HASC Corpus: Large Scale Human Activity Corpus for the Real-World Activity Understandings

Nobuo Kawaguchi

[1]Graduate School of Engineering, Nagoya University
1, Furo-Cho, Chikusa-ku, Nagoya, 464-8603, Japan
kawaguti @ nagoya-u.jp

ABSTRACT
Understandings of human activity through the wearable sensor will enable the next-generation human-oriented computing. However, most of researches on the activity recognition so far are based on small number of subjects, and not well adapted for real world application. To overcome the situation, we have started a project named "HASC Challenge" to collect a large scale human activity corpus. As the mid of 2011, by the collaboration of over 20 teams, we have gathered 4897 accelerometer data with 116 subjects and compose them as HASC2011corpus. We make HASC2011corpus into public for the research community to use it as a common ground of the Human Activity Recognition. We also developed a tool named "HASC Tool" for management, evaluation and collection of the large number of activity sensor data.

1. INTRODUCTION
Recent advancement of MEMS technology enables the installation of small sized accelerometers or gyroscopes on the various kinds of information devices. By using such activity sensors, these devices can estimate its posture or status. However, most of current devices only utilize these sensors for simple orientation and gesture recognition. More deep understandings and recognition of human activity through these sensors will enable the next-generation human-oriented computing. To enable the real-world application by these kinds of wearable sensors, a large scale human activity corpus might play an important role. Human activities vary from person to person. So it is not easy to find suitable features of activity signals for the robust recognition. Most of researches on the activity recognition so far [8]~[26] are based on small number of subjects, and not well adapted for real world application.

To overcome the situation, we have made a consortium named "HASC: Human Activity Sensing Consortium", and started a collaboration project for gathering large scale human activity corpus. This project is named "HASC Challenge". In this paper, we first explain the importance of the large scale corpus, and then report the first result of HASC Challenge 2010[27]. We also defined the standard data formats for activity corpus, and developed an open source data management tool (HASC Tool). HASC Challenge2010 has a result with more than 6000 activity files. After the challenge, we have added more 20 subjects and remove the non-uniform subjects' data and formed HASC2011corpus for convenience to the researchers. Now we open the whole HASC2010/2011corpus to the public.[1]

2. Importance of Large Scale Corpus for the Activity Recognition
There are several research fields which are boosted by the large scale corpus.

2.1 Large Scale Corpus in the Other Field
In the field of speech recognition, large scale corpus play a really important role. Most of recent speech recognizers are based on HMM acoustic-model and statistical language model[1]. These models are developed through the machine learning through the large scale speech corpus. So, the qualities or recognition-rates of the speech recognizers are proportional with the scale of the underlined speech corpus. To advance the recognition quality or to tackle the different situation, development of a high quality and large scale corpus is a key. Most popular example of the speech corpus is "DARPA TIMIT" corpus [2,3]. From the development of this corpus, a lot of continuous speech recognizer works better.

Also in-car activity field, we already gathered a huge scale corpus(more than 800 subjects) for various purposes [4,5]. These corpus enables in-car related researches.

In the field of image recognition, the most popular example is the area of face recognition [6,7]. There are several databases for facial images. Some of them are freely available. This situation makes the fast improvement of the facial recognition technologies.

2.2 Corpus for Activity Recognition
As far as we know, there is no public large scale corpus for human activity recognition area. Of course, there are several researches with the some scale of experiments [7,8]. However, none of them are not published the data nor the algorithms. As we mention in the above, sharing the databases/corpus is a key of the improvement in the pattern recognition research and development field.

3. HASC Challenge
To overcome the situation, we have decided to create a large scale corpus for human activity recognition with founding the

[1] If you want to obtain/contribute to the HASC corpus, please visit http://hub.hasc.jp/.

consortium HASC (Human Activity Sensing Consortium). However, there are several issues which cannot simply decide such as sensor types, number of sensors, sampling frequency, position of sensors and what kind of activities we should share. 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, it is not simple. As the result of the 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. By publishing the data with various kinds of sensors, we believe researchers can find the better configuration of activity recordings.

So, we have planned "HASC Challenge" to gathering the corpus and technological evaluation.

3.1 Rules of HASC Challange2010

In the planning phase of the challenge, we consider the following issues. (We slightly change the rule for HASC Challenge2011. Details are shown in Section 5.)

- HASC Challenge is not a contest. It is a "Technology challenge".
- Each participant should gather at least five subjects with the following activities.
  - 5 series of 20sec of 6 activities : stay, walk, jogging, skip, stair-up, stair-down.
  - 120sec of labeled activity sequence which includes all of above 6 activities. (each activity should longer than 5 sec).
- Each participant can use any kind of sensor but it should be available in the market.
- Activity data must be in the HASC data format.
- Each participant will get the all activity data but label data of 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.

3.2 Result of HASC Challenge2010

HASC Challenge 2010 symposium was held in Dec 8, 2010. Before the symposium, 24 teams have submitted the activity data. Finally, we got 540 subjects activity data and 96 subjects with full dataset. Total number of activity data file is 6791 and total size is 966Mbytes (Table 1). The major types of sensors are iPhone / iPod Touch, and WAA-series (ATR).

We also got the recognition result from 6 teams. However, this challenge is the first time and the most of participants are not expert of the activity recognition technology. So the recognition results are ranged from 38%~72%.

From the experience of HASC Challenge2010, we confirm the strong requirements of the activity corpus and also the toolkit for the activity recognition.

Table 1. Number of Acquired Files in HASC Challenge2010

Gender	Number of Subjects	Number of Files
Male	89	4032
Female	12	341
Unknown	439	2418
Total	540	6791

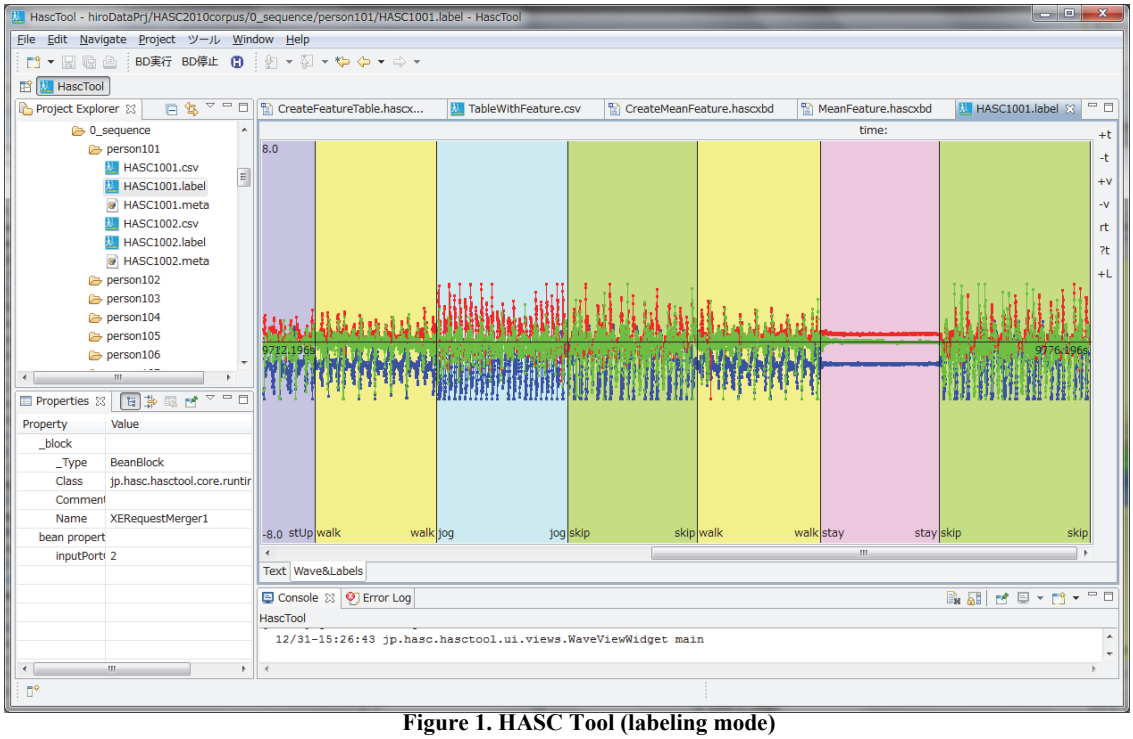

Figure 1. HASC Tool (labeling mode)

.

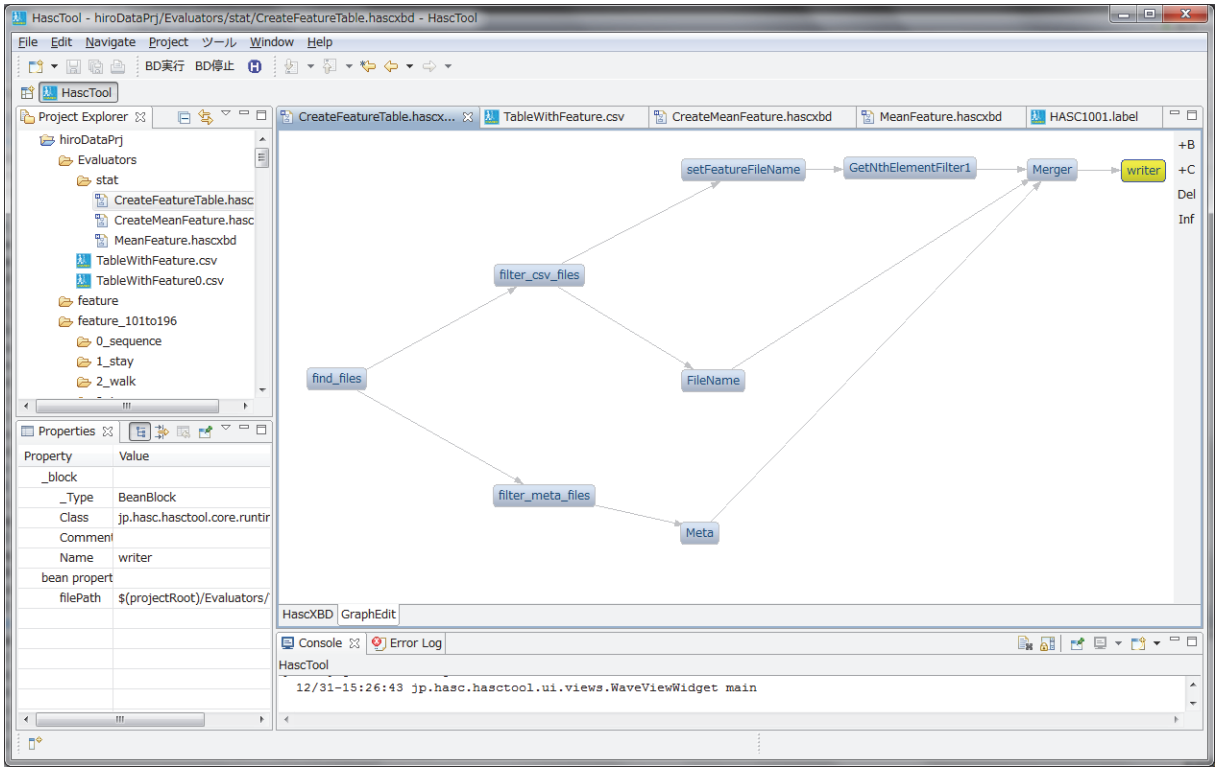

Figure 2. HASC Tool (graph editing mode).

4. Toolkit for Activity Recognition

To boost the data handling and trial and error process of the signal processing, we have developed a new tool named "HASC Tool"[30]. Fig. 1 shows a screen image of HASC Tool[2]. HASC Tool is based on the famous IDE called Eclipse.

HASC tool has following features.

- Showing wave and label data (Fig.1)
- Create a process block diagram graph called "XBD". By using "XBD", one can easily automate the various signal processing and file processing.(Fig.2) Without this kind of automation, handling thousands of files are not simple.
- Real time / offline data acquisition with wireless sensor
- Connection with Weka Toolkit[3]
- Distributed data processing over the network.

By using HASC Tool, we can exchange the processing way of activity recognition using XBD files.

4.1 HASC Data Format for Activity Data

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.

- Activity Accelerometer data (.csv)
  We defined accelerometer data as a simple csv format with time stamp and x, y, z axis-acceleration. 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/s^2).
- Meta information format (.meta)
  For each acceleration data, related information of the subject and the data acquisition condition are important. We defined a meta data 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".
- 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.

5. Findings from HASC2011corpus

We have performed a simple investigation on the effect of number of subjects using HASC2010corpus[28]. We have updated the result using HASC2011corpus[29]. HASC2011corpus is a new corpus which includes additional 20 subjects and omits some in-complete data from HASC2010corpus. So, number of subjects and total files are decreased from the HASC2010corpus but the quality and the consistency are increased in HASC2011corpus. Table 2 shows a basic statistics of HASC2011corpus.

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

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

For the feature selection, we use mean and variance of norm of 3 accelerometers (2features), and power of 4 band FFT with zero-cross rate (7 features), and integration of acceleration (8 features). We learned the segmented data from the corpus and evaluate the sequence data. Table 3 shows a result of recognition experiment using 80 subjects for learning using 8 features. In this experiment, we randomly selected 80 subjects for learning and use the other subjects' sequence data for testing. (User-independent analysis) Figure 3 shows the result of the experiment with x-axis showing the number of subjects and y-axis showing the recognition rate.

Table 2. Statistics of HASC2011corpus

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

Table 3. Confusion matrix of using 80 subjects from HASC2011corpus in the user-independent analysis.

	stay	walk	jog	skip	stUp	stDown
stay	86.7	4.4	2.0	0.7	2.2	3.9
walk	1.5	54.3	0.6	0.8	12.7	30.0
jog	1.3	0.9	49.9	32.4	0.2	15.3
skip	1.7	0.0	10.9	82.8	0.4	4.1
stUp	1.6	7.4	0.0	0.0	78.3	12.7
stDown	2.8	2.0	5.3	2.3	13.8	73.8
Overall	71.0					

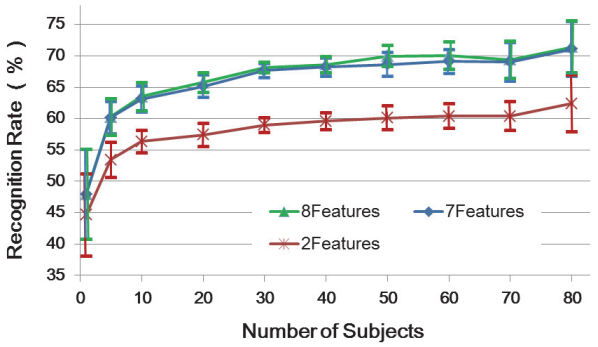

Figure 3. Effects of number of subjects in the user-independent analysis on the human activity recognition.

6. HASC Challenge2011

This year, we are now working for the HASC Challenge2011. We slightly change the rules (or types and length of collecting data) from the experience of the last challenge.

- Each participant should gather at least five subjects with the following activities.
  - 5 series of 20sec of 6 activities : stay, walk, jogging, skip[4], stair-up, stair-down.
  - 300sec(5min) of labeled activity sequence which includes all of above 6 activities. (each activity should longer than 10 sec).
- Each participant can use any kind of sensor but it should be available in the market.
  - Sensor types include accelerometer, gyroscope, magnetic field sensor, GPS.
- Activity data must be in the HASC data format.
- Each participant will get the all activity data but label data of sequence data. They can submit the result of the recognition in the label data format or submit the recognition algorithm.
- Additional data gathering: "Commuting activities"
  - Movement activities from a place to another place around 5~10minutes.
  - More real-world like movement.
- Labeling should be more precise.
  - We do not specify the detail of labeling in the last challenge. However, the recognition rate is heavily affected by the quality of the label. In HASC Challenge 2011, we specify the method of labeling to make it more precise and accurate. Unlike the last year's label like in Figure 1, we asked to the participants to label the activity like Figure 4. (Do not connect each activity. Transition part should not labeled.)

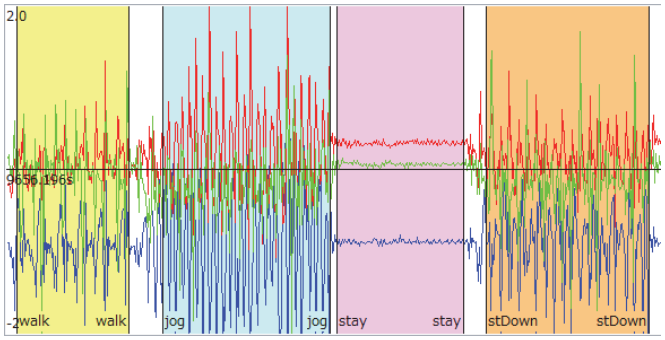

Figure 4. Labeling without activity transition segment.

6.1 Hub.hasc.jp: Supporting site

In HASC Challenge2010, one of the biggest difficulties was to handle the huge number of activity files. Some of the participants do not understand / or mistakenly understand the HASC format and submit in the wrong format. So, we had to check and collect all of them. In the HASC Challenge 2011, we are preparing the web site name "Hub.hasc.jp." to automate the data check.

7. SUMMARY

In HASC Challenge2010, we have gathered more than 6400 activity accelerometer data from 24 teams. From the experience, we confirm the strong demands for activity corpus from researchers and companies. Now we open the HASC2010/2011corpus to the public. We hope a large number of

[4] "Skip" is an activity like jogging with rhythmic jumps. You may find it at http://www.youtube.com/watch?v=aLD7zzX3c48

researchers will use HASC corpus to develop/evaluate their own recognition technologies. We have also developed HASC Tool for data management and processing for activity recognition.

ACKNOWLEDGMENTS

This work is partially supported by CREST Kashin-Free project.

REFERENCES

[1] A. Waibel and K. F. Lee. "Readings in Speech Recognition.", Morgan Kaufmann, 1990.

[2] Fisher, William M.; Doddington, George R. and Goudie-Marshall, Kathleen M., "The DARPA Speech Recognition Research Database: Specifications and Status". Proceedings of DARPA Workshop on Speech Recognition. pp. 93–99.(1986).

[3] Garofolo, J. S.; Lamel, L. F.; Fisher, W. M.; Fiscus, J. G.; Pallett, D. S., "DARPA TIMIT acoustic-phonetic continous speech corpus", CD-ROM, NIST speech disc.

[4] Nobuo Kawaguchi, Shigeki Matsubara, Hiroyuki Iwa, Shoji Kajita, Kazuya Takeda, Fumitada Itakura and Yasuyoshi Inagaki, Construction of Speech Corpus in Moving Car Enviroment , The 6th International Conference on Spoken Language Processing, ICSLP2000, pp.362-365(2000).

[5] Nobuo Kawaguchi, Shigeki Matsubara, Kazuya Takeda, Fumitada Itakura, CIAIR In-Car Speech Corpus: Influence of Driving Status, IEICE Transactions on Information and Systems. v. E88-D, n.3, 2005, p.578-582

[6] R. Gross, Face Databases, "Handbook of Face Recognition", Stan Z. Li and Anil K. Jain, ed., Springer-Verlag, February 2005

[7] Face-Recognition Homepage , http://www.face-rec.org/databases/

[8] Sasank Reddy, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, Mani Srivastava, "Using Mobile Phones to Determine Transportation Modes", ACM Transactions on Sensor Networks, Vol.6, No.2, pp.13:1-13:27(2010).

[9] Sunny Consolvo, David W.McDonald, Tammy Toscos, Mike Y. Chen, Jon Froehilich, Beverly Harrios, Perdrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, Ian Smith, James A.Landay, "Activity Sensing in the Wild: A Field Trial of UbiFit Garden", Proc. of the conference on Human factors in computing systems (CHI '08), 1797-1806 (2008).

[10] Yuki Arase, Fei Ren and Xing Xie, "User Activity Understanding from Mobile Phone Sensors", in Proc. of Ubicomp2010.

[11] Carlijn V. C. Bouten, Karel T. M. Koekkoek, Maarten Verduin, Rens Kodde, and Jan D. Janssen, "A Triaxial Accelerometer and Portable Data Processing Unit for the Assessment of Daily Physical Activity", IEEE Transaction on Biomedical Engineering, Vol. 44, No. 3, pp. 136-147 (1997).

[12] O. Perrin, P. Terrier, Q. Ladetto, B. Merminod, and Y. Schutz, "Improvement of walking speed prediction by accelerometry and altimetry, validated by satellite positioning", Medical and Biological Enginnering and Computing, Vol.38, No. 2, pp. 164-168 (2000).

[13] Holger Junker, Paul Lukowicz, and Gerhard Troster, "Sampling Frequency, Signal Resolution and the Accuracy ofWearable Context Recognition Systems", ISWC 2004. 8th International Symposium on Wearable Computers, pp. 176-177 (2004).

[14] Stephen J Preece, John Y Goulermas, Laurence P J Kenney, Dave Howard, Kenneth Meijer and Robin Crompton, "Activity identification using body-mounted sensors - a review of classification techniques", Physiological Measurement, Volume 30, Number 4, pp. 1-33 (2009).

[15] Matthias Budde, Dawud Gordon, Hedda R. Schmidtke and Michael Beigl, "ActiServ: Activity Recognition Service for Mobile Phones",ISWC 2010. 14th International Symposium on Wearable Computers (2010).

[16] Minoru Fujimoto, Naotaka Fujita, Yoshinari Takegawa, Tsutomu Terada and Masahiko Tsukamoto, A Motion Recognition Method for a Wearable Dancing Musical Instrument, ISWC 2009. 13th International Symposium on Wearable Computers, pp. 11-18 (2009).

[17] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman, "Activity Recognition from Accelerometer Data", In Proceedings of the Seventeenth Conference on Innovative Applications of Artificial Intelligence, IAAI-05, pp. 1541-1546 (2005).

[18] Ling Bao and Stephen S. Intille, "Activity Recognition from User-Annotated Acceleration Data", Pervasive 2004. In Proceedings of second International Conference on Pervasive Computing, pp. 1-17 (2004).

[19] Keng-hao Chang, Mike Y. Chen, and John Canny, "Tracking Free-Weight Exercises", UbiComp 2007. Ubiquitous Computing, pp. 19-37 (2007).

[20] Jonny Farringdona, Andrew J. Moorea, Nancy Tilburyb, James Churchb and Pieter D. Biemonda, "Wearable Sensor Badge and Sensor Jacket for Context Awareness", In Proceedings of the Third International Symposium on Wearable Computers, pp. 107-113 (1999).

[21] Nicky Kern, Bernt Schiele and Albrecht Schmidt, "Multi-Sensor Activity Context Detection for Wearable Computing", Ambient Intelligence, LNCS 2875, pp. 220-232, (2003).

[22] Jani Mantyjarvi, Johan Himberg, Tapio Seppanen, "Recognizing Human Motion With Multiple Acceleration Sensors", Systems, Man, and Cybernetics, 2001 IEEE International Conference, VOL.2, pp. 747-752 (2001).

[23] Maja Stikic, Kristof Van Laerhoven and Bernt Schiele, "Exploring Semi-Supervised and Active Learning for Activity Recognition", ISWC 2008. 12th International Symposium on Wearable Computers, pp. 81-88 (2008).

[24] Andreas Zinnen, Ulf Blanke and Bernt Schiele, An Analysis of Sensor-Oriented vs. Model-Based Activity Recognition, ISWC 2009. 13th International Symposium on Wearable Computers, pp.93-100 (2009).

[25] Jonathan Lester, Tanzeem Choudhury, Nicky Kern, Gaetano Borriello and Blake Hannaford, A Hybrid Discriminative /Generative Approach for Modeling Human Activities, IJCAI 2005. Proceedings of the 19th international joint conference on Artificial intelligence, pp. 766-772 (2005)

[26] Daniel Siewiorek, Asim Smailagic, Junichi Furukawa, Neema Moraveji, Kathryn Reiger and Jeremy Shaffer, “SenSay: A Context-Aware Mobile Phone”, ISWC 2003. 7th International Symposium on Wearable Computers, pp. 248-249 (2005).

[27] Nobuo Kawaguchi, Nobuhiro Ogawa, Yohei Iwasaki, Katsuhiko Kaji, Tsutomu Terada, Kazuya Murao, Sozo Inoue, Yoshihiro Kawahara, Yasuyuki Sumi, Nobuhiko Nishio, “HASC Challenge: Gathering Large Scale Human Activity Corpus for the Real-World Activity Understandings”, Proceedings of the 2nd Augmented Human International Conference(AH2011), 27, 5pages(2011).

[28] Nobuhiro Ogawa, Katsuhiko Kaji, Nobuo Kawaguchi, “Effects of Number of Subjects on Activity Recognition - Findings from HASC2010corpus –“,1st International Workshop on Frontiers in Activity Recognition using Pervasive Sensing, pp. 48-51(2011).

[29] Nobuo Kawaguchi, Ying Yang, Tianhui Yang, Nobuhiro Ogawa, Yohei Iwasaki, Katsuhiko Kaji, Tsutomu Terada, Kazuya Murao, Sozo Inoue, Yoshihiro Kawahara, Yasuyuki Sumi, Nobuhiko Nishio, “HASC2011corpus: Towards the Common Ground of Human Activity Recognition”, Proceedings of the 13th International Conference on Ubiquitous Computing(UbiComp2011), pp.571-572(2011).

[30] Nobuo Kawaguchi, Nobuhiro Ogawa, Yohei Iwasaki, Katsuhiko Kaji, “Distributed Human Activity Data Processing Using HASC Tool”, Proceedings of the 13th International Conference on Ubiquitous Computing(UbiComp2011), pp.603-604(2011).