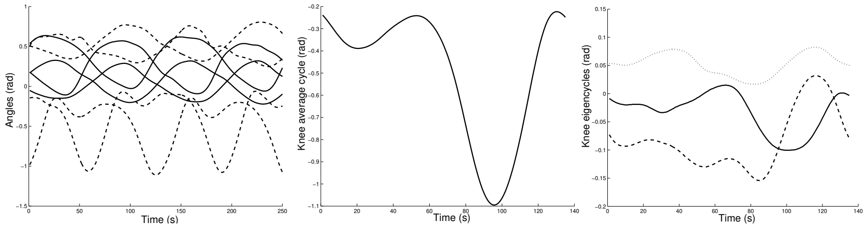


Fig. 3. Learning a walking model. *(a)* Joint angles of different people walking were acquired with a motion capture system. Curves are segmented into walking cycles manually and an eigenmodel of the cycle is constructed. *(b)* Mean angle of left knee as a function of time. *(c)* First three eigenmodes of the left knee $\mathbf{B}_j$, $j \in [1, 3]$, scaled by their respective variance λ_j. $(1 = solid, 2 = --, 3 = \cdots)$

are given by $\boldsymbol{\phi}_t = [\boldsymbol{\tau}^g_t, \boldsymbol{\theta}^g_t, \boldsymbol{\theta}^l_t]$ where $\boldsymbol{\tau}^g_t$ and $\boldsymbol{\theta}^g_t$ represent the translation and rotation that map the body into the world coordinate system and $\boldsymbol{\theta}^l_t$ represents the relative angles between all pairs of connected limbs. Let $\mathbf{V}_t = [\dot{\boldsymbol{\tau}}^g_t, \dot{\boldsymbol{\theta}}^g_t, \dot{\boldsymbol{\theta}}^l_t]$ represent the corresponding velocities. The physical limits of human movement are modeled as hard constraints on the individual quantities such that $\boldsymbol{\phi}_t \in [\boldsymbol{\phi}_{\min}, \boldsymbol{\phi}_{\max}]$.

Our smooth motion model assumes that all elements $\phi_{k,t} \in \boldsymbol{\phi}_t$ and $V_{q,t} \in \mathbf{V}_t$ are independent. The dynamics are represented by

$$p(\phi_{i,t} \mid \phi_{i,t-1}, V_{i,t-1}) = \begin{cases} G(\phi_{i,t} - (\phi_{i,t-1} + V_{i,t-1}), \sigma^\phi_i) & \text{if } \phi_{i,t} \in [\phi_{i,\min}, \phi_{i,\max}] \\ 0 & \text{otherwise} \end{cases}$$

$$p(V_{i,t} \mid V_{i,t-1}) = G(V_{i,t} - V_{i,t-1}, \sigma^V_i),$$

where $G(x, \sigma)$ denotes a Gaussian distribution with zero mean and standard deviation σ, evaluated at x. The standard deviations σ^ϕ_i and σ^V_i are empirically determined. The joint angles of heavy limbs typically have lower standard deviations than those in lighter limbs.

This model works well for tracking individual body parts that are relatively low dimensional. This is demonstrated in Section 7 for tracking arm motion (cf. [7]). This is a relatively weak model for constraining the motion of the entire body given the current sampling framework and limited computational resources. In general, one needs a variety of models of human motion and a principled mechanism for choosing among them.

6.2 Action Specific Model: Walking

In order to build stronger models, we can take advantage of the fact that many human activities are highly constrained and the body is often moved in symmetric and repetitive patterns. In what follows we consider the example of walking motion.

Training data corresponding to the 3D model parameters was acquired with a commercial motion capture system. Some of the data are illustrated in Figure 3. From the data, $m = 13$ example walking cycles from 4 different subjects (professional dancers) were segmented manually and scaled to the same length. These cycles are then used to train a walking model using Multivariate Principal Component Analysis (MPCA) [3, 19, 23]. In addition to the joint angles, we model the speed, $\mathbf{v}_i$ of the torso in the direction of the walking motion i. This speed, $v_{i,\mu}$, at time step μ in the cycle is $v_{i,\mu} = \|\boldsymbol{\tau}^g_{i,\mu+1} - \boldsymbol{\tau}^g_{i,\mu}\|$. The curves corresponding to the speed of the torso and the relative angles of the limbs, $\boldsymbol{\phi}^l_i$, are concatenated forming column vectors $\mathbf{A}_i$ for each training example $i = 1 \ldots m$. The mean vector $\hat{\mathbf{A}}$ is subtracted from all examples: $\mathcal{A}_i = \mathbf{A}_i - \hat{\mathbf{A}}$. Since the walking speed [m/frame] and the joint angles [radians] have approximately the same scales they need not be rescaled before applying MPCA.

The eigenvalues λ_j and eigenvectors $\mathbf{B}_j$, $j \in [1, m]$ of the matrix $\mathcal{A} = [\mathcal{A}_1, \cdots, \mathcal{A}_m]$ are now computed from $\mathcal{A} = \mathbf{B}\boldsymbol{\Sigma}\mathbf{D}^T$ using Singular Value Decomposition (SVD) where $\mathbf{B} = [\mathbf{B}_1, \cdots, \mathbf{B}_m]$ and $\boldsymbol{\Sigma}$ is a diagonal matrix with λ_j along the diagonal. The eigenvectors represent the principal modes of variation in the training set, while the eigenvalues reflect the variance of the training set in the direction of the corresponding eigenvector. The eigenvectors $\mathbf{B}_j$ can be viewed as a number of eigencurves, one for each joint, stacked together. Figure 3c shows three eigencurves corresponding to the left knee walking cycle.

The smallest number d of eigenvectors $\mathbf{B}_j$ such that $\sum_{j=1}^{d} \lambda_j^2 > 0.95$ is selected; in our case $d = 5$. With $\tilde{\mathbf{B}} = [\mathbf{B}_1, \cdots, \mathbf{B}_d]$ we can, with d parameters $\mathbf{c} = [c_1, \cdots, c_d]^T$, approximate a synthetic walking cycle $\mathbf{A}^*$ as:

$$\mathbf{A}^* = \hat{\mathbf{A}} + \tilde{\mathbf{B}}\mathbf{c}. \tag{13}$$

The set of independent parameters is now $\{\mathbf{c}_t, \mu_t, \boldsymbol{\tau}^g_t, \boldsymbol{\theta}^g_t\}$ where μ_t denotes the current position (or phase) in the walking cycle. Thus, this model reduces the original 25-dimensional parameter space, $\boldsymbol{\phi}$, to a 12-dimensional space.

Recall that the global translation and rotation, $\boldsymbol{\tau}^g_t$, $\boldsymbol{\theta}^g_t$, can be expressed as a homogeneous transformation matrix $\mathbf{T}$. We also define v_{t-1} to be the learned walking speed at time $t-1$. The parameters are propagated in time as:

$$p(\mathbf{c}_t \,|\, \mathbf{c}_{t-1}) = G(\mathbf{c}_t - \mathbf{c}_{t-1}, \boldsymbol{\sigma}^c I_d) \tag{14}$$
$$p(\mu_t \,|\, \mu_{t-1}) = G(\mu_t - \mu_{t-1}, \sigma^\mu) \tag{15}$$
$$p(\boldsymbol{\tau}^g_t \,|\, \mathbf{T}_{t-1}, \mathbf{c}_{t-1}) = G([\boldsymbol{\tau}^g_t, 1]^T - \mathbf{T}^{-1}_{t-1}[v_{t-1}\ 0\ 0\ 1]^T, \boldsymbol{\sigma}^{\boldsymbol{\tau}} I_3) \tag{16}$$
$$p(\boldsymbol{\theta}^g_t \,|\, \boldsymbol{\theta}_{t-1}) = G(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}, \boldsymbol{\sigma}^{\boldsymbol{\theta}} I_3) \tag{17}$$

where σ^μ, $\boldsymbol{\sigma}^{\boldsymbol{\tau}}$ and $\boldsymbol{\sigma}^{\boldsymbol{\theta}}$ represent the empirically determined standard deviations, I_n is an $n \times n$ identity matrix, and $\boldsymbol{\sigma}^c = \varepsilon\boldsymbol{\lambda}$ where ε is a small scalar with $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_d]^T$. ε is expected to be small since we expect the $\mathbf{c}$ parameters to vary little throughout the walking cycle for each individual

From a particular choice of $\{\mu_t, \mathbf{c}_t\}$, the relative joint angles are $\boldsymbol{\theta}^l_t = \mathbf{A}^*(\mu_t) = \hat{\mathbf{A}}(\mu_t) + (\tilde{\mathbf{B}}\mathbf{c})(\mu_t)$, where $\mathbf{A}^*(\mu_t)$ indicates the interpolated value of each joint cycle, $\mathbf{A}_i^*$, at phase μ. The angular velocities, $\dot{\boldsymbol{\theta}}_t^l = \mathbf{A}^*(\mu_t{+}1){-}\mathbf{A}^*(\mu_t)$, are not estimated independently and the velocities $\dot{\boldsymbol{\tau}}_t^g, \dot{\boldsymbol{\theta}}_t^g$ are propagated as in the smooth motion case above. The Gaussian distribution over μ_t and $\mathbf{c}_t$ implies a Gaussian distribution over joint angles which defines the distribution $p(\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1}, \mathbf{V}_{t-1})$ used in the Bayesian model.

7 Experiments

We present examples of tracking people or their limbs in cluttered images. On an Ultra 1 Sparcstation the C++ implementation takes approximately 5 minutes/frame for experiments with 10,000 state samples. At frame 0, the posterior distribution is derived from a hand-initialized 3D model. To visualize the posterior distribution we display the projection of the 3D model corresponding to the expected value of the model parameters: $\frac{1}{N}\sum_{i=1}^{N} p_i \boldsymbol{\phi}_i$ where p_i is the normalized likelihood of state sample $\boldsymbol{\phi}_i$.

Arm Tracking. The smooth motion prior is used for tracking relatively low dimensional models such as a single arm as illustrated in Figure 4. The model has 8 parameters corresponding to the orientation and velocity of the 3 shoulder angles and the elbow angle.

The twist of the upper arm θ_z is ambiguous when the arm is straight since the only information about the change in θ_z in that situation is the rotation of the texture pattern on the upper arm. If the upper arm texture is of low contrast (as in Figure 4) this will provide a very weak cue. This ambiguity is easily represented in a particle filtering framework. In our case, θ_z is assigned a uniform starting distribution. Some frames later (around frame 20), the arm bends slightly, and the distribution over θ_z concentrates near the true value. The rotation of a straight arm is an example of a kinematic singularity [7, 17].

Tracking Walking People. The walking model described in Section 6.2 is used to track a person walking on a straight path parallel to the camera plane over 50 frames (Figure 5). The global rotation of the torso was held constant, lowering the number of parameters to 9: the 5 eigencoefficients, $\mathbf{c}$, phase, μ, and global 3D position, $\boldsymbol{\tau}^g$. All parameters were initialized manually with a Gaussian prior at time $t = 0$ (Figure 5, frame 0). As shown in Figure 5 the model successfully tracks the person although some parts of the body (often the arms) are poorly estimated. This in part reflects the limited variation present in the training set.

The next experiment involves tracking a person walking in a circular path and thus changing both depth and orientation with respect to the camera. Figure 6 shows the tracking results for frames from 0 to 50. In frame 50 notice that the model starts to drift off the person since the rotation is poorly estimated. Such drift is common with optical flow-based tracking methods that rely solely on the the relative motion between frames. This argues for a more persistent model of object appearance. Note that, while a constant appearance model (i.e.

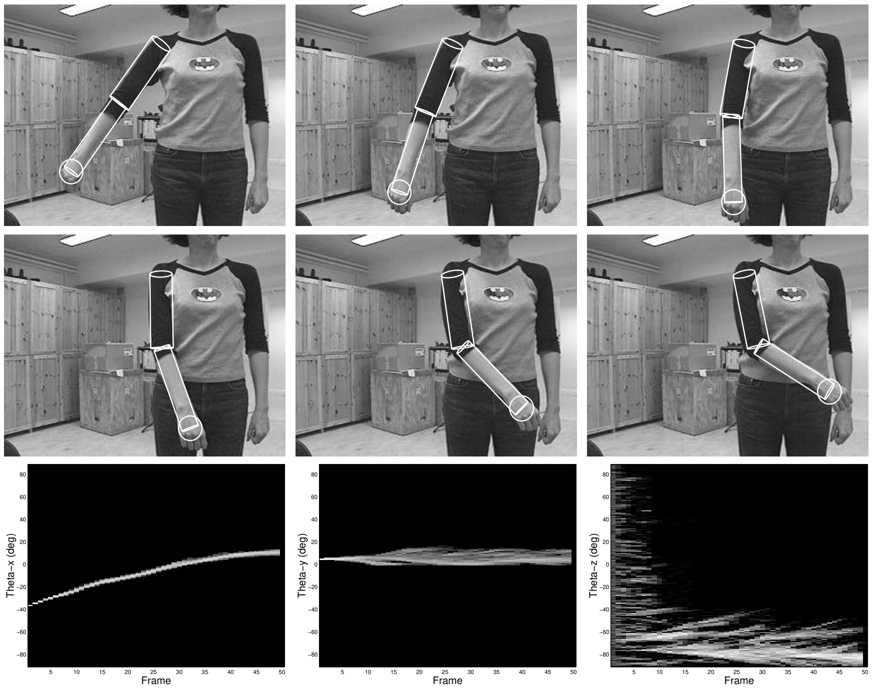


Fig. 4. Tracking of one arm (2000 samples). Upper rows: frames 0, 10, 20, 30, 40 and 50 with the projection of the expected value of model the model parameters overlaid. Frame 0 corresponds to the manual initialization. Lower row: distributions of the shoulder angles θ_x, θ_y and θ_z as function of frame number. Brightness values denote the log posterior distribution in each frame.

a template) would not suffer the same sort of drift it would be unable to cope with changes in view, illumination, and depth. Note also that the training data only contained examples of people walking in a straight line. While the circular walking motion here differs significantly, the temporal model is sufficiently general that it can approximate this new motion.

How significant is the temporal walking prior model? Figure 7 illustrates the effect of repeating the above experiment with a uniform likelihood function, so that the evolution of the parameters is determined entirely by the temporal model. While the prior is useful for constraining the model parameters to valid walking motions, it does not unduly affect the tracking.

8 Conclusion

This paper has presented a Bayesian formulation for tracking of articulated human figures in 3D using monocular image motion information. The approach employs a generative model of image appearance that extends the idea of parameterized optical flow estimation to 3D articulated figures. Kinematic singu-

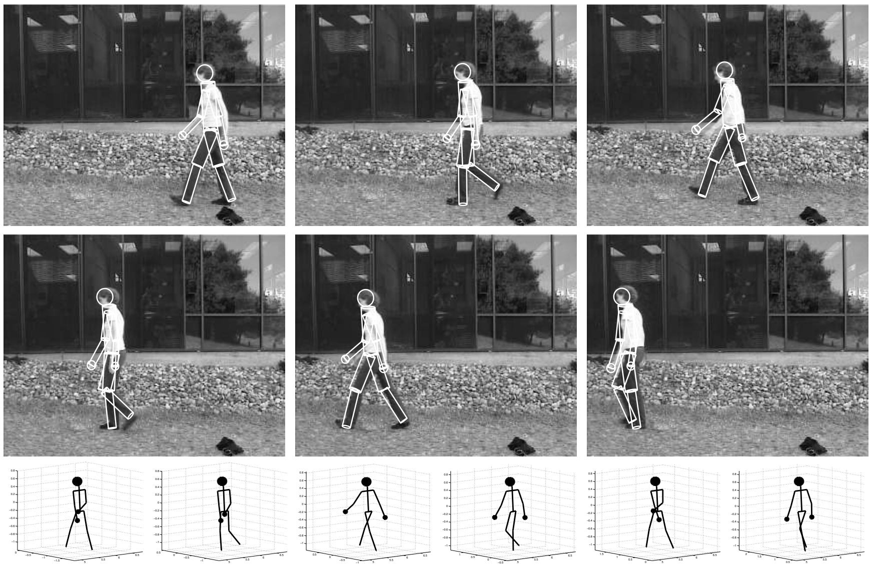


Fig. 5. Tracking a human walking in a straight line (5000 samples, no rotation). Upper rows: projection of the expected model configuration at frames 0, 10, 20, 30, 40 and 50. Lower row: 3D configuration for the expected model parameters in the same frames.

larities, depth ambiguities, occlusion, and ambiguous image information result in a multi-modal posterior probability distribution over model parameters. A particle filtering approach is used to represent and propagate the posterior distribution over time, thus tracking multiple hypotheses in parallel. To constrain the distribution to valid 3D human motions we define prior probability distributions over the dynamics of the human body. Such priors help compensate for missing or noisy visual information and enable stable tracking of occluded limbs. Results were shown for a general smooth motion model as well as for an action-specific walking model.

A number of outstanding issues remain and are the focus of our research. The current model is initialized by hand and will eventually lose track of the object. Within a Bayesian framework we are developing a fully automatic system that samples from a mixture of initialization and temporal priors. We are also developing new temporal models of human motion that allow more variation than the eigencurve model yet are more constrained than the smooth motion prior. We are extending the likelihood model to better use information at multiple scales and to incorporate additional generative models for image features such as edges. Additionally, the likelihood computation is being extended to model the partial occlusion of limbs as in [21]. Beyond this, one might replace the cylindrical limbs with tapered superquadrics [9, 15] and model the prior distribution over these additional shape parameters. Finally, we are exploring the representation of the posterior as a mixture of Gaussians [6]. This provides a more compact

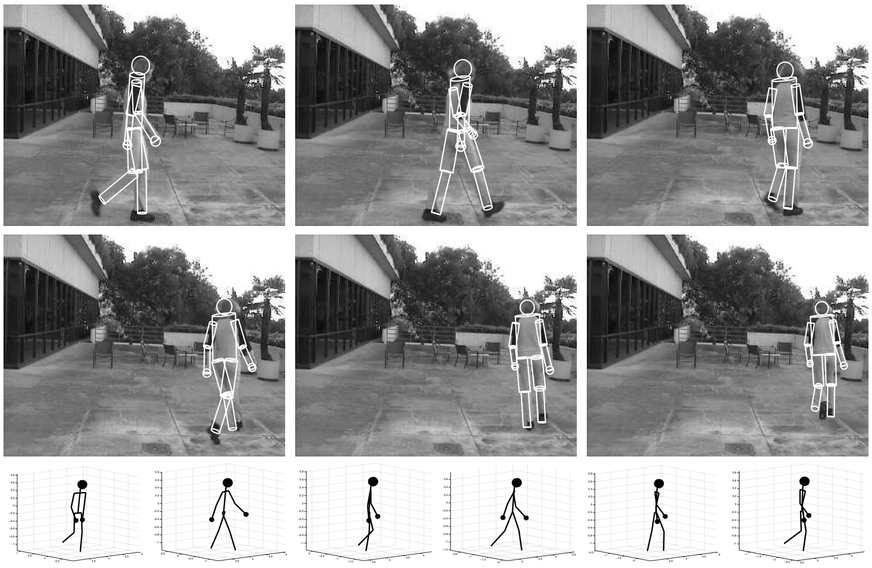


Fig. 6. Person walking in a circle (15000 samples). Upper rows: frames 0, 10, 20, 30, 40, 50 with the projection of the expected model configuration overlaid. Lower row: expected 3D configuration in the same frames.

representation of the distribution and interpolates between samples to provide a measure of the posterior in areas not covered by discrete samples.

Acknowledgments. This work was partially sponsored by the Foundation for Strategic Research under the "Center for Autonomous Systems" contract. This support is gratefully acknowledged. We would also like to thank Jan-Olof Eklundh, Dirk Ormoneit and Fernando De la Torre for helpful discussions and Michael Gleicher for providing the 3D motion capture data.

References

1. M. J. Black. Explaining optical flow events with parameterized spatio-temporal models. *CVPR*, pp. 326–332, 1999.
2. M. J. Black and D. J. Fleet. Probabilistic detection and tracking of motion discontinuities. *ICCV*, pp. 551–558, 1999.
3. A. Bobick and J. Davis. An appearance-based representation of action. *ICPR*, 1996.
4. M. Brand. Shadow puppetry. *ICCV*, pp. 1237–1244, 1999.
5. C. Bregler and J. Malik. Tracking people with twists and exponential maps. *CVPR*, 1998.
6. T-J. Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. *CVPR*, pp. 239–245, 1999.
7. J. Deutscher, B. North, B. Bascle, and A. Blake. Tracking through singularities and discontinuities by random sampling. *ICCV*, pp. 1144–1149, 1999.

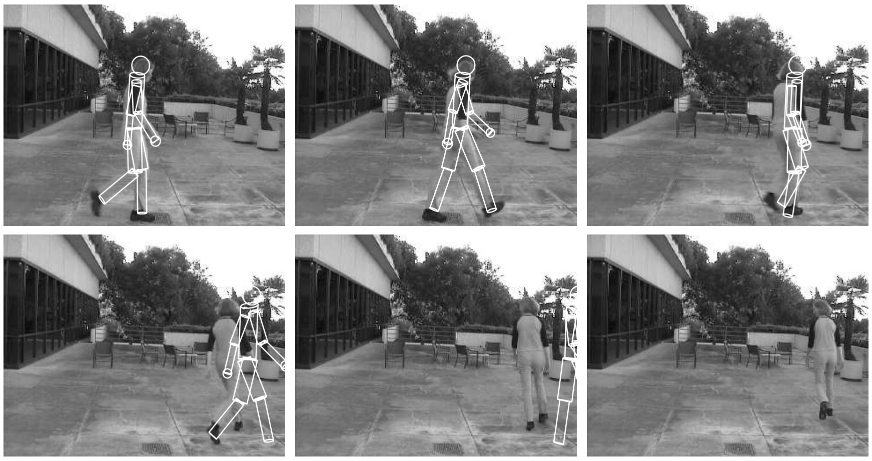


Fig. 7. How strong is the walking prior? Tracking results for frames 0, 10, 20, 30, 40 and 50, when no image information is taken into account.

8. D. M. Gavrila. The visual analysis of human movement: a survey. *CVIU*, 73(1):82–98, 1999.
9. D. M. Gavrila and L. S. Davis. 3-D model-based tracking of humans in action: A multi-view approach. *CVPR*, pp. 73–80, 1996.
10. L. Goncalves, E. Di Bernardi, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3D. *ICCV*, 1995.
11. N. Gordon, D. J. Salmond, and A. F. M. Smith. A novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. Radar, Sonar and Navigation*, 140(2):107–113, 1996.
12. D. Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.
13. M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. *ECCV*, pp. 343–356, 1996.
14. S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. *Int. Conf. on Automatic Face and Gesture Recognition*, pp. 38–44, 1996.
15. I. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. *CVPR*, pp. 81–87, 1996.
16. M. E. Leventon and W. T. Freeman. Bayesian estimation of 3-d human motion from an image sequence. TR–98–06, Mitsubishi Electric Research Lab, 1998.
17. D. Morris and J. M. Rehg. Singularity analysis for articulated object tracking. *CVPR*, pp. 289–296, 1998.
18. 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.
19. J. O. Ramsay and B. W. Silverman. *Functional data analysis.* New York: Springer Verlag, 1997.
20. H. Sidenbladh, F. de la Torre, and M. J. Black. A framework for modeling the appearance of 3D articulated figures. *Int. Conf. on Automatic Face and Gesture Recognition*, 2000.

21. S. Wachter and H. H. Nagel. Tracking persons in monocular image sequences. *CVIU*, 74(3):174–192, 1999.
22. C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *PAMI*, 19(7):780–785, 1997.
23. Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities in temporal surfaces. *CVIU*, 73(2):232–247, 1999.
24. Y. Yacoob and L. Davis. Learned temporal models of image motion. *ICCV*, pp. 446–453, 1998.
25. M. Yamamoto, A. Sato, S. Kawada, T. Kondo, and Y. Osaki. Incremental tracking of human actions from multiple views. *CVPR*, pp. 2–7, 1998.