Usability Engineering for Augmented Reality: Employing User-based Studies to Inform Design

Joseph L. Gabbard and J. Edward Swan II, *Member, IEEE*

Abstract— A major challenge, and thus opportunity, in the field of human-computer interaction and specifically usability engineering is designing effective user interfaces for emerging technologies that have no established design guidelines or interaction metaphors or introduce completely new ways for users to perceive and interact with technology and the world around them. Clearly, augmented reality is one such emerging technology. We propose a usability engineering approach that employs user-based studies to inform design, by iteratively inserting a series of user-based studies into a traditional usability engineering lifecycle to better inform initial user interface designs. We present an exemplar user-based study conducted to gain insight into how users perceive text in outdoor augmented reality settings and to derive implications for design in outdoor augmented reality. We also describe "lessons learned" from our experiences conducting user-based studies as part of the design process.

Index Terms— H.5.2: User Interfaces — Ergonomics, Evaluation / Methodology, Screen Design, Style Guides, H.5.1: Multimedia Information Systems — Artificial, Augmented, and Virtual Realities,

1 Introduction and Context

A major challenge, and thus opportunity, in the field of human-computer interaction (HCI) and specifically usability engineering (UE) is designing effective user interfaces for emerging technologies that have no established design guidelines or interaction metaphors, or introduce completely new ways for users to perceive and interact with technology and the world around them. Clearly, augmented reality (AR) is one such emerging technology. In these design contexts, it is often the case that user-based studies, or traditional human factors studies, can provide valuable insight. However, a literature survey we conducted in 2004 [1] found that user-based studies have been underutilized in AR, and we posit that this underutilization extends well beyond this specific technology. Our survey found that, in a total of 1104 articles on augmented reality, only 38 (~3%) addressed some aspect of HCI, and only 21 (~2%) described a formal user-based study. As a community, how can we expect to design and deploy effective application-level user interfaces and interaction techniques when we have too little understanding of human performance in these environments? We assert that the most effective user interfaces for emerging technologies will be grounded on user-based studies that aim to understand fundamental perceptual and cognitive factors, especially for those technologies that fundamentally alter the way humans perceive the world (e.g., VR, AR, etc.).

In this paper, we propose a usability engineering approach that employs user-based studies to inform design, by iteratively inserting a series of user-based studies into a traditional usability engineering lifecycle to better inform initial user interface designs. Under this approach, user performance can be explored against combinations of design parameters (i.e., experimental factors and levels), to discover what combinations of parameters support the best user performance under various conditions. What makes this approach different than traditional HCI approaches is that basic user interface and/or interact issues are explored vis-à-vis user-based studies *as part of* the usability engineering of a specific application, as opposed to application developers drawing from a body of established guidelines produced in the past by others performing low-level, or generic, user-based studies.

We have applied this approach as part of the usability engineering and software development of the BARS. Following a domain analysis [3], we began to identify over 20 scientific challenges, over half of which were user interface design challenges that required insight from conducting user-based studies. Since that time, most of our user-based studies have focused on three of these areas; (1) the representation of occlusion [4], (2) understanding depth perception in optical see-through AR [5],[6] and (3) text legibility in outdoor optical see-through AR [7],[8],[9].

We first present our usability engineering approach, and justify the importance of employing user-based studies to inform design. We then present an exemplar user-based study, which we conducted to gain insight into how users perceive text in outdoor AR settings and to

- *J.L. Gabbard is with the Center for Human Computer Interaction at Virginia Tech, Blacksburg, VA, 24061. E-mail: jgabbard@vt.edu.*
- *J. Ediward Swan II is with the Department of Computer Science & Engineering, Mississippi State University, Starkville, MS 39762. E-mail: swan@acm.org.*

Manuscript received (August 3, 2007). Please note that all acknowledgments should be placed at the end of the paper, before the bibliography.

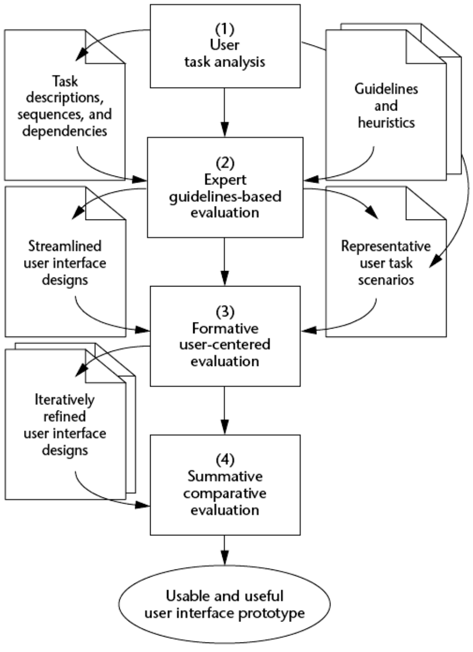

Fig. 1. The user-centered design and evaluation methodology for virtual environment user interaction described by Gabbard et al. [14].

derive implications for design in outdoor AR[1]. Lastly, we describe some "lessons learned" from our experiences conducting user-based studies as part of the design process.

2 Usability Engineering Approaches to Designing User Interfaces

To date, numerous approaches to software and user interface design have been developed and applied. The waterfall model, developed by Royce [10], was the first widely-known approach to software engineering. This model takes a top-down approach based on functional decomposition. Royce admitted that while this process was designed to support large software development efforts, it was inherently flawed since it did not support iteration; a property that he eventually added it to the model.

The spiral model [11] was the first widely recognized approach that utilized and promoted iteration. It is useful for designing user interfaces (as well as software), because it allows the details of user interfaces to emerge over time, with iterative feedback from evaluation sessions feeding design and redesign. As with usability engineering approaches, the spiral model first creates a set of user-centered requirements through a suite of traditional domain analysis activities (e.g., structured interviews, participatory design, etc.). Following requirements analysis, the second step simply states that a "preliminary design is created for the new system".

Hix and Hartson [12] describe a star life cycle that is explicitly designed to support the creation of user interfaces. The points of the star represent typical design/development activities such as "user analyses", "requirements/usability specifications", "rapid prototyping", etc, with each activity connected through a single center "usability evaluation" activity. The points of the start are not ordered, so one can start at any point in the process, but can only proceed to another point via usability evaluation. The design activities focus on moving from a conceptual design to a detailed design.

Mayhew [13] describes a usability engineering lifecycle that is iterative and centered on integrating users throughout the entire development process. With respect to design, the usability engineering lifecycle relies on screen design standards, which are iteratively evaluated and updated. Both the screen design standards as well as the detailed user interface designs rely on style guides that can take the form of a "platform" style guide (e.g., Mac, Windows, etc.), "corporate" style guide (applying a corporate "look and feel"), "product family" style guide (e.g., MS Office Suite), etc.

Gabbard, Hix, and Swan [14] present a cost-effective, structured, iterative methodology for user-centered design and evaluation of virtual environment (VE) user interfaces and interaction. Fig. 1 depicts the general methodology, which is based on sequentially performing:

1. user task analysis,
2. expert guidelines-based evaluation,
3. formative user-centered evaluation, and
4. summative comparative evaluations.

While similar methodologies have been applied to traditional (GUI-based) computer systems, this methodology is novel because we specifically designed it for—and applied it to—VEs, and it leverages a set of heuristic guidelines specific to VEs. These sets of heuristic guidelines were derived from Gabbard's taxonomy of usability characteristics for VEs [15].

A shortcoming of this approach is that it does not give much guidance for design activities. The approach does not describe how to engage in design activities, but instead asserts that initial designs can be created using input from task descriptions, sequences, and dependencies as well as guidelines and heuristics from the field. Since this methodology assumes the presence of guidelines and heuristics to aid in designs to be evaluated during the "expert guidelines-based evaluation" phase, it is not applicable to emerging technologies such as augmented reality, where user interface design guidelines and heuristics have not yet been established.

When examining many of the approaches described above – and specifically the design and evaluation activities – in most cases, design activities rely on leveraging existing metaphors, style guides or standards in the field (e.g., drop down menus, a browser's "back" button, etc.). However, in cases where an application falls within an emerging technological field, designers often have no existing metaphors or style guides, much less standards on which to base their design. Moreover, in cases where the technology provides novel approaches to user interaction or fundamentally alters the way users perceive the interaction space (i.e., where technology and the real-world come together), designers often have little under-

[1] This study has been previously described by Gabbard et al. [9].

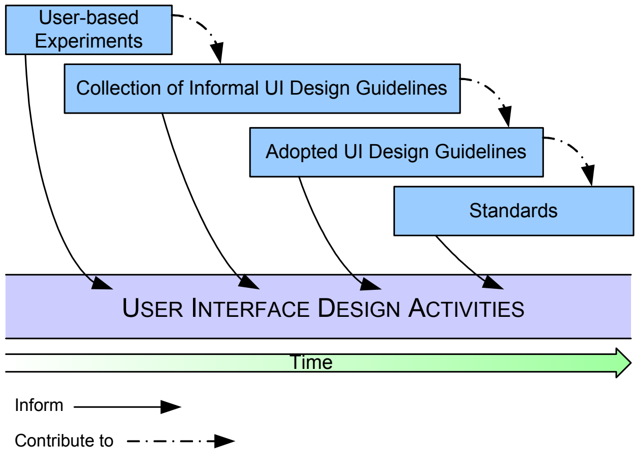
User-based Experiments
Collection of Informal UI Design Guidelines
Adopted UI Design Guidelines
Standards
User Interface Design Activities
Time
Inform
Contribute to

Fig. 2. User-based experiments are a critical vehicle for discovery and usability early in an emerging field's development. Over time, contributions from the field emerge, leading eventually to adopted user interface design guidelines and standards.

standing of the perceptual or cognitive ramifications of "best guess" designs.

As a result, a process is needed to help designers of novel user interfaces iteratively create and evaluate designs, to gain a better understanding of effective design parameters, and to determine under what conditions these parameters are best applied. Without this process, applications developed using traditional usability engineering approaches can only improve incrementally from initial designs — which again, are often based on developers' best guesses, given the absences of guidelines, metaphors, and standards.

2.1 User Interface Design Activities for Augmented Reality and other Emerging Technologies

As shown in Fig. 2, it can be argued that user-based experiments are critical for driving design activities, usability, and discovery early in an emerging technology's development (such as AR). As a technological field evolves, lessons learned from conducting user-based studies are not only critical for the usability of a particular application, but provide value to the field as a whole in terms of insight into a part of the user interface design space (e.g., of occlusion or text legibility). As time progresses, contributions to the field (from many researchers) begin to form a collection of informal design guidelines and metaphors from which researchers and application designers alike can draw. Eventually, the informal design guidelines are shaken down into a collection of tried-and-true guidelines and metaphors that are adopted by the community. Finally, the guidelines and metaphors become "defacto" standards or at best deemed "standards" by appropriate panels and committees.

The context of the work reported here, however, falls within the application of user-based studies to inform user interface design; the left, upper-most box of Fig. 2. Based on our experiences performing usability engineering, and specifically design and evaluation activities for the BARS, we propose an updated approach to user interface design activities for augmented reality systems. This approach emphasizes iterative design activities in between the *user task analysis phase*, where requirements are gathered and user tasks understood, and the *formative user-centered evaluation phase*, where an application-level user interface prototype has been developed and is under examine. With this approach, we couple the expert evaluation and user-based studies to assist in the user interface design activity (Fig. 3). These user-based studies differ from traditional approaches to application design, in that their scope addresses basic user interface or interaction design in lieu of established design guidelines. Expert evaluations can be iteratively combined with well-designed user-based studies to refine designers' understanding of the design space, understanding of effective design parameters (e.g., to identify subsequent user-based studies), and most importantly to refine user interface designs. A strength of this approach is that interface design activities are driven by a number of activities; inputs from the user task analysis phase, user interface design parameters correlated with good user interface performance (derived from user-based studies), and expert evaluation results.

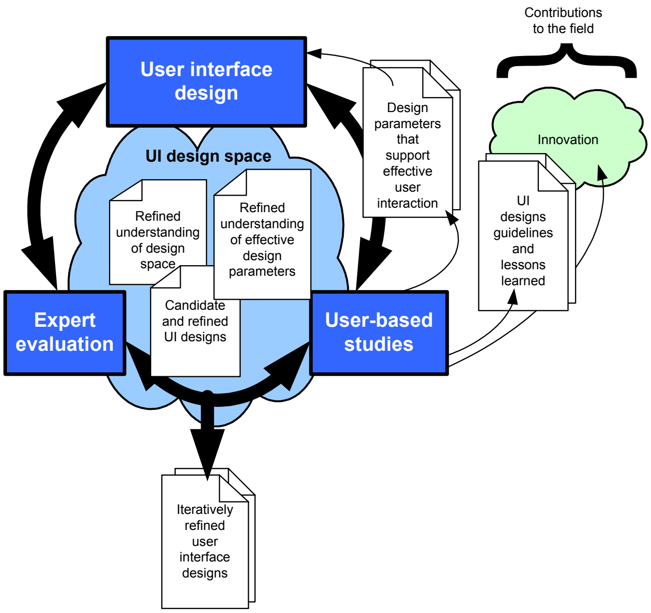
Contributions to the field
User interface design
UI design space
Design parameters that support effective user interaction
Innovation
Refined understanding of design space
Refined understanding of effective design parameters
UI designs guidelines and lessons learned
Candidate and refined UI designs
Expert evaluation
User-based studies
Iteratively refined user interface designs

Fig. 3. We applied the depicted user-centered design activities as part of our overall usability engineering approach. With this methodology, expert evaluations along with user-based studies are iteratively applied to refine the user interface design space. It is the scope of the user-based studies that make this approach unique in that that the user-based studies address basic user interface and interaction issues (as opposed to application-level user interface issues) in lieu of established design guidelines.

Of the three main activities shown in Fig. 3, there are two logical starting points: *user interface design* and *user-based studies*. An advantage of starting with user interface design activities is that designers can start exploring the design space prior to investing time in system development, and moreover, can explore a number of candidate designs quickly and easily. In the past, we have successfully used PowerPoint mockups to examine dozens of AR design alternatives. If mocked up correctly, the static designs can be presented through an optical see through display, which allows designers to get an idea of how the designs may be perceived when viewed through an AR display in a representative context (e.g., indoors versus outdoors).

Once a set of designs has been created, expert evaluations can be applied to assess the static user interface designs, culling user interface designs that are likely to be less effective than others. The expert evaluations are also useful in terms of further understanding the design space by identifying potential user-based experimental factors and levels. Once identified, user-based studies can be conducted to further examine those factors and levels to determine, for example, if the findings of the expert evaluation match that of user-based studies.

In cases where the design space is somewhat understood and designers have specific questions about how different design parameters might support user task performance, designers may be able to conduct a user-based study as a starting point. Under this approach, designers start with experimental design parameters as opposed to specific user interface designs. As shown in Fig. 3, user-based studies not only identify user interface design parameters to assist in UI design, but also have the potential to produce UI design guidelines and lessons learned, as well as generate innovation, which provides both tangible contributions to the field while also improving the usability of a specific application.

Ultimately, a set of iteratively refined user interface designs are produced that are the basis for the overall application user interface. This design can then be evaluated using formative user-centered evaluation, as described by Hix, Gabbard, and Swan [14].

3 Case Study: A User-based Study Examining Text Legibility in Outdoor AR

In this section we describe, as a case study, a user-based experiment that seeks to better our understanding of how users perceive text in outdoor AR settings. Note that this study was first reported at VR 2007 [9]. However, this case study extends the VR07 work by including an analysis of pair-wise contrast comparisons as described below

This study is one of many user-based studies that we have conducted as part of our BARS design activities. As depicted in Fig. 3, these this user-based study was part of the proposed iterative design cycle, which included expert evaluation of PowerPoint mockups as well as a prior user-based study on text legibility (first reported in [7]). From these experiences, as well as insights from the graphics art field, we identified a number of important design parameters used to drive the design of this study.

3.1 Outdoor Augmented Information

Presenting legible augmenting information in the outdoors is problematic, due mostly to uncontrollable environmental conditions such as large-scale fluctuations in natural lighting and the various types of backgrounds on which the augmenting information is overlaid. There are often cases where the color and/or brightness of a real-world background visually and perceptually conflicts with the color and/or contrast of graphical user interface (GUI) elements such as text, resulting in poor or nearly-impossible legibility. This issue is particularly true when using optical see-through display hardware.

Several recent studies in AR have begun to experimentally confirm that which was anecdotally known amongst outdoor AR practitioners, but not yet documented — namely, that text legibility is significantly affected by environmental conditions, such as color and texture of the background environment as well as natural illuminance at both the user's and background's position [8],[7],[16],[17],[18].

One strategy to mitigate this problem is for visual AR representations to actively adapt, in real-time, to varying conditions of the outdoor environment. Following this premise, we created a working testbed to investigate interactions among real-world backgrounds, outdoor lighting, and visual perception of augmenting text. We have termed this testbed a "visually active AR testbed". This testbed senses the condition of the environment using a real-time video camera and lightmeter. Based on these inputs, we apply active algorithms to GUI text strings, which alter their visual presentation and create greater contrast between the text and the real-world backgrounds, ultimately supporting better legibility and thus user performance. This concept easily generalizes beyond text strings to general GUI elements.

We conducted a study that examined the effects on user performance of outdoor background textures, text colors, text drawing styles, and text drawing style algorithms for a text identification task. We captured user error, user response time, and details about text drawing and real-world background colors for each trial.

3.2 A Visually Active AR Testbed

Our recent instantiation of a visually active AR user interface serves as a testbed for empirically studying different text drawing styles and active text drawing algorithms under a wide range of outdoor background and illuminance conditions. Fig. 4 shows our testbed, which employs a real-time video camera to capture a user's visual field of view and to specifically sample the portion of the real-world background on which a specific user interface element (e.g., text) is overlaid. It also employs a real-time lightmeter (connected via RS232) to provide real-time natural illuminance information to the active system. The user study reported in this paper only actively uses the camera information; the testbed recorded lightmeter information but did not use it to drive the active algorithms. We anticipate developing algorithms that are actively driven by the lightmeter in the future.

As shown in Fig. 4, the AR display, camera and lightmeter sensor are mounted on a rig, which in turn is mounted on a tripod (not shown in the figure). Participants sat in an adjustable-height chair so that head positions are consistent across all participants. Our testbed did not use a motion tracking system. For this study, we

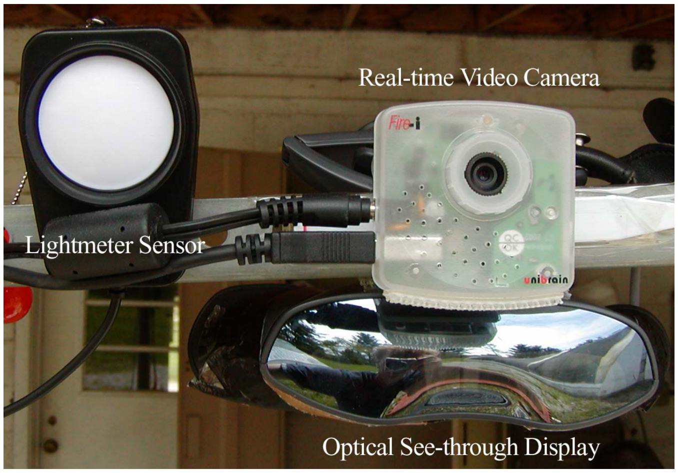

Fig. 4. AR display, video camera and lightmeter components of our visually active AR testbed.

fixed the participants' field-of-view on different backgrounds by repositioning the rig between background conditions. We used previously captured camera images of backgrounds to assist in the positioning procedure and to ensure that each participant's FOV is the same for each background.

Our testbed uses the text's screen location and font characteristics to compute a screen-aligned bounding box for each text string. It then computes the average color of this bounding box, and uses this color to drive the active text drawing algorithms – which in turn determine a text drawing style color. For example, if using a billboard drawing style (see Fig. 7), the active text drawing algorithm uses the sampled background color as an input to determine what color to draw the billboard. The specific text drawing styles and text drawing style algorithms are discussed in more detail below.

Our testbed was implemented as a component of the BARS, and uses an optical see-through display, a real-time video camera, a lightmeter, and a mobile laptop computer equipped with a 3D graphics card. The optical see-through display was a Sony Glasstron LDI–100B biocular optical see-through display, with SVGA resolution and a 28° horizontal field of view in each eye. We used a UniBrain Fire-i firewire camera (with settings of YUV 4:2:2 format, 640 X 480 resolution, 30Hz, and automatic gain control and exposure timing). The lightmeter is an Extech 407026 Heavy Duty Light Meter with RS232 interface to measure illuminance at the user's position. Our laptop system (and image generator) was a Pentium M 1.7 GHz computer with 2 gigabytes of RAM and an NVidia GeForce4 4200 Go graphics card generating monoscopic images, running under Windows 2000. We used this same computer to collect user data. Fig. 4 shows the HMD, camera, and lightmeter components.

3.3 Task and Experimental Setup

We designed a task that abstracted the kind of short reading tasks, such as reading labels, which are prevalent in many proposed AR applications. For this study, we purposefully designed the experimental task to be a low-level visual identification task. That is, we were not concerned with participants' semantic interpretation of the data, but simply whether or not they could quickly and accurately read information. Moreover, the experimental task was designed to force participants to carefully discern a series of random letters, so that task performance was based strictly on legibility. The task was a relatively low-level cognitive task consisting of visual perception of characters, scanning, recognition, memory, decision-making, and motor response.

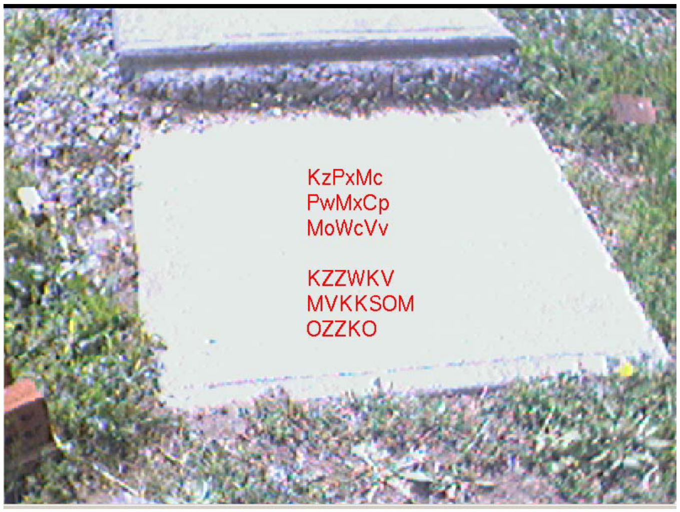

Fig. 5. Our experimental task required participants to identify the pair of identical letters in the upper block (e.g., "Vv"), and respond by pressing the numeric key that corresponds to the number of times that letter appears in the lower block (e.g., "2"). Note that this image is a screen capture (via camera) of the participants' field of view and overlaid text, and is not an exact representation of what participants viewed through the AR display.

As shown in Fig. 5, participants viewed random letters arranged in two different blocks. The upper block consisted of three strings of alternating upper- and lower-case letters, while the lower block consisted of three strings of upper-case letters. We instructed the participant to first locate a target letter from the upper block; this was a pair of identical letters, one of which was upper case and the other lower case (e.g., "Vv" in Fig. 5). Placement of the target letter pair in the upper block was randomized, which forced participants to carefully scan through the block. We considered several other visual cues such as underlining, larger font size, and bold text for designating the target letter; however, we realized that this would result in a "pop-out" phenomenon wherein the participant would locate the target without scanning the distracting letters.

We used the restricted alphabet "C, K, M, O, P, S, U, V, W, X, Z" to minimize variations in task time due to the varying difficulty associated with identifying two identical letters whose upper and lower case appearance may or may not be similar. A post-hoc analysis showed an effect size of d = .07 error for letter, which is small when compared to the other effect sizes reported in this paper.

After locating the target letter, the participant was then instructed to look at the lower block and count the number of times the target letter appeared in the lower block. Placement of the target letters in the lower block was ran-

Table 1. Summary of variables studied in experiment.

Independent Variables

participant	24	*counterbalanced*
outdoor background texture (Fig. 6)	4	brick, building, sidewalk, sky
text color	4	white, red, green, cyan
text drawing style (Fig. 7)	4	none, billboard, drop shadow, outline
text drawing style algorithm	2	maximum HSV complement, maximum brightness contrast
repetition	3	1, 2, 3

Dependent Variables

response time	in *milliseconds*
error	0 (*correct*), 1, 2, 3 (*incorrect*)

domized. Participants were instructed that the target letter would appear 1, 2, or 3 times. The participant responded by pressing the "1", "2", or "3" key to indicate the number of times the target letter appeared in the lower block. In addition, participants were instructed to press the "0" key if they found the text completely illegible.

To minimize carryover effects of fatigue, a rest break was also provided every 28 trials; participants were instructed to close their eyes and relax. The length of the rest break was determined by each participant. After each rest break, the next task was presented to the participant in a similar manner. The entire study consisted of 336 trials for each participant.

We wanted to conduct the study under outdoor illuminance conditions, because while indoor illuminance varies by ~3 orders of magnitude, outdoor illuminance varies by ~8 orders of magnitude [19]. However, we could not conduct the study in direct sunlight, because graphics on the Glasstron AR display become almost completely invisible. We also needed to protect the display and other equipment from outdoor weather conditions. We addressed these issues by conducting our study in a covered breezeway overlooking an open area. Since this location required participants to face south (i.e., towards the sun as it moves across the sky), we positioned the participant at the edge of the breezeway, so that their heads (and thus the display) were shaded from the sun, but their vertical field of view was not limited by the breezeway's roof structure. We ran the experiment between April 6th and May 10th, 2006, in Blacksburg, Virginia, during which time the sun's elevation varied between 23° and 68° above the horizon.

We conducted studies at 10am, 1pm, and 3pm, and only on days that met our pre-determined natural illuminance lighting requirements (between 2000 and 20,000 lux). Using the lightmeter displayed in Fig. 4, we measured the amount of ambient illuminance at the participant's position every trial. Our goals were to quantify the effect of varying ambient illumination on task performance, and to ensure that ambient illuminance fell into our established range. However, our current finding is that between-subjects illumination variation, which represents differences in the weather and time of day, was much larger than the variation between different levels of experimental variables. Therefore, we do not report any effects of illuminance as collected at the user's position in this paper.

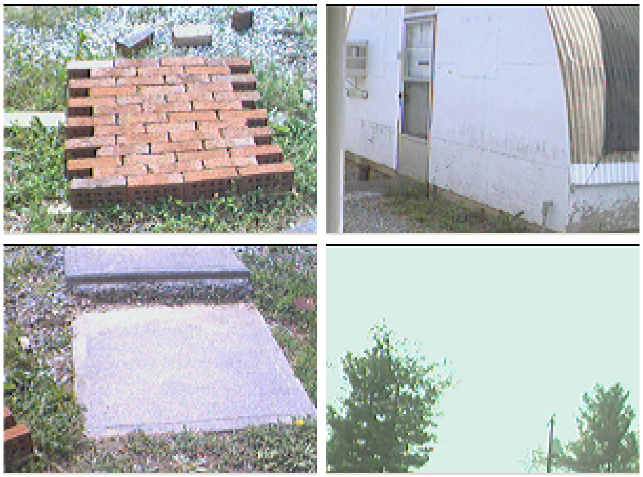

Fig. 6. We used four real-world outdoor background textures for the study. Shown above are (clockwise starting in upper left): brick, building, sky, and sidewalk. Stimulus text strings (both upper and lower blocks) were completely contained within the background of interest (as shown in Fig. 5). The images represent participants field of view when looking through the display.

3.4 Independent Variables

A summary of our independent variables is presented in Table 1. Details of each independent variable follow.

Outdoor Background Texture: We chose four outdoor background textures to be representative of commonly-found objects in urban settings: brick, building, sidewalk, and sky (Fig. 6). Note that three of these backgrounds (all but building) were used in our previous study [7], [8] but at that time were presented to the participant as large posters showing a high-resolution photograph of each background texture. In this new study, we used actual real-world backgrounds. Stimulus strings were positioned so that they were completely contained within each background (Fig. 5).

We kept the brick and sidewalk backgrounds covered when not in use, so that their condition remained constant throughout the study. The sky background varied depending upon cloud cover, haze, etc., and in some (rare) cases would vary widely as cumulus clouds wandered by. We considered including a grass background, but were concerned that the color and condition of the grass would vary during the months of April and May, moving from a dormant green-brown color to a bright green color.

Text Color: We used four text colors commonly used in computer-based systems: white, red, green, and cyan. We chose white because it is often used in AR to create labels and because it is the brightest color presentable in an optical see-through display. Our choice of red and green was based on the physiological fact that cones in the human eye are most sensitive to certain shades of red and green [20], [21]. These two text colors were also used

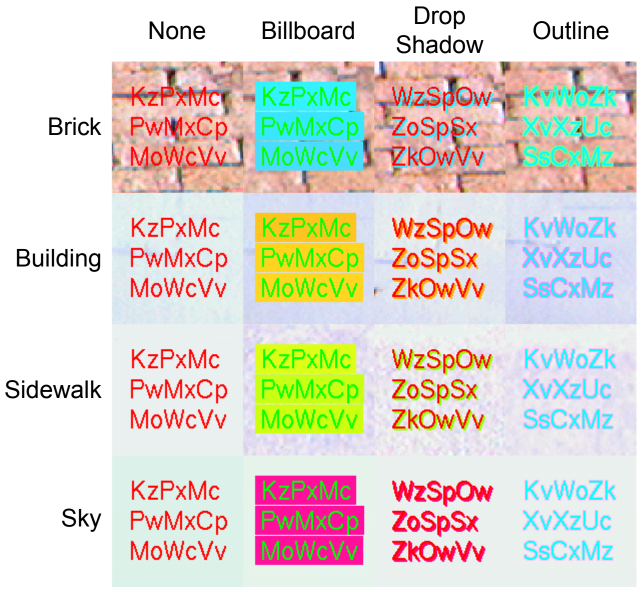

Fig. 7. We used four text drawing styles: none, billboard, drop shadow and outline (shown on the four outdoor background textures). Note that the thumbnails shown above were sub-sampled from the participant's complete field of view.

in our first study. We chose cyan to represent the color blue. We chose not to use a "true" blue (0, 0, 255 in RGB color space), because it is a dark color and is not easily visible in optical see-through displays.

Text Drawing Style: We chose four text drawing styles (Fig. 7): none, billboard, drop shadow, and outline. These are based on previous research in typography, color theory, and human-computer interaction text design. We used a sans serif font (Helvetica), and presented the text at a size that appeared approximately two inches tall at a distance of two meters. Text size did not vary during the experiment. None means that text is drawn "as is", without any surrounding drawing style. We included the billboard style because it is commonly used in AR applications and in other fields where text annotations are overlaid onto photographs or video images; arguably it is one of the standard drawing styles used for AR labels. We used *billboard* in our previous study as well [7]. We included *drop shadow* because it is commonly used in print and television media to offset text from backgrounds. We included *outline* as a variant on drop shadow that is visually more salient yet imposes only a slightly larger visual footprint. The outline style is similar to the "anti-interference" font described by Harrison and Vicente [22]. Another motivation for choosing these drawing styles was to compare text drawing styles with small visual footprints (drop shadow, outline) to one with a large visual footprint (billboard).

Text Drawing Style Algorithm: We used two active algorithms to determine the color of the text drawing style: maximum HSV complement, and maximum brightness contrast. These were the best active algorithms from our previous study [7]. As discussed above, the input to these algorithms is the average color of the screen-aligned bounding box of the augmenting text. We designed the *maximum HSV complement algorithm* with the following goals: retain the notion of employing color complements, account for the fact that optical see-through AR displays cannot present the color black, and use the HSV color model [23] so we could easily and independently modify saturation. We designed the *maximum brightness contrast algorithm* to maximize the perceived brightness contrast between text drawing styles and outdoor background textures. This algorithm is based on MacIntyre's maximum luminance contrast technique [24], [25]. Both algorithms are described in detail by Gabbard et al. [7].

Repetition: We presented each combination of levels of independent variables three times.

3.5 Dependent Variables

Also as summarized in Table 1, we collected values for two dependent variables: response time and error. For each trial, our custom software recorded the participant's four-alternative forced choice (0, 1, 2, or 3) and the participant's response time. For each trial, we also recorded the ambient illuminance at that moment in time, the average background color sampled by the camera, as well as the color computed by the text drawing style algorithm. This additional information allowed us to calculate (post-hoc) pair-wise contrast values between text color, text drawing style color, and background color. In this paper we report on analyses of error and response time data, as well as the pair-wise contrast ratio.

3.6 Experimental Design and Participants

We used a factorial nesting of independent variables for our experimental design, which varied in the order they are listed in Table 1, from slowest (participant) to fastest (repetition). We collected a total of 24 (participant) × 4 (background) × 4 (color) × [1 (drawing style = none) + [3 (remaining drawing styles) × 2 (algorithm)]] × 3 (repetition) = 8064 response times and errors. We counterbalanced presentation of independent variables using a combination of Latin Squares and random permutations. Each participant saw all levels of each independent variable, so all variables were within-participant.

Twenty-four participants participated, twelve males and twelve females, ranging in age from 18 to 34. All participants volunteered and received no monetary compensation; some received a small amount of course credit for participating in the study. We screened all participants, via self-reporting, for color blindness and visual acuity. Participants did not appear to have any difficulty learning the task or completing the experiment.

3.7 Hypotheses

Prior to conducting the study, we made the following hypotheses:

(1) The brick background will result in slower and less accurate task performance because it is the most visually complex.

(2) The building background will result in faster and more accurate task performance because the building wall faced north and was therefore

shaded at all times.

(3) Because the white text is brightest, it will result in the fastest and most accurate task performance.

(4) The billboard text drawing style will result in the fastest and most accurate task performance since it has the largest visual footprint, and thus best separates the text from the outdoor background texture.

(5) Since the text drawing styles are designed to create visual contrast between the text and the background, the presence of active text drawing styles will result in faster and more accurate task performance than the *none* condition.

3.8 Results

For error analysis we created an error metric e that ranged from 0 to 3:

$$e = \begin{cases} |c-p| & \text{if } p \in \{1,2,3\} \\ 3 & \text{if } p = 0 \end{cases}, \quad (1)$$

where e = 0 to 2 was computed by taking the absolute value of c, the correct number of target letters, minus p, the participant's response. e = 0 indicates a correct response, and e = 1 or 2 indicates that the participant miscounted the number of target letters in the stimulus string. e = 3 is used for trials where users pressed the "0" key (indicating they found the text illegible). We first analyzed a signed-error term, but did not find any interesting or significant finding regarding over- versus under-counting, therefore, we used an absolute error term that considers over-counts and under-counts of the same magnitude as equivalent error values. This error metric was used because it is a more robust measure of error (as compared to simply capturing whether a response was correct or incorrect) as it provides a measure of *how* incorrect a response is, and more to the point, is a better indicator of how difficult the text is to read. Our rationale for using the value 3 for an unreadable stimulus string is that not being able to read the text at all warranted the largest error score, since it gave the participant no opportunity to perform the task. Our error analysis revealed a 14.9% error rate across all participants and all 8064 trials. This error rate is composed of 5.2% for e = 1, 0.5% for e = 2, and 9.2% for e = 3.

For response time analysis, we removed all repetitions of all trials when participants indicated that the text was illegible (e = 3), since these times were not representative of tasks performed under readable conditions. This resulted in 7324 response time trials (~91% of 8054 trials). Overall, we observed a mean response time of 5780.6 milliseconds (msec), with a standard deviation of 3147.0 msec.

We used repeated-measures Analysis of Variance (ANOVA) to analyze the error and response time data. We strove for an experiment-wide alpha level of 0.05 of less to denote a main effect. For this ANOVA, the participant variable was considered a random variable while all other independent variables were fixed. Because our design was unbalanced (the text drawing style *none* had no drawing style algorithm), and because we removed trials for the response time analysis, we could not run a full factorial ANOVA. Instead, we separately tested all main effects and two-way interactions of the independent variables. When deciding which results to report, in addition to considering the p value, the standard measure of effect significance, we considered d, a simple measure of effect size. d = *max* – *min*, where *max* is the largest mean and *min* the smallest mean of each result. d is given in units of either error or msec.

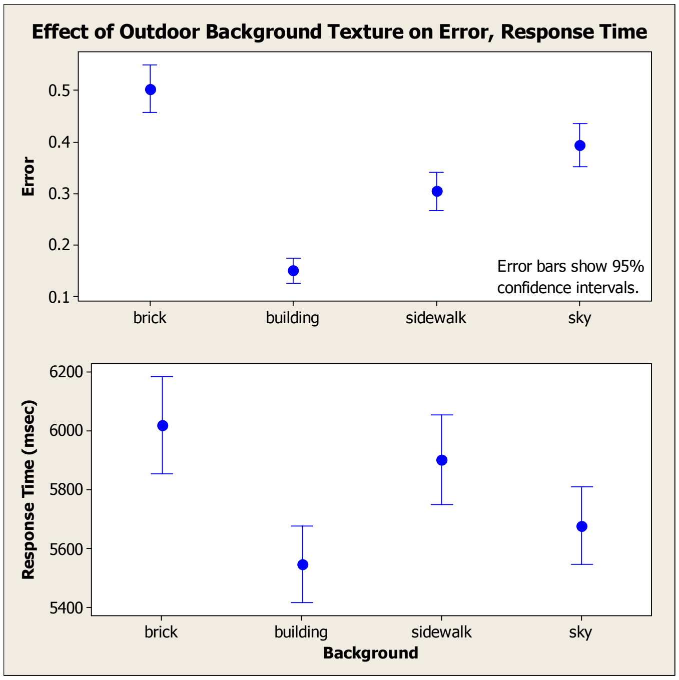

Fig. 8. Effect of background on error (N = 8064) and response time (N = 7324). In this and future graphs, N is the number of trials over which the results are calculated.

We also analyzed the pair-wise contrast ratios between text color and background color, text color and drawing style color, and drawing style color and background color. We performed ANOVA and correlation analysis, focusing on the luminance contrast ratio, calculated using the Michelson definition [26]:

$$\frac{(L_{max} - L_{min})}{(L_{max} + L_{min})} \quad (2)$$

where L_{max} and L_{min} are taken from the Y value in CIE XYZ color space, and represent the highest and lowest luminance.

Main Effects

Fig. 8 shows the main effect of background on both error (F(3, 69) = 23.03, p < .001, d = .353 error) and response time (F(3, 69) = 2.56, p = .062, d = 471 msec). Participants performed most accurately on the building background, and made the most errors on the brick background. A similar trend was found for response time. These findings are consistent with hypothesis 1 and hypothesis 2.

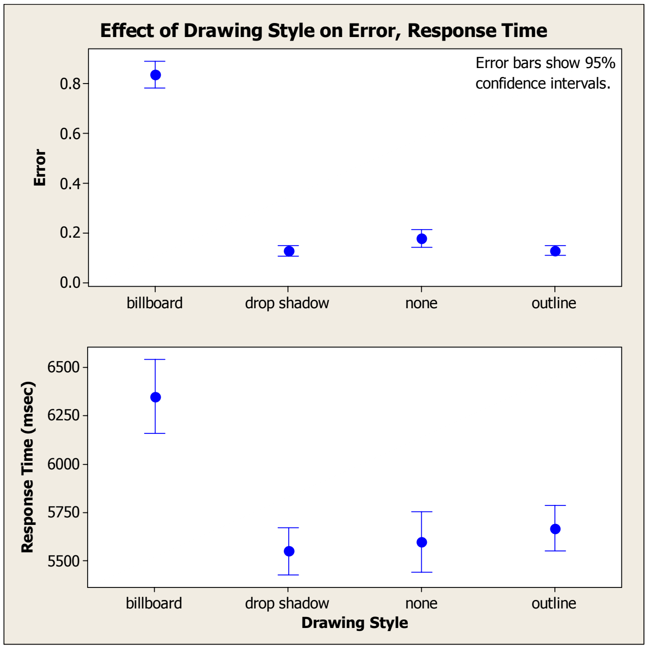

Fig. 9. Effect of text drawing style on error (*N* = 8064) and response time (*N* = 7324).

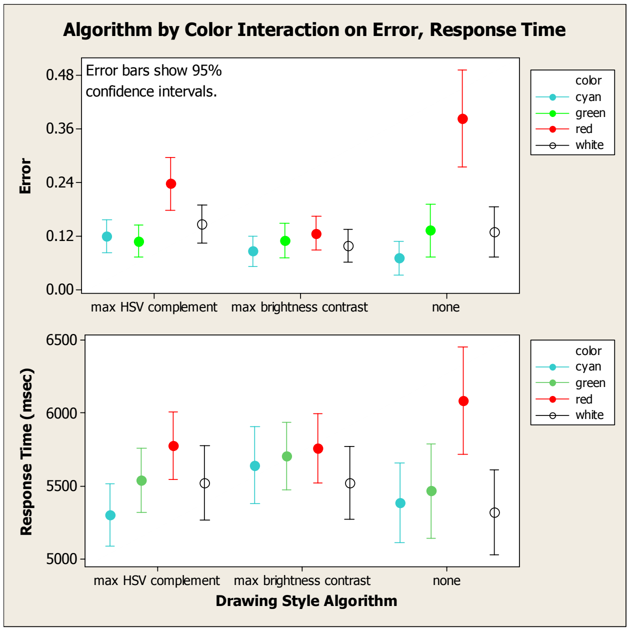

Fig. 10. Effect of drawing style algorithm by text color on error (*N* = 5760) and response time (*N* = 5615) for the trials where drawing style ≠ *billboard*. The right-hand column shows the effect of text color on error (*N* = 1152) and response time (*N* = 1109) for the trials were drawing style = *none*.

There was little difference in error under sidewalk and sky conditions (*d* = .089 error), and similar results for response time (*d* = 225 msec). We observed a relatively large amount of illuminance reflecting off the brick background, and we hypothesize that this illuminance, as well as the complexity of the brick background texture, explain why brick resulted in poor performance. Similarly, we hypothesize that the lack of reflected sunlight and homogeneity of the building background account for the lower errors and faster response times.

Contrary to hypothesis 3, there was no main effect of text color on either error (*F*(3, 69) = 2.34, *p* = .081, *d* = .075 error) or response time (F(3, 69) = 1.81, *p* = .154, *d* = 253 msec). However, when we examined the subset of trials where drawing style = *none*, we found significant main effects of both error (*F*(3, 69) = 5.16, *p* = .003, *d* = .313 error) and response time (*F*(3, 69) = 8.49, *p* < .001, *d* = 1062 msec). As shown in the right hand side of Fig. 10 (where algorithm=none), participants performed less accurately and more slowly with red text, while performance with the other text colors (cyan, green, white) was equivalent (*d* = .063 error, *d* = 166 msec). This result may be due to the luminance limitations of the Glasstron display, resulting in less luminance contrast for red text as compared to cyan, green, and white text. This result is consistent with the finding in our pervious study that participants performed poorly with red text [7],[8] and provides further design guidance that pure red text should be avoided in see-through AR displays used in outdoor settings. Furthermore, together with the lack of an effect of text color over all of the data, these findings suggest that our active drawing styles may enable more consistent participant performance across all text colors, which would allow AR user interface designers to use text color to encode interface elements.

Fig. 9 shows the main effect of text drawing style on both error (*F*(3, 69) = 152, *p* < *.001*, *d* = .711 error) and response time (*F*(3,69) = 11.6, *p* < *.001*, *d* = 797 msec). In both cases, participants performed less accurately and more slowly with the billboard text drawing style, while performance across the other text drawing styles (drop shadow, outline, none) was equivalent (*d* = .051 error, *d* = 118 msec). These findings are contrary to hypothesis 4. As explained in Section 4.3, our active text drawing style algorithms use the average background color as an input to determine a drawing style color that creates a good contrast between the drawing style and the background. Furthermore, the drawing style is a graphical element that surrounds the text, either as a billboard, drop shadow, or outline. A limitation of this approach is that it does not consider the contrast between the text color and the surrounding graphic. Both drop shadow and outline follow the shape of the text letters, while billboard has a large visual footprint (Fig. 7). Therefore, it is likely that in the billboard case, the contrast between text color and the billboard color is more important than the contrast between billboard color and background color (as discussed below), while the opposite is likely true for the drop shadow and outline styles.

Additionally, we propose that there are (at least) two contrast ratios of interest when designing active text drawing styles for outdoor AR: that between the text and the drawing style, and that between the text drawing style and the background. Both the size of the text drawing style and whether or not it follows the shape of the letters likely determines which of these two contrast ratios is more important.

Since our billboard style was not compatible with our back-ground-based drawing style algorithms, and because it exhibits a large effect size, we removed the bill-

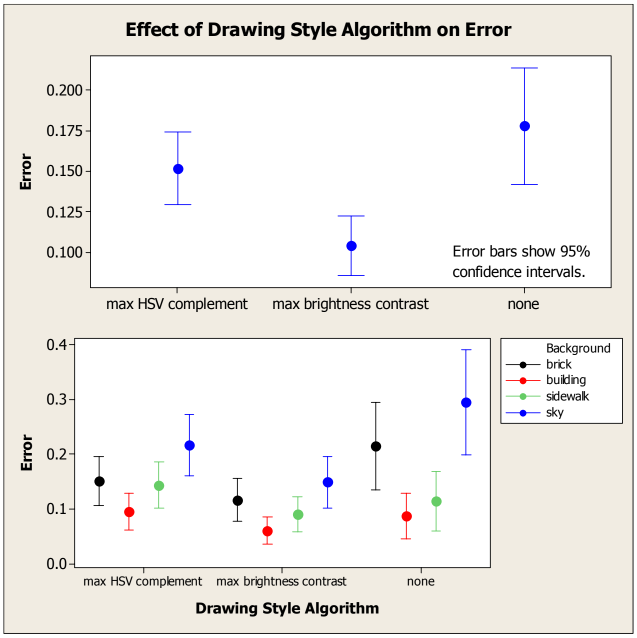

Fig. 11. Effect of text drawing style algorithm on error (N = 5760) for the trials where drawing style ≠ *billboard*.

board drawing style and performed additional analysis on the remaining data set.

Fig. 10 shows that drawing style algorithm interacted with text color using this subset of data, on both error (F(6, 138) = 2.96, p = .009, d = .313 error) and response time (F(6, 138) = 2.95, p = .010, d = 1062 msec). The effect size of text color was the smallest with the maximum brightness contrast algorithm (d = .040 error, d = 221 msec), followed by the maximum HSV complement algorithm (d = .129 error, d = 589 msec), and followed by text drawn with no drawing style and hence no algorithm (d = .313 error, d = 1062 msec). Fig. 11 shows that drawing style algorithm also had a small but significant main effect on error (F(2, 46) = 3.46, p = 0.04, d = .074 error). Participants were most accurate when reading text drawn with the maximum brightness contrast algorithm, followed by the maximum HSV complement algorithm, and followed text drawn with no algorithm. Tukey HSD posthoc comparisons [27] verify that *maximum brightness contrast* is significantly different than the other algorithms, while *maximum HSV complement* and *none* do not significantly differ.

It is important to note that the maximum brightness contrast drawing style algorithm does not exist by itself, but instead is manifested within the drawing style. More importantly, the algorithm resulted in fewer errors for the sky and brick background conditions (see Fig. 11, bottom), suggesting that there are some backgrounds where the addition of active drawing styles can provide a real benefit (although we did not find an algorithm-by-background interaction for this data set (F(6, 138) = 1.21, p = .304, d = .234 error)). Similar to the findings for text color, the effect size of background was the smallest with the maximum brightness contrast algorithm (d = .089 error), followed by the maximum HSV complement algorithm (d = .122 error), and followed by text drawn with no drawing style and hence no algorithm (d = .208 error).

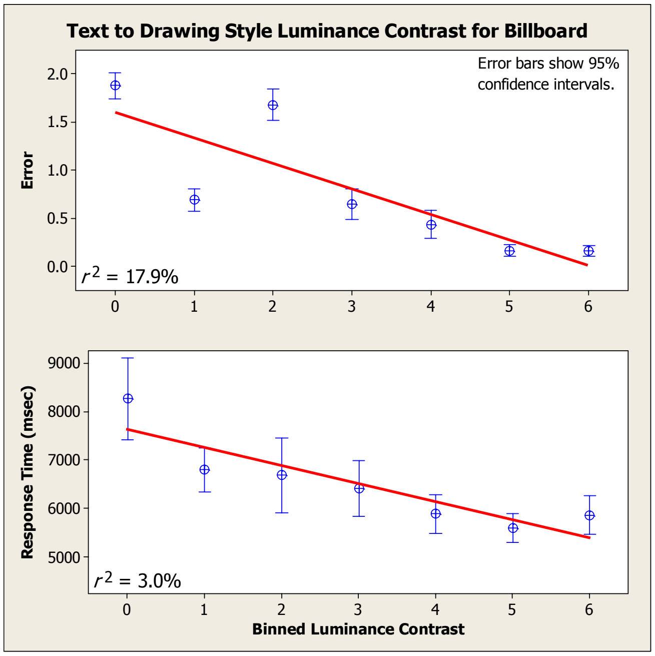

Fig. 12. For the billboard drawing style, correlation between binned luminance contrast ratio (calculated using text color luminance and text drawing style luminance) and error (N = 2304) and response time (N = 1709).

Taken together, these results show that when drawing style ≠ *billboard*, the maximum brightness contrast algorithm resulted in the overall best error performance (Fig. 11, top), as well as the least variation in performance over color for error and response time (Fig. 10), and the least variation over background for error (Fig. 11, bottom). More generally, these results suggest that the presence of active text drawing styles can both decrease errors and reduce variability over the absence of any text drawing styles (i.e., the *none* condition) — especially those active drawing styles that employ the maximum brightness contrast drawing style algorithm.

Contrast Ratio Analysis

To assist in out contrast ratio analysis, we first calculated all pair-wise luminance contrast ratios using (2). We then "binned" the luminance contrast ratios into numbered integer bins ranging from 0 to 10, by multiplying each ratio by 10 and then rounding to the nearest integer. For example, a luminance contrast ratio of 0.32 was assigned to bin 3, a luminance contrast ratio of 0.67 was assigned to bin 7, and so on.

The most compelling results of these analyses were for the billboard drawing style. As hypothesized above, we found that the contrast ratio between the text and the drawing style (i.e., billboard) affected user performance more so than, for example, the contrast ratio between text drawing style and background.

For the billboard drawing style, Fig. 12 shows a correlation between binned luminance contrast (calculated using text color luminance and text drawing style luminance) and both error (r^2 = 17.9%, F(1, 2302) = 504.3, p < *.001*) and response time (r^2 = 3.0%, F(1, 1707) = 53.0, p <

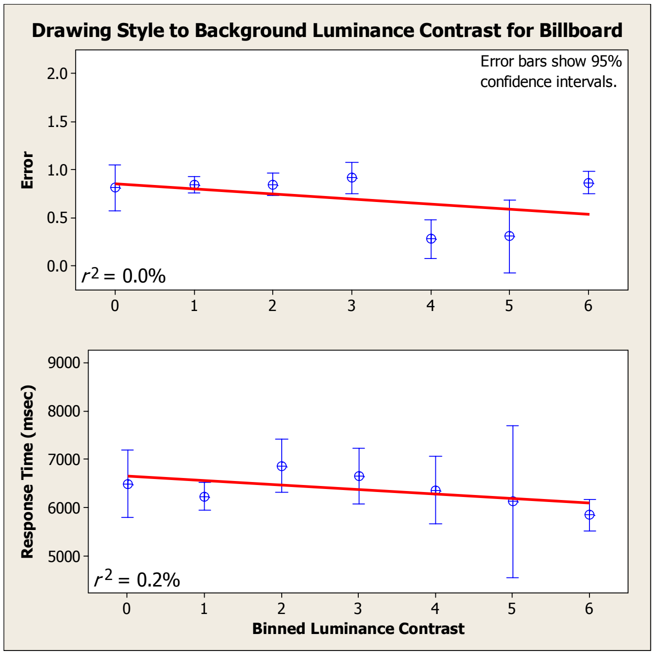

Fig. 13. For the billboard drawing style, correlation between drawing style to background luminance contrast for error (N = 2304) and response time (N = 1709).

.001). As the luminance contrast between the text and drawing style increased, observers both made fewer errors (d = 1.596 error) and became faster (d = 2226 msec). Fig. 13 shows the same analysis, this time conducted between the drawing style to background luminance contrast. Here the correlations were comparatively very weak (error: r^2 = 0.0%, F(1, 2302) = .01, p = .912; response time: r^2 = 0.2%, F(1, 1707) = 3.75, p = .053). As the luminance contrast between the drawing style and the background increased, observer errors decreased by d = 0.318 error, and response time decreased by d = 564 msec. Similar findings were found when we examined the contrast ratio between text and background for the billboard condition. Thus, for drawing styles with larger visual footprints (e.g., billboard), we can conclude that the luminance contrast ratio between the text and the billboard is a better predictor of user performance than the luminance contrast ratios between text and background, and drawing style and background.

3.9 Implications for Design

Our empirical findings suggest that the presence of active drawing styles effects user performance for text legibility, and that as we continue to research and design active drawing styles, we should take into account at least two kinds of contrast ratios: the contrast ratio between the text and the drawing style, as well as the contrast ration between the drawing style and the background. Although not explicitly explored here, there are likely times where a third contrast ratio (text color to background) is of interest – and indeed, in active systems may indicate whether or not an intervening drawing style is even needed at all!

Our findings also suggest that when using a billboard drawing style, maximizing the luminance contrast ratio between the desired text color and the billboard color supports better user performance on text reading tasks.

A finding consistent with our previous study [8], is clear empirical evidence that user performance on a visual search task, which we believe is representative of a wide variety of imagined and realized AR applications, is significantly affected by background texture (Fig. 8), text drawing style (Fig. 9), text color (Fig. 10), and active drawing style algorithm (Fig. 10 & Fig. 11). These findings suggest that more research is needed to understand how text and background colors interact, and how to best design active systems to mitigate performance differences.

4 Lessons Learned from Performing User-Based Studies to Inform Design

As part of the design process, and in preparation for user-based studies, it is advantageous to develop sets of design concepts using PowerPoint or other static mockups presented through an AR display, which can help form and refine an understanding of the design space. Moreover, these mockups can help designers identify design parameters that are good candidates for user-based studies. In some cases, it is possible to empirically cull candidate designs and identify candidates that are likely to result in better user performance (as compared to the designs that are culled). This was the approach we used when examining occlusion in the BARS [28].

User-based studies should employ user tasks that are representative but not so specific such that findings cannot be applied throughout the target application or domain. For example, our user task described in this paper required users to visually scan text, discriminate letters, identify patterns, and count target letters. While this task is not an actual task that would be performed in the BARS application, it is representative of visual scanning tasks that employ text as the main user interface element.

User-based studies should be conducted using the equipment that is most likely to be used in the application setting. By doing so, results of studies are more likely to be applicable to the final application and its supporting hardware. This is especially true for optical see through displays in outdoor settings, where the brightness, color gamut, and optical settings can vary widely. It is also true for any novel input devices that have form factors and or button arrangements. Our studies were run using a Glasstron display, which was not optimal for outdoor use. Specifically, the graphics displayed through the Glasstron are not sufficiently bright for all outdoor conditions. Moreover, we had to construct and affix "horse blinds" to the sides of the display to keep glare from entering participants' eyes through the sides of the display. In the Glasstron's defense, it was not designed for outdoor use but was the only display we had available at that time.

User-based studies should be conducted in the environment that is most likely to be used in the application setting. This is especially true for outdoor settings where lighting can vary depending upon location, time-of-day, etc. For our observations, lighting issues, setting, context, and so on, all potentially affect user performance. As a result, researchers should strive to match the experimental setting to the application setting as much as possible.

Another lesson we learned: do not employ tracking unless you need to! In cases where the scientific inquiry does not center around or hinge upon tracking (e.g., mobile settings, dynamic viewing of objects from various angles, etc.), we have found that eliminating tracker integration expedites the entire process and generally makes setting up and conducing user-based experiment much easier. Instead of tracking, we have either (1) fixed the users head position comfortably using some type of apparatus [5] or (2) presented pre-captured static images of the scene in order to physically align the user's view to a controlled view. The latter approach was used in the study presented herein.

When designing user-based studies to inform design, we recommend striving to keep the experimental designs small. Smaller experimental designs help force designers to focus on the most important user interface design factors. Indeed, the design space of the study presented herein initially had eleven independent variables that resulted in just over 6000 trials! With a mean response time of 5-6 seconds, that would have taken a subject at least eight hours for a fully within-subjects design! Here is another situation where the use of static mockups can help narrow the design space to a tractable set of factors and levels. Since smaller experiments equate to less time per subject (a maximum of 2 hours from time of arrival to exit is our rule of thumb), they afford running more subjects, which generally enhances the experiment's validity and power.

Moreover, smaller experimental designs are quicker to design, develop, and run, and are also faster and easier to analyze. Along these lines, when performing analysis, focus on the main effects as well as 2-way interactions. Look for the most obvious findings and then move on. The successful application of user-based studies within larger a usability engineering approach relies on the ability to iterate and evolve quickly.

Lastly, we have learned that by iteratively evolving a design space through user-based studies and evaluation, it is possible to gain insight on novel approaches to solving user interface design problems identified as part of the design/evaluation process. For example, the case study presented herein was the second of two studies performed to examine text legibility in outdoor AR. As described above, the second study employed an active AR testbed to alter the text in real time based on the real-world background texture. The need for an active systems resulted from our analysis of our first user-based study. A related example is the identification of the need for an optical see-through display that can display black (today's optical see through displays use black as transparent)[2]. Mobile outdoor AR would benefit greatly from a display that could present a larger color gamut, specifically in the darker regions. Indeed, both of these examples show that iteratively evolving the design space through user-based studies at least introduces the potential for innovation.

5 Applicability to Other Emerging Technologies

Along with augmented reality, there are other emerging technologies that would likely benefit from a usability engineering approach that utilizes user-based studies to optimize user interface designs. For example, as the use of cell phones and handhelds increase we see designs moving away from the standard WIMP metaphor towards more novel interaction techniques, such as the iPhone's use of accelerometers and touch sensing.

Handhelds are also starting to serve as the platform for mobile handheld augmented reality. Here again, interacting with information overlaid onto the real world with a small form factor will introduce some interesting design challenges – solved either through inspiration or empirical observations of users working with suites of candidate designs.

As ubiquitous computing matures, the notion of having access to computing power at all times, but without the bother of cumbersome cords or fixed location will require the development of novel display (in the broadest sense of the word) and interaction techniques. While some user interface and interaction techniques can be leveraged from related technologies, it is likely the case that guidelines for design much less standards for design will emerge overnight.

6 Conclusion

We have presented a modified usability engineering approach to design that employs a combination of user interface design, user-based studies and expert evaluation to iteratively design a usable user interface as well as refine designers' understanding of a specific design space. We have presented a case study involving text legibility in outdoor AR to illustrate how user-based studies can inform design. Finally, we have presented lessons learned in terms of the product (i.e., specific design recommendations and guidelines) as well as the process (i.e., recommendations on how to conduct a user-based experiment to inform design).

In the near term, we will be fleshing out more details of the proposed usability engineering approach and identify specific modifications needed to support design, development and evaluation of emerging technologies. Specifically, we will be examining some of the challenges of performing domain analyses as well as formative usability evaluation using these technologies.

We have recently conducted a follow-on study that systematically varied the contrast ratio between text color and text drawing style color. The goal of this study was to gain more insight on minimum contrast needed between text and a billboard background for effective task performance on text legibility tasks. We intend to analyze the results of this study, and optimally identify contrast

[2] Some optical see-through AR displays that support true optical occlusion, and hence can display black, have been developed as research prototypes (e.g., Rolland [29]).

thresholds or ranges in which user performance is unhindered. Assuming we identify these thresholds, we will use this knowledge to inform more sophisticated drawing style algorithms and to determine appropriate text drawing styles under varying environmental conditions.

With respect to further understanding the design space of text legibility in outdoor AR, we intend to design and conduct further studies on this topic. Specifically, we intend to design a study that systematically varies and controls the pair-wise contrast ratios between text color, text drawing style color, and outdoor background textures. By designing a study that explicitly controls these factors, we hope to be able to better understand the relative importance of each pair-wise contrast ratio for our given text drawing styles (including the none drawing style).

ACKNOWLEDGMENT

Our work on active drawing styles for outdoor AR has been supported by the Office of Naval Research (ONR) and the Naval Research Lab, under Program Managers Behzad Kamgar-Parsi and Larry Rosenblum. We also thank Debroah Hix for her continued appreciation and interest in "process-level" usability discussions and musings. We acknowledge Mark Livingston of the Naval Research Lab, as well as Steve Feiner of Columbia University, for loaning us the Glasstron AR display, without which this work could not have happened. Lastly, we thank Phil Keating, Lloyd Hipkins, and Claude Kenley for supporting our research and allowing us to conduct the user-based study described herein at their "Virginia Tech Weed Science – Turfgrass Noncrop Aquatics" facilities.

REFERENCES

[1] J.E. Swan II and J.L. Gabbard, "Survey of User-Based Experimentation in Augmented Reality", In *Proceedings of 1st International Conference on Virtual Reality, HCI International 2005*, Las Vegas, Nevada, USA, July 22-27, 2005.

[2] M.A. Livingston, L. Rosenblum, S.J. Julier, D. Brown, Y. Baillot, J.E. Swan II, J.L. Gabbard, and D. Hix, "An Aug-mented Reality System for Military Operations in Urban Terrain", In *Proceedings of the Interservice / Industry Training, Simulation, & Education Conference* (I/ITSEC '02), Orlando, FL, December 2–5, 2002.

[3] J.L. Gabbard, J.E. Swan II, D. Hix, M. Lanzagortac, M. Livingston, D. Brown, and S.J. Julier, "Usability Engineering: Domain Analysis Activities for Augmented Reality Systems". In *Proceedings SPIE, Stereoscopic Displays and Virtual Reality Systems IX*, Vol. 4660, p. 445-457, Andrew J. Woods; John O. Merritt; Stephen A. Benton; Mark T. Bolas; Eds. Photonics West 2002, Electronic Imaging conference, San Jose, CA, January 19-25, 2002.

[4] M.A. Livingston, J.E. Swan II, J.L. Gabbard, T.H. Höllerer, D. Hix, S.J. Julier, Y. Baillot, and D. Brown, "Resolving Multiple Occluded Layers in Augmented Reality", In *Proceedings of the 2nd International Symposium on Mixed and Augmented Reality (ISMAR)*, October 7-10, 2003.

[5] J.E. Swan II, M.A. Livingston, H.S. Smallman, D. Brown, Y. Baillot, J.L. Gabbard and D. Hix, "A Perceptual Matching Technique for Depth Judgments in Optical, See-Through Augmented Reality", Technical Papers, *Proceedings of IEEE Virtual Reality 2006*, Alexandria, Virginia, USA, March 25-29, pages 19-26, 2006.

[6] J.E. Swan II, A. Jones, E. Kolstad, M.A. Livingston, H.S. Smallman, "Egocentric Depth Judgments in Optical, See-Through Augmented Reality", *IEEE Transactions on Visualization and Computer Graphics*, Volume 13, Number 3, May/June 2007, pages 429–442.

[7] J.L. Gabbard, J.E Swan II, D. Hix, R.S. Schulman, J. Lucas, and D. Gupta, "An Empirical User-Based Study of Text Drawing Styles and Outdoor Background Textures for Augmented Reality", In *Proceedings of IEEE Virtual Reality 2005*, pp. 11-18, 2005.

[8] J.L. Gabbard, J.E. Swan II, and D. Hix, "The Effects of Text Drawing Styles, Background Textures, and Natural Lighting on Text Legibility in Outdoor Augmented Reality", Invited paper to *Presence: Teleoperators & Virtual Environments*, Vol. 15, No. 1, Pages 16-32, Spring 2006.

[9] J.L. Gabbard, J.E. Swan II, D. Hix, S. Kim, and G. Fitch, "Active Text Drawing Styles for Outdoor Augmented Reality: A User-Based Study and Design Implications", In Proceedings *IEEE Virtual Reality Conference, VR '07*. 10-14 March 2007, Page(s):35 – 42.

[10] W. Royce, "Managing the Development of Large Software Systems", *Proceedings of IEEE WESCON*, pp 1-9, August, 1970.

[11] B.W. Boehm, "A Spiral Model of Software Development and Enhancement", *IEEE Computer*, 21(5), 61-72, 1988.

[12] D. Hix and H.R. Hartson, "*Developing User Interfaces: Ensuring Usabililty Through Product and Process*", New York, NY, John Wiley & Sons, 1993.

[13] D.J. Mayhew, "*The Usability Engineering Lifecycle, a Practitioner's Handbook for User Interface Design*", San Francisco, CA, Morgan Kaufmann Publishers, 1999.

[14] J.L. Gabbard, D. Hix and J.E. Swan II, "User-Centered Design and Evaluation of Virtual Environments". Invited Paper to *IEEE Computer Graphics and Applications*, Volume 19, Number 6, pages 51-59, November / December, 1999.

[15] J.L. Gabbard, "*Taxonomy of Usability Characteristics in Virtual Environments*". Master's thesis. Department of Computer Science, Virginia Tech, 1997.

[16] A. Leykin and M Tuceryan, "Automatic determination of text readability over textured backgrounds for augmented reality systems", In *Proceedings of the 3rd IEEE and ACM Symposium on Mixed and Augmented Reality* (ISMAR 2004), pages 224 – 230, 2004.

[17] R. Azuma and C. Furmanski, "Evaluating label placement for augmented reality view management". In *Proceedings of the 2nd IEEE and ACM International Symposium on Mixed and Augmented Reality* (ISMAR 2003), pages 66–75, 2003.

[18] W. Piekarski and B. Thomas, B. "ARQuake: The Outdoor Augmented Reality Gaming System", *Communications of the ACM*, Volume 45, Issue 1, pp. 36–38, 2002.

[19] C.P. Halsted, "Brightness, Luminance and Confusion", *Information display*, March, 1993.

[20] E. Hecht, *Optics (2nd edition)*, Addison Wesley, 1987.

[21] S.J. Williamson and H.Z. Cummins, *Light and Color in Nature and Art*, Wiley and Sons, NY, 1983.

[22] B.L. Harrison and K.J. Vicente, "An Experimental Evaluation of Transparent Menu Usage", In *Proceedings CHI '96*, pp. 391–398, 1996.

[23] J.D. Foley, A. van Dam, S.K. Feiner, J.F. Hughes, and R.L. Phillips, *Introduction to Computer Graphics (2nd edition)*, Reading, MA, Addison-Wesley, 1993.

[24] B. MacIntyre, "A Constraint-Based Approach To Dynamic Colour Management For Windowing Interfaces", Master's thesis, University of Waterloo, Available as Department of Computer Science Research Report CS-91-55, 1991.

[25] B. MacIntyre and W. Cowan, "A Practical Approach to Calculating Luminance Contrast on a CRT", *ACM Transactions on Graphics*, Volume 11, Issue 4, pp. 336–347, 1992.

[26] A. Michelson, *Studies in Optics*. U. of Chicago Press. 1927.

[27] D.C. Howell, *Statistical Methods for Psychology (5th edition)*, Duxbury, 2002.

[28] D. Hix, J.L. Gabbard, J.E. Swan II, M.A. Livingston, T.H. Höllerer, S.J. Julier, Y. Baillot, and D. Brown, "A Cost-Effective Usability Evaluation Progression for Novel Interactive Systems", In *Proceedings of the Hawaii International Conference on Systems Sciences*, Big Island, Hawaii, January 5-8, 2004

[29] O. Cakmakci, H. Yonggang Ha, and J.P. Rolland, "A compact optical see-through head-worn display with occlusion support", In Proceedings *Mixed and Augmented Reality, ISMAR 2004*. Third IEEE and

ACM International Symposium on 2-5 Nov. 2004. Page(s):16 – 25, 2004.

Joseph L. Gabbard obtained his M.S. degree in Computer Science from Virginia Tech, with a specialty in Human Computer Interaction and Virtual Environments. He also holds a B.S. degree in Computer Science as well as a B.A. degree in Sociology. Mr. Gabbard is currently a research faculty in the Center for Human-Computer Interaction at Virginia Tech. Prior to this work, he was lead scientist at Virtual Prototyping and Simulation Technologies, Inc. His current research interests include usability engineering techniques for novel technologies, augmented reality, mobile computing, and embodied interaction.

J. Edward Swan II is an Associate Professor of Computer Science and Engineering, and an Adjunct Associate Professor of Psychology, at Mississippi State University. He holds a B.S. (1988) in computer science from Auburn University and M.S. (1992) and Ph.D. (1997) degrees in computer science from Ohio State University, where he studied computer graphics and human-computer interaction. Before joining Mississippi State University in 2004, Dr. Swan spent seven years as a scientist at the Naval Research Laboratory in Washington, D.C. Dr. Swan's research has been broad-based, centering on the topics of virtual and augmented reality, computer graphics, empirical methods, visualization, human-computer interaction, and human factors. Dr. Swan is a member of ACM, IEEE, and the IEEE Computer Society.