The Benefit of Activity Recognition for Mobile Phone Based Nursing Documentation: A Wizard-of-Oz Study

Bashar Altakouri, Gerd Kortuem
School of Computing and Communications
University of Lancaster, UK
{b.altakrouri,kortuem}@comp.lancs.ac.uk

Agnes Grünerbl, Kai Kunze, Paul Lukowicz
Embedded Systems Lab
University of Passau, Germany
{agnes.gruenerbl,paul.lukowicz}@uni-passau.de

Abstract

In a Wizard-of-Oz experiment we investigate to what degree automatic activity recognition could support the use of prioritized lists for nursing documentation on hand held mobile devices. The study involved 15 nurses, 60 patients records and over 250 documented processes at a geriatric care ward. Based on time effort, interaction complexity, error rate and subjective system perception our wizard of oz study shows that activity recognition is a key factor in the usability and acceptance of the system. We also study the impact of varying simulated activity recognition errors and demonstrate that error rates of up to 25% do not aversely affect the documentation process.

1 Introduction

Smart phones are often cited as a wearable computing success story. However they still differ from a "true" Wearable in one important way: a wearable system should be usable without interfering with the user's actions in the physical world. Thus, the user interface should minimize the cognitive load, it should not distract the user's attention from the real world and should not impair his physical freedom of action. Even though those requirements are not yet fulfilled, the latest generation of devices has moved closer to the wearable ideal as many phones can be operated with a single hand and can be used while walking or doing other basic tasks. This has led to a new type of device usage that we call "filling up the dead cycles". Instead of taking dedicated time for accessing information on a mobile device, time slots of "forced inactivity" or "forced low intensity activity" are utilized. Examples are walking from one room to another and waiting for the elevator, a print job or coffee.

In this paper we investigate if context and activity recognition can widen the scope of functionalities that can be accessed in the "dead cycles" mode. To this end we have first performed a series of work place observations and user interviews. It was followed by two Wizard-of-Oz experiments (the context recognition has been simulated), conducted in a real nursing ward with real nurses and real documentation work flows. The studies and experiments were designed to answer the following questions: *Can nursing documentation be done during "dead cycles" such as walking between patient rooms ? How much could automatic recognition of activities, which reduces the documentation to mere confirmation and correction of individual items, help this process ?* and *How much do recognition errors impact the utility of automatic recognition for the documentation process ?*.

Related Work. A number of studies have investigated the use and usefulness of context in wearable computing [8]. Most relevant, Kunze et al. [4] describe a quantitative Wizard-of-Oz study of using context recognition for a wearable maintenance. However, this work focused on real-time, context-aware help and task guidance, not documentation. Another similar work [6] points out the risks of using check list (when none was used before), as this can lead to missing key elements when they are not present in the list.

A broader overview of pervasive healthcare can be found in [7]. As far as the use of mobile technology to support nursing activities is concerned there is relatively little work. Focusing on documentation, Ammenwerth et al. [2] investigate factors affecting user acceptance. Hyun et al. [3] describe the development and clinical trial of a mobile nursing information system, and in [5] Kuwahara et al. propose a wearable auto-event-recording system of medical nursing. Adamer et al. [1] report the development and evaluation of a wearable support system for hospital ward rounds.

2 Background and Workplace Observation

The work described in this paper was done at the Mainkofen Psychiatric Hospital in Germany. The specific ward deals with elderly dementia patients with behavioral disorders and severe somatic diseases. During the morning

Table 1. Dead-Cycles

d-cycle activity	#days	av.sec/day	av.sec/activity	av.occur./day
go to patient. room	22	241.6	36.6	6.6
go to ward. office	17	110.5	30.8	3.6
go to bathroom	13	19.1	14.7	1.3
fill sink/bowl	21	63.9	28.6	2.2
bring sth. back	15	52.1	36.5	1.4

rounds a nurse attends to the assigned patients (4 to 5 patients, 30 min each), helps them get up, dress, and wash, and performs required medical tests and administers medications. Next the nurses help the patients with breakfast, or attend to assorted ward business. Around noon they document the morning activities using a shared PC.

In an initial study, a researcher shadowed nurses at one ward for a month, five times a week resulting in 25 annotated days, with a total of around 100 attended patients. The key observations are:

The Activities: In total 47 types of high level activities were identified which were further subdivided into 192 individual activities. Typical examples include morning examinations (measuring blood pressure, pulse, ...) and morning hygiene (dressing, washing, brushing teeth, ...).

Current Documentation process: Documentation is performed in three steps. Hand-written notes while performing the activity. Vital signs are noted during breaks which later entered as part of the official health record. At the end of the shift each nurse performs documentation in a computer-based health record system (20 to 40 minutes - not including time waiting for an available PC).

"Dead Cycles" and the use of Mobile Devices: The nurses spend a significant amount of time walking from one room to the next, waiting for sinks to fill up or for patients to finish tasks. Table 1 lists the most frequent types of "dead-cycles" with their typical length and number of observations. Since the hands of the nurse are mostly unoccupied during those dead cycles, they might be used to make entries into a hand-held or body-worn documentation system. Since individual "dead cycle" time slots are rather short and do not always allow full attention on a mobile device, an activity recognition system reducing the documentation to confirmation/correction of a recorded list appears helpful.

3 Experimental Study

We have conducted semi-structured interviews with 9 nurses to determine the nurses general attitude towards dead-cycles based documentation. All nurses stated that they would welcome such a system. Thus, from the observation study and the initial interviews we hypothesize that activity recognition based documentation is potentially useful and would likely be accepted by the users. To verify the hypothesis quantitatively in an empirical study, the experiment entails a group of nurses test a Wizard-of-Oz simulation (with respect to activity recognition) of such a system under realistic conditions and evaluate the documentation efficiency, documentation errors and the subjective perception of the system. We compare system variants without activity recognition, with simulated perfect recognition and with different (simulated) recognition error rates. Overall 272 instances of the documentation process were collected under different circumstances and evaluated.

As performing the experiment during the shift was considered disruptive by the hospital administration, we have recruited 15 nurses to use a test system immediately after the end of their shift. The documentation concerned real tasks actually performed during the shift. The study was conducted in the ward environment. The nurses were asked to perform the documentation during most common "dead cycle" (based on the observation study).

3.1 Documentation User Interface

We have designed the documentation interface using the iPhone scroll and select mechanism (Figure 1(a)). Searching for an activity is done by scrolling up and down, adding it to the care documentation by tapping and, if needed, entering a numerical value with the on screen keyboard.

Following a typical order of tasks the activities were grouped by care categories (i.e. morning hygiene, morning examination, ..) first and then ordered alphabetically. They are described by both text and easily recognizable icons.

The primary goal of the study was to understand the difference between documentation with and without activity recognition. Therefore we compared four versions of the basic documentation interface described above, one assuming activity recognition and three without:

Context list: A context-dependent activity list with only the activities that the nurse actually performed. The interface thus simulates perfect activity recognition.

Clearly, the benefit of activity recognition depends on the complexity of the selection task(=length of the list from which the tasks need to be selected). Thus, we used three different manually predefined list lengths:

Short list: this lists contains documented activities mandated by the national health authority. This list only comprises 17 activities.

Medium list: this list contains all (27) activities that nurses frequently perform (as observed). Need to document those activities was indicated by nurses.

Long list: All activities that have been observed (43).

The second study goal was to understand the impact of activity recognition accuracy on documentation overhead. Thus, we introduced random errors into the simulated activity recognition as follows: 100% correct recognized

"100%", 25% false positives "fp 25%" (activities not performed were recognized and had to be deleted), 50% false positives "fp 50%" , 25% false negatives "fn 25%" (activities performed were missing and had to be entered), and 50% false negatives "fn 50%".

3.2 Experiment Design

Test Scenarios: From mentioned observations the following dead cycle activities were defined: go to bathroom and fill sink (walk: 30 sec., fill sink: 43 sec.), go to kitchen and prepare coffee (walk: 42 sec., prepare coffee: 26 sec.), go to bathroom, empty sink, bring something (walk: 30 sec. empty sink: 13 sec.), return used items (to kitchen and bathroom, distance kitchen - bathroom: 10 sec), and check toilet (walk: 8 sec one way). The nurses were asked use mobile documentation while performing the above mentioned scenarios to resemble the real work situation.

Experimental Procedure: The subjects were divided into two groups. A group of 7 nurses was given the perfect activity recognition based interface and 3 different length lists (short, medium, and long). The second group of 8 nurses was given 5 activity recognition based lists with different accuracy. Each nurse had to perform the documentation for each patient with each list. Each run of each nurse was assigned to a different "dead cycles" scenario. The "lattice square approach" was used to vary the combination of test subject, scenarios and interface condition.

Overall this led to 16 documentation runs (4 patients each with 4 interface conditions) with 28 instances (7 nurses times 4 patients) in the first group and 20 runs (4 patients, 5 conditions) with 32 instances (8 nurses times 4 patients) in the second. A training run for each nurse, performed to minimize learning effects, was also used as the ground truth for the Wizard-of-Oz simulation of activity recognition.

Evaluation Criteria: The dependent variables were chosen to reflect usability aspects of the interfaces and characterize to which extent documentation can be a peripheral activity: first, documentation overhead (time additionally needed to document); second, interaction complexity (number of interaction steps required for documentation); third, documentation error rate (number of wrongly included or omitted activities). In addition we evaluated the subjective user experience through questionnaires.

4 Results

Perfect Recognition: Documentation Overhead: Figure 1(b) (dark gray) illustrates the average documentation overhead. There was a significant difference in the scores for context list against short, medium, and long lists (Table 2 a)). This suggest that activity recognition can significantly reduce the effort required to document care activities.

Table 2. Results (perfect/imperfect recognition)

a) Doc.overhead - perf.	mean (sec)	std	t	p
context	2.06	0.85		
short	4.21	2.03	-4.175	<= 0.01
medium	3.48	1.25	-5.659	<= 0.001
long	3.79	0.94	-8.360	<= 0.001
b) Inter.complex. - perf.	mean (step)	std	t	p
context	2.86	0.47		
short	5.61	1.72	-5.176	<= 0.01
medium	4.48	0.83	-10.661	<= 0.001
long	4.02	0.60	-9.992	<= 0.001
c) Doc.overhead - imp.	mean (sec)	std	t	p
100%	1.15	0.79		
false positive "fp" %25	4.73	1.60		
false negative "fn" %25	2.81	1.65	-2.171	<= 0.001
false positive "fp" %50	1.68	1.14	-5.223	<= 0.05
false negative "fn" %50	2.64	1.83	-5.223	<= 0.001
d) Inter.complex. - imp.	mean (step)	std	t	p
100%	3.60	1.26		
false positive "fp" %25	4.73	3.72	-2.239	<= 0.05
false negative "fn" %25	6.15	2.74	-7.235	<= 0.000
false positive "fp" %50	4.37	1.88	-2.541	<= 0.05
false negative "fn" %50	6.21	3.50	-4.631	<= 0.000

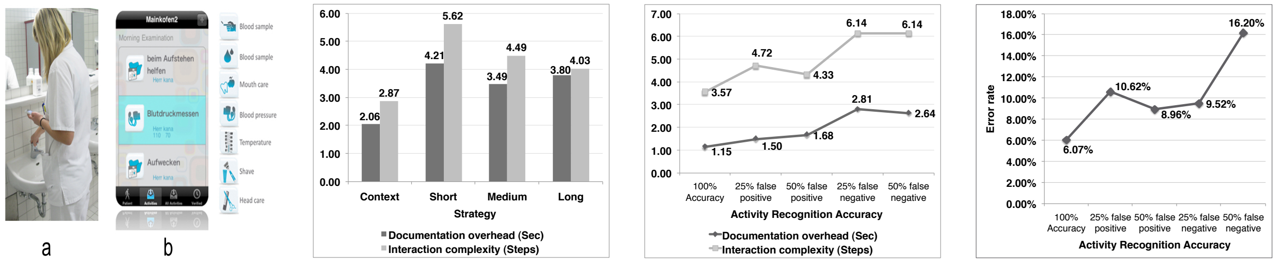
Perfect Recognition: Interaction Complexity: Figure 1(b) (light gray) illustrates the average interaction complexity. There was a significant difference in the scores for context list against short, medium, and long lists (Table 2 b)). This suggests that activity recognition can significantly reduce interaction complexity.

Table 3. System perception (1 - positive, 5 -negative)

Iphone:	glob. usage: 1.14		for documentation: 1.21	
pref. mode:	context list: 6 nurses		short list: 1 nurse	
dead-cycles:	documentation: 1.14			
d-cycle type:	while waiting: 1.71		while walking: 4.21	
list rating:	contex: 1.14	short: 1.79	medium: 4	long: 4.43

Imperfect Recognition: Documentation Overhead: Figure 1(c) (dark gray) illustrates the average documentation overhead for various recognition accuracies. There was no significant difference in the scores for 100% activity recognition accuracy against fp 25%. However, there was significant difference for 100% activity recognition accuracy against fp 50%, fn 25%, and fn 50% - see Table 2 c). These results suggest that the documentation overhead is significantly decreased with the increase of activity recognition accuracy. Furthermore, the results show that the false negative activity recognition errors have greater negative impact on the documentation overhead than false positive activity recognition error.

Imperfect Recognition: Interaction Complexity: Figure 1(c) (light gray) illustrates the interaction complexity for various recognition accuracies. There was a significant difference in the scores for 100% activity recognition ac-



(a) Mobile doc. interface (a. dead-cyle, b. the interface)

(b) doc. overh. (time) and interac. compl. (steps)

(c) Imp. Recognition: doc. overh. (time) and interac. compl. (steps)

(d) Documentation errors

Figure 1. Interface and quantitative Evaluation Results

curacy against 25% fp, 50% fp, 25% fn, and 50%fn - see Table 2 d). These results suggest that activity recognition accuracy does have an effect on number of interaction steps needed to verify health care records. Specifically, our results suggest that the interaction steps are significantly decreased with the increase of activity recognition accuracy, therefore the overall documentation is easier and less disruptive to the task in hand. The results also show that the negative effect of the activity recognition accuracy on the interaction steps is greater for false negatives.

Subjective System Perception: When asked by a short questionnaire, the nurses expressed satisfaction with the overall concept and rated the context interface most suitable for their needs (see Table 3). The preferred type of deadcycles for documentation was "while waiting". Unsurprisingly, the 100% accuracy interface was preferred. Moreover, the false positive interfaces were prefered to the false negative ones (5 out of 8 nurses). There is effectively no difference between the fn 25% and the 50% versions, while the fp 25% was rated slightly better then the fp 50%.

5 Conclusion

Our study shows that documentation tasks that can be mapped into "browse and select" list interaction, can be performed during or in between simple every day activities on mobile devices. On nearly all metrics we have seen a significant performance drop on the transition from activity recognition to non activity recognition based list selection and only slower, gradual performance degradation with increasing complexity of the selection task (as given by the relation between list length and the number of items that were to be selected). A similar trend could be observed with respect to activity recognition errors. The difference between 100% accuracy and 25% errors was much more pronounced then the difference between 25% and 50% errors. Nonetheless, all measures indicate that even systems with error rates around 25% are useful and would be accepted by the users (e.g. "would use" score of 1.3 and "tolerability" score of 1.89). This a very significant insight towards practical deployment of activity recognition, as such error rates are well within the current state of the art [1].

References

[1] K. Adamer et al. Developing a wearable assistant for hospital ward rounds: An experience report. *The Internet of Things*, pages 289–307.

[2] E. Ammenwerth, U. Mansmann, C. Iller, and E. Eichstdter. Factors affecting and affected by user acceptance of computer-based nursing documentation: Results of a two-year study. *Journal of the American Medical Informatics Association*, 1(10):69–84, January-February 2003.

[3] S. Hyun, J. Choi, J. Chun, S. Lee, D. Shin, D. Kim, and D. Kim. Implementation of mobile computing system in clinical environment: Mobilenurse™. In *Proc. AMIA Symposium*, page 1036. AMIA, 2000.

[4] K. Kunze, F. Wagner, E. Kartal, E. Morales Kluge, and P. Lukowicz. Does Context Matter? -A Quantitative Evaluation in a Real World Maintenance Scenario. *Pervasive Computing*, pages 372–389.

[5] N. Kuwahara, H. Noma, K. Kogure, N. Hagita, N. Tetsutani, and H. Iseki. Wearable auto-event-recording of medical nursing. In *Proc., 9th IFIP TC13 Int. Conf. on Human-Computer Interaction*, 2003.

[6] J. Ockerman and A. Pritchett. Preliminary investigation of wearable computers for task guidancein aircraft inspection. In *Second International Symposium on Wearable Computers, 1998. Digest of Papers*, pages 33–40, 1998.

[7] C. Orwat, A. Graefe, and T. Faulwasser. Towards pervasive computing in health care–A literature review. *BMC Medical Inf. and Decision Making*, 8(1):26, 2008.

[8] A. Schmidt, M. Beigl, and H. Gellersen. There is more to context than location. *Computers and Graphics*, 23:893–901, 1998.

[1]This work was supported by the European Union in the ALLOW project