Report 2: Detect Human Activities

Seyed Mohammad Mehdi Hosseini, s301769,
ICT for Health attended in A.Y. 2022/23

January 17th 2023

1 Introduction

The objective of this study is to design and evaluate a system for automatic recognition of human activity with the minimum number of devices and computational resources available. The system has been collected sensor data from multiple locations on the body, including the torso (T), right arm (RA), left arm (LA), right leg (RL), and left leg (LL), at a frequency of 25 Hz, using 3-axis accelerometers, gyroscopes, and magnetometers. The data collected results in 45 values per sample (sensor features). This dataset is available publicly at [1]

The activities dataset includes 8 participants in total, each of whom performed a set of 19 activities for 5 minutes each. In this research, we are going to specifically focus on the subject number 2 activities. The data is then divided into 5-second segments, resulting in files with 125 rows and 45 columns. The proposed system uses a combination of data pre-processing algorithms and K-Means clustering technique to analyze the sensor data and classify the performed activity with high accuracy using clustering algorithms.

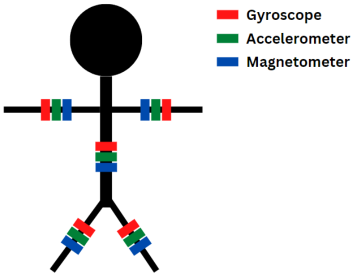


Figure 1: Sensors placement on subject's body

2 Activities Dataset

In achieve the available dataset, there are three types of sensors (Accelerometer, Gyroscope and Magnetometer) have been installed on Torso, Right and Left Arm and Right and Left

Leg of subject's body. Figure. 1 shows how the sensors are installed on subject's body. Here is a brief description about each type of sensors utilized in data collection:

An accelerometer is a device that measures linear acceleration, which is the rate of change of velocity over time. Accelerometers typically measure acceleration in one, two, or three dimensions, depending on the design of the device which is only in one dimension (x-axis, y-axis and z-axis) in our research. In provided dataset, the accelerometer features can be found through this naming convention: [T_(x,y,z)mag, RA_(x,y,z)mag, LA_(x,y,z)mag, RL_(x,y,z)mag, LL_(x,y,z)mag]

A gyroscope is a device that measures angular velocity, which is the rate of change of angular displacement over time. Gyroscopes typically measure acceleration in one, two, or three dimensions, depending on the design of the device which is only in in one dimension (x-axis, y-axis and z-axis) in this research. In provided dataset, the accelerometer features can be found through this naming convention: [T_(x,y,z)gyro, RA_(x,y,z)gyro, LA_(x,y,z)gyro, RL_(x,y,z)gyro, LL_(x,y,z)gyro]

A magnetometer is a device that measures the strength and direction of a magnetic field. It can be used to determine the orientation of a device with respect to Earth's magnetic field. Magnetometers typically measure acceleration in one, two, or three dimensions, depending on the design of the device which is only in one dimension (x-axis, y-axis and z-axis) in this research. In provided dataset, the accelerometer features can be found through this naming convention: [T_(x,y,z)acc, RA_(x,y,z)acc, LA_(x,y,z)acc, RL_(x,y,z)acc, LL_(x,y,z)acc]

3 Data Preprocessing

Before going to trial and error with the training data, first it's appropriate to evaluate each type of sensor performance in terms of comparing the distance of the centroid of each activity with the standard deviation of the measures by that sensors. Comparing the results in Figure 2. the accelerometer measurements has large value of standard deviation for each activity. The distance between the considered centroid (mean of values) and mean distance from points to centroids for siting, standing, lying it completely acceptable but on the contrary, the rest of activities have a large distance so in further analysis, these activities can not be easily detected. Considering the pure gyroscope measurements, all of the centroids are near to the zero, measurements variations are considerably lower than the accelerometer data but as long as the variations are high, there can not be a good setup to use gyroscope sensors. Magnetometer measurements has a significant better results, comparing to the other sensors. There are a huge part of sensors that have minimum centroid distance higher than the mean distance from points to centroid. It should be noted that these are the raw measurement and it following sections, the results will be improved using data processing techniques.

3.1 Rolling Mean

Rolling mean is a technique used in signal processing and time series analysis that involves calculating the mean of a set of observations over a fixed period of time, and then updating this mean as new observations become available. This technique is useful for smoothing out short-term fluctuations in data. By taking the mean of a set of observations, the rolling mean effectively filters out high-frequency noise and random variation, allowing for a clearer view of the underlying patterns in the data. Figure. 3 shows the results of using rolling mean with window size of 50. This value has been found through trials and errors with different window sizes. Note that the window sizes can be lower or higher, for example it can be 25 (the maximum number of samples in each second) but it passes more high frequency noises through data. Higher values of window size also lead to mean value of the whole measurements during the time.

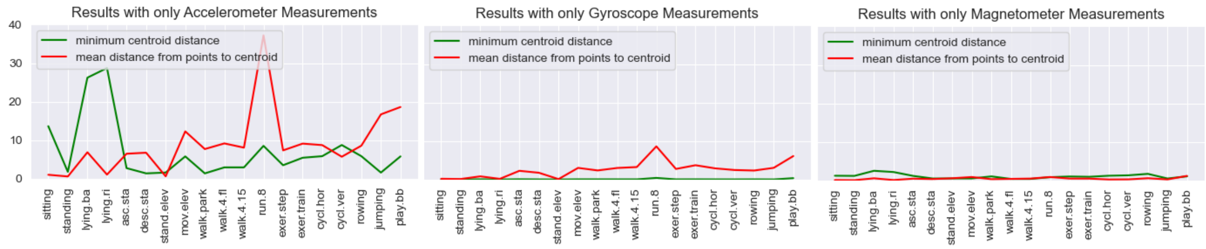


Figure 2: Comparing Minimum Centroid Distance with Mean of the distance between points and centroids

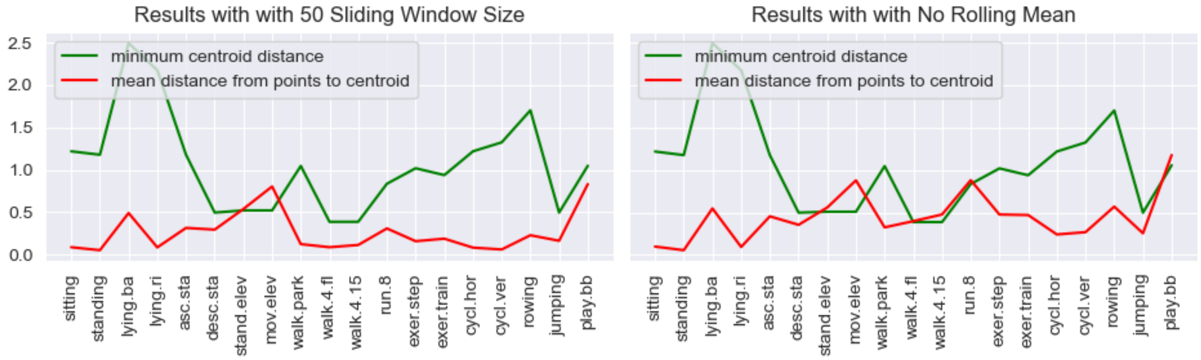


Figure 3: Improvements using rolling mean with size of 50

$$X_{m,N} = \frac{X_m - \mu_m}{\sigma_m} \tag{1}$$

4 K-Means Clustering

K-Means Clustering is a widely used unsupervised machine learning technique for grouping similar observations together based on their feature values. The algorithm partitions a dataset into K clusters, where K is a user-specified parameter, by iteratively assigning each observation to the cluster with the closest mean, which is calculated as the centroid of the observations in the cluster. The process is repeated until the cluster assignments no longer change or a stopping criterion is met. To do this we are benefiting from Scikit-Learn package to implement the algorithm. There are several hyperparameters to set before executing the algorithm. we have set the $'ninit'$ to $auto$, $'init'$ method to $'kmeans++'$ initialization which is based on an empirical probability and $'algorithm'$ to $lloyd$ algorithm.

4.1 K-Means Result Mapping with Actual Data

The most frequent label approach, also known as the majority voting method, is a technique used to map the results of K-Means clustering to actual labels. The most frequent label approach involves assigning each cluster the label that is most commonly found among the observations in that cluster. This is done by counting the number of observations in each cluster that belong to each label, and then selecting the label with the highest count as the cluster label. This approach assumes that the observations in a cluster are more similar to each other than to observations in other clusters, and therefore the most frequent label among the observations in a cluster is likely to be the correct label.

5 Comparison and Hyperparameter Tuning

5.1 Comparison Criteria

In order to compare different models during the training and tuning phase, we are using the mean of all accuracies among all activities (Equation 2.) where n is the number of activities, y_i is the true label of the $i-th$ observation, $\hat{y_i}$ is the predicted label of the i-th observation and $P(y_i|\hat{y_i})$ is the probability that the i-th observation is correctly classified.

$$Acc = \frac{1}{n}\sum_{i=1}^{n} P(y_i|\hat{y_i}) \tag{2}$$

5.2 Feature Selection

After finding the Magnetometer sensors as the best choice to follow, we have utilized a greedy search to find the best 5 features with highest training accuracies among all 15 Magnetometers features to have the least number of features. Table. 1 lists the top 3 results among all possible combinations:

List of Features	Train Accuracy	Test Accuracy
$('T_ymag',' RA_xmag',' LA_ymag',' LA_zmag',' LL_ymag')$	0.884	0.795
$('RA_xmag',' T_zmag',' LA_ymag',' LA_zmag',' LL_ymag')$	0.873	0.776
$('T_ymag',' RA_xmag',' LA_ymag',' LA_zmag',' RL_zmag')$	0.852	0.769

Table 1: Results with five sensors

As the test accuracies are not sufficient and are too low, another greedy search deployed to find two best sensors among accelerometers to add to this current sensor setup. The results show that by adding $'RL_zacc','RL_xacc'$ features, the accuracies raised up to 0.887 and 0.847 for train and test data respectively.

5.3 Training Slices

There are 60 slices of data in total for each activity. It is possible to consider the 30 slices as the maximum possible number of slices for the training and the rest for test. Table. 2 shows the results for three different training slice numbers. It can be conceived that high the number of training slices, the model is going to overfit on the training data and the low number of training slices will lead to underfit. The best choice among these options is 20 slices for training. By using 20 slices as training, Figure. 4 shows the confusion matrix for training and test predictions. Also, Table. 3 show the final results for each activity in the test data set. Figure 5 shows the final comparison between mean of centroids and mean of points distances to the their associated centroids.

Number of Training Slices	Train Accuracy	Test Accuracy
30	0.844	0.805
20	0.887	0.847
10	0.902	0.769

Table 2: Results with different number of training slices

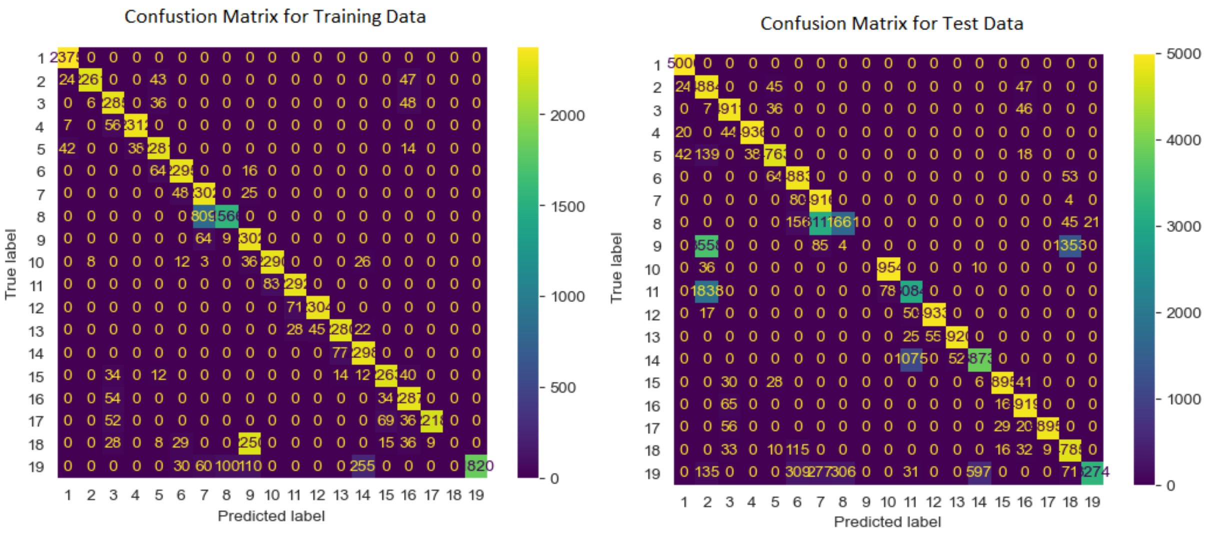


Figure 4: Confusion Matrix for Training and Test Data

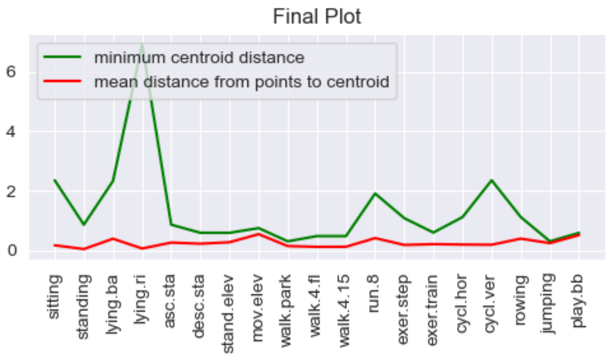
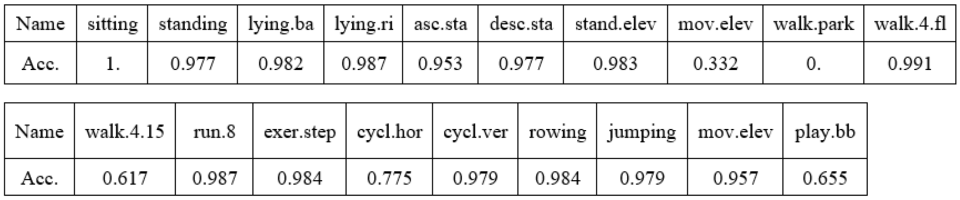


Figure 5: Comparing Minimum Centroid Distance with Mean of the distance between points and centroids Using Final Sensor Setup



Name	sitting	standing	lying.ba	lying.ri	asc.sta	desc.sta	stand.elev	mov.elev	walk.park	walk.4.fl
Acc.	1.	0.977	0.982	0.987	0.953	0.977	0.983	0.332	0.	0.991

Name	walk.4.15	run.8	exer.step	cycl.hor	cycl.ver	rowing	jumping	mov.elev	play.bb
Acc.	0.617	0.987	0.984	0.775	0.979	0.984	0.979	0.957	0.655

Table 3: Accuracy for each activity on test data set

6 Conclusions

In the current study, data pre-processing and hyperparameter tuning were conducted to determine the optimal set of sensors and features for predicting human activities. The final selection consisted of 5 sensors and 7 features, specifically ['RL_zacc', 'RL_xacc', 'RA_xmag', 'T_zmag', 'LA_ymag', 'LA_zmag', 'LL_ymag']. It is worth noting that the feature 'LL_ymag' could be replaced with 'LA_xmag', which would reduce the number of devices required while still maintaining a high level of accuracy. The goal of these efforts was to minimize computation costs while utilizing an appropriate number of features and algorithms for data processing. However, it should be noted that the 'walk.park' activity was not able to be predicted due to confusion with 'standing' and 'jumping' activities. Additionally, 'mov.elev' activity was also found to be confused with 'stand.elev'. Despite these limitations, the results obtained using the K-Means algorithm were promising.

References

[1] https://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities