Dataset documentation for the paper "When Your Fitness Tracker Betrays You: Quantifying the Predictability of Biometric Features Across Contexts"

Abstract

This report contains the documentation of the biometric dataset collected and used in the paper "When Your Fitness Tracker Betrays You: Quantifying the Predictability of Biometric Features Across Contexts" [1]. The dataset contains five different behavioural biometric traits (eye movements, mouse movements, touch dynamics, gait and ECG) from 30 users, measured in two separate session. Additionally, each biometric trait has been measured in different contexts. The report first presents the general structure of the dataset, then for each biometric describes the details of the feature extraction methodology and the format and content of each file.

Sensor Placement	chest	watch	arm	hand	pocket
Device Name	Movisens ekgMove	Garmin VivoActive HR	Blu Vivo 6	Blu Vivo 6	Blu Vivo 6

Table 1: Accelerometer sensor placement and corresponding devices used for the Gait measurements.

1 Dataset Structure

Folder Hierarchy. The dataset for each biometric contains a single folder for each user, each user's folder may then contain additional subfolders for individual contexts as documented in the following sections. As an example, folder `gait/12345/pocket` contains the gait data (accelerometer) for session 12345 as measured by the sensor located in the user's pocket. The two sessions for each user have distinct ids (distinct subfolders). A metadata file `meta.csv` is provided, and contains additional user's information, including the relationship between first and second session. For each user, general information like height (cm), weight (kg) and gender is recorded. In addition, the user's self-reported amount of weekly exercise (in hours) is shown for general exercise and running in particular.

File Format. All files are provided in `.csv` format, the first row is a header that indicates the content of the columns. If there is no `timestamp` column, it means that the sampling rate was constant and can be found in this document in the relative subsection.

2 Gait

Methodology. The user wore five different sensors that measured accelerometer data, and were instructed to (i) walk from a start point to an end point, (ii) come back to the start point, then (iii) jog to the end point and (iv) jog back to the start point. The measurements were taken in University Parks in Oxford, Figure 1 shows a map with start and end point. Table 1 shows the sensor placement and the devices used.

Data Format. Accelerometer data are found in `gait/<userid>`. Data was collected through the `Motion Sensor` Android API and the Movisens API. Each file filename is in the form `<subf>/<activity><n>.csv`:

- `<subf>` identifies the sensor placement (`chest`, `watch`, `arm`, `hand` or `pocket`);
- `<activity>` identifies the measured activity (`walk` or `jog`);
- `<n>` identifies whether the user was going from start to end, or viceversa (`1` or `2`).

The sampling rate of the ekgMove device was 64 Hz.

Figure 1: Map showing where the Gait data was measured. Users were instructed to walk from start to end point, walk back to start, then jog to end, then jog back to start.

3 ECG

Methodology. The Lead I,II,III measurements were collected through a Heal Force Prince 180B ECG monitor and adhesive electrodes. As the monitor only supports a single lead, measurements were taken one after another (i.e., not simultaneous). The same device was also used to collect the palm measurement using the built-in electrodes.

As the monitor software does not support exporting the raw ECG data, the data provided was extracted from the plot shown by the software, details of this can be found in [2]. As such, the measurements do not correspond to any specific unit. The sampling rate of the data was 150 Hz.

The chest strap data was collected through the Movisens ekgMove at a sampling rate of 1024 Hz.

The mobile dataset was collected through the AliveCor KardiaMobile device, similar to the Prince 180B the data was extracted from plots at a sampling rate of 300 Hz.

Data from the Nymi Band was collected through the SDK following authentication[1]. The data consistently exhibits high baseline drift, this can be easily removed through a low-pass filter as documented in [2]. The data sampling rate is 250 Hz. Following initial enrolment, each user made up to two attempts to authenticate to the band. The outcome of this is documented in the `meta.csv` file: 1 meaning first attempt successful, 2 meaning second attempt successful

[1]Note that this functionality is no longer available in later versions of the SDK or the Nymi Band's consumer version

and 0 meaning both failed.

Data Format ECG data are found in `ecg/<userid>`. Each file filename is in the form `<dev>_<activity>.csv`:

- `<dev>` identifies the device used for the measurement (`ekgMove`, `Lead`, `mobile`, `nymi` or `Palm`);
- `<activity>` identifies the measured activity (`rest`, `walk` or `jog`);

4 Touch Dynamics

Methodology. The user played a "spot the difference" game on a smartphone, where he swiped between two images with subtle differences and was requested to spot them (taken from [3]). The number of differences found by each user is documented in the `meta.csv` file The process was repeated for three times with a different smartphone each time: M5 Smart Phone, Motorola Moto G3 and Blue Vivo 6. For each smartphone, the user played the game for three minutes. The Android application used is available online[2]. The images (Figure 2) have been taken from the Allstarpuzzles website[3] with the author's consent.

Data Format. Touchscreen data are found in `touch/<userid>`. All Data was collected through the `MotionEvent` Android API. Filenames are in the form `<device>_<imageset>.csv`:

- `<device>` identifies the smartphone used for the measurement (`ttsim`, `moto` or `vivo`);
- `<imageset>` identifies the pair of images used (`set1`, `set2` or `set3`, shown in Figure 2).

One should note that due to the different devices, some values in the data have different granularity (e.g., the API value returned for pressure was constant for the M5 Smart phone). In addition, the update frequency of the data varies between devices.

5 Mouse Movements

Methodology. Each participant was asked to play a mole clicking game (see Figure 3) twice, once with a mouse and once with a laptop trackpad. After the user clicks on the mole, it switches to a random location. The random sequence is kept identical for all participants. After completing 250 iterations, the user is then asked to complete the same game using the laptop's trackpad. Mouse movement and click data is recorded using the Windows hooking API through the `pyHook` python module.

[2] https://github.com/giuliolovisotto/touchscreen-collector/
[3] allstarpuzzles.com/spotdiff/index.html

Data Format. Each user's folder contains two files, `trackpad.csv` and `mouse.csv`. Each file contains the coordinates, timestamps (in milliseconds) and types of events (mouse movement, click-ups and click-downs). When using the trackpad, some participants used the built-in hardware button, while other "tapped" on the tracking area. The latter results in mostly constant up-down intervals.

6 Eye Movements

Methodology. Eye movement data was collected through an SMI Red500 eye tracker. Before commencing tracking, the eye tracker needs to be calibrated for the current user. Within a single session, each user undergoes two separate collection phases, A and B. In phase A, we perform a calibration for the user before the data collection. In phase B, we instead load the calibration settings from a previous user (i.e., a random calibration for all intents and purposes) before the data collection. To avoid bias, the order of the two phases is randomized, and documented in the `meta.csv` file. Column `note-cal-first` indicates whether phase A was carried out before phase B (1) or vice-versa (0). Column `note-eyes-calibration` indicates the user id of the calibration settings used in phase B.

During the data collection, users were asked to consecutively complete five tasks, with directions being shown on the screen, with each task lasting approximately 180 seconds:

1. Reading an excerpt from the book Game of Thrones;
2. Typing a segment of the previous Game of Thrones' text;
3. Watching a movie trailer for the film "Baby Driver"[4];
4. Complete a web browsing game. The user is presented with a random wikipedia article and is asked to use links within the article to reach the article "University of Oxford"[5];
5. Watch an educational video "What if there was a black hole in your pocket?"[6].

Data Format. Each user's folder contains two subfolders, `calibrated` and `uncalibrated`. Each of these folders then contains a `samples` and `fixations` file for each of the five tasks. The `samples` file contains the raw gaze positions reported by the eye tracker, the pupil diameter and the timestamp in microseconds. The maximum sampling rate is 500 Hz, since the data only contains valid samples the actual sampling rate is usually significantly lower. This can be a result of blinks, poor calibration or the user not looking directly at the screen. The fixations file contains a list of fixations calculated by the eye tracking software. Samples can be matched to fixations by using the sample's timestamp and the fixation's start and end times.

The `meta.csv` file also shows whether each individual was wearing glasses or contact lenses.

[4] https://www.youtube.com/watch?v=D9YZw_X5UzQ&t=14s
[5] https://en.wikipedia.org/wiki/University_of_Oxford
[6] https://www.youtube.com/watch?v=8nHBGFKLHZQ

7 Citing this work

The dataset has been collected and first used for the paper in [1]. If you use this dataset we ask you to cite the paper:

```
@inproceedings{seberz2018,
  title={When Your Fitness Tracker Betrays You:
  Quantifying the Predictability of Biometric Features Across Contexts},
  author={Eberz, Simon and Lovisotto, Giulio and Patan\‘e, Andrea
  and Kwiatkowska, Marta and Lenders, Vincent and Martinovic, Ivan},
  booktitle={Proceedings of the 2018 IEEE Symposium on Security and Privacy},
  year={2018},
  organization={IEEE}
}
```

References

[1] S. Eberz, G. Lovisotto, A. Patanè, M. Kwiatkowska, V. Lenders, and I. Martinovic, "When your fitness tracker betrays you: Quantifying the predictability of biometric features across contexts," in *Proceedings of the 2018 IEEE Symposium on Security and Privacy*. IEEE, 2018.

[2] S. Eberz, N. Paoletti, M. Roeschlin, A. Patan, M. Kwiatkowska, and I. Martinovic, "Broken hearted: How to attack ecg biometrics," in *24th Annual Network and Distributed System Security Symposium*, 2017.

[3] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song, "Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 136–148, 2013.

(a) Set 1, Image 1.

(b) Set 1, Image 2.

(c) Set 2, Image 1.

(d) Set 2, Image 2.

(e) Set 3, Image 1.

(f) Set 3, Image 2.

Figure 2: Set of images used in the Touch Dynamics data collection. Users were asked to swipe between pair of images with subtle differences in order to spot them. Between each pair of images there are 15 differences.

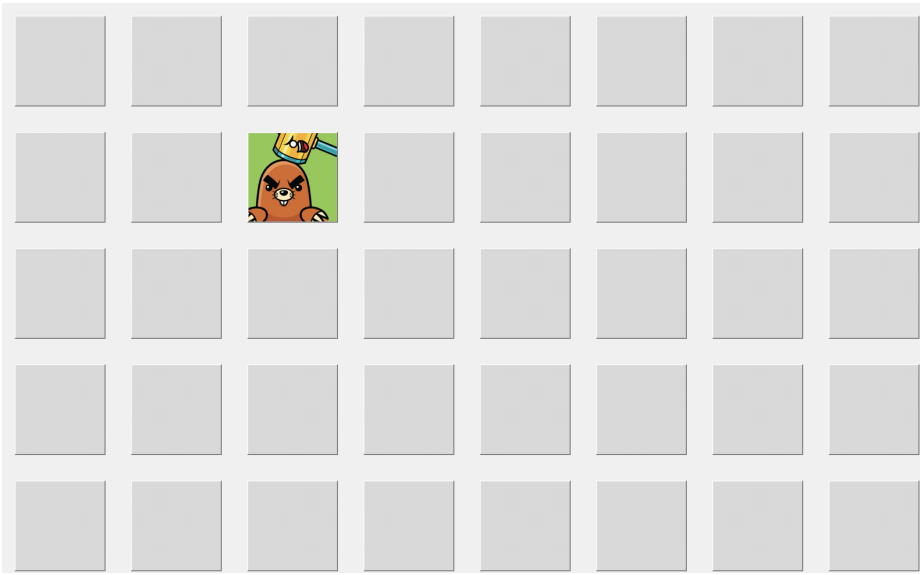

Figure 3: Mole clicking game used to collect mouse movement data.