Natural Metrics and Least-Committed Priors for Articulated Tracking

Søren Hauberg[a], Stefan Sommer[a], Kim Steenstrup Pedersen[a]

[a] *eScience Centre, Dept. of Computer Science, University of Copenhagen, Universitetsparken 5, Copenhagen, Denmark*

Abstract

In articulated tracking, one is concerned with estimating the pose of a person in every frame of a film. This pose is most often represented as a kinematic skeleton where the joint angles are the degrees of freedom. Least-committed predictive models are then phrased as a Brownian motion in joint angle space. However, the metric of the joint angle space is rather unintuitive as it ignores both bone lengths and how bones are connected. As Brownian motion is strongly linked with the underlying metric, this has severe impact on the predictive models. We introduce the spatial kinematic manifold of joint positions, which is embedded in a high dimensional Euclidean space. This Riemannian manifold inherits the metric from the embedding space, such that distances are measured as the combined physical length that joints travel during movements. We then develop a least-committed Brownian motion model on the manifold that respects the natural metric. This model is expressed in terms of a stochastic differential equation, which we solve using a novel numerical scheme. Empirically, we validate the new model in a particle filter based articulated tracking system. Here, we not only outperform the standard Brownian motion in joint angle space, we are also able to specialise the model in ways that otherwise are both difficult and expensive in joint angle space.

Keywords: Articulated Tracking, Brownian Motion on Riemannian Manifolds, Manifold-valued Stochastic Differential Equations, Numerical Solutions to SDEs

1. Introduction

This paper is concerned with least-committed priors for probabilistic articulated tracking, i.e. estimation of human poses in sequences of images (Poppe, 2007). When treating such problems, a maximum *a posteriori* estimate is typically found by solving an optimisation problem, and the optimisation is then guided by a prior model for predicting future motion. For such statistical models of human motion, it is common to express the model as a kinematic skeleton (see fig. 1). This "stick figure" model is complex enough to be descriptive and simple enough to give tractable algorithms. Most of the resulting models are, however, expressed in a space with rather unnatural metric properties, which is also apparent in the models. Specifically, the applied metrics most often only study changes in joint angles; the "size" of a movement is simply measured by summing how much each joint was bent. This ends up with *the flick of a finger* being just as large a motion as *waving an arm*, even though one would expect the latter to be much larger (see fig. 2). This rather unintuitive behaviour occurs as the metric ignores both the length of the individual bones and the hierarchical nature of the human body (the arm bone is connected to the shoulder bone, the shoulder bone is connected to the back bone, etc.). Often this problem is mitigated by weighting the joints, but, as we will show, this cannot lead to a spatially consistent metric.

In this paper, we define a representation of the kinematic skeleton with natural metric properties. Instead of studying joint angles, we explicitly model *joint positions*, such that our representation consists of the three dimensional spatial coordinates of all joints. As bone lengths are constant, the distance between connected joints is also constant. This constraint confines our representation to a manifold embedded in the Euclidean space consisting of all joint positions. By inheriting the metric from the embedding space, we get a metric corresponding to the length of the spatial curves that joint positions follow during the movement. Interestingly, this natural metric is well in tune with how humans plan, think about and discuss motion (Morasso, 1981; Abend et al., 1982).

Using our spatial representation, we define a Brownian motion model on the Riemannian representation manifold that reflects the metric. The Brownian motion model is expressed as a manifold-valued stochastic differential equation (SDE), for which we need numerical solvers. We present a novel scheme for solving the SDE, which we apply as a least-committed prior in a particle filter based articulated tracking system. Furthermore, we show how the spatial nature of the model allows us to model interactions with the environment; something that is often ignored when the model is expressed with joint angles.

Email addresses: hauberg@diku.dk (Søren Hauberg), sommer@diku.dk (Stefan Sommer), kimstp@diku.dk (Kim Steenstrup Pedersen)

Accepted for publication in *Image and Vision Computing* (http://dx.doi.org/10.1016/j.imavis.2011.11.009).

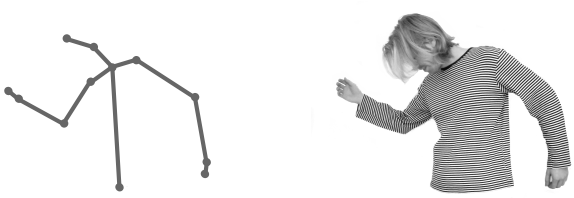


Figure 1: Left: a rendering of the kinematic skeleton. Each bone position is computed by a rotation and a translation relative to its parent. Right: an image showing a human in the pose represented by the skeleton to the left.

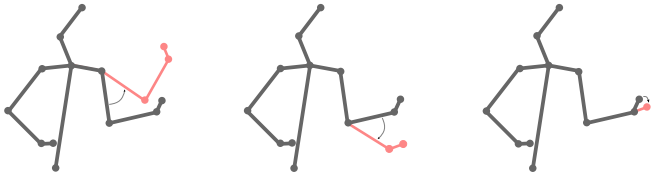


Figure 2: Examples of three motions that are equally large in the commonly used angular metric. All examples have a 45 degree angular distance to the initial pose.

1.1. Organisation of the Paper

We start the paper by discussing relevant background material and related work with emphasis on the non-spatial joint angle metric most often used when modelling human motion. We continue by defining a spatial manifold-valued pose representation with a natural and intuitive metric. The next step is to define a least-committed stochastic process on this manifold for predicting human motion; in sec. 3.2 we define a Brownian motion model that serves this purpose. In order to apply the model in real-world scenarios we need a suitable numerical scheme for working with this stochastic process; in sec. 4 we show how the underlying manifold-valued stochastic differential equation can be simulated. We then incorporate the predictive model in an articulated tracking system and compare with the standard Brownian motion in joint angle space. Furthermore, we show how interaction with the environment can trivially be included in the motion model due to the spatial nature of our framework. Finally, the paper is concluded with a discussion in sec. 6.

2. Background and Related Work

Probabilistic articulated tracking concerns the maximum *a posteriori* estimate of the pose of a person in every frame of a film. This requires a representation of human poses and a framework for computing the statistics of the observed poses. As we are seeking a posterior estimate, we need a prior model of the motion. This prior is the focus of this paper. In this section, we describe the pose representation, the probabilistic framework, the standard priors and other related work.

2.1. The Kinematic Skeleton

To represent the human body, we use the *kinematic skeleton* (see fig. 1), which is by far the most common choice (Poppe, 2007). This representation is a collection of connected rigid bones organised in a tree structure. Each bone can be rotated at the point of connection between the bone and its parent. We will refer to such a connection point as a *joint*. Elbow joints will be represented using one parameter while all other joints will be represented using three parameters.

We model the bones as having known constant length and, therefore, the angles between joints constitute the only degrees of freedom in the kinematic skeleton. We may collect all these joint angle vectors into one large vector θ, which will be confined to the N dimensional torus $\mathbb{T}^N$.

From joint angles it is straightforward to compute joint positions using *Forward Kinematics* (Erleben et al., 2005). This process starts at the skeleton root and recursively computes a joint position by translating its parent in the direction encoded by the joint angles, i.e.

$$\mathbf{a}_l = \mathbf{R}_l \left(\mathbf{a}_{l-1} + \mathbf{t}_l \right) \quad , \tag{1}$$

where $\mathbf{a}_l$ is the end-point of the l^{th} bone, and $\mathbf{R}_l$ and $\mathbf{t}_l$ denotes a rotation and a translation respectively. The rotation is parametrised by the relevant components of the pose vector θ and the length of the translation corresponds to the known length of the bone. We shall denote the vector containing all spatial joint coordinates as $F(\theta)$. The forward kinematics function F, thus, encodes bone lengths, bone connectivity as well as joint types.

In the human body, bones cannot move freely. A simple example is the elbow joint, which can approximately only bend between 0 and 160 degrees. To represent this, θ is confined to a subset $\mathbf{\Theta}$ of $\mathbb{T}^N$. For simplicity, $\mathbf{\Theta}$ is often defined by confining each component of θ to an interval, i.e. $\mathbf{\Theta} = \prod_{n=1}^{N} [l_n, u_n]$, where l_n and u_n denote the lower and upper bounds of the n^{th} component. More realistic joint constraints are also possible, e.g. the implicit surface models of Herda et al. (2004). For our purposes any hard constraint model is applicable, though the choice will have an impact on the computational requirements.

2.2. Probabilistic Motion Inference

The objective in *articulated human motion estimation* is to infer θ in each observation in a sequence (Poppe, 2007). To make things practical, it is common to assume that the joint angles follow a first order Markov chain and that observations are conditionally independent given the true joint configuration. From all observations seen so far, the current joint angles can then be estimated from (Cappé et al., 2007)

$$p(\theta_t|\mathbf{Z}_{1:t}) \propto p(\mathbf{Z}_t|\theta_t) \int p(\theta_t|\theta_{t-1}) p(\theta_{t-1}|\mathbf{Z}_{1:t-1}) \mathrm{d}\theta_{t-1} \quad , \tag{2}$$

where $\theta_{1:T} = \{\theta_1, \ldots, \theta_T\}$ and $\mathbf{Z}_t$ denotes the observation at time t.

When only using a single camera or a narrow baseline stereo camera, $p(\mathbf{Z}_t|\theta_t)$ becomes multi-modal due to self-occlusions and visual ambiguities. For this reason, we apply the particle filter (Cappé et al., 2007) for inferring pose parameters. Briefly put, this algorithm recursively draws samples $\theta_{t+1}^{(j)}$ from the motion prior $p(\theta_{t+1}|\theta_t)$ and assigns weights to these according to the likelihood $p(\mathbf{Z}_{t+1}|\theta_{t+1})$. These weighted samples form an approximation of $p(\theta_{t+1}|\mathbf{Z}_{1:t+1})$; the mean of which can be estimated from

$$\mathrm{E}[\theta_{t+1}|\mathbf{Z}_{1:t+1}] \approx \sum_{j=1}^{J} w_j \theta_{t+1}^{(j)} \quad , \tag{3}$$

where $w_j \propto p(\mathbf{Z}_{t+1}|\theta_{t+1})$ are normalised likelihoods that sum to one.

2.3. Brownian Motion of Joint Angles

The focus of this paper is motion priors, i.e. $p(\theta_{t+1}|\theta_t)$. When no specific motion is being modelled, it is common to assume that θ_t follows an Euclidean Brownian motion, i.e.

$$p(\theta_{t+1}|\theta_t) \propto \exp\left(-\frac{1}{2} d_\theta^2(\theta_{t+1}, \theta_t)\right) \quad , \tag{4}$$

where $d_\theta(\theta_{t+1}, \theta_t) = \|\theta_{t+1} - \theta_t\|$ is the Euclidean distance in joint angle space. In practice it is common to scale the individual joint angles to encode that some joints move more than others. This corresponds to introducing a covariance matrix in eq. 4. Formally, this makes the model an *Itô diffusion* (Øksendal, 2000), but we will simply treat it as a Brownian motion in the scaled coordinate system. However, as we shall see, the Brownian motion model in angle space has some rather unintuitive properties, which cannot be avoided by scaling the coordinates.

Formally, Euclidean Brownian motion, also known as the Wiener process, is defined (Sato, 1999) as a stochastic process W_t on $\mathbb{R}^d$ having independent increments, such that for any partitioning, $n \geq 1$ and $0 \leq t_0 < t_1 < \ldots < t_n$, $W_{t_0}, W_{t_1} - W_{t_0}, \ldots, W_{t_n} - W_{t_{n-1}}$ are independent random variables. Furthermore, the increments are zero mean Gaussian distributed $W_{s+t} - W_s \sim \mathcal{N}(0, t\mathbf{I})$ for all $s, t > 0$. Hence, we may intuitively think of Euclidean Brownian motion as the result of time integration of zero mean Gaussian white noise, that is an infinite sum of i.i.d. infinitesimal Gaussian steps. As such the Euclidean Brownian motion is both a d-dimensional Gaussian and a Levy process (Sato, 1999).

Brownian motion is generally considered the *least-committed* motion model as it 1) assumes no knowledge of the past motion given our current position and 2) takes steps with maximum entropy under the constraint of a fixed finite variance. The last point arises from the fact that the steps are Gaussian distributed, which is the maximum entropy distribution constrained by a finite variance and known mean value.

Furthermore, Brownian motion lies at the heart of stochastic calculus and the theory of stochastic differential equations (Øksendal, 2000). It allows for the formulation of general stochastic process models, including the Kalman-Bucy filter, the continuous time formulation of the Kalman filter. Brownian motion also forms the basis of most other models of interest for articulated tracking.

2.4. The Joint Angle Metric

The Euclidean Brownian motion model in eq. 4 is strongly linked to the metric. Specifically, eq. 4 assumes that $d_\theta(\theta_t, \theta_{t-1}) = \|\theta_t - \theta_{t-1}\|$ is a suitable metric for comparing poses. While this model might seem reasonable at first glance, we shall soon see that it exhibits several unnatural properties.

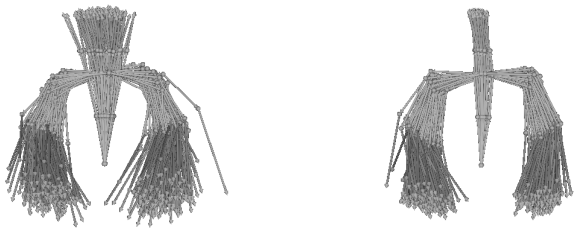
As a motivating example of the behaviour of d_θ, we show three movements of "equal size" in fig. 2. In all movements one joint has been moved 45 degrees, while the remaining have been kept constant. While the actual numerical changes from the initial positions are the same, the movements appear to be substantially different, with the movement on the left of the figure appearing to be much larger than the one on the right. The example in fig. 2 just scratches the surface of the unnatural behaviour of d_θ. The main causes of difficulty with d_θ are due to two phenomena.

First, the metric ignores the length of the bones in the body. As such, even a small change in the angle of a joint connected to a long bone can lead to large spatial changes. This problem can be avoided by assigning a weight to each joint angle according to the length of the bone controlled by the joint.

The second phenomena, is that the metric ignores the order of the joint in the kinematic chain. By bending one joint, the position of all joints further down the kinematic chain is altered, while the position of joints closer to the root of the kinematic tree remain unaltered. From a probabilistic point of view, this means that the variance of joint positions increases as the kinematic chains are traversed. Hence the joint angle model artificially increases the spatial variance, which means that the model is bound to perform poorly as a temporal low-pass filter.

These phenomena effectively means that some joint angles have much more influence than others. In practice this often leads to unstable predictive models. To mitigate this instability, it is common to introduce a covariance Σ_θ in joint angle space that influences the relative importance of each joint. To illustrate this, we learn the covariance of a Brownian motion in joint angle space corresponding to a person waving his arms. In fig. 3a, we then show samples from this distribution. As can be seen, the variance of each joint position increases with the distance to the skeleton root. This increase in variance is an inherent part of the model and does not come from the motion data.

To gain further insight into the spatial behaviour of the joint angle model, we approximate the covariance of joint positions defined by the forward kinematics function $F(\theta)$



(a) **(b)**

Figure 3: Samples from two Brownian motion models, where the covariance is learned from the same motion data. (a) Samples from a model in joint angle space. (b) Samples from a spatial manifold-valued Brownian motion. Note that the joint angle model is inherently more variant.

(Hauberg et al., 2010). Linearising $F(\cdot)$ around θ gives an approximation of this covariance,

$$\text{cov}[F(\theta)] \approx \mathbf{J}_\theta \Sigma_\theta \mathbf{J}_\theta^T \ , \tag{5}$$

where $\mathbf{J}_\theta$ denotes the Jacobian of $F(\cdot)$ in θ. From this expression, we see that if we want a consistent model of joint positions, expressed in terms of joint angles, we need different covariance matrices for every value of θ because $\mathbf{J}_\theta$ varies. A model where the covariance smoothly varies is essentially a model on a Riemannian manifold. This is the approach we will study in this paper.

To conclude, using d_θ as the underlying metric when defining Brownian motion priors leads to the rather unnatural above-mentioned phenomena. These cannot be avoided by introducing a single covariance matrix in joint angle space, instead a Riemannian approach is needed.

2.5. Modelling Interaction with the Environment

In practice, articulated tracking systems are often based on particle filters due to the multi-modality of the likelihood. Unfortunately, the particle filter scales exponentially with the dimensionality of the state space. One solution is to specialise the motion prior $p(\theta_{t+1}|\theta_t)$ to the studied motion. This can "guide" the filter through the multiple modes of the likelihood.

Humans are constantly interacting with the environment: picking up objects, leaning against walls, touching the ground plane, and so on. Hence, an immediate way to improve motion models is to include this knowledge. When motion models are expressed in terms of joint angles, it is, however, difficult to incorporate knowledge of the environment into the models. As the environment is inherently spatial, the relationship between joint angles and the environment is given by the non-linear forward kinematics function F. Due to this non-linearity, only limited work has been done to build models that incorporate environmental knowledge. One notable exception include the work of Yamamoto and Yagishita (2000), where the forward kinematics function is linearised. This approach shows promise in constrained situations, even though the linearised function is highly non-linear. Brubaker et al. (2010) also model interaction with the ground plane as part of a biomechanical model of walking; their model is, however, only capable of describing walking.

Kjellström et al. (2010) has suggested a more general object interaction model. They model a person interaction with a stick with known position, which gives them information about the position of the hands. They then suggest a motion model consisting of angular Brownian motion subject to the constraint that the hands attain the known positions. Kjellström et al. samples approximately from this model using rejection sampling. While this approach works, the rejection sampling is, computationally very demanding due to the high dimensionality of the angle space. We will consider this model further in the experimental section of the paper.

2.6. Manifold Learning in Motion Analysis

Another way to craft motion models is to learn a manifold in angle space and confine the motion to this manifold. A predictive motion model can then be learned on this manifold. Sidenbladh et al. (2000) learned a low-dimensional linear subspace using *Principal Component Analysis* and used a linear motion model in this subspace. Sminchisescu and Jepson (2004) use *Laplacian Eigenmaps* (Belkin and Niyogi, 2003) to learn a nonlinear motion manifold. Similarly, Lu et al. (2008) use a *Laplacian Eigenmaps Latent Variable Model* (Carreira-Perpinan and Lu, 2007) to learn a manifold. All three learning schemes can be phrased in terms of pair-wise distances between training data, where the metric is the joint angle distance discussed in sec. 2.4.

The above approaches learn a manifold and then ignore the training data. A reasonable alternative is to also use the data for learning a predictive model on the manifold. Urtasun et al. (2005) suggested to learn a prior distribution in a low dimensional latent space using a *Scaled Gaussian Process Latent Variable Model* (Grochow et al., 2004). This not only restricts the tracking to a low dimensional latent space, but also makes parts of this space more likely than others. The approach, however, ignores all temporal aspects of the training data. To remedy this, both Urtasun et al. (2006) and Wang et al. (2008) suggested learning a low dimensional latent space *and* a temporal model at once using a *Gaussian Process Dynamical Model*. The learning algorithms in both approaches, however, require regularisation to give stable results. This regularisation is in practice based on the joint angle metric.

All manifold learning approaches discussed in this section rely on the joint angle metric. As we have discussed in sec. 2.4, this metric has several undesirable properties, which will influence the learning. In this paper, we take a step back and design a sensible metric along with a compatible least-committed motion model. This will allow us to fix the problems with the joint angle metric and the related angular Brownian motion. It should be stressed that

we will not be *learning* any manifolds; we will analytically be designing one.

3. A Spatial Metric

For years, experimental neurologists have studied how people move (Morasso, 1981; Abend et al., 1982) and have found strong evidence that humans plan motion in terms of the spatial location of limbs. This unsurprising conclusion complements the fact that both the surrounding environment and images thereof are inherently spatial as well. We, thus, set out to model how joint *positions* change over time. This will allow us to improve upon the joint angle metric and will also ease modelling that includes knowledge of the environment. As we will see, the constraints imposed by constant bone lengths confines the collection of all joint positions to a smooth manifold. As most statistical tools have been developed for Euclidean spaces, defining a probabilistic model on the manifold is not straightforward. There is, e.g., no direct generalisation of the normal distribution to the Riemannian domain. For this reason, we turn to the underlying stochastic differential equation (SDE) of Brownian motion. This SDE has the nice property that it can be generalised to the Riemannian domain (see e.g. (Hsu, 2002)). One problem with SDE's on manifolds, is that, to the best of our knowledge, no general literature exists on their numerical treatment. Later in the paper, we will introduce a novel method for simulating the manifold valued SDE's numerically and use this for predicting human motion in an articulated tracking system.

In (Hauberg et al., 2010; Hauberg and Pedersen, 2011b), we introduced the kinematic manifold and showed that it is suitable for modelling interactions with the environment. In these papers, a somewhat *ad hoc* predictive model was defined where motion was modelled in the embedding space followed by a projection onto the manifold. In contrast to this, the model developed here has a solid foundation in the well-known Brownian motion model.

3.1. The Metric and the Kinematic Manifold

The joint angle representation has at least two good properties. First, it is fairly simple to create statistical models in joint angle space. Secondly, as long as the joint limits are respected, the resulting pose is valid. As previously mentioned, the metric in angle space is, however, not as well-behaved as one would like, which gives rise to unstable statistical models.

As we are studying images of motion, we want a metric where the size of a movement is determined by "how large" it appears. To achieve this, we consider the physical length of the spatial curves that joint positions follow when going from one pose to another. To properly define these curves, we first consider the set of spatial joint coordinates of all possible poses as the image of the forward kinematics function F. The resulting set

$$\mathcal{M} \equiv \{F(\theta) \mid \theta \in \Theta\} \ . \tag{6}$$

is a subset of the space $\mathbb{R}^{3L}$ with L denoting the number of bone end-points counting only one for each joint. Hence, a point in $\mathcal{M}$ is a vector of spatial joint positions. Since the angle space is compact and F is an injective function with a full-rank Jacobian, $\mathcal{M}$ is a compact differentiable manifold with boundary embedded in $\mathbb{R}^{3L}$. We denote $\mathcal{M}$ the *kinematic manifold*. It should be stressed that $\mathcal{M}$ is topologically equivalent to the angle space Θ, but has a different geometry. In other words, the two representations capture the same set of poses, but have different metrics.

The distance between two poses on the kinematic manifold is given by the manifold metric and is therefore defined as the length of the shortest curve on $\mathcal{M}$ connecting the poses. Formally, for poses $x, x' \in \mathcal{M}$, we have

$$\text{dist}_{\mathcal{M}}(x, x') = \min_{\substack{c(\tau)\in\mathcal{M}, \\ c(0)=x, c(1)=x'}} \int_0^1 \|\dot{c}(\tau)\| \mathrm{d}\tau \ , \tag{7}$$

with $\|\dot{c}(\tau)\|$ denoting the size in $\mathbb{R}^{3L}$ of the curve derivative $\dot{c}(\tau)$. Hence, the integral corresponds to the ordinary curve length. The distance between two poses, thus, is the shortest of all curves on $\mathcal{M}$ that connect the poses. As a curve on $\mathcal{M}$ is a sequence of poses, this metric corresponds to the minimal combined physical distance that the joints need to move. This gives the metric a strong physical interpretation as it measures distances directly in the world coordinate system. This is in stark contrast to the joint angle metric, which measures distances in terms of an intrinsic set of parameters.

From the definition of $\mathcal{M}$ (eq. 6) it is clear that poses on $\mathcal{M}$ encodes all knowledge of the forward kinematics function F. This includes both bone lengths and connectivity. The manifold metric, thus, incorporates knowledge of the skeleton layout when measuring the size of a movement. This is a quite natural requirement for a "movement metric", yet the joint angle metric is inherently unable to include such knowledge.

3.2. Manifold-Valued Brownian Motion

Having a natural metric for measuring movements, the next step is to define a least-committed temporal model that respects this metric. We will define a manifold-valued Brownian motion model for this. While the normal distribution provides a Brownian motion model in the Euclidean case, no such simple model is available in the general Riemannian domain. We, thus, turn to stochastic differential equations for such models.

The Brownian motion model is completely characterised by its mean and covariance function. The temporal evolution of these moments are given by the Kolmogorov backward equation (Øksendal, 2000), i.e. by a diffusion governed by the infinitesimal generator of the process. For the Euclidean Brownian motion process, this generator is half

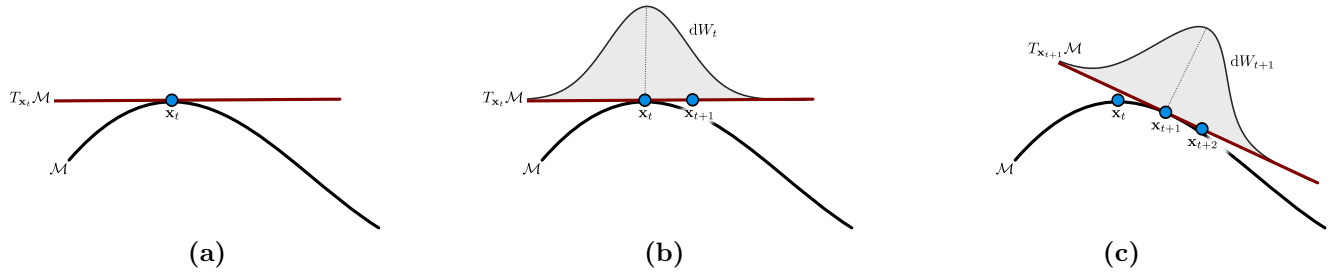


Figure 4: Steps in the Brownian motion model. (a) The manifold $\mathcal{M}$ along with the tangent space $T_{\mathbf{x}_t}\mathcal{M}$ at $\mathbf{x}_t$. (b) A normal distribution $\mathrm{d}W_t$ in the embedding space with mean value $\mathbf{x}_t$ is projected to the tangent space. The value $\mathbf{x}_{t+1}$ is sampled from the projection of $\mathrm{d}W_t$. Note that for infinitesimal variances, $\mathbf{x}_{t+1}$ stays on the manifold. (c) A normal distribution $\mathrm{d}W_{t+1}$ with mean value $\mathbf{x}_{t+1}$ is again projected to the tangent space at $\mathbf{x}_{t+1}$. A new position $\mathbf{x}_{t+2}$ is sampled and the procedure is repeated.

the Laplace operator $1/2\Delta$. Similarly, Brownian motion on a manifold is generated by half the Laplace-Beltrami operator $1/2\Delta_{\mathcal{M}}$ (Hsu, 2002), which, in coordinates, is defined by

$$\Delta_{\mathcal{M}} f = \sum_{i,j=1}^{\dim \mathcal{M}} \frac{1}{\sqrt{\det g}} \partial_i \left(\sqrt{\det g}\, g^{ij} \partial_j f \right) \tag{8}$$

for smooth scalar valued functions $f : \mathcal{M} \to \mathbb{R}$. Here g^{ij} denotes the components of the metric tensor g which determines the geometry of $\mathcal{M}$. For embedded manifolds such as the kinematic manifold, the Laplace-Beltrami operator has a particularly simple form. Let $P^\alpha(\mathbf{x}_t)$ denote the projection of the αth coordinate unit vector in the embedding space $\mathbb{R}^{3L}$ to the tangent space of $\mathcal{M}$ at $\mathbf{x}_t$. Then

$$\Delta_{\mathcal{M}} f = \sum_{\alpha=1}^{3L} \partial^2_{P^\alpha} f \ , \tag{9}$$

i.e., the operator differentiates twice in each direction P^α before summing the results. Using this form, Brownian motion is a solution to the stochastic differential equation

$$\mathrm{d}\mathbf{x}_t = \sum_{\alpha=1}^{3L} P_\alpha(\mathbf{x}_t) \circ \mathrm{d}W_t^\alpha \ , \tag{10}$$

in the embedding space $\mathbb{R}^{3L}$. Here W_t is a Euclidean Brownian motion in the embedding space with W_t^α denoting the αth coordinate, and the equation is written using the Stratonovich integral (Hsu, 2002; Øksendal, 2000) as indicated by the notation $\circ\mathrm{d}$. It is interesting to note that while the geodesic distance played an important part when the model was defined it does not appear in eq. 10; for this reason it need not be computed in the numerical implementation.

Because the projection of a Gaussian distribution into a linear subspace is still a Gaussian, the above SDE can be interpreted as taking infinitesimal Gaussian steps in the tangent space. It is important to note that solutions to eq. 10 will stay on the manifold even though the infinitesimal steps are taken in the tangent space, i.e.

$$P\left(x_t \in \mathcal{M} \;\middle|\; x_0 \in \mathcal{M}\right) = 1 \ . \tag{11}$$

An illustration of this model can be seen in fig. 4. New steps along the Brownian path are generated by following an infinitesimal Euclidean Brownian motion in the tangent space at the current position of the path. These steps are then integrated over time to generate the final path.

As with the joint angle model, it is often convenient to be able to express that some bones move more than others. This can be achieved by scaling the coordinates in the embedding space resulting in a model which, technically, is not a Brownian motion on the manifold, but instead an instance of Itô diffusion.

3.3. Spatially Constrained Brownian Motion

When building motion models, it can be practical to constraint certain bone positions. This can be used to ensure that the feet are touching the ground plane, that the hands are holding on to an object of known position and so forth. As a point on the kinematic manifold consists of the spatial position of individual bone end points, it is trivial to incorporate such knowledge into the model. If, for instance, we wish to keep the hand positions fixed, we can force the relevant entries of $\mathrm{d}W_t$ to zero. More complicated constraints can be encoded in the same way as long as they are physically possible.

3.4. Relations to Directional Statistics

A large part of the work on manifold-valued statistics has been done on spheres; this is known as *directional statistics* (Mardia and Jupp, 1999). Here easy-to-use Brownian motion models are available in the Von Mises distribution. In sequential analysis, this has found uses in such different areas as multi-target air plane tracking (Miller et al., 1995) and white matter tracking in Diffusion Tensor MRI (Zhang et al., 2007). Except for the special case of the kinematic skeleton consisting of only one bone, the kinematic manifold is not spherical and hence the Von Mises distribution is not applicable. The more general Brownian motion model defined using the Laplace-Beltrami operator is nevertheless compatible with directional statistics in the sense that the definition coincides with the Von Mises model for spherical manifolds.

4. Numerical Scheme

So far we have defined a Brownian motion model that respects the manifold metric. We now set out to simulate this model using the SDE in eq. 10. While there exists literature on both simulating SDE's in Euclidean spaces (Kloeden and Platen, 1992) and solving ODE's on manifolds (Hairer et al., 2004), to the best of our knowledge, no general solvers for manifold-valued SDE's have been described in the literature.

The most basic scheme for simulating Stratonovich SDE's in Euclidean domains is the Euler-Heun scheme, which is an ordinary first-order scheme for the Stratonovich integral (Kloeden and Platen, 1992). Given the current end-position x_t of the Brownian path, the next position x_{t+1} can be simulated in N steps with N controlling the precision of the scheme. For the SDE in eq. 10, a step in the Euler-Heun scheme takes the form

$$\begin{aligned} x_{t+1/N} &= x_t + \frac{1}{2}\left[P_{x_t} + P_{\tilde{x}_t}\right]\frac{\Delta W_t}{\sqrt{N}} \\ \tilde{x}_t &= x_t + P_{x_t}\frac{\Delta W_t}{\sqrt{N}} \ , \end{aligned} \tag{12}$$

where ΔW_t is normally distributed in $\mathbb{R}^{3L}$ and P_x is the orthogonal projection operator to the tangent space $T_x\mathcal{M}$. Letting U_x be a matrix with columns constituting an orthonormal basis of $T_x\mathcal{M}$, we can get the projection as

$$P_x = U_x U_x^T \ . \tag{13}$$

Unfortunately, the scheme in eq. 12 fails to ensure that the Brownian path stays on the manifold. We handle this issue by projecting each step to the manifold, resulting in the scheme

$$\begin{aligned} x_{t+1/N} &= \mathrm{proj}_{\mathcal{M}}\left(x_t + \frac{1}{2}\left[P_{x_t} + P_{\tilde{x}_t}\right]\frac{\Delta W_t}{\sqrt{N}}\right) \\ \tilde{x}_t &= \mathrm{proj}_{\mathcal{M}}\left(x_t + P_{x_t}\frac{\Delta W_t}{\sqrt{N}}\right) \ . \end{aligned} \tag{14}$$

Similar methods are used for ODE's on manifolds where a simple argument shows that the solution to the modified equation converges to the solution of the original ODE (Hairer et al., 2004, Chap. IV). The situation is more complex for the less well-behaved SDE's. Though the Euler-Heun scheme without the projection converges to a solution to the SDE (Kloeden and Platen, 1992), we have at this point no theoretical proof of convergence of the scheme in eq. 14.

In fig. 3b we show samples generated using this numerical scheme. The spatial covariance has been learned from the same data as the angular Brownian motion shown in fig. 3a. Comparing the two set of samples shows that the spatial Brownian motion model has smaller variance than the angular model. As the two models are learned from the same data, this clearly shows that the angular model artificially increases the variance. This makes the manifold-valued Brownian motion model a superior temporal low-pass filter.

4.1. Simulating Spatially Constrained Brownian Motion

As discussed in sec. 3.3 it can be practical to spatially constrain the Brownian motion, such that e.g. the hands attain known positions. The numerical scheme in eq. 14 easily allows for such extensions. Before projecting back to the manifold, the relevant entries of the joint position vector can be fixed to attain the desired positions. This will result in a simulated human pose where the constraints are approximately fulfilled: the projection can lead to minor violations of the constraints.

4.2. Manifold Projection

In order to implement the numerical scheme, we need a method for projecting points onto the manifold. We do this by defining projection as a search for the nearest point on the manifold. Specifically, let $\hat{\mathbf{x}}_t$ denote a sample from the distribution in embedding space; we now seek $\hat{\theta}_t$ such that $F(\hat{\theta}_t) = \mathrm{proj}_{\mathcal{M}}[\hat{\mathbf{x}}_t]$. We perform the projection in a direct manner by seeking

$$\hat{\theta}_t = \arg\min_{\theta_t} \left\|\hat{\mathbf{x}}_t - F(\theta_t)\right\|^2 \qquad \text{s.t.} \qquad \theta_t \in \boldsymbol{\Theta} \ , \tag{15}$$

where the constraints correspond to the joint limits. Solving this problem corresponds to finding a pose in a kinematic skeleton such that the joint positions are as close as possible to a given set of positions. This is known as *inverse kinematics* (Erleben et al., 2005) in the animation and robotics literature. As this is an important tool in much applied research, much work has gone into finding good solvers; we apply a projected steepest descent with line-search (Nocedal and Wright, 1999), as empirical results have shown it to be both fast and stable (Engell-Nørregård and Erleben, 2011). The search is started in θ_{t-1}, which practically ensures that a good optimum is found as the numerical simulation of Brownian motion only makes small incremental changes to the previous pose.

The optimisation problem in eq. 15 is defined as finding a set of joint angles corresponding to the projected point on the manifold. This shows that while our model is phrased spatially, it can be implemented in terms of joint angles in kinematic skeletons, which simplifies development.

5. Experiments

Having designed a numerical scheme, we now experimentally validate the least-committed spatial motion model by 1) comparing it to a least-committed model in joint angle space and 2) showing how the model can be extended to include knowledge of the environment. First, we briefly describe the tracking system where the motion model is used.

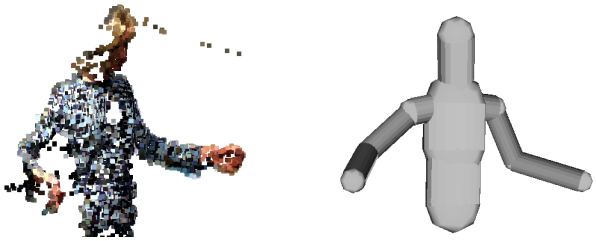


Figure 5: Left: an input data example. Noisy three dimensional points are scattered around the surface of the human body. Right: the skin model. Each bone is assigned a capsule and the collection of capsules describes the skin.

5.1. The Articulated Tracking System

As previously mentioned we build an articulated tracking system using a particle filter (Cappé et al., 2007). For the predictive model, $p(\theta_{t+1}|\theta_t)$, we will compare different models in the following sections. We describe the likelihood system next; this likelihood was previously described in (Hauberg and Pedersen, 2011a).

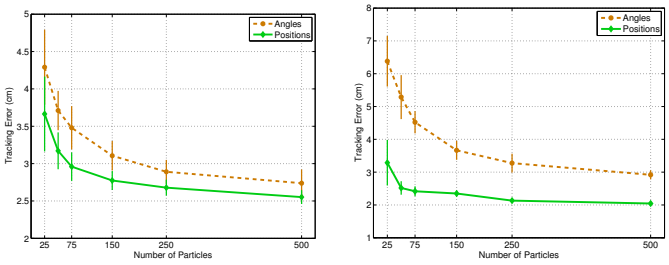
We use a small baseline consumer stereo camera[1] for acquiring data. At each time instance we, thus, get a set of three dimensional points $\mathbf{Z}_t = \{\mathbf{z}_t^{(1)}, \dots, \mathbf{z}_t^{(K)}\}$ that are mostly scattered around the surface of the human as well as around the surrounding environment (see fig. 5). In order to compare a given pose hypothesis θ_t to this data, we need a description of the surface of the pose. We assign a capsule to each bone in the skeleton with a radius corresponding to the width of the bone. This collection of capsules will serve as our surface (or skin) model (see fig. 5). We then define our likelihood measure as

$$p(\mathbf{Z}_t|\theta_t) \propto \exp\left(-\frac{\sum_i \|\mathbf{z}_t^{(i)} - \text{proj}_{\text{skin}(\theta_t)}(\mathbf{z}_t^{(i)})\|^2}{2\sigma^2}\right), \quad (16)$$

where σ is a parameter and $\text{proj}_{\text{skin}(\theta_t)}(\cdot)$ denotes projection of a point onto the surface of the pose parametrised by θ_t. This projection can easily be performed in closed-form as the skin consists of a set of capsules.

5.2. Experiment 1: Comparing Priors

In our first experiment, we compare the Brownian motion model in angle space with the Brownian motion model on the kinematic manifold. In both models, we scale the individual coordinates to encode that some joints move more than others. For both models, the scaling parameters are learned from separate training data. We perform tracking on an image sequence where a person is standing in place while waving a stick around. The sequence consists of 300 frames and the tracking is manually initialised. In general, both motion models allows for successful tracking of the motion, except for the part where the person moves both arms behind the head; here the data



(a) **(b)**

Figure 7: A comparison of Brownian motion in joint angle space versus Brownian motion on the kinematic manifold. The latter consistently outperforms the angular model. The vertical lines correspond to the standard deviation of the error measure over several runs of the particle filter, while the curve itself corresponds to the mean value.

do not provide strong enough clues for successful tracking. This is shown in fig. 6, where several frames are available; frame 192 shows the just mentioned situation. The angular Brownian motion is able to capture the trends of the motion, but it is rarely very accurate. The spatial Brownian motion, on the other hand, captures the motion very well. This is evident in both fig. 6 and in the supplementary film.

In order to quantify the above observations, we place markers on the arms of the person and estimate their three dimensional position using a commercial motion capture system[2]. As an error measure, we measure the average distance between the motion capture markers and the capsule skin of the estimated pose. This measure is then averaged across frames, such that the error measure becomes

$$\mathcal{E}(\theta_{1:T}) = \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M} \|(\text{skin}(\theta_t) - \mathbf{v}_{mt}\| \ , \quad (17)$$

where $\|\text{skin}(\theta_t) - \mathbf{v}_{mt}\|$ is the shortest Euclidean distance between the m^{th} motion capture marker and the skin at time t. We vary the number of particles from 25 to 500 and report this error measure for both prior models in fig. 7a. As can be seen, the Brownian motion model on the kinematic manifold consistently outperforms the angular Brownian motion model. This is also visually evident in the supplementary film.

We repeat the above experiment on a different sequence where the person is standing in place while moving his upper body. The resulting errors are shown in fig. 7b and selected frames are available in fig. 8. Again, the results clearly shows that the Brownian motion on the kinematic manifold improves results noticeably compared to the angular Brownian motion. This is also evident in the supplementary film.

5.3. Experiment 2: Object Interaction

To illustrate models that incorporate environmental knowledge, we replicate an experiment suggested by Kjell-

[1] http://www.ptgrey.com/products/bumblebee2/

[2] http://phasespace.com/

Frame 122 / 300 Frame 192 / 300 Frame 250 / 300 Frame 294 / 300

Figure 6: Selected frames from the tracking results using two different priors. The tracking is performed using 75 particles. The top row contains frames from the angular Brownian motion model, and the bottom row contains frames from the Brownian motion model on the kinematic manifold.

Frame 31 / 300 Frame 123 / 300 Frame 192 / 300 Frame 292 / 300

Figure 8: Selected frames from the tracking results using two different priors. The tracking is performed using 75 particles. The top row contains frames from the angular Brownian motion model, and the bottom row contains frames from the Brownian motion model on the kinematic manifold.

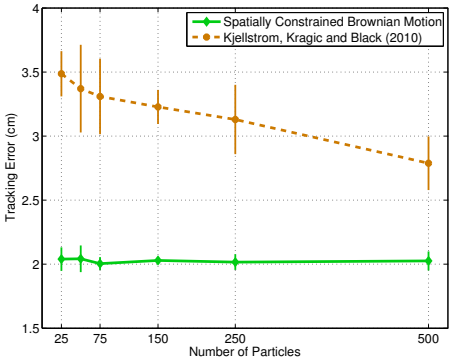


Figure 10: Tracking error for the spatially constrained Brownian motion used for modelling object interaction.

ström et al. (2010): the tracked person keeps both hands on a stick-like object and a separate tracking system is used to determine the position of the object. This knowledge can then be used to constrain the tracker by enforcing the hands to be on the object. Kjellström et al. achieve this by sampling for the angular Brownian motion and rejecting those samples where the attained hand positions are too far away from the object. While this strategy works, the need for brute-force techniques such as rejection sampling clearly shows that the joint angle space is not well-suited for this type of models. In contrast it is straightforward to model this problem in the spatial domain as described in sec. 3.3.

In the experiment, we track a person waving a stick in a sword-fighting manner. We attain the position of the stick by placing motion capture markers at the end-points. We then compare the rejection sampling strategy of Kjellström et al. with our spatial model. In fig. 9 we show selected frames from the sequence with results from the two trackers. As can be seen, both methods provides fairly good results, though the rejection sampling looses track of the arms in some frames (frame 103 in the figure). This error occurs when too many rejections are needed in order to fulfil the spatial constraints; in our implementation, we give up on fulfilling the constraints after 5000 rejections. As in the previous experiment, we plot the tracking error of the two methods against the number of particles (fig. 10). As can be seen the spatial model consistently achieves an error around 2 centimetre, while the rejection sampling approach is in the range of 3.5 to 3 centimetre. Computationally, the rejection sampling approach is fairly expensive: on average it needs 32.2 times as many resources as the spatial Brownian motion. Our spatial model is, thus, more accurate and computationally more efficient than current state-of-the-art.

6. Conclusion

We have discussed one of the most fundamental aspects of statistical models of human motion: the underlying metric. We have questioned the commonly used joint angle metric, which we feel has several unnatural properties. These occur as the metric specifically ignores both bone lengths and connectivity. As the metric greatly influences the statistical models, we have designed a metric that has a nice physical interpretation: it is the combined spatial distance travelled by the joints. This metric is tightly linked to both bone lengths and connectivity.

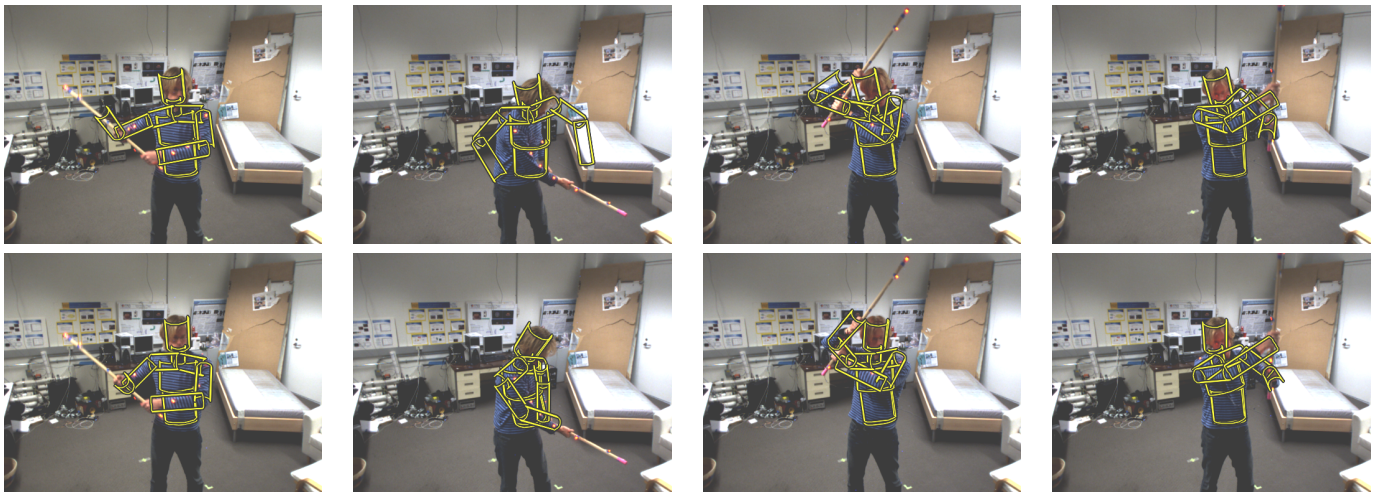
In order to design the metric, we introduced the kinematic manifold consisting of the position of all joints in the kinematic skeleton. This manifold allows us to apply techniques from Riemannian geometry when designing motion models. Our specific focus has been on predictive stochastic processes for describing human motion. We have defined a Brownian motion model on the kinematic manifold and demonstrated its usefulness. Moreover, since Brownian motion is the most basic building block of stochastic calculus, the work paves the way for even better models using more complex stochastic processes on manifolds.

We have applied the spatial Brownian motion model in an articulated tracking system, where we have theoretically and empirically shown that this model has a tighter covariance than the ordinary angular Brownian motion. In our experiments this leads to better tracking results as the new model performs better as a temporal low-pass filter. Furthermore, we have shown how interaction with the environment can trivially be modelled in the spatial domain, something that has previously required rather expensive techniques. These observations makes us believe the spatial domain is a more natural space for designing models of human motion.

To apply the Brownian motion model in an articulated tracking system, we used a particle filter, which requires us to simulate the stochastic differential equation of Brownian motion. To the best of our knowledge, no general-purpose numerical schemes exists for SDE's on manifolds. We have, thus, suggested an Euler-Heun scheme with projection steps for this simulation. This is a general scheme that allows the stochastic process to be simulated on other embedded Riemannian manifolds. Our approach can, thus, be carried on to other domains than human motion analysis. It is interesting to note that while Brownian motion is strongly linked to the underlying metric, the numerical scheme never requires distances to be calculated. This simplifies development substantially.

With our focus on Brownian motion, we have derived a motion agnostic model. As previously mentioned, motion specific models are often crafted by learning manifolds to which the motion is confined. An obvious next step is, thus, to learn a submanifold of the kinematic manifold $\mathcal{M}$ using e.g. *Principal geodesic analysis* (Fletcher et al., 2004) or *Geodesic PCA* (Huckemann et al., 2010). This can then be used to restrict the tracking system.

In this paper, we have focused exclusively on models of human motion. The Brownian motion model is, however, applicable to many other domains. Since the suggested numerical scheme works for any embedded Riemannian manifold, our work is directly transferable.



Frame 29 / 300 Frame 103 / 300 Frame 208 / 300 Frame 291 / 300

Figure 9: Selected frames from the tracking results using the spatially constrained motion models for object interaction. The top row corresponds to the rejection sampling approach by Kjellström et al. (2010) and the bottom row corresponds to our spatial model.

References

Abend, W., Bizzi, E., Morasso, P., 1982. Human arm trajectory formation. Brain 105 (2), 331–348.

Belkin, M., Niyogi, P., 2003. Laplacian Eigenmaps for dimensionality reduction and data representation. Neural Computation 15 (6), 1373–1396.

Brubaker, M. A., Fleet, D. J., Hertzmann, A., 2010. Physics-based person tracking using the anthropomorphic walker. International Journal of Computer Vision 87 (1-2), 140–155.

Cappé, O., Godsill, S. J., Moulines, E., 2007. An overview of existing methods and recent advances in sequential Monte Carlo. Proceedings of the IEEE 95 (5), 899–924.

Carreira-Perpinan, M. A., Lu, Z., 2007. The Laplacian Eigenmaps Latent Variable Model. JMLR W&P 2, 59–66.

Engell-Nørregård, M., Erleben, K., January 2011. A projected backtracking line-search for constrained interactive inverse kinematics. Computers & Graphics In Press, Accepted Manuscript.

Erleben, K., Sporring, J., Henriksen, K., Dohlmann, H., August 2005. Physics Based Animation. Charles River Media.

Fletcher, T. P., Lu, C., Pizer, S. M., Joshi, S., 2004. Principal geodesic analysis for the study of nonlinear statistics of shape. Trans. on Medical Imaging 23 (8), 995–1005.

Grochow, K., Martin, S. L., Hertzmann, A., Popović, Z., 2004. Style-based inverse kinematics. ACM Transaction on Graphics 23 (3), 522–531.

Hairer, E., Lubich, C., Wanner, G., 2004. Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations. Springer.

Hauberg, S., Pedersen, K. S., 2011a. Predicting articulated human motion from spatial processes. International Journal of Computer Vision 94, 317–334.

Hauberg, S., Pedersen, K. S., 2011b. Stick it! articulated tracking using spatial rigid object priors. In: Kimmel, R., Klette, R., Sugimoto, A. (Eds.), ACCV 2010. Vol. 6494 of Lecture Notes in Computer Science. Springer, Heidelberg, pp. 758–769.

Hauberg, S., Sommer, S., Pedersen, K. S., September 2010. Gaussian-like spatial priors for articulated tracking. In: Daniilidis, K., Maragos, P., , Paragios, N. (Eds.), ECCV 2010. Vol. 6311 of Lecture Notes in Computer Science. Springer, Heidelberg, pp. 425–437.

Herda, L., Urtasun, R., Fua, P., 2004. Hierarchial implicit surface joint limits to constrain video-based motion capture. In: Pajdla, T., Matas, J. (Eds.), Computer Vision - ECCV 2004. Vol. 3022 of LCNS. Springer, pp. 405–418.

Hsu, E., February 2002. Stochastic Analysis on Manifolds. American Mathematical Society.

Huckemann, S., Hotz, T., Munk, A., 2010. Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. Statist. Sinica 20 (1), 1–58.

Kalman, R., 1960. A new approach to linear filtering and prediction problems. Transactions of the ASME-Journal of Basic Engineering 82 (D), 35–45.

Kjellström, H., Kragić, D., Black, M. J., 2010. Tracking people interacting with objects. In: CVPR '10: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 747–754.

Kloeden, P. E., Platen, E., May 1992. Numerical Solution of Stochastic Differential Equations. Springer.

Lu, Z., Carreira-Perpinan, M., Sminchisescu, C., 2008. People Tracking with the Laplacian Eigenmaps Latent Variable Model. In: Platt, J. C., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems 20. MIT Press, pp. 1705–1712.

Mardia, K. V., Jupp, P. E., January 1999. Directional Statistics. Wiley.

Miller, M. I., Srivastava, A., Grenander, U., 1995. Conditional-mean estimation via jump-diffusion processes in multiple target tracking/recognition. IEEE Transactions on Signal Processing 43, 2678–2690.

Morasso, P., April 1981. Spatial control of arm movements. Experimental Brain Research 42 (2), 223–227.

Nocedal, J., Wright, S. J., 1999. Numerical optimization. Springer Series in Operations Research. Springer-Verlag.

Øksendal, B., 2000. Stochastic Differential Equations: An Introduction with Applications, 5th Edition. Springer.

Poppe, R., 2007. Vision-based human motion analysis: An overview. Computer Vision and Image Understanding 108 (1-2), 4–18.

Sato, K.-I., 1999. Lévy Processes and Infinitely Divisible Distributions. Cambridge University Press.

Sidenbladh, H., Black, M. J., Fleet, D. J., 2000. Stochastic tracking of 3d human figures using 2d image motion. In: Proceedings of ECCV'00. Vol. II of Lecture Notes in Computer Science 1843. Springer, pp. 702–718.

Sminchisescu, C., Jepson, A., 2004. Generative modeling for continuous non-linearly embedded visual inference. In: ICML '04: Proceedings of the twenty-first international conference on Machine learning. ACM, pp. 759–766.

Urtasun, R., Fleet, D. J., Fua, P., 2006. 3D People Tracking with Gaussian Process Dynamical Models. In: CVPR '06: Proceedings

of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 238–245.

Urtasun, R., Fleet, D. J., Hertzmann, A., Fua, P., 2005. Priors for people tracking from small training sets. In: Tenth IEEE International Conference on Computer Vision. Vol. 1. pp. 403–410.

Wang, J. M., Fleet, D. J., Hertzmann, A., 2008. Gaussian Process Dynamical Models for Human Motion. IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2), 283–298.

Yamamoto, M., Yagishita, K., 2000. Scene constraints-aided tracking of human body. In: CVPR. Published by the IEEE Computer Society, pp. 151–156.

Zhang, F., Goodlett, C., Hancock, E., Gerig, G., 2007. Probabilistic fiber tracking using particle filtering and von mises-fisher sampling. In: Yuille, A., et al. (Eds.), Energy Minimization Methods in Computer Vision and Pattern Recognition. Vol. 4679 of Lecture Notes in Computer Science. Springer, pp. 303–317.