Data in Brief 37 (2021) 107164

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

ContextLabeler dataset: Physical and virtual sensors data collected from smartphone usage in-the-wild

Mattia Giovanni Campana*, Franca Delmastro

Institute of Informatics and Telematics, National Research Council of Italy, Pisa, Italy

ARTICLE INFO

Article history:
Received 7 May 2021
Accepted 19 May 2021
Available online 21 May 2021

Keywords:
Context-aware
Smartphone sensing
Social sensing
User context
Mobile computing

ABSTRACT

This paper describes a data collection campaign and the resulting dataset derived from smartphone sensors characterizing the daily life activities of 3 volunteers in a period of two weeks. The dataset is released as a collection of CSV files containing more than 45K data samples, where each sample is composed by 1332 features related to a heterogeneous set of physical and virtual sensors, including motion sensors, running applications, devices in proximity, and weather conditions. Moreover, each data sample is associated with a ground truth label that describes the user activity and the situation in which she was involved during the sensing experiment (e.g., *working, at restaurant*, and *doing sport activity*). To avoid introducing any bias during the data collection, we performed the sensing experiment in-the-wild, that is, by using the volunteers' devices, and without defining any constraint related to the user's behavior. For this reason, the collected dataset represents a useful source of real data to both define and evaluate a broad set of novel context-aware solutions (both algorithms and protocols) that aim to adapt their behavior according to the changes in the user's situation in a mobile environment.

DOI of original article: 10.1016/j.eswa.2021.115124
* Corresponding author.
E-mail address: mattia.campana@iit.cnr.it (M.G. Campana).
Social media: (M.G. Campana)

https://doi.org/10.1016/j.dib.2021.107164

Specifications Table

Subject	Computer Science
Specific subject area	User context sensing and modeling for detecting human complex activities and situations in mobile environments.
Type of data	Comma Separated Files (CSV)
How data were acquired	Android mobile application
Data format	Raw sensors data with user annotations
Parameters for data collection	The experiment has been designed to collect context data into the wild. In other words, to avoid introducing biases during the data acquisition, we did not define any constraints for the user behavior during the experiment. For example, we encouraged the volunteers to use their smartphones without worrying about the positions of the device (e.g., trousers' pockets, or hand).
Description of data collection	In order to take into account the diversity of different devices, the volunteers have installed the sensing application on their smartphones. The mobile application has been designed for Android OS, and it collects data generated by a heterogeneous set of sensors, including both physical and virtual sensors. The collected data was stored in the internal storage unit of the mobile device. The volunteers were able to start and stop the sensing application whenever they wanted, and they freely annotated the collected data by choosing among a set of predefined daily life activities.
Data source location	The dataset has been mainly generated within the geographical area defined by the following 3 cities located in the Tuscany region, Italy: Pisa, Lucca and Pontedera.
Data accessibility	Direct URL to data: https://github.com/contextkit/ContextLabeler-Dataset
Related research article	M.G. Campana, F. Delmastro, "COMPASS: Unsupervised and Online Clustering of Complex Human Activities from Smartphone Sensors", accepted for publication in Elsevier Journal of Expert Systems with Applications, https://doi.org/10.1016/j.eswa.2021.115124.

Value of the Data

- The presented dataset provides a broad set of sensors data describing human complex activities collected from the use of commercial smartphones into-the-wild. All the data samples have been freely annotated by the users in order to specify their daily life activities. Moreover, since we have not defined any constraint for the user behaviour, the presented dataset is not affected by biases that can be introduced by performing predefined actions in controlled environments (e.g., laboratory).
- Researchers can use this dataset to analyze and automatically recognize the situation in which the user is currently involved by using commercial smartphones.
- This dataset provides a valuable starting point for the automatic detection of the user context in a mobile setting. Specifically, it can be used to evaluate novel context-aware solutions, including recommender systems, activity recognition algorithms, and wireless communication protocols.
- The dataset presents also additional values: (i) the data has been collected in real environments, without defining any sort of constraints related to the user behaviour nor to the interactions between the user and her mobile device; (ii) each data sample is represented by a high-dimensional vector composed by more than 1K features extracted from a heterogeneous set of mobile sensors (both physical and virtual); (iii) the data has been freely annotated by the users according to their daily life activities.

1. Data Description

The dataset contains smartphone sensors data collected from the personal devices of 3 volunteer users in their usual environment. It is released in the form of a set of comma-separated (CSV) files, one for each volunteer, and they are respectively named as follows: **user_1.csv, user_2.csv**, and **user_3.csv**. The CSV files contain time series of sensors data collected from the users' devices through a mobile application specifically designed for the sensing experiment. This application has been used by the volunteers for two weeks to annotate the collected sensed data with labels that describe their daily life activities. In total, we collected 45681 data samples that are distributed among the three files as follows: 8456 samples in user_1.csv, 17882 samples in user_2.csv, and 19343 in user_3.csv.

The dataset contains both physical and virtual sensors data that can be used to characterize all the different aspects of the user context in a mobile setting. Physical sensors are implemented in the hardware equipment of the mobile phone (e.g., accelerometer), while virtual sensors represent data sources that describe the device's status, the surrounding environment, and the interactions between the user and her device.

Each data sample is composed by 1332 features, with both continuous and categorical values, describing a heterogeneous set of sensors. According to the type of sensors they describe, we can divide the collected features in 13 categories. In the following, we describe in details the data collected for each of the sensor categories, along with the number of columns in which they are located in the CSV source files:

- **Date and Time**, columns 1–7: each data sample is associated with a Unix timestamp that represents the instant in which our sensing application has captured the sensors data. Starting from the timestamp, we also extracted 6 categorical features related to both day and time, i.e., *weekday, weekend, morning, afternoon, evening*, and *night*.
- **User gait**, columns 8–15: 8 categorical features that represent the user's gait detected by the Android Activity Recognition API[1]:
  - *activity_rec_in_vehicle*: the user is in a vehicle (e.g., a car),
  - *activity_rec_on_bicycle*: the user is riding a bicycle,
  - *activity_rec_on_foot*: the user is walking or running,
  - *activity_rec_running*: the user is running,
  - *activity_rec_still*, the device is not moving,
  - *activity_rec_tilting*, the user is rapidly moving the device,
  - *activity_rec_walking*, the user is walking,
  - *activity_rec_unknown*, the Google API is not able to recognize the current user's activity
- **Running applications**, columns 16–71: 56 categorical features that represent the possible main application categories, according to the Google Play Store (e.g., *ART_AND_DESIGN, BUSINESS*, and *ENTERTAINMENT*). The value of a feature represents the number of running applications that belong to the corresponding application category.
- **Weather conditions**, columns 72–133**:** based on the user's location, we defined a total of 62 features by using the information collected from the OpenWeatber API service[2]. More specifically, we defined the following 8 continuous features:
  - *weather_temp*: the current temperature in Celsius,
  - *weather_temp_min*: the minimum temperature of the day,
  - *weather_temp_max*: the maximum temperature of the day,
  - *weather_humidity*: the percentage of humidity,
  - *weather_pressure*: the atmospheric pressure in hPa,
  - *weather_wind_speed*: the wind speed in meter/sec,
  - *weather_wind_direction*: the wind direction in degrees,
  - *weather_cloudiness*: the percentage of cloudiness

[1] https://developers.google.com/android/reference/com/google/android/gms/location/DetectedActivity.html.
[2] https://openweathermap.org/api.

In addition, we defined a total of 54 categorical features derived from the weather conditions codes defined by the OpenWeather service[3].

- **Audio**, columns 134–145: a set of 12 categorical and continuous features related to the current smartphone's audio settings. Specifically, we defined 4 categorical features to represent the ringer mode (i.e., *audio_ringer_mode_silent*, *audio_ringer_mode_vibrate*, and *audio_ringer_mode_normal*), and the following 5 categorical features for other audio characteristics: *audio_bt_sco_on* and *audio_headset_on*, that respectively represent whether a bluetooth and a wired headset is connected to the device or not; *audio_music_active* and *audio_speaker_on* that respectively indicate if the music and the speaker are active; and *audio_mic_mute*, that represent if the microphone is on or off. In addition, we defined the following continuous features to characterize the level of different audio settings:
  - *audio_alarm_volume*: the alarm volume,
  - *audio_music_volume*: music volume,
  - *audio_notification_volume*: the volume level set for the notifications,
  - *audio_ring_volume*: the ringtone volume
- **Battery**, columns 146–149: 4 categorical features related to the battery information. Specifically, a feature that represents whether the device is connected to a power source or not (i.e., *battery_unplugged*), and 3 features to characterize the type of power source: *battery_plugged_ac* (an AC charger), *battery_plugged_usb* (a USB port), and *battery_plugged_wireless* (an inductive wireless charger).
- **Bluetooth connections**, columns 150–204: 55 categorical features that characterize the type of the first 5 Bluetooth devices connected to the user's smartphone. Specifically, based on the Bluetooth Major ID number, the possible device categories are the following: *audio_video, computer, health, imaging, misc, networking, peripheral, phone, toy, wearable*, and *uncategorized*.
- **Bluetooth devices in proximity**, columns 205–259: 55 categorical features that characterize the type of the first 5 Bluetooth devices in proximity.
- **Display status**, columns 260–270: 11 categorical features that represent information about the display status. Specifically, the following features describe the current display state:
  - *display_status_on*: the display is on,
  - *display_status_off*: the display is off,
  - *display_status_doze*: the display is in a low-power state: the display shows only system-provided content while the device is non-interactive,
  - *display_status_doze_suspended*: the display is in a suspended low-power state, where the CPU is no more updating it,
  - *display_status_vr_mode*: the display is optimized for the Virtual Reality (VR) mode,
  - *display_status_on_suspended*: the display is in a full-power mode, but the display is not updating it,
  - *display_status_unknown*: the system is not able to recognize the current display status,
  - while the following features characterize the rotation angle of the display: *display_rotation_0* (natural -vertical- orientation), *display_rotation_90* (horizontal mode), *display_rotation_180* (vertical and rotated by 180 degree), and *display_rotation_270* (horizontal and rotated by 270 degree).

- **Location**, columns 271–1193: two continuous features that respectively represent the geographical coordinates (i.e., latitude and longitude) of the user's current location. Moreover, based on the user's location, we downloaded the category of the most probable venue according to the Foursquare Places API (e.g., Art Gallery or Italian Restaurant)[4]. Therefore, we also defined 921 categorical features that represent the main venue categories defined by Foursquare[5].

[3] https://openweathermap.org/weather-conditions.
[4] https://developer.foursquare.com/docs/api/venues/search.
[5] https://developer.foursquare.com/docs/resources/categories.

- **Wi-Fi**, column 1194: a categorical feature that represents whether the mobile device is currently connected to a Wi-Fi Access Point or not.
- **Physical Sensors**, columns 1195–1330: a set of 136 continuous features that represent several descriptive statistics related to the following physical sensors: light, accelerometer, gravity, gyroscope, linear acceleration, rotation, and proximity. More specifically, for each sensor we collected 200 data samples and we calculated the following statistics: minimum, maximum, and average values; quadratic mean; 25th, 50th, 75th, and 100th percentiles. Moreover, for those sensors that are composed of multiple components (e.g., a 3-axis gyroscope), we calculated the same set of statistics for each component.
- **Multimedia**, columns 1331–1332: 2 categorical features that represent whether the user was taking a picture or recording a video with her smartphone.

Finally, each data sample is associated with its Ground Truth (column 1333): the label specified by the user to describe the type of situation in which she was involved when the application collected the sensors' data. The labels specified by the users are the following: *Working, Restaurant, Lunch Break, Shopping, Break, Home, Nightlife, Sleep, Physical exercise*, and *Free time*.

2. Experimental Design, Materials and Methods

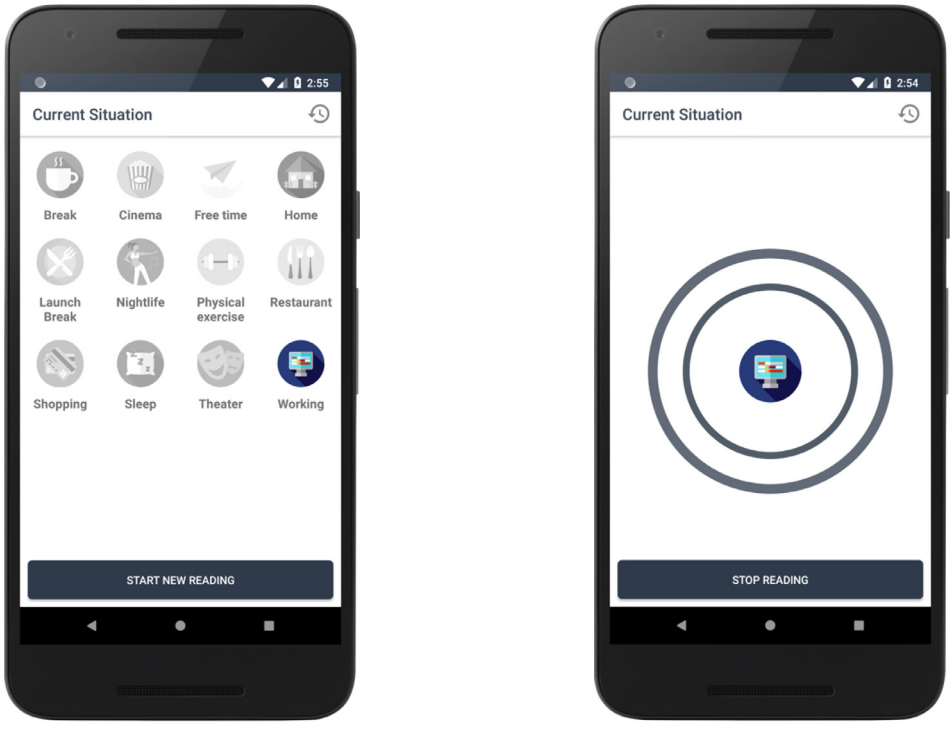
The dataset we release with this work is the result of a data collection campaign designed to capture the complexity of the user context in a mobile environment. With the term context we mainly refer to the activities performed by a person during her daily life and the situations in which she can be involved. Examples of possible contexts are the following: *attending a lecture, being at home*, and *taking a coffee with friends*.

According to the literature [1], simple human activities (e.g., *running* or *walking*) can be characterized by using a small set of sensors embedded in personal and wearable devices like, for example, the accelerometer and the gyroscope. On the contrary, complex activities are characterised by higher-level semantics and require a combination of heterogeneous sources of data. Therefore, to infer the user situation by using the sensing capabilities of her mobile and personal devices, we need to take into account a broad set of heterogeneous data sources. To this aim, the simple physical sensors are not enough, but we also need to exploit the so-called virtual sensors, i.e., those data sources that characterize the user-device interactions as long as the surrounding environment (e.g., running applications and devices in proximity).

Research studies in the area of activity recognition and human behavior modeling usually base their results on experiments performed in controlled environments (e.g., a research laboratory) [2]. During the data collection process (often performed with the same device), volunteers are asked to perform some activities that have been previously defined by researchers. However, in the real world, we have heterogeneous devices and different users may have different ways of doing the same activity; thus the experimental results usually diverge from those obtained in the lab [3].

To build a realistic and valuable dataset, we enrolled three voluntary users equipped with heterogeneous commercial mobile devices, with different characteristics and sensors: a Nexus 5 with Android 6.0.1, a Xiaomi Mi 5 with Android 7.1.2, and a Reader P10 with Android 6.0. To collect the dataset we developed Context Labeler, an Android application that allows the volunteers to freely annotate the sensed data. More specifically, we asked the volunteers to install the sensing application on their smartphones and to select their daily life activities among the following set of labels: *Break, Cinema, Free time, Home, Lunch Break, Nightlife, Physical exercise, Restaurant, Shopping, Sleep, Theatre*, and *Working*. Fig. 1a shows the User Interface offered by Context Labeler to specify the label associated with the current user's context. After the activity selection, Context Labeler starts ContextKit[6], our sensing framework that monitors a broad set of sensors, both physical and virtual [4].

[6] https://contextkit.github.io.



(a) Label selection. (b) Reading sensors.

Fig. 1. User Interface of the Context Labeler mobile application.

In order to avoid affecting the user behavior and the interactions with her device, the data collection is completely performed unobtrusively in the background. When the current activity ends, the user manually stops the data reading using a specific button (Fig. 1b) and both the sensed data and the selected label are stored into the device's hard drive.

Before running the application, the users signed an informed consent including all the policies adopted for personal data storage, management, and analysis, including the publication of the anonymized dataset, according to the EU GDPR. In addition, to avoid introducing biases during the data acquisition, we did not define any constraints for the user behavior during the experiment. On the contrary, we encouraged the volunteers to use their smartphones without worrying about the positions of the device (e.g., trousers' pockets, or hand).

Table 1 shows the sampling rate of each sensor's category that we used in Context Labeler during the data collection. When the user starts the sensing procedure, the application collects new data samples every 1–5 minutes for most of the sensors; while it downloads the weather conditions every hour from the OpenWeather service. Moreover, both the *Bluetooth Connections* and the *Multimedia* sensors react to specific events. Specifically, when the user connects or disconnects a Bluetooth device to her smartphone, or she takes a photo or records a video, Context Labeler saves such information on the log files.

ContextKit stores the sensed data in dedicated log files, one for each monitored sensor, alongside with the reading timestamps. However, different sensors or events monitored by the framework may have different sampling rates. Therefore, even if two different sensor data refer to the same user context, they may have slightly different timestamps. Moreover, each label collected by the application is stored, together with its duration, in a separate log file.

Table 1
Sensors categories and their sampling rates used during the data collection.

Category	Sampling rate (sec.)
User gait	60
Running applications	300
Weather conditions	3600
Audio	60
Battery	60
Bluetooth connections	On event
Bluetooth scans	60
Location	300
Wi-Fi	180
Physical sensors	60
Multimedia	On event

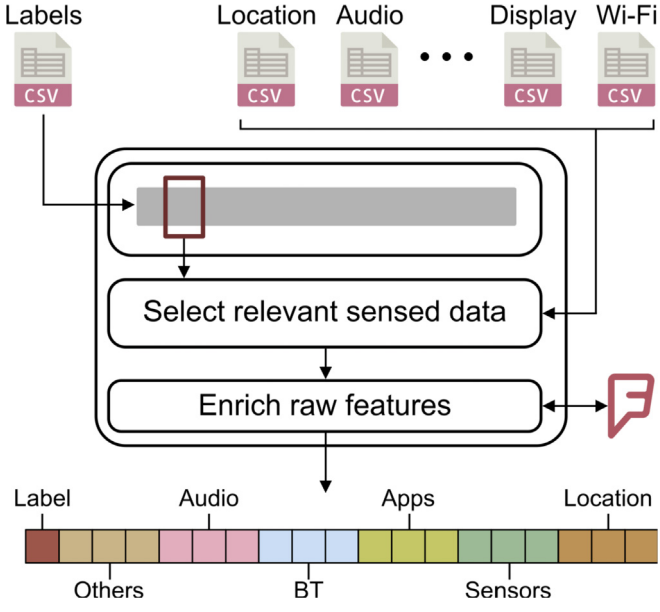


Fig. 2. Dataset building process.

In order to generate a dataset which is ready to be used for research purposes (e.g., to evaluate context-recognition algorithms), we processed the log files as shown in Fig. 2. First, we used a sliding window approach to split the duration of each user situation into slots of 1 second each. Second, for every time slot, we fetched from the raw log files only the sensor data with the closest reading timestamp to the starting time of the current slot. In this way, we kept only dense feature vectors, and we discard data samples wit missing values. Then, we enriched the raw features with additional categorical information. For example, using the *Foursquare APIs*[7], we extended the location features by retrieving the category of the most probable venue according to the GPS coordinates. Finally, we have created the final feature vector by combining the categorical features with the continuous ones derived from physical sensors values, and we associate the corresponding situation's label indicated by the user.

[7] https://developer.foursquare.com.

Since an ordinal relationship among categorical features does not exist, we include the categorical features into the features vector by using the well-known One Hot Encoding technique, which creates a binary feature for each possible category. For example, according to the Android Framework[8], the possible display orientation modes are the following: 0, 90, 180, and 270 degrees. Assuming that, for a given context snapshot, the display was held by the user in portrait mode, we have created 4 different features for describing the display status, where one of them (i.e., the one corresponding to 0 degrees) is set to 1, while the others are set to 0. The resulting dataset contains 45681 labeled samples, where each sample is composed by 1332 features.

Ethics Statement

An informed consent has been obtained by each participant before the data collection. In addition, data are fully anonymized.

CRediT Author Statement

Mattia G. Campana: Conceptualization, Methodology, Software, Data curation, Writing - Original draft preparation, Writing - Reviewing and Editing; **Franca Delmastro:** Conceptualization, Methodology, Supervision, Writing - Reviewing and Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

This work has been carried out in the framework of the INTESA project (CUP CIPE D78I16000010008), co-funded by the Tuscany Region (Italy) and MIUR under the Programme FAR FAS 2007-2013.

References

[1] L. Peng, et al., Complex activity recognition using acceleration, vital sign, and location data, IEEE Trans. Mob. Comput. 18 (7) (2018) 1488–1498, doi:10.1109/TMC.2018.2863292.
[2] D. Micucci, M. Mobilio, P. Napoletano, Unimib shar: a dataset for human activity recognition using acceleration data from smartphones, Appl. Sci. 7 (10) (2017) 1101.
[3] R. Mafrur, I. Gde Dharma Nugraha, D. Choi, Modeling and discovering human behavior from smartphone sensing life-log data for identification purpose, Hum.-Centric Comput. Inf. Sci. 5 (1) (2015) 31.
[4] M.G. Campana, Lightweight modeling of user context combining physical and virtual sensor data, in: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, 2018, doi:10.1145/3267305.3274178.

[8] https://developer.android.com/reference/android/view/Display#getOrientation().