Video-Integrated System for Testing Augmented Reality (VISTA): A Rapid Testing Methodology for AR Platforms

Dylan Green[1,2(✉)], Kenton Bach[1], Audrey L. Aldridge[1], Christa M. Chewar[1], Michael Novitzky[1], and Thomas A. Babbitt[1]

[1] United States Military Academy, West Point, NY 10996, USA
{kenton.bach,christa.chewar,michael.novitzky,thomas.babbitt}@westpoint.edu, ala214@msstate.edu
[2] Brown University, Providence, RI 02912, USA
Dylan_Green@Brown.edu

Abstract. This paper evaluates a methodology aimed at enhancing the iterative development of augmented reality (AR) systems. We introduce the Video-Integrated System for Testing Augmented Reality (VISTA), a solution addressing two key challenges for AR developers: slow application deployment and the requirement for users to wear an AR headset for testing. Our approach leverages Holographic Remoting software and the Device Portal application to create short demonstration videos, enabling users to observe users interacting with and the physical world without deploying the application to a device or requiring headset use. To pilot VISTA, we surveyed novice software developers for feedback on an AR application we created, which yielded encouraging results.

Keywords: Augmented Reality · Mixed Reality · Iterative Development · User Testing

1 Introduction

Augmented Reality (AR) and Mixed Reality (MR) technologies offer seamless digital integration into physical environments, with applications ranging from neurosurgery to military operations [4]. However, integrating these systems into dynamic settings remains challenging [8,11]. Despite their potential, AR and MR systems face commercial adoption hurdles due to technical limitations and lingering negative perceptions, exacerbated by Google Glass's failure in 2013 [8]. While AR research is extensive, it largely focuses on complete applications rather than optimizing usability testing. Conducting early and frequent user testing remains difficult due to the need for individual calibration, high device costs, and risks like motion sickness [5]. Formative testing also faces barriers such as safety concerns, lengthy approval processes, and higher effort requirements compared to testing fully developed applications.

C. Stephanidis et al. (Eds.): HCII 2025, CCIS 2522, pp. 83–93, 2025.
https://doi.org/10.1007/978-3-031-94150-4_9

To address these challenges, we propose Video-Integrated System for Testing Augmented Reality (VISTA), a novel methodology enabling incremental feedback without full-scale user tests. By analyzing video recordings of AR applications, developers can gather usability insights without deploying software to AR devices or requiring participants to wear and calibrate them - an approach largely unexplored in AR research. To test our novel evaluation methodology, we developed an AR-focused usability and user experience survey, as no validated survey specifically for AR evaluation was found.

2 Related Work

Research on rapid prototyping and testing for AR systems is limited, with early efforts focused on pre-made asset toolkits like the Designer's Augmented Reality Toolkit (DART), which streamlined AR scene development [9]. More recent approaches emphasize low-fidelity prototypes, such as paper-based mock-ups that rely on the 'Wizard of Oz' methodology, though they fail to simulate real AR interactions [1]. In 2023, Garcia and Andujar explored user feedback on rapidly developed mock-up application designs, showing that users could provide valuable insights despite the lack of perfectly simulating an AR environment [3].

While these methods support different stages of development, they fall short in conveying the look and feel of a user's experience when interacting with AR applications and their surrounding environment. This challenge is further compounded by the difficulty of designing seamless AR displays, as they must be unobtrusive, maintain a user's visibility without blocking surroundings, and provide constant utility. These design challenges are why usability and user experience testing are critical in the design and development processes. UI evaluation requires both technical expertise and an understanding of human psychology, leading to the development of several psychology-oriented evaluation methodologies commonly used in the state of the art. Existing tools like SUMI [6], SUS [7], and PSSUQ [2] provide structured ways to assess usability while prioritizing end-user insights over developer expertise. With PSSUQ offering the most relevant insights for AR by evaluating usability, information quality, and interface quality, we primarily based our survey methodology on these factors.

Despite advancements in AR development tools, current methods overlook the value of showing users an application running on an AR device. We argue that usability evaluation should go beyond simplified prototypes. Given the success of psychology-driven UI assessments and the robust tools integrated into AR platforms, we propose a rapid usability evaluation methodology (UEM) that allows users to assess live AR applications more efficiently.

3 Methodology

To test our UEM (VISTA), we extracted five distinct features from a map program we had developed for the Microsoft HoloLens that would allow users to

mark friendly and enemy dispositions on a map. No previous usability testing had been performed, but we were hopeful that VISTA could be used to rapidly refine these features in an on-going development process.

In human-subjects studies for rapid AR prototyping and design evaluation, inherent risks with AR like motion sickness [5], can make it challenging for researchers to promptly attain Internal Review Board (IRB) approval. To mitigate these risks and accelerate AR prototyping, we devised a unique solution that could also streamline the IRB approval process - videos of experiment staff performing tasks in AR could be created for participants to watch. We administered our software video demonstration and survey to participants. We compared their responses to four AR experts who physically tested the software and completed a near-identical survey (with the exception of one question which was omitted). Their prior AR use and familiarity with the potential risks still allowed for expedited approval.

In both the demonstration video production and AR expert testing procedure, we sought to bypass the lengthy application deployment process; we used Microsoft's Holographic Remoting tool to run our application on a Hololens over WiFi. We then leveraged Windows' built-in Device Portal to screen record our interactions with the HoloLens application, enabling rapid video creation. This video demonstration approach offered two key benefits for iterative development: eliminating the need for software deployment and allowing participants to evaluate the application without wearing the HoloLens. While remoting introduced minor jitters and stutters, it did not significantly impact user feedback quality and greatly accelerated development and testing speed.

3.1 AR Usability and User Experience Survey Creation

To solicit user feedback, we created a novel usability/user experience survey consisting of 9 statements and 3 questions that review five different features of our AR application. Attached to the survey was a two-minute video thoroughly demonstrating all of the use cases and features within the application. At the beginning of each section was an image of the specific feature to indicate the feature being evaluated. The statements in the survey were designed to assess both the feedback of the user and how confident they were that they could provide feedback on the application. Seven of the statements in our novel survey are similar to established software evaluation surveys (SUMI, SUS, and PSSUQ), while two statements were specifically generated for AR systems and our methodology. Various statements from the established software evaluation surveys were tailored to fit the context of our iterative evaluation approach, as the individuals taking the survey did not actually use the application. These statements include the following.

S1. **"Overall, I am satisfied with how easy it looks to use this feature."** - derived from PPSUQ statement 1 [2]
S2. **"The location of this feature distracted me from the surrounding environment."** - generated for our rapid evaluation methodology

S3. **"The information provided by the feature was clear."** - derived from PSSUQ statement 9 [2]
S4. **"I would imagine that most people would learn how to interact with this feature quickly."** - derived from SUS statement 7 [7]
S5. **"The feature provided immediate and understandable feedback."** - loosely derived from PSSUQ statements 7, 8, and 11 [2] and from SUMI statements 3 and 23 [6]
S6. **"This feature looked awkward to use."** - derived from SUS statement 8 [7] and SUMI statement 47 [6]
S7. **"The interactions were consistent with how I expected the feature should work."** - loosely derived from SUMI statement 41 [6]
S8. **"The feature's design prevents incorrect actions from being taken during its use."** - loosely derived from SUMI statements 23, 26, and 28 [6]
S9. **"I believe I can provide effective feedback on this feature despite not interacting with it myself."** - generated for our rapid evaluation methodology

The purpose of these nine statements was to address 1) a unique obstacle presented by AR with integrating visuals into a dynamic environment, and 2) how confident participants felt in the overall assessment they were giving. Finally, at the end of the survey are three free-response prompts:

Q1. **"What improvements would you like to see in the demonstration video to have a better understanding of how to use the application?"**
Q2. **"Do you believe you can provide a software developer with feedback on their application based on this type of video? If so, what would you change in the demonstrated application?"**
Q31. **"Did you feel the questions allowed you to convey your full thoughts about each feature? Are there any other questions you would have liked to been asked?"**

These questions provided a holistic software assessment, acknowledging both standard software challenges, AR-specific issues, and insight into how confident users felt despite not physically interacting with the application.

3.2 Participants

The test population for this study consisted of 35 senior-level Computer Science (CS) and Cyber Science (CY) majors at the United States Military Academy. As this study was conducted using graduating seniors, most of the participant pool had completed nearly every core class in their technical majors, providing them with an understanding of both general coding principles and software development. Of the 35 potential participants, nine volunteered to participate. Four software experts who served as the control group for the study included faculty members who were either part of the Army Cyber Institute or West Point's Simulation Center. All have extensive software engineering or design backgrounds and are highly familiar with AR platforms.

3.3 Protocol

We distributed a Qualtrics form containing the software video demonstration, the usability/user experience survey, and corresponding instructions to our target population. The experts completed the survey (excluding S9) using knowledge gained from *actual interaction* with the application rather than with the video produced for the study. This served as the ground-truth evaluation of the application, independent of both the remoting tool and screen recording methodology used to create the videos. Approval from the United States Military Academy's Human Research Protection Program, reference number CA-2024-130, was obtained prior to conducting the study.

4 Results

In reviewing the results, we analyzed both the participants' survey responses as well as the comparison of participant feedback to the expert panel's responses. Four key questions guided the survey response analysis. The first regards the participants' capability and their perceived capability to identify bugs or potential improvements that would help enhance the specific AR features in question (see Sect. 4.2). The second question asks for an evaluation of the overarching trend in feedback, whether it be positive or negative toward the application (see Sect. 4.2). The third involves comparing participants' feedback to the control profile (see Sect. 4.3). Finally, the fourth question investigates whether the materials provided in the survey were sufficient for capturing participants' feedback about the software (see Sect. 4.4).

4.1 Reading the Likert-Scale Data

The graphs (seen in Fig. 1) include survey results where S2 and S6 are negatively phrased statements, meaning a 'strongly agree' response indicates dissatisfaction. In the bidirectional bar graphs, the red outlines represent the expert panels' responses (the control profile), while the thick vertical red lines mark control responses absent from participant data. With a study population of nine, bar percentages reflect response distribution. Note, in S6 of the Home Screen UI graph, one participant left a response blank; however, their data was retained as the omission did not affect overall validity. Because S9 was not an applicable question to ask the software experts, there is no control profile visualization for S9 in any of the figures.

4.2 Overarching Data and Trend Analysis

Upon looking at the data as a collective, the responses to the software were generally positive. Most importantly, users felt they were capable of providing feedback on the application despite not using it themselves. Throughout the entire survey, there are only two instances in which a participant "somewhat

disagreed" with S9 of our survey, as seen in the Map UI Platoon Icon interaction graphs. For each feature of the AR application, a majority of participants felt they could provide effective feedback. This is important to recognize, as it supports participants' other survey responses. Another interesting commonality across the results is that at least 40% of users found the *interaction* with AR objects to be awkward (S6, Map interaction and Platoon Icon interaction). On the other hand, the *UIs* were generally well regarded with a majority of participants responding that they were not awkward to use (S6, Home UI, Map, and Platoon Icon UIs). Additionally, out of all five features, there was only one with an instance in which a majority of users responded with negative feedback; as seen in S6 of the Platoon Icon Interaction graph, 77% of users felt that placing and moving icons looked awkward to perform.

Out of 35 responses that had clear majorities, 31 of the expert panel's responses were the same or one Likert degree different than the majority response from participants. With almost 90% consistency between the majority of responses from participants and the feedback of expert users who physically used the software, the software video demonstration appears to convey software usability effectively. The most significant disparity between the control and participant feedback is found in the Map User Interface. This highlighted one potential drawback of the VISTA methodology: behavior can seem intuitive when watching someone perform the intended actions, but it may be more ambiguous when not given explicit instructions.

4.3 Evaluating Specific Features

Individually, the feedback differed substantially based on the feature. For the following analysis, Likert scale values that are equal-to or off-by-one between the control and participant populations are considered to be consistent with each other.

Home Screen, Map, and Platoon Icon User Interfaces
The Home Screen was generally thought of as clear and easy to use, with individuals strongly agreeing that users would learn how to interact with the feature quickly, that it was responsive and provided understandable feedback, and that the interactions were consistent with the expected behavior. This menu was generally simple, and in 7 out of 7 interactions, the user feedback was consistent with the control profile. Experts seemed to think that interactions and feedback from the menu could be improved (S5 and S7) though they were generally happy with the feature.

While 7/7 results were consistent with the Platoon Icon Menu, the results regarding the Map Interaction Menu showed the most significant difference between the control and user populations. With the map menu, only 5/8 results were consistent between the control and participants, the lowest agreement observed in the study. We believe the map menu behavior was highly intuitive when watching the video demonstration, as the researchers knew exactly how to interact with the feature they designed. The AR experts, however, seemed

unable to decide if the feature looked like it worked as intended, was easy to use, or was awkward. (S1, S6, S7). We initially expected expert feedback to align most similarly with users for the User Interface features, and the idea that video demonstrations potentially can "explain away" unintuitive behavior stood out as an important takeaway for the VISTA methodology.

Apart from this, experts with substantial backgrounds in design and familiarity with AR software may have higher expectations for software UIs than our average participants. S6 of the Home Screen UI and S1 of Platoon Icon UI were the only 2 out of 22 statements regarding UIs where the expert's feedback was more positive than the participants' majority response. Though expert feedback was generally consistent with participants for User Interfaces, experts were slightly more negative and rarely more positive than participants who watched the demonstration video. Specifically, with the Map User Interface, experts identified substantial room for improvement.

Map Object Interaction. When evaluating the interaction with the physical Map Object, participants and software experts demonstrated consistency on 6/7 statements, both ascertaining the awkwardness of interaction and the potential distraction the object has from the environment. Most users saw the information on the map clearly, which was unexpected as we noticed blurriness on multiple occasions. The control differed from the user responses in S1, S6, and S8. Interestingly, S8 had a four-way tie, making the users seem uncertain about whether the feature adhered to fail-proof design principles, i.e., preventing users from taking incorrect action. Ultimately, the shakiness and the ease at which the system permits mistakes likely resulted in the control profile's less positive feedback with the Map Object (S1). Both individuals who watched the videos and the software experts who interacted with this feature seemed to recognize the challenges with this system.

Platoon Icon Interaction. The feedback regarding platoon icon interaction was both consistent and positive, with 6/6 statements with a clear majority demonstrating consistency. AR experts and video observers seemed to understand the use case for the feature and how the objects were supposed to behave.

The AR experts seemed slightly less confident than participants after physically interacting with the platoon icons that the information provided was clear and that users would learn how to interact with them quickly (S3 and S4). Additionally, the slightly more negative feedback in S6 indicated that the feature seemed awkward to use. Our research team expected this discrepancy; even as developers, it took our team time to get used to the interaction with the platoon icons.

4.4 Open Responses

The most consistent point of feedback participants gave when prompted with Q1 was that a narration or voice over would be helpful. Three individuals suggested including an audio track while two others suggested that they were content with the video as it demonstrated all the necessary features. Users seemed to feel

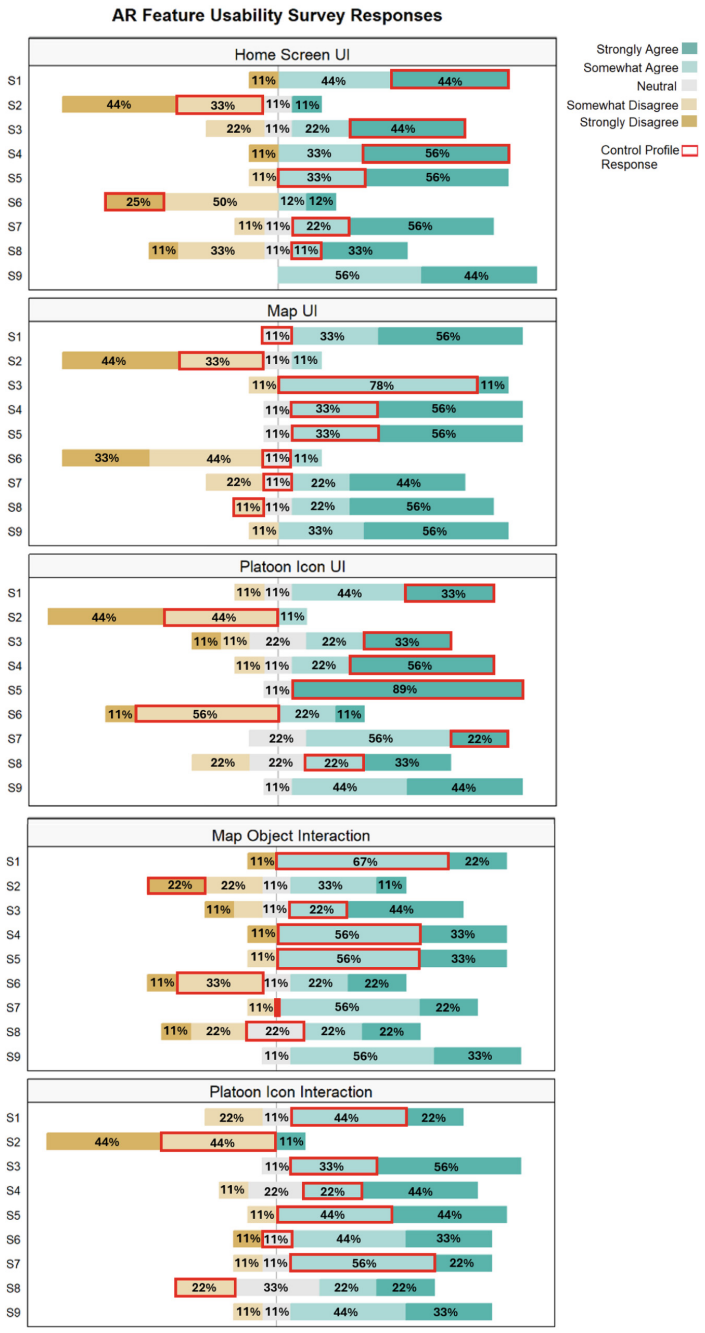

Fig. 1. Comparison of participant feedback to expert responses (red outline), after reviewing five different features in an AR prototype. Experts experienced the feature through the AR device, while regular participants used the VISTA method. (Color figure online)

like the stand-alone video was adequate, but that a voice over would enhance the overall experience. The majority of participants believed the demonstration would be useful for a software developer. Some participants suggested features like toggle buttons, updated icons, and a markup capability - all of which we had previously explored, attempted, or considered for future implementation given more time. The open-ended feedback demonstrated the clear ability for individuals who were familiar with computer science to provide general software feedback. Participants responded to Q2 and Q3 primarily with positive feedback. One participant suggested that S8 was generally not applicable from their experience with the evaluation. Others felt the questions were all-inclusive or wished they could provide written feedback on each feature. Although there were various answers to the questions offered in the survey, the general sentiment of the open-ended responses was that the videos served as a valuable resource for providing software feedback.

5 Discussion

With this initial study, we wanted to understand how well users could perceive, interpret, and articulate feedback regarding features under development in an AR system. We also wanted to demonstrate that a resource that could be produced quickly could also be useful for feedback generation. No script was needed for the video, and an individual could produce a similar quality resource in a matter of minutes. We were highly encouraged by the results, achieving an 88.5% similarity rating between software experts' and participants' feedback while lowering the accessibility barriers to software testing for AR systems.

These results indicate that AR software developers can efficiently solicit critically necessary user feedback without engaging in lengthy IRB approval processes or calibrating and running tests with physical AR devices. This methodology could be extended to context-specific applications that would be uniquely difficult to achieve study approval for. For example, to realistically evaluate the map software in the intended usage context, the researchers might have asked participants to test the application while walking through the woods wearing the AR headset–introducing significant risk to personal safety and requiring test administration time. Achieving such specificity during an iterative development process is substantially more feasible with VISTA.

Apart from the value offered by the VISTA methodology, we recognized areas for improvement. Instances in which the control profile deviated from participants' responses seemed to suggest that behavior that appeared trivial in the video demonstration was slightly less intuitive when physically interacting with the system. One way to address this would be to ensure video demonstrations deliberately show awkward behavior rather than only the intended behavior, allowing for more holistic system feedback.

Consistent with the feedback from the open-ended responses, we agree that including a voice-over and explaining the application would further improve the consistency between individual responses and the control profile of a developer

who actually worked with the application. The experience of administering the study allowed us to recognize that adding a voice-over would help developers identify specific challenges experienced with the application. Additionally, making videos that are far more in-depth about specific features would likely be insightful to developers as they could understand and walk through the flaws of various systems in the application. In conjunction with a voice-over, this would leave a user with a clearer picture of the challenges present within the system.

The optional nature of the survey and the short response-time lead to a relatively small population size. Per Nielsen's classic study on Heuristic Evaluation of UIs, three to five evaluators can provide similar quality feedback to that of a much larger cohort [10]. Thus, we concluded our sample population of nine participants and our cohort of four experts was sufficient. From this study, we determined that participants were reliably able to analyze features that required interaction with virtual objects by using a visual aide and that this pilot use of VISTA was successful. As such, this study suggests that individuals with technical backgrounds can successfully identify software shortcomings in AR platforms and identify usability issues without ever having to wear the headset. We believe this testing methodology could be effectively extended to individuals with more diverse academic backgrounds.

6 Conclusion and Future Work

This paper introduces VISTA, the Video-Integrated System for Testing Augmented Reality, a novel testing methodology that was evaluated in a pilot user test. VISTA uses software demonstration videos made with the holographic remoting tool and the device portal's recording software as a means for users to rapidly provide AR developers with critical feedback. Leveraging this, the methodology permits and encourages iterative development of AR applications - something that would have been immensely useful in previous development efforts.

Using a novel software assessment survey in VISTA's pilot evaluation (which can be generalized for other usability evaluations), participants were not only confident in their ability to provide feedback but demonstrated the ability to effectively critique applications with almost 90% of their feedback maintaining consistency with the control profile generated by experts. While participants had some challenges identifying awkward interactions, they generally recognized improvements that could be made to the application, as recorded by the open-ended questions of the survey. Importantly, this insight was gained through expedited IRB review and only 12 h of total test administration time, making it practical for iterative design evaluation during a development sprint. The quality of feedback generated by participants in this short time frame highlights the value of the VISTA methodology as a discounted usability test.

Future work proposed for continuing this research and for further improving and evaluating VISTA's feasibility includes adding voice-overs to the demonstration videos, creating videos on more specific features rather than on entire

applications, and developing a distributable survey for AR software assessment. Continued exploration of VISTA and other testing methodologies tailored for efficient use within iterative AR development will be essential for AR platforms to achieve their potential and become prevalent in our modern society.

Acknowledgements. The opinions in the work are solely of the authors and do not reflect those of the U.S. Army, the U.S. Military Academy, or the Department of Defense.

References

1. Freitas, G., Pinho, M.S., Silveira, M.S., Maurer, F.: A systematic review of rapid prototyping tools for augmented reality. In: 2020 22nd Symposium on Virtual and Augmented Reality (SVR), pp. 199–209. IEEE (2020)
2. Fruhling, A., Lee, S.: Assessing the reliability, validity and adaptability of PSSUQ. In: AMCIS 2005 Proceedings, p. 378 (2005)
3. Garcia, S., Andujar, M.: Capturing quantitative data from UI prototypes for AR and VR using online remote user testing. In: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1543–1548 (2023). https://doi.org/10.1109/SMC53992.2023.10394035
4. Guha, D., Alotaibi, N.M., Nguyen, N., Gupta, S., McFaul, C., Yang, V.X.: Augmented reality in neurosurgery: a review of current concepts and emerging applications. Canadian J. Neurol. Sci./Journal Canadien des Sciences Neurologiques **44**(3), 235–245 (2017). https://doi.org/10.1017/cjn.2016.443
5. Kaufeld, M., Mundt, M., Forst, S., Hecht, H.: Optical see-through augmented reality can induce severe motion sickness. Displays **74**, 102283 (2022)
6. Kirakowski, J.: The software usability measurement inventory: background and usage. In: Usability Evaluation in Industry, pp. 169–178 (1996)
7. Lewis, J.R.: The system usability scale: past, present, and future. Int. J. Hum.–Comput. Interact. **34**(7), 577–590 (2018)
8. Liao, T., Iliadis, A.: A future so close: mapping 10 years of promises and futures across the augmented reality development cycle. New Media Soc. **23**(2), 258–283 (2021). https://doi.org/10.1177/1461444820924623
9. MacIntyre, B., Gandy, M., Dow, S., Bolter, J.D.: Dart: a toolkit for rapid design exploration of augmented reality experiences. In: Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology, pp. 197–206 (2004)
10. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 249–256 (1990)
11. Stewart, A.: Microsoft's Hololens-like army device gets poor marks from soldiers (2022). https://www.businessinsider.com/microsoft-hololens-like-army-device-gets-poor-marks-from-soldiers-2022-10. Accessed 1 Mar 2024