University at Buffalo Keystroke Dataset

This dataset is collected from 148 subjects in 3 separate laboratory sessions (session 0, session 1 and session 2). Each session takes about 50 minutes, and contains about 5.7k keystrokes. There are 28 days in average time intervals between sessions. Four different types of keyboards are used across sessions.

There are two subsets of users divided based on the keyboards they use: baseline subset (ID from 001 to 075), with 75 users using the same type of keyboard across 3 sessions; and rotation subset (ID from 075 to 148), with 73 users using 3 different types of keyboard across 3 sessions.

The data are the result of each subject performing the following tasks in a laboratory (for each session):

Task 0: Transcription of Steve Jobs' Commencement Speech split in three pieces

Task 1: Free Text Questions

a. 2 survey style questions, 1 scene description
b. series of routine work in order to mimic the realistic daily working scenario, e.g., checking email, sending email and web surfing.

There is one folder for each session. Under the session folder, there are 2 folders named baseline subset and rotation subset. Under the subset folder there are 2 files for each subject, task 0 and task 1, named by the subject ID. The naming rules are shown as follows.

File ID				
Sequence No.	1-2-3rd	4th	5th	6th
Assignment	User ID	Session #	Keyboard type code	Task #
Value	001 ~ 148	0 ~ 2	0 ~ 3	0 ~ 1

Keyboard type code	
Code	**Keyboard**
0	Lenovo keyboard
1	HP wireless keyboard
2	Microsoft Keyboard
3	Apple Bluetooth keyboard

For example, file 123121.txt stores the data collected from user 123 in session 1, task 1 using Microsoft keyboard.

Within each file, there are multiple lines, each corresponding to one typing event.

The format of each line is explained as follows, using the lines we take from 123121.txt as an example:

```
A KeyDown 63578429792961

A KeyUp 63578429793054

M KeyDown 63578429793257

M KeyUp 63578429793382

…
```

where the entities are separated by space in each line. The first entity is the name of the key. The second entity is the key event (key down or key up). The third entity is the time stamp in milliseconds.

Gender information is included in file “GenderInfo.txt”. For more details, please refer to the reference paper.

The dataset is intended for research purposes only and as such cannot be used commercially. In addition, reference must be made to the following publication when this dataset is used in any academic and research reports.

Reference

Yan Sun, Hayreddin Ceker and Shambhu Upadhyaya, “Shared Keystroke Dataset for Continuous Authentication”, *8th IEEE International Workshop on Information Forensics and Security, Abu Dhabi, UAE, December 2016.*