ACM SIGGRAPH / Eurographics Symposium on Computer Animation (2018)
T. Beeler and N. Thuerey (Editors)

Dilated Temporal Fully-Convolutional Network for Semantic Segmentation of Motion Capture Data

Noshaba Cheema[1,2,3], Somayeh Hosseini[1,2†], Janis Sprenger[1,2], Erik Herrmann[1,2], Han Du[1,2], Klaus Fischer[1] and Philipp Slusallek[1,2]

[1]DFKI Saarbrücken, [2]Saarland University, [3]Max-Planck Institute for Informatics; Germany

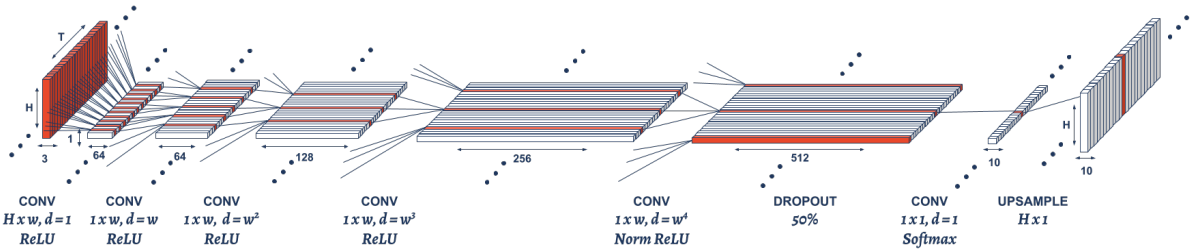

Figure 1: *Our dilated temporal fully-convolutional neural network (DTFCN) for motion capture segmentation. The initial layer consists of a traditional 2D convolutional layer. The next layers are 1D temporal acausal convolutions with dilation. The final dilated conv layer uses a normalizing ReLU activation function. Finally, we add a Softmax layer and upsample the output.*

Abstract

Semantic segmentation of motion capture sequences plays a key part in many data-driven motion synthesis frameworks. It is a preprocessing step in which long recordings of motion capture sequences are partitioned into smaller segments. Afterwards, additional methods like statistical modeling can be applied to each group of structurally-similar segments to learn an abstract motion manifold. The segmentation task however often remains a manual task, which increases the effort and cost of generating large-scale motion databases. We therefore propose an automatic framework for semantic segmentation of motion capture data using a dilated temporal fully-convolutional network. Our model outperforms a state-of-the-art model in action segmentation, as well as three networks for sequence modeling. We further show our model is robust against high noisy training labels.

CCS Concepts

•***Computing methodologies*** → ***Motion processing;*** *Motion capture; Image processing;*

1. Our Proposed Architecture

Recurrent neural networks (RNN) are the go-to method to model time-dependent sequences. However, one of their major drawbacks is the exploding and vanishing gradient problem and the difficulty to parallelize their training. Additionally, [BKK18] have shown that temporal convolutional networks (TCN) perform just as well or even better than RNNs in sequence modeling tasks. Hence, we introduce a model, which is inspired by traditional image segmentation approaches [LSD15] and recent advances in sequence modeling [BKK18] for semantic segmentation of motion capture data.

In a preprocessing step, we first transform our motion capture data to an RGB image domain, much in the spirit of [LB*17]. Each column of the image represents a frame in the motion sequence. The rows represent the joints and the RGB values are the scaled XYZ Euclidean coordinates of each corresponding joint. Such a *motion image* can be seen in Fig. 4 (top). We then pass it to our network (Fig. 1). Akin to the five areas in our visual cortex (V1 - V5) [Rem12], our model has a total of five convolution layers.

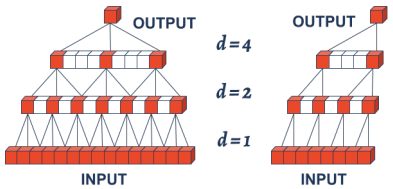

Figure 2: *Dilated convolution. Left: acausal dilation. Right: causal dilation. Systematic dilation increases the receptive field size exponentially.*

† First two authors contributed equally; email: ncheema@mpi-inf.mpg.de

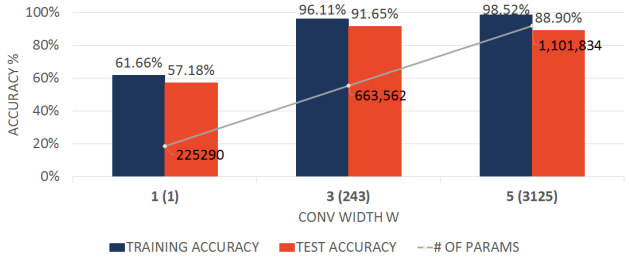

Figure 3: *Number of parameters (line) vs. train (dark blue) and test (red) accuracy for different convolution widths. The receptive field size with dilations after the first five layers is written in parentheses.*

	ED-TCN	WaveNet	TDNN	LSTM	Ours
Train	90.05%	90.64%	87.22%	86.32%	**96.11%**
Test	88.69%	88.47%	85.54%	81.95%	**91.64%**

Table 1: *Comparison against other models. ED-TCN: [L*17], WaveNet: [VDO*16], TDNN: [W*90], LSTM: [HS97]*

The initial layer consists of a traditional 2D convolutional layer which is only applied in the time dimension. To do that, we set the kernel height to the height of the image. Every layer has the same convolution width w with stride 1. The next four layers are 1D temporal acausal convolutions with dilation. The dilation rate d increases with each layer l, according to $d = w^{l-1}$. A convolutionized dense layer with a Softmax activation function is added after that. We found that a normalizing ReLU function [L*17] before the Softmax layer increases accuracy.

Fig. 2 shows how dilated convolutions increase the receptive field exponentially without loss of resolution for acausal and causal convolutions. Causal convolutions are used for temporal data, where the output depends on previous samples only. Since our goal is to distinguish motions like *left step* (step while walking) from *begin* and *end left step* (step from/to standing position), we use acausal convolutions, as these motion types rely on past and future information.

2. Experiments and Results

Our motion capture dataset consists of 70 sequences with 10 motion labels: *standing*, *left/right step*, *begin/end left step*, *begin/end right step*, *reach*, *retrieve* and *turn*. Our sequences reach up to 1500 frames. In all of our experiments, we use non-randomized 7-fold cross-validation. We use the Adam [KB14] optimizer with 100 epochs for training.

In order to determine the optimal receptive field size (RFS), we test our model on different convolution kernel widths w. Fig. 3 shows that even though a width $w = 5$ (RFS: 3125 frames) covers the entire sequence, the accuracy does not differ much from using $w = 3$ (RFS: 342 frames). A width $w = 3$ uses ∼438K fewer parameters, however. Since our model has to be robust against human error due to wrongly-classified labels, we further train our model on noisy labels and test it on the true labels. Fig. 5 shows despite adding 80% noisy labels in the training data, an accuracy of over 88% is reached on the true test labels for $w = 3$.

We test our model ($w = 3$) against another state-of-the-art TCN model [L*17] for action segmentation and three commonly used neural network models [VDO*16, HS97, W*90] for sequence modeling and classification using our dataset without noisy labels, and show that our model is superior to these models (Tab. 1).

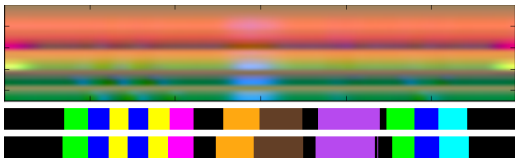

Figure 4: *Top: Motion capture sequence in RGB image domain. Middle: True labels. Bottom: Our predictions.*

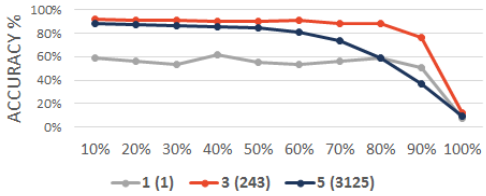

Figure 5: *Test accuracies for different receptive field sizes depending on noise level.*

With this work, we have shown that our model provides a fruitful segmentation tool for motion capture segmentation. To support various types of motion, we further plan on increasing our motion image database and do more experiments on noisy data.

Acknowledgements

This work is funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 642841; and by the German Federal Ministry of Education and Research (BMBF) through the project Hybr-iT under the grant 01IS16026A.

References

[BKK18] BAI S., KOLTER J. Z., KOLTUN V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018). 1

[HS97] HOCHREITER S., SCHMIDHUBER J.: Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780. 2

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 2

[L*17] LEA C., ET AL.: Temporal convolutional networks for action segmentation and detection. *2017 IEEE CVPR* (2017), 1003–1012. 2

[LB*17] LARABA S., BRAHIMI M., ET AL.: 3d skeleton-based action recognition by representing motion capture sequences as 2d-rgb images. *Computer Animation and Virtual Worlds 28*, 3-4 (2017). 1

[LSD15] LONG J., SHELHAMER E., DARRELL T.: Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE CVPR* (2015), pp. 3431–3440. 1

[Rem12] REMINGTON L. A.: Chapter 13 - visual pathway. In *Clinical Anatomy and Physiology of the Visual System*, Remington L. A., (Ed.), 3rd ed. Butterworth-Heinemann, 2012, pp. 233 – 252. 1

[VDO*16] VAN DEN OORD A., ET AL.: Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016). 2

[W*90] WAIBEL A., ET AL.: Phoneme recognition using time-delay neural networks. In *Readings in speech recognition*. 1990, pp. 393–404. 2