Sequential Image Classification of Human-Robot Walking Environments using Temporal Neural Networks

Bohdan Ivaniuk-Skulskyi[1,4], Andrew Garrett Kurbis[2,4], Alex Mihailidis[3,4], and Brokoslaw Laschowski[3,4]

[1]National University of Kyiv-Mohyla Academy, Ukraine; [2]University of Waterloo, Canada; [3]University of Toronto, Canada; [4]KITE-Toronto Rehabilitation Institute, Canada

Introduction

Robotic prosthetic legs and exoskeletons require real-time and accurate state estimation of the walking environment for smooth transitions between different locomotion mode controllers. However, previous studies have mainly been limited to static image classification, therein ignoring the temporal dynamics of human-robot walking.

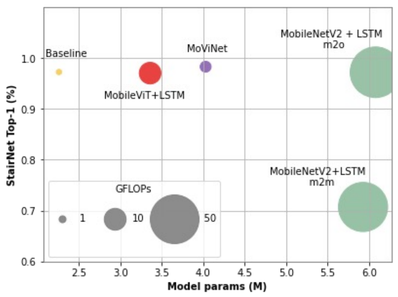


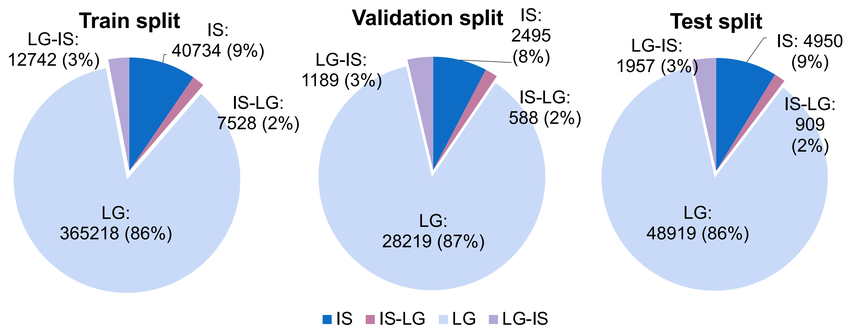
Figure 1. Comparison of model size, number of operations, and prediction accuracy. The 3D-CNN MoViNet model achieved better accuracy with slightly more FLOPs and parameters than the state-of-the-art 2D-CNN baseline.

Objectives

Here we developed different temporal neural networks to compare the performances of static vs. sequential image classification of real-world walking environments in terms of prediction accuracy, and computational and memory storage requirements [1].

Methods

Building of *ExoNet* and *StairNet*, the large-scale datasets of first-person videos of real-world walking environments, we organized the data into sequences of images for four classes, including level-ground terrain, incline stairs, and transitions to and from incline stairs.



Sequential data requires specialized deep learning models that account for temporal dynamics. We studied different encoders, including VGG, EfficientNet, MobileNetV2, ViT, and MobileViT, each coupled with a temporal long short-term memory (LSTM) backbone. We also built and studied MoViNet - a new video classification model designed for mobile and embedded devices with limited computational resources [2].

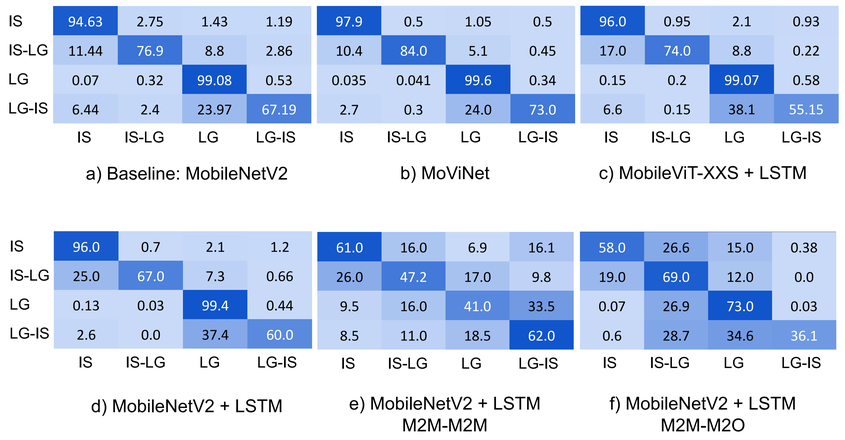


Table 1. Confusion matrices on StairNet using different temporal neural networks.

Results

The 3D-CNN MoViNet network outperformed the 2D-CNN encoders with LSTM backbones and the 2D-CNN baseline model in terms of prediction accuracy, suggesting that network architecture plays an important role in performance besides simply the incorporation of temporal data.

Model	Parameters	GFLOPs	Accuracy	F1	NetScore	FPS
MoViNet [2]	4.03	2.5	**0.983**	**0.982**	69.67	13.45
MobileViT - LSTM	3.36	9.84	0.970	0.968	64.28	17.55
MobileNet - LSTM	6.08	53.96	0.973	0.970	54.35	17.65
MobileNet - LSTM (seq2seq) - M2O	5.93	50.97	0.707	0.799	49.18	17.55
MobileNet - LSTM (seq2seq) - M2M	5.93	50.97	0.432	0.538	40.62	17.55
Baseline [1]	2.26	0.61	0.972	0.972	**78.12**	22.58

Discussion

In this study, we compared static vs. sequential image classification of walking environments. Although 3D-CNN networks achieve higher prediction accuracy compared to 2D-CNN encoders with LSTM backbones, they also have higher computational and memory storage requirements, which can be disadvantageous for robotic prosthetic legs and exoskeletons using edge computing devices

References

1. Kurbis, et al. (2022). "Stair Recognition for Robotic Exoskeleton Control using Computer Vision and Deep Learning". bioRxiv.
2. Kondratyuk, et al. (2021). "MoViNets: Mobile Video Networks for Efficient Video Recognition". arXiv.

Contact: brokoslaw.laschowski@uhn.ca

@KITE_UHN