Does Context Matter ? - A Quantitative Evaluation in a Real World Maintenance Scenario

Kai Kunze[1], Florian Wagner[1], Ersun Kartal[2], Ernesto Morales Kluge[3],
and Paul Lukowicz[1]

[1] Embedded Systems Lab, University of Passau,
Innstr 43, 94032 Passau, Germany
www.wearable-computing.org, www.wearcomp.eu
first.lastname@uni-passau.de
[2] Carl Zeiss AG, Konzernfunktion Forschung und Technologie (KFT-TV)
73446 Oberkochen, Germany
[3] Bremer Institut für Produktion und Logistik, University of Bremen,
Hochschulring 20, 28359 Bremen, Germany
www.biba.uni-bremen.de
mer@biba.uni-bremen.de

Abstract. We describe a systematic, quantitative study of the benefits using context recognition (specifically task tracking) for a wearable maintenance assistance system. A key objective of the work is to do the evaluation in an environment that is as close as possible to a real world setting. To this end, we use actual maintenance tasks on a complex piece of machinery at an industrial site. Subjects for our study are active Zeiss technicians who have an average of 10 years job experience.

In a within subject Wizard of Oz study with the interaction modality as the independent variable we compare three interaction modalities: (1) paper based documentation (2) speech controlled head mounted display (HMD) documentation, and context assisted HMD documentation. The study shows that the paper documentation is 50% and the speech only controlled system 30% slower then context. The statistical significance of 99% and 95% respectively (one sided ANOVA test). We also present results of two questionnaires (custom design and standard NASA TLX) that show a clear majority of subjects considered context to be beneficial in one way or the other. At the same time, the questionnaires reveal a certain level of uneasiness with the new modality.

1 Introduction

Since early conceptual work on the use of context in pervasive systems (e.g. [5, 14, 11]) much research aims at implementing context and activity recognition systems. Interestingly, researchers devoted little work to a systematic, quantitative evaluation of the benefit such systems bring to diverse applications. This paper presents such a systematic, quantitative evaluation.

We focus on the domain of wearable maintenance systems. Many such systems are proposed and implemented since the early days of wearable computing (e.g. [1, 13, 15, 16]). These systems aim to provide maintenance personnel with access to complex

H. Tokuda et al. (Eds.): Pervasive 2009, LNCS 5538, pp. 372–389, 2009.

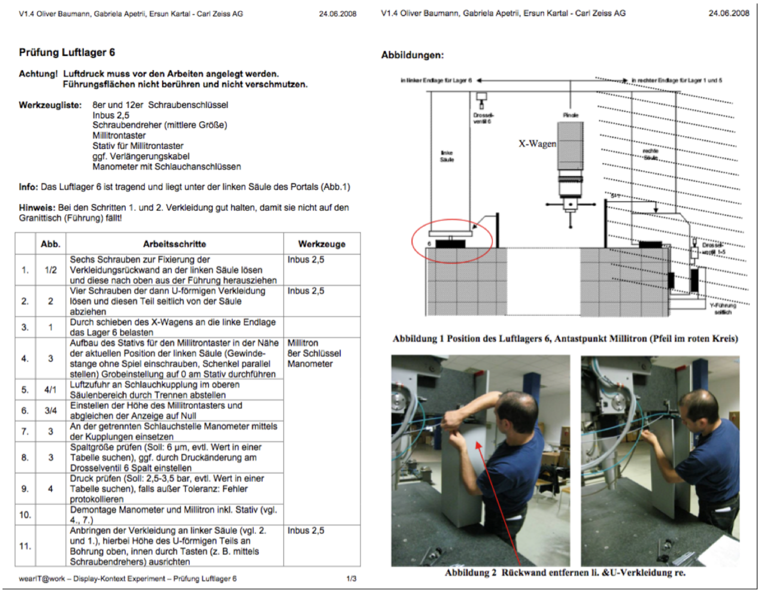
electronic information with as little interference as possible to the primary task at hand. Typically, they rely on head mounted displays (often with augmented reality), input modalities that minimize hand use (e.g. speech, special gloves) and interfaces that focus on reducing the cognitive load on the user.

It is widely believed that wearable maintenance systems can benefit from automatic work progress tracking. Main uses for such tracking are 'just in time' automatic delivery of information (seeing the manual page you need without having to explicitly demand it), error detection (e.g. "you forgot to fasten the last screw"), and warnings (e.g. " do not touch this surface"). In this paper we present a quantitative, statistically significant benefits evaluation of such functionality in a real industrial setting. The study involves 18 real technicians, most with over 10 years of job experience, doing 3 different, real maintenance task on a complex piece of industrial machinery. We compare three types of systems: (1) paper based documentation, (2) documentation displayed on a head mounted display controlled by a speech interface, and (3) context controlled documentation displayed on a head mounted display. Both speech recognition and context recognition are simulated using the Wizard of Oz technique to ensure perfect system performance and avoid system quality related artifacts. Key results are that the average time per task is around 50% longer when using paper based documentation than when using the context (level of confidence 99%) supported system. The speech controlled HMD system is a bit faster but still around 30% slower then the context controlled version (confidence level 95%). We also present and discuss the results of two questionnaires (NASA TLX and custom) assessing the subjective view of the participants.

2 Related Work and Paper Contributions

Following early conceptual work on the usefulness of context (e.g. [5,14,11]) there has been an extensive body of work on tracking a multitude of activity types from fitness, through furniture assembly to health care related issues. By contrast, only very few projects deal with the evaluation of the benefit that context recognition brings to different applications. Bristow et. al. demonstrate how context information speeds up access to environment related information from the Internet ([2]). A similar study related to the use of physical context for information retrieval was given by Rhodes ([10]). Smailagic et. al. study a mobile phone that provides the caller information about other persons context ([12]). They show that using such information to prompt the caller to speak slowly or make pauses reduces the risk of using the phone while driving. Some qualitative discussion of a context sensitive tourist guide application has been presented in [4]. In another qualitative study context sensitivity has been evaluated in a wearable nursing support system([6]).

Somewhat related to this paper is research evaluating wearable maintenance assistance systems in general without including the context issue. Examples include a wearable remote collaboration system evaluated on a bicycle repair task, an evaluation of wearable system for aircraft maintenance, studies using HMD technology for guided instructions in a medical setting and early work from in Mizell and Caudell that hints at the usefulness of HMDs in maintenance scenarios in a qualitative way [3,8,9,12].



V1.4 Oliver Baumann, Gabriela Apetrii, Ersun Kartal - Carl Zeiss AG 24.06.2008

Prüfung Luftlager 6

Achtung! Luftdruck muss vor den Arbeiten angelegt werden.
Führungsflächen nicht berühren und nicht verschmutzen.

Werkzeugliste: 8er und 12er Schraubenschlüssel
Inbus 2,5
Schraubendreher (mittlere Größe)
Millitrontaster
Stativ für Millitrontaster
ggf. Verlängerungskabel
Manometer mit Schlauchanschlüssen

Info: Das Luftlager 6 ist tragend und liegt unter der linken Säule des Portals (Abb.1)

Hinweis: Bei den Schritten 1. und 2. Verkleidung gut halten, damit sie nicht auf den Granittisch (Führung) fällt!

	Abb.	Arbeitsschritte	Werkzeuge
1.	1/2	Sechs Schrauben zur Fixierung der Verkleidungsrückwand an der linken Säule lösen und diese nach oben aus der Führung herausziehen	Inbus 2,5
2.	2	Vier Schrauben der dann U-förmigen Verkleidung lösen und diesen Teil seitlich von der Säule abziehen	Inbus 2,5
3.	1	Durch schieben des X-Wagens an die linke Endlage das Lager 6 belasten	
4.	3	Aufbau des Stativs für den Millitrontaster in der Nähe der aktuellen Position der linken Säule (Gewindestange ohne Spiel einschrauben, Schenkel parallel stellen) Grobeinstellung auf 0 am Stativ durchführen	Millitron 8er Schlüssel Manometer
5.	4/1	Luftzufuhr an Schlauchkupplung im oberen Säulenbereich durch Trennen abstellen	
6.	3/4	Einstellen der Höhe des Millitrontasters und abgleichen der Anzeige auf Null	
7.	3	An der getrennten Schlauchstelle Manometer mittels der Kupplungen einsetzen	
8.	3	Spaltgröße prüfen (Soll: 6 μm, evtl. Wert in einer Tabelle suchen), ggf. durch Druckänderung am Drosselventil 6 Spalt einstellen	
9.	4	Druck prüfen (Soll: 2,5-3,5 bar, evtl. Wert in einer Tabelle suchen), falls außer Toleranz: Fehler protokollieren	
10.		Demontage Manometer und Millitron inkl. Stativ (vgl. 4., 7.)	
11.		Anbringen der Verkleidung an linker Säule (vgl. 2. und 1.), hierbei Höhe des U-förmigen Teils an Bohrung oben, innen durch Tasten (z. B. mittels Schraubendrehers) ausrichten	Inbus 2,5

wearIT@work – Display-Kontext Experiment – Prüfung Luftlager 6 1/3

V1.4 Oliver Baumann, Gabriela Apetrii, Ersun Kartal - Carl Zeiss AG 24.06.2008

Abbildungen:

Abbildung 1 Position des Luftlagers 6, Antastpunkt Millitron (Pfeil im roten Kreis)

Abbildung 2 Rückwand entfernen li. &U-Verkleidung re.

Fig. 1. A sample of the paper manual for the task 'checking bearing pressure'

Paper Contributions. The paper presents a quantitative, statistically relevant study showing the benefit of context recognition in wearable maintenance support systems. It does so in a real world, industrial environment with real, professional technicians and real maintenance task on a complex piece of machinery. This clearly goes beyond what has so far been investigated with respect to the usefulness of activity tracking in wearable maintenance systems. As sketched above, it also goes beyond much previous work on the benefits of context recognition in general.

We carefully describe out experimental design including a discussion of key considerations allowing other groups to learn from our experience. In addition to the quantitative results we describe a range of interesting qualitative observations. All results are discussed and put into perspective. We believe that this work constitutes an important piece of information for people designing context aware maintenance support systems as well as for more general context aware, assistive systems. It also provides a strong argument for continued research and development of such systems.

3 Experiments

Subsequently, we give an overview of the tasks, the selection process, the experimental design and setup.

3.1 Tasks

Task Selection Process. Task selection is a crucial step in our experimental design. We want to use a real maintenance task representative of the subjects' daily work, no artificial 'toy activities'. We want a complexity level that makes the use of some sort of documentation unavoidable. This also means that the specific task should be unfamiliar to our subjects (although of a general type to which they are used). In addition, the task needs to be not too short so that differences in performance can be resolved. On the other hand, the task can not be too long and too complex because the amount of technician time that we are allocated was limited. We also want the task to be doable by a professional without additional training. On the practical side, we need to find a machine that could be 'spared' for a couple of days and where a maintenance task could be performed repeatedly without fear of causing significant damage.

Finally, we need not one, but three tasks. We require each technician to use each of the three modalities (paper, HMD without context, HDM with context as described below). Yet, we want to avoid learning effect on the tasks.

The selection process involved several visits to the Zeiss facility, discussions with the responsible personnel and test runs of task candidates with a technician that was familiar with them. This was followed by test runs with ourselves and novice technicians.

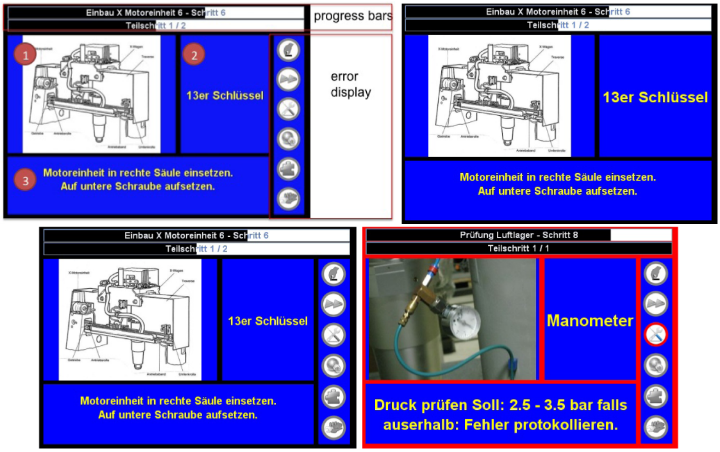
Task Overview. Finally, we selected three maintenance tasks at a metrology system, more specifically at a Zeiss UMC 850 coordinate measuring machine (CMM) as shown in Figure 4. The specific machine is an older model taken back from a customer in an upgrade deal. This means that most technicians are not familiar with it and that it is 'not critical' in terms of any damage resulting from the experiments. The procedures that we selected can be summarized as follows:

1. Checking the bearing pressure on the left column. This task involved removing the casing, straining the air bearing, assembling the measurement apparatus, cutting the air supply, adjusting the militron, insertion of the manometer, measuring the gap size, measuring the pressure, disassembling the measuring apparatus, fixing the casing.
2. X-Motor installation with the following steps: attaching the belt pulley, attaching the oscillating element, fixing the belt at the engine shaft, attaching the basis plate, installing the motor, connecting the electronic cables.
3. Y-Gears installation with the following steps: affixing the gears, attaching the belt, threading the belt through the gears, attaching the belt, checking the friction clutch and the deflections towards the x and y-axis and adjusting the belt accordingly.

As seen from the descriptions, the tasks require being a trained technician to even understand them, not to speak of being able to perform them. They are by no means simple.

3.2 The Support Modalities

We investigate three support modalities: (1) traditional paper based documentation, (2) a speech controlled wearable support system with a head mounted display, and (3) a wearable support system with speech control and context aware support. Thus, we can determine how much of the improvement over paper based systems comes from the wearable system in general and how much is actually due to context.



Motoreinheit in rechte Säule einsetzen.
Auf untere Schraube aufsetzen.

Druck prüfen Soll: 2.5 - 3.5 bar falls
auserhalb: Fehler protokollieren.

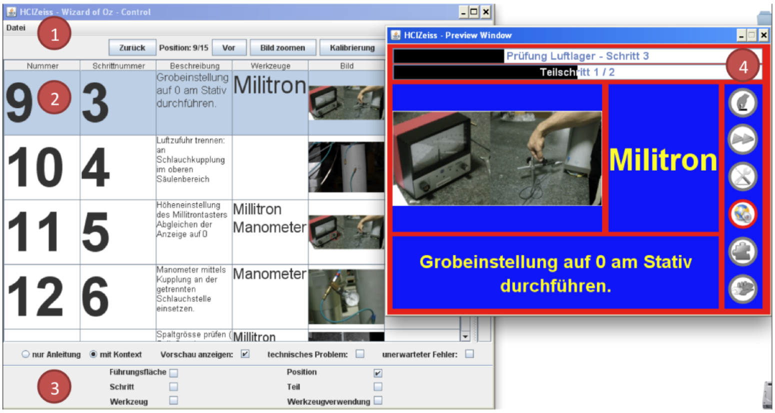
Fig. 2. The HMD UI as seen from the technician, first an overview, the second picture shows the UI in speech mode only, the last two with context recognition/error detection active

Depending on the modality, paper, speech or speech with context (we will refer to the later one as context for the rest of the paper), the technician has specific help to perform all three maintenance tasks.

To develop the UI and control application we used an iterative approach testing it during 2 test runs with 2 experienced technicians doing incremental improvements. Of course, these technicians have not participated in the later user study.

Paper Manual. As the official maintenance documentation manual contains over 800 pages and the information useful to the chosen maintenance tasks is spread throughout the manual, we decided to gather all relevant information and compile it into single compact document for each maintenance task. Our paper manual, a sample is depicted in Figure 1, contains general information on the top, a list of tools to use, the task steps in a table with the tools to use and references to pictures and pictures on the following pages. We evaluated our manual instructions with the 2 novice technicians, to be sure that they include all the necessary information to complete the three tasks.

Speech Controlled HMD GUI. The paper manuals are the basis for the instructions displayed in the HMD. We used only the information, pictures and text provided in the paper manual for the GUI instructions. No additional material/animation/video etc. is presented in the HMD GUI. The HMD user interface is depicted in Figure 2. The UI shows a task overview first, like a table of contents. Then, each step is displayed. If available, the technician sees a picture of the task at hand (1), the tools he needs to use (2) and a short description of what to do (3). On the top of the screen there are two progress bars, one for the overall task and one for the subtask as given in the table of contents. The technician has the following speech commands to navigate between task steps: next, previous, index (to display the table of contents), go to step no., zoom in and zoom out (for the images, one level of magnification only).


HCIZeiss - Wizard of Oz - Control
Datei
1
Zurück
Position: 9/15
Vor
Bild zoomen
Kalibrierung
2
3
HCIZeiss - Preview Window
4
Prüfung Luftlager - Schritt 3
Teilschritt 1 / 2
Militron

Nummer	Schrittnummer	Beschreibung	Werkzeuge	Bild
9	3	Grobeinstellung auf 0 am Stativ durchführen.	Militron	
10	4	Luftzufuhr trennen: an Schlauchkupplung im oberen Säulenbereich		
11	5	Höheneinstellung des Millitrontasters Abgleichen der Anzeige auf 0	Millitron Manometer	
12	6	Manometer mittels Kupplung an der getrennten Schlauchstelle einsetzen.	Manometer	
		Spaltgrösse prüfen (	Millitron	

○ nur Anleitung ● mit Kontext Vorschau anzeigen: ☑ technisches Problem: □ unerwarteter Fehler: □

Führungsfläche □ Position ☑
Schritt □ Teil □
Werkzeug □ Werkzeugverwendung □

Grobeinstellung auf 0 am Stativ durchführen.

Fig. 3. The Wizard-Of-Oz control gui with preview enabled

The HMD UI is naturally constrained by the actual HMD we use. The colors we use, yellow on blue, provide the strong contrast ratios on the display and are save against ghost images, a problem we faced using black on white, for example.

The Context Aware HMD System. If context recognition is enabled an error bar is placed on the right side of the screen. The technician gets an alert if one of the following errors happen:

1. Touching the bearing surface. The bearing surface is not supposed to be touched as this can result in the need for recalibrate the machine which takes several hours.
2. Wrong task step. The technician missed an essential task previous to what he is currently working on.
3. Wrong tool. The technician is holding the wrong tool for the current task step.
4. Wrong position. The technician is not at the correct position relative to the machine to perform the task at hand.
5. Wrong part. The technician is operating/using at a wrong part of the machine.
6. Wrong tool usage. The technician uses the correct tool in a wrong way.

This error information is conveyed to the technician using the pictograms on the right. If a technician finishes a step successfully in context mode the UI switches to the next maintenance step.

The Wizard of Oz Control. We control the HMD display using a desktop application with the wizard of oz (WOZ) approach. This way, we avoid artifacts from the speech and context recognition system performance. This also saves us the effort of implementing the context recognizer which is highly non trivial.

The code for the HMD UI and the WOZ control application are written in Java. The complete implementation is open-sourced and published under http://sourceforge.net/projects/jwoz. The control application allows the wizard of oz to steer the UI the technician sees. He can display every task step and error information.

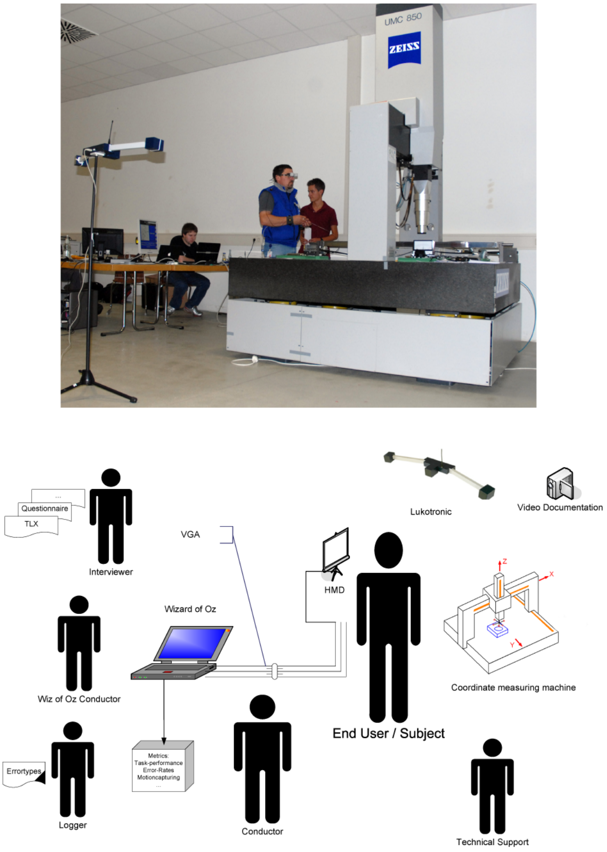


Fig. 4. The UMC 850 coordinate measuring machine and the experimental setup overview

The control gui is depicted in Figure 3. The wizard is able to load a given manual into the application and step with next and previous through the maintenance task steps (1). The task steps are displayed in a table view (2), highlighting the currently selected one. Below the table are radio buttons and checkboxes (3) for enabling context or speech mode, logging an unexpected error (an error that can not be put in one of the categories described above, in that case nothing is given as feedback, it is just written in the logfile), logging a technical problem and logging/displaying the error categories described on top. The Wizard of Oz is also able to display a preview window on the screen of the HMD screen. For the experiments we used a separate monitor. The control application is able to detect the monitor and convey the technician screen to it.

3.3 The Setup

For the experiments, we used the following setup shown in Figure 4. The test subject performed one of the three maintenance tasks at the machine using either the paper manual, the HMD with speech control or the HMD with speech control and context. A conductor introduces the test subject to the technology, he is the one contact person for the technician, if questions/technical problems etc. arise. He also signals to the wizard of oz the errors and completions of task steps. Over a separate monitor, the conductor can see the current UI the technician is looking at. The wizard of oz controls the UI over a desktop, which is directly connected over a VGA cable to the HMD and the monitor for the conductor. Another person is responsible for doing periodic consistency checks on the logs of the Wizard of Oz application, as well as running and maintaining the systems for ground truth information. We deployed two systems for ground truth, a simple video camera capturing the machine and the lukotronic active infrared marker system to capture the movements of the test subject (the information from this system was not evaluated in this paper). The logger also checks the error types helping the conductor/wizard of oz. A interviewer is located in a separate room interviewing the test subject using a custom questionnaire and the Nasa Task Load Index (TLX) after each experimental trial. In addition to this, at least one person is on stand-by for technical support. The test subject wears a vest on which the HMD, batteries and the cabeling is fixed together with 2 infrared markers for the Lukotronic systems. The subject also wears easy to strap on wrist bands with the remaining infrared marker (3 for each hand) for each trial.

3.4 The Studies

The Subjects. We selected a total of 18 subjects for participation – 16 males and 2 females aged between 17-56 years (mean 38.9 years). They were performed the tasks as part of their normal job. All 18 are professional maintenance operators from the Zeiss maintenance facility actively working in the maintenance field. Yet, they are not familiar with the CMM they have to perform the experiments on.

Study Design. The study uses a within subjects design with the interaction modality as the independent variable, meaning that all subjects will test every interaction modality in one of the three maintenance tasks.As there are three interaction modalities to test (and three different maintenance tasks to avoid bias caused by a learning effect), a Greco-Latin Square of the same order is used to distribute the 18 participants.

The LEGO Practice Round. To familiarize with the HMD, the speech commands and the error displays, we let each test subject perform several steps in building a simple LEGO Technic Forklift in a separate room. We picked the LEGO bricks assembly as it has nothing in common with the actual task and it is simple to explain the working of the UI and the error displays using this setup.

The Experimental Session. A test session consists of one practice round where the subject gets to practice each interaction modality, followed by the experimental rounds during which data is collected and interviews (questionnaires and TLX) are carried out for analysis. The practice round uses the same modalities of the experiment and an abstracted build task that allows subjects to get familiar with the modalities and that avoids fatigue. As explained above, the practice rounds are conducted using a LEGO

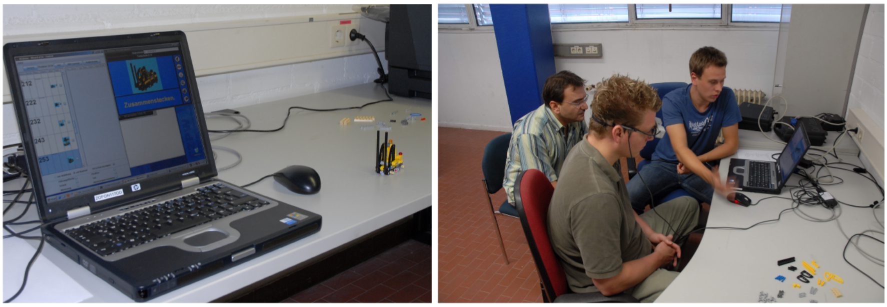


Fig. 5. The lego practice round

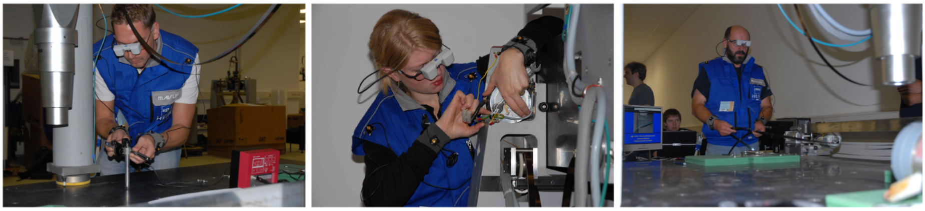


Fig. 6. Maitenance pictures from the experimental sessions. From left to right: Checking bearing pressure, installing the xmotor and ygears.

mockup. The total time required for a session is around 210 minutes. The three maintenance tasks are around 20-30 min each, however adding the time for the mockup and the interviews it takes substantially longer.

4 Quantitative Results

4.1 Objective Performance Metrics

Obviously, the two key objective performance metrics for a maintenance task are the time needed to perform the procedure and the number of mistakes.

Time. Figure 7 shows the average time per task for the three different modalities. The averages are taken over all workers and tasks. Since each worker performed each task only once and used each of the three modes only once we have 18 data points for each mode. These 18 instances are approximately equally distributed over the three tasks. Using paper documentation is on average around 50% (22.0 vs. 14.6 min) slower than context assisted HMD based documentation. The HMD without context is in between the two: 19.5 min, 30 % slower than context assisted case and about 15% faster than paper documentation. This is a significant difference, which however has to be seen in the context of high variation of individual times, in particular in case of paper manual.

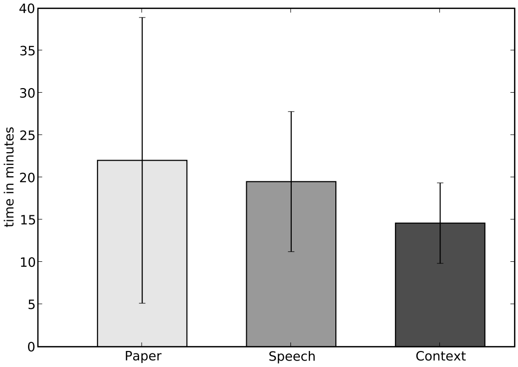


Fig. 7. The average time needed over all maintenance tasks split in modalities with standard deviation

As a consequence, when applying the one way ANOVA (F-Test)[1] to asses the statistical significance of the results we arrive at a confidence level of 99% (p-value: 0.01) for the comparison between context and paper,95% (p-value:0.04) between context and HMD, and only 80% (p-value:0.22) between HMD and paper.

In summary, we can say that the use of context improves the efficiency of our tasks in a relevant, statistically significant way. For HMD alone the results strongly point towards an improvement but the sample size and large variation between subjects mean that the results are not statistically significant (typically 95% is picked as the threshold for statistical significance).

4.2 Mistakes

The total number of mistakes made by all subjects in all tasks was 48 for paper, 31 for HMD without context and 29 for the context driven system. This clearly shows that the use of a wearable HMD based systems reduces the number of mistakes. It also indicates that the key factor in the error reduction is not context, but easy access to the information on the HMD.

Where context does make a significant difference is in the average time needed to recover from the mistakes, depicted in Figure 8. On average it has taken the workers about double as long to recover from a mistake when using paper then when using context. The use of HMD without context was only insignificantly faster then paper documentation. The statistical significance of the comparison between paper and context and HMD with and without context are both 99% respectively.

It is interesting to note that despite the huge relative difference in error recovery times, error recovery is not a relevant factor in the speedup in the overall average execution time. The error recovery speedup was in the range of 25sec, whereas the overall speedup is about 7min.

[1] The one-way ANalysis Of VAriance is used to test for differences among two or more independent groups and provides a likelihood to reject the null-hypothesis.

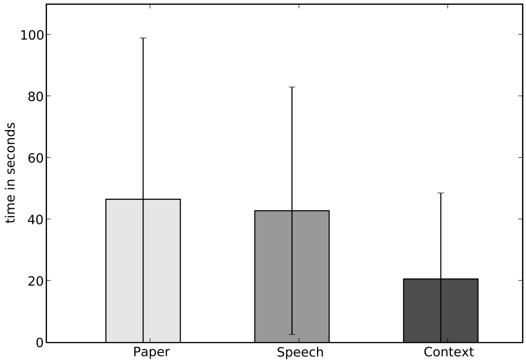


Fig. 8. The average time spent for resolving errors split in modalities with standard deviation

4.3 System Perception

We assess the subjective perception of the system with two questionnaires: one custom questionnaire designed for our specific study, and the standard NASA TLX questionnaire [7].

Custom Questionnaire. The custom questionnaire with a summary of the answers is shown in Figure 1. It contains 5 question groups. The first refers to the qualifications and background of the workers. Because of the small sample size correlating different backgrounds with the usefulness of context and HMD makes little sense, although it would be an interesting scientific question. Instead those questions are meant to provide an overview of the type of subjects that we were working with. It can be seen that (with the exception of two trainees) all consider themselves highly skilled in the repair of machines and mostly have 10+ years of experience on the job. Most, 15 out of 18, rate themselves as very good to average in terms of computer skills. The groups has a good age mix.

The second is the most relevant group of questions that reflects the perception of the context sensitive assistance system. The key questions are:

1. Overall impression of the system. On a scale from 1 (very good) to 6 (very bad) we have an average of 2,7 with 10 times 2, 5 times 3 , 2 times 4 and a single 6.
2. The favorite mode (paper, HMD with speech, HMD with context). Of the 18 participants 8 chose context, 6 chose HMD without context, 1 chose paper and 1 was undecided between context and paper.
3. The worst mode (paper, HMD with speech, HMD with context). Here 17 of the 18 participants named paper and one was undecided between HMD with and without context.
4. The improvement brought by context. On a scale from 1(very good) to 6(none) 12 participants picked 1, 3 picked 2 and 2 picked 3 while 1 participant failed to answer the question. This gives an average of 1.4. This might be see as an inconsistency with the question on favorite modality where only 8 subjects picked context. On the

Table 1. A summary of the questionaire filled out during the interviews with each participant

Questionaire	
Background	
How long are you employed in maintenance?	mean 14.81 years
How old are you?	mean 38.90 years
How much computer experience do you have?	mean 2.42 (Scale 1 to 6)
Are you satiesfied with your results?	mean 2.19 (Scale 1 to 6)
Perception of the context sensitive system	
How was your overall impression?	mean 2.72 (Scale 1 to 6)
Wich modality did you like best	P:1 HS:7 HC:10 counts
Which modality did you dislike most	P:16 HS+HC:1 counts
Wearing comfort	
How comfortable was the HMD?	mean 2.78 (Scale 1 to 6)
Would you wear the system for daily work?	yes:13 indifferent:4 no:1 counts
Do you felt relieved after taking of the HMD?	yes: 9 indifferent:1 no:8 counts
Do youl felt the system as obtrusive?	yes:6 indifferent: 2 no:10 counts
Do you had problems with the weight of the system?	yes: 4 no:14 counts
Would you have liked more help (inside the application)?	yes:13 indifferent:2 no:3 counts
How easy/difficult was it to put on the HMD?	mean 2.00 (Scale 1 to 6)
Was the screen big enough?	mean 2.11 (Scale 1 to 6)
How sharp/unsharp was the HMD image?	mean 2.83 (Scale 1 to 6)
How light/dark was the HMD image?	mean 1.89 (Scale 1 to 6)
Interface quality	
Did you need the offered resources and tools (system)?	yes:16 indifferent:2 no:0 counts
How comfortable was the navigation?	mean 2.11 (Scale 1 to 6)
How good/bad could you read the text?	mean 2.56 (Scale 1 to 6)
How much did you like the choice of colours and fonts?	mean 2.17 (Scale 1 to 6)
Motivation	
How motivated dou you felt during the experiment?	mean 2.00 (Scale 1 to 6)
How tensed dou you felt during the experiment?	mean 2.58 (Scale 1 to 6)

other hand, it is not unusual for people to understand that something might improve their work and still feel more comfortable with a different solution.

5. Willingness to use such a system in every day work. Again on a scale from 1 to 6 we get an average of 1.3 with 13 times 1, 4 times 2 and a single 3. Note that this question did not differentiate between a system with and without context.

In summary the above questions group indicates a positive subjective perception of the system.

The third questions group concentrates on the perception of the head mounted display. It was included out of two reasons, one due to interest of Zeiss (who manufactures the displays) and second, it should show whether there is reason to believed that the results were skewed by poor HMD quality. It can bee seen in Table 1 that most subjects rated the comfort and weight of the HMD to be average and were roughly evenly split between those say there were relieved to take the display off and those who say they were not. Thus, we conclude that there is no obvious indication for a HMD related skew.

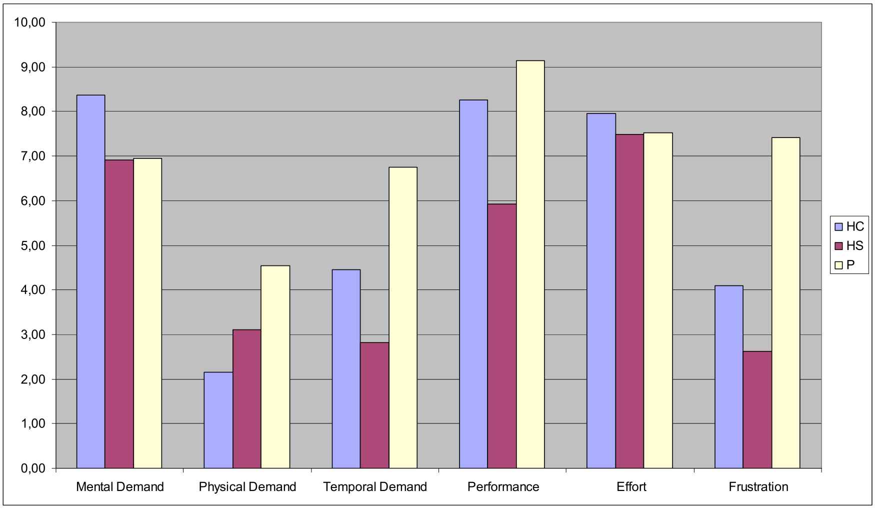


Fig. 9. Nasa TLX (see [7] for detailed description) results. Large values are ”worse”.

The fourth questions group dealt with the quality of the interface. Like the HMD related questions, it was meant to establish if there was reason to believe that the results were somehow skewed by the interface. Again the questionnaire shows no indication of such a skew. On all counts including navigation and readability the users rate the interfaces mostly between good and average.

The final two questions relate to the motivation and level of comfort during the experiments. On both issues the answers are mostly 2 and 3.

NASA TLX Questionnaire. From the background of the objective performance metrics and the custom questionnaire described above the results of the TLX questionnaire [7] shown in Figure 9 seem surprising at first. On every question with the exception physical effort the average score of the context based system is worse than that of a speech only HMD system (large score is worse for all items). In fact, in some cases (mental demand and effort) context even scores worse then paper. This is particularly surprising with respect to mental demand, since reducing mental demand is one of the key goals of context awareness.

However, when we examine the answers in more detail, a different picture emerges. Table 2 shows the breakdown of the answers by the number of subject who rate one modality to be better (lower score !), equal, or worse (higher score !) then the other on each of the criteria. It can be seen that with respect to mental load, effort and frustration more subjects consider context to be better then speech than the other way around (9 vs. 4, 7 vs. 4, and 15 vs. 3). The same is true for the comparison of context an paper. However, it seems that the comparison of paper with context is more polarizing. For mental load and effort there are 5 (4) subjects who consider both equal. For context and paper this numbers are 0 and 1.

It is interesting to note how clearly context wins against both speech only and paper with respect to frustration (15 vs. 3 against speech and 17 vs. 1 against paper).

Table 2. The number of people who have rated one modality better then the other on the different NASA TLX metrics

Lower score is better so $A > B$ means A scores worse then B !						
	$HC > HS$	$HC = HS$	$HC < HS$	$HC > P$	$HC = P$	$HC < P$
Mental	4	5	9	7	0	11
Temporal	10	0	8	4	2	12
Performance	9	2	7	5	1	12
Effort	4	4	10	7	1	10
Frustration	3	0	15	1	0	17

Even in the breakdown in table 2 context ”looses” with respect to two criteria. The majority (10 vs. 8) think they were faster with the speech only interface. In reality 11 subjects were actually faster with context and only 8 with speech. Here it is important to note that subjects performed different tasks with different modalities. At the same time, tasks were not equally long so that even if a subject was in principle more efficient with a given modality, he might have scored a better absolute time with a different modality. Therefore, these results must only be viewed as a trend over all users.

The second criterion where context ”looses” is performance. Is is defined as the users subjective perception of how well he/she performed. The score here is clearly consistent with the skewed perception of the actual timing with different modalities.

In summary, the averaged TLX scores are a result of the subjects being against context are so by a large margin. Broken down by the number of people that score one modality better then another, the TLX results are reasonably well in line with the results of the custom questionnaire described in the previous section. The majority of the users see benefits from the context system (most clearly on the level of frustration). On the other hand, the subjective perception of the context systems seems to be worse then the objective data. To a degree, this is understandable as it is well known that people tend to perceive interaction modalities that are new and unusual for them in a more negative way. The key question, which we can not answer in this study, is whether there are other issues beyond context being new and strange. Thus, the question is if the negative bias will go away on its own, once people have worked with context aware systems long enough or whether there are some fundamental reasons why some people are uneasy about context controlled systems (e.g. they feel stressed by the system doing something on its own).

5 Qualitative Results and Observations

In this section we summarize some observations which are not backed by statistically relevant numbers but which we nonetheless consider noteworthy.

5.1 Most Proficient Subject

The overall fastest test subject managed the xmotor task in 9 min with the speech interface. With context he needed only 12 min for the ygears task. Of course, the two tasks are not really comparable. The subject did not do any mistakes, which is another indication of his skills.

It was interesting to see how this skilled subject used the system. We needed to relabel a lot of his data using video analysis, as he already skipped forward to the next steps to get an overview while working on earlier task steps.

This subject was also one of the few using the zoom function extensively. He used 15 times the picture zoom combined over all maintenance tasks, compared to an average mean of 3 zooms over all other participants.

5.2 Errors

Using the speech and context modalities not many bearing surface errors happened (an total of 3 compared to 9 in paper). We assume that although the paper has the information about the bearing surface in bold over the steps, it is often overlooked and not read carefully. Yet, in the speech and context mode the subjects are directly confronted with the text on screen and also often they asked the moderator again where the bearing surface is exactly located.

The errors in the context and speech are more deterministic than the ones during paper (happening at the same task steps), suggesting that an improved version of the system could fix this.

During paper and speech trials a total of 3 part errors for connecting the wrong electronic cables on the xmotor happened. This would have caused the motor to be destroyed when the machine is switched on. A context sensitive system detecting such errors can mitigate this problem.

5.3 HMD

One direct implication we saw, is that with HMD the technician focuses very much at the step displayed (and the last steps they saw). This is an important aspect for designing HMD systems. For example, one common error was in step 10 of checking the air pressure: The technician first employs two measuring devices and then needs to check the values for both of them. Most technicians built up both and measured first with the last one, which was the wrong one to use.

6 Conclusion

We have presented a systematic quantitative evaluation of the usefulness of context in a wearable maintenance assistance scenario. A key objective of the work is to perform the evaluation in an environment that is as close as possible to the setting in which such system would be deployed in a real world scenario. We have achieved this by picking real maintenance tasks on a complex piece of machinery, and using real technicians (who did not volunteer but did this as part of their normal job). We have carefully selected the procedures to be complex enough not to be doable without instructions but not too complex to be performed without extensive prior training. We also made sure that the procedures were long enough to resolve the effects we were looking for.

Despite the constraints of an industrial environment we managed to get enough subjects and runs to achieve statistical significance on the objective, quantitative metrics.

As an indication of the effort involved in the experiment consider the fact that initial discussion with Zeiss started more then a year before completing the experiment. This was followed by about 10 visits to the site to search for adequate machinery and tasks, evaluate manuals and interfaces with different technicians, and test the overall technical setup. The actual experiment took over a week during which we recorded and analyzed over 60 hours of data.

We believe that the experimental procedures described in this paper together with the scripts and software available from our www site are a significant additional contribution of the paper, from which other groups aiming at real life experiments can learn.

6.1 Key Conclusions

The most important conclusions from the study can be summarized as follows:

1. On average, the use of context information speeds up the procedures in a significant, statistically relevant way. Paper is 50% slower then context, speech control is 30% slower.
2. The amount of errors is already significantly reduced by the use of speech controlled HMD documentation. The addition of context brings a minimal additional improvement. However context reduces the time needed to correct the errors by an average of nearly 100% as compared to both speech and paper.
3. From our qualitative observations it seems that context is more useful for technicians who are less proficient. The technicians who were fastest with paper were the once who made no mistakes. The fastest technician achieved the result with the speech controlled system and used it in such a way (looking ahead of the task) that default context control (display manual for current task) would not have work. In fact, he was slower with context. An alternative conclusion could be that we need to rethink the way we use context.
4. A clear majority of subjects considered context to be beneficial in one or the other way. This was reflected in the respective questions of the custom questionnaires (12 participants very strongly and 3 strongly agreeing that context brought a benefit to their work) and in the answer counts (how many people though one modality was better then the other) in the TLX Frustration, Mental Load and Effort metrics.
5. However, the subjectively perceived advantages of context are less clear then the time measurement might suggest. Participants tend to subjectively underestimate their performance when using the context controlled system, which was reflected in the Temporal and Performance TLX metrics. Also, significantly less participants picked the context as their preferred modality that strongly agreed that context was beneficial (8 vs. 15). This suggests a degree of uneasiness towards the context control. At this stage, we can not say if this is the usual uneasiness towards the unknown or whether there are more fundamental issues behind it.
6. Interestingly, there was much less uneasiness towards the speech controlled HMD system. Context 'lost" mostly to speech, nearly never to paper. The only participant who picked paper as preferred modality wore special contact lenses and was unable to see clearly on the HMD.

7. When people felt that context was inferior, they tended to consider it significantly inferior. Similar applied to those who considered paper to be better. This is reflected in the average TLX scores, were context scored worse then speech controlled HMD.

6.2 Open Questions and Future Work

We are convinced that the study presented in this paper is a valid and relevant 'data point'. However, without doubt, it leaves a number open questions which need to be studied in the future. For one, it is unclear how much our interface design and the quality of the existing paper documentation influenced the results. We believe that our approach of taking the paper documentation 'as is' (except putting the relevant pieces in one document) and using it as 'blueprint' for the HMD interface is reasonable. However, it is conceivable that a more elaborate interface and/or documentation design might have lead to different results. Similar can be said about the way we use context and the choice of voice as 'non context' interface.

Another interesting question is how the results change when the subjects get used to the new modalities. Was the uneasiness towards context the usual result of the modality being new and strange, or was there something more fundamental? Finally, it is unclear if and how context interfaces can be adapted to help more proficient users, who seemed to benefit least in our study according to qualitative observations.

Acknowledgment

This work was funded by the European Commission through the WearIT@Work Project under grant 004216.

References

1. Boronowsky, M., Nicolai, T., Schlieder, C., Schmidt, A.: Winspect: A case study for wearable computing-supported inspection tasks. In: Fifth International Symposium on Wearable Computers (ISWC 2001), pp. 8–9 (2001)
2. Bristow, H., Baber, C., Cross, J., Wooley, S.: Evaluating contextual information for wearable computing. In: Proceedings of the Sixth International Symposium on Wearable Computers, pp. 179–185 (2002)
3. Caudell, T., Mizell, D.: Augmented reality: an application of heads-up display technology tomanual manufacturing processes. System Sciences (January 1992)
4. Cheverst, K., Davies, N., Mitchell, K., Friday, A., Efstratiou, C.: Developing a context-aware electronic tourist guide: some issues and experiences. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 17–24. ACM Press, New York (2000)
5. Dey, A., Abowd, G.: Towards a Better Understanding of Context and Context-Awareness. In: CHI 2000 Workshop on the What, Who, Where, When, and How of Context-Awareness (2000)
6. Drugge, M., Hallberg, J., Parnes, P., Synnes, K.: Wearable Systems in Nursing Home Care: Prototyping Experience. In: IEEE Pervasive Computing, pp. 86–91 (2006)
7. Hart, S., Staveland, L.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Human Mental Workload 1, 139–183 (1988)

8. Nilsson, S., Johansson, B.: Acceptance of augmented reality instructions in a real work setting. In: CHI 2008: CHI 2008 extended abstracts on Human factors in computing systems (April 2008)
9. Ockerman, J., Pritchett, A.: Preliminary investigation of wearable computers for task guidancein aircraft inspection. In: Second International Symposium on Wearable Computers, 1998. Digest of Papers, pp. 33–40 (1998)
10. Rhodes, B., Innovations, R., Menlo Park, C.: Using physical context for just-in-time information retrieval. IEEE Transactions on Computers 52(8), 1011–1014 (2003)
11. Schmidt, A., Beigl, M., Gellersen, H.: There is more to context than location. Computers & Graphics 23(6), 893–901 (1999)
12. Smailagic, A., Siewiorek, D.: Application design for wearable and context-aware computers. IEEE Pervasive Computing 1(4), 20–29 (2002)
13. Smith, B., Bass, L., Siegel, J.: On site maintenance using a wearable computer system. In: Conference on Human Factors in Computing Systems, pp. 119–120. ACM Press, New York (1995)
14. Starner, T., Schiele, B., Pentland, A.: Visual contextual awareness in wearable computing. In: Proceeding of the Second Int. Symposium on Wearable Computing, Pittsburgh (October 1998)
15. Sunkpho, J., Garrett Jr., J., Smailagic, A., Siewiorek, D.: MIA: A Wearable Computer for Bridge Inspectors. In: Proceedings of the 2nd IEEE International Symposium on Wearable Computers, p. 160. IEEE Computer Society Press, Washington (1998)
16. Webster, A., Feiner, S., MacIntyre, B., Massie, W., Krueger, T.: Augmented reality in architectural construction, inspection and renovation. In: Proc. ASCE Third Congress on Computing in Civil Engineering, pp. 913–919 (1996)