Factorized 3D-CNN for Skeleton-based Action Recognition

Nadhira Noor and In Kyu Park
Department of Electrical and Computer Engineering, Inha University
Incheon 22212, Korea
{nadhirannoor@gmail.com, pik@inha.ac.kr}

***Abstract*—Most CNN-based action recognition heavily relies on appearance information by taking an RGB sequence of entire image regions as an input. While RGB-based methods excel in capturing contextual information, they are unable to explicitly comprehend human motion, which limits their robustness. Prior skeleton-based works employed GCN, however it has limitation in computational heaviness. To address this problem, we propose skeleton-based action recognition with factorized 3D-CNN. As proposed in [1], we represent human skeleton pose as pseudo joint heatmap. This leads to a lighter neural network as we require no large spatial size. Furthermore, we factorize 3D-CNN to 2D spatial convolution and 1D temporal convolution. As found in [2] decomposition of 3D-CNN, facilitates the optimization.**

I. Introduction

In action recognition, CNN-based approaches have traditionally leaned on RGB sequences, effectively capturing contextual information. Duan *et al*. [1], successfully incorporates skeleton-based into 3D-CNN. However, it is still computationally heavy. In response, we propose a factorized 3D-CNN for skeleton-based action recognition. The factorization of 3D-CNN, as demonstrated in [2], enhances the optimization process. We opt to factorizing 3D-CNN not only for its ease of optimization but also aligns with future deployment on edge devices, enhancing computational efficiency and enabling real-time processing. This choice offers benefits such as reduced parameters, adaptability, and scalability.

II. Proposed Method

As proposed in [1], we represent our input into stacked pseudo joint heatmaps. First, we employ top-down pose estimator to extract 2D human pose. Then we generate the pseudo joint heatmap by generating gaussian maps that are centered at every joint location, example of generated joint heatmap shown in Figure 1(a). Lastly, we sample 32 frames uniformly and discard the remaining frames. Therefore, our input shape is $C \times T \times H \times W$ where C is channel, T is number of frames, H, W is the height and width of the frame.

We utilize (2+1)D [2] which decompose 3D-CNN into two separate convolutions, a 2D spatial convolution and 1D temporal convolution. Additionally, we incorporate non-linear rectification between these two operations. However unlike [2], we represent the human motion with pseudo joint heatmap that does not require large spatial size, which allowing us to reduce the channel filter size and make the network computationally much lighter while maintaining high accuracy.

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University) and No.2022-0-00981, Foreground and Background Matching 3D Object Streaming Technology Development and No.2021-0-02068, Artificial Intelligence Innovation Hub).

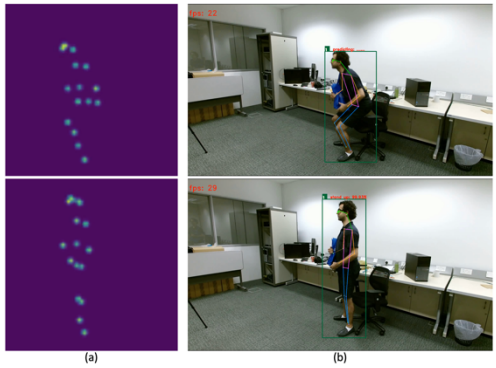

Figure 1. (a) Generated pseudo joint heatmap (b) Our method results.

III. Discussions

In this section, we discuss the performance of our method based on the results of our experiments. We present quantitative comparison results between our proposed method and prior works [1, 2]. In Figure 1, we display our qualitative results of our method tested on NTU RGB+D 60 dataset [3]. These results demonstrate our method performs well, able to classify actions correctly and lastly, our method achieves 25 FPS on real-time action recognition.

IV. Conclusion

We demonstrate the robustness and efficiency of factorized 3D-CNN for skeleton-based approach for action recognition. We show that representing human motion through stacked joint heatmaps allows for the efficient design of a convolutional neural network that does not require a large spatial size while preserving essential features. Additionally, by employing a factorized 3D-CNN, we can retain both spatial and temporal features, leading to higher accuracy and faster processing times.

References

[1] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai. Revisiting skeleton-based action recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2959–2968, 2022.

[2] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[3] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016.