Skeleton-Based Human Activity Understanding

Abstract

Human activity understanding in videos is a fast developing research area due to its wide applications in intelligent surveillance, human-computer interaction, robotics, and so on. In recent years, human activity analysis based on 3D skeleton data has attracted a lot of attention due to its succinctness, robustness, and view-invariant representation.

Recent works attempted to utilize recurrent neural networks to model the temporal dependencies between the 3D positional configurations of human body joints for better analysis of human activities in the skeletal data. As the first work of this thesis, we extend this idea to spatial domain as well as temporal domain to better analyze the hidden sources of action-related information within the human skeleton sequences in both of these domains simultaneously. Based on the pictorial structure of Kinect's skeletal data, an effective tree-structure based traversal framework is also proposed. In order to deal with the noise in the skeletal data, a new gating mechanism within LSTM module is introduced, with which the network can learn the reliability of the sequential data and accordingly adjust the effect of the input data on the updating procedure of the long-term context representation stored in the unit's memory cell. The experimental results demonstrate the effectiveness of the proposed method.

As not all skeletal joints are informative for action recognition, and the irrelevant joints often bring noise which can degrade the performance, we need to pay more attention to the informative ones. However, the original LSTM network does not have explicit attention ability. In the second work, we propose a new class of LSTM network, global context-aware attention LSTM, for skeleton-based action recognition, which is capable of selectively focusing on the informative joints in each frame by using a global context memory cell. To further improve the attention capability, we also introduce a recurrent attention mechanism, with which the attention performance of our network can be enhanced progressively. Besides, a two-stream framework, which leverages coarse-grained attention and fine-grained attention, is also introduced. The proposed method achieves state-of-the-art performance on five challenging datasets for skeleton-based action recognition.

In action prediction (early action recognition), the goal is to predict the class label of an ongoing action using its observed part so far. In the third work, we focus on online action prediction in streaming 3D skeleton sequences. A dilated convolutional network is introduced to model the motion dynamics in temporal dimension via a sliding window over the time axis. As there are significant temporal scale variations of the observed part of the ongoing action at different progress levels, we propose a novel window scale selection scheme to make our network focus on the performed part of the ongoing action and try to suppress the noise from the previous actions at each time step. Furthermore, an activation sharing scheme is proposed to deal with the overlapping computations among the adjacent steps, which allows our model to run more efficiently. The extensive experiments on two challenging datasets show the effectiveness of the proposed action prediction framework

Publication List

[1] **Jun Liu**, Amir Shahroudy, Dong Xu, Alex C. Kot, Gang Wang, Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates, IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), 2018.

[2] **Jun Liu**, Gang Wang, Ling-Yu Duan, Ping Hu, Alex C. Kot, Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM Networks, IEEE Transactions on Image Processing (**TIP**), 2018.

[3] **Jun Liu**, Amir Shahroudy, Gang Wang, Ling-Yu Duan, Alex C. Kot, SSNet: Scale selection network for online 3D action prediction, IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2018.

[4] **Jun Liu**, Gang Wang, Ping Hu, Ling-Yu Duan, Alex C. Kot, Global context-aware attention LSTM networks for 3D action recognition, IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2017.

[5] **Jun Liu**, Amir Shahroudy, Dong Xu, Gang Wang, Spatio-temporal LSTM with trust gates for 3D human action recognition, European Conference on Computer Vision (**ECCV**), 2016.

[6] Ping Hu, Bing Shuai, **Jun Liu**, Gang Wang, Deep level sets for salient object detection, IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2017.

[7] Amir Shahroudy, **Jun Liu**, Tian-Tsong Ng, Gang Wang, NTU RGB+D: A large scale dataset for 3D human activity analysis, IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2016.

[8] **Jun Liu**, Henghui Ding, Amir Shahroudy, Lingyu Duan, Xudong Jiang, Gang Wang, Alex C. Kot, Feature Boosting Network for 3D Pose Estimation, Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), under major revision.

[9] **Jun Liu**, Amir Shahroudy, Gang Wang, Ling-Yu Duan, Alex C. Kot, Skeleton-Based Online Action Prediction Using Scale Selection Network, Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), under major revision.

[10] **Jun Liu**, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot, NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding, Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), under review.