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

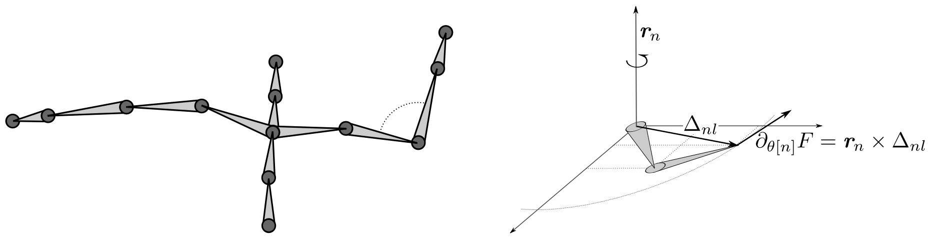
1 Articulated Tracking

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*.


$\boldsymbol{r}_n$
Δ_{nl}
$\partial_{\theta[n]}F = \boldsymbol{r}_n \times \Delta_{nl}$

(a) **(b)**

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 $\varDelta_{nl}$ 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 $\partial_{\boldsymbol{\theta}[n]} F_l = \boldsymbol{r}_n \times \varDelta_{nl}$. 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 $\boldsymbol{\varTheta}$ of $\mathbb{T}^N$. With this further restriction, $\mathcal{M}$ becomes a manifold with boundary.

For simplicity, $\boldsymbol{\varTheta}$ is often defined by confining each component of $\boldsymbol{\theta}$ to an interval, i.e. $\boldsymbol{\varTheta} = \prod_{n=1}^{N} [l_n, u_n]$, 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

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

where $\mathcal{U}_{\boldsymbol{\varTheta}}$ denotes the uniform distribution on $\boldsymbol{\varTheta}$ enforcing the angular constraints and the subscript t denotes time. We shall call this model the *Angular Prior*. In practice, $\varSigma_{\boldsymbol{\theta}}$ 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 $\text{cov}[F(\boldsymbol{\theta}_t)]$. 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

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

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.

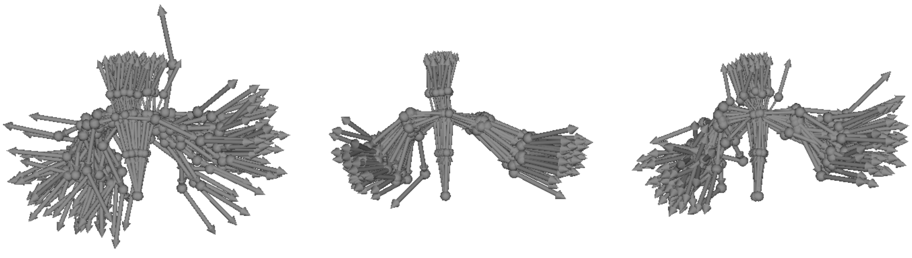

(a) p_{angle} **(b)** p_{proj} **(c)** $p_{\mathcal{T}}$

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 $F(\hat{\boldsymbol{\theta}}_t) = \mathrm{proj}_{\mathcal{M}}[\boldsymbol{x}_t]$. 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 $\Sigma_{\mathcal{T}} = \mathbf{J}_{\boldsymbol{\theta}_{t-1}} \Sigma_{\boldsymbol{\theta}} \mathbf{J}_{\boldsymbol{\theta}_{t-1}}^T$. 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

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

where $\mathbf{J}_{\boldsymbol{\theta}_{t-1}}^{\dagger} = (\mathbf{J}_{\boldsymbol{\theta}_{t-1}}^T \mathbf{J}_{\boldsymbol{\theta}_{t-1}})^{-1} \mathbf{J}_{\boldsymbol{\theta}_{t-1}}^T$ 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 $\mathbf{J}_{\boldsymbol{\theta}_{t-1}}^{\dagger}$ 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 $\boldsymbol{\theta}_t = \left(\mathbf{J}_{\boldsymbol{\theta}_{t-1}}^{\dagger}\right)^T \boldsymbol{x} + \boldsymbol{\theta}_{t-1}$. 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.

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

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.

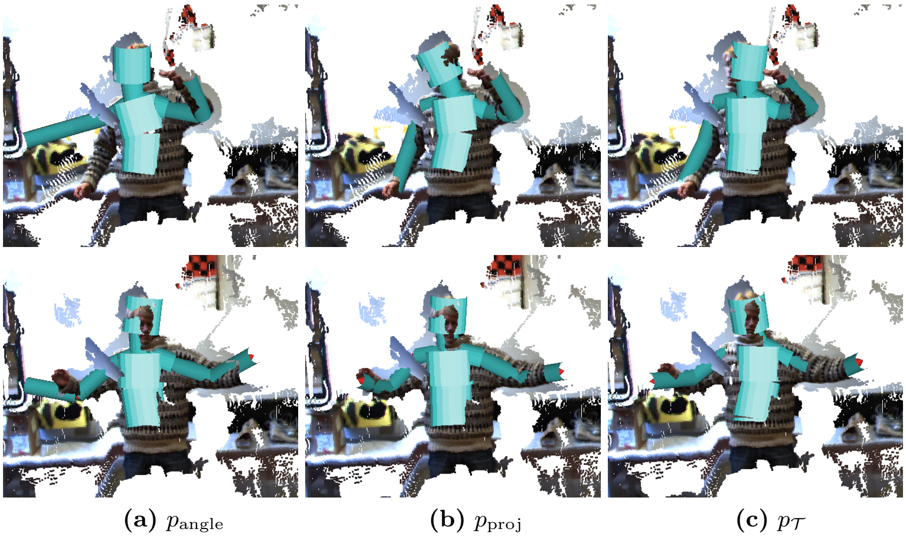

(a) p_{angle} **(b)** p_{proj} **(c)** $p_{\mathcal{T}}$

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.

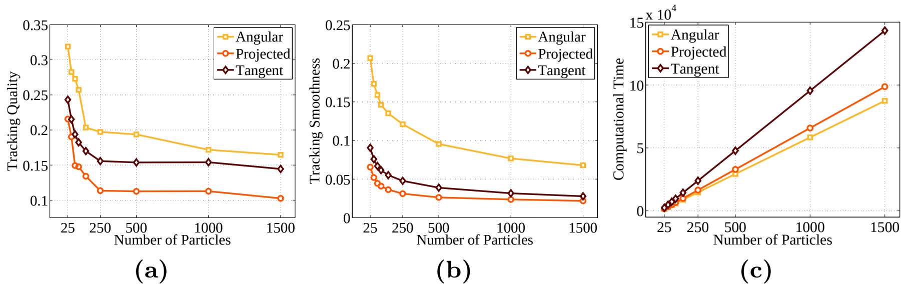


Fig. 4. Performance of the three priors. All reported numbers are averaged over several trials. (a) The error measure $\mathcal{E}$ as a function of the number of particles. The average standard deviation of $\mathcal{E}$ with respect to the trials are 0.018 for the angular prior, 0.008 for the projected 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.

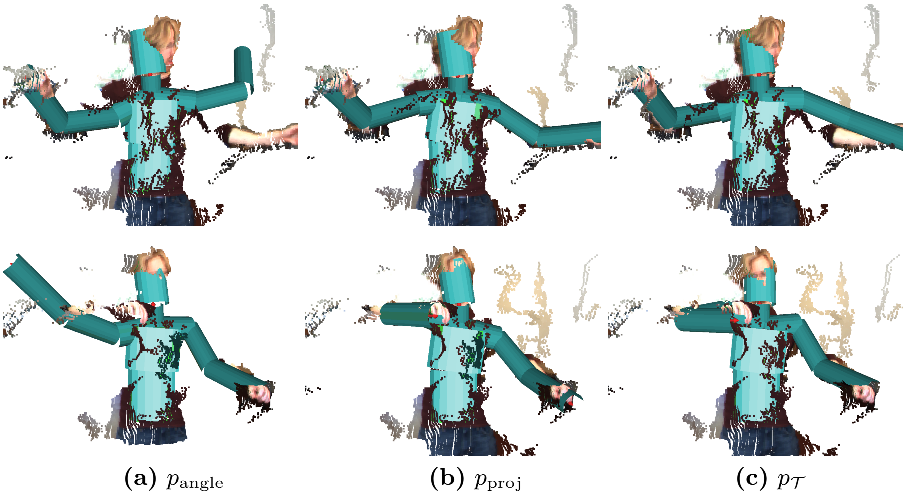

(a) p_{angle}
(b) p_{proj}
(c) $p_{\mathcal{T}}$

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

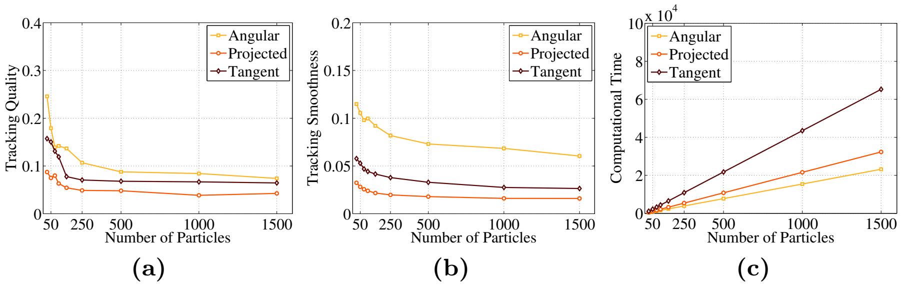


Fig. 6. Performance of the three priors. All reported numbers are averaged over several trials. (a) The error measure $\mathcal{E}$ as a function of the number of particles. The average standard deviation of $\mathcal{E}$ with respect to the trials are 0.021 for the angular prior, 0.007 for the projected 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)