# Introducing a Human Activity Recognition Dataset Gathered on Real-Life Conditions

Daniel Garcia-Gonzalez, Enrique Fernandez-Blanco, Daniel Rivero, and Miguel R. Luaces

Department of Computer Science and Information Technologies, University of A Coruna, CITIC, 15071 A Coruna, Spain
Correspondence: d.garcia2@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.08*

*Abstract*: Human activity recognition (HAR) has garnered significant scientific interest in recent years. The widespread use of smartphones enabled convenient and cost-effective data collection, eliminating the need for additional wearables. Given that, this paper introduces a novel HAR dataset in which participants had freedom in choosing smartphone orientation and placement during activities, ensuring data variability. It also includes contributions from diverse individuals, reflecting unique smartphone usage habits. Moreover, it comprises measurements from accelerometer, gyroscope, magnetometer, and GPS, corresponding to one of four activities: inactive, active, walking, or driving. Unlike other datasets, the collected data in this study were obtained from smartphones used in real-life scenarios.

## Introduction

Emerging as a distinct field of study, human activity recognition (HAR) has gained substantial attention due to its precise classification of diverse human actions. This field focuses on classifying the activities executed by different individuals by analysing data obtained from an array of sensors (Aggarwal and Xia, 2014). These sensors capture information while subjects carry out predetermined movements like nodding, raising a hand, walking, running, or driving. Wearable devices like activity wristbands and smartphones have proven immensely valuable in generating such data. More specifically, smartphones, with their abundant sensors and convenient dimensions, offer a user-friendly means of gathering high-quality data. Additionally, the insights derived from people's behavioural patterns, as tracked by these sensors, provide for various domains such as healthcare, fitness, and home automation, thus enhancing the capabilities of these sectors (Zahin et al., 2019; Zhu et al., 2015). This fusion of widespread sensing facilitated by smartphones and the consequent model development has given rise to a flourishing research domain that has garnered escalating interest in recent years (Demrozi et al., 2020; Lara and Labrador, 2012).

However, there are two main challenges encountered in this field. Firstly, efficiently managing the vast amount of data produced by those devices, alongside their temporal interdependence, is a significant hurdle. Secondly, establishing the correlation between this data and predefined movements remains an ongoing enigma. Some methodologies have demonstrated impressive outcomes in extracting insights from sensor data (Attal et al., 2015; Hassan et al., 2018). Nonetheless, it is noteworthy that these studies often involve adapting devices for specific placement, like attaching them to various body parts such as the wrist or waist. As a result,

[1] This document is heavily based on the contents of Garcia-Gonzalez et al. (2020).

the effectiveness of these models might be skewed, given the controlled setting in which data is gathered, encompassing limited activities and particular device orientations.

Moreover, the controlled environment established for those purposes significantly deviates from real-world scenarios. Users, especially those who employ smartphones, exhibit diverse manners of carrying those devices. In addition, individuals may vary their clothing choices, resulting in discrepancies in the orientation and placement of the devices. This variability in user behaviours and situations clearly contrasts with the hypothetical ideal. Notably, the artificial intelligence (AI) models introduced so far are markedly influenced by factors like placement and orientation. This reliance restricts their widespread applicability and obstructs their seamless integration into practical situations. In this way, the lack of a smooth transition to real-life circumstances constitutes a significant disparity. Nowadays, the AI models developed for HAR are intricately linked to specific orientations and positions. As a result, these models lack the flexibility to adapt universally, thereby limiting their extension to diverse user types. Consequently, the effort to personalise AI models for human activity recognition across a large spectrum of individuals remains an ongoing avenue of investigation. In fact, this exploration has persisted for nearly ten years (Solis Castilla et al., 2020; Weiss and Lockhart, 2012).

For the previously mentioned reasons, a new dataset was gathered, looking to close the gap with the real-life application. Specifically, it was collected using the sensors of 19 individuals' smartphones, with almost complete freedom. In this way, the differences in the orientation and placement of the used devices are various, as well as the physical characteristics of each participant.

## Data collection

The data collection process was made through a custom Android application developed by the authors, which streamlined the process of recording, categorising, and storing data. In this way, an initial data collection phase was undertaken, spanning approximately one month. The objective here was double: comprehending the nature of the acquired data and performing preliminary assessments. Then, a more intensive data gathering effort was conducted over a period of about one week. That served to rectify the imbalances and weaknesses identified in the preceding phase. The 19 participants in the study were instructed to specify the activity they were about to engage in using the previously mentioned Android application before commencing data collection. In such a manner, upon activity selection, data acquisition began automatically and stopped when the user indicated the conclusion of the activity. As a result, each recorded session corresponded to a distinct activity undertaken by a specific individual. The activities performed were classified into four categories:

- Inactive: not having the smartphone in motion. This consisted of any activity that involved not carrying the smartphone.
- Active: moving with the smartphone, without a specific destination. Activities such as preparing dinner, attending concerts, shopping for groceries, or doing household chores fell under the "active" category.
- Walking: any movement towards a defined location. Activities like running or jogging were categorised as "walking".
- Driving: every movement via motorised transportation, without the need to be the person driving. This included vehicles like cars, buses, motorbikes, trucks, and similar modes of conveyance.

Given that, data gathering originated from four distinct sensors: accelerometer, gyroscope, magnetometer, and GPS. The selection of accelerometer and gyroscope was predicated on their prevalence in the existing literature and their demonstrated efficacy. Furthermore, the magnetometer and GPS were also incorporated, proposing their utility in addressing this challenge. Specifically, GPS could play a crucial role in distinguishing those activities by detecting the user's movement velocity while carrying the smartphone. Anyhow, the data from the accelerometer, gyroscope, and magnetometer were stored with their tri-axial values. In the context of GPS data, the device's latitude, longitude, and altitude increments were stored alongside the measurement's bearing, velocity, and accuracy. In addition, the accelerometer data were refined by using the gravity sensor. That involved subtracting the last reading of the latter sensor from the observations of the former one, yielding refined accelerometer values (referred to as linear accelerometer values). These values remained unaltered by the smartphone's orientation, resulting in a dataset independent of the user's location and the device's orientation, as initially intended.

On the other hand, before local data storage, a series of filters were applied. Concerning the accelerometer and magnetometer, a low-pass filter was employed to mitigate excessive noise within the measurements from these sensors. In contrast, for the gyroscope, which confronts the well-recognised gyro drift issue, a high-pass filter was adopted as a workaround. Nevertheless, there was a challenge posed by Android, as each sensor could not be uniformly set to the same frequency. This situation proved particularly intricate in this context, with the need to merge data from sensors with highly disparate frequencies, such as the high-frequency accelerometer and the low-frequency GPS. While the accelerometer can yield as many as ten or even fifty measurements per second, the GPS provides new measurements approximately every ten seconds. Regrettably, Android's inherent constraints require the acceptance of receiving values as provided by the system, leading to potential data gaps. These gaps are particularly pronounced in the GPS data, wherein instances might arise when no new measurements are captured for over a minute (albeit this could be attributed to challenges related to enclosed environments). Similar gaps also emerge in accelerometer, gyroscope, and magnetometer data, notwithstanding their respective frequencies of around 10, 5, and 8 measurements per second under stable conditions. These gaps typically span 1 to 5 seconds, predominantly at the outset of each data collection session, although they occur less frequently than in GPS readings. Nevertheless, the average count of recordings per second for each sensor and activity is showcased in Table 1, along with the resulting mean frequency. A smaller font size beneath each average value outlines the corresponding standard deviation for each category. Notably, activities entailing movement, such as "active" or "walking", experience an elevation in these measurements, particularly noticeable with the accelerometer. The smartphone's sensors automatically heighten their frequency to derive maximum information from movements detected during these activities. This augmentation extends to the "driving" activity as well, possibly attributed to vehicular vibrations that the smartphone sensors might also detect. Furthermore, in instances of "walking" and "active" activities, intermittent periods of inactivity, such as waiting at traffic lights or moments of standing engagement, contributed to a moderation in these average frequencies.

Table 1: Mean recordings per second for each sensor and every measured activity.

| Activity | Accelerometer Hz. | Gyroscope Hz. | Magnetometer Hz. | GPS Hz. |
|---|---|---|---|---|
| **Inactive** | 11.00<br>±16.38 | 4.66<br>±0.74 | 7.91<br>±11.72 | 0.13<br>±0.35 |
| **Active** | 32.55<br>±24.80 | 4.46<br>±1.44 | 9.13<br>±13.64 | 0.06<br>±0.23 |
| **Walking** | 31.24<br>±27.47 | 6.24<br>±11.86 | 8.16<br>±12.05 | 0.06<br>±0.23 |
| **Driving** | 51.16<br>±31.59 | 4.66<br>±2.42 | 17.00<br>±20.01 | 0.04<br>±0.20 |

Table 2: Data distribution for each measured activity in the dataset.

| Activity | Recorded Time (s) | Number of Recordings | Number of Observations | Percentage of Data |
|---|---|---|---|---|
| **Inactive** | 292,213 | 147 | 7,064,757 | 24.25% |
| **Active** | 178,806 | 99 | 8,918,021 | 30.62% |
| **Walking** | 98,071 | 200 | 4,541,130 | 15.59% |
| **Driving** | 112,226 | 128 | 8,602,902 | 29.54% |
| **Overall** | 681,316 | 574 | 29,126,810 | 100% |

As a result, the ultimate distribution of activities within the dataset is presented in Table 2. This table illustrates the total recorded time, the number of recordings, the number of observations, and the corresponding data percentage (the latter being related to the number of observations) for each designated activity. Here, a "recording" corresponds to an entire activity session, from the start of an action to its end. In contrast, an "observation" corresponds to an individual sensor measurement. Notably, a relatively lower number of observations is observed in "inactive" activities relative to the overall recorded time. This discrepancy arises from the sensors' increased frequency during activities involving more movement, a phenomenon explained earlier. In this way, the general data distribution, when measured by total percentage, might misperceive the actual scenario when sliding windows are introduced. That is because, with sliding windows employed for feature computation, the number of observations becomes secondary in importance, with total recorded time taking precedence. The more extensive the recorded time, the greater the number of computed sliding windows and resultant samples for a given class. Consequently, an imbalance becomes clearly evident in the dataset, wherein the "inactive" activity contains thrice as many samples as the "walking" category. As for the number of recordings made, a notable disparity exists, with the "walking" activity featuring significantly more recordings than the others. Regardless, the dataset is deemed valuable and feasible for developing models capable of discerning these activities. Furthermore, the study engaged 19 individuals, contributing a diverse array of behaviours that inherently enrich the potential models that could be crafted in the subsequent stages.

Furthermore, the data acquisition process sought to contain a spectrum of individuals with various attributes, encompassing differences in physical traits, usage routines for their smartphones, and the device models utilised. As a result, the study engaged 19 participants, spanning an age range of roughly 25 to 50 years. This approach was adopted to ensure the inclusion of an extensive array of behavioural patterns that could significantly contribute to developing subsequent models. However, it is worth noting that gender diversity remains limited, with merely two female participants. Nevertheless, the participants exhibit a wide array of physical attributes, habits, and smartphone preferences, with their manner of use and device placement. This scope of variation signifies that while there might be room for enhancement in terms of variability, a noteworthy level of diversity remains inherent within the dataset, as pursued initially.

Nonetheless, an additional difficulty arises within the Android framework, as not all devices are equipped with both a gyroscope and a magnetometer. Although an accelerometer and GPS are requisite, older Android versions do not mandate the inclusion of a gyroscope or magnetometer. Consequently, certain users stored measurements without the involvement of these sensors. Tables 3 and 4 provide an overview of the number of observations lacking a gyroscope or both a gyroscope and magnetometer concurrently. It is worth noting the contrast between the relationship of number of observations to recorded time depicted in these tables compared to the report in Table 2. Specifically, the number of observations is considerably higher compared to the recorded time. This peculiarity might account for the previously noted unusual data in Table 1, wherein the accelerometer's frequency potentially increases more extensively as it becomes the sole sensor detecting motion. In addition, the percentages displayed in these tables are derived from the entire dataset quantity, as outlined in Table 2. Fortunately, these percentages remain pretty modest, and the dataset's integrity is relatively unharmed by this issue. Nevertheless, it is prudent to bear this in mind while preparing the data for its application in forthcoming AI models.

Table 3: Distribution of the dataset for each measured activity excluding the gyroscope.

| Activity | Recorded Time (s) | Number of Recordings | Number of Observations | Percentage of Data |
|---|---|---|---|---|
| **Inactive** | 11,523 | 8 | 668,536 | 2.29% |
| **Active** | 13,866 | 7 | 619,913 | 2.13% |
| **Walking** | 4,169 | 15 | 584,262 | 2.01% |
| **Driving** | 25,718 | 23 | 3,776,468 | 12.97% |
| **Overall** | 55,276 | 53 | 5,649,179 | 19.40% |

Table 4: Distribution of the dataset for each measured activity without the gyroscope and magnetometer.

| Activity | Recorded Time (s) | Number of Recordings | Number of Observations | Percentage of Data |
|---|---|---|---|---|
| **Inactive** | 5,409 | 2 | 269,710 | 0.93% |
| **Active** | 10,286 | 2 | 90,487 | 0.31% |
| **Walking** | 0 | 0 | 0 | 0% |
| **Driving** | 0 | 0 | 0 | 0% |
| **Overall** | 25,695 | 4 | 360,197 | 1.24% |

## Actual data behaviour example

In this section, an example is provided to illustrate the behaviour of the previously collected data in a real-world scenario, wherein the data had undergone prior preprocessing, encompassing the removal of outliers and other observations that could be deemed corrupted.

Accordingly, without delving deeper into such preprocessing steps, Table 5 presents each sensor's mean and standard deviation values for each studied activity. Given that, to accurately comprehend the values in that table, it is worth explaining what each sensor measures. First, the accelerometer values correspond to the acceleration force applied to the smartphone along the three physical axes $(x, y, z)$ in $\mathrm{m/s^2}$. Then, the gyroscope measures the rotation speed of the smartphone around each of the three physical axes $(x, y, z)$ in rad/s. Regarding the magnetometer, it measures the environmental geomagnetic field along the three physical axes $(x, y, z)$ of the smartphone in $\mu$T. Concerning the GPS, its values include increments in longitude and latitude coordinates relative to the previous measurement, as well as increments in altitude in meters. Moreover, the values of speed, bearing, and accuracy were also considered. In this way, speed, measured in m/s, represents the smartphone's velocity. As for the bearing, it indicates the horizontal direction of the smartphone's travel in degrees. Finally, the accuracy values indicate the deviation from the actual smartphone location, expressed in meters, with smaller values indicating higher measurement accuracy. Back to Table 5, note that each cell contains the mean values at the top and the corresponding standard deviation values below in smaller font size. Each pair of values corresponds to each sensor set, where the accelerometer, gyroscope, and magnetometer refer to their respective axes ($X$, $Y$ and $Z$). For GPS, the set includes latitude increments (*Lat.*), longitude increments (*Long.*), altitude increments (*Alt.*), speed (*Sp.*), bearing (*Bear.*), and accuracy (*Acc.*) measurements. There, it is worth noting some rare data, such as those associated with the GPS "inactive" activity, which exhibit unexpectedly high values. That can be attributed to the fact that such action was often performed indoors, which may limit GPS accessibility. Nevertheless, noticeable differences exist between the activities, indicating the potential for identification with future models.

Table 5: Mean and standard deviation values of sensors for each recorded activity.

| | | Activity | | | |
|---|---|---|---|---|---|
| | | **Inactive** | **Active** | **Walking** | **Driving** |
| **Accelerometer** | $X$ | 0.11761 ±0.45934 | −0.01338 ±1.30277 | 0.09425 ±3.33422 | −0.04747 ±0.83290 |
| | $Y$ | 0.06136 ±0.26764 | 0.07598 ±1.45440 | −0.37604 ±4.35808 | −0.12936 ±0.93828 |
| | $Z$ | 0.84318 ±2.66926 | 0.13008 ±1.70294 | 0.07353 ±4.09859 | 0.18127 ±1.24042 |
| **Gyroscope** | $X$ | −0.00004 ±0.03828 | −0.00001 ±0.36806 | 0.00760 ±1.31125 | 0.00080 ±0.19224 |
| | $Y$ | 0.00004 ±0.04719 | −0.00102 ±0.40959 | −0.00020 ±0.89244 | 0.00277 ±0.19835 |
| | $Z$ | 0.00001 ±0.03526 | 0.00055 ±0.24528 | −0.00560 ±0.53685 | −0.00243 ±0.16678 |
| **Magnetometer** | $X$ | 25.93805 ±56.45617 | 6.03153 ±30.00980 | −0.28182 ±27.03210 | −5.96356 ±46.08005 |
| | $Y$ | −19.62683 ±85.70343 | −0.02890 ±28.76398 | 18.73800 ±29.63926 | 10.73609 ±40.46829 |
| | $Z$ | −56.60425 ±33.19593 | 9.56310 ±39.76136 | 0.64541 ±25.55331 | −2.93043 ±29.45994 |
| GPS | **Lat.** | 0.00075 ±0.00166 | 0.00112 ±0.00234 | 0.00047 ±0.00220 | 0.00175 ±0.00365 |
| | **Long.** | 0.00125 ±0.00285 | 0.00118 ±0.00314 | 0.00056 ±0.00300 | 0.00204 ±0.00420 |
| | **Alt.** | 32.59169 ±53.06269 | 30.77538 ±48.65634 | 34.06931 ±42.51933 | 41.59391 ±54.74934 |
| | **Sp.** | 0.37222 ±0.82495 | 0.12109 ±0.81007 | 0.79924 ±0.71835 | 10.82191 ±11.82733 |
| | **Bear.** | 57.25005 ±105.49576 | 14.69719 ±56.00693 | 124.85103 ±119.80663 | 118.88108 ±118.78510 |
| | **Acc.** | 265.44485 ±494.66499 | 214.57640 ±429.81169 | 75.54539 ±259.59907 | 192.90736 ±508.87285 |

In any case, to depict the actual distribution that may arise when processing the data to feed the relevant AI models, an illustrative example is presented in Table 6. In this instance, the data corresponds to the application of a 20-second window with a 19-second overlap, derived from the dataset resulting from the simultaneous utilisation of all sensors. As discernible, a bias towards the "inactive" activity is ultimately observed, as previously discussed, owing to the ease of collecting such data in comparison to the rest. Nonetheless, an adequate number of samples exists for all the studied activities. Consequently, their subsequent classification is viable, necessitating only an awareness of this issue and its ensuing mitigation.

Table 6: Total sample count generated using all sensors with a 20-second sliding window and 19-second overlap.

| Activity | | | | |
|---|---|---|---|---|
| **Inactive** | **Active** | **Walking** | **Driving** | **Overall** |
| 214,130 (43%) | 140,060 (28%) | 83,376 (17%) | 61,710 (12%) | 499,276 |

## Acknowledgements

This work was funded by CITIC is funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS), Xunta de Galicia/FEDER-UE (ConectaPeme, GEMA: IN852A 2018/14), MINECO-AEI/FEDER-UE (Flatcity: TIN2016-77158-C4-3-R) and Xunta de Galicia/FEDER-UE (AXUDAS PARA A CONSOLIDACION E ESTRUTURACION DE UNIDADES DE INVESTIGACION COMPETITIVAS.GRC: ED431C 2017/58 and ED431C 2018/49).

## Bibliography

J. K. Aggarwal and L. Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70–80, 2014.

F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 15(12):31314–31338, 2015.

F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi. Human activity recognition using inertial, physiological and environmental sensors: a comprehensive survey. *arXiv preprint arXiv:2004.08821*, 2020.

D. Garcia-Gonzalez, D. Rivero, E. Fernandez-Blanco, and M. R. Luaces. A public domain dataset for real-life human activity recognition using smartphone sensors. *Sensors*, 20(8):2200, 2020.

M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren. A robust human activity recognition system using smartphone sensors and deep learning. *Future Generation Computer Systems*, 81:307–313, 2018.

O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, 15(3):1192–1209, 2012.

R. Solis Castilla, A. Akbari, R. Jafari, and B. J. Mortazavi. Using intelligent personal annotations to improve human activity recognition for movements in natural environments. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2020.

G. Weiss and J. Lockhart. The impact of personalization on smartphone-based activity recognition. *AAAI Publications, Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

A. Zahin, R. Q. Hu, et al. Sensor-based human activity recognition for smart healthcare: A semi-supervised machine learning. In *International Conference on Artificial Intelligence for Communications and Networks*, pages 450–472. Springer, 2019.

N. Zhu, T. Diethe, M. Camplani, L. Tao, A. Burrows, N. Twomey, D. Kaleshi, M. Mirmehdi, P. Flach, and I. Craddock. Bridging e-health and the internet of things: The sphere project. *IEEE Intelligent Systems*, 30(4):39–46, 2015.