Supplementary of Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition

Lei Shi[1,2] Yifan Zhang[1,2*] Jian Cheng[1,2,3] Hanqing Lu[1,2]
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
{lei.shi, yfzhang, jcheng, luhq}@nlpr.ia.ac.cn

1. The learning rate scheduler and the data preprocessing methods.

The original performance of ST-GCN on the NTU-RGBD dataset is 88.3, in which the learning rate is multiplied by 0.1 at the 10_{th} and 50_{th} epochs. The training process is ended in 80_{th} epoch. We rearrange the learning rate scheduler from [10, 50, 80] to [30, 40, 50] and obtain the better performance (marked as "before preprocessing" in Tab. 1).

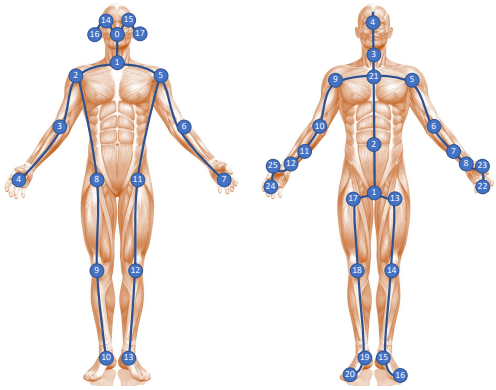

Figure 1. The left sketch shows the joint label of the Kinetics-Skeleton dataset and the right sketch shows the joint label of the NTU-RGBD dataset.

Moreover, we use some preprocessing strategies on the NTU-RGBD dataset. The body tracker of Kinect is prone to detecting more than 2 bodies, some of which are objects. To filter the incorrect bodies, we first select two bodies in each sample according to the energy of each body, which is defined as the summation of the skeleton's standard deviation across each channel.

Subsequently, each sample is normalized to make the distribution of the data for each channel unified. In detail, the coordinates of each joint are subtracted from the coordinates of the "spine joint" (the 2_{nd} joint in the left sketch in Fig. 1).

Finally, as different samples may be captured in different viewpoints, similar to [1], we translate the original 3D location of the body joints from the camera coordinate system to body coordinates. For each sample, we perform a 3D rotation to fix the X axis parallel to the 3D vector from the "right shoulder" (5_{th} joint) to the "left shoulder" (9_{th} joint), and the Y axis toward the 3D vector from the "spine base (21_{st} joint) to the spine (2_{nd} joint). Fig. 2 shows an example of the preprocessing.

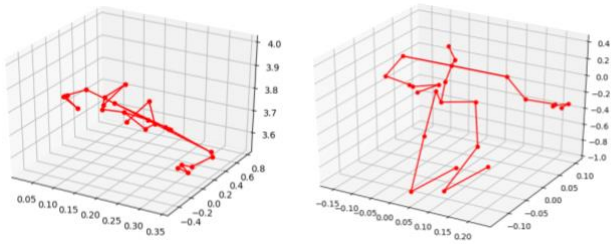

Figure 2. Example of the data preprocessing on the NTU-RGBD dataset. The left is the original skeleton, and the right is the preprocessed skeleton.

Tab. 1 compares the performances that before and after the preprocessing. It shows that the preprocessing considerably helps the recognition, which may because that the original data are noisy.

References

[1] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A Large Scale Dataset for 3d Human Activity Analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

*Corresponding Author

Methods	Accuracy (%)
original performance in [2]	88.3
before preprocessing	90.1
after preprocessing	92.7

Table 1. Comparisons of the validation accuracy using rearranged learning-rate scheduler and data preprocessing.

[2] S. Yan, Y. Xiong, and D. Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI*, 2018. 2