Usability of the ARgh! Augmented Reality system for on-the-job learning
Kaj Helin\textsuperscript{a}, Timo Kuula\textsuperscript{a}, Jaakko Karjalainen\textsuperscript{a}, Iina Aaltonen\textsuperscript{a}, Fridolin Wild\textsuperscript{b}, Juhani Viitaniemi\textsuperscript{c}, Antti Väätänen\textsuperscript{a}
\textsuperscript{a}VTT Technical Research Centre of Finland Ltd, Finland; \textsuperscript{b}Open University, UK

This contribution reports on design decisions and evaluation procedures as well as findings studying an on-the-job Augmented Reality learning system, namely the ‘ARgh!’ system. The objective of this paper is to assess system usability of the visual symbols developed that guide the user through physical action step-by-step, live, and on-the-job. This study contributed to the research programme of the EC-funded ICT project TELLME (‘Technology Enhanced Learning Livinglab for Manufacturing Environments’) which aims to develop and trial in authentic contexts (SME-driven human-centric and service-oriented manufacturing workplaces) an innovative cross-enterprise methodology and IT platforms for continuous education and training in heterogeneous business ecosystems, blending lifelong learning and participative co-creation aspects in ways that can address more business needs than traditional training.

Practitioner Summary: The main findings of the study lead to conclude, that most of the test subjects were able to perform the given task by following the symbol-based AR-system guidance with minimal or no intervention from the researcher. The results indicate that supplementary information (text/video/audio/pictures) is needed to facilitate effective and independent performance, enabling end users to better learn the meaning of all symbols. Based on the results, future development of the ARgh! system is needed in two areas: (1) the system should give feedback to the user during task execution, and (2) the user should be able to go back and forward in the work sequence.

Keywords: Augmented Reality (AR), Symbols, Mixed Reality (MR), Usability, Learning

1. Introduction
The test case of this study is maintenance work on heavy machinery, i.e. a rock crusher. In the test case setup, all manually manipulated objects in the driver’s cockpit have been made available as hardware, while the rock crusher, however, is visualised only virtually on the surrounding video walls. The test subjects are not familiar with the work task, thus resembling situations where people face new, previously unencountered challenges in their work. Existing skills and the use of new approaches such as opportunistic collaboration are needed to respond and to improve attitudes towards learning and change. Subjects are deploying the ARgh! system using a tablet see-through video-overlay (see Figure 1).

2. Methods
2.1 AR system and symbols
The ARgh! AR-system tablet version (see Figure 1 on the right-hand side) was used in this study. The tablet GUI enables the presentation with both rich HTML5 content and simple, contextualised AR content. This means that we have been able to test both presentation types at the same time and transfer the required functionalities also to an HMD version once they have been tested. The tablets are technically very suitable for AR. It is possible to track AR objects reasonably well thanks to the powerful processors and high resolution cameras typically found on tablet devices. Since the user is looking at the environment through the live video feed produced by the tablets camera, the overlays can be placed accurately without the need of any offset corrections. Since the AR layer is added using this handheld ‘window’, the user can simply put the AR view aside whenever it is not needed. Problematic for tablet use are those cases where you need both of your hands together (while following instructions). Since the majority of the on-the-job learning cases require...
hands free operation, ARgh! running on tablets cannot be the only option, although the tablet might offer the most reliable AR experience.

Figure 1. Pre-test during design phase in MR. Left: MR setup view of the virtual rock crusher. Middle: Performing task with the ARgh! AR-system. Right: View from ARgh! system tablet version.

Based on the literature review, there are no standardised AR-based symbols for work support available. A work ontology including task verbs was generated, extracted from use-cases, technical documentation, and human expert input. Matching AR-symbols were developed jointly with AR, learning, and human factors experts. The proposed symbols are depicted in Figure 2.

Figure 2. ARgh! system verbs in symbol form.

2.2 Usability methods and Questionnaires

The purpose of the methods was to provide information on user experience including technology acceptance, learnability, and usability, in particular about the use of AR-symbols developed. The methods applied for collecting data were observation, questionnaire, and interview. The observation was performed during the actual test situation by a group of researchers with special emphasis on use and readability of the symbols. While performing the tasks, the test participants were encouraged to think aloud, especially when facing problems. The user experience questionnaire consisted of question patterns from SUS, the System Usability Scale (Brooke, 1996) and QUIS, the Questionnaire for User Interaction Satisfaction (Chin et al., 1988). Two additional statements were added using a Likert scale:

- “The AR symbols were seamlessly integrated into the task.” (inspired by Gabbard & Hix, 1997)
- “I was aware of the phase of the task at all times during the execution of the task.”

Readability of the AR symbols was studied by showing participants a sheet of numbered pictures of the symbols. The form included all symbols listed in Figure 2, excluding the “Add XX” symbols, and the test participants were asked to explain the meaning of each symbol in one to three words in their mother tongue. It was permitted to give “no answer” as an explanation, if the symbol was not understood. A researcher wrote down the explanations and additionally marked down if giving the explanation was ‘easy’, ‘difficult’, or ‘confusing’, i.e. the participant had difficulties in arriving at a conclusion. In addition, the researcher marked down if the test participant mentioned recognising the symbol from the test situation (‘used in test’). The
order of explaining the symbols could be chosen freely, but was also noted down by the researcher. The readability of the symbols was evaluated by grouping the meanings given by the participants into four categories: ‘congruent’ (with respect to the intended meaning, see Figure 2), ‘comparable’, ‘incongruent’, and ‘no answer’.

As a final step, the participants were interviewed by asking them to freely answer the following questions:

- How did they feel about the AR system; what was good and what should be improved
- Did they manage to complete the task and how they would describe their task accomplishment
- What is their attitude and opinions towards audio and tactile feedback in this kind of solutions.

2.3 Laboratory setup and work task

Technical set up of Mixed Reality lab in VTT Tampere (see Figure 3 - left) includes: a three-screen system with active stereo (Barco RLM-W12 stereo projectors), 3D sound system, and optical tracking system (Vicon T20). Virtual rock crusher was running Unity3D (Helin et al., 2015). Manually manipulated objects were realised as real physical parts and devices. An iPad tablet was deployed for running the ARgh! system, with the application connected to a server for fetching the data about work task sequences. All the symbols were visualized on the tablet (see Figure 3 - right).

Main objective of the task was to assemble a hydraulic block and install it behind the maintenance hatch. The main idea of the task was the ability to perform it merely using the symbol-based instruction provided. Textual information on the ARgh! system was minimized to include only the number of the current step and the instruction to follow the symbols. The performed task was split in six action phases (See Table 1).

The test group consisted of seven (7) participants (6 male, 1 female), all working as researchers at VTT Tampere (Finland). None of the participants had previous experience with the ARgh! AR-system.

<table>
<thead>
<tr>
<th>Task step</th>
<th>Symbol(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take hydraulic block out from the right side box and place it on the table.</td>
<td>Unpack &amp; Lower</td>
</tr>
<tr>
<td>Take bolt out from the plastic box and screw it to right side of the hydraulic block</td>
<td>Pick, Screw &amp; Point</td>
</tr>
<tr>
<td>Locate the door and open it</td>
<td>Locate &amp; Open</td>
</tr>
<tr>
<td>Place / assemble the hydraulic block into the maintenance hatch</td>
<td>Assemble &amp; video</td>
</tr>
<tr>
<td>Close the door</td>
<td>Close</td>
</tr>
<tr>
<td>Plug correct cable to electric box</td>
<td>Plug, Allow &amp; Forbid</td>
</tr>
</tbody>
</table>

Figure 3. Left: Layout of lab-test set-up. Right: View from test subject point-of-view
3. Results

The results presented in this chapter are divided in four sections including user experience, AR symbols, observation, and audio and tactile feedback. These sections are presented in four subsections.

3.1 User experience

The user experience questionnaire consisted of two question patterns: SUS and QUIS. Additionally, two statements using Likert scale were included.

SUS (System Usability Scale) is a tool for measuring both usability and learnability (Brooke, 2013). The SUS scores (0–100) calculated from individual questionnaires represent the system usability. A Likert scale from 1–7 was used in this study for more accurate results than with 1–5 scale.

The mean SUS score for the Argh! AR-system was 58.3 (lowest 28.3; highest 81.7). According to validation studies, the acceptable SUS score is about 70 (Brooke, 2013; Bangor et al., 2009), and thus the system usability was clearly below the acceptable. From the individual SUS scores (from individual questionnaires), five out of seven (5/7) were below 70.

QUIS (Questionnaire for User Interaction Satisfaction) measures usability, learnability and technology (user) acceptance. QUIS was modified, using only the relevant questions from the viewpoint of this study (QUIS version 5.5 was used as a basis for this study). Altogether 21 questions with a scale mapped to numeric values of 1 to 9 were used.

<table>
<thead>
<tr>
<th>Overall Reaction</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terrible – Wonderful</td>
<td>6.1</td>
</tr>
<tr>
<td>Difficult – Easy</td>
<td>5.6</td>
</tr>
<tr>
<td>Frustrating – Satisfying</td>
<td>5.1</td>
</tr>
<tr>
<td>Inadequate power – Adequate power</td>
<td>5.6</td>
</tr>
<tr>
<td>Dull – Stimulating</td>
<td>6.6</td>
</tr>
<tr>
<td>Rigid – Flexible</td>
<td>6.3</td>
</tr>
</tbody>
</table>

Table 2. QUIS overall reactions (all items scale from 1-9)

The mean values of overall reactions to the system varied from 5.1 to 6.6 (Table 2), suggesting that the system is acceptable, but not showing great usability. It is noteworthy, that some test participants mentioned that they were sometimes thinking about the idea and potential of the system rather than the actual system on hand, when filling in the questionnaire. This could be seen for example in the scores indicating that the system is more stimulating (6.6) than easy to use (5.6).

According to the rest of the QUIS scores, the participants were most happy with the system speed (mean 7.7). The statement “Learning to operate the system” scored relatively high as well (mean 6.4), but not with very significant difference to the rest of the statements. Statements with lowest scores were “Experienced and inexperienced users' needs are taken into consideration” (4.6), and “Messages on screen which prompt user for input: confusing - clear” (4.7). The latter statement was mainly related to the AR-symbols. The remaining QUIS scores varied very little, from 5.0 to 6.1, supporting the idea of a system that is tolerable but far from excellent.

The two additional statements (scaled from 1-7) yielded the following results:

- “The AR symbols were seamlessly integrated into the task”: 3.0 (mean)
- “I was aware of the phase of the task during the entire execution of the task”: 3.9 (mean)
The low score related to AR symbols’ integration to the task seemed to be consistent with the low score from QUIS question “Messages on screen which prompt user for input”: The AR symbols were not experienced as informative or accessible enough on their own without further textual, visual, or auditive instruction.

The test participants were interviewed after filling in the questionnaire. Below is a summary of the interview answers. The system has potential despite of the problems in overall user experience. The interactive system was seen as better than a traditional manual. The short video clip for describing a specific sub task was generally appreciated. The system usability improved after learning the logic of the use of the system.

More instructions and guidance was needed for moving from current step (phase) to the next step. This was experienced probably the most critical defect of the system. However, it was also mentioned as a strength that the system let the user decide when to move on to the next step. Feedback from performing the task was hoped for as well. The system didn’t inform how successful the performance of the (sub-) task was. The goals or end-results of the whole test task and sub tasks were not explained, thus the test participants couldn’t know what is expected from them.

Argh! AR symbols were mainly considered as difficult to understand (see Section 3.2 for more detailed results). The advantage of the symbols is that they can be used universally without any language or cultural issues. The tablet was experienced as difficult to handle during task performance, since both hands were required for using the tablet. Some ideas for further development included the following:

- Virtual glasses or simply a smart phone could be easier to use than a tablet
- A ‘virtual AR hand’ that shows what to do could be better than AR symbols
- Audio feedback after for example a faulty performance
- An ‘infosymbol’ for explaining the meaning of a chosen symbol
- A combination of symbols and text (or audio)
- The view through the tablet should be as similar to the reality as possible; the interface should be integrated more seamlessly with the real world.

### 3.2 AR symbols

The AR symbols were analysed by comparing the meanings given by the participants with the intended meanings (as in Figure 2). Four categories were used: congruent, comparable, incongruent, and no answer; and the results are shown in Figure 4. To exemplify the categories, the symbol Lubricate was categorized as follows: “oiling”, “squeeze oil”, “put oil”, “lubrication” (all congruent), “put substance from tube” (comparable), and “splattering substance” (incongruent). There was also one participant who gave no answer.

Altogether 18 (of 27) symbols were given congruent meanings by over 50% of the participants. The participants were unanimous on the meanings of symbols Lift, Lower, Pack, Plug and Unplug. The symbols Lift and Lower, however, were often supplemented with phrases “with your hands,” “using two hands,” or “and keep it level”. On the other hand, none of the participants gave congruent meanings to the symbols Point, Unfasten, Screw, and Unscrew. Table 3 lists example misinterpretations.

Eleven symbols were rated with “no answer”. The symbol Drill was left unnamed by three participants, and the symbols Lubricate, Pick, Screw, Place, Assemble, Unscrew and Disassemble each by one participant. The “no answer” options were not equally distributed: 3 participants gave 10/13 of these answers.

Most of the participants chose to name the symbols row by row, and the researchers could not observe any symbol being significantly easier to explain than others. Some participants mentioned that they remembered seeing a certain symbol in the test, and they knew it’s meaning from that context, but might have otherwise answered differently. The symbol Allow, for example, was well remembered and its meaning was inferred from the context of seeing a green (Allow) and a red (Forbid) hand at the same time. Although four participants named it differently afterwards, two of them were uncertain whether the meaning was “allow”, and ended up choosing an alternative meaning. The pose of the hand seemed less relevant.

Other symbols that the researchers evaluated as having been difficult or confusing were Rotate CCW and Rotate CW. Many participants mentioned that it resembled the common symbol used for recycling and paid
less attention to the direction of the arrow. In addition, the symbol Unfasten was interpreted as “Use a tool” (or similar) by all participants.

![Figure 4. AR symbols arranged in descending order with respect to the congruence between the meanings given by participants and the symbol’s intended meaning. An asterisk (*) denotes that the symbol was used in the test setup.](image)

Table 3. Examples of alternative ways to interpret the symbols.

<table>
<thead>
<tr>
<th>Symbol</th>
<th># participants with congruent meaning</th>
<th># participants with incongruent meaning</th>
<th>Examples of other given meanings (# of answers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>5</td>
<td>2</td>
<td>Exit [1], Take out [1]</td>
</tr>
<tr>
<td>Assemble</td>
<td>5</td>
<td>1</td>
<td>Close the box [1]</td>
</tr>
<tr>
<td>Close</td>
<td>4</td>
<td>3</td>
<td>Open door [1], Enter [1], Put in between [1]</td>
</tr>
<tr>
<td>Allow</td>
<td>3</td>
<td>4</td>
<td>Item available [2], Drop [1], Lower [1]</td>
</tr>
<tr>
<td>Rotate CCW</td>
<td>2</td>
<td>4</td>
<td>Redo [1], Show again [1], Revolve CCW [1], …</td>
</tr>
<tr>
<td>Screw</td>
<td>6</td>
<td></td>
<td>Turn item CW [1], Twist and push down [1], …</td>
</tr>
<tr>
<td>Point</td>
<td>7</td>
<td></td>
<td>Lift [3], Look up [1], Go fwd [1]</td>
</tr>
<tr>
<td>Unfasten</td>
<td>7</td>
<td></td>
<td>Use a tool [3], Tool needed [2]</td>
</tr>
</tbody>
</table>

3.3 Observation

The test participants were given minimal information at the beginning of the task. The researchers needed to emphasize that the symbolic instructions were to be operated on physical objects instead of the tablet screen, which needed touching only to acknowledge a phase was completed. Many of the subjects first thought that the system is interactive and gives feedback, although the task script actually did not. This created confusion, especially when acknowledging a completed phase, but the system did not prompt to do so.

Based on observations and the participants’ monologue, the symbols were more often acted upon the basis of the symbol itself, and the effect of the context became apparent only after there was a conflict between the physical objects and the task. None of the symbols clearly stood out, but the symbols Locate, Open, and Allow & Forbid & Plug seemed to be the easiest and most unanimously comprehended.
The observations revealed that the participants usually noticed one dominant feature in a symbol, neglected other features, and relied more on the context. An arrow was especially strong a feature, and if the context allowed any interpretation, the participant acted on the first interpretation. Only after they felt the task was not proceeding as expected, they noticed the details such as hands. Sometimes the participant proceeded with the task without even looking at the symbols, and used the symbols more to check that all the actions were done as the symbols instructed. A somewhat similar thing occurred after the video was played. The symbol Assemble was barely noticed, because the preceding video was found very realistic and drew all the participants’ attention, and they struggled to remember what the video had shown. Some expressed a wish that the video could be replayed.

Some subjects found it confusing that two symbols (e.g. Unpack and Lower) were displayed simultaneously, and suggested that the symbols could be marked with numbers indicating the order of execution. Furthermore, some subjects did not notice the second symbol at all, because it might have been partially or completely out of view of the tablet’s display.

### 3.4 Audio and tactile feedback

All participants perceived the idea of utilising sounds as part of AR solutions’ feedback positively. Sounds are an explicit and instant feedback method for indicating that a phase of work is done. Some of the participants mentioned that audio feedback should be used context sensitively. Sounds should not be used in very noisy surroundings and the selection between loudspeakers or headphones as output devices should be based on case-by-case basis. In some contexts sounds can be useless or ignored. The participants suggested three different ways to utilise sounds: 1) Appropriate sounds for indicating positive or negative outcomes, 2) only one simple sound for catching one’s attention or acknowledging that a task is executed, and 3) speech output for giving detailed information and providing verbal instructional guidance.

The participants were encouraged to name good feedback sound examples, but there was no specific favourite referenced for positive and negative sounds describing, e.g., successful or failed action. Positive sound examples were easier to mention and sounds like chinks and jingle sounds were named commonly. Negative sounds were, e.g., buzzes reacting to ignoring action. Speech is appropriate, where more detailed verbal information, especially guidance, is needed and visual information is not available or appropriate.

Tactile feedback was preferred to be suitable for alerting one’s attention when needed. And after tactile notice more information could be given visually and/or aurally. Every participant responded that tactile feedback should not be the primary method and it is very difficult to use tactile sense for distinguishing positive and negative outcomes of the system.

### 4. Discussion

The overall user experience of the ARgh! AR-system could be interpreted as tolerable with many usability issues highlighted by the test results. However, most of the test participants liked the idea of an interactive learning system and considered it to be of great potential for assembly tasks as long as the usability issues are resolved. The most critical system defects affecting the user experience were lack of instructions for moving from current step to the next step, lack of task performance feedback, lack of descriptions of overall and subtask goals, and the poor readability of the AR symbols. It should be noted, that a ‘raw’ version of ARgh! AR-system was created for this specific test in order to study especially the readability and usability of the AR symbols. The main development version of the system already includes features that for example support understanding of the task better (such as providing proper textual description of the whole task and its phases), but these were intentionally removed.

Performing a rather simple assembly task mainly instructed by AR symbols unknown to the user and with very limited other information (text/video) proved to be difficult for test participants. The researcher’s guidance was necessary in order to accomplish the given task. The results thus indicate that supplementary information (text/video/audio/pictures) is needed for effective and independent performance. When such information is provided and the context of the performance is better defined, it could be assumed that the readability and usability of the AR symbols might increase as well. Naturally, understanding the logic of use
of the ARgh! AR-system would improve the overall performance as well. It is likely that the users' low level of performance was not affected that much by the AR symbols, but rather resulted from the lack of supplementary textual instruction (and lacking previous experience).

Most participants could explain the AR symbols relatively quickly and with no hesitation when they were shown on paper. The results indicate that most of the symbols are acceptable as such, and some of them can be used with minor improvement. Only one symbol (Unfasten) was given a different meaning ("Use a tool") from what was intended. The symbol Drill was very poorly recognized and in need of improvement, although it is difficult to estimate, whether an animated symbol would be more easily understood. In this symbol test, not all of the symbols were shown in context or animated. Although animations could help in understanding the symbols, showing the participants static symbols may give more realistic results to whether the symbol is acceptable of not. There seemed to be no difference in how congruently or unanimously the symbols were evaluated whether they had been used in the test setup or not.

The context of use allows for many misinterpretations. For example, the symbol Screw was interpreted by some of the participants so that the object (where the screw was to be placed) was rotated over two different axes, or rotated and pushed along the table. In addition to the lack of textual instructions on the symbols, the participants were given very little information on the task itself, which complicated the task even further. This may contribute to these misplaced actions, but the possibility of misinterpretations should be considered regardless. As mentioned above, adding supplementary information may reduce this risk.

When designing the symbols, and especially the context of use, three factors should be considered carefully. Firstly, there seems to be one dominant feature in many of the symbols that catches the eye, and other details are neglected if the context allows any inference. For example, the colours red and green seemed dominant over the hand pose, and arrows dominate other details such as hands. Secondly, when a symbol indicating opening or closing something, the direction of the arrow and the door or lid should match the reality, if possible. Thirdly, if more than one symbol is shown simultaneously, there can be confusion in which order the tasks are to be executed. The order could be indicated, for example, numerically or using animation, which would also help noticing if a symbol was left out of the tablet's view. In general, the usability of the AR system would be also improved if there was a chance to move a step backwards or replay the video to avoid loading the participant's work memory.

Results related to audible and tactile feedback indicated that they are useful to integrate into upcoming versions of the ARgh! AR-system. They can be used to supplement visual information or they can be utilised separately as notification or acknowledgement of proceeded work phase. Finally, user evaluations of how users perceive sounds and tactile in action is needed.

Acknowledgements

This study was partially supported by the European Commission’s Seventh Framework Programme FP7 under grant agreement 318329 ‘TellMe-project’.

References