Evaluation of DeepLabCut as a Human Markerless Motion Capture Tool

by

W. Seth Daley

Submitted in partial fulfilment of the requirements

for the degree of Master of Science

at

Dalhousie University

Halifax, Nova Scotia

August, 2023

Dalhousie University is located in Mi'kma'ki, the

ancestral and unceded territory of the Mi'kmaq.

We are all Treaty people

Table of Contents

List of Tables

List of Figures

Abstract

There are a variety of motion capture methods available; however, many of them are not well suited for collections outside a laboratory setting. AI markerless motion capture may fit this need, but its implementation and accuracy need to be better understood. Therefore, the purpose of this research was to evaluate the tracking capabilities of DeepLabCut and conditions (complexity of the feature set and camera setup) that can affect its performance. Two markerless networks, a common joint center tracking set and a complex feature set, were trained using 40 participants completing 6 movements that were recorded by 8 cameras. Network retraining and performance evaluation (tested with 10 participants) occurred 3 times for each network. The results from this markerless motion capture research highlight the importance of choosing minimally occluded features of interest and camera positions that maximize the number of frames where the full feature set is visible.

List of Abbreviations Used

PSIS	Posterior Superior Iliac Spine
ASIS	Anterior Superior Iliac Spine
SFS	Simplified Feature Set network
CFS	Complex Feature Set network
DLC	DeepLabCut
IMU	Inertial Measurement Unit
2D	Two Dimensional
3D	Three Dimensional
EMS	Electromagnetic Motion analysis Systems
RMS	Root Mean Squared
RMSD	Root Mean Squared Difference
ISB	International Society of Biomechanics

Acknowledgements

I would like to start by expressing my thanks and gratitude to my supervisor, Dr Ryan Frayne, for his guidance and support through this process. His dedication to his students is inspiring and has fostered in me a love of learning and teaching that I had not anticipated. Thanks must also be expressed to Freddy who shared his dad's time with us and joined in for the occasional meeting.

I would like to extend thanks to everyone in the BENlab at Dalhousie for their support and accommodation during this process, to my lab mates Jessica Reddin and Dylan Sutherland for their feedback and assistance, to Michaela Title for long and esoteric conversations about statistics, and to Jasmine Kwan for some heroic participant recruitment and assistance with data collection.

Lastly, I would like to thank my family and friends for their love and support along this journey. The encouragement of my parents has been a rock as I navigated returning to school as a mature student. I cannot convey here how thankful I am for my partner Connor for being there and supporting me through highs and lows equally.

Chapter 1: Introduction

Motion capture, the process of quantifying human body movement and its characteristics, is a critical component for disciplines, such as rehabilitation science, sport science, and ergonomics (1). The use of motion capture technology to quantify human movement is also an intrinsic component of biomechanics research (1). There are multiple motion capture methods: inertial measurement units, electromagnetic tracking, image processing, and active or passive optical motion tracking systems. Each can be used to quantify human body motion, but have strengths and limitations based on the system. As technology continues to evolve there are new motion capture methods available which can then open new avenues for research. For example, almost all major sports now benefit from kinematic data collected using motion capture to determine sport specific performance metrics (2).

In the 19th century the newly developed technology of high-speed photography provided the ability to distinguish characteristics of movements and lead to new discoveries in the fields of movement analysis (3). Using still images taken by a sequence of cameras, photographer Eadweard Muybridge helped pioneer manual digitization, where features of interest in the photograph could be identified over multiple sequential images and their locations used to identify change in joint position (**Figure 1**). Muybridge used these techniques to identify novel characteristics of movements and developed new understandings of human and animal gait kinematics. A contemporary of Muybridge, Etienne-Jules Marey, who had already published the seminal work on displaying data in graphical form (4), saw the research potential in sequential

photography. He developed the first high-speed camera that took images at regular intervals and thus could be of direct use in kinematic experiments of the day.

Techniques that can be considered under the umbrella of image processing continued to evolve and in the early 20th century; the process of rotoscoping was developed by animators (5). By tracing over frames of film they were able to represent human motion more accurately in animated films. While this was not of immediate applicability to biomechanics research, these developments led directly to early optical motion capture technologies used by the film industry in the 1960's and 70's (5).

Manual digitization progressed alongside video technology and the ability to capture images. As a motion capture modality manual digitization requires that the features of interest are present in sequential images and when a calibration is available with objects of known length the position of those features can be reconstructed into spatial coordinates (6). This method is still used in some modern data collection processes as it does not require the attachment of markers, making it a valuable tool in sport biomechanics where the analysis of movement during training and competition is important (6). However, manual digitization is time consuming; making it impractical when video acquisition uses high sampling frame rates because the number of video frames that would have to be analyzed could number in the thousands and would overwhelm the manual process (7). There are additional concerns of subjective user error with manual digitization because of the necessity of having a human identify the points of interest. This questions the method's accuracy and the need for interrater reliability assessments. These challenges have led to the development of the various motion capture systems with improved tracking capabilities.

Figure 1 *"Animal Locomotion" (Muybridge Plate 626)* Sequence with jockey on horseback by Eadweard Muybridge (1830 - 1904). Images depicting high-speed photography of a running horse and providing the opportunity to study gait.

Most modern motion capture applications are optoelectronic systems, which involve active (light-emitting), or passive (light-reflecting; Figure 2) markers placed on the participant that are tracked by a series of cameras. These motion capture systems are highly accurate (8) and reliable (9). These systems drastically reduce the human-hour investment when quantifying movement data compared to manual digitization; however, they are often limited to a laboratory setting because of the specialized equipment and the participant markers sometimes impede movement (10). Unfortunately for movement research, a laboratory setting may not have the same ecological validity as collecting the data in the field.

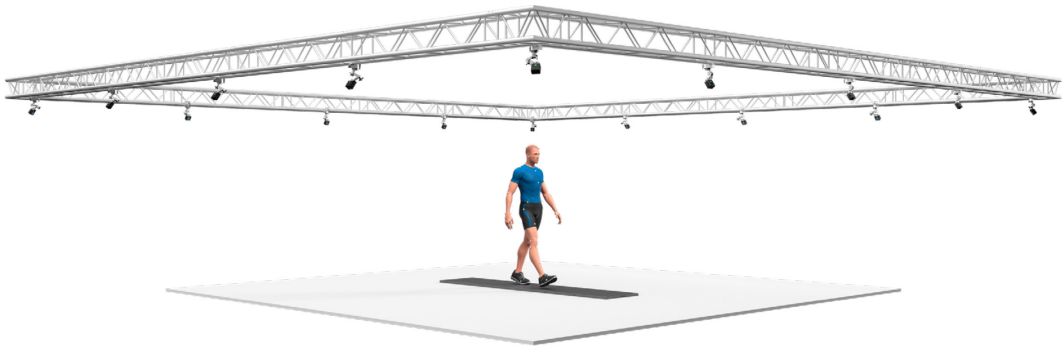

Figure 2 Example of a 16-camera prime-13 passive motion capture system (OptiTrack Inc. USA) with a series of cameras set up around the participant.

A possible solution to the motion capture ecological validity problem is to use inertial measurement units (IMUs), which are a wearable technology that consist of an accelerometer, a magnetometer, and a gyroscope. These devices are effective at capturing kinematic data in the field, but they also possess some limitations such as accuracy, being affected by drift error and magnetic fields (11,12). Therefore, there remains a need to have a data collection tool that can quantifying human movement outside of the laboratory environment (e.g., Sport arenas, factories, etc.) without the use of markers or the need for lengthy digitization processes.

Recent technologies such as AI-driven markerless motion capture have the potential to perform field-based data collection with laboratory precision (13). Markerless motion capture techniques represent the newest development in motion analysis and present significant potential benefits relative to previous methods. Movement can be analyzed during a wide range of activities and settings without participant preparation and laborious manual digitization processing (6). Such systems may address the previously required trade-off between accuracy and ecological validity. Recently, a few markerless motion capture systems such as the Kinect (14) have become publicly

available; however, most such systems have been developed for the entertainment industry (14–16) and use widely different methods for recording movement. While new research continues to explore the accuracy of entertainment-driven markerless systems (17) not all such systems are suitable for biomechanics research (5). The challenges in using these entertainment-based systems can be attributed to the different accuracy requirements between the entertainment industry and biomechanical research where small movements must be tracked accurately (6). In summary, markerless motion capture systems are highly variable and;

> "There is no general consensus regarding the minimum accuracy requirement of [markerless] motion analysis systems and the magnitudes of the inevitable measurement errors will vary depending on the context (laboratory vs. field), the characteristics of the movement and the participant, the experimental setup and how the human body is modelled." (6)

The world of markerless motion capture underwent a significant shift with the increased access to machine learning programs which can automate the process of accurately identifying features on an image or still frame from a video. The recent advances in the field of machine learning have provided a novel approach to the earlier motion capture technique of manual digitization and may offer the accuracy required in research settings (18–20). Machine learning is not a recent development; algorithms related to the field began appearing in the 1970's (18). The changes in computing power have allowed this resource to be applied to larger and more complex problems. For example, machine learning's ability to leverage feature identification in the analysis of images can be used for motion analysis. Supervised machine learning, a subset of

machine learning applications, uses algorithms that are trained by providing both the data set to be analyzed (such as a series of video frames) and the correct output (the location of each feature of interest in the frame). The algorithm modifies itself to best process the data and arrive at the correct output (21). Network training is an iterative process that improves feature detection accuracy with each repetition of self-modification. The capacity for large-scale rapid data analysis can then be applied to video files. A network trained to identify user-defined features can process still images at a rate that keeps pace with larger modern data volumes and, in theory, can even match the live production of the videos (allowing for real-time motion capture) (13).

DeepLabCut (DLC), an open-sourced supervised machine learning program, was produced as a tool for harnessing the existing power of feature identification within images and to automate the manual digitization process of videos, frame by frame. Although it was designed primarily for animal motion capture where the features being tracked may vary wildly between applications, its adaptability may make it useful for human motion capture in diverse settings. As with other supervised machine learning programs DLC's network 'learns' marker locations by being trained first on user-labelled data (13). Additional programs like Anipose have been suggested by the DLC authors as add-ons that can leverage the ability of DLC to identify locations in 2D space and triangulate those locations using multiple cameras to transition to a three-dimensional data collection (22). However, studies into various markerless applications have yet to examine the effect of camera positioning and complexity of the set of features being tracked on network performance (23,24). Additionally, the open-source, trainable options such as DLC have yet to be validated against gold standards for 3D data collection.

The purpose of this research is to explore the tracking capabilities of DLC and the various conditions that can affect its performance with the goal of evaluating the 3D outputs against traditional marker-based motion capture. Within that goal are three sub-objectives. First to examine the effects that tracking a complex set of features, such as those typically used in optical motion capture (25), has on the performance of DLC compared to a more commonly used (simplified) marker set (26). Second to examine the effects of varied camera angles, and the resulting feature occlusion, on tracking success/network performance. Third to compare the Anipose-derived 3D hip and knee joint position and angle data from DLC to traditional marker-based motion capture.

Chapter 2: Literature Review

2.1 Overview of motion capture

Motion capture is an integral part of biomechanics (27,28), both in research and in practical applications such as computer animation (29). From the earliest applications using Muybridge's photographic analysis (1) to Marey's invention of high-speed photography (30) and then to modern camera and marker-based systems (31), the ability to identify the positions of anatomical landmarks that can be used to estimate joint centers, or other features of interest is imperative to biomechanical human motion analysis. Kinematic analysis is commonly performed by measuring the spatiotemporal characteristics of rigid bodies (32). However, the systems and models used to gather and quantify 3D positional data can be quite different. Since Muybridge's early work, a variety of motion capture techniques have been developed that include electromagnetic systems, IMUs, and optical motion capture.

2.1.1 Electromagnetic:

Electromagnetic motion analysis systems (EMSs) measure the time taken for electromagnetic waves to travel from the device to base stations (33). EMSs can determine the position of an object(s) in large capture volumes without requiring that the sensor remain in view of a detecting device (ie. a camera) (33). These devices are used in scenarios where the capture volume area precludes other techniques from determining position, such as satellite position sensors when there is a roof (27). EMSs have greater accuracy limitations compared to optical motion capture systems (27) and can suffer from magnetic materials interference (34). Additionally, these systems tend to have low

sampling frequencies which limit their usefulness when analyzing higher frequency movements (27). Common EMS systems include the WASP which uses tags on the participant and fixed anchor nodes to determine 2D position (27). Accuracy of this system varies with studies showing errors of between 0.48m and 0.7m during an indoor basketball game (35). A more accurate form of EMS is the Radio Frequency Identification Device which uses an electromagnetic field to transfer the data from an active or passive tag to the base station. Researchers using a passive tag system found accuracies of 0.17m within a 5.4m^2 volume (36).

2.1.2 Inertial Measurement Units:

Inertial measurement unit systems (IMUs; **Figure 3**) are a popular method for collecting movement data that combine an accelerometer, gyroscope, and in some cases a magnetometer. They are capable of quantifying rapid motion and for that reason can be useful in sports science (37,38). The current best option for many field collection scenarios, multiple highly capable commercial products such as Notch and ActivPal are available that use this technology (39–41); however, their ability to determine precise joint-center locations and three-dimensional joint angles can be limited by the effect of the participant's acceleration on the device with both rapid acceleration causing errors and low accelerations causing increased drift error (27). The current gold standard of multi-IMU systems which are able to determine joint angles by comparing outputs of adjacent IMU devices is the Xsens. Validation studies have shown is is capable of producing movement waveforms that match optical motion capture outputs, however during high levels of acceleration the joint angle error can increase up to 15.9º (42).

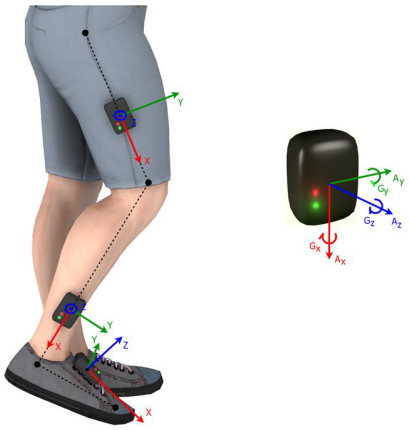

Figure 3 Placements of a 3 IMU set for lower limb kinematic data collection.

2.1.3 Optical Motion Capture:

Optical motion capture systems are the gold standard for motion capture, both in terms of reliability and accuracy (10). These systems identify specific body segments by tracking markers located on anatomical landmarks/features. These camera systems detect light, typically infrared, either produced by light-emitting active markers such as with the Optotrack system, or reflected off passive retroreflective markers from cameras like with Optitrack and Vicon (**Figure 4**) (27).

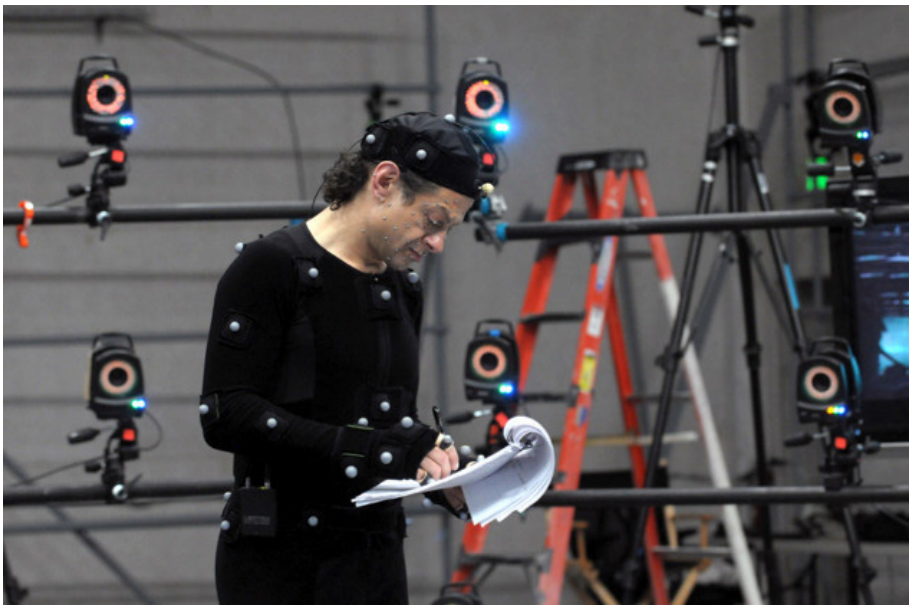

Figure 4 Retroreflective passive marker-based motion capture system used to track movement during a filming process.

Accuracy of these systems is dependent on the number of cameras and the size of the capture volume, with a high number of cameras and smaller volumes typically resulting in the greatest accuracy (43). Typically considered the gold-standard within the OMS group, Vicon, has been shown to possess millimetre (< 0.007m) accuracy when measuring fixed objects (44–46). While these systems are accurate, they bring with them numerous practical difficulties. The capture volume is determined by the number of cameras; therefore, if a large collection space is required, then there is an increased need for more cameras set up around the participant. Also, movement interference may occur from the presence of markers, especially after markers that are typically wired to a control unit. Constant line-of-sight between the marker and a minimum of three cameras is necessary to track 3D positions; however, marker occlusion happens frequently depending on the movement, participant clothing or equipment. Therefore, data

collection protocols typically must limit certain clothing, equipment, or movements which then decreases the ecological validity of the motion capture data.

2.1.3.1 Marker placement:

The choice of where to place optical motion capture markers is an important consideration that can influence the kind of data that can be collected as well as its accuracy. Sufficient markers on each body segment are required to calculate relative segment position (47). Marker placement standards, using various boney landmarks on each body segment, have been established to estimate joint centers and segment orientations (23,47). Marker clusters are multiple markers attached to a rigid surface which is then attached to a body segment. They provide a potentially more accurate method for calculating joint movement but still require calibration with markers placed on boney landmarks to identify the distance and orientation from joint centers. Landmark identification, although requiring training, can be a potential source of error because of its subjectivity and dependence on researcher's experience (48).

2.1.3.2 Camera Placement:

Camera placement is crucial for accurate motion capture as it directly affects the quality and reliability of the captured data. The positioning and configuration of cameras determine their field of view, depth perception, and ability to track markers or features on the subject (49). Placing cameras strategically around the collection space ensures maximum coverage of the capture area. Cameras should be positioned and angled to cover the entire capture volume, ensuring motions or actions are not missed during the capture process. Camera placement should consider the distance between the subject and

the cameras, as well as the inter-camera spacing, to enable accurate depth perception and triangulation. This allows the system to reconstruct the subject's movement accurately in 3D space (49).

2.1.3.3 Optical Motion Capture limitations:

Marker-based motion capture systems, while more reliable than many other forms of motion capture, have limitations. The sometimes-expensive costs, combined with complex camera and computer equipment required for data collections make them an unrealistic and cumbersome choice for many data collection environments outside of a laboratory (50). The reliance of these optical motion capture systems on markers presents additional challenges, such as bony landmark identification and marker placement errors from researchers (13,51). Improper marker locations and inconsistency in marker placements creates inter-operator and inter-session data variability, impacting the accuracy of the results (52,53). Attaching markers to non-rigid structures, like skin and clothing, results in tracking errors because of motion artifact between the marker and the item being tracked (54–56). These errors may compound potentially decreasing the accuracy of the optical motion capture system, which would be a problem if the tracked movements were small or complex.

2.1.4 Image Processing Systems:

Image processing systems (IPS) use still images, typically taken from video , to locate features and then track those features across subsequent video frames. The original Muybridge photography could be classified in this way, but most modern systems use some form of computer vision to automatically perform the segment orientations (27).

Early computer IPS used model-based tracking where the computer system had a predetermined 3D model of the tracked object, but this required extensive environmental information, making these systems challenging to use (57). Another form of IPSs are feature-based systems which focus on tracking elements of an image; for example, markers or other identifiable features, like printed targets (58).

Historically IPS are limited in their accuracy; therefore, they often require restricted conditions such as very specific camera angles during data collection and/or long post-processing times (14,59). An example of a more modern IPS software that is commonly used in biomechanical research is Kinovea (60–62). This software has been shown to produce accurate results (63); however, the accuracy is conditional to the camera view being orthogonal to the plane of the movement of interest (63), identifiable markers need to be applied to the participant, and the program has no capacity for 3D data collection. These limitations decrease the utility of Kinovea for researchers without the ability to create their own calibration/triangulation code.

Recent advances in the field of machine learning have made image processing more efficient, accurate and, accessible to researchers by limiting the computation demands (22). One such system, Theia3D Markerless, uses an adaptable deep-learning method at its core; however, it is limited to tracking a pre-determined set of joint coordinates (64). Theia3D has been shown to be an accurate alternative in situations where optical motion capture is not feasible (64–66); however, validations suggest caution in its use for clinical situations where accuracy thresholds may be smaller (67,68). Furthermore, its fixed set of tracked locations mean that any research question including non-standard features, or objects that need to also be tracked (e.g. PPE) cannot

use the program. Therefore, researchers tracking more than 3D human body movements require a more adaptable form of 3D markerless motion capture that is as accurate as marker-bases system.

2.1.5 Summary Table of Motion Capture Modalities

Mocap Modality	**Description**	**Strengths**	**Weaknesses**	**Accuracy**	**Field applicability**
Electromagnetic motion analysis (EMS)	-Finds position of transponders - Measures flight time of EM waves	- Line of sight not required - Large capture volume	- Sensitive to magnetic interference - Increased noise at greater distances - Limited to 2D	- Varies greatly between systems - [1]0.18m – [2]0.7m	- Useful for tracking participant position within large spaces - Cannot produce joint angle data
Inertial measurement unit (IMU)	- Accelerometer and gyroscope - Multiple IMUs can track body position	- Least intrusive of all wearable systems - Can detect rapid accelerations for a single unit - No limit on capture volume	- Subject to drift over time - Error can compound with multiple units[3]	- Multi-IMU compared with OMS[3] - Comparable mvmt waveforms - RMSD 5.2° - 15.9° - > error with > acceleration	- Useful for collecting kinematic data in field when precision accuracy is not required
Optical Motion Capture (OMS)	- Reflected or emitted light detected by cameras	- Able to track large number of points - Highest accuracy of common systems	- Limited to fixed locations - Sensitive to light contamination - Line of sight - Accuracy can be affected by marker movement	- High accuracy when tracking fixed objects - [4]0.0067m – [5]0.0003m - Dependent on camera position	- Complex setup for outdoor data collection - Costly infrastructure - Sensitive to light interference - Challenging system type for field conditions
Image Processing System (IPS)	- Computer vision used to track humans or markers	- More accurate than EMS - Larger possible collection volume than OMS	- Limited advantages over other systems. - Insufficient accuracy for precise joint angles	- Significant variability across different systems - [6]0.04m – [7]0.22m	- Useful in large volume outdoor spaces -When high accuracy is not required
AI-based Image Processing System (AI-IPS)	- Computer neural networks trained to track sets of features	- Adaptable to experimental needs -	- requires calibrated cameras for 3D – limits capture volume - Only track pre-set features	- Limited research comparing raw coordinate locations - errors < 8° for joint angles[8]	- High accuracy - Can track large numbers of people quickly - Less accurate than marker-based systems

Table 1: Summary of motion capture modalities. [1]Shirehjini & Shirmohammadi, 2012 (36), [2]Sathyan et al., 2012 (35), [3]Nijmeijer et al., 2023 (42), [4]Spörri, Schiefermüller, & Müller, 2016 (45), [5]Maletsky, Sun, & Morton, 2007 (43), [6]Corazza et al., 2010 (69), [7]Liu et al., 2009 (70), [8]Wren, Isakov, & Rethlefsen, 2023 (67)

2.2 DeepLabCut:

DeepLabCut, is an adaptable machine learning application designed for markerless motion capture. First published in Nature, it uses pose estimation and feature identification to processes video files and track features with minimal training data (13). DLC is an open-source project that uses a trained network that learns which features to track by being trained on user-labelled data.

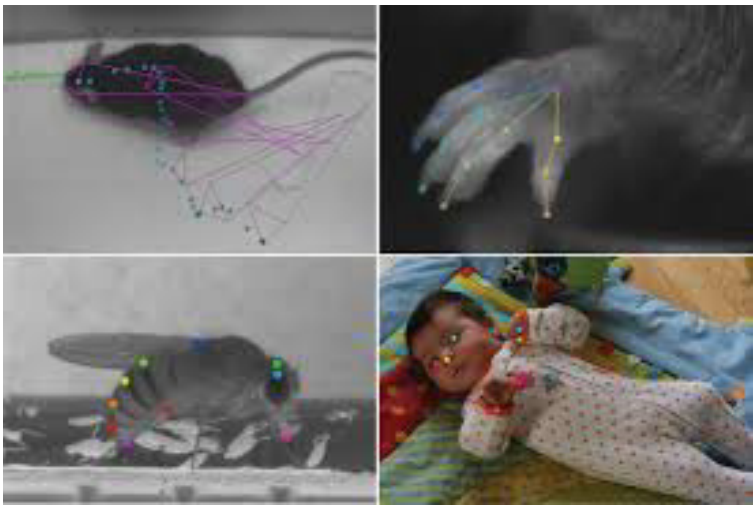

Figure 5 Examples of DeepLabCut applications in research (13)

The program arose out of a need to track movements on animals where markers would not be possible. DLC uses a robust deep-learning framework to decrease the number of trained frames required for a network to accurately identify points. By decreasing the required workload for training, it became possible for labs to train their own network; resulting in DLC being used across a wide range of applications (Mathis, 2020). DLC has the potential to bring the versatility and unencumbered data acquisition of image processing up to the accuracy and utility of optical motion capture.

For DLC to be usable for research, its validity and accuracy must be quantified

(71). DLC has been validated in multiple studies (26,71–75) but they share a common limitation of only being validated in two dimensions (72), which is of limited use for most biomechanical kinematic research applications. Also, DLC's accuracy has been commonly compared to that of human labellers rather than against a gold standard like optical motion capture (26,74). In the studies that have compared DLC directly to marker-based optical systems, researchers have used training images with markers present (72). This raises the strong probability that the network will have learned to rely on the presence of those markers, and so the results can only be extended to other situations where the person tracked also has markers applied (72). This creates a research need to move toward validating DLC in a 3D human environment.

2.2.1 DeepLabCut commonly tracked features

DLC was written with the intention of tracking simplified sets of features. In contrast with optical motion capture systems where upwards of 11 markers could be used to track specific features around a single limb (76), DLC typically aims to track/estimate joint centers. Therefore, when it was validated against marker-based systems, the DLC network was trained to track different points (joint centers and segment ends) than the system it is compared to (72). Joint angle outcomes were compared rather than raw coordinate locations. Unlike marker-based systems where the marker can be place on any location that a researcher wishes to track, features tracked by DLC tend to be visually distinct from surrounding tissue, such as visible bony landmarks, or points of connection between two segments. 3D position tracking requires multiple points on each segment (76) and so there is need to explore DLC's performance with tracking complex sets of features.

2.2.2 DeepLabCut Camera position

DLC was designed to solve challenges in animal research and as such involved relatively simple camera setups (13); for example, sagittal views while recording gait, and overhead views tracking movement through a confined space. Therefore, research validating DLC output has been constrained to these simplified viewing angles, with both studies validating its application on humans using sagittal views only (26,72). The sagittal view has the advantage of maintaining all features of interest within view for the duration of data collection, except for very occasional marker occlusion during a gait cycle (26). Further research is required to determine if the sagittal view validations extend to the oblique views found in three-dimensional motion capture.

2.2.3 Anipose

Programs like DLC that track in two dimensions are useful for specific cases, such as within animal research where behaviours are being tracked; however, complex whole-body movements, in both animals and humans, often require three dimensions to be fully assessed (23,77). Therefore, DLC has included an inbuilt 2-camera triangulation package that permits a 3D data acquisition, but this portion of DLC has not been studied to determine the accuracy of movements that include changes of depth (towards or away from the field of view), Instead, the DLC authors recommend the use of Anipose (7).

Anipose is a separate (from DLC), open-sourced, software toolkit designed to leverage DLC's ability to track movements from 2D video and collect multiple video angles to produce 3D joint angles and segment position outputs. Anipose should resolve some of the challenges faced by DLC (77) because the additional cameras may improve movement capture and occlusion issues (23,77). Anipose takes each camera's DLC

output and combines them together to then extract 3D kinematic data. This process is performed in three steps: calibration, refinement of 2D keypoints, and then determination of 3D position.

Calibration:

Calibration is accomplished by recording video of a ChArUco board (**Figure 6**), a version of a checkerboard with unique combinations of pixels placed in the white squares (78). The board is moved through the motion capture space ensuring that at least two cameras can always see the board. The unique images inserted into the white checkerboard space reduce the potential of occlusion error that occur in some camera frames because the calibration program infers locations for the occluded parts of the board. The calibration videos are then provided to Anipose where pre-trained neural networks are used for edge detection of the ChArUco board (79). When using this method more than 90% of frames had an error of less than 20 μm and 1° in angle. This is an improvement even over human manual detection of edges in an image.

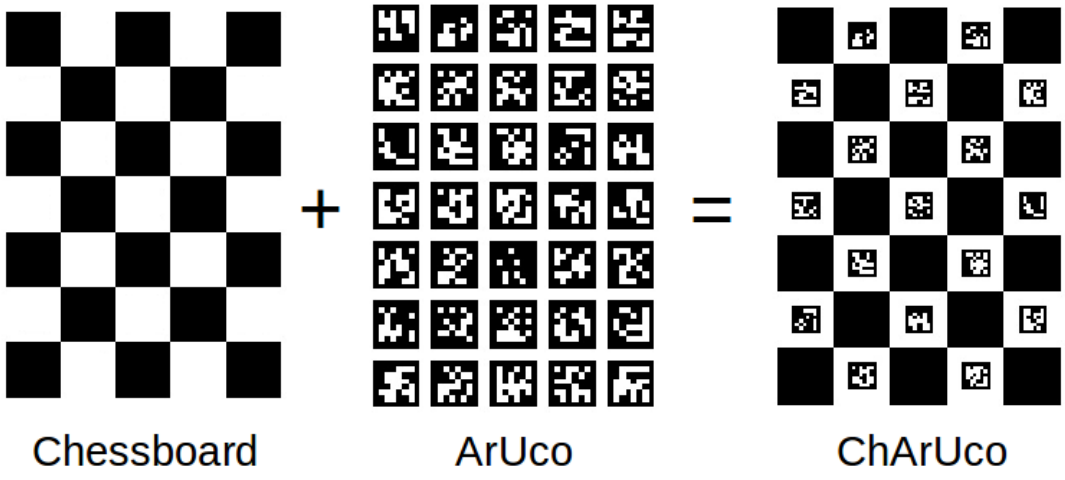

Figure 6 ChArUco is the combination of the chessboard-style calibration image and the ArUco image which is more robust to occlusions.

Refinement:

Anipose uses the feature identification network that a researcher has trained on DLC and then applies a series of three filters to improve the tracking results provided (77). A median filter removes outlier data that deviates outside a user defined window and then uses an interpolation algorithm to replace the removed data. A Viterbi filter refines joint position by comparing movement from frame to frame and removes frames where the position jumps for small incorrect periods. Lastly, Anipose uses an autoencoder filter to modify the confidence score of the identified joint, decreasing it based on the scores of the other joints. As a final step in the refining process, frames with poor tracking, typically when the feature is not visible, are removed altogether so that other camera views during that frame can be relied upon (77).

3D position:

Videos from each calibrated camera are loaded by Anipose and then run through the DCL network to identify 2D positions from that camera's viewing angle. 3D positions are calculated with the triangulation matrix established during calibration, filtered using the various layers of Anipose filtering, and then refined with the limb and joint prediction model. The estimated positions are smoothed to ensure consistent limb length (77).

Anipose calculates 3D joint flexion angles using the inverse cosine of the dot product of the two adjacent segments such as the thigh and shank. To estimate abduction and rotation angles, Anipose assumes that abduction only happens at the most proximal joint (eg hip), and that rotation cannot happen at the most distal (eg ankle). The Anipose

authors solve the remaining joint angles by first estimating the rotation of each joint based on the proximal segments coordinate system (77).

2.3 Prior research validating markerless motion capture

Markerless motion capture may eliminate the need to manually identify bony landmarks on participants and utilizes a less complex equipment set-up (13). However, until recently, markerless motion capture systems had limitations in both accuracy and optical range, which limited their adoption. For example, markerless motion capture systems were sensitive to changes in lighting conditions, required a simplified background free of objects and, systems that tracked the outlines of humans were sensitive to even small changes of shape caused by clothing (7,13,22,80,81). Newer and more robust markerless motion capture systems are becoming as accurate as optical motion capture and may replace other forms of motion capture (26,32,65,66,82), but the popular markerless systems all use pre-trained networks that limit the researcher to tracking only the features the program was designed/trained to track. Should a researcher wish to track additional points, a different method of motion capture would have to be used.

In an early example of the work done to validate markerless motion capture methods, Ceseracciu, Sawacha, & Cobelli (2011) developed a method to automate the manual digitization of swimming kinematics. This scenario represented an ideal application of markerless motion capture as the water refraction would distort any optical motion capture attempt. When tracking the wrist, a maximum root mean square distance (RMSD) of 56mm was reported compared to manual digitization. While this error is

significantly larger than those between comparable optical motion capture (31) it demonstrated that markerless systems could be used to track human movement.

The methods used by Ceseracciu, Sawacha, & Cobelli involved a silhouette markerless motion capture system. It did, however, require the additional step of background subtraction which would not be required by later systems. The results of their case study found an average (RMSD) of 11.75 deg in knee flexion-extension which accounted for 18% of the range of motion in their study. This was better than the results for the ankle and hip which showed an error comprising 45% range of motion and 33% of their range of motion, respectively. These results were not accurate enough to support the use of this markerless motion capture approach in research.

KinaTrax (KinaTrax Inc., Boca Raton, FL), a markerless motion capture system more closely resembling DLC (i.e., use of deep learning), was validated against both force plate data (gait spatiotemporal parameters) and a marker-based system (50). The marker system used eight infrared cameras (SMART-DX; BTS Bioengineering, Milan, Italy), while the markerless motion capture system used 8 video cameras. Twenty-two participants completed three walking trials and 9 parameters were compared between the systems. In order to compare all three systems only the gait and timing results were compared and so the 3D position and joint angle data was not directly compared between the markerless and marker-based systems. Between-system absolute agreements and relative consistencies were assessed using ICCs (Intraclass correlations) and Bland-Altman plots. The authors reported overall excellent agreement between methods, with stride length having the strongest agreement of $R = 0.710$ but reported poor agreement with swing time. The authors suggest that this poor result was due to the simplicity of the

foot model in the KineTrax system and indicated that additional tracking of the heel should be implemented in future work.

This further highlights the weakness that many currently available markerless systems have; pre-trained networks cannot track new objects or body landmarks that were not included in the network development/training. This makes the current systems unusable for any scenario where atypical features or items beyond typical human features, are to be tracked (eg. protective equipment). It also emphasized the potential challenges with extracting 3D joint angle data from a small number of tracked features, commonplace in markerless systems.

Theia3D is a commercial deep learning markerless motion capture system based on an architecture similar to DLC (65). Whereas DLC trains a new network for each unique laboratory setup, Theia3D uses a pre-trained network to track over 51 human features. This trade-off means that Theia is useful for human data but cannot track beyond the pre-determined feature set. Kanko et al. (32) validated Theia3D by collecting gait data on two systems simultaneously, marker-based seven camera Qualisys 3+ (Qualisys AB, Gothenburg, Sweden) system, and eight additional Qualisys video cameras to record video. Thirty participants conducted ten consecutive trials of four second treadmill walking. Knee and hip joint angles and global segment angles were produced for each system. The average RMS difference range in hip and knee angles for flexion and abduction was 1 – 2.2° for both joints, in both planes of motion. The hip internal-external rotation differences were greater, 8.5 – 13°. Joint centers were also calculated, and the two systems differed by less than 3cm in absolute position. The authors noted that

occlusion by the treadmill bar and blurry video caused by the speed of distal limb movement had a negative effect on the markerless motion capture accuracy.

While recent validation work into markerless motion capture has displayed promise there remain pertinent areas that have not been examined. For example, as markerless systems become capable of 3D data collection and increasingly used for research, it will be important to understand the effect of various setup conditions have on the systems accuracy. Currently there is a lack of evidence understanding the effects of increasingly complex features of interest (eg. higher number of features and/or boney landmarks), and the effect of camera angle have on the accuracy of the system.

2.3.1 Feature sets.

Previous applications of DLC on human movement have focused on simple sets of features tracked, typically lower-limb joint centers from a sagittal view (26,72). These features are sufficient for 2D collection and tracking movement along a single plane, but to track and compute 3D motion/orientations like a marker-based system, the position of additional features (boney landmarks) must be known (47). To the best of the authors knowledge, no studies have examined the effect of increased complexity of marker sets (features of interest) on the accuracy of a DLC trained network. A notable difference between markerless and marker-based tracking targets is that the markerless systems will attempt to track the same joints center point regardless of viewing angle, whereas marker-based systems will track fixed points placed on the skins surface. Applying commonly used marker set locations to DLC training will facilitate direct comparisons between the systems but may reduce DLC network performance.

2.3.2 Camera angle

The viewing angle of a camera changes the visibility of features on a participant and can decrease occlusion. In traditional marker-based 3D motion capture researchers need to track features that may be regularly out of view from a single camera, so multiple cameras with different views are required. Currently, there is little evidence identifying the effects of alternative/multiple camera angles on the accuracy of a DLC trained network. Following traditional marker-based tracking theory, multiple and varied camera views will be required for accurate 3D markerless motion tracking (77). However, it remains unknown how the system will learn to identify a feature if it is rarely visible to specific cameras. Therefore, the effect of various camera angles on network performance should be explored.

2.4 Research Problem

Markerless motion capture is a captivating solution to the ecological validity concerns of traditional marker-based 3D motion capture systems (83). However, typical experimental setups (complex feature or object tracking) make most commercial markerless programs unsuitable. This highlights the usefulness of open-sourced markerless tracking programs like, DLC, because they provide a possible solution to tracking more complex features.

Unfortunately, previous research on DLC has been limited to tracking features that remain visible to the camera either constantly or with only occasional occlusion. Studies have yet to attempt to use the more robust set of features/markers commonly seen in optical motion capture. Of interest are the capabilities of DLC to track the more complex sets of features consistent with current marker-based motion capture protocols

which have known results when calculating joint angles (47). While authors have reported success at producing 3-dimensional joint kinematics on humans from tracking only joint centers (77) rather than the more complex sets, these results were not validated against a gold standard.

Past DLC research has also only used only single cameras collecting data from the angle most appropriate for the movement being tracked, i.e., sagittal view for gait studies, overhead view for behaviour tracking. Motion capture of complex features often requires cameras placed at multiple viewing angles but the effect that sub-optimal camera views with high occlusion percentage have on DLC's performance has not been quantified. Lastly, most DLC validation studies have been in two dimensions, but most applications of motion capture require 3D positions to determine joint angles. Therefore, DLC recommends a multi-camera 3D software package Anipose (77), which has yet to be validated in the literature.

AI markerless motion capture faces common challenges, some shared with other modalities, others unique to AI. The robustness of the network is dependent on its ability to learn to track the features of interest. The visual distinctiveness of the feature matters, so features like the mid-point along a person's thigh would be difficult to track, while a feature like the lateral knee, which is visually distinct from the surrounding tissue is less difficult. The amount of 'noise' in the background can affect network robustness, with simple backgrounds facilitating feature identification and complex backgrounds with other untracked people present is much more difficult. The number of features tracked is another factor that can influence the performance of a network. Marker-based systems track large numbers of points to quantify joint position and angle, but trainable AI

systems like DLC typically only track joint centers or other large identifiable features. A final common problem is feature occlusion, when the feature being identified is not in view. DLC and other similar programs can handle small numbers of occlusions, but when a large number of features is being tracked, an individual camera might never see a particular feature. It is unclear how DLC will respond in such a scenario, but it would likely result in poor performance.

Of particular importance to the successful implementation of DLC as a 3D motion capture tool are clear understanding of its practical implementation. Specifically what effect a feature set resembling ones commonly used in motion capture (CFS) will have on DLC's ability to track as compared to the more common SFS. Additionally, while camera position and its effect on accuracy are known for optical motion capture, this has not been addressed in any literature for DLC.

2.5 Purpose

The purpose of this research is to explore the tracking capabilities of DeepLabCut and some of the various conditions that can affect its performance with the goal of evaluating the 3D outputs against traditional marker-based motion capture. Three sub-objectives were set to achieve this research goal.

Sub-objective 1:

Evaluate the effects of a complex 18 feature set, then comparing it to a simplified 10 feature set in line with common DLC targets. Performance differences between the networks will demonstrate the performance effect of increased numbers of tracked features on DLC.

Sub-objective 2:

Examine the network performance effects of varied camera angles, and the resulting feature occlusion, on tracking success/network performance.

Sub-objective 3:

Evaluate the Anipose-derived 3D outputs from DLC processed video using both the complex and simplified feature sets trained networks. Then compare the 3D hip and knee joint positions and angle data from each network to traditional marker-based motion capture and identify if 3D joint angle data accuracy is compromised when only using a 3 or 5-camera verses an 8-camera setup.

Chapter 3: Methodology

3.1 Participants

50 participants were recruited for the study in accordance with a sample size calculation (84) identifying the need for eight testing participants. This calculation was based on previous ICC results comparing DeepLabCut (DLC) to optical motion capture (72). The sample calculation was conducted by custom R code using the calculateIccSampleSize() function within the 'ICC.Sample.Size' package based on the work of Zou et al. (84). Inputs included a predicted ICC output of 0.8, which was based on conservative predictions from previous DLC validation work (72) and the number of comparisons for each participant (6). Default values were accepted for desired alpha (0.05) and power (0.8) Previous research into DLC using human participants suggest training the network with four participants for each novel testing participant (26,72). To account for attrition and this 4:1 ratio of training to testing data, 50 participants in total were recruited for this project, with 40 participant's data used for network training and 10 for validation testing purposes. Participant homogeneity can lead to a lack of robustness in the trained network (22,85); therefore, deliberate effort was made to ensure the diversity of the participant group. Subjective visual assessment was used to identify perceived gender presentation and to estimate the inclusion of persons of visible minorities.

Participants were excluded if they expressed any difficulty performing the required movements or if they disclosed any musculoskeletal conditions that would preclude any of the movements. Participants provided written informed consent to participate in the study (Appendix A). The experimental design and protocols were

approved by the Dalhousie Health Sciences Research Ethics Board (REB file#: 2022-5974; Appendix B).

3.2 Hardware

3.2.1 Markerless Motion Capture

Video recordings of all participants conducting movements for DLC processing, either to train networks or to test them, were gathered using seven GoPro Hero8 cameras and one GoPro Hero9 camera recording at 1080p-60fps. All videos of the training group participants were recorded in separate files for each participant in each of the six movement conditions. Testing group participant video was recorded continuously for all movement conditions to facilitate synchronization between the GoPro camera videos and the marker-based Optitrack output.

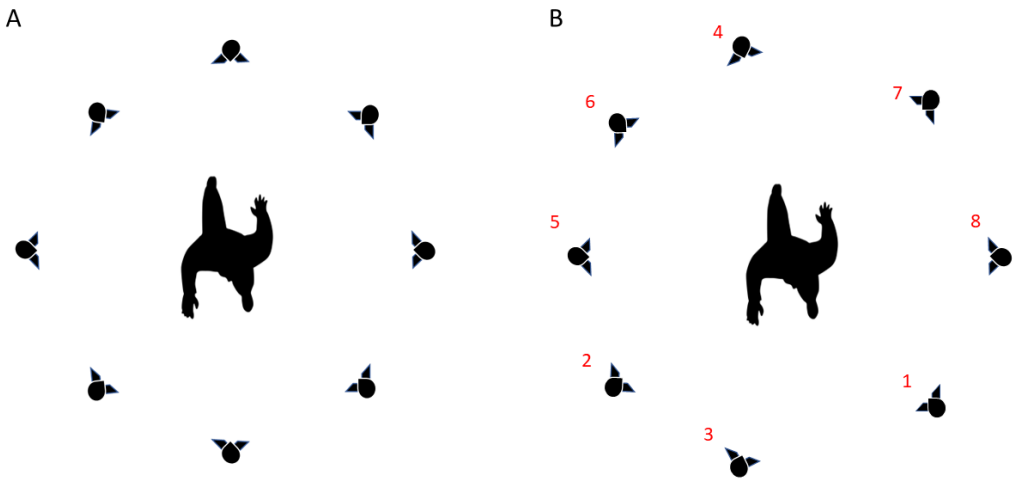

Figure 7 Camera position during data collection. (A) Training group collection position shown with cameras placed equidistant around the participant. (B) Testing group collection positions shown with cameras placed such that five cameras have a view of one side of the participant and the remaining three are equidistant in the remaining space. Numbers indicate the identification number assigned to each camera for comparison of performance.

During all training group data collections, the cameras were placed equidistant around the participant to get a diverse collection of movement images (Figure 7A). During all testing group data collections, the cameras were placed such that a subset of five cameras were located in a semicircular shape forming a sagittal total view angle of approximately 150° (Figure 7B). This alteration in camera angles, for the testing group, was to ensure that the right side of the participant features remained in the field of view of 5 cameras. By grouping the 5 sagittal viewing cameras the intent was to provide the opportunity to calibrate a 3-camera, 5-camera, and 8-camera calibration for anipose so that number of cameras could be explored as a factor in anipose accuracy. This would also improve the Anipose calibration ensuring that at minimum 5 cameras could see the

ChArUco board simultaneously (Figure 6). The final three cameras were placed equidistant in the remaining volume. GoPro camera control was achieved via Bluetooth using the "Camera tools for Heros" app. Prior to each data collection session, camera calibration video was recorded using a 0.91m by 0.61m ChArUco (78) board with 8x12 squares and a 4x4 ArUco dictionary (86).

3.2.2 Optical Motion Capture

A 14-camera OptiTrack passive motion capture (Prime 13 OptiTrack Cameras; Natural Point, Oregon, USA) configuration and associated reflective markers were used to obtain kinematic data of the 10 testing participants. Markers were placed on the anterior superior iliac spine (ASIS), posterior superior iliac spine (PSIS), greater trochanter, medial & lateral epicondyles, medial and lateral malleoli, calcaneus, and the base of the 3rd metatarsal for both legs. Locations were selected to satisfy ISB guidelines regarding lower-limb data collection with the minimal number of markers (47). Data were collected in a single continuous collection for each participant at 60Hz using Motive 2.1.1 (Optitrack, Natural Point, Oregon, USA) to match the DLC camera settings.

3.3 Software

3.2.1 Markerless Motion Capture

GoPro videos for markerless motion capture were processed to a compatible codex (H.264) using ffmpeg (87), then edited and synchronized using DaVinci Resolve 18. DeepLabCut 2.1 was then used to train the markerless motion capture networks with the 40 training group participants and to analyze the performance of networks on the training group participants. These steps were completed for both the simplified (joint center) and complex feature networks (complex feature set). All videos for the testing

group participants were selected for the analyze_videos function within the DLC GUI. Options to produce .csv output files along with tracked videos were selected.

3.2.2 Optical Motion Capture

Optical motion capture data collected for the 10 testing group participants was accomplished using Motive motion analysis software. Post-processing of the lower limb marker data for each testing participant was completed using this software, then exported to Excel (Microsoft. Washington, USA) for analysis.

3.4 Experimental protocol

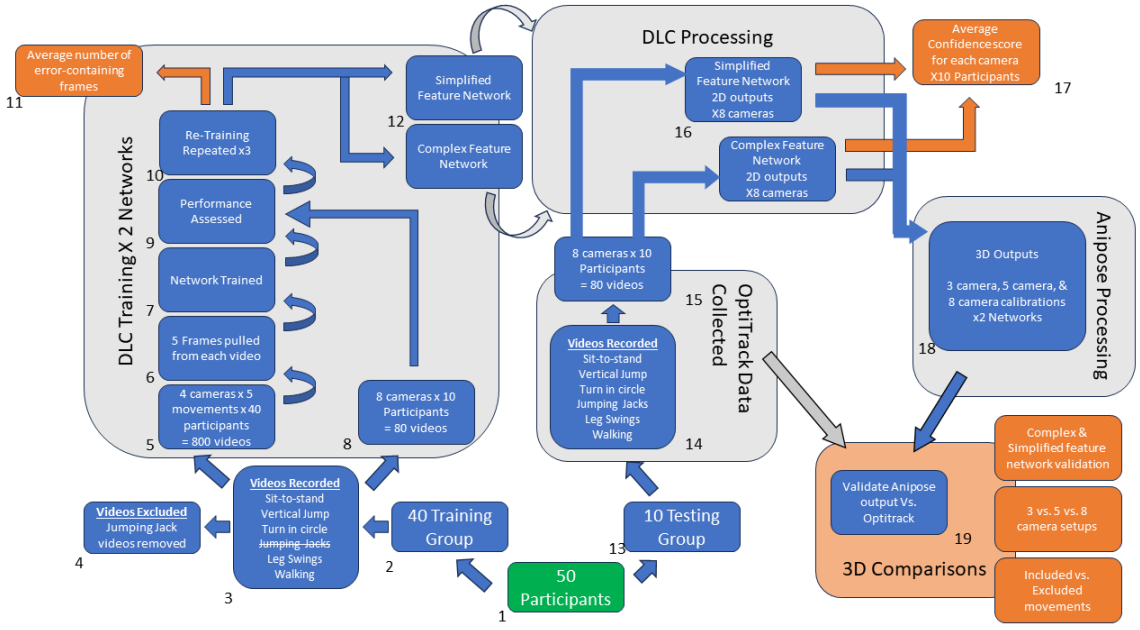

Figure 8 Project workflow. Green indicates start of workflow. Orange indicates data collected for sub-objectives. [1]50 total participants. [2]40 participants assigned to the training group. [3]6 videos recorded by 8 cameras for each participant. [4]Jumping jack videos removed. [5]Videos from 4 cameras provided to DCL for initial training of two networks. [6]5 frames from each video selected by DLC. [7]Networks trained. [8]10 participants from training group selected and [9]combined videos used to assess network performance. [10]Videos used to retrain network. [11]Average number of error-containing frames recorded for sub-objective 1. [12]Two final networks created. [13]10 participants assigned to training group. [14]Training group has single video recorded by 8 cameras of 6 movements while simultaneously recorded by OptiTrack. [15]Videos of testing group are [16]provided to the two networks for 2D analysis. [17]2D outputs including confidence scores are used for sub-objective 2. [18]2D outputs are used by Anipose to calculate 3D position for each feature. [19]Anipose and OptiTrack outputs are compared for sub-objective 3.

Upon arrival at Dalhousie University's Biodynamics Ergonomics and Neuroscience Laboratory (BENlab) (2661 South St. Halifax, NS. Canada), participants were given time read and ask questions about the experimental protocol then provide consent in accordance with the Dalhousie Health Sciences Research Ethics Board.

Anthropometric data was then collected (Height and segment lengths for foot, shank, and thigh) using a stadiometer and anthropometers.

Training group:

The training group performed six (6) movements in a pseudo-randomised block design order at the instruction of the researcher: vertical jump, walking through the capture area, rotating in a circle, leg swings in a sagittal plane, sit-to-stand, & jumping jacks. For a randomly determined 6 participants, passive motion caption markers were placed on the participant to decrease the likelihood of error when the network encounters the testing videos with markers. In these special participant cases, the markers were placed bilaterally on the 3rd distal metatarsal, on the calcaneus, on the medial and lateral malleoli, on the medial and lateral epicondyles, on the greater trochanter, on the ASIS, and on the PSIS (47).

Testing Group:

Participants were asked in advanced to arrive at the lab in snug-fitting clothing to facilitate marker placement. Prior to recording any data reflective markers were placed on the participant (3rd distal metatarsal, calcaneus, medial and lateral malleoli, medial and lateral epicondyles, greater trochanter, ASIS, and PSIS). With reflective markers on, each testing group participant performed the same 6 movements in a randomized order. Participants from this group were recorded simultaneously by the GoPro cameras for markerless motion capture and the marker-based OptiTrack system. Synchronization of the GoPro video and the Optitrack system was accomplished by dropping a ball prior to

any participant movements. The ball drop created a common fixed time-point for later cropping and synchronization.

3.5 Data Analysis

3.5.1 DeepLabCut

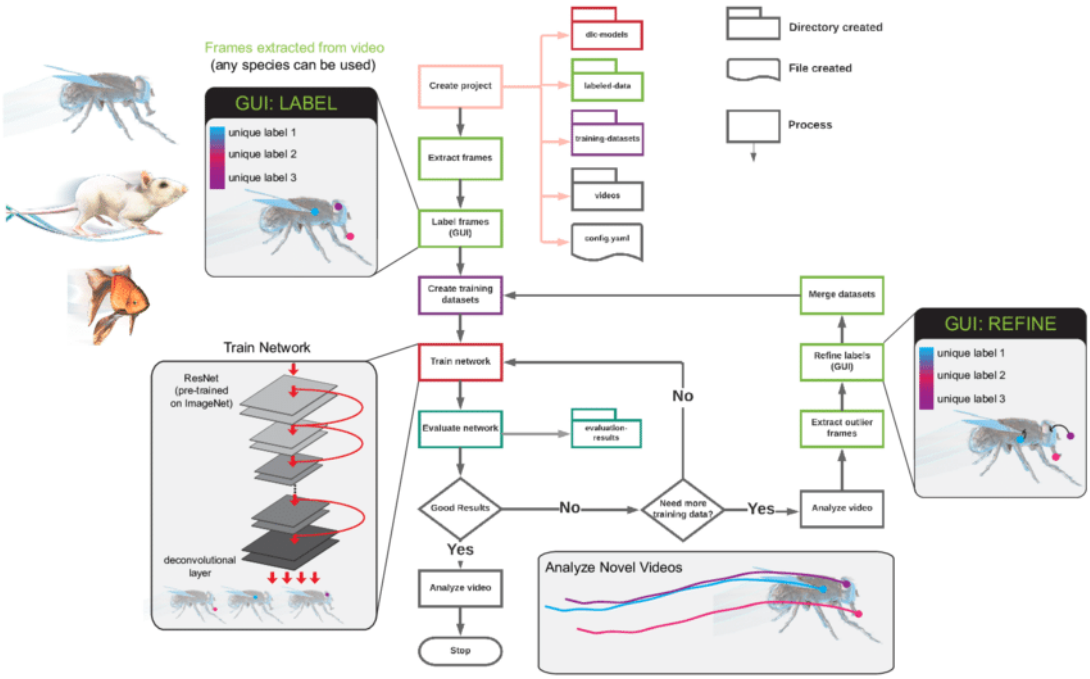

Figure 9. Chart of DeepLabCut workflow from project creation to the analysis of novel videos (7).

Training Group:

The current study differed, in several aspects, from previous human DLC validation research (72). For example, the network was not primarily trained on images which included passive optical motion capture markers, which eliminated the likelihood that the network learned to identify the reflective markers, rather than the desired body feature/landmark. Of interest is the DLC network's ability to track movements that have not been explicitly included in the training data. To test this, one of the six movements,

jumping jacks, was randomly selected and removed from the training dataset. This would allow later testing of DLC's joint center tracking performance during a movement that was absent from training data. A total of 800 videos remained available for training (40 participants x 5 included movements x 4 cardinally opposed cameras). DLC's inbuilt algorithm (7,13) identified 5 frames that were representative of the range of positions present in each video, and trained labelers identified the 18 features of interest (complex feature set) corresponding with the OptiTrack marker positions. Then, a second (simplified) network consistent with previous DLC tracking (7,26,72,77) was trained tracking joint centers of the ankle, knee, and hip, along with the ASISs and PSISs. The two networks will allow for assessment of DLC's ability to track a set of features consistent with current marker-based motion capture protocols (sub-objective 1) as well as assessment of the simplified feature set's ability to produce accurate 3D joint angles (sub-objective 3).

Training Group – Initial network training

To train the networks, separate projects within DLC were created for each network. Config.yaml files were edited to include a list of the features to be tracked for that network along with the number of frames to be selected from each training video (five). Plotting configuration was set to connect tracked features in the eventual output. While this is a cosmetic step for DLC, later processing by Anipose will treat these connections as fixed lengths. The DLC graphical user interface (GUI) was used exclusively during this project. Videos of the training group were loaded within the 'Manage Project' tab. When extracting frames, default DLC settings were chosen such that an automatic extraction method used a kmeans algorithm to identify frames with

varied content. This algorithm was chosen for its stated ability to detect novel movements which when used in training lead to a more robust network (7). When creating the network, the default ResNet-50 was used as the base network. While Mathis et al. found that networks with deeper layers had a small reduction in error (RMSE of 3.09 ± 0.04 vs. 2.90 ± 0.09 with ResNet-101) it was decided that the default setting were most appropriate for an initial validation study.

The networks were each trained on a computer with an Intel i9-9900K processor running at 3.6GHz using 32 GB of RAM with a NVIDA RTX 2080Ti, running Windows version 10.0.19045 Pro. Training was capped at a maximum of 1030000 iterations (the default for DLC). A total of 4,000 frames were used, 5 from each video, 5 videos from each camera, 4 cameras for each participant for the 40 training group participants.

Training Group - Retraining

Network performance, for both the simplified and complex feature set, was evaluated as per DLC workflow (Figure 8) by subjective examination of the tracking results. 10 participants were selected from the training group, two of which had markers present, and their videos spliced together such that each camera angle had a separate video that included all five training movements. This was done as the initial training process had shown difficulty with large numbers of videos. The ‘Analyze videos’ tab on the DLC GUI was used to select those 80 videos and track the features across the entirety of the video. When participant video showed multiple tracking errors, the researcher initiated the retraining process.

During the retraining process the ‘Extract outlier frames’ tab of the GUI was used and the 80 videos showing tracking errors were provided to the program. Multiple algorithms exist to detect outlier frames; however, the default option of ‘jump’ was used which looks for frames where the feature detected moves more than a set distance from its location in a previous frame. These ‘jumps’ of feature location were identified in earlier piloting as a common error and so the default was chosen. Given that large numbers of errors were found, it was a satisfactory choice. During the extract outlier frames process DLC identified the number of outlier frames (misidentified features) in these reanalyzed videos and reported to the user. This metric was then used as a proxy for network performance; however, its initial intent was as a feedback tool to guide the user through improving the network.

In the next step of the retraining process, DLC selected a representative sample of those outlier frames using a kmeans algorithm for the user to correct. 18 frames were chosen from 4 camera views of each of the subset of 10 participants to ensure a robust sample. The ‘Refine labels’ tab then opens a graphical interface where 720 video frames were presented to the labelers who corrected any misidentified features. This new training data was added to the initial network training data and a new round of training using the same settings was initiated. This retraining process was repeated three times for each network leading to a final tally of 6,160 frames used per network. The number of outlier frames was obtained and recorded from each retraining iteration for both the simplified and complex feature set (used for sub-objective one analysis). Outlier frame counts were then compiled in excel and analyzed using custom R code.

Testing Group:

The videos of the 10 participants of the testing group who were recorded both by video and Optitrack were processed by both the simplified and complex feature networks. Two-dimensional feature coordinates from each camera view were exported as a .csv file by DLC for analysis along with a confidence score for each feature in each frame (used for sub-objective two analysis). As a machine learning algorithm develops during the training process and, in the case of DLC, learns to identify the features of interest, it compares its results against the user-annotated images. With each training iteration, of which there are hundreds of thousands, the algorithm improves and can identify how confident it is with each prediction(88). It outputs this confidence level as a confidence score.

Confidence scores are a common output in machine learning, typically expressed as a number between 0 and 1 which indicates how confident the network is in the output it provided with 0 indicating low or no confidence and 1 indicating high confidence. While the exact methodology used by DLC to calculate confidence are not published, the industry approach to their application is understood (89) and they are often used as threshold values when interpreting machine learning outputs (88). Individual confidence scores can be problematic as a network can be confidently incorrect in identifying a feature (false positive) or not confident in a correct identification (false negative) (89); therefore, as an aggregate of all features across all frames the values are potentially more robust.

Both DLC and Anipose use confidence values to identify outlier frames and so the author was comfortable in their application within this study. The raw files were processed by custom MATLAB code which extracted the confidence scores from the .csv file and exported to R for statistical analysis.

3.5.2 Optical Motion Capture

The 14 OptiTrack passive motion capture Prime-13 cameras were set up around the capture space and controlled by the Motive 2.1.1 Motion analysis software (Natural Point, Oregon, USA). This software was used to trigger and collect all of the testing group participants, then used to post-process each participant's movement trials to ensure all 18 markers were identified throughout the movements. Cleaned Optitrack data was then exported to custom MATLAB code for joint angle and segments length calculations. This code used a cubic spline method to fill any missing data, then filtered it using a 2^{nd} order 6Hz 2-pass Butterworth filter(1). 3D joint angles were calculated, and then average lengths calculated between segment endpoints.

3.6 Statistics

Sub-objective 1, network performance improvement and comparison between complex (CFS) and simple (SFS) feature sets, utilized the DLC "outlier frames found" data to identify network quality. Data for each feature set network and their respective retraining iterations were tested for normality with Shapiro-Wilk's tests. The number of error frames was normally distributed at each retraining iteration ($p > 0.05$). QQ plots were constructed and as all points fall along the reference line, we can assume normality (Figure 9). A two-way repeated measures ANOVA was calculated to determine the effects of network type on frames containing outliers/errors over the three iterations of

retraining. Effect of network type at each iteration was determined using pairwise comparisons. Effect of retraining iterations, for both networks, were tested with one-way ANOVAs.

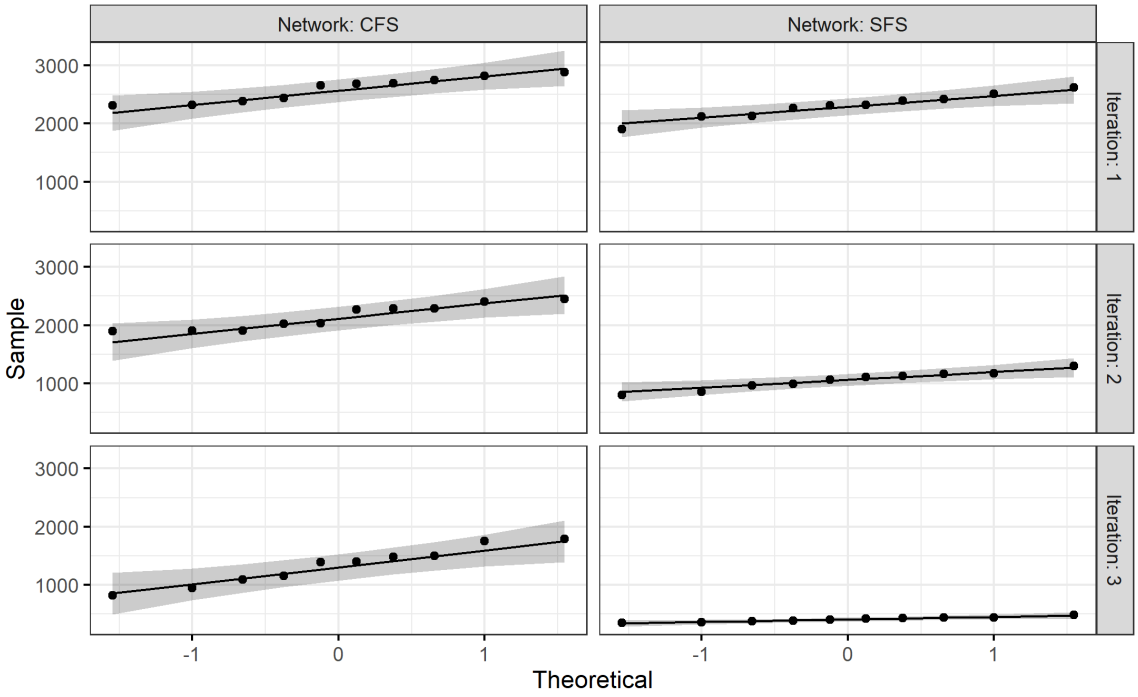

Figure 10 Grid comparing Quantile-quantile plots showing the empirically observed quantiles of error-containing frames (y-axis) as a function of quantiles expected from a normal distribution with the same mean and variance as the empirical distribution (x-axis).

Sub-objective 2, camera angle performance (how well a particular camera view was able to track the features of interest), was determined by obtaining the average camera feature tracking confidence scores (obtained from DLC) for each testing participant. Data across all cameras was tested for normality with a QQ plot of an ANOVA model's residuals and with Shapiro-Wilk's test (Figure 10). As not all points fall along the reference line normality was not assumed, which was also supported by a

significant ($p < 0.05$) Shapiro-Wilk test. Therefore, to compare camera angle performance, a Friedman test was chosen as a non-parametric alternative to a one-way repeated measures ANOVA test. In the case of a significant Friedman test, a post-hoc pairwise Wilcoxon signed-rank test using Bonferroni correction will be completed to identify which camera positions are different. Effect size was calculated using Kendall's W (90).

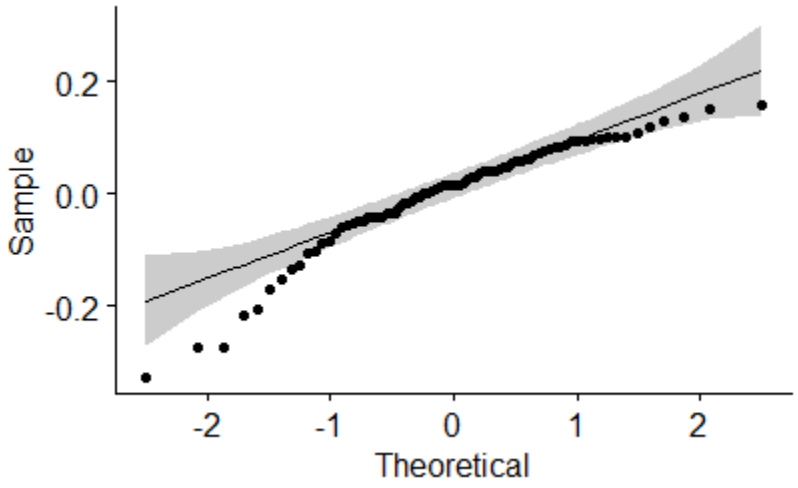

Figure 11 Quantile-quantile plot showing the empirically observed quantiles of mean network confidence residuals (y-axis) as a function of quantiles expected from a normal distribution with the same mean and variance as the empirical distribution (x-axis). Data following a normal distribution would be expected to fall along the prediction line.

Sub-objective 3, had the data been available, would have been analyzed using interclass correlation coefficients to quantify agreement between raw coordinate position. Statistical parametric mapping would have been used to show differences in joint angle calculated between Anipose and Optitrack during movement cycles.

Chapter 4: Results

4.1 Demographics

The training population was representative of the testing population with similar proportions in physical presentation of gender traits and inclusion of persons of visible minorities (*Table 1*).

Table 2: Participant demographics for the training and testing groups.

Group	Mass	Stature	Gender Presentation	
Training n = 40	73.8 ± 17.40 Kg	1.72 ± 0.09 m	26 ♀ presenting 14 ♂ presenting	20% persons of visible minority
Testing n = 10	75.4 ± 19.42 Kg	1.73 ± 0.08 m	4 ♀ presenting 6 ♂ presenting	20% persons of visible minority

4.2 Sub Objective 1: Comparison of networks

There is a statistically significant two-way interaction between network type and retraining iteration, $F(2,36) = 57.6$, $p < 0.05$. Pairwise comparisons show that the mean number of error-containing frames (outliers) was significantly different between CFS and SFS networks, with SFS outperforming the CFS at training iteration 1 ($p = 0.005$), retraining iteration 2 ($p < 0.001$), and retaining iteration 3 ($p < 0.001$). One-way ANOVA results showed a significant effect of iteration on error-containing frames in the CFS

network ($F(2,18) = 234$, p-adj < 0.001) and in the SFS network ($F(2,18) = 689$, p-adj < 0.001).

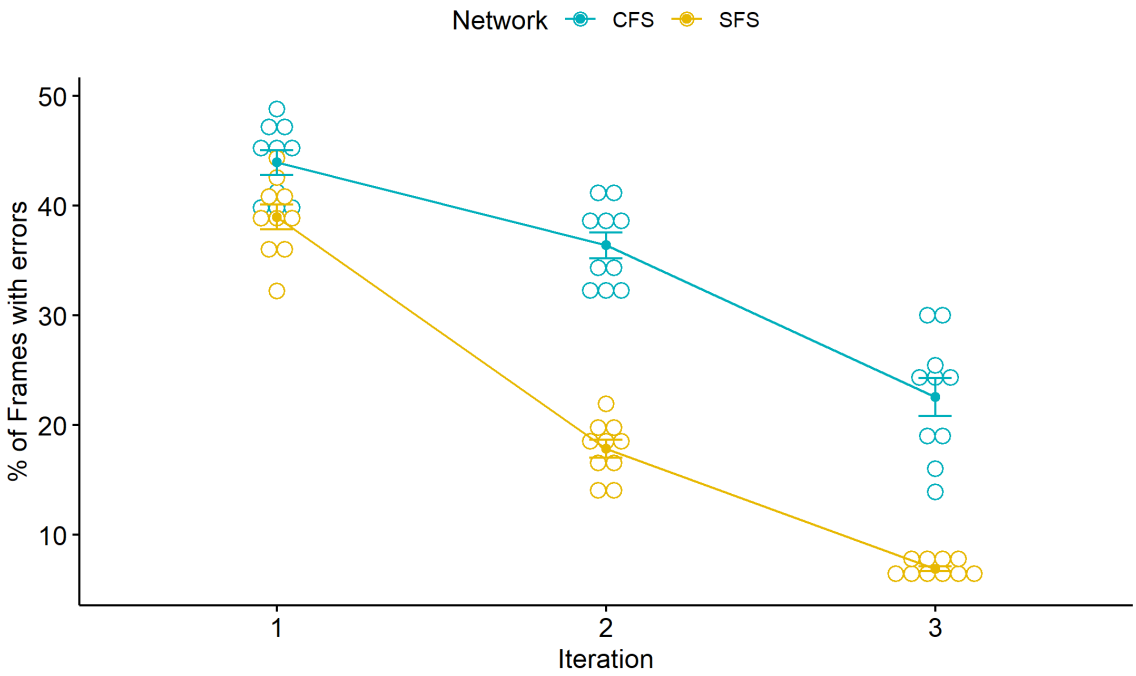

Figure 12. Error-containing frames (y-axis) over successive retraining iterations (x-axis). Complex feature set (CFS) network displayed in teal, simplified joint center network in yellow.

4.3 Sub-objective 2: Comparison of cameras

Feature identification confidence was significantly different between the different camera angles using Friedman test, $\chi^2(7) = 54.1$, $p < 0.05$, $n = 10$.

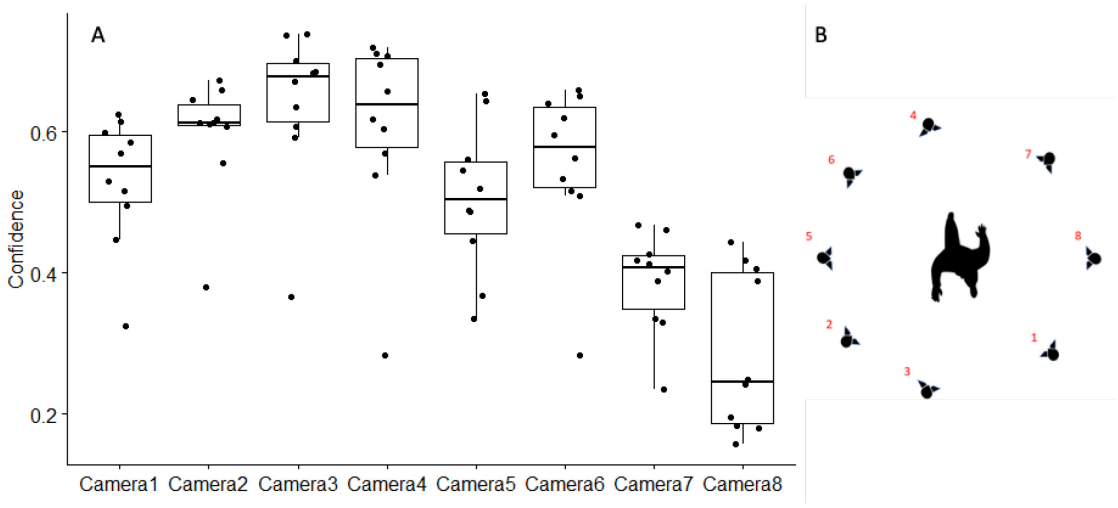

Figure 13 Image A shows Confidence scores (y-axis) between each camera position (x-axis) showing effect of viewing angle. Points indicate mean confidence scores for each participant. Image B indicates the raw camera ID numbers associated with identification in Freidmans's test.

Pairwise Wilcoxon signed rank test using a Bonferroni correction between cameras revealed no statistically significant differences in confidence score (Table 2). Effect size was calculated as W = 0.77 which indicates a large effect.

Table 3. Results of pairwise Wilcoxon signed rank tests between camera positions. The statistic column represents the Wilcoxon signed rank test statistic. Each test had 10 degrees of freedom.

Group 1	Group 2	Statistic	p	p-adjusted
Camera1	Camera2	1	0.021	0.602
Camera1	Camera3	0	0.002	0.055
Camera1	Camera4	1	0.021	0.602
Camera1	Camera5	5	1.000	1.000
Camera1	Camera6	5	1.000	1.000
Camera1	Camera7	9	0.021	0.602
Camera1	Camera8	10	0.002	0.055
Camera2	Camera3	2	0.109	1.000
Camera2	Camera4	4	0.754	1.000
Camera2	Camera5	9	0.021	0.602
Camera2	Camera6	7	0.344	1.000
Camera2	Camera7	10	0.002	0.055
Camera2	Camera8	10	0.002	0.055
Camera3	Camera4	8	0.109	1.000
Camera3	Camera5	9	0.021	0.602
Camera3	Camera6	9	0.021	0.602
Camera3	Camera7	10	0.002	0.055
Camera3	Camera8	10	0.002	0.055
Camera4	Camera5	8	0.109	1.000
Camera4	Camera6	8	0.109	1.000
Camera4	Camera7	10	0.002	0.055
Camera4	Camera8	10	0.002	0.055
Camera5	Camera6	4	0.754	1.000
Camera5	Camera7	10	0.002	0.055
Camera5	Camera8	10	0.002	0.055
Camera6	Camera7	10	0.002	0.055
Camera6	Camera8	10	0.002	0.055
Camera7	Camera8	10	0.002	0.055

4.4 Sub-Objective 3: 3D joint angles and segment length comparison

Technical difficulties with the Anipose software meant that no three-dimensional data was generated from either DeepLabCut feature set (simplified or complex). While initial pilot testing in 2020 produced positive results using the sample data provided by Anipose, recent attempts to use the program encountered difficulty. Unfortunately, while the program did create the desired sub-folders to store processed data, no data was created in the folder, and no error code was provided by the program. When the Anipose sample data was used again, a similar result was found. Re-installation of the program, the creation of new environments to run the program, and attempts to return the PC to the state it was in when the program ran in 2020, all produced the same results. Installation on other computers as well as a full format of an off-site computer all failed to change the Anipose outcomes. Without any identifiable error to troubleshoot and without successfully running the program on any location it was determined that Anipose could not be made to work in time for this project. As a result, no 3D joint angles or segment length comparisons to the optical motion capture outcomes were possible.

Chapter 5: Discussion

Results showed that the simplified feature set (SFS) network had superior performance at each stage of the retraining process, and that both networks improved significantly as retraining occurred. Camera location was shown to have a large effect on the mean confidence levels of the identified points. Cameras with anterior/posterior views outperformed cameras with sagittal views.

5.1 Comparing marker sets

The number of error frames with misidentified features, although not its intended purpose, was a useful metric for measuring and comparing network performance. DeepLabCut (DLC) has a consistent methodology (e.g., DLC identifies features with low identification confidence and/or unexpected location jumps from frame to frame) for identifying these outlier frames. Both network types improved after each retraining iteration, as demonstrated by the decreased number of misidentified frames (Figure 11). Specifically, each retraining for the SFS network decreased the misidentified frames by 50%. However, in both the SFS and the CFS training networks there were features not visible from some camera angles for the entire data collection.

Prolonged feature occlusions or complete absences of features from the entire video were much more common in the CFS which was trained to mimic the full set of markers used in optical motion capture. For example, within the CFS network, the sagittal viewing cameras on the right side of the participant may briefly glimpse the medial femoral condyle of the right leg during the rotating in a circle movement; however, for all other movements it is obscured entirely. These occlusion problems persisted for this sagittal camera, not able to view the left side ASIS, PSIS, greater

trochanter, lateral femoral condyle, and lateral malleolus as well as the right medial malleolus. That means that of the 18 features, only 7 are in view of the sagittal camera and 4 are intermittently obstructed (left calcaneus, left 3rd metatarsal, left medial malleolus, and left medial epicondyle). Contrasting with the SFS network, the same view would have 5 locations always in view (right ankle, knee, hip, ASIS, and PSIS), 2 intermittently obstructed (left ankle, and knee) and only 3 completely occluded (left hip, ASIS, and PSIS).

Typically for DLC projects, the features tracked are estimated joint centers, rather than boney landmarks. The joint center can be identified from multiple angles which limit the chances for occlusion. The effect of the increased number of marker occlusions in CFS compared to SFS is apparent in the network performance improvements seen in Figure 12 as well as in the results of the pairwise t-tests which showed significant differences at each measured point. After the first round of retraining the number of frames with low confidence decreased by 50% in the SFS network and only by 20% in the CFS. The second round of retraining decreased the error further for both networks as shown by the one-way ANOVA tests, but the SFS network still significantly outperformed the CFS. These performance scores may have been further improved with an increased number of retraining iterations.

The DLC retraining process is necessary to improve networks, and its importance is crucial when being trained on images/videos with significant rates of feature occlusion. However, network training guidelines, in situations of high feature occlusions, are not well documented or researched within the current literature. It would be expected that continuing to retrain the networks would improve the networks until a performance

plateau was achieved, but the number of training iterations that would be necessary for a complex feature set to achieve high tracking performance is unclear and should be standardized. For example, for very simple feature sets (i.e., tracking only lateral hip, knee, ankle) a high-quality tracking network can be achieved when trained on 17 participants, for 400 frames, with no retraining iterations (26). While these networks may be easy to train, the accuracy of the three-dimensional joint angle output that can be derived from a simplified feature set must be established to understand its role in field biomechanics research.

5.1.1 Selecting features of interest

The choice of features is one of the most impactful elements when training a network for markerless motion capture. The number of training images required will be directly dependent on the ease with which the network can learn to identify the features, and the number of required training images will affect the feasibility of DLC as a motion capture tool. Features with strong visual contrast to the surrounding area, or in areas where movement occurs, like joint centers, will be much easier to track than areas with no discernible distinction to their surrounding (7). Features that are the center of a particular body part rather than a superficial aspect of that body part will be easier to track, as we saw with the greater performance from the SFS tracking joint centers.

DLC is designed to handle occlusions when features are momentarily blocked from view, for instance by another limb passing over the area. However, it was unclear how DLC would perform if the feature was totally absent from the analyzed video. This project identified that CFS and sagittal cameras angles often caused contralateral features to remain occluded for the entirety of the movement, resulting in poor network

performance scores and camera tracking confidence scores, respectively. Although the CFS was intended to match that of common optical motion capture marker sets (76), DLC may not be at a stage to track the many common boney features needed for accurate segment position and orientation tracking. The relative success of the SFS within this study is encouraging but their remains a need to validate whether these simplified feature sets can produce accurate 3D joint angle data.

The ease with which some features can be tracked compared to others was apparent during the retraining process for both the SFS and CFS networks. In this project, the greatest difficulty for feature tracking was found for the medial malleoli and medial femoral condyles. They were both specific and small superficial features that were often occluded. The greater trochanter, despite being a feature without much visual distinction from surrounding tissue, was well tracked, potentially because it represented a mid-point for the pelvis and because it was a point where movement appeared to occur. Consideration should be given to the factors that effect how easily a feature can be tracked in future work, as the effect on network robustness is significant.

5.2 Effect of camera angle on performance

Camera angle confidence score results showed that there was a statistically significant difference in network tracking performance between the different camera angles. Cameras with oblique and sagittal views (Cameras 5 & 8) performed poorly compared to those with anterior/posterior views. This is most likely due to feature occlusion as the networks struggled to identify points that were not visible. As a result, the network would incorrectly locate them, causing decreased confidence scores. For example, in both networks the cameras with purely sagittal views (5 and 8; Figure 12)

were consistently unable to see the contralateral greater trochanter, ASIS, and PSIS locations; resulting in mislabelled contralateral leg feature locations (Figure 13). This resulted in large performance differences, specifically between camera 8 and all other cameras, that were statistically significant until corrected for multiple comparisons. The anterior/posterior cameras (3 and 4: Figure 12) had unobstructed views of ankle, knee, hip locations, and would only have the opposite pelvis markers (ASIS or PSIS) hidden from the cameras.

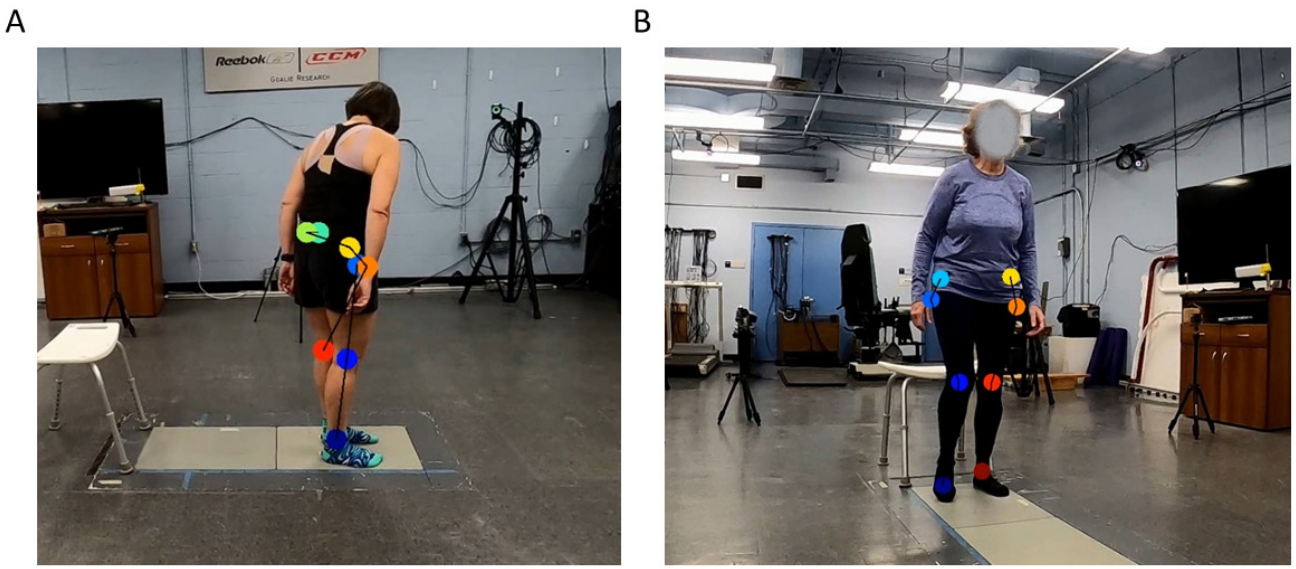

Figure 14 Image A showing the SFS network misidentifying left hip marker (orange) and left ASIS (yellow) when those features were not visible in camera 5. are not visible from this angle (camera 5). Image B shows correctly identified features as viewed from camera 3.

The common practice, in optical motion capture, of surrounding the subject with cameras so that each marker is visible by at least two cameras (23) does not translate to AI image processing. Optical marker-based systems process all cameras simultaneously whereas in image processing each camera is treated, and processed, separately (7) then the triangulation process is performed later (7,77). Since, DLC trains the network with

single camera views, it becomes imperative that camera location and camera views are strategically considered for the object or participant being recorded. Recently reported (reported after data collection was complete for this study) camera location research identified that other AI systems recommend two to six front facing cameras (23). Some of these other AI systems, such as openCap, also rely on human models to predict the location of occluded features, which improves the network's quality (23).

A predictive model is not available with DLC and may partially explain why cameras 7 and 8 performed worse than their camera counterparts (cameras 5 and 6). The potential reasons for these conflicting camera results are: 1) the sit-to-stand movement required a stool which may have obstructed camera 7's view of specific features, 2) cameras were set at varying heights, with camera 8 set lower to the ground than camera 5, and 3) camera 8 was the GoPro hero 9 rather than a GoPro hero 8 (all remaining cameras), which while set to the same collection parameters as the other cameras, may have had intrinsic qualities that differed from the other cameras.

It is important to note that while we found that a frontal plane viewing camera performed better than the sagittal viewing cameras, to collect robust 3D data of movements that occur mostly in the sagittal plane a single frontal viewing camera would not perform well. Multiple cameras with oblique views to one another are needed for 3D data collection, at least three for Anipose (77). Networks based on fixed human models have however been shown to collect accurate data with two, if the angle between them is 70° (23).

5.3 Challenges of Anipose

Anipose is an add-on application for extending DLC's data collection into multi-camera 3D space. It takes a DLC trained network and uses it to identify features recoded by 3+ calibrated cameras. It then outputs 3D joint angles between tracked segments as well as feature position. However, despite following the guidelines for installation (77), Anipose did not function properly which resulted in no three-dimensional joint angle and segment length data from the markerless motion capture networks. Despite multiple attempts to get Anipose to function (described in Section 4.4), Anipose was unable to provide any 3D output. An alternative to Anipose could have potentially been achieved by writing custom triangulation code to obtain 3D joint positions from 3, 5, and 8 camera setups. Once camera parameters are found then the 2D output from DLC could be combined via custom MATLAB code. Subsequent joint angle calculation for the CFS would follow typical code for Optitrack outputs but for the SFS this would not be possible. A novel approach using assumptions like the procedure used by Anipose and described in section 2.2.3: 3D position would be required to calculate joint angles.

Packages exist that can perform camera calibration as well as the subsequent triangulation; however, Anipose was designed specifically for DLC output. The Anipose calibration functionality uses a neural network trained to identify the edges and vertices of a calibration board, and specifically a type of board that assists in managing occluded views (77). The algorithms used surpass the accuracy of any other commonly employed method. Its refinement filtering algorithms were designed specifically to manage the types of output and challenges found with DLC and its 3D joint calculations are based on

the kinds of networks commonly tracked by DLC. Despite the inability to apply the program to this work, it remains the best option for future 3D DLC applications.

5.4 Limitations of DLC

DLC is designed to use a minimal number of training images so that it can be useful to researchers conducting smaller scale projects; however, this benefit also presents a limitation regarding the representativeness of the training videos compared to the videos being analyzed. Typical networks require thousands of training images to accomplish a human level of feature identification (72) accuracy (<5 pixels). While DLC reduces the number of training images to more manageable amounts, (300-500) (13), it introduces a potential for increased error if the analyzed images are too unlike the images used to train the model. Specifically, the program's authors note that the random component of the method used to select training frames from provided videos increases the possibility both that a novel movement might not be included and that a feature being tracked is occluded during the selected frame (13). While increasing the number of images used in the training process could reduce the prevalence of these errors, this decreases the very strength that DLC was designed to have. The authors recommended using diverse environmental, demographics and movement training videos (7). Error rates would be expected to decrease if the selected training videos had more diverse parameters than the videos being analyzed.

Another potential source of error was the small number of frames per video used in the initial round of network training (5 frames). This number was selected to keep the total number of training frames within a practical size while still using diverse camera views of all participants. This small number combined with the chosen kmeans selection

algorithm may have resulted in a less diverse training set. Uniform selection of frames, or larger numbers of initial frames (where there is little movement captured by the cameras) should be explored to identify changes in performance.

A further potential limitation, that may be more pronounced in DLC compared to other reinforced learning AI applications is the random error associated with inaccurate feature identification within the training data. With the large training sets of conventional networks, individual errors occur in a randomly distributed manor around the mean feature location. With an application such as DLC where the number of training images can be small, a skew in the error distribution around feature identification can result in larger misidentifications/deviations from the intended feature.

Many markerless motion capture programs like, openCap and Theia3D, that are designed to track human movement use built-in human models to improve the performance of their network. DLC does not use a built-in model in this way, which may partially explain why it had difficulty with partially occluded features. Unfortunately, this reduces DLC's robustness compared to other markerless applications. However, DLC's authors intentionally did not incorporate a model because the program is intended to be applied to a large range of things (eg. Humans, animals, equipment, machines; (13)). Therefore, DLC excels in its versatility but may require further training iterations to successfully track the wide variety of objects.

5.5 Future Directions

The results of sub-objectives 1 and 2 indicate that future work with DLC may want to avoid complex feature sets that mirror that of typical optical motion capture

systems, rather focus on features that can be tracked from multiple viewing angles. Additionally, front facing cameras should be used so that features of interest are visible from the frontal plane and throughout the movement. It would be recommended that 5 cameras are used in a 150° arc in front of the participant or object to be tracked. The between camera positions post-hoc analysis with Bonferroni correction may have been overly conservative; therefore, future work should consider grouping cameras with equivalent views to facilitate a smaller number of tests, ie sagittal and frontal views.

Network performance improved with retraining, but an inflection point, where the gains at each step of retraining decrease, was not reached. Future work would benefit from examining at what point diminishing returns occur during the retraining process. Since Anipose was unable to be successfully implemented for this work, the next step will be to reach out to the program authors so that 3D movement data can be produced from DLC and then compared to marker-based systems.

Without the 3D data from Anipose the researcher was unable to compare the results against the Optitrack data. Therefore, the next plan for this work is to directly contact the creators of Anipose to identify the rate limiting step that could not be solved prior to the completion of this work.

5.6 Conclusion

This research evaluated the tracking capabilities of DLC and how feature set complexity and camera locations may affect markerless motion capture network performance. The simplified feature set (SFS) had significantly better performance (fewer outlier frames) than the complex feature set (CFS), and the anterior/posterior facing cameras had significantly better feature identification confidence scores compared

to traditional sagittal plane camera views. The results from both sub-objectives 1 and 2 highlight the importance of choosing an experimental setup that will support strong network performance. Networks may drastically improve their performance if the features of interest are infrequently occluded from view, as well as forward facing camera positions that maximize the visibility of features of interest. These findings will inform future markerless motion capture research design by minimizing the time required for data collection and network training.

References

1. Winter DA. Biomechanics and motor control of human movement. 4th ed.. Hoboken, N.J.: Hoboken, N.J. : Wiley; 2009.

2. McGinnis PM. Biomechanics of sport and exercise. Fourth edi. Champaign, IL: Champaign, IL : Human Kinetics; 2020.

3. Bernshteĭn NA. The Co-ordination and Regulation of Movements [Internet]. Pergamon Press; 1967. Available from: https://books.google.ca/books?id=kX5OAQAAIAAJ

4. Smith LD, Best LA, Cylke VA, Stubbs DA. Psychology Without p Values: Data Analysis at the Turn of the 19th Century. Am Psychol. 2000;55(2):260–3.

5. Conditt J. 100 years of motion-capture technology. Engadget. 2018;

6. Colyer SL, Evans M, Cosker DP, Salo AIT. A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. Sports Med Open. 2018;4(1):24.

7. Nath T, Mathis A, Chen AC, Patel A, Bethge M, Mathis MW. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nat Protoc [Internet]. 2019;14(7):2152–76. Available from: http://dx.doi.org/10.1038/s41596-019-0176-0

8. Nagymáté G, Tuchband T, Kiss RM. A novel validation and calibration method for motion capture systems based on micro-triangulation. J Biomech. 2018;74:16–22.

9. Ehara Y, Fujimoto H, Miyazaki S, Mochimaru M, Tanaka S, Yamamoto S. Comparison of the performance of 3D camera systems II. Gait Posture [Internet]. 1997;5(3):251–5. Available from: http://www.sciencedirect.com/science/article/pii/S0966636296010934

10. Richards JG. The measurement of human motion: A comparison of commercially available systems. Hum Mov Sci [Internet]. 1999;18(5):589–602. Available from: http://www.sciencedirect.com/science/article/pii/S0167945799000238

11. Aminian K, Najafi B. Capturing human motion using body-fixed sensors: Outdoor measurement and clinical applications. Comput Animat Virtual Worlds. 2004 May;15(2):79–94.

12. Sekar LPN, Santos A, Beltramello O. IMU Drift reduction for augmented reality applications. Vol. 9254, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015. p. 188–96.

13. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci. 2018;21(9):1281–9.

14. Bonnechère B, Jansen B, Salvia P, Bouzahouene H, Omelina L, Moiseev F, et al. Validity and reliability of the Kinect within functional assessment activities: Comparison with standard stereophotogrammetry. Gait Posture. 2014;39(1):593–8.

15. Kotsifaki A, Whiteley R, Hansen C. Dual Kinect v2 system can capture lower limb kinematics reasonably well in a clinical setting: Concurrent validity of a dual camera markerless motion capture system in professional football players. BMJ Open Sport Exerc Med. 2018;4(1):1–9.

16. Schmitz A, Ye M, Boggess G, Shapiro R, Yang R, Noehren B. The measurement of in vivo joint angles during a squat using a single camera markerless motion capture system as compared to a marker based system. Gait Posture [Internet]. 2015;41(2):694–8. Available from: http://dx.doi.org/10.1016/j.gaitpost.2015.01.028

17. Kurillo G, Hemingway E, Cheng ML, Cheng L. Evaluating the Accuracy of the Azure Kinect and Kinect v2. Sensors. 2022;22(7).

18. Louridas P, Ebert C. Machine Learning. IEEE Softw. 2016;33(5):110–5.

19. Mathis A, Mamidanna P, Abe T, Cury KM, Murthy VN, Mathis MW, et al. Markerless tracking of user-defined features with deep learning. 2018;1–14. Available from: http://arxiv.org/abs/1804.03142

20. Wu JJS, Hung A, Lin YC, Chiao CC. Visual Attack on the Moving Prey by Cuttlefish. Front Physiol. 2020;11.

21. Nabi J. Machine Learning —Fundamentals - Towards Data Science. Medium [Internet]. 2019 May 24; Available from: https://towardsdatascience.com/machine-learning-basics-part-1-a36d38c7916

22. Mathis A. 3DeepLabCut [Internet]. 2020. Available from: https://github.com/DeepLabCut/DeepLabCut/blob/master/docs/Overviewof3D.md

23. Gow R. Evaluation of a Smartphone-based Motion Capture System for Athletic Movement Screening. UC San Diego Electronic Theses and Dissertations. UNIVERSITY OF CALIFORNIA SAN DIEGO; 2023.

24. Uhlrich SD, Falisse A, Kidziński Ł, Muccini J, Ko M, Chaudhari AS, et al. OpenCap: 3D human movement dynamics from smartphone videos. bioRxiv [Internet]. 2022;(650):2022.07.07.499061. Available from: https://www.biorxiv.org/content/10.1101/2022.07.07.499061v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.07.07.499061v1.abstract

25. Collins TD, Ghoussayni SN, Ewins DJ, Kent JA. A six degrees-of-freedom marker set for gait analysis: Repeatability and comparison with a modified Helen Hayes set. Gait Posture. 2009;30(2):173–80.

26. Cronin NJ, Rantalainen T, Ahtiainen JP, Hynynen E, Waller B. Markerless 2D kinematic analysis of underwater running: A deep learning approach. J Biomech [Internet]. 2019;87:75–82. Available from: http://www.sciencedirect.com/science/article/pii/S0021929019301551

27. van der Kruk E, Reijne MM. Accuracy of human motion capture systems for sport applications; state-of-the-art review. Eur J Sport Sci. 2018;18(6):806–19.

28. Sutherland DH. The evolution of clinical gait analysis: Part II Kinematics. Gait Posture [Internet]. 2002;16(2):159–79. Available from: https://www.sciencedirect.com/science/article/pii/S0966636202000048

29. Wang X, Yang C, Wang L. Applying Motion Capture in Computer Animation Education. International Journal of Engineering and Manufacturing. 2011;1(4):44–52.

30. McIntyre N. Etienne Jules Marey (1830–1904). J Med Biogr. 2005;13(3):141.

31. Eichelberger P, Ferraro M, Minder U, Denton T, Blasimann A, Krause F, et al. Analysis of accuracy in optical motion capture – A protocol for laboratory setup evaluation. J Biomech [Internet]. 2016;49(10):2085–8. Available from: http://dx.doi.org/10.1016/j.jbiomech.2016.05.007

32. Kanko RM, Laende E, Selbie WS, Deluzio KJ. Inter-session repeatability of markerless motion capture gait kinematics. J Biomech. 2021 May 24;121.

33. Stelzer A, Pourvoyeur K, Fischer A. Concept and Application of LPM—A Novel 3-D Local Position Measurement System. Vol. 52, IEEE transactions on microwave theory and techniques. New York, N.Y. : Professional Technical Group on Microwave Theory and Techniques, Institute of Electrical and Electronics Engineers,; 2004. p. 2664–9.

34. Day JS, Dumas GA, Murdoch DJ. Evaluation of a long-range transmitter for use with a magnetic tracking device in motion analysis. Vol. 31, Journal of biomechanics. [Oxford] ; [New York] : Pergamon; 1998. p. 957–61.

35. Sathyan T, Shuttleworth R, Hedley M, Davids K. Validity and reliability of a radio positioning system for tracking athletes in indoor and outdoor team sports. Behav Res Methods. 2012;44(4):1108–14.

36. Shirehjini AAN, Yassine A, Shirmohammadi S. An RFID-Based Position and Orientation Measurement System for Mobile Objects in Intelligent Environments. IEEE Trans Instrum Meas. 2012;61(6):1664–75.

37. Zohlandt C, Walk L, Nawara W. Classification of Vault Jumps in Gymnastics. 2012.

38. Lee JB, Burkett BJ, Thiel D v, James DA. Inertial sensor, 3D and 2D assessment of stroke phases in freestyle swimming. Procedia Eng. 2011;13:148–53.

39. Wearnotch | Wearable 3D motion capture | Free apps | Mobile SDK | Template Code [Internet]. [cited 2022 Nov 14]. Available from: https://wearnotch.com/

40. Home - Xsens 3D motion tracking [Internet]. [cited 2022 Nov 14]. Available from: https://www.xsens.com/

41. Perception Neuron Motion Capture | Motion Capture for All – NeuronMocap [Internet]. [cited 2022 Nov 14]. Available from: https://neuronmocap.com/

42. Nijmeijer EM, Heuvelmans P, Bolt R, Gokeler A, Otten E, Benjaminse A. Concurrent validation of the Xsens IMU system of lower-body kinematics in jump-landing and change-of-direction tasks. J Biomech. 2023;154:111637.

43. Maletsky LP, Sun J, Morton NA. Accuracy of an optical active-marker system to track the relative motion of rigid bodies. Vol. 40, Journal of biomechanics. [Oxford] ; [New York] : Pergamon; 2007. p. 682–5.

44. Windolf M, Götzen N, Morlock M. Systematic accuracy and precision analysis of video motion capturing systems—exemplified on the Vicon-460 system. J Biomech. 2008;41(12):2776–80.

45. Sporri J, Schiefermuller C, Muller E. Collecting Kinematic Data on a Ski Track with Optoelectronic Stereophotogrammetry: A Methodological Study Assessing the Feasibility of Bringing the Biomechanics Lab to the Field. PLoS One. 2016;11(8):e0161757–e0161757.

46. Monnet T, Samson M, Bernard A, David L, Lacouture P. Measurement of three-dimensional hand kinematics during swimming with a motion capture system: a feasibility study. Sports engineering. 2014;17(3):171–81.

47. Wu G, Siegler S, Allard P, Kirtley C, Leardini A, Rosenbaum D, et al. ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—part I: ankle, hip, and spine. J Biomech. 2002;35(4):543–8.

48. Della Croce U, Leardini A, Chiari L, Cappozzo A. Human movement analysis using stereophotogrammetry Part 4: Assessment of anatomical landmark misplacement and its effects on joint kinematics. Vol. 21, Gait and Posture. Elsevier Ireland Ltd; 2005. p. 226–37.

49. Rahimian P, Kearney JK. Optimal Camera Placement for Motion Capture Systems. IEEE Trans Vis Comput Graph. 2017;23(3):1209–21.

50. Ripic Z, Signorile JF, Kuenze C, Eltoukhy M. Concurrent validity of artificial intelligence-based markerless motion capture for over-ground gait analysis: A study of spatiotemporal parameters. J Biomech. 2022;143:111278.

51. della Croce U, Leardini A, Chiari L, Cappozzo A. Human movement analysis using stereophotogrammetry Part 4: Assessment of anatomical landmark misplacement and its effects on joint kinematics. Vol. 21, Gait and Posture. Elsevier Ireland Ltd; 2005. p. 226–37.

52. Gorton GE, Hebert DA, Gannotti ME. Assessment of the kinematic variability among 12 motion analysis laboratories. Gait Posture. 2009 Apr;29(3):398–402.

53. Levin I, Lewek MD, Giuliani C, Faldowski R, Thorpe DE. Test-retest reliability and minimal detectable change for measures of balance and gait in adults with cerebral palsy. Gait Posture. 2019 Jul 1;72:96–101.

54. Camomilla V, Bonci T, Cappozzo A. Soft tissue displacement over pelvic anatomical landmarks during 3-D hip movements. J Biomech. 2017 Sep 6;62:14–20.

55. Cereatti A, Bonci T, Akbarshahi M, Aminian K, Barré A, Begon M, et al. Standardization proposal of soft tissue artefact description for data sharing in human motion measurements. J Biomech. 2017 Sep 6;62:5–13.

56. Holden JP, Orsinib' JA, Siegelc KL, Kepplec TM, Gerber\ LH, Stanhopec SJ. Surface movement errors in shank kinematics and knee kinetics during gait 1. Vol. 5, Gait & Posture. 1997.

57. Ceseracciu E, Sawacha Z, Fantozzi S, Cortesi M, Gatta G, Corazza S, et al. Markerless analysis of front crawl swimming. Vol. 44, Journal of biomechanics. [Oxford] ; [New York] : Pergamon; 2011. p. 2236–42.

58. Akman O. Robust augmented reality [Internet]. Delft University of Technology; 2012. Available from: http://resolver.tudelft.nl/uuid:3adeccef-19db-4a06-ab26-8636ac03f5c0

59. Sholukha V, Bonnechere B, Salvia P, Moiseev F, Rooze M, van Sint Jan S. Model-based approach for human kinematics reconstruction from markerless and marker-based motion analysis systems. J Biomech [Internet]. 2013;46(14):2363–71. Available from: http://dx.doi.org/10.1016/j.jbiomech.2013.07.037

60. Ortega B, Salazar O, Valdivia G, Escobedo C. Therapeutic Motion Analysis of Lower Limbs Using Kinovea. International Journal of Soft Computing and Engineering (IJSCE). 2013;(3):2231–307.

61. Moral-Munoz JA, Esteban-Moreno B, Arroyo-Morales M, Cobo M, Herrera-Videma E. Agreement between face-to-face and free software video analysis for Assessing hamstring flexability in adolescents. J Strength Cond Res. 2015;29(9):2661–5.

62. Charmant J, contributors. Kinovea (Version 0.9.5) [Internet]. 2021. Available from: https://www.kinovea.org

63. Puig-Diví A, Escalona-Marfil C, Padullés-Riu JM, Busquets A, Padullés-Chando X, Marcos-Ruiz D. Validity and reliability of the Kinovea program in obtaining angles and distances using coordinates in 4 perspectives. PLoS One. 2019;14(6):1–14.

64. Keller VT, Outerleys JB, Kanko RM, Laende EK, Deluzio KJ. Clothing condition does not affect meaningful clinical interpretation in markerless motion capture. J Biomech. 2022;141:111182.

65. Kanko RM, Laende EK, Davis EM, Selbie WS, Deluzio KJ. Concurrent assessment of gait kinematics using marker-based and markerless motion capture. J Biomech. 2021 Oct 11;127.

66. Kanko RM, Laende EK, Strutzenberger G, Brown M, Selbie WS, DePaul V, et al. Assessment of spatiotemporal gait parameters using a deep learning algorithm-based markerless motion capture system. J Biomech. 2021 Jun 9;122:110414.

67. Wren TAL, Isakov P, Rethlefsen SA. Comparison of kinematics between Theia markerless and conventional marker-based gait analysis in clinical patients. Gait Posture [Internet]. 2023;104:9–14. Available from: https://www.sciencedirect.com/science/article/pii/S0966636223001443

68. D'Souza S, Fohanno V. Comparison between Theia3D markerless and CAST model marker-based systems in pathological lower-body 3D gait kinematics in adults and children. Gait Posture. 2023;100:40–1.

69. Corazza S, Mündermann L, Gambaretto E, Ferrigno G, Andriacchi TP. Markerless Motion Capture through Visual Hull, Articulated ICP and Subject Specific Model Generation. Int J Comput Vis. 2010;87(1–2):156–69.

70. Liu G, Tang X, Cheng HD, Huang J, Liu J. A novel approach for tracking high speed skaters in sports using a panning camera. Pattern Recognit. 2009;42(11):2922–35.

71. Moore DD, Walker JD, MacLean JN, Hatsopoulos NG. Validating markerless pose estimation with 3D X-ray radiography. J Exp Biol. 2022;225(9).

72. Drazan JF, Phillips WT, Seethapathi N, Hullfish TJ, Baxter JR. Moving outside the lab: Markerless motion capture accurately quantifies sagittal plane kinematics during the vertical jump. J Biomech. 2021;125:110547.

73. Li R, St George RJ, Wang X, Lawler K, Hill E, Garg S, et al. Moving towards intelligent telemedicine: Computer vision measurement of human movement. Comput Biol Med [Internet]. 2022;147(April):105776. Available from: https://doi.org/10.1016/j.compbiomed.2022.105776

74. Williams S, Zhao Z, Hafeez A, Wong DC, Relton SD, Fang H, et al. The discerning eye of computer vision: Can it measure Parkinson's finger tap bradykinesia? J Neurol Sci [Internet]. 2020;416(February):117003. Available from: https://doi.org/10.1016/j.jns.2020.117003

75. Zhan W, Zou Y, He Z, Zhang Z. Key Points Tracking and Grooming Behavior Recognition of Bactrocera minax (Diptera: Trypetidae) via DeepLabCut. Math Probl Eng. 2021;2021.

76. Wu G, Cavanagh PR. ISB recommendations for standardization in the reporting of kinematic data. J Biomech. 1995;28(10):1257–61.

77. Karashchuk P, Rupp KL, Dickinson ES, Walling-Bell S, Sanders E, Azim E, et al. Anipose: A toolkit for robust markerless 3D pose estimation. Vol. 36, Cell Reports. 2021.

78. An G, Lee S, Seo MW, Yun K, Cheong WS, Kang SJ. Charuco Board-Based Omnidirectional Camera Calibration Method. Electronics (Basel). 2018;7(12):421.

79. Bradski G. The OpenCV Library. Dr Dobb's Journal of software tools [Internet]. 1998;130(2):556. Available from: http://dx.doi.org/10.1016/j.jaci.2012.05.050

80. Antico M, Balletti N, Laudato G, Lazich A, Notarantonio M, Oliveto R, et al. Postural control assessment via Microsoft Azure Kinect DK: An evaluation study. Comput Methods Programs Biomed. 2021 Sep 1;209.

81. Pantzar-Castilla E, Cereatti A, Figari G, Valeri N, Paolini G, della Croce U, et al. Knee joint sagittal plane movement in cerebral palsy: a comparative study of 2-dimensional markerless video and 3-dimensional gait analysis. Acta Orthop. 2018 Nov 2;89(6):656–61.

82. Riazati S, McGuirk TE, Perry ES, Sihanath WB, Patten C. Absolute Reliability of Gait Parameters Acquired With Markerless Motion Capture in Living Domains. Front Hum Neurosci. 2022;16:867474.

83. Drory A, Li H, Hartley R. A Learning-based Markerless Approach for Full-body Kinematics Estimation In-Natura from a Single Image. J Biomech. 2017;55:1–10.

84. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med. 2012;31(29):3972–81.

85. Jush FK, Biele M, Dueppenbecker PM, Maier A. Deep Learning for Ultrasound Speed-of-Sound Reconstruction: Impacts of Training Data Diversity on Stability and Robustness. 2022;

86. Garrido-Jurado S, Muñoz-Salinas R, Madrid-Cuevas FJ, Marín-Jiménez MJ. Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognit. 2014 Jun;47(6):2280–92.

87. Tomar S. Converting video formats with FFmpeg. Linux Journal. 2006;2006(146):10.

88. Jung Alexander. Machine Learning The Basics . 1st ed. 20. Springer Nature eBook. Singapore: Springer Singapore; 2022. (Machine Learning: Foundations, Methodologies, and Applications).

89. Poggi M, Tosi F, Mattoccia S. Quantitative Evaluation of Confidence Measures in a Machine Learning World. In: ICCV. IEEE; 2017. p. 5238–47.

90. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. Trends Sport Sci [Internet]. 2014;1(21):19–25. Available from: http://www.wbc.poznan.pl/Content/325867/5_Trends_Vol21_2014_ no1_20.pdf

Appendix A

DALHOUSIE UNIVERSITY

CONSENT FORM

Project title: 3D Validation of DeepLabCut as a Markerless Motion Capture Tool

Lead researcher
Seth Daley, Health and Human Performance
Email: s.daley@dal.ca
(902) 456-7564

Other researchers
Ryan Frayne, PhD
Email: ryan.frayne@dal.ca
(902) 494-6499
Assistant Professor:
School of Health and Human Performance
Dalhousie University
Halifax, Nova Scotia, Canada

Introduction
We invite you to take part in this research study being conducted by myself, Seth Daley, and Dr. Ryan Frayne at Dalhousie University as part of my Master of Science (Kinesiology) degree. The choice to participate in this study is completely your decision. There will be no impact on your academic studies, athletic performance or services provided should you choose not to participate in this research. The information below explains to you what is involved in the research and what you will be expected to do, about any benefit, risk, inconvenience or discomfort you may experience.

If you have any questions about the study please contact the lead researcher, Seth Daley or Dr. Ryan Frayne (Supervisor). Please ask as many questions as you like.

Research Study Background and Purpose
Motion capture is a commonly used tool in research but often requires the use of expensive systems as complicated marker sets placed on the participant. DeepLabCut is a machine learning application that can track movement from video files without the need for markers or laboratory setups. For scientists to be confident using the tool however, it needs to be shown to be as accurate as the current best systems by comparing the results against each other.

Who Can Take Part in the Research Study
Anyone between the ages of 18 and 65 who are comfortable preforming basic movements like standing jumps, bodyweight squats, walking, and reaching are able to participate.

What You Will Be Asked to Do
This study will consist of one session at the Dalplex, 6160 South Street, Halifax, NS. The lead researcher will meet you at the front door of the building and escort you to the biomechanics ergonomics neuroscience lab (BENlab) located inside the Kinesiology Department of the Dalplex. You will be asked to wear snug fitting comfortable clothing.

You will have your height and the length of your arms and legs measured. You will then be randomly assigned to one of two groups. The first group will perform 5 separate movements, each recorded by a series of cameras, and then a 6^{th} movement with a series of small markers placed at various parts of your legs, hips, and shoulders. The second group will perform all 6 movements with the markers applied and will be recorded both by cameras and by an optical motion capture system. The movements will include the following, in a randomized order: walking along a flat surface, jumping into the air, turning in a circle, three repetitions of a jumping jack, swinging a leg through the air, and standing up from a seated position.

Possible Benefits, Risks and Discomforts
As the movements performed will be simple ones without any notable effort there will be no anticipated benefits from participation beyond contributing to the understanding of markerless motion capture.

There are no anticipated risks or discomforts from participation in this study.

How your information will be protected:
Confidentiality: Once you have enrolled in the study you will be assigned an anonymous study number to protect your confidentiality. All data collection forms and participant data will only use the study number to identify you. Only the investigators will have access to these numbers.

Privacy: The lead researcher and supervisor may be present during the data collection to ensure protocols run smoothly. If a participant has a personal inquiry or question regarding the study they will be given an opportunity to speak directly to the lead researcher.

Data retention: All paperwork will be stored in a locked cabinet owned by the lead researcher's supervisor. Electronic materials (including data and images) will be stored on a secured, password-protected computer with images edited to remove identifying

features. Videos will be edited after analysis to remove any identifying features. Data will be retained for 5 years, with all data being destroyed after this time period by the lead researcher.

If You Decide to Stop Participating

You are free to leave the study at any time. If you decide to stop participating at any point in the study, you can also decide whether you want any of the information that you have contributed up to that point to be removed or if you will allow us to use that information. You can also decide for up to 1 month if you want us to remove your data. After that time, it will become impossible for us to remove it because it will already be analyzed.

How to Obtain Results

A summary report including graphical presentations will be prepared and made available to you and all participants if you wish to know the outcome of the study.

Questions

We are happy to talk with you about any questions or concerns you may have about your participation in this research study. Please contact Seth Daley (s.daley@dal.ca) or Ryan Frayne (ryan.frayne@dal.ca) at any time with questions, comments, or concerns about the research study.

If you have any ethical concerns about your participation in this research, you may also contact Research Ethics, Dalhousie University at (902) 494-1462, or email: ethics@dal.ca (and reference REB file # 2022-5974).

Signature Page

Project Title: 3D Validation of DeepLabCut as a Markerless Motion Capture Tool

Lead Researcher
Seth Daley, Health and Human Performance, Dalhousie University
Email: s.daley@dal.ca

I have read the explanation about this study. I have been given the opportunity to discuss it and my questions have been answered to my satisfaction. I understand that I have been asked to take part in one session that will occur at Dalplex and that physiological measures and video will be recorded during this session. I agree to take part in this study. My participation is voluntary, and I understand that I am free to withdraw from the study at any time and can ask for my results to be removed from the study up until 1 month after the session is completed.

_____ _____ _____

Name Signature Date

I wish to have the outcomes of the study sent to me.

_____ _____ _____

Email Address Signature Date

Appendix B

DALHOUSIE UNIVERSITY

Health Sciences Research Ethics Board
Letter of Approval

February 11, 2022

Seth Daley
Health\School of Health and Human Performance

Dear Seth,

REB #: 2022-5974
Project Title: 3D Validation of DeepLabCut as a Markerless Motion Capture Tool

Effective Date: February 11, 2022
Expiry Date: February 11, 2023

The Health Sciences Research Ethics Board has reviewed your application for research involving humans and found the proposed research to be in accordance with the Tri-Council Policy Statement on *Ethical Conduct for Research Involving Humans.* This approval will be in effect for 12 months as indicated above. This approval is subject to the conditions listed below which constitute your on-going responsibilities with respect to the ethical conduct of this research.

Effective March 16, 2020: Notwithstanding this approval, any research conducted during the COVID-19 public health emergency must comply with federal and provincial public health advice as well as directives from Dalhousie University (and/or other facilities or jurisdictions where the research will occur) regarding preventing the spread of COVID-19.

Sincerely,

Dr. Lori Weeks, Chair