whuGAIT Dataset

The **whuGAIT** dataset contains inertial data of 118 subjects collected by smartphones. It consists of 8 sub-datasets. Datasets 1 to 4 are for person identification, and Datasets 5 and 6 are for authentication. Datasets 7 and 8 are used for separating walking data from non-walking data. Among the 118 subjects, 20 subjects collect data in two days, with thousands of samples for each, and the other 98 subjects collect data in one day, with hundreds of samples for each. A data sample contains the 3-axis accelerometer data and the 3-axis gyroscope data. The sampling rate of the sensors is 50 Hz.

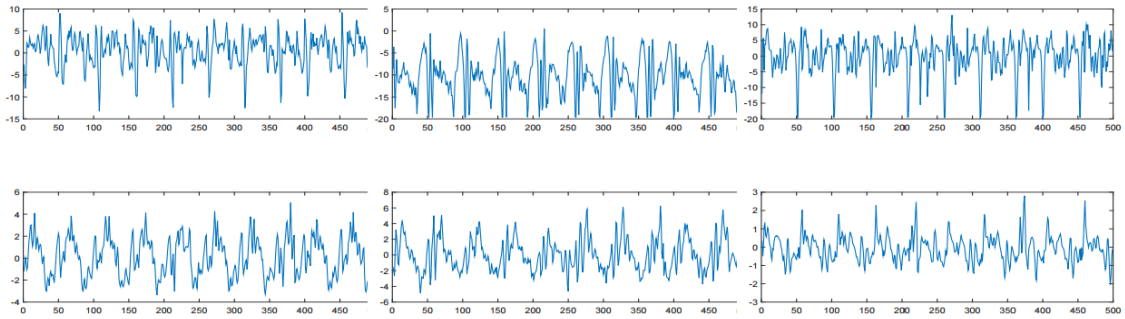

Figure 1 An example of the data sample

- **Datasets for Person Identification & Authentication**
  - ***Dataset #1***

This dataset is collected on 118 subjects. The collected gait data have been annotated into every two steps. Meanwhile, a single sample is interpolated into a fixed length of 128 (using Linear Interpolation function). In order to enlarge the scale of the dataset, we make a one-step overlap between two neighboring samples for all subjects. In this way, a total number of 36,884 gait samples are collected. We use 33,104 samples for training, and the rest 3,740 for test.

  - ***Dataset #2***

This dataset is collected on 20 subjects. We also divide the gait curve into two-step samples and interpolate them into the same length of 128. As each subject in this dataset has a much larger amount of data as compared to the that in Dataset #1, we do not make overlap between the samples. Finally, a total number of 49,275 samples are collected, in which 44,339 samples are used for training, and the rest 4,936 for test.

  - ***Dataset #3***

This dataset is collected on the same 118 subjects as in Dataset #1. Different from Dataset #1, we divide the gait curve by using a fixed time length, instead of a step length. Exactly, we collect a sample with a time interval of 2.56 seconds. While the frequency of data collection is 50Hz, the length of each sample is also 128. Also, we make an overlap of 1.28 seconds to enlarge the dataset. A total number of 29,274 samples are collected, in which 26,283 samples are used for training, and the rest 2,991 for test.

  - ***Dataset #4***

This dataset is collected on 20 subjects. We also divide the gait curve in an interval of 2.56 seconds. We make no overlap between the samples. Finally, a total number of 39,314 samples are collected, in which 35,373 samples are used for training, and the rest 3,941 for test.

  - ***Dataset #5***

This dataset is used for authentication. It contains 74,142 authentication samples of 118 subjects, where the training set is constructed on 98 subjects and the test set is constructed on the

other 20 subjects. There are 66,542 samples and 7,600 samples for training and test, respectively. Each authentication sample contains a pair of data sample that are from two different subjects or one same subject. The data sample consists of a 2-step acceleration and gyroscopic data, which are interpolated in the way as described in Dataset #1 and Dataset #2. The two data samples are horizontally aligned to create an authentication sample.

- ***Dataset #6***

This dataset is also used for authentication. The authentication samples are constructed as the same as in Dataset #5. The only difference is that, in authentication sample construction, two data samples from two subjects are vertically aligned instead of horizontally aligned.

The information of datasets #1 - #6 are detailed as below:

TABLE III
DETAIL INFORMATION OF THE SIX DATASETS.

Dataset Name	Usage	Number of Subjects	Time-fixed or Interpolation	Overlap in Sampling	Samples for Training	Samples for Test	Alignment
Dataset #1	Classification	118	Interpolation	1 step	33,104	3,740	N/A
Dataset #2	Classification	20	Interpolation	0	44,339	4,936	N/A
Dataset #3	Classification	118	Time-fixed	1 step	26,283	2,991	N/A
Dataset #4	Classification	20	Time-fixed	0	35,373	3,941	N/A
Dataset #5	Authentication	118	Interpolation	1 step	66,542	7,600	Horizontal
Dataset #6	Authentication	118	Interpolation	1 step	66,542	7,600	Vertical

*Note: there is no overlap between the training sample and the test sample for all datasets.

Datasets for Gait Data Segmentation

- ***Dataset #7***

We took 577 samples from 10 subjects, with data shape of 6×1024. 519 of them were used for training and 58 were used for testing. Both the training and testing datasets have data from these 10 subjects. There is no overlap between the training sample and the test sample.

- ***Dataset #8***

We took 1,354 samples from 118 subjects, with data shape of 6×1024. In order to make the training and test data come from different subjects, we use 1022 samples from 20 subjects as training data and 332 samples from other 98 subjects for testing.

The construction and content of dataset #7 and dataset #8 is shown as below:

TABLE IV
DETAIL INFORMATION OF THE GAIT-DATA EXTRACTION DATASETS

Dataset Name	Number of Subjects	Samples for Training	Samples for Test
Dataset #7	10	519	58
Dataset #8	118	1022	332

Download

BaiduYun: https://pan.baidu.com/s/1epYd7YENDHAwLQ-V7qeaTQ code: gnoi
Or
OneDrive: https://1drv.ms/f/s!AittnGm6vRKLyh3yWS7XaXfyUNQp

Citation (If you use this dataset in your own work)

Zou, Q., Wang, Y., Zhao, Y., Wang, Q., and Li, Q. (2018). Deep Learning Based Gait Recognition Using Smartphones in the Wild. *arXiv preprint arXiv:1811.00338*.

Contact

Dr. Qin Zou (If any problem accessing or using the dataset, email to qzou@whu.edu.cn)