Bachelor thesis

Artificial Intelligence

Radboud University 

Integrated 2D video-based analysis using DeepLabCut pose estimation

Author:
Tim Hiemstra
s1020403

First supervisor:
prof. dr. W.P. Medendorp
Donders Institute; Radboud University Nijmegen
p.medendorp@donders.ru.nl

Second supervisor:
C.M.J. Willemsen MSc
Donders Institute; Radboud University Nijmegen
c.willemsen@donders.ru.nl

Second reader:
Dr Y. Güçlütürk
Donders Institute; Radboud University Nijmegen
y.gucluturk@donders.ru.nl

June 18, 2021

Abstract

Readily available pose estimation algorithms often require extensive calibration methods for 3D pose estimation whereas 2D pose estimation does not always capture all motion. At present, 3D pose estimation approaches often require non-trivial deep learning methods and there is a lack of sufficient 3D datasets. In this study, the feasibility of integrating two different camera views of human locomotion using DeepLabCut was investigated. A significant perspective difference in most x-components was observed, while all y-components did not have a significant perspective difference. This suggests that integrating 2D pose estimation is not feasible using this approach.

Contents

1 Introduction

Recent advances in video-based pose estimation have provided practical frameworks for automating video-based analysis of human- and animal movement [1, 2, 3, 4]. These markerless pose estimation software packages are publicly available and can simplify the process of video-based gait analysis. Previously, gait analysis could only be performed by manually annotating body parts or using marker-based motion capture. Marker-based motion capture is very accurate but it also has many disadvantages.

First of all, marker-based motion capture requires specific hardware, such as reflective markers and infrared cameras, which are expensive. Additionally, determining the correct amount of markers and their location is time-consuming and hard to reproduce. A possible solution to the aforementioned flaws of marker-based motion capture is the use of markerless pose estimation.

Markerless pose estimation can speed up the process of capturing motion by allowing movement to be analyzed outside of laboratories [5]. Nevertheless, markerless pose estimation is still lacking in some cases. Namely, 2D pose estimation often cannot detect the complete movement as it is a 2D view of 3D motion. Furthermore, 3D pose estimation often requires extensive calibration that is not present in many datasets. However, integrating multiple 2D perspectives can aid in detecting complex movement. The aim of this study is to find out whether 2D video integration is a feasible approach to pose estimation of human locomotion, through the use of DeepLabCut.

DeepLabCut is a deep convolutional network which combines pretrained ResNets with deconvolutional layers. The network consists of a variant of ResNet, which was trained using transfer learning on the object recognition benchmark called ImageNet [3, 6]. When using the DeepLabCut network it is first trained on frames in which specific body parts are labeled. The trained network can in turn analyze videos and predict labels on new videos. The network outputs the labeled videos with the pixel coordinates and confidence values of these labels.

DeepLabCut is available for free, and only requires a video capturing device in addition to being applicable in out-of-lab environments. Furthermore, DeepLabCut requires relatively little coding experience as there are many great examples available on their repository. DeepLabCut is mostly used for animal applications, but through its easy use it can aid clinical applications as well as smaller research labs in analyzing human movement.

In this study, DeepLabCut was trained on videos of walking humans. These

videos were provided in the GPJATK dataset[7]. DeepLabCut analyzed both perspectives and labeled them. Afterwards, both perspectives were compared to each other to assess whether integrating 2D perspectives is a feasible approach to 2D video-based pose estimation.

2 Methods

2.1 Dataset

I used the publicly available multi-view video and motion capture GPJATK dataset of walking data sequences from 32 participants (10 women and 22 men) [7]. The dataset contains motion capture, gait and video data of walking sequences. The walking sequences were traversed in a straight line or diagonally. Only a subset of this dataset has been used. Specifically, only the videos showcasing left-to-right and right-to-left movements in a straight line were used. Two trials per participant contained this movement. However, one participant did not participate in these two trials and in one trial the participant abruptly stopped moving, turned their body and then returned to walking. This resulted in the exclusion of these trials. Therefore, I analyzed 61 trials with an average duration of 5.01 ± 0.76 seconds.

The walking space was captured by four RGB cameras that recorded both the sagittal plane views as well as the frontal plane views, which can be seen in Figure 1. The videos of cameras C1 and C3 were used in this study. The cameras recorded at 25 Hz with a 960x540 pixel resolution. The videos of the cameras are all synchronized, so the videos of the different cameras have the same duration and contain the same trial.

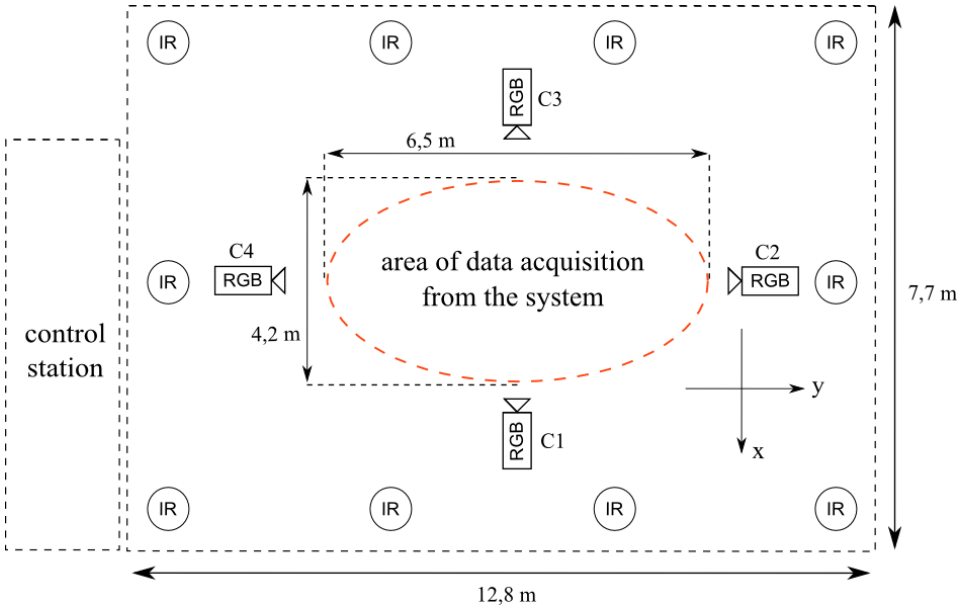


Figure 1: Camera layout [7]

2.2 Training DeepLabCut

20 out of the 61 right-to-left walking sequence videos were randomly selected and split into frames. DeepLabCut then randomly selects 10 of these frames. In each of these frames, I labeled the neck, hip, knee, ankle, heel and the center of the toes of the side of the body facing the camera. The labeled data was then used to train the DeepLabCut network for 200.000 iterations.

2.3 DeepLabCut analysis

To integrate the two sagittal planes, the opposite views of the 61 right-to-left data sequences were mirrored horizontally. This resulted in 122 videos of right-to-left walking sequences, which were then analyzed and labeled by the trained DeepLabCut network.

2.4 Post processing

As the movement in the data sequences is from right to left, I inverted the coordinate system so that it has the origin in the bottom right corner of the frame. This is done so that the x-values are increasing instead of decreasing over time.

In the evaluated videos, DeepLabCut occasionally labeled the wrong limb. The false positive rates for the knee, ankle, heel and toes are about 0.9%,

1.2%, 0.8% and 1.6% respectively. Most of these errors were detected automatically by discarding values above a certain margin and fixed through inter- and extrapolation. Interpolation was done by first checking whether the x-value of the current label point was different from the previous and next frame by a certain margin and if so by changing the current point to the average of the previous and next frame. Extrapolation was done by first checking if the x-value of the current label point is different from the previous point by a certain margin and if the x-value of the current label point is close to the position of the other leg.

2.5 Gait cycles

To calculate the gait cycles, the x-component of the limb of interest is subtracted by the x-component of the neck. The neck was chosen as ground truth for the position of the body, as this label seemed to be the most consistent body marker. Heel strikes and toe offs were then identified based on the distance the ankle was from the body. One gait cycle is the movement in between heel strikes from the same leg. These gait cycles were then interpolated to the same size and averaged per trial.

2.6 Statistical analysis

I used a two-way repeated measures ANOVA to determine the difference of the sagittal views for the neck, hips, knees, ankles, heels and toes. The within-subject factors were time and perspective. This test was performed per limb per dimension, resulting in 12 different tests. The level of significance was set at 0.05.

3 Results

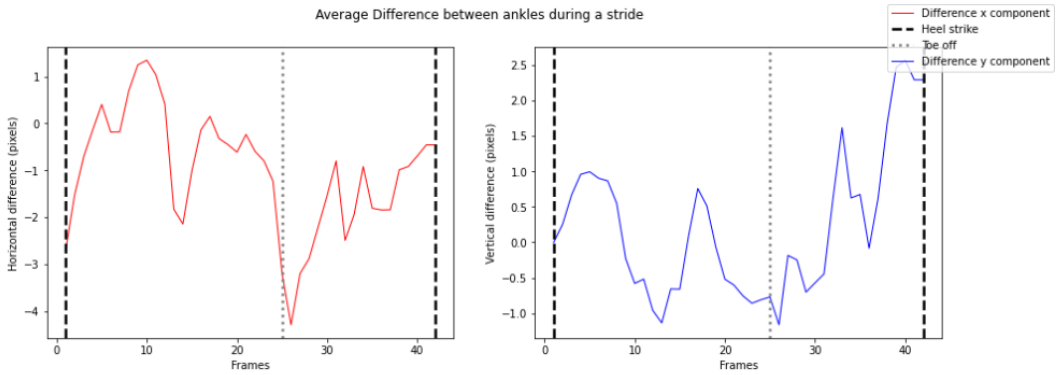
Figure 2 describes the following differences between perspectives in pixels for every labeled body part, with the most notable difference being that of the y-component of the neck. Table 1 illustrates the p-values, which were also calculated by two-way repeated measures ANOVA with within-subject factors time and perspective. The effect of time was significantly different for each labeled body part and axis. The interaction of time and perspective was only significantly different for the x-axis of the hip. The effect of perspective was significantly different for the x-axis of the ankle, toe, knee, hip and neck.

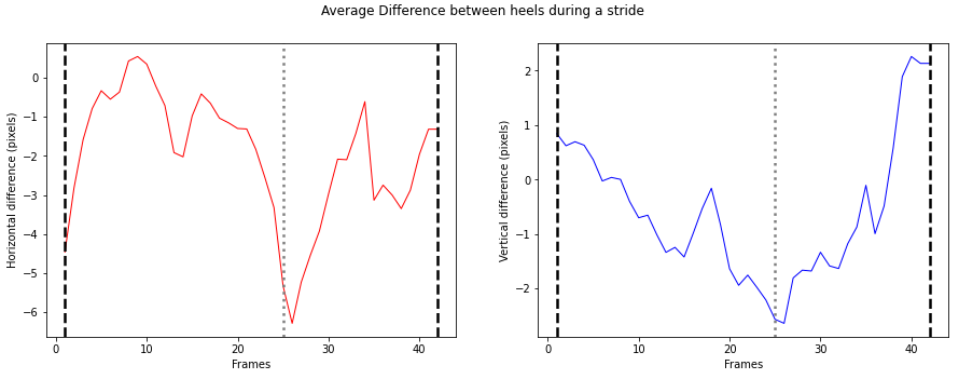
Effect of perspective		
Body part	F-value	p-value
x-value of ankle	4.487041	0.038304
y-value of ankle	0.000354	0.985057
x-value of toe	6.231604	0.015314
y-value of toe	0.024829	0.875323
x-value of heel	3.033732	0.086675
y-value of heel	0.268083	0.606524
x-value of knee	7.429682	0.008397
y-value of knee	0.019486	0.889449
x-value of hip	4.077560	0.047930
y-value of hip	0.260430	0.611698
x-value of neck	5.124007	0.027226
y-value of neck	0.027377	0.869139

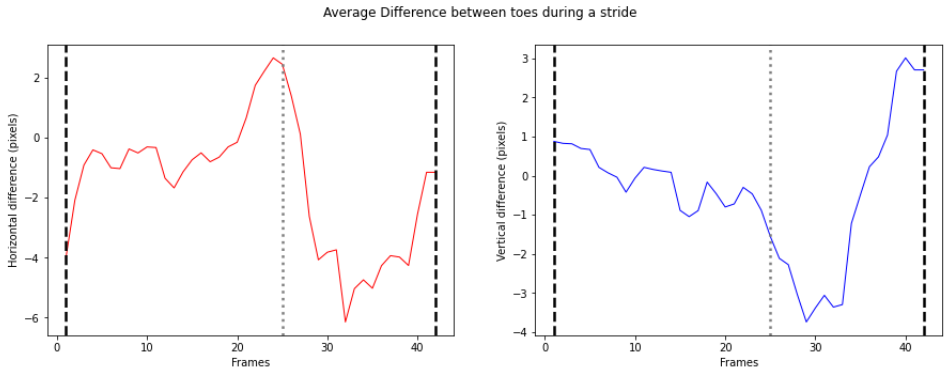
Effect of time		
Body part	F-value	p-value
x-value of ankle	3453.72	8.18e-124
y-value of ankle	902.65	5.38e-162
x-value of toe	3831.51	9.43e-117
y-value of toe	1273.57	2.87e-137
x-value of heel	3529.13	3.40e-118
y-value of heel	817.06	3.33e-131
x-value of knee	1538.61	4.29e-148
y-value of knee	341.31	2.73e-101
x-value of hip	73.14	3.04e-45
y-value of hip	13.75	9.88e-17
x-value of neck	10389.02	2.22e-130
y-value of neck	28.41	2.33e-28

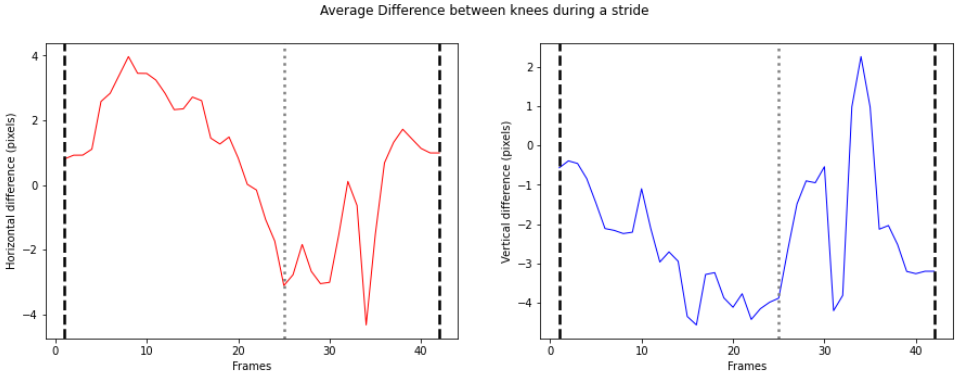
Effects of within-subject factors		
Body part	F-value	p-value
x-value of ankle	0.75	0.50
y-value of ankle	1.71	0.16
x-value of toe	0.88	0.44
y-value of toe	1.89	0.12
x-value of heel	0.57	0.62
y-value of heel	0.94	0.42
x-value of knee	0.63	0.59
y-value of knee	0.83	0.50
x-value of hip	2.74	0.01
y-value of hip	1.45	0.17
x-value of neck	0.79	0.48
y-value of neck	2.57	0.10

Table 1: Statistical test results


Average Difference between ankles during a stride
Difference x component
Heel strike
Toe off
Difference y component
Horizontal difference (pixels)
Vertical difference (pixels)
Frames


Average Difference between heels during a stride
Horizontal difference (pixels)
Vertical difference (pixels)
Frames


Average Difference between toes during a stride
Horizontal difference (pixels)
Vertical difference (pixels)
Frames


Average Difference between knees during a stride
Horizontal difference (pixels)
Vertical difference (pixels)
Frames

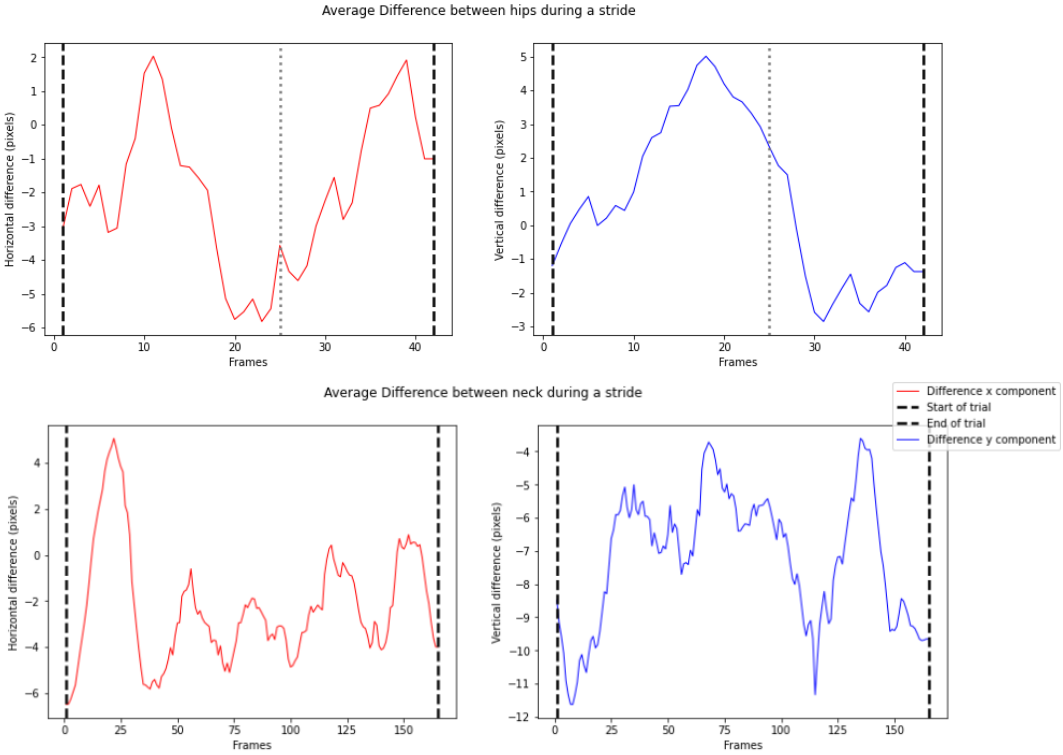


Figure 2: Gait cycle perspective differences averaged across subjects

4 Discussion

In this study, I investigated the feasibility of integrating 2D pose estimation using DeepLabCut. The results suggest that the workflow that was used does not provide a feasible approach to integrating 2D pose estimation. I found that there is a significant difference in perspectives for the x-component of the ankle, toe, knee, hip and neck. However, there is no significant difference in perspectives for the y-components of the ankle, toe, heel, knee, hip and neck.

Furthermore, the effect of time was significantly different for every limb and axis, which is not surprising since movement in a gait cycle is not constant. Lastly, the interaction of time and perspective is only significantly for the x-component of the hip. A possible reason for this is marker occlusion, since the hands occasionally obstruct the hip label during a gait cycle. This produces low confidence labels. From these results, the effect of perspectives is the most relevant for investigating the feasibility of integrating two camera views.

The effect of perspectives not being significant for the y-components of all labeled body parts, suggests that a more extensive method can provide results that show a less significant difference for the x-components. Namely, since the x-component of a labeled body part moves considerably more than the y-component, it is more error prone. Improvements to the methodology such as a better dataset, trained network or post processing procedure could be used to try to label these x-components more accurately. For each of these parts in the methodology there are several improvements that can be made.

First of all, the dataset used RGB cameras that recorded at 25 Hz with a 960x540 pixel resolution. Most modern smartphones record at 30 Hz with a 1980x1080 pixel resolution, suggesting that the accuracy can improve using normal household devices since this would provide clearer images and less sporadic movement. Furthermore, the walking space inside the laboratory was around 5 meters, resulting in only a couple of gait cycles per trial. A larger walking space could provide more walking time, leading to a better average gait cycle per trial. Moreover, the clothing of the participants in some trials had too little contrast, resulting in the knee and hip marker being difficult to place in some frames. Additionally, the pixel length of a landmark which could be seen in the videos, was not consistent across videos and perspectives, suggesting that the camera was slightly moved between trials. This can also cause a difference between perspectives as the other camera might be a couple of pixels off for the duration of the trial.

Second of all, the DeepLabCut network was trained on 200 frames for 200.000 iterations. This could be improved as 200.000 is the suggested minimum number of iterations. Training for the maximum number of iterations instead, namely 1.300.000, could provide better estimates of label positions. Increasing the number of annotated frames could also improved the quality of the generated labels, as 500 images should achieve less pixel error [4].

Finally, post processing could be improved. The detection of errors was fully automatic and thus could not always find minor errors. The interpolation and extrapolation also only used 2 samples to fill the error, resulting in less accurate estimates. Moreover, the error detection was only done for the x-component of all limbs. Thus, a more thorough automatic error detection could also improve label estimates. Last of all, the neck label was used as ground truth for the position of the body. Thus, differences between perspectives of the neck label also translated to differences in the toe, heel, ankle, knee and hip.

The many improvements that could be made on integrating 2D pose estimation suggest that the approach could prove to be feasible. Future work could be done on the comparison of markerless integrated 2D pose estimation and marker-based motion capture techniques. Moreover, creating a custom neural network that can take both perspectives could also be a feasible approach to integration, automating this process.

5 Conclusion

I observed that integrating 2D pose estimation using DeepLabCut is not feasible using this paper's methodology. The effect of the perspectives is significantly different for the x-components of the ankle, toe, knee, hip and neck. However, the effect of the perspective was not significantly different for any of the y-components of the ankle, toe, heel, knee, hip and neck. I have identified and discussed attributes that might have influenced the accuracy of the pose estimation and proposed solutions to these problems. I am optimistic about the feasibility of integrating 2D perspectives for measuring human locomotion and expect that these methods will continue to improve.

6 References

[1] J. Stenum, C. Rossi, and R. T. Roemmich, "Two-dimensional video-based analysis of human gait using pose estimation," *PLoS computational biology*, vol. 17, no. 4, p. e1008935, 2021.

[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[3] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature neuroscience*, vol. 21, no. 9, pp. 1281–1289, 2018.

[4] T. Nath, A. Mathis, A. C. Chen, A. Patel, M. Bethge, and M. W. Mathis, "Using deeplabcut for 3d markerless pose estimation across species and behaviors," *Nature protocols*, vol. 14, no. 7, pp. 2152–2176, 2019.

[5] A. Drory, H. Li, and R. Hartley, "A learning-based markerless approach for full-body kinematics estimation in-natura from a single image," *Journal of biomechanics*, vol. 55, pp. 1–10, 2017.

[6] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*, pp. 34–50, Springer, 2016.

[7] B. Kwolek, A. Michalczuk, T. Krzeszowski, A. Switonski, H. Josinski, and K. Wojciechowski, "Calibrated and synchronized multi-view video and motion capture dataset for evaluation of gait recognition," *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 32437–32465, 2019.