Nonparametric Object and Parts Modeling with Lie Group Dynamics

David S. Hayden, Jason Pacheco, John W. Fisher III
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge, MA 02139
{dshayden, pachecoj, fisher}@csail.mit.edu

Abstract

Articulated motion analysis often utilizes strong prior knowledge such as a known or trained parts model for humans. Yet, the world contains a variety of articulating objects–mammals, insects, mechanized structures–where the number and configuration of parts for a particular object is unknown in advance. Here, we relax such strong assumptions via an unsupervised, Bayesian nonparametric parts model that infers an unknown number of parts with motions coupled by a body dynamic and parameterized by SE(D), the Lie group of rigid transformations. We derive an inference procedure that utilizes short observation sequences (image, depth, point cloud or mesh) of an object in motion without need for markers or learned body models. Efficient Gibbs decompositions for inference over distributions on SE(D) demonstrate robust part decompositions of moving objects under both 3D and 2D observation models. The inferred representation permits novel analysis, such as object segmentation by relative part motion, and transfers to new observations of the same object type.

1. Introduction

The world is full of moving objects comprised of articulating parts. Despite the wide range and complexity of such objects, humans have a remarkable ability to accurately discern both the number of articulating parts and their relation to the whole with few observations. We are interested in developing reasoning methods and algorithms that mimic this ability. While one might consider supervised methods that rely on large amounts of labeled training data about every conceivable object, articulated motion and view, the task of data collection seems daunting and unnecessary.

Consequently, we develop a generative model that infers an object decomposition solely from brief observations of the object in motion. Specifically, we propose a parts-based representation that leverages Bayesian nonparametric dynamical models while eschewing strong assumptions about the number or configuration of parts. The model simultaneously infers a dynamic body frame, the number of parts, and their motion relative to the body frame. Diverse inputs are supported–2D image sequences, 2.5D depth sequences and 3D point cloud or mesh sequences–without need for restrictive assumptions about target appearance or the existence of specially-placed markers or sensors.

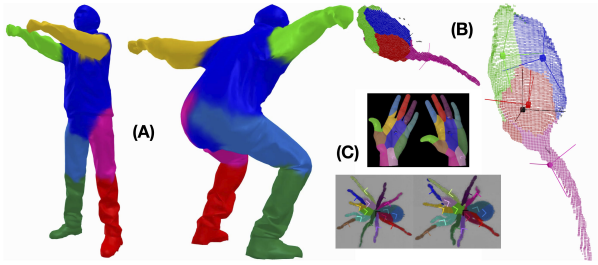


Figure 1. The number, rotation, translation, and shape of an object's parts are learned from a small number of observations of that object in motion. Motion of the body and parts is parameterized by the Lie group of rigid transformations in 3D or 2D. Supported data sources include sequences of meshes / point clouds (A, human), depth data (B, marmoset), and 2D images (C, hand, spider).

Objects and parts are assumed to rotate and translate smoothly in space, leading to a natural parameterization in SE(D), the Lie group of rigid transformations. By representing statistics of motion in the Lie algebra se(D), we derive closed-form Gibbs updates on translation dynamics, and an efficient sampler for rotation dynamics.

Contributions. We specify a novel Bayesian nonparametric model that is well-suited to the properties of articulated objects in motion (part persistence, rigid transformation dynamics, unknown number of parts). We demonstrate a novel decomposition of inferring translations and rotations in posterior distributions on SE(D) with concentrated Gaussian [35] priors. We validate our methods on 2D and 3D sequences containing different object types. We show that the parts in one data sequence transfer to other data sequences of the same object type (but different instance). Finally, we present novel analysis of the motion of different

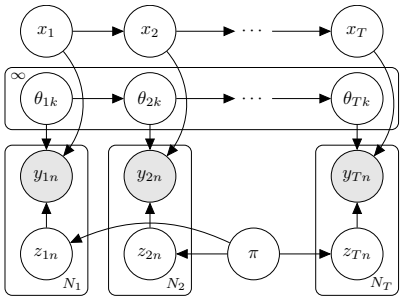


Figure 2. Simplified graphical model for an unknown number of time-varying parts $\{\theta_{tk}\}_{t=1,k=1}^{T,\infty}$ coupled by shared dynamics $\{x_t\}_{t=1}^{T}$. Observations y_{tn} are generated by part k if $z_{tn} = k$. Stick weights $\{\pi_k\}_{k=1}^{\infty}$ influence the observation counts for each part. Priors and $\{Q, \omega_k, S_k, W_k, E_k\}$ omitted for clarity.

regions of an object, such as segmenting an object based on the motion of its parts relative to the body frame.

2. Related Work

This work draws on body/parts models, Bayesian nonparametric dynamical models and Lie groups. Each contain a rich literature so we highlight only the most relevant details. Importantly, we are aware of no work that models body and part motion over time with Lie group dynamics, that is also unsupervised and nonparametric in the parts.

Body and Part Models The many treatments of part-based modeling begin with the pioneering work on human models of pictorial structures [10] and cardboard people [18]. Later work on deformable parts models [9] removes the need to define object-specific part configurations. Building on the success of offline analysis, real-time human pose tracking is now possible as well [27, 13]. All of these methods require specifying the number of parts. More detailed shape and pose models have been developed for a variety of objects, using a combination of known body models, mesh representations and sophisticated collection schemes including multiple cameras, IMUs, lasers and/or specially-painted targets [21, 41, 3, 24].

Unsupervised methods [36, 20, 26, 41, 40] have significant restrictions such as working for only 2D or only 3D data, or requiring annotated landmarks or point correspondences. In contrast, our unsupervised method works for 2D and 3D inputs, requires only a single sensor observing an object in motion, requires no distinctive or annotated object markings, and no observation correspondences.

Lie Groups Our work relies on the Lie group SE(D), the space of rigid transformations, for representing body and part motion. Lie groups have been used extensively in robotics and computer vision tasks such as SLAM [6], navigation [19], and parts-based models [5, 14, 17]. Defining observation models for Lie groups is challenging since the group is not a vector space. As such, notions of distance (and therefore distributions) require special care [23, 37], *e.g.*, simple additive noise models violate the group topology. Our approach defines a distribution (Gaussian) in the tangent plane about an element of the group [35]. Most works that model dynamics with SE(D) perform inference with approximate filters or smoothers, commonly the EKF [4] or UKF [6]. One exception that does full posterior inference is [28], though that work is not a dynamical model. See [8] for an accessible introduction to Lie groups, and [16] for a more thorough introduction.

Nonparametric Models Sequential models extending the well-known Dirichlet process (DP) [1] include the HDP-HMM [31], sticky HDP-HMM [12], infinite HMM [2], and infinite factorial HMM (ifHMM) [15]. Each of these permit an infinite number of states, but are restricted to discrete labels. Extensions to continuously-varying latent states include the HDP-SLDS [11], dynamic HDP [25], mixture of DPs [7], and the evolutionary HDP [38]. While each has the desirable property of shared global dynamics, none capture component persistence allowing new atoms at each time instance. This is undesirable for parts modeling as objects do not tend to acquire and lose parts over time and nonparametric priors already risk creating duplicate parts [12].

Closely related is the infinite factorial dynamical model [32], a continuous extension of the ifHMM which only permits shared global binary on/off states, and the Transformed Dirichlet Process [30], a DP allowing multiple groups of observations to share the same set of atoms (but with no dynamics). Most relevant, and what we use for comparison, is the Bayesian nonparametric model of Zhou et al. [39], a linear dynamical model where parts are independently sampled from a Dirichlet process at each time (but with no part persistence or Lie group representation).

3. Model

Let $t = 1, \ldots, T$ index time, $k = 1, \ldots, \infty$ index parts, and $n = 1, \ldots, N_t$ index observations at time t. Most generally, our nonparametric parts model (Figure 2) takes as its sole input observations $\{y_t\}_{t=1}^{T}$ where the t^{th} batch $y_t = \{y_{tn}\}_{n=1}^{N_t}$ contains N_t observations with unknown correspondence. There is a global (body) dynamic with time-varying parameters x_t and time-fixed parameter Q. There are an unknown number of components (parts) with time-varying parameters θ_{tk} and time-fixed parameters $\{\omega_k, S_k, E_k, W_k\}$. Stochastic dynamics models f, g and stochastic observation model h are, for each t, k, n,

$$x_t \sim f(x_{t-1}, Q) \qquad \theta_{tk} \sim g(\theta_{(t-1)k}, \omega_k, S_k)$$
$$y_{tn} \sim h(x_t, \theta_{tz_{tn}}, \omega_{z_{tn}}, E_{z_{tn}})$$

where $z_{tn} = k$ indicates that observation y_{tn} was generated by component k. A prior probability of association is given

by the discrete distribution of stick weights π (for $\alpha > 0$):

$$z_{tn} \sim \pi \qquad \pi \sim \text{GEM}(\alpha) \tag{1}$$

To specialize for object and parts modeling we must further specify the domain of random variables $\{y_{tn}, x_t, \theta_{tk}, \omega_k, S_k, E_k, W_k, Q\}$, the form of priors $\{H_x, H_\theta, H_\omega, H_S, H_E, H_W, H_Q\}$ and the forms of $\{f, g, h\}$. First, we introduce distributions on Lie groups.

3.1. Lie Groups

A Matrix Lie group G is a continuous group whose elements can be described by matrices with special structure. In this work, $G = \text{SE}(D)$, the space of rigid transformations on $\mathbb{R}^D$ or $G = \text{SO}(D)$ the space of D-dimensional proper rotations. Associated to G is Lie algebra $\mathfrak{g}$, which can be viewed as a local vector space approximation about the identity element of G. This approximation can be made with respect to any element in G because group elements compose via matrix multiplication and each element has an inverse. For $b, \mu \in G$ we call the local vector space approximation about μ the *tangent space* of μ, denoted $T_\mu G$. Mappings to and from the tangent space of μ are accomplished via the left-invariant Riemannian logarithm and left-invariant Riemannian exponential,

$$\text{Log} : G \times G \to \mathfrak{g} = \text{Log}_\mu b = \log_G(\mu^{-1} b) \tag{2}$$

$$\text{Exp} : G \times \mathfrak{g} \to G = \text{Exp}_\mu v = \mu \exp_G(v) \tag{3}$$

where $v \in \mathfrak{g}$ is a *tangent vector* in the tangent space of μ, and $\log_G, \exp_G$ are the Lie group logarithm and exponential maps, which can be computed using the matrix logarithm and matrix exponential. Note that v is called a tangent vector even though it is represented as a matrix: this is because a bijective mapping exists between a matrix and vector representation of v. We omit additional notation for brevity.

3.1.1 Distributions on Lie Groups

Constructing a distribution on G to reason over body and parts models is complicated by the fact that G is not a vector space. Exploiting the maps between group elements and their tangent spaces, we can define a distribution with location parameter $\mu \in G$ by mapping its support to the tangent space of μ. Define the left-invariant concentrated Gaussian $\text{N}_\text{L}(\cdot)$ in terms of the multivariate Gaussian $\text{N}(\cdot)$:

$$\text{N}_\text{L}(b|\mu, \Sigma) = \text{N}(\text{Log}_\mu b|0, \Sigma) \tag{4}$$

In similar fashion to [29] this can be thought of as a Gaussian in the tangent space about mean $\mu \in G$. The covariance Σ exists in the tangent space and can be understood to operate the same as in the typical Euclidean case except that vectors in $T_\mu G$ must be mapped back to the group by Eqn. 3. See Figure 3 for a visualization on $\text{SO}(2)$.

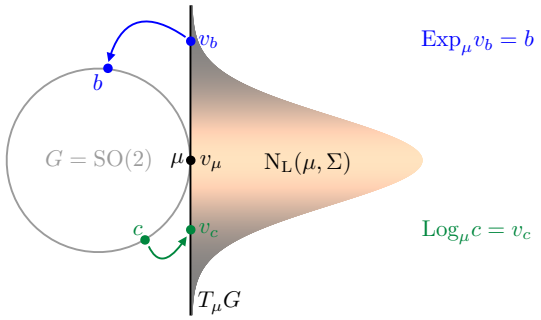


Figure 3. Location and scale distributions such as Gaussians can be locally defined about element μ in Lie group G by mapping their support into the local vector space approximation $T_\mu G$.

3.1.2 $\text{SE}(D)$ Actions

Represent $b, c \in G = \text{SE}(D)$ as real block matrices,

$$b = \begin{pmatrix} R_b & d_b \\ 0 & 1 \end{pmatrix} \quad c = \begin{pmatrix} R_c & d_c \\ 0 & 1 \end{pmatrix} \tag{5}$$

The R_* are $D \times D$ rotation matrices (with determinant $+1$) in $\text{SO}(D)$ and the d_* are translations in $\mathbb{R}^D$. Henceforth, we use this notation to represent the rotation and translation components of any element in $\text{SE}(D)$; for example, if $x_t \in \text{SE}(D)$ then it has rotation R_{x_t} and translation d_{x_t}.

Elements of $\text{SE}(D)$ compose via matrix multiplication (maintaining group closure), and act as a change of basis for homogeneous coordinates $\tilde{p}$ of point $p \in \mathbb{R}^D$:

$$bc\tilde{p} = \begin{pmatrix} R_b (R_c p + d_c) + d_b \\ 1 \end{pmatrix} \tag{6}$$

If $\tilde{p}$ are coordinates in frame c then $c\tilde{p}$ can be interpreted as its coordinates in frame b and $bc\tilde{p}$ can be interpreted as its coordinates in the standard (or world) basis. In general, changes of bases are best viewed as composing from left to right, but acting on points from right to left.

3.2. Body and Parts Model

Let $G = \text{SE}(D)$ for dimension $D \in \{2, 3\}$. We seek to infer a parts decomposition of an articulating object by directly observing it in motion. Specifically, we model the inputs $y_{tn} \in \mathbb{R}^D$ as being random collections of points sampled *within* the object as it moves across time. Variable (including no) observations are supported at each time, and no correspondence between observations is assumed. Diverse inputs are supported, including foreground pixels of 2D image sequences, unprojected points from depth sequences, and 3D point clouds sampled within mesh sequences.

We assume part persistence–an object does not gain or lose parts over time. We also assume that parts move smoothly through space but remain close (in an L2 sense) to a common body which also moves smoothly. The relation between body and part motion could be modeled in

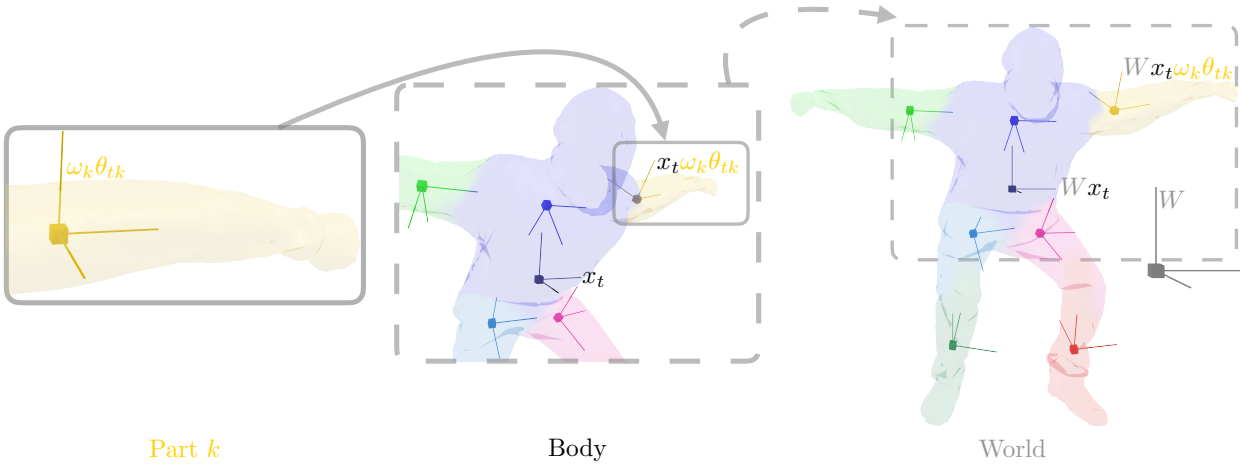


Figure 4. The frames that comprise an object at time t. Per-time body frames x_t are rigid transformations from world frame W. Each part k contains a time-fixed canonical part frame ω_k and a per-time part frame θ_{tk}. ω_k are a rigid transformation from body frame x_t while θ_{tk} are a rigid transformation from ω_k. Using stabilized random walk dynamics, each per-time part frame θ_{tk} is designed to transform smoothly over time but remain near the origin of their respective canonical part frame ω_k.

many ways: one naive extreme would be to model them as floating bodies with linear dynamics while the other extreme would be to model them as existing in a skeletal network of joints. Linear dynamics would fail to capture part articulation while skeletal networks are overly restrictive.

We take a middle ground: parts $\{\theta_{tk}, \omega_k, S_k, E_k, W_k\}$ are modeled as floating bodies that rotate and translate smoothly through space about a body frame $x_t \in G$, but whose origins tend to remain near the origin of a canonical part frame $\omega_k \in G$ through stabilized random walk dynamics. Canonical part frames are close to the body frame and remain fixed across time but parts also have per-time frames $\theta_{tk} \in G$. Parts are not fixed in their spatial extent; instead, they have a probabilistic, ellipsoidal shape model governed by Gaussian covariance E_k. Part dynamics are governed by covariance S_k and body dynamics are governed by covariance Q. The dispersion of canonical part frames about the body frame is governed by covariance W_k. Figure 4 graphically depicts how body and part frames compose.

3.2.1 Body and Part Dynamics

Body frames x_t and parts evolve independently, but are implicitly coupled through the observation model. In particular, the body frame stochastic dynamics model is:

$$x_t \sim \mathrm{N_L}(x_t \mid x_{t-1}, Q) \tag{7}$$

Object dynamics are a non-linear random walk on G whose noise covariance Q exists in the tangent space about the body frame at the previous time. Canonical part frames ω_k are dispersed about the body frame with covariance W_k,

$$\omega_k \sim H_\omega = \mathrm{N_L}(\cdot \mid I, W_k) \tag{8}$$

where $I \in G$ is the identity element (no translation or rotation) and covariance W_k can be thought to (implicitly) exist in the tangent space of x_t. Each part has per-time dynamics θ_{tk} with driving noise covariance S_k governed by:

$$\theta_{tk} = \begin{pmatrix} \mathrm{Exp}_{R_{\theta_{(t-1)k}}} \phi_{tk} & A\, d_{\theta_{(t-1)k}} + B\, m_{tk} \\ 0 & 1 \end{pmatrix} \tag{9}$$

with constants $A = \mathrm{diag}\left(\sqrt{a}, \ldots, \sqrt{a}\right)$, $B = \mathrm{diag}\left(\sqrt{1-a}, \ldots, \sqrt{1-a}\right)$ and Exp in Eqn. 9 the Riemannian exponential for $\mathrm{SO}(D)$. $\phi_{tk} \in \mathrm{so}(D)$ is a vector in the tangent space of $R_{\theta_{(t-1)k}}$. Part translation driving noise m_{tk} and rotation driving noise ϕ_{tk} are jointly distributed:

$$(m_{tk}, \phi_{tk}) \sim \mathrm{N}\left(0, S_k\right) \tag{10}$$

As proven in the supplemental, carefully chosen coefficients of matrices A, B ($a = 0.95$) cause the asymptotic covariance of the part translation $d_{\theta_{tk}}$ to equal the covariance of translation driving noise m_{tk}. This form enables parts to transform smoothly, but never too far from their canonical location, and mitigated part confusion during inference.

All driving noise covariances are drawn from Inverse-Wishart distributions, where we note that our model supports arbitrary correlations between translation and rotation for object, canonical part, and part transformations:

$$Q \sim H_Q = \mathrm{IW}(\cdot|v_{Q_0}, \Lambda_{Q_0}) \tag{11}$$
$$S_k \sim H_S = \mathrm{IW}(\cdot|v_{S_0}, \Lambda_{S_0}) \tag{12}$$
$$W_k \sim H_W = \mathrm{IW}(\cdot|v_{W_0}, \Lambda_{W_0}) \tag{13}$$

And initial body and part frames are drawn according to:

$$x_1 \sim H_x = \mathrm{N_L}(\cdot|x_0, \Sigma_x) \tag{14}$$
$$\theta_{1k} \sim H_\theta = \mathrm{N_L}(\cdot|\theta_0, \Sigma_\theta) \tag{15}$$

3.2.2 Observation Models for 3D and 2D data

Input y_{tn} is assumed to be in world coordinate system W, which is assumed to be aligned with the sensor's coordinate system (hence, W has no rotation or translation and is henceforth omitted). Parts generate observations in their respective part coordinate systems and are mapped to world coordinates via θ_{tk}, ω_k and the body frame x_t. That is, part k generates point $e_{tn} \sim \mathrm{N}(0, E_k)$ which is then mapped to world coordinates $\tilde{y}_{tn} = x_t \omega_k \theta_{tk} \tilde{e}_{tn}$ if $z_{tn} = k$ (where $\tilde{(\cdot)}$ is a homogeneous projection of $(\cdot)$). The transformation is linear in $\tilde{e}_{tn}$ allowing straightforward mean and variance computations of the homogeneous point in world coordinate $\tilde{y}_{tn}$, yielding the following observation model (for $z_{tn} = k$)

$$\tilde{y}_{tn} \sim \mathrm{N}(\tilde{y}_{tn} | x_t \omega_k \theta_{tk} \tilde{0}_{\mathbb{R}}, x_t \omega_k \theta_{tk} \tilde{E}_k \theta_{tk}^\top \omega_k^\top x_t^\top) \quad (16)$$

where $\tilde{0}_{\mathbb{R}}$ is the homogeneous zero vector in $\mathbb{R}^D$ and $\tilde{E}_k$ is a degenerate block covariance matrix E_k with a zero row and column (a covariance in homogeneous coordinates). Without homogeneous coordinates, this is (via Eqn. 5):

$$y_{tn} \sim \mathrm{N}\left(y_{tn} \mid \mu_{tk}, \Sigma_{tk}\right) \quad (17)$$

$$\mu_{tk} = R_{x_t} R_{\omega_k} \left(d_{\theta_{tk}} + d_{\omega_k}\right) \quad (18)$$

$$\Sigma_{tk} = R_{x_t} R_{\omega_k} R_{\theta_{tk}} E_k R_{\theta_{tk}}^\top R_{\omega_k}^\top R_{x_t}^\top \quad (19)$$

While simple, it accommodates image plane observations in 2D, depth observations in 2.5D and XYZ observations in 3D. Incorporating additional terms (*e.g.*, appearance) is straightforward, but were not needed for our purposes. As with most generative models, robustness to missing data (common for depth sensors) is handled seamlessly.

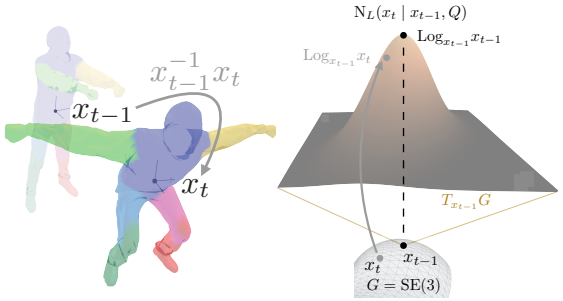
The observation covariance Σ_{tk} for y_{tn} is some rotation of E_k for $z_{tn} = k$ due to the composition of body and part frames. Consequently, E_k is constrained to be diagonal (*i.e.* axis-aligned) so as to avoid ambiguity. While the use of E_k implies a probabilistic, ellipsoid part shape model, its primary function is to yield robust associations z_{tn} of observations to parts. Here, we use the following prior:

$$E_k \sim H_E = \mathrm{IW}(\cdot | v_{E_0}, \Lambda_{E_0}) \quad (20)$$

4. Inference

We wish to sample from the posterior, which has likelihood proportional to the product of Equations (1, 7, 8, 10, 11, 12, 13, 14, 15, 17, 20) over all t, k, n. This is accomplished with Markov Chain Monte Carlo (MCMC) inference that exploits Gaussian statistics in the tangent space for efficient updates while simultaneously respecting the geometry of the Lie group via Eqns. 2 and 4. This is accomplished by sampling from the full conditional distributions of each latent variable, grouped in order of discussion,

$$(x_t, \theta_{tk}, \omega_k) \qquad z_{tn} \qquad (\pi, E_k, S_k, Q) \quad (21)$$



Body Frame Projection **Tangent Distribution**

Figure 5. *Left:* Object dynamics of body frame at time t are projected into coordinates at time $t-1$ by the Lie group operation $x_{t-1}^{-1} x_t$. *Right:* The projection is in $SE(3)$ with Gaussian statistics in the tangent plane of x_{t-1}. The figure notionally depicts two degrees of freedom, whereas $SE(3)$ would have 6 degrees.

where $t = 1, \ldots, T, k = 1, \ldots, \infty, n = 1, \ldots, N_t$ and omitted leading subscripts are taken to mean joint dependence (*i.e.* $y = \{y_t\}_{t=1}^T$ and $y_t = \{y_{tn}\}_{n=1}^{N_t}$). Inference complexity is linear in the number of observations and parts. In our experiments, chains were generally mixed after about 300 samples, with approximately 1 minute per sample being the worst-case timing for any data we tested on.

In the sequel we sketch the sampling of body transformations x_t. Full details are in the supplement, along with sampling of the canonical parts ω_k and part transformations θ_{tk} which take a similar form. We also discuss sampling part associations z_{tn}, which are conjugate except when sampling assignments to the base measure. The conditionals in the third grouping (π, E_k, S_k, Q) can be sampled analytically due to conjugate priors, which we defer to the supplement. Finally, parts $\{\theta_{tk}, \omega_k, S_k, W_k, E_k\}$ can be sampled in parallel across k and z_{tn} can be sampled in parallel across t, n.

4.1. Decomposition of Lie Group dynamics

We exploit the Lie algebra to develop an efficient Gibbs sampler for dynamical terms $\{x_t, \omega_k, \theta_{tk}\}$. For example, the operation $x_{t-1}^{-1} x_t$ transforms the body frame at time t into that of the body frame at time $t-1$ (Fig. 5, left). This operation is an element of $SE(D)$:

$$x_{t-1}^{-1} x_t \triangleq \begin{pmatrix} R_{x_{t-1}^{-1}, x_t} & d_{x_{t-1}^{-1}, x_t} \\ 0 & 1 \end{pmatrix}, \quad (22)$$

where $R_{x_{t-1}^{-1}, x_t} = R_{x_{t-1}}^T R_{x_t}$ and $d_{x_{t-1}^{-1}, x_t} = R_{x_{t-1}}^T (d_{x_t} - d_{x_{t-1}})$. Elements in the frame x_t are mapped to the tangent space of x_{t-1} via the Riemannian Log map (Fig. 5, right):

$$\mathrm{Log}_{x_{t-1}} x_t \triangleq \log_G(x_{t-1}^{-1} x_t) = \begin{pmatrix} V_{x_{t-1}^{-1}, x_t}^{-1} d_{x_{t-1}^{-1}, x_t} \\ \phi_{x_{t-1}^{-1}, x_t} \end{pmatrix} \quad (23)$$

The first entry $V_{x_{t-1}^{-1}, x_t}^{-1} d_{x_{t-1}^{-1}, x_t}$ are tangent space coordinates of translation and the second entry $\phi_{x_{t-1}^{-1}, x_t}$ is a rota-

tion vector. The invertible linear operator $V^{-1}_{x^{-1}_{t-1},x_t}$ is computable from rotation $R_{x^{-1}_{t-1},x_t}$ (or from $\phi_{x^{-1}_{t-1},x_t}$). This is well-defined for $x^{-1}_{t-1}x_t$ sufficiently close to identity and consistent with small incremental motions.

4.2. Gibbs Sampling Updates

Recall that Eqns. (22) and (23) map x_t to the tangent space of x_{t-1}. When conditioned on rotation, this mapping is linear in the translation component d_{x_t}. This observation, combined with Gaussian statistics in the tangent space, yields closed-form Gibbs updates for translation. To see this, observe that the distribution over dynamics in the tangent space is (Fig. 5, right),

$$\mathrm{N}_L(x_t \mid x_{t-1}, Q) = \mathrm{N}\left(\begin{pmatrix} Cd_{x_t} + u \\ \phi_{x^{-1}_{t-1},x_t} \end{pmatrix} \middle| 0, Q\right) \quad (24)$$

where $C = V^{-1}_{x^{-1}_{t-1},x_t} R^\top_{x_{t-1}}$ and $u = -V_{x^{-1}_{t-1},x_t} R^\top_{x_{t-1}} d_{x_{t-1}}$. Conditioned on rotation R_{x_t} and previous body frame x_{t-1}, the corresponding rotation vector $\phi_{x^{-1}_{t-1},x_t}$ and matrix $V_{x^{-1}_{t-1},x_t}$ are fixed quantities. This renders C and u computable and yields a Gaussian conditional distribution for d_{x_t}. This conditional constitutes our prior belief about d_{x_t} given R_{x_t}, x_{t-1} and covariance Q. Similar logic allows us to derive a Gaussian conditional on d_{x_t} given future transformation x_{t+1}. These can be analytically combined to provide a Gaussian distribution for $d_{x_t} \mid R_{x_t}, x_{t-1}, x_{t+1}$. Because this is Gaussian, and the observation model is also a product of Gaussians whose parameters are known given $\{\omega_k, E_k, \theta_{tk}\}_{k=1}^{\infty}$ and $\{z_{tn}\}_{n=1}^{N_t}$, it follows that the posterior on d_{x_t} is also Gaussian, and analytically computable.

In contrast, sampling of rotation parameters lacks a closed form. We utilize univariate slice sampling [22] for the full conditional of each rotation parameter, along with a fixed number of MCMC proposals to correct for known rotational symmetries. Details are in the supplement.

Part Association The conditional distribution for a single assignment to an existing part $k \geq 1$ is given by $p(z_{tn} = k \mid y_{tn}, x_t, \omega, \theta_t, \pi, E) \propto \pi_k \, p(y_{tn} | x_t, \omega_k, \theta_{tk}, E_k)$. Conversely, association to a new part is given by,

$$p(z_{tn} = -1 \,|\, y_{tn}, x_t, \omega, \theta_t, \pi, E) \propto \quad (25)$$
$$\pi_* \int p(y_{tn} | x_t, \omega_*, \theta_{t*}, E_*) p(\omega_*, \theta_{t*}, E_*) \,\mathrm{d}(\omega_*, \theta_{t*}, E_*)$$

where π_* is the stick weight corresponding to the base measure (*i.e.* all uninstantiated parts). This is not analytic in our model, but can be effectively approximated by Monte Carlo sampling of parts (need only be done once) or approximation by a constant (since the predictive distribution of parts will be broad, but centered at the object frame of reference). We obtain satisfactory results with both approaches.

5. Results

We present several experimental results. We compare quantitatively and qualitatively to nonparametric and parametric baselines in 5.1. We present results on dynamic mesh data in 5.2. We demonstrate object segmentation based on relative part motion in 5.3. We show transfer of learned representations to a novel dataset and synthesize motion from the learned representation in the supplement. The video supplemental animates these results.

5.1. Quantitative Comparison

We examine *part discovery* performance on three object motion datasets and compare to manually-annotated ground-truth. We emphasize that annotations are not incorporated into the inference procedure. We refer to the datasets as `hand`, `spider`, and `marmoset`. `hand` and `spider` are 2-D image data, while `marmoset` is 3D data unprojected from a depth camera. Inference utilizes $12 - 44$ frames (depending on the dataset) and results are compared to five manually-annotated ground-truth frames (where ground truth is the number of parts and their segmentations – examples are in the supplement). In each dataset, parts have nearly indistinguishable appearances and none of the compared methods use an appearance model. Consequently, part discovery is achieved via analysis of motion dynamics. Inputs only contain foreground (*i.e.* background is removed), as is done in related works [20].

We report multi-object tracking and segmentation (MOTS) metrics [34], which measure how well the part associations overlap with groundtruth part segmentations (MOTSA, sMOTSA, MOTSP) and how stable the part associations are over time (IDS). These metrics are intended for segmenting multiple targets, but we repurpose them to segment multiple parts. Comparisons are with IoU 0.3.

We compare against two baselines: the Bayesian nonparametric model of [39] (discussed in Section 2), which we call the nonparametric extents model `npe`, and a parametric modification of [39], so that it is given the advantage of knowing the true number of parts. We call this the parametric extents model `pe`. Neither `npe` nor `pe` consider part *persistence* over time (as we do), so for these methods we use the Hungarian algorithm to compute part correspondences between pairs of timesteps on the distance (in the body frame) of component means.

Taken together, our model, and the two baselines, constitute an ablation study in which we consider unknown number of parts with Lie group dynamics, and unknown / known number of parts, without Lie group dynamics. In all cases, we compute mean and standard deviation of MOTS statistics on 100 samples taken from a Markov chain of 1000 samples, use data-dependent priors (specified in the supplemental), and set concentration parameter $\alpha = 0.1$. Figure 6

Dataset	Method	IDS	MOTSA	MOTSP	sMOTSA
hand	ours	$\mathbf{0.00 \pm 0.00}$	$\mathbf{2.79 \pm 0.30}$	$\mathbf{0.71 \pm 0.01}$	$\mathbf{1.34 \pm 0.24}$
	npe	4.45 ± 1.84	1.93 ± 0.8	0.51 ± 0.01	-4.2 ± 0.78
	pe	4.03 ± 2.11	1.57 ± 0.44	0.47 ± 0.01	-0.33 ± 0.37
spider	ours	$\mathbf{5.14 \pm 1.49}$	$\mathbf{3.44 \pm 0.25}$	$\mathbf{0.55 \pm 0.02}$	$\mathbf{1.26 \pm 0.18}$
	npe	19.6 ± 2.88	-4.4 ± 0.92	0.51 ± 0.01	-6.72 ± 0.9
	pe	17.28 ± 3.06	1.73 ± 0.31	0.52 ± 0.01	-0.24 ± 0.27
marmoset	ours	1.24 ± 0.65	1.39 ± 0.89	$\mathbf{0.49 \pm 0.02}$	-0.47 ± 0.71
	npe	3.18 ± 1.28	-32.44 ± 2.78	0.35 ± 0.01	-34.06 ± 2.72
	pe	$\mathbf{0.43 \pm 0.51}$	$\mathbf{3.86 \pm 0.17}$	0.48 ± 0.00	$\mathbf{1.77 \pm 0.17}$
average	ours	$\mathbf{2.12 \pm 0.71}$	$\mathbf{2.54 \pm 0.48}$	$\mathbf{0.58 \pm 0.02}$	$\mathbf{0.71 \pm 0.38}$
	npe	9.07 ± 2.0	-11.63 ± 1.5	0.46 ± 0.01	-15.0 ± 1.47
	pe	7.25 ± 1.89	2.39 ± 0.31	0.49 ± 0.01	0.39 ± 0.27

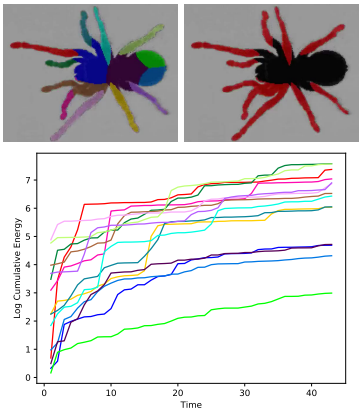


Figure 6. (**Left**): Quantitative comparison of our nonparametric parts model (ours) with nonparametric baseline npe and parametric baseline pe using MOTS metrics. Lower IDS is better, higher MOTSA, MOTSP, and sMOTSA is better. Best-performing method is emboldened. (**Right**): Object segmentation based on part motion across time. Whereas the parts nearest the body center exhibit little motion (in the body frame), the extremities of spider exhibit large amounts of motion. (Top-Left): Part associations. (Top-Right): Part segmentation based on motion energy. (Bottom): Log cumulative part motion energy across time (color-coordinated to associations).

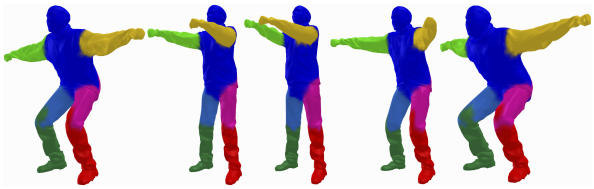


Figure 7. Dynamic mesh segmentation. By using points sampled inside a mesh as the input to our nonparametric parts model, then computing associations to mesh vertices, our model can learn parts and dynamics from mesh data. Additional views in Figure 1

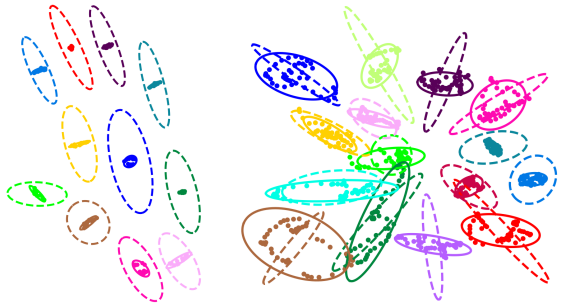


Figure 8. Part posteriors for *hand* and *spider*. Dotted ellipses are the mean part covariance, solid ellipses visualize the part posterior location covariance. Points are observed part locations used for the posterior updates. The leg locations of spider are smeared due to their articulation whereas the fingers of the hand are concentrated.

(left) shows quantitative results while Figure 9 show qualitative comparisons between our method and the baseline.

Our model outperforms the nonparametric baseline in all datasets and metrics. The pe baseline (which benefits from knowing the number of parts) outperforms our method on label switches (IDS) and overall quality (sMOTSA) on the 3D marmoset data. This is largely due to noisy data from the depth sensor generating observations from the background that are distant from the object, but not so distant as to be relegated to the base measure. We see very little ID switching (IDS) and relatively high precision (MOTSP) in our model, which we attribute to the canonical parts ω_k enforcing that each part transformation θ_{tk} move stably. Visually, part assignments correspond best to ground-truth parts that are extremities (fingers, legs, tails), but tend to oversegment large object interiors (palms, bodies). We attribute this to the ellipsoidal observation model but find that, for the purposes of part analysis, it has no obvious negative impact.

5.2. Dynamic Mesh Segmentation

We apply our method to the squat1 sequence of the articulated mesh dataset from [33] decomposing the mesh sequence data into individual parts as shown in Figure 7. Note that legs are segmented into two parts each, while arms are segmented into one part. This result is consistent with the movement in this sequence where the legs bend, but the arms are held straight. Some artifacts appear when, for instance, the lower-left leg (red) has small numbers of associations above the knee when the person is squatting, but not when standing straight up. Qualitatively, the results conform to human part interpretation.

5.3. Motion Analysis

We show how our model facilitates novel object / part analysis. Beginning with Figure 8, we visualize part diagrams for hand and spider. Dotted ellipses show the observation noise model E_k for each part (in the object frame), while solid ellipses show the covariance for that part's translation across time. Because the part translation covariances are spatially separated, the model resists label switching between parts because they tend to stay proximate to their

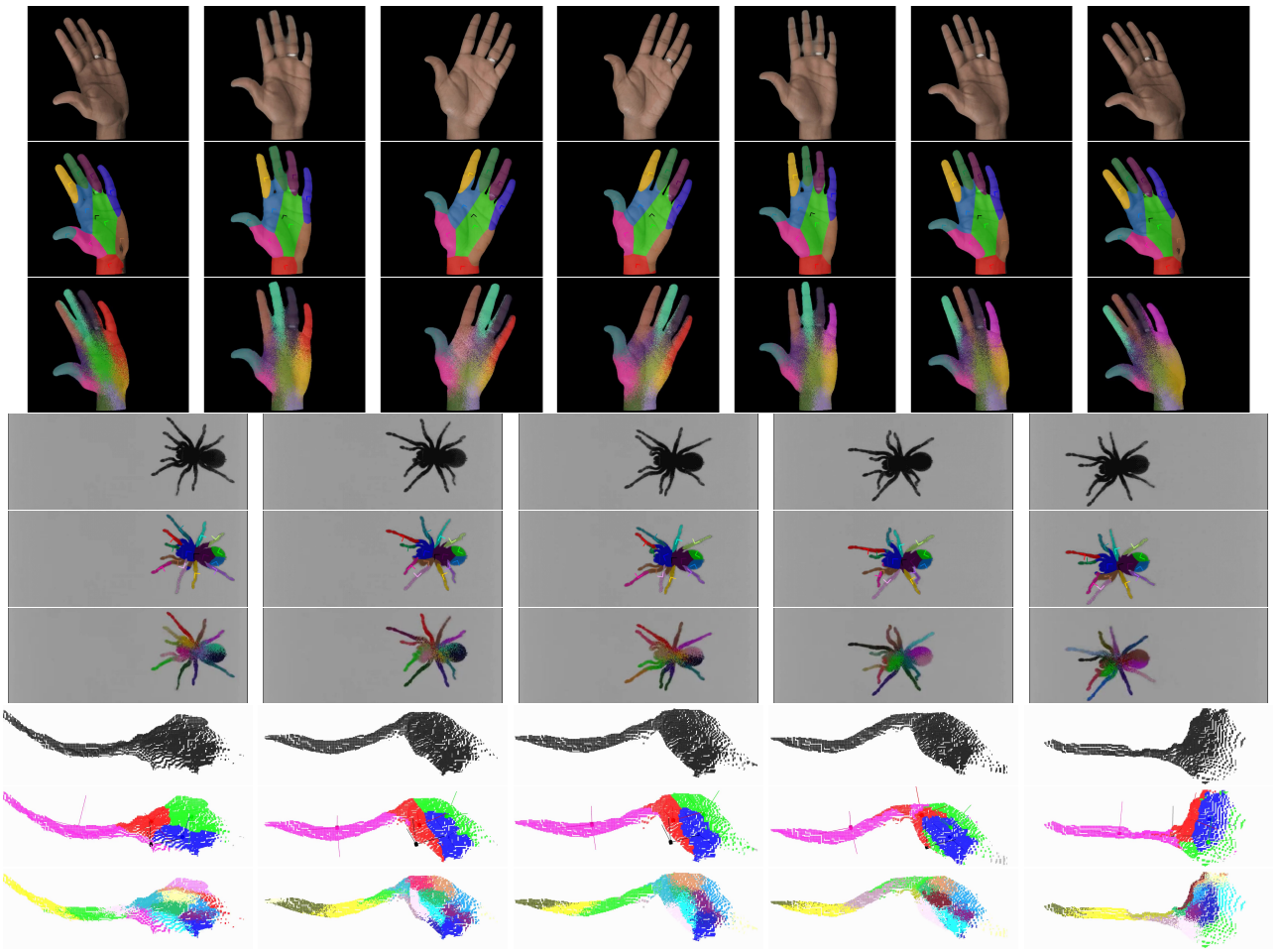


Figure 9. **Part associations** for a subset of frames in each sequence. For each sequence we show the original video (*top-row*) with part associations and object/part coordinate frames overlaid (*middle-row*) and baseline associations (bottom-row). We find estimated parts to be largely consistent over time, even for the highly articulated spider legs.

canonical frame. We observe that the part translation covariances are tight for the hand, but horizontally smeared for the spider–this is expected, because the fingers moved very little in `hand` compared to the legs in `spider`.

One analysis that our model enables is the comparison of part motions in the body frame (i.e. motion not from the object moving, but from its parts). By integrating each part's motion over time within the body frame we can determine which areas of an object experience high or low *relative* motion. Figure 6 (right) shows that, for `spider`, the legs are able to be segmented from other parts.

6. Discussion

In this work we demonstrated that our nonparametric representation of kinematic bodies infers meaningful part decompositions of objects in an unsupervised way, by simply observing them in motion. Furthermore, our Lie group representation constrains articulations of moving parts to physically plausible kinematic states, without the requirement of object-specific knowledge such as skeletal structures. Part decompositions are learnable on very short sequences, and generalize to other datasets and instances of the same object type. In contrast to methods which rely on extensive training data and/or object-specific 2D/3D models, we were able to demonstrate robust analysis by direct observation of single instances of an object.

Our model simplifies inference and motion analysis while suggesting straightforward extensions. For example, part persistence ensures that the representation of parts persists over a video sequence, even if parts become occluded. A hierarchical model over multiple videos of similar objects would thus be robust to occlusions in any single video. Furthermore, Gaussian tangent-space conditionals allows closed-form Gibbs updates for translation, efficient slice sampling of rotation, and proves sufficient for motion analysis. Explicit models of part shape may avoid oversegmenting large regions and is the focus of current work.

Acknowledgements This work was partially supported by the ONR (N00014-17-1-2072) and the NIH (5R01MH111916).

References

[1] Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974. 2

[2] Matthew J Beal, Zoubin Ghahramani, and Carl E Rasmussen. The infinite hidden markov model. In *Advances in neural information processing systems*, pages 577–584, 2002. 2

[3] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic faust: Registering human bodies in motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6233–6242, 2017. 2

[4] Guillaume Bourmaud, Rémi Mégret, Marc Arnaudon, and Audrey 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. 2

[5] C. Bregler and J. Malik. Tracking people with twists and exponential maps. *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, pages 8–15, 1998. 2

[6] Martin Brossard, Silvere Bonnabel, and Jean-Philippe Condomines. Unscented kalman filtering on lie groups. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2485–2491. IEEE, 2017. 2

[7] David B. Dunson. Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, 7(4):551–568, 2006. 2

[8] Ethan Eade. Lie groups for computer vision. *Cambridge Univ., Cam-bridge, UK, Tech. Rep*, 2014. 2

[9] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010. 2

[10] Martin A Fischler and Robert A Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on computers*, (1):67–92, 1973. 2

[11] Emily Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. Bayesian nonparametric inference of switching dynamic linear models. *IEEE Transactions on Signal Processing*, 59(4):1569–1585, 2011. 2

[12] Emily B Fox, Erik B Sudderth, Michael I Jordan, Alan S Willsky, et al. A sticky hdp-hmm with application to speaker diarization. *The Annals of Applied Statistics*, 5(2A):1020–1056, 2011. 2

[13] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015. 2

[14] Oren Freifeld, Alexander Weiss, Silvia Zuffi, and Michael J Black. Contour people: A parameterized model of 2d articulated human shape. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 639–646. IEEE, 2010. 2

[15] Jurgen V Gael, Yee W Teh, and Zoubin Ghahramani. The infinite factorial hidden markov model. In *Advances in Neural Information Processing Systems*, pages 1697–1704, 2009. 2

[16] Brian C Hall. Lie groups, lie algebras, and representations. In *Quantum Theory for Mathematicians*, pages 333–366. Springer, 2013. 2

[17] Søren Hauberg, François Lauze, and Kim Steenstrup Pedersen. Unscented kalman filtering on riemannian manifolds. *Journal of Mathematical Imaging and Vision*, 46(1):103–120, 2013. 2

[18] Shanon X Ju, Michael J Black, and Yaser Yacoob. Cardboard people: A parameterized model of articulated image motion. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 38–44. IEEE, 1996. 2

[19] Giuseppe Loianno, Michael Watterson, and Vijay Kumar. Visual inertial odometry for quadrotors on SE(3). *Proceedings - IEEE International Conference on Robotics and Automation*, 2016-June(3):1544–1551, 2016. 2

[20] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10955–10964, 2019. 2, 6

[21] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5442–5451, 2019. 2

[22] Radford M Neal et al. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003. 6

[23] Frank C. Park. Distance Metrics on the Rigid-Body Motions with Applications to Mechanism Design. *Journal of Mechanical Design*, 117(1):48–54, 1995. 2

[24] 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 *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 663–670. IEEE, 2010. 2

[25] Lu Ren, David B. Dunson, and Lawrence Carin. The dynamic hierarchical Dirichlet process. *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 824–831, 2008. 2

[26] David A Ross, Daniel Tarlow, and Richard S Zemel. Unsupervised learning of skeletons from motion. In *European Conference on Computer Vision*, pages 560–573. Springer, 2008. 2

[27] J Shotton, A Fitzgibbon, M Cook, T Sharp, M Finocchio, R Moore, A Kipman, and A Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304. IEEE Computer Society, 2011. 2

[28] Julian Straub, Jason Chang, Oren Freifeld, and John Fisher III. A dirichlet process mixture model for spherical data. In *Artificial Intelligence and Statistics*, pages 930–938, 2015. 2

[29] Julian Straub, Oren Freifeld, Guy Rosman, John J Leonard, and John W Fisher III. The manhattan frame model–manhattan world inference in the space of surface normals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3

[30] Erik B Sudderth, Antonio Torralba, William T Freeman, and Alan S Willsky. Describing visual scenes using transformed dirichlet processes. In *Advances in neural information processing systems*, pages 1297–1304, 2006. 2

[31] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 2

[32] Isabel Valera, Francisco Ruiz, Lennart Svensson, and Fernando Perez-Cruz. Infinite factorial dynamical model. In *Advances in Neural Information Processing Systems*, pages 1666–1674, 2015. 2

[33] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics (TOG)*, volume 27, page 97. ACM, 2008. 7

[34] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7942–7951, 2019. 6

[35] Yunfeng Wang and Gregory S Chirikjian. Error propagation on the euclidean group with applications to manipulator kinematics. *IEEE Transactions on Robotics*, 22(4):591–602, 2006. 1, 2

[36] Zhenjia Xu, Zhijian Liu, Chen Sun, Kevin Murphy, William T Freeman, Joshua B Tenenbaum, and Jiajun Wu. Unsupervised discovery of parts, structure, and dynamics. *International Conference on Learning and Representation*, 2019. 2

[37] Milos Zefran, Vijay Kumar, and Christopher Croke. Choice of Riemannian Metrics for Rigid Body Kinematics. *ASME Design Engineering Technical Conference and Computers in Engineering Conference*, (3):1–11, 1996. 2

[38] Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, page 1079, 2010. 2

[39] Yang Zhou, Jitendra Sharma, Qiong Ke, Rogier Landman, Jingli Yuan, Hong Chen, David S Hayden, John W Fisher, Minqing Jiang, William Menegas, et al. Atypical behaviour and connectivity in shank3-mutant macaques. *Nature*, page 1, 2019. 2, 6

[40] Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3955–3963, 2018. 2

[41] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6365–6373, 2017. 2