Using Context-Aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices

Joyce Ho and Stephen S. Intille
Massachusetts Institute of Technology
77 Massachusetts Avenue (NE18-4FL)
Cambridge, MA 02139 USA
intille@mit.edu

ABSTRACT

The potential for sensor-enabled mobile devices to proactively present information when and where users need it ranks among the greatest promises of ubiquitous computing. Unfortunately, mobile phones, PDAs, and other computing devices that compete for the user's attention can contribute to interruption irritability and feelings of information overload. Designers of mobile computing interfaces, therefore, require strategies for minimizing the perceived interruption burden of proactively delivered messages. In this work, a context-aware mobile computing device was developed that automatically detects postural and ambulatory activity transitions in real time using wireless accelerometers. This device was used to experimentally measure the receptivity to interruptions delivered at activity transitions relative to those delivered at random times. Messages delivered at activity transitions were found to be better received, thereby suggesting a viable strategy for context-aware message delivery in sensor-enabled mobile computing devices.

Author Keywords

Interruption, context-aware computing, human-computer interface, mobile computing

ACM Classification Keywords

H5.m Information interfaces and presentation (e.g. HCI): Miscellaneous.

INTRODUCTION

Mobile computing devices present a conundrum for the interface designer. On one hand, current mobile computing devices such as phones and PDAs are contributing to feelings of information overload and to what might be called "interruption irritability." On the other hand, new sensor-enabled mobile devices will permit the mobile developer to create innovative applications that proactively deliver information to people when and where they need it. Many of the new applications that have the potential to add the most value to mobile computing devices are also those that will require proactive prompting of the user, yet proactive prompting could contribute to feelings of interruption and annoyance.

Mobile computing devices will increasingly deliver phone calls, reminders, email, task lists, instant messages, news, and other time and/or place-based information. They will also soon run context-aware applications such as location or activity based friend-finders, activity-triggered to-do reminders, and imaginative new phone services. The algorithms that determine when and what information to present will not make flawless decisions. Delivering the proactive interruptions the user wants, therefore, will inevitably increase the number of unwanted interruptions the user must endure.

Each time a device proactively provides information, it is competing for the user's attention and possibly interrupting the ongoing tasks. As others have observed, although computing power will continue to improve, permitting more powerful mobile devices, human attention is a limited resource [7]. Determining a good time to interrupt requires a complex assessment of context and message content. For example, consider an office worker sitting at a desk discussing a report with a supervisor. If the phone rings and it is a co-worker with updated information for the report, the office worker is likely to be receptive to the phone call. However, if the phone call is from from a friend to discuss plans for the weekend, then the office worker is likely to be less receptive. On the other hand, the office worker might be receptive to the phone call from the friend if the phone displays the message visually instead of using the ring to signal the interruption. The visual notification is less likely to disrupt the flow of the current conversation, perhaps lowering the perceived burden of the interruption for both people in the room.

There are at least 11 factors that impact the perceived burden of an interruption, as listed in Table 1. Developing a system that weighs these complex factors would require activity and discourse recognition systems that are well beyond the state of the art. Fortunately, recent work suggests that a small set of sensors gathering key information for a particular domain such as the office may provide useful information about interruptibility [17] and receptiveness to context-aware information [19]. Applications that can infer interruptibility from sensors can defer non-time-critical prompts to the times that are likely to be least disruptive.

Factor	Description of the Factor	Prior work
Activity of the user	The activity the user was engaged in during the interruption	[8, 9, 4, 22, 2]
Utility of message	The importance of the message to the user	[4, 31]
Emotional state of the user	The mindset of the user, the time of disruption, and the relationship the user has with the interrupting interface or device	[32, 16, 13, 21]
Modality of interruption	The medium of delivery, or choice of interface	[32, 1, 27, 14]
Frequency of interruption	The rate at which interruptions are occurring	[27]
Task efficiency rate	The time it takes to comprehend the interruption task and the expected length of the task	[9, 27, 31]
Authority level	The perceived control a user has over the interface or device	[16, 29]
Previous and future activities	The tasks the user was previously involved in and might engage in during the future	[12]
Social engagement of the user	The user's role in the current activity	[18, 15]
Social expectation of group behavior	Activities and expected reaction to interruption of nearby people	[18]
History and likelihood of response	The type of pattern the user follows when an interruption occurs	[24, 26]

Table 1. Eleven factors that influence a person's interruptibility at a given moment.

One strategy that has been suggested to minimize the perceived burden of an interruption is to present reminders immediately following the completion of some action [25]. The assumption is that at activity transitions, memory load may be low, because a person may be between evaluation of the last activity and formation of a new goal. In this work, we study the possibility that prompts from mobile devices may be perceived as less disruptive if they are presented at times when the user is transitioning between different *physical* activities. Our experiment is motivated by the casual observation that a transition between two different physical activities may strongly correlate with a task switch, and a task switch may be a better time to prompt the user with an interruption than an otherwise random time. When switching physical activities, the user has often just completed one task and is moving onto the next, or the user is "self interrupting" the current task to embark on a new activity. This may lower the user's resistance to an interruption from a mobile computing device that presents new information.

Not every physical activity transition will mark a good time to interrupt. For instance, it is possible that prompts at physical activity transitions could disrupt task planning and cause a person to lose track of what he or she was planning to do. Or, it is possible that by the time a person has changed to a new physical activity, he or she is already deeply immersed in a new task and not receptive to a new prompt. Finally, it is possible that it is more difficult to respond to prompts at the same time one is changing physical activity than at other times. Nonetheless, it seems plausible that, on average, prompts that coincide with activity changes may be perceived as less disruptive than prompts presented at random times.

In the remainder of the paper we describe a device we built to measure simple activity transitions and an experiment we conducted to compare user receptivity to messages delivered by mobile computing devices at activity transitions with those presented at other times.

RELATED WORK
The majority of prior work on interruption has focused on desktop computing applications and office environments rather than mobile computing.

Modeling Interruption and defining interruptibility
An interruption is an event that breaks the user's attention on the current task to focus on the interruption temporarily [27]. In an office environment, interruptions range from e-mails to impromptu meetings in the hallway. Interruptions are not always disruptive; some are beneficial to the user. When a person takes a coffee break or uses the restroom, it is a self-initiated interruption from current work that can help the person refocus on the task at hand.

The 11 factors in Table 1 all contribute to a user's evaluation of perceived burden of an interruption. An exhaustive model of interruptibility would include a weighted sum of these factors, where the context is detected automatically by appropriate sensors. No work to date has considered more than a few factors simultaneously because of the limited sensing options available. Some factors are particularly challenging, such as predicting the user's future activity.

Prior work varies in the explicit or implicit definition of "interruptibility." Some of the metrics that have been used to evaluate interruptibility are listed in Table 2. We will measure interruptibility by asking the user to rate *receptiveness* to receiving a particular type of message at a particular time. We expect that a drop in receptiveness will correspond to an increase in perceived burden of the interruption, but in pilot testing we found users could more easily understand the question when framed in terms of receptiveness. The *per-*

ceived burden of the interruption is not equivalent to the *actual* disruptiveness of the interruption. An interruption welcomed by the user may still negatively impact overall productivity, but in this work we focus on user perception at the time of the interruption itself.

Detecting interruptibility with sensors

Fortunately, recent prior work suggests that a few key sensors placed in an office environment may provide enough information to improve interruptibility detection without explicit detection of each of the 11 factors in Table 1 [17]. Hudson *et al.* studied an office environment and proposed an interruptibility model incorporating the activity of the user, the emotional state of the user and the social engagement of the user. They argue that these factors can be tracked using an audio sensor, the time of the day, and monitors for telephone, keyboard, and mouse usage [17]. These factors were sufficient to determine interruptibility with an accuracy of 75-80% when using simulated sensors. A followup study with functioning sensors also achieved good results [6].

Another recent study used the social engagement and activity of the user, as well as social expectation of group behavior, to build a model of interruption based upon switching among desktop computing applications. This study showed that the cost of an interruption for a user can be determined with a 73% accuracy using a computerized meeting scheduler, ambient acoustics in the office, visual analysis of the user's pose to obtain a model of attention, and activity on the computer desktop [12].

These studies suggest a small set of sensors may provide valuable information about interruptibility in the office setting, but mobile computing applications offer additional challenges. Mobile devices cannot rely as heavily on keyboard and mouse monitoring or text analysis to determine user activity. Mobile computing applications also must use sensors worn or carried by the user, who is likely to roam throughout the home, workplace, and community.

Detecting interruptibility in mobile applications

Hinckley and Horvitz modeled interruptibility by considering the user's likelihood of response and the previous and current activity. Three sensors necessary to detect these factors were built in a mobile computing device: a two-axis linear accelerometer for tilt sensing, a capacitive touch sensor to detect if the user was holding the device, and an infrared proximity sensor that detected if the head was in close proximity to the device [10].

Kern *et al.* estimated a mobile computer user's personal interruptibility using the activity of the user, the social engagement of the user, and the social expectation of group behavior. The sensor network to determine these factors included a two-axis accelerometer attached to a user's right thigh to measure a user's activity, a microphone that detected auditory context for the social situation, and a wireless LAN access point to determine the user's location within the building as well as outdoors. It was found that this model

could determine interruptibility with 94.6% accuracy in a pilot study with the experimenter as the test subject [18].

Siewiorek *et al.* created a mobile phone application that adjusted the modality of the interruption (i.e. vibration, ringer) based on the activity of the user and the social expectation of group behavior (detected using light, accelerometers, and microphones) [26]. The device was tested using a Wizard of Oz protocol simulating sensor behavior.

Liu used inputs from an accelerometer, a heart rate monitor, and a pedometer to trigger interruptions from a mobile device at "non-stressful" moments to study the impact of an emotionally-friendly interface on interruptibility. Subjects in the study were most receptive to interruptions when the system was emotionally friendly and triggering off non-stressful activities [20].

In this work, we study the relationship between *physical activity transitions* and perceived burden of interruption in a study with 25 subjects who carried a fully-functional sensor-enabled mobile computing device outside of a laboratory setting, each for an entire workday.

MOTIVATION

Imagine a scenario where an office worker has been sitting for some period of time, presumably working. A task-list program running on her mobile computing device activates with the goal of providing a proactive reminder to bring home a book. Consider two message delivery approaches: (1) the reminder is delivered at a random time (e.g. 24 hours after it was initially entered) which occurs while the person is sitting still and presumably working, or (2) the reminder is not delivered immediately but the next time the person gets up and walks for a few seconds. Our approach is motivated by the assumption that – given no other information about the message content or the user's situation – the second delivery approach would be perceived as less disruptive.

The second option is not always the most appropriate strategy. For example, the person may get up from her seat and walk to the end of a long room to give a presentation, or the message could be a time-critical one that cannot be delayed, even for a few minutes. Further, it is possible that an interruption at the beginning of a new task could be more costly than when the user has been engaged in a task for some time [5]. We set out to explore whether the second approach will minimize disruption when other contextual information is not available.

An interruption that is placed at the end of a task will usually be less disruptive and annoying than an interruption placed during a user's task [2]. When physical transitions occur, mental transitions are also likely.[1] Therefore, when a physical activity transition occurs, the user may already be in the process of interrupting his/her current activity and consequently may be more likely to be receptive to an interruption.

---

[1]The converse is much less likely to be true.

Authors	Definition of Interruptibility	Measure of Interruptibility
Bailey et al. [2]	Waiting for an opportune moment to avoid disruption on the primary task	The amount of time necessary to complete the interruption task and the original task while maintaining accuracy
Horvitz et al. [12]	Cost of interruption based on the user's model of attention, such as high-focus solo activity	Willingness to pay to avoid the disruption
Hudson et al. [17]	Perceived burden of interruption	Self-reports of interruptibility (scale of 1-5)
McFarlane [24]	Cognitive limitations to work during an interruption	Completion time, performance accuracy, and number of task switches
Kern [18]	Value of the notification	Self-annotation of the value of a notification
McCrickard et al. [23]	Unwanted distraction to primary task	Accuracy
Speier et al. [27]	Ability to facilitate decision making	Performance on decisions
Hess et al. [9]	Cognitive activity disruption	Accuracy and reaction time

Table 2. Definitions of interruptibility and evaluation metrics used in some recent prior work.

Of the many types of messages that could be delivered – reminders, tasks, phone calls, instant messages, rest breaks, etc. – even those that are time or place specific could often be delayed for a short period of time. Therefore, we envision a system where the mobile device is actively scheduling interruptions to correspond to physical activity transitions and (potentially) reducing the perceived burden of their delivery. An alternative approach would be to use activity transitions to compute a priority for incoming messages, perhaps in combination with other information about the message content and user context. The priority score could be used by an application to perform negotiation-based coordination.

EXPERIMENTAL FRAMEWORK AND TECHNOLOGY

To test the validity of our assumption that interruptions paired with activity transitions will result in a lower perceived burden on the user than otherwise randomly presented interruptions, we created a context-aware computing device that can detect activity transitions in real-time. 25 subjects were asked to wear small, unencumbering accelerometers throughout an entire workday. We used a within-subjects design. Each subject received two types of randomly intermixed interruptions: activity-transition triggered interruptions and random interruptions. Participants rated their receptivity to either receiving a "reminder" or a "phone call" on a 1-5 scale.

Activity detection

We began by building a mobile system consisting of a PocketPC PDA (iPAQ), a special iPAQ sleeve with an embedded wireless receiver, and two wireless accelerometers. An activity classification algorithm that has shown good performance on activity recognition of household activities using multiple accelerometers [3] was implemented for the PDA. The accelerometers used in this work are small and lightweight, and each runs on a coin cell battery that provides a sufficient power for a working day [30]. The activity detection system was limited to accelerometers to test the system under non-laboratory conditions. This allowed the subjects complete freedom of motion without any encumbering or overtly privacy invasive sensors.

The accelerometers each send real-time, 3-axis, 0-10G motion information at 100Hz to the PDA. Software running on the PDA computes mean, energy, entropy, and correlation features on 256 wide sample windows of acceleration data, with 128 samples overlapping between consecutive windows. At the sampling frequency of 100 Hz per accelerometer, each window represents 1.28 seconds, thereby resulting in a responsive algorithm with only a small lag. Features computed on each sliding window are passed through a previously trained C4.5 decision tree supervised learning classifier for activity recognition. In this work, the C4.5 classifier was trained to detect only three postural and ambulation activities: sitting, standing, and walking. These three activities had particularly high recognition rates in prior work [3] and are therefore suitable to use to robustly detect changes in gross movement for interruptibility evaluation. For more detail on the feature calculation, activity detection algorithm, and classifier training, see [11].

We selected the following four transitions to target in this work: sitting to walking, walking to sitting, sitting to standing, and standing to sitting. To reduce spurious activity transition detection, temporal smoothing is used. The algorithm only registers a transition between two physical states when the activity prior to the transition is detected for 10 seconds and the activity after the transition is detected for 3 seconds. This definition introduces a 3 second lag (in addition to the 1.5 second lag due to feature computation) but filters "spurious" transitions such as when a person rapidly moves from sitting to walking through standing. Ideally, the technology would have no latency. However, the system runs on a relatively slow PDA computer. Despite the latency, the lab pilot testing suggested that the system generally feels responsive. One still feels as if he/she is still in the transition period when the activity-triggered prompts occur.

Interruption triggering

Software was developed for the PDA to interrupt the subject once every 10-20 minutes throughout the day, either randomly or at an activity transition. The software ensures that each subject is presented with approximately the same number (plus or minus 2) of each interruption type throughout a day. The total number of interruptions received in a day is determined by the frequency of activity transitions detected

Figure 1. Screen shots of the two message prompts.

Figure 2. Reminder attached to the back of the iPAQ.

(i.e. a person who is sedentary receives fewer random interruptions to keep the interruption types in balance).

At each interruption trigger, soft chimes sound and gradually increase in volume for 30 seconds. If the user does not respond, the chimes change to a beep for another 30 seconds that gradually increases to the maximum volume. The software randomly choses one of the following two questions to display on the screen: "How receptive are you to a phone call?" or "How receptive are you to a reminder?" as shown in Figure 1. A subject answers by clicking the large rating button and then "ok" with a finger. No stylus was provided. At that point the screen is turned off until the next interruption. The device can be muted for up to 1 hour by turning it on at any time and clicking a "Mute" button.

In some prior work, subjects have been asked about interruptibility independent of content. However, in our pilot work, people we interviewed found this difficult to do. Pilot testers preferred to rate receptiveness or interruptibility in the context of a particular type of message that they were familiar with. We therefore selected two familiar message types to use in this study. Prior to the start of the experiment, the two notification types were explained. Subjects were told that the "reminder" is a reminder of something on a to-do list, but that it is not time critical (e.g. it does not include a reminder to attend a meeting in 5 minutes). Subjects were told that the phone call is a standard call but that there is no caller ID and so there is never a way to know who the caller is.[2] Subjects were asked to answer the receptivity question using a scale of 1-5, with 1 being "not at all receptive" and 5 being "extremely receptive." The reminder sheet pictured in Figure 2 was taped to the back of the PDA and used to explain the rating scale to subjects.

As the PDA chimes or beeps, a tap on the screen turns off the sound. If the subject does not answer the question within one minute, the answer is logged as "no response." There are two ways to interpret the no response (NR) in the data. An NR can occur when a subject is not receptive but able to

---

[2]An incoming phone call is a time-dependent message. We selected it because it was familiar to pilot testers, not because we expect that it will be the type of message that would most likely be time shifted in emerging mobile applications.
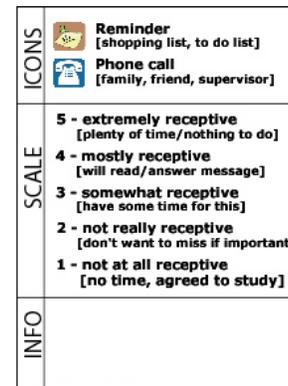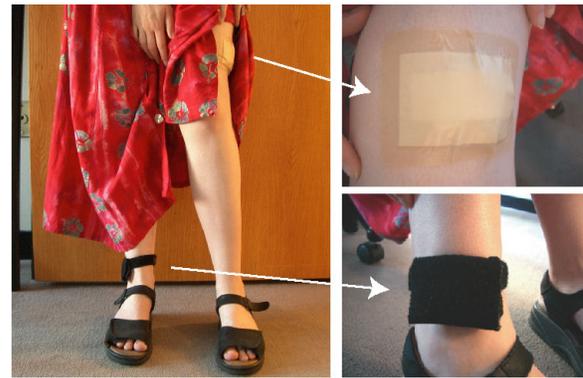

Figure 3. Placement of the wireless accelerometers.

silence the device. In this case an NR is equivalent to a "not at all receptive." Alternatively, a "no response" can result from situations where the screen is accidentally tapped just prior to hearing the sound or where the user has put down the PDA and is out of earshot of the prompt.

Setup procedure

Each subject was asked to wear two wireless accelerometers. With the help of a research assistant, one was attached to the outside of the right ankle using a small Velcro pouch and the other was attached to the outside of the left thigh just above the knee using an adhesive bandage. The placement of the sensors is shown in Figure 3. These locations were picked based on prior work suggesting that these positions were effective for posture discrimination and walking detection. Researchers are actively working on developing algorithms for detecting activities from accelerometers located in more convenient locations, such as in watches and shoes. The locations used here were selected only to provide maximal robustness to test the paper's main hypothesis, not to suggest these are the sensor locations that would ultimately be required.

The accelerometers are small and lightweight, roughly the size of a quarter [30]. Figure 4 shows the accelerometers and a special housing created for the iPAQ that holds the wireless accelerometer receiver and connects to the iPAQ serial port.

Figure 4: The 3-axis wireless accelerometers (with a U.S. quarter) and the iPAQ with the receiver casing.

Subjects were asked to carry the iPAQ with them at all times, either in a small pouch that attached to a belt or in a small travel bag that was provided by the researchers.

Subjects were given the iPAQ and the wireless accelerometers at the beginning of their workday and instructed on how to wear them. They were also told to answer each question based only on the particular situation at the time of the beep and asked not to consider any previous questions. Subjects were also asked to maintain their normal work schedule. At the end of the day, a 30-minute exploratory wrap-up interview was conducted.

Subjects
The study protocol was approved by the investigator's review board on the use of human subjects. Subjects were recruited through posters placed in nearby office parks, stores, and public spaces. The posters contained the following text: "Carry a cell phone? Help Researchers learn how to design user-friendly mobile devices." E-mails were also sent with the same text to local mailing lists.

Twenty-five subjects (9 male, 16 female) participated in this study. Two additional potential subjects were dropped. One subject stopped the study because the device was found to be too disruptive; another subject did not push the "OK" button after responding to the questions, preventing the system from logging any responses. The subjects were between the ages of 19 and 36 (mean=25.6, SD=3.3). Subject occupations were as follows: Administrative Staff (3), Lab Researcher (5), Office Professional (12), Field Professional (4), and Customer Service (1). Each of the subjects owned a mobile phone, and none were affiliated with the authors' research group. The subjects carried the PDA for full work day (mean=8.42 hours, SD=1.3 hours). Each subject was compensated with a ten-dollar gift certificate.

Subjects were not told that they would experience two different types of interruptions but were told that we were interested in making mobile computing devices easier to use. Subjects experienced 16-48 interruptions (mean=28.8, SD=7.1) spaced out over the course of the day. The rates of "no re-

	False pos	False neg	Incorrect classification	Real-trans. accuracy	Classifier accuracy
Mean	3.1%	12.3%	5.7%	82.6%	91.2%
SD	7.9%	4.5%	3.6%	7.0%	8.7%

Table 3: Summary of the means and standard deviations (SD) for the activity transition detection algorithm evaluated on the 5 subjects not used to train the classifier.

sponse" answers ranged from 0-28% (mean = 9.4%, SD = 7.8%).

RESULTS
First we present results of our evaluation of the real-time activity transition detection algorithm. Next we present the findings from the interruption study itself.

Verification of transition detection algorithm
The validity of the interruption experiment is dependent upon the quality of the real-time activity transition detection algorithm. Ten colleagues who were not subjects in the interruption study were used to train the classifier. These volunteers were asked to wear the sensors and to repeatedly perform the three target activities.

After the classifiers were trained, five people involved in the training plus an additional five who did not wear the sensors in a validation phase were used to evaluate activity detection performance. Two iPAQ PDAs were calibrated to have identical clock times. One ran the activity detection algorithm, and the other ran an application that allowed the user to mark his/her activity by choosing one of the three target activities. 393 known physical activity transitions were recorded during this testing phase, against which the algorithm was evaluated.

It was difficult for subjects to indicate a transition precisely when it occurred. We therefore consider a classification valid if the difference between the self-annotated transition time and the activity transition detection algorithm time differs by no more than 10 seconds.

False positives are cases in which the algorithm detected a transition when one did not occur. False negatives are cases in which a physical transition occurred but the algorithm did not detect any type of transition. Incorrect classifications are cases when the algorithm detected the transition but detected the wrong type. Real-transition accuracy is the percentage of real transitions the classifier was able to detect correctly. For instance, if a subject transitioned from standing to sitting, this accuracy represents the probability that this particular transition was detected by the algorithm given that the transition actually occurred. The classifier-transition accuracy is the percentage of transitions the algorithm correctly classified. For example, if the algorithm detected a sitting to walking transition, the classifier-transition accuracy is the probability that the subject was actually transitioning from sitting to walking. In other words, given that the classifier detected a transition, this indicates the percentage of the time the transition was correct.

Table 3 summarizes the means and standard deviations of the algorithm's performance on the five subjects not used to train the classifier. The classifier did not perform consistently for all the subjects, as illustrated by the standard deviation. One subject fidgeted his leg for a period of time, leading to a high number of false positives. In addition, the subject acknowledged that he missed recording some of the transitions. He specified the time at which this occurred, and any transitions during this time period were not used in the evaluation. However, it is possible that a few unmarked transitions remained in the data, resulting in an artificially low classifier and real-transition accuracy. The algorithm was also tested on the 5 subjects used for acquiring training data. As expected, the algorithm has a higher detection accuracy for the trained subjects. However, the accuracy only drops by 8% on untrained subjects. In this work, the untrained subject results are used because the subjects in the interruption study did not contribute any training data.

The interruption experiment requires a low false positive and incorrect classification percentage. Therefore, the key measure is the 91.2% accuracy with respect to the classifier. It is important that when the algorithm detects a transition that the transition actually occurred and was correctly classified. The relatively high number of false negatives will not affect the interruption study protocol results because an interruption will not be triggered at that particular moment. The false negatives may, however, have resulted in fewer interruptions being delivered to subjects who did not regularly make activity transitions.

The possibility does exist that what is intended to be a random, rather than an activity-triggered, interruption could coincidentally occur during an activity transition. However, over the course of the day, it was assumed this situation would be exceedingly rare and would not have a significant effect on the overall receptivity results.

Interruption study

The number of interruptions experienced by subjects ranged from 16-48 (mean=28.8, SD=7.1). The data were aggregated on a subject level by calculating the mean for each subject's responses to each type of message for the entire day, and then using that mean to represent the overall receptivity of that particular subject. This is an approach common when using ecological momentary assessment (EMA) sampling [28]. A paired t-test was used to compare the mean of the subjects' means to evaluate if there was a difference in overall receptivity between the two cases of interest: random versus triggered interruptions.

A user's failure to respond (i.e. "no response" answers) was handled in two ways. The first method simply dropped "no responses" from the analysis, as is frequently done in EMA research [28]. This method is appropriate since it is unknown whether a subject failed to answer the question because s/he was unreceptive or because s/he did not hear the audio prompt. In the second method, a "no response" was treated as an "extremely unreceptive" response under the assumption that the subject was too busy to respond or to carry

the PDA and therefore not at all receptive to an interruption. It is possible, however, that the volunteer simply left the PDA behind and was not within earshot.

A standard t-test would not account for the uncertainty introduced from the classifier, which was measured to have 91.2% accuracy on the volunteers who did not train the algorithms (see Table 3). Therefore, we adjusted the raw data to account for the possibility that the detection algorithm misclassified activity changes. The accuracy adjustment was simulated by repeatedly randomly switching 9% of the activity transition responses to random interruption responses, thus simulating algorithm error. Each time, new means were computed for all subjects. Results from 25 iterations of the simulation were then averaged. The iteration minimizes the variation due to the removal of particular responses. The same procedure was followed for an estimated worst case scenario, by assuming the accuracy of the classifier to be only 82.4%, one standard deviation below the mean accuracy.

The two-tailed, paired t-tests used a confidence interval of 95% with the significance level defined at p = 0.05. Table 4 summarizes the results of the paired t-tests using the raw data (which assumes the activity detection is flawless) and the adjusted activity-detection algorithm performance rates of 91.2% and the assumed worst case, 82.4%.

Table 5 contains the mean and standard deviation of the number of triggered interruptions experienced by the subjects once the assumed classifier accuracy is factored in.

The results are strongly significant and show that subjects rated random responses as more disruptive than activity triggered responses. This trend holds regardless of how the "no responses" are treated. In addition, it holds even assuming the worst case scenario for the activity classifier algorithm. Furthermore, other analysis showed that the results are significant when computed for male subjects only, female subjects only, and either message type only.

DISCUSSION

The results support the strategy of using activity transitions as a trigger for non-time-critical interruptions to potentially reduce feelings of information overload. By delaying interruptions that are not time-sensitive until a physical activity transition, the mobile computing device may lower the perceived interruption burden of some of the messages. We have also shown that two 3-axis wireless accelerometers can reliably detect a user's activity transition in real time and be used to determine the activity transitions.

The wrap-up interviews were used to elicit more information about what factors were salient in subjects' decision making. Subjects were first asked to estimate the number of interruptions they experienced and whether they would recommend the study to a friend. The difference between actual and estimated interruptions ranged from underestimating by 14 and overestimating by 71 (mean=-1.8, SD=16). Nineteen of the subjects indicated they would recommend the study

	Accuracy	Signif.	Mean random score	St.Dev	Mean transition score	St.Dev
"NR" OMITTED						
All msg types	100%	<.001	2.83	.51	3.34	.57
All msg types	91.2%	<.001	2.92	.49	3.30	.52
All msg types	82.4%	<.001	2.94	.48	3.31	.52
"NR" INCLUDED						
All msg types	100%	<.001	2.57	.42	3.00	.51
All msg types	91.2%	<.001	2.70	.38	3.09	.49
All msg types	82.4%	<.001	2.74	.38	3.07	.49

Table 4: Summary of two-tailed, paired t-tests results (N=25) for classifiers assuming 100%, 91.15%, and 82.4% accuracy of the activity transition detector algorithm. Shown is significance, mean raw score for the random and activity-transition triggered conditions with standard deviations. Regardless of how "no responses" are treated, the results indicate signficant or highly significant trends that subjects were more receptive to prompts tied with activity transitions than those presented at a random time.

to a friend, 3 indicated they would possibly recommend the study, and 3 indicated they would not recommend the study.

The subjects were also asked about their impressions of each type of reminder. Nine of the 25 subjects favored a reminder while the remaining 16 preferred phone calls. Most subjects who favored the reminder estimated an average phone call to take at least 10 times as long as a reminder. Subjects commented that it was difficult to differentiate the two types of questions, the phone call and the reminder. They would have liked the system to highlight the difference using two different set of chimes. Additionally, five subjects stated that they did not use reminders and found it difficult to rate their response because they had no previous experience on which to base their receptivity. These comments illustrate the difficulty of assessing interruptibility in a general way.

Only two of the 25 subjects muted the study, each for a total of 1 hour. One of the first subjects wore headphones a significant portion of the day. The headphones prevented the subject from hearing the audio prompt, and the subject had to be notified by neighboring coworkers that the iPAQ was signaling an interruption. As a result, the subject answered "extremely unreceptive" to these interruptions because of the possible disruption to coworkers. During the wrap-up interview, the subject stated that the disruption experienced by the coworkers did not change his receptivity rating because he was preoccupied at the moment anyway and would have made the same response. Furthermore, this subject did not skew the data toward favoring the activity transition triggered interruptions so his data was included in the final analysis. Future subjects were asked to avoid the use of headphones for the day, however.

The wrap-up interviews frequently indicated that the reasoning for choosing "not at all receptive" was that the subject was talking to his/her supervisor. Subjects were asked whether an interruption of a different type (maybe breaking news, an e-mail message, or a stress reduction exercise) or a different medium of delivery would make a difference. Some subjects responded positively to the suggestion of using vibration to notify the user of the interruption, but they acknowledged that they still would be unable to respond to the interruption immediately.

Some of the five volunteers who worked as lab researchers complained that interruptions occurred while they were conducting experiments. Two subjects reported that they had to remove their gloves to answer the questions. A few lab researchers had considered not carrying the PDA because they were involved in work that required precise measurements and could not afford to be interrupted. Additionally, another lab researcher noted that s/he was interrupted more frequently during an appointment with a patient. The higher frequency resulted from the researcher's common behavior of walking to attend to a patient and then sitting down to perform tests multiple times during an appointment. This behavior could trigger an activity-change interruption and is an example of a situation where different activity transition strategies may be required.

An additional office professional commented that the PDA seemed to deliver prompts more frequently at inappropriate times. An example situation was when the subject was leading a board discussion with several coworkers and clients and was frequently interrupted. During this time, the subject would stand to write on a white-board but then sit back down to continue the discussion with the rest of the group. This too, is a situation that violates the assumption on which our hypothesis is based.

Several subjects commented that interruptions occurred while they were driving. Two of the subjects stated that this was actually a good time for an interruption because they were "just driving," but other subjects considered this a distraction at an inappropriate time. These reminders were most likely randomly triggered, because the activity-transition algorithm would not generally detect activity changes during sedentary driving activity.

Some subjects indicated that even though they were provided with carrying cases for the PDA, sometimes they would forget to bring the device with them, leading to "no response" answers to randomly triggered prompts (activity triggered prompts do not occur if the accelerometers are moved more than 20 feet from the PDA).

When subjects were informed the nature of the study, 5 subjects noted that the algorithm should consider monitoring their computer since there were periods during the day when

	Trigger 91%	Trigger 82%	Phone 91%	Reminder 82%	Phone 91%	Reminder 81%
Mean	12.6	11.4	6.8	5.8	6.3	5.1
SD	3.6	3.3	2.9	2.4	2.7	2.4

Table 5: The means and standard deviations (SD) for the number of triggered responses experienced by the subjects broken down by two different classifier accuracy values.

they had nothing to do and were surfing the Internet. They described these moments as times when they would be extremely receptive to any interruption since it would keep them occupied.

An issue to consider in future work is whether there are especially poor times to interrupt users. If so, these times may lower the mean receptiveness of the random condition our tests. We have not investigated if the physical activity transitions are in some way dependent upon poor interruption times. To exclude these times would require that a detector be built that can recognize contextual cues for especially poor receptiveness.

The qualitative interviews offered further support that the algorithm was operating as intended throughout the study. Overall, our results show that our volunteers were significantly (in the statistical sense) more receptive to messages delivered at activity transitions than those delivered randomly. However, whether this difference is large enough to impact long-term opinions about a mobile computing device providing proactive messages remains an open question. After months of use, will a mean difference of approximately .5 on the scale from "extremely receptive" (5) to "not at all receptive" (1) lead to a change in the user's overall evaluation of a device, or will a few outliers – poorly timed, memorable prompts – dominate the user's impression of the device's performance? Could a larger library of activity transition types further reduce the user's burden and lead to average numbers that are 1-2 points higher overall, and would a longitudinal study show receptivity falloff rates and other novelty effects? We leave these questions for future research.

SUMMARY

A change in physical activity may sometimes correlate with a self-initiated task interruption. This study suggests that proactive messages delivered by a mobile computing when the user is transitioning between two physical activities (e.g. sitting to walking) may be received more positively than the same messages delivered at random times. The results suggest that the perceived burden of context-aware mobile computing devices may be minimized by time-shifting some proactive messages to moments when the user is already transitioning between different physical activities.

ACKNOWLEDGMENTS

REFERENCES

1. E. Arroyo, T. Selker, and A. Stouffs. Interruptions as multimodal outputs: which are the less disruptive? In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, pages 479–482. IEEE Press, 2002.

2. B.P. Bailey, J.A. Konstan, and J.V. Carlis. Measuring the effects of interruptions on task performance in the user interface. In *IEEE Conference on Systems, Man, and Cybernetics 2000 (SMC 2000)*, pages 757–762. IEEE Press, 2000.

3. L. Bao and S.S. Intille. Activity recognition from user-annotated acceleration data. In A. Ferscha and F. Mattern, editors, *Proceedings of PERVASIVE 2004*, volume LNCS 3001, pages 1–17. Springer-Verlag, Berlin Heidelberg, 2004.

4. E. Cutrell, M. Czerwinski, and E. Horvitz. Effects of instant messaging interruptions on computing tasks. In *Extended Abstracts of CHI 2000, Human Factors in Computing Systems*. ACM Press, 2000.

5. E. Cutrell, M. Czerwinski, and E. Horvitz. Notification, disruption and memory: Effects of messaging interruptions on memory and performance. In M. Hirose, editor, *Human-Computer Interaction–Interact '01*, pages 263–269. IOS Press, 2001.

6. J. Fogarty, S.E. Hudson, and J. Lai. Examining the robustness of sensor-based statistical models of human interruptibility. In *Proceedings of the 2004 Conference on Human Factors in Computing Systems*. ACM Press, New York, NY, 2004.

7. D. Garlan, D. Siewiorek, A. Smailagic, and P. Steenkiste. Project Aura: toward distraction-free pervasive computing. *IEEE Pervasive Computing*, April-June:22–31, 2002.

8. T. Gillie and D. Broadbent. What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological Research*, 50(4):243–50, 1989.

9. S.M. Hess and M. Detweiler. Training to reduce the disruptive effects of interruptions. In *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, volume 2, pages 1173–1177. Human Factors & Ergonomics Society, Santa Monica, CA, 1994.

10. K. Hinckley and E. Horvitz. Toward more sensitive mobile phones. In *Proceedings of the 14th annual ACM Symposium on User Interface Software and Technology*, pages 191–192. ACM Press, 2001.

11. J. Ho. *Interruptions: Using Activity Transitions to Trigger Proactive Messages*. M.Eng. Thesis, Massachusetts Institute of Technology, 2004.

12. E. Horvitz and J. Apacible. Learning and reasoning about interruption. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, pages 20–27. ACM Press, New York, NY, 2003.

13. E. Horvitz, A. Jacobs, and D. Hovel. Attention-sensitive alerting. In K.B. Laskey and H. Prade, editors, *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 305–513. Morgan Kaufmann, 1999.

14. E. Horvitz, C.M. Kadie, T. Paek, and D. Hovel. Models of attention in computing and communications: from principles to applications. *Communications of the ACM*, 46(3):52–59, 2003.

15. E. Horvitz, P. Koch, C. Kadie, and A. Jacobs. Coordinates: probabilistic forecasting of presence and availability. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI-02)*, pages 224–233. Morgan Kaufmann Publishers, 2002.

16. J.M. Hudson, J. Christensen, W.A. Kellogg, and T. Erickson. 'I'd be overwhelmed, but it's just one more thing to do:' availability and interruption in research management. In *Proceedings of the Confererence on Human Factors in Computing Systems*. ACM Press, 2002.

17. S.E. Hudson, J. Fogarty, C.G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J.C. Lee, and J. Yang. Predicting human interruptability with sensors: a Wizard of Oz feasibility study. In *Proceedings of the Conference on Human Factors and Computing*, pages 257–264. ACM Press, New York, NY, 2003.

18. N. Kern and B. Schiele. Context-aware notification for wearable computing. In *Proceedings of the 7th International Symposium on Wearable Computing*, pages 223–230. IEEE Press, 2003.

19. N. Kern, B. Schiele, and A. Schmidt. Multi-sensor activity context detection for wearable computing. In *European Symposium on Ambient Intelligence (EUSAI)*, volume LNCS 2875, pages 220–232. Springer-Verlag, Heidelberg, 2003.

20. K.K. Liu. *A Personal, Mobile System for Understanding Stress and Interruptions*. S.M. Thesis, Massachusetts Institute of Technology, 2004.

21. P. Lukowicz, H. Junker, M. Staeger, T. von Bueren, and G. Troester. WearNET: a distributed multi-sensor system for context aware wearables. In *Proceedings of the 4th International Conference on Ubiquitous Computing*, pages 361–370. Springer, 2002.

22. C.E. McCarthy and M.E. Pollack. A plan-based personalized cognitive orthotic. In M. Ghallab,

J. Hertzberg, and P. Traverso, editors, *Proceedings of the Sixth International Conference on Artificial Intelligence Planning Systems*, pages 243–252. AAAI Press, 2002.

23. D.S. McCrickard and C.M. Chewar. Attuning notification design to user goals and attention costs. *Communications of the ACM*, 45(3):67–72, 2003.

24. D.C. McFarlane. Coordinating the interruption of people in human-computer interaction. In A. Sasse and C. Johnson, editors, *Human-Computer Interaction - INTERACT'99*, pages 295–303. IOS Press, 1999.

25. Y. Miyata and D.A. Norman. Psychological issues in support of multiple activities. In D.A. Norman and S.W. Draper, editors, *User-Centered System Design*, pages 265–284. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

26. D.P. Siewiorek, A. Smailagic, J. Furukawa, A. Krause, N. Moraveji, K. Reiger, J. Shaffer, and F.L. Wong. SenSay: a context-aware mobile phone. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, pages 248–249. IEEE Press, 2003.

27. C. Speier, J.S. Valacich, and I. Vessey. The effects of task interruption and information presentation on individual decision making. In *Proceedings of the Eighteenth International Conference on Information Systems*, pages 21–36. Association for Information Systems, 1997.

28. A.A. Stone and S. Shiffman. Capturing momentary, self-report data: a proposal for reporting guidelines. *Annals of Behavioral Medicine*, 24(3):236–243, 2002.

29. N.A. Storch. Does the user interface make interruptions disruptive? A study of interface style and form of interruption. UCRL-JC-108993, Lawrence Livermore National Laboratory, 1992.

30. E.M. Tapia, N. Marmasse, S.S. Intille, and K. Larson. MITes: Wireless portable sensors for studying behavior. In *Proceedings of Extended Abstracts Ubicomp 2004: Ubiquitous Computing*. 2004.

31. M. van Dantzich, D. Robbins, E. Horvitz, and M. Czerwinski. Scope: providing awareness of multiple notifications at a glance. In *Proceedings of AVI 2002, ACM Conference on Advanced Visual Interfaces*. ACM Press, 2002.

32. R. van Solingen, E. Berghout, and F. van Latum. Interrupts: just a minute never is. *IEEE Software*, 15(5):97–103, 1998.