Modeling and Discovering Human Behavior from Smartphone Sensing Life-Log Data

Rischan Mafrur and Deokjai Choi*

School of Electronic and Computer Engineering, Chonnam National University, Buk-gu, Gwangju, 500-757, Korea

Advanced Network Lab,
Chonnam National University
Gwangju, South Korea

Introduction

Today, personal data is becoming a new economic asset. Personal data which generated from our smartphone can be used for many purposes such as identification, recommendation system, and etc. On other side, smartphones capability have increased significantly. Smartphone has equipped with high processor, bigger memory, bigger storage and etc. With this equipment, smartphones have capability to running complex applications. Many sensors also have embedded to the smartphone. With those sensors and log capability of smartphone, we can develop many useful systems or applications in different domains such as healthcare (elderly monitoring system [1, 2]) human fall detection [3, 4], transportation (monitoring road and traffic condition [5]), human happiness [6] and etc. To develop such systems, we have to collect user personal data and then analyze it.

In this research, our research purposes are to discover human behavior based on their smartphone life log data. Then we want to build behavior model which can be used for human identification. In this research, we have collected user personal data from 37 students during less than 2 months which consist of 19 kind of data sensors. When we are working in this field, we have to consider about realistic dataset. The definition of realistic dataset are: ,

1. In realistic environment, user has different types and brand of smartphone and each smartphone has different types of sensors and hardware specification and capabilities.
2. We could not expect the human actions and their activities, they will do actions and activities as they want.
3. There is no ideal data collection platform that can record user personal data for every day 24 hour non-stop, it will drain the battery and spend smartphone resources.
4. There is no ideal data collection that can record all of data without any data loss.

We have developed a new approach to build human behavior model which can deal with those situations. Our contribution in this work are:

(1) We have developed an application data collector which can collect user personal data and its following opportunistic method. This application does not bothering users, there is nothing to do after user install this application.
(2) We have developed system that can identify human behavior based on their smartphone personal data.
(3) Instead of identifying human behavior we also have developed system which can create human behavior model.

Conclusion

In this paper, we proposed approach that can used for user identification by building human behavior model. We use and combine of many sensors instead only focus on one sensors because we realize that sometimes user does not has data from one or more sensors. Based on our result, we can see that our approach is good enough for user identification.

We have tried also to remove one or more features and then observe the accuracy values. The result shows that even one or more features have been removed but our system still can be used for identification. It means our system can handle the problem if one or more data sensors from users smartphone not available. Some of result from our system can achieve up to more than 80 % accuracy but any four of them have less than 30 % accuracy. In this paper, we have explained also why four students have bad accuracy. The reasons are students who have bad accuracy, their dataset are too small and they have different behavior for almost each day which our approach does not capable to handle it. Despite some of accuracy values are under 30 % but those values still can be used for identification because those values are the highest one compared to others. It means that our approach still good enough for identification system.

References

1. T. Faetti and R. Paradiso, "*A Novel Wearable System for Elderly Monitoring*," Advances in Science and Technology, vol. 85, pp. 17-22, 2013.
2. P. Pierleoni, L. Pernini, A. Belli and L. Palma, "*An Android-Based Heart Monitoring System for the Elderly and for Patients with Heart Disease*," International Journal of Telemedicine and Applications, vol. 2014, p. 11, 2014.
3. L. Tong, Q. Song, Y. Ge and M. Liu, "*HMM-Based Human Fall Detection and Prediction Method Using Tri-Axial Accelerometer*," IEEE,Sensors Journal, vol. 13, no. 5.
4. O. Aziza, E. J. Parkc, G. Morid and S. N. Robinovitch, "*Distinguishing the causes of falls in humans using an array of wearable tri-axial accelerometers*," Gait and Posture, pp. 506-512, 2014.
5. P. Zhou, Y. Zheng and M. Li, "*How long to wait?: predicting bus arrival time with mobile phone based participatory sensing*," in MobiSys '12 Proceedings of the 10th international conference on Mobile systems, applications, and services.
6. A. Bogomolov, B. Lepri and F. Pianesi, "*Happiness Recognition from Mobile Phone Data*," in BioMedCom 2013, 2013.

Discussion

The data that we have is time series data, we have data for 1 month 20 days. First, we process those data, we explain in Data Processing section, then we applied our method to build human behavior model. In this research, we tried to looking the similarity pattern between human activities data by comparing between days. Fig. 1a shows that we use static window size (w=2), means we compare between two days, and find the similarity data. The way to find the similarity pattern can be seen on Fig. 1b and the details of our Algorithm can be seen on Algorithm 1.

Fig. 1a

Time	Sensor Name	Sensor Value
13:00	location	same
13:00	wifi	1-AP, iptime
14:00	runapps	kakao
14:00	location	long
15:00	runapps	kakao
15:00	location	little

Time	Sensor Name	Sensor Value
13:00	location	same
13:00	wifi	1-AP, iptime
14:00	battery	charging
14:00	wifi	D-link
15:00	runapps	kakao
15:00	location	little

Group-1 = 13:00,location,same | 13:00,wifi,1-AP,iptime
Group-1 = 13:00,location,same | 13:00,wifi,1-AP,iptime
Group-2 = 15:00,runapps,kakao | 15:00, location, little
Group-2 = 15:00,runapps,kakao | 15:00, location, little

Fig. 1b

Algorithm (Similarity Detection)

```
Data : D, w
Result : All Detected Group in a Window
grpAll, grpTemp, grpPrevious <- NULL
dataValue, dataValueNext <- NULL

while (D in w) for all of D do
  dataValue <- D.current.day
  dataValueNext <- D.next.day
  grpTemp <- findingSimilarPatterns(dataValue,
  dataValueNext)

  if (grpTemp in grpPrevious)then
     grpNew <- merge(grpPrevious, grpTemp)
     grpAll <- add(grpNew)
  else
     grpAll <- add(grpTemp)
```

Algorithm 1

Data Collection and Processing

Personal Data Collector Application

In this research we did not develop our application from scratch, we use ***Funf library***. This application is following opportunistic sensing, means it does not need user intervention. User only install this application and everything is done, this application will automatically collected all of their sensor data. This application collected 19 types of sensor data, but in this research, we only use nine sensors data are activity, GPS, Nearby Bluetooth, Nearby Wi-Fi, call, SMS, Battery, Current Run Apps, Phone Screen.

Data Processing

Our dataset is quite big enough, when we load all of those data in the same time it will spend computer resource especially RAM because to process data, R environment system load all of data that will be process in RAM. To handle that problem, we have to define what kind of data that we want to use and store those data to another file (temporary file), in this case, we use csv files. We have three kind of preprocessing modules and each module will store new data to csv file.

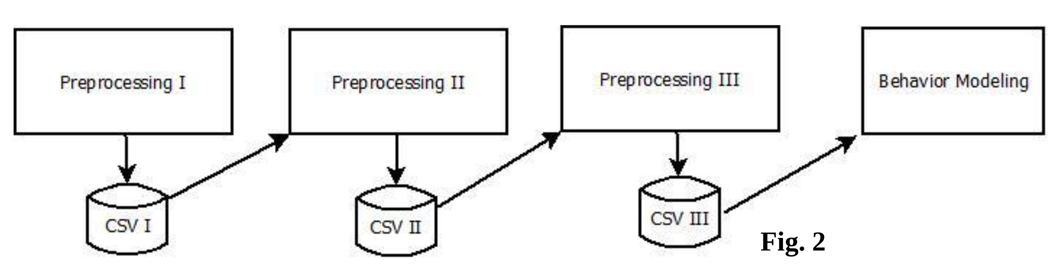

Fig. 2

Figure 2 shows preprocessing process and dataset transformation from preprocessing I until behavior modeling module. Preprocessing I will load all of raw data, removing duplication data, cleansing data, and select the most important data that have been defined. Preprocessing I will store the result data to the CSV I database. Preprocessing II will load the CSV I data not the raw data, in this process features extraction applied. The result of Preprocessing II stored in CSV II. Preprocessing III load the CSV II data and transform the data to the fit format before creating behavior model applied. This ways will reduce time processing and computer resource's usage.

Human Behavior Identification

What is the human behavior in case of smartphone sensing?.

Human daily activities which carried out continuously

In terms of human daily activities, we have to consider about four things:

- **What kind of activity (e.g meeting, studying, exercising)**
- **When (e.g around 9 AM)**
- **Location (e.g Lab)**
- **Human Interaction (e.g all lab's members)**

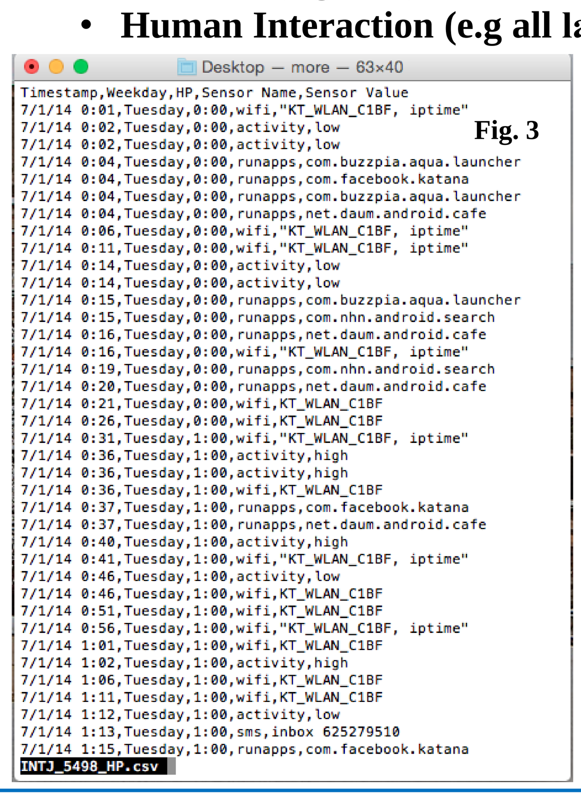

Fig. 3

Based on this definition, we use 9 sensors data which I have mentioned before then we extracting the features from those data.

The output of Preprocessing III (before modeling behavior applied can be seen on Fig. 3).

After we got that output, we applied our method to build behavior model using Algorithm 1, and the way for finding similarity pattern can be seen on Fig. 1b.

Experiment and Result

The total of dataset which collected around 1 month and 20 days. We divide all dataset to two parts, first month for creating model (first dataset) and remaining dataset for testing performance (second dataset). We use first dataset for modeling human behavior, we call it B1 (behavior model/profile). Then, we extract and process second dataset, applying similarity detection to second dataset with same setting, and we call the result is B2 (set of behavior/group activities from second data set). The question are, is all of B2 identified by B1?, how many set of group activities (B2) which identified by B1, then calculate the percentage of data identified.

Result of User Identification

	Table 1	TEST					
MODEL		ENFP_0719	ENFP_2012	INTJ_5498	ISTJ_3052	ESTJ_5190	ESFP_4634
	ENFP_0719	**67.922**	0	0.4	2.187	0	1.943
	ENFP_2012	0	**83.582**	0	0	0	0
	INTJ_5498	2.178	0	**75.977**	2.087	0	3.401
	ISTJ_3052	2.289	0	0.4	**93.439**	8.232	1.943
	ESTJ_5190	0	0	0	0.099	**22.866**	0
	ESFP_4634	2.289	0	0.977	2.087	0	**89.686**

Table 1 is not confusion matrix table, it just looks like confusion matrix table. The value means the percentage of B2 (behavior data from test dataset) which is successfully identified by B1 (behavior model). We can see that our proposed features and our approach can be used for identification. Based on the result and our observation, our approach can achieved good enough accuracy even some of users has bad accuracy (under 30%).
Full Table can be seen on Appendix, page : 42.

Model Performance Testing

Performance Evaluation by Removing some Sensors Data

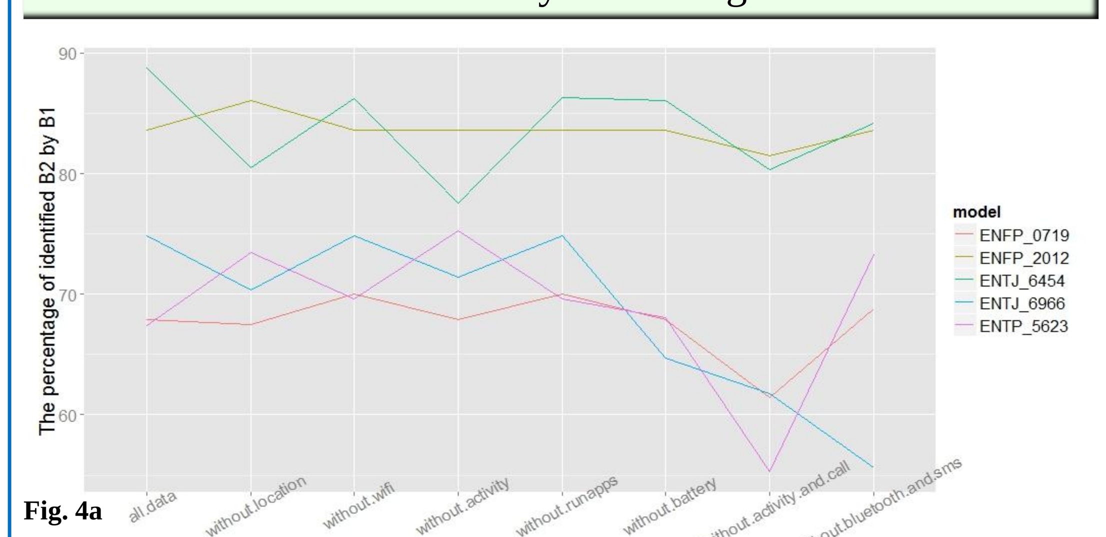

Fig. 4a

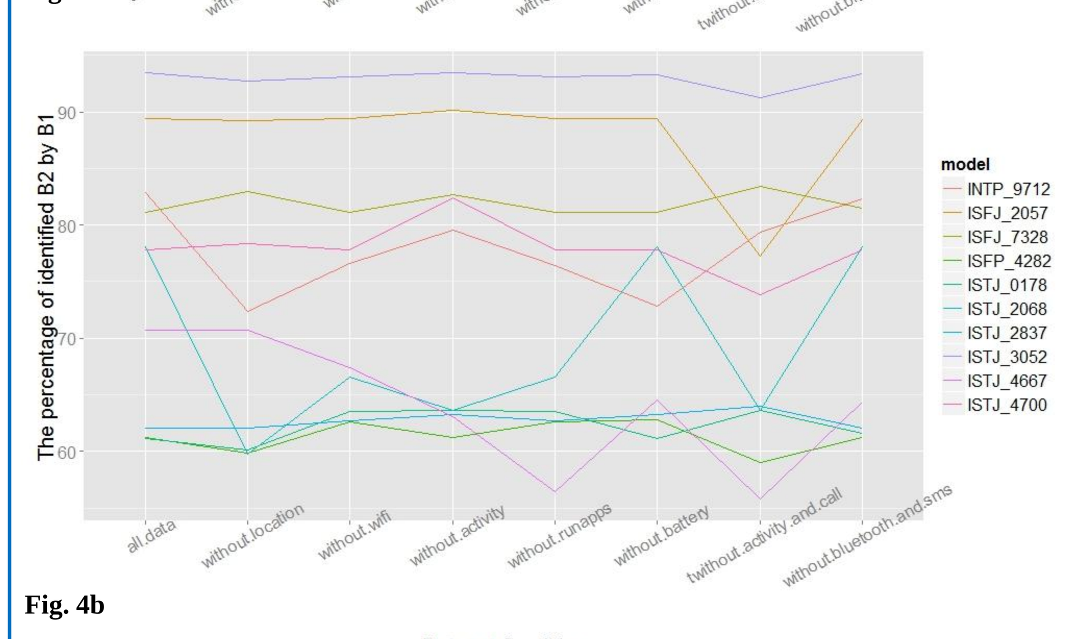

Fig. 4b

Full Figure can be seen on Appendix, page : 48.

Performance Evaluation by Removing some Sensors Data. When we doing research in this field, we have to realize that some sensors probably does not supported by users smartphone or probably user does not have any data in one of sensor such as user does not have SMS and call log. We have to consider about that, if we focus only one sensor, it will be problem. We want our approach can dealing well with realistic data, so we tried to remove one and more sensors data and then observe the accuracy. Based on the result, we can conclude that by removing one or two features our approach still good enough for user identification.

Chonnam National University
Advanced Network Lab