Spatio-temporal priors in 3D human motion

Anna Deichler*
KTH Royal Institute of Technology

Kiran Chhatre*
KTH Royal Institute of Technology

Jonas Beskow
KTH Royal Institute of Technology

Christopher Peters
KTH Royal Institute of Technology

When we practice a movement, human brains creates a motor memory of it. These memories are formed and stored in the brain as representations which allows us to perform familiar tasks faster than new movements. From a developmental robotics and embodied artificial agent perspective it could be also beneficial to exploit the concept of these motor representations in the form of spatial-temporal motion priors for complex, full-body motion synthesis. Encoding such priors in neural networks in a form of inductive biases inherit essential spatio-temporality aspect of human motion. In our current work we examine and compare recent approaches for capturing spatial and temporal dependencies with machine learning algorithms that are used to model human motion.

The examined algorithms operate on motion capture data, which contain human motion in the form of data structured as $X \in \mathbb{R}^{T \times V \times C}$, where V denotes the number of joints, C is the dimension of the time series (dependent on the motion representation eg. 3D position, joint angles) and T is the number of frames in the motion clips. In [1], Graph Convolutional Networks (GCN) are used to model human motion as an undirected spatio-temporal graph $G = \{V, E\}$, where V represents the joint nodes and E represents the edges for inter as well as intra-frame connections. In GCN, the adjacency matrix (A) creates explicit connections between the joints. The prior of having a static adjacency matrix, A such that: $A \in \{e_{i,j} = 1\}$, where the indices denotes two connected joints, has the disadvantage that it models only partial high level features and disregards the fact that the contribution of the connected joints is not equal. Additionally, the implicit relationships between the non-connected joints plays an important role while modeling human motion. To counter this, in [2] a new form of adjacency matrix A_p is initialized from original matrix A of the same dimension and a global graph matrix Q is initialized to learn all implicit relationships between the non-connected human joints (Equation 1).

$$f_{out} = \Lambda^{-\frac{1}{2}}[(A_p \circ M) + Q]\Lambda^{-\frac{1}{2}} f_{in} W, \tag{1}$$

where f_{in} and f_{out} are the input and output features between two consecutive layers, Λ is the matrix to normalize the self connected adapted adjacency matrix. The product between Λ and $(A_p \circ M) + Q$ is to normalize the connectivity between the nodes, whereas the product between the features f_{in} and linear projection W is to project the trajectory graph from a higher to a lower dimension.

On the other hand, the Transformer architecture was first used for modeling natural language, but since has gained popularity in a wide range of fields. The transformer model with its self-attention mechanism dynamically builds relations between and within joints of the input sequence (Equation 2).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where Q, K, and V are the query, key, and value embeddings for the skeleton joints.

Recently [3] and [4] have used Transformers to capture spatial-temporal dependencies in human motion capture data. [3] builds on [1] and motivates the use of Transformers to confront structural limitations of the ST-GCN architecture in capturing correlations. Spatial and temporal correlations are captured through the separate spatial and attention weight matrices, which express how much attention is given to joints in a single frame as well as across the frames for a single joint. In contrast the architecture in [4] only builds on the Transformer architecture and motivates its use for generating motions on much larger time horizons than previous methods.

In GCN based methods correlations are captured via the adjacency matrix and its variants, whereas in the Transformer based approaches the attention module contributes as the prior. These methods have consistently outperformed previous recurrent neural network based architectures for the motion recognition and prediction tasks. In future work we aim at building on these architecture and leverage them in the form of spatial-temporal priors for motion synthesis models in embodied artificial agents that exhibit more natural motion.

* authors contributed equally, {deichler, chhatre}@kth.se

REFERENCES

[1] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition." in AAAI, 2018

[2] Q. Cui, H. Sun and F. Yang, "Learning Dynamic Relationships for 3D Human Motion Prediction," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6518-6526, doi: 10.1109/CVPR42600.2020.00655.

[3] C. Plizzari, M. Cannici, M. Matteucci, Skeleton-based action recognition via spatial and temporal transformer networks, Computer Vision and Image Understanding, Volumes 208–209, 2021, 103219, ISSN 1077-3142, https://doi.org/10.1016/j.cviu.2021.103219.

[4] E. Aksan, P. Cao, M. Kaufmann, and O. Hilliges, (2020). Attention, please: A Spatio-temporal Transformer for 3D Human Motion Prediction. ArXiv, abs/2004.08692.