Use of Virtual Reality for the Evaluation of Human-Robot Interaction Systems in Complex Scenarios

Valeria Villani, Beatrice Capelli, Lorenzo Sabattini

Abstract— **Human-robot interaction has gained a lot of attention in recent years, since the use of robots can complement and improve human capabilities. To make such interaction smooth, proper interaction approaches are needed. Customarily these are tested in simplified scenarios and tame laboratory environment, since reproducing complex real use cases is often difficult. Achieved results are then not representative of actual interaction in reality and do not scale to complex scenarios. To overcome this issue, in this paper we consider the use of virtual reality as an alternative tool to assess HRI in those scenarios that are difficult to reproduce in reality. To this end, we compare the interaction experience for the same task, which is carried out in both virtual reality and real environment. To assess user's interaction in the two scenarios, we consider quantitative task related metrics, mental workload sustained, and subjective reporting. Results show that virtual reality allows to reproduce a faithful interaction experience and, thus, can be used to reliably validate human-robot interaction approaches in complex scenarios.**

I. Introduction

Recent progresses in robot control and safety systems have allowed the access, in the last few years, of robots in our daily life, to support humans in performing a great variety of tasks. Complex tasks, impractical for humans, such as large area coverage for environmental monitoring or search and rescue activities, can be delegated to robots and multi-robot systems, which multiply and distribute the capabilities of a single robot over a team with multiple different capabilities.

Moving from the initial scenario where robots were expected to operate autonomously, collaborative approaches, where humans and robots share their skills and work together towards the same goal, have gained a lot of attention. The presence of a human operator allows to combine human flexibility and cognitive and soft skills to the advantages of robots. This, however, implies that proper means need to be provided to let the human communicate with the robot in a smooth and intuitive way.

To validate such human-robot interaction (HRI), it is fundamental that, during usability tests, the real conditions in which the human is supposed to interact with the robot are faithfully reproduced. In particular, this applies to the interaction task, the environment and any situational constraints, such as reduced visibility, rush or psychological pressure. However, novel HRI approaches are customarily validated in tamed laboratory settings under several simplistic assumptions. This is due to practical limitations related to costs, risks and need for appropriate infrastructure and equipment to replicate realistic use cases representative of HRI. Achieved results, hence, might be misleading and unrealistic and do not scale well to complex scenarios, in which the stimuli the user is exposed to during real interaction vary notably and cannot be reliably reproduced. In particular, this is true when complex scenarios are considered, such as when big infrastructures are needed, as in, e.g., [1], or human-multi-robot systems are considered [2]. In this last case, experimental set-ups are typically limited in terms of number of robots composing the team and test scenario (consider, e.g., replicating a scenario of search and rescue over an area damaged and partially unreachable) [3].

The aim of this paper is to consider the use of virtual reality (VR) as a tool to overcome the above discussed issues. VR provides a totally immersive environment where the user's senses are under control of the system [4]. Some preliminary examples of the use of VR in robotics have been proposed considering, for example, training users to plan the optimal sequence of operations in assembly tasks [5]. Additionally, it has been proposed for the ergonomic assessment of such assembly tasks [6]–[9]. Robot programming by VR has also been proposed [10], since, with respect to traditional lead-through and off-line programming, it allows safe programming in a more intuitive manner and provides rapid feedback to the user.

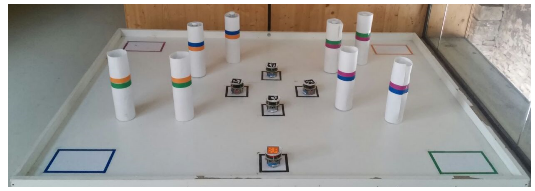
Technologies based on VR exploit human spatiality, which is humans' innate ability to act in physical space and interact with physical objects that are replicated in the virtual environment. Hence, they are well suited to study how the human interacts with a robotic system that is simulated in a totally immersive and realistic manner. By creating a virtual replica of reality, it is then possible to test the use of robotic systems in complex scenarios and get insights on how the interaction means is perceived by the user and how effective it is when the surrounding environment is included in the interaction.

Moving along these lines, it becomes fundamental to characterize the reliability of VR in this regard, in order to ascertain if and how the interaction experience can be accurately estimated in the virtual environment and, then, scaled to real scenarios. The aim of this paper is to assess the reliability of VR as an alternative to real (complex) scenarios in HRI tests. Specifically, considering that a difference in the perception of interaction is very likely to be found, since a virtual replica is compared to real world, we aim at investigating how remarkable is such a difference, to understand whether it should be preferred to test interaction in a real simplistic scenario or in a simulated realistic one. As

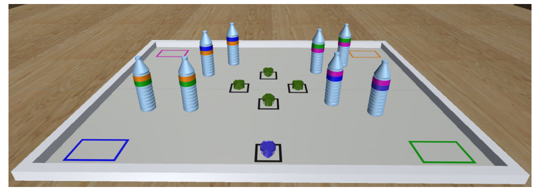
The authors are with Department of Sciences and Methods for Engineering (DISMI), University of Modena and Reggio Emilia, Reggio Emilia, Italy {name.surname}@unimore.it

a representative example of complex interaction task, in this paper we consider a human-multi-robot system interaction case. The same task is carried out in real and virtual environments and the interaction experience is qualitatively and quantitatively characterized in both conditions. Our aim is to provide a benchmark for the use of VR to validate HRI approaches. Results show that, although some unavoidable differences are found by test subjects, VR is a relevant tool for validating the interaction with robotic systems in complex scenarios.

The paper is organized as follows. The methodology considered for our comparative analysis is described in Sec. II. Results of the analysis are extensively presented in Sec. III, where quantitative and subjective assessment of interaction is reported. Finally, Sec. IV follows with a discussion of the achieved results and some concluding remarks.



(a) Real environment



(b) Virtual environment

Fig. 1. Experimental scenarios.

II. Methods

The aim of our experiments is to assess the reliability of VR as a replica of real environment (RE) for HRI. To this end, great effort was put in the identification of a dual scenario that is as much identical as possible in the two conditions, namely VR and RE. Moreover, the degree of similarity of the two was quantified considering both the performance in the interaction task and how the user perceives the interaction experience.

A. Experimental Scenarios

A dual experimental scenario was designed, consisting in a team of mobile robots moving in an arena populated with obstacles. Fig. 1 shows the experimental set-up in RE and VR. Specifically, in RE we considered 5 e-pucks, which are differential drive wheeled mobile robots [11]. In VR these were replicated exploiting the physics engine NVIDIA PhysX (NVIDIA, California, USA) to develop a realistic dynamic model of the robots. The arena measured 2.0 m $\times$ 1.5 m: the two environments were populated with the same objects in the same positions, as shown in Fig. 1. In particular, objects in the arena are the start positions for the robots (black small rectangles), the targets (colored big rectangles) and the gates (each of which is defined by a pair of poles, marked with the same colored stripes).

B. Experimental Task

Test participants were asked to perform the same task in the two experimental scenarios. During the session, 4 e-pucks were moving randomly in the arena, and the user was in charge of driving the other robot. The robot under user's control was marked with a different color than the others, as shown in the two panels of Fig. 1. A joypad was used to let the user command the robot's motion. The joypad was selected being a standard interaction device, in order to minimize any effect on the performance and perception of the test scenarios related to the interaction means.

The goal of the interaction task was to drive the robot through the gates to the colored rectangles at the corners of the arena. No predefined order was considered in the rectangles to reach; however, a matching between gates and rectangles was established. The user was allowed to reach a specific target only after passing through the two gates identified with the target's color (e.g., the robot could reach the blue target only after having passed through the gates with pink/blue and blue/orange poles). Moreover, all the rectangles had to be visited before reaching one of them once again. The total duration of the interaction task was 5 min. At least four hours occurred between the two experimental sessions for each subject.

C. Test Subjects

A total of $N = 30$ users (6 females, 24 males, age 27.4 ± 5.6 y.o.) were enrolled in the experiments. Participants are researchers working at our engineering department, in different research fields. All of them were completely new to the experimental task and goals.

According to counterbalanced measures design, users were exposed to the two experimental scenarios (VR and RE) in random order, to compensate for any learning effect on the task. For each experiment, the subject was asked to read a description of the experiment and sign a consent form. After signing the consent form, the experimental protocol was explained to the subject and demographic information was collected.

Information about previous acquaintance with joypad and VR was also collected. As regards the use of the joypad, 10 subjects reported that they were *very good* at using it, 9 *good*, 9 *fair* and, finally, 2 reported that they were *poor*. As regards any previous use of VR tools, almost half of the enrolled subjects (16 out of 30) reported that they had never used it before these experiments.

D. Metrics

In both experimental sessions the same metrics were registered. In particular, the effectiveness in the task was quantitatively computed, and the mental fatigue of the subject during interaction was estimated by physiological measurements. Additionally, participants were solicited to verbalize their impressions and perception of the interaction scenarios.

1) Performance Metrics: As quantitative assessment of performance, we considered as success count the number of rectangles the user was able to reach during the test. An error was detected when the robot commanded by the user touched the arena or the gates (fixed obstacles) or the other robots (moving obstacles).

2) Physiological Measurements: To account for differences in the ease of the interaction between the two experimental scenarios, we considered the subject's cognitive burden during the task. To this end, heart rate was measured, since heart rate variability (HRV) is known to be influenced by cognitive processing and mental workload [12]–[18]. HRV is the variation over time of the interval between consecutive heart beats and is measured from the RR series, which is the series of occurrence times of heart beats.

Customarily, short-term HRV analysis is performed on recording windows of 5 min duration [19]. Commonly used metrics are classified into time domain metrics, which can be statistical or geometrical, or frequency domain metrics that evaluate power, or ratios of power, in certain spectral bands [19]. In particular, it was found that, among others, one of the effects of mental workload on HRV is a decrease of the mean value of the RR series, denoted by $\overline{RR}$ in the following [14], [15].

In our experiments, subject's photoplethysmography-derived heart rate was recorded by a non-invasive wrist-worn device, as reported below in Subsec. II-E. The parameter $\overline{RR}$ was considered as an index of mental fatigue during the two experimental conditions [20].

In particular, according to [20], an increase in mental fatigue is detected if $\overline{RR}$ decreases, namely

$$\mathcal{S} = -\,\mathrm{sign}(\overline{RR}_k - \overline{RR}_{k-1}) \tag{1}$$

where $\mathcal{S}$ denotes the change in mental fatigue and $\overline{RR}_k$ is $\overline{RR}$ computed on the $k-$th time window. Specifically, considering two adjacent recording windows for the RR series, $\mathcal{S} = 1$ ($\mathcal{S} = -1$) means that an increase (decrease) in mental fatigue has occurred during the second recording window, with respect to the first one. If $\mathcal{S} = 0$, no change in mental fatigue has been detected.

3) Subjective Reporting: Subjective reporting was solicited to collect feedback about how each experimental scenario was perceived by test participants. At the end of each session (VR and RE), the participants were administered a questionnaire asking for subjective assessment. The questionnaire was designed ad hoc for the experiment, adapting that in [21]. Specifically, using a 5-point Likert scale, the subject was asked to give her/his opinion about the following features of the tested scenario: ease of use of the set-up, (perceived) efficiency in task completion, evoked anxiety, satisfaction in the use of the set-up (not for the task), learnability, complexity and ease of providing inputs to the robot. Moreover, after the second experimental trial, subjects were questioned about similarities and differences found in the two set-ups and were asked to express their preference for one of them.

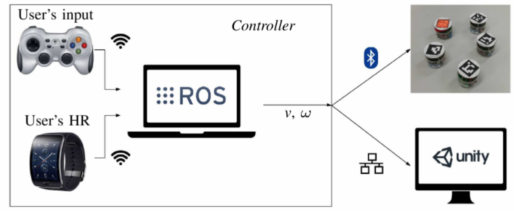


Fig. 2. Architecture used to implement the considered experimental scenarios in VR and RE.

In addition, during the interaction task, *think aloud* was prompted to the users to record specific issues or difficulties encountered when using the system. It consists in asking participants to verbalize their thoughts as they complete the task, without attempting to interpret their actions and words, in order to achieve objective understanding of their opinion about design.

E. Implementation

The above described scenarios and task were implemented considering the set-up shown in Fig. 2.

A common joypad for gaming was used in both scenarios for letting the user interact with the robots. Moreover, the user was wearing a Samsung Gear S smartwatch (Samsung, South Korea), which was used to record heart rate. Data from the smartwatch and the joypad were collected and processed on an external computer: the control inputs for the robots were computed from joypad output using ROS.

The scenario of VR was developed in Unity (Unity Technologies, California, USA), a free game development platform with a built-in physics and rendering engine. As mentioned above, the robots models were developed with NVIDIA PhysX. The Oculus Rift headset (Oculus VR, USA) was used for immersive VR.

III. Results

In the following we report the results of our experiments about HRI tests in VR and RE. Results are expressed in terms of quantitative measurement of performance, physiological parameters monitoring user's mental fatigue and subjective reporting.

A. Objective assessment of performance

Fig. 3 shows the quantitative assessment of performance, according to the performance metrics in Subsec. II-D.1. Specifically, in the left panel we report the histogram of targets reached in VR compared to the number of targets reached in RE. In the right panel, the same is applied to the number of errors. Thus, just to cite an example, a total of 5 users reached the same number of targets in VR and RE, whereas 7 subjects made the same number of errors in the two experimental sessions. In formal terms, the histograms in Fig. 3 are computed for the following quantities:

$$\begin{array}{r} \Delta_{\text{targets}}(i) = \text{targets}_{VR}(i) - \text{targets}_{RE}(i) \\ \Delta_{\text{errors}}(i) = \text{errors}_{VR}(i) - \text{errors}_{RE}(i) \end{array}\ ,\quad i = 1, \ldots, N \tag{2}$$

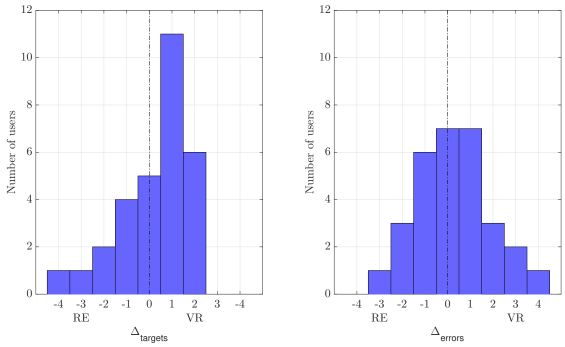


Fig. 3. Histograms of objective assessment of performance: count of reached targets (left) and errors (right). With reference to Eq. (2), values are expressed in terms of difference between VR and RE: positive values denote that the number of targets or errors was higher in VR than RE.

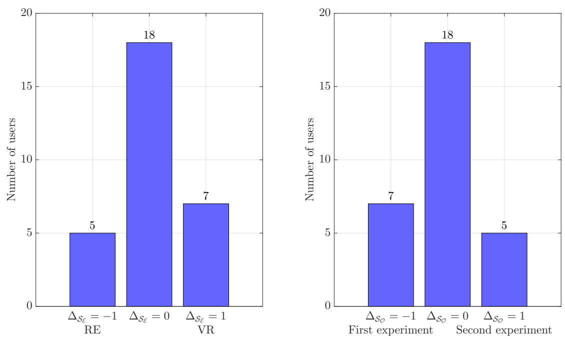


Fig. 4. Histogram of the difference in mental fatigue derived from HRV analysis. Left: comparison between VR and RE. Right: comparison between the first and the second experiment performed by each test subject.

where $N = 30$ is the number of enrolled subjects, $\text{targets}_{VR}(i)$, $\text{targets}_{RE}(i) \in \mathbb{N}$ is the number of targets reached by user i in VR and RE, respectively, and $\text{errors}_{VR}(i)$, $\text{errors}_{RE}(i) \in \mathbb{N}$ is the number of errors made by user i in VR and RE, respectively.

B. Physiological parameters

Fig. 4 summarizes the objective measurements of mental fatigue during the interaction task, measured according to the metrics introduced in Subsec. II-D.2. Specifically, the left panel refers to the comparison between VR and RE, and the right panel refers to the comparison between the first and second experiment performed by each test subject, irrespective of the experimental scenario. The histograms report the different reaction of test subjects to experiments in terms of difference in mental fatigue between the two conditions (VR and RE in the left panel, and first and second test session in the right panel). In formal terms, the histograms refer to the following quantities:

$$\Delta_{\mathcal{S}_\mathcal{E}}(i) = \mathcal{S}_{\text{VR}}(i) - \mathcal{S}_{\text{RE}}(i) \qquad \Delta_{\mathcal{S}_\mathcal{O}}(i) = \mathcal{S}_{\text{II}}(i) - \mathcal{S}_{\text{I}}(i) \tag{3}$$

where

$$\mathcal{S}_\omega(i) = -\operatorname{sign}(\overline{RR}_\omega(i) - \overline{RR}_{\omega_{\text{bl}}}(i)), \quad \omega \in \{\text{VR}, \text{RE}, \text{II}, \text{I}\} \quad i = 1, \dots, N \tag{4}$$

Considering the comparison between mental fatigue experienced with VR and RE, in Eq. (3) $\mathcal{S}_{\text{VR}}(i)$ and $\mathcal{S}_{\text{RE}}(i)$ denote the possible change in mental fatigue experienced by test subject i during the interaction tasks in VR and RE, respectively. Such change in mental fatigue is computed according to Eq. (4), namely as the difference between $\overline{RR}$ during the interaction task ($\overline{RR}_{\text{VR}}(i)$ and $\overline{RR}_{\text{RE}}(i)$) and $\overline{RR}$ computed at baseline, before the beginning of the test ($\overline{RR}_{\text{VR}_{\text{bl}}}(i)$ and $\overline{RR}_{\text{RE}_{\text{bl}}}(i)$). The same applies to the comparison related to the order of the experiments, i.e., $\Delta_{\mathcal{S}_\mathcal{O}}(i)$, whose histogram is reported in the right panel of Fig. 4.

From Fig. 4 it follows that 18 users experienced the same amount of mental fatigue when using VR or RE, 5 users were more stressed in the RE experiment, while 7 users were more stressed in the VR experiment. As regards the effect of the execution order on the mental fatigue, irrespective of the experimental scenario, 18 users experienced the same amount of mental fatigue during the first and the second experiment, 7 users were more stressed during the first experiment, while 5 users were more stressed during the second one.

C. Subjective reporting

Figures 5 and 6 show the results of the subjective reporting, collected by means of questionnaires as discussed in Subsec. II-D.3.

Specifically, in Fig. 5 each panel reports the histogram of agreement (given on a 5-point Likers scale) to each of the following statement:

(a) I feel comfortable using this set-up
(b) I am able to efficiently complete my task using this set-up
(c) I felt anxious during the experiment
(d) I enjoyed the task performed with this set-up
(e) I think I would like to use this set-up frequently
(f) I think most people would learn to use this set-up very quickly
(g) I found this set-up unnecessarily complex
(h) I like using the input control of this set-up

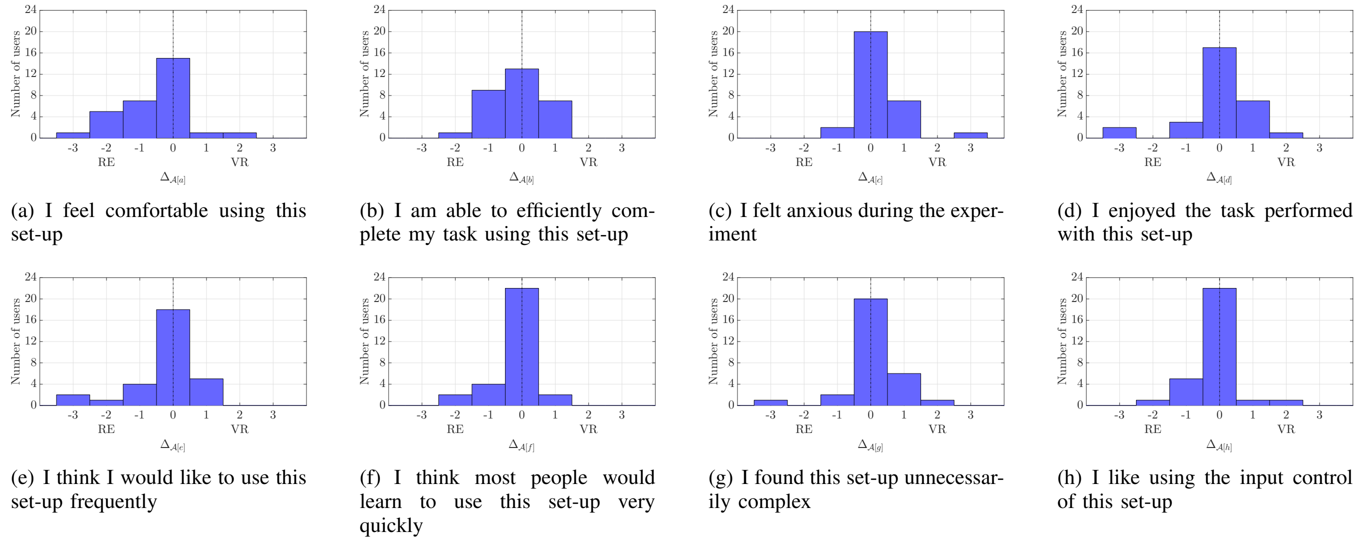
For each statement, the histogram reports the difference between the answer given for VR compared to the answer given for RE. Thus, as an example, considering panel (a), 15 users reported the same level of comfort in VR and RE. In formal terms, the histograms in Fig. 5 are computed for the following quantities:

$$\Delta_{\mathcal{A}[\kappa]}(i) = \mathcal{A}_{VR}[\kappa](i) - \mathcal{A}_{RE}[\kappa](i), \qquad i = 1, \dots, N \quad \kappa \in \{a, b, c, d, e, f, g, h\} \tag{5}$$

where $N = 30$ is the number of enrolled subjects, and $\mathcal{A}_{VR}[\kappa](i)$, $\mathcal{A}_{RE}[\kappa](i) \in [1, 5]$ are the answers given by user i to statement κ for VR and RE, respectively.

In addition, after that each test subject had performed both the experiments, she/he was asked the following questions, which solicited a comparison between the two scenarios:

(a) I have found significant differences between the two set-ups
(b) Which set-up did you prefer?



(a) I feel comfortable using this set-up

(b) I am able to efficiently complete my task using this set-up

(c) I felt anxious during the experiment

(d) I enjoyed the task performed with this set-up

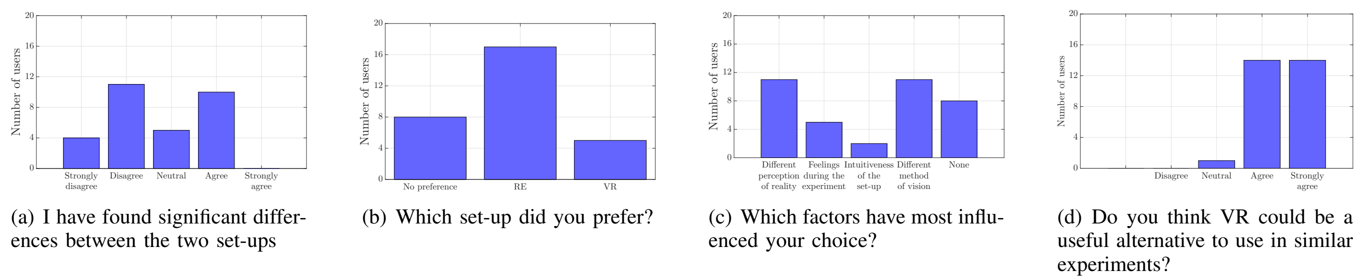
(e) I think I would like to use this set-up frequently

(f) I think most people would learn to use this set-up very quickly

(g) I found this set-up unnecessarily complex

(h) I like using the input control of this set-up

Fig. 5. Histograms of the results of subjective reporting: comparison of answers to questionnaire between VR and RE.



(a) I have found significant differences between the two set-ups

(b) Which set-up did you prefer?

(c) Which factors have most influenced your choice?

(d) Do you think VR could be a useful alternative to use in similar experiments?

Fig. 6. Answers to questions about the comparison between RE and VR.

(c) Which factors have most influenced your choice?
(d) Do you think VR could be a useful alternative to use in similar experiments?

For each of them, a set of possible replies was given as an option to study participants. Collected answers are reported in Fig. 6.

IV. Discussion

The results reported in Sec. III clearly show high levels of similarity between VR and RE: this similarity is noticeable considering all the proposed metrics, which provide a thorough assessment of the performance of the interaction system from all the relevant perspectives. From the experimental results, it is then possible to conclude that, from the user's perspective, the implementation of the same interaction task in VR or in RE does not cause large differences.

The task is, in fact, performed with comparable performances by the users in the two situations, as shown in Fig. 3. Furthermore, from the analysis of the *think aloud*, the small differences that can be measured in terms of performance are due to the fact that, in RE, users are more concerned with safety issues, such as damaging the robots: this indicates that, while VR is a good alternative for general HRI tasks, it might not be well suited for safety critical problems.

Considering the effect the task has on the user in terms of cognitive workload, Fig. 4 shows that the influence of the set-up is negligible. In fact, for the majority of the users there is no difference in the cognitive workload generated by the task in VR and in RE. Furthermore, the number of users that perceived a difference is (almost) equally distributed, considering those who were subject to a higher workload in VR with respect to RE. This is quite independent of possible learning or boredom effects, as confirmed by the right panel of Fig. 4. Indeed, the order of execution of the experimental tasks does not introduce major effects on the users' cognitive workload.

As regards the results of subjective reporting, Fig. 5 shows that the greatest majority of test subjects gave the same answer for the two scenarios (i.e., $\Delta_{\mathcal{A}[\kappa]} = 0$). This suggests that VR and RE solicit the same emotive response and interaction experience for most of the subjects. Thus, when it is required to assess how HRI tasks are perceived by users, VR can be used as a reliable alternative to experiments in RE. However, it should be noted that, with respect to Fig. 6, 10 subjects out of 30 reported significant differences between the two set-ups. Such differences were mainly due to the different method of vision, as reported also during *think aloud*. In particular, some reported that the perception of depth was notably reduced in VR and it was also reported that the resolution of the VR headset is quite low, thus making it more remarkable the difference in perception of VR and RE. Further, some test participants commented on the fact that no realistic audio was reproduced in VR scenario. Although explicit audio feedback was not relevant for the considered task, the lack of the sense of sound affected the realism of the scene experienced in virtual environment.

The result about the differences found between VR and RE, together with the fact that most of the subjects prefer RE to VR, is not surprising, given human innate ability to perceive and physically interact with objects and surrounding environment. It is also worthwhile mentioning that the headset used for VR is quite cumbersome and heavy[1]. In other words, it is definitely not transparent to the user who is wearing it and represents an unfavorable feature of VR over RE. In this regard, the answers reported in Fig. 6(d) are significant, since, despite the above mentioned disadvantages, almost all the test subjects reported that VR is a useful alternative for HRI experiments.

V. Conclusions

The aim of this paper was to assess the reliability of VR to assess HRI in complex scenarios that are difficult to reproduce in reality for evaluation purposes. In such cases VR can be a useful tool to replicate the interaction experience and validate novel interaction approaches.

To this end, we devised an interaction task and faithfully reproduced it both in RE and VR. To assess user's interaction in the two scenarios, we considered three different metrics: quantitative task related metrics, mental workload through physiological measurements, and subjective reporting about how individuals perceive the two interaction experiences. Reported results show that, on the one side, some differences between interaction in VR and RE exist, which is not surprising given the human innate spatiality and the still limited technological readiness of VR devices. Nevertheless, on the other side, it was found that a very similar interaction is experienced by test subjects, in terms of all the considered metrics.

These results confirm that VR is a reliable tool for the validation of HRI experiments. When testing in RE is difficult to implement, or for early assessment of prototypal interaction approaches, VR offers a comparable interaction experience.

Acknowledgement

The research is carried out within the "Smart and adaptive interfaces for INCLUSIVE work environment" project, funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement N723373.

References

[1] J. W. S. Chong, S. Ong, A. Y. Nee, and K. Youcef-Youmi, "Robot programming using augmented reality: An interactive method for planning collision-free paths," *Robotics and Computer-Integrated Manufacturing*, vol. 25, no. 3, pp. 689–701, 2009.

[2] A. Kolling, P. Walker, N. Chakraborty, K. Sycara, and M. Lewis, "Human interaction with robot swarms: A survey," *IEEE Trans. Human-Machine Systems*, vol. 46, no. 1, pp. 9–26, 2016.

[3] J. Cacace, A. Finzi, V. Lippiello, M. Furci, N. Mimmo, and L. Marconi, "A control architecture for multiple drones operated via multimodal interaction in search & rescue mission," in *IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2016, pp. 233–239.

[4] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, "Augmented reality: A class of displays on the reality-virtuality continuum," in *Photonics for industrial applications*. International Society for Optics and Photonics, 1995, pp. 282–292.

[5] A. Seth, J. M. Vance, and J. H. Oliver, "Virtual reality for assembly methods prototyping: a review," *Virtual reality*, vol. 15, no. 1, pp. 5–20, 2011.

[6] G. Chryssolouris, D. Mavrikios, D. Fragos, V. Karabatsou, and K. Alexopoulos, "A hybrid approach to the verification and analysis of assembly and maintenance processes using virtual reality and digital mannequin technologies," in *Virtual and augmented reality applications in manufacturing*. Springer, 2004, pp. 97–110.

[7] D. Mavrikios, V. Karabatsou, M. Pappas, K. Alexopoulos, and G. Chryssolouris, "An integrated VR-based simulation environment for the assessment of process and human factors in assembly," in *1st CIRP Int. Seminar on Assembly Systems*, 2006, pp. 301–305.

[8] M. Pappas, V. Karabatsou, D. Mavrikios, and G. Chryssolouris, "Ergonomic evaluation of virtual assembly tasks," in *Digital enterprise technology*. Springer, 2007, pp. 511–518.

[9] C. Pontonnier, A. Samani, M. Badawi, P. Madeleine, and G. Dumont, "Assessing the ability of a VR-based assembly task simulation to evaluate physical risk factors," *IEEE Trans. Visualization and Computer Graphics*, vol. 20, no. 5, pp. 664–674, 2014.

[10] J. Aleotti, S. Caselli, and M. Reggiani, "Leveraging on a virtual environment for robot programming by demonstration," *Robotics and Autonomous Systems*, vol. 47, no. 2, pp. 153–161, 2004.

[11] F. Mondada, M. Bonani, X. Raemy, J. Pugh, C. Cianci, A. Klaptocz, S. Magnenat, J. C. Zufferey, D. Floreano, and A. Martinoli, "The e-puck, a robot designed for education in engineering." in *Proceedings of the 9th Conference on Autonomous Robot Systems and Competitions*, 2009, pp. 59–65.

[12] L. Bernardi, J. Wdowczyk-Szulc, C. Valenti, S. Castoldi, C. Passino, G. Spadacini, and P. Sleight, "Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability," *J. Am. Coll. Cardiol.*, vol. 35, no. 6, pp. 1462–1469, 2000.

[13] U. R. Acharya, K. P. Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: a review," *Med. Biol. Eng. Comput.*, vol. 44, no. 12, pp. 1031–1051, 2006.

[14] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia, "Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis," *Biomedical Signal Process. Control*, vol. 18, pp. 370–377, 2015.

[15] R. Castaldo, W. Xu, P. Melillo, L. Pecchia, L. Santamaria, and C. James, "Detection of mental stress due to oral academic examination via ultra-short-term HRV analysis," in *Proc. IEEE 38th Annu. Int. Conf. Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 3805–3808.

[16] A. Hoover, A. Singh, S. Fishel-Brown, and E. Muth, "Real-time detection of workload changes using heart rate variability," *Biomedical Signal Process. Control*, vol. 7, no. 4, pp. 333–341, 2012.

[17] P. Rani, J. Sims, R. Brackin, and N. Sarkar, "Online stress detection using psychophysiological signals for implicit human-robot cooperation," *Robotica*, vol. 20, no. 06, pp. 673–685, 2002.

[18] J. F. Thayer, F. Åhs, M. Fredrikson, J. J. Sollers, and T. D. Wager, "A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health," *Neuroscience & Biobehavioral Rev.*, vol. 36, no. 2, pp. 747–756, 2012.

[19] Task Force of The European Society of Cardiology & The North American Society of Pacing and Electrophysiology, "Heart rate variability - Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, vol. 17, pp. 354–381, 1996.

[20] V. Villani, L. Sabattini, C. Secchi, and C. Fantuzzi, "A framework for affect-based natural human-robot interaction," in *27th IEEE Int. Symp. Robot and Human Interactive Communication (RO-MAN)*, IEEE, Ed., 2018.

[21] T. S. Tullis and J. N. Stetson, "A comparison of questionnaires for assessing website usability," in *Usability professional association conference*, 2004, pp. 1–12.

[1] However, the same applies to the other existing devices for VR.