Towards Recognizing Person-Object Interactions using a Single Wrist Wearable Device

Juhi Ranjan

University of Virginia
Charlottesville, VA 22903, USA
juhi@virginia.edu

Kamin Whitehouse

University of Virginia
Charlottesville, VA 22903, USA
whitehouse@virginia.edu

Abstract

Activity recognition (AR) is an important part of context-aware applications. In this paper, we focus on an indirect AR method: by sensing the objects that a person is using. Objects that provide functional utility to their user, also indicate the type of activity that their user is doing. For example, the use of a hair dryer indicates that its user is grooming their hair. In this paper, we discuss an approach to sense the objects that the person interacts with, using only a single wearable device on the wrist of the person. Wearable devices typically have an IMU sensor, which can sense several aspects of the person's hand gestures, such as acceleration, and orientation. We collect a dataset of 17 different object interaction gestures using 5 participants in a test home. We evaluate the object gestures using supervised and unsupervised machine learning approaches. Our study reveals that we can recognize object interactions with 83-91% accuracy in the supervised approach, and 58-66% accuracy in the unsupervised approach.

Author Keywords

Person-Object Interaction, Activity Recognition, Wearable Devices, Supervised Learning, Unsupervised Learning

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

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. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
UbiComp/ISWC'16 Adjunct, September 12-16, 2016, Heidelberg, Germany
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4462-3/16/09...\$15.00.
DOI: <http://dx.doi.org/10.1145/2968219.2968279>

Introduction

The ability to perform unsupervised activity recognition has long been the holy grail for many different context-aware computing researchers. And for a good reason. Activity recognition can enable quantification of a person's self-care abilities and daily routine. Health care professionals are interested as it can help them track if elderly are able to live independently in homes and perform Activities of Daily Living (ADLs) [12]. Self-Quantification is a also big thrust these days, as people want objective insights on how they spend their time. While the context of a person's activities, is helpful in a multitude of ways, expecting the end-user to invest a significant time in training an AR system can prove to be one of the biggest deterrent to its commercial success. In fact, some of the most widely accepted AR systems are all equipped with unsupervised learning techniques, and can work out of the box. For example, Google Fit automatically detects if a person is Walking/Running/Biking etc., and gives an estimated calorie expenditure based on the activity performed.

In the current state of the art, most of the commercial AR products are able to automatically track the ambulatory activities such as walking, running, swimming etc. Part of the success of these products is due to the fact that they do not require 1. training, or 2. extra devices as part of the setup. Using a device is as easy as buying it or installing an application on a smartphone. However, there is limited commercial success in developing systems that can automatically recognize ADLs. The main difference in these two types of activities is that ambulation typically involves a more predictable, repetitive patterns of motion, typically carried out for continuous period of time, which is not the case with ADLs. The challenge in detecting ADLs such as grooming, cooking etc. arises from the fact that they typically do not have a repetitive motion, or a movement pattern

that is consistent across all people.

Due to this challenge, researchers have considered two different types of approaches in detecting ADL - 1. Direct inference - this method uses on-body or infrastructural sensors in determining the unique set of motion and posture of different body parts to detect what ADL they are performing, 2. Object-Interaction - this method uses sensors to determine the objects that a person is interacting with, in order to infer the activity that they may be doing. ADL recognition solutions can be implemented as a combination of two main factors: 1. location where devices are installed (on body vs. infrastructure), and 2. how many devices are required (single vs. multiple). Typically direct inference based AR uses multiple, on-body sensor devices [2], while object-interaction based AR uses a single on-body, and multiple infrastructural sensors [10].

In this paper, we discuss an object-interaction based approach to detect activities that a person is performing using only a single wearable device on the wrist of the person. Given that many different people have already started using a single fitness tracking device on their body that can directly sense their ambulatory activities, we would like to leverage this technology and explore its potential in recognizing activities. The wearable device has an IMU sensor on it which can sense several aspects of the person's hand gestures. We perform feature extraction on the IMU data and evaluate its performance in determining how accurately we can recognize different object interaction gestures. We instrumented 12 objects in a home, including lights, water fixtures, and major appliances, and annotate 17 different object-interaction gestures. We conducted a study with 5 participants, who used the different objects in the test home. Our study reveals that our approach can recognize object interactions with 83-91% accuracy using a super-

vised learning approach, and 58-66% accuracy using an unsupervised learning approach.

Background: Person-Object Interactions

There are different approaches for recognizing what object is being used by a person, that vary in terms of the sensing technology being used and Some methods propose instrumenting the objects with sensors that can detect who is interacting with them. For example, Hodges et. al. show that it is possible to recognize an individual based on how they touch objects by equipping objects with pressure sensors or other smart surfaces [7]. Chang et. al use accelerometers embedded in a television remote control or mobile device to identify household members, based on the unique way each person uses the remote [4, 13].

Other proposed methods recognize an object user in a house by instrumenting all objects or fixtures with RFID tags [3, 9], and RFID reader devices on the occupants' hands. Cheng et. al use coordinate level tracking in a study and determine which person is using an object or appliance, they use a proximity detection system where users carry a magnetic beaconing system which is detected by special receivers near appliances [5].

Lee et al. use tracking systems to monitor people's movements in the rooms of a house [8]. They identify the occupants' individual bedrooms based on their frequency of movement to the rooms, and assign all object usage in those rooms to the person. In shared spaces, they use tracking to identify people present in the room when an object is used. If multiple people are present in the room, they simply assign the object usage equally between all people present in the room. In another work [10], researchers explore the idea of using heuristics augmented coarse-grained room level localization system to identify the object

user and achieve results comparable to more fine-grained localization system. In our previous work [11]. we explored the idea that we can use hand gesture made by people to determine who used an object. In this method, when an object, such as microwave, is used, we look for the person who made the gesture required to use the object, such as the microwave door pulling gesture. The biggest drawback of this method is that it requires some system that can monitor when every object is used in a space. In this paper, we perform a detailed analysis of the different object gestures per participant, using both supervised and unsupervised methods. We also explore a richer set of IMU features to characterize object gesture. We finally discuss the challenges that needs to be solved for object-interaction recognition to work out of the box in near future.

Study

In this section, we describe the study that we perform in order to evaluate supervised and unsupervised method of recognizing object interaction gestures.

Experimental Setup

We instrumented 12 major appliances, light and water fixtures in a two-bedroom apartment. The relative locations of all objects are shown in Figure 1.

We placed direct sensors on all objects to monitor their usage. We used magnetic reed sensors on appliances with hinged doors, such as Fridge (Frid), Microwave (Micr), Freezer (Freez). We also used magnetic reed sensors to instrument the bathroom hot (B-F-H) and cold (B-F-C) faucets, as well as the kitchen hot (K-F-H) and cold (K-F-C) faucets. The reed sensors were plugged into HOBO UX90-001 data loggers. We used light on/off sensing HOBO UX90-002M data loggers to sense the Lights (L) in the house. These loggers are installed right next to the light

Figure 2: Light switches K-L and L-L

Figure 3: Hot and cold water kitchen faucets K-F-C and K-F-H

Figure 4: Hot and cold water bathroom faucets B-F-C and B-F-H

Figure 5: Fridge (Frid), Freezer (Freez) and Microwave (Micr)

Figure 1: Object layout: We instrumented 5 lights, 4 faucets and 3 appliances in the test home

bulbs, and have a programmable threshold for light intensity to detect when a light is on.

To sense the hand gestures, we used LG G watch, which runs on Android Wear platform. We wrote an app for the watch which collected IMU based sensor data (accelerometer, gravity and orientation vector) at 33 Hz and transmitted it to its paired smartphone. The smartphone had a listener app which logged the received data on the phone. For each sensor data sample, the app recorded the timestamp, the sensor type, and the set of raw sensor values along x, y and z axes.

Data Collection

We invited 5 participants (3 F, 2 M) to use the objects of the instrumented home. The procedure of this study required each participant to follow a script that made them operate each of the 12 objects in the house in a fixed order. The participants were free to operate each object in a manner they liked. All the objects in the home were labeled for easy reference. To discourage participants from mechanically using the objects in the same way, we designed the script

to have consecutive object usages in different rooms. They were asked to perform the entire object usage script ten times, therefore logging 120 object usages each. In order to obtain clean gesture data for this study, the participants were instructed to pause for 5 seconds before and after using each object. This made annotation of gesture data extremely easy, and we could automate the process. Every time that an object was used, a +/- 5 seconds segment of sensor data from the object usage timestamp was extracted. The only gesture in this time window was the one used for operating the object.

Challenging Dataset: Similar Objects

The dataset that we collected was challenging from a recognition perspective, as many of the objects that were instrumented were similar. Overall, 12 objects were instrumented, out of which 5 were light fixtures, 4 were faucets (including hot and cold), and 3 were appliances with doors.

The 5 light fixtures were - Living room light (L-L), Kitchen light (K-L), Bathroom light (Ba-L), Bedroom light (Be-L), and Study light (S-L). The challenges with lights are:

1. All these lights are operated by the same type of switches, and the switch for Living room light and Kitchen light, are located on the same switch board.
2. For each of these lights L, we want to recognize their 'On' gesture (L-On), as well as 'Off' gesture (L-Off). As the 'On' and 'Off' gestures are very similar - 'On' is flicking up the switch, and 'Off' is flicking down the switch, it makes this dataset further challenging.

The 4 faucets were - Bathroom hot water faucet (B-F-H), Bathroom cold water faucet (B-F-C), kitchen hot water faucet (K-F-H), and kitchen cold water faucet (K-F-C). The challenges with the faucets are:

1. Bathroom hot and cold water faucets are similar fixtures, which are located on the same sink, making their usage

gestures very similar. Same is case with Kitchen hot and cold water faucets.

2. Kitchen hot and cold faucets were knobs that opened the same way, so their usage gesture was exactly the same.

The three appliances - Microwave (Micr), Fridge (Frid) and Freezer (Freez) had challenges as well:

1. All three appliances had doors, which means the gesture to interact with them was a door pull gesture

2. All three appliances were installed against the same wall, and faced the same direction

Feature Extraction

Since objects have a unique combination of interface, orientation and location within a space, the hand gestures for using them in terms of the hand's acceleration signature, tilt and compass direction are also unique. In order to sense these three parameters of a hand's motion, we use the nine-axis Inertial Measurement Unit (IMU) sensor of a wrist wearable device. Data from accelerometer, gyroscope and magnetometer can be merged to infer the 3-D compass direction of the hand. In our work, we used pre-processed direction information available from Android Wear API [1].

We then perform feature extraction for each data segment, and then evaluate it using supervised and unsupervised learning approaches.

Segmentation of the sensor data is based on the the observation that every time a hand interacts with an interface, there is a small pause before and after the interface is used. This results in a peak in the acceleration. Therefore, to segment the data, we detect the clusters of peaks of acceleration, and characterize them by performing feature extraction.

Features were extracted for all 9 data streams - Accelerometer - x, y and z, Gravity - x, y and z, and Orientation - x, y and z. Features extracted were of two types - 1. features of the raw signal data, and 2. features of peaks detected in the data

1. Raw Signal - Mean, Median, Std. Deviation, 25th Percentile, 75th Percentile, and Zero Crossing Rate

2. Peaks - Total number of Peaks, Min, Max, Median, Std. Deviation of peak attributes such as width, prominence, and height.

Supervised Learning Approach

We first evaluate the ability of a single wearable device based system to recognize different object interaction gestures. As was discussed earlier, we instrumented 12 different objects, which included 3 appliances, 4 faucets and 5 lights. For each light fixture, we include the 'On' as well as 'Off' gesture. Therefore, we evaluate the recognition of 17 gestures in total.

To evaluate the feasibility of our approach, we first implement a supervised machine learning technique in differentiating between the 17 object interaction gestures. For the machine learning algorithm, we selected Random Forest. We chose this as it has been shown to have a high classification accuracy in a variety of datasets [6]. We perform 10-fold cross validation of the gestures for each participant's dataset. We use cross validation per person, as we imagine each person's gestures for using different objects will be unique because of the differences in their heights, force applied, length of arm. etc. Therefore the results are indicative of the situation where a person buys a wearable device, and trains it individually before using it.

Figure 6 shows the confusion matrix obtained in using IMU based features for classification for every participant. Over-

Figure 6: Supervised learning: Classification accuracy of using IMU based features for each person varied from 83-91%

all, the classification accuracy ranged from 83-91% accuracy. Considering that the dataset consists of very similar object interaction gestures, and the fact that we are using only one device on the wrist of the person, the IMU based features perform very well.

A further analysis of the results show the strengths and weaknesses of this approach. The biggest challenge in this dataset was the presence of similar object gestures. In many cases, the machine learning algorithm was able to

detect the differences between very similar gestures. In Figure 6(a), it was able to recognize 'On' and 'Off' gesture for Bathrooms light (Ba-L-On and Ba-L-Off) with 95% and 90% accuracy. It was able to recognize Kitchen hot water faucet (K-F-H) with 100% accuracy. It was also able to recognize appliances Microwave (Micr), Fridge (Frid) and Freezer (Freez) with an accuracy of 100%, 100% and 89% accuracy respectively. It was able to recognize Bathroom hot water faucet (Ba-F-H) with 90% accuracy. In Figure 6(b), it was able to recognize gestures for Bathroom and Bedroom

Figure 7: Unsupervised Learning: Clustering the gestures using K-Means, with K=17 resulted in a 33-42% clustering errors

light 'On' and 'Off' gestures (Ba-L-On, Ba-L-Off, Be-L-On, Be-L-Off), and Bathroom hot and cold water faucets (B-F-H and B-F-C) with 100% accuracy. It was able to recognize appliances Micr, Frid and Freez with an accuracy of 90%, 100% and 90% respectively. In Figure 6(c), it was able to recognize Kitchen hot and cold water faucets (K-F-H, K-F-C), appliances Micr, Fridge, and Bathroom hot water faucet with 100% accuracy. In Figure 6(d), it was able to recognize the Kitchen hot water faucet (K-F-H) with 100% accuracy. It recognized some of the other gestures such as Ba-L-On,

Ba-L-Off, Be-L-On, K-F-H, B-F-H and B-F-C with 90% accuracy. In Figure 6(e), it recognizes Ba-L-On with 95% accuracy.

While the approach was able to recognize many gestures accurately, it made mistakes, especially when it came to discerning between object gestures that have interfaces which are alike. For example, in Figure 6(a), 50% of K-L-On were classified as L-L-On, and similarly 30% L-L-On were classified as K-L-On. Also, 10% of B-F-H were classified as

B-F-C, and 22% of B-F-C were classified as B-F-H. In Figure 6(b), 20% of K-F-H gestures were classified as K-F-C, and 30% of K-F-C were classified as K-F-H. In Figure 6(c), 40% of Be-L-Off are classified as Ba-L-Off. In Figure 6(d), many of the light on/off gestures get errors in identifying the correct fixtures, although they do identify the gesture as 'On' or 'Off' correctly. For example, this happens with gestures of Be-L-Off, K-L-On, K-L-Off, and L-L-On. In Figure 6(e), 11% of Micr gestures are classified as Freez, and 11% of Freez gestures are classified as Frid. 20% of B-F-H gestures are classified as B-F-C, and 10% of B-F-C are classified as B-F-H.

Unsupervised Learning Approach

While a supervised machine learning approach requires the use of a training dataset, for an activity recognition device to work out of the box, it is important to explore ways to discover object interaction gestures of a person, in an unsupervised manner. To test how an unsupervised learning approach works with our IMU features, we use a k-means clustering algorithm. We set the number of clusters, or k , to 17, based on the number of object gestures that are present in the dataset. We apply the clustering algorithm to each participant's set of object interaction gestures.

The results of the clustering algorithm are shown in the form of confusion matrices in Figure 7. The clusters formed are along x-axis, while the actual object gestures are along y-axis. The number in each cell represents the percentage of gestures for a given gesture (row) present in the given cluster (column). Overall, the set of IMU features resulted in 33-42% clustering error, where clustering error is defined as the ratio of total number of non-majority class elements present in a cluster, to the total number of elements.

In Figure 7(a), we observe that some of the light 'Off' gestures are clustered in the same group. For example, Ba-L-Off, Be-L-Off and K-L-Off are clustered into same groups. We also observe some light 'On' gestures that are clustered in same groups, for e.g. K-L-On and L-L-On. Some 'On' and 'Off' gestures for the same object are also clustered together, for e.g. S-L-On and S-L-Off, and Be-L-On and Be-L-Off. The bathroom hot and cold faucets B-F-H and B-F-C are clustered together, but in two separate groups. Same is the case with kitchen hot and cold faucets, K-F-H and K-F-C. In Figure 7(b), sometimes gestures of two different object types are sometimes clustered together, for e.g. 1. Micr and Frid are clustered with L-L-Off and K-L-Off, and 2. K-F-C is clustered with K-L-On, 3. Frid is clustered with S-L-On, and 4. Freez is clustered with B-F-H. In Figure 7(c), the gesture for B-F-H and B-F-C is clustered with 100% precision. 80% of the Frid gestures clustered in its own group, while 20% are clustered with K-F-H. 40% of Micr gestures are clustered with Freez. While the gesture for light 'On' and 'Off' have split across multiple groups, the 'On' gestures for different lights have clustered together in the same groups, as have the 'Off' gestures. In Figure 7(d), we observe that the bathroom faucets B-F-H and B-F-C are clustered correctly into individual groups, while the kitchen faucets K-F-H and K-F-C are clustered in one group. As was the case with Figure 7(c), the gestures for the same light have split across multiple groups, whereas 'On' and 'Off' gestures across different light have clustered together. In Figure 7(e), the bathroom faucet gestures B-F-C and B-F-H cluster together in two different groups, and K-F-C and K-F-H cluster in the same group. The gestures of Micr, Frid and Freez are split into different groups, and clustered together.

Figure 6 revealed that it is possible to differentiate between similar object gestures such as the hot and cold faucets of bathroom sink, different lights etc. However, when cluster-

ing is performed, many of the gestures of similar object get grouped together. While there seem to be many clustering errors in the dataset, it shows some promise in that the gestures of different object types such as lights, and appliances are clustered separately.

Challenges

While detecting person-object interactions using smart wearable devices seems promising specially given that smart devices are finding greater commercial acceptance, there is still a long way to go before ADL recognition can be integrated into a robust product. There are several signal processing challenges which must be addressed to get there.

Segmentation

Activities which current devices are successful in detecting, such as walking, running, biking etc., have one thing in common - they are all repetitive by nature. Object interaction gestures on the other hand are not. Therefore, one of the main technical challenges that needs to be solved is to figure out a method for segmenting the object gesture, given a continuous signal with other gestures and noise present.

Recognizing non-object gestures

While object interaction gestures have a fixed and repetitive pattern present, so do other gestures. In performing unsupervised learning of patterns of gesture, it is important to know how to discern between object and non-object gestures.

Using a single device

Using a single point of measurement, for e.g. the wrist provides limited context about the gesture. For example, It becomes challenging to determine what posture a person is in when performing a certain gesture. Also, other challenges

with using a single wrist wearable device is that, if the person uses the other arm to interact with an object, the gesture may not be detected by the device

Bootstrapping the device

The list of the types of objects that a person can possibly interact with in order to perform ADLs is quite large, and as such expecting the wearable device to start working out of the box without any gesture reference database, seems challenging. It would be easier if a representative set of gestures could be identified in order to apply a semi-supervised learning approach.

Recognizing objects with variable locations

In the current scope of the work, we are only looking at features required to recognize objects with fixed locations. The heuristics of orientation, and action will have to change significantly when we broaden our scope to include objects which can be moved, such as water kettle, comb, etc.

Conclusions

In this paper, we discuss an approach to detect activities that a person is performing by sensing the objects that the person interacts with. We propose doing so using only a single wearable device on the wrist of the person. The wearable device has an IMU sensor on it which can sense several aspects of the person's hand gestures. We perform feature extraction on the IMU data and evaluate its performance in determining how accurately we can recognize different object interaction gestures. We apply supervised and unsupervised learning methods to the data, and determine that while supervised method far outperforms the unsupervised approach, there is still promise in performing further exploratory studies in unsupervised methods of recognizing object interactions.

Acknowledgements

This work is supported by the National Science Foundation under Grants 1038271 and 1305362 and the National Science Foundation Graduate Research Fellowship Program under Grant 0809128.

REFERENCES

1. Android 2015. Android Wear Orientation Sensor. (2015). http://developer.android.com/reference/android/hardware/Sensor.html#TYPE_ROTATION_VECTOR.
2. Ling Bao and Stephen S Intille. 2004. Activity recognition from user-annotated acceleration data. In International Conference on Pervasive Computing. Springer, 1–17.
3. Eugen Berlin, Jun Liu, Kristof Van Laerhoven, and Bernt Schiele. 2010. Coming to grips with the objects we grasp: detecting interactions with efficient wrist-worn sensors. In Proceedings of the fourth international conference on Tangible, embedded, and embodied interaction. ACM, 57–64.
4. Keng-hao Chang, Jeffrey Hightower, and Branislav Kveton. 2009. Inferring identity using accelerometers in television remote controls. In Pervasive Computing. Springer, 151–167.
5. Yun Cheng, Kaifei Chen, Ben Zhang, Chieh-Jan Mike Liang, Xiaofan Jiang, and Feng Zhao. 2012. Accurate real-time occupant energy-footprinting in commercial buildings. In Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings. ACM, 115–122.
6. Manuel Fernández-Deado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? The Journal of Machine Learning Research 15, 1 (2014), 3133–3181.
7. Mark R Hodges and Martha E Pollack. 2007. An object-use fingerprint: The use of electronic sensors for human identification. In UbiComp 2007: Ubiquitous Computing. Springer, 289–303.
8. Seungwoo Lee, Daye Ahn, Sukjun Lee, Rhan Ha, and Hojung Cha. 2014. Personalized energy auditor: Estimating personal electricity usage. In Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on. IEEE, 44–49.
9. Matthai Philipose, Kenneth P Fishkin, Mike Perkowitz, Donald J Patterson, Dieter Fox, Henry Kautz, and Dirk Hahnel. 2004. Inferring activities from interactions with objects. Pervasive Computing, IEEE 3, 4 (2004), 50–57.
10. Juhi Ranjan, Erin Griffiths, and Kamin Whitehouse. 2014. Discerning electrical and water usage by individuals in homes. In Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings. ACM, 20–29.
11. Juhi Ranjan and Kamin Whitehouse. 2015a. Object hallmarks: Identifying object users using wearable wrist sensors. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 51–61.
12. Juhi Ranjan and Kamin Whitehouse. 2015b. Rethinking the fusion of technology and clinical practices in functional behavior analysis for the elderly. In Human Behavior Understanding. Springer, 52–65.
13. Matt Wolff. 2013. Behavioral Biometric Identification on Mobile Devices. In Foundations of Augmented Cognition. Springer, 783–791.