A Module Selection-based Approach for Efficient Skeleton Human Action Recognition

Shurong Chai[1], Rahul Kumar Jain[1], Jiaqing Liu[1], Tomoko Tateyama[2], Yinhao Li[1], and Yen-Wei Chen[1]

[1]Graduate School of Engineering, Ritsumeikan University, Japan
[2]Department of Intelligent Information Engineering, Fujita Health University, Japan
E-mail: is0538kr@ed.ritsumei.ac.jp

Human action recognition is one of the major issues in the field of human-computer interaction. It is useful in a wide range of applications including video surveillance, virtual reality and computer-aided surgery. In recent years, human pose estimation techniques have achieved a greater success in estimating the human skeleton joints. Skeleton-based methods have proven to be more robust, reliable and accurate against diverse backgrounds in comparison to RGB images. Previously proposed approaches have used frameworks based on Recurrent Neural Networks (RNNs) or Transformer-based network considering the skeleton data as a sequential data. Some studies employ Convolutional Neural Networks (CNNs) transforming the skeleton data into pseudo images. Recently, Graph Fourier Transforms was proposed extending the Fourier analysis to the graph context in spectral domain. Subsequently, Graph Convolutional Networks (GCNs) were introduced that can generalize the graph convolution from the spectral domain to spatial domain. Spatial GCN-based framework significantly reduces the computational cost and various uses of spatial GCNs are investigated recently. GCN-based networks effectively capture the long-range dependencies while maintaining the topological human body structure in comparison to the conventional RNN and CNN-based networks.

Due to their superior performance, existing methods heavily use Graph Convolutional Networks (GCN) and Temporal Convolutional Networks (TCN) to extract spatial and temporal feature information, respectively. These methods use a stack of blocks containing the spatial and temporal feature extraction modules to extract enhanced high-level feature information for prediction. But these approaches have two major drawbacks. First, the number of extractor modules are fixed for all types of human actions whereas the optimal number of modules should vary depending on the particular action. For some specific actions such as ‘reading’ where the consecutive frames only slightly differ from one-another, we can emphasize spatial information using a larger number of spatial modules. For actions such as ‘clapping’, where temporal frames are quickly altered, we need to emphasize both spatial and temporal features. Second, the order of these extractor modules is designed manually. Thus, the framework may not have an optimal design and takes a high inference time.

On the other hand, a low-spec device needs to be used to implement a human action recognition task for real-world applications. Therefore, low computational cost is required for this task. Inspired by recent dynamic methods, we propose a module selection strategy for efficient skeleton-based human action recognition. Our method lowers the cost of computing while learning to choose the network structure adaptively. We formulate decision-making strategy investigating at local and global level to determine an optimal network structure. We add a decision-making network before each spatial and temporal module to decide whether it needs to be kept or dropped for feature extraction.

In this session, we will discuss the role of data science in the medical field and the two items we have described above.