DEEP LEARNING IN GAIT ANALYSIS: THE EFFECT OF MARKER PRESENCE IN NEURAL NETWORK TRAINING TO KINEMATIC OUTCOMES

Roope Uitto

Master's thesis in Biomechanics
Unit of Biology of Physical Activity
University of Jyväskylä
Spring 2021
Supervisors: Neil Cronin & Janne Avela

ABSTRACT

Uitto, R. 2021. Deep learning in gait analysis: the effect of marker presence in neural network training to kinematic outcomes. Faculty of Sport and Health Sciences, University of Jyväskylä, Master's thesis in Biomechanics, 84 pp., 3 appendices.

Accurate quantification of human movement is usually performed with optoelectronic motion capture systems inside a laboratory by tracking reflective markers on the subject. This is accurate and reliable method, but the markers and laboratory environment can restrict motion, the required equipment is expensive, and preparation of a subject takes time. With recent advances in computer vision, the use of deep learning -based methods for human pose estimation has increased and they've been shown to reach human-level labelling accuracies in 2D and 3D. Some studies have compared deep learning -based methods to optoelectronic systems for acquiring joint kinematics, but those studies haven't implemented anatomically relevant keypoints or biomechanically valid kinematic models in their analyses. Neither is the effect of marker presence been studied in these settings. Thus, the purpose of this study was to investigate how the presence of markers in training data affects the predicting performance of a deep learning -based methods in 3D with proper kinematic model.

18 healthy subjects were recruited to walk and run on a treadmill at moderate speeds, while their gait was recorded simultaneously by two systems: an 8-camera Vicon motion capture system at 300 Hz, and an 8-camera GoPro system at 60 Hz. Two deep learning models were trained with data from GoPros, one with data where subjects wore markers, and the other with data where markers were removed. Trials with markers and without markers were analysed by both models and compared to data from Vicon. Additionally, the between-trial reliability of Vicon data was calculated to give insight into the amount of variability due to intraindividual differences in gait.

The 3D analysis failed due to poor performance of the models in some camera views, but 2D analysis on the right ankle and knee could be performed. The marker presence in the training data didn't clearly affect the performance of models, while performing similarly with data that had markers and failing to produce any results from the trials without markers. Also, when compared to data from Vicon, one supra-threshold cluster was found in the knee and eight in the ankle during SPM analyses. The between-trial reliability was most of the time reasonable for clinical measurements (ICC > 0.9), but pointwise analysis showed clear differences in reliability at different points of the gait cycle. There is promise for deep learning -based methods to be used in clinical gait analysis, if the conditions are appropriate. Additional training of the pretrained neural networks with markers present or absent from data didn't seem to make a difference to the performance of the model.

Key words: deep learning, gait analysis, kinematics, reliability

TIIVISTELMÄ

Uitto, R. 2021. Deep learning in gait analysis: the effect of marker presence in neural network training to kinematic outcomes. Liikuntabiologian tieteenalaryhmä, Jyväskylän yliopisto, Biomekaniikan pro gradu -tutkielma, 84 s., 3 liitettä.

Ihmisen tuottaman liikkeen määrittämiseen käytetään yleensä optoelektronisia liikkeenkaappausjärjestelmiä, jotka perustuvat kohteen iholle kiinnitettävien valoa heijastavien markkerien seurantaan. Nämä laboratorio-olosuhteissa tarkat ja luotettavat järjestelmät ovat kuitenkin kalliita, mittaustapahtuman valmistelu vie reilusti aikaa, ja markkerit voivat estää kohteen luonnollisen liikkumisen. Konenäön kehittymisen myötä syväoppimiseen perustuvien menetelmien käyttö ihmisen asennon määrittelyssä on yleistynyt ja niiden on osoitettu olevan ihmisen kanssa yhtä tarkkoja merkitsemään avainkohtia kuviin. Aikaisemmissa tutkimuksissa suoritetuissa kinemaattisissa vertailuissa avainpisteet eivät ole olleet anatomisesti tarkkoja, biomekaaniset mallit ovat olleet erilaiset menetelmien välillä, eikä markkerien vaikutusta syväoppimismallien suorituskykyyn ole huomioitu. Tämän tutkimuksen tarkoituksena oli tutkia miten markkerien läsnäolo harjoitusnäytteissä vaikuttaa mallien suorituskykyyn, kun niitä sovelletaan 3D-liikeanalyysiin ja verrataan optoelektroniseen järjestelmään.

18 koehenkilöä käveli ja juoksi juoksumatolla eri vauhdeilla samalla, kun heidän liikkumistaan tallennettiin kahdella kahdeksankameraisella järjestelmällä: optoelektroninen Vicon-järjestelmä (300 Hz) ja GoPro-järjestelmä (60 Hz). Kaksi syväoppimismallia kehitettiin GoPro-järjestelmällä tallennettujen harjoitusnäytteiden perusteella siten, että toiseen käytettiin näytteitä, joissa markkerit olivat läsnä, ja toiseen näytteitä ilman markkereita. Kävely- ja juoksunäytteet analysoitiin molemmilla malleilla ja kinemaattisia tuloksia verrattiin tuloksiin Vicon-järjestelmästä. Lisäksi, Viconilla mitattujen kinemaattisten muuttujien toistettavuus samalla koehenkilöllä laskettiin, jotta yksilön askelten välinen vaihtelu voitiin määrittää.

3D-liikeanalyysi epäonnistui mallien heikon suorituskyvyn takia, mutta sagittaalitason 2D-liikeanalyysi voitiin suorittaa oikean jalan nilkalle ja polvelle. Markkerien läsnäololla ei ollut selkeää vaikutusta mallien suorituskykyyn. Mallien suorituskyky oli lähes yhtäläinen markkerien ollessa läsnä näytteissä, eikä kumpikaan kyennyt kelvollisesti analysoimaan näytteitä, joissa ei ollut markkereita. SPM-analyysi paljasti, että menetelmien välillä havaittiin yksi tilastollisesti merkitsevä joukko polvessa ja kahdeksan nilkassa. Toistettavuus oli suurimmassa osaa näytteitä riittävä kliinisiin mittauksiin (ICC > 0.9), mutta pisteittäinen analyysi osoitti toistettavuuden vaihtelevan askelsyklin aikana. On mahdollista, että syväoppimismalleja voidaan tulevaisuudessa käyttää 3D-liikeanalyysissä, jos olosuhteet ovat asianmukaiset. Tässä tutkimuksessa markkerien läsnäololla harjoitusnäytteissä ei ollut vaikutusta syväoppimismallien suorituskykyyn sovellettaessa 2D-liikeanalyysiin.

Avainsanat: syväoppiminen, liikeanalyysi, kinematiikka, luotettavuus

LIST OF COMMON ABBREVIATIONS

2D	two dimensional
3D	three dimensional
AI	artificial intelligence
CGM	Conventional Gait Model
CNN	convolutional neural network
DL	deep learning
DLT	direct linear transformation
FLIC	Frames Labeled In Cinema -dataset
GCS	global coordinate system
GOM	global optimization method
GPU	graphics processing unit
HPE	human pose estimation
LCS	local coordinate system
LSP	Leeds Sport Pose -dataset
ML	machine learning
MPII	Max Planck Institute for Informatics -dataset
STA	soft tissue artefact

TABLE OF CONTENTS

1 INTRODUCTION

For over a century, humans have captured motion by varying techniques and tried to quantify its nature. The techniques have improved as the technical fields like photography, electronics and computer science have made advances in their respective fields. After the use of television systems became popular in movement sciences in the late 1960s and optoelectronic systems began spreading commercially in 1980s, human movement analysis became relatively accurate, although the work was quite laborious. Today, motion analysis is automated to a large extent and a myriad of different movements have been studied in humans and animals.

Marker-based systems are the current gold standard in 3D motion analysis, where the human body is modeled by tracking reflective markers attached to specific landmarks on the skin. Although giant steps have been taken forward in preparation and processing times, the motion is hampered by the markers and subjects are forced to move in confined laboratories. Alternative vision-based systems have been explored already in the 20th century (Moeslund et al. 2006), but not until after recent advancements in computer vision and deep learning (Krizhevsky et al. 2012), the markerless motion analysis has improved markedly.

Deep learning -based human pose estimation methods could provide noninvasive and flexible ways to analyse human movement outside laboratory settings and with more affordable equipment. Recent studies have shown that these methods are comparable to marker-based systems in measuring kinematic variables during different movements (Moro et al. 2020; Nakano et al. 2020; Stenum et al. 2020; Van Den Bogaart et al. 2020; Zago et al. 2020). However, these methods are rather task-specific and sensitive to the data they are trained on. Furthermore, in many such approaches the body model used isn't biomechanically valid, or they are trained on public datasets which might not have anatomically relevant annotations. Therefore, this thesis aims to gain more data on how training data affects the predicting performance of a deep learning -based human pose estimation method with proper body model.

2 MARKER-BASED MOTION ANALYSIS

Motion analysis has enabled quantification of even complex movements and made possible to answer very precise questions about movement kinematics. Kinematic data is numerical information about movement of an object, where external or internal forces experienced by the object are not of interest. The movement is thus observed without regard to what's causing it. However, the kinematics can be used in inverse dynamics process as inputs for computing forces and moments acting across joints. (Robertson et al. 2014, 9.) In human movement, kinematic variables describe positions, velocities and accelerations of body segments, and analogous angular variables between segments. These kinematic variables are based on spatiotemporal motion of certain landmarks on segment surfaces, which are usually indicated by tracking markers and identified with camera systems. (Medved 2001, 47). This chapter summarizes the basics of how motion capture systems are used to estimate 3D poses of body segments and their kinematics. Also, a brief history of selected works in the field of motion analysis is presented.

2.1 History of motion analysis

After the invention of photography and Muybridge's work capturing people and animals in action in the 19th century (figure 1), significant advances were made in quantifying locomotion (Allard et al. 1998, 12; Jarrett 1976, 10 – 12). Braune and Fischer in 1895 (Allard et al. 1998, 15 – 16 & 87) were the first to perform extensive 3D-analysis of human gait with one subject, four cameras and chronophotographic technique. The preparation of their subject alone, took 6 – 8 hours, which speaks of the great methodological leaps taken in the next century. Television systems for movement analysis became popular in the late 1960s, where coordinates of detectable points could be identified from television scan lines (Allard et al. 1998, 88; Jarrett 1976, 47 – 48).

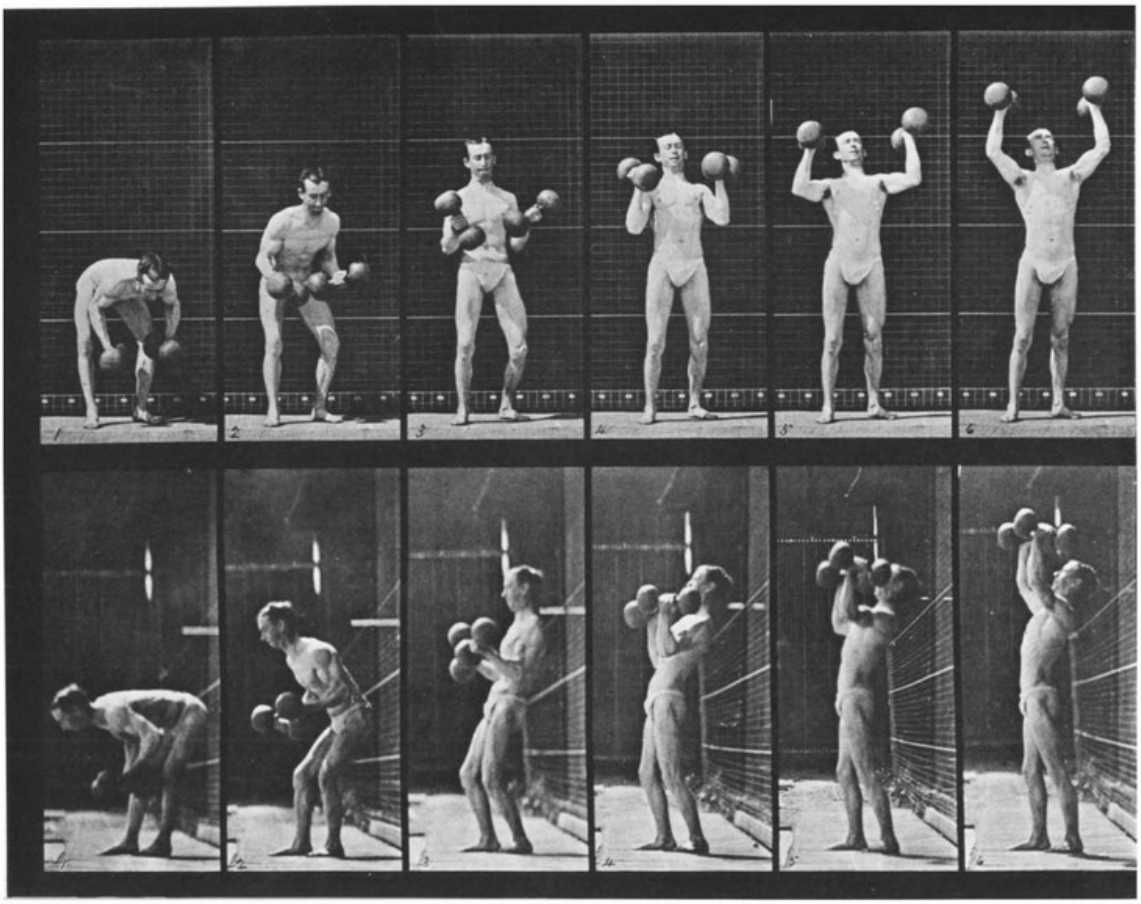


FIGURE 1. A man lifting dumbbells (Muybridge 1955). Originally, picture was published in Muybridge's Animal Locomotion (1887) by University of Pennsylvania and taken between 1872 – 1885.

Dinn et al. (1970) published one of the first computer interfaces (CINTEL) used in human motion studies, which could convert analog television signal to digital form differentiating up to 32 different point intensities. The CINTEL system was used for studying gait in crippled children (Winter et al. 1972, according to Winter 2009, 60 – 61 & Jarrett 1976, 51 – 52), where reflective hemispheres of ping pong balls were used as markers and intensity resolution was one bit (black and white). The potential of TV-computer -systems were later further demonstrated with another CENTIL-based research (Winter et al. 1974, according to Jarrett 1976, 52), where the authors were able to automate data analysis and study walking gait in healthy subjects with relative ease. Another television-based system was developed in 1974 at University of Strathclyde (Allard et al. 1998, 88) and further refined to include multiple cameras

for 3D-analysis (Jarrett 1976). This system was later adopted by Oxford Medical Systems, and yet improved to become the first commercial television/computer-based motion analysis system Vicon in 1980 (Allard et al. 1998, 89; Sutherland 2002; Winter 2009, 61).

In modern motion analysis, optoelectronic systems are the most popular (Chiari et al. 2005; Robertson et al. 2014, 12) and accurate (Van Der Kruk & Reijne 2018) systems. The general working principle in optoelectronic systems is the conversion of incoming electromagnetic radiation (light) to voltage (Jarrett 1976, 32; Medved 2001, 57). In modern optoelectronic systems, this is done by optical sensors, like charge-coupled devices (CCD) and position-sensitive devices (PSD) (Medved 2001, 70 & 87; Allard et al. 1998, 88 – 89). First PSD-based optoelectronic imaging system for studying movement was the Swedish SELSPOT, which was able to determine 2D-positions of multiple infrared LEDs and self-identify each marker (Lindholm 1974; Jarrett 1976, 34 – 36; Allard et al. 1998, 89). One major advantage of PSDs was their nonaddressable sensors: scanning the entire image wasn't necessary as was with television systems (Medved 2001, 87).

2.2 Extracting 3D-coordinates of markers

With a single camera it's possible to capture movement in 2D, but this restricts observations to one plane and it's critical to place the camera correctly with respect to the movement observed. Human movement is rarely confined to a single plane, although in some movements one plane can contain most of the useful information. Multiple cameras can be used to gather 3D kinematics and more accurate description of a movement. To capture 3D movement, it's necessary to have line-of-sight to every marker from at least two cameras. Usually, imaging systems comprise of more than two cameras to factor in the inevitable visual blocking of markers during human motion. (Robertson et al. 2014, 12 – 15.) Contrary to 2D analysis, initial placement and orientation of cameras in 3D system isn't as exact, since a calibration object will refine the system parameters (Alem et al. 1978).

Markers used in motion capture can be passive or active. Passive markers are less invasive for the subject and coated with retroreflective paper to reflect incoming light to the sensor. Passive

spherical markers are visible from every direction and aren't limited by rotation. Most systems illuminate the markers themselves (infrared LEDs), but in principle any light is sufficient. In passive systems the sensors see the markers simultaneously, which requires marker identification with labelling software. Linear and higher-order extrapolation techniques have been used from the late 1960s to identify markers in successive frames (Medved 2001, 75 – 76). In contrast, active markers produce the light themselves and require a power supply. They mostly work in the near-infrared wavelengths and in time-multiplexed mode. Time-multiplexity means that the markers activate for short period of time in an orderly manner, which makes marker identification automated. Disadvantages of active markers are that they are rotationally constrained because the light-emission angles are restricted in LEDs and they usually require wiring for power. (Allard et al. 1995, 59 – 61; Robertson et al. 2014, 12.)

Threshold detection is a popular method to recognize markers in an image. In essence, all pixel values that are higher than the threshold are detected as a marker. The threshold must be adjusted such that the number of false positives (reflections, bright objects etc.) is low, but all the markers are detected. (Allard et al. 1995, 62 – 63.) In the earlier described CINTEL-system, the marker center was determined by calculating the mean of all coordinate points that were occupied by one marker. With large enough markers encompassing approximately 10 points, accurate spatial resolution of 1 mm could be achieved (Medved 2001, 72 – 74.) Centroid of a marker can also be found by detecting the edges of the marker and then using iterative fitting to find the center of a circle. This method is used in the Vicon system. (Medved 2001, 79 – 80.) Ferrigno & Pedotti (1985) presented a marker detection algorithm based on shape recognition. In their approach, a grayscale input image (4-bit) is cross-correlated with a predetermined mask to identify markers. Centroids are then calculated based on the cross-correlation values instead of actual pixel intensities.

Cameras can record 2D projections of 3D objects, therefore, the goal of 3D-analysis is to reconstruct the 3D object from multiple planar projections. This process is called photogrammetry or stereometric method. (Allard et al. 1995, 9; Medved 2001, 50 – 51.) The calculation of 3D coordinates with photogrammetric reconstruction requires knowledge of specific optical parameters of the cameras. This information might not be available or can be hard to measure, therefore, algorithmic approaches have been developed for the 3D

reconstruction. (Allard et al. 1995, 11.) Direct linear transformation (DLT) (Abdel-Aziz & Kamara 1971) is a popular reconstruction algorithm that uses linear equations to determine the three unknown marker coordinates. DLT method assumes that there is a linear relationship between 3D coordinates of the object and it's 2D projections in the cameras (figure 2) (Robertson et al. 2014, 35 – 36). Each camera provides two equations; thus, at least two cameras are required to solve the 3D coordinates. Cameras must be calibrated beforehand, which requires solving linear equations with 11 unknown parameters. The necessary parameters are acquired by calibrating the cameras with a 3D calibration frame that is instrumented with markers. (Alem et al. 1978.) The calibration process of DLT approach allows cameras to be positioned randomly in global space in 3D-analysis (Medved 2001, 108).

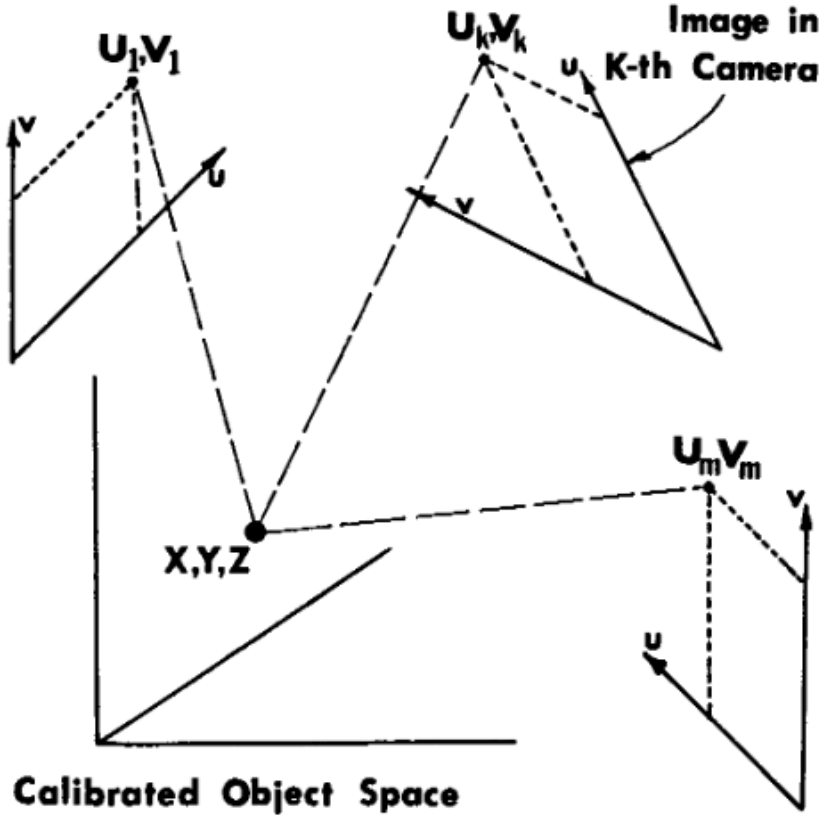


FIGURE 2. Coordinates of the 3D object are projected to image planes of each camera in DLT (Alem et al. 1978).

The accuracy of motion capture systems has been found to increase, when calibration object includes more control points (markers) and they are evenly spaced in the capture volume. Use of 3D calibration frames becomes impractical, when capture volume increases and adequate accuracy is still desired. An alternative approach for calibration is a simultaneous multi-frame analytical calibration (SMAC), originally presented by Woltring (1980), where an object with planar control points is observed by the cameras in different positions. This is more suitable calibration method for larger volumes. (Chiari et al. 2005.) Vicon systems use a wand equipped with 5 coplanar markers with known relative distances to calibrate the cameras. The calibration calculations are done by the software and the user is only needed to wave the wand inside the desired capture volume. (Vicon documentation 2020.)

2.3 Coordinate systems and pose estimation

Data collected with a motion capture system yields a set of x, y and z coordinates of tracking markers in a global coordinate system (GCS), which has a fixed origin in the laboratory space. Coordinates in GCS describe how the markers are moving through the laboratory space. Often when human movement is examined, researchers and practitioners are interested in the orientations of specific segments or relative orientation of two segments, i.e., joint angles. To determine segment orientations and joint angles, the x, y and z coordinates in GCS must be transformed to local coordinate systems (LCS), which presents the coordinates in relation to anatomical axes of segments. The LCS is fixed to the segment and it moves as the segment moves. Hence, the orientation of LCS to GCS defines the segment orientation in the 3D-space (figure 3). Conventions for naming the axes may vary between laboratories and motion analysis systems. (Winter 2009, 176 – 177; Robertson et al. 2014, 36 – 37.)

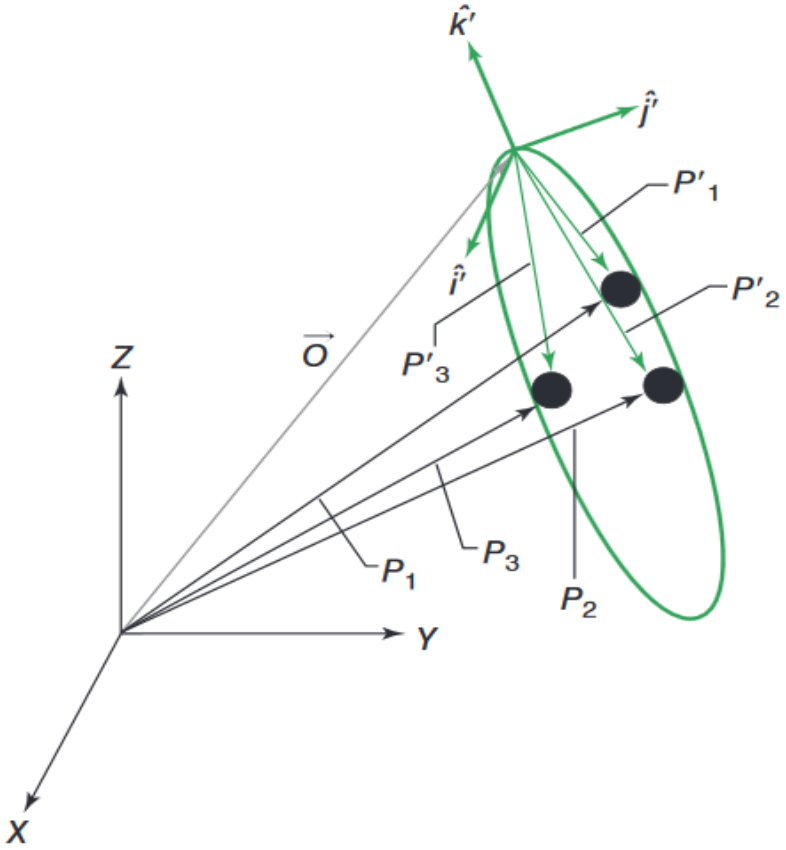


FIGURE 3. Markers (black circles) on a segment presented in GCS (*XYZ*) as P_1, P_2 and P_3, and in LCS (*ijk*) as P'_1, P'_2 and P'_3. The position of LCS origin in GCS is marked with *O*. (Robertson et al. 2014, 48.)

The LCS, or the bone-embedded frame, is defined analytically from anatomical and technical markers placed on the segment. The markers allow us to estimate positions of anatomical landmarks in the body. Sometimes markers can be placed directly on the landmarks (anatomical markers), if they are close to the skin surface. Some landmarks are more internal to the body and placing markers directly over them isn't possible. Then, technical markers can be used, which usually don't have direct anatomical relevance. (Allard et al. 1998, 131 – 136.) At least three noncollinear markers are required per segment to describe the 3D pose (position and orientation) of the segment (Winter 2009, 180). Assumption in using markers to estimate pose of a segment is that the markers are not moving relative to the bone or each other (Robertson et al. 2014, 45). Anatomical axes (LCS) can be constructed from the locations of anatomical landmarks. The Conventional Gait Model (CGM) (Baker et al. 2017) is the most popular

method for defining the LCS for studying gait and other movements, but alternatives have also been proposed (Cappozzo et al. 1995).

Conversion of coordinates between GCS and LCS happens through linear and rotational transformations. If we consider any marker in the figure 3, linear transformation from GCS to LCS is described as:

$$\overrightarrow{P'} = \vec{P} - \vec{O} \tag{1}$$

where P' and P are the coordinate vectors of same marker in LCS and GCS, respectively, and O is the vector describing the position of LCS origin in GCS. Rotational transformation between coordinate systems is computed by a rotation matrix. A rotation matrix is made up of orthogonal unit vectors that rotate the coordinate system about its origin to align itself with the other coordinate system. Usually, there is both, translation and rotation, so the full transformation from GCS to LCS is described as:

$$\overrightarrow{P'} = R(\vec{P} - \vec{O}) \tag{2}$$

where R is the 3x3 rotation matrix. (Robertson et al. 2014, 37 – 38.)

For quantification of movement, the human body with its complexities must be reduced down to idealized mechanical models. Usually, the human body is modelled as a set of rigid segments interconnected by joints, which together form a kinematic chain. (Allard et al. 1995, 3; Medved 2001, 15 – 16.) The biomechanical model selected must have adequate complexity for the modelled movement. The CGM might be the best option for analysing walking or running, but it probably isn't intricate enough to study knee ligament forces with appropriate accuracy. Additionally, mechanical centers of rotations of human joints can move during movements (Smith et al. 2003), which should be considered if particularly accurate results are desired.

Rigidity of markers relative to bone is an ideal assumption, and in reality, markers attached to the skin can translate and rotate relative to the underlying bone, causing errors in the pose

estimation (Cappozzo et al. 1996). This soft tissue artefact (STA) should therefore be taken into account when skin marker -based modeling is performed. If body segments are treated separately during biomechanical modeling, it leads to apparent dislocation of joints mainly caused by STA. (Lu & O´Connor 1999.)

Lu & O´Connor (1999) presented a global optimization method (GOM) for pose estimation, which included joint constraints and decreased the STA errors in modeling compared to other optimization algorithms. The method searches for an optimal pose that minimizes the difference between measured and model-based marker coordinates in a least squares sense. The GOM can be tailored to each subject by performing a static calibration, where skin movement can be assumed to be absent, and thus the calibration data can be used as a reference frame. The method allows for setting varying weights on each marker, depending on how susceptible the marker is to STA. Lower weights are usually assigned to markers that are prone to STA. (Lu & O´Connor 1999.)

Computer-simulated gait trials were performed to compare the accuracy of the GOM to other methods. Joint constraints forced adjacent segments to stay together in the global optimization method and therefore the problem of joint dislocations was removed. Other methods that didn't include joint constraints had approximately 1 to 3 cm of joint dislocations in the hip and knee joints, measured by the distance between adjacent segment end points. The GOM performed best out of other three in locating the joint centers of hip, knee and ankle relative to absolute truth, especially in hip joint. When analyzing joint angles, the GOM improved the accuracy most notably in frontal and transverse planes. (Lu & O´Connor 1999.)

Joint angles and their derivatives are the variables that many utilizers of research data want to know and apply in their occupation. The relative orientation of one segment LCS to another defines the joint angle between segments. The Cardan system is commonly used in biomechanics to define the relative 3D orientation of frames, where rotations happen in x-y-z -sequence. First, the coordinate system is rotated about the x-axis, followed by y- and z-axes, in that order. This is one method to compute the 3D angle between the segments. Planar joint angles are relatively easy to comprehend, but a 3D angle is more abstract and will not be

untangled here. One point that should be pressed is that 2D projected angles don't equal to a 3D angle. (Robertson et al. 2014, 50 – 54.)

2.4 Strengths and limitations of marker-based motion analysis

Errors in the computed segment pose arise from two sources: instrumental errors and experimental artefacts. Instrumental errors originate from the photogrammetric system and is tied to the used instruments. Experimental artefacts are caused by a particular method and one of the most prominent errors in this category is the STA. (Cappozzo et al. 1996.) The effect of STA wasn't noticed or regarded as significant factor in biomechanics when marker-based motion analysis began in 1960s. It was still underestimated in the 1980s, even though it was recognized to occur during human movement. In the 1990s the issue of STA started to get proper attention. Cappozzo (1991) has stated about STA that it's "a major source of error in the experimental procedures using optoelectronic systems". (Camomilla et al. 2017.)

Motion of the bones is generally the focus of human movement researchers, which has been methodological problem from early on, since we're unable to observe bone movement directly. Therefore, the position and orientation of an assumed rigid body segment is usually based on skin markers on the segment surface. Filtering can be used to eliminate instrumental errors but not STA, because it has mostly the same frequency content as the bone movement. Therefore, the positions of the markers on the skin should be selected such that the unwanted movement is minimized. Also, optimization algorithms (e.g., Lu & O´Connor 1999) can mathematically reduce the amount of STA. (Cappozzo et al. 1996.)

Cappozzo et al. (1996) demonstrated that STA in lower limb skin markers varies from few millimetres to even 4 cm during walking and cycling. Markers attached close to anatomical landmarks near the joints were most susceptible to higher STA, and thus, these locations were deemed as unsuitable. Markers located further away from joints in the lateral side may be expected to exhibit lower STA. The amount of STA was also dependent on the joint angles. The effect of STA on the inaccuracies in knee joint kinematics were found to be much more pronounces in frontal and transverse planes than in the flexion-extension direction.

Richards (1999) critically reviewed the accuracy of the most significant motion capture systems by field tests with SAMSA device, which has seven markers attached to it with known relative distances. More recently, Topley & Richards (2020) compared current higher-resolution motion capture systems with same tests with SAMSA device in laboratory environments. The review included relatively new models from popular manufacturers: Vicon, Qualisys, OptiTrack and Motion Analysis Corporation. Errors in measurements were clearly smaller in modern systems compared to results from 1999. Absolute average errors when measuring the distances between two markers on SAMSA device were all under 0.6 mm in modern systems compared to over 1 mm in older systems. Variability (standard deviation) in the measured distances were only around 0.1 mm with modern systems compared to clearly over 1 mm in older systems. Also, modern systems could perform tracking in near real time. Generating coordinate data might have taken from minutes to even several hours in 1999. Topley & Richards (2020) demonstrated in their review that current motion capture systems can track markers on a moving body very precisely in a confined laboratory setting. The authors also conclude that errors associated with modern system hardware and software are relatively small compared to errors originating from STA and incorrect marker placement.

In conclusion, marker-based motion analysis is accurate but rather restrictive tool for motion analysis. Accurate pose estimation requires expensive equipment and time-consuming subject preparation, and still, the movement must be performed inside a specified capture volume with external markers attached to the skin. Marker-based motion capture has also been mostly constrained to laboratory settings for strict control of environmental factors. Although in recent years the systems have evolved and motion capture in outdoor settings has been performed (Colyer et al. 2018).

2.5 Kinematic analysis of walking and running in humans

The word gait is used to describe the manner or style of walking and running, which will vary between individuals. Whittle (2007, 48) defines walking and running together as *"a method of locomotion involving the use of two legs, alternatively, to provide both support and propulsion"*. To differentiate running from walking, he points out that in walking there always

must be at least one foot in contact with the ground. While this is true for normal gait, it might not hold for some forms of pathological gait. Furthermore, "normal" gait is always relative to the group in question and even normal gait varies stride-to-stride. (Whittle 2007, 47 – 48.)

Prior to automated human movement measurements, the marker locations were digitized by hand from cine film. It was very laborious but still promising for the analysis of gait and other movements. The real breakthrough came after Hans Furnée, PhD, started developing automated systems for recording marker positions in the Netherlands around 1967. Furnées work motivated John Paul, PhD, to occupy his students to work with a 3D system for movement analysis (Jarrett 1976), which later evolved to commercial Vicon system and marked the beginning for faster and more practical clinical gait analysis. (Sutherland 2002.)

Observing gait while on a treadmill allows researchers to control gait speed and other variables more easily. It's also faster to gather data with treadmills and they might even be instrumented with force plates to gather reaction forces. Even though studying gait is more convenient on a treadmill, it might affect the gait compared to overground gait. The subject might shorten their stride on a treadmill compared to overground, if the belt is small. Additionally, the initial contact can decelerate the belt and push off can accelerate the belt, if the motor of the treadmill isn't powerful enough. (Whittle 2007, 133.)

Reliability in gait analysis means the extent to which the gait measurements are free from variation. When gait is analysed before and after an intervention, the difference in the gait variables might be present for two reasons: the intervention had a real effect to gait or there is measurement variation. (McGinley et al. 2009.) Variability or error in gait data can arise from intrinsic and extrinsic sources. Intrinsic variability (stride-to-stride) occurs naturally as the participants strides aren't perfectly identical every time. Intrinsic variability is free of methodological errors and it can only be observed and managed. Extrinsic variability can be caused by intra- or inter-observer differences, e.g., in palpation of anatomical landmarks and anthropometrics measurements. In hierarchical biomechanical models, the errors propagate "downstream", usually from pelvis towards the ankles in lower body models. (Schwartz et al.

2004.) Knowing the variability in gait is essential to make valid interpretations from the data and avoid type I and II errors.

DeVita & Bates (1988) found that at least 25 ground contacts were required to obtain stable mean parameters for ground reaction force data in running. They concluded that multiple trials are needed to measure subject's mean performance in repetitive tasks, and the number of trials will probably vary between activities. Diss (2001) demonstrated that inter-session reliability for kinematic variables was higher with five trials compared to one in running. Similarly, Monaghan et al. (2007) pointed out that larger number (10) of gait trials improved the intra-rater reliability in walking between two sessions on separate days. Schwartz et al. (2004) measured the stride-to-stride variability using 5 strides in walking gait from two subjects and showed that the peak standard deviation during the gait cycle was below 2° for all except sagittal knee (~3°) and ankle (~2.5°) motion. Riley et al. (2008) found out that 10-12 strides were enough to reach stable mean in kinematic and kinetic peak values during treadmill running. McGinley et al. (2009) concluded in their review that the appropriate reliability in gait kinematics must be assessed relating to the intended application. However, they present that less than 2° can be considered acceptable, 2 – 5° is reasonable and over 5° should raise concern. The values they present are drawn from various studies, which present the absolute variation as either standard deviation or standard error. Today, normative data from different research centers can be expected to be compatible, if standards are followed in coordinate system definitions, marker placements and other procedures that might cause variability (Baker 2013, 180).

Conventional Gait Model is a name for a group of biomechanical models that emerged in 1980s (Baker et al. 2017). There are many implementations of the CGM (e.g., Newington, Helen Hayes and Plug-In Gait) and it is by far the most used model in clinical gait analysis. There are alternative models (e.g., Cappozzo et al. 1995) but there isn't convincing evidence that they are better than the CGM. (Baker 2013, 29.) The CGM has seven segments; femur, tibia and foot for both sides and one pelvis, which are linked by ball joints in a hierarchical manner, where pelvis is highest in the order. Joint angles in CGM aren't calculated directly from marker locations, but from estimated joint centers. (Baker et al. 2017.) The CGM marker set is illustrated in figure 4 and detailed guidance for marker placements and definition of segment

models is described in Baker (2013, 30 – 43). When the subject is wearing shoes and palpation is not possible, the foot marker should be place on the midline of the forefoot section. Also, the line between heel and foot marker should be parallel with the pitch angle of the shoe. (Baker 2013, 43 – 44.) Misplacing markers on the skin will change some of the joint kinematics because the marker locations are directly affecting the calculated joint centers (Baker 2013, 152 – 159).

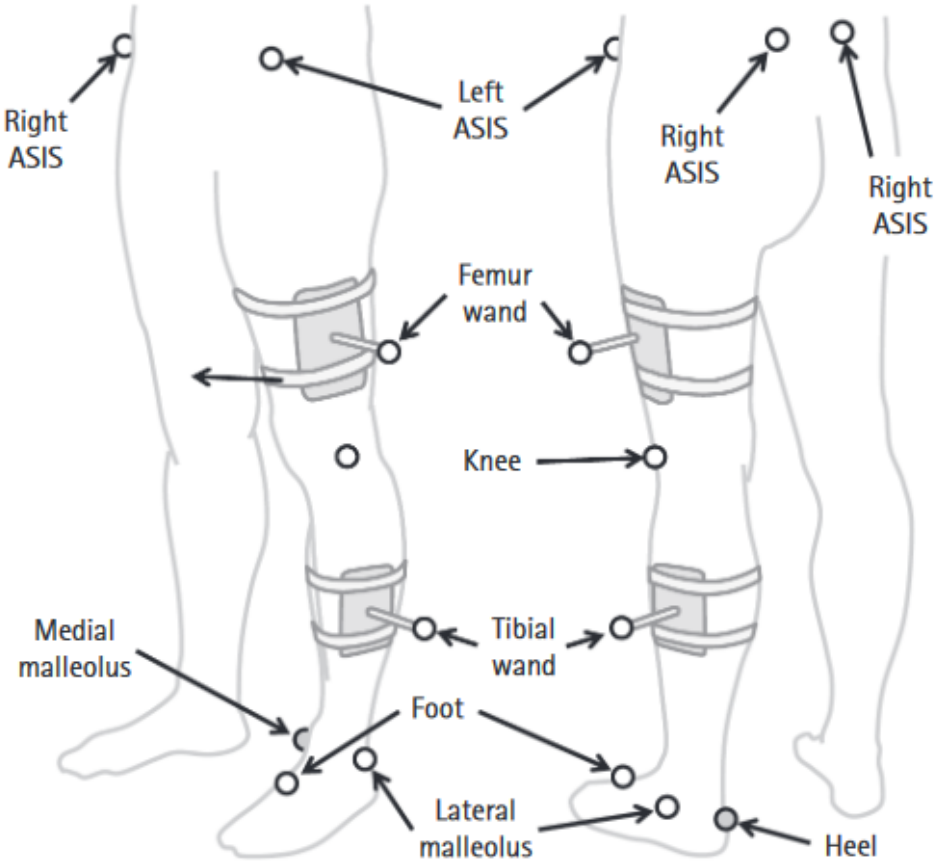


FIGURE 4. Markers required for the CGM. For clarity, only markers on the left limb are shown since right side is similar. Grey markers indicate markers that are only needed for static calibration. (Baker 2013, 46.)

3 NEURAL NETWORKS AND DEEP LEARNING

The visual cortex is believed to process information at increasing levels of abstraction. The standard model suggests that the brain starts simple by first recognizing edges, and then moving towards more complex shapes. (Murphy 2012, 995.) It is suggested that such deep architectures are required in the field of artificial intelligence (AI) to solve complicated tasks, like object or speech recognition (Bengio 2009). Machine learning (ML) is a subfield of AI, risen from the need to automate data analysis in the era of big data. ML is defined as a set of methods to predict future data by detecting patterns in existing data sets. (Murphy 2012, 1.) Deep learning (DL) is a special subfield of ML, where an algorithm can learn complicated concepts without human operator specifying exactly what to do (Goodfellow et al. 2016, 1). With mathematical functions and appropriate training data, DL algorithms can learn to detect patterns in data and perform a specific task accurately (e.g. recognizing handwritten digits or locations of joint centers) (Nielsen 2015).

In principle, a neural network "learns" by first providing the network training data as examples with correct labels, then allowing it to train with new data, and finally giving it feedback about how it performed labeling the new data. The network changes its behavior based on the feedback and then the training process can be repeated. These networks have specific structure and properties, which allow them to change their response to some input. The internal properties of these networks are modified with automated algorithms, like gradient descent and back-propagation, which allow the networks to be trained with large amounts of data and learn to perform seemingly intelligent tasks. The next sections describe briefly how neural networks are structured and how they are able to learn.

3.1 Structure and function of neural networks

A neural network is made of artificial functional units called neurons and connections between them (figure 5). Neurons are organized to layers and in feedforward neural networks output from one layer is used as an input to the next layer. In general, a neuron receives multiple numerical inputs, processes them and outputs a single numerical value. Layers are organized

hierarchically in the network. First layer of a neural network is the input layer and activation of the input neurons is essentially the data fed to the network. For example, in computer vision, the input data could be color intensities of single pixels in an image. Last layer of a network is the output layer. Activation of neurons in output layer determines the outcome of the neural network with specific input parameters. The activations of neurons in the output layer could represent probabilities of the input image belonging to a certain class, like a dog, or probability of specific joint being present in the pixel. Layers between input and output layers are called hidden layers. The number of hidden layers and neurons in them define the complexity of relations between input and output. (Nielsen 2015.) Training data defines what the neural network should output with given input, but not what other layers should do. The learning algorithm must define the behaviour of hidden layers to produce the correct output. Goodfellow et al. (2016, 165) argue, that because the correct output provided by training data is not given directly to these layers, they are called hidden layers.

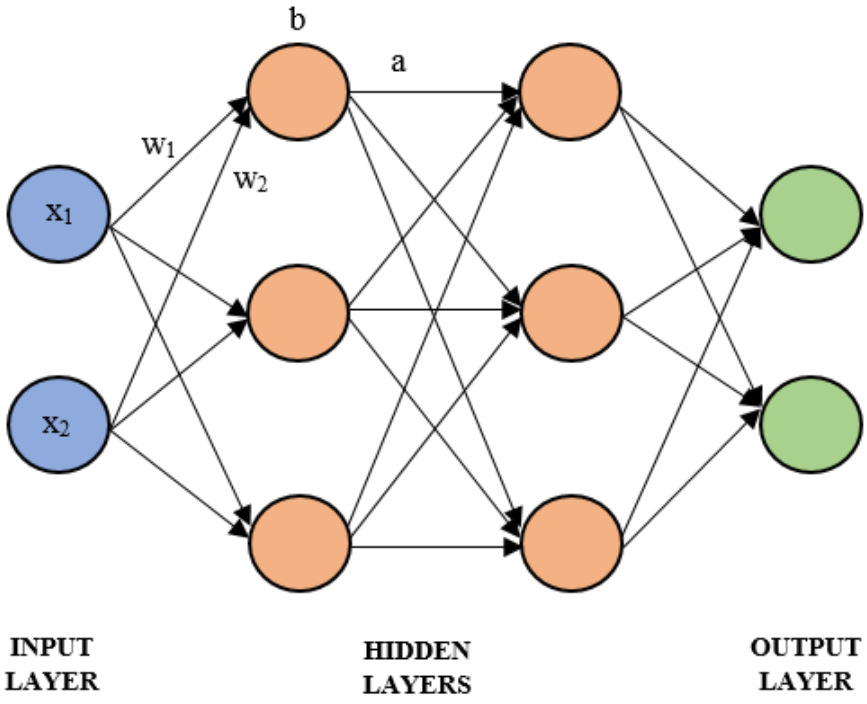


FIGURE 5. A neural network which has two neurons on input and output layers, and three neurons on two hidden layers. Inputs to the network (x_1 and x_2) and weights (w_1 and w_2), bias (b) and activation (a) of one neuron on first hidden layer are shown.

The neural network model is a function with set inputs and desired outputs controlled by adjustable parameters. By adjusting these parameters (weights and biases), output of the network can be modified to fit the desired task. (Bishop 2006, 228.) Functional units in the context of neural networks are the neurons. If we consider one neuron from figure 5 (top neuron on the first hidden layer), it has two inputs (x_1 and x_2), one weight (w_1 and w_2) for each input and a single bias (b). Output of a neuron is called its activation (a), which is usually fed forward as an input to the next layer. By Nielsen (2015), weights are real numbers that express the importance of the input it's tied to, and the bias can be thought as a measure of how easy it's to activate the neuron. Activation of a neuron is defined by an activation function $f(z)$, where z is defined as the sum of weighted sum of inputs and the bias as follows (Bishop 2006, 227):

$$z = \sum wx + b \tag{3}$$

It's worth noting that z isn't the activation of the neuron, but rather the input to the activation function $f(z)$, which outputs the activation, a.

Activation functions used in feedforward neural networks are nonlinear functions. If linear functions would be used as activation functions, the output of the network would be linear function of its input. Nonlinear functions are more versatile to describe complex relationships. (Goodfellow et al. 2016, 168.) The activation function used depends on the data and assumed distribution of output variables, but there isn't a rule of thumb for picking the right function. (Sharma et al. 2020). Activation functions are generally chosen to be sigmoidal, which refers to the function having an "S"-shaped curve. (Bishop 2006, 227 – 228). The logistic sigmoid function outputs a value between 0 and 1 based on the parameter given to the function (figure 6). If the parameter given to the logistic sigmoid function is very positive or negative, the activation of the neuron approaches 1 or 0. Even a relatively big change of the parameter value at the "tails" of the function affects the activation of the neuron only slightly. In contrast, small changes in the parameter value can have significant difference in the activation, if the parameter value is close to 0. (Nielsen 2015.) The form of logistic sigmoid function is:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{4}$$

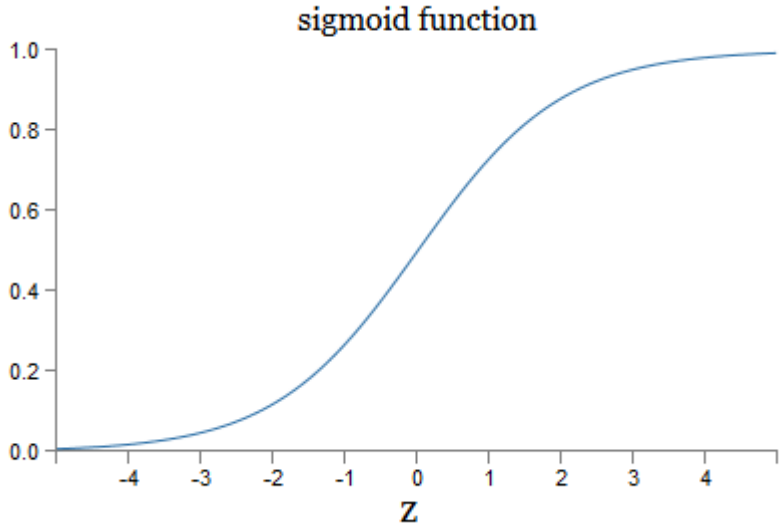


FIGURE 6. The logistic sigmoid function, which outputs a value between 0 and 1. The parameter *z* is calculated from equation 3. (Nielsen 2015.)

Sigmoidal functions perform well in classification problems but suffer from a vanishing gradient problem where learning slows down significantly because the slope of the function becomes very small at the tails of the function. ReLU (rectified linear unit) activation function is also widely used and performs better than others in many cases (figure 7). ReLU is more efficient than other activation functions since all the neurons aren't always activated simultaneously. However, the zero derivative when input is negative might be problematic in some cases. Softmax activation function is a combination of multiple sigmoid functions and used most often in the output layer of a neural network. It is used in multiclass classification problems where for each class there is one neuron in the output layer representing the probability of the input belonging to the specific class. Some other examples of activation functions are Binary Step Function, Tanh, Leaky ReLU and Swish. (Sharma et al. 2020.)

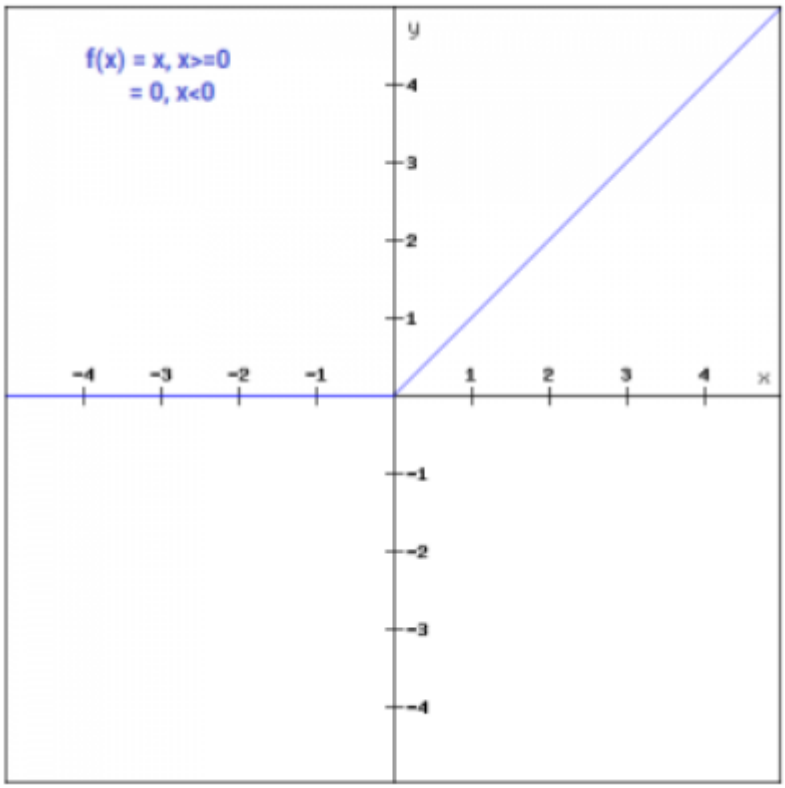


FIGURE 7. ReLU activation function where $y = x$, when $x \geq 0$ and $y = 0$, when $x < 0$. The y parameter represents the activation and x is the input to the function (equation 3).

3.2 Learning of neural networks

The end goal of a feedforward neural network is to find weights and biases that allow the network to perform well at a given task. For example, consider that the task is to recognize all images with a cat in them, and some function $y = h(x)$ does that. For every image (x) given to the function as a parameter, it outputs $y = 1$ for all images including a cat, and $y = 0$ if there isn't a cat. Neural networks goal is to define another function $y = g(x, w, b)$ that approximates the correct results given by function $y = h(x)$, where w and b are the weight and bias vectors of the network, respectively. By manipulating weights and biases, the neural network can learn to output $y = 1$ or $y = 0$ depending if there is a cat. (Goodfellow et al. 2016, 164 – 165).

To make correct adjustments to weights and biases, the learning algorithm must evaluate how well the network currently performs and how to improve it. This process usually involves

minimizing a cost (also loss or error) function (Goodfellow et al. 2016, 80). The mean squared error is a simple example of a cost function, which is the average squared difference between some desired and measured outputs. Going back to previous example of recognizing cat images, the algorithm would sum over all squared differences $(h(x) - g(x, w, b))^2$, and take the average of that to represent the mean error of the model. Using mean squared error as a cost function usually causes the learning process to slow down at some point, but this can be avoided by using different functions. One such a function is called cross-entropy loss function, which is regularly used in DL. (Nielsen 2015.) Cross-entropy is a method to quantify the difference between two probability distributions, which in case of neural networks are the desired and actual outputs of the network (Goodfellow et al. 2016, 174.)

The technique often implemented to minimize a cost function is called the gradient descent. The gradient of a function is determined by taking the derivative of the function at specific point. When a function only has one parameter, the gradient is the slope of the function. In case of multiple parameters (even thousands in neural networks), the gradient is defined by partial derivatives, which quantify how the function is changing when only one parameter is changed. The gradient defines which way we must move to decrease the function. Since we want to decrease the cost function, we have to actually move in the direction of negative gradient. An important factor affecting the learning speed and performance of the network is the learning rate. It is a scalar value, which determines how big of a step the algorithm takes in the direction of the negative gradient (figure 8). (Goodfellow et al. 2016, 80 – 84.)

Overfitting happens when the cost of the network decreases, but the performance (ability to classify images or detect keypoints) of the network stalls. This is a significant problem especially for modern networks, which often have numerous weights and biases. If there is limited amount of training data, the network might adapt too strongly to specific input patterns. More training data usually reduces overfitting. (Nielsen 2015.) Data augmentation can be used to increase training data artificially via geometric transformations and image manipulations to original images (Mathis et al. 2020). A regularization technique called “dropout” can also reduce overfitting. The dropout technique randomly omits the output of a neuron with a set probability and prevents neurons relying on other neurons too strongly. (Hinton et al. 2012).

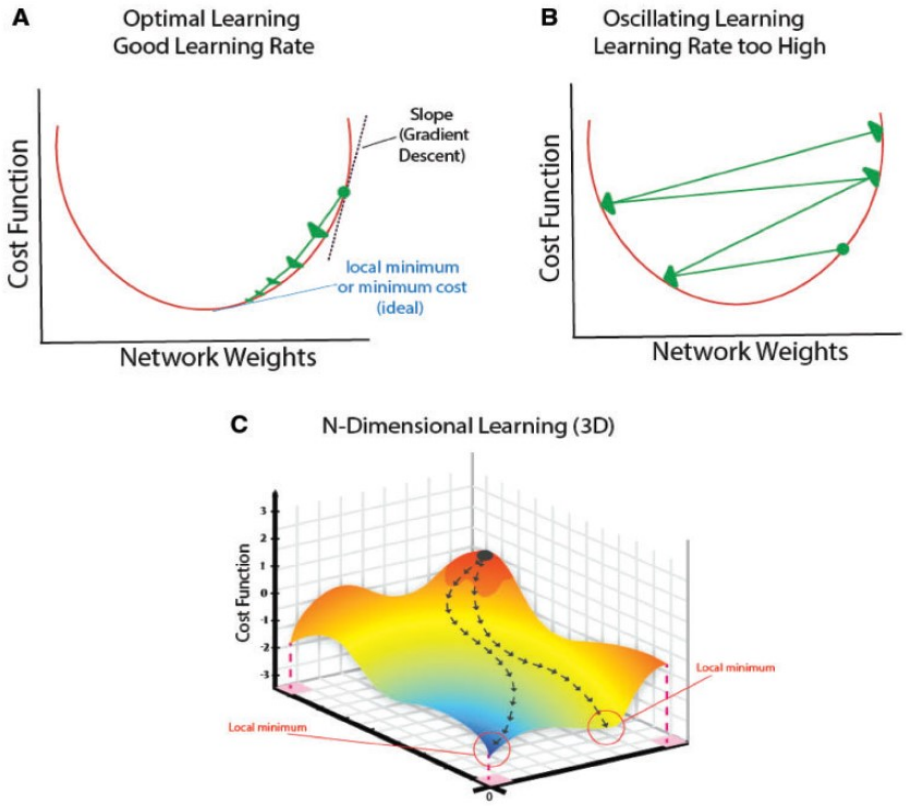


FIGURE 8. Gradient descent and the effect of learning rate visualized in 2D (A and B) and 3D (C). With appropriate learning rate, the algorithm approaches the local minimum of the cost function (A). With too high learning rate, the algorithm can remain oscillating around the minimum (B). (Murphy 2012, 247.) The cost function might have multiple minimas and the ”route” of the algorithm during minimizing process can differ depending on the parameters, such as the learning rate (C). Figure: Krittanawong et al. (2019).

Modern deep neural networks can have multiple layers, which might be formed from thousands of neurons per layer. Thus, the computational power required becomes very large to compute the cost and gradient of the network multiple times during the training process. An algorithm called back-propagation uses relatively simple and inexpensive procedure to solve derivatives of any function, and it is widely used for computing the gradient in multi-layer neural networks (Goodfellow et al. 2016, 200). As the name suggests, the algorithm progresses backwards in the network: it starts from the output of the network and works backwards visiting every neuron to obtain the gradient of the cost function (Bishop 2006, 244).

In brief, the learning process of a feed forward neural network proceeds in the following way:

1. The network is provided with examples of correctly labelled training data.
2. The network is given new data, which results an output from the network.
3. Cost of the network is calculated with appropriate cost function.
4. Gradient of the cost function is defined with back-propagation algorithm.
5. Weights and biases are modified with gradient descent algorithm to decrease the cost function.
6. Repeat steps 2 to 5 or stop training the network.

3.3 Convolutional neural networks

Convolutional neural networks (CNN) are specialized networks, which have been extensively used in image recognition (Krizhevsky et al. 2012; Touvron et al. 2020). The progress in human pose estimation in the recent years is mostly thanks to success of CNNs (Desmarais et al. 2020). Previously described learning process steps can be applied to CNNs with slight modifications (Bishop 2006, 268 – 269). Layers in neural networks described earlier are fully connected (Bishop 2006, 361), which means that each neuron has a connection to all neurons in the preceding and following layer. If input data to a network with fully connected layers are pixel intensities, this would mean that far off pixels affect each other the same way as neighbouring pixels. This architecture ignores spatial structure in data and here CNNs can make a difference.

A CNN has at least some convolutional layers, but it might also have fully connected layers. A convolutional layer is composed of feature maps, which can conceptually be thought as 2D-grids, much like an image is 2D-grid of pixels (figure 9). A single unit (neuron) on a feature map in the first hidden layer is defined by a small subregion of the input image. This subregion is called the local receptive field (LRF) and each unit in a feature map has same shared weights and a bias. If the LRF is a 3x3 grid, then the feature map has 9 shared weights and one shared bias. The LRF scans over the whole input image in specified steps and outputs activations to one feature map. As stated, each feature map in a convolutional layer is defined by individual set of shared weights and a bias. This architecture enables the CNN to detect different patterns

across the whole image and the information doesn't disappear, if the pattern is shifted to different part of the image. Subsequent convolutional layers have the same principle, but the LRF is now applied to feature maps in previous layers. (Bishop 2006, 267 – 269). It's also common for CNNs to have pooling layers, which simplify the feature maps. For example, a max pooling layer would pick the highest activation in a small grid (e.g. 2x2) and disregard the rest (Murphy 2012, 1005).

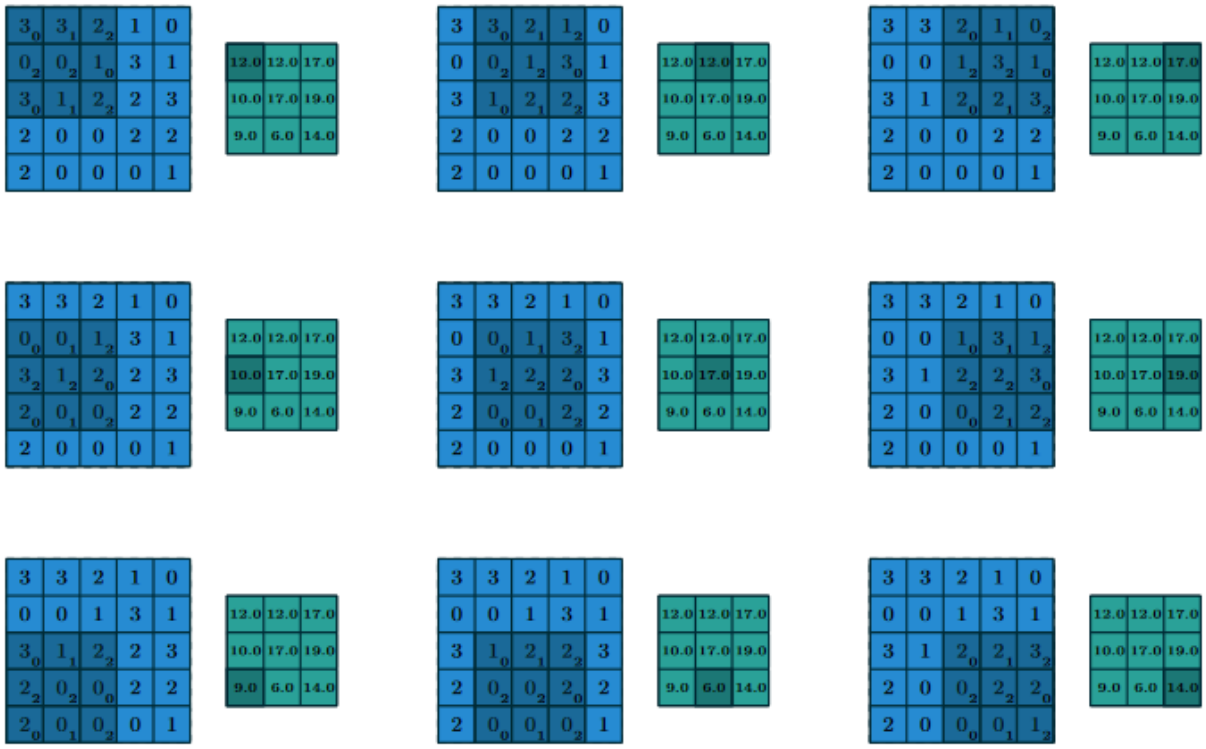


FIGURE 9. Process of an LRF (subscripted numbers) moving over the input image (blue) outputting values to a feature map (green) in a convolutional layer. Weights of the LRF are shown here as subscripts, bias is set 0 and step size is 1. Larger step size would preserve less information but decrease the number of parameters in the network. Convolutional layers usually have multiple feature maps which each has their own LRF. (Dumoulin & Visin 2018.)

This chapter wasn't an exhaustive description of all structural and functional parts of deep neural networks, but it includes the most important parts to understand basics of DL. The ability of a deep neural network to learn efficiently and accurately depends on many adjustable components. For example, how many layers and neurons there are, which activation function is used or how the weights and biases are initialized. Fundamentally, if already existing and

proper neural network architecture is used, the outcome of the network depends heavily on the training data. The learning algorithm tries to perform well related to the training data it is given. Precision in manual labelling and diverse training data expressing as many achievable scenarios as possible increase the probability of successful DL (Mathis et al. 2020).

4 DEEP LEARNING –BASED MOTION ANALYSIS

As was concluded in chapter 2, marker-based motion analysis is expensive and time-consuming, which makes it inaccessible for many practitioners. As such, methods that require less resources could enable the use of motion analysis for a larger variety of individuals and groups, like small businesses, solo practitioners and sports coaches. Different markerless methods have been developed with varying accuracies and purposes (Colyer et al. 2018). DL –based methods are just one branch in the tree of markerless approaches but they are deemed the most powerful method for pose estimation (Mathis et al. 2020).

Human pose estimation (HPE) by DL –based methods has evolved rapidly in the past few years, partly because of open-source datasets and codes for algorithms (Mathis et al. 2020). HPE methods can be categorized to different approaches and Chen et al. (2020) have reviewed the current state of DL –based 2D and 3D HPE methods. They reviewed the literature on each method, but this thesis will mainly focus on fundamentals of HPE and some specific methods that are relevant.

4.1 Fundamentals of HPE

DL –based HPE can be categorized to generative (model-based) and discriminative (learning- or example-based) methods. Discriminative methods are usually faster, but their accuracy is highly dependent on the robustness of the training data. Top-down approaches in HPE work from higher level stages (e.g., person detection) to lower level (e.g., keypoint detection), and vice versa for bottom-up approaches. (Chen et al. 2020.) Multi-person HPE problems can be solved with either top-down approaches by first detecting persons and then performing single person HPE (e.g., Iqbal & Gall 2016), or with bottom-up approaches where all keypoints are detected first and body part configurations are determined in later processing (e.g., Insafutdinov et al. 2016).

HPE can further be divided to regression-based and detection-based methods. In regression-based methods the keypoint coordinates are directly mapped into the image (Toshev & Szegedy

2014; Carreira et al. 2016; Sun et al. 2017). Detection-based methods represent the keypoints by image patches and heatmaps (Tompson et al. 2014; Newell et al. 2016; Sun et al. 2019). Heatmaps are usually pixelated 2D maps and generated by a neural network for each joint, which present the probabilities of a given joint being present in the image (Jain et al. 2014). Detection-based methods provide more robust pixel information, but resolution gets decreased because of pooling in the CNNs. In general, detection-based methods are more suitable and accurate in HPE, consequently, most of the recent HPE research is based on heatmaps. Lastly, methods with one-stage processing (end-to-end networks) aim to produce the human pose from the input image without intermediate supervising. In multi-stage methods the process has separate stages, e.g., predictions of joint coordinates in 2D and further processing for 3D coordinates. Training of one-stage methods is easier, but with less control than in multi-stage approaches. (Chen et al. 2020.)

One of the most essential part in HPE, and DL in general, is the network training. In HPE the ground truth is determined by labels in the training data images, where the labels are regularly placed on the articulating joints. It's obvious that if the ground truth is incorrect, the network learns to label incorrectly. It would be comparable for a piano instructor to teach wrong notes for their pupil and expect them to learn the right melody. It's rather common to use publicly available and readily annotated datasets for training and network evaluation (Chen et al. 2020). For 2D HPE some of the most popular datasets are FLIC (Sapp & Taskar 2013) that contains over 5000 images from Hollywood movies, LSP (Johnson & Everingham 2010) that includes 2000 images of different sport poses, and MPII (Andriluka et al. 2014) which has almost 25000 frames from various YouTube videos. Popular databases for 3D HPE are HumanEva-1 and 2 (Sigal et al. 2010) which include around 40000 frames from common activities by four subjects with markers, and Human3.6M (Ionescu et al. 2014) that includes 3.6 million 3D poses of 11 actors with markers doing daily activities.

Where 2D HPE estimates planar keypoint locations in the camera view, 3D HPE must predict a third dimension, which is depth. This is more challenging due to the extra dimension, but also because the training data is harder to obtain. Many different strategies have been developed for 3D HPE, but mainly they can be divided to model-free and model-based methods. Further, the

approaches can also be split to methods which either predict 3D poses from intermediary 2D poses, or directly predict the 3D pose in the image. (Chen et al. 2020.)

Recently, open-source pose estimation toolboxes like DeepLabCut (Mathis et al. 2018; Nath et al. 2019), Anipose (Karashchuk et al. 2020), pose3D (Sheshadri et al. 2020) and Visual Gait Lab (Fiker et al. 2020) have been developed. The goal of these toolboxes is to speed up the analysis by making it more automated and easier to use without computer science experience. Additionally, a commercial software called Theia3D (Theia Markerless Inc., Kingston, ON) has been developed for 3D HPE analysis with CNNs. Kanko et al. (2020a) did show that Theia3D can measure joint kinematics more reliably than marker-based systems and Kanko et al. (2020b) demonstrated that Theia3D is a valid tool to measure spatiotemporal gait parameters compared to a marker-based system. However, the system is not freely available.

4.2 Overview of 2D and 3D HPE

LeCun et al. (1989) were the first to apply backpropagation to CNNs and realize the effectiveness of this method, in their case, in recognizing handwritten zip code digits. The working principles of CNNs were able to decrease the computing demands by forcing the hidden neurons to combine only local information of minimally pre-processed input images. Two decades later, with CNN-based AlexNet (Krizhevsky et al. 2012), the whole computer vision community took a giant leap forward in object recognition and other subfields of computer vision. The AlexNet won the ImageNet LSVRC-2012 image classification competition (Russakovsky et al. 2015) by a significant margin, with top-5 test error rate of 15.3 % compared to 26.2 % for the runner up. The top-5 error rate means that the correct class wasn't in the five classes that the network ranked most probable. In the competition the goal of the network is to classify over a million images to ~1000 different classes.

What made AlexNet successful, was the unprecedentedly large CNN with over 60 million parameters. The size of the network was actually limited by the amount of memory available on then existing state-of-the-art graphics processing units (GPU). Even though two GPUs were working parallel, the network took between five and six days to train. The authors stated that

their results could be improved by just waiting for faster GPUs. (Krizhevsky et al. 2012.) As a remark about the rapid development in the field, the current state-of-the-art network (Touvron et al. 2020) in ImageNet classification benchmark reaches top-5 test error rate of only 1.3 % with approximately 480 million parameters.

Two years later, Toshev & Szegedy (2014) presented DeepPose (figure 10), a regression-based multi-stage HPE method, which was based on AlexNet due to its success in image classification. DeepPose was the first method which applied an original deep neural network to HPE and showed that a generic CNN, initially designed for image classification, could be applied for joint localization task. The network was evaluated to FLIC and LSP datasets and it achieved similar or better results as the current state-of-the-art networks at that time. The network was trained with specific training data from the evaluation datasets. For example, they used 11000 annotated training examples from the LSP dataset to train the network.

FIGURE 10. Schematic view of the 2D HPE process of DeepPose (Toshev & Szegedy 2014). The network makes an initial estimation of the joint location from a down-sampled input image (left). In stage *s,* the network refines the initial estimation from a cropped higher-resolution input image.

In DeepPose, the initial regression estimate is refined through subsequent stages. The advantage in the initial stage is that the joint location estimation is based on the full image, but the small resolution limits the accuracy of the estimate. Increasing the input image size could result better estimates, but that would escalate the size of already big pool of parameters. Instead, the full estimation is performed as a cascade of stages, where the input for stage *s* is a cropped and higher-resolution image from the previous stage. (Toshev et al. 2014.)

Some other regression-based HPE methods have included an Iterative Error Feedback framework that recursively self-corrects the initial solutions (Carreira et al. 2016) and a compositional pose regression, which is a structure-aware method that uses bones instead of joints and constrains the solution with joint connection structures (Sun et al. 2017). Estimating joint coordinates by regression with only few constrains is a very hard task and it simplifies the problem too much (Sun et al. 2017). Thus, most of the recent HPE research is based on heatmaps representation since it's more robust than the coordinate representation (Chen et al. 2020).

Jain et al. (2014) proposed a novel detection-based method for 2D HPE where they used heatmaps for joint localization and, additionally, temporal information for refining the predictions. Simply put, optical-flow maps were computed from image pairs, where active regions indicated relative motion between the images. Thus, temporality could be taken into account in the analysis. The authors also used a simple spatial model to make the joint predictions stronger by masking out the incorrect heatmaps (figure 11). For training the network, they used almost 4000 training images. Papandreou et al. (2017) proposed an enhanced heatmap prediction method by adding an offset vector field for each keypoint. When the heatmaps and corresponding offset field were combined in a weighted manner, they produced very highly localized activation maps. Cao et al. (2017) presented an efficient method to detect 2D poses of multiple persons in an image, which was later released as an open-source system OpenPose (Cao et al. 2019). First, the system provides heatmaps of body part locations and part affinity fields describing the orientation of limbs. Second, the heatmaps and part affinity fields are analysed to compute connected limbs for each person in the image. The presented method is very efficient and able to detect 2D poses of multiple people in real-time.

The field of HPE is constantly evolving with new and better methods regularly emerging from different research groups. Therefore, the previously presented methods might be outdated in the sense of predicting performance but have provided a short overview on the fundamental principles of DL –based HPE methods.

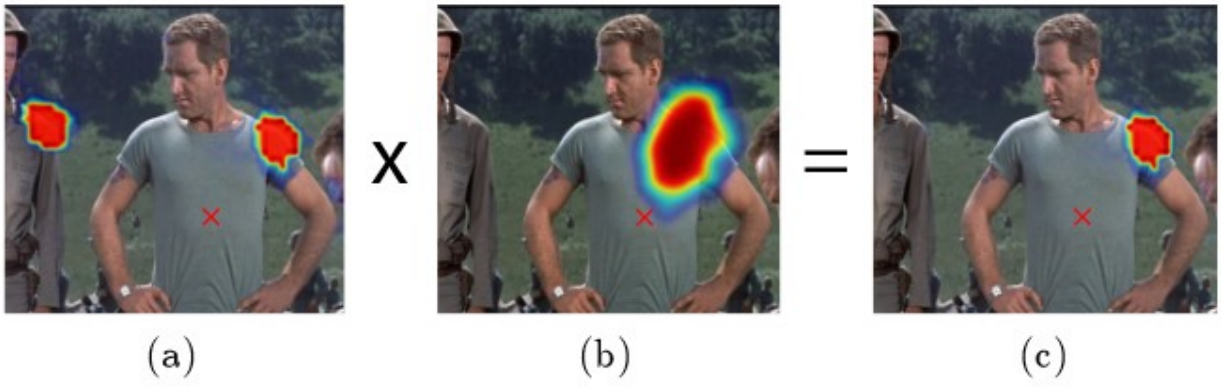


FIGURE 11. Spatial model implemented in Jain et al. (2014) for masking out the incorrect joint heatmaps. A joint mask (b) is used to describe the possible joint locations when the torso is centered in the mask. In (a) there are two separate heatmaps for left shoulder from which the incorrect one is masked out in (c).

Lately, user-friendly toolboxes implementing DL –based pose estimation features have been developed. DeepLabCut (Mathis et al. 2018; Nath et al. 2019) was originally developed as a Python toolbox for animal pose estimation, which utilizes feature detectors of DeeperCut (Insafutdinov et al. 2016) for keypoint localization. Mathis et al. (2018) demonstrated that they were able to reach human-level labelling accuracy with DeepLabCut using only ~200 frames of annotated training data. The low number of training frames was possible because of phenomenon called transfer learning (Yosinski et al. 2014). In transfer learning the network has first been pretrained to detect general features and later tailored to a specific task. It's common that modern HPE algorithms are pretrained with higher level datasets like ImageNet (Deng et al. 2009), which includes 14.2 million images from 21000 classes (Mathis et al. 2020). Cronin et al. (2019) have verified the power of transfer learning by showing that 300 – 400 labelled images were enough to reach human-level labelling accuracy in underwater running.

Karashchuk et al. (2020) have introduced Anipose, a Python toolkit for 3D pose estimation in animals and humans. The Anipose pipeline (figure 12) streamlines the data processing in pose estimation and the authors also provide tutorials and help for new users. Anipose triangulates and refines 2D detections to estimate 3D poses and it can be used together with any neural network –based 2D method. Camera system used must be calibrated and the authors recommended a precision manufactured checkerboard for it.

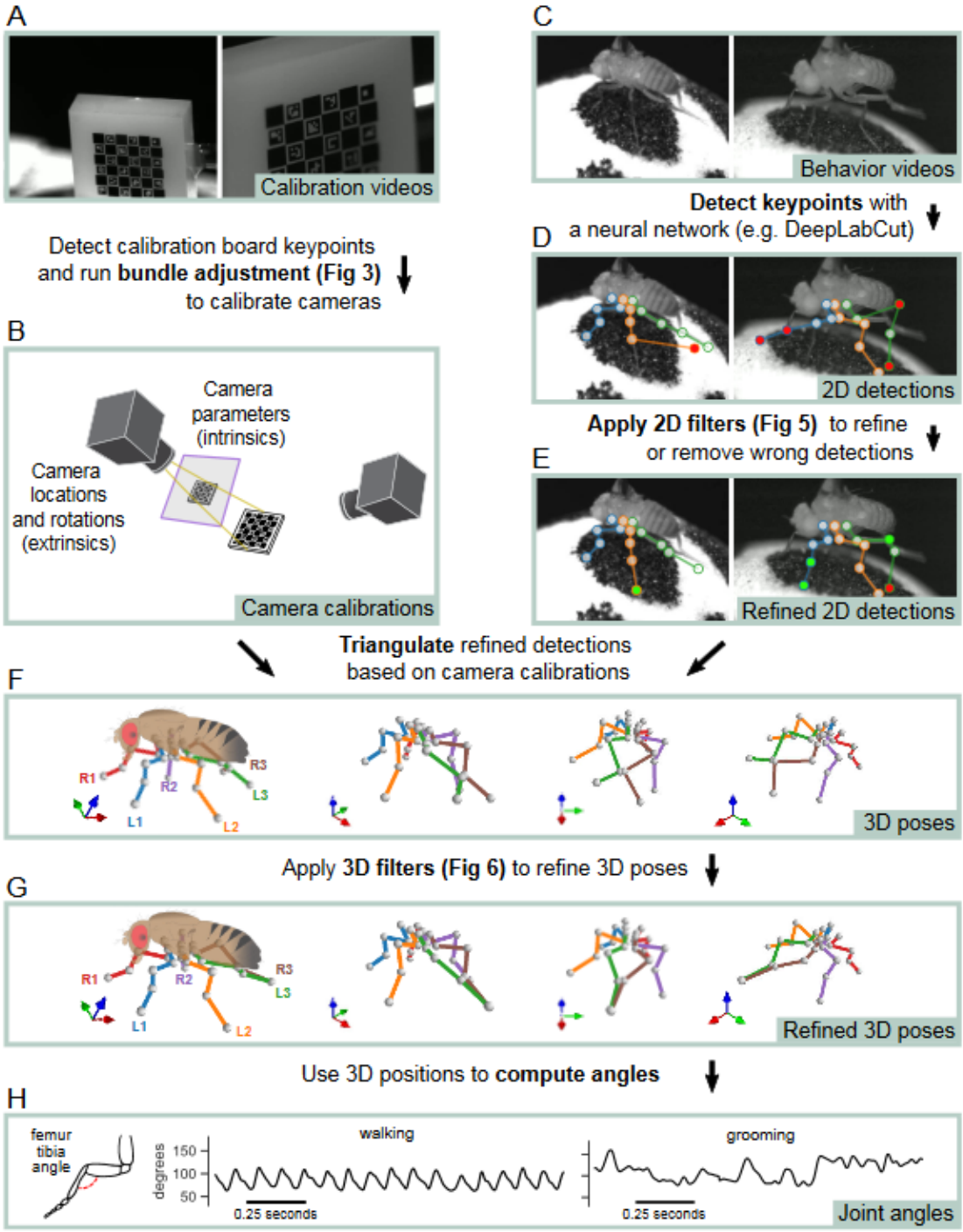


FIGURE 12. An overview of the Anipose pipeline. The multi-camera system is calibrated with the calibration board (A and B) and the intended behaviour is captured with the same camera system (C). The 2D keypoints are estimated with any neural network –based method (D) and those keypoints are refined with filtering (E). The 2D detections are triangulated to estimate the 3D poses (F), which are then passed through additional filters (G). Lastly, joint kinematics are computed (H). (Karashchuk et al. 2020.)

4.3 Strengths and limitations of HPE methods

The most obvious strength of DL –based HPE methods are their noninvasiveness. They don't require markers and without markers there isn't STA either. Theoretically the data can be captured anywhere, e.g., clinical office or sports environment, and already existing data can also be analysed or used for training, if applicable. The equipment for data capture is less expensive than for marker-based systems, and under certain conditions even cameras of modern smartphones can be sufficient. In addition, analyzing new videos gets very effortless after the network has been trained once and its verified to perform well.

Training a network gets computationally very demanding when the depth of the network and number of parameters grows higher and higher in search of better performance. Luckily, the available computing power is also increasing with more powerful GPUs every few years. The required computing power can currently be harnessed also remotely via internet (Mathis et al. 2020). Cloud computing services, e.g., Google Cloud GPUs (Google Cloud 2020), offer the computing power of fast GPUs in their facilities to be used for machine learning and other demanding computing tasks.

Transfer learning has reduced the number of training examples required to few hundred (Mathis et al. 2018; Cronin et al. 2019) but the training data still must characterize the behavior as well as possible. Pretraining saves training time, increases robustness and decreases the amount of training data needed (Mathis et al. 2020). Anatomical relevance of detected keypoints is one of the problems that is faced when DL –based HPE methods are used for biomechanical analysis. If anatomically valid joint kinematics is the goal of the analysis, labelling the keypoints must be meticulous. Further, the annotations in the public datasets might not always be anatomically relevant. DL –based methods are becoming more flexible with toolboxes like DeepLabCut and the users can determine more accurately what keypoints they want to track.

To decrease some of the errors and problems that might emerge with DL –based methods, few practical guidelines should be considered. Video quality should be good enough to identify the keypoints, but not too high to keep training time reasonable and to spare disk space. Labeling

should be as accurate as possible, and conventions should be agreed on if multiple people are doing the labelling. Training data should represent the full behavior in question and usually it's better to select more videos and use less frames per video than select the same number of frames but from fewer videos. Data augmentation can be useful to create artificial training data and reduce overfitting. (Mathis et al. 2020.)

Seethapathi et al. (2019) have reviewed the field of HPE studies and criticized that current pose estimation methods aren't prioritizing the right things for movement sciences. They argue that most of the current algorithms treat consecutive frames as statistically independent, although the motion of human body obeys spatial and temporal laws of motion. If position, velocity and acceleration are known in few consecutive frames, that is a strong prior for the next frames. They also point out that the field needs more anatomically meaningful ground truths and collaboration with movement scientists.

4.4 Validity of HPE methods for measuring gait kinematics

The previously described public datasets (FLIC, LSP, MPII, HumanEva and Human3.6M) are used for fair comparison between algorithms and can quantitatively separate which method is the best. Different metrics are used for comparison depending on the dataset. For example, the 2D MPII dataset uses Percentage of Correct Keypoints (PCK) which indicates the relative number of joints that are detected correctly. Correctness is determined by a threshold and in MPII dataset they use 50 % of the head segment length (PCKh-0.5) as the threshold. The current state-of-the-art (Bulat et al. 2020) for MPII achieves PCKh-0.5 of 94.1 %. In one the 3D datasets, Human3.6M, Mean Per Joint Position Error (MPJPE) is used as a metric. It reports the Euclidean distance between the predicted 3D coordinates and the ground truth averaged over all joints. The MPJPE for the current state-of-the-art (Iskakov et al. 2019) in Human3.6M is 17.7 mm. (Chen et al. 2020.)

Moro et al. (2020) have shown that there aren't statistically significant differences between lower limb 2D kinematics during walking in stroke survivors when analysed with marker-based system and DeepLabCut (Mathis et al. 2018). The mean Euclidean distances between systems

for lower limb keypoints were approximately 11 – 18 mm. Van Den Bogaart et al. (2020) used DeepLabCut and Anipose to track 48 marker positions in 3D of subjects standing on a balance board. The network was trained with 440 labelled frames and it achieved an error under 10 mm compared to human labeler.

Stenum et al. (2020) compared spatiotemporal gait parameters and sagittal lower-body joint kinematics during overground walking between marker-based system (Vicon) and OpenPose (Cao et al. 2019). They found out that hip and knee angles in sagittal plane correlated well between Vicon and OpenPose, but ankle kinematics weren't as similar (figure 13). Limitations in the paper were the facts that there was difference between actual marker positions (anatomically relevant) and keypoints detected with OpenPose (general) (figure 15), and that OpenPose was used "off-the-shelf" without further training. Training the network with additional data and having keypoints that are anatomically relevant could improve the results. The authors also noticed that that some of the spatiotemporal gait parameters got poorer when the subject was close to the image edges.

Another study comparing OpenPose (Cao et al. 2019) and marker-based system (Motion Analysis Corporation) was done by Nakano et al. (2020). They recorded two subjects with multiple cameras (3D) while walking, jumping and throwing a ball. The mean absolute error (MAE) was calculated between the time series produced by OpenPose and marker-based system (figure 14). Of all MAEs, 47 % were <20 mm and 80 % were <30 mm, but 10 % were >40 mm. As can be seen from figure 14, OpenPose tracked the keypoints rather well in some joints and planes but performed poorly in other cases. Alike in study by Stenum et al. (2020), the keypoints of OpenPose don't fully match with markers on the skin, which might be one reason for the larger MAEs.

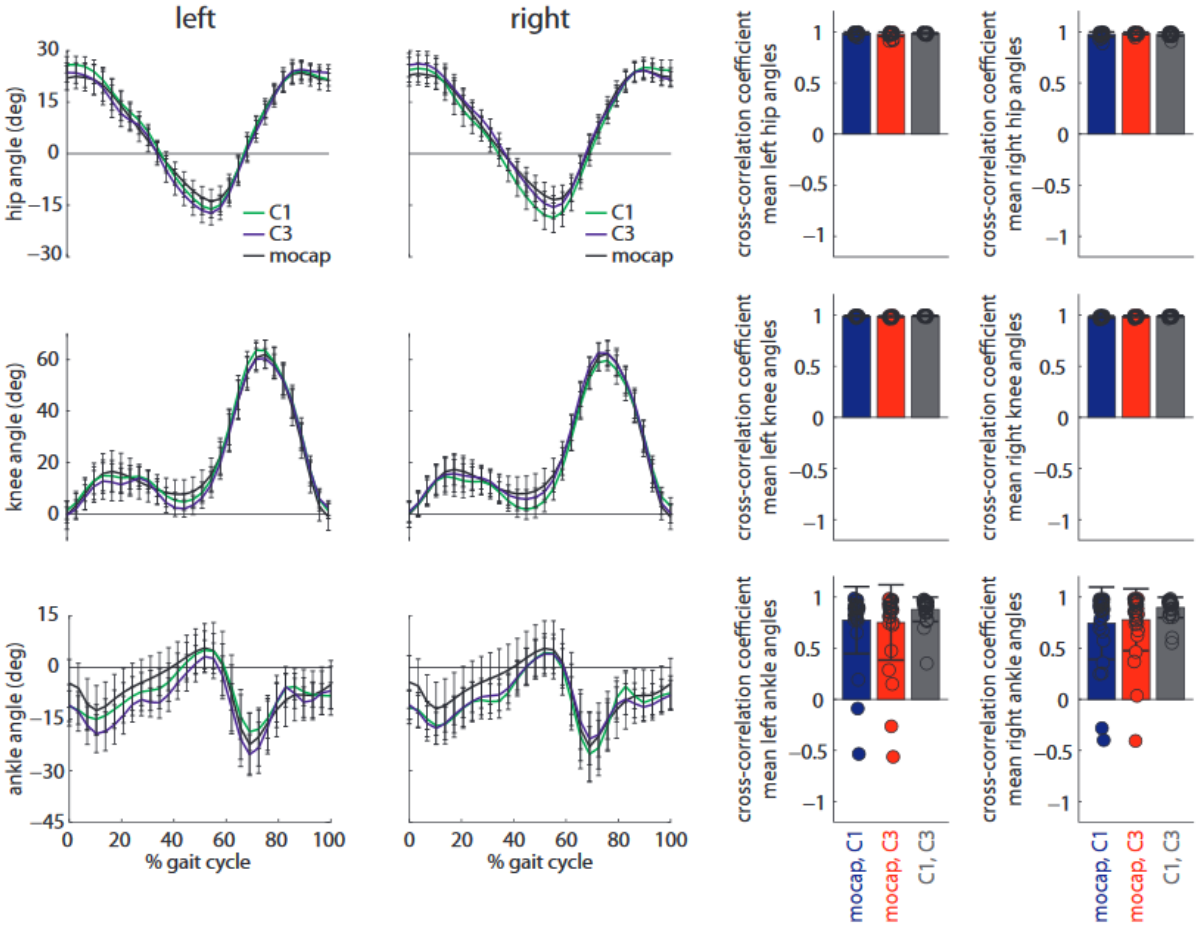


FIGURE 13. On the left are sagittal lower-body joint angles for left and right limbs between Vicon (mocap) and OpenPose from digital video cameras on both sides (C1 and C3). On the right are the cross-correlations between Vicon and OpenPose. (Stenum et al. 2020.)

Zago et al. (2020) also evaluated the performance of OpenPose (Cao et al. 2019) relative to marker-based system. They specifically studied the effect of three factors on the gait variables: relative distance of two cameras, gait direction and video resolution. Lowest RMS error (20.8 mm) between keypoint trajectories from OpenPose and marker-based system over all keypoints was achieved with cameras 1.8 m apart, straight gait (towards or away from cameras) and high resolution (1312x736 pixels).

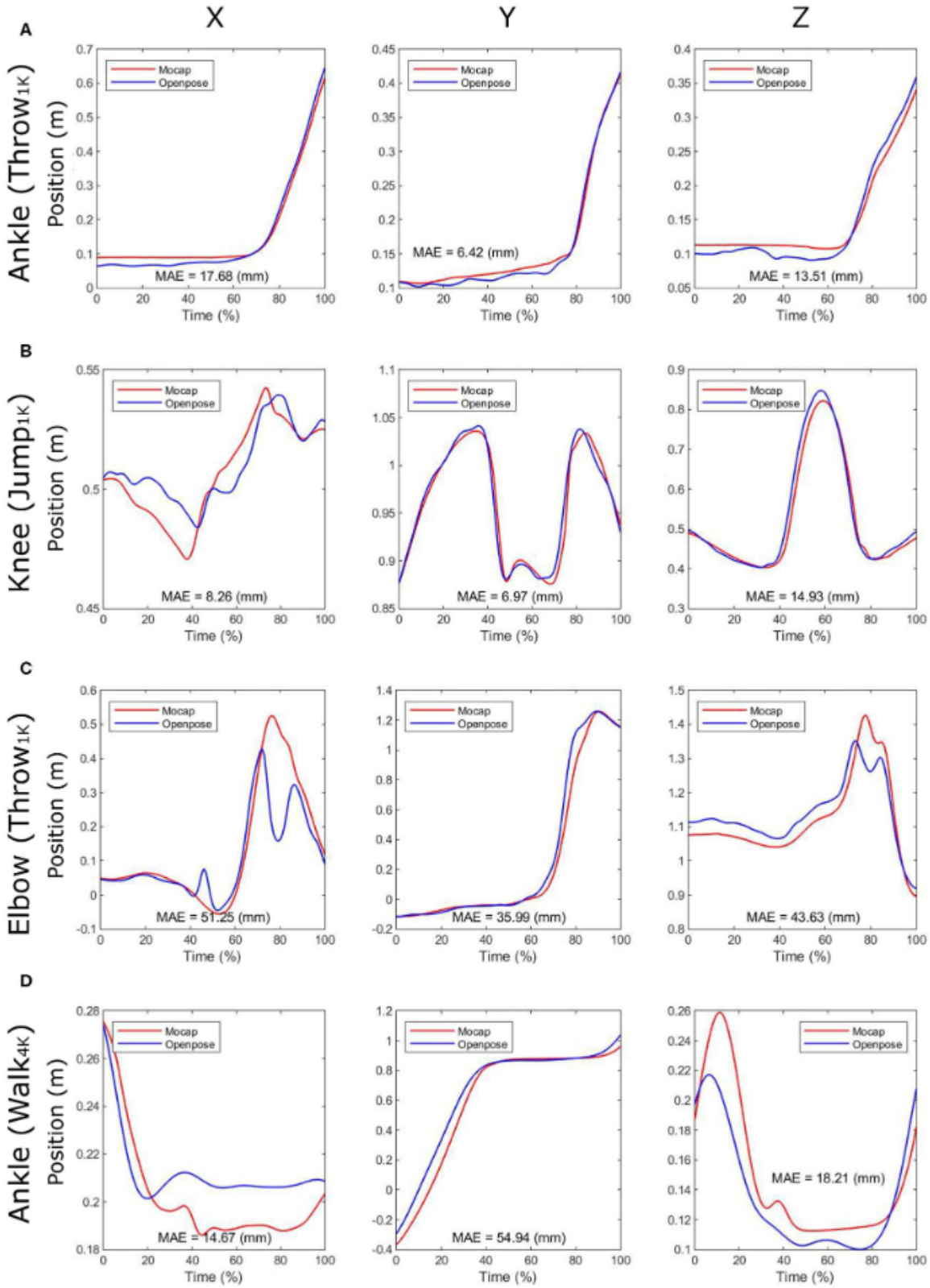


FIGURE 14. Positions of different joints from four different actions (A-D) separated to three components. Mean absolute error (MAE) is also presented on each graph. X is medial/lateral (ML), Y is anterior/posterior (AP) and Z is vertical component. (Nakano et al. 2020.)

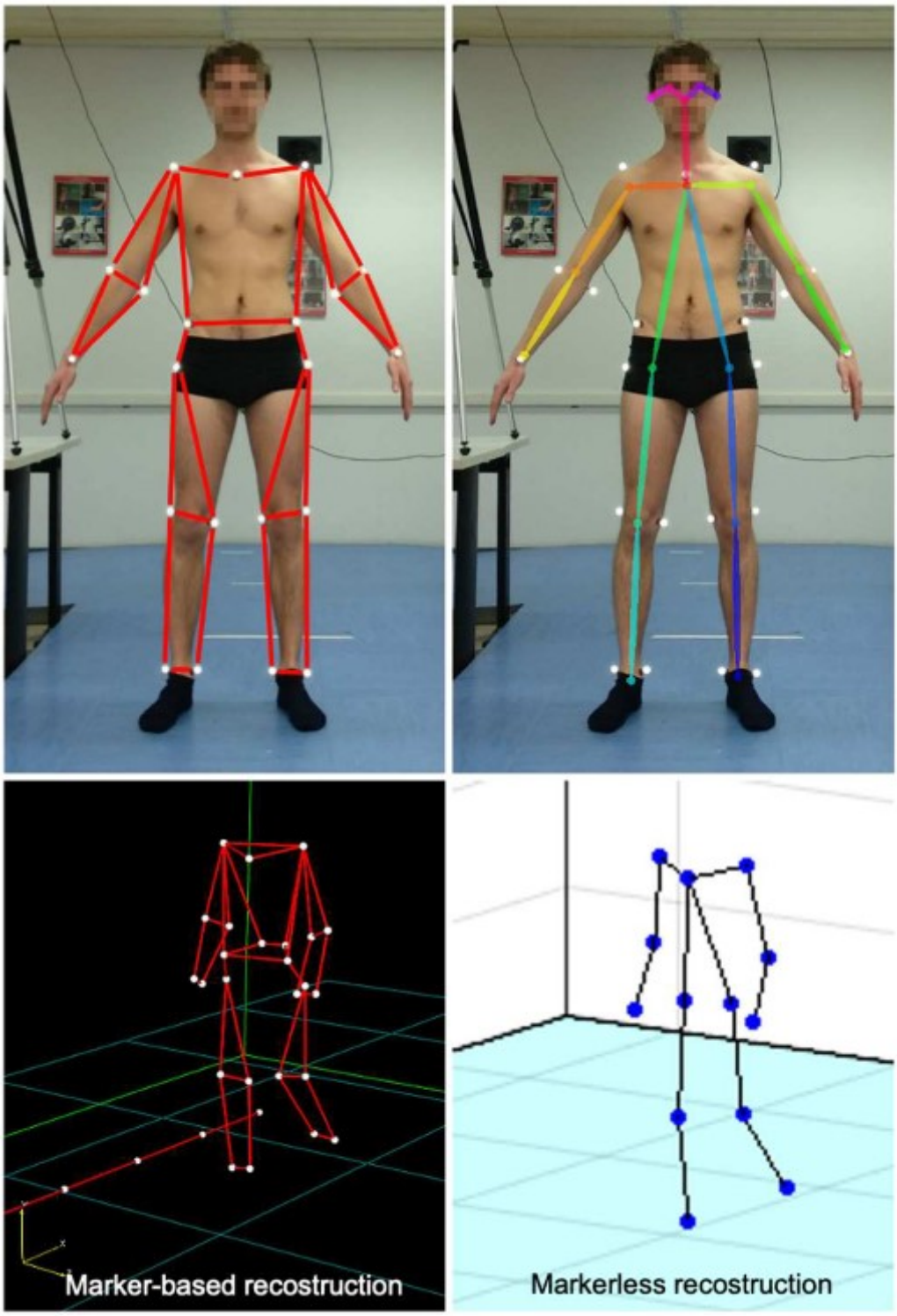


FIGURE 15. Differences between keypoint locations and reconstructed skeleton models with marker-based system (left) and OpenPose (right) in Zago et al. (2020).

The most recent study investigating the performance of markerless pose estimation algorithms was conducted by Needham et al. (2021). They analysed three different DL -based methods (OpenPose, AlphaPose and DeepLabCut) and compared their performance in estimating joint centre locations in 3D to marker-based motion capture during walking, running and jumping. Again, no additional training for the models was performed and keypoints didn't match the marker locations of criterion method. They found systematic differences of 30 – 50 mm in 3D joint centre locations for knee and hip joints, and 1 – 15 mm for ankle joint depending on the activity.

5 PURPOSE OF THE STUDY

Training data for a neural network must represent the full intended behavior in question. In this project, the neural networks were required to perform well with different body types, gait patterns, statures and clothing. Therefore, training data should include these variations so that the model becomes robust. To do this, males and females were recruited to walk and run at various velocities. The participants' naturally have some variation in their size, clothing and gait patterns. Since one goal of this research project was to evaluate the concurrent validity of neural network -based approach and Vicon for measuring joint kinematics, reflective markers had to be attached to the subjects. The markers can be visually very distinct on the videos and they become one variable that might affect the predictive performance of the model. Thus, the effect of marker presence in the training data to neural network performance was evaluated.

In previous studies (Nakano et al. 2020; Stenum et al. 2020; Zago et al. 2020), the keypoints analysed by a neural network have been different from marker locations. Same keypoints should be located between the systems so the comparison becomes more unbiased. This was achieved by using marker locations as keypoints in this study. Also, the analysis partly required comparing kinematics from different trials: kinematics analysed from trials without markers by neural networks were compared to kinematics analysed from trials with markers by Vicon. Thus, the between-trial reliability of joint kinematics when analysed by Vicon was assessed to estimate how much of the observed variation between different trials was due to stride-to-stride variability.

Thus, the main purposes of this study were to evaluate how marker presence in the training data affects the performance of neural networks and what is the role of stride-to-stride variability when kinematic data is compared between different trials. The research questions of this study were:

1) How does the presence or absence of reflective markers in the training dataset affect neural network predictive performance?

Hypothesis was that neural network performance would be better when it was used to analyse similar trials as what it was trained with. It was estimated that having markers in the training data could make the network rely on identifying marker features and perform poorly with data without markers. In theory, the network trained with data without markers should've learned other features and still perform well with data that has markers. Thus, it was hypothesized that the network trained with data without markers would be more versatile than the other network and could perform decently also with data that has markers.

2) What is the between-trial reliability of joint kinematics in walking and running when analysed with Vicon and how it changes with number of included strides?

Hypothesis was that between-trial reliability would increase as more strides were used per trial and reliability would reach reasonable levels for clinical measurements.

6 METHODS

To answer the research questions, participants were recruited to walk and run on a treadmill while their gait was recorded with two camera systems. The study was conducted in the facilities of University of Jyväskylä and with equipment owned by the university. The study was approved by the ethic committee of the University of Jyväskylä and was conducted in accordance with the Declaration of Helsinki.

6.1 Participants

Participants were recruited from local students and residents through word of mouth, email and social media. In total, 18 participants volunteered to take part in the study of which 11 were males and 7 females (table 1). Inclusion criteria for participants were: 1) 18-35 year-old male or female, 2) able to run at least 2500 meters in a 12 minute running test (Cooper's test), 3) no injuries affecting walking or running during the previous year and 4) no chronic cardiovascular or respiratory diseases. Cooper's test wasn't tested during measurements and it was used as a criterion to exclude novice runners from the study. It was estimated that being able to run at least 2500 meters in the test would indicate adequate cardiovascular fitness to prevent fatigue during our protocol. Other criteria were included to exclude any effects that age or injuries might have on kinematics and to minimize the risks in participating to the study. Participating in the study was voluntary and didn't include compensation. The participant was informed about their rights, testing protocol and risks involved in conducting the measurements. This information was provided during recruitment via email or personally, during the measurement session and with an informed consent at the beginning of the measurement session.

TABLE 1. Subject demographics by gender. Data are presented as mean ± SD.

Gender	Age (years)	Height (cm)	Weight (kg)
Male (n = 11)	26.1 ± 4.4	178.8 ± 6.4	80.2 ± 11.8
Female (n = 7)	24.9 ± 3.3	163.0 ± 2.8	59.0 ± 5.5

6.2 Study design

Participants' joint kinematics were analysed by neural networks from videos obtained by GoPro cameras and by Vicon motion analysis system. Two neural networks were trained with slightly different training data: one with frames from trials where participants were wearing markers and another one with frames without markers. Kinematics obtained through neural networks were compared to kinematics obtained from Vicon system, which was treated as the golden standard system. See figure 16 for visualization of the logic and appendix 1 for the experimental protocol.

When neural networks were used to analyse trials without markers, it wasn't possible to compare data directly to data from Vicon. Thus, trials were performed twice with markers to evaluate the between-trial repeatability of kinematic variables when analysed with Vicon. If the kinematics obtained by a neural network from trials without markers had similar variability, then it could be said that the neural network performs similarly to Vicon even when analysing trials with no visible markers.

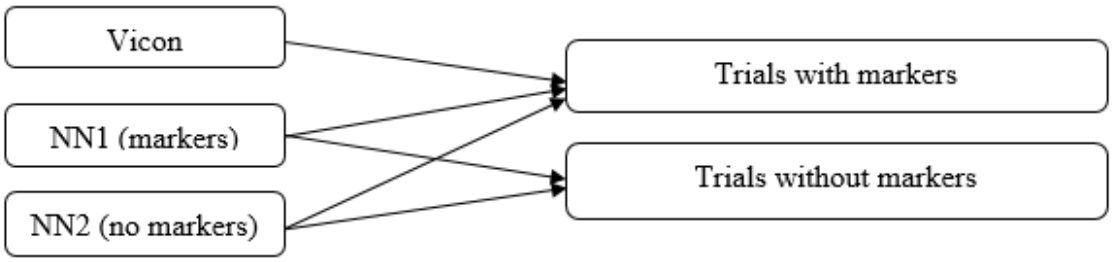


FIGURE 16. A logical structure of which systems were used to analyse which trials. All trials were analysed by both neural networks and only trials with markers were analysed by Vicon. There were three sets of analysed kinematics for each trial with markers and two sets for each trial without markers.

6.3 Experimental protocol

In the study, each participant attended one measurement session taking approximately 2 hours. Small marks were made on the participants skin with a marker, where the reflective markers would be attached for motion capture (table 2). The list doesn't include medial malleoli that were used only to measure leg lengths. Also, marks weren't made for heels and toes since those markers were attached on the participants shoes. Anthropometric measures were taken next, which were required by the biomechanical model in Vicon. These included participants standing height, weight, ankle and knee width, and leg length. The participant was standing during all the measurements. Height was measured with a meter attached to a wall while the participant had their heels, back and back of the head against the wall. Weight was measured by a scale with the participant wearing light sports clothing but no shoes. Ankle and knee width were measured with a caliper as the distance between medial and lateral malleoli, and, as the width across the line of the knee axis, respectively. Leg length was measured as the distance between anterior superior iliac spine and medial malleolus. All the anthropometric measures were taken by following guidelines provided by Vicon Nexus documentation (Vicon documentation 2020).

Markers were then attached to the marked locations on the skin and shoes. For shoes, the toe marker was attached before the heel marker. If markers were blocked by participants shorts or shirt, they were taped where necessary. If there were unwanted reflections visible in Nexus software due to reflective areas in participants clothing or shoes, those were covered with tape. To keep the participant in the correct area of the treadmill belt, there was a large cross taped on the wall directly in front of the participant. Additionally, unwanted AP drifting on the belt was tried to mitigate by instructing the participant to keep one pair of the cameras directly on their right side. The participant was wearing a harness during measurements, which also "pulled" the participant back on correct area if they drifted too far. Verbal cueing was also used, if previous methods failed.

TABLE 2. Marker names and their locations in the Plug-in Gait lower body model (Vicon documentation 2020).

Marker	Segment	Anatomical location
LASI & RASI	Pelvis	On anterior superior iliac spine.
LPSI & RPSI	Pelvis	On posterior superior iliac spine immediately below the sacro-iliac joints.
LTHI & RTHI	Femur	On the lateral surface of the thigh. On the proximal third on the right leg and distal third on the left.
LKNE & RKNE	Femur	On lateral side of the flexion-extension axis of the knee.
LTIB & RTIB	Tibia	On the lateral surface of the tibia. On the proximal third on the right leg and distal third on the left.
LANK & RANK	Tibia	On the lateral malleolus on the imaginary transmalleolar axis.
LTOE & RTOE	Foot	On the second metatarsal head on the mid-foot side of the equinus break between fore-foot and mid-foot.
LHEE & RHEE	Foot	On the calcaneus at the same height as the toe marker.

The main protocol performed can be seen in appendix 1. It consisted of six 10-12 -minute blocks of walking or running with a 2-minute rest after walking and a 4-minute rest after running. The participant walked and ran with three different velocities during each block: 4, 5 and 6 km/h for walking, and 8, 10 and 12 km/h for running. The order of velocities was chosen arbitrary by the researchers for each participant so that the order was different in each block. Only the last minute of each velocity was captured. The familiarization period before capturing varied between 2 and 5 minutes so that it was longer for the first velocity and shorter for the last two. The very first periods for each gait type were the longest. It's recommended to include familiarization periods of 8 min for running (Van Hooren et al. 2020) and 6 min for walking (Meyer et al. 2019) to reach stable kinematics, but such periods would extend the protocol length significantly. It was estimated that the selected times were long enough for the kinematics to stabilize relatively well, but short enough to prevent accumulation of fatigue late in the protocol. Data collection with Vicon and GoPro systems were started simultaneously.

If there was a camera or other system malfunction, loose or obstructed marker or other reason to pause the protocol, a running clock started at the beginning was paused and the problem was fixed. If the protocol had to be paused during a capture, there was a recapture with additional minute of familiarization before it. The camera systems were recalibrated, if fixing the problem required moving the cameras. Loose or detached markers were reattached to the marked positions. If the detached marker was toe or heel marker, the correct position was palpated or measured again. Additional tape was sometimes added to attach the markers more firmly.

6.4 Gait analysis

Eight pairs of cameras recorded each participant's gait while they were walking and running on the treadmill. The approximate locations and orientations of these camera pairs relative to the treadmill belt are illustrated in figure 17. Each pair consisted of a GoPro video camera and a Vicon Vero infrared camera, which were placed as close to each other as practically possible (figure 18). Distances between two cameras varied from 15 to 30 cm.

The room where the measurements were made wasn't originally designed to be used for motion capture and was relatively small (figure 19). The treadmill (Telineyhtymä Kotka OJK-1, 1989) used in the measurements was operated by one of the researchers from a control panel. Inclination of the treadmill was always set to 0 degrees and velocities were manually set with a turning knob. The length and width of the treadmill belt were 4 m and 2 m, respectively. The treadmill also had a bulky frame with horizontal and vertical railings to which the safety harness was attached. The small room size, the treadmill frame and few attachment sites for cameras caused major issues with trying to find unobstructed views for each pair of cameras.

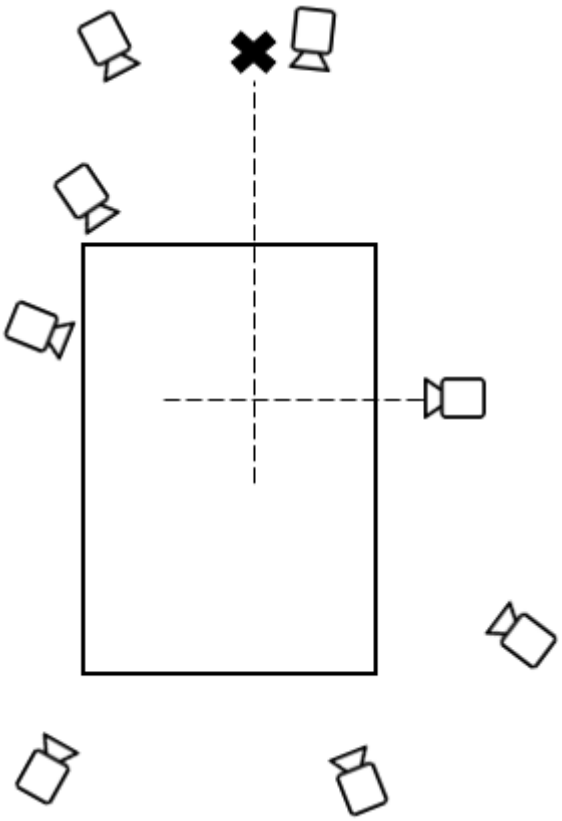


FIGURE 17. Relative location and orientation of cameras relative to the treadmill belt during measurements. Each camera icon represents a pair of cameras (Vicon Vero and GoPro). The intersection of dashed lines indicates the point on the treadmill belt where the participant was instructed to stay.

FIGURE 18. Examples of camera attachments and relative locations of cameras in a pair. Distances between cameras in a pair varied from 15 to 30 cm.

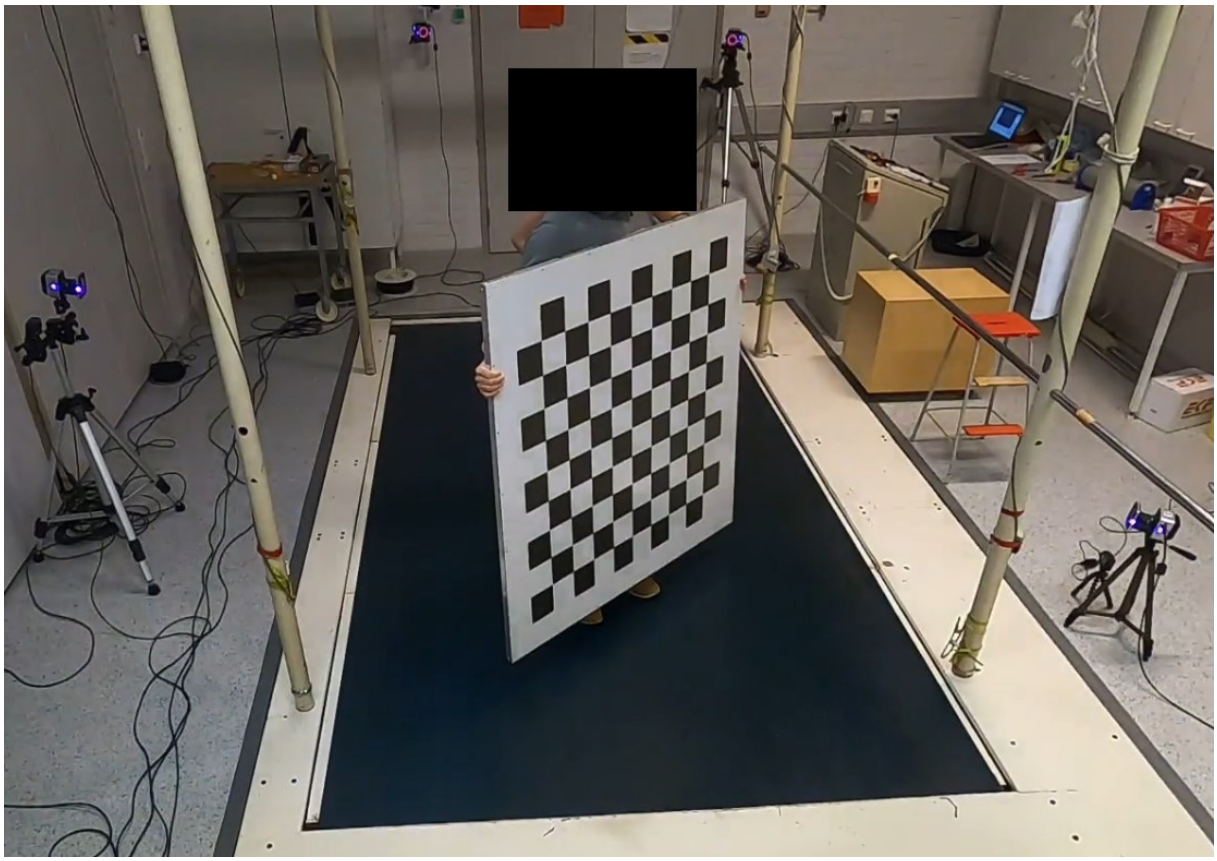


FIGURE 19. A cropped view from one of the cameras on the wall railing showing the measurement room, the treadmill and part of its' frame, camera pairs and the calibration board.

6.4.1 Vicon motion capture system

Eight Vicon Vero v2.2 infrared cameras were used to record locations of reflective markers attached to the participant at 300 Hz. Around the lens of a Vero camera there is a ring of strobing infrared LEDs illuminating the markers for the infrared sensors in the camera. The cameras were connected to a 26-port PoE switch (D-Link, DGS-1026MP) by Ethernet cables and the switch was connected to a laptop also by an Ethernet cable. Data collection was performed with a data capture software Nexus 2.10.2, which was run in the laptop.

Before data collection, the cameras were orientated towards the capture volume and camera parameters were adjusted such that markers inside the capture volume were in focus and illuminated properly. Parameters were readjusted before each data collection session, if cameras were moved or calibration accuracy was poor. Any unwanted marker detections in Nexus from

other cameras or reflections were masked out. Cameras were calibrated by waving a T-shaped 5-marker wand inside the capture volume so that each camera captured at least 1500 frames containing all the markers on the wand. The volume origin was set with the wand approximately to the point where the participant would be on the belt. The wand was laid flat on the ground so that the leftmost edge of the horizontal part of the T-shape was placed at a 1050 mm distance from the left edge of the treadmill. It was aligned in the AP direction with a small slit on the left side of the treadmill. Before starting the walking and running trials, the participants subject skeleton template in Nexus was calibrated by capturing 1-2 seconds of participant standing still.

Passive spherical markers were used with a diameter of 14 mm and they were attached to the participant's skin with two-sided tape. The 16-marker Plug-in Gait lower body model was used as the biomechanical model for calculation of kinematics (table 2). The model requires some subject specific parameters, which are listed in section 6.3.

6.4.2 GoPro camera system

Eight GoPro HERO 8 Black -cameras were used to record videos of the participants' gait. All the cameras were recording at 60 Hz and with 1920x1080 pixel resolution. Zoom setting was set to 1.0x and HyperSmooth was set off. Charging cables were connected to the cameras constantly during data collection and videos were save to an SD-card (SanDisk Extreme, UHS-1 U3 / V30, A2 64 GB). The cameras were connected to a GoPro Smart Remote, which was used to start and stop all the cameras simultaneously. Two of the cameras were mounted on small tripods and the rest were attached to railings with GoPro Pole Mounts. A checkerboard was used to calibrate the system by rotating it inside the capture volume and capturing its' locations with all the cameras simultaneously (figure 19).

6.5 Data analysis

This section describes how data was processed after data collection, what systems were used and what statistical analyses were performed to the processed data to answer the research questions.

6.5.1 Neural network analysis

DeepLabCut toolbox (Nath et al. 2019) was used to label training data, train and evaluate the model, and predict keypoint locations from trial videos. Three subjects were chosen randomly and data from them were only used in keypoint predictions. Data from the remaining 15 subjects were used to train and test the neural network model. Two different models were trained: one with data where subjects wore markers (DLC-M) and the other where markers were taken off (DLC-NoM).

Labelling

Each trial from each subject was recorded by eight GoPro-cameras and from each usable and full-length video 5 frames were sampled for labelling. During data collection, some video files were corrupted and couldn't be used. Also, due to unsolved issues in the GoPro system, some cameras tended to turn off during data collection and less frames were sampled from these videos. In total, 9974 frames were labelled. All four pelvis markers and the markers from the right leg were used as keypoints and labelled in the frames, if visible (figure 20). In frames where markers were taken off, the label was placed on a location where the marker would be.

Training and evaluation

Both models used ResNet-101 neural network (He et al. 2015), which had been pretrained on ImageNet (Deng et al. 2009). For training and evaluating the DLC-M and DLC-NoM models, a total of 6649 and 3325 frames were used, comprising of about 831 and 416 frames per camera, respectively. Frames were split into train (95 %) and test (5 %) datasets, with train dataset only used for training the models and test dataset for evaluating them. Frames were also rescaled by a factor of 0.8 before being passed into the network. Both models were trained with batch size of 1, but DLC-M with 400000 iterations and DLC-NoM with 450000 iterations. Learning rate was 0.005 for the first 10000 iterations, 0.02 until 430000 iterations and 0.002 for rest of the training. DeepLabCut also augmented "new" data for the network by cropping, rescaling or otherwise modifying the original frames. Training times were approximately 46 hours for DLC-

M and 52 hours DLC-NoM. Both models were trained on a NVIDIA GeForce RTX 3060 12 GB GPU.

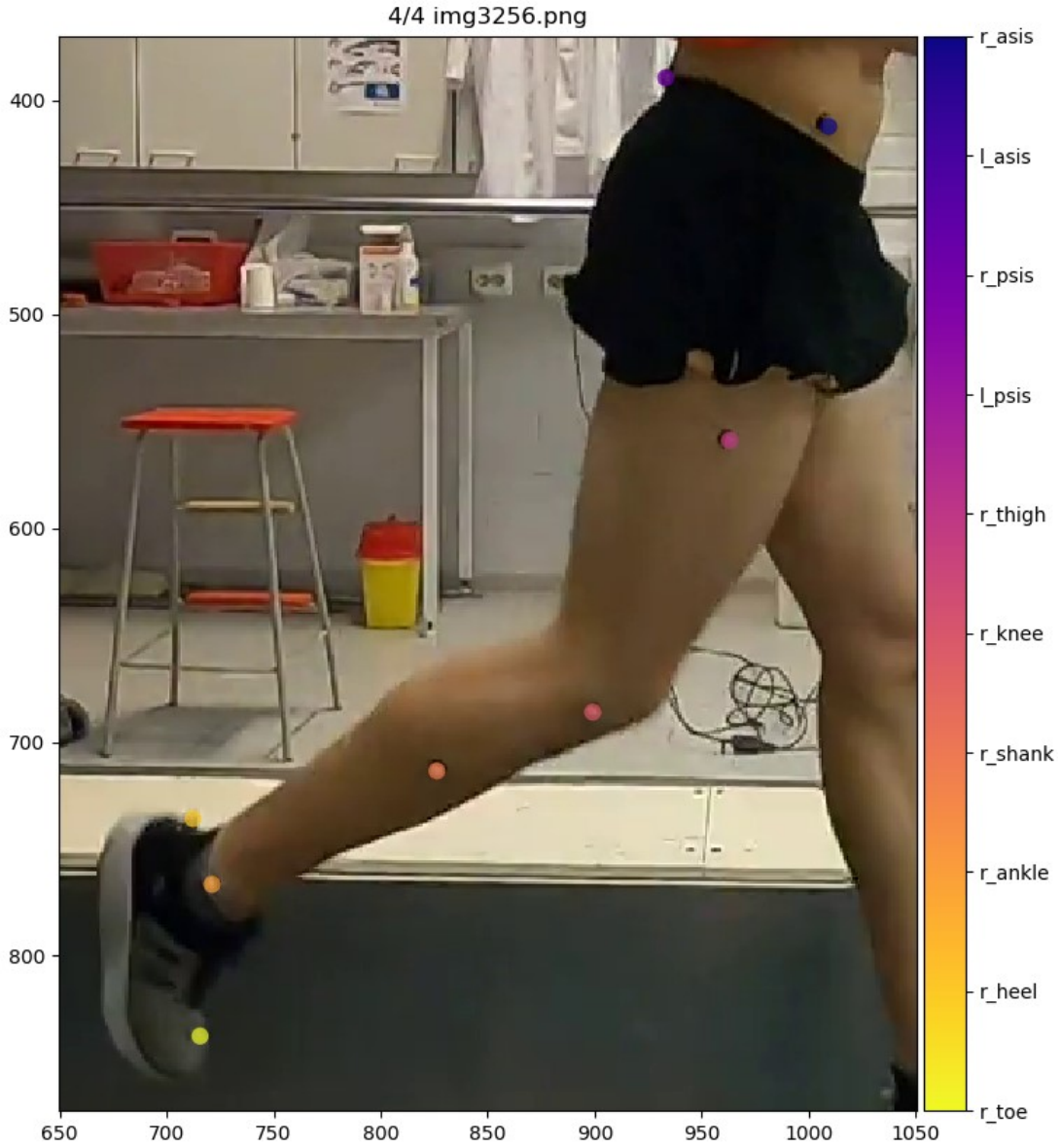


FIGURE 20. An example of a frame (zoomed in) and labels, which were placed over each visible keypoint.

Predicting keypoints

The models were run on a different GPU (NVIDIA GeForce GTX 960 2 GB) with less memory when predicting keypoints, which forced to crop the resolution of original videos to 1100 x 800 pixels, but so that all keypoints were still visible on all frames. For each analysed video, the model produced a list of most probable coordinates for each keypoint in each frame of the video. At this stage, a median filter with window length of 5 was applied to each keypoint trajectory to remove large jumps caused by tracking errors.

6.5.2 Vicon analysis

All analyses described in this section were performed in the Nexus software. First, all trials were labelled and gap filled to obtain continuous 3D coordinates during the full trial. Due to poor capture quality in part of the data, trials that couldn't be labelled correctly or had too large gaps to fill with reasonable accuracy were dropped out from analysis. Running trials from three subjects and walking trials from one subject had to be excluded. After this, data from 17 and 15 subjects were left for walking and running reliability analysis, respectively. To analyse results from DeepLabCut and Vicon as similarly as possible, 3D coordinates of the previously selected three random subjects were exported from Nexus and further processing was conducted the same way for both (see section 6.5.3). Data that was dropped out didn't include data from these three subjects. The remainder of this section applies only to reliability analysis conducted for Vicon trials.

Before any modeling, the real marker trajectories were filtered with a Woltring filter (quintic spline). Joint centers locations were determined first before relative segment angles could be calculated. The Plug-In Gait model in Vicon uses Newington-Gage method (Davis et al. 1991) to calculate the hip joint center based on locations of pelvis markers, leg length and marker radius. Knee and ankle joint centers in the model are calculated by using a chord function (figure 21). For example, the knee joint center is determined based on the locations of hip joint center, knee and thigh markers, and knee width. The joint centers and the knee marker lie on the periphery of a circle and thigh marker lies on the plane formed by the three other points. LCS

for each segment was determined based on joint centers and marker locations after which relative segment orientations could be compared. Joint angles were calculated from Cardan angles (YXZ). (Vicon documentation 2020.) These steps were performed with an automated pipeline inside the Nexus software. Finally, sagittal joint angle data of ankle, knee and hip was exported out of the software for further analysis (see section 6.5.4).

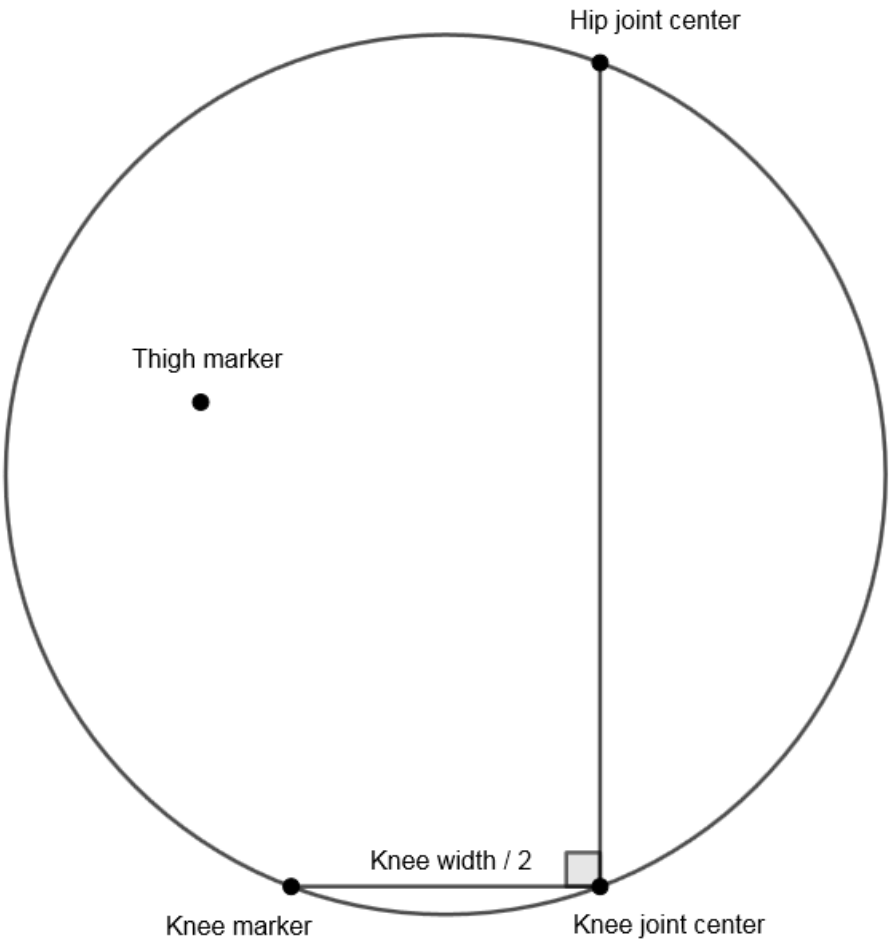


FIGURE 21. Visualization of the chord function applied to determining knee joint center (anterior view). The line between the joint centers and the line between knee joint center and knee marker form a right angle, and they all also lie on the periphery of a circle. Additionally, the thigh marker is in the same plane formed by the joint centers and the knee marker. With these conditions it's possible to calculate the location of the knee joint center.

6.5.3 2D angle analysis

Due to poor performance of the DL models in some cameras, it wasn't possible to triangulate 3D locations of the keypoints and perform 3D analysis for DeepLabCut data. Thus, a 2D analysis for the right leg was performed using data from the camera positioned perpendicular to sagittal plane of the subject (figure 20). Similarly, only the vertical and AP components of 3D marker positions from Vicon data were used in the analysis. This decision was made to try to reduce the data to 2D and eliminate the possible effect of extra dimension, if the sagittal component from 3D analysis would've been used. Data processing here was implemented in Visual Studio Code 1.56.2 using Python 3.8.3 and scientific computing library SciPy 1.6.2.

To further reduce the number of spikes from DeepLabCut data, coordinate predictions with probability of lower than 0.2 – 0.4 were removed. The number was chosen by visually inspecting the changes in the trajectories and selecting a number, which didn't remove correct data. Empty gaps were then interpolated with a cubic spline. All marker trajectories were then smoothed with a fourth-order low-pass Butterworth filter

Knee flexion angle was calculated as the angle between two lines, which were determined from the locations of markers. The first line passed through the knee and ankle markers, and the second line through the knee and thigh markers. Zero degrees of knee flexion denoted straight angle between the lines i.e. full knee extension. Ankle angle was determined similarly: the first line passed through the ankle and knee markers, and the second line through the ankle and toe markers. Zero degrees was reached when the lines formed a right angle and positive values denoted plantarflexion.

Riley et al. (2008) reported reaching stable means in kinematic and kinetic peak values in treadmill running by using average of 10-12 strides. Thus, the first 15 full gait cycles from heel strike to heel strike were averaged and normalized to 101 points to represent joint angles relative to percentage of gait cycle duration. Heel strike detection was conducted with a similar method than in Zeni et al. (2008), but with a slight modification. In their original method AP components of heel and toe markers were used to identify gait events. Because of the moving

surface, the AP component of heel marker starts to decrease immediately after contact and heel strikes can be found at the signal peaks. In healthy population walking on a treadmill, Zeni et al. (2008) showed that with this method 94 % of gait events were within one frame (0.0167 s) of the event determined from ground reaction force. Data quality of heel and toe markers from DeepLabCut were lower than of ankle marker, thus, it was decided that ankle marker would be used for detecting heel strikes.

6.5.4 Reliability analysis

The exported joint angle data from Nexus were processed in Visual Studio Code 1.56.2 using Python 3.8.3. To evaluate the effect of number of selected strides to between-trial reliability, the data was analysed by using the average of 2, 5, 10 and 15 strides. Detection of heel strikes and time normalization was conducted similarly as in previous section, but now the heel marker could be used to identify heel strikes as in the original method by Zeni et al. (2008).

6.6 Statistical analysis

Statistical parametric mapping (SPM) (Friston et al. 1995) was used to compare joint angle curves from Vicon and DeepLabCut models. Although originally used in neuroimaging, its use has increased in continuous n-dimensional biomechanical data analysis (Pataky 2010). In this study, a SPM two-tailed paired t-test ($\alpha = 0.05$) was done to assess differences between curves at each time point. A p-value was assigned for each supra-threshold cluster (adjacent points exceeding the threshold) to indicate the probability that with which the cluster could've emerged from an equally smooth random process. SPM analyses were implemented with an open-source spm1d 0.4.6 software package in Python 3.8.3.

Between-trial reliability of Vicon data was assessed with intraclass correlation coefficients (ICC) and standard errors of measurement (SEM). These methods that are intended for univariate data were extended to analyse continuous curve data as described in Pini et al. (2019). As such, both were applied pointwise to time normalized data, and reported pointwise, as well as an average over the full gait cycle. ICC analyses were implemented with an open-source

pingouin 0.3.12 statistics software package and SEM analyses with standard numerical computation, both in Python 3.8.3.

Following conventions by Shrout and Fleiss (1979), ICC(3,k) with absolute agreement was used, which was selected because the rater (Vicon) is fixed in this test-retest setting, mean of k number of strides were used and the absolute agreement between trials is of interest. The criteria used for interpreting reliability from ICC values were the following: < 0.75 poor to moderate, ≥ 0.75 good, and > 0.9 "reasonable for clinical measurements" (abbreviated as "reasonable" in text) (Portney & Watkins 2009, according to Trevethan 2017). By Harvill (1991), the SEM is defined as "the standard deviation of errors of measurement that are associated with test scores from a particular group of examinees" and expressed as:

$$SEM = SD * \sqrt{1 - ICC}$$

where SD is the standard deviation of the full sample (test and retest) and ICC is the ICC-value calculated between the test and retest samples.

7 RESULTS

Evaluating all predictions by DLC models resulted errors of 3.74 px (train) and 4.70 px (test) for DLC-M, and 50.23 px (train) and 43.99 px (test) for DLC-NoM. A sample of analysed marker coordinate trajectories from two walking trials of one subject are plotted in figure 22 (trial with markers) and figure 23 (trial without markers).

In figures 24 and 25 are ensemble graphs and results from SPM analyses for walking and running, respectively. When comparing time normalized and averaged joint angle data from walking between methods with an SPM analysis, one supra-threshold cluster was found in knee angle at 78 – 83 % of stride duration ($p < 0.001$) for DLC-M. In ankle angle, three clusters were found at 2 – 13 % ($p < 0.001$), 25 – 43 % ($p < 0.001$) and 86 – 88 % ($p = 0.015$) for DLC-M, and two clusters at 10 – 55 % ($p < 0.001$) and 82 – 90 % ($p < 0.001$) for DLC-NoM. In data from running, no clusters were found in knee angle. In ankle angle, one cluster was found at 35 – 53 % ($p < 0.001$) for DLC-M, and two clusters at 8 – 12 % ($p = 0.005$) and 65 – 80 % ($p < 0.001$) for DLC-NoM.

Time normalized and averaged (2 and 15 strides) sagittal joint angle data from Vicon are visualized in appendices 2 and 3 for walking and running, respectively. Both trials from each subject are plotted with the same colour.

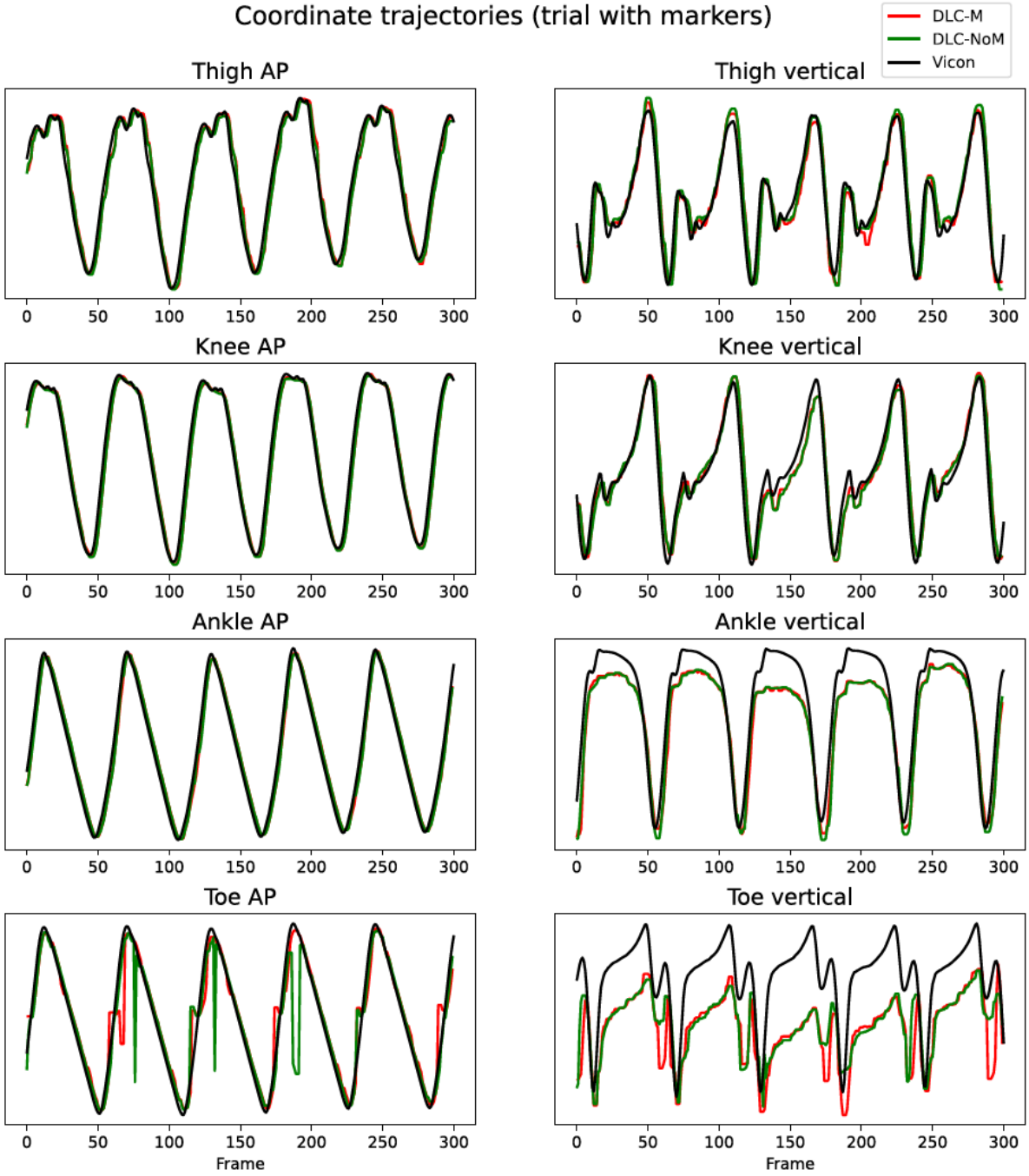


FIGURE 22. Coordinate trajectories from a 5-second portion of one walking trial where markers were attached. Due to different units in y-axis, Vicon data was scaled such that it fitted thigh marker data from DLC models. Scaling parameters were kept the same on all trajectories.

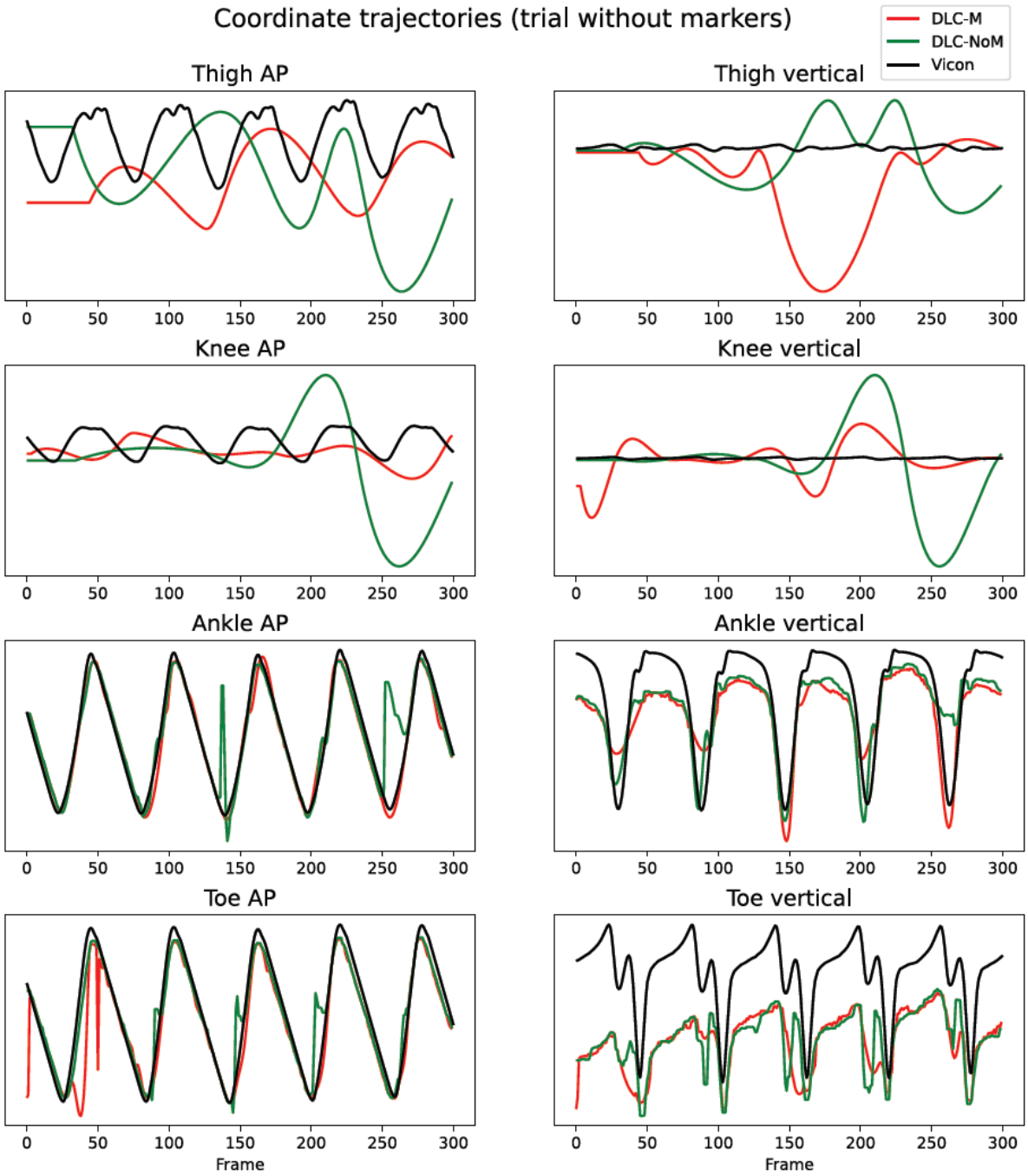


FIGURE 23. Coordinate trajectories from a 5-second portion of one walking trial where markers were removed. For reference, Vicon data from figure 22 is plotted, which is from same subject, but with markers. Same scaling parameters were used as was for data in figure 22.

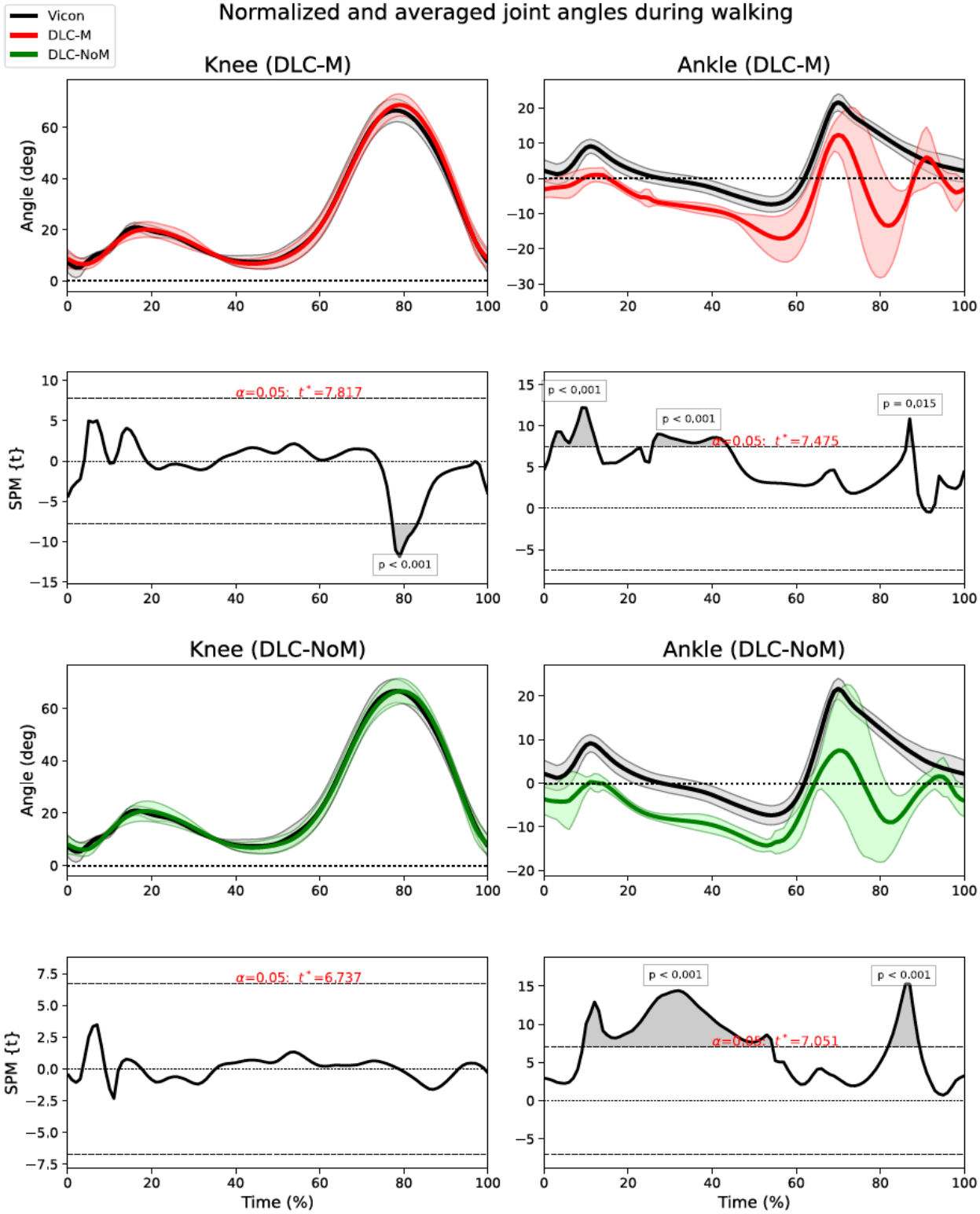


FIGURE 24. SPM analyses and ensemble graphs of joint angles during walking. In joint angle graphs the thick line represents the mean and shaded area 1 SD for each method. Shaded areas accompanied with p-values in SPM graphs represent supra-threshold clusters indicating statistically significant differences in the waveforms.

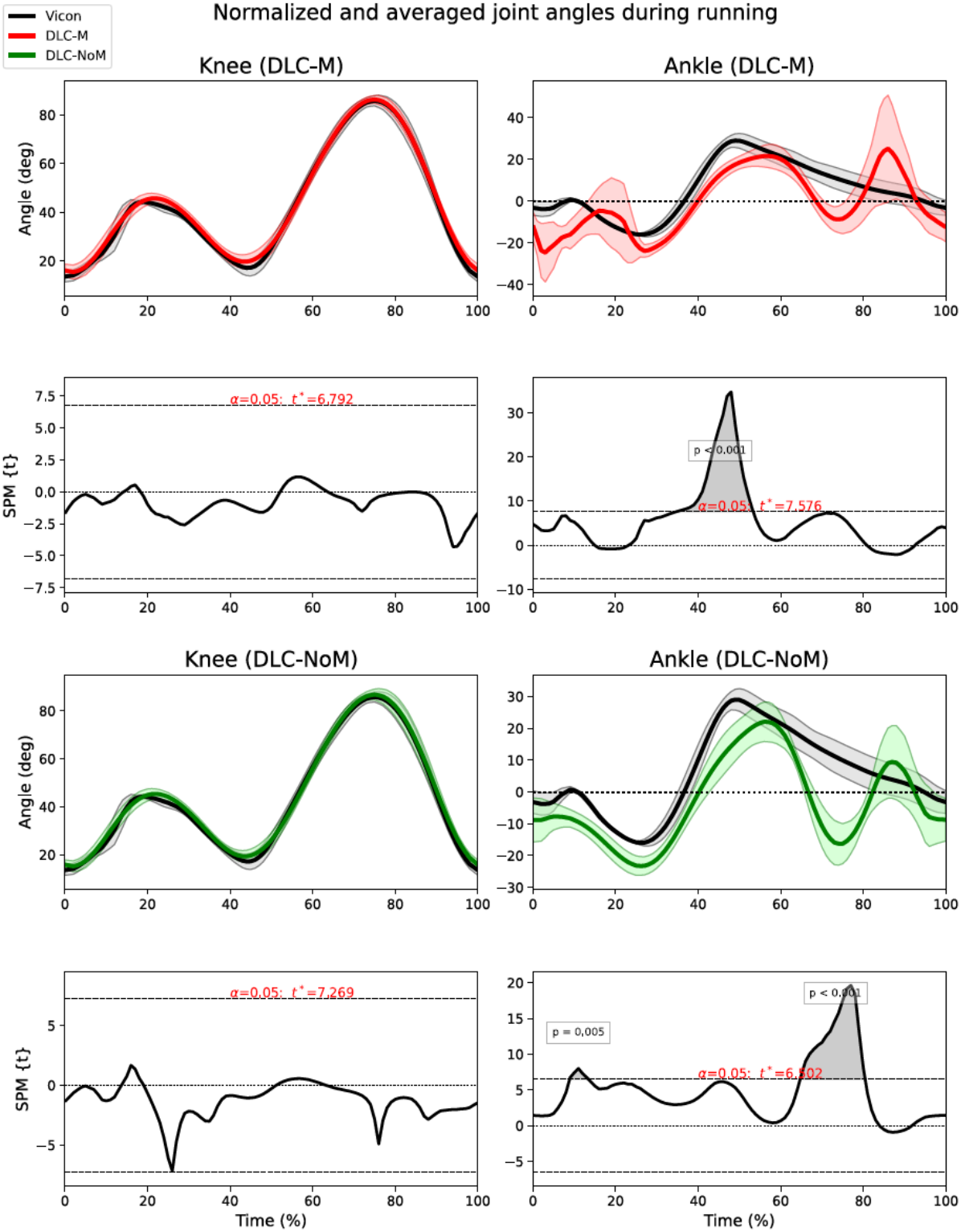


FIGURE 25. SPM analyses and ensemble graphs of joint angles during running. In joint angle graphs the thick line represents the mean and shaded area 1 SD for each method. Shaded areas accompanied with p-values in SPM graphs represent supra-threshold clusters indicating statistically significant differences in the waveforms.

All reliability analyses were performed only to Vicon data. Average ICC and SEM values of between trial sagittal joint angles averaged across different number of strides and time normalized to one gait cycle can be seen in tables 3 and 4 for walking and running, respectively. Average ICC values reached "reasonable" level of reliability (< 0.9) for all cases except for ankle when running with mean of 2 strides (0.87). The average SEM was < 1 deg for all cases when walking and for hip when running, and < 2 deg for ankle and knee when running.

Pointwise values for ICC (with 95% CIs) and SEM across whole gait cycles are plotted in figures 26 and 27 for walking and running, respectively. When looking at 15 averaged strides, the reliability according to ICC with 95% CIs was "reasonable" for knee and hip when walking and for hip when running, varied from good to "reasonable" for ankle when walking and for knee when running, and varied from poor to "reasonable" for ankle when running. Peak SEM was < 1.5 deg for all cases when walking and for hip when running, and around 3 deg for knee and ankle when running.

TABLE 3. Average ICC and SEM across whole gait cycle in walking for each joint and different number of averaged strides. Analysis was done to Vicon data.

	Ankle		Knee		Hip	
Strides	ICC	SEM (deg)	ICC	SEM (deg)	ICC	SEM (deg)
2	0.96	0.74	0.98	0.85	0.99	0.63
5	0.97	0.62	0.99	0.65	0.99	0.65
10	0.97	0.60	0.99	0.65	0.99	0.63
15	0.98	0.57	0.99	0.64	0.99	0.60

TABLE 4. Average ICC and SEM across whole gait cycle in running for each joint and different number of averaged strides. Analysis was done to Vicon data.

	Ankle		Knee		Hip	
Strides	ICC	SEM (deg)	ICC	SEM (deg)	ICC	SEM (deg)
2	0.87	1.96	0.96	1.78	0.98	0.96
5	0.92	1.52	0.97	1.59	0.98	0.84
10	0.93	1.40	0.98	1.43	0.99	0.77
15	0.94	1.35	0.98	1.40	0.99	0.75

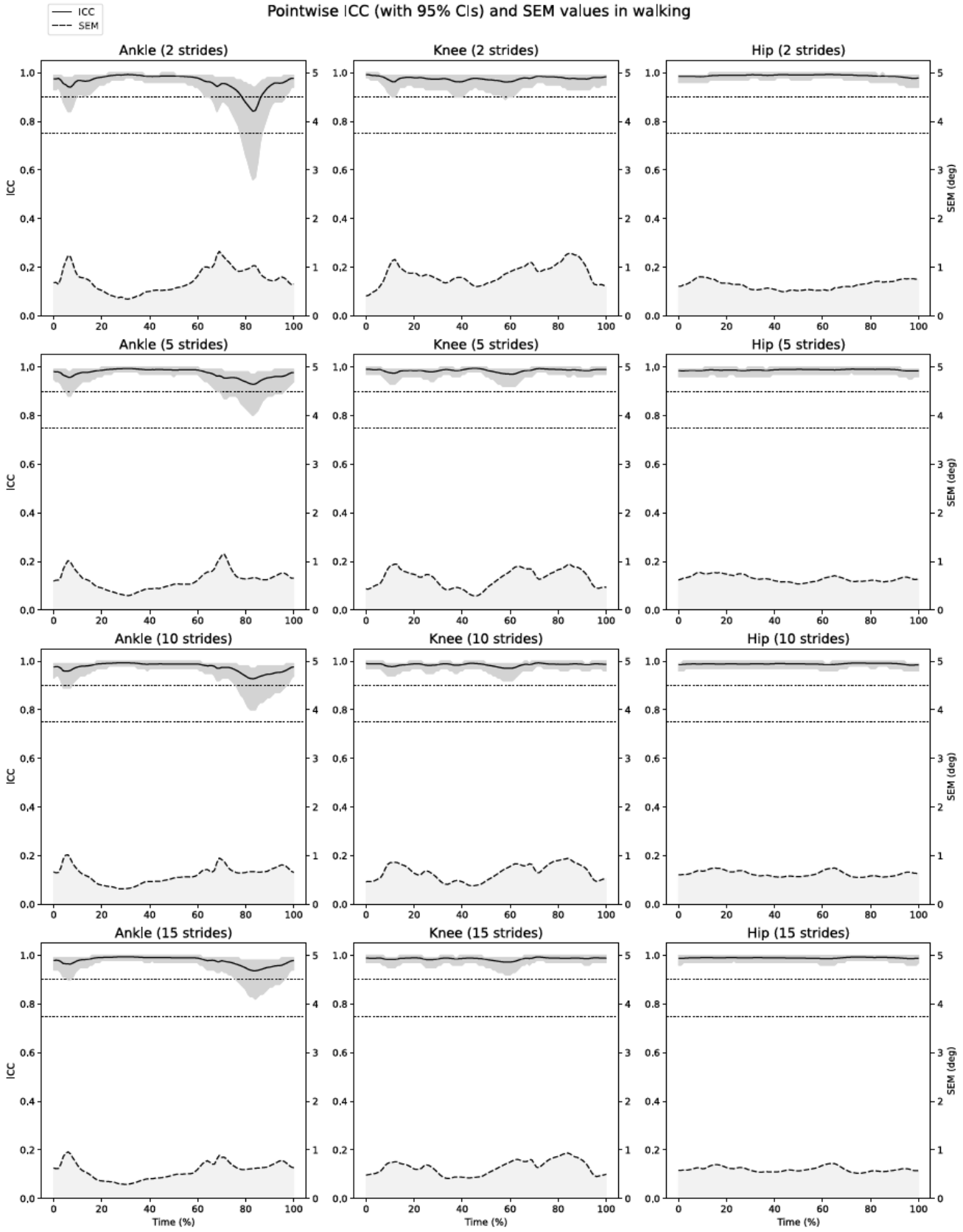


FIGURE 26. Pointwise ICCs with 95% CIs (solid line with grey areas) and SEMs (dashed line) across time normalized gait cycles from Vicon data. Dotted lines are at 0.9 and 0.75 ICC.

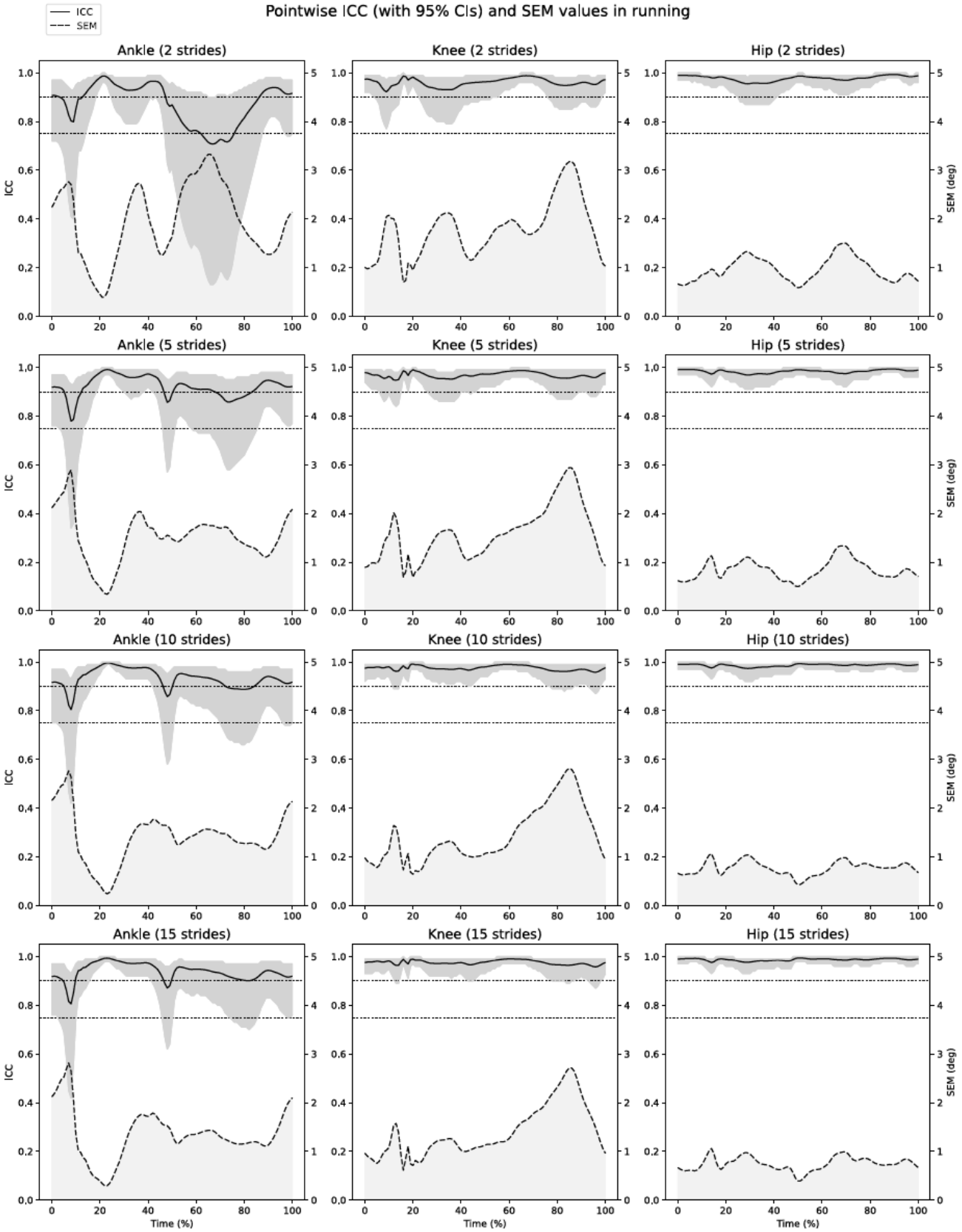


FIGURE 27. Pointwise ICCs with 95% CIs (solid line with grey areas) and SEMs (dashed line) across time normalized gait cycles from Vicon data. Dotted lines are at 0.9 and 0.75 ICC.

8 DISCUSSION

The performance of both DLC models was better when analysing trials with markers, which was against the hypothesis in the case of DLC-NoM. The quality of coordinate data from trials without markers (figure 23), especially for knee and thigh markers, was too low for angle data to show any patterns or to be reported. Agreement between both DLC models and Vicon was good in the knee data with only one supra-threshold cluster ($p < 0.001$) found in walking between DLC-M and Vicon, covering 3 % of all knee data. There were visible offsets and inconsistencies in the ankle data and multiple supra-threshold clusters were found for both DLC models, covering 27 % for DLC-M and 38 % for DLC-NoM of all the ankle data. In the reliability analysis of Vicon data, the average reliability reached "reasonable" levels of ICC in all cases when averaged over at least 5 strides. Pointwise analysis showed that lower level of 95 % CI of ICC with 15 averaged strides stayed completely or most of the time at "reasonable" level on all cases except in then ankle when running. The average SEM values in all cases reached acceptable level (< 2 deg) according to McGinley et al. (2009), but pointwise analysis revealed that SEM increased to reasonable level (2 – 5 deg) in some parts of ankle and knee data when running.

The very similar performance of both DLC models was surprising and against the expected results. It was likely and expected that DLC-M would have problems tracking the keypoints in trials where markers were removed, since it was trained with data where the keypoints were always labelled over a marker. DLC-NoM was expected to outperform DLC-M in data without markers, but that wasn't the case, when both models had poor performance on that data. The most unexpected result was the fact that DLC-NoM's performance was better with data having markers and roughly equal to DLC-M with the same data. Somehow the presence of a marker made it easier for DLC-NoM to recognize the keypoint, even though there wasn't a single frame with markers processed during training. This indicates that the models can learn features that are not intuitive to human mind.

From figure 23 it's noticeable that the models actually performed with decent accuracy on ankle and toe keypoints, with knee and thigh keypoints being the problematic ones. The smooth

appearance of knee and thigh trajectories is probably caused by the median filter applied by the algorithm in DLC. The knee and thigh keypoints didn't have any distinct landmarks around them, except the light-coloured skin and the edges of the leg. In contrast, participant's shoe was always there as a landmark for ankle and toe keypoints, no matter if there was a marker or not. This might explain the observed performance difference between keypoints. Also, training a single model for all cameras might've caused there to be too much variation in the labelled data and made the models unable to generalize well to all cameras. It can only be speculated, if the performance of models would be different, if the models would've been trained with data only from one camera.

The SPM analyses of the knee angle show that both models performed almost identically to Vicon, which indicates that DL -based human pose estimation methods have potential to be used in gait analysis. Although, it must be considered that only six trials in total were analysed from three subjects by both methods, which makes any statistical inference unreliable. The SPM analyses of ankle show that there were significant differences in the data at multiple locations. These observations are partly similar to sagittal angles obtained by Stenum et al. (2020), where data from knee correlated more strongly with data from criterion method, than ankle data (figure 13). There is also a noticeable offset between methods, where data from DLC models show more dorsiflexion across most of the data. This is most likely due to projection error from the GoPro camera. The location of the camera was approximately at the hip height and relatively close to the subject. This caused the keypoints from the foot to be projected at an angle relative to true sagittal plane, because the foot was low in the camera's view. This effect can be seen from figures 22 and 23, where the range of vertical component of ankle and toe is clearly smaller than in Vicon's data. AP component is unaffected probably because the subject is horizontally in the middle of the frame, which doesn't cause distortion or projection error in AP direction. Even though this offset would be fixed, there are still clear differences in the pattern of data in late stride during swing phase. During swing phase of the right leg, the feet of the subject pass each other such that both feet are visible to the camera, and this caused the models to accidentally label keypoints from the left foot during the pass. This is probably one reason for the differences in the patterns of ankle data.

Before discussing the results from the reliability analysis of Vicon data, the meaning of reliability in this case should be made clear. The initial purpose of the reliability analysis in this study was to investigate what is the intrasubject variability between trials during the same session. Then, if we could've analysed trials without markers with DLC models, we would've had data about the intrinsic variability in subject's gait and could've made estimates on how much of the difference between data from Vicon and DLC models is due to the method and not the subject when analysing different trials. With this kind of setting for reliability test, the variability in the data is not due to different raters, because the same Vicon system is used to analyse both trials and the markers aren't reattached for the second trial. The participants were instructed only before starting the protocol to walk or run normally, thus, it can be assumed that the variability isn't originating from different movement strategies. The variability is assumed to originate from instrument error in the Vicon system, STA and the stride-to-stride differences in the subject's gait. Topley & Richards (2020) have shown that modern optoelectronic systems can measure distances with an average error of 0.6 mm and, additionally, instrumental errors can be partly removed by filtering. If instrument error is assumed to be negligible, the variability in the data is due to STA and stride-to-stride differences. Optimization algorithms can be used to reduce STA (Lu & O´Connor 1999) and the Plug-In Gait model in Nexus probably implements an algorithm for this, but it must be recognized that it's part of the observed variability. Therefore, the reliability results represent how much there is variability due to intraindividual stride differences and STA.

The between-trial reliability analysis showed that sagittal plane angle data from ankle, knee and hip reached "reasonable for clinical measurements" -level (> 0.9), when averaging over at least five strides and considering only the average ICC value. However, the reliability is different depending on which discrete point at the curve is selected. The pointwise graphs (figures 26 and 27) show that the ankle data is least reliable in both gait types, but especially when running. Because of the nature of Plug-In Gait model, the errors propagate downstream from hip to ankle (Schwartz et al. 2004) and might explain partly the lower reliability in the ankle. Also, there is lower reliability in the ankle after the foot hits the ground, which might be caused by STA when the accelerations of foot markers are high. The 95 % confidence intervals for the ankle in running show that reliability could decrease to poor to moderate levels in some parts of the stride. Increasing the number of used strides did only have minor effect on the averaged ICC

levels, but qualitatively increased the reliability when looking at the pointwise confidence bands, especially in the ankle.

In addition to ICC, it's important to know what the absolute reliability in the data is, if clinical conclusions would be made based on the data. The SEM was selected to describe the absolute reliability in this study. In walking, SEMs were low even with only two strides and pointwise graphs show that with 15 strides the SEMs were ≤ 1 deg throughout the stride. In running, there were distinct peaks in the SEM for ankle and knee, while SEM in the hip stayed relatively constant throughout the stride, especially with more strides. The highest SEMs were close to 3 degrees for ankle and knee. These results clearly show that the average measures of reliability over the full curve can "hide" unreliable parts of the curve. Using criteria by McGinley et al. (2009), The SEM values in walking stayed at acceptable level (< 2 deg) across the stride, but in running the SEM increased occasionally to reasonable level (2 – 5 deg) in ankle and knee. The SEM can also be used to estimate minimum detectable change ($SEM * Z * \sqrt{2}$), which can be helpful when determining if an intervention made a real difference to subject's gait (Weir 2005). If using 95% CIs (Z = 1.96), then minimum detectable change would vary between 1.7 and 8.3 degrees with SEM values between 0.5 and 3 degrees. This almost five-fold difference shows that if one would want to make conclusions about if there is a real difference between data, using average SEM value could potentially raise type 1 or 2 errors.

This was the first study to the author's knowledge that attempted to track anatomically valid keypoints in human gait in 3D with neural network models by retraining already pretrained networks with manually labelled data. Van Den Bogaart et al. (2020) did successfully track 3D locations of markers on a subject with eight cameras, but in a static setting where the subject stood on a balance board. Also, this was the first study trying to apply a biomechanically similar model compared to criterion method for assessing the performance of those models. In previous studies (Nakano et al. 2020; Stenum et al. 2020; Zago et al. 2020; Needham et al. 2021), the keypoints predicted by the DL -based model were different from the marker locations of criterion method and therefore also the kinematic model didn't match with the criterion. However, after labelling and training the models in this study, their performance wasn't good enough to accurately track the keypoints from all camera views, therefore, it wasn't possible to

triangulate 3D locations of the keypoints and use those as input to the biomechanical model. There are multiple reasons related to environment, data collection methods, training data and network parameters, which can be considered to have had an effect to the final performance of the models.

The environment, specifically lighting and clothing, could've affected the visibility and distinctiveness of markers. From figure 28 it's clearly visible that markers are dimmer in some frames and e.g. light-coloured shoes could make it hard to distinguish markers from the foot, although that's not the case in these examples. The properties and settings of the cameras used in data collection are also related to the quality of marker data. It's apparent from the frames in figure 28, that the resolution of cameras makes the markers only few pixels in size for the cameras further away, although the resolution was relatively high at 1920x1080. Also, the capture rate of 60 Hz is too low to completely avoid motion blur, mostly in running trials. It's hard to say if better lighting and more accurate cameras would've changed the outcome in this study, but this shows that with nonoptimal quality the models couldn't perform as hoped.

The fact that both models were trained with data from all of the cameras and expected to generalize well to all of the camera views might have been too challenging task for the models, at least with the amount and quality of training data in this study. It has been shown that 300-400 frames (Cronin et al. 2019) and even just 200 frames (Mathis et al. 2018) have been enough to reach human-level labelling accuracy in 2D settings. Also, Van Den Bogaart et al. (2020) successfully used a total of only 440 frames from four cameras in their study, but their measurement condition was static. Although there were about 416 and 831 frames per camera view in this study, which should be enough for a single 2D view, it might not have been enough for training a single model for eight 2D views capturing dynamic movement. It's also possible that the amount of training data wasn't the problem per se, but other properties and quality of it. Full size frames were fed into the network during training without additional cropping, which caused the actual subject to cover only a small area of the frame. Therefore, there was a lot of "noise" in the data. Additionally, due to lack of GPU computing power, the trial videos during analysis had to be cropped before they were analysed. In practice this meant that the training data didn't fully correspond to the data that was analysed in the final stage. However, cropping

only removed part of the “noise” around the meaningful data and probably didn’t affect the performance significantly.

FIGURE 28. Frames from different GoPro-cameras from the same subject. These are zoomed in views so that the subject is roughly the same size in all of them.

The network chosen for a model and parameters of training can also affect the performance of the model. Only one network, ResNet-101, was tested and it's possible that the outcome could've been different with some other network. However, the chosen network has one of the more complex structures and should be able to learn complex patterns. Training parameters of the model, such as batch size, number of iterations, learning rate and decay of learning rate, can also potentially change the performance of the model. By training several models with different networks and training parameters, it would be possible to compare their performance on the same data and select the one that has the best performance. Thus, with more time devoted to fine tuning the model, it would've been possible to rule out the possibility that the best model wasn't used for the data.

Based on the observations from this study, some recommendations and suggestions can be made for future studies attempting to use deep learning methods for gait analysis in a similar way. It should be made sure that the quality of captured data is adequate by having proper lighting and cameras with appropriate resolution and frame rate for the setup in question. However, it's always dependent on the goal: if the model is required to perform in a poorly lit environment and with motion blur, then the training data should represent these requirements. Also, to make sure that the model is appropriate and the best option available, some expertise and time is needed for fine tuning the models' parameters. It's also probably a good idea to try and minimize the amount of unnecessary data being passed into the network when training. In this study, cropping the data before labelling could have made it easier for the models to learn and perform better. Another option to pursue would be to train individual models for each camera view while this has stronger support from previous studies. As Cronin et al. (2019) and Mathis et al. (2018) have shown, labelling accuracies by the model are comparable to human when a model is trained with 200-400 manually labelled frames from the same view.

9 CONCLUSION

Based on the 2D analyses in this study it can be stated that the presence of markers in the training data didn't have a clear effect on the performance of those models. Both models failed to track keypoints in data where markers were taken off, which was surprising for DLC-NoM that was trained without markers present in the data. DLC-M performed slightly better than DLC-NoM in ankle data, when comparing the total coverage of supra-threshold clusters of the time normalized data from SPM analyses (27 % vs 38 %). The opposite was true for knee data, but the absolute coverage of supra-threshold clusters was low for both models (3 % vs 0 %). Although the 3D analysis failed due to poor quality in keypoint predictions from most of the camera views, this study showed that DL -based models can be used to calculate lower-body sagittal plane angles with good accuracy, if the conditions are appropriate. This study also showed that training a DL -based model for 3D gait analysis isn't a trivial task or necessary possible with nonoptimal training data, when aiming for high anatomical and biomechanical accuracy.

The reliability analysis conducted on Vicon data showed that the between-trial reliability depends strongly on what part of the stride is observed, especially in running. The reliability during walking reached best criteria levels based on ICC and SEM almost throughout the full gait cycle in all joints. There was more variability in reliability values in running, especially in the ankle, but mostly the reliability stayed inside the best criteria level with few drops below it at some points of gait cycle. The ICC and SEM values averaged over the pointwise values across full stride duration only represent the mean, which could lead to type 1 or 2 errors when inferring conclusions from data. The pointwise values of reliability showed that reliability is different at different points of the gait cycle, and that it's probably not the best practice to use mean reliability values from full gait cycle when making statistical inference on discrete points of the gait cycle. Also, a reliability value at some point of the stride shouldn't be used at other points of the stride.

REFERENCES

Abdel-Aziz, Y. I. & Karara, H. M. 1971. Direct Linear Transformation from Comparator Coordinates into Object Space Coordinates in Close-Range Photogrammetry. Proceedings of the Symposium on Close-Range Photogrammetry, 26–29 January 1971, Urbana, IL, 1 – 18.

Alem, N. M., Melvin, J. & Holstein, G. L. 1978. Biomechanics applications of direct linear transformation in close-range photogrammetry. Proceedings of the Sixth New England Bioengineering Conference, pp. 202 – 206.

Allard, P., Stokes, I. A. F. & Blanchi, J-P. 1995. Three-dimensional analysis of human movement. Champaign, IL: Human Kinetics.

Allard, P., Cappozzo, A., Lunberg, A. & Vaughan, C. L. 1998. Three-dimensional analysis of human locomotion. Chichester, UK: John Wiley & Sons.

Andriluka, M., Pishchulin, L., Gehler, P. & Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. IEEE Conference on Computer Vision and Pattern Recognition, 3686 – 3693.

Baker, R. 2013. Measuring walking: a handbook of clinical gait analysis. London: Mac Keith Press.

Baker, R., Leboeuf, F., Reay, J. & Sangeux, M. 2017. The Conventional Gait Model: The Success and Limitations. In: Müller, B. & Wolf, S. 2018. Handbook of Human Motion, pp. 489 – 508. Cham: Springer.

Bengio, Y. 2009. Learning Deep Architectures for AI. Foundations and Trends in Machine Learning 2 (1), 1 – 127.

Bishop, C. M. 2006. Pattern recognition and machine learning. New York: Springer.

Bulat, A., Kossaifi, J., Tzimiropoulos, G. & Pantic, M. 2020. Toward fast and accurate human pose estimation via soft-gated skip connections. arXiv:2002.11098.

Camomilla, V., Dumas, R. & Cappozzo, A. 2017. Human movement analysis: the soft tissue artefact issue. Journal of Biomechanics 62, 1 – 4.

Cao, Z., Simon, T., Wei, S-E. & Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1302 – 1310.

Cao, Z., Hidalgo, G., Simon, T., Wei, S-E. & Sheikh, Y. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv:1812.08008.

Cappozzo, A. 1991. Three-dimensional analysis of human walking: experimental methods and associated artifacts. Human Movement Science 10, 589 – 602.

Cappozzo, A., Catani, F., Della Croce, U. & Leardini, A. 1995. Position and orientation in space of bones during movement: anatomical frame definition and determination. Clinical Biomechanics 10 (4), 171 – 178.

Cappozzo, A., Catani, F., Leardini, A., Benedetti, M. G. & Della Croce, U. 1996. Position and orientation in space of bones during movement: experimental artefacts. Clinical Biomechanics 11 (2), 90 – 100.

Carreira, J., Agrawal, P., Fragkiadaki, K. & Malik, J. 2016. Human Pose Estimation with Iterative Error Feedback. arXiv:1507.06550.

Chen, Y., Tian, Y. & He, M. 2020. Monocular Human Pose Estimation: A Survey of Deep Learning-based Methods. Computer Vision and Image Understanding 192, 102897.

Chiari, L., Croce, U. D., Leardini, A. & Cappozzo, A. 2005. Human movement analysis using stereophotogrammetry: Part 2: Instrumental errors. Gait & Posture 21 (2), 197 – 211.

Colyer, S. L., Evans, M., Cosker, D. P. & Salo, A. I. T. 2018. A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. Sports Medicine Open 4 (1), 24.

Cronin, N. J., Rantalainen, T., Ahtiainen, J. P., Hynynen, E. & Waller, B. 2019. Markerless 2D kinematic analysis of underwater running: A deep learning approach. Journal of Biomechanics 87, 75 – 82.

Davis, R. B., Õunpuu, S., Tyburski, D. & Cage, J. R. 1991. A gait analysis data collection and reduction technique. Human Movement Science 10, 575 – 587.

Deng, J., Dong, W., Socher, R., Li, L-J., Li, K. & Li, F-F. 2009. Imagenet: A large-scale hierarchical image database. IEEE Conference on computer vision and pattern recognition, 248 – 255.

Desmarais, Y., Mottet, D., Slangen, P. & Montesinos, P. 2020. A review of 3D human pose estimation algorithms for markerless motion capture. arXiv:2010.06449.

Dinn, D. F., Winter, D. A. & Trenholm, B. G. 1970. CINTEL-Computer Interface for Television. IEEE Transaction on Computers C-19 (11), 1091 – 1095.

Diss, C. E. 2001. The reliability of kinetic and kinematic variables used to analyse normal running gait. Gait & Posture 14, 98 – 103.

Dumoulin, V. & Visin, F. 2018. A guide to convolution arithmetic for deep learning. arXiv:1603.07285v2.

Ferrigno, G. & Pedotti, A. 1985. Elite: A Digital Dedicated Hardware System for Movement Analysis Via Real-Time TV Signal Processing. IEEE Transactions on Biomedical Engineering BME-32 (11), 943 – 950.

Fiker, R., Kim, L. H., Molina, L. A., Chomiak, T. & Whelan, P. J. 2020. Visual Gait Lab: A user-friendly approach to gait analysis. Journal of Neuroscience Methods 341.

Friston, K. J., Holmes, A., Worsley, K., Poline, J., Frith, C. & Frackowiak, R. 1994. Statistical parametric maps in functional imaging: a general linear approach. Human Brain Mapping 2, 189 – 210.

Goodfellow, I., Bengio, Y. & Courville, A. 2016. Deep learning. Cambridge, Massachusetts: The MIT Press.

Google Cloud. Retrieved on 18.11.2020 from: https://cloud.google.com/gpu.

Harvill, L. M. 1991. An NCME Instructional Module on. Educational Measurement: Issues and Practice 10, 33 – 41.

He, K., Zhang, X., Ren, S. & Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M. & Schiele, B. 2016. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In European Conference on Computer Vision 34 – 50.

Ionescu, C., Papava, D., Olaru, V. & Sminchisescu, C. 2014. Human3.6m: Largescale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 1325 – 1339.

Iqbal, U., Milan, A. & Gall, J. 2016. Posetrack: Joint multi-person pose estimation and tracking. arXiv:1611.07727.

Iskakov, K., Burkov, E., Lempitsky, V. & Malkov, Y. 2019. Learnable Triangulation of Human Pose. arXiv:1905.05754.

Jain, A., Tompson, J., LeCun, Y. & Bregler, C. 2014. MoDeep: A Deep Learning Framework Using Motion Features for Human Pose Estimation. arXiv:1409.7963.

Jarrett, M. O. 1976. A television/computer system for human locomotion analysis. Doctoral thesis.

Johnson, S. & Everingham, M. 2010. Clustered pose and nonlinear appearance models for human pose estimation. Proceedings of the British Machine Vision Conference, 1 – 11.

Kanko, R., Laende, E., Selbie, S. & Deluzio, K. 2020a. Inter-session repeatability of Theia3D markerless motion capture gait kinematics. bioRxiv 2020.06.23.155358.

Kanko, R., Strutzenberger, G., Brown, M., Selbie, S. & Deluzio, K. 2020b. Assessment of spatiotemporal gait parameters using a deep learning algorithm-based markerless motion capture system. https://doi.org/10.31224/osf.io/j4rbg.

Karashchuk, P., Rupp, K. L., Dickinson, E. S., Sanders, E., Azim, E., Brunton, B. W. & Tuthill, J. C. 2020. Anipose: a toolkit for robust markerless 3D pose estimation. bioRxiv 2020.05.26.117325.

Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., Baber, U., Min, J. K., Tang, W. H. W., Halperin, J. L. & Narayan, S. M. 2019. Deep learning for cardiovascular medicine: A practical primer. European Heart Journal. 40 (25), 1 – 15.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25 (2), 1097 – 1105.

Leardini, A., Belvedere, C., Nardini, F., Sancisi, N., Conconi, M. & Parenti-Castelli, V. 2017. Kinematic models of lower limb joints for musculo-skeletal modelling and optimization in gait analysis. Journal of Biomechanics 62, 77 – 86.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. Neural Computation 1 (4), 541 – 551.

Lindholm L.E. 1974. An optoelectronic instrument for remote on-line movement monitoring. In: Nelson R. C. & Morehouse C. A. 1974. Biomechanics IV. International Series on Sport Sciences. Palgrave, London.

Lu, T.-W. & O´Connor, J. J. 1999. Bone position estimation from skin marker co-ordinates using global optimisation with joint constraints. Journal of Biomechanics 32, 129 – 134.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W. & Bethge, M. 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature neuroscience 21, 1281 – 1289.

McGinley, J. L., Baker, R., Wolfe, R. & Morris, M. E. 2009. The reliability of three-dimensional kinematic gait measurements: a systematic review. Gait & posture, 29 (3), 360 – 369.

Medved, V. 2001. Measurement of human locomotion. Boca Raton, FL: CRC Press.

Meyer, C., Killeen, T., Easthope, C. S., Curt, A., Bolliger, M., Linnebank, M., Zörner, B. & Filli, L. 2019. Familiarization with treadmill walking: How much is enough? Scientific Reports 9 (1).

Moeslund, T. B., Hilton, A. & Krüger, V. 2006. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104, 90 – 126.

Monaghan, K., Delahunt, E. & Caulfield, B. 2007. Increasing the number of gait trial recordings maximises intra-rater reliability of the CODA motion analysis system. Gait & Posture 25, 303 – 315.

Moro, M., Marchesi, G., Odone, F. & Casadio, M. 2020. Markerless gait analysis in stroke survivors based on computer vision and deep learning: a pilot study. Proceedings of the 35th Annual ACM Symposium on Applied Computing.

Murphy, K. P. 2012. Machine learning: A probabilistic perspective. Cambridge, MA: MIT Press.

Muybridge, E. 1955. The human figure in motion. Mineola, NY: Dover Publications, Inc.

Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M. & Mathis, M. W. 2019. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nature Protocols 14 (3), 1 – 25.

Nakano, N., Sakura, T., Ueda, K., Omura, L., Kimura, A., Iino, Y., Fukashiro, S. & Yohsioka, S. 2020. Evaluation of 3D Markerless Motion Capture Accuracy Using OpenPose With Multiple Video Cameras. Frontiers in Sports and Active Living 2, article 50.

Needham, L., Evans, M., Cosker, D. P., Wade, L., McGuigan, P. M., Bilzon, J. L. & Colyer, S. L. 2021. Human Movement Science in The Wild: Can Current Deep-Learning Based Pose 1 Estimation Free Us from The Lab? bioRxiv 2021.04.22.440909.

Nielsen, M. A. 2015. Neural Networks and Deep Learning. Determination Press. http://neuralnetworksanddeeplearning.com/

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C. & Murphy, K. 2017. Towards Accurate Multi-person Pose Estimation in the Wild. arXiv:1701.01779.

Pataky, T. C. 2010. Generalized n-dimensional biomechanical field analysis using statistical parametric mapping. Journal of Biomechanics 43 (10), 1976 – 1982.

Pini, A., Markström, J. L. & Schelin, L. 2019. Test-retest reliability measures for curve data: an overview with recommendations and supplementary code. Sports Biomechanics.

Pishchulin, L., Jain, A., Andriluka, M., Thormählen, T. & Schiele, B. 2012. Articulated people detection and pose estimation: reshaping the future. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition.

Portney, L. G. & Watkins, M. P. 2009. Foundations of Clinical Research: Applications to Practice. 3th ed. Upper Saddle River, NJ: Pearson Education.

Richards, J. G. 1999. The measurement of human motion: A comparison of commercially available systems. Human Movement Science 18 (5), 589 – 602.

Riley, P. O., Dicharry, J., Franz, J., Croce, U. D., Wilder, R. P. & Kerrigan, D. C. 2008. A Kinematics and Kinetic Comparison of Overground and Treadmill Running. Medicine & Science in Sports & Exercise 40 (6), 1093 – 1100.

Robertson, D. G. E., Caldwell, G. E., Hamill, J., Kamen, G. & Saunders, N. W. 2014. Research methods in biomechanics. 2th ed. Champaign, IL: Human Kinetics.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpahty, A., Khosla, A., Bernstein, M., Berg, A. C. & Li, F-F. 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115, 211 – 252.

Sapp, B. & Taskar, B. 2013. Modec: Multimodal decomposable models for human pose estimation. IEEE Conference on Computer Vision and Pattern Recognition, 3674 – 3681.

Schwartz, M. H., Trost, J. P. & Wervey, R. A. 2004. Measurement and management of errors in quantitative gait data. Gait & Posture 20, 196 – 203.

Seethapathi, N., Wang, S., Saluja, R., Blohm, G. & Kording, K. P. 2019. Movement science needs different pose tracking algorithms. arXiv:1907.10226.

Sharma, Sid., Sharma, Sim. & Athaiya, A. 2020. Activation functions in neural networks. International Journal of Engineering Applied Sciences and Technology 4 (12), 310 – 316.

Sheshadri, S., Dann, B., Hueser, T. & Scherberger, H. 2020. 3D reconstruction toolbox for behavior tracked with multiple cameras. Journal of Open Source Software 5 (45), 1849.

Shrout, P. E. & Fleiss, J. L. 1979. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86 (2), 420 – 428.

Sigal, L., Balan, A. O. & Black, M. J. 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International journal of computer vision 87, 4 – 27.

Smith, P. N., Refshauge, K. M. & Scarvell, J. M. Development of the concepts of knee kinematics. Archives of Physical Medicine and Rehabilitation 84 (12), 1895 – 1902.

Stenum, J., Rossi, C. & Roemmich, R. T. 2020. Two-dimensional video-based analysis of human gait using pose estimation. bioRxiv 2020.07.24.218776.

Sun, X., Shang, J., Liang, S. & Wei, Y. 2017. Compositional Human Pose Regression. arXiv:1704.00159.

Sutherland, D. H. 2002. The evolution of clinical gait analysis: Part II Kinematics. Gait & Posture 16, 159 – 179.

Topley, M. & Richards, J. G. 2020. A comparison of currently available optoelectronic motion capture systems. Journal of Biomechanics 106.

Toshev, A. & Szegedy, C. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. IEEE Conference on Computer Vision and Pattern Recognition, 1653 – 1660.

Touvron, H., Vedaldi, A., Douze, M. & Jégou, H. 2020. Fixing the train-test resolution discrepancy: FixEfficientNet. arXiv:2003.08237.

Trevethan, R. 2017. Intraclass correlation coefficients: Clearing the air, extending some cautions, and making some requests. Health Services and Outcomes Research Methodology 17 (2), 127 – 143.

Van Den Bogaart, M., Bruijn, S. M., Spildooren, J., van Dieën, J. H. & Meyns, P. 2020. Using deep learning to track 3D kinematics. Gait & Posture 81, 369 – 370.

Van Der Kruk, E. & Reijne, M. M. 2018. Accuracy of human motion capture systems for sport applications; state-of-the-art review. European journal of sport science 18 (6), 806 – 819.

Van Hooren, B., Fuller, J. T., Buckley, J. D., Miller, J. R., Sewell, K., Rao, G., Barton, C., Bishop, C. & Willy, R. W. 2020. Is Motorized Treadmill Running Biomechanically

Comparable to Overground Running? A Systematic Review and Meta-Analysis of Cross-Over Studies. Sports Medicine 50 (4), 785 – 813.

Vicon documentation for Vicon Nexus 2.5. Retrieved on 20.10.2020 from: https://docs.vicon.com/display/Nexus25/Nexus+Documentation.

Weir, J. P. 2005. Quantifying Test-Retest Reliability Using the Intraclass Correlation Coefficient and the SEM. The Journal of Strength and Conditioning Research, 19 (1), 231 – 240.

Whittle, M. 2007. Gait analysis: An introduction. 4th ed. Edinburgh; NY: Butterworth-Heinemann.

Winter, D. A., Greenlaw, R. K. & Hobson, D. A. 1972. Television-Computer Analysis of Kinematics of Human Gait. Computers and Biomedical Research 5 (5), 498 – 504.

Winter, D. A., Quanbury, A. Q, Hobson, D. A., Sidwall, H. G., Reimer, G., Trenholm, B. G., Steinke, T. & Shlosser, H. 1974. Kinematics of Normal Locomotion - A Statistical study based on T. V. Data. Journal of Biomechanics 7 (6), 479 – 486.

Winter, D. A. 2009. Biomechanics and motor control of human movement. 4th ed. Hoboken, NJ: John Wiley & Sons, Inc.

Woltring, H. J. 1980. Planar control in multi-camera calibration for three-dimensional gait studies. Journal of Biomechanics 13 (1), 39 – 48.

Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. 2014. How transferable are features in deep neural networks?. Advances in Neural Information Processing Systems 27, 3320 – 3328.

Zago, M., Luzzago, M., Marangoni, T., De Cocco, M., Tarabini, M. & Galli, M. 2020. 3D Tracking of Human Motion Using Visual Skeletonization and Stereoscopic Vision. Frontiers in Bioengineering and Biotechnology 8, article 181.

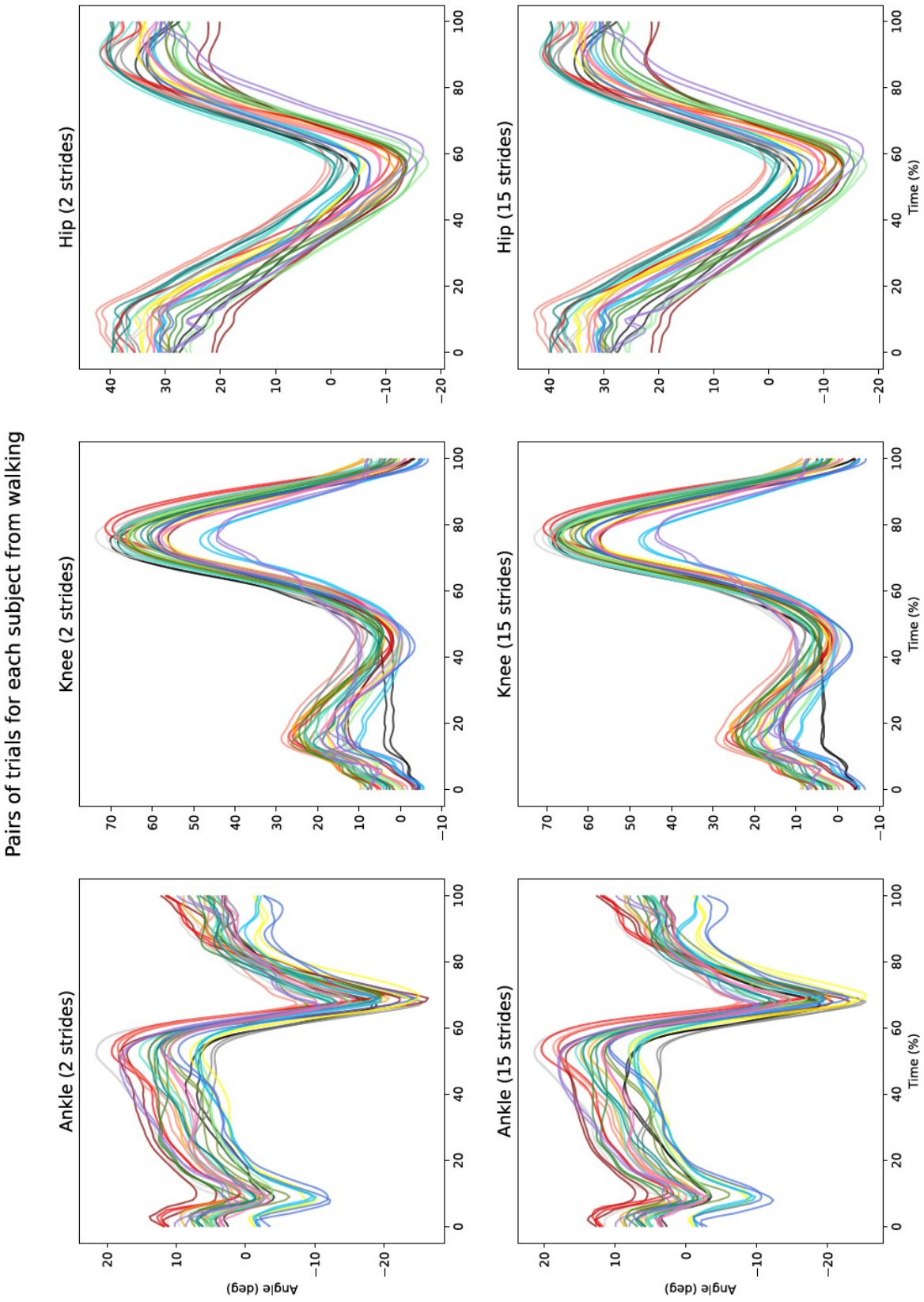
Zeni, J. A., Richards, J. G. & Higginson, J. S. 2008. Two simple methods for determining gait events during treadmill and overground walking using kinematic data. Gait & Posture 27 (4), 710 – 714.

APPENDICES

APPENDIX 1. A template of measurement protocol. Velocities were chosen arbitrary for each participant by the researchers so that the order of velocities was different during each block of walking or running.

Time	Duration	Gait type	Velocity
0 – 6	6	Walking	
6 – 9	3	Walking	
9 – 12	3	Walking	
12 – 14	2	REST	
14 – 20	6	Running	
20 – 23	3	Running	
23 – 26	3	Running	
26 – 30	4	REST	
30 – 34	4	Walking	
34 – 37	3	Walking	
37 – 40	3	Walking	
40 – 42	2	REST	
42 – 46	4	Running	
46 – 49	3	Running	
49 – 52	3	Running	
52 – 56	4	REST (remove markers)	
56 – 60	4	Walking	
60 – 63	3	Walking	
63 – 66	3	Walking	
66 – 68	2	REST	
68 – 72	4	Running	
72 – 75	3	Running	
75 – 78	3	Running	

APPENDIX 2. Time normalized and averaged sagittal joint angle data from walking measured with Vicon. Both trials from each subject is represented with one colour.



APPENDIX 3. Time normalized and averaged sagittal joint angle data from running measured with Vicon. Both trials from each subject is represented with one colour.