Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP CMS 2019 (2019) 000–000

www.elsevier.com/locate/procedia

CIRP Manufacturing Systems Conference 2019

A Virtual Reality Assembly Assessment Benchmark for Measuring VR Performance & Limitations

Michael Otto[a*], Eva Lampen[a], Philipp Agethen[a], Mareike Langohr[a], Gabriel Zachmann[b], Enrico Rukzio[c]

[a]*Daimler Research & Development, 73760, Ulm, Germany*
[b]*University of Bremen, 28359, Bremen, Germany*
[c]*Ulm University, Institute of Media Informatics, 73760, Ulm, Germany*

* Corresponding author. *E-mail address:* michael.m.otto@daimler.com

Abstract

With an increasing product complexity in manufacturing industry, virtual reality (VR) offers the possibility to immersively assess assembly processes already in early product development stages. Within production validation phases, engineers visually assess product part assembly and interactively validate corresponding production processes. Nevertheless, by now research does not give answers on how VR assembly system's performance can be measured with respect to its technical limitations. The proposed Virtual Reality Assembly Assessment (VR2A) benchmark is an open, standardized experiment design for evaluating the overall VR assembly assessment performance in terms of sizes and clearances instead of measuring single technical impact factors within the interaction cycle, such as tracking, rendering and visualization limitations. VR2A benchmark focusses on the overall production engineer's assessment objective generating quantifiable metrics. Using VR2A, users gain practical insights on their overall VR assessment system's performance and limitations. An in-depth evaluation with production engineers (N=32) revealed, that negative clearances can be detected more easily than positive ones, part sizes directly correlate with the assessment performance. Additionally, the evaluation showed that VR2A is easy to use, universally usable and generates objective insights on the applied VR system.

Peer-review under responsibility of the scientific committee of the 52nd CIRP Conference on Manufacturing Systems.

Keywords: Virtual Reality; Benchmark; Assembly; Assessment; Usabitity; Score; Production Validation;

1. Introduction

With the vast availability of low-cost HMDs and novel tracking technologies, virtual reality conquered broad new industries, such as gaming, entertainment, sales and of course manufacturing industry, even though research on this topic is carried out already for several decades [1]. Each professional VR application follows an overall purpose, such as excitement in gaming [2], positive emotions for point of sale applications [3], novel rehabilitation methods [4] in medicine, more effective learning in schools [5], [6] and of course higher quality validation results in manufacturing industry [7].

Today, automotive industry already vastly utilizes VR technology for product, process and resource assessments for higher quality in planning results. For example, these virtual methods are used in planning departments of automotive final assembly. Each novel products is assessed multiple times during the product development process. Within production validation workshops multiple aspects are optimized, such as packaging, visibility, assemblability, production ergonomics, process quality, process efficiency, logistics, walk paths and many more. Therefore, the novel product is built up multiple times using digital methods – just like in the physical domain. Using VR assembly simulations, parts are dynamically inserted in the virtual product at its respective manufacturing state. By performing such a virtual assembly, the above mentioned objectives are validated. In comparison to the physical domain, research currently does not give any answers on how to

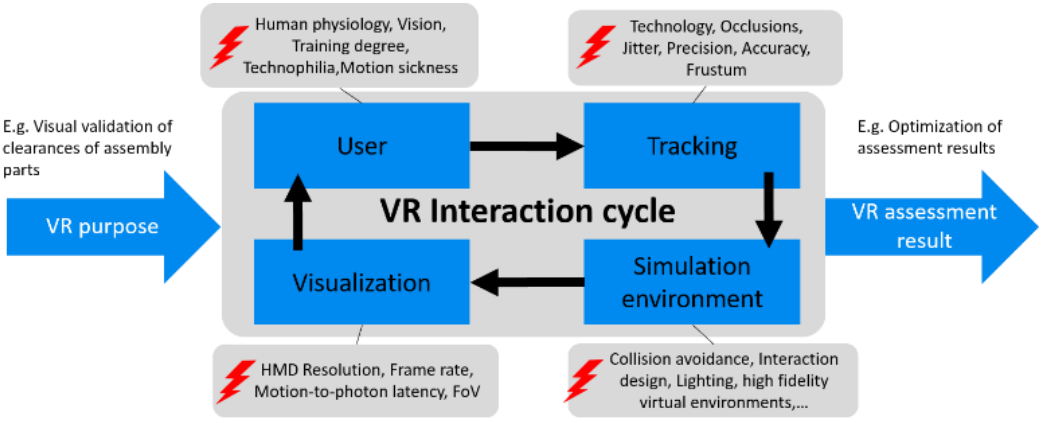

Fig. 1. Block diagram of VR interaction cycle including error influence factors

measure the limitations of such a virtual assembly simulation system.

Using VR in production validation, immersion is not an end in itself, but has to be beneficial to the overall assessment goal. In literature, immersion is described as one of the main advantages of VR, but for professional use of VR, users expect to achieve their goals either in a higher quality or in a more efficient way. Many papers propose using VR for better immersion and better spatio-temporal understanding of the upcoming production process (compare Bowman et al. [8]). In this paper the "VR assembly assessment" (VR2A) benchmark is proposed as a unified experiment design, in order to quantify the practical VR system's performance without measuring the VR interaction cycle influence parameters. More precisely, VR2A measures the user's ability to visually assess the assemblability of the digital mock-up (DMU) with respect to two independent variables: Assembly part sizes and clearances. The user represents both the operator and the product assessor at the same time, just as in real production validation workshop situations. VR2A measures whether production engineers can achieve their assessment goals even with small parts and low clearances and how small both may get.

2. State of The art

Research presents many publications on the use of VR in the context of automotive production. Zimmermann presents a brief overview on VR use cases in his survey [9] as well as Ottoson [10] throughout the product development process. Lawson et al. discusses future directions of VR for automotive manufacturers in a survey with 11 engineers, where they show up further VR development needs [11]. Berg and Vance present an overview on the application scenarios in product design and manufacturing [12]. Multiple academic publications on VR in automotive production are presented in the following topics: Production verification and maintenance (see Gomes de Sá and Zachmann [13]), training use cases [14], [15], product design and packaging [12] and continuous improvement process [16].

All of these use cases share the same goal, that they apply VR technology for a better spatio-temporal understanding and immersive effects for the users. Basic VR research present the effects on how immersion influences the behavior in virtual environments (VE) and its effectiveness. Immersion creates a feeling of presence in the VE or a feeling of "being there" and is often described as "the outcome of a good [gaming] experience" [17]. Jennett et al. presented research on the experience of immersion in games and found that immersion can be measured both subjectively using questionnaires and objectively by measuring task completion time or eye movements [17]. Interestingly, Ellis [18] doubts that presence might directly lead to better task performance, for instance when a more abstract view of an environment is required, for instance in flight control use cases, for achieving the goal. Beforehand, Witmer et al. present a well-known "presence questionnaire", which became a standard for measuring presence in VR [19], which is also applied in this study. Bowmann and McMahan ask, how much immersion is enough in VR [8] and give an overview on empirical studies which show, that full immersion is not always necessary.

Overall, literature does not yet present a uniform experiment design as a benchmark for VR assembly assessments for quantifying the VR system's limitations. Most closely, Funk et al. present a uniform experiment design as a benchmark for evaluating interactive instructions using augmented reality for assembly tasks [20], which differs in the benchmark scope, since Funk et al. evaluate task completion times whereas VR2A is intended to quantify the geometric limitations.

3. Influence parameters on the overall VR purpose

The VR interaction cycle consists of tracking devices, simulation software, rendering pipeline, hardware devices and of course the user itself. Each of those components inherits various sources of errors, unpredictable behavior and influence parameters. Fig. 1 depicts a simplified VR interaction cycle including exemplary error influence parameters of each component. The following exemplary error sources limit the overall VR system's performance:

- *Stable and precise tracking* is crucial for a good VR experience. All tracked components need precise 6DoF tracking. Typical limitations of the tracking system are optical occlusions, limited spatial frustum and limited tracking precision, jitter and accuracy.
- The *simulation software* also introduces multiple sources of errors in the interaction cycle, such as unsuitable usability, rendering issues, scene lighting, simulation software properties and missing collision detection and avoidance.
- *VR visualization devices* such as HMDs have a limited field of views, limited motion-to-photon latency, limited

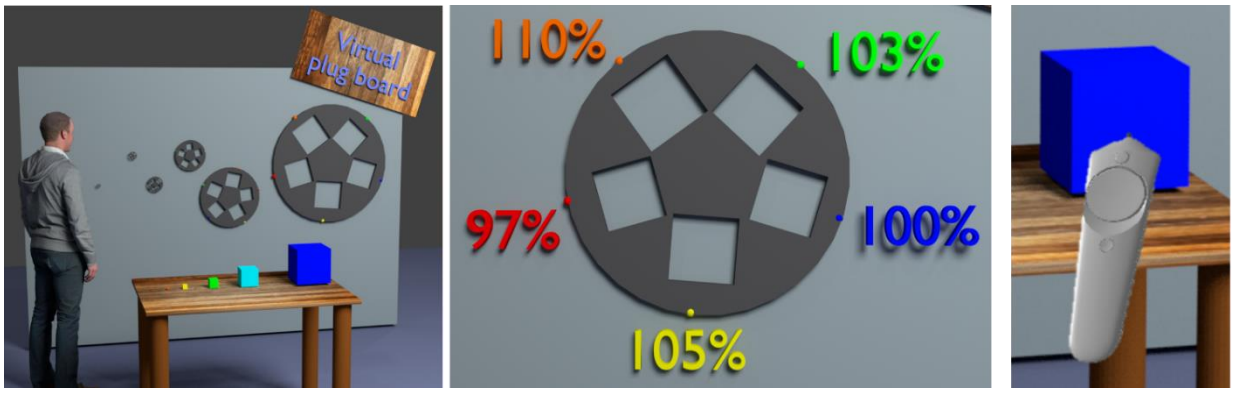

Fig. 2. Left: Rendering of the open virtual environment with six differently sized cubes. Middle: Explanation of disc cavities relative to the corresponding cube sizes, which are not visible to the user. Right: Controller with sharp grasping point

framerate and resolution. This is why visualization additionally induces errors in the interaction cycle itself.

- Finally yet importantly, one major influence factor on the overall system performance is the *user herself / himself*. For fulfilling the overall VR simulation purpose, he /she has to be able to interact with the whole system, so his training degree can be a potential source of errors. Additionally limitations in his physiology, vision and perception in general will influence the overall VR assessment results, such as human tremble or uncorrected vision.

As the abovementioned non-exhaustive list of errors shows, there are too many influence parameters to control every single one of it. Nevertheless, the users are not interested in quantifying these various VR system's properties, but want to know, if they can reach their VR assessment goals efficiently. Concisely, this is why from a production engineer's perspective, each single error parameter presented in Fig. 1 is less important than the overall VR system's performance. The respective error parameters in the interaction cycle can be regarded as a black box with an overall limitation for reaching the assessment task. Therefore, using VR2A benchmark, the system is tested for its applicability towards its native purpose.

4. The Virtual Reality Assembly Assessment benchmark

VR2A is proposed as an open, standard experiment design to evaluate a VR system's overall geometric limitations for assembly assessment scenarios and is considered to be "quick and easy". The VR2A scene is available here: https://skfb.ly/6FQOV

Two parameters are varied in an abstract assembly task: Clearance and Assembly part sizes. By conducting the VR2A benchmark, the user gains quantified insights on how small the assembly parts and clearances can be reliably be assessed by production engineers to still get reliable assembly assessment results. On purpose, VR2A abstracts all above-mentioned influence and error parameters within the interaction cycle and only focusses on the assembly relevant assessment results: Assessment of clearances and part size limitations.

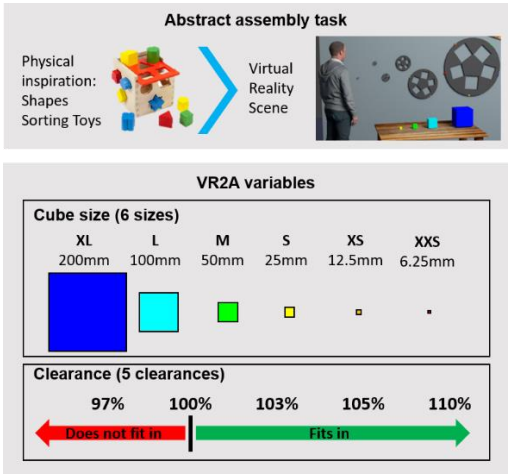

Fig. 3. Overview on the standard experiment design of VR2A and the two independent variables: Size and Clearance

VR2A carries out an abstract assembly task, inspired by kid's game called "shapes sorting toy" (see Fig. 3). The virtual reality scene has been published to set VR2A as a standard benchmark. As depicted in Fig. 2, within the virtual environment, there is a static table, six static discs each with five cavities on a wall. On the table, six dynamic (graspable) cubes are placed with the following sizes:

- XXS (6.25mm, red)
- XS (12.5mm, orange)
- S (25mm, yellow)
- M (50mm, green)
- L (100mm, cyan)
- XL (200mm, blue)

All six discs are placed on the wall, which are horizontally rotated and flipped in randomized angles. Each disc contains five cavities corresponding relatively to the sizes of the cubes (see Fig. 3.). Each disc has five cavities at the size of 97%, 100%, 103%, 105% and 110% relatively to the corresponding cube size (see Fig. 3. right). For example, the XL disc's 100% cavity matches exactly the size of the XL cube. The L cube does not fit in the respective "97% L cavity", but the S cube does fit in the respective "S 103% cavity".

The procedure of the benchmark is designed straight forward: Each participant inserts all six cubes in each of the

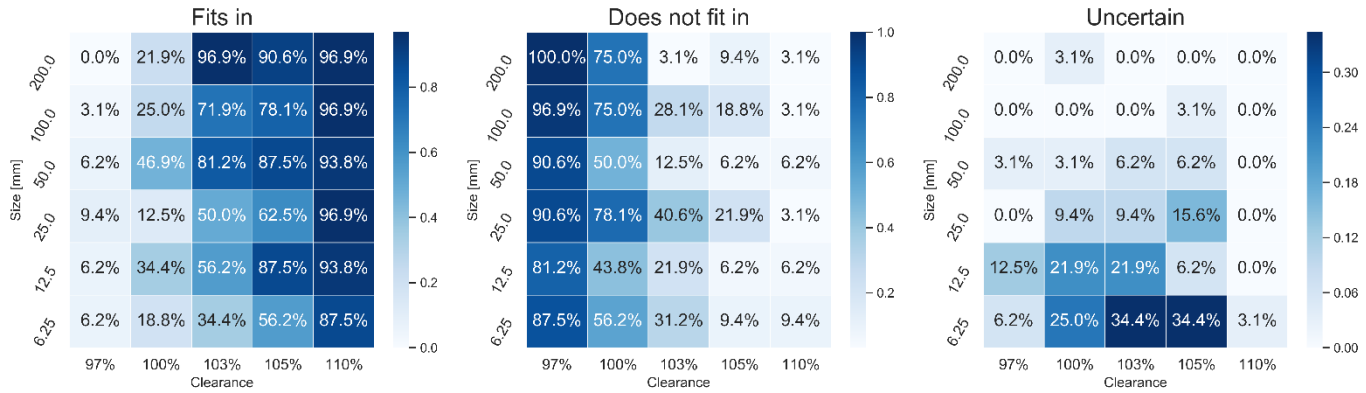

Fig. 4. Relative frequencies of the participant's answers in VR2A benchmark over the different scenarios

five corresponding cavities of the matching disc size. For example, the L cube has to be assembled in all five L disc's cavities in randomized order. The user does not know the correct answer in contrast to the experimenter. His answer possibilities are "Fits in", "Does not fit in" and "I can't assess it". The experimenter tells the participant, that it is not the goal to place the cube without collision, but to assess correctly whether it could be mounted that way – according to real production validation tasks. Task completion time is not in the scope of this task.

The results are calculated as follows: Each of the three answer possibilities are sorted into matrices containing the relative frequency for each condition. The relative frequencies of answers "Fit in" ($A_{Positive}$), "Does not fit in" ($A_{Negative}$) and "I'm unsure" ($A_{Neutral}$) are calculated. (1) calculates the relative homogeneity of answers between the assessments. If $S_{homogenity}$ equals zero in the matrix, the value of 0% would indicate, that the same amount of people state "Fits in" and "Does not fit in". Therefore, the assembly assessment would not include any reliable results.

$$S_{homogenity} = abs(A_{Positive} - A_{Negative}) \quad (1)$$

The overall VR2A score S_{VR2A} additionally penalizes "I'm unsure" feedbacks by the participants (see (2)). Therefore, VR2A score can be interpreted as the overall uncertainty for each variation of size and clearance.

$$S_{VR2A} = (abs(A_{Positive} - A_{Negative}) - A_{Neutral}) \quad (2)$$

Therefore, S_{VR2A} can theoretically range from -100% to 100%. Using these results, the overall VR system limitations can be explored using VR2A. Setting an individual threshold of for example 80% VR2A, gives a clear understanding, how small assembly parts and clearances may get in order to achieve the personal VR assessment purpose.

4.1. Evaluation using VR2A

In this study we use the VR2A benchmark to evaluate the overall performance of a VR assembly simulation system applied in automotive industry. Therewith, validations on assemblability are carried out. Even though, automotive products and the resulting assembly paths can be more complex, this abstracted assembly task gives useful insights on the system's performance.

4.2. Setup, stimuli and design

The hardware setup consists of a HTC Vive Business Edition (110° field of view, 2.160 x 1.200 resolution) attached to a high-performance Intel Core i7-8700k PC, 16GB RAM with a GTX 1080 TI graphics card. The tracking devices are calibrated in accordance with the technical specifications. The open VR2A scene is loaded in an proprietary assembly simulation software veo:IPV. This software natively supports the HTC Vive headset via OpenVR. Assembly parts (VR2A cubes) are set to dynamic objects. No physics, collision detection or gravity are turned on during the evaluation. The participant's use the HTC vive VR controller. Its virtual representation is visualized 1:1, but ending in a sharp cone as the root point, to allow for as precise grasping as possible for the participants (see Fig. 2. right).

4.3. Participants

For this study 32 production validation workshop participants were selected on a voluntary basis, such as research engineers, ergonomics experts, production engineers and students all working for several departments planning departments in an automotive OEM company. Therefore, this study was directly carried out with the intended key users of the system. They did not get any extra rewards for taking place in this study. 24 male and 8 female participants took part, all ranging from 18 to 51 years (M=28.2, SD=6.7). All participants reported normal to corrected vision.

4.4. Procedure

The experiment consists of two parts, namely the VR2A experiment and a final questionnaire. The experiment took about 25 minutes per user. 10 minutes for the VR2A evaluation itself and 15 minutes to fill out the questionnaires.

The experimenter warmly welcomed the user and described the assembly task in a standardized way. The participants are asked to get familiar the VR environment, the controllers, the virtual scene and dynamic handling of the cubes by playing around with them. When the participant felt confident in manipulating the virtual scene, he absolves all 30 VR2A

assembly tasks. Starting with the biggest cube (XL) through the smallest (XXS), each cube is inserted in all five corresponding cavities of each disc, but the experimenter randomizes the order of the cavities. For each cavity, the user verbally tells the experimenter the result of his visual assessment, if the cube fits into the cavity without collision. If required by the VR user, the experimenter adjusts the vertical height so that the user always has a comfortable viewpoint on the discs.

After finishing the assembly task, the participant fills out questionnaires, consisting of five non-standardized assembly experience questions and two standardized questionnaires, the "Prescence Questionnaire" and the "System Usability Scale".

4.5. Results

VR2A benchmark gives insights on the limitations of size and clearance performing a VR assembly assessment tasks. Fig. 4 depicts the relative frequencies of the according answers "Fits in", "does not fit in", and "uncertain". Hence, for clearances >100%, the objectively correct answer is "Fits in" whereas for <100% clearance scenario, the objectively correct answer is "does not fit in", since cubes overlap with the disc. For 100% clearance scenario, the expected answer would be "uncertain", since theoretically, the cube fits in, practically in VR the cubes cannot be placed mathematically correct position without any overlap. Interestingly, for the "100% clearance scenario", in mean 63.02% of the participants decide for the answer "Does not fit in" whereas only 26.56% decide for "Fits in". Only 10.42% decide for "I don't know".

The data presented in Fig. 4 is the source data to calculate VR2A score using equation (2)**Fehler! Verweisquelle konnte nicht gefunden werden.**. Results are depicted in Fig. 5. Low scores indicate high uncertainty and inhomogeneity of answers. The lowest VR2A value can be found in scenario 6.25mm sized cube with 103% clearance with the value of -31.2%. Highest values have been found for the biggest cube in 97% scenario: All participants recognized correctly, that the 200% cube does not fit in.

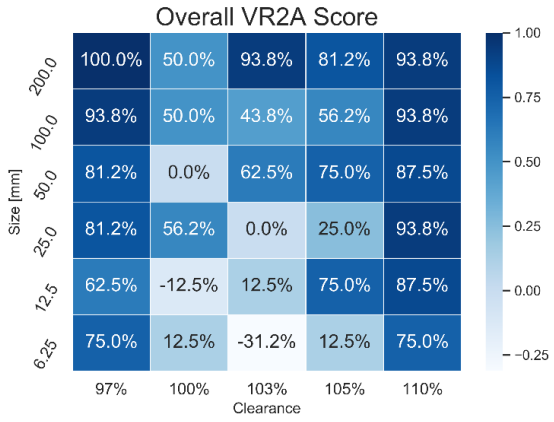

Fig. 5. Results of the VR assembly assessment score. Low values indicate high uncertainty or inhomogeneity of answers.

Plotting the mean VR2A scores over one of the two independent variables gives interesting insights on the assessment performance of the participants. Fig. 6 plots mean VR2A scores over the cube sizes in non-percentage values. One can clearly see that the VR2A positively correlates with the size of the cubes, as indicated by the 2nd polynomial regression. For 6.25mm cube size the mean score is only 28.75% whereas the 200 mm cube averages at 83.75%.

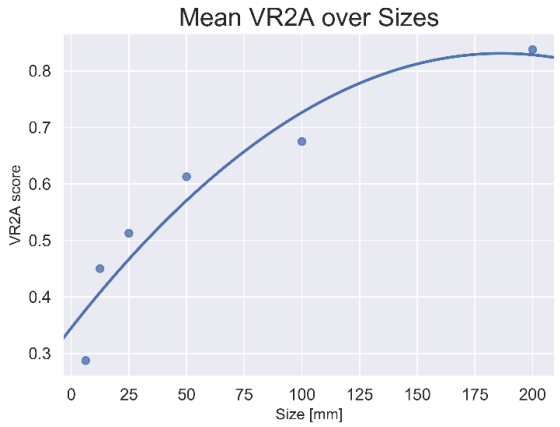

Fig. 6. Mean VR2A score over size scenarios with the respective 2nd polynomial regression.

Fig. 7 plots the mean VR2A results over the absolute clearance scenarios. Low mean VR2A scores can be found for the scenarios 100% (26.04%), 103% (30.21%) and 105% (54.1%). For both scenarios 97% and the 110% the scores are higher 82.29% and 88.54% respectively.

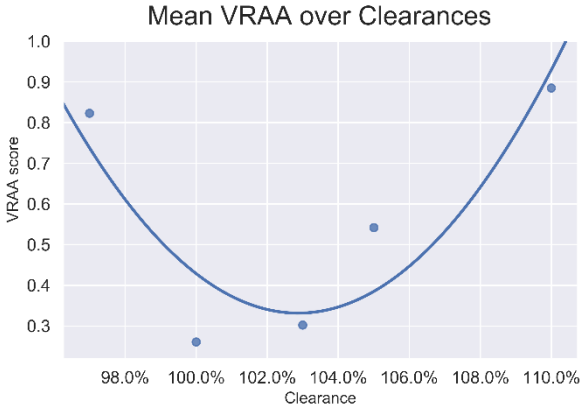

Fig. 7. Mean VR2A score over clearance scenarios with the respective 2nd polynomial regression.

4.6. Practical conclusions using VR2A insights

On a practical basis, VR2A benchmark helps production engineers to decide on how reliable their assessment has to be. They can define their own personal threshold, and therefore can easily derive, how small parts and their clearances may get. For example, for critical parts with paint coating they need high confidence in the assembly assessment. Therefore, he can set his personal VR2A threshold to 80% and can get a rough estimation, if the assembly part can be assessed correctly, e.g. 150mm or positive clearances should be bigger than 110% (see Fig. 6. and Fig. 7.). In contrast, for robust parts with large clearances, the VR2A threshold can be set lower to 50%.

4.7. Discussion

Results also indicate that collisions can be detected more easily compared to small clearances. The mean VR assembly score for 97% percent overlap performed a lot better than the 103% clearances. Even when comparing 97% overlap to 110% clearance values, they almost performed identically in terms of mean VR assembly score (see Fig. 5.). In general, the maximum uncertainty was expected at no tolerance scenarios (100% clearance), whereas the 103% clearance cavity led to the overall smallest VR2A values. Additional research has to find out, whether this entropy is highest for all assessments.

Results indicate, that even though people are encouraged to tell that "I can not assess it" is a valid answer, people still tend to give a judgment answer "Fits in" or "Does not fit in", even though there is no clearance at all.

Subjective feedback of the participants indicate potential reasons for this system's limitations: Human tremble and resolution of VR HMD: For the cube sizes XS (12.5mm) and XXS (6.25 mm) the vast majority of participants started hold the VR controller in both hands in order to reduce human tremble. Tracking accuracy still seems to be more stable than human tremble for small cube sizes. Therefore, in this evaluation, human tremble is currently the limiting factor for improving assessment performance (in comparison with HTC Vive precision and accuracy see also Niehorster et al. [21]). Additionally, for the smallest cube size (6.25 mm), all clearances are in sub-millimeter scale. Even though the participants could move their head as close as necessary to the discs, the VR HMD resolution was mentioned to be the subjectively limiting factor. On the other hand, four participants actively told the experimenter, that assessing large cubes is harder than small cubes due to necessary big head movement for assessing clearances. Even though in this evaluation collision avoidance has been disabled, VR2A still works with enabled collision avoidance. Further research has to be carried out using VR2A with collision detection.

5. Summary and Outlook

The "Virtual Reality Assembly Assessment" (VR2A) benchmark is a standardized, open source experiment design, to evaluate the overall VR system's assembly assessment performance and limitations. VR2A can be universally applied for different environments, simulation software and VR hardware devices. All readers are encouraged to assess their own assembly assessment system using the open source VR2A scene. Therewith, production engineers can gain practical insights on their next VR assembly assessment simulation. The evaluation showed, that VR2A is a reliable benchmark for quantifying the overall assessment performance and for revealing its limitations in assembly. By using VR2A in production validation of automotive and manufacturing industry, validation results get more reliable.

In future, there additional research will be carried out on the effects of more complex assembly part geometries, for example balls, stars, toruses, triangles or screw-shaped geometries, and other parameters, such as task-completion time. On purpose, these additional degrees of freedom are not considered in this evaluation. Furthermore, VR2A will be evaluated in broader studies towards its robustness using other VR technologies, simulation software and participant populations. As all researchers are encouraged to conduct VR2A themselves, third-party research will be integrated in later works.

References

[1] I. E. Sutherland, „A head-mounted three dimensional display", in *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, 1968, S. 757–764.

[2] J. Gugenheimer, E. Stemasov, J. Frommel, und E. Rukzio, „ShareVR: Enabling Co-Located Experiences for Virtual Reality between HMD and Non-HMD Users", in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, S. 4021–4033.

[3] H. Van Kerrebroeck, M. Brengman, und K. Willems, „Escaping the crowd: An experimental study on the impact of a Virtual Reality experience in a shopping mall", *Comput. Hum. Behav.*, Bd. 77, S. 437–450, Dez. 2017.

[4] K. E. Laver, S. George, S. Thomas, J. E. Deutsch, und M. Crotty, „Virtual reality for stroke rehabilitation", *Cochrane Database Syst. Rev.*, Nr. 2, 2015.

[5] A. Brown und T. Green, „Virtual reality: Low-cost tools and resources for the classroom", *TechTrends*, Bd. 60, Nr. 5, S. 517–519, 2016.

[6] Z. Merchant, E. T. Goetz, L. Cifuentes, W. Keeney-Kennicutt, und T. J. Davis, „Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis", *Comput. Educ.*, Bd. 70, S. 29–40, Jan. 2014.

[7] M. Otto, M. Prieur, P. Agethen, und E. Rukzio, „Dual Reality for Production Verification Workshops: A Comprehensive Set of Virtual Methods", *Procedia CIRP*, Bd. 44, S. 38–43, 2016.

[8] D. A. Bowman und R. P. McMahan, „Virtual Reality: How Much Immersion Is Enough?", *Computer*, Bd. 40, Nr. 7, S. 36–43, Juli 2007.

[9] P. Zimmermann, „Virtual reality aided design. A survey of the use of VR in automotive industry", in *Product Engineering*, Springer, 2008, S. 277–296.

[10] S. Ottosson, „Virtual reality in the product development process", *J. Eng. Des.*, Bd. 13, Nr. 2, S. 159–172, Juni 2002.

[11] G. Lawson, D. Salanitri, und B. Waterfield, „Future directions for the development of virtual reality within an automotive manufacturer", *Appl. Ergon.*, Bd. 53, S. 323–330, März 2016.

[12] L. P. Berg und J. M. Vance, „Industry use of virtual reality in product design and manufacturing: a survey", *Virtual Real.*, Bd. 21, Nr. 1, S. 1–17, März 2017.

[13] A. Gomes de Sá und G. Zachmann, „Virtual reality as a tool for verification of assembly and maintenance processes", *Comput. Graph.*, Bd. 23, Nr. 3, S. 389–403, Juni 1999.

[14] A. Dünser, K. Steinbügl, H. Kaufmann, und J. Glück, „Virtual and Augmented Reality As Spatial Ability Training Tools", in *Proceedings of the 7th ACM SIGCHI New Zealand Chapter's International Conference on Computer-human Interaction: Design Centered HCI*, New York, NY, USA, 2006, S. 125–132.

[15] A. Stork *u. a.*, „Enabling virtual assembly training in and beyond the automotive industry", in *2012 18th International Conference on Virtual Systems and Multimedia (VSMM)*, 2012, S. 347–352.

[16] J. C. Aurich, H. Hagen, D. Ostermayer, und M. Bertram, „VR-unterstützter KVP-Workshop", *Wt Online*, Nr. 95.

[17] C. Jennett *u. a.*, „Measuring and defining the experience of immersion in games", *Int. J. Hum.-Comput. Stud.*, Bd. 66, Nr. 9, S. 641–661, 2008.

[18] S. R. Ellis, „Presence of mind: A reaction to Thomas Sheridan's "further musings on the psychophysics of presence"", *Presence Teleoperators Virtual Environ.*, Bd. 5, Nr. 2, S. 247–259, 1996.

[19] B. G. Witmer und M. J. Singer, „Measuring presence in virtual environments: A presence questionnaire", *Presence*, Bd. 7, Nr. 3, S. 225–240, 1998.

[20] M. Funk, T. Kosch, S. W. Greenwald, und A. Schmidt, „A Benchmark for Interactive Augmented Reality Instructions for Assembly Tasks", in *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, New York, NY, USA, 2015, S. 253–257.

[21] D. C. Niehorster, L. Li, und M. Lappe, „The Accuracy and Precision of Position and Orientation Tracking in the HTC Vive Virtual Reality System for Scientific Research", *-Percept.*, Bd. 8, Nr. 3, S. 2041669517708205, Juni 2017.

Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP CMS 2019 (2019) 000–000

www.elsevier.com/locate/procedia

Peer-review under responsibility of the scientific committee of the 52nd CIRP Conference on Manufacturing Systems.