Lie Algebra-Based Kinematic Prior for 3D Human Pose Tracking

Edgar Simo-Serra, Carme Torras, and Francesc Moreno-Noguer
Institut de Robòtica i Informàtica Industrial (CSIC-UPC). Barcelona, Spain
{esimo,torras,fmoreno}@iri.upc.edu

Abstract

We propose a novel kinematic prior for 3D human pose tracking that allows predicting the position in subsequent frames given the current position. We first define a Riemannian manifold that models the pose and extend it with its Lie algebra to also be able to represent the kinematics. We then learn a joint Gaussian mixture model of both the human pose and the kinematics on this manifold. Finally by conditioning the kinematics on the pose we are able to obtain a distribution of poses for subsequent frames that which can be used as a reliable prior in 3D human pose tracking. Our model scales well to large amounts of data and can be sampled at over 100,000 samples/second. We show it outperforms the widely used Gaussian diffusion model on the challenging Human3.6M dataset.

1 Introduction

Tracking humans in videos is an area of computer vision that has been seeing a lot of research recently. This is due to the appearance of cheap depth cameras that have allowed the creation of many new databases with 3D human pose for tasks such as 3D human pose estimation itself or action recognition in which 3D human pose estimation becomes a simple feature. The scope and size of these new datasets require development of new tools that can scale well for these tasks.

In this work we propose modelling 3D human pose and kinematics in a single Riemannian manifold which is able to fully capture individual-independent pose and motion efficiently. We do this by first defining a manifold on the joint angles for the pose and extending the manifold with its own Lie algebra. We then learn a joint model using a recently proposed unsupervised clustering method for data on known Riemannian manifolds [13]. The mixture can then be conditioned on a given pose to obtain a distribution of velocities for that pose which, as we show, can be used as a reliable prior for 3D human pose tracking. An example is shown in Fig. 1.

The most simple traditionally used kinematic prior has been Gaussian diffusion [2, 3, 4, 10, 16]. This consists in simply searching in a small area defined by a Gaussian from the previous pose, i.e., $x_t = x_{t-1} + \epsilon$, where x_t would be the pose at time t and ϵ would be a Gaussian with 0 mean and diagonal covariance. This prior is considered to be action independent as it is a hyperparameter not tuned for a specific action. While this approach has proven to be fairly effective, by learning stronger motion models much better and more efficient algorithms can been obtained. More efficient algorithms allow achieving both higher performance as well as being much faster due to avoiding the need of thoroughly sampling the solution space.

A large number of approaches have been using the family of Gaussian Process models for learning mo-

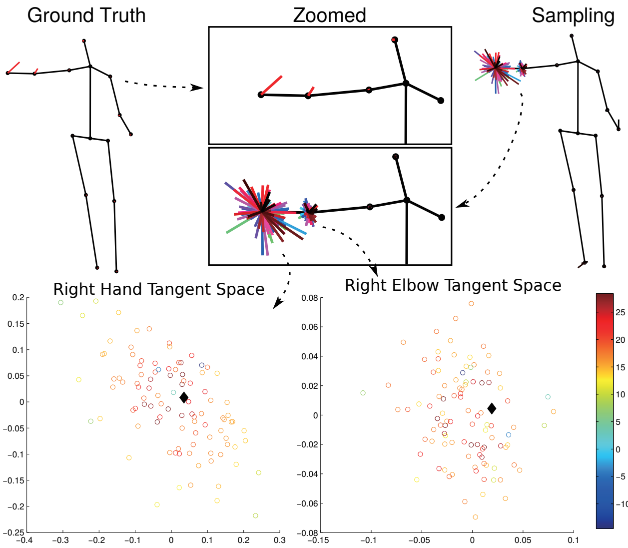

Figure 1. Example of our motion prior. We show 100 samples of predictions of our model from a particular pose. For visualization purposes the samples are multiplied by 3. We additionally show the samples for some of the joints with their associated log-likelihood, where the ground truth is shown with a black diamond.

tion based on latent spaces (GPLVM) [7]. One of the most well known approaches is the Gaussian Process Dynamic Model (GPDM) proposed by Wang et al. [18, 19, 20]. Hierarchical variants (hGPLVM) have also been used in a tracking by detection approach [1]. However, Gaussian Processes do not scale well to large datasets due to their $O(n^3)$ complexity for prediction. Sparse approximations do exist [9], but in general do not perform as well. In contrast our algorithm has a $O(1)$ complexity for sampling.

There have been other approaches such as learning Conditional Restricted Boltzmann Machines (CRBM) [17]. However, these methods have a very complex learning procedure that makes use of several approximations and thus it is not easy to train good models. Li et al. [8] proposed the Globally Coordinated Mixture of Factor Analyzers (GCMFA) model which is similar to the GPLVM ones in the sense it is performing a strong non-linear dimensionality reduction. Yet, as GPLVM it does not scale well to large datasets such as the ones we consider in this work.

We would like to point out that none of the aforementioned approaches are consistent with the manifold of human motion. Some of them use directly the 3D points of the joints while others use angles. In the case of considering 3D points the limb length may vary during the tracking, which is neither realistic nor desirable. In the case of angle representations, they have an inherent periodicity and thus are not a vector space even though they are usually treated as such. Two nearby angles may have very different values, e.g., 0

Table 1. Comparison of different pose priors in the literature for tracking. For complexity we take into account the number of hyperparameters and the difficulty of learning the model. Models are considered to scale if they can handle well large amounts of data (~100k samples) and to be consistent if they use geodesic distances instead of other metrics.

Prior	Complexity	Scales	Consistent
Gaussian diff.	Low	**Yes**	No
GPLVM [7]	Low	No	No
GPDM [19]	Medium	No	No
hGPLVM [1]	Medium	No	No
CRBM [17]	High	**Yes**	No
GCMFA [8]	High	No	No
GFMM (Ours)	Low	**Yes**	**Yes**

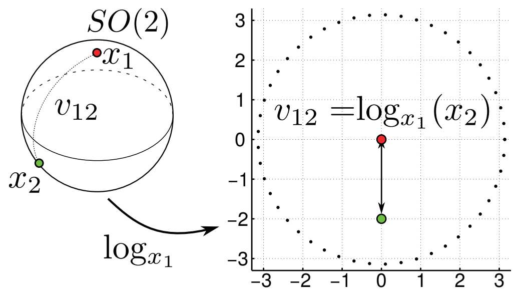

Figure 2. Visualization of the velocity. Velocities correspond to points on the tangent space at x_1. Given a consecutive point x_2, the velocity is the curve going from x_1 to x_2 which is equivalent to a straight line in the tangent space. The modulus of v_{12} on the tangent space corresponds to the geodesic distance. The conditional distribution $p(v|x,\theta)$ is also defined on this tangent space.

and 2π. In this case the distance using the angular value would be 2π while the true geodesic distance is 0. Our approach can handle both these limitations.

We show an overview of different models in Table 1. We can see that our model scales well while being consistent with the manifold, and has low complexity, i.e., it just considers a single hyperparameter and can be easily learnt using an Expectation-Maximization algorithm. It is worth noting here that our model is also the fastest of them for sampling (it is $\mathcal{O}(1)$).Our Matlab implementation allows obtaining over 100,000 samples per second.

2 Kinematic Prior

We will now describe the way we use the manifold to learn a model which can then be used as a strong kinematic prior for tracking.

2.1 Joint Pose and Kinematic Manifold

We model 3D human pose using the $SO(2)^n$ manifold [13], where n is the number of joints. This representation consists of modeling each joint as a unit sphere in which we have only two rotational degrees of freedom. By not taking into account the limb lengths (distance between two neighboring joints) we obtain an individual-agnostic representation. The natural metric for comparing poses is the geodesic distance, i.e., the shortest distance between two points on that manifold. Note the fact angles periodicity has no effect on this metric: 0 and 2π have a distance of 0.

The tangent plane to $SO(2)^n$ is its Lie algebra $\mathfrak{so}(2)^n$. The tangent plane and consequently the Lie algebra is local to a specific point, and in our case this point corresponds to a particular pose. Given two consecutive poses x_1 and x_2 at a constant framerate, we can compute the velocity v_{12} between x_2 and x_1 on the Lie algebra using the logarithm map at x_1 as

$$v_{12} = \log_{x_1}(x_2), \quad x_1, x_2 \in SO(2)^n, \quad v_{12} \in \mathfrak{so}(2)^n$$

where $\|v_{12}\|$ is the geodesic distance between both points. Please see Fig. 2 for a visual representation.

The joint manifold for both pose and kinematics will therefore become $SO(2)^n \times \mathfrak{so}(2)^n$. In order to be able to perform clustering we have to define both the exponential map and logarithm map for the manifold. In this case the pose and kinematic submanifolds can be handled independently. For the pose manifold refer to [13]. The kinematic manifold is defined already on the tangent space and therefore Euclidean metrics can be directly used (remember that geodesic and Euclidean distances are the same in this case). Therefore we can define the maps as

$$\log_{v_1}(v_2) = v_2 - v_1 \quad \text{and} \quad \exp_{v_1}(v_2) = v_2 + v_1$$

This will ensure that the mean of the data on the tangent space at the geodesic mean will be 0.

2.2 Probabilistic Model

With data on a known manifold we can use the publicly available algorithm of [13] to perform unsupervised clustering while taking into account the underlying manifold structure. By using the proposed pose and kinematic manifold we are effectively learning the joint probability

$$p(x, v|\theta) = \sum_{k=1}^{K} \alpha_k p(x, v|\theta_k) = \sum_{k=1}^{K} \mathcal{N}_{\mu_k}(0, \Gamma_k)$$

where $\theta = (\mu, \Gamma)$ and α are the parameters of the model and K is the number of clusters. Each $p(x, v|\theta_k)$ corresponds to a cluster on a different tangent plane centered on μ_k. In particular, we model each cluster as a Gaussian with zero mean and concentration matrix Γ_k. Note that while the mean is zero, the cluster is centered on a tangent space which effectively makes the point μ_k the mean of the Gaussian.

The model parameters are learnt by a variant of the Expectation-Maximization (EM) algorithm with a Minimum Message Length (MML) criterion that is also able to select the number of clusters K. This is done by initializing the number of clusters to a large value and then proceeding to run the EM algorithm until convergence. Afterwards the weakest cluster is eliminated and the EM algorithm is repeated. At any point of the optimization, clusters that are not well supported by the data can be eliminated. Finally, the model with the lowest overall energy (including the MML criterion) is chosen. By doing this the method is able to find a good balance between complexity and expressiveness.

2.3 Conditional Distribution

Even though we estimate the joint model, we are interested in computing the conditional probability distribution

$$p(v|x,\theta) = \frac{p(x,v|\theta)}{p(x|\theta_x)} = \frac{\sum_{k=1}^{K} \alpha_k p(x|\theta_{k,x}) p(v|x,\theta_k)}{\sum_{k=1}^{K} \alpha_k p(x|\theta_{k,x})} \ .$$

Observe that this is indeed a new mixture model $p(v|x,\theta) = \sum_{k=1}^{K} \pi_k p(v|x,\theta_k)$, where the weights have changed to

$$\pi_k = \frac{\alpha_k p(x|\theta_{k,x})}{\sum_{j=1}^{K} \alpha_j p(x|\theta_{j,x})} \quad .$$

It is important to note that while the Gaussians were originally centered at 0, that is no longer necessarily the case. In general, for $p(v|x,\theta_k) = \mathcal{N}_{\mu_v}(\mu_{v|x}, \Gamma_{v|x})$ where $\Gamma_k = \begin{bmatrix} \Gamma_{k,x} & \Gamma_{k,vx} \\ \Gamma_{k,vx} & \Gamma_{k,v} \end{bmatrix}$. We can compute these new distributions as

$$p(v|x,\theta_k) = \mathcal{N}_{\mu_v}(\Gamma_{k,vx}\Gamma_{k,x}^{-1} \log_{\mu_{k,x}}(x_x), \\ \Gamma_{k,v} - \Gamma_{k,vx}\Gamma_{k,x}^{-1}\Gamma_{k,vx}) \ .$$

Computing these conditional probability models is done in closed form and hence, very efficiently.

The model can then be run in two different ways: pure generative fashion to sample hypotheses with complexity $O(1)$; or in a discriminative manner in which the log-likelihood of a sample is computed with complexity $O(K)$. That is, we can either generate hypotheses or score them. Although the number of clusters K is generally low and thus estimating the log-likelihood not too computationally expensive, sampling is extremely fast and is the preferred approach.

3 Results

We have evaluated our approach on the Human3.6M dataset [5, 6] which is a large dataset containing 11 actors performing various actions. A motion capture system is used to provide an accurate 3D ground truth. We show both qualitative and quantitative results in which we compare our log-likelihood model against the widely used Gaussian diffusion model.

We model the pose as a an articulated body of 15 joints. We normalize the data by using the center of the hip as the coordinate origin and rotate it so that the normal of the plane formed by the hip joints and the neck joint is aligned with the principal axis. The secondary axis is defined by the line from the center of the hip joints to the neck joint. This gives us a unique local representation of the pose which we are then able to model with the $SO(2)^{12} \times SO(1)^2$ manifold as two joints only have one degree of freedom each. Extending this with the Lie algebra for the joint kinematics we finally obtain the $SO(2)^{12} \times SO(1)^2 \times \mathfrak{so}(2)^{12} \times \mathfrak{so}(1)^2$ manifold we use. Instead of using a full covariance matrix, we simplify by using the block diagonal approach as in [13]. Therefore, each covariance matrix has 92 degrees of freedom. Note that the kinematics and the pose are very different in magnitude. In order to avoid fitting the model to the dominant data we scale them both in the tangent space so they are roughly consistent. In particular we multiply the kinematics in the tangent space by a constant factor of 30.

Table 2. Comparison of different priors. We compare against the widely used Gaussian diffusion, trained both globally for all joints and individually for each joint. For our model, in parentheses we show the percentage of the training set we are considering, and the final number of estimated clusters. For the Gaussian diffusion models we do not perform subsampling of the training set.

Method	Log-likelihood Train	Test
Samples	465,325	62,064
Gaussian diffusion	5.4325	5.4349
local Gaussian diffusion	6.4193	6.4206
Ours (30%, 211 clusters)	9.3382	11.7874
Ours (15%, 147 clusters)	8.9544	**11.8714**

We split the dataset using a leave-one-person-out scheme. That is, we use all 15 categories of actions, each comprised of two subcategories, for actors 5, 6, 7, 8, 9, and 11 for the training set, and use actor 1 for the test set. The diversity of the actions makes the dataset very challenging to learn. This gives us 465,325 frames for training and 62,064 frames for testing. Since the frames are highly correlated because motion are smooth, we perform a random subsampling before training our model.

Additionally, we provide quantitative results by looking at the expected log-likelihoods on the test dataset. We compare against the Gaussian diffusion approach both trained on a global level (single Gaussian is averaged for all joints) and on a more local level (single Gaussian is averaged for each joint independently). We additionally train several kinematic models with different degrees of subsampling of the training data, and show the results in Table 2. A subsampling of 15% corresponds to 69,799 training samples, roughly the same amount as the test set. We can see that the local Gaussian diffusion model outperforms the Gaussian diffusion model with a single parameter. However, our model outperforms both of them by a considerable margin. It is interesting to note that the log-likelihood of the test set is higher than that of the training set. This can be explained by the presence of actors that are outliers and are not as well captured by the model. On the other hand actor 1 seems to be well represented by the other actors. Increasing the number of samples does increase the number of clusters in the model, but does not significantly change the performance on the test set. This is an indication that subsampling might be an easy way to obtain more simple models that still can generalize well for datasets in which there is a high correlation between poses due to the temporal component.

We finally depict some qualitative examples in Fig. 3. We sample directly from $p(v|x,\theta)$ for several frames. It is worth noting that we can obtain 100,000 samples in 0.85 seconds on a Intel Core i7 2.93GHz CPU using a Matlab implementation.

4 Conclusions

We have presented a novel kinematic prior for 3D human pose tracking based on extending a pose man-

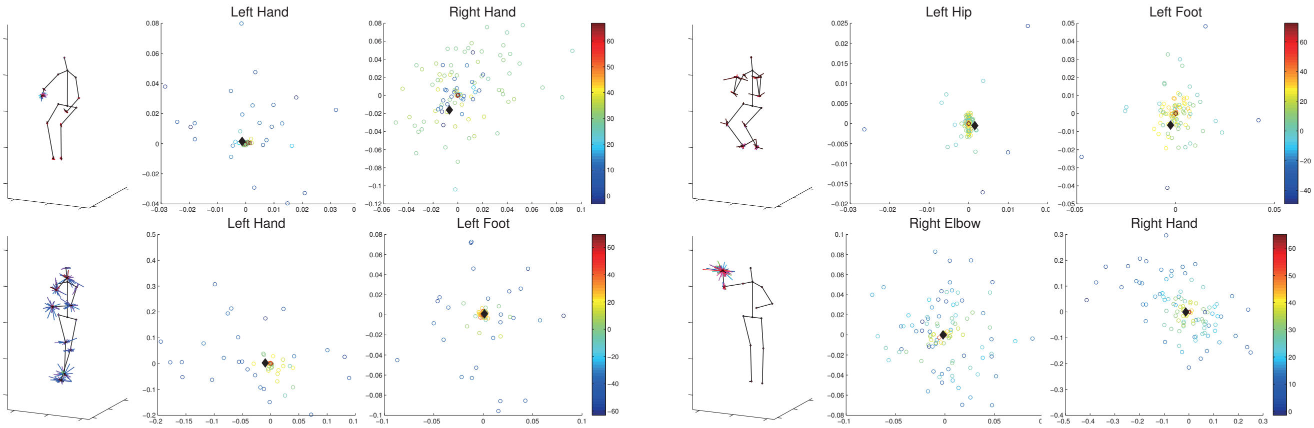

Figure 3. Qualitative examples. We show several examples from the test set using the 15% subsampled model with 147 clusters. We visualize the ground truth and 100 samples from our model in 3D. For visualization purposes the velocity is scaled by 10 and the samples are scaled by 3. We also show the distribution of the samples on the tangent space for some of the joints, scored by their log-likelihood with the ground truth as a black diamond.

ifold with its Lie algebra. By exploiting the fact that the pose manifold is well known and defined, we can use a simple mixture model defined on the manifold. We show that our approach is able to scale well to large datasets and can be sampled at a rate of over 100,000 samples per second, making it ideal for real-time applications. We show quantitative results that demonstrate a large improvement over the widely used Gaussian diffusion models. Furthermore, it is straightforward to extend existing 3D human pose estimation algorithms [11, 12] to tracking using the proposed prior using stronger image features [14, 15].

While we have centered this work on 3D human pose tracking, the framework we presented is general for tracking data on other manifolds. Additionally, it would be simple to extend our model to predict a pose from multiple previous frames, to also modeling acceleration or other higher derivatives. We believe the simplicity and the results of the proposed approach make it a powerful tool for improving any sampling-based tracking method.

Acknowledgements: This work has been partially funded by Spanish Ministry of Economy and Competitiveness under projects PAU+ DPI2011-27510, TextilRob 201550E028, and by the ERA- Net Chistera project ViSen PCIN-2013-047.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D Pose Estimation and Tracking by Detection. In *CVPR*, 2010.

[2] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2005.

[3] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *IJCV*, 87:75–92, March 2010.

[4] S. Hauberg, S. Sommer, and K. S. Pedersen. Natural metrics and least-committed priors for articulated tracking. *Image and Vision Computing*, 30(6):453–461, 2012.

[5] C. Ionescu, F. Li, and C. Sminchisescu. Latent Structured Models for Human Pose Estimation. In *ICCV*, 2011.

[6] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014.

[7] N. D. Lawrence. Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *JMLR*, 6:1783–1816, 2005.

[8] R. Li, T.-P. Tian, S. Sclaroff, and M.-H. Yang. 3d human motion tracking with a coordinated mixture of factor analyzers. *IJCV*, 87(1-2):170–190, 2010.

[9] J. Quiñonero-candela, C. E. Rasmussen, and R. Herbrich. A Unifying View of Sparse Approximate Gaussian Process Regression. *JMLR*, 6:1939–1959, 2005.

[10] L. Sigal, M. Isard, H. W. Haussecker, and M. J. Black. Loose-limbed People: Estimating 3D Human Pose and Motion Using Non-parametric Belief Propagation. *IJCV*, 98(1):15–48, 2012.

[11] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer. A Joint Model for 2D and 3D Pose Estimation from a Single Image. In *CVPR*, 2013.

[12] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer. Single Image 3D Human Pose Estimation from Noisy Observations. In *CVPR*, 2012.

[13] E. Simo-Serra, C. Torras, and F. Moreno-Noguer. Geodesic Finite Mixture Models. In *BMVC*, 2014.

[14] E. Simo-Serra, C. Torras, and F. M. Noguer. DaLI: Deformation and Light Invariant Descriptor. *IJCV*, pages 1–19, 2015.

[15] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, and F. M. Noguer. Fracking Deep Convolutional Image Descriptors. *CoRR*, abs/1412.6537, 2014.

[16] C. Sminchisescu and B. Triggs. Estimating Articulated Human Motion With Covariance Scaled Sampling. *IJRR*, 22(6):371–391, 2003. Special issue on Visual Analysis of Human Movement.

[17] G. Taylor, L. Sigal, D. Fleet, and G. Hinton. Dynamical binary latent variable models for 3d human pose tracking. In *CVPR*, 2010.

[18] R. Urtasun, D. J. Fleet, and P. Fua. 3D People Tracking with Gaussian Process Dynamical Models. In *CVPR*, 2006.

[19] J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models. In *NIPS*, 2005.

[20] A. Yao, J. Gall, L. V. Gool, and R. Urtasun. Learning probabilistic non-linear latent variable models for tracking complex activities. In *NIPS*, 2011.