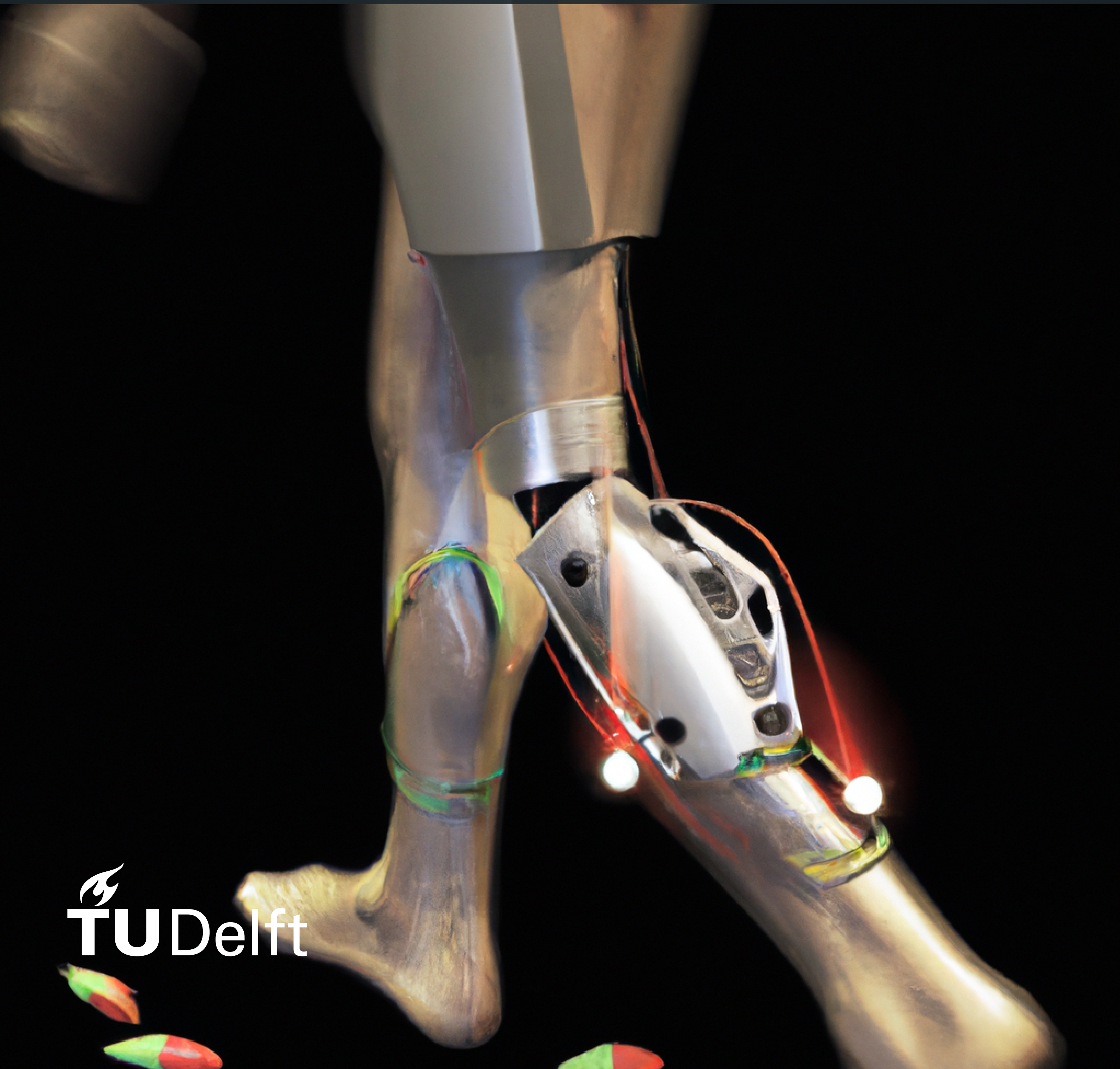
Msc Thesis

A deep learning approach to (semi-) automatically track bone movement in ultrasound images of patients with a unilateral transtibial prosthesis

Daniël Donse



Msc Thesis

A deep learning approach to (semi-) automatically track bone movement in ultrasound images of patients with a unilateral transtibial prosthesis

Msc Thesis

by

Daniël Donse

to obtain the degree of
Master of Science
In Mechanical Engineering
At the
Department of Cognitive Robotics
Delft University of Technology
To be defended publicly on:
December 13, 2022, at 15:00

Thesis committee:	
Chair:	Dr. ing. L. Marchal-Crespo
Supervisors:	Dr. ir. N. Tümer
	N. Jonkergouw, Msc.
External examiner:	Dr. A. Zgonnikov
Place:	Faculty of Mechanical Engineering, Delft
Project Duration:	December 2021 - December 2022
Student number:	4369246

Faculty of Mechanical Engineering · Delft University of Technology

Delft University of Technology

Preface

The procedure to fit a prosthetic socket to a patient, which can assure the patient's comfort during activities of daily living, is labour intensive. When a slight displacement of the residual limb occurs in the socket, the socket control is considered high. More room for movement of the residual limb relative to the socket could lead to impaired control and decreased usage comfort. The demand for an automated data-driven method to measure the in-socket bone movement in patients with a unilateral transtibial prosthesis is high. This thesis tries to answer the question of whether an automated deep learning model could be of use for tracking in-socket bone displacement from Ultrasound recordings, compared to a semi-automated procedure that requires some time from a prosthetist. I conducted this research at the Delft University of Technology in collaboration with the Military Rehabilitation Centre Aardenburg (MRC) in Doorn. The academic part of this research project was done in correspondence with the Delft University of Technology. At the same time, MRC provided me with office space to work during the covid pandemic. The Ultrasound measurements and human experimental procedure were carried out at de VU, in Amsterdam, together with my daily supervisor from MRC. Without our collaboration and the expert knowledge provided by my supervisors, this project would not have been feasible. This thesis starts with a research paper describing this study's results and relevance. The appendices with tables that include metrics follow up after the references from the research paper to provide an understanding of the results obtained during the experiments.

I want to thank my supervisors at the TU Delft, Nazli and Laura, for their expertise and quick response to my questions. Nazli was always there if a pressing matter was at hand and inspired me to look at the project from a different perspective after every time we met. Laura made sure I could contact anyone with in-depth knowledge of any subject in the department of Cognitive Robotics and was open to any discussion regarding this or any other scientific study. I would also like to thank MRC for making it possible to do my internship and graduation there, especially during covid. I want to thank Niels for the time spent conducting this research and for always being available to answer any technical questions with his specialized knowledge of prosthetics.

Furthermore, I want to show my appreciation to all my close relatives for their love and support during the project. I would especially like to thank Joris for our interesting discussions at times these were needed the most. Thank you, Robert, for the countless times to pinpoint bugs and time spent explaining what I could improve. I would like to thank Michael for the birds-eye view you have given during the final stage of this project. Thank you, Tomas and Lana, for proofreading my thesis and being there for me. A word of thanks to Jesse, Friso and Myron for the musical enrichment and the interesting conversations during this phase. Furthermore, I would like to thank Leidy for the emotional support and distractions at the right time. A word of thanks can not be left out for all my housemates and previous housemates who were the best company at times it was necessary. I would also like to thank my family, who were always there for good conversations and support. Your support made the work that is in front of you possible. This is the grand finale of my studies in Delft, which would not have been possible without all of you. Thank you to the reader, I hope you enjoy reading my work.

D. Donse
Delft, December 2022

Contents

A deep learning approach to (semi-) automatically track bone movement in ultrasound images of patients with a unilateral transtibial prosthesis

D. Donse
Cognitive Robotics Dept., Faculty of 3mE, Delft University of Technology

Abstract

Background: The procedure to fit a prosthetic socket to a patient, which can assure the patient's comfort during activities of daily living, is labour intensive. Such a lengthy procedure could benefit from an automated and more efficient data-driven method capable of automatically tracking the relative movement between the patient's tibia and the prosthetic socket. To investigate such a method, we acquired in-socket bone displacement data during the physical activities of the prosthetic user. Manually tracking the location of the tibia from, e.g., B-mode (imaging) ultrasound (US) sequences might be a solution, but this is time-consuming, and the interpretation of the sequences is highly operator dependent. Therefore, an automated and efficient method to assess socket fit in US sequences is needed.

Methods: We used an existing 3D U-Net with a long short-term memory module (LSTM) and compared its ability to track a landmark location point on the tibia in US recordings by comparing the displacement and similarity in shape with data obtained from a semi-automatic single-point tracker. To evaluate the performance of the developed automated workflow, we obtained experimental data from three participants who performed three repetitive stepping tasks with their prosthetic leg in a sideways, forward, and backward motion. Three deep learning models were trained with a varying hold-out method (66% training data, 34 % test data) to test the ability to track a landmark location on the tibia in unseen data from one participant. To find the similarity of the deep learning models compared to a semi-automated single point tracker, the normalised root mean squared error (NRMSE) was calculated. We also evaluated the normalised maximum cross-correlation (NMCC) to account for the maximum similarity in displacement trajectory when a delay occurred between the true trajectory and that from the automated model. We analysed the repeatability of each step task per participant with the standard deviation from the mean tibia's landmark location trajectories.

Results: Due to the delay between the semi-automated single-point tracker and the DL model, the NRMSEs ranged between 27% and 90%. The similarity threshold (0,95) was reached for five trajectories of the tracked point on the tibia in the anterior-posterior direction, with a delay between 1,5% and 8,5% of the step duration. The similarity in the anterior-posterior direction of the tibia's landmark location trajectory was higher than that in the lateral-medial direction. The SD for all participants was around 1 mm but varied proportionally to the amount of movement observed per participant. The SD of the DL models was similar to that of the semi-automated single-point tracker.

Conclusion: We conclude that a DL model from a 3D U-Net with an LSTM module has the potential to assist prosthetists and researchers in tracking in-socket tibial bone movement in the anterior-posterior direction.

I. Introduction

Within daily practice, certified prosthetists/orthotists (CPO) must adjust the prosthesis of a unilateral transtibial amputee (UTA) due to changes in the shape and/or volume of the stump. This change in stump dimensions would result in a change in stump-socket interactions and disuse of the prosthesis due to friction or shear caused by an incorrect stump-socket pressure distribution [1]. Friction and increased pressure may cause soft tissue damage, and the user is prone to a lack of control of the prosthetic socket when the shape of the socket and that of the stump are non-complementary [2]. Consequently, an incorrect socket fit could impair the full functionality of the prosthesis. To determine the correct shape and size of the socked, CPOs use iterative external observations and questions to assess the patient's (subjective) perception of the prosthetic fit. Although a broad range of evaluation methods are available, the demand for a more efficient and data-driven approach to analyse bone displacement in the sockets during activities of daily living rises [3].

However, a standardised method for measuring the stump displacement within a transtibial prosthetic socket during activities of daily living (ADL) does not exist. Researchers have tried to develop several methods to gain insight into the kinematics inside the prosthetic socket, such as marker motion capture [4], [5], combined with musculoskeletal modelling [6], pressure detection [7], finite element models from computed tomography (CT) [8], and magnetic resonance imaging (MRI) [9]. However, these methods are unsuitable for this task since soft tissue deformation, between-layer movement, and sweat production might displace markers from marker motion models or give a distorted view of the coupling between the skeletal and prosthetic movement in other methods [3]. Additionally, using marker motion models requires undesirable prosthetic adjustments in clinical care (e.g., drilling holes in the socket). Furthermore, CT, MRI, and X-ray are not applicable in prosthetic adjustment procedures since these methods introduce a strong magnetic field or harmful radiation and, thus, are not suitable for repeatedly measuring test participants in motion [10], [11].

Therefore, we aim to measure the in-socket movement of the tibia using diagnostic ultrasound (US). US systems are relatively low-cost compared to other medical imaging methods and have no known side effects [11]. Laprè et al. designed a prosthetic socket with an opening to attain direct skin contact with the US transducer [3]. Previous studies have used ultrasound measurements to detect the location of the residual femur in a water-filled socket [11], [12]. They reported that mounting the transducers and time-consuming analysis of the ultrasound images could be an obstacle to applicability in a clinical setting. Convery and Murray successfully used two socket-mounted US transducers to analyse the motion of a residual femur of one test subject while performing several activities [11]. They found inconsistencies in the motion patterns obtained from a sequence of US images from the system recording. They attributed them to the fact that they did not measure physical activity during a simple repetitive motion of the participants [11]. To test whether intra-subject repetitive motion patterns could be observed from the tibia trajectory, we ran a pilot study to monitor bone movement from US sequences in unilateral transtibial amputees during ten repetitive step motions in one of five different directions (one direction per step task). Three of these tasks involved steps that lifted the prosthetic foot from the ground.

The location of the tibia within a US recording may be tracked by manually appointing a landmark location on each US image (e.g., a frame from the US recording). However, manually tracking the location of the tibia in US recordings is time-consuming. Furthermore, deciding which point to follow is prone to differences in interpretation of the US image between diagnosticians [13]. A reliable automated approach might address these limitations. Several algorithms exist for automatic landmark tracking trained with data labelled by surgeons and radiologists to provide a standardised, automated medical image analysis method [14]. Automation is widely used in classifying and segmenting bone structures, enabling algorithms to recognise bone landmarks and obtain a detailed bone contour from US imaging [15] [16]. Importantly, automation might be a solution to reduce labour intensity and interpretation variety of analysing US sequences. Here, we investigated the accuracy of an automated method for in-socket bone location tracking.

Automatically detecting and tracking a bone structure from US imaging involves two steps: i) **segmentation** throughout the sequence and ii) **landmark location tracking**. Machine learning (ML) and deep learning (DL) have generally shown the most promising results in several US tracking challenges [17], [18]. Most DL segmentation methods rely on convolutional neural networks (CNNs) or recurrent neural networks (RNNs) [19], [20], [21]. CNN iteratively scan an image for recognisable properties while focusing on a small region of interest that shifts after each iteration and produces a feature map of lower dimensionality. After all iterations, the original image is deconstructed into several feature maps, called encoding. A U-Net is a commonly used CNN architecture in medical image recognition, where the architecture is shaped like a 'U'. A U-net reconstructs the scanned regions into a complete image as output, termed decoding, causing it to be more receptive to surrounding features of the scanned area. U-Nets also need less training data than conventional CNNs [22]. Recurrent neural networks excel in analysing the sequential type of data, with a long short-term memory module (LSTM) in particular for video tracking [23]. Belikova et al. proved that an adapted combination of a 3D U-Net with an LSTM outperformed a standalone 3D U-Net and U-Net in tracking landmark locations in a US sequence of temporomandibular joint movement [14].

In this work, we decided to explore an existing 3D U-net with the LSTM module to track the tibia movement by tracking a landmark location on the tibia from US sequences of the residual tibial bone movement within the prosthetic socket. The 3D U-Net can learn to recognise abstract patterns in the US frames, while the LSTM could be necessary for memorising the temporal track of the tibia throughout the sequence. Furthermore, we evaluated the deep learning network tracking performance compared to the semi-automatic single-point tracker from After Effects (Adobe, United States of America (USA)). Since a model is only an approximation of reality, we assessed the relation between the in-socket tibial displacement and the participant's physical movement based on ground-truth data from semi-automatically tracking a landmark location to answer the following research question:

How does a model of a 3D U-Net with LSTM module perform compared to a semi-automated single point tracker when tracking in-socket residual bone movement from US sequences in unilateral transtibial amputees doing repetitive step tasks with their prosthetic leg?

The following sub-questions will help find the answer:

- What is the similarity in terms of the shape of the 3D U-Net with LSTM models with a semi-automated single-point tracker?
- What is the repeatability in terms of standard deviation from the mean of the results from the semi-automated single-point tracker and the automated models?

II. Methodology

A scheme of the general methodology followed in this work can be found in figure 1.

A. Experimental setup

The experiments were carried out with an HM70A US system (Samsung Electronics, South Korea) with an LA3-16AD transducer (see Appendix A). The US images were recorded by connecting a US probe horizontally (marked side pointing in the lateral direction) to a custom-designed prosthetic socket made of Thermolyn Clear (OttoBock, Germany) with a Limblogic active vacuum pump (WillowWood, USA) for each test participant, see Appendix A for dimensions. The 60-second adhesive (Fabtech Systems LLC, USA) held the probe in place, and all gaps between the probe, prosthetic socket and stump were filled with US transmission gel, see figure 2. At the location of the probe connection, the Alpha Duo liner (WillowWood, USA) had a hole. The base configuration of the prosthetic foot alignment with the socket was determined by observational gait analysis to ensure that the participants were walking stable and even.

The test procedure was also recorded with an Ipad 2 and iPhone X (Apple, USA) that recorded the participants' movements at 60 frames per second (fps).

B. Participants

We recorded US videos of three male participants with unilateral transtibial amputation. All participants provided written consent before the start of the study (ethical approval was acquired from the Medical Ethical Review Board of the University Medical Centre Groningen, Groningen, the Netherlands (NL74038.042.21)). Participant one has a left-sided amputation, unlike the other two participants with a right-sided amputation. Participant one was 51 years old, 84 kg, 1.82 m, participant two was 50 years old, 110 kg, 1.92 m, and participant three was 47 years old, 86 kg, 1.93 m.

C. Human experimental procedure

While wearing the test prosthesis, the participants performed step tasks of 10 step repetitions each. In total, they performed five step tasks:

1) weight shifting from one leg to the other,
2) steps to the right side,
3) steps to the left side,
4) steps forward, and

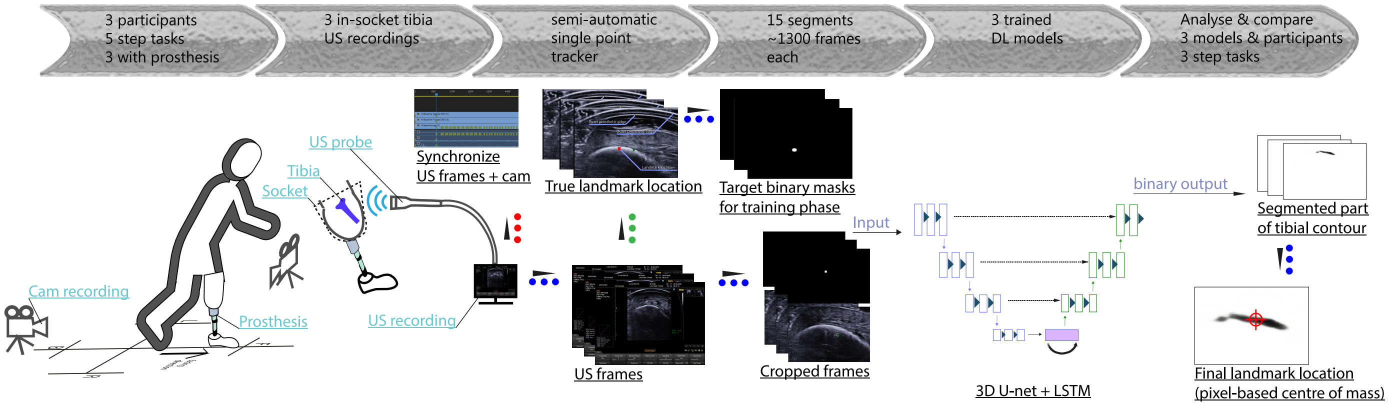


Figure 1. General methodology from physical experiment to data analysis. Red dots = Adobe Premiere Pro (Adobe, USA), Green dots = Adobe After effects (Adobe, USA), Blue dots = Python (Python software foundation, USA) processing

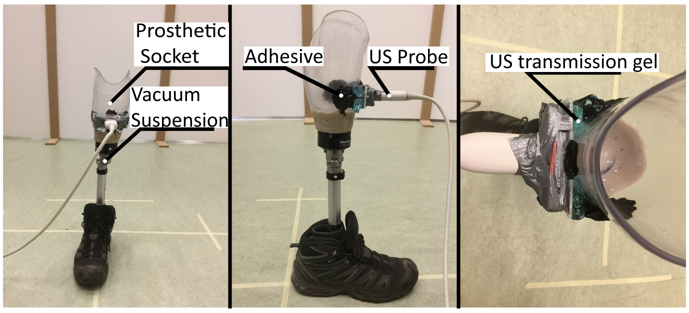


Figure 2. Connection of US probe to the custom-designed prosthesis.

5) steps backwards.

A metronome regulated the step frequency at 100 beats per minute (bpm), and the CPO gave a short demonstration of the step task. A cross on the ground indicated the direction (forward, backward, and sideways) and recommended length of each step, see figure 3. The desired size of the steps was calculated per participant based on their body height. The forward and sideways steps had a length of 33 % of body height, while the backward step was half the length of the other directions. During each directional task, one leg was static on the ground. In forward motion, one step entails; starting stance in the initial position (figure 3, right, A), middle: one step forward (figure 3, right, B) and end: back at the initial location (figure 3, right, C).

All data from the five tasks are used for training the DL model; however, since the focus of this research is on monitoring bone movement while the prosthetic foot is in motion, we will focus the analysis of the similarity between the semi-automated- and DL tracker on the following three step tasks:

- Sideways
  - Start: standing on two feet initial position in the middle of the cross from figure 3
  - middle: one step sideways with the prosthesis
  - end: back at the initial position (standing on two feet in the middle of the cross from figure 3)
- Forward
  - Start: standing on two feet initial position in the middle of the cross from figure 3
  - middle: one step forward with the prosthesis
  - end: back at the initial position (standing on two feet in the middle of the cross from figure 3)
- Backwards
  - Start: standing on two feet initial position in the middle of the cross from figure 3
  - middle: one step backwards with the prosthesis
  - end: back at the initial position (standing on two feet in the middle of the cross from figure 3)

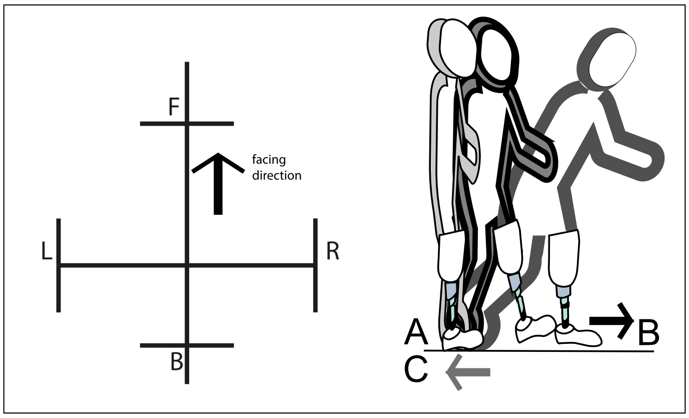


Figure 3. Left: cross indicating desired step length per direction (F= forward, L = left, B = backwards, R = right). Right: Representation of one step from steps forward procedure with the prosthesis. Initial position (A) is in the middle of the cross; one loading response moves the prosthetic foot to (B) and then returns to the initial position (C) in the middle of the cross. The black arrow represents the movement direction of the prosthetic foot from A to B, and the grey arrow is the movement direction from B to C. The end pose is back at the exact location as the start pose with two feet parallel on the ground pointing forward.

D. Video recordings

The test procedure was also recorded by an Ipad 2 and an iPhone X (Apple, USA), each placed approximately three meters from the initial position of the participant. The camera on the side captured the physical movement of the participant in the anterior-posterior (frontal-backside) direction, and the camera on the front- or back-side recorded the physical movement in the lateral-medial (outside-inside) direction. The cameras captured the lower limbs of each participant, so we could capture the moment of loading response and push back to the initial rest position. We also recorded the moment of starting the US recording, which was synchronised with that moment in actual time in the videos captured by the two cameras and compared per set of recordings in Premiere Pro (Adobe, USA) for each participant, see figure 4.

We trimmed the recording, so the duration was equal to the procedures from start to end of each step task of 10 repetitions; please see figure 4. Every instance in which the participant was in starting pose, stance and end pose for a step was marked so that the frame number could be related to the frame number in the US data set corresponding to that instance in time.

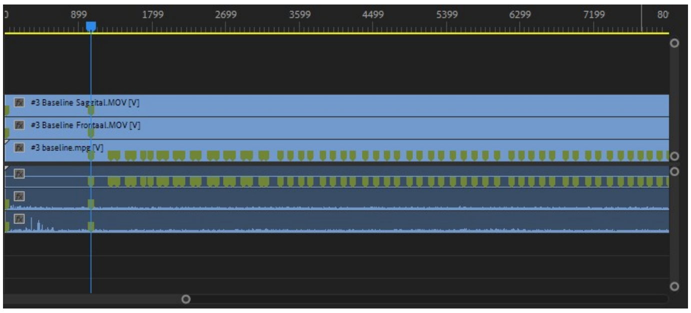


Figure 4. Timeline synchronisation of the start of videos from two cameras and US recording with marked points for the start, middle and end of each step

E. Semi-automatic single-point tracking

A CPO appointed a landmark location on the tibia contour for each unedited US frame, shown in figure 5 left and middle, with a semi-automatic single-point tracker in After Effects (Adobe, USA). Incorrect instances of the tracked point, located by the semi-automatic single-point tracker, were detected and corrected by the CPO. These landmark locations are the basis of the target images used as references for the automatic landmark tracking algorithm. The target images were constructed as a matrix in Python (Python Software Foundation, USA). Each value in the matrix corresponds to a pixel, where white pixels have a value of one and black pixels are valued at zero, and the target matrices are called binary masks. In the binary masks we made as a matrix in Python with a size equal to the US frames (1024 x 768), a white pixel corresponds to a pixel at the desired landmark location. These masks were used as the targets for the automatic tracking model. The targets are depicted as white dots with a radius of 5 pixels, which covers the thickness of the tibia at the landmark location on a black background, see the right side of figure 5.

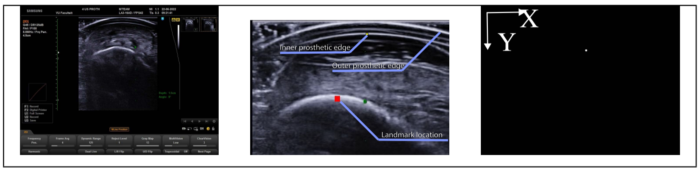


Figure 5. Left: Unedited US frame. Middle: Cropped US frame with landmark location (red square) and an indication of prosthetic edges. Right: Binary mask with a white dot functioning as target label for the automatic tracking detection model, including the X-Y coordinate system used for calculating the pixel-based centre of mass (positive x-direction = lateral direction, positive y-direction = posterior direction)

The total US data set contained 19912 frames obtained from all three participants, from all step tasks, for ten repetitions per task. The participants followed the experimental procedure in five different directions each. Therefore, the data was split into 15 raw US segments representing each of the five motion patterns of all participants combined. The target data, constructed as binary masks, contained an equal amount of frames as the US data set. The target data was split into the same segments as well.

F. Automatic landmark tracking algorithm

1. Data preprocessing

The training of the DL models was done on Ubuntu 18.04, running an NVIDIA GeForce RTX 2080 Ti GPU (11GB RAM). All unedited US frames were converted to a grey scale matrix (each pixel represents a value between 0 and 255 dependent on the intensity of the pixel; 0 is black, and 255 is white) using OpenCV (Intel, USA) Python (Python Software Foundation, USA). The original US frames had a size of 1024 x 768 pixels. They were automatically cropped to decrease the processing time and limit the amount of surrounding noise around the contour of the tibia based on the area of movement observed during semi-automatically tracking of the landmark location on the tibia. The cut frames had a size of 360 x 240 pixels, with the top left at 119 pixels from the top and 281 pixels from the left of the original US frame.

The ground-truth binary masks (an additional 19912 frames) were cropped to the same dimensions as the input frames, so each location of a value in the binary mask corresponds to the exact location in the input matrix. It took 2 minutes and 22 seconds to crop all 39824 images (19912 US frames and 19912 corresponding binary masks) and divide the data into 30 separate segments representing separate step tasks (15 cropped US segments and 15 cropped binary masks). The cropped US frames are used as input to the DL network, and the cropped binary masks function as target images during training, which the network should reproduce.

2. Deep learning network

We used an existing 3D U-Net LSTM module [14], depicted in figure 7, for training the segmentation models. The network consists of three main parts: the convolutional encoding and decoding units and the LSTM module in the middle. The basic convolution operations are performed, followed by a rectified linear unit (ReLU) activation and batch normalisation in the encoding and decoding parts of the network. These operations ensure that each feature map consists of binary values indicating the segmented values with a value of one and make the network more stable during training [24]. For down-sampling in the encoding unit, 1x2×2 max-pooling operations are performed to find the maximum values for each patch, producing feature maps with half the dimension of the input sample.

The values in the LSTM module pass through several gates (input, output, and forget gate) and overwrite the internal states (cell and hidden state) of the LSTM cell [25]. The cell can forget irrelevant information, which is controlled by the gates and determines which information is passed from the input to the output. Static weight matrices W and U for the input and hidden state, respectively, along with bias vectors b, are used in each of these gates, constructed during the training phase. For each time step, these weights and biases remain equal. This results in the following equations in an LSTM cell for one time step t:

$$i_t = \sigma_r \left(W_i \cdot X_t + U_i \cdot h_{t-1} + b_i\right) \tag{1}$$

$$f_t = \sigma_r \left(W_f \cdot X_t + U_f \cdot h_{t-1} + b_f\right) \tag{2}$$

$$c'_t = \sigma_a \left(W_c \cdot X_t + U_c \cdot h_{t-1} + b_c\right) \tag{3}$$

$$O_t = \sigma_r \left(W_o \cdot X_t + U_o \cdot h_{t-1} + b_o\right) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot c_t \tag{5}$$

$$h_t = O_t \odot \sigma_a \left(c_t\right) \tag{6}$$

where i is the input gate, f the forget gate, c' the cell gate (c_t = current state, c_{t-1} = previous state), o the output gate, c the cell state, h the hidden state, σ_a the activation for the states, and σ_r the recurrent activation for the gates, where $\odot$ represents an element-wise multiplication.

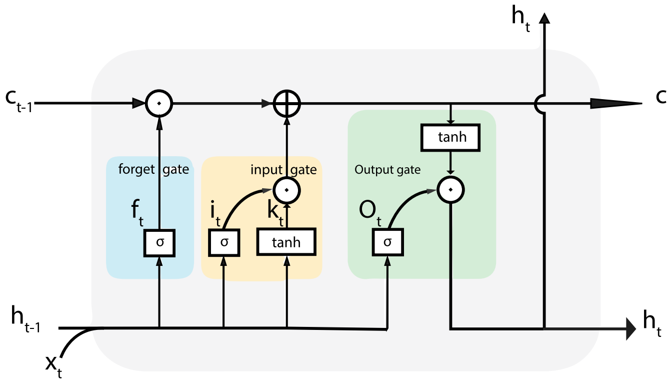


Figure 6. LSTM cell for one time step t; i is the input gate, f the forget gate, c_t = current cell state, c_{t-1}= previous state, o the output gate, c the cell state, h the hidden state, σ the activation for the states, and tanh the recurrent activation for the gates, where $\odot$ represents an element-wise multiplication and $\oplus$ a sum, retrieved and adjusted from [26]

In the decoding phase, 1x2x2 up-convolution operations are performed to up-sample the feature maps, extrapolating the maximum values over the whole dimension of each patch. Each up-convolution is concatenated with a sample from the same-sized max pooling layer. The network's final output produces binary masks with a segmented contour of the tibia. We used an Adam optimiser with a weight decay regularisation parameter of 0.0005 and a learning rate of 0.00001; these were determined iteratively before training these models. For all chosen hyperparameters, we referred to a previous study on tracking bone in US sequences [14], see Appendix A.

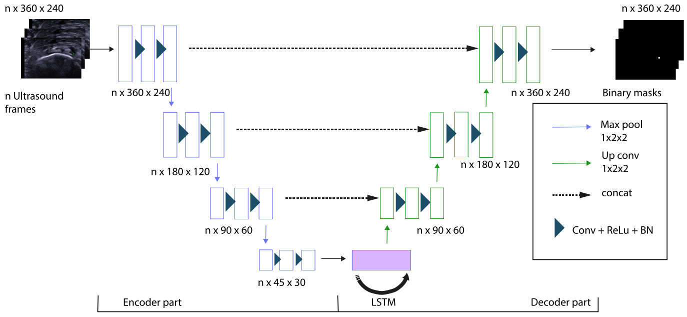


Figure 7. The deep learning architecture of the 3D U-Net with an LSTM (long short-term memory module), where n is the number of frames, Max pool = max pooling, up conv = up convolution, concat = concatenation, conv = convolution, ReLu = rectified linear unit and BN = Batch normalisation. Adapted from [14]

3. Training, validation, and evaluation procedure

We trained the DL network in three different configurations (models). In this training procedure, all 15 segments of 1300 frames on average were split into sequences of eight frames since previous literature used the same sequence length [14]. The models processed input data with dimensions $n \times 360 \times 240$ pixels, as shown in figure 8, where n is the number of frames per sequence. The target data had the same dimensions as the input data and were used as a reference for the output of the models.

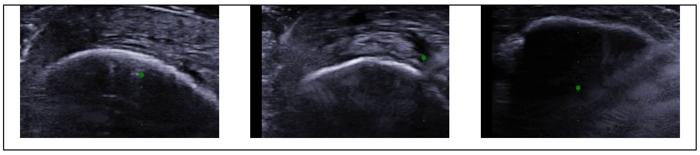


Figure 8. From left to right: cropped frame of data from participants one, two, and three used as input for the DL algorithm

Each model used a different train-test split; a hold-out test only fed the model with data it had never processed during training to test the actual performance and prevent data leakage from the train to the test set. This hold-out test (66% training data, 34% test data) was performed with the following configurations for the test and training sets, please see figure 9:

- Model one
  - Training set: segments six to 15 (from participants two and three)
  - Test set: segments one to five (from participant one)
- Model two
  - Training set: segments one to five and 11 to 15 (from participants one and three)
  - Test set: segments six to 10 (from participant two)
- Model three
  - Training set: segments one to 10 (from participants one and two)
  - Test set: segments 11 to 15 (from participant three)

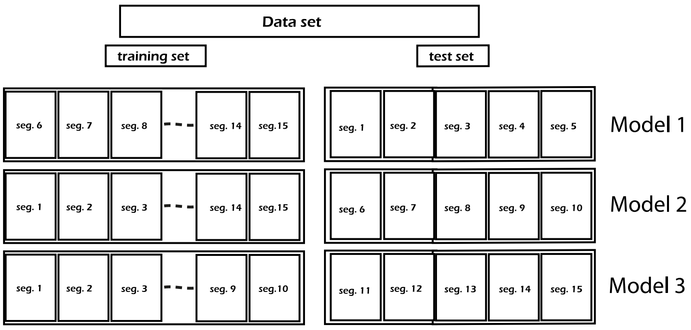


Figure 9. Training/test procedure per model with a varying train-test split of all segments in the data set

Several image transformations were performed; a slight random rotation, flipping relative to the timeline by shifting half of the n images in a sequence, and random contrast as data augmentation techniques to minimise the overfitting problem with a limited training set. The amount of data used as input to the model remained the same while the orientation, contrast and rotation were randomly applied to increase the variance of the input data. We flipped the images randomly with Numpy.flip in Python over the timeline with an execution probability of 0.5, thus, there is a 50% chance of each input sequence being flipped. The images were rotated randomly over an angle spectrum of five degrees positive to five degrees negative rotation with Scipy.ndimage.rotate (Enthought, USA). We changed the contrast of the input images by a factor valued between one and two with a 50% chance of execution with RandomContrast, for any pixel x in the channel, the output will be;

$(x - mean)*$ factor $+mean$ where $mean$ is the mean value of the channel [27].

The networks were trained for 100 epochs during training before the hold-out test. As a loss function, we use Binary Cross-Entropy (BCE) [28], shown in the equation below:

$$L_{BCE}(y_t, y_m) = -(y_t \log(y_m) + (1 - y_t) \log(1 - y_m)) \quad (7)$$

where L_{BCE} is the loss function, y_m is the predicted value produced by the model, whereas y_t is the target value for the landmark location. The target value can either be one (corresponding to $y_t \log(y_m)$) or zero (corresponding to $(1 - y_t)\log(1 - y_m)$)). A value of one indicates a white pixel corresponding to the desired landmark location, while a zero indicates a black pixel. If each predicted value is the same as the value in the target binary mask, the loss becomes zero, so the algorithm is trained to get the loss function as close to zero as possible. After softmax activation of the log probabilities, the loss values are scaled between zero and one. Since we had limited data available and wanted to avoid leakage of training or validation data into the test set, we excluded the validation phase. We used the test data to evaluate each model's accuracy in predicting the final landmark location and similarity in terms of shape with the semi-automated single-point tracker for all step tasks with the prosthesis. We calculated the final landmark location during postprocessing since the DL architecture was not designed to produce landmark location coordinates as output. Therefore, we could not make a valid comparison between training loss and test results. Consequently, we did not include the loss values in the results.

G. Postprocessing of the output from the deep learning network

After obtaining a segmentation of the landmark location area (see figure 10) with the models, automatic postprocessing was performed, including the Otsu thresholding method from OpenCV (Intel, USA) Python (Python software foundation, USA) to filter any remaining noise from the background of the models' outputs [29]. The Otsu classification-based binarisation method searches for the threshold that minimises the intra-class variance, defined as a weighted sum of variances of the two classes (background and foreground). It assumes the image consists of only the object (foreground) and background, and the heterogeneity and diversity of the scene are ignored [29].

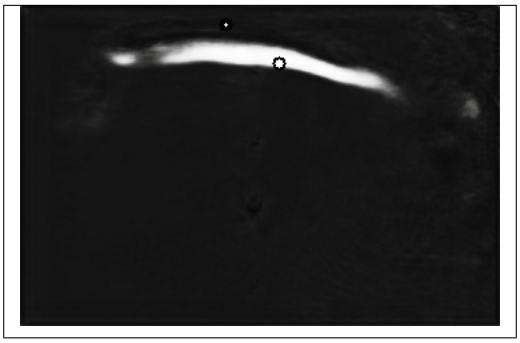


Figure 10. One frame from the output of model three with a segmentation of the landmark location area (depicted in white) of the tibia, used to determine the final landmark location (white dot with a black lining)

To obtain the final landmark location, the segmented contour of the tibia is detected with OpenCV Python, and the final landmark location was computed as a pixel-based weighted centre of mass for the predicted output [30] using equations:

$$C_y = \frac{M_{10}}{M_{00}} \qquad C_x = \frac{M_{01}}{M_{00}} \tag{8}$$

where $\mathrm{C_x}$ is the x coordinate and $\mathrm{C_y}$ is the y-coordinate of the centroid in the coordinate system depicted in figure 5, and M denotes the image moment:

$$M_{ij} = \sum_x \sum_y y^i x^j I(y,x) \tag{9}$$

where I(x,y) is the pixel intensity, with x and y the x- and y-coordinate for that pixel in the segmented contour of the tibia. The summation extends over all the elements on that axis in the US frame. Here M_{00} is the total amount of pixels in the segmented part of the tibia, M_{10} the total value of the intensity for all pixels in the y-axis multiplied by the index values, and M_{01} is the total value of the intensity of all pixels summed over the x-axis multiplied by the index values.

Outliers were removed by calculating the z-score of all coordinate values (separated in x- and y-direction) per step of each step task. The values with a z-score of three or higher were removed and accounted for 1% of the data from the predicted landmark locations.

H. Mean landmark location and trajectory calculation

The displacement of the landmark location during each step was split into anterior-posterior and lateral-medial translations to analyse the models' accuracy. We converted the coordinate values from pixels to mm by using the scale used in each recording; 350 pixels accounted for 35 mm in the recordings from participants two and three, while the scale for participant one was 350 pixels for 45 mm, please see table 1.

Participant	Scale
One	$\frac{45}{35} \frac{pixels}{mm}$
Two	$10 \frac{pixels}{mm}$
Three	$10 \frac{pixels}{mm}$

Table 1. Conversion table for pixels to mm for the US image data per participant

Since the recorded steps varied in duration, we calculated a step percentage from the start at 0% (point A in figure 3) to the end of the step at 100% (point C in figure 3). We used the step percentage to calculate the mean amplitude (with reference point: the initial position of the tibia's landmark location point at the start of each step) of the landmark location on the tibia for each step task. To calculate the mean amplitude at each step percentage, each measurement per step needed an equal amount of data points. We collected a set of discrete data points with the semi-automatic motion tracker with unequal data size per step (ranging from 101 to 196 data points per step). Since a curved trajectory best represents a physical movement pattern of the tibia's landmark location, we used the second-order B-spline interpolation from Cox and de Boor [31], [32], according to the following equation:

$$\mathrm{B}_{i,2}(t) = \begin{cases} \frac{1}{2}(t - t_i)^2 & t_i \leq t < t_{i+1} \\ \frac{1}{2}((t - t_i)(t_{i+2} - t) + (t - t_{i+1})(t_{i+3} - t)) & t_{i+1} \leq t < t_{i+2} \\ \frac{1}{2}(t_{i+3} - t)^2 & t_{i+2} \leq t < t_{i+3} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

where $B_{i,2}(t)$ is the i^{th} second order b-spline at point t (a coordinate in either x- or y-direction), over equally spaced knots t_i in the range t_0 to t_{199}, since all steps had a duration between 101 and 196 frames. The space between all knots of the b-spline was 0,5% of the step percentage so that the mean translation could be related to the step percentage from 0% to 100%.

During the task, all participants used each direction's first and last repetitions to accommodate a different position, leading to a deviation in their movement pattern. Therefore, we excluded these steps from the analysis.

I. Analysis of the automatic tibia's landmark location trajectory

1. Similarity of tracking methods

To evaluate the absolute tracking error between the semi-automatic single point tracker and the DL models' automatic detection, we calculated the Euclidian distance between the (x,y) coordinate location of the predicted landmark location and the true location per step task.

The results from the semi-automatic single-point tracker, referred to as the ground truth, were compared to those obtained from the automatic landmark tracking algorithm models. The similarity between the two was computed using normalised root mean square errors (NRMSE) between all true points obtained directly from the semi-automated tracker and the automated landmark location tracker [33]:

$$NRMSE = \frac{\sqrt{\sum_{i=1}^{n} \frac{\left(y_{pr,i} - y_{gt,i}\right)^2}{n}}}{y_{pr,max} - y_{pr,min}} \times 100\% \qquad (11)$$

where n is the number of frames per step, $y_{pr,i}$ is the predicted value of the landmark location in one of two directions (y-direction (anterior-posterior) or x-direction (lateral-medial)) at time i, and $y_{gt,i}$ is the ground truth value at step percentage i, $y_{pr,max}$ = maximum amplitude of the landmark location point on the tibia of the ground truth data during each step, $y_{pr,min}$ = minimum amplitude of the ground truth data.

Evaluation of the shape similarity of the results from the mean of the models and the ground truth data was performed with a time shift invariant method: the normalised maximum cross-correlation (NMCC) [33]:

$$NMCC = \frac{\max |(f \star g)[i]|}{\sqrt{\sum_{i=1}^{n} f[i]^2} \cdot \sqrt{\sum_{i=1}^{n} g[i]^2}} \qquad (12)$$

where f and g are functions with a length of n, per step percentage i and $(f \star g)[i]$ is the cross-correlation of functions f and g that represent the mean tibia's landmark location trajectory of the semi-automated single point tracker and the DL model's automatic detection. The mean trajectory of the tibia's landmark location estimated by the models was assumed to represent that of the ground truth sufficiently if the calculated NMCC was within a margin of 5% (*i.e.*, NMCC $\geq$ 0.95) [33].

A delay was observed between the ground truth and predicted trajectories of the tibia's landmark location and estimated from the step percentage value where the NMCC was maximal.

2. Repeatability of the semi-automatic single-point tracker and deep learning models

The repeatability was evaluated by calculating the standard deviations from the mean of the tibia's landmark location trajectory of the semi-automatic single-point tracker and deep learning models. A higher SD indicates lower repeatability and vice versa. We calculated the average of the standard deviation (SD) from the mean of the tibia's landmark location trajectory per step task and direction and interquartile range of the SD to find the repeatability of all landmark location trajectories between steps per participant for each step task for both the ground-truth results from the landmark locations and those obtained from the models.

III. Results

A. Segmentation of the landmark location area from the DL model

The DL algorithms successfully segmented the data in general; please see figure 11.

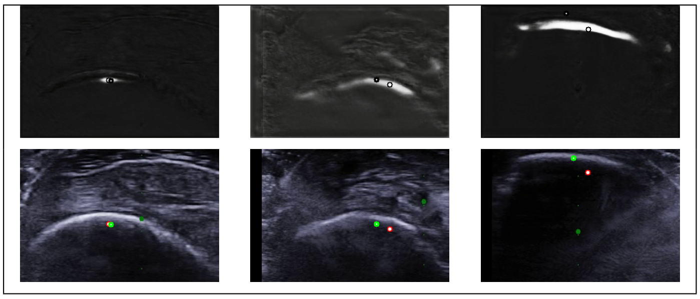


Figure 11. Three frames from the DL model's output segmented part of the tibia's contour (upper) and the cropped input US frames (lower) with landmark locations from the semi-automated (black dot (upper)/ green dot (lower)) and DL model's (white dot with black lining (upper)/ red dot with white center (lower)) output; from left to right: model one, model two and model three

B. Similarity and NRMSE of the tibia's landmark location trajectories between semi-automatic and automatic methods

The results corresponding to the output from the tibia's landmark location trajectories of the semi-automatic single-point tracker for the three participants (referred to as participant one, participant two, and participant three) and the automatic deep learning models (referred to as model one, model two, and model three) were compared for all step tasks involving steps with the prosthetic foot. A boxplot of the absolute tracking error per step task can be found in figure 12. The NRMSE, delay, and NMCC values can be found in table 2. The translations are shown in figure 13 as the mean amplitude of the landmark point on the tibia from all steps per participant per step task:

- Sideways task
  - Anterior-posterior translation (figure 13A)
  - Lateral-medial translation figure 13B)
- Forward task
  - Anterior-posterior translation (figure 13C)
  - Lateral-medial translation figure 13D)
- Backward task
  - Anterior-posterior translation (figure 13E)
  - Lateral-medial translation figure 13F)

The average standard deviation and its interquartile range from the mean tibia's landmark location trajectory for the sideways, forward and backward tasks can be found in table 3, 4, and 5, respectively.

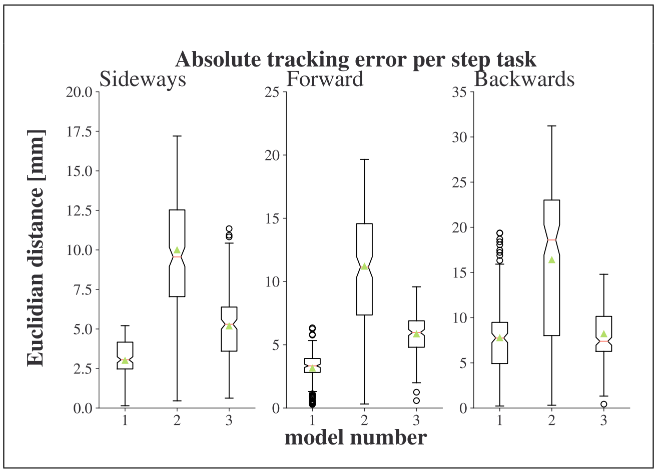


Figure 12. The Euclidian distance in [mm] between the predicted location by each model and the true location obtained with a semi-automated single-point tracker per step task, the orange line is the median and the green triangle points at the mean value for each model per task

C. Similarity between semi-automatic single point tracker and deep learning models

For all models, a significant delay of around 2,0% to 13,0% on average can be observed in the tibia's landmark location trajectory compared to the semi-automatic single-point tracker results. This is also reflected in the NRMSE values, ranging from 27,7% to 77,3% for the anterior-posterior direction and from 38,9% and 90.0% in the lateral-medial trajectory. The maximum NRMSE of model one (i.e. 90%) is found for the prediction of the lateral-medial translation during the forward task. The mean approximation and the standard deviation of model one do not overlap with that of participant one during the translation of the tibia's landmark location (figure 13 D). For model two, the maximum NRMSE (84.9%) is found for the lateral-medial translation of the backward task (figure 13 F). In that trajectory, the true translation of the tibia's landmark location trajectory shows the most extensive amplitude changes in the step percentage intervals of 15% to 40% and 70% to 80%, ranging from 0 mm to 9 mm. Compared to all of its other trajectories, model three shows the highest NRMSE (i.e. 77.3%) for approximating the anterior-posterior course of the tibia's landmark location for the sideways task, where the amplitude changes are the smallest (0 mm to 1.5 mm) compared to the other trajectories observed in participant three.

1. Model one

The trajectory of the anterior-posterior translation of the landmark location on the tibia of model one does represent the shape of the semi-automatically tracked course for participant one during the sideways- and forward tasks (figures 13 A and C). The NMCC values of 0.99 confirm this observation. The trajectory for the backward task in the anterior-posterior direction shows a large phase shift and significant differences in shape, verified by the NRMSE (58,4%) and insufficient NMCC value (0.91).

In the lateral-medial direction, the approximation of the ground truth is less accurate, resulting in a sufficient NMCC of 0.97 for the forward step task but with a high NRMSE of 90%, insufficient NMCC of 0.89 for the sideways and 0.22 for the backward task.

2. Model two

The tibia's landmark location trajectories approximated by model two do not reach the NMCC threshold.

The highest amplitude changes (+4 mm to -4 mm within 15% of the step) of the tibia's anterior-posterior landmark location for the semi-automatic single-point tracker were observed in participant two during the forward and backward tasks (figure 13 C and E). The model reached an NMCC of 0.93 for the forward task and 0.9 for the backward task. For all trajectories except the sideways, anterior-posterior translation of the tibia's landmark location, participant two showed the highest amplitude and changes in amplitude. For the sideways task, model two did not approach the shape and trajectory of the anterior-posterior translation trajectory for participant two well and reached an NMCC of 0.71.

The lateral-medial translation of the landmark location on the tibia for the backward task of participant two shows the most extensive changes in amplitude of all measurements (± 4 mm change within 2% of the step). Model two cannot track this trajectory in shape and similarity, which is reflected in the NRMSE (i.e. 84,9%) and NMCC (i.e. 0.22) values.

3. Model three

During the approximation of the anterior-posterior trajectory of the tibia's landmark location for the sideways task, the NRMSE was the highest, where the amplitude changes were the smallest compared to the other trajectories observed in participant three. Still, the NMCC was within the margin, indicating a good representation of the ground truth. Model three showed a relatively similar shape of the tibia's landmark location trajectory for participant three in the anterior-posterior direction for all tasks and reached an NMCC of 0.95 for the sideways task and 0.99 for the forward- and backward step-task. In the lateral-medial direction, the approximation of the tibia's landmark location trajectory was less accurate, resulting in NMCC ranging between 0.31 and 0.78.

Participant three showed the smallest amplitude (up to 4 mm) and changes in amplitude (up to 1 mm per %)of the tibia's landmark location trajectory obtained from the semi-automatic single-point tracker compared to that of the other two participants but also stayed within the NMCC margin for all anterior-posterior trajectories (higher than 0.95). Participant two showed the largest amplitude (+10 mm and -10 mm) and changes in amplitude (up to 6 mm per %), and model two did not reach the NMCC threshold for any of the trajectories. The tibia's landmark location trajectory tracked by the semi-automatic single-point tracker of participant one reached a slightly larger amplitude (up to 5 mm) than that of participant three (up to 4 mm) and a somewhat smaller amplitude than participant two (up to 10 mm). Compared to the other tasks, participant one showed the largest amplitude (up to 5 mm) and changes in amplitude (up to 2 mm per %) in the anterior-posterior direction during the backward step task, where the NMCC (i.e. 0.91) did not reach the threshold. Participant One showed a slightly smaller amplitude for the sideways and forward tasks (up to 3 mm). In contrast, the NMCC values (i.e. 0.99 for both) showed a sufficient representation of the shape of the actual trajectory.

D. Repeatability of the semi-automatic single-point tracker and deep learning models

The interquartile range of the SD shows a more extensive range when the SD (from the mean of the tibia's landmark location trajectory per step task and direction) is higher for all trajectories.

1. Sideways task

For the true trajectory of participant one, the SD values are around 0.8 mm and comparable for both directions (anterior-posterior and lateral-medial). In the lateral-medial direction, participant two's

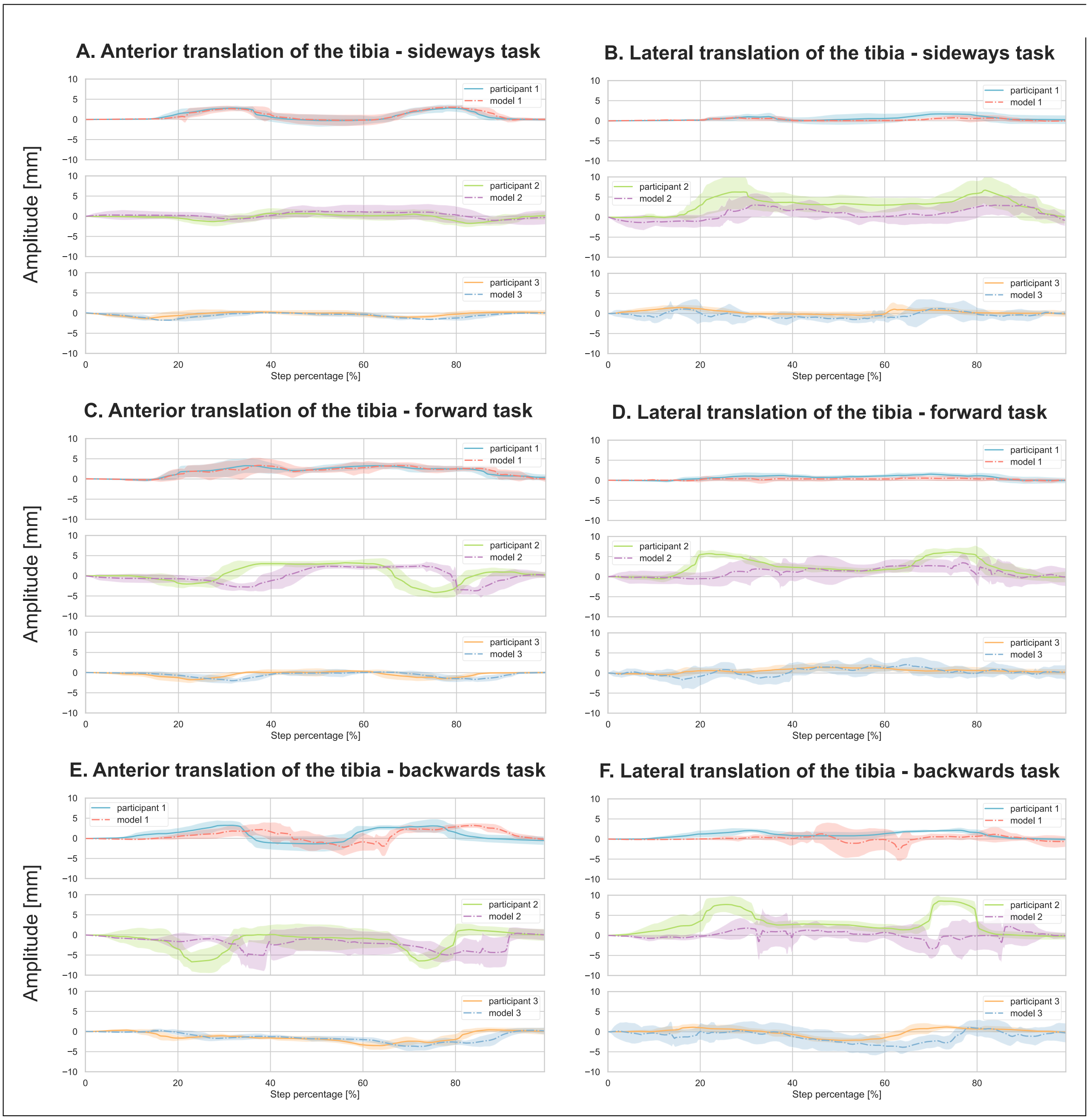


Figure 13. The mean trajectories of the Anterior (A, C, and E) and Lateral (B, D, and F) translation of the tibia w.r.t the starting pose in mm over the step percentage, where 0% is the start of the step, and 100% is the end of each step. A negative amplitude in the anterior direction = positive amplitude in the posterior direction, and a negative amplitude in the lateral direction = a positive amplitude in the medial direction.

SD (2.34 mm) is more than twice as high as in the anterior-posterior direction. The SD (0.78 mm) in the lateral-medial direction for participant three is about 30% higher than in the anterior-posterior direction (0.62 mm).

The SD for participant two is the highest, while that of participant three is the lowest in both directions for the true landmark location trajectory. The SD for most models is around the same order of magnitude as that of the true trajectories per participant, except for the lateral-medial trajectory of model three, which shows a higher SD than the ground truth.

2. Forward task

For participant one, the SD is about 40% higher in the anterior-posterior direction (0.98 mm) than in the lateral-medial direction (0.58 mm). The SD of participant two shows a similar value for both directions (±1.20 mm). In contrast, participant three's SD is about 30% higher for the lateral-medial direction (0.78 mm) than for the anterior-posterior direction (0.62 mm).

The SD for participant two is the highest, while that of participant three is the lowest in the anterior-posterior direction for the true landmark location trajectory. In the lateral-medial direction, the SD for participant two is the highest, while that of participant one is the lowest. The SD for most models is around the same order of magnitude as that of the true trajectories per participant, except for the lateral-medial trajectory of model three, which shows a higher SD than the ground truth obtained from the semi-automatic single-point tracker.

3. Backwards task

During the backwards task, the SD for participant one is about twice as high in the anterior-posterior direction (1.24 mm) as in the lateral-medial direction (0.61 mm). Participant two also shows a higher SD in the anterior-posterior direction (2.43 mm) than in the lateral-medial direction (1.71 mm), a difference of around 40%. Participant three shows a similar SD value for both directions (0.71 mm and 0.75 mm) of the tibia's landmark location trajectory. Similar to the SD values of the forward step task, the SD for participant two is the highest, while that of participant three is the lowest in the anterior-posterior direction for the true landmark location trajectory. In the lateral-medial direction, the SD for participant two is the highest, while that of participant one is the lowest. The SD for most models is around the same order of magnitude as that of the true trajectories per participant anterior-posterior direction. In the lateral-medial direction, the SD values of the models are higher than those from the true trajectories.

IV. Discussion

A DL network was trained in three configurations by splitting the data into several train- and test batches. For each model, the data from two participants were used for training, and the remaining data were held out for the test phase. We tested the corresponding models' ability to track a landmark location on the tibia in sequences of US images with similarity measures of the true shape and trajectory. The displacement of a landmark location assigned to the tibia was observed with a semi-automatic single-point tracker for all participants and used to validate the displacement detected by the DL models. All displacements are decomposed into anterior-posterior and lateral-medial translations of the tracked point on the tibia to analyse the similarity (with NRMSE and NMCC) and standard deviations of each model per participant.

A. Segmentation of the landmark location area from the DL model

All segmented landmark location areas in figure 11have a different shape. Especially the width of each body varies, while the thickness of the segmented regions is similar. Since we determined the landmark location point by calculating the centre of mass of the landmark location area, a variance in the width of the segmented part of the tibia could lead to a variance in the lateral-medial translation of the final landmark location (figure 11, right side). A curved shape of the segmented area could also lead to a displacement of the final landmark location in the anterior-posterior direction (figure 11, middle). If the final step of determining the landmark location area were incorporated differently in the DL model, the model could eventually learn to correct the location where necessary [34]. A possibility would be to design a different DL architecture to produce the final landmark locations as output, as done in corn kernel detection [35]. Hence, the model learns to specify the correct coordinates as output.

B. Similarity between semi-automatic single point tracker and deep learning models

1. Anterior-posterior translation of the landmark location point

All models showed a delay in the mean landmark location trajectory (phase shift) compared to the results from the semi-automatic single-point tracker for all participants, resulting in an NRMSE higher than 27% for all models.

However, NMCC values do not account for phase shift. They show a sufficient approximation within a margin of 5% similarity in shape (i.e. NMCC $\geq$ 0.95) for some of the models in the anterior-posterior direction. Model one represents the mean of the actual trajectory of participant one sufficiently for the sideways and forward task with a phase shift of 2,0% and 1,5%, respectively. Model three shows a sufficient representation for all step tasks in the anterior-posterior direction with a delay between 5,5% and 8,5%, while model two does not reach the NMCC threshold.

The delays of the models that reach the similarity threshold are between 1,5% and 8,5% compared to the semi-automatic single-point tracker. On average, the delay of the models that reach the similarity threshold is 4,9%. Since the upper- and lower bound of these delays differ by 7%, shifting the model's results by 5% would not be a robust solution. A modification of the training procedure or architecture would be a better solution (e.g., training with a more extensive data set, using k-fold cross-validation or dividing the data into smaller segments).

2. Lateral-medial translation of the landmark location point

The only lateral-medial trajectory that reaches the NMCC threshold is from model one's approximation of the forward step task of participant one. The course of the tibia's landmark location trajectory from model one is relatively flat, while the true trajectory also shows a flat shape. However, the mean trajectory does not overlap with the true trajectory for participant one, resulting in the highest NRMSE of 90%. All models showed significant deviations from the ground truth trajectory for the remaining mean trajectories of the tibia's landmark location in the lateral-medial direction.

3. Train- and test procedure

Since model one was trained on the true trajectories of participants two and three, it was trained on small and large translations of the tibia's landmark location observed in the US recordings of participants two and three. Model two was trained on more minor translations (those of participants one and three) than the translations observed in the actual trajectory of the tibia's landmark location for participant two. Finally, model three was trained on the data segments from participants one and two, which showed larger amplitudes and changes in amplitude than the data model three processed for testing. A validation run was not performed to check the loss development per epoch. Investigating this could provide helpful insights into the model's performance, and a distinction could be made whether it is overfitting or underfitting the data. The optimal weights could then be obtained when the model's validation loss is at its minimum [36]. Since a validation run was not performed due to the limited availability of data, it is dangerous to derive conclusions from these observations. However, a recommendation for further research can be given.

Two possible explanations can be given for the errors in the results of the DL models compared to those from the semi-automated single-point tracker;

1) Covariate shift; a difference in input data distribution could have caused the model to perform poorly when tested on data from participant two [37]. The data used for the training procedure of models one and three contained a large variety in amplitude magnitudes and the rate of change in amplitude. At the same time, models one and three show a higher NMCC during testing than model two's results. If a model is trained on different distributed data than the data used during the test phase, there would be a risk of impaired performance [37]. Additionally, DL models have a high risk of overfitting on small and noisy data sets due to the complexity of the DL architecture [36]. Reducing the complexity of the model by removing layers during the training process could provide a way to improve the performance of this model [36]. An overfitted model could result in a higher loss and error during testing and implementation, especially when using a small data set [38]. A good training data set contains enough variability to prepare the model for processing new data [38]. Since only limited data were available for training these models, we did not have enough data to train all models with a representative training set, causing a covariate shift.

2) The DL network might not be able to follow fast changes in the amplitude of the tibia's landmark location in long segments of data in more than one direction. We used relatively long segments compared to a previous study on spatio-temporal localisation of the temporomandibular joint in US videos with a similar DL architecture that found a mean Euclidian distance of 2,14 mm (2,86 SD) for segments of 60 frames each [14]. The results are hardly comparable since they study a different body part. Therefore, the motion pattern is probably not the same. However, it is essential to look at the absolute tracking error of their version of this DL model (trained with a similar architecture). When compared to the boxplots in figure 12, it followed that the absolute tracking error of our models, expressed in Euclidian distance, was higher for 75% of all tracked points (the lower quantile of each model for each task exceeds 2.5 mm, while their model's mean absolute tracking error was 2.14mm). Each segment contained one step (one anterior translation of the jaw and back again), processed separately during each simulation, contrary to processing eight steps consecutively in the simulation of our DL models (eight displacements in the anterior-posterior and lateral-medial domain). The temporomandibular joint only showed a large translation in the anterior-posterior direction. On the contrary, the translation in the perpendicular direction is negligible. In this study, the movement of the tibia shows translations in two directions; anterior-posterior and lateral-medial. The resulting NRMSE in the lateral-medial direction of our models was mainly higher than in the anterior-posterior direction. That might indicate impairment of this DL architecture for following the displacement in two directions instead of a translation.

C. Repeatability of the semi-automatic single-point tracker and deep learning models

A higher SD indicates lower repeatability of the step task. The SD of most true trajectories tracked by the semi-automatic single-point tracker is around 1 mm. Participant two shows the highest SD, which is proportional to the amount of tibial movement shown in the ultrasound recordings. Since the shape of all trajectories from the tibia's landmark location shows two peaks in one direction, the SD will be higher during tibial movement. The movement will mainly occur when the prosthesis is lifted from the ground (during the swing phase) and just before. When the participants fully load their weight on the prosthetic foot, the residual limb is pushed into the socket, stagnant in one location. Since all standard deviations are averaged during the step, this stagnation of the tibia in the exact location might give a biased result when assessing the repeatability of these measurements. Only looking at the parts of the step where the residual limb is in movement during the swing phase might improve this approach.

During the forward and backward step task, the participants had to shift their weight forward or backwards while placing the prosthetic foot farther in the anterior or posterior direction, respectively. During the total loading of the prosthetic foot, one of the knees is in extension during both tasks when balancing. At the same time, in extension, their muscles exert various activation patterns to accommodate this motion during the forward or backward step task [39].

During the sideways step task, the participants shift their centre of gravity sideways while balancing on their prosthesis in the stance phase. At this point, the centre of pressure (COP) is located on the lateral side of the prosthetic foot. After landing on the prosthetic foot, the task was to move the prosthetic foot back to the initial position by shifting the COP to the medial side of the prosthetic foot and stepping back. A recent study on dynamic balancing responses in UTA during slow treadmill walking (0,5 m/s) showed that in-stance COP modulation strategies while in the stance phase heavily rely on the ankle mechanism in the control group without amputation [40]. Since ankle muscles are absent in the prosthesis of UTA, their natural response is to adjust the placement of the non-amputated leg.

D. Limitations

1. Human experimental procedure

Only three subjects were involved in this study, limiting the versatility and availability of data produced during this experiment. The experimental procedure was demonstrated, but each participant could interpret and exert the demonstration themselves, which might reduce the repeatability of the task. The forward- and backward tasks were motion patterns that could also occur during gait, thus might be more familiar to the participants.

Due to the step exercise where the non-amputated leg should remain in the initial location, the adjustment response of the non-amputated leg was not possible. Therefore, the participants had to adapt their strategy to balance the stance phase on their prosthetic foot. Since the kinematic motion patterns for the step tasks are not temporally and kinematically constrained (e.g., a fixed movement path of the prosthetic foot, assisted with lights to indicate the desired position of the foot in time) [39], a variance in the exercise might be more evident during the sideways step task than during the forward- or backward task [11].

Also, one CPO appointed the landmark locations in the semi-automated motion tracker. Following one point by observation proved to be tedious since the US images were distorted with noise, and the range of view in the recording did not cover the complete tibia.

2. Deep learning

In this study, the metrics used for the evaluation of the training procedure and the testing procedure were not the same. Also, a validation phase was not performed after training the models due to the limited availability of data, so evaluating the loss per epoch was impossible and optimising the weights per epoch could not be done. Since only 66% of the data obtained from three participants is used for training, a small data set remains with little variability. Therefore, the complex models were likely to overfit the noisy training data [36].

All models show insufficient similarity for the lateral-medial trajectories of the landmark location point of the tibia. This could, for instance, be caused by a variance in the width of the segmented part of the tibia, as shown in figure 11. Furthermore, the study on tracking the temporomandibular joint only entailed a large amplitude in one translatory direction and achieved a lower tracking error than the DL models investigated in this research. Consequently, that might indicate impairment of this DL architecture to track displacement of the tibia's landmark location point instead of translation. A study on tracking liver lesions in respiratory motion from US sequences with a CNN of a different architecture showed similar amplitude magnitudes in two directions and reported a mean Euclidian distance of 0.69 mm (0.67 SD) [34]. Their model was trained on 24 and tested on 39 sequences. The duration of the US sequences ranged from four seconds to 10 minutes. However, the frequency of the motion is about four times lower than for the translation of the tibia's landmark location. They included locating the actual landmark location in the DL approach, which could also be an improvement to our model. Their model achieved high accuracy and their training data set was more extensive than the set used in this study. Hence, a larger training data set obtained from a variant participant pool is needed to discover the full potential of this DL architecture.

E. Recommendations

1. Human experimental procedure

It would be interesting to investigate whether a higher constraint of the kinematic motion pattern would lead to lower standard deviations of the mean trajectory of the tibia's landmark location point. Since one CPO in this study labelled the ground truth, the appointed landmark location is still prone to human error. Therefore, for future research, we recommend defining a standard landmark location visible in all US recordings, like a bony landmark, that could be used to train a DL model for automated detection of this location on the tibia.

Since we have shown that measuring the displacement of a landmark location on the tibia in a prosthetic socket is feasible with a semi-automated motion tracker, further studies could also try to measure during gait. Using a second US probe could be interesting to enable estimation of the angular range of motion of the residual limb during gait [11].

2. Deep learning

For further research, it is essential to include a training, validation and testing phase with the same metrics for evaluation of the performance of the models in all stages. The optimal weights can be retrieved when the validation loss is minimal [36]. Hyperparameters that influence the complexity of the models (e.g., dropping out layers, adjusting batch size and regularisation) can also be tuned during this phase.

Another good starting point for future research might be to investigate whether this DL network achieves better accuracy when trained on a more extensive data set with a larger variety in tibial trajectories, using k-fold cross-validation with hyperparameter search as a training procedure [38]. To avoid deviations from the desired point due to a variance in the shape of the segmented part of the tibia, using deep learning to find the final landmark location could also be an improvement for future research [34].

V. Conclusion

To conclude, a 3D U-net with an LSTM model can track the tibia's landmark location trajectory in the anterior-posterior direction with a delay with respect to the semi-automatic single-point tracker. However, the similarity of the landmark location trajectory of the models compared to a semi-automated single point tracker did not reach an NMCC within a sufficient margin for the lateral-medial trajectories of the tibia's landmark location trajectory. So, the DL model has the potential to assist researchers in tracking the anterior-posterior trajectory of a landmark location on the tibia.

Ethical approval

This study was approved by the Medical Ethical Review Board of the University Medical Centre Groningen, Groningen, the Netherlands (NL74038.042.21). All participants provided written consent prior to the start of the study.

Funding

This work was supported by the Stichting ZiektekostenVerzekering Krijgsmacht [grant number 20-0231].

References

[1] R. Safari, "Lower limb prosthetic interfaces: Clinical and technological advancement and potential future direction," *Prosthetics and Orthotics International*, vol. 44, no. 6, pp. 384–401, 2020, ISSN: 17461553. DOI: 10.1177/0309364620969226.

[2] V. Rajtukova, R. Hudak, J. Zivcak, P. Halfarova, and R. Kudrikova, "Pressure distribution in transtibial prostheses socket and the stump interface," in *Procedia Engineering*, vol. 96, Elsevier Ltd, Jan. 2014, pp. 374–381. DOI: 10.1016/j.proeng.2014.12.106.

[3] A. K. Laprè, V. Q. Nguyen, U. Baspinar, M. White, and F. C. Sup, "Capturing prosthetic socket fitment: Preliminary results using an ultrasound-based device," *IEEE International Conference on Rehabilitation Robotics*, pp. 1221–1226, Aug. 2017, ISSN: 19457901. DOI: 10.1109/ICORR.2017.8009416.

[4] W. L. Childers and S. Siebert, "Marker-based method to measure movement between the residual limb and a transtibial prosthetic socket," *Prosthetics and Orthotics International*, vol. 40, no. 6, pp. 720–728, Dec. 2016. DOI: 10.1177/0309364615610660. [Online]. Available: https://journals-lww-com.tudelft.idm.oclc.org/poijournal/Fulltext/2016/40060/Marker_based_method_to_measure_movement_between.9.aspx.

[5] J. Tang, M. Mcgrath, N. Hale, *et al.*, "A combined kinematic and kinetic analysis at the residuum/socket interface of a knee-disarticulation amputee," *Medical Engineering and Physics*, vol. 49, pp. 131–139, 2017. DOI: 10.1016/j.medengphy.2017.

08.014. [Online]. Available: http://dx.doi.org/10.1016/j.medengphy.2017.08.014.

[6] A. K. LaPrè, M. A. Price, R. D. Wedge, B. R. Umberger, and F. C. Sup, "Approach for gait analysis in persons with limb loss including residuum and prosthesis socket dynamics," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 34, no. 4, 2018, ISSN: 20407947. DOI: 10.1002/cnm.2936.

[7] V. Noll, J. Wojtusch, J. Schuy, M. Grimmer, P. Beckerle, and S. Rinderknecht, "Measurement of biomechanical interactions at the stump-socket interface in lower limb prostheses," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2015-November, pp. 5517–5520, Nov. 2015. DOI: 10.1109/EMBC.2015.7319641.

[8] A. van Heesewijk, A. Crocombe, S. Cirovic, M. Taylor, and W. Xu, "Evaluating the effect of changes in bone geometry on the trans-femoral socket-residual limb interface using finite element analysis," *IFMBE Proceedings*, vol. 68, no. 2, pp. 587–591, 2018. DOI: 10.1007/978-981-10-9038-7{_}109.

[9] J. W. Steer, P. R. Worsley, M. Browne, and A. Dickinson, "Key considerations for finite element modelling of the residuum–prosthetic socket interface," *Prosthetics and Orthotics International*, 2020. DOI: 10.1177/0309364620967781.

[10] T. Heckman, J. C. Krumm, and P. Spie, "Searching for contours," no. March 1996, 2021. DOI: 10.1117/12.234745.

[11] P. Convery and K. D. Murray, "Ultrasound study of the motion of the residual femur within a trans-femoral socket during daily living activities other than gait," *Prosthetics and Orthotics International*, vol. 25, no. 3, pp. 220–227, 2001. DOI: 10.1080/03093640108726605.

[12] C. SY and R. O, "Exploring the Use of Non-Image-Based Ultrasound to Detect the Position of the Residual Femur within a Stump," *PloS one*, vol. 11, no. 10, Oct. 2016, ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0164583. [Online]. Available: https://pubmed-ncbi-nlm-nih-gov.tudelft.idm.oclc.org/27764120/.

[13] S. N. Friedman, M. Grushka, H. K. Beituni, M. Rehman, H. B. Bressler, and L. Friedman, "Advanced Ultrasound Screening for Temporomandibular Joint (TMJ) Internal Derangement," *Radiology Research and Practice*, vol. 2020, pp. 1–10, May 2020, ISSN: 2090-1941. DOI: 10.1155/2020/1809690.

[14] K. Belikova, A. Zailer, S. V. Tekucheva, S. N. Ermoljev, and D. V. Dylov, "Deep Learning for Spatio-Temporal Localization of Temporomandibular Joint in Ultrasound Videos," pp. 1257–1261, Jan. 2022. DOI: 10.1109/BIBM52615.2021.9669857.

[15] L. J. Brattain, B. A. Telfer, M. Dhyani, J. R. Grajo, and A. E. Samir, "Machine learning for medical ultrasound: status, methods, and future opportunities," *Abdominal Radiology*, vol. 43, no. 4, pp. 786–799, 2018. DOI: 10.1007/s00261-018-1517-0. [Online]. Available: https://doi.org/10.1007/s00261-018-1517-0.

[16] I. Hacihaliloglu and A. Rasoulian, "Statistical Shape Model to 3D Ultrasound Registration for Spine Interventions," pp. 361–368, 2013.

[17] X. Dai, Y. Lei, J. Roper, *et al.*, "Deep learning-based motion tracking using ultrasound images," *Medical Physics*, vol. 48, no. 12, pp. 7747–7756, Dec. 2021, ISSN: 2473-4209. DOI: 10.1002/MP.15321. [Online]. Available: https://onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/full/10.1002/mp.15321%20https://onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/abs/10.1002/mp.15321%20https://aapm-onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/10.1002/mp.15321.

[18] C. Mayer, M. Danelljan, D. P. Paudel, and L. Van Gool, "Learning Target Candidate Association to Keep Track of What Not to Track," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 13 424–13 434, 2021, ISSN: 15505499. DOI: 10.1109/ICCV48922.2021.01319.

[19] I. Hacihaliloglu, R. Abugharbieh, A. J. Hodgson, and R. N. Rohling, "Bone Surface Localization in Ultrasound Using Image Phase-Based Features," *Ultrasound in Medicine and Biology*, vol. 35, no. 9, pp. 1475–1487, 2009, ISSN: 03015629. DOI: 10.1016/j.ultrasmedbio.2009.04.015.

[20] I. Hacihaliloglu, "Enhancement of bone shadow region using local phase-based ultrasound transmission maps," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 6, pp. 951–960, 2017, ISSN: 18616429. DOI: 10.1007/s11548-017-1556-y.

[21] A. Rangamani, T. Xiong, A. Nair, T. D. Tran, and S. P. Chin, "Landmark Detection and Tracking in Ultrasound using a CNN-RNN Framework," *Conference on Neural Information Processing Systems (NIPS)*, no. April 2017, 2016. [Online]. Available: https://www.researchgate.net/publication/316441246.

[22] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation," Feb. 2018. DOI: 10.48550/arxiv.1802.06955. [Online]. Available: https://arxiv.org/abs/1802.06955v5.

[23] M. X. Jiang, C. Deng, Z. G. Pan, L. F. Wang, and X. Sun, "Multi-object tracking in videos based on LSTM and deep reinforcement learning," *Complexity*, vol. 2018, 2018, ISSN: 10990526. DOI: 10.1155/2018/4695890.

[24] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, pp. 448–456, Feb. 2015. DOI: 10.48550/arxiv.1502.03167. [Online]. Available: https://arxiv.org/abs/1502.03167v3.

[25] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "Skeleton-based human activity recognition using ConvLSTM and guided feature learning," *Soft Computing*, vol. 26, no. 2, pp. 877–890, Jan. 2022, ISSN: 14337479. DOI: 10.1007/S00500-021-06238-7.

[26] W. Chen, F. Zheng, S. Gao, and K. Hu, "An LSTM with Differential Structure and Its Application in Action Recognition," *Mathematical Problems in Engineering*, vol. 2022, 2022, ISSN: 15635147. DOI: 10.1155/2022/7316396.

[27] K. Lee, J. Zung, P. L. Google, V. J. Google, and H. S. Seung, "Superhuman Accuracy on the SNEMI3D Connectomics Challenge," [Online]. Available: http://brainiac2.mit.edu/SNEMI3D/home.

[28] S. Jadon, "A survey of loss functions for semantic segmentation," *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2020*, Oct. 2020. DOI: 10.1109/CIBCB48159.2020.9277638.

[29] N. Otsu, "THRESHOLD SELECTION METHOD FROM GRAY-LEVEL HISTOGRAMS.," *IEEE Trans Syst Man Cybern*, vol. SMC-9, no. 1, pp. 62–66, 1979, ISSN: 00189472. DOI: 10.1109/TSMC.1979.4310076.

[30] L. Rocha, L. Velho, and P. C. P. Carvalho, "Image moments-based structuring and tracking of objects," *Brazilian Symposium of Computer Graphic and Image Processing*, vol. 2002-January, pp. 99–105, 2002, ISSN: 15301834. DOI: 10.1109/SIBGRA.2002.1167130.

[31] M. G. Cox, "The Numerical Evaluation of £-Splines*," *J. Inst. Maths Applies*, vol. 10, pp. 134–149, 1972. [Online]. Available: https://academic.oup.com/imamat/article/10/2/134/687696.

[32] C. de Boor, "On calculating with B-splines," *Journal of Approximation Theory*, vol. 6, no. 1, pp. 50–62, 1972, ISSN: 10960430. DOI: 10.1016/0021-9045(72)90080-9.

[33] N. Sarkalkan, A. J. Loeve, K. W. Van Dongen, G. J. Tuijthof, and A. A. Zadpoor, "A Novel Ultrasound Technique for Detection of Osteochondral Defects in the Ankle Joint: A Parametric and Feasibility Study," *Sensors 2015, Vol. 15, Pages 148-165*, vol. 15, no. 1, pp. 148–165, Dec. 2014, ISSN: 1424-8220. DOI: 10.3390/S150100148. [Online]. Available: https://www.mdpi.com/1424-8220/15/1/148/htm%20https://www.mdpi.com/1424-8220/15/1/148.

[34] F. Liu, D. Liu, J. Tian, X. Xie, X. Yang, and K. Wang, "Cascaded one-shot deformable convolutional neural networks: Developing a deep learning model for respiratory motion estimation in ultrasound sequences," *Medical Image Analysis*, vol. 65, p. 101 793, Oct. 2020, ISSN: 1361-8415. DOI: 10.1016/J.MEDIA.2020.101793.

[35] S. Khaki, H. Pham, Y. Han, A. Kuhl, W. Kent, and L. Wang, "Convolutional neural networks for image-based corn kernel detection and counting," *Sensors (Switzerland)*, vol. 20, no. 9, May 2020, ISSN: 14248220. DOI: 10.3390/S20092721.

[36] X. Ying, "An Overview of Overfitting and its Solutions," *Journal of Physics: Conference Series*, vol. 1168, no. 2, Mar. 2019, ISSN: 17426596. DOI: 10.1088/1742-6596/1168/2/022022.

[37] M. Sugiyama, T. Suzuki, S. Nakajima, *et al.*, "Direct importance estimation for covariate shift adaptation," *Ann Inst Stat Math*, vol. 60, pp. 699–746, 2008. DOI: 10.1007/s10463-008-0197-x.

[38] A. Géron, *1. The Machine Learning Landscape | Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*, 2019. DOI: 9781492032649. [Online]. Available: https://learning-oreilly-com.tudelft.idm.oclc.org/library/view/hands-on-machine-learning/9781492032632/ch01.html#idm45022180452984.

[39] K. Zhao, Z. Zhang, H. Wen, A. Scano, and A. Adamatzky, "Intra-Subject and Inter-Subject Movement Variability Quantified with Muscle Synergies in Upper-Limb Reaching Movements," 2021. DOI: 10.3390/biomimetics6040063. [Online]. Available: https://doi.org/10.3390/biomimetics6040063.

[40] A. Olenšek, M. Zadravec, H. Burger, and Z. Matjačić, "Dynamic balancing responses in unilateral transtibial amputees following outward-directed perturbations during slow treadmill walking differ considerably for amputated and non-amputated side," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, no. 1, Dec. 2021, ISSN: 17430003. DOI: 10.1186/S12984-021-00914-3.

Appendix A: Tables

Similarity and standard deviations

Model	*Model 1*			*Model 2*			*Model 3*		
Direction	NRMSE [%]	Delay [step %]	NMCC	NRMSE [%]	Delay [step %]	NMCC	NRMSE [%]	Delay [step %]	NMCC
Sideways									
Anterior-posterior	33,4	2,0	0,988	61,0	6,5	0,707	77,3	5,5	0,953
Lateral-medial	85,1	2,5	0,894	74,7	5,0	0,914	38,9	7,5	0,307
Forward									
Anterior-posterior	27,7	1,5	0,996	68,3	6,5	0,928	67,7	7,0	0,986
Lateral-medial	90	2,5	0,966	76,9	-	0,771	53,3	7,0	0,775
Backwards									
Anterior-posterior	58,4	9,5	0,912	66,2	10,5	0,897	50,6	8,5	0,998
Lateral-medial	57,1	11,0	0,219	84,9	-	0,221	64,7	13,0	0,739

Table 2. Normalised root mean square errors, delays in step percentage and normalised maximum cross-correlation for the trajectory of all three automated models compared to the course of the landmark location on the tibia determined by the semi-automated model (for participant one, two, and three)

Direction	subject	Standard deviation average [mm]	Interquartile range	Model	Standard deviation average [mm]	Interquartile range
Anterior-posterior	participant 1	0.83	[0.54, 1.13]	model 1	0.84	[0.45, 1.24]
Anterior-posterior	participant 2	0.97	[0.82, 1.24]	model 2	1.33	[0.97, 1.74]
Anterior-posterior	participant 3	0.62	[0.47, 0.72]	model 3	0.50	[0.24, 0.64]
Lateral-medial	participant 1	0.80	[0.66, 1.12]	model 1	0.43	[0.26, 0.56]
Lateral-medial	participant 2	2.34	[1.74, 3.32]	model 2	2.05	[1.57, 2.53]
Lateral-medial	participant 3	0.78	[0.69, 0.81]	model 3	1.42	[0.96, 1.75]

Table 3. For the sideways task: The standard deviation average and interquartile range for all subjects (semi-automatically tracked) and models (automatically tracked)

Direction	subject	Standard deviation average [mm]	Interquartile range	Model	Standard deviation average [mm]	Interquartile range
Anterior-posterior	participant 1	0.98	[0.77, 1.21]	model 1	1.11	[0.71, 1.53]
Anterior-posterior	participant 2	1.23	[0.71, 1.81]	model 2	1.38	[0.83, 1.98]
Anterior-posterior	participant 3	0.62	[0.33, 0.87]	model 3	0.61	[0.36, 0.82]
Lateral-medial	participant 1	0.58	[0.46, 0.75]	model 1	0.53	[0.48, 0.66]
Lateral-medial	participant 2	1.19	[0.79, 1.47]	model 2	2.24	[1.64, 2.73]
Lateral-medial	participant 3	0.78	[0.69, 0.93]	model 3	1.57	[1.22, 1.89]

Table 4. For the forward task: The standard deviation average and interquartile range for all subjects (semi-automatically tracked) and models (automatically tracked)

Direction	subject	Standard deviation average [mm]	Interquartile range	Model	Standard deviation average [mm]	Interquartile range
Anterior-posterior	participant 1	1.24	[0.98, 1.72]	model 1	1.23	[0.67, 1.87]
Anterior-posterior	participant 2	2.43	[1.66, 3.12]	model 2	2.76	[1.95, 3.71]
Anterior-posterior	participant 3	0.71	[0.54, 0.90]	model 3	0.72	[0.50, 0.91]
Lateral-medial	participant 1	0.61	[0.38, 0.90]	model 1	1.41	[0.60, 2.18]
Lateral-medial	participant 2	1.71	[1.02, 2.26]	model 2	2.92	[1.79, 4.22]
Lateral-medial	participant 3	0.75	[0.57, 0.92]	model 3	2.10	[1.84, 2.4]

Table 5. For the backward task: The standard deviation average and interquartile range for all subjects (semi-automatically tracked) and models (automatically tracked)

3D U-Net and LSTM module

Hyperparameters	Values
Number of layers	2
Batch size	256
Hidden size	1024
Optimizer	Adam
Learning rate	0.00001
Weight decay regularization	0.0005
Sequence length	8
Number of epochs	100
Dropout probability	0.2
Activation function	Softmax

Tracker results

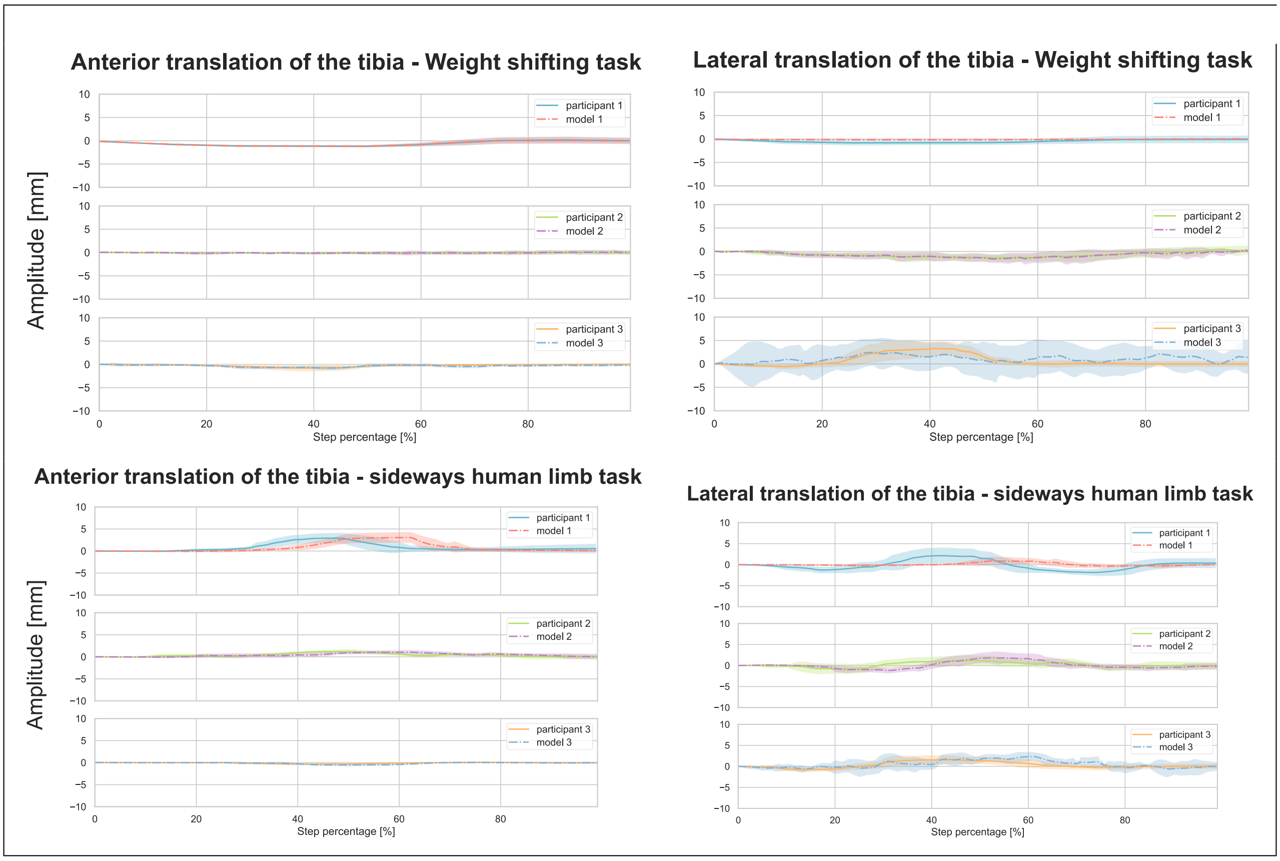


Figure 14. Mean trajectories of the Anterior (left side of the figure) and Lateral (right side of the figure) translation of the tibia w.r.t the starting pose in mm over the step percentage for each task, where 0% indicates stance on two feet at the start of the step and 100% indicates the return to the initial position on two feet.

US settings

Setting	Values
Frequency	8.8 MHz/Penetration
Frame avg	4
Dynamic range	125 dB
Reject level	1
Gray map	13
Multivision	Low
Clearvision	3
scan area	100%
2D Image Size	90
Tissue	1500
Edge enhance	0
Focus number	1
Power	100 VAC

US probe placement

subject	Knee [mm]	transducer [mm]
One	510	420
Two	555	465
Three	560	482

Table 6. Distance from the ground to the knee and US transducer

Appendix B: Test-retest reliability

an ICC (Two-way mixed effects, mean of k measurements, absolute agreement) analysis was performed on the trajectories found for each task per subject. The test-retest reliability of the landmark trajectories from the semi-automatic single-point tracker and the models were compared for all steps, each movement direction (anterior-posterior or lateral-medial), and step task (sideways, forward or backwards) per participant with the Intraclass Correlation Coefficient (ICC) defined in table 7.

Model	Two-way mixed effects
Type	Mean of k measurements
Definition	Absolute agreement

Table 7. ICC, intraclass correlation coefficient for test-retest reliability [1]

According to the following equation:

$$ICC = \frac{\mathrm{MS_R} - \mathrm{MS_E}}{\mathrm{MS_R} + \frac{\mathrm{MS_C} - \mathrm{MS_E}}{n}} \tag{13}$$

where MSR = mean square for rows (each row represents one of 200 data points per step); MSE = mean square for error between steps; MSC = mean square for columns (each column represents one step); n = number of participants, in this case one. ICC values less than 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability [2]. After interpolating (second order cardinal b-spline) all landmark location points per step over 200 knots, the ICC (Two-way mixed effects, mean of k measurements, absolute agreement) was calculated. K is the amount of steps per step task.

Test-retest reliability results

ICC(Two-way mixed effects, mean of k measurements, absolute agreement) values less than 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability and values greater than 0.90 indicate excellent reliability [2]. The standard deviation median and its interquartile range of the mean tibia's landmark location trajectory and ICC values for the sideways, forward and backward tasks can be found in table 8, 9, and 10, respectively. The ICC for the sideways and the forward task is generally higher in the anterior-posterior direction than the lateral-medial direction for all participants. The ICC of the DL models is lower than the ICC of the semi-automatic single-point tracker for all participants except for the anterior-posterior course tracked by model three during the sideways step task.

- **Participant one:** In the anterior-posterior direction, participant one achieves good to excellent test-retest reliability for all step tasks. For the lateral-medial direction, the ICC is good for the sideways task, good for the forward task, and good to excellent for the backwards task.
- **Model one:** The test-retest reliability of model one in the anterior-posterior direction is slightly lower than that of participant one but still achieves a good ICC for all step tasks. However, in the lateral-medial direction, the ICC indicates poor reliability for the backwards and forward tasks and poor to moderate reliability for the sideways task.
- **Participant two:** In the anterior-posterior direction, the sideways task trajectory of the tibia's landmark location trajectory for participant two shows moderate to good test-retest reliability; for the forward task the ICC reaches excellent score while for the backwards task, it achieves good to excellent reliability. In the lateral medial direction, the ICC is good for the sideways task and excellent during the forward and backwards tasks.
- **Model two:** In the anterior-posterior direction, the ICC for model two is poor to moderate for the sideways task, good to excellent for the forward task and moderate for the backwards task. In lateral-medial direction, the score is moderate to good for approximating the tibia's landmark location trajectory of the sideways task, poor to moderate for the forward task and poor for the backwards task.
- **Participant three:** In the anterior-posterior direction, the test-retest reliability of the tibia's landmark location trajectory is moderate to good for the sideways task, good for the forward task and excellent for the backwards task. In lateral-medial direction, it is moderate to good for the sideways and forward task and good to excellent for the backwards task.
- **Model three:** Model three achieves good to excellent ICC for the anterior-posterior direction of the sideways, good for the forward and excellent for the backwards task trajectory of the tibia's landmark location. The lateral-medial directional sideways task trajectory's ICC is poor for the sideways and forward tasks and moderate to good for the backwards task.

The ICC(Two-way mixed effects, mean of k measurements, absolute agreement) was calculated to find the test-retest reliability of the tibia's landmark location trajectories obtained from the semi-automated single-point tracker and the three DL models. A visual representation of the average ICC values per step task can be found in figure 15.

Direction	subject	Standard deviation average [mm]	Interquartile range	ICC	95% confidence intervals	Model	Standard deviation average [mm]	Interquartile range	ICC	95% confidence intervals
Anterior-posterior	participant 1	0.83	[0.54, 1.13]	0.91	[0.89 0.93]	model 1	0.84	[0.45, 1.24]	0.90	[0.88 0.92]
Anterior-posterior	participant 2	0.97	[0.82, 1.24]	0.71	[0.62 0.78]	model 2	1.33	[0.97, 1.74]	0.59	[0.38 0.72]
Anterior-posterior	participant 3	0.62	[0.47, 0.72]	0.79	[0.62 0.88]	model 3	0.5	[0.24, 0.64]	0.88	[0.82 0.91]
Lateral-medial	participant 1	0.8	[0.66, 1.12]	0.66	[0.51 0.76]	model 1	0.43	[0.26, 0.56]	0.60	[0.49 0.68]
Lateral-medial	participant 2	2.34	[1.74, 3.32]	0.81	[0.75 0.86]	model 2	2.05	[1.57, 2.53]	0.72	[0.61 0.79]
Lateral-medial	participant 3	0.78	[0.69, 0.81]	0.71	[0.63 0.78]	model 3	1.42	[0.96, 1.75]	0.37	[0.24 0.49]

Table 8. For the sideways task: The standard deviation average and interquartile range with ICC values with 95% confidence intervals for all subjects (semi-automatically tracked) and models (automatically tracked)

Direction	subject	Standard deviation average [mm]	Interquartile range	ICC	95% confidence intervals	Model	Standard deviation average [mm]	Interquartile range	ICC	95% confidence intervals
Anterior-posterior	participant 1	0.98	[0.77, 1.21]	0.89	[0.85 0.92]	model 1	1.11	[0.71, 1.53]	0.85	[0.8 0.89]
Anterior-posterior	participant 2	1.23	[0.71, 1.81]	0.95	[0.93 0.96]	model 2	1.38	[0.83, 1.98]	0.91	[0.88 0.93]
Anterior-posterior	participant 3	0.62	[0.33, 0.87]	0.85	[0.81 0.88]	model 3	0.61	[0.36, 0.82]	0.87	[0.83 0.89]
Lateral-medial	participant 1	0.58	[0.46, 0.75]	0.83	[0.73 0.89]	model 1	0.53	[0.48, 0.66]	0.33	[0.16 0.47]
Lateral-medial	participant 2	1.19	[0.79, 1.47]	0.95	[0.93 0.96]	model 2	2.24	[1.64, 2.73]	0.45	[0.32 0.57]
Lateral-medial	participant 3	0.78	[0.69, 0.93]	0.70	[0.57 0.78]	model 3	1.57	[1.22, 1.89]	0.63	[0.5 0.72]

Table 9. For the forward task: The standard deviation average and interquartile range with ICC values with 95% confidence intervals for all subjects (semi-automatically tracked) and models (automatically tracked)

Direction	subject	Standard deviation average [mm]	Interquartile range	ICC	95% confidence intervals	Model	Standard deviation average [mm]	Interquartile range	ICC	95% confidence intervals
Anterior-posterior	participant 1	1.24	[0.98, 1.72]	0.89	[0.87 0.92]	model 1	1.23	[0.67, 1.87]	0.88	[0.84 0.9]
Anterior-posterior	participant 2	2.43	[1.6, 3.12]	0.88	[0.83 0.91]	model 2	2.76	[1.95, 3.71]	0.74	[0.56 0.75]
Anterior-posterior	participant 3	0.71	[0.54, 0.9]	0.95	[0.92 0.96]	model 3	0.72	[0.50, 0.91]	0.96	[0.91 0.96]
Lateral-medial	participant 1	0.61	[0.38, 0.9]	0.89	[0.86 0.91]	model 1	1.41	[0.60, 2.18]	0.40	[0.28 0.52]
Lateral-medial	participant 2	1.71	[1.02, 2.26]	0.95	[0.94 0.96]	model 2	2.72	[1.79, 4.22]	-0.26	[-0.53 -0.03]
Lateral-medial	participant 3	0.75	[0.57, 0.92]	0.92	[0.89 0.94]	model 3	2.10	[1.84, 2.40]	0.70	[0.51 0.81]

Table 10. For the backward task: The standard deviation average and interquartile range with ICC values for all subjects (semi-automatically tracked) and models (automatically tracked)

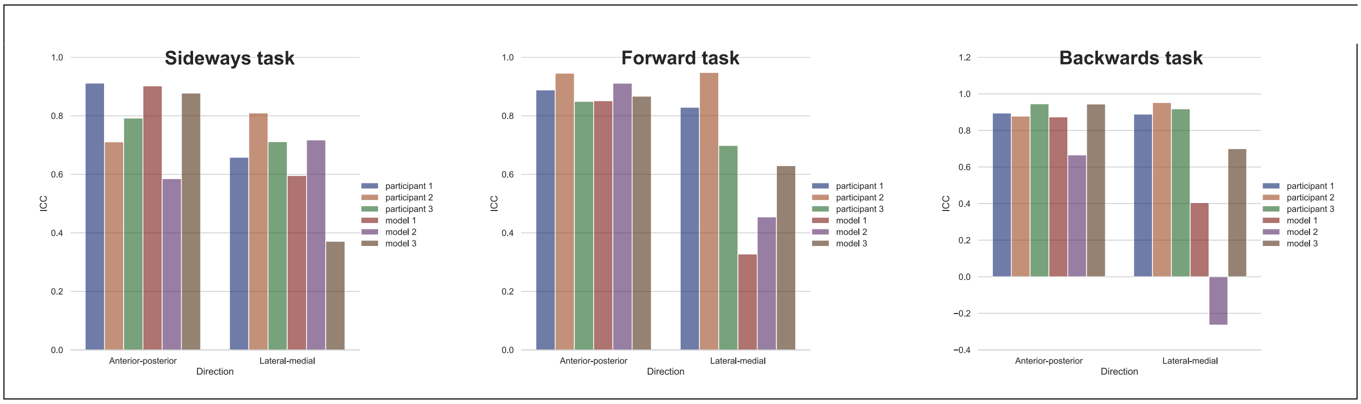


Figure 15. The mean ICC for all participants (semi-automatic tracker) and models (automatic tracker) per task.

These are the qualitative scores of the true trajectories for tracking the tibia's landmark location point;

- **Excellent:**
  - The anterior-posterior, forward task trajectory of participant one and two and backwards task trajectory of participant three
  - The lateral-medial forward task trajectory of participant three.
- **Good to excellent:**
  - The anterior-posterior, sideways and forward task trajectory of participant one and the backwards task trajectory of participant one and two.
  - The lateral-medial, backwards task trajectory of participants one and three.
- **Good:**
  - The anterior-posterior forward trajectory of participant three.
  - The lateral-medial, sideways trajectory of participant two

- **Moderate to good:**
  - The anterior-posterior, sideways trajectory for participants two and three
  - The lateral-medial, sideways and forward trajectories of participants one and three.

Test-retest reliability discussion
So, four of the 18 semi-automatically tracked landmark location trajectories reach excellent, and six have excellent test-retest reliability. The ICC of the sideways task is lower than those for the backwards and forward tasks. This part is not included in the main body of the paper since the ICC from only three participants should be taken with caution [2]. A too optimistic outcome might occur when calculating an ICC score according to 200 points for each step, since the start, middle, and final part of each trajectory of the landmark location point on the tibia might occur around the same values. An ICC might add more value when only focused on the part of the step where the displacement and change in displacement of the tibia is of the highest magnitude. However, it would be interesting to investigate whether a higher constraint of the kinematic motion pattern would lead to more excellent ICC scores in that case.

Test-retest reliability conclusion
The test-retest reliability for the true anterior-posterior trajectories of the backward and forward steps reached good to excellent scores. However, in the lateral-medial direction, the scores were lower as well as for the DL models.

Appendix C: Additional visualizations

Amplitude change

The amplitude changes for the true trajectories of the tibia's landmark location point obtained with the semi-automatic motion tracker can be found in figure 16. The gradient of the mean tibia's landmark location trajectory is calculated by dividing the increase/decrease in amplitude in [mm] by the difference in step % for each step task, direction and participant.

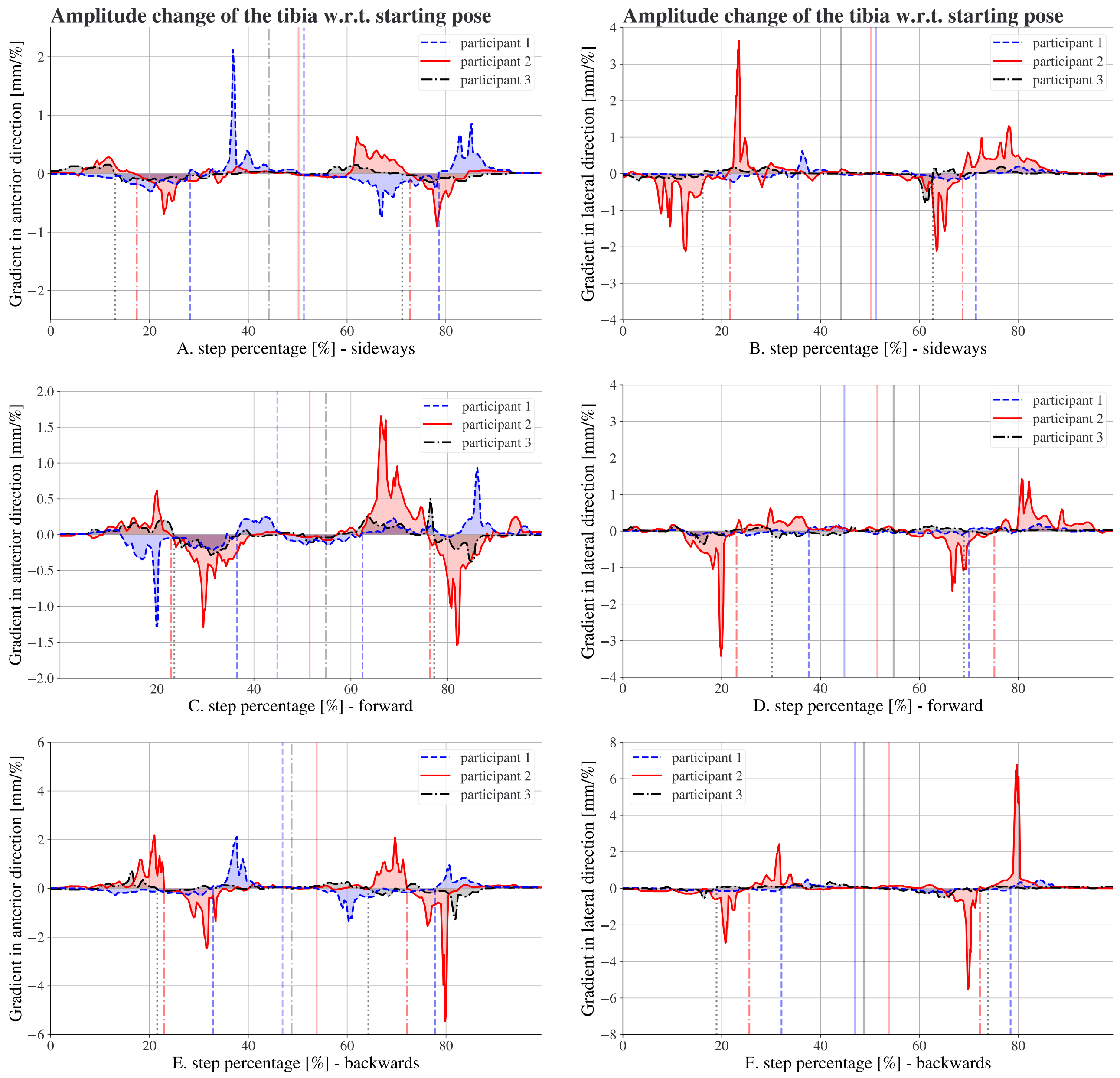


Figure 16. The amplitude changes of the mean trajectories of the Anterior (A, C, and E) and Lateral (B, D, and F) translation of the tibia w.r.t the starting pose in $[mm/\%]$ over the step percentage, where 0% is the start of the step, and 100% is the end of each step. A negative amplitude change in the anterior direction = positive amplitude change in the posterior direction, and a negative amplitude change in the lateral direction = a positive amplitude change in the medial direction.

Landmark location point trajectory in two dimensions with respect to the prosthetic edge

The mean total displacement of the tibia's landmark location point relative to the prosthetic edge for each step task and participant is shown in figure 17.

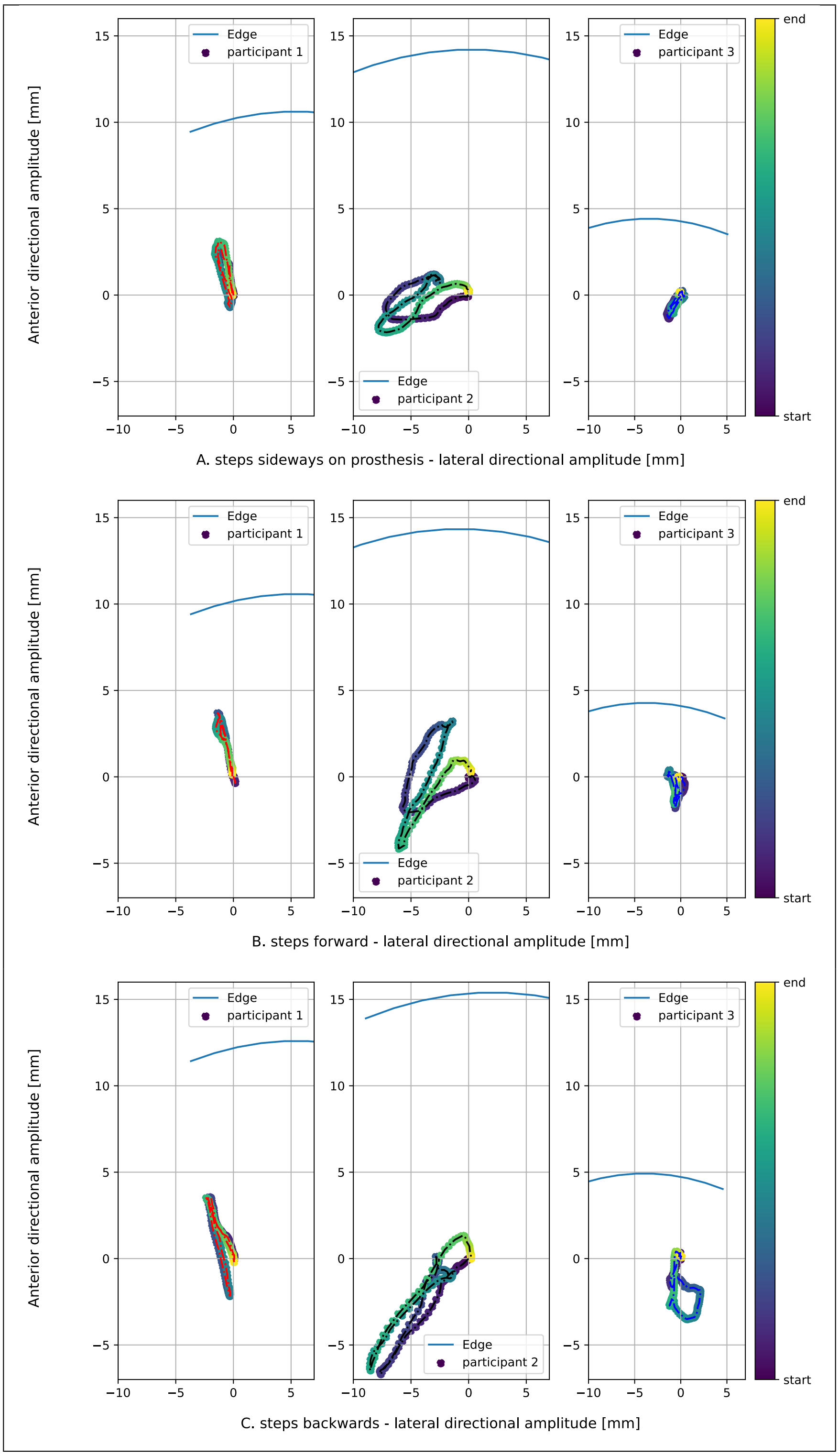


Figure 17. The mean trajectories in x- (Anterior-posterior) and y-(Lateral-medial direction all participants (semi-automatic tracker) and models (automatic tracker) per task.

References

[1] K. O. Mcgraw, "Forming Inferences About Some Intraclass Correlation Coefficients Intrinsic motivation View project," 1996. DOI: `10.1037/1082-989X.1.1.30`. [Online]. Available: `https://www.researchgate.net/publication/232568057`.

[2] T. K. Koo and M. Y. Li, "Cracking the Code: Providing Insight Into the Fundamentals of Research and Evidence-Based Practice A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," 2016. DOI: `10.1016/j.jcm.2016.02.012`. [Online]. Available: `http://dx.doi.org/10.1016/j.jcm.2016.02.012`.