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

Bogdan Ivanyuk-Skulskyi, Andrew Garrett Kurbis, Alex Mihailidis, and Brokoslaw Laschowski

***Abstract* – Here we studied the effects of temporal neural networks on sequential image classification of human-robot walking environments, with an emphasis on stair recognition. We developed a dataset featuring sequences of images of level-ground terrain, incline stairs, and transitions to and from incline stairs. We then studied the performances of different 2D deep learning encoders, each coupled with an LSTM backbone, as well as a state-of-the-art 3D video classification model. Although the 3D neural network achieved higher prediction accuracy compared to the 2D neural networks with LSTM backbones, it also has higher computational and memory storage requirements, which can be disadvantageous for human locomotion with robotic legs using edge computing.**

I. 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. 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].

II. Methods

Building on 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 deep learning 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].

III. Results and Discussion

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

References

[1] A. G. Kurbis, B. Laschowski, and A. Mihailidis, "Stair recognition for robotic exoskeleton control using computer vision and deep learning," in 2022 IEEE International Conference on Rehabilitation Robotics (ICORR), 2022, pp. 1–6

[2] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "MoViNets: Mobile video networks for efficient video recognition," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[3] S. Mehta and M.Rastegari, "MobileVit: light-weight, general-purpose, and mobile-friendly vision transformer," arXiv, 2021.

Table 1. Comparison of temporal neural networks used for sequential image classification of human-robot walking environments. Best values for each category as shown in bold.

Deep Learning Model	Parameters	GFLOPS	Accuracy	F1-score	NetScore	Frames per second
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**

*Research supported by AGE-WELL Networks of Centres of Excellence (NCE) Program, Canada.

B. Ivanyuk-Skulskyi is with the Department of Mathematics, National University of Kyiv-Mohyla Academy, Kyiv, Ukraine (e-mail: ivanyuk.skulskiy@ukma.edu.ua).

A. G. Kurbis is with the Department of Systems Design Engineering, University of Waterloo, ON, Canada, on placement in the Temerty Faculty of Medicine, University of Toronto, ON, Canada (e-mail: agzkurbis@uwaterloo.ca).

A. Mihailidis is with the Temerty Faculty of Medicine, University of Toronto, ON, Canada, and the Toronto Rehabilitation Institute, ON, Canada (e-mail: alex.mihailidis@utoronto.ca).

B. Laschowski is with the Department of Mechanical Engineering, University of Toronto, ON, Canada, and the Toronto Rehabilitation Institute, ON, Canada (e-mail: brokoslaw.laschowski@utoronto.ca).