HASC-PAC2016: Large Scale Human Pedestrian Activity Corpus and Its Baseline Recognition

Haruyuki Ichino
Graduate School of Engineering, Nagoya University, Japan
icchi@ucl.nuee.nagoya-u.ac.jp

Katsuhiko Kaji
Faculty of Information Science, Aichi Institute of Technology, Japan
kaji@aitech.ac.jp

Ken Sakurada
Graduate School of Engineering, Nagoya University, Japan
sakurada@nagoya-u.jp

Kei Hiroi
Institutes of Innovation for Future Society, Nagoya University, Japan
k.hiroi@ucl.nuee.nagoya-u.ac.jp

Nobuo Kawaguchi
Institutes of Innovation for Future Society, Nagoya University, Japan kawaguti@nagoya-u.jp

Ubicomp/ISWC'16 Adjunct , September 12 - 16, 2016, Heidelberg, Germany

ACM 978-1-4503-4462-3/16/09···$15.00
DOI: http://dx.doi.org/10.1145/2968219.2968277

Abstract
Human activity recognition by wearable sensors will enable a next-generation human-oriented ubiquitous computing. However, most of the existing research on human activity recognition is based on a small number of subjects, and lab-created-data. To overcome this problem, we hold HASC Challenge as a technical challenge to collect the data for activity recognition. In addition to HASC Challenge, we collected indoor pedestrian sensing data of 107 people with a balance of gender and age (HASC-IPSC). Through these data collection, we gathered 111,968 sensor files of 510 subjects. For the convenience of the future researchers in this field, we combined them as a single corpus named HASC-PAC2016 and make it public. Baseline recognition result of HASC-PAC2016 segmented data is 73.4% accuracy for overall, 81.4% for limited by terminal position, and 85.1% with file-based recognition. For sequence data, we only get 73.4% even for limited subjects. This shows we need further research of activity recognition using HASC-PAC2016.

Author Keywords
Activity Recognition; Activity Understandings; Wearable Computing; Accelerometer; Wearable Sensor; Large Scale Corpus; HASC; Smartphone; Sensor.

ACM Classification Keywords
I.5.m [Pattern Recognition]: Miscellaneous

Introduction

Most of the researches on human activity recognition are based on a small number of subjects and lab-created data. Therefore, it is difficult to compare the methods/algorithms of each research. To overcome this problem, we have started a project named HASC[1] Challenge [1] to collect large-scale human activity corpus.

HASC Challenge is a technical challenge which shares the data and tools, and enhances know-how to share sensor information and knowledge about activity recognition. In HASC Challenge, participated teams collect body-mounted sensor data recorded by smart-phone and submit it. We hold HASC Challenge 5 times and composed submitted data as HASC corpora. In addition, we collected indoor pedestrian sensing data of 107 participants with a balance of gender and age, and composed them as HASC-IPSC[2] [2].

We combined activity data in all HASC corpora and HASC-IPSC, and composed them as HASC-PAC2016[3]. We open the HASC-PAC2016 public for researchers to evaluate their methods and compare with others'.

We show baseline recognition using HASC-PAC2016. We use 2 kinds of data type, segmented data and sequence data (These are described in Section "HASC-PAC2016") for recognition. Segmented data is 73.4% accuracy for overall, 81.4% for limited by terminal position, and 85.1% with file-based recognition. For sequence data, we only get 73.4% even for limited subjects.

This paper is organized as follows: Section "Related Work" describes large-scale corpus in activity recognition and other recognition fields and indicates its problem. Section "HASC Challenge" describes HASC Challenge and gathered data. Section "HASC-PAC2016" describes this corpus such as the files included it and data formant. Section "Baseline Recognition Experiment" shows the baseline recognition using some dataset of HASC-PAC2016.

Related Work

Large-scale corpus including many subjects is important in human information processing because the more subjects the corpus includes the better the recognition performance would be [3]. There are some corpora in some fields in human information processing like:

- Speech recognition: PASL-DSR [4], UT-ML [5], TMW [6]
- Image processing: Face [7], TRECVID [8]
- Natural language processing: CSJ [9], BCCWJ [10]

However, there is no large corpus contains a lot of subjects in the field of activity recognition until 2009. Therefore, we founded HASC in 2009. At present, there are some corpora in the field of activity recognition in addition to HASC. UCI has provided "Machine Learning Repository" [11] which maintains five datasets for activity recognition as a service to the machine learning community. For example of dataset in this repository,

[1] HASC: Human Activity Sensing Consortium
[2] HASC-IPSC: HASC Indoor Pedestrian Sensing Corpus
[3] HASC-PAC2016: HASC Pedestrian Activity Corpus 2016

Anguita [12] composed a human activity recognition dataset built from the recordings of 30 subjects performing six activities while carrying a waist-mounted smartphone. Roggen [13] composed a large-scale corpus called "OPPORTUNITY" for activity recognition, and proposed recognition system using it. Also, Lockhart [14] composed a corpus collected by 59 users wearing a smartphone to propose the benefits of personalized smartphone-based activity recognition models.

However, these corpora do not include more than 59 subjects. We need the dataset which includes a lot of subjects, since the characteristics of human activity are different among the individuals, and it greatly affects the performance of activity recognition as demonstrated by Lockhart [14]. To overcome this problem, we hold HASC Challenge.

HASC Challenge 2010-2014

We hold "HASC Challenge" to gather the corpus and technological evaluation. Up to now, HASC Challenge has been held five times in 2010, 2011, 2012, 2013, and 2014. Trough each HASC Challenge, we composed each HASC Corpus. Table 1 shows the number of teams, subjects and submitted data in each HASC Challenge.

	Number of teams (submitted data)	Number of subjects	Number of sensor files
HC2010	21	116	4,898
HC2011	19	101	7,662
HC2012	17	80	8,572
HC2013	18	61	9,993
HC2014	11	45	47,934

Table 1: Number of teams/subjects/files in HASC Challenges

We composed six HASC corpora (HASC2010corpus, HASC2011corpus [3], HASC2012corpus [15], HASC2013corpus, HASC2014corpus, HASC2015corpus) as the results of each HASC Challenge. HASC2011corpus is composed from HASC2010corpus as the result of HASC Challenge 2010 in addition to another 20 subjects. Moreover, we collected indoor pedestrian sensing data of 107 people with a balance of gender and age, and composed them as HASC-IPSC.

HASC-PAC2016

We combined activity data in all HASC corpora and HASC-IPSC, and composed them as HASC-PAC2016. HASC-PAC2016 is targeted for basic human activity, which includes 6 activities: no activity(stay), walk, jog, skip, going up stairs(stUp), going down stairs(stDown) because they are performed regularly by many people in their daily routines. We make the HASC-PAC2016 for the research community to use it as a common ground of the human activity recognition. You can download the corpus via http://hub.hasc.jp.

HASC Activity Data Format

To share the activity data or processing functions among the researchers and developers, activity data format must be standardized. We have defined the following data format as HASC activity data format for activity understandings. The name of each file in HASC-PAC2016 consists of file ID No and extension of file type such as csv, label and meta. In case that the file is sensor data file, the file name also contains sensor type. (e.g. HASC0500010-acc.csv)

SENSOR DATA (.CSV)

We defined sensor data file format as a simple csv format with time stamp and sensor values. For the

acceleration data, it contains: time stamp, x, y and z axis-acceleration values for each row. Time stamp is in seconds with floating point. Therefore any sampling rate data can be stored with this format. Accelerations are in the gravitational acceleration unit ($1G = 9.80665m/s^2$).

e.g. HASC0500010-acc.csv

```
1361771068.581000,0.195312,-0.982422,-0.132812
1361771068.591000,0.189453,-0.992188,-0.111328
1361771068.600000,0.172852,-0.993164,-0.111328
1361771068.610000,0.175781,-0.991211,-0.110352
...
```

LABEL DATA FORMAT (.LABEL)

For each continuous activity data, "tag" or "label" is required to put on the activity time period. We defined a ".label" data format as a csv format with start-time, end-time and label-name. Line starts from "#targetfile:" denotes the reference to the accelerometer data. This helps HASC Tool to show the label with the signal data. By using this format, one can easily add any kind of label onto the time-series data. However, definition of the labeling is not easy. We need further research on this area.

e.g. HASC0500010.label

```
#targetfile:HASC0500010-acc.csv
#targetfile:HASC0500010-gyro.csv
#targetfile:HASC0500010-mag.csv
1.3617710722246E9„start
1.3617710740270002E9,1.3617710989520001E9,stay
1.3617711005690002E9,1.3617711127801E9,walk
1.3617711609210002E9,1.3617711188512E9,jog
1.3617711923560002E9,1.3617712195030003E9,skip
...
```

META INFORMATION FORMAT (.META)

For each sensor data, related information of the subject and the data acquisition condition are important. We defined a meta information file format to record log version, subject's number, gender, generation, weight, height, shoes, terminal mount, terminal position, route, activity type, terminal's ID, terminal's type, client version, sampling rate and source corpus. The style of the format is simple "attribute:value" pair.

e.g. HASC0500010.meta

```
LogVersion: 2
Person:Person1206
Gender:male
Generation:20;late
Height(cm):173
Weight(kg):70
Shoes:sneakers
TerminalMount:free
TerminalPosition:wear;outer;chest
Route: ra00
Activity: sequence
TerminalID: 4ac98181d65082ae
TerminalType: GT-I9300;SDK=16;VI=I9300XXELLB
ClientVersion: Logger+Wifi for Android;1.0
Frequency: 100
SourceCorpus:HASC-IPSC
```

To visualize the sensor data of these formats, we developed a tool named HASC Tool[4] (Figure 1). HASC Tool has the following features including visualization to

[4] HASC Tool is Apache 2.0 Licensed open source software. You can download it from http://en.sourceforge.jp/projects/hasc

boost the data handling and trial-and-error process of the signal processing.

- Creating a process block diagram graph called "XBD". By using "XBD", one can easily automate the processing of various signals and file processing. Without this kind of automation, handling thousands of files is not easy.
- Both "real time" and "offline" data acquisition with wireless sensors.
- Connection with Weka[5] Toolkit.

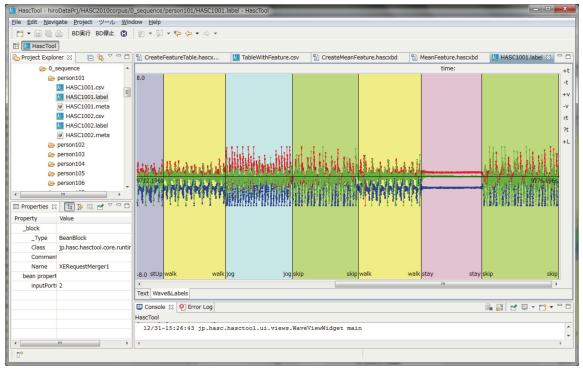

Figure 1: HASC Tool (labeling mode)

Files included in HASC-PAC2016

HASC-PAC2016 contains the following files (Table 2 and Table 3). The number of subjects is 510 people. Some subjects have several sensors simultaneously such as accelerometer(Acc), gyroscope(Gyro), magnetometer (Mag), location(Loc), barometric pressure(Pressure), proximity(Proxi) and WiFi, both in segmented data and sequence data. Also, the number of subjects is 81 people in Real World data. We will describe about segmented, sequence and real world data. Data size is 9.26GB in total.

Number of subjects	Sensor Type	Number of files
510 (Male 390, Female 120)	Acc	40,702
	Gyro	28,644
	Mag	25,361
	Loc	1,881
	Pressure	6,810
	Proxi	534
	Light	559
	WiFi	6,536
	Total	111,027

Table 2: Statistics of basic activity data in HASC-PAC2016

Number of subjects	Sensor Type	Number of files
81 (Male 65, Female 16)	Acc	307
	Gyro	296
	Mag	270
	Loc	40
	Pressure	5
	Proxi	11
	Light	12
	WiFi	0
	Total	941

Table 3: Statistics of real world data in HASC-PAC2016

HASC-PAC2016 consists of segmented, sequence and real world data. These data types are described as follows.

[5] Weka Toolkit is a data mining/machine learning tool developed by Waikato Univ. (http://www.cs.waikato.ac.nz/ml/weka/)

SEGMENTED DATA (MOSTLY FOR TRAINING DATA)
For each file, more than 20 sec continuous single activity is recorded.

SEQUENCE DATA (MOSTLY FOR TEST DATA)
The data file of each subject includes an activity sequence of 300 sec of all the previously mentioned 6 labeled activities. (Each activity should be longer than 10 sec.) Therefore, sequence data is similar to the actual behavior of human basic activity.

REAL WORLD DATA
The label data of real world data includes the movement between landmarks. Each movement between landmarks is consisted of many consecutive events. Therefore, we can try to estimate the route using landmark information.

We have improved data format of HASC Challenge since HASC Challenge2011 because of the lesson from HASC Challenge 2010. The changes are as follows:

- Detailed terminal position with metadata.
- The extended the duration of (data reading/activity recognition) from 2mins to 5mins.
- Collected the data from all sensors. (Not only acceleration sensor.)
- Collected real world data.

All sensor data in HASC-PAC2016 have metadata that contains some tags. There are many types of tags of the meta information file in HASC-PAC2016. These tags are varied according to each subject. Figure 2 shows the rates of the terminal position in HASC-PAC2016. Figure 3 shows the rates of smartphone manufacturers in HASC-PAC2016. Figure 4 shows subject's weight and height in HASC-PAC2016.

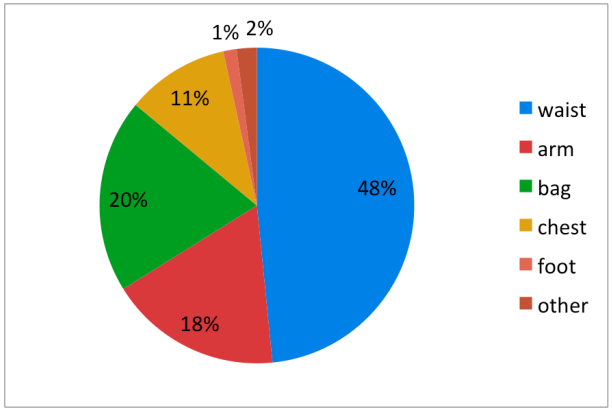

Figure 2: The rate of terminal position among metadata files on HASC-PAC2016

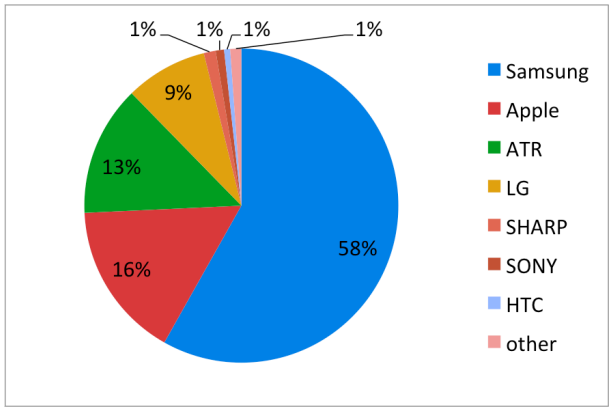

Figure 3: The rate of smartphone manufacturers in HASC-PAC2016

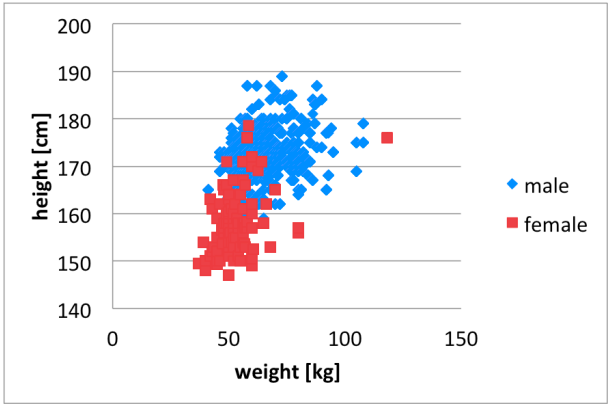

Figure 4: Subject's weight and height on HASC-PAC2016

Baseline Recognition Experiment

By using a large-scale human activity corpus, we can perform various research tasks. In this section we show the baseline recognition through simple experiment for evaluation using HASC-PAC2016. In this experiment, we use 2 kinds of dataset. One contains all terminal position (40702 acceleration files of 510 subjects), another is the data which is collected with the terminal being put in the right pocket. (2,998 acceleration files of 62 subjects) We use acceleration data for activity recognition in all experiments because the number of subjects decreases in case that we extract the data which contains both of accelerometer and gyroscope. To recognize the user activities, we use random forest algorithm by Scikit-learn [16], because HASC Tool does not include random forest algorithm yet. (We will implement new algorithm such as random forest to HASC Tool.)

Data Pre-processing & Feature Computation

Because HASC-PAC2016 contains various sampling frequency, we resample acceleration data to 100Hz. Some of HASC-PAC2016 sensor data file contains initialization or terminating behavior of data collection, we remove the first 2 seconds and the last 5 seconds for each file.

We use the following features [17,18] on 4 seconds windows of acceleration data with 2 seconds overlapping between consecutive windows for x, y, z, and norm. Thus, we use 52 features (13×4).

- Average
- Standard Deviation
- The difference between "maximum value" and "minimum value"
- Power Spectrum:
  Power spectrum is computed from the result of FFT. From 0.5Hz to 5Hz(0.5Hz intervals)

To maximize the recognition result, we examine the effects of the power spectrum for recognition. In previous research Bao [17] uses low frequency of power spectrum because it is effective for human activity recognition. On contrast, this examination uses the specific frequency divided 13 patterns (From 0.5Hz to 50Hz as shown in Figure 5) to find out effective range of power spectrum for activity recognition. In this examination, we use the data which is collected with the terminal being put in the right pocket. (2,998 acceleration files of 62 subjects). The result is shown in Figure 5. Figure 5 shows that the performance peak is on 0.5~5Hz and tends to decrease gradually over

0.5~10Hz. Therefore, we used the range of power spectrum between 0.5Hz and 5Hz for activity recognition.

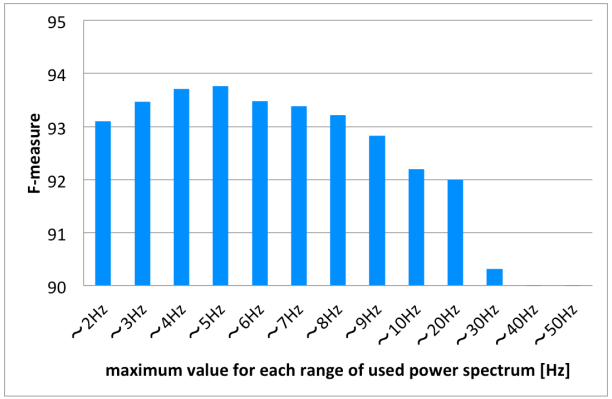

Figure 5: F-measure for each range of used power spectrum (From 0.5Hz to 50Hz as maximum)

Recognition using Segmented Data

WINDOW-BASED RECOGNITION

We use only segmented data for training and testing. 80% of the data is for training, and other 20% is for testing as 5-fold cross validation. The result of each window is shown in Table 4. This result indicates that most of errors are a result of confusing walking with going up or down stairs and skip with jog. These two confusions may be due to the similar period between footsteps in these activities.

In the recognition as above, we use the dataset contains all terminal position. However, sensor data recorded in each terminal position is different. Therefore, we extract the data recorded in right pocket from HASC-PAC2016 to arrange experiment condition, and use it for recognition. The result is shown in Table 5.

		Classified Class						Recall [%]
		a	b	c	d	e	f	
Actual Class	a = stay	**23755**	383	296	295	242	190	96.9
	b = walk	1682	**28310**	402	690	8772	5654	62.2
	c = jog	1205	670	**38391**	1236	503	2063	87.1
	d = skip	1253	849	2229	**35785**	782	3064	81.4
	e = stUp	1216	8114	246	785	**23047**	6702	57.5
	f = stDown	1325	6757	959	1777	8431	**20795**	51.9
Precision [%]		86.8	62.8	90.3	88.2	55.2	54.1	**73.4**

Table 4: Confusion matrix on window-based recognition using the data contains all terminal position (40702 acceleration files of 510 subjects)

		Classified Class						Recall [%]
		a	b	c	d	e	f	
Actual Class	a = stay	**2315**	21	40	1	30	5	96.0
	b = walk	52	**1598**	11	9	401	246	69.0
	c = jog	15	26	**2034**	61	34	59	91.3
	d = skip	9	3	168	**2043**	14	9	91.0
	e = stUp	7	211	1	14	**1662**	285	76.2
	f = stDown	4	194	19	23	246	**847**	56.0
Precision [%]		96.4	77.8	89.5	95.0	64.7	58.4	**81.4**

Table 5: Confusion matrix on window-based recognition using the data which is collected with the terminal being put in the right pocket. (2,998 acceleration files of 62 subjects)

FILE-BASED RECOGNITION

In segmented data, each file only contains single activity. Therefore, we apply file-based recognition using most classified activity in each file. Table 6 shows the confusion matrix on file-based recognition. Table 6 shows the activity recognition performance for each activity. As a result of file-based recognition, performance for each activity is improved. The average accuracy increases from 81.4% to 85.1%.

		Classified Class						Recall [%]
		a	b	c	d	e	f	
Actual Class	a = stay	**546**	1	4	0	0	0	99.1
	b = walk	4	**384**	4	1	59	57	75.4
	c = jog	0	14	**463**	5	3	26	90.6
	d = skip	0	0	27	**491**	2	0	94.4
	e = stUp	0	9	0	1	**347**	55	84.2
	f = stDown	1	35	2	6	110	**198**	56.3
Precision [%]		99.1	86.7	92.6	97.4	66.6	58.9	**85.1**

Table 6: Confusion matrix on file-based recognition using the data which is collected with the terminal being put in the right pocket. (2,998 acceleration files of 62 subjects)

Recognition using Sequence Data

We use segmented data for training and sequence data for testing. In this evaluation, we do not use the file-based recognition as majority rule because we could not determine the period of activity unlike the case of using segmented data for testing. The result of this recognition is shown in Table 7. The performance is 73.4% accuracy.

		Classified Class						Recall [%]
		a	b	c	d	e	f	
Actual Class	a = stay	**1544**	15	7	9	39	31	93.9
	b = walk	10	**1034**	24	83	358	467	52.3
	c = jog	9	4	**872**	67	14	17	88.7
	d = skip	5	8	58	**798**	25	18	87.5
	e = stUp	12	108	9	26	**669**	207	64.9
	f = stDown	8	105	34	24	156	**488**	59.9
Precision [%]		97.2	81.2	86.9	79.2	53.1	39.7	**73.4**

Table 7: Confusion matrix using sequence data for testing

Conclusion

In this paper, we introduce large-scale corpus "HASC-PAC2016" which contains 510 subjects with 111,968 sensor data files as the result of HASC Challenge and HASC-IPSC over 5 years. We also evaluated the performance as baseline for activity recognition using HASC-PAC2016. We demonstrated baseline recognition result of HASC-PAC2016 segmented data is 73.4% for overall, 81.4% for limited by terminal position, and 85.1% with file-based recognition. For sequence data, we only get 73.4% even for limited subjects. This shows we need further research of activity recognition using HASC-PAC2016. We would be glad if this data could be used in basic and applied research and contribute further to advancement in the fields of activity recognition.

References

1. Kawaguchi, N., Ogawa, N., Iwasaki, Y., Kaji, K., Terada, T., Murao, K., ... & Nishio, N. (2011, March). HASC Challenge: gathering large scale human activity corpus for the real-world activity understandings. In *Proceedings of the 2nd*

Augmented Human International Conference (p. 27). ACM.

2. Kaji, K., Watanabe, H., Ban, R., & Kawaguchi, N. (2013, September). HASC-IPSC: indoor pedestrian sensing corpus with a balance of gender and age for indoor positioning and floor-plan generation researches. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication* (pp. 605-610). ACM.
3. Kawaguchi, N., Yang, Y., Yang, T., Ogawa, N., Iwasaki, Y., Kaji, K., ... & Sumi, Y. (2011, September). HASC2011corpus: towards the common ground of human activity recognition. In *Proceedings of the 13th international conference on Ubiquitous computing* (pp. 571-572). ACM.
4. Spoken Language and the DSR projects speech corpus (PASL-DSR). http://research.nii.ac.jp/src/en/PASL-DSR.html.
5. University of Tsukuba Multilingual Speech Corpus (UT-ML). http://research.nii.ac.jp/src/en/UT-ML.html.
6. Tohoku University - Matsushita Isolated Word Database (TMW). http://research.nii.ac.jp/src/en/TMW.html.
7. FACE RECOGNITION HOMEPAGE. 2005. DATABASES. http://www.face-rec.org/databases/.
8. TRECVID. http://trecvid.nist.gov/trecvid.data.html.
9. Corpus of Spontaneous Japanese (CSJ). http://pj.ninjal.ac.jp/corpus_center/csj/misc/preliminary/index_e.html.
10. Balanced Corpus of Contemporary Written Japanese (BCCWJ). http://pj.ninjal.ac.jp/corpus_center/en/kotonoha.html.
11. Asuncion, A., & Newman, D. (2007). UCI machine learning repository.
12. Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2012, December). Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *International Workshop on Ambient Assisted Living* (pp. 216-223). Springer Berlin Heidelberg.
13. Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Förster, K., Tröster, G., ... & Doppler, J. (2010, June). Collecting complex activity datasets in highly rich networked sensor environments. In *Networked Sensing Systems (INSS), 2010 Seventh International Conference on* (pp. 233-240). IEEE.
14. Lockhart, J. W., & Weiss, G. M. (2014). The Benefits of Personalized Smartphone-Based Activity Recognition Models. In *SDM* (pp. 614-622).
15. Kawaguchi, N., Watanabe, H., Yang, T., Ogawa, N., Iwasaki, Y., Kaji, K., ... & Sumi, Y. (2012, April). Hasc2012corpus: Large scale human activity corpus and its application. In *Proceedings of the IPSN* (Vol. 12).
16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825-2830.
17. Bao, L., & Intille, S. S. (2004, April). Activity recognition from user-annotated acceleration data. In *International Conference on Pervasive Computing* (pp. 1-17). Springer Berlin Heidelberg.
18. Kwapisz, J. R., Weiss, G. M., & Moore, S. A. (2011). Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, *12*(2), 74-82.