Projects in VR

Editors: Larry Rosenblum and Michael Macedonia

Evaluating Human Factors in Augmented Reality Systems

Mark A. Livingston

Naval Research Laboratory

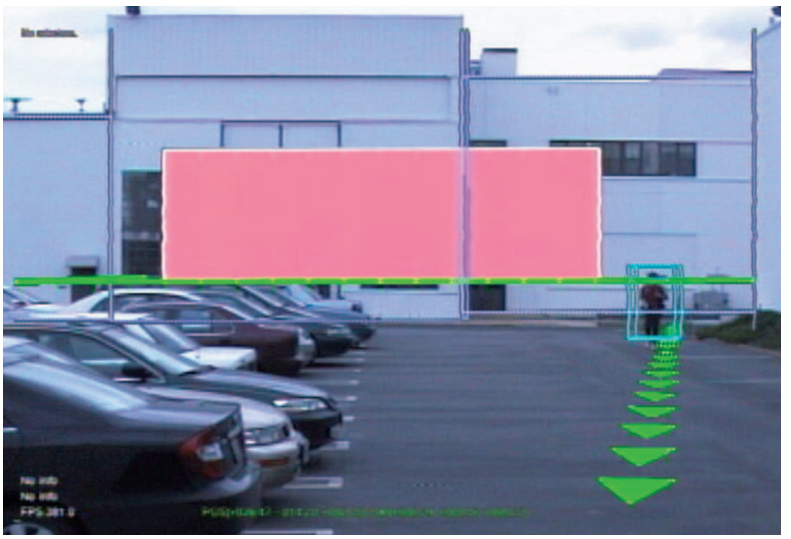
Augmented reality (AR) has been part of computer graphics methodology for decades. A number of prototype AR systems have shown the possibilities this paradigm creates. Mixing graphical annotations and objects in a user's view of the surrounding environment offers a powerful metaphor for conveying information about that environment. AR systems' potential still exceeds the practice. In fact, most AR systems remain laboratory prototypes. There are several reasons for this; two of the most prominent are that researchers need more advanced hardware than currently available to implement the systems, and (the subject of this article) the AR research community needs to resolve human factors issues. AR systems are usually interactive; thus, we must verify usability to determine if the system is effective.

Challenges

There are a number of difficulties to overcome when investigating human factors in AR systems. The hardware factors that often make AR systems difficult to use also impede human factors studies. Various display devices suffer from deficiencies in resolution, field of view, brightness and contrast, stereo vision (for example, no interpupillary distance or vergence adjustment), ergonomics, proper occlusion between real and virtual objects, or focus adjustment. Tracking systems struggle to provide range, accuracy, robustness, latency, ergonomic comfort, and user-friendly calibration. While certain systems have achieved good performance in one or some of these characteristics, it often comes at the expense of another—for example, high tracking accuracy over a short range, or high resolution display over a narrow field of view. This might work for some application classes—for example, AR over a desktop area—but clearly there is work to do on both displays and tracking systems.

Given these limitations, it should be little wonder that few user studies have been conducted using AR systems. Separating whether the performance (good or bad) is due to user interface factors, hardware factors, or other issues is a difficult challenge. At the Naval Research Laboratory, my workgroup has employed a methodology for our application that overcomes this inherent difficulty.

Our application (see Figure 1) assists a dismounted soldier in achieving and maintaining situation awareness (SA). SA is understanding the spatial environment, knowing what is happening, and predicting what might happen in the near future in your environment. We performed a domain analysis to determine which AR capabilities most naturally lend themselves to useful features for our end users. Our highest priority feature was implementing the metaphor of x-ray vision—being able to "see through" walls. This could solve the important problem of knowing the location of friendly forces in an operating environment.



1 Snapshot from our application, presenting information about the environment's geometric structure, showing a user's location and intended route. (The user is visible for illustrative purposes.)

This domain analysis gave us a cognitive task (achieving SA), which would be aided by a perceptual task (understanding depth relationships presented through graphical representations). For example, when achieving SA, a user might want to rendezvous with another user; this requires perceiving the other user's location and then performing the (cognitive) task of

 Published by the IEEE Computer Society

navigation. This vertical division of tasks helped structure our investigations of human factors in AR.

Identifying AR human factors issues

Human–computer interaction researchers have long understood that a usable interface lacking powerful features will generally enable the user to perform better on any given task than a poorly designed interface rich in features. User performance depends more heavily on being able to use the features of the software and hardware than on having many features available. This remains a problem for AR systems; the features are not fully tested.

For example, outdoor AR systems often enable both the typical head-up AR view and a 2D map. Suppose a user wants to instruct another user (perhaps at a later time) on how to navigate the environment. Theoretically, we could let the user draw a route through a complex urban environment either on the 2D map or in the head-up view. However, until we have solved the x-ray vision problem and implemented a suitable 3D cursor with which the user can specify 3D locations, the head-up drawing will be unusable. This is true despite the fact that the map registration to the real world occurs in the user's head, and the 2D map can't specify a route that goes up a staircase. The registration of a map to the world is a task our users—and many other people—are accustomed to performing, and these users will continue to create iconography for the 2D interface to represent climbing until we develop a complete and usable 3D AR interface. For many applications, until AR systems are more robust and fully functional, there is little to gain from application studies without solving perceptual issues such as the x-ray vision problem for this example.

This observation creates a quandary for AR system and hardware designers. If we attempt to test user performance with an AR system against traditional methods, we risk testing an interface that has not undergone rigorous testing of perceptual design and application utility against one that has. On the other hand, if we as a field do not demonstrate that applications benefit from AR presentations, then it will be difficult to justify funding the fundamental research on perceptual effects of presentation techniques that will result in well-designed AR interfaces. This leads us to the following two questions:

- How do we determine the most important perceptual needs of the AR user and the best methods of meeting those needs with AR interfaces?
- For which cognitive tasks are AR methods better than conventional methods?

From the previous logic, it seems we must answer these questions simultaneously.

Our solution is to conduct limited tests—perceptual tests of designs and task-based tests that use only the well-designed part of the user interface—to obtain the best insight into both good design for human perception and utility for cognitive tasks. Human perception is innate; variations in performance on perceptual tasks with different interfaces reflect the user interface's usability. Cognitive task evaluation will enable us to determine how users are performing in applications when compared to conventional methods. Performance on higher-level tasks can only be reliably tested against traditional methods for solving these tasks after the results of the perceptual-level tests inform the system design.

AR task building blocks

This strategy of limited tests will be more powerful if we identify a set of tasks that make up the functions any AR user would perform, both at the perceptual level and (via abstractions) at the cognitive level. This will enable transfer of techniques between applications and across perceptual mechanisms. AR has traditionally been primarily visual, but devices for the auditory and tactile senses are in use as well. Devices for smell and taste are not yet in common use.

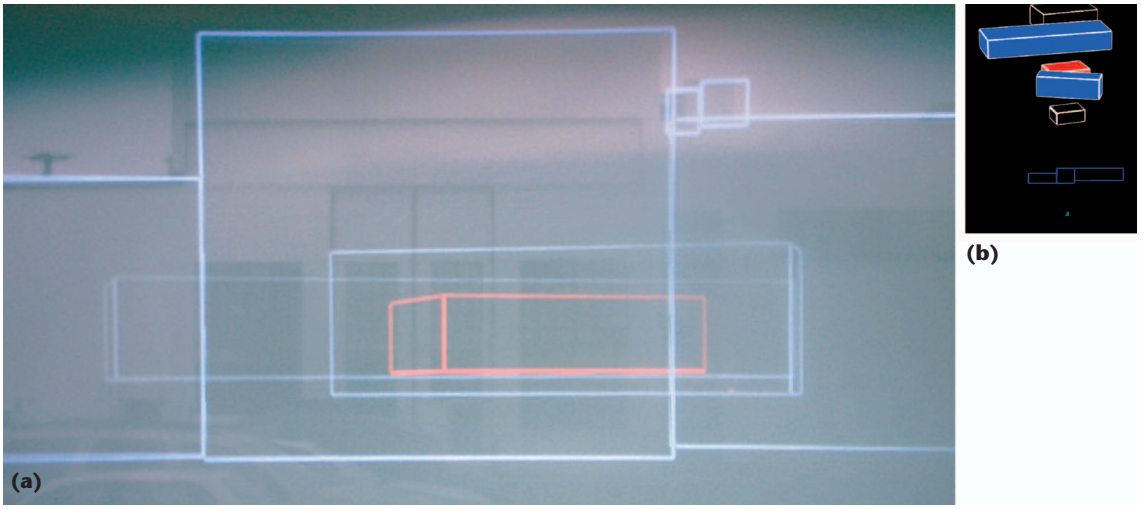
Visual tasks. Visual perception begins with resolving an object—knowing something is present, distinct from the surrounding environment. The next step is to identify (recognize) an object, its intrinsic properties (for example, brightness or color, size, or shape), and extrinsic properties (for example, position, orientation, or motion). Searching is a task based on these perceptual capabilities. You might also try to predict future states of the object based on current states (for example, predict time or place of collision with another part of the environment). Navigation is a task that we primarily accomplish visually (although it can be performed with tactile and auditory senses too).

Auditory tasks. Human perceptual capabilities are frequently shared by the senses; we can identify and determine distance of objects via hearing. Many systems commonly employ sounds as alert mechanisms to attract the user's attention. Analogous stimuli exist for visual (for example, blinking) or tactile (for example, a shoulder tap) senses, but many user-interface designers choose sounds for this. Humans naturally use type, distance, and direction of sounds to predict dangerous situations (for example, oncoming traffic or an approaching attacker), and this sort of prediction could form the basis for situation awareness or—in combination with visual cues—a navigation task.

Tactile tasks. Via haptic devices, we can apply virtual forces in an AR system. Similarly to the visual and auditory senses, we can also resolve, recognize, or describe some properties of objects via touch. We grasp and move objects in either precise or gross adjustments using tactile and kinesthetic senses—understanding a joint's position and motion as well as the forces present. When coupled with a concrete goal in the environment, manipulation becomes a cognitive task (for example, assembling a puzzle versus setting an object on a surface or holding it upright).

AR's fundamental characteristics

With this preliminary list of canonical AR tasks, we can begin to investigate performance specifications that



2 (a) Example of depth ambiguity created by the AR system, which can show the locations of objects or people hidden by urban infrastructure.[1] (b) Overhead view showing the correct and candidate solutions; such a view requires further integration and cognitive load to select the correct interpretation. One of our goals is to make the AR view as natural as an overhead view, but without requiring the cognitive effort to interpret depth, even for complex scenes.

the system must achieve or, more precisely, that the user must achieve with the system, which will in turn require certain system performance levels. Much work remains to be done on codifying system requirements to accomplish cognitive tasks, but we can identify perceptual tasks for which the AR system must determine user and system performance requirements for successful use of the system.

Resolving and identifying objects and their attributes requires sufficient acuity and contrast to discern objects from their surroundings. Thus, we need to determine the necessary user performance on these fundamental aspects of sensation. Tests of visual acuity, such as the standard Snellen eye chart, are well-known; similar tests of other visual and auditory measures exist. If tactile cues provide information to the user, analogous tests must be employed. These will lead to requirements for resolution and range of the displays (visual, auditory, or tactile).

AR systems display some real objects' attributes with virtual cues. To perceive these properly, the virtual cues must be aligned to the real environment—that is, the system must achieve a level of registration accuracy. How much registration error can exist before it degrades user performance to an unacceptable level will depend on the task. Whether the user can spend cognitive effort associating graphical elements with the real environment depends on the workload of the task. While the presumption remains that advancement of tracking systems will continue to reduce registration errors, there can be no expectation that perfect registration will be achieved. Representations of objects that adapt to registration error can enable the user to understand the intent of the system. The combination of representation and reduced error in registration that will lead the user to a correct perception of the spatial relationships and object properties has yet to be determined in user studies of our application or any other of which we are aware.

Task division for situation awareness and depth perception

We have applied this methodology to the situation awareness task and the x-ray vision metaphor described earlier.

Through a series of studies, we have attempted to discover the best methods to present far-field depth information in optical see-through AR. Our first pilot experiment used two domain expert categories—AR user interface designers and US Marines. Using experts from these fields helped us verify that solving the problem of x-ray vision is important and tractable. We asked these experts questions to measure their SA, as well as get qualitative feedback about the system. Subsequent studies elucidated exact parameters that improve user performance on the specific task of identifying ordinal or metric depth perception (see Figure 2).[1]

Our subsequent studies have shown that users' depth perception (measured by accuracy and variance with increasing depth) of virtual objects behaves similarly to depth perception of real objects, but that seeing virtual objects in the x-ray vision condition still significantly degrades depth perception (manuscript in submission). With this improved understanding of depth perception of virtual objects, we are planning to return to the structure of our pilot experiment where users looked at virtual objects hidden among urban infrastructure and identified which were closest, their directions of motion, and other properties within the real environment.

Sometimes we've simplified the system to precisely study factors. For some depth-perception tests, we eliminated the tracking system. Depth perception exists without motion parallax, although it's limited by the absence of such a cue. However, we get useful results about the utility of graphical representations and the other depth cues that remain, and we don't need to wonder what perceptual effects the jitter from a noisy tracking system has on the user.

Along the way, we found further problems to study at various points along the continuum of tasks discussed in the previous section. Head-worn displays used for augmented or virtual reality reduce the user's effective visual acuity.[2] How severe this problem is and how much diminished capability the user can tolerate and still perform a task are largely open questions. The task will heavily affect the requirements. This is an example of a problem that came up within the context of our research, but was not originally on our list of things to study.

As discussed previously, SA often leads to navigation, perhaps to come to the aid of another user. Thus, we assisted in the development of a search-and-rescue navigation task to test our system; this task asks a user to traverse the entire area and then retrieve an object in the area, returning by the shortest route.[3] The relevant observation from this test was not that users were slower but completed more of the required traversal with AR, but rather that details such as text layout and legibility clearly impeded the users' performance. This emphasizes the difficulty of performing such high-level tests without solving such details, but also shows the difficulty in foreseeing (and motivating the study of) such details before designing the test. Thus, we advocate iterating between low-level perceptual tests and high-level application tests.

Discussion

The long-term goal for many AR researchers is to build usable applications that people prefer over conventional methods and with which they perform tasks better. This requires identifying applications that can benefit from AR and building good interfaces to support those tasks. These dual objectives are sometimes at odds: How can we know which applications will benefit users until good interfaces exist? This shouldn't discourage researchers from trying new applications, but it does prescribe caution in understanding why an application was or was not successful.

Performance requirements on cognitive tasks remain open for most AR systems. Which tasks are appropriate for investigation depends on the application, but certain core tasks are emerging in the literature: visual search, navigation (especially for outdoor, mobile AR), manipulation (especially for desktop AR), and situation awareness. Accuracy and time are obvious performance metrics for almost any task; they apply easily to navigation and manipulation. SA has an accuracy metric, although the acceptable level is difficult to foresee for even a single scenario and might depend on an individual user's task, which could be different than another's in the same environment. Time in achieving the acceptable level is also a good metric, but the dynamic nature of the problem makes this a more difficult quantity to measure.

We have benefited from having ready access to domain experts who helped us identify the most likely tasks to improve good performance metrics. We then developed AR interfaces for those specific tasks. This helped us realize the issue of depth perception between real and virtual objects. For this, as in many tasks, improving the interface might benefit from simplifying the task to isolate performance factors that depend on the software from performance factors that require hardware capabilities we have yet to develop. We also tried to identify components of even this perceptual task, such as resolving objects, that could lead to even more low-level perceptual tasks for study. When this is complete, we will return to application-level tests that can demonstrate the utility and convince researchers in the field and potential users outside the AR field of the value of further study. In analyzing the application tests, we will look for factors that we had not previously identified, and iterate the design-and-testing cycle.

Thus, we balance our goals of finding applications that will benefit from the use of AR and studying these applications by abstracting the component tasks the user must perform. What the field still lacks is a battery of basic tests that determines the benefits of AR use and what hardware and software capabilities these would assume. The software capabilities should lead us to some further perceptual tasks to develop and test. Doing so would lead to an improved understanding of the application characteristics for which AR is a useful interface methodology and lead to better AR methods for these tasks. Ultimately, we must meet both of these goals to ensure the acceptance of AR for exactly those tasks for which it can be useful. ■

References

1. M.A. Livingston et al., "Resolving Multiple Occluded Layers in Augmented Reality," *Proc. Int'l Symp. Mixed and Augmented Reality* (ISMAR), IEEE Press, 2003, pp. 56-65.
2. M.A. Livingston et al., "Objective Measures for the Effectiveness of Augmented Reality," *Proc. IEEE Virtual Reality (Poster Session)*, IEEE CS Press, 2005, pp. 287-288.
3. M.A. Livingston et al., "Applying a Testing Methodology to Augmented Reality Interfaces to Simulation Systems," *Int'l Conf. Human–Computer Interface Advances for Modeling and Simulation* (SIMCHI), Soc. for Computer Simulation Int'l, 2005.

Readers may contact Mark A. Livingston at markl@ait.nrl-navy.mil.

Readers may contact Larry Rosenblum at lrosenbl@nsf.gov.