Heuristic Evaluation Adaptation to Virtual Reality Systems for Visualisation of Robot Sensor Data in Nuclear Decommissioning

Paul Bremner, Manuel Giuliani, Alan Winfield, Anthony G. Pipe
paul.bremner@brl.ac.uk,manuel.giuliani@brl.ac.uk,alan.winfield@brl.ac.uk,tony.pipe@brl.ac.uk
Bristol Robotics Laboratory, University of the West of England
Bristol

ABSTRACT

We are currently investigating the use of VR for robot teleoperation in nuclear decommissioning, particularly for the presentation of robot sensor data using a virtual reality (VR) environment. Here we present an adaptation of the well known heuristic evaluation methodology for use to evaluate the user interface for such VR information-visualisation systems. The aim of which is design improvements prior to evaluation by a full user study. We present a heuristic set and application methodology for VR information-visualisation systems, and demonstrate its application with a case study. Through analysis of this case study we determine that heuristic evaluation is of use in this context, and produce some general guidelines to its application in systems of this type.

CCS CONCEPTS

• **Human-centered computing** → **Visualization design and evaluation methods**; **User interface design**; • **Computer systems organization** → Robotics.

KEYWORDS

Virtual Reality, Robotics, Heuristic Evaluation, Nuclear Decommissioning

ACM Reference Format:
Paul Bremner, Manuel Giuliani, Alan Winfield, Anthony G. Pipe. 2020. Heuristic Evaluation Adaptation to Virtual Reality Systems for Visualisation of Robot Sensor Data in Nuclear Decommissioning. In *VAM-HRI '20: International Workshop on Virtual, Augmented and Mixed Reality for Human-Robot Interaction, March 23, 2020, Cambridge.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The decommissioning challenges facing the nuclear industry are immense. A key domain of this decommissioning process are legacy nuclear facilities, these have often been closed off for many decades and are frequently poorly understood, with inventory records and design drawings either incomplete, erroneous or unavailable [14]. Decommissioning such facilities requires characterisation of these environments, a process that involves gathering data on the structures as well as environmental factors such as radiation, temperature etc. Currently this is a process undertaken by skilled workers for whom working in such conditions is stressful, physically demanding, awkward because of the need to wear protective clothing, and a potential health risk. Robotics presents a potential alternative means of carrying out this characterisation process.

Figure 1: Two robots from the multi-agent SLAM system in a test environment.

We are currently working on a characterisation solution where a heterogeneous swarm of robots performs multi-agent simultaneous localization and mapping (SLAM), constructing a map of the structures as well as gathering environmental data (Figure 1). The safety case of nuclear decommissioning requires there to be a human in the loop directing operations of the robots and interpreting the data that they generate to perform the needed characterisation process. Hence, the gathered data needs to be presented to such an operator in a way that is intuitive to understand, and facilitates characterisation and robot tele-operation. We are investigating the use of virtual reality (VR) for this purpose.

In our VR based system the data gathered by the multi-robot system is used to construct an environment that can be navigated in VR. We are investigating VR for this purpose as an immersive interface has potential advantages over a standard 'desktop' interface: "...it leads to a demonstrably better perception of a datascape geometry, more intuitive data understanding, and a better retention of the perceived relationships in the data." [4]. Indeed, the highly spatial nature of the data, and the subsequent tasks that must be conducted lend themselves particularly well to VR. Intuitiveness of the interface, and ease of data understanding are of particular importance in the nuclear decommissioning context due to the high

---

cost of operator training time, as well as high cognitive load and stress of the decommissioning process.

As with any VR or information visualisation (InfoVis) application design specifications are context specific, and the impact of design choices cannot be fully understood a priori. Thus, a common approach to evaluating the impact of design choices is by way of user studies. The utility of such user studies can be improved by means of expert evaluation and subsequent refinement of the system prior to user study deployment [11]. An approach which we have followed for our system under development. A hybrid approach similar to this has been found to improve identification of usability issues [11]

One means of conducting expert reviews is a process called heuristic evaluation (HE). The process of HE involves having a small set of expert evaluators examine a system and judge its compliance with a priori identified usability principles (heuristics) [8]. The utility of expert reviews by HE has been demonstrated for several systems including VR [13] and InfoVis [5]. However, upon examination of the heuristics proposed for evaluation of VR and InfoVis systems it is clear that not all of them are appropriate in our use case; further, a combined list of all proposed heuristics would be unwieldy for use in system evaluation. Hence, we present a heuristic set for use in VR Infovis comprised of applicable heuristics from the literature. We then detail its application to our nuclear decommissioning use case to demonstrate the application procedure, results analysis, and evaluate the utility of HE and our proposed heuristics for our use case. In doing so we also demonstrate the potential of our design approach for VR InfoVis.

2 BACKGROUND AND RELATED WORK

Heuristic evaluation (HE) is a method from usability engineering whereby experts in user interface (UI) design identify problems in a given user interface; identified problems can then be attended to as part of an iterative design process [8, 9]. The process of HE involves expert evaluators performing a set of tasks as might be undertaken by a typical user, and identifying problems arising as a result of deviations from defined usability principles (heuristics). Identified issues are attributed to the heuristic from which deviation has been observed, either by evaluators in the case of written evaluator produced reports, or an observer in the case of verbal problem identification. The observations of multiple evaluators are then compiled and analysed to produce design recommendations. We have opted to use verbal problem identification as it fits better with the VR use case.

It has been shown that the majority of usability issues can be identified with a small number of evaluators, typically 3-5 [8]. The need for only a few evaluators, results in low evaluation time, and hence low cost relative to typical user studies; thus, it is considered a very efficient method for use in iterative design [8]. However, it is not without issue, it has been shown that for some use cases expert review performs significantly worse at problem identification than user testing (UT) e.g. for testing a virtual environment [2]. Hence, we view it as a precursor to UT rather than a replacement.

The original set of heuristics identified by Nielson are intended for general application across many HCI contexts [8, 9]. A need for more specifically applicable heuristics has been demonstrated across many domains of HCI, and they have been found to better identify domain specific issues (see [6] for a review). In the review article of Hvannberg et al. [7] (examining HE for virtual environments) they demonstrate the utility of adapting previously identified heuristics to a specific use case, and taking a 'patchwork' approach to utilising heuristics from multiple sources; one caveat to this approach is to not have a too large set of heuristics as its application becomes problematic. Here we follow this 'patchwork' heuristic set composition approach taking applicable heuristics from sets composed for VR and InfoVis evaluation.

A number of previous works have considered the application of HE to VR. In their seminal work Sutcliffe and Gault propose a heuristic set specifically for VR [13]. Their heurisitc set is based on a principled adaptation of Nielsen's original heuristics [8**?**]; they demonstrate its utility through evaluation of a VR chess game. A number of other authors have used their heuristics in whole or in part, adapted for their particular use case (see [7] for a review).

Munoz and Chalegre propose a set of heuristics for use in evaluating virtual world applications (e.g., Second Life), their set has many similarities to those in [13]. Similarly, Desurvire and Kreminski propose a set of heuristics for VR game design [3]. More recent work has iterated on these heuristic sets, refining them for particular use cases [10, 12]. These works demonstrate the utility of modifying previously established VR heuristic sets for a particular use case. As the most generalized set of VR heuristics we use those proposed in [13] as a basis for our work. The other sets reported on here are specialised to use cases differing significantly from ours. It is also important to note that previous applications of HE to VR have utilised written reports produced by evaluators after task completion (as this requires leaving the VR environment), rather than during task verbal reporting as we use here.

3 METHODOLOGY

In the following section we present the selected heuristics, their descriptions as provided in the source paper, and the reasons for their selection. Additionally, we describe the application procedure, adapted from the methodology proposed by Nielsen [9], for their use in VR-Infovis evaluation.

3.1 Heuristic selection

In order to compile a suitable set of heuristics for evaluating VR-InfoVis systems we selected the most applicable heuristics from validated sets used for evaluating VR [13] and InfoVis [5]. Sutcliffe and Gault highlight several example studies where it was found that the different interaction modality of VR applications made application of standard evaluation measures such as Nielsen's heuristics problematic; they proposed a set of heuristics for use in VR application evaluation [13]. Similarly, Forsell and Johansson propose a set of heuristics for evaluating InfoVis systems that account for the characteristics of such systems [5].

The core selection criteria were based on the requirement for ease of data comprehension and task actions. These system features are underpinned by an intuitive control scheme, and data presentation approach that minimises operational cognitive load and training time.

3.2 InfoVis Heuristics

Heuristics in this section were taken from those proposed and validated in [5].

Information coding. Perception of information is directly dependent on the mapping of data elements to visual objects. This should be enhanced by using realistic characteristics/techniques or the use of additional symbols. ⟹ Correct perception of the information is important for interpretation and taking appropriate actions.

Minimal actions. Workload with respect to the number of actions necessary to accomplish a goal or a task. ⟹ Fewer actions means more efficient task completion and less cognitive load in order to do so.

Flexibility. The number of possible ways of achieving a given goal. ⟹ Having more ways to achieve a goal allows the user to select the most appropriate for a particular task. However, this does have a cost in terms of training time to learn the available affordances.

Spatial organization. Users' orientation in the information space, the distribution of elements in the layout, precision and legibility, efficiency in space usage and distortion of visual elements. ⟹ The spatial characteristics and relationships between data features are important to be easily perceived.

Consistency. The way design choices are maintained in similar contexts, and are different when applied to different contexts. ⟹ Data presentation must be consistent in all elements of the environment and navigation tools so they can be easily interpreted.

Recognition rather than recall. The user should not have to memorize a lot of information to carry out tasks. Minimize the user's memory load by making objects, actions, and options visible. ⟹ Requiring recall increases cognitive load.

Remove the extraneous. Whether any extra information can be a distraction and take the eye away from seeing the data or making comparisons. ⟹ Correct, straight forward perception of the data important to perform tasks with the data and navigate the environment. Too much information increases cognitive load and learning effort.

3.3 VR Heuristics

Heuristics in this section were taken from those proposed and validated in [13].

Compatibility with the user's task and domain. the VE and behaviour of objects should correspond as closely as possible to the user's expectation of real world objects; their behaviour; and affordances for task action. ⟹ The data corresponds to objects in the real world, and the tasks undertaken will correspond to real world actions.

Natural expression of action. the representation of the self/presence in the VE should allow the user to act and explore in a natural manner and not restrict normal physical actions. ⟹ The control system should be as intuitive as possible.

Close coordination of action and representation. the representation of the self/presence and behaviour manifest in the VE should be faithful to the user's actions. Response time between user movement and update of the VE display should be less than 200 ms to avoid motion sickness problems. ⟹ Observed action should match the action taken, this directly relates to intuitiveness of the control scheme, i.e., a performed control input has an expected system output. Avoiding motion sickness is key for being able to use VR for tasks uninterrupted.

Faithful viewpoints. the visual representation of the virtual world should map to the user's normal perception, and the viewpoint change by head movement should be rendered without delay. ⟹ Visual distortions would affect the perception and comprehension of the sensor data.

Navigation and orientation support. the users should always be able to find where they are in the VE and return to known, preset positions. ⟹ The spatial components of presented data makes navigation key for data comprehension and task performance.

Consistent departures. when design compromises are used they should be consistent and clearly marked, e.g. cross-modal substitution and power actions for navigation. ⟹ Minimise confusion while facilitating ease of use.

3.4 Evaluation Procedure

From Nielsen's original HE methodology [9] we have selected to implement observer transcription and compilation of verbally delivered expert evaluations (rather than evaluator completed written reports). To do so evaluators are given a set of typical user tasks to complete in the VR environment, and asked to perform them in an exploratory fashion to test system support for task performance. They are then given the set of heuristics, and are asked to verbally report identified problems as they are encountered; evaluators remarks are recorded for later transcription. For this stage the heuristics are used as prompts to guide evaluators in their examination of the system as during task performance evaluators are unable to look at them. Evaluators need not remember more than the rough outlines of the heuristics as identified problems will be fitted to heuristics by an observer during transcription of their recorded verbal feedback. Following task performance evaluators give additional feedback on the system while being able to review the heuristics. This post-hoc feedback aims to catch any problems that might have not been reported during task performance due to forgotten heuristics, or problems that need more consideration of the system as a whole. As with the in-task evaluations, post-hoc evaluation is given verbally, and transcribed by an observer. Transcribed reports from all evaluators are compiled into a single report and analysed to formulate recommendations for system improvements.

A key feature of this approach to HE is that the process is relatively unobtrusive to evaluators experiencing the system like a user would, while still providing structure for the feedback. In particular the immersive nature of VR is maintained during in-task evaluation. As the onus for fitting problems to heuristics and report compilation is on an observer, it should improve consistency of assignment of problems to heuristics. Additionally, time commitment of the experts is reduced, making the process more cost efficient (assuming experts are compensated for their time). However, there are some

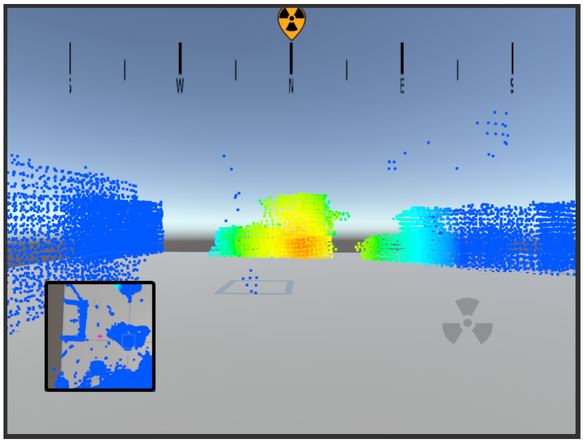

Figure 2: User view in radiation view mode, the highest radiation level is red radiation decreasing along the standard hue spectrum to blue for background radiation levels.

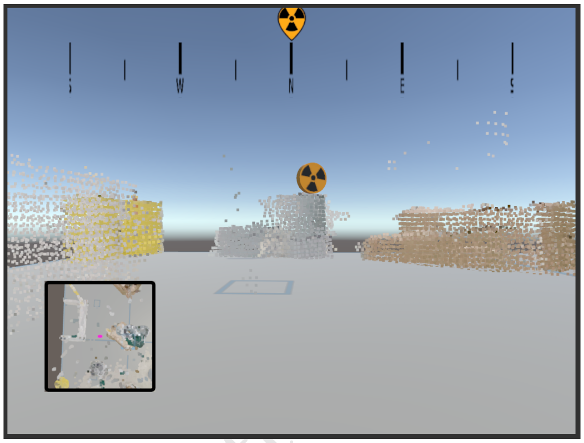

Figure 3: User view (RGB view mode) showing HUD elements: compass bar and navigation marker (top); map (bottom-left). A floating radiation marker can also be seen.

limitations: primarily that there may be observer error in report interpretation; relatedly, there is no severity ratings from problems reported (as there is in written reports), so the importance of issues is up to the interpretation of the observer.

4 CASE STUDY

In order to demonstrate the application of our approach to HE for VR-InfoVis systems we detail here a case study of its use for our nuclear decommissioning data presentation system. It allows us to demonstrate the process of conducting the expert system review, and how to produce a structured report based on evaluator transcripts that can result in actionable system improvements. Four expert evaluators were recruited from the University of the West of England Immersion network of academics working in the field of virtual reality. Within the field of VR the application context, and thus, area of particular expertise varied across evaluators. The domains of expertise were: interactive art design, computer game design, application in therapy, application in data visualisation. Hence, our recruited experts provided a broad spectrum of lenses through which the HE was conducted.

4.1 System description

Our in development system for robot mapping of legacy nuclear decommissioning sites generates a map consisting of a 3D pointcloud of spatially located sensor readings. In this initial work we only use the data from two sensors: an RGB-D camera and a collimated radiation sensor. Thus, each point has a 3D location, an RGB value, and a radiation value. The pointcloud data (PCD) is used to create a 3D environment using the Unity game engine[1] where each point is rendered as a mesh vertex (points are arbitrarily assigned to container meshes for operational simplicity). Two matched sets of vertex points are coloured depending on the sensor associated with the set: RGB value from the RGB-D sensor, radiation values normalised and mapped to colours following a standard heatmap paradigm (see Figure 2). The visible point set can be switched between by the user to allow viewing of the different data related to observed points. Increased pointsize of the vertices is used to make object perception easier in the case of a relatively sparse pointcloud. One of the final goals of the system is to enable pointclouds to be updated with new data as the robots continue to sense their environment. Hence, we are interested in using the data only as points rather than investigating auto-meshing due to performance issues, particularly related to accurate perception of the data (incorrect meshing will create distortions in data perception), and processing resources required for re-rendering large environments.

For the purposes of HE we created a small (relative to that planned for the user study) initial test environment consisting of a range of objects captured using an Intel RealSense D415 RGB-D camera. The objects used aimed to be representative of the sort of industrial objects found at legacy nuclear sites, though at a quarter scale. The camera was moved around by hand[2], and a composite pointcloud map of the whole environment was generated using the RTabMap ROS package [1]. The resulting pointcloud was scaled up to create an environment with objects of the needed size. A radiation source point was selected, and radiation values for all points were procedurally generated (for the generation of the radiation map): the radiation source was given a radius of contamination, with radiation most intense at the centre, and decreasing linearly with distance, down to background levels at the edge of the radius.

To aid navigation areas of raised radiation are denoted by floating markers present in RGB view mode. These markers are then utilised within a set of 'heads-up-display' (HUD) style navigation elements that overlay the VR camera viewpoint (see Figure 3). The HUD consists of a top down map view of the environment centered on the user, and a compass bar to indicate the relative direction of navigational goals (e.g., radiation markers). These display elements are based on common paradigms used in computer games, where

---

[1] www.unity.com

[2] Work on the multi-robot SLAM system is ongoing, so we have used artificial data for this initial visualisation system test, doing so will inform design considerations for the real system.

they are supporting similar navigational tasks as we have in our system. We are investigating whether the utility, and intuitiveness of these tools translate between application contexts, i.e., from games to InfoVis.

Our current implementation uses the HTC Vive Pro, hence the control system for navigating in the environment utilises the Vive controllers (though the principle of the controls could be mapped to other controllers). The left hand d-pad controls altitude, and body rotation (continuous rotation of the virtual body, head orientation is relative to body direction). The right hand d-pad allows translational movement on the ground plane relative to the direction of gaze. The right hand trigger is used for movement by teleportation (a common VR paradigm), shooting an arc to a teleport location that is travelled to on trigger release. The left hand trigger is used for changing view mode, currently this toggles between RGB and radiation coloured pointcloud visibility; in radiation view mode a semi-transparent radiation symbol is added to the HUD (Figure 2).

4.2 Task Description

The principle purposes of our system is data comprehension and radiation source characterisation. Radiation source characterisation is an important step in the decommissioning process whereby sources of radiation must be identified, along with identification of objects in the environment contaminated by radiation. Hence, the tasks to be performed for system evaluation are navigation to a radiation source, and identification of objects contaminated by radiation, including determining which object is the source. This task set is performed by the expert evaluator, as well as forming the basis for the tasks to be performed in a future user study.

4.3 Instructions to Evaluators

Expert evaluators are given an initial tutorial in a sparse environment so that they can gain an initial understanding of the affordances of the system that they are to evaluate. The tutorial is conducted prior to explaining the heuristics and HE procedure, which, combined with the sparse environment, aims to minimise pre-evaluation opinions of the system. Following this tutorial they are given the set of heuristics to read and instructed as to the tasks and HE procedure. Included in the explanation is context details including that the environment is created from the sensor data from a small robot, and that we are interested in if people are able to comprehend and operate in point cloud environments.

4.4 Results and Analysis

In order to synthesise a report on the evaluations of the four VR experts the transcript of each was first separated into distinct comments relating to a particular problem; some comments from evaluators identified good features rather than problems, these are included in the results analysis as useful for future system evaluations and development. Each comment was then attributed to one (or more) heuristics. In the following sections a summary of the comments for each heuristic are presented, along with an analysis of the design implications and potential system developments to address them.

Information coding. Two comments were made regarding noise inherent in the data output by RTabMap and rendered as part of the pointcloud. Firstly that floating points (i.e., those seperate from objects) seemed strange, and secondly that some of the noise could be assumed to be part of objects in the environment. Filtering the noise at the rendering stage is a non-trivial problem, particularly given our intent to dynamically update the map data as the robots continue to explore the environment. Indeed, it is one of the aims of our follow up user study to test if users can mentally filter out this noise and interpret the data correctly.

One evaluator reported that the pointcloud was too coarse for good object recognition, and commonly the evaluators struggled to identify some objects in the environment. To address these issues we are looking at two solutions, firstly to have RTabMap output a denser cloud (it downsamples the data from the RGB-D camera to constrain the number of points in the final map), secondly to use voxels rather than mesh vertexes to render the data. Voxels have the added advantage that we can have more control over their appearance, and filter out noise as data from several close points is displayed as one voxel.

Finally one evaluator reported that the lack of collisions with the pointcloud could misrepresent data as not belonging to solid objects. This can be addressed if we opt to use voxels for data representation as they allow procedural generation of colliders. However, in the first instance we will be evaluating whether this lack of collisions is a problem in a follow up user study.

Minimal actions. Two evaluators reported that they *strafed* around objects (i.e., sideways translation with body rotation to keep an object in view) in order to take advantage of the parallax effect to observe the relative location of points. This use of the navigation affordances to better understand the data shows promise for presenting data in this way. However it goes somewhat against the minimal actions heuristic, and could add significantly to operation time with a larger environment. The need to do this rotation could be reduced by using voxels to render the data: as 3D objects lighting effects should allow some observation of relative positions without the need to move as much.

Flexibility. Two evaluators suggested alternative means by which the data could be observed, this would add flexibility to the system. One suggested method was to have procedurally generated teleport points around objects of interest so that they could be more easily viewed from different directions. The second method was to allow section of groups of points that could then be manipulated for inspection. Both of these methods are interesting alternatives that may be of use. However, they are substantial deviations from the current design and would merit investigation into their utility before integrating into the system.

Spatial organization. One evaluator commented that the addition of a virtual floor aided in interpretation of noisy floor data points, i.e., ones that were not all on one level due to pointcloud fusion errors. This highlights the benefit of augmenting the data with visual references expected by users.

Consistency. One evaluator described their process for identification of objects of which they were uncertain: they used comparisons with other more recognisable objects in the environment. This highlights the importance in consistency of representation, needed for such comparisons to be useful.

Recognition rather than recall. Though one evaluator noted the utility of vertical travel, they did not always remember to use it. Relatedly another observer often forgot how much they had moved vertically so lost their sense of scale. Both of these issues might be addressed with the addition of a display of the current height on the HUD. Hints of available affordances might also be displayed, however, these would need to be unobtrusive, and will likely become irrelevant as users become experienced with the system.

One evaluator noted the need for a reminder as to the current task and procedure during system operation. Such reminder information could easily provided with an action trigger associated with it.

Remove the extraneous. No reported problems fitted this heuristic.

Compatibility with the user's task and domain. The majority of evaluators noted that identifying objects from the rendered pointclouds was made more challenging due to a lack of knowledge of what objects to expect. This can be easily addressed by adding pictures of objects typically found in the environment to the initial instructions. This maintains ecological validity as end users for the system would have knowledge of typical objects that appear in the environment, or could be provided with pictures taken in the environment surveyed by the robots.

One evaluator noted that a possible reason the controls were intuitive to him was his prior computer gaming experience, and end users might not have a similar frame of reference. To evaluate if gaming experience impacts system usability we will be including questions on computer gaming experience in the demographics questionnaires of future user studies of the system.

One evaluator commented that the expectation on reaching the radiation marker (the navigation goal) they were expecting some ability to mark task completion. Adding the facility to mark navigation as completed will not only help to match user expectations, but also facilitate evaluation of task completion times for future user studies. Moreover, marking of navigation accomplishment could be used to automate recording of characterisation of the reached radiation source.

Natural expression of action. Two evaluators noted that strafing (sideways translation with body rotation) induced feelings of motion sickness, though they continued to do so as an intuitive way of moving around in the environment. One possible reason for this behaviour is that it was induced by the available motion controls. A solution we will be investigating is to modify body rotation to be in discrete jumps (by clicking the edges of the dpad) rather than continuous motion: it is hoped this will lead to less nausea inducing behaviour.

All evaluators noted the intuitiveness of the controls. This is a promising finding for our planned user study.

Close coordination of action and representation. One evaluator had problems with the dpad, both with some apparent inaccuracies of use (though they were able to resolve this over time) and in over sensitivity. These controller limitations will need to be further tested to see if they are problematic beyond the initial time a user has with the system.

Faithful viewpoints. The majority of evaluators reported issues relating to camera height: having to make assumptions as to what height their viewpoint corresponded to, both small robot and human were reported; the lack of height information made it hard to determine the size of objects, and this made identification difficult. Two possible solutions to this issue could be investigated, one is some sort of altimeter displaying their current camera height, the other is displaying a reference object of a known height to compare objects to.

Navigation and orientation support. Though one evaluator noted that the compass bar and marker system was intuitive, another found it unclear how it might be used. This highlights failings in the tutorial for those whom have no frame of reference for such navigation aids. Hence, the tutorial will be redesigned for future studies to ensure all users have as similar levels of understanding as possible.

Consistent departures. As mentioned under the flexibility heuristic, two evaluators suggested alternative means for viewing the data which are clearly departures from naturalistic viewing of the data. In addition to the points made previously the evaluator suggesting the viewing teleport points noted that the disadvantage of such a system would be the difficulty of tracking points from one view to another. This reinforces our suggestion that such design ideas need to be tested as independently as possible for their utility.

5 DISCUSSION

The primary desired outcome from HE is formative feedback that can be utilised as part of an iterative design process to improve the efficacy of UI designs. The specific outcomes for our particular system and application of HE are detailed above; here we discuss some more generalisable conclusions that can be drawn from our results analysis.

It is important to avoid violating assumptions that a user might make about a UI, as it will impact how intuitive the system is. Moreover, in VR doing so can impact user immersion, and this likely also has knock-on effects for task performance and data perception. However, determining what a user might assume about a particular system is challenging to fully achieve a priori. Hence, we suggest user studies that probe for user assumptions of system prototypes would be of value early in the design cycle.

In determining if particular evaluator comments need to be actioned, it was important to consider how much of a deviation they suggested the system was from the heuristics (bearing in mind system design goals). Additionally, we suggest that comments that require radical changes to system design need to be carefully considered, and likely result in iterative investigations of the utility of potential new system features.

When utilising the approach of transcribed verbal feedback, a key part of the HE process is assignment of comments to heuristics for analysis. It is instructive to reflect on this process of assigning evaluator feedback, and analysing comments with reference to the heuristic to which they are assigned. Although the process of heuristic assignment introduces some noise to the HE process, requiring as it does the judgement of evaluator comments post-hoc, it is essential to the principled analysis of evaluator feedback. In the results presented above there was several instances where a comment could be logically assigned to more than one heuristic (due in part to some conceptual overlaps with heuristics from the two domains from which they were selected). Indeed in several instances comments were analysed with regards to more than one heuristic. Hence, it could be considered that post-hoc assignment allows fuller, less constrained feedback from evaluators, albeit harder to analyse. Additionally, as the analysis and report synthesis was conducted by a system designer, it resulted in a more in depth reflection of how well the system matched up to the selected design heuristics. One limitation of our approach was that there could be errors in consistency of heuristic assignment due to reliance on the judgement of a single observer. Hence, in future work we will be investigating the use of multiple observers with results analysed for consistency, and combined to synthesise a single report.

Additional useful outcomes, beyond the primary aims of HE were forthcoming from the analysis of the evaluator transcripts. While evaluators were only instructed to comment on problems that they encountered, they additionally gave positive feedback on UI features that worked, and design suggestions to overcome particular problems they observed. Both of these types of feedback are useful in determining how design iteration might be undertaken. Similarly useful and tangential to the primary instructions, some evaluator comments related to features connected to task and instruction design, i.e., protocol for the follow up user study. This highlights the importance of appropriate task selection, particularly with our intention to complete our system evaluation with a follow up user study. In future work we will investigate how the HE procedure might be modified to encourage and structure such feedback to better take advantage of observers expertise.

6 CONCLUSION

In this paper we have detailed our approach to Heuristic Evaluation for Virtual Reality Information-Visualisation systems. By composing a set of heuristics through principled selection from an established set in each domain (VR and InfoVis) we were able to demonstrate how to apply HE to the aforementioned hybrid domain. We then describe our method for conducting HE in VR while maintaining evaluator immersion, and demonstrate its application in our use case of nuclear decommissioning robot sensor data presentation. We suggest that our methodology has applications to other VR InfoVis contexts; one caveat being it is important to consider the applicability of a given heuristic to a particular context, and re-selection of heuristics might be required for particular contexts.

Our case study allowed us to clearly demonstrate the process of evaluator feedback analysis and report synthesis. Further, it allowed us to reflect on our process, and come to some generalisable conclusions about key factors in the design of VR InfoVis systems.

In future work we will look to refine our HE process using multiple observers for report synthesis, and extend HE to formalise the process of human-factors study piloting.

7 ACKNOWLEDGMENTS

This work was funded by the EPSRC in the Robotics for Nuclear Environments project under grant number EP/P01366X/1.

We would like to thank the UWE academics from the Immersion Network for their participation as expert evaluators.

REFERENCES

[1] [n.d.]. rtabmap_ros - ROS Wiki. http://wiki.ros.org/rtabmap{_}ros
[2] Cedric Bach and Dominique L. Scapin. 2010. Comparing inspections and user testing for the evaluation of virtual environments. *International Journal of Human-Computer Interaction* 26, 8 (jul 2010), 786–824. https://doi.org/10.1080/10447318.2010.487195
[3] Heather Desurvire and Max Kreminski. 2018. Are game design and user research guidelines specific to virtual reality effective in creating a more optimal player experience? Yes, VR PLAY. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10918 LNCS. Springer Verlag, 40–59. https://doi.org/10.1007/978-3-319-91797-9_4
[4] Ciro Donalek, S. G. Djorgovski, Alex Cioc, Anwell Wang, Jerry Zhang, Elizabeth Lawler, Stacy Yeh, Ashish Mahabal, Matthew Graham, Andrew Drake, Scott Davidoff, Jeffrey S. Norris, and Giuseppe Longo. 2015. Immersive and collaborative data visualization using virtual reality platforms. In *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*. Institute of Electrical and Electronics Engineers Inc., Waishigton, 609–614. https://doi.org/10.1109/BigData.2014.7004282
[5] Camilla Forsell and Jimmy Johansson. 2010. An heuristic set for evaluation in Information Visualization. In *Proceedings of the Workshop on Advanced Visual Interfaces AVI*. ACM Press, New York, New York, USA, 199–206. https://doi.org/10.1145/1842993.1843029
[6] Setia Hermawati and Glyn Lawson. 2016. Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? , 34–51 pages. https://doi.org/10.1016/j.apergo.2015.11.016
[7] Ebba Thora Hvannberg, Gyda Halldorsdottir, and Jan Rudinsky. 2012. Exploitation of heuristics for virtual environments. In *NordiCHI 2012: Making Sense Through Design - Proceedings of the 7th Nordic Conference on Human-Computer Interaction*. 308–317. https://doi.org/10.1145/2399016.2399065
[8] J Nielsen. 1994. Usability Inspection Methods. In *New York: John Wiley & Sons*, J Nielsen and R.L Mack (Eds.). Chapter Heuristic.
[9] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 249–256. https://doi.org/10.1145/97243.97281
[10] Eduardo Jose Goncalves De Oliveira, Francisco Paulo Magalhaes Simoes, and Walter Franklin M. Correia. 2017. Heuristics evaluation and improvements for low-cost virtual reality. In *Proceedings - 19th Symposium on Virtual and Augmented Reality, SVR 2017*, Vol. 2017-Novem. Institute of Electrical and Electronics Engineers Inc., 178–187. https://doi.org/10.1109/SVR.2017.31
[11] Martin Schmettow, Cedric Bach, and Dominique Scapin. 2014. Optimizing usability studies by complementary evaluation methods. In *Proceedings of the 28th International BCS Human Computer Interaction Conference: Sand, Sea and Sky - Holiday HCI, HCI 2014*. BCS Learning and Development Ltd., 110–119. https://doi.org/10.14236/ewic/hci2014.12
[12] Daniel Smith, Jasmine Granados, and Joel Suss. 2019. Evaluating the Comprehensiveness of VR PLAY Guidelines Using Elder Scrolls : Skyrim VR. In *Proceedings of the Human Factors and Ergonomics Society 2019 Annual Meeting*, Vol. 63. 2287–2291. https://doi.org/10.1177/1071181319631164
[13] Alistair Sutcliffe and Brian Gault. 2004. Heuristic evaluation of virtual reality applications. *Interacting with Computers* 16, 4 (aug 2004), 831–849. https://doi.org/10.1016/j.intcom.2004.05.001
[14] World Nuclear Association. November 2018. Decommissioning Nuclear Facilities. http://www.world-nuclear.org/information-library/nuclear-fuel-cycle/nuclear-wastes/decommissioning-nuclear-facilities.aspx Last accessed 11 February 2020.