Deep learning semantic segmentation for indoor terrain extraction: Toward better informing free-living wearable gait assessment

Jason Moore
Dept. Computer and Information Sciences
Northumbria University
Newcastle Upon Tyne, UK
jason.moore@northumbria.ac.uk

Fraser Young
Dept. Computer and Information Sciences
Northumbria University
Newcastle Upon Tyne, UK
fraser.young@northumbria.ac.uk

Peter McMeekin
Dept. Nursing, Midwifery and Health
Northumbria University
Newcastle Upon Tyne, UK
peter.mcmeekin@northumbria.ac.uk

Richard Walker
Northumbria Healthcare NHS Foundation Trust
Newcastle Upon Tyne, UK
richard.walker@northumbria-healthcare.nhs.uk

Sam Stuart
Dept. Sport, Exercise and Rehabilitation
Northumbria University
Newcastle Upon Tyne, UK
sam.stuart@northumbria,ac.uk

Alan Godfrey
Dept. Computer and Information Sciences
Northumbria University
Newcastle Upon Tyne, UK
alan.godfrey@northumbria.ac.uk

Contemporary approaches to gait assessment use wearable devices within free-living environments to capture habitual information, which is more informative compared to data capture in the lab. Wearables range from inertial to camera-based technologies but pragmatic challenges such as analysis of big data from heterogenous environments exist. For example, wearable camera data often requires manual time-consuming subjective contextualization, such as labelling of terrain type. There is a need for the application of automated approaches such as those suggested by artificial intelligence (AI) based methods. This pilot study investigates multiple segmentation models and proposes use of the *PSPNet* deep learning network to automate a binary indoor floor segmentation mask for use with wearable camera-based data (i.e., video frames). To inform the development of the AI method, a unique approach of mining heterogenous data from a video sharing platform (YouTube) was adopted to provide independent training data. The dataset contains 1973 image frames and accompanying segmentation masks. When trained on the dataset the proposed model achieved an Instance over Union score of 0.73 over 25 epochs in complex environments. The proposed method will inform future work within the field of habitual free-living gait assessment to provide automated contextual information when used in conjunction with wearable inertial derived gait characteristics.

***Clinical Relevance*—Processes developed here will aid automated video-based free-living gait assessment.**

Keywords— Deep Learning, Gait Analysis, IMU, Terrain Classification, Floor Segmentation

I. INTRODUCTION

Within the field of neurological disorders such as Parkinson's disease (PD), gait assessment has emerged as a pragmatic (clinical) tool for assessing disease onset and progression [1]. Typically, gait assessment is performed within a purpose-built environment specifically for the examination of people
with PD or other movement disorders [2]. However, controlled facilities are not able to capture data from the person that is representative of their natural/habitual ability/performance [3]. Furthermore, free-living/habitual gait data capture is key to overcome the impact of observer phenomenon/affect [4] and could have the advantage of being able to increase the speed of diagnosis and improve strategies to e.g., reduce falls [5].

Typically, traditional approaches to gait assessment rely on direct observation (from an experienced clinician) or use of an instrumented walkway [6, 7]. However, contemporary approaches now focus on inertial measurement units (IMU's), enabling use within the clinic and beyond (i.e., free-living) [8]. IMU's combine accelerometers and gyroscopes to collect accurate and high-resolution data pertaining to the specific dynamic movements of the wearer. The measurement of gait can then be used to classify disturbances such as freezing [9] or to inform fall prediction based on abnormal patterns [10]. Although IMU's provide objective and accurate data, they remain limited to fully inform free-living gait as they are "blind": although IMU's provide robust data, the context of the walking environment is unknown.

Contextual insight could help better inform IMU-based gait characteristic interpretation. For example, to determine if high/increased step time variability is due to intrinsic (pathology) issue(s) or extrinsic environmental condition(s). Although an attempt to explore contextualization directly from raw (sample level) IMU data has been conducted [11], the approach is limited by the amount/extent of contextual information provided. Without a wearable camera, intrinsic or extrinsic factors cannot be absolutely determined [12].

Contextual data from wearable cameras would provide absolute context, enabling a better understanding of gait variations. However, video-based analysis is challenging, typically involving a researcher manually examining many hours of video and

labelling/classifying each scene [12]. That process could be streamlined using artificial intelligence (AI) to automatically classify video data. The AI approach could also alleviate issues relating to confidentiality as an automated classification could negate (or at worst greatly minimize) privacy concerns pertaining to human viewing (i.e., researcher witnessing sensitive video data from someone's home).

The fundamental requirement to better understand free-living gait centre's on information to the ground. Accordingly, developing an AI approach is driven by one question: what terrain is someone is walking on? Consequently, this pilot study proposes a new methodology/model using available and trainable deep learning algorithms, which given their ability to train on more representative data including edge cases for a robust and diverse classification methodology.

II. METHODS

The proposed model aims to identify and produce a binary segmentation mask of any indoor terrain type within a video frame. The model utilizes the TensorFlow deep learning library complemented with OpenCV for image processing tasks all within a Python 3.8 environment.

A. Dataset creation

Given the novel application of the algorithm (i.e., extract and classify terrains from within indoor representative living environments), pre-existing public datasets are not available for this type of classification task. Accordingly, a new dataset was created for both the segmentation of just the floor within indoor environments along with the type of terrains themselves (wood, carpet, tile). For segmentation, the input images are needed for training but also the accompanying segmentation masks or a file (typically JSON) including the polygonal coordinate points for the mask.

Gathering the required training video-based frames was completed using the unique approach of a video sharing platform (www.youtube.com), where videos are available under the CC-BY license [13] and useable within research projects if attribution is provided. The platform was mined for real estate-based (i) walkthroughs and (ii) tours, recorded from a first-person perspective (i.e., a wearable camera view).

B. Mining

A data mining pipeline was established pertaining to the creation of a custom Python script using the YouTube-dl library. Upon execution a researcher inputs a link to the desired YouTube video which was downloaded in 1280×720px. The script then adds metadata (channel, video title, video link) to a CSV file for attribution requirements. Frames were manually filtered removing any not pertaining to the interior of a property or which did not have a visible terrain.

C. Video data processing

All videos were downloaded at a uniform resolution of 1280×720px at 24 frames/s with the VGG Image Annotator [14] used to create the accompanying binary segmentation masks. Videos were segmented into frames (every 30 frames) with the assumption being major terrain differentiations are unlikely to occur over a course of 30 frames (1.25s) when someone is walking. Equally, 30 frames were used to prevent overfitting to specific scenes given the relatively low number of training images. Each extracted frame was imported into the VGG Image Annotation Tool and manually labelled such that within a given frame a polygon containing the outline of floor was drawn and masked (Fig 1).

D. Choosing system architecture

TensorFlow deep learning library was used within a Python 3.8 Environment to create all models. Different contemporary approaches to image segmentation were explored, with a pragmatic train/test ratio of 80:20. The dataset was run across 4 different segmentation models to determine an optimal approach: (i) *U-Net* [15], (ii) *FPN* [16], (iii) *LinkNet* [17] and (iv) *PSPNet* [18]. All models were trained on an Apple device (M1 Pro, 16GB RAM). The trained network takes in an input image of 1280×720px and downscaled to 256×256px utilizing the OpenCV resize function and dimensionality was also reduced to one color channel (grayscale). The image is then input to the model resulting in an output matching 256×256px binary segmentation mask utilizing the pixel wise *SoftMax* function (Equation 1) of the floor within the frame.

$$p_k(X) = \frac{\exp(a_k(\mathbf{x}))}{\sum_{k'=1}^{K} \exp(a_{k'}(\mathbf{x}))} \quad (1)$$

where $a_k(\mathbf{x})$ is the activation in feature channel k at the pixel position $\mathbf{x}$ and K is the number of classes.

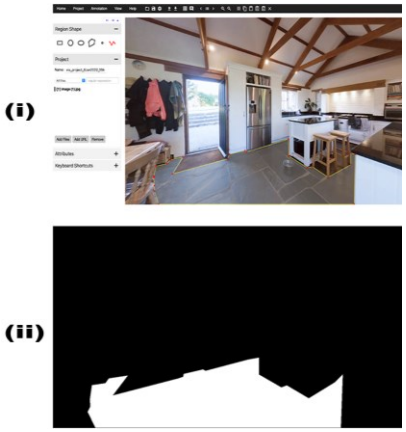

Figure 1: (i) Original input image labeled/annotated within VGG image anotator (ii) Output binary segmentation mask of the floor (white)

The binary segmentation mask is then upscaled to match the original image again using the OpenCV library resize function (Algorithm 1). Typically, upscaling an image results in quality loss but this is not relevant here given the low-resolution requirement on the segmentation mask.

Algorithm 1: Provide playback of segmented floor
Require: Input Image of 1280 × 720px
Ensure: Segmentation of floor
1: for frame in video
2: if frame number is divisible by 30
3: resize frame to 128 × 128px
4: convert frame to grayscale
5: get image mask from segmentation model
6: resize mask to 1280 × 720px
7: apply bitwise and to mask and original frame
return bitwise and resulting image

I. Analysis
A standardized Intersection-over-Union (IoU) benchmark was used to evaluate object detection and tracking model performance, providing a percentage of overlap (i.e., how close are the labelled and predicted segmentation masks) [19].

E. Architecture evaluation
All models were tested using the new dataset of 1973 images over a period of 25 epochs and a batch size of 16. *PSPNet* outperformed all other models with an IoU of 0.73, Table 1. This was also the case with the processing speed where *PSPNet* outperformed other comparative models at 1 FPS.

F. PSPNet architecture
First an input image (256×256×1) is input to the network with feature maps generated using a DeepLab dilated network strategy model [20], giving a feature map 1/8 of the size of the input image, Fig 2ii. Generated feature maps are then fed forward into a pyramid pooling module, pooling the feature map into a vertical chain of average pooling outputs (1×1, 2×2, 3×3 and 6×6), Fig. 2iii. All average pooling outputs are then convolved by 1×1 to further reduce the dimensionality of the output pooling. To up sample feature maps, bilinear interpolation is applied to each low dimensional feature map to match original size. All upscaled feature maps are concatenated and convolved to produce a final prediction.

TABLE 1: PERFORMANCE COMPARISON OF COMPETING SEGMENTATION ARCHITECTURES

Architecture	***IoU***	***FPS***
U-Net	0.71	0.80
FPN	0.72	0.73
LinkNet	0.71	0.77
PSPNet	*0.73*	*1.00*

G. Terrain classification
Here, the segmentation algorithm will produce a binary mask of a video frame showing the extracted region of floor, Figs. 2iv and 2v. The Python OpenCV library is then utilized to extract the contours of the binary mask allowing for the quantification of overall terrain area, Fig. 2vi. An arbitrary threshold of 2000 was set for the mask with any area below classified as the frame not having a terrain visible. Cases ≥2000 are then used to map a 500×500px rectangle crop from the image including only the terrain (Fig. 3).

H. Reference and evaluation
To evaluate the proposed method, it was compared against a rudimental un-segmented approach taking an arbitrary image crop defined by a generic 500×500px region taken from the bottom of all video frames to coincided with the possible floor location.

A random sample of 60 frames was selected from the dataset and supplemented with additional frames from a university environment during a brief walk. All frames were manually labelled.

III. Results

A. Dataset

Currently the dataset consists of 1973 numbers of raw training images and accompanying amount of binary segmentation masks. The dataset contains frames spanning many types of interior terrain types (wood, tile, carpet) and covers a range of lighting and video quality conditions.

B. Terrain classification

An accuracy improvement of 5% is achieved when using the proposed method compared against the rudimental approach (i.e., generic image cropping), Table 2.

TABLE 2: ACCURACY COMPARISONS OF SEGMENTATION VS UNSEGMENTED IMAGES

Method	**Accuracy**
Proposed	76%
Rudimental	71%

IV. Discussion

This paper has shown the viability of deep learning to produce accurate and generalizable floor segmentations and classifications across a range of indoor types. A network is created using readily available tools such as TensorFlow and OpenCV within Python, by chaining layers of down sampling convolutions and average pooling before reversing the process using further convolutions and concatenation. The methodology presented here could be useful to aid automated analysis of free-living environments to help inform habitual gait assessment as it provides terrain classification only by removing unnecessary clutter or noise to the full scene view. The methodology has applicability to wearable cameras which could be used in conjunction with wearable IMU for a more rounded and fully informed free-living gait assessment.

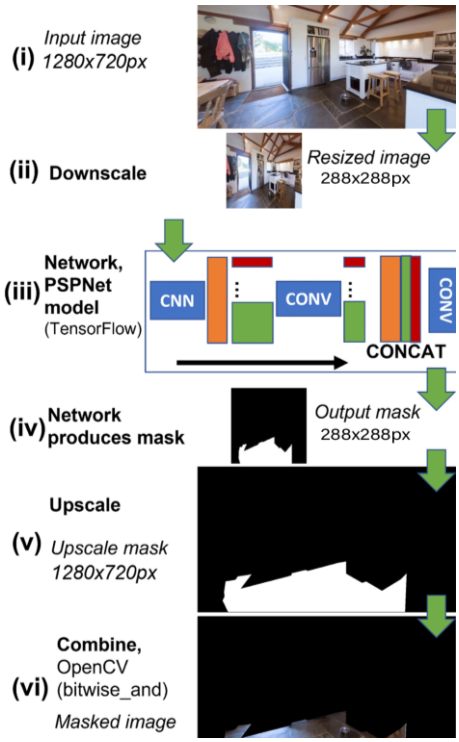

Figure 2: Model flow (i) original input image (ii) downscaled (iii) downscaled image into PSPNet model (iv) PSPNet outputs predicted segmentation mask at 288×288px (v) output segmentation mask upscaled to match original image (vi) OpenCV overlays the mask on the original image utilizing the bitwise and function to subtract all elements except floor.

A. Toward better informed free-living IMU gait

IMU's are valuable tools to quantify habitual/free-living gait in those with a PD. However, current trends in IMU-based gait assessment beyond the lab examine all walking/gait data, inferring intrinsic assumptions with no knowledge of extrinsic factors. IMU gait supplemented with wearable camera data can provide a rounded and better habitual gait assessment. Current attempts to analyze free-living camera data broadly rely on manual, time consuming annotation. Additionally, manual annotation is fraught with ethical concerns given the requirements for researchers to sift through video from a person's daily life, which may often be quite sensitive e.g., toilet breaks. Robust automated (AI) approaches alleviate researcher burden while protecting participant confidentiality.

Design and implementation of floor segmentation algorithms have been implemented within autonomous robots using

lightweight methods e.g., CANNY line detection [21] or Superpixels [22]. Whilst their lightweight nature is useful for edge computing devices in ideal environmental conditions, their application within more general (free-living, home) environments may not be advisable given the greater complexity (e.g., clutter on the ground or strong variations in lighting conditions) and the algorithms inability to learn from previous examples. Unlike the previously mentioned studies the algorithm presented here has an ability to further learn and improve accuracy with accumulation of further training data.

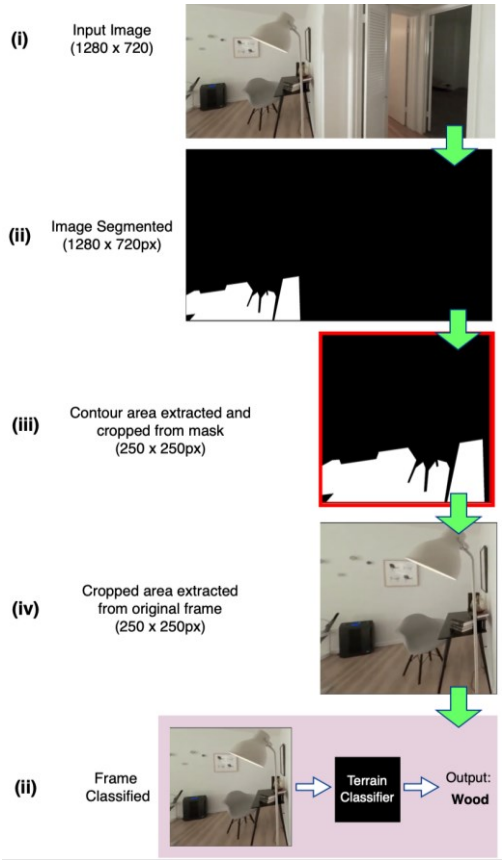

Figure 3: Segmentation to terrain pipeline: (i) Input image (ii) image segmented to binary segmentation mask (iii) contour detection extracts terrain area and region to produce a cropped mask (iv) cropped mask region applied to original image to provide a crop from original frame (v) final crop is fed to terrain classification.

B. Limitations, new opportunities, and future work

Computational complexities and time requirements of the chosen algorithm (1 fps) are a limitation. High complexity requires processing to be completed off-device and on a remote server, necessitating an internet connection and no current possibility of on-device real time processing.

To overcome lack of participant recruitment (COVID), an alternative source was investigated. Here we adopted a unique and alternative approached by mining data from a social media platform, which may be useful for other researchers in the field who struggle to recruit participants to grow datasets. However, mined video data may not be fully representative of the true conditions lived in by those with PD who may be from low socio-economic groups. Future work for the model would be to use of more representative environments through the capture of primary videos within homes.

C. Increasing dataset

With a wealth of available data to be mined from online video sharing platforms the only constraint on the acquisition of the training data is the manual labelling process (for ground truth labelling) of extracted frames. As work continues to increase the volume of the dataset the accuracy (IoU results) of the model proposed should increase further.

V. CONCLUSION

This paper presents a valid and functioning deep learning-based floor segmentation and classification model to automatically interpret indoor floor terrains from videos. The methodology is built upon a new dataset mined from a video sharing platform. The proposed model will aid increasing accuracies as it enables terrain classifiers to cover the entire floor span without including noise or cluttering objects within the periphery. The model will be used in conjunction with wearable IMU's to better inform free-living gait.

VI. FUNDING

This research is co-funded by a grant from the National Institute of Health Research (NIHR) Applied Research Collaboration (ARC) North East and North Cumbria (NENC). This research is also co-funded by the Faculty of Engineering and Environment

at Northumbria University.

REFERENCES

[1] R. Morris, et al, "Gait and cognition: mapping the global and discrete relationships in ageing and neurodegenerative disease," *Neuroscience & Biobehavioral Reviews,* vol. 64, pp. 326-345, 2016.

[2] Y. Celik, et al, "Gait analysis in neurological populations: Progression in the use of wearables," *Med Eng & Physics,* vol. 87, pp. 9-29, 2021.

[3] D. Powell, et al, "Investigating the AX6 inertial-based wearable for instrumented physical capability assessment of young adults in a low-resource setting," *Smart Health,* vol. 22, p. 100220, 2021.

[4] S. Del Din, et al, "Free-living gait characteristics in ageing and Parkinson's disease: impact of environment and ambulatory bout length," *Journal of NeuroEng & Rehab,* vol. 13, no. 1, pp. 1-12, 2016.

[5] M. Nouredanesh, et al, "Fall risk assessment in the wild: A critical examination of wearable sensor use in free-living conditions," *G&P,* vol. 85, 178-190, 2021.

[6] R. Cutlip, et al, "Evaluation of an instrumented walkway for measurement of the kinematic parameters of gait," *G&P,* 12(2), 2000.

[7] F. Coutts, "Gait analysis in the therapeutic environment," *Manual therapy,* vol. 4, no. 1, pp. 2-10, 1999.

[8] Y. Celik, et al, "Multi-modal gait: A wearable, algorithm and data fusion approach for clinical and free-living assessment," *Information Fusion,* vol. 78, pp. 57-70, 2022.

[9] S. Del Din, et al, "Free-living monitoring of Parkinson's disease: Lessons from the field," *Movement Disorders,* 31(9), 2016.

[10] S. Lord, et al, "Moving forward on gait measurement: toward a more refined approach," *Movement Disorders,* 28(11), 1534-1543, 2013.

[11] M. Hashmi, et al, "What lies beneath one's feet? terrain classification using inertial data of human walk," *Applied Sciences,* 9(15), 2019.

[12] A. Weiss *et al.*, "Does the evaluation of gait quality during daily life provide insight into fall risk? A novel approach using 3-day accelerometer recordings," *NeuroReh & Neural repair,* 27(8), 2013.

[13] C. Commons. "ABOUT CC LICENSES - CREATIVE COMMONS." https://creativecommons.org/about/cclicenses/ (access 11/01, 2022).

[14] A. Dutta, et al, "VGG image annotator (VIA)," *URL: http://www. robots. ox. ac. uk/~ vgg/software/via,* 2016.

[15] O. Ronneberger, et al, "U-net: Convolutional networks for biomedical image segmentation," in *International Conf Medical image computing and computer-assisted intervention*, 2015: Springer, pp. 234-241.

[16] T.-Y. Lin, et al, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117-2125.

[17] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," *IEEE Visual Communications and Image Processing (VCIP)*, 2017: IEEE, pp. 1-4.

[18] H. Zhao, et al, "Pyramid scene parsing network," *Proceedings IEEE conf on computer vision and pattern recognition*, 2017, 2881-2890.

[19] D. Zhou *et al.*, "Iou loss for 2d/3d object detection," in *2019 International Conference on 3D Vision (3DV)*, 2019: IEEE, 85-94.

[20] L.-C. Chen, et al, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE trans on pattern analysis and machine intelligence,* 40(4), 2017.

[21] G. C. Barcelo, et al, "Image-based floor segmentation in visual inertial navigation," *IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2013: IEEE, pp. 1402-1407.

[22] F. G. Rodriguez-Telles, et al, "A fast floor segmentation algorithm for visual-based robot navigation," in *2013 International Conference on Computer and Robot Vision*, 2013: IEEE, pp. 167-173.