ACM SIGGRAPH / Eurographics Symposium on Computer Animation (2018)

DILATED TEMPORAL FULLY-CONVOLUTIONAL NEURAL 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[1,2] & Philipp Slusallek[1,2]
[1]DFKI Saarbrucken, [2]Saarland University, [3]Max-Planck Institute for Informatics; Germany

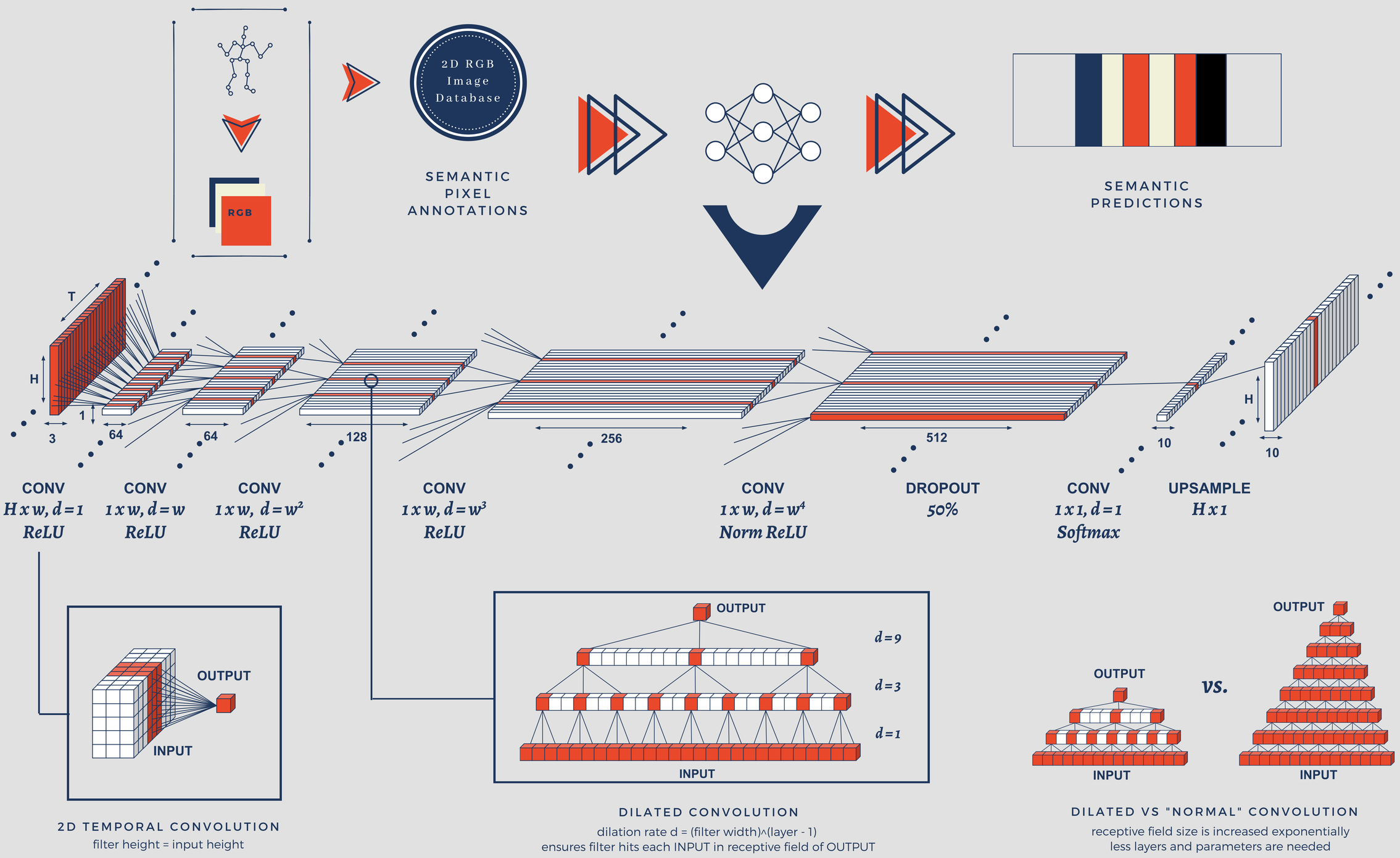
First two authors contributed equally.

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.

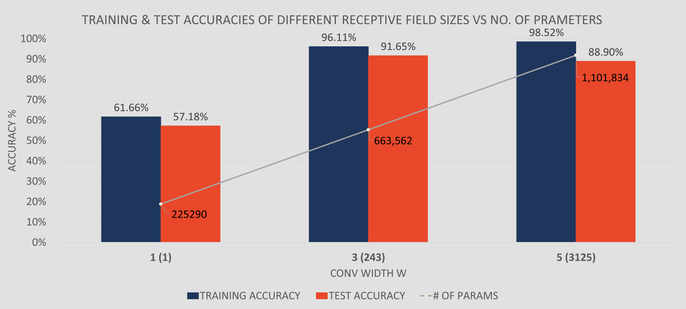
APPROACH

We first transform motion capture data into a "motion image" (columns = frames, rows = joints, RGB = scaled XYZ Euclidean coordinates), which allows us to apply common convolutional neural networks on motion data, to partition our image into smaller segments. To be able to extract temporal information, we make use of convolutions that are only applied in the time domain. Additionally, we take advantage of dilated convolutions to enlarge the receptive field of our model exponentially using comparably few layers and parameters.

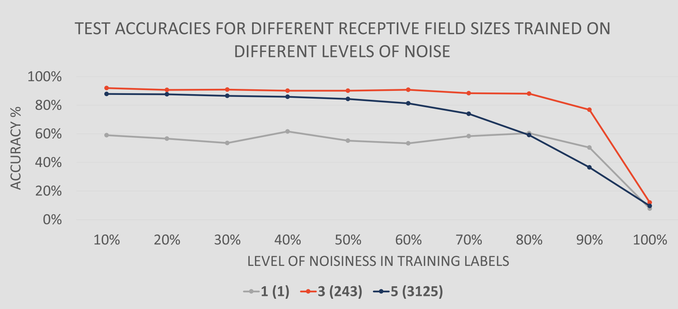


EXPERIMENTAL RESULTS

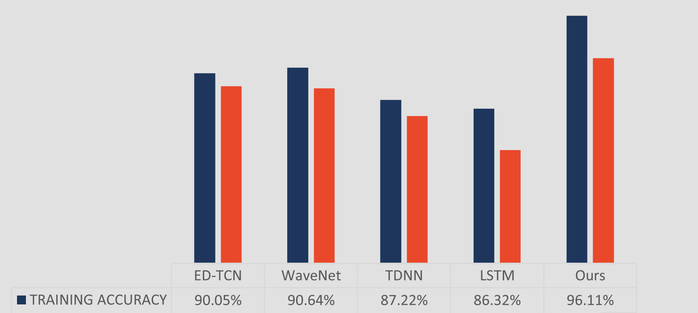
Our motion capture dataset consists of 70 sequences with 10 motion labels: standing, left/right step, begin/end left step, begin/end right step, reach, retrieve and turn. Our sequences reach up to 1500 frames. In all of our experiments, we use 7-fold cross-validation. We use the Adam optimizer with 100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



We test our model on different convolution kernel widths w and find that w = 3 works best, despite having a receptive field size of just 342 frames, which is less than the total sequence lengths.



For robustness against human-error, we train our models on noisy labels. We again find that model w = 3 works best, as it constantly reaches an accuracy of over 88% up until noise levels of 80%.



We further compare our model (w = 3) against another state-of-the-art TCN model (ED-TCN) for action segmentation and three commonly used methods (WaveNet, LSTM, TDNN) for sequence modeling.

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.