Long-term Human Motion Generation and Reconstruction Using Graph-based Normalizing Flow

Wenjie Yin[1], Hang Yin[1], Danica Kragic[1], Mårten Björkman[1]

***Abstract*— Modeling human motion remains challenging for synthesizing high-quality samples robustly from imperfect input. We propose a probabilistic generative model to generate and reconstruct motion sequences given past information and control signals. The model adapts MoGlow by leveraging graph-based model to capture the spatial and temporal information of skeleton. We evaluate the model performance on a mixture of human locomotion dataset with foot-step and bone-length analysis. The results show comparable results on generating long-term pose sequences, with improved robustness when generating and reconstructing from imperfect inputs.**

I. INTRODUCTION

Capturing human motion patterns is essential in animating synthetic characters [1] and understanding behaviors for social robotics [2]. Recent deterministic motion synthesis methods [3], [4] are often limited to generate average poses and fail to capture the natural variability. Probabilistic generative models allow modelling of the full space without collapsing to an stereotypical pose [5], [6]. Normalizing flow based methods allow tractable likelihood evaluation and efficient parameterization, however, have rarely been explored for human motion compared to alternatives. Our model builds upon MoGlow [1], an autoregressive normalizing flow model.

One challenge that generative models are facing is imperfect information. Under less controlled environments, motion capture (MoCap) systems inevitably suffer from missed markers [7]. Unfortunately, most previous works are unsatisfactory in generating stable motion under such conditions. Our model exploits the invariant spatial correlation of human skeletons with a graph model to address this limitation.

The proposed framework conditions on control signals and can generate diverse long-term human motion sequences. When the past sequences are incomplete, the generation is still robust and the missing markers can be reconstructed by reversing the generation. We evaluate our framework on a mixture of human locomotion datasets. The evaluation shows a generation and reconstruction quality close to ground truth, outperforming baseline under imperfect input data.

II. METHOD

Fig. 1 gives an overview of the graph-based motion glow for generation and reconstruction. Our flow model includes the three main reversible transformation layers of Glow, but extended to graph structures. We further use ST-GCN [9] to extract features from the autoregressive history input. More implementation details can be found in [10].

[1]Robotics, Perception and Learning lab, KTH Royal Institute of Technology, Sweden. {yinw, hyin, dani, celle}@kth.se.

The input X and output z are represented as tensors with spatial dimension M, channel dimension C and temporal dimension T_h. In MoGlow, the coordinates of one frame skeleton data are concatenated to one vector. We convert the skeleton data into an undirected graph $G = (V, E)$, where V is the set of nodes and E is the set of edges. The skeleton graph used in this work is illustrated in Fig. 2. With a spatial graph neural network (S-GCN), a spatial graph convolution can then be defined as:

$$y_i = \sum_{v_j \in S_i} \frac{x_j}{D_{v_i}(v_j)} w(l_{v_i}(v_j)), \tag{1}$$

where x_i and y_i are the feature vector of node v_i before and after the convolution, S_i is the neighbor set of v_i, and w is a weight function.

The actnorm layer performs data-dependent initialization and invertible 1×1 convolution layer performs a soft permutation of channels. X_a and X_b are the outputs of these two layers. The affine coupling layer transforms half of the input X_{b1} based on the other half X_{b2} and conditioning information, h is the output of this layer. The coupling can then be written

$$[h_1, h_2] = [X_{b1}, (X_{b2} + \mathbf{b}) \odot \mathbf{s}], \tag{2}$$

where $\odot$ is a Hadamard product and the scaling $\mathbf{s}$ and bias $\mathbf{b}$ are computed with S-GCN, ST-GCN and LSTM:

$$g_t = SGCN(X_{b1,t}), \tag{3}$$

$$p_t = STGCN(\hat{X}_{(t-T_h):(t-1)}), \tag{4}$$

$$[\mathbf{s}_t, \mathbf{b}_t] = LSTM(g_t, p_t, C_{(t-T_h):(t)}). \tag{5}$$

Here, S-GCN captures the spatial graph information g_t from markers in the current time step t, ST-GCN extracts spatial-temporal features p_t from the past sequence, and LSTM produces the scaling and bias with dependencies over time.

For generation, we can generate a future pose from the trained model using a latent vector z_t since the flow model is reversible. The generated X_t then becomes a part of conditioning information for the next pose X_{t+1}. During training, the motion data was augmented by lateral mirroring and time-reversion. We reconstructed the incomplete poses using the same model by reversing the generated sequences $\mathbb{X}_{(t_0):(T_h)}$ and control signal $C_{(t_0-T_h):(T_h)}$ to $\mathbb{X}_{(t_h):(T_0)}$ and $C_{(T_h):(t_0-T_h)}$. The reversed sequences are regarded as control information to generate markers to fill the holes of missing markers.

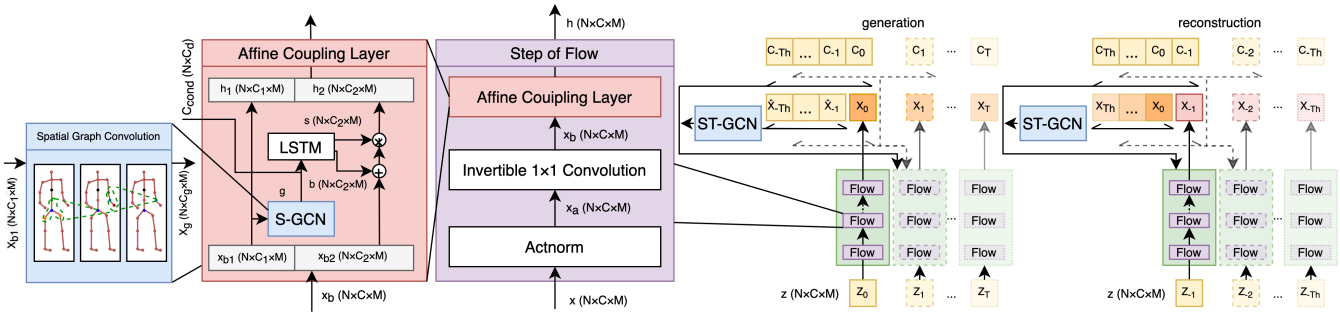
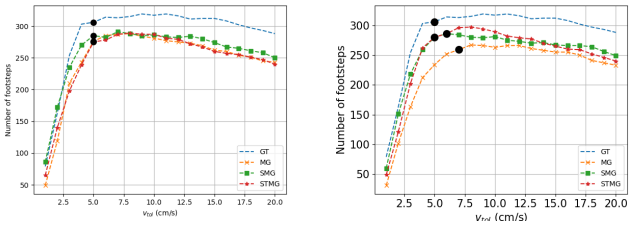


Fig. 1: The overview of the framework for skeleton-based motion generation and reconstruction.

Fig. 2: The spatial graph of the human skeleton in this paper. Each node represents the body marker on the right [8].



(a) (b)

Fig. 3: Footstep analysis for complete (a) and incomplete (b) past input: footstep count f_{est} for each tolerance v_{tol}. Black dots indicate v_{tol}^{95}.

Miss	Model	f_{est}	v_{tol}^{95}	μ	σ
-	GT	5	306	0.315	0.273
Complete	MG	5	276	0.298	0.318
	SMG	5	**285**	0.294	0.242
	STMG	5	275	**0.316**	**0.267**
Incomplete	MG	7	256	0.357	0.329
	SMG	5	280	0.307	0.252
	STMG	6	**286**	**0.315**	**0.253**

TABLE I: Results of foot-step analysis for motion generation: total number of footsteps f_{est}, speed tolerance for capturing 95% steps v_{tol}^{95}, mean μ and standard deviation σ of step-duration. The numbers closest to the ground truth are shown in bold.

Data	Model	Generation		Reconstruction	
		RMSE	σ	RMSE	σ
Complete	MG	0.597	0.067	-	-
	SMG	**0.191**	**0.039**	-	-
	STMG	0.779	0.073	-	-
Incomplete	MG	787638	12.138	80931	2.897
	SMG	**0.542**	**0.044**	6.589	0.092
	STMG	0.938	0.080	**1.072**	**0.065**

TABLE II: Results of bone-length analysis for human motion generation and reconstruction. The best values are in bold.

III. RESULTS

We consider a human locomotion dataset preprocessed in [1]. The pose is represented by 21 marker coordinates as shown in Figure 2 and the trajectory is represented 3 scalar control signals include forward, sideways and rotational velocities for each frame. To generate incomplete MoCap frames, we set some markers to zero. We consider three configurations. Our proposed graph-based model denoted as STMG, the SMG without temporal convolution and MG uses no graph structure. A video with generated examples can be found at this link[2]

Footsteps analysis (see [1] for detailed definitions) is used to evaluate foot-sliding artifacts in locomotion synthesis. Footsteps can be detected as time intervals where the horizontal speed of the heel joints is below a tolerance value v_{tol}. We incremented the tolerance v_{tol} in small steps. We note that in Fig. 3-(a), without missing markers, the curves are close. However, when the given past data is incomplete, in Fig. 3-(b) the curves of our STMG are closer to the curve of ground truth. For evaluation, we detect the first tolerance value v_{tol}^{95}, for which at least 95% of the maximum number of footsteps are estimated. These values are shown as black dots in the figures. The results are also shown in Table I.

We further perform bone-length analysis to detect artifacts such as flying-apart joints. As illustrated in Table II, with complete past frames, all models achieve relative small RMSE and σ of bone lengths. Again, given untrained imperfect data, MG performs poorly on this indicator, exhibiting huge bone-length artifacts. For both SMG and STMG, the RMSE and σ of bone-lengths are still small.

IV. CONCLUSION

We propose a graph-based normalizing flow model to tackle the limitations in earlier models of human motion generation and reconstruction. This new modelling framework is an extension of MoGlow that is probabilistic and allows inference of the exact likelihood. It utilizes graph convolutional networks to improve the robustness of generation. In the future, we plan to extend the graph-based motion glow model to multiple scales to tackle more complex motions.

[2]https://kth.box.com/s/ylilqe0b5rbygo73rbi8gqetb2mhl7ls

ACKNOWLEDGEMENTS

This research has received funding from the EC Horizon 2020 research and innovation program under grant agreement n. 824160 (EnTimeMent).

REFERENCES

[1] G. E. Henter, S. Alexanderson, and J. Beskow, "Moglow: Probabilistic and controllable motion synthesis using normalising flows," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–14, 2020.

[2] R. Murakami, L. Y. Morales Saiki, S. Satake, T. Kanda, and H. Ishiguro, "Destination unknown: Walking side-by-side without knowing the goal," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 471–478.

[3] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4346–4354.

[4] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom, "Deep representation learning for human motion prediction and classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6158–6166.

[5] Z. Wang, J. Chai, and S. Xia, "Combining recurrent neural networks and adversarial training for human motion synthesis and control," *IEEE transactions on visualization and computer graphics*, vol. 27, no. 1, pp. 14–28, 2019.

[6] H. Ahn, J. Kim, K. Kim, and S. Oh, "Generative autoregressive networks for 3d dancing move synthesis from music," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3500–3507, 2020.

[7] T. Kucherenko, J. Beskow, and H. Kjellström, "A neural network approach to missing marker reconstruction in human motion capture," *arXiv preprint arXiv:1803.02665*, 2018.

[8] I. Habibie, D. Holden, J. Schwarz, J. Yearsley, and T. Komura, "A recurrent variational autoencoder for human motion synthesis," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[9] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[10] W. Yin, H. Yin, D. Kragic, and M. Björkman, "Graph-based normalizing flow for human motion generation and reconstruction," 2021.