Investigating the Impact of Real-World Environments on the Perception of 2D Visualizations in Augmented Reality

Marc Satkowski
Interactive Media Lab Dresden
Technische Universität Dresden
Dresden, Germany
msatkowski@acm.org

Raimund Dachselt*†
Interactive Media Lab Dresden
Technische Universität Dresden
Dresden, Germany
dachselt@acm.org

Figure 1: Different aspects of the design of our user studies: (A) and (B) show a subject sitting in front of two background configurations. (B) illustrates subjects solving a task on the shown visualization in front of background *BG4*, while (C) depicts a question on *BG3*. (D) shows three additional displays (indicated by purple squares) with different signal colors and numbers that simulate a secondary observation task for the second study.

ABSTRACT

In this work we report on two comprehensive user studies investigating the perception of Augmented Reality (AR) visualizations influenced by real-world backgrounds. Since AR is an emerging technology, it is important to also consider productive use cases, which is why we chose an exemplary and challenging industry 4.0 environment. Our basic perceptual research focuses on both the visual complexity of backgrounds as well as the influence of a secondary task. In contrast to our expectation, data of our 34 study participants indicate that the background has far less influence on the perception of AR visualizations. Moreover, we observed a mismatch between measured and subjectively reported performance. We discuss the importance of the background and recommendations for visual real-world augmentations. Overall, our results suggest that AR can be used in many visually challenging environments without losing the ability to productively work with the visualizations shown.

KEYWORDS

User Study, Augmented Reality, Visual Perception, Immersive Analytics, In-Situ Visualization, AR Visualization, Industrial Scenario, Industry 4.0

ACM Reference Format:
Marc Satkowski and Raimund Dachselt. 2021. Investigating the Impact of Real-World Environments on the Perception of 2D Visualizations in Augmented Reality. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3411764.3445330

1 INTRODUCTION

Augmented Reality (AR) is of rising interest due to significant technological advances and commercial products being increasingly available. In the future, the use of AR can be expected to pervade many different areas of our daily life. As this technology gets more common in everyday life, the importance to understand how information displayed in AR should be presented increases. In the past, AR research often focused on displaying additional 3D objects or simple visual elements like text [48]. In comparison, we can observe a rising interest in the effective integration of more complex and even abstract information into real-world scenes. This is exemplified by trends such as situated, embedded [16, 48, 56], and immersive analytics [37] as new areas of data visualization.

*Also with, Centre for Tactile Internet with Human-in-the-Loop (CeTI), Technische Universität Dresden.
†Also with, Cluster of Excellence Physics of Life, Technische Universität Dresden.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445330

The immersive visualization research community is quite active with regard to these topics [20, 27]. However, we are still faced with many open questions [5, 27] to answer, among them how humans perceive information presented in AR [17, 30]. To contribute to answering this question, we focused on how real-world environments, which are always present in AR, might influence this very perception. This understanding can further help us to make informed decisions on how and where information can be presented in AR. This is especially necessary if we want to elevate AR technology to a tool that cannot only be used in casual everyday situations, but also in productive scenarios, like modern production plants. For such a tool, it is essential to present information not only with simple visual representations (e.g., text), but also in a more advanced and complex presentation style as it is common with data visualizations.

Information visualizations (e.g., line charts or scatterplots) show abstract and often complex data by means of combining different rather simple visual elements (e.g., text, lines, bars). The perception of such elements has already been studied, and design recommendations were presented as well [19, 24]. However, to the best of our knowledge, a deeper understanding on how more complex visualizations are perceived in AR is lacking. This gap is especially crucial to be closed for immersive analytics [37] applications that use AR to generate insights often directly linked to real-world objects. However, other application domains would also benefit from findings in this research area. Traditional use cases for AR range from instructions, over manuals to training (an overview can be found in [53]). Tourism [28], education [57], and even grocery shopping [1, 7] are further possible scenarios. All those differ regarding their real-world environments, including the background, noise, or lighting. Some use cases are more demanding and challenging than others. While use cases like tourism often focus on a more casual interaction, AR application for industrial scenarios have the clear goal to use AR as a productive tool. When operators have to observe complex data (e.g., trend charts) augmented to several machines, they are also faced with complex visual backgrounds. Inspired by the wish for productive tools in the challenging industrial context of AR visualizations, but not limited to this domain, we want to better understand the influence of the background scene in the perception of AR visualizations.

In this paper we present two user studies with a total of 34 participants conducted in an experimental production plant (see Fig. 1). Our first study solely focused on the influence of the visual background. However, this does not fully reflect a real-world use case, since operators often have to interact not only with the presented AR content, but also with the machines in front of them. For this reason, we added a secondary observation task coupled to the background for the second study to simulate such a scenario. Most of the current research that investigates AR takes place in strictly controlled laboratory environments and often uses 2D pictures of 3D scenes as visual backgrounds (as seen in [34, 35]). Instead, we focused on one real-world scenario for our studies: industrial production plants. However, we took care to enable the generalization of our findings to other settings and domains. The results of our studies show an unexpected result: real-world backgrounds have far less influence on the measured performance than expected. Having said that, the subjective reports of our participants reveal that the perceived influence differs from the measured one. We understand our work as part of a research agenda to make Augmented Reality – and AR visualizations in particular – truly usable as a productive tool in real-world contexts.

In summary, with this paper we contribute:

- Two user studies conducted in a real-world experimental production plant. Both use industrial 3D scenes as backgrounds.[1]
- Insights on how strong the influence of the visual background on AR augmentation is. This includes both the subjectively perceived and objectively measured influence.
- Findings about how the influence of the visual background is altered with the introduction of a secondary observation task.
- A discussion of the results and recommendations for visual real-world augmentations.

2 BACKGROUND AND RELATED WORK

Our studies focus on the perception of information visualization in Augmented Reality (AR). Subsequently, our work relates to three main topics: perception in AR, AR visualizations, and AR in industrial scenarios. The last topic was chosen to reflect on an exemplary usage of AR visualizations as a productive tool.

2.1 Perception in Augmented Reality

AR technology enables the enrichment of the real world with additional digital information. This can be done either by overlaying visual elements in the field of view of the user (optical see-through) or by embedding the new information in a video stream that will be shown to the user (video see-through). Even though the AR research community is quite active [27], there are still open question to be answered as proposed by Kruijff et al. [30], Billinghurst et al. [5], Kim et al. [27]. Those surveys in addition to Erickson et al.'s review [17] show that only a few research projects focus on information visualizations and their perception in AR. The existing research papers mostly investigated fundamental topics, such as depth perception and handling of occlusion [22], color perception [33], and automatic color correction [12]. Lu et al. [34, 35] had a deeper look on how attention can be controlled through subtle cues in AR scenes. In general, perceptual experiments mostly investigated those effects on *pictures* of 3D scenes. Lu et al. [34] recommended conducting perceptual experiments with optical-see-through devices and therefore in real-world scenes. Further, they used Feature Congestion (FC) [46] as a measurement of how cluttered a visual background is. This value is based on color, luminance, and orientation features of the image. We also use this value as a guideline for selecting appropriate background scenes.

A more applied view on perception in AR can be seen in different research projects with focus on basic visual elements, like text. Thereby the style of text and background [14, 19, 23–25, 29] or the placement of text labels [4, 36, 43] were often investigated. There is also work attempting to automatically determine the legibility of

[1] We provide most of the study data as supplemental material, including the questionnaire data, task descriptions, logged study data, and analysis scripts on our project page https://imld.de/ar-vis-perception.

text in correlation to the background it is placed on [31]. In general, those research works show that white text on blue background is the best choice, but it is not clear how the findings can be transferred to different and more complex visual objects like information visualizations. Further, the placement of labels only focused on connections between the real-world environment and the digital data. Information visualizations often contain a combination of text and additional visual elements, and use visual variables like color as a property to represent or highlight data.

One exemplary paper that investigated such visualizations can be seen in the work of Büschel et al. [8]. In their work, they focused on edge visualization in graphs and how different edge types perform in a graph analysis task in AR. They did not investigate how the visual background could influence their findings. However, they asked the participants if they were influenced by the background, which was answered with no. Another recent research project from Whitlock et al. [55] investigates how data visualizations are interpreted by the users. They compare different display types including a head-mounted AR display (video-see-through). The participants in their study had to solve different tasks for 2D and 3D scatterplots and bar charts. The results show that navigation in AR performed better than other display types, while color was harder to distinguish due to the influence of the color of real-world objects.

2.2 Information Visualization in AR

AR can make digital information immersive by introducing it directly into real-world environments. This capability is researched by situated, embedded [16, 48, 56], and immersive analytics [37, 50]. Situated visualization "is a data representation whose physical presentation is located close to the data physical referent" [56], while embedded visualization directly places the information as an overlay on the corresponding physical referent [56]. Marriott et al. [37] describe immersive analytics as "the use of engaging, embodied analysis tools to support data understanding and decision making". Visualizations, like line charts or network graphs, can be one component of such analysis tools. A few examples exist that extend a display-based analysis software with augmented reality content [9, 45, 54]. The former investigate analysis in a collaborative scenario, while Wang et al. [54] try to understand the general concepts of hybrid analysis setups. To truly be able to use and to optimally design such visualizations in AR, it is necessary to understand how the environment can influence perception. Therefore, we designed our studies to contribute to the knowledge about this influence.

To easily create visualizations for AR devices like the Microsoft HoloLens, several frameworks were introduced in the past years. All are based on the Unity 3D engine. While ImAxes [11] only creates axes that can be linked with one another, DXR [49] and IATK [10] allow the generation of axis-based visualizations with different visual marks for the data. u2Vis [45] presents a more generic framework which can also be used for and in combination with large displays and desktop environments. Lastly, CollARVis [6] presents a visualization framework for collaborative scenarios. To our knowledge, these frameworks were not evaluated with regard to their perception.

2.3 AR for Industrial Use Cases

AR head-mounted displays can be used in various application areas due to their form factors. They enable the user to free both hands for the interaction with the real-world environment, show customized representations, and create a greater immersion due to visualizations placed inside the real-world. Those properties seem to be especially useful for scenarios, where such applications should be used as a productive tool in industrial scenarios [13, 21, 39]. Several research projects focus on the use of AR in such use cases and investigate text legibility [19, 24], the recognition of industrial machines and modules in front of AR applications [2, 51], or guidelines and challenges for the user interface design [42]. Paelke et al. [42] see a greater need for user-friendly applications (similar to web and mobile apps) and assistance for a wide range of users. Additionally, they present a set of interaction and visualization techniques that are needed in such systems, including process-oriented and context-sensitive information representations. Other research focused on more specific use cases like maintenance [39], assembly [53] or CNC milling [32]. All the current research papers also show that a wide variety of dynamic parameters, application areas, users, and specialized use cases make industrial-scenarios quite challenging. Lastly, AR tools in such use cases have the overall goal to create a productive working environment. For this, Cardoso et al.'s survey [13] show that AR reduces execution time and improves the overall quality besides other factors.

3 STUDY 1: INFLUENCE OF BACKGROUNDS

AR applications in an industrial scenario will always exhibit various real-world backgrounds. It is necessary to investigate the influence of those backgrounds to allow AR visualizations and application to reach and even surpass currently used productive tools in performance and user-friendliness. However, we believe that the understanding of this influence can also be transferred to different use cases. Every AR visualization is and will be placed in different real-world environments and therefore has to be fitted to the respective backgrounds. Additionally, the concept of visual perceptual load, which describes "that perception has a limited capacity, which automatically proceeds until exhausted" [40], could influence the visual primary tasks presented in AR since "task-irrelevant stimuli are still processed to an extent that enables them to affect performance in a primary task" [44]. With this goal in mind, we created the first study. In order to investigate if the background has an influence on the measured and perceived performance, our subjects analyze visualizations with varying numbers of data points shown on different background scenes.

3.1 Design & Hypotheses

Since we aimed to answer the question wether the background in AR applications has an influence on the perception of information visualizations, we focused on two independent variables. Those are the *background* (*BG1*, *BG2*, and *BG3* as seen in Fig. 2) and the visualization *complexity* (presented by the number of data points in each visualization: 40, 50, 60a, 60b, and 70). The different *background* configurations, which were motivated by the wish to better understand the influence of dynamic and real-world backgrounds

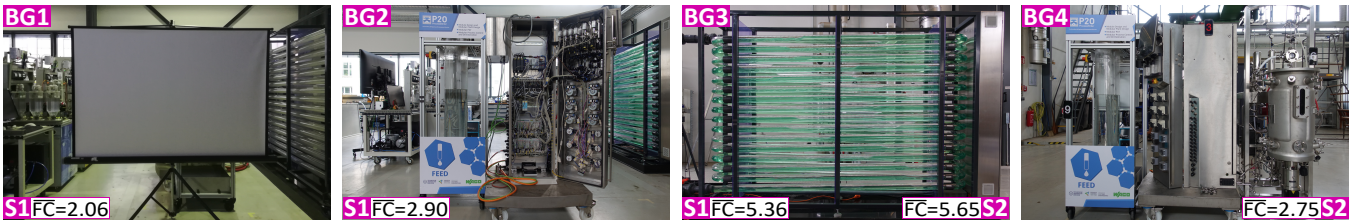

Figure 2: All background configurations used in both studies. *BG1*, *BG2* and *BG3* were used in the first study. *BG3* and *BG4* were used in the second study. *BG4* also shows the placement of the additional displays. The placement is similar to *BG3* (see Fig. 1d). The displayed FC values are the mean of all images taken from the HoloLens v1 in each session for each study (S1 for Study 1, S2 for Study 2).

[17, 27, 34], were used as a between-subject variable. The visualization *complexity*, motivated by the visual complexity which can increase the overall visual load [38, 41], was altered within each session. We measured the following dependent variables: *task completion time*, *error* values (as *absolute error* and *percentage error*), and questionnaire data. Lastly, we distinguished between different *analysis tasks* based on the low-level analysis tasks presented by Amar et al. [3].
We generated four hypotheses to guide our research:

- **H1** Solving tasks on *backgrounds* with elements that are more complex (by means of visual clutter) performs worse. A more distracting *background* will interfere with the overall perception of the shown visualizations.
- **H2** Solving tasks on visualizations with higher *complexity* perform worse. With more data points to analyze, the subjects will take longer and eventually give more incorrect answers.
- **H3** The negative effect of the *background* will be more visible while the subjects solve tasks on visualizations with higher *complexity* due to the overall increase of visual complexity.
- **H4** Subjects perceive the *backgrounds* with more clutter as more distracting.

3.2 Participants

We recruited 21 unpaid participants per word-of-mouth for our study. We had to exclude three runs due to headaches of one subject and problems with the state of one of the background modules for the other two. The remaining 18 participants (11 female, 7 male) were students from media informatics, media research, and computer science of our local university. The average age was 21 years ($M = 20.89$ years, $SD = 2.11$ years) and the self-reported height ranged from 162 cm to 198 cm ($M = 173.05$ cm, $SD = 9.38$ cm). No specific knowledge was required to participate in this study. All participants had normal or corrected-to-normal vision and had no color vision defect or spatial perception difficulties. On a five-step scale, all participants had less experience with AR in general ($M = 2.17, SD = 0.92$), no experience with AR via head-mounted displays ($M = 1.17, SD = 0.38$), no experience with Virtual Reality ($M = 1.61, SD = 0.70$) and some experience with the use of visualizations ($M = 2.72, SD = 0.89$) in general.

3.3 Setup

The study was conducted in an experimental production hall with a real-life modular chemical plant inside. We created our backgrounds (see Fig. 2) based on the Feature Congestion (FC) value [46] and some visual characteristics. The FC is a single value that is computed through a combination of the color, luminance, and orientation map of a given image (for this, we used the Piranhas Toolkit [15]). The smaller the value, the less clutter is present in the image. In general, we not solely relied on the FC value to define our backgrounds, but we also used our human judgment based on different characteristics, like motion, uniformity, or overall color. After we tested several background configurations in the industrial production hall, we chose two sufficiently different backgrounds (*BG2*, *BG3*), while the third (*BG1*) was added as a baseline. *BG3* shows moving green water, additional LED stripes, and an overall uniform design, with slightly angled pipes. On the other hand, *BG2* has two different sections. The left side is quite uniform, while the right side is highly cluttered with cables. The FC values for the *backgrounds*, calculated as an average out of all images taken in the session for each participant, are: $\overline{FC(BG1)} = 2.06$, $\overline{FC(BG2)} = 2.90$, and $\overline{FC(BG3)} = 5.36$ (see Fig. 2). As typical for many real-world scenarios (outside a clean, controlled laboratory) the noise of the environment and machines, the temperature in the production hall, the lighting, and the presence of other people were hard to control. To minimize disturbances in our study sessions, we took some precautions like informing the staff and blocking the specific area in the hall. However, the module used in *BG3* created a constant noise and some people worked occasionally at the same time the experiments were conducted.

For our study, we built two different applications: an AR client application used by the subjects and a server application controlled by the investigator. The AR application was developed for the Microsoft HoloLens v1, which has good image quality but a rather small field of view of approximately 30 degrees diagonally. The client application was implemented using the Unity 3D engine and the IATK framework [10], which we modified to better fit to our needs. To interact with the application, the participants used a Microsoft Clicker. We chose the Clicker to minimize gesture recognition problems, fatigue of the arm and the learning procedure for the subjects. While the participants were seated at a distance of approximately 2.5 m in front of the backgrounds (see Fig. 1a), the application showed questions (see Fig. 1c) and the visualization

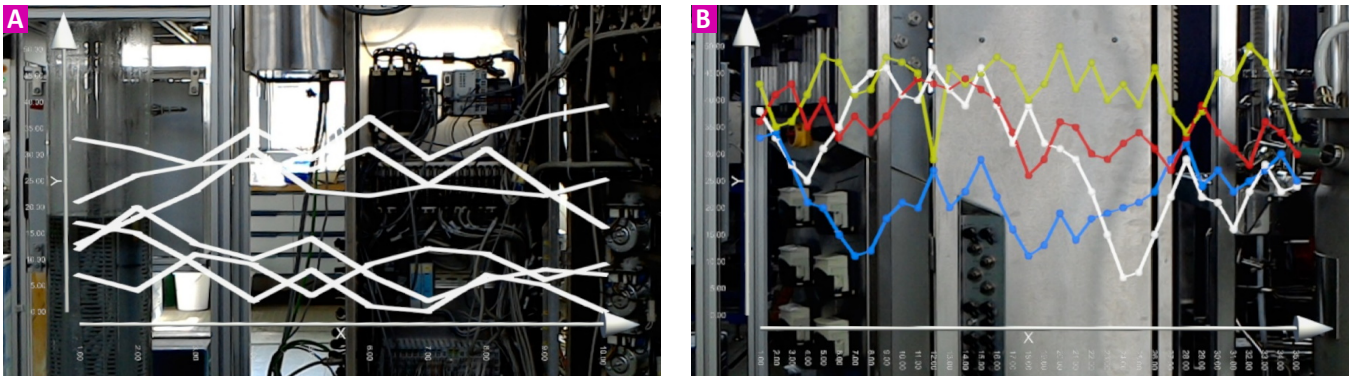

Figure 3: Line charts in front of the background from both studies. (A) was used in the first study (*complexity* of 60 data points with 10 x-values and 6 lines), while (B) was used in the second one (with a static amount of 35 x-values and 4 lines). Further, (A) shows *BG2* and (B) *BG4*.

(see Fig. 1b) at a height of 1.15 m above the floor. Additionally, the visualizations almost filled the field of view of the HoloLens entirely. The server application was implemented in C# with WPF and helped the investigator to control the experiments. This application allows monitoring the current state of the AR application and the subjects' interaction. Further, it logged all interactions and information the participants generated in the experiment.

3.4 Procedure

Each session consisted of the following phases: (1) A short introduction to the production hall and to modular industrial factories; (2) A questionnaire and a declaration of consent; (3) An explanation of the study tasks and interaction vocabulary; (4) The calibration of the HoloLens to the subject and a short training (four tasks); (5) The conduction of the experiment; (6) Final questionnaire regarding the experiment and the perceived influence of the background. One investigator who was in the production hall but was seated outside the field of view of the participants led the experiment. He only interacted with the participants during the experiment if they had any questions, wanted to answer an *analysis task*, or some technical problem occurred (e.g., with the input device).

Each click with the Clicker advanced the session to a different state. A hold of 1 s allowed the subject to orally give the answer to the previously shown *analysis task*. We chose an oral answer method to reduce the number of needed interaction devices and vocabulary. The session advanced to the next *analysis task* or question block after the investigator entered the given answer. Altogether, each session lasted approximately 51 min (M = 51.38 min, SD = 8.03 min), of which 25 min (M = 25.07 min, SD = 5.40 min) were needed for phase (5) of the study.

3.5 Tasks

The participants of our study had to solve *analysis tasks* on line charts (see Fig. 3a). We chose 2D line charts since they are widely established (as considered a basic visualization by Saket et al. [47]) and often used as trend charts in industrial factories. This decision allowed us to reduce the complexity of our experiment. The mentioned *analysis tasks* the subjects had to solve were based on Amar et al.'s [3] low-level analysis tasks. They describe that those tasks "largely capture people's activities while employing information visualization tools for understanding data". We chose a subset of six primitive tasks to generate a more natural set of questions. Those are: *Retrieve Value*, *Filter*, *Compute Derived Value*, *Find Extremum*, *Determine Range*, and *Find Anomalies*.

To generate a set of questions, we created a total of four block types. Of those, three types contained four, while the last type contained only three *analysis tasks*. The questions in one block built on the answers of the preceding *analysis task* to allow a more complex analysis. Each block was repeated five times, based on the *complexity*. In each repetition, we altered small details, like the highest or smallest value for a *Find Extremum* task (see Tab. 1). Each line chart was created based on one *complexity* level that defines the number of shown data points in steps of 10, altered through the number of data points per line and the number of lines. We chose this approach since the "Visual complexity is mainly represented by the perceptual dimensions of quantity of objects [and] clutter" [38, 41]. The y-value range remained constant between 0 and 50 for each *complexity*. In total, we created five *complexity* levels: 40 (5 lines with 8 x-values); 50 (5 lines with 10 x-values); 60a (5 lines with 12 x-values); 60b (6 lines with 10 x-values); 70 (7 lines with 10 x-values). The data used in those charts were generated through a Python script. Altogether, each participant had to solve 75 *analysis tasks* (5 repetitions, each with 15 tasks).

We minimized the possible bias through training effects by counterbalancing the order in which each participant had to solve all 20 blocks (5 *complexity* x 4 block types). For this, we used a latin-square for both factors separately. Therefore, each participant had to solve one block type with the five possible *complexities* before the next block type was presented. Lastly, each participant was randomly assigned to only one of the three *backgrounds*. In total, each value of the independent variable *background* was tested by six different subjects.

3.6 Measurements & Derived Data

As part of our study applications, we logged timestamps (e.g., start of a block, toggle between visualization and question), different

Low-Level-Analysis Task	Number of Questions Study 1	Study 2	Example Question
Retrieve Value	4	3	What is the Y-value at the given X-value?
Filter	3	1	How many lines have a Y-value **higher** than **45**?
Compute Derived Value	3	2	What is the average Y-value for all Y-values of those lines?
Find Extremum	3	1	At what X-value is the **highest** Y-value of the selected lines?
Sort		1	What is the descending order of all lines according to their **highest** Y-value?
Determine Range	1	1	What is the Y-value range for those lines?
Find Anomalies	1		At what X-value lies an outlier point?
Cluster		1	At what X-value is the distance of all lines to each other the **smallest**?
Compare		2	Does the mentioned Y-value lie **below** the average Y-value of all lines?

Table 1: All used *analysis tasks* (based on the work of Amar et al.[3]) in our study. The numbers in the middle columns show how many of the questions in each repetition were mapped to those *analysis tasks*. Lastly, the bold words in the example questions were altered between different repetitions (e.g., higher to smaller).

events, as well as the given answers of each participant. With the timestamps, we could calculate the time spent completing one *analysis task* (in the following mentioned as *task completion time*). This time was measured from first seeing the visualization (after a new question) to the time the answer of the subject was entered into the system. On the other hand, the answers allowed us to calculate the *error*. Here we differ between *absolute error*, for tasks with numerical outcome (e.g., *Retrieve Value*), and *percentage error*, for answers that could only be correct or not (e.g., *Filter*). In the following, we will refer to both *errors* and the *task completion time* as performance. Each session was also video recorded while the investigator observed the participants and took notes.

The questionnaires contained a total of 40 questions. 13 questions focused on demographic data and 19 were closed questions based on a rating scale. The last eight were open questions. In general, we were guided in the creation of the post-study questionnaire by three main questions we liked to be answered: (1) Does the task solving process have an influence on the physical state of the user? (2) How is the background perceived by the user? (3) How were the visualizations and their associated tasks perceived? The closed questions are about the physical state (fatigue, concentration, motivation, headache, dry and irritated eyes) before and after the study, the overall recognition of the lines and axes, as well as the perception of the visualization and the background (see Fig. 5). A NASA TLX was also performed and is assigned to this category. Among the open questions were: "Which particular areas of the background caused problems to you?", "To what extent did you notice the background when solving the tasks?", and "Were there any other influences that distracted you in the study?".

3.7 Data Analysis

Before we started to analyze our data, we checked for outliers and replaced those with the following formula: $M + 2 * SD$ (as proposed by Field et al. [18]). A total of 7.5 % data points for *absolute error* and 2.7 % for *task completion time* were replaced. Afterwards, our preliminary test show that our data is not normally distributed (Shapiro-Wilk). Additionally, we checked the equality of variance (Levene) for the *backgrounds* while we calculated the sphericity (Mauchly) on the *complexities*. All tests showed that the data has no violation on the equality and one violation on the sphericity for *complexity* on *absolute error*. Following, we used one-way ANOVAs on the *background* (grouped over all questions per block), one-way repeated measurements ANOVAs for the *complexity* with a Greenhouse-Geisser correction for the violation of sphericity, and two-way mixed ANOVAs for the interaction of *background* and *complexity*. Further, we used Kruskal-Wallis H test for the questionnaire answers on a rating scale. In general, if we found any significance, we calculated either pairwise t-tests or Mann-Whitney U test with Benjamini/Hochberg FDR correction.

3.8 Results

In this section, we will present the results grouped by the generated hypotheses, followed by additional findings concerning subjective experiences. The results of our study is depicted in several diagrams of our measured data (see Fig. 4) and questionnaire data (see Fig. 5 and Fig. 6).

H1 (clutter on performance) The statistic tests reveal no significance influence on the performance, as seen by *absolute error* ($F(2, 50) = 0.281, p = 0.7561, \eta_p^2 = 0.011$), *percentage error* ($F(2, 51) = 0.086, p = 0.9175, \eta_p^2 = 0.003$), and *task completion time* ($F(2, 104) = 2.408, p = 0.0950, \eta_p^2 = 0.044$). Additionally, Fig. 4(a) shows that tasks solved in our baseline condition *BG1* (lowest FC value) have the fewest *absolute error*. Overall, solved tasks on *BG3* (dynamic background with highest FC value) were the fastest, while the second fastest condition varies for different *analysis tasks*. In summary, *no significant influence of background clutter on the overall performance was found.*

H2 (complexity on performance) The statistic tests show a significance effect on the performance, as seen by *absolute error* ($F(4, 68) = 3.901, p < 0.05, \eta_p^2 = 0.187$) and *percentage error* ($F(4, 68) = 6.362, p < 0.001, \eta_p^2 = 0.272$), while *task completion time* ($F(4, 68) = 1.144, p = 0.3426, \eta_p^2 = 0.063$) did not. The t-tests for *absolute error*, visualized in Fig. 4(b), show that both *complexity* with 60 data points have the least *errors*. For *percentage errors* the t-tests present different significant combinations (see Fig. 4(b)). In summary, *our data displays that the complexity has an influence on the error but not on the task completion time.*

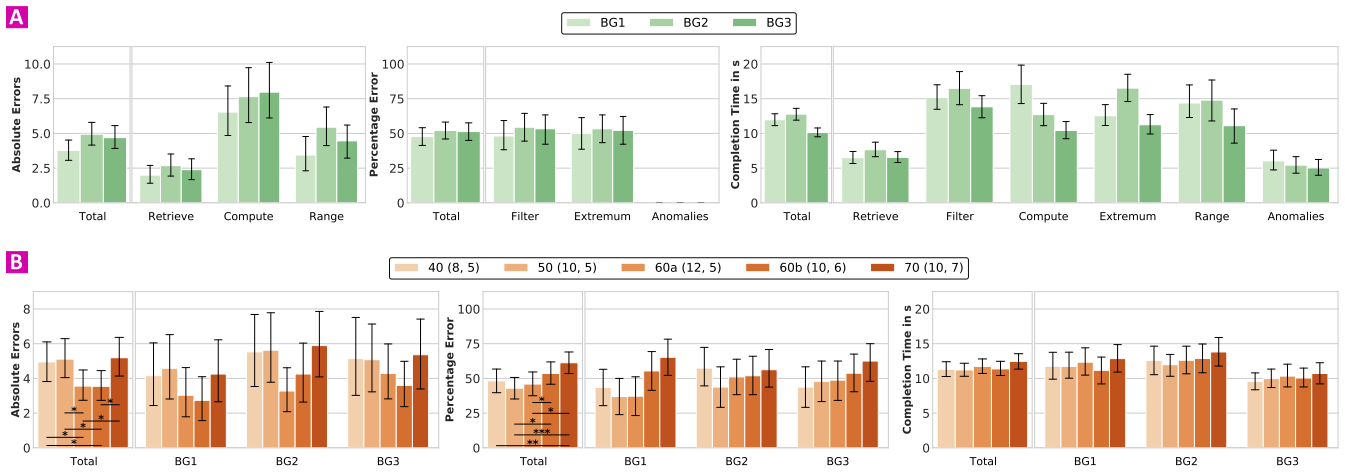

Figure 4: Different data analysis visualizations for the first study. (A) presents the *absolute error*, *percentage error* and *task completion time* grouped by each *analysis task* and *background*. (B) shows the same dependent variables for the different *complexities* and *backgrounds*. Both display the mean value and a 95% confidence interval with the whiskers. (*: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$ for pairwise t-tests)**

H3 (complexity & clutter on performance) The statistic tests demonstrate no significant interaction between *background* and *complexity* on *absolute error* ($F(8, 60) = 0.398, p = 0.9171, \eta_p^2 = 0.05$), *percentage error* ($F(8, 60) = 1.518, p = 0.1702, \eta_p^2 = 0.168$), and *task completion time* ($F(8, 60) = 0.202, p = 0.9895, \eta_p^2 = 0.026$). Fig. 4(b) shows again that *BG1* has the least amount of *errors*. In summary, *the combination of background and complexity shows no significant interaction and therefore no significant effect.*

H4 (clutter on perceived distraction) The questions focused on the perceived distraction and performance (see Fig. 5, PQ = perception questions) show that the *background* has a significant effect on the perceived influence (*PQ2* and *PQ3*) and how well the visualizations could be read (*PQ5*, *PQ6*, *PQ7*). All these questions signify that *BG1* is the *background* with the least influence and distraction. The additional open question "*To what extend did you notice the background during the tasks?*" also supports this. Subjects on *BG1* mentioned that the *background* was not noticeable (5 out of 6), while *BG2* was perceived as partly noticeable (3 out of 6). The participants on *BG3* rated the *background* as quite difficult (6 out of 6). The open question "*What areas of the visual background caused the most difficulties?*" shows that the reflection, the lighting and the cables (1 out of 6 for each) of *BG2* were recognized. For *BG3* the light stripes (2 out of 6) and the movement (2 out of 6) made it hard to read the visualizations correctly (2 out of 6). This also caused the visual merging of lines and the *background* (2 out of 6). We can conclude *that participants perceive the background conditions with a higher FC value as more distracting.*

Further findings Additionally, we analyzed the NASA TLX and the change of the physical state over the course of the study. The first showed no significance, while only the question regarding dry and irritated eyes showed that the background has a significant influence ($H(2) = 8.675, p < 0.05$) with a possible effect between *BG1* and *BG3* ($U = 7.5, p = 0.0814$) and a significant effect between *BG1* and *BG2* ($U = 2.5, p < 0.05$). Lastly, a few subjects gave negative comments on the field of view of the HoloLens (4 out of 18), while some perceived the constant noise produced by *BG3* as distracting (5 out of 18).

3.9 Discussion

Overall, our results show that the *background* has no influence on the measured performance (**H1**). However, we can see interesting differences in each measured value separately. *BG1*, the *background* with the least clutter, shows the best results in *absolute error* and *percentage error*, while the other two backgrounds are quite close to each other. Interestingly, the background with the highest clutter, *BG3*, has the fastest completion time. One explanation could be that the subjects did not invest the same effort on solving the tasks on this *background* since they perceived the distraction and their own performance on this background as rather bad. Also 3 out of 6 subjects reported slight dizziness on *BG3*. This could also explain why they wanted to finish the experiment on this *background* quickly. The performance values for the different *complexities* also reveal an interesting result. While the *task completion time* increases steadily, but not significantly, the *errors* show a significant effect for the *complexities* (**H2** and **H3**). Especially the *complexities* with 60 data points have the lowest number of errors. However, we cannot explain why this number of data points performed better than the conditions with less data points (40 and 50). In general, the influence seen through the measured values differ from the perceived influence of the participants (**H4**). The subjects found it difficult to read values on the axes for *BG3* due to the light stripes, movement of the water, and the contrast difference in the scene. However, the measured *error* reveal no such effect while those even performed the fastest. Overall, we have seen that the subjects perceived backgrounds with a higher FC value as more distracting while the measured data show no support for this.

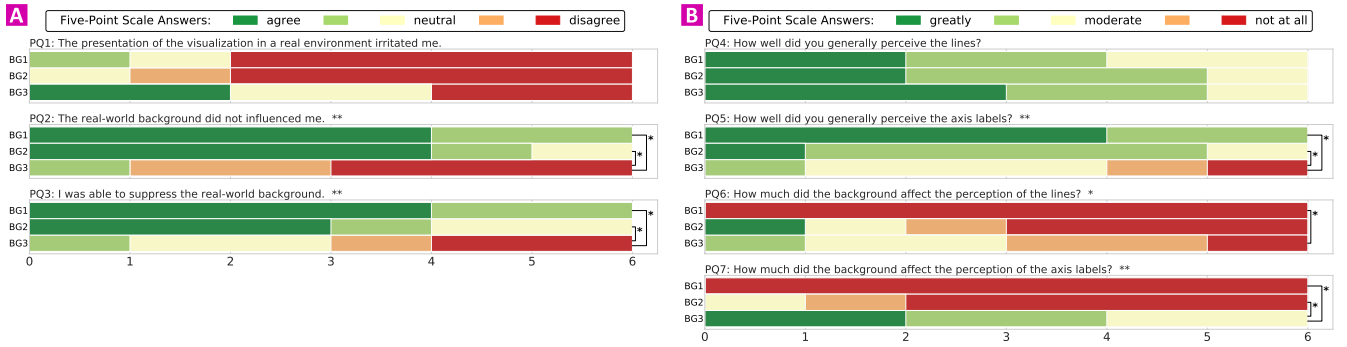

Figure 5: Survey results of questions scored on a five-point scale regarding the overall perception (PQ = perception question). The x-axis presents the number of participant that voted for each answer respectively. (: $p < 0.01$, *: $p < 0.05$)**

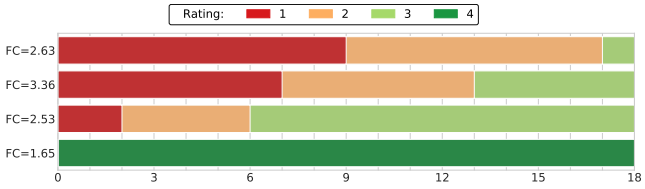

Figure 6: The order the subjects chose for four additional images presenting different background configurations. Each image has its own FC value. The rating 1 represents the image with the highest perceived distraction level.

Lastly, we wanted to get a feeling on how well the Feature Congestions (FC) can represent the perceived clutter of an image. Therefore, we presented the subjects four images of background configurations with fixed FC values which were partly not used in the study.[2] The subjects had to rate those based on how distracting the shown backgrounds would be to work on (see Fig. 6). The background with the highest distraction level should be placed in the first place. The ratings reveal that the FC values do not always match the perceived most distracting background. This can be seen by the images with close FC values of $FC = 2.63$ and $FC = 2.53$. While the $FC = 2.63$ is in the first place ($M = 1.56$), $FC = 2.53$ is in the third place ($M = 2.56$). We think that not only the specific FC value of each image defines how distracting subjects perceive a background. Also the specific characteristics, like the light stripes (2 out of 6 participants), moving water (2 out of 6) or the cables (1 out of 6) are quite important.

4 STUDY 2: INFLUENCE OF A SECONDARY TASK

In our first study, the background was decoupled from the AR visualization and was only used as a visual distraction. In most of the real-world environments, this would not be the case. Especially immersive and embedded visualizations [16, 37, 48, 50, 56] try to connect the shown virtual information with the real-world, as is also the case with research focusing on AR assisted assembly tasks [42, 53]. Additionally, machines in real-life production plants often already possess displays to show important information [51]. Therefore, users have to alter their focus between the background and the virtual information quite often. Overall, the second study aims to deepen the understanding of the influence of the background on information visualization while taking into account the forced attention switch between the background and the AR content caused by the introduction of a secondary task.

4.1 Design & Hypotheses

To understand the additional influence of a secondary task, we designed a user study that was also conducted in a real experimental production plant. We used the following two independent variables: the *background* (*BG3* and *BG4* as seen in Fig. 2) and *focus type* (*single focus* and *split focus*). The goal of the *focus type* was to increase the attention the participants had to pay to the background, which should increase the overall influence of the background itself. Both variables were used in a within-subject design and allowed us to measure the following dependent variables: *completion time*, *error* (as *absolute error* and *percentage error*), and questionnaire data. Further, we used a different subset of *analysis tasks* than in Study 1. We focused on the following hypotheses:

- **H5** The clutter of the *background* has no influence on the performance of the subjects (inverted **H1** from Study 1).
- **H6** The performance of solving the primary task gets worse when it is performed in the *split focus* condition.
- **H7** Solving tasks on a *background* with more clutter in the *split focus* condition decreases the performance.
- **H8** The *background* is more noticeable for subjects when the secondary task is introduced.
- **H9** Subjects perceive the *background* with more clutter as more distracting (the same as **H4** from Study 1).

4.2 Participants

We recruited 18 participants through e-mail and word-of-mouth for this study. All of them did not participate in Study 1 and were not compensated. The gathered data of only 16 volunteers could be used for our data analysis. One sessions had to be aborted due to

[2] Please refer to the supplemental material for the presented images.

external interruptions in the production hall, while the second subject showed major difficulties at understanding the presented tasks. All of the remaining 16 volunteers (11 male, 5 female) were students or post-graduates with engineering background. The average age of the participants was 24 year ($M = 23.81$ year, $SD = 3.08$ year) and the self-reported height ranged from 162 cm to 198 cm ($M = 178.81$ cm, $SD = 11.65$ cm). Again, no specific knowledge was required to participate in this study. All participants had normal or corrected-to-normal vision and had no spatial perception difficulties. We adjusted the colors for one subject due to red-green weakness. On a five-step scale, all participants had generally some experience with AR ($M = 2.44, SD = 0.89$), little experience with AR via head-mounted display ($M = 2.00, SD = 1.03$), no experience with VR ($M = 1.75, SD = 0.77$) and were generally quite experienced at working with visualizations ($M = 3.88, SD = 0.72$). The volunteers also had to solve a small introductory question on how well they could define elements in a line chart with a total score of eight points. Only one subject had one error (mixed up the keywords "x-value" and "x-axis") while the participant with the comprehension problems had four errors.

4.3 Setup

Overall, this study has the same setup as Study 1 but changed and altered the used backgrounds. To enable our secondary task design, we added three 6 inch smartphones (two Huawei Honer 9 and one Samsung Galaxy S4) as simulated displays of real-life modules to our backgrounds. We therefore used the following backgrounds with their respective FC values: $\overline{FC(BG3)} = 5.65$ and $\overline{FC(BG4)} = 2.75$. We removed *BG1* to reduce the complexity of the study and enable a within-subject design. We transformed *BG2* to *BG4* to create a wider background for a uniform smartphone placement between both backgrounds. This also allowed us to increase the difference between *BG4* and *BG3*, since *BG4* is more uniform by reducing the open cables of *BG2*. The height at which the additional smartphones were attached to the backgrounds differs between each position with approximately 95 cm on the left, 170 cm on the center, and 140 cm on the right (see Fig. 2d).

Our AR application is the same as in Study 1, while the server application, in addition, handles the events of the smartphones. The smartphone application was also implemented in Unity and only displayed a different number for each device to allow the simulation of a secondary observation task. Those ranged from 0 to 9 and could be colored in yellow, white, blue, or red, while green was used as a signal color (white and green were swapped for the subject with red-green weakness) (see Fig. 1d). The number changed in an interval of 5 s to 10 s for each display while the signal color appeared with the next network message to any device after a period of 90 s.

We tried to control possible disturbing factors in our setup. However, *BG3* still created a constant background noise, in addition to some people that worked in the production hall at the same time.

4.4 Procedure

Each session followed the following phases: (1) A small introduction to the experimental environment; (2) A declaration of consent and a first questionnaire; (3) An explanation of the *analysis tasks*, the interaction with the system and both *focus types*; (4) Short training with six questions and a preceded calibration of the HoloLens; (5) The first half of the experiment on the first *background*; (6) A short break with a second questionnaire regarding the first *background*; (7) The second half of the experiment with the second *background*; (8) A final and third questionnaire with questions connected to the second *background* as well as overall questions. The experiment was led by one investigator who was in the production hall and was seated partly inside the field of view of the participants, due to technical reasons.

The subjects had to interact with the system as described in Study 1. Additionally, the subjects had to raise their hand and speak aloud the value as soon as they saw a green number on one of the displays in the *split focus* condition. Each session lasted on average 78 min ($M = 77.98$ min, $SD = 10.63$ min), whereby 39 min ($M = 39.35$ min, $SD = 8.57$ min) were needed for phases (5) and (7).

4.5 Tasks

The subjects of the study had to solve *analysis tasks* on line charts, like in Study 1. Our set of task types consisted of seven primitive low-level analysis tasks [3]. Those were *Retrieve Value*, *Filter*, *Compute Derived Value*, *Find Extremum*, *Sort*, *Determine Range*, and *Cluster*. Additionally, we added *Compare* as a higher-level task which often uses different low-level analysis tasks (see Tab. 1). As in Study 1, we created three different blocks that contained four questions each. Furthermore, the questions built onto the answers of the preceding questions in the same block. To simulate a secondary observation task, we introduced the *focus type*. The *focus type* had two different states: *single focus* and *split focus*. In *single focus*, the participants only had to focus on solving the primary task presented for the visualization (as in Study 1), while in *split focus* they also had to observe the background as a secondary task. The subjects had to recognize and speak out the value of a green number (out of 5 different colors) on one of three displays placed inside the background configuration. The state of the *focus type* was switched after half of the tasks (six blocks) in phase (5) and (7) were solved.

Like in Study 1, we used line charts as our visualization type of choice. However, all charts were created with the same parameters: four colored lines (red, blue, yellow, white), x-value range of 0 to 35, and y-value range of 0 to 50. The color made the lines more distinguishable (see Fig. 3b). We created two different visualizations for each combination of *background*, *focus type* and block. In total, subjects had to solve 96 tasks (2 *backgrounds* x 2 *focus types* x 2 visualization x 3 blocks x 4 *analysis tasks*).

We again minimized a possible bias through training effects by counterbalancing the tasks. For each subject we ordered the four conditions created from the *backgrounds* and the *focus type*, finishing one *background* before switching to the second one (for example the order of: *BG3 single focus*, *BG3 split focus*, *BG4 single focus*, *BG4 split focus*). We cycled through all eight possible order combinations. Lastly, we ordered the 12 task blocks through a latin-square, split them into groups of three, and assigned each group to each condition, while keeping the order of *analysis tasks* in each task block.

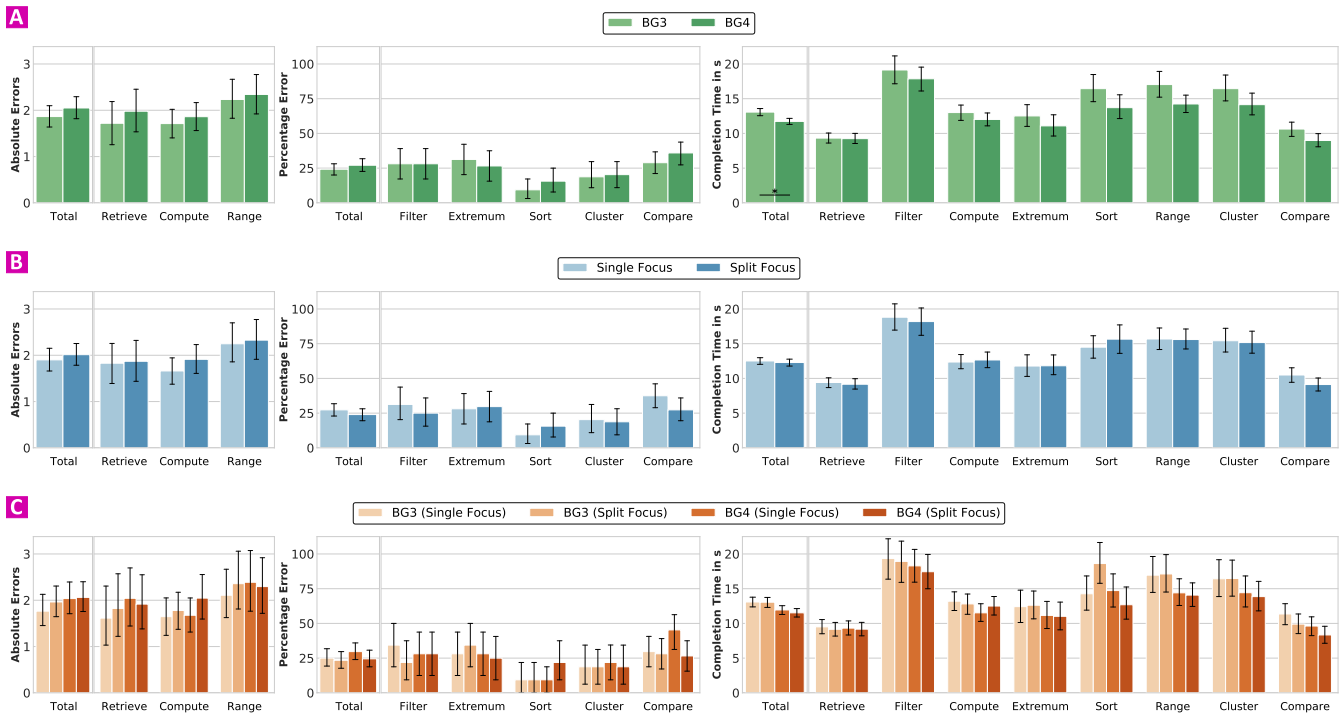

Figure 7: Different data analysis visualizations for the second study. (A) presents the *absolute error*, *percentage error* and *task completion time* grouped by each *analysis task* and *background*. (B) shows the same dependent variables for different *analysis tasks* and *focus type*. Lastly, (C) also displays all three dependent variables grouped by each *analysis tasks* and the combination of *background* and *focus type*. All bars show the mean value and a 95% confidence interval with the whiskers. (*: $p < 0.05$ in (A) for *task completion time*)

4.6 Measurements & Derived Data

As in Study 1, we collected the answers and the timestamps for each participants. Additionally, we logged the events connected to the smartphone (value updates and answers for the *split focus*). The derived data are also the same with *absolute error*, *percentage error*, and *task completion time*. Each session was again video recorded while the investigator took notes accordingly. The three questionnaires contained a total of 57 questions. 14 were assigned to demographic data, 31 were questions based on a rating scale and 12 were open questions. As in Study 1, we used the same guiding questions. However, in comparison to Study 1, we added questions regarding the *split focus*, while we repeated the questions for physical state, the recognition of visual elements, and the perception of the background after each individual *background*. Lastly, we added five closed question for the *focus task* regarding the influence on the main task (see Fig. 8).

4.7 Data Analysis

As in our first study, we checked and replaced outliers (2.5 % for *absolute error*, 4.3 % for *task completion time*) in our data with the following formula: $M + 2 * SD$ (as proposed by Field et al. [18]). We checked our data with preliminary tests before we analyzed them further. For this, we tested for normal distribution (Shapiro-Wilk), which was never the case, and sphericity (Mauchly), which was never violated, for the *background* and *focus type* separately. We used one-way ANOVAs on the *background* and *focus type*, while calculating two-way repeated measurements ANOVAs on the combination of *background* and *focus type*. Further, we use a Friedman F test for the questionnaire answers on a rating scale. In general, if we found any significance we calculated pairwise t-tests or Wilcoxon W tests.

4.8 Results

This section reveals the results of our second study, ordered by the presented hypotheses. They are depicted in visualizations of our measured data (see Fig. 7) and questionnaire data (see Fig. 8 and Fig. 9).

H5 (clutter on performance) The statistic tests show no significance effect on *absolute error* ($F(1, 15) = 0.418, p = 0.528, \eta_p^2 = 0.027$) and *percentage error* ($F(1, 15) = 0.850, p = 0.371, \eta_p^2 = 0.054$) while a significant influence can be seen for the *task completion time* ($F(1, 15) = 5.407, p < 0.05, \eta_p^2 = 0.265$). Further, Fig. 7(a) shows that *BG3* has always less *absolute error* than *BG4*, while this order changes for *percentage error* with different *analysis tasks*. In summary, *the performance is only partly influenced by the background clutter.*

H6 (focus condition on performance) The statistic tests display

no significance effect on the performance values *absolute error* ($F(1, 15) = 0.514, p = 0.484, \eta_p^2 = 0.033$), *percentage error* ($F(1, 15) = 1.338, p = 0.266, \eta_p^2 = 0.082$), and *task completion time* ($F(1, 15) = 0.298, p = 0.593, \eta_p^2 = 0.019$). However, Fig. 7(b) depicts that less *absolute errors* appear in the *single focus*, while this changes between different *analysis tasks* for the *percentage error*. In summary, *the focus type shows no significant influence on the performance of the subjects.*

H7 (clutter & focus type on performance) The statistic tests reveals no significant interaction between *background* and *focus type* on *absolute error* ($F(1, 15) = 0.022, p = 0.6454, \eta_p^2 = 0.014$), *percentage error* ($F(1, 15) = 0.657, p = 0.430, \eta_p^2 = 0.042$), and *task completion time* ($F(1, 15) = 0.305, p = 0.589, \eta_p^2 = 0.020$). Fig. 7(c) shows that the *analysis tasks* perform differently in *split focus* than in *single focus*. Interestingly, the way how the secondary task influenced the *analysis tasks* differs. Some increased the error and needed more time, while others reduced both. In summary, *there is no significant interaction between the background and the focus type regarding the performance of the subjects.*

H8 (focus type on perceived distraction) The questionnaire contained five questions regarding the additional *focus type* (FQ = focus question), as seen in Fig. 8. Overall, the subjects perceived that the *split focus* condition made it more difficult to solve the main task (*FQ1*: $M = 2.38, SD = 0.96$) as it needed some additional attention (*FQ4*: $M = 2.69, SD = 1.12$). The secondary task increased the perceived *task completion time* (*FQ2*: $M = 3.06, SD = 1.12$), while the given answers were only slightly influenced (*FQ3*: $M = 1.88, SD = 0.89$). We can conclude *that the split focus condition increased the awareness and the influence of the background for the subjects.*

H9 (clutter on perceived distraction) Our collected questionnaire data regarding the perception (PQ = perception question) of the *background* (see Fig. 9) shows no significant influence on the perceived distraction or performance. However, a small difference between both *backgrounds* exists. *BG4* is overall perceived less distracting than *BG3*, the background with dynamic elements and additional light. This is especially true for *PQ2* ($Q(1) = 3.267, p = 0.0707$), *PQ3* ($Q(1) = 1.923, p = 0.1655$), and *PQ7* ($Q(1) = 2.571, p = 0.1088$). We further asked the participants to choose which of the *backgrounds* was more difficult to work in. 13 out of 16 participants chose *BG3*. The light stripes (4 out of 16), the pipes (4 out of 16), the water movement (6 out of 16), and green color (4 out of 16) were stated as the main reasons for this. On the other hand, *BG4* was perceived as only slightly distracting (6 out of 16), which changed with the introduction of the *split focus* (3 out of 16). The sun light through a window in the background (5 out of 16), as well as the reflection on the smartphones (2 out of 16) were the most distracting factors. We conclude *that the background clutter has no significant influence on the perception. However, the open questions and comments depict a strong tendency that BG3 was perceived as more distracting.*

Further findings Additionally, we analyzed the physical state between each *background* as we did in Study 1. This time, we could not find any significant influence regarding the different questions. We also looked at the number of times the participants reported

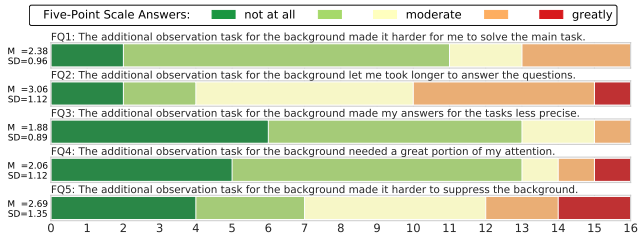

Figure 8: Survey results of questions scored on a five-point agreement scale regarding the *split focus* condition (FQ = focus questions). Each question also presents the mean score (M) and standard deviation (SD). The x-axis presents the number of participant that voted for each answer respectively.

a number with the correct signal color in the *split focus* condition. Here we can see that the subjects recognized the numbers less often on *BG3* ($M = 75.42\ \%, SD = 22.22\ \%$) than on *BG4* ($M = 83.10\ \%, SD = 20.09\ \%$). Some participants gave us negative comments about the field of view of the HoloLens v1 (6 out of 16) and some mentioned that they could not suppress the constant noise of *BG3* (8 out of 16). Lastly, a few participants wished for more support through the visualizations with a grid, ticks on the axis, and better overlap handling of the lines (2 out of 16 for each).

4.9 Discussion

Our results reveal that the *background* affects the *task completion time*, while the *errors* are not affected (**H5**). Tasks that were solved on more cluttered *backgrounds*, like *BG3*, took longer to solve than on a *background* with less clutter. Interestingly, with longer *task completion times* the subjects gave less error prone answers. This allows us to assume that harder to suppress backgrounds increase the duration that users need to find the correct answer. However, the answers themselves seem not to be influenced by the backgrounds. The collected subjective questionnaire answers also support that backgrounds with a higher FC value are perceived as more distracting and cluttered (**H9**). Further, in the questionnaire regarding *BG3*, some subjects found it hard to read values on the axes (4 out of 16). However, the answers given for tasks on this *background* were slightly better than for *BG4* (see *absolute errors* in Fig. 7(a)). Following, we think that users are quite capable to overcome the faced perceptual issues.

As we investigated our second independent variable, the *focus type*, we could not find any significance in the measured performance (**H6**). This is especially interesting since we believed that the increased task load through the *split focus* (**H8**) also would decrease the overall performance (**H7**). The participants also perceived the introduced *focus type* as more demanding. The questionnaire answers (see Fig. 8) show that the participants perceived a higher task load and increased *task completion time*. Overall, the *split focus* setup successfully increased the awareness of the *background* for the participants (*FQ5*: $M = 2.69, SD = 1.35$), which was also stated directly from some of the subjects (3 out of 16). Again, we can verify that the perceived and measured performance of our subjects greatly differ. Lastly, the difference of the recognition rate for the

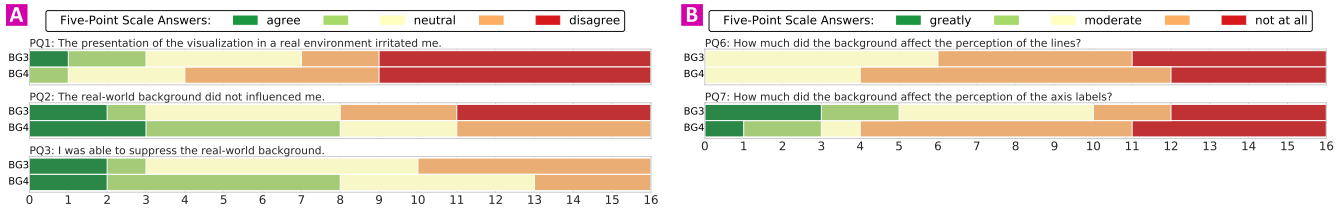

Figure 9: Survey results of questions scored on a five-point agreement scale focused on the perceptional questions (PQ = perception question). The x-axis presents the number of participant that voted for each answer respectively.

split focus condition on both backgrounds can be explained through the change blindness [26], as the signal color was the same as the color of the moving water in *BG3*.

5 OVERALL DISCUSSION & LIMITATIONS

In this section, we will discuss the results of our two studies in combination. For this, we will first look at the limitations of the studies and then present an overall discussion.

5.1 Limitations

With our presented user studies, we created a mixture of controlled lab and in-the-wild studies. Those allowed us to investigate the influence of real-world backgrounds on the perception of AR visualizations. However, our overall design decisions also introduced several limitations that we want to discuss.

For our investigation, we deliberately chose only one specific visualization type, line charts. Other visualization techniques and even small changes of visual parameters of e.g, bar charts or scatter plots might produce different results. This could be particularly true for 3D visualizations. In addition, our chosen visualizations only show static data. However, dynamic data sets, which are more common in industrial scenarios, can also change the overall perception and therefore the outcome of studies. The tasks performed on the shown visualization were cognitively challenging as shown by our NASA TLX on mental demand ($M = 6.09, SD = 1.73$). Tasks which are less demanding could also change how strong the influence of the background is perceived.

In our studies, we used the Microsoft HoloLens v1 to present AR content to the subjects. The usage of a head-mounted display is more practicable since it frees the hands for interaction but also comes with drawbacks. Especially the field of view is a limiting factor on how big visualizations can be and on how much information can be perceived at the same time. In our studies, 10 out of 34 participants mentioned the small field of view. The HoloLens was also perceived as quite uncomfortable (12 out of 34). The subjects reported pressure (5 out of 34) or pain (5 out of 34) on the nasal bridge. However, we think that manufacturers, as already seen with the HoloLens 2, will address those limitations.

Our user studies were placed in a real-world experimental production plant that introduced many possible uncontrollable factors like the temperature, the presence of other people, or noises. Our findings could slightly differ in an environment with varying influencing factors and background configurations. The backgrounds we created were use case specific and contained both, static and dynamic elements. Different backgrounds with more distinct properties, like color gradients, stronger movement, or additional and bigger information displays, could change the overall perception. The Feature Congestion (FC), in combination with other environmental characteristics, was used to create our background configurations. However, the FC value is calculated on static images and only takes into account three different image properties [46]. Those properties make it difficult to fully represent the complete background and environment AR applications could be used in.

Our overall setup was quite restricting. The experiments duration was rather short and the participants were instructed to sit for the whole duration. However, in a real scenario, users will move around, interact with the environment, change their posture, and their viewing angle. All those factors might change a users' perception with regard to the environment and the presented visualizations.

We are aware of the rather small sample size of both studies. However, our studies already show first important insights on how the background of a real-world scene could have or have not an influence on the perception of visualizations. Lastly, the participants of our studies only reflect a small spectrum of the relevant population as visible through the age distribution and the background the subjects were coming from. This is generally a problem if we want to analyze how the perception is influenced in an industrial scenario.

5.2 Discussion

This discussion is split into four different topics that we believe are the most important outcomes of our research.

The background has only marginal influence on the perception of visualizations shown in AR. Both our studies reveal that the background has no influence on the answers given to the presented tasks (**H1** and **H5**). However, the *task completion time* was influenced in Study 2, while it was not affected in Study 1. Overall, participants were able to ignore the influence of the presented background configurations. This makes us believe that AR visualizations can be used in many real-life scenarios without a big effect on users' performance. This can also be seen by the percentage of how often subjects recognized the signaled number in the *split focus* condition in Study 2, since the higher cluttered background (*BG3*) had a lower recognition value in comparison. However, this also shows that important information placed in such a background should be more prominent to be recognizable. Additionally, we believe that it is possible that users compensated for the increased difficulty by increasing their *task completion time*, as seen in Study 2. However,

we cannot see this compensation in Study 1 as well. In comparison, the *task completion time* in Study 1 was the fastest on *BG3*. This, as well as the rather high *error* rate in Study 1, could be explained by the different demographics of our study participants, the overall task difficulty (visualizations in Study 2 contained more data points), or the different visual representation of the visualization (only white lines in Study 1). Overall, we do not consider it advisable to prioritize adjustments to the real environment, as the decrease in performance based on more distracting environments is minimal.

There is more than visual clutter that defines a background and its influence on the perception. We used the Feature Congestion (FC) [46] as a guidance while we created background configurations for our studies. In general, the FC value can create a rather good definition on how cluttered or distracting a background scene is. However, it only uses images and therefore loses specific characteristics of a real-world scene (**H4** and **H9**). Such characteristics, as stated by the participants, can include the overall lighting (11 out of 34), reflections (3 out of 34), movement (8 out of 34), special areas like the pipes and their attachments (4 out of 34), or the cables of *BG2* (1 out of 34). The perception of those characteristics can further change, if users have to interact with objects in the background as well (**H8**). We think that a purely automatic calculation of visual clutter and distraction based on images (e.g., Feature Congestion) with regard to real-life environments is not enough, since specific characteristics of a background are not considered by it. Possible extensions of the FC algorithm should consider crowding [52], depth perception [17] of real-world scenes, or the dynamics of the background [17]. Lastly, also a learned classifier working with human labeled and rated images could be beneficial as well.

User perceive a visually cluttered background as more distracting than it actually affects their overall performance. Participants reported that the background (**H4** and **H9**), as well as the secondary task of Study 2 (**H8**) have an influence on their performance. The subjects had the feeling that they took longer and were less precise in their given answers while working on a background with a higher FC value. They also felt a higher task load or had a harder time to read values of the visualizations, like the axes. However, this perceived influence was not visible in their measured performance values. We think that in longer working scenarios than simulated in our experiments, this perception can also have an influence on the real performance. Additionally, since the perception is coupled to the user, the overall user satisfaction should also be considered. This is because a higher satisfaction could compensate for the loss in performance. Overall, the *backgrounds* resulted in an increase of cognitive load, while the answers given were not affected. This could change with longer working sessions or even over several sessions. Further, the safety of the user and the real-world environment should be a concern as well, since the subjects in our experiments were able to suppress the background rather well. This can be especially dangerous if they ignore warning signs in the background as simulated with the *focus type* in Study 2. In summary, we can see a difference between the measured and perceived performance of the users. This makes us believe that a greater focus on user experience could help in the design phase of user interfaces or visualizations for such AR applications. Lastly, a more user-centric design could help users to perceive their own work as more correct and productive.

Various task or visualization parameters could have an influence on the real and perceived performance. In both studies, we investigated an additional secondary independent variable, the *complexity* of the visualization in Study 1 and an additional secondary task in Study 2. The *complexity* showed a significant effect on the performance, while the interaction with the *background* had no such effect (**H2** and **H3**). The *focus type* of Study 2 reveals no significance, either individually or in interaction with the *background* (**H6** and **H7**). However, the secondary task increased the awareness of the background for the users (**H8**). In general, we believe that other parameters of visualizations or the presented tasks could have different effects on the user. The usage of various visual parameters (e.g, visualizations types, visual marks) or another secondary task, like changing parameters of a running production machine, could enhance or decrease the influence of the environment such an AR application is used in. Lastly, it seems advisable to not prioritize an adaptation to a rather simple secondary task.

6 CONCLUSION

In this paper, we presented the results of our investigations on how real-world environments can influence the perception of visualizations in Augmented Reality (AR), based on different background configurations created in an experimental production plant. In addition, we also investigated the complexity of visualizations and the introduction of a secondary parallel task in combination with the created background scenes. Both studies showed that the background has only a marginal influence on the measured task performance, while the perceived performance was affected by the real-world backgrounds. We discussed the results of the studies, as well as their limitations and provided insights and recommendations.

With this research project, we hope to contribute to a general understanding of the way how visualizations in AR are perceived. We think that further research work on perceptual issues can also yield important and interesting insights. Possible research directions could be the user experience on different visualizations types, visual variables, or user characteristics (e.g., age, visualization expertise). Overall, we hope that our work can be used to make informed design and development decisions and to bring AR a step closer to becoming a productive tool in real-life working scenarios.

ACKNOWLEDGMENTS

The authors wish to thank Wolfgang Büschel for his support and guidance in regards to the statistical analysis, the Process-To-Order Lab (see https://tu-dresden.de/ing/forschung/bereichs-labs/P2O-Lab) for granting us access to their facility, as well as the anonymous reviewers for their constructive feedback.

This work was funded in part by "Deutsche Forschungsgemeinschaft" (DFG, German Research Foundation) under grant number 319919706/RTG2323 "Conducive Design of Cyber-Physical Production Systems", under project number 389792660 as part of TRR 248 (see https://perspicuous-computing.science), under Germany's Excellence Strategy - EXC-2050/1 - Project ID 390696704 - Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden, and under Germany's

Excellence Strategy - EXC-2068 - 390729961 - Cluster of Excellence "Physics of Life" of Technische Universität Dresden.

REFERENCES

[1] Andrea Albarelli, Augusto Celentano, Luca Cosmo, and Renato Marchi. 2015. On the Interplay between Data Overlay and Real-World Context using See-through Displays. In *Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter - CHItaly 2015*. ACM Press, New York, New York, USA, 58–65. https://doi.org/10.1145/2808435.2808455

[2] Markus Aleksy and Johannes Schmitt. 2018. Utilizing Augmented Reality and Wearables in Industrial Applications. In *Advances in Intelligent Systems and Computing*. Vol. 612. Springer, Cham, 147–155. https://doi.org/10.1007/978-3-319-61542-4_13

[3] Robert Amar, James Eagan, and John Stasko. 2005. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* IEEE, 111–117. https://doi.org/10.1109/INFVIS.2005.1532136

[4] Ronald Azuma and Chris Furmanski. 2003. Evaluating label placement for augmented reality view management. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE Comput. Soc, 66–75. https://doi.org/10.1109/ISMAR.2003.1240689

[5] Mark Billinghurst, Adrian Clark, and Gun Lee. 2015. A Survey of Augmented Reality. *Foundations and Trends® in Human–Computer Interaction* 8, 2-3 (2015), 73–272. https://doi.org/10.1561/1100000049

[6] Wolfgang Büschel, Georg Eckert, and Raimund Dachselt. 2020. Challenges in Collaborative Immersive Visualization. *4th Workshop on Immersive Analytics: Envisioning Future Productivity for Immersive Analytics* (2020).

[7] Wolfgang Büschel, Annett Mitschick, and Raimund Dachselt. 2018. Here and Now : Reality-Based Information Retrieval. *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* May 2015 (2018), 171–180. https://doi.org/10.1145/3176349.3176384

[8] Wolfgang Büschel, Stefan Vogt, and Raimund Dachselt. 2019. Augmented reality graph visualizations. *IEEE Computer Graphics and Applications* 39, 3 (2019), 29–40. https://doi.org/10.1109/MCG.2019.2897927

[9] Simon Butscher, Sebastian Hubenschmid, Jens Müller, Johannes Fuchs, and Harald Reiterer. 2018. Clusters, Trends, and Outliers. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, New York, 1–12. https://doi.org/10.1145/3173574.3173664

[10] Maxime Cordeil, Andrew Cunningham, Benjamin Bach, Christophe Hurter, Bruce H. Thomas, Kim Marriott, and Tim Dwyer. 2019. IATK: An Immersive Analytics Toolkit. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 200–209. https://doi.org/10.1109/VR.2019.8797978

[11] Maxime Cordeil, Andrew Cunningham, Tim Dwyer, Bruce H Thomas, and Kim Marriott. 2017. ImAxes: Immersive Axes As Embodied Affordances for Interactive Multivariate Data Visualisation. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 71–83. https://doi.org/10.1145/3126594.3126613

[12] Juan David Hincapié-Ramos, Levko Ivanchuk, Srikanth Kirshnamachari Sridharan, and Pourang Irani. 2014. SmartColor: Real-time color correction and contrast for optical see-through head-mounted displays. In *ISMAR 2014 - IEEE International Symposium on Mixed and Augmented Reality - Science and Technology 2014, Proceedings*. IEEE, 187–194. https://doi.org/10.1109/ISMAR.2014.6948426

[13] Luís Fernando de Souza Cardoso, Flávia Cristina Martins Queiroz Mariano, and Ezequiel Roberto Zorzal. 2020. A survey of industrial augmented reality. *Computers & Industrial Engineering* 139 (2020), 106159. https://doi.org/10.1016/j.cie.2019.106159

[14] Saverio Debernardis, Michele Fiorentino, Michele Gattullo, Giuseppe Monno, and Antonio Emmanuele Uva. 2014. Text Readability in Head-Worn Displays: Color and Style Optimization in Video versus Optical See-Through Devices. *IEEE Transactions on Visualization and Computer Graphics* 20, 1 (2014), 125–139. https://doi.org/10.1109/TVCG.2013.86

[15] Arturo Deza, Emre Akbas, and Miguel P Eckstein. 2016. Piranhas Toolkit: Peripheral Architectures for Natural, Hybrid and Artificial Systems. https://github.com/ArturoDeza/Piranhas

[16] Neven A.M. Elsayed, Christian Sandor, and Hamid Laga. 2013. Visual Analytics in Augmented Reality. In *2013 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2013*. IEEE, 1–4. https://doi.org/10.1109/ISMAR.2013.6671817

[17] A Erickson, K Kim, G Bruder, and G Welch. 2020. A Review of Visual Perception Research in Optical See-Through Augmented Reality. *ICAT-EGVE* (2020), 1–9. https://sreal.ucf.edu/wp-content/uploads/2020/10/DarkModeSurvey_ICAT_EGVE_2020-2.pdf

[18] A Field, J Miles, and Z Field. 2012. *Discovering statistics using R*. Sage Publications Ltd.

[19] Michele Fiorentino, Saverio Debernardis, Antonio E. Uva, and Giuseppe Monno. 2013. Augmented Reality Text Style Readability with See-Through Head-Mounted Displays in Industrial Context. *Presence: Teleoperators and Virtual Environments* 22, 2 (2013), 171–190. https://doi.org/10.1162/PRES_a_00146

[20] Adrien Fonnet and Yannick Prie. 2019. Survey of Immersive Analytics. *IEEE Transactions on Visualization and Computer Graphics* (2019), 1–22. https://doi.org/10.1109/tvcg.2019.2929033

[21] Emanuele Frontoni, Jelena Loncarski, Roberto Pierdicca, Michele Bernardini, and Michele Sasso. 2018. Cyber Physical Systems for Industry 4.0: Towards Real Time Virtual Reality in Smart Manufacturing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10851 LNCS. Springer, Cham, 422–434. https://doi.org/10.1007/978-3-319-95282-6_31

[22] Chris Furmanski, Ronald Azuma, and Mike Daily. 2002. Augmented-reality visualizations guided by cognition: perceptual heuristics for combining visible and obscured information. In *Proceedings. International Symposium on Mixed and Augmented Reality*. IEEE, 215–320. https://doi.org/10.1109/ISMAR.2002.1115091

[23] Joseph L. Gabbard and J. Edward Swan. 2008. Usability engineering for augmented reality: Employing user-based studies to inform design. In *IEEE Transactions on Visualization and Computer Graphics*, Vol. 14. 513–525. https://doi.org/10.1109/TVCG.2008.24

[24] Michele Gattullo, Antonio E. Uva, Michele Fiorentino, and Joseph L. Gabbard. 2015. Legibility in Industrial AR: Text Style, Color Coding, and Illuminance. *IEEE Computer Graphics and Applications* 35, 2 (2015), 52–61. https://doi.org/10.1109/MCG.2015.36

[25] Michele Gattullo, Antonio Emmanuele Uva, Michele Fiorentino, and Giuseppe Monno. 2015. Effect of Text Outline and Contrast Polarity on AR Text Readability in Industrial Lighting. *IEEE Transactions on Visualization and Computer Graphics* 21, 5 (2015), 638–651. https://doi.org/10.1109/TVCG.2014.2385056

[26] Melinda S. Jensen, Richard Yao, Whitney N. Street, and Daniel J. Simons. 2011. Change blindness and inattentional blindness. *Wiley Interdisciplinary Reviews: Cognitive Science* 2, 5 (2011), 529–546. https://doi.org/10.1002/wcs.130

[27] Kangsoo Kim, Mark Billinghurst, Gerd Bruder, Henry Been Lirn Duh, and Gregory F. Welch. 2018. Revisiting trends in augmented reality research: A review of the 2nd Decade of ISMAR (2008-2017). *IEEE Transactions on Visualization and Computer Graphics* 24, 11 (2018), 2947–2962. https://doi.org/10.1109/TVCG.2018.2868591

[28] Chris D. Kounavis, Anna E. Kasimati, and Efpraxia D. Zamani. 2012. Enhancing the tourism experience through mobile augmented reality: Challenges and prospects. *International Journal of Engineering Business Management* 4, 1 (2012), 1–6. https://doi.org/10.5772/51644

[29] Ernst Kruijff, Jason Orlosky, Naohiro Kishishita, Christina Trepkowski, and Kiyoshi Kiyokawa. 2018. The Influence of Label Design on Search Performance and Noticeability in Wide Field of View Augmented Reality Displays. , 2821–2837 pages. https://doi.org/10.1109/TVCG.2018.2854737

[30] Ernst Kruijff, J. Edward Swan, and Steven Feiner. 2010. Perceptual Issues in Augmented Reality Revisited. In *9th IEEE International Symposium on Mixed and Augmented Reality 2010: Science and Technology, ISMAR 2010 - Proceedings*. IEEE, 3–12. https://doi.org/10.1109/ISMAR.2010.5643530

[31] Alex Leykin and Mihran Tuceryan. 2004. Automatic determination of text readability over textured backgrounds for augmented reality systems. In *ISMAR 2004: Proceedings of the Third IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE, 224–230. https://doi.org/10.1109/ISMAR.2004.22

[32] Chao Liu, Sheng Cao, Wayne Tse, and Xun Xu. 2017. Augmented Reality-assisted Intelligent Window for Cyber-Physical Machine Tools. *Journal of Manufacturing Systems* 44 (2017), 280–286. https://doi.org/10.1016/j.jmsy.2017.04.008

[33] Mark A. Livingston, Joseph L. Gabbard, J Edward Swan Ii, Ciara M. Sibley, and Jane H. Barrow. 2013. Basic Perception in Head-Worn Augmented Reality Displays. In *Human Factors in Augmented reality environments*. Springer, New York, 35–65. https://doi.org/10.1007/978-I-461

[34] Weiquan Lu, Henry Been Lirn Duh, Steven Feiner, and Qi Zhao. 2014. Attributes of Subtle Cues for Facilitating Visual Search in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2014), 404–412. https://doi.org/10.1109/TVCG.2013.241

[35] Weiquan Lu, Dan Feng, Steven Feiner, Qi Zhao, and Henry Been-Lirn Duh. 2014. Evaluating subtle cueing in head-worn displays. In *Proceedings of the Second International Symposium of Chinese CHI on - Chinese CHI '14*. ACM Press, New York, 5–10. https://doi.org/10.1145/2592235.2592237

[36] Jacob Boesen Madsen, Markus Tatzqern, Claus B. Madsen, Dieter Schmalstieg, and Denis Kalkofen. 2016. Temporal Coherence Strategies for Augmented Reality Labeling. *IEEE Transactions on Visualization and Computer Graphics* 22, 4 (2016), 1415–1423. https://doi.org/10.1109/TVCG.2016.2518318

[37] Kim Marriott, Falk Schreiber, Tim Dwyer, Karsten Klein, Nathalie Henry Riche, Takayuki Itoh, Wolfgang Stuerzlinger, and Bruce H. Thomas. 2018. *Immersive Analytics*. Lecture Notes in Computer Science, Vol. 11190. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-01388-2

[38] Eleni Michailidou, Simon Harper, and Sean Bechhofer. 2008. Visual complexity and aesthetic perception of web pages. In *SIGDOC 2008 - Proceedings of the 26th ACM International Conference on Design of Communication*. 215–223. https://doi.org/10.1145/1456536.1456581

[39] Thomas Moser, Markus Hohlagschwandtner, Gerhard Kormann-Hainzl, Sabine Pölzlbauer, and Josef Wolfartsberger. 2019. Mixed Reality Applications in Industry: Challenges and Research Areas. In *Lecture Notes in Business Information Processing*, Vol. 338. Springer, Cham, 95–105. https://doi.org/10.1007/978-3-030-05767-1_7

[40] Sandra Murphy, Nick Fraenkel, and Polly Dalton. 2013. Perceptual load does not modulate auditory distractor processing. *Cognition* 129, 2 (nov 2013), 345–355. https://doi.org/10.1016/j.cognition.2013.07.014

[41] Aude Oliva, Michael L Mack, Mochan Shrestha, and Angela Peeper. 2004. Identifying the perceptual dimensions of visual complexity of scenes. In *Proc. of the 26th Annual Meeting of the Cognitive Science Society*. 1041–1046. https://escholarship.org/uc/item/17s4h6w8

[42] Volker Paelke and Carsten Röcker. 2015. User Interfaces for Cyber-Physical Systems: Challenges and Possible Approaches. In *DUXU 2015*, Vol. 9186. Springer, Cham, 75–85. https://doi.org/10.1007/978-3-319-20886-2_8

[43] Stephen D. Peterson, Magnus Axholt, and Stephen R. Ellis. 2008. Label segregation by remapping stereoscopic depth in far-field augmented reality. In *Proceedings - 7th IEEE International Symposium on Mixed and Augmented Reality 2008, ISMAR 2008*. IEEE, 143–152. https://doi.org/10.1109/ISMAR.2008.4637341

[44] Giulia Pugnaghi, Daniel Memmert, and Carina Kreitz. 2020. Loads of unconscious processing: The role of perceptual load in processing unattended stimuli during inattentional blindness. *Attention, Perception, and Psychophysics* 82, 5 (jul 2020), 2641–2651. https://doi.org/10.3758/s13414-020-01982-8

[45] Patrick Reipschlager, Tamara Flemisch, and Raimund Dachselt. 2020. Personal Augmented Reality for Information Visualization on Large Interactive Displays. *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. https://doi.org/10.1109/TVCG.2020.3030460 arXiv:2009.03237 [cs.HC]

[46] Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. 2007. Measuring visual clutter. *Journal of Vision* 7, 2 (2007). https://doi.org/10.1167/7.2.17

[47] Bahador Saket, Alex Endert, and Cagatay Demiralp. 2019. Task-Based Effectiveness of Basic Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 7 (2019), 2505–2512. https://doi.org/10.1109/TVCG.2018.2829750 arXiv:1709.08546

[48] Dieter Schmalstieg and Tobias Hollerer. 2016. *Augmented Reality: Principles and Practice*. Addison-Wesley Professional. 13–18;239–270; pages.

[49] Ronell Sicat, Jiabao Li, Junyoung Choi, Maxime Cordeil, Won-Ki Jeong, Benjamin Bach, and Hanspeter Pfister. 2019. DXR: A Toolkit for Building Immersive Data Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 715–725. https://doi.org/10.1109/TVCG.2018.2865152

[50] Richard Skarbez, Nicholas F. Polys, J. Todd Ogle, Chris North, and Doug A. Bowman. 2019. Immersive Analytics: Theory and Research Agenda. *Frontiers in Robotics and AI* 6 (2019), 82. https://doi.org/10.3389/frobt.2019.00082

[51] Hanas Subakti and Jehn-Ruey Jiang. 2018. Indoor Augmented Reality Using Deep Learning for Industry 4.0 Smart Factories. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*. IEEE, 63–68. https://doi.org/10.1109/COMPSAC.2018.10204

[52] Ronald van den Berg, Frans W. Cornelissen, and Jos B.T.M. Roerdink. 2009. A crowding model of visual clutter. *Journal of Vision* 9, 4 (apr 2009), 24–24. https://doi.org/10.1167/9.4.24

[53] X. Wang, S. K. Ong, and A. Y.C. Nee. 2016. A comprehensive survey of augmented reality assembly research. *Advances in Manufacturing* 4, 1 (2016), 1–22. https://doi.org/10.1007/s40436-015-0131-4

[54] Xiyao Wang, David Rousseau, Lonni Besançon, Mickael Sereno, Mehdi Ammi, and Tobias Isenberg. 2020. Towards an Understanding of Augmented Reality Extensions for Existing 3D Data Analysis Tools. In *ACM CHI 2020*. https://doi.org/10.1145/3313831.3376657

[55] Matt Whitlock, Stephen Smart, and Danielle Albers Szafir. 2020. Graphical Perception for Immersive Analytics. *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2020), 616–625. https://cmci.colorado.edu/visualab/3DPerception/3DPerception.pdf

[56] Wesley Willett, Yvonne Jansen, and Pierre Dragicevic. 2017. Embedded Data Representations. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 461–470. https://doi.org/10.1109/TVCG.2016.2598608

[57] Hsin-Kai Wu, Silvia Wen-Yu Lee, Hsin-Yi Chang, and Jyh-Chong Liang. 2013. Current status, opportunities and challenges of augmented reality in education. *Computers & Education* 62 (2013), 41–49. https://doi.org/10.1016/j.compedu.2012.10.024 arXiv:1204.1594