Int J Comput Vis (2010) 87: 140–155
DOI 10.1007/s11263-009-0274-5

Physics-Based Person Tracking Using the Anthropomorphic Walker

Marcus A. Brubaker · David J. Fleet · Aaron Hertzmann

Received: 30 March 2008 / Accepted: 16 July 2009 / Published online: 6 August 2009

Abstract We introduce a physics-based model for 3D person tracking. Based on a biomechanical characterization of lower-body dynamics, the model captures important physical properties of bipedal locomotion such as balance and ground contact. The model generalizes naturally to variations in style due to changes in speed, step-length, and mass, and avoids common problems (such as footskate) that arise with existing trackers. The dynamics comprise a two degree-of-freedom representation of human locomotion with inelastic ground contact. A stochastic controller generates impulsive forces during the toe-off stage of walking, and spring-like forces between the legs. A higher-dimensional kinematic body model is conditioned on the underlying dynamics. The combined model is used to track walking people in video, including examples with turning, occlusion, and varying gait. We also report quantitative monocular and binocular tracking results with the HumanEva dataset.

Keywords Tracking people · Physics · Passive-based walking

This work was financially supported in part by NSERC Canada, the Canadian Institute for Advanced Research (CIFAR), the Canada Foundation for Innovation (CFI), the Alfred P. Sloan Foundation, Microsoft Research and the Ontario Ministry of Research and Innovation. A preliminary version of this work appeared in Brubaker et al. (2007).

M.A. Brubaker (✉) · D.J. Fleet · A. Hertzmann
Department of Computer Science, University of Toronto, Toronto, Canada
e-mail: mbrubake@cs.toronto.edu

D.J. Fleet
e-mail: fleet@cs.toronto.edu

A. Hertzmann
e-mail: hertzman@dgp.toronto.edu

1 Introduction

Most current methods for recovering human motion from monocular video rely on *kinematic* models learned from motion capture (mocap) data. Generative approaches rely on density estimation to learn a prior distribution over plausible human poses and motions, whereas discriminative models typically learn a mapping from image measurements to 3D pose. While the use of learned kinematic models clearly reduces ambiguities in pose estimation and tracking, the 3D motions estimated by these methods are often physically implausible. The most common artifacts include jerky motions, feet that slide when in contact with the ground (or float above it), and out-of-plane rotations that violate balance.

The problem is, in part, due to the relatively small amount of available training data, and, in part, due to the limited ability of such models to generalize well beyond the training data. For example, a model trained on walking with a short stride may have difficulty tracking and reconstructing the motion of someone walking with a long stride or at a very different speed. Indeed, human motion depends significantly on a wide variety of factors including speed, step length, ground slope, terrain variability, ground friction, and variations in body mass distributions. The task of gathering enough motion capture data to span all these conditions, and generalize sufficiently well, is prohibitive.

As an alternative to learned kinematic models, this paper advocates the use of *physics-based models*. We hypothesize that physics-based dynamics will lead to natural parameterizations of human motion. Dynamics also allows one to model interactions with the environment (such as ground contact and balance during locomotion), and it generalizes naturally to different speeds of locomotion, changes in mass distribution and other sources of variation. Modeling the underlying dynamics of motion should result in more accurate

tracking and produce more realistic motions which naturally obey essential physical properties of human motion.

In this paper, we consider the important special case of walking. Rather than attempting to model full-body dynamics, our approach is inspired by simplified biomechanical models of human locomotion (Collins and Ruina 2005; Collins et al. 2001; Kuo 2001; McGeer 1992). Such models are low-dimensional and exhibit stable human-like gaits with realistic ground contact. We design a generative model for people tracking that comprises one such model, called the *Anthropomorphic Walker* (Kuo 2001, 2002), with a stochastic controller to generate muscle forces, and a higher-dimensional kinematic model conditioned on the low-dimensional dynamics.

Tracking is performed by simulating the model in a particle filter, producing physically plausible estimates of human motion for the torso and lower body. In particular, we demonstrate stable monocular tracking over long walking sequences. The tracker handles occlusion, varying gait styles, and turning, producing realistic 3D reconstructions. With lower-body occlusions, it still produces realistic reconstructions and infers the time and location of ground contacts. We also applied the tracker to the benchmark HumanEva dataset and report quantitative results.

2 Related Work

The 3D estimation of human pose from monocular video is often poorly constrained, and, hence, prior models play a central role in mitigating problems caused by ambiguities, occlusion and measurement noise. Most human pose trackers rely on *articulated kinematic models*. Early generative models were specified manually (e.g., with joint limits and smoothness constraints), while many recent generative models have been learned from motion capture data of people performing specific actions (e.g., Choo and Fleet 2001; Herda et al. 2005; Pavlović et al. 1999; Sidenbladh et al. 2000; Sminchisescu and Jepson 2004; Urtasun et al. 2006; Wachter and Nagel 1999). Discriminative models also depend strongly on human motion capture data, based on which direct mappings from image measurements to human pose and motion are learned (Agarwal and Triggs 2006; Elgammal and Lee 2004; Rosales et al. 2001; Shakhnarovich et al. 2003; Sminchisescu et al. 2007).

In constrained cases, kinematic model-based trackers can produce good results. However, such models generally suffer from two major problems. First, they often make unrealistic assumptions; e.g., motions are assumed to be smooth (which is violated at ground contact), and independent of global position and orientation. As a result, tracking algorithms exhibit a number of characteristic errors, including rotations of the body that violate balance, and *footskate*, in which a foot in contact with the ground appears to slide or float in space. Second, algorithms that learn kinematic models have difficulty generalizing beyond the training data. In essence, such models describe the probability of a motion by comparison to training poses; i.e., motions "similar" to the training data are considered likely. This means that, for every motion to be tracked, there must be a similar motion in the training database. In order to build a general tracker using current methods, an enormous database of human motion capture will be necessary.

To cope with the high dimensionality of kinematic models and the relative sparsity of available training data, a major theme of recent research on people tracking has been dimensionality reduction (Elgammal and Lee 2004; Rahimi et al. 2005; Sminchisescu and Jepson 2004; Urtasun et al. 2005, 2006). It is thought that low-dimensional models are less likely to over-fit the training data and will therefore generalize better. They also reduce the dimension of the state estimation problem during tracking. Inspired by similar ideas, our physics-based model is a low-dimensional abstraction based on biomechanical models. Such models are known to accurately represent properties of human locomotion (such as gait variation and ground contact) that have not been demonstrated with learned models (Blickhan and Full 1993; Full and Koditschek 1999; Kuo 2001). We thus aim to gain the advantages of a physics-based model without the complexity of full-body dynamics, and without the need for inference in a high-dimensional state space.

A small number of authors have employed physics-based models of motion for tracking. Pentland and Horowitz (1991) and Metaxas and Terzopoulos (1993) describe elastic solid models for tracking in conjunction with Kalman filtering, and give simple examples of articulated tracking by enforcing constraints. Wren and Pentland (1998) use a physics-based formulation of upper body dynamics to track simple motions using binocular inputs. For these tracking problems, the dynamics are relatively smooth but high-dimensional. In contrast, we employ a model that captures the specific features of walking, including the nonlinearities of ground contact, without the complexity of modeling elastic motion. Working with 3D motion capture data and motivated by abstract passive-dynamic models of bipedal motion, Bissacco (2005) uses a switching, linear dynamical system to model motion and ground contact. We note that, despite these attempts, the on-line tracking literature has largely shied away from physics-based prior models. We suspect that this is partly due to the perceived difficulty in building appropriate models. We show that, with judicious choice of representation, building such models is indeed possible.

It is also notable that the term "physics-based models" is used in different ways in computer vision. Among these,

physics is often used as a metaphor for minimization, by applying virtual "forces" (e.g., Chan et al. 1994; Delamarre and Faugeras 2001; Kakadiaris and Metaxas 2000; Kass et al. 1987; Terzopoulos and Metaxas 1990); unlike in our work, these forces are not meant to represent forces in the world.

Physics-based models of human motion are also common in computer animation where two main approaches have been employed. The Spacetime Constraints approach (Witkin and Kass 1988) solves for a minimal-energy motion that satisfies animator-specified constraints, and has recently shown some success at synthesizing full-body human motion (Liu et al. 2005; Safonova et al. 2004). However, such batch optimization is unsuitable for online tracking. Controller-based methods (e.g., Hodgins et al. 1995; Yin et al. 2007) employ on-line control schemes for interaction with physical environments. Our control mechanism is similar, but we use a minimal motion model with stochastic control for probabilistic 3D tracking. Finally, the model we develop is perhaps most similar to motion editing methods where low-dimensional physical constraints (Kovar et al. 2002; Popović and Witkin 1999; Shin et al. 2003) are applied to a high-dimensional kinematic model. Here we do not require example data to be transformed, and it is important to note that for tracking we do not need a fully-realistic dynamical model.

3 Motivation and Overview

Our primary goal is to track human locomotion from monocular video sequences. We employ a probabilistic formulation which requires a prior density model over human motion and an image likelihood model. The key idea, as discussed above, is to exploit basic physical principles in the design of a prior probabilistic model.

One natural approach is to model full-body dynamics as is sometimes done in humanoid robotics and computer animation. Unfortunately, managing the dynamics of full-body human motion, like the control of complex dynamical systems in general, is extremely challenging. Nonetheless, work in biomechanics and robotics suggests that the dynamics of bipedal walking may be well described by relatively simple *passive-dynamic walking* models. Such models exhibit stable, bipedal walking as a natural limit cycle of their dynamics. Early models, such as those introduced by McGeer (1990a), were entirely passive and could walk downhill solely under the force of gravity. Related models have since been developed, including one with a passive knee (McGeer 1990b), another with an upper body (Wisse et al. 2007), and one capable of running (McGeer 1992).

More recently, powered walkers based on passive-dynamic principles have been demonstrated to walk stably on level-ground (Collins et al. 2005; Kuo 2001, 2002). These models exhibit human-like gaits and energy-efficiency. The energetics of such models have also been shown to accurately predict the preferred relationship between speed and step-length in human walking (Kuo 2001). In contrast, traditional approaches in robotics (e.g., as used by Honda's humanoid robot *Asimo*), employ highly-conservative control strategies that are significantly less energy-efficient and less human-like in appearance, making them a poor basis for modeling human walking (Collins et al. 2005; Pratt 2000).

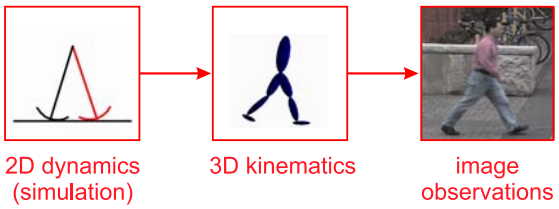


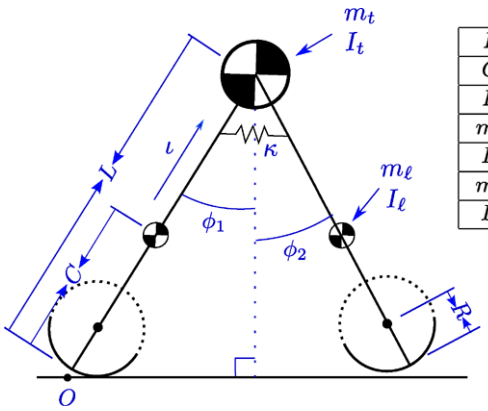
Fig. 1 A cartoon outline of the graphical model used for visual tracking. Conditioned on the control parameters one can simulate the equations of motion for the planar model to produce a sequence of 2D poses. The 3D kinematic model is conditioned on the 2D dynamics simulation. The image likelihood function then specifies the dependence of the image measurements on the kinematic pose

These issues motivate the form of the model sketched in Fig. 1, the components of which are outlined below.

Dynamical Model Our walking model is based on the *Anthropomorphic Walker* (Kuo 2001, 2002), a planar model of human locomotion (Sect. 4.1). The model depends on active forces applied to determine gait speed and step length. A prior distribution over these control parameters, together with the physical model, defines a distribution over planar walking motions (Sect. 4.2).

Kinematic Model The dynamics represent the motion of the lower body in the sagittal plane. As such it does not specify all the parts of the human body that we wish to track. We therefore define a 3D *kinematic model* for tracking (see Fig. 1). As described in Sect. 4.3, the kinematic model is constrained to be consistent with the planar dynamics, and to move smoothly in its remaining degrees of freedom (DOF).

Image Likelihood Conditioned on 3D kinematic state, the likelihood model specifies an observation density over image measurements. For tracking we currently exploit foreground and background appearance models as well as optical flow measurements (explained in Sect. 5.1). With the prior generative model and the likelihood, tracking is accomplished with a form of sequential Monte Carlo inference.



L	1.0m
C	0.645m
R	0.3m
m_t	0.678
I_t	0.167
m_ℓ	0.161
I_ℓ	0.017

Fig. 2 The planar Anthropomorphic Walker and inertial parameters. The model parameters in the table are taken from Kuo (2002). Units of mass are given as a proportion of the total mass of the walker

4 Dynamic Model of Human Walking

Our stochastic walking model is inspired by the minimally-powered *Anthropomorphic Walker* of Kuo (2001, 2002). Shown in Fig. 2, the Anthropomorphic Walker is a planar abstraction with two straight legs of length L and a rigid torso attached at the hip with mass m_t and moment of inertia I_t. The "feet" are circles of radius R, which roll along the ground as the model moves. Each leg has mass m_ℓ and moment of inertia I_ℓ, centered at distance C from the foot. The origin of the global frame of reference is defined to be the ground contact point of the stance foot when the stance leg is vertical.

The legs are connected by a torsional spring to simulate muscle torques at the hips. The spring stiffness is denoted κ. During normal walking, the *stance leg* is in contact with the ground, and the *swing leg* swings freely. The walker also includes an impulsive "toe-off" force, with magnitude ι, that allows the back leg to push off as support changes from the stance foot to the swing foot.

4.1 Dynamics

As in a Lagrangian formulation, we define generalized coordinates representing the configuration of the walker at a given instant: $\mathbf{q} = (\phi_1, \phi_2)^{\mathrm{T}}$, where ϕ_1 and ϕ_2 are the global orientations of the stance and swing legs, respectively. The state of the walker is given by $(\mathbf{q}, \dot{\mathbf{q}})$, where the generalized velocities are $\dot{\mathbf{q}} \equiv \frac{d\mathbf{q}}{dt}$. The equations of motion during normal walking are then written as a function of the current state:

$$\mathcal{M}(\mathbf{q})\ddot{\mathbf{q}} = \mathcal{F}(\mathbf{q}, \dot{\mathbf{q}}, \kappa) \tag{1}$$

where $\mathcal{M}(\mathbf{q})$ is known as the generalized mass matrix, $\mathcal{F}(\mathbf{q}, \dot{\mathbf{q}}, \kappa)$ is a generalized force vector which includes gravity and the spring force between the legs, and κ denotes the spring stiffness. This equation is a generalization of Newton's Second Law of Motion. Solving (1) at any instant gives the generalized acceleration $\ddot{\mathbf{q}}$. The details of (1) are given in Appendix A.

An important feature of walking is the collision of the swing leg with the ground. The Anthropomorphic Walker treats collisions of the swing leg with the ground plane as impulsive and perfectly inelastic. As a consequence, at each collision, all momentum of the body in the direction of the ground plane is lost, resulting in an instantaneous change in velocity. Our collision model also allows for the characteristic "toe-off" of human walking, in which the stance leg gives a small push before swinging. By changing the instantaneous velocity of the body, toe-off helps to reduce the loss of momentum upon ground contact.

The dynamics at ground collisions, as explained in Appendix B, are based on a generalized conservation of momentum equation which relates pre- and post-collision velocities of the body, denoted $\dot{\mathbf{q}}^-$ and $\dot{\mathbf{q}}^+$, and the magnitude of the impulsive toe-off, ι; i.e.,

$$\mathcal{M}^+(\mathbf{q})\dot{\mathbf{q}}^+ = \mathcal{M}^-(\mathbf{q})\dot{\mathbf{q}}^- + \mathcal{I}(\mathbf{q}, \iota) \tag{2}$$

where $\mathbf{q}$ is the pose at the time of collision, $\mathcal{M}^-(\mathbf{q})$ and $\mathcal{M}^+(\mathbf{q})$ are the pre- and post-collision generalized mass matrices, and $\mathcal{I}(\mathbf{q}, \iota)$ is the change in generalized momentum due to the toe-off force. The impulsive toe-off force depends on the angle at which the swing foot strikes the ground and on magnitude of the impulse, ι.

Given κ and ι, the dynamics equations of motion (1) can be simulated using a standard ODE solver. We use a fourth-order Runge-Kutta method with a step-size of $\frac{1}{30}$ s. When a collision of the swing foot with the ground is detected, we switch the roles of the stance and swing legs (e.g., we swap ϕ_1 and ϕ_2), and then use (2) to solve for the post-collision velocities. The simulation is then restarted from this post-collision state.

4.2 Control

The walking model has two control parameters $\theta = (\kappa, \iota)$, where κ is the spring stiffness and ι is the magnitude of the impulsive toe-off. Because these parameters are unknown prior to tracking, they are treated as hidden random variables. For effective tracking, we desire a prior distribution over θ which, together with the dynamical model, defines a distribution over motions. A gait may then be generated by sampling θ and simulating the dynamics.

One might learn a prior over θ by fitting the Anthropomorphic Walker to human mocap data of people walking with different styles, speeds, step-lengths, etc. This is challenging, however, as it requires a significant amount of mocap data, and the mapping from 3D kinematic description used for the mocap to the abstract 2D planar model is not

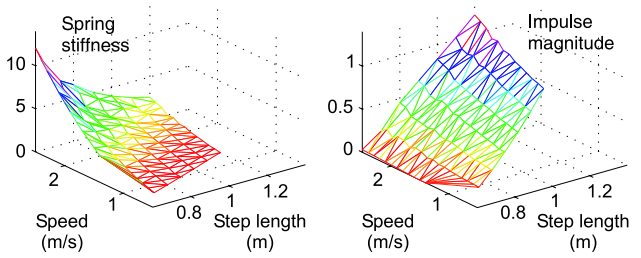


Fig. 3 Optimal stiffness κ (*left*) and impulse magnitude ι (*right*) as functions of speed and step length are shown. These plots illustrate the flexibility and expressiveness of the model's control parameters. Parameters were found by searching for cyclic motions with the desired speed and step length

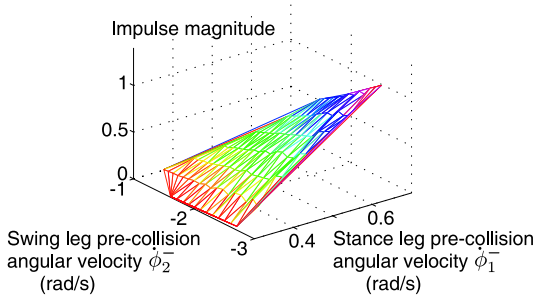


Fig. 4 Impulse magnitude ι of the optimal cyclic gaits plotted versus pre-collision velocities $\dot{\mathbf{q}}^- = (\dot{\phi}_1^-, \dot{\phi}_2^-)$. During tracking, a bilinear fit to the data shown here is used to determine the conditional mean for a Gamma density over ι at the beginning of each stride

obvious. Rather, we take a simpler approach motivated by the principle that walking motions are characterized by stable, cyclic gaits. Our prior over θ then assumes that likely control parameters lie in the vicinity of those that produce cyclic gaits.

Determining Cyclic Gaits The first step in the design of the prior is to determine the space of control parameters that generate cyclic gaits spanning the natural range of human walking speeds and step-lengths. This is readily formulated as an optimization problem. For a given speed and step-length, we seek initial conditions $(\mathbf{q}_0, \dot{\mathbf{q}}_0)$ and parameters θ such that the simulated motion ends in the starting state. The initial pose $\mathbf{q}_0$ can be directly specified since both feet must be on the ground at the desired step-length. The simulation duration T can determined by the desired speed and step-length. We then use Newton's method to solve

$$\mathcal{D}(\mathbf{q}_0, \dot{\mathbf{q}}_0, \theta, T) - (\mathbf{q}_0, \dot{\mathbf{q}}_0) = 0, \tag{3}$$

for $\dot{\mathbf{q}}_0$ and θ where $\mathcal{D}$ is a function that simulates the dynamics for duration T given an initial state $(\mathbf{q}_0, \dot{\mathbf{q}}_0)$ and parameters θ. The necessary derivatives are computed using finite differences. In practice, the solver was able to obtain control parameters satisfying (3) up to numerical precision for the tested range of speeds and step-lengths.

Solving (3) for a discrete set of speeds and step-lengths produces the control parameters shown in Fig. 3. These plots show optimal control parameters for the full range of human walking speeds, ranging from 2 to 7 km/h, and for a wide range of step-lengths, roughly 0.5–1.2 m. In particular, note that the optimal stiffness and impulse magnitudes depend smoothly on the speed and step-length of the motion. This is important as it indicates that the Anthropomorphic Walker is reasonably stable. To facilitate the duplication of our results, we have published Matlab code which simulates the model, along with solutions to (3), at http://www.cs.toronto.edu/~mbrubake/permanent/awalker.

Stochastic Control To design a prior distribution over walking motions for the Anthropomorphic Walker, we assume noisy control parameters that are expected to lie in the vicinity of those that produce cyclic gaits. We further assume that speed and step-length change slowly from stride to stride. Walking motions are obtained by sampling from the prior over the control parameters and then performing deterministic simulation using the equations of motion.

We assume that the magnitude of the impulsive toe-off force, $\iota > 0$, follows a Gamma distribution. For the optimal cyclic gaits, the impulse magnitude was very well fit by a bilinear function $\mu_\iota(\dot{\mathbf{q}}^-)$ of the two pre-collision velocities $\dot{\mathbf{q}}^-$ (see Fig. 4). This fit was performed using least-squares regression with the solutions to (3). The parameters of the Gamma distribution are set such that the mean is $\mu_\iota(\dot{\mathbf{q}}^-)$ and the variance is 0.05^2.

The unknown spring stiffness at time t, κ_t, is assumed to be nearly constant throughout each stride, and to change slowly from one stride to the next. Accordingly, within a stride we define κ_t to be Gaussian with constant mean $\bar{\kappa}$ and variance σ_κ^2:

$$\kappa_t \sim \mathcal{N}(\bar{\kappa}, \sigma_\kappa^2) \tag{4}$$

where $\mathcal{N}(\mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2. Given the mean stiffness for the ith stride, the mean stiffness for the next stride $\bar{\kappa}^{(i+1)}$ is given by

$$\bar{\kappa}^{(i+1)} \sim \mathcal{N}(\beta\mu_\kappa + (1-\beta)\bar{\kappa}^{(i)}, \sigma_{\bar{\kappa}}^2) \tag{5}$$

where μ_κ is a global mean spring stiffness and β determines how close $\bar{\kappa}^{(i)}$ remains to μ_κ over time. We use $\beta = 0.85$, $\sigma_\kappa^2 = 1.0$, $\mu_\kappa = 0.7$ and $\sigma_{\bar{\kappa}}^2 = 0.5$.

During tracking, $\bar{\kappa}$ does not need to be explicitly sampled. Instead, using a form of Rao-Blackwellization (Doucet et al. 2000; Khan et al. 2004), $\bar{\kappa}$ can be analytically marginalized out. Then, only the sufficient statistics of the resulting Gaussian distribution over $\bar{\kappa}$ needs to be maintained for each particle.

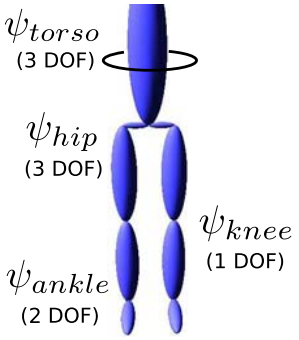


Fig. 5 The 3D kinematic model is conditioned on the 2D planar dynamics of the Anthropomorphic Walker

Because the walking model is very stable, the model is relatively robust to the choice of stochastic control. Other controllers may work just as well or better.

4.3 Conditional Kinematics

The model above is low-dimensional, easy to control, and produces human-like gaits. Nevertheless, it is a planar model, and hence it does not specify pose parameters in 3D. Nor does it specify all parameters of interest, such as the torso, knees and feet. We therefore add a higher-dimensional 3D kinematic model, conditioned on the underlying dynamics. The coupling of a simple physics-based model with a detailed kinematic model is similar to Popović and Witkin's (1999) physics-based motion editing system.

The kinematic model, depicted in Fig. 5, has legs, knees, feet and a torso. It has ball-and-socket joints at the hips, a hinge joint for the knees and 2 DoF joints for the ankles. Although the upper body is not used in the physics model, it provides useful features for tracking. The upper body in the kinematic model comprises a single rigid body attached to the legs.

The kinematic model is constrained to match the dynamics at every instant. In effect, the conditional distribution of these kinematic parameters, given the state of the dynamics, is a delta function. Specifically, the upper-leg orientations of the kinematic model in the sagittal plane are constrained to be equal to the leg orientations in the dynamics. The ground contact of stance foot in the kinematics and rounded "foot" of the dynamics are also forced to be consistent. In particular, the foot of the stance leg is constrained to be in contact with the ground. The location of this contact point on the foot rolls along the foot proportional to the arc-length with which the dynamics foot rolls forward during the stride.

When the simulation of the Anthropomorphic Walker predicts a collision, the stance leg, and thus the contact constraint, switches to the other foot. If the corresponding foot of the kinematic model is far from the ground, applying this constraint could cause a "jump" in the pose of the kinematic model. However, such jumps are generally inconsistent with image data and are thus not a significant concern. In general, this discontinuity would be largest when the knee is very bent, which does not happen in most normal walking. Because the Anthropomorphic Walker lacks knees, it is unable to handle motions which rely on significant knee bend during contact, such as running and walking on steep slopes. We anticipate that using a physical model with more degrees-of-freedom should address this issue.

Table 1 The parameters of the conditional kinematic model used in tracking. The degrees of freedom not listed (Hip X) are constrained to be equal to that of the Anthropomorphic Walker

Joint	Axis	α[a]	k	$\bar{\psi}$	σ	$(\psi^{\min}, \psi^{\max})$
Torso	Side	0.9	5	0	25	$(-\infty, \infty)$
	Front	0.9	5	0	25	$(-\infty, \infty)$
	up	0.75	0	0	300	$(-\infty, \infty)$
Hip	Front	0.5	5	0	50	$(-\frac{\pi}{8}, \frac{\pi}{8})$
	up	0.5	5	0	50	$(-\frac{\pi}{8}, \frac{\pi}{8})$
Stance knee	Side	0.75	20	0	50	$(0, \pi)$
Swing knee	Side	0.9	15	[b]	300	$(0, \pi)$
Ankle	Side	0.9	50	0	50	$(-\frac{\pi}{8}, \frac{\pi}{8})$
	Front	0.9	50	0	50	$(-\frac{\pi}{8}, \frac{\pi}{8})$

[a]Values of α shown here are for $\Delta t = \frac{1}{30}$ s. For $\Delta t = \frac{1}{60}$ s, the square roots of these values are used

[b]$\bar{\psi}_{\text{swing knee}}$ is handled specially, see text for more details

Each remaining kinematic DOF $\psi_{j,t}$ is modeled as a smooth, 2nd-order Markov process:

$$\psi_{j,t} = \psi_{j,t-1} + \Delta t \alpha_j \dot{\psi}_{j,t-1} + \Delta t^2 (k_j(\bar{\psi}_j - \psi_{j,t-1})) + \eta_j) \quad (6)$$

where Δt is the size of the timestep, $\dot{\psi}_{j,t-1} = (\psi_{j,t-1} - \psi_{j,t-2})/\Delta t$ is the joint angle velocity, and η_j is IID Gaussian noise with mean zero and variance σ_j^2. This model is analogous to a damped spring model with noisy accelerations where k_j is the spring constant, $\bar{\psi}_j$ is the rest position, α_j is related to the damping constant and η_j is noisy acceleration. Joint limits which require that $\psi_j^{\min} \leq \psi_j \leq \psi_j^{\max}$ are imposed where appropriate and η_j is truncated (Robert 1995) to satisfy the joint limits.

The joint evolution parameters $\alpha, k, \bar{\psi}$ and σ^2 are fixed to the values shown in Table 1, with the exception of the knee rest position of the swing leg. Due to the sharp bend in the knee immediately after toe-off, a simple smoothness prior has difficulty modelling this joint. To account for this, we define $\bar{\psi}_{\text{swing knee}} = 5\psi_{\text{hip}}$ where ψ_{hip} is the sagittal angle between the two legs. This encourages a bent knee at the beginning of a stride and a straight knee towards the end of a stride.

It is interesting to note that, while most existing methods for people tracking rely heavily on learned models from motion capture data, our model does not use any motion capture data. However, it is clear that the kinematic model in general, and of the knee in particular, is crude, and could be improved greatly with learning, as could other aspects of the model.

5 Sequential Monte Carlo Tracking

Pose tracking is formulated with a state-space representation. The state $\mathbf{s}_t$ at time t comprises dynamics parameters, $\mathbf{d}_t$, and the kinematic DOFs, $\mathbf{k}_t$; i.e., $\mathbf{s}_t = (\mathbf{d}_t, \mathbf{k}_t)$. The dynamics parameters comprises 2 continuous joint angles and their angular velocities, a binary variable to specify the stance foot, and two variables for the sufficient statistics for the mean spring stiffness as described at the end of Sect. 4.2. The kinematic state comprises 3 DOFs for the global torso position, 3 DOFs for global torso orientation, and 12 DOFs for remaining joint angles. Note that, while the dynamics contain the joint angles and angular velocities of the Anthropomorphic Walker, they are deterministic given the previous state and current control parameters. In essence, inference is done over the control parameters in lieu of the pose parameters.

With the Markov properties of the generative model given in Sect. 4, and conditional independence of the measurements, one can write the posterior density over motions recursively;

$$p(\mathbf{s}_{1:t}|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{s}_{t-1})p(\mathbf{s}_{1:t-1}|\mathbf{z}_{1:t-1}) \tag{7}$$

where $\mathbf{s}_{1:t} \equiv [\mathbf{s}_1, \ldots, \mathbf{s}_t]$ denotes a state sequence, $\mathbf{z}_{1:t} \equiv [\mathbf{z}_1, \ldots, \mathbf{z}_t]$ denotes the observation history, $p(\mathbf{z}_t|\mathbf{s}_t)$ is the observation likelihood, and $p(\mathbf{s}_t|\mathbf{s}_{t-1})$ is derived from the generative model in Sect. 4.

By the definition of the generative model, the temporal state evolution can be factored further; i.e.,

$$p(\mathbf{s}_t|\mathbf{s}_{t-1}) = p(\mathbf{k}_t|\mathbf{d}_t, \mathbf{k}_{t-1})p(\mathbf{d}_t|\mathbf{d}_{t-1}). \tag{8}$$

Here $p(\mathbf{d}_t|\mathbf{d}_{t-1})$ is the stochastic dynamics of the Anthropomorphic Walker described in Sects. 4.1 and 4.2 and $p(\mathbf{k}_t|\mathbf{d}_t, \mathbf{k}_{t-1})$ is the conditional kinematic model explained in Sect. 4.3. Thus, to sample from $p(\mathbf{s}_t|\mathbf{s}_{t-1})$, the dynamics state $\mathbf{d}_t$ is sampled according to $p(\mathbf{d}_t|\mathbf{d}_{t-1})$ and, conditioning on $\mathbf{d}_t$, the kinematic state $\mathbf{k}_t$ is then sampled from $p(\mathbf{k}_t|\mathbf{d}_t, \mathbf{k}_{t-1})$. The likelihood function and the inference procedure are described below.

5.1 Likelihood

The 3D articulated body model comprises a torso and lower limbs, each of which is modeled as a tapered ellipsoidal cylinder. The size of each part is set by hand, as is the pose of the model in the first frame of each sequence. To evaluate the likelihood $p(\mathbf{z}_t|\mathbf{s}_t)$, the 3D model is projected into the image plane. This allows self-occlusion to be handled naturally as the visibility of each part can be determined for each pixel. The likelihood is then based on appearance models for the foreground body and the background, and on optical flow measurements (Fleet and Weiss 2005).

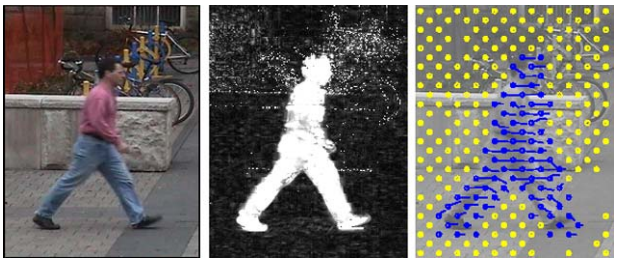


Fig. 6 (Color online) A cropped image (*left*) is shown with a example of the background negative log likelihood (*middle*), and a grid of motion trajectories (*blue/yellow* depict large/small speeds)

A background model, learned from a small subset of images, comprises mean color (RGB) and intensity gradients at each pixel and a single 5×5 covariance matrix (e.g., see Fig. 6 (middle)). The foreground model assumes that pixels are IID in each part (i.e., foot, legs, torso, head), with densities given by Gaussian mixtures over the same 5D measurements as the background model. Each mixture has 3 components and its parameters are learned from hand labeled regions in a small number of frames.

Optical flow is estimated at grid locations in each frame (e.g., see Fig. 6 (right)), using a robust M-estimator with non-overlapping regions of support. The eigenvalues/vectors of the local gradient tensor in each region of support provide a crude approximation to the estimator covariance Σ. For the likelihood of a flow estimate, $\mathbf{v}$, given the 2D motion specified by the state, $\mathbf{u}$, we use a heavy-tailed Student's t distribution (chosen for robustness). The log-likelihood is given by

$$\log p(\mathbf{v}|\mathbf{u}) = -\frac{\log|\Sigma|}{2} - \frac{n+2}{2}\log(1+e^2) + c \tag{9}$$

where $e^2 = \frac{1}{2}(\mathbf{v}-\mathbf{u})^{\mathrm{T}}\Sigma^{-1}(\mathbf{v}-\mathbf{u})$ and $n = 2$ is the degrees of freedom, and c is a constant. Because the camera is not moving in our image sequences, we define the log-likelihood of a flow measurement on the background as given by (9) with $\mathbf{u} = \mathbf{0}$.

The visibility of each part defines a partition of the observations, such that $\mathbf{z}_t(i)$ are the measurements which belong to part i. The background is simply treated as another part. Then the log-likelihood contribution of part i is

$$\log p(\mathbf{z}_t(i)|\mathbf{s}_t) = \sum_{\mathbf{m}\in\mathbf{z}_t(i)} \log p(\mathbf{m}|\mathbf{s}_t) \tag{10}$$

where the sum is over the measurements belonging to part i. To cope with large correlations between measurement errors, we define the appearance and flow log-likelihood to be the weighted sum of log-likelihoods over all visible measurements for each part

$$\log p(\mathbf{z}_t|\mathbf{s}_t) = \sum_i w_i \log p(\mathbf{z}_t(i)|\mathbf{s}_t) \tag{11}$$

where the weights are set inversely proportional to the expected size of each part in the image.[1] If multiple cameras are available, they are assumed to be conditionally independent given the state $\mathbf{s}_t$. This yields a combined log-likelihood of

$$\log p(\mathbf{z}_t^1, \mathbf{z}_t^2, \ldots|\mathbf{s}_t) = \sum_i \log p(\mathbf{z}_t^i|\mathbf{s}_t) \tag{12}$$

where $\mathbf{z}_t^i$ is the observation from camera i.

5.2 Inference

Using a particle filter, we approximate the posterior (7) by a weighted set of N samples $\mathcal{S}_t = \{\mathbf{s}_{1:t}^{(j)}, w_t^{(j)}\}_{j=1}^N$. Given the recursive form of (7), the posterior $\mathcal{S}_t$, given $\mathcal{S}_{t-1}$, can be computed in two steps; i.e.:

1. Draw samples $\mathbf{s}_t^{(j)} \sim p(\mathbf{s}_t|\mathbf{s}_{t-1}^{(j)})$ using (8) to form the new state sequences $\mathbf{s}_{1:t}^{(j)} = [\mathbf{s}_{1:t-1}^{(j)}, \mathbf{s}_t^{(j)}]$; and
2. Update the weights $w_t^{(j)} = c w_{t-1}^{(j)} p(\mathbf{z}_t|\mathbf{s}_t^{(j)})$, where c is used to normalize the weights so they sum to 1.

This approach, without re-sampling, often works well until particle depletion becomes a problem, i.e., where only a small number of weights are significantly non-zero. One common solution to this is to re-sample the states in $\mathcal{S}_t$ according to their weights. This is well-known to be suboptimal since it does not exploit the current observation in determining which states should be re-sampled (i.e., survive). Instead, inspired by the auxiliary particle filter (Pitt and Shephard 1999), we use future data to predict how well current samples are likely to fare in the future. This is of particular importance with a physics-based model, where the quality of a sample is not always immediately evident based on current and past likelihoods. For instance, the consequences of forces applied at the current time may not manifest until several frames into the future.

In more detail, we maintain an approximation $\mathcal{S}_{t:t+\tau} = \{\mathbf{s}_{t:t+\tau}^{(j)}, w_{t:t+\tau}^{(j)}\}_{j=1}^N$ to the marginal posterior distribution over state sequences in a small temporal window of $\tau + 1$ frames, $p(\mathbf{s}_{t:t+\tau}|\mathbf{z}_{1:t+\tau})$. The sample set is obtained by simulating the model for $\tau + 1$ time steps, given $\mathcal{S}_{t-1}$, evaluating the likelihood of each trajectory and setting

$$w_{t:t+\tau}^{(j)} = c w_{t-1}^{(j)} \prod_{\ell=t}^{t+\tau} p(\mathbf{z}_\ell|\mathbf{s}_\ell^{(j)}) \tag{13}$$

where c is set such that the weights sum to one.

Following Doucet et al. (2000) and Kong et al. (1994), when the effective number of samples,

$$N_{\text{eff}} = \left(\sum_j (w_{t:t+\tau}^{(j)})^2\right)^{-1}, \tag{14}$$

becomes too small we re-sample $\mathcal{S}_{t-1}$ using importance sampling; i.e.:

1. Draw samples $\mathbf{s}_{t-1}^{(k)}$ from the weights $\{\hat{w}_{t-1}^{(j)}\}_{j=1}^N$ where $\hat{w}_{t-1}^{(j)} = (1-\gamma) w_{t-1}^{(j)} + \gamma w_{t:t+\tau}^{(j)}$ and γ represents our trust in our approximation $\mathcal{S}_{t:t+\tau}$;
2. Set the new weights to be $w_{t-1}^{(k)} / \hat{w}_{t-1}^{(k)}$, and then normalize the weights so they sum to 1.

The importance re-weighting (step 2) is needed to maintain a properly weighted approximation to the posterior (7). Below we use $\tau = 3$ and $\gamma = 0.9$. With this form of importance sampling, resampling occurs once every 4 or 5 frames on average for the experiments below.

6 Results

Here we present the results of four experiments with our model. The first three experiments use the same set of parameters for the kinematic evolution and the same prior over the control parameters for the dynamics. The parameters for the fourth experiment were set to similar values, but adjusted to account for a difference in frame rate (30 frames per second for experiments one through three and 60 frames per second for experiment four). These parameters were empirically determined. Finally, for each image sequence, we determine the camera intrinsics and extrinsics with respect to a world coordinate frame on the ground plane based on 10–12 correspondences between image locations and ground truth 3D locations in each scene. The direction of gravity is assumed to be normal to the ground plane.

All experiments used 5000 particles, with resampling when $N_{\text{eff}} < 500$. Experimentally we have found that, while as few as 1000 particles can result in successful tracking of some sequences (e.g., Experiment 1), 5000 particles was necessary to consistently track well across all experiments. Excluding likelihood computations, the tracker runs at around 30 frames per second. The body geometry was

[1] To avoid computing the log-likelihood over the entire image, we equivalently compute log-likelihood ratios of foreground versus background over regions of the image to which the 3D body geometry projects.

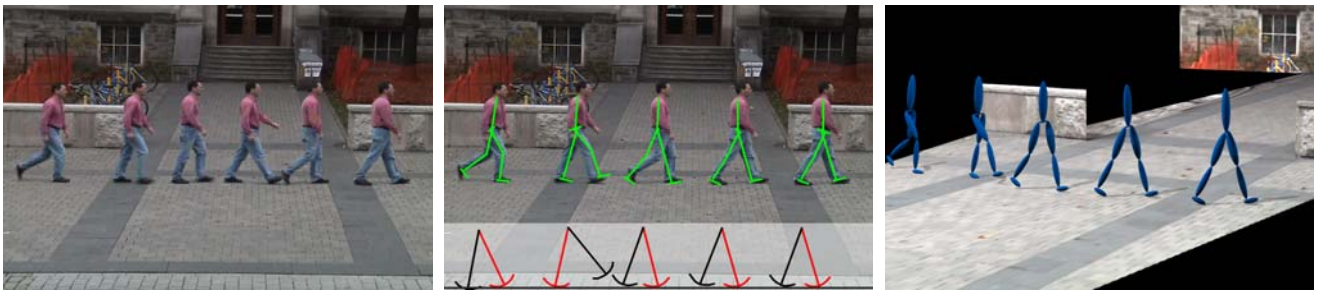


Fig. 7 (Color online) Composite images show the subject at several frames, depicting the motion over the 130 frame sequence: (*left*) the original images; (*middle*) the inferred poses of the MAP kinematics overlayed on the images, with the corresponding state of the Anthropomorphic Walker depicted *along the bottom* (the stance leg in *red*); (*right*) a 3D rendering of MAP poses from a different viewpoint

set by hand and the mean initial state was coarsely hand-determined. Initial particles were sampled with a large variance about that mean state. The inference procedure results in a set of particles that approximate the posterior distribution $p(\mathbf{s}_{1:t}|\mathbf{z}_{1:t})$ for a given time t. Our demonstration of the results will focus mainly on the *maximum a-posteriori* (MAP) trajectory of states over all T frames,

$$\mathbf{s}_{1:T}^{\mathrm{MAP}} = \arg\max_{\mathbf{s}_{1:T}} p(\mathbf{s}_{1:T}|\mathbf{z}_{1:T}). \tag{15}$$

This is crudely approximated by choosing the state sequence associated with the particle at time T with the largest weight. We present the MAP trajectory because it ensures that the sequence of poses is consistent with the underlying motion model.

Experiment 1: Changes in Speed Figure 7 (left) shows a composite image of a walking sequence in which the subject's speed decreases from almost 7 to 3 km/h. Figure 8 shows the recovered velocity of the subject over time in the solid blue curve. Also shown with the dashed green curve is the posterior probability of which leg is the stance leg. Such speed changes are handled naturally by the physics-based model. Figure 7 (middle) shows the recovered MAP trajectory from the original camera position while Fig. 7 (right) shows that the recovered motion looks good in 3D from other views.

Figure 9 shows cropped versions of tracking results for a short subsequence, demonstrating the consistency of the tracker. Weakness in the conditional kinematic model at high speeds leads to subtle anomalies, especially around the knees, which can be seen in the early frames of this subsequence.

Experiment 2: Occlusion We simulate occlusion by blacking out an image region as shown in Fig. 10. The silhouette of the lower body is therefore lost, and we discard all flow measurements that encroach upon the occluder. Nevertheless, the subtle motion of the torso is enough to track the person, infer foot positions, and recover 3D pose.

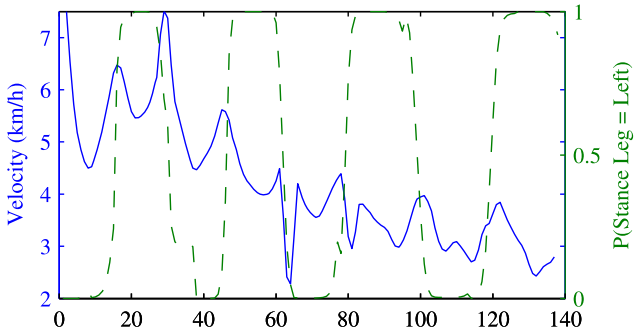


Fig. 8 (Color online) Inferred speed as a function of time for the MAP trajectory in Experiment 1 (*blue*). The *dashed green line* is $p(\text{stance leg} = \text{left}|\mathbf{z}_{1:t})$, the probability of the left leg being the stance leg given the data up to that frame

It is particularly interesting to examine the posterior distribution $p(\mathbf{s}_t|\mathbf{z}_{1:t})$ which can be seen in the bottom row of Fig. 11. These images show colour coded points for the head, hip, knees and feet for each particle in the posterior. The brightness of each point is proportional to its log weight. While there is increased posterior uncertainty during the occlusion, it does not diffuse monotonically. Rather, motion of the upper body allows the tracker to infer the stance leg and contact location. Notice that, soon after ground contacts, the marginal posterior over the stance foot position tends to shrink.

Finally, during occlusion, leg-switching can occur but is unlikely. This is visible in the posterior distribution as an overlap between yellow (right foot) and white (left foot) points. However, the ambiguity is quickly resolved after the occlusion.

Experiment 3: Turning While the Anthropomorphic Walker is a planar model we are still able to successfully track 3D walking motions because of the conditional kinematics. As can been seen in Fig. 14, the model successfully tracks the person through a sharp turn in a sequence of more than 400 frames. Despite the limitations of the physical model, it is able to accurately represent the dynamics of the motion

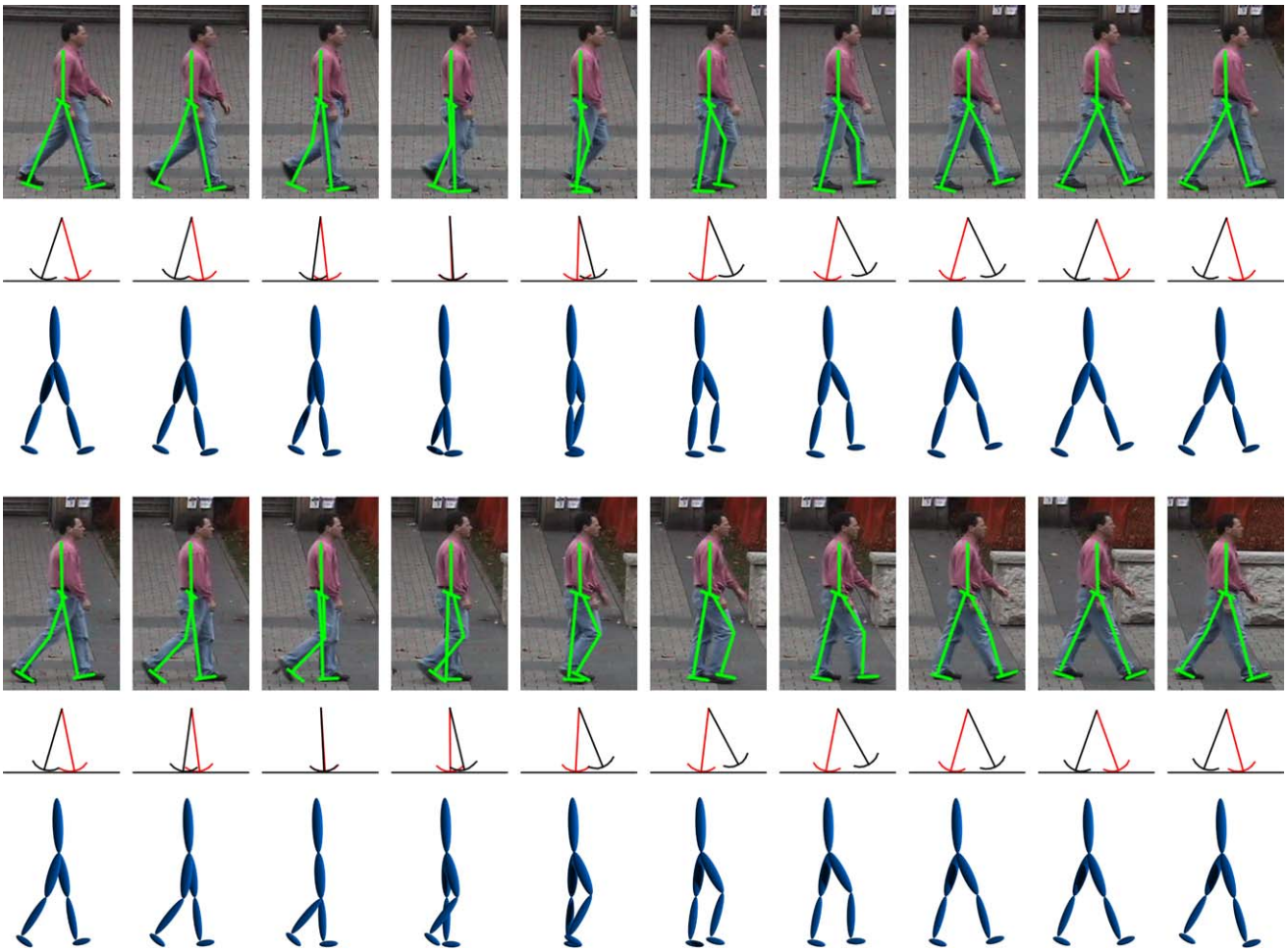


Fig. 9 (Color online) Two rows of cropped images showing every second frame of the MAP trajectory in Experiment 1 for two strides during change of speed: (*top*) the kinematic skeleton is overlayed on the subject; (*middle*) the corresponding state of the Anthropomorphic Walker is shown with the stance printed in *red*; (*bottom*) a 3D rendering of the kinematic state

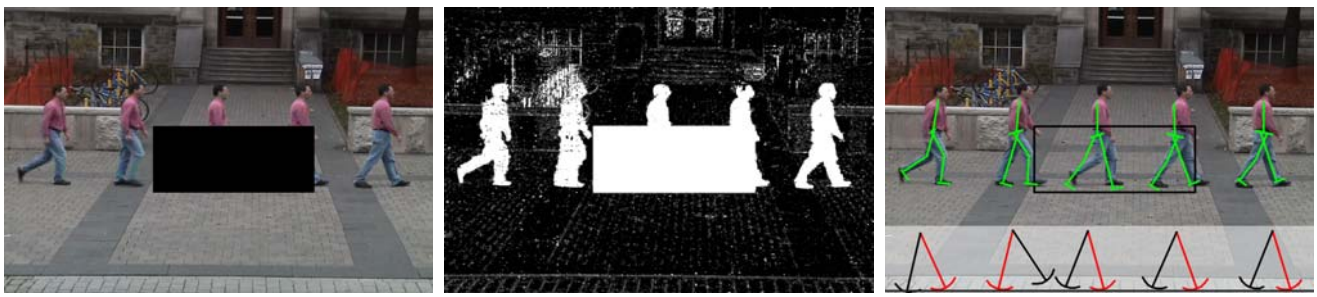


Fig. 10 Composite images show the input data (*left*), background model (*middle*) and MAP trajectory (*right*) at several frames for Experiment 2. Only the outline of the occluder is shown for illustration

in 2D while the conditional kinematic model represents the turning motion.

Figure 13 shows the speed of the subject and the posterior probability of which leg is the stance leg. Between frames 250 and 300 there is significant uncertainty in which leg is in contact with the ground. This is partly because, in these frames which correspond to the middle row in Fig. 14, there are few visual cues to disambiguate when a foot has hit the ground.

Experiment 4: HumanEva To quantitatively assess the quality of tracking, we also report results on the HumanEva

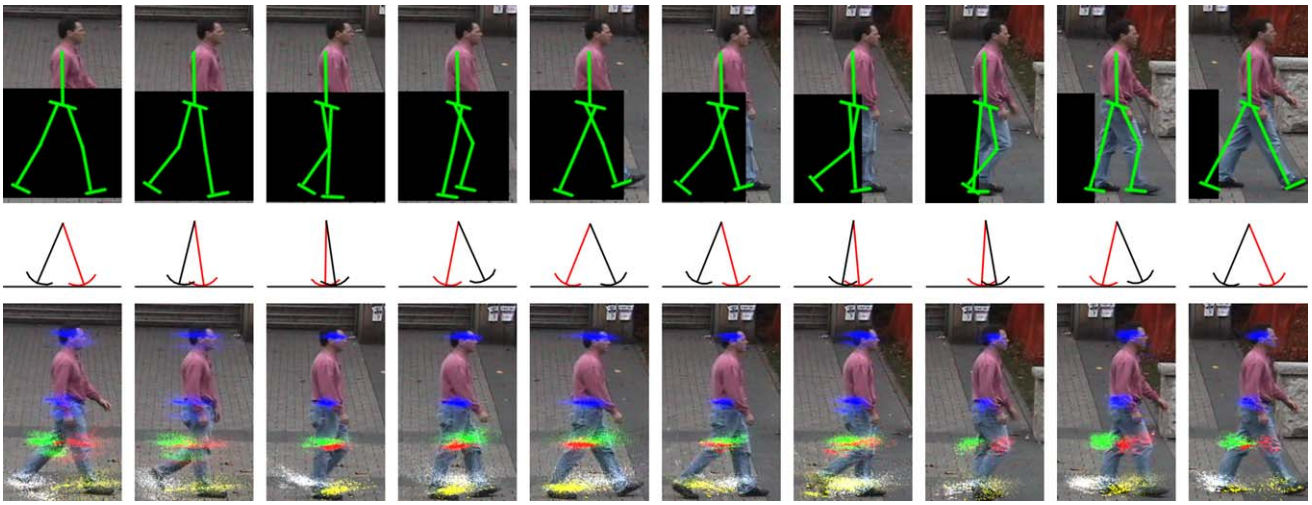


Fig. 11 (Color online) Cropped images showing every 4th frame of the MAP trajectory (*top*), the corresponding state of the Anthropomorphic walker (*middle*) and the posterior distribution (*bottom*) in Experiment 2. In the posterior points on the head (*blue*), left and right feet (*white* and *yellow*), left and right knees (*green* and *red*) and hip (*blue*) are plotted for each particle with intensity proportional to their log weight

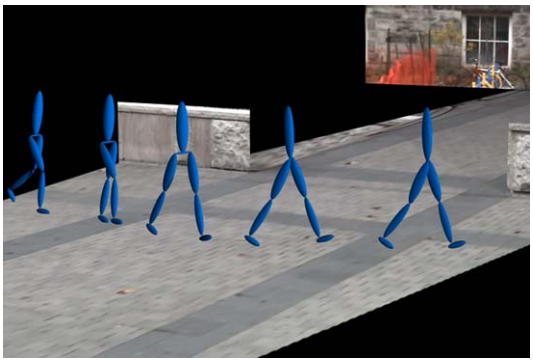


Fig. 12 3D rendering of the MAP trajectory in Experiment 2

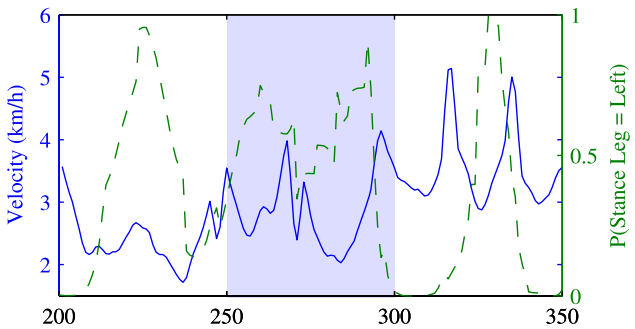


Fig. 13 (Color online) MAP trajectory velocity (*blue*) and stance leg posterior $p(\text{stanceleg} = \text{left}|\mathbf{z}_{1:t})$ (*dashed green*) for the times shown in Fig. 14. The highlighted region, corresponding to the middle row of Fig. 14, exhibits significant uncertainty about which leg is the stance leg

benchmark dataset (Sigal and Black 2006). This dataset contains multicamera video, synchronized with motion capture data that can be used as ground truth. Error is measured as the average Euclidean distance over a set of defined marker positions. Because our method does not actively track the head and arms, we report results using only the markers on the torso and legs.

As above, tracking was hand initialized and segment lengths were set based on the static motion capture available for each subject. The camera calibration provided with the dataset was used and it was assumed that the ground plane was located at $Z = 0$. We report monocular and binocular results on subjects 2 and 4 from HumanEva II. Error is measured from the poses in the MAP trajectory of states over all T frames. The results are summarized in Table 2 and errors over time are plotted in Figs. 15 and 16.

It is important to note that the same model (dynamics and kinematics) is used to track the two HumanEva subjects as well as the subject in the preceding experiments. Only the body size parameters were different. This helps to demonstrate that the model can generalize to different subjects.

In this paper, both relative and absolute 3D error measures are reported. Absolute error is computed as the average 3D Euclidean distance between predicted and ground truth marker positions (Sigal and Black 2006). Following HumanEva, relative error is computed by translating the pelvis of the resulting pose to the correct 3D position before measuring the 3D Euclidean distance. This removes gross errors in depth.

The type of error reported is significant, as different measures make meaningful comparisons difficult. Both error types are reported here to allow a more direct comparison with other methods. For example, relative error is often used by discriminative methods which do not recover absolute 3D depth.

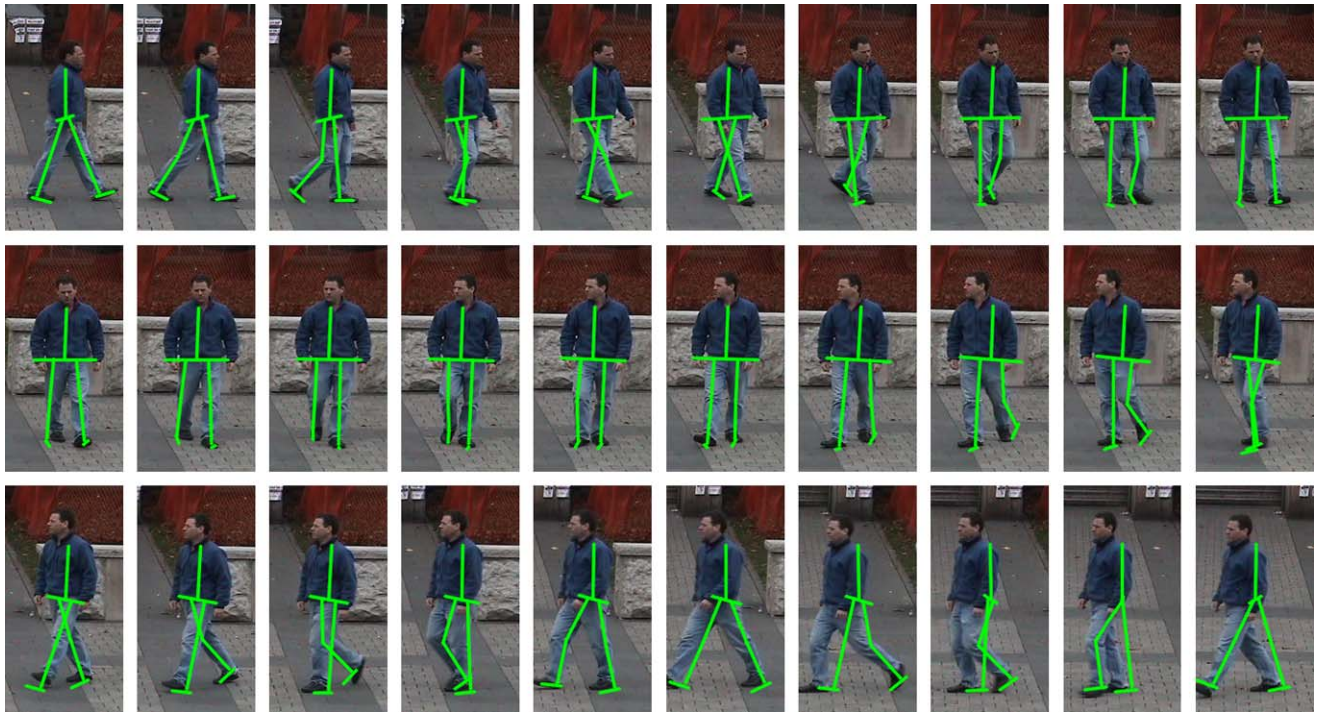


Fig. 14 (Color online) Cropped images showing every 5th frame of the MAP trajectory through an acceleration and sharp turn, starting at frame 200. The skeleton of the kinematic model is overlayed in *green*. The *middle row* corresponds to the shaded portion of Fig. 13

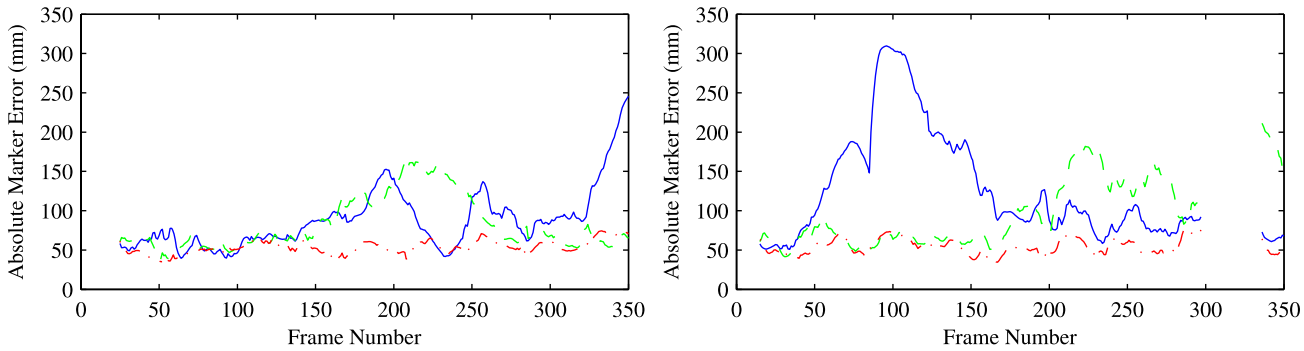


Fig. 15 (Color online) Average absolute marker error over time for subject 2, Combo 1 (*left*) and subject 4, Combo 4 (*right*). Plots are shown for monocular tracking with camera 2 (*solid blue*) and camera 3 (*dashed green*) as well as binocular tracking with cameras 2 and 3 (*dot-dashed red*)

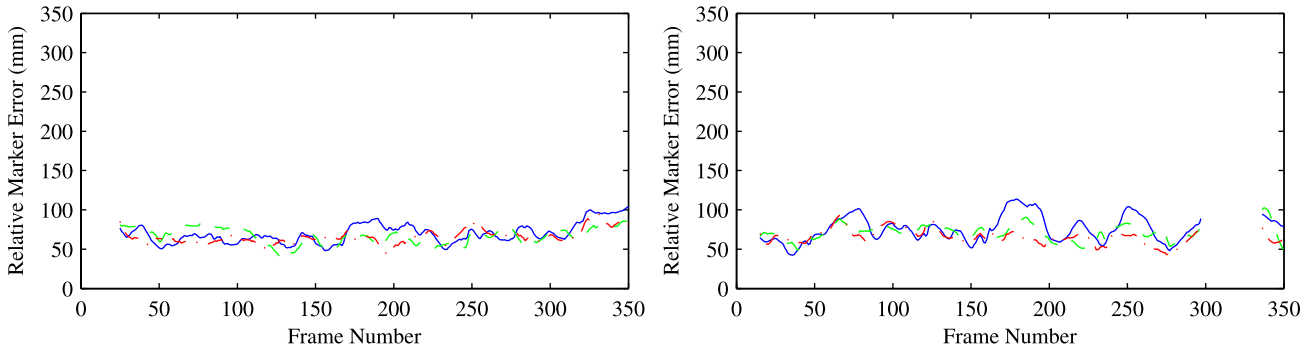
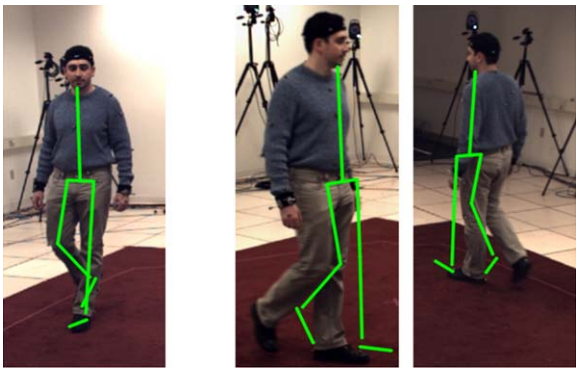


Fig. 16 (Color online) Average relative marker error over time for subject 2, Combo 1 (*left*) and subject 4, Combo 4 (*right*). Plots are shown for monocular tracking with camera 2 (*solid blue*) and camera 3 (*dashed green*) as well as binocular tracking with cameras 2 and 3 (*dot-dashed red*)

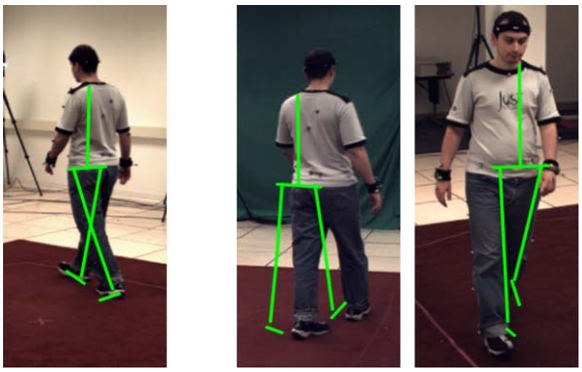
Table 2 Quantitative results on sequences from HumanEva II

Sequence	Error type	Monocular (camera 2)		Monocular (camera 3)		Binocular (cameras 2 and 3)	
		Median	Mean	Median	Mean	Median	Mean
Subject 2, Combo 1, Frames 25–350	absolute	82 mm	88 mm ± 38	67 mm	82 mm ± 34	52 mm	53 mm ± 9
	relative	67 mm	70 mm ± 13	67 mm	67 mm ± 11	64 mm	66 mm ± 9
Subject 4, Combo 4, Frames 15–350[a]	absolute	98 mm	127 mm ± 70	77 mm	96 mm ± 42	52 mm	54 mm ± 10
	relative	74 mm	76 mm ± 17	71 mm	70 mm ± 10	65 mm	66 mm ± 10

[a]As noted on the HumanEva II website, frames 298–335 are excluded from the calculation due to errors in the ground truth motion capture data



(a) Subject 2, Combo 1, Camera 3. The pose at frame 225 of the MAP trajectory is shown from camera 3 on the left. On the right are the views from cameras 2 and 4 respectively



(b) Subject 4, Combo 4, Camera 2. The pose at frame 125 of the MAP trajectory is shown from camera 2 on the left. On the right are the views from cameras 3 and 4 respectively

Fig. 17 Monocular tracking errors due to depth ambiguities. In both examples, the model appears to fit well in the view from which tracking is done. However, when viewed from other cameras the errors in depth become evident

The difference between the relative and absolute errors is also indicative of the nature of errors made by the tracker. Table 2 shows that, unsurprisingly, absolute errors are lower when using two cameras. In contrast, the plots in Fig. 16 suggest a negligible gain in relative error when using two cameras. Taken together, these results suggest that depth uncertainty remains the primary source of monocular tracking error. With these depth errors removed, the errors in binocular and monocular tracking are comparable.

This is further illustrated in Figs. 17(a) and 17(b) which show frames from the monocular trackers. The pose of the subject fits well in 2D and is likely to have a high likelihood at that frame. However, when viewed from other cameras, the errors in depth are evident.

Table 2 also reveals that relative error can be higher than absolute error, particularly for binocular tracking. This peculiar result can be explained with two observations. First, while relative error removes error from the pelvic marker, it may introduce error in other markers. Further, direct correspondences between positions on any articulated model and the virtual markers of the motion capture may not be possible as the motion capture models have significantly more degrees of freedom. These correspondence errors can then be magnified by the translation of the pelvic marker, particularly if there are errors in the pelvic marker itself.

Interestingly, the monocular tracking errors shown in Fig. 15 (the green and blue curves) tend to have significant peaks which fall off slowly with time. While evident in all experiments, this can be most clearly seen when tracking subject 4 from camera 2. These peaks are the combined result of depth uncertainty and a physically plausible motion model. According to the motion model, the only way the subject can move in depth is by walking there. If a foot is misplaced it cannot gradually slide to the correct position, rather the subject must take a step. This results in errors persisting over at least one stride. However, this is also the same behaviour which prevents footskate and ensures more realistic motions.

7 Discussion and Future Work

In this paper we showed that physics-based models offer significant benefits in terms of accuracy, stability, and generality for person tracking. Results on three different subjects in a variety of conditions, including in the presence of severe occlusion, are presented which demonstrate the ability of the

tracker to generalize. Quantitative results for monocular and binocular 3D tracking on the HumanEva dataset (Sigal and Black 2006) allows for direct comparison with other methods.

Here we used a simple powered walking model, but we are currently exploring more sophisticated physical models (Brubaker and Fleet 2008) which may yield even more general trackers for other types of motion. There will, generally, be a trade-off between model generality and the difficulty of designing a controller (Vondrak et al. 2008). We note that, while control of humanoid dynamical models is a challenging problem, there is a substantial literature in robotics and animation from which to draw inspiration.

Although our approach employs online Bayesian inference, it should also be possible to incorporate physical laws within other tracking frameworks such as discriminative methods. Models similar to this may also be used for modelling and tracking other animals (Full and Koditschek 1999).

Acknowledgements Thanks to Zoran Popović and Allan Jepson for valuable discussions. Thanks to Jack Wang for some initial software.

Appendix A: Equations of Motion

Here we describe the equations of motion for the *Anthropomorphic Walker*, shown in Fig. 2. While general-purpose physics engines may be used to implement the physical model and the impulsive collisions with the ground, most do not support exact ground constraints, but instead effectively require the use of springs to model static contact. In our experience it is not possible to make the springs stiff enough to accurately model the data without resulting in slow or unstable simulations. Hence, we derive equations of motion which exactly enforce static contact constraints. These equations produces stable simulations which allow (3) to be solved efficiently.

In order to derive the equations of motion for the walking model, we employ the TMT method (van der Linde and Schwab 2002), a convenient recipe for constrained dynamics. The TMT formulation is equivalent to Lagrange's equations of motion and can be derived in a similar way, using d'Alembert's Principle of virtual work (Goldstein et al. 2001). However, we find the derivation of equations of motion using the TMT method simpler and more intuitive for articulated bodies.

We begin by defining the kinematic transformation, which maps from the generalized coordinates $\mathbf{q} = (\phi_1, \phi_2)$ to a 6×1 vector that contains the linear and angular coordinates of each rigid body which specify state for the Newton-Euler equations of motion. The torso is treated as being rigidly connected to the stance leg and hence we have only two rigid parts in the Anthropomorphic Walker. The kinematic transformation can then be written as

$$\mathbf{k}(\mathbf{q}) = \begin{bmatrix} -R\phi_1 - (C_1 - R)\sin\phi_1 \\ R + (C_1 - R)\cos\phi_1 \\ \phi_1 \\ -R\phi_1 - (L - R)\sin\phi_1 + (L - C)\sin\phi_2 \\ R + (L - R)\cos\phi_1 - (L - C)\cos\phi_2 \\ \phi_2 \end{bmatrix} \tag{16}$$

where $C_1 = \frac{(Cm_\ell + Lm_t)}{m_\ell + m_t}$ is the location along the stance leg of the combined center rigid body. Dependence of angles on time is omitted for brevity. The origin, O, of the coordinate system is on the ground as shown in Fig. 2. The origin is positioned such that, when the stance leg is vertical, the bottom of the stance leg and the origin are coincident. Assuming infinite friction, the contact point between the rounded foot and the ground moves as the stance leg rotates.

The equations of motion are summarized as

$$\mathbf{T}^{\mathrm{T}}\mathbf{M}\mathbf{T}\ddot{\mathbf{q}} = \mathbf{f} + \mathbf{T}^{\mathrm{T}}\mathbf{M}(\mathbf{a} - \mathbf{g}) \tag{17}$$

where the matrix $\mathbf{T}$ is the 6×2 Jacobian of $\mathbf{k}$, i.e., $\mathbf{T} = \partial\mathbf{k}/\partial\mathbf{q}$. The reduced mass matrix is

$$\mathbf{M} = \mathrm{diag}(m_1, m_1, I_1, m_\ell, m_\ell, I_\ell) \tag{18}$$

where $m_1 = m_\ell + m_t$ is the combined mass of the stance leg. The combined moment of inertia of the stance leg is given by

$$I_1 = I_\ell + I_t + (C_1 - C)^2 m_\ell + (L - C_1)^2 m_t \tag{19}$$

The *convective acceleration* is

$$\mathbf{g} = \frac{\partial}{\partial\mathbf{q}}\left(\frac{\partial\mathbf{k}}{\partial\mathbf{q}}\dot{\mathbf{q}}\right)\dot{\mathbf{q}} \tag{20}$$

and $\mathbf{a} = g[0, -1, 0, 0, -1, 0]^{\mathrm{T}}$ is the generalized acceleration vector due to gravity ($g = 9.8m/s^2$). The generalized spring force is $\mathbf{f} = \kappa[\phi_2 - \phi_1, \phi_1 - \phi_2]^{\mathrm{T}}$. By substitution of variables, it can be seen that (17) is equivalent to (1), with $\mathcal{M}(\mathbf{q}) = \mathbf{T}^{\mathrm{T}}\mathbf{M}\mathbf{T}$ and $\mathcal{F}(\mathbf{q}, \dot{\mathbf{q}}, \kappa) = \mathbf{f} + \mathbf{T}^{\mathrm{T}}\mathbf{M}(\mathbf{a} - \mathbf{g})$.

Appendix B: Collision and Support Transfer

Since the end of the swing leg is even with the ground when $\phi_1 = -\phi_2$, collisions are found by detecting zero-crossings of $\mathcal{C}(\phi_1, \phi_2) = \phi_1 + \phi_2$. However, our model also allows the swing foot to move below the ground,[2] and thus a zero-crossing can occur when the foot passes above the ground.

[2]Because the Anthropomorphic Walker does not have knees, it can walk only by passing a foot through the ground.

Hence, we detect collisions by detecting zero-crossings of $\mathcal{C}$ when $\phi_1 < 0$ and $\dot{\mathcal{C}} < 0$.

The dynamical consequence of collision is determined by a system of equations relating the instantaneous velocities immediately before and after the collision. By assuming ground collisions to be impulsive and inelastic the result can be determined by solving a set of equations for the post-collision velocity. To model toe-off before such a collision, an impulse along the stance leg is added. In particular, the post-collision velocities $\dot{q}^+$ can be solved for using

$$\mathbf{T}^{+\mathrm{T}}\mathbf{M}\mathbf{T}^{+}\dot{\mathbf{q}}^{+} = \mathbf{T}^{+\mathrm{T}}(\mathbf{v} + \mathbf{M}\mathbf{T}\dot{\mathbf{q}}^{-}) \tag{21}$$

where $\dot{\mathbf{q}}^-$ are the pre-collision velocities, $\mathbf{T}$ is the pre-collision kinematic transfer matrix specified above,

$$\mathbf{k}^{+}(\mathbf{q}^{-}) = \begin{bmatrix} -R\phi_2 - (L-R)\sin\phi_2 + (L-C)\sin\phi_1 \\ R + (L-R)\cos\phi_2 - (L-C)\cos\phi_1 \\ \phi_1 \\ -R\phi_2 - (C_1-R)\sin\phi_2 \\ R + (C_1-R)\cos\phi_2 \\ \phi_2 \end{bmatrix} \tag{22}$$

is the post-collision kinematic transformation function, $\mathbf{T}^+ = \partial\mathbf{k}^+/\partial\mathbf{q}$, is the post-collision kinematic transfer matrix, $\mathbf{M}$ is the mass matrix as above and

$$\mathbf{v} = \iota[-\sin\phi_1, \cos\phi_1, 0, 0, 0, 0]^{\mathrm{T}} \tag{23}$$

is the impulse vector with magnitude ι. Defining

$$\mathcal{M}^{+}(\mathbf{q}) = \mathbf{T}^{+\mathrm{T}}\mathbf{M}\mathbf{T}^{+\mathrm{T}}, \tag{24}$$

$$\mathcal{M}^{-}(\mathbf{q}) = \mathbf{T}^{+\mathrm{T}}\mathbf{M}\mathbf{T}, \tag{25}$$

$$\mathcal{I}(\mathbf{q}, \iota) = \mathbf{T}^{+\mathrm{T}}\mathbf{v} \tag{26}$$

and substituting into (21) gives (2).

At collision, the origin of the coordinate system shifts forward by $2(R\phi_2 + (L-R)\sin\phi_2)$. The swing and stance leg switch roles; i.e., ϕ_1 and ϕ_2 and their velocities are swapped. Simulation then continues as before.

References

Agarwal, A., & Triggs, B. (2006). Recoving 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(1), 44–58.

Bissacco, A. (2005). Modeling and learning contact dynamics in human motion. In *Proceedings of IEEE conference on computer vision and pattern recognition* (Vol. 1, pp. 421–428).

Blickhan, R., & Full, R. J. (1993). Similarity in multilegged locomotion: bouncing like a monopode. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, *173*(5), 509–517.

Brubaker, M. A., & Fleet, D. J. (2008). The kneed walker for human pose tracking. In *Proceedings of IEEE conference on computer vision and pattern recognition*.

Brubaker, M. A., Fleet, D. J., & Hertzmann, A. (2007). Physics-based person tracking using simplified lower-body dynamics. In *Proceedings of IEEE conference on computer vision and pattern recognition*.

Chan, M., Metaxas, D., & Dickinson, S. (1994). Physics-based tracking of 3D objects in 2D image sequences. In *Proceedings of ICPR* (pp. 432–436).

Choo, K., & Fleet, D. J. (2001). People tracking using hybrid Monte Carlo filtering. In *Proceedings of IEEE international conference on computer vision* (Vol. II, pp. 321–328).

Collins, S. H., & Ruina, A. (2005). A bipedal walking robot with efficient and human-like gait. In *Proceedings of IEEE conference on robotics and automation*.

Collins, S. H., Wisse, M., & Ruina, A. (2001). A three-dimensional passive-dynamic walking robot with two legs and knees. *International Journal of Robotics Research*, *20*(7), 607–615.

Collins, S., Ruina, A., Tedrake, R., & Wisse, M. (2005). Efficient bipedal robots based on passive-dynamic walkers. *Science*, *307*(5712), 1082–1085.

Delamarre, Q., & Faugeras, O. (2001). 3D articulated models and multiview tracking with physical forces. *Computer Vision and Image Understanding*, *81*(3), 328–357.

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, *10*(3), 197–208.

Elgammal, A., & Lee, C.-S. (2004). Inferring 3D body pose from silhouettes using activity manifold learning. In *Proceedings of IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 681–688).

Fleet, D., & Weiss, Y. (2005). Optical flow estimation. In *The handbook of mathematical models of computer vision* (pp. 239–258). Berlin: Springer.

Full, R. J., & Koditschek, D. E. (1999). Templates and anchors: neuromechanical hypotheses of legged locomotion on land. *Journal of Experimental Biology*, *202*, 3325–3332.

Goldstein, H., Poole, C. P., & Safko, J. L. (2001). *Classical mechanics* (3rd ed.). Reading: Addison-Wesley.

Herda, L., Urtasun, R., & Fua, P. (2005). Hierarchical implicit surface joint limits for human body tracking. *Computer Vision and Image Understanding*, *99*(2), 189–209.

Hodgins, J. K., Wooten, W. L., Brogan, D. C., & O'Brien, J. F. (1995). Animating human athletics. In *Proceedings of SIGGRAPH* (pp. 71–78).

Kakadiaris, L., & Metaxas, D. (2000). Model-based estimation of 3D human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1453–1459.

Kass, M., Witkin, A., & Terzopoulos, D. (1987). Snakes: active contour models. *International Journal of Computer Vision*, *1*(4), 321–331.

Khan, Z., Balch, T., & Dellaert, F. (2004). A Rao-blackwellized particle filter for eigentracking. In *Proceedings of IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 980–986).

Kong, A., Liu, J. S., & Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, *89*(425), 278–288.

Kovar, L., Schreiner, J., & Gleicher, M. (2002). Footskate cleanup for motion capture editing. In *Proceedings of symposium on computer animation*.

Kuo, A. D. (2001). A simple model of bipedal walking predicts the preferred speed-step length relationship. *Journal of Biomechanical Engineering*, *123*(3), 264–269.

Kuo, A. D. (2002). Energetics of actively powered locomotion using the simplest walking model. *Journal of Biomechanical Engineering*, *124*(1), 113–120.

Liu, C. K., Hertzmann, A., & Popović, Z. (2005). Learning physics-based motion style with nonlinear inverse optimization. *ACM Transactions on Graphics*, *24*(3), 1071–1081.

McGeer, T. (1990a). Passive dynamic walking. *International Journal of Robotics Research*, *9*(2), 62–82.

McGeer, T. (1990b). Passive walking with knees. In *Proceedings of international conference on robotics and automation* (Vol. 3, pp. 1640–1645).

McGeer, T. (1992). Principles of walking and running. In *Advances in comparative and environmental physiology* (Vol. 11, Chap. 4, pp. 113–139). Berlin: Springer.

Metaxas, D., & Terzopoulos, D. (1993). Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(6), 580–591.

Pavlović, V., Rehg, J., Cham, T.-J., & Murphy, K. (1999). A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Proceedings of IEEE international conference on computer vision* (pp. 94–101).

Pentland, A., & Horowitz, B. (1991). Recovery of nonrigid motion and structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(7), 730–742.

Pitt, M. K., & Shepard, N. (1999). Filtering via simulation: auxiliary particle filters. *Journal of the American Statistical Association*, *94*, 590–599.

Popović, Z., & Witkin, A. (1999). Physically based motion transformation. In *Proceedings of SIGGRAPH* (pp. 11–20).

Pratt, G. A. (2000). Legged robots at MIT: what's new since Raibert? *IEEE Transactions on Robotics and Automation*, *7*(3), 15–19.

Rahimi, A., Recht, B., & Darrell, T. (2005). Learning appearance manifolds from video. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 868–875).

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, *5*(2), 121–125.

Rosales, R., Athitsos, V., Sigal, L., & Sclaroff, S. (2001). 3D hand pose reconstruction using specialized mappings. In *Proceedings of IEEE international conference on computer vision* (Vol. 1, pp. 378–385).

Safonova, A., Hodgins, J. K., & Pollard, N. S. (2004). Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Transactions on Graphics*, *23*(3), 514–521.

Shakhnarovich, G., Viola, P., & Darrell, T. (2003). Fast pose estimation with parameter-sensitive hashing. In *Proceedings of IEEE international conference on computer vision* (pp. 750–757).

Shin, H. J., Kovar, L., & Gleicher, M. (2003). Physical touchup of human motions. In *Proceedings of Pacific graphics* (pp. 194–203).

Sidenbladh, H., Black, M. J., & Fleet, D. J. (2000). Stochastic tracking of 3D human figures using 2D image motion. In *Proceedings of IEEE European conference on computer vision* (Vol. 2, pp. 702–718).

Sigal, L., & Black, M. (2006). *HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion*. Technical report CS-06-08, Computer science, Brown University.

Sminchisescu, C., & Jepson, A. (2004). Generative modeling for continuous non-linearly embedded visual inference. In *Proceedings of international conference on machine learning* (pp. 96–103).

Sminchisescu, C., Kanaujia, A., & Metaxas, D. (2007). *BME*^{3}: Discriminative density propagation for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(11), 2030–2044.

Terzopoulos, D., & Metaxas, D. (1990). Dynamic 3D models with local and global deformations: deformable superquadrics. In *Proceedings of IEEE international conference on computer vision* (pp. 606–615).

Urtasun, R., Fleet, D. J., Hertzmann, A., & Fua, P. (2005). Priors for people tracking from small training sets. In *Proceedings of IEEE international conference on computer vision* (Vol. 1, pp. 403–410).

Urtasun, R., Fleet, D. J., & Fua, P. (2006). 3D people tracking with Gaussian process dynamical models. In *Proceedings of IEEE conference on computer vision and pattern recognition* (Vol. 1, pp. 238–245).

Van der Linde, R. Q., & Schwab, A. L. (2002). *Lecture notes multibody dynamics B*, wb1413, course 1997/1998. Lab. for Engineering Mechanics, Delft University of Technology.

Vondrak, M., Sigal, L., & Jenkins, O. C. (2008). Physical simulation for probabilistic motion tracking. In *Proceedings of IEEE conference on computer vision and pattern recognition*.

Wachter, S., & Nagel, H. H. (1999). Tracking persons in monocular image sequences. *Computer Vision and Image Understanding*, *74*(3), 174–192.

Wisse, M., Hobbelen, D. G. E., & Schwab, A. L. (2007). Adding an upper body to passive dynamic walking robots by means of a bisecting hip mechanism. *IEEE Transactions on Robotics*, *23*(1), 112–123.

Witkin, A., & Kass, M. (1988). Spacetime constraints. In *Proceedings of SIGGRAPH* (Vol. 22, pp. 159–168).

Wren, C. R., & Pentland, A. (1998). Dynamics models of human motion. In *Proceedings of automatic face and gesture recognition*.

Yin, K., Loken, K., & van de Panne, M. (2007). SIMBICON: simple biped locomotion control. *ACM Transactions on Graphics*, *26*(3).