# DETERMINATION OF GAIT EVENTS FROM 2D VIDEO USING LONG SHORT-TERM MEMORY NEURAL NETWORKS

**Marion Mundt[1], Steffi Colyer[2], Jacqueline Alderson[1]**

**[1]UWA Tech & Policy Lab, The University of Western Australia, Australia**
**[2]The Centre for the Analysis of Motion, Entertainment Research and Applications, University of Bath, United Kingdom**

The purpose of this study was to automatically identify the key gait events, foot-strike and foot-off, from 2D video data. Markerless motion capture and pose estimation have become accepted tools in many biomechanics applications to automatically analyse 2D videos of human movement. However, the accurate detection of gait events from various camera views is still a challenge. We trained a long short-term memory neural network to identify foot-strike and foot-off events in walking and running trials captured from nine different camera views based on 2D pose estimation keypoint labels. We achieved a detection accuracy of 86.3-96.1% (F1 score 76.2-92.5%). These results show the applicability of machine learning tools for the automatic detection of key event frames, which will help practitioners to easily identify frames of interest for further biomechanical analyses.

**KEYWORDS:** machine learning, pose estimation, markerless.

**INTRODUCTION:** In clinical and sports biomechanics servicing environments the full automation of 2D video-based biomechanical analyses is desired. The recent developments in deep learning have made two- and three-dimensional (2D, 3D) markerless motion capture an accepted tool in multiple applications to replace the laborious manual digitisation processes previously required for video analyses (Needham et al., 2021, Wade et al., 2023). While 3D markerless motion capture usually relies on multiple calibrated cameras, 2D analysis can be performed from a single camera view. However, the accurate detection of anatomical landmarks by pose estimation models is just the first step to automate the full biomechanical 2D video analysis. These pipelines further include the accurate detection of foot-strike and foot-off events to determine frames and phases of interest and the calculation of relevant kinematic and kinetic parameters. The absence of force plates, which are the gold-standard tool for foot-strike and foot-off detection, in on-field analyses creates a new challenge. A multitude of algorithms based on thresholds in 3D optical foot marker position, velocity or acceleration have been suggested to determine key gait events in running and walking (e.g. Maiwald et al., 2009, Handsaker et al., 2016) using the vertical trajectory of markers placed on the heel, first metatarsal head, and second toe or shoe tip.
These algorithms encounter a couple of limitations when applied to 2D videos: for the majority of 2D off-the-shelf pose estimation models, no or a limited number of keypoints on the foot segment are available, usually missing the toe/shoe tip, which is used in threshold-based algorithms to accurately detect the foot-off event (e.g. Maiwald et al., 2009). These keypoints also suffer from noise, which makes the application of threshold-based algorithms challenging. Another limitation is the need for a perfectly sagittal camera view and no out-of-plane movements of the participant so that the vertical trajectory of the 2D keypoint is the same as the vertical 3D marker trajectory.
These limitations have driven the use of machine learning instead of threshold-based algorithms. Gait events have been successfully determined from different 2D camera views for walking using video data as inputs to convolutional neural networks (Jamsrandorj et al., 2021). In treadmill running, long short-term memory neural networks (LSTMs) have achieved good results in estimating gait events based on 3D motion capture data inputs (Rivadulla et al., 2021). In this study, we use pose estimation keypoints as inputs to an LSTM to estimate foot-strike and foot-off events for both, overground running and walking trials and compared the outcomes to manually detected events or events detected by the force plate.

**METHODS:** The dataset used in this study was previously published by Needham et al., 2021 and comprises ten running and ten walking trials each from 15 participants captured by nine 2D video cameras (n=2700), 3D motion capture, and force plates (n=300).
OpenPose, an off-the-shelf pose estimation model (Cao et al., 2021), was used to detect 25 keypoints in all 2D videos. 3D marker trajectories representing the OpenPose keypoints were projected to 2D camera views using the video cameras' intrinsic and extrinsic parameters to represent the markers in 2D (as in Wade et al., 2023). Projected and OpenPose keypoint trajectories were low pass filtered ($4^{th}$ order Butterworth filter, cut-off frequency 10 Hz, measurement frequency 200 Hz) (Wade et al., 2023) after linearly interpolating missing values. Keypoints were normalised to the mid-hip keypoint to account for different positions of the person in the global 2D image view, resulting in an input matrix of size $n$ x $s$ x 48, with $n$ being the number of samples, $s$ the sequence length, and 48 features (x-, y-component of all keypoints minus the mid-hip keypoint).
Ground-truth foot-strike and foot-off events were obtained from the down-sampled vertical force with a threshold of 20 N. Since trials included more ground contacts than those on the force plate, additional foot-strike of foot-off events were manually digitised using the 3D marker trajectories of foot markers. The output matrix was constructed fitting a Gaussian distribution around the event to avoid an unbalanced dataset containing mostly zeros for no events with the event frame having a probability of 1. This resulted in an output matrix of $n$ x $s$ x 4, with foot-strike and foot-off for the left and right as features (Jamsrandorj et al., 2021).
Sequences of $s$ = 40 frames were created using a sliding window as input data for a bi-directional LSTM consisting of two hidden layers with 64 neurones each. The LSTM was trained for a maximum number of 300 epochs with early stopping being implemented to prevent overfitting. An ADAM optimiser was used with a mean-squared error loss function.
Three different input datasets were tested: only walking trials ($n$ = 22146), only running trials ($n$ = 15357), and the combination of running and walking trials ($n$ = 37503). LSTMs were trained using all data of one participant for testing, all data of another participant for validation, and the data of 13 participants for training. The training was repeated 15 times, using each participants' data for testing once (leave-one-subject-out cross-validation). All LSTMs were tested separately on running and walking trials and using projected and OpenPose keypoints.
For evaluation, the maximum value exceeding a probability threshold of 0.6 was determined in the estimated time series as the event frame. True positives, true negatives, false positives, and false negatives were calculated. The influence of detection differences of 1-10 frames on accuracy, precision, recall, and F1 score were investigated. Based on the defined threshold, the dataset was unbalanced with 25-30% of the data being positives. Therefore, accuracy is not the ideal measure for the performance but the F1 score is the more reliable measure.

**RESULTS:**

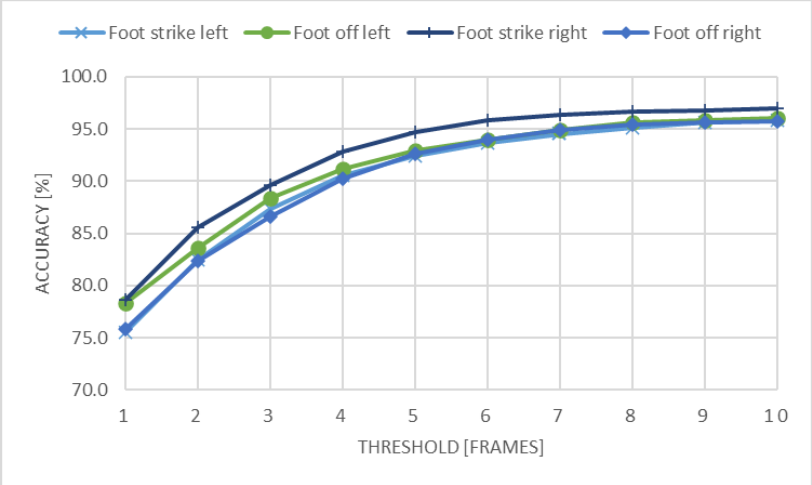


**Figure 1 LSTM model estimation accuracy of foot-strike and foot-off detection for time differences of up to ten frames (50 ms).**

The average estimation accuracy of all four gait events exceeded 90% for a threshold of four frames (20 ms) (Figure 1) when trained on running and walking data and was therefore chosen for the following investigation. The dataset also included more walking (60%) than running data.

The average accuracy, precision, recall and F1 score of the gait event detection for the different training and test inputs is presented in Figure 2. An LSTM solely trained on walking trials, achieved the best results for the determination of foot contacts in walking (F1 score 84.3% for OpenPose and 92.5% for projected keypoints). However, it could not generalise to running (F1 score 0% for OpenPose and projected keypoints). This result was similar for an LSTM that had been trained on running: the F1 score was 76.2% for OpenPose and 79.3% for projected keypoints when evaluated on running, but 10.2% for OpenPose and 20.8% for projected keypoints when evaluated on walking trials. The LSTM trained on both, running and walking, resulted in good estimation for running (F1 score 74.7% for OpenPose and 83.3% for projected keypoints) and walking (F1 score 82.2% for OpenPose and 90.9% for projected keypoints). The detection results using projected keypoints as inputs was up to 10% better than when using OpenPose keypoints. Gait events for walking were estimated with a higher accuracy and F1 score than for running. The detection of foot-strike events showed about 3% better results than the detection of foot-off with no difference between left and right.

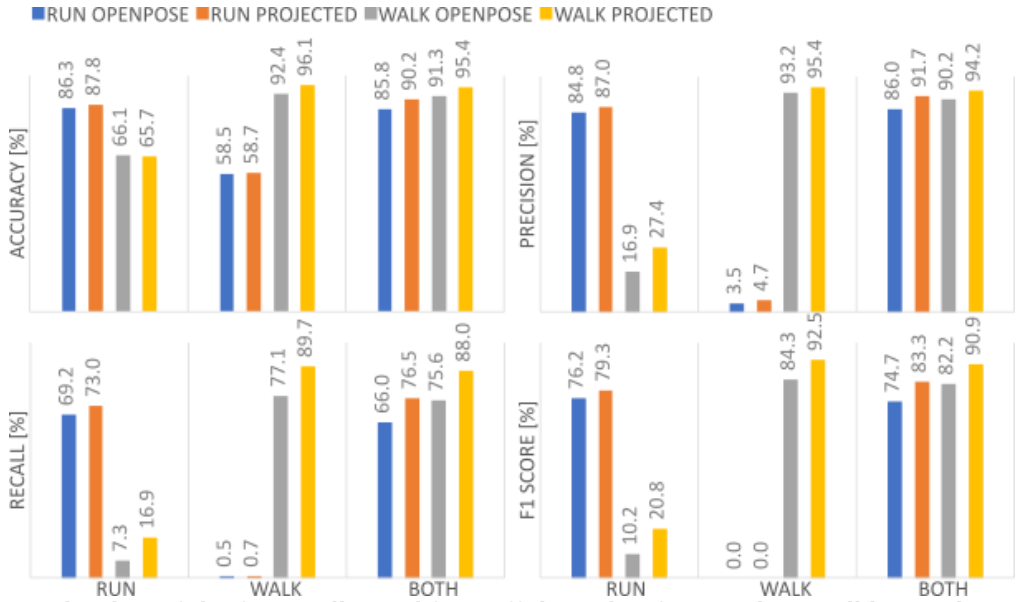


**Figure 2 Evaluation of the foot-strike and foot-off detection for running, walking and a combined input dataset tested of running and walking, for projected and OpenPose keypoints separately.**

**DISCUSSION:** The aim of this study was to apply a machine learning method to automatically detect foot-strike and foot-off in 2D video data. The combination of running and walking trials in one dataset resulted in accurate event detection for both movements (F1 score 74.7-90.9%), while a machine learning model that was trained on a single movement could not generalise to the other movement (F1 score 0-20.8%). This indicates that a model needs to be trained on data of the specific task. Including a larger variety of movements might allow for the development of a more generalisable model but this needs to be further investigated. The higher accuracy of the model when using projected keypoints as inputs shows that markerless motion capture cannot achieve the same accuracy as marker-based motion capture yet.

We found a threshold tolerance in time differences of four video frames (20 ms) to result in high detection accuracy. In another study (Jamsrandorj et al., 2021), the same four frame cut off was used for analysing walking only and resulted in similar accuracy of about 90%. However, the dataset used in this study was captured at 200 Hz while the previous study used a frame rate of 60 Hz; hence the cut off for a successful detection in this study is 20 ms while it is close to 70ms in the other study. It should be further evaluated whether the detection time difference is a constant bias, e.g. a constant delayed detection that might be caused by filtering

of the input keypoints. A random bias could be due to the manual detection of many gait events to enlarge the dataset to include steps in the absence of force plates. A bias between different raters of two to three frames has been previously reported (Bruening and Ridge, 2014), while very high interclass correlation coefficients (>0.90) were found for intra-rater accuracy (Åberg et al, 2021). Both studied used videos captured at 25 Hz, hence it is likely that the intra-rater accuracy is lower in this study for the annotation of 200 Hz videos because of the increased number of frames available.

Using an LSTM for the detection of foot-strike, foot-off, and contact time from a limited number of 3D marker trajectories, velocities and accelerations resulted in good agreement for the event detection from an instrumented treadmill for running (foot-strike bias = 0 [–10, 7] ms, RMSE = 5 ms; foot-off bias = 0 [–10, 10] ms, RMSE = 6 ms, contact time bias = 0 [–15, 15] ms, RMSE = 8 ms) (Rivadulla et al., 2021). The use of kinematic information of a limited number of keypoints and the application to overground running should therefore be further investigated.

In a sports context, camera views might vary. In this study, all nine camera views surrounding the field of interest were used to train the machine learning model. It should be analysed if the model is able to generalise to unknown camera views to ensure that changes in the camera perspective do not influence the result of the detection.

**CONCLUSION:** This study showed the applicability of an LSTM model to detect foot-strike and foot-off events in running and walking videos. Since there are small, but relevant, errors in the detection of gait foot-strike and foot-off events, this model could be used to support the digitisation process of sports videos by providing an estimate of the correct frame to analyse. This would speed up, but not replace, the manual digitisation process of videos for biomechanists. The relevance of a detection offset of 20 ms needs to be considered dependent on the application: while in slow walking the difference in secondary parameters might hardly change, this difference can be relevant in high-speed movements, especially in high-performance sports where error margins are small.

## REFERENCES

Jamsrandorj, A., Jung, D., Kumar, K. S., Arshad, M. Z., Lim, H., Kim, J., & Mun, K. R. (2023). View-independent gait events detection using CNN-transformer hybrid network. *Journal of Biomedical Informatics*, 147. https://doi.org/10.1016/j.jbi.2023.104524

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2021). OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

Rivadulla, A., Chen, X., Weir, G., Cazzola, D., Trewartha, G., Hamill, J., & Preatoni, E. (2021). Development and validation of FootNet; a new kinematic algorithm to improve footstrike and toe-off detection in treadmill running. *PLoS ONE.* 16(8), https://doi.org/10.1371/journal.pone.0248608

Bruening, D.A. & Trager Ridge, S. (2014). Automated event detection algorithms in pathological gait. *Gait & Posture*. 39 (1), 472-477, https://doi.org/10.1016/j.gaitpost.2013.08.023

Åberg, A.C., Olsson, F., Bozkurt Åhman, H., Tarassova, O., Arndt, A., Giedraitis, V., Berglund, L. Halvorsen, K. (2021). Extraction of gait parameters from marker-free video recordings of Timed Up-and-Go tests: Validity, inter- and intra-rater reliability. *Gait & Posture.* 90, 489-495, https://doi.org/10.1016/j.gaitpost.2021.08.004

Needham, L., Evans, M., Cosker, D. P., Wade, L., McGuigan, P. M., Bilzon, J. L., & Colyer, S. L. (2021). The accuracy of several pose estimation methods for 3D joint centre localisation. *Scientific Reports*, 11(1). https://doi.org/10.1038/s41598-021-00212-x

Wade, L., Needham, L., Evans, M., McGuigan, P., Colyer, S., Cosker, D., & Bilzon, J. (2023). Examination of 2D frontal and sagittal markerless motion capture: Implications for markerless applications. *PLOS ONE*, 18(11), e0293917. https://doi.org/10.1371/journal.pone.0293917

Maiwald, C., Sterzing, T., Mayer, T. A., & Milani, T. L. (2009). Detecting foot-to-ground contact from kinematic data in running. *Footwear Science*, 1(2), 111–118. https://doi.org/10.1080/19424280903133938

Handsaker, J. C., Forrester, S. E., Folland, J. P., Black, M. I., & Allen, S. J. (2016). A kinematic algorithm to identify gait events during running at different speeds and with different footstrike types. *Journal of Biomechanics*, 49(16), 4128–4133. https://doi.org/10.1016/j.jbiomech.2016.10.013

**ACKNOWLEDGEMENTS:** This project was partially funded by the Australian Institute of Sport (AIS Research Grant Number 0003223).