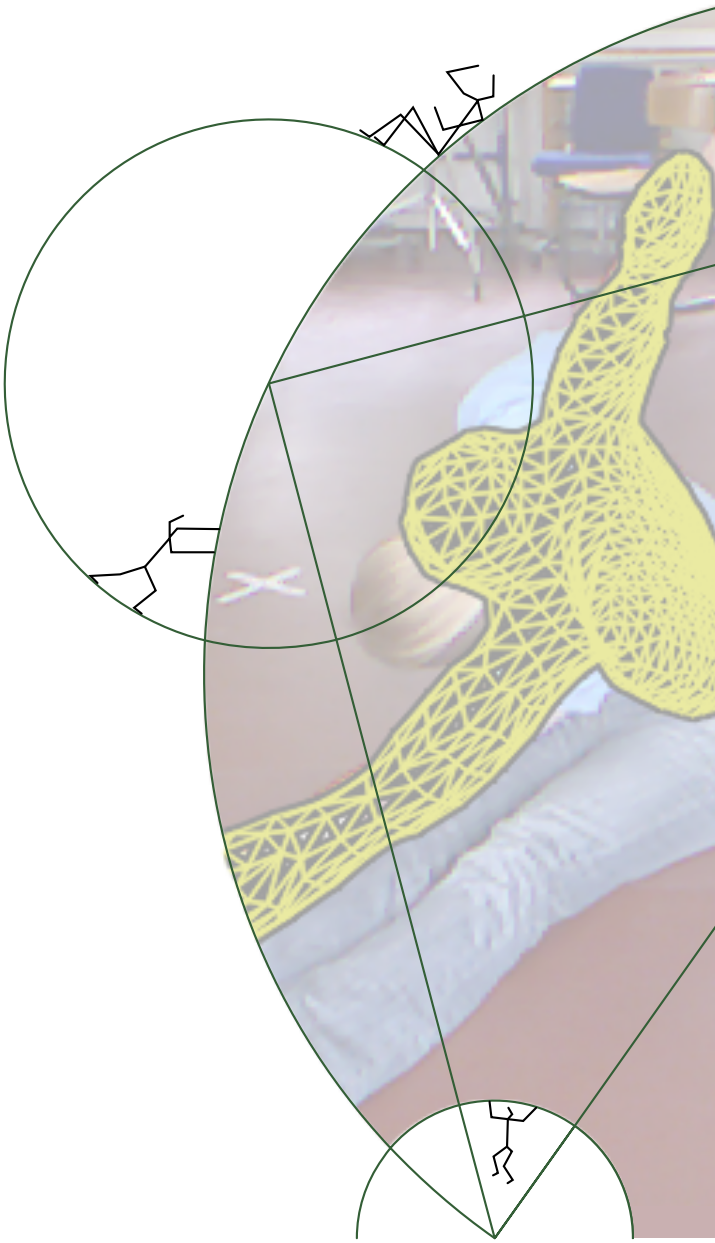
FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

PhD Thesis

Søren Hauberg

Spatial Models of Human Motion



Supervisor: Kim Steenstrup Pedersen

October 31, 2011

"Simplicity is the ultimate sophistication"

Leonardo da Vinci

Notation

In this thesis the following notation will be used mostly consistently; in individual papers notation might differ slightly. In general, lower case **bold face** will denote vectors, while upper case bold face denotes matrices.

θ_t	The joint angles of the kinematic skeleton at time t.
$F(\theta_t)$	The forward kinematics function applied to the pose vector θ_t.
$\mathbf{\Theta}$	The set of legal joint angles.
$\mathcal{M}$	The kinematic manifold.
$\mathbf{J}_\theta$	The Jacobian of the forward kinematics function evaluated in θ.
$\mathbf{g}_t$	The goal position of selected limbs.
$\mathbf{Z}_t$ or $\mathbf{X}_t$	The observation at time t.
$a \propto b$	a is proportional to b; often the constant of proportionality is irrelevant.
$a \equiv b$	a is defined to be equal to b.
$\mathbf{A}_{1:t}$	Short-hand notation for the sequence $\{\mathbf{A}_1, \ldots, \mathbf{A}_t\}$.

Thesis Overview

This document is a PhD thesis in Computer Science from the University of Copenhagen, Denmark. The thesis is concerned with statistical models of human motion that are suitable for performing articulated tracking. Our main hypothesis is that it is not only possible but also beneficial to create motion models in the spatial domain, rather than in the traditional joint angle space. The thesis contributes to the field by providing both statistical models expressed in the spatial domain as well as practical algorithms for working with these models. Furthermore, several of the contributed algorithms are suitable for solving general filtering problems on Riemannian manifolds.

This thesis consists of a set of papers written in the last three years along with an introductory text that motivates the modelling decision behind the papers. The papers are presented in their original form with an exception of page numbers. The introductory text is given in the first chapter of the thesis, while the papers are in the following ones. Finally, a few finishing remarks are given in the last chapter.

All Stories have a Beginning...

This thesis started with me being told to implement an *off the shelf* articulated tracking algorithm; *"then we'll have something to compare with"* my supervisor said. This was a frustrating experience: too much problem-specific tweaking was needed to get things working. The main practical problem I kept facing was that the prior (or regularisor) was very sensitive, which made parameter estimation unstable. We came up with a quick "hack" [Hauberg et al., 2009], where we changed the state space of the tracker from the joint angles into the coordinates of the hands and head; the intuition was that this reduced the dimensionality of the state space, which should improve stability. We then used an inverse kinematics system to compute joint angles from hand and head positions. Surprisingly, this worked remarkably well. The major focus of the thesis then became to determine why it worked. This opened up a box of fascination that finally sent me into the interplay of statistics and geometry.

Included Papers

Most chapters of this thesis are papers that I have first-authored during my PhD studies. These papers are as follows.

Søren Hauberg and Kim Steenstrup Pedersen. Predicting Articulated Human Motion from Spatial Processes. *International Journal of Computer Vision*, 94:317–334, 2011a.

Søren Hauberg, Stefan Sommer, and Kim Steenstrup Pedersen. Gaussian-like Spatial Priors for Articulated Tracking. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV*, volume 6311 of *LNCS*, pages 425–437. Springer, 2010.

Søren Hauberg and Kim S. Pedersen. Stick It! Articulated Tracking using Spatial Rigid Object Priors. In *ACCV 2010*. Springer-Verlag, 2010.

Søren Hauberg and Kim Steenstrup Pedersen. Data-Driven Importance Distributions for Articulated Tracking. In Yuri Boykov et al., editors, *Energy Minimization*

Methods in Computer Vision and Pattern Recognition, Lecture Notes in Computer Science. Springer, 2011b.

Søren Hauberg, Stefan Sommer, and Kim Steenstrup Pedersen. Natural Metrics and Least-Committed Priors for Articulated Tracking. *Image and Vision Computing* **(under review)**, 2011a.

Søren Hauberg, François Lauze, and Kim Steenstrup Pedersen. Unscented Kalman Filtering on Riemannian Manifolds. *Journal of Mathematical Imaging and Vision* **(under review)**, 2011b.

Excluded Papers

While I admit to being a simple-minded creature, I cannot claim to have a one track mind. While this thesis is concerned with articulated tracking I have also collaborated with others on other projects; this has resulted in the following publications. These papers make interesting points and contributions, but, alas, they either do not fit into the story told in this thesis or I was not first author, so they had to be excluded.

Aasa Feragen, Søren Hauberg, Mads Nielsen, and François Lauze. Means in spaces of tree-like shapes. In *International Conference on Computer Vision*, 2011.

Stefan Sommer, François Lauze, Søren Hauberg, and Mads Nielsen. Manifold Valued Statistics, Exact Principal Geodesic Analysis and the Effect of Linear Approximations. In K. Daniilidis, P. Maragos, , and N. Paragios, editors, *ECCV '10: Proceedings of the 11th European Conference on Computer Vision*, Lecture Notes in Computer Science, pages 43–56. Springer, Heidelberg, September 2010.

Rune Møllegaard Friborg, Søren Hauberg, and Kenny Erleben. GPU Accelerated Likelihoods for Stereo-Based Articulated Tracking. In *The CVGPU workshop at European Conference on Computer Vision*, 2010.

Søren Hauberg, Jérome Lapuyade, Morten Engell-Nørregård, Kenny Erleben, and Kim Steenstrup Pedersen. Three Dimensional Monocular Human Motion Analysis in End-Effector Space. In Daniel Cremers et al., editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, Lecture Notes in Computer Science, pages 235–248. Springer, August 2009.

Peter Mysling, Søren Hauberg, and Kim Steenstrup Pedersen. An Empirical Study on the Performance of Spectral Manifold Learning Techniques. In T. Honkela et al., editors, *ICANN 2011*, volume 6791 of *Lecture Notes in Computer Science*, pages 347–354. Springer, Heidelberg, 2011.

Anders Boesen Lindbo Larsen, Søren Hauberg, and Kim Steenstrup Pedersen. Unscented Kalman Filtering for Articulated Human Tracking. In Anders Heyden and Fredrik Kahl, editors, *Image Analysis*, volume 6688 of *Lecture Notes in Computer Science*, pages 228–237. Springer, 2011.

Morten Engell-Nørregård, Søren Hauberg, Jérome Lapuyade, Kenny Erleben, and Kim Steenstrup Pedersen. Interactive Inverse Kinematics for Monocular Motion Estimation. In *Proceedings of VRIPHYS'09*, 2009.

Praises

During the last three years much have been discovered and experienced, but my wonderful girlfriend stands out as the single most important discovery. Wrestling for a *Cold Bitch* with my then-soon-to-be-girlfriend, Aasa, was a fight that makes this thesis feel like a walk in the park. My unknown legend! I owe her a great many praises for many things, but in particular for the introduction of *the PhD account*, which allowed me to behave like an idiot whenever such behaviour seemed rational.

My family also deserves many praises for always being immensely helpful. In particularly, for the time immediately after August 2nd, 2011, when we discovered we no longer had a home. Help was on the way!

Inspiration has always been part of family life. As a child, my father and I crafted many strange, yet useful, structures in LEGO together. He always reminded me that being playful, and not always make the same assumptions as everyone else, is the best foundation for creativity. I guess, this is part of the reason why I keep returning to the basics of the problems I work on. However, being playful often means I find myself on unstable grounds. Here my mother has always been a constant source of stability and support, which has ensured that my experiments never felt dangerous.

One of the great things about working is that you get to take breaks. Here friends play essential parts, and I want to thank mine for making me consume delicious beverages, carry crazy-heavy objects or just having nonsense conversations that shall never be remembered. I also want to apologise for not spending more time this way.

Another good thing about working is that you get to work with people. Colleagues quickly becomes a large part of your life and the Image Group is no exception. Always a fun and friendly place, where I've most likely spend more time just hanging out then I'd like to admit. Here my different office-mates have played delightful parts and I thank for all the disturbance. The cake club and the Friday beer should be thanked for helping me gain weight. Likewise the running club deserves thanks for helping me run away from potential weight gains.

One colleague, unsurprisingly, stands out: my supervisor, Kim, deserves great thanks not only for standing my strange thoughts and ideas, but also for "baby lobsters" and the "always an open door"-style of supervision.

Speaking of supervisors, I want to thank Ruzena Bajcsy for hosting me during my stay in Berkeley. What a wonderful and insightful woman! And I must not forget to mention the home-made cakes.

Finally, I want to thank my previous supervisor, Peter Johansen, for introducing me to computer vision and showing me that computer science can be a playful form of art.

Thank You

Summary

The thesis is concerned with estimation of human poses in image sequences (articulated tracking). The focus is on designing spatial models of human motion, when poses are represented using the kinematic skeleton. This data structure enforces constant bone length, which makes the space of joint positions a non-linear Riemannian manifold. The thesis contributes with two basic models of how joint positions move and a series of algorithms and specialisations.

The first model is a Bayesian interpretation of inverse kinematics. Assuming that we know where the human wishes to place his/her hands, we derive the natural conditional distribution of joint angles that reaches this goal. To work with the model in practice, we derive a second order approximation to the bootstrap filter. This allows us to create both general-purpose activity independent models as well as a motion specific physiotherapeutic model and a model of motion during object interaction.

The second model takes a geometric point of view to the problem. We show that the space of all joint positions is a Riemannian manifold, which we call the *kinematic manifold*. In this model, inverse kinematics can be viewed as a projection operator from embedding space to the manifold. The metric on this manifold is the physically natural measure of how far individual joints move in the world coordinate system, which provides a good base for doing statistics. For motion models we both derive a Brownian motion model and a projected model, which can be viewed as an efficient approximation to the Brownian model. These models allow us to easily control the variance of joint positions, which gives rise to good low-pass filtering properties. This is an inherently hard problem in the joint angle domain. Furthermore, we show how the model can be extended to describe interactions with the environment and how it can be used to provide efficient data-driven importance distributions for particle filters.

We also contribute a generalisation of the unscented Kalman filter to Riemannian manifolds. The resulting filter is very general and only requires information about how to compute parallel transports, exponential maps and logarithm maps. As these can be computed numerically on many manifolds, the suggested filter is widely applicable. A nice property of the filter is that it is remarkably simple, which is a hint that perhaps Riemannian models need not be overly complicated in practice.

Contributions

The overall contribution of the thesis is an argument that human motion should be described in terms of spatial joint positions. This is in contrast to existing models that describe how joint angles change over time. We argue for models of joint positions using neurology, filtering theory as well as practical considerations. We phrase several mathematical models and practical algorithms that realise the goal of modelling joint positions. Several of the presented algorithms are general and can be applied in other scenarios than human motion.

Specifically, the thesis makes the following novel contributions to the field:

- A probabilistic interpretation of inverse kinematics, which is suitable for tracking [21].

- A geometric interpretation of the kinematic skeleton in the kinematic manifold. In this line of thinking, inverse kinematics becomes a projection operator [24]. The metric on this manifold becomes a physically natural measure of the size of motions [26].
- Different models of human motion during interaction with the environment [19, 21].
- An approximation strategy for designing data-driven importance distributions for articulated tracking [20].
- A Brownian motion model on embedded manifolds along with a novel numerical scheme for simulating the underlying manifold-valued stochastic differential equation [26].
- A Riemannian generalisation of the unscented Kalman filter, which provides general-purpose tools for both filtering and optimisation on Riemannian manifolds [25].

Resumé (in Danish)

Denne afhandling omhandler estimering af menneskers positurer i billedsekvenser (artikuleret tracking). Fokus er på design af rumlige statistiske modeller af menneskets bevægelser, når disse repræsenteres vha. det kinematiske skelet. Denne datastruktur håndhæver at knoglers længde forbliver konstant, hvilket gør at rummet af samtlige ledpositioner bliver en ikke-lineær Riemannsk mangfoldighed. Afhandlingen bidrager med to fundamentale modeller af hvorledes ledpositioner bevæger sig, samt flere algoritmer og specialiseringer.

Den første model giver en Bayesiansk fortolkning af invers kinematik. Ud fra en antagelse om at vi ved hvor mennesket ønsker at placere sine hænder udleder vi den naturlige betingede fordeling af ledvinkler der sikrer at hænderne når sine mål. Som en praktisk algoritme udleder vi en anden ordens approksimation af Bootstrap filteret. Dette giver os mulighed for at konstruere både generalle modeller samt en specialiseret fysioterapeutisk model og en model for bevægelser mens mennesket interagerer med omgivelserne.

Den anden model anskuer problemet fra et geometrisk synspunkt. Vi viser at rummet af alle ledpositioner udgør en Riemannsk mangfoldighed, som vi kalder *den kinematiske mangfoldighed* (*eng: kinematic manifild*). Under denne model bliver invers kinematik en projektionsoperator. Afstandsmålet på denne mangfoldighed svarer til det fysisk naturlige mål af hvor langt de enkelte ledpositioner bevæger sig, hvilket skaber et godt grundlag for statistiske overvejelser. Vi opskriver både en Brownsk bevægelsesmodel samt en projiceret model, der kan opfattes som en effektiv approksimation af den Brownske model. Med disse modeller bliver det let af kontrollere variansen af ledpositionerne, hvilket skaber grundlag for gode lavpasfiltre. Dette er fundamentalt svært for modeller udtrykt i termer af ledvinkler. Vi viser desuden hvordan modellerne kan udvides til at beskrive interaktion med omgivelserne samt til at designe data-drevne *importance* fordelinger til effektive partikel filtre.

Vi bidrager desuden med en generalisering af *the unscented Kalman filter* til Riemannske mangfoldigheder. Dette filter er yderst generelt og kræver kun at parallel transporter, eksponential afbildninger og logaritme afbildninger kan beregnes. Eftersom disse ofte kan beregnes numerisk kan filteret anvendes bredt. I praksis er filteret meget enkelt, hvilket giver os tro på at Riemannske modeller ikke behøver at være besværlige at arbejde med i praksis.

Videnskabelige bidrag

Det overordnede videnskabelige bidrag i denne afhandling er en argumentation for at menneskelige bevægelser beskrives bedst ved rumlige ledpositioner. Dette står i kontrast til eksisterende modeller der beskriver hvorledes ledvinkler ændres over tid. Vi argumenterer for ledpositionsmodeller ud fra både psykologiske, filtreringsteoretiske samt praktiske overvejelser. Vi opstiller flere matematiske modeller samt praktiske algoritmer der realiserer målet om at beskrive ledpositioners bevægelse. Flere af de opstillede algoritmer er meget generelle og vil kunne benyttes i andre scenarier.

Mere konkret har afhandlingen følgende videnskabelige bidrag:

- En sandsynlighedsteoretisk fortolkning af invers kinematik, som er velegnet til tracking [21].

- En geometrisk fortolkning af det kinematiske skelet, hvilket giver den kinematiske mangfoldighed. Fra denne synsvinkel kan invers kinematik opfattes som en projektionsoperator [24]. Afstandsmålet på mangfoldigheden giver et naturligt fysisk mål for størrelsen af bevægelser [26].
- Forskellige modeller af menneskets bevægelser under interaktion med omgivelserne [19, 21].
- En generel approksimationsstrategi til at designe data-drevne *importance* fordelinger til artikuleret tracking [20].
- En model for Brownske bevægelser på indlejrede mangfoldigheder samt en numerisk metode til at simulere den underliggende stokastiske partielle differential ligning [26].
- En Riemannsk generalisering af *the unscented Kalman filter*, hvilket giver et bredt anvendeligt værktøj til både filtrering og optimering på Riemannske mangfoldigheder [25].

Contents

Chapter 1

Modelling Concerns in Articulated Tracking

In this thesis we attack the problem of *articulated tracking*, which is also known as *human motion capture*. This is the process of estimating the pose of a person in each frame of an image sequence [41]; an example of the type of results we are looking for is shown in fig. 1.1. Uses of such tracking systems range from rehabilitation and biomechanics to film production and human computer interaction, e.g. in computer games. We can think of this problem as "fitting a skeleton" to the image data, which can be done in a multitude of ways as will be reviewed in the following section.

First, we briefly introduce the skeleton representation; a more detailed description will be provided in later chapters. As the basic pose representation, we use the *kinematic skeleton* [14], which is a "stick figure" model, where individual bones have constant length (see fig. 1.2). Due to the constant bone lengths, only the *joint angles*, i.e. the angles between connected bones, appear as degrees of freedom in the model; the vector containing all these angles is denoted θ_t, where the subscript indicates a temporal frame index. From joint angles, the position of each joint can be computed using *Forward Kinematics* [14]. We will denote these positions $F(\theta_t)$, where F denotes the *forward kinematics function*. This function consists of a series of rotations and translations – we will define it in a later chapter, but for now it suffices to note that it is highly non-linear.

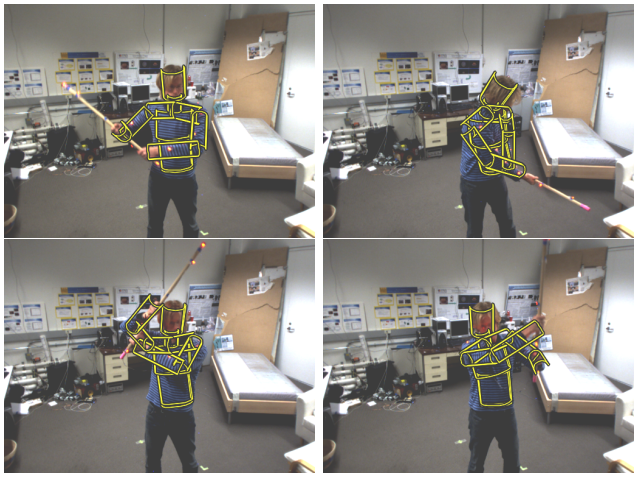


Figure 1.1 *Examples of results attained with an articulated tracker.*

1.1 Different Approaches

One of the first successful approaches to articulated tracking was that of Bregler and Malik [6], who phrased the problem as one of differential motion estimation. This approach is similar in spirit to the classical approach in *optical flow* [27], where changes in pixel intensities can be related to flow vectors from an intensity constancy assumption. Using the *twists* and *products of exponential maps* framework from the robotics literature [36], Bregler and Malik showed how to relate changes in pixel intensities to changes in joint angles. This leads to a differential equation which can be solved to estimate the human pose sequentially. In practice, this approach is not very flexible as it is hard to change the criterion being optimised, and the gradient descent style optimisation often diverges.

To overcome these issues, Sidenbladh et al. [45] suggested a Bayesian model consisting of two parts: 1) a *likelihood model* that relates the image information to the skeleton, and 2) a *motion prior* that encodes any information available about the observed motion.

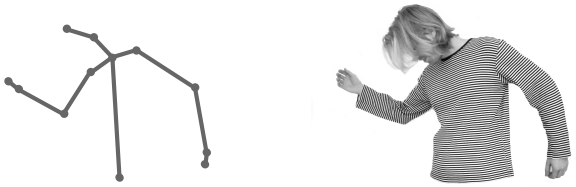


Figure 1.2 *Left: An illustration of the kinematic skeleton used to represent human poses. Bone lengths are kept constant such that joint angles are the only degrees of freedom in the model. Right: an image showing a human in the pose represented by the skeleton to the left.*

This split makes the approach highly flexible as both the likelihood and the motion prior can easily be replaced. To perform maximum a posteriori (MAP) estimation, Sidenbladh et al. suggested using a particle filter, which allows several possible optima to be traced simultaneously. This makes the approach very general, but as it is based on sampling, the computational demands grows exponentially with the dimensionality of the state space.

1.1.1 Dimensionality Reduction

One of the most fundamental issues with the general solution based on a particle filter is the dimensionality of the problem. In the most general case, we need to infer all joint angles in the human body, which often gives rise to 25–50 degrees of freedom [4, 10, 16, 45]. This makes the approach computationally demanding as well as potentially unstable due to the many chances of failure in high dimensional spaces. The most obvious solution to this problem is to reduce the number of degrees of freedom in the model, which has been done in many different ways.

Manifold Learning

One solution was presented in the original paper by Sidenbladh et al. [45]: from motion capture data of a specific activity (often *walking*) learn a low-dimensional probabilistic model and use this as the motion prior. This immediately lowers the dimensionality of the problem and tracking becomes fast and stable. In the seminal work of Sidenbladh et al., it was suggested to apply a linear motion model in a linear subspace of the joint angle domain that was learned using PCA. This fairly straight forward approach turned out to work quite well: better results could be attained using fewer computational resources. The authors, however, questioned the strategy by providing an experiment in their paper, where the image information was neglected by the tracker. The results, which are replicated in fig. 1.3, provide insights into the prior. As can be seen, the prior alone is actually able to track the motion for several frames without considering the data at all. This indicates that the prior provides a good description of the motion. The resulting tracker, however, simply "plays back the prior" rather than interpreting the data, so we should be careful when applying such priors to new sequences.

Since the work of Sidenbladh et al., the idea of learning the motion prior has been extended in several ways. Sminchisescu and Jepson [47] use Laplacian Eigenmaps [5] to learn a nonlinear motion manifold, and Lu et al. [33] use a Laplacian Eigenmaps Latent Variable Model [8]. The methods used for learning the manifolds assume that data are densely sampled on the manifold, which either requires vast amounts of data or a low-dimensional manifold [37]. In the mentioned papers [33, 47], a one-dimensional manifold corresponding to *walking* is studied. It remains to be seen if the approach scales to higher dimensional manifolds.

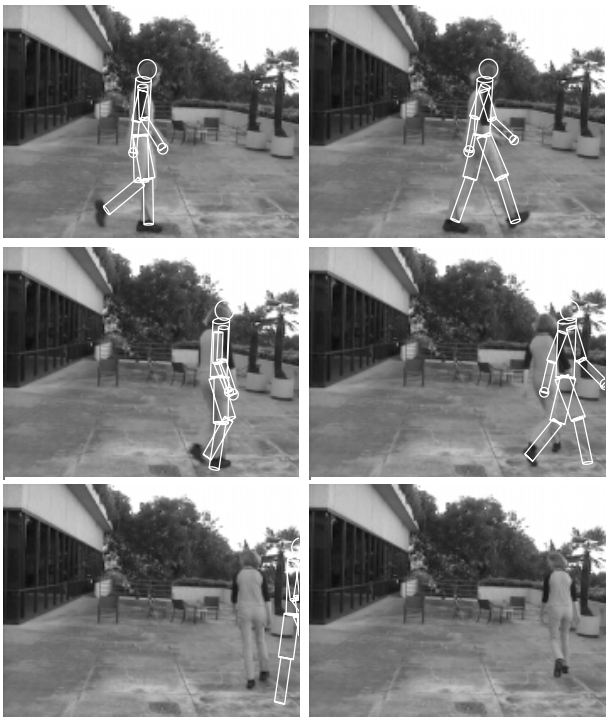


Figure 1.3 *How strong is the walking prior? Tracking results for frames 0, 10, 20, 30, 40, 50, when no image information is taken into account. Figure adapted from [45]* (courtesy of Michael J. Black).

Instead of just learning a manifold and restricting the tracking to this, it seems reasonable to also use the density of the training data on this manifold. Urtasun et al. [48] suggested to learn a prior distribution in a low dimensional latent space using a *Scaled Gaussian Process Latent Variable Model* (Scaled GPLVM) [18]. This not only restricts the tracking to a low dimensional latent space, but also makes parts of this space more likely than others. Our experience, however, seems to indicate that this type of prior is too "tight" to generalise. Fig. 1.4 shows a two-dimensional model learned with a Scaled GPLVM learned from walking data for three different persons. As can be seen, the model essentially learns a tri-modal distribution, with each mode corresponding to the walk-cycle of each person. As such, the model fails to learn the similarities between how different people walk, which indicates that the model will not be able to generalise to unseen data.

Environment Interaction

An alternative to learning a low dimensional model is to design a model with fewer degrees of freedom. One approach is to model the fact that most human motion happens during interaction with the environment: humans often touch the ground plane, pick up objects, lean against walls and so forth. This knowledge can

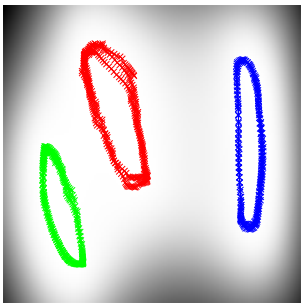


Figure 1.4 *A Scaled GPLVM model of walking performed by three different persons. Each point corresponds to a pose and colours indicate different persons. Note that the learned model is tri-modal, meaning it failed to capture the similarities between persons.* Figure courtesy of Anders Boesen Lindbo Larsen [32].

be used to reduce the degrees of freedom in the model, which can improve tracking quality and robustness.

An obvious phenomena to include in the model is interaction with the ground plane as this often is a flat surface, which simplifies the mathematical model. Yamamoto and Yagishita [53] do this using a linear approximation of the motion path by linearising the forward kinematics function. Since this is a highly non-linear function and motion paths in general are non-linear, this modelling decision seems to be made out of sheer practicality. Promising results are, however, shown on constrained situations, such as when the position and orientation of a persons feet is known.

When modelling interactions with the environment, inverse kinematics is an essential tool, as it provides a mapping from the spatial world coordinate system to the joint angle space. Rosenhahn et al. [43] use this to model interaction with sports equipment, such as bicycles and snowboards. They use the previously mentioned formalism of twists and products of exponential maps to derive an approximate gradient descent scheme for estimating poses. In a similar spirit, Kjellström et al. [30] model interaction with a stick-like object. They solve the inverse kinematics problem using rejection sampling in joint angle space, leading to a computationally expensive approach due to the high dimensionality of the joint angle space.

Physics-based Models

A further development in the direction of including the environment is to model the physics of the situation. One such approach is given by Vondrak et al. [51], who suggest using a physical simulation as part of the motion prior. This allows for modelling of human motion dynamics, ground contact and environment interaction, resulting in a high dimensional state space. To make optimisation feasible, an exemplar-based model is used to reduce the effective dimensionality of state space.

A more constrained model was suggested by Brubaker et al. [7] who describe a biomechanical model of *walking*. This is a three-dimensional extension of a planar model, which allows for modelling of both balance and ground contact. This gives a robust model that gives good results even when occlusions occur. The model does, however, not easily generalise to other types of motion.

1.1.2 Improved Optimisation Schemes

Instead of improving the models, one might ask if we simply need to improve the tools? Many have reported improved tracking results by replacing the particle filter with alternative optimisation tools.

One of the most popular choices is to use the *Annealed Particle Filter* suggested by Duetscher et al. [10]. This filter borrows ideas from simulated annealing and iteratively samples from a prior with smaller and smaller variance, while exponentially increasing the likelihood such that modes become more exaggerated. This forces the particles to focus more on the modes of the likelihood, which most often improves tracking quality substantially. One downside is that the Bayesian interpretation is lost, in the sense that we only get mode estimates and not an estimate of the entire *a posteriori* distribution. This can be of interest if the human pose is a latent variable in a larger system, such that the pose variable should be marginalised.

In a similar spirit, Gall et al. [15] suggested using a multi-layer approach, where simulated annealing is used in the first layer to roughly estimate mode positions and local optimisation is used in the second layer to refine the mode estimates. Again, the Bayesian interpretation is lost.

Another approach is to use the hierarchical structure of the kinematic skeleton to improve the particle filter. Bandouch and Beetz [3] suggest using a hierarchical sampling scheme such that the torso is fitted to the data before the arms. This approach is potentially more efficient, but it is more sensitive to local minima as one incorrectly fitted body part makes it impossible to fit the remaining parts.

One problem with the particle filter is that particles are often sampled directly from the motion prior. This disregards the likelihood such that particles are often far away from the likelihood modes. To overcome this issue Poon and Fleet [40] suggested using a *hybrid Monte Carlo filter*, where samples are moved closer to the modes of the likelihood according to the gradient of the likelihood. In some sense this is similar to the framework suggested by Gall et al. [15], but the hybrid Monte Carlo approach preserves the Bayesian interpretation. On the downside, the filter is highly specialised to the specific likelihood model.

1.1.3 Non-visual Sensors

One of the sources for many problems in visual articulated tracking is that certain joint configurations are inherently hard to determine visually. For example, the rotation of limbs around their own axes are practically impossible to determine from silhouette images. One way around such difficulties is to use extra sensors that are capable of determining this information. One example of this approach is the work of Pons-Moll et al. [38] where inertial orientation sensors are used to determine the orientation of hands and feet. This constraint is incorporated into a gradient descent scheme similar to the work of Bregler and Malik [6]. A later version of the work extended this approach to be part of a more robust particle filter [39].

1.1.4 Part Detection

Another way to constrain the tracker, and thereby reduce the degrees of freedom, is to detect positions of selected body parts and only consider pose estimates that are in tune with the detections.

One way to achieve this was suggested by Sigal et al. [46], who loosen the hard "constant bone length" constraint of the kinematic skeleton to a soft constraint. Individual limbs are then detected and combined to a complete pose using *loopy belief propagation* in a Bayesian network, where nodes correspond to joint positions and edges to connections in the skeleton structure. This effectively allows for tracking without the need for initialisation. A similar approach was taken by Ramanan and Forsyth [42], who also show that the method is robust with respect to occlusions and that it can recover from failures.

Ganapathi et al. [16] suggest detecting selected body parts and using inverse kinematics to relate the detected positions to the joint angles of the skeleton. This is incorporated in a Bayesian tracker using a probabilistic interpretation of inverse kinematics that is based on the unscented transform [28]. The resulting system is simple enough to be able to run in almost real time.

The, so far, most successful approach to tracking based on part detectors is the work of Shotton et al. [44]. In the previous approaches, the part detectors looked at small patches; in contrast, Shotton et al. classify every single pixel into a set of body parts. From this, body parts are detected by looking at clusters of pixels of the same body part and a skeleton model is extracted. The resulting tracker runs in real time, is quite robust to tracking errors and requires no initialisation. The solution is, however, designed for the Microsoft Kinect depth camera and does not immediately generalise to colour cameras.

1.2 Our Approach

In this thesis we start from the work of Sidenbladh et al. [45], but instead of building upon it, we stop to re-examine the most basic motion prior. This should encode whatever knowledge of the motion we posses, so we start by reviewing neurological studies of human motion. As it turns out, these studies points in a direction that gives rise to good temporal low-pass filters, as well as models that makes it strikingly easy to model interaction with the environment and so forth.

In the rest of this chapter, we discuss some of the modelling thoughts that went into designing our new motion priors. The remaining chapters present individual models and applications of these.

1.3 How do Humans Move?

As we will focus on the motion prior, we first need to acquire some insights into how people move; this can then serve as a design guideline. For these insights, we turn to experimental neurology, where the subject of motion planning has been researched for years.

One of the first experiments was performed by Morasso [35], who measured joint angles and hand positions while people moved their hand to a given target. Later, Abend et al. [1] expanded on this experiment, and introduced obstacles that the hand had to move around. Both made the same observation: joint angles show great variation between both persons and trials, while hand positions consistently followed the same path with low variation. Recently, Ganesh [17] made similar observations for actions like *punching* and *grasping* in a computer vision based tele-immersion system. One interpretation of these results is that the hand trajectories describe a *generic*, i.e. person independent, part of the motion.

This result is a strong indication that humans plan body motion in terms of spatial hand trajectories rather than joint angles. It, thus, seems reasonable to express the motion prior in terms of these trajectories. One can arrive at the same conclusion by taking a purely statistical point of view: the above-mentioned experiments showed less variation in hand positions compared to joint angles. Thus, it is more robust to build models in terms of spatial hand positions.

1.4 The Gold Standard

As we have seen that there are dangers associated with learning motion specific priors, we take a step backwards to analyse activity independent models. Such models appear both in activity independent trackers and as regularisers for learning schemes. However, not much work has gone into designing good models for activity independent motion priors.

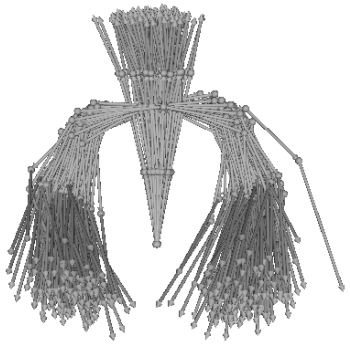


Figure 1.5 *Human poses sampled from eq. 1.1 where the covariance is learned from motion data. Notice how the variance accumulates along the spine.*

The most popular model is a Brownian motion model[1] in joint angle space, i.e.

$$p(\theta_t|\theta_{t-1}) = \mathcal{N}(\theta_t|\theta_{t-1}, \Sigma) \ , \tag{1.1}$$

where Σ denotes the (often diagonal) covariance. This model has, amongst many, been used by itself in [2, 4, 30, 45] and in [33, 45, 47–50, 52] where the prior is used as part of learning schemes.

At a first glance, eq. 1.1 seems like the obvious model choice: it is simple, easy to work with and easy to understand. However, it has several undesirable properties, which both lead to poor performance and make the model hard to extend.

1.4.1 Connections to the Human Plan

As discussed in sec. 1.3, there is evidence that humans plan their motion in the spatial domain. It is our belief that the best possible prediction of human motion can be crafted by "guessing" at the next step in the human plan. In this line of thinking, we should be predicting in the spatial domain, i.e. model joint positions rather than joint angles.

1.4.2 Accumulated Variance

The model in eq. 1.1 provides a covariance in *joint angle* space. In the end, we are, however, more interested in the covariance of *joint positions* as this is what is observed in the data. So, how do joint positions vary when we change the joint angles?

This question immediately leads to two observations:

1. **Long bones move more.** When a joint angle is changed, it affects the bone connected to the joint. If the connected bone is short, then the change in joint position will be small, but if the bone is long, the positional change will be large. In principle, we can compensate for this by scaling the covariance appropriately, but this would conflict with learning a covariance from motion data.

2. **Variance accumulates.** A change in a joint angle in the beginning of a kinematic chain affects all joint positions further down the chain. As an example, consider changing an angle in the shoulder joint: this moves the position of all joints in the arm. This means that joint positions further down the kinematic chains will exhibit larger variance than those in the beginning of the chains. As an example, the hand position will have a larger variance than the shoulder position as it depends on more joint angles.

Neither of these phenomena seem particularly natural, and both are consequences of the fact that eq. 1.1 ignores the skeleton configuration, i.e. bone sizes and how they are connected to each other. Their effect is quite evident in fig. 1.5, where samples from eq. 1.1 are shown; notice how the variance increases along the spine.

While these phenomena are not particularly natural, one might ask whether they matter? Both phenomena lead to increases in the variance of joint positions. In practice, the main purpose of the motion prior is to act as a temporal low-pass filter to suppress noise in the likelihood function. As the noise in the likelihood stems from noise in the observed image data it is essential that the motion prior has good low-pass properties in the image domain, i.e. the spatial domain. The large joint position variance of eq. 1.1 is simply another way of saying that this motion prior performs poorly as a low-pass filter in the spatial domain. As this is the data domain, the large spatial variance essentially means that eq. 1.1 fails at its most fundamental task!

1.4.3 Unnatural Metric

The motion prior in eq. 1.1 is a Brownian motion model in joint angle space, which is well-established as *the* least-committed motion prior. This model is sensitive to the underlying distance measure, so one way to gain insights into the unnatural variance behaviour of the model is to study the metric in joint angle space. If we assume an isotropic covariance, then the joint angle metric can be written as

$$\begin{aligned} \text{dist}^2(\theta_a, \theta_b) &= \|\theta_a - \theta_b\|^2 \\ &= (\theta_a - \theta_b)^T(\theta_a - \theta_b) \ , \end{aligned} \tag{1.2}$$

i.e. a measure of change in angles. This measure has little physical intuition, which is illustrated in fig. 1.6. Here, three motions of size 45 degrees are shown; each motion has been constructed by changing one joint angle. While the motions numerically are equally "big", they appear substantially different. This observation is quite troublesome, as practically all statistical models (esp. eq. 1.1) are highly dependent on the metric. The

[1] Technically, it is an instance of *Itô diffusion* due to the non-isotropic covariance, but we shall informally call it a Brownian motion.

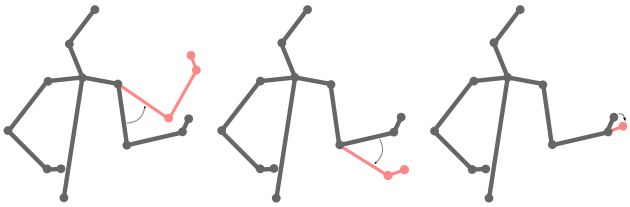


Figure 1.6 *Three motions of "equal size" under the isotropic joint angle metric.*

unnatural variance behaviour and the unnatural distance measure are both due to the fact that the joint angle model ignores the skeleton structure. This gives us the hint that we can improve the variance behaviour by improving the metric.

1.4.4 Model of Intrinsic Parameters

The basic idea behind eq. 1.1 is to assume a simple model in the parameter space of the kinematic skeleton. As we have seen, this choice leads to an increase in spatial variance due to an unnatural metric. While one might be tempted to accept this due to the benefits of having such a simple model, there are other downsides to the model. As the relationship between the parameter space (joint angle space) and the spatial domain is inherently non-linear it quickly becomes non-trivial to expand upon the model in eq. 1.1.

As human motion only rarely happens in empty spaces, the first extension we shall consider is to model interaction with the environment. Most often humans are interacting with their surroundings, they are picking up objects, moving them, putting them down and so forth. It, thus, seems reasonable to include knowledge of object interaction in the models. As the surroundings are inherently spatial, it becomes non-trivial to model the interaction in the joint angle domain due to the non-linear relationship between angles and positions. On the other hand, it turns out to be (almost embarrassingly) simple to perform such modelling when we model joint positions rather than angles.

On the more technical side, *particle filters*, which are used for estimating the pose, can be drastically improved by drawing samples from the *optimal importance distribution* [9, 22]. This distribution is on the form $p(\theta_t|\theta_{t-1}, \mathbf{Z}_t)$, where $\mathbf{Z}_t$ denotes the current observation. In other words: we can improve the particle filter if we are able to draw samples from a distribution that depends on both the previous pose as well as the current observation. Again, we observe that such distributions can be simple to design in the spatial domain, as this can easily be linked to the observation.

1.5 Modelling Summary

So far, we have seen that the commonly used Brownian motion model in joint angle space has several downsides. First, the choice of metric leads to large variances in joint position space, which makes the model perform poorly as a low-pass filter. Second, the choice of modelling joint angles makes it hard to extend the model. Third, the model seems to have little in common with how neurologists believe humans plan their motion. All these issues can swiftly be solved if we are able to create motion priors that are expressed directly in the spatial joint position domain. This will be the main focus of the included papers.

1.5.1 Thesis Organisation

The rest of this thesis is a set of papers followed by some concluding remarks in the last chapter. Each paper is introduced with a short statement that highlights the connections between the different papers and how the specific paper fits into the larger picture.

In the first paper, we formalise the notion of *probabilistic inverse kinematics*. In an attempt to build spatial motion priors, we introduce spatial goal positions of a few selected joints. We then derive the natural distribution of joint angles that reach this goal as closely as possible. Furthermore, we derive a sampling scheme for this distribution that allows us to use the model as a motion prior in a particle filter.

In the first paper, we depend on the joint angle metric for regularisation, which is somewhat dissatisfying. The second paper addresses this issue by modelling all joint positions instead of a few selected ones. This strategy turns out to put our models on a Riemannian manifold, which we call *the kinematic manifold*. From this view-point inverse kinematics becomes a projection operator onto the manifold. This allows us to define a new statistical model, called the *projected prior*, where a Gaussian is projected onto the manifold. As before, we show how this can be simulated and thereby used in particle filters.

The third paper is a reaction to another paper [30], published around the same time as the second paper. This paper is concerned with interaction with the environment. The third paper argues that such models should be expressed in the spatial domain rather than in terms of joint angles. This leads to improvements in both speed and accuracy, while the model is conceptually remarkably simple.

The fourth paper tries to improve the particle filter by changing the importance distribution such that samples are drawn closer to the modes of the likelihood. The main point of the paper is that such work is vastly simplified when working directly in the spatial domain. In practice, we present two similar approaches that improve accuracy quite a bit, while requiring little extra computational effort.

The second, third, and fourth papers are based on a somewhat *ad hoc* statistical model where a Gaussian is projected onto a Riemannian manifold. The fifth paper introduces a Brownian motion model expressed directly on the manifold along with a numerical scheme for working with this model in practice. This allows us to use more well-known statistical models even if we are working in a non-Euclidean domain.

All of the previous papers have been based on particle filters. When working on non-trivial Riemannian manifolds, such *Monte Carlo*-style methods have been the only available tool for filtering problems. This can be seen as an argument against using Riemannian models, as having more simple tools available can improve algorithmic development. In order to counter this argument, we show in the sixth paper how to generalise the unscented Kalman filter to Riemannian manifolds.

Finally, the thesis is concluded with a brief summary of the presented results and a discussion of future work.

Paper 1: Predicting Articulated Human Motion from Spatial Processes

Authors: Søren Hauberg and Kim S. Pedersen.
Status: Published at the *International Journal of Computer Vision* [21].

The model presented in the first paper is inspired by the theory stated by Morasso [35] and Abend et al. [1], saying that humans plan many actions based on spatial goals. More specifically, it was found that humans plan their motion by determining where in space they want to position selected body parts (e.g. the hands). This will be the starting point of the first model: we will assume that we know a goal and will then derive the distribution of joint angles that reach for this goal. The resulting model is called *Probabilistic Inverse Kinematics*.

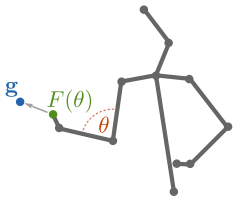


Figure 2.1 *An illustration of the goal-based model: a few selected limbs are assigned a goal position. On the figure only one hand is assigned a goal.*

The paper contributes with, to the best of our knowledge, the first Bayesian model of temporal inverse kinematics. In order to apply the model in practice, we derive a second order approximation to the Bootstrap filter. Early versions of this paper were presented in [12, 23], but the included paper supersedes these papers completely.

Accepted for International Journal of Computer Vision
Final copy available at http://www.springerlink.com/content/6710v0w2252m6520/

Predicting Articulated Human Motion from Spatial Processes

Søren Hauberg · Kim Steenstrup Pedersen

the date of receipt and acceptance should be inserted later

Abstract We present a probabilistic interpretation of inverse kinematics and extend it to sequential data. The resulting model is used to estimate articulated human motion in visual data. The approach allows us to express the prior temporal models in spatial limb coordinates, which is in contrast to most recent work where prior models are derived in terms of joint angles. This approach has several advantages. First of all, it allows us to construct motion models in low dimensional spaces, which makes motion estimation more robust. Secondly, as many types of motion are easily expressed in spatial coordinates, the approach allows us to construct high quality application specific motion models with little effort. Thirdly, the state space is a real vector space, which allows us to use *off-the-shelf* stochastic processes as motion models, which is rarely possible when working with joint angles. Fourthly, we avoid the problem of accumulated variance, where noise in one joint affects all joints further down the kinematic chains. All this combined allows us to more easily construct high quality motion models. In the evaluation, we show that an activity independent version of our model is superior to the corresponding state-of-the-art model. We also give examples of activity dependent models that would be hard to phrase directly in terms of joint angles.

Keywords Motion Analysis · Articulated Human Motion · Articulated Tracking · Prediction · Inverse Kinematics · Particle Filtering

S. Hauberg
Dept. of Computer Science, University of Copenhagen
E-mail: hauberg@diku.dk

K.S. Pedersen
Dept. of Computer Science, University of Copenhagen
E-mail: kimstp@diku.dk

1 Introduction

Three dimensional articulated human motion analysis is the process of estimating the configuration of body parts over time from sensor input (Poppe, 2007). One approach to this estimation is to use motion capture equipment where e.g. electromagnetic markers are attached to the body and then tracked in three dimensions. While this approach gives accurate results, it is intrusive and cannot be used outside laboratory settings. Alternatively, computer vision systems can be used for non-intrusive analysis. These systems usually perform some sort of optimisation for finding the best configuration of body parts. Such optimisation is often guided by a system for predicting future motion. This paper concerns a framework for building such predictive systems. Unlike most previous work, we build the actual predictive models in spatial coordinates, e.g. by studying hand trajectories, instead of working directly in the space of configuration parameters. This approach not only simplifies certain mathematical aspects of the modelling, but also provides a framework that is more in tune with how humans plan, think about and discuss motion.

Our approach is inspired by results from neurology (Morasso, 1981; Abend et al, 1982) that indicates that humans plan their motions in spatial coordinates. Our working hypothesis is that the best possible predictive system is one that mimics the motion plan. That is, we claim that predictions of future motion should be phrased in the same terms as the motion plan, i.e. in spatial coordinates. This is in contrast to most ongoing research in the vision community where predictions are performed in terms of the pose representation, e.g. the joint configuration of the kinematic skeleton.

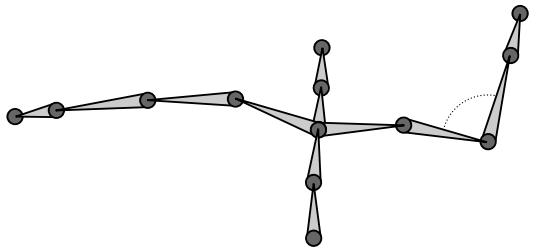


Fig. 1 A rendering of the kinematic skeleton. Each bone is computed as a rotation and a translation relative to its parent. Any subset of the circles, are collectively referred to as the *end-effectors*.

For long, researchers in computer animation have arrived at similar conclusions (Kerlow, 2003; Erleben et al, 2005). It is quite difficult to pose a figure in terms of its internal representation. For most work, animators instead pose individual bones of the figure in spatial coordinates using an *inverse kinematics* system. Such a system usually seeks a configuration of the joints that minimises the distance between a goal and an attained spatial coordinate by solving a nonlinear least-squares problem. In this paper, we recast the least-squares optimisation problem in a probabilistic setting. This then allows us to extend the model to sequential data, which in turn allows us to work with motion models in spatial coordinates rather than joint angles.

2 The Pose Representation

Before discussing the issues of human motion analysis, we pause to introduce the actual representation of the human pose. In this paper, we use the *kinematic skeleton* (see fig. 1), which, amongst others, was also used by Sidenbladh et al (2000) and Sminchisescu and Triggs (2003). The 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 point of connection as a *joint*.

Since bone lengths tend to change very slowly in humans (e.g. at the time scale of biological growth), these are modelled as being constant and effectively we consider bones as being rigid. Hence, the direction of each bone constitute the only degrees of freedom in the kinematic skeleton. The direction in each joint can be parametrised with a vector of angles, noticing that different joints may have different number of degrees of freedom. We may collect all joint angle vectors into one large vector $\boldsymbol{\theta}$ representing all joint angles in the model. Since each element in this vector is an angle, $\boldsymbol{\theta}$ must be confined to the N-dimensional torus, $\mathbb{T}^N$.

2.1 Forward Kinematics

From known bone lengths and a joint angle vector $\boldsymbol{\theta}$ it is straight-forward to compute the spatial coordinates of the bones. Specifically, the purpose is to compute the spatial coordinates of the end points of each bone. This process is started at the root of the tree structure and moves recursively along the branches, which are known as the *kinematic chains*.

The root of the tree is placed at the origin of the coordinate system. The end point of the next bone along a kinematic chain is then computed by rotating the coordinate system and then translating the root along a fixed axis, 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 l^{th} end point, 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 $\boldsymbol{\theta}$ and the length of the translation corresponds to the known length of the bone. We can repeat this process recursively until the entire kinematic tree has been traversed. This process is known as *Forward Kinematics* (Erleben et al, 2005).

The rotation matrix $\mathbf{R}_l$ of the l^{th} bone is parametrised by parts of $\boldsymbol{\theta}$. The actual number of used parameters depends on the specific joint. For elbow, knee and angle joints, we use one parameter, while we use three parameters to control all other joints. These two different joint types are respectively known as *hinge joints* and *ball joints*.

Using forward kinematics, we can compute the spatial coordinates of the end points of the individual bones. These are collectively referred to as *end-effectors*. In fig. 1 these are drawn as circles. In most situations, we shall only be concerned with some subset of the end-effectors. Often one is only concerned with body extremities, such as the head and the hands, hence the name *end*-effectors. We will denote the spatial coordinates of these selected end-effectors by $F(\boldsymbol{\theta})$.

2.2 Joint Constraints

In the human body, bones cannot move freely at the joints. A simple example is the elbow joint, which can approximately only bend between 0 and 160 degrees. To represent this, $\boldsymbol{\theta}$ is confined to a subset $\boldsymbol{\Theta}$ of $\mathbb{T}^N$. For simplicity, this subset is often defined by confining each component of $\boldsymbol{\theta}$ to an interval, i.e.

$$\boldsymbol{\Theta} = \prod_{n=1}^{N} [l_n, u_n] \quad , \tag{2}$$

where l_n and u_n denote the lower and upper bounds of the n^{th} component. This type of constraints on the angles is often called *box constraints* (Erleben et al, 2005). More realistic joint constraints are also possible, e.g. the implicit surface models of Herda et al (2004).

3 Challenges of Motion Analysis

Much work has gone into human motion analysis. The bulk of the work is in *non-articulated* analysis, i.e. locating the position of moving humans in image sequences and classifying their actions. It is, however, beyond the scope of this paper to give a review of this work. The interested reader can consult review papers such as the one by Moeslund et al (2006).

In recent years, focus has shifted to *articulated* visual human motion analysis in three dimensions (Poppe, 2007). Here, the objective is to estimate $\boldsymbol{\theta}$ in each image in a sequence. When only using a single camera, or a narrow baseline stereo camera, motion analysis is inherently difficult due to self-occlusions and visual ambiguities. This manifests itself in that the distribution of the human pose is multi-modal with an unknown number of modes. To reliably estimate this distribution we need methods that cope well with multi-modal distributions. Currently, the best method for such problems is the particle filter (Cappé et al, 2007), which represents the distribution as a set of weighted samples. Unfortunately, the particle filter is smitten by the curse of dimensionality in that the necessary number of samples grow exponentially with the dimensionality of state space. The consequence is that the particle filter is only applicable to low dimensional state spaces. This is in direct conflict with the fact that the human body has a great number of degrees of freedom.

The most obvious solution to these problems is to introduce some activity dependent model of the motion, such that the effective degrees of freedom is lowered. Here it should be noted that the actual number of degrees of freedom in the human body (independently of representation) is inherently large. So, while this approach works it does force us into building models that does not generalise to other types of motion than the ones modelled.

3.1 Dimensionality Reduction in Motion Analysis

Motion specific models can be constructed by reducing the dimensionality of the angle space by learning a manifold in angle space to which the motion is restricted. 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. In the paper a model of *walking* is learned, which is a periodic motion, and will therefore be performed in a nonlinear cyclic subspace of the angle space. The choice of a linear subspace therefore seems to come from sheer practicality in order to cope with the high dimensionality of the angle space and not from a well-founded modelling perspective.

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. The methods used for learning the manifolds, however, assumes that data is densely sampled on the manifold. This either requires vast amounts of data or a low-dimensional manifold. In the mentioned papers, low dimensional manifolds are studied. Specifically, one-dimensional manifolds corresponding to *walking*. It remains to be seen if the approach scales to higher dimensional manifolds.

Instead of learning the manifold, Elgammal and Lee (2009) suggested learning a mapping from angle space to a *known* manifold. They choose to learn a mapping onto a two dimensional torus, which allows for analysis of both periodic and aperiodic motion. By enforcing a known topology, the learning problem becomes more tractable compared to unsupervised methods. The approach is, however, only applicable when a known topology is available.

Instead of just learning a manifold and restricting the tracking to this, it seems reasonable also to use the density of the training data on this 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*. This approach seems to provide smooth priors that are both suitable for animation and tracking.

This approach, however, gracefully ignores the topology of the angle space. Specifically, the approach treats the angle space as Euclidean, and thus ignores both the periodic nature of angles and the constraints on the joints. To deal with this issue, Urtasun et al (2008) suggested changing the inner product of the before-

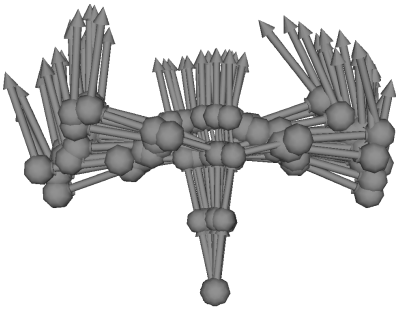


Fig. 2 One hundred random poses generated by sampling joint angles independently from a Von Mises distribution with concentration $\kappa = 500$, corresponding to a circular variance of approximately 0.001 (for details of this sampling, see sec. 7.1.1). Notice how the variance of spatial limb positions increase as the kinematic chains are traversed.

mentioned Gaussian process to incorporate joint constraints.

3.2 The Case Against Predictions in Angle Space

When representing a pose as a set of joint angles $\boldsymbol{\theta}$, it is tempting to build motion models $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ in terms of joint angles. This approach is, however, not without problems.

The first problem is simply that the space of angles is quite high dimensional. This does not rule out manual construction of models, but it can make model learning impractical.

The second problem is the relationship between spatial limb positions and limb orientations implied by forward kinematics. The spatial position of a limb is dependent on the position of its parent. Thus, if we change the direction of one bone, we change the position of all its children. Now, consider a predictive model in angle space that simply adds a little uncertainty to each angle. For this simple model, we see that the variance of the spatial position of limbs increases as we traverse the kinematic chains (see fig. 2). In other words: the position of a hand is always more uncertain than the position of the elbow. This property does not seem to come from a well-founded modelling perspective.

The third problem is that the angle space is topologically different from $\mathbb{R}^N$. If the joint angles are unconstrained, such that each angle can take on values in the circular domain $[0, 2\pi)$, they live on the N-dimensional torus. Thus, it is necessary to restrict the motion models to this manifold, which rules out models that are designed for $\mathbb{R}^N$.

If the joint angles are instead constrained, the topology of the angle space changes significantly. If e.g. box constraints are enforced the set of legal angles becomes a box in $\mathbb{R}^N$. Specifically, it becomes the product space $\prod_{n=1}^{N}[l_n, u_n]$, where l_n and u_n are the lower and upper constraints for the n^{th} angle. Again, we cannot apply motion models designed for $\mathbb{R}^N$, without taking special care to ensure that the pose adheres to the constraints.

In either case, we cannot use motion models designed for $\mathbb{R}^N$. This means that as long as we model in terms of joint angles, it is not mathematically well-defined to learn motion models using e.g. PCA. This problem can be alleviated by picking a suitable inner product for the angle space as suggested by Urtasun et al (2008). While a carefully designed inner product can solve the mathematical problems, it does not solve the rest of the above-mentioned problems.

From a modelling point of view, these problems all lead to the same fundamental question: *which space is most suitable for predicting human motion?*

3.3 Experimental Motivation

For many years, experimental neurologists have studied how people move. Morasso (1981) measured joint angles and hand positions while people moved their hand to a given target. Abend et al (1982) expanded on this experiment, and introduced obstacles that the hand had to move around. Both made the same observation: joint angles show great variation between both persons and trials, while hand positions consistently followed the same path with low variation. Recently, Ganesh (2009) made similar observations for actions like *punching* and *grasping* in a computer vision based tele-immersion system. One interpretation of these results is that the end-effector trajectories describes the *generic*, i.e. the person independent, part of the motion.

This result is a strong indication that humans plan body motion in terms of spatial limb trajectories rather than joint angles. It is our claim that we can achieve better predictions of human motion if we do so in the same space as where the motion was planned. Hence, it makes sense to build predictive models in end-effector space rather than the space of joint angles.

One can arrive at the same conclusion by taking a purely statistical point of view. The above-mentioned experiments showed less variation in end-effector positions compared to joint angles. Thus, it seems more robust to build or learn models in terms of end-effectors as this describes the generic part of the motion.

3.4 Our Approach

Inspired by the results from neurology, we turn to modelling human motion in terms of a few selected end-effectors. Which end-effectors we choose to model de-

pends on the studied motion. Our first assumption is that we know some stochastic process that describes the motion of these end-effectors. Our goal is then to provide a framework for moving this process back into the angle space. From a practical point of view, we aim at describing the pose distribution in angle space given the end-effector positions. This problem is closely related to inverse kinematics, which seeks a joint configuration that attains the given end-effector positions.

This approach has several advantages.

1. Since we are only modelling few selected end-effectors their spatial coordinates are more low-dimensional than all angles. While the degrees of freedom in the human body remains unchanged, the modelling space becomes more low-dimensional. This makes manual model crafting more practical, but more importantly it also makes model *learning* much more robust as fewer parameters need to be estimated.
2. Many types of motion can be easily described in spatial coordinates, e.g. *move foot towards ball* is a description of kicking a ball, whereas the same motion is hard to describe in terms of joint angles. These types of motions are typically *goal oriented*. In computer animation, this line of thinking is so common, that inverse kinematics systems are integrated with practically all 3D graphics software packages.
3. The stochastic motion process is expressed in spatial coordinates, which is a real vector space, instead of angles, which is on the N-dimensional torus. This makes it easier to use *off-the-shelf* stochastic processes as the motion model, since most processes are designed for real vector spaces.

3.5 Inverse Kinematics in Tracking

The most basic fact of computer vision is that the world is inherently spatial and that we study projections of this spatial world onto the image plane. Articulated motion can be estimated as a direct optimisation in the image plane, which requires the derivative of the likelihood with respect to the pose parameters. As the image data is inherently spatial this gradient depends on a mapping from the spatial domain into the joint angle space, i.e. an inverse kinematics system. Bregler et al (2004) formalises this in terms of *twists* and *products of exponential maps* as is common in the robotics literature (Murray et al, 1994). This leads to an iterative scheme for optimising a likelihood based on the grey value constancy assumption. This assumption works well on some sequences, but fails to cope with changes in the lighting conditions. Knossow et al (2008) avoids this issue by defining the likelihood in terms of the chamfer-distance between modelled contours and observed image contours. To perform direct optimisation Knossow et al finds the derivative of the projection of the contour generator into the image plane with respect to the pose parameters – again, this requires solving an inverse kinematics problem. These direct methods works well in many scenarios, but does not allow for an immediate inclusion of a prior motion model.

In this paper, we focus on statistical models of human motion expressed in the spatial domain. This idea of modelling human motion spatially is not new. Recently, Salzmann and Urtasun (2010) showed how joint positions can be modelled, while still adhering to the constraints given by the constant limb lengths. In a similar spirit, Hauberg et al (2010) has proposed that the constraint problem can be solved by projection onto the nonlinear manifold implicitly defined by enforcing constant limb lengths. These ideas has also been explored successfully in the motion compression literature, where Tournier et al (2009) showed how spatial limb positions could be compressed using principal geodesic analysis (Fletcher et al, 2004). These approaches are different from ours as we model goal positions of a few selected end-effectors rather than considering all joints in the kinematic skeleton.

When motion is estimated from silhouettes seen from a single view-point inherent visual ambiguities creates many local minima (Sminchisescu and Triggs, 2003). Using a simple closed-form inverse kinematics system allows Sminchisescu and Triggs to enumerate possible interpretations of the input. This results in a more efficient sampling scheme for their particle filter, as it simultaneously explores many local minima, which reduces the chance of getting stuck in a single local minimum. However, their approach suffers from the problems discussed in sec. 3.2, which makes it difficult to incorporate application specific motion priors.

When modelling interactions with the environment, inverse kinematics is an essential tool as it provides a mapping from the spatial world coordinate system to the joint angle space. Rosenhahn et al (2008) uses this to model interaction with sports equipment, such as bicycles and snowboards. They use the previously mentioned formalism of twists and products of exponential maps to derive an approximate gradient descent scheme for estimating poses. In a similar spirit, Kjellström et al (2010) models interaction with a stick-like object. They solve the inverse kinematics problem using rejection sampling in joint angle space, leading to a computational expensive approach due to the high dimensionality of the joint angle space.

Previously, we (Hauberg et al, 2009; Engell-Nørregård et al, 2009) successfully used end-effector positions

as the pose representation, which provided a substantial dimensionality reduction. When comparing a pose to an image, we used inverse kinematics for computing the entire pose configuration. This paper provides the full Bayesian development, analysis and interpretation of this approach.

Courty and Arnaud (2008) have previously suggested a probabilistic interpretation of the inverse kinematics problem. We share the idea of providing a probabilistic model, but where they solve the inverse kinematics problem using importance sampling, we use inverse kinematics to define the importance distribution in our particle filter.

3.6 Organisation of the Paper

The rest of the paper is organised as follows. In the next section, we derive a model, that allows us to describe the distribution of the joint angles in the kinematic skeleton given a set of end-effector goals. In sec. 5 we show the needed steps for performing inference in this model as part of an articulated motion analysis system. These two sections constitute the main technical contribution of the paper. To actually implement a system for articulated motion analysis, we need to deal with observational data. A simple system for this is described in sec. 6 and in sec. 7 we show examples of the attained results. The paper is concluded with a discussion in sec. 8.

4 Deriving the Model

The main objective of this paper is to construct a framework for making predictions of future motion. Since we are using the kinematic skeleton as the pose representation, we need to model some distribution of $\boldsymbol{\theta}$. As previously described we wish to build this distribution in terms of the spatial coordinates of selected end-effectors. To formalise this idea, we let $\mathbf{g}$ denote the spatial coordinates of the *end-effector goals*. Intuitively, this can be thought of as the position where the human wishes to place e.g. his or her hands. Our objective thus becomes to construct the distribution of $\boldsymbol{\theta}$ given the end-effector goal $\mathbf{g}$. Formally, we need to specify $p(\boldsymbol{\theta}|\mathbf{g})$.

Given joint angles, we start out by defining the likelihood of an end-effector goal $\mathbf{g}$ as a "noisy" extension of forward kinematics. Specifically, we define

$$p(\mathbf{g}|\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{g}|F(\boldsymbol{\theta}), \mathbf{W}^{-1}\right) \ , \tag{3}$$

where $\mathbf{W}$ is the precision (inverse covariance) matrix of the distribution. Here, the stochasticity represents that one does not always reach ones goals.

We do not have any good prior information about the joint angles $\boldsymbol{\theta}$ except some limits on their values. So, we take a least-commitment approach and model all legal angles as equally probable,

$$p(\boldsymbol{\theta}) = \mathcal{U}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \ , \tag{4}$$

where $\boldsymbol{\Theta}$ is the set of legal angles, and $\mathcal{U}_{\boldsymbol{\Theta}}$ is the uniform distribution on this set. In practice, we use box constraints, i.e. confine each component of $\boldsymbol{\theta}$ to an interval. This gives us

$$p(\boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{U}_{[l_n,u_n]}(\boldsymbol{\theta}[n]) \ , \tag{5}$$

where l_n and u_n denote the lower and upper bounds of the n^{th} component $\boldsymbol{\theta}[n]$.

Using Bayes' theorem, we combine eq. 3 and eq. 5 into

$$p(\boldsymbol{\theta}|\mathbf{g}) = \frac{p(\mathbf{g}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{g})} \propto p(\mathbf{g}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{6}$$

$$= \mathcal{N}\left(\mathbf{g}|F(\boldsymbol{\theta}), \mathbf{W}^{-1}\right) \prod_{n=1}^{N} \mathcal{U}_{[l_n,u_n]}(\boldsymbol{\theta}[n]) \ . \tag{7}$$

As can be seen, we perform a nonlinear transformation F of the $\boldsymbol{\theta}$ variable and define its distribution in the resulting space. In other words we effectively define the distribution of $\boldsymbol{\theta}$ in the end-effector goal space rather than angle space. It should be stressed that while $p(\boldsymbol{\theta}|\mathbf{g})$ looks similar to a normal distribution, it is not, due to the nonlinear transformation F.

In the end, the goal is to be able to extract $\boldsymbol{\theta}$ from observational data such as images. We do this using a straight-forward generative model,

$$p(\mathbf{X}, \boldsymbol{\theta}, \mathbf{g}) = p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{g})p(\mathbf{g}) \ , \tag{8}$$

where $\mathbf{X}$ is the observation. This model is shown graphically in fig. 3. It should be noted that we have yet to specify $p(\mathbf{g})$; we will remedy this at a later stage in the paper.

4.1 Relation to Inverse Kinematics

Taking the logarithm of eq. 7 gives us

$$\begin{aligned} \log p(\boldsymbol{\theta}|\mathbf{g}) = &-\frac{1}{2}(\mathbf{g} - F(\boldsymbol{\theta}))^T \mathbf{W}(\mathbf{g} - F(\boldsymbol{\theta})) \\ &+ \sum_{n=1}^{N} \log \mathcal{U}_{[l_n,u_n]}(\boldsymbol{\theta}[n]) + \text{constant} \ . \end{aligned} \tag{9}$$

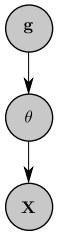


Fig. 3 Graphical representation of the *Probabilistic Inverse Kinematics* model given in eq. 8.

Maximising this corresponds to minimising

$$d^2\left(F(\boldsymbol{\theta}), \mathbf{g}\right) = \frac{1}{2}(\mathbf{g} - F(\boldsymbol{\theta}))^T \mathbf{W}(\mathbf{g} - F(\boldsymbol{\theta})) \ , \tag{10}$$

subject to $\mathbf{l} \leq \boldsymbol{\theta} \leq \mathbf{u}$, where $\mathbf{l}$ and $\mathbf{u}$ are the vectors containing the joint limits. This is the inverse kinematics model presented by Zhao and Badler (1994). Thus, we deem eq. 7 *Probabilistic Inverse Kinematics (PIK)*.

It should be noted that due to the nonlinearity of F, this optimisation problem is nonlinear, rendering maximum a posteriori estimation difficult. Since the end-effector space is often much more low-dimensional than the angle space, the Hessian of eq. 10 does not have full rank. As a consequence the optimisation problem is not guaranteed to have a unique solution. In fact, the minima of eq. 10 often form a continuous subspace of $\boldsymbol{\Theta}$. This can be realised simply by fixing the position of a hand, and moving the rest of the body freely. Such a sequence of poses will be continuous while all poses attain the same end-effector position.

4.2 Sequential PIK

As previously mentioned we aim at building predictive distributions for sequential analysis based on the end-effector goals. To keep the model as general as possible, we assume knowledge of some stochastic process controlling the end-effector goals. That is, we assume we know $p(\mathbf{g}_t|\mathbf{g}_{1:t-1})$, where the subscript denotes time and $\mathbf{g}_{1:t-1} = \{\mathbf{g}_1, \ldots, \mathbf{g}_{t-1}\}$ is the past sequence. In sec. 7 we will show examples of such processes, but it should be noted that any continuous process designed for *real* vector spaces is applicable.

While we accept any continuous model $p(\mathbf{g}_t|\mathbf{g}_{1:t-1})$ in end-effector goal space, we do prefer smooth motion in angular space. We model this as preferring small temporal gradients $\left|\left|\frac{\partial \boldsymbol{\theta}}{\partial t}\right|\right|^2$. To avoid extending the state with the temporal gradient, we approximate it using finite differences,

$$\left|\left|\frac{\partial \boldsymbol{\theta}}{\partial t}\right|\right|^2 \approx ||\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}||^2 \ . \tag{11}$$

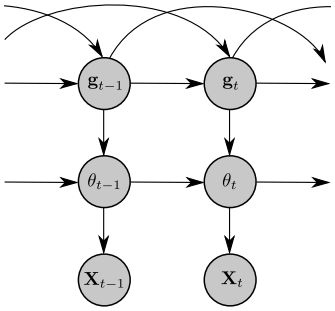


Fig. 4 Graphical representation of the *Sequential Probabilistic Inverse Kinematics* model given in eq. 13.

So, we introduce a first order Markov model in angle space and define

$$\begin{aligned}\log p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1}) = &-\frac{1}{2}(\mathbf{g}_t - F(\boldsymbol{\theta}_t))^T \mathbf{W}(\mathbf{g}_t - F(\boldsymbol{\theta}_t)) \\ &- \frac{\lambda}{2}||\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}||^2 \\ &+ \sum_{n=1}^{N} \log \mathcal{U}_{[l_n, u_n]}(\boldsymbol{\theta}_t[n]) + \text{constant} \ ,\end{aligned} \tag{12}$$

where λ controls the degree of temporal smoothness. This is effectively the same as eq. 9 except poses close to the previous one, $\boldsymbol{\theta}_{t-1}$, are preferred. This slight change has the pleasant effect of isolating the modes of $p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})$, which makes maximum a posteriori estimation more tractable. Consider the previously given example. A person can put his or her hands in a given position in many ways, but no continuous subset of the pose space can attain the given hand position and be closest to the previous pose at the same time.

In summary, we have the following final generative model

$$\begin{aligned}p(\mathbf{X}_{1:T}, \boldsymbol{\theta}_{1:T}, \mathbf{g}_{1:T}) = &\ p(\mathbf{X}_1|\boldsymbol{\theta}_1) p(\boldsymbol{\theta}_1|\mathbf{g}_1) p(\mathbf{g}_1) \\ &\prod_{t=2}^{T} p(\mathbf{X}_t|\boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t|\mathbf{g}_t, \theta_{t-1}) p(\mathbf{g}_t|\mathbf{g}_{1:t-1}) \ ,\end{aligned} \tag{13}$$

where $\mathbf{X}_{1:T} = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ denotes the sequence of observations. This model is illustrated in fig. 4.

4.3 Designing Spatial Processes

One of the advantages of the Sequential PIK model is that given a spatial model $p(\mathbf{g}_t|\mathbf{g}_{1:t-1})$ of the end-effector goals, we can work with the corresponding model in terms of joint angles. However, so far, we have avoided the question of how to define such spatial models. We will give specific examples in sec. 7, but until then, it can be instructive to consider how one might design such models.

Consider a person grasping for an object. Neurologists (Morasso, 1981; Abend et al, 1982) have shown that people most often move their hands on the straight line from the current hand position towards the object. This can easily be modelled by letting the goal position of the hand move along this line. The speed of the hand could be modelled as a constant or it could be controlled by another process. It should be stressed that the immediate goal $\mathbf{g}_t$ follows the mentioned line and hence is different from the end goal (the position of the object).

As a second example, consider a person walking. Many successful models of such motions have been derived in terms of joint angles as discussed in sec. 3.1. Most of them does, however, not consider interaction with the ground plane, which is an intrinsic part of the motion. This interaction actually makes the motion non-smooth, something most models cannot handle. By modelling the goals of the feet it is easy to ensure that one foot always is in contact with the ground plane and that the other never penetrates the ground. The actual trajectory of the moving foot could then be learned using, e.g. a Gaussian process (Rasmussen and Williams, 2006). Again, the immediate goal $\mathbf{g}_t$ would move along the curve represented by the learned process.

5 Inference

In the previous section a model of human motion was derived. This section focuses on the approximations needed to perform inference in the model. Due to the inherent multi-modality of the problem, we use a *particle filter* to perform the inference. In the next section this algorithm is described from an *importance sampling* point of view as our choice of *importance distribution* is non-conventional. Results will, however, not be proved; instead the reader is referred to the review paper by Cappé et al (2007).

We will assume that a continuous motion model $p(\mathbf{g}_t|\mathbf{g}_{1:t-1})$ is available and that we can sample from this distribution. Specific choices of this model will be made in sec. 7, but it should be stressed that the method is independent of this choice, e.g. further constraints such as smoothness may be added.

We will also assume that a system $p(\mathbf{X}_t|\boldsymbol{\theta}_t)$ for making visual measurements is available. Such a system will be described in sec. 6, but the method is also independent of the specifics of this system.

5.1 Approximate Bayesian Filtering

The aim of approximate Bayesian filtering is to estimate the filtering distribution $p(\boldsymbol{\theta}_t|\mathbf{g}_t, \mathbf{X}_{1:t})$ by means of samples. Instead of drawing samples from this distribution, they are taken from $p(\boldsymbol{\theta}_{1:t}|\mathbf{g}_{1:t}, \mathbf{X}_{1:t})$ and then all values of the samples but $\boldsymbol{\theta}_t$ are ignored. Since the filtering distribution is unknown, we turn to importance sampling (Bishop, 2006). This means drawing M samples $\boldsymbol{\theta}_{1:t}^{(m)}$ from an *importance distribution* $q(\boldsymbol{\theta}_{1:t}|\mathbf{g}_{1:t}, \mathbf{X}_{1:t})$ after which moments of the filtering distribution can be estimated as

$$\begin{aligned}\bar{h} &= \int h(\boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t|\mathbf{g}_t, \mathbf{X}_{1:t}) \mathrm{d}\boldsymbol{\theta}_t \\ &\approx \sum_{m=1}^{M} w_t^{(m)} h\left(\boldsymbol{\theta}_t^{(m)}\right)\end{aligned} \tag{14}$$

for any function h. Here we have defined the *importance weights* as

$$w_t^{(m)} \propto \frac{p\left(\boldsymbol{\theta}_{1:t}^{(m)}\middle|\mathbf{g}_{1:t}^{(m)}, \mathbf{X}_{1:t}\right)}{q\left(\boldsymbol{\theta}_{1:t}^{(m)}\middle|\mathbf{g}_{1:t}^{(m)}, \mathbf{X}_{1:t}\right)} \quad , \quad \sum_{m=1}^{M} w_t^{(m)} = 1 \ . \tag{15}$$

Unsurprisingly, eq. 14 is exact when $M \to \infty$.

The key to making this strategy work is to choose an importance distribution, which ensures that the resulting algorithm is recursive. With this in mind, it is chosen that the importance distribution should be factorised as

$$\begin{aligned}q(\boldsymbol{\theta}_{1:t}|\mathbf{g}_{1:t}, \mathbf{X}_{1:t}) = \ & q(\boldsymbol{\theta}_{1:t-1}|\mathbf{g}_{1:t-1}, \mathbf{X}_{1:t-1}) \\ & q(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1}, \mathbf{X}_t) \ .\end{aligned} \tag{16}$$

With this choice, one can sample from $q(\boldsymbol{\theta}_{1:t}|\mathbf{g}_{1:t}, \mathbf{X}_{1:t})$ recursively by extending the previous sample $\boldsymbol{\theta}_{1:t-1}^{(m)}$ with a new sample $\boldsymbol{\theta}_t^{(m)}$ from $q(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1}, \mathbf{X}_t)$. The weights $w_{t-1}^{(m)}$ can also be recursively updated by means of

$$w_t^{(m)} \propto w_{t-1}^{(m)} \ p\left(\mathbf{X}_t\middle|\boldsymbol{\theta}_t^{(m)}\right) r^{(m)} \tag{17}$$

with

$$r^{(m)} = \frac{p\left(\boldsymbol{\theta}_t^{(m)}\middle|\mathbf{g}_t^{(m)}, \boldsymbol{\theta}_{t-1}^{(m)}\right)}{q\left(\boldsymbol{\theta}_t^{(m)}\middle|\mathbf{g}_t^{(m)}, \boldsymbol{\theta}_{t-1}^{(m)}, \mathbf{X}_t\right)} \ . \tag{18}$$

When extending the previous sample, we need to draw a sample from $q(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1}, \mathbf{X}_t)$. This, however, assumes that the true value of $\boldsymbol{\theta}_{t-1}$ is known, which is not the case. Several strategies can be used to approximate this value. *Sequential Importance Sampling* assumes that the previous sample positions were the true value, i.e. $\boldsymbol{\theta}_{t-1} = \boldsymbol{\theta}_{t-1}^{(m)}$. This is usually not stable, since small differences between the sample and the

true value accumulates over time. The *particle filter* approximates the distribution of $\boldsymbol{\theta}_{t-1}$ with the weighted samples from the previous iteration, i.e.

$$p(\boldsymbol{\theta}_{t-1}|\mathbf{g}_{t-1}, \mathbf{X}_{1:t-1}) \approx \sum_{m=1}^{M} w_{t-1}^{(m)} \delta\left(\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_{t-1}^{(m)}\right) . \quad (19)$$

The value of $\boldsymbol{\theta}_{t-1}$ is then approximated by sampling from this distribution. This simply corresponds to a resampling of the previous samples, where samples with large weights have a high probability of surviving. Since these samples are assumed to come from the true distribution $p(\boldsymbol{\theta}_{t-1}|\mathbf{g}_{t-1}, \mathbf{X}_{1:t-1})$, the associated weights have to be reset, i.e. $w_{t-1}^{(m)} = 1/M$ for all m.

We have still to choose the importance distribution $q(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1}, \mathbf{X}_t)$. The most common choice is inspired by eq. 18. Here, we note that $r^{(m)} = 1$ if we set $q(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1}, \mathbf{X}_t) = p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})$, which simplifies the weight update. With this choice, the resulting filter is called the *Bootstrap filter*. This cannot, however, be applied for the model described in this paper, as we cannot sample directly from $p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})$. A different importance distribution will be presented next.

5.2 The Importance Distribution

Inspired by the Bootstrap filter, we drop the observation $\mathbf{X}_t$ from the importance distribution, such that $q(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1}, \mathbf{X}_t) = q(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})$. It would seem tempting to use $p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})$ as the importance distribution. It is, however, not straightforward to draw samples from this distribution, so this choice does not seem viable. Instead, we seek a distribution that locally behaves similarly to $p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})$.

In sec. 4.2 we noted that $p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})$ has isolated modes. Thus, it seems reasonable to locally approximate this distribution using a Laplace approximation (Bishop, 2006), which is a second order Taylor approximation of the true distribution. This boils down to fitting a normal distribution around the local mode $\boldsymbol{\theta}_t^*$, using the Hessian of $-\log p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})$ as an approximation of the precision matrix. Assuming that $F(\boldsymbol{\theta}_t^*) = \mathbf{g}_t$, i.e. the located pose actually reaches the given end-effector goal, this Hessian matrix attains the simple form

$$\mathbf{H} = -(\mathbf{g}_t - F(\boldsymbol{\theta}_t^*))^T \mathbf{W} \frac{\partial \mathbf{J}}{\partial \boldsymbol{\theta}} + \mathbf{J}^T \mathbf{W} \mathbf{J} + \lambda \mathbf{I} \quad (20)$$

$$= \mathbf{J}^T \mathbf{W} \mathbf{J} + \lambda \mathbf{I} \ , \quad (21)$$

where $\mathbf{J}$ is the Jacobian of F at $\boldsymbol{\theta}_t^*$. This Jacobian consists of a row for each component of $\boldsymbol{\theta}_t$. Each such row can be computed in a straightforward manner (Zhao and Badler, 1994). If $\mathbf{r}$ is the unit-length rotational axis of the n^{th} angle and $\Delta_{\mathbf{g}}$ is the vector from the joint to the end-effector, then the row is computed as $\frac{\partial F}{\partial \boldsymbol{\theta}_t[n]} = \mathbf{r} \times \Delta_{\mathbf{g}}$. This is merely the tangent of the circle formed by the end-effector when rotating the joint in question as is illustrated in fig. 5.

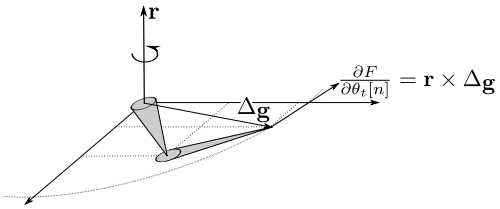


Fig. 5 The derivative of a joint is found as the cross product of the rotational axis $\mathbf{r}$ and the vector from the joint to the end-effector $\mathbf{g}$.

We, thus, have an easy way of computing the Laplace approximation of $p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})$, which we use as the importance distribution. That is, we pick

$$q(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1}, \mathbf{X}_t) = \mathcal{N}\left(\boldsymbol{\theta}_t \middle| \boldsymbol{\theta}_t^*, \left(\mathbf{J}^T \mathbf{W} \mathbf{J} + \lambda \mathbf{I}\right)^{-1}\right) \mathcal{U}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t) \ . \quad (22)$$

Hence, we are using a local second order approximation of the Bootstrap filter, while adhering to the joint constraints. It should be noted that we can easily sample from this distribution using *rejection sampling* (Bishop, 2006). When using box constraints, this rejection sampling can be performed one joint angle at a time and as such does not impact performance in any measureable way.

5.2.1 Solving the Nonlinear Least-Squares Problem

One assumption made in the previous section was that we can compute the mode $\boldsymbol{\theta}_t^*$ of $p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})$. Since the modes of this distribution are isolated, we can easily locate a mode, using a nonlinear constrained least-squares solver. In this paper, we are using a simple, yet effective, gradient projection method with line search (Nocedal and Wright, 1999).

To perform the optimisation, we need to compute the gradient of $\log p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})$. This is given by

$$\frac{\partial \log p(\boldsymbol{\theta}_t|\mathbf{g}_t, \boldsymbol{\theta}_{t-1})}{\partial \boldsymbol{\theta}_t} = (\mathbf{g}_t - F(\boldsymbol{\theta}_t))^T \mathbf{W} \mathbf{J} - \lambda(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) \ . \quad (23)$$

When solving the nonlinear least-squares problem, we start the search in $\boldsymbol{\theta}_{t-1}$, which usually ensures convergence in a few iterations.

6 Visual Measurements

In this section we define $p(\mathbf{X}_t|\boldsymbol{\theta}_t)$, i.e. we describe how we compare an observation with a pose hypothesis. This allows us to compute weights for the particle filter, which then optimises the posterior. Since this paper is focused on the prediction aspect of a tracker, we deliberately keep this part as simple as possible.

6.1 General Idea

To keep the necessary image processing to a minimum, we use a small baseline consumer stereo camera from Point Grey[1]. At each time-step, this camera provides a set of three dimensional points as seen from a single view point (see fig. 6). The objective of $p(\mathbf{X}_t|\boldsymbol{\theta}_t)$ is then essentially to measure how well $\boldsymbol{\theta}_t$ fits with these points. Due to the small baseline, visual ambiguities will occur, which leads to local maxima in the likelihood. This is one reason for using a particle filter for performing inference.

Let $\mathbf{X}_t = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ denote the set of three dimensional points provided by the stereo camera. Our first simplifying assumption is that these are independent and identically distributed, i.e.

$$p(\mathbf{X}_t|\boldsymbol{\theta}_t) = \prod_{k=1}^{K} p(\mathbf{x}_k|\boldsymbol{\theta}_t) \ . \tag{24}$$

We then define the likelihood of an individual point as

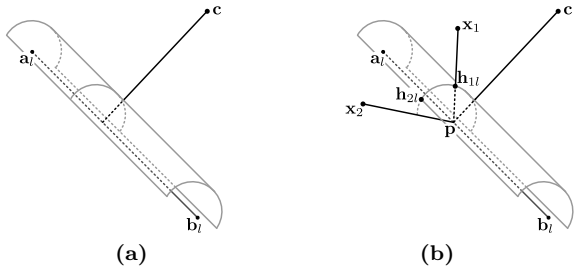
$$p(\mathbf{x}_k|\boldsymbol{\theta}_t) \propto \exp\left(-\frac{D^2(\boldsymbol{\theta}_t, \mathbf{x}_k)}{2\sigma^2}\right) \ , \tag{25}$$

where $D^2(\boldsymbol{\theta}_t, \mathbf{x}_k)$ denotes the squared distance between the point $\mathbf{x}_k$ and the surface of the pose parametrised by $\boldsymbol{\theta}_t$.

We, thus, need a definition of the surface of a pose, and a suitable metric.

6.2 The Pose Surface

The pose $\boldsymbol{\theta}_t$ corresponds to a connected set of L bones, each of which have a start and an end point. We, respectively, denote these $\mathbf{a}_l$ and $\mathbf{b}_l$ for the l^{th} bone; we can compute these points using forward kinematics. We then construct a capsule with radius r_l that follows the line segment from $\mathbf{a}_l$ to $\mathbf{b}_l$. The surface of the l^{th} bone is then defined as the part of this capsule that is visible from the current view point. This surface model of a bone is illustrated in fig. 7a. The entire pose surface is then defined as the union of these bone surfaces. This is essentially the same surface model as was suggested by Sidenbladh et al (2000), except they used cylinders instead of capsules. In general, this type of surface models does not describe the human body particularly well. The capsule skin can, however, be replaced with more descriptive skin models, such as the articulated implicit surfaces suggested by Horaud et al (2009).



(a) (b)

Fig. 7 (a) An illustration of the surface of a bone. Here $\mathbf{a}_l$ and $\mathbf{b}_l$ denotes the end points of the bone, while $\mathbf{c}$ denotes the camera position. For illustration purposes, we only show the cylindric part of the capsules. (b) An illustration of the computation of the point on the bone surface from a data point. To keep the figure simple, we show data points $\mathbf{x}_1$ and $\mathbf{x}_2$ that share the same nearest point $\mathbf{p} = \mathbf{p}_{1l} = \mathbf{p}_{2l}$ on the line segment between $\mathbf{a}_l$ and $\mathbf{b}_l$. The vectors $\mathbf{w}_{1l}$ and $\mathbf{w}_{2l}$ are the vectors from $\mathbf{p}$ pointing towards $\mathbf{h}_{1l}$ and $\mathbf{h}_{2l}$.

In the end, our objective is to compute the distance between a data point and the surface. We do this by first finding the nearest point on the surface and then compute the distance between this point and the data point. Since we define the pose surface bone by bone, we can compute this distance as the distance to the nearest bone surface, i.e.

$$D^2(\boldsymbol{\theta}_t, \mathbf{x}_k) = \min_l \left(d^2(\mathbf{x}_k, \mathbf{h}_{kl})\right) \ , \tag{26}$$

where $\mathbf{h}_{kl}$ is the nearest point (in the Euclidean sense) on the l^{th} bone and $d^2(\mathbf{x}_k, \mathbf{h}_{kl})$ is the squared distance between $\mathbf{x}_k$ and $\mathbf{h}_{kl}$. Note that the minimisation in eq. 26 can be trivially performed by iterating over all L bones.

6.2.1 Finding Nearest Point on the Bone Surface

We thus set out to find the point on a bone surface that is nearest to the data point $\mathbf{x}_k$. We start by finding the nearest point on the capsule with radius r_l around the line segment from $\mathbf{a}_l$ to $\mathbf{b}_l$. We let $\mathbf{p}_{kl}$ denote the point on the line segment that is nearest to $\mathbf{x}_k$, and then the nearest point on the capsule can be found as $\mathbf{p} + r_l \frac{\mathbf{x}_k - \mathbf{p}_{kl}}{||\mathbf{x}_k - \mathbf{p}_{kl}||}$.

We now turn our attention to the part of the capsule that can be seen from the camera. Points on this part of the capsule can be described as the points where the angle between the vectors from $\mathbf{p}_{kl}$ to the camera and

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

Fig. 6 A rendering of the data from the stereo camera from different views.

to the point on the surface should be no greater than 90 degrees. This is formalised by requiring $(\mathbf{x}_k - \mathbf{p}_{kl})^T(\mathbf{c} - \mathbf{p}_{kl}) \geq 0$, where $\mathbf{c}$ denotes the position of the camera.

If the nearest point is not on the visible part, then it is on the border of the surface. Hence, the vector from $\mathbf{p}_{kl}$ to the nearest point must be orthogonal to the line segment formed by $\mathbf{a}_l$ and $\mathbf{b}_l$, and the vector from $\mathbf{p}_{kl}$ to $\mathbf{c}$. In other words, the nearest point on the surface can then be computed as

$$\mathbf{h}_{kl} = \mathbf{p}_{kl} + r_l \frac{\mathbf{w}_{kl}}{||\mathbf{w}_{kl}||} \quad , \tag{27}$$

where

$$\mathbf{w}_{kl} = \begin{cases} \mathbf{x}_k - \mathbf{p}_{kl}, & (\mathbf{x}_k - \mathbf{p}_{kl})^T(\mathbf{c} - \mathbf{p}_{kl}) \geq 0 \\ \text{sgn}\left(\mathbf{x}_k^T \mathbf{v}\right) \mathbf{v}, & \text{otherwise} \end{cases} , \tag{28}$$

with $\mathbf{v} = (\mathbf{c} - \mathbf{p}_{kl}) \times (\mathbf{b}_l - \mathbf{a}_l)$. The geometry behind these computations is illustrated in fig. 7b.

6.3 Robust Metric

We are now able to find the point on the surface of a bone that is nearest to a data point $\mathbf{x}_k$. Thus, we are only missing a suitable way of computing the squared distance between the data point and the nearest point on the bone surface. The most straight-forward approach is to use the squared Euclidean distance. This is, however, not robust with respect to outliers. Looking at fig. 6, we see that the data from the stereo camera contains many outliers; some due to mismatches and some due to other objects in the scene (i.e. the background).

To cope with these outliers, we could use any robust metric. Here, we choose to use a simple thresholded squared Euclidean distance, i.e.

$$d^2(\mathbf{x}_k, \mathbf{h}_{kl}) = \min\left(||\mathbf{x}_k - \mathbf{h}_{kl}||^2, \tau\right) \quad . \tag{29}$$

7 Results

We now have a complete articulated tracker, and so move on to the evaluation. First, we compare our predictive model to a linear model in angle space. Then, we show how the model can be extended to model interactions between the person and objects in the environment. Finally, we provide an example of an activity dependent model from the world of physiotherapy.

In each frame, the particle filter provides us with a set of weighted hypotheses. To reduce this set to a single hypothesis in each frame, we compute the weighted average, i.e.

$$\hat{\boldsymbol{\theta}}_t = \sum_{m=1}^{M} w_t^{(m)} \boldsymbol{\theta}_t^{(m)} \quad , \tag{30}$$

which we use as an estimate of the current pose. This simple choice seems to work well enough in practice.

7.1 Linear Extrapolation in Different Spaces

We start out by studying an image sequence in which a test subject keeps his legs fixed while waving a stick with both hands in a fairly arbitrary way. This sequence allows us to work with both complex and highly articulated upper body motions, without having to model translations of the entire body. A few frames from this sequence is available in fig. 8. This sequence poses several problems. Due to the style of the motion, limbs are often occluded by each other, e.g. one arm is often occluded by the torso. As the sequence is recorded at approximately 15 frames per second we also see motion blur. This reduces the quality of the stereo reconstruction, which produces ambiguities for the tracker.

7.1.1 Angular Motion Model

For comparative purposes we build an articulated tracker in which the predictive model is phrased in terms of

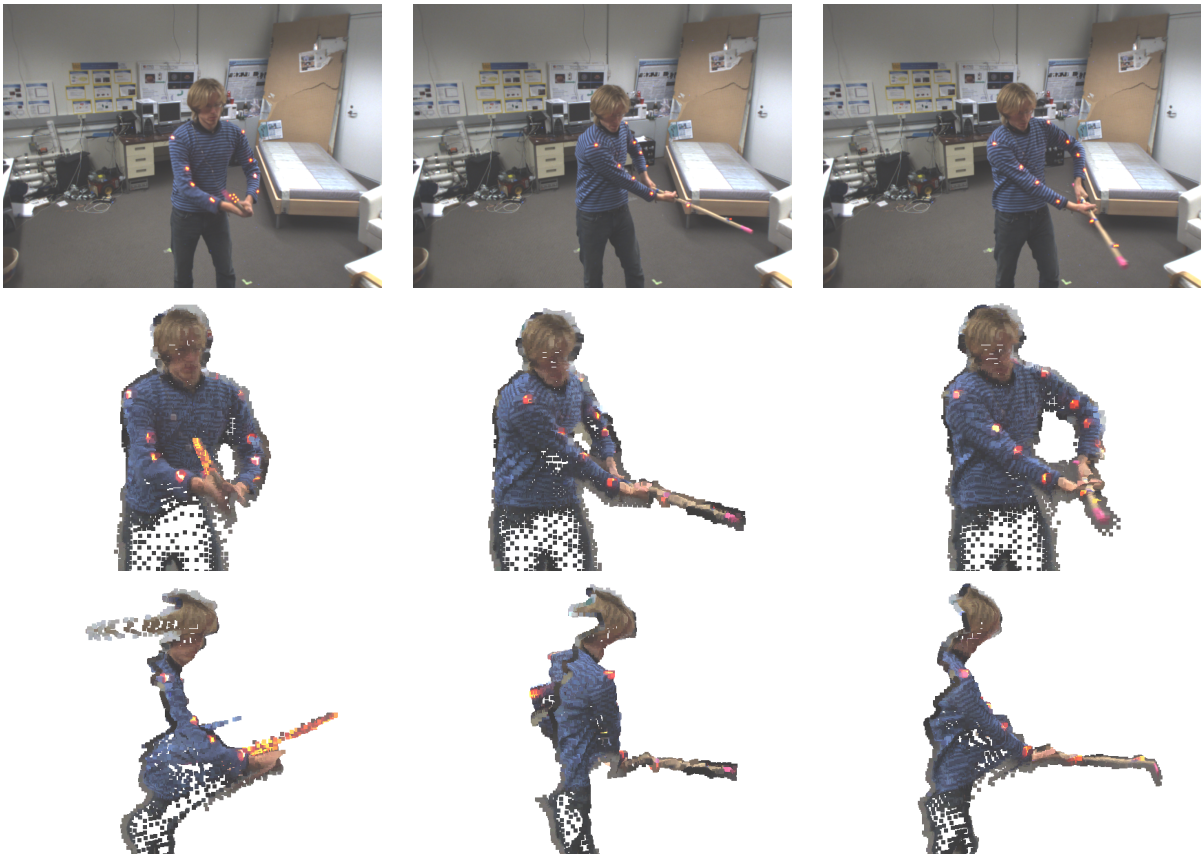


Fig. 8 Frames 91, 159 and 268 from the first test sequence. On the top is one image from the camera; in the middle and in the bottom is a rendering of the stereo data from two different view points. The final objective is to extract $\boldsymbol{\theta}_t$ from such data.

independent joint angles. Specifically, we linear extrapolate the joint angles, i.e. we define the mean of the predictive distribution as

$$\bar{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}_t + (\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) \ . \tag{31}$$

The predictive distribution is then defined as a Von Mises distribution (Bishop, 2006) with the above mean, which is constrained to respect the joint limits. Precisely, the predictive distribution is defined as

$$\begin{aligned} &p\big(\boldsymbol{\theta}_{t+1} \,\big|\, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}\big) \\ &\quad \propto \mathcal{U}_\Theta(\boldsymbol{\theta}_{t+1}) \prod_{n=1}^{N} \exp\left(\kappa_n \cos(\boldsymbol{\theta}_{t+1}[n] - \bar{\boldsymbol{\theta}}_{t+1}[n])\right) \end{aligned} \tag{32}$$

This model is conceptually the one proposed by Poon and Fleet (2002), except our noise model is a Von Mises distribution whereas a Normal distribution was previously applied.

7.1.2 End-effector Motion Model

We compare the linear predictor in angle space to a linear predictor in the space of end-effector goals. Specifically, we focus on the spatial positions of the head and the hands, such that $\mathbf{g}_t$ denotes the goal of these. We then define their motion as

$$p(\mathbf{g}_{t+1}|\mathbf{g}_t, \mathbf{g}_{t-1}) = \mathcal{N}(\mathbf{g}_{t+1}|\mathbf{g}_t + (\mathbf{g}_t - \mathbf{g}_{t-1}), \sigma^2\mathbf{I}). \tag{33}$$

The predictive distribution in angle space is then created as described in sec. 4.

7.1.3 Experimental Setup

To evaluate the quality of the attained results we position motion capture markers on the arms of the test subject. We then 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

$$\mathcal{E}(\boldsymbol{\theta}_{1:T}) = \frac{1}{TM} \sum_{t=1}^{T} \sum_{m=1}^{M} D(\boldsymbol{\theta}_t, \mathbf{v}_{mt}) \ , \tag{34}$$

where $D(\boldsymbol{\theta}_t, \mathbf{v}_{mt})$ is the shortest Euclidean distance between the m^{th} motion capture marker and the skin at time t.

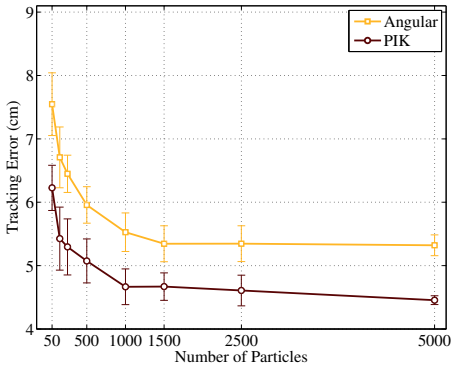


Fig. 13 The error measure $\mathcal{E}(\boldsymbol{\theta}_{1:T})$ plotted as a function of the number of particles. The shown results are averaged over several trials; the error bars correspond to one standard deviation.

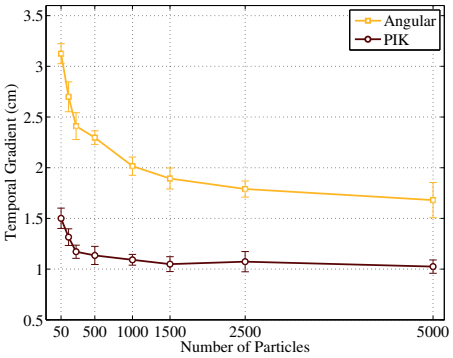


Fig. 14 The smoothness measure $\mathcal{S}(\boldsymbol{\theta}_{1:T})$ plotted as a function of the number of particles. Low values indicate smooth trajectories. The shown results are averaged over several trials; the error bars correspond to one standard deviation.

If the observation density $p(\boldsymbol{\theta}_t|\mathbf{X}_t)$ is noisy, then the motion model tends to act as a smoothing filter. This can be of particular importance when observations are missing, e.g. during self-occlusions. When evaluating the quality of a motion model it, thus, can be helpful to look at the smoothness of the attained pose sequence. To measure this, we simply compute the average size of the temporal gradient. We approximate this gradient using finite differences, and hence use

$$\mathcal{S}(\boldsymbol{\theta}_{1:T}) = \frac{1}{TL}\sum_{t=1}^{T}\sum_{l=1}^{L}||\mathbf{a}_{lt} - \mathbf{a}_{l,t-1}|| \tag{35}$$

as a measure of smoothness.

7.1.4 Evaluation

To see how the two motion models compare we apply them several times to the same sequence with a variable number of particles. The tracker is manually initialised in the first frame. Visualisations of the attained results for a few frames are available in fig. 9–12. Movies with the same results are also available on-line[2]. Due to the fast motions and poor stereo reconstructions, both models have difficulties tracking the subject in all frames. However, visually, it is evident that the end-effector motion model provides more accurate tracking and more smooth motion trajectories compared to the angular motion model.

In order to quantify these visual results, we compute the error measure presented in eq. 34 for results attained using different number of particles. Fig. 13 shows this. Here the results for each number of particles has been averaged over several trials. It is worth noticing that our model consistently outperforms the model in angle space.

[2] `http://humim.org/pik-tracker/`

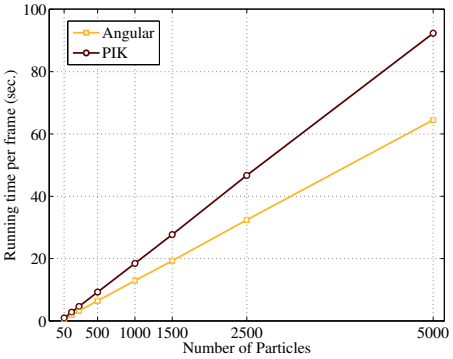
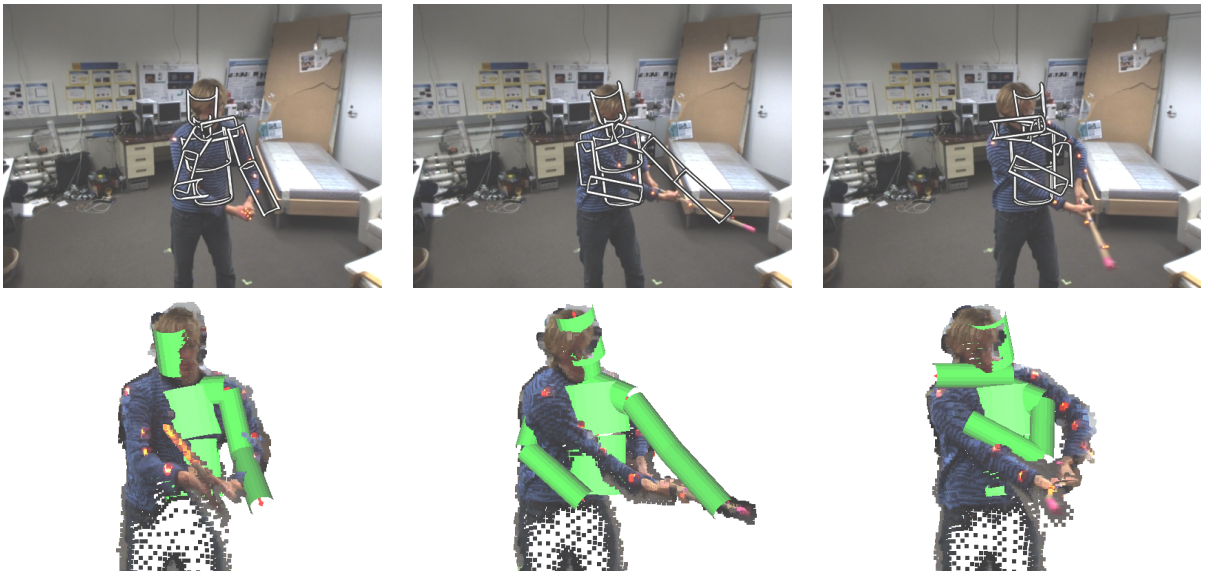


Fig. 15 The running time per frame of the tracking system when using different motion models. The end-effector based model approximately requires 1.4 times longer per frame.

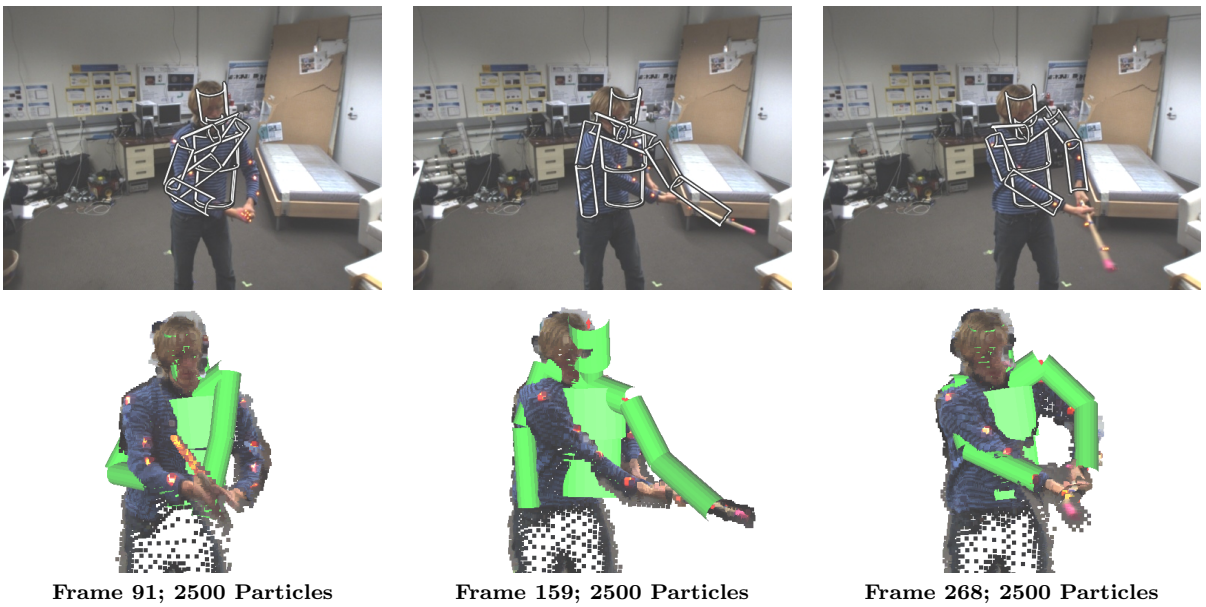
We also measure the smoothness $\mathcal{S}$ of the attained pose sequences as a function of the number of particles. This is plotted in fig. 14. As can be seen, our model is consistently more smooth compared to the linear model in angle space. This result is an indication that our model is less sensitive to noise in the observational data.

In summary, we see that our model allows for improved motion estimation using fewer particles compared to a linear model in angle space. This, however, comes at the cost of a computationally more expensive prediction. One might then ask, when this model improves the efficiency of the entire program. The answer to this question depends on the computation time required by the visual measurement system as this most often is the computationally most demanding part of tracking systems. For our system, the running time per frame is plotted in fig. 15. Comparing this with fig. 13, we see that our model produces superior results for fixed computational resources.



Frame 91; 500 Particles **Frame 159; 500 Particles** **Frame 268; 500 Particles**

Fig. 9 Results in frame 91, 159 and 268 using the angular model with 500 particles.



Frame 91; 2500 Particles **Frame 159; 2500 Particles** **Frame 268; 2500 Particles**

Fig. 10 Results in frame 91, 159 and 268 using the angular model with 2500 particles.

7.2 Human–Object Interaction

In the previous section we saw that the activity independent end-effector model improved the results of the angular model. However, results were still not perfect due to the poor data. Inspired by the work of Kjellström et al (2010) we now specialise the end-effector model to include knowledge of the stick position. Assume we know the person is holding on to the stick and that we know the end points of the stick. As the stick is linear, we can write the hand positions as a linear combination of the stick end points.

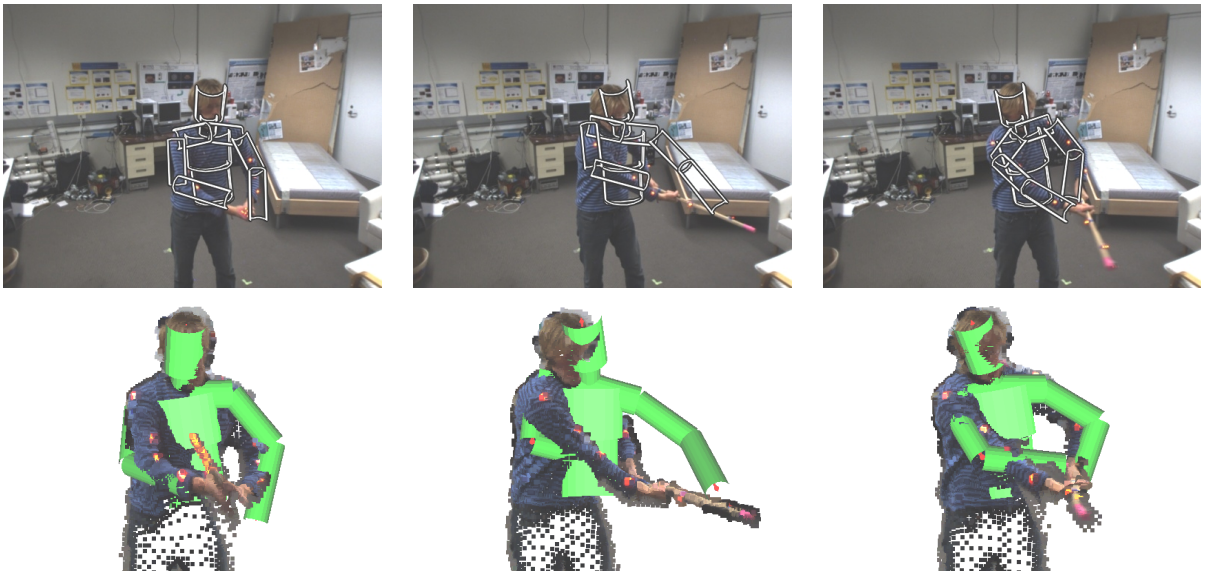
We now model the goal positions of the hands as such a linear combination, i.e. we let

$$\mathbf{g}_t = \mathbf{s}_t^{(1)}\gamma_t + \mathbf{s}_t^{(2)}(1-\gamma_t) \ , \tag{36}$$

where $\mathbf{s}_t^{(i)}$ denotes the end points of the stick. Here we let γ_t follow a normal distribution confined to the unit interval, i.e.

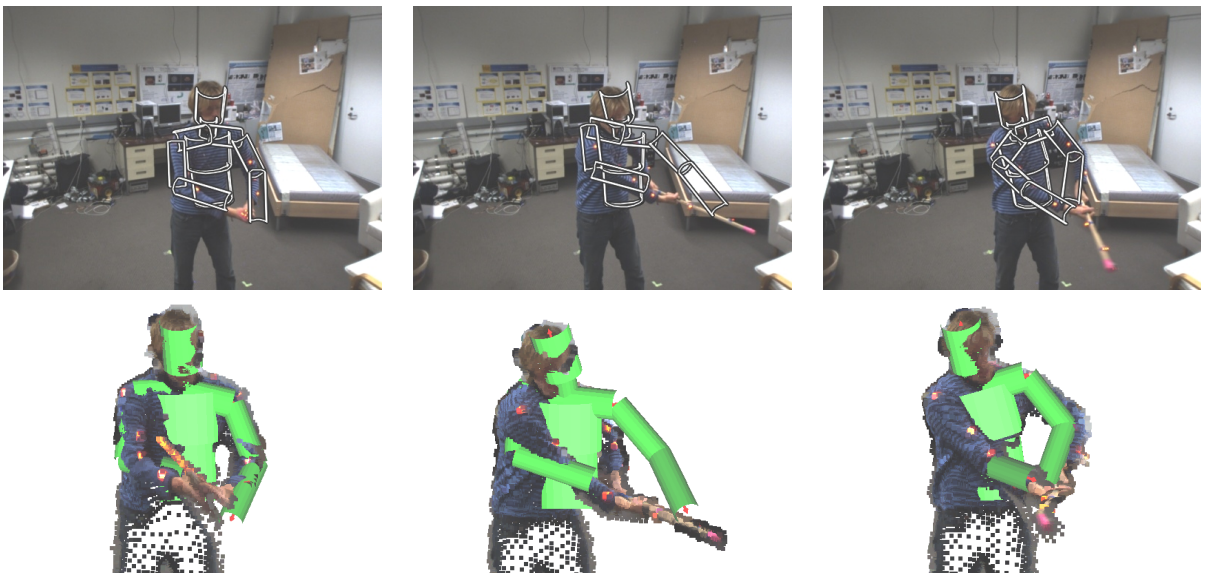
$$p(\gamma_t|\gamma_{t-1}) \propto \mathcal{N}\left(\gamma_t|\gamma_{t-1},\sigma^2\right)\mathcal{U}_{[0,1]}(\gamma_t) \ . \tag{37}$$

We now have a motion model that describes how the person is interacting with the stick. To apply this model



Frame 91; 500 Particles **Frame 159; 500 Particles** **Frame 268; 500 Particles**

Fig. 11 Results in frame 91, 159 and 268 using the end-effector model with 500 particles.

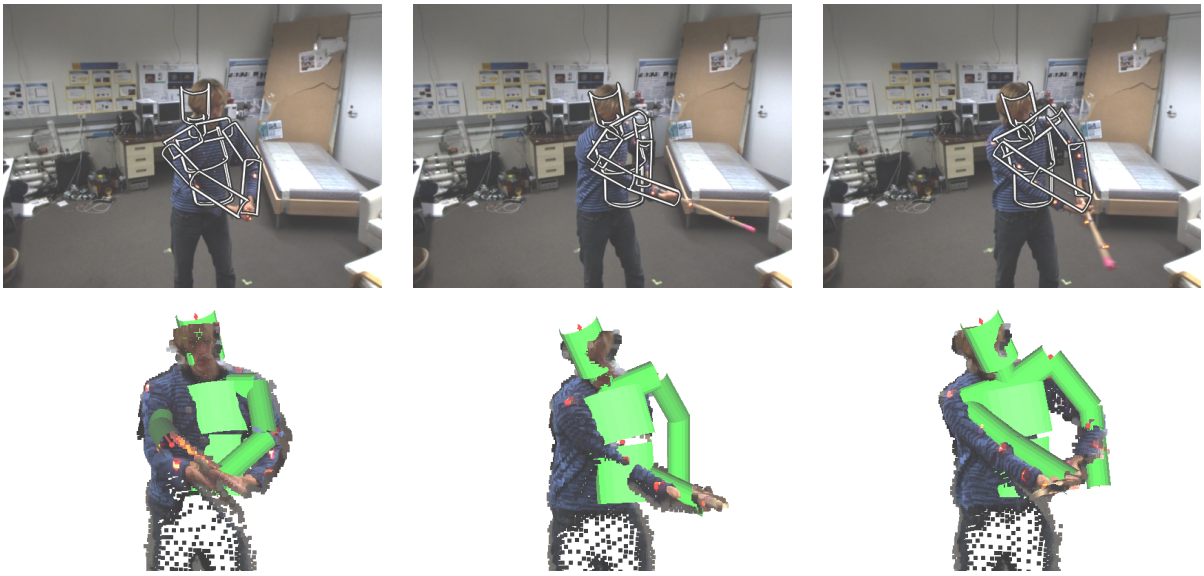


Frame 91; 2500 Particles **Frame 159; 2500 Particles** **Frame 268; 2500 Particles**

Fig. 12 Results in frame 91, 159 and 268 using the end-effector model with 2500 particles.

we need to know the end points of the stick at each frame. Here, we simply attain these by placing markers from the optical motion capture system on the stick. In practical scenarios, one would often only have the two dimensional image positions of the stick end points available. The model can be extended to handle this by restricting the hand goals to the plane spanned by the lines starting at the optical centre going through the stick end points in the image plane (Hauberg and Pedersen, 2011).

We apply this object interaction model to the same sequence as before. In fig. 16 we show the attained results using 500 particles on the same frames as before. As can be seen, the results are better than any of the previously attained results even if we are using fewer particles. This is also evident in fig. 17, where the tracking error is shown. In general, it should be of no surprise that results can be improved by incorporating more activity dependent knowledge; the interesting part is the ease with which the knowledge could be incorporated into the model.



Frame 91; 500 Particles **Frame 159; 500 Particles** **Frame 268; 500 Particles**

Fig. 16 Results in frame 91, 159 and 268 using the object interaction model.

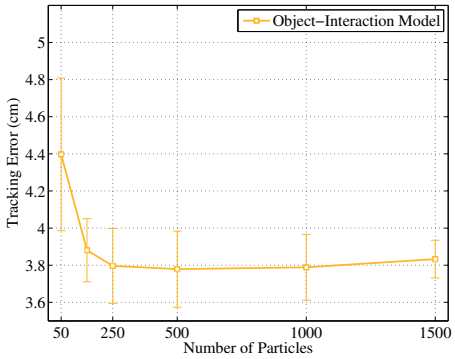


Fig. 17 The error measure $\mathcal{E}(\boldsymbol{\theta}_{1:T})$ when using the object interaction model. The shown results are averaged over several trials; the error bars correspond to one standard deviation.

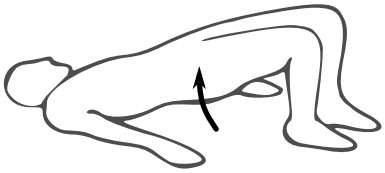


Fig. 18 An illustration of the *pelvic lift* exercise. The patient lies on a flat surface with head, feet and hands fixed. The pelvic region is lifted and lowered repeatedly.

7.3 The Pelvic Lift

We have seen that the suggested modelling framework can be useful to create activity independent motion models and to model interactions with objects. The main motivation for creating the framework is, however, to be able to easily model physiotherapeutic exercises. As a demonstration we will create such a model for the *the pelvic lift* exercise. The simple exercise is illustrated in fig. 18. The patient lies on the floor with bend knees. He or she must then repeatedly lift and lower the pelvic region.

To model this motion we focus on the position of the feet, the hands and the pelvic region. We fix the goals of the feet and the hands, such that they always aim at the position in which the tracking was initialised, i.e.

$$\mathbf{g}_{t+1} = \mathbf{g}_1 \qquad \text{for hand and feet.} \tag{38}$$

The root of the kinematic skeleton is placed at the pelvic region. We model the motion of the root as moving mostly up or downwards, by letting

$$\mathbf{root}_{t+1} = \mathbf{root}_t + \boldsymbol{\eta} \ , \tag{39}$$

where $\boldsymbol{\eta}$ is from a zero-mean Normal distribution with covariance $\mathrm{diag}(\sigma^2, \sigma^2, 16\sigma^2)$. Here, the factor 16 ensures large variation in the up and downwards directions. We further add the constraint that $\mathbf{root}_{t+1}$ must have a positive z-value, i.e. the root must be above the ground plane. This model illustrates the ease with which we can include knowledge of both the environment and the motion in the predictive process.

To illustrate the predictions given by this model, we sample 10 particles from the model. These are drawn in fig. 19. As can be seen, the position of the head, hands and feet show small variation, whereas the position of both the pelvic region and the knees shows more variation. It should be noted that the knees vary as we did

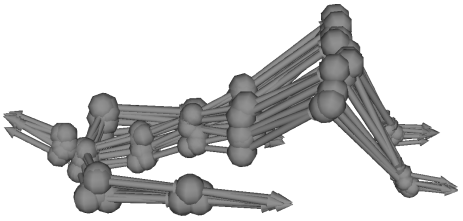


Fig. 19 Ten samples from the tracking of the pelvic lift exercise. Notice how the spine and pelvic region shows large variation in the vertical direction, the knees shows large variation in all directions, while other limbs show low variation. The precision matrix of the importance distribution used to generate one of these samples is shown in fig. 20.

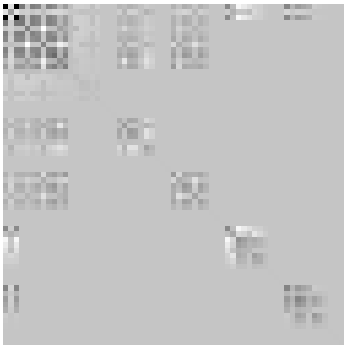


Fig. 20 Precision matrix of the importance distribution used for predicting one of the samples shown in fig. 19. Dark entries correspond to positive values, light entries correspond to negative values while the light grey covering most entries correspond to zero.

not model their motion. To gain further insight into the prediction, we plot the precision matrix of the importance distribution in fig. 20. It can be seen that this matrix has a block-structure indicating that clusters of joints are correlated.

We have applied this predictive model to a sequence where a test subject performs the pelvic lift exercise. The exercise is repeated 6 times and is successfully tracked using 100 particles. A few selected frames are available in fig. 21 and a movie is available on-line[2].

8 Conclusion

In this paper we have presented a probabilistic extension of inverse kinematics. With this, we are able to build predictive models for articulated human motion estimation from processes in the spatial domain. The advantages of this approach are many.

First, we have empirically demonstrated that our spatial motion models improve the tracking using far less particles, while they at the same time provide more smooth motion trajectories compared to simple models in the space of joint angles. In our experience, the traditional motion models in joint angle space actually provide little to no predictive power. The basic issue is that the spatial variance of limb coordinates tends to accumulate with these models as the kinematic chains are traversed. From a practical point of view this means that limbs which are far down the kinematic chains are rarely predicted correctly, meaning many particles are required. Our model does not suffer from this issue as we control the end positions of the chains. We believe this is the main cause of the models efficiency.

Secondly, we saw that the model allows us easily to take the environment into account. We saw this with the stick example, where we could trivially incorporate the stick position into the model. We also saw this with the pelvic lift example, where we modelled the ground plane.

Thirdly, our approach makes it easier to construct high quality models of a large class of motions. Specifically, *goal oriented* motions are usually easy to describe in spatial coordinates. This was demonstrated on the *pelvic lift* exercise, which is trivial to describe spatially, but complicated when expressed directly in terms of joint angles.

Fourthly, our models mostly works in the low dimensional space of end-effector goals, which is simply an ordinary Euclidean space. This makes it more practical to build motion models as we do not need to deal with the topology of the space of joint angles.

We feel there is great need for predictive motion models that are expressed spatially as this seems to mimic human motion plans more closely. It is, however, not clear if our approach of modelling end-effector *goals* is the best way to achieve spatial motion priors. It could be argued that it would be better to model the actual end-effector positions rather than their goals. Our strategy do have the advantage that the resulting optimisation problems are computationally feasible. It also allows us to study stochastic processes in ordinary Euclidean spaces. Had we instead chosen to model the actual end-effector positions, we would be forced to restrict our motion models to the reachable parts of the spatial domain, making the models more involved.

At the heart of our predictive models lies an inverse kinematics solver that computes the mode of eq. 12. In the future this solver should be extended with a collision detection system, such that self-penetrations would be disallowed. We are also actively working on determining more restrictive joint limit models (Engell-Nørregård et al, 2010). Both extensions would reduce the space of possible poses, which would allow us to reduce the number of particles even further.

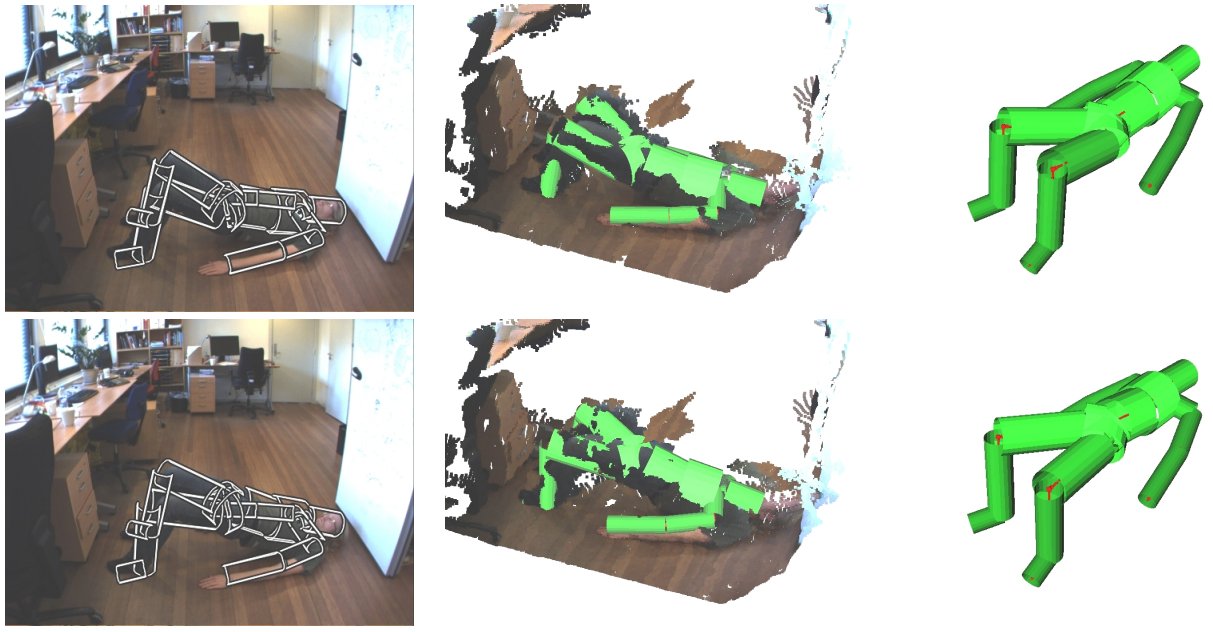


Fig. 21 Frames 71 and 159 from a sequence with a pelvic lift exercise. The exercise is repeated 6 times during approximately 200 frames. The video was recorded at approximately 10 frames per second.

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

Bishop CM (2006) Pattern Recognition and Machine Learning. Springer

Bregler C, Malik J, Pullen K (2004) Twist based acquisition and tracking of animal and human kinematics. International Journal of Computer Vision 56:179–194

Cappé O, Godsill SJ, 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 MA, Lu Z (2007) The Laplacian Eigenmaps Latent Variable Model. JMLR W&P 2:59–66

Courty N, Arnaud E (2008) Inverse kinematics using sequential monte carlo methods. In: Articulated Motion and Deformable Objects: 5th International Conference, Springer-Verlag New York Inc, pp 1–10

Elgammal AM, Lee CS (2009) Tracking People on a Torus. IEEE Transaction on Pattern Analysis and Machine Intelligence 31(3):520–538

Engell-Nørregård M, Hauberg S, Lapuyade J, Erleben K, Pedersen KS (2009) Interactive inverse kinematics for monocular motion estimation. In: Proceedings of VRIPHYS'09

Engell-Nørregård M, Niebe S, Erleben K (2010) Local joint–limits using distance field cones in euler angle space. In: Computer Graphics International

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

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

Ganesh S (2009) Analysis of goal-directed human actions using optimal control models. PhD thesis, EECS Dept., University of California, Berkeley

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

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

Hauberg S, Lapuyade J, Engell-Nørregård M, Erleben K, Pedersen KS (2009) Three Dimensional Monocular Human Motion Analysis in End-Effector Space. In: Cremers D, et al (eds) Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer, LNCS, pp 235–248

Hauberg S, Sommer S, Pedersen KS (2010) Gaussian-like spatial priors for articulated tracking. In: Daniilidis K, Maragos P, , Paragios N (eds) ECCV 2010, Springer, LNCS, vol 6311, 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, LCNS, vol 3022, Springer, pp 405–418

Horaud R, Niskanen M, Dewaele G, Boyer E (2009) Human motion tracking by registering an articulated surface to 3d points and normals. IEEE Transactions on Pattern Analysis and Machine Intelligence 31:158–163

Kerlow IV (2003) Art of 3D Computer Animation and Effects, 3rd edn. John Wiley & Sons

Kjellström H, Kragić D, Black MJ (2010) Tracking people interacting with objects. In: CVPR '10: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition

Knossow D, Ronfard R, Horaud R (2008) Human motion tracking with a kinematic parameterization of extremal contours. International Journal of Computer Vision 79(2):247–269

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

Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104(2):90–126

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

Murray RM, Li Z, Sastry SS (1994) A Mathematical Introduction to Robotic Manipulation. CRC Press

Nocedal J, Wright SJ (1999) Numerical optimization. Springer Series in Operations Research, Springer

Poon E, Fleet DJ (2002) Hybrid monte carlo filtering: Edge-based people tracking. IEEE Workshop on Motion and Video Computing 0:151

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

Rasmussen CE, Williams C (2006) Gaussian Processes for Machine Learning. MIT Press

Rosenhahn B, Schmaltz C, Brox T, Weickert J, Cremers D, Seidel HP (2008) Markerless motion capture of man-machine interaction. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 0:1–8

Salzmann M, Urtasun R (2010) Combining discriminative and generative methods for 3d deformable surface and articulated pose reconstruction. In: Proceedings of CVPR'10

Sidenbladh H, Black MJ, Fleet DJ (2000) Stochastic tracking of 3d human figures using 2d image motion. In: Proceedings of ECCV'00, Springer, Lecture Notes in Computer Science 1843, vol II, 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

Sminchisescu C, Triggs B (2003) Kinematic Jump Processes for Monocular 3D Human Tracking. In: In IEEE International Conference on Computer Vision and Pattern Recognition, pp 69–76

Tournier M, Wu X, Courty N, Arnaud E, Reveret L (2009) Motion compression using principal geodesics analysis. Computer Graphics Forum 28(2):355–364

Urtasun R, Fleet DJ, 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

Urtasun R, Fleet DJ, 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 DJ, Geiger A, Popović J, Darrell TJ, Lawrence ND (2008) Topologically-constrained latent variable models. In: ICML '08: Proceedings of the 25th international conference on Machine learning, ACM, pp 1080–1087

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

Zhao J, Badler NI (1994) Inverse kinematics positioning using nonlinear programming for highly articulated figures. ACM Transaction on Graphics 13(4):313–336

Paper 2: Gaussian-like Spatial Priors for Articulated Tracking

Authors: Søren Hauberg, Stefan Sommer and Kim S. Pedersen.
Status: Published at the *European Conference on Computer Vision* 2010 [24].

The "work horse" of the *Probabilistic Inverse Kinematics* model was the importance distribution [21, eq. 22],

$$q(\theta_t|\theta_{t-1}, \mathbf{X}_t) \propto \mathcal{N}\left(\theta_t|\theta_t^*, \left(\mathbf{J}^T\mathbf{W}\mathbf{J} + \lambda\mathbf{I}\right)^{-1}\right)\mathcal{U}_\Theta(\theta_t) \ .$$

While the issue was subtle, this equation was troubling me. The inverse covariance matrix of q consists of two terms: $\mathbf{J}^T\mathbf{W}\mathbf{J}$ and $\lambda\mathbf{I}$. The first term is not full rank as $\mathbf{J}$ is non-square. This explains why the second term is needed: it is effectively a regularisation term added in the joint angle domain. The need for this joint angle regularisation was somewhat disappointing, given our goal of developing a model in the joint position domain. One question kept reappearing: could we avoid regularising in the joint angle domain? The immediate answer was to avoid regularisation in the first place. As the rank of $\mathbf{J}^T\mathbf{W}\mathbf{J}$ is 3 times the number of end-effectors, the regularisation can be avoided simply be increasing the number of end-effectors in the model. The extreme case is to use *all* joints as end-effectors, which leads to an interesting geometric point of view.

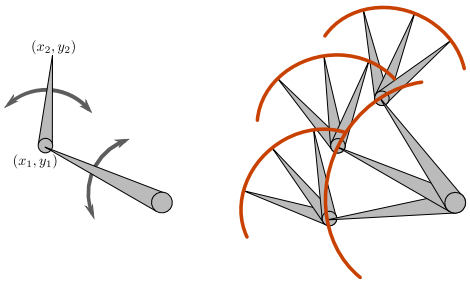


Figure 3.1 *An illustration of the kinematic manifold $\mathcal{M}$ for a planar two-bone skeleton.* Left: *the two-bone skeleton, with arrows indicating how joints rotate.* Right: *different configurations of the skeleton; the orange circular paths indicate parts of points on the manifold. On this figure, the embedding space is 4-dimensional corresponding to (x_1, y_1, x_2, y_2) while $\mathcal{M}$ is 2-dimensional due to two constraints (one for each bone).*

Instead of thinking of the kinematic skeleton as parametrised by a set of joint angles, we can think of it as being parametrised by the joint positions. These joint positions are, however, constrained by the configuration of the skeleton, i.e. the hand must be at a certain distance from the elbow. Hence, the vector of all joint positions is confined to a subset $\mathcal{M}$ of the Euclidean space containing the joints

$$\mathcal{M} \equiv \{F(\theta) \ | \ \theta \in \mathbf{\Theta}\} \ .$$

As the forward kinematics function is quite well-behaved, $\mathcal{M}$ turns out to be a Riemannian manifold embedded in the Euclidean space. This gives rise to a geometric interpretation of both inverse kinematics and the above expression for the inverse covariance of q. Inverse kinematics becomes a projection operator while $\mathbf{J}^T\mathbf{W}\mathbf{J}$ is strongly related to the tangent space of $\mathcal{M}$. This observation allowed us to study variance properties of Gaussian priors in the joint angle space as well as define priors directly on the manifold.

A word on terminology: in Computer Vision the word *manifold* is being used somewhat loosely, at times to indicate a *subset* of the embedding space. In this thesis, the word manifold is used in the Riemannian sense, i.e. a subset with a *smoothly* varying metric.

Gaussian-like Spatial Priors for Articulated Tracking

Søren Hauberg, Stefan Sommer, and Kim Steenstrup Pedersen

{hauberg, sommer, kimstp}@diku.dk,
The eScience Centre, Dept. of Computer Science, University of Copenhagen

Abstract. We present an analysis of the spatial covariance structure of an articulated motion prior in which joint angles have a known covariance structure. From this, a well-known, but often ignored, deficiency of the kinematic skeleton representation becomes clear: spatial variance not only depends on limb lengths, but also increases as the kinematic chains are traversed. We then present two similar Gaussian-like motion priors that are explicitly expressed spatially and as such avoids any variance coming from the representation. The resulting priors are both simple and easy to implement, yet they provide superior predictions.

Key words: Articulated Tracking · Motion Analysis · Motion Priors · Spatial Priors · Statistics on Manifolds · Kinematic Skeletons

1 Articulated Tracking

Three dimensional articulated human motion tracking is the process of estimating the configuration of body parts over time from sensor input [1]. One approach to this estimation is to use motion capture equipment where e.g. electromagnetic markers are attached to the body and then tracked in three dimensions. While this approach gives accurate results, it is intrusive and cannot be used outside laboratory settings. Alternatively, computer vision systems can be used for non-intrusive analysis. These systems usually perform some sort of optimisation for finding the best configuration of body parts. This optimisation is often guided by a system for predicting future motion. This paper concerns such a predictive system for general purpose tracking. Unlike most previous work, we build the actual predictive models in spatial coordinates, rather than working directly in the space of configuration parameters.

In the computer vision based scenario, the objective is to estimate the human pose in each image in a sequence. When only using a single camera, or a narrow baseline stereo camera, this is inherently difficult due to self-occlusions. This manifests itself in that the distribution of the human pose is multi-modal with an unknown number of modes. To reliably estimate this pose distribution we need methods that cope well with multi-modal distributions. Currently, the best method for such problems is the particle filter [2], which represents the distribution as a set of weighted samples. These samples are propagated in time using a predictive model and assigned a weight according to a data likelihood. As

such, the particle filter requires two subsystems: one for computing likelihoods by comparing the image data to a sample from the pose distribution, and one for predicting future poses. In terms of optimisation, the latter guides the search for the optimal pose. In practice, the predictive system is essential in making the particle filter computationally feasible, as it can drastically reduce the number of needed samples.

1.1 The Kinematic Skeleton

Before discussing the issues of human motion analysis, we pause to introduce the actual representation of the human pose. In this paper, we use the *kinematic skeleton* (see Fig. 1a), which is by far the most common choice [1]. 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*.

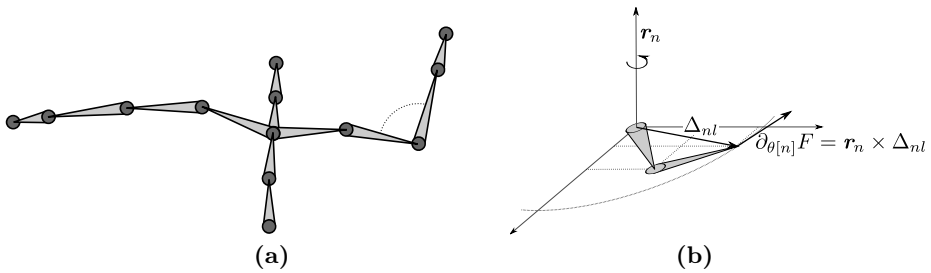


Fig. 1. (a) A rendering of the kinematic skeleton. Each bone position is computed by a rotation and a translation relative to its parent. The circles, are collectively referred to as the *end-effectors*. (b) The derivative of an end point with respect to a joint angle. This is computed as the cross product of the rotational axis $\boldsymbol{r}_n$ and the vector from the joint to the end-effector.

We model the bones as having known constant length (i.e. rigid), so the direction of each bone constitute the only degrees of freedom in the kinematic skeleton. The direction in each joint can be parametrised with a vector of angles, noticing that different joints may have different number of degrees of freedom. We may collect all joint angle vectors into one large vector $\boldsymbol{\theta}$ representing all joint angles in the model. This vector will then be confined to the N dimensional torus $\mathbb{T}^N$.

Forward Kinematics From known bone lengths and a joint angle vector $\boldsymbol{\theta}$, it is straight-forward to compute the spatial coordinates of the bones. Specifically, the purpose is to compute the spatial coordinates of the end points of each bone. This process is started at the root of the tree structure and moves recursively along the branches, which are known as the *kinematic chains*.

The root of the tree is placed at the origin of the coordinate system. The end point of the next bone along a kinematic chain is then computed by rotating the coordinate system and translating the root along a fixed axis relative to the parent bone, i.e.

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

where $\boldsymbol{a}_l$ is the l^{th} end point, and $\mathbf{R}_l$ and $\boldsymbol{t}_l$ denotes a rotation and a translation respectively. The rotation is parametrised by the relevant components of the pose vector $\boldsymbol{\theta}$ and the length of the translation corresponds to the known length of the bone. We can repeat this process recursively until the entire kinematic tree has been traversed. This process is known as *Forward Kinematics* [3].

The rotation matrix $\mathbf{R}_l$ of the l^{th} bone is parametrised by parts of $\boldsymbol{\theta}$. The actual number of used parameters depends on the specific joint. For elbow joints, we use one parameter, while we use three parameters to control all other joints. These two different joint types are respectively known as *hinge joints* and *ball joints*.

Using forward kinematics, we can compute the spatial coordinates of the end points of the individual bones. These are collectively referred to as *end-effectors*. In Fig. 1a these are drawn as circles. We will denote the coordinates of all end-effectors by $F(\boldsymbol{\theta})$. We will assume the skeleton contains L end-effectors, such that $F(\boldsymbol{\theta}) \in \mathbb{R}^{3L}$.

It should be clear that while $F(\boldsymbol{\theta}) \in \mathbb{R}^{3L}$, the end-effectors does not cover all of this space. There is, for instance, an upper bound on how far the hands can be apart. Specifically, we see that $F(\boldsymbol{\theta}) \in \mathcal{M} \subset \mathbb{R}^{3L}$, where $\mathcal{M}$ is a compact differentiable manifold embedded in $\mathbb{R}^{3L}$ (since $\mathbb{T}^N$ is compact and F is an injective function with full-rank Jacobian).

Derivative of Forward Kinematics Later, we shall be in need of the Jacobian of F. This consists of a column for each component of $\boldsymbol{\theta}$. Each such column can be computed in a straightforward manner [4]. Let $\boldsymbol{r}_n$ denote the unit-length rotational axis of the n^{th} angle and Δ_{nl} the vector from the joint to the l^{th} end-effector. The entries of the column corresponding to the l^{th} end-effector can then be computed as $\partial_{\boldsymbol{\theta}[n]} F_l = \boldsymbol{r}_n \times \Delta_{nl}$. This is merely the tangent of the circle formed by the end-effector when rotating the joint in question as is illustrated in Fig. 1b.

Joint Constraints In the human body, bones cannot move freely. A simple example is the elbow joint, which can approximately only bend between 0 and 120 degrees. To represent this, $\boldsymbol{\theta}$ is confined to a subset $\boldsymbol{\Theta}$ of $\mathbb{T}^N$. With this further restriction, $\mathcal{M}$ becomes a manifold with boundary.

For simplicity, $\boldsymbol{\Theta}$ is often defined by confining each component of $\boldsymbol{\theta}$ to an interval, i.e. $\boldsymbol{\Theta} = \prod_{n=1}^{N} [l_n, u_n]$, where l_n and u_n denote the lower and upper bounds of the n^{th} component. This type of constraints on the angles is often called *box constraints* [5].

1.2 Related Work

Most work in the articulated tracking literature falls in two categories. Either the focus is on improving the image likelihoods or on improving the predictions. Due to space constraints, we forgo a review of various likelihood models as this paper is focused on prediction. For an overview of likelihood models, see the review paper by Poppe [1].

Most work on improving the predictions, is focused on learning motion specific priors, such as for *walking* [6–12]. Currently, the most popular approach is to restrict the tracker to some subspace of the joint angle space. Examples include, the work of Sidenbladh et al [10] where the motion is confined to a linear subspace which is learned using PCA. Similarly, Sminchisescu and Jepson [8] use spectral embedding to learn a non-linear subspace; Lu et al [9] use the Laplacian Eigenmaps Latent Variable Model [13] to perform the learning, and Urtasun et al [14] use a Scaled Gaussian Process Latent Variable Model [15]. This strategy has been improved even further by Urtasun et al [12] and Wang et al [7] such that a stochastic process is learned in the non-linear subspace as well. These approaches all seem to both stabilise the tracking and make it computationally less demanding. The downside is, of course, that the priors are only applicable when studying specific motions.

When it comes to general purpose priors, surprisingly little work has been done. Such priors are not only useful for studying general motion but can also be useful as hyperpriors for learning motion specific priors. The common understanding seems to be that the best general purpose prior is to assume that the joint angles follow a Gaussian distribution. Specifically, many researchers assume

$$p_{\text{angle}}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \propto \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \Sigma_{\boldsymbol{\theta}})\, \mathcal{U}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t) \ , \tag{2}$$

where $\mathcal{U}_{\boldsymbol{\Theta}}$ denotes the uniform distribution on $\boldsymbol{\Theta}$ enforcing the angular constraints and the subscript t denotes time. We shall call this model the *Angular Prior*. In practice, $\Sigma_{\boldsymbol{\theta}}$ is often assumed to be diagonal or isotropic. This model has, amongst others, been applied by Sidenbladh et al [10], Balan et al [16] and Bandouch et al [17]. At first sight, this model seems quite innocent, but, as we shall see, it has a severe downside.

1.3 Our Contribution and Organisation of the Paper

In Sec. 2 we provide an analysis of the spatial covariance of the common motion prior from Eq. 2. While the formal analysis is novel, its conclusions are not surprising. In Sec. 3, we suggest two similar motion priors that are explicitly designed to avoid the problems identified in Sec. 2. This work constitutes the main technical contribution of the paper. In order to compare the priors we implement an articulated tracker, which requires a likelihood model. We briefly describe a simple model for this in Sec. 4. The resulting comparison between priors is performed in Sec. 5 and the paper is concluded in Sec. 6.

2 Spatial Covariance Structure of the Angular Prior

While the covariance structure of $\boldsymbol{\theta}_t$ in Eq. 2 is straight-forward, the covariance of $F(\boldsymbol{\theta}_t)$ is less simple. This is due to two phenomena:

1. **Variance depends on distance between joint and end-effector.** When a joint angle is changed, it alters the position of the end point of the limb attached to the joint. This end point is moved on a circle with radius corresponding to the distance between the joint and the end point. This means the end point of a limb far away from the joint can change drastically with small changes of the joint angle.
2. **Variance accumulates.** When a joint angle is changed, all limbs that are further down the kinematic chain will move. This means that when, e.g., the shoulder joint changes both hand and elbow moves. Since the hand also moves when the elbow joint changes, we see that the hand position varies more than the elbow position.

Neither of these two phenomena seem to have come from well-founded modelling perspectives.

To get a better understanding of the covariance of limb positions, we seek an expression for $\mathrm{cov}[F(\boldsymbol{\theta}_t)]$. Since $F(\boldsymbol{\theta}_t)$ lies on a non-linear manifold $\mathcal{M}$ in $\mathbb{R}^{3L}$, such an analysis is not straight-forward. Instead of computing the covariance on this manifold, we compute it in the tangent space at the mean value $\bar{\boldsymbol{\theta}}_t = \mathbb{E}(\boldsymbol{\theta}_t)$ [18]. This requires the Logarithm map of $\mathcal{M}$, which we simply approximate by the Jacobian $\mathbf{J}_{\bar{\boldsymbol{\theta}}_t} = \partial_{\boldsymbol{\theta}_t} F(\boldsymbol{\theta}_t)|_{\boldsymbol{\theta}_t = \bar{\boldsymbol{\theta}}_t}$ of the forward kinematics function, such that

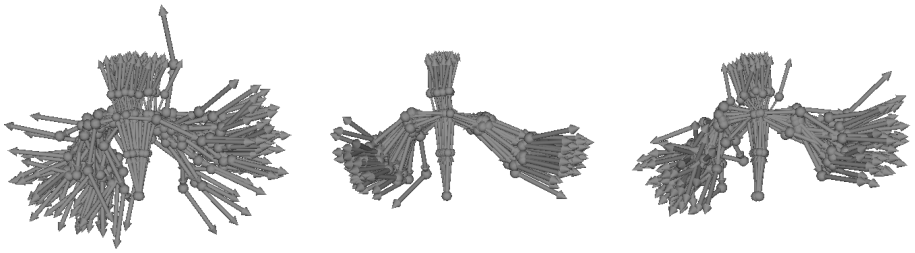
$$\mathrm{cov}[F(\boldsymbol{\theta}_t)] \approx \mathrm{cov}[\mathbf{J}_{\bar{\boldsymbol{\theta}}_t}\boldsymbol{\theta}_t] = \mathbf{J}_{\bar{\boldsymbol{\theta}}_t}\,\mathrm{cov}[\boldsymbol{\theta}_t]\mathbf{J}_{\bar{\boldsymbol{\theta}}_t}^T = \mathbf{J}_{\bar{\boldsymbol{\theta}}_t}\Sigma_{\boldsymbol{\theta}_t}\mathbf{J}_{\bar{\boldsymbol{\theta}}_t}^T \ . \tag{3}$$

As can be seen, the covariance of the limb positions is highly dependent on the Jacobian of F. A slightly different interpretation of the used approximation is that we linearise F around the mean, and then compute the covariance.

We note that $\left\|\partial_{\boldsymbol{\theta}_t[n]} F_l\right\| = \|\Delta_{nl}\|$, meaning that the variance of a limb is linearly dependent on the distance between the joint and the limb end point. This is the first of the above mentioned phenomena. The second phenomena comes from the summation in the matrix product in Eq. 3. It should be stressed that this behaviour is a consequence of the choice of representation and will appear in any model that is expressed in terms of joint angles unless it explicitly performs some means of compensation. We feel this is unfortunate, as the behaviour does not seem to have its origins in an explicit model design decision. Specifically, it hardly seems to have any relationship with natural human motion (see the discussion of Fig. 2a below).

In practice, both of the above mentioned phenomena are highly visible in the model predictions. In Fig. 2a we show 50 samples from Eq. 2. Here, the joint angles are assumed to be independent, and the individual variances are learned from ground truth data of a sequence studied in Sec. 5. As can be seen the spatial variance increases as the kinematic chains are traversed. In practice, this behaviour reduces the predictive power of the model drastically; in our

experience the model practically has no predictive power at all. Bandouch et al [17] suggested using *Partitioned Sampling* [19] to overcome this problem. This boils down to fitting individual limbs one at a time as the kinematic chains are traversed, such that e.g. the upper arm is fitted to the data before the lower arm. While this approach works, we believe it is better to fix the model rather than work around its limitations. As such, we suggest expressing the predictive model directly in terms of spatial limb positions.



(a) p_{angle} **(b)** p_{proj} **(c)** $p_{\mathcal{T}}$

Fig. 2. Fifty samples from the different priors. The variance parameters for these distributions were assumed independent and was learned from ground truth data for a sequence studied in Sec. 5. (a) The angular prior p_{angle}. (b) The projected prior p_{proj}. (c) The tangent space prior $p_{\mathcal{T}}$.

3 Two Spatial Priors

Informally, we would like a prior where each limb position is following a Gaussian distribution, i.e.

$$p_{idea}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \mathcal{N}(F(\boldsymbol{\theta}_t)|F(\boldsymbol{\theta}_{t-1}), \Sigma) \ . \tag{4}$$

This is, however, *not* possible as the Gaussian distribution covers the entire $\mathbb{R}^{3L}$, whereas $F(\boldsymbol{\theta}_t)$ is confined to $\mathcal{M}$. In the following, we suggest two ways of overcoming this problem.

3.1 Projected Prior

The most straight-forward approach is to define $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ by projecting Eq. 4 onto $\mathcal{M}$, i.e.

$$p_{\text{proj}}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \text{proj}_{\mathcal{M}}\left[\mathcal{N}(F(\boldsymbol{\theta}_t)|F(\boldsymbol{\theta}_{t-1}), \Sigma_{\text{proj}})\right] \ . \tag{5}$$

When using a particle filter for tracking, we only need to be able to draw samples from the prior model. We can easily do this by sampling from Eq. 4 and projecting the result onto $\mathcal{M}$. This, however, requires an algorithm for performing the projection.

Let $\boldsymbol{x}_t$ denote a sample from Eq. 4; we now seek $\hat{\boldsymbol{\theta}}_t$ such that $F(\hat{\boldsymbol{\theta}}_t) = \text{proj}_{\mathcal{M}}[\boldsymbol{x}_t]$. We perform the projection in a direct manor by seeking

$$\hat{\boldsymbol{\theta}}_t = \min_{\boldsymbol{\theta}_t} \left|\left|\boldsymbol{x}_t - F(\boldsymbol{\theta}_t)\right|\right|^2 \qquad \text{s.t.} \qquad \boldsymbol{l} \leq \boldsymbol{\theta}_t \leq \boldsymbol{u} \ , \tag{6}$$

where the constraints corresponds to the joint limits. This is an overdetermined constrained non-linear least-squares problem, that can be solved by any standard algorithm. We employ a projected steepest descent with line-search [5], where the search is started in $\boldsymbol{\theta}_{t-1}$. To perform this optimisation, we need the gradient of Eq. 6, which is readily evaluated as $\partial_{\boldsymbol{\theta}_t} \|\boldsymbol{x}_t - F(\boldsymbol{\theta}_t)\|^2 = 2(\boldsymbol{x}_t - F(\boldsymbol{\theta}_t))^T \mathbf{J}_{\boldsymbol{\theta}_t}$.

In Fig. 2b we show 50 samples from this distribution, where Σ_{proj} is assumed to be a diagonal matrix with entries that have been learned from ground truth data of a sequence from Sec. 5. As can be seen, this prior is far less variant than the Gaussian prior p_{angle} on joint angles.

3.2 Tangent Space Prior

While the projected prior provides us with a suitable prior, it does come with the price of having to solve a non-linear least-squares problem. If the prior is to be used as e.g a regularisation term in a more complicated learning scheme, this can complicate the models substantially. As an alternative, we suggest a slight simplification that allows us to skip the non-linear optimisation. Instead of letting $F(\boldsymbol{\theta}_t)$ be Gaussian distributed in $\mathbb{R}^{3L}$, we define it as being Gaussian distributed in the tangent space $\mathcal{T}$ of $\mathcal{M}$ at $F(\boldsymbol{\theta}_{t-1})$. That is, we define our prior such that

$$p_{\mathcal{T}}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \mathcal{N}_{\mathcal{T}}(F(\boldsymbol{\theta}_t)|F(\boldsymbol{\theta}_{t-1}), \Sigma_{\mathcal{T}}) \ , \tag{7}$$

where $\mathcal{N}_{\mathcal{T}}$ denotes a Gaussian distribution in $\mathcal{T}$. A basis of the tangent space is given by the columns of the Jacobian $\mathbf{J}_{\boldsymbol{\theta}_{t-1}}$. From Eq. 3 we know that the covariance structure near $F(\boldsymbol{\theta}_{t-1})$ in this model is $\Sigma_{\mathcal{T}} = \mathbf{J}_{\boldsymbol{\theta}_{t-1}} \Sigma_{\boldsymbol{\theta}} \mathbf{J}^T_{\boldsymbol{\theta}_{t-1}}$. In general, $\mathbf{J}_{\boldsymbol{\theta}_{t-1}}$ is not square, so we cannot isolate $\Sigma_{\boldsymbol{\theta}}$ from this equation simply by inverting $\mathbf{J}_{\boldsymbol{\theta}_{t-1}}$. Instead, we take the straight-forward route and use the pseudoinverse of $\mathbf{J}_{\boldsymbol{\theta}_{t-1}}$, such that

$$p_{\text{tang}}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \propto \mathcal{N}\left(\boldsymbol{\theta}_t \middle| \boldsymbol{\theta}_{t-1}, \mathbf{J}^{\dagger}_{\boldsymbol{\theta}_{t-1}} \Sigma_{\mathcal{T}} \left(\mathbf{J}^{\dagger}_{\boldsymbol{\theta}_{t-1}}\right)^T\right) \mathcal{U}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t) \ , \tag{8}$$

where $\mathbf{J}^{\dagger}_{\boldsymbol{\theta}_{t-1}} = (\mathbf{J}^T_{\boldsymbol{\theta}_{t-1}} \mathbf{J}_{\boldsymbol{\theta}_{t-1}})^{-1} \mathbf{J}^T_{\boldsymbol{\theta}_{t-1}}$ denotes the pseudoinverse of $\mathbf{J}_{\boldsymbol{\theta}_{t-1}}$. If we consider $\mathbf{J}_{\boldsymbol{\theta}_{t-1}}$ a function from $\mathbb{T}^N$ to $\mathcal{T}$ then $\mathbf{J}^{\dagger}_{\boldsymbol{\theta}_{t-1}}$ is indeed the inverse of this function. One interpretation of this prior is that it is the normal distribution in angle space that provides the best linear approximation of a given normal distribution in the spatial domain.

To sample from this distribution, we generate a sample $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathcal{T}})$. This is then moved into the joint angle space by letting $\boldsymbol{\theta}_t = \left(\mathbf{J}^{\dagger}_{\boldsymbol{\theta}_{t-1}}\right)^T \boldsymbol{x} + \boldsymbol{\theta}_{t-1}$. In order to respect joint limits, we truncate joint values that exceeds their limitations. This simple scheme works well in practice.

In Fig. 2c we show 50 samples from this distribution, where $\Sigma_{\mathcal{T}}$ is the same as the projected prior in Fig. 2b. As can be seen, this prior behaves somewhat more variant than p_{proj}, but far less than p_{angle}.

4 Visual Measurements

To actually implement an articulated tracker, we need a system for making visual measurements, i.e. a likelihood model. To keep the paper focused on prediction, we use a simple likelihood model based on a consumer stereo camera[1]. This camera provides a dense set of three dimensional points $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K\}$ in each frame. The objective of the likelihood model then becomes to measure how well a pose hypothesis matches the points. We assume that each point is independent and that the distance between a point and the skin of the human follows a zero-mean Gaussian distribution, i.e.

$$p(\mathbf{Z}|\boldsymbol{\theta}_t) \propto \prod_{k=1}^{K} \exp\left(-\frac{D^2(\boldsymbol{\theta}_t, \boldsymbol{z}_k)}{2\sigma^2}\right) \quad , \tag{9}$$

where $D^2(\boldsymbol{\theta}_t, \boldsymbol{z}_k)$ denotes the square distance between the point $\boldsymbol{z}_k$ and the skin of the pose $\boldsymbol{\theta}_t$. To make the model robust with respect to outliers in the data we threshold the distance function D such that it never exceeds a given threshold.

We also need to define the skin of a pose, such that we can compute distances between this and a data point. Here, we define the skin of a bone as a cylinder with main axis corresponding to the bone itself. Since we only have a single view point, we discard the half of the cylinder that is not visible. The skin of the entire pose is then defined as the union of these half-cylinders. The distance between a point and this skin can then be computed as the smallest distance from the point to any of the half-cylinders.

5 Experimental Results

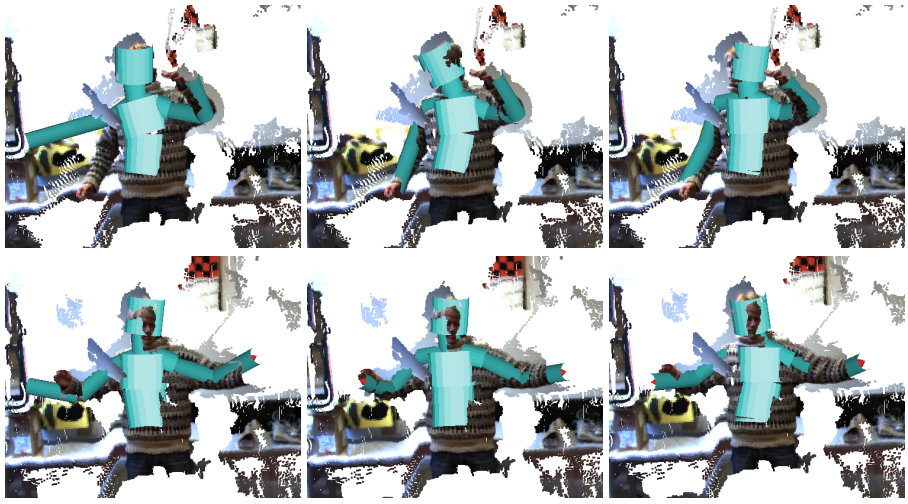
To build an articulated tracker we combine the likelihood model with the suggested priors using a particle filter. This provides us with a set of weighted samples from which we estimate the current pose as the weighted average.

We seek to compare the three suggested priors, p_{angle}, p_{proj} and p_{tang}. As the base of our comparison, we estimate the pose in each frame of a sequence using a particle filter with 10.000 samples, which is plenty to provide a good estimate. This will then serve as our ground truth data. As we are studying a general purpose motion model, we assume that each prior has a diagonal covariance structure. These variances are then learned from the ground truth data to give each prior the best possible working conditions.

We apply the three prior models to a sequence where a person is standing in place and mostly moving his arms. We vary the number of particles in the

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

three tracking systems between 25 and 1500. The results are available as videos on-line[2] and some selected frames are available in Fig. 3. The general tendency is that the projected prior provides the most accurate and smooth results for a given number of particles. Next, we seek to quantify this observation.



(a) p_{angle} **(b)** p_{proj} **(c)** $p_{\mathcal{T}}$

Fig. 3. Results attained using 150 and 250 samples superimposed on the image data. Top row is using 150 particles, while bottom row is using 250 particles. (a) Using the angular prior. (b) Using the projected prior. (c) Using the tangent space prior.

To compare the attained results to the ground truth data, we apply a simple spatial error measure [16, 20]. This measures the average distance between limb end points in the attained results and the ground truth data. This measure is then averaged across frames, such that the error measure becomes

$$\mathcal{E} = \frac{1}{TL} \sum_{t=1}^{T} \sum_{l=1}^{L} ||\boldsymbol{a}_{lt} - \boldsymbol{a}'_{lt}|| \ , \tag{10}$$

where $\boldsymbol{a}_{lt}$ is the spatial end point of the l^{th} limb at time t in the attained results, and $\boldsymbol{a}'_{lt}$ is the same point in the ground truth data. This measure is reported for the different priors in Fig. 4a. As can be seen, the projected prior is consistently better than the tangent space prior, which in turn is consistently better than the angular prior. One explanation of why the projected prior outperforms the tangent space prior could be that $\mathcal{M}$ has substantial curvature. This explanation is also in tune with the findings of Sommer et.al [21].

If the observation density $p(\mathbf{Z}_t|\boldsymbol{\theta}_t)$ is noisy, the motion model acts as a smoothing filter. This can be of particular importance when observations are

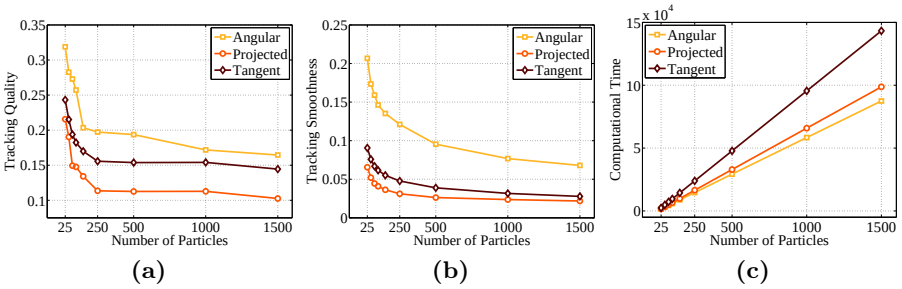
[2] `http://humim.org/eccv2010/`

missing, e.g. during self-occlusions. Thus, when evaluating the quality of a motion model it can be helpful to look at the smoothness of the attained pose sequence. To measure this, we simply compute the average size of the temporal gradient. We approximate this gradient using finite differences, and hence use

$$\mathcal{S} = \frac{1}{TL} \sum_{t=1}^{T} \sum_{l=1}^{L} ||\boldsymbol{a}_{lt} - \boldsymbol{a}_{l,t-1}|| \tag{11}$$

as a measure of smoothness. This is reported in Fig. 4b. It can be seen that the projected prior and the tangent space prior give pose sequences that are almost equally smooth; both being consistently much more smooth than the angular prior. This is also quite visible in the on-line videos.

So far we have seen that both suggested priors outperform the angular prior in terms of quality. The suggested priors are, however, computationally more demanding. One should therefore ask if it is computationally less expensive to simply increase the number of particles while using the angular prior. In Fig. 4c we report the running time of the tracking systems using the different priors. As can be seen, the projected prior is only slightly more expensive than the angular prior, whereas the tangent space prior is somewhat more expensive than the two other models. The latter result is somewhat surprising given the simplicity of the tangent space prior; we believe that this is caused by choices of numerical methods. In practice both of the suggested priors give better results than the angular prior at a fixed amount of computational resources, where the projected prior is consistently the best.

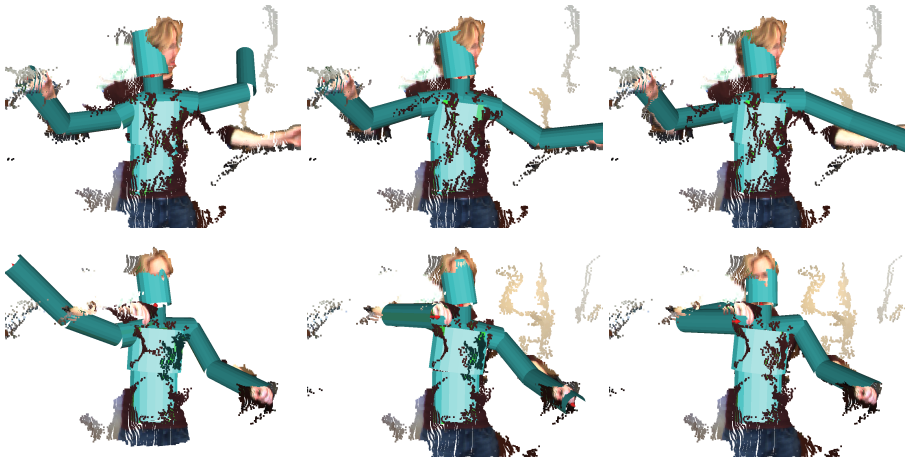


(a) **(b)** **(c)**

Fig. 4. Performance of the three priors. All reported numbers are averaged over several trials. (a) The error measure $\mathcal{E}$ as a function of the number of particles. The average standard deviation of $\mathcal{E}$ with respect to the trials are 0.018 for the angular prior, 0.008 for the projected prior and 0.009 for the tangent space prior. (b) The smoothness measure $\mathcal{S}$ as a function of the number of particles. The average standard deviation of $\mathcal{S}$ with respect to the trials are 0.0028 for the angular prior, 0.0009 for the projected prior and 0.0016 for the tangent space prior. (c) The computational time as a function of the number of particles.

We now repeat the experiment for a second sequence, using the same parameters as before. In Fig. 5 we show the tracking results in selected frames for the

three discussed priors. As before, videos are available on-line[2]. Essentially, we make the same observations as before: the projected prior provides the best and most smooth results, followed by the tangent space prior with the angular prior consistently giving the worst results. This can also be seen in Fig. 6 where the error and smoothness measures are plotted along with the running time of the methods. Again, we see that for a given amount of computational resources, the projected prior consistently provides the best results.



(a) p_{angle} **(b)** p_{proj} **(c)** $p_{\mathcal{T}}$

Fig. 5. Results attained using 150 and 250 samples superimposed on the image data. Top row is using 150 particles, while bottom row is using 250 particles. (a) Using the angular prior. (b) Using the projected prior. (c) Using the tangent space prior.

6 Discussion

We have presented an analysis of the commonly used prior which assumes Gaussian distributed joint angles, and have shown that this behaves less than desirable spatially. Specifically, we have analysed the covariance of this prior in the tangent space of the pose manifold. This has clearly illustrated that small changes in a joint angle can lead to large spatial changes. Since this instability is ill-suited for predicting articulated motion, we have suggested to define the prior directly in spatial coordinates.

Since human motion is restricted to a manifold $\mathcal{M} \subset \mathbb{R}^{3L}$, we, however, need to define the prior in this domain. We have suggested two means of accomplishing this goal. One builds the prior by projecting onto the manifold and one builds the prior in the tangent space of the manifold. Both solutions have shown to outperform the ordinary angular prior in terms of both speed and accuracy. Of the two suggested priors, the projected prior seems to outperform the tangent

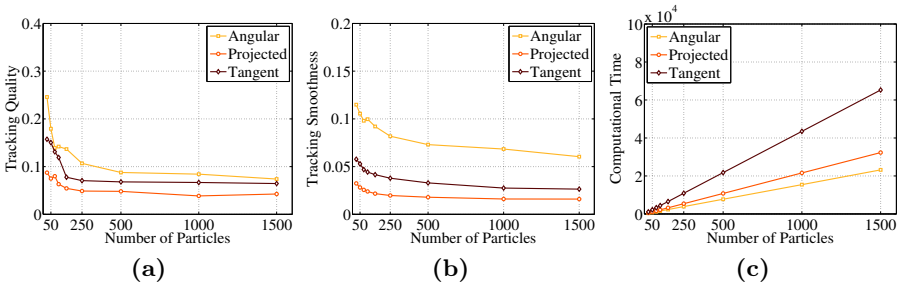


Fig. 6. Performance of the three priors. All reported numbers are averaged over several trials. (a) The error measure $\mathcal{E}$ as a function of the number of particles. The average standard deviation of $\mathcal{E}$ with respect to the trials are 0.021 for the angular prior, 0.007 for the projected prior and 0.015 for the tangent space prior. (b) The smoothness measure $\mathcal{S}$ as a function of the number of particles. The average standard deviation of $\mathcal{S}$ with respect to the trials are 0.002 for the angular prior, 0.0004 for the projected prior and 0.001 for the tangent space prior. (c) The computational time as a function of the number of particles.

space prior, both in terms of speed and quality. The tangent space prior does, however, have the advantage of simply being a normal distribution in joint angle space, which can make it more suitable as a prior when learning a motion specific model.

One advantage with building motion models spatially is that we can express motion specific knowledge quite simply. As an example, one can model a person standing in place simply by reducing the variance of the persons feet. This type of knowledge is non-trivial to include in models expressed in terms of joint angles.

The suggested priors can be interpreted as computationally efficient approximations of a Brownian motion on $\mathcal{M}$. We therefore find it interesting to investigate this connection further along with similar stochastic process models restricted to manifolds. In the future, we will also use the suggested priors as building blocks in more sophisticated motion specific models.

References

1. Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding **108** (2007) 4–18
2. Cappé, O., Godsill, S.J., Moulines, E.: An overview of existing methods and recent advances in sequential Monte Carlo. Proceedings of the IEEE **95** (2007) 899–924
3. Erleben, K., Sporring, J., Henriksen, K., Dohlmann, H.: Physics Based Animation. Charles River Media (2005)
4. Zhao, J., Badler, N.I.: Inverse kinematics positioning using nonlinear programming for highly articulated figures. ACM Transaction on Graphics **13** (1994) 313–336
5. Nocedal, J., Wright, S.J.: Numerical optimization. Springer Series in Operations Research. Springer-Verlag (1999)
6. Brubaker, M.A., Fleet, D.J., Hertzmann, A.: Physics-based person tracking using the anthropomorphic walker. Int. J. of Comp. Vis. **87** (2010) 140–155

7. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian Process Dynamical Models for Human Motion. Pattern Analysis and Machine Intelligence **30** (2008) 283–298
8. Sminchisescu, C., Jepson, A.: Generative modeling for continuous non-linearly embedded visual inference. In: ICML '04: Proceedings of the twenty-first international conference on Machine learning, ACM (2004) 759–766
9. Lu, Z., Carreira-Perpinan, M., Sminchisescu, C.: People Tracking with the Laplacian Eigenmaps Latent Variable Model. In Platt, J., et al., eds.: Advances in Neural Inf. Proc. Systems. Volume 20. MIT Press (2008) 1705–1712
10. Sidenbladh, H., Black, M.J., Fleet, D.J.: Stochastic tracking of 3d human figures using 2d image motion. In: Proceedings of ECCV'00. Volume II of Lecture Notes in Computer Science 1843., Springer (2000) 702–718
11. Elgammal, A.M., Lee, C.S.: Tracking People on a Torus. IEEE Transaction on Pattern Analysis and Machine Intelligence **31** (2009) 520–538
12. Urtasun, R., Fleet, D.J., Fua, P.: 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. (2006) 238–245
13. Carreira-Perpinan, M.A., Lu, Z.: The Laplacian Eigenmaps Latent Variable Model. JMLR W&P **2** (2007) 59–66
14. Urtasun, R., Fleet, D.J., Hertzmann, A., Fua, P.: Priors for people tracking from small training sets. In: Int. Conf. on Comp. Vis. Volume 1. (2005) 403–410
15. Grochow, K., Martin, S.L., Hertzmann, A., Popović, Z.: Style-based inverse kinematics. ACM Transaction on Graphics **23** (2004) 522–531
16. Balan, A.O., Sigal, L., Black, M.J.: A quantitative evaluation of video-based 3d person tracking. Visual Surveillance and Performance Evaluation of Tracking and Surveillance **0** (2005) 349–356
17. Bandouch, J., Engstler, F., Beetz, M.: Accurate human motion capture using an ergonomics-based anthropometric human model. In: Proc. of the 5th int. conf. on Articulated Motion and Deformable Objects, Springer (2008) 248–258
18. Pennec, X.: Probabilities and statistics on riemannian manifolds: Basic tools for geometric measurements. In: NSIP. (1999) 194–198
19. MacCormick, J., Isard, M.: Partitioned sampling, articulated objects, and interface-quality hand tracking. In: ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part II, Springer-Verlag (2000) 3–19
20. Sigal, L., Bhatia, S., Roth, S., Black, M.J., Isard, M.: Tracking loose-limbed people. In: CVPR '04: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 1. (2004) 421–428
21. Sommer, S., Lauze, F., Hauberg, S., Nielsen, M.: Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations. In: ECCV 2010: Proceedings of the 1th European Conference on Computer Vision, Springer-Verlag (2010)

Paper 3: Stick It! Articulated Tracking using Spatial Rigid Object Priors

Authors: Søren Hauberg and Kim S. Pedersen.
Status: Published at the *Asian Conference on Computer Vision* 2010 [19].

Shortly after the previous paper was accepted, I stumbled upon a preprint of a paper that was accepted at the *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) 2010. The paper, *Tracking people interacting with objects* [30], was studying how tracking could be improved by incorporating knowledge of the surrounding environment. The authors suggested using rejection sampling in joint angle space to enforce constraints on hand positions, i.e. to solve an inverse kinematics problem. This seemed like a poor solution to the problem and we set out to show that our spatial model was more suitable.

The resulting model is remarkably simple, yet it outperforms state-of-the-art in both speed and accuracy. As we are modelling joint positions, our strategy is simply to move joints that are interacting with the environment to the point of interaction; remaining joints are treated as in the previous paper.

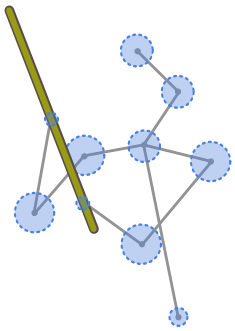


Figure 4.1 *An illustration of the spatial interaction model; blue circles indicate covariances. When a joint is known to be interacting with the environment, its mean is set to to the point of interaction.*

Stick It! Articulated Tracking using Spatial Rigid Object Priors

Søren Hauberg and Kim Steenstrup Pedersen

{hauberg, kimstp}@diku.dk,
The eScience Centre, Dept. of Computer Science, University of Copenhagen

Abstract. Articulated tracking of humans is a well-studied field, but most work has treated the humans as being independent of the environment. Recently, Kjellström et al. [1] showed how knowledge of interaction with a known rigid object provides constraints that lower the degrees of freedom in the model. While the phrased problem is interesting, the resulting algorithm is computationally too demanding to be of practical use. We present a simple and elegant model for describing this problem. The resulting algorithm is computationally much more efficient, while it at the same time produces superior results.

1 Introduction

Three dimensional articulated human motion tracking is the process of estimating the configuration of body parts over time from sensor input [2]. A large body of work have gone into solving this problem by using computer vision techniques without resorting to visual markers. The bulk of this work, however, completely ignores that almost all human movement somehow involves interaction with a rigid environment (people sit on *chairs*, walk on the *ground*, lift the *bottle* and so forth). By incorporating this fact of life, one can take advantage of the constraints provided by the environment, which effectively makes the problem easier to solve.

Recently, Kjellström et al. [1] showed that taking advantage of these constraints allows for improved tracking quality. To incorporate the constraints Kjellström et al., however, had to resort to a highly inefficient rejection sampling scheme. In this paper, we present a detailed analysis of this work and show how the problem can be solved in an elegant and computationally efficient manner. First we will, however, review the general articulated tracking framework and related work.

1.1 Articulated Tracking

Estimating the pose of a person using a single view point or a small baseline stereo camera is an inherently difficult problem due to self-occlusions. This manifests itself in that the distribution of the human pose is multi-modal with an unknown number of modes. Currently, the best method for coping with such

distributions is the particle filter [3]. This aims at estimating the state of the system, which is represented as a set of weighted samples. These samples are propagated in time using a predictive model and assigned a weight according to a data likelihood. As such, the particle filter requires two subsystems: one for computing likelihoods by comparing the image data to a sample from the hidden state distribution, and one for predicting future states. In practice, the predictive system is essential in making the particle filter computationally feasible, as it can drastically reduce the number of needed samples. As an example, we shall later see how the predictive system can be phrased to incorporate constraints from the environment.

For the particle filter to work, we need a representation of the system state, which in our case is the human pose. As is common [2], we shall use the kinematic skeleton (see Fig. 1). 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 model the bones as having known constant length (i.e. rigid), so the direction of each bone constitute the only degrees of freedom in the kinematic skeleton. The direction in each joint can be parametrised with a vector of angles, noticing that different joints may have different number of degrees of freedom. We may collect all joint angle vectors into one large vector $\boldsymbol{\theta}_t$ representing all joint angles in the model at time t. The objective of the particle filter, thus, becomes to estimate $\boldsymbol{\theta}_t$ in each frame.

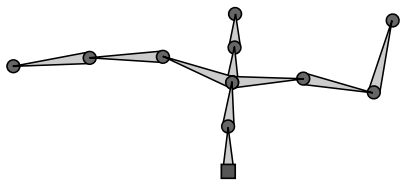


Fig. 1. An illustration of the kinematic skeleton. Circles correspond to the spatial bone end points and the square corresponds to the root.

To represent the fact that bones cannot move freely (e.g. the elbow joint can only bend between 0 and 120 degrees), we restrict $\boldsymbol{\theta}_t$ to a subset $\boldsymbol{\Theta}$ of $\mathbb{R}^N$. In practice, $\boldsymbol{\Theta}$ is chosen such that each joint angle is restricted to an interval. This is often called box constraints [4].

From known bone lengths and a joint angle vector $\boldsymbol{\theta}_t$, it is straight-forward to compute the spatial coordinates of the bones. The root of the kinematic tree is placed at the origin of the coordinate system. The end point of the next bone along a branch in the tree is then computed by rotating the coordinate system and translating the root along a fixed axis relative to the parent bone. The rotation is parametrised by the angles of the joint in question and the length of the translation corresponds to the known length of the bone. We can repeat

this process recursively until the entire kinematic tree has been traversed. This process is known as Forward Kinematics [5].

1.2 Related Work

Most work in the articulated tracking literature falls in two categories. Either the focus is on improving the vision system or on improving the predictive system. Due to space constraints, we forgo a review of various vision systems as this paper is focused on prediction. For an overview of vision systems, see the review paper by Poppe [2].

Most work on improving the predictive system, is focused on learning motion specific priors, such as for *walking* [6–12]. Currently, the most popular approach is to restrict the tracker to some subspace of the joint angle space [7–10, 13]. Such priors are, however, action specific. When no action specific knowledge is available it is common [1, 10, 14, 15] to simply let $\boldsymbol{\theta}_t$ follow a normal distribution with a diagonal covariance, i.e.

$$p_{\text{gp}}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \propto \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \text{diag})\ \mathcal{U}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t)\ , \tag{1}$$

where $\mathcal{U}_{\boldsymbol{\Theta}}$ is a uniform distribution on the legal set of angles that encodes the joint constraints. Recently, Hauberg et al. [16] showed that this model causes the spatial variance of the bone end points to increase as the kinematic chains are traversed. In practice this means that with this model the spatial variance of e.g. the hands is always larger than of the shoulders. We will briefly review a solution to this problem suggested by Hauberg et al. in Sec. 1.3, as it provides us a convenient framework for modelling interaction with the environment.

In general, as above, the environment is usually not incorporated in the tracking models. One notable environmental exception seems to be the ground plane [6, 17]. Yamamoto and Yagishita [17] use a linear approximation of the motion path by linearising the forward kinematics function. As this is a highly non-linear function and motion paths in general are non-linear this modelling decision seems to be made out of sheer practicality. Promising results are, however, shown on constrained situations, such as when the position and orientation of a persons feet is known. Brubaker et al. [6] explicitly model the ground plane in a biomechanical model of walking. Their approach is, however, limited to interaction with the ground while walking.

Of particular importance to our work, is the paper by Kjellström et al. [1]. We will therefore review this in detail in Sec. 2.

1.3 Projected Spatial Priors

Recently, an issue with the standard general purpose prior from Eq. 1 was pointed out by Hauberg et al. [16]. Due to the tree structure of the kinematic skeleton, the spatial variance of bone end point increase as the kinematic chains are traversed. To avoid this somewhat arbitrary behaviour it was suggested to build the prior distribution directly in the spatial domain.

To define a predictive distribution in the spatial domain, Hauberg et al. first define a representation manifold $\mathcal{M} \in \mathbb{R}^{3L}$, where L denotes the number of bones. A point on this manifold corresponds to all spatial bone end points of a pose parametrised by a set of joint angles. More stringent, $\mathcal{M}$ can be defined as

$$\mathcal{M} = \{F(\boldsymbol{\theta}) | \boldsymbol{\theta} \in \boldsymbol{\Theta}\} \ , \tag{2}$$

where F denotes the forward kinematics function for the entire skeleton.

Once this manifold is defined, a Gaussian-like distribution can be defined simply by projecting a Gaussian distribution in $\mathbb{R}^{3L}$ onto $\mathcal{M}$, i.e.

$$p_{\text{proj}}(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}) = \text{proj}_{\mathcal{M}} \left[\mathcal{N}(F(\boldsymbol{\theta}_t) | F(\boldsymbol{\theta}_{t-1}), \Sigma) \right] \ . \tag{3}$$

When using a particle filter for tracking, one only needs to be able to draw samples from the prior model. This can easily be done by sampling from the normal distribution in $\mathbb{R}^{3L}$ and projecting the result onto $\mathcal{M}$. This, however, requires an algorithm for performing the projection. This is done by seeking

$$\hat{\boldsymbol{\theta}}_t = \min_{\boldsymbol{\theta}_t} \left|\left| \boldsymbol{x}_t - F(\boldsymbol{\theta}_t) \right|\right|^2 \qquad \text{s.t.} \qquad \boldsymbol{\theta}_t \in \boldsymbol{\Theta} \ , \tag{4}$$

where $\boldsymbol{x}_t$ denotes a sample from the normal distribution in $\mathbb{R}^{3L}$. This is an overdetermined constrained non-linear least-squares problem, that can be solved by any off-the-shelf optimisation algorithm [4]. We shall later see that the spatial nature of this prior is very helpful when designing priors that take the environment into account.

2 The KKB Tracker

Kjellström et al. [1] consider the situation where a person is holding on to a stick. It is assumed that the 3D position of the stick is known in each frame. In practice they track the stick using 8 calibrated cameras. They define the stick as

$$\text{stick}(\gamma_t) = \gamma_t \boldsymbol{a} + (1 - \gamma_t)\boldsymbol{b}, \quad \gamma_t \in [0, 1] \ , \tag{5}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the end points of the stick.

The state is extended with a γ_t for each hand, which encodes the position of the respective hand on the stick. The state, thus, contains $\boldsymbol{\theta}_t$, $\gamma_t^{(\text{left})}$ and $\gamma_t^{(\text{right})}$. The goal is then to find an algorithm where the hand positions implied by $\boldsymbol{\theta}_t$ corresponds to the hand positions expressed by the γ_t's.

Kjellström et al. take a rejection sampling approach for solving this problem. They sample $\boldsymbol{\theta}_t$ from Eq. 1 and compute the attained hand positions using forward kinematics. They then keep generating new samples until the attained hand positions are within a given distance of the hand positions encoded by the γ_t's. Specifically, they keep generating new $\boldsymbol{\theta}_t$'s until

$$\|F_{\text{left}}(\boldsymbol{\theta}_t) - \text{stick}(\gamma_t^{(left)})\| < T_E \quad \text{and} \quad \|F_{\text{right}}(\boldsymbol{\theta}_t) - \text{stick}(\gamma_t^{(right)})\| < T_E \ , \tag{6}$$

where F_{left} is the forward kinematics function that computes the position of the left hand, F_{right} is the equivalent for the right hand and T_E is a threshold. We will denote this prior p_{kkb}, after the last names of its creators.

The γ_t's are also propagated in time to allow for sliding the hands along the stick. Specifically, Kjellström et al. let

$$p\left(\gamma_t^{(\text{left})}|\gamma_{t-1}^{(\text{left})}\right) \propto \mathcal{N}\left(\gamma_t^{(\text{left})}|\gamma_{t-1}^{(\text{left})}, \sigma^2\right)\ \mathcal{U}_{[0,1]}\left(\gamma_t^{(\text{left})}\right)\ , \tag{7}$$

where $\mathcal{U}_{[0,1]}$ is the uniform distribution on $[0, 1]$. $\gamma_t^{(\text{right})}$ is treated the same way.

The advantage of this approach is that it actually works; successful tracking was reported in [1] and in our experience decent results can be attained with relatively few particles. Due to the rejection sampling, the approach is, however, computationally very demanding (see Sec. 5, in particular Fig. 4). The approach also has a limit on how many constraints can be encoded in the prior, as more constraints yield smaller acceptance regions. Thus, the stronger the constraints, the longer the running time. Furthermore, the rejection sampling has the side effect that the time it takes to predict one sample is not constant. In parallel implementations of the particle filter, such behaviour causes thread divergence, which drastically lessens the gain of using a parallel implementation.

3 Spatial Object Interaction Prior

We consider the same basic problem as Kjellström et al. [1], that is, assume we know the position of a stick in 3D and assume we know the person is holding on to the stick. As Kjellström et al., we extend the state with a γ_t for each hand that encodes where on the stick the hands are positioned using the model stated in Eq. 5. As before these are propagated in time using Eq. 7.

Following the idea of Hauberg et al. [16], we then define a motion prior in the spatial domain. Intuitively, we let each bone end point, except the hands, follow a normal distribution with the current bone end point as the mean value. The hands are, however, set to follow a normal distribution with a mean value corresponding to the hand position implied by $\gamma_t^{(\text{left})}$ and $\gamma_t^{(\text{right})}$. The resulting distribution is then projected back on the manifold $\mathcal{M}$ of possible poses, such that the final motion prior is given by

$$p_{\text{stick3d}}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \text{proj}_{\mathcal{M}}\left[\mathcal{N}(F(\boldsymbol{\theta}_t)|\boldsymbol{\mu}, \Sigma)\right]\ , \tag{8}$$

where $\boldsymbol{\mu}$ indicates the just mentioned mean value. Samples can then be drawn from this distribution as described in Sec. 1.3.

3.1 Two Dimensional Object Information

When we defined p_{stick3d} we assumed we knew the three dimensional position of the stick. In the experiments presented in Sec. 5, we are using an active motion capture system to attain this information. While this approach might be feasible

in laboratory settings it will not work in the general single-viewpoint setup; in practice it is simply too hard to accurately track even a rigid object in 3D. It is, however, not that difficult to track a stick in 2D directly in the image. We, thus, suggest a trivial extension of p_{stick3d} to the case were we only know the 2D image position of the stick.

From the 2D stick position in the image and the value of $\gamma_t^{(left)}$ we can compute the 2D image position of the left hand. We then know that the actual hand position in 3D must lie on the line going through the optical centre and the 2D image position. We then define the mean value of the predicted left hand as the projection of the current left hand 3D position onto the line of possible hand positions. The right hand is treated similarly. This is sketched in Fig. 2. The mean value of the remaining end point is set to their current position, and the resulting distribution is projected onto $\mathcal{M}$. We shall denote this motion prior p_{stick2d}.

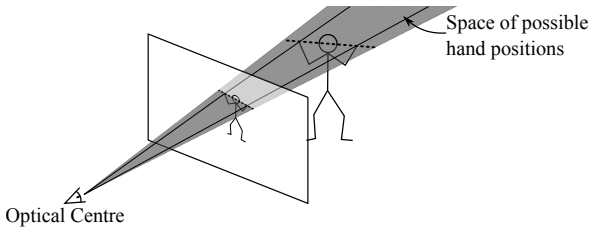


Fig. 2. An illustration of the geometry behind the p_{stick2d} model. The stick is detected in the image and the hands are restricted to the part of $\mathbb{R}^3$ that projects onto the detected stick.

4 Visual Measurements

To actually implement an articulated tracker, we need a system for making visual measurements. To keep the paper focused on prediction, we use a simple vision system [16] based on a consumer stereo camera[1]. This camera provides a dense set of three dimensional points $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K\}$ in each frame. The objective of the vision system then becomes to measure how well a pose hypothesis matches the points. We assume that points are independent and that the distance between a point and the skin of the human follows a zero-mean Gaussian distribution, i.e.

$$p(\mathbf{Z}|\boldsymbol{\theta}_t) \propto \prod_{k=1}^{K} \exp\left(-\frac{\min\left[D^2(\boldsymbol{\theta}_t, \boldsymbol{z}_k),\ \tau\right]}{2\sigma^2}\right) , \tag{9}$$

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

where $D^2(\boldsymbol{\theta}_t, \boldsymbol{z}_k)$ denotes the squared distance between the point $\boldsymbol{z}_k$ and the skin of the pose $\boldsymbol{\theta}_t$ and τ is a constant threshold. The minimum operation is there to make the system robust with respect to outliers.

We also need to define the skin of a pose, such that we can compute distances between this and a data point. Here, we define the skin of a bone as a capsule with main axis corresponding to the bone itself. Since we only have a single view point, we discard the half of the capsule that is not visible. The skin of the entire pose is then defined as the union of these half-capsules. The distance between a point and this skin can then be computed as the smallest distance from the point to any of the half-capsules.

5 Experimental Results

Using the just mentioned likelihood model we can create an articulated tracker for each suggested prior. This gives us a set of weighted samples at each time step, which we reduce to one pose estimate $\hat{\boldsymbol{\theta}}_t$ by computing the weighted average.

We record images from the previously mentioned stereo camera at 15 FPS along with synchronised data from an optical motion capture system[2]. We place motion capture markers on a stick such that we can attain its three dimensional position in each frame. In the case of $p_{stick2d}$, we only use the marker positions projected into the image plane.

To evaluate the quality of the attained results we also position motion capture markers on the arms of the test subject. We then measure the average distance between the motion capture markers and the capsule skin of the attained results. This measure is then averaged across frames, such that the error measure becomes

$$\mathcal{E} = \frac{1}{TM} \sum_{t=1}^{T} \sum_{m=1}^{M} D(\hat{\boldsymbol{\theta}}_t, \boldsymbol{v}_m) \ , \tag{10}$$

where $D(\hat{\boldsymbol{\theta}}_t, \boldsymbol{v}_m)$ is the Euclidean distance between the m^{th} motion capture marker and the skin at time t.

In the first sequence we study a person who moves the stick from side to side and finally move the stick behind his head. This type of motion utilises the shoulder joints a lot, which is typically something that can cause difficulties for articulated trackers. We show selected frames from this sequence with the estimated pose superimposed in Fig. 3. Results are shown for the three different priors that utilise knowledge of the stick position. For reference, we also show the result of the standard model p_{gp} that assumes independent normally distributed joint angles. In all cases, 500 particles was used. As can be seen, the three stick-based priors all track the motion successfully, whereas the general purpose prior fail. This is more evident in the videos, which are available online[3].

[2] `http://www.phasespace.com/`

[3] `http://humim.org/accv2010`

To quantify the quality of the results, we compute the error measure from Eq. 10 for each of the attained results. This is reported along with the computation time in Table 1. As can be read, p_{stick3d} gives the most accurate results, closely followed by p_{kkb} and p_{stick2d}. However, when it comes to computation speed, we note that the p_{kkb} prior is 7.2 times slower than the general purpose angular prior, whereas our priors are both only 1.1 times slower.

Upon further study of the results attained by the p_{kkb} prior we note that in a few frames the pose estimate does not actually grab onto the stick. To understand this phenomena, we need to look at the details of the rejection sampling scheme. If we keep rejecting samples until Eq. 6 is satisfied, we have no way of guaranteeing that the algorithm will ever terminate. To avoid infinite loops, we stop the rejection sampling after a maximum of 5000 rejections. We found this to be a reasonable compromise between running times and accuracy. In Fig. 4a we plot the percentage of particles meeting the maximum number of rejections in each frame. As can be seen this number fluctuates and even reaches 100 percent in a few frames. This behaviour causes shaky pose estimates and even a few frames where the knowledge of the stick position is effectively not utilised. This can also be seen in Fig. 5 where the generated particles are shown for the different priors. Videos showing these are also available online[3]. Here we see that the p_{kkb} prior generates several particles with hand positions far away from the stick. We do not see such a behaviour of neither the $p_{stick3d}$ nor $p_{stick2d}$ priors.

We move on to the next studied sequence. Here the person is waiving the stick in a sword-fighting-manner. A few frames from the sequence with results superimposed are available in Fig. 6. While $p_{stick3d}$ and $p_{stick2d}$ are both able to successfully track the motion, p_{kkb} fails in several frames. As before, the reason for this behaviour can be found in the rejection sampling scheme. In Fig. 4b we show the percentage of particles reaching the maximum number of rejections. As before, we see that a large percentage of the particles often reach the limit and as such fail to take advantage of the known stick position. This is the reason for the erratic behaviour. In Table 2 we show accuracy and running time of the different methods, and here it is also clear that the p_{kkb} prior fails to track the motion even if it spends almost 10 times more time per frame than $p_{stick3d}$ and $p_{stick2d}$.

Table 1. Results for the first sequence using 500 particles.

Prior	Error (std.)	Computation Time
p_{kkb}	2.7 cm (1.3 cm)	687 sec./frame
p_{stick3d}	2.4 cm (1.0 cm)	108 sec./frame
p_{stick2d}	2.9 cm (1.5 cm)	108 sec./frame
p_{gp}	4.2 cm (2.3 cm)	96 sec./frame

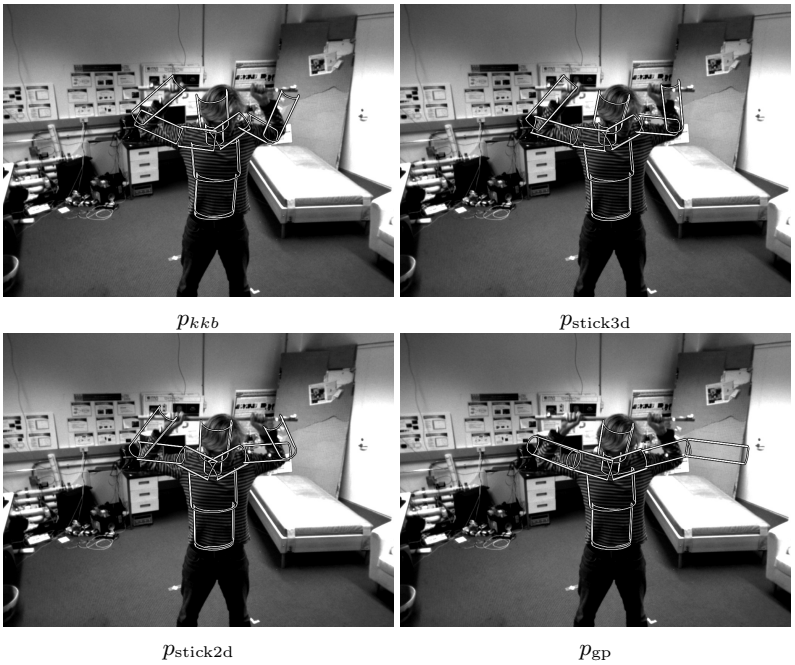
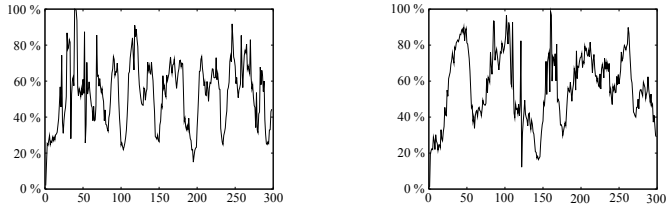


Fig. 3. Frame 182 from the first sequence. Image contrast has been enhanced for viewing purposes.



(a) First sequence (b) Second sequence

Fig. 4. Percentage of particles which reached the limit of the rejection sampling.

Table 2. Results for the second sequence using 500 particles.

Prior	Error (std.)	Computation Time
p_{kkb}	8.4 cm (1.9 cm)	782 sec./frame
p_{stick3d}	2.2 cm (0.8 cm)	80 sec./frame
p_{stick2d}	2.8 cm (1.7 cm)	80 sec./frame
p_{gp}	8.4 cm (2.2 cm)	68 sec./frame

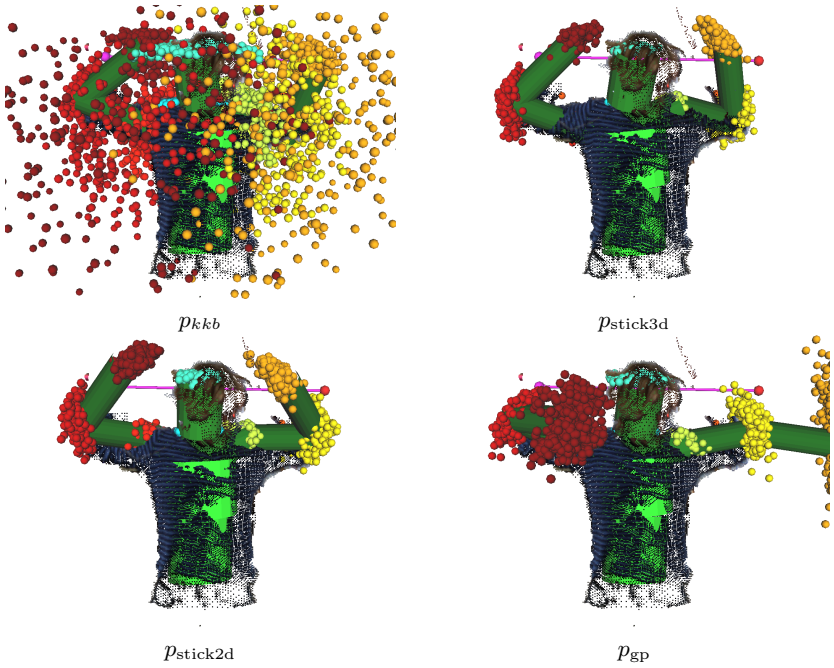


Fig. 5. The particles active in frame 182 in the first sequence.

6 Discussion

In this paper we have analysed an algorithm suggested by Kjellström et al. for articulated tracking when environmental constraints are available. We argued, and experimentally validated, that the algorithm is computationally too demanding to be of use in real-life settings. We then presented a simple model for solving the same problem, that only comes with a small computational overhead. The simplicity of our method comes from the decision to model the motion spatially rather than in terms of joint angles. This provides us with a general framework in which spatial knowledge can trivially be utilised. As most environmental knowledge is available in this domain, the idea can easily be extended to more complex situations.

In practice, much environmental information is not available in three dimensions, but can only be observed in the image plane. As such, we have suggested a straight-forward motion prior that only constraint limb positions in the image plane. This provides a framework that can actually be applied in real-life settings as it does not depend on three dimensional environmental knowledge that most often is only available in laboratory settings.

The two suggested priors are both quite simple and they encode the environmental knowledge in a straight-forward manner. The priors, thus, demonstrate

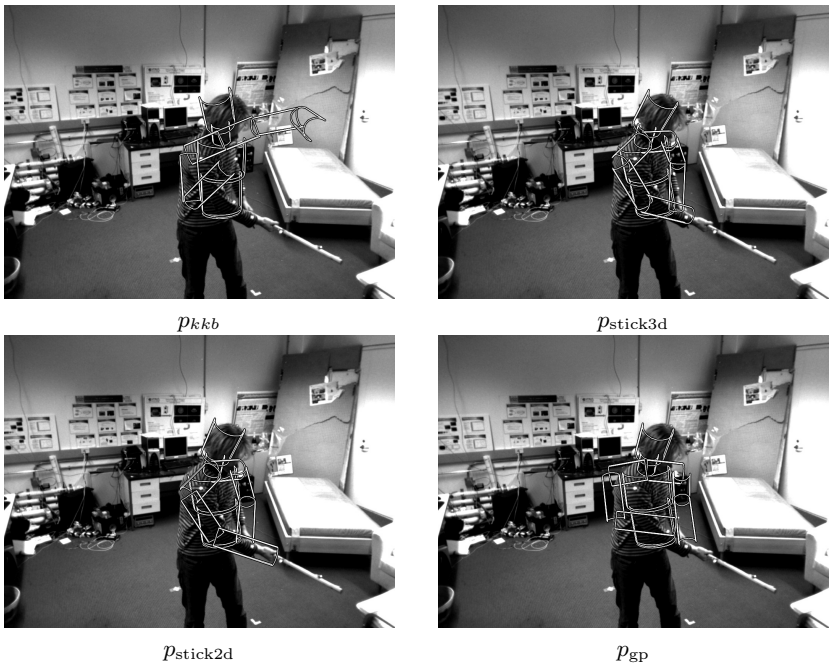


Fig. 6. Frame 101 from the second sequence. Image contrast has been enhanced for viewing purposes.

the ease of which complicated problems can be solved when the motion is modelled spatially rather than in terms of joint angles. As spatial models have been shown to have more well-behaved variance structure than models expressed in terms of joint angles [16], we do believe spatial models can provide the basis of the next leaps forward for articulated tracking.

References

1. Kjellström, H., Kragić, D., Black, M.J.: Tracking people interacting with objects. In: CVPR '10: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2010)
2. Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding **108** (2007) 4–18
3. Cappé, O., Godsill, S.J., Moulines, E.: An overview of existing methods and recent advances in sequential Monte Carlo. Proceedings of the IEEE **95** (2007) 899–924
4. Nocedal, J., Wright, S.J.: Numerical optimization. Springer Series in Operations Research. Springer-Verlag (1999)
5. Erleben, K., Sporring, J., Henriksen, K., Dohlmann, H.: Physics Based Animation. Charles River Media (2005)

6. Brubaker, M.A., Fleet, D.J., Hertzmann, A.: Physics-based person tracking using the anthropomorphic walker. International Journal of Computer Vision **87** (2010) 140–155
7. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian Process Dynamical Models for Human Motion. IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (2008) 283–298
8. Sminchisescu, C., Jepson, A.: Generative modeling for continuous non-linearly embedded visual inference. In: ICML '04: Proceedings of the twenty-first international conference on Machine learning, ACM (2004) 759–766
9. Lu, Z., Carreira-Perpinan, M., Sminchisescu, C.: People Tracking with the Laplacian Eigenmaps Latent Variable Model. In Platt, J., Koller, D., Singer, Y., Roweis, S., eds.: Advances in Neural Information Processing Systems 20. MIT Press (2008) 1705–1712
10. Sidenbladh, H., Black, M.J., Fleet, D.J.: Stochastic tracking of 3d human figures using 2d image motion. In: Proceedings of ECCV'00. Volume II of Lecture Notes in Computer Science 1843., Springer (2000) 702–718
11. Elgammal, A.M., Lee, C.S.: Tracking People on a Torus. IEEE Transaction on Pattern Analysis and Machine Intelligence **31** (2009) 520–538
12. Urtasun, R., Fleet, D.J., Fua, P.: 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. (2006) 238–245
13. Urtasun, R., Fleet, D.J., Hertzmann, A., Fua, P.: Priors for people tracking from small training sets. In: Tenth IEEE International Conference on Computer Vision. Volume 1. (2005) 403–410
14. Bandouch, J., Engstler, F., Beetz, M.: Accurate human motion capture using an ergonomics-based anthropometric human model. In: AMDO '08: Proceedings of the 5th international conference on Articulated Motion and Deformable Objects, Springer-Verlag (2008) 248–258
15. Balan, A.O., Sigal, L., Black, M.J.: A quantitative evaluation of video-based 3d person tracking. Visual Surveillance and Performance Evaluation of Tracking and Surveillance **0** (2005) 349–356
16. Hauberg, S., Sommer, S., Pedersen, K.S.: Gaussian-like spatial priors for articulated tracking. In Daniilidis, K., Maragos, P., , Paragios, N., eds.: ECCV 2010. Volume 6311 of Lecture Notes in Computer Science., Springer, Heidelberg (2010) 425–437
17. Yamamoto, M., Yagishita, K.: Scene constraints-aided tracking of human body. In: CVPR, Published by the IEEE Computer Society (2000) 151–156

Paper 4: Data-Driven Importance Distributions for Articulated Tracking

Authors: Søren Hauberg and Kim S. Pedersen.
Status: Published at the *Energy Minimization Methods in Computer Vision and Pattern Recognition* conference 2011 [20].

The papers so far have focused on variance properties of different priors, the geometry of the models and interactions with the environment. Oddly enough, the hidden agenda of the Riemannian approach was something entirely different. A fascinating aspect of particle filters is the idea of drawing samples near the modes of the likelihood via changes in the importance distribution. This approach has not seen a lot of attention in the articulated tracking literature because of the non-linear relationship between the space of observed images and the joint angle space. If we instead consider the joint position space, then this is more directly related to the space of observed images. This simplifies many parts of the optimisation and in particular allows for better importance distributions.

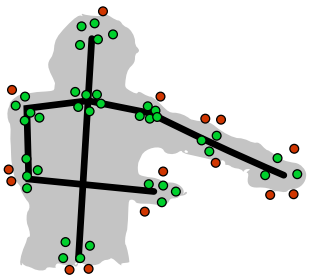


Figure 5.1 *A sketch of the simplest data-driven importance distribution: individual joint positions are sampled using rejection sampling to ensure that they are inside the observed silhouette.*

In the next paper we provide two examples of simple data-driven importance distributions that improve tracking quality substantially. These distributions are based on silhouette and depth data, respectively. These are, however, meant as examples of a general strategy; the major point of the paper is to show how easy it is to design data-driven importance distributions once the pose representation is spatial.

Data-Driven Importance Distributions for Articulated Tracking

Søren Hauberg and Kim Steenstrup Pedersen

{hauberg, kimstp}@diku.dk,
The eScience Centre, Dept. of Computer Science, University of Copenhagen

Abstract. We present two data-driven importance distributions for particle filter-based articulated tracking; one based on background subtraction, another on depth information. In order to keep the algorithms efficient, we represent human poses in terms of spatial joint positions. To ensure constant bone lengths, the joint positions are confined to a non-linear representation manifold embedded in a high-dimensional Euclidean space. We define the importance distributions in the embedding space and project them onto the representation manifold. The resulting importance distributions are used in a particle filter, where they improve both accuracy and efficiency of the tracker. In fact, they triple the effective number of samples compared to the most commonly used importance distribution at little extra computational cost.

Key words: Articulated tracking · Importance Distributions · Particle Filtering · Spatial Human Motion Models

1 Motivation

Articulated tracking is the process of estimating the pose of a person in each frame in an image sequence [1]. Often this is expressed in a Bayesian framework and subsequently the poses are inferred using a particle filter [1–11]. Such filters generate a set of sample hypotheses and assign them weights according to the likelihood of the observed data given the hypothesis is correct. Usually, the hypotheses are sampled directly from the motion prior as this vastly simplifies development. However, as the motion prior is inherently independent of the observed data, samples are generated completely oblivious to the current observation. This has the practical consequence that many sampled pose hypotheses are far away from the modes of the likelihood. This means that many samples are needed for accurate results. As the likelihood has to be evaluated for each of these samples, the resulting filter becomes computationally demanding.

One solution, is to sample hypotheses from a distribution that is not "blind" to the current observation. The particle filter allows for such *importance distributions*. While the design of good importance distributions can be the deciding point of a filter, not much attention has been given to their development in articulated tracking. The root of the problem is that the pose parameters are related to the observation in a highly non-linear fashion, which makes good importance distributions hard to design. In this paper, we change the pose parametrisation and then suggest a simple approximation that allows us to design highly efficient importance distributions that account for the current observation.

1.1 Articulated Tracking using Particle Filters

Estimating the pose of a person using a single view point or a small baseline stereo camera is an inherently difficult problem due to self-occlusions and visual ambiguities. This manifests itself in that the distribution of the human pose is multi-modal with an unknown number of modes. Currently, the best method for coping with such distributions is the particle filter [12]. This relies on a prior motion model $p(\theta_t|\theta_{t-1})$ and a data likelihood model $p(\mathbf{Z}_t|\theta_t)$. Here θ_t denotes the human pose at time t and $\mathbf{Z}_t$ the observation at the same time. The particle filter approximates the posterior $p(\theta_t|\mathbf{Z}_{1:t})$ as a set of weighted samples. These samples are drawn from an *importance distribution* $q(\theta_t|\mathbf{Z}_t, \theta_{t-1})$ and the weights are computed recursively as

$$w_t^{(n)} \propto w_{t-1}^{(n)} p\left(\mathbf{Z}_t|\theta_t^{(n)}\right) r_t^{(n)} \qquad \text{s.t.} \qquad \sum_{n=1}^{N} w_t^{(n)} = 1 \ , \tag{1}$$

where the superscript (n) denotes sample index and the *correction factor* $r_t^{(n)}$ is given by

$$r_t^{(n)} = \frac{p\left(\theta_t^{(n)}|\theta_{t-1}^{(n)}\right)}{q\left(\theta_t^{(n)}|\mathbf{Z}_t, \theta_{t-1}^{(n)}\right)} \ . \tag{2}$$

In practice, it is common use the motion prior as the importance distribution, i.e. to let $q(\theta_t|\mathbf{Z}_t, \theta_{t-1}) = p(\theta_t|\theta_{t-1})$ as then $r_t^{(n)} = 1$ which simplifies development. This does, however, have the unwanted side-effect that the importance distribution is "blind" to the current observation, such that the samples can easily be placed far away from the modes of the likelihood (and hence the modes of the posterior). In practice, this increases the number of samples needed for successful tracking. As the likelihood has to be evaluated for each sample, this quickly becomes a costly affair; in general the likelihood is expensive to evaluate as it has to traverse the data.

To use the particle filter for articulated tracking, we need a human pose representation. As is common [1], we shall use the kinematic skeleton (see fig. 1). 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 model the bones as having known constant length, so the angles between connected bones constitutes the only degrees of freedom in the kinematic skeleton. We collect these into one large vector θ_t representing all joint angles in the model at time t. To represent constraints on the joint angles, they are confined to a subset $\mathbf{\Theta}$ of $\mathbb{R}^N$.

From known bone lengths and a joint angle vector θ_t, the joint positions can be computed recursively using *forward kinematics* [13]. We will let $F(\theta_t)$ denote the joint positions corresponding to the joint angles θ_t. In this paper, we will make a distinction between joint *angles* and joint *positions* as this has profound impact when designing data-driven importance distributions.

1.2 Related Work

In articulated tracking, much work has gone into improving either the likelihood model or the motion prior. Likelihood models usually depend on cues such as *edges* [2–4],

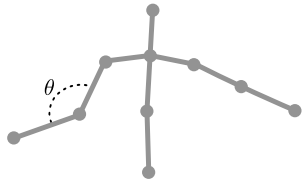


Fig. 1. An illustration of the *kinematic skeleton*. Bones are connected in a tree structure where branches have constant length. Angles between connected bones constitute the only degrees of freedom in the model.

optical flow [4, 11] or *background subtraction* [3, 5, 14–18]. Motion priors are usually crafted by learning activity specific priors, such as for *walking* [6, 7, 19, 20]. These approaches work by restricting the tracker to some subspace of the joint angle space, which makes the priors activity specific. When no knowledge of the activity is available it is common [5, 6, 18, 21] to simply let θ_t follow a normal distribution with a diagonal covariance, i.e.

$$p_{\text{gp}}(\theta_t|\theta_{t-1}) \propto \mathcal{N}(\theta_t|\theta_{t-1}, \text{diag}) \ \mathcal{U}_{\boldsymbol{\Theta}}(\theta_t) \ , \tag{3}$$

where $\mathcal{U}_{\boldsymbol{\Theta}}$ is a uniform distribution on the legal set of angles that encodes the joint constraints. Recently, Hauberg et al. [8] showed that this model causes the spatial variance of the joint positions to increase as the kinematic chains are traversed. In practice this means that with this model the spatial variance of e.g. the hands is always larger than of the shoulders. To avoid this somewhat arbitrary behaviour it was suggested to build the prior distribution directly in the spatial domain; a solution we will review in sec. 3.

In this paper we design data-driven importance distributions; a sub-field of articulated tracking where little work has been done. One notable exception is the work of Poon and Fleet [9], where a hybrid Monte Carlo filter was suggested. In this filter, the importance distribution uses the gradient of the log-likelihood, which moves the samples closer to the modes of the likelihood function (and, hence, also closer to the modes of the posterior). This approach is reported to improve the overall system performance.

In the more general filtering literature, the *optimal particle filter* [12] is known to vastly improve the performance of particle filters. This filter incorporates the observation in the importance distribution, such that samples are drawn from $p(\theta_t|\theta_{t-1}, \mathbf{Z}_t)$, where $\mathbf{Z}_t$ denotes the observation at time t. In practice, the optimal particle filter is quite difficult to implement as non-trivial integrals need to be solved in closed-form. Thus, solutions are only available for non-linear extensions to the Kalman filter [12] and for non-linear extensions of left-to-right Hidden Markov models with known expected state durations [22].

2 A Failed Experiment

Our approach is motivated by a simple experiment, which proved to be a failure. In an effort to design data-driven importance distributions, we designed a straight-forward importance distribution based on silhouette observations. We, thus, assume we have a

binary image $\mathbf{B}_t$ available, which roughly separates the human from the scene. When sampling new poses, we will ensure that joint positions are within the human segment. We model the motion prior according to eq. 3, i.e. assume that joint angles follow a normal distribution with diagonal covariance.

Let $\mathcal{U}_{\mathbf{B}_t}$ denote the uniform distribution on the binary image $\mathbf{B}_t$, such that background pixels have zero probability and let $\text{proj}_{im}[F(\theta_t)]$ be the projection of joint positions $F(\theta_t)$ onto the image plane. We then define the importance distribution as

$$\tilde{q}(\theta_t|\mathbf{B}_t, \theta_{t-1}) \propto \mathcal{N}(\theta_t|\theta_{t-1}, \text{diag})\ \mathcal{U}_{\mathbf{\Theta}}(\theta_t)\ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)])\ . \tag{4}$$

The two first terms correspond to the motion prior and the third term ensures that sampled joint positions are within the human segment in the silhouette image. It is worth noticing that the correction factor $r_t^{(n)}$ (eq. 2) becomes constant for this importance distribution and hence can be ignored.

It is straight-forward to sample from this importance distribution using *rejection sampling* [23]: new samples can be drawn from the motion prior until one is found where all joint positions are within the human segment. This simple scheme, which is illustrated in fig. 2, should improve tracking quality. To measure this, we develop one articulated tracker where the motion prior (eq. 3) is used as importance distribution and one where eq. 4 is used. We use a likelihood model and measure of tracking error described later in the paper; for now details are not relevant. Fig. 3a and 3b shows the tracking error as well as the running time for the two systems as a function of the number of samples in the filter. As can be seen, the data-driven importance distribution *increases* the tracking error with approximately one centimetre, while roughly requiring 10 times as many computations. An utter failure!

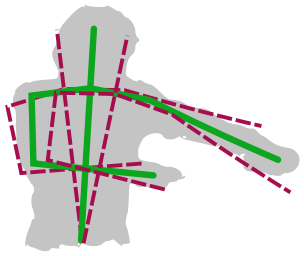


Fig. 2. An illustration of the rejection sampling scheme for simulating the importance distribution in eq. 4. The green skeleton drawn in full lines is accepted, while the two red dashed skeletons are rejected as at least one joint is outside the silhouette.

To get to the root of this failure, we need to look at the motion prior. As previously mentioned, Hauberg et al. [8] have pointed out that the spatial variance of the joint positions increases as the kinematic chains are traversed. This means that e.g. hand positions are always more variant than shoulder positions. In practice, this leads to rather large spatial variances of joint positions. This makes the term $\mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)])$ dominant

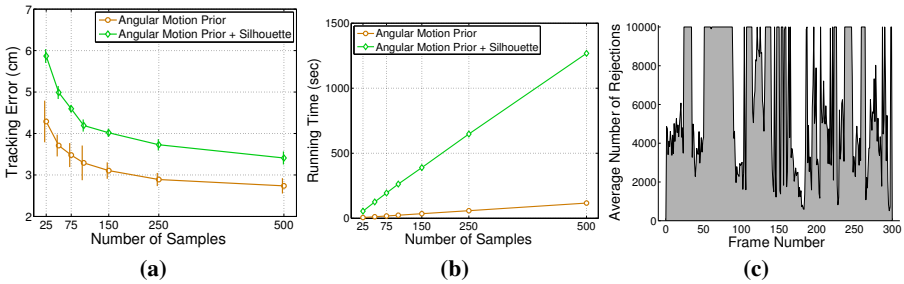


Fig. 3. Various performance measures for the tracking systems; errorbars denote one standard deviation of the attained results over several trials. (a) The tracking error measured in centimetre. (b) The running time per frame. (c) The average number of rejections in each frame.

in eq. 4, thereby diminishing the effect of the motion prior. This explains the increased tracking error. The large running time can also be explained by the large spatial variance of the motion prior. For a sampled pose to be accepted in the rejection sampling scheme, *all* joint positions need to be inside the human silhouette. Due to the large spatial variance of the motion prior, many samples will be rejected, leading to large computational demands. To keep the running time under control, we maximally allow for 10000 rejections. Fig. 3c shows the average number of rejections in each frame in a sequence; on average 6162 rejections are required to generate a sample where all joint positions are within the human silhouette. Thus, the poor performance, both in terms accuracy and speed, of the importance distribution in eq. 4 is due to the large spatial variance of the motion prior. This indicates that we should be looking for motion priors with more well-behaved spatial variance. We will turn to the framework suggested by Hauberg et al. [8] as it was specifically designed for controlling the spatial variance of joint positions. We shall briefly review this work next.

3 Spatial Predictions

To design motion priors with easily controlled spatial variance, Hauberg et al. [8] first define a spatial pose representation manifold $\mathcal{M} \subset \mathbb{R}^{3L}$, where L denotes the number of joints. A point on this manifold corresponds to all spatial joint positions of a pose parametrised by a set of joint angles. More stringently, $\mathcal{M}$ can be defined as

$$\mathcal{M} = \{F(\theta) \mid \theta \in \mathbf{\Theta}\} \ , \tag{5}$$

where F denotes the forward kinematics function for the entire skeleton. As this function is injective with a full-rank Jacobian, $\mathcal{M}$ is a compact differentiable manifold embedded in $\mathbb{R}^{3L}$. Alternatively, one can think of $\mathcal{M}$ as a quadratic constraint manifold arising due to the constant distance between connected joints. It should be noted that while a point on $\mathcal{M}$ corresponds to a point in $\mathbf{\Theta}$, the metrics on the two spaces are different, giving rise to different behaviours of seemingly similar distributions.

A Gaussian-like predictive distribution on $\mathcal{M}$ can be defined simply by projecting a Gaussian distribution in $\mathbb{R}^{3L}$ onto $\mathcal{M}$, i.e.

$$p_{\text{proj}}(\theta_t|\theta_{t-1}) = \text{proj}_{\mathcal{M}}\left[\mathcal{N}(F(\theta_t)|F(\theta_{t-1}), \Sigma)\right] \quad . \tag{6}$$

When using a particle filter for tracking, one only needs to be able to draw samples from the prior model. This can easily be done by sampling from the normal distribution in $\mathbb{R}^{3L}$ and projecting the result onto $\mathcal{M}$. This projection can be performed in a direct manner by seeking

$$\hat{\theta}_t = \arg\min_{\theta_t} \left\|\hat{\mathbf{x}}_t - F(\theta_t)\right\|^2 \qquad \text{s.t.} \qquad \theta_t \in \mathbf{\Theta} \ , \tag{7}$$

where $\hat{\mathbf{x}}_t \sim \mathcal{N}(F(\theta_t)|F(\theta_{t-1}), \Sigma)$. This is an *inverse kinematics* problem [13], where all joints are assigned a goal. Eq. 7 can efficiently be solved using gradient descent by starting the search in θ_{t-1}.

4 Data-Driven Importance Distributions

We now have the necessary ingredients for designing data-driven importance distributions. In this paper, we will be designing two such distributions: one based on silhouette data and another on depth data from a stereo camera. Both will follow the same basic strategy.

4.1 An Importance Distribution based on Silhouettes

Many articulated tracking systems base their likelihood models on simple background subtractions [3, 5, 14–18]. As such, importance distributions based on silhouette data are good candidates for improving many systems. We, thus, assume that we have a binary image $\mathbf{B}_t$ available, which roughly separates the human from the scene. When predicting new joint positions, we will ensure that they are within the human segment.

The projected prior (eq. 6) provides us with a motion model where the variance of joint positions can easily be controlled. We can then create an importance distribution similar to eq. 4,

$$q_{\text{bg}}(\theta_t|\mathbf{B}_t, \theta_{t-1}) \propto p_{\text{proj}}(\theta_t|\theta_{t-1}) \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)]) \ . \tag{8}$$

While the more well-behaved spatial variance of this approach would improve upon the previous experiment, it would still leave us with a high dimensional rejection sampling problem. As this has great impact on performance, we suggest an approximation of the above importance distribution,

$$\begin{aligned} q_{\text{bg}}(\theta_t|\mathbf{B}_t, \theta_{t-1}) &\propto p_{\text{proj}}(\theta_t|\theta_{t-1}) \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)]) && (9)\\ &= \text{proj}_{\mathcal{M}}\left[\mathcal{N}(F(\theta_t)|F(\theta_{t-1}), \Sigma)\right] \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)]) && (10)\\ &\approx \text{proj}_{\mathcal{M}}\left[\mathcal{N}(F(\theta_t)|F(\theta_{t-1}), \Sigma) \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)])\right] \ . && (11) \end{aligned}$$

In other words, we suggest imposing the data-driven restriction in the embedding space before projecting back on manifold. When the covariance Σ is block-diagonal, such that the position of different joints in embedding space are independent, this importance distribution can be written as

$$q_{\mathrm{bg}}(\theta_t|\mathbf{B}_t, \theta_{t-1}) \approx \mathrm{proj}_{\mathcal{M}} \left[\prod_{l=1}^{L} \mathcal{N}(\mu_{l,t}|\mu_{l,t-1}, \Sigma_l)\ \mathcal{U}_{\mathbf{B}_t}(\mathrm{proj}_{im}[\mu_{l,t}]) \right] , \qquad (12)$$

where $\mu_{l,t}$ denotes the position of the l^{th} joint at time t and Σ_l denotes the block of Σ corresponding to the l^{th} joint. We can sample efficiently from this distribution using rejection sampling by sampling each joint position independently and ensuring that they are within the human silhouette. This is L three dimensional rejection sampling problems, which can be solved much more efficiently than one $3L$ dimensional problem. After the joint positions are sampled, they can be projected onto the representation manifold $\mathcal{M}$, such that the sampled pose respects the skeleton structure.

A few samples from this distribution can be seen in fig. 4c, where samples from the angular prior from eq. 3 is available as well for comparative purposes. As can be seen, the samples from the silhouette-driven importance distribution are much more aligned with the true pose, which is the general trend.

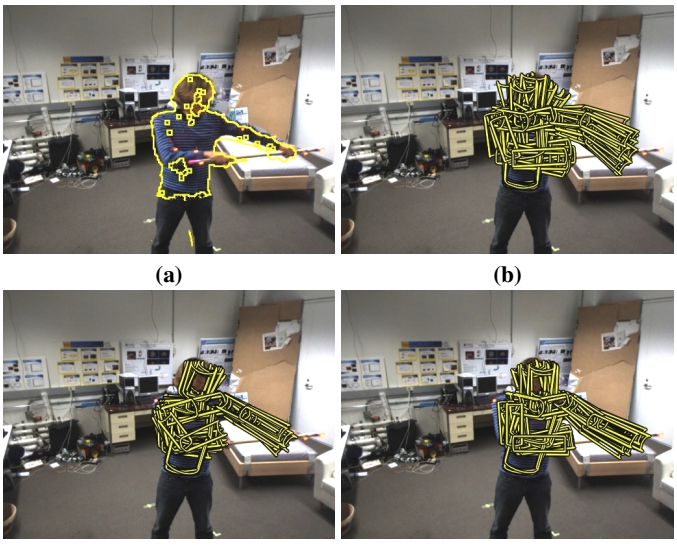


Fig. 4. Samples from various importance distributions. Notice how the data-driven distributions generate more "focused" samples. (a) The input data with the segmentation superimposed. (b) Samples from the angular prior (eq. 3). (c) Samples from the importance distribution guided by silhouette data. (d) Samples from the importance distribution guided by depth information.

4.2 An Importance Distribution based on Depth

Several authors have also used depth information as the basis of their likelihood model. Some have used stereo [8, 10, 24] and others have used time-of-flight cameras [25]. When depth information is available it is often fairly easy to segment the data into background and foreground simply by thresholding the depth. As such, we will extend the previous model with the depth information. From depth information we can generate a set of points $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_K\}$ corresponding to the surface of the observed objects. When sampling a joint position, we will simply ensure that it is not too far away from any of the points in $\mathbf{Z}$.

To formalise this idea, we first note that the observed surface corresponds to the skin of the human, whereas we are modelling the skeleton. Hence, the joint positions should not be directly *on* the surface, but a bit away depending on the joint. For instance, hand joints should be closer to the surface than a joint on the spine. To encode this knowledge, we let $\mathbf{Z}^{\oplus r_l}$ denote the set of three dimensional points where the shortest distance to any point in $\mathbf{Z}$ is less than r_l, i.e.

$$\mathbf{Z}^{\oplus r_l} = \{\mathbf{z} \mid \min_k(\|\mathbf{z} - \mathbf{z}_k\|) < r_l\} \ . \tag{13}$$

Here the r_l threshold is set to be small for hands, large for joints on the spine and so forth. When we sample individual joint positions, we ensure they are within this set, i.e.

$$\begin{aligned} q_{\text{depth}}(\theta_t|\mathbf{Z}, \theta_{t-1}) &\propto p_{\text{proj}}(\theta_t|\theta_{t-1}) \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)]) \ \mathcal{U}_{\mathbf{Z}^{\oplus}}\big(F(\theta_t\big) \\ &\approx \text{proj}_{\mathcal{M}}\left[\prod_{l=1}^{L} \mathcal{N}(\mu_{l,t}|\mu_{l,t-1}, \Sigma_l) \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[\mu_{l,t}]) \ \mathcal{U}_{\mathbf{Z}^{\oplus r_l}}(\mu_{l,t})\right] \end{aligned} \tag{14}$$

where $\mathcal{U}_{\mathbf{Z}^{\oplus r_l}}$ is the uniform distribution on $\mathbf{Z}^{\oplus r_l}$. Again, we can sample from this distribution using rejection sampling. This requires us to compute the distance from the predicted position to the nearest point in depth data. We can find this very efficiently using techniques from kNN classifiers, such as k-d trees [26].

Once all joint positions have been sampled, they are collectively projected onto the manifold $\mathcal{M}$ of possible poses. A few samples from this distribution is shown in fig. 4d. As can be seen, the results are visually comparable to the model based on background subtraction; we shall later, unsurprisingly, see that for out-of-plane motions the depth model does outperform the one based on background subtraction.

5 A Simple Likelihood Model

In order to complete the tracking system, we need a system for computing the likelihood of the observed data. To keep the paper focused on prediction, we use a simple vision system [8] based on a consumer stereo camera[1]. This camera provides a dense set of three dimensional points $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_K\}$ in each frame. The objective of the vision system then becomes to measure how well a pose hypothesis matches the points. We

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

assume that points are independent and that the distance between a point and the skin of the human follows a zero-mean Gaussian distribution, i.e.

$$p(\mathbf{Z}|\theta_t) \propto \prod_{k=1}^{K} \exp\left(-\frac{\min\left[D^2(\theta_t, \mathbf{z}_k),\ \tau\right]}{2\sigma^2}\right) \ , \tag{15}$$

where $D^2(\theta_t, \mathbf{z}_k)$ denotes the squared distance between the point $\mathbf{z}_k$ and the skin of the pose θ_t and τ is a constant threshold. The minimum operation is there to make the system robust with respect to outliers.

We also need to define the skin of a pose, such that we can compute distances between this and a data point. Here, the skin of a bone is defined as a capsule with main axis corresponding to the bone itself. Since we only have a single view point, we discard the half of the capsule that is not visible. The skin of the entire pose is then defined as the union of these half-capsules. The distance between a point and this skin can then be computed as the smallest distance from the point to any of the half-capsules.

6 Experimental Results

We now have two efficient data-driven importance distributions and a likelihood model. This gives us two systems for articulated tracking that we now validate by comparison with one using the standard activity independent prior that assumes normally distributed joint angles (eq. 3) as importance distribution. We use this motion prior as reference as it is the most commonly used model. As ground truth we will be using data acquired with an optical marker-based motion capture system.

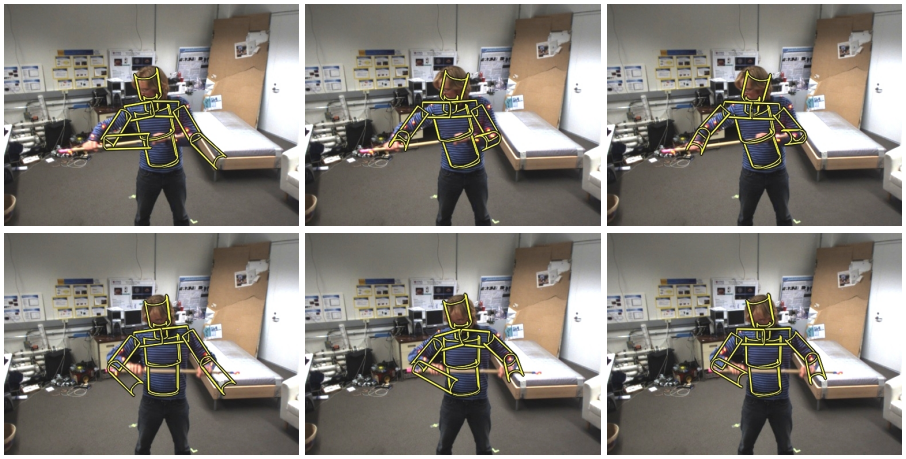
We first illustrate the different priors on a sequence where a person is standing in place while waving a stick. This motion utilises the shoulders a lot; something that often causes problems for articulated trackers. As the person is standing in place, we only track the upper body motions.

In fig. 5 we show attained results for the different importance distributions; a film with the results are available as part of the supplementary material. Visually, we see that the data-driven distributions improve the attained results substantially. Next, we set out to measure this gain.

To evaluate the quality of the attained results we position motion capture markers on the arms of the test subject. We then measure the average distance between the motion capture markers and the capsule skin of the attained results. This measure is then averaged across frames, such that the error measure becomes

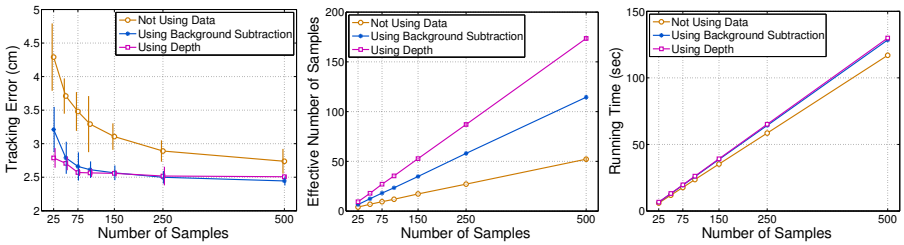
$$\mathcal{E} = \frac{1}{TM} \sum_{t=1}^{T} \sum_{m=1}^{M} D(\hat{\theta}_t, \mathbf{v}_m) \ , \tag{16}$$

where $D(\hat{\theta}_t, \mathbf{v}_m)$ is the orthogonal Euclidean distance between the m^{th} motion capture marker and the skin at time t. The error measure is shown in fig. 6a using between 25 and 500 particles. As can be seen, both data-driven importance distributions perform substantially better than the model not utilising the data. For a small number of samples,



(a) **(b)** **(c)**

Fig. 5. Results from trackers using 150 particles with the different importance distributions. The general trend is that the data-driven distributions improve the results. (a) The angular prior from eq. 3. (b) The importance distribution guided by background subtraction. (c) The importance distribution guided by depth information.



(a) **(b)** **(c)**

Fig. 6. Various performance measures for the tracking systems using different importance distributions on the first sequence. Errorbars denote one standard deviation of the attained results over several trials. (a) The tracking error $\mathcal{E}$. (b) The effective number of samples N_{eff}. (c) The running time per frame.

the model based on depth outperforms the one based on background subtraction, but for 150 particles and more, the two models perform similarly.

In the particle filtering literature the quality of the Monte Carlo approximation is sometimes measured by computing the *effective number of samples* [12]. This measure can be approximated by

$$N_{eff} = \left(\sum_{n=1}^{N} w_t^{(n)} \right)^{-1} , \tag{17}$$

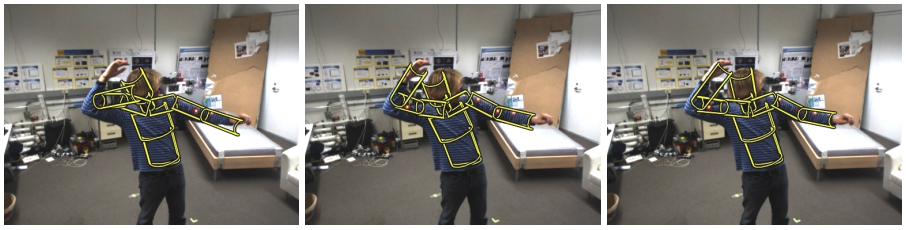
where $w_t^{(n)}$ denotes the weight of the n^{th} sample in the particle filter. Most often this measure is used to determine when resampling should be performed; here we will use it to compare the different importance distributions. We compute the effective number of samples in each frame and compute the temporal average. This provides us with a measure of how many of the samples are actually contributing to the filter. In fig. 6b we show this for the different importance distributions as a function of the number of particles. As can be seen, the data-driven importance distributions gives rise to more effective samples than the one not using the data. The importance distribution based on background subtraction gives between 1.6 and 2.2 times more effective samples than the model not using data, while the model using depth gives between 2.3 and 3.3 times more effective samples.

We have seen that the data-driven importance distributions improve the tracking substantially as they increase the effective number of samples. This benefit, however, comes at the cost of an increased running time. An obvious question is then whether this extra cost outweigh the gains. To answer this, we plot the running times per frame for the tracker using the different distributions in fig. 6c. As can be seen, the two data-driven models require the same amount of computational resources; both requiring approximately 10% more resources than the importance distribution not using the data. In other words, we can triple the effective number of samples at 10% extra cost.

We repeat the above experiments for a different sequence, where a person is moving his arms in a quite arbitrary fashion; a type of motion that is hard to predict and as such also hard to track. Example results are shown in fig. 7, with a film again being available as part of the supplementary material. Once more, we see that the data-driven importance distributions improve results. The tracking error is shown in fig. 8a; we see that the importance distribution based on depth consistently outperforms the one based on background subtraction, which, in turn, outperforms the one not using the data. The effective number of samples is shown in fig. 8b. The importance distribution based on background subtraction gives between 1.8 and 2.2 times more effective samples than the model not using data, while the model using depth gives between 2.8 and 3.6 times more effective samples. Again a substantial improvement at little extra cost.

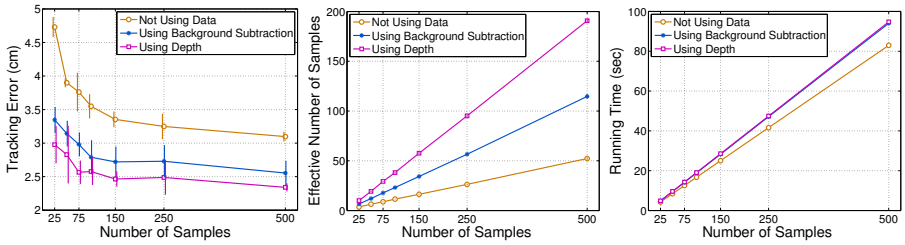
7 Conclusion

We have suggested two efficient importance distributions for use in articulated tracking systems based on particle filters. They gain their efficiency by an approximation that allows us to sample joint positions independently. A valid pose is then constructed by a projection onto the manifold $\mathcal{M}$ of possible joint positions. While this projection might



(a) **(b)** **(c)**

Fig. 7. Results from trackers using 150 particles with the different importance distributions. The general trend is that the data-driven distributions improve the results. (a) The angular prior from eq. 3. (b) The importance distribution guided by silhouette data. (c) The importance distribution guided by depth information.



(a) **(b)** **(c)**

Fig. 8. Various performance measures for the tracking systems using different importance distributions on the second sequence. Errorbars denote one standard deviation of the attained results over several trials. (a) The tracking error $\mathcal{E}$. (b) The effective number of samples N_{eff}. (c) The running time per frame.

seem complicated it merely correspond to a least-squares fit of a kinematic skeleton to the sampled joint positions. As such, the suggested importance distributions are quite simple, which consequently means that the algorithms are efficient and that they actually work. In fact, our importance distributions triple the effective number of samples in the particle filter, at little extra computational cost. The simplicity of the suggested distributions also makes them quite general and easy to implement. Hence, they can be used to improve many existing tracking systems with little effort.

References

1. Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding **108** (2007) 4–18
2. Sminchisescu, C., Triggs, B.: Kinematic Jump Processes for Monocular 3D Human Tracking. In: In IEEE International Conference on Computer Vision and Pattern Recognition. (2003) 69–76
3. Duetscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: CVPR, Published by the IEEE Computer Society (2000) 2126
4. Sminchisescu, C., Triggs, B.: Estimating articulated human motion with covariance scaled sampling. The International Journal of Robotics Research **22** (2003) 371
5. Kjellström, H., Kragić, D., Black, M.J.: Tracking people interacting with objects. In: IEEE CVPR. (2010)
6. Sidenbladh, H., Black, M.J., Fleet, D.J.: Stochastic tracking of 3d human figures using 2d image motion. In: ECCV. Volume II of LNCS 1843., Springer (2000) 702–718
7. Urtasun, R., Fleet, D.J., Fua, P.: 3D People Tracking with Gaussian Process Dynamical Models. In: IEEE CVPR. (2006) 238–245
8. Hauberg, S., Sommer, S., Pedersen, K.S.: Gaussian-like spatial priors for articulated tracking. In Daniilidis, K., Maragos, P., , Paragios, N., eds.: ECCV. Volume 6311 of LNCS., Springer (2010) 425–437
9. Poon, E., Fleet, D.J.: Hybrid monte carlo filtering: Edge-based people tracking. IEEE Workshop on Motion and Video Computing **0** (2002) 151
10. Hauberg, S., Pedersen, K.S.: Stick it! articulated tracking using spatial rigid object priors. In: ACCV 2010, Springer-Verlag (2010)
11. Brubaker, M.A., Fleet, D.J., Hertzmann, A.: Physics-based person tracking using the anthropomorphic walker. International Journal of Computer Vision **87** (2010) 140–155
12. Doucet, A., Godsill, S., Andrieu, C.: On sequential monte carlo sampling methods for bayesian filtering. Statistics and computing **10** (2000) 197–208
13. Erleben, K., Sporring, J., Henriksen, K., Dohlmann, H.: Physics Based Animation. Charles River Media (2005)
14. Balan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: IEEE CVPR. (2007) 1–8
15. Vondrak, M., Sigal, L., Jenkins, O.C.: Physical simulation for probabilistic motion tracking. In: CVPR, IEEE (2008)
16. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture. International Journal of Computer Vision **87** (2010) 75–92
17. Bandouch, J., Beetz, M.: Tracking humans interacting with the environment using efficient hierarchical sampling and layered observation models. In: Computer Vision Workshops (ICCV Workshops). (2009)

18. Balan, A.O., Sigal, L., Black, M.J.: A quantitative evaluation of video-based 3d person tracking. Visual Surveillance and Performance Evaluation of Tracking and Surveillance **0** (2005) 349–356
19. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian Process Dynamical Models for Human Motion. IEEE PAMI **30** (2008) 283–298
20. Lu, Z., Carreira-Perpinan, M., Sminchisescu, C.: People Tracking with the Laplacian Eigenmaps Latent Variable Model. In Platt, J., Koller, D., Singer, Y., Roweis, S., eds.: Advances in Neural Information Processing Systems 20. MIT Press (2008) 1705–1712
21. Bandouch, J., Engstler, F., Beetz, M.: Accurate human motion capture using an ergonomics-based anthropometric human model. In: AMDO '08: Proceedings of the 5th international conference on Articulated Motion and Deformable Objects, Springer-Verlag (2008) 248–258
22. Hauberg, S., Sloth, J.: An efficient algorithm for modelling duration in hidden markov models, with a dramatic application. J. Math. Imaging Vis. **31** (2008) 165–170
23. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
24. Ziegler, J., Nickel, K., Stiefelhagen, R.: Tracking of the articulated upper body on multi-view stereo image sequences. In: IEEE CVPR. (2006) 774–781
25. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S.: Real time motion capture using a single time-of-flight camera. In: IEEE CVPR. (2010) 755–762
26. Arya, S., Mount, D.M.: Approximate nearest neighbor queries in fixed dimensions. In: Proc. 4th ACM-SIAM Sympos. Discrete Algorithms. (1993) 271280

Paper 5: Natural Metrics and Least-Committed Priors for Articulated Tracking

Authors: Søren Hauberg, Stefan Sommer and Kim S. Pedersen.
Status: Submitted to *Image and Vision Computing* [26].

The previous papers in the thesis have suggested and extended a projected prior on the kinematic manifold. The idea of projecting a Gaussian distribution to a Riemannian manifold has seen some use in directional statistics [34] as an approximation to Riemannian Brownian motion. It is, however, a less known model, and its behaviour has not seen much study. To remedy this, we show how to work with Brownian motion on embedded manifolds, such as the kinematic manifold. This stochastic model was chosen because it is one of the most well-known models available, and also it forms the basis of most stochastic calculus.

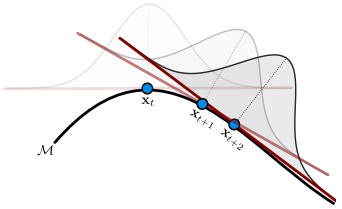


Figure 6.1 *An illustration of Riemannian Brownian motion. Infinitesimal Gaussian random steps are taken in the tangent space.*

Technically, we work with Brownian motion expressed as a Stratonovich stochastic differential equation and show how to simulate this on embedded manifolds. To the best of our knowledge, this is the first numerical solution to stochastic differential equations on manifolds.

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)

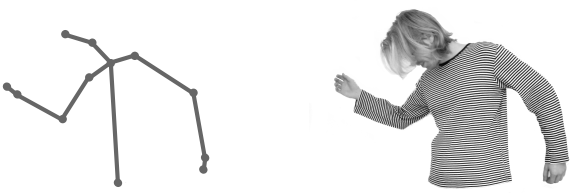


Figure 1: Left: a rendering of the kinematic skeleton. Each bone position is computed by a rotation and a translation relative to its parent. Right: an image showing a human in the pose represented by the skeleton to the left.

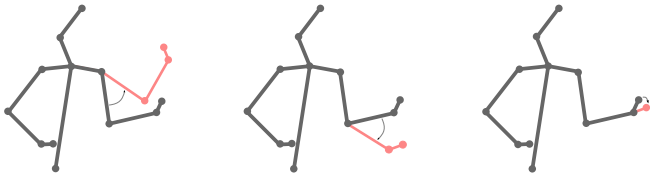


Figure 2: Examples of three motions that are equally large in the commonly used angular metric. All examples have a 45 degree angular distance to the initial pose.

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.

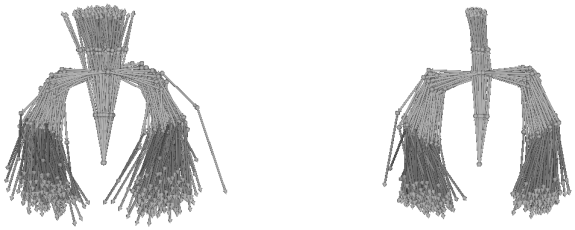
As a motivating example of the behaviour of d_θ, we show three movements of "equal size" in fig. 2. In all movements one joint has been moved 45 degrees, while the remaining have been kept constant. While the actual numerical changes from the initial positions are the same, the movements appear to be 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

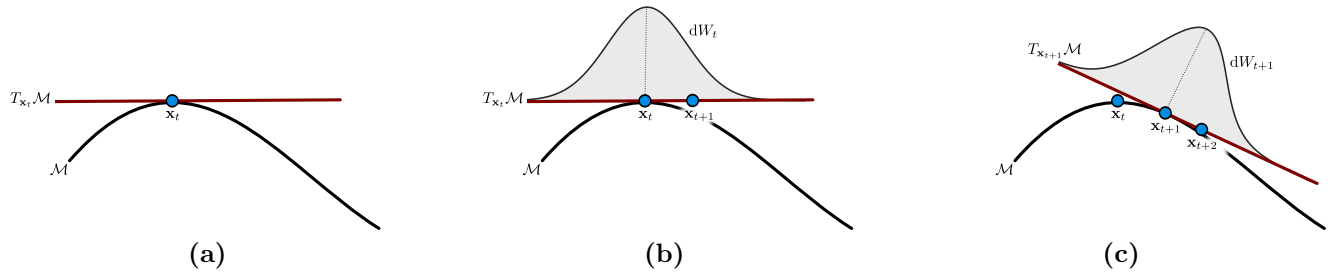


Figure 4: Steps in the Brownian motion model. (a) The manifold $\mathcal{M}$ along with the tangent space $T_{\mathbf{x}_t}\mathcal{M}$ at $\mathbf{x}_t$. (b) A normal distribution $\mathrm{d}W_t$ in the embedding space with mean value $\mathbf{x}_t$ is projected to the tangent space. The value $\mathbf{x}_{t+1}$ 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.

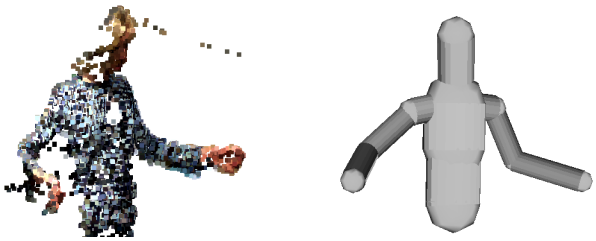


Figure 5: Left: an input data example. Noisy three dimensional points are scattered around the surface of the human body. Right: the skin model. Each bone is assigned a capsule and the collection of capsules describes the skin.

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).

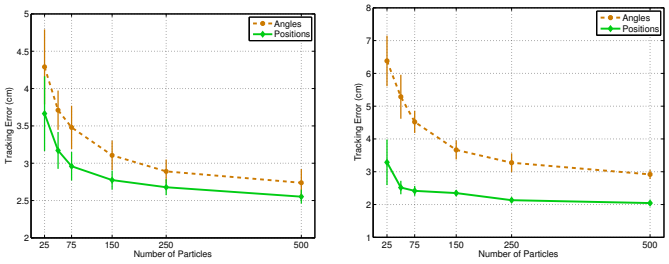
We use a small baseline consumer stereo camera[1] for acquiring data. At each time instance we, thus, get a set of three dimensional points $\mathbf{Z}_t = \{\mathbf{z}_t^{(1)}, \dots, \mathbf{z}_t^{(K)}\}$ that are mostly scattered around the surface of the human as well as around the surrounding environment (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

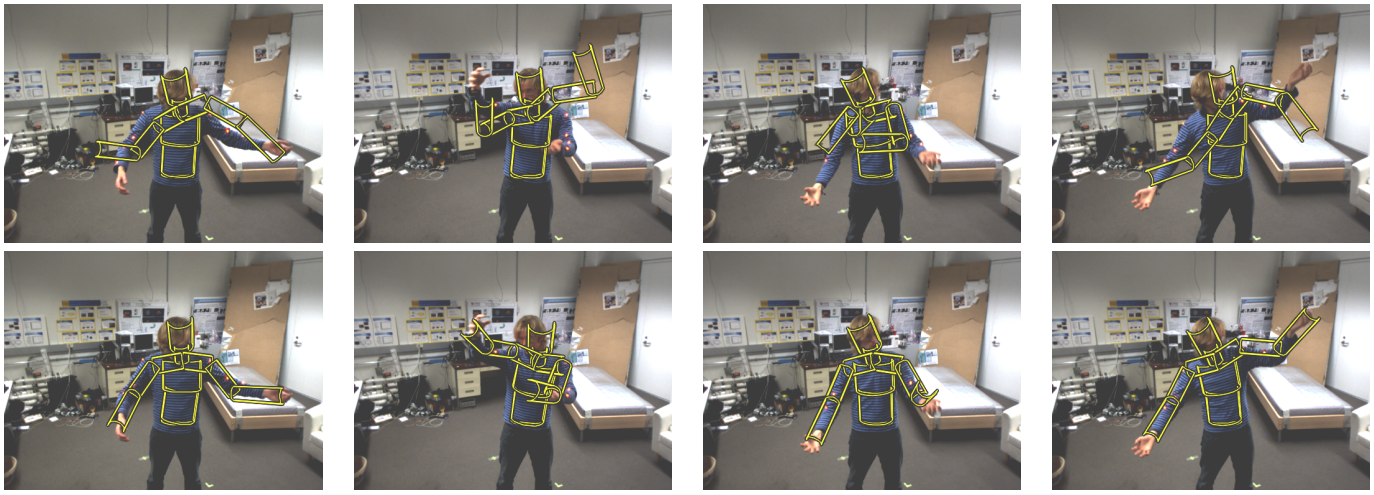
To illustrate models that incorporate environmental knowledge, we replicate an experiment suggested by Kjell-

[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.

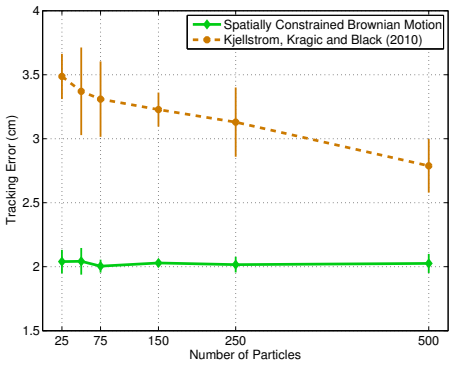


Figure 10: Tracking error for the spatially constrained Brownian motion used for modelling object interaction.

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.

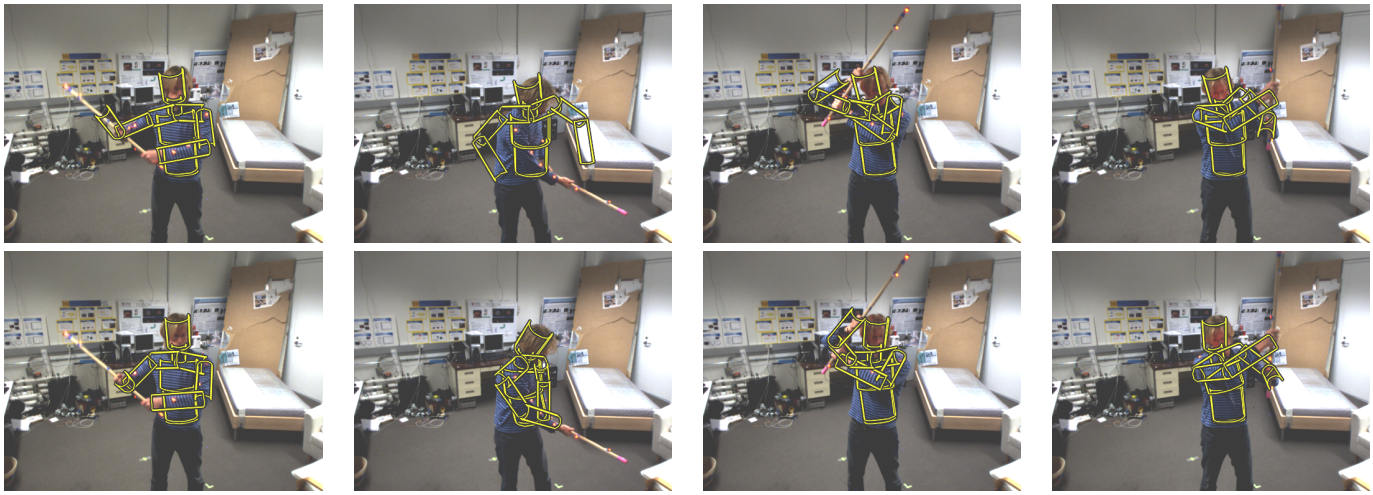
In order to design the metric, we introduced the kinematic manifold consisting of the position of all joints in the kinematic skeleton. This manifold allows us to apply techniques from Riemannian geometry when designing motion models. Our specific focus has been on predictive stochastic processes for 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.

Paper 6: Unscented Kalman Filtering on Riemannian Manifolds

Authors: Søren Hauberg, François Lauze and Kim S. Pedersen.
Status: Submitted to *Journal of Mathematical Imaging and Vision* [25].

The previous papers have all been concerned with performing tracking on a Riemannian manifold using the particle filter. The need for such a "brute-force" Monte Carlo approach is somewhat dissatisfying, as closed-form solutions often appear as parts of more complicated models. For this reason, we investigated whether the classic Kalman Filter [29] could be generalised to Riemannian manifolds. As it turns out, the Unscented Kalman Filter [28] was perfectly suited for this as is shown in the next paper. Besides providing a general tool on Riemannian manifolds, the paper also provides a hint that we should not be afraid of using Riemannian manifolds for modelling, as the algorithms need not be complicated.

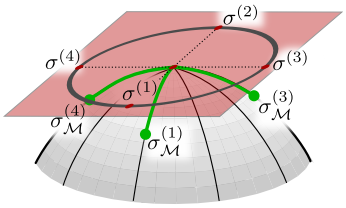


Figure 7.1 *An illustration of the Riemannian unscented transform. The unscented transform is performed in the tangent space and the individual sigma points are send to the manifold.*

The basic insight of the paper is that the unscented transform is also applicable in Riemannian domains. This gives us the benefit that we need not be concerned with more complicated aspects of the manifold than geodesics, parallel transports and exponential and logarithm maps. This makes our approach very general as these can be computed numerically on many manifolds.

From a modelling point of view, this paper bears great resemblance to the previously described Brownian motion model. The algorithmic approach is, however, different, as the Kalman filter is a deterministic algorithm, whereas the Brownian motion model was derived to be used with Monte Carlo methods such as the particle filter.

Submitted to Journal of Mathematical Imaging and Vision

Unscented Kalman Filtering on Riemannian Manifolds

Søren Hauberg · François Lauze · Kim Steenstrup Pedersen

the date of receipt and acceptance should be inserted later

Abstract In recent years there has been a growing interest in problems, where either the observed data or hidden state variables are confined to a known Riemannian manifold. In sequential data analysis this interest has also been growing, but rather crude algorithms have been applied: either Monte Carlo filters or brute-force discretisations. These approaches scale poorly and clearly show a missing gap: no generic analogues to Kalman filters are currently available in non-Euclidean domains. In this paper, we remedy this issue by first generalising the unscented transform and then the unscented Kalman filter to Riemannian manifolds. As the Kalman filter can be viewed as an optimisation algorithm akin to the Gauss-Newton method, our algorithm also provides a general-purpose optimisation framework on manifolds. We illustrate the suggested method on an articulated tracking problem as well as a pose optimisation problem, where constraints on joint positions impose a manifold structure.

1 Modelling with Manifolds

In many statistical problems it is becoming increasingly common to model non-linearities by confining parts of the model to a Riemannian manifold. This often provides better and more natural metrics, which has direct impact on the statistical models. The benefits of having good metrics have led manifolds to be used in a wide variety of models. Sometimes the observed data itself lives on a manifold, e.g. in Diffusion Tensor Imaging [7, 24] and shape analysis [8, 15]. Other times the hidden state variables of a generative model are confined to a non-Euclidean domain, e.g. in image segmentation [6] and human motion modelling [10, 12].

One notable downside to working with manifolds is the lack of many basic tools known from the Euclidean domain. While generalisations of mean values and covariances [23], as well as principal component analysis [8, 29] are available, most remaining tools are still missing. In this paper we tackle one of the most fundamental models for sequential data analysis: the Kalman filter. Our approach is based on the unscented Kalman filter (UKF) [13], which is a widely applied generalisation of the linear Kalman filter. This turns out to be a perfect fit for Riemannian manifolds as the unscented transform is readily generalisable. Furthermore, our approach has the advantage that only limited knowledge of the manifold is needed to apply the filter.

This paper is structured as follows. In the next section we discuss related work on filtering on manifolds and thereafter we provide a brief introduction to Riemannian manifolds including basic statistics (sec. 2). We present the theoretical contribution of the paper by presenting a generalisation of the unscented transform and the unscented Kalman filter for Riemannian manifolds (sec. 3). To illustrate the applicability of the new filter, we develop an articulated tracker for estimating the pose of a moving person over time (sec. 4). The paper is concluded in sec. 5 with a discussion of further developments.

S. Hauberg
Dept. of Computer Science, University of Copenhagen
E-mail: hauberg@diku.dk

F. Lauze
Dept. of Computer Science, University of Copenhagen
E-mail: francois@diku.dk

K.S. Pedersen
Dept. of Computer Science, University of Copenhagen
E-mail: kimstp@diku.dk

1.1 Filtering on Manifolds

Filtering is the task of estimating the moments of the hidden state variable of a non-linear dynamical system [4],

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}) \equiv \hat{f}(\mathbf{x}_{t-1}, \mathbf{v}_{t-1}) \ , \tag{1}$$

$$\mathbf{y}_t = h(\mathbf{x}_t) \quad \equiv \hat{h}(\mathbf{x}_t, \mathbf{n}_t) \ , \tag{2}$$

where $\mathbf{x}_t$ is the hidden state and $\mathbf{y}_t$ is the observation. The *process noise* $\mathbf{v}_t$ and *observation noise* $\mathbf{n}_t$ determine the stochastic nature of the system. To ease notation, we shall omit the noise terms in the rest of the paper. The (often non-linear) functions f and h respectively determine the system dynamics and relate the state to the observation. In the Euclidean case, this problem can be solved in closed-form when $\mathbf{x}_t$ is discrete using hidden Markov models [26], and by the Kalman filter [14] when the noise is additive and Gaussian with f and h linear. For other models, approximation schemes, such as particle filters [4] or extended [20] and unscented Kalman filters [13], are required.

When $\mathbf{x}_t$ is confined to a Riemannian manifold $\mathcal{M}$ the scenario is more difficult due to the inherent non-linearities of the state space. One option, advocated by Tidefelt and Schön [30], is to discretise $\mathcal{M}$ and use an ordinary hidden Markov model on the discrete domain. Others [12, 21, 33] have solved the problem using particle filters, which are easily generalised to manifolds. Both approaches are affected with the curse of dimensionality and the complexity scales exponentially with the dimension of the state space; the former approach requires more discretisation bins and the latter more particles as the dimension increases. For many problems the computational burden of these approaches becomes too much and alternatives are needed.

In Euclidean domains, the Kalman filter provides an efficient and robust solution that scales well. In general, this filter is, however, not applicable to Riemannian manifolds, though some work has been done on selected Lie groups [17, 18, 31]. Tyagi and Davis [31] have shown how this filter can be applied for a specific dynamical model on the Lie group of positive definite symmetric matrices. Similarly, Kraft [17] and Kwon et al. [18] show how to apply the unscented Kalman filter on the Lie groups of Quaternions and $SO(3)$ and $SE(3)$, respectively. While these approaches has some similarity to our work, they are based on specific knowledge of the Lie groups and cannot easily be generalised to other domains.

In this paper we provide a filter more general than the ones suggested for Lie groups, as it works for at least any geodesically complete Riemannian manifolds. First, we pause to review some basic tools from Riemannian geometry as these are needed to fully grasp the details of the filter. It should, however, be noted that these details are fully encapsulated in the filter and are not needed to apply the filter in practice.

2 Basic Tools on Riemannian Manifolds

In this section we recall some of the elementary aspects of Riemannian geometry and more details can be found in appendix A. Riemannian geometry [5] studies smooth manifolds endowed with a Riemannian metric. A metric on a manifold M is a smoothly varying inner product given in the tangent space $T_{\mathbf{x}}\mathcal{M}$ at each point $\mathbf{x}$ on the manifold. The tangent space at a point x of $\mathcal{M}$ is a Euclidean space, which locally approximates the manifold. For this reason, the inner product provides an infinitesimal metric, which can be integrated along the manifold. Thus, the length L of a curve $\alpha : [0, 1] \to \mathcal{M}$ connecting two points on $\mathcal{M}$ is defined by integrating the size of the curve derivative with respect to the local metric,

$$L(\alpha) = \int_0^1 \|\alpha'(\tau)\| \mathrm{d}\tau \ , \tag{3}$$

where $\alpha' \in T_\alpha\mathcal{M}$ denotes the curve derivative. The shortest curve connecting two points is known as a *geodesic*; the distance between two points is defined as the length of a geodesic joining them.

Many operations are defined in the Euclidean tangent space $T_{\mathbf{x}}\mathcal{M}$ and one have mappings back and forth between the manifold and the tangent space. The Riemannian *exponential map* at a point $\mathbf{x}$ of $\mathcal{M}$ maps a tangent vector $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$ to the point $\mathbf{y} = \mathrm{Exp}_{\mathbf{x}}(\mathbf{v})$ such that the curve $t \mapsto \mathrm{Exp}_{\mathbf{x}}(t\mathbf{v})$ is a geodesic going from $\mathbf{x}$ to $\mathbf{y}$ with length $\|\mathbf{v}\|$. It is in general only defined in a neighbourhood of the origin of $T_{\mathbf{x}}\mathcal{M}$. The inverse mapping, which maps $\mathbf{y}$ to $\mathbf{v}$, is the Riemannian *logarithm map*, denoted $\mathrm{Log}_{\mathbf{x}}(\mathbf{y})$. It is in general only defined in a neighbourhood of $\mathbf{x}$. $\mathrm{Exp}_{\mathbf{x}}$ is the straightest local parametrisation of $\mathcal{M}$ in a neighbourhood of $\mathbf{x}$ in the sense that it is the one that locally least deforms distances around $\mathbf{x}$. Given a curve $\alpha : [0, 1] \to \mathcal{M}$ there exists *isometries* (thus preserving inner products) $P_t : T_{\alpha(0)}\mathcal{M} \to T_{\alpha(t)}\mathcal{M}$ called the *parallel transport along* α. Parallel transport extends naturally to more general objects than vectors, called tensors. The parallel transport is the straightest, or least deforming way to move geometric objects along curves, and coupled with the exponential map, it provides the straightest way to move a geometric object from one point of the manifold to a neighbouring one via the geodesic curve that joins them. This is usually defined via the *Levi-Civita connection* and associated *covariant derivatives*

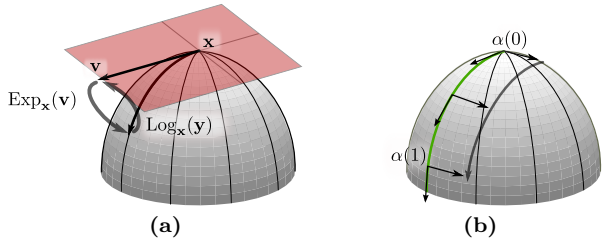


Fig. 1 Graphical illustration of basic manifold operations. (a) The exponential and logarithm maps. (b) The parallel transport for moving vectors along a curve.

uniquely associated to the metric. They are defined in appendix A and illustrated in fig. 1a and 1b.

When $\mathrm{Exp}_{\mathbf{x}}$ is defined on all of $T_{\mathbf{x}}\mathcal{M}$, for each $\mathbf{x} \in \mathcal{M}$, the manifold is said to be *geodesically complete* and the Hopf-Rinow theorem asserts that for any pair of points $\mathbf{x}$ and $\mathbf{y}$ of $\mathcal{M}$, there exist a geodesic joining them with length $d(\mathbf{x}, \mathbf{y})$; for that reason we will assume $\mathcal{M}$ is geodesically complete in the rest of this paper.

Generalising basic statistics to Riemannian manifolds is straightforward [23]. The empirical mean of a set of data points $\mathbf{x} = \{x_1 \dots, x_K\}$ is defined as the point on the manifold that minimises the sum of squared distances:

$$\mathrm{E}[\mathbf{x}] = \arg\min_{\mu} \sum_{k=1}^{K} d^2(\mathbf{x}_k, \mu) \ . \tag{4}$$

Unlike the Euclidean case, such a mean is not necessarily unique; local optima of eq. 4 are known as *Karcher means*, while a global optimum is called the *Fréchet mean*. When $\mu = \mathrm{E}[\mathbf{x}]$ exists, the empirical covariance is generalised as

$$\mathbf{P}_{\mu} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{Log}_{\mu}(\mathbf{x}_k) \, \mathrm{Log}_{\mu}(\mathbf{x}_k)^T \ . \tag{5}$$

In order to represent it as a matrix, a basis in the tangent space is needed. Alternatively, it can also be seen as a bilinear map on $T_{\mu}\mathcal{M}$ in which case a basis is unnecessary. This viewpoint of a *covariant 2-tensor* is useful as tensors can be parallel transported from one point to a neighbouring one. In the case of a symmetric tensor, like the covariance, the following construction provides this parallel transport. When $\mathbf{P}_{\mu}$ is seen as a symmetric matrix, it has an orthonormal basis $(v_1, \dots, v_M)$ made of eigenvectors, with corresponding eigenvalues λ_m. Given a curve α starting at μ, we can parallel transport the eigenvectors along it to obtain a new family $v_m(t)$ in $T_{\alpha(t)}\mathcal{M}$ for each t and we have the following result.

Proposition 1 *The matrix*

$$\boldsymbol{P}_{\alpha(t)} = \sum_{m=1}^{M} \lambda_m v_m(t) v_m(t)^T \tag{6}$$

is the parallel transport of $\boldsymbol{P}_{\mu}$ *along* α *to* $\alpha(t)$.

Proof. The proof relies only on concepts presented above and in the appendix; we refer the reader to the appendix for properties of covariant derivatives and parallel transport.

In the sequel D/dt denotes the covariant derivative along the curve α. By the definition of parallel transport we have $Dv_m(t)/dt = 0$, $m = 0 \dots M$. Furthermore, the compatibility of D/dt with the metric implies that the $v_m(t)$s form a moving orthonormal frame along α. Setting

$$\omega(t)(a(t), b(t)) = \sum_{m=1}^{M} \lambda_m a(t)^T v_m(t) v_m(t)^T b(t) \ , \tag{7}$$

where a and b are two vector fields, i.e., $\omega(t)$ is the bilinear form associated to $\mathbf{P}_{\alpha(t)}$[1]. Hence, it is sufficient to show that

$$\left(\frac{D\omega}{dt}\right)(v_m(t), v_n(t)) = 0, \quad 1 \leq m, n \leq M \ . \tag{8}$$

But

$$\left(\frac{D\omega}{dt}\right)(v_m(t), v_n(t)) = \frac{d}{dt}\left(\omega(t)(v_m(t), v_n(t))\right) \tag{9}$$

$$-\omega(t)\left(\frac{Dv_m}{dt}, v_n(t)\right) - \omega(t)\left(v_m(t), \frac{Dv_n}{dt}\right) \ . \tag{10}$$

The terms in (10) vanish because the fields $v_m(t)$ are parallel and $\omega(t)(v_m(t), v_n(t)) = \lambda_m \lambda_n \delta_{mn}$ is independent of t (δ_{mn} is the Kronecker symbol). □

3 The Manifold UKF

We now have the preliminaries settled and are ready to design the unscented Kalman filter on Riemannian manifolds. We shall first generalise the unscented transform and then we provide the new filter.

3.1 The Unscented Transform

The *unscented transform* [13] is a method for estimating the mean and covariance of a distribution undergoing a non-linear transformation. Given a stochastic variable $\mathbf{x}$, the idea is to pick a set of *sigma points* that fully describe the mean and covariance of $\mathbf{x}$ and then let each sigma point undergo a non-linear transformation f. The mean and covariance of $f(\mathbf{x})$ can then be estimated

[1] The association between a matrix and a covariant 2-tensor requires in general the use of *musical* isomorphisms [19], as it is metric dependent. The use of an orthonormal frame simplifies the situation here.

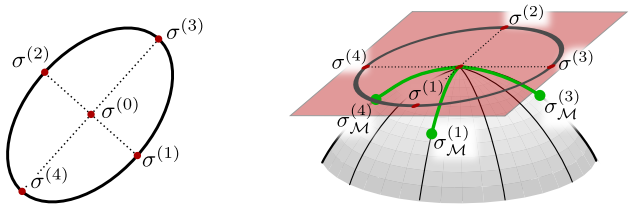


Fig. 2 An illustration of the sigma points. The ellipse represents a covariance. (a) Sigma points in the Euclidean case. (b) Sigma points in the tangent space and on the manifold.

by computing the sample mean and covariance of the transformed sigma points.

In more detail, let $\bar{\mathbf{x}}$ and $\mathbf{P}$ denote the mean and covariance of the M dimensional variable $\mathbf{x}$. The sigma points are then calculated as

$$\sigma^{(0)} = \bar{\mathbf{x}} \tag{11}$$

$$\sigma^{(m)} = \bar{\mathbf{x}} \pm \left(\sqrt{(M+\lambda)\mathbf{P}}\right)_m , \quad m = 1, \ldots, 2M, \tag{12}$$

where $\left(\sqrt{\cdot}\right)_m$ denotes the m^{th} column of the Cholesky decomposition and λ is a parameter for controlling the distance between the sigma points and the mean value. The sigma points are illustrated in fig. 2a. The mean and covariance of $f(\mathbf{x})$ can then be estimated as

$$\mathrm{E}[f(\mathbf{x})] \approx \mu = \sum_{m=0}^{2M} w_m f(\sigma^{(m)}) , \tag{13}$$

$$\mathrm{cov}[f(\mathbf{x})] \approx \sum_{m=0}^{2M} w_m (f(\sigma^{(m)}) - \mu)(f(\sigma^{(m)}) - \mu)^T, \tag{14}$$

where the weights are defined as

$$w_0 = \frac{\lambda}{\lambda + M} , \tag{15}$$

$$w_m = \frac{1}{2(\lambda + M)} , \quad m > 0 . \tag{16}$$

These equations are enough to approximate the correct mean to third order and covariance to the second order [13].

3.1.1 Generalisations

We now consider two different approaches to generalising the unscented transform for Riemannian manifolds; both approaches are based on the same basic observation. Consider a stochastic variable $\mathbf{x} \in \mathcal{M}$ with mean value $\bar{\mathbf{x}}$ and covariance $\mathbf{P}$ expressed in the basis of the tangent space at $\bar{\mathbf{x}}$. Let $\sigma^{(0:2M)} = \{\sigma^{(0)}, \ldots, \sigma^{(2M)}\}$ denote the sigma points of the covariance $\mathbf{P}$ calculated in the (Euclidean) tangent space. Inspection of eq. 5 reveals that

$$\sigma^{(m)}_{\mathcal{M}} = \mathrm{Exp}_{\bar{\mathbf{x}}}(\sigma^{(m)}) , \quad m = 0, \ldots, 2M \tag{17}$$

captures both the mean and covariance. We, thus, have two sets of sigma points that capture the statistics; one set on the manifold and one in the tangent space. These are illustrated in fig. 2b. This gives rise to two different, but equally useful, unscented transforms that we shall both discuss.

First, we consider the case where the non-linear mapping, $f : \mathcal{M}_1 \to \mathcal{M}_2$, moves the sigma points from the manifold to a possibly different (possibly Euclidean) manifold. The mean value can then be estimated by computing the average of the transformed sigma points as discussed in sec. 2, i.e.

$$\begin{aligned} \mathrm{E}[f(\mathbf{x})] &\approx \mu_{\mathcal{M}_2} \\ &= \arg\min_{\mathbf{q} \in \mathcal{M}_2} \sum_{m=0}^{2M} w_m d^2(f(\sigma^{(m)}_{\mathcal{M}_1}), \mathbf{q}) , \end{aligned} \tag{18}$$

where $d(\cdot, \cdot)$ denotes geodesic distance on $\mathcal{M}_2$. The covariance can be estimated in the tangent space of $\mu_{\mathcal{M}_2}$ using eq. 5,

$$\begin{aligned} &\mathrm{cov}[f(\mathbf{x})] \\ &\approx \sum_{m=0}^{2M} w_m \mathrm{Log}_{\mu_{\mathcal{M}_2}}(f(\sigma^{(m)}_{\mathcal{M}_1})) \mathrm{Log}_{\mu_{\mathcal{M}_2}}(f(\sigma^{(m)}_{\mathcal{M}_1}))^T . \end{aligned} \tag{19}$$

A second generalisation considers the case where the non-linear mapping, $f : T_{\bar{\mathbf{x}}}\mathcal{M} \to T_{\bar{\mathbf{x}}}\mathcal{M}$, moves the sigma points in the tangent space. After this transformation a new mean and covariance can be calculated using ordinary Euclidean techniques in the tangent space, i.e.

$$\mathrm{E}[f(\mathbf{x})] \approx \mu_{T_{\bar{\mathbf{x}}}\mathcal{M}} = \sum_{m=0}^{2M} w_m f(\sigma^{(m)}) , \tag{20}$$

$$\mathrm{cov}[f(\mathbf{x})] \approx \sum_{m=0}^{2M} w_m (f(\sigma^{(m)}) - \mu_{T_{\bar{\mathbf{x}}}\mathcal{M}})(f(\sigma^{(m)}) - \mu_{T_{\bar{x}}\mathcal{M}})^T . \tag{21}$$

The mean value $\mu_{T_{\bar{\mathbf{x}}}\mathcal{M}}$ can readily be transferred back to the manifold as $\mu = \mathrm{Exp}_{\bar{\mathbf{x}}}(\mu_{T_{\bar{x}}\mathcal{M}})$; the covariance is transported using parallel transport defined in eq. 6, along the geodesic path $t \mapsto \mathrm{Exp}_{\bar{\mathbf{x}}}(t\mu_{T_{\bar{x}}\mathcal{M}})$, c.f. proposition 1.

3.2 The Unscented Kalman Filter

Before stating the Riemannian generalisations of the Kalman filter, we review the Euclidean techniques for optimal minimum mean-squared error (MMSE) filtering as it provides the basis for the generalisations.

Consider the dynamical system in eq. 1 and 2 with known initial mean and covariance

$$\bar{\mathbf{x}}_0 = \mathrm{E}[\mathbf{x}_0] \ , \tag{22}$$

$$\mathbf{P}_0 = \mathrm{cov}[\mathbf{x}_0] \ . \tag{23}$$

The Euclidean optimal minimum mean-squared error estimate of $\mathbf{x}_t$ can then be written as a linear interpolation between the prediction $\hat{\mathbf{x}}_t$ of $\mathbf{x}_t$ and the predicted observation $\hat{\mathbf{y}}_t$ [20],

$$\bar{\mathbf{x}}_t = \hat{\mathbf{x}}_t + \mathbf{K}(\mathbf{y}_t - \hat{\mathbf{y}}_t) \ , \tag{24}$$

where $\mathbf{K}$ is the so-called *Kalman gain*. Here the terms are calculated as

$$\hat{\mathbf{x}}_t = \mathrm{E}[f(\mathbf{x}_{t-1})] \ , \tag{25}$$

$$\mathbf{K} = \mathbf{P}_{\mathbf{xy}}\mathbf{P}_{\mathbf{yy}}^{-1} \ , \tag{26}$$

$$\hat{\mathbf{y}}_t = \mathrm{E}[h(\hat{\mathbf{x}}_t)] \ , \tag{27}$$

where $\mathbf{P}_{\mathbf{yy}}$ denotes the covariance of $\hat{\mathbf{y}}_t$ and $\mathbf{P}_{\mathbf{xy}}$ the cross-covariance of $\hat{\mathbf{x}}_t$ and $\hat{\mathbf{y}}_t$. The covariance of the state estimate can also be propagated as

$$\mathbf{P}_t = \hat{\mathbf{P}}_t - \mathbf{K}\mathbf{P}_{\mathbf{yy}}\mathbf{K}^T \ , \tag{28}$$

where $\hat{\mathbf{P}}_t = \mathrm{cov}[f(\mathbf{x}_{t-1})]$. The above equations can, however, only be solved in closed-form when f and h are linear functions. One common approximation for the non-linear scenario is the unscented Kalman filter (UKF) [13].

3.2.1 UKF: The Euclidean Case

Let $\bar{\mathbf{x}}_{t-1}$ and $\mathbf{P}_{t-1}$ denote the mean and covariance of the state estimate at time $t-1$. A set of sigma points, $\sigma^{(0:2M)}$, can be calculated from these using the unscented transform, which allows us to estimate $\hat{\mathbf{x}}_t$

$$\hat{\mathbf{x}}_t \approx \sum_{m=0}^{2M} w_m f(\sigma^{(m)}) \ , \tag{29}$$

$$\hat{\mathbf{P}}_t \approx \sum_{m=0}^{2M} w_m (f(\sigma^{(m)}) - \hat{\mathbf{x}}_t)(f(\sigma^{(m)}) - \hat{\mathbf{x}}_t)^T \ . \tag{30}$$

Likewise, the unscented transform can be applied to estimate the effects of h:

$$\hat{\mathbf{y}}_t \approx \sum_{m=0}^{2M} w_m h(\sigma^{(m)}) \ . \tag{31}$$

The covariance and cross-covariance needed to compute the Kalman gain can also be readily approximated

$$\mathbf{P}_{\mathbf{yy}} \approx \sum_{m=0}^{2M} w_m (h(\sigma^{(m)}) - \hat{\mathbf{y}}_t)(h(\sigma^{(m)}) - \hat{\mathbf{y}}_t)^T \ , \tag{32}$$

$$\mathbf{P}_{\mathbf{xy}} \approx \sum_{m=0}^{2M} w_m (f(\sigma^{(m)}) - \hat{\mathbf{x}}_t)(h(\sigma^{(m)}) - \hat{\mathbf{y}}_t)^T \ . \tag{33}$$

These estimates are second-order accurate [13].

3.2.2 UKF: The Riemannian Case

We now use the same approach to generalise the Kalman filter to Riemannian state spaces. Here we shall assume that $h : \mathcal{M} \to \mathcal{M}_{\mathrm{obs}}$. The filter can then be expressed as the following steps, which will be elaborated later:

1. Use the Riemannian generalisation of the unscented transform to estimate the predicted state mean, $\hat{\mathbf{x}}_t = \mathrm{E}[f(\mathbf{x}_{t-1})]$, and covariance $\hat{\mathbf{P}}_t = \mathrm{cov}[f(\mathbf{x}_{t-1})]$.
2. Compute the Riemannian generalisation of the unscented transform of $\hat{\mathbf{P}}_t$ to estimate $\hat{\mathbf{y}}_t$, $\mathbf{P}_{\mathbf{yy}}$ and $\mathbf{P}_{\mathbf{xy}}$. Here, $\hat{\mathbf{y}}_t \in \mathcal{M}_{\mathrm{obs}}$, $\mathbf{P}_{\mathbf{yy}}$ describes covariance in $T_{\hat{\mathbf{y}}_t}\mathcal{M}_{\mathrm{obs}}$ and $\mathbf{P}_{\mathbf{xy}}$ describes the cross-covariance between the sigma points in $T_{\hat{\mathbf{x}}_t}\mathcal{M}$ and $T_{\hat{\mathbf{y}}_t}\mathcal{M}_{\mathrm{obs}}$.
3. Compute state updates $\bar{\mathbf{x}}_t$ and $\mathbf{P}_t$ according to eq. 24 and 28. These will be expressed in $T_{\hat{\mathbf{x}}_t}\mathcal{M}$.
4. Move $\bar{\mathbf{x}}_t$ to the manifold as $\mathrm{Exp}_{\hat{\mathbf{x}}_t}(\bar{\mathbf{x}}_t)$ and parallel transport $\mathbf{P}_t$ to the tangent space at this point.

The above steps are essentially straight-forward generalisations of the Euclidean case. However, as two different generalisations of the unscented transform are available (one in the tangent space and one on the manifold), some details need further attention. In the next two sections, we discuss the first two steps in the above filter in greater detail.

3.2.3 Step 1: Dynamical Models

When the system is predicted according to the dynamical model, two general types of models are worth considering.

The first is when the dynamical model f moves the sigma points directly on the manifold, i.e. $f : \mathcal{M} \to \mathcal{M}$. In this case, each sigma point $\sigma_{\mathcal{M}}^{(n)}$ is propagated through f and a mean and covariance can be estimated according to eq. 18 and 19. This requires knowledge of the logarithm map on $\mathcal{M}$, but not of the parallel transport.

The second class of dynamical models worth considering, is when the sigma points are moved in the tangent space, i.e. $f : T_{\bar{\mathbf{x}}_t}\mathcal{M} \to T_{\bar{\mathbf{x}}_t}\mathcal{M}$. In general, the dynamical model happens directly on the manifold and

should not be expressed in the tangent space. However, if dynamics are simple (i.e. the identity function) or time-steps are small a first-order approximation in the tangent space can be convenient. When the dynamics are expressed in tangent space, the predicted mean and covariance can be estimated using ordinary Euclidean techniques. The mean value can be moved back to the manifold using the exponential map and then the covariance can be parallel transported to this mean value. This does not require knowledge of the logarithm map, but does require the parallel transport.

3.2.4 Step 2: Observation Models

The unscented Kalman filter is a generative model: the h function must generate an observation for each input sigma point. It is, thus, reasonable to require that this function is given a valid state as input, i.e. the input should be confined to the state manifold $\mathcal{M}$. Hence, we will not consider the case where the input is in the tangent space of $\mathcal{M}$, giving $h : \mathcal{M} \to \mathcal{M}_{\text{obs}}$.

Let $\sigma_{\mathcal{M}}^{(0:2M)}$ denote the sigma points corresponding to $\hat{\mathbf{P}}_t$. The mean $\hat{\mathbf{y}}_t$ of $h(\sigma_{\mathcal{M}}^{(0:2M)})$ is computed using eq. 18. The transformed sigma points are then lifted to the tangent space at $\hat{\mathbf{y}}_t$, and $\mathbf{P_{yy}}$ and $\mathbf{P_{xy}}$ are estimated as

$$\mathbf{P_{yy}} \approx \sum_{m=0}^{2M} w_m \operatorname{Log}_{\hat{\mathbf{y}}_t}(h(\sigma_{\mathcal{M}}^{(m)})) \operatorname{Log}_{\hat{\mathbf{y}}_t}(h(\sigma_{\mathcal{M}}^{(m)}))^T, \quad (34)$$

$$\begin{aligned}\mathbf{P_{xy}} &\approx \sum_{m=0}^{2M} w_m \operatorname{Log}_{\hat{\mathbf{x}}_t}(\sigma_{\mathcal{M}}^{(m)}) \operatorname{Log}_{\hat{\mathbf{y}}_t}(h(\sigma_{\mathcal{M}}^{(m)}))^T \\ &= \sum_{m=0}^{2M} w_m \sigma^{(m)} \operatorname{Log}_{\hat{\mathbf{y}}_t}(h(\sigma_{\mathcal{M}}^{(m)}))^T \ .\end{aligned} \quad (35)$$

When the observation manifold is $\mathbb{R}^N$, the logarithm map is no longer necessary and the above equations reduce to

$$\mathbf{P_{yy}} \approx \sum_{m=0}^{2M} w_m (h(\sigma_{\mathcal{M}}^{(m)}) - \hat{\mathbf{y}}_t)(h(\sigma_{\mathcal{M}}^{(m)}) - \hat{\mathbf{y}}_t)^T \ , \quad (36)$$

$$\mathbf{P_{xy}} \approx \sum_{m=0}^{2M} w_m \sigma^{(m)} (h(\sigma_{\mathcal{M}}^{(m)}) - \hat{\mathbf{y}}_t)^T \ . \quad (37)$$

When, moreover, the dynamics can be modelled in the tangent space of the current estimate $T_{\bar{\mathbf{x}}_t}\mathcal{M}$, no logarithm map is involved. This has great practical importance, as apart from relatively simple manifolds, computing the logarithm map generally requires solving an optimal control problem [5].

4 An Example in Articulated Tracking

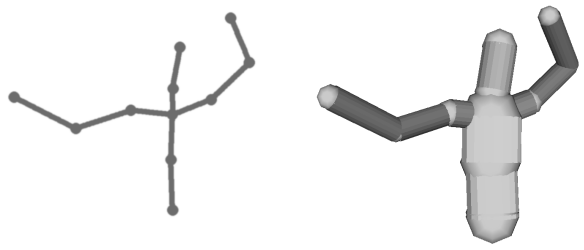
As an example, we now build an articulated tracking system using the suggested filter. The objective of such a system is to estimate the pose of a moving person in each frame of an image sequence [25]. As is common [25], we represent human poses using the *kinematic skeleton* (see fig. 3a), which is a collection of rigid bones connected in a tree structure. As the bones have a constant length the joint angles between connected bones are the only degrees of freedom. Therefore it is common [1, 2, 16, 27] to let the system dynamics be given by a normal distribution in joint angle space. The metric in this space measures the distance between two poses by looking at the difference between individual joint angles. As a consequence the movement of an entire arm by a change in the shoulder joint appears as large as the movement of a finger by a change in a finger joint. This rather unnatural metric makes the dynamical model unstable as some limb positions becomes inherently more variant than others [11, 12].

To avoid such issues, we model the position of all joints, which we concatenate into one vector $\mathbf{x}$ The constraint that each bone has constant length confines $\mathbf{x}$ to a non-linear Riemannian sub-manifold $\mathcal{M}$ of $\mathbb{R}^N$ [12]. The distance measure on this manifold corresponds to the physical distance that joint positions move, which gives the measure a clear physical interpretation. This natural metric gives stable dynamical models [12]. Furthermore, the Riemannian approach has been shown to be very suitable for modelling interaction with the environment due to its spatial nature [10].

4.1 Observation Model

To define the filter, we need to be able to generate observations $h(\sigma_{\mathcal{M}}^{(n)})$ corresponding to each sigma point. We use data from a consumer stereo camera[2], which provides us with a set $\mathbf{Z}_t$ of three dimensional points in each frame. Hence, $h(\sigma_{\mathcal{M}}^{(n)})$ should generate a set of comparable three dimensional points. As the camera observes the surface (or skin) of the person we shall let $h(\sigma_{\mathcal{M}}^{(n)})$ generate a set of three dimensional points on the skin of the pose parametrised by $\sigma_{\mathcal{M}}^{(n)}$. We will generate these points by projecting the actual observed points onto the skin of the pose. This requires a skin model. To simplify the projection, we define the skin of a pose by associating a capsule with each bone (see fig. 3b). The projection of a point onto the skin can then be defined as finding the closest point on the nearest capsule, which can easily be solved in closed-form.

[2] `http://www.ptgrey.com/products/bumblebee2/`



(a) **(b)**

Fig. 3 (a) The kinematic skeleton used for representing poses. (b) The skin model.

In summary, we define the generative observation model as

$$h(\sigma_{\mathcal{M}}^{(n)}) = \text{proj}_{\text{skin}(\sigma_{\mathcal{M}}^{(n)})}(\mathbf{Z}_t) \ . \tag{38}$$

This is essentially the likelihood system presented in [11].

4.2 Dynamical Model

For the dynamical model, we shall use the simplest of all to predict the motion, i.e.

$$f(\mathbf{x}_{t-1}) = \mathbf{x}_{t-1} \ , \tag{39}$$

as experience indicates that such models work better than e.g. second order models [1]. This model can easily be expressed in the tangent space such that no logarithm maps are required, which simplifies development substantially.

With these model choices, we only need the exponential map and parallel transport on the manifold. As $\mathcal{M}$ is a rather complicated manifold, these operations are not available in closed-form and numerical techniques are required. We use a standard forward Euler scheme based on the *standard projection method* [9] for exponential maps and *Schild's Ladder* [22] for parallel transports.

4.3 Results

We apply the filter on two sequences consisting of 300 frames each; both are available as part of the supplementary material and a few frames are shown in fig. 4 and 5. In both sequences the person is standing in place and only moving his upper body. We, thus, only model the upper body parts.

In both sequences, the person moves his arms both parallel and orthogonal to the image plane, causing self-occlusions, which is a challenge. The filter is able to successfully track this motion, though some jitter is observed due to the numerical exponential maps and parallel transports. In a few frames the filter looses track, but it quickly recovers. In the second sequence the person is wearing motion capture markers, which allows us to estimate the tracking error. On average these markers are 2.7cm from the surface of the estimated pose. A particle filter with the same observational and dynamical model more than doubles this error at 5.7cm at the same computational load (75 particles). To achieve the same accuracy as the unscented Kalman filter, the particle filter needs 10 times the computational resources (750 particles).

4.4 Optimisation on Manifolds

It is well-known that the extended Kalman filter can be viewed as a single step in a *Gauss Newton* optimisation scheme [3]. Both the extended and the unscented Kalman filter have been used to solve optimisation problems, such as weight learning in neural networks [28, 32]. An immediate question is then if the Riemannian generalisation of the unscented Kalman filter can be used as a general-purpose optimisation scheme on manifolds. We illustrate this potential on a pose fitting problem related to articulated tracking. We position a skeleton far away from the true pose (see fig. 6a) and optimise the likelihood used in the tracking example by repeated iterations of the unscented Kalman filter. After approximately 60 iterations this converges to the correct pose (see fig. 6b). A video showing the iterations are available in the supplementary material and the optimised error measure is shown in fig. 6c.

5 Conclusion

In this paper, we have introduced an extension of the unscented transform and the unscented Kalman filter to Riemannian manifolds. The idea of working with sigma points seems to be a perfect fit for the Riemannian extension as the approach is both practical and gives descriptive results. The suggested filter has the advantage that only limited knowledge of the manifold is needed for an implementation: in the most general case, only the exponential map, the logarithm map and the parallel transport are required. These are available in closed-form for simple manifolds and numerical techniques exists for more complex scenarios. This makes the filter readily applicable for a wide range of problems. We have successfully illustrated the filter on an articulated tracking problem, where constraints on joint positions impose a non-trivial manifold structure.

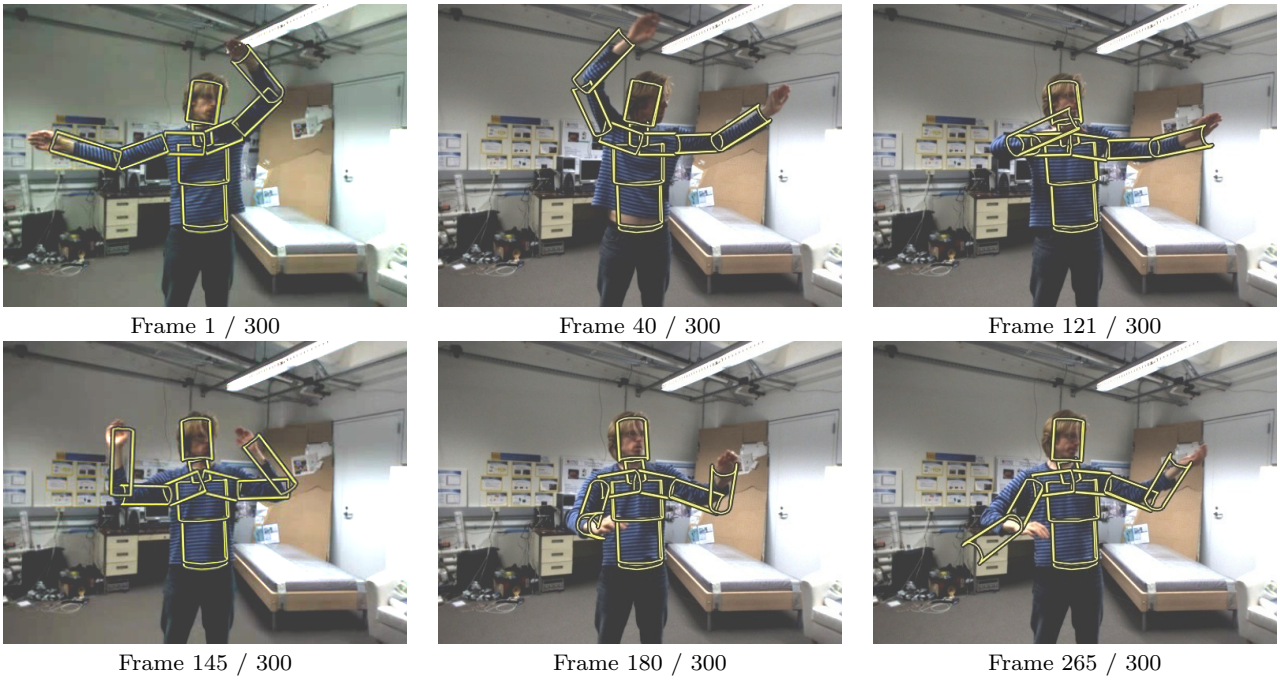


Fig. 4 Six frames from the first sequence; the entire film is available in the supplementary material.

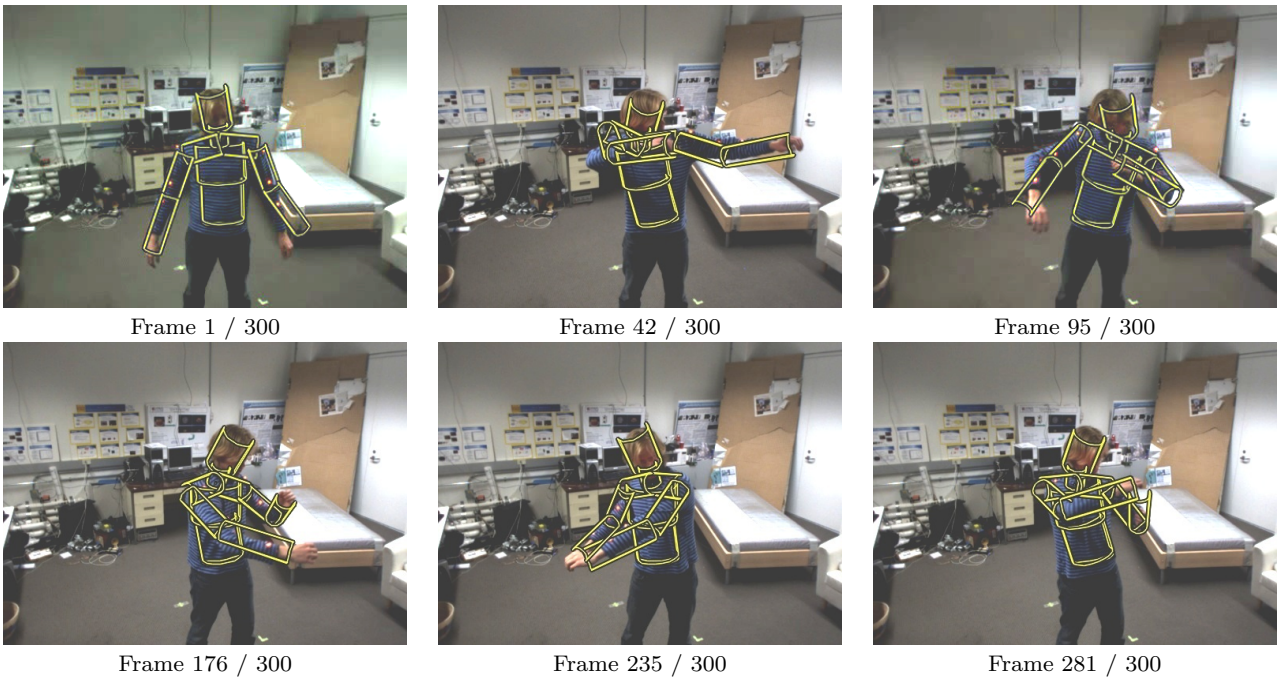
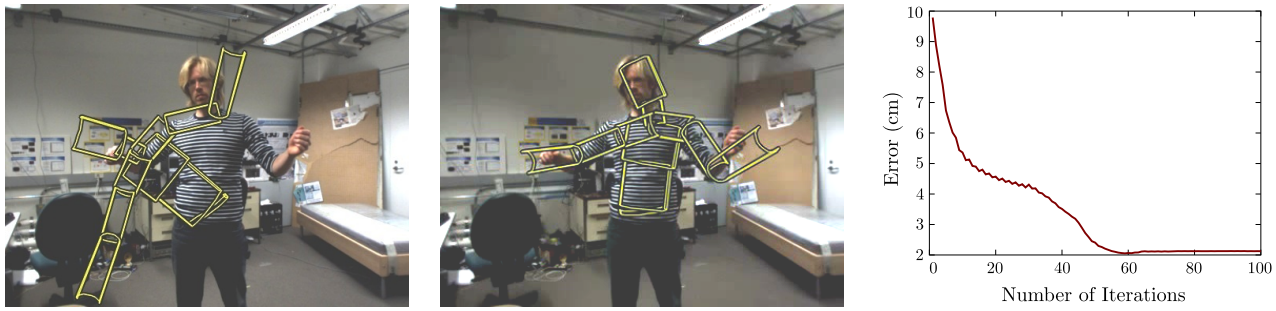


Fig. 5 Six frames from the second sequence; again, the entire film is available in the supplementary material.

We have also shown how the filter can be used as a general-purpose optimisation scheme on manifolds. As the filter does not require much analytical knowledge of the manifold, it is easy to apply on many optimisation problems posed on manifolds, which leads to many interesting applications. This is, however, left for future work.

References

1. Balan AO, Sigal L, Black MJ (2005) A quantitative evaluation of video-based 3d person tracking. Visual Surveillance and Performance Evaluation of Tracking and Surveillance 0:349–356



(a) **(b)** **(c)**

Fig. 6 Using the UKF as an optimisation scheme on manifolds. (a) The initialisation of the optimisation scheme. (b) The located optimum. (c) The error measure of the optimisation as a function of the number of iterations.

2. Bandouch J, Engstler F, Beetz M (2008) Accurate human motion capture using an ergonomics-based anthropometric human model. In: AMDO '08: Proceedings of the 5th international conference on Articulated Motion and Deformable Objects, Springer-Verlag, pp 248–258
3. Bell BM, Cathey FW (1993) The iterated kalman filter update as a Gauss-Newton method. IEEE Trans on Automatic Control 38:294–297
4. Cappé O, Godsill SJ, Moulines E (2007) An overview of existing methods and recent advances in sequential Monte Carlo. Proceedings of the IEEE 95(5):899–924
5. do Carmo MP (1992) Riemannian Geometry. Birkhäuser Boston
6. Caselles V, Kimmel R, Sapiro G (1997) Geodesic active contours. IJCV 22:61–79
7. Fletcher PT, Joshi S (2007) Riemannian geometry for the statistical analysis of diffusion tensor data. Signal Processing 87:250–262
8. Fletcher PT, Lu C, Pizer SM, Joshi S (2004) Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE TMI 23(8):995–1005
9. Hairer E, Lubich C, Wanner G (2004) Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations. Springer
10. Hauberg S, Pedersen KS (2010) Stick it! articulated tracking using spatial rigid object priors. In: ACCV 2010, Springer, LNCS, vol 6494
11. Hauberg S, Pedersen KS (2011) Predicting articulated human motion from spatial processes. International Journal of Computer Vision 94:317–334
12. Hauberg S, Sommer S, Pedersen KS (2010) Gaussian-like spatial priors for articulated tracking. In: ECCV, Springer, LNCS, vol 6311, pp 425–437
13. Julier SJ, Uhlmann JK (1997) A new extension of the kalman filter to nonlinear systems. In: Int. Symp. Aerospace/Defense Sensing, Simul. and Controls, pp 182–193
14. Kalman R (1960) A new approach to linear filtering and prediction problems. Transactions of the ASME-Journal of Basic Engineering 82(D):35–45
15. Kendall DG (1984) Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. Bulletin of the London Mathematical Society 16(2):81–121
16. Kjellström H, Kragić D, Black MJ (2010) Tracking people interacting with objects. In: IEEE CVPR
17. Kraft E (2003) A Quaternion-based Unscented Kalman Filter for Orientation Tracking. In: Proc. of the Sixth International Conference on Information Fusion, pp 47–54
18. Kwon J, Lee KM (2010) Monocular SLAM with locally planar landmarks via geometric Rao-Blackwellized particle filtering on Lie groups. In: Proceedings of CVPR'10
19. Lee JM (1997) Riemannian Manifolds: An Introduction to Curvature, Graduate Texts in Mathematics, vol 176. Springer
20. Lewis FL (1986) Optimal Estimation: With an Introduction to Stochastic Control Theory. Wiley
21. Li R, Chellappa R (2010) Aligning spatio-temporal signals on a special manifold. In: ECCV, Springer, LNCS, vol 6315, pp 547–560
22. Misner C, Thorne K, Wheeler J (1973) Gravitation. W. H. Freeman and Company
23. Pennec X (1999) Probabilities and statistics on riemannian manifolds: Basic tools for geometric measurements. In: NSIP, pp 194–198
24. Pennec X, Fillard P, Ayache N (2004) A riemannian framework for tensor computing. IJCV 66:41–66
25. Poppe R (2007) Vision-based human motion analysis: An overview. Computer Vision and Image Understanding 108(1-2):4–18
26. Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. In: Proceedings of the IEEE, pp 257–286

27. Sidenbladh H, Black MJ, Fleet DJ (2000) Stochastic tracking of 3d human figures using 2d image motion. In: ECCV, Springer, LNCS 1843, vol II, pp 702–718
28. Singhal S, Wu L (1989) Training multilayer perceptrons with the extended kalman algorithm. In: Advances in neural information processing systems 1, pp 133–140
29. Sommer S, Lauze F, Nielsen M (2010) The differential of the exponential map, jacobi fields and exact principal geodesic analysis. CoRR abs/1008.1902
30. Tidefelt H, Schön TB (2009) Robust point-mass filters on manifolds. In: Proc. of the 15th IFAC SYSID
31. Tyagi A, Davis JW (2008) A recursive filter for linear systems on riemannian manifolds. In: CVPR'08, pp 1–8
32. Wan EA, van der Merwe R (2002) The unscented kalman filter for nonlinear estimation. In: Adaptive Systems for Signal Processing, Communications, and Control Symposium, IEEE, pp 153–158
33. Wu Y, Wu B, Liu J, Lu H (2008) Probabilistic tracking on riemannian manifolds. In: ICPR, pp 1–4

A Definitions from Differential geometry

We give definitions of some concepts from differential geometry that we use in the paper (mainly from [5]) for the convenience of the reader.

1. **Differentiable Manifolds:**
   A differentiable manifold of dimension M is a set $\mathcal{M}$ and a family of injective mappings $\mathcal{T} = \{x_i : U_i \subset \mathbb{R}^M \to \mathcal{M}\}$ of open sets U_i of $\mathbb{R}^M$ into $\mathcal{M}$ such that
   - $\bigcup_i x_i(U_i) = \mathcal{M}$, i.e. the open sets cover $\mathcal{M}$.
   - For any pair i, j with $x_i(U_i) \bigcap x_j(U_j) = W \neq \phi$, the mapping $x_j^{-1} \circ x_i$ is differentiable.
   - The family $\mathcal{T}$ is maximal, which means that if (y, V), $y : V \subset \mathbb{R}^M \to \mathcal{M}$ is such that: for each element of $\mathcal{T}$, (x_i, U_i) with $x_i(U_i) \cap y(V) \neq 0$ implies that $y^{-1} \circ x_i$ is a diffeomorphism, then in fact $(y, V) \in \mathcal{T}$.
2. **Directional derivative of a function along a vector field:**
   A vector field X on $\mathcal{M}$ is a map that associates to each $p \in \mathcal{M}$ an element $X(p) \in T_p\mathcal{M}$, where $T_p\mathcal{M}$ is the tangent space of $\mathcal{M}$ at p. The space of smooth vector fields on $\mathcal{M}$ is denoted $\mathfrak{X}(\mathcal{M})$. Let $f : \mathcal{M} \to \mathbb{R}$ be a differentiable function of $\mathcal{M}$ and X a vector field on $\mathcal{M}$. The directional derivative $X.f$ is the function $\mathcal{M} \to \mathbb{R}$,
   $$(X.f)(p) = df_p(X(p)) \tag{40}$$
   the differential of f at p evaluated at vector $X(p)$.
3. **Covariant tensors:**
   A p-covariant tensor h is a $\mathcal{C}^\infty$ p-linear map
   $$\underbrace{T\mathcal{M} \times \cdots \times T\mathcal{M}}_{p \text{ times}} \to \mathcal{C}^\infty(\mathcal{M}) \tag{41}$$
   i.e., for all $x \in \mathcal{M}$, $x \mapsto h_x$:
   $$v_1, \ldots, v_p \in T_x\mathcal{M} \mapsto h_x(v_1, \ldots, v_p) \in \mathbb{R} \tag{42}$$
   is p-linear and for vector fields $X_1, \ldots, X_p \in \mathfrak{X}$, the map $x \mapsto h_x(X_1(x), \ldots, X_p(x))$ is smooth.
4. **Riemannian Metric:**
   A Riemannian metric on a manifold $\mathcal{M}$ is a covariant 2-tensor g which associates to each point $p \in \mathcal{M}$ an inner product $g_p = \langle -, - \rangle_p$ on the tangent space $T_p\mathcal{M}$, i.e., not only is it bilinear, but symmetric and positive definite and thus define a Euclidean distance on each tangent space. In terms of local coordinates, the metric at each point x is given by a matrix, $g_{ij} = \langle X_i, X_j \rangle_x$, where X_i, X_j are tangent vectors to $\mathcal{M}$ at x, and it varies smoothly with x. A *Geodesic curve* is a local minimizer of arc-length computed with a Riemannian metric.
5. **Affine connection:**
   An affine connection ∇ on a differentiable manifold $\mathcal{M}$ is a mapping
   $$\nabla : \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M}) \tag{43}$$
   which is denoted by $\nabla(X, Y) \to \nabla_X Y$ and which satisfies the following properties:
   - $\nabla_{fX+gY} Z = f\nabla_X Z + g\nabla_Y Z$.
   - $\nabla_X(Y + Z) = \nabla_X Y + \nabla_X Z$.
   - $\nabla_X(fY) = f\nabla_X Y + X(f)Y$.

   in which $X, Y, Z \in \mathfrak{X}(\mathcal{M})$ and f, g are $\mathcal{C}^\infty(\mathcal{M})$. This gives a notion of directional derivative of a vector field defined on the manifold. An affine connection extends naturally to more than vector fields, and especially of interest here, covariant tensors: if h is a covariant p-tensor and $X \in \mathfrak{X}(\mathcal{M})$, $\nabla_X h$ is defined as follows. Given p vector fields $Y_1, \ldots, Y_p \in \mathfrak{X}(\mathcal{M})$,
   $$\begin{aligned}(\nabla_X h)(Y_1, \ldots, Y_p) &= X.(h(Y_1, \ldots, Y_p)) \\ &\quad - \sum_{i=1}^{p} h(Y_1, \ldots, \nabla_X Y_i, \ldots, X_p)\end{aligned} \tag{44}$$
6. **Covariant derivatives:**
   Let $\mathcal{M}$ be a differentiable manifold with affine connection ∇. There exists a unique correspondence which associates to a vector field V along the differentiable curve $c : I \to \mathcal{M}$ another vector field $\frac{DV}{dt}$ along c, called the covariant derivative of V along c, such that
   - $\frac{D}{dt}(V + W) = \frac{DV}{dt} + \frac{DW}{dt}$, where W is a vector field along c.
   - $\frac{D}{dt}(fV) = \frac{df}{dt}V + f\frac{DV}{dt}$, where f is a differentiable function on I.
   - If V is induced by a vector field Y, a member of the tangent bundle of $\mathcal{M}$, i.e. $V(t) = Y(c(t))$, then $\frac{DV}{dt} = \nabla_{\frac{dc}{dt}} Y$.

   The covariant derivative extend to covariant tensors via the extension of the connection to them: Given a covariant p-tensor h defined along c and vector fields $U_1, \ldots, U_p$ along c,
   $$\begin{aligned}\frac{Dh}{dt}(U_1(t), \ldots, U_p(t)) &= \frac{D}{dt}\left(h_t(U(t), V(t)\right) \\ &\quad - \sum_{i=1}^{p} h_t\left(U_1(t), \ldots \frac{DU_p}{dt}, \ldots, U_p(t)\right)\end{aligned} \tag{45}$$
7. **Parallel transport:**
   Given a vector $P \in T_{c(0)}\mathcal{M}$, the differential equation
   $$\begin{cases}\frac{DP(t)}{dt} &= 0 \\ P(0) &= P\end{cases} \tag{46}$$

admits a unique solution, called the parallel transport of P along c. The induced map $P \mapsto P(t)$ from $T_{c(0)}\mathcal{M}$ to $T_{c(t)}\mathcal{M}$ is a linear isomorphism.

8. **Levi-Civita connection:**
Given a Riemannian metric g on the manifold $\mathcal{M}$, there exists a unique affine connection ∇ such that
   - compatibility with the metric:

$$X.g(Y,Z) = g(\nabla_X Y, Z) + g(X, \nabla_X Z) \tag{47}$$

   - symmetry:

$$\nabla_X Y - \nabla_Y X = [X,Y] \tag{48}$$

   ($[X,Y]$ is the Lie bracket of X and Y).

   ∇ is the *Levi-Civita connection associated to* g. Note that from the previous items, one has $\nabla_X g = 0$ for any $X \in \mathfrak{X}(\mathcal{M})$ and that the parallel transport in that case is a *linear isometry*.
   The compatibility of ∇ and the metric g can be expressed in term of covariant derivatives: if $X(t) = X(c(t)$ and $Y(t) = Y(c(t))$ are two vector fields along the curve c, and D/dt is the covariant derivative along c,

$$\frac{d}{dt} g(X(t), Y(t)) = g\left(\frac{DX(t)}{dt}, Y(t)\right) + g\left(X(t), \frac{DY(t)}{dt}\right). \tag{49}$$

9. **Christoffel symbols:**
In a parametrized manifold, where the curve $c(t)$ is represented as $(x^1(t), \ldots, x^M(t))$, the covariant derivative of a vector field v becomes

$$\frac{Dv}{dt} = \sum_m \left\{ \frac{dv^m}{dt} + \sum_{i,j} \Gamma_{ij}^m v^j \frac{dx^i}{dt} \right\} \frac{\partial}{\partial x_m} \tag{50}$$

   where the Γ_{ij}^m are the *coefficients of the connection* also known as the *Christoffel symbols* Γ. In particular, the parallel transport equation above becomes the first-order *linear* system

$$\frac{dv^m}{dt} + \sum_{i,j} \Gamma_{ij}^m v^j \frac{dx^i}{dt} = 0, \quad m = 1 \ldots M. \tag{51}$$

   For the Levi-Civita connection associated with the metric g, the corresponding Christoffel symbols are given by

$$\Gamma_{ij}^m = \frac{1}{2} \sum_l \left\{ \frac{\partial}{\partial x_i} g_{jm} + \frac{\partial}{\partial x_j} g_{mi} - \frac{\partial}{\partial x_m} g_{ij} \right\} g^{ml} \tag{52}$$

   g_{ij} is the ij^{th} element of the metric, and g^{ij} is the ij^{th} element of its inverse. A curve is geodesic if the covariant derivative of its tangent vector field is zero everywhere on it, which means that a geodesic curve has zero tangential acceleration. Such a curve c satisfies the second order system of ODEs, which, with the above parametrization becomes

$$\frac{d^2 x^m}{dt^2} + \sum_{ij} \Gamma_{ij}^m \frac{dx^i}{dt} \frac{dx^j}{dt} = 0, \quad m = 1 \ldots M. \tag{53}$$

10. **Exponential map:**
The exponential map is a map $\mathrm{Exp} : T\mathcal{M} \to \mathcal{M}$, that maps $v \in T_q\mathcal{M}$ for $q \in \mathcal{M}$, to a point $\mathrm{Exp}_q v$ in $\mathcal{M}$ obtained by going out the length equal to $|v|$, starting from q, along a geodesic which passes through q with velocity equal to $\frac{v}{|v|}$. Given $q \in \mathcal{M}$ and $v \in T_q\mathcal{M}$, and a parametrization $(x_1, \ldots, x_n)$ around q, $Exp_q(v)$ can be defined as the solution at time 1 of the above system of ODEs (53) with initial conditions $(x^m(0)) = q$ and $(\frac{dx^m}{dt}(0)) = v$, $m = 1, \ldots, M$. The geodesic starting at q with initial velocity t can thus be parametrized as

$$t \mapsto \mathrm{Exp}_q(tv). \tag{54}$$

11. **Logarithm map:**
For $\tilde{q}$ in a sufficiently small neighborhood of q, the length minimizing curve joining q and $\tilde{q}$ is unique as well. Given q and $\tilde{q}$, the direction in which to travel geodesically from q in order to reach $\tilde{q}$ is given by the result of the logarithm map $\mathrm{Log}_q(\tilde{q})$. We get the corresponding geodesics as the curve $t \mapsto \mathrm{Exp}_q(t \,\mathrm{Log}_q \tilde{q})$. In other words, Log is the inverse of Exp in the neighborhood.

Chapter 8

Conclusion

This thesis has presented a series of papers along with a discussion of the basic thoughts behind the presented models. From a modelling point of view, the major point of the thesis is that many things are more easily expressed in terms of joint positions rather than joint angles. The reason behind this point is that we are studying *images*, which are inherently spatial objects. Hence, we should strive to build models in the spatial domain. Interestingly, neurologists believe that humans plan their motion in the spatial domain [1, 35], which we believe is another good reason for building models in the spatial domain. Furthermore, as the surrounding environment is inherently spatial, we get yet another reason: spatial models make it trivial to take the environment into account.

Motivated by these thoughts, we have presented two different basic models of human motion in the spatial domain. The first was a probabilistic interpretation of inverse kinematics, which allowed us to avoid several practical problems that appears in the standard joint angle model. Furthermore, the model allows us to trivially incorporate knowledge of the environment, such as ground plane contact and object interaction into the model. From a theoretical point of view, it was, however, dissatisfying that the model required a regularisation in the joint angle domain.

To scratch this itch, we suggested a second model, without the need for such regularisation. This led to the introduction of the kinematic manifold, which is the space of all possible joint positions. This new geometric view of the kinematic skeleton provided us with a physically natural distance measure between poses. In order to respect this distance measure, we developed a Brownian motion model on the kinematic manifold along with a numerical scheme for simulating the underlying stochastic differential equation. In addition to being theoretically nice, this approach also works quite well in practice. It does, however, require some implementation effort and the numerical scheme could be faster. For our practical applications, we, thus, use a similar model, which we call the *projected prior*. This model consists of Gaussian projected onto the kinematic manifold. For all practical purposes, this works just as well as the Brownian motion model, but it is quite a bit easier to implement and several times faster.

One limiting aspect of the suggested models is that their applications require Monte Carlo methods, which can be computationally demanding. To show that alternatives are indeed possible, we have developed an extension of the unscented Kalman filter to general Riemannian manifolds. This turns out to be remarkably simple, which gives us confidence in our strategy.

Throughout this thesis, we have argued that spatial models are more natural than models in joint angle space and we have provided practical algorithms for working with the spatial models. While we believe in the spatial models, they have their down-sides. The most obvious issue is the computational requirements: it takes more resources to simulate the spatial models than to simulate a joint angle model. In our implementation, the likelihood evaluation dominates the computational resources required and the overhead of the spatial models have little impact, but such observations depend on the choice of likelihood model.

Another potential pitfall when working in the Riemannian space of joint positions is that this space is not Euclidean. This has the downside that the vast array of statistical techniques available in Euclidean domains are not always available in Riemannian domains. This is, however, not a problem for the *probabilistic inverse kinematics* model.

8.1 Contributions

The thesis makes several contributions to state-of-the-art, which can be summarised with the following (also repeated in the summary).

- A probabilistic interpretation of inverse kinematics suitable for tracking [21].
- A geometric interpretation of the kinematic skeleton in the kinematic manifold. In this line of thinking, inverse kinematics becomes a projection operator [24]. The metric on this manifold becomes a physically natural measure of the size of motions [26].
- Different models of human motion during interaction with the environment [19, 21].

- An approximation strategy for designing data-driven importance distributions for articulated tracking [20].

- A Brownian motion model on embedded manifolds along with a novel numerical scheme for simulating the underlying manifold valued stochastic differential equation [26].

- A Riemannian generalisation of the unscented Kalman filter, which provides a general-purpose tool for both filtering and optimisation on Riemannian manifolds [25].

These contributions show that it is possible to work in the spatial domain and that it is beneficial both in terms of filtering properties and for extending the models.

8.2 Scientific Outlook

While we have presented a series of spatial models, the work does not end here. The models opens the door to new applications and extensions that could be interesting directions for future research.

The first extension of the spatial models we would like to consider is to take the surrounding environment more fully into account. Imagine constructing a three dimensional model of the surrounding scene using a stereo camera or some other depth sensor. Can we then use this model to guide the tracker? Already in the first paper in this thesis, we showed how simple it is to ensure that body parts do not penetrate the ground plane, but can we go further? One idea would be to model the inherent "laziness" of most humans: we tend to lean against walls, sit on flat surfaces and similar actions that give rise to contacts between the human and its surrounding environment. Using the spatial models it would be quite straight forward to model this by predicting the individual joint positions closer to the surface of the surroundings whenever said joint is close enough to this surface.

Besides taking the environment into account, the spatial models make is easy to take advantage of image cues to guide the tracker. The data-driven importance distributions was an example of this type of guidance. A similar extension would be to develop body part detectors and use these to guide the tracker. If we e.g. detect the right hand at a given three dimensional position, we can predict the hand near this position with little extra work. If we are only given a two dimensional body part position in image coordinates, then we can follow the approach presented in the third paper and predict the part to be on the three dimensional line going through the optical centre and the image point of the body part. Such an extension should be fairly easy assuming that good body part detectors can be developed.

The spatial models all depend on an inverse kinematics solver, either for locating the mode of the second order bootstrap approximation or for computing projections onto the kinematic manifold. We use a simple gradient descent algorithm, which includes a projection step to cope with joint limits. It would be interesting to extend this solver in several ways. The first extension is to handle more sophisticated joint limits that do not treat different angles independently. Successful extensions to the solver exists [13], but they have not been tried for tracking purposes. Another extension is to ensure that the solver does not generate solutions with self-penetrating limbs. This requires taking the body shape into account when solving the inverse kinematics problem. This problem has not seen much attention in the literature, but it should be solvable for simple shape models such as the capsule-based skin model applied in the papers in this thesis.

We have seen that Brownian motion can be simulated on the kinematic manifold. As this model provides the basis of most stochastic calculus, it would be interesting to generalise more involved stochastic processes to the kinematic manifold. One such example could be the *Gegenbauer process* [31]. This is a pseudo-periodic process, which would be potentially well-suited to describe that many human motion patterns, e.g. *walking*, are highly periodic. It is quite interesting to note that while many models of *walking* have been developed [11, 33, 45, 47–50, 52], only few are actually periodic by design.

8.3 And All Stories have an End

Large parts of this thesis have been accepted for publications at respected conferences and journals. While reviewers have generally been pleased, we have observed a trend: our work is often criticised for being "simple". We consider this to be one of the major advantages of the suggested spatial models: they are fairly simple to both understand and implement. To quote one of the old masters

> *"Simplicity is the ultimate sophistication"*
>
> Leonardo da Vinci

And with these words the thesis ends.

Bibliography

[1] William Abend, Emilio Bizzi, and Pietro Morasso. Human arm trajectory formation. *Brain*, 105(2):331–348, 1982.

[2] Alexandru O. Balan, Leonid Sigal, and Michael J. Black. A quantitative evaluation of video-based 3d person tracking. *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 0:349–356, 2005.

[3] Jan Bandouch and Michael Beetz. Tracking humans interacting with the environment using efficient hierarchical sampling and layered observation models. In *Computer Vision Workshops (ICCV Workshops)*, 2009.

[4] Jan Bandouch, Florian Engstler, and Michael Beetz. Accurate human motion capture using an ergonomics-based anthropometric human model. In *AMDO '08: Proceedings of the 5th international conference on Articulated Motion and Deformable Objects*, pages 248–258. Springer-Verlag, 2008.

[5] Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. ISSN 0899-7667.

[6] Christoph Bregler and Jitendra Malik. Tracking people with twists and exponential maps. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '98. IEEE Computer Society, 1998.

[7] Marcus A. Brubaker, David J. Fleet, and Aaron Hertzmann. Physics-based person tracking using the anthropomorphic walker. *International Journal of Computer Vision*, 87(1-2):140–155, 2010.

[8] Miguel A. Carreira-Perpinan and Zhengdong Lu. The Laplacian Eigenmaps Latent Variable Model. *JMLR W&P*, 2:59–66, 2007.

[9] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3): 197–208, 2000.

[10] Jonathan Duetscher, Andrew Blake, and Ian Reid. Articulated body motion capture by annealed particle filtering. In *CVPR*, page 2126. Published by the IEEE Computer Society, 2000.

[11] Ahmed M. Elgammal and Chan-Su Lee. Tracking People on a Torus. *IEEE PAMI*, 31(3):520–538, March 2009.

[12] Morten Engell-Nørregård, Søren Hauberg, Jérome Lapuyade, Kenny Erleben, and Kim Steenstrup Pedersen. Interactive Inverse Kinematics for Monocular Motion Estimation. In *Proceedings of VRIPHYS'09*, 2009.

[13] Morten Engell-Nørregård, Sarah Niebe, and Kenny Erleben. Local joint–limits using distance field cones in euler angle space. 2010.

[14] Kenny Erleben, Jon Sporring, Knud Henriksen, and Henrik Dohlmann. *Physics Based Animation*. Charles River Media, August 2005. ISBN 978-1584503804.

[15] Juergen Gall, Bodo Rosenhahn, Thomas Brox, and Hans-Peter Seidel. Optimization and filtering for human motion capture. *International Journal of Computer Vision*, 87(1):75–92, 2010. ISSN 0920-5691.

[16] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real time motion capture using a single time-of-flight camera. In *IEEE CVPR*, pages 755–762, 2010.

[17] Sumitra Ganesh. *Analysis of Goal-directed Human Actions using Optimal Control Models*. PhD thesis, EECS Dept., University of California, Berkeley, May 2009.

[18] Keith Grochow, Steven L. Martin, Aaron Hertzmann, and Zoran Popović. Style-based inverse kinematics. *ACM Transaction on Graphics*, 23(3):522–531, 2004. ISSN 0730-0301.

[19] Søren Hauberg and Kim S. Pedersen. Stick It! Articulated Tracking using Spatial Rigid Object Priors. In *ACCV 2010*. Springer-Verlag, 2010.

[20] Søren Hauberg and Kim Steenstrup Pedersen. Data-Driven Importance Distributions for Articulated Tracking. In Yuri Boykov et al., editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, Lecture Notes in Computer Science. Springer, 2011.

[21] Søren Hauberg and Kim Steenstrup Pedersen. Predicting Articulated Human Motion from Spatial Processes. *International Journal of Computer Vision*, 94: 317–334, 2011.

[22] Søren Hauberg and Jakob Sloth. An Efficient Algorithm for Modelling Duration in Hidden Markov Models, with a Dramatic Application. *J. Math. Imaging Vis.*, 31:165–170, July 2008.

[23] Søren Hauberg, Jérome Lapuyade, Morten Engell-Nørregård, Kenny Erleben, and Kim Steenstrup Pedersen. Three Dimensional Monocular Human Motion Analysis in End-Effector Space. In Daniel Cremers et al., editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, Lecture Notes in Computer Science, pages 235–248. Springer, August 2009.

[24] Søren Hauberg, Stefan Sommer, and Kim Steenstrup Pedersen. Gaussian-like Spatial Priors for Articulated Tracking. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV*, volume 6311 of *LNCS*, pages 425–437. Springer, 2010.

[25] Søren Hauberg, François Lauze, and Kim Steenstrup Pedersen. Unscented Kalman Filtering on Riemannian Manifolds. *Journal of Mathematical Imaging and Vision* **(under review)**, 2011.

[26] Søren Hauberg, Stefan Sommer, and Kim Steenstrup Pedersen. Natural Metrics and Least-Committed Priors for Articulated Tracking. *Image and Vision Computing* **(under review)**, 2011.

[27] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. Technical report, Cambridge, MA, USA, 1980.

[28] Simon J. Julier and Jeffrey K. Uhlmann. A new extension of the kalman filter to nonlinear systems. In *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls*, pages 182–193, 1997.

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

[30] Hedvig Kjellström, Danica Kragić, and Michael J. Black. Tracking people interacting with objects. In *IEEE CVPR*, 2010.

[31] Paul M. Lapsa. Determination of Gegenbauer-type random process models. *Signal Processing*, 63:73–90, 1997.

[32] Anders Boesen Lindbo Larsen. Predicting articulated motion. *DIKU Student Project*, 2011.

[33] Zhengdong Lu, Miguel Carreira-Perpinan, and Cristian Sminchisescu. People Tracking with the Laplacian Eigenmaps Latent Variable Model. In John C. Platt, Daphne Koller, Yoram Singer, and Sam Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1705–1712. MIT Press, 2008.

[34] Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley, January 1999.

[35] Pietro Morasso. Spatial control of arm movements. *Experimental Brain Research*, 42(2):223–227, April 1981.

[36] Richard M. Murray, Zexiang Li, and S. Shankar Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, March 1994.

[37] Peter Mysling, Søren Hauberg, and Kim Steenstrup Pedersen. An Empirical Study on the Performance of Spectral Manifold Learning Techniques. In T. Honkela et al., editors, *ICANN 2011*, volume 6791 of *Lecture Notes in Computer Science*, pages 347–354. Springer, Heidelberg, 2011.

[38] Gerard Pons-Moll, Andreas Baak, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bodo Rosenhahn. Multisensor-fusion for 3d full-body human motion capture. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[39] Gerard Pons-Moll, Andreas Baak, Juergen Gall, Laura Leal-Taixe, Meinard Mueller, Hans-Peter Seidel, and Bodo Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[40] Eunice Poon and David J. Fleet. Hybrid monte carlo filtering: Edge-based people tracking. *IEEE Workshop on Motion and Video Computing*, 0:151, 2002.

[41] Ronald Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108(1-2):4–18, 2007. ISSN 1077-3142.

[42] Deva Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:467, 2003. ISSN 1063-6919.

[43] Bodo Rosenhahn, Christian Schmaltz, Thomas Brox, Joachim Weickert, Daniel Cremers, and Hans-Peter Seidel. Markerless motion capture of man-machine interaction. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 0:1–8, 2008.

[44] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[45] Hedvig Sidenbladh, Michael J. Black, and David J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV*, volume II of *LNCS 1843*, pages 702–718. Springer, 2000.

[46] Leonid Sigal, Sidharth Bhatia, Stefan Roth, Michael J. Black, and Michael Isard. Tracking loose-limbed people. In *CVPR '04: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 421–428, 2004.

[47] Cristian Sminchisescu and Allan Jepson. Generative modeling for continuous non-linearly embedded visual inference. In *ICML '04*, pages 759–766. ACM, 2004.

[48] Raquel Urtasun, David. J. Fleet, Aaron Hertzmann, and Pascal Fua. Priors for people tracking from small training sets. In *ICCV*, volume 1, pages 403–410, 2005.

[49] Raquel Urtasun, David J. Fleet, and Pascal Fua. 3D People Tracking with Gaussian Process Dynamical Models. In *IEEE CVPR*, pages 238–245, 2006.

[50] Raquel Urtasun, David J. Fleet, Andreas Geiger, Jovan Popović, Trevor J. Darrell, and Neil D. Lawrence. Topologically-constrained latent variable models. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 1080–1087. ACM, 2008. ISBN 978-1-60558-205-4.

[51] Marek Vondrak, Leonid Sigal, and Odest Chadwicke Jenkins. Physical simulation for probabilistic motion tracking. In *CVPR*. IEEE, 2008.

[52] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian Process Dynamical Models for Human Motion. *IEEE PAMI*, 30(2):283–298, 2008.

[53] Masanobu Yamamoto and Katsutoshi Yagishita. Scene constraints-aided tracking of human body. In *CVPR*, pages 151–156. Published by the IEEE Computer Society, 2000.