ESTABLISHING TRAINING PARAMETERS FOR A DEEP NEURAL NETWORK TO ASSESS 2D, FRONTAL PLANE KINEMATICS

Caleb D. Johnson[1], Lauren K. Sara[1], Taylor Kofton[1], William Marshall[1], Julie M. Hughes[2], Stephen A. Foulis[2], Irene S. Davis[1]

[1]Harvard Medical School, Department of Physical Medicine and Rehabilitation, Cambridge, MA, USA, [2]Military Performance Division, U.S. Army Research Institute of Environmental Medicine, Natick MA, USA

The purpose of this study was to establish the optimal training parameters to assess frontal plane, 2D kinematics using DeepLabCut. DeepLabCut is an open-source platform that allows the user to train neural networks for customized feature detection in 2D videos. Deep neural networks were trained using frontal plane videos from 41 participants who completed single- and double-leg drop landings. Training was initiated from a pretrained network (ResNet50) and for identification of 10 landmarks. Networks were trained with an increasing number of training iterations (25-250k) and frames (200-800). Train/test errors were calculated as the mean Euclidean error between network-based predictions and manually digitized landmarks. Our results indicate that 175k training iterations and 400 training frames were adequate for stable network performance (training/test errors= 2.8/3.7 pixels).

KEYWORDS: motion analysis, DeepLabCut, markerless, deep learning

INTRODUCTION: Several open-source platforms for markerless motion capture offer the ability to track 2D kinematics using a single digital video camera. These platforms leverage deep learning, starting from pre-trained neural networks, to fine-tune new neural networks for customized feature detection. These networks allow for the rapid, automated digitization of videos to generate 2D coordinate data, which can then be used to calculate 2D kinematics. One of these platforms, DeepLabCut (DLC), has been shown to produce reliable/accurate 2D coordinate data, resulting in reliable and accurate 2D kinematics for human movement.(Drazan et al., 2021; Papic et al., 2021) However, only one study has reported on the necessary network training parameters to optimize DLC performance.(Cronin et al., 2019)

Cronin et al. (2019) found that 300-400 training frames and 200,000 training iterations yielded optimal network performance for tracking underwater running. However, this study, as well as those by Drazan (2021) and Papic (2021), focused on the sagittal plane. Frontal plane kinematics have shown higher errors even for marker-based motion capture systems (Schmitz et al., 2014) In addition, more landmarks, inherently necessary to track both limbs for the frontal plane, can be more challenging for neural networks.

Therefore, our purpose was to establish the DLC training parameters for optimal tracking of frontal plane kinematics during a commonly studied movement pattern; drop landings. Drop landings are a commonly studied movement pattern in active populations (e.g., athletes, soldiers, etc.), providing a measure of dynamic lower extremity control. We hypothesized that a larger number of training frames and iterations, in comparison to previous work, would be necessary to achieve stable network performance.

METHODS: Participants included 41 US Army recruits (24/17 women/men, age= 20±4yrs, height= 1.7±0.9m, mass= 65.1±11.2kg). Approval was obtained from the overseeing Institutional Review Board and written informed consent was obtained from all participants. Participants were first prepped with orange markers placed on landmarks that would be difficult to identify for video analysis later. These included the bilateral anterior superior iliac spine (hip markers) and heads of the 1st/5th metatarsals (medial/lateral feet markers). Participants performed single- and double-leg drop landings from a 30.5/45.7cm box, to a 61x61cm landing area marked with tape on the ground (Figure 1). A camera (GoPro Hero9, resolution= 1080p, rate= 240fps) was positioned in the frontal plane, approximately 96cm from the landing area and at a height of 46cm. Participants completed 3 trials for each type of movement (total videos= 9). The full movement was recorded; from stepping off the box through landing. A single video for each participant was selected for network training and analysis.

A total of 800 training frames (18-20 per video) were extracted from videos using the k-means algorithm for random selection of frames within each video.(Nath et al., 2019) These frames were manually annotated with digital landmarks, including bilateral hip, knee, ankle, and medial/lateral foot markers (Figure 1). Four neural networks were then trained in an iterative fashion, using sets of 200, 400, 600, and 800 training frames. Training was performed using an Nvidia® GeForce® RTX 2070 GPU (Nvidia Corp, Santa Clara, CA) and a laptop computer (Dell Inspiron G7 15-7500, RAM: 32GB, CPU: 1TB SSD). All networks used the ResNet50 pretrained network for initial weights and a 90/10 training/test fraction, meaning 90% of frames were randomly selected to include in the training set and 10% were left out to test the model after training. Each network was trained for a total of 250,000 iterations, with networks benchmarked every 25,000 iterations (i.e., 25k, 50k, 75k, etc.)

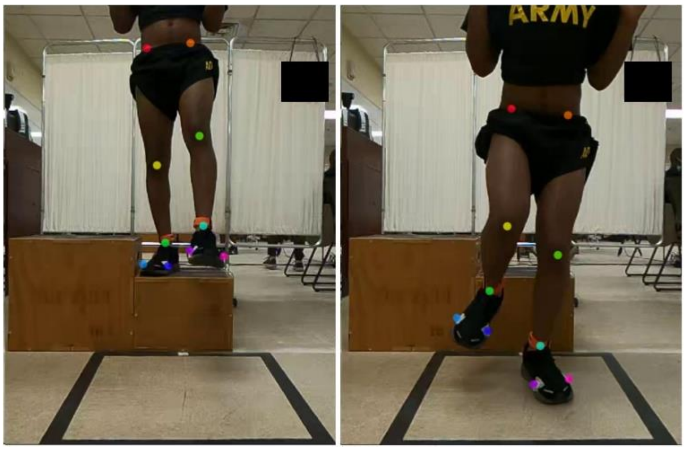

Figure 1: Example annotated images for DLC training sets

To assess network performance, training and test errors (pixels) were calculated as the mean Euclidean error between marker locations that were predicted by the network and placed manually by the user. Training error refers to the error on frames used to train the network, while the test error refers to the error on frames left out of the training set (i.e., novel frames). Errors were calculated separately for each marker, network (i.e., networks trained using 200/400/600/800 frames), and training benchmark (i.e., 25k iterations, 50k iterations, etc.). Mean errors, averaged across the 4 networks, were plotted by the number of training iterations. A 1x10 repeated measures ANOVA was used to test for the main effect, with post-hoc pairwise comparisons performed using Bonferroni adjusted p-values. Using the number of training iterations identified through this analysis (175,000), mean errors were then plotted by number of training frames for each marker. The minimum number of training frames for stable network performance was determined subjectively, similar to the previous analysis by Cronin et al. (2019). Briefly, this was based on plateauing and limiting the spread of errors, as well as convergence between the types of eror (training/test).

RESULTS: The main effect of increasing training iterations on mean training/test errors is depicted in Figure 2. Both types of error appeared to stabilize around 150k-175k iterations. Results of repeated measures ANOVAs confirmed this. There was a significant main effect of training iterations on mean training (F= 57.4, $p< 0.01$) and test (F= 24.4, $p< 0.01$) errors. Post-hoc testing demonstrated that significant differences in mean training and test errors between increasing levels of iterations were resolved after 175k iterations. In other words, after 175k, mean differences between additional levels of iterations (i.e., 200k, 225k, 250k), did not reach significance (p= 0.25-1.00).

The effect of increasing training frames on network performance (using 175,000 training iterations) is shown in Figure 3. Mean training errors were found to be slightly lower for networks trained with less frames (200/400) compared to those with more (600/800). The difference was small, about 0.68px on average. For test errors, there was a fairly significant decrease from 200 to 400/800 frames, of approximately 1.55px. However, with 600 training frames, the test errors were similar to 200 (difference= 0.24px). Also of note, errors were markedly higher for the knee marker in comparison to all others, especially for test errors. The average test error for the knee was 7.1px, with the next closest marker, the ankle, at 5.4px.

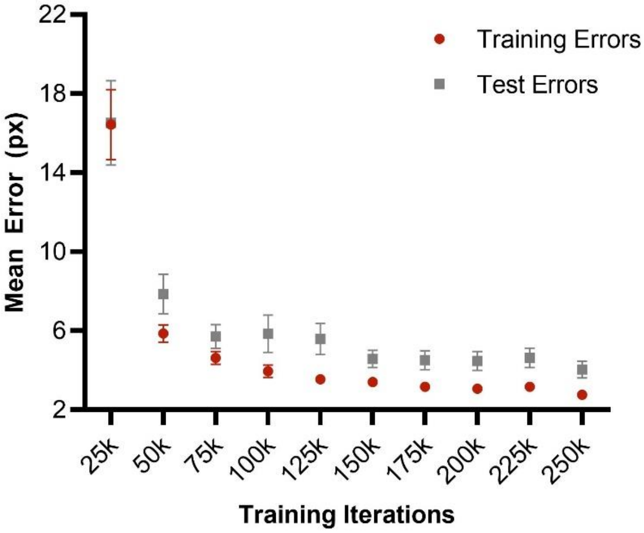

Figure 2: Mean network errors (± standard deviation), across networks/landmarks, by training iterations

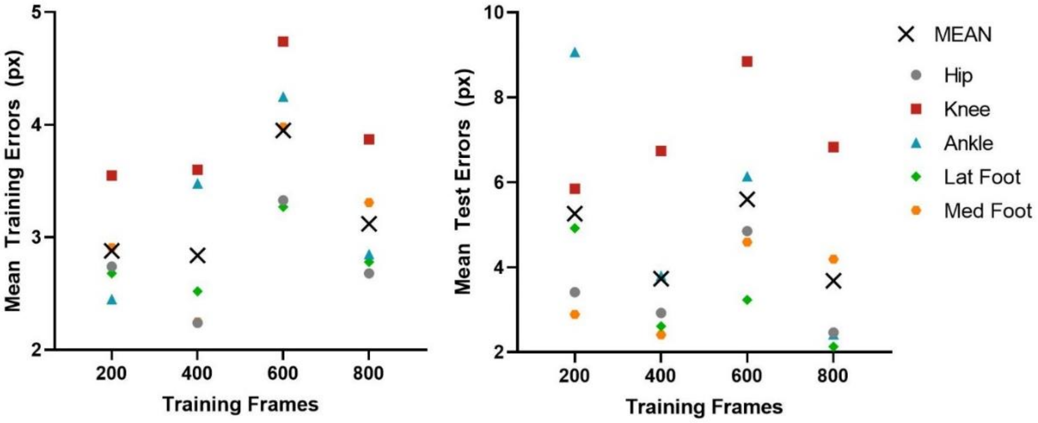

Figure 3: Mean network errors (using 175k iterations) by training frames

DISCUSSION: Our purpose was to establish the DLC training parameters for optimal tracking of 2D frontal plane kinematics during drop landings. To do this, we studied the effect of increasing training iterations (25k-250k and frames (200-800) on training and test errors for the resulting neural network.

We first found an exponential decrease in errors with increasing training iterations, plateauing at approximately 175k. With additional iterations yielding no significant differences in errors, we would interpret this as the minimum number of training iterations necessary. Counter to our hypothesis, this is lower than the ideal number of training iterations (200k) identified by Cronin (2019) for a sagittal view of underwater running. However, in their paper, the authors only identify this number subjectively and do not perform statistical testing to confirm it. In addition, it does appear as if their average network performance plateaus earlier; around 160k-165k iterations. In any case, our results generally agree with this prior study, providing more support for a minimum number of training iterations in the range of 175k-200k.

Our second finding was that network training and test performance were affected differently by increasing the number of training frames. Overall, training performance was slightly poorer and test performance significantly better, for networks with a higher number of training frames. Similar to the findings of Cronin (2019), this indicates overfitting of the 200-frame network. In other words, a network may be less “flexible” with inadequate training frames, becoming overly adept at making predictions for images it has seen at the expense of predictions on novel frames. It is difficult to explain why test performance decreased with 600 training frames in comparison to 200/400. However, Cronin (2019) also found this discrepancy, to some degree. Their networks showed small decreases in training performance between 300-500 training frames, as well as decreases in test performance between 100-200 frames. This is likely due to differences in the training/test data sets used for each individual network, due to both the elimination of training frames to form reduced data sets, as well as the random train/test split. Overall, our results showed that 400 training frames is likely a good minimum threshold for stable performance of a trained network, yielding mean training/test errors of 2.8/3.7px (≈ 3.6/4.8mm). This was also counter to our hypotheses, as Cronin (2019) found that 300-400 frames were ideal for the sagittal plane and underwater running. Our network errors were similar to those reported by Cronin (2019); training/test errors of 1.4/2.9px (≈ 4.8/10mm).

Our last finding was that errors were highest for the knee marker. This is likely due to a lack of discernible landmarks around the knee joint center in the frontal plane, leading to high labelling variability. Physical markers were placed to identify the hip and foot markers. While no marker was placed for the ankle joint center, labelling for this landmark was aided by referencing shoes and the crease between the foot and shank. It is interesting that the second highest errors were found for the ankle marker, highlighting the impact of placing physical markers to aid with labelling consistency. In addition, the physical markers also would provide a consistent pixilation pattern at these landmarks between participants, further aiding in predictions by the trained networks. However, there is a trade-off with using markers and time/efficiency.

CONCLUSION: In summary, our results demonstrate that 175,000 training iterations and 400 training frames should be used when training DLC networks for 2D frontal plane kinematics. Importantly, this is with the qualifications that similar testing procedures and equipment, as well as number of landmarks, are used as in the current study. While counter to our hypothesis, these findings are in line with previous work by Cronin et al. (2019), suggesting that the training parameters for optimal network performance in human movement applications may be robust to the plane of view and movement pattern being assessed. Future research should focus on validating frontal plane kinematic variables derived from trained DLC networks, as well as cross-validation of our network results in novel samples.

REFERENCES

Cronin, N. J., Rantalainen, T., Ahtiainen, J. P., Hynynen, E., & Waller, B. (2019). Markerless 2D kinematic analysis of underwater running: A deep learning approach. *J Biomech*, *87*, 75-82. https://doi.org/10.1016/j.jbiomech.2019.02.021

Drazan, J. F., Phillips, W. T., Seethapathi, N., Hullfish, T. J., & Baxter, J. R. (2021). Moving outside the lab: markerless motion capture accurately quantifies sagittal plane kinematics during the vertical jump. *J Biomech*, 110547. https://doi.org/10.1016/j.jbiomech.2021.110547

Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., & Mathis, M. W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols*, *14*(7), 2152-2176. https://doi.org/10.1038/s41596-019-0176-0

Papic, C., Sanders, R. H., Naemi, R., Elipot, M., & Andersen, J. (2021). Improving data acquisition speed and accuracy in sport using neural networks. *J Sports Sci*, *39*(5), 513-522. https://doi.org/10.1080/02640414.2020.1832735

Schmitz, A., Ye, M., Shapiro, R., Yang, R., & Noehren, B. (2014). Accuracy and repeatability of joint angles measured using a single camera markerless motion capture system. *J Biomech*, *47*(2), 587-591. https://doi.org/10.1016/j.jbiomech.2013.11.031

ACKNOWLEDGEMENTS: We would like to thank Takashi Nagai, Peter Frykman, Kyle Murdock, Michelle Aube, and David Zeppetelli, as well as the US Army recruit volunteers for their assistance with this study. This study was supported by the US Department of Defense award W81XWH-20-C-0123 P0001. **Disclaimer:** The opinions or assertions in this manuscript are the private views of the authors and should not be construed as official policy or reflecting the views of the US Army.