A Benchmark for Interactive Augmented Reality Instructions for Assembly Tasks

Markus Funk[1], Thomas Kosch[1], Scott W. Greenwald[2], Albrecht Schmidt[1]
[1]University of Stuttgart (Pfaffenwaldring 5a, 70569 Stuttgart, Germany)
[2]MIT Media Lab (75 Amherst St., Cambridge, MA 02139 USA)
firstname.lastname@vis.uni-stuttgart.de[1] – scottgwald@media.mit.edu[2]

ABSTRACT
With the opportunity to customize ordered products, assembly tasks are becoming more and more complex. To meet these increased demands, a variety of interactive instruction systems have been introduced. Although these systems may have a big impact on overall efficiency and cost of the manufacturing process, it has been difficult to optimize them in a scientific way. The challenge is to introduce performance metrics that apply across different tasks and find a uniform experiment design. In this paper, we address this challenge by proposing a standardized experiment design for evaluating interactive instructions and making them comparable with each other. Further, we introduce a General Assembly Task Model, which differentiates between *task-dependent* and *task-independent* measures. Through a user study with 12 participants, we evaluate the experiment design and the proposed task model using an abstract pick-and-place task and an artificial industrial task. Finally, we provide paper-based instructions for the proposed task as a baseline for evaluating Augmented Reality instructions.

CCS Concepts
•Human-centered computing → HCI design and evaluation methods;

Author Keywords
Evaluation, Experiment Design; Benchmark; Instruction Giving; Remote Collaboration; Augmented Reality.

INTRODUCTION
Providing task instruction and imparting task-specific knowledge have long been important topics in HCI [4]. Today these are gaining particular practical importance in the area of assembly tasks. These tasks are more complex and more varied, and hence more demanding for workers. To overcome the complexity and to cognitively support the workers, assistive systems have been introduced [1, 5, 13]. We distinguish three main categories of technology-enabled task instruction and learning systems. First, remote collaborative work systems where e.g. a technician is in the field and is guided by an remote expert [8], Second Virtual Reality (VR) learning systems, e.g. for enabling risk-free learning in safety-critical tasks [2]. And third, context-aware stationary assistive systems that enable learners to get instructions for a specific manufacturing task [6]. Usually those systems use Augmented Reality (AR) technology for providing instructions, e.g. by augmenting a workplace with a projector [1].

Although many different approaches for teaching assembly tasks have been proposed, comparing them with each other is very cumbersome. This is mainly because each proposed type of instruction introduces a different reference task. Therefore, comparing instructions across different papers is nearly impossible because the used measures e.g. task completion time or number of errors, are *task-dependent* and cannot be easily compared across different scenarios. In order to address this problem, better models are needed – and a basic precondition for the construction of general models is a means for comparing the performance of different systems. Other areas in the field of HCI have already recognized the need for standardized tasks. For example, for evaluating text-entry techniques, the phrase set of MacKenzie et al. [11] is considered the standardized task. However, there is still no standardized task for providing instructions for assembly tasks. Therefore there is a need for a shared benchmark to allow the community to better compare system performance, and advance the state-of-the-art more rapidly.

To address this need, our paper provides the following two contributions: (1) We introduce a General Assembly Task Model consisting of *task-dependent* and *task-independent* measures, and (2) propose a benchmark experiment design consisting of two cheap and easily reproducible assembly tasks. Further, we provide paper-based instructions for the two assembly tasks that can be used by other researchers as a baseline for comparing new approaches.

RELATED WORK
Augmenting the workplace with AR instructions has been the topic of many research projects. Caudell et al. [4] first described displaying AR instructions for assembly processes in 1992. They use a manufacturing task inspired by aircraft manufacturing where the user has to drill holes. Their head-mounted display (HMD) indicates drilling positions and distances. Moreover, Zhou et al. [16] use a projector-based AR approach for visualizing welding spots. In their study, they use a metal car body part as welding task. Henderson et

MUM '15, November 30-December 02, 2015, Linz, Austria

DOI: http://dx.doi.org/10.1145/2836041.2836067

al. [10] use three-dimensional arrows and text-labels to instruct a worker in maintaining an armored personnel carrier turret. They display the instructions on a HMD. To interact with the instructions the user operates a wrist-worn controller that advances the feedback. With the proliferation of Google Glass, Zheng et al. [15] compared AR approaches for providing instructions using a centered HMD, a peripheral HMD, a tablet, and printed instructions. In their study they use a car maintenance task – checking a car's oil level and changing a light bulb. Another approach is demonstrated by the TeleAdvisor system [8], which uses a camera projector system to enable a remote helper to give instructions to an on-site learner. They use a task where the workers are connecting cables at a TV setup scenario. Furthermore, Gauglitz et al. [7] use AR instructions for remote collaboration in a Boeing 737 cockpit.

Previous research recognized the need for simplifying manufacturing tasks to compare instructions. Tang et al. [13] ran a study comparing printed manuals, computer assisted instructions, and AR-instructions using a HMD. They introduced an abstract pick-and-place task using Lego Duplo bricks. Both Sakata et al. [12] and Funk et al. [6] also use Lego Duplo bricks for giving instructions for assembly tasks. Sakata et al. [12] use a remote collaboration scenario whereas Funk et al. [6] provide projected in-situ feedback at a workplace for cognitively impaired workers. However, the works by Funk, Sakata, and Tang were using different assembly tasks.

A GENERAL ASSEMBLY TASK MODEL

All research projects mentioned above provide ways of giving instructions to workers. However, each presented approach uses its own assembly task. This makes the different instructions very difficult to compare against each other, as each new assembly task introduces a different complexity and different times to assembly assemble a product. Unfortunately, most research only reports the total time that workers need to assemble a product, which is only one *task-dependent* measure. This problem in comparing different instruction approaches with one another could be solved by finding *task-independent* measures or by standardizing experiments.

In order to find *task-independent* measures, we analyzed the tasks that were used in related work and constructed an equation which can be applied to each presented task. As related work identified the four relevant component actions for assembly tasks to be *reach, grasp, move position, and release* [3], we abstract the basic actions to high-level actions that require cognitive effort and are affected by an instruction. Thus, we suggest the General Assembly Task Model (see Figure 1), which calculates the total time a workers needs to assemble a product t_{total} by measuring four sub-times (t_{locate_part}, t_{pick}, t_{locate_pos}, and $t_{assemble_x}$).

$$t_{total} = n\left(t_{locate\_part} + t_{pick} + t_{locate\_pos} + t_{assemble_x}\right)$$

Figure 1. The equation for calculating the assembly time according to the General Assembly Task Model.

Where n is the number of assembly steps required. The time to locate the bin in which the next part is stored is t_{locate_part}. It includes both the cognitive time to process the feedback and the time to move the arm to the target bin. The latter can be treated as a constant value as workplaces are usually designed in a way that the distance to pick parts is within an arms reach [14]. The time that a worker needs to pick a part is t_{pick}. If the specific task does not include picking a part, t_{locate_part} and t_{pick} are 0. t_{locate_pos} is the time that the worker needs to locate the assembly position of the part that the worker is currently holding or the part that the worker needs to modify. Finally, $t_{assemble_x}$ is the time needed by the worker to perform the assembly task associated with a specific part x.

task-dependent measures: $t_{assemble_x}$ is *task-dependent* since different parts might take more or less time to assemble. It is also a measure for instruction quality, as instructions can communicate how to use tools during an assembly step, how to correctly assemble a part, and which details to focus on. t_{pick} is a *task-dependent* measure, as it depends on the size and weight of the part that needs to be grasped. However, we don't consider t_{pick} to be a measure for instruction quality since grasping is usually not instructed.

task-independent measures: We consider t_{locate_part} and t_{locate_pos} as indicators for the *task-independent* instruction quality, as they quantify the cognitive effort that the worker needs to transfer the given instructions to the workplace.

REQUIREMENTS FOR A STANDARDIZED TASK

In order to find a standardized assembly task for instruction giving, we analyzed tasks that were used in previous work and identified two categories of tasks. Some related approaches [1, 6, 12, 13] recognized Lego Duplo as abstraction for industrial pick-and-place tasks. The major benefit of such an abstract pick-and-place task is that the time that the worker needs is mainly t_{locate_part} and t_{locate_pos}, where $t_{assemble_x}$ is relatively low, as the brick is just placed at a position and no further assembly is required. On the other hand, related work uses specific industrial assembly tasks [4, 10, 15, 16]. Compared to a pick-and-place task, $t_{assemble_x}$ is much higher when using industrial assembly tasks. As the most time is used to perform the assembly itself, the t_{locate_part} and t_{locate_pos} only account for a small part of t_{total}. Overall, we recognize two different types of tasks, pick-and-place tasks and industrial assembly tasks. Thereby, each work step belonging to either type consists of one or many of the following three actions: picking items, placing items, and assembling them.

To sum up the requirements for a standardized task for comparing assembly instructions with one another, we define the following four criteria for designing a uniform assembly task:

- **cheap to setup:** the proposed task has to consist of off-the-shelf components that are affordable and easy to purchase.
- **easy to replicate:** the assembly tasks have to be described in a way that they are easily replicable.
- **representative:** the tasks have to cover the three main actions that can be found in assembly scenarios in the industry (i.e. picking parts, placing parts, and assembling parts).
- **easy to scale up:** the number of working steps have to be easily changeable without changing the nature of the task.

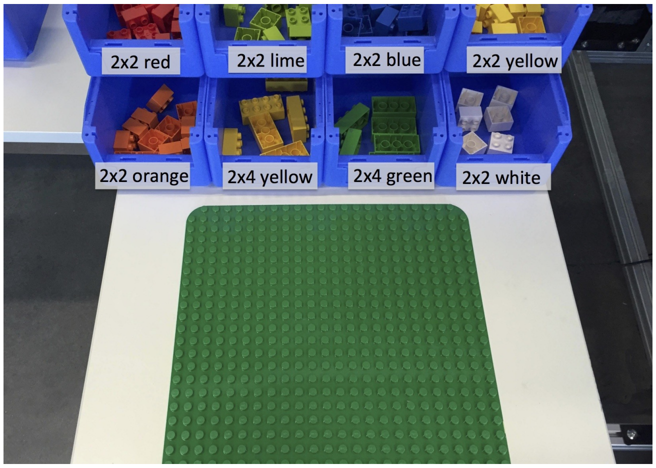

Figure 2. The apparatus of the duplo task: the bricks used for the task are stored in blue picking bins and are arranged as depicted. The plate is in front of the picking bins.

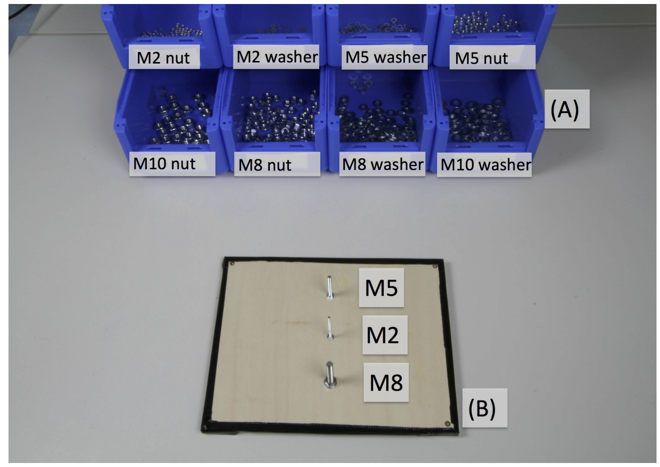

Figure 3. The apparatus of the assembly task. The assembly parts are stored in blue picking bins (A), the workpiece carrier (B) is placed in front of the boxes. The screws are aligned vertically that the worker can use both hands to for assembling the nuts and washers.

A UNIFORM EXPERIMENT DESIGN

Inspired by approaches from prior work, we propose a uniform experiment design as a benchmark for assembly instructions that fulfills the requirements mentioned above. We implement two assembly tasks which follow the proposed design: A pick-and-place task using Lego Duplo, and an industrial task that requires a precise assembly of components.

The benchmark follows a repeated measures design with the number of assembly steps as the only independent variable. We consider tasks with 4, 8, 16, and 32 assembly steps for both tasks. As dependent variables, we suggest measuring the assembly time according to the General Assembly Task Model (GATM) (t_{locate_part} ,t_{pick} ,t_{locate_pos}, $t_{assemble_x}$), the errors that were made during assembly, and the perceived cognitive load using the NASA-TLX [9] questionnaire. As the time and errors are dependent on the number of assembly steps, we normalize all times and errors by dividing them by the number of assembly steps.

In order to measure the exact assembly times, the experiment has to be recorded and evaluated in a post-analysis. The time t_{locate_part} is measured from the moment the instruction is shown until the hand of the worker is in the correct bin. Second, t_{pick} is the time the participants' hand is inside the bin. Next, t_{locate_pos} is the time from when the hand leaves the bin until the part arrives in the right location for the assembly task. Finally, $t_{assemble_x}$ is the time at the appropriate site to perform the assembly task. Further, the order of the number of assembly steps should be counterbalanced according to the Balanced Latin Square.

Duplo bricks

The first task is an abstract pick-and-place task using Lego Duplo bricks. For the task, we equip a workplace with a 26×26 green Lego Duplo plate (see Figure 2). The Duplo bricks that are used for the assembly are stored in 8 picking bins that are arranged in a 2×4 grid. The exact arrangement and content of the bins is depicted in Figure 2. We use 8 different Lego Duplo bricks (size 2×2: orange, yellow, blue, red, white, lime; size 4×2: yellow, and green).

Artificial industry task

In the industry, assembly pieces are usually manufactured on so called workpiece carriers, which hold the piece in a position that facilitates assembly. Because these workpiece carriers are product-dependent, we propose creating an artificial workpiece carrier from a wooden plate and screws.

We use a 30cm×24cm wooden plate. In the middle of the plate we drill holes to fit three types of metric screw threads[1]: M5, M2, and M8 . All holes have a distance of 6cm between them, resulting in an (x/y) position of M5(15/6), M2 (15,12), and M8 (15,18). The carrier is depicted in Figure 3 (B). The parts to assemble are stored in 8 picking bins that are arranged in a 2×4 grid. The position of the different parts is depicted in Figure 3. We use washers and nuts for each metric thread. Additionally, we use M10 nuts and washers as distracting elements.

Baseline: paper-based instructions

As an easy-to-reproduce baseline, we took pictures of each work step from the worker's perspective. In the Lego Duplo task, we displayed the brick to pick next in the upper left corner and highlighted the position of the brick to place using a red arrow. In the industrial assembly task, we highlighted the position of the nut or washer that needs to be picked next using a red rectangle. Further we highlighted the position to assemble the part using a red circle. The pictorial instructions can be downloaded by other researchers from our website[2].

Evaluation

To evaluate the GATM and provide baseline values for our suggested experiment, we conducted a user study with 12 participants (6 male, 6 female). The participants were in the age range from 22 to 31 (M=24.83, SD=3.15) and were recruited via a mailing list. Participants were undergraduate students with various majors. The study took approximately 30 minutes, including assembly tasks and filling out questionnaires.

[1]ISO 68-1:1998 - http://www.iso.org/iso/catalogue_detail.htm?csnumber=3707 (last access 08-15-15)

[2]Paper instruction download: http://www.hcilab.org/ar-instruction-benchmark

#steps	4	8	16	32
t_{locate_part}	1.84 (1.52)	2.22 (1.57)	2.19 (1.13)	2.29 (1.36)
t_{pick}	0.86 (0.43)	0.86 (0.39)	0.87 (0.42)	0.97 (0.63)
t_{locate_pos}	1.40 (1.07)	1.26 (1.05)	1.17 (0.48)	1.19 (0.75)
$t_{assemble\_x}$	1.30 (1.63)	0.95 (0.43)	0.96 (0.47)	0.96 (0.50)
# errors	0.17 (0.58)	0.33 (0.65)	0.33 (0.65)	0.5 (0.52)

Table 1. Average time in seconds for the task-independent and task-dependent measures for the Lego Duplo task and the average number of errors made. Standard deviation is depicted in parenthesis.

#steps	4	8	16	32
t_{locate_part}	2.28 (1.09)	2.34 (1.61)	2.14 (1.46)	2.15 (1.46)
t_{pick}	1.31 (0.58)	1.48 (0.81)	1.52 (0.99)	1.60 (1.09)
t_{locate_pos}	1.15 (0.63)	1.28 (0.61)	1.29 (0.91)	1.35 (0.85)
$t_{assemble\_x}$	6.23 (7.06)	6.16 (7.03)	5.83 (9.02)	6.01 (9.61)
# errors	0.08 (0.28)	0.16 (0.39)	0.08 (0.29)	0.92 (1.24)

Table 2. Average time in seconds for the task-independent and task-dependent measures for the industrial task, and the average number of errors made. Standard deviation is depicted in parenthesis.

We statistically compared the average time for $t_{assemble_x}$, t_{pick}, t_{locate_part}, and t_{locate_pos}, the number of errors, and the NASA-TLX between the two proposed tasks using a one-way repeated measures ANOVA. The average times are depicted in Table 1 for the Lego Duplo task, and in Table 2 for the industrial assembly task.

The mean time to locate a part t_{locate_part} over all step-sizes was 2.14s (SD=0.88s) for the Lego Duplo task and 2.17s (SD=0.47s) for the industrial assembly task. A repeated measures ANOVA did not reveal a significant difference ($p > 0.05$). Further, we compared t_{locate_part} within each task across the different step sizes. However, the test did not find a statistically significant difference between the two tasks (all $p > 0.05$).

The average time to pick a part t_{pick} over each step size was 0.89s (SD=0.20s) for the Lego Duplo task and 1.46s (SD=0.34s) for the industrial assembly task. A repeated measures ANOVA showed a significant difference for t_{pick} between the two tasks $F(1,11) = 33.22$, $p < .001$. Further, we compared t_{pick} within each task across the different step sizes. However, a repeated measures ANOVA did not find a significant difference for both tasks (all $p > 0.05$).

The average time to find the location of a picked part t_{locate_pos} over each step-sizes was 1.25s (SD=0.29s) for the Lego Duplo task and 1.25s (SD=0.35s) for the industrial assembly task. A repeated measures ANOVA did not reveal a significant difference ($p > 0.05$). We further compared t_{locate_pos} within each task across the different step sizes. For the Lego Duplo task, we did not find a significant difference between the different step sizes ($p > 0.05$). Additionally, we compared the different step-sizes within the tasks using a one-way repeated measures ANOVA. Mauchly's test showed that the sphericity assumption was violated ($\chi^2(5) = 12.927$, $p = .025$). Therefore, we used the Greenhouse-Geisser correction to adjust the degrees of freedom ($\varepsilon = .62$). Interestingly, we found a significant difference between the step sizes in the industrial assembly task $F(1.887, 20.753) = 3.952$, $p = 0.037$. Pairwise comparisons revealed a significant difference between 4 and 8 steps.

The average time to assemble a part $t_{assemble_x}$ over all step sizes was 1.04s (SD=0.29s) for the Lego Duplo task and 6.04s (SD=1.27s) for the industrial assembly task. A repeated measures ANOVA revealed a significant difference for $t_{assemble_x}$ between the two tasks $F(1,11) = 212.32$, $p < 0.001$. We compared $t_{assemble_x}$ within each task across the different step sizes. However, a repeated measures ANOVA did not find a significant difference for both tasks (all $p > 0.05$).

We also compare the number of errors made across the different tasks. The mean number of errors made across the four different levels of complexity was 1.33 (SD=1.67) for the Lego Duplo task, and 1.25 (SD=1.54) for the industrial assembly task. The analysis did not reveal a significant difference in the mean number of errors between the tasks ($p > 0.05$). Further, we compared the number of errors within each task across the different step sizes. The number of errors for each complexity is shown in Table 1 for the Lego Duplo task, and in Table 2 for the industrial assembly task. However, the ANOVA did not reveal a statistically significant difference (all $p > 0.05$).

Considering the perceived cognitive load according to the average NASA-TLX score [9] the Lego Duplo task reached a score of 33.16 (SD=15.52), whereas the industrial assembly task reached a score of 43.91 (SD=16.77). The analysis revealed a significant difference regarding the perceived cognitive load between the two tasks $F(1,11) = 13.05$, $p = 0.004$.

DISCUSSION

The results of the study show that both proposed *task-dependent* measures $t_{assemble_x}$ and t_{pick} are significantly different between the two tasks. Accordingly, the two *task-independent* measures t_{locate_part} and t_{locate_pos} were not significantly different between the tasks using paper-based instructions. The results support our proposed GATM.

Further, we found a significant difference in t_{locate_pos} between the 4 and 8 steps industrial assembly tasks. We believe that this difference occurred because participants worked at a faster pace when assembling the 4 steps scenario. One participant (P8) stated that *"[He] wanted to finish very quickly when seeing that the instruction only consists of a little number of steps."* For all other measures and all other step-sizes the analysis did not find a significant difference. This suggests that the two assembly tasks and the used dependent variables are step-size independent and that experiments consisting of 8 work steps might be sufficient.

CONCLUSION

In this paper, we introduce the General Assembly Task Model using *task-dependent* and *task-independent* measures and provide a uniform experiment design as a benchmark for evaluating assembly instructions. Further, we suggest two tasks that are cheap to build, easily reproducible, and covering most tasks that are found at assembly workplaces in industrial settings. We provide a paper-based baseline for the two proposed tasks that can be downloaded by other researchers. We believe that introducing a uniform benchmark will make different instruction approaches more comparable to one another.

ACKNOWLEDGMENTS

This work is funded by the German Federal Ministry for Economic Affairs and Energy in the project motionEAP, grant number 01MT12021E and the NSF under the grant number IIA-1358096.

REFERENCES

1. A Bannat, F Wallhoff, G Rigoll, F Friesdorf, H Bubb, S Stork, HJ Müller, A Schubö, M Wiesbeck, and MF Zäh. 2008. Towards optimal worker assistance: a framework for adaptive selection and presentation of assembly instructions. In *Proceedings of the 1st international workshop on cognition for technical systems, Cotesys*. DOI:http://dx.doi.org/10.1.1.364.3359

2. Martin Barrett, Jonathan Blackledge, and Eugene Coyle. 2011. Using virtual reality to enhance electrical safety and design in the built environment. (2011).

3. Rainer Bokranz and Kurt Landau. 2012. Handbuch Industrial Engineering. *Schäffer-Poeschel, Stuttgart* (2012).

4. Thomas P Caudell and David W Mizell. 1992. Augmented reality: An application of heads-up display technology to manual manufacturing processes. In *System Sciences, 1992. Proceedings of the Twenty-Fifth Hawaii International Conference on*, Vol. 2. IEEE, 659–669. DOI:http://dx.doi.org/10.1109/HICSS.1992.183317

5. Markus Funk, Andreas Bächler, Liane Bächler, Oliver Korn, Christoph Krieger, Thomas Heidenreich, and Albrecht Schmidt. 2015a. Comparing Projected In-Situ Feedback at the Manual Assembly Workplace with Impaired Workers. In *Proceedings of the 8th International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, New York, NY, USA. DOI:http://dx.doi.org/10.1145/2769493.2769496

6. Markus Funk, Sven Mayer, and Albrecht Schmidt. 2015b. Using In-Situ Projection to Support Cognitively Impaired Workers at the Workplace. In *Proceedings of the 17th international ACM SIGACCESS conference on Computers & accessibility*. ACM. DOI: http://dx.doi.org/10.1145/2700648.2809853

7. Steffen Gauglitz, Cha Lee, Matthew Turk, and Tobias Höllerer. 2012. Integrating the physical environment into mobile remote collaboration. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*. ACM, 241–250. DOI: http://dx.doi.org/10.1145/2371574.2371610

8. Pavel Gurevich, Joel Lanir, Benjamin Cohen, and Ran Stone. 2012. TeleAdvisor: a versatile augmented reality tool for remote assistance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 619–622. DOI: http://dx.doi.org/10.1145/2207676.2207763

9. Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183. DOI: http://dx.doi.org/10.1016/S0166-4115(08)62386-9

10. Steven J Henderson and Steven Feiner. 2009. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on*. IEEE, 135–144. DOI: http://dx.doi.org/10.1109/ISMAR.2009.5336486

11. I Scott MacKenzie and R William Soukoreff. 2003. Phrase sets for evaluating text entry techniques. In *CHI'03 extended abstracts on Human factors in computing systems*. ACM, 754–755. DOI: http://dx.doi.org/10.1145/765891.765971

12. Nobuchika Sakata, Takeshi Kurata, and Hideaki Kuzuoka. 2006. Visual assist with a laser pointer and wearable display for remote collaboration. (2006). DOI: http://dx.doi.org/10.1.1.123.9324

13. Arthur Tang, Charles Owen, Frank Biocca, and Weimin Mou. 2003. Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 73–80. DOI: http://dx.doi.org/10.1145/642611.642626

14. Hans-Peter Wiendahl, Jürgen Reichardt, and Peter Nyhuis. 2015. *Handbook Factory Planning and Design*. (pp. 150–153) Springer. DOI: http://dx.doi.org/10.1007/978-3-662-46391-8

15. Xianjun Sam Zheng, Cedric Foucault, Patrik Matos da Silva, Siddharth Dasari, Tao Yang, and Stuart Goose. 2015. Eye-wearable technology for machine maintenance: Effects of display position and hands-free operation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2125–2134. DOI:http://dx.doi.org/10.1145/2702123.2702305

16. Jianlong Zhou, Ivan Lee, Bruce Thomas, Roland Menassa, Anthony Farrant, and Andrew Sansome. 2011. Applying spatial augmented reality to facilitate in-situ support for automotive spot welding inspection. In *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry*. ACM, 195–200. DOI: http://dx.doi.org/10.1145/2087756.2087784