Contents lists available at ScienceDirect

Int. J. Human–Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

An investigation of using mobile and situated crowdsourcing to collect annotated travel activity data in real-word settings

Yung-Ju Chang[a,b,*], Gaurav Paruthi[a], Hsin-Ying Wu[a], Hsin-Yu Lin[a], Mark W. Newman[a]

[a] School of Information, University of Michigan, Ann Arbor, MI, USA
[b] Department of Computer Science, National Chiao Tung University, Hsingchu City, Taiwan

ARTICLE INFO

Keywords:
Mobile crowdsourcing
Crowdsensing
Wearable camera
Annotated activity data collection
Travel activity

ABSTRACT

Collecting annotated activity data is vital to many forms of context-aware system development. Leveraging a crowd of smartphone users to collect annotated activity data in the wild is a promising direction because the data being collected are realistic and diverse. However, current research lacks a systematic analysis comparing different approaches for collecting such data and investigating how users use these approaches to collect activity data in real world settings. In this paper, we report results from a field study investigating the use of mobile crowdsourcing to collect annotated travel activity data through three approaches: *Participatory*, *Context-Triggered In Situ,* and *Context-Triggered Post Hoc*. In particular, we conducted two phases of analysis. In Phase One, we analyzed and compared the resulting data collected via the three approaches and user experience. In Phase Two, we analyzed users' recording and annotation behavior as well as the annotation content in using each approach in the field. Our results suggested that although Context-Triggered approaches produced a larger number of recordings, they did not necessarily lead to a larger quantity of data than the Participatory approach. It was because many of the recordings were either not labeled, incomplete, and/or fragmented due to the imperfect context detection. In addition, recordings collected by the Participatory approach tended to be more complete and contain less noise. In terms of user experience, while users appreciated automated recording and reminders because of their convenience, they highly valued having the control over what and when to record and annotate that the Participatory approach provided. Finally, we showed that activity type (Driver, Riding as Passenger, Walking) influenced users' behaviors in recording and annotating their activity data. It influenced the timing of recording and annotating using the *Participatory approach*, users' receptivity using the Context-Triggered In Situ approach, and the characteristics of the content of annotations. Based on these findings, we provide design and methodological recommendations for future work that aims to leverage mobile crowdsourcing to collect annotated activity data.

1. Introduction

The design of context-aware systems has been a topic of long-standing concern in the HCI and Ubiquitous Computing communities (Abowd et al., 1999; Chen et al., 2000). Researchers and practitioners in these fields are seeking to develop systems aware of users' context and activity, thereby providing relevant information and/or services to the users. A common practice in context-aware system development is collecting contextual data representing user activities and contextual conditions that the system is expected to encounter when they are deployed in the field (Newman et al., 2010). Such data are needed for training and evaluating recognizers that detect important contextual states and trigger system responses (Dey et al., 2001), and are also important for use in the prototyping and evaluation stages of system development (MacIntyre et al., 2004; Newman et al., 2010). An essential step in collecting these activity data is to collect labels and annotations describing the data. These metadata not only allow developers to train and test their recognizers but also enable them to more easily filter and select suitable sets of data for testing the functionality of the system.

Researchers have used different ways to collect annotated contextual and activity data, including recording and annotating data on their own (DeVaul and Dunn, 2001) and using a structured participant-based approach, i.e. recording and annotating data with a small sample of people performing predefined activities in a controlled environment under the researchers' guidance (Bao and Intille, 2004; Kwapisz et al.,

[*] Corresponding author at: Department of Computer Science, National Chiao Tung University, Hsingchu City, Taiwan.
E-mail addresses: armuro@cs.nctu.edu.tw (Y.-J. Chang), gparuthi@umich.edu (G. Paruthi), hsinying@umich.edu (H.-Y. Wu), hyclin@umich.edu (H.-Y. Lin), mwnewman@umich.edu (M.W. Newman).

http://dx.doi.org/10.1016/j.ijhcs.2016.11.001
Received 25 February 2016; Received in revised form 2 November 2016; Accepted 8 November 2016
Available online xxxx

1071-5819/ © 2016 Elsevier Ltd. All rights reserved.

2011). As sensor-laden smartphones have become pervasive, researchers have started exploring ways to leverage a larger number of mobile smartphone users—sometimes referred to as the mobile crowd—to record and annotate targeted activities using their smartphones in real-world settings (Abowd et al., 1999; Chang et al., 2012; Newman et al., 2010). Two broad approaches are commonly used for crowd-based data collection and annotation. P*articipatory* data collection (Ganti et al., 2011; Kanhere, 2011) refers to the process in which mobile users actively participate in collecting data; they manually control an instrument to collect data based on their interpretation of researchers' needs and instructions (Paxton and Benford, 2009). In contrast, *Opportunistic* data collection (Ganti et al., 2011; Lane et al., 2010) refers to the process in which the instrument automates data recording: the mobile users carry an instrument that records data itself based on a certain sampling heuristic—continuous, randomized, schedule-based, or context-triggered (Froehlich et al., 2007; Meschtscherjakov et al., 2010). To obtain users' annotations, instruments can be programmed to prompt users to annotate during an activity to obtain *in situ* annotations, or to prompt users afterwards to obtain *post hoc* annotations.

Using these methods to collect contextual and activity data via the mobile crowd in real-world settings has a considerable advantage compared to a controlled data collection method: the collected data are more diverse, naturalistic, and representative of users' real-life behaviors. However, because the data collection process is not under researchers' supervision, the quantity and the quality of the collected data have been a major concern of researchers. At present, we have limited understanding of which approaches can reliably and effectively produce high quantity and quality of data, and this fact in turns limits the usefulness of mobile crowdsourcing for collecting activity data.

Due to this limitation, in recent years, research has started assessing the quality of labeled activity data collected in the field. For example, Cleland et al. (Cleland et al., 2014) showed that collecting labeled physical activity data in the field using a Context-Triggered Experience Sampling Method (ESM) approach obtained equally accurate labels compared to those obtained in a controlled lab study. However, in this study, the authors neither analyzed the quantity and quality of the collected activity recordings nor analyzed user experience and behavior in using the approach. In addition, the controlled lab studies they compared were not performed in real word settings, meaning that the Context-Triggered ESM was the sole approach being performed in the field. Thus, it remains unclear whether or not the Context-Triggered ESM is a more reliable and effective approach for collecting activity data compared to other approaches such as the Participatory approach. The purpose of this paper, therefore, is to compare the effectiveness of different approaches for collecting annotated activity data via the mobile crowd and understand users' behavioral pattern in using these approaches to record and annotate activity data. We believe such an investigation can shed lights on how to design better an activity data collection tool and method and for collecting annotated activity data via mobile crowdsourcing.

In this paper, we report findings from a two-week field study involving the mobile crowd using three approaches to collect annotated travel activity data in real-world settings, namely, *Participatory (PART)*, *Context-Triggered In Situ* (SITU), and *Context-Triggered Post Hoc* (POST). These three approaches were performed by 37 smartphone users to collect their travel activity data when they were traveling outdoors using our instrument. To obtain the ground truth of their travel activity during the study, we asked the study participants to wear a wearable camera all day during the study and collected their location and activity traces. We also asked them to self-report in daily diary regarding challenges they encountered in collecting their travel activity data and to reflect on the their collected activity recordings. After they completed their participation, we conducted an interview with each participant to understand their overall experiences, strategies, and preferences with respect to each approach. Finally, we also collected logs of participants interacting with the instrument to capture their actual interaction with the instrument to perform each approach to compensate their self-reported behaviors from the interviews.

We conducted two phases of analysis. In Phase One, we compared the quantity and quality of the resulting data among the three approaches and participants' experience in using each approach. Our results provide two highlights. First, in our study, although Context-Triggered approaches produced a larger number of recordings, they did not lead to a larger quantity of data than the Participatory approach. It was because many of these recordings were not labeled, were incomplete, and/or were fragmented due to the imperfect context detection. In addition, the data collected using the Participatory approach were more complete and contained less noise. Second, while participants appreciated automated recording and reminders for convenience, they highly valued having the control over what and when to record and annotate. As a result, we conclude that user burden and user control are two important aspects a future tool in mobile crowdsourcing should take into consideration.

In Phase Two, we analyzed participants recording and annotation behavior using the interaction logs we collected on our instrument. The logs showed how and when participants' used the instrument to record and annotate their activity using each of the approaches in the field. Through analyzing the logs, we were able to also examine how the specific nature of the activities being captured affected their recording and annotation behavioral pattern respectively. Analyzing participants' behaviors, as Dumais et al. (2014) has suggested, enabled us to obtain a more complete and accurate picture of the participants' behavioral patterns that they would have not been able to remember and articulate accurately. In addition, we analyzed the characteristics of participants' annotations to understand whether annotations would differ according to the type of activity being collected. We also analyzed participants' diary entries to understand what contributed to unlabeled, mislabeled, and erroneous activity recordings. To summarize, we found that the type of activity being captured influenced recording and annotation timing, participants' receptivity, and characteristics of annotations. Moreover, these factors were impacted by the nature of transitions between activities, the attentional requirements of each activity, and the context of the activity. Based on the findings, we provide a set of design and methodological suggestions regarding the approach, tool, and instruction for collecting activity data via mobile crowdsourcing.

Note that the results of the first phase analysis have been reported in Chang et al. (2015). However, we chose to also present the results again in this paper because presenting both provides a more complete picture of the phenomenon of collecting-annotated-activity-data-through-mobile-crowdsourcing. And because of this complete picture, we are able to discuss the phenomenon as a whole and derive holistic design and methodological suggestions that consider findings from both Phases of analysis instead of just either one (the latter case would take readers additional effort to integrate the two standalone suggestions). Therefore, while new contributions are mainly from the Phase Two analysis, we include the Phase One analysis to help readers better understand the context of the discussion and design and methodological suggestions, since we believe they are more important takeaways for people who need concrete and practical suggestions on leveraging mobile crowdsourcing to collect activity data. To distinguish the present paper from Chang et al. (2015), below we list the novel contributions and new findings of this paper:

- We show that activity type affects users' recording and annotation timing, which in turn affects the quality of the data.
- We show that activity type affects users' receptivity to annotation requests for Context-Triggered ESM.
- We show that activity type affects both the length and content of notes in annotations.
- We present reasons for unrecorded, unlabeled, and erroneous activity data

- We provide practical suggestions for future data collection tools and methods to collect better-quality annotated activity data in the wild.

The remainder of the paper is organized as follows: We discuss related work in Section 2. We present the field study and explain our research methods in Section 3. We describe our general data processing and coding process in Section 4. We describe the analysis and present and discuss the findings of Phase One and Phase Two in Sections 5 and 6, respectively. Then in Section 7 we provide a general discussion, including the design and methodological implications and the study limitations. Finally, we conclude the paper in Section 9.

2. Literature background

2.1. Leveraging the Mobile Crowd to Collect Data

Leveraging a crowd of workers to perform tasks in the mobile environment has been gaining attention in recent years because of the wide availability of smartphones and mobile Internet. Since most modern smartphones are equipped with various sensors, many researchers have attempted to develop applications and platforms to collect sensor data from smartphone users, a method known as mobile crowdsensing (Ganti et al., 2011; Khan et al., 2013; Lane et al., 2010), and citizen science (Silvertown, 2009). Participatory Sensing (Kanhere, 2011), in particular, is a well known and widely used approach to collecting sensor data in the wild in mobile crowdsensing (Ganti et al., 2011; Khan et al., 2013; Lane et al., 2010). The idea of Participatory Sensing is that participants initiate data collection with guidelines provided by task requesters (usually researchers) and use an instrument to capture data of interest for data requesters. Because researchers need to rely on participants to cooperate and to provide good quality data, much of prior research in Participatory Sensing focused on supporting participants, including protecting participants' privacy (De Cristofaro and Soriente, 2011; Ganti et al., 2008; Sakamura et al., 2014), reducing participants' effort by requesting data only from those who are in relevant locations (Linnap and Rice, 2014) or are moving to the target area (Konomi and Sasao, 2015), and improving the data quality (Huang et al., 2010; Reddy et al., 2007; Sheppard et al., 2014).

Mobile and situated crowdsourcing, an emerging area that aims to overcome the limitation of online crowdsourcing on performing tasks beyond the desktop, is not limited to collecting sensor data. For example, Goncalves et al. (2014) used public displays as a crowdsourcing platform to gather keywords to describe locations; Heimerl et al. (2012) used a vending machine for performing locally relevant tasks; Agapie et al. (2015) involved local workers to report local events. Hosio et al. (2014), on the other hand, used a kiosk to offer a variety of crowd tasks, including typical crowdsourcing tasks such as identifying and annotating objects in images (Nowak and Rüger, 2010) and videos (Vondrick et al., 2012). However, the fact that the kiosk is deployed in the field allowed the workers to perform field tasks such as describing the environment.

However, most, if not all, of these mobile crowdsensing and crowdsourcing applications primarily focus on performing tasks wherein the work can be assessed and validated by other peer workers and experts. For example, tasks such as sensing public phenomena and reporting locally relevant information are relatively easy to be verified with multiple workers by assigning them the same task. However, when it comes to collecting individuals' personal contextual and activity data, it is much more challenging to assign peer workers to verify the data collected by a worker. After all, it would be infeasible to assign peer workers to follow and observe data collectors recording his or her daily activities. Perhaps because of this challenge, there has been a lack of study evaluating the effectiveness of mobile crowdsourcing for collecting individual activity data. However, we argue that this gap needs to be filled due to an increasing need for collecting annotated contextual and activity data in real word settings.

In our own study, we assessed the collected travel activity data by asking participants to wear a wearable camera to capture their outdoor travel activities. Then, we used passively logged location and activity traces on their smartphones along with the photos captured by the wearable camera to reconstruct their "ground truth" travel activity histories during the study. Although combining the three sources was laborious, it enhanced the validity and the reliability of the travel activity histories, enabling us to use them as a gold standard for evaluating participants' collected annotated travel activity data using each data collection approach.

2.2. Acquiring annotations on recorded activity data

Researchers in context and activity recognition routinely collect labeled contextual and activity data for building training, and testing their systems. While it is impossible to conduct a comprehensive review of this line of research, we rather focus on the research that particularly aims at supporting acquiring annotations. One focus of obtaining annotations is to leverage video to help with recognizing collected activities. For example, CRAFT (Nazneen et al., 2012) adopts both *in situ* and *post hoc* approaches to capture behaviors of children in a video. However, in their study, p*ost hoc* annotations were added by experts to validate *in situ* annotations added by parents. The annotators were not the people who performed the activities. In addition, the study was not aimed at comparing performances of different approaches in the field.

Another topic relating to annotation acquisition is reducing the effort required to provide annotations. One approach is asking users to speak rather than type to annotate. (Harada et al., 2008; Lane et al., 2011) Another is using a Context-Triggered ESM prompt to ask users to label activities (Cleland et al., 2013, 2014). Cleland et al. (2014) compared the accuracy of labels using this approach with using both structured and semi-structured approaches where researchers annotate the activities. They found that the accuracy of labels obtained using the Context-Triggered *in situ* approach was similar to the structured approach. However, in this study only the Context-Triggered ESM approach was not conducted in a controlled setting. In addition, the authors neither analyzed the quantity and the quality of *recordings* nor analyzed users' experience and behavior in using the approaches. To the best of our knowledge, our study is the first study providing a systematic analysis of different approaches to collecting annotated activity data and investigating participants' behavioral of collecting activity data in the field. We also further provide thorough suggestions on the approach, tool and instruction for using mobile crowdsourcing to collect activity data.

2.3. Validity assessment of research methods

Another research area related to this study is assessing the validity of approaches to collect behavioral data. In this line of research, methods often being assessed are usually retrospective methods such as surveys (Sonnenberg et al., 2012) and interviews (Klumb and Baltes, 1999) because they are generally believed to be subject to recall errors. To validate data collected via these methods, researchers have used ESM or Ecological Momentary Assessment (EMA) as a gold standard to compare with retrospective methods because ESM and EMA are considered to accurately reflect participants' *in situ* and momentary experiences and behaviors. In addition, the daily construction method (DRM) (Kahneman et al., 2004), an approach proposed for allowing participants to reconstruct the sequence of activities that occur during a day, has also been assessed using ESM/EMA (Dockray et al., 2010; Kim et al., 2013). However, data collection for context-aware systems development introduces new concerns that go beyond validity as compared to a gold standard, for example, the quantity and the temporal alignment of collected activity data compared to the actual activity.

2.4. Mobile Receptivity and Interruptibility

Finally, finding opportune moments to request users to perform data collection is critical for maximizing users' data contribution. This topic has received attention from a number of researchers, including those employing an ESM approach for issuing requests to obtain data for developing machine learning models (Turner et al., 2015). When using an ESM to prompt users to respond to annotation task (e.g. a questionnaire), one question is: *how receptive are users to an annotation task on mobile phones*? Research on receptivity has focused on developing models for predicting users' interuptibility (Rosenthal et al., 2011), attentiveness to communication (Dingler and Pielot, 2015; Pielot et al., 2014b), availability for calls (Pielot, 2014) and boredom (Pielot et al., 2015).

On the other hand, recent research has also investigated users' attentiveness to mobile notifications. Overall, the literature suggests that mobile users are quite attentive to mobile notifications. For example, Sahami Shirazi et al. (2014) suggests that mobile users valued notifications related to people and events more highly than otherwise. Both Pielot, et al. (2014a) and Chang and Tang (2015) found that mobile users attend to notifications typically within several minutes; Chang and Tang (2015) further suggested that mobile users are more likely to attend to messages within a minute when their phone is not silent than when their phone is silent. In addition, Dingler and Pielot (2015) found that mobile users were attentive to messages 12.1 h a day, and they would return to their attentive state within 5 min after inattentiveness.

Recent research also explores opportune moments to deliver notifications on mobile phones. For example, Fischer et al. (2011) suggested that at the endings of making calls and receiving SMS indicated breakpoints on the use mobile phones. Poppinga et al. (2014) developed a model for predicting opportune moments to deliver notifications. They suggested that phone position, time in a day, and location were good indicators of opportune moments. Pejovic and Musolesi (2014) explored opportune moments for delivering questionnaires and suggested that good indicators of opportune moments included physical activity, location, time of day, and engagement. Sarker et al. (2014) found that location, emotion, physical activity, time of day, and day of the week played an important role in predicting availability for answering an ESM questionnaire. Finally, Okoshi et al (2015) developed Attelia || and showed that adding physical activity-based breakpoint detection to UI event-based breakpoint detection could result in significant reduction of users' perception of workload of dealing with notifications. This result suggests a potentially effective approach to increase users' receptivity to annotation tasks requested from researchers.

However, while these research works suggested that mobile users are attentive to mobile notifications and indicated several features indicative of users' receptivity to messages and questionnaires, none of these research was addressing mobile users' receptivity to *data collection tasks*, especially when the task involves users for annotating the activity. As Turner et al., (2015) point out, most works in this line of research has focused on particular scenarios, making the applicability of the features predictive to people's receptivity to other scenarios uncertain. The scenario studied here—collecting and annotating activity data on the go—has not been studied previously despite its increasing importance. Because of such a gap, we include receptivity analysis in our study, and our results suggest that, in the context of collecting and annotating travel activity data, mobile users' receptivity was significantly lower when the users were in an activity requiring their high attention (e.g. driving) than in an activity requiring their low attention (riding as a passenger). We believe that these findings are important to using ESM for delivering mobile crowdsourcing tasks, especially to mobile users who are on the go.

In Section 3 below, we present our field study investigating the mobile crowd using three different data collection approaches to collect activity data in the field.

3. The field study

3.1. Collecting travel activity

We chose *travel activity* as the target activity to record and annotate. We had considered other types of contextual/activity data collected in prior research, including home activity, phone placement, noise, and body motions. We set up a list of criteria to evaluate each choice, including: 1) the data collection task is challenging enough but not too difficult so that users' performances could be distinguished; 2) the task could be performed for several days, so that there is diversity within the to-be-recorded activity; 3) a known method exists for approximately detecting the to-be-recorded activity with a reasonable accuracy so that we could use it for implementing Context-Triggered approaches and 4) the occurrence of the to-be-recorded activity should be frequent enough so that failing to detect an instance of it will not lead to significant user frustration and a delay of the study. After evaluating each alternative, we chose to collect travel activity: *participants recording and annotating their travel activity when they are traveling outdoors,* as shown in Fig. 1.

3.2. Choices of approach to compare: PART, SITU, POST

We chose to compare three approaches to collect travel activity data: *Participatory Sensing (PART), Context-Triggered In Situ (SITU), and Context-Triggered Post Hoc (POST)*. We chose these three approaches for several reasons. First, PART and POST are commonly adopted and discussed techniques in mobile crowdsensing (Ganti et al., 2011; Khan et al., 2013; Lane et al., 2010). SITU implements a Context-Triggered ESM approach, which is commonly used for collecting contextual and behavioral data (e.g. Froehlich et al., 2007). Second, PART, POST, and SITU impose different kinds and levels of effort on users, namely, 1) the effort of operating the system to record and to annotate data; 2) the effort of remembering to start and stop recording data, 3) the effort of responding to a prompt in time and then returning to the original task if the current task is interrupted, and 4) the effort of recalling and reconstructing what happened during the recorded activity. We assume the differences in these aspects would influence user burden and compliance, and the quality of the recorded data. Finally, all PART, SITU, and POST have been used in collecting travel activity data with users' inputs (Auld et al., 2009; Froehlich et al., 2009; Reddy et al., 2010). Later we will describe the implementation of the three approaches in our study.

3.3. Instrument for data collection: Minuku

For this study we used *Minuku* to collect data. Minuku is an Android data collection tool developed in our lab and is supported by a backend for data storage. While the study was conducted, Minuku mainly supported between Android 4.0 and 4.4. It could passively record contextual data (e.g. location, activity), trigger actions such as delivering questionnaires based on the context, and schedule daily diary prompts at designated times. These features were necessary for SITU and POST: Minuku needed to automatically initiate recording data when it detected a user likely traveling using a particular transportation mode. Furthermore, In SITU, Minuku needed to additionally prompt users to annotate their trips about their travel

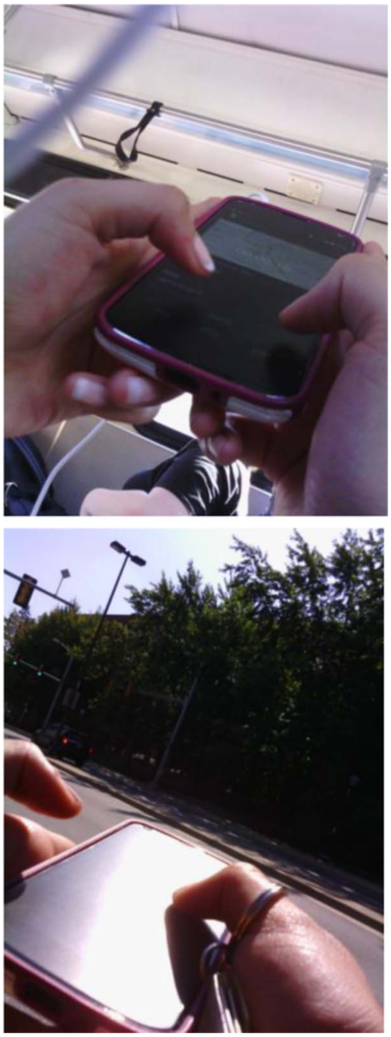


Fig. 1. Study participants recorded and annotated their trips when they traveled outdoors.

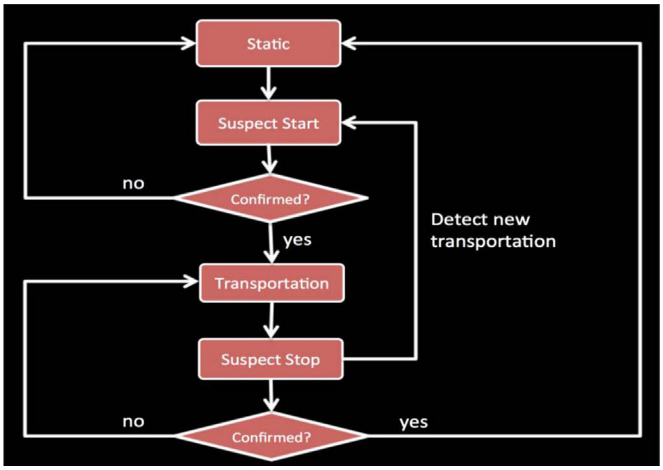


Fig. 2. We built a finite state machine (FSM) to process activity labels received from the Google Activity Recognition service and to determine whether a user is in a certain transportation or is stationary. The FSM has four states: Static, Suspected Start, Transportation, and Suspected Stop.

activity.

3.3.1. Transportation detection

Minuku utilizes Google Activity Recognition service[1] to generate activity logs and uses the log to generate a first approximation of users' transportation mode. Specifically, Minuku extracts the *in vehicle*, *on foot*, *on bicycle*, and *still* labels from the service, and ignores the *tilting* and *unknown* labels because they are not informative to the users' movement (these also only appear sporadically among the dense log of transportation labels). Each label is accompanied by a confidence value, indicating how confident the service thinks the user, or the phone, is performing the activity. The service may include one or more activity labels, of which the all confidence values add up to 100. For example, *{in_vehicle:100}* shows that the service is 100% confident that the user is in a vehicle; *{in_vehicle:77, on_bicycle:23}* shows that the user is most probably in a vehicle, with a small likelihood of biking. Considering that these labels might be subject to positioning errors and noises, and possibly also affected by some unexpected or errant traffic conditions, we built an additional layer, a finite state machine (FSM), to process received activity labels and to determine whether a user is in a certain transportation or is stationary. At a high level, the FSM considers both the current and the previous activity labels received within certain period of time to determine the user's current transportation mode. The purpose of "looking back" at previous labels is to raise the threshold for transitioning the user from one transportation state to another, so that the transition would be more resistant to noisy labels. We give more information about the implementation below.

As shown in Fig. 2 above, initially, a user is in a *Static* state, indicating the user is not traveling. When Minuku receives an activity label indicating a movement (e.g. on foot, in a vehicle, or on a bicycle), the FSM transitions the user to a *Suspect Start* state, signifying that Minuku suspects that the user is traveling in that transportation. In this state, the FSM examines previous labels within a period of time such as 20 s (we call it a *window time*). If the frequency or percentage of transportation mode labels is higher than a predefined threshold, the FSM transitions the user to a *Transportation* state. In this state, the FSM starts to look for labels indicating non-movement (e.g. still) or a different transportation mode. It enters a *Suspect Stop* state if it receives any such a label. The FSM returns to the *Static* state if over another window of time, lower than a percentage of previous labels are the current transportation mode, i.e. the user is less likely to be moving in the current TM anymore. However, if the FSM continuously receives the same label of a new different transportation mode and the percentage passes the threshold for detecting a start of that transportation mode, it skips the *Static* state and directly enters the *Suspect Start* state for that transportation mode (e.g. on foot - > in vehicle instead of on foot - > still - > in vehicle). As a result, in this FSM, there are four key parameters that control the transition between the states: *a window time* within which the FSM checks previous activity labels for starting and stopping a transportation mode, respectively; the thresholds for confirming a start and a stop of a transportation, respectively. All of these thresholds were arbitrarily set initially, but tested and modified iteratively in our pilot testing and study.

The pilot testing and study were important to the field experiment because while a low threshold would cause Minuku to repeatedly prompt users during the same travel activity (over-segmentation), a high threshold would impose a significant delay before Minuku detected a start of a travel activity. It is noteworthy our FSM adopted a fairly simple heuristic for determining transportation. But we realized that such a detection would never be perfect (which is true for any current activity recognizer) due to the variety of transportation situations that individuals would encounter in their real lives. In

[1] https://developer.android.com/reference/com/google/android/gms/location/ActivityRecognitionApi.html.

addition, we assume that the audience of the paper—researchers who have a desire to collect real-life and diverse activity data via the mobile crowd to develop a new recognizer or to make their present recognizer more accurate and reliable—may not yet have had a reliable recognizer for performing a Context-Triggered approach. Instead, they may have a basic recognizer simply capable enough for performing a Context-Triggered approach like our FSM did. Therefore, we decided not to pursue an optimized transportation recognizer for our study and stopped testing until the transportation detection was robust and accurate enough in our own pilot testing and study. However, it is our belief that our transportation recognizer is reliable enough for the study purpose because our FSM is built on top of the Google Activity Recognition service, of which the reliability and accuracy can be assumed to be tested and verified.

The final values of parameters used were as follows: For detecting a start of any transportation, we used 60% as the percentage threshold and used 20 s as the window time. For detecting a stop of a transportation, we used 20% as the percentage threshold. And we used 60 s as a window time for *on foot*; 90 s for *on bicycle*; 150 s for *in a vehicle*. We start to describe our study design and procedure below.

3.4. Study design and procedure

We adopted a within-subject design for this study, i.e., each participant collected data using each method: PART, SITU, and POST. We chose this design because we anticipated that people would have a varied number of travel activities in a day and different commute routines. To mitigate the order effect, we randomly assigned participants to one of the six possible orderings of the three approaches. The number of participants in each order was balanced.

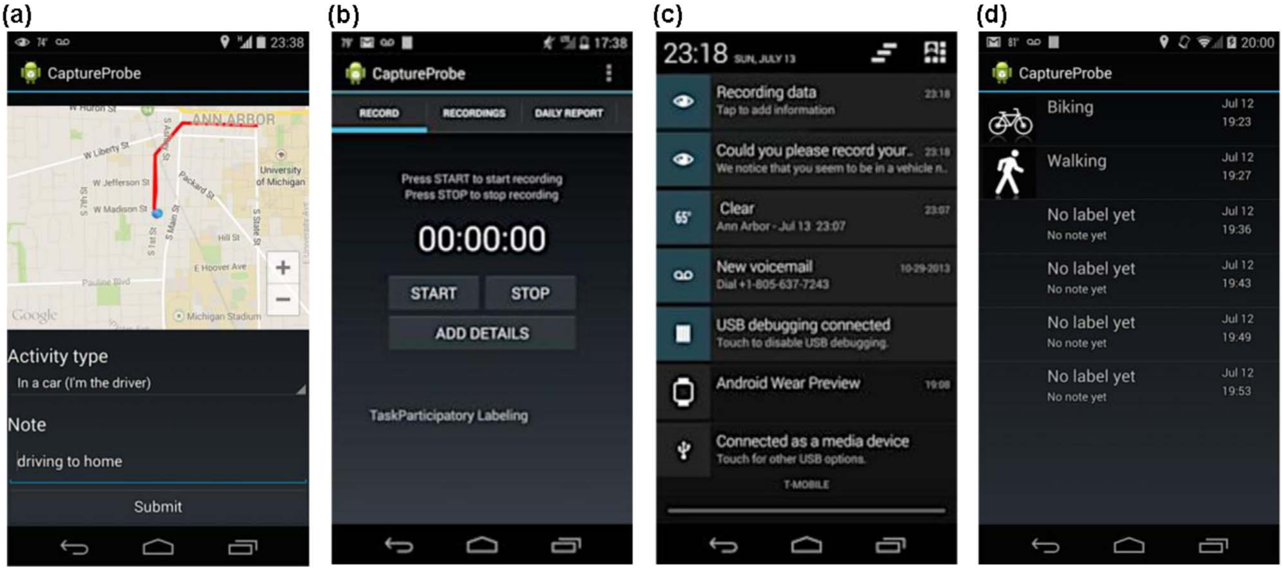
3.4.1. Collecting travel activities using Minuku when traveling outdoors

We asked participants to record and annotate their *trips* when they were traveling outdoors (i.e., between locations). We told the participants (also put in the study instruction) that a trip is a journey with an origin, a destination, and a certain transportation mode at least 3 min long. This instruction is particularly put for the PART approach because participants could only control recording in the PART condition. However, we also told them that the definition of a trip was flexible, and the main purpose was to avoid them recording very short travel activity. The annotation interface was same for all three conditions and is shown in Fig. 3a. Participants were asked to choose an activity type (i.e. a transportation mode label) as a *label*, and to add a *note* to describe their trip. Both labels and notes are considered a type of annotation in the study. We considered a recording being annotated if it was either labeled or added a note. Specifically, we told participants, "The note field is optional. However, it would be great if you could let us know what the trip is about, especially when the trip is atypical, such as you are stuck in a traffic jam." We consider that a recorded trip is annotated if it is either labeled or is added a note. One intent of this instruction was to *encourage,* rather than *require* the participants to describe their trips to reduce their burden. Additionally, participants were given the freedom and flexibility in typing a note so that we could explore the kinds of information participants thought would be relevant to their travel activities. Furthermore, we also told participants that when a trip was being recorded, an ongoing notification icon would reside in the notification bar of the phone, and they could access the annotation interface by choosing that notification, as long as they saw that notification.

Participants were asked to record and annotate at least two trips per day. At the end of each day, we tracked the number of recordings that participants annotated, and transitioned them to the next study condition once they had aggregated four days of annotated trips in the current condition. When the transition occurred, we sent them a new version of Minuku customized for the subsequent condition. We told them that the four days of recordings did not need to be consecutive, and they should travel as they would normally do. We provided them with $24 for completing the three conditions. Participants were also rewarded 25 cents for recording each extra trip beyond the two required daily trips, and they could earn up to $10 for the extra trips.

3.4.2. Performing PART, SITU, AND POST

For the PART condition, we told participants that they manually started and stopped recording their trips using the interface shown in Fig. 3b. They were instructed: *"Hit the Start button when you start your trip; hit the Stop button when you end your trip."* They could also pause and resume a recording. Clicking the "Add Details" button brought them to the annotation interface. We told them that they could modify labels and notes for their trips in the Recording Tab, in which they could also see all recordings. In addition, we instructed them how to handle transitions between trips with examples and told them not to intentionally split a trip in the same transportation mode



Activity type
In a car (I'm the driver)
Note
driving to home
Submit

Press START to start recording
Press STOP to stop recording
00:00:00
TaskParticipatory Labeling

23:18 SUN, JULY 13
65°
Recording data 23:18
Tap to add information
Could you please record your... 23:18
We notice that you seem to be in a vehicle n...
Clear 23:07
Ann Arbor - Jul 13 23:07
New voicemail 10-29-2013
Dial +1-805-637-7243
USB debugging connected
Touch to disable USB debugging.
Android Wear Preview 19:08
Connected as a media device
Touch for other USB options.
T-MOBILE

Biking	Jul 12 19:23
Walking	Jul 12 19:27
No label yet No note yet	Jul 12 19:36
No label yet No note yet	Jul 12 19:43
No label yet No note yet	Jul 12 19:49
No label yet No note yet	Jul 12 19:53

Fig. 3. From left to right are: (a) The interface for labeling and adding notes, (b) PART: users manually record their trips, (c) SITU: prompting users to annotate their trips, and (d) POST: users reviewing and annotating trips afterwards.

into multiple recordings. We also clarified that whenever they switched to a different transportation mode (e.g., walking after parking a car), they were starting a new trip, since they needed to choose a difference label.

In the SITU condition, we told participants that Minuku automatically detected their travel activities and would prompt them a phone notification to annotate their current trip as soon as a trip using a new transportation mode was detected (as shown in Fig. 3c). We told them that choosing that notification took them to the same annotation interface and that a notification was automatically dismissed when they were detected as having ended the current trip. We emphasized that they could only annotate during the trip because there was no recording tab in this condition, but they should annotate while they were in a safe situation (e.g. not while driving).

For the POST condition, we told participants that Minuku automatically detected their travel activities but would not prompt them to annotate during a trip. Instead, any trips they completed would appear in the Recordings Tab (as shown in Fig. 3d), and Minuku would remind them every day at 9 pm to annotate. This approach is similar to a daily diary study and the day reconstruction method (DRM) used for reflecting on life experience (Kahneman et al., 2004). The method has also been termed *prompted recall survey* in transportation research (Auld et al., 2009). We told participants that they could annotate their trips in the Recordings Tab at any time.

3.4.3. Collecting “Ground truth” data

To assess the quantity and the quality of participants' recordings, it is necessary to know when they started and stopped traveling outdoors. Therefore, we used Minuku to passively log participants' location and activity traces. While activity traces were passively logged at all times, location traces were logged only when participants were detected to be moving (i.e. not stationary) so that we could minimize the power consumption of Munuku. However, because location and activity traces are not always reliable and accurate, we asked participants to wear a wearable camera called Narrative Clip[2] during the study period. The camera is “always on” and takes a photo every 30 s. It is intended to be attached to the front of one's clothing, and to capture whatever the wearer is looking at. Wearable cameras have previously been used to validate travel diaries in transportation research (Doherty et al., 2013; Kelly et al., 2014). Inspired by this line of research, we intended to combine photos and logs to cross-validate and to generate *Ground Truth Trips* for each participant during the study.

We had considered recording continuous video, however, during the study, there was no wearable camera that could continuously record a video for an entire day or take still photos at a rate higher than 2 Hz. We asked participants to wear the camera at all times if possible and emphasized to them that it was important for the study that they wore it whenever they started to move. However, for ethical reasons, we told them that they could take off the camera if they were uncomfortable with wearing it in particular settings such as in a private meeting or in a bathroom. We told participants that photos were important for the analysis, but we did not tell them that photos were used for reconstructing the ground truth of their travel activities. In addition, Minuku logged participants' any actions related to recording and annotation and the times when Context-Triggered annotation prompts were generated.

3.4.4. Daily diary and post-study interview

In all three conditions, we sent participants a *diary prompt* e-mail at 9:30 pm daily to have them reflect on recordings. The diary prompt contained a list of recordings captured that day, with the start time, end time, and a transportation mode label next to it. We asked them to review and correct any incorrect recordings. For any unlabeled recording, we asked them to choose a reason from a list of reasons why the recording was unlabeled and also provide context about the recording. We also asked them to list trips that they took but did not appear in the recording list, and to choose a reason for why the trip did not appear. We interviewed each participant after they completed all three conditions. We first asked them about their commute process in a typical day and how they decided which trips to record. Subsequent questions were focused on, for each approach, how they annotated, the challenges they encountered, their subjective preferences, and their suggested improvements.

3.5. Participants

We recruited participants that regularly commuted to work or school by posting flyers on campus, sending department-wide e-mails, and advertising on social media. Respondents completed a screening survey to provide their 1) commute behaviors, 2) experience in using an Android phone; and 3) anticipated out-of-town travel plans in the near future. We filtered out participants who traveled fewer than 4–5 days in a week, whose typical commute time was less than 5 min, and who were planning to travel out of town for more than a couple of days during the study timeframe. We attempted to balance gender, age, and primary commute transportation mode among participants. While we started the study with 37 participants, only 29 completed participation (16 males, 13 females). There were several reasons why participants dropped out of the study: the app did not work with their phone, they lost the wearable camera, or they stopped responding. Fourteen participants' ages were 18–25; twelve were 26–35, three two were 36–45, and one was over 55. We refer to them as P1-P29 throughout this paper. P13 and P19′s data were excluded from the quantitative analysis because their data were incomplete. Thirteen participants reported that their primary commute mode was “car,” while ten reported “bus,” four “walk”, and two “bike.”

4. Data processing and coding

4.1. Cleaning, merging, and processing recordings

It is important to distinguish between the terms *trip* and *recording* to understand the performance of the participants using each method, i.e. how much of their travel activity was recorded by them, and how much noise were contained in the data they recorded. As we put in the instructions for the participants aforementioned, we define a trip as an actual journey in which a participant departed from an origin to a destination (e.g. from home to work). A recording, on the other hand, refers to a data representation of a trip generated in Minuku. That is, when Minuku records a trip, either via a context trigger or via manual activation by a participant, it generates a recording of the trip. As a result, it is important to note that a recording does not necessarily perfectly represent an actual trip a participant took. For example, a recording is likely to capture only some portion of a trip if Minuku starts recording before the trip starts, and/or ends recording before the trip ends. Likewise, a recording also likely contains data beyond a trip (referred to below as *noise*) if it starts recording before the trip starts, and/or ends recording after the trip ends. Moreover, it is also possible that multiple recordings are associated with one trip if for any reason Minuku stops and restarts recording during the same trip. We determined whether a recording belonged to a certain trip based on participants' diary entries (many participants explicitly flagged recording as split trips) and based on our observation on the continuity of nearby recordings. We collected in total 3070 recordings generated by Minuku. Among these recordings, we removed duplicate recordings generated due to Minuku's error and the false recordings (e.g. flagged by a participant in the diary, e.g. walking indoors). We also merged split recordings that participants flagged. Through this data cleaning and merging process, we obtained 2587 *valid recordings* (84.3% of all

[2] http://getnarrative.com/.

recordings), including both labeled and non-labeled ones.

4.2. Generating ground truth trips

To evaluate participants' performance, we reconstructed *Ground Truth Trips* taken by the participants from approximately 117,000 captured photos and from activity and location traces passively logged on Minuku. Several participants mentioned in the interview that they did not wear the camera at work or private places. There were also a few diary entries where participants said they forgot to wear the camera during a few trips. Thus, while we asked participants to wear the camera at all times if possible, we could not assert that Ground Truth Trips captured "all" participants' trips during the study. We describe the process of coding Ground Truth Trips as follows:

Two coders independently coded participants' Ground Truth Trip times and transportation modes from photos and trace logs. Specifically, for photos, coders were trained to determine a transportation mode and when a participant started and ended a trip based on the movement changes observed among photos. For trace logs, they were trained to inspect participants' travel activity by using the Google Earth for Desktop[3] to playback location traces to observe participants' movements. For any persistent movement observed in the Google Earth Desktop, the coders also searched within activity trace logs to find a block of activity labels corresponding to the movement. As mentioned earlier, these labels are generated by the Google Activity Recognition service that infers a transportation mode of the movement of the phone at a particular time. From the activity log, the coders also inferred an approximate start time and end time of the movement by considering the continuity, the persistence, and the confidence level of that transportation mode determined by the service. From these two sources, the coders decided a start and an end time of all travel activities observable from the logs. Therefore, as a summary of the reconstruction, we used photos as a primary source to determine when a travel activity occurred, and we used trace logs to determine a more precise start and end time. Considering both sources is crucial to ensure the accuracy of Ground Truth Trips because of the temporal precision required. Although trace logs provide ambiguous information regarding transportation mode, it offers better temporal precision than photos do. While the wearable camera used a 30-s shooting rate, Minuku sampled location traces every two seconds.

Taking both photos and logs into account, the coders finalized a transportation mode, a start time, and an end time for each Ground Truth Trip. This process took each coder more than 200 h to complete. To ensure consistency between the coders, we developed a standardized coding protocol. In addition, the first author met with the coders weekly to discuss and to resolve any uncertainty on coded times. We randomly chose a subset (644) from the coder's' coded times and ran the intra-class coefficient (ICC) test between them. The ICC score was .87, indicating high reliability between the two coders. After the test, each coder coded a subset of the rest of the photos and trace logs (randomly assigned). The two coders generated 1,414 Ground Truth Trips, and paired each of them with participants' recordings by comparing their start time, end time, and transportation mode. As mentioned earlier, a number of Ground Truth Trips were paired with multiple recordings. During the pairing process, we considered a Ground Truth Trip being *collected* if, and only if, it has at least one paired recoding with a collect label (i.e. the label of the recording matched the coded transportation mode of the Ground Truth Trip); *unlabeled recordings* (recordings without an attached label) and *mislabeled recordings* (recordings incorrectly labeled) were not paired to any Ground Truth Trip.

4.3. Analyzing data in two phases

Because of the number and the variety of data sources, we conducted two phases of data analysis. The first phase of analysis (referred to as Phase One) was primarily focused on the comparison among the annotation approaches, including comparing the quantity and quality of the resulting data collected through each approach, as well as users' preferences of and experiences in using each approach.

The second phase of analysis (referred to as Phase Two) was focused on the participants' recording and annotation behavior when using PART and SITU in the field. The analysis includes a behavioral log analysis, a content analysis of participants' annotations, and qualitative analysis on participants' diary data. Additionally, we revisited interview data with a new theme focused on participants' overall strategies and behaviors in using each approach.

In Section 5, we first present the analysis, results, and discussion in Phase One. Then in Section 6, we follow the same structure to report the new findings and offer new insights into users' behaviors of recording and annotation in the field obtained in Phase Two.

5. Phase one: comparing the annotation approaches

In Phase One, we analyzed the quantity and the quality of the recordings obtained in each condition by comparing them to the Ground Truth Trips we reconstructed. For quantity, we measured the coverage of recordings. For quality, we measured the completeness and the precision of the recordings. In addition, we computed overall performance measures such as *the number of recordings*, *recording labeling ratio*, and *recording annotating ratio*. We give more details in the section below.

5.1. Measures in quantitative analysis

5.1.1. Overall performance measures

We computed measures indicating participants' overall performance in producing recordings using each of the three annotation approaches. The measures are:

1. *Number of valid recordings.*
2. *Recording labeling ratio*: The ratio of valid labeled recordings to total valid recordings.
3. *Recording annotating ratio*: The ratio of annotated valid recordings to total valid recordings.

5.1.2. Coverage and trip labeling ratio

Coverage of recordings measures the length of data being recorded and *correctly labeled* in absolute time (seconds) and percentage of total time (percentage) per day. For example, if a participant traveled 70 min in a day and recorded 56 min, the coverage length is 56 min, and the percentage is 80%. Thus, the higher these two measures are for a particular approach, the greater quantity of data participants collected through that approach. Another measure we calculated was the *trip-labeling ratio (T-LR)* per day. This measure indicates the ratio of participants' actual trips being recorded and correctly labeled to total trips per day. For example, if the number of Ground Truth Trips for a participant on a certain day is 8 but the participant only recorded and correctly labeled 4 of them, the T-LR for that day is 50%. For this measure, we hypothesized that the T-LR in PART is lower than the T-LR in SITU and POST, because, in PART, participants had to initiate recording on their own, whereas in both SITU and POST Minuku recorded a trip whenever it detected movement in a targeted transportation mode.

5.1.3. Completeness

Completeness measures the percentage of a trip being recorded and labeled (Fig. 4, top). For example, if 15 min out of a 20-min trip is

[3] http://www.google.com/earth/explore/products/desktop.html.

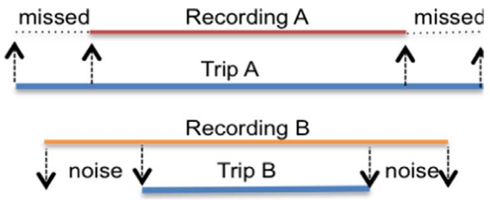


Fig. 4. (Top) Completeness of recordings measures the percentage of a trip being recorded and labeled. (Bottom) Precision of recordings measures the percentage of a recording that precisely reflects its label.

recorded and labeled, the completeness of the recording is 75%. Two other related measures are: 1) length of missed portions at the beginning and 2) length of missed portions at the end of a trip (seconds). For example, if a recording starts ten seconds *after* a trip starts, it misses ten seconds at the beginning; if it ends ten seconds *before* a trip ends, it misses ten seconds at the end. For this measure, we expect to see some missed beginning in the recordings in SITU and POST because in these two conditions Minuku needs to detect movement of participants, which thus was likely to delay recording.

5.1.4. Precision

Precision measures the percentage of a recording that precisely reflects its label (Fig. 4, bottom). For example, if a recording being labeled as "driving" starts one minute earlier than a 9-min trip, it contains one minute of noise at the beginning, and its precision is 90%. Similarly, if the recording ends one minute later than the end of the trip, the recording contains one minute of noise at the end. Also due to the detection delay, we expect to see noise at the end of recordings in SITU and POST.

5.2. Methods of data analysis

We used a Chi-Square Test to examine whether participants had significant differences in overall performance in producing recordings across three approaches. For measures related to coverage, completeness, and precision, we examined the main effect of variables of interest, including *condition, transportation mode, the day of a week,* and *user* using an analysis of variance (ANOVA). The *user* variable was included to account for individual differences. We included the *periods of day* variable for trip level analysis such as completeness and precision. The periods we used are: morning (6–11 am), noon (11 am–2 pm), afternoon (2–6 pm), evening (6–9 pm), night (9 pm–1 am), and midnight (1–6 am). These periods were determined based on our knowledge of participants' typical daily travel patterns obtained from the interviews. We also included the interaction effect between condition and transportation mode to examine whether certain combinations between the two would have an impact on coverage and precision. We used the Tukey HSD Test for post-ANOVA pairwise comparisons.

We also conducted qualitative analysis on the interview and diary data. Specifically, we transcribed interviews and coded the transcriptions and daily diary entries using an iterative process of generating, refining, and probing emergent themes. In this data analysis phase, the coding themes were focused on the topics of participants' likes and dislikes about each approach and their preferences and challenges of using the approaches.

5.3. Results: quantity and quality of activity data

5.3.1. Overall performance

We first present the results of overall performance. Among the 2587 valid recordings, 1919 (74.2%) were labeled (i.e., were assigned a transportation mode), and 994 (38.4%) were annotated (i.e. contained a free-text note). As expected, the number of labeled recordings in PART (424) is noticeably lower than of SITU (723) and POST (772). In terms of the ratio of labeled recordings to total recordings, from highest to lowest are: PART (91.6%), POST (76.8%), and SITU (64.9%), and all of the differences between any two approaches are statistically significant using the Chi-Square Test for pairwise comparisons (PART vs. SITU: χ^2=109.9, p < .001; SITU vs. POST: χ^2=33.4, p < .001; PART vs. POST: χ^2=40, p < .001). This suggests that participants labeled less percentage of recordings using the Context-Triggered approaches (i.e. SITU and POST) than when they used the PART approach. In addition, PART also had the highest ratio of annotations to recordings (58.2%), which is statistically significantly higher than it of SITU (31.6%, χ^2=25.1, p < .001) and of POST (36.8%, χ^2=28.3, p < .001). No significant difference was found between SITU and POST. This suggests that participants also were mostly likely to attach a note to a recording when they used the PART approach.

There are several things to note regarding these results. First of all, the SITU approach, i.e. asking users to label during activity, led to the lowest ratio of labeled recordings. We think this may be linked to the issue of interruption in SITU. It was also likely that participants missed the prompt often or did not want to label on purpose. Secondly, the ratio of annotated recordings for POST is roughly as low as SITU. We speculate that this is because, in a post hoc review, it might be easier for participants to recall (or reason) the transportation mode of a trip than to recall the detail of a trip, making them less likely to describe those recordings in a note. Third, SITU and POST produced more valid recordings than PART because they employed automated recording. However, we learned in the interviews that participants sometimes were prompted to label a trip more than once in SITU and POST when Minuku falsely detected them stopping and restarting a new trip. Regardless of these reasons, Fig. 5 shows that as the level of user effort increased (i.e. labeling and adding annotations), the advantage of Context-Triggered approaches was diminished with respect to producing a larger number of annotated recordings. The decrease in the rate of adding notes is especially apparent, possibly because we only *encouraged* instead of *required* participants to add a note to the recordings, making this action more dispensable than the other requested actions.

5.3.2. Coverage of recordings

In this section, we show that more labeled recordings, however, does not necessarily lead to a greater quantity of annotated activity data. We compared the ratio of actual trips being labeled to the total number of actual trips *per day* (T-LR) and the coverage of recordings

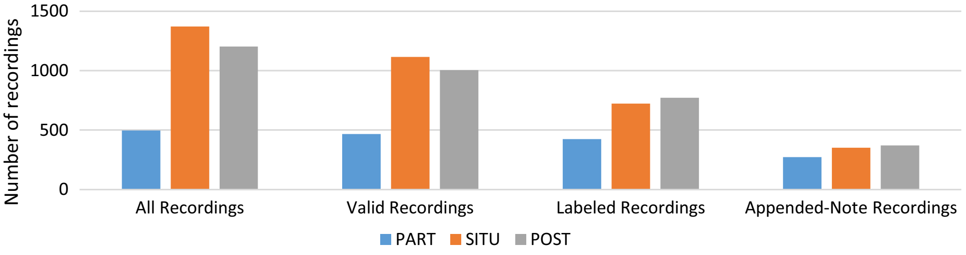


Fig. 5. The differences in the number of recordings decreased as users' effort increased.

among the three approaches. For T-LR, our results indicated a main effect of transportation mode (F[5,454]=5.3, p < .001). In a post hoc analysis, we found the T-LR of walking trips (59.6%) to be lower than it of bus trips (77.9%, p < .001) and car trips (71.6%, p=.02), respectively. We think this may have been because participants considered car and bus trips more like "real trips," and thus may have been more likely to record and label them. However, we feel surprised at not seeing a significant difference in the T-LR among conditions (PART: 70.9%; SITU: 70.7%; POST: 62.6%). This result was unexpected because although PART took more effort for initiating a recording, participants did not label fewer actual trips per day.

For coverage length, our results showed main effects of both condition (F[2,454]=4.9, p=.007) and transportation mode (F[6,454] =18.6, p < .001). In a post hoc analysis, unexpectedly, we found that the total coverage (absolute time) of PART is greater than that of SITU (p=.02) and POST (p=.02), suggesting that participants produced most travel activity data in PART among the three approaches. A similar result was also found in coverage percentage: we found a main effect of condition on coverage percentage (F[2,454] =12.9, p < .001). The overall coverage percentage of PART (67.7%) was greater than that of SITU (52%, p < .001) and of POST (50.3%, p < .001). In addition, we also found a main effect of transportation mode on coverage percentage (F[5,454]=2.8, p=.02). Specifically, the coverage percentage of walking trips (55.3%) is significantly lower than car (74%) and bus (78.5%) trips. This result is consistent with the results of T-LR mentioned above. In other words, in terms of length of time, only approximately half of the walking activity were recorded and labeled by the participants. In contrast, approximately three quarters of the vehicles activity were recorded and labeled.

We found these results interesting and surprising. First, although participants were not more likely to label a larger number of *trips* using any approach in a day, they produced a larger quantity of annotated travel activity data in terms of *length of time* using PART than using SITU and POST. Based on our observation of the characteristics of recordings, we conjecture that this might be because many of the recordings generated in SITU and POST were fragmented, while the recordings generated in PART were more complete and precise (which we will show later). In particular, Fig. 6 shows that the differences among the three conditions seem to more apparent in bus and car data than in walk data. This suggests that PART might be more advantageous for collecting vehicle activities than collecting walk activities. Later we analyzed completeness and precision of recordings.

5.3.3. *Completeness of recordings*

As a reminder, *completeness* refers to the percentage of a Ground Truth Trip being recorded and labeled. For example, if 15 min out of a 20-min trip is recorded and labeled, the completeness of the recording is 75%. Our results showed main effects of both condition (F[2,1365] =35.2, p < .001) and transportation mode (F[5,1365]=8.2, p < .001). A post hoc analysis showed that completeness of recordings in PART (68.2%) was significantly higher than that it of SITU (48.1%, p < .001) and POST (47.4%, p < .001), as shown in Fig. 7 (left). This result supports our hypothesis that recordings in PART were most complete among the three conditions. We also found completeness of recordings for walking trips (45.2%) lower than it of car trips (59.8%, p < .001) and bus trips (59.7%, p < .001), respectively. In particular, this large difference is mainly due to the interaction effect between condition and transportation mode (F[4,1365]=3.8, p=.004). That is, when participants were using PART, completeness of recordings of walking trips (51.8%) was significantly lower than of bus trips (80.5%, p < .001) and car trips (76.2%, p < .001), respectively. This result indicates that when the participants were recording trips by themselves, there was a larger disagreement between participants and our coders regarding when a walking trip started and ended than car trips and bus trips. We think this result may partially explain a previous result that the coverage of walking trip is lowest in the PART condition.

We further looked into what led to incomplete recordings. Regarding missed portion at the beginning of a trip, we found main effects of condition (F[2,901]=31.3, p < .001) and transportation mode (F[4,901]=7.2, p < .001), as well as an interaction effect between the two (F[4,901]=3.9, p=.004). Specifically, the recordings in PART, as shown in Fig. 7 (middle), missed significantly shorter portions at the beginning (29.8 s) than those of SITU (140.4 s, p < .001) and POST (144.1 s, p < .001). This suggests that the delay of recording caused by the transportation detection did lead to longer missed portions at the beginning. In addition, recordings of walking trips missed longer portions at the beginning than of car trips (p < .001). This missed portion may be mainly responsible for the lower overall completeness of recordings of walk trips.

Regarding missed portion at the end, Fig. 7 (right) seems to suggest that recordings in PART missed least portions. However, we did not see any statistically significant difference among the approaches. In fact, the missed portions across all three approaches were limited. This result is not surprising because we expected that Context-Triggered approaches would tend to stop recording after a trip because of the detection delay. On the other hand, this result also implies that when participants used PART and chose to stop recording before a trip ended, they did not stop recording too early. However, it should be noted that this result does not imply that participants stopped recording before the end of the trip. This missed-portion-at-the-end analysis did not consider recordings that ended later than their corresponding actual trips. We investigated participants' recording behavior in Phase Two. In the section below, we present a precision analysis that shows an overall measurement of the noise contained in participants' recordings.

5.3.4. *Precision of recordings*

As a reminder, *precision* measures the percentage of a recording of which the content reflects its transportation mode label. For example, if a 10-min recording labeled as "driving" starts one minute earlier than

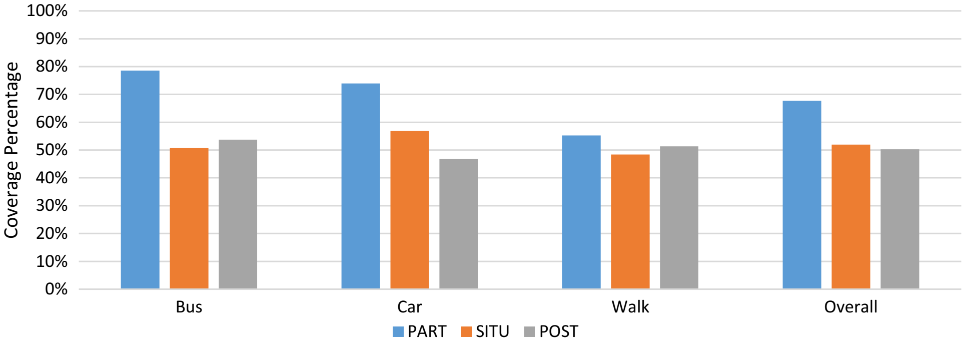


Fig. 6. The coverage percentage by transportation mode and condition. The overall coverage percentage of PART (67.7%) was greater than that of SITU (52%, p < .001) and of POST (50.3%, p < .001). The differences between the approaches are more apparent in vehicle (car and bus) data than in walk data.

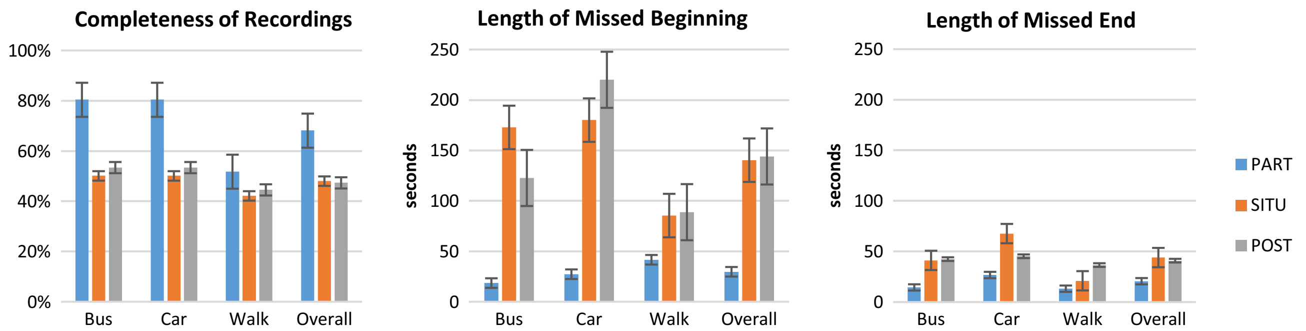


Fig. 7. Completeness of Recordings (Left), length of missed portion at the beginning (Middle), and length of missed portion at the end (Right) across approaches and transportation modes. The error bars show standard error.

the start of an 8-min trip and ends one minute after the trip ends, it contains one minute of noise at the beginning and one minute of noise at the end, respectively. Thus, its precision is 80%, i.e. only the 8 min in the middle precisely reflect the label "driving." Our results showed main effects of condition (F[2,901]=32.1, p < .001) and transportation mode (F[4,901]=16.5, p < .001) on precision. Specifically, we found that the precision of recordings in PART (81.4%) is higher than it in SITU (67.2%, p < .001) and POST (67.4%, p < .001), as shown in Fig. 8 (left). As with the completeness result discussed above, this difference was probably also caused by the detection delay. Furthermore, we found that the precisions of the recordings of walking trips in both SITU (57.8%) and POST (58.1%) were lower than any other combination of transportation mode and condition (all p-values are below .001). With further investigation, we found that such low precisions were mainly because of the noise at the end of trips, as shown in Fig. 8 (right). Specifically, not only that recordings in SITU and POST contained significantly more noise at the end than in PART (both p-values < .001), but also that both recordings of car trips (p=.005) and walking trips (p < .001) contained significantly more noises at the end than bus trips. We think these results may be because the ends of car trips and walking trips were more ambiguous than the end of bus trips from the perspective of our transportation recognizer. That is, the driving and walking patterns among people may be more various than the bus movement patterns. Therefore, while this result may suggest improvement we could make further on our transportation recognizer, it also suggests that researchers may see different performances of a Context-Trigger approach for collecting different activities in real-life settings.

To summarize, our quantitative analysis indicates three results of particular interest. First, although SITU and POST produced a larger number of labeled travel activity recordings, PART produced a greater quantity of labeled travel activity data in terms of length of time. Second, recordings in PART were more complete (less missed data at the beginning) and more precise (less noise at the end) than the recordings in SITU and POST. Third, it seems that walking trips are most ambiguous among all the activity types regarding when a trip started and ended, regardless whether a PART approach or a Context-Triggered approach is used. However, it is important to note that these results did not suggest any tendency of participants' behavior in terms of when they would record a travel activity relative to the timing of the activity. It is because these analyses did not investigate completeness and precision together, but instead, separated the two measures and focused on the contrast among the three approaches. We will dig more into participants' behavior in recording and annotation in the Phase Two analysis.

5.4. Results: qualitative experience in using PART, SITU, and POST

5.4.1. Challenges encountered

According to participants, the greatest challenge of using PART was to remember to record a trip. Most participants reported that they had forgotten to record their trips once or more. Furthermore, many participants reported that it was easier to forget to start recording than forget to stop because once they had started a recording, they were aware that Minuku was recording and would remember to stop it. Some participants also mentioned they took off the camera while they went indoors, and this action reminded them to stop the recording.

The greatest challenge of using SITU was being able to annotate before the prompt disappears during an activity that requires high attention when the activity requires high attention. For instance, whereas most participants said it was not troublesome to annotate while walking, participants who commute by car reported that when driving they had to find a good time to label when getting prompted, usually at stoplights. In order not to miss the prompt, several participants said they tended to wait for the prompt once they started moving, but this gave them pressure and anxiety. For example, P5 said: "it made me so anxious, like 'I've got to record this.'" She continued: "... at first I thought 'oh, [SITU] sounds like the easiest one' but it was actually annoying. [...] there was no way to go back and redo it afterwards, [so the] pressure was like 'I've got to record while I'm doing it or I'll miss it.'".

The most-cited challenge of using POST was being unable to recognize a trip. While sometimes it was because when reviewing a recording on the map the trajectory did not make sense to them, at

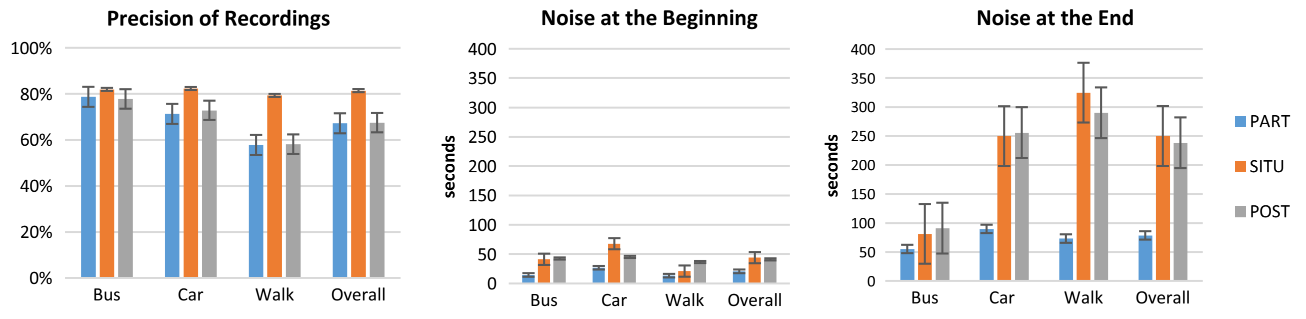


Fig. 8. Precision of Recordings (Left), noise at the beginning (Middle), and noise at the end (Right) across approaches and transportation modes, shown with standard error bars.

other times they said they simply could not recall what that recording was about. For example, P26 said: “[...] I did not recall anything, but it recorded itself. But at the end of the day I had to remember as to what I did at that point, what I did not do at that point.” Interestingly, when reviewing recordings, whereas some participants said that they relied on the map to recognize a trip, others said they mainly relied on the time of a trip. When asked about their rationale, participants who mainly relied on the time indicated that their schedule and travel pattern were regular and predictable; thus time was sufficient for them to recognize their trips. On the other hand, participants who often had irregular travels tended to rely on the map view to recognize their trips. However, participants generally agreed that both maps and time were useful, and noted they had used both for labeling at some point during the study. It is noteworthy that participants often “reasoned” a trip rather than recalling it. For example, P22 reasoned her trips largely based on the time, “I definitely looked at the times a lot because I know I'm walking between 4:30 and 4:45, and then I know I'm driving between 4:45 and 5:00 something, and then if I knew it was an evening trip, I'd remember if I drove or someone else drove.” P24, on the other hand, used trajectories to reason her trips: “[...] like when the line is clearly on the bus route that I take, [it] is very obvious, so that's very reliable, and the same for a car and walking.”

5.4.2. Likes and dislikes

Most participants liked PART because they had complete control over what and when to record. For example, P18 said, “I guess the good part about participatory is that I wouldn't have to respond to three-minute walking trips 'cause those seemed not important.” In addition, they thought the PART approach produced the most accurate recordings among the three. Participants disliked PART mostly because they had to remember to start and stop on their own. For example, P5 said, “You had to remember to press. [...] so if you were forgetful you wouldn't want to have that burden.”

Some participants disliked SITU because they were prompted multiple times during a single trip. Although this issue only occurred to some of the participants, it gave them additional burden and interruption. For example, P10 complained about getting prompts whenever he encountered a stop sign: “By the time I get to the stop sign, it was [like]: ‘Perfect, you got a stop sign.’ And then, [the prompt] would then pop up. I was like, "You stupid [app], [Do] not give me the notification.” Another commonly cited problem was being unable to prevent the app from recording the movement they did not want to record. For example, P13 said, “[...] especially when I didn't wanna record a trip, it would constantly be nagging me. Like when I work, I deliver stuff.” P5 also complained: “... it would record me walking inside, [...]. I was like ‘ugh, just leave me alone.’” Furthermore, participants felt that they lacked control over when to annotate in SITU, as P24 reported: “I didn't like how I couldn't go back to my trips at the end of the day. Like I said, every now and again, I was concerned about not being able to record them... Icouldn't go back and see which ones I forgot to record.” These participants wished there had been a way for them to review and labeled their trips afterwards like in POST.

On the other hand, participants liked SITU for its prompting feature suggesting the current transportation mode. For example, P4 said, “I like that it did have that reminder, it was able to pre-judge what transportation I was actually using” P9 also said, “[...] it was pretty efficient the way that it only prompted when it was a long trip.” He later added, “I thought [it] was intelligent. It can detect when you're in a car, when you're walking, so, which was pretty good. [...] It was always accurate.” To summarize participants’ feelings about SITU, it seems that the detection accuracy of SITU has the major influence. Participants who liked SITU a lot were those for whom Minuku’s transportation detection was accurate.

Finally, participants liked POST in that they only needed to annotate their trips *once* at the end of the day or when they were free, as P34 said, “I really enjoyed being able to [...] fill it all out in one time. [...] It gave me a lot of flexibility. I could label it afterwards. I could label it at the very end of the day when I was sitting down charging the camera.” When asked to rank the three approaches, P28 described an improved version of POST by saying, “The best one would be: have an app which will do efficient tracking, and it will pop up only once in the night. It will do everything in the background, okay?” However, not all participants liked repeatedly annotating their trips all at once, which may have led to less effort being directed towards the annotation task. For example, P29 illustrated this issue in the interview: “Submit. Submit. Submit. [laughter]. Most people will be more diligent so they'll take more time to fill out the reports.”

Another often mentioned dislike about POST was seeing a number of errors, such as recordings that were too short or hard to recognize. For example, P9 said: “Prompting me for a lot of trips which weren't trips actually. [...]. Icouldn't remember what they were, because the map would show like 10 feet or something, like a dot.” It is noteworthy that many of these recordings were the instances of participants walking indoors. Many of these location traces were only shown as “a dot” on the map because participants’ location were provided by a Wifi router.

5.5. Discussion of findings in phase one

We draw on the findings and discuss the pros and cons of PART, SITU, and POST in three aspects vital to collecting annotated activity data through the mobile crowd: *characteristics of the collected data*, and *user experience*.

5.5.1. The characteristic of the collected data

One question for a Participatory approach (PART) versus a Context-Triggered approach (SITU and POST) is: Does automated recording lead to a greater quantity of data compared to manual recording? Our results do not suggest such an advantage. We discussed this unexpected results by considering the *mechanism* and the *implementation* of the approaches. First, we were not surprised to see that when using the two Context-Triggered approaches, participants labeled a considerably larger number of recordings than when they used PART because we assumed PART required more users’ effort. However, we also had assumed that we would have observed a higher percentage of travel activities being recorded and labeled in SITU and POST than in PART for the same reason. Yet, it turned out that we did not observe any significant difference in the trip labeling ratio per day among the approaches in our data. In other words, within a four-day period, participants captured a similar *number of annotated trips of their own*, regardless of whether they used PART or the two Context-Triggered approaches. However, we wonder: if participants could capture a similar number of their trips per day, why did we obtain an obviously larger number of recordings in the SITU and POST condition than in the PART condition? And why did it turn out that the coverage of recordings in PART was larger than it of SITU and POST?

We think these two questions may be explained by the *implementation* of the approaches. Specifically, in the PART condition, participants were able to produce recordings of which the timings were rather close to actual travel activities. And there was a clear one-to-one match between recordings and the actual trips because participants were asked not to split recording intentionally (i.e. there were rare instances where multiple recordings were associated with on trip). In contrast, in the SITU and POST conditions, most recordings, if not all, missed some data at the beginning of a travel activity because of the detection delay, i.e. the window time we used for detecting a start of a trip. In addition, some of these recordings were fragmented and were parts of the same travel activity because of the over-aggressive segmentation caused by false transportation detection. As a result of these issues, we saw a large number of incomplete and fragmented recordings that, overall, led to lower coverage of activity data compared to the PART condition. It is likely that if we had used a different set of parameters—the window times and the percentage thresholds for detection—the results might have been different. For example, not only the length of the missed portion and noises would differ, but also the frequency of a trip being split would vary. However, we argue that these issues caused by the detection error would have no easy solution for our audience. As mentioned earlier, whereas a low detection threshold would cause repeated prompts during the same activity (over-segmentation), a high threshold would impose a significant delay in determining a start or a stop of an activity. Unfortunately, finding a good balance between these two would be simpler in a lab testing/experiment than in a real-world deployment with diverse participants, who have various traveling patterns and situations. More importantly, we believe that many of our audience desire to collect data via the mobile crowd because they have not had a full-fledged context and activity recognizer that is accurate or intelligent enough for field deployment. In addition, the transportation detection of Minuku was developed on top of the Google Activity Recognition service with improvements on accuracy; it would be unrealistic to assume that a similar service would be available for the audience to use for collecting any other kinds of activity. Given these reasons, we think that instead of arguing for not to use a Context-Triggered approach until an accurate context or activity recognizer has been developed, we think a more important takeaway is to understand the potential characteristic of the data being collected so that readers know how to deal with the collected data. After all, the current study was conducted in a setting where each condition only lasted four days. If the duration of the data collection had been longer, whether participants' high performance in PART would have sustained (e.g. using PART for 10 days) is questionable.

5.5.2. *The user experience*

According to the qualitative findings, we identify two key aspects of user experience particularly vital to collecting annotated activity data: *user burden* and *user control*. Regarding user burden, participants generally felt PART least convenient because they needed to remember to record their trips. In contrast, they appreciated the convenience of SITU and POST because of their automated recording and prompt, especially that in POST, they did not need to annotate during the activity in the field as they needed for the PART and SITU approaches.

Regarding user control, participants highly valued being able to control when and what to annotate and record. The fact that participants could only annotate during a trip in SITU made participants anxious about missing a prompt, especially when an activity required their attention (e.g. driving). They favored the flexibility of deciding when to annotate in POST because they could annotate whenever they were free. In addition, participants wanted to control the instrument so that it did not record a trip they were reluctant or did not need to record. However, as mentioned earlier, these issues are specific to Context-Triggered approaches and can be challenging to address due to the lack of a full-fledged context detection system for employing this approach. On the other hand, we think these issues are crucial to address because inaccurate detection is likely to annoy users over time with recurring prompts and thus decrease users' compliance. One solution is allowing users to take control over the recording process when context-detection is not accurate. As context detection improves, users may be willing to cede more control to the system. To summarize, we think it is important that future mobile crowdsourcing tools take both user burden and user control into account to assure good users' experience in recording and annotating activity data. Neglecting either of these two aspects may result in a decrease of users' compliance. However, it is also noteworthy that these two aspects are in tension with each other because more control may lead to more burden. Future research would be needed to explore an ideal combination of the two aspects to make users' compliance more sustainable.

6. Phase two: user behavior analysis

In the first phase of analysis, we focused on comparing the three approaches in terms of the resulting travel activity data collected and the user experience. In Phase Two, we primarily focused on understanding participants' recording and annotation behavior in the field, i.e. how participants recorded and annotated their activity using PART and SITU. It is important to note that, in the previous phase, we observed some behavioral aspects of the participants. For example, participants generally were able to record their activity precisely using the PART approach. However, in this section, we dig more deeply into participants' recording and annotation behavior by inspecting their interactions with Minuku in using the two approaches. In Phase Two, we focus on *activity type* instead of *transportation mode* in terms of the impact on annotation behavior. As an example, instead of distinguishing between cars and bus, we distinguish between Drivers and Passengers. We make this distinction because we think these two activity types demand different degrees of attention from participants, which we think would be influential on when and how participants would annotate their travel activity when traveling. These activity types were provided by the participants' assigned labels to each trip.

In addition, it is important to note that we did not we analyze participants' behaviors in POST despite the fact that we did collect and organize the data in this condition. We chose only to focus on PART and SITU because we were mainly interested in understanding user behaviors "in the field," i.e. when participants were mobile and situated in a travel activity. Although participants sometimes annotated their recordings when they were on the go in the POST condition, most of our participants, according to the interviews and based on our preliminary inspection on the behavioral logs, more often annotated their recordings at the end of the day at home (usually after receiving the annotation reminder). Below, we provide more details of the analysis and the findings.

6.1. *Behavior log analysis*

As mentioned in Section 3.4.3, we collected participants' usage logs in Minuku, representing all actions that participants performed within the tool. Analyzing these logs allowed us to understand *when* and *how* participants recorded and annotated their travel activities. Specifically, we measured: a) when participants started and stopped recording in PART, b) when participants started, submitted, and completed annotations in both PART and SITU, and c) how many sessions (a series of actions performed in a continual manner) participants undertook to complete the entire annotation process. After obtaining these measures, we examined the influence of *activity type* on these measures, i.e. whether a difference in these measures existed among different travel activities. The activity type for each recorded trip was determined by the participants' assigned label and was classified into three categories: *Driving*, *Passenger* (whether by bus or by car), and *Walking*. These are common travel activities, yet they demand different degrees of attention from participants. As a result, we expect to observe some differences in participants' annotation timings during different travel activities. In addition, we also compared participants' recording times with the Ground Truth Trips to examine whether they tended to start/stop recording their trip earlier or later.

In analyzing participants' recording and annotation behaviors, we had different specific research questions for PART and SITU because of their different mechanisms for collecting annotated data. For PART, we analyzed the influence of activity type on users' annotation completion time—the elapsed time of participants' last annotation submission in relation to the start of recording. That is, we aim to investigate whether participants would tend to finish the task right after they started recording, during the trip, or after the trip. Because the elapsed time is highly correlated to the length of the trip, we classified the completion time into three levels of an ordinal measure: START (3)—completing

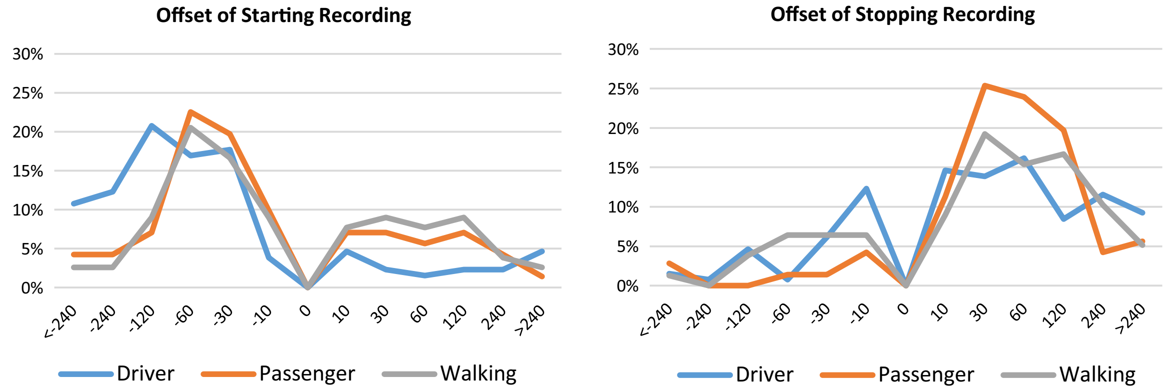


Fig. 9. (Left) Most recordings were started before the actual trips started, and that Drivers started earlier than other two activity types. (Right) Most recordings were stopped after the actual trips ended, though there is no statistical significant difference among activity types in terms of when late recordings were stopped.

annotation within 60 s after the start of the recording; DURING (2) —completing annotation between one minute later the recording and before the end of recording; AFTER (1)—completing annotation after the end of the recording. The 60-second threshold was decided based on two observations: a) a typical duration that were sufficient for participants to type a note with enough details (e.g. “'traffic was VERY heavy due to rush hour'”), and b) the distribution of the annotation completion times in PART (participants' annotation submissions started to scatter throughout the recording after 60 s (shown in Fig. 10, right). We made sure that the duration was long enough for participants to type details because we would examine the influence of annotation timing on the length of notes in a statistical analysis. We refer to this ordinal measure as *Annotation Completion Timing* in the rest of the paper. A higher rank indicates an earlier time for completing the annotation task.

For SITU, we investigated participants' receptivity to annotation task requests. For our purposes, an annotation task was “responded to” by a participant when the participant started to annotate through the prompt. Our measures included: a) the percentage of annotation prompts responded to by the participants, b) how quickly the participants responded to the prompts, and c) how quickly the participants completed the annotation tasks. These measures displayed how receptive participants were when they were requested to collect annotated travel activity data via the mobile phone. We did not measure participants' recording times in the receptivity analysis because Minuku automatically started recording on its own in the SITU condition. We also grouped participants' Annotation Completion Timings into START and DURING using the same 60-second threshold (there was no AFTER for SITU).

Finally, we inspected participants' behavioral logs to look for emergent patterns that recurred and were distinct from participants' typical patterns in recording and annotation. From this inspection, we were able to uncover issues causing erroneous activity data. We also measured the length of annotations and investigated the influence of activity type and Annotation Completion Time on the length and content of notes.

We ran mixed-effects regression models for all of the quantitative analysis. Specifically, we ran mixed-effects linear regression on numeric dependent variables (e.g. recording time, annotation time, the length of note), mixed effects logistic regression on binary dependent variables (e.g. whether an annotation prompt is responded to), and mixed-effects ordinal logistic regression on ordinal dependent variables (*Annotation Completion Timing*). For all analysis but one we included Activity (Driver, Passenger, Walking), periods of the day, and day of the week as fixed-effect independent variables. We used transportation mode (car, bus, walk) rather than activity (Driving, Riding as Passenger, Walking) for the analysis of response rate because we did not know whether or not a participant was a driver or passenger if they did not respond to the annotation task (note that we rely on their labels to know the activity type). We could have inferred this information from the ground truth photos of the wearable cameras. However, this inference would be unreliable. When coding photos, we found that it was difficult to distinguish between driving and being a passenger from some photos where the camera was not facing toward to the front but the car ceiling.

6.2. Qualitative analysis

6.2.1. Content analysis of annotations

We conducted a content analysis of participants' free text annotations (i.e. notes that participants added to recordings). Note that users were given freedom as to whether to provide a note and what to write in a note. Surprisingly, even when the participants were aware that the note was optional, they provided 272 notes in the PART condition (64% of 424 labeled recordings) and 352 notes in the SITU condition (49% of 723 labeled recordings) for SITU. Two co-authors of the paper independently coded the recorded notes obtained in PART and SITU. The codes were categorized into various categories such as routes (departure, destination), the context of the trip, intent behind/purpose of the trip, routineness, and errors. We assessed the inter-rater reliability (IRR) of the codes and obtained a Cohen's kappa value of .90, which indicates a high agreement between the two coders on the coded content and characteristics of annotations.

6.2.2. Diary and interviews

We also analyzed participants' diary entries and revisited the interview data with a new focus on user behavior. Specifically, for diary entries, we focused on reasons why participants did not record and/or annotate their travel activities. For interview data, we sought to understand participants' overall recording and annotation behaviors and strategies in using PART and SITU as well as the issues they encountered that might have interfered with their recordings and annotations.

6.3. Results: recording and annotation behavior

6.3.1. Recording timing in PART

Our first result is regarding recording timing. We want to examine, overall, whether our participants would tend to record before or after the start of a trip. When we compared participants' recordings in PART with Ground Truth Trips, we found that, on an average, participants started recording their trips 46.2 s earlier than the start of the trip (Median=28, SD=221), and stopped recording 58 s after the end of the trip (Median=28, SD=218). In particular, 72.4% of recordings started

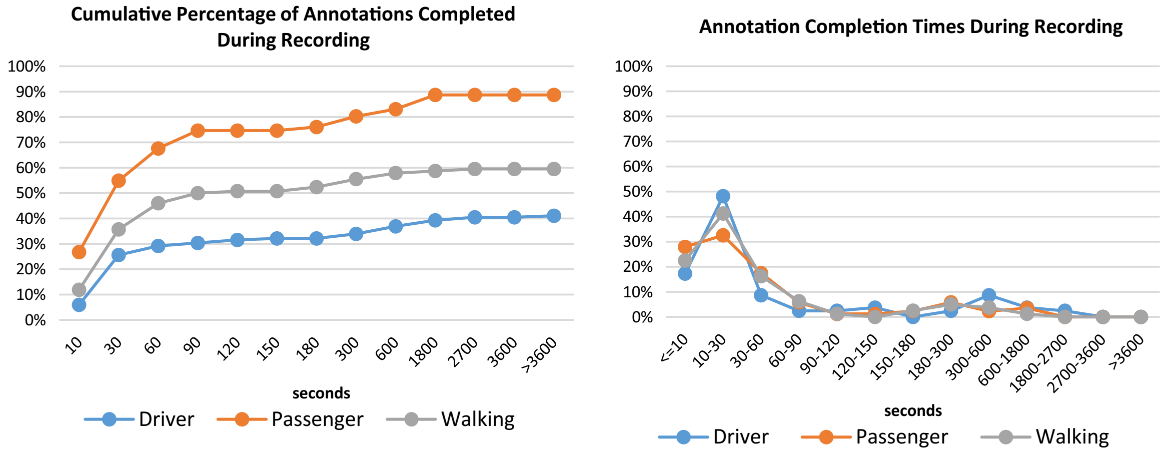


Fig. 10. Annotation Completion Timing using PART. (Left): Cumulative percentage of annotations completed during recording. The remaining percentage represents annotations completed after recording. For example, as Passengers, 88.7% of annotations were completed during recording. 11.3% were completed after recording. (Right): Percentage of annotations completed between certain times during recording.

earlier than Ground Truth Trips (Fig. 9, left), and 78.5% of their recordings ended later than Ground Truth Trips (Fig. 9, right). Furthermore, we found when participants were Drivers, they recorded their trips earlier than when they were Passengers or Walking (Fig. 9, left), and the difference between Drivers and Walking was statistically significant (t(X)=2.6, p=.01). We did not observe any statistically significant differences across activity types in terms of when recordings were stopped. These results complement well the results obtained in Phase One. That is, although in the PART condition participants produced recordings both with missed portions (i.e. record after the activity starts) and noises (i.e. record before the activity starts), respectively, overall, participants more often recorded before than after the start of activity, regardless of activity type). The impact of activity type seems to be mainly on *how much earlier* participants started the recording. As a result, researchers may expect to see noise more often than missed portions at the beginning of recordings when participants use a PART approach.

6.3.2. Annotation completion timing in PART

In the analysis of Annotation Completion Timing in PART, we found a strong effect of activity type on Annotation Completion Timing. Specifically, we found that participants were more likely to complete annotation tasks at the start of recording or during recording when they were Passengers (M=2.51, SD=.89) than when they were Drivers (M=1.70, SD=.74, p=.01) and Walking (M=2.10, SD=.93, p=.05). Specifically, when participants were Passengers (the orange line), 88.7% of the annotation tasks were completed during recording, and only 11.3% were completed after recording. However, when they were Walking (the gray line) or Drivers (the blue line), only 59.5% and 41.1% of annotation tasks were completed during recording, respectively. In other words, when participants were Drivers, nearly 60% of annotation tasks were completed after recording. Fig. 10 (left) shows such a pattern. It plots a cumulative percentage of annotations completed during recording among all recordings (including recordings annotated before and after recording). This finding is also supported by the number of sessions participants spent to complete annotations. Our results showed that when participants were Passengers, 94% of annotations were completed in one session; in contrast, only 60% and 64% of annotations were completed in one session when users were Drivers or Walking, respectively. The differences between Passengers and Drivers was statistically significant (t(389)=2.4, p=.02), and between Passengers and Drivers was marginal (t(389) =1.78, p=.07). More interestingly, we observed that among annotations completed during recording, participants tended to complete annotations sooner rather than later. As one can see in Fig. 10 (right), among the annotation completed during recording, the majority of them were completed within one minute (Driver: 71%, Passenger: 77.5%, Walking: 77.3%). In addition, as a Passenger, participants 27.9% of annotations were completed within 10 s. However, when they were Drivers or Walking, only 17.3% and 22.5% of annotations were completed within 10 s, respectively. Thus, this shows that even among instances where participants could complete annotations soon, activity type still had an impact on how quickly participants completed annotations.

In the interviews, we asked participants when they annotated their recordings. Many of them reported that they preferred to annotate soon so that they would not forget later. Participants especially mentioned that they would annotate soon when they were taking buses or walking because did not need to concentrate as they needed to when driving. For example, P9 reported: "I'm sitting in a bus anyway, so there's nothing to do. You can just quickly do it if you're sitting in the bus. [...] Walking also, there's nothing to do, right? You only have to walk." In contrast, when participants were Drivers, they needed to concentrate on driving; they reported that when driving they annotated while they were at breakpoints (e.g. stoplight) or after they stopped their trips. As P26 said, "If I'm walking I do it pretty [much] right away because it's not much of a deviation. If imagine I'm in a car, thenIgenerally respond to it whenever I think it is safe or whenever I kind of stop the car."

6.3.3. Users' receptivity to annotation requests in SITU

For the SITU approach, we mainly analyzed participants' receptivity to the annotation task. Specifically, we analyzed three aspects of receptivity: a) response rate of the annotation prompt, b) how quickly participants responded to the prompts, c) and how quickly participants completed the annotation tasks. For response rate in particular, as noted earlier, we had to use transportation mode (Car, Bus, Walking) rather than activity type because we did not have reliable information about whether participants were Driver or Passenger in a car in the recordings they did not label. This is because photos are not a reliable source for inferring this information since some photos only showed the car ceiling. We found that on an average, participants had high response rate to annotation prompts across all transportation modes (Car: 86.7%, Bus: 88.9%, Walking: 81.1%). We think such a high response rate might be because participants had expected to be prompted whenever they were traveling, thereby becoming more receptive to the phone notification.

However, among the prompts that were responded to, where we had the activity label provided by the participants, participants responded more quickly when they were Passengers and Walking than when they were Drivers (Passenger vs. Driver: (t(418)=−3.79 p < .001);

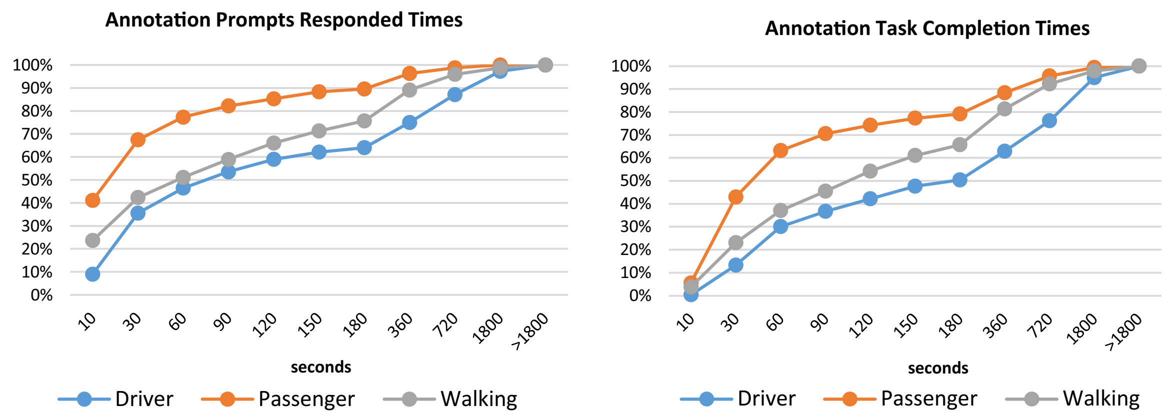


Fig. 11. (Left): Cumulative percentages of annotation prompts that were responded to within certain time in the SITU condition. (Right): Cumulative percentages of annotation tasks that were responded to and completed within certain time.

Walking vs. Driver: (t(418)=−3.31 p < .001)). As shown in Fig. 11(left), specifically, 67.5% of prompts were responded to within 30 s when participants were Passengers; but only 42.4% and 35.6% of prompts were responded to within 30 s when participants were Walking and Drivers, respectively.

In addition, participants also *completed* annotation tasks more quickly when they were Passengers (t(417)=−2.8 p=.006) and Walking (t(417)=−1.9 p=.06) than when they were Drivers, as shown in Fig. 11(right). Specifically, 42.9%of annotation tasks were completed and submitted within 30 s when participants were Passengers, but only 23.1% and 13.3% of annotation tasks were completed and submitted within 30 s when participants were Walking and Drivers, respectively.

We further looked into how quickly participants completed annotation tasks once they had responded to the prompts. Interestingly, we found that most of the time participants completed annotation tasks within a minute after they responded (Passenger: 95.7%, Walking: 94.7%, Driver: 89.45%). And Moreover, the differences between Passengers and Drivers and between Walking and Drivers are both marginally significant (Driver vs. Passenger: t(417)=1.94, p=.05; Driver vs. Walking t(417)=2.00, p=.05).

These results suggest at least two things. First, although activity type appeared to influence how quickly participants completed an annotation task once they have responded to it, its main impact on receptivity seems more related to how quickly participants *could respond to* the prompt. Second, similar to the PART condition, the fact that most of the time participants completed annotation within a minute regardless of activity type indicates that participants also tended to complete annotation sooner rather than later in SITU.

From the interviews, we also asked participants about when they annotated their trips in SITU. Most participants were well aware that they were in the study and would expect to get prompts when they were traveling. A typical explanation for their immediate response to the prompt is as what P4 said: “I know it's gonna pop up sometime here soon. I just kept looking at my phone. I gotta remember that it's going to come up.” Many users added reasons why they preferred to annotate immediately. Similar to using PART, the main reason for performing it early was to prevent them from forgetting to do it later. P15 said: “I immediately respond, so that I don't forget it later so, ‘Okay. I've seen the notification so let me get over with it now.’” P5 also said: ”I just want to get it done. I didn't want to miss it. [...]. I was very careful at the beginning and then I was worried, because I wanted to do it right away.” However, whether such a high response rate would sustain if the study duration were longer remains a question.

On the other hand, when participants were driving, they deferred the response to a point where they felt safe to complete it, as P22 said: “I'd get the notification while I'd be in the process of driving so I'd have to wait 'til I was at a light or something, and kinda answer it or try to remember not to hit "Submit" [laughter] and then set it down, then go about my business.”However, sometimes it was hard for users to anticipate how long a breakpoint (e.g. stop light) is; thus, some users would defer it until the end of the trip before the notification disappeared. For example, U36 stated: ”In many cases that I don't know how much time I have at the light. And rather than, just leave it in the middle, I'd wait till Iwasn't traveling anymore.” Taking these results together, participants seemed to prefer to respond to the prompt and complete the annotation task early if they are not preoccupied with the activity. We think this behavior cannot be all attributed to the fact that they could only annotate during the trip because participants also displayed the same tendency in using the PART approach, for which they could annotate whenever they wanted to.

6.3.4. Characteristics of participants’ notes

To understand how annotation timing and other factors would affect the characteristics of participants’ notes, we conducted a content analysis of all submitted notes and analyzed the effect of activity type on the length of the notes. For the former, we observed an interaction effect between activity type and Annotation Completion Timing on the length of the notes, as shown in Fig. 12. Specifically, we found that when participants annotated AFTER recording when they were

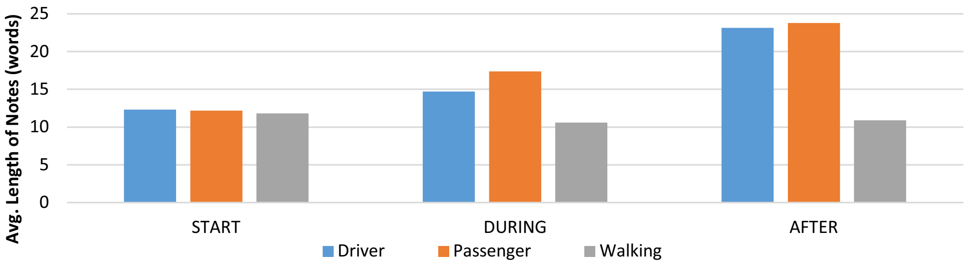


Fig. 12. Participants generally wrote short notes when they annotated at the beginning. When users annotated AFTER recording when they were Passengers and Drivers, they tended to put longer notes than when they were Walking.

Passengers and Drivers, they tended to put longer notes than when they were Walking (Passengers: t(1046)=1.95, p=.05; Driver: t(1046) =2.37, p=.02). In addition, when participants annotated During recording, they also put longer notes when they were Passengers than when they were Walking (t(1046)=2.88, p=.004). However, we did not observe an effect of activity type when participants annotated at the START of recording. Instead, while participants annotated at the START of recording during walking, their notes were generally short. These results suggest that participants seemed to put short notes when they were walking, regardless of the annotation timing and that when they put notes early in the trip, they tended to put shorter notes. They were able to put longer notes when they were Passengers, which we think might be because they had more attention available compared to other activity types. Finally, when they put notes after the activity, they tended to put longer notes, except when they were walking, perhaps because walking was too a routine activity for them.

Regarding the content of notes, we found different characteristics of participants' notes according to transportation mode, activity type, and annotation timing. Specifically, we found some types of information appeared in annotations for one transportation mode more often than in those for other transportation modes. For example, participants more often described multiple-destinations (e.g. 'driving daughter to school then work') and purpose of trips in annotations of car trips than in annotations of the bus and walking trips. We suspect that this might be because participants' bus and walking trips were more routine trips, whereas participants had car trips for more diverse purposes. We also found that participants more often included information of transportation mode when they were walking (*e.g. "walking to the library where I volunteer twice a week"*) than when they were in a car or on the bus. Furthermore, when participants were Drivers and Walking, the annotations made at the Start contained fewer words and categories of information describing their trips. That is, whereas the notes added at the Start mostly contained destinations and purposes of the trips, notes added later included more details such as with whom the users were traveling, details of the route, and events occurring during the trip. However, when participants were Passengers, they use similar categories and number of words to describe their trips regardless of when the annotation was created. We think this might be because as a Passenger, participants had abundant time and cognitive resource to annotate during a travel activity. In contrast, when participants were Drivers or Walking, in which they had to spend more attention resources on performing the travel activity itself, participants did not seem to be able to include more information during the activity but after. Finally, we also found that participants more often used shortened descriptions when they were Drivers, indicating their tendency of making notes as efficient as possible. Taken these results together, it seems that two major reasons are mainly responsible for explaining the characteristics of notes: what context is relevant to annotate about the activity, and the availability of participants for annotating the activity.

Another interesting observation we had is that some participants would assume a common ground shared with researchers, which made them shorten descriptions over time or referenced a trip to previous trips (e.g. *"to recycling from home"* ⇒ *"more recycling"* or *"walking to great clips to get haircut"* ⇒ *"'still walking to hair cut place"*). Some of these occurred because users were prompted multiple times in one trip in SITU.

6.4. Reasons for unrecorded, unlabeled, and erroneous activity data

Finally, we analyzed diary entries, interview data, and inspected behavioral logs to identify reasons and patterns that caused unrecorded, unlabeled, and erroneous activity data. We found that forgetting and missed notifications were responsible for most of the unrecorded, unlabeled and erroneous activity data. Specifically, from diary entries, we found that the major reason contributing to unrecorded activity data was participants forgetting to record (18 out of 38 unrecorded trips). Other often cited reasons included feeling it was troublesome to record (8 out of 38) and feeling it was inconvenient to record (7 out of 38).

For unlabeled activity data, in PART the main reason reported by participants was forgetting to label (10 out of 23); in SITU, the main reasons were not part of their plan to annotate (93 out of 250) and missed notifications (88 out of 250). These results show that participants could either intentionally and unintentionally not respond to a prompt, and it seems that both reasons, at least in our study, seemed to be of equal importance. However, it is important to note that many of the unresponded prompts were generated upon false detection such as detecting indoor walking, to which participants did not need to respond.

As to erroneous activity data, we learned from the interviews that many participants forgot to stop recording their trips after they had ended their trips when they used PART. This sometimes resulted in unnecessarily long recordings, large portions of which were incorrectly labeled. According to the participants, the main reason causing them to forget starting and stopping the recording was distractions in the moment or that they had been preoccupied with other things, as P22 said, "So and then the one time I forgot to stop and transition from walking to driving... I just had a lot on mymind so I just didn't think about so I went all auto pilot." U15 also reported, "because you have to get down, you have to cross the street, you have to choose which shop to go to. So yes, I tend to forget here." In particular, one common source of distractions reported was interacting with other people, as U20 said: "Because oftentimes, I'd be wrapped up in what I'm supposed to be doing, or maybe I met a friend when I was walking, and we're walking together, and then I forgot."

As to SITU, we observed from the behavior logs that many labeling errors occurred at transitions between travel activities. One typical case was that participants did not respond to the prompt until they were about to start a new trip. Another case was that participants changed labels because they thought they were transitioning to a new trip. For example, P10 commented his strategy of labeling his trip in SITU: "If I got the notification right away when I was driving, then I'd put 'driving.' But sinceI would go back and I would always check it numerous times, so then over to walking instead of the driving, then I'd probably go and switch it to the walking." In SITU, these issues seemed to be related to the delay of annotation prompts when participants transitioned to a new trip, and the issues often occurred when the transition was short such as walking to a car. While participants were instructed to provide the transportation mode that was current as of when the prompt was issued, not when it was received, the participants would provide the mode when they responded to the prompt.

Below, we discuss the findings and conclude with implications for the design of a mobile crowdsourcing tool for collecting annotated activity data from individual mobile workers.

6.5. Discussion of findings in phase two

6.5.1. Possible reasons behind the influence of activity type

Our findings suggest that activity type influenced participants' recording and annotation timing, receptivity, and the characteristic of their notes added to the recordings. Here we discuss the possible reasons for such influences.

First of all, regarding the recording timing, when participants were Drivers, they tended to record their trips earlier than when they are Passengers or Walking. We conjecture this might be related to the length of transition to the trip. For example, a transition to a car involves multiple stages (e.g. opening a door of the car, sitting in a car, and waiting for the car to move) and thus is longer than a transition to walking. As a result, participants would have more time to record at transitions to driving than at transitions to walking. Another reason that might explain the differences in the recording time would be participants' perception of the amount of attention required during the travel activity. Drivers might perceive a challenge of recording their trips precisely at the moment when they start traveling and thus tend to start recording earlier. On the other hand, although the transitions to bus trips might be longer than the transition to walking trips, the fact that being a Passenger requires limited attention to the travel activity might explain why the participants did not tend to record as early as for Driving.

The impact of the amount of attention required to perform an activity is also evident in the differences in the Annotation Completion Timing among different activity types. For instance, in both the PART and SITU conditions, although our results suggest that participants tended to annotate early rather than later, participants completed the annotation task quickest when they were Passengers and slowest when they were Drivers. Moreover, participants also more often used multiple sessions to complete the annotation task when they were Drivers and Walking. Drivers also completed their annotation after the recording in about half of the cases. In the SITU condition, participants also had a lower receptivity to annotation prompts when they were Walking or Drivers than when they were Passengers. These results taken together indicate that the level of attention required by an activity has an impact on user's annotation completion timing. This observation was also supported by many participants' self-reports that they would annotate during breakpoints (e.g. stoplights) or after driving when they were a driver.

Finally, our results suggest that both Annotation Completion Timing and the context of activity might have an impact on the content of notes added to the recordings. For the former, notes added at the START contained limited categories of information of the activity compared to those created later. For the latter, participants more often described purposes of the trips and included multiple destinations when they were in car trips than they were on a bus and walking. This difference might be because the bus and walking trips users recorded were more routine trips, whereas users went to more diverse places when they were in cars. Finally, the context in which activity is performed might also affect whether and to what extent users were distracted or preoccupied. This might in turn influence how likely users would be to remember to stop recording.

6.5.2. Anticipating characteristics of collected data

Following the discussion above, we summarize four features of an Activity that may influence the quality and the characteristics of activity recordings and annotations. These features are a) length of transitions b) degree of attention required for performing the activity, c) distribution and lengths of breakpoints during the activity, d) possible contexts in which the activity is performed. Specifically, based on our observations of the results, we first conjecture that recording timing mainly correlates to the length of transition to start and stop the activity and the degree of attention required for performing the activity. The longer the transitions are, the more likely users would be to record earlier and stop recording later. Second, we conjecture that Annotation Completion Time mainly correlates to the degree of attention required and the distribution and lengths of breakpoints during the activity. The more the attention required for users to perform the activity and the fewer and shorter the breakpoints are, the more likely the users would annotate late or after the activity. Third, we conjecture that content and characteristics of annotations mainly correlate to the degree of attention required, the distribution and lengths of breakpoints, and the context in which the activity is performed. In other words, content and characteristics of annotations depend not only on how much time users can spend on annotation but also on what information is relevant to the current activity. The latter is especially true when the activity is more routine activity in users' daily lives. The users not only may have limited categories of information to describe a routine activity but also, may feel bored by annotating same information repeatedly. As we showed earlier, participants would shorten annotations when they found the researchers had known the activity.

Fig. 13 shows a presentation of an *activity* with these four features. It is important to note that activity is a complex phenomenon (Nardi, 1996) and Fig. 13 is only a provisional and simplified representation of an activity for the purpose of introducing the four features we found to be critical to collecting annotated activity data. This schematic may help researchers anticipate the type of errors that may occur in data collection and the characteristics of collected data, such as how long the noise is, how long the recording misses a travel activity, what information would be included in annotation, and so on. For instance, if there tends to be a long transition to the activity of interest and the activity demands some attention from the user, researchers may anticipate that the user is likely to record before the activity starts and that the recording would contain noise in the beginning. If the researcher anticipates that the activity of interest does not demand much and continual attention or it does but contains many breakpoints, the researcher may anticipate that the user annotates early in the activity. One potential issue with early-made annotations is that the user may not mention events occurring later in the activity in the annotation unless they are explicitly instructed to do so. On the other hand, late-made annotations are also likely to neglect events that occurred early on, if additional salient events occurred later. Finally, researchers may be able to predict whether users are likely to forget to start and stop recording using the PART approach by anticipating possible distractions during transitions before and after the activity. They may also anticipate how likely the user mislabels previous activity using the SITU approach, given the length of transitions and the current context-detection method researchers use.

In the last section below, we conclude our findings with a list of suggestions for future work aiming to use mobile crowdsourcing to collect individual annotated activity data.

7. General discussion

7.1. Towards a better practice of collecting annotated activity data

In this paper, we present a field study that aimed to identify an approach that would be reliable and effective for collecting annotated activity data through the mobile crowd. Our study shows several

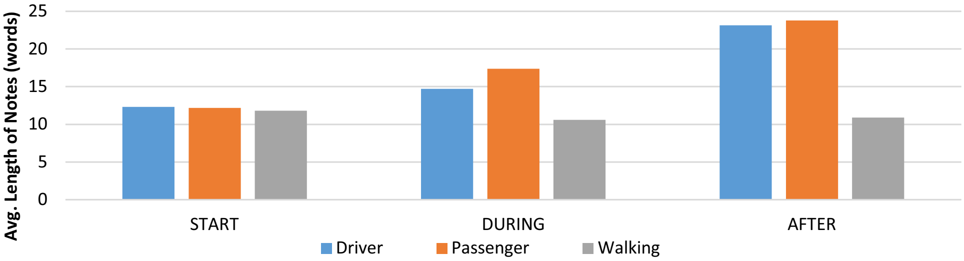


Fig. 13. A simplified presentation of an activity with four features: a) length of transitions b) degree of attention required for performing the activity, c) distribution and lengths of breakpoints during the activity, and d) possible contexts in which the activity is performed.

important takeaways that shed lights on the approach, tool, and instruction that will make mobile crowdsourcing appealing for collecting annotated activity data.

First of all, despite the fact that the Context-Triggered approaches may produce a larger number of recordings, we show that many of these recordings may be fragmented and contain noise, with the result that Context-Triggered approaches may not necessarily produce a greater quantity of annotated activity data in terms of length of time. In addition, because of the presence of noise, data requesters would need to process and clean the data further to make them directly usable. In contrast, despite the smaller number of recordings, the data produced by the Participatory approach would be more complete and contain less noise. As a result, the quantity of the collected data, in fact, is possibly larger than the quantity of the collected data produced by Context-Triggered approaches.

On the other hand, regarding user experience, whereas the Participatory approach is generally more burdensome than the Context-Triggered approaches because it requires more users' effort, the SITU approach can sometimes be considered annoying if it prompts the user repeatedly or at the time when the user does not want to annotate the trip. It may also produce data that users cannot recognize nor annotate. However, a Context-Triggered approach is still worth the hassle to develop and employ because it reduces users' burden of data collection. We believe that in the long run, such a reduction would be necessary for sustaining users' compliance.

Because of these tradeoffs between Participatory approach and the Context Triggered approaches, we do not conclude that one approach is *better* than the others. Rather, we think a more important lesson is that we must understand the strengths and weaknesses of each approach so as to develop better practices for collecting annotated activity data via the mobile crowd. Our proposal is to use a hybrid approach that combines the strength of the Participatory and the Context-Triggered approaches. We will be present details of our proposal in the next section.

Another important takeaway from the study that we must understand users' behaviors in using each approach with respect to the nature of the activities being collected, so that we can better anticipate the characteristics of the collected data. In turn, this understanding will help us know how to process the collected data and use it later. These understandings also inform us about how to improve the design of data collection campaigns and how better to instruct participants to collect data. For example, our results suggest that participants tend to add annotations sooner rather than later. Moreover, the annotations created at the start tend to contain limited categories of information compared to those created later. While we cannot assert that these differences should all be attributed to annotation timing, the fact that most of the time participants rarely revisited annotations made at the start implies that events occurring later in activities would be less likely to be mentioned in the annotation. These pieces of information (e.g. encountering traffic jam later in the activity), however, may be valuable for researchers to later sort out the collected data or to get inspiration for the activity they can collect later. In addition, when using the Participatory approach, users may tend to start recording before the start rather than after the start of the activity. As a result, researchers may expect noise at the beginning of a recording. When using the SITU approach, users would be more receptive to annotation tasks when they are performing an activity that demands less attention. As a result, when the researchers deliver a data collection task to the mobile crowd in the field, the particular activity that the user is currently performing is vital to take into consideration. Finally, despite the high response rate in SITU, being non-responsive to a prompt is still reported as a major reason for not labeling a recording using the SITU approach. It can also be a potential cause of mislabeling the previous activity if they user label it in the next activity. As a result, a future approach should design a better mechanism to increase users' responsiveness or to better handle non-responded annotation tasks to minimize the mislabel error.

7.2. Design and methodological suggestions

Based on the takeaways mentioned above as well as other findings reported, we propose a list of design and methodological suggestions that aim to inform the approach, tool, and instruction for using mobile crowdsourcing to collect annotated activity data. Our goal for these implications is to improve the overall quantity and quality of the collected data as well as to sustain users' compliance. Because the tool is the instrument which users use to perform the approach, we combined the implications for approach and tool in Section 7.2.1. Then in Section 7.2.2, we provide suggestions on instructions.

7.2.1. Suggestions for the approach and tool for activity data collection

Our high-level suggestion on the approach and tool is to employ a hybrid approach, using the Participatory approach as the main approach to grant user control and use a Context-Triggered technique as a support to ease user burden, to remind users, and to prevent data collection errors. While granting user control and easing user burden can be seen as a design tradeoff, our experiences convince us that these two elements can be balanced to improve not only user experience but also the quantity and quality of data collected.

Specifically, we suggest researchers to encourage users to manually record their activity to increase the accuracy of data as well as to provide user control; meanwhile, a Context-Triggered function, if available, can run as a fall-back to deliver reminders and to enable automation when it is necessary. Regardless of whether the Context-Triggered function is activated or not, the tool should allow users to control when they want this function to be on and off to prevent the tool from recording and prompting them when they do not want to be bothered.

The Context-Triggered function provides several important benefits. First, it can trigger reminders when it detects that participants have forgotten to start recording their activity. The tool then can remind the users to annotate. Similarly, when it detects that participants have forgotten to stop recording an activity, it can automatically stop recording. Although this may result in some portions of activity not being recorded, it would help reduce unlabeled data and prevent a long period of noise at the end of the recording, thus making the data cleaner. In addition, the reminder notification should reside in the notification center even after the trip has ended. A reminder residing in the notification center during the activity will increase users' awareness of an ongoing recording and allow them to annotate it at breakpoints. Leaving the reminder in the notification center after the activity ends provides users with more control of when to annotate. It also avoids the unnecessary pressure and anxiety of needing to complete the annotation task during an activity that demands attention. Furthermore, the annotation reminder can indicate an aggregated number of recordings that are waiting for users' responses. This may make them be mindful of the presence of unannotated recordings and can remind them to annotate sooner while they still have a fresh memory of what happened during the activities.

To ameliorate the issue of mislabeled recordings, a Context-Triggered function can detect whether an activity to be annotated is likely to be a transition (e.g. a short walk to taking a bus). When detecting such an instance, a reminder can ask users to verify whether their label should be associated with the transition activity (walk) or the next activity (bus). Another alternative to avoid mislabeling errors is to let the instrument start recording only after the users have responded to the annotation prompt instead of at the moment of detecting the activity. This will assure that the label provided by the users correctly reflects the activity being recorded at the moment when the users see the prompt.

Finally, to further ease user burden, the Context-Triggered feature can suggest a label where possible, meaning that users only need to change the label if it is incorrect. When the tool detects the users being in the same activity consecutively, it could ask whether this is a continued activity and if yes, it could automatically connect the current recording to the previous one. Detecting an opportune moment [32] for delivering the prompt during or after an activity can also avoid interrupting the user.

7.2.2. Suggestions on the instructions for activity data collection

Regarding instructions, because users tend to start recording before the activity and stop recording after the activity, respectively, we suggest that researchers explicitly instruct users to be as precise about the recording timing as possible to reduce noise in recordings. However, as it is not always convenient for the users to operate the tool at when the activity starts and ends (e.g. driving), the tool may allow researchers to enter anticipated lengths of noise at the beginning and the end of the recording, respectively, and then trim the recording accordingly. In addition, because users tend to annotate sooner rather than later in the activity and do not often revisit the annotations, we suggest that researchers instruct users to be mindful of the events occurring after their annotations and encourage them to revisit annotations after the activity. On the other hand, we also suggest researchers interested in knowing more about the semantics of the activity instruct users to include the intent behind or the purpose for the activity in the annotation, especially early in the activity because they will remember it better. From our experience in analyzing the content of the annotations, we found this information particularly helpful to understand the personal meanings of the activity, which would be difficult to infer from the raw data. Although the intent information may not be essential for detecting the activity per se, it is useful for distinguishing among variances within the same travel activity, such as identifying personally significant places, predicting where the user is departing for, and recommending places of interest (e.g. Andrienko et al., 2010; Ashbrook and Starner, 2003; Baltrunas et al., 2011; Bhattacharya et al., 2012; Cao et al., 2010; Liao et al., 2007).

With these improvements proposed, our future work includes both implementing these features on Minuku, the instrument we used for the study, to implement the hybrid approach, and to examine whether the proposed features would increase the effectiveness and improve the user experience of the process of collecting annotated travel activity data with the mobile crowd. We plan to employ the tool and the approach to collect other activity types. Meanwhile, we hope that these design suggestions will enable researchers and practitioners interested in using mobile crowdsourcing to collect activity data to collect a greater quantity and quality of activity data and annotations.

7.3. Limitations of the study

It is important to note that the study is subject to several limitations. First, the Ground Truth Trips were reconstructed generated where photos were available. As a result, despite the fact that we instructed participants to wear the wearable camera all day, we could not be 100% sure that they did so as anticipated. This might make the photos be subject to a systematic bias related to the availability of photos. Second, the sampling rate of the camera is one photo per 30 s. Although we used logs to establish more precise times of Ground Truth Trips, there might be still some imprecision on the start/end times. Third, our analysis was based on a relatively small sample of smartphone users in a particular area. Their behavior may not be representative of the general mobile user population, especially when it comes to the differences in the dynamics of transportation activities in different geographic areas. Fourth, we do not know whether users were passengers or not in a car when they did not respond to an annotation prompt. As a result, in the analysis of response rate we had to use the transportation mode information from Ground Truth Trips instead. Fifth, participants only used each approach for four days. Their compliance was likely to change if the study had been longer. For example, participants were likely to be less compliant in using the PART approach that was considered more burdensome to perform. We think it is worthwhile to study the same phenomenon with a longer duration. Sixth, in our study we only asked participants to collect travel activities; it is possible that the results of the comparison might have been different if we had chosen to collect other types of activities (e.g., exercise). However, in analyzing the impact of activity type on labeling behavior, we sought to tease out the essential features of an activity that causes the impact as we show in Fig. 11, so that the insights can be more generalizable to collecting annotations for other types of activity. Thus, for example, while our study showed specifically that drivers tend to annotate after the trip, the more important takeaway is that that users tend to annotate after the completion of an activity requiring high attention and containing few and short breakpoints. These characteristics (attentional demand and the nature of breakpoints) are present to different degrees in other activity domains, including activities of daily living in the home, physical exercise and other health-related activities, general time usage, etc. Thus, we believe our findings relating activity features to users' data collection behavior around those activities can help researchers anticipate the characteristics of collected data based on qualities of the activities being studied. Finally, the findings and the errors we presented might be specific to the instruction, tool, and the approaches we used to collect activity data. However, it is important to note that the ultimate goal of the paper is to inform the tool and the approach for using mobile crowdsourcing to collect annotated activity data. We believe the design and the instructional implications we draw from the findings advance toward the goal in this regard.

8. . Conclusions

In this paper, we present results from a field study in which participants used a Participatory (PART), a Context-Triggered In Situ (SITU), and a Context-Triggered Post Hoc (POST) approach to record and annotate their personal travel activity data. Our objective of the study is to understand the effectiveness of these approaches in collecting personal activity data via the mobile crowd, to understand users' recording and annotation behavioral patterns in using these approaches, to understand users' experience in collecting their own activity using each of the approaches, and most importantly, to inform the design of a better approach, tool, and instruction for collecting annotated activity data via the mobile crowd. It remains to be our belief that readers will gain the most benefit if we offer concrete and practical suggestions regarding how to make the tool, approach, and instruction more effective and efficient so that they can collect high quantity and quality annotated activity data easily. As a result, instead of concluding a winner among the compared approaches, we instead sought to analyze the pros and cons of different approaches, with a hope to create a better tool and approach based on the lessons learned. By analyzing users' recording and annotation behavioral patterns as well as the impact of activity on these patterns, we are able to inform readers about what to anticipate on the characteristics of collected activity data so that a better decision as to how to deal with them can be made. Finally, it is important to understand users' experience and challenges in using each of the approaches so as to know how to motivate users and sustain their willingness to contribute their activity data over time. The current paper is the first research work attempting to and being able to offer a set of design and methodological suggestions on making mobile crowdsourcing not only a viable but also an appealing approach to collecting diverse and realistic personal annotated activity data. By offering these suggestions, we hope to contribute to both mobile crowdsourcing and the development of context-aware systems, as well as any research fields needing to collect diverse annotated behavioral data. However, as mentioned in the limitation section, it would be important to examine the effectiveness of the approaches in the context of collecting other activity data. As a next step of this research work, we have started implementing the proposed hybrid approach, and evaluating it with the Participatory and the Context-Triggered approaches in different activity contexts.

Acknowledgement

This work was funded in part by NSF award IIS-1149601 and the University of Michigan M-Cubed program, Project ID: 349.

References

Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P., 1999. Towards a better understanding of context and context-awareness. In: Handheld and Ubiquitous Computing. Springer, Berlin Heidelberg, 304–307.

Agapie, E., Teevan, J., Monroy-Hernández, A., 2015. Crowdsourcing in the Field: A Case Study Using Local Crowds for Event Reporting, In: Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing.

Andrienko, G., Andrienko, N., Mladenov, M., Mock, M., Pölitz, C., 2010. Discovering bits of place histories from people's activity traces, In: 2010 IEEE Symposium on Visual Analytics Science and Technology (VAST). Presented at the 2010 IEEE Symposium on Visual Analytics Science and Technology (VAST), IEEE, pp. 59–66. 〈http://dx.doi.org/10.1109/VAST.2010.5652478〉.

Ashbrook, D., Starner, T., 2003. Using GPS to learn significant locations and predict movement across multiple users. Pers. Ubiquitous Comput. 7, 275–286.

Auld, J., Williams, C., Mohammadian, A., Nelson, P., 2009. An automated GPS-based prompted recall survey with learning algorithms. Transp. Lett. 1, 59–79.

Baltrunas, L., Ludwig, B., Peer, S., Ricci, F., 2011. Context-aware places of interest recommendations for mobile users. Des. User Exp. Usability Theory Methods Tools Pract. pp. 531–540

Bao, L., Intille, S.S., 2004. Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (Eds.), Pervasive Computing. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–77.

Bhattacharya, T., Kulik, L., Bailey, J., 2012. Extracting significant places from mobile user GPS trajectories: a bearing change based approach, In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '12. ACM, New York, NY, USA, pp. 398–401. 〈http://dx.doi.org/10.1145/2424321.2424374〉.

Cao, X., Cong, G., Jensen, C.S., 2010. Mining significant semantic locations from GPS data. Proc. VLDB Endow. 3, 1009–1020.

Chang, Y.-J., Hung, P.-Y., Newman, M., 2012. TraceViz: brushing for location based services, In: Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services Companion, MobileHCI '12. ACM, New York, NY, USA, pp. 219–220. 〈http://dx.doi.org/10.1145/2371664.2371717〉.

Chang, Y.-J., Paruthi, G., Newman, M.W., 2015. A field study comparing approaches to collecting annotated activity data in real-world settings, In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, pp. 671–682.

Chang, Y.-J., Tang, J.C., 2015. In: vestigating Mobile Users' Ringer Mode Usage and Attentiveness and Responsiveness to Communication, In: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI '15. ACM, New York, NY, USA, pp. 6–15. 〈http://dx.doi.org/10.1145/2785830.2785852〉.

Chen, G., Kotz, D., et al., 2000. A survey of context-aware mobile computing research. Tech. Rep. TR2, 000–381, (Dept. of Computer Science, Dartmouth College).

Cleland, I., Han, M., Nugent, C., Lee, H., Zhang, S., McClean, S., Lee, S., 2013. Mobile Based Prompted Labeling of Large Scale Activity Data. In: Nugent, C., Coronato, A., Bravo, J., (Eds.), Ambient Assisted Living and Active Aging, Lecture Notes in Computer Science. Springer International Publishing, pp. 9–17.

Cleland, I., Han, M., Nugent, C., Lee, H., McClean, S., Zhang, S., Lee, S., 2014. Evaluation of prompted annotation of activity data recorded from a smart phone. Sensors 14, 15861–15879.

De Cristofaro, E., Soriente, C., 2011. Short paper: PEPSI—privacy-enhanced participatory sensing infrastructure, In: Proceedings of the Fourth ACM Conference on Wireless Network Security. pp. 23–28.

DeVaul, R., Dunn, S., 2001. Real-Time Motion Classification for Wearable Computing Applications.

Dey, A.K., Abowd, G.D., Salber, D., 2001. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Hum.-Comput Inter. 16, 97–166.

Dingler, T., Pielot, M., 2015. I'll be there for you: quantifying attentiveness towards mobile messaging, In: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services. ACM, pp. 1–5.

Dockray, S., Grant, N., Stone, A.A., Kahneman, D., Wardle, J., Steptoe, A., 2010. A Comparison of affect ratings obtained with ecological momentary assessment and the day reconstruction method. Soc. Indic. Res. 99, 269–283. http://dx.doi.org/10.1007/s11205-010-9578-7.

Doherty, A., Kelly, P., Foster, C., 2013. Wearable cameras: identifying healthy transportation choices. IEEE Pervasive Comput 12, 44–47. http://dx.doi.org/10.1109/MPRV.2013.21.

Dumais, S., Jeffries, R., Russell, D.M., Tang, D., Teevan, J., 2014. Understanding user behavior through log data and analysis, In: Olson, J.S., Kellogg, W.A. (Eds.), Ways of Knowing in HCI. Springer New York, pp. 349–372.

Fischer, J.E., Greenhalgh, C., Benford, S.,, 2011. Vestigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. pp. 181–190.

Froehlich, J., Chen, M.Y., Consolvo, S., Harrison, B., Landay, J.A., 2007. MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In: Proceedings of the 5th International Conference on Mobile Systems, Applications and Services. pp. 57–70.

Froehlich, J., Dillahunt, T., Klasnja, P., Mankoff, J., Consolvo, S., Harrison, B., Landay, J. A., 2009. UbiGreen: investigating a mobile tool for tracking and supporting green transportation habits, In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI '09. ACM, New York, NY, USA, pp. 1043–1052. 〈http://dx.doi.org/10.1145/1518701.1518861〉.

Ganti, R.K., Ye, F., Lei, H., 2011. Mob. crowdsensing: Curr. State Future Chall. Commun. Mag. IEEE 49, 32–39.

Ganti, R.K., Pham, N., Tsai, Y.-E., Abdelzaher, T.F., 2008. PoolView: stream privacy for grassroots participatory sensing, In: Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems. pp. 281–294.

Goncalves, J., Hosio, S., Ferreira, D., Kostakos, V., 2014. Game of Words: Tagging Places Through Crowdsourcing on Public Displays, In: Proceedings of the 2014 Conference on Designing Interactive Systems, DIS '14. ACM, New York, NY, USA. pp. 705–714. 〈http://dx.doi.org/10.1145/2598510.2598514〉.

Harada, S., Lester, J., Patel, K., Saponas, T.S., Fogarty, J., Landay, J.A., Wobbrock, J.O., 2008. VoiceLabel: using speech to label mobile sensor data, In: Proceedings of the 10th International Conference on Multimodal Interfaces, ICMI '08. ACM, New York, NY, USA, pp. 69–76. 〈http://dx.doi.org/10.1145/1452392.1452407〉.

Heimerl, K., Gawalt, B., Chen, K., Parikh, T., Hartmann, B., 2012. CommunitySourcing: engaging local crowds to perform expert work via physical kiosks, In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 1539–1548.

Hosio, S., Goncalves, J., Lehdonvirta, V., Ferreira, D., Kostakos, V., 2014. Situated Crowdsourcing Using a Market Model, In: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, UIST '14. ACM, New York, NY, USA, pp. 55–64. 〈http://dx.doi.org/10.1145/2642918.2647362〉.

Huang, K.L., Kanhere, S.S., Hu, W., 2010. Are you contributing trustworthy data?: the case for a reputation system in participatory sensing, In: Proceedings of the 13th ACM International Conference on Modeling, Analysis, and Simulation of Wireless and Mobile Systems. pp. 14–22.

Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., Stone, A.A., 2004. A survey method for characterizing daily life experience: the day reconstruction method. Science 306, 1776–1780.

Kanhere, S.S., 2011. Participatory Sensing: Crowdsourcing Data from Mobile Smartphones in Urban Spaces, In: Mobile Data Management (MDM), 2011, 12th IEEE International Conferenceon. pp. 3–6.

Kelly, P., Doherty, A., Mizdrak, A., Marshall, S., Kerr, J., Legge, A., Godbole, S., Badland, H., Oliver, M., Foster, C., 2014. High group level validity but high random error of a self-report travel diary, as assessed by wearable cameras. J. Transp. Health 1, 190–201. http://dx.doi.org/10.1016/j.jth.2014.04.003.

Khan, W.Z., Xiang, Y., Aalsalem, M.Y., Arshad, Q., 2013. Mobile phone sensing systems: a survey. IEEE Commun. Surv. Tutor 15, 402–427, 〈http://dx.doi.org/1.1109/SURV.2012.031412.00077〉.

Kim, J., Kikuchi, H., Yamamoto, Y., 2013. Systematic comparison between ecological momentary assessment and day reconstruction method for fatigue and mood states in healthy adults. Br. J. Health Psychol. 18, 155–167. http://dx.doi.org/10.1111/bjhp.12000.

Klumb, P.L., Baltes, M.M., 1999. Validity of retrospective time-use reports in old age. Appl. Cogn. Psychol. 13, 527–539.

Konomi, S. 'ichi, Sasao, T., 2015.The use of colocation and flow networks in mobile crowdsourcing, In: Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers, UbiComp/ISWC'15 Adjunct. ACM, New York, NY, USA, pp. 1343–1348. 〈http://dx.doi.org/10.1145/2800835.2800967〉

Kwapisz, J.R., Weiss, G.M., Moore, S.A., 2011. Act. Recognit. Using Cell phone accelerometers. ACM SigKDD Explor. Newsl. 12, 74–82.

Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T., 2010. A survey of mobile phone sensing. Commun. Mag. IEEE 48, 140–150.

Lane, N.D., Xu, Y., Lu, H., Hu, S., Choudhury, T., Campbell, A.T., Zhao, F., 2011. Enabling large-scale human activity inference on smartphones using community similarity networks (csn), In: Proceedings of the 13th International Conference on Ubiquitous Computing. ACM. pp. 355–364.

Liao, L., Fox, D., Kautz, H., 2007. Extracting places and activities from gps traces using hierarchical conditional random fields. Int. J. Robot. Res. 26, 119–134.

Linnap, M., Rice, A., 2014. Managed Participatory Sensing with YouSense. J. Urban Technol. 21, 9–26. http://dx.doi.org/10.1080/10630732.2014.888216.

MacIntyre, B., Gandy, M., Dow, S., Bolter, J.D., 2004. DART: a toolkit for rapid design exploration of augmented reality experiences, In: Proceedings UIST 2010, UIST '04. pp. 197–206. 〈http://dx.doi.org/10.1145/1029632.1029669〉.

Meschtscherjakov, A., Reitberger, W., Tscheligi, M., 2010. MAESTRO: orchestrating user behavior driven and context triggered experience sampling, In: Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research. p. 29.

Nardi, B.A., 1996. Context and consciousness: activity theory and human-computer interaction. The MIT Press.

Nazneen, N., Rozga, A., Romero, M., Findley, A.J., Call, N.A., Abowd, G.D., Arriaga, R.I., 2012. Supporting parents for in-home capture of problem behaviors of children with developmental disabilities. Pers. Ubiquitous Comput. 16, 193–207.

Newman, M.W., Ackerman, M.S., Kim, J., Prakash, A., Hong, Z., Mandel, J., Dong, T., 2010. Bringing the field into the lab, In: Proceedings of the Proc. UIST 2010. Presented at the the 23nd annual ACM symposium, New York, New York, USA, p. 105. ⟨http://dx.doi.org/10.1145/1866029.1866048⟩.

Nowak, S., Rüger, S., 2010. How Reliable Are Annotations via Crowdsourcing: A Study About Inter-annotator Agreement for Multi-label Image Annotation, In: Proceedings of the International Conference on Multimedia Information Retrieval, MIR '10. ACM, New York, NY, USA, pp. 557–566. ⟨http://dx.doi.org/10.1145/1743384.1743478⟩

Okoshi, T., Ramos, J., Nozaki, H., Nakazawa, J., Dey, A.K., Tokuda, H., 2015. Reducing Users' Perceived Mental Effort Due to Interruptive Notifications in Multi-device Mobile Environments, In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15. ACM, New York, NY, USA, pp. 475–486. ⟨http://dx.doi.org/10.1145/2750858.2807517⟩.

Paxton, M., Benford, S., 2009. Experiences of participatory sensing in the wild, In: Proceedings of the 11th International Conference on Ubiquitous Computing. ACM, pp. 265–274.

Pejovic, V., Musolesi, M., 2014. InterruptMe: designing intelligent prompting mechanisms for pervasive applications, In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, pp. 897–908.

Pielot, M., Church, K., de Oliveira, R., 2014a. An In-Situ Study of Mobile Phone Notifications, In: Proc. MobileHCI.

Pielot, M., de Oliveira, R., Kwak, H., Oliver, N., 2014b. Didn't you see my message? Predicting attentiveness to mobile instant messages, In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems. ACM, pp. 3319–3328.

Pielot, M., Dingler, T., San Pedro, J., Oliver, N., 2015. When attention is not scarce-detecting boredom from mobile phone usage, in: Proc. of UbiComp

Pielot, M., 2014. Large-scale evaluation of call-availability prediction, In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, pp. 933–937.

Poppinga, B., Heuten, W., Boll, S., 2014. Sensor-based identification of opportune moments for triggering notifications. IEEE Pervasive Comput 13, 22–29, ⟨http://dx.doi.org/1.1109/MPRV.2014.15⟩.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M., 2010. Using mobile phones to determine transportation modes. ACM Trans. Sens. Netw. TOSN 6, 13.

Reddy, S., Burke, J., Estrin, D., Hansen, M., Srivastava, M., 2007. A framework for data quality and feedback in participatory sensing, In: Proceedings of the 5th International Conference on Embedded Networked Sensor Systems. pp. 417–418.

Rosenthal, S., Dey, A.K., Veloso, M., 2011. Using decision-theoretic experience sampling to build personalized mobile phone interruption models. In: Lyons, K., Hightower, J., Huang, E.M. (Eds.), Pervasive Computing, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 170–187.

Sahami Shirazi, A., Henze, N., Dingler, T., Pielot, M., Weber, D., Schmidt, A., 2014. Large-scale Assessment of Mobile Notifications, In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14. ACM, New York, NY, USA, pp. 3055–3064. ⟨http://dx.doi.org/10.1145/2556288.2557189⟩.

Sakamura, M., Yonezawa, T., Nakazawa, J., Takashio, K., Tokuda, H., 2014. Help Me!: Valuing and visualizing participatory sensing tasks with physical sensors, In: Proceedings of the 2014 International Workshop on Web Intelligence and Smart Sensing, IWWISS '14. ACM, New York, NY, USA, pp. 3:1–3:6. ⟨http://dx.doi.org/10.1145/2637064.2637095⟩.

Sarker, H., Sharmin, M., Ali, A.A., Rahman, M.M., Bari, R., Hossain, S.M., Kumar, S., 2014. Assessing the availability of users to engage in just-in-time intervention in the natural environment, In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14. ACM, New York, NY, USA, pp. 909–920. ⟨http://dx.doi.org/10.1145/2632048.2636082⟩.

Sheppard, S.A., Wiggins, A., Terveen, L., 2014. Capturing quality: retaining provenance for curated volunteer monitoring data, In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM, pp. 1234–1245.

Silvertown, J., 2009. A new dawn for citizen science. Trends Ecol. Evol. 24, 467–471.

Sonnenberg, B., Riediger, M., Wrzus, C., Wagner, G.G., 2012. Measuring time use in surveys – Concordance of survey and experience sampling measures. Soc. Sci. Res. 41, 1037–1052. http://dx.doi.org/10.1016/j.ssresearch.2012.03.013.

Turner, L.D., Allen, S.M., Whitaker, R.M., 2015. Interruptibility Prediction for Ubiquitous Systems: Conventions and New Directions from a Growing Field.

Vondrick, C., Patterson, D., Ramanan, D., 2012. Efficiently scaling up crowdsourced video annotation. Int. J. Comput. Vis. 101, 184–204. http://dx.doi.org/10.1007/s11263-012-0564-1.