Digital Biomarkers

Research Reports - Research Article

Digit Biomark 2022;6:9–18
DOI: 10.1159/000520732

Received: July 22, 2021
Accepted: November 2, 2021
Published online: January 13, 2022

Video-Based Pose Estimation for Gait Analysis in Stroke Survivors during Clinical Assessments: A Proof-of-Concept Study

Luca Lonini[a, b] Yaejin Moon[a] Kyle Embry[a, b] R. James Cotton[a, b]
Kelly McKenzie[a] Sophia Jenz[a] Arun Jayaraman[a, b]

[a]Shirley Ryan Ability Lab, Chicago, IL, USA; [b]Dept. of Physical Medicine and Rehabilitation, Feinberg School of Medicine, Chicago, IL, USA

Keywords
Pose estimation · Video analysis · Deep learning · Stroke · Gait analysis

Abstract
Recent advancements in deep learning have produced significant progress in markerless human pose estimation, making it possible to estimate human kinematics from single camera videos without the need for reflective markers and specialized labs equipped with motion capture systems. Such algorithms have the potential to enable the quantification of clinical metrics from videos recorded with a handheld camera. Here we used DeepLabCut, an open-source framework for markerless pose estimation, to fine-tune a deep network to track 5 body keypoints (hip, knee, ankle, heel, and toe) in 82 below-waist videos of 8 patients with stroke performing overground walking during clinical assessments. We trained the pose estimation model by labeling the keypoints in 2 frames per video and then trained a convolutional neural network to estimate 5 clinically relevant gait parameters (cadence, double support time, swing time, stance time, and walking speed) from the trajectory of these keypoints. These results were then compared to those obtained from a clinical system for gait analysis (GAITRite®, CIR Systems). Absolute accuracy (mean error) and precision (standard deviation of error) for swing, stance, and double support time were within 0.04 ± 0.11 s; Pearson's correlation with the reference system was moderate for swing times (r = 0.4–0.66), but stronger for stance and double support time (r = 0.93–0.95). Cadence mean error was −0.25 steps/min ± 3.9 steps/min (r = 0.97), while walking speed mean error was −0.02 ± 0.11 m/s (r = 0.92). These preliminary results suggest that single camera videos and pose estimation models based on deep networks could be used to quantify clinically relevant gait metrics in individuals poststroke, even while using assistive devices in uncontrolled environments. Such development opens the door to applications for gait analysis both inside and outside of clinical settings, without the need of sophisticated equipment.

Introduction

Gait analysis is a key aspect of clinical assessments for quantifying functional outcomes following a neurological or musculoskeletal disease. A variety of health conditions, such as stroke, Parkinson's disease, cerebral palsy, and spinal cord injury, often cause impairments of gait,

karger@karger.com
www.karger.com/dib

Correspondence to:
Arun Jayaraman, ajayaraman@sralab.org

whose common clinical outcome measures such as 6-min or 10-meter walk tests are unable to quantify in detail. In contrast, kinematic gait analysis provides rich, high-resolution data on movement patterns and is useful for monitoring progress during rehabilitation training and optimizing interventions. However, gait analysis requires the availability of a dedicated gait lab, with specialized and expensive equipment, such as high-speed cameras, instrumented walkways, or dedicated wearable sensors like inertial measurement units [1]. Additionally, extensive amounts of training and experience are required to accurately collect, process, and interpret the data from such a setup.

On the other hand, recent advances in computer vision and deep learning, and the availability of annotated datasets of people and body landmarks have enabled the automated detection of body landmarks (keypoints) from single videos. As such, several open-source algorithms, for 2D and 3D [2] pose estimation and body shape estimation [3], are available, which opens the door to applications in clinical and movement sciences. Such automated video analysis has the potential to become an inexpensive and easy-to-use tool to quantify movement, which could reduce the barrier to obtaining quantitative data on gait outcomes and even enable remote clinical assessments.

Despite the growing number of algorithms for reconstructing body poses, translational research of these models to movement science and gait analysis is still in its infancy. Most studies so far used pretrained models, such as OpenPose [4], to investigate whether gait parameters and clinical outcome measures can be estimated from single videos. Stenum et al. [5] validated the accuracy of spatiotemporal gait parameters estimated from OpenPose, against 3-dimensional motion capture, and found errors within 0.02 s and 0.05 m for temporal and spatial parameters, respectively. However, they limited their analysis to healthy human gait. Ng et al. [6] computed spatiotemporal parameters from videos of older adults with dementia, showing significant associations with balance and fall-risk measures. Sato et al. [7] showed that estimated cadence in individuals' with Parkinson's was correlated with disease status, while Kidziński et al. [8] trained models at predicting multiple gait metrics relevant for treatment planning in children with cerebral palsy, showing moderate to good agreement with the values obtained from motion capture data.

While these studies show the potential of existing pose estimation algorithms at quantifying gait measures, several challenges remain before these systems can be deployed for clinical gait analysis. First, individuals with sensorimotor impairments display idiosyncratic gait patterns [9], which differ widely from those of healthy individuals, thereby rendering the estimation of gait parameters from the sequence of detected keypoints challenging. Second, these individuals may use assistive devices, such as leg braces, walkers, or even robotic exoskeletons during therapy or community mobility, that occlude body parts and may impact the quality of the pose estimation. Third, during clinical gait tests, it is common to have medical personnel assisting the patient. For example, a physical therapist may walk along with the patient to provide assistance or guard them, while they perform a walking test, which requires the pose estimation algorithm to accurately identify and track the patient throughout the video. Finally, if the model uses videos captured with a handheld camera, the gait inference should be robust to different viewing angles and hand motions. Such challenges still need to be addressed and require further adaptation and validation of existing pose estimation models and training data [10].

Here we estimate temporal gait parameters from 82 below-waist videos of stroke survivors undergoing multiple sessions of gait training with a therapist and wearing different types of leg braces. Temporal parameters are among the outcomes characterizing the hemiparetic gait dysfunction of stroke survivors, including reduced walking speeds, decreased cadence, prolonged swing time, and reduced stance time on the paretic side, compared with those parameters of healthy subjects or with the nonparetic side [11]. Additionally, traditional and emerging gait training methods often have aimed to improve the temporal gait parameters and recorded them as treatment response biomarkers [12, 13], or biomarkers related to fall risk [14]. To circumvent the problem of reliably tracking the patient, we used an approach based on transfer learning: we manually annotated the positions of 5 landmarks (keypoints) on the leg and foot in 2 frames for each video, and fine-tuned a pretrained deep learning model on these data using DeepLabCut [15], an open source framework for animal and human pose estimation, at detecting these keypoints in the videos. We then trained a convolutional neural network at predicting 5 parameters (cadence, swing and stance time, double support time, and walking speed) from the 2D trajectories of the keypoints and compare the accuracy and precision of our model to a gold standard system for gait analysis. While our work is still a proof of concept and needs further development on a larger dataset, its main contributions are the following:

Table 1. Participants' demographics, clinical data, and number of videos per patient

Subject ID	Gender	Age, years	Height, cm	Time since stroke, years	Stroke type	Paretic side	Gait speed, m/s	Sessions (videos), *n*
SS02	Female	46	174	9	Isc	R	0.57	6
SS04	Male	67	173	5	Hemo	L	0.37	12
SS06	Female	56	163	2	Isc	L	0.81	11
SS07	Male	59	177	2	Isc	R	0.75	10
SS13	Male	61	163	9	Hemo	R	0.84	10
SS18	Male	64	180	6	Isc	L	0.51	12
SS20	Female	47	163	8	Isc	L	0.56	12
SS22	Male	63	165	6	Hemo	L	0.99	9
AVG	–	**57.9**	**169.8**	**5.9**	–	–	**0.68**	
SD	–	**7.8**	**7.0**	**2.8**	–	–	**0.21**	

Isc, ischemic; Hemo, hemorrhagic; L, left; R, right; AVG, average; SD, standard deviation.

- Temporal gait parameters in stroke survivors can be estimated from single videos framing the lower half of the body using deep learning models.
- A pose estimation model trained on 2 labeled frames per video using DeepLabCut can track the patient foot and leg keypoints.
- A deep network trained on the keypoint sequences can estimate gait parameters from a nonstationary camera and across individuals with different levels of impairments.

Materials and Methods

Participants

This proof-of-concept study consisted of a convenience sample of 8 individuals with stroke (Table 1) who represented a subsample of participants in ongoing gait rehabilitation investigations. Inclusion criteria for all participants were (1) 18 years of age or older, (2) at least 6 months poststroke, (3) hemiparesis/hemiplegia after a single stroke, (4) Functional Ambulation Category [16] of 2 or greater, (5) no presence of severe lower-limb spasticity, (6) no presence of painful musculoskeletal dysfunction, (7) no history of seizures, and (8) no metal implants in the spine or back. Each participant provided informed consent. These procedures were approved by the Northwestern University Institutional Review Board.

Experimental Setup

We used secondary data of a clinical trial where each participant received gait training for 3 days per week for 8 weeks (24 sessions in total). Walking assessments at pre-, mid (i.e., after 12 sessions)-, post-, and 3-month follow-up were video recorded. Participants walked at their self-selected pace along an 8-m instrumented walkway (GAITRite® Gold, CIR Systems, West Conshohocken, PA, USA) with or without a leg brace. In total, we analyzed 82 walking trials from the above clinical study.

The GAITRite® walkway recorded each footfall's initial and final contact time of steps with a reported spatial accuracy of ±1.25 cm and a temporal accuracy of ±8.3 ms at the sampling rate (120 Hz) that was selected. The associated software was then used to compute several spatiotemporal gait measurements. We limited this analysis to the following temporal parameters:

- walking speed (m/s)
- cadence (steps/min)
- swing time (time between final contact of the current cycle and initial contact of the next cycle on the same foot)
- stance time (time between initial contact and final contact of the same cycle on the same foot)
- double support time (sum of the time elapsed when both feet are in contact with the ground during a gait cycle)

Swing and stance time were estimated for each leg and reported separately for the paretic (P) and nonparetic (NP) leg for each participant. Average values of the parameters across all the strides occurring on the walkway were yielded by the GAITRite software and used as the ground truth values for training the subsequent model (see "Gait Estimation model").

A digital RGB video camera recorded the left- or right-side sagittal plane views of the walking sequence at 30 frames per sec (fps) and at a resolution of 1,280 × 720 pixels. The camera was mounted on a tripod placed on a tripod dolly, and the height of the camera placement was set to about 75 cm. The video dataset does not contain identifiable participant information as the videos were taken from the waist down (example video frames with keypoints superimposed shown in Fig. 1). As a participant walked, a researcher tracked the participant by pushing the tripod dolly along the participant's trace. In most sequences, a therapist walked next to the subject to guard the participant and ensure their safety. Both the right and left views of each participant were captured, as subjects performed 3 trials changing the direction between each trial. Recording both views allowed researchers to observe each leg motion without any obstruction, when analyzing the videos of each subject–.

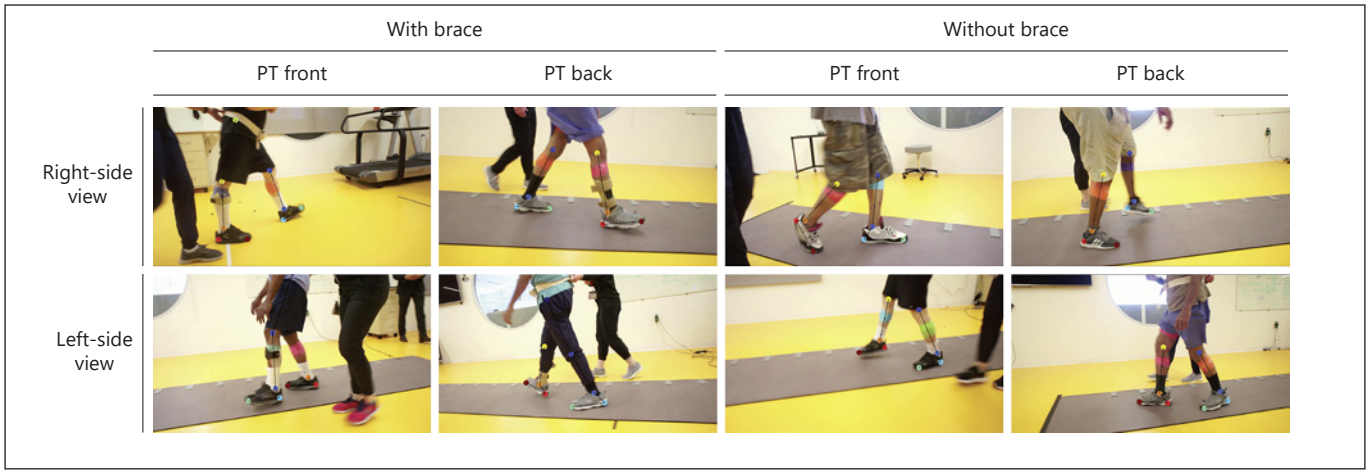

Fig. 1. Example video frames with overlapped keypoint detection from the pose estimation. Participants could wear a variety of clothing and could use different types of leg braces to stabilize their walking. Right- and left-side views were captured. A therapist can be seen walking next to the participant in each frame. PT, physical therapist.

Pose Estimation

We used DeepLabCut to train a model at detecting 5 patient landmarks (keypoints) on each leg in any given video; the tracked keypoints were then used to estimate the gait parameters. DeepLabCut is an open source algorithm for pose estimation, which allows the user to label the key points in their own data and fine-tune a pretrained ResNet or MobileNet architecture [17] at detecting the annotated keypoints in the video. This approach aided the tracking of the patient leg keypoints, regardless of the presence of assistive devices and other people in the scene. We manually annotated 2 frames in each of the 82 videos, for a total of 164 annotated frames: this constituted the training data for the pose estimation model. Five keypoints for each leg (hip, knee, ankle, heel, and toe) were manually labeled in each of the training frames. We did not use the hip key points for the gait parameter estimation, as the hip joint was not always visible in the videos and resulted in noisy detections. Frames to label were selected by a k-means clustering algorithm implemented in DeepLabCut, which chose frames based on dissimilarity. The model was trained for 100,000 epochs using the default Adam optimizer and a variable learning rate schedule (1e-5, 0–15,000 epochs; 5e-5, 15,000–24,000 epochs; and 1e-5, 100,000), and the batch size was set to 4. Data augmentation was used by applying random rotations (±30°), image scaling (zoom between ×0.5 and ×1.5), and scaling the image contrast as provided in the "Imgaug" Python package.

Gait Parameter Estimation

The position of 4 keypoints (knee, ankle, heel, and toe) from both legs throughout the video was used to estimate the aforementioned gait parameters, averaged across all the strides recorded in the video. For healthy individuals, the detection of gait events (e.g., foot contacts, swing, and stance phase) from kinematic data could be achieved with peak-detection algorithms as healthy gait patterns are remarkably similar; in contrast, detecting these events in people with gait impairments and using assistive devices can be significantly more challenging, due to the idiosyncrasies of their gait, and thus benefits from the use of data-driven approaches.

We used a convolutional neural network to predict the average gait parameters for a video from the leg and foot keypoint time series (Fig. 2). The architecture consists of 2 blocks of convolutional and pooling layers. Each block contains two 1D convolutional layers (32 filters, kernel size = 8), followed by a max-pooling layer (stride = 2). The kernel size of 8 corresponds to a window of 0.24 s. Batch normalization was applied after each block of 2 convolutional layers and before the pooling layer. The last 2 dense layers contain, respectively, 10 neurons with a ReLu activation function and 1 output neuron with linear activation function. A single dropout layer (rate = 0.1) was added before the 2 dense layers. This network architecture has been used successfully in previous work to estimate gait parameters from sensor [18] and video [8] data. An input data point had a size of 120 × 8 (time steps × keypoints); the input depth was 8 as it comprised 4 keypoints from each leg, and the output was a single gait parameter (e.g., cadence). We trained 1 network per gait parameter (i.e., single output network), as this resulted in more accurate results than when training the same network at predicting all parameters simultaneously (i.e., multiple outputs network). Mean squared error was chosen as the network loss function and RMSProp as the optimization algorithm (learning rate = 0.001). A batch size of 32 was chosen and the network was trained using early stopping for a total of 200 epochs.

The network was trained on 4 s long (120 samples) time series of the left and right heel, toe, ankle, and knee x-coordinates (the horizontal axis of each video), as output by the network trained with DeepLabCut. Since all videos were >4 s (mean duration = 13.3 s, standard deviation = 5.9 s), we split the time series into consecutive 4-s sequences (with 75% overlap) to increase the amount of data available to train the network. This yielded a total of 656 data points (keypoint time series sequences). The prediction for a video was obtained by averaging predictions from all the individual 4-s sequences extracted from the video. We chose a duration of 4 s, as this value was about half the duration of the shorter videos, and long enough to capture the whole gait cycle for all the individuals. We did not use the y-coordinates of the key points, as it did not provide any significant advantage on model accuracy.

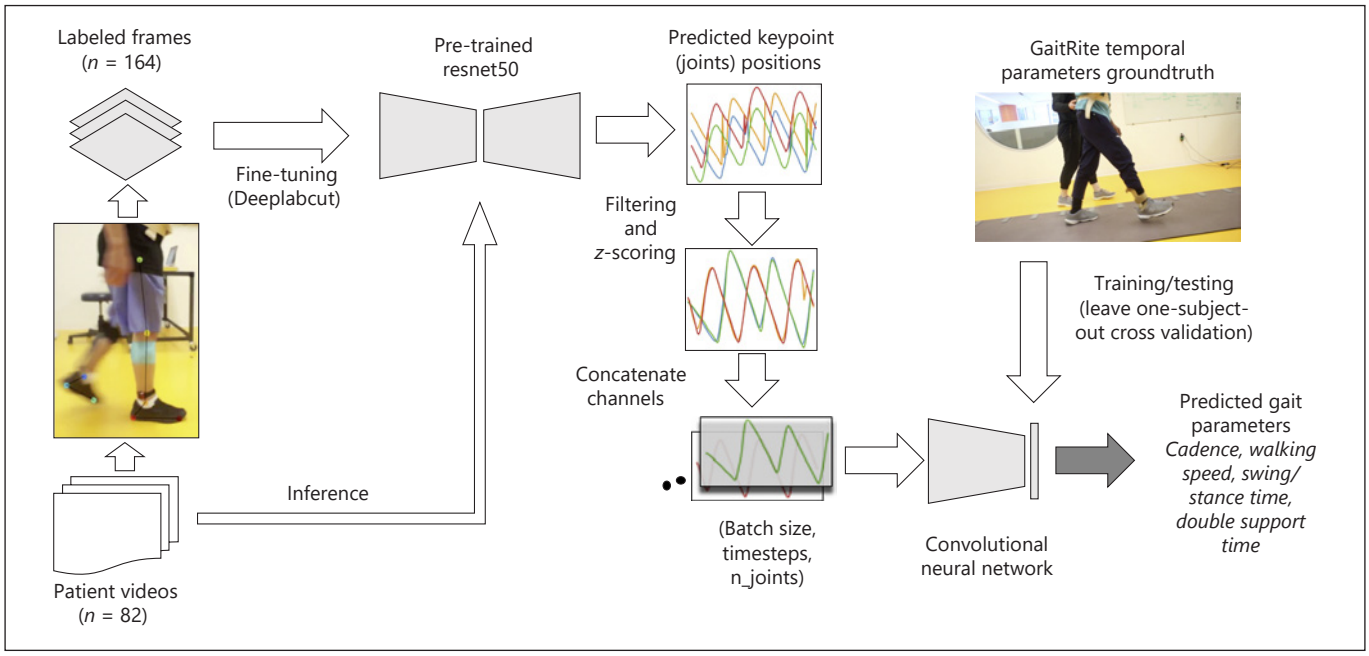

Fig. 2. Data pipeline: we labeled the location of 5 keypoints on each leg and foot in 164 frames drawn from a set of 82 videos and use the labeled frames to fine tune a pretrained ResNet50 deep network architecture through DeepLabCut. At inference time, we used the fine-tuned model to track the keypoints in the video. The keypoint sequences of each joint (different colors indicate different joints) were filtered, normalized by z-scoring, and concatenated along the third dimension to train a second convolutional neural network at predicting 5 gait temporal parameters. Ground truth values for gait parameters were derived from data collected using a reference clinical system (GAITRite). A leave-1-subject-out cross-validation was used to evaluate the model performance.

The keypoint time series were affected by low-frequency noise due to camera movements while following the participant, as well as high-frequency noise and missed detections due to inaccuracies in the pose estimation model. As such, keypoint time series from each video were linearly interpolated, filtered with a high-pass 8-th order Butterworth filter (cutoff frequency 0.25 Hz) to detrend the signal, and passed through a median filter (window size = 5 frames). We finally rescaled the signal by subtracting the mean and dividing by the standard deviation of the time series and applied a Gaussian filter (sigma = 1 frame) to remove remaining high-frequency noise. We inverted the sign of the time series signal for the videos where participants were walking toward the left so that these signals matched those from videos that were recorded from the right.

Statistical Analysis

A leave-1-subject-out cross-validation was used to train and evaluate the network performance: we split data such that training and test folds contained videos from different subjects. Average ground truth values for the gait parameters were obtained from the GAITRite system as mentioned above. For each gait parameter, we computed the Pearson's correlation coefficients as a measure of agreement between the GAITRite measurements (truth value) and the convolutional network output (estimated value). As additional evaluation metrics, we calculated the mean and standard deviation of the error for each parameter, which correspond to the accuracy ± precision of the model. Specifically, we computed 2 types of errors, absolute (Abs) and relative (Rel) errors, on the videos in the test set, where the error on video *i* is as follows:

$$\mathrm{Abs}_{\mathrm{error}_i} = \mathrm{estimated}_i - \mathrm{truth}_i, \quad (1)$$

$$\mathrm{Rel}_{\mathrm{error}_i} = \frac{\mathrm{estimated}_i - \mathrm{truth}_i}{\mathrm{truth}_i} \times 100\,. \quad (2)$$

For each type of error, we calculated its mean and standard deviation (i.e., accuracy and precision), where accuracy quantifies how far the predicted parameter deviates from the reference GAITRite measure, and precision quantifies how consistent are the predictions with each other.

$$\mathrm{Accuracy}_{\mathrm{abs}} = \mu\,(\mathrm{Abs}_{\mathrm{error}});\ \mathrm{Accuracy}_{\mathrm{rel}} = \mu\,(\mathrm{Rel}_{\mathrm{error}}), \quad (3)$$

$$\mathrm{Precision}_{\mathrm{abs}} = \sigma\,(\mathrm{Abs}_{\mathrm{error}});\ \mathrm{Precision}_{\mathrm{rel}} = \sigma\,(\mathrm{Rel}_{\mathrm{error}}). \quad (4)$$

Results

Gait Parameter Estimation

Participants' walking abilities varied widely: gait parameter values, as measured by the GAITRite system, spanned a broad range across participants and conditions (e.g., use of a leg brace vs. no-brace or pre- vs. post gait

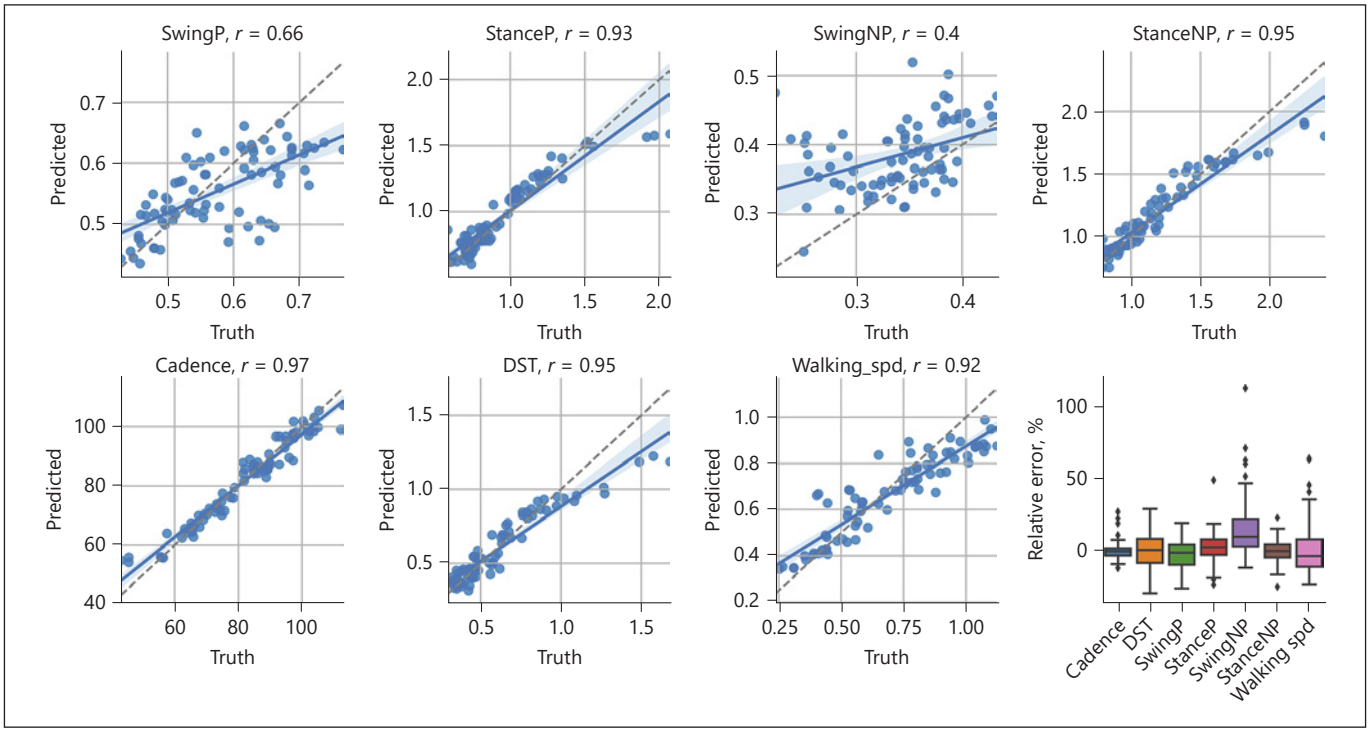

Fig. 3. Gait parameter estimation: scatter plots showing the correlations between the gait parameters estimated by the neural network (y-axis, predicted) and those yielded by the GAITRite system (x-axis, ground truth). Each dot represents the datum for a single video in the test set. Most of the estimated parameters showed high agreement with the GAITRite; lower correlations were obtained for swing times for the paretic (P, $r = 0.66$) and nonparetic legs (NP, $r = 0.4$) legs. The lower right panel shows a box plot of relative errors for each parameter. DST, double support time; P, paretic side; NP, nonparetic side.

Table 2. Distribution of gait parameter values measured by the reference system (GAITRite©) across participants

Parameter	Mean	SD	Min	Max
Cadence, steps/min	82.5	16	43.1	112.9
DST, s	0.61	0.30	0.29	1.68
Walking speed, m/s	0.68	0.24	0.24	1.12
Swing paretic side, s	0.58	0.09	0.43	0.77
Stance paretic side, s	0.95	0.31	0.58	2.07
Swing non paretic side, s	0.34	0.05	0.22	0.43
Stance non paretic side, s	1.18	0.36	0.79	2.40

DST, double support time.

training). Table 2 shows the summary statistics for all the measured parameters. For example, the mean cadence of participants ranged from 43 to 113 steps/min (mean = 83 steps/min, SD = 16), while mean walking speed was 0.68 m/s and ranged from 0.24 to 1.12 m/s.

To compare the accuracy of our pose estimation-based model, we first looked at Pearson's correlations between the estimated parameters and those from the reference GAITRite system (Fig. 3). Most parameters showed high agreement with the reference system: correlations for cadence ($r = 0.97$), double support time (DST, $r = 0.95$), stance times (paretic side $r = 0.93$; nonparetic side $r = 0.95$), and walking speed ($r = 0.92$) were higher than those for swing (paretic side $r = 0.66$). Correlation of swing times for the nonparetic side was the lowest overall (nonparetic side $r = 0.4$).

We measured the network absolute and relative error in terms of accuracy ± precision (Table 3) with respect to the GAITRite system for each gait parameter (Fig. 4). Cadence had an absolute accuracy of −0.25 steps/min (±3.9) and a relative accuracy of 0.3% (±5.9%). Absolute accuracy for swing and stance parameters was of the order of 1 frame time (i.e., 0.03 s), while absolute precision for swing times (paretic: ±0.06 s; nonparetic: ±0.05 s) was higher than that of stance times (paretic: ±0.11 s; nonparetic: ±0.12 s). Relative precision for swing and stance times was within ±10.5%,

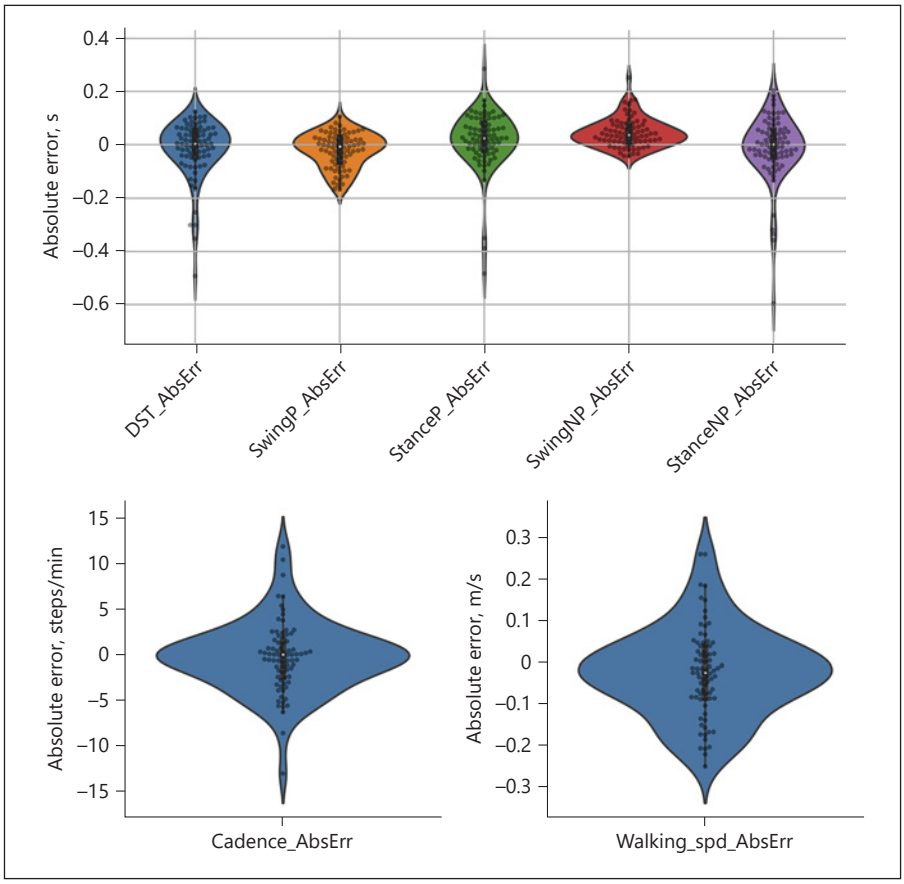

Fig. 4. Estimation errors: violin plots showing the distribution of absolute errors for swing, stance, and double support time (top panel) as well as for cadence and walking speed (bottom panels). Each dot represents the estimated value for a single video in the test set (leave 1 subject out cross-validation). DST, double support time; P, paretic side; NP, nonparetic side.

Table 3. Absolute and relative errors of gait parameters estimated by the model with respect to the GAITRite reference system

Parameter	Absolute error		Relative error [%]	
	accuracy	precision	accuracy	precision
Cadence, steps/min	−0.25	3.88	0.3	5.9
Double support time, s	−0.02	0.11	0.6	13.5
Swing (P), s	−0.02	0.06	−2.5	10.5
Stance (P), s	0.02	0.11	2.7	10.1
Swing (NP), s	0.04	0.05	15	20.6
Stance (NP), s	−0.01	0.12	0.2	8.0
Walking speed, m/s	−0.02	0.11	0.7	18.0

P, paretic side; NP, nonparetic side.

except for the nonparetic side swing time, which had the lowest relative precision (±20.6%); as swing times were shorter for the nonparetic side, the relative error was higher. Absolute accuracy for double support time was in a similar range to that of stance times (−0.02 ± 0.11 s; relative 0.6 ± 13.5%). Walking speed absolute accuracy was −0.02 ± 0.11 m/s (relative: 0.7 ± 18%).

Discussion/Conclusion

Markerless pose estimation based on deep learning holds the promise of becoming an easy-to-access method to analyze human movements without the use of a motion capture lab. Here we used DeepLabCut, a framework for human and animal pose estimation, to fine-tune an archi-

tecture based on a deep residual network (ResNet50) on 82 videos of stroke survivors and track 4 keypoints on each leg and foot, as participants walked in a gait lab during clinical assessments. We then trained a second convolutional neural network on the keypoint trajectories to estimate a set of temporal gait parameters, validating the results against a clinical gold standard system for gait analysis.

Labeling only 2 frames per video allowed us to track the body landmarks with sufficient accuracy for training a convolutional network to infer the temporal gait parameters from the keypoint trajectories. Most parameters, except for swing times, showed correlations equal or greater than 0.92 with the reference GAITRite system. Swing times for the nonparetic limb yielded the lowest correlation ($r = 0.40$). Absolute precision for swing times ranged from ±0.05 s to ±0.06 s (paretic and nonparetic side); as such, relative precision was worse for the nonparetic side (±20.5%) than for the paretic side (±10.5%), as swing times for the nonparetic limb are of shorter durations. Absolute precision for stance times and double support time was about half that of swing time (±0.11 s to ±0.12 s); similar values have been reported when training a model to segment strides in individuals with gait impairments from inertial measurement unit data [18]. Cadence was the parameter with the lowest error overall (−0.25 ± 3.7 steps/min, relative precision ±5.9%). Mean precision for walking speed was 0.11 m/s, which is less than the minimal clinically important difference (= 0.14 m/s) for changes in gait speed for individuals with stroke [19]. This is remarkable, also considering that the model was agnostic to the camera geometry.

Our methodology resembles that of Kidziński et al. [8], as a convolutional neural network was used to predict gait parameters and markers of gait pathology from the sequence of 2D keypoints in videos of children with cerebral palsy; however, they used OpenPose to extract the keypoints from the videos, and only chose videos where the entire person was in the frame, subject clothing was similar, and no other people were present. They reported correlations, relative to a motion capture system, of 0.73 and 0.79 for walking speed and cadence. Moro et al. [20] used DeepLabCut to estimate 4 spatiotemporal gait parameters in stroke survivors from lateral videos. They compared their estimation to a marker-based motion capture system, reporting spatial errors in the order of centimeters, and found that differences from the reference system were not statistically significant. Their setup though used a fixed RGB camera that only captured 1 stride for each of the 10 subjects, thereby limiting the generalization of their results.

We chose to use DeepLabCut because of the nature of our data; in contrast to previous studies, our videos presented several simultaneous challenges, including the presence of a therapist walking next to the participant, varying clothing, and the presence of leg braces worn by the participants. In addition, to preserve the privacy of the participants, the videos only framed the lower half of the body, which hinders the ability of pretrained pose estimation models to track body landmarks. While the use of pretrained algorithms does not require manually labeling and training the model on the labeled data, our initial attempts using OpenPose and AlphaPose [21] resulted in unreliable tracking of the legs and foot keypoints in our data. This is also in line with previous reports about the current limits of pretrained pose estimation models [10], which were not trained on datasets that include individuals with mobility impairments or validated in clinical settings.

We trained a convolutional network to estimate the gait variables from the tracked foot keypoints. Simpler methods [22] based on peak detection and pixel velocity [23, 24] of the foot could have been used here to segment the gait phases and compute the gait parameters directly; however, these methods rely on manually tuning threshold parameters and can be less reliable in the presence of abnormal gait patterns, such as those seen in patients with stroke. Indeed, our participants displayed widely different kinematics due to variable neurological impairments, as well as because of the presence/absence of a leg brace. In such a scenario, approaches that directly predict the gait parameters from the kinematic data may provide better results [25, 26].

A primary limitation of our study is the limited sample size and the fact that we had to manually label frames in each video to train the pose estimation model. Due to the limited number of subjects and videos and the uncontrolled recording conditions, generalization to new videos is known to be a challenge when fine-tuning a deep network using DeepLabCut [27]. We did not quantify the generalization error of the pose estimation model and the accuracy of the system at estimating gait parameters in new unseen videos, which remains as future work. Second, we trained the network to predict the average gait parameters from all the strides captured by the GAITRite in a walking trial; however, the model input was a 4-s-window, which would only contain one or few strides. As stride-to-stride variability is common in populations with neurological impairments [28, 29], this may have caused inaccuracies in the training signal and the resulting model predictions. We limited our analysis to gait

temporal parameters, although spatial parameters could have been considered as well. Estimating spatial parameters would require 2 or more camera views to determine the 3D position of points in the scene; alternatively, a different class of pose estimation models that can reason about the 2D joint estimation to infer the 3D configuration of the body from a single frame could be used: these latter models include "3D lifting" approaches [30, 31] that map from 2D to 3D joint locations, or those that use 3D body models [32, 33]. However, these models require a complete view of the person and have not been validated on clinical populations.

Future work should explore how many annotated data are necessary to provide sufficient generalization on new videos, which is known to be a common problem with transfer learning models [34]. This problem could be mitigated by employing label propagation methods, which leverage sparse labeled frames to predict the poses of a person in neighboring frames [35, 36]. Similarly, dealing with the presence of a second person in the video (e.g., therapist), as well as with truncated or occluded views of the body, requires accurate identification and tracking of the target patient and their body landmarks. While pose estimation models are starting to incorporate person identification and temporal tracking of each person [37], their accuracy for ambient monitoring in a clinical environment [38] where clinicians may interact or assist the patient during the visit remains to be seen.

Acknowledgment

The authors would like to thank Jasmine Hunt for helping with annotating the video dataset used in the analysis.

Statement of Ethics

All participants provided written informed consent. The study protocol was reviewed and approved by the Northwestern University IRB (Study No. STU00206430).

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

This study was funded by the Frankel Family Foundation, the Max Nader Lab, and the Shirley Ryan AbilityLab. All 3 funding sources contributed equally in study design, implementation, and dissemination.

Author Contributions

L.L. conceived of the study, performed data analysis, and wrote the manuscript; Y.M. recorded the data, performed data analysis, and contributed to writing the manuscript; K.E. performed data analysis and revised the manuscript; J.R.C. provided expertise on the analysis design and revised the manuscript; K.M. enrolled participants, recorded the data, and revised the clinical aspects of the manuscript. S.J. preprocessed and organized the video data and revised the manuscript; and A.J. conceived of the study, supervised the project, and revised the manuscript.

Data Availability Statement

The video data used in the study can be made available from the authors upon reasonable request.

References

1 Chen S, Lach J, Lo B, Yang GZ. Toward pervasive gait analysis with wearable sensors: a systematic review. IEEE J Biomed Health Inform. 2016;20(6):1521–37.
2 Chen Y, Tian Y, He M. Monocular human pose estimation: a survey of deep learning-based methods. Comput Vis Image Underst. 2020;192(March):1–23.
3 Kolotouros N, Pavlakos G, Black M, Daniilidis K. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. Proc IEEE Int Conf Comput Vis. 2019;2019:2252–61.
4 Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell. 2021 Jan 1;43(1):172–86.
5 Stenum J, Rossi C, Roemmich RT. Two-dimensional video-based analysis of human gait using pose estimation. PLoS Comput Biol. 2021;17(4):e1008935.
6 Ng KD, Mehdizadeh S, Iaboni A, Mansfield A, Flint A, Taati B. Measuring gait variables using computer vision to assess mobility and fall risk in older adults with dementia. IEEE J Transl Eng Health Med. 2020;8(May):2100609.
7 Sato K, Nagashima Y, Mano T, Iwata A, Toda T. Quantifying normal and parkinsonian gait features from home movies: practical application of a deep learning–based 2D pose estimator. PLoS One. 2019;14(11):1–15.
8 Kidziński Ł, Yang B, Hicks JL, Rajagopal A, Delp SL, Schwartz MH. Deep neural networks enable quantitative movement analysis using single-camera videos. Nat Commun. 2020;11(1):1–10.
9 Nadeau S. Understanding spatial and temporal gait asymmetries in individuals post stroke. Int J Phys Med Rehabil. 2014;2:3.
10 Seethapathi N, Wang S, Saluja R, Blohm G, Kording KP. Movement science needs different pose tracking algorithms. 2019 Jul 24. [cited 2021 Jul 15].
11 Patterson KK, Gage WH, Brooks D, Black SE, McIlroy WE. Changes in gait symmetry and velocity after stroke: a cross-sectional study from weeks to years after stroke. Neurorehabil Neural Repair. 2010 Nov [cited 2021 Oct 16];24(9):783–90.

12 Holleran CL, Straube DD, Kinnaird CR, Leddy AL, Hornby TG. Feasibility and potential efficacy of high-intensity stepping training in variable contexts in subacute and chronic stroke. Neurorehabil Neural Repair. 2014 Feb 10 [cited 2021 Oct 16];28(7):643–51.
13 Shin SY, Lee RK, Spicer P, Sulzer J. Does kinematic gait quality improve with functional gait recovery? A longitudinal pilot study on early post-stroke individuals. J Biomech. 2020 May 22 [cited 2021 Oct 18];105:109761.
14 Ehrhardt A, Hostettler P, Widmer L, Reuter K, Petersen JA, Straumann D, et al. Fall-related functional impairments in patients with neurological gait disorder. Sci Rep. 2020 Dec 3 [cited 2021 Oct 17];10(1):1–10.
15 Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci. 2018;21(9):1281–9.
16 Mehrholz J, Wagner K, Rutte K, Meissner D, Pohl M. Predictive validity and responsiveness of the functional ambulation category in hemiparetic patients after stroke. Arch Phys Med Rehabil. 2007 [cited 2021 Jun 30];88(10):1314–9. Available from:
17 Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B. Deepercut: a deeper, stronger, and faster multi-person pose estimation modelLNCS. Lect Notes Comput Sci. 2016;9910: 34–50.
18 Hannink J, Kautz T, Pasluosta CF, Gasmann KG, Klucken J, Eskofier BM. Sensor-based gait parameter extraction with deep convolutional neural networks. IEEE J Biomed Health Inform. 2017;21(1):85–93.
19 Bohannon RW, Glenney SS. Minimal clinically important difference for change in comfortable gait speed of adults with pathology: a systematic review. J Eval Clin Pract. 2014 [cited 2021 Jun 1];20:295–300.
20 Moro M, Marchesi G, Odone F, Casadio M. Markerless gait analysis in stroke survivors based on computer vision and deep learning: a pilot study. Proc ACM Symp Appl Comput. 2020:2097–104.
21 Xiu Y, Li J, Wang H, Fang Y, Lu C. Pose flow: efficient online pose tracking. Br Mach Vis Conf. 2018 Feb 3 [cited 2021 Jul 15].
22 Hreljac A, Marshall RN. Algorithms to determine event timing during normal walking using kinematic data. J Biomech. 2000;33(6):783–6.
23 Fellin RE, Rose WC, Royer TD, Davis IS. Comparison of methods for kinematic identification of footstrike and toe-off during overground and treadmill running. J Sci Med Sport. 2010 Nov 1;13(6):646–50.
24 Auvinet E, Multon F, Aubin CE, Meunier J, Raison M. Detection of gait cycles in treadmill walking using a Kinect. Gait Posture. 2015 Feb 1;41(2):722–5.
25 Haji Ghassemi N, Hannink J, Martindale CF, Gaßner H, Müller M, Klucken J, et al. Segmentation of gait sequences in sensor-based movement analysis: a comparison of methods in Parkinson's disease. Sensors. 2018;18(1):1–15.
26 Kidziński Ł, Delp S, Schwartz M. Automatic real-time gait event detection in children using deep neural networks. PLoS One. 2019;14(1):1–11.
27 Mathis A, Biasi T, Schneider S, Yüksekgönül M, Rogers B, Bethge M, et al. Pretraining boosts out-of-domain robustness for pose estimation. 2021 [cited 2021 Jul 15]. Available from: http://horse10.
28 Balasubramanian CK, Neptune RR, Kautz SA. Variability in spatiotemporal step characteristics and its relationship to walking performance post-stroke. Gait Posture. 2009 Apr [cited 2021 Jul 15];29(3):408.
29 Sanchez N, Schweighofer N, Finley JM. Different biomechanical variables explain within-subjects versus between-subjects variance in step length asymmetry post-stroke. IEEE Trans Neural Syst Rehabil Eng. 2021 Jun 17;29:1188–98.
30 Martinez J, Hossain R, Romero J, Little JJ. A simple yet effective baseline for 3d human pose estimation. Proc IEEE Int Conf Comput Vis. 2017 Dec 22;2017:2659–68.
31 Pavllo D, Zürich E, Feichtenhofer C, Grangier D, Brain G, Auli M. 3D human pose estimation in video with temporal convolutions and semi-supervised training. 2019 [cited 2021 Oct 17]. Available from: https://github.com/.
32 Kanazawa A, Black MJ, Jacobs DW, Malik J. End-to-end recovery of human shape and pose input reconstruction side and top down view part segmentation input reconstruction side and top down view part segmentation. [cited 2021 Oct 17]. Available from: https://akanazawa.github.io/hmr/.
33 Kocabas M, Athanasiou N, BlackVIBE MJ. : video inference for human body pose and shape estimation. Oct 17] In (pp. 5253-5263). https://github.com/mkocabas/VIBE.
34 Raghu M, Zhang C, Brain G, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging. 2019.
35 Bertasius G, Feichtenhofer C, Tran D, Shi J, Torresani L. Learning temporal pose estimation from sparsely-labeled videos. Adv Neural Inf Process Syst. 2019 Jun 6 [cited 2021 Jul 15]:32.
36 Wu A, Buchanan EK, Whiteway M, Schartner M, Meijer G, Noel JP, et al. Deep graph pose: a semi-supervised deep graphical model for improved animal pose tracking. Adv Neural Inf Process Syst. 2020 [cited 2021 Jul 15];33:6040–52. Available from: https://github.com/paninski-lab/deepgraphpose.
37 Wang M, Tighe J, Modolo D. Combining detection and tracking for human pose estimation in videos. 2020.
38 Haque A, Milstein A, Fei-Fei L. Illuminating the dark spaces of healthcare with ambient intelligence. Nature. 2020 [cited 2021 Jul 3];585:193–202.