HASC2012corpus: Large Scale Human Activity Corpus and Its Application

Nobuo Kawaguchi
Hodaka Watanabe
Tianhui Yang
Nobuhiro Ogawa
Yohei Iwasaki
Katsuhiko Kaji
Nagoya University,
1Furo-Cho, Chikusa-ku,
Nagoya 464-8603, Japan
+81-52-789-4388
{kawaguti,hodaka,tenki,hiro,iwasaki,kaji}@ucl.nuee.nagoya-u.ac.jp

Tsutomu Terada
Kazuya Murao
Kobe University
1-1Rokkodai-machi,
Nada-ku, Kobe, 657-8501, Japan
{tsutomu,murao}@eedept.kobe-u.ac.jp

Hisakazu Hada
Keio University
5322 Endo, Fujisawa,
Kanagawa 252-0882, Japan
hada@sfc.wide.ad.jp

Sozo Inoue
Kyushu Institute of Technology1-1 Sensui-cho, Tobata-ku,
Kitakyushu 804-8550, Japan
sozo@mns.kyutech.ac.jp

Yasuyuki Sumi
Future University Hakodate
116-2 Kamedanakano-cho,Hakodate, Hokkaido
041-8655, Japan
sumi@acm.org

Yoshihiro Kawahara
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-0033, Japan
kawahara@akg.t.u-tokyo.ac.jp

Nobuhiko Nishio
Ritsumeikan University
1-1-1 Noji-higashi,
Kusatsu, Shiga 525-8577, Japan
nishio@cs.ritsumei.ac.jp

ABSTRACT
Human activity recognition through the wearable sensor will enable a next-generation human-oriented ubiquitous computing. However, most of researches on the activity recognition so far are based on small number of subjects, and non-public data. To overcome the situation, we have formed a consortium named HASC and create corpora for the research community. In HASC Challenge2010, we have made HASC2011corpus with 116 subjects and 4896 accelerometer data. In HASC Challenge 2011, we have gathered 7,668 sensor data with 136 subjects and compose them as HASC2012corpus. In the field of pattern recognition, it is very important to evaluate and to improve the recognition methods by using the same dataset as a common ground. We make these corpora into public for the research community to use it as a common ground of the Human Activity Recognition. We also show several facts and results of obtained from the corpus.

Categories and Subject Descriptors
H.5.2 [**User Interfaces**]: Interaction styles

General Terms
Measurement, Experimentation, Human Factors, Standardization.

Keywords
Activity Recognition, Activity Understandings, Accelerometer, Wearable Sensor, Large Scale Corpus, HASC

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IPSN'12, April 16–20, 2012, Beijing, China.

Copyright 2012 ACM 978-1-4503-1227-1/12/04…$10.00.

1. INTRODUCTION
Most of researches on the human activity recognition so far are based on small number of subjects and lab-created private data[1]. So it is difficult to compare the methods/algorithms across the literatures. To overcome the situation, we have started a project named "HASC Challenge[2]" to collect a large scale human activity corpus. In HASC Challenge 2010, by the collaboration of more than 24 teams, we have gathered 4,897 carefully and precisely labeled accelerometer data with 116 subjects. As a result, we have composed them as HASC2011corpus. We make the HASC2011corpus into public for the research community to use it as a common ground of the Human Activity Recognition. We have continued the data collection effort as HASC Challenge 2011. In HASC Challenge 2011, we have expanded the sensor type and collection types. As a result, we have gathered 7,668 sensor data (not only accelerometer but with gyro, GPS and magnetic field sensor) with 136 subjects and compose them as HASC2012corpus.

In the following, we first explain about HASC Challenge. Then we explain HASC activity data format which enables simple and easy data exchange. We also explain about HASC Logger and HASC Tool[7] which is a basic toolkit for activity recognition research. Then we show the basic information of HASC2011corpus and HASC2012corpus. We also explain the current preliminary findings from the HASC2012corpus about route estimation by the sequence of activity recognition.

2. HASC CHALLENGE 2010 AND 2011
In 2009, we have organized a consortium named "HASC: Human Activity Sensing Consortium", and started a collaborative project for gathering a large scale human activity corpus. This project is named "HASC Challenge." However, there are several issues which cannot simply be decided such as sensor types, number of sensors, sampling frequency, position of sensors and the kinds of activities. If we decide the application of the corpus, these

parameters can be easily decided. However, our purpose of the corpus gathering is to boost the research of the activity recognition. So fixing the application area is not the solution. In HASC Challenge 2010, as a result of long discussion, we have decided to start gathering a single accelerometer sensor data of simple activities with various kinds of sensors, positions and sampling rates. Also, we restrict activities in 6 types such as stay(standing), walk, jogging, skip, stair-up and stair-down. Restriction of activities is not good for various kinds of applications. However, these activities are inherently core activities (except skip), so it is very important to robustly recognize them in the initial stage.

From the taking place of HASC Challenge 2010, we have obtained follows important experiences.

- It is important to have a large number of subjects to obtain a higher accuracy on a user-independent recognition [3].
- Labeling for activities is very important. Especially for the transition between the activities.
- Difference of the placement of sensors and sensors itself have severe effects on recognition accuracy.
- Richer sensor data is required for more accurate recognition.

In HASC Challenge 2011, we extend sensor types from single accelerometer to any sensors which are equipped in the smartphone such as gyroscope, magnetic field sensor and GPS. We also add another type of data collection named "real world activity" data. By using "real world activity" data, we might be possible to test the robustness of the future activity recognition methods.

2.1 Basic Rules of HASC Challenge

For each participants of HASC Challenge, we announced the following issues.

- HASC Challenge is not a contest. It is a "Technology challenge."
- Each participant should gather some activity dataset.
- In HASC Challenge 2010, all of participants should submit at least five subjects with the following "lab-controlled activity".
    - 5 accelerometer data file of 6 activities(length 20 seconds): stay(standing), walk, jogging, skip, stair-up and stair-down.
    - Labeled sequence of activities which include all of above 6 activities. (Each activity should be longer than 5 seconds and total length should be longer than 120 seconds.)
- In HASC Challenge 2011, participants can select to gather the above "lab-controlled activity" data or "real-world activity" data.
    - For lab-controlled dataset, we have slightly modified the condition for sequence data. We asked participants to collect more than 300 seconds of sequence data and each activity should be longer than 10 seconds. This is because of difficulty in the recognition of small length activities.
    - For real-world data, we asked participants to collect a data of movement between landmarks such as stations, stores, offices or etc. The activity data should be labeled with activity tags and event tags. Typical length of real-world data is around 300 to 600 seconds.
- Each participant can use any kind of sensor but it should be available in the market.
- Activity data must be described in the HASC data format with meta-information.
- After the submission deadline, each participant will get all the activity data from other participants without label data of the sequence data. They can submit the result of the recognition in the label data format or submit the recognition algorithm.
- HASC steering member will evaluate the recognition rate.

3. HASC ACITIVITY 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 data format for activity understandings. All of HASC Challenge data are in this format.

3.1 Sensor data (.csv)

We defined sensor data file format as a simple csv format with time stamp and sensor values. For the accelerometer data, it may contain: time stamp, x, y and z axis-acceleration values for each row. Time stamp is in the second time scale with floating point. So any sampling rate data can be stored with this format. Accelerations are in the gravitational acceleration unit (1G = 9.80665m/s2). To indicate the sensor type, we simply add the extension to the file name. For accelerometer, gyroscope, magnetic field sensor, and GPS, we add "-acc", "-gyro", "-mag" and "-loc" respectively.

3.2 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 subject's gender, weight and height, and sensor's type, sampling rate and position. The style of the format is simple "attribute:value" pair. We add several attributes in HASC Challenge 2011.

3.3 Label data format (.label)

For each continuous activity data, "tag/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. By using this format, one can easily add any kind of label onto the time-series data. In HASC Challenge 2011, we have employed hierarchical label type. We also introduce the "event" tag which has no duration like "pushing elevator button", "opening-door", etc.

4. HASC Logger and HASC Tool

To simplify the data collection using several sensors, we have developed and deployed "HASC Logger" for both iPhone AppStore and Android Market(Fig1, Fig2). HASC Logger has a feature to record several sensor data and send it to PC. It also has a feature to send sensor data using UDP over WiFi in real-time. iPhone version of HASC Logger can record audio with sensor data. This helps labeling the data. We have also developed a PC tool named "HASC Tool.(Fig 3)" By using HASC Tool, one can easily handle the huge number of activity data. HASC Tool has several visualization features such as raw-signal, labeled signal, and spectrogram. It is fully conformed to HASC Activity Data Format, one can easily view the activity sensing data using graph tool.

Figure 1. HASC Logger for iPhone / iPod Touch.

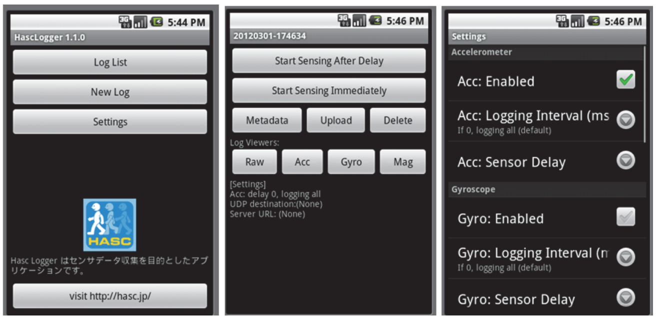

Figure 2. HASC Logger for Android.

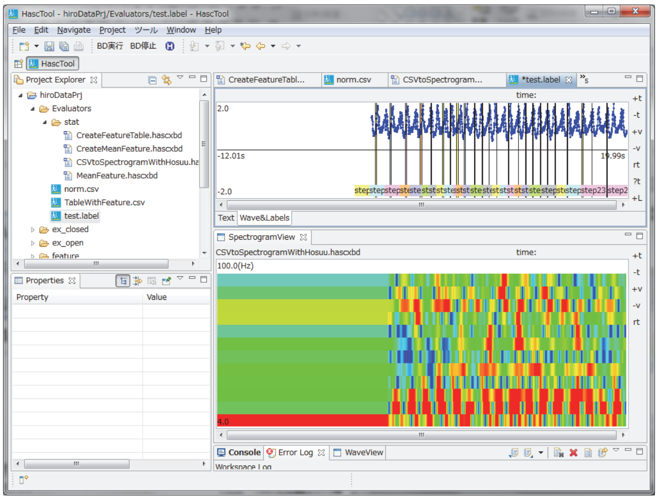

Figure 3. Screen shot of HASC Tool.

5. HASC2011corpus and HASC2012corpus

As a result of HASC Challenge 2010, we have composed HASC2011corpus[6] (Table 1). We have already made the HASC2011corpus public through our web site[8]. For the benefit of participants in HASC Challenge, we do not make the corpus into public just after the challenge. Only participants can obtain the corpus just after the challenge. However, if you want to obtain the corpus, at any time, you can commit the same portion of dataset, and then you will get the latest corpus.

As a result of HASC Challenge 2011, we have constructed a corpus named HASC2012corpus. HASC2012corpus consists of 136 subjects data with 7,668 sensor data(Table 2). HASC2012corpus also include various types of sensor devices. Maker of sensors are shown in Table 3.

Table 1. Statistics of HASC2011corpus

Gender	Number of Subjects	Number of Files
Male	102	4464
Female	14	434
Total	116	4898

Table 2. Statistics of HASC2012corpus

Type	Number of Subjects	Sensor Types	Number of Files
Lab controlled 6 type	96 (Male 85, Female 11)	ACC	4495
		GYRO	2521
		GPS	240
		MAG	310
RealWorld	40 (Male 34, Female 6)	ACC	40
		GYRO	30
		GPS	27
		MAG	5
Total	136		7668

Table 3. Sensor Devices in HASC2012corpus

Sensor Maker	Number of Files	Percentage (%)
ATR /WAA-00X	1854	40.9
Apple / iPhone, iPodTouch	1759	38.8
Samsung	662	14.6
HTC	118	2.6
Sharp	95	2.1
Sony	32	0.7
LG	15	0.3

6. APPLICATION OF THE CORPUS

In HASC Challenge 2011, we have collected "real-world activity" data for the commuting task between landmarks. In HASC2012corpus, we have 40 traces of commuting activity. In this section, we will propose a new area of research using this kind of sensor data of moving human.

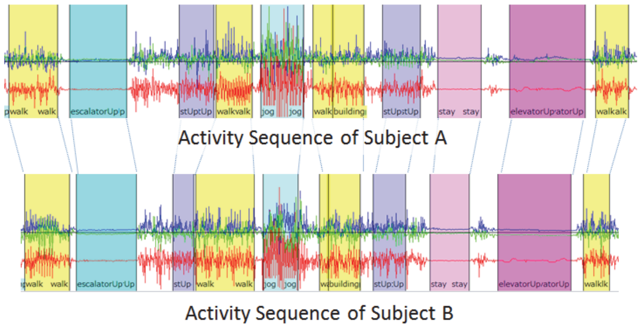

Figure 4. Activity Sequence of Same Route.

Figure 5. Activity Sequence of Recognition Result.

6.1 Similarity of the Activity Sequence

By using "commuting activity data", we think it might be possible to estimate the user's route, location or direction just using activity sensor data. Figure 4 shows a different subject's activity sequence of the same route. We ordered subjects A and B to move through the same commuting route between the exit of subway to the office at the 5th floor. The activity sequences of them are labeled by hand. You can see, there are similarities between two activity sequences. If it is possible to recognize these activity sequences using activity recognition technology, we can calculate "similarity" between different activity sequences. Especially, if a sequence contains elevator, escalator or stairs, these activities make the sequence more unique. By using some kind of matching technology, we can match activity records of unknown route with route-known activity sequences. The matching is like a "video clip search". If we can find out the similar route, that means we can estimate the current user's route or location only from the sensing data.

If we can successfully develop this "route estimation technology from sensing data", we can improve "life logging", "context aware services", or "location aware services".

6.2 DP Matching of Activity Sequence

In this section, we just show the current status of our progress on this area. To estimate the similarity between the activity sequences, we employ DP-Matching. We also utilize the cost function which is related with recognition accuracy. By using DP-matching with the cost function, we can define the similarity between the sequences.

First, we compare the recognition result of the sequences with the hand-defined label. However, recognition rate of the current system[4] only performs around 60% in this real world activity data. In Figure 5, we show the recognition result of Subject B's activities of Figure 4. In Figure 4, escalator and elevator are labeled collect. However, current our recognizer cannot robustly recognize them. Also in Figure 4, not of all sensor signals are labeled with activities. Because there is transition between the activities, it is not clear that how to label these transitions. So, the matching between hand-defined labels with machine recognized labels does not work well.

Second, we change our mind from using hand-labeled activity sequences to machine recognized label. We collect another activity dataset which contains activity sequences of 5 subjects (A-E) with 3 routes. All of the sequences are labeled with recognition engine and then calculate the similarity between each of them.

Figure 7 shows a table of similarity (smaller is more similar) between different sequences. From this table, you can read that in most of the routes, we can find other sequence of same route by other subjects. This result suggests that if we have large scale data of activity sequences with route information, we can estimate the user's route.

7. CONCLUSION

We have made a basic step for improving human activity recognition technology. Currently, we mostly have a dataset for simple activities. However, we believe even these simple activity data can improve the recognition technology. By exploring the corpus, there will be more rich results.

We also show a sample application of HASC2012corpus. By using real world activity data, we start a new research about "estimation of user's route by sensing data". This is a new application of activity recognition technology. We think there will be more spaces of research in this area.

Additionally, we have obtained a grant for gathering large scale activity dataset, so we can continue to gather human activity corpus to make and improve the common ground.

table	stay	stay	walk	stay	stay	stay	walk	walk	stUp	stDown	stay	stUp	stay	stay
stay	1	32	154	185	216	247	369	491	613	735	766	888	919	950
walk	45	41	34	103	172	241	249	280	349	418	487	556	625	694
walk	89	85	40	74	143	212	243	251	320	389	458	527	596	665
stay	95	91	137	42	73	104	226	336	344	413	391	513	529	560
stay	101	97	184	48	44	75	197	319	429	437	397	484	515	531
stay	107	103	190	54	50	46	168	290	412	522	403	490	486	517
walk	151	147	105	98	94	90	48	79	148	217	286	355	424	493
walk	195	191	111	142	138	134	54	50	119	188	257	326	395	464
walk	239	235	117	151	182	178	60	56	90	159	228	297	366	435
walk	283	279	123	157	191	222	66	62	96	130	199	268	337	406
walk	327	323	129	163	197	231	72	68	102	136	170	239	308	377
walk	371	367	135	169	203	237	78	74	108	142	176	210	279	348
walk	415	411	141	175	209	243	84	80	114	148	182	216	250	319
walk	459	455	147	181	215	249	90	86	120	154	188	222	256	290
stay	465	461	244	149	180	211	187	183	179	213	156	278	224	255
walk	509	505	250	193	189	220	193	189	223	219	200	196	265	264
walk	553	549	256	237	233	229	199	195	229	263	244	240	236	305
stay	559	555	353	243	239	235	296	292	288	322	250	337	242	238
walk	603	599	359	287	283	279	237	268	332	328	294	290	286	282
stUp	653	649	409	337	333	329	287	283	270	345	344	296	336	332
stUp	703	699	459	387	383	379	337	333	276	316	391	302	342	382
stay	709	705	556	393	389	385	434	430	373	369	318	399	304	335
stay	715	711	653	399	395	391	478	527	470	466	324	411	310	306
stay	721	717	750	405	401	397	484	571	567	563	330	417	316	312
stay	727	723	810	411	407	403	490	577	664	660	336	423	322	318
walk	771	767	725	455	451	447	405	436	505	574	380	376	366	362

Figure 6. Sample of Activity Sequence DP-Matching

	A-1	B-1	C-1	D-1	E-1	A-2	B-2	C-2	D-2	E-2	A-3	B-3	C-3	D-3	E-3
A-1	2	7	20	25	16	24	27	20	30	29	58	30	30	39	39
B-1	9	2	18	26	24	27	38	20	26	30	54	29	30	55	42
C-1	30	29	1	45	31	16	19	15	17	21	36	34	23	30	31
D-1	12	7	19	2	19	27	35	20	23	25	65	33	32	54	55
E-1	24	23	25	26	2	28	34	22	24	31	58	30	30	40	42
A-2	68	58	34	100	69	1	10	24	34	10	38	35	35	35	40
B-2	64	58	32	71	62	8	1	46	35	20	25	28	35	36	31
C-2	49	51	25	71	42	26	28	1	26	26	35	38	33	36	40
D-2	71	60	27	95	48	9	10	20	1	8	25	33	29	32	37
E-2	77	62	38	99	75	31	10	30	32	1	35	37	32	37	44
A-3	74	69	46	80	92	40	39	38	43	40	1	17	20	14	24
B-3	57	55	45	75	70	28	26	38	40	19	8	1	25	10	23
C-3	57	59	44	70	70	28	27	37	38	21	10	24	1	10	14
D-3	52	56	52	70	72	41	40	39	40	37	7	21	19	1	20
E-3	63	58	44	74	66	39	31	39	47	29	7	15	16	9	1

Figure 7. Similarity of DP-Matched Sequences

8. ACKNOWLEDGMENTS

This work is partly supported by KAKENHI(23240014), Grant-in-Aid for Scientific Research(A) and JST CREST Kashin-Free project. We all have great thanks to HASC Challenge participants.

9. REFERENCES

[1] A.J. Bernheim Brush, John Krumm, James Scott, and T. Scott Saponas. 2011. Recognizing Activities from Mobile Sensor Data: Challenges and Opportunities. Mobile Sensing: Challenges, Opportunities and Future Directions, Ubicomp2011 Workshop.

[2] Nobuo Kawaguchi, Nobuhiro Ogawa, Yohei Iwasaki, Katsuhiko Kaji, Tsutomu Terada, Kazuya Murao, Sozo Inoue, Yoshihiro Kawahara, Yasuyuki Sumi, and Nobuhiko Nishio. 2011. HASC Challenge: gathering large scale human activity corpus for the real-world activity understandings. In Proceedings of the 2nd Augmented Human International Conference (AH '11). ACM, New York, NY, Article 27 , 5 pages. DOI= http://doi.acm.org/10.1145/1959826.1959853

[3] HASC Website: http://hasc.jp/en

[4] Nobuhiro Ogawa, Katsuhiko Kaji, Nobuo Kawaguchi, Effects of Number of Subjects on Activity Recognition - Findings from HASC2010corpus -, International Workshop on Frontiers in Activity Recognition using Pervasive Sensing(IWFAR2011), pp. 48-51.

[5] Yuichi Hattori, Sozo Inoue, Go Hirakawa, A Large Scale Gathering System for Activity Data with Mobile Sensors, IEEE International Symposium on Wearable Computers (ISWC2011),pp.97-100.

[6] Nobuo Kawaguchi, Ying Yang, Tianhui Yang, Nobuhiro Ogawa, Yohei Iwasaki, Katsuhiko Kaji, Tsutomu Terada, Kazuya Murao, Sozo Inoue, Yoshihiro Kawahara, Yasuyuki Sumi and Nobuhiko Nishio, HASC2011corpus: Towards the Common Ground of Human Activity Recognition, In Proc. the 13th int. conf. on Ubiquitous Computing (UbiComp '11). ACM, New York, NY, USA, 571-572. DOI= 10.1145/2030112.2030218.

[7] Nobuo Kawaguchi, Nobuhiro Ogawa, Yohei Iwasaki, and Katsuhiko Kaji. 2011. Distributed Human Activity Data Processing using HASC Tool. In Proceedings of the 13th international conference on Ubiquitous computing (UbiComp '11). ACM, New York, NY, USA, 603-604. DOI= http://doi.acm.org/10.1145/2030112.2030234.

[8] HASC Hub: http://hub.hasc.jp/