Instructions for accessing and using stereotypical motor movement (SMM) dataset to replicate and extend results of Goodwin et al published in the proceedings of Ubicomp' 14.

*This document provides detail descriptions of the **SMM dataset** in addition to step-by-step instructions for replicating results presented in Goodwin et al (2014). Please cite the paper in any presentations and/or publications that result from accessing this data.*

Table of contents

1. Download the package of paper, dataset, source code, and results

1.1 Goodwin's paper in Ubicomp' 14

Matthew S. Goodwin, Marzieh Haghighi, Qu Tang, Murat Akcakaya, Deniz Erdogmus, and Stephen Intille. 2014. Moving towards a real-time system for automatically recognizing stereotypical motor movements in individuals on the autism spectrum using wireless accelerometry. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14). ACM, New York, NY, USA, 861-872. DOI=10.1145/2632048.2632096 http://doi.acm.org/10.1145/2632048.2632096

1.2 Package of dataset, MATLAB, R code, and results

https://bitbucket.org/mhealthresearchgroup/stereotypypublicdataset-sourcecodes/downloads

2. Package structure

- Root Path
  - data
    - Study 1
      - URI-00X-MM-DD-YY
    - Study 2
      - 00X-YYYY-MM-DD
  - src
    - MATLAB
    - R
  - results
    - SVM
    - DT

Note that

1) **`data`** folder contains the raw data and annotation files, as well as intermediate .mat files required to replicate results. **`src`** folder contains MATLAB and R source code used in the paper. **`results`** folder stores all results generated by the source code, separately in a SVM classifier folder and Decision Tree (DT) folder.

2) Session name pattern: X represents participant ID, MM is month (01-12), DD is day code (01-31), YY is two digit year code and YYYY is the four digit year code.

3. Dataset details

3.1 Studies and sessions

Study 1 and Study 2 sessions were collected 2-3 years apart.

Participant ID	Study 1	Study 2
1	URI-001-01-18-08 URI-001-01-25-08	001-2010-05-25 001-2010-05-28 001-2010-06-01
2	URI-002-01-18-08 URI-002-01-24-08	002-2010-06-04 002-2011-06-02
3	URI-003-01-18-08 URI-003-02-08-08	003-2010-05-07 003-2011-05-23
4	URI-004-01-17-08 URI-004-02-07-08	004-2010-04-27 004-2010-05-11 004-2011-03-22
5	URI-005-01-16-08 URI-005-02-08-08	005-2010-05-17 005-2011-05-25
6	URI-006-01-15-08 URI-006-01-23-08	006-2010-03-12

3.2 Files in one session

Inside each session folder you can find:

Study type	Study 1	Study 2
Session name	URI-00X-MM-DD-YY	00X-YYYY-MM-DD

Raw data files (in .csv)	`MITes_01_RawCorrectedData_Trunk.RAW_DATA.csv`
	`MITes_08_RawCorrectedData_Left-wrist.RAW_DATA.csv`
	`MITes_11_RawCorrectedData_Right-wrist.RAW_DATA.csv`
		`Wocket_00_RawCorrectedData_Right-Wrist.csv`
		`Wocket_01_RawCorrectedData_Left-Wrist.csv`
		`Wocket_02_RawCorrectedData_Torso.csv`

Annotation files (in .xlsx, .xml)	`Annotator1Stereotypy.annotation.*`
	`Phone.annotation.*`
		`AnnotationPhoneIntervals.xlsx`
		`AnnotationVideo1Intervals.xlsx`
		`AnnotationVideo2Intervals.xlsx`
		`Annotator1Stereotypy.annotation.*`
		`Annotator2Stereotypy.annotation.*`
		`Phone.annotation.*`

MATLAB intermediate files	`preprocessedDataAndLabels.mat`
	`featureVectorAndLabels.mat`
		`preprocessedDataAndLabels.mat`
		
		`featureVectorAndLabels.mat`

R intermediate files	`.Rcache folder`	`.Rcache folder`

Note that:

1) Raw data is stored in `.csv` format. The file naming convention contains information about sensor type, sensor id, and sensor location, separated by "_". For example: in `MITes_01_RawCorrectedData_Trunk.RAW_DATA.csv`, `MITes` is the type of sensor used, `Trunk` is the sensor location, which means `Torso`. The same convention applies to Study 2 directories.
2) Annotation files are stored in several different formats, including .xlsx, .xml, and .csv, with the same contents, just for the sake of users' convenience.
    a) `Annotator1Stereotypy.annotation` or `Annotator2Stereotypy.annotation` are offline annotations, `Phone.annotation` is online annotation.
    b) Files with pattern `Annotation*Intervals.xlsx` (containing the same information with `*.annotation.*` files) are currently used in the preprocessing script for Study 2 data due to legacy reasons, but should be changed to use files `*.annotation.*` in the future.
    c) Study 1 has only one online and one offline annotation while Study 2 has two online and one offline annotations.
    d) Experiments in the paper only used `Annotator1` offline annotations, but one can refer to our previous Ubicomp' 09 paper [1] for discussion about online and offline annotations.

3) `preprocessedDataAndLabels.mat` is the intermediate file used by the scripts to store all the preprocessed data and annotations. `featureVectorAndLabels.mat` is the intermediate file used by the scripts to store all the computed feature sets along with the labels. Users can refer to this file directly using their own classifiers (Detailed description of the format can be found in file `preExp.m` in folder `src/MATLAB`).

3.2.1 Raw data format

The raw data .csv file has four columns representing unix timestamp, raw x value, raw y value, and raw z value, respectively. There is no header row. The raw value is an integer in the range of 0 - 1023 for Wockets and 0 - 512 for Mites, with a dynamic range ±2g for MITes sensors and ±4g for Wockets sensors.

```
1274781775000,475,436,435
1274781775009,474,437,436
1274781775017,474,438,433
…
1274782392242,501,469,415
1274782392253,501,468,416
1274782392263,504,466,418
```

3.2.2 Annotation file format

For **.xml** file,

Annotation data format (*.annotation.xml)
Annotated activity labels are encoded in a human and computer readable XML file. This is stored in the Annotation subdirectory inside the session folder. Some datasets may be labeled by more than one annotator using more than one category or set. Each file uses the XML format.

At the top of the file you may find some comments assigned automatically during generation of this document by a source application.

The first XML node defines the name/session for the dataset, along with the

annotator and other relevant details about this document:

```
<ANNOTATIONS DATASET="HomeLife" ANNOTATOR="JPN" EMAIL="jpn@neu.edu" DESCRIPTION="Labels for activities that occurred at home on the afternoon of Jan, 25, 2011" METHOD="Video" NOTES="Only the participant was at home at the time. Received a phone call and requested audio be deleted between 13:30 and 13:42">
```

The following properties may be defined:

- DATASET: A short name of the protocol to be used in the visualization software
- ANNOTATOR: Identification of the person who performed the annotation
- EMAIL: Annotator's contact information
- DESCRIPTION: A clear, concise description of the protocol/annotation set
- METHOD: Indication of how the annotation was done. Options are
    - direct_observation_paper: start/stop times were written down on paper as activities were performed
    - direct_observation_mobile start/stop times were entered directly into a computer using custom annotation software in real-time
    - video: annotations were coded from a video recording
    - audio: annotations were coded from an audio recording
- NOTES. Any special notes should be included here. In particular, a full description of how the intensity values (if used) were determined should be included here.

The main content of the the annotation.xml file comprises time-stamped label data. Each ANNOTATION node can be assigned a Globally Unique Identifier (GUID) to allow synchronization between multiple annotators.

```
<ANNOTATION GUID="79c0d9a4-589b-4536-9875-4f6e176267b4">
```

The LABEL node can also be assigned a Globally Unique Identifier (GUID) which allows lookup in a database of labels to find the specific definition of the label in use at the time this document was generated

```
<LABEL GUID="44abf157-a692-4235-b7b7-28251ec2d77d">searching for an item</LABEL>
```

The START_DT and STOP_DT date-time nodes receive values in the form [YEAR]-[MO]-[DY] [HR]:[MI]:[SE].[MSe].

```
<START_DT>2011-01-25 11:42:54.396</START_DT>
<STOP_DT>2011-01-25 11:43:02.396</STOP_DT>
```

The RATINGS NODES are optional timestamps used to allow annotators to express how certain they feel the observed label represented the activity they observed. The

SETNAME and GUID can be used to map meaningful labels onto ratings values. A METARATING is also available for applications where a second rating values is desired, such as intensity.

```
<RATINGS GUID="15c28dfe-47bb-487c-a5e8-f5c8f844f864" SETNAME="BLP">
        <RATING TIMESTAMP="2011-01-25 11:42:54.396" VALUE="2" METARATING="3" />
        <RATING TIMESTAMP="2011-01-25 11:43:00.000" VALUE="3" METARATING="3" />
        <RATING TIMESTAMP="2011-01-25 11:43:02.396" VALUE="0" METARATING="2" />
</RATINGS>
```

A final PROPERTIES node contains tracking information about the annotation, such as the annotation set of origin and the time of creation/modification

```
    <PROPERTIES ANNOTATION_SET="HomeLife.JPN" LAST_MODIFIED="2011-01-25 13:24:27.573" DATE_CREATED="2011-01-25
13:24:27.573" />
</ANNOTATION>
```

An example AnnotationData file can be found here: annotation.xml.
Although activity labels are experiment-specific, researchers would be encouraged to reuse the annotation schema of others whenever possible in order to facilitate data sharing. Examples from prior experiments would be featured on the data sharing website.

(Cited from lab inner documentation)

`.xlsx` files are converted from `.xml` files, so user should be able to find the corresponding column name with the node name described above. `.csv` files have a header row that describes the meaning of each column.

3.2.3 MATLAB intermediate file format

1) preprocessedDataAndLabels.mat (see also file `preExp.m` in `src/MATLAB` folder)

This file contains 3 cells for right, left, and torso data and each cell is a matrix (time samples x 4) where the first column is time (starting from 0), second column is acceleration data for x-axis, third column is acceleration data for y-axis, and forth column is acceleration data for z-axis.

2) featureVectorAndLabels.mat (see also file `preExp.m` in `src/MATLAB` folder)

This file is structured with the following fields:

- `fv` with dimension: `number of features × dimension of features(stockwell)`
- `fvt` with dimension: `number of features × (dimension of features + 1)(stockwell + time)`
- `fvNew` with dimension: `number of features × dimension of features(stockwell+ baseline features)`
- `videoLabelvec` with dimension: `number of features × offline(video) labels`
- `phoneLabelvec` with dimension: `number of features × online(phone) labels`

4. How to replicate results

The source code is in MATLAB and R. MATLAB was used to preprocess data, prepare feature sets, run experiments with SVM, and compute all kinematic parameters. R was used to run experiments using decision trees. Before running scripts, make sure your hard disk has more than 8GB free space needed to store all intermediate files (this does not include space required to run MATLAB and R software).

4.1 Prepare run environment

4.1.1 Prepare MATLAB requirement

1) Install MATLAB version R2014+
2) Install Windows SDK 7.1 for .mex file compilation
3) Install LIBLINEAR Package for MATLAB: The source code is already included in the src/MATLAB folder, please follow its own instructions to install (Source code can also be downloaded from: http://www.csie.ntu.edu.tw/~cjlin/liblinear/).

4) Install ITE toolbox for MATLAB: The source code is already included in the src/MATLAB folder, please follow its own instructions to install (Source code can also be downloaded from: https://bitbucket.org/szzoli/ite/downloads).

4.1.2 Prepare R environment

1) Install latest version of RStudio with R version greater than 3.0.2
2) Install latest JRE and set "JAVA_HOME" environment
3) Make sure your system has at least 8G RAM, better to have 16G

4.2 Configure run environment

1) In MATLAB, change to src/MATLAB folder,
    a) Reset all parameter values in `stereotypyParameter.m` to default values.
    b) Change `rootPath` variable in `setRoot.m` file to the path where you unzip this package, make sure it's the root as shown in 2. Package structure instead of any subfolder such as `src`.
    c) Run `preExp.m` script to generate all required intermediate files (`preprocessedDataAndLabels.mat` and `featureVectorAndLabels.mat` files) in the data folder for each subject/session (Please note that running this script on a computer with *8GB/8x Intel Xeon CPU E5-1620 @ 3.70Hz/Windows 8 or Ubuntu 14.04* took approximately half an hour. On a slower computer with less memory it could take substantially longer).
2) Then in R, change to `src/R` folder,

a) Check the current working directory by `getwd()`, make sure it's at the root path of the unzipped package, if not, set correct path by using `setwd()`.

b) Run `init.R` script to install and load all required packages and cache feature sets, all cached feature sets will be stored in `.RData` format in a folder called `.Rcache` which can be found in the current R working directory (Please note that running this script on a computer with *8GB/8x Intel Xeon CPU E5-1620 @ 3.70Hz/Windows 8 or Ubuntu 14.04* took approximately half an hour. On a slower computer with less memory it could take substantially longer).

c) If `rJava` package installation fails, please refer to this post for a solution http://stackoverflow.com/questions/3311940/r-rjava-package-install-failing

4.3 TABLE - 1 & 2 (SMM kinematic parameters and statistics)

In order to compare performance between our different classifiers and feature combinations, we generated kinematic parameters that quantitatively described observed SMM. For each participant we have multiple sessions of data collection, and we denote the total number of sessions as K and the duration of the k^{th} session as T_k. The total duration of pooled data is $\sum_{k=1}^{K} T_k = T$. We define a contiguous time range in which an individual is engaged in SMM as a bout, and denote the accelerometer data corresponding to the n^{th} bout for a specific SMM $m \in$ {rock, flap, flaprock} in the pooled data as ${B_{n_s}}^m(t)$, which is a function of time, $t \in D_{{B^m}_{n_s}}$, where $D_{{B^m}_{n_s}}$ is the duration of the n^{th} bout corresponding to the SMM m, and $s \in$ {right wrist, left wrist, torso} is the location of the accelerometer sensor. $D_{{B^m}_{n_s}}$'s are equal for all sensor locations, s; therefore, for simplicity, we denote them as $D_{{B^m}_n}$. If the total number of m bouts is ${N^m}_s$, than we have a set of bouts $\{ {B_{1_s}}^m(t), \ldots, {B_{N_s}}^m(t) \}$. Moreover, $L_{({B_{n_s}}^m, {B_{(n+1)_s}}^m)}$ is the time delay between the end of a bout ${B_{n_s}}^m$ and the beginning of a bout ${B_{(n+1)_s}}^m$, which we call latency. Note that number of bouts and latencies corresponding to the

m^{th} SMM are equal for different sensor locations. We define the intensity of the n^{th} bout corresponding to the m^{th} SMM as:

$$I^m{}_n = \frac{1}{D_{B_n}} \sum_{t \in D_{B_n}} (\| B_{n_{right\ wrist}}(t) \|^2 + \| B_{n_{left\ wrist}}(t) \|^2 + \| B_{n_{torso}}(t) \|^2)$$

The estimated movement frequency (EMF) in the n^{th} bout corresponding to the m^{th} SMM was defined as:

$$EMF = \frac{total\ number\ of\ moves\ in\ B^m{}_{n_{selected\ sensor}}(t)\ for\ t \in D_{B^m{}_n}}{D_{B^m{}_n}}$$

To calculate EMF, we chose the sensor with the highest variance in acceleration during a specific bout of interest. We then subtracted the mean of the measured acceleration and calculated zero crossings in the resulting signal to count the number of moves.

Proportion of the total duration in which a participant was engaged in a specific SMM m was defined as Engagement Proportion (EP) and calculated as:

$$EP^m = \frac{\sum_{n=1}^{N^m} D_{B^m{}_n}}{T}$$

1) Run script `experiment6p.m` from folder src/MATLAB (Please note that running this script on a computer with *8GB/8x Intel Xeon CPU E5-1620 @ 3.70Hz/Windows 8 or Ubuntu 14.04* took approximately 10 minutes. On a slower computer with less memory it could take substantially longer).

2) Check file `kinectTables.mat` in folder for results. It's a structure with the following 6 fields:

    a) Field `table`: same structure as TABLE 2 in the paper.

    b) Field `table1`: same structure as TABLE 1 in the paper.

c) Field `table2`: anova on kinematic parameters for each participant for each study for each session and across different sessions.

d) Field `table3`: t-test on kinematic parameters of each participant across studies.

e) Field `table4`: pairwise t-test p-values for kinematic parameters for each participant and each study across different sessions.

f) Field `table5`: detailed version of Table 1 (info for each session).

4.4 Experiment 1 (Individual leave-one-session-out)

1) Run script `experiment1.m` from folder `src/MATLAB` (Please note that running this script on a computer with *8GB/8x Intel Xeon CPU E5-1620 @ 3.70Hz/Windows 8 or Ubuntu 14.04* took approximately 10 minutes. On a slower computer with less memory it could take substantially longer).

2) Check file `exp1tableS1S2.mat` in folder `results/SVM` for results of SVM.

3) SVM Results contains 2 cells ({study1},{study2}), each cell contains a matrix with the following dimensions: `3(feature sets) × 24 (each 4 columns for one participant).`

	P1				P2	
Feature set	Accuracy	TPR	FPR	Precision	Accuracy	...
Baseline feature						...
Stockwell feature						...
Combined feature						...

4) Run script `Experiment1_study1_DT.R` and `Experiment1_study2_DT.R` from folder `src/R` (Please note that running this script on a computer with *8GB/8x Intel Xeon CPU E5-1620 @ 3.70Hz/Windows 8 or Ubuntu 14.04* took approximately half an hour for each one. On a slower computer with less memory it could take substantially longer).

5) Check files in folder `results/DT/experiment1` for results of decision tree

6) Each file is the result of a specific feature set on the specific study. The result file has the following naming convention: `{feature set}.{study type}.{experiment timestamp}.csv`. Each file is a `m(participants) × n(metrics)` matrix

4.5 Experiment 2 (Individual cross-validation)

1) Run script `experiment2.m` from folder `src/MATLAB` (Please note that running this script on a computer with *8GB/8x Intel Xeon CPU E5-1620 @ 3.70Hz/Windows 8 or Ubuntu 14.04* took approximately 10 minutes. On a slower computer with less memory it could take substantially longer).

2) Check file `exp2tableS1S2.mat` in folder `results/SVM` for results. The result file has 2 cell ({study1},{study2}), each cell contains a matrix with the following dimensions: `3(feature sets) × 24 (each 4 columns for one participant).`

	P1				P2	
Feature set	Accuracy	TPR	FPR	Precision	Accuracy	...
Baseline feature						...
Stockwell feature						...
Combined feature						...

4.6 Experiment 3 (Leave-one-subject-out)

1) Run script `experiment3.m` from folder `src/MATLAB` (Please note that running this script on a computer with *8GB/8x Intel Xeon CPU E5-1620 @ 3.70Hz/Windows 8 or Ubuntu 14.04* took approximately 20 minutes. On a slower computer with less memory it could take substantially longer).

2) Check file `exp3tableS1S2.mat` in folder `results/SVM` for results. The results file contains two cells ({study1},{study2}), each cell contains a matrix with the following dimensions: `3(feature sets) × 24 (each 4 columns for one participant).`

	P1				P2	
Feature set	Accuracy	TPR	FPR	Precision	Accuracy	...
Baseline feature						...
Stockwell feature						...
Combined feature						...

4.7 Experiment 4 (Train on Study 1, test on Study 2)

1) Run script `experiment4.m` from folder `src/MATLAB` (Please note that running this script on a computer with *8GB/8x Intel Xeon CPU E5-1620 @ 3.70Hz/Windows 8 or Ubuntu 14.04* took approximately 10 minutes. On a slower computer with less memory it could take substantially longer).

2) Check file `exp4table.mat` in folder `results/SVM` for results. The results file contains a matrix with the following dimensions: `3(feature sets) × 24 (each 4 columns for one participant).`

	P1				P2	
Feature set	Accuracy	TPR	FPR	Precision	Accuracy	...
Baseline feature						...
Stockwell feature						...
Combined feature						...

5. Contact us

If you have any question about the dataset or source codes, please contact Dr. Matthew Goodwin: m.goodwin@neu.edu.

6. References

1. Albinali, Fahd, Matthew S Goodwin, and Stephen S Intille. "Recognizing stereotypical motor movements in the laboratory and classroom: a case study with children on the autism spectrum." *Proceedings of the 11th international conference on Ubiquitous computing* 30 Sep. 2009: 71-80.