Lauren Cairco Dukes
Graduate Research Assistant
Virtual Environments Group,
School of Computing,
Clemson University,
Clemson, SC 29632
e-mail: LCairco@clemson.edu

Amy Ulinski Banic
Department of Computer Science,
University of Wyoming,
Laramie, WY 82071
e-mail: abanic@cs.uwyo.edu

Jerome McClendon
Graduate Research Assistant
e-mail: jmcclen@clemson.edu

Toni Bloodworth Pence
Graduate Research Assistant
e-mail: tbloodw@clemson.edu

Virtual Environments Group,
School of Computing,
Clemson University,
Clemson, SC 29632

James Mathieson
Graduate Research Assistant
e-mail: jmathie@clemson.edu

Joshua Summers
Associate Professor
Mem. ASME
e-mail: jsummer@clemson.edu

CEDAR Lab,
Department of Mechanical Engineering,
Clemson University,
Clemson, SC 29632

Larry F. Hodges
Professor
School of Computing,
Clemson University,
Clemson, SC 29632
e-mail: lfh@clemson.edu

Evaluation of System-Directed Multimodal Systems for Vehicle Inspection

Multimodal systems have been previously used as an aid to improve quality and safety inspection in various domains, though few studies have evaluated these systems for accuracy and user comfort. Our research aims to combine our software interface designed for high usability with multimodal hardware configurations and to evaluate these systems to determine their user performance benefits and user acceptance data. We present two multimodal systems for using a novel system-directed interface to aid in inspecting vehicles along the assembly line: (1) wearable monocular display with speech input and audio output and (2) large screen display with speech input and audio output. We conducted two evaluations: (a) an experimental evaluation with novice users, resulting in accuracy, timing, user preferences, and other performance results and (b) an expert-based usability evaluation conducted on and off the assembly line providing insight on user acceptance, preferences, and performance potential in the production environment. We also compared these systems to current technology used in the production environment: a handheld display without speech input/output. Our results show that for visual and tactile tasks, benefits of system-directed interfaces are best realized when used with multimodal systems that reduce visual and tactile interaction per item and instead deliver system-directed information on the audio channel. Interface designers that combine system-directed interfaces with multimodal systems can expect faster and more efficient user performance when the delivery channel is different from channels necessary for task completion. [DOI: 10.1115/1.4023004]

1 Introduction and Motivation

Quality inspections during manufacturing are important for reducing costs and improving the safety and quality of the final product. Multimodal wearable computer systems have previously been used as an aid for inspection in various manufacturing domains, although aside from limited usability testing, little research has been conducted to experimentally evaluate these systems with respect to accuracy and user comfort [1–4]. Additionally, the software interface used for aiding these inspections may not fulfill certain usability and user experience goals, which in turn might hinder the inspection process. The aim of our research was to determine the relative performance benefits and design implications when combining a system-directed software interface and multimodal hardware configurations. Additionally, our research aims to determine accuracy and comfort benefits over existing systems in an industrial vehicle manufacturing domain.

The BMW Manufacturing Corporation, LLC., incorporates various technologies such as in-house software, handheld devices, and standard PCs to aid employees (also called associates) in the vehicle inspection process [5]. Vehicle inspection involves associates checking various items or areas on the vehicles for defects that should be fixed. If a defect is not found and fixed prior

Contributed by the Computers and Information Division of ASME for publication in the Journal of Computing and Information Science in Engineering. Manuscript received February 13, 2012; final manuscript received July 30, 2012; published online January 7, 2013. Editor: Bahram Ravani.

to vehicle delivery, the correction cost significantly increases. The accuracy goals expressed by BMW management were to reduce the number of defects that still exist when vehicle assembly is complete while still allowing associates to complete inspection within the time a vehicle passes through an inspection area along the assembly line. Initially, we visited the BMW manufacturing plant to observe the inspection line associates conduct vehicle inspection tasks with the current systems. The inspector must visually and/or manually check the fidelity or functionality of 12 to 15 parts in various locations on the inside and outside of each vehicle within less than 2 min while the vehicle moves along the assembly line. To record results, inspectors currently use either a stylus-based handheld device or a desktop computer located approximately 6 ft away from the side of the assembly line. The software interface for each of these devices is spreadsheet-based and the small font and lack of visual status indicators do little to aid the inspector in quickly and accurately entering data. These observations led us to address the issues of interface design and hardware configurations. We designed a new system-directed graphical user interface (GUI) for aid in inspection as well as designed two multimodal hardware configurations and suggested an alternative handheld configuration similar to the technology currently used in this setting [6].

In this paper, we present these three systems for inspecting vehicles along the assembly line, and the results from two system evaluations: (1) an experimental evaluation with novice users, or those with little to no industrial vehicle inspection experience, performing an abstracted inspection task, resulting in accuracy, timing, user preferences, and other performance results, and (2) an expert-based usability evaluation with BMW industrial vehicle inspectors, providing insight on user acceptance, preferences, and performance potential in the production environment. Our first hypothesis is that our two multimodal hardware configurations will provide performance benefits over the comparable handheld configuration. We also hypothesize that associates will report our new interface as an improvement over their current interface.

2 System Aided Inspection

2.1 Multimodal Systems for Aiding Inspection Tasks. Several researchers have successfully implemented multimodal wearable computer inspection systems. For example, the Mobile Inspection Assistance (MIA) bridge inspection system allows civil engineers to inspect a bridge while in the field. It supports hands-free interaction by allowing users to enter in data by taking pictures through a wearable camera and by speaking into a wearable microphone, and also provides a visual display on a computer worn around the inspector's waist [4]. Another example is Winspect, a wearable computer designed to support inspection of moving cranes while climbing them. In the traditional process, the user must inspect the cranes and remember their results until return on the ground, since it would be hazardous to carry anything during inspection. Instead, Winspect displays inspection items that are near the user's position on a head mounted display (HMD) and allows the user to select inspection points and enter results using a tracked glove [1].

Ockerman and Pritchett [3] used voice recognition in combination with a head mounted display to support pilots in preflight aircraft inspections. They implemented software that showed pilots one inspection item at a time and allowed pilots to use voice commands to mark items as checked and to review items that had not yet been checked. In a preliminary study, the authors found that there were no significant differences in the number of faults the pilots detected, whether they used the wearable computer or inspected the aircraft using traditional methods. The pilots liked the idea of a hands free checklist but recommended lighter equipment and allowances for customization since items were always presented in a specific sequence. Carnegie Mellon University, in collaboration with Bosch, built a wearable computer used for standardized car inspection in Germany [2]. A wearable system was evaluated as an alternative to a paper-based approach. The system included a wearable computer, a head mounted display, a handheld display, and a wireless adapter, and used speech recognition for input. The device was field tested by inspectors who were trained in performing the inspection as well as by students working on the project. The results showed that there was moderate to good acceptance of the system by the inspectors, although the inspectors thought the computer was too bulky and were worried about damaging it during inspection. Users did not like the head mounted display and instead opted to use the handheld display. The speech recognition performed sufficiently well even in noisy environments, although the learning curve for speech recognition was frustrating to the inspectors due to the long list of acceptable commands. No data on inspection accuracy or timing were reported.

More recent related work includes products of the WearIT@ work initiative, which is a project funded by the European Union with the goal of encouraging the integration of wearable computing in industrial settings [7]. Volkswagen's Skoda Auto division is a partner in this effort and has supported the development of wearable computers for assembly training as well as quality control during manufacturing [7]. For the assembly training application, sensors in a stationary vehicle and radio frequency identification (RFID) markers on tools allowed the system to be aware of the user's location and task. Users wore a textile keyboard, a data glove, and a Bluetooth headset with a microphone and headphone and could interact with the system through speech recognition or the textile keyboard. The system automatically recognizes the tasks that the user performs and offers real-time feedback and aid. It also allows a supervisor to monitor performance so that the supervisor can offer aid if necessary. The experimenters evaluated two different display options: a large screen or a HMD. While using the wearable system in comparison to paper-based training, users improved task performance but did not learn faster. The workers preferred voice-based interaction and preferred graphical information over text information and overall found the system useful. When comparing the display options, task performance was better using the large screen but most users preferred the head mounted display [8]. For the quality control application, Lukowicz et al. [7] mention that a system has been developed and data have been gathered at ETH Zurich; however, we were unable to find a published system description or data analysis [9]. Aleksy et al. [10] provides a review of other existing literature on wearable computers in industrial use, noting that many papers have identified the potential value of wearable systems but few provide comprehensive evaluation and results.

2.2 Current Systems for Aiding Inspection at BMW. To record inspection results along the assembly line, currently BMW employs one of two systems for each inspector: a handheld device or a standalone PC. Both of these systems require the user to initially scan a bar code appearing on a paper version of the inspection task list, which contains 12–15 parts to be inspected on the vehicle. We observed that inspectors read the items from the paper list instead of the screen, inspected the car, and then went to the PC or handheld device to record the inspection results. It seemed that users attempted to remember all 12–15 items to inspect and their inspection results until inspection was complete, then recorded them into the input device.

The handheld system requires stylus input into a checklist-like interface. Based on a preliminary study, we determined that its interface was problematic for several reasons. The users needed two hands to operate the device, so they could not properly perform any bimanual inspections while holding it. The text was small and difficult to read, and the target buttons were too small to be easily tapped by the stylus [6]. The standalone PC system required keyboard and mouse input and presented the inspection items in a small font within a spreadsheet-like interface. This

Fig. 1 Display for multimodal configurations using an abstracted task list. The display is rendered on one of three hardware configurations and is paired with audio, touch, and/or speech recognition.

alternative was also problematic: the stationary location of the computer discouraged inspectors from entering their inspection results one by one as they moved along the assembly line, since traveling back and forth to the computer between each checkpoint would not allow for inspection completion within time constraints. Additionally, based on traditional human–computer interaction principles, we determined that neither interface provided useful interaction feedback or affordances [11].

In an initial field study, we observed associates using each of these alternatives and gathered their opinions about the systems [6]. We based our new systems' design on their input and the recommendations of management. We chose two primary goals for our system design: (1) to enforce a systematic check on each of the items in the list and (2) to facilitate mobility so that inspectors can check items and enter inspection results concurrently. Other considerations in system design included preserving associate safety, providing meaningful interaction cues, and ensuring user comfort. We incorporated many aspects of other inspection systems including software-directed inspection, speech recognition, and augmented reality hardware, as well as support for mobile inspection. Additionally, we present performance and user preference evaluations for our software and its implementation on three hardware alternatives.

3 System Design

Our multimodal interface aims to enforce systematic checks by presenting the items to be inspected for each vehicle in the order that the items would appear as the inspector moves around the car and by requiring a response for each checkpoint before advancing to the next checkpoint. This system-directed design was recommended by BMW management to encourage inspection of every checkpoint on the list. We incorporated solutions into the design to improve readability, to permit user-directed flexibility, and to easily correct errors.

The main GUI loads in the checklist of 15 parts on the vehicle designated by the initial configuration data. In order to reduce cognitive demand in relation to the number of items the inspectors have to remember on the list, the user interface displays a maximum of five items at a time [12]. The five items displayed are the item most currently inspected, the item currently under inspection, and the next three items to be inspected (Fig. 1). When the software begins, only the first four items to be inspected are displayed. Yellow, red, and green boxes, respectively, indicate items that have not been inspected, items that failed inspection, and items that passed inspection. Although we chose these colors to enhance understanding and accuracy, it is possible that this may introduce difficulties for individuals who are red/green colorblind. In the future, we intend to incorporate user profiles allowing for individual color selection. All items are displayed in large, centered text for readability, and the box and text for the item currently under inspection is larger than the others. A progress bar is displayed at the bottom of the screen as an additional visual indicator of the inspection progress and results (Fig. 1). As the inspector proceeds, the rectangles composing the progress bar are shaded with colors corresponding to the inspection result–red for failed inspection or green for passed inspection. Our design is simple in order to help inspectors focus on their task instead of distracting them with unnecessary visual information.

Audio indicators are also provided to communicate inspection status. When it is time for an item to be inspected, the item is read to the inspector through text-to-speech and he/she records the inspection result using a keyword. Sound effects confirm whether the inspector's speech was recognized. The inspector speaks keywords to enter inspection results or change program status. The inspector can disable speech input by saying "pause" and enable speech input by saying "resume," and can say "start" to begin inspection and "stop" to complete inspection. The inspector says pass or "fail" to indicate checkpoint status during inspection. As a result, each item of the checklist moves up one position on the screen and the box of the currently inspected item changes color to indicate its inspection status (Fig. 1). The inspector can also say "back" to return to the item most currently inspected, allowing the inspector to correct a mistake by removing inspection results for that item and resetting it to a "not yet checked" state. Although we anticipate mistakes to typically be corrected before another item is inspected, the inspector may say back multiple times to move backward through the list of checked items. Giving the user flexibility in checkpoint ordering, the inspector may say "skip" to delay inspection of the current item until the end of inspection, placing the current item at the end of the list and advancing to the next item to be inspected. Using these keywords, the inspector may advance through the list. Although there is no built-in aid for helping users remember keywords, BMW associates indicated that pass and fail were commonly used terms to indicate item status and should cause little additional memory load for inspectors. The other keywords are less familiar but are also used much more infrequently during inspection. In the future, we could allow inspectors to save profiles with their own sets of appropriate keywords to further reduce memory load or could add a "help" keyword to show a menu with keyword options.

Our multimodal software was implemented as a Windows dialog-based application. Graphics were rendered full screen at an 800×600 resolution using OPENGL API. Text-to-speech was generated using Microsoft SAPI, and we used Dragon Naturally Speaking for speech recognition. Further details on implementation and design justification can be found in our previous work [6].

3.1 Hardware Configurations. For prototyping purposes, we chose three hardware configurations: handheld (H), large screen display (L), and monocular display (M). We implemented

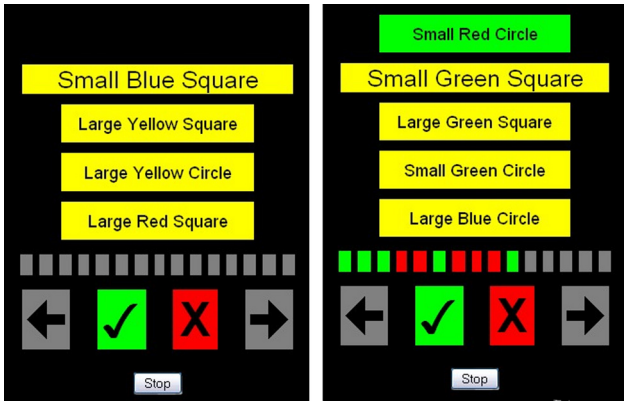

Fig. 2 Visual display for handheld configuration using an abstracted task list. Left: the handheld GUI when the application starts. Right: the GUI after several items has been checked off. The most recently checked item is green because it passed inspection, and the progress bar is color coded according to the inspection results of all items checked so far.

the software for monocular and large screen configurations on an Asus EEE OC 1005HAB PC laptop and used a Palm Pre device for the handheld configuration.

3.1.1 Handheld Configuration (H). For our evaluation, we implemented an alternative unimodal implementation of the software for a handheld configuration on a Palm Pre cell phone. Although this alternative is not multimodal, we chose to configure a handheld system to act as a hardware baseline by closely matching the touch-based interaction type that is currently utilized by the handheld devices which BMW uses to aid in vehicle inspection. We aimed to improve usability by implementing the interface on a device that can be used with one hand. The Palm Pre has a 3.1 in. multitouch screen with a 320×480 resolution. Due to the limitations of this device, we made a separate web-based implementation of our software interface. Speech input and audio output were not permitted in this application in order to closely match the current BMW touch-based system. Instead, the interface has been modified for best interaction using touch-based icons. There are four buttons at the bottom of the screen: a left arrow provides the same functionality as the back keyword, a right arrow provides the same functionality as the skip keyword, a check mark on a green background represents pass, and an "X" on a red background represents fail. Buttons begin and end inspection (Fig. 2).

3.1.2 Large Screen Configuration (L). The large screen configuration uses a RCA 52 in. flat panel LCD television for display and an Andrea Bluetooth headset for audio output and speech input (Fig. 3(*a*)). Since we use the same laptop for the monocular and large screen configurations, the screen resolution is restricted to 800×600, preserving the proportion of the display but tailoring the size of the interface to effectively utilize space.

3.1.3 Monocular Display Configuration (M). The monocular display configuration uses a monocular display with the laptop worn by the user and an Andrea Bluetooth headset for audio output and speech input. We modified a hunter's vest to hold the monocular device connector, the small computer, and the cabling on the user's back. Velcro straps hold the laptop and device connector against the back of the vest, which has a zippered fabric cover over it. Cabling can be placed in the back pouch of the vest (Fig. 4). The entire assembly weighs approximately 5 lb. An eMagin z800 3DVisor with a 800×600 resolution and a right eye piece is used for display. Through the optics, the display is

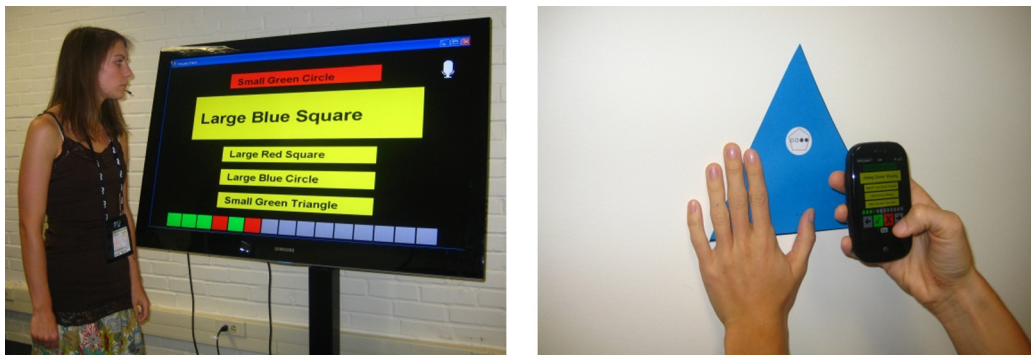

Fig. 3 (*a*) Large screen configuration and (*b*) one-handed touch for an abstracted task list using the handheld device

Fig. 4 (*a*) and (*b*) Monocular display hardware configuration; (*c*) example of a two-handed touch using the monocular display configuration

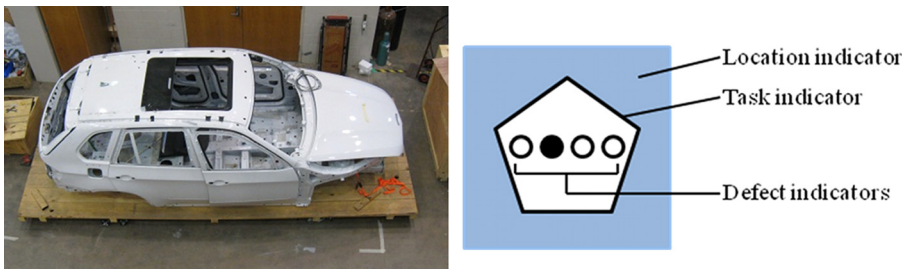

Fig. 5 (*a*) Vehicle body used for experimental evaluation. (*b*) Example of shapes used for abstracted inspection task. This item would be called "Small Blue Square," would require a one-handed touch for the inspection action since the shape is a pentagon and would pass inspection since an odd number of dots are shaded.

equivalent of a 105 in. screen viewed at 12 ft. Design considerations make our software suitable for monocular displays [6].

4 Experimental Evaluation With Novice Users

In order to evaluate the usability of our hardware choices for factory vehicle inspections and the effectiveness of our software, we conducted a user study for the performance evaluation of an abstracted inspection task using each of the three hardware configurations. The evaluation consisted of a 3×3 within subjects design, where participants performed three inspection tasks utilizing each of the three hardware configurations (H, L, and M), for a total of nine inspection tasks. Presentation orders of the hardware configuration conditions were created using balanced Latin squares, and participants were randomly assigned to one of the following six orders to use the inspection devices: (1) LMH, (2) MHL, (3) HLM, (4) HML, (5) LHM, (6) MLH.

4.1 Setup and Apparatus. We conducted this evaluation on Clemson University campus in a shared workshop in the mechanical engineering building. In one area of the workshop, there is a BMW X5 car frame on a slightly elevated platform for student experimental use with sufficient space to walk comfortably around the vehicle. We chose 20 locations on and inside the passenger area of the vehicle as inspection points and affixed Velcro markers to those locations to attach inspection items. We placed the large screen TV 4 ft from the passenger side and centered horizontally on the vehicle. We marked off a perimeter with tape on the floor around the car to designate a safe space for the participant to move in and also marked the starting location on the floor for beginning inspection 4 ft from the back passenger side door.

We used a Wizard-of-Oz approach for speech input since it was not the goal of this evaluation to measure speech recognition accuracy. We listened to their speech through transmitted Bluetooth and controlled the software interface through Virtual Network Computing (VNC)[1] by pressing keys corresponding to keywords.

4.2 Task Design. Since this experiment was conducted using novice participants without vehicle inspection experience, we abstracted the inspection task. Three concentric geometric shapes comprised one symbol that simulated a checkpoint, called an "inspection item marker" (Fig. 5(*b*)). The inspection item markers were used as indicators for the location and status of the item and how to perform inspection. The outermost shape, called the "location indicator," matched a shape description on the checklist. Location indicators had three parameters: size (large, inscribed inside an 8.5×11 in. piece of paper, or small, inscribed inside a quarter-sheet of paper), color (red, yellow, green, or blue), and shape (triangle, circle, or square). For example, through text-to-speech and/or screen output, a participant is informed that the current item to check is a "Large Red Triangle." The participant would then find that shape on the car. No two location indicators were the same on any checklist so that the location of the current checkpoint was unambiguous. Once the participant found the shape, the participant would then look at the shapes within the location indicator. The next concentric shape, called the "task indicator," indicated how the participant should inspect the item. Depending on whether the shape is a square, pentagon, or hexagon, the participant should either look at but not touch the marker, touch the marker with only one hand, or touch the marker with two hands, respectively. Finally, within the task indicator, there was a set of four dots in a row called the "defect indicator." The shading of these four dots indicated to the participant whether an item had a defect. An odd number or zero shaded dots indicated an item should pass inspection while an even number of shaded dots indicated a defect, which should fail inspection.

In designing the abstracted task, our goal was to simulate the cognitive load and time requirements for actual vehicle inspection. We chose three-word item identifiers to roughly match the phrase length of items that occur on a typical inspection checklist and the time it would take text-to-speech to read them. Requiring the user to look at shapes to determine what action to take, and whether the item passed inspection, simulated the time and cognitive load it would take an inspector to determine the status of an actual inspection item, since in real vehicle inspection an associate must look at a part, recall how to inspect it, and then determine whether the part passes or fails inspection. Inspection item markers were placed throughout the vehicle frame to simulate the various checkpoint locations in a vehicle. The percentage of defective items per checklist approximated the defect rate reported through actual vehicle inspection at BMW. Since inspectors are provided with reference material at their stations should they forget how to inspect a particular checkpoint, we provided our participants with a lanyard holding a reference sheet for the meanings of the task and defect indicators.

For each trial, there were 20 inspection item markers placed on the vehicle, providing the 15 items on the checklist plus five items as distracters, since expert inspectors would not inspect each item present on a vehicle. The participant was presented with one checklist of 15 items, ordered by their location moving counterclockwise around the vehicle. This simulated the actual inspection process, since an inspector is directed to check only certain vehicle features. Participants were given 106 s to complete the checklist, based on the time frame standard for factory inspections. If the participant finished before time was up, he or she said stop to complete the trial. If the participant did not finish in time, the system automatically stopped accepting input. We created unique checklists for ten trials. The first checklist was used for a practice trial before using any devices for input. In this first trial, the participant inspected all 20 checkpoints with the help of an experimenter to help familiarize them with the inspection item markers and the locations of the checkpoints on the vehicle. The user then completed nine trials, with three trials per device.

4.3 Measures. For each participant, we gathered data through pre- and postquestionnaires, a debriefing interview,

[1]RealVNC: http://www.realvnc.com/.

experimenter logs, and software logs. The prequestionnaire gathered information about demographics, level of use of various hardware configurations, vehicle knowledge, and learning styles. We recorded the number of items inspected and not inspected, the number of items inspected correctly and incorrectly, the time it took for inspection completion, and each voice command the user spoke. While the participant conducted an inspection, an experimenter marked each checkpoint on a clipboard to indicate if the user had inspected the correct item with the correct action (look, one-hand touch, or two-hand touch). Finally, in the postquestionnaire and debriefing interview, we asked questions related to usability, preferences, and effectiveness of the interfaces and hardware configurations.

4.4 Results. Of 25 participants from Clemson University, there were 10 females and 15 males, aged 18–53 (mean = 23). Participants rated themselves as having low ($N = 12$), average ($N = 8$), and high ($N = 5$) levels of vehicle knowledge. Twenty-four participants were college students and one participant was a postdoctoral fellow.

Since the time to complete each trial was limited to 106 s, many participants did not complete all trials for an input device. Overall, 13 participants completed all three trials using the handheld configuration (H), 14 participants completed all three trials with the large screen configuration (L), and 18 participants completed all three trials with the monocular display configuration (M). The task performance data were treated with a repeated measures 3×3 analysis of variance (ANOVA) to test for the within subject effects of hardware configurations and the within subject effects by trial. Data reported from the postquestionnaires were analyzed using the Chi-square test. The F and χ^2 tests that are reported for analysis used an alpha level of 0.05 to indicate significance.

4.4.1 Accuracy. The accuracy percentage of correctly checked items was determined by dividing the number of items that were correctly checked by the total number of items checked in each trial. Correctly checked items are those that the participant both performed the correct inspection action and reported the correct inspection result. There was no significant main effect of hardware configuration type for mean accuracy percentage of correctly checked items, $F(2,38) = 1.58$, $p = 0.22$, $n^2 = 0.08$. All configurations allowed for high accuracy with monocular display (M) as the highest and handheld (H) being the lowest. There was no significant main effect found among the sets of the trials or interaction effect of device by trial.

The defect detection percentage was determined by dividing the number of defects that were correctly detected by the total number of defects for each trial. Since the number of defects varied over each trial, the total accuracy of defect detection for each trial was averaged across trials and participants, and then analyzed using a one-way ANOVA. There was no significant main effect found for the defect detection accuracy nor any significant interaction effect of device by trial. However, all configurations allowed for high accuracy as listed from highest to lowest: monocular display (M), large screen (L), and handheld (H). No significant differences were found for accuracy grouped by vehicle knowledge, device usage, or learning style. Unfortunately, no accuracy data are recorded for BMW inspectors, so we could not compare our accuracy results to the baseline accuracy achieved in the manufacturing environment.

4.4.2 Task Completion Times. Analysis revealed a significant main effect for hardware configuration type for overall task completion time. The handheld configuration (H) allowed for significantly faster overall completion time than the monocular display configuration (M) and the large screen configuration (L) (Table 1). In addition, participants' overall performance became significantly faster by the third and last trial, $F(2,38) = 4.38$, $p = 0.019$, $n^2 = 0.19$. There was no significant interaction effect of hardware configuration by trial.

A few participants discontinued the inspection task accidentally, indicating that participants were having difficulty or accidentally executed a command. A participant possibly discontinued the task accidentally if the overall task completion time was less than 105.5 s (due to rounding error) and did not check all 15 items on the list. There were eight accidental discontinuations for the handheld configuration (H) possibly due to difficulties with the touch screen interface, while there were no accidental discontinuations of the task for the monocular (M) or large screen (L) configurations, likely due to the Wizard-of-Oz setup. As a result of several participants not completing the full trial, it may be more informative to analyze completion time per individual item. This was calculated as a result of each participant's overall time divided by the number of items each participant actually inspected. We did not find a significant main effect for hardware configuration type for task completion time per item, $F(2,38) = 1.83$, $p = 0.18$, $n^2 = 0.09$. However, the level of power for the statistical test dropped from 0.98 to 0.36, so the trend indicates that this result is likely to become significant with more trials. The completion times per item for the hardware configurations are listed from fastest to slowest: large screen (L), handheld (H), and monocular (M). Additionally, there was no effect by trial, $F(2,38) = 1.08$, $p = 0.35$, $n^2 = 0.05$. The completion times per item for the hardware configurations are listed from fastest to slowest: large screen (L), handheld (H), and monocular (M). To address the concern of the possibility of skewed data due to participants accidentally executing commands or discontinuing, we conducted an analysis without data from participants who were experiencing difficulty. We found no significant differences. Therefore, our conclusions on the differences between the hardware configurations are valid even with participants having some difficulty.

4.4.3 Postexperimental Questionnaire. Based on participants' self reported data, the large screen configuration (L) was the

Table 1 User preference results for the experimental evaluation. Configurations with the highest percentage are unshaded, configurations with the lowest percentage are shaded dark gray, and configurations with a middling percentage are shaded light gray. * denotes a significant difference.

	Handheld (%)	Large screen (%)	Monocular (%)	$\chi^2(2, 25)$	p
Easiest to use	28	56	16	6.32	0.04
Most useful for task	36	36	28	—	—
Best facilitated in task	16	*68	16	13.52	<0.001
Helped complete task most	40	*52	8	10.16	0.01
Most comfortable	36	*64	0	15.44	<0.001
Most awareness of checklist	24	20	*56	5.84	0.05
Helped remember more of checklist	36	48	16	—	—
Best task performance	32	48	20	—	—
Fewest mistakes	12	52	36	—	—
Best for proper inspection actions	8	*60	32	10.16	0.01
Best completion time	36	*56	8	8.72	0.01

Table 2 Performance results for the experimental evaluation

	Handheld		Large screen		Monocular			
	Mean	SD	Mean	SD	Mean	SD	Main effect	Interaction
Checkpoint accuracy	99.89%	0.00	99.55%	0.01	98.7%	0.04	No, $p = 0.34$	No
Defect detection accuracy	91.67%	0.29	99.33%	0.18	100%	0.00	No, $F < 1$	No
Overall completion time	93.03 s	19.81	102.20 s	7.13	101.84 s	6.80	Yes, $p < 0.001$	No
Completion time per item	10.62 s	9.54	8.06 s	2.06	9.30 s	4.03	No, $p = 0.18$	No

easiest to use, best facilitated in task, most comfortable, best helped to properly inspect the items, and allowed for the best completion time (Table 2). The large screen configuration was reported as allowing for the best performance on the task, though the difference was not significant. The monocular display configuration was reported to allow for the most awareness of the checklist significantly more than the other configurations. Participants felt that all three configurations were equally useful for the task overall. Participants ranked the large screen configuration (L) as their first choice significantly more than others, $\chi^2(2, 25) = 12.08$, $p = 0.002$ and monocular display configuration as their third choice significantly more than others, $\chi^2(2, 25) = 10.64$, $p = 0.005$. Participants reported that speech was the most useful method of input, $\chi^2(2, 25) = 15.08$, $p = 0.01$, and the monocular display with audio was the most useful method of output, $\chi^2(2, 25) = 13.76$, $p = 0.001$.

4.4.4 Participant Comments on Hardware Configurations. Overall, participants preferred the large screen configuration:

- "The TV facilitated… the most because I did not have to carry anything or close my eye to see the instructions [in reference to the monocular display]."
- "It eliminated the need to have any extra gear on you and was hands free."

However, participants appreciated aspects of the other configurations and made recommendations for how to improve them. Participants liked the convenience of the handheld device, with one participant commenting that, unlike the other alternatives, "it did not get in the way of your vision or range of motion and you didn't have to turn back to look at it." However, other participants acknowledged the inconvenience of having to carry the device, saying that "the handheld was hard to use two hands at once, so I had to put it down, but at times there was nowhere to put it," and suggesting that "it needs to be on a necklace." Additionally, participants commented on the stop button being too close to the other buttons on the touch screen. Participants also liked the convenience of the monocular display but were concerned about adverse effects of wearing the display, saying that it "hindered [their] vision," "hindered depth perception, making it hard to touch what you need," and "gave [them] a headache… I couldn't see the screen unless I squinted uncomfortably."

4.4.5 Discussion on Experimental Evaluation. None of the three configurations seemed to provide significant benefits in terms of inspection accuracy, although each configuration supported a high accuracy rate (98.7% or greater) and a high defect detection accuracy rate (91.67% or greater). Although the handheld configuration (H) provided a significantly lower time to task completion, all three configurations allowed for inspection completion within the allotted time frame, suggesting that any of our configurations are suitable for use from an accuracy and time perspective. Based on participant rankings and comments on usability, it seems that the large screen (L) configuration was overall the most preferred, possibly due to the inclusion of visual and auditory output and speech input without any cumbersome equipment to wear. Additionally, participants may have already been familiar with using large screen televisions and Bluetooth headsets, making interaction natural and reducing any learning curve. Using unfamiliar equipment such as a monocular display may have been more difficult for participants to adjust to within the trial time.

Based on participant comments and our observations, there are several modifications we can make to our interfaces to further improve usability. One simple correction is to move the stop button on the handheld device away from the other buttons so that it is not selected accidentally or to provide an option for returning to the inspection in case the participant accidentally presses the stop button. We could also consider some kind of necklace, holster, or wrist mount for the handheld display so that the inspector could have both hands free while inspecting, even if they were using the handheld device. Improvements for usability of the monocular display configuration might include identifying and using a less cumbersome monocular display or a see-through display to support better depth perception.

Additional research should be conducted to determine whether our experimental results extend to performance in a manufacturing setting. Based on our observations of assembly line operations, we attempted to mimic the inspection process with our abstracted inspection task. However, comparison data for task performance and workload in the actual vehicle inspection setting were not available.

5 Usability Evaluation With Expert Users

5.1 Preliminary System Modifications. We made several changes to our software and hardware systems based on feedback from BMW associates after they participated in a preliminary offsite demonstration of our system in order to make it more usable for associates and more suitable for use in the BMW plant setting. We enabled additional options for checkpoint status: "fixed," which means that an inspector found and corrected a defect, and "found," which means that the item was originally defective but was corrected by another associate. The voice keywords found and fixed were added to the monocular and large screen configurations. For the handheld display, when the user pressed the button to indicate a defective item, three buttons were presented to record the type of defect: "Not fixed," which corresponds to fail, "I fixed it," which corresponds to fixed, and "Already fixed," which corresponds to found. We also associated new colors with each option. For a fixed item, the rectangle in the progress bar and the rectangle for the completed item are a low saturation green, and for a found item, they are a medium saturation green. The green color indicates that these items have in effect passed inspection but are differentiated from the high saturation green that indicates a normal "passed" item. Finally, for all three configurations when an inspector chose to stop an inspection, we gave an option to return to the inspection, since many users expressed frustration when they accidentally hit the stop button on the handheld device, both in the experimental evaluation and in the preliminary demonstration.

We used Dragon Naturally Speaking for voice input during our field test instead of a Wizard-of-Oz setup as in the experimental evaluation due to interest in the feasibility of speech recognition in the industrial setting as well as restrictions on Bluetooth communication. For checkpoints, we used a set of 6 actual 12-item inspection lists provided by BMW.

5.2 Field Site. Expert inspection associates evaluated our system at the BMW manufacturing plant in conditions similar to their everyday work environments. Associates tested our system in two different locations. In the nonproduction location, the vehicle that the participants inspected was stationary, and the environment was moderately quiet except for occasional construction noise. In this location, there were no restrictions on where the user could walk and no enforced time constraint, since this was not a production setting. We were able to test all three hardware configurations in this setting. The large screen was centered horizontally along the vehicle and placed 6 ft away from the driver's side of the vehicle.

The production location was at an inspection station along the production assembly line with vehicles that were currently being assembled and inspected on a moving platform. In this location, the inspector inspected the vehicle slightly behind the location where another associate was doing the actual production inspection. The inspector had to be more mindful of her location and events occurring around her, since the assembly line was moving; other people were performing tasks around them, and notifications were played over loudspeakers. This environment was very noisy. Due to safety and quality assurance concerns, we were only permitted to test the handheld and large screen configurations in this setting—supervisors were concerned that the monocular display would inhibit inspectors' vision, which could cause them to walk into the path of moving parts on the assembly line and were also worried that the hanging zipper tags and cables could get caught in other moving parts or scratch the car body. For the large screen configuration, the screen was placed a vehicle's length ahead of the location where inspection began, and as the inspection proceeded, the user moved closer to the display as the vehicle moved along the assembly line.

5.3 Procedure and Measures. The day before the field test, each inspector completed voice training in a quiet environment, and we tested speech recognition for each of the keywords, training individual keywords that were problematic until we felt the recognition rate was sufficient. The inspector was then given a handout with labeled screenshots of our software interface and a listing of the keywords, along with a written description of how to use the systems. The following day, we evaluated our systems at the two sites described above: nonproduction and production, with all three hardware configurations evaluated in the nonproduction environment and only the handheld (H) and large screen (L) alternatives evaluated in the production setting.

Each inspector completed at least three inspection lists with each device. On the first inspection for each device, we assisted the inspector with any questions and problems and followed less closely on the remaining inspections in order to reduce task inhibition. Because it was not possible to evaluate inspection accuracy since we could not know the state of the checkpoints of the vehicle to be inspected, we focused on gathering inspector preferences and recommendations for improvement. After the inspector used all three systems, we interviewed them concerning their opinions on our systems.

5.4 Results. Our evaluation included a total of five experienced BMW associates who conduct vehicle inspections on a daily basis. Two associates participated only in the production environment, two participated only in the nonproduction environment, and one associate participated in both environments. There were two female and three male associates. The participants are knowledgeable in car part locations and have experience using BMW's current inspection reporting systems.

5.4.1 Expert Users' Preferences. The three participants who used all three devices in the nonproduction environment were asked which of the three hardware configurations they preferred. Two preferred the large screen (L) and one preferred the handheld device (H). The participant who chose the handheld device (H) made that choice based on her previous experience with handheld devices. However, she later commented that she preferred speech input over other modalities. The three participants who used only two devices (handheld and large screen) in the production environment preferred the large screen (L) configuration.

Out of all five participants across both environments, three preferred speech recognition for input, while the remaining two would have preferred to be able to use both speech recognition and touch for input. Three participants preferred listening to the items read out loud by text-to-speech alone, while two participants preferred to use both visual and audio cues for items. No participants preferred any other alternatives.

5.4.2 Comparison to Current Systems. Overall, participants felt that the way the system delivered the list on a screen or through a headset (regardless of hardware configuration) was an improvement from the PC configuration currently used in inspection. Participants commented on how our configurations reduce the need to travel to the computer, saving time and memory load, as well as reducing mistakes:

- "[I] love the screen; I can't see the PC while checking the car."
- "[I] don't have to go back and forth to the computer, [so I] save steps [and am] less apt to miss something."
- "I go by memory now, [but the] visual and audio aids pinpointing inspection points and [will] improve quality of inspection."
- "[During manufacturing they] add check[points] different days, so [with this interface you] don't have to worry about remembering if you have to check [the items]."

Participants felt that our interface software alone was a major improvement, due to its touch-based input (instead of stylus), larger text, and fewer items displayed at a time, leading to better readability:

- "[The interface is] excellent now, compared to what I was doing."
- "Everything is there [on our improved system], [on the current handheld system] you have to hit it with a little pen [and] it doesn't go directly to pass, fail or fixed; you have to go through several things before you get to it."
- "[I] liked input and how easy it was to read [and the] bright screen. The current handheld system is hard to read and touch is very small; [you] have to use the stylus."
- "I liked that there were less words on the screen, in [the current handheld device] all of the items are displayed really tiny at once and I have to scroll."

Participants additionally commented positively on the multimodality of the other configurations:

- "Inspection needs to be hands free."
- "The audio with the large screen utilizes both [output modalities]. [I] don't like the handheld because I have to lean into the car and pull or push so I have to put it down at times."
- Several participants commented that the system would be useful for training new associates and for helping experienced associates returning from vacation or changing inspection stations to regain familiarity with the inspection system.
- "When [you] have a day off, you have to come in and look at the book. This is better because it is automatically within the system."

5.4.3 Observations on Participant Performance. For the large screen and monocular configurations, participants generally conducted one-handed and two-handed inspections properly. In contrast, when participants used the handheld device, while one or two inspectors put down the device, the majority did not and only used one hand for inspections that should require two hands, possibly endangering the quality of the inspection. For example, one

participant used a pen light in one hand and felt the surfaces under inspection with his other hand while using the large screen and monocular configurations, but while using the handheld device, held the phone in one hand and the pen light in the other hand and did not touch the surface as necessary for proper inspection. This shows that it is important to have a hands-free aid in inspection because carrying something can be distracting, inhibiting how the inspection is conducted, increasing the time it takes if inspectors need to put the device down and possibly endangering the surface of the vehicle.

5.4.4 Discussion on Expert User Evaluation. Regardless of hardware configuration, expert users viewed our software interface as an improvement on their current software due to its improved readability [6] and its aid in remembering items on the checklist for improved accuracy and increased learnability. They preferred our handheld configuration over the existing handheld device due to its readability and large buttons, but overall preferred the hands-free interaction afforded by the other hardware configurations. Additionally, we observed that hands-free interaction supported proper inspection actions while the handheld configuration inhibited inspection performance.

Although we were unable to test the monocular configuration within the production environment due to safety and quality concerns, our experimental study as well as the evaluation from the experts in the nonproduction environment indicate that the large screen and handheld configurations are likely the best in terms of user preference. It is possible that using alternative hardware with less cabling could remedy safety and quality concerns and allow the system to be tested in production.

Due to the nature of a functional manufacturing environment, we could not measure data for accuracy and task completion time and so could not compare any task performance data between the existing systems and our systems. Further research should be conducted to determine whether our systems meet the goal of increasing inspection accuracy within the production setting.

6 Conclusion

Our research sought to determine the performance differences among system-directed multimodal systems for vehicle inspection. Evaluations conducted in the experimental setting with novice users helped to reveal performance benefits. Among novice users, although no hardware configuration significantly improved accuracy or reduced completion time, it was reported that the multimodal configurations aided in conducting two-handed inspection tasks and improved memory load, possibly reducing mistakes. When designing a system, it is not only important to understand and observe how people are conducting their tasks in the actual environment with expert users but also evaluations conducted with expert users in the task environment can help reveal user preferences and benefits relating to the specific domain. Expert users reported that our software interface facilitated in the vehicle inspection process better than the current software due to improved readability, system-directed task delivery, and fewer items displayed at a time. Additionally, experts commented that the system-directed format can help new associates get started more quickly or assist existing associates when they return from a long absence from the job. Observation of systems in use can also reveal information that users might not be aware of, such as in this case where users were modifying how they inspected an item that needed two hands when one hand was occupied.

In conclusion, our results show that for visual and tactile tasks, benefits of system-directed interfaces are best realized when used with multimodal systems that reduce visual and tactile interaction per item and instead deliver system-directed information on the audio channel. Interface designers that combine system-directed interfaces with multimodal systems can expect faster and more efficient user performance when the delivery channel is different from channels necessary for task completion.

Acknowledgment

This research was funded, in part, by a grant from BMW Manufacturing Corporation, LLC and student resources were provided in part by BMW Information Technology Research Center (ITRC). Without the support of Josef Schwab, this project would have been unthinkable. This project was also supported by the National Science Foundation Graduate Research Fellowship (fellow IDs 2011095211 and 2009080400).

References

[1] Boronowsky, M., Nicolai, T., Schlieder, C., and Schmidt, A., 2001, "Winspect: A Case Study for Wearable Computing-Supported Inspection Tasks," Proceedings Fifth International Symposium on Wearable Computers, IEEE, pp. 163–164.
[2] Burgy, C., Garrett, J. H., Jr., and Klausner, M., 2000, "Speech-Controlled Wearable Computer: A Mobile System Supporting Inspections in Garages." Availabe at: http://www.ce.cmu.edu/~wearables/docs/scwc_projectreport_03_2000.pdf (Accessed 9 December 2012).
[3] Ockerman, J., and Pritchett, A., 1998, "Preliminary Investigation of Wearable Computers for Task Guidance in Aircraft Inspection," Digest of Papers, Second International Symposium on Wearable Computers, IEEE, pp. 33–40.
[4] Sunkpho, J., Garrett, J., Jr., Smailagic, A., and Siewiorek, D., 1998, "MIA: A Wearable Computer for Bridge Inspectors," Digest of Papers, Second International Symposium on Wearable Computers, IEEE, pp. 160–161.
[5] BMW Manufacturing Co., Spartanburg South Carolina, 2012.
[6] Cairco, L., Ulinski, A., McClendon, J., Bloodworth, T., Matheison, J., Hodges, L., and Summers, J., 2010, "Interface Design and Display Modalities to Improve the Vehicle Inspection Process," Proceedings on WINVR, ASME.
[7] Lukowicz, P., Timm-Giel, A., Lawo, M., and Herzog, O., 2007, "Wearit@work: Toward Real-World Industrial Wearable Computing," IEEE Pervasive Comput., **6**(4), pp. 8–13.
[8] Maurtua, I., Kirisci, P., Stiefmeier, T., Sbodio, M., and Witt, H., 2007. "A Wearable Computing Prototype for Supporting Training Activities in Automotive Production," 4th International Forum on Applied Wearable Computing (IFAWC), VDE, pp. 1–12.
[9] ETH-Ife-Wearable Computing–Wearit@Work Project Homepage, 2012.
[10] Aleksy, M., Rissanen, M., Maczey, S., and Dix, M., 2011, "Wearable Computing in Industrial Service Applications," Proc. Comput. Sci., **5**, pp. 394–400.
[11] Rogers, Y., Sharp, H., and Preece, J., 2011, *Interaction Design: Beyond Human-Computer Interaction*, Wiley, New York.
[12] Miller, G., 1956, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," Psychol. Rev., **63**(2), pp. 81–97.