# Vertex Feature Encoding and Hierarchical Temporal Modeling in a Spatio-Temporal Graph Convolutional Network for Action Recognition

ICPR 2020
Milan, 10-15 January 2021

uni.lu | SnT
https://cvi2.uni.lu/

Konstantinos Papadopoulos, Enjie Ghorbel, Djamila Aouada, Björn Ottersten
SnT, University of Luxembourg
Email: {firstname.lastname}@uni.lu

## Introduction

In this paper, we introduce two novel modules for Spatio-temporal Graph Convolutional Networks (ST-GCN) [1], namely, the Graph Vertex Feature Encoder (GVFE) and the Dilated Hierarchical Temporal Convolutional Network (DH-TCN). GVFE learns appropriate vertex features for action recognition by encoding raw skeleton data into a new feature space, while DH-TCN is capable of capturing both short-term and long-term temporal dependencies using a hierarchical dilated convolutional network. The use of GVFE and DH-TCN results in a smaller number of layers and parameters; thus the required training time and memory are reduced.

## Motivation

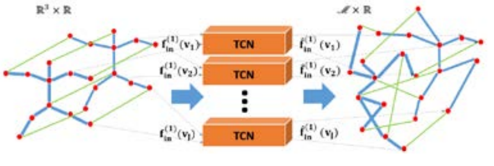
Spatio-temporal Graph Convolutional Networks (ST-GCNs) [1] have shown great performance. However:

- Vertex features containing raw skeleton data might be not discriminative enough, since they are not learned in an end-to-end-manner.
- Temporal dependencies are modeled by a single temporal convolutional layer and, consequently, critical long-term dependencies might not be consistently described.
- They make use of a considerable number of ST-GCN blocks (10 in most cases).

## Proposed Approach

### A: GVFE

- GVFE is directly placed before the first ST-GCN block.
- It is trained in an end-to-end manner with the entire network.
- It maps 3D skeleton coordinates from the Cartesian coordinate system $\mathbb{R}^3$ to a learned feature space $\mathcal{M} \subseteq \mathbb{R}^{C_{out}}$ of higher dimensionality.
- This module preserves the skeleton structure.



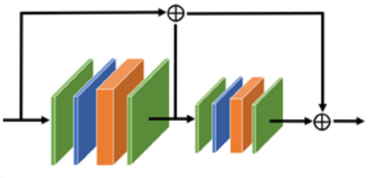

- $\hat{f}_{in}^{(1)}(v_i) = W_i^{TCN} * f_{in}^{(1)}(v_i)$, where $\{W_i^{TCN}\}$ is the collection of tensors containing the Temporal Convolutional Network (TCN) kernel filters.

Applicable to any graph-based network, better generalization, more sufficient feature space for action recognition.

Luxembourg Nationa
Research Fund

### B: DH-TCN

- DH-TCN is composed of $N$ successive dilated temporal convolutions and it replaces the temporal convolutions in the last ST-GCN block.
- Each layer output $f_{temp}^{(k,n)}$ of order $n$ of DH-TCN is obtained as: $f_{temp}^{(k,n)} = F\left(W_i^{DH} *_i f_{temp}^{(k,n-1)}\right)$, with $f_{temp}^{(k,0)} = f_{out}^{(k)}$, $f_{out}^{(k)}$ the output feature map from the Spatial GCN block and $\{W^{DH}\}$ the trainable temporal filters.



Green: **BatchNorm**
Blue: **ReLU**
Orange: **2D Conv**

Encodes both short-term and long-term dependencies.
Both GVFE and DH-TCN require fewer ST-GCN blocks.

## Experimental Results

| Method | NTU-60 | NTU-120 | Kinetics |
|---|---|---|---|
| | Xsub / Xview | Xsub / Xview | Top1 / Top5 |
| Skelemotion | 76.5 / 84.7 | 67.7 / 66.9 | - |
| Pose Ev. Map | **91.7 / 95.3** | 64.6 / 66.9 | - |
| ST-GCN (10b) [1] | 81.5 / 88.3 | 72.4 / 71.3 | 30.7 / 52.8 |
| **Ours (ST-GCN) (4b)** | 79.6 / 88.0 | 72.3 / 71.7 | 29.0 / 50.9 |
| AS-GCN (10b) [2] | 86.8 / 94.2 | 77.7 / 78.9 | **34.8 / 56.5** |
| **Ours (AS-GCN) (4b)** | 86.4 / 92.9 | **79.2 / 81.2** | - |

## Conclusion

In this paper, two novel modules for ST-GCN based methods have been proposed called GVFE and DH-TCN. These modules enable the reduction of the number of needed blocks and parameters while conserving almost the same or improving the recognition accuracy.

## References

[1]: Yan et al. "Spatial temporal graph convolutional net-works for skeleton-based action recognition", AAAI 2018.
[2]: Li, et al. "Actional-Structural Graph Convolutional Net-works for Skeleton-based Action Recognition", CVPR 2019.

## Acknowledgements

This work was funded by the National Research Fund (FNR), Luxembourg, under the project C15/IS/10415355/ 3D-ACT/Björn Ottersten. We would, also, like to thank Christian Hundt from NVIDIA AI Technology Center Luxembourg for his valuable input and fruitful discussions.