Comparing performance and situation awareness in USAR unit tasks in a virtual and real environment

Corine H.G. Horsch
Delft University of Technology
C.H.G.Horsch@tudelft.nl

Nanja J. J. M. Smets
TNO/ Delft University of Technology
Nanja.Smets@tno.nl

Mark A. Neerincx
TNO/ Delft University of Technology
Mark.Neerincx@tno.nl

Raymond H. Cuijpers
Eindhoven University of Technology
R.H.Cuijpers@tue.nl

ABSTRACT

A convenient way to test Urban Search And Rescue (USAR) robots would be in virtual environments (VEs). Evaluations in VEs are generally accepted as alternative for real scenarios. There are obvious differences between operation in a real and virtual environment. Nonetheless, the current experiment showed no significant differences in situation awareness (SA) and performance during several elementary tasks (e.g. slalom) between a virtual world and a previous experiment in reality (Mioch, Smets, & Neerincx, 2012). Only small dependencies between the unit tasks were found. The effect of individual differences (like gender, km driven per year, and gaming experience), were significant for certain elementary tasks. Testing robots in virtual environments could still be useful even if differences between VE and reality exist, since comparisons of different conditions in VE seems to have the same results as the same comparison in the field (Bishop & Rohrmann, 2003; Van Diggelen, Looije, Mioch, Neerincx, & Smets, 2012).

Keywords

Test USAR robots, comparison virtual reality and reality, experiment, elementary task, situation awareness, performance

INTRODUCTION

When man-made structures collapse, Urban Search and Rescue teams (USAR) come to search for victims trapped under the rubble. Such an environment is dangerous for humans since collapses may still occur, gas leaks and exposed electric cables also pose a serious risk (Murphy, 2003). Robots can especially be helpful in mapping parts of the disaster area that is restricted for humans. But before a robot can be deployed in a real USAR situation, it is useful to evaluate the robot. Scenes can be set up to simulate real situations. However, live scenarios are often costly because of high personnel and time demands for preparation, execution, analysis and debriefing (Jenvald & Morin, 2004). Instead robots could also be evaluated in a virtual environment (VE). Advantage of evaluations in VEs is that the robot does not need to be completely finished before the evaluation can be performed, the environment is fully controlled and experiments can be exactly replicated (Fox, Arena, & Bailenson, 2009). VEs have some negative features as well, like insufficient depth perception and lack of haptic feedback (Wang, Hwang, Fang, Sheu, Leong, & Ma, 2011). Despite the drawbacks of VEs, evaluations in VEs are a good alternative to real scenarios since it is possible to represent a natural environment with higher realism than in a laboratory setting. But how reliable can a virtual experiment represent an experiment in an analogue setting (such as exercises in training facilities)?

Mioch and colleagues (2012) developed unit tasks for USAR tasks and situations and investigated if the unit tasks were able to predict performance in a training scenario. The unit tasks are part of the usage centered evaluation methodology by Mioch and colleagues (2012), where a relative simple set of abstract tasks are used to assess basic aspects of human-robot collaboration. The unit tasks resemble basic functionality of human-robot collaboration in envisioned USAR scenarios and consisted of: a slalom, drive through a narrow hallway, stop as close to a wall as possible, and find five signs in a room (detect objects task; see Figure 1). The training scenario represented a car accident in a tunnel and the participants were asked to gather information about the situation (Mioch, Smets, & Neerincx, 2012). During the unit tasks and the scenario the performance of the human-robot system and the situation awareness (SA) of the operator was measured. SA is the perception of the elements in the environment (comprehension, meaning and projection) and important to USAR tasks (Endsley, 1995). The results of the experiment showed that the difference in collisions (SA measure) in the detect objects task explained 40% of the variance in the scenario. For performance a correlation was found between the number of collisions in the detect objects task and the scenario. Furthermore, individual differences turned out to influence performance and SA both in the unit tasks as well as in the scenario. In summary, the detect objects task is most capable of predicting SA in the scenario. However, individual characteristics may not be ignored.

Figure 1: a) slalom task, b) narrow hallway and stop before collision task, c) detect objects task, d) operator controlling robot with game controller and head mounted display (Mioch et. al, 2012).

To investigate how reliable a virtual experiment can represent an experiment in reality, this paper evaluates to what extent the unit tasks of Mioch and colleagues (2012) in a virtual reality can represent the results of the unit tasks in a physical experiment. It is important to know which variables will differ in the virtual versus real environment and in what direction they differ. We expect there to be a difference in performance and SA of the human-robot system between the real environment and the virtual environment.

Unit tasks can be developed at certain levels while they still are simple atomic tasks. The higher level tests thereby depend on lower level tests (Van Diggelen, Looije, Mioch, Neerincx, & Smets, 2012). The slalom, the narrow hallway, and the stop before collision task are of a lower level than the detect objects task. The performance of the latter task partially depends on the performance of the low level tasks. The lower level unit tasks should be able to explain some of the variance in the detect objects task. So, the performance and SA in the detect objects task depends on the performance and SA in the low level unit tasks.

The experiment of Mioch and colleagues (2012) showed that the age of the operator influenced the performance and SA in the scenario and in the unit tasks (Mioch, Smets, & Neerincx, 2012). We therefore expect that age will also influence the performance and SA in a virtual environment. Moreover, we also expect that gaming experience will effect performance and SA, because performing tasks in a virtual environment will probably feel as a game.

METHOD

The experimental design is a within subject design. The order of the tasks was counterbalanced across subjects. In this experiment 12 persons participated, from which six were male and six were female. The participants were on average 25 years old and seven of them possessed a driver's license for 5,7 years on average. They had no USAR operations experience and none to minimal robot operating experience. They did play first person shooter games (varying from once per month to daily).

Materials

For the virtual experiment a laptop was used on which Unreal Development Kit and UsarSim were installed. UsarSim is a high fidelity robotic environment developed as a simulation of robots for USAR operations (such as reconnaissance). It uses Epic Games' Unreal Engine 3 to provide a high fidelity simulator. The P3AT robot was used to perform the tasks and was controlled with a standard game controller. The robot in Mioch's experiment was also controlled with a game controller, but did have a head mounted display (Figure 1d).

Procedure and tasks

The participants filled out the demographical questions. Then the experiment began with a training session. The participants were asked to get familiar with the controls of the robot for 15 minutes while checking out some objects. After the training session the participants were ready for the unit tasks. Each unit task was explained beforehand. Screenshots of the virtual unit tasks are depicted in figure 2. The slalom consisted of five traffic cones (distance 118 centimeters). The narrow hallway consisted of a 76 centimeters wide hall in a L-shape with little space to maneuver in the corner. At the end of the hallway, the robot has to stop as close to the wall as possible without touching the wall. The last unit task is the detect objects task. During this task the robot is placed at the ‘entrance’ of a room and has to detect and classify as many objects (green and pink signs) as possible in two minutes. In total five signs were placed in the room at three different height levels: low, medium high, and high, respectively 4 cm, 80 cm, and 170 cm from the ground. After the task the participants filled in a questionnaire to measure SA. The observer evaluated the performance on the task. Finally the participants filled out an end questionnaire that measured overall SA.

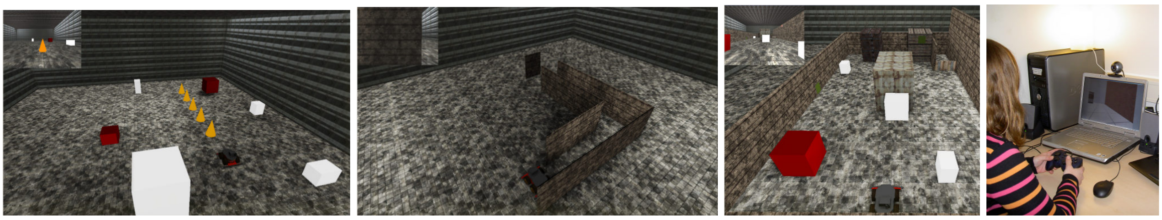

Figure 2: a) slalom task, b) narrow hallway and stop before collision task, c) detect objects task, d) operator with game controller and laptop.

Measurements

The following questionnaires and measurements were taken . General questionnaire (Gender, age, driving and gaming experience, driver’s license and km per year). The objective performance was measured by time to finish, number of collisions, and number of found objects. The subjective performance was rated by an observer, driving efficiency (5-point scale) and in the hallway task the observer judged if the robot went in a straight line or not. Situation awareness (SA) consisted of an SA difference measure, the difference between the number of collisions the operator thought he made during the task and the number of actual collisions (Mioch, Smets, & Neerincx, 2012). Second, questions for assessing SA (Mioch, Smets, & Neerincx, 2012). Third, after the slalom task SA was measured by asking the participants how many colored blocks there were in the environment, after the detect objects task they were asked what the color of the majority of the blocks was. And finally in the end questionnaire, this covered overall SA and a small evaluation of the robot. Part of the SA questions consisted of the validated SPASA (Short Post-Assessment of Situation Awareness) questionnaire (Gatsoulis, Virk, & Dehghani-Sanij, 2010).

RESULTS

During the slalom task participants in the virtual experiment bumped on average more into cones (M=5.1, SE=0.73), than participants in the real experiment (M=3.0, SE=0.41, $t(16.7)=-2.48$, $p=.024$, $r=.5$), see Figure 3. The other performance measures were not significantly different between the real and virtual environment in any task. Nonetheless, a medium effect size ($r=.3$) was found in the detect objects tasks where participants in the virtual environment bumped more into the walls (M=1.33, SE=.26) than in the real environment (M=.67, SE=.44).

None of the SA measures were significantly different between the real and virtual environment. However, a medium effect size ($r=.34$) was found in the mismatch of distance (difference between the real distance to the wall and the estimated distance to the wall) between the real and virtual experiment in the collision task. Despite the non significant difference, the mismatch was bigger in the virtual environment (M=7.37, SE=5.35) than in the real environment (M=-1.00, SE= 4.39). A negative value means that the participant estimated that he was closer to the wall than he actually was.

The end questionnaire regarding overall SA showed that on average, participants reported a better overall SA in the real experiment (M=4.04, SE=.25), than participants in the virtual experiment (M=3.29, SE=.24, $t(18)=2.13$, $p=.047$, $r=.45$). The overall SA questionnaire (from Mioch, Smets, & Neerincx, 2012) did not correlate with the SPASA questions (from Gatsoulis, Virk, & Dehghani-Sanij, 2010) ($r=.467$, ns).

The second region of interests was the relationship between the low level unit tasks (slalom, hallway, and collision) and the high level unit task (detect objects task). A multiple regression analysis showed that the SA (difference in collision) in the slalom accounts for 38,4% of the variation in performance (signs found) in the detect objects task (Sig F Change = .032). The performance measures in the low level unit tasks were not able to predict the performance in the high level unit task better than just taking the mean of the performance in the high level task. Furthermore, the results showed that the amount of game experience of the participant accounts for 39,9% (Sig. F Change = .028) of the variation in number of collisions (performance) and for 36,4% of the variation in efficiency (performance) in the slalom task. Whereas the game experience is an important predictor for the performance during the slalom task, the number of km driven is a significant predictor for the performance in the hallway task. The number of km driven by the participant accounts for 34,1% of the variation in time (performance) needed to finish the hallway task (Sig. F Change = .046). The number of km driven by the participant also accounts for 51,7% of the variation in number of collisions (performance) in the hallway task (Sig. F Change = .008). Lastly, the number of km driven by the participant accounts for 40,3% of the variation in efficiency (performance) in the hallway task (Sig. F Change = .026). In short the performance in the slalom can be predicted well by game experience, while the hallway performance can be predicted by the kilometers driven.

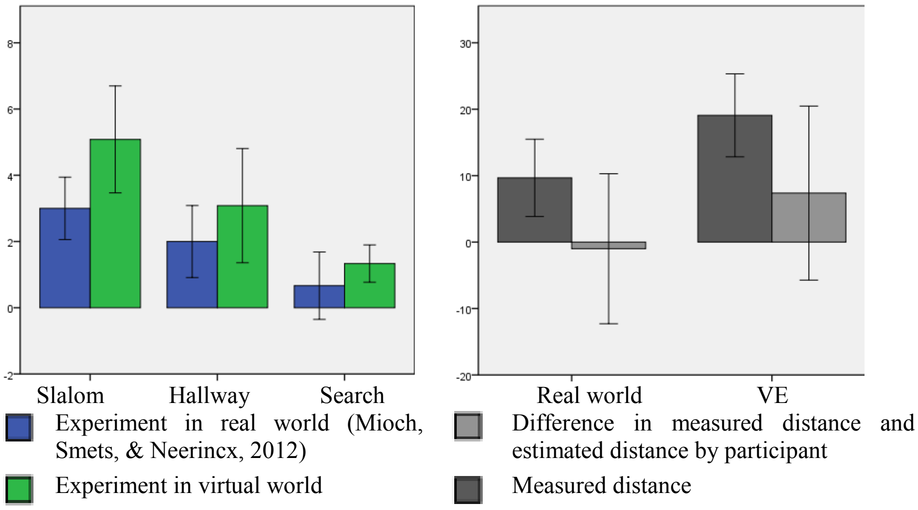

Figure 3: left) number of collisions; right) distance to the wall in the stop before collision task. Error bars: 95% CI.

The gender of the participant accounts for 47,4% of the variation in the mismatch of the number of blocks (difference between the real number of blocks and the reported number of blocks) in the slalom (Sig. F Change = .013). Males had a bigger mismatch (M=2, SE=.37) than females (M=0.5, SE=.34, t(10)=3.0, p=.013).

DISCUSSION

In general, no significant differences between the previous (Mioch, Smets, & Neerincx, 2012) and current, virtual experiment were found in SA or performance. However, participants reported a better overall SA afterwards in Mioch's real experiment than participants in this virtual experiment. Hypothesis 1 stated that there are differences in SA and performance between the real and virtual experiment per unit task. This hypothesis cannot be confirmed nor rejected. The only difference found is that there are more collisions in the slalom in the VE than in the real experiment. This may be due to the fact that in the virtual experiment participants were unaware of the shape of the robot.

Although the brief examination of the unit tasks showed a fundamental difference between the detect objects task and the other unit tasks, this claim is not supported by the results (H2). The performance in the low level unit tasks could not predict the performance in the detect objects task. And the SA in the low level unit tasks could also not predict the SA in the detect objects task. However, the SA in the slalom could partially predict the performance in the detect objects task. SA has broadly been acknowledged as a crucial factor for successful performance. The result of this experiment gives a little support for this idea. However, other researchers say that "*it is important to note that SA is distinct from performance, SA can be maintained even when there is no performance to be observed*" (Wickens, Lee, Liu, & Gordon Becker, 2004). Furthermore, Endsley herself critized explicit SA measures because even if SA is directly measured, the measure does not provide any clues about how participants will perform (Endsley, 1995). In summary, there is no clear consensus on the relation between SA and task performance and the discussion remains open after this experiment.

The last two hypotheses considered the effects of individual differences (H3). The results did not show that age is significantly related to performance or SA, unlike the hypothesis stated. This could be caused by the small range in age (min=22, max=30). However, the results did show that the number of km driven affects the performance on the hallway task. A correlation is found between age and how many km a person drives per year (r=.67, p=.017). So indirectly age is related to the performance on the hallway task.

Furthermore, an unexpected effect of gender was found, which indicates that women had a better SA (number of blocks) in the slalom task. The number of blocks is fairly similar to the question about how many traffic cones were present in the real experiment. In that question however, no differences between males and females were found. The results can be explained by the selectivity model (Meyers-Levy, 1989). According to the selectivity model, males tend to rely on subsets of available cues for achieving a task. In the slalom they probably focused on the traffic cones, since the traffic cones are highly relevant for the task. Females, on the other hand, process information more comprehensively. In this case women probably also processed information about less relevant objects, like the blocks, besides processing information about the traffic cones.

CONCLUSION AND FUTURE WORK

Previous research has shown that the results of an experiment in reality and simulation differ on certain variables. The result of this experiment did not show those differences. In the USAR domain it could be the case that a virtual environment resembles reality enough, so that no differences are found. When an operator controls the robot in a real USAR situation he is physically not in the same room and has to rely on sensory input from the robot, which is similar to operation in VR. Even if differences between VE and reality exist, testing robots in virtual environments could still be useful, since comparisons of different conditions in VE seems to have the same results as the identical comparison in the field (Bishop & Rohrmann, 2003; Van Diggelen, Looije, Mioch, Neerincx, & Smets, 2012). Furthermore, it is shown that individual differences, like gender, age, possessing a driving license, driving experience, and gaming experience influence situation awareness and performance. In the future it would still be valuable to keep comparing real versus virtual operation of robots. Since it could make for a cost effective, safe and controllable way of testing early in the development process of robots. But also training for USAR robots and its operators is a good possibility of a future application of VE.

ACKNOWLEDGMENTS

This paper describes research done under the EU-FP7 ICT247870 NIFTi project (http://www.nifti.eu).

REFERENCES

1. Bishop, I., & Rohrmann, B. (2003). Subjective responses to simulated and real environments: a comparison. *Landscape and urban planning*, 261-277.
2. Van Diggelen, J, Looije, R., Mioch, T., Neerincx, M.A., & Smets N.J.J.M. (2012). A Usage-Centered Evaluation Methodology for Unmanned Ground Vehicles. The Fifth International Conference on Advances in Computer-Human Interactions (ACHI 2012, January 30 - February 4). Xpert Publishing Services (XPS).
3. Endsley, M. (1995). Measurement of situation awareness in dynamic systems. *Human factors* , 65-84.
4. Fox, J., Arena, D., & Bailenson, J. (2009). Virtual reality. A survival guide for the social scientist. *Journal of Media Psychology*, 95-113.
5. Gatsoulis, Y., Virk, G., & Dehghani-Sanij, A. (2010). On the measurement of situation awareness for effective human-robot interaction in teleoperated systems. *Journal of cognitive engineering and decision making*, 69-98.
6. Jenvald, J., & Morin, M. (2004). Simulation-supported live training for emergency respons in hazardous environments. *Simulation & Gaming*, 363-377.
7. Meyers-Levy, J. (1989). Gender differences in information processing: A selectivity interpretation, in Cafferata, P.,& Tybout, A. (Eds.), *Cognitive and Affective Responses to Advertising*, Lexington Books, Lexington, MA, pp. 219-260.
8. Mioch, T, Smets, N.J.J.M. & Neerincx, M.A. (2012). Predicting Performance and Situation Awareness of Robot Operators in Complex Situations by Unit Task Tests. *The Fifth International Conference on Advances in Computer-Human Interactions* (ACHI 2012, January 30 - February 4). Xpert Publishing Services (XPS).
9. Murphy, R. (2003). Human Robot Interaction in rescue Robotics. *Invited paper*.
10. Wang, C., Hwang, W., Fang, J., Sheu, C., Leong, I., & Ma, H. (2011). Comparison of virtual reality versus physical reality on movement characteristics of persons with parkinson's disease: effects of moving targets. *Arch Phys Med Rehabil*, 1238-1245.
11. Wickens, C., Lee, J., Liu, Y., & Gordon Becker, S. (2004). *An introduction to human factors engineering.* New Jersey: Pearson Education