

BACHELOR THESIS  
ARTIFICIAL INTELLIGENCE

**Radboud University**



---

**Markerless tracking of head movement kinematics:  
testing a sensorimotor integration model**

---

*Author:*  
Giorgia Corrado  
s1017255

*First supervisor:*  
prof. dr. W.P. Medendorp  
Donders Institute for Brain,  
Cognition and Behaviour;  
Radboud University Nijmegen  
p.medendorp@donders.ru.nl

*Second supervisor:*  
S. Willemsen, MSc  
Donders Institute for Brain,  
Cognition and Behaviour;  
Radboud University Nijmegen  
c.willemsen@donders.ru.nl

*Second reader:*  
dr. Y. Güçlütürk  
Donders Institute for Brain,  
Cognition and Behaviour;  
Radboud University Nijmegen  
y.gucluturk@donders.ru.nl



June 18, 2021

## **Abstract**

Assessing movement kinematics in naturalistic conditions is an important challenge in neuroscience. Here I studied head motion kinematics using DeepLabCut, a valuable tool for 3D markerless pose estimation that is based on deep neural networks. The input to this network are videos capturing the gait of several participants. The output consists of labeled videos in which each label represents the probability that a pixel belongs to a specific body part. I used these videos to extract head motion kinematics. By means of these results, I studied the finding that predictability of head motion can be used to determine how the brain integrates sensory and motor signals. In particular, I found peaks in the motor-to-sensory noise graph that correspond approximately to toe-off and heel strike. These peaks indicate the moments during which head movement is not well predicted. Thus, sensory signals are expected to play a more crucial role during this time frame in the gait cycle.

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Methods</b>	<b>6</b>
2.1	Dataset . . . . .	6
2.2	Labeling and Training . . . . .	6
2.3	DeepLabCut to Movement Kinematics . . . . .	7
2.4	Kinematic Analysis . . . . .	7
<b>3</b>	<b>Results</b>	<b>9</b>
<b>4</b>	<b>Discussion</b>	<b>13</b>
	<b>References</b>	<b>15</b>
<b>A</b>	<b>Appendix</b>	<b>16</b>

# Chapter 1

## Introduction

Moving through an environment is one of the most complex activities carried out by living beings. Walking is an ability that needs to be learnt by babies and needs to be restored by recovering stroke patients who cannot properly move their body anymore.

How can something perceived so natural be at the same time very complex in terms of neural control? There is a whole subfield of mechanics, kinematics, that describes motion of bodies without taking into account the forces that make them move. How does the brain control the kinematics of gait? Every time we perform an action, the brain needs to integrate sensory feedback (coming from the external sensory world) and efference copy signals (generated by an organism's motor system) into a coherent framework to control the motion (Medendorp & Heed, 2019).

If the brain relies on both sensory and motor involvement, a follow-up question would be: how much does each of them contribute during everyday actions? A formal model of sensorimotor integration for walking has been described by MacNeilage and Glasauer (2017). Their goal is to provide a model to quantify how much we rely on sensory and motor signals during locomotion. They asked ten participants to perform various activities for 6 minutes while they were wearing an IMU (Inertial Measurement Unit) on the head that measured head acceleration.

For obtaining sensory and motor noise from kinematic data they made some assumptions. Since they did not have access to the efference copy signals, they assumed that average head motion is a conservative estimate of the motor predictions. Secondly, deviation from the average head motion is due to external perturbation, intended deviation, and motor noise. They assumed that the influence of the first two components is very small thus the residual variability (deviation from the stride cycle attractor) can be estimated as

motor noise. A third point is about the sensory signals. They assumed that the only source of sensory feedback comes from the vestibular system, while in reality the motor control system also receives sensory feedback from other sources. Moreover, they assumed that noise scales with the magnitude of the signal (i.e. the system has signal-dependent noise).

Each participant’s IMU recording was divided into strides and averaged. They found that peaks in mean head acceleration over one stride in the vertical direction appeared right after heel-strike and before toe-off (**Figure 1.1**) when the foot was completely on the ground (midstance period). They then calculated the motor noise:

$$SS(t)_{res} = \frac{1}{N} \sum_{i=1} \sum_d (m(t)_{d,i} - f(t)_d)^2 \quad (1)$$

and sensory noise:

$$SS(t)_{tot} = \frac{1}{N} \sum_{i=1} \sum_d (m(t)_{d,i} - \bar{m}_d)^2 \quad (2)$$

as the deviation of head motion from the average head motion for a specific stride time and dimension ( $f(t)_d$ , also called stride-cycle attractor) and from the average head motion along one dimension ( $\bar{m}_d$ ) respectively. The ratio between  $SS(t)_{res}$  and  $SS(t)_{tot}$  gives the motor-to-sensory noise ratio

$$V(t)_{res} = SS(t)_{res}/SS(t)_{tot}. \quad (3)$$

Given that the numerator and denominator of  $V_{res}$  represent noise, high values of  $V_{res}$  indicate more reliability on the vestibular system (i.e. the motor noise has a high value) while lower values of  $V_{res}$  indicate that efference-copy signals are expected to play a more crucial role, meaning that head motion can be predicted based mainly on motor signals. Considering that  $V_{res}$  values are low during midstance and higher just outside this range (**Figure 1.2**), they concluded that right after heel strike and before toe-off we mainly rely on motor signals and in the moments of most uncertainty, when the foot is not on the ground, we rely more on sensory signals.

In contrast to physical recording devices as IMUs, in recent years deep neural networks have become very popular because of their versatility, and they turned out to be an accurate method for pose estimation (Mathis, Schneider, Lauer, & Mathis, 2020). An example of such a network is DeepLabCut, an open-source package to compute 3D pose estimations from video input data (Mathis et al., 2018). DeepLabCut uses DeeperCut’s feature detectors, which are residual neural networks with deconvolutional layers that compute the probability that a body part is in a specific location (Mathis et

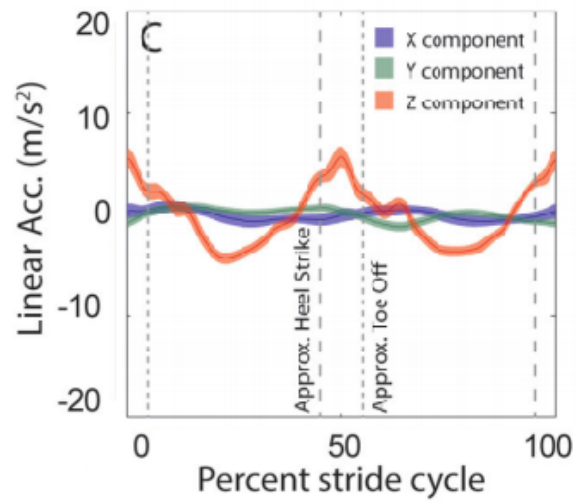


Figure 1.1: Average linear head motion during one stride: example subject (MacNeilage & Glasauer, 2017).

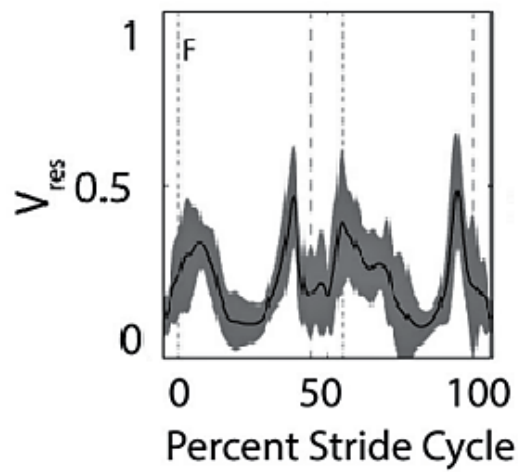


Figure 1.2: Across-subject average residual variance ( $V_{res}$ ) for linear head movement during walking (MacNeilage & Glasauer, 2017).

al., 2018). So far, this software package has been mainly applied in animal studies. In this research I use it to analyze human movement kinematics by feeding the network with recordings from the GPJATK dataset (Kwolek et al., 2019). The output of the network is a series of x-y body coordinates over frames. Thus, converting the output to movement kinematics is an important step.

The relevance of this study can be described by combining several factors. The advantage of using markerless instead of marker-based pose estimation methods to measure head motion is that in the aforementioned method the amount and location of the markers do not have to be determined prior to the experiment. Moreover, when using physical devices, the accuracy of some recorded measurements might be impaired due to calibration errors or blocking by other objects and stability problems can arise. Furthermore, the ratio between the motor and sensory noise can be interpreted as a measure of gait variability. It can explain the degree of vestibular reliance during a gait cycle and could be seen as an indicator of fall risk (MacNeilage & Glasauer, 2017). Finally, if we want to use DeepLabCut, we only need one camera to gather data, thus patients and elderly subjects can be studied more easily in their home environments.

The current study aims to use DeepLabCut to measure head motion during gait and test the sensorimotor integration model from MacNeilage and Glasauer (2017) based on these data.

# Chapter 2

## Methods

### 2.1 Dataset

This is a quantitative study using secondary data taken from the GPJATK dataset (Kwolek et al., 2019). The dataset contains 166 sequences capturing 32 participants (10 women and 22 men) while walking. In 24 sequences 6 participants wear different clothes, in 14 sequences 7 participants are wearing a backpack, and in the remaining 128 sequences the participants are all wearing their own clothes. The videos are collected using 10 motion capture cameras and 4 RGB cameras (**Figure A.1**, see appendix **A**), the latter at a frequency of 25Hz.

MacNeilage and Glasauer (2017) study three head components, x, y and z, but only the vertical, z-axis linear acceleration seems to give important insights about the predictability of head motion. For this reason, I analyzed the data recorded from the two cameras capturing the participant's side view, which contain information about the horizontal and vertical directions (cameras C1 and C3 in **Figure A.1**, see appendix **A**). This is the data DeepLabCut is trained on.

### 2.2 Labeling and Training

DeepLabCut is trained following the workflow shown in **Figure A.2** (see appendix **A**). The labeling was performed offline. The body parts of interest for this study are the head, the toe and the heel. More information on the Python commands needed throughout the workflow can be found in Table 1 in the paper by Nath et al. (2019). The labeled data was used to train the network for 200.000 iterations.

Given that MacNeilage and Glasauer (2017) show their results for one example subject as well as across subjects, I trained two networks: one fed

with the video of one participant and the other fed with videos from multiple participants. If we want to extract kinematic information as accurately as possible, then the one-participant network would be more suited because DeepLabCut analyzes only one specific subject. However, there is a risk of overfitting since the network is tested on the data it is trained on. On the other hand, training DeepLabCut on multiple videos better generalizes, i.e. it is able to label body parts of videos never seen before.

In the first network 220 frames were extracted from only one video. For the second network, 20 out of 61 sequences of different subjects from right to left were chosen and split randomly into 10 frames. The entire training and evaluation process was run in Google Colab using the notebook that can be found on the **DeepLabCut GitHub** repository. The function `analyze_videos()` returns the coordinates of the labeled body parts over frames in a CSV file.

## 2.3 DeepLabCut to Movement Kinematics

In order to extract kinematic information from the coordinates, the first step is to convert pixels to meters by calculating a scaling factor (Stenum, Rossi, & Roemmich, 2020) as :

$$s = \frac{distance}{pixel\ length}$$

I followed the analysis by Stenum et al. (2020), who measured a *distance* of 6.3 meters. The *pixel length* is taken as the horizontal pixel length between the midpoints of each strip of tape (**Figure A.3**, see appendix A). The multiplication of the scaling factor by the pixel coordinates gives the resulting *x* and *y* coordinates in meters:

$$x_{new} = x_{pixel} \cdot s$$

$$y_{new} = y_{pixel} \cdot s$$

## 2.4 Kinematic Analysis

MacNeilage and Glasauer (2017) identify a stride as two consecutive footfalls of the same foot, i.e. from the heel strike of one foot to the next heel strike of the same foot. To determine the moments of heel strike and toe-off I used two thresholds applied on the foot speed (Huitema, Hof, & Postema, 2002). As can be seen in **Figure 2.1**, the toe-off occurs when the speed of the foot is above 30% of the maximal foot velocity and heel strike when the speed drops below 35% of the maximal velocity.

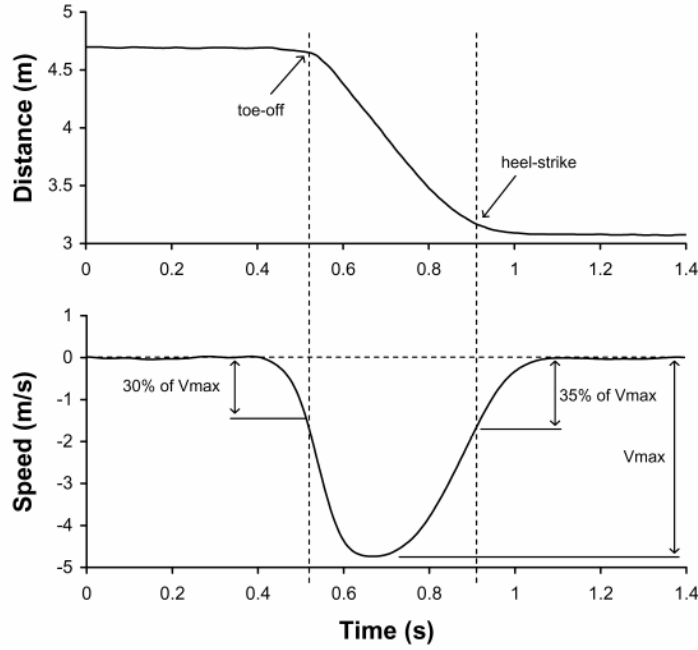


Figure 2.1: Toe-off and heel strike thresholds (Huitema et al., 2002).

The head acceleration was calculated by differentiating the head velocity graph. The latter was identified as the rate of change of the head position over frames.

At this point I identified the head motion during stride  $i$  along dimension  $d$  ( $m(t)_{di}$ ) and computed the average head motion in a specific stride time along one dimension to find the stride-cycle attractor ( $f(t)_d$ ). Another important quantity is the average head motion along a specific dimension across all strides and times ( $\bar{m}_d$ ). From these quantities I derived  $SS_{res}$  (Eq. 1) and  $SS_{tot}$  (Eq. 2). The ratio between the last two measures gives the motor-to-sensory noise ratio ( $V(t)_{res}$ ), Eq. 3) which is plotted in Figure 3.5 across frames.

It is worth mentioning that DeepLabCut was trained by labeling only one side of the body, thus the other foot position, i.e. "Right toe off", was estimated based on an approximate division of the gait cycle in 6 phases (Figure A.4, see Appendix A).

## Chapter 3

# Results

The moments of heel strike and toe-off are plotted on top of the head acceleration graph (**Figure 3.1**). I split the gait cycle in three parts (from one red line to the next) to extract single strides. Subsequently, I interpolated the strides to make them all equal length and time normalized (**Figure 3.2**). Average linear head acceleration across the three strides is visualized in **Figure 3.3**. This graph shows the mean head acceleration from one subject's recording used to train the first network. The same analysis can be done across multiple participants (using the second trained network). The result is shown in **Figure 3.4**.

The average linear head acceleration (stride-cycle attractor) is used to calculate the predictability of head motion ( $V_{res}$ , **Figure 3.5**) using equations 1, 2, and 3 formulated in the Introduction.

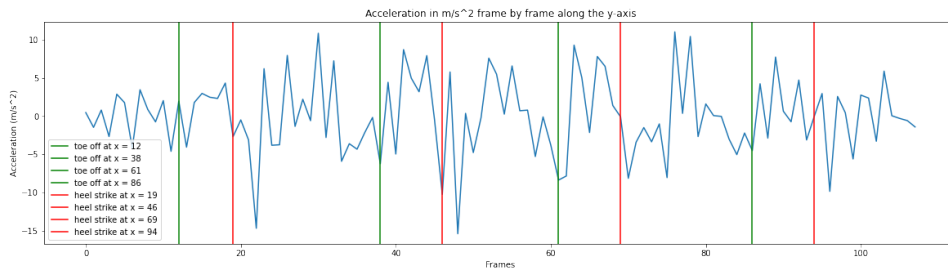


Figure 3.1: Head acceleration ( $m/s^2$ ) along the y-axis.

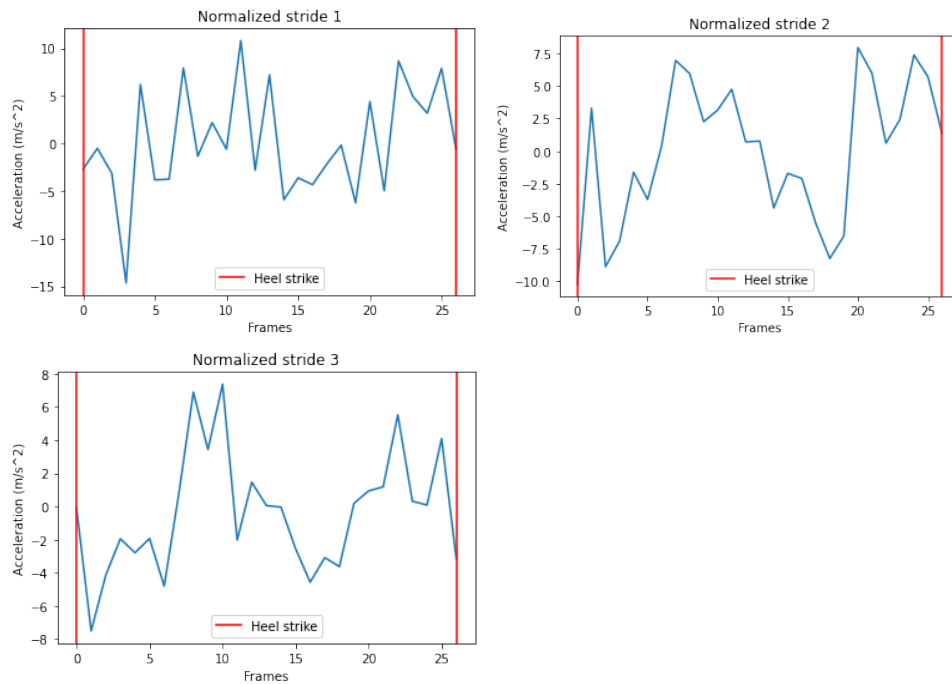


Figure 3.2: Stride data resampled to a total of 27 frames for three different strides of one participant via interpolation in Python. (The method used is `interp1d()` from `scipy.interpolate`.)

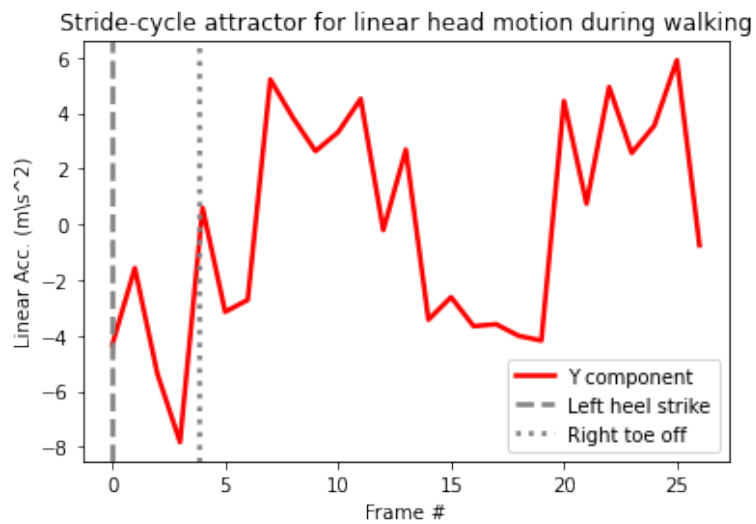


Figure 3.3: Average linear head acceleration during one stride: same subject as Figure 3.2.

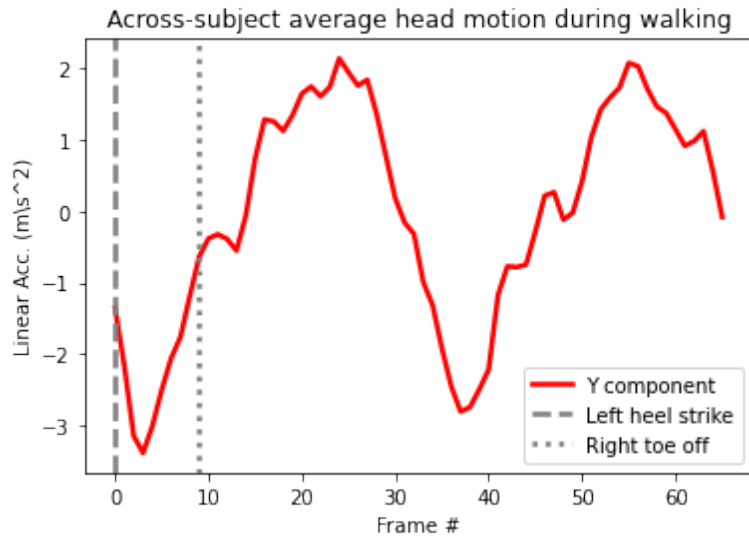


Figure 3.4: Across-subject average stride cycle during walking normalized to 66 frames.

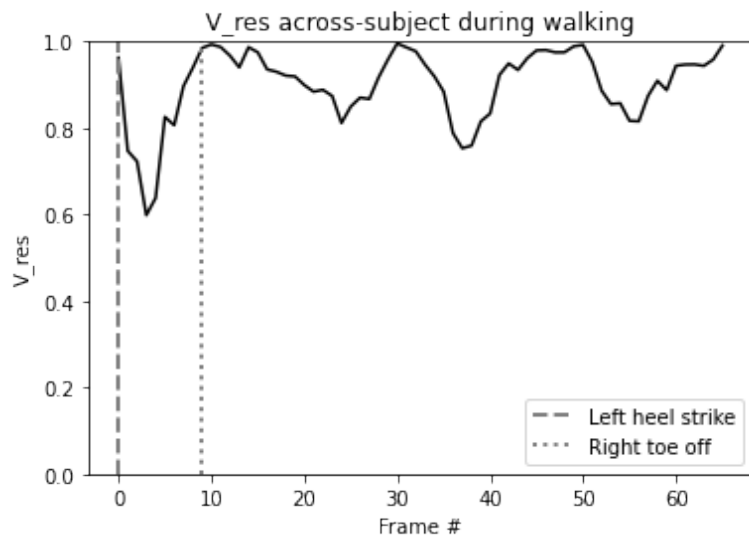


Figure 3.5: Average proportion of residual variance during walking: across-subject.

### Statistical Analysis

In **Figure 3.5** it can be noticed that there is a drop at around frame number 4. Is it significant? To assess this question I ran a T-test by comparing all the drops with each other and the null-hypothesis being that their pair-

wise difference is equal to 0. I determined four distributions by taking samples before and after the lowest point of the drops and assumed that they are normally distributed. I found that the only p-values less or equal to 0.05, were the ones resulting from the T-tests run by comparing the first distribution with the others ( $p = 0.002$ ,  $p = 0.019$ ,  $p = 0.05$ ). The differences between the other pair-wise T-tests, not including the first one, were not significant (p-values bigger than 0.05).

## Chapter 4

# Discussion

DeepLabCut can indeed be used to measure head motion kinematics and further test the sensorimotor weighting model of MacNeilage and Glasauer (2017).

Earlier research has found that that the head acceleration is maximal during midstance, i.e. between heel strike and toe-off (Mulavara & Bloomberg, 2002). MacNeilage and Glasauer (2017) used this assumption to determine heel strike and toe-off from the IMU data (**Figure 1.1**). On the contrary, I used two thresholds applied on the foot speed (Huitema et al., 2002) to determine these two quantities and showed that the head acceleration is indeed maximal between heel-strike and toe-off as can be seen in **Figure 3.3** and **3.4**.

**Figure 3.5** can be compared to **Figure 1.2** where the motor-to-sensory noise ratio can be interpreted as a measure of vestibular reliance during walking. From the graph two peaks can be identified at the moments of heel strike and toe-off. Given that  $V_{res}$  represents the ratio between motor and sensory noise ( $V_{res} = SS_{res}/SS_{tot}$ ), a higher value on the y-axis means that vestibular signals are expected to play a more critical role. By contrast, when there is a significant drop (during midstance) we rely much more on motor predictions, and head motion can be predicted based mainly on efference-copy signals.

The study has some limitations. When comparing the plots showed in this study with those produced by MacNeilage and Glasauer (2017), it is clear that the video-based kinematics graphs are much more noisy. In particular, the significant drop in **Figure 3.5** is roughly at 0.6 on the y-axis while in **Figure 1.2** the drop is around at 0.2. Moreover, the peaks at heel strike and toe-off in **Figure 3.5** are at 1, while in **Figure 1.2** appear approximately at 0.6. There are two main reasons for why this is the case. The first is

that they measure the head acceleration using an IMU which transmitted 3-degrees-of-freedom (DOF) measurements of linear acceleration, orientation, and angular velocity at 150 Hz sampling frequency. Their analysis delivers a linear head acceleration over 3 axis, i.e.  $x$ ,  $y$ , and  $z$ . DeepLabCut, instead, was trained on 2D videos recorded at a sampling frequency of 25 Hz. Having a lower resolution makes the data less smooth and less precise. The second reason is that MacNeilage and Glasauer (2017) used much more data. They recorded the performance of participants performing activities for 6 minutes, while in my case each subject was walking 5 seconds on average, meaning that fewer strides could be extracted.

In conclusion, the study showed how DeepLabCut was successful in tracking head movement kinematics and further test the sensorimotor weighting model by MacNeilage and Glasauer (2017). In future studies, it could be used to investigate which kinematic strategies humans use when they move body parts other than the head.

# References

- Huitema, R., Hof, A., & Postema, K. (2002). Ultrasonic motion analysis system—measurement of temporal and spatial gait parameters. *Journal of biomechanics*, *35* 6, 837-42.
- Kim, J.-J., Kim, H., Lee, C., & Kim, J.-Y. (2018). Power assistance and evaluation of an end-effector typed walking rehabilitation robot. *2018 18th International Conference on Control, Automation and Systems (ICCAS)*, 1353-1355.
- Kwolek, B., Michalczuk, A., Krzeszowski, T., Switonski, A., Josinski, H., & Wojciechowski, K. (2019). Calibrated and synchronized multi-view video and motion capture dataset for evaluation of gait recognition. *Multimedia Tools and Applications*, *78*, 32437 - 32465.
- MacNeilage, P., & Glasauer, S. (2017). Quantification of head movement predictability and implications for suppression of vestibular input during locomotion. *Frontiers in Computational Neuroscience*, *11*.
- Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V., Mathis, M., & Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*, 1281-1289.
- Mathis, A., Schneider, S., Lauer, J., & Mathis, M. (2020). A primer on motion capture with deep learning: Principles, pitfalls, and perspectives. *Neuron*, *108*, 44-65.
- Medendorp, W., & Heed, T. (2019). State estimation in posterior parietal cortex: Distinct poles of environmental and bodily states. *Progress in neurobiology*, 101691.
- Mulavara, A., & Bloomberg, J. (2002). Identifying head-trunk and lower limb contributions to gaze stabilization during locomotion. *Journal of vestibular research : equilibrium & orientation*, *12* 5-6, 255-69.
- Nath, T., Mathis, A., Chen, A., Patel, A., Bethge, M., & Mathis, M. (2019). Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nature Protocols*, *14*, 2152-2176.
- Stenum, J., Rossi, C., & Roemmich, R. T. (2020). Two-dimensional video-based analysis of human gait using pose estimation. *bioRxiv*.

# Appendix A

# Appendix

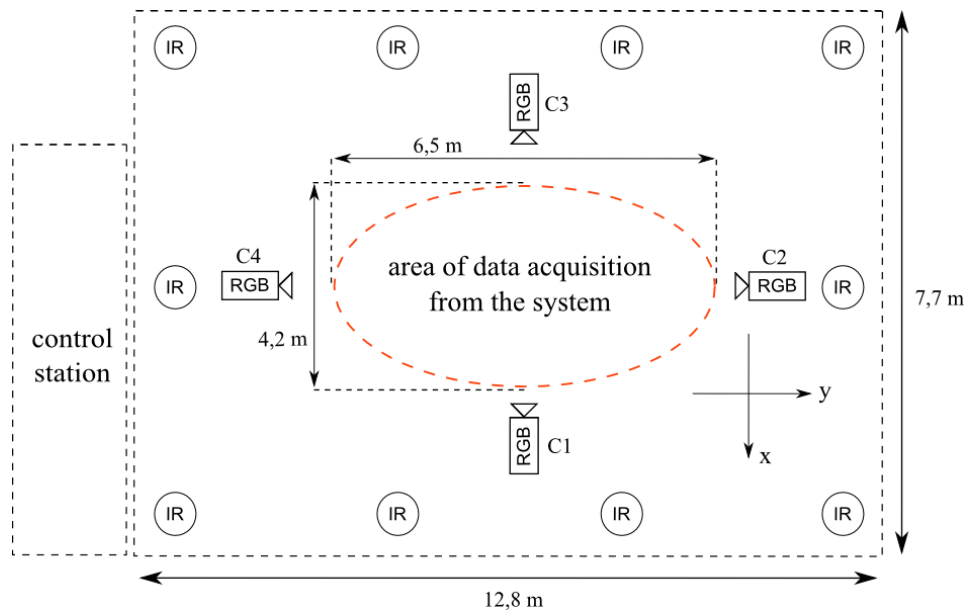


Figure A.1: Camera layout. IR = infrared camera used to track the moCap markers. RGB = RGB camera (Kwolek et al., 2019).

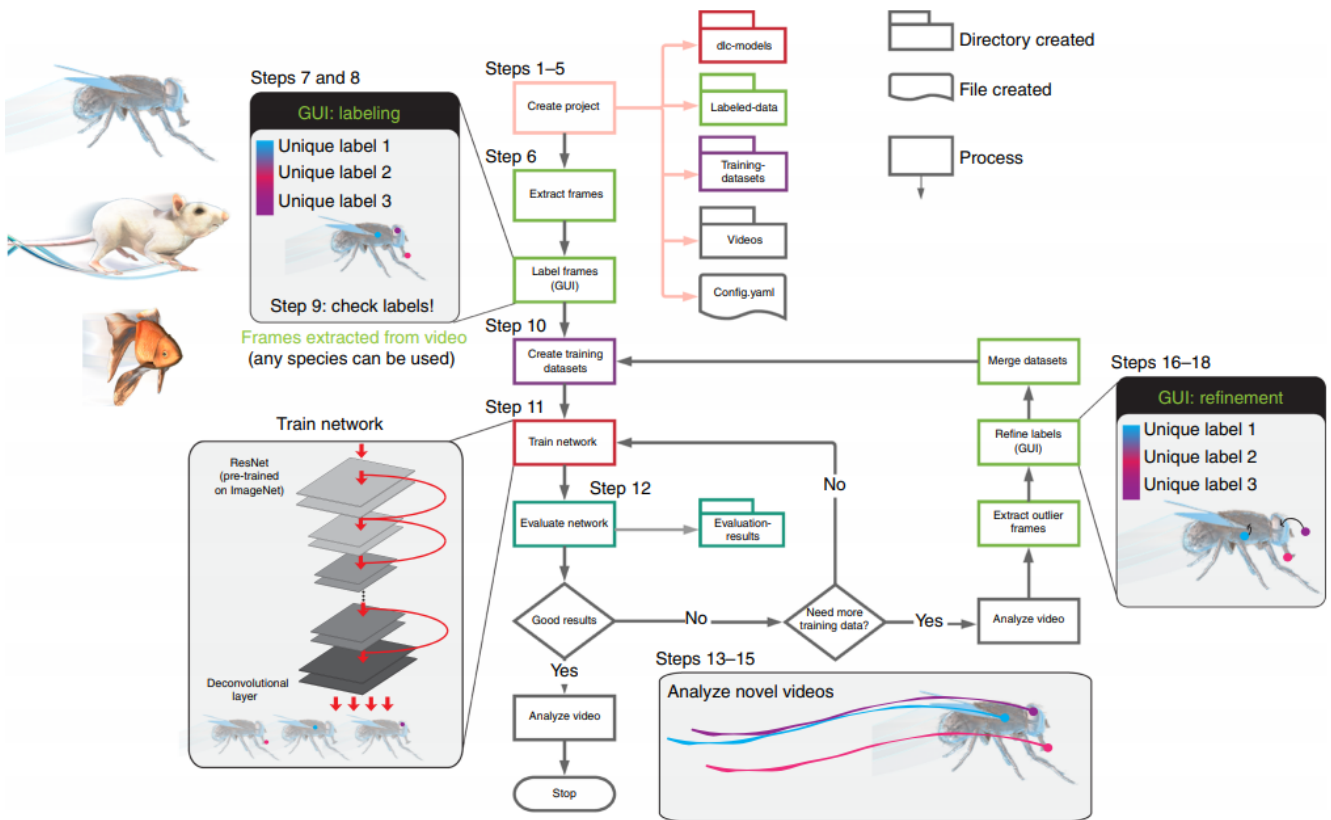


Figure A.2: DeepLabCut workflow (Nath et al., 2019).



Figure A.3: Recording from camera 1 (C1). Distance between the red crosses: 6.3m. Screenshot taken from one of the videos in the GPJATK dataset (Kwolek et al., 2019).

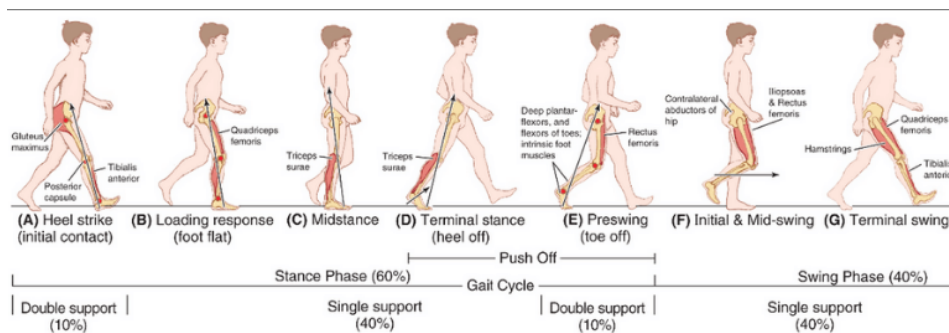


Figure A.4: Gait cycle (Kim, Kim, Lee, & Kim, 2018).