Group Project - Walking Analysis

Group Ace

Hongrui Qu | 301331940
Yuting Yang | 301417395
Choo Carmen | 301573484

CMPT 353– D100 | Computational Data Science
Instructor: Greg Baker

December 9th, 2022

Abstract

Even though gait analysis is an important field for learning the overall movement of the human body, many people have asymmetrical gaits for years without knowing it. This report will introduce a method for analyzing an individual's gait pattern and determine wherever the pattern is symmetric.

Problem Analysis

Before this project, none of the team members questioned about their own walking patterns, which brought deep interest in the topic of gait pattern analysis. To further extend on the topic of git pattern analysis, we also developed an interest in the difference between hand movement and leg movement. Therefore we focus on three questions:

1. Does the two legs' movement symmetric during walking? Or in another word, does the left leg move at a different scale compared to the right leg?
2. If someone got injured, does our method successfully determine the asymmetric of the movement of the legs?
3. Does the hand and leg from the same side move at different scales during walking?

Data Collection

The setting of our data collection is the third floor of West Mall Center at SFU Burnaby campus, where the floor is flat without any slope. We also try our best to walk in straight lines and avoid turning left or right. In total, we collect 30 normal datasets and 7 injury datasets. However, when we investigate the data, we notice two of the right foot datasets are missing so we delete the datasets for trial 16 and trial 36. Therefore the valid number of datasets is 29 normal datasets and 7 injury datasets. For software, we choose to use the linear accelerometer of Physics Toolbox Sensor Suite, which collects the movement data in three dimensions.

The procedures:

1. Download the Physics Toolbox Sensor Suite on all of our phones and familiarize ourselves with working with it.

2. Borrowed a walking stick to help simulate the injury walking data.
3. Go to the setting we choose and make sure there are not too many people to avoid interruption during the data collection process. We collected the data starting at 9:40 am and thankfully there were not too many students in West Mall Center.
4. Start near the end of West Mall Center and leave a water bottle to mark it as the starting point of our data collection.
5. We use two phone running bands to tie our phone to both the left ankles and right ankles of the individual who takes the task.
6. Open Physics Toolbox Sensor Suite and select the “linear accelerometer” function.

For regular data collection:

7. Ask the individual to hold the phone in her hands during walking.

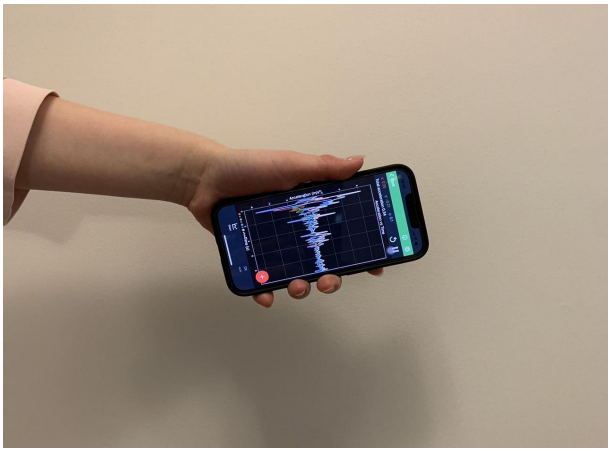
For injury data collection

7. Ask the individual to hold the walking stick and simulate being injured on one of her feet.

Since one of her hands holds the walking stick, we did not collect the hand data for injury data.

8. Set the timer for 70 seconds, which rang after the timing finished.
9. Try to start all of the physics toolbox sensors on the phones tied to the left ankle, the right ankle, the hand, and the timer.
10. After the timer rings, try to tap the stop button on all of the sensors.
11. Naming the dataset and outputting the dataset for later use.

Here are pictures of how we tied the phone to collect the data:



The numbers contained by the name of our datasets are the same as walking trial numbers. Trials 1-10 are collected by Yuting. Trials 11-20 are collected by our friend, Wenyu. Trials 23-29 are the simulated injury data collected by Hongrui. Trials 21,22 and 30-38 are the normal walking data collected by Hongrui. Additionally, for trials 1 to 20, the hand data is collected from the right hand. For trials 21, 22 and 30-38, the hand data is collected from the left hand.

Data Explanation

Explanation of data file names:

The names of data files from the legs of normal trials contain the word ‘regular’, such as “sensor01L_regular.csv”. The names of data files from the legs of simulated injury trials contain the word “injury”, such as “sensor23L_injury”. The names of data files from the hands contain the word “hand”, such as “hand01R.csv”. As the examples mentioned above, all of our data files contain either “L” or “R” , while “L” means left side of the body and “R” means right side of the body.

By using the app, all of our data created the same five columns which are “time”, “ax (m/s^2)”, “ay (m/s^2)”, “az (m/s^2)” and “aT (m/s^2)”. “ax” shows the movement of left-right dimensions. “Ay” shows the movement of back-forth dimensions and “az” shows up-down dimensions.
We did not actually use the “at” which means the total calculation of the acceleration which is not the interest of this project.

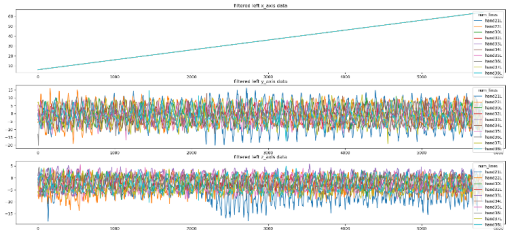
Data Cleaning

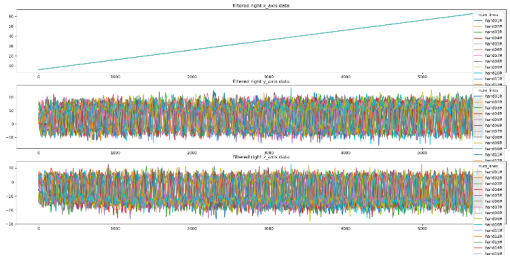
Before using the data for analysis, we first investigate the datasets. We notice that dataset 16 of the right foot is missing, therefore we delete all left foot and hand datasets for trial 16 since they can not be used for analysis. Then we decide to remove the first 4 seconds and last 6 seconds of the data since the sensors are not started or stopped at the exact same time and therefore will increase the bias of data. We also align all datasets to have the same rows of data. As a result, we keep 5799 rows of data and 102 columns.

Considering the large size of the data frame, we decided to work with Sparks. After extracting data from all the datasets and combining them into a Spark data frame, we notice there are some missing values in the rows. Luckily the missing value is less than 3 percent and therefore we simply remove all the missing values. After formatting the data to x y z data frames, we employ a Butterworth filter to eliminate the noise of phone sensors.

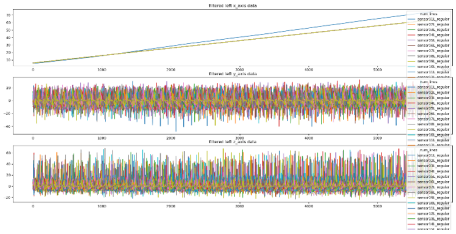
We manually tested the frequency from 0.03 to 0.9 and selected the number that seems to filter the noise while not overfitting or underfitting. The graphs below show the data before and after applying the filter. An order of 3 and a frequency of 0.12 is used for a_x data. An order of 3 and a frequency of 0.07 is used for a_y data. An order of 3 and a frequency of 0.17 is used for a_z data.

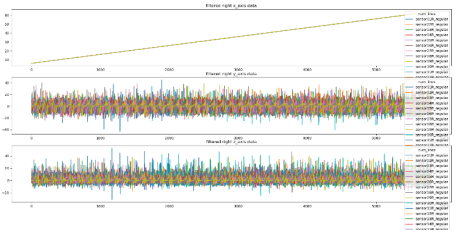
The plot of hand data on all datasets:



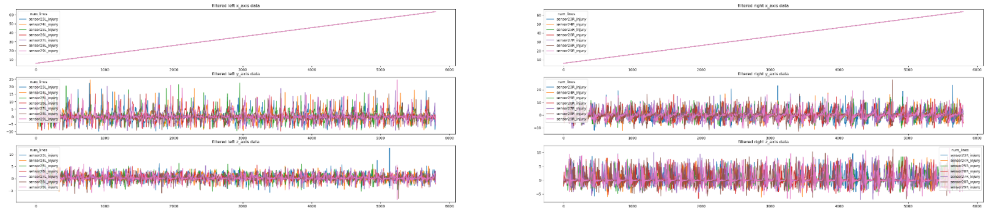


The plot of normal data on all datasets:

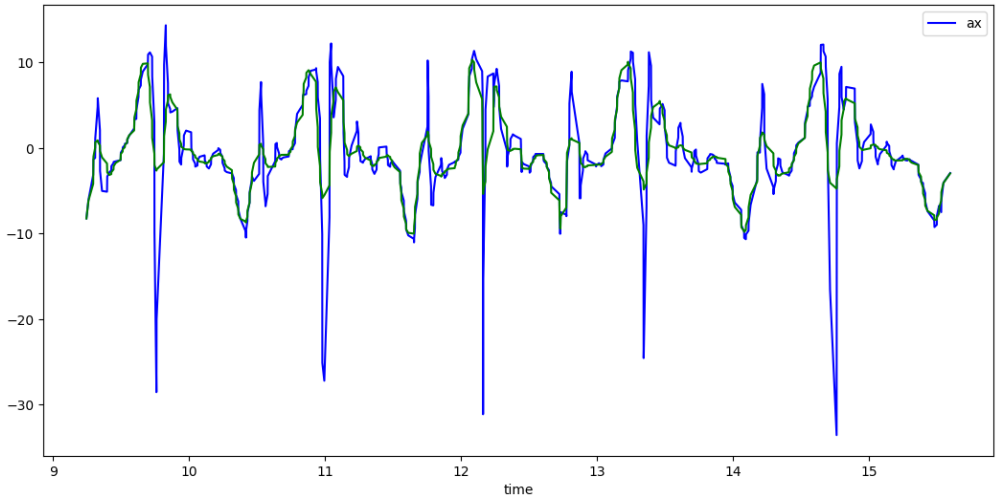




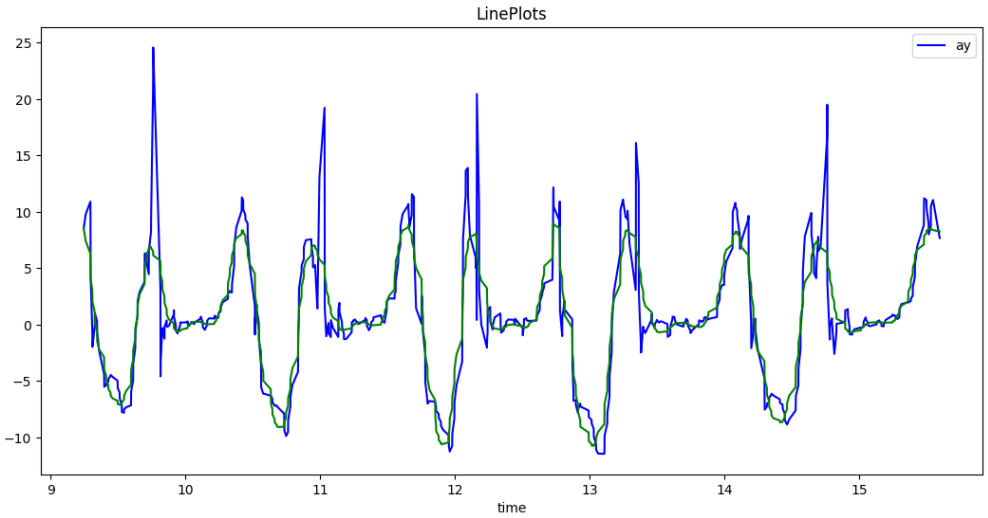
The plot of injury data on all datasets:



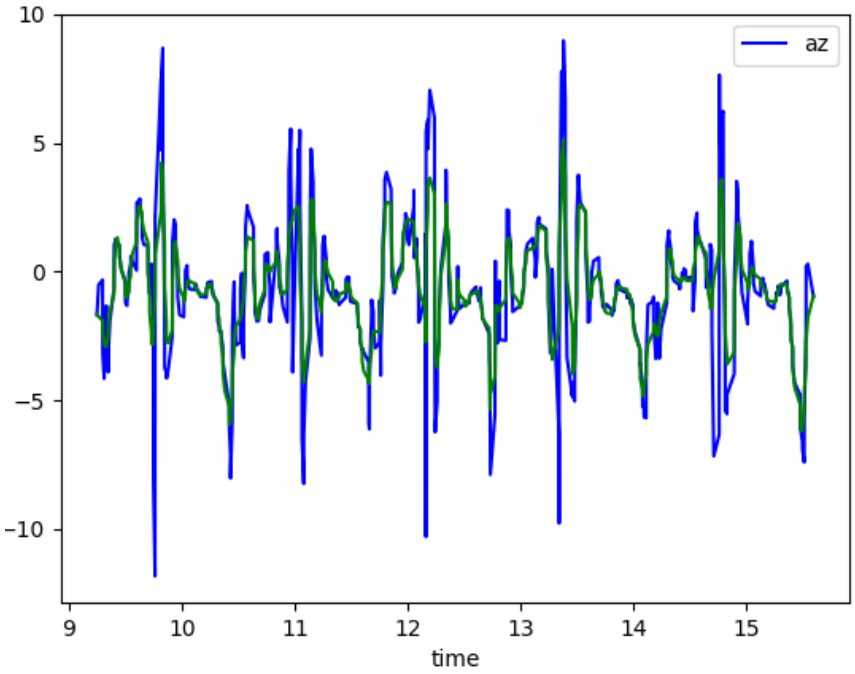
While the plots on all datasets show more information it is overwhelming for the explanation. Therefore we include the plot for 6 seconds for one dataset as below to help explain.



Blue line represents: Ax before applying filter
Green line represents : Ax after applying butterworth filter



Blue line represents : Ay before applying filter
Green line represents : Ay after applying butterworth filter



Blue line represents : Az before applying filter
Green line represents : Az after applying butterworth filter

Data Analysis

After we cleaned the data, we used the statistic method to examine the datasets. Since only one foot leaves the ground at a time during walking, we can not compare the overall data pattern between the left foot and right foot. Instead, we decide to compare the variance of our data to determine if the foot or hand moves at a similar distance scale during walking. However, when we investigate the datasets, we notice a certain time range of data points has a significantly large variance compared to other time ranges. Therefore instead of calculating a single variance for all data points for the same dataset, we decide to split the dataset into a 1.5 seconds time range. We choose the number 1.5 seconds because it is relatively similar to the period of the x y z data. Then for each dataset, we have 40 variances. According to the central limit theorem, and because these datasets are not too far from being normal, we can use the tests that assume normality. We choose to use the t-test to compare our variances.

For question 1: The null hypothesis is that the variance of the left foot and the right foot is the same. The alternative hypothesis is that the variance of the left foot and the right foot is not the same.

For question 2:
The null hypothesis is that the variance of the left injured foot and the right injured foot is the same. The alternative hypothesis is that the variance of the left injured foot and the right injured foot is not the same.

For question 3:
For hand data collected from left hand,
The null hypothesis is that the variance of the left hand data and variance of the left foot is the same. The alternative hypothesis is that the variance of the left hand data and variance of the left foot is not the same.
For hand data collected from the right hand,
The null hypothesis is that the variance of the right hand data and variance of the left foot is the same. The alternative hypothesis is that the variance of the right hand data and variance of the left foot is not the same.

Result

The python program helps us to output a CSV file containing all the p-values for each of our trials. We choose the confidence interval as 0.05, which means the p-value < 0.05 will reject the null hypothesis so we draw a conclusion of the alternative hypothesis as mentioned early. If the p-value > 0.05, we fail to reject the null hypothesis and will draw a conclusion as the null hypothesis.

For datasets 1-10, which are the normal walking data for Yuting. 10 out of 10 p-values for y and z dimensions are smaller than 0.05. This means that Yuting has a different variance for the left foot and right foot. These statistical results provide evidence to conclude that Yuting's y and z dimension of the left foot and right foot moves at different scales and therefore Yuting' two leg movements are NOT symmetric during walking.

For datasets 11-20 (no dataset 16), which are the normal walking data for Wenyu. 9 out of 9 p-values for y and z dimensions are smaller than 0.05. This means that Wenyu has a different variance for the left foot and right foot. These statistical results also provide evidence to conclude that Wenyu's y and z dimension of the left foot and right foot moves at different scales and therefore Wenyu' two leg movements are also NOT symmetric during walking.

For datasets 23-29, which are the simulated injury data. 7/7 p-values for z dimension and 6/7 p-values for y dimension are smaller than 0.05. While 1/10 p-value for y dimension is bigger than 0.05. We believe the 9/10 p-value still provides sufficient evidence to include that the y and z dimension of the left foot and right foot moves at different scales. Therefore our method successfully determines the asymmetric of the movement of the legs for an injured person.

For datasets 21, 22, and 30-38(no dataset 36), which are the normal walking data for Hongrui. 5 out of 10 p-values for z-dimension and 2/10 p-values for y-dimension are smaller than 0.05. We think this provides moderate statistical evidence to conclude that Hongrui's z and y dimension of the left foot and right foot moves at different scales therefore Hongrui's two leg movements are also NOT symmetric during walking. However, we are less confident in this conclusion.

For all of the x-dimension p-values, only 2 results are smaller than 0.05. The reason is most likely that we walk in straight lines and therefore not move left and right for both our feet and hand. However, this is reasonable so the x-dimension p-value is not necessary for drawing conclusions.

Limitation

Firstly, the degree of freedom for our p-value is 38 (since we have 40 variances and use a two-sample t-test). Also for Yuting, 10 datasets were analyzed. For Wenyu 9 datasets were analyzed. For Hongrui's normal walking, 10 datasets were analyzed. The degree of freedom and number of datasets for each individual provides moderate confidence for our results. While more data points and more datasets will provide more confidence for our analysis, we did not collect more datasets considering the time cost.

Secondly, the injury data was simulated by our team members. We try our best to simulate by using a walking stick, however, this still reduces the reliability of the data and reduces the reliability of our results too. Additionally, only 7 datasets were collected on injury data, which provides less confidence for the results on question 2 as well.

Thirdly, we consider the psychological effect. Since we are walking on the purpose of collecting data instead of walking inattentively. Some of us are too nervous to walk and started to walk like robots, which negatively impacts the reliability of our results.

Fourthly, the validation of the results heavily depends on the Physics Toolbox Sensor, which we used to collect data.

Accomplishment

All of the team members contribute equally to this project. All three of us gathered together to write the code and report for this project. While Hongrui and Yuting collected the dataset, Carmen is not responsible for data collection since she was sick on the day of data collection. Additionally, we would like to thank our friend, Wenyu, who helps us with data collection.