A Spatial-Temporal Graph Convolutional Networks-based Approach for the OpenPack Challenge 2022

3rd Place

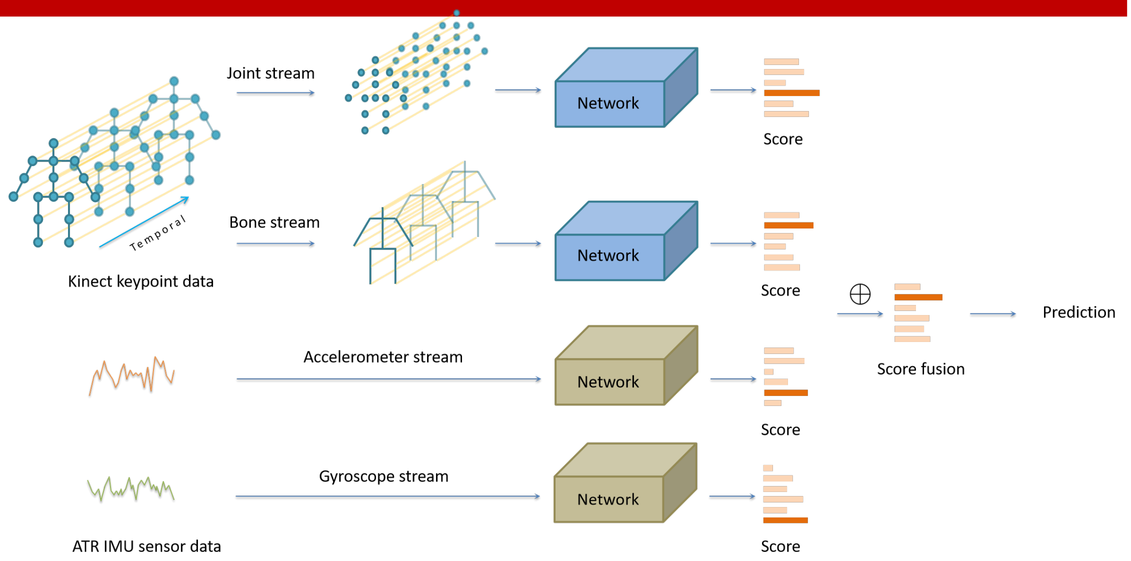
Shurong Chai, **Jiaqing Liu**, Rahul Jain, Yinhao Li, Tomoko Tateyama, Yen-Wei Chen

Team: Ritsumei

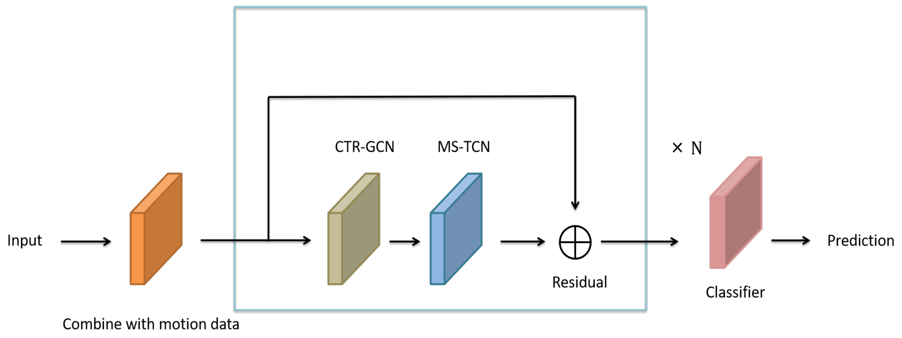
Method Overview

Novelty:

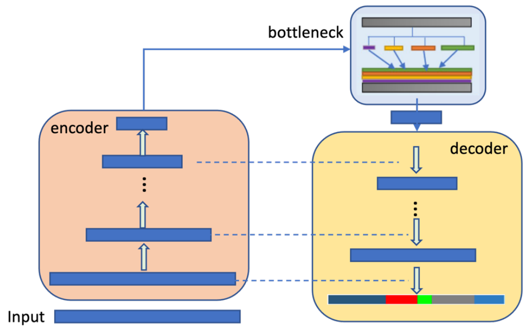
- Multimodal Feature extraction => joint, bone, accelerometer, gyroscope
- Motion data changes dramatically from one action to another=> Combine normal and motion data
- Capture the long-range dependencies among temporal dimensions =>propose a multi-scale temporal convolution network employing large-size kernels
- Over segmentation problem => smoothing loss



Keypoint data stream



Sensor data stream



Accelerometer, Gyroscope stream

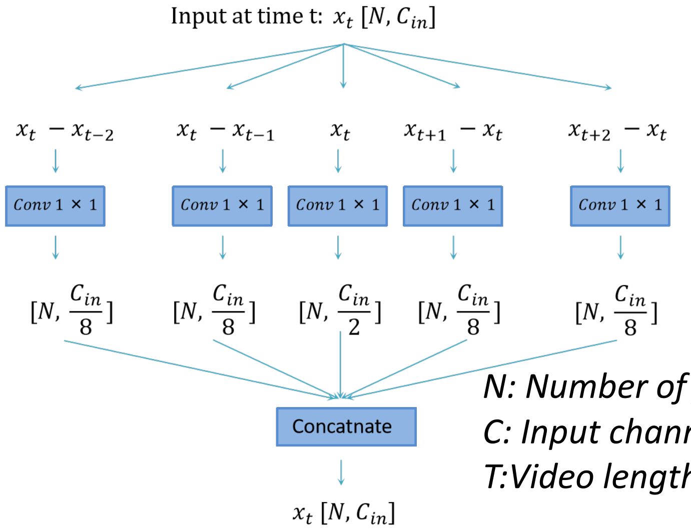
Capture information at different resolutions

Add a classifier to predict the action boundary[1]

Motion-aware input

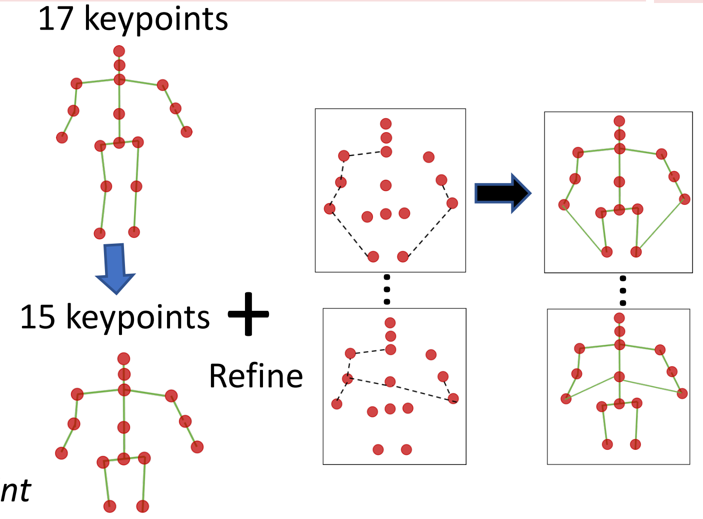
Small-scale temporal difference

Larger-scale temporal difference



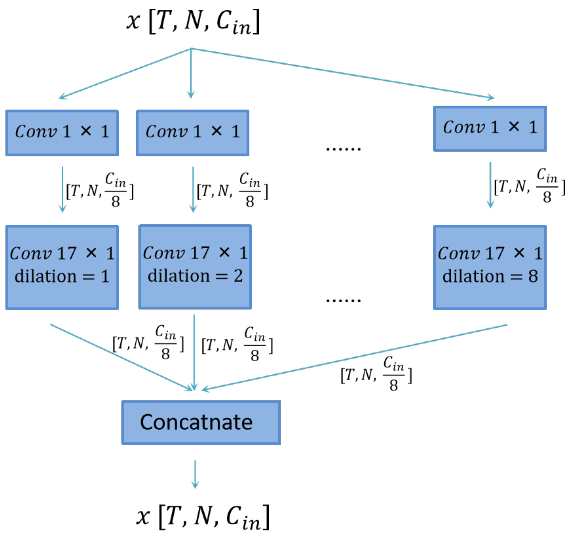
N: Number of joint
C: Input channels
T:Video length

Spatial feature extraction



Channel-wise topology refinement graph convolution[2]

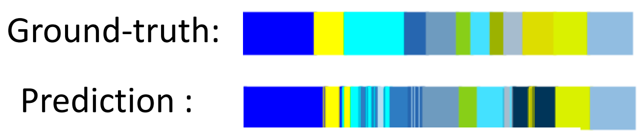
Temporal feature extraction


$x\ [T, N, C_{in}]$
$Conv\ 1 \times 1$
$[T, N, \frac{C_{in}}{8}]$
$Conv\ 17 \times 1$ dilation = 1
$Conv\ 17 \times 1$ dilation = 2
$Conv\ 17 \times 1$ dilation = 8
......
Concatnate
$x\ [T, N, C_{in}]$

Multi-scale temporal convolution

Loss functions

Over segmentation problem [3]



$$L_{keypoint} = L_{\text{CrossEntropy}} + L_{\text{TMSE}}$$

$$L_{\text{TMSE}} = \frac{1}{TC}\sum_{t,c}\tilde{\Delta}_{t,c}^2 \qquad \tilde{\Delta}_{t,c} = \begin{cases}\Delta_{t,c}: & \Delta_{t,c} \leq \tau \\ \tau: & otherwise\end{cases}$$

$$\Delta_{t,c} = |\, log\ y_{t,c} - log\ y_{t-1,c}\, |$$

T: Video length
C: Number of classes
$y_{t,c}$: Probability of class c at time t
τ = 16

Results

F1 score = 0.924

References

[1] Singhania, D., Rahaman, R., & Yao, A. (2021). Coarse to fine multi-resolution temporal convolutional network. arXiv preprint arXiv:2105.10859.

[2] Chen, Yuxin, et al. "Channel-wise topology refinement graph convolution for skeleton-based action recognition." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[3]Farha, Y. A., & Gall, J. (2019). Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3575-3584).