Toward Continuous-Time Representations of Human Motion

Weiyu Du(✉), Oleh Rybkin, Lingzhi Zhang, and Jianbo Shi

University of Pennsylvania, Philadelphia, USA
{weiyudu,oleh,zlz,jshi}@seas.upenn.edu

Abstract. For human motion understanding and generation, it is common to represent the motion sequence via a hidden state of a recurrent neural network, learned in an end-to-end fashion. While powerful, this representation is inflexible as these recurrent models are trained with a specific frame rate, and the hidden state is further hard to interpret. In this paper, we show that we can instead represent the *continuous* motion via latent parametric curves, leveraging techniques from computer graphics and signal processing. Our parametric representation is powerful enough to faithfully represent continuous motion with few parameters, easy to obtain, and is effective when used for downstream tasks. We validate the proposed method on AMASS and Human3.6M datasets through reconstruction and on a downstream task of point-to-point prediction, and show that our method is able to generate realistic motion. See our demo at www.github.com/WeiyuDu/motion-encode.

Keywords: Motion generation · 3D Motion · Sequence representation

1 Introduction

Human motion understanding and generation techniques commonly employ recurrent neural networks [2,6] that aggregate information across the temporal domain by processing each frame step-by-step [1,4,5,11,14]. While this allows using powerful black-box neural networks for the task at hand, such as action classification or future prediction, the representation is also inflexible in several key ways. It requires operating at a fixed frame rate, while real-world data often have different or even variable frame rates, which makes current systems cumbersome in practice. Similarly, when used for motion generation, the generated motion is restricted to the training frame rate, and different networks need to be trained for different temporal resolution.

Instead of taking this per-frame perspective, we argue that human motion should be represented holistically in a *continuous* manner, using latent parametric curves. We introduce two motion encoding schemes based on classical techniques from computer graphics and signal processing, namely Bezier and Sine Motion Encoding that represent the motion with Bezier and Sine curves respectively. Crucially, we apply this parametrization in a latent space instead of

A. Bartoli and A. Fusiello (Eds.): ECCV 2020 Workshops, LNCS 12540, pp. 543–548, 2020.
https://doi.org/10.1007/978-3-030-65414-6_37

the original joint space. This combination of a powerful per-step latent encoding with a simple and interpretable parametric temporal encoding makes our approach both powerful and flexible. Our approach can be used with input sequences of any framerate, or even variable framerate. When used for generation, it further enables us to generate frames at any desired rates and timestamps. Moreover, as these curves only require a few parameters, they provide a compact representation compared to the original full-size embedding.

We also study the task of controllable human motion generation with endpoints as an application of our proposed method. In animation production, in order to generate an animated motion, artists usually define key frames for the character's pose and design a trajectory using spline interpolation. Motion generation with end points can greatly expedite this process as it can automatically fill in the blanks between two poses with relatively long, realistic motion and reduces the number of key frames needed from the artists.

Experiments on AMASS [10] and Human3.6M [7] data show that our model with Bezier and Sine Motion Encoding beats latent linear interpolation baseline by a large margin both visually and with joint angle mean squared error. Our model generates realistic and smooth motion on the AMASS and Human3.6 datasets. Please see video results on the demo website.

2 Method

We want to encode and represent a motion sequence $x_{1:M}$ and time stamps $t_{1:M}$, where $x_m \in R^N$ is an individual pose at time t_m. First, we trained a Variational Auto-Encoder (VAE, [8,13]) that encodes individual frames $x_{1:M}$ into per-frame latent codes $z_{1:M}$. Given this per-frame encoding, we want to find a representation of the continuous sequence F such that $F(t)$ approximates the latent sequence $z_{1:M}$.

2.1 Bezier Motion Encoding

We first evaluate Bezier Motion Encoding, a technique inspired by classic computer graphics techniques [3]. The encoding is defined as the set of control points P for Bezier curves in the latent space of pose VAE. To ensure maximal expressiveness, we model a time channel in addition to the latent dimensions. To generate pose at different timestamps, we discretize the curve by taking 1000 samples in time channel and take the latent code with closest matching time. Formally, Bezier Motion Encoding is defined as $F_{bezier}(t) = B(s)\ s.t.\ t = T(s), B(s) = \sum_{i=0}^{n} \binom{n}{i} (1-s)^{n-i} s^i P_i$ where t is time, $T(s)$ is the matching process in time domain, $0 \leq s \leq 1$, P is the set of control points and n is its size. We do not use s to represent time because s is not evenly distributed along the curve. Taking s as time flattens the curve, which limits the expressiveness of the encoding.

Bezier Motion Encoding has several advantages: 1) The curve begins at P_0 and ends at P_n, which is desirable in the controllable motion generation application. 2) Displacement of control point in a direction corresponds to a smooth drag of the curve. However, Bezier curve has global control points, which makes it hard to adjust the curve locally in detail. This can result in overly smooth latent trajectories, which have trouble modeling highly subsampled motion sequences.

2.2 Sine Motion Encoding

Motivated by the shortcomings of the Bezier encoding, we further evaluate Sine Motion Encoding that represents the curve via the most salient frequences in the frequency domain, as common in signal processing techniques [12]. The Sine encoding is defined as a linear combination of Sine curves $F_{sin}(t) = \sum_{i=0}^{n} A\sin(\omega t + \phi)$, where n is the number of Sine curves, A is amplitude, ω is angular frequency and ϕ is phase. Sine curves are periodic and smooth and a linear combination of them can model complex signals with few parameters. We can also increase the level of complexity and details in encoding by using more Sine curves. This is hard to achieve by Bezier encoding.

2.3 Optimization

Given motion sequence $x_{0:T}$, we obtain Bezier or Sine Motion Encoding from the following optimization: $\min_F \sum_m ||F(t_m) - z_m||^2$ where F is either F_{bezier} or F_{sin} defined in the above sections.

2.4 Controllable Human Motion Generation

To evaluate the proposed encodings on a downstream task, we study the task of controllable human motion generation with endpoints. Given a pair of poses (x_1, x_M), the task is to fill in the motion sequence $x_{2:M-1}$ in between.

We first embed the input pose pair with the pre-trained pose VAE to latent codes (z_0, z_M). Then we use a Multilayer Perceptron (MLP) that takes this as input and outputs a Gaussian distribution for control points P in the case of Bezier Motion Encoding and A, ω, ϕ in the case of Sine Motion Encoding. We use the ground truth latent trajectory and motion sequence as supervision. The loss is formulated as follows

$$\mathcal{L}_{gen} = ||F_{0:M} - z_{0:M}||^2 + ||\mathrm{Dec}(F_{1:M}) - x_{1:M}||^2 - \mathrm{D}_{\mathrm{KL}}(\mathcal{N}(\hat{\mu}, \hat{\sigma})||\mathcal{N}(0,1)) \quad (1)$$

where Dec is the pose VAE decoder, $\hat{\mu}, \hat{\sigma}$ is the output of MLP. The first two terms are reconstruction loss and the last term helps as regularization.

3 Experiments

We experiment on Human3.6M [7] and AMASS [10] datasets in SMPL [9] format. We use linear interpolation in the latent space of pre-trained pose VAE as baseline. We split the AMASS data into 30-frame sequences with interval of 10 frames. On Human3.6M, we subsample the data 5 times to evaluate our method

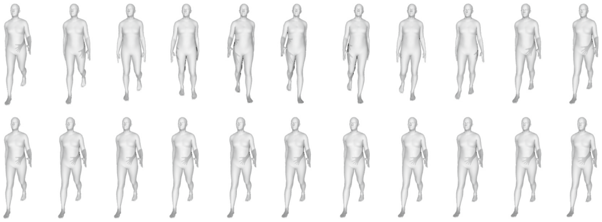

Fig. 1. Visualization on AMASS test set. Our Sine Motion Encoding (top) encodes walking motion while baseline (bottom) fails. On the website, we show further generations, including generating frames at a higher frequency than training data.

on long-term motion, then split it using the same scheme. We use 4 free control points for Bezier Motion Encoding and 3 curves for Sine Motion Encoding. To model more local and detailed movements, we experiment with adding 3 more curves to Sine Motion Encoding, where they only contribute to local portions of the sequence. We use MSE on joint angles as evaluation metric.

3.1 Motion Representation

We first evaluate the ability of our method to represent continuous motion. Our results are shown in Table 1, first row, as well as on the demo website. We see that our method is able to faithfully reproduce the encoded motion, demonstrating potential for sequence representation learning for many downstream tasks.

Table 1. Row 1: motion reconstruction error from Bezier and Sine encoding on AMASS test set. Row 2–3: Motion generation error in joint angle MSE per sequence. Our motion encodings generate motions with better visual quality and smaller error.

	LERP	Bezier	Sine (3 curves)	Sine (6 curves)
Reconstruction	–	1.89	2.87	1.32
AMASS	194.13	15.91	14.90	**14.43**
Human3.6M	307.66	30.25	29.28	**27.19**

3.2 Controllable Motion Generation

We show that our representation is suitable for point-to-point motion generation in Table 1. In Fig. 1 and on the demo website, we see that latent linear interpolation baseline (LERP) can only generate smooth transitions, while parametric latent curves capture the diversity in the data, and can even model cyclic motion where the initial and ending poses are similar. Also qualitatively, we observe that only the Sine (6 curves) representation models the Human3.6 data faithfully.

This method is more powerful, and better captures high-frequency details (such as in fast walking) that contribute little to quantitative metrics but are crucial for visual quality.

3.3 Variable Frequency Motion Generation

An advantage of our approach for motion generation is that it is possible to sample the generated curve at a frequency different than training data. To demonstrate this, we show the same predicted sequence that is sampled with 12 frames per second, 30-frame-long (training data) and 24 frames per second, 60-frame-long from our predicted representation. Our method produces good quality motion even on higher temporal resolution. This improves the visual quality of the motion by making it smoother. The video result of this experiment is shown on the demo website.

References

1. Aksan, E., Kaufmann, M., Hilliges, O.: Structured prediction helps 3D human motion modelling. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7144–7153 (2019)
2. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1724–1734. Association for Computational Linguistics, October 2014. https://doi.org/10.3115/v1/D14-1179. https://www.aclweb.org/anthology/D14-1179
3. Foley, J.D., et al.: Computer Graphics: Principles and Practice, vol. 12110. Addison-Wesley Professional, Boston (1996)
4. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4346–4354 (2015)
5. Ghosh, P., Song, J., Aksan, E., Hilliges, O.: Learning human motion models for long-term predictions. In: 2017 International Conference on 3D Vision (3DV), pp. 458–466. IEEE (2017)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
7. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
8. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. CoRR abs/1312.6114 (2014)
9. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (2015)
10. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: archive of motion capture as surface shapes. In: International Conference on Computer Vision, pp. 5442–5451, October 2019
11. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2891–2900 (2017)

12. Proakis, J.G.: Digital Signal Processing: Principles Algorithms and Applications. Pearson Education India (2001)
13. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and variational inference in deep latent Gaussian models. In: International Conference on Machine Learning, vol. 2 (2014)
14. Wang, T.H., Cheng, Y.C., Lin, C.H., Chen, H.T., Sun, M.: Point-to-point video generation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 10491–10500 (2019)