A Test Platform for the Evaluation of Augmented Reality Head Mounted Displays in Industrial Applications

Volker Paelke[1(✉)], Carsten Röcker[2,3], and Jendrik Bulk[1]

[1] Hochschule Bremen, Flughafenallee 10, 28199 Bremen, Germany
{volker.paelke, jendrik.bulk}@hs-bremen.de
[2] inIT - Institut für industrielle Informationstechnik, Hochschule Ostwestfalen-Lippe Langenbruch 6, 32657 Lemgo, Germany
carsten.roecker@hs-owl.de
[3] Fraunhofer IOSB-INA, Langenbruch 6, 32657 Lemgo, Germany
carsten.roecker@iosb-ina.fraunhofer.de

Abstract. This paper presents a test platform for the systematic evaluation of head-mounted displays (HMDs). The focus is on an augmented reality (AR) test application for assembly tasks, which supports tests that are flexible in terms of complexity and scope, thus enabling the realistic assessment of usability, comfort and ergonomics by the test users.

Keywords: Head-Mounted Displays (HMDs) · Augmented Reality (AR)
Test platform · Evaluation · Assembly

1 Introduction

There is a great deal of interest in augmented reality (AR) applications in research and teaching, for entertainment, as well as for use in industry [1–3]. Besides stationary [4] and hand-held systems [5], in particular head-mounted displays (HMDs) are gaining increased attention. For professional use in industry it is necessary to select displays that meet the requirements of operators and users in terms of usability, ergonomics, functionality and robustness. The evaluation of AR HMDs is complex, covering aspects like the quality of the display (including aspects such as field of view, resolution, contrast, visibility under different lighting conditions), handling, reliability, robustness and usability, as well as the wearing comfort and ergonomics (weight, balance, run time on batteries) [6]. While some of this information is provided as technical data by manufacturers this is usually insufficient to assess more complex aspect like comfort and ergonomics.

In this paper we report on a test platform that includes a suite of test applications for the systematic evaluation of AR HMDs in industrial settings. The test applications allow conducting user tests with standardized tasks, thus enabling easy comparison between different displays. To realistically assess aspects like wearing comfort and ergonomics users have to work with a system in a realistic setting for extended periods – this in turn requires test tasks that are scalable in complexity and duration.

The test suite supports the creation of customized tests tasks that cover a large range of task complexity and task duration.

2   State of the Art

Researchers recognized at an early stage that the quality of the display is of critical importance for the usability of an AR system. Already in 1995 Rolland et al. [7] examined both the technological issues involved in AR display usability, as well as central perceptual and human factors such as depth perception, user acceptance, and safety. In 1997 Azuma [8] discussed the advantages and disadvantages of various display technologies and identified important display qualities such as resolution and contrast. Numerous evaluation studies of AR systems exist in the literature and many of these include the evaluation of different displays (e.g., [9] or [10]).

Existing studies can provide general guidance in the selection of an AR display, such as selecting a suitable display type for an application (e.g. HMD vs. Handheld vs. Projection [11]), but are often of limited use in the selection of a display for a specific development project in industry. A key factor is the rapid development of AR hardware in recent years and the fact that AR hardware has only recently reached the maturity and reliability for productive use in industrial applications. Published evaluation and test results often cover out-of-date hardware or prototype systems.

The aim of our platform is therefore to support the evaluation of HMDs by means of a standardized and partially automated procedure. The use of AR displays to assist in picking and assembly tasks has been investigated in research for a long time [12, 13] and is now established as a product in the market (e.g. [14]), which makes this application well suited for test purposes.

3   Evaluation Platform Requirements

As part of our project, a test platform for HMDs is developed consisting of two parts: a hardware test bed that allows to measure key parameters of HMDs in a largely auto-mated way and an interaction part with standardized user tests. In this paper, we focus on the interaction part that is currently being validated in user tests with common AR HMDs like the Epson Moverio, Microsoft Hololens and Vuzix glasses. The hardware test bed has recently been completed as a functional prototype and consists of a platform on which a mannequin head is mounted with two high-resolution cameras at the eye positions (Fig. 1).

The platform is equipped with sensors and actuators. The sensors measure the mass and balance of the HMD, and the motion actuators allow to move the head to measure latency and lag. The cameras make it possible to automatically measure several important aspects such as field of view, contrast and occlusion.

The interaction part of the evaluation platform was designed to enable user tests that measure important qualities of AR HMDs that are difficult or impossible to derive from the raw data provided by the hardware test bed, e.g. wearing comfort and user fatigue. Wearing comfort and fatigue can only be realistically assessed by human users
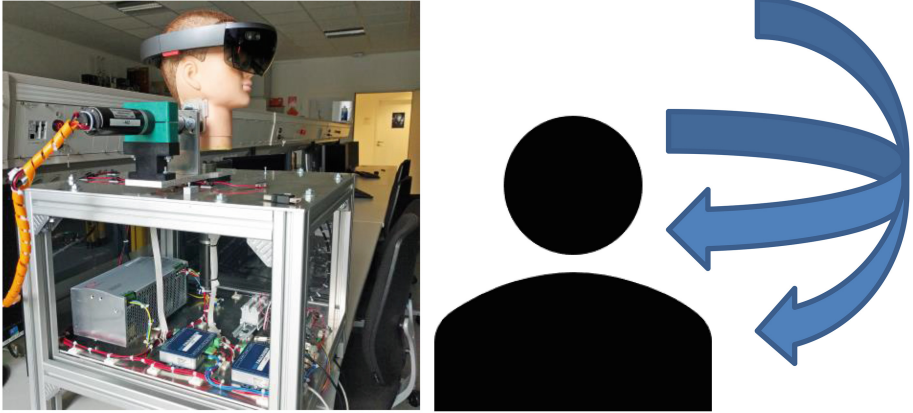
Fig. 1. Hardware part of the test platform and moving directions of the platform

after prolonged use of a HMDs in a realistic setting. Other qualities like readability of information and recognizability of graphics in different places in the field of view are also more realistically assessed by user tests.

Requirements for the test tasks were derived from in-depth discussions with practitioners using AR in industrial applications in industry, research and university settings. From the results the following requirements were established regarding the test tasks, the test environment and the implementation (Table 1).

The same process was also used to establish a collection of variables (qualities) that should be determined in the tests with users (Table 2).

Based on these requirements solutions were developed in an iterative user centered design process to cover the tasks, tests, implementation and data collection as presented in the following subsections.

4 Test Tasks

Typical tasks that are currently supported by industrial AR applications are navigation, information visualization and assembly or maintenance guidance. While these are all suitable as potential realistic test tasks our initial focus is on assembly and maintenance guidance tasks because these allow to address the additional test requirements of limited training, replication and limited setup in an easier way. Navigation tasks typically require extended test areas with significant set up costs and are therefore difficult to replicate. Information visualization tasks can be difficult to scale, especially if they are to be performed by untrained users. While simple information visualization tasks can be performed without previous training it is difficult to derive realistic and useful test scenarios that extend usage time to an hour or more. More complex visualization tasks are suitable for longer tests, but usually require previous training and domain expertise. Assembly and maintenance tasks can be designed to be performed by users with limited training and experience and are well suited for tests. In our first test

Table 1. Overview over requirements for the test tasks

Type of requirement	Description
Task requirements	• Realistic tasks: The tasks should be realistic tasks that are representative of typical industrial usage scenarios • Wide range of task complexity and duration: The tasks should be easily scalable in their complexity and duration
Test requirements	• No or limited training required: Test users should be able to complete a task without the need for previous (professional) training • No tools: Test users should be able to complete a task without the need for expensive and potentially dangerous tools • Limited setup: The task environment should be limited and self-contained, so that tests can be performed both in controlled lab settings as well as working production sites • Safe tasks: Test tasks should be safe to perform by large and diverse user groups with different backgrounds and experience • Replication: It should be possible to replicate the test environment at different locations
Implementation requirements	• Efficiently expandable: The design of the evaluation platform should allow for efficient creation of new test tasks • Reusable and affordable: The materials used by test users in the evaluation should be affordable and reusable across a large number of tests • Portability: It should be possible to adapt the test tasks to a wide variety of AR HMDs with different tracking and interaction modalities and port them to different operating system environments

Table 2. Overview over variables for user tests

Variable	Explanation
Completion time	Time required to complete a task
Errors	Errors made/corrected during the task and errors persisting in the final result
Fatigue	Measure of user fatigue
Attention	Measure of user attention/distraction
Cybersickness	Collection of data on cybersickness symptoms
Acceptance	Feedback on technological and social acceptance of the AR HMD
Perceived utility	Feedback on the perceived utility of the AR system
Perceived usability	Feedback on the perceived usability of the AR system
User experience	Feedback on the user experience of the AR system

application we have implemented several picking and assembly tasks. The assembly of objects makes for a clear and motivating scenario for test users and the duration and complexity of a test can (within reasonable limits) be adjusted both through the size and complexity of the object to be assembled and through repetition of the assembly

task itself. In future versions we plan to expand the scope of tasks to include information visualization/read-out tasks that can be combined with the assembly and maintenance tasks.

5 Tests

One potential issue with using industrial assembly tasks in user tests is that the tests have to take occupational health and safety issues into account, especially if (power) tools are involved. Regulations often require previous training, supervision and special insurance, which make it difficult to recruit large and diverse user groups for testing. Since 2001, we have successfully used construction toys like fischertechnik [15], Lego [16] and Makeblock [17] in augmented reality demonstrators and test applications. A large and diverse group of users has used these construction toys in AR applications over the last 3 years in which we conducted public outreach activities in the Mobile-GameLab. The MobileGameLab is a STEM laboratory in Bremen (Germany), where children, students, senior citizens and other groups can gather experience with emerging technologies like AR and also create their own applications and projects. Based on these experiences and requirements we developed the initial set of picking and assembly tests based on components from fischertechnik and Lego construction toys. As toys these are certified as 'safe-to-play-with' for age groups 8 and up. The toys are requiring no tools to assemble and allow for easy replication of tests across multiple instances and sites. For the creation of assembly test tasks, the different construction set systems we have identified specific advantages and disadvantages (Table 3):

Due to these differences we currently use Makeblock only for mobile robotics projects and focus on the use of Lego and fischertechnik for the test platform. Fischertechnik seems to be best suited for a generic test platform, because a flexible subset of elements (suitable for placement in a typical assembly workplace) allows to cover a wide variety of models. Lego seems to be well suited for more specialized casual applications, e.g. use in demonstrators and fair presentations, where a specific set of blocks can be acquired for the specific assembly scenario.

6 Implementation

The use of construction toys allows for an easy modification and extension of assembly tasks to adapt to specific requirements and test scenarios. The elements of both fischertechnik and Lego models can be easily reused across many test and are affordable, especially if second hand elements can be purchased in bulk. This allows to create a library of 3D models that can then be used to create picking and assembly instructions. To support a wide variety of AR HMDs with different tracking and interaction modalities on different platforms and different operating systems the test applications are implemented in Unity, because Unity support is available for all current mainstream AR HMDs. Unity also provides support for other platforms like mobile devices which can often be of interest as an independent reference.

Table 3. Overview over different construction set systems with specific advantages and disadvantages

Construction set	Advantages/disadvantages
Lego	• individual bricks widely available in bulk both new and second hand • extremely large collection of objects with assembly instructions available • well known and familiar as toys to many test participants, easy to get started for new users • small give-aways can be great motivators, especially for casual tests in fair settings, losses in less controlled settings common (everyone knows someone who could use a hand full of Lego bricks…) • well suited for objects with a focus on appearance, somewhat limited for complex mechanics (e.g. mobile robots) due to limited solidity of models • very large number of different bricks available, making the selection of a flexible subset suitable for a large set of different tasks difficult • parts are very simple to assemble, it is not possible to show a more different task than sticking bricks together
Fischertechnik	• individual elements widely available in bulk both new and second hand in Germany, less common in other parts of the world • large collection of objects with assembly instructions available, active communities in Germany and the Netherlands • not that well known to many test users, viewed as less toy like and more professional, easy to get started for new users • more difficult to create small affordable give-aways, less losses - possibly due to perception as professional parts vs. toys • well suited for complex mechanics (robust and durable assemblies), more limited for models with focus on nice appearance • limited number of standard elements, allowing for a flexible subset from which many different models can be constructed • many different parts with special techniques for assembling but still simple enough to do without any Fischertechnik experience
Makeblock	• currently no established second-hand market for individual elements • smaller community, strong focus on mobile robotics models • largely unknown to many test users, viewed as less toy like and more professional, requires some instruction to get started for new users • requires some tool use, assembly more involved and time intensive • very difficult to create small affordable give-aways, less losses - possibly due to perception as professional parts vs. toys • very well suited for mobile robots (very robust and durable assemblies), less flexibility for other models • limited number of standard elements, but also limited coverage of possible models

The Unity development environment and the large collection of tools available for it allow for fast and effective creation and modification of test tasks. For future versions we plan to further automate the creation of appropriate visualization and instruction elements from the 3D models.

7   Data Collection

There is a large collection of variables (qualities) that can be of interest in the evaluation of an AR HMD. A small set of these can be captured directly in the test application, e.g. Task Completion Time. However, most variables are best measured by prompting users to provide feedback, e.g. user experience (UX) and technology acceptance, while others can be measured either through user questionnaires or sensor instrumentation, e.g. fatigue and user attention.

In the current version of the test environment we record those variables that can be derived directly from user interaction in the test software (#of user interactions, task completion time) and use questionnaires to capture the remaining variables. There is a set of widely used standardized questionnaires that address the variables of interest, e.g. IsoNorm and SUS to rate usability related variables, AttrakDiff for UX related variables and Nasa-TLX to measure the task load perceived by the user. However, presenting users with a complete set of these questionnaires that cover all variables of interest has proven to be impractical. Test participants would have to answer far to many different questions that are sometimes overlapping in the different questionnaires and the differences in wording can cause additional confusion, prompting users to abandon the questionnaire and drop out of the test. We have therefor developed an initial questionnaire (Fig. 2; currently in German only) that aims to cover the variables of interest in a coherent way, with wording adapted for industrial AR applications and with the number of questions reduced to a practical minimum. For the future we aim to refine the questionnaire by enabling test designers to limit the number of questions depending on the variables of interest in a specific test setup, by validating the questionnaires results against the established standard questionnaires and by translating the questionnaires to other languages, starting with English.

8   Experience with the Test Setup in Different Configurations

Figures 3, 4 and 5 show different test setups, using a variety of AR glasses (and alternative techniques like projection) in a variety of settings. Our approach is flexible enough to enable user tests in all these settings with minimal adaptation. A simple ad-hoc setup (Fig. 4) allows conducting tests everywhere, without the need for additional infrastructure. Such an approach is especially useful in early exploratory stages of evaluation. A low-cost workspace setup provides a realistic simulation of an industrial workspace in a way that can be easily and cheaply replicated (Fig. 5). Such a setting is especially useful to conduct tests with large numbers of participants or at different locations. Tests in a real factory environment (Smartfactory OWL, Fig. 3) enable more realistic tests and are especially useful to validate the external validity of previous test results. Different setups have been used with multiple Lego and fischertechnik models in different tests and demonstrations, e.g. at the Hannover Industry Fairs in 2014, 2015, 2016, 2017, in the SmartFactory OWL since 2016 and in the MobileGameLab since 2017 with a wide variety of users and test tasks.

Fig. 2. Questionnaire (currently in German only)

9   Observations and Outlook

Experiences in tests with users have shown that even inexperienced users can quickly understand how to carry out complex assembly tasks using AR visualization. The central objective in our test scenarios is currently a realistic use of the HMDs over a long period of time in order to enable the test users to be able to evaluate aspects such as usability, wearing comfort and ergonomics of the HMD.

This goal has been achieved. We have conducted extensive tests with AR displays including the Microsoft HoloLens, Epson Moverio and Vuforia, as well as custom built displays.

Fig. 3. AR assembly system with AR glasses (left) and projection (right) in SmartFactoryOWL
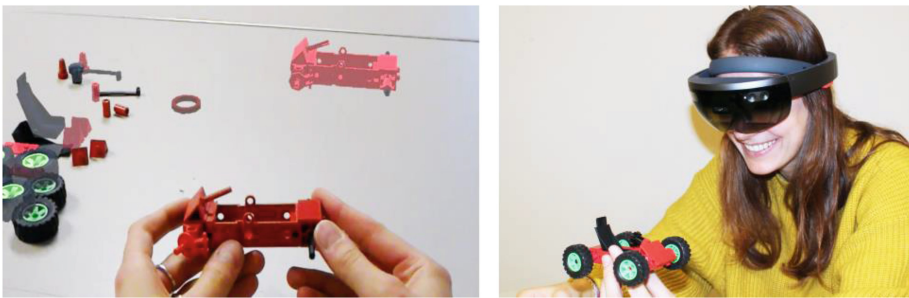

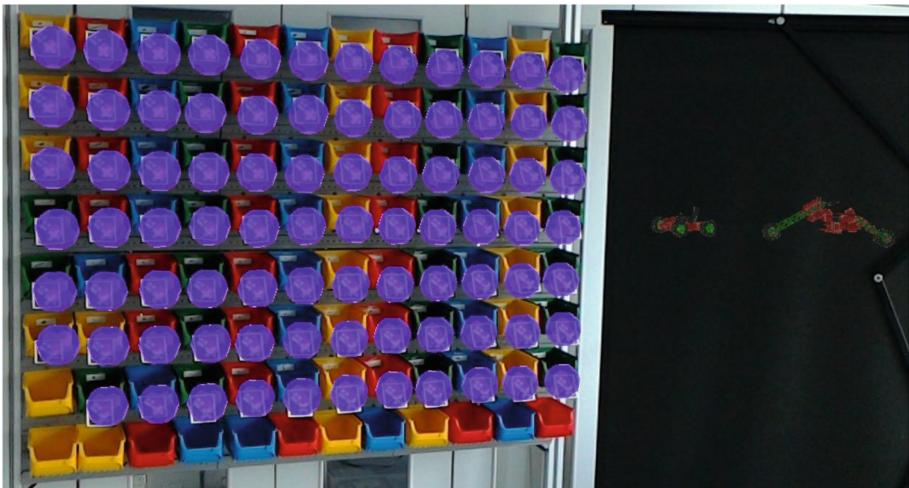**Fig. 4.** Minimalistic test setup with Microsoft Hololens

Fig. 5. Low-cost workspace with Microsoft Hololens

The use of assembly tasks with construction toy systems as an application scenario has proven to be valuable, since it allows efficient creation of test tasks at different levels of complexity and cost-effective creation of test workspaces.

The visualization techniques used in the assembly tasks have so far not been subjects of investigation and we have used simple techniques that are easy to implement and that have proven to be easy to understand in previous demonstrators. In the minimalistic setup (Fig. 4) we use simple animation of augmented 3D 'doppelganger' objects of the building blocks to be moved. This presentation is easy to create and well understood, but can be tiring in longer test tasks, as the users see a lot of animation all the time. Test users have also commented that it would be useful if the last component that was assembled would be highlighted.

Rack based test workspaces with boxes (Figs. 3 and 5) can use more simplistic information to indicate the required part (from box number to visual augmentation of the box) that is less tiring for users in longer test tasks. Such a setup also allows to extend tests to wearable displays without camera and tracking functionality, since the clearly identifiable boxes also allow picking by direct instruction (e.g., box A3) and without AR visualization. For the future we plan to extend the test framework with a wider set of visualization options to experiment with the usability of visualization techniques in addition to the display hardware.

References

1. Büttner, S., Sand, O., Röcker, C.: Extending the design space in industrial manufacturing through mobile projection. In: Adjunct Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI 2015), pp. 1130–1133. ACM Press (2015)
2. Büttner, S., Mucha, H., Funk, M., Kosch, T., Aehnelt, M., Robert, S., Röcker, C.: The design space of augmented and virtual reality applications for assistive environments in manufacturing: a visual approach. In: Proceedings of the International Conference on Pervasive Technologies Related to Assistive Environments (PETRA 2017), pp. 433–440. ACM, New York (2017)
3. Fellmann, M., Robert, S., Büttner, S., Mucha, H., Röcker, C.: Towards a framework for assistance systems to support work processes in smart factories. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) Machine Learning and Knowledge Extraction. Proceedings of the International Cross Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE 2017). LNCS, vol. 10410, pp. 59–68 Springer, Heidelberg (2017)
4. Flatt, H., Koch, N., Röcker, C., Günter, A., Jasperneite, J.: A context-aware assistance system for maintenance applications in smart factories based on augmented reality and indoor localization. In: Proceedings of the 20th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA 2015), Luxembourg, 8–11 September 2015, pp. 1268–1271 (2015)
5. Sand, O., Büttner, S., Paelke, V., Röcker, C.: smARt.Assembly – projection-based augmented reality for supporting assembly workers. In: Lackey, S., Shumaker, R. (eds.) Virtual, Augmented and Mixed Reality. LNCS, vol. 9740, pp. 643–652. Springer International Publishing, Cham (2016)

6. Paelke, V., Röcker, C., Koch, N., Flatt, H.: User interfaces for cyber-physical systems: expanding the designer's toolbox. In: at – Automatisierungstechnik, vol. 63, no. 10, pp. 833–843. De Gruyter, Oldenbourg (2015)
7. Rolland, J.P., Holloway, R.L., Fuchs, H.: Comparison of optical and video see-through, head-mounted displays. In: Proceedings of SPIE 2351, Telemanipulator and Telepresence Technologies, vol. 293, 21 December 1995
8. Azuma, R.: A survey of augmented reality. Presence Teleoperators Virtual Environ. **6**(4), 355–385 (1997)
9. Özbek, C., Giesler, B., Dillmann, R.: Jedi training: playful evaluation of head-mounted augmented reality display systems. In: Proceedings of SPIE, Stereoscopic Displays and Virtual Reality Systems XI, San Diego, USA, vol. 5291, pp. 454–463, May 2004
10. Büttner, S., Röcker, C., Sand, O.: Using head-mounted displays and in-situ projection for assistive systems – a comparison. In: Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments (PETRA 2016). ACM, New York (2016). Article No. 44
11. Paelke, V., Büttner, S., Mucha, H., Röcker, C.: A checklist based approach for evaluating augmented reality displays in industrial applications. In: 8th International Conference on Applied Human Factors and Ergonomics (AHFE 2017). Springer, Los Angeles, July 2017
12. Baird, K.M., Barfield, W.: Evaluating the effectiveness of augmented reality displays for a manual assembly task. Virtual Reality **4**, 250–259 (1999)
13. Radkowski, R., Herremaa, J., Olivera, J.: Augmented reality based manual assembly support with visual features for different degrees-of-difficulty. Int. J. Hum. Comput. Interact. **31**(5), 337–349 (2015)
14. Ubimax Frontline Productpage. www.ubimax.com/en/portfolio/#frontline
15. fischertechnik Homepage. www.fischertechnik.de
16. lego Homepage. https://www.lego.com
17. Makeblock Homepage. https://makeblockshop.eu