Articulated Motion Estimation From a Monocular Image Sequence Using Spherical Tangent Bundles

Spyridon Leonardos, Xiaowei Zhou and Kostas Daniilidis

Abstract— **We propose a second order stochastic dynamical model for generic articulated objects whose state space is a Riemannian manifold naturally suggested by the articulation constraints. We derive the equations of a Riemannian Extended Kalman Filter to perform the structure estimation from an image sequence captured by a perspective camera. In order to theoretically validate our approach, we prove that the proposed model is locally weakly observable. Finally, we report quantitative results on both synthetic data and on real sequences from the CMU Mocap dataset.**

I. INTRODUCTION

Articulated motion reconstruction from a monocular video sequence is a central problem of geometric computer vision and robotics which has received an increasing amount of attention during the last decade. It has widespread applications in computer vision, robotics and graphics since many objects such as robots and humans can be categorized as articulated objects.

Articulated bodies are described by a graph structure, usually a tree, termed the *articulation graph*, where each node represents a joint and each edge between nodes in the tree corresponds to a link between joints. Links have constant length at all time instances. We refer to such a constant distance constraint as an *articulation constraint*. Our goal is to estimate the 3D trajectories of the joints from noisy projections assuming a full perspective projection model.

The main challenge in this problem arises from the intrinsic ambiguity in monocular vision: a 3D point may lie at any position along the ray connecting the camera center and the 2D projection in the image plane. Even with the articulation constraint, there are two distinct solutions for the 3D position of a node relative to its parent node at each time instance.

In this paper, we develop a sequential approach for articulated motion estimation from a monocular image sequence. The proposed algorithm takes as input a sequence of noisy projections of the joints of an articulated body and outputs estimates of their 3D positions. We model the variation of the angular velocity of a joint as a zero-mean Gaussian noise process and propose a second order stochastic dynamical model whose state space is a Riemannian manifold naturally suggested by the articulation constraints.

Our contributions are mainly twofold: (1) We derive the equations of a Riemannian Extended Kalman Filter (REKF) which is a non-trivial generalization of the standard Extended Kalman Filter (EKF) [4]. Unlike most previous methods assuming a weak perspective camera model, we use a full perspective camera model. (2) We provide a theoretical analysis to show that the proposed model is locally observable and explain the choice of second order dynamics.

The remainder of this paper is organized as follows. First, we make a review of the prior art in Section II. Then, we review preliminaries from differential and Riemannian geometry in Section III. We present the proposed dynamical model in Section IV. Next, we describe in detail the implementation of the filter in Section V. Observability of the proposed model is discussed in Section VI. Finally, we present the experimental evaluation in Section VII.

II. RELATED WORK

There have been a few efforts towards articulated motion estimation from a monocular image sequence. In their seminal paper, Bregler and Malik [8] parametrized the human body using the product of exponentials for kinematic chains [23] and proposed a tracking algorithm for the case of scaled orthographic projection model. Products of exponentials for articulated motion reconstruction were also used by Drummond and Cipolla [13] who efficiently solved an inference problem by propagating the statistics between one adjacent pair of links at a time. A serious drawback of this approach is the potential error accumulation at the end of the kinematic chain. In addition, none of the two above described methods investigate the problem of multiple inverse kinematics solutions. Several methods of investigating alternative solutions were proposed by Sminchisescu and Triggs [29], [30]. Although the proposed methods are designed to avoid getting stuck in the wrong local minimum forever, there are no theoretical guarantees that they converge to the correct minimum either. The latter techniques can be used on top of a tracking system such as the one we propose in this work. A detailed discussion on the ambiguities and singularities of visual tracking of articulated objects can be found in [26].

More recently, Park and Sheikh [25] proposed a predefined trajectory basis approach in which they incorporate articulation constraints. Given the 3D position of a node in the articulation tree, the 2D projection of a child node on the image plane and the distance between them in 3D, then there are, in general, two distinct solutions for the 3D position of the child node. To model this two-fold ambiguity, they introduced a binary variable for each node in each frame. The resulting problem is a combinatorial optimization problem which is in general NP-hard to solve. They solve it

The authors are with the Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA. {spyridon,xiaowz,kostas}@seas.upenn.edu

Support by the grants NSF-DGE-0966142, NSF-IIP-1439681, NSF-IIS-1426840, ARL MAST-CTA W911NF-08-2-0004, and ARL RCTA W911NF-10-2-0016 is gratefully acknowledged.

using branch-and-bound techniques which are in general of exponential complexity. To address the computational burden of the branch-and-bound optimization, Valmadre *et al.* [34] proposed a dynamic programming approach combining articulation constraints with temporal smoothness. But the success of these methods rely on the representability of the predefined basis shapes and unrealistically fast camera motion. Moreover, they require that all frames are available at once and not suited for sequential processing.

Apart from optimization-based approaches for articulated motion estimation, there have been numerous probabilistic ones. Brookshire and Teller [9] proposed a particle filtering approach that takes into account the non-linear articulation constraints. However, it is assumed that the state of the articulated body can be uniquely determined by available observations. This is not the case with perspective projection observational model as we shall shortly explain. Particle filtering approaches have been used in the past [11] for the purpose of human motion tracking. Hauberg *et al.* [18] proposed Gaussian priors for limb positions either on the embedding space of the so-called pose manifold followed by a projection operation or directly on the tangent space of the pose manifold. Another line of works includes learning-based techniques for human pose estimation [28], [32], [33]. However, most of the above mentioned approaches require either 3D data as input or training data, whereas the proposed method does not require any training apart from tuning the filter and takes as input direct image data. More importantly, we aim to develop a framework applicable to any articulated objection instead of being specific to humans.

Finally, our work is relevant to a large body of literature on nonrigid SFM. A non-exhaustive list of this body of work includes approaches based on shape bases [7], trajectory bases [3], probabilistic principal component analysis [31], rank minimization [10], smoothness priors [24], trajectory grouping [14] and sparse coding [35]. However, none of the above methods incorporate articulation constraints.

III. PRELIMINARIES AND NOTATION

In this section, we briefly review several elementary facts from Riemannian geometry. For a more detailed and rigorous treatment, we refer the reader to [12], [15], [21].

A *Riemannian manifold* (M, g) is a manifold whose tangent spaces are equipped with a smoothly varying inner product, termed the *Riemannian metric*. We use $g(\xi, \zeta)$ to denote the inner product of two vectors ξ, ζ in the tangent space of $\mathcal{M}$ at x, denoted by $T_x\mathcal{M}$. An *affine connection* ∇ generalizes the concept of usual directional differentiation of vector fields. Given two vector fields ξ, ζ on $\mathcal{M}$, the covariant derivative $\nabla_\xi \zeta$ expresses the change of ζ in the direction of ξ. The *acceleration vector field* $\frac{D^2}{dt^2}\gamma$ on γ is defined by $\frac{D^2}{dt^2}\gamma(t) \doteq \nabla_{\dot\gamma(t)}\dot\gamma(t)$. A *geodesic curve* on $\mathcal{M}$ is the generalization of a straight line, that is, a curve with zero acceleration. We denote by $\gamma_{x,\xi}(t)$ the geodesic emanating from x in the direction of $\xi \in T_x\mathcal{M}$. The *exponential map* $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ is defined as $\exp_x(\xi) \doteq \gamma_{x,\xi}(1)$. The *logarithm map* $\log_x : \mathcal{M} \to T_x\mathcal{M}$ is the inverse of the exponential map and is generally defined only in a neighborhood of x.

A vector field ξ along a curve $\gamma(t)$ is said to be *parallel* if $\nabla_{\dot\gamma(t)}\xi_{\gamma(t)} = 0$ for every t. Given $t_0 \in \mathbb{R}$ and $\xi_0 \in T_{\gamma(t_0)}\mathcal{M}$ there is a unique parallel vector field ξ on γ such that $\xi_{\gamma(t_0)} = \xi_0$. The mapping $P_\gamma^{t\leftarrow t_0}\xi$ sending $\xi_{\gamma(t_0)}$ to $\xi_{\gamma(t)}$ is called *parallel transport along* γ. Given three vector fields ξ, ζ, η on a Riemannian manifold $\mathcal{M}$, the Riemannian curvature tensor R is defined by $R(\xi, \zeta)\eta = \nabla_\zeta \nabla_\xi \eta - \nabla_\xi \nabla_\zeta \eta + \nabla_{[\xi,\zeta]}\eta$.

Let $F : \mathcal{M} \to \mathcal{N}$ be a smooth map between manifolds $\mathcal{M}$ and $\mathcal{N}$. The linear mapping $DF(x) : T_x\mathcal{M} \to T_{F(x)}\mathcal{M} : \xi \mapsto DF(x)[\xi]$ is called the *differential* of F at x. For any curve $\gamma(t)$ on $\mathcal{M}$ we have $DF(\gamma(t))[\dot\gamma(t)] = dF(\gamma(t))/dt$. The *Jacobian* is the matrix representation of the differential in local coordinates. Finally, since manifolds of interest have a natural embedding in a Euclidean space, we write $DF(x)$ for the matrix representation of the differential of F at x, in the coordinates of the embedding space.

In the context of this work, we heavily use the spherical tangent bundle $T\mathbb{S}^{n-1}$ of the unit sphere $\mathbb{S}^{n-1}$ of $\mathbb{R}^n$. In general, the *tangent bundle* of a manifold $\mathcal{M}$ is the set $T\mathcal{M} = \{(x, \xi) : x \in \mathcal{M}, \xi \in T_x\mathcal{M}\}$ and is a manifold of double dimension. The *vertical space* $\mathcal{V}_{(x,\xi)}$ is the linear subspace of $T_{(x,\xi)}T\mathcal{M}$ given by $\mathcal{V}_{(x,\xi)} = \ker(D\pi(x, \xi))$, where π is the *canonical projection* onto the first component. The *horizontal space* is the orthogonal complement of $\mathcal{V}_{(x,\xi)}$ in $T_{(x,\xi)}T\mathcal{M}$. Intuitively, horizontal curves in the tangent bundle $T\mathcal{M}$, *i.e.* curves with horizontal tangents, correspond to parallel vector fields on $\mathcal{M}$ and vertical curves correspond to curves on $T_x\mathcal{M}$. Each tangent vector in $T_{(x,\xi)}TM$ can be uniquely decomposed as the sum of its horizontal and vertical components $\zeta^h + \eta^v$ where ζ^h, η^v are, respectively, the horizontal and vertical lifts of $\zeta, \eta \in T_x\mathcal{M}$. Therefore, we represent a tangent vector $\zeta^h + \eta^v \in T_{(x,\xi)}T\mathcal{M}$ by a pair $(\zeta, \eta) \in T_x\mathcal{M} \times T_x\mathcal{M}$. A metric for $T\mathbb{S}^{n-1}$ can be naturally defined from the metric of $\mathbb{S}^{n-1}$. A particular choice is the so-called Sasaki metric [27]. For a rigorous treatment of the geometry of tangent bundles, we refer the reader to [16].

IV. PROPOSED MODEL

In this section, we present a second order stochastic dynamical model for articulated systems that enables us to use tools from estimation theory and, in particular Kalman filtering [20], to solve the estimation problem at hand. The first issue one encounters is that the exact dynamics are unknown. However, the angular velocity of joints does not vary arbitrarily from frame to frame. We compensate for the acceleration of joints by employing a statistical model. The second obstacle is the non-linearity of the state space. Joint positions and velocities do not lie on a linear space. To model this nonlinearity, we parametrize the configuration space as a properly chosen Riemannian manifold.

For simplicity, we present our proposed model for the case of a single kinematic chain. Generalization for the case of kinematic trees is straightforward. Consider the articulated chain of Fig.1. The articulation graph is in this

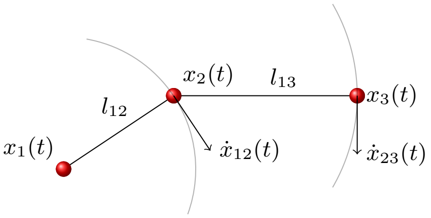

Fig. 1: Example of an articulation chain with three joints with positions $x_1(t)$, $x_2(t)$ and $x_3(t)$.

case $\mathcal{G} = (\mathcal{V}, \mathcal{E}) = (\{1, 2, 3\}, \{(1, 2), (2, 3)\})$. We assume that the root position $x_1(t)$ can be measured directly with some uncertainty[1]. We model $x_1(t)$ as noisy input and not as part of the state vector. Moreover, we assume that the joint lengths are known. These are standard assumptions which are adopted by competitive methods, *e.g.* [25], [34] as well. The articulation constraints read

$$\|x_i(t) - x_j(t)\|_2 = l_{ij} = \text{const.} \tag{1}$$

for all $(i, j) \in \mathcal{E}$. We use the shorthand notation $x_{ij} \doteq \frac{x_j - x_i}{\|x_j - x_i\|_2} \in \mathbb{S}^2$ for the orientation of the link between joints i and j. Under our assumptions, the position of the second joint $x_2(t)$ is uniquely determined by the orientation $x_{12}(t) \in \mathbb{S}^2$. In addition, we include second order information in our dynamical system to improve the accuracy of our model. The velocity of the second joint can be readily computed as $\dot{x}_2(t) = \dot{x}_1(t) + l_{12}v_{12}(t)$, where $v_{12}(t) \doteq \dot{x}_{12}(t)$. Thus, the velocity of $x_2(t)$ is uniquely determined by the angular velocity $v_{12}(t) \in T_{x_{12}(t)}\mathbb{S}^2$. Naturally, the state manifold for a single link is $T\mathbb{S}^2$ and $(T\mathbb{S}^2)^n$ for the case of n links.

It remains to write the dynamics of the system. Since the angular velocity of a joint does not change arbitrarily from frame to frame and the exact dynamics are unknown, we employ a statistical model for its time derivative, that is

$$(x_{ij}(t{+}1), v_{ij}(t{+}1)) = \exp_{(x_{ij}(t), v_{ij}(t))}(v_{ij}(t), w_{ij}(t)) \tag{2}$$

for all $(i, j) \in \mathcal{E}$, where $\exp$ denotes the exponential map of $T\mathbb{S}^2$, $(x_{ij}, v_{ij}) \in T\mathbb{S}^2$ and $w_{ij}(t) \in T_{x_{ij}(t)}$ is zero-mean isotropic Gaussian noise. Intuitively, when $w_{ij}(t) = 0$, x_{ij} travels on a geodesic of the sphere and $v_{ij}(t)$ is a parallel vector field along x_{ij}. In general, the exponential map of $T\mathbb{S}^2$ has no closed form expression. However, in the noiseless case, *i.e.* when $w_{ij}(t) = 0$, it is given by

$$\begin{aligned} x_{ij}(t+1) &= \exp_{x_{ij}(t)}(v_{ij}(t)) \\ v_{ij}(t+1) &= P_{x_{ij}}^{t+1 \leftarrow t} v_{ij}(t) \end{aligned} \tag{3}$$

for all $(i, j) \in \mathcal{E}$. In other words, the the exponential map of $T\mathbb{S}^2$ for the case of zero vertical component, can be computed from the exponential map and the parallel transport along geodesics of $\mathbb{S}^2$ which both have closed form expressions (see Appendix B).

Finally, the measurement model for the set of joints, excluding the root of the articulation chain, is given by the standard perspective projection model

$$y_i(t) = \pi(x_i(t)) + \eta_i(t), \quad i \in \mathcal{V} \setminus \{1\} \tag{4}$$

where $\eta_i(t)$ is a zero-mean isotropic Gaussian noise process modeling the uncertainty of the joint localization and $\pi(x)$ is the standard perspective function $\pi \colon \mathbb{R}^3 \to \mathbb{R}^2$. For the measurement of the root, we assume isotropic zero-mean Gaussian noise as well.

V. RIEMANNIAN EXTENDED KALMAN FILTER

In this section, we derive the equations for the Riemannian Extended Kalman Filter (REKF), which, as we shall shortly see, is a generalization of the widely used Extended Kalman Filter (EKF) [4]. Extended Kalman Filters for Riemannian manifold possessing a Lie group structure have been described in [5], [6]. Unfortunately, they are not applicable in the current setting because the state manifold does not have a Lie group structure. For arbitrary Riemannian manifolds, the Uscented Kalman Filter (UKF) was introduced by Hauberg *et al.* [17]. In theory, the UKF of Hauberg *et al.* [17] is applicable in our case. However, in practice, the logarithm computation for $T\mathbb{S}^2$ requires to globally solve an optimization problem on $T\mathbb{S}^2$ for all sigma points at every UKF iteration, which would make our approach computationally inefficient and impractical.

Intuitively, we linearize the dynamical model at the tangent space of the current estimate. The measurement update takes place on the tangent space of the current estimate and then, we use the exponential map to obtain the update on the manifold. The covariance of the estimation error is propagated by parallel transporting its eigenvectors from the tangent space of the predicted estimate to the tangent space of the updated estimate. To achieve this procedure, we develop a numerical method for computing the exponential map of $T\mathbb{S}^2$ and the parallel transport of a tangent vector along the corresponding geodesic of $T\mathbb{S}^2$.

Unfortunately, geodesics of $T\mathbb{S}^2$ do not have a closed form expression. Therefore, an one step Euler integration method, proposed by Muralidharan and Fletcher [22], is employed. We generalize it, in order to compute the parallel transport of a tangent vector along the same geodesic.

Proposition 5.1: The differential equations of a geodesic curve on $T\mathbb{S}^2$, equipped with the Sasaki metric, emanating from $(x, \xi) \in T\mathbb{S}^2$ in the direction of $(\zeta, \eta) \in T_{(x,\xi)}T\mathbb{S}^2$ read

$$\nabla_\zeta \zeta = -R(\xi, \eta)\zeta \tag{5}$$

$$\nabla_\zeta \eta = 0 \tag{6}$$

where ∇ is the connection of $\mathbb{S}^2$ compatible with its standard metric. The differential equations for parallel transport of a tangent vector (μ, ν) along the same geodesic read

$$\nabla_\zeta \mu = -\frac{1}{2}R(\xi, \nu)\zeta - \frac{1}{2}R(\xi, \eta)\mu \tag{7}$$

$$\nabla_\zeta \nu = \frac{1}{2}R(\zeta, \mu)\xi \tag{8}$$

[1] For example, the root position for a human skeleton can be estimated by structure from motion algorithms assuming that the torso is rigid, and the root position for a robot arm can be measured offline if the base is fixed.

Detailed proof of Proposition 5.1 is included in Appendix C. Based on Proposition 5.1, we compute the geodesic and the parallel transport using one step Euler integration. Specifically, $\exp_{(x_0,\xi_0)}(\zeta_0,\eta_0)$ and the parallel transport of $(\mu_0,\nu_0) \in T_{(x_0,\xi_0)}\mathbb{S}^2$ from (x_0,ξ_0) to $\exp_{(x_0,\xi_0)}(\zeta_0,\eta_0)$ along the corresponding geodesic are computed using the following iterative scheme for $k = 0, 1, \dots, N-1$

$$x_{k+1} = \exp_{x_k} \epsilon\zeta_k \tag{9}$$

$$\xi_{k+1} = P_x^{k+1\leftarrow k}\xi_k + \epsilon\eta_k \tag{10}$$

$$\zeta_{k+1} = P_x^{k+1\leftarrow k}\zeta_k - \epsilon R(\xi_k,\eta_k)\zeta_k \tag{11}$$

$$\eta_{k+1} = P_x^{k+1\leftarrow k}\eta_k \tag{12}$$

$$\mu_{k+1} = P_x^{k+1\leftarrow k}\mu_k - \frac{\epsilon}{2}(R(\xi_k,\nu_k)\zeta_k + R(\xi_k,\eta_k)\mu_k) \tag{13}$$

$$\nu_{k+1} = P_x^{k+1\leftarrow k}\nu_k + \frac{\epsilon}{2}R(\zeta_k,\mu_k)\xi_k \tag{14}$$

where $\epsilon = 1/N$. Note that the exponential map, the parallel transport and the Riemannian curvature tensor in the above equations are all operations of $\mathbb{S}^2$ with closed form expressions [2], [22].

Finally, the following proposition from [17] provides us with a method of parallel transporting covariances along geodesics of a Riemannian manifold $\mathcal{M}$:

Proposition 5.2: Let $\gamma_x(t)$ be a geodesic of the Riemannian manifold $\mathcal{M}$ with $\gamma_x(0) = x$. Let $\{v_1,\dots,v_m\}$ is an orthonormal basis for the tangent space $T_x\mathcal{M}$. If $v_i(t) \doteq P_{\gamma_x}^{t\leftarrow 0}v_i$, then the parallel transport of the symmetric bilinear form with eigendecomposition $A = \sum_{i=1}^m \lambda_i v_i v_i^T$ is given by $P_{\gamma_x}^{t\leftarrow 0}A \doteq \sum_{i=1}^m \lambda_i v_i(t) v_i(t)^T$.

In the context of this work, we consider discrete time dynamical systems which evolve on a Riemannian manifold $\mathcal{M}$ and are of the form

$$\begin{aligned} x(t+1) &= \exp_{x(t)}\left(\log_{x(t)}(f(x(t))) + w(t)\right) \\ y(t) &= h(x(t), u(t) + \nu(t)) + \eta(t) \end{aligned} \tag{15}$$

where $f\colon \mathcal{M} \to \mathcal{M}$ corresponds to the dynamical model in the absence of process noise, $w(t)$ is the process noise on the tangent space $T_{x(t)}\mathcal{M}$, $\eta(t)$ is the measurement noise and $\nu(t)$ is the additive noise of the (unknown) input $u(t)$. The Riemannian Extended Kalman Filter equations follow:

- **Linearization:**

$$\begin{cases} F(t) = Df(\widehat{x}(t|t)) \\ C(t) = D_x h(\widehat{x}(t|t-1), u(t) + \nu(t)) \\ D(t) = D_u h(\widehat{x}(t|t-1), u(t) + \nu(t)) \end{cases} \tag{16}$$

- **Update:**

$$\begin{cases} \widehat{x}(t|t) = \exp_{\widehat{x}(t|t-1)} L(t)(y(t) - \widehat{y}(t)) \\ \Omega(t) = C(t)\Sigma(t|t-1)C(t)^T \\ \qquad + D(t)\Sigma_\nu(t)D(t)^T + \Sigma_\eta(t) \\ L(t) = \Sigma(t|t-1)C(t)^T\Omega(t)^{-1} \\ \Sigma(t|t) = P_{\widehat{x}(t|\cdot)}^{t\leftarrow t-1}(I - L(t)C(t))\Sigma(t|t-1) \end{cases} \tag{17}$$

- **Prediction:**

$$\begin{cases} \widehat{x}(t+1|t) = f(\widehat{x}(t|t)) \\ \Sigma(t+1|t) = F(t)\Sigma(t|t)F(t)^T + \Sigma_w(t+1) \end{cases} \tag{18}$$

VI. Observability Analysis

In this section, we discuss the observability of the proposed model. The observability problem refers to whether the initial condition of a dynamical system is uniquely determined by a sequence of measurements. Observability concepts have been extensively used for analyzing the performance of Kalman filters. In the first subsection, we show that the proposed model is locally weakly observable and in the second subsection, we discuss the limitations of a first order model. We present analysis for the case of an articulation chain with two joints. Generalizing the following results for more than two joints can be trivially done in a recursive fashion.

A. Local weak observability of articulated motion

Given a nonlinear continuous time dynamical system with state vector $x \in \mathbb{R}^n$, two initial conditions are *indistinguishable* if they produce the same output for all time instances and inputs. A system is *observable at* $x_0 \in \mathbb{R}^n$ if x_0 is not indistinguishable from any other point. This type of observability is sometimes referred to as *global observability*. However, for nonlinear systems global observability is too much to ask for and hard to prove in general. For this reason, the notion of *local weak observability* was introduced in [19]. A system is *locally weakly observable* at a point $x_0 \in \mathbb{R}^n$ if there exists a neighborhood U of x_0 such that x_0 is not indistinguishable from any of its neighbors in U.

At this point, we state the main result of this section regarding the observability of the proposed model.

Theorem 6.1: The dynamical model (3) with observational model (4) is locally weakly observable at a configuration $(x_{12}(0), v_{12}(0)) \in T\mathbb{S}^2$ when the ray from the camera center to the second joint is not tangential to the sphere centered at the first joint and having radius l_{12}. If the ray is tangential, then $x_{12}(0)$ has a unique global solution but the initial velocity $v_{12}(0)$ is not uniquely determined from the projection of the second joint and its time derivative (optical flow) at $t = 0$.

A detailed proof of Theorem 6.1 can be found in Appendix A. The results of the theorem hold for any dynamical model including both link orientations and angular velocities. If a model makes further assumptions, *e.g.* zero angular acceleration as in our case, then the model can be proved to be locally weakly observable everywhere. Since observability is a property of a model not of the underlying physical system, we keep the conditions of Theorem 6.1 as general as possible.

B. The two-fold ambiguity

One natural question is whether the model is globally observable apart from locally observable. However, it is clear that global observability is too much to ask for since, in general, there are two distinct solutions for the 3D position of a node relative to its parent node in the articulation tree at each time instance

The two-fold ambiguity has been well recognized in the computer vision community [30], [26], [25], [34]. Recently,

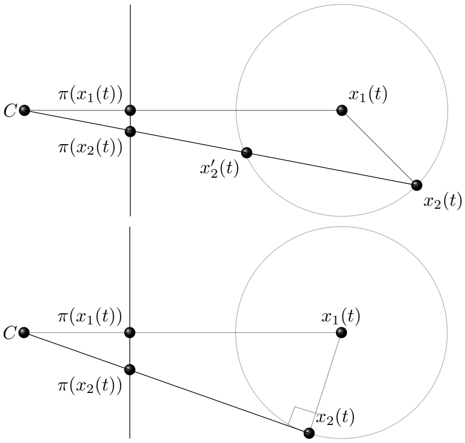

Fig. 2: Illustration of the two-fold ambiguity. A unique solution exists when the ray from the camera to x_2 is tangential to the sphere of radius l_{12} centered at x_1.

temporal smoothness was proposed to resolve this ambiguity [25], [34]. However, there is no guarantee that the true 3D trajectory of a joint is the smoothest among all trajectories having identical image projections. For sequential estimation, the temporal smoothness prior corresponds to a first order dynamical model. In practice, we have observed that 3D trajectories of joints producing identical image projections may eventually intersect at some time instance. Even with accurate initialization, it would impossible to disambiguate near the time instances of intersection and tracking would possibly fail for the case of a first order model. However, augmenting the state space to include higher order information resolves the problem, in general, since the trajectories of the augmented system no longer intersect. Finally, in the proposed framework, all articulation constraints are considered at once, rather than in a recursive fashion as in [13], [25], [34], which has the advantage of significantly reducing the number of possible trajectories and the error accumulation at the end of an articulated chain.

VII. EXPERIMENTS

A. Experiments on synthetic data

We use synthetic data to demonstrate the convergence rate of our filter and to test its robustness against input noise (error of the root position in 3D) and measurement noise (error of the joint position in 2D). We use an articulation chain consisting of three joints and of unit length links. The root joint is static at position $(0, 0, 5)$. To show convergence, we perturb the initial orientations of the links around their true values by an angle value sampled from a zero-mean Gaussian distribution with standard deviation equal to 45^o and uniform direction. We repeat the experiment 200 times. Results are presented in Fig. 3. Convergence is achieved in less than 5 iterations even for very rough initialization. Robustness against input noise, *i.e.* error in the position of the root, and against measurement noise (projection error) is presented, respectively, in Fig. 4 and Fig. 5. The experiment is repeated 50 times for each value of σ_ν^2 in m^2 and σ_η^2 in square focal length units. The implemented filter produces accurate results for root position error up to the order of centimeters and for projection error up to a few pixels or equivalently up to 10^{-6} to 10^{-5} square focal length units.

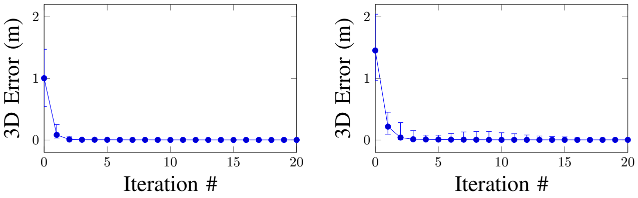

Fig. 3: 3D Error in meters of the second joint (left) and the third joint. Noise parameters: $\sigma_w^2 = 10^{-6}$, $\sigma_\eta^2 = 10^{-8}$, $\sigma_\nu^2 = 10^{-6}$.

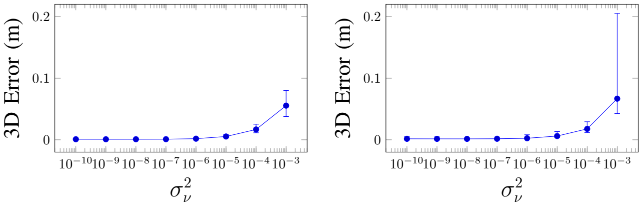

Fig. 4: 3D Error in meters of the second joint (left) and the third joint as a function of input noise (root measurement) variance. Other noise parameters include $\sigma_w^2 = 10^{-6}$, $\sigma_\eta^2 = 10^{-8}$.

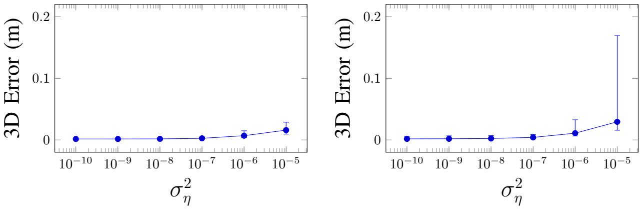

Fig. 5: 3D Error in meters of the second joint (left) and the third joint as a function of measurement noise variance. Other noise parameters include $\sigma_w^2 = 10^{-6}$, $\sigma_\nu^2 = 10^{-6}$.

B. Human pose reconstruction

We validate our method with the CMU motion capture datasets [1] and compare it with the alternative method [25]. The 2D observations are synthesized by projecting 3D joints to 2D with a perspective camera model. Similar to previous work [25], [34], the camera parameters and trajectories of the rigid body (marker set $\{1, 2, 5, 8, 10, 13\}$, see Fig. 7) are provided to the algorithms. We measure the 3D error in meters for the non-rigid part of the human body, *i.e.* maker set $\{3, 4, 6, 7, 11, 12, 14, 15\}$. We use 10 sequences of 500 frames in total. Specifically we use sequences #2-#11 of subject #15 [1] which include a variety of actions such as walking, dancing, hand signals. We perturb the ground truth using isotropic zero-mean Gaussian noise with large standard deviation, *i.e.* 10cm, to initialize our filter. Our method significantly outperforms the state of the art method [25] in all sequences. The achieved frame rate is about 10 frames per second for this dataset.

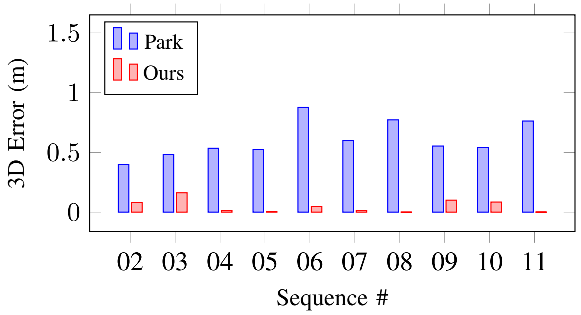

Fig. 6: Left: Comparison of our method with [25]. Average 3D error in meters over joints $\{3, 4, 6, 7, 11, 12, 14, 15\}$.

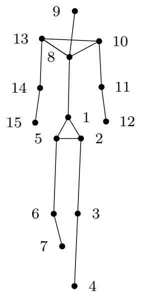

Fig. 7: Human model of CMU Mocap databaset.

VIII. CONCLUSIONS

In this paper, we presented a second order stochastic dynamical model for articulated motion. We demonstrated the usefulness of including second order information and we proved the local observability of the proposed model. Finally, we derived the equations of a REKF and verified experimentally its performance compared to a state of the art method.

APPENDIX

A. *Proof of Theorem 6.1*

We prove Theorem 6.1 for the case of two joints. Let $x_1(t) \in \mathbb{R}^3$ denote the trajectory of the root joint, which is modeled as an input to the proposed dynamical model, and $(x_{12}(t), v_{12}(t)) \in T\mathbb{S}^2$ denote the state vector. The projection of the second joint $y_2(0)$, its time derivative $\dot{y}_2(0)$ and the implicit constraints defining $T\mathbb{S}^2$ as a subset of $\mathbb{R}^6$ read

$$x_{12}(0)^T x_{12}(0) - 1 = 0 \tag{19}$$

$$x_{12}(0)^T v_{12}(0) = 0 \tag{20}$$

$$\pi(x_1(0) + l_{12}x_{12}(0)) - y_2(0) = 0 \tag{21}$$

$$\left.\frac{d\pi(x_1(t) + l_{12}x_{12}(t))}{dt}\right|_{t=0} - \dot{y}_2(0) = 0 \tag{22}$$

The above constraints define a map $\Theta\colon \mathbb{R}^6 \to \mathbb{R}^6$. To show that the model is locally weakly observable at a point $(x_{12}(0), v_{12}(0))$, it suffices to show that Θ is locally invertible (locally a diffeomorphism). By properly rearranging the constraints, it is not hard to see that the Jacobian $\mathcal{O} \doteq D\Theta \in \mathbb{R}^{6\times6}$ is lower block triangular of the form

$$\mathcal{O} = \begin{bmatrix} \mathcal{O}_1 & \mathbf{0}_{3\times3} \\ \mathcal{O}_2 & \mathcal{O}_1 \end{bmatrix} \tag{23}$$

with

$$\mathcal{O}_1 = \begin{bmatrix} 2x_{12}(0)^T \\ \dfrac{\partial\pi(x_1(0) + l_{12}x_{12}(0))}{\partial x_{12}(0)} \end{bmatrix} \in \mathbb{R}^{3\times3} \tag{24}$$

Matrix $\mathcal{O}$ has full rank if and only if the submatrix $\mathcal{O}_1 \in \mathbb{R}^{3\times3}$ has full rank since $\det(\mathcal{O}) = \det(\mathcal{O}_1)^2$. But A is rank deficient if and only if the nullspace of $x_{12}(0)$ and $\frac{\partial\pi(x_1(0)+l_{12}x_{12}(0))}{\partial x_{12}(0)}$ have a non-trivial intersection, *i.e.*

$$x_{12}(0)^{\perp} \cap \operatorname{span}\{x_1(0) + l_{12}x_{12}(0)\} = \{0\} \tag{25}$$

or equivalently if the ray from the camera to the second joint is not contained in the plane tangent to the sphere of radius l_{12} centered at the first joint. In this case, Θ is a local diffeomorphism.

In the opposite case, we have a unique global solution for $x_{12}(0)$ which can be computed from the point of tangency. However, velocity $v_{12}(0)$ is not uniquely determined since any velocity of the form $v_{12}(0) + cx_2(0)$ for some scalar c yields the same optical flow of the second joint.

B. *Linearization of the model*

Before proceeding to the linearization of the dynamical model, we need a remark concerning the representation of tangent vectors of $T\mathbb{S}^2$ in a computer program. Let $z_h(t) = (x(t), P_x^{t\leftarrow0}v(0)) \in T\mathbb{S}^2$ any horizontal curve on the tangent bundle $T\mathbb{S}^2$. Its time derivative at $t = 0$ is given by

$$\dot{z}_h(0) = (\dot{x}(0), -x(0)v(0)^T\dot{x}(0)) = \dot{x}(0)^h \tag{26}$$

where we used the analytic expression for the parallel transport along geodesics of the sphere (see [2]). Now, let $z_v(t) = (x(0), u(t)) \in T\mathbb{S}^2$ any vertical curve. We have that

$$\dot{z}_v(0) = (0, \dot{u}(0)) = \dot{u}(0)^v \tag{27}$$

Thus, any vector $(\dot{x}(0), \dot{v}(0)) \in T\mathbb{S}^2$ tangent to a curve $(x(t), v(t)) \in T\mathbb{S}^2$ has the form

$$\begin{bmatrix} \dot{x}(0) \\ \dot{v}(0) \end{bmatrix} = \begin{bmatrix} I & 0 \\ -x(0)v(0)^T & I \end{bmatrix} \begin{bmatrix} \dot{x}(0) \\ \dot{u}(0) \end{bmatrix} \tag{28}$$

where $\dot{u}(0)$ is the component of $\dot{v}(0)$ in $T_{x(0)}\mathbb{S}^2$. In a computer program, we do not use $(\dot{x}(0), \dot{v}(0))$ but $(\dot{x}(0), \dot{u}(0))$ which are both in $T_x\mathbb{S}^2$.

Now, let the map $f\colon T\mathbb{S}^2 \to T\mathbb{S}^2$ defined by

$$\begin{bmatrix} x \\ v \end{bmatrix} \mapsto \begin{bmatrix} \cos(\|v\|)x + \operatorname{sinc}(\|v\|)v \\ -\|v\|\sin(\|v\|)x + \cos(\|v\|)v \end{bmatrix} \tag{29}$$

which is the proposed dynamical model in the absence of noise. By taking the time derivative of $f(x(t), v(t))$ at $t = 0$, where $(x(t), v(t)) \in T\mathbb{S}^2$ is a smooth curve, one can easily compute the matrix $A(x(0), v(0))$ such that

$$\left.\frac{df(x(t), v(t))}{dt}\right|_{t=0} = A(x(0), v(0)) \begin{bmatrix} \dot{x}(0) \\ \dot{v}(0) \end{bmatrix} \tag{30}$$

Based on the previous discussion, the matrix representation of the differential of f at a point $(x, v) \in T\mathbb{S}^2$ is

$$Df(x, v) = \begin{bmatrix} I & 0 \\ xv^T & I \end{bmatrix} A(x, v) \begin{bmatrix} I & 0 \\ -xv^T & I \end{bmatrix} \qquad (31)$$

C. Proof of Proposition 5.1

First, we need the following proposition from [16].

Proposition 8.1: Let $(\mathcal{M}, g)$ be a Riemannian manifold with Riemannian curvature tensor R and $\overline{\nabla}$ denote the Levi-Civita connection of the tangent bundle $(T\mathcal{M}, \overline{g})$ equipped with the Sasaki metric. Then, for all $\zeta, \eta \in \mathfrak{X}(\mathcal{M})$ we have the following

$$\begin{cases} (\overline{\nabla}_{\zeta^h} \eta^h)_{(x,\xi)} = (\nabla_\zeta \eta)^h_{(x,\xi)} - \dfrac{1}{2}(R(\zeta, \eta)\xi)^v \\ (\overline{\nabla}_{\zeta^h} \eta^v)_{(x,\xi)} = (\nabla_\zeta \eta)^v_{(x,\xi)} + \dfrac{1}{2}(R(\xi, \eta)\zeta)^h \\ (\overline{\nabla}_{\zeta^v} \eta^h)_{(x,\xi)} = \dfrac{1}{2}(R(\xi, \zeta)\eta)^h \\ (\overline{\nabla}_{\zeta^v} \eta^v)_{(x,\xi)} = 0. \end{cases} \qquad (32)$$

At this point, we can proceed to the main proof using Proposition 8.1. The differential equation for the geodesics of $T\mathbb{S}^2$ is $\overline{\nabla}_{\zeta^h + \eta^v}(\zeta^h + \eta^v) = 0$ and for the parallel transport $\overline{\nabla}_{\zeta^h + \eta^v}(\mu^h + \nu^v) = 0$. By linearity of the connection and Proposition 8.1, we get

$$\begin{aligned} \overline{\nabla}_{\zeta^h + \eta^v}(\zeta^h + \eta^v) &= \overline{\nabla}_{\zeta^h}\zeta^h + \overline{\nabla}_{\zeta^h}\eta^v + \overline{\nabla}_{\eta^v}\zeta^h + \overline{\nabla}_{\eta^v}\eta^v \\ &= (\nabla_\zeta \zeta + R(\xi, \eta)\zeta)^h + (\nabla_\zeta \eta)^v \end{aligned}$$

which proves the first claim. Moreover, we have

$$\begin{aligned} \overline{\nabla}_{\zeta^h + \eta^v}(\mu^h + \nu^v) &= \overline{\nabla}_{\zeta^h}\mu^h + \overline{\nabla}_{\zeta^h}\nu^v + \overline{\nabla}_{\eta^v}\mu^h + \overline{\nabla}_{\eta^v}\nu^v \\ &= (\nabla_\zeta \mu + \frac{1}{2}R(\xi, \nu)\zeta + \frac{1}{2}R(\xi, \eta)\mu)^h \\ &\quad + (\nabla_\zeta \nu - \frac{1}{2}R(\zeta, \mu)\xi)^v \end{aligned}$$

which proves the second claim.

ACKNOWLEDGMENT

We thank Hyun Soo Park for assistance with experiments, Roberto Tron for fruitful discussions and the anonymous reviewers for their constructive comments.

REFERENCES

[1] MoCap: Carnegie Mellon University Motion Capture Database. http://mocap.cs.cmu.edu/.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[3] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Nonrigid structure from motion in trajectory space. In *Advances in Neural Information Processing Systems*, 2008.

[4] B. D. Anderson and J. B. Moore. *Optimal filtering*. Courier Corporation, 2012.

[5] S. Bonnabel, P. Martin, and E. Salaün. Invariant Extended Kalman Filter: theory and application to a velocity-aided attitude estimation problem. In *The IEEE Conference on Decision and Control (CDC)*, 2009.

[6] G. Bourmaud, R. Mégret, M. Arnaudon, and A. Giremus. Continuous-discrete extended Kalman filter on matrix Lie groups using concentrated Gaussian distributions. *Journal of Mathematical Imaging and Vision*, 51(1):209–228, 2015.

[7] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[8] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1998.

[9] J. Brookshire and S. Teller. Articulated pose estimation via over-parametrization and noise projection. In *Robotics: Science and Systems*, 2014.

[10] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[11] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000.

[12] M. do Carmo. *Riemannian Geometry*. Mathematics (Boston, Mass.). Birkhäuser, 1992.

[13] T. Drummond and R. Cipolla. Real-time tracking of highly articulated structures in the presence of noisy measurements. In *The IEEE International Conference on Computer Vision (ICCV)*, 2001.

[14] K. Fragkiadaki, M. Salas, P. Arbelaez, and J. Malik. Grouping-based low-rank trajectory completion and 3D reconstruction. In *Advances in Neural Information Processing Systems*, 2014.

[15] J. Gallier and J. Quaintance. Notes on differential geometry and Lie groups. *University of Pennsylvannia*, 2016.

[16] S. Gudmundsson and E. Kappos. On the geometry of tangent bundles. *Expositiones Mathematicae*, 20(1):1–41, 2002.

[17] S. Hauberg, F. Lauze, and K. S. Pedersen. Unscented Kalman filtering on Riemannian manifolds. *Journal Mathematical Imaging and Vision*, 46(1):103–120, 2013.

[18] S. Hauberg, S. Sommer, and K. S. Pedersen. Gaussian-like spatial priors for articulated tracking. In *The European Conference on Computer Vision (ECCV)*. 2010.

[19] R. Hermann and A. J. Krener. Nonlinear controllability and observability. *IEEE Transactions on Automatic Control*, 22(5):728–740, 1977.

[20] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.

[21] J. M. Lee. Introduction to smooth manifolds. 2000.

[22] P. Muralidharan and P. T. Fletcher. Sasaki metrics for analysis of longitudinal data on manifolds. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[23] R. M. Murray, Z. Li, S. S. Sastry, and S. S. Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 1994.

[24] S. I. Olsen and A. Bartoli. Implicit non-rigid structure-from-motion with priors. *Journal of Mathematical Imaging and Vision*, 31(2-3):233–244, 2008.

[25] H. S. Park and Y. Sheikh. 3D reconstruction of a smooth articulated trajectory from a monocular image sequence. In *The IEEE International Conference on Computer Vision (ICCV)*, 2011.

[26] J. M. Rehg, D. D. Morris, and T. Kanade. Ambiguities in visual tracking of articulated objects using two-and three-dimensional models. *International Journal of Robotics Research*, 22(6):393–418, 2003.

[27] S. Sasaki. On the differential geometry of tangent bundles of Riemannian manifolds. *Tohoku Mathematical Journal, Second Series*, 10(3):338–354, 1958.

[28] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *The European Conference on Computer Vision (ECCV)*, 2000.

[29] C. Sminchisescu and B. Triggs. Building roadmaps of local minima of visual models. In *The European Conference on Computer Vision (ECCV)*. 2002.

[30] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3D human tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[31] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):878–892, 2008.

[32] R. Urtasun and T. Darrell. Sparse probabilistic regression for activity-independent human pose inference. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[33] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[34] J. Valmadre, Y. Zhu, S. Sridharan, and S. Lucey. Efficient articulated trajectory reconstruction using dynamic programming and filters. In *The European Conference on Computer Vision (ECCV)*, 2012.

[35] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3D shape estimation from 2D landmarks: A convex relaxation approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.