A Data Set for the Study of Human Locomotion with Inertial Measurements Units

C. Truong[1,2], R. Barrois-Müller[2], T. Moreau[3], C. Provost[2,4], A. Vienne-Jumeau[2],
A. Moreau[2], P.-P. Vidal[2], N. Vayatis[1,2], S. Buffat[2,5,6], A. Yelnik[2,4], D. Ricard[2,7,5],
L. Oudre[8]

[1] CMLA, ENS Paris-Saclay, CNRS, Université Paris-Saclay, 94235, Cachan, France
[2] Cognition and Action Group, Cognac-G, CNRS, Université Paris Descartes, Service de Santé des Armées, Paris, France
[3] Parietal Team, INRIA Saclay, Université Paris-Saclay, Saclay, France.
[4] PRM Department, GH St Louis Lariboisière F. Widal, AP-HP, Paris Diderot University, Paris, France
[5] École d'application du Val-de-Grâce, Service de Santé des Armées, Paris, France
[6] Institut de Recherche Biomédicale des Armées, Service de Santé des Armées, Bretigny-sur-Orge, France
[7] Service de Neurologie, Hôpital d'Instruction des Armées de Percy, Service de Santé des Armées, Clamart, France
[8] L2TI, Université Paris 13, Villetaneuse, France

Communicated by Luis Álvarez

Abstract

This article thoroughly describes a data set of 1020 multivariate gait signals collected with two inertial measurement units, from 230 subjects undergoing a fixed protocol: standing still, walking 10 m, turning around, walking back and stopping. In total, 8.5 h of gait time series are distributed. The measured population was composed of healthy subjects as well as patients with neurological or orthopedic disorders. An outstanding feature of this data set is the amount of signal metadata that are provided. In particular, the start and end time stamps of more than 40,000 footsteps are available, as well as a number of contextual information about each trial. This exact data set was used in [Oudre et al., Template-based step detection with inertial measurement units, Sensors 18, 2018] to design and evaluate a step detection procedure.

Source Code

The source code contains the signals and metadata of the data set described in this article, and is available on this web page[1]. Usage instructions are included in the `README.txt` file of the archive. Additional functions to load and manipulate the data (in Python and R) are provided on a separate repository[2].

Keywords: physiological signals; biomedical data set; multivariate time series

---

[1]https://doi.org/10.5201/ipol.2019.265
[2]https://github.com/deepcharles/gait-data

1 Introduction

The study of human gait is a central problem in medical research with far-reaching consequences in the public health domain. This complex mechanism can be altered by a wide range of pathologies (such as Parkinson's disease, arthritis, stroke,...), often resulting in a significant loss of autonomy and an increased risk of fall. Understanding the influence of such medical disorders on a subject's gait would greatly facilitate early detection and prevention of those possibly harmful situations. To address these issues, clinical and bio-mechanical researchers have worked to objectively quantify gait characteristics. A natural strategy is to measure a patient's movements while walking, compute a number of features from the collected time series and explore significant connections between those features and the patient's medical condition. To that end, gait is described as a succession of footsteps or gait cycles, which are further decomposed into a stance phase and a swing phase (see for instance [11]). During the stance, a sequence of gait events takes place, based on the foot position (namely, heel-strike, toe-strike, heel-off, toe-off). The ability to identify, in a signal, the various cycles, phases and events of the gait, is essential to the understanding of human walking. In the literature, numerous studies have been proposed, using many different measurement apparatus: surface electromyography sensors [6], instrumented treadmills [12], motion capture camera [9], etc. In this paper, as in several other published works [3, 2, 5, 13, 14], we resort to foot-worn Inertial Measurement Units (IMUs) to measure subjects' movements. Those sensors were chosen because they are easy to deploy in routine clinical practice, and therefore allowed researchers to collect data from a large population. Precisely, a total of 230 subjects have been monitored while undergoing a given protocol during one or several clinical consultations. The protocol consisted in performing simple activities (standing still, walking 10 m, turning around and walking back 10 m) in a fixed order.

Strikingly, compared to the volume of published articles, very little of the collected data is freely available and documented enough so as to be reusable. The objective of open and curated sets of gait data is twofold. First, clinicians would be able to test and fairly compare clinical hypotheses, such as the discriminative power of walking patterns in a faller/non-faller population [4, 5]. Second, bio-engineers would be able to design and measure the accuracy of algorithmic procedures. In gait analysis, such procedures include, for instance, methods to detect, in a signal, the temporal boundaries (i.e. the starts and ends) of footsteps [14]. Several initiatives have emerged to achieve both objectives. A well-known example of freely available gait signals is the Daphnet data set [1], which consists of more than 8 h of time series, recorded with IMUs, from patients with Parkinson's disease. The authors also provide the start and end time stamps of a certain type of event, called "gait freeze". In the same way, the HuGaDB database [8] contains 10 h of signals, from IMUs and electromyography sensors. The associated time series are annotated with the activities performed by the 18 participants, for instance walking, going up stairs, etc. Similarly, in [7, 10], 27 subjects have been monitored with the inertial sensor contained in a smartphone, resulting in more than 3 h of signal. Here, only the number of footsteps per trial is given by the authors. In contrast to existing data sets, the one we describe includes the exact start and end time stamps of all footsteps recorded by the sensors (more than 40,000 in total). Overall, our data set contains around 8.5 h of gait signals, collected with foot-worn IMUs from 230 subjects. To the best of our knowledge, this is the largest data set freely available, in terms of participants and footstep annotations. To illustrate the richness of this data set, several articles have already used part of it, in computer science and clinical research [3, 2, 5, 13]. One article has also used this *exact* data set [14].

Contributions. This paper describes a set of 1020 time series, each associated with a number of contextual metadata. In detail, we distribute around 8.5 h of gait signals, collected from 230 subjects performing a sequence of simple movements (such as standing, walking and turning) in a fixed order.

Pathology group	Number of subjects	Number of trials	Gender (M/F)	Age (years)	Height (cm)	Weight (kg)
Healthy	52	242	35/17	36.4 (20.6)	173.4 (10.8)	70.7 (12.2)
Orthopedic	53	243	26/27	60.1 (19.3)	169.2 (10.2)	77.7 (16.8)
Neurological	125	535	80/45	61.5 (13.2)	169.8 (8.7)	72.7 (15.6)
Total	230	1020	141/89	55.5 (19.6)	170.5 (9.7)	73.4 (15.4)

Table 1: Participants' characteristics. For the age, height and weight, means and standard deviations are displayed.

The measured population is composed of healthy subjects as well as patients with neurological and orthopedic disorders. One outstanding feature of this data set is the quantity of signal annotations that are provided. In particular, more than 40,000 footsteps have been manually annotated, meaning that the start and end time stamps are given. By its size, richness and quality of annotation, this data set constitutes an important contribution to the field of human motion studies. In addition to the data files and this article, code snippets to access, visualize and perform basic analysis are available online (github.com/deepcharles/gait-data[3]) for several standard programming languages.

2 Acquisition

This section describes the participating population as well as the acquisition method and the measurement apparatus.

2.1 Participants

The data have been collected between April 2014 and October 2015 by monitoring healthy (control) subjects and patients from the following medical departments: Service de chirurgie orthopédique et de traumatologie de l'Hôpital Européen Georges Pompidou, Assistance Publique des Hôpitaux de Paris, Service de médecine physique et de réadaptation de l'Hôpital Fernand Widal, Assistance Publique des Hôpitaux de Paris, Service de neurologie de l'Hôpital d'Instruction des Armées du Val de Grâce, Service de Santé des Armées. The study was validated by a local ethic committee (Comité de Protection des Personnes Île de France II, CPP 2014-10-04 RNI) and both patients and control subjects gave their written consent to participate. Healthy subjects had no known medical impairment. Non-healthy participants are divided into two groups, depending on the pathology: the orthopedic group and the neurological group. The orthopedic group is composed of 2 cohorts of distinct pathologies: lower limb osteoarthrosis and cruciate ligament injury. The neurological group is composed of 4 cohorts: hemispheric stroke, Parkinson's disease, toxic peripheral neuropathy and radiation induced leukoencephalopathy. For the data set that is presented here, 230 subjects participated (141 males and 89 females). Participants' characteristics are provided in Table 1.

2.2 Protocol and Equipment

All subjects underwent the same protocol which is now described. First, two IMUs (for Inertial Measurement Unit) that recorded accelerations and angular velocities were attached to the body: one on each foot (dorsal face)[4]. All signals have been acquired at 100 Hz with two brands of IMUs: **XSens$^{\text{TM}}$** (autonomy range ±2000 deg/s, dynamic accuracy roll/pitch 0.75 deg RMS, dynamic accuracy heading 1.5 deg 6 h, device dimension $47 \times 30 \times 13$ mm, weight 16 g, acceleration range ±160 m/s^2, angular velocity RMS) and **Technoconcept$^®$** (I4 Motion, autonomy 4 h Li-Ion battery, device dimensions 4.9 cm $\times$ 3.8 cm $\times$ 1.9 cm, acceleration range ±6 g, angular velocity range ±500

---

[3]https://github.com/deepcharles/gait-data

[4]Two other sensors have also been used, one on the lower back (L4L5 vertebra) and one on the forehead, but there are not considered in the remaining of this article.
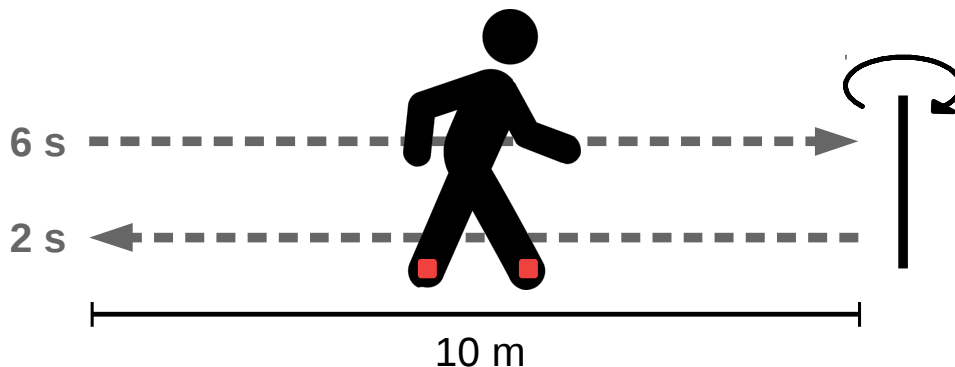
Figure 1: Schematic view of the protocol. Red squares indicate the locations of the IMUs.

deg/s, sampling frequency 100 Hz, and angular velocity measurement error < 1 deg/s). Both were attached to the body with manufacturer-designed adhesive straps. Second, after sensor fixation, participants were asked to perform the following sequence of activities:

- stand for 6 s,

- walk 10 m at preferred walking speed on a level surface to a previously shown turn point,

- turn around (without previous specification of a turning side),

- walk back to the starting point,

- stand for 2 s.

Subjects walked at their comfortable speed with their shoes and without walking aid. This protocol is schematically illustrated in Figure 1. Subjects executed the described protocol between 1 and 27 times. On average, a protocol lasted 30.1 s, with a maximum of 186.4 s and a minimum of 11.7 s.

3   Data Description

This section presents in detail the collected data, which are composed of time series associated with metadata. The file format is also described.

3.1   Time Series

Each IMU records its acceleration and angular velocity in the (X, Y, Z, V) set of axes defined in Figure 2. The V axis is aligned with gravity, while the X, Y and Z axes are attached to the sensor, as shown in Figures 2-b and 2-c. In total, the two IMUs that the participants wore provided $\mathbb{R}^{16}$-valued signals, sampled at 100 Hz. In this setting, each dimension is defined by a sensor ("R" for right foot, "L" for left foot), the signal type ("A" for acceleration, "R" for angular velocity) and the axis ("X", "Y", "Z" or "V"). For instance, "LRX" denotes the angular velocity around the X-axis of the left foot sensor. Accelerations are given in m/s^2 and angular velocities, in deg/s. An illustrative signal example is displayed in Figure 3 (only 8 dimensions are shown out of the available 16). The "flat part" at the beginning of each dimension is the result of the participants standing still for a few seconds before walking (see protocol description). The same behaviour can be seen at the end of each dimension, though for a quite smaller duration. In total, the data set contains more than 8.5 hours of gait signals.
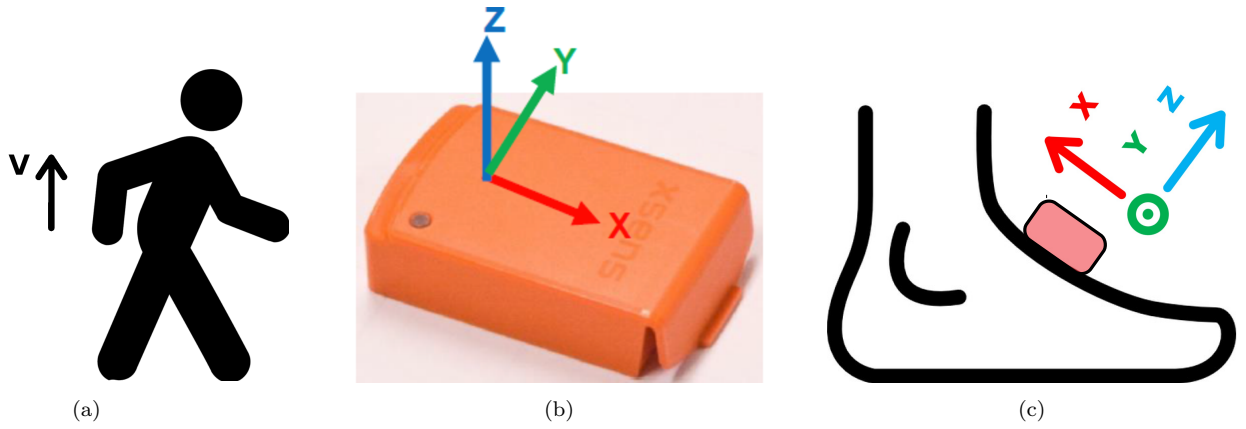
Figure 2: Axes definition: (a) vertical axis, (b) X, Y and Z axes on a sensor, (c) X, Y and Z axes sensor once placed on a foot.

(a) Acc. along the (Ox) axis (RAX)

(b) Rot. around the (Ox) axis (RRX)

(c) Acc. along the (Oy) axis (RAY)

(d) Rot. around the (Oy) axis (RRY)

(e) Acc. along the (Oz) axis (RAZ)

(f) Rot. around the (Oz) axis (RRZ)

(g) Vertical acc. (RAV)
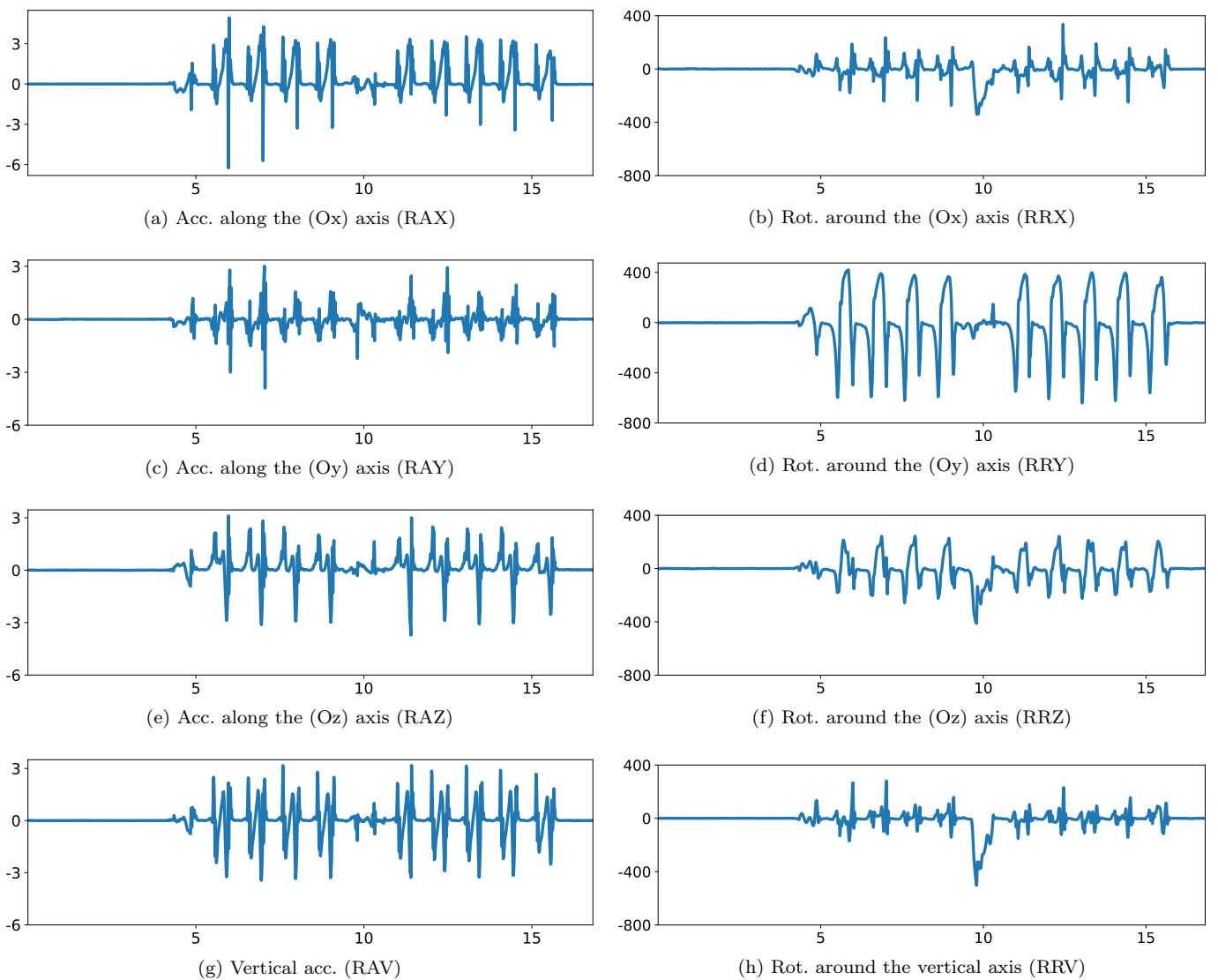
(h) Rot. around the vertical axis (RRV)

Figure 3: Signal examples. Accelerations (acc.) and angular velocities (rot.) for the right foot sensor are shown. Time is in second (s); accelerations are in m/s²; angular velocities are in deg/s.

(a) Gait cycle definition
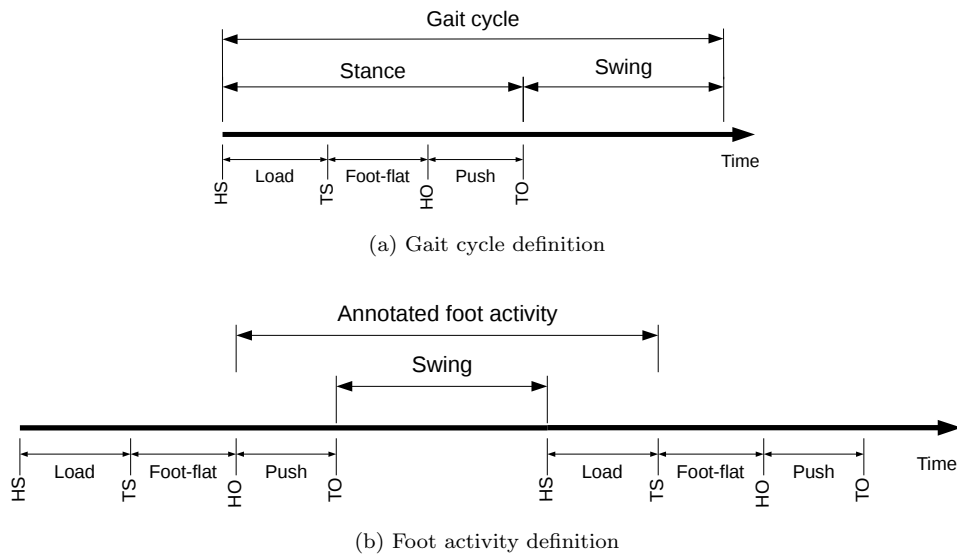

(b) Foot activity definition

Figure 4: (a) Medical definition of a gait cycle [11]. HS, TS, HO and TO respectively stand for heel-strike, toe-strike, heel-off and toe-off. (b) Adopted definition of the annotated foot activity, which is the period during which the foot is moving.

3.2   Metadata

A number of metadata are provided with the signals. Overall there are 16 variables, either numerical or categorical. One of the most important metadata consists in the footstep annotations, manually provided by medical experts. In gait studies, a footstep (or a gait cycle) is formally defined as the period between two heel-strikes of the same foot (see for instance [11]). A gait cycle is further decomposed into a stance phase, during which the foot touches the ground, and a swing phase, during which the foot does not touch the ground. The stance phase is also described as the succession of the following four gait events: heel-strike (HS), toe-strike (TS), heel-off (HO) and toe-off (TO). This decomposition is schematically displayed in Figure 4-a. In this data set, as in [14], a slightly modified definition of a footstep is considered: a footstep is the period during which the foot is moving. Those footsteps are separated by periods when the foot is still and flat on the floor. Therefore, in our setting, a footstep starts with a HO and ends with the following TS of the same foot. The relationship between the medical definition of a gait cycle and its modified version is illustrated in Figure 4-b. In the remaining of the article, to avoid any confusion, the period between a HO and a TS of the same foot is called a "foot activity period". All available metadata are now listed.

1. **Subject** (from 1 to 230). Number of the subject.

2. **Trial** (from 1 to 27). Number of the trial.

3. **Code.** Unique identifier for the trial. It is equal to "**Patient**-**Trial**" (for instance, "10-2" for the subject n° 10 and the trial n° 2).

4. **Age** (in years).

5. **Gender.** Male ("M") or female ("F").

6. **Height** (in meters).

7. **Weight** (in kilograms).

8. **BMI** (in kg/m^2). Body mass index.

(a) Complete signals
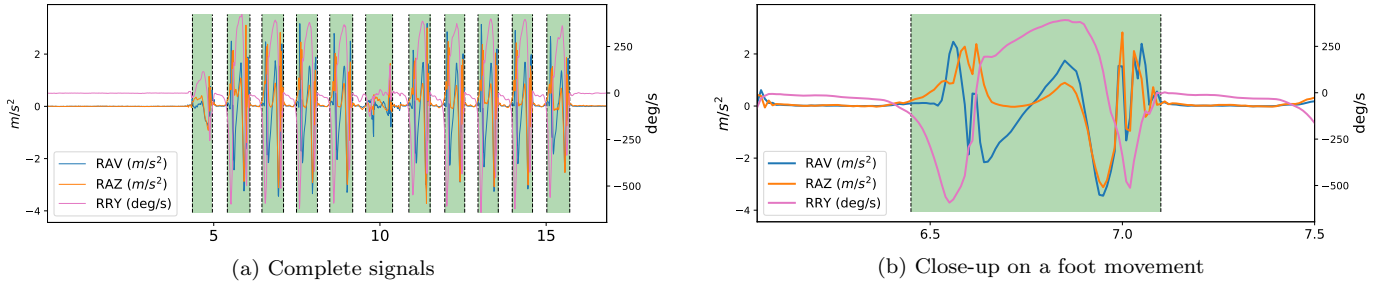
(b) Close-up on a foot movement

Figure 5: Signal example with foot activity annotations (here, **RightFootActivity**). (a) For the right foot sensor, the vertical acceleration (RAV), the Z-axis acceleration (RAZ) and the Y-axis angular velocity (RRY) are shown. (b) A close-up on a single foot movement is displayed.

9. **Laterality.** Subject's "footedness" or "foot to kick a ball" ("Left", "Right" or "Ambidextrous").

10. **Sensor.** Brand of the IMU used for the recording ("XSens" or "TCon").

11. **WalkedDistance** (in meters). Crude visual assessment of the distance walked by the subject from the start to the turn (therefore half the total walked distance).

12. **WalkingSpeed** (in km/h). Average walking speed of the subject during the trial (based on **WalkedDistance**).

13. **PathologyGroup.** Grouping described in [14] and based on the subject's pathology. This variable takes value in {"Healthy", "Orthopedic", "Neurological"}.

14. **IsControl.** Whether the subject is a control subject ("Yes" or "No").

15. **LeftFootActivity.** List $\left[[s_1^{\text{left}}, e_1^{\text{left}}], [s_2^{\text{left}}, e_2^{\text{left}}], \ldots\right]$ of start (HO) and end (TS) *indexes* of the successive activity periods of the left foot.

16. **RightFootActivity.** List $\left[[s_1^{\text{right}}, e_1^{\text{right}}], [s_2^{\text{right}}, e_2^{\text{right}}], \ldots\right]$ of start (HO) and end (TS) *indexes* of the successive activity periods of the right foot.

A few general comments can be made about those metadata.

- *Any* of the described variables can take the value "NC" which stands for "Not Communicated". This label replaces missing data and depending on the variable may affect up to 2% of the database. Also, because a few subjects have undergone multiple trials over several months, the variables **Age**, **Weight**, **Height** and **BMI** can change, for a given participant, between trials.

- For the **PathologyGroup** variable, healthy (or control) subjects had no known medical impairment. The orthopedic group is composed of 2 cohorts of distinct pathologies: lower limb osteoarthrosis and cruciate ligament injury. The neurological group is composed of 4 cohorts: hemispheric stroke, Parkinson's disease, toxic peripheral neuropathy and radiation.

- All foot activity, either of the left foot (**LeftFootActivity**) or of the right foot (**RightFootActivity**), was manually annotated by specialists using a software that displayed the signals from the relevant sensor (left or right foot) and allowed to point with the mouse the starts and ends. An illustrative example is provided in Figure 5: the repeated patterns correspond to periods when the foot is moving. During the non-annotated periods, the foot is flat and not moving and the signals are constant. Overall, this data set contains 40,465 annotated periods (20,225 for the left foot, and 20,240 for the right foot). More statistics can be found in Table 2.

Pathology group	Total count	Counts per trial	Duration	Walking speed
Healthy	7078	29.2 (4.7)	0.71 (0.09)	2.4 (0.5)
Neurological	25228	47.2 (17.4)	0.80 (0.16)	1.6 (0.5)
Orthopedic	8159	33.6 (13.1)	0.74 (0.13)	1.7 (0.5)
Total	40465	39.7 (16.4)	0.77 (0.15)	1.8 (0.6)

Table 2: Counts and durations (in seconds) of all gait activity (union of **LeftFootActivity** and **RightFootActivity**) and walking speeds (in km/h) for each pathology group (**PathologyGroup**). Means and standard deviations are given.
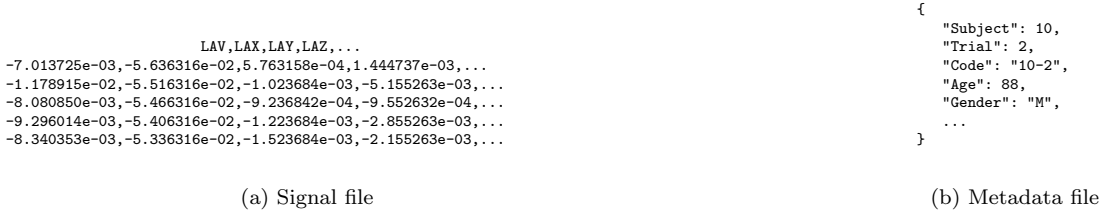
```
                    LAV,LAX,LAY,LAZ,...
-7.013725e-03,-5.636316e-02,5.763158e-04,1.444737e-03,...
-1.178915e-02,-5.516316e-02,-1.023684e-03,-5.155263e-03,...
-8.080850e-03,-5.466316e-02,-9.236842e-04,-9.552632e-04,...
-9.296014e-03,-5.406316e-02,-1.223684e-03,-2.855263e-03,...
-8.340353e-03,-5.336316e-02,-1.523684e-03,-2.155263e-03,...
```

```
{
    "Subject": 10,
    "Trial": 2,
    "Code": "10-2",
    "Age": 88,
    "Gender": "M",
    ...
}
```

(a) Signal file   (b) Metadata file

Figure 6: Excerpts from a signal file (ending in .csv) and a metadata file (ending in .json).

3.3   Data Format

Data are distributed in universal data structures which are supported by all modern programming languages, namely Comma-Separated Values (CSV) and JavaScript Object Notation (JSON). In detail, each trial is associated with two files. The first one has a .csv extension and contains the time series. The second one has a .json extension and contains the metadata. Files are identified by the **Code** variable which uniquely determines a trial and is equal to "**Patient-Trial**". For instance, the trial n° 2 of subject n° 10 is associated with the two following files: 10-2.csv and 10-2.json. As a result, the complete data set has 2040 files, equally distributed in .csv and .json files. A signal file (ending in .csv) has $T + 1$ rows, where T is the number of samples, and $D = 16$ columns. The additional row, located at the beginning of the file, is a header and contains the column names, separated by commas ("LAV,LAX,LAY,LAZ,...", see Section 3.1). The columns are organized according to the lexicographical ordering. Accelerations and angular velocities are respectively expressed in m/s² and deg/s, both with a six-digit numeric precision. A metadata file contains the names and values of the metadata described in Section 3.2 and follows the JSON format (www.json.org[5]). In a nutshell, each file is an unordered set, enclosed by braces, of name/value pairs. Pairs are separated by commas and are arranged as follows: "name: value" (for instance, "**Age**: 54"). Excerpts of such files are displayed in Figure 6.

3.4   Data Availability

This data set is distributed under a Creative Commons CC-BY-NC-SA license (creativecommons.org[6]).

4   Conclusion

In this article, we have described a set of 1020 time series, each associated with contextual metadata. Overall, 8.5 h of gait signals, collected from 230 subjects following a fixed protocol, are provided. The measured population is composed of healthy subjects and patients with neurological or orthopedic disorders. In addition, the metadata contain the start and end time stamps of more than 40,000 gait events. This data set can be used to test and compare clinical hypotheses [2, 5] as well as assess the performance of algorithmic procedures, step detection [14] for instance. The data are made available under a CC-BY-NC-SA license, in universal file formats (JSON and CSV). Code snippets to access,

---

[5]https://www.json.org/
[6]https://creativecommons.org/licenses/by-nc-sa/3.0/

visualize and perform basic analysis are available online (github.com/deepcharles/gait-data[7]) for several standard programming languages.

References

[1] M. Bächlin, M. Plotnik, D. Roggen, I. Maidan, J. M. Hausdorff, N. Giladi, and G. Tröster, *Wearable Assistant for Parkinson's Disease Patients With the Freezing of Gait Symptom*, IEEE Transactions on Information Technology in Biomedicine, 14 (2010), pp. 436–446. https://doi.org/10.1109/TITB.2009.2036165.

[2] R. Barrois-Müller, T. Gregory, L. Oudre, T. Moreau, C. Truong, A. Aram Pulini, A. Vienne, C. Labourdette, N. Vayatis, S. Buffat, A. Yelnik, C. de Waele, S. Laporte, P.-P. Vidal, and D. Ricard, *An automated recording method in clinical consultation to rate the limp in lower limb osteoarthritis*, PLoS One, 11 (2016), p. e0164975. https://doi.org/10.1371/journal.pone.0164975.

[3] R. Barrois-Müller, L. Oudre, T. Moreau, C. Truong, N. Vayatis, S. Buffat, A. Yelnik, C. de Waele, T. Gregory, S. Laporte, P. P. Vidal, and D. Ricard, *Quantify osteoarthritis gait at the doctor's office: a simple pelvis accelerometer based method independent from footwear and aging*, Computer Methods in Biomechanics and Biomedical Engineering, 18 Suppl 1 (2015), pp. 1880–1881. https://doi.org/10.1080/10255842.2015.1072414.

[4] R. Barrois-Müller, D. Ricard, L. Oudre, L. Tlili, C. Provost, A. Vienne, P.-P. Vidal, S. Buffat, and A. Yelnik, *Étude observationnelle du demi-tour à l'aide de capteurs inertiels chez les sujets victimes d'AVC et relation avec le risque de chute*, Neurophysiologie Clinique/Clinical Neurophysiology, 46 (2016), p. 244. https://doi.org/10.1016/j.neucli.2016.09.019.

[5] ——, *Observational study of 180 turning strategies using inertial measurement units and fall risk in poststroke hemiparetic patients*, Frontiers in Neurology, 8 (2017). https://dx.doi.org/10.3389/fneur.2017.00194.

[6] G. Bovi, M. Rabuffetti, P. Mazzoleni, and M. Ferrarin, *A multiple-task gait analysis approach: kinematic, kinetic and EMG reference data for healthy young and adult subjects*, Gait & Posture, 33 (2011), pp. 6–13. https://doi.org/10.1016/j.gaitpost.2010.08.009.

[7] A. Brajdic and R. Harle, *Walk detection and step counting on unconstrained smartphones*, in Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing, Zurich, Switzerland, 2013, pp. 225–234. https://doi.org/10.1145/2493432.2493449.

[8] R. Chereshnev and A. Kertész-Farkas, *HuGaDB: human gait database for activity recognition from wearable inertial sensor networks*, in Proceedings of the International Conference on Analysis of Images, Social Networks and Texts (AIST), Moscow, Russia, 2017, pp. 131–141. https://doi.org/10.1007/978-3-319-73013-4_12.

[9] D. A. dos Santos, C. A. Fukuchi, R. K. Fukuchi, and M. Duarte, *A data set with kinematic and ground reaction forces of human balance*, PeerJ, 5 (2017), p. e3626. https://doi.org/10.7717/peerj.3626.

---

[7]https://github.com/deepcharles/gait-data

[10] R. Harle and A. Brajdic, *Research data supporting "Walk detection and step counting on unconstrained smartphones"*, 2017. https://doi.org/10.17863/CAM.12982.

[11] B. Mariani, H. Rouhani, X. Crevoisier, and K. Aminian, *Quantitative estimation of foot-flat and stance phase of gait using foot-worn inertial sensors*, Gait & Posture, 37 (2013), pp. 229–234. https://doi.org/10.1016/j.gaitpost.2012.07.012.

[12] J. K. Moore, S. K. Hnat, and A. J. van den Bogert, *An elaborate data set on human gait and the effect of mechanical perturbations*, PeerJ, 3 (2015), p. e918. https://doi.org/10.7717/peerj.918.

[13] L. Oudre, R. Barrois-Müller, T. Moreau, C. Truong, R. Dadashi, T. Grégory, D. Ricard, N. Vayatis, C. De Waele, A. Yelnik, and P.-P. Vidal, *Détection automatique des pas à partir de capteurs inertiels pour la quantification de la marche en consultation*, Neurophysiologie Clinique/Clinical Neurophysiology, 45 (2015), p. 394. https://doi.org/10.1016/j.neucli.2015.10.015.

[14] L. Oudre, R. Barrois-Müller, T. Moreau, C. Truong, A. Vienne-Jumeau, D. Ricard, N. Vayatis, and P.-P. Vidal, *Template-based step detection with inertial measurement units*, Sensors, 18 (2018). https://dx.doi.org/10.3390/s18114033.