An Empirical Task Analysis of Warehouse Order Picking Using Head-Mounted Displays

Kimberly A. Weaver
GVU Center, School of
Interactive Computing
Georgia Tech
Atlanta, GA 30332
kimberly.weaver@gatech.edu

Hannes Baumann
TZI, Universität Bremen
Am Fallturm 1
D-28359 Bremen, Germany
hannes@tzi.de

Thad Starner
GVU Center, School of
Interactive Computing
Georgia Tech
Atlanta, GA 30332
thad@cc.gatech.edu

Hendrick Iben
TZI, Universität Bremen
Am Fallturm 1
D-28359 Bremen, Germany
hiben@tzi.de

Michael Lawo
TZI, Universität Bremen
Am Fallturm 1
D-28359 Bremen, Germany
mlawo@tzi.de

ABSTRACT

Evaluations of task guidance systems often focus on evaluations of new technologies rather than comparing the nuances of interaction across the various systems. One common domain for task guidance systems is warehouse order picking. We present a method involving an easily reproducible ecologically motivated order picking environment for quantitative user studies designed to reveal differences in interactions. Using this environment, we perform a 12 participant within-subjects experiment demonstrating the advantages of a head-mounted display based picking chart over a traditional text-based pick list, a paper-based graphical pick chart, and a mobile pick-by-voice system. The test environment proved sufficiently sensitive, showing statistically significant results along several metrics with the head-mounted display system performing the best. We also provide a detailed analysis of the strategies adopted by our participants.

Author Keywords

order picking, wearable computers, head-mounted display

ACM Classification Keywords

H.5.2 Information interfaces and presentation (e.g., HCI): User Interfaces–Evaluation/methodology.

General Terms

Experimentation, Human Factors, Standardization

INTRODUCTION AND RELATED WORK

The fields of wearable and ubiquitous computing have evolved from the creation of laboratory prototypes to examining systems deployed in workers' and consumers' everyday lives. In wearable computing, researchers focused on the tasks of inspection, maintenance, manufacturing, and repair as potential areas where wearable computing might prove beneficial [15, 13, 7]. Ockerman describes many of these applications as part of the task guidance domain [9]. She defines task guidance systems as "unsensored, computer-based systems which remind a user of the actions required to complete a task." As is often the case with new technologies, most user studies compare the status quo to the use of a new device or interface (see Siewiorek et al. [14] for an overview). More subtle effects, such as what might occur from a change in visualization or display modality, may be overlooked or go unreported as researchers attempt to show value in the new paradigm itself. Furthermore, practical application domains are often complex, causing more ecologically valid user studies to be relatively insensitive to interface variations. Ideally, we seek an experimental environment where a wearable computing or ubiquitous computing approach can demonstrate improvements over traditional methods and metrics and still have sufficient sensitivity to show the effect of changes to the interface. In order to interest industry, the application domain should be commercially relevant and of high value. As we discuss below, order picking meets all of these requirements.

About 750,000 warehouses worldwide distribute approximately $1 US trillion in goods [6]. Warehouses represent approximately 20% of the logistics costs for many businesses [6], and picking accounts for 55% [1] to 65% [3] of the total operational costs of a warehouse. Picking is the process of collecting items from an assortment in inventory and represents one of the main activities performed in warehouses. There are a wide variety of picking methods, ranging from fully automatic systems where thousands of objects are handled per hour to relatively infrequent picks performed by hand from an inventory shelf. According to an estimate by de Koster et al. [6], 80% of the warehouses in Western Europe are picked manually from storage racks or bins (low-level picking), where the picker moves among the parts

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CHI 2010, April 10 – 15, 2010, Atlanta, Georgia, USA
Copyright 2010 ACM 978-1-60558-929-9/10/04...$10.00.

(picker-to-parts) and picks multiple objects based on a particular order. Most research papers in the field focus on more automatic systems [6], possibly because they are more amenable to analysis. Yet, most picking is still done manually, presumably due to the cost and difficulty of making a robotic system that can handle the large variety of parts typical in such tasks. The focus in this paper is on the evaluation of a head-mounted display (HMD) system which can be easily integrated into current warehouse layouts without costly modifications to accommodate automated systems.

Typically, order picking begins with a paper picking list specifying the location of each type of item, the number of items to be picked, and sequence in which the items will be picked. A worker collects the items from stock and transports the items to a specific location for later delivery to a customer or to an assembly line. Errors in picking can jeopardize customer relations or stop an assembly line. Thus, while picking should be time efficient, it should also be accurate.

According to Tompkins [17], typically 50% of a picker's time is spent traveling, 20% searching, 15% picking, 10% in setup, and 5% performing other tasks. Research of manual picking systems focuses on optimizing travel time. Besides efficient path planning (which resembles the Traveling Salesman Problem [6]), orders requiring similar parts may be grouped together (proximity batching). Similarly, items that are normally picked together may be clustered on the shelves (family grouping). In order to avoid picker travel, automation may bring shelves of items to the picker based on the requirements of the order, resulting in a very small pick area. Here, we assume that the picker's travel time has already been optimized. Instead, we focus on optimizing the presentation of pick lists to improve setup, search, and pick times and accuracy.

Industry has created many approaches to assisting the order picker. Instead of paper lists, hand held mobile data terminals (MDTs) equipped with bar code scanners may be used to display the next pick and confirm a correct pick as it occurs. Alternatively, "pick-by-voice" wearable computer systems cue the picker as to the next pick and free the picker's hands for manipulating the items [16]. Such systems often use limited speech recognition for the picker to give commands such as next pick, repeat, back, or "empty" to indicate that the item was not where it was expected. "Pick-by-light" systems cue the picker as to the next pick with lights under each bin to be picked and under the appropriate order bin (if multiple orders are being picked at the same time). In some systems, proximity sensors or laser scanners sense a picker's actions and automatically proceed to the next pick or warn of an incorrect pick. Pick-by-light can also be implemented using projectors or lasers, though such systems can be difficult to deploy in practice as the picker may often obscure the beam while performing his tasks.

Researchers have begun to use HMDs to assist order picking. Reif et al. [11] report on an augmented reality (AR) "pick-by-vision" system that guides the picker to each item using arrows and "attention tunnels" overlaid on the user's visual field as they transverse the pick area. Unfortunately, the pick speed was only approximately 3.7% over a paper list and pick error improvement did not show a statistically significant result (probably due to the small number of errors observed). In addition, the AR system requires motion tracking hardware that is currently impractical to deploy in many environments. Iben et al. [5] compare a text-based paper pick list to a text pick list rendered on a HMD. Context sensing was shown to help limit pick errors. However, the main improvement in pick speed was attributed to the wearable computer's ability to order picks more optimally based on which set of shelves the picker was currently attending. Both of these HMD-based user studies, while attempting to replicate reasonable scenarios in industry, had many confounds that potentially limited the quantitative effects that could be attributed to the interfaces.

We created an experimental design that should be better suited to isolate the variables involved in order picking while still maintaining reasonable similarity to order picking environments we (and others) have observed in industry. To provide a baseline to non-technological approaches, we optimized the paper picking list, choosing a graphical representation that should speed the picker's identification of which items need to be picked. We compare this paper-based graphical representation to a HMD system using the same representation, a paper-based text list, and a pick-by-voice system. We present our results, discuss possible improvements to the study design, and suggest future work using the study design to compare other potential aids for order picking.

EXPERIMENTAL DESIGN

This section describes the layout for the warehouse used in the study as well as a general definition of the picking task. The four picking methods are described in detail.

Warehouse Layout and Task Description

The study took place on a simulated factory floor at a university in Germany in an area dedicated to warehouse picking. The warehouse environment in which the participants were working is shown in Figure 1. The layout consisted of two shelving units (A and B). Each row of a shelving unit housed three part containers and each shelving unit had four rows. The part bins were represented with a two digit number. The first digit indicated the row of the part bin (1 being the top row and 4 being the bottom row). The second digit indicated the position in that particular row (1 indicated the left side, 2 indicated the middle and 3 represented the right side of the row). A part with the number 31 would be the row one up from the bottom and on the left side.

A task consisted of picking the required parts for three orders at the same time. A pick is defined as reaching into a part bin and removing one or more parts from the bin. A place is defined as putting all of the items currently being carried into an order bin. An order was generated by randomly selecting four parts bins from a shelving unit. One part each was picked from three of those part bins, and the fourth bin required a pick of two of the same parts. Four picks were chosen because four independent items plus or

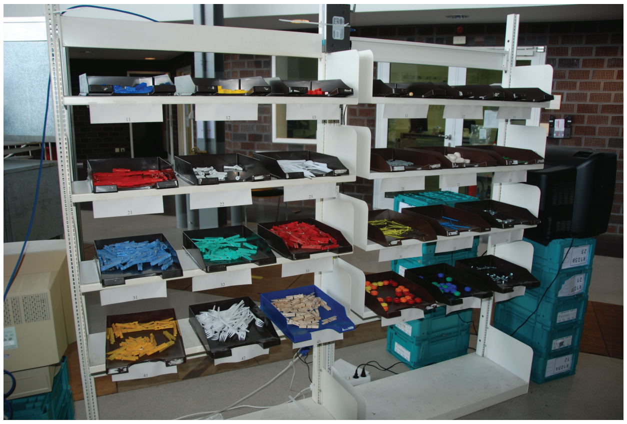


Figure 1: Arrangement of parts and shelves. Two sets of shelves, A and B each contain 4 rows of 3 part bins.

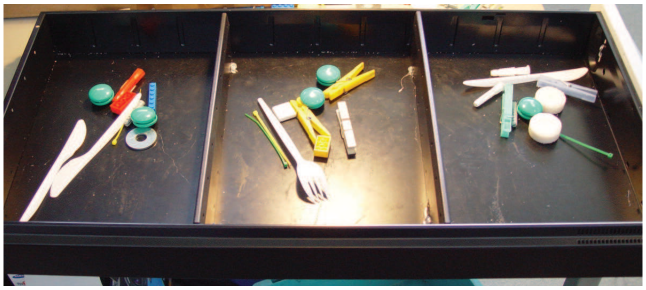


Figure 2: Parts in the three order bins from a completed task.

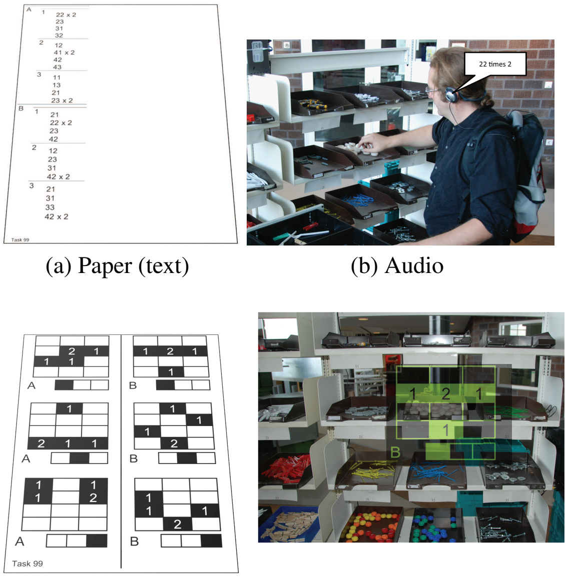
minus one is the number of unconnected things that a person can keep in short term memory simultaneously [2]. A person can reasonably remember all four parts by receiving the information once and then pick all of the parts from the shelf. The process of randomly selecting part bins was repeated for both shelves which resulted in a completed order being ten total parts. Each task required five parts from each shelf for each order. Figure 2 shows the parts collected for all three orders in a task. The parts were small enough so that the participants could hold all of the parts for a single order in their hands so that only one place per order was necessary at each shelf. Two shelves were employed instead of one in order to add complexity to the task and try to induce the participants to make errors. The next section will describe how the participants complete a task with each of the four picking methods tested in the study.

Picking Methods

Figure 3 shows the four picking methods tested in the study. Each image shows how information would be presented for the same task for easier comparison between the four picking methods. The completed result of the task represented in Figure 3 is depicted by the order bins in Figure 2.

Text-Based Paper Picking

The text-based paper picking method can be seen in Figure 3a. Using this method, participants were asked to retrieve a piece of paper from a plastic bin which contained a list of all



(a) Paper (text) (b) Audio

(c) Paper (graphical) (d) HMD

Figure 3: The four picking methods. The same task is displayed with each method. The audio and HMD versions only show what is accessible to the picker while filling part of order 1 on shelf B.

of the parts needed for a single task. Parts were first separated by shelf (A or B) and then by order number (1, 2 or 3). Within an order section, was a list of four part numbers. The part that needed to be picked twice was indicated with a "x 2" after the part number. Each order section was separated by a horizontal line. After completing the task, participants handed the completed parts list to the experimenter.

Audio Picking

In this method, participants wore a backpack containing a Sony Vaio UX ultra mobile computer. The computer was connected to headphones which provided the picking instructions. A Wizard of Oz approach was used for speech recognition. A human wizard listened for voice commands and initiated the appropriate computer response. Information was provided in a list manner much like that in the text-based paper method. In order to get the next line of instruction, participants were asked to say "okay". Upon starting a task, participants were told "Regal A (shelf A)". The system then went through the list of parts for shelf A order 1 individually. For the part that was picked twice, the participant was told the bin number followed by "mal zwei (times two)". Once all of the parts were picked for order 1, the participant was told "fertig eins (completed 1)". Upon completing the last order for shelf A, the participant was told "fertig 3; Regal B (completed 3; shelf B)". The audio method was literally a spoken copy of the text-based paper picking method with the exception that in the audio version, the order number was given just prior to placement in the bin instead of at the beginning of the picking. Participants were also allowed to repeat a command if they were unable to hear it by saying "noch mal (repeat)" or to step back to the previous command by saying "zurück (back)". In addition, participants were instructed that they could say “okay” in advance and in quick succession to avoid delays in picking. The audio picking method can be seen in Figure 3b. The instructions shown are for a single part on shelf B in order 1.

Graphical Paper Picking

The graphical paper picking method did not rely on part bin numbers to indicate the desired parts in a task. Instead, a grid consisting of 3 columns and 4 rows was displayed to represent the layout of the shelf. The bins to pick from were represented by black cells with a number inside indicating the number of parts to grab from that particular bin. Below the grid was a single row with three columns to represent the layout of the order bins. Again the black cell indicated the relevant order bin. This representation resulted in 6 images one for each order on each shelf for a single task. These images were arranged on a single piece of paper as seen in Figure 3c. The graphical representations for shelf A were in one column along the left side and the graphical representations for shelf B were along the right side of the paper.

HMD Picking

In this method, participants wore a backpack containing a Sony Vaio UX series ultra mobile computer. A monocular HMD device (MicroOptical SV-6) placed over the picker’s dominant eye was connected to the computer to provide the visual instructions. The HMD system repeated the representation of the graphical paper method but instead of showing all of the images for a task at the same time, participants were shown a single image for a single order on one shelf. To provide better perceptibility on the HMD, black was used as the background color (instead of white) and yellow as the foreground color (instead of black). Figure 3d shows the participant’s view for order 1 on shelf B. In order to see the next image, participants said “okay”. If the participant wished to go back in the task in order to correct mistakes and see previous images, they can say “zurück (back)”. Although the focus of the experiment is this final HMD picking method, it was important to include both graphical representations methods in the experiment to determine if any advantages found in the new HMD picking method were due to the graphical nature of the representation alone or due to other factors unique to the HMD.

Environment

Figure 4 shows the general layout of the people and equipment in the environment. Two video cameras were used to record the experiments. One camera faced the back of the shelves to capture the participant picking items from the part bins. The second camera looked along the shelves so that it had a view of the participant placing items into the order bins. Two monitors were used, one facing each camera. The monitors always displayed a running clock so that the videos for the two cameras could be synchronized with each other and with the logs from the computers in the experiment. The monitors displayed most of the information which was being saved into the logs during the course of the experiment to aid in synchronizing the video feed and the raw data in post-study analysis. In the case of the audio method, the monitor displayed text versions of what the participant was hearing as well as interactions from the user. For the HMD method, the monitors showed the participant’s current view in the HMD. For all four interaction methods, the monitors would show when the participant had placed parts in the order bins based on the wizard’s input. In Figure 4, the Camera 1 Monitor shows that the participant is working on task 93. He has just finished placing objects in order bin number 2. The paper list bin is where the participant retrieved the paper task forms for the text-based paper and graphical paper picking methods.

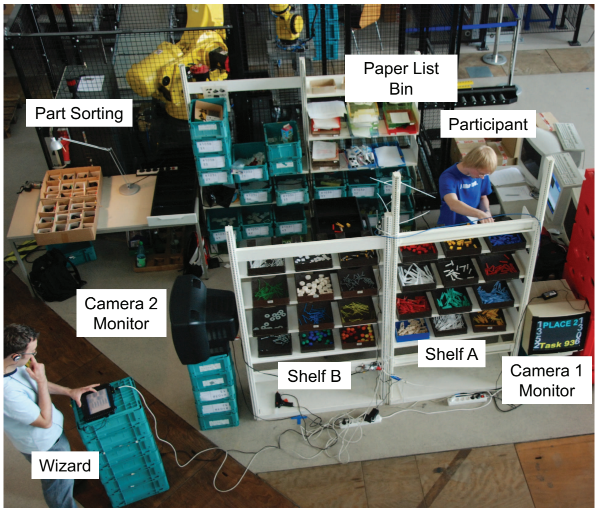


Figure 4: View of people, parts, and equipment in the experimental setup

Two researchers were required for this study. The wizard in the lower left corner of Figure 4 presses buttons on the tablet PC (a teXXmo Kaleo GX) to indicate when the participant has begun the task, placed objects in an order bin, or finished the task. For the audio and HMD methods, the wizard is also responsible for responding to verbal commands from the participant to initiate the proper response from the computer system. The second researcher is stationed in the part sorting area in the upper left hand corner of Figure 4. This person is responsible for taking pictures of the Camera 2 Monitor at the beginning of each task so that photographic data can be connected to its relevant task. Upon the completion of a task, this person retrieves the filled order bin and takes a picture of the parts inside for accuracy analysis. Parts were then sorted into their appropriate compartments in two trays for easy return to the part bins. The second researcher refilled the part bins between each method while the wizard debriefed the participant on the previous method and helped prepare them for the next method.

Method

Twelve participants (eight male, four female) were recruited for the study from a university in Germany. Ten participants were right-handed and two participants were left handed. All participants were tested for eye-dominance to determine placement of the HMD which only covered one eye. The participants held their thumbs out at arms length and closed

Die Aufgabe war leicht zu erlernen. *The task was easy to learn.*
Die Aufgabe war unangenehm auszuüben. *The task was uncomfortable to perform.*
Ich konnte die Aufgabe schnell ausüben. *I could perform the task quickly.*
Ich machte Fehler beim ausüben der Aufgabe. *I made mistakes while performing the task.*

Table 1: List of Likert scale statements.

one eye. If the position of the thumb moved relative to the background, then that eye is dominant. Ten participants were right-eye dominant, one participant was left-eye dominant, and one participant was uncertain but used the left eye for the purposes of the study. This proportion of right-eye to left-eye dominant participants is consistent with that of the general population [4]. While it is not certain that eye dominance will impact performance, some studies show there is potential impact [10]. All subjects were native German speakers. Although everything is described in English for the purpose of this paper, all instructions, interactions with the picking methods, and survey instruments were provided to the participants in German during the study.

Due to the participants' unfamiliarity with the four picking techniques to be tested and with warehouse picking in general, participants first completed a training phase. During the training phase, the experimenters explained each method in turn and allowed the participant to perform five tasks (involving a total of 120 picks) using each of the methods. The order of presentation of the four picking methods during the training phase was text-based paper, graphical paper, audio, and finally HMD. After completing the training sessions, the participants then began the testing session of the study. During the testing phase, the order the participant used each picking method was determined by the a balanced Latin Square. The balanced Latin square created four unique orders of presentation. By using twelve participants, we ensured that each order was used by three participants in the testing session and thus reduced ordering effects in the data. All statistics provided in this paper are derived solely from the order-counterbalanced testing phase. Participants performed ten picking tasks with each picking method during the testing session. Times were recorded for each interaction with the interface in the case of the HMD and audio picking versions as well as the start and end times of a task. After using each picking method, participants were asked to complete a NASA-TLX survey and to rate the learnability, comfort, speed, and accuracy of the method on a seven-point Likert scale (shown in Table 1). At the conclusion of the testing phase, participants were asked to rank the methods from best (1) to worst (4) based on overall preference, learnability, comfort, speed and accuracy. Accuracy data was also collected from photographs of the order bin after each task.

RESULTS

We designed the experiment with the paradigm of evaluating one new picking system (in this case the HMD-based system) in comparison with other methods. The HMD was evaluated against the two paper-based picking methods and the audio picking method using both performance and usability measures.

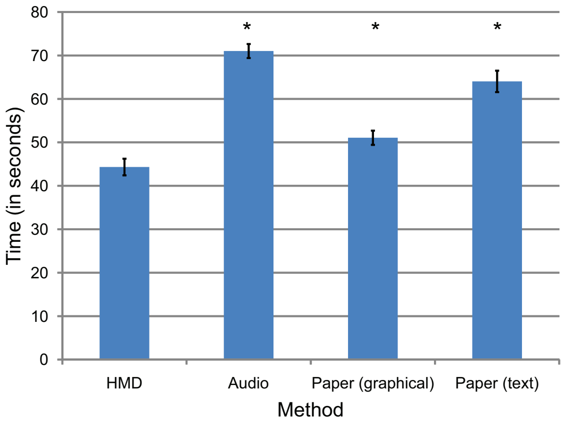


Figure 5: Time per task. A * indicates a significantly slower time than the HMD.

Performance Measures

Task Times

To achieve accurate task times, the start and end times extracted from the log files were verified and corrected by a self-written video annotation tool. For the paper-based picking methods, start time was defined as when the participant picked up the task paper. The start time for the audio picking method was defined as when the first instruction to pick a part was played. For the HMD-based picking method, the start time was defined as when the first shelf-order combination was displayed. For all methods, the end time was determined by when the last item was placed into the order bin. The average time per task for each of the picking methods can be seen in Figure 5. The error bars represent the standard error of the mean. A one-tailed paired samples t-test with Bonferroni correction for multiple comparisons was used to compare the average task time for each of the picking methods. The average time per task when using the HMD ($M = 44.33, SD = 6.63$) was significantly faster than the average time per task when using any of the other three methods: the graphical paper version ($M = 51.07, SD = 5.68), t(11) = 7.24, p < 0.05$ (one-tailed), the text-based paper version($M = 64.03, SD = 8.53), t(11) = 24.40, p < 0.05$ (one-tailed), and the audio version ($M = 71.03, SD = 5.59), t(11) = 14.43, p < 0.05$ (one-tailed).

Figure 6 shows the average time required for participants to complete the last eight tasks in the testing session for each of the four picking methods. Each of the lines is relatively straight indicating that by the last eight tasks the participant had reached a consistent performance level in each condition. Learning effects seem minimized.

Accuracy

Pictures were taken of the order bins after each task to evaluate per task accuracy based on number of substitutions, insertions, and deletions. Substitutions are when one part

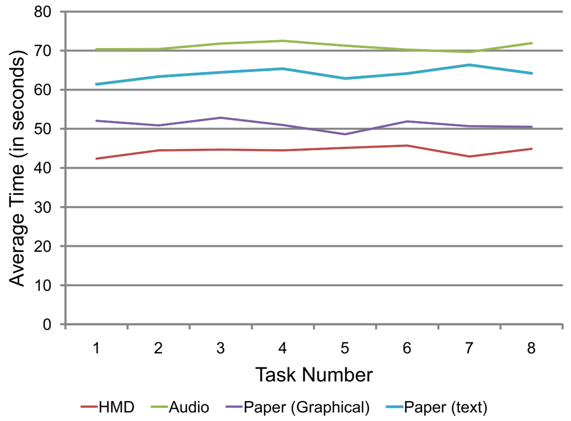
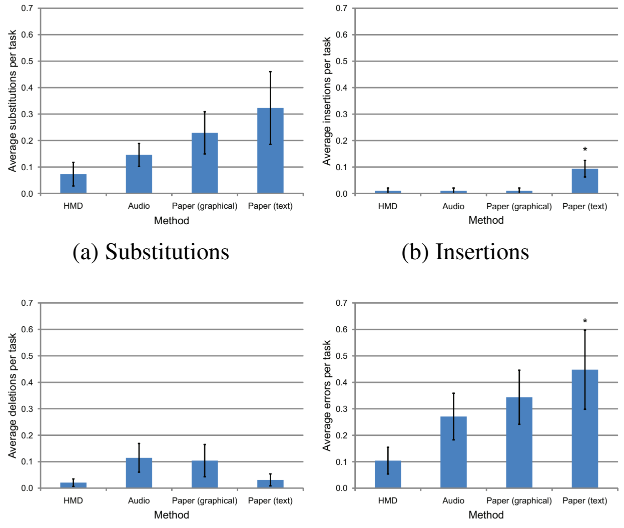


Figure 6: Comparison of the average time to complete each task per method

was swapped for another part, insertions are when an unrequested part was put in an order bin and all other requested parts were correctly picked, and deletions are when a part was forgotten and not replaced by another object. When an error was detected, it was confirmed through review of the video from the pick. This analysis helped determine the cause of an error. One common error was placing the items from an order into the wrong order bin. In the graphical picking methods, participants sometimes started picking from the wrong part bin and thus all of the subsequent picks were misaligned as well. In the audio picking method one participant would place parts from order 2 into the order 3 bin in shelf B and then skip order 3 completely. In some cases participants only picked one part instead of two from the bin where duplicates were required.

The total number of substitutions, insertions, and deletions in a task was combined to create a per task error value. A one-tailed paired samples t-test with a Bonferroni correction was used to compare participant's average per task accuracy based on substitutions, insertions, deletions and errors for all 4 picking methods. The HMD ($M = 0.010, SD = 0.036$) resulted in significantly fewer insertions than the text-based paper method ($M = 0.094, SD = 0.108$), $t(11) = -2.60, p < 0.05$ (one-tailed). With regards to overall errors, the sum of all insertions, deletions and substitutions, the HMD ($M = 0.104, SD = 0.175$) resulted in significantly fewer errors than the text-based paper method ($M = 0.448, SD = 0.518$), $t(11) = -2.45, p < 0.05$ (one-tailed). Figure 7 shows the comparison between substitutions, insertions, deletions and total errors across each of the four picking members. The error bars show the standard error of the mean. Based on Figure 7c, it appears that the text-based paper version also performs pretty well in reducing errors due to deletions.



(a) Substitutions

(b) Insertions

(c) Deletions

(d) Errors

Figure 7: Accuracy. A * represents a significantly higher number of errors than the HMD.

Measures	Picking Method			
	HMD	Audio	Paper (graphical)	Paper (text)
Overall	1.0	2.0 *	2.5 *	4.0 *
Learnability	2.5	2.5	2.0	4.0
Comfort	1.0	2.0	3.0	4.0 *
Speed	1.0	3.0 *	2.0 *	4.0 *
Accuracy	2.0	2.0	3.0	4.0 *

Table 2: Post-study rankings. A * indicates a significantly lower rank than the HMD.

Usability Measures

Post-Study Rankings

The median post-study ranks for overall preference, learnability, comfort, speed, and accuracy for all four picking methods is shown in Table 2. The ranks were compared using a Wilcoxon Signed Rank Test, the non-parametric equivalent of a paired samples t-test, with a Bonferroni correction for multiple comparisons. The HMD was ranked significantly higher overall than the other three order picking methods: audio, $z = -2.44, p < 0.05$ (one-tailed), with a large effect size ($r = 0.50$), graphical paper, $z = -2.86, p < 0.05$ (one-tailed), with a large effect size ($r = 0.58$), and text-based paper, $z = -3.21, p < 0.05$ (one-tailed), with a large effect size ($r = 0.66$). The HMD ($Md = 1.0$) was ranked significantly higher than the text-based paper version ($Md = 4.0$) with regards to comfort, $z = -2.92, p < 0.05$ (one-tailed), with a large effect size ($r = 0.60$). On the speed measure, the HMD method was ranked significantly higher than audio, $z = -2.39, p < 0.05$ (one-tailed), with a medium effect size ($r = 0.49$), graphical paper, $z = -3.28, p < 0.05$ (one-tailed), with a large effect size ($r = 0.67$), and text-based paper, $z = -3.15, p < 0.05$ (one-tailed), with a large effect size ($r = 0.64$). When asked to rank each of the methods in order of resulting accuracy, the participants ranked the HMD ($Md = 2.0$) better than the text-based paper version ($Md = 4.0$), $z = -2.93, p < 0.05$ (one-tailed), with a large effect size ($r = 0.63$).

Measures	Picking Method			
	HMD	Audio	Paper (graphical)	Paper (text)
Learnability	7.0	7.0	7.0	6.0 *
Comfort	6.0	5.0	5.0	4.0
Speed	6.0	6.0	6.0	5.0 *
Accuracy	4.5	5.5	4.0	3.0 *

Table 3: Likert scale responses. A * indicates a significantly lower (worse) score than the HMD.

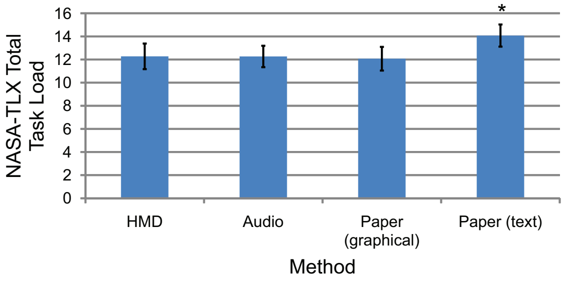


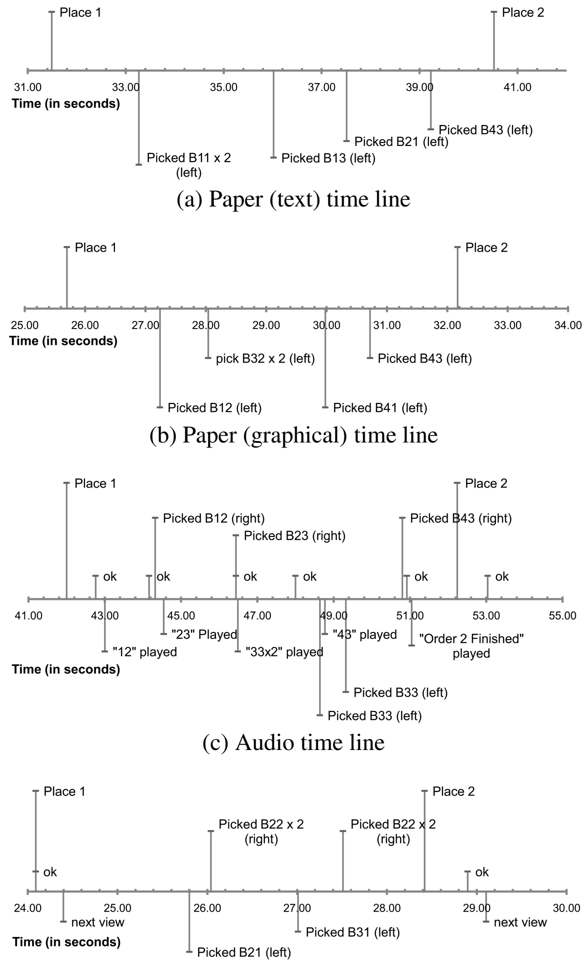
Figure 8: Overall task load

Picking Method Likert Scale Responses

Two of the Likert scale statements were positively worded and two of the statements were negatively worded. For the statistical tests in this paper, we flipped the responses for the negatively worded statements so that 1 is always the worst and 7 is always the best. A one-tailed Wilcoxon Signed Rank Test was used with a Bonferroni correction. The HMD received a significantly higher score ($Md = 7.0$) than the paper text version ($Md = 6$) with regards to learnability, $z = -2.16, p < 0.05$ (one-tailed), with a medium effect size ($r = 0.44$). The HMD ($Md = 6.0$) was also given a better score for speed than the text-based paper version ($Md = 5.0$), $z = -2.70, p < 0.05$ (one-tailed), with a large effect size ($r = 0.55$). On the accuracy measure, the HMD ($Md = 4.5$) was also given a better score than the text-based paper version ($Md = 3.0$), $z = -2.3, p < 0.05$ (one-tailed), with a medium effect size ($r = 0.46$). The median scores reported by the users for all parameters and all picking methods are shown in Table 3.

NASA-TLX

The NASA Task Load Index Survey (NASA-TLX) was administered after each picking method's testing phase. A one-tailed paired samples t-test with a Bonferroni correction for multiple comparisons was used to compare the overall task load for each method and each of the task load sub-scales. The total task load when using the head-mounted display ($M = 12.3, SD = 3.8$) was significantly lower than the total task load when using the text-based paper version ($M = 14.1, SD = 3.3), t(11) = 4.27, p < 0.05$ (one-tailed). The HMD did not show a significant improvement over the graphical paper method or the audio method with regards to the total task load. None of the other comparisons resulted in a significant improvement for the HMD. Figure 8 shows a graph comparing the overall task load.



(a) Paper (text) time line

(b) Paper (graphical) time line

(c) Audio time line

(d) HMD time line

Figure 9: Time lines for each of the four picking methods

Time Lines

The data from a participant who was comparatively fast for all of the picking methods was selected for an analysis of picking strategies. Figure 9 shows detailed time lines of all of the picks, placements and interactions for each modality for the first order from shelf B on one of the tasks. Figures 9a and 9b, which show the time lines for text-based paper and graphical paper, are highly similar. The participant only picks objects with his left hand. The paper task lists are being held with the right hand. The text-based paper time line (Figure 9a) shows that the objects were being picked at a fairly even rate indicating that it takes approximately the same amount of time to interpret the instructions and move to the next picking location. The graphical paper time line (Figure 9b) shows a more punctuated picking rate. The first two objects are picked, and then the second two. The participant may have used the graphical nature of the presentation to remember the first two picks because they were in the same column and then the second two picks because they were both in the same row allowing for faster picking.

The audio and HMD picking methods allowed for hands-free interaction with the order data. Correspondingly, participants used both hands for picking. Figure 9c shows the time line for the audio picking method. The timings for picking from the part bins and placing in the order bins are shown, as well as the voice commands from the participant and instructions from the audio device. The first thing that can be noticed is that participants do indeed use both their right and left hands with the audio picking method even though instructions are only being provided serially. The serial presentation of picking information is one disadvantage of an audio presentation of order information in a densely packed warehouse environment. If the participant needed to travel farther to the next picking location, there would be sufficient time to receive the instructions before the part bin is reached. This participant did use some optimizations to pick more quickly. The "okay" command was given slightly before the instruction for part B12 so that the next instruction could begin. The gap between "okay" command and pick was even larger after the instruction to pick part B33 twice. Unfortunately in this case, the participant was not as attentive to the quantity of the first command. The participant started by picking one object from bin 33 and then realized that two parts were required and had to quickly reach into the part bin again. There was also considerable delay between placing parts in the order bin and giving the command for the next instruction after order 1 and order 2.

Figure 9d shows the HMD time line. Here the participant actually says "okay" at the same time that parts are being placed in the order 1 bin. The participant is also alternating picking with right and left hand. As with the graphical paper picking method (Figure 9b), the participant is picking with a more punctuated rate: picking both objects from row 2 first, and then picking the two objects in column 1. The combination of being able to use both hands and always having the display visible makes picking faster than any of the other picking methods.

DISCUSSION

The HMD performed considerably better than the traditional method of text-based paper for many measures. In this study, we showed that the HMD resulted in significantly faster picking times not only for the text-based paper version, but also for the graphical paper and audio versions. Because the HMD was faster than the graphical paper version, we can see that it is not just the graphical presentation of the information in the HMD that results in faster picking. Another clear advantage for the HMD was that it allowed the pickers to use two hands to collect parts (Figure 10). When using the graphical paper version, only one hand could be used because the other hand was holding the paper. Some participants suspended the paper from the shelves while picking, but these participants also made many mistakes. The HMD display was also adjusted so that the part bins and the display were at the same focal distance, which meant that unlike with the two paper versions, in the HMD method participants could maintain constant focus.

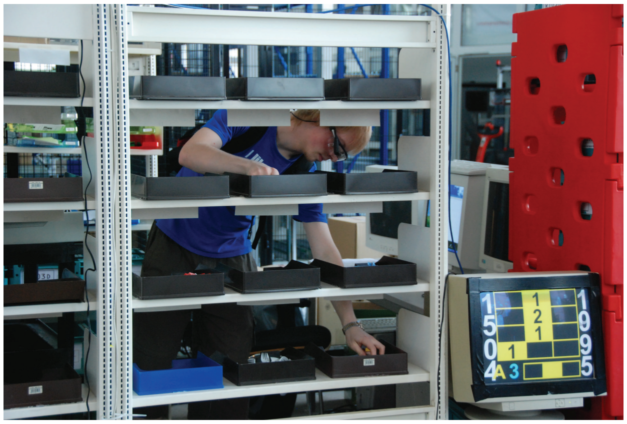


Figure 10: A participant uses two hands to collect parts while wearing a HMD.

The audio version was the slowest picking method. Our warehouse layout was not the best for testing the desirability of an audio interface in warehouse picking. If the parts had been distributed among multiple banks of shelves, the audio method would have had less of a disadvantage. Participants also had more opportunities to optimize their picking with the audio method, but did not take advantage of them. Participants could have requested information for the next part while in motion for the current part, but instead most waited until after the current part had been picked. There may be some concern over the possible processing delay in the audio interface due to the reaction time of the wizard with regards to the results in this study. The average time from user request to computer action in the audio method was 0.4 seconds. Processing time, while currently unknown, would also be required for speech recognition systems to interpret the commands. Because the participants did not take advantage of all of the optimization possibilities when using the audio picking method and the fact that the total duration of all audio prompts for a task without any delay between commands is 47 seconds, there was little chance for the audio version to outperform the HMD version in this study. A new study focusing on audio interfaces using the same experimental method described in this paper could be implemented in order to better understand the causes of delay with this picking method.

The participants in the study did not necessarily have previous experience with HMDs, which could have affected their reception to the novel device. According to the two measures of usability we collected, ranking and Likert responses, the HMD was not significantly harder to learn than any of the other methods. In fact, participants reported that the text-based paper version was harder to learn.

There were no significant differences between the HMD and the three other picking methods for any of the sub-scales in the NASA-TLX survey, but the HMD did show significant improvement over the text-based paper version in overall task load. This improvement is consistent with expectations. The audio method and the text-based paper method are the only two methods that require the participant to pay attention to the labels on the part bins. The reduction in task load for the audio method and not for the text-based paper version is most likely due to the audio method's presentation of only one part at a time. In the text-based paper version, the entire list of parts for the task is available simultaneously and the participant must keep track of what has already been picked.

The participants were fairly capable of evaluating their own performance when using the four picking methods. Participants felt that they were fastest using the HMD, and this method indeed proved fastest. The paper graphical method was predicted and shown to be the second fastest. The participants did feel like they were faster on the audio method than with the text-based paper method, when in fact the opposite was true. This conflict between perception and ground truth is a positive endorsement for the audio method because it indicates that participants did not feel like they were being slowed down while waiting for the audio instructions. Participants were also able to correctly evaluate their accuracy. Participants felt that their accuracy suffered more in the text-based paper version, and was also demonstrated in their actual accuracy scores. This consistency of user impressions and actual performance is important.

Evaluation of User Study Design

The study we created was sensitive to the differences in the four picking methods, more so than both Iben et al.[5] and Reif et al.[11] with regards to the time measure. The user study was less able to differentiate between the picking methods in terms of errors. One common error in warehouse picking is when the picker loses track of which shelf they are at, causing them to pick the parts from the wrong shelf. Some participants divided the parts they had picked from the shelves so that one shelf's parts were at edge of the order bin closest to them and the parts from the other shelf were at the farthest part of the order bin allowing them to keep track of which parts had been picked from which shelf and for which order. A possible modification to the experimental set up would be to provide smaller order bins which does not allow this division or to find new ways to make the task more difficult by incorporating more shelves.

The method of synchronizing the time with the computers for the picker, wizard, and for the display monitors was a success. This synchronization made it very simple to evaluate the data at the end of the study. The time stamp information in the logs and on the video feed was invaluable for consulting the video to verify inconsistencies with the accuracy data. It was also very important to always ensure that there were at least 20 parts in each of the part bins. The participant never had to struggle to pick parts from a particular bin and helped to guarantee that picking times during the beginning of a method's testing phase stayed consistent with the picking times at the end of the testing phase.

The user study was sufficient for discriminating between the four picking methods based on efficiency and usability factors. However, the sociality of the workplace and the interactions between other pickers in the environment play an important role as well [8]. It may be possible to extend the experimental environment described in this paper to incorporate multiple pickers and capture some of the effects of the social work environment. Other study designs would be necessary to investigate such factors as large scale deployment and effects of fatigue that may occur from long term use of a HMD. However, this experimental setup succeeds in allowing the interaction designer to accurately compare and discriminate among many task guidance systems simultaneously, something which may not be possible with a more ethnographic-centered or long term study design.

One advantage of the experimental protocol in this study is that it provides the researchers with a wide range of performance and usability data for task guidance systems. The incorporation of video cameras to record the experimental session allows for easy recovery of experimental data in the case that the logging mechanism in the computers fails. The experimental protocol has shown that the HMD allows for significantly faster picking times than the audio and text-based paper methods. It also showed that it is not just a matter of the novel graphical presentation, because the HMD was also significantly faster than the graphical paper picking method. The training session on each of the four methods allowed the participants to reach expert rates before evaluating the performance and usability of the methods. Usability data showed that the participant's views of the technology were well aligned with the evidence from the performance data. The creation of time lines from the log data allowed the experimenters to start to understand some of the behaviors that were likely with the different picking methods.

FUTURE WORK

The protocol and environment described in this paper can be easily extended to explore the effects of changes in the user interface for the HMD and audio picking method or for novel picking methods. One method which is beginning to be implemented in warehouses is pick-by-light. In a pick-by-light system, warehouse bins are wired with lights and sensors. As a picker walks around the shelves, bins near their current location will light up. In a pick-by-order system (such as the one described in this paper), after the part has been picked, another light will appear on the order bin where the part should be placed. A future study is planned using this experimental protocol to compare the pick-by-light method against an HMD along with the traditional text-based paper method.

This planned study can be further expanded by looking at the effects of adding context sensing to the environment. There are two ways that a warehouse inventory system can know that a part has been picked: either the picker has performed some action, such as saying a command or pushed a button on the part bin, or the system can use sensors to determine automatically when a part has been picked. Such fully instrumented systems (which provide more automation than defined by the term "task guidance systems") can be expensive. A study based on the described experimental protocol could help companies determine whether there is benefit to the added expense.

The experimental protocol will be useful not only in helping to improve industrial warehouse picking methods, but it can also be used as an evaluation task for wearable and ubiquitous computing technologies in general. The warehouse picking task can be easily standardized and replicated. The protocol described in this paper could be valuable for evaluating new sensing methods, graphical representations and visual displays. A first study could be a comparison between the HMD display used in this paper and the augmented reality (AR) pick-by-vision system described by Schwerdtfeger and Klinker and Reif et al. [12, 11]. The AR system provides visual wayfinding cues to guide the picker to the appropriate picking bin. This study could show whether the advantage in picking over traditional text methods is due to the visual nature of the display or due to other factors. A more interesting comparison between the HMD method in this paper and the AR method would be with regards to errors. Because the AR method guides the picker to the exact part bin, this might lead to less errors than the HMD method in this paper. In order to provide a possible contrast between the methods, the experimental protocol would have to be adapted to ensure that picker errors are more common.

CONCLUSION

In this paper we provided a comparison of various mobile computing interfaces and paper-based methods to support the task of warehouse picking. We articulated an empirically grounded exploration of how head-mounted displays can support such a complex real-world task. In addition we provided a detailed analysis of the motivations and strategies adopted by participants when using the various interfaces. The environment we devised to test the interfaces proved sufficiently sensitive, showing statistically significant results along several metrics including task time and overall usability. Picking times were significantly faster using the HMD over all of the other picking methods. The HMD supports graphical-hands free guidance which allowed users to combine the benefits of the graphical paper and audio picking methods in a single device.

REFERENCES

1. J. Bartholdi and S. Hackmann. Warehouse and distribution science release 0.89. Technical report, Georgia Institute of Technology, January 2009.
2. N. Cowan. The magical number 4 in Short-Term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(01):87–114, 2001.
3. J. Coyle, E. Bardi, and C. Langley. *The Management of Business Logistics: A Supply Chain Perspective*. South-Western College, Cincinnati, OH, 2002.
4. B. Crider. A battery of tests for the dominant eye. *The Journal of General Psychology*, 31:179–190, 1944.
5. H. Iben, H. Baumann, T. Starner, C. Ruthenbeck, and T. Klug. Visual based picking supported by context awareness: Comparing picking performance using paper-based lists versus lists presented on a head mounted display with contextual support. In *ICMI-MLMI*, Cambridge, MA, USA, November 2009. ACM.
6. R. de Koster, T. Le-Duc, and K. Roodbergen. Design and control of warehouse order picking: a literature review. Technical Report ERS-2006-005-LIS, Rotterdam, The Netherlands, January 2006.
7. D. Mizell. *Fundamentals of Wearable Computers and Augmented Reality*, chapter Boeing's wire bundle assembly project, pages 447–467. Lawrence Erlbaum & Associates, Philadelphia, PA, 2001.
8. M. J. Muller. Invisible work of telephone operators: An ethnocritical analysis. *Computer Supported Cooperative Work (CSCW)*, 8(1):31–61, Mar. 1999.
9. J. Ockerman. *Task guidance and procedure context: aiding workers in appropriate procedure following*. PhD thesis, Georgia Institute of Technology, Atlanta, GA USA, April 2000.
10. E. Peli. Visual issues in the use of a head-mounted monocular display. *Optical Engineering*, 29(8):883–892, 1990.
11. R. Reif, W. A. Günthner, B. Schwerdtfeger, and G. Klinker. Pick-by-vision comes on age: evaluation of an augmented reality supported picking system in a real storage environment. In *AFRIGRAPH '09: Proceedings of the 6th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*, pages 23–31, New York, NY, USA, 2009. ACM.
12. B. Schwerdtfeger, T. Frimor, D. Pustka, and G. Klinker. Mobile information presentation schemes for supra-adaptive logistics applications. In *Advances in Artificial Reality and Tele-Existence*, pages 998–1007. 2006.
13. D. Siewiorek, A. Smailagic, L. Bass, J. Siegel, R. Martin, and B. Bennington. Adtranz: a mobile computing system for maintenance and collaboration. In *International Symposium on Wearable Computers*, pages 25–32. IEEE Computer Society Press, 1998.
14. D. Siewiorek, A. Smailagic, and T. Starner. *Application Design for Wearable Computing*. Morgan Claypool, San Rafael, CA, 2008.
15. A. Smailagic, D. Siewiorek, R. Martin, and J. Stivoric. Very rapid prototyping of wearable computers: a case study of custom versus off-the-shelf design methodologies. *Design Automation for Embedded Systems*, 3(1):217–230, March 1998.
16. T. E. Starner. Wearable computers: No longer science fiction. *IEEE Pervasive Computing*, 1(1):86–88, 2002.
17. J. A. Tompkins, J. A. White, Y. A. Bozer, E. H. Frazelle, and J. M. A. Tanchoco. *Facilities Planning*. John Wiley and Sons, NJ, 2003.